集成光学二十年

JI CHENG GUANG XUE SAN SHI NIAN

陈益新 著

上海交通大学出版社

集成光学三十年

陈益新 著

上海交通大学出版社

图书在版编目(CIP)数据

集成光学三十年/陈益新著. - 上海:上海交通大学出版社,1999

ISBN 7-313-02238-7

Ⅰ. 集… Ⅱ. 陈… Ⅲ. 集成光学 - 现状 Ⅳ. TN25 - 1

中国版本图书馆 CIP 数据核字(1999)第 24014 号

集成光学三十年

陈益新 著

上海交通大学出版社出版发行

上海市番禺路 877 号 邮政编码 200030

电话 64281208 传真 64683798

全国新华书店经销

常熟市印刷二厂·印刷

开本:787×1092(mm) 1/16 印张:19.5 字数:476 千字

版次:1999 年 10 月 第 1 版

印次:1999 年 10 月 第 1 次

ISBN 7 - 313 - 02238 - 7/TN·079

定价:32.00 元

前　言

自1956年起，我在上海交通大学执教40余年，教学工作和科学研究先后涉及的领域和学科有电力电缆、通信电缆、电气绝缘材料、电介质物理、高分子物理、超导材料及器件（如超导磁铁、超导薄膜和器件）、半导体器件（如高频硅可控整流器）、固态电子学、集成光学、光电子器件（如锗雪崩光电二极管、Ⅲ－Ⅴ族半导体光晶体管、半导体超晶格MSM光电探测器）、光纤通信、光计算机和光互连、垂直磁记录和薄膜磁头、微加工、微机械以及纳米技术等。单独编写、主编或参加编写的教材和专著有《电介质物理》、《高分子物理》、《集成光学——理论和技术》、《固态电子学》、《固体电子学中的等离子体技术》、《光计算》以及《毫微加工——物理、技术、应用》等，发表论文120余篇，培养了硕士和博士研究生共50名。

在许多同事、同学和学生们的建议和鼓励下，我挑选了1980年以来有一定代表性的章节和论文编成这部选集。一时想不出合适的书名，只能借用最近应邀撰写的一篇综述性论文的题目：集成光学三十年。不过，就全书的内容而言，除了垂直磁记录和薄膜磁头这部分偏离较远外，其余的内容与集成光学及其相关领域都有不同程度的联系。因为集成光学从其基本理论、器件结构、工作原理、材料和加工、测试和应用所涉及的学科和领域是十分广泛和相互交叉的。实际上，即使垂直磁记录也不是与集成光学毫不相关的。记录信息的磁畴与介质表面垂直，从物理上看，它不仅可由磁场直接磁化形成，也可以在较弱偏置磁场下由激光束加热形成，这就是磁光记录的原理。磁光记录中信息的写入和读出，都由受控制的聚焦激光束来实现，因而"光头"相当于磁记录装置中的磁头，是磁光记录装置中的核心部件，如果这光头能用集成的微光电机系统来构成就可使光头的结构更为紧凑、可靠、体积小、质量轻，提高响应速度，适合批量制造，降低成本。

本书的内容分二部分，即教学篇和研究篇。前者都选自出版的教材和专著中由本人撰写的对不同学科的领域有一定代表性的章节。后者选自主要由本人撰写的论文和报告，其中大部分已在期刊或学术会议上公开发表过，只有少数几篇是第一次发表。为了使读者查阅方便，研究篇先按内容分成集成光学和光纤通信、光计算和光互连、垂直磁记录和薄膜磁头以及微米纳米技术及其应用四部分，每一类中大致以时间先后分序。

本书不仅可作为在党和人民培育下，在社会主义制度下成长起来的一个教育工作者和科技工作者毕生的工作总结和汇报。同时，也从一特定的角度上多少能示踪出近二十年来我国高等教育和科学研究发展所经历的过程和取得的进步。发展像集成光学这类多学科交叉的高新科技，需要把自然科学与工程技术熔合成一体。培养新一代的应用科学人才就是要善于在这两者间架起桥梁。希望本书也能对作者这一贯的教育思想有所反映。由于本人学术水平有限，对书中的差错和不当之处，欢迎读者批评指正。

本书所包含的内容决不全是我个人的贡献，它也凝聚了我的许多同事和学生们长期辛勤的工作和巨大的努力。在出版过程中又得了上海交通大学领导和有关部门的多方关心和支持。作者对此一并表示衷心感谢。

陈益新

序

《集成光学三十年》是陈益新同志40多年来在上海交通大学从事教学工作和科学研究工作的辛勤劳动成果。1952年我国高等学校进行院系调整后，交通大学设置了一些新专业，电气绝缘与电缆技术专业就是其中之一，这个专业需要电工和电子、物理、化学以及机械等多学科交叉的知识。陈益新同志是这个专业第一届毕业生，他毕业后留校任教，师从我国著名的无线电前辈陈季丹教授，担任电介质物理的教学和科研工作。1979年，他作为我国第一批访问者前往美国圣地亚哥加州大学，在国际知名教授张慎四博士的指导下进行集成光学研究。1980年底回国后，他作为上海交通大学集成光学的学科带头人带领了一批青年教师和学生拼搏在教学科研第一线，取得了显著成绩。他致力于把我国集成光学的研究推向当前国际上发展的前沿，他不断在国内学术会议上率先提出新的研究方向，如三维集成光学、非线性集成光学和微光电机集成系统等。同时他也积极组织国内和国际的学术交流，以进一步推动我国集成光学的发展。他的辛勤工作，得到了国内外许多同行学者的高度评价。陈益新同志已成为推动我国集成光学学科发展的先驱者之一。

陈益新同志在上海交通大学任教期间，根据工作需要，多次负责筹建新学科，他先后从事的学科领域有：电气绝缘、电线电缆、电介质、高分子、半导体、超导、磁记录、光电子和集成光学、固态电子学、微加工、微机械和纳米技术等。他理论基础坚实，知识面较广，富有开拓精神；他在教学中努力使理科与工科的知识相互渗透，积极倡导理工结合；他在科学研究中重视理论和实际应用相结合；他在学生中反复强调要善于在自然科学与工程技术之间构筑桥梁，这样学生才能更好发挥创造性。他的教学思想和学术观点在科学技术突飞猛进的今天，对培养面向21世纪的科技人员是具有积极意义的。本书主要选辑了这一时期的作品，读者不难从本书的各部分内容中体会到上述学术思想和观念。

《集成光学三十年》书名虽似偏专业化，但本书内容覆盖的学科相当广泛。内容涉及：物理、化学、材料学、电子学、计算机、光子学和光电子学、集成光学和光纤通信、磁记录、固态电子学、微加工、微米和纳米技术等。相信本书的出版将会引

起多方面读者的兴趣，包括高等学校的教师、高年级大学生和研究生，研究机构和工厂企业的科技人员，乃至教育科技管理部门的干部等。读者不仅可在具体的知识内容上得到益处，也能从如何加速自我知识更新，跟上科技发展日新月异的进展，树立适应新时代的教学和科学研究新观念方面有所启迪。

中国科学院院士
上海交通大学教授　张照

1999 年 5 月

目　　录

第一篇　教学篇

固态电子学是沟通科学和工程的桥梁 …… 3
电介质的电子性质 …… 5
集成光路及其应用 …… 54
光计算导论 …… 93
微米纳米加工导论 …… 118

第二篇　研究篇

集成光学和光纤通信

Planar and Channel Single-Mode $LiNbO_3$ Waveguides Fabricated by Ion Exchange …… 159
Characterization of $LiNbO_3$ Waveguides Exchanged in $TlNO_3$ Solution …… 160
$LiNbO_3$ Waveguides by Electrically Enhanced Ion Migration and a Comparison of Techniques … 164
A New Concept of 3-Dimensional Integrated Optics …… 166
Nonlinear Integrated Optics …… 168
光通信传输速度的进展引人注目 …… 180
以网络和应用为导向的光纤通信 …… 184
集成光学三十年回顾及展望 …… 189

光计算和光互连

二维空间光调制器的研究和应用 …… 202
二维 Si/PLZT 混合集成空间光调制器 …… 209
未来计算系统中的光互连 …… 214
光互连技术 …… 215

垂直磁记录和薄膜磁头

磁记录技术的现状和趋向 …… 226
垂直磁记录的新进展 …… 235
A New Concept of Magnetic Thin-Film Heads with Superconductors …… 243
Characterization of the Film Heads for Perpendicular Magnetic Recording …… 248
The Thin Film Magnetic Recording …… 250

微米/纳米技术及其应用

新世纪的纳米科学技术 …… 258
近场扫描光学显微镜中探针优化设计的考虑 …… 277
微光电机械系统的技术和应用 …… 283
Thin Film Technologies for Micro-Opto-Electro-Mechanical System Applications …… 292

第一篇　教　学　篇

固态电子学是沟通科学和工程的桥梁

固态电子学是一门以固体中的电子过程为基础，阐明固态电子材料的各种特性和效应，以及研究如何利用这些特性和效应构成固态电子器件的基本原理的应用基础学科。以晶体管和半导体激光器为代表的固态器件的发明和发展而导致的集成电路、电子数字计算机、光纤通信、集成光路和光电子集成等许多应用技术领域的突飞猛进，成为当代信息技术的主要支柱。这是把固态物理和近代物理学中的成就与电气和电子工程应用相结合的最好实例。固态电子学是以沟通科学和工程的桥梁而获得了强大的生命力。自从60年代开始在国外一些著名的大学中开设这门课以来，一直受到越来越广泛的重视，在内容上也有很大的充实和更新；它已成为电气和电子工程、材料科学和工程、应用物理和应用化学等许多方面的工程技术人员及学生从固态电子器件的角度出发对固态电子材料的电、磁、光、声和热等各种特性和效应及其机理进行研究，并将这些基本科学知识转变为技术进步和创造发明的不可缺少的课程。

由于历史原因，我国高等学校的课程中至今只有“固态物理学”，从未开设“固态电子学”。而在有关的专业课程中，对固态电子材料和器件的学习，几乎都是分门别类的，例如半导体材料和器件、电介质材料和器件、磁性材料和器件以及激光材料和器件。这样造成的后果往往使学生对固体中的电子运动和对固态电子材料和器件及基本物理过程缺少完整和统一的了解，知识面不广，思路比较局限，不利于培养学生的创造性。特别应看到的是随着技术的不断进步，固态电子器件已从应用单功能材料进而应用多功能材料或复合功能材料，从利用单一效应发展到利用多种效应，例如：采用铁电晶体薄膜代替MOS半导体器件中的氧化层电介质，就可获得非挥发性的半导体存储器；光子器件与电子器件的集成不仅可以大大改善器件的原有性能，而且还包括能达到单一器件所不可能获得的功能；在未来的集成光路中不仅包括了光电子集成的光源和探测器，而且还利用了如电光效应、声光效应、磁光效应和热光效应等许多相关效应所构成的调制器和其他器件。最大限度地利用固体材料的电、磁、光、声和热等自效应和相关效应将是固态电子材料和器件发展的一个重要趋向。

出于上述的认识，我们从1981年开始，参照了国外一些“固态电子学”的教科书以及有关期刊，对物理系的高年级大学生和一年级研究生开设了“固态电子学”课程，同时也作为其他系，如材料科学和工程系、电子工程系等的学生的选修课。7年的教学实践表明，“固态电子学”课程对扩大学生知识面，培养学生的创造性和提高学生对高科技工作的适应性是有帮助的，因而引起了更多师生的兴趣和重视。

由于迄今尚没有“固态电子学”的中文教科书，开始两次讲课都选用几本英文课本。为了进一步把这门课程提高和推广，就在历次讲课提纲和讲稿的基础上整理成“固态电子学”讲义，共分上下两册。本教科书是在讲义的基础上，根据多年的教学实践修改而成的。希望通过广泛使用和积累更多的教学经验，对其加以进一步的提高和完善。

本书适用于电工和计算机、电子工程、无线电和通信、材料科学和工程、应用物理、应用化学和生物工程等系科三、四年级大学生。由于这是一本带有引论性质的教科书，所以在内容上力求避免用较多的数学演算而能给出清晰的物理概念。只要具备一些近代物理和量子力学的

基本知识和基本概念，就不难学习这门课程。上述系科的低年级研究生也可选用本教材，教学中必要时可补充内容，学生也可多阅读一些每章所附的参考书和文献。适合于研究生用的高层次的"固态电子学"待取得更多经验后再行编写。

全书共分七章。第一章介绍金属和半导体电子学，这里所讨论的特性和效应都与电子的输运过程有关；第二章讨论利用这些特性和效应构成各种半导体器件的基本原理；第三章光电子和光电器件中所述及的那些效应和现象虽与电子输运过程有关，但起主导作用的往往是电子的跃迁过程，超导现象也可看作电子的输运过程，但与金属和半导体的导电特性差别极大，特别近年来高临界温度超导材料的发现，对超导现象和应用又提出了许多新的研究课题，这部分内容作为第四章；与电子的自旋和轨道运动过程相联系的磁性和磁效应及其在磁电子器件中的应用将在第五章阐述；束缚状态的电子发生位移时形成极化，当电场从直流变到光频以上（甚至可包括γ射线在内）时，极化过程所伴随着的各种效应都在第六章讨论；最后一章介绍固体中相关效应，主要讨论电介质中电、光、声和热效应之间的相互作用，这里虽然不能包括固体中全部的相关效应，但读者从这一章可以得到启迪，应该如何进一步发现和运用固体中一切可能出现的效应来发展新一代固态电子器件。贯穿全书各章的统一思想是以固体中电子运动的各种形式和过程作为认识的基础来理解固体所呈现的全部特性和效应，同时又把这些电子过程和效应的形成与固体微观结构的特征紧密相联系。因而，读者将获得对固态电子学，包括物质结构、电子过程、特性和效应、材料和器件较完整的认识。由于本书是一套固体物理教学参考书中的一本，为避免内容重复，在把讲义修改成书时，有关晶体结构、能带理论等一些详细内容已予删节。

电介质的电子性质

从能带结构来看，如果固体的价带与导带之间具有较宽的能隙，例如数电子伏或更高，这类固体几乎完全没有导电性，是电绝缘体，或叫电介质。在外电场作用下，原子、分子或晶体中的电子或其他电荷不能自由迁移，只能作有限的相对位移，形成电偶极，这叫做电介质的极化。如果电场以很高的频率交变、极化将引起能量的损耗和色散现象。当频率范围从工频、射频、微波直到光波范围，或波长更短时，电介质的极化和能量损耗可采用不同的宏观参数描述，如复电容率、复传播常数和复折射率等。从固体在交变电场下呈现极化特性这点来说，电介质的含意不能狭义地限于绝缘体，只要不是金属，或即使是金属在特殊情况下，都可看成是广义的电介质。

实际的电介质具有一定的漏泄，其所能承受的电场强度也不能超过某极限，否则电介质中的电流将剧增，引起电介质的损坏这叫电介质击穿。另外，电介质除了由于外场感应产生极化外，还可能由于分子或晶体结构的不对称，或由于应力和温度等外界条件造成极化，这些特性和效应有许多已在实际中广泛应用。

1 电介质极化

1.1 极化的宏观现象

当一平行板电容器在真空中充有电荷 Q_0、电极间电压 V，则这真空电容器的电容 C_0 为

$$C_0 = \frac{Q_0}{V} \tag{1}$$

电容 C_0 与电极的几何形状和尺寸有关，对平行板电容器来说，如果忽略边缘效应，C_0 可表示为

$$C_0 = \varepsilon_0 \frac{A}{d} \tag{2}$$

式中：A 为极板面积；d 为极板间距；ε_0 为真空电容率或介电常数。在国际单位制中，$\varepsilon_0 = 8.855\text{pF/m}$。

如果在电容器中充以电介质，外施电压保持不变，则可观察到电容器所充的电荷增大为 Q，表明电容器的电容增大为 $C = Q/V$。电容 C 的增加，不是由于电容器几何尺寸的改变，而是因为电介质的电容率 ε 比真空电容率 ε_0 大，$C = \varepsilon(A/d)$。电容器由于充入电介质使电容增加的相对值为

$$\frac{C}{C_0} = \frac{Q}{Q_0} = \frac{\varepsilon}{\varepsilon_0} = \varepsilon_r \tag{3}$$

式中 ε_r 是电介质的相对电容率或叫相对介电常数。

为什么当有电介质放进电容器后，电极上所带电荷会增加？这是由于电介质在电场作用

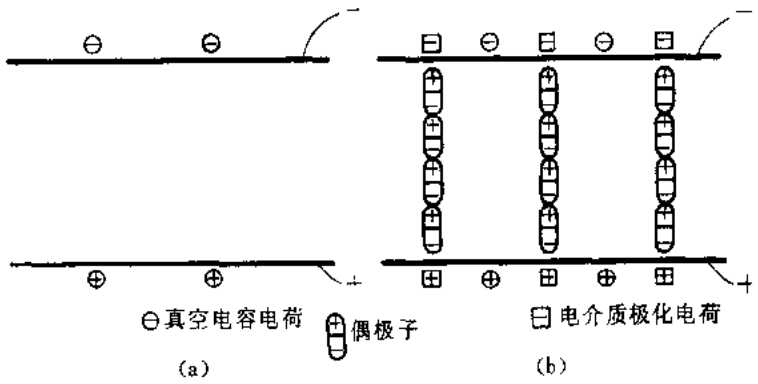


图1 真空中(a)和放入电介质(b)的平行板电容器

下极化，这极化在电介质表面形成的束缚电荷在电极上引起了附加的自由电荷，如图1所示。当电容器中放入电介质后，电极上的总电荷 Q 包括两部分：一部分为真空电容电荷 Q_0，$Q_0 = Q/\varepsilon_r$；另一部分电荷是由于放入电介质后，要补偿极化偶极子作用而增加的电荷

$$Q' = Q - Q_0 = Q(1 - 1/\varepsilon_r) \tag{4}$$

为了进一步说明这些电荷的物理意义及它们在空间产生的作用，我们在下面引进三个电场中的矢量来表示电荷密度，总电荷密度为

$$\sigma = \frac{Q}{A} \tag{5}$$

定义矢量 $\boldsymbol{D}$(电通密度或介电位移)表示电极上总电荷密度 σ，表面电荷密度与 $\boldsymbol{D}$ 的垂直分量相等，矢量从正电荷指向负电荷。与此相似，用 $\varepsilon_0\boldsymbol{E}$ 表示真空电容电荷密度 σ/ε_r，矢量 $\boldsymbol{E}$ 称电场强度，矢量 $\boldsymbol{P}$ 叫极化强度，用来表示电介质极化后在电极上引起的附加电荷密度 $\delta(1 - 1/\varepsilon_r)$，它等于电介质表面极化电荷密度。

从 $\boldsymbol{D}$ 和 $\boldsymbol{E}$ 的定义可知，它们具有关系：

$$\boldsymbol{D} = \varepsilon_0\varepsilon_r\boldsymbol{E} = \varepsilon\boldsymbol{E} \tag{6}$$

从式(4)可以导出电介质极化的宏观关系式为

$$\boldsymbol{P} = \boldsymbol{D} - \varepsilon_0\boldsymbol{E} = (\varepsilon - \varepsilon_0)\boldsymbol{E} = (\varepsilon_r - 1)\varepsilon_0\boldsymbol{E} = \chi\varepsilon_0\boldsymbol{E} \tag{7}$$

式中 χ 为电介质的电极化系数，$\chi = \varepsilon_r - 1$。χ 等于电介质极化后引起的附加电荷与真空电容电荷之比值，χ 和 ε_r 相似，它们都是表示电介质极化能力的宏观参数。当 $\varepsilon_r \gg 1$ 时，$\chi \approx \varepsilon_r$。

三个矢量中的 $\boldsymbol{D}$ 和 $\boldsymbol{P}$ 具有单位面积上的电荷，即电荷密度的量纲，而 $\boldsymbol{E}$ 具有不同的量纲与物理意义，根据静电学原理，试验电荷 q 在电场强度为 $\boldsymbol{E}$ 时受到的作用力为

$$\boldsymbol{F} = q\boldsymbol{E} \tag{8}$$

因而 $\boldsymbol{E}$ 的大小和方向与单位电荷所受力的大小和方向相同，由此可以得出介电常数的量纲为

$$[\varepsilon] = \left[\frac{\text{单位面积上的电荷}}{\text{单位电荷所受的力}}\right]$$

1.2 极化的基本形式[1]

电介质极化的宏观现象是在表面出现极化电荷。电介质极化的微观过程是其中的分子形成电矩或叫偶极矩 $\boldsymbol{p}$。偶极子是等量的正负电荷相对位移形成的，偶极矩是表示偶极子基本

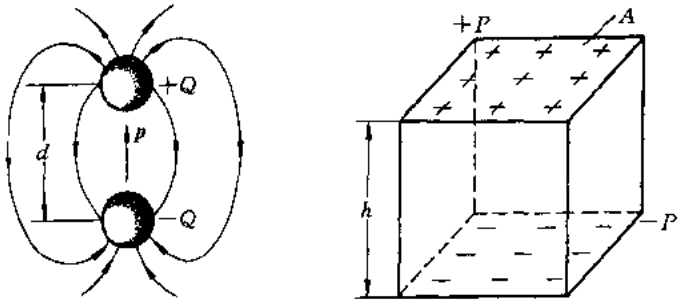


图 2　偶极子(a)的电矩和电介质的电矩(b)示意图

特性的参数,其大小等于电荷与位移距离的乘积,其方向由负电荷指向正电荷(见图 2)。偶极矩 $\boldsymbol{p}$ 可表示为

$$\boldsymbol{p} = Q\boldsymbol{d} \tag{9}$$

如果电介质单位体积内的分子数为 N,每个分子具有相同的偶极矩 $\boldsymbol{p}$,则面积为 A、高为 h 的电介质内的总电矩为

$$\boldsymbol{p}' = N\boldsymbol{p}Ah \tag{10}$$

从前面极化强度的定义,即单位面积表面上的极化电荷,也可以得出此电介质具有的总电矩为

$$\boldsymbol{p}' = \boldsymbol{P}Ah \tag{11}$$

从物理意义上说,式(10)和(11)都是表示电介质电矩,因而 $\boldsymbol{P} = N\boldsymbol{p}$。这是极化强度的另一个物理意义,即 $\boldsymbol{P}$ 为单位体积内的电矩。

电介质在电场作用下,其分子极化的基本形式有电子极化、原子极化、偶极极化和空间电荷极化四种,如图 3 所示。各种极化的特性可用极化率 α 来表征,它是分子极化形成的感应电矩 $\boldsymbol{p}$ 与作用于分子的电场强度 $\boldsymbol{E}'$ 的比例系数

$$\boldsymbol{p} = \alpha\boldsymbol{E}' \tag{12}$$

不同的极化形式,具有不同的极化率。

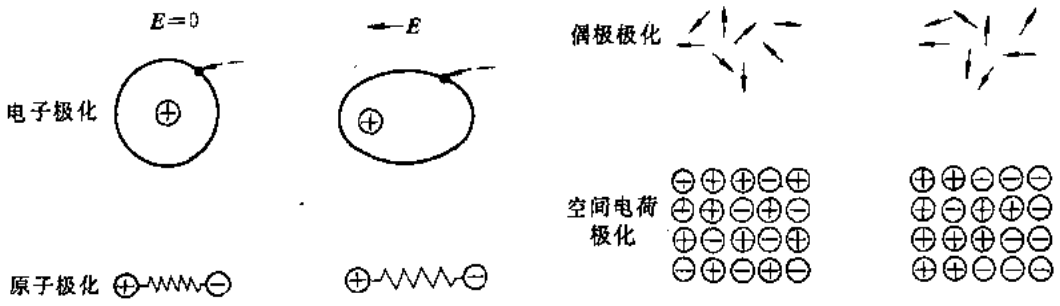


图 3　电介质极化的基本形式

电子极化是原子外围的电子云相对于原子核的位移。这种极化形式出现于所有各种电介质,由于电子质量轻,电子极化的响应时间很短,在 0.1 ~ 1fs 范围,因而电场交变频率进入光频也能响应。电子极化率 α_e 应包括核外全部电子位移的结果,但最外层的价电子束缚最弱,对

极化率的贡献最大。应用量子力学方法可以对原子的电子极化率作精确计算，但对多电子体系来说，这种计算十分繁复，如将电子云看成在核外是均匀分布，则根据静电学原理可以作近似估算，结果为

$$\alpha_e = 4\pi\varepsilon_0 r^3 \tag{13}$$

式中 r 为原子半径。由式(13)可见，原子的电子极化率与其体积成比例，量子力学计算也有相似的结果。

原子极化或离子极化是构成分子或晶体的正负离子或负电性不同的原子之间的位移。由于原子的质量比电子大得多，极化的响应时间比较长，一般为 1～0.1ps，响应频率相当于红外光谱范围。外电场作用下使正负离子相对位移，这不仅与正负电荷的库仑引力有关，也与电子云之间的斥力有关。从正负离子间的引力和斥力，用静电学可以算出离子极化率为

$$\alpha_i = 4\pi\varepsilon_0 \frac{a^3}{n-1} \tag{14}$$

式中：a 为正负离子间距离，即晶格常数；n 为电子壳层之间的斥能与距离关系的指数。n 值可由实验测定，n 值大，表示电子壳层不易变形，离子在外场作用下位移小。

偶极极化(或叫转向极化)是极性分子在电场作用下的转动而不是位移。无电场作用时，极性分子由于无规则热运动，它们在偶极方向作杂乱的随机分布，因而宏观上不呈现极性，在电场力作用下，这些极性分子都将不同程度地转向电场方向，这就形成介质的极化。除去电场，在热运动驱动下，极性分子又最终回复到随机分布。必须指出，偶极极化的转向过程与前两种极化的位移过程是有区别的，前者是一种与热运动密切联系的弛豫过程，而后者则是与弹性回复力有关的谐振过程。一般来说，这种与热运动有关的弛豫过程比较缓慢，其响应频率分布很广，可以从工频到射频。偶极极化率为

$$\alpha_d = \frac{p_0^2}{3k_B T} \tag{15}$$

式中：p_0 为极性分子的固有偶极矩；k_B 为玻耳兹曼常数。

α_d 与绝对温度 T 成反比，温度升高，分子热运动加剧，电场方向的偶极矩分量减小，因而 α_d 随之减小。另外，在推导式(15)时有一假设，即偶极子在电场中的势能远比分子的热运动能小。这样，分子在电场方向的平均偶极矩可以近似与电场强度成正比，α_d 为一常数。如果偶极子在电场中势能超过了分子的功能，则大部分偶极子都已转到电场方向，趋于饱和状态，这时极化与电场强度不再是线性关系。但实际上，电场强度即使高达 3MV/m，上述的假设仍能满足。

空间电荷极化是由于那些受阻的载流子在电介质内迁移所形成的。这些载流子一般都处于较深的势阱内，在热运动激发下也能迁移，但几率较小，结果使这些载流子在介质内均匀分布。在电场力作用下，迁移不再是随机分布，具有一定的方向性，形成空间电荷极化，电场除去后，空间电荷慢慢消失。这种极化的响应时间非常长，一般只有在超低频或直流下才能明显地观察到。空间电荷极化与转向极化相似，也是一种与温度有关的弛豫过程，因而具有与转向极化相似的特性，只是各有不同的响应频谱。所以下面我们不再对这种极化作专门讨论。

如果电介质内上述各种极化都存在，则其总的极化率为

$$\alpha = \alpha_e + \alpha_i + \alpha_d \tag{16}$$

1.3 内电场

从上面的讨论得知，表示电介质极化的宏观特性是介电常数 ε 或极化系数 χ，电介质受

到的作用是宏观电场强度 E；表示电介质分子极化的微观特性是极化率 α，分子受到的作用是内电场或叫有效电场 E'。我们也看到，极化率 α 与电介质的分子结构特征有关，找到 ϵ 与 α 的关系，也就能知道电介质极化特性与材料结构的关系。要建立这一关系必须弄清 E' 和 E 的关系。

为什么分子受到的内电场 E' 不同于外电场 E？这是因为电介质内的每一个分子除了受外电场作用外，还受到周围分子极化形成的偶极子电场的作用。只有在气态下，分子相距甚远，相互作用可以近似地忽略不计，因而 $E' = E$。在凝聚态下，则必须考虑分子间的相互作用。

为了计算方便，取如图4所示的计算模型。为了计算 A 点处分子所受的有效电场，以 A 为圆心作一球，此球是虚拟的，把电介质分子极化形成的偶极矩对 A 点的作用分成球内和球外两部分。球的半径比分子尺寸足够大，使球外分子对 A 的作用可以作宏观处理；但球的半径又不能太大，使球内有限个分子对 A 的作用可以逐个计算。这样，作用于 A 的有效电场 E' 可表示为

$$E' = E_0 + E_1 + E_2 + E_3 \tag{17}$$

式中：E_0 为电介质二面电极上总电荷对 A 作用的电场；

E_1 为电介质极化在二表面形成的极化电荷对 A 作用的电场，这电场与使电介质极化的电场 E_0 方向相反，也叫去极化场；

E_2 为球表面极化电荷对 A 作用的电场，也叫洛仑兹电场，E_1 和 E_2 是用宏观电荷的方法代表球外所有分子的偶极矩电场对 A 的作用；

E_3 为球内分子偶极矩电场对 A 的作用的矢量和，可根据晶格结构和各分子的具体位置，逐个进行计算。

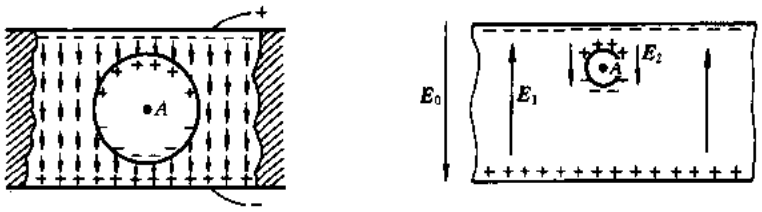


图4　计算内电场的假设模型

根据前面定义的几种电荷面密度与电场中矢量的关系，不难找出如下关系：

$$E_0 = D/\varepsilon_0, E_1 = -P/\varepsilon_0$$

$$E_0 + E_1 = (D - P)/\varepsilon_0 = E \tag{18}$$

计算 E_2 可将球面上的元电荷对 A 作用的电场进行积分，元面积可取与轴线夹角为 θ 和 $\theta + d\theta$ 之间的球面环（见图5）：

$$dA = 2\pi \sin\theta r^2 d\theta \tag{19}$$

dA 元面积上极化电荷对 A 的作用电场为

$$dE_2 = \frac{P\cos\theta}{\varepsilon_0 4\pi r^2} dA \tag{20}$$

沿整个球面积分（θ 从 0 变到 π），则

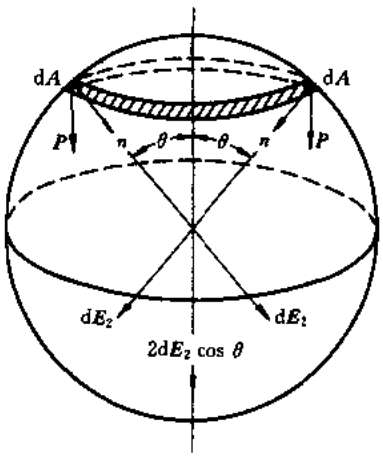


图5　计算有效电场分量 E_2

$$\boldsymbol{E}_2 = \int_0^{\pi} \frac{\boldsymbol{P}\cos^2\theta}{\varepsilon_0 4\pi r^2} 2\pi r^2 \sin\theta \mathrm{d}\theta = \frac{1}{3}\frac{\boldsymbol{P}}{\varepsilon_0} = \frac{1}{3}(\varepsilon_r - 1)\boldsymbol{E} \tag{21}$$

计算 $\boldsymbol{E}_3$ 需知道周围每个偶极子的确定位置，但对高度对称的立方晶格来说，格点上分子的偶极矩都在相同方向，则不难证明，A 点周围各对称点上偶极子对其作用的电场互相对消，因而

$$\boldsymbol{E}_3 = 0 \tag{22}$$

这情况也可以适用于完全随机分布的非极性电介质。将式(18)、(21)和(22)代入(24)得

$$\boldsymbol{E}' = \boldsymbol{E} + \boldsymbol{E}_2 = \frac{\varepsilon_r + 2}{3\varepsilon_0(\varepsilon_r - 3)}\boldsymbol{P} = \frac{\varepsilon_r + 2}{3}\boldsymbol{E} \tag{23}$$

如果不属于上述 $\boldsymbol{E}_3 = 0$ 的情况，则 $\boldsymbol{E}_3$ 的计算将十分繁复，一般情况下，可认为 $\boldsymbol{E}_3$ 正比于 $\boldsymbol{P}$，因而

$$\boldsymbol{E}' = \boldsymbol{E} + \frac{\boldsymbol{P}}{3\varepsilon_0} + \boldsymbol{E}_3 = \boldsymbol{E} + \frac{\beta \boldsymbol{P}}{\varepsilon_0} \tag{24}$$

式中 β 为内场系数，它与电介质的结构有密切联系。对于对称性低的电介质，通常 β 值比较大，内场比外场明显地大。

1.4 介电常数与极化率的关系

从式(11)和(12)可得

$$\boldsymbol{P} = N\alpha \boldsymbol{E}' \tag{25}$$

式(23)可改写为

$$\boldsymbol{P} = \frac{3\varepsilon_0(\varepsilon_r - 1)}{\varepsilon_r + 2}\boldsymbol{E}' \tag{26}$$

式(25)、式(26)是分别用微观参数和宏观参数表示的电介质极化，显然，它们表示的同一物理过程整理后可得介电常数 ε 与极化率的关系为

$$\frac{\varepsilon_r - 1}{\varepsilon_r + 2} = \frac{\varepsilon - \varepsilon_0}{\varepsilon + 2\varepsilon_0} = \frac{N\alpha}{3\varepsilon_0} \tag{27}$$

此关系式就是著名的 C-M(Clausius-Mossotti)方程。必须指出，该方程是在考虑洛仑兹电场的基础上导出的，在其他情况下，需要对此方程进行修正，或导出其他关系式。

在光频下，除了电子极化以外，其他的极化形式都不能响应，因而极化率中只有电子极化率 α_e。这时的介电常数可用折射率 n 来表示。因为光在真空中的传播速度为 $c = 1/\sqrt{\varepsilon_0\mu_0}$。其中 μ_0 是真空磁导率。光在介质中的传播速度为 $v = 1/\sqrt{\varepsilon\mu}$。其中 μ 是介质的导磁率。折射率定义是光在真空中和在介质中传播速度的比：

$$n = \frac{c}{v} = \sqrt{\frac{\varepsilon\mu}{\varepsilon_0\mu_0}} \tag{28}$$

对于非磁性的电介质，一般 $\mu = \mu_0$，所以

$$n = \sqrt{\varepsilon/\varepsilon_0} = \sqrt{\varepsilon_r} \text{ 或 } \varepsilon_r = n^2 \tag{29}$$

这样，C-M 方程具有如下的推广式：

$$\frac{n^2 - 1}{n^2 + 2} = \frac{N\alpha}{3\varepsilon_0} \tag{30}$$

直流电压下的静态介电常数通常应包括所有极化形式的贡献，而 n^2 则主要是电子极化的贡献，所以可以根据比较这两者的数值来估价其他极化形式对介电常数的贡献。表 1 列出了几种常用电介质和半导体的静态相对介电常数 ε_r 和折射率 n。

表 1　常用电介质和半导体的静态相对介电常数 ε_r 与折射率 n

材料	ε_r	n	n^2	材料	ε_r	n	n^2
聚苯乙烯	2.5	1.55	2.40	Si	11.7	3.45	11.90
Nacl(单晶)	6.3	1.54	2.37	Ge	16.3	4.09	16.73
PbO_2(单晶)	26.0	2.60	6.76	GaAs	13.2	3.40	11.56
TiO_2(多晶)	114.0	2.86	8.18	ZnO	8.1	1.93	3.72
冰	78	1.32	1.74	CdS	11.6	2.43	5.90

2　电介质损耗和色散

2.1　电介质损耗参数

由于任何一种极化的建立都需要一定时间，因而当电场交变时，建立极化所引起的电流在时间相位上出现两个分量。今设加于电容器上的交变电压

$$V = V_0 e^{i\omega t} \tag{31}$$

在真空下，电容器所充电荷

$$Q = C_0 V = C_0 V_0 e^{i\omega t} \tag{32}$$

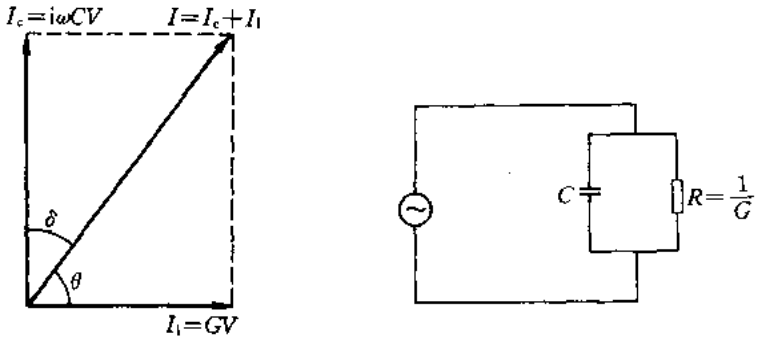
则充放电所引起的电流

$$I_c = dQ/dt = i\omega C_0 V = I_0 e^{i(\omega t + \pi/2)} \tag{33}$$

即充电电流前导电压 $\pi/2$ 是无功电流，不会引起电能的损耗。当电容器中放入电介质后，电流的导前分量为电容电流 I_c，如图 6 所示。I_c 可表示为

$$I_c = i\omega CV = i\omega C_0 \frac{\varepsilon}{\varepsilon_0} V \tag{34}$$

另一电流分量与电压同位相，它是损耗电流：



(a)　　(b)

图 6　(a)电介质中电流向量图(b)等值电路

$$I_1 = GV \tag{35}$$

式中 C 为电介质等值电路中的电导。总电流

$$I = I_c + I_1 = (\mathrm{i}\omega C + G)V \tag{36}$$

图 6 中相位角 δ 可以用作表征介质损耗的参数，叫介质损耗角，即

$$\tan\delta = \left|\frac{I_1}{I_c}\right| = \frac{1}{\omega CR} \tag{37}$$

两电流分量 I_c 和 I_1 也可通过定义一个复数介电常数 ε 来表示。从式(36)和(37)可得

$$I = \mathrm{i}\omega(\varepsilon - \mathrm{i}\varepsilon\tan\delta)\frac{C_0}{\varepsilon_0}V \tag{38}$$

ε 可表示为

$$\varepsilon = \varepsilon' - \mathrm{i}\varepsilon'' \tag{39}$$

ε 的实部 ε' 具有原来介电常数的物理意义，其虚部 ε'' 表示介质损耗的参数，称损耗因数，ε'' 与 $\tan\delta$ 的关系为

$$\varepsilon'' = \varepsilon'\tan\delta \tag{40}$$

把式(38)改写为电流密度与电场强度的关系式，得

$$\boldsymbol{J} = \mathrm{i}\omega(\varepsilon' - \mathrm{i}\varepsilon'')\boldsymbol{E} = \varepsilon\,\mathrm{d}\boldsymbol{E}/\mathrm{d}t \tag{41}$$

当电磁波在电介质内传播时，电介质的极化和损耗特性也可用传播常数描述。根据麦克斯韦电磁场理论，当电磁波为正弦波时，电介质中沿 x 方向传播的平面波可表示为

$$\frac{\partial^2\boldsymbol{E}}{\partial x^2} = \varepsilon\mu\frac{\partial^2\boldsymbol{E}}{\partial t^2}, \quad \frac{\partial^2\boldsymbol{H}}{\partial x^2} = \varepsilon\mu\frac{\partial^2\boldsymbol{H}}{\partial t^2} \tag{42}$$

对电介质来说，复数导磁率等于真空导磁率，则

$$\frac{\partial^2\boldsymbol{E}}{\partial x^2} = \varepsilon\mu_0\frac{\partial^2\boldsymbol{E}}{\partial t^2} \tag{43}$$

求解得

$$\boldsymbol{E} = \boldsymbol{E}_0\mathrm{e}^{(\mathrm{i}\omega t - \gamma x)} \tag{44}$$

式中 γ 为传播常数，

$$\gamma = \mathrm{i}\omega\sqrt{\varepsilon\mu_0} = \mathrm{i}\omega\sqrt{(\varepsilon' - \mathrm{i}\varepsilon'')\mu_0} = \alpha + \mathrm{i}\beta \tag{45}$$

代入式(44)，得

$$\boldsymbol{E} = \boldsymbol{E}_0\mathrm{e}^{-\alpha x}\mathrm{e}^{\mathrm{i}2\pi(ft - \beta x/2\pi)} \tag{46}$$

根据式(46)可写出在时间 t_1 时的电场分量 $\boldsymbol{E}_y$ 在 x 方向的空间波列为

$$\boldsymbol{E}_y = \boldsymbol{E}_1\mathrm{e}^{-\mathrm{i}x} = \boldsymbol{E}_1\mathrm{e}^{-\alpha x}\mathrm{e}^{-\mathrm{i}2\pi\frac{x}{\lambda}} \tag{47}$$

$\boldsymbol{E}_y$ 在空间 x 方向的分布如图 7(a)所示。由此可以明显看出传播常数 γ 的两个分量的物理意义。α 表示电场振幅沿传播方向 x 衰减的指数，叫衰减因子；β 表示电场矢量在空间单位长度上的位相变化，叫位相因子。在 x 方向位相变化的周期就是波长 λ，所以图 7(b)是 $\boldsymbol{E}_y$ 变化的极坐标形式。在这里 $\boldsymbol{E}_y$ 矢量以角度 φ 顺时针方向旋转代表波沿 x 方向的传播：

$$\beta = \frac{2\pi}{\lambda} \tag{48}$$

在光频下，电介质具有的极化和损耗特性将表现为光速变化和吸收，这时可以用复数折射率表征：

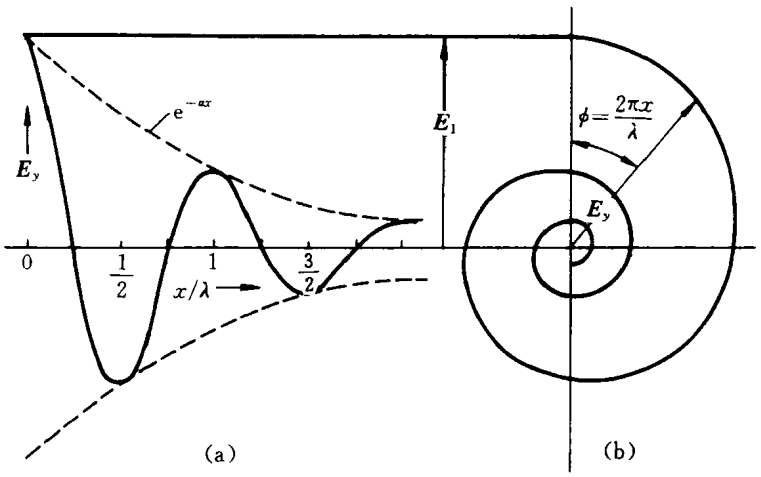


图 7　空间分布的电场波

$$\dot{n} = n(1 - \mathrm{i}\alpha) = n - \mathrm{i}n\alpha = n - \mathrm{i}k \tag{49}$$

式中：α 为吸收系数；k 为消光系数。

复数介电常数 ε、传播常数 γ 和复数折射率 $\dot{n}$ 及它们的两个分量都是表示电介质极化和损耗的参数，相互可以转换。如何选用这几套参数，可以根据不同的频率范围而定。

这些参数的实验测量在不同频率范围有不同的方法。一般说来，低频下可以采用电桥法；高频和超高频范围采用谐振电路；微波区可用驻波测定法；红外以上则采用行波法测量。这些测量方法的基本原理如图 8 所示。

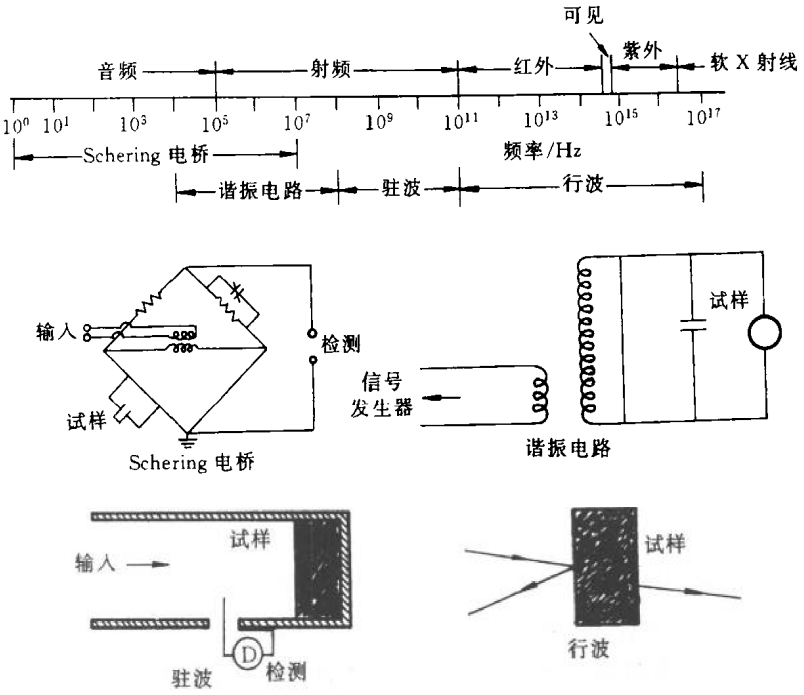


图 8　不同频率范围的测量技术

2.2 弛豫损耗及介电谱[2]

不同的极化形式造成损耗的机理不同，其特性也不同。通常电介质损耗机理可区分为两类：一类是弛豫损耗，这是在那些与热运动直接有关的极化形式，如转向极化和空间电荷极化过程中产生的；另一类是谐振吸收，那是由与弹性力有关的极化形式，如电子极化和离子极化过程所引起的。这里首先讨论弛豫损耗。

在气体或液体中，极性分子可以自由地旋转，当外电场交变时，这些偶极子将跟随着电场方向的更换而往返转动。显然这种运动，特别对液体来说，由于分子间的内摩擦必然引起能量损耗，即把电场能转变成分子的热运动能。在固体中，由于相邻分子的相互作用，其分子不再能自由转动取任意方向，但一般存在少数可以稳定的方向，这决定于电介质的分子结构。在热运动激发下，固体中的偶极子可以从一个方向转到另一个方向。没有外电场作用时，偶极子在这些方向上的分布是完全随机的，因而电介质在宏观上不呈现极性。只有在外电场的驱动下，改变了偶极子的分布，造成极化。

为了能作进一步的定量分析，对上述过程采用双阱模型来描述。为简单起见，设偶极子只有 $+x$ 和 $-x$ 两个取向是其平衡状态，这时其受周围分子作用的势能最低，即处在势能曲线的两个势阱 A 和 B 中，在这两势阱之间，有势垒 ϕ_0，如图 9(a)所示。在热运动激发下，$-x$ 方向的偶极子可以转到 $+x$ 方向；反之，也可。它们需克服的势垒都是 ϕ_0，这时偶极子在两个方向上的转动几率相等，根据玻耳兹曼统计分布，这转动几率也就是单位时间内每个极性分子平均转向的次数：

$$P = v\exp\left(-\frac{\phi_0}{k_B T}\right) \tag{50}$$

式中 v 为分子在平衡位置上热振动的频率。由于转向几率相等，因而偶极子在这两方向的分布几率也相等。

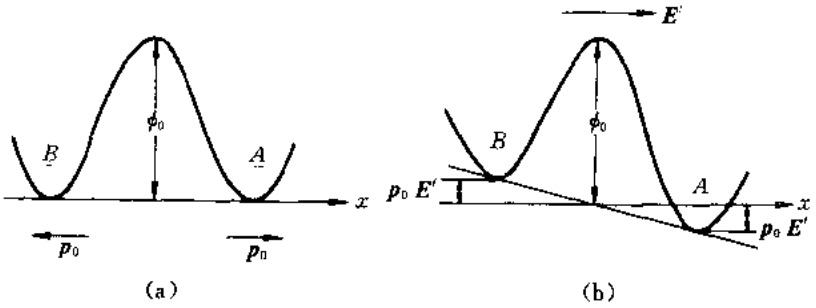


图 9　固体中偶极极化的双阱模型

加电场以后，偶极子 $\boldsymbol{p}_0$ 在电场 $\boldsymbol{E}'$ 中具有势能：

$$\phi = -\boldsymbol{p}_0\boldsymbol{E}'\cos\theta \tag{51}$$

式中 θ 为偶极矩矢量 $\boldsymbol{p}_0$ 与电场强度 $\boldsymbol{E}'$ 之夹角。不同方向的偶极子具有不同的势能，在 $+x$ 方向的偶极子，$\theta = 0°$，$\phi = -\boldsymbol{p}_0\boldsymbol{E}'$；$-x$ 方向的偶极子，$\theta = 180°$，$\phi = \boldsymbol{p}_0\boldsymbol{E}'$，如图 9(b)所示。因此，偶极子从 A 转向 B 需克服的势垒增加为 $\phi + \boldsymbol{p}_0\boldsymbol{E}'$；偶极子从 B 转向 A 的势垒减小到 $\phi - \boldsymbol{p}_0\boldsymbol{E}'$，即极性分子从 B 转向 A 的几率大于由 A 转向 B。结果，在两个方向上分布的偶极子数量不再相等。设：n_1 为在 $+x$ 方向单位体积内的偶极子数；n_2 为在 $-x$ 方向单位体积内的偶极子数；

P_{12}为偶极子从 A 转向 B 的几率；P_{21}为偶极子从 B 转向 A 的几率。显然，两个方向上偶极子数量的差随时间的变化可表示为

$$\frac{\mathrm{d}}{\mathrm{d}t}(n_1 - n_2) = 2(-P_{12}n_1 + P_{21}n_2) \tag{52}$$

一般情况，偶极子在电场内的势能远比其热运动能小，即 $\boldsymbol{p}_0\boldsymbol{E}' \ll k_B T$，在此条件下，可取近似

$$\exp\left(\pm\frac{\boldsymbol{p}_0\boldsymbol{E}'}{k_B T}\right) \approx \left(1 \pm \frac{\boldsymbol{p}_0\boldsymbol{E}'}{k_B T}\right) \tag{53}$$

因而，P_{12}和 P_{21}可作如下简化：

$$\left.\begin{aligned} P_{12} &= v\exp\left(-\frac{\phi + \boldsymbol{P}_0\boldsymbol{E}'}{k_B T}\right) = P\exp\left(-\frac{\boldsymbol{P}_0\boldsymbol{E}'}{k_B T}\right) = P\left(1 - \frac{\boldsymbol{p}_0\boldsymbol{E}'}{k_B T}\right) \\ P_{21} &= v\exp\left(-\frac{\phi - \boldsymbol{p}_0\boldsymbol{E}'}{k_B T}\right) = P\exp\left(\frac{\boldsymbol{p}_0\boldsymbol{E}'}{k_B T}\right) = P\left(1 + \frac{\boldsymbol{p}_0\boldsymbol{E}'}{k_B T}\right) \end{aligned}\right\} \tag{54}$$

代入式(52)得

$$\frac{\mathrm{d}}{\mathrm{d}t}(n_1 - n_2) = 2P\left[-(n_1 - n_2) + \frac{\boldsymbol{p}_0\boldsymbol{E}'}{k_B T}(n_1 + n_2)\right] \tag{55}$$

作适当变换，可得

$$\frac{1}{2P}\frac{\mathrm{d}}{\mathrm{d}t}\frac{\boldsymbol{p}_0(n_1 - n_2)}{3(n_1 + n_2)} = -\frac{\boldsymbol{p}_0(n_1 - n_2)}{3(n_1 + n_2)} + \frac{\boldsymbol{p}_0^2\boldsymbol{E}'}{3k_B T} \tag{56}$$

定义

$$\tau = \frac{1}{2P} = \frac{1}{2v}\exp\left(\frac{\phi}{k_B T}\right), \quad \bar{m} = \frac{\boldsymbol{p}_0(n_1 - n_2)}{3(n_1 + n_2)}$$

代入式(56)可得

$$\tau\frac{\mathrm{d}\bar{m}}{\mathrm{d}t} = -\bar{m} + \frac{\boldsymbol{p}_0\boldsymbol{E}'}{3k_B T} \tag{57}$$

式中：τ 为松弛时间，是偶极子二次转向的间隔时间，τ 与温度 T、势垒 U 和振动频率 v 有关，是转向极化的重要参数，τ 大，说明转向极化响应较慢；$\bar{m}$ 为每个极性分子在电场方向产生的平均极化偶极矩，这里假设偶极子在空间三个坐标方向是平均分布的，因而单位体积内的偶极子总数为 $3(n_1 + n_2)$。

式(57)即为电场 $\boldsymbol{E}'$ 变化时转向极化的动态特性方程，当 $\boldsymbol{E}'$ 为直流电场时，有

$$\bar{m} = A\exp\left(-\frac{t}{\tau}\right) + \frac{\boldsymbol{p}_0}{3k_B T}\boldsymbol{E}' \tag{58}$$

式中 A 为积分常数。如考虑极化建立过程，边界条件为：$t = 0$ 时，$\bar{m} = 0$，可求出 $A = -\boldsymbol{p}_0^2/(3k_B T)$。代入式(58)得

$$\bar{m} = \left[1 - \exp\left(-\frac{t}{\tau}\right)\right]\frac{\boldsymbol{p}_0}{3k_B T}\boldsymbol{E}' \tag{59}$$

直流电场下，当 $t \to \infty$，极化趋于稳定，有

$$\bar{m} = \frac{\boldsymbol{p}_0}{3k_B T}\boldsymbol{E}' \tag{60}$$

由此可见，直流电场下的转向极化率为

$$\bar{m} = \frac{\boldsymbol{p}_0}{3k_B T} \tag{61}$$

当电场 $\boldsymbol{E}'$ 作正弦交变，有

$$\boldsymbol{E}' = \boldsymbol{E}' \exp(i\omega t) \tag{62}$$

式(59)的稳态解为

$$\bar{m} = \frac{\alpha_{d0}}{1 + i\omega\tau} \cdot \boldsymbol{E}'_0 \exp(i\omega t) \tag{63}$$

这说明交流电场下转向极化率出现复数

$$\dot{\alpha}_d = \frac{\alpha_{d0}}{1 + i\omega\tau} = \alpha_d' - i\alpha_d'' \tag{64}$$

其中

$$\alpha_d' = \frac{\alpha_{d0}}{1 + \omega^2\tau^2} = \frac{1}{1 + \omega^2\tau^2} \frac{\boldsymbol{p}_0^2}{3k_B T}$$

$$\alpha_d'' = \frac{\omega\tau\alpha_{d0}}{1 + \omega^2\tau^2} = \frac{\omega\tau}{1 + \omega^2\tau^2} \frac{\boldsymbol{p}_0^2}{3k_B T}$$

α_d' 和 α_d'' 随 $\omega\tau$ 变化的曲线如图 10 所示。这种极化率随频率而变化的特性称为弥散。转向极化的弥散区出现在 $\omega\tau = 1$(或 $\tau = 1/\omega$)附近，即电场交变的周期与极化响应时间相对应时。低于这频率，转向极化能完全响应，极化率 α_d' 接近于直流下的静态极化率；高于这频率很多，转向极化将几乎不能响应，故 α_d' 逐渐趋于零。同时，表示能量损耗的因子 α_d'' 则在 $\omega\tau = 1$ 的附近出现峰值。

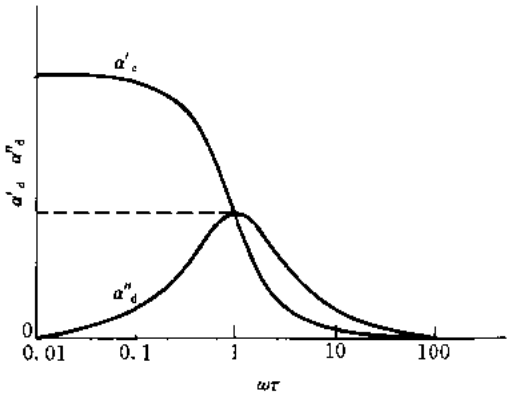


图 10　转向极化率的弥散特性

与此相似，介电常数也呈现弥散特性。假设洛仑兹电场仍适用，并且除了转向极化外，还有电子极化，则式(27)表示的 C-M 方程在正弦交变电压下具有

$$\frac{\varepsilon_r - 1}{\varepsilon_r + 2} = \frac{N}{3\varepsilon_0}\left(\alpha_e + \frac{\boldsymbol{p}_0^2}{3k_B T} \frac{1}{1 + i\omega\tau}\right) \tag{65}$$

当 $\omega \to 0$ 时，$(\varepsilon_r)_{\omega\to 0} = \varepsilon_{r0}$，可得

$$\frac{\varepsilon_{r0} - 1}{\varepsilon_{r0} + 2} = \frac{N}{3\varepsilon_0}\left(\alpha_e + \frac{\boldsymbol{p}_0^2}{3k_B T}\right) \tag{66}$$

当 $\omega \to \infty$ 时，$(\varepsilon_r)_{\omega\to\infty} = \varepsilon_{r\infty}$，这时转向极化消失，即

$$\frac{\varepsilon_{r\infty} - 1}{\varepsilon_{r\infty} + 2} = \frac{N}{3\varepsilon_0}\alpha_e \tag{67}$$

将式(66)和(67)代入式(65)，可得著名的德拜方程：

$$\varepsilon_r = \varepsilon_{r\infty} + (\varepsilon_{r0} - \varepsilon_{r\infty})/(1 + i\omega\tau_0) = \varepsilon_r' - i\varepsilon_r'' \tag{68}$$

式中的 τ_0 与 τ 的关系为：$\tau_0 = [(\varepsilon_{r0} + 2)/(\varepsilon_{r\infty} + 2)\tau]$。通常 τ_0 较 τ 为大。这是考虑了转向极化引起内电场改变的结果，如果 $\varepsilon_{r0} \approx \varepsilon_{r\infty}$，即转向极极化对介电常数的贡献并不很大时，则

$\tau_0 \approx \tau$。

ε_r 有两部分组成，即

$$\varepsilon'_r = \varepsilon_{r\infty} + \frac{\varepsilon_{r0} - \varepsilon_{r\infty}}{1 + \omega^2\tau_0^2}, \quad \varepsilon''_r = \frac{(\varepsilon_{r0} - \varepsilon_{r\infty})\omega\tau_0}{1 + \omega^2\tau_0^2} \tag{69}$$

$$\tan\delta = \frac{\varepsilon''_r}{\varepsilon'_r} = \frac{(\varepsilon_{r0} - \varepsilon_{r\infty})\omega\tau_0}{\varepsilon_{r0} + \varepsilon_{r\infty}\omega^2\tau_0^2} \tag{70}$$

消去 $\omega\tau_0$，可得

$$\left(\varepsilon'_r - \frac{\varepsilon_{r0} + \varepsilon_{r\infty}}{2}\right)^2 + \varepsilon''^2_r = \left(\frac{\varepsilon_{r0} - \varepsilon_{r\infty}}{2}\right)^2 \tag{71}$$

在以 ε'_r 为横坐标和以 ε''_r 为纵坐标的复平面上，式(71)的轨迹为一半圆，其圆心在 $((\varepsilon_{r0} + \varepsilon_{r\infty})/2, 0)$ 点，圆的半径为 $((\varepsilon_{r0} - \varepsilon_{r\infty})/2)$，半圆上的每一点表示某一特定的频率，从坐标原点到该点的矢量是复数介电常数 ε_r，此矢量在横坐标上的投影为 ε'_r，在纵坐标上投影为 ε''_r，如图 11 所示。此图常被叫做 Cole-Cole 图。

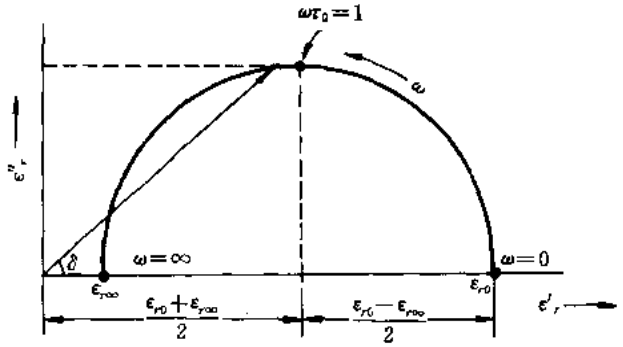


图 11　复数介电常数的 Cole-Cole 图

式(68)所表示的德拜方程及所对应的 Cole-Cole 图都是理想化的，因为所有偶极子的转向具有相同的弛豫时间。但对实际的电介质来说，各偶极子转向时所遇到的势垒并不相同，呈分散分布，因而弛豫时间 τ 也不是单一值，是一系列的分布值。在德拜方程中可用一个经验系数 β 来表示 τ 的分散度，有

$$\dot{\varepsilon}_r = \varepsilon_{r\infty} + \frac{\varepsilon_{r0} - \varepsilon_{r\infty}}{(1 + i\omega\tau_0)\beta} \tag{72}$$

对非晶态固体如聚合物等，β 值常在 0.3～0.7。这时的 Cole－Cole 图不再是一个半圆，而成为一段圆弧，这时圆心位于横坐标以下。圆心离横坐标距离愈远，反映 τ 的分散度愈大，与此相应，这时弛豫谱中的 ε'_r 变化比较缓慢，ε''_r 的峰也趋于平坦，图 12 是一种硫化橡皮的实验曲线，虚线表示理想的德拜弛豫谱，以资比较。

对于空间电荷极化来说，也具有类似的弛豫谱特性，它们也可用双阱模型来解释。所不同的是这时的极化过程不再是偶极子的热转向，而是处于热阱底的离子向相邻势阱的热迁移。

对结构复杂的电介质来说，产生弛豫损耗的不止一种机构。这时在弛豫谱上可以出现多个损耗峰，每一个峰与某一种特定的弛豫机构相对应，如果频率保持恒定，使温度作连续变化，可以得到弛豫特性的温度谱，它与弛豫特性的频率谱具有对应的关系。前面曾证明极化的弥散出现在 $\omega\tau = 1$ 附近，如果保持 ω 不变，改变温度 T 就相当于改变弛豫时间 τ，因而介电常数

的温度谱能出现与频率谱相似的弛豫峰。这种温度谱的测量广泛采用了热激放电法(TSD)，先将试样在加外电场的条件下从高温冷却到低温，然后除去外场再将试样加热，使它均匀缓慢地从低温上升到高温，并在这过程中测量试样两面电极的电流变化。

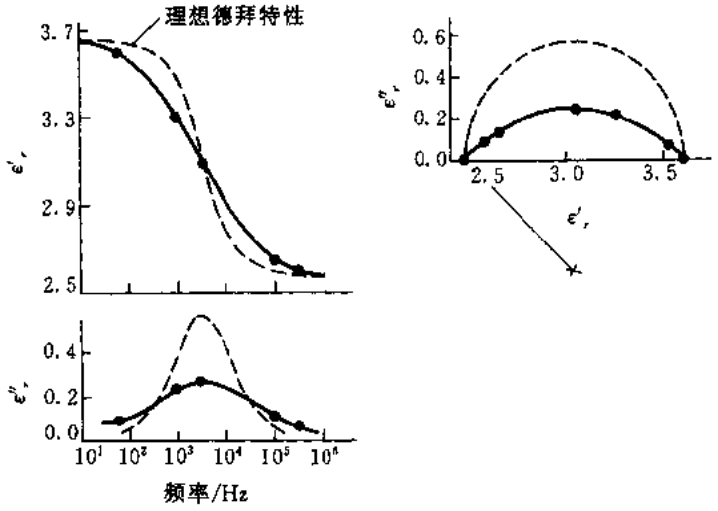


图 12　硫化橡皮的介电常数频率特性和 Cole-Cole 图

2.3　谐振极化及色散[2]

电子极化和离子极化与偶极转向极化的区别是极化位移与正负电荷间的库仑力有关，这是一种决定于变形大小的弹性力。因而采用经典的方法，可将交变电场中的电子和离子极化作为谐振体系来处理。

我们可以把带电质点的运动方程写为

$$m\frac{\mathrm{d}^2x}{\mathrm{d}t^2}+\beta\frac{\mathrm{d}x}{\mathrm{d}t}+\gamma x=e\boldsymbol{E}'_0\exp(\mathrm{i}\omega t) \tag{73}$$

式中：e、m 分别是带电质点的电荷和质量；β 是质点运动时受到的阻尼常数，这里假定阻尼力与速度成正比；γ 是弹性系数，故弹性力与位移成正比。该体系的自然振动频率

$$\omega_0=\sqrt{\gamma/m} \tag{74}$$

从式(73)可解得极化位移 x 的稳态解为

$$x=\frac{e\boldsymbol{E}'_0/m}{\omega_0^2-\omega^2+\mathrm{i}\omega\beta/m}\exp(\mathrm{i}\omega t) \tag{75}$$

交变电场下的位移极化率成为复数

$$\alpha_{ei}=\frac{ex}{\boldsymbol{E}'_0\exp(\mathrm{i}\omega t)}=\frac{e^2/m}{\omega_0^2-\omega^2+\mathrm{i}\omega\beta/m}=\alpha'_{ei}-\mathrm{i}\alpha''_{ei} \tag{76}$$

式中

$$\alpha'_{ei}=\frac{e^2(\omega_0^2-\omega^2)/m}{(\omega_0^2-\omega^2)^2+(\omega\beta/m)^2},\quad \alpha''_{ei}=\frac{e^2\omega\beta/m^2}{(\omega_0^2-\omega^2)^2+(\omega\beta/m)^2}$$

极化率两个分量的频率特性如图 13 所示。

对电子极化来说，ω_0 大约处在紫外光频率范围，例如，$\omega_0=4\times10^{16}\,\mathrm{rad/s}$。在 $\omega_0+\Delta\omega>$

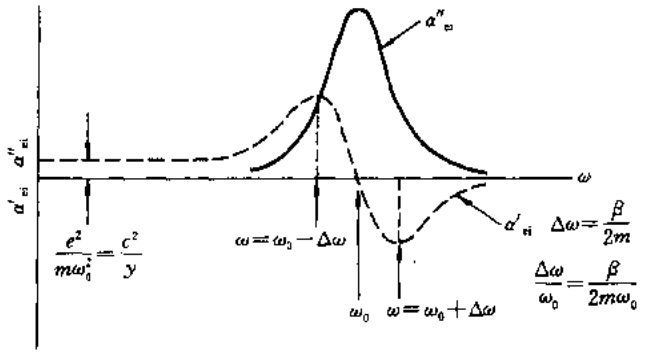


图 13　位移极化率的频率特性

$\omega > \omega_0 - \Delta\omega$ 频谱范围内，α''_e的值相当大，可以把这范围叫吸收区。对原子极化来说，α'_i和 α''_i具有完全相似的频率特性，只是谐振频率通常位于红外光频率范围，因而也叫红外吸收。

如果将式(73)代入前面已导出的关系：$\boldsymbol{P} = Ne\boldsymbol{r}$，$\boldsymbol{E}' = \boldsymbol{E} + \boldsymbol{P}/3\varepsilon_0$。这里再一次假设能适用洛仑兹电场，则可得

$$\frac{\mathrm{d}^2\boldsymbol{P}}{\mathrm{d}t^2} + \frac{\beta}{m}\frac{\mathrm{d}\boldsymbol{P}}{\mathrm{d}t} + \left(\omega_0^2 - \frac{Ne^2}{3m\varepsilon_0}\right)\boldsymbol{P} = \frac{Ne^2}{m}\boldsymbol{E}_0\exp(\mathrm{i}\omega t) \tag{77}$$

极化强度的稳态解为

$$\boldsymbol{P} = \boldsymbol{P}_0\exp[\mathrm{i}(\omega t + \psi)] = \frac{Ne^2/m}{\omega'^2_0 - \omega^2 + \mathrm{i}\omega\beta/m}\boldsymbol{E} \tag{78}$$

式中

$$\omega'^2_0 = \omega_0^2 - \frac{Ne^2}{3m\varepsilon_0}$$

由此可得电介质由于位移极化而出现复数介电常数

$$\varepsilon_r = 1 + \frac{\boldsymbol{P}}{\varepsilon_0\boldsymbol{E}} = 1 + \frac{Ne^2/\varepsilon_0 m}{\omega_0^2 - \omega^2 + \mathrm{i}\omega\beta/m} = \varepsilon'_r - \mathrm{i}\varepsilon''_r \tag{79}$$

式中

$$\varepsilon'_r = 1 + \frac{(\omega'^2_0 - \omega^2)Ne^2/\varepsilon_0 m}{(\omega'^2_0 - \omega^2)^2 + (\omega\beta/m)^2}$$

$$\varepsilon''_r = \frac{\omega Ne^2\beta/\varepsilon_0 m^2}{(\omega'^2_0 - \omega^2)^2 + (\omega\beta/m)^2}$$

式(78)中的 ψ 为极化强度 $\boldsymbol{P}$ 与外电场 $\boldsymbol{E}$ 之间的相角，ε'_r、ε''_r和 ψ 的频率特性如图 14 所示。

比较图 10 与图 13，或比较图 12 和图 14 可以看出，偶极转向极化与位移极化的介电常数的频率特性具有明显区别，前者在弥散区是随频率的增加单调地减小；后者在色散区则是先随频率增加而上升，而后迅速下降，到达最低点后再上升并趋于一定值。介电常数随频率增加而升高的现象叫做正常色散，这是因为在历史上最早发现白光经过玻璃棱镜时的色散是频率高的紫光具有较大的折射率，并将这样的色散叫做正常色散。后来又发现充满碘蒸气的三棱镜对紫光的折射比红光小，因而被称为反常色散。实际上，凡是具有损耗的物质都存在介电常数随频率增加而减小的特性，并不“反常”。由图 10 或图 12 可见，对偶极转向极化来说，只存在反常弥散。而图 13 和图 14 表示出位移极化的全部色散区，既包含正常色散又包括反

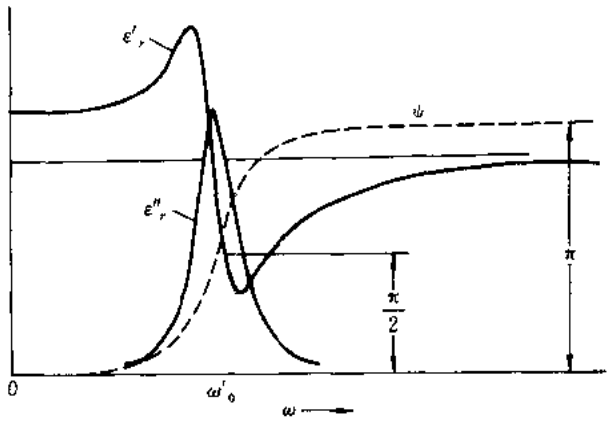


图 14　ε'_r、ε''_r和 ψ 的频率特性

常色散。

另外，比较两类极化的 ε''_r的频率特性可见，弛豫损耗的德拜峰比谐振吸收峰要平坦得多，即谐振吸收峰只占很窄的频带。这一特性也是弛豫过程与谐振过程的明显区别。由于这个原因，利用材料的紫外到红外波段的吸收谱可以分析其化学结构和成分。

在研究这两类极化的频率特性时，通常可以采用等值电路来模拟，如图 15 所示。其中阻容电器(a)具有弛豫极化的频率特性，电容和电感的电路(b)具有谐振极化的频率特性。

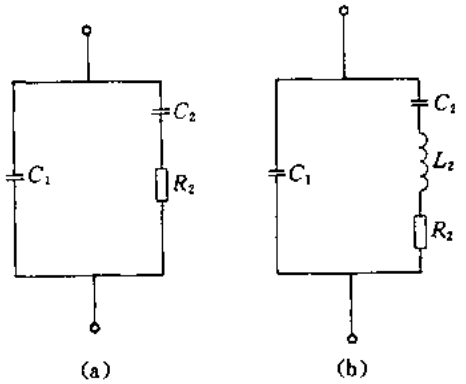


图 15　弛豫极化(a)和谐振极化(b)的等值电路

综上所述，电介质的偶极极化、原子极化和电子极化将在不同的频率范围出现弥散或色散现象，同时伴随着明显的介质损耗或吸收。这种极化的动态特性可以由极化率的频谱来表征，如图 16 所示。介电常数的频谱特性与极化率基本相似。

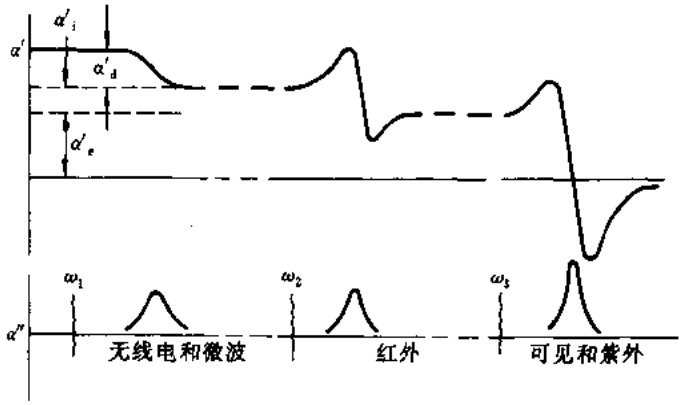


图 16　电介质中各种形式极化的频率特性

3 电介质电导和击穿

3.1 电介质中的导电现象[2]

理想的绝缘体是完全没有电导的，但是实际的电介质多少具有一点导电性。对大多数电介质来说，与导体和半导体相比，它们的电导往往是小到可以忽略不计。但有些情况下，即使电介质的电导很小，也会大大影响到器件的性能。因此，研究电介质电导的现象及其机理是十分必要的。

图17给出了电介质、半导体和导体的电阻率区分的范围，电介质的电阻率一般在$10^8 \sim 10^{20}$ Ω·cm甚至更高；金属导体的电阻率约10^{-6}Ω·cm；大部分半导体的电阻率则在$10^{-2} \sim 10^4$Ω·cm。

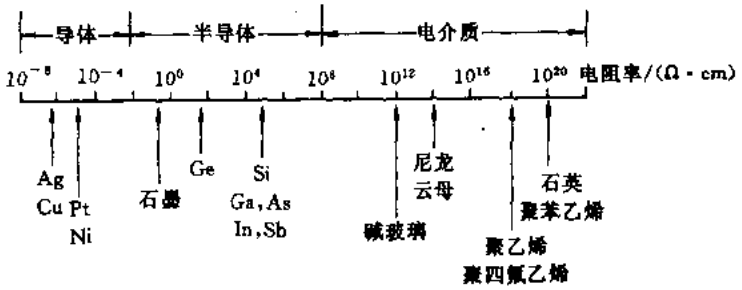


图 17 电介质、半导体和导体的电阻率范围

电介质的电导率除了在数量上与金属导体相差悬殊外，还有其他许多特性。电介质电导随温度的变化与金属相反，而与半导体相似，电导率在相当大的温度范围内随温度作指数式增长，通常具有

$$\log\rho = A + B/T \tag{80}$$

式中：ρ 为电阻率；T 为绝对温度；A 和 B 为两常数。

另一特性是电介质中的电流随时间的延长而缓慢变化，如图 18 所示。我们可将总电流分成三部分：

$$I = I_1 + I_2(t) + I_3 \tag{81}$$

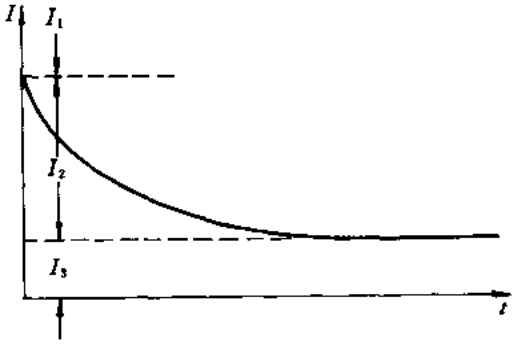


图 18 电介质中电流与时间的关系

式中：I_1 为瞬时充电电流，这由几何电容充电以及电子极化和原子极化所引起的电流，它变化很快，几乎在加电压的瞬时完成。$I_2(t)$为吸收电流，它随时间缓慢下降，可以在几秒钟内达到稳定，也可能延长至几分钟、几小时甚至更长。可认为这是由较缓慢的极化所引起。主要是空间电荷极化(包括界面极化在内)。I_3 是电导电流。但 I_2 往往比 I_3 大得多，并且变化较慢，这对电导的测量和材料的性能都带来影响。

电介质的电导与外施电压的大小有关。直流电压下，电介质中电流与电压成线性关系，满足欧姆定律。在强电场下，电流开始迅速增加，为非线性区。当接近击穿电场强度时，电流的

增长更加急剧，而且往往难于测得其确切的数值，如图 19 所示。

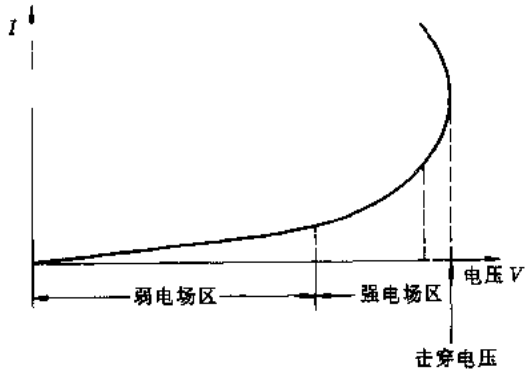


图 19　电介质的电流-电压特性

需指出的是，电介质的电导对外界环境条件十分敏感，除了前面已提到的温度外，湿度和水分的影响也很明显，杂质的沾污，例如能离解出碱离子的盐类等都将使电导引起几个数量级的增加。特别是表面电导的增加，将直接造成器件的失效。

3.2　离子电导

上述电介质电导的一系列特性都与电介质的结构因素有关，具体说是决定于电介质内载流子的产生和类型。一般地说，晶体电介质的导带与价带之间有很宽的能隙，因而在电场不太大时，导带中出现电子和价带中形成空穴的几率很小，即电子(空穴)电导不可能是主要的，这时往往由热离解引起的离子充当载流子，这是离子电导。只有在强电场下，例如到达10^6V/cm，电介质的电导可能转变为以电子电导为主。

电介质中离子的来源有两种情况：如果电介质的组成中离子成分在热运动激发下，一部分离子将脱离原来的位置充当载流子，这叫本征离子；另一种在电介质内迁移的离子是外来的，这叫杂质离子，例如进入 SiO_2 的 Na^+，或进入有机材料的 OH^- 等。在晶体中由于热运动而形成的缺陷有两种；一种是离子离开格点进入间隙位置，然后向相邻的间隙迁移，这叫弗仑克尔缺陷；另一种是离子离开晶体内部而迁移到表面，在体内引起空位，离子可向这空位迁移，这叫肖特基缺陷。这两种热缺陷如图 20 所示，它们都造成离子电流。根据玻耳兹曼统计分布，弗仑克尔缺陷的浓度

$$n_F = \sqrt{NN'}\exp\left(-\frac{W_F}{2k_BT}\right) \tag{82}$$

肖特基缺陷的浓度

$$n_S = N\exp\left(-\frac{W_S}{2k_BT}\right) \tag{83}$$

式中：N 为单位体积内离子总数；N' 为单位体积内填隙离子总数；W_F 为形成弗仑克尔缺陷需作的功；W_S 为形成肖特基缺陷需作的功。

离子在电场方向迁移的速度 v 的计算可以参考图 21 所示的关系，设离子从一平衡位置向相邻位置迁移需克服的势垒为 ϕ，在电场作用下，造成正、反向迁移需克服的势垒不等，相差 $\pm\frac{eEa}{2}$。其中：e 为离子电荷；E 为电场强度；a 是每次迁移的行程。设离子在平衡位置上振

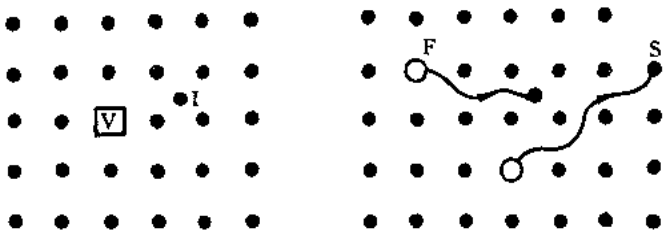


图 20　晶体片的热缺陷

V—空位；　I—填隙离子；　S—肖特基缺陷；　F—弗仑克尔缺陷

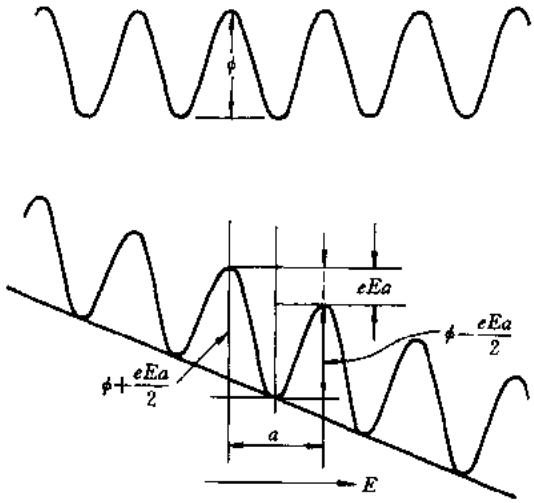


图 21　计算离子迁移率

动频率为 ν,则

$$v = a\nu\left[\exp\left(\frac{-\phi+eEa/2}{k_BT}\right)-\exp\left(\frac{-\phi-eEa/2}{k_BT}\right)\right]=$$

$$2a\nu\exp\left(-\frac{\phi}{k_BT}\right)\sinh\frac{eEa}{2k_BT} \tag{84}$$

在离子电导的范围内,即外电场 $<10^6$V/cm 时,一般有 $eEa\ll k_BT$(T 为室温),则可取如下近似

$$\sinh\frac{eEa}{2k_BT}\approx\frac{eEa}{2k_BT} \tag{85}$$

将式(85)代入(84),得

$$v\approx a\nu\frac{eEa}{k_BT}\exp\left(-\frac{\phi}{k_BT}\right) \tag{86}$$

由此可得离子迁移率 μ 为不随电场强度变化的常数,即

$$\mu=\frac{v}{E}=\frac{e\nu a^2}{k_BT}\exp\left(-\frac{\phi}{k_BT}\right) \tag{87}$$

如果电介质中的离子浓度为式(83)表示的肖特基缺陷,则电介质的电导率

$$\sigma=en\mu=\frac{Nve^2a^2}{k_BT}\exp\left(-\frac{W_S+\phi}{k_BT}\right) \tag{88}$$

或简化为

$$\sigma = C\exp(-B/T) \tag{89}$$

式中 C 和 B 为两个常数，因而电阻率

$$\rho = \frac{1}{\sigma} = \frac{1}{C}\exp\left(\frac{B}{T}\right) \tag{90}$$

电阻率 ρ 的温度特性与式(80)所表示的实验曲线相符合。从这里可看到，在 $\log\rho$ 与 $1/T$ 图上所求得的斜率 B，其物理意义是离子生成和迁移的激活能。

3.3 电子电导[2,4]

固体电介质的结构可分为晶体和非晶体或无定形体两类，晶体可以是离子晶体如 LiF 等卤化物，或共价晶体如金刚石之类，无定形结构如玻璃等。另外也有许多是非完整的晶体，或是颗粒很小的微晶。因此很难用统一的模型来描述电介质内电子电导的过程。一般来说，电子电导应包括电子激发到导带以及导带中电子的迁移两步。下面先讨论电子的发射，然后再讨论电子的迁移。

电介质在弱电场下，由于热运动激发而进入导带的电子数极少，因而弱电场下电介质内的电子电流小到可以忽略不计；否则，就成为半导体了。但是，当电场强度超过某临界值时，可能发生其他形式的电子发射，使导带中的电子浓度明显增长。由电场引起的电子发射形式大致有六种不同情况，如图 22 所示。

图 22 表示了电介质放在金属 A 和金属 B 两个电极之间，由于电场的作用，使能带发生倾斜，这就引起了不同形式的电子发射。图中过程①是电子从电介质的价带穿过禁区直接进入导带，这叫曾纳效应。过程②是电子从禁区内的杂质能级直接进入导带；电子也可以从金属电极的导带直接进入电介质的导带③。或电子由电介质的价带进入电极金属的导带④。以上四种形式是属于量子力学隧道发射过程。另外，电子可以从电极的导带越过势垒进入电介质的导带⑤，这叫肖特基效应；电子也可以从电介质的杂质能级越过势垒进入导带⑥，这叫 Poole-Frenkel 效应。

各种隧道效应引起的电流，用量子力学方法可得到 Fowler-Nordheim 公式

$$J = AE^2\exp(-B/E) \tag{91}$$

式中 A 和 B 是与电极和电介质的功函数有关的常数。隧道效应电流一般与温度没有明显关系，$\ln T/E^2$ 与 $1/E$ 呈负斜率的直线关系。

肖特基发射电流的估算可采用图 23 中所表示的关系。若给金属内的电子以一定能量，电子将越过势垒 ϕ_D 逸出金属表面进入电介质。如果加上电场，势垒将减小，使电子发射容易发生，这种在金属与电介质界面上的场致发射，叫肖特基发射。图 23 中的 $\phi_0(x)$ 为镜像势垒，是由于逸出电子与感应的正电荷之间的镜像力所造成的：

$$\phi_0(x) = -\frac{e^2}{16\pi\epsilon x} \tag{92}$$

加电场后，电子在电介质内的势能分布为

$$\phi = \phi_0 - eEx \tag{93}$$

由 $\partial\phi/\partial x = 0$，可求得 ϕ_{max} 和 x_{max} 分别为

$$\phi_{max} = -\sqrt{\frac{e^3E}{4\pi\epsilon}}, \qquad x_{max} = \sqrt{\frac{e}{16\pi\epsilon E}} \tag{94}$$

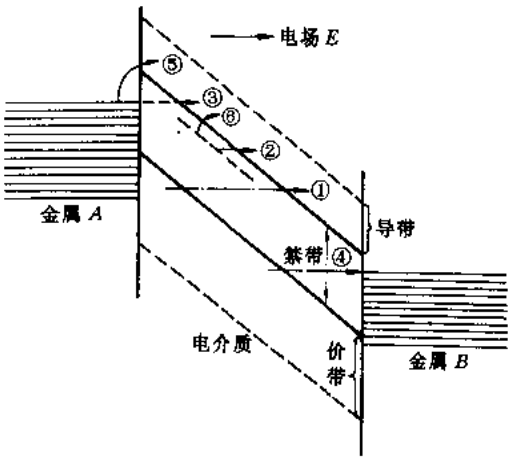


图 22　电介质内电子发射过程

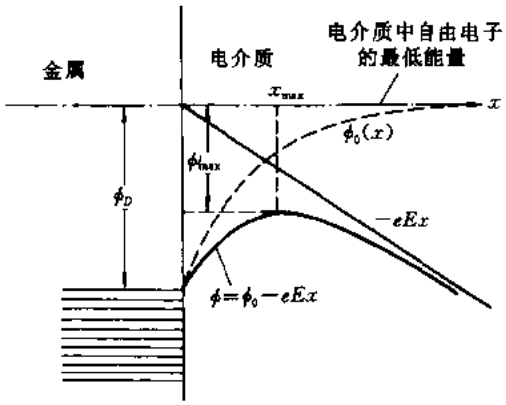


图 23　肖特基发射效应

这时势垒下降为 $\phi_D-\sqrt{\dfrac{e^3E}{4\pi\varepsilon}}$。

金属表面的热发射电流密度由 Richardson-Dushman 公式表示：

$$J=AT^2\exp(-\phi/k_BT) \tag{95}$$

式中：A 为常数；ϕ 为功函数。如果把 ϕ 用 $\phi_D-\sqrt{\dfrac{e^3E}{4\pi\varepsilon}}$ 代入，可得肖特基发射电流密度：

$$J=AT^2\exp\left(\frac{\beta_SE^{1/2}-\phi_D}{k_BT}\right) \tag{96}$$

式中

$$\beta_S=\sqrt{\frac{e^3}{4\pi\varepsilon}}$$

$\ln J$ 与$\sqrt{E}$的关系呈直线，不再符合欧姆定律。

在电介质内杂质局部能级上的电子越过势垒形成的 Poole-Frenkel 发射电流的机理如图 24 所示，图中表示具有正电荷的施主能级上所俘获的电子激发到导带的情况。如在施主能级与导带之间的能隙为 ϕ_D，弱电场下也可能有一部分电子由于热激发而进入导带引起电流密度：

$$J_0=A\exp\left(-\frac{\phi_D}{2k_BT}\right) \tag{97}$$

式中 A 为常数。这时施主正电荷与金属和电介质界面上的镜像势不同，形成库仑势垒：

$$\phi_D(x)=-\frac{e^2}{4\pi\varepsilon x} \tag{98}$$

有外电场时，有效势垒改变为

$$\phi=\phi_0(x)-eEx \tag{99}$$

由$\partial\phi/\partial x=0$，可求得 ϕ_{max}和 x_{max}分别为

$$\phi_{max}=-\sqrt{\frac{e^3E}{\pi\varepsilon}},\qquad x_{max}=\sqrt{\frac{e}{4\pi\varepsilon E}} \tag{100}$$

将 $\phi_0-\sqrt{\dfrac{e^3E}{\pi\varepsilon}}$取代 J_0 中的 ϕ_D，得

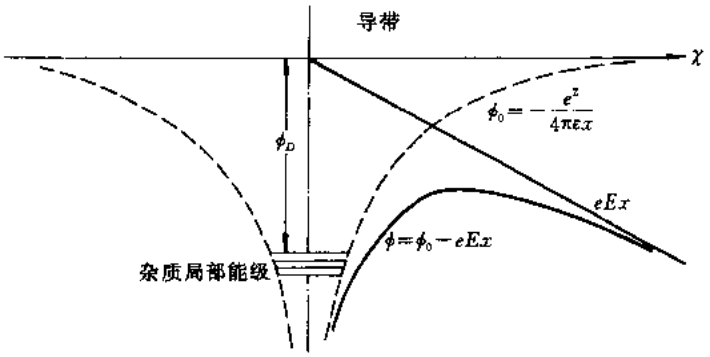


图 24　Poole-Frenkel 发射电流

$$J = A\exp\left(\frac{-\phi_0+\sqrt{e^3E/\pi\varepsilon}}{2k_BT}\right) = J_0\exp\left(\frac{\beta_{PF}E^{1/2}}{2k_BT}\right) \tag{101}$$

式中
$$\beta_{PF}=\sqrt{\frac{e^3}{\pi\varepsilon}}=2\beta_S$$

如作 $\ln J$ 与$\sqrt{E}$的关系图，得到的是斜率为$\frac{\beta_{PF}}{2k_BT}$的直线，与肖特基电流的特性基本相似，但后者将随金属功函数不同而变。

固体电介质中电子密度的迅速增长的另一种过程是电子雪崩，电子雪崩的出现常常是电介质击穿的先兆，故这一过程将在后面讨论。

电介质中电子的迁移可以用两种模型来解释，对于长程有序的晶体电介质一般可用能带模型，而对于短程有序长程无序的无定形固体来说采用跳跃模型。对能带模型来说，电介质与本征半导体类似，导带中的电子与价带中的空穴数相等，可表示为

$$n_c = n_p = \sqrt{N_C N_V}\exp\left(\frac{E_g}{2k_BT}\right) \tag{102}$$

式中：N_C 和 N_V 分别为导带底和价带顶中有效状态密度；E_g 为能隙。

在电场作用下，电子在导带中的迁移率为

$$\mu = \frac{e}{m}\tau \tag{103}$$

式中：e 和 m 分别为电子的电荷和质量；τ 为散射的弛豫时间，即二次碰撞的时间间隙。

式(103)推导如下：电子在电场作用下受力为 $-eE$，由此在反电场方向产生平均漂移速度 $\bar{v}_x$。设平均动量

$$\bar{p}_x = m\bar{v}_x \tag{104}$$

根据经典力学定律，电场造成电子的动量变化等于电场力，有

$$\left(\frac{dp_x}{dt}\right)_{电场} = m\frac{d\bar{v}_x}{dt} = -eE \tag{105}$$

假设电子在每次碰撞中其动量全部失去，电子二次碰撞的时间间隔为 τ，则单位时间内的碰撞次数为 $1/\tau$。因而，单位时间内因碰撞而引起的动量变为

$$\left(\frac{dp_x}{dt}\right)_{碰撞} = -\frac{1}{\tau}p_x = -\frac{m\bar{v}_x}{\tau} \tag{106}$$

经过相当长时间后，电场加速与碰撞减速达到平衡状态，使电子在 $-x$ 方向的迁移速度 v_x 为恒定值，即

$$\left(\frac{\mathrm{d}\bar{p}_x}{\mathrm{d}t}\right)_{\text{电场}}+\left(\frac{\mathrm{d}\bar{p}_x}{\mathrm{d}t}\right)_{\text{碰撞}}=0 \tag{107}$$

将(106)和式(107)代入式(105)，得

$$\bar{v}_x=-\frac{e}{m}\tau E \tag{108}$$

由式(108)即可导出式(103)。由此可知，电子在导带中运动的迁移率不随电场而变。

对于非完整的晶体或非晶体来说，不能简单地用能带模型来解释。这时可以把能带看成是分段的。在每一段内部认为电子共有化运动与晶体内一样容易，而在相邻段之间由于原子相距较远，存在一定的势垒，不能自由迁移，而需获得能量越过势垒，这过程称为跳跃，电子迁移的跳跃模型如图 25 所示。一般来说，电子由于这种跳跃过程而有的迁移率的临界值大约为 $1\mathrm{cm}^2/(\mathrm{V}\cdot\mathrm{s})$。大于此值可认为是能带电导；反之，即为跳跃电导。

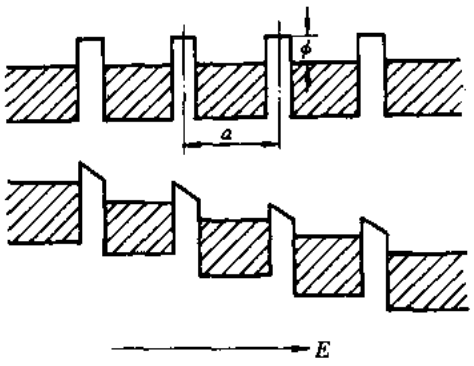


图 25　电子电导的跳跃模型

3.4　本征击穿[2,3]

任何电介质所能承受的电场强度都有一定限度，超过此临界值，电介质中的电流急剧增长并出现不稳定，最终导致介电状态的破坏，这就是电介质的击穿。电介质的击穿具有不同的过程和机理，这由电介质的结构特征所决定。一般来说，如果电介质的结构均匀致细，不含气泡和杂质，且电导和损耗很低，这种情况下，电介质的击穿过程直接与电场作用下的电子过程有关，这叫本征击穿或电击穿。如果电介质具有较高的电导和损耗，或结构松疏含有小气泡，在电场强度尚不足引起本征击穿以前，可能由于热的因素或是局部气体放电的影响使电介质击穿，这叫热击穿。还有其他的击穿形式。

电介质本征击穿的特点是，击穿过程发展的时间非常短，常常在 $10^{-1}\mathrm{s}$ 以下。如果电路中有较大的限流电阻，使击穿后的电流受限制，则击穿的痕迹将局限在很小的部位，有的甚至肉眼不能发现。固体电介质的电击穿电场强度通常都在 1～10MV/cm。

已经提出的各种本征击穿理论都是考虑电场的作用使电介质内自由电子增长，但各人提出的击穿开始的判据和条件不同。较有代表性的电击穿理论有四种，即单电子击穿、集合击穿、雪崩击穿和场致击穿。前两者击穿的判据只考虑电子能量的平衡，而不考虑自由电子增长的过程；后两者则主要从电子增长过程来确定击穿的条件。

3.4.1　自由电子的能量平衡

在电场作用下，自由电子在电场中迁移从而获得能量，但同时电子又受到晶格的散射而使一部分能量损耗。单位时间内电子从电场获得能量

$$A=e\mu E^2 \tag{109}$$

式中

$$\mu = \frac{e}{m}\tau(\varepsilon) \tag{110}$$

式中：μ 为电子迁移率；τ 为弛豫时间，即电子迁移过程中受晶格散射的平均时间间隔，不同能量 ε 的电子，其 τ 也不同，故 $\tau(\varepsilon)$是 ε 的函数。将式(110)代入式(109)，得

$$A = \frac{e^2}{m}\tau(\varepsilon)E^2 = A(\varepsilon, T_0, E) \tag{111}$$

式中 T_0 为环境温度，也表征未受电场加速前的初始能量。

电子单位时间内因受晶格散射而损失的能量

$$B = \Delta\varepsilon/\tau(\varepsilon) = B(\varepsilon, T_0) \tag{112}$$

式中 $\Delta\varepsilon$ 为每次散射失去的能量。

3.4.2 单电子击穿理论

如果电子从电场获得能量的速率不超过损失于晶格散射的速率，即 $A \leqslant B$，则击穿不可能发生。但如果 $A > B$，则电子将不断累积能量，最终导致自由电子大量增长，使击穿到来。不同能量的电子其 $A(\varepsilon)$和 $B(\varepsilon)$是不同的，$A(\varepsilon)$和 $B(\varepsilon)$随 ε 的变化曲线如图 26 所示。

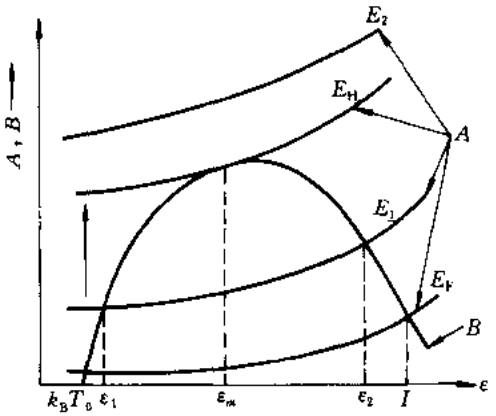


图 26 电子的能量平衡

从图 26 可知，如果电场强度为 E_1，则能量在 ε_1 以下的电子，都具有关系 $A > B$，这些电子被电场加速；当能量在 ε_1 和 ε_2 之间，具有关系 $A < B$，因而电子被散射而减速；只有能量大于 ε_2 的电子，则完全满足 $A > B$ 的条件，能不断加速，产生新的自由电子。

Von Hipple 提出，只有当电场强度增加到 E_H，能使 $\varepsilon = \varepsilon_m$ 的电子加速时，即满足

$$\left.\begin{aligned} &A(\varepsilon, T, E_0) = B(\varepsilon, T_0) \\ &\varepsilon = \varepsilon_m \end{aligned}\right\} \tag{113}$$

作为击穿条件。式(113)称为 Hipple 击穿判据。从图 26 可知，ε_m 是曲线 A 和 B 的切点，从式(113)可以解得 Hipple 击穿电场强度 E_H。显然，当 $E = E_H$ 时，各种能量的电子都一直被电场加速，因而，实际上这一判据是充分的。但是非必要的。

Fröhlich 则认为电场强度只要能使电介质中任何可能出现的电子一直加速，击穿就可发生。他假设，由于电子的能量统计分布存在着各种不同能量的电子，只是不超过晶格的电离能 I，因而他提出的击穿条件为

$$\left.\begin{aligned} &A(\varepsilon, T_0, E) = B(\varepsilon, T_0) \\ &\varepsilon = I \end{aligned}\right\} \tag{114}$$

这叫 Fröhlich 击穿判据，由此关系可求出 Fröhlich 击穿电场强度 E_F。显然，E_F 是使电介质中有电子一直被电场加速的最低的电场强度，因而，这一判据是完全必要的，但非足够充分。

3.4.3 集合电子击穿理论

理论和实验的差别说明只考虑某种能量电子的个别作用是不够的。实际上，不同能量的

电子相互有影响，特别当电子密度较高时，电子相互作用的几率将超过电子受晶格散射的几率。换句话说，不应简单地考虑电子的个别作用，而应将电子作为一个集合系统来处理，也有人将此集合电子系统叫电子气。

这时不再考虑个别电子的能量，而是考虑这集合中电子的平均能量，并以电子温度 T_e 来表征。这样，在电场作用下 T_e 将大于周围环境（晶格）温度 T_0。如果电场强度 E 达到某临界值，使电子气从电场获得能量 $\bar{A}$ 超过了电子气因散射而损耗于晶格的能量 $\bar{B}$ 时，击穿可以发生，所以这时的击穿判据为

$$\bar{A}(T_e, T_0, E) = \bar{B}(T_e, T_0) \tag{115}$$

式中 $\bar{A}$ 和 $\bar{B}$ 对 T_e 的关系，与前面的 A 和 B 对 ε 的关系相似，可以解出击穿电场强度 E_C。由于在该理论中考虑了不同能量电子的平均作用，可以预料，所计算得到的 E_C 一定具有如下关系：

$$E_H > E_C > E_F \tag{116}$$

3.4.4 电子雪崩击穿理论

上述理论只是考虑了能使起始自由电子不断获得能量从而产生大量新生电子所需要的电场强度。完全没有计及电子大量增生的过程。电子雪崩击穿理论从电子碰撞过程出发考虑达到击穿所需要的电场强度。这种电子碰撞电离所形成电子雪崩的过程，可以像气体中那样把它看成是电子的个别作用，但在固体中，电子相互作用，也应将不同能量的电子作为一体集合电子系统来处理。

下面仅就单电子近似的电子雪崩理论作简单讨论，如有一随机的单电子从阴极出发进入电介质，如果电场足够大，发生碰撞电离，使晶格中造成大量增生的新自由电子。设此电子在进入阳极前发生了 i 代的碰撞电离，总共生成的自由电子有 2^i 个。这些电子集合在一个电子雪崩内，电子雪崩范围的半径由电子的热扩散所决定，扩散长度

$$r = \sqrt{2Dt} \tag{117}$$

式中 D 为扩散系数。如取 $D = 1\text{cm}^2/\text{s}$ 和时间 $t = 1\mu\text{s}$，可算出 r 大约为 10^{-3}cm。长 1cm，半径为 10^{-3}cm 的圆柱体的体积为 $10^{-6}\pi\text{cm}^3$，在该体积内的原子数大约为 10^{17} 个。如设这部分晶格点阵的原子由于达到某临界温度而分散，每原子需获得 10eV 能量，因而在体积内电子将损耗 10^{18}eV 的能量。如果外电场强度是 10^6V/cm，再假设电子从电场所获得的能量全部转交给晶格，则电子雪崩中的总数不应小于 $10^{18}/10^6 = 10^{12}$ 个，$2^i = 10^{12}$，即 $i \approx 40$，这就是说电子从阴极到阳极的路径上至少要产生 40 代新生自由电子，才能导致击穿，这就是著名的 Seitz40 代理论。

当然这是十分粗略的估计，在固体击穿的电子雪崩理论中，如果考虑不同能量的电子之间相互作用，以及由于碰撞电离形成的晶格正电荷（空穴）的影响，计算将会变得相当复杂，这方面的理论工作尚有待进一步完善。

3.4.5 场发射击穿理论

由于量子力学的隧道效应，价带电子可以进入导带，由于对晶格转移了能量，使其温度增高，达到某临界值时，导致击穿发生。这种击穿也叫隧道击穿或叫齐纳击穿。

电子从价带进入导带的几率与禁带宽有明显关系，对禁带宽的电介质来说，根据场发射原理算出的击穿电场强度大大超过实验结果。这说明这种理论对禁带比较窄的电介质或半导

体较为适用，对禁带宽的晶体电介质和无定形电介质来说，其他形式的电击穿将比隧道击穿早来到。

3.5 热击穿和放电击穿

实验结果表明，当电介质发生电击穿时，击穿电场强度与温度的关系不明显。电介质热击穿时，不仅击穿电场强度比电击穿低得多，而且击穿电场强度的数值随温度增加而显著下降，同时还与电压作用时间有关。这些特点都说明电介质的热击穿与电场使电介质加热的过程密切有关。

电介质热击穿理论最早是在 1922 年由 Wagner 提出的，他认为电介质具有电导和损耗，电场增强时，当消耗在介质内的电能转变为热能来不及向周围发散时，出现热不稳定，导致电介质的击穿。若对某种绝缘结构中的发热和散热进行计算即可得出其击穿电压。这种击穿电压常常不是这种材料的固有数值，而与周围环境的参数和散热条件有关。一些电气设备或电子器件在受潮、高温等情况下，很容易发生绝缘的击穿就是这原因。

固体电介质的放电击穿有两种情况：一是电介质内部含有微小气孔，在电场作用下，气孔中发生气体放电；另一种是电极设计不合理，在边缘出现局部电场集中引起气体放电。这种电介质内部或电极边缘放电的结果会造成电场的畸变，放电产生的电子、离子、射线和化学活性物质使材料老化，这些影响都造成击穿电压下降。

4 电介质材料和应用

固体电介质材料从晶体到无定形，从无机材料到有机材料，其品种远比导体和半导体多得多。下面我们将分为单晶、玻璃和陶瓷以及聚合物电介质两大类来讨论。

4.1 单晶、玻璃和陶瓷电介质材料

晶体从其化学组成来分有元素的单质晶体和多元素化合物晶体。从元素周期表可知，非金属元素包括Ⅲ族的 B；Ⅳ族的 C 和 Si；Ⅴ族的 P 和 As；以及Ⅵ族的 S，Se 和 Te 都可以成为单质晶体，但其中只有 C 构成的金刚石晶体是典型的电介质，其禁带宽为 5.2eV，乃是典型的元素半导体。还有一些元素因为晶体熔点低，化学稳定性差或难于生成单晶，很少能作为材料实际应用。

化合物晶体的电介质则品种很多。如果化合物的原子或离子的价电子壳层中的 s 和 p 轨道完全充满的话，都是属于非金属晶体。换句话说，化合物的结合是由于原子间电子的转移而形成离子键，或是由于电子共有而形成共价键，而构成非金属，其中大部分是电介质，有一部分是半导体。例如二元化合物的碱卤晶体，二元和多元化合物的金属与非金属的氧化物晶体，氮化物晶体等大多数是电介质。特别值得指出，一些结构复杂的多元化合物晶体由于对称性较低，往往呈现出与某种结构对称性密切有关的特性。例如，$BaTiO_3$、$LiNbO_3$ 等晶体表现出良好的铁电和压电等特性。一般来说，晶体电介质由于其生长和加工比较困难，使其用作电绝缘材料受到限制，但用它们的特种性能制成的电介质器件则有很大的发展潜力。

广义上说，一切非晶体都可以称玻璃体。但通常说的玻璃体是指未结晶的无机氧化物和复合物，它们是由多种无机氧化物熔融后的过冷液体。最基本的氧化物称为玻璃的基体，由它

们构成玻璃的网格结构，其他的氧化物可以加入这种玻璃结构，但它们自身不能单独形成网格，这氧化物叫调节剂或助熔剂；另外还有一类中间氧化物，它们自身也不能形成网格，但可以在网格中取代一部分基体氧化物。玻璃中最常用的基体氧化物有 SiO_2 和 B_2O_3，其他的基体氧化物如 P_2O_5、GeO_2、As_2O_5 等。中间氧化物主要有 Al_2O_3、PbO，也可能是 TiO_2 和 BeO 等。助熔剂有 Na_2O、K_2O、Li_2O、CaO 和 MgO 等。

图 27 示出了石英晶体和石英玻璃的二维结构示意图，这两种结构都包括了 SiO_2 的四位体单元，即 Si 原子位于四面体中心，四顶角上是氧原子，而每个氧原子又被两个四面体所共有，因而构成了网格结构。图 27(a)中 SiO_2 四面体的排列有严格周期性，由六个四面体构成六角环。图 27(b)中 SiO_2 四面体的排列则没有严格规则，只是短程有序，即每个 SiO_2 仍作四面体排列，但不是长程有序，即四面体之间的连接有随机性，有的可以四个四面体构成一个环，也可以有五个、六个、七个甚至八个四面体连成一个环。图 28 是普通窗玻璃的结构图。其成分除了 SiO_2 外，还包括了一定比例的 Na_2O。由于 Na_2O 的加入使网格中一部分氧原子的连接中断，从而使玻璃的熔融温度大大降低，这给加工带来不少方便。但是一价 Na^+ 的引入会使电介质的性能恶化，例如即使少量 Na^+ 的存在也会引起电导率和介质损耗的显著增加。特别是在半导体 MOS 器件的 SiO_2 绝缘层中，哪怕极少量 Na^+ 的沾污，也会造成能带中表面态密度增大，使器件性能下降。

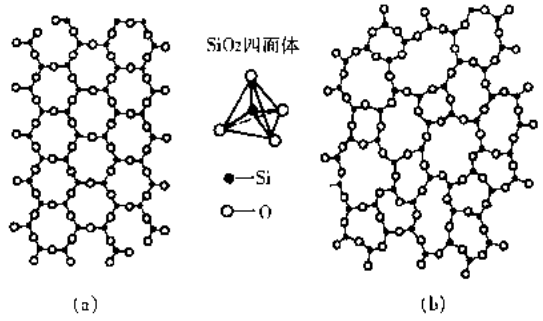


图 27　石英晶体(a)和石英玻璃(b)的结构比较示意图

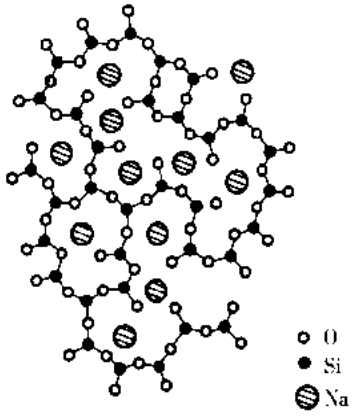


图 28　钠硅玻璃的结构

一般来说，玻璃都是良好的电介质。它们的介电性能与其化学组成和均匀性有密切关系。一价碱金属氧化物的加入将造成介电性能下降，如果用二价金属氧化物代替上述碱金属氧化物作为助熔剂，介电性能可得到明显改善。电工绝缘材料采用无碱玻璃的原因就在于此。

玻璃不仅是好的电介质，同时也是好的光介质。各种光学玻璃往往要求具有高的折射率，如果在玻璃成分中加入那些电子极化率大的原子，就能有效地提高玻璃的折射率，例如加入一定比例的 PbO，就具有相当高的折射率，因而称为铅玻璃。含有大量 Pb 的玻璃，对 X-射线具有很强的吸收作用，因而可以用作防护。近年来迅速发展起来的光导纤维，目前大都采用石英玻璃拉制而成，如果在石英玻璃中含有极少量的—OH 基，就会导致光纤在某一红外波段的吸收出现很高峰值。为了提高光纤纤芯的折射率，通常可在石英玻璃中掺入少量的 GeO 等氧化物。

陶瓷也是无机材料，但它与玻璃的加工过程和结构都不相同。陶瓷制造的一般工艺过程是将粉末状原料压成坯，然后在高温下烧结而成。陶瓷的结构是由大小不等的晶粒结合而成，相邻晶粒之间界面叫晶界，每一颗晶粒是一单晶体，但不同的晶粒具有不同的晶向。在一定的高温下保持相当时间，晶粒的尺寸能逐步扩大。由于陶瓷的加工与晶体生长相比较具有制备方便和成本低等优点，所以被广泛地应用。

陶瓷电介质从其电特性来分主要有电绝缘陶瓷和铁电、压电陶瓷。电工和电子绝缘陶瓷的主要化学成分是硅酸盐，铁电和压电陶瓷的种类较多，常用的如钛酸钡、锆钛酸铅、铌酸锂和钽酸锂等。以上这些陶瓷组成大部分都是氧化物。另外，有许多不含氧的陶瓷不断被合成出来，它们往往表现出不寻常的铁电和反铁电特性。

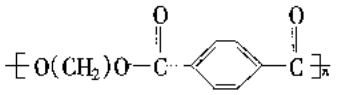
4.2 聚合物电介质材料

聚合物是相对分子质量很高的有机化合物，天然的聚合物有橡胶、棉和丝等，现在用作电介质的大量是合成的聚合物。聚合物都是由大量低分子的单体通过聚合化学反应生成。因而聚合物分子中包括了大量相同的重复单元，重复单元的数目叫聚合度。例如聚乙烯可由乙烯气体直接聚合而成：

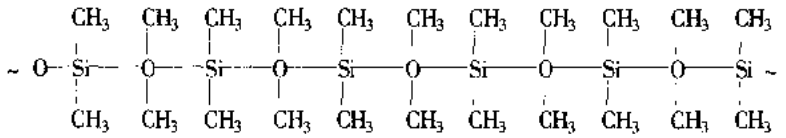
$$\underset{\text{乙烯}}{n\mathrm{CH_2=CH_2}} \xrightarrow{\text{聚合反应}} \underset{\text{聚乙烯}}{\left[\mathrm{CH_2-CH_2}\right]_n}$$

聚合物按材料的性质可以分为塑料、橡胶和纤维三类，塑料又可以分为热塑性塑料和热固性塑料。前者为长链线型（或支型）聚合物，加热时可以软化和流动，能够多次重复地塑化成型，例如聚乙烯、聚氯乙烯等；后者为体型聚合物，一经固化成型后，不再能重新塑化，这种聚合物有酚醛树酯、环氧树脂等。

根据高分子主链构成的化学组成可分为碳链、杂链和元素高分子三类。碳链高分子的主链全由碳原子构成，上面所举出的聚乙烯就是最典型的碳链高分子；主链上除碳原子外还有氧、氮、硫等原子的杂链高分子，例如聚酯和环氧树脂等。聚对苯二甲酸乙二醇酯的化学式为



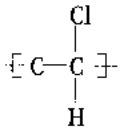
另一类是元素高分子，其主链不一定含有碳原子，而由硅、氧、氮、铝、硼、磷和钛等元素构成，如有机硅树脂的化学式为



聚合物的物理状态和性能不仅决定于化学组成，而且与其结构有密切关系。与低分子物质相比，聚合物结构的主要特征是多层次，聚合物所具有的某些独特性能如高弹性，就是由于结构的多层次，通常可将聚合物结构分成三个层次。

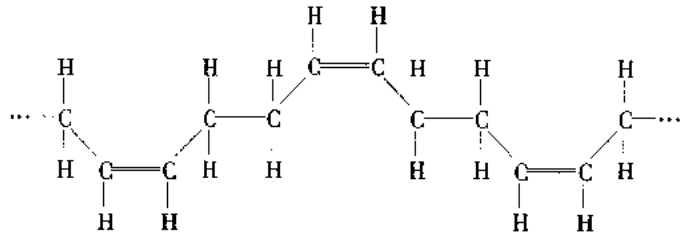
一次结构：是高分子链节的结构，也就是高分子中的重复单元的结构，这包括化学结构、空间构型，链节连接的序列和链段的分支或交联等，这是聚合物最基本的结构。主链两侧基团的极性大小和对称与否，往往对其介电性有显著影响，当链节的结构不对称时，就可能出现两种不同的构型，例如聚氯乙烯的两种链节构型：

d-链节

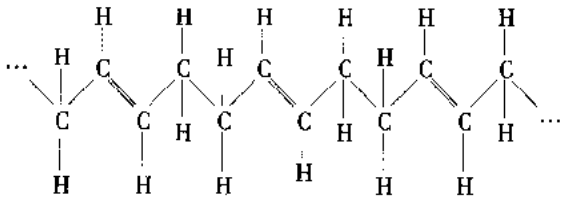


这样，整个分子链中链节的排列就可能出现三种不同的有规立构：等规立构（或全同立构）：…*ddd*…或…*lll*…；间规立构（或间同立构）：…*dldldl*…；无规立构：…*dldddldlld*…。另外，当链节中有双链存在时可以出现顺式和反式两种几何立构，以聚1,4-丁二烯为例：

反式聚1,4-丁二烯



顺式聚1,4-丁二烯



这两种不同的几何立构聚1,4-丁二烯具有完全不同的物理状态和性能，后者是一种富有高弹性的橡胶体；前者则为高度结晶的坚硬固体。

二次结构：是高分子链由于主链价键的内旋转和链段的热运动而形成的各种构象。图29(a)表示了高分子链中的局部运动，称为链段运动。产生这种局部运动的原因是高分子链中的每一个C—C链都可以相对地旋转，叫内旋转，如图29(b)所示。分子链段热运动的结果使整个分子链在平衡状态下不可能完全伸直，处于卷曲状，这时分子两端的距离叫末端距，如图29(c)所示。这种末端距的方根值可用来表示分子链的柔性，即均方根末端距越小，表示分子链柔性越大。

分子链的柔性大小对聚合物的二次结构和高次结构有显著影响。而分子链的柔性又是决定于上述一次结构，如果链节结构不对称，或具有极性基团，或具有支链，这些因素都使分子链

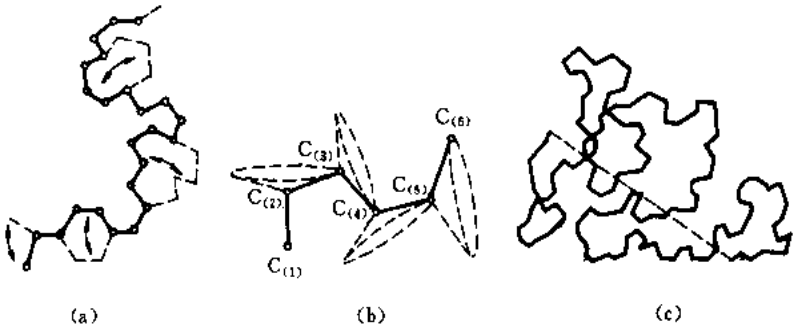


图29 高分子链热运动形成的构象

(a) 链段运动；(b) 内旋转；(c) 卷曲状

柔性减小。

由于分子链节的结构和分子链的柔性不同，聚合物的二次结构通常呈四种基本构象，即：伸直构象、卷曲构象、折叠构象和螺旋构象，如图 30 所示。

图 30　聚合物二次结构——高分子链构象

(a) 伸直；(b) 卷曲；(c) 折叠；(d) 螺旋

三次结构：是聚合物的聚集态结构。聚合物是由无数高分子链聚集而成。这种链聚集起来的形态和结构叫聚集态结构，可分为非晶态或无定形态，晶态和部分晶态。聚合物和低分子晶体结构的明显差别是很少成为完全晶体，常用结晶度这个参数来说明聚合物中晶态和非晶态所占的比例。不同的聚合物，不同的结晶条件，生成的晶态结构也不同。聚合物的晶态结构主要有片晶、球晶和螺旋晶态结构等。

图 31 表示聚乙烯单晶片形成过程的示意图。这是至今发现的聚合物中最完整的晶体，单晶片最长可达 50μm，厚为 10nm，通常是在溶液中缓慢生长而成。

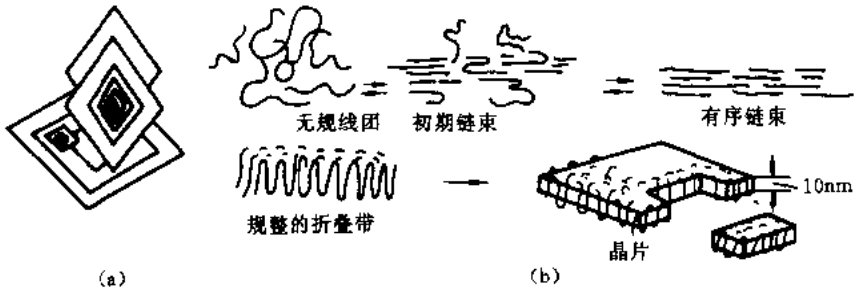


图 31　聚乙烯单晶片(a)及形成过程(b)示意图

当线型聚合物从熔融态慢慢冷却下来时生成球晶，如图 32 所示。整个结构是球晶晶粒分散在非晶态结构内。球晶本身是球形界面内部结构复杂的多晶，在球内部同样存在相当部分的无定形结构。

束晶的结构如图 33(a)所示，分子链中一部分作有规排列成束晶，另一部分则任意排列为非晶结构。如果对这类聚合物进行拉伸定向，可以使各束晶在拉伸方向定向，由此可以获得良好的薄膜或纤维。

聚合物由于结构的多层次，其物理状态及其热行为也与低分子物质有明显区别。低分子物质在温度改变时可处于固态、液态或气

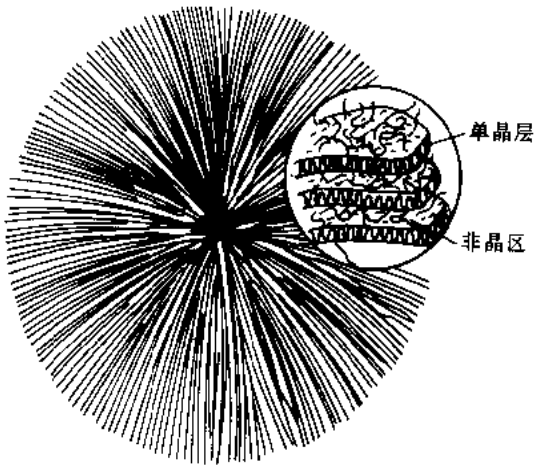


图 32　聚合物球晶

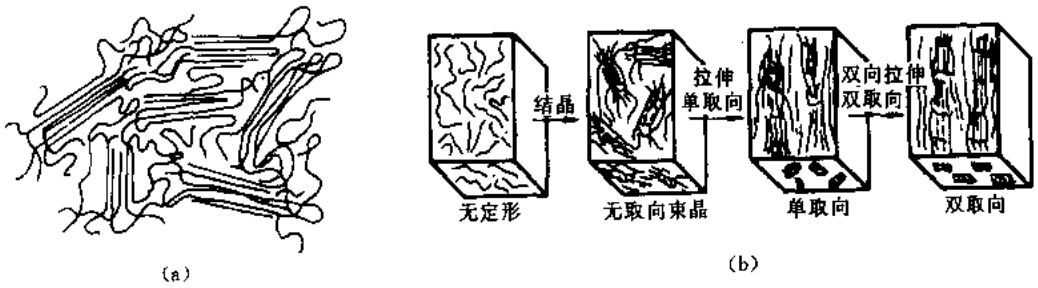


图 33 (a)束晶模型和(b)拉伸生成定向束晶

态。聚合物没有气态,只有固态和液态,而且聚合物处在液态下其粘性很大,故叫粘流态,以资与低分子物质液态的区别。

聚合物的物态和力学性质随温度的变化通常具有如图 34 所示的曲线。图中 T_B 是脆点温度。低于 T_B,聚合物成脆性。超过 T_B,聚合物是质硬超过其韧性的固体,这时分子链的运动形式主要是原子基团和小链段的短程热振动,因此形变很小,属于普通弹性形变,当温度较高时即超过 T_g,分子链段可能在外力作用下发生长程运动,可以产生很大的形变(例如可达 1000%),故叫高弹态。T_g 是聚合物从高弹态向玻璃态转化的温度,叫玻璃化温度。如果温度再升高超过粘流温度 T_f 时,形变不再回复,聚合物进入粘流态,T_f 高低对聚合物的热塑成型很重要。

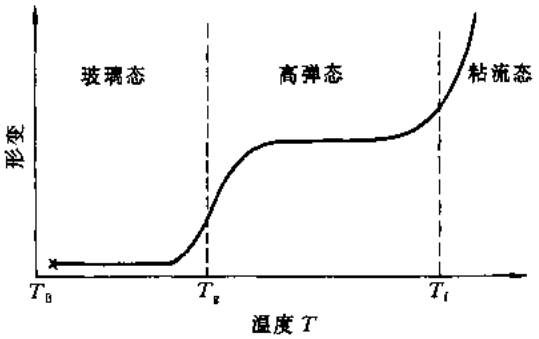


图 34 聚合物的温度形变特性

聚合物这种温度形变特性同样也反映在聚合物介电性能与温度的关系,前面曾提到聚合物的介质损耗的温度曲线上有许多峰值出现,其中有些峰就与聚合物的状态有关,很可能低温 α 峰出现在玻璃态时,对应于较大的松弛时间,β 峰可能出现在高弹态时,对应于弛豫时间的减小。

由于聚合物具有以上这些不同于无机电介质的特有状态和特性,而且加工成型简便,因而得到了越来越广泛的应用。

4.3 电介质的应用

电介质在电工和电子器件和装置中的用途极广泛,根据电介质在器件和装置中所起的作用可分为三类,电绝缘材料、电容器介质及电介质功能器件。

4.3.1 电绝缘材料

凡是应用电的地方都少不了电绝缘,电绝缘的作用是保证带电部分具有所需要的电势和电流,保护带电体之间或对地不发生漏泄或短路。因而绝缘材料必须满足以下各方面的要求:

① 能承受所作用的电压或电场强度,这就要求电介质具有高的击穿场强。在一些电子器件中,有时工作电压仅数伏,但由于其绝缘层厚只有微米或小于微米,因而其电场强度相当高。

② 漏电流低于允许的范围。要求有高的绝缘电阻，而且在工作环境条件下，特别在高温高湿条件下不会恶化。必要时电绝缘表面需作防潮处理。

③ 介质损耗小。特别在高频条件工作的电绝缘，如介质损耗高，不仅品质因数差，而且容易引起热击穿。高频电介质一般都是非极性或极性很小、没有松弛离子、结构均匀的材料。

④ 具有良好的耐热性和耐老化性。特别是对有机电介质材料，一般其允许工作温度较低，如果超温工作将使材料加速老化。

⑤ 易于加工成形，原料丰富，成本低。

根据上述要求，发展的趋向是合成的聚合物电介质在电工绝缘中的应用越来越广泛，许多天然有机电介质如低棉纤维、丝和天然橡胶，越来越多地为塑料薄膜、合成纤维、合成橡胶等产品所取代。电机、电器、电线和电缆中应用的薄膜、纤维、油漆、橡胶和塑料无一不属此类。即使对电子器件的电绝缘，现在也开始对聚合物电介质感兴趣，例如已开始用聚酰亚胺膜作多层布线的绝缘层等。

玻璃可用来制成各种绝缘子和绝缘套管，现在大量用于拉制玻璃纤维，以便于编织，可与树脂组成复合材料，如环氧玻璃层压板广泛用于印刷电路的绝缘基板。

陶瓷电介质具有良好的介电性能和机械强度，又有良好的耐热性和耐气候性，虽然因其加工困难和成本高的原因，已有一部分应用被聚合物电介质所代替，但在高电压、高频场合，仍大量采用陶瓷作绝缘子和其他应用。可靠性要求高的集成电路也仍用陶瓷封装。

单晶材料由于加工成型困难以及价格高，除非常特殊的情况外，几乎很少用于电绝缘。唯一广泛采用的天然材料云母，一直作为大容量电机的重要绝缘材料。

4.3.2 电容器的介质

晶体电介质如果其结构的对称性差，这时会出现自发极化现象，或压电和热电现象等，而且伴随着极化的非线性效应。利用这些特性和效应可以制成电介质功能器件，包括光、声和热等传感器或换能器、放大器件、信息存贮器件和显示器件等。这些器件都是用晶体电介质包括聚合晶体在内来制备的。这些电介质材料除了上面讨论的这些介电性能外，还有一些其他特性。

5 压电性[5,6]

5.1 电压效应

电介质的压电性早在1880年由居里兄弟首先发现，他们观察到当一石英晶体受外力作用时在表面有电荷出现。当时只是作为一种物理现象来研究，并没有发现它的应用价值。直到第一次和第二次世界大战期间，相继发现了一些压电和铁电晶体，才重视了对它的应用研究。1919年有人把罗息盐做成电声元件，后来又用水晶制成谐振器、滤波器、换能器等。到40年代，由于军事上需要，利用压电晶体制造声纳取得成功。到现在为止，已有大量性能更好的压电晶体和陶瓷被发现和应用。

现已查明，在三十二种点群中，只有二十种点群不具有对称中心的电介质晶体才表现出压电性。点群没有对称中心只是必要条件，但还必须满足晶体结构中要有分别带有正负电荷的质点，以及晶体不导电这一充分条件。

晶体的压电效应可用图35来说明其原理。图35(a)表示没有外力作用时晶体中的电荷分

布，正负电荷作用中心相重合，晶体的总电矩等于零，没有极化现象，晶体表面不出现电荷。当对晶体沿某方向施加压力而产生压缩形变时，如图 35(b)所示，这时正负电荷中心不再重合，从而形成电矩，晶体表面出现极化电荷。同样，图 35(c)表示晶体受拉伸时的情况也相仿，只是极化方向相反。如果将这类晶体放在电场中，由于其中电荷受到电场力作用产生位移而造成形变和应力，这效应称为逆压电效应。需要指出，逆压电效应与电介质通常所具有的电致伸缩效应是两种机理，后者与晶体是否有对称中心无关，甚至伸缩的尺寸远比逆压电效应小。

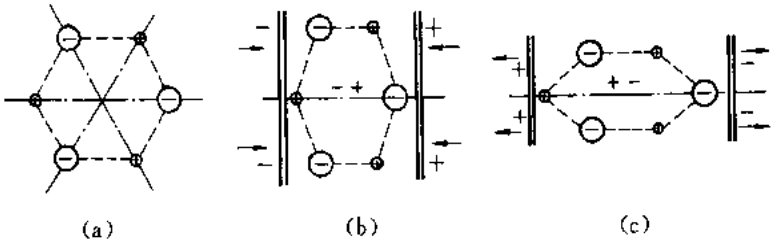


图 35　压电效应示意图

5.2　压电方程组及特性参数

压电现象是电介质极化与弹性形变的相关效应，可用一组方程加以描述，称为压电方程组。各向异性的电介质的电极化具有关系：

$$\mathrm{d}D_i = \varepsilon_{ij}\mathrm{d}E_j \tag{118}$$

式中：D_i 和 E_j 分别是矢量 $\boldsymbol{D}$ 和 $\boldsymbol{E}$ 的分量。$\boldsymbol{D}$ 和 $\boldsymbol{E}$ 是一阶张量，所以介电常数 ε 为二阶张量。如果以 $\boldsymbol{S}$ 表示应变，$\boldsymbol{\sigma}$ 表示应力，则弹性形变与应力的关系为

$$\mathrm{d}S_{ij} = s_{ijkl}\mathrm{d}\sigma_{kl} \tag{119}$$

因为 $\boldsymbol{S}$ 和 $\boldsymbol{\sigma}$ 都是二阶张量，所以弹性柔顺系数 s 是四阶张量。

晶体的压电性或逆压电性是电极化和机械形变之间的转换，可以有不同表示方式，两者之间的关系如图 36 表示。图中给出了不同变量之间的关系常数，通过这些常数可将 $\boldsymbol{D}$、$\boldsymbol{E}$、$\boldsymbol{S}$ 和 $\boldsymbol{\sigma}$ 的关系写成方程组，根据所选自变数的不同，方程组可以有不同形式。

如果取 E_j 和 σ_l 为自变量，可以得到第一类压电方程组：

$$\left.\begin{aligned} D_i &= \sum_{j=1}^{3}\sigma_{ij}^{\mathrm{T}}E_j + \sum_{l=1}^{6}d_{il}\sigma_l \\ S_t &= \sum_{j=1}^{3}d_{tjk}E_j + \sum_{l=1}^{6}s_{kl}^{\mathrm{E}}\sigma_l \end{aligned}\right\} \tag{120}$$

或简写为：

$$\left.\begin{aligned} \boldsymbol{D} &= \varepsilon^{\sigma}\boldsymbol{E} + d\boldsymbol{\sigma} \\ \boldsymbol{S} &= d_{\mathrm{t}}\boldsymbol{E} + s^{\mathrm{E}}\boldsymbol{\sigma} \end{aligned}\right\} \tag{121}$$

式中：ε^{σ} 为恒定压力下的介电常数，也叫自由介电常数；s^{E} 为恒定电场中的弹性柔顺系数，也叫短路弹性柔顺系数；d 为压电应变常数，通常简称压电常数；d_{t} 为 d 的转置矩阵。

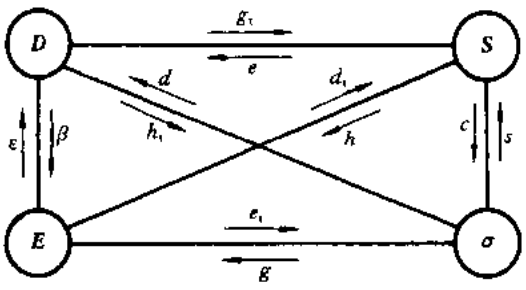


图 36　电极化和机械变形之间的联系

如果取 E_j 和 S_i 为自变量，可得第二类压电方程组，其简单表示形式为

$$\left.\begin{aligned}\boldsymbol{D} &= \varepsilon^{s}\boldsymbol{E} + e\boldsymbol{S}\\ \boldsymbol{\sigma} &= e_{t}\boldsymbol{E} + c^{E}\boldsymbol{S}\end{aligned}\right\} \tag{122}$$

式中：ε^s 为恒定应变时的介电常数，也叫受夹介电常数；c^E 为恒定电场下的(短路)弹性刚度系数；e 为压电应力常数，单位是 C/m^2；e_t 为 e 的转置矩阵。

类似的方法，取 D_j 和 σ_l 为自变量，可得第三类方程组。取 D_j 和 s_l 为自变量，则得第四类压电方程组。由图 37 和上述压电方程组可知，在不同情况下可以选用不同的压电常数 d、e、h 和 g，这四个常数都有转置矩阵 d_t、e_t、h_t 和 g_t。一般来说，对于非铁电性晶体，没有自发极化，常常以电场强度作自变量比较方便，即可以用第一类或第二类压电方程组。对于铁电性晶体，通常取电感应强度为自变量，即采用第三或第四压电方程组较方便。

由于上述介电常数 ε、弹性系数 s 和压电常数 d 都是张量，对具体晶体电介质材料来说，其晶体结构的对称性不同，这些张量中的独立分量中的数目也不同，例如 $LiNbO_3$ 和 $LiTaO_3$ 等晶体为 $3m$ 点群，其 s、d 和 ε 张量分别为

$$\begin{pmatrix} s_{11} & s_{12} & s_{13} & s_{14} & 0 & 0\\ s_{12} & s_{11} & s_{13} & -s_{14} & 0 & 0\\ s_{13} & s_{13} & s_{33} & 0 & 0 & 0\\ s_{14} & -s_{14} & 0 & s_{44} & 0 & 0\\ 0 & 0 & 0 & 0 & s_{44} & 2s_{14}\\ 0 & 0 & 0 & 0 & 2s_{14} & x \end{pmatrix} \tag{123}$$

式中 $x = 2(s_{11} - s_{12})$。

$$\begin{pmatrix} 0 & 0 & 0 & 0 & d_{15} & -2d_{22}\\ -d_{22} & d_{22} & 0 & d_{15} & 0 & 0\\ d_{31} & d_{31} & d_{33} & 0 & 0 & 0 \end{pmatrix} \tag{124}$$

$$\begin{pmatrix} \varepsilon_{11} & 0 & 0\\ 0 & \varepsilon_{11} & 0\\ 0 & 0 & \varepsilon_{33} \end{pmatrix} \tag{125}$$

在实际应用时，这些张量可以在手册中查到。

除了上述描述的晶体材料的弹性、介电常数和电压常数外，还有三个参数对压电材料来说也是十分重要的，即介电损耗 $\tan\delta$、机械品质因素 Q_m 和描述机械能和电能互相转换的机电耦合系数 K。关于介质损耗，前面已有详细讨论。下面就 Q_m 和 K 作简单介绍。

当压电晶体产生弹性谐振时，因克服内摩擦而损耗能量，造成机械损耗，Q_m 表示机械损耗的程度，定义为

$$Q_m = 2\pi \frac{\text{振子在谐振时储存的机械能量}}{\text{振子在谐振 1 周内的机械损耗能量}}$$

机电耦合系数 K 的定义为

$$K = \sqrt{\frac{\text{通过压电效应转换的电能}}{\text{储入机械能总量}}}$$

也可以表示为

$$K=\sqrt{\frac{\text{通过逆压电效应转换的机械能}}{\text{储入电能总量}}}$$

K 不仅由材料性能所决定，而且与振子的几何形状和振动模式有关。

5.3 压电材料及其应用

压电材料可分为单晶和陶瓷两类。50 年代前主要应用压电晶体，有水晶（石英晶体）、罗息盐（$NaKC_4H_4O_6\cdot 4H_2O$，缩写为 R_s）、磷酸二氢钾（KH_2PO_4，缩写为 KDP）等。50 年代后压电陶瓷获得迅速发展，在许多应用方面取代了上述这些水溶性压电晶体，特别是钛酸钡一类的压电陶瓷，在声学和低频超声换能器以及低中频滤波器等方面广泛地使用。进入 60 年代后，随着高频和超高频技术的发展，压电陶瓷材料因其高频损耗太大而显得不能满足新的需要。因而又有一批新的压电晶体材料被发现和使用，例如镓酸锂、锗酸铋、锗酸锂等。另外也使用新的铁电晶体具有的压电性，如铌酸锂、钽酸锂等，还有不少半导体压电体如硫化镉、氧化锌等。随着压电器件频率不断提高，特别是进入微波声学以及器件小型化的要求，人工生长的压电晶体和压电薄膜就成为主要的材料。下面就几种典型的压电材料作简要介绍。

到目前为止，石英晶体仍然是高频振子的主要材料，SiO_2 结晶可以有三种形态，即石英、鳞石英和方石英。这些晶体的每一种又有高温型（β-型）和低温型（α-型）两种类型，作为压电材料的主要是 α-石英。α-石英和β-石英的温度转变点为573℃，β-型比α-型具有较高的对称性。水晶的人工生长目前最成功的是用“水热合成法”（简称水热法）。这种方法是在高温高压条件下，利用水溶液的温度梯度去溶解在通常条件下不溶于水的物质，并使之结晶。

锗酸铋（$Bi_{12}GeO_{20}$）的熔点为 935℃，属立方晶系点群 $23mm$，在熔点以下无其他相变。压电效应是由于在 GeO_2 正四面体中 Ge 离子在外力作用下沿[111]方向位移而引起的。因为不具有唯一单向极轴，故没有热电性和铁电性。由于立方晶格结构可消除热膨胀系数变化的影响，所以这种晶体用来产生纯振动模式是比较理想的，是一种很好的高频换能器材料。另外，其声波传播速度低，约为 $LiNbO_3$ 的 1/2，是微声延迟线的理想材料。锗酸铋具有一次和二次电光效应和光电导效应，但这方面的参数并不优越，目前很少用作电光或光电导材料，但可以作为声光材料使用。

化合物半导体如Ⅱ-Ⅵ族的 ZnS、CdTe、CdS 和 ZnO 等以及Ⅲ-Ⅴ族的 GaAs、InSb 和 InP 等都具有压电性。这类晶体的晶格结构一般有两种：一是立方晶系闪锌矿结构，点群属 43m，与金刚石结构十分相似，CoTe 和 GaAs 等晶体属于这类；另一种是六角晶系纤维锌矿结构，点群为 6mm，CdS 和 ZnO 等晶体属于这类。ZnS 晶体的结构这两种都可能。这类晶体的压电效应可用等效电荷理论模型来解释。对许多Ⅱ-Ⅵ族化合物半导体的压电常数测量表明，压电效应大小与相对原子质量之间有一定关系。实验结果表明，压电系数随Ⅱ族金属元素相对原子质量的增大而增大，随Ⅵ族非金属元素相对原子质量的增大而减小。压电常数和“等效原子电荷”之间有简单关系。计算所得到的等效原子电荷值是一个电子电荷的几分之一或稍少于电子电荷，这等效原子电荷随金属相对原子质量的增大而增大，随非金属相对原子质量的增大而减小，与化学负性完全一致。这说明这类化合物的化学键是极性共价键，介于共价键和离子键之间，离子性愈强，即等效原子电荷愈大，压电效应愈明显。

压电半导体主要用作高频超声以及 300MHz 以上的微波声学的体声波和表面声波换能器

和声放大器材料。声放大器要求同时利用晶体的压电效应和半导体特性，其工作原理如图 37 所示。高频输入电信号经过输入换能器变成超声波在晶体内传播，由于晶体有压电效应，传播的超声波激发出纵向电场行波。与此同时，晶体两端加上直流或脉冲的漂移电场，使压电半导体内的载流子随声波前进。当载流子的漂移速度超过声速时，发生声子-电子的相互作用，能量就从载流子转移到声波，使信号放大。但是，这种放大器对材料各方面性能要求很高，往往噪声大、损耗大，给实用化带来困难。后来又发展了表面声波放大器，可以在压电材料上生长一层半导体薄膜，或在半导体基板上生长一层压电薄膜而制成。

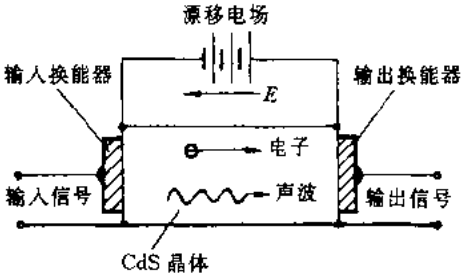


图 37　压电半导体声行波放大器工作原理

近年来发现聚合物晶体薄膜如聚偏氟乙酸也具有压电性，这留待后面讨论。

6　热电性[5,6]

6.1　热电效应

除了上述由机械应力的作用引起电极化的压电性外，其中有一部分晶体可以由于温度的变化而产生电极化。热电效应与温差电效应不同，后者是由于物体的不同部位具有温度差而引起了电势差。热电效应则是由于晶体本身存在自极化的结果。这种自发极化与前面所讨论的由外电场产生的感应极化或外力引起的压电效应不同，这是由于材料本身的结构中在某些方向上正负电荷作用中心不重合而造成的。因而即使没有外电场或外应力，这种自发极化也会产生。当温度恒定时，晶体的自发极化不易发现，因为自发极化使材料表面所带电荷往往被体内的自由电荷或吸附的表面电荷所补偿。只有当温度改变时，自发极化将随温度而变，这时材料表面吸附的电荷也随之而变，这样才有可能发现和测量。

不是所有不具对称中心的晶体都具有自发极化。实际上，在具有压电性的 20 种没有对称中心的点群中，只有 10 种点群呈现热电性。例如前面提到的 α-石英，它具有明显的压电性，但没有热电性。原因是在(0001)面上三个极轴是完全相同的，温度改变时，三个极轴上正负电荷作用中心的位移相同，如图 38 所示。这说明晶体有一个与其他极轴不相同的极轴时，才有可能由于热膨胀而造成总电矩的变化，并从而出现热电性。这 10 种晶体因而也叫电极性晶体。

热电性现象最早在电气石$(Na,Ca)(Mg,Fe)_3B_3Al_6Si_6(O,OH,F)_3$中观察到。当均匀加热这晶体时，能在唯一的三重旋转对称轴两端形成等量异号的电荷，如果晶体冷却，得到的电荷符号相反。后来又不断发现了许多具有热电性的晶体，如钛酸钡、硫酸三甘肽等。早在 1938 年曾经有人提出利用热电效应探测红外辐射，但长期以来没有引起重视，直到 60 年代，才有较多的研究。由于激光、红外扫描成象技术的迅速发展，要求寻找不用冷却的新的红外探测方法，这对军事上有重要意义。在工业、医疗等方面要求非接触式温度测量，以及其他方面都要求使用小型高性能的红外探测元件。热电晶体具有高灵敏、快响应和宽频谱等优点，具有很大的应用潜力。

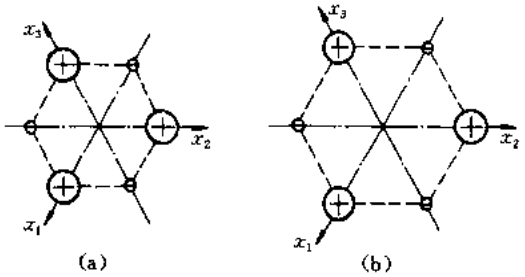


图 38　α-石英热膨胀时电荷分布

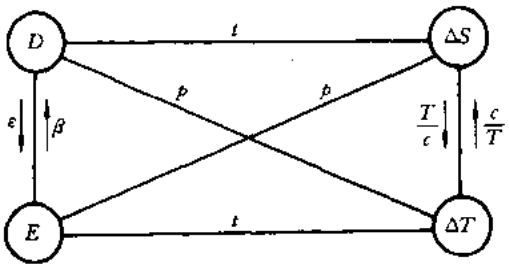


图 39　电性参数与热性参数关系

6.2　热电方程组和特性参数

与压电方程组相类似，可以用热电方程组来描述电极化热行为之间的相关性。如果用熵变 ΔS 和温度变化 ΔT 表示热行为的参数，则电性参数和热性参数之间的联系可用图 39 表示。图中 c 表示比热，p 和 t 为热电系数。

如果晶体内有微小温度变化(是均匀变化，没有温度梯度)，则极化强度的变化与之成正比：

$$\mathrm{d}P_i = p_i \mathrm{d}T \tag{126}$$

式中极化强度 $\boldsymbol{P}$ 是一阶张量，温度 T 为标量，所以热电系数 $\boldsymbol{p}$ 也是一阶张量，应有三个分量。

应该指出，热电体总是压电体。当温度变化时，自由晶体的体积发生变化。从而由压电性产生二次极化，并叠加于一次热电极化上，使热电现象大大复杂化。如果加热不均匀，则晶体内造成热应力，将会产生三次热电效应。所以相应有一次热电系数、二次热电系数和三次热电系数。为简单起见，我们下面只讨论一次热电系数，它相当于晶体加热，而形状和大小保持不变时的热电系数，这状态称为束缚晶体。对应于前面提到的自由晶体，如温度改变时，保持晶体中电场不变，则

$$\mathrm{d}D_i = \boldsymbol{p}_i \mathrm{d}T \tag{127}$$

式中 $\boldsymbol{p}$ 的单位是 $\mathrm{C}/(\mathrm{m}^2 \cdot ℃)$。

如果以 E_i 和 T 作为独立变量，可以写出方程组

$$\left.\begin{aligned} \mathrm{d}D_i &= \left(\frac{\partial D_i}{\partial E_i}\right)_T \mathrm{d}E_i + \left(\frac{\partial D_i}{\partial T}\right)_{E_i} \mathrm{d}T \\ \mathrm{d}S &= \left(\frac{\partial S}{\partial E_i}\right)_T \mathrm{d}E_i + \left(\frac{\partial S}{\partial T}\right)_{E_i} \mathrm{d}T \end{aligned}\right\} \tag{128}$$

根据热力学定律，系统的吉布斯自由能函数为

$$G = H - TS = U - TS - E_i D_i - T_j S_j \tag{129}$$

式中 H 为焓，$H = U - E_i D_i - T_j S_j$，系统假设为束缚晶体，忽略 $T_j S_j$ 项，对 G 求微分可得

$$\mathrm{d}G = -S\mathrm{d}T - D_i \mathrm{d}E_i \tag{130}$$

式中 G 为 E_i 和 T 的函数，G 的全微分为

$$\mathrm{d}G = \left(\frac{\partial G}{\partial T}\right)_{E_i} \mathrm{d}T + \left(\frac{\partial G}{\partial E_i}\right)_T \mathrm{d}E_i \tag{131}$$

式中 $\left(\frac{\partial G}{\partial T}\right)_{E_i} = -S$，$\left(\frac{\partial G}{\partial E_i}\right)_T = -D_i$

求导可得

$$-\frac{\partial^2 G}{\partial T \partial E_i} = \left(\frac{\partial D_i}{\partial T}\right)_{E_i} = \left(\frac{\partial S}{\partial E_i}\right)_T = p_i \tag{132}$$

如计及$\left(\frac{\partial D_i}{\partial E_i}\right)_T = \varepsilon_{ij}^T$，$\left(\frac{\partial S}{\partial T}\right)_{E_i} = \frac{c^E}{T}$。其中：$\varepsilon_{ij}^T$为等温介电常数；$c^E$为恒定电场下的比热。

当 E_i 和 T 变化范围不大时，式(128)可写为

$$\left.\begin{aligned} D_i &= \varepsilon_{ij}^T E_j + p_i \Delta T \\ \Delta S &= p_i E_i + \left(\frac{c^E}{T}\right)\Delta T \end{aligned}\right\} \tag{133}$$

这就是以 E_i 和 ΔT 为自变量的热电方程组，选用不同的自变量，也可写成不同的热电方程组。如果同时考虑电、热和弹性三者之间的相关效应，则具有更复杂的关系，这里不再作进一步的讨论。

6.3 热电材料及其应用

热电材料目前主要用于红外探测，包括工业和空间技术所用的各类辐射计、光谱仪以及红外激光探测和热成象管等。图 40 是一种热电探测器的工作原理图。热电探测器的突出优点是能在室温下工作，比光探测器有更宽的频谱，具有高频响应，适用于 X 射线到毫米波范围内

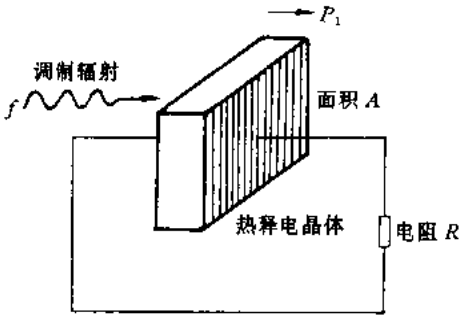


图 40 热电探测器工作原理

下面简单分析热电探测的工作原理及对材料的性能要求。热电晶体的自发极化在表面上形成的束缚电荷可被体内和外部来的自由电荷所补偿，这过程的弛豫时间

$$\tau = \varepsilon\rho \tag{134}$$

式中：ε 为介电常数；ρ 为电阻率。大多数热电晶体的 τ 在 1 ~ 1000s。所以这类红外探测器对所测的辐射要进行调制。图 40 中的 f 为红外辐照的调制频率，因而造成晶体温度、晶体自发极化以及由此引起的束缚电荷均随频率 f 而周期变化。如果频率很低，$f < 1/\tau$，束缚电荷始终被体内自由电荷所中和，探测器不能工作；若 $f > 1/\tau$，体内电荷跟不上束缚电荷的变化，晶体两端出现交流电势。如果接通负载，将有电流流过，这就是热电探测器的工作原理，设探测器面积为 A，负载电阻为 R，则输出信号为

$$\Delta V = AR_L \frac{dP_s}{dt} = AR_L \frac{dP_s}{dT}\cdot\frac{dT}{dt} = AR_L p \frac{dT}{dt} \tag{135}$$

式(135)表明，输出信号与热电系数 p 成正比，而且与温度变化速度成正比，而不取决于是否达到热平衡。热电系数 $p = dP_s/dT$ 的值可以在 P_s—T 曲线上求得。p 大，表示自发极化随温度变化显著。图 41 给出了硫酸三甘肽(TGS)和钛酸钡($BaTiO_3$)两种类型的 P_s—T 曲线。TGS 的 P_s 在居里点附近连续变化到零，这是二级相变的情况；$BaTiO_3$ 的 P_s 在居里点附近产生突变，这相当于一级相变的情形。

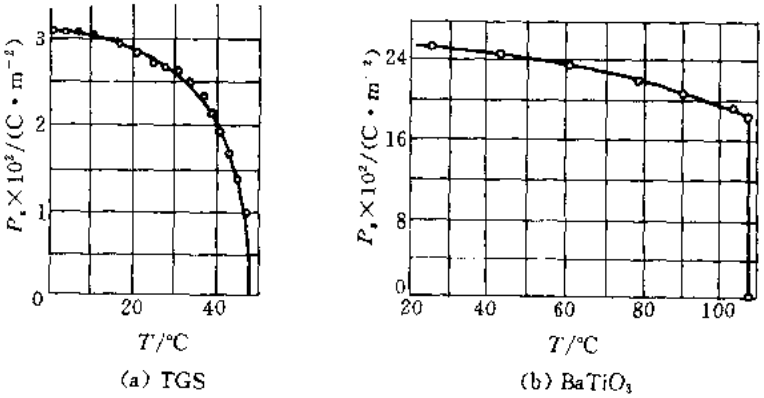


图 41　自发极化随温度的变化

比较上述两种情况，$BaTiO_3$ 在远离居里点时，dP_s/dT 太小，在居里点附近虽然 dP_s/dT 变得很大，但起伏太大，且晶体易退极化，并且噪声大。

对红外探测材料除要求热电系数高外，还要求受红外辐射后温度上升要快，这就要材料对红外线的吸收大、比热小、密度轻，容易加工成很薄的晶片。另外，还要求介电常数 ε 小，介电损耗 $\tan\delta$ 小，这对高频应用特别重要

迄今为止，能满足上述各方面要求的实用材料数量并不很多。用得较多的还是 TGS 晶体，其优点是易于生长和加工成大面积的晶片，在大的频率范围内灵敏度高。缺点是临界温度低，晶体容易退极化，而且水溶性晶体易于受潮解。

硫酸锂($Li_2SO_4 \cdot H_2O$)不是铁电体，没有居里点，不会退极化，其介电常数低，适宜于高频下大面积用。其缺点也是一种水溶性晶体。

铌酸锶钡($Sr_{1-x}Ba_xNa_2O_6$)SBN 晶体在 $0.25 \leqslant x \leqslant 0.75$ 呈四方钨青铜结构，属于 4mm 点群，具有较高热电系数和一次电光系数，半波电压低，不潮解，机械性能好。但介电常数大，高频品质因数比 TGS 小一个数量级，晶体生长比较困难，不易得大截面积晶体，故宜于在低频和小面积条件下使用。

近年来，对能用于红外探测的材料又不断有新的发现，例如，认为锗酸铅($Pb_5Ge_3O_{11}$)和硒砷化铊(Tl_3AsSe_3)这两种材料用于红外探测器有很好的潜力。Tl_3AsSe_3 的热电系数是迄今最高的，介电常数小，非铁电体，没有居里点和退极化问题，低频品质因数比 TGS 高，只是电阻率较低。另外，聚二氟乙烯(PVF_2)和聚氟乙烯(PVF)等聚合物薄膜用于红外探测有很多优点，可以大面积加工成任意形状。

在选用热电材料时，要考虑到各方面的因素。首先是要求材料的热电品质因数大。由于影响元件性能的材料特性参数很多，很难用一个简单的品质因数充分反映。目前常用三个品质因数进行评价：第一品质因数或叫第一优值 $Q_{pr1} = p/c\varepsilon$。其中 p、c 和 ε 分别为热电系数、比热和介电常数，它们反映热电效应的电压灵敏度，通常用它来比较在高频和大面积使用条件下的各种材料；第二品质因数或第二优值 $Q_{pr2} = p/c$。它反映热电效应的电荷灵敏度，适用于评价在低频小面积条件下使用的材料；第三品质因数或第三优值 $Q_{pr3} = \dfrac{p}{c\sqrt{\varepsilon}}$。适用的频率范围介于前两者之间。

其他方面的考虑应包括，材料没有或不易退极化，能抗辐射，材料的化学和物理性能稳定，

易于加工成型，成本低廉等。

7 铁电性[4,5]

7.1 铁电现象

铁电体与上述热电体一样是具有自发极化的材料，但铁电体与热电体相比有两个明显区别的现象。

(1) 电滞回线。像铁磁体那样，铁电体的极化强度与电场强度之间也呈回线，叫电滞回线，如图 42 所示。铁电性因而得名，实际上铁电材料都不含 Fe 元素。热电体虽有自发极化，但不出现回线。这说明铁电体中自发极化在外电场作用下能转向，具有相同自发极化方向的区域叫电畴。

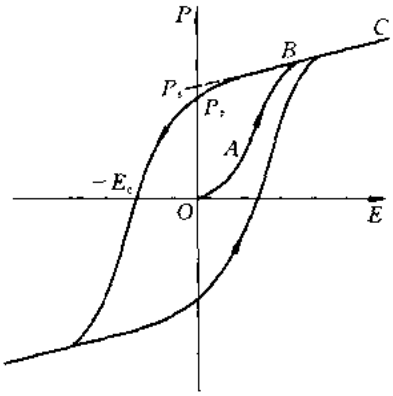


图 42 铁电体的电滞回线

图 42 为铁电体的电滞回线。当没有外场时，晶体的总电矩为零，相当于图上的 O 点。加上电场后，有一部分不在电场方向的电畴转向电场方向，因而极化强度沿 A 迅速增大。当达到 B 点时，整个晶体成为一个单畴体，自发极化的转向已趋饱和，如果电场 E 继续增大，只有感应极化能继续增大。从 B 到 C，一般具有线性关系，如将 CB 延长在纵坐标轴上的截距 P_s，即为自发极化强度。但如果从 C 开始减小电场，晶体的 P 并不能沿 CB 的延长线到 P_s，而是与纵轴相交于 P_r。P_r 值比 P_s 低，叫剩余极化强度，这说明当电场去除后，有一部分自发极化将离开电场方向，但不是全部。只有当电场在反向达到某值 $-E_c$ 时，晶体的总电矩才为零。此 $-E_c$ 叫矫顽场。P_s、P_r 和 E_c 是铁电性的重要参数，P_s 愈大，说明自发极化强，P_r 与 P_s 相差愈小，表明材料愈接近单畴体，E_c 愈大，说明铁电性愈稳定，但如果矫顽场强 E_c 超过材料的击穿电场强度，则无法使自发极化转向，这就很难说这材料是铁电体。

(2) 临界特性。铁电体与热电体现象的另一重要区别是前者具有临界特性，铁电体呈现电滞回线具有特定的温度范围。当温度超过某临界值或临界范围时，晶体发生相变，铁电性消失。这临界温度叫居里温度或居里点，常用 T_c 表示。一般说，在居里点以上，晶体属于非极性结构，也叫顺电结构。晶体的铁电结构通常是由顺电结构发生微小的晶格变化而产生的，居里点是晶体顺电-铁电相变温度。以 $BaTiO_3$ 为例，在 T_c 以上，顺电相的晶体结构属于立方晶系。氧原子构成氧八面体，金属原子 Ba 和 Ti 分别位于不同的氧八面体中心。由于这种晶格结构是对称的，没有自发极化。

当温度低于 120℃时，$BaTiO_3$ 晶格结构改变，这是由于 Ti 和 O 原子的位置产生了相对位移，这时 $BaTiO_3$ 已从立方晶系转变为四方晶系，即其中 c 轴的长度不再与其他两轴相等，对称性下降，属于 $4mm$ 点群，晶体在 c 轴方向出现自发极化并具有铁电性。

如果温度更低，在 -5℃附近，$BaTiO_3$ 晶体将又一次发生相变，这时晶格转变为正交晶系，对称性进一步下降，为点群 mm^2，三个晶轴都不相等，晶体同样有自发极化，只是方向转到原来立方晶系的[011]方向。

$BaTiO_3$ 还有第三个相变温度，在 -80℃附近晶格变成三角晶系，为点群 3m，这时元胞的三个棱相等，但不垂直 $a=b=c$ 及 $\alpha=89°52'$，自发极化沿原来立方晶格的[111]方向。$BaTiO_3$ 自发极化 P_s 和方向随温度的变化如图 43 所示。在三个相变点中，只有温度最高的那个是顺电-铁电相变点，可叫居里点。其余两个相变点则是从一种铁电相到另一种铁电相的转变，不能算居里点。

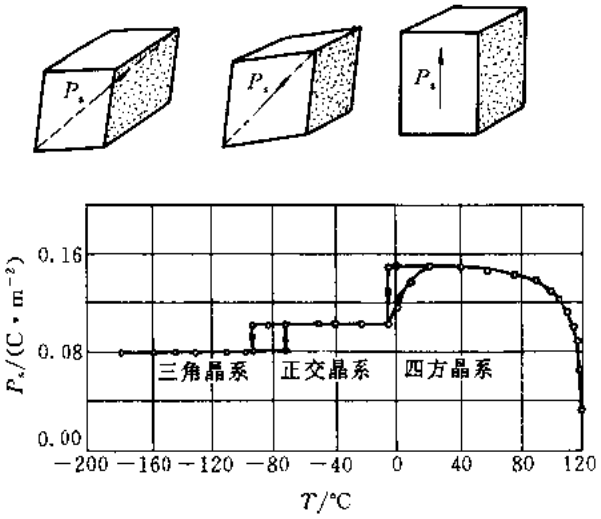


图 43　$BaTiO_3$ 的相变和 P_s 的温度特性

也有一些铁电体具有上下两个居里点 T_{c_1} 和 T_{c_2}，只有在 $T_{c_2}<T<T_{c_1}$ 范围内为铁电相，在这两温度范围以外，$T<T_{c_2}$ 和 $T>T_{c_1}$ 都是顺电相。图 44 所表示的罗息盐就是一个例子。罗息盐有两个居里点 +24℃和 -18℃，在这两温度之间是铁电相属单斜晶系，点群 $2mm$。在这温度范围以外为顺电相，是正交晶系，点群为 222，有压电性。如罗息盐的部分 H 原子被同位素 D 取代，成为氘代罗息盐 $NaKCH_4D_2O_6\cdot 4H_2O$，其铁电相的温度范围扩大到 +35 ~ -22℃。

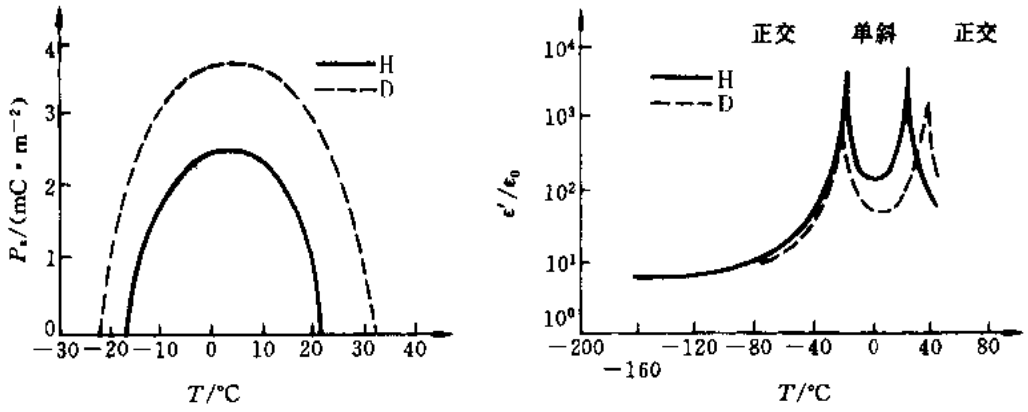


图 44　罗息盐 P_s 和 $\varepsilon/\varepsilon_0$ 的温度特性

一般来说，晶体顺电-铁电相变有两种类型，即一级相变和二级相变。发生一级相变的铁电体其比热随温度出现突变，并伴随有潜热产生，自发极化在居里温度下突然下降到零。$BaTiO_3$就是一级相变铁电体的例子。如果发生二级相变，只有比热突变而无潜热出现，自发极

化也将连续变到零，罗息盐的相变就属于此类。

7.2 铁电理论

为了对上述铁电现象的机理能作出解释，提出了不少的理论和模型，概括起来大致可分成三类：微观模型理论、唯象理论和晶格动力理论。各种理论都能从某一方面对铁电现象的机理进行描述，相互补充，但迄今还不能说铁电理论已十分完整。

7.2.1 微观模型理论

按照自发极化机构的不同，通常把铁电体分成二类：位移型铁电体和有序无序型铁电体。对位移型铁电体来说，在发生顺电-铁电相变时，离子从高对称的平衡位置向低对称的位置发生位移，从而造成自发极化，钙钛矿结构的晶体如 $BaTiO_3$、$SrTiO_3$ 和 $LiNbO_3$ 等都属于这一类。

有序无序铁电体大部分是水溶性晶体，如罗息盐 KH_2PO_4、KDP、$(CH_2NH_2COOH)_3H_2SO_4$ - $6H_2O$(TGS)和亚硒酸锂($LiH_3(SeO_3)_2$)等，这些类型的铁电体都含有 O—H—O 或 O—H—N 氢键，这些氢键把负离子团如$(SO_4)^{2-}$、$(SeO_4)^{2-}$等吸引在一起形成偶极。在居里点以上，这些偶极作无序排列，温度降到居里点以下时，由于氢键作有序排列，从而使这些负原子团也进入有序状态，从而形成自发极化。

在位移型铁电体中，无论是顺电相或是铁电相，离子的平衡位置只有一个，而在有序无序铁电体中偶极的定向方式有两个以上的自由度，这两种铁电体在电场作用下自发极化转向的机构也不同，前者是由于离子的位移，后者是因为偶极转向，所以响应的频率范围有很大差别。例如，$BaTiO_3$ 在 24GHz 以上才出现明显的介电损耗，而罗息盐在 3 GHz 已几乎显示不出铁电性质。

不论是对位移型铁电体或是有序无序铁电体，不同的作者都针对具体的晶体结构提出了各种形成自发极化的微观模型或假设。例如对 $LiNbO_3$ 来说，当温度下降到居里点(1240℃)以下时，Li^+ 和 Nb^{5+} 相对于氧原子层有微小位移，因而正、负电荷作用中心沿 c 轴方向偏离，出现自发极化。图 45 表示 $LiNbO_3$ 从顺电相转变为铁电相时 Li^+ 和 Nb^{5+} 位移的情况。从图上可看出，这两种离子都沿一方向移动，$LiNbO_3$ 晶体的自发极化仅沿 $+c$ 或 $-c$ 轴向，其他方向不产生电矩，所以这种晶体只出现正反平行取向的两种电畴。

一般来说，铁电体的微观模型理论能对自发极化机构提供具体的物理图像，但由于这模型

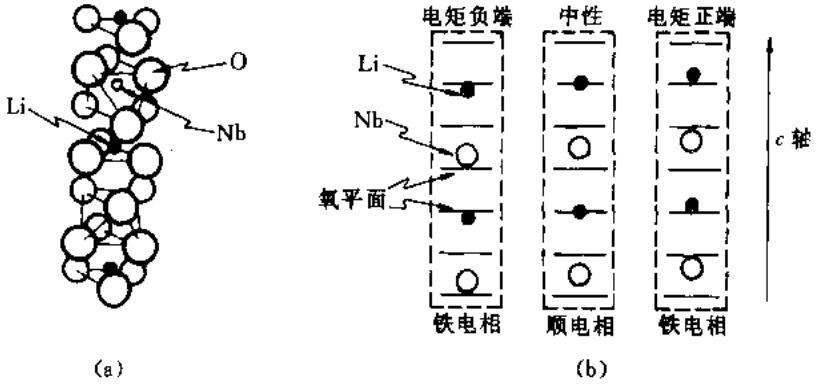


图 45 (a) $LiNbO_3$ 晶格结构；(b) 铁电相的离子位移

是以特定的晶体结构出发而建立的，所以得到的结论往往有很大的局限性。

7.2.2 热力学唯象理论[4]

唯象理论和上述的微观理论相反，几乎完全不考虑晶体的具体结构，只是从宏观特性对热力学状态函数的关系进行分析，从而对铁电现象作出解释，具有较大的普遍性。下面就 Devonshire 的唯象铁电理论作简单介绍。

假设在晶体中，一离子偏离平衡位置产生微小位移，形成偶极，产生局部电场。如果电场力超过了弹性回复力，离子将进一步位移。当位移达到可观程度，出现了非谐回复力，这时离子处于新的平衡位置，晶体转变为铁电相。因而，力的平衡关系为

$$A E' = Bx + Cx^3 + Dx^5 \tag{136}$$

式中：E'为实际作用于离子的局部有效电场；x 为离子的位移；A、B、C 和 D 分别为比例常数；x^3 和 x^5 就是上述的非谐分量。由于产生的偶极矩和极化强度 P 正比于位移，所以式(136)也可表示为

$$E' = aP + bP^3 + cP^5 \tag{137}$$

根据内电场关系，采用 $E' = E + fP$(f 表示内电场系数)，代入式(137)，得

$$E = (a - f)P + bP^3 + cP^5 \tag{138}$$

因而，晶体由于电极化而具有的内能为

$$U = \int E\mathrm{d}P = \frac{(a - f)}{2}P^2 + \frac{b}{4}P^4 + \frac{cP^6}{6} \tag{139}$$

下面首先讨论铁电相。有两种情况是我们特别感兴趣。

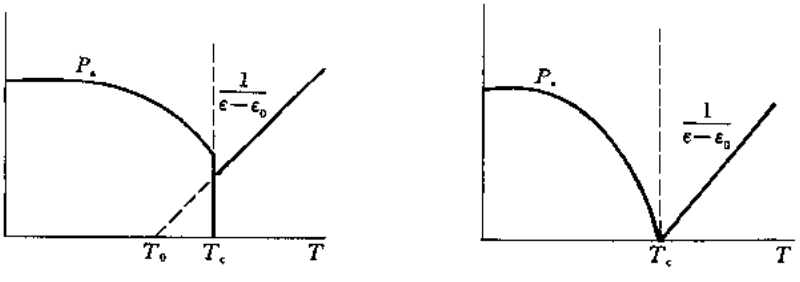
如图 46 所示，在情况 I 中，$(a - f)$及 c 为正值，b 为负值。自发极化 P_s 在居里温度 T_c 时从零值不连续地跳到某一定值。而零值相当于介电相(顺电相)，非零值相当于铁电相，因为这两状态在 $T = T_c$ 时同时存在，即在 $T = T_c$ 时必须满足：

$$\frac{(a - f)P_s^2}{2} + \frac{bP_s^4}{4} + \frac{cP_s^6}{b} = U_o = U_p = 0 \tag{140}$$

式中：P_s是自发极化；足标 p 和 o 分别表示极化和非极化。P_s的存在不依靠外电场 E，因而从式(138)可得

$$(a - f) + bP_s^2 + cP_s^4 = 0 \tag{141}$$

从式(140)和(141)可求出以下的解：



(a) 情况 I　　(b) 情况 I

图 46　两种相交情况的 P_s 和 $1/(\varepsilon - \varepsilon_0)$与温度的关系

$$\left.\begin{aligned} P_s^2(T_c) &= -\frac{3b}{4c} = -\frac{4(a-f)}{b} \\ P_s^4(T_c) &= \frac{3(a-f)}{c} \end{aligned}\right\} \tag{142}$$

在情况Ⅱ中，$(f-a)$和 b 是正值，而 c 可以忽略不计。显然，当 $c=0$ 时，式(140)和式(141)在 $T=T_c$ 时的解有

$$P_s^2(T_c)=0, \quad f-a=0 \tag{143}$$

其次，我们再来讨论介电相，考虑到介电相时，P 比较小，高次项 P^3 和 P^5 与线性项 P 相比可以忽略不计，这样式(138)的微分为

$$\varepsilon-\varepsilon_0=\frac{\partial P}{\partial E}=\frac{1}{(a-f)} \tag{144}$$

对于情况Ⅱ，因为在 $T=T_c$ 时，$a-f=0$，可以将 $a-f$ 展开成 $T-T_c$ 的幂级数。如果只保留一次项，则 $\varepsilon-\varepsilon_0$ 具有如下形式

$$\varepsilon-\varepsilon_0=\frac{\beta}{T-T_c} \tag{145}$$

对于情况Ⅰ，在 $T=T_c$ 时，$(a-f)$不等于零，因而必须在另一个温度 T_0 下使 $a-f=0$。所以这时具有

$$\varepsilon-\varepsilon_0=\frac{\beta}{T-T_0} \tag{146}$$

以上所得到的 P_s 和 ε 作为温度的函数可以用图 46 表示。图(a)和(b)表示的情况Ⅰ和情况Ⅱ在热力学中分别叫一级相变和二级相变，情况Ⅰ中包含有相变潜热，情况Ⅱ中则没有。

下面我们再进一步讨论自发极化和外加电场的关系。式(138)能变换成以归一化量表示的普遍式

$$e=2pt+4p^3+2p^5 \tag{147}$$

式中 $p=\left(-\frac{2c}{b}\right)^{1/2}P$，$e=-\frac{4}{b}\left(-\frac{2c}{b}\right)^{3/2}E$，$t=-\frac{2}{\beta b}\left(-\frac{2c}{b}\right)(T-T_0)$

当 t 具有不同数值时，p 与 e 的关系如图 47 表示。$\partial e/\partial p$ 为负值的区域表示不稳定状态，

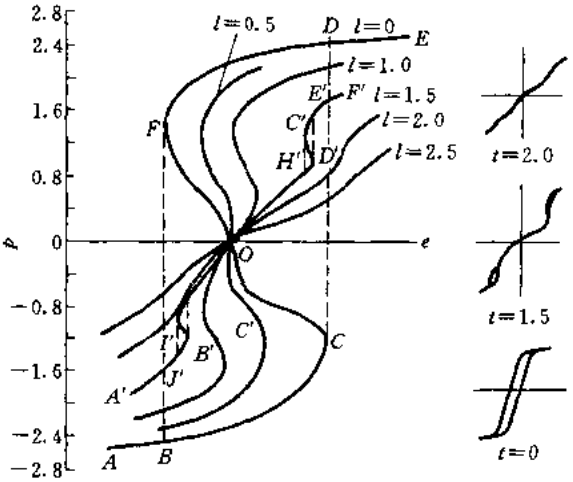


图 47　不同温度下极化与电场强度的归一化曲线

可以观察到两种不同形式的电滞回线，一种回线在图上沿着 *ABCDEDFBA* 轨迹，另一回线沿 *A'B'C'OD'E'F'G'H'OI'J'A'*，它们取定于 t 或 $T-T_0$ 的值。这种蝶形回线已从实验中观察到。

7.2.3 铁电软模理论

上面所讨论的唯像理论虽然可对铁电性的主要现象作出符合实际的解释。但这是一种宏观处理的方法，不能把这些参量与晶体的结构联系起来。60 年代初期，有人从晶格动力学出发提出了铁电性的软模理论，其既具有一定的普遍性，又能与晶体微观结构相联系，对位移型铁电体可以作出非常满意的解释。

在晶格动力学中，有一个著名的 LST(Lyddane-Sachs-Teller)关系：

$$\frac{\omega_{L0}}{\omega_{T0}}=\sqrt{\frac{\varepsilon_s}{\varepsilon_\infty}} \tag{148}$$

式中：ω_{L0}是晶格振动纵波光学模的频率；ω_{T0}是晶格振动横波的频率；ε_s 和 ε_∞ 分别为静态和高频下的介电常数。在前面讨论中我们已知在居里点时，ε_s 将趋于无限大，因而这时将出现 $\omega_{T0}\to 0$ 的情况，即横波光学模的频率接近于零。

$\omega_{T0}\to 0$ 的物理意义是弹性系数很小，当离子偏离平衡位置时，几乎没有弹性力使它们回复原位，这将引起晶格结构的改变。这说明晶体从顺电相转变为铁电相，必然与晶格振动横波光学模中某一模的频率将随温度的下降而减小到零相对应，并将此模称为软模，相应的声子叫光学软声子。

7.3 铁电材料及其应用

与热电晶体相同，铁电晶体所属的点群有 10 种。目前已发现具有铁电性的晶体有 1000 多种，从化学组成来看，品种较多的是：双氧化物(如钛酸盐、铌酸盐等)、硫酸盐、硝酸盐、磷酸盐、砷酸盐和酒石酸盐等。自发极化强度从 $10^{-3}\sim 10^{2}\mathrm{C/m^2}$。

在早期，对铁电材料的应用主要是利用它们的压电性、热电性、电光性能以及高的介电常数等，而对铁电性的应用发展不快，曾有人提出用钛酸钡单晶的铁电畴开关来作为存储器件，但效果不理想，一直没有大量应用。近年来，随着许多新的铁电材料的发现和薄膜生长技术的发展，使铁电性在信息存储，图像显示和全息照相等方面的应用有了新的开端。

对铁电性的应用有三种情况。第一种情况是单独应用铁电体自发极化反转所表现出的开关特性。例如用于计算机的矩阵编址存储器、移位寄存储器、电荷转换器，以及利用某种铁电材料的自反转效应器件，如快速非破坏读出(NDR)存储器等。

第二种情况是应用极化转向和铁电体具有的其他特性复合工作的器件，例如铁电性-压电性复合工作器件、铁电-光复合工作器件以及温度自稳非线性介电元件 TANDEL(Temperature Auto-stabilishing Non-linear Dielectric Element)等。

第三种情况是把铁电材料和其他材料构成复合结构，例如铁电－电致发光器件、铁电－光电导器件和铁电－半导体器件等。下面我们仅以铁电－半导体器件为例，说明这类复合结构的工作原理。如果将铁电体与半导体直接相接界，就可能构成几种重要的场效应器件，其中半导体的基本特性会由于相邻的铁电体而改变。图 48 表示 N 型半导体由于接界的铁电体处于三种不同极化状态下($P=0$，$P=+P_r$ 和 $P=-P_r$)所发生的能带畸变，图 48(a)表示铁电体未极化时，半导体能带平直，表面的载流子密度与体内相等；图 48(b)表示铁电体具有反向剩余极化 $-P_r$ 时，半导体面附近的电子密度超过体内，因而其能带在界面处向下弯曲；图

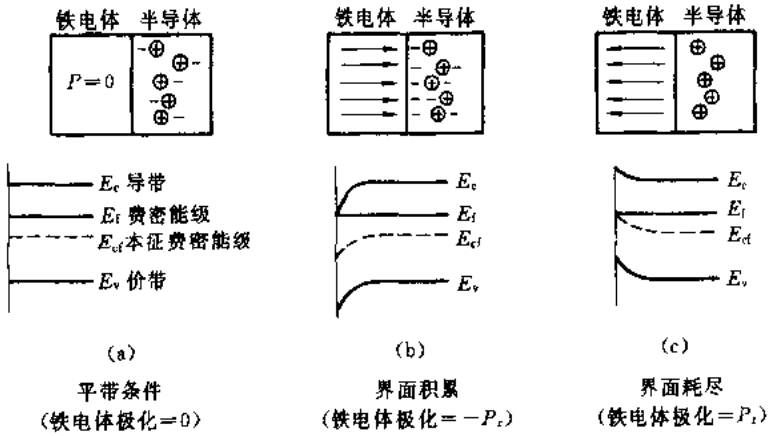


图 48 铁电-半导体界面能带畸变

48(c)的情况与此相反，铁电体具有正向剩余极化 $+P_r$，半导体界面附近的载流子减少，能带向上弯曲。

在早期，铁电-半导体器件是将 Te、CdS 或 CdSe 等半导体薄膜淀积在铁电基板上，如 $BaTiO_3$或 TGS。也有将 SnO 薄膜淀积在 PZT 铁电陶瓷基板上。最简单的器件结构如图 49 所示。其工作原理如下：在源和漏之间半导体通道的电导率直接受与半导体相接触的铁电体表面的极化状态的影响。例如，若半导体为 N 型，当铁电栅处于负极化状态时表面电子密度增大，使半导体电导率增大；相反，当铁电栅处于正极化状态，使半导体电导率减小。由于铁电体的电滞回线特征，源与漏之间的电流对栅电压也具有电磁回线型式。因为铁电体可以表现为部分开关特性，这就能将源漏之间的电导率设置在任意的中间数值。利用这种器件可以做成放大器，其增益可以通过电压加以控制在 1～1000。另外，这类器件具有很大的优点是用来做成非挥发性和模拟的存储器。图 49(b)所示的复合结构器件存在半导体薄膜中俘获电荷的不稳定性，使其电导率随存放时间而改变，改进的方法之一是不采用半导体薄膜而采用铁电薄膜。这种器件的结构如图(b)所示，这是在硅圆片上用溅射的方法淀积一层 3μm 的 $Bi_4Ti_3O_{12}$ 薄膜。这种 $Bi_4Ti_3O_{12}$-Si 晶体管十分稳定，在几个早期的测量中其开路状态的漏极电流没有变化。

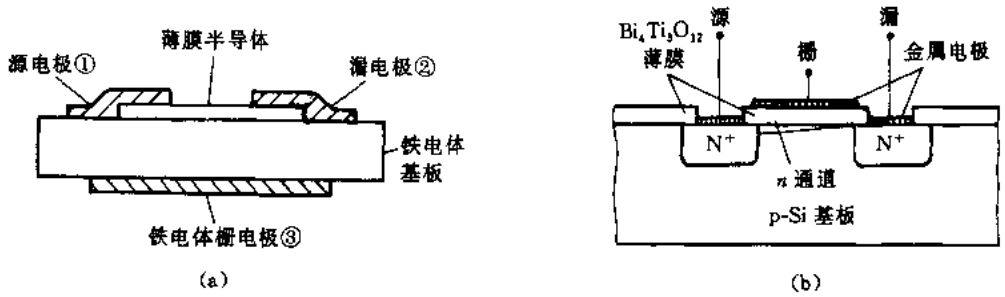


图 49 (a) 铁电体基-板半导体薄膜器件；(b) 铁电薄膜-半导体基板器件

8 驻电性[7,8]

8.1 驻电现象

驻电体是呈现准永久电荷的介电材料。这里"准永久"的含义是反映电荷衰减特性的时间常数比对驻电体作实验观察或作器件使用的期限要长得多。驻电体 electret 与永磁体 magnet 的英文词是有对偶性的,从这一点就可对驻电体的特性有一个粗略的概念,有些地方将 electret 译成驻极体,类似地驻电体是准永久带电的电介质,不同的是,驻电体既可以是带正负电荷形成双极性的,也可以只带正电荷或负电荷形成单极性。

虽然驻电体的性质早在 1732 年就由 Gray 作过初步的描述,并在 1892 年由 Heaviside 首先取名叫驻电体。但真正比较系统地研究驻电体的性质是由日本物理学家 Eguhi 从 1919 年开始的,当时所应用的材料基本上与 Gray 所用的相似,是棕榈腊和松香的混合物,采用热充电法制成。由于这种材料做成的驻电体寿命短、灵敏度低、机械强度差,没有大量实际应用,直到聚合物驻电体的出现,使对驻电体的研究和应用进入了一个崭新的时期。

驻电体的电荷分布一般可用图 50 来表示。图中在驻电体的一个表面上淀积有金属电极。电介质的表面可能驻有表面电荷,体内有空间电荷,或作定向排列的电偶极子,由于驻电体中电荷的吸引,金属电极上会出现异号的补偿电荷。只有表面电荷或体内电荷的叫空间电荷驻电体;只有偶极的叫偶极驻电体。

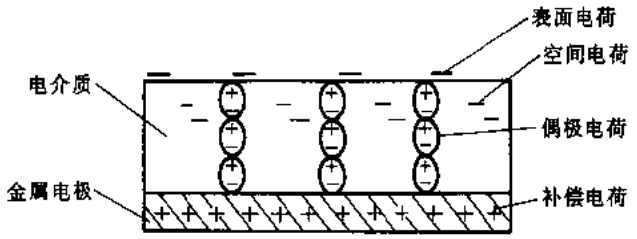


图 50 单面金属化的驻电体

8.2 驻电体形成方法

8.2.1 摩擦起电法

当两物体表面接触后分开或摩擦可以造成起电。在真空中进行摩擦起电的研究表明,当排除了大气的影响后,电介质摩擦时的起电是由于电子从电介质逸出或从与其接触的表面进入。影响这过程的参数是材料的功函数,对电介质材料来说,通常借用金属或半导体的接触理论。但是这种起电方法对制造驻电体没有实用意义,主要原因是缺乏重复性。

8.2.2 热充电法

将材料加热到某一温度后施加直流电场,并在直流电场作用下让材料冷却到室温。对腊来说一般加热到熔点,对聚合物通常加热到玻璃化温度以上,但低于其熔点。施加电压的电极可以紧贴在试样表面,也可保留一很小间隙。热充电过程,可能出现以下几种不同情况:①由于偶极子作定向排列,或体内电荷的分离,使试样表面出现的驻电荷符号与电极极性相反,这

种电荷叫异电荷。②通过相接触的电极使电荷注入介质内，这时介质表面的电荷符号与电极极性相同，叫同电荷。③如果在间隙内产生气体放电，使电荷淀积在介质表面，这时注入介质内的电荷符号也与邻近的电极极性相同，也是同电荷。

8.2.3 *等温电荷淀积法*

在常温下由于空气隙中的放电而产生电荷转移称为等温电荷淀积法。由于没有加温，不会出现因电介质极化而造成异电荷。这种电荷淀积法现已被普遍采用，因为方法简单且速度快，特别适用于聚合物薄膜的充电。

用得最广泛的是电晕充电法，采用不均匀电场在大气压下产生电晕放电。通常用针尖状或刀口形的上电极放在试样上某一距离的高度上，试样的另一面放在接地电极，如果上电极接负电位，这时负电荷（大部分是电子）飞向电介质表面，例如对聚四氟乙烯薄膜的充电就常采用这种电压偏置法。

8.2.4 *部分穿透的电子束和离子束*

电子透入的深度与电子束的能量有关。实验表明，能量在 0.5～1MeV 的电子束在大气压下轰击，其深度为 0.1cm 或更大。因此对于薄膜材料要用低能电子轰击，通常要求在真空下进行，为了使注入均匀，可以用电子束来进行扫描，加速电压在 5～50kV。

这种方法的优点是可以准确地控制电荷注入的深度、横向电荷的分布和电荷密度。因而，这是目前驻电体充电最完善的方法，广泛用于制作微音器用的驻电薄膜，具有很长的衰减时间。

8.3 驻电体材料及其应用

对能实用的驻电体材料的性能要求主要有电导率、迁移率、偶极弛豫频率以及压电和热电常数。

如果驻电体具有本征载流子，则产生本征电导率 σ，σ 的大小将决定材料体内驻电荷衰减的快慢。在最简单的情况下，如果驻电荷位于开路试样（例如没有淀积电极的电介质），通过材料体内而衰减，衰减时间常数可表示为 $\tau_1 = \varepsilon/\sigma$。而对于偶极驻极体的衰减时间常数可表示为 $\tau_2 = 1/\alpha$。其中 α 叫偶极弛豫频率。图 51 表示几种聚合物薄膜由热充电法形成的驻电体在室温下和干燥大气中的衰减特性。其中 PTFE 是原四氟乙烯；PC－KL 是聚碳酸脂；PP 是聚丙烯，PPPO 是聚 2,6－二苯基－1,4－苯撑氧；PET 是聚酯。其中 PTFE 的衰减时间最长。

偶极驻电体具有压电性和热电性，这可用图 52 来说明。当驻电体上两电极短路时，如果

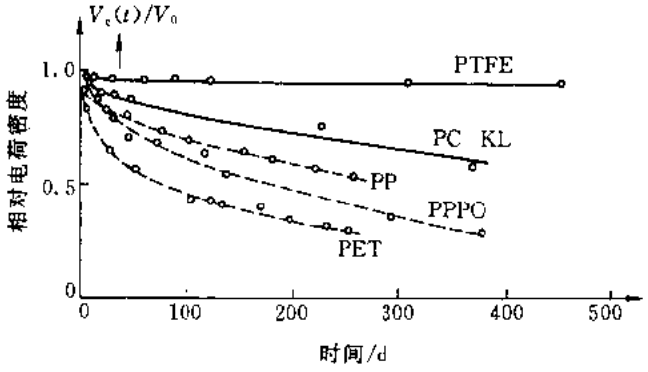


图 51 聚合物薄膜驻电体的电荷衰减

增加静压力或降低温度使驻电体收缩时，由于两电极距离减小，为了保持二电极的电势差为零，则有电荷通过电极间的连接线。因而驻电体不仅可利用其驻电性，同时也可利用其压电性和热电性。

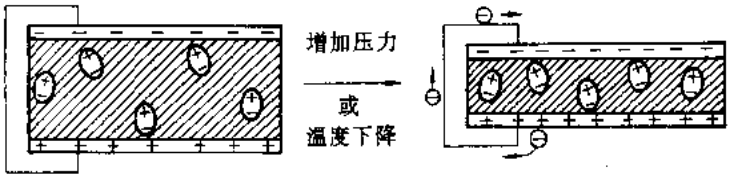


图 52　偶极驻电体的压电性和热电性

到目前为止，驻电体的应用已具有十分广泛的领域。例如可以用于高传真微音器、耳机和扬声器，许多种的换能器；应用驻电体作静电印刷，静电记录；也可利于辐射剂量探测，红外探测和成像等，还可用于制作驻电体马达和发电机，静电过滤器等。

聚合物薄膜驻电体制成的微音器，不仅具有高传真、高灵敏度、响应频带宽、体积小等优点，而且，即使没有电源也能工作，它相当于一架电信号发生器，其工作原理如图 53 所示。

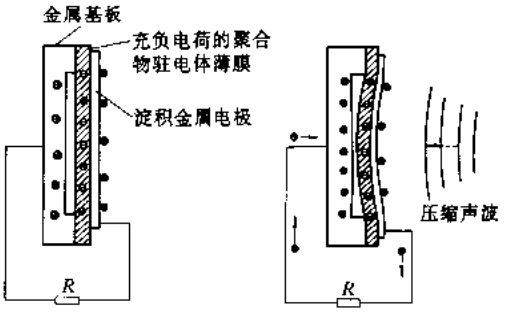


图 53　薄膜驻电体微音器工作原理

参 考 文 献

[1] A. K. von Hippel. Dielectric and Waves. New York: John Wiley & Sons, Inc., 1954
[2] 陈益新.电介质物理(讲义).上海:上海交通大学,1964
[3] 陈季丹,刘子玉.电介质物理学.北京:机械工业出版社,1980
[4] Shyh Wang. Solid State Electronics. Chapter 7. McGraw-Hill Inc., 1966
[5] J. C. Burfoot, G. W. Taylor. Polar Dielectrics and Their Applications. London and Basingstoke: Macmillan Press Ltd., 1979
[6] 许煜寰,等.铁电与压电材料.北京:科学出版社,1978
[7] G. M. Sessler. Electrets. New York: Springer-Verlag Berlin Heidelberg, 1980
[8] B. Hilczer. J. Malecki. Electrets Elsevier, Warszawa, 1986

集成光路及其应用

像集成电路一样，理想的集成光路是将各种光波导器件都制作在同一种材料的衬底上，这叫做单片集成光路(Monolithic integrated optical circuit)。光源、探测器和光放大器这类器件需要用半导体材料作衬底。但迄今所用的半导体材料(如 Si、Ga As 和 InP 等)的电光和声光特性远不如铁电晶体(如 $LiNbO_3$ 等)电介质材料，因而需要用不同材料作衬底制成性能最佳的器件或组件，然后将它们以一定的耦合方式结合成完整的系统，这叫做混合集成光路(Hybrid integrated optical circuit)。单片集成光路不仅指半导体衬底的集成光路，也包括电介质衬底的集成光路(如 $LiNbO_3$ 衬底的光开关阵列等)。另外，在许多集成光路系统中也包括了光纤波导元件在内，如许多光纤传感器是由平面波导元件与光纤元件组合而成。

1 光纤系统

1.1 光纤通信系统

光纤通信自 1970 年开始发展以来，已在邮电通信、广播电视、闭路电视、电子计算机数据通信、科学研究、工业、交通和国防等许多领域获得广泛应用。

光纤通信系统的层次有干线(trunk)、市内网、地区网(local area network)和用户环(subscriber loop)。干线用于城市或地区间的长途干线和大城市内较短距离的中继线；地区网把各种计算机和智能终端连接起来，实现工厂、企业和机关等的自动化，以提高工作效率和经济效益；用户环除了传统的电话和数据业务外，还增加电视电话、电视会议、闭路电视、远距离服务(如医疗、教育)以及高传真立体声广播等，以达到所谓综合服务的目标。不同的光纤通信系统由于传输的距离和容量不同，使用的光纤和光器件也各不相同。图 1[1]表示不同应用场合下

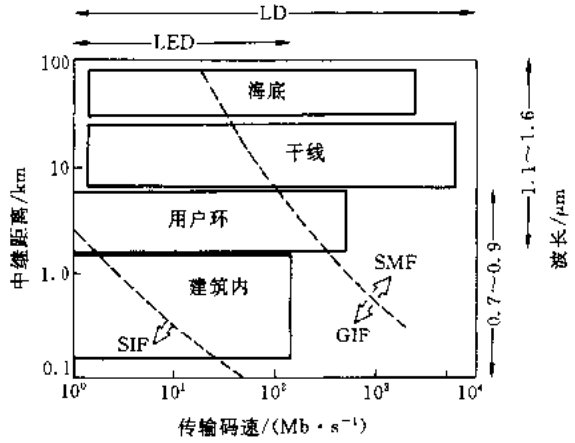


图 1 光纤传输系统的应用范围

SMF—单模光纤，GIF—渐变折射率光纤，SIF—阶跃折射率光纤

LD—激光二极管，LED—发光二极管

光纤传输系统所选用的光纤类型、光源、波长、传输码速和中继距离间的关系。

就现有水平来说，0.8～0.9μm 短波长光纤通信系统在技术上已相当成熟，并进入商业应用阶段，光纤、光源、探测器和各种无源器件的性能可以满足实用化的基本要求，可靠性有明显提高。由于生产数量的增加，价格开始下降，所以经济上也是合理的，只要传输码速超过 32Mb/s，则比同轴电缆系统更经济。随着光纤和光器件的大批生产，光纤系统的费用还会有较大幅度的下降。这种系统可用于市内电话网和上述新发展的用户环路，采用多路复用技术则使系统更为经济。

1.3～1.6μm 的长波长光纤通信系统有多模和单模两种。对以石英材料为基础的光纤来说，短波长多模光纤的传输损耗一般为 2.5～3dB/km，带宽为 200～800MHz·km，长波长多模光纤(1.3μm)的损耗为 0.6～1dB/km，带宽大于 1GHz·km。近年来，由于长波长光纤和光器件性能不断改进，长波长光纤系统已进入实用化现场试验阶段，许多国家开始用于通信干线上，如日本的 F－6M 系统[1]和美国的 FT3C 系统[2]，有趋势表明，今后长波长光纤系统将在更大程度上取代短波长系统。

单模光纤不存在模色散，可以大大增加传输带宽，单模光纤系统的发展比预计的要快。目前，长波长单模光纤的传输损耗为 0.2～0.5dB/km，带宽大于 3GHz·km。美国贝尔实验室在 Atlanta 试验了 432Mb/s 和 144Mb/s，全长 240km，分为 8 个中继站的单模光纤通信系统[3]。采用的光纤是内包层折射率下降的单模光纤，在 1.31μm 时传输损耗为 0.46dB/km。采取波分复用，波长为 1.275μm 和 1.335μm，串音为 40dB。应用 432Mb/s 时，最大中继距离为 64.9km，用 144Mb/s 时为 70.1km，最小色散的波长为 1.30μm。引起普遍注意的是码速为 1Gb/s、波长为 1.55μm、长度为 84km 的单模光纤传输试验[4]。该系统采用 C^3 型(Cleaved-Coupled-Cavity)稳模激光器[5]，它由两个长度分别为 135μm 和 125μm，中间间隔为 5μm 的 FP(Febry Perot)腔激光器集成在一起，其中一只作为激光器，另一只作为调制器。

单模光纤通信能用于越洋海底光缆系统。美国“TAT-8”海底光缆计划是代替原计划将要敷设的大西洋海底电缆系统，全长 1 万余 km，采用 274Mb/s 的传输码速，利用时分多路复用将通信容量扩大到原来的 3 倍，即 12000 路电话。为筹建这条越洋海底光缆通信线路，贝尔实验室于 1982 年初成功地进行了一次 101km 无中继的传输试验[6]。试验结果证明，中继距离主要是受损耗而不是色散的限制。而后，又在新英格兰海岸大西洋海底 5.5km 深处完成了第一次带有中继器的海底光缆通信系统的现场试验[7]，试验表明，光缆在深海水温和压力下承受了 150kN/m 的张力，但对它的传输性能没有妨碍。

1.2 军用光纤系统

由于光纤通信具有抗电磁干扰能力强、保密性好、体积小、质量轻、功耗低和安装简单等一系列独特的优点，十分适宜于军事通信上的战略和战术应用。因此，实际上光纤通信一出现就受到了国防方面的重视，特别是美国三军联合战术通信局在 70 年代初就把光纤通信作为发展军用通信的基本方针。从 1974 年起，美国三军就根据各自军种的特点，分别建立了庞大的研究机构，其军用光纤通信系统的应用列于表 1。MX 计划是军用光纤通信系统的最大计划，MX 是一种可移动的陆军武器系统，配有 200 枚核导弹，由地下铁道往返运送到各发射点上，导弹掩体共 4600 个，整个系统需用光缆 1 万 km。还有一种光纤制导系统，在导弹头部安装一架摄像机，从导弹尾部引出光缆将目标范围内的视频信号传输给发射点，以命中发射目标。在这情

况中导弹可从流动发射架、直升飞机或地下发射出去。

表 1　美国军用光纤通信系统举例

军　种	安　装　地　点	系　统　性　能
海　军	USS Kitty Hawk 航空母舰	可视系统
海　军	USS Little Rock 巡洋舰	声学系统
海　军(ALOFT)	A-7 战斗机	13 根光纤，115 通道代替扭绞对试验飞行
空　军(TIFSF-DADS)	C-131(改型)	0.5Mb/s，16 通道
陆　军	便携式战场系统	18.7Mb/s，8km；2.3Mb/s，64km
海　军	Meade 要塞	20 Mb/s，18 通道，2km，卫星接收数据地面传输
陆　军	MX 导弹	正在开发中
陆　军(ADOCS)	UH-60A Black Hawk 直升飞机	计划飞行试验于 1984 年进行
空　军(AN/GRC-206)	前方控制室与遥控雷达发射机之间的地面连接	2 根光纤，200Mb/s，正在开发中

在美国，近年来军事应用占了光纤市场的 40%，而邮电通信约占 30%，其他各种用途共约占 30%。但是，在 80 年代拟计划发展的长距离邮电通信系统，可能会使光纤市场上军用和商用的比例发生变化。

1.3　医用光纤系统

光纤技术在医学上最早的一种应用是光纤内窥镜，它能使医生从内部看到咽喉、支气管、心脏、结肠和其他器官。这种内窥镜由一相干的光纤束构成，即光纤在横截面上的位置沿着器件的长度保持不变，因而可以传输光学图像。通常，用另一光束与内窥镜合在一起用来提供照明，这传光光束中的光纤的排列不必是相干的。有些内窥镜也包括用来切除组织标本的外科手术器械。

应用比较高的激光功率的光纤波导可以进行显微外科手术，例如切除肿瘤或损坏血管的结块等。通常，这种情况下采用 0.5μm 波长的氩激光器，或用 1.06μm 的 Nd:YAG 激光器。采用光纤波导可以使外科手术精确地控制激光光斑的位置。在进行激光外科手术时也常常采用内窥镜，这样就可清楚地看到激光的手术操作。

光纤可作为传感器测量某些医疗参数，例如，器官的局部温度、血液的流速及化学成分等。虽然这方面的研究尚不很深入，但已有某些器件在实验室中得到应用[8]。一旦技术有了发展，很可能将有更多的光纤传感器和遥测仪在医学上应用，也有可能与集成光路结合起来用作信号处理。

1.4　其他应用

在低损耗光纤出现以前，早在 60 年代就开始应用普通光导纤维作为传像、导光和折像。其中一类刚性光纤器件即光纤面板和微通道板可构成电子光学器件，如增强器和夜视器。

2 半导体单片集成光路器件

半导体单片集成光路的最终目标是把光源、波导、探测、调制、放大等光器件甚至电子器件集成在单一材料的衬底上，实现特定的功能，以满足光纤通信、光信号处理等应用的需要。现阶段，由于材料和工艺方面的限制，还不能完全达到上面的目标。但是适用于单片集成的各种光波导元件以及少量元件的集成已取得相当进展，其中有些集成器件具有一定实用价值。这些第一代的半导体集成光路包括：由一些相同光器件的集成、不同光器件的集成以及光子器件和电子器件的集成。

2.1 相同光器件的单片集成

多路复用的光纤系统需要多种波长的集成光源。Aiki 等[9,10]采用两步液相外延法，把 6 个不同波长的 DFB 激光器做在 5mm 方形 GaAs 衬底上，工作波长间隔为 2nm。激光器具有分离限制异质结(SCH)[11]结构。由化学刻蚀方法在表面造成三级光栅，应用的掩模是由全息光刻产生的。激光通过直接传输而耦合到未掺杂的 $Ga_{0.9}Al_{0.1}As$ 波导层，如图 2 所示。激光器和波导的横向尺寸的限制是由台面刻蚀到 GaAs 衬底而形成 20μm 宽和 30μm 厚的条，激光器的间距为 300μm。为了使它们一起进入汇合耦合器，波导做成弯曲形，最小半径为 4mm，如图 3 所示。耦合器的输出通过一单通道波导与光纤作端接耦合。

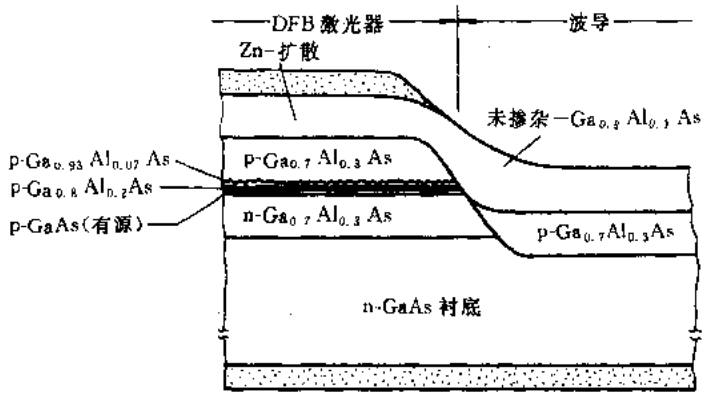


图 2 DFB 激光器由直接传输而耦合进 GaAlAs 波导

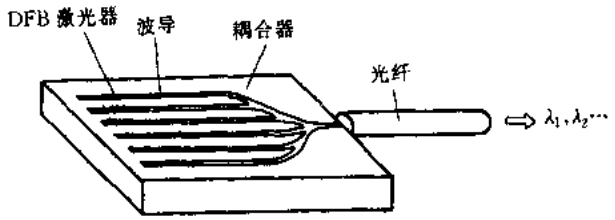


图 3 波长复用光源的输出

激光器在外施重复速率为 1kC 的 100ns 脉冲电流下工作，微分量子效率测量结果为 7%，波导的损耗系数约 $5cm^{-1}$。室温下激光器的阈值电流密度为 3～6kA/cm^2。各激光器波长间隔

测量值为(2±0.5)nm,偏差的原因是用液相外延生长异质结时造成芯片上的不均匀性造成的。激光束谱宽的典型值是0.05nm。分别调制各激光器不会有困难,同时在发射输出端测得的总微分量子效率约为30%。因此,这种集成光源经过改进是可以实用的。

最近,在InP衬底上制成了长波长单片集成DFB激光二极管[12]。集成的两个DFB激光器波长为1.3μm,可以在40℃以下连续工作,实验证明它们的波长差可以精确控制在1nm以内。InP衬底上的波纹用全息光刻后经化学刻蚀制成,全息干涉系统由计算机控制有可能精确形成不同的光栅周期,误差在0.05nm以内。集成的GaInAsP/InPDFB激光器采用常规隐埋异质结以实现对横模的控制。为了避免晶体生长过程中表面波纹的损坏,把外延生长的温度控制在600℃以下。

这类集成光源除了多路复用以外,也可简单地作为大功率光源使用,贝尔实验室曾将18个激光器依次排在宽约0.27mm的GaAs芯片上;也可将集成的激光器作为备用光源,当一个激光器失效以后,另一个激光器可以立即自动替换上去,从而大大提高了可靠性。此外,这种集成光源除了做成横向排列的以外,也可以做成垂直排列结构[13]。

与多种波长光源的功用相似,多路复用光纤系统也需要集成探测器。集成的器件与相同数量以分立元件装配的器件相比较,具有体积小并易于与光纤相耦合,而且在同一单片上制备的元件之间性能比较均匀。图4为一个由10个InGaAsP/InP光电二级管单片集成在一个单元的结构图[14]。10个光电二极管在衬底上排成一行,单管所占面积为$(400\times400)\mu m^2$,光敏面直径为150μm。用液相外延工艺在掺Sn的InP晶片上连续生长InP缓冲层(15μm)、掺Cd的InGaAsP吸收层(2.4μm)和掺Zn的InP窗口层(15μm)。四元系的禁带宽度为0.92eV,对应的波长为1.35μm。各个单管的击穿电压和响应时间的偏差很小,分别为±1%和±6%。

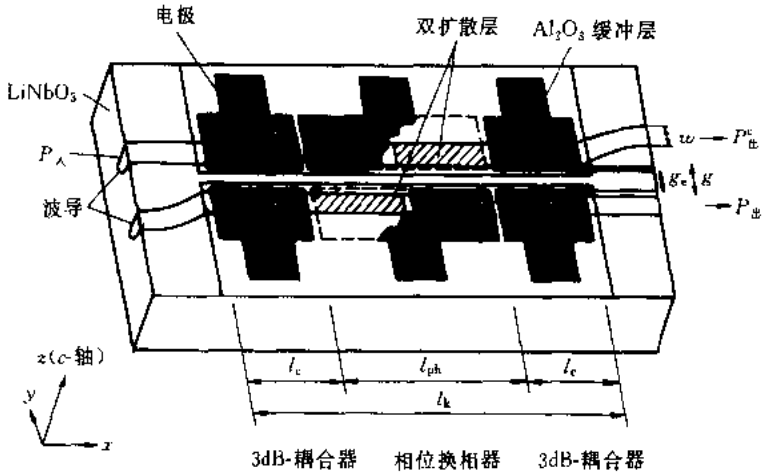


图4 InGaAsP/InP集成光电二极管

2.2 不同光器件的单片集成

现阶段这类集成包括:光源、波导、调制器和接收器之间多种形式的集成,目的是为了改善原有器件的特性和获得新的功能。例如,新近引人注目的一种复合腔结构可以改善激光器输出特性和实现新的激射功能;采用共腔双区双异质结结构制成的激光器,可用来进行选模和实现单纵模激光输出;也可用来产生ps超短脉冲以及实现光学双稳态效应获得光放大功能等。

一种典型的复合腔结构如图 5 所示[15]，这是具有两个腔长分别为 l_1 和 l_2 的 GaInAsP/InP。

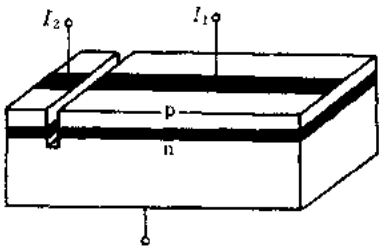


图 5　耦合腔集成激光器

DH 激光器的集成器件[15,16]可以获得纵模光谱的控制与有效的调制。其中短腔（l_2）起有源标准具（active etalon）作用，改变其偏流可以对激射光谱精确控制。当短腔上的偏流增大时，对不需要模的抑止性能也随之加强。这种情况下，中间模可以抑止到 28dB。当 $l_1 \approx 300\mu m$，$l_1/l_2 \geq 10$ 时，这种激光器可以单模工作，实验结果如图 6 所示。这种器件可以应用于单波长光纤通信系统。

迄今已提出了许多方法来加工这种耦合腔器件的中间反射面。例如，可以采用择优湿法化学刻蚀技术[17~19]；也可采用干法反应离子刻蚀（RIE）[20]代替常规解理技术获得反射境面。以前曾有人提出通过氧化物窗口进行选择外延的方法[21]。新近又发展了一种微解离（microcleavage）[22,23]技术，可以在不解理衬底的情况下在晶片的任意局部位置上制备出微解理面。其方法是利用常规光刻工艺，通过对 GaAs 衬底的定向择优刻蚀与组分选择刻蚀得到异质结构的悬臂或桥，然后由超声机械振动使臂或桥的截面解理，图 7 表示微解理前形成的悬臂；这种方法适用于在各种集成光路中制备出任意腔长的激光器。

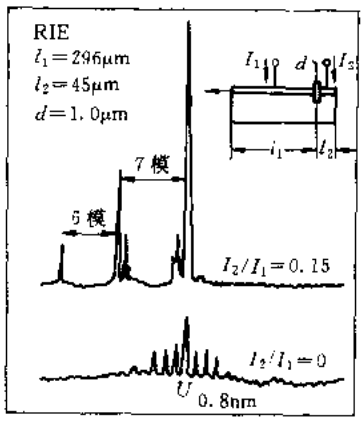


图 6　双腔激光器在不同偏流时的光谱特性

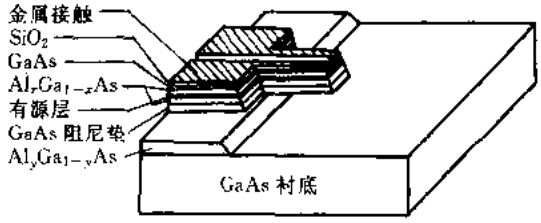


图 7　对异质结构的悬臂采用微解离技术（微解理前的双异质结悬臂）

这种耦合腔结构也可用于激光器与探测器、激光器与调制器的集成[13]。在激光器－探测器集成器件中，激光器可从前后两个面发射激光，前向光可直接耦合进入光纤传输，后向光耦合进探测器转换成电信号反馈到使激光器稳定工作的电路中，例如可以补偿由温度变化而引起的光功率漂移。图 8 表示一种用微解理技术制备的 AlGaAs/GaAs 双异质结激光器与结型光电二极管的集成器件[23]。由于激光器腔可以做得很短，所以阈值电流低和纵模稳定，而且获得了高精度的监控特性。

另一个光晶体管与其他光器件的集成例子也受到相当重视。InGaAsP/InP 异质结光晶体管（HPT）在长波长光纤通信系统中作为探测器与雪崩二极（APD）管相比，其突出的优点是具有高光增益，而不需要像 APD 那样高的偏压，也没有由于雪崩效应造成的剩余噪声。而 HPT 的不足是响应时间较长和灵敏度不够高，这是目前正努力设法解决的问题。但是 HPT 具有高光增益和大集电极电流可用来发展新的集成器件。例如，光放大器[25,26]、图像转换器[27]和波长转换器[28]都是由一个 HPT 与一个双异质结激光器或 LED 构成的单片集成光路。在这些器件中 LD 和 LED 需要由 HPT 输出的大电流驱动。通过控制对 HPT 的掺杂分布，特别是基区的掺杂水平，可以获得170mA 的集电流输出[24]，这时入射波长为1.15μm，输入光功率为155μW，HPT 所用的偏压仅2V，HPT 的光增益相当于1180。同时也证实这种集成器件不仅具有光放大功能，而且还能作为一种开关器件或

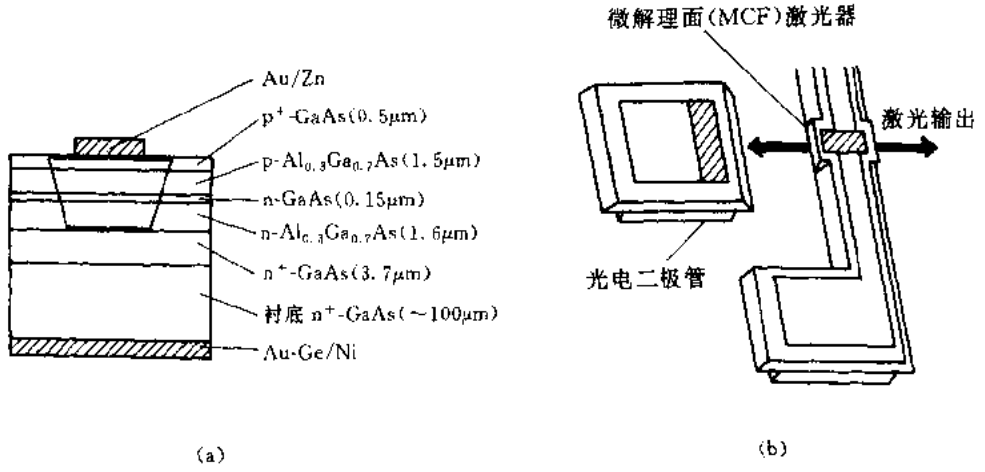


图 8　用微解理技术制备的集成器件

(a) 微解理激光器的截面图；(b) 激光器 - 光电二极管集成器件

光双稳器件，这是通过对单片集成器件(六层结构)中电或光的耦合，即在 HPT 和 LD 或 LED 之间的正反馈的控制来实现的。图9中给出了单片集成六层结构的截面及其能带分布。实验结果表明，这种单片集成器件的光放大增益在1.2μm 波长下为1.3。最近，通过调节掺杂浓度和在 HPT 和 LED 之间加进二层吸收层以改进反馈的抑制，使器件的光放大增益提高到3.3，在 μW 功率范围光双稳态下增益为7.7。可以预计，器件的性能可以进一步改善。

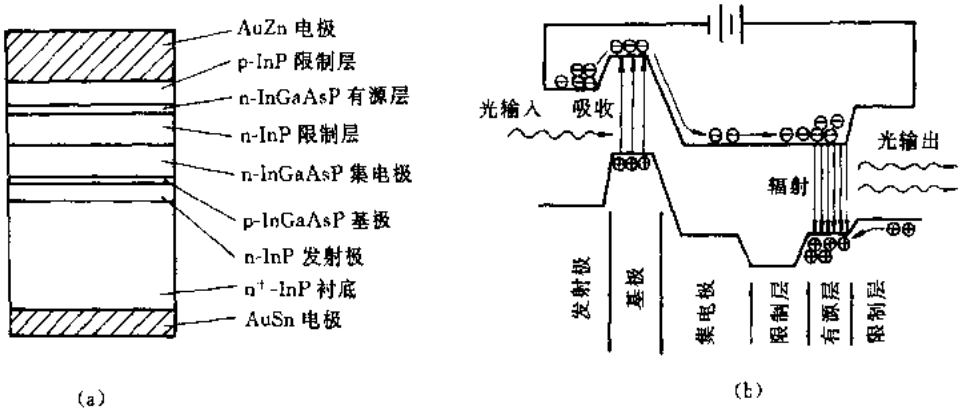


图 9　HPT-LED 单片集成器件的六层结构

(a) 截面图；(b) 能带图

2.3　光电子单片集成

将光器件与电子器件做在同一衬底上，这是半导体单片集成的发展趋向，称为光电子集成回路，(简称 IOEC 或 OEIC)。这种器件的优点除了尺寸小、坚固和可靠性高等优点外，并由于寄生电抗的减小，可使光电子回路的速度和噪声得到有效的改善。

光电子集成首先遇到的问题是采用导电的还是半绝缘衬底。在 GaAs 衬底上制备的注入型激光器是在衬底面上生长不同成分 $Ga_{1-x}Al_xAs$ 的外延层，通常采用的是重掺杂的高电导n-型衬底。但另一方面，Gunn 二极管和 MESFET 等电子器件在半绝缘衬底上生成的 n-型层上制备的，电子器件的电流横向流过 n-型层。半绝缘衬底对于其上面所制备的器件之间有良好的

隔离作用，因而在实现光电子集成以前，首先解决在半绝缘衬底上制备激光器的问题。遇到的困难是，注入型激光器具有高电流密度，在很小体积内产生大量的热。通常采用的散热方法是将激光器倒装在热沉上，这种办法在光电子集成中显然是不适用的，因为这里有许多电的连接，所以必须通过衬底散热。这就限制了激光器的功率消耗，要选用低阈值电流的激光器设计。第二个困难是常规的激光器用解理面作为腔的反射，因而芯片的尺寸受到限制。前面已经提到有几种非解理衬底来形成反射面的方法。例如，刻蚀镜面、生长镜面、离子铣镜面、分布布拉格反射器以及微解理镜面等。

最早在半绝缘衬底上做成的激光器如图 10 所示，它利用常规的液相外延方法在半绝缘衬底上生长五层的结构。它的激射是由于载流子的群集效应，故称为群集效应（crowding-effect）激光器[29]。通过 p-n 结的电流密度从边缘随横向距离而下降，这是由于开始二层的薄层电阻产生的电压降所致。这样就在台面边缘附近形成一狭条的有效增益区（近场的半宽一般为 5μm）。这种群集激光器的结构有些类似于常规的双异质结激光器，p-n 结平行于外延层，其载流子和光的横向限制由电流的群集来提供。但是这种器件有两个缺点：在空气与半导体界面附近载流子的表面复合损耗；由于刻蚀表面的光散射损耗。这就影响了激光器的性能，一般它具有相当高的阈值电流（> 100mA）和低的外微分量子效率。

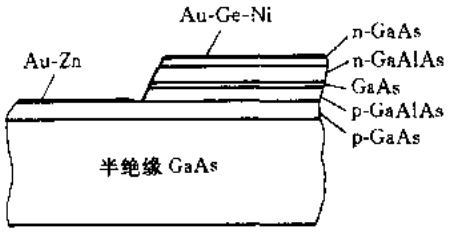


图 10　群集效应激光器

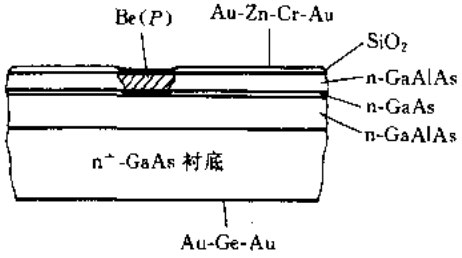


图 11　Be 注入的几何条形激光器

改进上述激光器的方法是把激射有源区从台面边缘移至内部，以消除表面载流子复合和光散射的影响。电流注入区的形成是通过对 n-型双异质结注入受主离子（例如 Be），改进后的激光器如图 11 所示[30]。Be 注入的能量为 100keV，剂量为 $3\times10^{15}/cm^2$，而后在 800℃下退火 40min。结果使注入的离子（4μm 条宽）扩散到 GaAs 有源区，在大多数情况下，p-n结处于有源区内。在正向偏置时，电流通过 p-n 结注入有源区。这种激光器阈值电流一般为 40mA，腔长 250μm，近场图为单横模，电流与光功率特性呈线性，直到输出光功率为 10mW 时不出现扭曲。

横结条形（TJS）激光器[31,32]如图 12 所示，其结构与前两者不同。电流经 p-n 结从横向注入有源层，p-型区是对 n-型双异质结的一部分扩 Zn 而形成的，电流横向通过结从 p 区流到 n

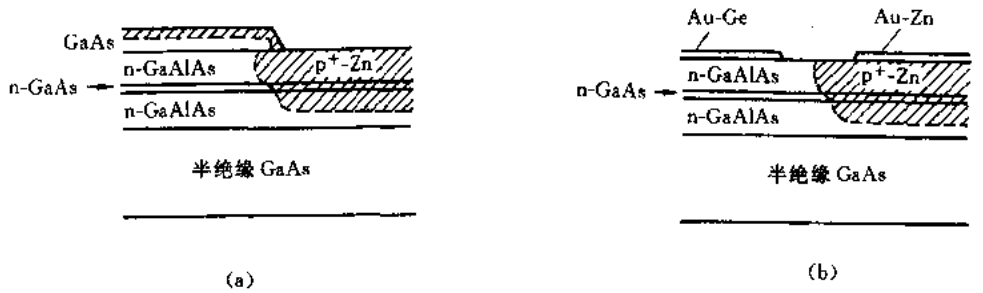


图 12　在半绝缘衬底上的横结条形激光器

（a）第一次扩散后激光器截面图；（b）最后结构的截面图

区。因为 GaAlAs 的带隙比 GaAs 的宽，载流子主要通过在 GaAs 有源层内的 p-n 结。这是由于有源区电导率较限制层高，有源层的载流子浓度高达 $2\times10^{18}cm^{-3}$。这种横结条形激光器具有良好的光模稳定性和较低的阈值电流(15～50mA)。

隐埋异质结激光器的结构较为复杂，如图 13 所示，它由两步液相外延工艺制成。第一步是在半绝缘衬底上生长形成四层常规的双异质结。然后在〈110〉方向刻蚀成窄台面(1～2μm 宽)，并在整个片子上再生长 p-GaAlAs 和 n-GaAlAs 二层隐埋层。器件在扩 Zn 后第二次刻蚀台面和金属化。这样制备的隐埋异质结激光器的阈值电流，当腔长为 300μm 时，每 μm 的条宽上为 8mA，微分量子效率为 55%[33]。

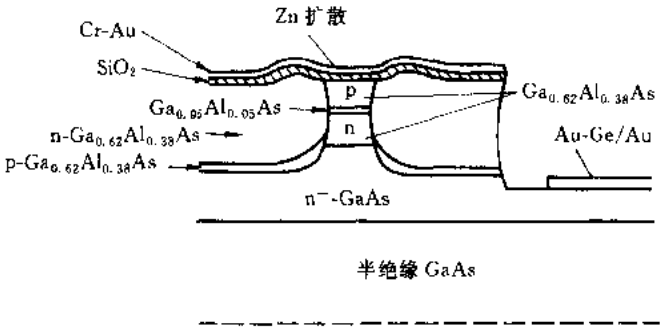


图 13　半绝缘衬底上的隐埋双异质结激光器截面图

在半绝缘的 GaAs 衬底上最早实现的光电子集成是把一个群集效应激光器与一个 Gunn 振荡器做在一起[34]。Gunn 器件的电流通过激光二极管进行调制，得到的输出光频率达 1GHz，调制深度为 70%。

在光纤通信系统中使信号加载于光束最方便的方法是对通过激光器的电流进行直接调制。GaAs MESFET 是理想的激光器的驱动器，因为它具有亚纳秒的开关速度。图 14 表示在半绝缘衬底上把 MESFET 与离子注入的激光器集成在一起的结构[30]。晶体管由半绝缘衬底上生长的 n-型 GaAs 层构成。电流串联通过激光器和晶体管并通过栅压进行调制。最近，又利用隐埋激光器与 MESFET 相类似的集成[35]，晶体管是凹形结构，栅极长度为 1.5μm，其典型的特性是：激光器阈值电流为 20～30mA，晶体管的夹断电压为 3～5V，电导为 5～15Ω^{-1}。激光器的频率响应取决于强度涨落噪声谱，在 2 倍阈值电流下达到 4.5GHz。

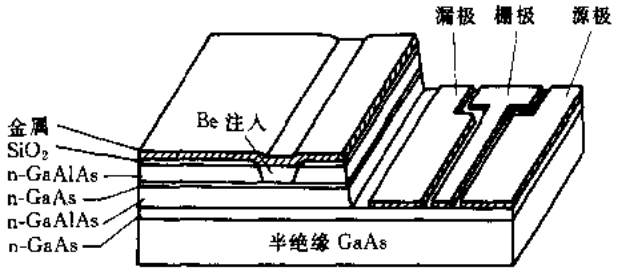


图 14　Be 注入激光器及 MESFET 光电子集成器件

GaAs MESFET 也可用来作为超高速光探测器(OPFET)。利用这种光探测器，电子放大器和注入激光器可以构成集成中继器[36]，其结构和电路如图 15 所示，它包含三个 MESFET 和一

个群集效应激光器。晶体管 Q_1、Q_2 和 Q_3 分别当作有源负载、光探测器和激光器的驱动器。通过光探测器 Q_2 的电流也必须通过其负载晶体管 Q_1。通过调节 Q_1 上的源和栅压，可以控制 S_1 与 D_1 之间的有效电阻。光探测器上出现的信号直接加到驱动晶体管 Q_3 的栅极上。实验中，激光器在没有光信号时，由外加电流源偏置在刚超过的阈值。中继器的放大部分在660MHz下有20dB的功率增益，整个器件的总增益为为10dB。

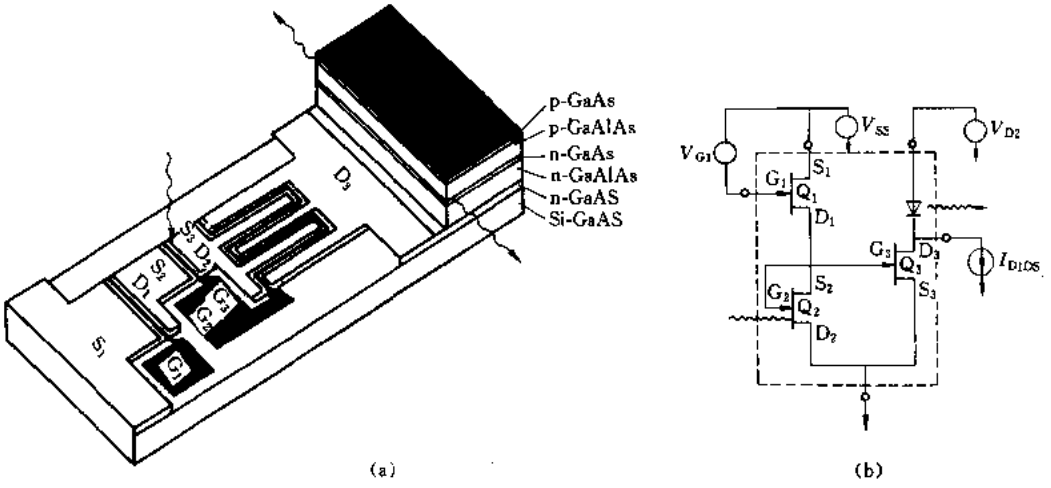


图 15　单片集成光中继器

(a) 结构图；(b) 电路原理图

最近又有关于激光器和驱动电路单片集成的报道[37]如图 16 所示。该集成芯片上包含了一个激光二极管，以及由四个场效应晶体管和一个电阻构成的驱动电路；另外还集成了一个光电二极管与二个场效应晶体管和一个电阻，它用作激光二极管发射光功率的自动控制(APC)，芯片尺寸为 0.6mm × 1.0mm。

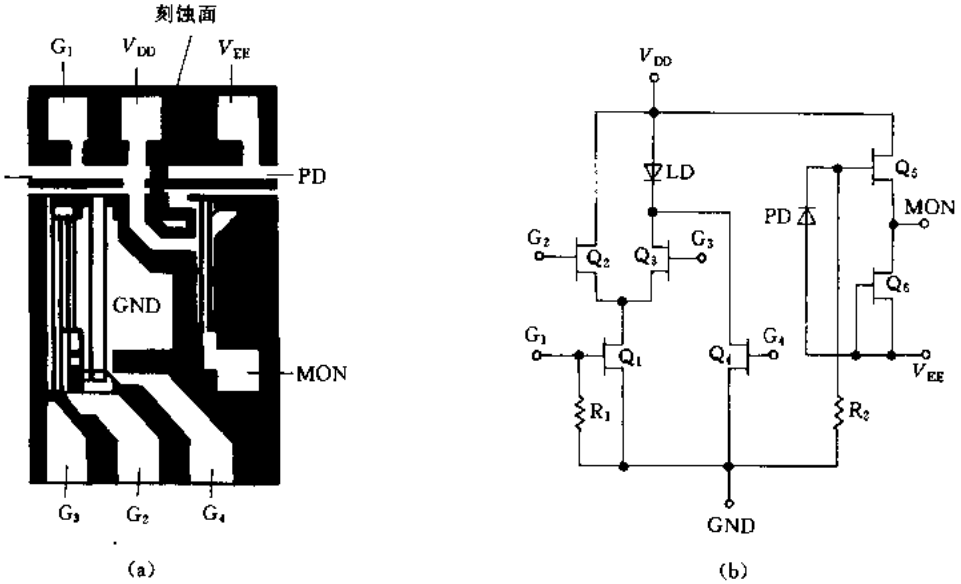


图 16　带驱动电路的激光二极管单片集成芯片

(a) 照相图；(b) 电路图

激光二极管是在具有台阶的半绝缘衬底上，用液相外延工艺生长 GaAs 的 n^+ 导电层和 GaAs/GaAlAs 异质结。一个镜面是用化学刻蚀方法制得的。在室温下脉冲驱动的阈值电流低至 75mA。

FET 是用离子注入方法做在半绝缘衬底表面上。该表面是由化学刻蚀除去生长在其面上的外延层而得到的。全部 FET 的栅极长为 2μm，$FETQ_4$ 的栅极宽为 720μm，其余 FET 的栅极宽都是 240μm。$FETQ_4$ 的漏-源电流是 144mA，其他的是 50mA，对应的夹断电压为 -3.5V。

3 开关网络

在光纤通信和光信号处理系统中需要应用小型和高速的开关网络，例如用作多路复用的连接；计算机网中光纤连接的母线等。一种 4×4 的集成光开关网络最早由 Taylor 提出[38]，后来 Schmidt 等在 1976 年用五个方向耦合器构成 4×4 光开关网络[39]，如图 17 所示。

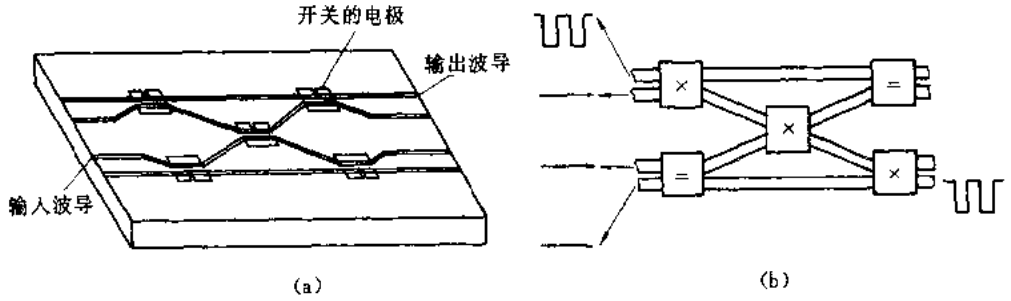


图 17 光开关网络及其工作状态

(a) 4×4 光开关网络；(b) 光信号由输入 1 接到输出 4 通道时的各开关工作状态

这种光开关网络的输入和输出端有四条通道波导，它们通过五个交换 $\Delta\beta$ 位相失配的方向耦合器开关使通道中光信号进行转换。这种开关的工作原理在前面已有详细讨论。简单地说，两组电极放在耦合波导的上面通过电光效应产生波导中的相速失配。开关具有直通和交叉两种工作状态，控制位相失配就能决定开关的工作状态。改变五个开关的状态就可以得到从输入端任意一通道中的光束进入输出端某一通道的转换方式。图 17(b)中表示了光信号从输入端通道 1 接到输出端通道 4 时各开关的工作状态。这种器件是在 $LiNbO_3$ 衬底上制备的，输出端通道间的串音为 18dB。

计算机之间的数据传输应用光连接有许多优点，例如：可以避免电传输时由于接线的电容和电阻引起的时延；不存在终端阻抗匹配的问题；可以消除由于接地回路引起噪声等。Becker 和 Chang 曾提出一种可用作计算机数据交换母线(communication bus)的光开关网络[40]，如图 18 所示。

这一集成光母线的工作原理如下：一组激光二极管分别由各微处理机的输出电压所驱动，由计算机 M_i 产生的数据流可相应地在源通道 S_i 中成为光脉冲流，通过不同的 S_iR_j 开关将光束通至接收通道 R_j，最后由光电探测器转换成电信号。所用交叉通道开关如图 18(b)所示，当操作电压为 10V 时，通态和断态相差 21dB。器件是用 $LiNbO_3$ 衬底经扩钛而成，钛层的条宽为 40μm，形成的波导是多模的，交叉波导的夹角小于 10°。

这一光开关网络具有速度高，因而传输码速可以在 Gb/s 以上。开关的通态和断态可相差

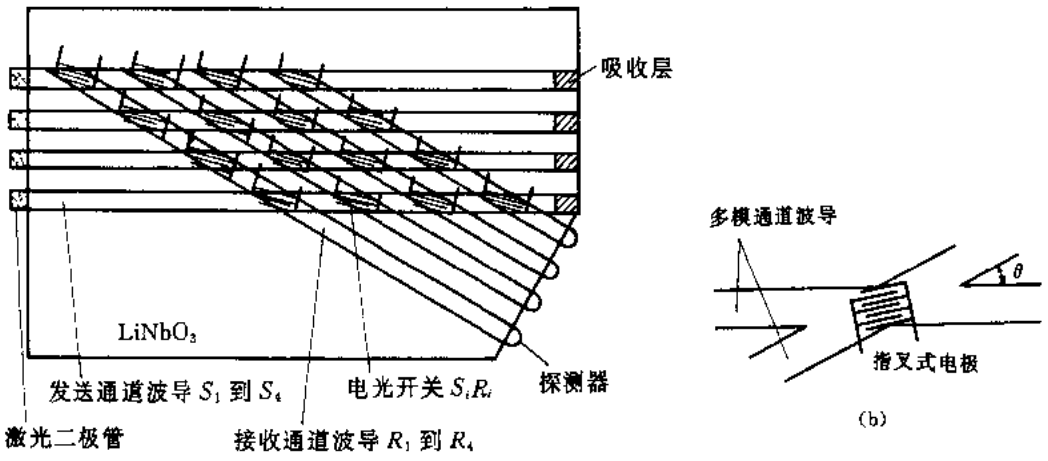


图 18　计算机数据交换母线的光开关网络

(a) 集成光通道波导交换母线；(b) 交叉通道开关

26dB,因而具有很小的误码率。而且通态的开关效率为 1% ~ 10%,因而光功率足以分配到所有的接收通道,这在广播式工作情况下是必要的。

交叉通道光开关除了采用叉指电极(IDE)形式外,Tsai 等[41]还提出了一种全内反射 TIR 电极,利用它可以做成开关网络,图 19 为这种开关构成的开关网络[42]。

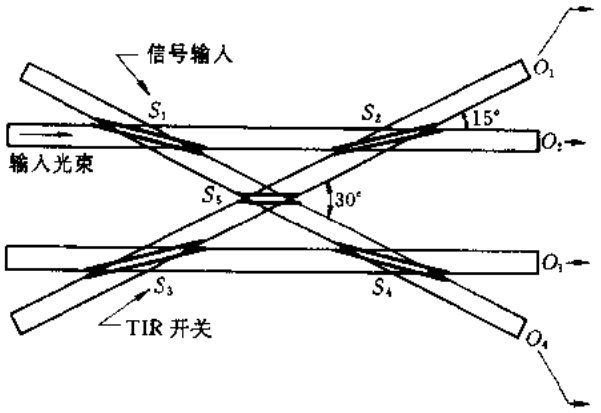


图 19　TIR 4×4 开关网络

这开关的原理是,在二通道波导的交叠范围内淀积一对靠得很近的平行电极,电极上没有电压时,波导内的光束直通传播,例如从端 1 至端 4。当电极上施加适当电压,由于线性电光效应,电极间波导的折射率减小,因而光在沿电极的折射率交界面上产生全内反射,光束不再直通,而射向交叉通道。为了获得低驱动电压和低的串音,在波导交叉区使折射率增加 1 倍,即为 $2\Delta n$。这种开关特性的典型例子:驱动电压 5V 时,开关效率为 93%,串音为 -15.7dB,波导的交叉角为 1.0°。用这类开关做成的 4×4 光开关网络包括五个 TIR 开关,做在 $LiNbO_3$ 衬底上总长仅 0.75cm,把相同的 TIR 开关串联起来,串音可减小 1 倍,即从 -15 ~ -30dB。

这类光开关的另一特点是电极间电容非常小,光的渡越时间很短,可以做成宽带器件。例如,通道波导宽为 10μm,夹角为 2.0°,电极长 0.6mm,相应的带宽达 33GHz。另外,器件的长度很短,而且通道波导没有转折,其插入损耗小,并且集成的密度比其他光开关器高。

最近，Neyer 等[43]报道了另一种单模 X 开关，其结构与上述 TIR 开关十分相似。其工作原理是，利用电光效应改变折射率，由外施电压控制波导交叉区中基模和一级横模之间的位相差。类似于分叉光波导有源(Bifurcation Optique Active, BOA)开关。因为在波导交叉区，不仅折射率增量大，而且波导宽，所以这里不仅存在基横模，还出现一级横模。因此，耦合到两个输出波导的光功率取决于交叉区两个模之间的位相差。位相差与通道波导的参数(如横向折射率的分布函数，宽度 W 和有效折射率增量 Δn)和夹角 θ 有关。如图 20 所示，耦合到输出端 3 和端 4 的相对光功率，根据其工作原理可作近似计算，考虑到交叉的长度为

$$L = 2W/\sin(\theta/2) \tag{1}$$

则 P_3 和 P_4 的相对值为

$$\frac{P_3}{P_1}, \frac{P_4}{P_1} \approx \sin^2\left[\frac{2W\Delta\beta}{\sin(\theta/2)} + \alpha_{3,4}\right] \tag{2}$$

式中：$\Delta\beta$ 表示两种模的传播常数之差；α 是相移常数。这一近似解与光束传播法(BPM)的数值计算十分接近。

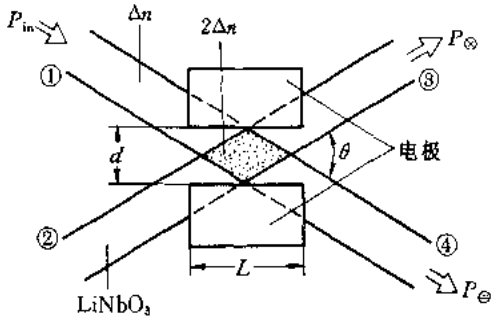


图 20　X 开关示意图

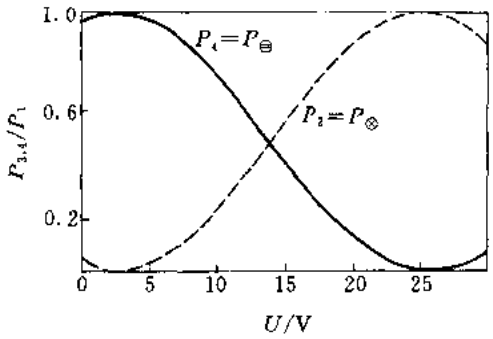


图 21　X 开关的电压特性

图 21 表示具有夹角 $\theta = 0.6°$ 的 X 开关的电压特性。电极间隔距离 $d = 1\mu m$，电极长度 $L = 1mm$，$P_\ominus$ 和 $P_\otimes$ 分别表示直通和转换状态时的光功率。在电压为 2.5V 时，直通功率达输入功率的 98%，而转换功率仅 0.1%，电压增加 22.5V 时，两种状态互相替换。2% 的光功率损耗可认为是由辐射所引起的。

这类开关具有较低的串音，而且能高密度集成，如果用它做上述计算机光连接母线，一个 10×10 的开关网络可以做在不超过 2cm × 2cm 的 $LiNbO_3$ 衬底上。

4　集成光学频谱分析器

Hamilton 等[44]提出的集成光学频谱分析器(IOSA)可以说是首先完成的由多种光波导元器件构成的混合光集成器件。这种频谱分析器的功能是使飞机驾驶员获得所接收到雷达信号的频谱，从而决定他的飞机是否已被地面站或空对空导弹等设施所跟踪。显然，如果要迅速采取有效的行动，这种信息的实时显示是十分必要的。当然，需要将可能遇到的全部敌方雷达信号频谱标本存贮在计算机内以作比较。不仅如此，据有关方面预测，今后在电子对抗中，IOSA 将成为新一代电子侦察接收机的主要形式。它与其他接收机相比较，具有以下主要优点：①能在高、中频带内(例如 2～4GHz)瞬时工作和实时显示；②信号截获率可高达 100%；③在宽带内具有同时识别各个信号的能力；④利用多波束技术，可把测向和测频在一台接收机中同时

完成；⑤处理方式简单，设备体积小、质量轻、成本低。

集成光学射频频谱分析器的工作原理示于图 22。从一激光源发出的光束耦合进入一平面波导。其中首先通过一波导透镜使准直。准直的光束通过布拉格声光调制器发生衍射，用来作频谱分析的射频信号加在声换能器的叉指电极上，所产生的表面声波因信号的频率不同而具有不同的空间周期。因此，调制器输出端光束的衍射角是射频信号频率的函数。第二个透镜是用来将光束聚焦到一个光电探测器列，如果射频信号中存在 1 个以上的频率分量时，光束将分成相应的分量会聚在探测器列中不同的单元上，每一探测器单元代表一特定的频道。从这里可以看到，集成光学频谱分析器与相应的电子设备相比，其特点是只需少量几个光学元件来实现这复杂的功能，而在电子设备中将应用成千上万个电子元件。

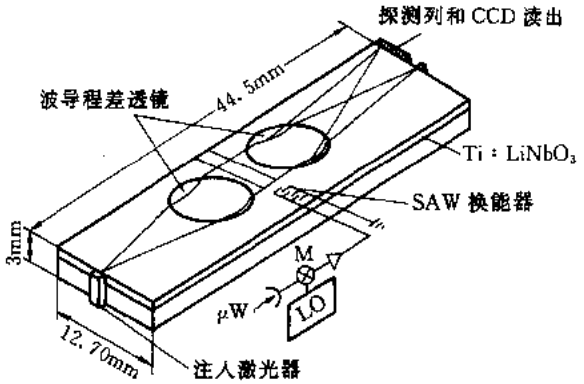


图 22　集成光学频谱分析器

IOSA 结构参数主要包括：光束宽 D；第二个透镜的焦距 F；探测器列中各单元的间隔（即中心距）S；射频分辨率 Δf 等。这些参数之间都互相有关，其最基本的关系是产生布拉格衍射必须满足布拉格条件，即

$$\delta = 2\arcsin\frac{\lambda_0}{2\Lambda n_{\text{eff}}} \tag{3}$$

式中：δ 为衍射光束与入射光束之间的夹角；λ_0 为光束在真空中的波长；n_{eff}为光波导的有效折射率；Λ 为表面声波的波长。

当 δ 很小时，上式可近似表示为

$$\delta \approx \frac{\lambda_0}{\Lambda n_{\text{eff}}} = \frac{\lambda_0 f}{n_{\text{eff}} V_s} \tag{4}$$

式中：f 为表面声波的频率；V_s 为表面声波的速度。

在参数 S、D 和 F 之间存在两个限制关系：一是要满足对射频分辨率 Δf 所提出的要求[45]，即

$$S \leqslant \frac{F\lambda_0}{n_{\text{eff}} V_s}\Delta f \tag{5}$$

另一是要满足光学分辨率的要求，有

$$S \geqslant \frac{g\lambda_0 F}{n_{\text{eff}} D}$$

式中 g 为与聚焦光斑的定义有关的系数$\left(\text{即}\frac{1}{e}\text{点、}\frac{1}{e^2}\text{点或全部斑}\right)$。

对 $LiNbO_3$ 材料来说，当 $n_{eff}=2.2$，$V_s\approx 3.5km/s$ 时，可以作出两条曲线（见图 23），其关系式为

$$\frac{Sn_{eff}}{F\lambda_0}=\frac{1}{V_s}\Delta f,\quad \frac{Sn_{eff}}{F\lambda_0}=\frac{g}{D}$$

根据上述这些关系就可确定器件的各种参数。

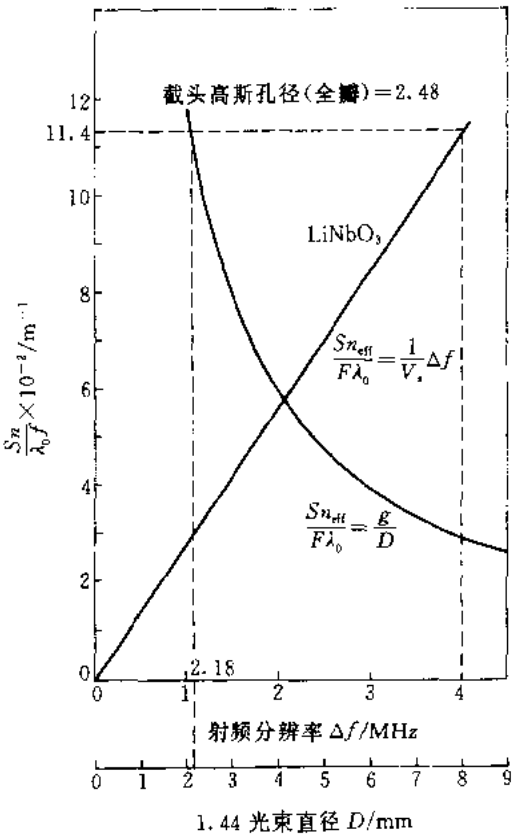


图 23　IOSA 的光路设计曲线

在设计和制造集成光学频谱分析器时，需要对各种元器件的材料和工艺作出适当的选择。

（1）激光器。目前较理想的是 GaAlAs 双异质结半导体激光器。它具有稳定的工作波长，其发射区的厚度与光波导接近，可获得较高的耦合效率。$LiNbO_3$ 光波导的光损伤阈值与波长有关，长波长下的光损伤阈值较高，因此可选用长波长激光器，但必须要有相应的探测器。

（2）波导衬底。曾用于试验的主要有两种：Si 衬底和 $LiNbO_3$ 衬底。采用 Si 衬底，可以把光电探测器集成在一起，但由于 Si 不具有压电性，必须在其上面淀积一层压电材料供产生表面声波（如 ZnO），通过这层压电材料可借助 SAW 换能器传递到光波导。但由于 Si 中 SAW 的衰减比 $LiNbO_3$ 中高 10 倍，因而产生布拉格衍射的效率很低。$LiNbO_3$ 是目前比较理想的衬底，具有高压电系数，又容易制作光波导，所以应用得较多，但无法用它制成激光器和探测器。

（3）透镜。在 Si 衬底上生成 SiO_2，然后淀积一层玻璃作为波导，这时可用 Ta_2O_5 等高折射率材料制成 Luneburg 透镜，但其轮廓不易控制。用 $LiNbO_3$ 作衬底时，由于其折射率较高，很难找到合适的材料做成 Luneburg 透镜，多数情况下是做成 Geodesic 透镜。采用单点金刚石车床可以加工出轮廓十分精确的透镜，其插入损耗低于 2dB。但这类透镜难以大批生产，因而成本较高。

（4）表面声波换能器。表面声波换能器基本结构是在压电介质上淀积叉指型金属电极，为了获得 GHz 以上的响应频带，曾设计了许多种不同的电极形式。

（5）探测器列。为提高器件的分辨率和动态范围，要求探测器单元的尺寸小，相邻单元间的串音低，信号可测的动态范围大，能与光路兼容，并在转换成电信号后能迅速读出，一般要求探测器列中每单元的读取时间必须在 μs 数量级以下，以充分体现出声光处理实时性的优点。目前采用的探测器有光电二极管列和电荷耦合器（CCD）列两种。

表 2 是美国 Westinghouse 高级技术实验室研制的 IOSA 参数和特性[46]。最初是用波长为 632.8nm 的 He－Ne 激光源作试验，得到的带宽为 400MHz，分辨率为 5.3MHz。后来改用波长为 0.83μm 的 GaAlAs 激光二极管作光源，使分辨率提高到 4MHz。

美国 Hughes 航空公司的实验室也有类似的研究，其器件的参数和性能列于表 3[45]。该室用工作波长为 0.82μm 的 GaAlAs 激光器与波导作端接耦合，探测器列采用硅电荷耦合器件，3dB 带宽为 380MHz，在最大射频功率 500mW 时的衍射效率为 5%。

表 2　Westinghouse 频谱分析器的参数和特性

衬　底	材　料	x-切割 $LiNbO_3$
	尺　寸	70mm × 25mm × 3mm
波　导	工　艺	扩钛，28nm
透　镜	型　式	Geodesic
	公　差	0.5μm
	插入损耗	2dB
	焦　距	24.5mm 及 27mm
SAW 换能器	中心频率	600MHz
	带　宽	400MHz
	衍射效率	5%（60mW 电动率时）
	动态范围	20dB
	型　式	两对叉指电极倾斜配置
探测器	型　式	自扫描光电二极管列
	单元数	140
	间　距	12μm
	时钟速率	5MHz
	动态范围	> 40dB
	相邻单元串音	> 15dB
	偏转灵敏度	2.2μm/MHz

表 3　Hughes 频谱分析器的参数和特性

衬　底	材　料	y-切割 $LiNbO_3$
	尺　寸	45mm × 12.7mm × 2.5mm
波　导	工　艺	Ti 内扩散
激光器	型　式	GaAlAs、隐埋、双异质结
	波　长	0.82μm
	谱　宽	< 0.1nm
	光斑尺寸	1μm（在 $1/e$ 点上）
	阈值电流	20mA
透　镜	型　式	Geodesic
	焦　距	18.8mm
	孔　径	7.4mm
	F　数	2.54
SAW 换能器	带　宽	750 ~ 1250MHz
	驱动电功率	500mW（最大值）
	A-O 偏转效率	5%（最大）
	插入损耗	12dB
探测器列及 CCD 读出	衬　底	n-Si
	型　式	埋沟式
	单元尺寸	8.0μm × 150μm
	读出方式	串 – 并
	读出速率	333kHz

自从 1980 年第一批 IOSA 问世后，又有许多新的研究不断报道，其中有一部分研究是设法改进 IOSA 的性能，基本原理和结构没有根本变化。例如，1983 年在日本东京召开的国际第四次集成光学和光纤通信会议上，展出的东芝频谱分析器[47]提高了射频分辨率，其 3dB 光斑尺寸相应的分辨频带为 2.7MHz，200 个通道的光电二极管探测器列接收信号的总带宽为 400MHz，每通道的频带为 2.0MHz。这一器件结构上的特点是准直透镜的焦距减小为 6.55mm，而傅里叶变换透镜的焦距增大到 52.7mm，光电二极管列中通道间隔为 11mm。

Westinghouse 频谱分析器的改进主要是增加动态范围[48]。以前器件的动态范围一般不超过 30dB，新结构中的透镜采用较短的焦距和减小曲率，并且改进波导抗光损伤的能力。实验结果表明，器件的动态范围可以超过 40dB。

针对程差透镜加工困难，可设法采用平面工艺来制造波导透镜。其中较成功的一例是，用变周期光栅反射镜装置在波导的两侧，构成所谓外透镜结构[49]。利用这种形式，并对 SAW 换能器加以改进，已制成带宽为 1GHz 的 IOSA，其结构如图 24 所示[50]。由于光在波导中的传播是来回反射的，所以波导尺寸可以做得很小，这一器件的 $LiNbO_3$ 衬底面积只有 $(12\times15)mm^2$。两个 SAW 换能器的中心频率分别为 500MHz 和 GHz，分辨率为 4MHz。

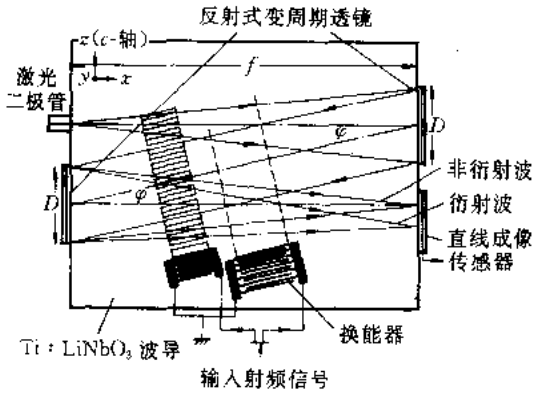


图 24　具有变周期光栅透镜的 IOSA

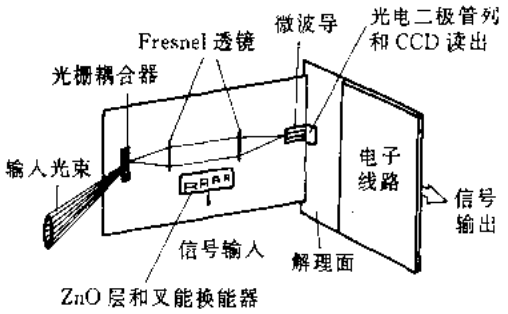


图 25　利用 Fresnel 透镜的 IOSA

上述外透镜虽能用平面工艺制作，但将它装配到波导上仍会有很大困难。较理想的将透镜直接利用平面工艺做在波导上。法国 Valette 等以硅为衬底，第一次成功地在其面上做成了 Fresnel 透镜[51]。这种 IOSA 的结构如图 25 所示。它是在硅衬底上生长一层 SiO_2，波导层是 Si_3N_4，波导层上面再淀积 ZnO，用来形成表面声波换能器。Fresnel 透镜是模拟的[52]，它具有良好的光学特性，具有接近衍射极限的光斑（直到 $F/3$），大的光场角（大于 6°），低的背景噪声（在离轴 1°处测量小于 -32dB），良好的重复性（dF/F 在 $10^{-2}\sim10^{-3}$，F 是焦距），以及可以采用平面工艺。为了避免在焦平面处精确地把硅片解理，在输入端做一个光栅耦合器，在第二个透镜的焦平面处制成一排 4μm 宽的通道波导。光栅耦合器、透镜和接收通道波导用同一块掩模做在 SiO_2 层内。表面声波由四个叉指换能器产生，带宽约为 250MHz。ZnO 薄膜淀积于 Si_3N_4 上，其质量的好坏可用光学参数来表示，其损耗低于 2～5dB/cm。从 SAW 换能器的发生区到光波导区的转移损耗，对一级声模来说低于 1dB，但对二级模则非常大（超过 20dB）。这种器件具有很大的潜在优点，不仅许多光学器件，如反射镜、分束器、扩束器等都可集成在一起，而且探测器及有关的电子电路也可以集成在同一衬底上。

另一种新颖的“无透镜”的频谱分析器，其结构如图 26 所示[53]，它的工作原理是通过电光

效应实现。由图可见，激光束从激光二极管耦合到平面波导，金属电极列上的电压使波导的折射率改变，引起光束的衍射。加在电极上的电压即为所要分析的信号。光束通过变周期电极后的焦距取决于信号的频率。这一实验虽是初步的，但它可以用最少的光学元件实现频谱分析的功能，特别是不要用分离的两个透镜，不过它至今还未达到 IOSA 实用化的目标。

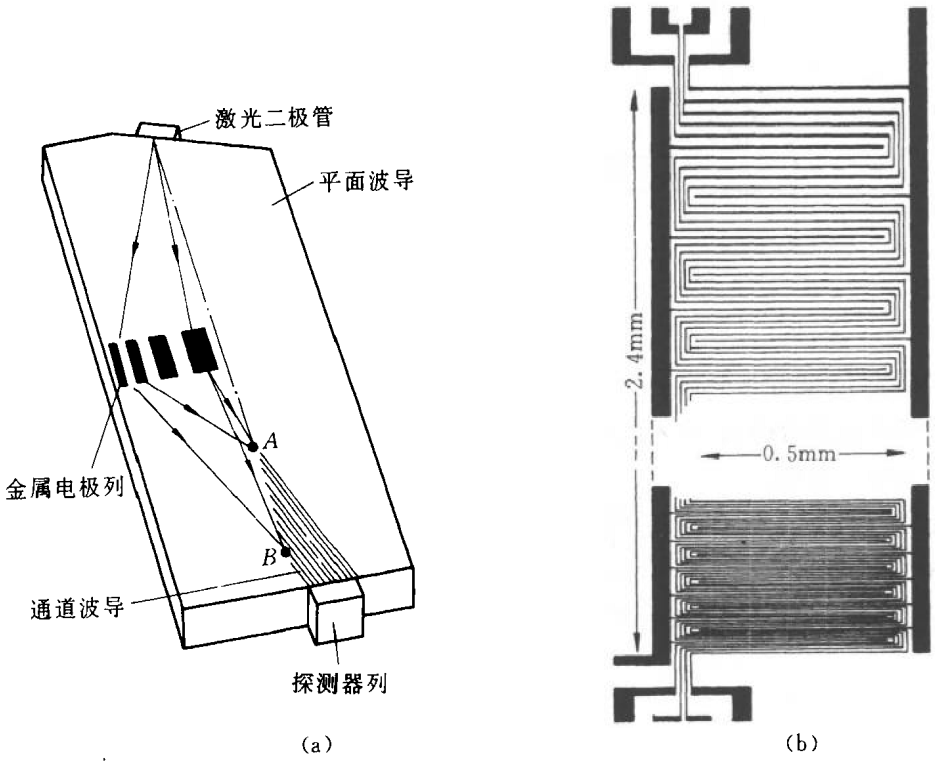


图 26 “无透镜”频谱分析器

(a) 工作原理；(b) 电极结构

集成光学频谱分析器引起如此大的兴趣，不仅由于军事上的应用，在光通信工程中也可用于多路复用及滤波等目的，在光纤传感器中也可用于干涉光路和外差接收等场合。

5 集成光学卷积器和光相关器

在信号处理中，卷积(convolution)和相关(correlation)运算都是十分重要的。完成这类运算通常用电子电路，也可用表面声波。但是，与上述频谱分析一样，应用集成光学的方法作信号的卷积和相关具有许多优点，集成光学器件除了体积小、结构坚固外(如用光导波和表面声波相互作用做成的卷积器)，还可得到相当大的动态范围和时间带宽积，这在雷达信号处理和单模光纤通信系统中都十分有用。

图 27 为一种应用声光效应的集成光学卷积器[54]。激光束通过输入棱镜耦合器进入光波导，与第一个表面声波相遇产生衍射。偏转的光束有两个特点：一是其频率由原来的 ω_0 增加到 $\omega_0+\omega_{ac}$(ω_{ac}是表面声波的频率)；二是假如声波是一个脉冲，那么光束也只是在这脉冲周期间发生偏转，即偏转的振幅与表面声波的包络成正比。如果再引进第二个表面声波，它只偏转已被第一个表面声波偏转的光束，对原来的光束不起作用。因为不满足布拉格条件，这样经过

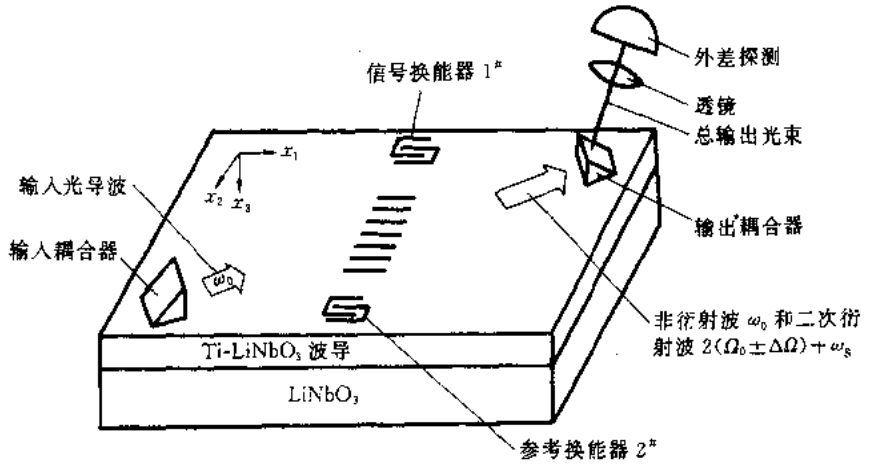


图 27　集成声光卷积器

两次偏折的光束，其光强与第一个表面声波包络和第二个表面声波包络的乘积成正比。如果将光束中的光能积分的话(用探测器接收)，它就对两个包络乘积积分，所得到的是卷积，即两个无线电信号的卷积。

设光导波的振幅为 A，则光导模的横电场为 $AE_m(x_3)$，其中 E_m 是归一化的。如果入射光束宽为 b，总的时间平均功率为 bAA^*。假设在一个光波波长内变化很慢，当满足布拉格条件时，被第一个表面声波偏转的光导波为

$$AE_3(x_3)F_1\left[t-\left(\frac{x_2}{v}\right)\right] \tag{6}$$

式中：v 是表面声波速度；F_1^2 是衍射效率，$F_1^2=\sin^2 g_1$。其中 g_1 正比于声脉冲波包的振幅 B_1，$g_1=\alpha B_1$。

从第二个换能器发出的表面声波在 $-x_2$ 方向传播，将与第一次衍射的光束再次位相匹配，第二次被衍射的光导波可表示为

$$AE_m(x_3)F_1\left(t-\frac{x_2}{v}\right)F_2\left(t-\tau+\frac{x_2}{v}+\frac{L}{v}\right) \tag{7}$$

式中

$$F_2^2=\sin^2 g_2=\sin^2\alpha B_2$$

B_2 是第二个表面声波波包的振幅；τ 是两个表面声波之间的时延；L 是第二个换能器沿 x_2 方向的位置。

从这一卷积器的几何关系上可看出，两次衍射光束与原始的入射光束相平行。由于这相互作用是声光效应，两次衍射光束的角频率具 $2(\Omega_0\pm\Delta\Omega)$ 数量的漂移。其中：Ω_0 是表面声波的中心频率；$\Delta\Omega$ 是表面声波信号的带宽。如果两次衍射光束及入射光束的其余部分允许被一个带宽大于 $2(\Omega_0\pm\Delta\Omega)$ 的探测器所接收，就成为光外差接收，其在频率范围 $2(\Omega_0\pm\Delta\Omega)$ 的输出电流为

$$i=\frac{\eta\eta' qG}{h\nu}\int_{-\infty}^{\infty}\int_{a-h/2}^{a+h/2}A^2E_m^2(x_3)F_1\left(t-\frac{x_2}{v}\right)\cdot F_2\left(t-\tau+\frac{x_2}{v}+\frac{L}{v}\right)\mathrm{d}x_2\mathrm{d}x_3 \tag{8}$$

式中：a 是从换能器到光束中心的距离；η 和 G 分别为探测器的内量子效率和内增益；η' 是波

导传输、输出耦合和探测光路上的总效率。对坐标 x_3 积分结果为

$$i = \frac{\eta q G}{h\nu} \cdot \eta' A^2 \int_{a-b/2}^{a+b/2} F_1\left(t - \frac{x_2}{v}\right) \cdot F_2\left(t - \tau + \frac{x_2}{v} + \frac{L}{v}\right) \mathrm{d}x_2 \tag{9}$$

以 B_1 和 B_2 取代 F_1 和 F_2，并作近似 $\sin g_2 \approx g_2$，最后再改变一个变数，可得所希望的关系

$$i = \frac{\eta q}{h\nu} \frac{\eta' P_{\mathrm{inc}}}{b} a^2 G v \int_{t-a/v-b/2v}^{t-a/v+b/2v} B_1(t') B_2\left(2t - \tau t' + \frac{L}{v}\right) \mathrm{d}t' \tag{10}$$

式中 P_{inc} 是在 $x_1 = 0$ 处波导内的光功率。式(10)表明如果束宽相当大，使 $b/2v$ 大大超过 $t - a/v$，则探测器的输出电流 i 与 B_1 和 B_2 的卷积积分成正比，在时间上压缩 1/2。式(10)清楚地表明，最大信号长度和表面声波时延都受束宽 b 与声速 v 的商的限制；动态范围只受输入光功率 P_{inc} 和光探测器的动态范围以及 F_1 和 F_2 成线性关系的表面声波振幅范围的限制；带宽将受换能器、光探测器和卷积过程的带宽的限制。这种卷积器初步获得的实验结果表明，最大时间带宽积在 1000 的量级或更大，最大的动态范围能到 83dB[54]。

集成光学卷积器与计算机做的卷积相比，其突出的优点是速度快，并且是实时处理；与单用表面声波做的卷积相比是时间带宽积大，亦即处理的精度高。当导波光束宽 b 为 1cm，声波带宽为 300MHz，则所得时间带宽积为 1000。单用 SAW 要达到这数值是十分困难的，如果光束宽为 3cm，则时间带宽积约为 3000，用 SAW 几乎是办不到的。

和卷积运算相似，也可用集成光学做信号的相关运算。通常有两种方法：一与上述卷积器大致相仿，不同的是两个表面声波不是相反方向传播，而是相同方向传播，因而二次产生的衍射也是同一方向的；另是将其中一个信号对光源进行调制，这种叫时间积分相关。一种新设计的集成声光时间积分相关器如图 28[55]所示。这一器件的工作原理是应用非各向同性的声光布拉格衍射。以前采用各向同性布拉格衍射的相关器[56]，需要应用一个准直透镜、两个成像透镜和一个空间滤波器做在同一衬底上，并且必须精确地对准。这种利用光导波的非各向同性布拉格衍射，由于采用一个薄膜起偏器把衍射光与非衍射光相分离，因此可以完全无必要再用成像透镜和空间滤波器。这样，相关器不仅体积更小和有更好的性能，而且更容易加工成集成光路。

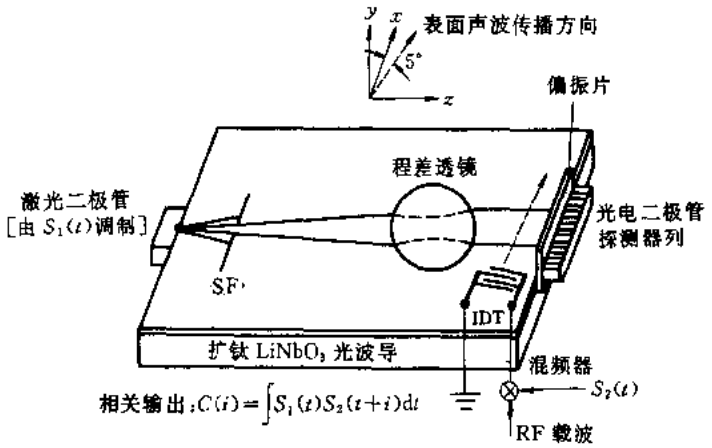


图 28　时间积分相关器集成光路

图 28 所示的混合集成声光相关器是制备在 y 切割的 $LiNbO_3$ 衬底上，其尺寸很小，为 15.4mm × 12mm × 2mm。在这器件中，激光源和带起偏器的光电探测器列分别与 $LiNbO_3$ 波导

的输入和输出端面作端接耦合，单个程差透镜用来使光束准直，焦距为 8mm，表面声波的传播方向与 x 轴成 5°。由于各向异性的声光布拉格衍射，衍射光的偏振与非衍射光的偏振相正交，所以薄膜起偏器可以阻挡非衍射光进入电荷耦合光电二极管列。由于各向异性声光布拉格衍射具有大的声光布拉格带宽，因而对入射光束的准直要求并不严格，这意味着对程差透镜的要求不需达到衍射极限，而且对准要求也不必十分准确。

工作时两个信号 $S_1(t)$和 $S_2(t)$之间的相关处理是，通过对激光二极管的调制和对表面声波换能器的激励来实现的。振幅调制的表面声波传播通过光学孔径，使入射光束衍射。结果，衍射光束在孔径截面上的近场光强分布与信号 $S_2(t)$完全相同，只是现在光强分布在光学孔径内以表面声波的速度传播。把偏振器和探测器列放置得十分靠近衍射光束的近场，就可使加载信号 $S_2(t)$的光强分布达到探测器列时不致引起畸变。最后，探测器列检测并积分衍射光强，所得的相关信号贮在电荷耦合探测器列的空域内，而后时钟取出并显示。由于这一器件不用成像透镜和空间滤波器，其尺寸比各向同性衍射的相关器减小 80%。换句话说，对同样大小的衬底，这种新器件的时间窗口将增大到 5 倍，因为准直透镜的焦距和用孔径的能增大到 5 倍。从实验得出的时间带宽积为 3.5×10^5。

6 集成光学模数转换器

模数转换器(ADC)的功能是将模拟信号转换成数字信号。ADC 的基本过程是重复地对时间变化的模拟信号波形取样，通常采用固定的时间间隔，并且发生一系列与取样值相近似的数字值。因为从各种传感器获得的信号常常是模拟量，必须通过 ADC 才能转换成电子计算机的数字化语言，因而 ADC 在通信系统、控制系统以及其他许多信号处理系统中有着极为广泛的用途。因此，对提高 ADC 的性能主要是转换速率和精度，以及降低费用等进行了大量的研究。如果系统要求非常高的转换速率，例如取样数超过 10^8 次/s，高速 ADC 就显得特别重要。这样高的转换速率在宽带雷达、电子对抗战，以及核试验监控等系统的信号处理中是十分必要的。

目前高速 ADC 主要用硅材料做成集成电路或混合电路，硅器件是比较成熟的技术，要在功能上有很大改进是不容易的。因而，能使 ADC 达到每秒千兆次取样所用的材料和技术正在开发中[57]。这些新的途径都是利用不同的物理效应，例如，利用轰击半导体靶的电子束的偏转，制成电子轰击半导体器件(EBS 技术)；利用砷化镓中的 Gunn 效应制成的转移电子逻辑器件(TELD)；以 Josephson 效应为基础的超导逻辑器件，以及利用光的电光调制和偏转。光 ADC 采用光波导器件和集成光路的许多优点。

模数转换器的工作原理可用图 29 的方框图来表示。输入 ADC 的是随时间变化的模拟电压波形$V(t)$，而输出是一组数字值 V_{ji}。输入的波形在时间 t_i 被取样，而模拟样本 $V(t_i)$通过编码器变换成数字输出。可能出现的数字化电压 V_j 共有 2^N 个(N 代表精度的位数)。所有电压 V_j 中的每一个可表示为唯一的二进制数，即

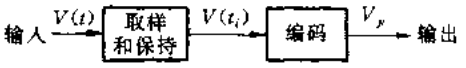


图 29 模数转换原理方框图

$$V_j = a_{1j}a_{2j}\cdots a_{(N-1)j}a_{Nj} \tag{11}$$

其中所有 a_{Nj}是二进制数，即“1”或“0”。在二进制编码中的最低有效位(LSB)定义为：当电压 V 改变时，其数值变化最快的位，而最高有效位(MSB)则随电压的改变而数值最慢变化。取样间

隔 Δt 定义为

$$\Delta t = t_{i+1} - t_i \tag{12}$$

电压差 ΔV 为

$$\Delta V = V_{j+1} - V_j \tag{13}$$

图 30 表示取样和保持电路使信号样本数值在一定时间内保持常数，使编码器足以执行转换。

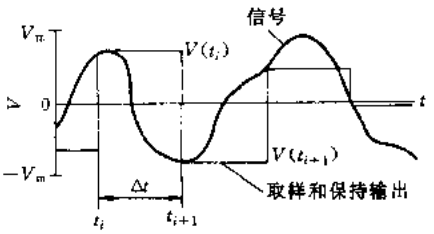


图 30　对时间变化波形的取样

一种应用四个 Mach-Zehnder 波导干涉调制器并列组成的四位模数转换器，如图 31 所示[57]。这种调制器的原理如图 32 所示，它是利用电光晶体的线性电光效应，目前最普遍采用 $LiNbO_3$ 晶体为衬底。构成调制器的通道波导是单模的，输入的激光束是线偏振的。一个需要转换成数字形式的模拟信号电压 $V(t)$ 加到每一个调制器的电极上，这信号电压通过线性电光效应使干涉器两臂中的光束产生相位相对延迟，结果造成输出光束的强度调制。第 n 个调制器的电光相互作用长度 L_n 是由信号电极的长度确定的，其关系为[58]

$$L_n = 2^{n-1}L_1 \qquad n = 1,2,\cdots,N \tag{14}$$

信号电压使干涉调制器的一个臂中光束位相迟后于另一臂中的位相，其位相差 $\Delta\Gamma_n$ 可表示为

$$\Delta\Gamma_n = 2^{n-1}\pi V/V_m \qquad n = 1,2,\cdots,N \tag{15}$$

式中 V_m 是最大电压振幅（$-V_m \leqslant V \leqslant V_m$）。从第 n 个调制器发射出来的光强为

$$I_n = A_n\cos^2\left(\frac{\Delta\Gamma_n}{2} + \frac{\Psi_n}{2}\right) + B_n \tag{16}$$

式中：Ψ_n 是静态相移；A_n 是调制幅值；B_n 是非调制的直流强度。对一种设计得好的调制器来说，$B_n \ll A_n$，即直流分量可以忽略。

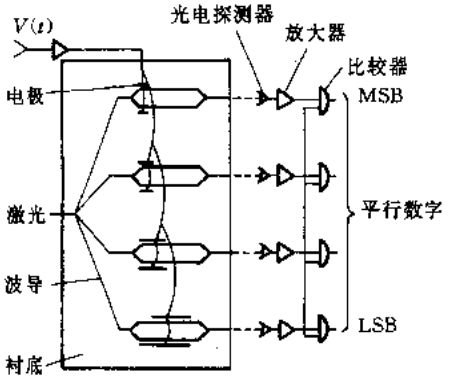


图 31　集成光学四位模数转换器

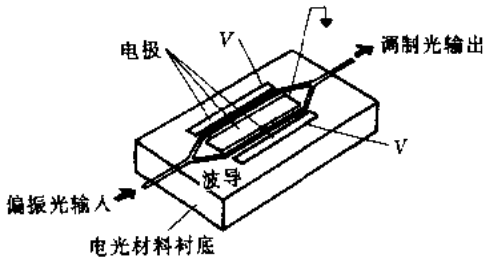


图 32　用于模数转换的光波导调制器

从每一个调制器出来的光束用光电探测器再转换成电信号，经放大后用比较器把 I_n 与一个阈值 I_t 相比较，而得到电压的二进制码数，即在第 n 位上输出“0”或“1”。二进制可以有不同的编码方式，其中有一种叫格雷码（Gray code），其特点是它所表示的任何两个相邻数间仅有一个数字不同，这在许多场合有很大优越性，所以在 ADC 中普遍采用。四位格雷码的模数转换器（ADC）的各光强分量 I_n 的不同情况如图 33 所示。由图可见，误码最可能出现于电压 V 的数值接近于曲线上 I_n 和 I_t 的交叉点处，对格雷码输出来说，任何 V 的微小变化不会同时引起二位数的改变，这可使误码率大大减少。

电光 *ADC* 与常规电路的 *ADC* 相比较有下列几点明显优点：首先是比较器的数量可以大大减少，对 N 位的 ADC 来说，可以从 2^N-1 个减少到 N 个（如对 6 位的 ADC，即从 63 个减少到 6 个）。其次，如果采用了重复脉冲锁模激光器作光源，可以省去取样和保持电路，而这些电路使 ADC 的转换速率和精度受限制。另外，光学 ADC 的一个显著的特点是编码器的光输出能直接记录在移动的照相底片上。这就可能把宽带的瞬时模拟信号直接永久性地记录下来，再用常规的速度进行处理，以获得对原始模拟信号波形能高精度地恢复。

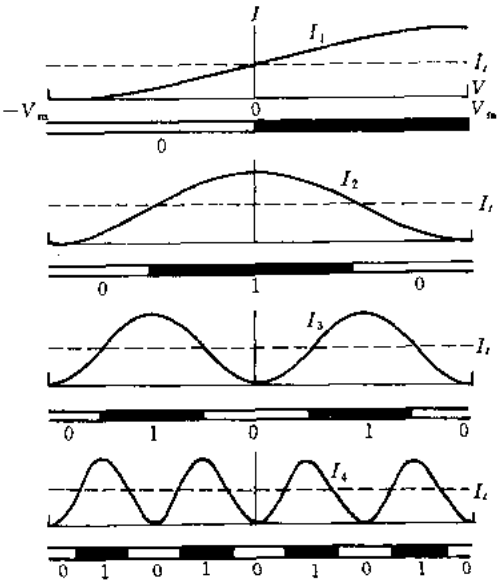


图 33　格雷码输出的四位 ADC 的光强与电压曲线

利用集成光学技术，设计制造取样速率在每秒千兆次以上的 ADC 是完全可能的。已经有实验成功的是四位 275MS/s 的电光模数转换器[59]。器件包含一个锁模 Nd:YAG 激光器，用扩 Ti 的 $LiNbO_3$ 光波导做成的干涉调制器列，锗雪崩光电二极管以及用作比较器和串并转换器的专门的高速 Si 集成电路。$LiNbO_3$ 波导干涉调制器列如图 34 所示。

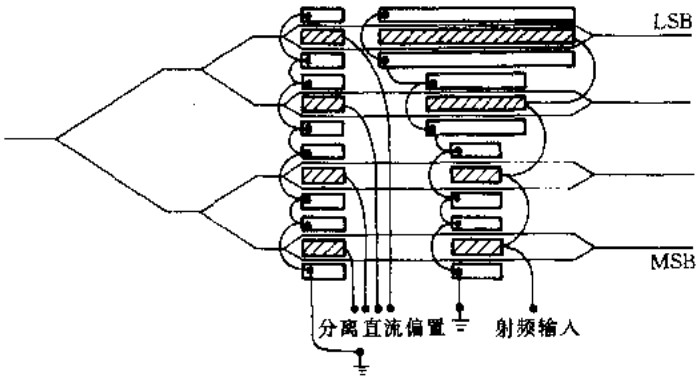


图 34　四位 275MS/s 电光 ADC 的 $LiNbO_3$ 波导调制器列

通过一个公共输入通道把激光束引进各调制器的多路分支。每个调制器具有直流偏压电极来调节干涉器的原始平衡。需要作数字化处理的射频信号同时加到四个调制器上。有三种电极长度相互差 2 倍。最长的电极为 18mm，相当于最低有效位（LSB）。采用格雷码输出，所以最短的两个电极是等长的，只是对其中一个加上引起 $\pi/2$ 相移的直流偏压。光波导宽 6μm，能在 1.06μm 波长下单模工作。$LiNbO_3$ 晶体为 x 切割，y 方向传播 TE 偏振模，利用的电光系数是 r_{33}。对最长的调制器实验测得 V_π 电压为 2.1V，所以使这四位 ADC 工作所要求的峰值电压为 8.4V。衬底全长 6cm，试验时采用端焦耦合（end fire coupling）。

电光 ADC 的另一种设计是可采用平衡桥式调制器结构[60]，如图 35(a)所示。这一调制器的优点是当光源脉冲强度涨落时，不易产生误码，由于对同一调制器的两个比较器的输入有相同的影响。利用平衡桥式调制器制成的 2 位 ADC，如图 35(b)所示[61]，实验的转换速率为

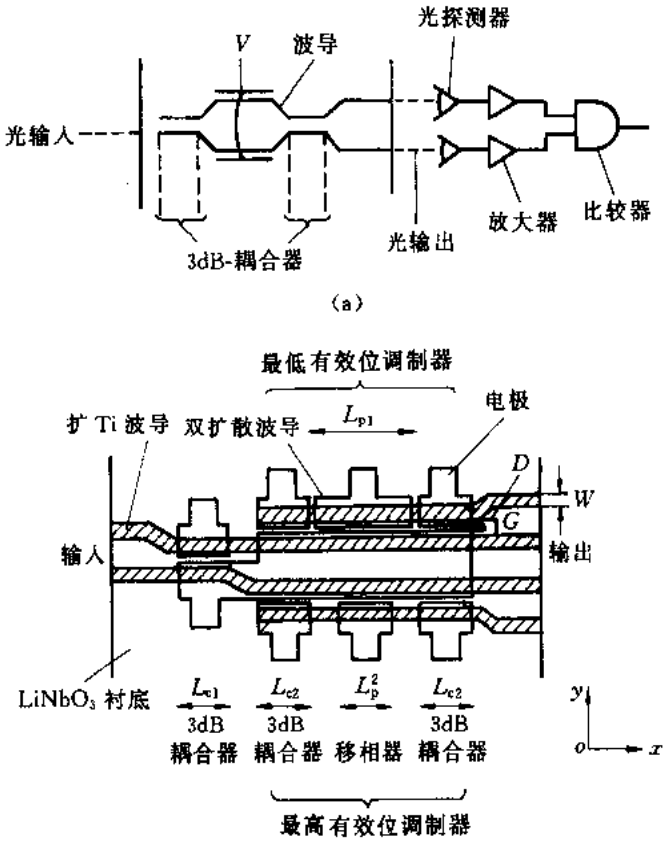


图 35　另一种电光 ADC

(a) 平衡桥式调制器；(b) 2 位模数转换器

240MHz/s。器件由一个 3dB 方向耦合器将光功率分配到两个平衡桥式调制器。每个调制器又包含两个 3dB 耦合器和一个移相器。为了能使移相器中的光功率不产生耦合，这部分的波导采用两次扩散的方法。整个器件长 29mm，耦合器和移相器的长度分别为 4mm 和 8mm。最低有效位(LSB)调制器中移相器的电极长为 8mm，最高有效位(MSB)的电极长为 4mm。波导宽 W、波导间隙 G 和电极间隙 D 分别为 4.6μm、6.2μm 和 5.5μm。试验时采用脉冲工作的 GaAlAs 激光二极管($\lambda = 0.83\mu m$)作光源和 Si 的 APD 为光探测器。实验结果预示，用这种调制器做成 4～6 位的 ADC 是完全可能的。如果比较器的集成电路得到改进，这种 ADC 的转换速率可望达1～2GS/s。平衡桥式调制器 ADC 的缺点是，结构图形比较复杂，由此带来工艺上的困难。

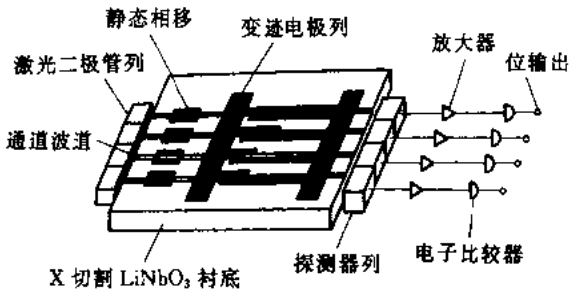


图 36　应用 Fabry-Perot 调制器构成的四位 A/D 转换器

应用通道波导 Febry-Perot 调制器[62]设计制成的四位电光模数转换器如图 36 所示[63]。由于每一个调制器只是一条直的通

道波导，所以有许多优点，如衬底尺寸小、几何图形简单以及光插入损耗低等。

7 逻辑集成光路

在光信号系统中，光逻辑运算和光计算的功能是必不可少的。要实现光逻辑功能，必须在两个或两个以上的光信号之间具有非线性相互作用，因此，Taylor 曾致力于研究不同的机理来获得两个激光束之间的非线性效应，但这些方法的缺点是每个逻辑元件消耗的电功率和光功率太大。已经证明无论对电子的或光的计算机来说，器件材料中所产生的热是决定其速度的重要因素。上述一些方法所实现的光逻辑元件所产生的热，比高速电子逻辑元件大几个数量级，因而没有实用意义。

后来应用集成光学方法做成的许多光波导调制器和开关具有速度高、电压低和功耗小等特点，用这些光波导元件来构成逻辑光路是有可能的。利用电光效应波导调制器和开关构成集成逻辑光路的优点是，门传输时延短和功耗低[64]。与上述利用非线性光学效应的逻辑元件相比，每个逻辑门的功耗将小几个数量级。这些光波导元件能构成各种功能，例如，与门、非门、或门等的集成光路。电光波导逻辑器件的原理如图 37 所示。从激光器发出的光束耦合进电光晶体衬底上的通道波导，光束通过的开关和调制器由输入的电信号电压控制，最后由波导输出的经过逻辑运算的光信号经探测和放大作为输出的电信号。

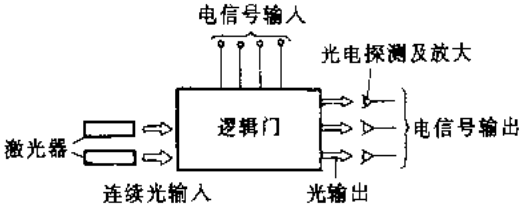


图 37 电光逻辑器件原理方框图

构成各种逻辑门的基本元件有干涉调制器和耦合波导开关，分别如图 38 和图 39 所示。假定所用的激光束是线偏振而且是单模工作的。

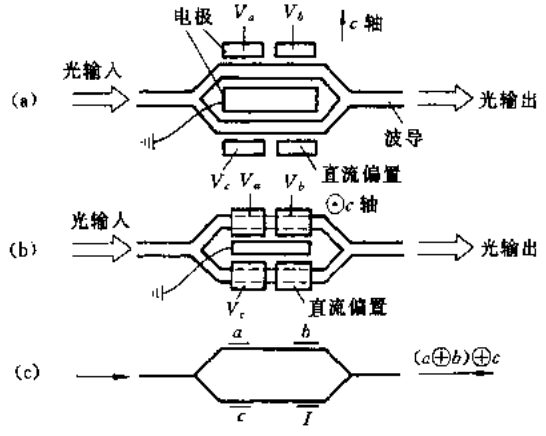


图 38 干涉调制器逻辑门

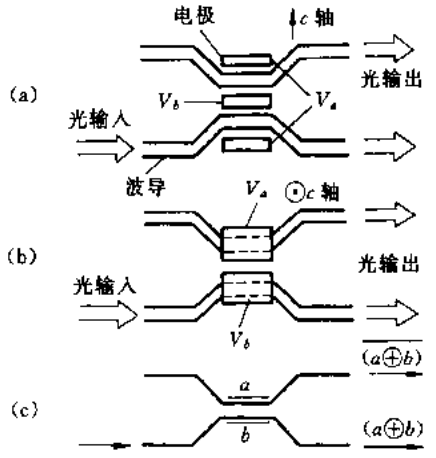


图 39 耦合波导开关逻辑元件

图 38(a)中的 c 轴平行衬底的表面，图(b)中的 c 轴则垂直于衬底表面，两种不同方向切割的晶体所设计的电极结构是不同的，其目的是使在两种情况下最大电场的方向都与 c 轴平行，这样对 $LiNbO_3$ 来说，可利用较大的电光系数 r_{33} 产生光束在波导中传输的相移。图(c)表示这一逻辑元件的简化符号图。在没有信号电压时，可认为干涉器两臂内光束的相移是对称的，当干涉器的四个电极上加上电压后，则两臂内的光束就产生一定相移差。如果相对相移是

$2N\pi$ 弧度，其中 $N=0,\pm1,\pm2\cdots$，则波导的输出有导模；如果相移差是 $(2N+1)\pi$ 弧度，$N=0,\pm1,\pm2\cdots$，则波导输出的导模消失，光功率由于辐射进衬底而损耗。

要把这种干涉调制器设计成二进位逻辑元件，必须使某一特定电压（例如 V_0）加在一信号电极上时产生相移差为 π 弧度。二进位的“1”代表电势 V_0，而二进位的“0”表示零电势，即地电势。只有假设“1”输入的数目是偶数时，调制器才有光传输。如果调制器由连续工作的激光作光源，它成为电信号输入和光信号输出的偶数宇称发生器。当器件在另外一个电极上加上直流偏置电压 V_0 时，可以转换成奇数宇称发生器，即只有“1”的输入数目为奇数时，光才能通过。

图 39 所示的耦合波导开关逻辑元件中，信号电压控制两个波导之间的功率耦合。在全部耦合区，波导靠得很近，只有几个波长的距离，并且波导尽可能做得相同。在不加信号时两波导中光的传播常数相等，输入到一个波导的光功率可以有效地耦合到另一个波导中去。能把功率全部转移过去的那段距离 L 称为耦合长度，在这情况下，可认为波导是同步的。当有信号作用于电极时，波导不再同步，即波导中光的传播常数不再相等，这将引起耦合效率减小和耦合长度的改变。在某一特定电压下，全部功率可以回到原来的波导。因此，在开关状态“1”，对应于不加电压，光从一个波导进去而从另一个波导射出；在另一状态，相应于有信号作用，光则从进去的那个波导射出。如果输入到两个信号电极都是“1”或都是“0”，则光功率从一个波导注入而从另一个波导射出。相反，如果一个电极输入“1”，另一个电极输入“0”，则光功率从同一个波导射出。当有一连续工作的激光束输入其中一个波导时，从同一个波导接收到的光束表示“异或门”功能，记为 $(a\oplus b)$；而从另一个波导接收到的光束是其互补的功能，即“异或非门”记为 $(\overline{a\oplus b})$。

从这两种波导逻辑元件可以构成许多基本逻辑运算门，如图 40 所示。这两种元件采用图 38 和图 39 中的符号。电极旁的字母 a 和 b 表示输入电信号，电极旁的“l”字表示直流偏压，它使干涉调制器产生 π 弧度相移。除了图中(f)输入的是光信号“a”和电信号“b”，其余的情况中输入的都是 CW 光束，“a”和“b”全部是电信号。图中(b)是表示“非门”$(\bar{a})$；(c)和(d)是“异或门”$(a\oplus b)$；(e)和(f)是“与门”$(a\cdot b)$；(g)是“或门”$(a+b)$。

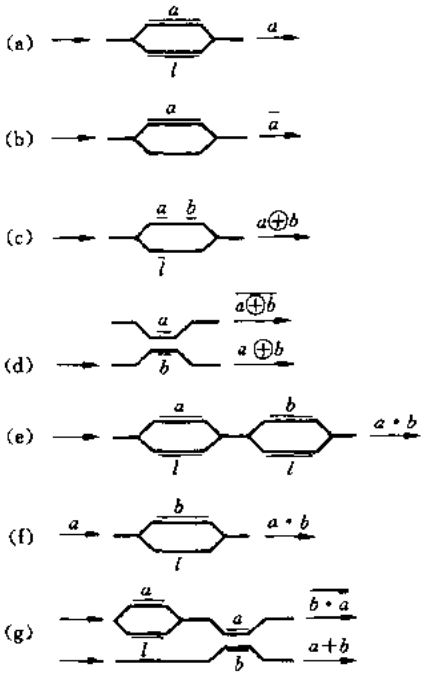


图 40　由干涉调制器和耦合波导开关构成的逻辑门

这类光逻辑器件的一个引人注目的特点是传输时延小。光经过一逻辑门所要求的时间 τ 表示为

$$\tau = nl/c \tag{17}$$

式中：n 是衬底材料的折射率；l 是门的长度；c 是真空中的光速。在 $LiNbO_3$ 中，$n=2.2$，则

$$\tau = 7.3l \quad (\text{ps/mm}) \tag{18}$$

门的长度 l 与信号电极长度 l_e 有关，它要求能产生 π 弧度相移，有

$$l_e = \frac{\lambda d}{n_e^3 r_{33} V_0} \tag{19}$$

式中：λ 是真空中光波波长；d 是电极间距离；n_e 是非

常光的折射率。式中假设波导区的电场是均匀的为 V_0/d，而实际上这电场是不均匀的。今设 $V_0=5\text{V}$，$d=6\mu\text{m}$ 和 $\lambda=0.8\mu\text{m}$，从材料手册中查出 $LiNbO_3$ 的 n_e 和 r_{33}，可算出 $l_c=3.0\text{mm}$。对耦合波导开关来说，电极长度略短一些。在图 40 中，(a)、(b)、(d)和(f)的门的总长度近似为 l_c，而其余的门的总长近似为 $2l_c$。由此，可以估算出每门的传输时延在 20～40ps。

上述基本逻辑门连接起来可以实现数字计算的功能。图 41 所示的逻辑集成光路所表现的功能是二进制数字第 n 位的加法器。这器件能计算和

$$C=A+B \tag{20}$$

式中 A 和 B 是以二进码表示的 n 位的加数，$A=a_n\cdots a_1$，$B=b_n\cdots b_1$。则 C 可表示为

$$C=c_{n+1}c_n\cdots c_1 \tag{21}$$

式中 a_n、b_n 和 c_n 是二进制数字，即“0”或“1”。加法运算分成二项进行，一是计算进位(carry bit)，另一是计算和位(sum bit)，进位 k_n 表示为

$$k_n=a_nb_n+k_{n-1}(a_n\oplus b_n) \tag{22}$$

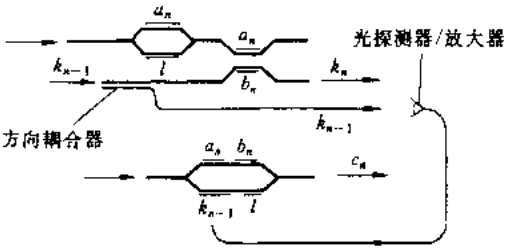


图 41　一位加法器的逻辑集成光路

和位 c_n 表示为

$$c_n=(a_n\oplus b_n)\oplus k_{n-1} \tag{23}$$

进位 k_n 与和位 c_n 都是 a_n、b_n 和 k_{n-1} 的函数，其关系见加法器的真值表(见表 4)。

表 4　二进制加法器真值表

输入			输出	
a_n	b_n	k_{n-1}	c_n	k_n
0	0	0	0	0
0	0	1	1	0
0	1	0	1	0
0	1	1	0	1
1	0	0	1	0
1	0	1	0	1
1	1	0	0	1
1	1	1	1	1

在图 41 中，输入电信号与 a_n、b_n 相当，以平行方式加在调制器和开关的电极上。进位依次分 N 级进行计算，从第 n 级得到的二进制光输入 k_n 传输到第 $n+1$ 级作为该级的输入。因为 $k_0=0$，所以第一级上没有代表进位的光输入。用一个方向耦合器把每一级的输出光信号分出一部分，供下一级的进位光信号输入。从第 $n-1$ 级产生的进位光信号 $k_{n-1}(n=2,\cdots,N)$用光电二极管探测并放大后，再作为电信号反馈输入到一个宇称发生器，用来计算第 n 级的和位 $c_n(n=2,\cdots,N)$。计算和位的其他电信号是 a_n、b_n 和一个直流偏压。N 个光宇称发生器的输出逐个进行探测，并放大平行构成二进制的和 c。最后一级进位输出 k_n 相当于 c_{N+1}，是和中的最高有效位(MSB)。

对这种加法器所需的电功率消耗估算表明，它略低于最快的电子逻辑电路。其传输时延极短，仅 20～40ps/门，但现有的探测和放大器的响应较慢，需要在光逻辑器件的总时延上另加

1ns左右。因此，需要在探测以前连续进行多次运算，充分利用光传输时延短的优点。

利用集成光学技术设计制备的逻辑光路还有许多不同的实验。例如，有一种 $LiNbO_3$ 衬底上制备的电光调制器是利用通道波导中传输的最低次模的截止条件的改变来实现的[65]，而且利用 CdS 光探测器把光信号转变成电信号，然后对另一通道波导中的光信号进行控制，这样就可实现光信号与光信号之间的逻辑运算。利用这种方案做成的集成光路半加法器，其特点是：(1)可在同一衬底上集成许多元件实现逻辑门运算，而不需其他附加的电子或光学元件；(2)单独的逻辑元件连接方便，因为它们的输出只有"0"或"1"；(3)全部高码速信号都是光信号，可以通过光纤与集成光路芯片耦合进去或耦合出来；(4)对制造的精度要求不高。

另一种方法是利用包含一个电光调制器的 Fabry-Perot 波导谐振腔[66]，调制器由一部分传输的光功率驱动，器件呈现出回线和双稳态，因而可利用作为逻辑运算的基本元件。

利用耦合波导开关中部分的输出光功率探测后转换成电功率，再反馈至开关的电极上可以获得光控四端双稳态开关器件[67]。还有其他许多光双稳态器件都有可能实现逻辑运算。

各种不同结构的光波导逻辑门继续不断在研究开发中，如不久前又制成一种新的全光异或门(XOR)[68]，这是扩钛 $LiNbO_3$ 波导做成的一种新结构的干涉器。

8 光波导和光纤传感器[69]

利用光导波作为信息来制成传感器可以说从 1977 年才真正开始。但是光波导和光纤传感器同传统的传感器相比具有一系列独特的优点，而且可能应用的领域极为广泛，所以许多实验室很快就投入了这方面的研究，在短短的几年时间内取得了惊人的进展，已经报道和正在研究中的光波导和光纤传感器有近百种，并且在 1983 年专门召开了光纤传感器的国际会议。

光波导和光纤传感器的基本原理是，利用环境对光导波的影响取作为传感信息，包括对光导波幅值和位相的影响。因而各种不同的光波导和光纤传感器可分成幅值(或光强)和位相(或干涉)传感器两大类。它们的主要优点如下：

(1) 具有比一般传感器高得多的灵敏度，特别是位相传感器。因为光波的波长比无线电波小几个数量级。

(2) 应用广泛。可以对许多物理量，如声、电磁波、温度、应力应变、转动和加速度，以及吸附等的微小变化产生响应，而且光纤传感器还可依不同的应用场合设计成不同的几何结构。

(3) 可用于一般传感器不能适应的特殊环境。因为光波导和光纤都是电介质器件，因而在高电压、电噪声严重、高温、化学腐蚀等环境中应用常规的金属材料或半导体材料制成的传感器是十分不利的。

(4) 与光纤遥测技术兼容。不论是光波导传感器或光纤传感器，都可以用损耗极小的光纤传输系统相连接实现遥测和遥控。

光强传感器的特点是比较简单，而且可以和多模光纤传输系统相兼容，但其灵敏度不如干涉传感器。在许多场合，权衡各方面的要求，如果对灵敏度要求不是非常高，往往可采用这种传感器，它与现有的各种传感器具有很大的优势。

位相传感器无论用于声、磁场或转动等场合，在理论上可以比现有的相应的传感器高几个数量级。以声传感器为例，利用光纤干涉器做成的声传感器已达到目前声测量的极限。光纤

制成的磁传感器可以在室温工作，其灵敏度等于或超过目前的低温磁测量技术，仅需 4～10K 的低温条件。所以位相传感器的高灵敏度和几何形状适应性具有很大竞争潜力。

光波导和光纤传感器虽已有很大发展，但离开实用化和大量生产还有许多技术问题有待研究解决，主要的是：噪声的抑制、探测系统的最优化、传感器的封装以及光纤的涂层等。可以相信，不久的将来这些问题都是完全可以解决的。

以下就声、磁、转动等位相传感器、光强传感器以及其他一些主要的光波导和光纤传感器的工作原理、设计和工艺、性能、试验等方面分别加以讨论。

8.1 声传感器

用于声传感器的光纤干涉器一般采用 Mach-Zehnder 干涉器结构[70,71]，如图 42 所示。从激光器发出的激光束分成二路：一束输入置于声场内的光纤传感器；另一束则进入参考光纤。在参考光路中采用某种方法产生光谱偏移，如采用布拉格调制器，或产生位相调制（如采用光纤展宽器或集成光路移相器）。两束光重新汇合后进入光探测器，并采用解调器探测原始的位相调制信号。已经试验过的各种解调器有：调频鉴频器、稳频零差接收、合成外差接收以及其他方法。

光纤干涉器做成的声传感器具有非常高的灵敏度，它与常规压电传感器灵敏度的比较见图 43[72]。该传感器所不同的是以光纤为塑料涂层，其信噪比由量子极限所决定。传感器的探测灵敏度与光纤长度有关，图中给出了人的听力极限和普通压电传感器的特性以作比较。由此可见，光纤长度仅 1m 的传感器，其灵敏度已超过普通的压电传感器。

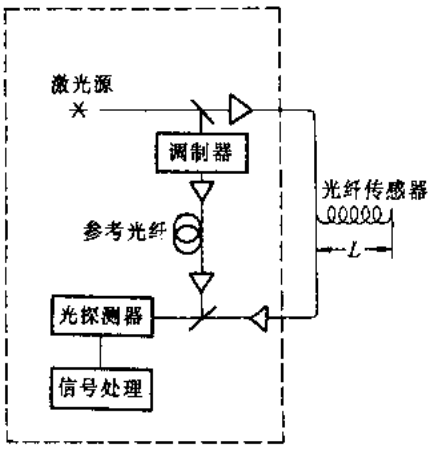


图 42　光纤干涉器的基本原理

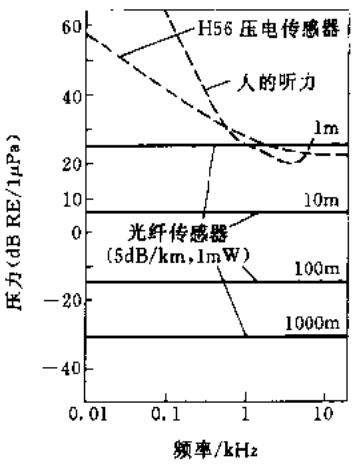


图 43　不同长度光纤干涉器的最小可探测压力

光纤传感器可以做成不同几何形状的元件，如做成平面的或线列型的；也可用单根光纤元件使其长度比声波波长长得多而成为高度定向的探测器。另外，利用两个靠近的光纤圈可以做成声场梯度传感器，其中一个圈即为图 42 中的参考光纤，如果两个光回路的压力灵敏度是匹配的，则这对光纤圈所响应的不是压力而是压力梯度。利用这种梯度传感器可以测量入射声波的方向。

1977～1978 两年中，光纤声传感器的灵敏度差不多提高了 100dB。改善信噪比的一个有效方法是采用了全光纤的干涉器，它的噪声水平在实验室条件下可以减小到 100Pa，声频可低

至数百赫兹[73]。

8.2 磁传感器

利用光纤传感器来测量微弱磁场有法拉第旋转法和磁致伸缩法，采用的光纤有常规的 SiO_2 单模光纤和掺杂的 SiO_2 光纤。前一种方法是将外磁场纵向作用于光纤上，使在线偏振方向发生旋转，这种方法做成的传感器[74]可用来测量 0～1kA 的大电流而不致造成电缆短路的危险。但是多数掺杂光纤的 Verdet 常数非常小（$\approx 1.5\times10^{-2}$min/A），因而只有在比较大的电流和强磁场情况才可采用此法。虽然可以采用在光纤中适当掺杂的方法来提高 Verdet 常数，如将顺磁离子掺入可获得比玻璃基体的上述常数来得大。稀土金属离子加入 SiO_2 光纤中可显著增强法拉第效应，但是由于这些离子在玻璃中的溶解度是有限制的，以及离子的引入增大了玻璃的光吸收，因而在实际上并不能有很大改善。这种稀土元素掺杂的光纤，在理论上其灵敏度达 0.1m G/m。总的来说，由于法拉第效应既要用特殊的材料又要求复杂的拉制工艺，因而其发展不如利用磁致伸缩方法来得快。

利用磁致伸缩材料来做磁传感器，不仅较为简单且有可能获得很高的灵敏度。这种传感器的基本工作原理是，在光纤外涂一层磁致伸缩材料的外套，或将光纤固定在磁致伸缩的材料上，因而在磁场作用下光纤经受纵向应变。最早的实验方案是在光纤外包上一层镍套[75]，传感器做成光纤干涉器，这完全和上述的声传感器一样，所不同的是声传感器的光纤采用塑料涂层。用体镍作为光纤包层做成的光纤传感器元件在频率 1～10kHz 下，第一次所获得的实验结果为 0.8nG/m[76]，而用镍的薄膜涂层光纤做成传感器的灵敏度就低于上述数值。利用磁致伸缩材料做成光纤偏振传感器同样可以获得很高的磁场测量灵敏度[77]。

光纤磁传感器当前研究的主要问题是，挑选合适的高磁致伸缩材料以及如何将它与光纤结合在一起。磁致伸缩材料可分成晶态金属和金属玻璃两类。磁致伸缩金属包括：Fe、Co 和 Ni 以及这三种元素的各种合金和化合物；金属玻璃现有的商品一般都是以 FeBSi 体系为基础的，已成功地用于制备高灵敏度传感器。

磁致伸缩材料与光纤结合的方式主要有三种，如图 44 所示。一种是将光纤绕在磁致伸缩材料的杆或管上；另一种是在光纤表面形成均匀的涂层或套；还有一种是将圆光纤固定在金属或金属玻璃条上。薄膜涂层可采用真空蒸涂或电镀方法，厚度一般约 10μm，为了消除金属包层的应力，传感器在工作以前需经过退火处理。上述三种中最满意的是采用光纤固定在金属玻璃条上的这种结构。这种非晶合金是将熔融的材料倒在冷的旋转飞轮上急速淬火而成为 3cm 宽、500μm 厚的带，光纤用环氧树脂粘结在该带上。采用镍涂层光纤制成的传感器在 5～100kHz 内的灵敏度为 1μG/m[76]。采用金属玻璃条粘结光纤制成的传感器的灵敏度达

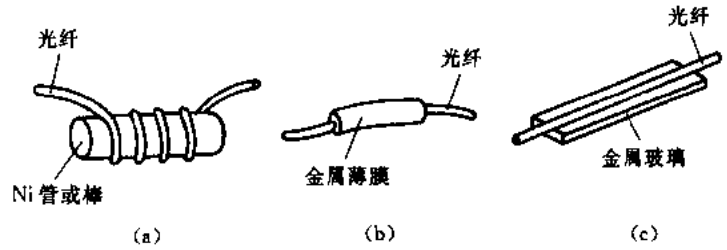


图 44　光纤磁传感器的基本结构

5nG/m。

坚固而高灵敏度室温磁场传感器目前最灵敏的磁测量装置是超导量子干涉器(SQUID),在实验室条件下它能测量 10～1pG 的弱磁场,但它要求复杂的低温装置,这在现场测试中往往是不具备的。光纤传感器不难在室温下测量 10^{-5}～10^{-10}G 范围的磁场,因而具有明显优越性。

8.3 光纤陀螺

最早提出可以利用无源光纤环干涉器作为惯性转动传感器是在 1976 年[78]。光纤陀螺的基本工作原理如图 45。单模光纤构成的环形光路可以作为 Sagnac 干涉器。陀螺围绕垂直于光纤环平面的轴转动,将在相反方向传播的两个光路之间产生非互易的位相差:

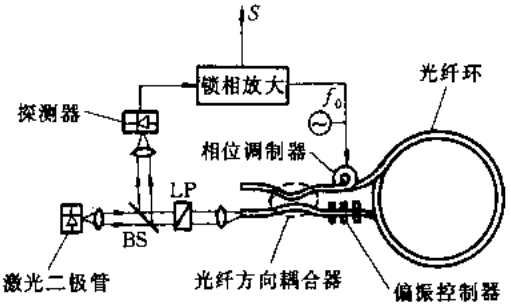


图 45 光纤陀螺基本原理

$$2\phi = \frac{8\pi NA\Omega}{\lambda c} \tag{24}$$

式中:NA 是光纤环包围的总面积;λ 和 c 分别是真空中光的波长和速度;Ω 是转速。当这两光速在光纤定向耦合中结合和在探测器中混合时,2ϕ 的相移导致光强随转动速度而改变

$$I = I_0(1+\cos 2\phi)/2 \tag{25}$$

在这种结构中,陀螺的灵敏度为

$$\left(\frac{1}{I_0}\right)\left[\frac{\mathrm{d}I}{\mathrm{d}(2\phi)}\right] = \frac{1}{2}\sin\phi \tag{26}$$

此式表明在转速低时陀螺的灵敏度也低。

为了使小信号获得最高灵敏度,在光纤环的相位调制器中,在相反传播的两个光路之间加进一个非互易的 $\pi/2$ 移相器。这样,探测器的灵敏度为

$$I = I_0(1-\sin 2\phi)/2 \tag{27}$$

在低转速时呈线性变化,而陀螺的灵敏度为其最大值的 1/2。这种非互易的 $\pi/2$ 移相器可以采用不同方法,可以由光纤来做或用光波导做。

对光纤陀螺灵敏度极限的理论估计是基于对探测器量子噪声的因素[79]。当然,从目前的实验结果看还有许多其他的噪声,但是随着技术的进展光纤的损耗已越来越小,这表明灵敏度极限决定于量子噪声的因素是有根据的。噪声所具有的光能是可以测量的,它与光子平均值的平方根成比例。只有当相应的光强变化大于这种量子噪声时,Sagnac 相移才可能测量出来。这就是,当干涉器在最大灵敏度下工作时,最小的可测 Sagnac 相移为

$$2\phi = \frac{\Delta I}{I_0} = \frac{1}{\sqrt{n_{\mathrm{ph}}}} \tag{28}$$

式中 n_{ph}是光子的平均值。由此可以写出陀螺的随机漂移为[69]

$$\theta_{\mathrm{drift}} = \frac{25\lambda c}{\pi LR}(10^{-\alpha L/10})^{-1/2}\left(\frac{P\lambda}{hc}\right)^{-1/2}(\mathrm{rad/h^{-1/2}}) \tag{29}$$

式中:L 和 R 分别为光纤长度和环的半径;α 是光纤的损耗,单位是 dB/km;h 是 Planck 常数;P 是注入光纤环的功率。在上式中,其他一些损耗,如耦合损耗和插入损耗等都忽略了。

根据式(29),把目前光纤损耗能达到的最低值代入,并把输入功率归一化为1mW,由此得到光纤陀螺灵敏度的理论值如图46所示。这些曲线说明光纤陀螺由于量子噪声极限所决定的灵敏度是十分高的。但是迄今为止,所有的实验结果都远较此理论值低。目前能达到的最高水平为1~0.1°/h。这样的灵敏度值已经可以在低灵敏度要求的场合下应用,在军事上已可用于战术导弹制导及飞行器导航等。

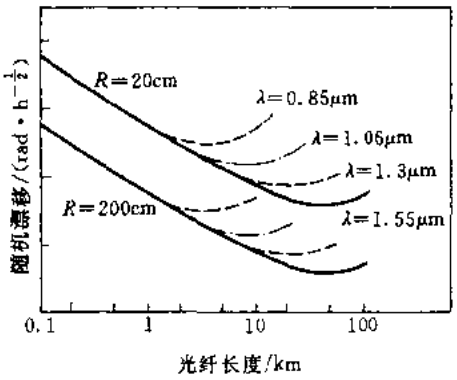


图46 光纤陀螺随机漂移的理论曲线

最近,大量的研究已发现有许多种噪声来源限制了光纤陀螺的灵敏度。这些噪声来源包括:两相反传播光路上的非互易效应,例如:法拉第磁光效应、光纤的双折射以及温度和压力梯度随时间的变化等所引起的噪声和不稳定性。为了使微小转动时获得最大的灵敏度,需要在相反传播的两光束之间引入非互易的稳定的$\pi/2$相移,这一相位偏置的任何不稳定性是噪声的重要来源。光源发射光束的偏振和幅值的不稳定性;光纤中光束在传播过程产生的背散射以及在连接或耦合处的反射;背散射光功率反馈到激光器产生的微扰、动态范围和码速超过了器件的最佳工作区等等因素也会引起噪声。

针对这些可能出现的噪声源,对激光器、探测系统、光纤以及各种光纤或波导器件如方向耦合器、偏振转换器、位相调制器、起偏器和退偏器等的设计和性能作了许多研究,因而获得了很大改进,致使光纤陀螺实验室样机的灵敏度差不多比最早的提高了近三个数量级,最近的实验结果其灵敏度提高到0.1°/h量程[80]。这一新的光纤陀螺是改进了的全光纤,结构如图47所示。其中耦合器、位相调制器和传感线圈都是由同一种高双折射的偏振保持的单模光纤做成。所用的光源是超发光二极管(Super Luminescent Diode, SLD)[81]。光纤具有3.8dB/km的损耗,双折射拍长为3.4mm。光纤耦合器产生49%~51%的分束,插入损耗为0.2dB,位相调制器是将光纤绕在压电材料的圆棒上形成,光纤线圈的直径为32cm,光纤长为430m。探测器是硅二极管。

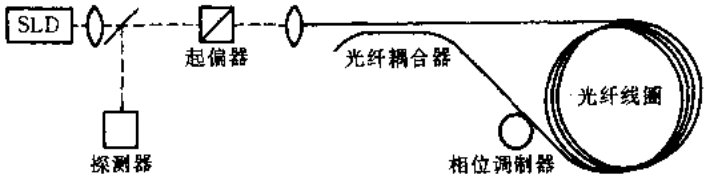


图47 全光纤激光陀螺

目前,在光纤陀螺方面继续为提高灵敏度和稳定性而努力,研究的重点是改善激发器的性能,采用偏振保持的单模光纤,选用最佳的探测系统和实现光路的集成比。

8.4 幅值传感器

上述几种都是基于干涉器原理的位相传感器,这一节讨论幅值传感器。这类传感器的原理是,所产生的光强变化正比于外加信号,一般说来,它比位相传感器简单。但是,如同位相传

感器一样，幅值传感器也能用来探测各种不同的场。下面仅以水下声传感器为例作简单讨论。

声探测阈值作为传感器之间的比较是最理想的参数，因为低的水下声背景使声传感器性能最需要知道。对任何一种光传感器来说，归一化调制指数是一个基本参数，可定义为

$$Q = \Delta I / I_0 p \tag{30}$$

式中：ΔI 是引起的光强变化；I_0 是激光器输出光功率；p 是声压力。调制指数 Q 所以值得重视，因为传感器响应和最小可探测阈值都可从这参数计算得出，这两参数是传感器必须要的特性。按照惯例，传感器响应是单位压力变化所引起的电压差，可表示为

$$S = qI_0RQ \tag{31}$$

式中：q 是探测器的响应，单位是 A/W；R 是探测器的负载电阻，单位是 Ω。

同样，以量子噪声为极限的探测器阈值可表示为

$$P_{\min} = \frac{1}{Q}\left(\frac{2eB}{qI_0}\right)^{1/2} \tag{32}$$

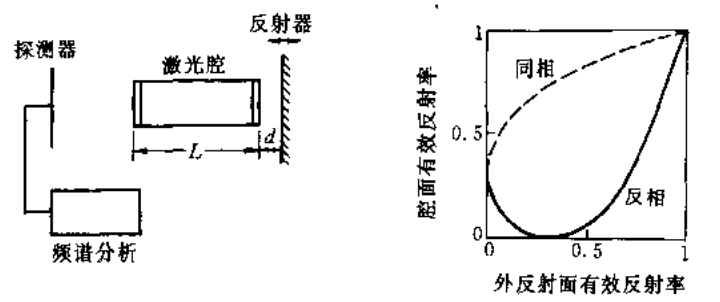
式中：e 是电子的电荷；B 是探测器的带宽。

根据幅值传感器不同的工作原理可算出其归一化调制指数。这些传感器包括光纤微弯曲传感器，由于压力的变化使光纤经受不同程度的弯曲。因而光束的损耗不同，输出的光功率成为压力的函数。消衰波或耦合波导传感器是将两根光纤中的一部分作平行放置而构成光纤方向耦合器，当压力变化时引起耦合条件变化。可动光纤传感器是两根端接耦合的光纤其中一根固定，另一根随压力而产生微小位移，因而改变耦合的光功率。偏振传感器是基于压力使单模光纤的双折射改变。还有许多不同的形式、不同的应用场合可以选择适当的传感器结构。

8.5 激光二极管传感器

许多场合对传感器的灵敏度要求并不高，但希望器件简单而小巧，激光二极管传感器[82]能满足这样的要求。它是利用外反射镜反馈进激光二极管腔内的光的位相调制。这种传感器具有的灵敏度可与现有的诸传感器相竞争，并且结构简单，能封装成很小的器件。这种器件也能成为声、磁、电流和加速度的传感器。

图 48(a)表示这种传感器最普通的实验结构，它包括 GaAlAs 的单模激光器和一个外反射器，外反射器的位置受被测的入射信号的调制。输出信号获得的途径可以采用靠近激光器端面的大面积光电二极管调制强度，也可以对二极管的驱动电流进行直接监控。在图(a)结构中，光的位相决定于外反射镜与二极管一端面之间的距离 d，当反馈的光与激光器腔内的光同



(a) 结构图　　(b) 腔面有效反射率的曲线

图 48　激光二极管传感器

位相时，就提高端面的反射率；反之，端面反射率下降。

二极管激光器腔（反射率一般约 0.3）与外反射器的有效反射率作为 d 的函数，可利用标准 Fabry-Perot 方程对具有三个不同反射镜面进行计算。腔面与外反射面合成的有效反射率是外反射面的反射率的函数，对于同相和反相有不同的关系，如图 48(b)所示。由图可见，如果外反射器的反射率与激光腔的相等，则当外反射器调节到某适当位置时激光器将熄灭。然而，在实验中，反射器的反射率低于 10%，所以不可能达到完全熄灭。当外反射器反射率不大时，腔面有效反射率可表示为

$$R = R_0 + 2(1 - R_0)\sqrt{rR_0}\cos\theta \tag{33}$$

式中：R_0 是没有外反馈时腔面的反射率；r 是外反射器的有效反射率，它应是外镜面反射率与耦合效率的乘积；θ 表示反射光的位相。

在激光阈值下的增益 g_0 可表示为

$$g_0 = \frac{1}{2L}\ln RR_0 \tag{34}$$

式中：L 是激光腔体长度。因而阈值电流随反馈光的位相而变化。保持激光电流不变（在相反条件下达到阈值），改变反馈光的位相将引起激光器输出的显著变化。这些效应就是激光二极管传感器的基本原理。只要外场对反射器的位置产生扰动，就能用这一方法作为传感。

应该指出，光纤可用作为外反射面与激光器之间的连接，这时传感器可以响应光纤特性的任何变化，做成一种外传感器，这样就把光纤传感器和激光传感器结合起来。

用玻璃薄膜作为反射面（反射率约 4%）可使二极管激光传感器成为声传感器。声波作用在膜片上调节膜片的位置，从而改变反馈光位相。也可用金淀积在聚酯薄膜上作为反射面而做水听器，其结构示于图 49(a)[83]。将反射器安放在具有磁致伸缩特性的镍管上，就能构成激光二极管的磁场传感器，如图 49(b)所示。用类似的设计，激光二极管传感器可应用于对电流和加速度的传感。

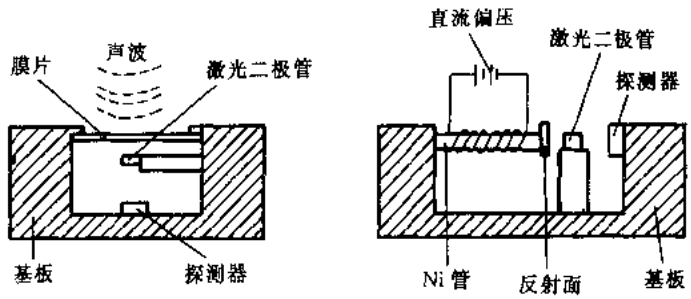


图 49　激光二极管传感器的应用

(a) 水听器；(b) 磁场传感器

8.6　介质波导传感器

介质波导不仅在光纤传感器系统中作为波导器件（如移相器）和光路连接的集成化，而且介质波导本身也可作为传感元件。其工作原理类似于光纤传感器，可以为位相的改变或是幅值的变化。

集成光路温度传感器[84]如图 50 所示，它由一组不等臂长的 $LiNbO_3$ 干涉器构成，全部是光连接，没有任何电连接。其设计的工作温度大于 700℃，分辨率为 2×10^{-3}℃。该传感器十分类似于光纤温度传感器，其特点是集成光路器件由单片集成，器件结构紧固，在要求不受电噪声影响的场合特别有用。这种传感器可用于测量压力、应变、电场、磁场和温度梯度。

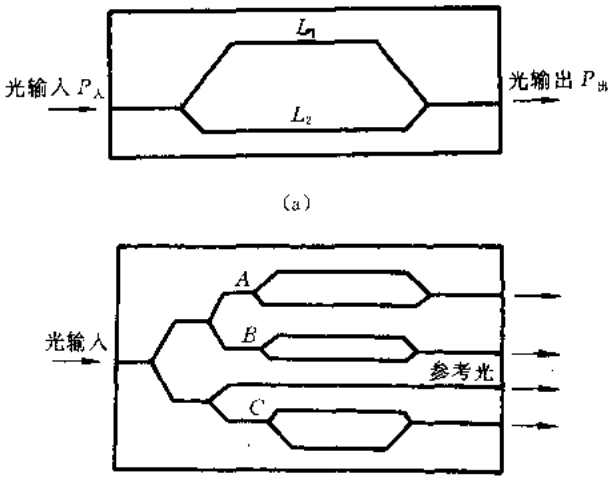


图 50　集成光路温度传感器

(a) 不等臂长波导干涉器；(b) 包括三个干涉器的结构

集成光路温度传感器的基本元件如图 50(a)所示，是一种波导形式的 Mach-Zehnder 干涉器。由一输入通道波导分成不等长的两臂 L_1 和 L_2，最后仍合起来与输出波导相连接。所有通道波导都是单模的，有效折射率为 n_{eff}。与等臂长的干涉器不同，路程差 $\Delta L = L_1 - L_2$，比光的波长大得多。在波长为 λ 时，两臂之间的相移差为

$$\Delta\phi = 2\pi n_{eff}\Delta L/\lambda \tag{35}$$

光传输(P_{out}/P_{in})将随 $\Delta\phi$ 作正弦变化。n_{eff}和 ΔL 都随温度而变化，因此，在这些参数与温度有线性关系的范围内，P_{out}/P_{in}与温度的关系可表示为

$$\frac{P_{out}}{P_{in}} = \frac{\gamma}{2}\left[1 + m\cos\left(\frac{2\pi}{\lambda}b\Delta LT + \Delta\phi_0\right)\right] \tag{36}$$

这样，器件温度的变化能通过光传输的变化来测定。式(36)中的比例常数 b 可表示为

$$b = \frac{\mathrm{d}n_{eff}}{\mathrm{d}T} + \frac{n_{eff}}{\Delta L}\frac{\mathrm{d}\Delta L}{\mathrm{d}T} \tag{37}$$

和 $\Delta\phi_0$ 是一常数，γ 和 m 分别表示干涉器的插入损耗和调制深度，对理想器件来说，$\nu = m = 1$。

单个干涉器不能同时获得宽的测量范围和高的分辨率。为了不致造成测量结果混淆，具有单个干涉器的传感器的工作温度范围不能超过$\frac{\lambda}{2}b\Delta L$。因此，增加 ΔL 可以增加灵敏度，但会减小测量范围。还有，在 $\Delta\phi = n\pi(n = 0, \pm1, \pm2, \cdots)$点的附近，单个干涉器的灵敏度很低。

高分辨率和宽量程的温度传感器可由几种不同路程差的干涉器并联而成，如图 50(b)所示，即为由 A、B 和 C 三个干涉器构成的集成光路作为传感器。

另一种基于 Mach-Zehnder 干涉器原理的集成光路可作为测量波前的传感器[85]，其基本元

件是等臂的干涉器，如图 51 所示。为了补偿制作时的偏差，可以在干涉器的一臂上做上电极，加直流偏压进行调制。同时，在两臂的近旁各做一条直波导，假设干涉器两臂中的光功率分别与其近旁的直波导中的功率 P_1 和 P_2 相同，这样干涉器的输出功率 P_ϕ 为

$$P_\phi = \frac{1}{2}\left[(P_1^{1/2} - P_2^{1/2}) + 4P_1^{1/2}P_2^{1/2}\cos^2\left(\frac{\phi}{2}\right)\right] B \tag{38}$$

式中：ϕ 为输入干涉器两臂的光波的相位差；B 是考虑干涉器附加损耗（例如波导的转折）的系数。当两臂的光功率相同时，即 $P_1 = P_2 = P$，式(38)可简化为

$$P_\phi = \left[2P\cos^2\left(\frac{\phi}{2}\right)\right] B \tag{39}$$

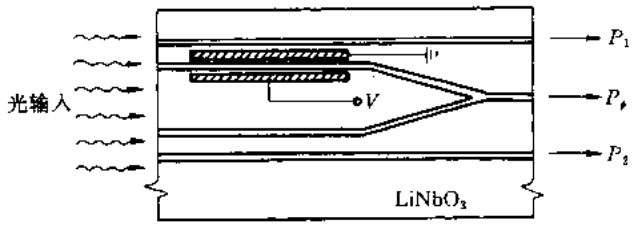


图 51　集成光路波前测量传感器

包括 18 个干涉器的实验器件的测量证明，上述理论关系与实验结果完全相符。现在，由 100 个干涉器组成的波前传感器正在设计制作中，而且器件可以在与波导面垂直的方向扫描，这样传感器就具有二维的测量孔径，波导的输出耦合到 CCD 器件而读出。

利用集成光路 Michelson 干涉器结构可以做成微位移传感器[86]和压力传感器[88]，其结构和原理分别如图 52 和图 53 所示。

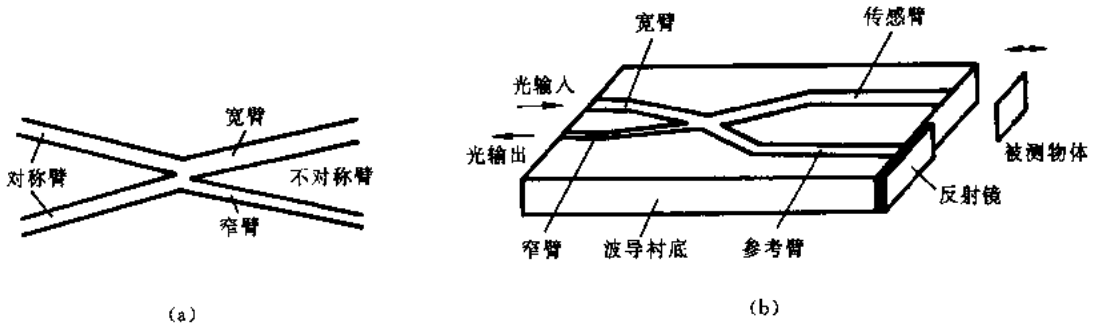


图 52　具有集成光路 Michelson 干涉器的传感器
(a) 基本元件混合耦合器；(b) 微位移传感器

这类传感器的基本元件是光波导混合耦合器，如图 52(a)所示。这是由一波导构成的半反射镜[87]，由四根单模波导连接成的结。左边的结是等臂宽的，而右边的两支具有不等的宽度。由这种混合耦合器构成的微拉移传感器如图 52(b)所示，不对称一边的宽和窄的波导分别作为输入和输出臂，对称一边的一支是传感臂，另一支为参考臂，直接把镜面粘在波导衬底端面上。为了测量物体的微拉移，将一反射镜粘在该物体上并靠近传感波导的端部，使传感的光束由原来的波导返回。进到输入端的光由混合耦合器分成束，它们从两个镜面反射并回到耦合器。通过传感臂的光经过的路程随镜面移动而变，而通过参考臂的光的路程是固定的，所产生的光波位相差在混合耦合器中转换成光强信号。从初步的实验结果得到这种传感器的分

辨率可在 10nm 以下。

用混合耦合器构成的集成光路压力传感器原理如图 53 所示[88]。其结构与上述微位移传感器十分相似。所不同的只是等宽的传感臂和参考臂光波导的端部都直接放上反射镜，而将需测量的压力垂直加载于传感波导直线部分的表面上。

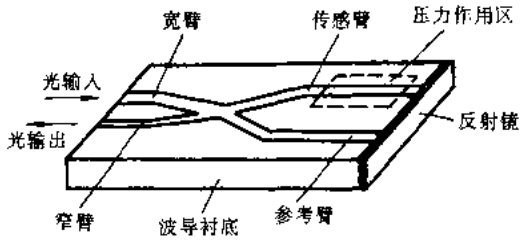


图 53　集成光学压力传感器

参 考 文 献

[1] E.I. Wahashi. IOOC,83 Technical Digest. Japan, Tokyo, 1983,28C1-1

[2] P.E.Rubin, J. R. Stauffer. IOOC, 83 Technical Digest. Japan, Tokyo,1983, 28C1-2

[3] S.S.Cheng, W. B. Gardner, C.J.McGrath. IOOC,83 Technical Digest. Japan, Tokyo,1983, 28C2-3

[4] R.A. Linke, B. L. Kasper,J-S Ko, et al. IOOC,83 Post-Deadline Paper Technical Digest. Japan, Tokyo, 1983, 29C5-1PD

[5] W.T.Tsang, N.A.Olsson, R.A.Logan. Appl. Phys. Lett., 1983,42:650～652

[6] M.M.Boenke, R.E.Wagner, D.J.Will. Elect. Lett., 1982,18:897～898

[7] R.E.Wagner, S.M.Abbott, P.R.Trischitta. IOOC,83 Technical Digest.Japan, Tokyo,1983,29A4-1

[8] R.Mack. Intern. Fiber Opt. and Commun,1981,2:21

[9] K.Aiki, M.Nakamura, J.Umada. Appl. Phys. Lett., 1976,29:506

[10] K.Aiki, M. Nakamura, J.Umada. IEEE J., 1977,QE-13:220

[11] K.Aiki,M. Nakamura, J.Umada. IEEE J., 1977,QE-13:597

[12] H.Okuda, J.Kinoshita, Y.Uematsu. IOOC,83 Technical Digest. Japan, Jokyo,1983,28B1-4

[13] P.K.Tien, J.A. Giordmaine. Bell Leb. Record,1981, Feb.38～45

[14] Takahashi. IEEE J., 1981,QE-17:239

[15] L.A. Coldren, K.Furuya, B.I.Miller, et al.Digest of Topical Meeting on Integrated and Guided-Wave Optics. 1982,WB1

[16] L.A. Coldren,B.I. Miller,J.E. Bowers, et al. IOOC,83 Technical Digest. Japan, Tokyo, 1983, 29B1-2

[17] J.L.Merz, R.A.Logan. J.Appl. Phys., 1976,47:3503～3509

[18] K.Iga, B.I. Miller. IEEE Quantum Electron.,1981,18:22～29

[19] N.Bouadma, J. Riou, J.C.Bouley. Electron. Lett., 1982,18:879～880

[20] L.A. Coldren, K. Iga, B.I.Miller, et al. Appl. Phys. Lett., 1980,37:681～683

[21] F.A.Blum, K.L.Lawley, F.H. Doerbeck, et al. Appl. Phys. Lett., 1974,37:681～683

[22] H.Blauvelt, N.Bar-Chaim, D.Fekete, et al. Technical Digest of Topical Meeting on Integrated and Guided-Wave Optics,USA FC4, 1982

[23] D.Wada, S. Yamakoshi, T.Fujii, et al. Electron. Lett., 1982,18:189～190

[24] A. Sasaki, K. Matsuda, Y. Kimura, et al. IEEE., 1982, ED-29: 1382

[25] H. Beneking, N.Grote, W.Roth, et al. Electron Lett., 1980,16:602～603

[26] A.Sasaki, M.Kuzuhara. Japan J.Appl. Phys., 1981,20:283～286

[27] H. Beneking. IEEE Electron Device Lett., 1981, EDL-2:99 ~ 100
[28] H. Beneking, N. Grote, M.N. Svilans. IEEE., 1981, ED-28:404 ~ 407
[29] C.P. Lee, S. Margalit, A. Yariv. Appl. Phys. Lett., 1977, 31:281 ~ 282
[30] D. Wilt, N. Bar-Chaim, S. Margalit, et al. IEEE J, 1980, OE-16:390 ~ 391
[31] C.P. Lee, S. Margalit, I. Ury, et al. Appl. Phys. Lett., 1978, 32:410 ~ 412
[32] H. Kumabe, T. Tanaka, H. Namizaki, et al. Appl. Phys. Lett., 1978, 33:38 ~ 39
[33] N. Bar-Chaim, J. Katz, I. Ury, et al. Electron Lett., 1981, 17:108 ~ 109
[34] C.P. Lee, S. Margalit, I. Ury, et al. Appl. Phys. Lett., 1978, 32:806 ~ 807
[35] N. Bar-Chaim, I. Ury, and A. Yariv. (to be published)
[36] M. Yust, N. Bar-Chaim, S.H. Izadpanah, et al. Appl. Phys. Lett., 1979, 35:795 ~ 797
[37] T.P. Tanaka, H. Matsueda, Shinya Sasaki, et al. In Proc. of Second European Conference on Integrated Optics. 1983, 21 ~ 23
[38] H.F. Taylor. Electron. Lett., 1974, 10:41 ~ 43
[39] R.V. Schmidt. Electron. Lett., 1976, 12:575 ~ 577
[40] R.A. Becker, W.S.C. Chang. Appl. Opt., 1979, 18(19):3296 ~ 3300
[41] C.S. Tsai, B. Kim, F.R. El - Akkari. IEEE J. Quantum Electron, 1978, QE-14:513 ~ 517
[42] C.F. Chang, C.S. Tsai. Digest of Topical Meeting on Integrated and Guided Wave Optics. 1982, Th D 2
[43] A. Neyer, W. Mevenkamp. Second European Conference on Integrated Optics. 1983, 136 ~ 139
[44] M.C. Hamilton, D.A Wille, W.J. Micelli. Opt. Engin., 1977, 16:475
[45] B. Chen, T.R. Ranganath, T.R. Joseph, et al. Digest of Topical Meeting on Integrated and Guided Wave Optics. 1980, ME3
[46] D. Mergerian, E.C. Malarkey. Microwave J., 1980, 23:37
[47] M. Kanazawa, T. Atsumi, M. Takami, et al. IOOC, 83 Technical Digest, 1983, 30B:
[48] D. Mergerian, E.C. Malarkey, R.P. Pautienus. IOOC, 83 Technical Digest, 1983, 30B:3 ~ 6
[49] T. Suhara, H. Nishihara, J. Koyama. IEEE J. of Quantum Electron, 1982, QE-18: 1057
[50] T. Suhara, H. Nishihara, J. Koyama. IOOC, 83 Post-Deadline Paper Technical Digest. 1983, 29C 5 - 5PD
[51] S. Valette, J. Lizet, P. Mottier, et al. IOOC, 83 Technical Digest, 1983, 30B:3 ~ 7
[52] S. Valette, A. Morque, P. Mottier. Electron. Lett., 1982, 18:13
[53] G. Arvidsson, L. Thylen. IOOC, 83 Technical Digest, 1983, 30B:3 ~ 8
[54] W.S.C. Chang, C.S. Tsai, R.A. Becker, et al. IEEE J. of Quantum Electron, 1977, QE-13:208 ~ 215
[55] K.Y. Liao, C.C. Lee, C.S. Tsai. Digest of Topical Meeting on Integrated and Guided Wave Optics. 1982, WA4
[56] C.S. Tsai, J.K. Wang, K.Y. Liao. SPIE-Symp. on Real-Time Signal Processing II. SPIE, 1979, 180:160 ~ 162
[57] H.F. Taylor. IEEE J. Quantum Electron, 1979, QE-15(4):210 ~ 216
[58] H.F. Taylor, M.J. Taylor, P.W. Bauer. Appl. Phys. Lett., 1978, 32:559 ~ 561
[59] F.J. Leonberger, C.E. Wood ward, R.A. Becker. Digest of Topical Meeting on Integrated and Guided Wave Oplics. 1982, WA3
[60] V. Ramas Wamy, M.D. Divino, R.V. Standley. Appl. Phys. Lett., 1978, 32:644 ~ 646
[61] S. Yamada, M. Minabata, J. Noda. Electro. Lett., 1981, 17:259 ~ 260
[62] P.W. Smith, I.P. Kaminow, P.J. Maloney, et al. Appl. Phys. Lett., 1978, 33:24
[63] C.L. Chang, C.S. Tsai. Appl. Phys. Lett., 1983, 43:22 ~ 24
[64] H.F. Taylor. Appl. Opt., 1978, 17(10):1493 ~ 1498
[65] L. Goldberg, S. H. Lee. Appl. Opt., 1979, 18, (12):2045 ~ 2051
[66] P.W. Smith, E.H. Turner. Appl. Phys. Lett., 1977, 30:280
[67] P.S. Cross, R.V. Schmidt, R.L. Thornton. Digest of Topical Meeting on Integrated and Guided Wave Optics. 1978, Tu B1

[68] H.A. Haus, E.P. Ippen, A. Lettes, et al. Appl. Phys., 1982, B-28:283
[69] T.G. Giaiorenzi, J.A. Bucaro, A. Dandridge, et al. J. IEEE, 1982, QE-18(4): 626 ~ 665
[70] J.A. Bucaro, H.D. Dardy, E. Carome. J. Acoust. Soc. Amer., 1977, 62:1302 ~ 1304
[71] J.H. Cole, R.L. Johnson, P.B. Bhuta. J. Acoust. Soc. Amer., 1977,62:1136 ~ 1138
[72] J.A. Bucaro, T.R. Hickman. Appl. Opt., 1979,18:938 ~ 940
[73] J.H. Cole, J.A. Bucaro. J. Acoust. Soc. Amer., 1980,67:2108 ~ 2109
[74] A.M. Smith. Appl. Opt., 1978,17:52 ~ 56
[75] A. Yariv H. Winsor. Opt. Lett., 1980,5:87 ~ 89
[76] A. Dandridge, A.B. Tveten, G.H. Sigel, et al. Electro. Lett., 1980,16:408 ~ 409
[77] S.C. Rashleigh. Opt. Lett., 1981,6:19 ~ 21
[78] V. Vali, R.W. Shorthill. Appl. Opt., 1976,15:1099 ~ 1100
[79] S.C. Lin, T.G. Giallorenzi. Appl. Opt., 1979,18:915 ~ 931
[80] W.K. Burns, R.P. Moeller, C.A. Villarruel, et al. IOOC.83 Technical Digest.1983,28C:3
[81] C.S. Wang, et al. Appl. Phys. Lett., 1982,41:587
[82] A. Dandridge, R.O. Miles, T.G. Giallorenzi. Electron. Lett., 1980,16:948 ~ 949
[83] A. Dandridge, R.O. Miles, A.B. Tveten, et al. Proc. of First European Conf. Integrated Optics.1981
[84] L.M. Johnson, F.J. Leonberger. Appl. Phys. Lett., 1982,41(2):134 ~ 136
[85] R.H. Rediker, T.A. Lind, F.J. Leonberger. Appl. Phys. Lett., 1983,42(8):647 ~ 649
[86] M. Izutsu, A. Enokihara, T. Sueta. Electron. Lett., 1982,(20):867 ~ 868
[87] M. Izutsu, A. Enokihara, T. Sueta. Optics Lett., 1982,7(11):549 ~ 551
[88] M. Izutsu, A. Enokihara, N. Mekada, et al. Proc. of Second European Conference on Integrated Optics. 1983, 144 ~ 147

光计算导论

由于用光子作为信息的载体具有一系列电子所不能比拟的优点，光计算在本世纪 60 年代初就成为一个有着巨大吸引力的课题。长期以来对光计算的研究更多的是看作一项潜在技术。可是在今天，越来越多的人认识到，光计算将光子技术用于计算已成为实际应用要求所推动的新技术了，最终它将显示出在计算速度、数据容量、实时性、人工智能等许多性能方面胜过电子计算。

1 电子计算机的进展及极限

1.1 电子计算机的演变[1,2]

自 1946 年第一台电子管计算机问世以来，电子计算机在不到半个世纪内已几经演变，当今人们正在为未来的第五代计算机而竞争。由于新器件的发明对电子计算机功能的提高起着决定性作用，通常就以此作为电子计算机历代划分的依据，即由电子管构成的为第一代；晶体管为第二代；采用中、小规模集成电路（SSI 和 MSI）的为第三代；发展到由大规模集成（LSI）和超大规模集成（VSLI）电路构成的为第四代。未来的第五代电子计算机除了采用集成度更高的 VSLI 或其他新器件以外，更重要的标志在于其智能化程度的大大提高。

电子计算机功能的提高除了上述器件因素外，计算机系统结构（computer architecture）的改进也是十分重要的，计算机系统由硬件（hardware）和软件（software）组成，硬件是计算机系统中的实体装置，它主要包括中央处理器、存储器、输入输出、外部设备及它们之间的通信互连装置等。软件一般指的是一组程序（如编译程序、监督程序、控制程序和应用程序等）及与操作有关的各种信息及使用维护手册、说明书及框图等。软件是计算机不可缺少的组成部分，它的作用是便于计算机的操作、提高效率和扩大硬件的功能。计算机的系统结构是包括硬件和软件在内的计算机系统整体的总称，它可与描述具有一定式样和风格的建筑物总称（architecture）相比拟。实际上，系统结构这个名词在计算机术语中是 70 年代从建筑学中移植过来而得到广泛使用的。

从应用的角度或从性能和价格的观点出发，可以对电子计算机进行分型。目前，电子计算机通常可分成巨型机、大型机、中型机、小型机和微型机等类型。分型的标准不能按体积的大小，也不能根据某一种特性参数的高低来确定。而且随着技术的进步，分型的标准也随之提高。过去需要占几个房间的机器的功能还不如现在一个机柜的小型机，15 ~ 20 年前的所谓大型机其主频和运算速度还比不上现有的小型机，甚至微型机。

一般来说，计算机应根据同一时期综合性能的高低来分型。所谓综合性能，在硬件方面包括计算速度、字长、数据的类型（目前微型机只有定点表示，小型机已增加浮点表示，而大型机、巨型机则不仅有定、浮点表示，还有向量、矩阵表示）、存储系统及容量、输入/输出能力（包括 I/O 处理器的处理能力和能连接的 I/O 设备数量）以及指令系统等。在软件方面，巨型机和大

型机比小型机和微型机配备更多种高级语言，更完善的操作系统、数据库、知识库网络通信软件和更多的用户程序包。大型机还具备其他许多功能，如纠错编码和诊断技术等各种提高计算机的可靠性所采用的措施。下面，我们以电子计算机的两极微型机和巨型机为例，就可清楚地看到电子计算技术的发展是何等迅速。

大家知道，1946 年诞生的第一台电子计算机 ENIAC（Electronic Numerical Integrator and Compter 的缩写，意即电子数值积分器与计算机）是一个庞然大物，共用 18×10^3 个真空管，质量近 30t，电力消耗 150kW，占地约 $170m^2$。其加法运算的速度为每次 0.2ms，乘法运算为每次 0.8ms。而 70 年代出现的微型机，其全套电路集成在一片或几片像小孩指甲大小的硅片上，一台完整的微型机与打字机差不多，功耗仅数瓦，而运算速度却比 ENIAC 高出几十倍或几百倍。微型机一出现就引起极大重视，掀起了计算机大普及的热潮，在短短的十多年内获得了巨大发展，并经历了四代技术更新，如表 1 所示。

表 1　微型机的技术更新

年　代	工艺技术	字　长	指令周期/ms	时钟/MHz	集成度 器件/芯片	产　品
1971～1973	PMOS	4～8	20	0.7～0.8	2k	Intel 4004 Intel 8008
1974～1977	NMOS	8	2	2～5	5～10k	Intel 8080 M6800、Z80
1978～1980	HMOS	16	0.5	5～10	30k	Intel 8086 M 68000 Z 8000
1981～	HMOS CMOS	32	0.3	8～18	100～500k	Intel 80386 Z 80000 HP－32

巨型机的研制是现代科学技术，特别是国防尖端技术和高技术发展的需要。核武器设计、星球大战系统、空间技术、气体动力学、长期天气预报、石油勘探、粒子束模拟计算、实时图像识别、机器人视觉和人工智能等许多问题都要求计算机不仅有很高速度，并且有很大容量。现有的一般大型通用计算机已远远不能满足要求。下面列举两例说明之。

以往研制飞机和航天器都要进行大量的风洞试验以采集设计数据，例如对航天飞机需要大约 45×10^3h 的风洞试验，十分费时、费钱。现在利用计算机模拟技术进行空气动力学计算，对整个航天飞机作三维形状的模拟，这类计算要求计算机具有 10BFLOPS（Billion Floating Logic Operation Per Second）的运算速度，即 10^9 次浮点逻辑运算/s。

在长期天气预报中，如果在经度方向取 100 点，纬度方向取 50 点，高度方向取 10 点，即总共取 5×10^4 个空间位置，计算时间间隔取 5min，这样要想得到 1 年后的气象数据，需进行 10^5 次计算。当要求这种计算能在若干小时内完成时，计算机的运算速度必须超过 100 BFLOPS。

随着科学技术的进步所提出的许多新的计算问题，要求计算的速度比上面所举的例子更高。目前研究开发的巨型机或超级计算机的运算速度已超过每秒十亿次，但离开实际的需要还有很大差距。

1.2 诺依曼机的基本结构[1,2,9]

最早的电子计算机 ENIAC 与以前的所有计算器械相比，最大的特点是采用了电子线路来执行算术运算、逻辑运算和数据存储。机器采用十进位制，它有 20 个加法器，每个加法器由十组环形计数器组成，可以保存一个字长十位的十进制数。它还有乘法器以及除法和开方装置供执行其他运算。这台计算机是在第二次世界大战期间，美国为了能迅速算出炮弹弹道轨迹而研制的。

虽然这是世界上第一台能实际运行的大型电子计算机，但它的基本结构和以往的机电式计算机没有本质差别。ENIAC 只是初步显示了电子元件在提高运算速度方面具有的优越性，却尚未最大限度地实现采用电子技术可能提供的巨大潜力。存在的主要缺陷：一是存储容量太小，至多只能储存 20 个字长十位的十进制数；另是它的程序是用线路连接方式实现的，为了进行几分钟的数字计算，程序准备工作要化几小时甚至 1～2 天。要想简单地采用增加存储器的办法来克服存储容量的不足，不仅不经济，而且在当时也不现实，因为 ENIAC 的存储器储存 1 个字长十位的十进制数需用 1 个十位环形计数器，而这类计数器使用的真空管达 600 个之多。至于如何简化复杂的开关程序操作，也只有改变计算机的基本结构才能真正解决。

著名的数学家冯·诺依曼(Von Neumann)及其合作者(大部分是 ENIAC 的研究组成员)1945 年提出了 EDVAC (Electronic Discrete Variable Automatic Computer，意为离散变量自动电子计算机)。这个方案有两个意义重大的改进：一是为了充分发挥电子元件的高速度而采用了二进制；二是提出了“存储程序”的概念，即能够自动地从一个程序指令进到下一个程序指令，其操作顺序可以通过一种称为“条件转移”的指令而自动完成。

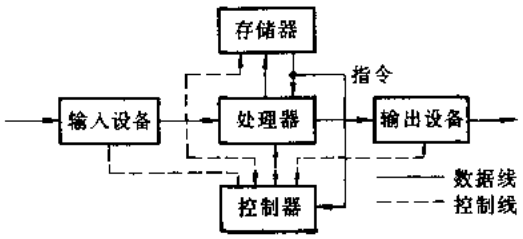


图 1　诺依曼计算机系统结构

这个概念被誉为计算机史上的一个里程碑。后来，人们就把这种基本结构称为诺依曼机。这种机器由五部分组成(见图 1)，包括处理器、控制器、存储器、输入装置和输出装置。诺依曼机的主要特点是：

(1) 计算机用二进码表示，采取二进制运算。

(2) 系统结构以处理器为中心，输入/输出设备与存储器之间的数据传输都通过处理器；处理器、存储器和输入/输出设备的操作及其通信都由控制器集中控制。

(3) 存储器是顺序线性编址的一维结构，操作时按坐标对每一个编址单位寻址。其运算速度与访问主存的次数有密切关系。

(4) 指令由操作码和地址码组成。操作码代表指令的操作类型，最基本的操作是算术运算，地址码代表操作数的地址。指令在存储器中是按执行顺序而存储，由指令计数规定每条指令所在单元的地址。每执行完一条指令，指令计数器自动顺序加 1，操作只能是顺序的。

由于诺依曼机中引进了对存储单元按地址寻址的机制，从而使处理器与存储器之间的数目大大减少。在这以前的计算机包括 ENIAC 统称经典有限态机(classical finite state machine)，其结构原理如图 2(a)所示，由一个组合逻辑单元 L、若干个存储单元 M 以及输入/输出(I/O)和各种互连器件所构成。所有的存储单元都平行地操作，不需要寻址，但当需要增加大量存储单元时，在逻辑单元和存储单元之间完成约 N 个互连已不现实，正如前面所指出的，需要耗用大

量电子元件。在诺依曼机(见图 2(b))中,由于采用了寻址机制 A 和二进制编码方式,使从逻辑单元到存储器的输出端从 N 个互连减少到 $\log_2 N$ 个,同时还利用公共回线大大减少了存储器和逻辑单元输入端的互连。这一结构的改进对电子计算机的发展有着极为重要的意义,虽然 30 多年来诺依曼机已有了许多发展和改进,但至今绝大多数电子计算机仍是建立在这个基本结构特点上的。

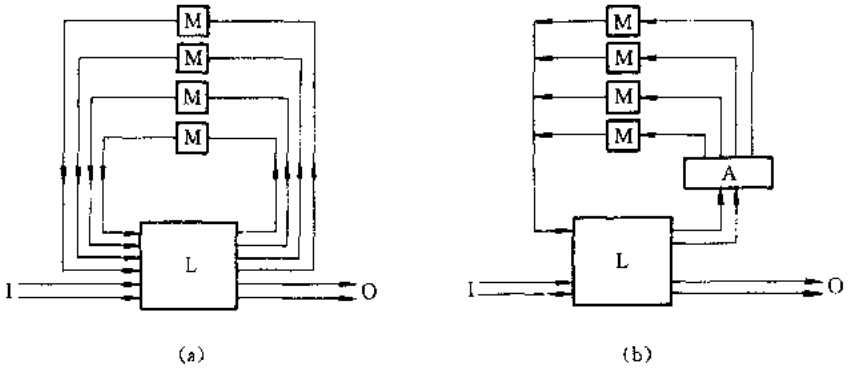


图 2 经典有限态机(a)与诺依曼机(b)的比较

1.3 计算机平行结构的发展[2~4]

富有戏剧性的事件出现了,曾对电子计算机的发展有巨大推动作用的诺依曼结构后来却又成为它进一步提高速度和扩大容量的主要障碍,被称为诺依曼"瓶颈"(bottleneck)。从图 2(b)可以看出,由于诺依曼机的操作要通过顺序寻址,在一个时间内只能对一个数据按一个指令进行操作,即只能串行操作。而经典有限态机由于不存在寻址机制,在处理器与存储器之间是直接互连的,可以同时进行运算,或称为并行操作。显然,比较一下 N 位串行计算机(serial computer)和 N 位并行计算机(parallel computer)的运算速度,不难看到,在元件具有同样性能的条件下,后者的速度差不多可以提高 N 倍。

随着科学技术的迅速发展,每秒运算数百万次的通用大型机仍远远不能满足实际的需要。突破传统的顺序处理的计算机应运而生,它采用了并行结构,也称为并行处理机。这种计算机的速度可以突破每秒十亿次计算,其他功能也大大增强,故列为巨型机。巨型机的成功,其物质基础主要归因于 VLSI 的出现。在一个不到 $1cm^2$ 的硅单晶芯片上可以集成数十万个晶体管、二极管、电阻和电容元件,从计算机发展的阶段而言,属于第四代计算机。前面介绍的 ENIAC 机实际上是可看作现代并行机的前身。

现代并行处理机中应用的并行性(parallelism)概念有多重含义。主要有两种实现途径:较直观的是设备重复或资源重复(resoure-replication)的概念,即采用大量相同的设备同时运算来提高速度;另一种是采用时间交错(time-interleaving)或设备共用或资源共享(resource sharing)的方法使设备在时间上充分利用,使速度和效率增加。

早在本世纪 20 年代,就有人为了解决气象预报中大量的数值计算,想让 6.4 万人坐在一个圆形剧场中同时进行计算,体现了朴素的并行计算思想。40 年代,美国原子能委员会曾组织了几千个台式计算机操作员同时进行核反应计算。这种利用大量重复设备的并行计算方案由于硬件价格昂贵的限制,在早期的电子计算机系统结构中不可能普遍采用。随着 VLSI 技术

的成熟和成本大幅度下降，设备重复的并行处理机，例如多存储单元的关联处理机(associative processor)、多操作单元的阵列处理机(array processor)等，近年来已获得迅速发展。

时间交错的概念是将多个处理过程在时间上互相错开，以便交替使用同一套硬件设备，以增加设备利用率来提高计算速度。流水线处理机(pipeline processor)就是这一类。与现代工厂中流水线生产的基本思想相类似，在这种处理机中将复杂的运算分解为基本的子运算。被运算的数或数对依次通过子运算单元，最后完成复杂的运算。由于各子运算单元可以同时操作，使运算速度成倍提高。例如，执行一条指令一般可粗略地分为取指令、指令译码、取操作数和执行指令四步。在非流水线的顺序处理计算机中，执行的时序如图 3(a)所示。在流水线处理机中指令执行过程如图 3(b)所示。图中 I 表示指令，右上角数字为指令序号，右下角的数字为指令执行的序号，在流水线处理机中也就是流水段序号。由图中不难看到两者的区别。在顺序处理机中必须等第一指令全部执行完成后才能执行第二指令，在流水线处理机中只要第一指令在第一流水段执行完成进入第二流水段时，第二指令就可以进入第一流水段，这样依次类推，每隔一个流水段时间 τ，就能输出一条指令结果，所以流水线分得越多，处理的速度也就越快。由于流水线上的各单元没有设备的重复，只是通过时间上的交错使设备得到了充分利用，所以这种系统有较高的性能价格比。但这种系统的平行性和速度提高是受限制的，不少计算机因而采用多重流水线，即不仅采用时间交错，也同时采用设备重复。实际上，有许多并行处理机中都综合利用了多种并行机制。流水线处理机中比较著名的例子如 Cray-1，1976 年提供产品，机器字长 64 位，最高能得到 100×10^6 次浮点运算/s。

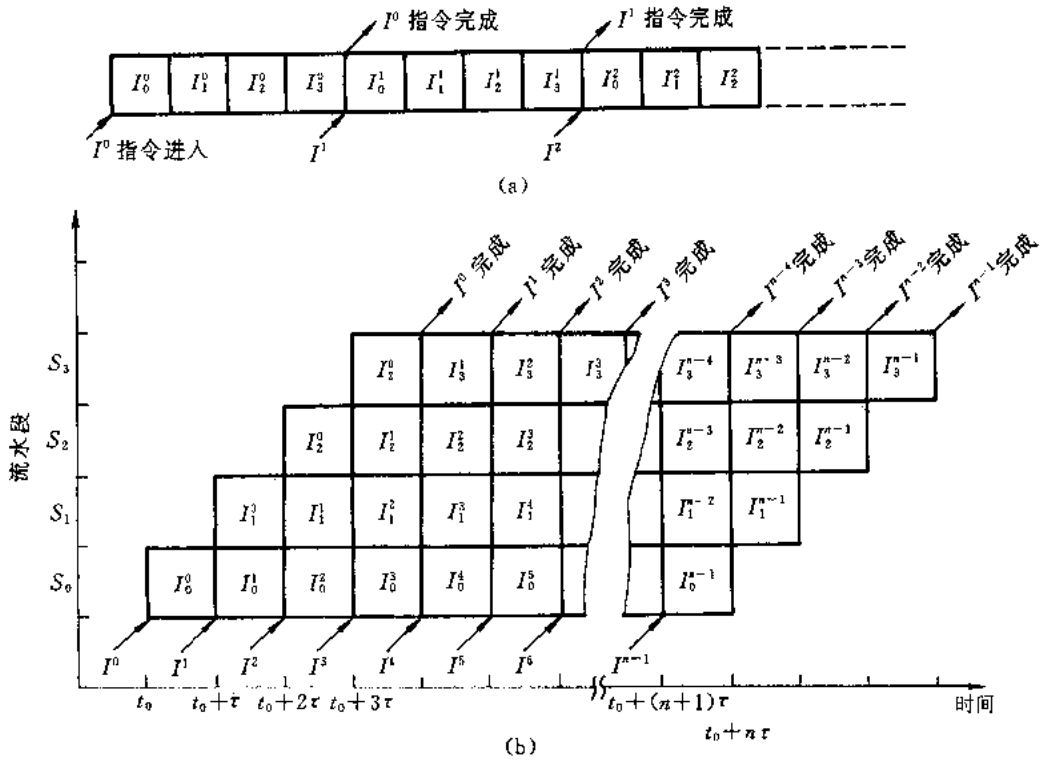


图 3　顺序处理指令执行过程(a)与流水线处理过程(b)的比较

下面我们讨论并行处理机的分类。从计算机的系统结构来看，可按在其不同层次引入并行性来分类。例如，上面提到的多存储单元的关联处理机是存储器操作这一级的并行处理。

一般来说，存储器的寻址是按照给定的几何坐标地址读出对应存储单元的数码，这就是按坐标寻址。有一种内部包括信息处理功能的存储器，它可按给定的信息内容的全部或部分特征把所有存储单元中内容与此特征相符合的全部数码一次判别出来。这种内容寻址存储器也称关联存储器(associative memory)。以关联存储器为核心，配备必要的运算部件、指令存储器、控制器和输入/输出接口就构成一台以存储器操作并行为特征的关联处理机，其结构框图如图 4 所示。它与普通顺序处理机结构的主要不同在于用关联存储器替代了坐标寻址存储器。由于每个存储单元不仅可存储信息，而且还有处理信息的功能。来自控制器的一条指令能对许多数据同时执行逻辑运算，完成并行操作。尽管关联存储器的信息处理速度可能比顺序操作的中央处理器的速度低，但由于它是大量重复设置而并行工作的，故其运算总的速度可以随关联存储器的并行数增加而提高。美国两种军用的巨型计算机系统 STARAN 和 PEPE 都是属于关联处理机一类，建造于 70 年代初期。

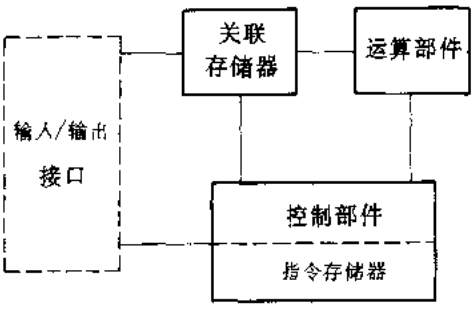


图 4　关联处理机

前面介绍的流水线处理机则是在处理器操作步骤这一层次上引入并行机制。更高层次的并行还有处理器级的并行，例如列阵处理机和多处理机系统等。

对并行处理计算机另一种较常用的分类方法是弗林(M.J.Flynn)分类法，即将数据流和指令流的并行性作为分类依据，关键在于系统能否实现指令级并行。数据流(data stream)表示待处理的数据组，包括输入数据和运算的中间结果；指令流(instruction stream)是计算机执行的指令组。Flynn 的分类方案是按数据流(D)和指令流(I)的并行性将计算机系统结构分为下列两类：

第一类单指令流单数据流系统(SISD)实际上就是传统的顺序处理计算机，其系统结构如图 5(a)所示。这类系统由于瓶颈效应使速度受到限制。至于多指令流单数据流(MISD)系统，即指令级并行，而数据级是串行的，可以把单一流水线处理机列为这一类，图 5(c)表示其系统结构。

第二类为已开发的并行处理机，多数是属于 SIMD 和 MIMD 结构，其系统结构如图 5(b)，(d)所示，因此它们能更高程度地发挥并行结构的优越性。上面讨论的关联处理机是更高一层次的 SIMD 系统。其中将处理单元排成一行或列阵并无原则区别，由指令存储器提供的一条指令，由指令单元同时送到各处理单元。各处理单元可以从数据存储器取不同的数据。列阵处理机的典型代表是 60 年代末美国依里诺大学设计的 ILLIAC Ⅳ 系统，它的 64 个处理单元在平面上排列成 8×8 的正方形列阵。

在 MIMD 结构中，主要包括多处理机(multiprocessor)和分布处理系统(distributed processing system)。多处理机一般可分为两类：一类是由相同类型的多个处理机组成的单一处理机系统，这称为同型(homogeneous)多处理机；另一类是由不同类型的多个处理机组成的计算机系统，称为异型(heterogeneous)多处理机。

在同型多处理机中，按照 MIMD 模式实现程序一级或任务一级的并行处理，将一道程序分解为若干相互独立的程序段或任务，分配到各处理机同时并行执行。这种并行任务在处理机

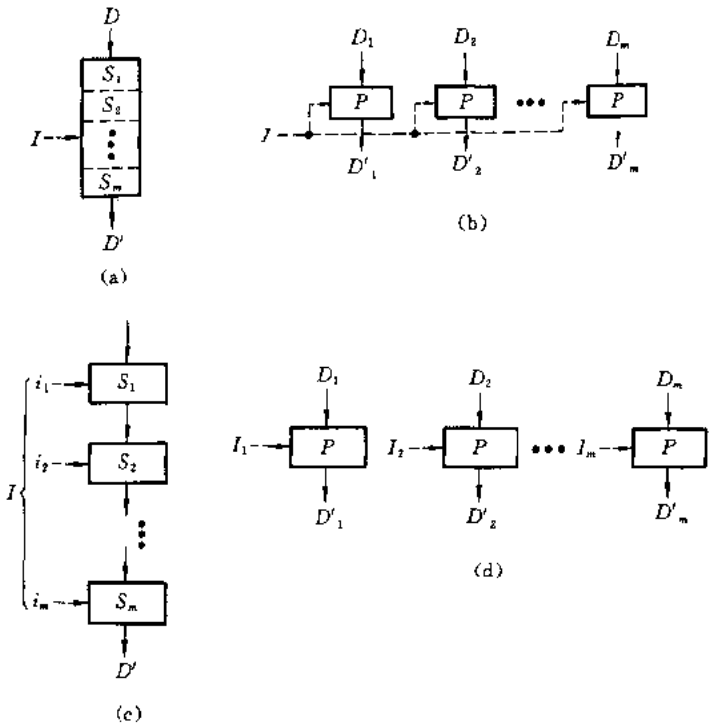


图 5　四种不同的系统结构

(a) 简单的 m 级顺序处理机，SISD 结构；
(b) 单指令流的 m 平行处理机，SIMD 结构；
(c) m 段流水线处理机，MISD 结构；
(d) 多指令流的 m 平行处理机，MIMD 结构

之间可随机进行调度，因此要求各处理机应具有同等功能。这类处理机是建立在设备重复原理的基础上，以极大提高处理速度为目标。

在异型多处理机中，它划分程序操作的原则与同型机明显不一样，不是分解为并行成分，而是依靠流水线原理来提高全系统的处理效率，分配给不同的处理机的是性质不同的任务，即各处理机是专业化的，因而要求各处理机具有不同的功能。例如 1979 年由美国 Burroughs 公司和依里诺大学联合开发成功的科学处理机 BSP，就是由专门完成数组运算、标量运算和系统管理等功能的三台处理机共同组成的系统，它的最高处理速度可达 5×10^7 次浮点运算/s。

分布处理系统则是更大范围的并行处理，但至今没有一个比较统一的定义。一般来说，这种系统具有如下共同特点：(1)具有多套设备，不是集中而是分布的；(2)各设备具有相对独立性，但由通信网络相互作用；(3)具有统一的操作系统。这种系统既具备多处理机的特点，又与计算机网的特征很相似。这种系统的优点主要是可以实现全系统的设备共用，包括程序和数据的共用，因而具有较高的性能价格化，同时系统的可靠性比单机系统高。近年来这种系统发展得很快。图 6 表示了并行处理机的运算速度随年代提高的情况。

近年来国内外正在研制的新一代计算机，目的在于处理知识信息的智能计算机，即更接近于人脑功能的计算机。它有以下几个特点：能高效率地支持非数值操作；能使用自然语言进行人机交互；能进行复杂的人工智能问题求解；能实行高速并行处理。

智能计算机主要由三个部分组成：自然语言理解、语音、图像和图形输入输出等的应用系

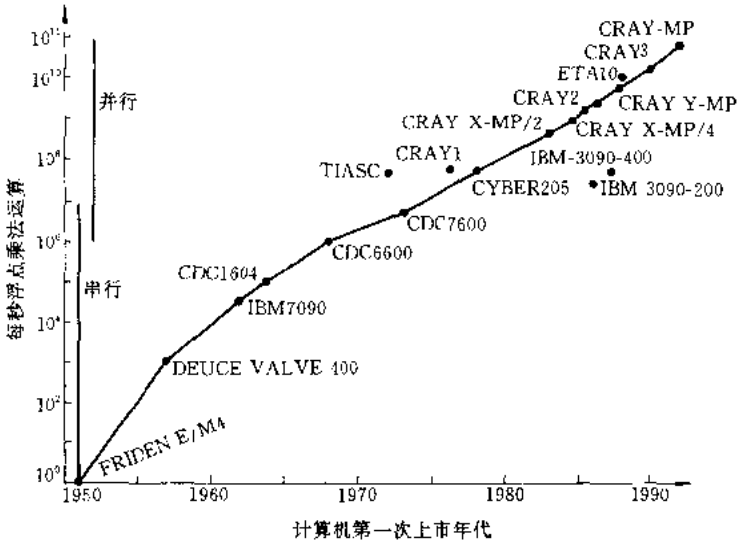


图 6　计算机速度的进展

统;知识库、推理和智能接口三个子系统等的软件系统;知识库、推理机和智能接口机等的硬件系统。

智能计算机的一个特点是容量大,为了有效地支持知识库的工作,存储器的容量应在 100 ~ 1000 GB。它的另一特点是运算速度快,在(10 ~ 1000) × 10^9 次/s。处理这样大的存储器容量和进行这样高速的运算不借助于并行处理是不可能实现的。

近几年来,人工智能技术已有了较大的进展。英国通信公司研究出声音应答自动翻译系统,它能翻译英、法、德和西班牙四种语言。专家系统已进入实用阶段,机器人的手脚已能灵活地自由活动,医疗界正在大力开发计算机诊断系统,模仿人脑的神经网络计算机也正在研究中。

1.4　电子计算面临的新困境及其限度[5,8]

由于 VLSI 技术的高度发展和并行处理系统结构的采用,使电子计算机摆脱了诺依曼瓶颈的困境,而又获得了新的发展。但由于电子电路的固有局限性使 VLSI 和并行性的潜力不能充分发挥,这主要是互连和通信的障碍。

任何连接导线都不可避免存在一定的电阻 R 和电容 C,因而信号通过导线的时间常数 RC 往往超过了晶体管的开关时间,这就大大限制了互连带宽的提高并造成了时钟扭歪(clock skew)。这互连通信问题不仅是各功能单元之间的连接,即使 VLSI 芯片内部的连接以及芯片之间的连接都成了难于逾越的障碍。

采用 VLSI 技术虽有很大优点,根据现有的技术,可以在单个硅片上制造出具有 10^{10} 个有源器件的系统,每个器件的开关速度可达 10^{-10} s。但遗憾的是这一惊人的容量不能与其通信带宽相匹配,连接导线拥塞,没有足够空间来排列。在 VLSI 芯片中,随着元件几何尺寸按比例缩小,导线的长度缩短了 α,而其截面减小了 α^2,因而导线的电阻增加了 α,电容减小了 α,结果是导线的时间常数 RC 乘积并没有变化。根据 VLSI 的参数可以估算出,信号在其中传输的速度只有光速的 0.5% 左右。

在电子计算机中,由于在不同通道中传输的信号存在不同的时延,因而不能同时到达同一

个逻辑门,或指令不能同时到达各操作单元,这就是通常所谓的时钟扭歪。当这种时钟扭歪超过一定范围时,运算和操作就产生误码。为了使电子计算机系统的各种层次互连的时钟扭歪能保持在一定的允许范围内,需要采取措施使不同长度的互连通过某种附加的延迟加以补偿。例如 CRAY-1 处理机,在电路板上最长的互连长度为 15.42cm,短于这个长度的互连必须通过门的延迟加以补偿,以使传输时间与互连最长的相同,使处理器的设计大大复杂化了。系统一级的互连,为保持信号上升前沿和下降前沿所需要的梯度及通信带宽,往往需要采用体积大、成本高的同轴电缆系统。

更重要的是在许多应用场合,如图像识别、人工智能、机器人视觉等,不仅要求运算速度高,并且要能同时对大量数据进行并行处理,有时要求处理的信息本身就是一个二维或三维巨大的数据流。在这种情况下,往往要 10^6 数量级的并行通道同时操作,在电子计算机系统中要达到如此高度并行的互连,实际上是难以实现的。

2 光子计算的特点

2.1 光子与电子计算的物理特性比较[5,6,8,14]

用光子作为信息的载体与电子相比,在传输方面带来的优点已在光纤通信上充分显示。今后,毋庸置疑,电缆通信即使不是全部也将是绝大部分被光纤通信所取代。把光子用于信息处理和计算方面的潜在作用在某种意义上说将比光子用于通信更为杰出。这是由于光子与电子的物理本质及特性所决定的。

首先,光子不像电子那样带有电荷。电子之间通过电磁场而相互作用,导致了电子信号很容易自身干扰或受外界干扰。光子之间很难相互作用,因此,光信号可以沿各自的通道传播,不论其通道互相平行或互相交叉都不会产生干扰,这就造成光学的固有并行性,这种特性对信息的并行处理和并行计算恰好配合,是电子学无法比拟的。

其次,光子不具静质量,它可以在真空中传播,也可以在介质中传播,并且很容易通过真空和介质的界面,不仅以光速传播,而且传播过程中能量损耗极小。电子一般限制在金属导线内传输。上面已经提到,由于存在电阻 R 和电容 C 的限制,电流的传输速度往往只有光速的千分之几,而且它的传输带宽也受到 RC 的限制。光具有很高频率,并且传输带宽不会有类似 RC 这种弛豫过程的限制。光学系统的空间带宽和时间带宽积很大,能够容纳大量独立的信道完成所需的各种操作。这一特性与光子互相不作用的特性结合起来解决计算系统中各级的互连和通信也将大大胜过电子学。

最后,光子很难相互作用既是很大优点,但同时也是光子的一个严重缺点,因为很难用光子来控制光子。电子由于容易相互影响,所以可以很方便地实现电子开关功能。自从发现某些材料具有很大的光学非线性系数后,情况有了根本改观。过去认为难于实现光开关和光逻辑操作的问题不再是光计算的壁垒。利用光子与某些材料相互作用所具有的三次非线性效应所设计的光双稳器件,其开关时间已低达纳秒量级,预期的理论极限在 $10^{-12} \sim 10^{-13}$ 范围,即为当今硅开关器件速度的 1000 倍。每个光脉冲开关能量现在能做到 <4pJ,理论值预期可低于 1fJ,相当于数千光子的能量。操作的光功率在 mW 的水平。这说明过去大家担心的光器件的能量消耗现在已能与最好的电子器件的能耗相接近。这样,利用光的高并行性($>10^6$)及快

响应($<10^{-9}$s),未来全光计算机的速度可以超过 10^{15}bit/s 的运算,这将是研究中的巨型计算机 CRAY-3 运算速度的 10000 倍。

2.2 光并行处理的潜力[7~9,14]

从上述电子计算机发展的过程已清楚地看到并行处理对提高计算机功能的重要作用。近年来,人们正在逐步揭开人脑神经系统之谜,它可看作是天然的并行结构的多处理器系统。下面我们不妨在大脑与并行处理机之间作一个有趣的类比,可以进一步看到高度并行对提高计算机功能是何等重要。大家很清楚,现阶段的电子计算机在实现智能操作,诸如推理和图像识别等方面所表现的能力是无法与人脑相比的。但实际上,就电子器件的开关速度而言,比组成人脑的开关元件的速度至少快 10^6 倍。作为生物开关的神经细胞,其响应时间一般在毫秒量级或更长。所不同的是生物开关具有高度并行性和高度互连。据现在研究的结果可知,一个神经网络大约有 10^{11}个神经细胞,每个神经细胞能有高达 10000 个突触(synapse),即生物连接器。而今天的电子计算机中的电子开关一般只与其他少数几个开关相连接,要像神经细胞那样众多的互连,在技术上是无法实现的。这充分说明,人脑高度的处理能力是由神经细胞的高度并行和高度互连所决定的,弥补了开关速度之不足。

未来的计算机能否达到如此的巨量并行性(massive parallelism)和高度互连呢?在光学上,一个好的透镜、棱镜或反射镜所传播的图像能够分解出 $10^7 \sim 10^8$ 个可分辨的光斑(spot),如果采用透镜列阵,几乎可以没有限制。在光处理或光计算系统中,每一个光斑就表示一个带宽极高的通道,这对二维和三维数据流的并行处理是非常有效的。当并行性下降到 10^4 个通道以下时,电子学就能与光学相竞争,而且常常可以优于光学。

根据并行处理机系统中每个处理单元,也称结点(node)的数量以及每个结点的复杂度(complexity)的高低,可以把并行处理机分成微型机列阵(microcomputer arrays)、计算列阵(computational array)、逻辑增强存储器(logic enhanced memories)以及人工神经网 ANS(artificial neural systems)等不同的层次,如图 7 所示。图中将普通单一的顺序处理机作为结点相对复杂度的参考坐标。当单处理器向多处理器发展时,其趋势是随着结点数的增加,每个处理的复杂度减小,直到另一个参考坐标随机存取存储器(RAM),它本身没有处理功能,复杂度最低,但可以达到最多的结点数。

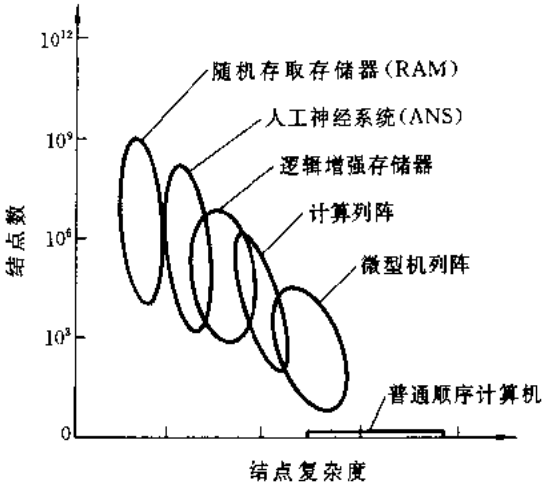


图 7 并行处理机按结点复杂度的分类

微型机列阵是由通信网连接起来的若干台计算机的并行系统。这种系统是一种松耦合,即个别计算机并没有公用的存储器和输入输出装置。这种松耦合的列阵已被用来处理一般能分解的问题,例如供应、后勤、人员等的管理,每台计算机具有自己所要求的功能,可并行操作,但需要保持各种计算过程之间的某种通信。

计算列阵系统中的每个处理单元具有的复杂度大致相当于浮点运算的功能。它通常设计用来完成如下任务:快速傅里叶变换(FFT)、图像识别中的特征提取以及计算机绘图的显示处

理等。脉动列阵(systolic arrays)处理机就是这一类的主要代表。该系统包含大量处理器,但每个处理器只有很少存储器。在这种处理机中,信息的数据流以流水线形式像心脏有节奏地间隙脉动那样通过高度规则排列的二维处理器列阵,脉动列阵因此而得名。操作的顺序可以建立在每个节点中,或以单指令多数据流(SIMD)的形式将指令传播到列阵中每个结点。

其他两类并行处理机包括逻辑增强存储器和人工神经系统,它们的结点除了能作信息存储外,还有不同程度的处理功能,以避免信息在分离的存储器和处理器之间传播所产生的延迟。对逻辑增强存储器来说,每个结点或处理单元(PE)可以包含数千比特的存储容量,一组寄存器以及一些关联逻辑,能按存储的内容操作以及能与其他单元直接通信。前面提到的关联处理机就属于这一类。

人工神经系统处理机有大量高度互连的非线性器件所构成,这种非线性处理单元(结点)相当于大脑的神经细胞。ANS 可用一组微分或差分方程作为其特征,通过对初始的或连续的输入产生其状态响应来处理信息。这是根据最简单的大脑模型设计的,与普通的计算机的工作原理有很大区别。例如在 ANS 中信息不是采用一定的二进制地址存储在某一空间位置,而是通过一定的互连图形(interconnection pattern)来存储。

逻辑增强存储器以及人工神经系统都是属于紧耦合和细粒度(fine-grained)的多处理机系统。由于它们要求巨量并行性的系统结构,这与光固有的并行性相匹配。因而近年来已成为光计算的重要研究方向。

建议中的一种有高度并行处理能力的全光多处理机的系统结构如图 8 所示。其输入是一个可独立寻址的激光二极管列阵或是一个二维空间光调制器。二极管列阵能达到极高的调制速度,但其驱动电路十分复杂。这些列阵都可能提供 10^6 个以上的并行通道。从光源发出的光束首先到达具有同样多通道的门列阵上。用作门列阵的器件在目前至少有两种:一是具有非线性响应的二维空间光调制器;另一是双稳光开关列阵。后者最终将能达到极高的开关速度,但现阶段这种器件的功耗太大难以实现。而后,光进到光束控制器,其功能类似于一个可以重构的全息元件,由于通道数量多,在这里也可能采用多平面的实时全息图列阵。光控制器是那些能不断改变连接图形的互连元件。为简单起见,计算机的三种互连系统,即处理器到处理器、处理器到存储器以及处理器到输入输出,都用一个方框表示,但实际上它们可以分别采

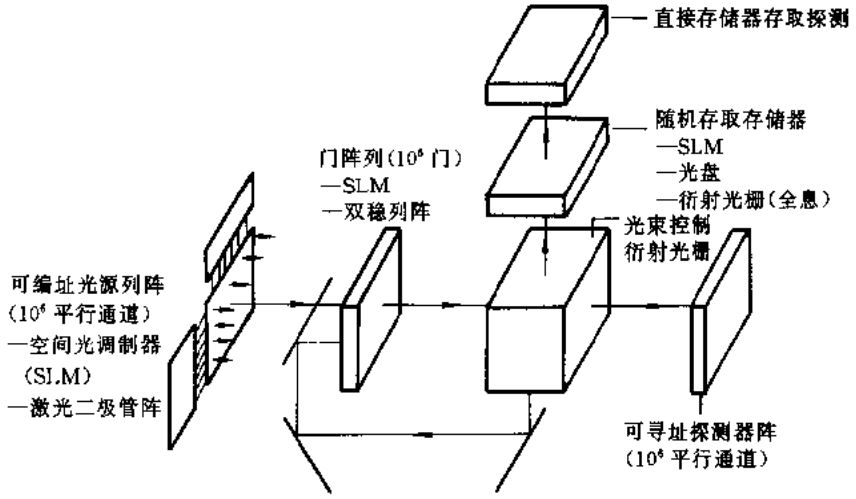


图 8　全光多处理机的系统结构示意图

用不同的元件。在理想情况下，光束控制器应能将门列阵发出的任何单光束传播到达随机存取存储器、探测器列阵或门列阵输入面上任何所要求的位置。

从技术上讲，这种系统包含 10^6 个并行通道是完全能实现的。例如由逻辑增强存储器所构成的光关联处理机就是这一类，其中每一个处理单元可以由若干个逻辑元件通过适当的互连元件而构成。例如，一个由 $n \times m$ 个逻辑元件(门)所构成的矩形列阵，其中有一个算术逻辑单元，几个寄存器以及还可能有几个缓冲存储器，这就构成一个处理单元。这种结构的一个实例如图 9 所示，其中 $n = m = 5$，即每个处理器结点包含 25 个逻辑元件，各个元件的功能已在图中标出。如果整个门列阵包含 1000×1000 个开关元件，这台处理机将有 40000 个结点。对这类多处理机来说，上述并行度在光计算中是不难实现的。

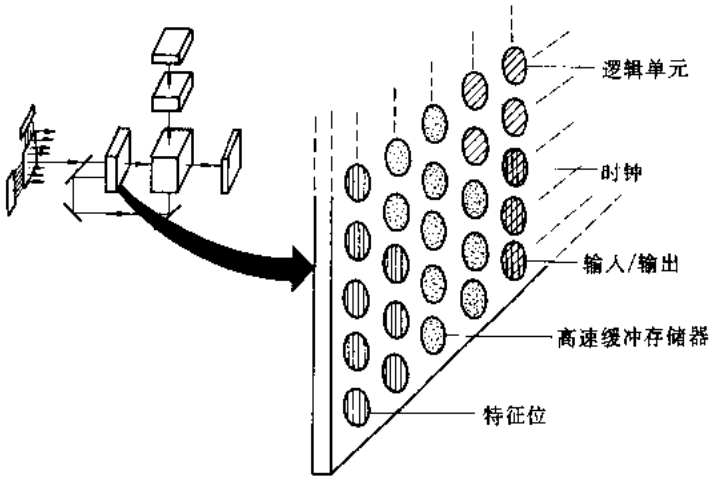


图 9　一个全光处理单元

2.3　光互连特点[7,18]

光子除了上述能提供巨量并行性外，另一个突出优点是实现高度的互连。光互连具有非常高的带宽，光学系统的空间带宽和时间带宽乘积很大，可以利用很多独立的通信通道；光互连可以提供大量的连接数，并能实现动态互连，这些都是电子互连无法做到的。由于光互连具有这些优点，不仅在全光计算系统中应用，而且也在电子计算机中引入光互连，被称为混合光/电子计算，或简称混合计算(hybrid computing)。

为了进一步提高电子计算机的速度和容量，正在发展集成度更高、速度更快的集成电路和新的系统结构。但是，不论在计算机系统的任何一级，采用传统的电子互连方法都已成为限制计算流通量的瓶颈。这种互连包括芯片内部的门与门之间、芯片与芯片之间、电路板与电路板之间，乃至较长距离的处理器与处理器之间，如果采用光互连方法，就可能根本上摆脱上述困境，使计算流通量大大提高。

以 VLSI 为例，过去 20 年中由于材料的加工技术的突破，使器件的集成度和速度大大增加了。现在预期最终可以使图形线宽达到 0.1μm 这个物理极限。但实际上在达到这个限值以前，首先要克服的困难是：(1)为芯片提供足够多的连接脚，芯片越大，集成度越高，要求的连接脚当然越多，但芯片上元件数随周长平方增加，脚数增加与周长是线性关系，即使现在的图形尺寸、芯片边上的接脚已非常拥挤无法再增加；(2)克服引线和接脚的分布电容和电阻所造

成的传输带宽限制，现在的 VLSI 的速度很大程度上是由芯片上的数据处理区与接脚之间的互连所决定的。特别在超高速 GaAs 集成电路中，互连带宽的问题更为严重。

芯片级采用光互连已有许多研究方案。现在最感兴趣的一个方面是克服现有芯片封装中连接端不足的困难，采用传统的电连接方法是无法实现的。例如大数量的 N 端随机存取存储器，即对芯片上所有存储单元能同时进行读/写，这对多处理机系统十分必要，那里有许多 CPU 需要对同一主存储器进行存取操作。在传统的封装中，由于接脚数量限制了地址线和数据线，这种器件无法制作。如果应用光互连，这种器件就可能实现，因为芯片通过光对外的互连数至少可以增加 10 倍以上。

光互连的技术在计算机系统结构的不同层次不可能完全相同。在电路板与电路板之间或局部地区网中可以采用分立器件的光源和探测器，以光纤作为连接。这一级光互连的最大优点在于长度-带宽积高，扇出量可以很大以及系统功耗低。在系统的较低层次，例如芯片与芯片之间或同一芯片的门与门之间，其主要的好处是能提供高密度通道。这一级的互连需要采用集成的光源和探测器、集成的波导和集成光路，或是自由空间互连。采用自由空间互连可以使互连的密度接近光的衍射极限，并能实现重构互连，即互连图形可以不断变更。

一种电光自由空间重构互连系统如图 10 所示。发光二极管列阵表示 n 个不同通道的输入，探测器列阵为 n 通道的输出。掩模可以用空间光调制器或其他二维光逻辑器件，通过可变程序可以使任一个输入光通道与任一个输出通道相连接，形成一个无阻塞的开关网。

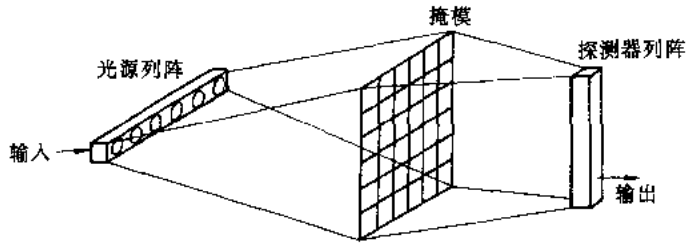


图 10　自由空间重构光互连系统

一个典型的采用光互连的混合计算机的系统结构原理如图 11 所示。为简单起见，没有包括全部系统，只是画出了其中二级处理器，每一个芯片上包含了若干个处理单元。作为这种光子/电子混合多处理机的实例，可以举出美国 MIT 开发的“连接机”(connection machine)。它属于一种细粒机，由一个很大的电路板列阵所组成，每块电路板上包含 512 处理单元，平均分配在 32 芯片内(即每芯片内包含 16 个处理单元)。图 11 所示每一电路板上的光电子芯片中包含有一个选频滤波器(全息图)，这些芯片上既有电子器件又有光器件。光器件为激光二极管和探测器，通过它们在芯片之间建立光通信。在装有芯片的电路板之间是重构衍射光栅的平面列阵，通过它完成互连

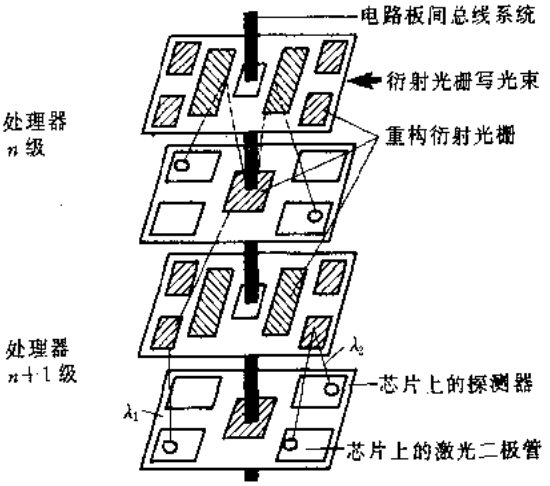


图 11　混合光/电子多处理机系统结构

过程中所需要的开关操作。这里还采用了波分复用技术把光数据流导向到规定的电路板上。图中标有 λ_1 的光束就表示这种操作，全息图直接位于发射芯片的上方，将光束导至下面一块电路板的中央，并在那里与主光束相叠加，然后可以传输到系统内所有的电路板。当到达所要求的电路板时，选频滤波器把光束衍射到电路板全息图上，通过它将光束导至最终探测器。电路板内或芯片内的互连由电路板上方的全息图来完成，如图中表示的 λ_2 光束。

3 光计算研究的进展和趋向

3.1 光计算的分类和历史背景[6,14,19]

严格地说，迄今对光计算的概念尚没有完全统一的确切定义。随着研究的深入，光计算所包含的内容在广度和深度上都已大大发展了。从历史上看，光计算概念的提出与一般所说的光信号处理内容有密切关系，即使在今天，光计算与光信号处理也并没有一个明显的界限。另一方面，参照电子数字计算的概念和内容，光计算乃是以光子和光学为工具进行各种形式的数据处理。也就是说，光计算既包括模拟计算，又包括了数字逻辑计算。如果把上面提到的利用光子互连通信与电子逻辑运算相结合的混合光子/电子计算也包括在光计算的范围内，则光计算的概念更为广泛。

实际上，图 8 中的“全光”处理机并不能完全脱离电子技术。因为可寻址光源列阵是由激光二极管组成的，则少不了要应用电子电路来驱动每一个激光器；作为逻辑门列阵的空间光调制器又少不了用电子电路来控制，等等。“全光”计算的含义是信息或数据一旦转换为光子流形式后，在传输、开关、运算、存储直到最后结果输出的全部操作过程中，光信号本身不再有光子到电子以及电子到光子形式的转变，这样可以充分利用光子的速度高、频带宽、无相互干扰和固有平行性等一系列特点。混合计算中为了实现以电子信号进行逻辑操作，以光子信号作为互连和通信，在全部操作过程中不可避免地有从电子到光子及光子到电子形式的转变，这种混合计算虽可兼有电子和光子两者的优点，但在每次转变中必然造成时间上的延迟和能量损耗。

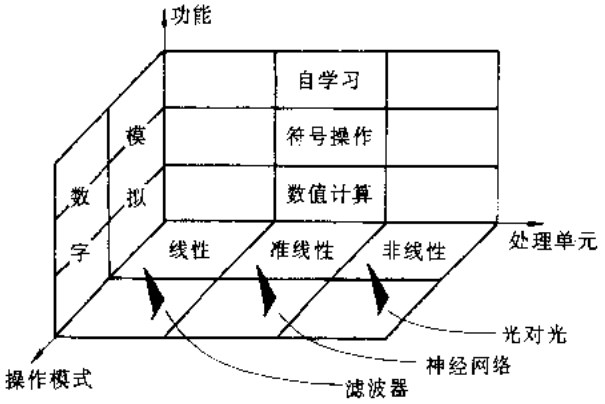


图 12 光计算分类图

这里讨论的光计算概念将是十分广义的。一种有价值的对光计算的分类方法如图 12 所示[14]。其中光计算用三种坐标给以分类：根据其操作模式，可分为模拟光计算和数字光计算；从处理单元的特性可分为线性、准线性和非线性三种光计算；从其功能的演变进程主要可分为数值计算、符号运算和自学习(self-learning)。

在早期的习惯观念上，光信号处理一般理解为仅仅是光学图像处理，而光计算也只限于数值计算。今天，光信号处理和光计算的概念都无疑大大地扩展了，而且在两者之间并不存在明显的分界，特别是对新一代的计算机要求的人工智能功能，如语音和图像的识别、推理、分类、

判断等都已不是简单的数值计算。所以许多著作中也不再以光信号处理与光计算来区分，而以模拟光计算和数字光计算两种不同模式的光计算来讨论，前者指的是光信号以连续函数形式进行各种处理；而后者则光信号成为离散的数据组或符号的形式进行处理或运算。现在用得最普遍的是二进制数，也可以是其他形式的，如余数制、十进制等。

不难理解，由于非线性光子器件的出现还仅仅是最近的事，早期发展的是应用透镜和电光、声光效应的线性光子器件实现的模拟光计算。相干光系统的一个重要应用就是信号的傅里叶变换，对傅里叶分量采用滤波器操作能完成信号的实时运算，这是普通电子数字处理系统不可能做到的。而且这种计算具有非常高的速度，即以光的速度从系统的输入到达其输出。例如相干光系统可在 ns 时间内完成包含 1024×1024 个光斑图像的二维傅里叶变换，由 1024×1024 个复数傅里叶分量与相同复杂程度的空间滤波器的相乘大约可在 3×10^{-14}s 内完成。滤波后的图像经过了傅里叶反变换，可以在 ns 时间后显示。当然，目前这种巨大的处理能力受输入/输出传感器带宽的限制。即使这样，每秒完成的运算次数仍是十分惊人的。

光处理的最初应用是衍射图形显示及通过简单的空间滤波操作使图像改善。后来，采用能控制初相位的连续调频滤波器进行微分、积分和拉氏变换运算。到 50 年代后期，光学线性系统分析已被广为熟知。许多研究者纷纷应用光学进行卷积、相关、匹配滤波以及频谱分析等方面的处理。

光处理应用最成功的实例是合成孔径雷达光处理机。这是一组以透镜作为所要求的匹配滤波器从采集到的数据产生雷达图像，由于这种专用处理器有着无可以拟的优良性能，至今仍然被使用着。

声光器件(通常称为布拉格元件(Bragg cell))是在 70 年代发展起来的，可以处理带宽扩展到 1GHz 范围甚至更高的信号。最初是采用一种空间积分结构进行瞬时频谱分析和相关处理。而后采用时间积分系统以增加频谱的分辨率或相关的增益。进一步又采用两个或多个布拉格元件能够充分发挥光学系统的二维本性。多通道布拉格元件具有解决更为复杂问题的潜在能力。

模拟光计算已得到了广泛的研究和许多应用。今后还将继续发展，应用也将进一步扩大。但模拟光计算存在两个主要缺点：一是通用计算能力较差，它只能进行一些简单的算术计算和某些变换；另一是计算的精度低和动态范围小，容易受噪声的响应。因而更广泛的应用，必须要发展数字光计算。数字系统的另一重要优点是有可能利用存储器系统，它能将信号的数据保持任意长时间而不出差错，这种稳定的存储器对模拟数据来说由于其受到噪声和漂移效应而不易实现。

60年代，人们首次对数字光计算进行了认真的考虑。然而由于当时的非线性光器件的限制，得出的结论是数字光计算技术无法与电子计算相抗衡。到70年代末，J. Goodman 指出光学互连和通信可能对现有的电子计算带来巨大影响。光纤在邮电通信方面的辉煌成就也给人们以无比鼓舞。更有决定意义的是在1976年前后，美国的 H. M. Gibbs 等人首次观察到了光双稳现象，英国的 S. D. Smith 等人发现 InSb 折射率的非线性比一般情况大10^9倍。因而到80年代初，已能制成室温工作的半导体光双稳器件，开启时间在 ps 量级。这种强非线性光学材料及器件的发现，使数字光计算系统有了实现的物质基础，并大大激发了人们对数字光计算研究的热情。

3.2 光计算研究的现状及趋向[5,21]

由于电子计算面临的困境及光子计算所展示的美好前景。进入 80 年代，美国、日本、前苏

联及欧洲诸国的政府、工业界及高等学校对光计算研究竞相争先，进入了一个新的高潮时期。有关光计算的大型国际会议和专题讨论会一个接着一个，许多杂志每年都编排有关光计算领域的专辑。美国的战略防御计划(SDI)，即人们常称的"星球大战计划"中也将光计算作为一项重要储备技术进行研究。

当前国际上对光计算研究的内容十分广泛，大致可以概括为三个领域，即系统结构及算法、器件以及材料和加工技术。

(1)系统结构及算法方面[6,9,14,16,19]。研究的主要方向是设法充分利用光的固有平行性和巨量互连的特点。现在已有许多光并行处理机的结构方案提出来了，虽然其中引用了一些电子并行机的基本概念，如单指令流多数据流(SIMD 系统)及多指令流多数据流(MIMD 系统)。显然，光处理机不能只是简单地仿照电子处理机的原理或两种器件的互易，必须在系统结构上根据光学特点和应用要求作新的设计。

早期的光计算系统是采用简单的和无源的(无功率消耗)元件，例如用透镜实现速度非常高的复杂的二维傅氏变换。利用这种系统结构可直接处理合成孔径雷达数据、图像的匹配滤波、探测及恢复接收到的雷达和声纳信号等。应该说这些方面的系统结构已研究得相当成熟，现在对光计算的研究已进入更高的发展阶段，把注意力转向更为广泛的问题。当前主要沿着三个重要方向：①采用光逻辑器件的数字光计算系统，使其高速计算和并行处理的能力，可以大大超过现有的电子计算机，以弥补其不足；②具有新系统结构的模拟光计算处理系统，它包含有巨量并行的算术运算和互连；③最近几年中，光计算的研究工作受到了对人类认识过程了解的强烈影响，仿造神经网络的模型，提出了一些系统结构的新概念。由比较简单和速度不快的元件(即神经细胞或神经元)所组成的巨量并行和高度互连的网络(即人脑)在认识(cognition)和感觉(perception)方面所表现出的惊人的能力，使人确信这刚好与上述第二方向的光子系统所具有的固有特长相吻合，因而这一新的方向特别引人注目。人工神经网和人工智能这一领域的研究尚处于初期，远没有信号处理和数字计算那样成熟，显然还缺乏指导这种系统结构发展的基本原理。然而，对人脑活动过程的研究已有近 40 年的历史，经过许多交叉学科的共同研究已获得一些很重要的概念，较有代表性的工作可参考 Grossber Kohonen 和 Hopfield 等的著作[14]。这些研究结果已对发展人工神经网系统结构的光计算有了很大推动。

众所周知，人脑的算术运算能力是远远及不上数字计算机的。但是当进行像关联(association)、归类(categorization)、综合(generalization)、分级(classification)、特征提取(feature extraction)、识别(recognition)及优选(optimization)等操作时，人脑绝对压倒一切现有功能最强的计算机。人脑在分析感觉数据和驾驶汽车技巧方面所表现出惊人的能力，更不用说在复杂的思维及智能推理方面，足以使它作为具有强烈迷惑力的模型应用于灵巧传感器(smart sensors)和自动识别机器人、自动控制以及其他许多人工智能系统中，这种神经网模型在处理感觉数据方面很重要的特点是使计算上十分复杂的问题，例如与视觉有关的一些问题容易解决。这些问题基本上属倒置问题，由于它们的不适定性(ill-posedness)，以往在计算技术上很伤脑筋。人脑具有的关联记忆能力，即使当提供的信息不完整时，仍能够成功地完成最近相邻搜索(nearest neighbor search)，这表明这种记忆能力是非常健全的，即使输入的数据有很大程度的失落或误差时仍能容许。人脑具有很强的补足失落信息的能力，这已启发人们可以应用这种模型来解决从不完整和噪声严重的数据中使信息复原的超分辨(super-resolution)及其他类似的问题。神经网模型能够用于优选问题的研究也已有报道。综上所述，人脑具有惊人的错误容差，虽然神

经细胞与身体其他部位的细胞不同，是不能再生的。尽管每个人有相当数量的神经细胞随时间而损失，然而直到 50 岁，我们的大脑功能仍保持正常。现在我们已能知道关于大脑如何处理信息的基本过程。它是由神经元形成的巨量互连网系统对信息作并行处理。大脑中的神经元大约在 $10^{10} \sim 10^{11}$，每个神经元与相邻的神经元有 $10^3 \sim 10^4$ 个突触相连，因而大脑的互连总数达 $10^{13} \sim 10^{15}$。即使我们假设神经元只有两个状态：即激发和抑制这种二进位制神经元状态，则大脑活动的自由度总数将达 $2^{10^{15}}$ 这样巨额的数字。每个神经元能独立地判别其所处状态并决定其是否要改变，这取决于它的突触（激发的或抑制的）输入总和是否超过规定的阈值，因而神经元表现出高度非线性（逻辑）运算。

光子的两大杰出贡献是能实现巨互连和平行性，因而光学对实现神经网模型有着特别重要的作用。在这种光学系统中，一般需要用到可编程序的非挥发性空间光调制器、光学的光放大器、光双稳态器件等作为可编程序的连接矩阵及光判定元件而构成多用途的人工神经网处理器。已有许多人对这种光处理器的系统结构作了大量的理论分析和实验验证。例如 Farhat 等人用光学系统研究了 Hopfield 模型，这是一个神经元数 $N = 32$ 的关联存储器，要使最近相邻的存储单元能完整和正确的记忆，即使有三个存储单元存在高达 30% 的输入误差还是允许的。甚至，如果有 10% 左右的神经元随机失效，也不致造成其性能的显著恶化。

(2)器件方面[10,11,14,17,20,21]。从图 13 所示的光处理器和光计算元件和系统的一般框图中可以看到作为输入输出器件、逻辑器件、存储器件以及互连和光束控制器件的主要的有光源和探测器、空间光调制器、全息器件和非线性器件等几大类。

① 首先，不论是全光的或是混合的光计算，光源和光探测器是必不可少的。为了适应光学并行性的特点，光源和探测器常常要求做成二维平面列阵，而且应该把发光二极管或激光二极管以及光电二极管与驱动和放大的电子器件集成起来。这种光电子集成器件目前有多种途径来实现。在混合光计算系统中，往往要求把以 Si 为基板的电子器件与以 GaAs 为基板的光

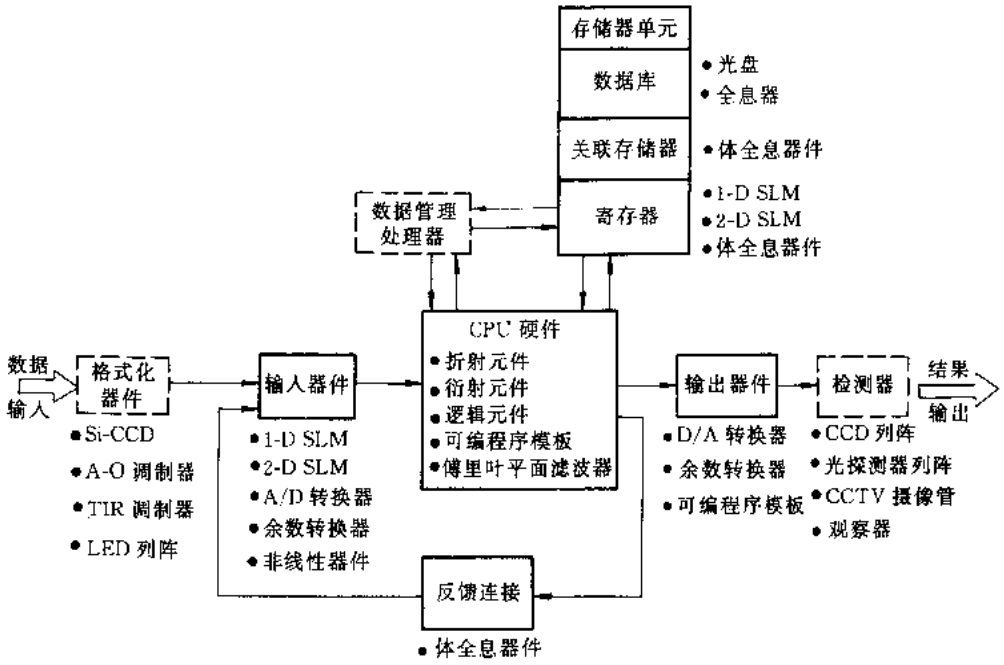


图 13　光处理器和光计算机元件及系统一般框图

子器件集成在一起，通常采用的方法是将不同的芯片浇铸在环氧基板上成为一体，最后在不同芯片间加以金属连线。这样得到的集成 Si/GaAs 光电子器件可以克服目前在 Si 和 GaAs 分立集成电路芯片间连接线电容过大的缺点。这种浇铸材料要求具有低电导率、低收缩率，与光致抗蚀剂的反应小，并有足够的平面度以保证光刻的质量，能全部满足这些要求的材料正在研究中。

另一方面，为了满足光电子集成的需要，正在寻找一种能将 Si 和 GaAs 这种结构不同的材料生长成异质结构的途径。存在的主要问题是由于晶格失配和热导的变化在界面处造成明显的应力，以及在一种材料的生长过程中保护好另一种材料不受影响。正在研究的各种方法有：采用保护性的缓冲层以及选用低温的激光束处理等。最近一项有突破性的成就是成功地将 GaAs/AlGaAs 双异质结 LED 及 Si 的 MOSFET 做成一种单片结构。将 Si 和 GaAs 生长在同一块蓝宝石基板上的试验也正在进行中。同时适合于光互连和光计算的光电子集成光源和探测器件也已提出了不少新的结构。

② 在光计算和光信号处理的系统中，空间光调制器是一类十分重要的器件[14,21]。一般来说，空间光调制器的功能是使空间光分布的相位、偏振、幅值或光强成为外施电信号或光信号的函数而变化。从上面的讨论已清楚地看到，在光计算系统中，这种能将信号加到一维或二维光数据场的器件对有效地利用光固有的高速、并行性和互连能力有着极为重要的作用。因而，近年来空间光调制器受到了有关光信号处理及光计算领域极大的注视，文献[14]中列出了国际上先后研究开发的各种型式和结构的空间光调制器不下 50 余种，现将其中主要的列于表 2。

表 2　各种空间光调制器的型式和结构

序号	名称/型号	调制材料	寻址方式		分辨率 lp/mm (No.Pixels)	灵敏度 /(μJ·cm⁻¹)	响应时间			研究单位
			光传感	电传感			写/ms	读/ms	存储	
1(O)	LCLV	向列液晶	CdS	–	30	6	10	15	15ms	Hughes
2(E)	Titus	KD_2PO_4	–	电子束	20	–	30	5	h	Sodern
3(O)	TP	热塑料	PVK 薄膜	–	200～1600	5	10	100	年	NRC 等
4(O)	PROM	BSO 或 BGO	BSO 或	–	6	5	<0.1	<0.1	<2h	Sumitomo
5(O)	MSLM	$LiNbO_3$	BGO	–	10	3×10	10	20	天或月	Optron 等
		KDP	光阴极			量子极限				
			和 MCP	电子束	1023×1023	–	–	–		GE
6(E)	Talaris	油膜	–	–	40	10	0.005	0.001	月	Singer
7(O)	Librascope	Smectic 液晶	液晶							
			(热吸收)	矩阵	128×128	–	10	10	年	Litton
8(E)	LIGHT－MOD	YIG(磁光)	–	电极						Semetex
	SIGHT－MOD			Si 电路	5000×1	–	<0.001	<0.001	<1μs	Xerox
9(E)	TIR	$LiNbO_3$	–	–	10	2	2	<0.5	5s	Lockheed
10(O)	Phototitus	KO_2PO_4	Si 二极管	–	10	10	<0.01	<0.01	s	UCSD
11(O)	PLZT	PLZT	Si 光							
			晶体管	–	40～120	30	5	4	15min	Xerox
12(O)	RUTICON	变形弹性膜	非晶 Se	–	128×128	2	0.025	0.04	200	TI
13(O)	DMD	变形膜	Si 光						ms	
			晶体管	–	40	3×10	0.01	0.01	天或月	NRL
14(O)	PEMLM	变形膜	光阴极			量子极限				
			和 MCP	电子束	20	–	10	10	年	USC
15(E)	VO_2	VO_2	–	矩阵	16×16		1	1	15ms	VARAD
16(E)	Optical Tunnel Array(OTA)	悬片	–	电极						

空间光调制器可以分成光寻址和电寻址两大类。图 14 为二维光寻址空间光调制器的基本结构，由光电导层、镜面反射和遮光层以及电光介质构成三层结构。工作时，有一偏压作用在三层结构上，当写入图形照射在光电导层上，由于图形的明暗区分使电导改变而引起双层介质上分压变化，电光介质层上电压的变化控制着读出光束的相位、幅值或偏振状态，从而获得调制的输出图形。电寻址空间光调制器的区别在于由电信号输入代替了上述的光信号输入。电寻址实现的方法有许多种，主要的如扫描电子束、电极矩阵、半导体效应等。

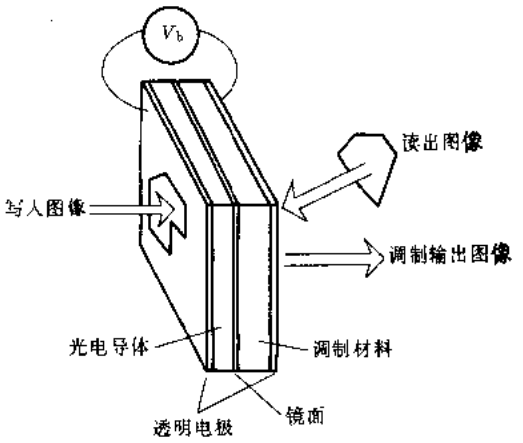


图 14　光寻址空间光调制器

空间光调制器具有不同的功能，除了调制和开关功能外，还可作为放大器、电信号和光信号变换的传感器、存储器，利用其非线性效应还可作为处理器或运算器这种器件的主要参数有分辨率，即每毫米长度有多少单元，可以在 $10 \sim 10^3$ 范围内；其次是灵敏度，单位常用 $\mu J/cm^2$，范围可从 $10^{-4} \sim 10^{-6}$变化；另一个重要参数是响应时间，包括写入、读出和存储的时间，采用不同的材料和结构，有很广的变化范围，例如读、写时间可从 1ps 变化到 10ms，存储时间可从微秒到若干年。

由于材料和加工的原因，目前能商品化的空间光调制器只有少数几种，例如液晶光阀(LCLV)、Pockels 读出光调制器(PROM)等，在分辨率、灵敏度和响应时间等性能方面都不能完全满足光信号处理和光计算的要求。当前研究的目标是希望能获得如下性能的空间光调制器：1000 × 1000 分辨单元，千赫频率的帧速，量子极限的灵敏度，小时以上的存储时间。

③ 体全息器件[10]，这对发展光计算系统也是十分重要的。这类器件的基本功能是由相干物(信息源)及参考光束在一定厚度的材料体内产生的空间变化光强图形，利用这种光强变化使材料的某种性质产生相应的空间变化，由此引起读出光束的相位或幅值产生足够的调制，从而使物信息得以重建，工作原理如图 15 所示。再与写入光束(信号及参考)相干涉形成材料的折射率调制，被衍射的读出光束沿着参考光束路径反向传输，这种器件的重要特征是可以把整页的信息进行全息存储，对存储的信息作布拉格选择角编码，工作区内兼有线性非线性两种响应特性以及能够作并行和顺序两种不同的编码。

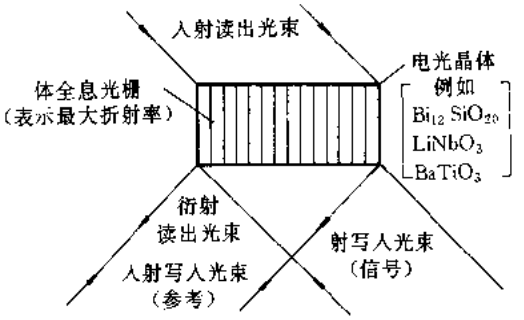


图 15　在光折变材料中体全息光栅的示意图

体全息光器件在光处理和光计算系统中可能应用的范围十分广泛。首先，它可以构成关联存储器，它允许完全并行和同时存取由前级计算结果所构成的多平面存储数据，而不需求助于中间地址编码和译码。其次，它可以完成实时光处理和光计算，例如进行相关和卷积运算以及图形边缘增强这类非线性运算。再次，体全息光器件能用作计算机发生全息的中间存储，从而能够顺序写入和并行再现。而且，有一定厚度的体全息器件所具有的布拉格选择性可以把

波长多路复用方案应用于信息的存储和检索。最后，这种器件能用于可编程序的光互连，并且在每一位置上对正交互连角和互连焦距两者作独立的控制。后一个特性可以允许多平面互连的系统结构。

在光处理和光计算系统中采用这种器件受到的最大限制是其每帧再编程序的速度。目前用来制造实时体全息器件的材料是一类光电导和电光晶体，或称为光折变(photorefractive)材料，如铋硅氧化物、铋锗氧化物及铌酸锂等。最近发现，GaAs、Inp 等Ⅲ-Ⅴ族半导体材料也具有这种光折变现象，但一般其电光系数很低，只有它们的量子阱材料具有明显增强的非线性效应。

这些器件在实用化以前，需要对它们的性能作进一步改进。除了要提高速度外，还要提高饱和衍射效率，这决定信息存储和检索的总效率，对器件的灵敏度也有影响。另外，最大信息容量也是一个重要性能，这表示单位体积能存储的信息量。

④ 光双稳器件[11,15,17,20,21]。光双稳器件的基本特点是输出光强对输入光强的响应是非线性的，对应于一个输入光强存在两个可以相互转换的稳态输出。图 16 是光双稳器件的典型特性，在图(a)中，输入光强 I_i 为某临界值 I_2 时，输出光强 I_o 从低值跃变为高值并随着 I_i 的增加一直保持；当 I_i 减小直到小于 I_2 的另一临界值 I_1 时，I_o 才从高值突变为低值。出现这种输入/输出光强的滞后回线是光双稳的基本特点。图(b)表示光双稳器件的这种非线性还具有光晶体管、光振荡器等功能，当输入光强在 I_g 附近，I_i 的微小变化能引起 I_o 有较大的改变。因而光双稳器件在光处理和光计算系统中有许多应用，例如可作为开关元件、存储元件、放大元件以及逻辑元件等。这些功能对全光计算的发展是十分关键的。

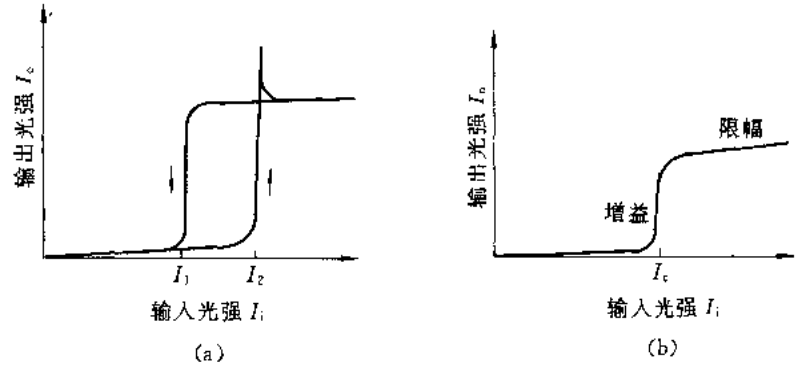


图 16　双稳器件典型特性

(a) 双稳特性(用作存储器)；(b) 高交流增益(用作光晶体管、鉴别器、限幅器)

实现光双稳器件可利用各种物理效应。从实用出发，光双稳器件应有很高的开关速度，很低的功率消耗，具有良好的级联特征，即上级的输出能直接作为下级的输入，最后，当光双稳器件制成高密度列阵时其串音应极小。有一种非线性标准具的器件是由三次非线性材料构成的 Fabry Perot 干涉器。这种器件的开关时间可达到 ps 量级，比现有的电子开关短了 1000 倍。但目前在功耗方面还不够低，需要数十毫瓦。

⑤ 分子电子器件(Molecular Electronic Devices，MED)[14]。分子电子器件是 80 年代开始研究的一种新器件。它的基本原理是通过分子的支链或功能团的结构变化来存储信息，这种可逆的结构变化包括键长或键角，或两者同时改变。例如在甲基吡唑($C_4H_6N_2$)分子中通过氢原子的位移引起的结构改变可作为一位信息的存储，其中不稳定的 H 原子的运动伴随着 π 电子

的重新排列。当单键变到双键时，键长通常缩短1.5～2.1nm。当上述分子从等式左面变到右面时，环上的甲基（$-CH_3$）由于H原子位移其键角将向下偏10°左右。

图17表示具有三个不稳定单元（L_1，L_2，L_3）的分子结构，通过光跃迁使分子结构改变。图中表示了分子从(000)到(100)再到(101)两次光跃迁，这种具有三个不稳定单元的分子可储存8位的信息。

分子电子器件可以用电信号或光信号作为输入，输出一般是光信号。在光计算机中，这种器件可作为开关、存储器或逻辑器件，在通信和生物传感器等系统中，这些功能都同样适用。由于这种器件是分子尺寸，因而具有很高的密度，可望达到10^{15}～10^{18} gates/cc，其平行性可大于10^7，开关时间$<10^{-12}$s，每位信息要求的能量仅10kT，即对10^7个元件的列阵其功率消耗仅10μW，而对同样数量的GaAs器件的列阵来说其功耗为1kW。

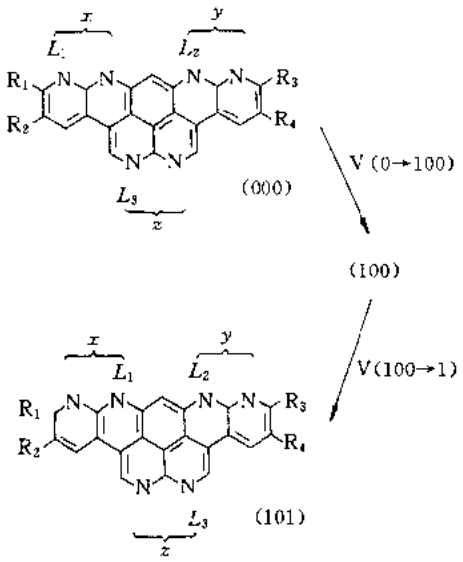


图17　可存储三位二进制信息的分子结构示意图

（3）材料和加工技术方面[10,17,20,21]。电子计算机应用的器件现阶段大多数是由硅单晶作为基板的，对超高速器件或其他特殊应用场合，也已开始采用化合物半导体GaAs材料。但是对光计算机来说，至今尚不能肯定那一种材料能占绝对优势。目前的状况是：一方面对已作过研究的材料进一步筛选和改进性能；另一方面还继续不断地致力于新材料的探索和开发。

从上面列举的光计算系统中各基本器件的功能，可以看出它们对材料的要求是各不相同的，情况要比集成电路复杂得多。除了现阶段在光纤通信系统中，用作光源和探测器比较成熟的Ⅲ-Ⅴ族化合物半导体材料和用作光波导的石英光纤材料在光计算系统仍可作为基本材料外，其他光计算器件性能提高很大程度上决定于非线性光学材料，特别是基于三阶光非线性效应的器件，如上述的光双稳器件和实时全息器件等。同时，应用电光效应和光折变效应的器件（如空间光调制器和开关列阵等）性能提高也取决于二阶光非线性材料的改进。由于光计算的发展一直受到材料的限制，这就促使了对光非线性材料的高度重视。

从器件功能的角度出发，对光非线性材料特性的要求最主要是光非线性系数大、开关能量小和开关时间短；其次，还要求没有色散和光损阈值高。从技术的角度出发，要求材料能做成薄膜和易于加工，机械强度足够和不易变形，工作温度范围宽，在大气环境中不易氧化或腐蚀。

当前，最受人注意的光非线性材料有三类：无机电介质、有机高分子以及无机超晶格。要了解材料的光非线性效应是什么引起的，需要弄清入射的光子对材料中的原子产生怎样的影响。光子进入材料后，将它的一部分或全部能量交给那些与核联系得不紧密的电子，并使它们与其原子分离，从而造成电荷分离。如果分离的电荷（例如带负电的电子和带正电的离子）能维持一段时间，则由此产生的电场将引起材料的非线性响应，这种非线性的大小与电荷分离的程度有关。下面我们讨论上述三类材料中电荷分离现象的特点。

对无机电介质来说，入射光子造成的电荷分离主要是导致产生的自由电子被俘获在材料中的其他位置上。当电子一旦获得足够能量，它就离开原来的原子而被邻近的原子所吸引并

参加这一原子的外层电子结构。在这类无机电介质中最常用的三种材料是:硅酸铋(BSO);铌酸锶钡和钛酸钡。虽然这类材料目前用得比另两类要普遍得多,但是其响应时间仅为 ms 量级。

对有机光非线性材料研究的历史还不长,但从初步的结果已能看出这类材料不仅可能获得比无机电介质大得多的光非线性系数,而且响应速度也非常快。这可解释为在有机高分子光非线性材料中,由于存在某种状态的电子(称为 π 电子),它与其原子结合的键能较低,因而容易造成电荷分离。而且,这种有机分子的结构明显地比无机分子大,就可能使自由电子沿着分子链移动,从而使正负电荷中心间距离加长。当前受到重视的材料有:一代或二代乙炔及二乙炔;蒽及其衍生物;染料;巨环化合物;聚苯并咪唑;聚苯并双噻唑及聚苯并双噁唑;聚酯及聚酯酰胺;聚醚酮;聚奎喔啉;卟啉及金属-卟啉络合物,TCNQ 或 TNAP 的金属络合物;以及尿素等。这类材料的主要缺点是对环境的稳定性比无机材料差,例如易于被氧化。为此,最近有人利用多孔玻璃(称[illegible])作为无机材料将[illegible]

需的能量，同时它本身也提供作为加工所需的材料源，如离子注入所用的离子束等。这种加工过程已不再像常规加工中是单纯的机械力学过程。而是粒子束与加工固体表面之间许多物理和化学作用的综合。例如用反应离子刻蚀光波导光栅透镜时，不仅有离子轰击固体表面产生溅射的物理过程，还包括了 Cl^- 或 F^- 等活性离子与固体表面产生易挥发性化合物的化学过程。

从加工形式来看，微加工通常包括三方面，即薄膜生长、图形的发生和转移(复印)以及刻蚀。

薄膜生长可以通过许多途径实现，许多金属膜可以通过真空蒸镀淀积而成，大多数氧化物介质膜可以由溅射淀积生长。对半导体单晶薄膜往往采用外延技术制备，这包括气相外延(VPE)、液相外延(LPE)、分子束外延(MBE)以及金属有机化学气相淀积(MOCVD)等。图 19 表示分子束外延装置原理图。多种分子束在超高真空室中从熔化喷炉射出到达加热的基板上生长成单晶外延膜。在此同时可以进行掺杂。这种外延技术突出的优点是外延层厚度可获精确控制，通过挡板操作可以生长出单原子层。这比 VPE 和 LPE 在生长厚度的控制精度上几乎大 100 倍。因而超晶格结构主要用 MBE 生长，各层的厚度和化学组成都可准确控制。这种方法的缺点是生长速度较慢，通常其速率约为 1μm/h。MOCVD 与 VPE 的主要区别在于前者采用了金属有机化合物作为生长源，因而有些无法用 VPE 生长的外延膜如 AlGaAs，可以用 MOCVD 生成。MOCVD 在生长厚度的控制上虽比不上 MBE，但它的生长速率比 MBE 高。因而，这两种外延技术对光电子器件特别重要，最近又有人结合 MBE 和 MOCVD 的优点加以改进，发展了一种化学束外延 CBE，它的生长速率比 MBE 有明显提高。

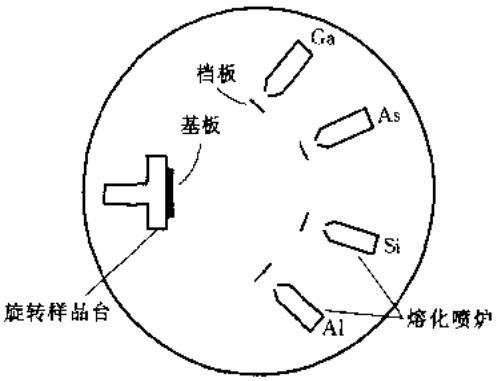


图 19　分子束外延装置原理图

制作掩膜和复印统称为制版，保证器件图形尺寸和公差与设计相一致，这是获得高性能器件必不可少的条件。最常用的制版技术是采用可见光波段。由于光的衍射效应，使图形的最小线宽受到限制，随着器件图形不断缩小，已发展了电子束制版，X-射线制版以及离子束制版，采用光学制版，图形线宽可以下降到 2 ~ 3μm。电子束可用来制造掩膜，也可以直接将图形做在圆片上。虽然电子束本身可以聚焦到数 nm，但由于电子束的背散射效应限制了图形的最小线宽目前的水平为 0.1μm。X-射线束制版可以将线宽减小到几十 nm，但要求掩膜必须对 X-射线有很强的吸收能力，并且目前尚缺乏适用于 X-射线的光学元件，很难使 X-射线聚焦和偏转，因而 X-射线还不能用于做掩膜，只能用于图形的复印。离子束由于其质量比电子重得多，因而其背散射不致影响它的分辨率，有可能加工纳米图形。另外，微离子束还能提供无掩膜制版，即按图形直接在基板上进行局部的掺杂、薄膜生长或刻蚀。这项新技术目前尚处于实验阶段。

刻蚀的目的是将不需要的部分除去。早期都采用湿法腐蚀，即要保留的部分用某种材料如抗蚀剂)保护起来，由化学溶液将不保护的部分腐蚀掉。因为这种湿腐蚀在一般情况下纵向和横向的腐蚀速率大致相同，这种各向同性的刻蚀导致产生严重的“钻蚀”效应，无法得到高宽比大的线条。后来发展了各种干法刻蚀技术，由于粒子束轰击这种物理过程有明显的各向异

性，其纵向刻蚀速率将大大超过横向的，可以获得非常陡削侧壁。这类干法刻蚀根据其工作物质的不同有：等离子刻蚀、反应离子刻蚀、离子束刻蚀、反应离子束刻蚀等。后两种刻蚀中，离子束具有较大动能，刻蚀呈明显的各向异性，可以获得纳米图形。

迄今为止，光计算材料和器件所应用的微加工技术，多数是沿用了微电子器件发展中所产生的技术。但是由于光集成和光电子集成器件中涉及的材料和微结构远比微电子器件复杂，因而现有的这些微加工技术已适应不了光计算系统中新器件的发展。例如根据器件功能的要求能在同一基板的不同区域生长不同材料的技术，或称薄膜局域生长技术，又如为了更好利用光的平行性和巨量互连的特性，需要发展加工三维光集成和光电子集成器件的技术等。对微加工技术进一步的了解，读者可参阅本书的“微米纳米加工概论。”

3.3 对光计算的展望[5,8,14,19]

应该看到，光计算的竞争对手是电子计算。正如第一节所讨论的，电子计算已获得了巨大的成就，它技术成熟、通用性强、成本低廉，熟悉电子学的工程师要比熟悉光学的工程师多得多，而一般光学工程师又很少精通计算机的，因此这场对抗赛不是轻而易举的。

从物理上看，光学效应的响应速度远比电子效应快，光计算系统完全有可能达到 10^{15}b/s 的运算速度，另外再加上允许有百万(1000×1000)通道以上的平行处理及可编程的巨量互连能力，这些都赋予光计算具有很大潜力。显然，沿用现有电子计算机的系统结构都无法充分开发这些潜力，这就首先要求光计算在系统结构上应有根本的改变，但也不要指望这场革命会在一个早上出现。在现阶段，不能忽视和排斥将光学部分引入电子计算机所带来的益处。

认为光计算能够全面取代电子计算的想法是不现实的，持有要光计算非全光型者莫属的观点也是不明智的。从整体上看，光学很难胜过电子学，但是利用光学固有的特长填补电子学之不足或电子学无法实现的领域，光学仍大有用武之地。光纤通信和光盘已取得的成就完全可以建立起这样的信心。在新一代计算机行列中，特别是在解决与人工智能有关的问题方面，光计算将以引人注目的姿态显现。

参 考 文 献

[1] 陈厚云，王云刚，等编著.计算机发展简史.北京：科学出版社，1985

[2] John P. Hayes. Computer Architecture and Organization. MeGraw-Hill Book Company, 1978；中译本.计算机结构和组织.上海：科学技术文献出版社，1982

[3] 金　兰，王鼎兴，沈美明.平行处理计算机结构.北京：国防工业出版社，1982

[4] Charles L. Seitz. Concurrent VLSI Architechures IEEE Trans on Computers. 1984, C-33(12):1247~1265

[5] Optics News. 1986,12(4)；中译本.光学计算时代的来临——80年代以来的十二种独特的观点.光机情报，1986,(12)

[6] A.W.Lohmann. Chances for Optical Computing; P.Chavel, et al. Architectures for a Sequential Optical logic Processor; A.Huang. Parallel Algori Thms for Optical Digital Computers; J.Tanida, Y.Ichioka. Optical Logic Array Processor; R.A.Athale. Optical Matrix Algebraic Processors. A Surocy, etc. Toth International Optical Computing Conference, MIT, Cambridge MA, 1983

[7] J.W. Goodman. Optical Interconnections in Micrcoelectronics; A.A.Sawchuk. Numerrical Optical Computing Techniques; A.Huang. Optical Digital Computers?; H.J.Caulfield, W.T.Rhodes. Digital Optical Systolic Array Processors; etc, IOC-13 Conference Digest, Japan, 1984

[8] A. Vander Lugt. A Review of Optical Signal Processing; C. Mead. Potential and Limitations of VLST; A. Lohmann. What Optics Can Do for the Digital Optical Computer; J. A. Neff. The Role of Optics in Future Computational Sytems; A. Huang. Why Use the Parallelism of Optics; etc. A Digest of Technical Papers Presented at the Topical Meeting on Optical Computing Incline, Village, Nevada, 1985

[9] A.A. Sawchuk. Digilal Optical Computing. Proceedings of the IEEE, 1984,72(7):758~779

[10] J.A. Neff. Optical Computing; A.R. Tanguay, Jr. Materials Requirements for Optical Processing and Computing Devices; N. Peyghambarian, H.M. Gibbs. Optical Bistability for Optical Signal Processing and Computing; etc. Special Issue of Optical Engineering, 1985,24(1)

[11] S.D. Smith. Lasers, Nonlinear Optics and Optical Computers. Nature, 1985,316:319~324

[12] R.P. Bocker. Photonic Computing; M. Takeda, J.W. Goodman. Neural Networks for Computation: Number Representations and Programming Complexity; etc. Applied Optics (Photonic Computing Special Issue),1986,25(18)

[13] S.H. Lee, et al. Architectures and Algorithms for Digital Optical Computing Systems with Applications to Numerical Transforms and Partial Differential Equations; W.T. Rhodes, et al. Optical Computing and Nonlinear Optics; P.R. Haugen, et al. Directions and Development in Optical Intereannect Technology; etc. Proceeding of SPIE, Optical Computing Los Angeles, 1986,625

[14] J.A. Ionson. Computing Challenges and the Principles of Innovations; J.P. Boris. Supercomputing at the U.S. Naval Research Laboratory; F.L. Carter. Molecular Computing and the Chemical Elements of Logic; R. Hecht-Nielsen. Performance Limits of Optical, Electro-Optical, and Electronic Neurocomputers; etc. SPIE Institutes for Advanced Optical Technologies, Second in the Series, 1986,634

[15] P.W. Smith. Digital Optics: Progress Toward Practical Applications; H.M. Gibbs, et al. Nonlinear Etalons and Oplical Conmuting; D. Casasent. Optical Artificial Intelligent Processors; etc. 1986 International Optical Computing Conference, Proceedings of SPIE 700,1986

[16] R. Arrathoon. Digital Optical Computing; G. Abraham. Multiple-Valued Logic for Optoelectronics; S.H. Lee. Optical Implementations of Digital Algorithms for Pattern Recognition. etc. Optical Engineering, 1986,25(1)

[17] C. Warde, U. Efron. Materials and Devices for Optical Information Processing; M. Dagenais, W.F. Sharfin. Extremely Low Switching Energy Optical Bistable Devices, etc. Optical Engineering, 1986,25(2)

[18] L.D. Hutcheson. Optical Interconnections; P.R. Haugen. Optical Interconnects for High Speed Computing; B.D. Clymer, J.W. Goodman. Optical Clock Distribution to Silicon Chip?; etc. Optical Engineering, 1986,25(10)

[19] R. Arrathoon. Historical Perspectives: Optical Eross Bars and Optical Computing; A.D. McAulay. Real-Time Probabilistic Oplical Expert System; K. Wagner. Multilayer Optical Learning Networks; etc. Proceedings of SPIE, Digital Optical Computing, Los Angeles, CA, 1987,752

[20] N. Peyghambarian. Optical Computing and Nonlinear Optical Signal Processing; J.A. Neff. Major Initiatives for Optical Computing; etc. Optical Engineering, 1987,26(1)

[21] U. Efron. Optical Information Processing: Systems, Materials, and Devices; D.A.B. Miller. Quantum Wells for Optical Information Processing; etc. Optical Engineering, 1987,26(5)

微米纳米加工导论

1 微米纳米加工与现代科学技术

1.1 微加工的方式和特点

加工方式和加工技术往往是社会生产力先进程度的重要基础和标志。人类的文明史从某种意义上来说是人类与自然的斗争中生产工具的演化史。迄今为止,人类一切生产活动中所使用的工具和器械其基本目的有两个:一是使人们从繁重或单调乏味的劳动中解放出来;二是延伸人类器官的功能以弥补其所不能。中国传统艺人可以在一粒大米上刻下唐诗宋词,但任何能工巧匠都无法靠手艺雕出集成电路芯片来。

无论是为了满足功能上的需求,或是从经济的角度出发,一个共同的发展趋向是元器件、装置或系统的尺寸的小型化。这种小型化或微型化的进程,现在已使加工的尺寸达到原子和分子量级,进入这个"微观王国",传统的机械加工方式几乎完全无能为力了。

随着元器件和系统不断向小型化进展,一个新的工程分支出现了,它被称为微结构(microstructures)工程。最为明显的例子是为了适应又一次新的技术和工业革命,不断地要求电子电路缩小尺寸和降低价格,从而导致微电子技术和工业的迅猛发展。与此相应,一门新的微加工(microfabrication)技术应运伴随而生。微加工方法几乎与现有的机械加工方式完全不同,它不再采用车刀、铣刀和钻头等这类整体的加工工具,其加工过程也不再是简单的机械切削或压力加工。在微加工中,往往利用具有一定能量的粒子束或射线包括电子束、离子束、原子或分子束、激光束和X-射线等与固体(工件)表面产生的物理和化学过程达到某种特定的加工目的。这种加工不仅包括图形的发生、转印和刻蚀,也包括单层或多层薄膜的淀积或外延生长,整片或局部的材料化学成分和微观结构的改变,以达到器件的设计要求。

微加工有两种不同的物理模式:一种是在整块的基片上将不需要的那部分材料去除,这有些像艺术家搞雕刻;另一种则是在需要的部位将材料添加上去,这类似于造房子那样在需要的地方把一块块砖堆砌起来,微加工中的砖块就是原子和分子,有人把这种技术称为分子工程(molecular engineering)。分子工程实际上就是高级阶段的微加工,毫微加工与分子工程的目标已相差不远了。

不同的微加工模式对粒子束功能的要求也不相同。一般来说对前面一种加工模式,加工过程主要靠粒子束提供足够的能量以造成原子从基片表面溅落,或产生分子化学键的断裂或交链等。而后一种加工模式,往往不仅要求粒子束的能量,而且粒子束本身就是新的材料来源,如离子注入、分子束外延等。如同电焊的焊条,它的作用使工件因接触处形成的电弧加热,又是填充焊缝的材料。因此,粒子束从其功能来区分可分成两类:电子束、激光束和X-射线等这类粒子束在微加工过程中主要作用是提供能量;离子束,原子和分子束这类粒子束兼有能量和物质供应双重作用,可能在分子工程中更有应用前景。

1.2 微加工技术的发展阶段

如前所述，微加工技术的发展是与微电子工业分不开的。从60年代开始，到90年代末的三十年中，集成电路的发展经历了小规模集成(SSI)、中规模集成(MSI)、大规模集成(LSI)和超大规模集成(VLSI)不同阶段。与此相应，微加工技术也经历了不同的发展时期，如图1所示。这是反映微加工发展阶段的一个重要标志。

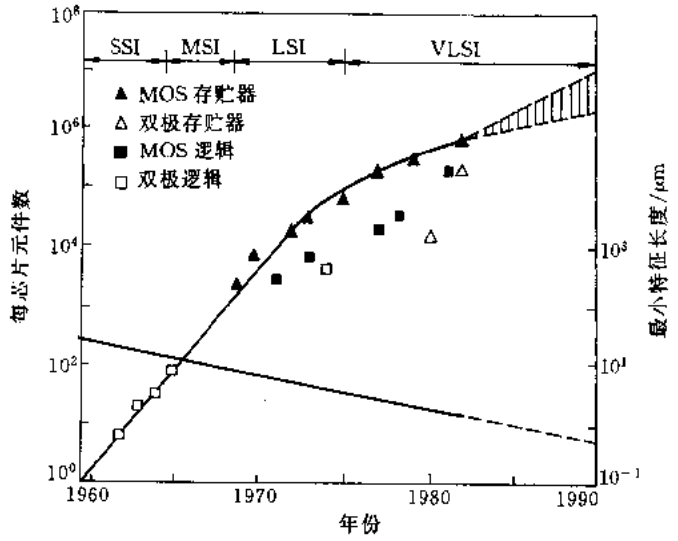


图1 微加工技术的发展图

小规模集成电路每芯片上的元件数在100以下，最小线宽在10μm以上；中规模集成电路的元件数不到1000，线宽不小于7μm；大规模集成电路的元件数可高达10^5，其线宽在3～5μm；超大规模集成电路芯片的元件数在10^5以上，图形的最小特征线宽不超过3μm，甚至缩小到亚微米。一般来说，线宽在5μm以上，制版可采用光学方法，刻蚀可采用湿法刻蚀，掺杂可沿用热扩散，微加工技术没有重大变革。当线宽小于3μm时，往往需要电子束制版、干法刻蚀和离子注入掺杂等新工艺。当图形尺寸进入亚微米，甚至纳米范围，则对微加工技术有更高的要求。1991年日本日立公司制成64Mb的动态RAM芯片在10mm×20mm的面积上包含有1.4×10^8个元件[1]，图形线宽为0.3μm。采用了电子束制版和X-射线曝光技术，研究小组克服了电子束制版中的邻近效应(proximity effect)，保证了图形尺寸的精度。他们预期，到1995年这种器件将可大量上市。

一般来说，微结构在实验研究阶段的尺寸要比工业生产中应用的导前一个量级。在80年代里，已有越来越多的大学实验室和大公司的研究所开始着手研究更小尺寸，即小于0.1μm的纳米结构(nanostructures)，这种尺寸的半导体器件也称为介观结构(mesoscopic structures)，这种结构所以具有极大的吸引力，因为在这样小的结构中所呈现的量子尺寸效应和电子增强现象可能导致性能更高、价格更便宜的称为纳电子学(nanoelectronics)的新一代集成电路器件。现在可以预期，采用常规技术的MOS集成电路，其最小的极限尺寸为0.25μm，利用电子量子效应的集成电路尺寸可缩小到0.010μm，如图2所示[2]。

在今后的十年中可以预期，从工业生产方面，微电子和光电子器件将进入亚微米尺度，有可能缩小到0.1μm。在实验室研究方面将集中于纳米加工技术的全面突破，首先是10nm量级的加工技术。

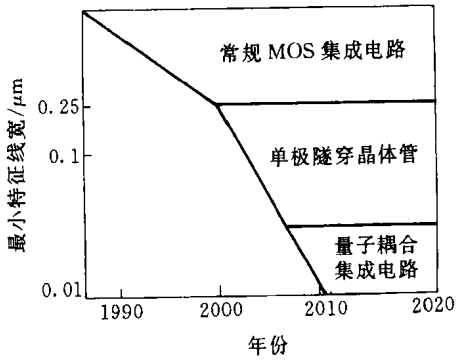


图 2　非低温集成电路技术的预测

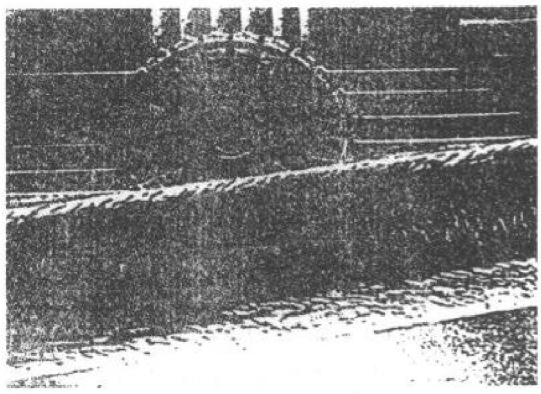


图 3　微电机的 SEM 照片(上部为头发)

1.3　微加工技术的应用及其重要性

虽然微加工技术是在微电子学的推动下迅速发展起来的。但集成电路取得如此明显的成就,在很大程度上取决于微加工技术的进步。这极大地鼓励人们考虑如何利用这种新一代的加工技术,来发展其他各种微结构元器件。最明显的例子就是在传统的光学领域提出了集成光学新概念,并大大推进了光电子器件和光电子集成的发展,这为光通信、光信息处理以及光计算等重要应用打下了基础。以光子作为信息载体不仅速度高,并具有多通道并行处理和巨互连能力。为充分利用光学的固有特性,最终将发展成为三维集成的集成光学和光电子集成器件[3]。这样,必将对现有主要用于集成电路的微加工技术的新发展有很大推动。

在仪器和测量方面,新近已利用现有的微加工技术制成了微型的传感器(sensors)、换能器(transducers)和执行器(actuators),这已形成了一门新技术,有人称之为微机子学(micromechatronics)或微动力学(microdynamics)[4,5]。这是利用微加工技术开辟了一个机械的微观世界。目前,已在硅片上做成了微型的传感器(如加速度传感器)、电机、阀和泵等。图 3 是在 Si 片上制成的微电机扫描电镜照片,其直径约为 0.1mm(照片上方为头发)。进一步的研究还可根据不同的需要将各种传感器和执行器以及有关的电路集成在同一基片上。毫不奇怪,通过这头发丝般的装置,能对像汽车、飞机之类的庞然大物进行精确的控制和操纵。

在材料领域,现阶段不仅已能利用分子束外延技术制成单原子层的超薄膜,也能形成各种超晶结构可获得许多可贵的材料性能,采用聚焦离子束技术可达到材料的微区生长和改性。但要达到能在三度空间内重构分子这个分子工程的最终目标,在技术上还有大量工作等待去做,包括还需开发更多新的技术途径。微加工技术的进步将把材料和器件界限打破,材料的设计和加工与器件的设计和制造将变得越来越不可分割。

即使在生物领域,人们已在考虑和研究如何利用微加工技术制成供人造肺和肾应用的膜和过滤器、人造视网膜以及人造鼓膜等,并将最终设法构成人造细胞。在活的细胞中,自然已赋予与我们极为广泛和错综复杂的分子尺寸器件的运行机制,它为分子工程提供了极好的样板,值得我们去理解和效仿。在分子工程成为现实以前,首先是要探索到这些器件功能的奥秘。这无疑将导致微加工技术新途径的不断开拓。

综上所述,可以毫不夸张地说,现在我们已进入了"微加工时代"。它与"电子时代"、"高分子时代"、"信息时代"等这些称号相比毫不逊色。30 年来微加工给微电子学造成的辉煌业绩,

今后也将同样地在集成光学、光电子学、微机子学、材料科学和新能源的开发及利用乃至在仿生材料、器件和生物工程广阔领域里创造出许许多多绚烂夺目的奇迹来。

2 微加工的物理和化学过程

2.1 引言

微加工中应用的粒子束有三类：光子束是不具静质量的粒子束，一般情况下主要呈现波动性，例如激光束和X-射线；电子束质量较轻，与质子的静质量相差3个量级，它的能量可以很高，多数情况下主要呈现微粒性；离子束包括原子束和分子束，它们是区分不同物质的最小单元，因而在微加工时有可能同时充当新的材料来源。

粒子束加工能力除了决定于粒子的类型，还决定于其能量和波长，这两者有一定的关系，对光子束来说，波长 λ 与光子能量 E_0 的关系为

$$\lambda = hc/E_0 \tag{1}$$

式中：c 为光速，$c = 3\times10^{10}$cm/s；h 为普朗克常数，$h = 4.1\times10^{-15}$eV·s；光子能量 E_0 用电子伏(eV)表示。则 $\lambda = 1.24/E_0(\mu m)$。

具有静质量的粒子束，按德布罗意实物粒子波的公式，其波长与能量的关系可表示为

$$\lambda = hc/\sqrt{E_0^2 - m^2c^4} \tag{2}$$

式中：m 为粒子的静质量。当粒子速度 v 甚小于光速 c 时，上式可简化为

$$\lambda = h/\sqrt{2mE_0} \tag{3}$$

对电子束，代入物质常数，则

$$\lambda = (1226/\sqrt{E_0})\times10^{-6}(\mu m) \tag{4}$$

对质子束

$$\lambda = (28/\sqrt{E_0})\times10^{-6}(\mu m) \tag{5}$$

表1列出不同能量的光子束、电子束和质子束所对应的波长。由表可知，在可见光波段(1.6～3.6eV)，其分辨率可小到1μm，因而其实用的图形尺寸为数微米。提高分辨率的途径是采用从远紫外到软X-射线(5～1000eV)的光子束，当光子能量超过1keV时，由于散射光子的射程增加，以致在微加工中的应用受到限制。电子束应用的能量范围一般在 $10^2\sim10^5$eV，其对应的波长在 $10^{-4}\sim10^{-5}\mu m$，故对微加工来说不存在电子束波长限制问题，实际所能达到的分辨率将决定于散射电子的射程。对离子束来说，由于质量更大，即使能量很低时，也没有波长限制。由于离子的尺寸与构成固体晶格的原子相近，与电子束相比，离子受到晶格阻挡，散射离子的射程小。

电子、离子及光子与固体表面作用时所产生的物理和化学过程与固体的表面状况有密切关系。固体表面附近的区域其原子的几何排列、电子结构以及化学成分可能与体内有较大差别。在大气条件下几乎所有的表面都被某种吸附层所沾污。既使在133μPa的真空中，由溅射或晶体断裂所造成的清洁表面，在1s时间内就能由气相的分子形成一层完整的单分子覆盖层(10^{15}原子/cm^2)。因此，若要研究原子级的清洁表面，需保持133μPa以上的超高真空。

根据电子衍射和场离子显微镜的研究，晶体表面的结构具有不同的情况如图4所示。图

表 1　不同能量的粒子束波长

E/eV	λ/μm		
	光　子	电　子	离　子
10^0	1.24	1.23×10^{-3}	2.87×10^{-5}
10^1	1.24×10^{-1}	3.88×10^{-4}	9.07×10^{-6}
10^2	1.24×10^{-2}	1.23×10^{-4}	2.87×10^{-6}
10^3	1.24×10^{-3}	3.88×10^{-5}	9.07×10^{-7}
10^4	1.24×10^{-4}	1.23×10^{-5}	2.87×10^{-7}
10^5	1.24×10^{-5}	3.88×10^{-6}	9.07×10^{-8}
10^6	1.24×10^{-6}	1.23×10^{-6}	2.87×10^{-8}

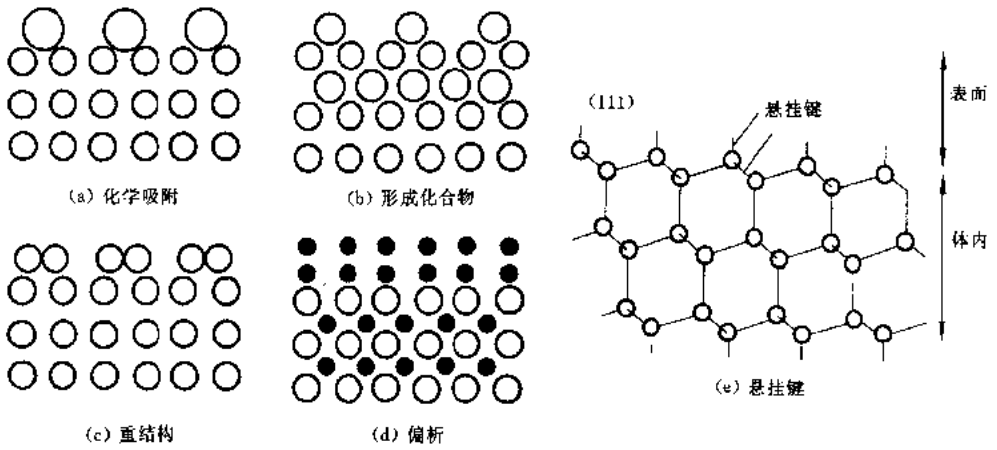


图 4　晶体表面结构的几种情况

4(a)表示晶体表面被不同的外来原子层所覆盖。也可能所吸附的活性物质与晶体表面原子形成三维表层化合物(b)。另一种情况会出现表面上原子的位移,形成新的横向周期性结构(c)。在合金的情况下有时会在表面出现某一种成分的偏析(d)。

表面的化学成分通常都与体内不同。例如,所有的金属表面几乎都有一层天然的氧化层,其厚度可在 1～10nm。如果不把这氧化层通过刻蚀或化学抛光去掉,则将会很大程度上影响其表面的化学和物理性质。一般来说,表面沾污包括吸附的气体、水、有机物和无机离子等。

晶体表面的电子结构与体内也有很大差异。有些电子轨道在晶体内部形成化学键而在表面没有键合,这些轨道其方向从表面向外伸出,称为"悬挂键"(dangling bonds),如图 4(e)所示。许多半导体如 Si、Ge 和 GaAs 等都存在这种情况,晶体内的化学键由相邻原子的两个电子共有化所形成,而在表面上留下了悬挂键,它只有一个电子。对应于这些悬挂键,造成了满的和空的两种电子表面态能级,它们完全不同于体内的电子能级。这种表面态对能带弯曲及与功函数有关的效应产生重要影响。固体的体与表面性质的差别之所以必须引起注意,原因是它们将影响到电子、离子和光子束对表面的轰击过程。下面我们对电子、离子和光子束与固体表面的相互作用分别作简要的阐述。

2.2　电子对固体的作用

当具有一定能量的电子束作用到固体薄膜样品时,如果薄膜的厚度比电子在靶材料中的射程小得多,有三种情况可能出现:电子束未经散射穿透薄膜;电子束穿透薄膜时产生弹性散射;电

子束遭受非弹性散射。三种情况下电子束穿过薄膜后的角度及能量分布如图 5 所示。

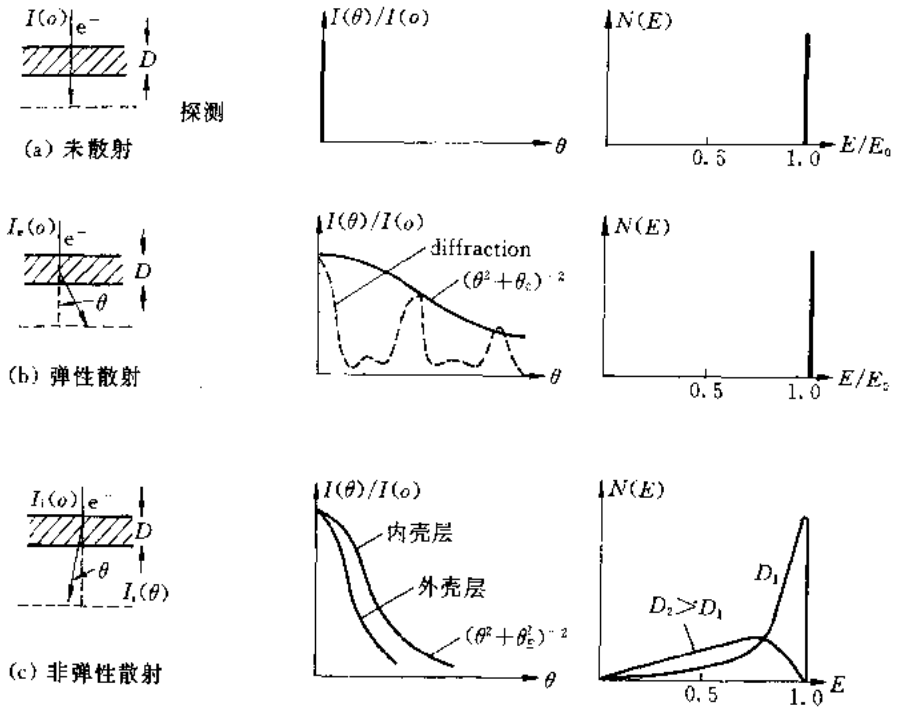


图 5　电子束穿过薄膜后的角度及能量分布

弹性散射主要是由核外电子云作用所引起。某些电子也可能受到卢了的散射，并有很小的能量损耗(1 ~ 10meV)，这与弹性散射电子无明显区别。弹性散射电子的偏角分布可用 Rutherford 散射公式计算。

设入射电子束为 $I(o)$，在 θ 角方向的散射束流为 $I(\theta)$，则散射束流的角分布为

$$I(\theta)/I(o) = (\theta^2 + \theta_0^2)^{-2} \tag{6}$$

式中 $\theta_0 = \lambda/2a\pi$ 为特征屏蔽角，a 为原子半径；λ 为入射电子波长。式(6)表示的角分布对非晶薄膜近似正确。但对结晶薄膜将受衍射效应的调制，衍射峰出现在满足 Bragg 条件 $\theta \approx \lambda/d$ (d 为晶格常数)的角度，如图 5(b)所示。由于电子受到散射，电子束的直径将随其在固体中进入的深度 z 而扩展。假设电子束中电子的空间密度呈高斯分布，可以得到如下公式：

$$r^2 = r_i^2 + \frac{4z^3}{3\delta} \tag{7}$$

式中：z 为与固体表面垂直的空间坐标；r_i 为 z 处入射电子束的高斯半径；r 为电子束经过厚度 z 受散射后的高斯半径；δ 为电子散射的平均自由行程。

在非弹性散射的情况，散射角 θ 取决于入射电子的能量损耗，其能量分布及角度分布曲线如图 5(c)所示，可表示为

$$I_i(\theta)/I_i(o) = (\theta^2 + \theta_E^2)^{-2} \tag{8}$$

式中 $\theta_E = E/pv$，E 是速度为 v，动量为 p 的入射电子的能量损耗。

电子束在固体表面和进入体内由非弹性散射引起的能量损耗有许多不同的过程，包括：产生低能(直到 50eV)二次电子，激发金属电子等离子体中的密度振荡，内壳层电离引起的 X-射

线和俄歇电子发射;产生电子空穴对并随后产生光子发射(阴极发光)、跃迁辐射及激发声子(晶格振动)等,如图6所示。

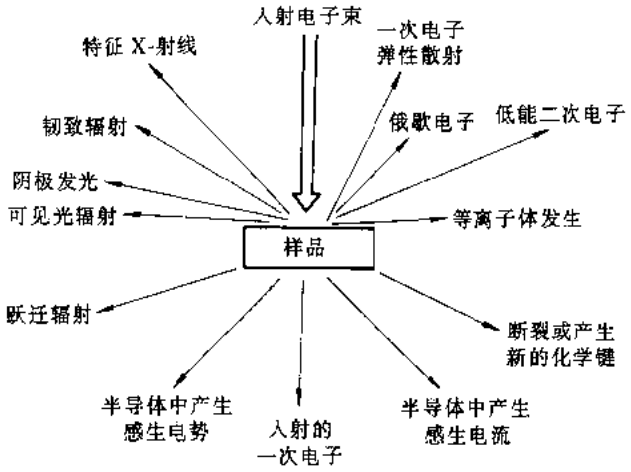


图6 微加工中产生的电子与靶的相互作用

所有非弹性散射所造成的能量损耗,一直沿用Bethe连续能量损耗关系式来描述:

$$\frac{dE}{ds}=\frac{N_A e^4}{2\pi\varepsilon_0^2}\frac{Z\rho}{A}\frac{1}{E}\ln\frac{1.66E}{J}=$$

$$-7.85\times10^4\frac{\rho Z}{AE}\ln\frac{1.66E}{J}\quad\left(\frac{\text{keV}}{\text{cm}}\right)\tag{9}$$

式中:dE 为具有能量 E(keV)的电子通过距离 ds 后的能量变化;Z 为原子数;A 为相对原子质量;ρ 为密度;J 为平均电离势,近似地随原子数正比增加。从式(9)可知,由于 Z/A 接近于常数,所以电子在行程上能量损耗速率随 ρ 增大而增加,随电子能量 E 的增大而减小。

对式(9)积分可以计算电子进入固体的射程,积分从 $s=0$ 开始,$E=E_0$,最大的距离为电子失去全部能量($E=0$)而停止下来:

$$R_B=\int_{E_0}^{0}\frac{dE}{dE/d(\rho s)}\tag{10}$$

式中 R_B 称为Bethe射程,这里用 ρs 表示可以把密度的变化考虑在内。电子在固体内实际达到的射程一般要比 R_B 小,因为在 R_B 的计算中没有考虑弹性散射。在 $E_0=1\sim100$keV 范围,实际的电子射程可用经验式近似表示:

$$R=10E_0^{1.43}\ (\mu g/cm^{-2})\tag{11}$$

电子束进入固体内产生的各种二次效应都将对电子射程和空间分辨率产生直接影响(见图7)。它表明由于背散射电子的横向扩展,成为提高电子束加工分辨率的主要

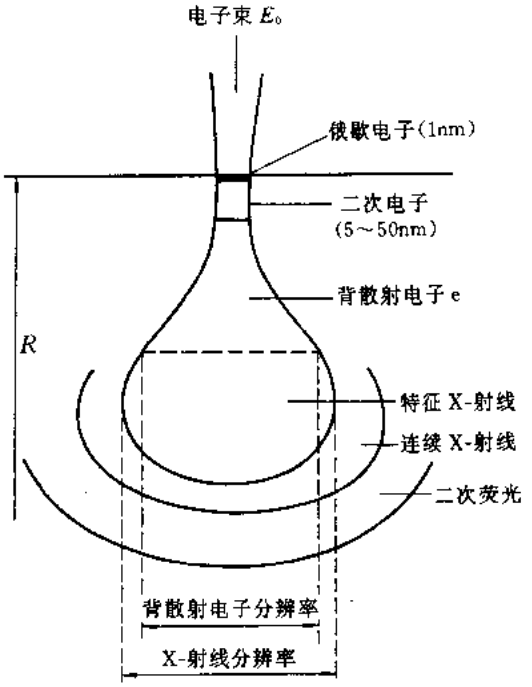


图7 电子束在固体靶内由二次效应引起的射程和空间分辨率

限制。

图 6 中所示的许多二次效应在微加工中可以应用。首先，电子束作用于有机薄膜造成化学键断裂使分子降解或引起交链形成大分子，这一效应已成功地用于电子束曝光；其次，电子束能量损耗产生的热效应可用于大面积或微区的退火；另外，俄歇电子、感生电势、电流和各种辐射可用于对材料和器件的成分、结构、形貌和特性等的表征和测定。

2.3 离子对固体的作用

荷能(1～100keV)离子束入射到固体表面所引起的物理和化学效应可用图 8 中的 10 种基本过程表示：(1)入射离子受到固体表面单个原子或原子团的背散射，通常这过程导致入射离子路径的改变，并在离子和靶原子之间发生能量交换。此能量交换可能是弹性的或非弹性，这决定于固体的组成粒子和离子的能量。(2)离子的能量足以把表面原子从晶格中束缚弱的位置移至束缚强的位置重新排列，这过程称原子位错。(3)能量较高的离子可引起样品体内的位错。(4)物理溅射是当离子轰击样品表面时有足够的动量转移引起一个或多个原子完全自由从而脱离表面。(5)离子注入是入射离子进入样品晶格当其能量耗完后被俘获。(6)化学溅射是由于离子能与表面原子产生化学反应，结果在固体表面生成新的化合物，能像气体分子那样从表面逸出，也称为反应溅射。(7)轰击的正离子能从表面获得一个电子而复合，并以中性原子形成反射出去。(8)离子被样品表面所束缚成为吸附离子。(9)离子轰击金属表面时在适当的条件下出现二次电子发射。(10)表面原子受激到电离态，引起二次离子发射。

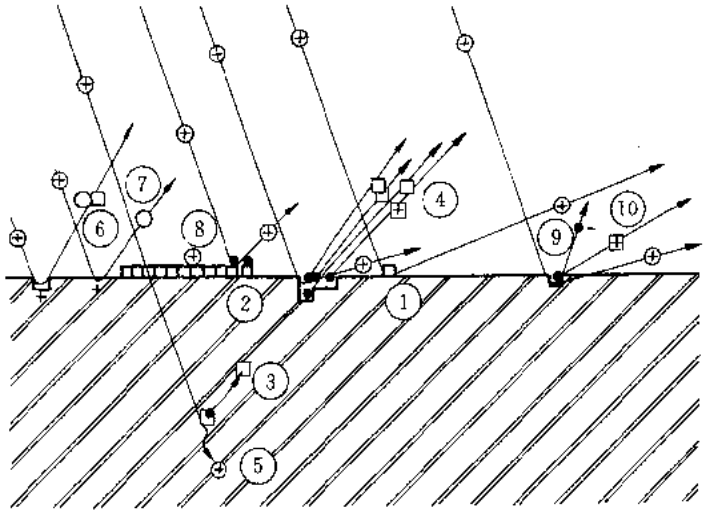


图 8 离子对固体的相互作用

① 离子-原子散射； ② 表面位错； ③ 体内位错； ④ 物理溅射；
⑤ 离子注入； ⑥ 化学溅射； ⑦ 电荷转移； ⑧ 离子吸附；
⑨ 电子发射； ⑩ 表面原子电离发射

入射离子把能量传递给固体的过程中其速度逐渐减小。在分析其能量损耗时，通常主要区分为电子碰撞和核碰撞两部分。另外，入射离子和靶原子之间的电荷交换也引起能量损耗。因此，总的能量损耗可表示为

$$dE/dz = (dE/dz)_n + (dE/dz)_e + (dE/dz)_{eh} \tag{12}$$

式中等式右边三项分别表示由核、电子和电荷交换引起的能耗。高能快离子的能耗主要由电子阻止所引起，与晶格电子云相互作用并引起激发和电离。由于电子密度高，产生碰撞频繁，这类似于电子的能耗，可看作为连续过程。核阻止的碰撞频率较低，低能离子的能耗主要由核阻止所引起，是造成离子束角分散的主要原因。一项有用的经验规则是当离子的能量 $E < A$keV 时，能耗主要由核阻止引起，A 为入射离子的相对原子质量。在中间能量范围，由电荷转移造成的能耗增加，但约占总能耗的 10%，一般可以忽略。各种能耗与离子能量的关系如图 9 所示。

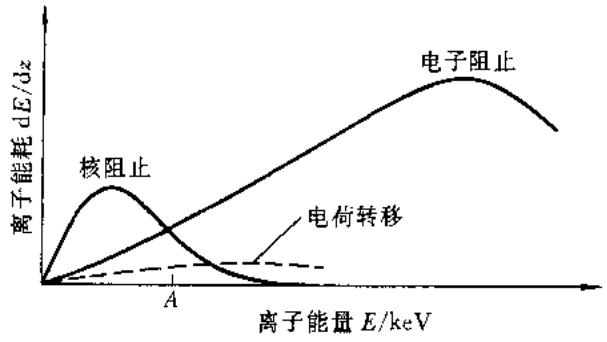


图 9　不同能耗与离子能量的关系

在计算离子射程时，可以假设电子阻止和核阻止过程是相互独立的，并略去电荷转移的影响：

$$R = \int_0^R dz = \int_{E_0}^0 \frac{dE}{(dE/dz)_n + (dE/dz)_e} \tag{13}$$

当核屏蔽势场采用方阱模型，可得近似式：

$$\left(\frac{dE}{dz}\right)_n = 278 \frac{Z_1 Z_2}{(Z_1^{2/3} + Z_2^{2/3})^{1/2}} \cdot \frac{M_1}{M_1 + M_2} N \ (\mathrm{eV \cdot nm^{-1}}) \tag{14}$$

式中：Z_1 和 M_1，Z_2 和 M_2 分别表示离子和晶格原子的原子序数及质量；N 为原子密度（原子数/nm^3）。电子阻止可采用 LSS 模型，假设离子的能耗与其速度成正比：

$$(dE/dz)_e = k' k^{1/2} \tag{15}$$

式中 k' 决定于离子靶和原子的物质常数：

$$k' = 3.28(Z_1 + Z_2) M^{-1/2} N \quad (\mathrm{eV^{1/2} \cdot nm^{-1}}) \tag{16}$$

把式(14)、(15)代入式(13)，积分后可得：

$$R \cong 2kE_0\left(1 - \frac{4}{3} kk' E_0\right) \tag{17}$$

式中

$$k = \frac{0.0018}{N} \cdot \frac{Z_1^{2/3} + Z_2^{2/3}}{Z_1 Z_2} \cdot \frac{M_1 - M_2}{M_1} \ (\mathrm{nm \cdot eV^{-1}}) \tag{18}$$

从表示离子射程的公式(17)不难看出，离子射程随原子序数 Z 和密度 N 的增加而减小。

离子入射到晶体时，在一定条件下出现沟道效应。晶体中的原子在空间作周期性的有序排列，使入射离子与这些原子产生相关碰撞，当这些相关碰撞沿着某一结晶方向时，离子可像

通过沟道一样深入晶格内部，如图 10 所示。晶体基片沟道效应的特性，决定于每个入射离子的方向角，也与离子和基片的特征有关。如果入射角 Ψ 大，振荡的幅值大，离子不能沿沟道通行，如图 10 中的轨迹(a)。要使离子能保持在沟道内通行，如轨迹(b)和(c)，存在一个最大允许幅值，Ψ_c 是对应于这一轨迹的临界角，可近似表示为

$$\Psi_c = \left(\frac{2Z_1 Z_2 e^2}{4\pi\varepsilon_0 Ed}\right)^{1/2} \tag{19}$$

由此式可知，离子能量 E 增加时，Ψ_c 减小，保持稳定的沟道轨迹较为困难；相反，原子堆砌密度高，即 d 减小，则 Ψ_c 增大，容易出现沟道效应。图 8 所表示的各种作用过程在微加工中有许多应用。如离子注入、离子束曝光、离子束淀积、溅射刻蚀、溅射淀积、表面处理和表面分析等。这些微加工技术留待下节讨论。

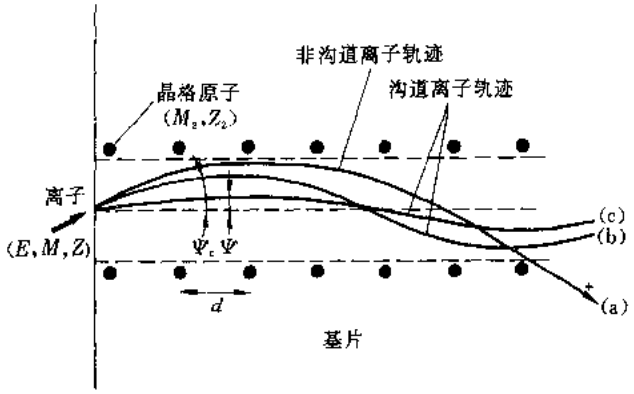


图 10　离子在晶体中的沟道效应

2.4　光子对固体的作用

微加工应用的光束主要有三个波段：330～430nm 的紫外波段；150～300nm 的远紫外波段；0.5～5nm 的软 X-射线波。光子束在微加工中主要用于制作掩膜版和图形转印的曝光过程，因而多数情况下是与称为抗蚀剂的光敏膜相互作用，只有当抗蚀剂中的光敏基团的化学键能与入射光子的能量相对应时，才能有效地引起化学键的断裂或交联。图 11 表示普遍应用的紫外光源高压汞弧灯的发射光谱曲线 $S(\lambda)$及几种常用的正性和负性抗蚀剂的光谱灵敏度曲线 $R(\lambda)$。如果光学系统的传输特性为 $F(\lambda)$，则达到抗蚀剂表面的光谱强度为 $S(\lambda)F(\lambda)$。再计及抗蚀剂中的光化学反应，则抗蚀剂表面在 $\Delta\lambda$ 波长范围内的有效光强将正比于

$$S(\lambda)F(\lambda)R(\lambda)\Delta(\lambda)$$

随着对图形线宽不断缩小的要求，这紫外波段的利用受到了重视。但由于光源和效率等因素，限制了它的应用范围。近年来，激光束技术的进步，特别是波长在紫外和远紫外范围的准分子(excimer)激光器的出现，使这种高功率和高效率的相干光子束对固体表面的热效应和光化学反应，在微加工中得到广泛的应用。包括曝光、激光诱导 CVD、激光刻蚀、表面掺杂和激光退火等。目前较为常用的准分子激光器有 XeCl(308nm)、KrF(248.4nm)和 ArF(193nm)等。

X-射线的光子具有较高能量，对固体的作用主要有三种过程：光电子过程、Compton 效应及生成电子对。完整地处理这些过程相当复杂，要求应用量子电动力学。但主要的实验结

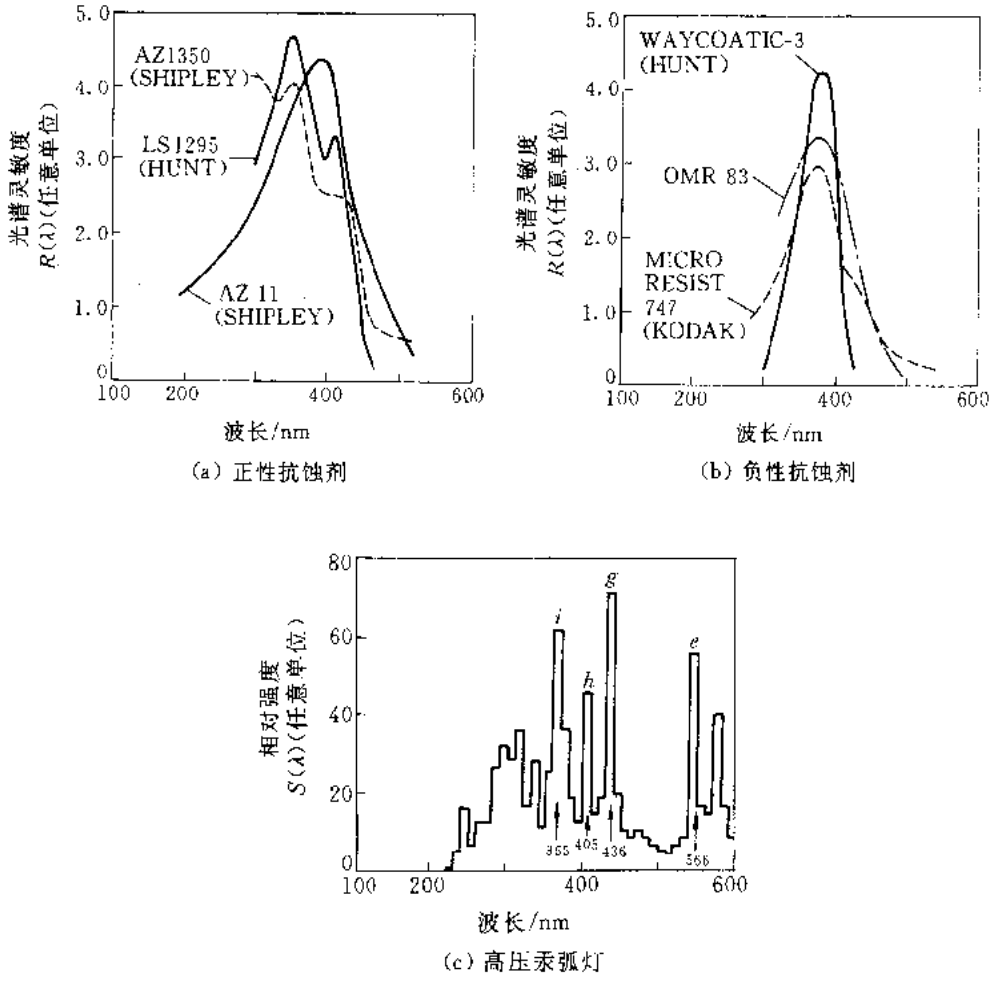


图 11　高压汞弧灯的发射光谱及几种正性和负性抗蚀剂的光谱灵敏度曲线的比较

果是简单的。在光电子过程中入射光子被原子所吸收，从而引起壳层中的电子发射。所谓 Compton 效应是指光子受到原子核外的散射。另一过程是入射光子被转换成电子和正电子对，这过程在自由空间是不可能出现的，因为当一个光子蜕变成两个有质量的粒子时，能量和动量不能同时守恒。但这过程能在核的 Coulomb 场中发生，因为这里能量和动量能够平衡。

这些过程的能量关系差别甚大。在小于数 keV 低能范围，光电子效应将是主要的，Compton 效应很弱，此能量不可能生成电子对。能量增加到 1MeV 就可能生成电子对并很快就占主导地位。三种过程中的两者，即光电子效应和生成电子对在相互作用时光子本身将消失。在 Compton 效应中，光子经散射后其能量递减。

光子束进入固体后的传输特性可用指数式描述：

$$N^{(z)} = N(o)e^{-\mu z} \tag{20}$$

式中 μ 是吸收指数，应是三项的和：

$$\mu = \mu_{\text{photo}} + \mu_{\text{comptor.}} + \mu_{\text{pair}} \tag{21}$$

传输的粒子数将按指数规律减小，不能确定它的射程。但是，粒子碰撞前通过的距离，即平均自由行程等于 $1/\mu$。

图 12 表示当入射光子能量在 $10^{-2}\sim10^{8}$eV 范围内单晶硅内光子吸收指数 μ(单位:cm^{-1})与光子能量的关系曲线。能量最小的光子能被晶体内的晶格声子所吸收。由杂质引起的吸收也可以在1eV下,这在图中没有表示出来。光电吸收开始的能量相应于越过带隙(E_g)产生激发的能量。能量接近于 E_g 的光激发电子将留在固体内,这过程称内光电效应,在高能量时,大约相当于自由原子的电离能,产生的光电子能离开固体,这叫外光电效应。能量在 E_g(1.2eV)附近的吸收截面数值决定于对固体能带结构的灵敏度。光子能量较高时,光电吸收是主导过程,直到出现 Compton 散射(>10eV)和生成电子对(>10eV)。

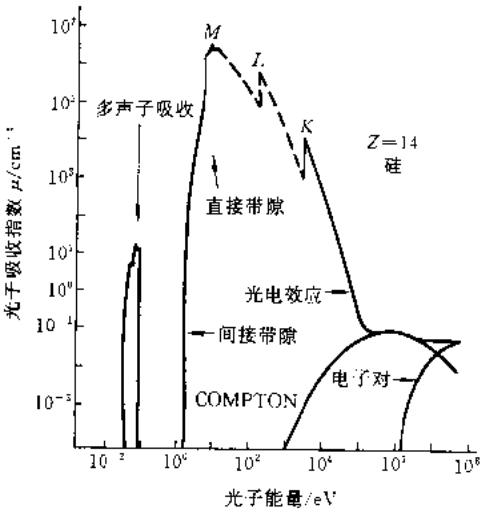


图 12 单晶硅中 X-射线的吸收能谱特性

表 2 列举了微加工中应用的主要几种粒子束的特征参数及与微加工有关的相互作用。

表 2 在微加工中应用的粒子束

粒子束类型	波长 λ/nm	射程	偏转特性	与物质的相互作用
光子 ($h\gamma$)	200~400	$I=I_0e^{-UZ}$	光学元件(透镜、反射镜)	化学键断裂和聚合反应(曝光);产生电子;热效应(退火);形成等离子体;几何对准;光子辅助下的淀积、外延、刻蚀和表面改性等
X-射线 ($h\gamma$)	0.2~5	决定于光电子射程	晶体;X 射线光学元件	化学键断裂和聚合反应(曝光);材料分析;对准
电子 (e^-,m_0)	~0.01	$10E_0^{1.43}(\mu g/cm^2)$ (E_0 以 keV 为单位)	电场和磁场	化学键断裂和聚合反应(曝光);产生二次电子;发射 X-射线;形成等离子体,感生电势和电流;热效应(退火),分析和形貌观察
离子 (e^+,M)	0.001	$2kE_0\left(1-\frac{4}{3}kk'E_0\right)$	电场和磁场	化学键断裂和聚合反应(曝光);溅射;淀积;刻蚀;离子注入;表面处理;分析和诊断
原子 (A,Z)	0.001	$2kE_0\left(1-\frac{4}{3}kk'E_0\right)$	梯度场 当 $\mu,p\neq0$	溅射;淀积;晶体生长;分子束外延

3 微加工技术

3.1 引言

现阶段微加工技术通常包括以下四方面:①薄膜技术:薄膜淀积,外延生长和表面改性;

②图形技术:图形发生和转印; ③刻蚀技术:除去不需要的部分,形成三维结构; ④诊断技术:结构和成分的测量和分析。这四方面技术是获得各种功能的微结构器件不可缺少的组成部分。

所有的微结构器件往往是薄膜的或是表面型的,作为器件在 z 方向(一维)加工的第一步,是在基片表面生长薄膜。而且,在器件加工的全过程中,可能用不同方法生长不同的薄膜,包括不同材料、不同厚度、结晶或非结晶、单层结构或多层结构等。图形技术是器件在 xy 二维平面方向产生实现器件功能所需的各种特定图形。通常先把图形做在掩膜版上,然后再转印到基片表面,这与摄影时先拍底片而后印成照片相仿。现在已能采用计算机控制的电子束、离子束等曝光技术把图形直接做在基片上而不再需用掩膜版,称为无掩膜技术,这不仅可以简化工艺过程,而且可减少加工误差。基片上生成的图形一般是由抗蚀剂构成的,刻蚀过程是按照已形成的抗蚀剂图形对基片表面或基片上的薄膜进行加工,把没有抗蚀剂保护的那些不需要的部分除去。对结构化比较复杂的器件,往往要把薄膜生长、图形制备和刻蚀三种加工过程多次交替使用。今后采用无掩膜技术,特别是无掩膜微区选择生长和刻蚀,有可能将上述三步加工过程一次完成,并使器件缩小到分子尺寸,逐步实现分子工程的期望目标。测试技术是微加工中不能忽视的重要方面,它对保证器件结构尺寸精确、材料成分准确及性能稳定可靠是不可缺少的,许多大型精密工艺设备中就附有许多分析测试装置,如分子束外延系统等。本节将就这四方面技术分别作一综述。

3.2 薄膜技术

薄膜材料是制造结构器件的基础,因而薄膜生长在微加工中占重要地位。不同的器件对膜厚的要求差别很大,可以从不到 1nm 的单分子直到几个光波波长、数微米或更大厚度。同时,薄膜的表面和界面状况、晶体构造和晶向排列、化学成分和膜层结构以及各种物理性能等都对器件的功能有直接影响。

薄膜形成的过程主要可分为三类:淀积膜、外延膜和表面改性。淀积膜与基片之间有明显的界面,例如在半导体基片上沉积金属膜或介质膜,膜层与基片的材料组成不同,可以是结晶的或非晶的。制备方法有真空蒸镀、溅射淀积、电离团束淀积、电镀和涂覆等。外延膜与基片之间有相同或非常接近的晶格结构,膜层的晶格通常是基片晶格的延伸。膜层与基片的材料的组成可以相同,例如硅片上外延硅;也可以不相同,如在 GaAs 基片上生长 GaAlAs 异质结。通常前者的界面不明显,而后者具有突变的界面。外延技术有气相外延(VPE),也称化学气相淀积(CVD);液相外延(LPE);分子束外延(MBE)和金属有机化学气相淀积(MOCVD)等。表面改性是通过基片的表面化学反应,如硅片氧化生成 SiO_2,或其他过程,如扩散、离子注入和离子交换等在基片表面形成化学组成、材料结构和性能参数与基片体内有明显差别的膜层,其特点是整体性好,但不易获得突变的界面,往往存在一定厚度的过渡区。

一般来说,在固体基片表面生成新的膜层,通常都是成膜的粒通过某种媒质作为载体与固体表面紧密接触而造成吸附或产生化学反应生成新化合物。这些成膜粒可以是原子、分子、离子,带电或不带电的原子团或分子片段等,载体媒质可以是固体、液体、气体或真空。

在固体作为载体的情况下,载体与基片间不易达到紧密接触,这就限制了应用范围。一个普通的例子是在制作扩 Ti 铌酸锂光波导,首先在基片表面淀积一层 Ti 膜,以保证其紧密接触。液体载体应用较广,因为许多化合物能溶解成溶液并能与基片表面良好接触,在以离子或其他

带电粒子状态溶解时，可施加电场以增加成膜质粒的迁移速度。根据粒子的平均自由行程可判断气体载体和真空载体。当气体载体具有相对高的压力时，成膜粒子再到达基片表面前往往与载体分子发生一系列碰撞，而在真空情况下，成膜粒子从源到基片表面的途中与残气分子碰撞几率极小。成膜粒子可以预先混合、悬浮或溶解于载体中，在淀积过程中逐渐消耗。成膜材料也可以从源连续地添加到载体中去。

因此，薄膜制备过程一般有如下几方面特征以资区别：

(1) 载体媒质(固体、液体、气体、真空)；

(2) 成膜质粒的种类(原子、分子、离子、晶粒)；

(3) 成膜质粒引入载体的方法(预先混合、溶解、蒸发、荷能粒子轰击靶表面等)；

(4) 表面反应(单纯凝聚、化学反应、电化学反应、注入等)；

(5) 成膜质粒从源到基片表面的输运机理(自由飞行、气相扩散、液相扩散)。

微加工中常用的薄膜制备方法及其特点列于表3。

表3 微加工常用薄膜生成方法及其特点

成膜方法	实 例	载体	成膜粒子	引入载体的方法	表面反应	输运机理
真空蒸镀	半导体上 Al 电极	真空	原子	源材料加热蒸发	凝聚	扩散
溅射	Si 上淀积 Au	气体	原子	载体正离子轰击	凝聚	扩散
电离团束淀积		真空	原子团	源材料加热蒸发部分电离	凝聚	电场加速及扩散
电镀	Cu 镀在 Fe 上	液体	离子	从铜电极电解	离子复合	电漂移
涂覆	抗蚀剂	液体	有机分子	溶解	溶剂挥发	液体流动
化学气相淀积	外延硅片	气体	分子	预先混合	表面化学反应	扩散
液相外延	InGaAsP/InP	液体	原子/分子	自身作载体	晶体生长	扩散
分子束外延	GaAlAs/GaAs	真空	原子/分子	源材料加热蒸发	凝聚	自由飞行
金属有机化学气相淀积	GaAlAs/GaAs	气体	分子	预先混合	表面化学反应	扩散
氧化	硅上生长 SiO_2	气体	原子	预先混合	表面化学反应	扩散
离子注入	B 注入 Si	真空	离子	等离子体或离子枪	注入	自由飞行
扩散	Ti 扩散 $LiNbO_3$	固体	原子	从基片表面扩散	扩散	扩散

薄膜生成过程和机理在理论上是一个十分复杂的问题。迄今为止对成膜过程的描述和模型尚停留在定性的讨论，还不能进行精确的定量计算。由电子显微镜观察和研究的结果，可认为从蒸汽源凝聚的薄膜生长过程经历了如下几个阶段：

(1) 成核：到达基片表面的原子或其他质粒凝聚成小核，直径约为 0.5nm，按统计规律分布于整个表面；

(2) 核生长：后来原子的不断加入使核从小变大，成为三维状的岛，常常具有微晶结构；

(3) 岛聚结：随着岛的长大，原来分离的岛互相聚结形成网络，但中间仍留有空的通道；

(4) 通道填满：形成连续的膜层。

薄膜生成过程的不同阶段可用图 13 示意。

当一个粒子碰撞固体表面时，可以发生各种不同的相互作用，它取决于表面原子与碰撞粒

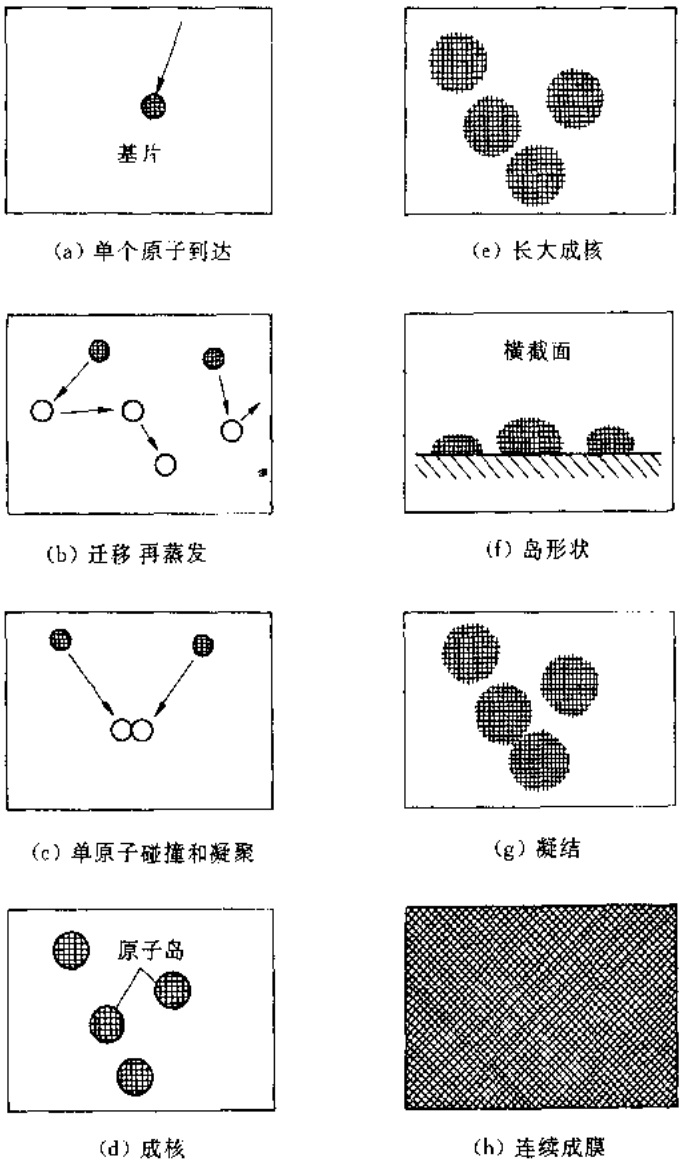


图 13　薄膜生成过程

子间的结合能。如,晶格表面的不对称引起的表面电场,碰撞粒子的极化率,基片和薄膜材料的晶格常数,基片温度以及碰撞粒子的能量等。

一般来说,我们可以假设碰撞粒子落到表面后,由于表面电场造成的极性力而留下。它过剩的能量使它能沿表面移动一定距离,直到它将能量损失于晶格而被束缚在固定位置上,或从晶格获得足够能量溢出表面电场的作用再蒸发。粒子在表面移动的途中,可以由某种弱化学键作用暂时受束缚。它在弱键位上停留的时间 τ' 取决于迁移活化能 ϕ_m 及表面温度 T,具有如下关系:

$$\tau' = \tau_0 \exp\left(\frac{\phi_m}{kT}\right) \tag{22}$$

粒子停留在表面的总时间 τ 由脱附活化能 ϕ_D 所决定：

$$\tau = \tau_0 \exp\left(\frac{\phi_D}{kT}\right) \tag{23}$$

通常 ϕ_D 比 ϕ_m 大得多。例如 Ba 在 W 基片表面的 ϕ_m 为 0.65eV，ϕ_D 为 3.8eV。当基片温度很高时，由于时间 τ' 很短，吸附粒子表现为二维气体。中等温度时，τ' 取某有限值，吸附粒子随机地从一个位置跳到另一个位置。低温下，τ' 可以非常长，粒子很快失去其过剩的能量几乎被永久束缚。Ba 在 W 表面的功函数低，在这种情况下，采用热离子发射显微镜可以显示 Ba 原子在 W 表面的运动的状况。

具有动能比基片温度高的碰撞粒子通常能被冷的表面捕获并沿着表面运动（如二维气体），直到它可能释放出这剩余能量凝聚成固体。只有存在成核的位置，且让这位置上吸附的粒能放出其剩余能量，凝聚才可能出现。如果碰撞粒子的供应超过其脱附速率，膜层成为过饱和，这时更容易发生凝聚。

开始出现凝聚的条件决定于脱附能 ϕ_D 与冷凝薄膜的升化热 ϕ_S，如果 $\phi_D \ll \phi_S$，凝聚出现时没有过饱和；如果 $\phi_D \approx \phi_S$，凝聚要求中等程度的过饱和；如果 $\phi_D > \phi_S$，只有在高度过饱和的条件下，才能开始凝聚。显然，当薄膜开始形成后，将立即达到 $\phi_D = \phi_S$。

凝聚过程的理论模型通常假设当吸附的原子相碰时，有一定的几率结合成对，原子进一步与原子对碰撞的结果形成更大的聚集体。与此同时，也有聚集体分解的逆过程。根据表面张力理论（capillary theory）[6]，在含有大量原子（>100）的聚集体情况下，可采用热力学方法加以分析。分析表明，当聚集体半径超过某一临界值 r^* 时，聚集体可以稳定存在，

$$r^* = \frac{2\sigma_{er} V}{kT \ln P/P_e} \tag{24}$$

式中：T 为表面温度；P 为非饱和蒸汽压；P_e 为其吸附状态的平衡压力；V 为从 P 到 P_e 凝聚的分子体积；σ_{er} 是对应于 V 的凝聚相与蒸汽界面的自由能。

上述理论针对非球形的情况作了修正，这是由于聚集体在基片表面受到表面张力所造成的。当核只包含少数几个（1～10）原子时，需要考虑到每个原子相互间以及与晶格间结合的影响。一种处理方法是对上述理论就这一影响进行修正；另一种途径是用统计物理方法分析聚集体的形成。这些理论在一定程度上与实验观察定性相符合。如果理论模型中的假设能更好地接近实际，今后就有可能使理论与实验值定量一致。

成核及长大的过程已被电子显微镜的观察所证实。在蒸发进行中，随机分布并互相分离的岛（直径 >0.5nm）大量增加，直到岛密度达饱和值，其范围为 $10^{10} \sim 10^{12}$ 原子/cm^2，这时岛与岛分开的距离为 10～100nm。随着蒸发过程的进一步发展，岛逐渐增大，岛密度由于相互聚结而减少。成核位置的密度能通过外界因素增加，如用电子轰击基片，成核也择优地出现在结晶表面的位错处、杂质存在的位置以及其他的晶格不规则处。这一效应可以用来显示单晶中的位错情况。

由溅射源造成的碰撞粒子的有效温度较蒸发源出发的高许多，因而能用来获得外延膜。在较低温度的基片上的成核密度也较高，这是因为大多数溅射系统中有电子和离子轰击表面的结果，采用反应方法淀积薄膜的成核过程显然就更为复杂，迄今尚缺少详细研究。

当基片上的成核位置都被占满后，在这些位置上的核开始长大成三维的岛。根据基片温度的高低，岛可以呈液滴状或成单晶。岛的熔点比体材料的熔点 T_m 低，实验结果为 $2/3\,T_m$。

假设岛的聚结只是出现在岛相互接触时，聚结有热量释放，所以许多材料的固体岛在接触时熔化。岛聚结后它们冷却下来形成新的晶体，其结晶方向往往与两个岛中较大的一个岛的原有结晶方向一致。

在单晶基片上生长外延膜或定向的单晶膜时，可以假设大多数成核岛的方向是相似，而且在聚结时的再结晶也具有择优方向。在实验中可观察到外延膜的结晶方向强烈依赖于基片的晶格结构，由此可以得到肯定。另一方面，也可能在非晶基片上，甚至在液体表面，像单晶基片那样生长定向单晶膜。

对每一个凝聚膜和基片体系来说，在一定的淀积速率下存着一个临界的基片温度，高于这一温度就能生长定向单晶膜，不论晶格失配的程度如何；低于这临界温度凝聚时，膜结构无序性增加，直到温度足够低（$T_m/3$）时，入射原子几乎就在碰撞的位置固定下来，产生非晶（高度无序）薄膜。

评估成膜质量首先是膜层形态和结构的观察和测量。膜层是否还包含三维的岛和存在沟道，或是密度均匀和表面平整的膜。最好的方法是采用高分辨率电子显微镜，但是光学衍射、电导及密度等测量也能提供膜层质量的信息。

要确定薄膜的晶体结构是单晶、单轴定向的多晶、随机定向多晶或是非晶，通常应用 X-射线衍射或电子衍射。非晶膜没有确定的衍射环；随机定向的多晶给出一定的衍射环，从衍射线的宽度可以确定晶体的平均尺寸 a，其关系式为

$$a=\frac{\lambda}{D\cos\theta}$$

式中：λ 为 X-射线的波长；D 是衍射线的角宽；θ 为布拉格角。

单晶膜或单方向取向的多晶膜能产生衍射斑（Laue 图），当多晶的方向分散时，光斑产生畸变成椭圆形，并出现新的光斑。在完全随机的极端情况下，椭圆光斑合在一起形成衍射环。如果电子束可以聚得足够细，成核后期个别岛的晶体结构可以用电子衍射来分析。

对薄膜化学成分的分析可以采用许多不同的方法。首先是常规的化学分析法，将膜层从基片上取下进行分析。这包括容积测定、质量测定和光学摄谱法等。其次是 X-荧光光谱、俄歇电子能谱和质谱分析。当薄膜受到高能电子轰击激发内层电子产生辐射，通过探测器接收可以显示发射 X-射线的能谱。采用聚焦电子束，可以对微区进行分析。当薄膜受到电子轰击发射二次电子（俄歇电子）的条件下，测量和分析其能谱可以对薄膜表面原子层作出评价。把薄膜逐步溅射刻蚀，再进行俄歇能谱分析，可以得到薄膜从表面到基片沿深度方向的成分分布。质谱分析是用荷能离子束轰击薄膜，再把溅射下来的原子经电离后进行质量谱分析。X-荧光分析、俄歇电子能谱和质谱仪都可配合扫描系统对薄膜的整个面积的成分变化进行测定。利用这些现代化仪器，分辨率可达亚微米级，或更小。但是，对含量低于百分之几的杂质除了采用常规分析外，其他方法都难于分析。

3.3 图形技术

虽然薄膜淀积作为器件在一维方向上的尺寸控制可以足够精确，直到1nm或更小，但目前其他二维的尺寸控制还不能达到同样的精度。在基片表面生成二维图形一般由光刻（lithography）过程来实现，这是从石版印刷技术演变而成的。微加工技术面临的最重的任务就是设法减小图形尺寸和提高精度。以 VLSI 为例，90年代生产商品器件要求的线宽的预测见表4。

表 4　90 年代 VLSI 产品预测

器件(DRAM)	设计规则/μm	曝光波长/nm	生产开始年	生产高峰年
1 Mbit	1.0～1.2	436	1988	1991
4 Mbit	0.7～0.8	436 或 365	1990	1993
16 Mbit	0.5～0.6	365	1992	1995
64 Mbit	0.3～0.4	远紫外或 X-射线	1994	1997

注：463nm－g 线，365nm－i 线

现在处于实验室研究阶段的高速电子器件、光子器件、光电子器件以及量子尺寸效应器件要求结构尺寸在 10～100nm，小于 10nm 的图形技术也已列入研究日程。

一般来说，图形是由抗蚀剂涂层的曝光和显影形成的。曝光可以采用光、X-射线、电子束和离子束等不同的源，需用不同类型的抗蚀剂使其敏感性能与曝光源相配合。光刻技术的分类及特征也可根据曝光源的不同分为光学光刻、X-射线光刻、电子束光刻和离子束光刻等。这些不同的光刻方法可用于掩膜版制备(mask making，通常叫制版)、图形转印(printing pattern transfer)和直接刻写(direct writing)，即无掩膜(mask less)光刻。从图形的设计数据开始到在基片上形成图形的各种光刻方法路线如图 14 所示。

图形技术全过程通常由计算机辅助版图设计开始，在早期，采用由计算机控制的绘图机，根

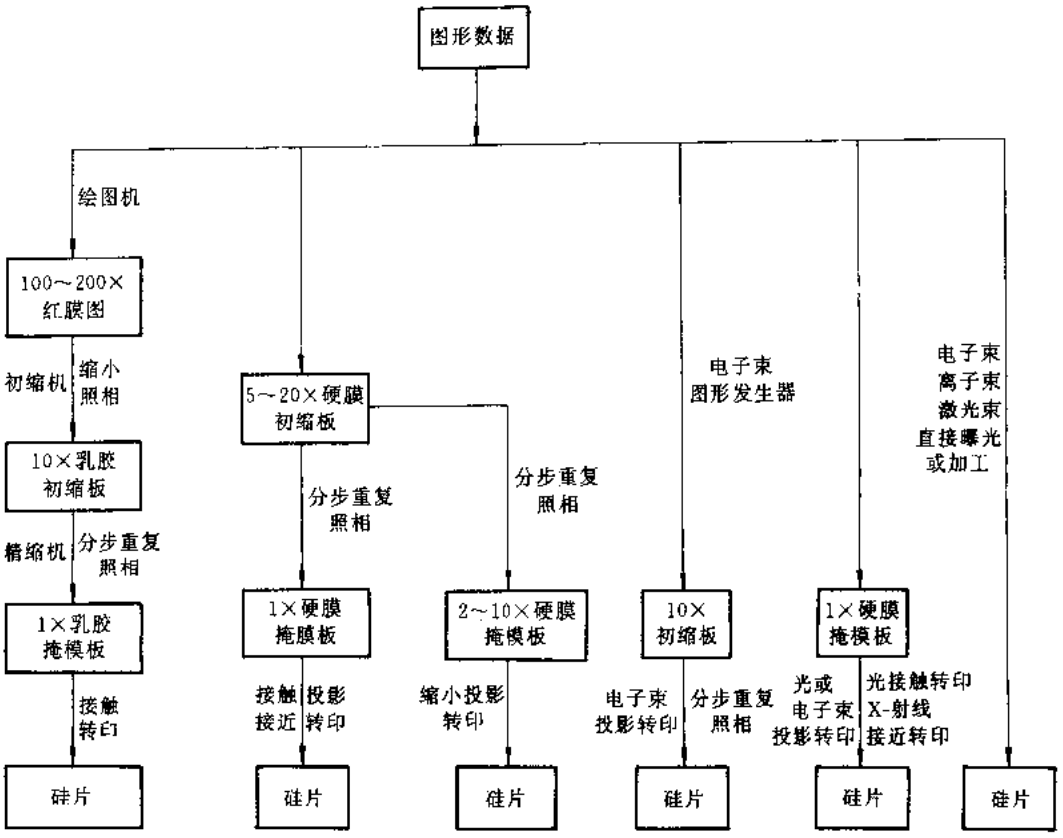


图 14　光刻技术分类图

据输入的设计数据在红模上刻出图形，经照相缩小成 1/2 ~ 1/10 的初缩版(reticle)，再经精缩机分步重复照相缩小到器件实际尺寸的掩膜，这就是制版过程。而后通过掩膜对基片上抗蚀剂曝光使图形转印。这种在红膜上刻图的方法将造成较大误差，而且图形的复杂程度及线宽都受到限制。后来就由图形发生器所代替，它是由计算机控制曝光源的扫描系统形成图形。图形发生器的曝光源有光束、电子束或离子束。目前，X-射线还难以使其聚焦和偏转，尚不能用于图形发生，但可用于图形转印。一般来说，图形通过掩膜转印一次将带来附加的误差，理想的方法是通过计算机控制直接在基片上扫描形成图形。这种直接描绘图形的无掩技术还可区分为不同的加工方式有：无掩膜光刻(曝光)是对基片上的抗蚀剂直接扫描曝光产生图形，以供后续加工如注入或刻蚀等；无掩膜加工包括无掩膜刻蚀、无掩膜注入、无掩膜淀积等，不再需要通过抗蚀剂来转印图形。这种无掩膜直接加工技术不仅大大简化了工艺过程，减少了加工误差，而且加工图形尺寸的缩小不受到抗蚀剂分辨率的限制，这对微加工具有特别重要意义。

评估各种光刻技术的性能参数主要有分辨率(resolution)、线宽(linewidth)、准确度(accuracy)、失真度(distortion)和套准精度(precision)，在工业生产中成品率(yield)和产出率(throughput)是不可忽视的。物理上对分辨率的定义是能清楚区分的两点之间距离，这在光刻中不常采用，工程上分辨率的定义通常用单位长度上可分辨的高反差线对数表示。线宽表示图形中线条最小的宽度，在光刻技术中分辨率经常用作线宽的同义语。准确度表示空间尺寸对标准值的偏差。失真度表示图形各部位尺寸的相对变化。套准精度表示相同过程产生的图形之间相吻合的程度。成品率表示合格产品与投入总数的百分比。产出率表示单位时间内的出产数量。

图 15 表示不同光刻技术物理模型的比较。光学光刻是当前用得最广泛的光刻技术，采用紫外光作为曝光光源时(g 线 463nm)，可得 1μm 的分辨率，±0.5μm 的套准精度和每小时曝光 100 片

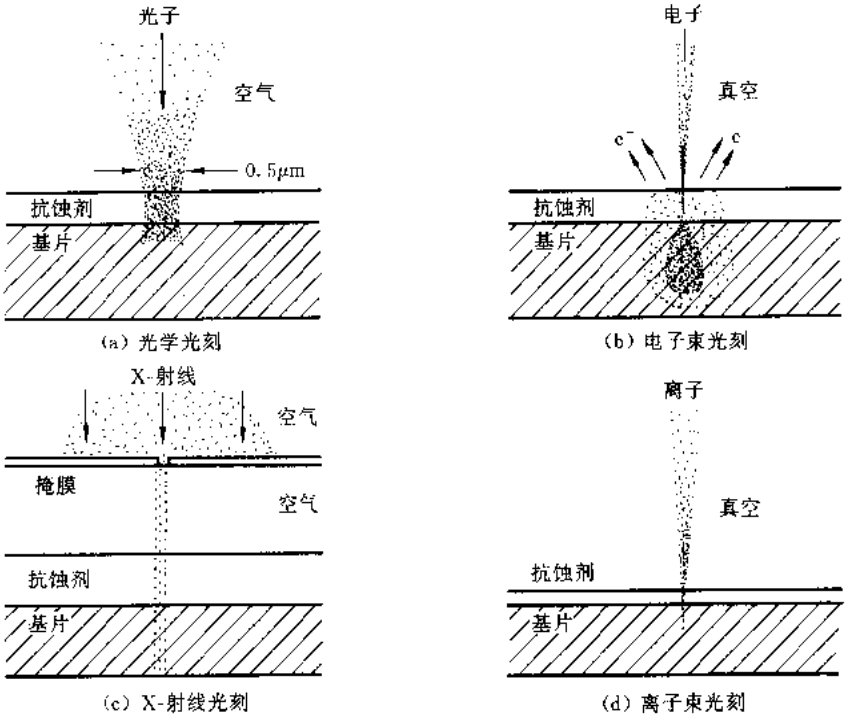
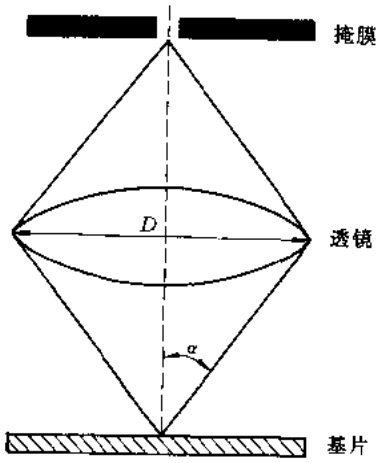


图 15　不同光刻技术的比较

的产出率。进一步提高分辨率受限于光的衍射，因而需要采用更短波长的光源，包括远紫外光和X-射线。对投影曝光装置来说，其光学参数的关系如图16所示。分辨率 ΔX 的表示式为

$$\Delta X = K\frac{\lambda}{NA} \tag{25}$$

式中：K 为与抗蚀剂材料和曝光工艺有关的常数，一般在0.6～0.8；λ 为光波波长；NA 为光学系统的数值孔径，通常设计在0.4～0.5。因而当线宽小于0.5μm时，需要用远紫外或X-射线光刻。



参　量	表 示 式
焦距	f
孔径	D
F 数	$fD = F$
数值孔径	$NA = n\sin\alpha = \frac{D}{2f}$
分辨率	$1.22\lambda F = \Delta x$
场深	$\pm 2\lambda F^2 = \pm\Delta a$

图16　投影曝光系统的光学参数

曝光用的软X-射线波长为0.4～500nm，可以避免常规光刻遇到的衍射问题，图形线宽最终可达数十纳米。由于缺乏实用的X-射线透镜和反射镜，不易得到使X-射线准直、聚焦和偏转的光学系统，因而X-射线曝光尚不能用于制版和投影转印，只能是接近转印，这影响了分辨率的提高。图17表示X-射线接近式曝光装置的工作原理，以及造成半影畸变和几何畸变的图解。半影畸变是由于源具有一定的直径 d 所引起的，可表示为

$$\Delta = s\frac{d}{D} \tag{26}$$

式中：s 为掩膜与样品的间距；D 为源与掩膜的距离。为了保护掩膜和避免掩膜接触基片造成缺陷，要求间隙 s 足够大，通常取10μm。对高分辨率系统，Δ 必须控制小于0.1μm。因此，要求 $d/D > 100$。

几何畸变是由于X-射线束的发散所产生的，样品上的投影像偏离掩膜图形尺寸的大小决定于像离束中心点的距离，如果样品半径为 W，则在样品边缘最大的几何畸变为

$$Z = s\frac{W}{D} \tag{27}$$

但这畸变对分辨率的影响并不明显。但当间隙改变 ds 时，则几何畸变的变化：

$$\Delta Z = \frac{W}{D}ds \tag{28}$$

对线宽有直接影响，在高分辨率系统中，同样要求 $\Delta Z < 0.1\mu m$。现在半导体生产中硅片尺寸已用到0.15cm(6in)以上，这就对 ds 提出了过高的要求。如果采用分步重复曝光的方法，这就

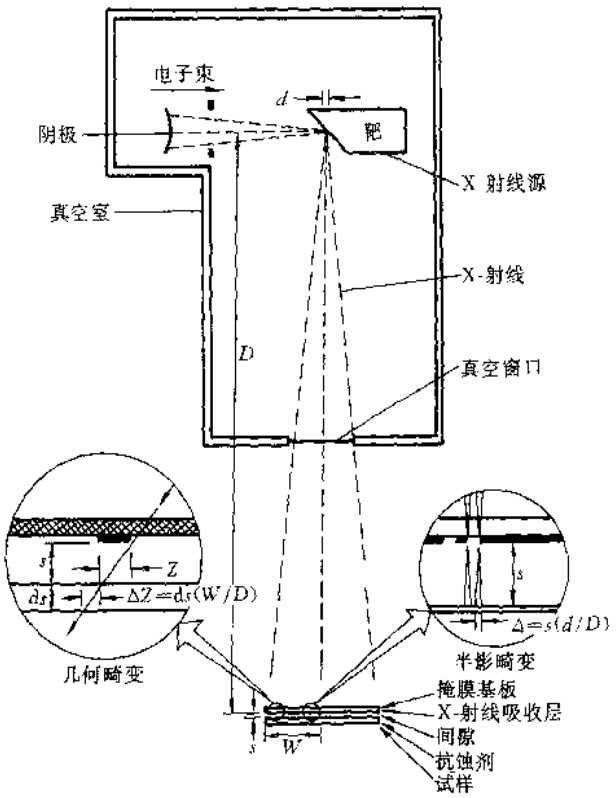


图 17　X-射线接近式曝光装置

可以使 W 保持比较小的范围。最近报道[7]，一家美国公司已首次出售了一种 X-射线步印光刻机(X-ray stepper)，分辨率为 0.35μm，据称其最终目标是 0.25μm 和 0.15μm，能满足 1Gbit 半导体存储器设计规则的要求。

X-射线光刻与射线源、掩膜版设计制造以及抗蚀剂的性能有密切关系，这将留待以后的专题中作详细讨论。

电子束用于曝光与 X-射线相比，不仅波长更短，而且能用电场或磁场对其偏转和速度调制作精确控制，另外还可以对到达抗蚀剂表面的电子能量及剂量在相当大的范围进行调节。因此电子束可以在计算机控制下直接产生图形，也可以通过特殊掩膜将图形转印。电子束可以将束斑聚焦到≤10nm。当束流足够大时，可在 10^{-7}s 时间内使抗蚀剂曝光。通常，每片样品曝光的总次数的典型值为 10^{10}，因而必需要求有非常高的曝光速度。

电子束曝光可采用两种不同的方法：平行曝光和扫描曝光。前者是把组成图形的所有像素在同一时刻曝光；后者是依次对各个像素分别曝光。平行曝光的投影系统通常具有高的产出率，并且装置没有扫描系统那样复杂。但要通过掩膜才能获得图形。扫描系统是通过计算机控制一束或多束电子产生图形，修正畸变和邻近效应误差以及对准基片位置。全部图形信息储存在磁带或磁盘中。这种扫描电子束曝光系统既可用来制作掩膜，也可以把图形直接描绘在基片上。这种方法已成为电子束曝光的主流。

电子束曝光装置工作原理如图 18 所示。电子束图形发生装置的电子光学系统与电子显微镜十分相似。从热阴极电子枪发射的电子束由静电场加速后再由磁场聚焦。最后，由磁场

及电场控制和偏转电子束以一定的轨迹形成所要求的图形。

电子束曝光中限制分辨率的一个重要因素是电子在抗蚀剂层中和在基片界面上的散射。用 Monte Carlo 统计方法通过计算机模拟，可以得出一系列电子无规则散射的运动轨迹。图 19 表示涂有 PMMA 抗蚀剂的硅片上对 100 个电子轨迹进行 Monte Carlo 模拟计算的结果。这种电子散射将给抗蚀剂的曝光图形造成严重影响。电子散射对图形轮廓的影响视抗蚀剂的类型而不同。图 20 表示由电子散射对负性和正性抗蚀剂产生的不同轮廓。对负性抗蚀剂来说，当入射束流密度和背散射剂量分布在抗蚀剂和基片的界面处呈高斯分布时，曝光所得图形的轮廓呈抛物线[8]。

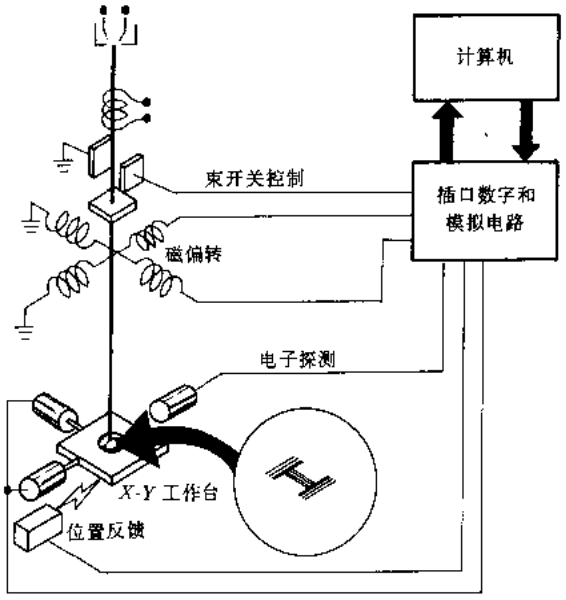
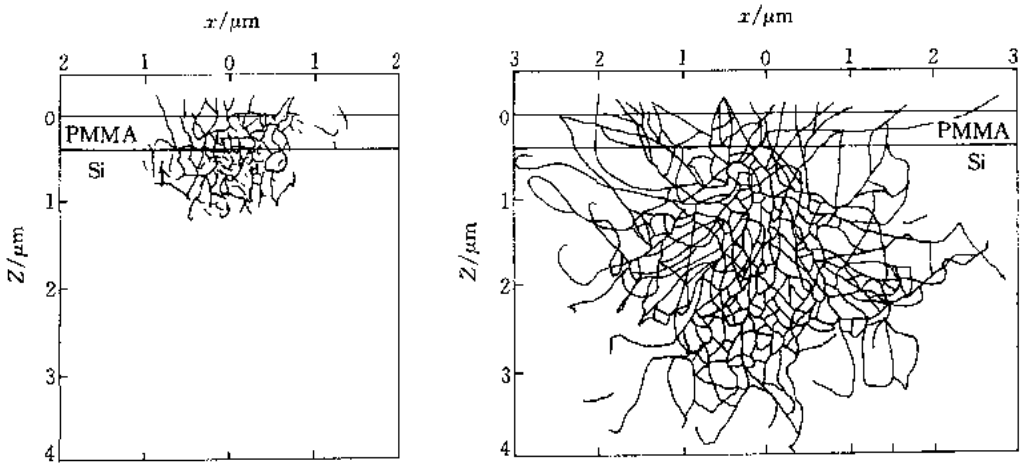


图 18　电子束曝光装置工作原理图



(a) 10kV　　(b) 20kV

图 19　硅片上的 PMMA 抗蚀剂中 100 个电子的 Monte Carlo 模拟轨迹

显影后抛物线底部宽 W 为

$$W = 2r\left(\ln\frac{D^0}{D_m}\right)^{1/2} \tag{29}$$

式中：r 为抗蚀剂和基片界面上的高斯半径(见式(1.8))；D^0 为抗蚀剂曝光后能获得 100% 初始膜厚所要求的照射剂量；D_m 为抗蚀剂开始形成凝胶的最小照射剂量。

抗蚀剂图形的高度 T_0(见图 20)可表示为

$$T_0 = T_i\frac{r}{\ln 10}\ln(D^0/D_m) \tag{30}$$

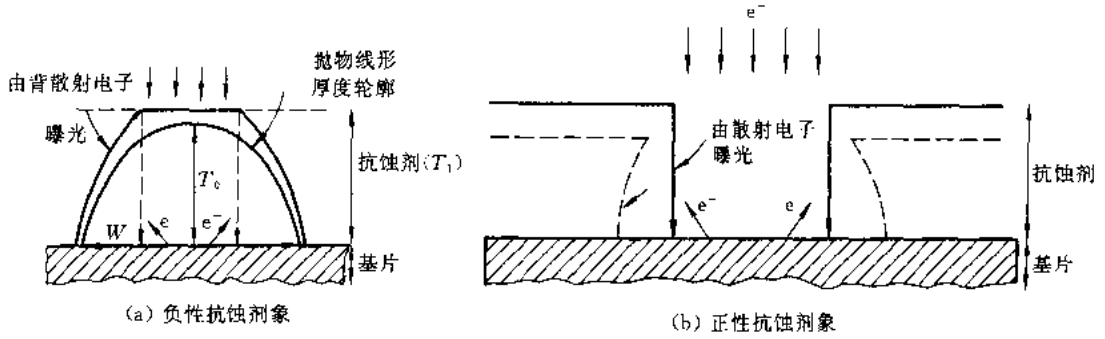


图 20　电子束曝光同电子散射对图形轮廓的影响

式中 r 为抗蚀剂的对比度，定义为

$$r = \left| \log \frac{D^0}{D_m} \right|^{-1} \tag{31}$$

这种由抗蚀剂中电子散射及基片中电子背散射所引起的对曝光图形的影响，通常称为邻近效应(proximity effects)。在严重情况下，邻近效应可使曝光的抗蚀剂线宽变化数微米。邻近效应的程度决定于电子束加速电压、抗蚀剂材料的对比度特性和厚度、基片材料以及显影工艺过程等因素，已经研究了多种途径设法限制邻近效应的影响，例如降低电子束加速电压，采用厚度为 3～100nm 单分子层抗蚀剂膜，也可以通过控制扫描速度和束流来改变到达图形各像素上的电荷量。

离子束光刻在亚微米和纳米加工中将变得越来越重要。这是因为离子束不仅可以代替电子束制作掩膜或直接把图形描绘在基片上，而且也可用于直接的离子注入、离子刻蚀、离子淀积等。离子轰击造成的损伤将可以增强刻蚀速率。利用此效应可以在不需任何抗蚀剂的情况下使氧化层产生图形，对 SiO_2 来说，轰击剂量达 2×10^6 质子/cm^2 就能使其刻蚀速率增加到饱和值。图 21 表示 SiO_2 在 H，D 或 He 离子轰击下造成的辐射损伤使刻蚀速率增强，在刻蚀时这部分损伤区首先被刻除。同样，离子束也能在金属膜，如 Ni 和 Mo 上不用抗蚀剂直接产生图形。因而 Ni 和 Mo 膜层可用作对离子辐射敏感的金属膜抗蚀剂。

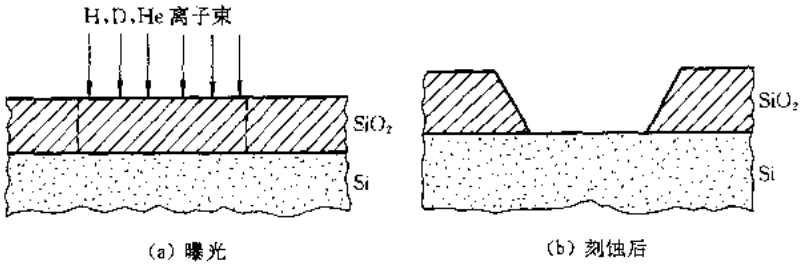


图 21　SiO_2 由 H，D 及 He 离子束的辐射损伤产生图形

离子束光刻具有分辨率高的固有特点，这是因为离子的质量大，离子受散射的射程比电子小得多，由离子束产生的二次电子则因能量低其扩散距离很短，实际上在离子束曝光时不会出现明显的背散。利用同样的 Monte Carlo 方法可以模拟离子束进入抗蚀剂和基片后个别离子的散射轨迹。图 22 表示具有能量为 60keV 的 H^+ 穿过 PMMA 抗蚀剂进入 Au、Si 和 PMMA 的不同轨迹 Monte Carlo 模拟结果。与图 19 所表示的电子轨迹相比，显然离子的散射不会导致对分辨率的限制。

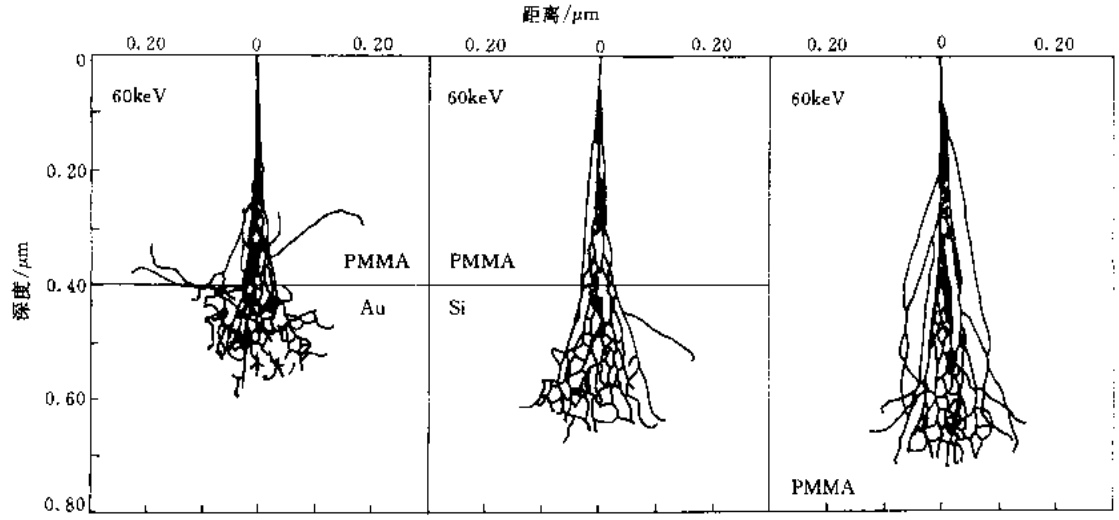


图 22　60keV H 穿过 PMMA 进入 Au、Si 和 PMMA 轨迹的 Monte Carlo 模拟

离子束曝光装置与电子束相仿也有平行和扫描辐射两种模式，采用大面积离子源的离子束投影光刻系统和离子束接近式光刻系统就是属于第一类模式。作为离子束投影系统的例子[9]是采用 H^+、He^+ 及 Ar^+ 射频离子源，电压在 100keV 范围，像场为 5mm × 5mm 一次曝光。该系统主要问题是体积庞大、耐振动性差，离子束光学系统需要改进才能满足 100nm 分辨率的要求。另一种离子束接近式曝光装置[10]，利用〈110〉Si 的沟道效应转印图形，如图 23 所示。在 3～6μm 的〈110〉Si 膜片表面由 100nm 厚的 Au 膜构成图形。采用准直的 He^+ 离子束，能量为 0.5～3keV，以沟道方向入射到〈110〉单晶 Si 膜片掩膜，没有 Au 膜阻挡的部分，入射离子穿过沟道使硅片上的抗蚀剂曝光。落在 Au 膜上的离子其能量在该部位被损耗。所有这些平行辐射

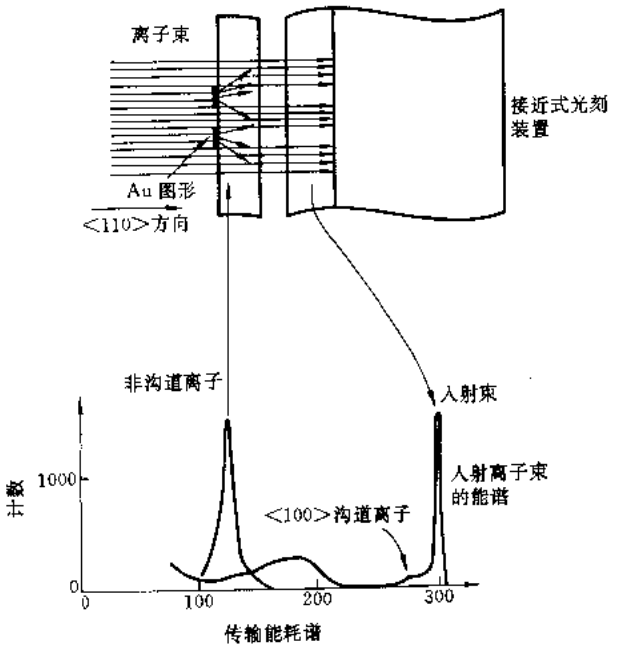


图 23　利用离子沟道效应传递图形的原理

的曝光系统由于离子源强度限制等原因，在实际生产中并没有广泛应用。

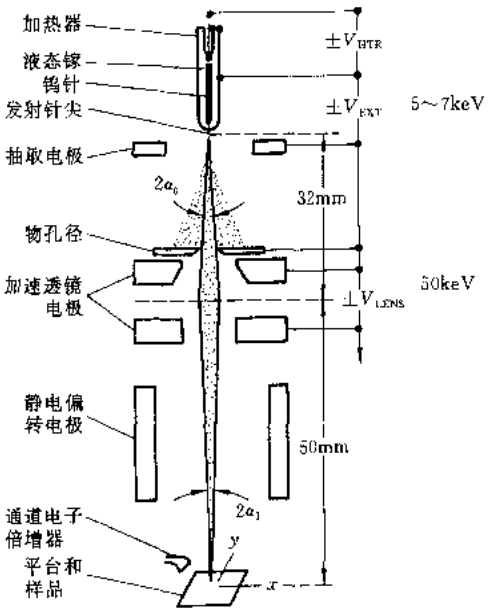


图 24　微聚焦离子束(57keV)扫描系统

采用液态金属离子源形成的微聚焦离子束(microfocused ion beam)构成的高强度扫描离子束系统能用于抗蚀剂曝光、材料淀积和刻蚀、改变表面的电和机械特性等。并且所有这些加工可以不用掩膜直接完成，它能充分发挥离子束的固有特长，因而近年来一直受到优先研究和发展。一个由液态金属镓离子源构成的微聚焦离子束扫描系统[11]如图 24 所示。这种离子源的角强度达 10^{-4}A/sr，较一般的气体放电离子源高出 100 倍。与 X-射线和电子束相比，离子束图形系统的分辨率为最高，可以达 10nm 以上。在纳米加工中它的潜力最大。最近，美国 Hughes 实验室报道了一个最新的聚焦离子束曝光装置[12]，采用同位素镓离子源和一个两透镜微探极系统产生的聚焦离子束直径为 8nm，它可用来产生 10nm 线宽的图形。在这情况下，限制分辨率的主要因素不再是离子束直径，而是抗蚀剂中的过程，如离子散射、原子反弹和曝光的统计涨落等。所采用的抗蚀剂可比普通的 PMMA 的灵敏度低，但要求更高的分辨率和对比度。

3.4　刻蚀技术

上面描述的图形技术在基片表面形成的抗蚀剂图形并不是器件的最终结构，而仅仅是作为对基片加工的样板，即把抗蚀剂的图形精确地转移到基片或基片表面上的薄膜。这一过程通常称为刻蚀。广义的刻蚀还包括基片表面均匀地去除若干原子层，以使表面清洁以及从基片上清除抗蚀剂(俗称去胶)等过程。

把抗蚀剂图形传递给基片可通过两种不同的工艺途径来达到。减法(subtractive)的工艺过程为先淀积一层薄膜，然后在上面涂覆抗蚀剂经光刻形成图形，最后通过刻蚀去除没有被抗蚀剂保护的那部分薄膜，如图 25(a)所示。加法(additive)的工艺过程则不同，先在基片上涂覆抗蚀剂经光刻形成图形，然后再淀积薄膜，这时一部分薄膜淀积在基片表面，另一部分淀积在抗蚀剂表面，最后在去除抗蚀剂时这部薄膜将随之一起被清除，这

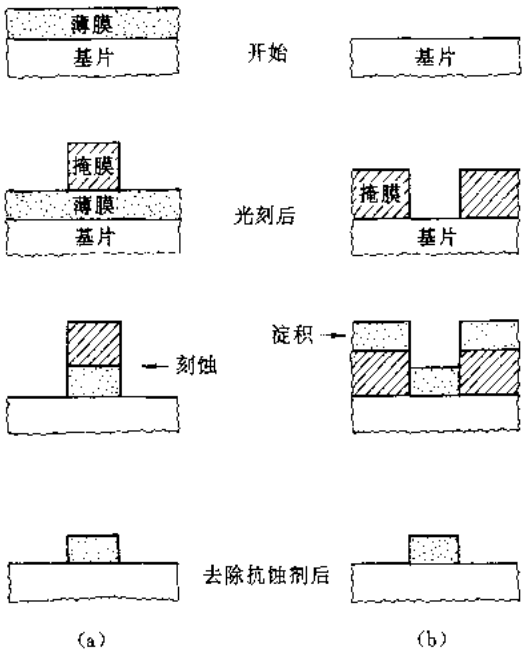


图 25　图形转移的工艺途径
(a) 减法；(b) 加法

过程也称为剥离(lift-off),如图 25(b)所示。必须指出,如要获得相同的最终图形,两种方法所采用的掩膜版是相反的。

薄膜刻蚀可看成是薄膜淀积的逆过程,一般可以通过化学反应、物理过程或两者结合的过程来实现。化学刻蚀是通过具有化学活性的分子、离子或基团与薄膜原子反应生成可溶性或挥发性物质而被去除。单纯的物理刻蚀是薄膜在荷能粒子的轰击下,其表面的原子被溅射出来。如果在溅射刻蚀中有活性物质存在,可以对刻蚀起到增强效应。通常根据刻蚀是液相过程或是气相过程又把刻蚀分为湿法和干法。湿法刻蚀主要是化学作用。干法刻蚀有主要是化学作用的等离子体刻蚀,主要是物理作用的离子束刻蚀(IBE)以及化学作用和物理作用互相增强的反应离子刻蚀(RIE)和反应离子束刻蚀(RIBE)。

刻蚀工艺的基本要求是能保真地转移图形,这可用偏差(bias)和公差(tolarance)两个参数来确定。偏差 B 表示为

$$B = d_f - d_m \tag{32}$$

式中:d_f 是抗蚀剂图形尺寸;d_m 是掩膜图形尺寸。如图 26 所示,刻蚀公差是基片上各点偏差的统计分布平均值,表示刻蚀的横向均匀性。这种偏差的造成是由于刻蚀的方向性。如果刻蚀只沿垂直于膜面的方向进行,称为各向异性刻蚀,如图 27(a)所示,这时抗蚀剂图形的侧面与掩膜图形一致,刻蚀偏差为零。如果刻蚀在薄膜的垂直和水平方向同时发生,称为各向同性刻蚀,如图 27(b)所示,这时抗蚀剂的侧面造成 1/4 圆弧,刻蚀偏差为膜厚的 1 倍。刻蚀各向异性的程度决定于横向和纵向的刻蚀速率的比值,并以各向异性因子 A_f 表示:

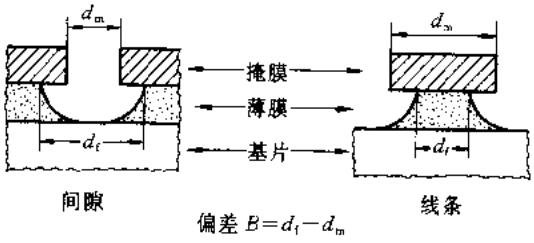


图 26 刻蚀偏差示意图

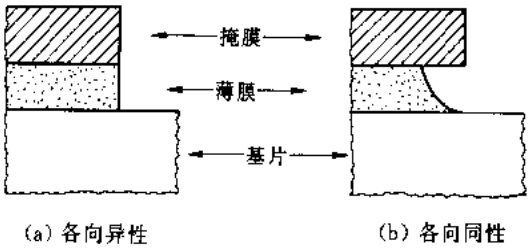


图 27 各向异性和各向同性的刻蚀

$$A_f = 1 - \frac{V_{fl}}{V_{fv}} \tag{33}$$

式中,V_{fl}和 V_{fv}分别表示横向和纵向的刻蚀速率。对各向同性刻蚀,$V_{fl} = V_{fv}$,$A_f = 0$。当 $1 \geqslant A_f > 0$ 时为各向异性刻蚀。如果 $V_{fl} = 0$,$A = 1$,这表示完全各向异性刻蚀。掩膜下的薄膜在横向被刻蚀的现象叫钻蚀,并由此引起抗蚀剂图形尺寸的偏差,它与薄膜刻蚀各向异向参量的关系为

$$A_f = 1 - \frac{|B|}{2h_f} \tag{34}$$

式中 h_f 为薄膜厚度。

实际上,在薄膜刻蚀时,抗蚀剂掩膜和基片也将同时被刻蚀,特别是干法刻蚀,只是它们的刻蚀速率有差异。通常把不同材料刻蚀速率之比称为刻蚀的选择性。例如薄膜对掩膜刻蚀的选择性 S_{fm}可表示为

$$S_{fm} = \frac{V_f}{V_m} \tag{35}$$

式中 V_f 和 V_m 分别为薄膜和掩膜的纵向刻蚀速率。由于掩膜受刻蚀,特别当掩膜的边缘具有

倾角时，将给图形转移造成附加的尺寸偏差，如图28所示。下面让我们分析一下掩膜和基片受刻蚀所产生的影响。

设薄膜厚度为 h_f，其全部面积上的百分比均匀性为 δ，因而最大的厚度为 $h_f(1+\delta)$。如果薄膜的平均刻蚀速率为 V_f，其百分比不均匀性为 ϕ_f，则各点刻蚀速率的变化范围为 $V_f(1+\phi_f)$。为要得到完整的图形，必须保证各部位的薄膜比掩膜早刻蚀完。因此要考虑薄膜最厚、刻蚀速率最小以及掩膜最薄、刻蚀速率最大的极端情况。这时，薄膜刻蚀结束的时间为

$$t_e = \frac{h_f(1+\delta)}{V_f(1-\phi_f)} \tag{36}$$

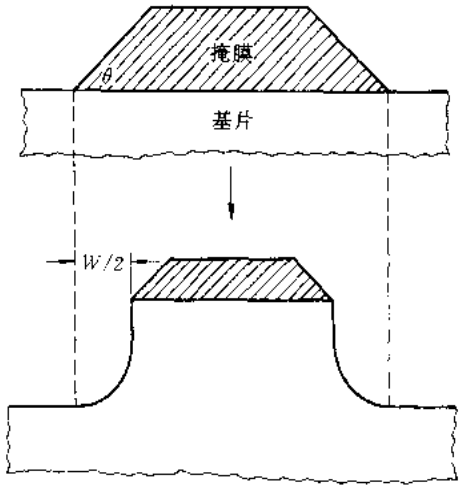


图28 掩膜刻蚀对图形尺寸的影响

在刻蚀工艺中，为了保证各部位都刻蚀完，需要增加一定时间叫过刻蚀时间，通常用时间、百分比 Δ 表示。因此刻蚀总时间为

$$t_t = \frac{h_f(1+\delta)(1+\Delta)}{V_f(1-\phi_f)} \tag{37}$$

在 t_t 时间内掩膜受到刻蚀在横向的尺寸减少为 W，由图28可得：

$$W/2 = [V_{mv}\cot\theta + V_{ml}]t_t \tag{38}$$

式中 V_{mv} 和 V_{ml} 分别为掩膜纵向和横向刻蚀速率的最大值。按式(36)的定义，代入式(37)、(38)可得：

$$S_{fm} = \frac{h_f}{W}U_{fm}[\cos\theta + (1-A_m)] \tag{39}$$

式中：$U_{fm} = [(1+\delta)(1+\Delta)(1+\phi_m)]/(1-\phi_f)$ 为均匀性因子；θ_m 为掩膜刻蚀速率的百分比不均匀性；$A_m = 1 - V_{ml}/V_{mv}$ 为掩膜刻蚀各向异性因子；θ 为抗蚀剂掩膜边缘的倾角，往往由光刻工艺和抗蚀剂材料所决定。扫描投影曝光所得倾角的典型值为60°，接触式曝光和多层抗蚀剂系统可得接近90°的直边掩膜。式(39)表示不同的 θ 和 A_m 情况下要达到一定的 h_f/W（膜厚对线宽减小比值），必须要求的选择比 S_{fm}。

对基片刻蚀所要求的选择性可以通过类似的方法来确定：

$$S_{fs} = \frac{h_f}{h_s}U_{fs} \tag{40}$$

式中：h_s 为基片最大的允许刻蚀深度；U_{fs} 为有关的均匀性因子。对基片刻蚀选择性的要求在过刻蚀时间增加时尤为重要，特别当各向异性刻蚀遇到台阶轮廓时是必不可少的。图29中假设刻蚀为完全各向异性（$A_f = 1$），当厚度为 h_2 的膜Ⅱ刻蚀终了时，在台阶处有厚度为 h_1 的残留，必须增加刻蚀时间才能除去，所需的过刻蚀时间将为 $\Delta = h_1/h_2$。

湿法刻蚀长期来得到广泛应用，技术成熟。它是利用某种对掩膜和基片不起作用而能溶解所需刻蚀的薄膜材料来实现的。这种方法使用的设备较简单、费用低，容易实现。湿刻蚀一般情况下表现为各向同性，引起明显的钻蚀，这是限制分辨率提高的主要因素，在半导体生产中通常只用于3μm以上的器件。但是如果采用剥离法（见图25(b)），则分辨率可以提

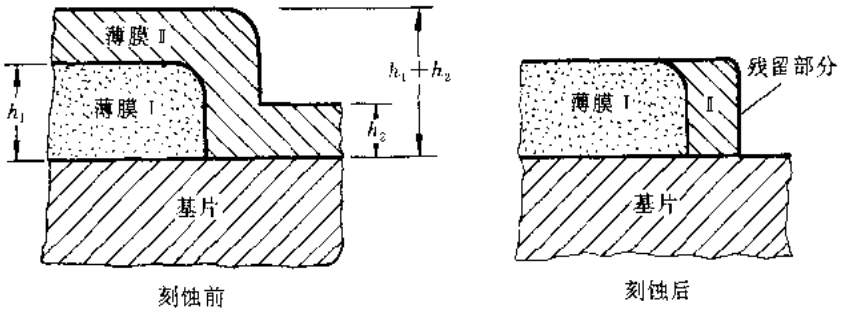


图 29　各向异性刻蚀时，需要过刻蚀将残留部分除去

(a) 刻蚀前；(b) 刻蚀后

高。这种方法对化学试剂不易作用的铂、金膜的刻蚀特别有利。但是，要保证薄膜淀积后抗蚀剂容易剥离，抗蚀剂要比较厚且边缘陡直，使基片上淀积的膜与抗蚀剂上的膜不连在一起。为更好地提高工艺的可靠性，对剥离法工艺作了进一改进，图 30 为几种实施方案。图 30(a)基片涂敷抗蚀剂后在氯苯中预浸一段时间，将表层中的低分子树脂和残余溶剂去掉，因而密度增加，在显影时尺寸收缩小，氯苯没有渗透到的深度，抗蚀剂显影时尺寸收缩大，因而造成 T 形截面。(b)采用两层对显影液溶解度不同的抗蚀剂材料，以达到相同的目的。(c)采用三层结构。

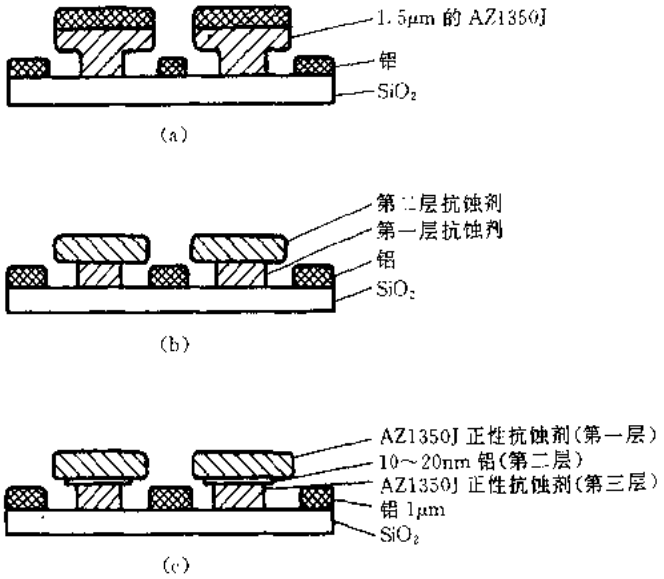


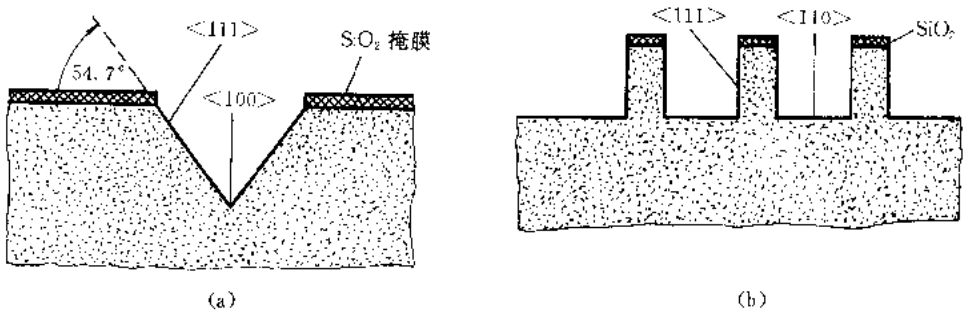
图 30　不同的剥离法(lift-off)工艺

(a) 抗蚀剂预浸一步法；(b) 二层不同显影溶解度的抗蚀剂；

(c) 二层相同的抗蚀剂之间增加中间金属层(Al)

有一些刻蚀剂对晶体的某些晶面的刻蚀速率比对其他晶面的刻蚀快得多，这时将出现各向异性刻蚀。例如对 Si 单晶来说，应用刻蚀液 EDA 在 100℃时对〈100〉、〈110〉和〈111〉晶面的刻蚀速率大致为 50:30:3(μm/h)。对〈100〉硅片用 SiO_2 作掩膜用 EDA 刻蚀可获得精确的 V 形槽，其边〈111〉面与〈100〉面成 54.7°，如图 31(a)所示。如果通过方形窗口对〈110〉硅片用 EDA

作择优定向刻蚀，可获得侧面陡直的矩形槽，如图 31(b)所示。利用这种方法已获得宽为 0.6μm，中心间距为 1.2μm，深度为 600μm 的槽，这是用其他方法一般难以达到。



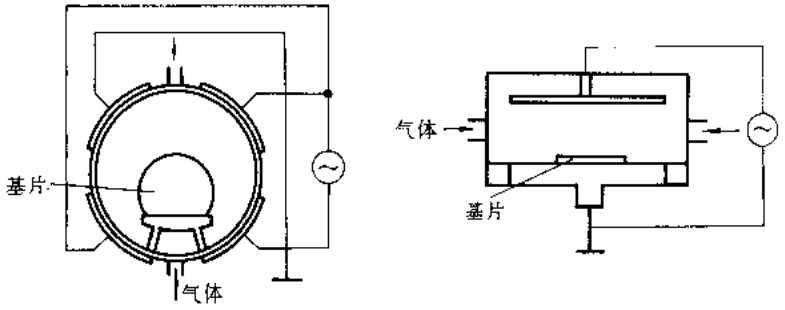
(a)　　(b)

图 31　不同晶面硅片的择优定向刻蚀

(a)〈100〉面；(b)〈110〉面

干法刻蚀包括等离子体刻蚀、溅射刻蚀、反应离子刻蚀、离子束刻蚀和反应离子束刻蚀，与湿法刻蚀不同，它们不是在液相条件下而是在气相状态下进行的。

等离子体刻蚀是利用稀薄气体辉光放电产生的等离子体所引起的化学反应。早在 60 年代，首先将这种方法用于除去抗蚀剂。因为抗蚀剂是 C 和 H 组成的有机化合物，等离子体中活泼的 O 原子把 C 氧化为 CO 和 CO_2，H 与 O 生成 H_2O，这些氧化物均可被真空系统抽走。这工艺称"等离子体灰化"(plasma ashing)。70 年代初，这种方法又被应用于其他材料，为此要寻找合适的放电气体，要求它能使除去的材料在辉光放电中形成挥发性产物，这方法被称为等离子体刻蚀(plasma etching)。用于等离子体刻蚀的装置主要有两种结构形式，即圆筒形和平板形，分别如图 32 所示。采用圆筒形反应器时，样品被包围在辉光放电的等离子体中，样品表面将受到光的辐照及电子和离子的轰击，包括在样品表面所发生的离子和电子的复合，并且伴随着热效应使样品温度升高，这给以抗蚀剂为掩膜的 Si 或 SiO_2 的刻蚀造成困难。为了不使样品表面遭受电子和离子的轰击，可用一个金属多孔或网格圆筒隧道放在反应器内作为屏蔽，使辉光放电限制在器壁与隧道壁之间。这样，放在隧道内的样品仍可与通过小孔进入隧道的活性原子和基团起作用，因而避免了电子和离子的轰击。当用 CF_4 气体对 Si 刻蚀时，这种装置能有效地抑制样品的温升，提高抗蚀剂的刻蚀选择比。图 33 所示是一种微波等离子体刻蚀装置，它将放电区和反应区分隔开来，通常采用 2.4GHz 的微波源产生辉光放电，所产生的活性原子团由气流输送到样品表面引起刻蚀，这样几乎还可以完全避免放电室



(a) 圆筒型　　(b) 平板型

图 32　等离子体刻蚀装置示意图

发生的光和热的影响。等离子体刻蚀主要是化学刻蚀，一般都无方向性，因而它像湿法刻蚀一样存在严重的钻蚀，限制了分辨率的提高。但可以获得较大的刻蚀选择比，刻蚀速率快。

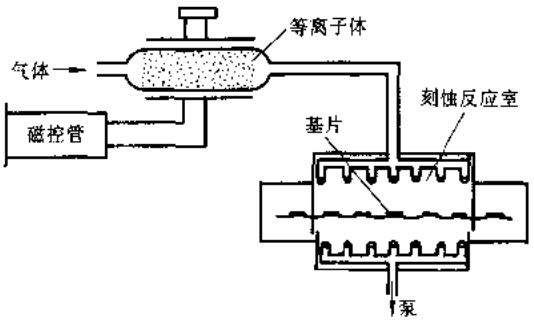


图 33　微波等离子体刻蚀装置示意图

溅射刻蚀是由气体放电形成的正离子在阴极偏压作用下直接轰击样品表面所引起。这种反应器基本上也是一套平板电极系统，样品放在阴极上。这种刻蚀有明显的方向性，不会产生钻蚀，可以加工亚微米图形。但由于溅射刻蚀是纯物理过程，对不同材料的选择性差，而且一般刻蚀速率较低。

反应离子刻蚀(Reactive Ion Etching, RIE)是采用含卤化物的气体代替惰性气体 Ar 的溅射刻蚀，所以也称为反应溅射刻蚀。这种刻蚀过程是离子轰击的物理效应与活性粒子的化学效应的相互增强，因而兼有等离子体刻蚀和溅射刻蚀两者的优点，不仅刻蚀速率高，而且具有良好的方向性和选择比，分辨率高。目前在超大规模集成电路的制造中广泛应用。反应离子刻蚀与平板等离子体刻蚀在装置上十分相似，两者的区别在于前者将样品置于电源电极(阴极)上，气体工作压强比较低，一般低于 10Pa，这就造成了使高能离子轰击样品表面的条件。后者将样品置于接地电极上，工作气压高于 10Pa，因而刻蚀中离子轰击不起明显作用。

离子束刻蚀(IBE)又称离子铣，其装置示意图如图 34 所示。在离子源部分，由气体放电产生的离子被电场加速后以高速离子束的形式轰击基片表面，通常离子束具有良好的平行性，可以通过倾斜样品台来控制入射角度。例如调节这倾斜角可加工出不同剖面的光栅，增加离子束能量可以提高刻蚀速率，刻蚀方向性很强，可获得非常高的分辨率，但选择性低并容易造成损伤，这就限制了在半导体器件上的应用范围。

反应离子束刻蚀 RIBE 的装置与工艺参数类似于离子铣，只是用等离子体刻蚀和反应离子刻蚀用的工作气体代替离子源中的惰性气体。与 RIE 一样，反应离子束刻蚀兼有离子束刻

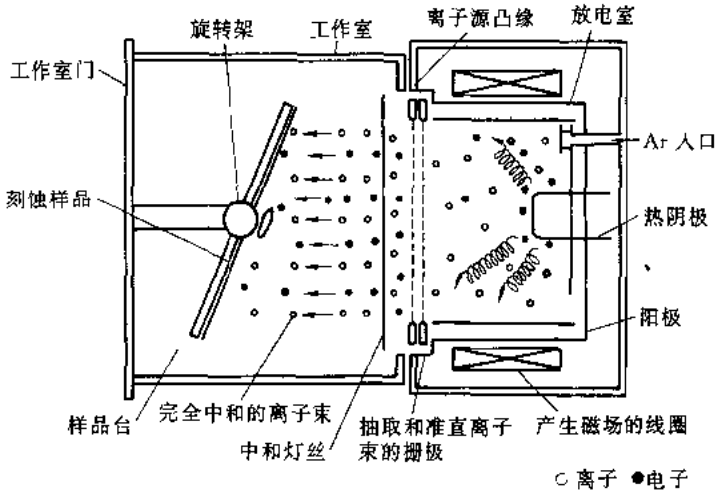


图 34　离子束刻蚀装置

蚀和等离子体刻蚀的优点。但是由于离子源装置相当复杂及维护原子源的困难，反应离子束刻蚀的应用受到一定限制。

各种刻蚀工艺的性能特点比较列于表5。

表5 各种刻蚀工艺的比较

工艺		湿法（化学溶液）	物理溅射或离子束刻蚀	等离子体刻蚀			反应离子刻蚀	反应离子束刻蚀
				圆筒形	圆筒形	平板形		
样品放置		浸入	阴极	等离子体区	隧道内	阳极	阴极	样品台
活性物质	高能离子		√					√
	低能离子			√		√		
	长寿命自由基			√	√	√	√	
	短寿命自由基			√		√	√	
	原子及分子			√	√	√	√	
各向异性（决定分辨率）		差	优	可改变	差	良—优	优—良	优
刻蚀速率		快	慢	快	快	快	中	中—快
样品温度控制		优	良	差	中	良	良	良
选择性		优	差	良	中	良	良	中—良
MOS绝缘损伤		无	重[a]	轻[b]	无	轻[b]	中[a]	中
再淀积		无	重[c]	轻[d]	无	轻[d]	中[c]	中
刻蚀均匀性		良	良	可改变	良	可改变	良	良
刻蚀尺寸极限/μm		3～4	<1	2～3	2～3	1～2	<1	<0.1

注：a)离子轰击；b)UV电子轰击，反向散射沾污；c)溅射出非挥发性物质反向散射；d)自由基的表面分解、沾污物的反向散射。

3.5 诊断技术

诊断技术是通过仪器分析方法来检验微加工是否符合要求的手段。对于加工完成的器件，必须通过诊断方法检验器件的结构是否符合设计的尺寸，其各部分材料的物理性质是否达到所提出的要求。在大量现有的诊断技术中，与微加工密切有关的仪器分析方法主要用于尺寸、成分、结构及性能的分析和测定。表6列出不同应用的仪器分析及其特点。

表 6　仪器分析方法的应用范围及特点

诊断项目	仪器分析技术	应用范围	特　点
形貌和尺寸	光学显微镜		场深小
	透射电子显微镜(TEM)	$>0.3\mu m$	要求薄膜样品
	扫描电子显微镜(SEM)	$>1nm$	适用于表面形态
	场发射电子显微镜(FEM)	$<3nm$	只能用于特定制备的样品
	场离子显微镜(FIM)	$<1nm$	
	扫描隧道显微镜(STM)	$>0.1nm$	
	原子力显微镜(AFM)	横向 0.2nm,纵向 0.005nm 横向 3nm,纵向 0.1nm	三维形貌
化学成分	化学分析电子能谱仪(ESEC(XPS))	分析极限 1×10^{-6}	测量深度达 $100\mu m$
	俄歇电子能谱仪(AES)	分析极限 100×10^{-6}	限于表面层 0.2～1nm
	二次离子质谱仪(SIMS)	分析极限 10×10^{-6}	分析直径 $1\mu m$ 以上
晶体结构	X-射线衍射仪(XRD)	相判别,常数测定等	测定晶格常数精确
	透射电子衍射仪(TED)	相判别,确定晶向等	能用于微晶区测定
	Rutherford 背散射谱(RBS)	晶相无定形相判别	能确定杂质原子在晶格中的位置

用作形貌观察和尺寸测量的显微镜可以用可见光束、电子束、离子束以及 X-射线来实现。最早应用的光学显微镜至今仍广泛地应用着。显微镜的功能通常有四个基本参数表征,即:视场,放大倍数(M),分辨率(δ)及焦深(D)。分辨率 δ 与光束波长 λ 的关系为:

$$\delta = \frac{\lambda}{NA} \tag{41}$$

式中 NA 为物镜的数值孔径。目前物镜的 NA 在空气中最大可达 0.95μm,在样品和物镜之间注油后可达 1.5μm。人的眼睛能反应的可见光波长范围为 0.4～0.7μm,即使采用最好的显微镜,能分辨的最小尺寸不小于 0.3μm。焦深是分辨能力不受散焦影响的沿光轴的距离,可表示为

$$D = \frac{\delta}{\sin\alpha} \tag{42}$$

2α 是成像光线的发散角,因此对高倍物镜来说,$\sin\alpha \approx 1$,表示光学显微镜的焦深与分辨率接近。

电子束即使能量仅 100eV,其波长也比光束短得多,虽然电子透镜做不到光学透镜那么完好,电子透镜的像差将造成杂散电子束。但是现在的电子显微镜的分辨率一般都可达到 1nm,经仔细调整电子光学及采用一些特殊的方法,分辨率可提高到 0.5nm,甚至原子尺寸(<0.3 nm)。虽然必须把观察的样品放进真空室,但现代商品电子显微镜并不难操作,已成为观察具有亚微米公差的器件的主要仪器,电子显微镜还具有许多附加的分析能力。

电子显微镜可分成三类,即透射式、反射式和发射式,它们又都可以两种不同模式操作,即投影式和扫描(飞点)式。这六种不同工作状态的电子显微镜都有它们特定的应用场合。其中利用二次电子发射的扫描式电镜(SEM)对微结构的观测应用得最为普遍,尽管其分辨率范围为 3～10nm。其显著的优点是具有记录三维物体表面结构的非常能力,而且能在低倍率下观察大的面积而后改变焦距对感兴趣的小范围用高倍率观察。同时,二次电子的能谱可获得表面层化学成分的信息,X-射线荧光光谱可给出有关体成分的信息,典型的扫描电镜的工作原理如图 35 所示。

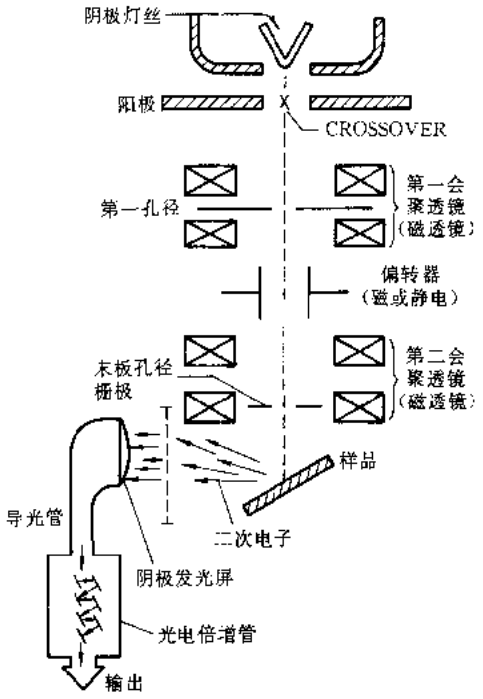


图 35　扫描电子显微镜工作原理

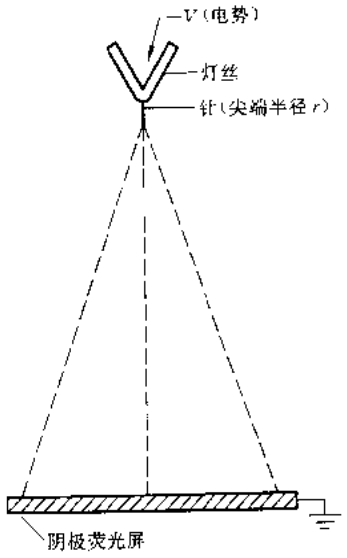


图 36　场发射电子显微镜

场发射电子显微镜(FEM)的工作原理如图 36 所示。其结构十分简单,它包括一枚端部非常尖的针作为样品和相隔一定距离的阴极荧光屏,在外施电压作用下,针尖产生场发射电子打到荧光屏上。针尖的电场强度 E 可表示为

$$E = V/kr \tag{43}$$

式中:V 为针的电势;r 为针尖半径;k 为决定针形状的常数,通常取 7。要使发射电流密度的范围为 $10^4 \sim 10^8 A/cm^2$,针尖电场强度应达 $10^7 \sim 10^8 V/cm$。如果 r 为 1μm,则需要加电压 10kV。电子发射的轨迹是沿径向的直线,放大倍数可简单地表示为

$$M = R/kr \tag{44}$$

式中 R 为荧光屏半径,当 $R = 10cm$ 时,$M = 10$。FEM 的分辨率决定于发射电子的速度分布及动量的涨落,一般可达 1nm,能够看到针尖吸附的有机分子。

场离子显微镜(FIM)的原理与 FEM 相似,只是由正离子代替了电子。如果一个原子或分子放在 $10^8 V/cm$ 的高电场中,首先是被极化,当电场力达到某临界值时,原子中的电子由隧道效应而逸出,形成带正电的离子。在针尖附近不仅电场强度高,而且电场梯度增强,电场梯度把极化的粒子推向针尖,进到一定的高电场区便发射电子形成离子,电场进一步作用于离子将其拉出表面,沿径向飞抵荧光屏。FIM 的分辨率可达 0.1nm,能清楚地观察到原子的排列,这是因为离子的质量大的缘故。

新近迅速发展起来的两种原子尺度的分析仪器扫描隧道显微镜(STM)[13]和原子力显微镜(AFM)[14]的工作原理和实验装置如图 37 所示。STM 的工作原理是基于物理上的隧道效应。由于电子固有的波动性,物质表面的电子密度并不突然下降为零,而是在很短的距离内(数 nm)按指数式迅速衰减,如图 37(a)所示。这衰减长度可表示为阻挡电子离开表面的势垒高度,因而"表面"对电子来说是一个模糊的界面,它只是定义为最外层原子的位置。如果两金属

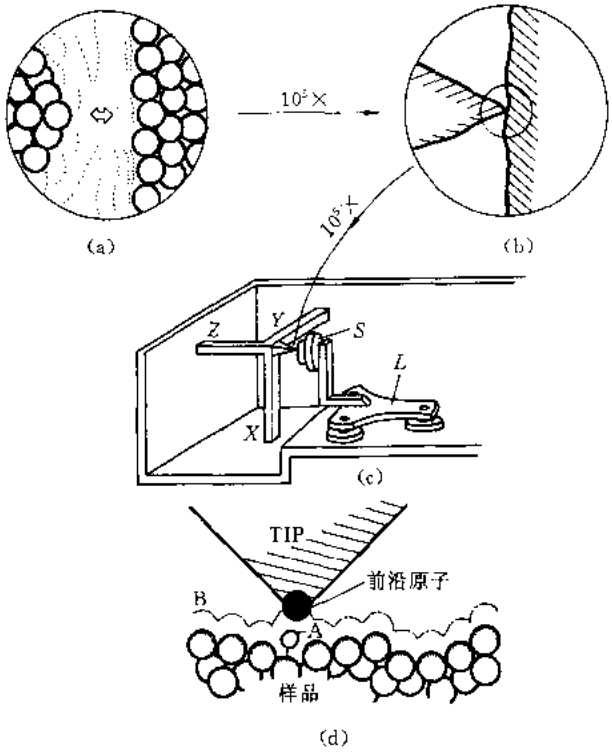


图 37　STM 和 AFM 的工作原理和装置

接近到几 nm 之内，其外围电子云发生重叠，只要在两金属间施加很小的电压，就能测得电流。这种隧道电流 J 反应出电子波函数重叠的程度，并与金属间距离 S 有很强的关系：

$$J \propto \exp(-AK_0S)$$

其中 K_0 为平均衰减长度的倒数，对两个均匀的金属表面，$K_0 \approx \sqrt{(\phi_1+\phi_2)/2}$，$\phi_1$ 和 ϕ_2 是对应的功函数，K_0^2 也表示隧道势垒高度。隧道电流 J 与 S 和 K_0 之间非常强的关系就是 STM 的基础。STM 的一对电极如图 37(b)所示，将一个非常尖的电极贴近样品表面，并在保持测得的隧道电流不变的条件下使极尖沿样品表面扫描，这时极尖所移动的轨迹是波函数等重叠的轮廓。在衰减长度恒定的情况下，极尖移动所描绘出的一系列轮廓将是表面原子位置的实像，即原子排列的形貌。在 STM 装置中，极尖与样品之间的距离的调节和沿样品表面的扫描是通过三维的压电微位移器实现的，如图 37(c)所示。图中 L 是静电滑动平台，供样品的初定位用。STM 不仅可用对表面结构的观察，同样可用于原子尺度的表面化学分析。现在达到的分辨率纵向为 0.005nm，这决定于机械稳定性，横向为 0.2nm，这由隧道通道的直径决定。虽然 STM 出现不久，但由于具有许多吸引人的性质，在短短几年中得到了极大发展和广泛应用。

AFM 的工作原理是 STM 的推广，如图 37(d)所示。当极尖沿轮廓 B 移动时保持隧道电流不变，这就是 STM 的情况；如果保持极尖与样品之间的相互作用力不变，得到的是 AFM 图形。AFM 是将 STM 和探针轮廓仪的原理结合起来，可以测量小到 10^{-18}N 的力，这样高的灵敏度可以测出单原子之间的作用力。这可用来分析导体和绝缘体的表面结构。初步在空气中获得的实验结果为横向分辨率 3nm，纵向分辨率为 0.1nm[14]。

不同元素的原子其外层电子和内层电子各有一定的结合能，因此可以用测定原子的特定

电子能谱作为分析材料化学成分的手段。图 38 是单电离的 Si 原子的部分能级图。当具有足够能量的电子入射时，能从内层电子壳层 K 中逐出电子形成空位。电子从较高的 L_1 能级跃迁到 K 层的空位时将释放能量，这能量可引起更高能级 $L_{2,3}$ 上的电子电离，这种二次电子称为俄歇电子。由于俄歇电子能量比较低，它们只能从表层原子中逸出。根据给定元素的俄歇峰，就可确定表面各部位的成分。

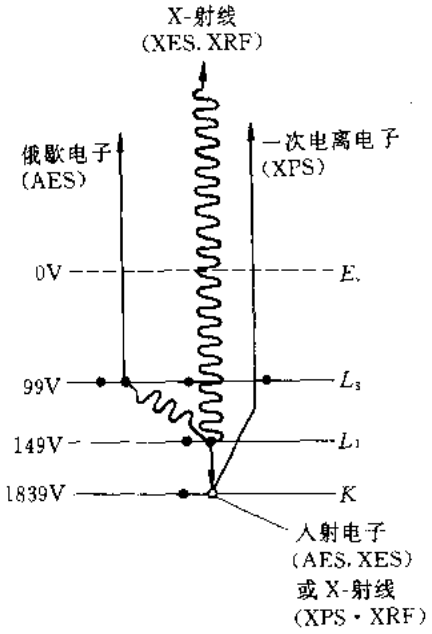


图 38　单电离的 Si 原子的部分能级图

电子从 L_1 能级跃迁到 K 层空位时能量释放也可以辐射 X-射线形式出现，根据这种 X-射线谱可对材料的成分作定性和定量分析，这方法称 X-射线发射谱分析(XFS)。在 SEM 和 TEM 装置上都可进行 XES 分析。

当 K 层电子电离是由入射的 X-射线造成时，可以用来进行 X-射线光电子能谱分析(XPS)或 X-射线荧光分析(XRF)。XPS 通常可替代 AES 或相互补充。XPS 的优点是：①由于 X-射线的散射截面所引起的吸引和分解比电子轰击时低得多，所以对材料的破坏性小；②因为 X-射线是中性的，不会造成样品表面充电，能用于绝缘体分析；③内层电子的能级受价态和化学键类型的影响，可从 XPS 数据获得有关化学键的信息。XPS 的不足是分析的面积不能太小。

分析化学成分另一类有效仪器是二次离子质谱分析(SIMS)，这种方法的典型装置如图

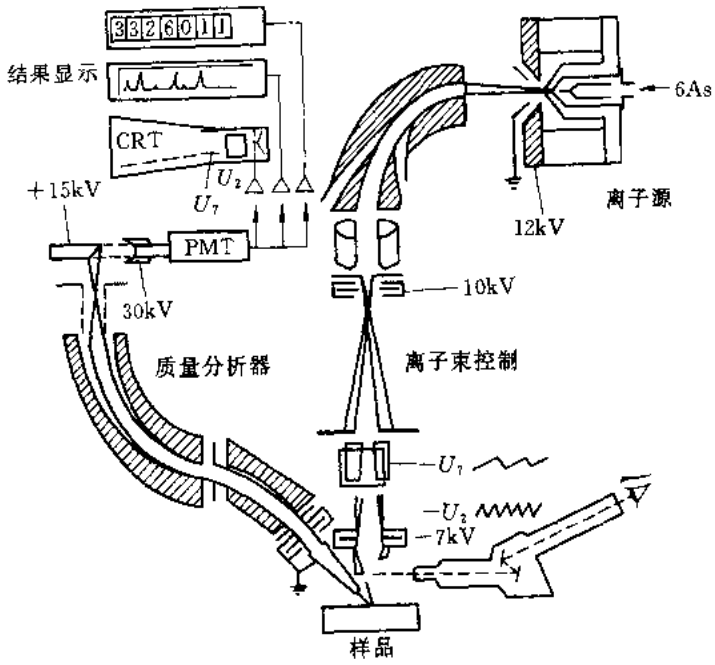


图 39　二次离子质谱分析装置原理图

39 所示。从离子源发射的一次离子束在样品表面溅射出二次离子通过电场抽取进入质谱仪，根据其荷质比对成分作出分析。多数情况是采用四极质谱仪。SIMS 能以三种模式工作：作为离子微探针采用聚焦离子束进行微区分析；入射离子在整个样品表面作栅状扫描，记录下二次离子束流变化形成物质二维分布图像；采用宽束对整个表面（$>100\mu m$）进行分析。这种分析方法的特点是包括 H 在内的元素周期表上所有元素都能进行分析，分析的面积可以小到直径 $1\mu m$，其灵敏度在大多数情况下可超过 AES 和电子探针，在一定条件下探测极限达 10×10^{-6}。

虽然为了确定微观尺度的形态、化学元素的组成、分子类型以及横向和纵向的分布，必须应用各种不同的微分析技术。而且，往往要求能同时确定表面和体内的成分，并具有亚微米级的分辨率。这种要求只有用联合分析系统才能实现。但是，对于纳米器件加工和特性的诊断，尚待研究开发信噪比非常高的新分析仪器。对亚微米的成分分析仪器也需要更高的灵敏度。因为现在普遍使用的分析仪器，当分析面积减小时，其灵敏度大大降低，从图 40 可看到当尺寸小于 100nm 时，材料的密度必须大于 10^{17}原子/cm^3 才能被探测到。

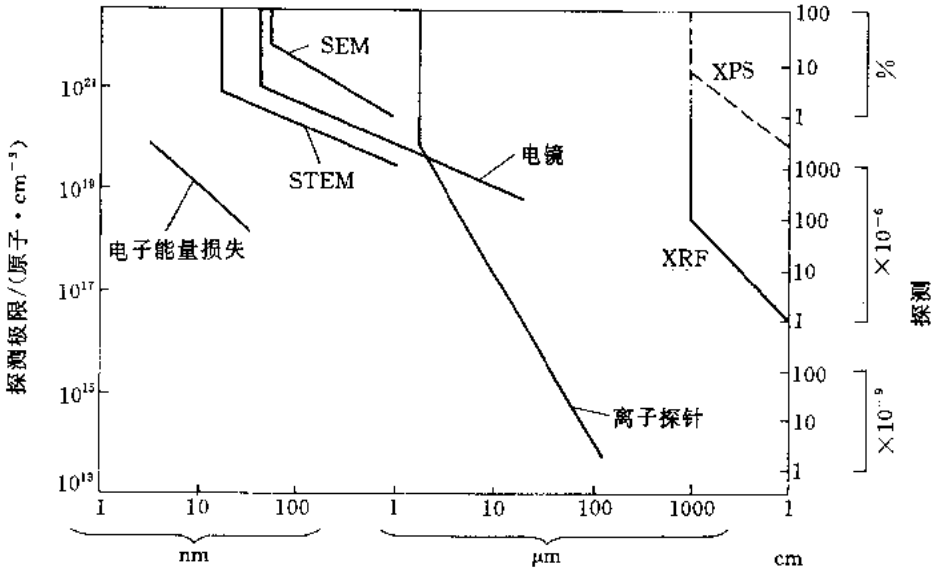


图 40　分析面积的直径对探测极限的影响

4　纳米加工的现状和未来

一般来说，器件尺寸的不断缩小最终将决定于下述两个极限因素：一是器件赖以工作的物理原理能有效的极限；二是达到器件要求的尺寸和公差的加工能力的极限。

可以说，迄今为止几乎所有的微器件都是基于材料的整体性质（bulk properties），即有关的物理性质由大块原子列阵集合效应所造成的。当列阵的尺寸减小到某临界值时，这种物理性质将不复存在或遵从不同的规律。例如，如果希望在光波导中传播一定波长的电磁波，当波导的截面小于某一临界尺寸时，则导波将被截止。表 7 列出几种当前感兴趣的器件及其尺寸缩小的主要极限因子和最小尺寸。

表 7　几种微器件尺寸缩小的主要极限因子

器　　件	主要极限因子	最小尺寸	
		原子数	nm
MOS 晶体管	由击穿电压决定的沟道长度	1000	250
双极晶体管	基区厚度和掺杂浓度	1000	250
磁泡存储器	由磁畴能量决定的磁泡直径	200	50
表面声波延迟线	表面薄层中衰减决定的表面波长	1000	250
电磁波波导	光子能量决定的波长(集成光波导器件)	1000	400
电荷存储器	由击穿电场和信息随机现象决定的储存在表面的电子数	300	100

在现在的平面工艺微加工技术中主要的限制是图形的发生和图形的刻蚀，因为薄膜技术在深度方向能控制的尺寸和公差对整体器件(bulk devices)来说已足够精确。前面已讨论到，目前电子束光刻的能力可达 300nm 左右的线宽，边缘陡度为 10nm。由于电子在基片中的散射，进一步提高分辨率遇到困难。但是离子束光刻其分辨率可提高到 10nm，同时离子束又能用于溅射刻蚀并能直接写出图形。这说明现有的微加工技术对整体效应器件来说可能在今后 10 年中(本世纪末)接近其极限尺寸。并可预测，这期间也将有许多新的整体效应器件被发明、制造和应用。这些器件不仅限于微电子和光电子器件，也包括许多传感、化学分析和生物功能的器件。

但是应该看到，整体器件极限(大约 1000 个原子)与自然单元尺寸(即单个原子)之间有一个明显的间隔，在这范围现有的整体效应将不再有效，这就是分子尺度的器件。虽然自然界早已存在许多各种各样复杂的基于分子效应并以生物细胞形式出现的器件，但至今尚无能力进行人工复制，这主要由于对这种尺度所表现出的效应缺乏科学上的了解。可以相信在不太远的将来，现有这些工作原理已知的器件和仪器能帮助我们获得对分子尺度现象的了解。

作为微加工的一个里程碑的纳米加工(nanofabrication)的研究和开发不仅是为了适应整体效应器件制造的需要，而且更是发展下一代分子尺度器件所必不可少和应当导前研究的技术。

通常把尺寸从 100～1nm 的加工作为纳米加工的范围。显然，现有的微加工技术中有相当一部分可以满足 100～10nm 加工的需要，在采取特殊措施的条件，少数加工技术能达 10nm 以下。因此，要完全满足纳米加工的要求，不仅要通过对现有技术的进一步提高和改进，还需要不断开拓新的途径。

正如前面已指出的，可供微加工选用的工具几乎全都是粒子束，不论是光子、X-射线、电子或离子，还是聚焦的或展宽的，这些束都是以光刻技术为基础进行加工的。因此作为纳米加工的第一步是致力于对现有光刻技术和设备性能的改进。但是这种光刻技术最终会达到它的极限，因为光刻所用的抗蚀剂本身的分子尺寸和排列会对加工造成严重影响。

采用微聚焦束能直接在局部淀积、刻蚀或改性而不必依赖掩膜或光刻。图 41 表示聚焦离子束(FIB)用于无掩膜直接刻蚀和直接淀积[15]，这可用于 VLSI 中电路的修正和诊断。(a)是 FIB 将掩埋 A1 电极上的纯化层刻蚀出孔；而后由 FIB 轰击表面吸附的分子使分解成挥发性物质蒸发，留下不挥发的物质淀积成膜，如图 4(b)所示。这里对电子束、离子束或 X-射线的聚焦都是利用特定的透镜实现的，例如对电子束和离子束的聚焦可用电磁透镜，对 X-射线的聚焦

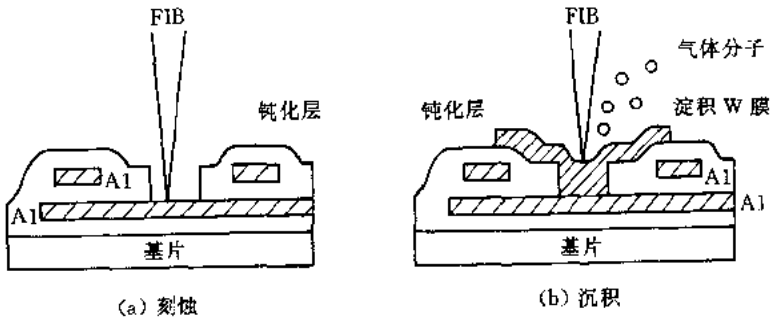


图 41　聚焦离子束无掩膜直接刻蚀和淀积

可用 Fresnel 带板或相位带板透镜。这种透镜聚焦束对毫微加工存在两个问题：①由于这些透镜没有光学透镜那样完美，聚焦束斑的直径受限制，例如目前达到的水平，对离子束来说大约 10nm，X-射线的聚焦不能小于 50nm，电子束的束斑虽可接近 1nm，但电子束在固体中的散射，使入射的束斑大大展宽。②聚焦电子束的能量往往在 10^3eV 范围，这样高的能量对分子或小原子团的影响难以得到很好的控制。

不同于透镜聚焦的另一种途径是邻近聚焦(proximity focusing)[16]。邻近聚焦束的获得是由于束源与靶非常贴近，因而束从源传输到靶的过程中不能产生有效的发散。扫描隧道显微镜 STM 是利用邻近聚焦最明显的例子。从极尖到样品的距离大约为 1nm，产生的隧道电子到达样品表面的束直径不超过 1nm，电子电流的空间限制是 STM 图像达到非常高的分辨率的关键。

获得邻近聚焦场发射源的重要基础是采用了压电精密位移装置。利用这种技术并结合点源发射器和用微加工技术制造的电子束微透镜，如图 42 所示[17]，以及其他类似的新装置是对电子束技术的根本改革，由于能获得低能高分辨率的聚焦束，对毫微加工和诊断的应用具有极大潜力。

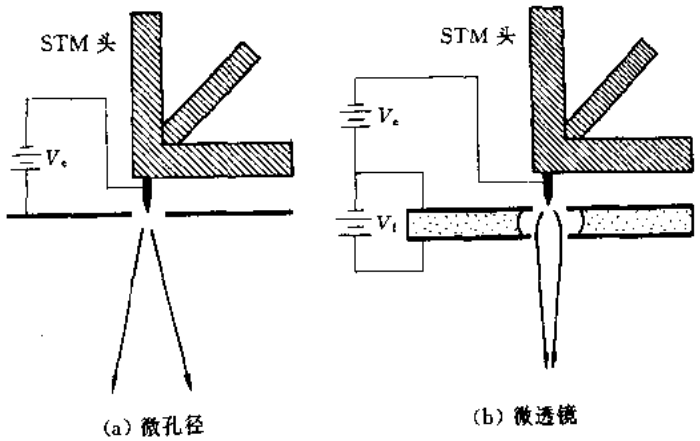


图 42　由 STM 头控制场发射极尖并带有简单孔径和微透镜的电子源

实际上任何点状源都能产生聚焦束，如果把源和样品贴得很近，离子[18]和光子[19]也能作邻近聚焦。但目前所得到的聚焦束直径尚在 10nm 以上，这比 STM 极尖或单原子点源所获得的有效束径(≤1nm)大得多。

这种荷电粒子的邻近聚焦束的重要优点是与常规的透镜聚焦束相比，相同束径的邻近聚

焦束粒子的能量低得多。粒子克服势垒从源表面逸出毋需很高的电势抽取。由于抽取电极(样品)非常贴近源电极,在十分低的外施电压作用下达到足够高的电场强度,大大减小了粒子从抽取电场获得的动能增量。所以需要低能粒子束,这是由于原子活化能低于10eV的某些过程,如迁移、断键和化学反应等要求能得到控制。隧道电子束的形成不是越过势垒而是穿过,因而基本上可获得零动能。邻近聚焦束的低能量、局域力和高电场这一系列吸引人的特点,有希望实现纳米加工的最终目标,即对个别的原子和分子按要求进行排列。新近最引人注目的报道[20]是利用 Van der Waals 力和静电力在低温条件下(4K)搬移 Ni(110)表面吸附的 Xe 原子排列成 I、B 和 M 字母,还将六个 Xe 原子排成间隔为 0.5nm 的直线链。这令人鼓舞的实验验证,可看成是进入分子工程时代的曙光。

参 考 文 献

[1] Watson.G. The 64 M-bit DRAM. IEEE Spectrum, 1991,28(1):30

[2] Bate R.T. Nanoelectronics. Nanotechnology, 1990,1(1):1

[3] Chen Y.X. A New Concept of 3-Dimensional Integrated Optics Optoelectronics-Devices and Technologies, 1990,5(1):109

[4] Howe R.T, Muller R.S, Gabriel K.J, Trimmer S.N. Silicon Micromechanics: Sensors and Sctruators on a Chip. IEEE Spectrum, 1990,27(7):29

[5] Muller R.S, Howe R.T. Technologies for Microdynamic Devices. Nano-Technology,1990,1(1):8

[6] Eckertova L. Physics of Thin Films. New York: Plenum Press, 1977.77

[7] X-ray Stepper Crafts Features Under 0.36μm. Electronic Design, 1991,39(11):23

[8] Greeneich J.S, Duzer T .Van. An Exposure Model for Electronsensitive Resists. IEEE Trans Electron Devices, ED-21,1974,(5):286

[9] Stengl G, et al. Ion Projection System for IC Production. J Vac Sci Technol,1883,1979,16:1883

[10] Rensch D.B, et al. Ion Beam Lithography for IC Fabrication with Submicrometer Features. J Vac Sci Technol. 1979,16:1897

[11] Seliger R.L, et al. A High-Intensity Scanning Ion Probe with Submicrometer Spot Size. J Appl phys Lett, 1979,34(5):310

[12] Focused Ion Beam Images Nanometer-Sized Objects. Electronic Design, 1991,39(12):28

[13] Binnig C, Rohrer H. Scanning Tunelling Microscopy. Surf Sci, 1985,152/153:17~26

[14] Binnig G, Quate C.F, Gerber C. Atomic Forec Microscope. Phys Rev Lett, 1986,56:930

[15] Namba S. Focused Ion Beam Processing. Nuclear Instruments and Methods in Physics Resarch B39, 1989,504~510

[16] Shedd G.M, Russell P.E. The Scanning Tunneling Microscope as a Tool for Nanofabrricstion. Nanotechnology. 1990,1(1):67

[17] Mccord M.A, Chang T.H.P, Kem D.P, Speidell J.L. A Novel Scanning Tunneling Microscope Controlled Field Emisssion Microlense Electron Source. J Vac Sci Technol, 1989,B7:1851

[18] Bell A.E, Rao K, Swanson L.W. Scanning Tunneling Microscope Liquid-Metal Ion Source for Microfabrication. J Vac Sci Tcehnol, 1988,B6:306

[19] Lieberman K, Harush S, Lewis A, Kopelman R. A Light Source Smaller than the Optical Wavelength. Science. 1990,247:59

[20] Eigler D.M, Schweitzer E .K. Positioning Atoms with a Scanning Tunnelg Microscope. Nature, 1990, 344:524

第二篇　研　究　篇

Planar and Channel Single-Mode $LiNbO_3$ Waveguides Fabricated by Ion Exchange

At present, most $LiNbO_3$ waveguides are fabricated by the titanium indiffusion process at 950 ~ 1000℃. The minimum node depth of such a single-mode waveguide (about 2 μm) is limited by the maximum diffusion time that can be allowed for a single propagating mode and the maximum prediffusion titanium film thickness that can be used without creating a residual TiO_2 layer after diffusion. The mode depth plays an important role in the coupling efficiency of semiconductor lasers to $LiNbO_3$ waveguides and in the defiection efficiency of electro-optical switches that have small electrode spacing. The in-plane random scattering noise of such a waveguide also limits the maximum dynamic range that can be obtained in optical signal processors such as the r.f. spectrum analyzer. It is most desirable to have an $LiNbO_3$ waveguide that has a smaller mode depth and/or less in-plane scattering noise.

Use of ion exchange was first reported by Shah[1] in 1975 and again by Jackel[2] However, Jackel's work is concerned primarily with multimode waveguides. We report here the results of the fabrication of single-mode waveguides using his method. Excellent attenuation (approximately 0.7 dB/cm)and in-plane scattering noise (for TE modes less than 30 dB at 3 mrad)were obtained. To be more specific, we shall report here the measured attenuation and Δn_{eff} of channel and planar waveguides at various temperatures and ion exchange times. In titanium-indiffused waveguides Δn_{eff}. is much smaller, indicating a smaller mode depth obtained by the titanium ion exchange process. Measured in-plane random scattering noise will be compared with that of the TE and TM modes in titanium-indiffused $LiNbO_3$ waveguides.

This work was supported in part by the National Science Foundation, USA

Reference

[1] Shah M L. Optical waveguide in $LiNbO_3$ by ion exchange technique. Appl Phys Lett, 1975,26:652

[2] Jackel J L. Ion exchange for optical waveguides in $LiNbO_3$ and $LiTaO_3$, Technical Dig. of Topical Neet. on Integrated and Guided-Wave Optics. 1980,WB4

Characterization of $LiNbO_3$ Waveguides Exchanged in $TlNO_3$ Solution

A single-mode $LiNbO_3$ waveguide with large Δn_e, high optical damage threshold, and small inplane scattering loss is very desirable for guided-wave optical signal processing. Large Δn_e implies that the optical waveguide will have a small depth suitable for efficient butt coupling to GaAlAs injection lasers[1] and for efficient electro-optical switching at large Bragg diffraction angles where the electric field penetration of the switching electrode may be very small[2]. Inplane scattering loss sets the background noise limit in planar waveguide signal processors such as the rf spectrum analyzer[3].

Ti-indiffused waveguides have a high sensitivity to optical damage, limiting their laser power handling capability, and it is not possible to obtain Ti-indiffused waveguides with mode depth less than 2 μm[1]. Waveguides obtained by Li_2O outdiffusion at high temperatures are able to handle considerably more laser power but have even larger mode depth ($\sim 5\mu$m). The use of ion exchange to form $LiNbO_3$ waveguides has been reported previously by Shah[4] and by Jackel[5]. Jackel reported that large $\Delta n_e = 0.12 \sim 0.13$, and a step in index profile can be created by this process. She also later reported[6] that such a $LiNbO_3$ waveguide would have a low optical damage sensitivity.

We have investigated the use of the ion exchange method in $TlNO_3$ for producing single-mode, large Δn_e, high damage threshold and low inplane scattering noise $LiNbO_3$ waveguides. Our results seem to be quite different than the results reported earlier, and they clarify some of the issues concerning the physical nature of the $LiNbO_3$ waveguides obtained in the $TlNO_3$ exchange process.

Following the reported procedure[5], we immersed x-cut $LiNbO_3$ samples (purchased from Crystal Technology Coup., Palo Alto, Calif., and double checked for crystalline orientation) in a molten solution of $TlNO_3$ contained in a quartz tube inside a furnace with a temperature control of ± 1℃. Table 1 shows the results obtained for some of the $LiNbO_3$ waveguides fabricated. It is clear that we were not able to obtain waveguides within the temperature range reported by Jackel[5]. n_e was determined at the 0.6328μm wavelength by measuring very carefully the angular position of the m lines coupled out from the output prism coupler. The accuracy of the measured angular position of the m line is 0.2 mrad. We deduce from these close n_e values that our n_e is much smaller than that reported by Jackel.

On the other hand, the attenuation rates of our samples, as shown in Table 1, are quite low(~ 0.6 dB/cm), comparable with those obtained in the Ti-indiffused and out-diffused waveguides. For the y direction of propagation, the measured inplane scattering noise(including the scattering noise of the prism input and output couplers) is less than 30 dB at 3 mrad angle. This is much better than the inplane scattering noise of the TE mode propagating in the x direction of a y-cut Ti-indiffused $LiNbO_3$ waveguide. In addition, we assessed the laser power handling capabilities of single-mode ion-exchanged waveguides at

Tab.1 Optical properties of $LiNbO_3$ waveguides fabricated by $TlNO_3$ exchange

Sample No.	Substrate	Processing			Modes	Attenuation dB/cm	n_{eff}	Remarks
		Melt	t/℃	T/h				
10	x-cut, y-prop.	$TlNO_3$	245	4	0			
11	x-cut, y-prop.	$TlNO_3$	245	26	0			
16	110	$TlNO_3$	340	7	2	1.4	2.2055,2.2022	
17	112	$TlNO_3$	340	21	3		2.2039,2.2026, 2.2022	
18	x-cut, y-prop.	$TlNO_3$	340	4	1	0.71	2.2022	
25	x-cut, y-prop.	$TlNO_3$	320	7	1	0.64	2.2014	
27	x-cut, y-prop.	$TlNO_3$	340	2	1	0.52	2.2014	
28	x-cut, y-prop.	$TlNO_3$	340	50	1		2.2039,2.2030	
29	x-cut, y-prop.	$TlNO_3$	340	21	4		2.2030,2.2022, 2.2018,2.2015	new $TlNO_3$
33	x-cut, y-prop.	$TlNO_3$	404	2	3			
34	x-cut, y-prop.	$TlNO_3$	414	2	1			
37	x-cut	$TlNO_3$	356	3	0			Sample put into the melt horizontally. Surface etched.

632.8 nm by the transmission method as recently reported[7]. We found that the performance of these waveguides was better than that of the Ti-indiffused waveguides and inferior to the out-diffused waveguides. At any given laser power, the power in the output beam of the waveguide degrades as a function of time, finally establishing a steady state in 20 ~ 200h. The steady-state power loss of an ion exchanged waveguide is given as a function of the laser power propagating in the guide in Fig. 1. It is shown in comparison to that of a titanium in-diffused and a Li_2O outdiffused waveguide.

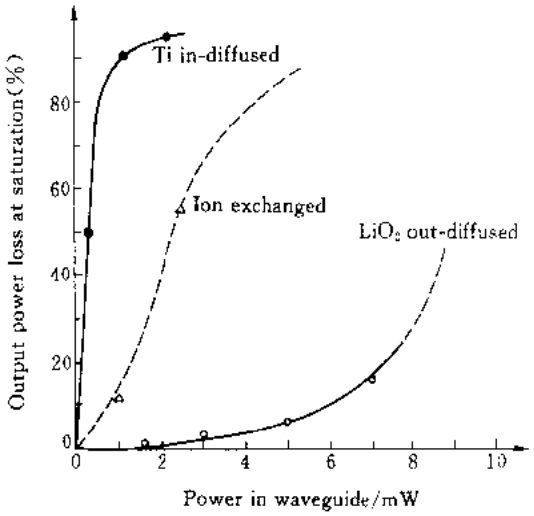
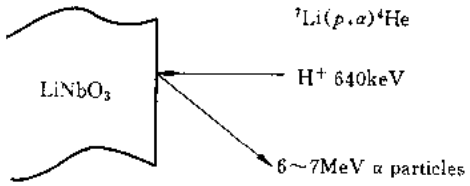


Fig.1 Effects of optical damage in $LiNbO_3$ waveguide

Intrigued by the discrepancy between our results and those obtained by Jackel, we proceeded to analyze(a)an unprocessed virgin sample, (b)our sample # 29, and(c)the $LiNbO_3$ sent to us by Jackel at Bell Laboratories that was used in Ref.[5] by MeV ion backscattering[8] and nuclear reaction techniques[9]. Measurements of the energy spectra of the backscattered ions from these samples with a 1.5 MeV $^4He^+$ ion analyzing beam in the random and x axis oriented directions showed the following results:(1) All three samples, (a),(b), and(c), were single crystals with reasonable crystalline quality as evidenced by channeling measurements[8]. This suggests that immersion of $LiNbO_3$ into a $TlNO_3$ bath did not cause any significant structural damage. (2) For samples (b)and(c), only barely detectable Tl signals from backscattering were found on the sample

surfaces. The estimated concentration of Tl from these signals is 5×10^{13} Tl/cm^2 (1 ~ 1/100 of a monolayer). This result indicates that Tl atoms did not migrate into the $LiNbO_3$ substrates after immersion in the $TlNO_3$ bath. (3) There was little or no change in Nb concentration in $LiNbO_3$ after the $TlNO_3$ immersion. There was some indication that the oxygen concentration had decreased as a result of immersion. However, we are not certain of the change of the oxygen concentration because of the low sensitivity of detecting oxygen by MeV $^4He^+$ backscattering. These results suggest that any observed waveguide effects cannot be due to Tl in-diffusion or Nb out-diffusion during the immersion.

Our next step of investigation was to utilize the $^7Li(p,\alpha)^4$ He reaction (using an 840 keV proton beam) to determine the Li concentration in the samples. Fig. 2 shows the α signal from Li atoms in samples(a)-(c). When we compare these signals with one another, the Bell Laboratory sample (c) clearly had a substantial loss of Li atoms (about 1/3). The scale sketched in Fig. 2 shows the depth profile of Li from the surface. We propose that the waveguiding of the sample reported in Ref. [5] may be provided due to a large amount of loss of Li or Li_2O over a depth of only 2 μm or less near the surface. In comparison, the signal from our sample(b) is very close to that of the virgin sample(a). Naturally, a small amount of Li or Li_2O out-diffusion over a depth of 5 ~ 10 μm will be difficult to detect. Our suggestion is in agreement with the weak loss-of-oxygen signals observed in backscattering



α signals from Li atoms in the samples
Each division of mg/cm² scale corresponds to 2~12 μm in thickness (for $LiNbO_3$)

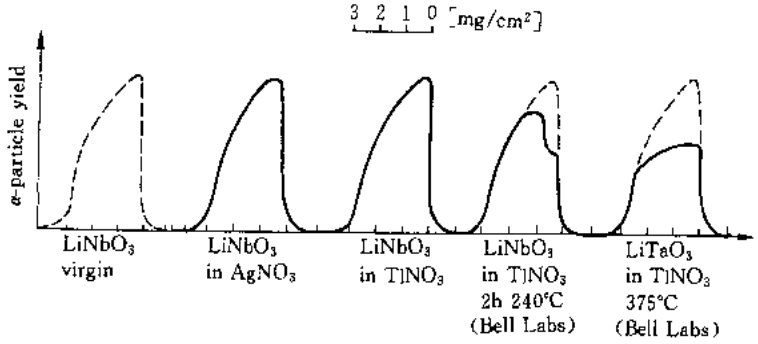


Fig. 2 α signals from Li atoms in the samples. The number of α particles is proportional to concentration of Li atoms in sample and the $^7Li(p,\alpha)^4$ He reaction cross section. The α particles produced in this reaction in deeper regions of sample have lower energy due to energy loss in the sample. The number of α particles produced in the deeper regions of a sample with uniform concentration decreases rapidly with increasing depth due to decreasing reaction cross section of lower proton energy

measurements since the loss of Li is most likely associated with the out-diffusion of Li_2O. It is in agreement with the observed high optical damage threshold, since Holman has pointed out[7] that Li_2O out-diffused waveguides have higher optical damage resistance if phase equilibrium and oxidizing conditions are maintained during the out-diffusion process. It is also in agreement with the observed low inplane scattering noise since low-temperature diffusion processes may create fewer optical defects. We have noticed further that samples exchanged in $TlNO_3$ at temperatures higher than 350℃ often have rough surfaces.

The unusual aspects of the thallium ion exchange are emphasized by comparison with the case of $LiNbO_3$ immersed in an $AgNO_3$ bath at 356℃ for 3 h. In that case, the α signal shown in Eig. 2 is almost identical with that of a virgin $LiNbO_3$ sample, and the channeled spectrum shows clearly the existence of Ag atoms as well as the distortion of the lattice near the surface. In Fig. 2, we have also shown the α signals obtained form a piece of $LiTaO_3$ that has been immersed in $TlNO_3$ at 375℃ at Bell Laboratories. Notice that there is again a substantial loss of Li atoms, but the depth of out-diffusion is about 5 ~ 10 μm, implying that the $LiNaO_3$ waveguide may have a much larger mode depth.

In conclusion, we have no explanation yet for the variations of the $LiNbO_3$ characteristics when it is immersed in $TlNO_3$. We have established, however, that the higher-index waveguiding is created by a loss of Li or Li_2O and not by an in-diffusion of Tl. This is contrary to the case of immersion in $AgNO_3$ where the in-diffusion of Ag seems to have played an impo tant role in creating the n_e. If we can succeed in developing a low-temperature process in which a substantial amount of loss of Li (or Li_2O) can be restricted to a small depth (e.g., 1 μm) and in which the surface of $LiNbO_3$ is not roughened by that process, then we would have obtained a much better $LiNbO_3$ waveguide than the waveguides obtained by either the Tl in-diffusion process or the Li_2O outdiffusion process at high temperatures.

The authors wish to thank J. Jackel of Bell Laboratories for discussion, and especially for furnishing us a chip of $LiNbO_3$ and $LiTaO_3$ waveguide for comparison with our samples. Yi-Xin Chen is a visiting scholar from Shanghai Jiao Tong University, the Peoples Republic of China.

References

[1] C.T. Mueller, C.T. Sullivan, W.S.C. Chang, et al. IEEE J. Quantum Electron, 1980, QE-16:363

[2] R.A. Becker, W.S.C. Chang. Appl. Opt., 1979, 18:3296

[3] D. Mugeriau, E.C. Malarkey. Technical Digest, Third International Conference on Integrated Optics and Optical Fiber Communication, April 27 – 79, 1981, San Francisco, Paper WH2

[4] M. Shah. Appl. Phys. Lett., 1975, 26:653

[5] J.L. Jackel. Appl. Phys. Lett., 1980, 37:739

[6] J.L. Jackel, D.H. Olsen, A.M. Glass. J. Appl. Phys., 1981, 52:4855

[7] R.L. Holman, P.J. Cressman. Technical Digest, Third International Conference on Integrated Optics and Optical Fiber Communication, April 27 – 29, 1981, San Francisco, Paper WB4

[8] W.K. Chu, J.W. Mayer, M-A. Nicolet. Backscattering Spectrometry Academic. New York, 1978

[9] W.K. Chu, J.W. Mayer, M-A. Nicolet, et al. This Solid Films, 1973, 17:1

$LiNbO_3$ Waveguides by Electrically Enhanced Ion Migration and a Comparison of Techniques

Abstract and Summary

Recently we have investigated the electrically enhanced migration technique for fabricating $LiNbO_3$ waveguides. With the aid of an analysis by an X-ray photo-electron spectroscopy (XPS) and a second ion mass spectroscopy (SIMS), we have found that the formation of a waveguide is mainly due to the migration of the Li atoms from the surface layer of the $LiNbO_3$ substrate.

The characteristics of three kinds of $LiNbO_3$ waveguide by different techniques have been measured for comparison, as given in Tab. 1. It shows that $LiNbO_3$ waveguides by ion migration are more resistive to optical damage (Fig. 1), with a higher index increase, a small mode depth and a nearly step-profile of effective index (Fig. 2).

It is also found that waveguides made by the ion-migration technique are compatible with Ti-indifussion waveguides. It is therefore possible to design waveguide devices with a multilayer structure and a certain planar pattern of various indices on a $LiNbO_3$ substrate.

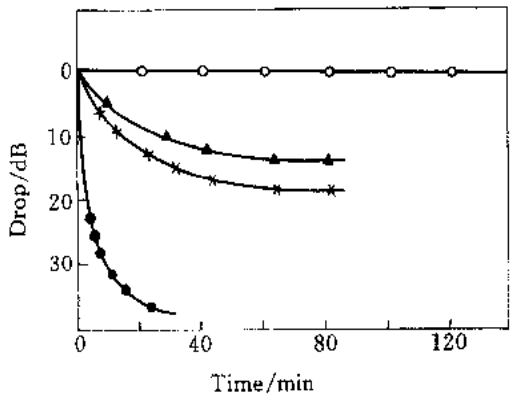


Fig. 1 Time-dependence of the drop of main peak of m-line

- o Ion migration
- ▲ Ti indiffusion in close crucible
- × Ti indiffusion in atmosphere
- • Ti indiffusion in wet oxygen

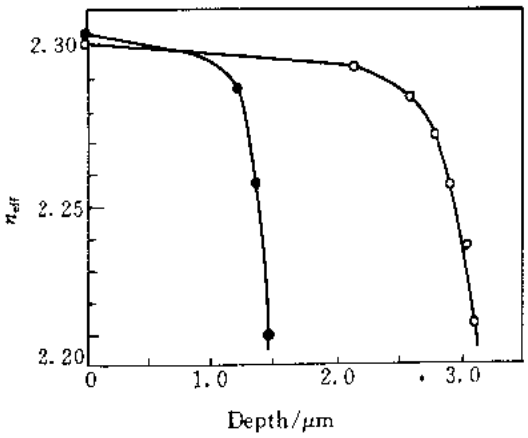


Fig. 2 Effective indices profile of ion-migration waveguide

Sample No. 15, 260℃, 5V, 125 min

- o Positive side
- • Negative side

Tab. 1 Comparison of various single-mode $LiNbO_3$ waveguides

Fabrication techniques	Ion-migration	Porton-exchange	Ti-In-diffusion
T/℃	260	200	1000
t/min	50	10	360
Δn_{eff}	0.04	0.06	0.01
$d_m/\mu m$	0.8 ~ 1.2	0.7	1.5 ~ 2.0
Attenuation/(dB·cm^{-1})	0.8	0.5	0.8
Scattering(mr, at 40dB)	1.7	1.3	2.5
Optical damage(632.8nm)	40mW, 5h, no damage	40mW, 5h, no damage	13mW, 2h, damage
n_{eff} profile	Nearly step	Nearly step	Complementary error function

A New Concept of 3-Dimensional Integrated Optics

It has been just 20 years since integrated optics was first named in 1969. In the passed two decades, not only has a lot of fruitful research works been accomplished in many laboratories, but also a number of integrated optical devices have begun to be used practically. With the rapid progress of technology in optical fiber communication and optical information processing, as well the fact that many new applications such as optical digital computing are widening unceasingly, it is becoming more and more clear that in a variety of fields light wave is superior to electric current as the media of propagation and processing of information. This is the reason why so many scientists and engineers are engaged enthusiastically in researching and designing the optical systems, devices, and materials for these applications. There is no doubt that these practical optical systems must be compact and reliable. Those requirements only can be met as the systems tend toward miniaturization, thin-film structure and integration. Hence, it is not surprising that the potential of integrated optics is great.

The term "integrated optics" implies that, first, the optical waveguide through which light beam can propagate effectively, the waveguide is formed on the substrate above which is a layer of media with higher index of refraction. Secondly, the variant functional devices can be structuralized by several kinds of waveguides: such as lasers, detectors, modulators, lenses, prisms, couplers and polaroids. Last, the practical optical systems or subsystems composed by the waveguide devices will be integrated on the same substrate. It is out of the question that proposal of the term "integrated optics" is really a great revolution compared to the huge and fragile traditional optical systems developed over the hundreds of years of history of optics.

Generally, the above concepts of integrated optics is two-dimensional; it can only deal with a one-dimensional spatial light beam. For instance, the spatial optical information obtained from of correlators, convolutors and spectrum analyzers of integrated optics using the acousto-optical effect of $LiNbO_3$ waveguides is only one-dimensional. However, most applications of mass data flow, image processing, artificial vision and optical neural computing systems are always with two-dimensional spatial optical signals.

In fact, the characteristics of propagation and processing of two-dimensional spatial optical signals with huge parallel light channels are the very places where optics has a prominent advantage over electronics. Thus, it is necessary to consider a new concept of three-dimensional integrated optics to meet those requirements. Of course, an integrated optical circuit has never been purely optical; it needs the supports given by various electronic circuits. The only difference is that the electronic circuits are separated or they can be integrated with optical circuits.

In principle expected three-dimensional integrated optical circuits should be the devices with multi-layer structures, every layer containing lots of optical components and devices. Those optical components and devices may be waveguide type or may not. For example, a two-dimensional microlens array which has been proposed to be made in the ion-exchange method is not a waveguidetype integrated optical device. Here, a light beam can go from an optical device to the others in the same layer; it can also propagate from one layer to another. The light beam propagating in the $x-y$ plane or along the z axis can be constrained in the waveguides. It can also propagate in free space or in media directly. For instance, a three-dimensional integrated optical circuit with different functional spatial light modulators (SLMs) for input and output, interconnecting, logic processing and memory probably could be used as a future optical parallel digital processing system.

In order to realize the new devices, three-dimensional integrated optics is confronted with a series of research topics both in theory and in experiment, such as coupling of waveguides and interconnecting of optical devices among the layers, the interaction of a large number of microbeams of light in free space or in waveguides; two-dimensional zonal growth of multi-layer heterogeneous materials, and various new physical effects, new materials and new devices for three-dimensional integrated optics. We have adequate reasons to believe that three-dimensional integrated optical devices will be realized eventually.

Nonlinear Integrated Optics

1 Introduction

The experiment on second harmonic generation by Franken and coworkers in 1961, in which red light from a ruby laser was frequency doubled into ultraviolet light, is generally considered as the beginning of nonlinear optics.

From the purely classical point of view, nonlinear optic can be traced back to much earlier times. Rayleigh well noticed the acoustical nonlinearities and had made extensive discussion of it in his famous book《Theory of Sound》. Lorentz could have made theoretical derivation of some nonlinear optical effects if he had allowed for a slight anharmonicity in his description of the electron as a harmonically vibrating bound particle. However, Lorentz lacked the stimulation of stimulated emission of radiation. It was after the invention of laser by Maiman in 1960 that one was able to achieve high light intensity required for experimental investigation and realization of nonlinear responses. From then on laser has been playing an very important role in the development of nonlinear optics.

The birth of laser also offered possibility to the initiation and development of guided-wave optics and integrated optics. The requirement from communications strongly stimulated the development of fiber technique, guided-wave optics as well as integrated optics. The first guided-wave nonlinear optical experiment was on second-harmonic generation in planar GaAs waveguides by Anderson and Boyd in 1971[1]. These early experiments took advantage of existing planar thin-film and microlithography techniques. Progress in this area of nonlinear guided-wave optics has concentrated on harmonic generation and parametric amplification, which utilize the large $\chi^{(2)}$ obtained in noncentrosymmetric waveguides. However, during the past few years there has been a great deal of interest in all-optical signal processing, which has stimulated work on third-order processes in integrated optics[2,3]. This evolution is in sharp contrast to the case of fiber nonlinear optics in which the usual fiber pulling technologies do not lead to noncentrosymmetric fiber cores and hence there is no second-harmonic activity. However, for fibers the propagation losses are so low that the small third-order nonlinearities in glasses could be compansated for by the long interaction length, and many third-order processes were reported. The initial experiments concentrated on phenomena such as stimulated Raman and Brillouin scattering and on self-phase modulation. The recent development of specialized fabrication techniques for producing crystal core fibers and the fortuitous discovery of second-harmonic generation in some glass fibers[4] have extended nonlinear fiber research into efficient $\chi^{(2)}$ processes.

Two salient features, i.e., high intensities and long interaction lengths, differentiate guided-wave

nonlinear optics form bulk nonlinear optics. To obtain high intensities, it is usually necessary to focus laser beams. However, in bulk media, the tighter the focus, the shorter the distance over which it can be maintained. Therefore there should be a trade-off in the efficiency between high intensity and interaction length. In a waveguide, however, the beam is confined in one(planar waveguide) or two(fiber or channel waveguide) dimensions to values of the order of the wavelength of light, for distances determined solely by the propagation losses. These are typically millimeters to centimeters in integrated-optics waveguides and meters to kilometers in fibers. Owing to the small cross-sections of beams in waveguides or fibers, very high intensities can be achieved even with moderate light powers. In the future it is possible to produce nonlinear effects in waveguides even with incoherent light. From what we have mentioned above, we conclude that optical waveguides are ideal for nonlinear interactions because they provide strong beam confinement over long propagation distances. From the practical point of view, the applications of optical nonlinear effects are expected to be realized with integrated and fiber devices. This review will concentrate on nonlinear integrated optics.

2 Nonlinear Optics in Waveguides

When one or several incident optical fields interact to each other in a medium, the induced polarization can be expanded in terms of the products of the interacting fields, giving

$$p^{\mathrm{NL}}(r,t) = \varepsilon_0\Big[\iint_{-\infty}^{\infty}\chi^{(2)}(t-t_1,t-t_2):E(r,t_1)E(r,t_2)\mathrm{d}t_1\mathrm{d}t_2 + \iiint_{-\infty}^{\infty}\chi^{(3)}(t-t_1,t-t_2,t-t_3):E(r,t_1)E(r,t_2)E(r,t_3)\mathrm{d}t_1\mathrm{d}t_2\mathrm{d}t_3\Big]$$

where $\chi^{(2)}$ and $\chi^{(3)}$ are the macroscopic second and third order susceptibilities respectively. These tensors depend on the materials and can cause obvious wavelength dispersion. This special expression applies for non-instantaneous responses of the material to the incident fields. For instantaneous response, for incident field.

$$E(r,t) = (1/2)E(\omega)\exp[\mathrm{i}(\omega t - \beta z)] + c.c$$

and the induced nonlinear polarization field

$$p^{\mathrm{LN}}(r,t) = (1/2)p^{\mathrm{NL}}(\omega)\exp[\mathrm{i}(\omega t - \beta z)] + c.c$$

the Fourier component of the polarization is

$$p^{\mathrm{NL}}(\omega) = \varepsilon_0 D^{(2)}(1/2)\chi^{(2)}_{ijk}(-\omega;\omega_1,\omega_2)E_j(\omega_1)E_k(\omega_2) + D^{(3)}(1/4)\chi^{(3)}_{ijkl}(-\omega;\omega_1,\omega_2,\omega_3)E_j(\omega_1)E_k(\omega_2)E_1(\omega_3)$$

where D's are degeneracy factors.

Optical waves can be guided by one or more surfaces when they are coupled into the media at the condition of resonance. For planar, channel or fiber waveguides these resonances are essentially geometric, being electromegnetic standing waves within the confined space. The waveguide modes correspond to the allowed resonances. Among various resonances, only the geometric ones can propagate for a long distance so that the guided wave can be of practical use to nonlinear optics.

Three commonly used optical waveguide structures are shown in Fig. 1 along with the corresponding

field distributions.

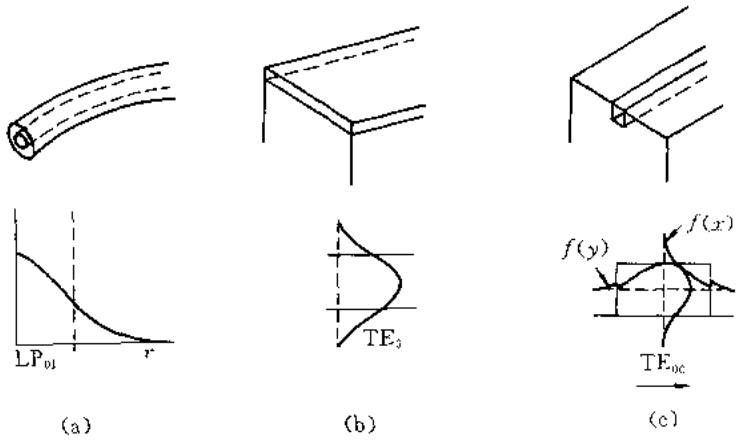


Fig. 1 Three common optical waveguide structures and corresponding field distributions

Channel waveguides and fibers have many similar features. The fields propagating in them can be expressed as

$$E_i(r,t) = (1/2) f_i^{(m,n)}(x,y) a^{(m,n)}(z) \exp[\mathrm{i}(\omega t - \beta^{m,n}) z] + c.c$$

where the guided wave wave vector $\beta = k_0 n_{\mathrm{eff}} = \beta^{m,n}$ is denoted with two integers m and n corresponding to two resonance conditions in two dimensiong. Here $f_i(x,y) = f_i^{(m,n)}(x,y)$ is the transverse field distribution of the ith component, normalized with

$$\iint f_i(x,y) \mathrm{f}_i^*(x,y) \mathrm{d}x \mathrm{d}y = 1$$

and $a(z) = a^{(m,n)}(z)$ is the mode amplitude, usually normalized to give the guided wave power as $|a(z)|^2$.

There are two orthogonally polarized mode families. For channel waveguides, $\mathrm{TE}_{(m,n)}$ denotes the modes with their major electric fields parallel to the surface, that is y-axis and $\mathrm{TM}_{(m,n)}$ denotes the modes with their major electric fields perpandicular to the surface, that is along x-axis. Both families of modes have only very small field component along the other two axes.

For fibers, the electric field distributions have circular symmetry. In almost all cases the approximate solutions of linearly polarized modes $\mathrm{LP}_{(m,n)}$ are adopted. Here one integer m denotes the cylindrically symmetric resonances across the core and another integer describes resonances along a circular path around the core center, expressed as $\exp[\mathrm{i}n\phi]$

For planar waveguides, the field distributions and dispersion relations are much simpler than that for channel waveguides and fibers. For example, the field is a plane wave along the y-axis, i.e., in the plane of the film, the field distribution is a function of x only and only one integer is enough to defined the mode. Thus one can substitute (m) for (m, n) in the above expressions and the modes are expressed as $\mathrm{TE}_m(E_y = 0, E_x = E_z = 0)$ and $\mathrm{TM}_m(E_x = 0,\ E_z = 0,\ E_y = 0)$.

Most nonlinear optical interactions involving plane waves can be analyzed with the slowly varying phase and amplitude approximation. For guided waves, the corresponding method is known as the coupled

mode theory which gives the growth rate of the complex amplitude of the signal beam as[5]:

$$\frac{\mathrm{d}}{\mathrm{d}z}a^{(m,n)}(z) = \mathrm{i}\frac{k^2(\omega)}{2\varepsilon_0\beta^{(m,n)}}\frac{\int_{-\infty}^{\infty}P_i^{\mathrm{NL}}(x,y)f_i^{(m,n)}(x,y)\mathrm{d}x\mathrm{d}y}{\int_{-\infty}^{\infty}|f^{(m,n)}(x,y)|^2\mathrm{d}x\mathrm{d}y}\exp[\mathrm{i}(\beta^{(m,n)}-\beta_p)z]$$

Basically, this described the average over the transverse coordinates of the projection of the nonlinear polarization onto the output guided wave field profile. It is this average over the transverse coordinates that introduces the overlap integral.

3 Second-Order Nonlinear Optical Guided-Wave Devices

Many $\chi^{(2)}$ phenomena have been demonstrated in planar guided-wave formats, including second-harmonic generation, difference-frequency generation, optical parametric amplification, and optical parametric oscillation. Of these, second-harmonic generation (SGH) has by far been the most widely studied interaction. The development in the last few years has been faster than that of the last two decades. An important factor is the possibility of frequency doubling of IR light from GaAs laser diodes into blue light for data storage and copying.

For the simplest case of SHG, a single fundamental guided wave is excited at $z=0$, propagating to $z=L$ where it leaves the waveguide along with the second harmonic generated between 0 and L. The second-harmonic power, $P(2\omega,z)$, is

$$P(2\omega,L) = (k_0L)^2\frac{d_{\mathrm{eff}}^2}{n_{\mathrm{eff}}^3}\frac{\sin\psi}{\psi^2}|K|^2P^2(\omega,L)$$

The incident (and harmonic) field is written as

$$E(r,t) = \frac{1}{2}\hat{e}E_i(\omega_i,\beta_i)f_i(x,y)\exp[\mathrm{i}(\omega_i t-\beta_i z)]+c.c$$

Where $f(x,y)$ is an appropriately normalized modal field. The waveguide figure of merit is given by $d_{\mathrm{eff}}^2/n_{\mathrm{eff}}^3$, where n is the effective index defined by $n_{\mathrm{eff}}k_0=\beta$.

In the expression for $P(2\omega,L)$, the overlap integral K, a concept unique to waveguides, is given by

$$K = \int_{-\infty}^{\infty}\frac{d_{ijk}}{d_{\mathrm{eff}}}E_i(2\omega)E_j(\omega)f_i(x,y)f_j(x,y)f_k(x,y)\mathrm{d}x\mathrm{d}y$$

If the product of the field distributions changes sign across the waveguide, interference effects occur in the integral, reducing the value of K. This overlap integral is usually small unless all the interacting modes have the same mode number. Fig. 2 shows two possible cases for SHG in slab waveguides. In the case of $TE_0(\omega)+TE_0(\omega)\rightarrow TE_0(2\omega)$, the optimum K can be obtained from the field product. On

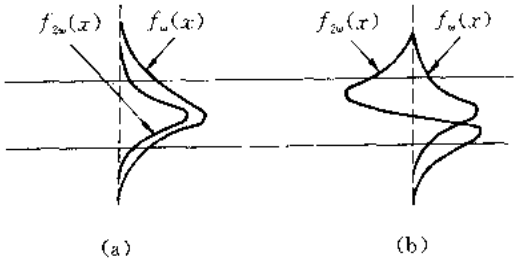


Fig.2 Two possible cases for SHG in planar waveguides
(a) $TE_0(\omega)+TE_0(\omega)\rightarrow TE_0(2\omega)$;
(b) $TE_0(\omega)+TE_0(\omega)\rightarrow TE_1(2\omega)$

the other hand, in the case of $TE_0(\omega) + TE_0(\omega) \rightarrow TE_1(2\omega)$, the field profiles contribute to K both positively and negatively, resulting in a small value of K. This leads to the conclusion that optimized efficiency of SHG can be obtained when both the fundamental and harmonic waves are in the modes of lowest order.

Another charactristic of guided wave interactions is the wavevector(phase) matching condition, that is $\psi = 0$, where

$$\psi = (1/2)L[\beta(2\omega) - \beta(\omega)] = (1/2)k_0 L[n_{\text{eff}}(2\omega) - n_{\text{eff}}(\omega)]$$

There are three commonly used methods of phase matching, as shown in Fig.3. The first method takes advantage of the birefringence of materials. When both orthogonal polarizations exist, the dispersion in frequency between the fundamental and harmonic waves is compensated for by material birefringence, $TE_0(\omega) + TE_0(\omega) \rightarrow TM_0(2\omega)$, as shown in Fig.3(a). Since there exist mixed polarizations, an off-diagonal nonlinear tensor element is required. In the case of quasi-phase-matching, either the refractive index or the nonlinearity is spatially modulated with period of Λ, as shown in Fig.3(b). Λ is chosen so that $2\beta(\omega) \pm 2\pi/L = \beta(\beta)$ holds. The third method is known as Cerenkov which requires that $2n(2\omega)\omega/C > 2\beta(\omega)$, as shown in Fig.3(c). Under this condition the second harmonic wave is no longer guided wave. It is radiated into the substrate at an angle θ which satisfies con $\theta = \beta(\omega)n_{\text{eff}}\ C/\omega n(2\omega)$. In this case, the overlap integral will include the overlap between guided and radiation fields.

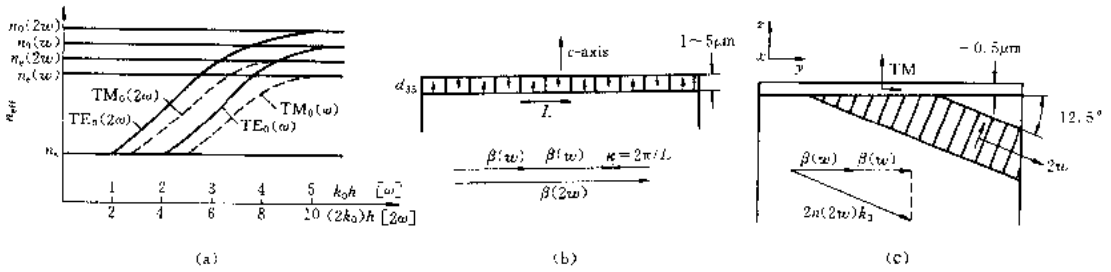


Fig.3 Three methods of phase matching for SHG in optical waveguides
(a) Material birefringence; (b) Quasi-phase-matching; (c) Cerenkov method

SHG in optical waveguides is most widely investigated experimentally with $LiNbO_3$ waveguides. In the case of quasi-phase matching and Cerenkov methods one makes use of the large d_{33} coefficient which is 7 times greater than d_{13} used for birefringent phase-matching. The normalized efficiencies obtained are $45\%/(W\cdot cm^2)$ and $170\%/(W\cdot cm^2)$, respectively. $170\%/(W\cdot cm^2)$ means that 17mW output of blue light can be generated for a 1 cm long device when the near-infrared input power is 100 mW. With quasi-phase-matching technique, the best result is $230\%/(W\cdot cm^2)$ achieved in domain-inverted KTP waveguides. It is possible to achieve an efficiency greater than $1000\%/(W\cdot cm^2)$ with nonlinear organic waveguides. Excellent results were also obtained in $LiNbO_3$ waveguides for difference-frequency generation, optical parametric amplification, and optical oscillators. For the optical parametric oscillators a maximum conversion efficiency of 15% was obtained for 15 W pump power at $\lambda = 0.6\ \mu m$. For parametric amplifiers, a gain of 16 dB was achieved with a pump ($\lambda = 0.65\mu m$) power of 150 W (pulsed). In the case of difference-frequency generation, efficiencies of 10^{-4} have been measured for

pump power of 50 mW ($\lambda = 1.4 \sim 1.6\mu m$) and idler power of 100 μW ($\lambda = 3.39\mu m$).

4 Third Order Nonlinear Guided-Wave Devices

The three-order optical nonlinear phenomena is a kind of all-optical interaction, that means the nonlinear polarizations are determined by the products of three optical fields.

In contrast to $\chi^{(2)}$ phenomena, for which imaginary components are undesirable since they imply loss, both real and imaginary components of $\chi^{(3)}$ can lead to interesting phenomena. For example, the imaginary part gives rise to stimulated Raman and Brillouin scattering, whereas the real part is responsible for the intensity-dependent refractive index, parametric mixing, degenerate four-wave mixing, etc.. Research in all these third-order guided-wave phenomena started in optical fibers, and degenerate four-wave mixing, coherent anti-Stokes scattering, and a variety of phenomena that depend on the intensity-dependent refractive index have also been investigated in integrated-optics waveguides.

About ten years ago it was realized that standard integrated optics devices could be operated in an all-optical mode by introducing waveguide media with intensity-dependent refractive indices. Such devices can be used for all-optical signal processing at speeds limited only by the "turn-off" time of the nonlinearity and material systems do exist with subpicosecond response times. Most of the present integrated optics devices, such as couplers, modulators and switches, rely either on wavevector matching between coupled fields, or interference effects between two guided waves whose relative phase has been modulated externally (typically via the electrooptic effect). Essentially every such linear device can be converted into an all-optical device by utilizing nonlinear materials in the waveguiding regions. To date, grating and prism couplers, grating reflectors, directional couplers and Mach-Zehnder interferometers have all been investigated.

When operated in an all-optical mode the guided wave wavevector βk_0 ($k_0 = \omega/C$) in all of these devices depends on the guided wave power, that is $\beta \rightarrow \beta(P)$. This is a consequence of waveguiding media having an intensity-dependent refractive index $n(I)$ with

$$n(I) = n_i + n_{2i}(I)I$$

Where I is the local intensity and the subscript i identifies the ith medium. This leads directly to an effective index given by

$$\beta = \beta_0 + \Delta\beta_0(P)P$$

and

$$\Delta\beta_0 = k_0 \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\, n_0^2(x,y)\, n_2(x,y)\, |f(x,y)|^4$$

Because the field is a maximum inside the guiding channel, $\Delta\beta_0$ is maximized when the nonlinear medium is the guiding medium. This also leads to the minimum operating power as well as to the best use of the available nonlinearity. Such a power-dependent wave vector can be used to alter a wave-vector conservation condition optically. For example, the Bragg condition and hence the reflectivity of a grating structure can be tuned optically. Similarly, the phase change accumulated by a guided wave after a propagation distance L also varies with the guided-wave power. Hence the interference between guided

waves can also be tuned optically.

A selection of all-optical guided wave devices, e.g., directional coupler, distributed-feedback grating. M-Z interferometer, mode sorter and prism coupler, as well as their response to optical power are shown in Fig.4. Also shown is the response of corresponding linear devices for the purpose of comparison. The harpest switching characteristics are obtained with the first two devices, namely, a nonlinear directional coupler and a nonlinear Bragg reflector. The response shown for the nonlinear directional coupler is typical of a whole class of devices involving the nonlinear coupling between two copropagating modes, usually derived for two weekly coupled parallel channels. The response of the nonlinear distributed-feedback grating is typical of two counterpropagating waves whose amplitudes are coupled.

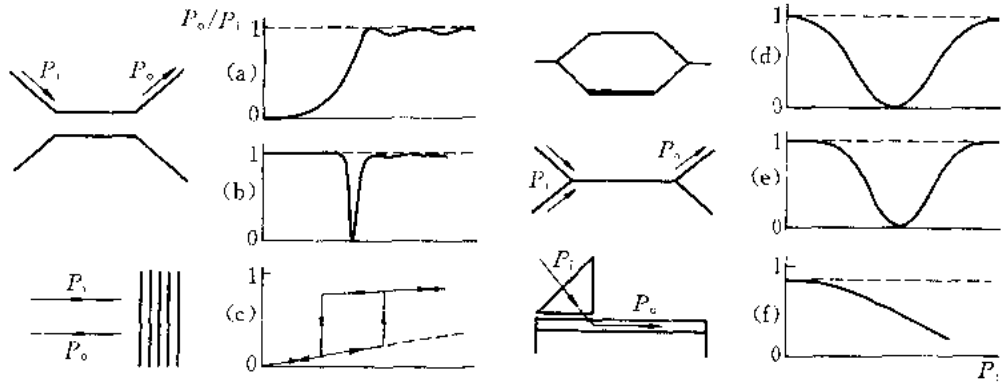


Fig.4 Integrated opticsdevices and their response fo optical power with and without nonlinearities

(a) 1/2-beat-length directional coupler; (b) 1-beat-length directional coupler;
(c) Distributed-feedback grating; (d) Mach-Zehnder interferometer;
(e) Mode sorter; (f) Prism coupler

5 Nonlinear Optical Waveguide Materials

The best results on SHG and related $\chi^{(2)}$ nonlinear interactions have been obtained with Ti-indiffused $LiNbO_3$ waveguides. This is because that this waveguide system is being successfully used for applications in electro-optic switching and phase modulators. The losses are of the order of 0.1 dB/cm in good waveguides, allowing multicentimeter interaction lengths. But there are also problems with nonlinear optics in $LiNbO_3$ waveguides. The material undergoes damage in the visible region of the spectrum at milliwatt power levels. Therefore research continues into better materials. Tab. 1 lists a number of different materials that are representative of material classes with potential for guided-wave

Tab.1 SHG figures of merit relative to $LiNbO_3$ for materials with potential for waveguide applications

Materials	$d_{eff}/(\times 10^{-8}$esu)	n	d^2/n^2
$LiNbO_3$	1.2	2.3	1
KTP	1.0	1.8	1.5
MNA[a]	7		75
NPP[b]	0.2		600
(PS)O - NPP[c]	2		
DCV/PMMA[d]	7		85

a) Metanitroaniline; b) N-(4-nitrophenyl)-L-prolino l; c) Chromophore functionalized Polymer; d) Dicyanovinyl azo dye polymethyl methacrylate

applications.

High-quality waveguides have already been made in KTP by ion exchange and SHG demonstrated. Although the improvement in the figure of merit over $LiNbO_3$ is only a factor of 1.5 this material is not easy to optical damage, which is the main problem with $LiNbO_3$, and is phase matchable over a much larger wavelength range.

A recent development is that of electric-field-poled molecules in polymeric glassy hosts. Molecules with large dipole moments (to facilitate molecular alignment in a d_c field) and large second-order hyperpolarizabilities are loaded into a host glassy polymer placed between two electrodes. The composite material is heated above its glass transition temperature, and a strong d_C electric field is applied to align the molecules, which are more or less free to rotate in the host. The sample is cooled with the electric field still on, and the molecules are locked into position with a preferential orientation. Typically the molecules constitute 10% ~ 50% of the material volume. As is evident from Tab. 1, large improvements in the figure of merit relative to $LiNbO_3$ have already been demonstrated. One of the problems being currently addressed is stabilization of the molecular orientation for long periods of time; in all the systems reported to date, there are two characteristic relaxation times for molecular orientation in the host. One occurs on a time scale of minutes to hours, and the second over months. Locking in of the molecular orientation by chemical bonding appears promising.

The major requirements for $\chi^{(3)}$ nonlinear materials which can be used in all-optical guided wave devices are: 1) the processing speed, 2) the fraction of power which can be switched, 3) the operating power, 4) heating effects, and 5) the device throughput. The minimum nonlinear phase shift, $\Delta\psi^{NL}$, and the minimum dimentionless material parameter W required for a number of typical nonlinear guided-wave devices are listed in Tab. 2.

Tab.2 Minimum nonlinear phase shift, $\Delta\psi^{NL}$, and mininum dimensionless material parameter W (>80% transmission) required for various nonlinear guided-wave devices

Nonlinear device	$\Delta\psi^{NL}$	W
Directional coupler	4π	10
1/2 beat length		
Directional coupler	$\cong 3.3\pi$	8
1 beat length		
Mach-Zehnder interferometer	2π	5
Distributed-feedback grating	π	2.5

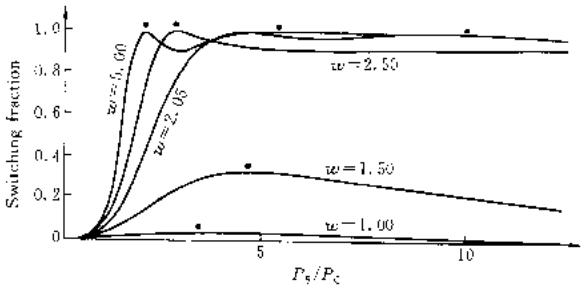


Fig. 5 Variation in the fraction of the incident power switched to the cross channel in a nonlinear directional coupler versus the saturation parameter w for various detunings of a tow-level saturable-absorber model

For any real material there is an upper limit to the refractive-index change that can be induced optically. This can be caused by saturation of some physical process for example, band filling in a semiconductor, or may be due to material damage in a glass or organic material. The effect of a saturable refrative index on the nonlinear directional coupler response is shown in Fig. 5. The parameter ω =

$\Delta\beta_{sat}L/\lambda$ clearly is critical, and $\omega > 2$ is required for complete switching, where L is the device length. This parameter does not, however, include the device throughput. High throughput requires that $I\alpha < 1$, where α is the attenuation coefficient. This leads to a materials figure of merit for a nonlinear directional coupler defined as

$$W = \Delta n_{sat}\frac{L}{\lambda}\frac{1}{I\alpha} = \frac{\Delta n_{sat}}{\lambda\alpha}$$

Where it is assumed that $\Delta\beta_{sat} = \Delta_{nsat}$, that is, that the saturation ineffective index is equal to the saturation in material index.

To be morespecific, we define $\Delta\phi^{NL}$ as the minimum phase change that is required from a single decoupled waveguide over the device length L. When this waveguide is used in a device configuration, usually a smaller nonlinear phase shift is actually used. However, the waveguide must be capable of $\Delta\phi^{NL}$ for the device to work properly.

Tab. 3 contains a summary of the pertinent material quantities for intensity-dependent nonlinear guided-wave devices. A large n_2 is important for low switching powers. The maximum operating speed of a device for serial pulses is determined by the nonlinarity response time. Any nonlinearity, even a thermal one, can be turned on instantaneously. However, in order to minimize cross talk it is important that the nonlinear change in index induced by one pulse relax back to zero before the next pulse arrives. Therefore the nonlinearity relaxation time is a critical material parameter.

Tab.3 Figures of merit, $\Delta n_{sat}/\alpha\lambda$, of different nonlinear materials for application of third-order nonlinearities to guided-wave devices

Material system	$n_2/(m^2\cdot W^{-1})$	τ/s	α/cm^{-1}	Δn_{sat}	W
Semiconductors					
GaAlAs(r)	-10^{-8}	10^{-8}	10^4	0.1	
(nr)	-10^{-12}	10^{-8}	30	$\cong 2\times10^{-3}$	0.9
(nr, theory)	-10^{-13}	10^{-8}	10	0.01	10
Doped glasses					
CdS_xS_{1-x}	-10^{-14}	10^{-11}	3	5×10^{-5}	0.3
Organics					
PTS(r)	2×10^{-15}	2×10^{-12}	10^5	$\cong 0.1$	
(nr)	10^{-16}		0.1	$>10^{-3}$	>100
Others(nr)	$10^{-16}\sim10^{-17}$	10^{-14}			
Glasses					
SiO_2(nr)	10^{-20}	10^{-14}	10^{-5}	$>10^{-6}$	$>10^3$

6 Measurements of Waveguide Nonlinearities

One of the precursors to implementing nonlinear waveguide devices is the measurement of the waveguide nonlinearity, including its magnitude, sign, and speed. There are two options. One is to measure the film nonlinearities on transmission through the film, for example by degenerate four-wave mixing. Alternatively, the diagnostic experiments can be carried out using guided-wave modes. This

includes nonlinear prism and grating coupling, degenerate four-wave mixing, pump-probe transmission measurements (for absorptive nonlinearities), intensity-dependent birefringence and hybrid Mach-Zehnder interferometers. Of these, the nonlinear prism coupler has by far been the most common diagnostic technique used experimentally.

It is well known that distributed coupling of an incident beam via a prism is most efficient when the projection onto the surface of the incident field wavevector matches that of the guided wave. Otherwise, the coupling efficiency is reduced. As the in-coupled guided wave power grows with propagation distance for an initially optimized coupler, the guided wave wavevector changes as follows:

$$\beta = \beta_0 + \Delta\beta_0(P)\Delta P$$

where P is the guided wave power. Hence the wavevector matching condition is lost and the coupling efficiency is reduced. Therefore one of the salient features of the nonlinear coupler is a decrease in coupling efficiency with increasing incident power[11].

The technique of degenerate four-wave mixing has been used for many years in plane wave nonlinear optics to measure the magnitude and the "turn-off" time of nonlinearities. It does not, however, yield the sign of the nonlinearity. Implementing this technique for waveguides involves guiding all four beams, as shown in Fig. 6[12]. The four-wave mixing signal leaves the interaction region along the direction of the probe beam, but traveling in the opposite direction. Assuming propagating pump (P_1 and P_2) and signal (P_3) beams traveling at right angles to each other, all L wide, the conjugate beam (P_4) guided wave power is given by

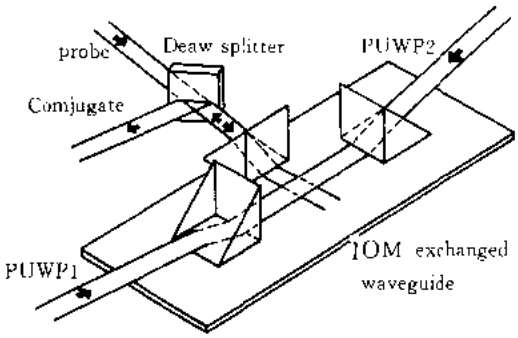


Fig.6 Waveguide degenerate four-wave mixing geometry

$$P_4 = |D^{NL}|^2 P_1 P_2 P_3$$

with

$$D^{NL} = j\frac{k_0 C^2 \varepsilon_0^2}{2}\frac{2+\cos^2\theta}{3}\int_{-\infty}^{\infty} dx n_2(x)|f^2(x)|^2$$

for all TE polarized waves. Hence the conjugate power is proportional to n_2^2.

The pump-probe technique for probing nonlinear waveguides is appropriate for studying nonlinearities related to absorption changes and hence can be investigated by monitoring the power dependence of the guided wave attenuation[13]. Examples are two-level saturable media, interband transitions in semiconductors, etc. In these cases, the absorption is bleached out at high powers. A simple example of how the absorption varies with intensity is

$$\alpha(I) = \alpha_b + \frac{\alpha^{NL}}{1 + I/I_{sat}}$$

where I_{sat} is the saturation intensity. Thus at high powers, the waveguide attenuation drops to the intrinsic waveguide value α_b. The experimental set-up is shown in Fig. 7. First, a high power, appropriately short pulse (relative to the nonlinearity "turn-off" time) is propagated through the

waveguide. It bleaches out the nonlinearity. The attenuation of a second weak probe pulse is measured as a function of delay time between the two pulses. Thus the fraction of the total loss attributable to the nonlinearity, and the relaxation of the nonlinearity can be measured.

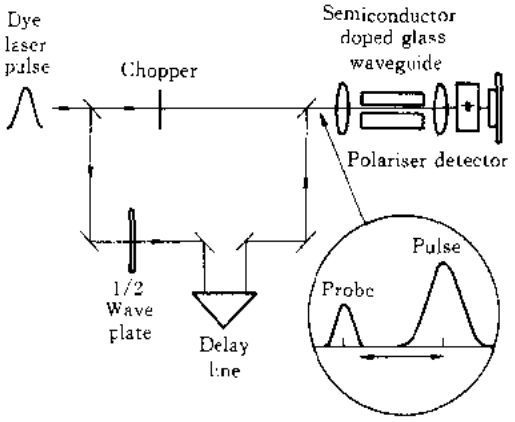


Fig.7 Pulse-probe experimental arrangement

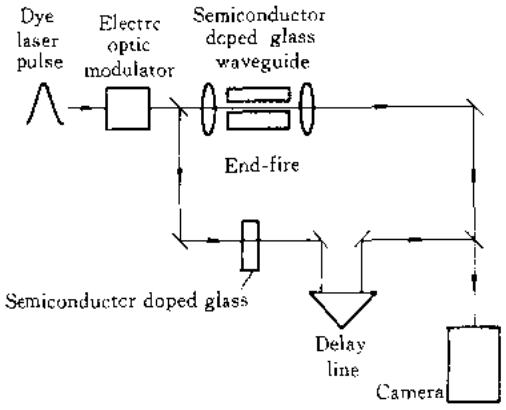


Fig.8 Nonlinear hybrid waveguide Mach-Zehnder interferometer

The nonlinear hybrid Mach-Zehnder interferometer technique was first used to measure the nonlinearities in MQW GaAs/GaAlAs strain-induced waveguides. The experimental set-up is shown in Fig. 8[14]. When one of the arms consists of a waveguide with an effective index which depends on the optical power, the power-dependent change in index can be measured via the power-dependent interference which occurs at the detector. Furthermore, the sign of the nonlinearity can also be determined from the direction which the fringes move with increasing power.

7 Research Fields and Prospect

Nonlinear guided wave optics is a rich and wide research field. It is a rapidly developing branch of nonlinear optics. Here the key point is that in waveguide devices large intensities can be maintained along great propagation distances, which is the key to efficient nonlinear interactions. In addition, beams are confined in regions of wavelength dimensions, which means that relatively low powers lead to large enough intensities, so that devices can be operated at sub-watt peak power.

Waveguide structures have led to the investigation of a number of new nonlinear optical effects which are very difficult to achieve in bulk media. Much work has to be done in order to put various nonlinear effects in practical use in waveguide devices. The researches that have to be carried out are: 1) Optimization of devices, including the simulation of device operation and the investigation into their characteristics; 2) Exploration of new materials. There will probably never be an "ideal material": that is simultaneously ideal for all applications. It is more realistic to optimize material properties for special $\chi^{(2)}$ and $\chi^{(3)}$ guided-wave devices; 3) Research on the application of nonlinear waveguide devices. In the next few years the application of nonlinear integrated optics devices to all-optical switching, all-optical signal processing and optical computing will be greatly developed.

References

[1] D.B.Anderson, J.T.Boyd. Appl. Phys. Lett., 1971, 19:266 ~ 268

[2] G.I.Stegeman, R.Zanoni, N. Finlayson, et al. IEEE J.Lightwave Technol., 1988, 6:953 ~ 970

[3] G.I.Stegeman, R.H.Stolen. J.Opt. Soc. Am., 1989, B6:652 ~ 662

[4] V.Osterberg, W.Margulis. Opt. Lett., 1986, 11:516 ~ 518, 1987, 12:57 ~ 59

[5] D.Marcuse. Theory of Dielectric Optical Waveguides. New York: Academic Press, 1974

[6] E.J.Lim, M. M.Fejer, R.L.Byer. Electron. Lett., 1987, 25:174

[7] K.Chikuma, S.Vmegki. J.Opt. Soc.Am., 1990, B7:768

[8] T.Ishigame, T.Sunara, N.Nishihara. Opt. Lett., 1991, 16:375

[9] J.D.Bielein. Digest of 1991 Integrated Photonics Research Topical Meeting, ThC1, Washington, OSA, 1991

[10] H.Hermann, W.Sohler, J.Opt. Soc. Am., 1988, B5:278

[11] C.Liao, G.I.Stegeman, et al. J.Opt. Soc. Amer., 1985, 590 ~ 594

[12] G.I.Stegeman, C.T.Seaton, C.Karaguleff. IEEE J.Quantum. Electron, 1986, QE-22:1344 ~ 1348

[13] C.N.Ironside, T.J.Cullen, et al. J.Opt. Soc. Amer.B., 1988, 5:492

[14] Li Kam Wa, P.N.Robson, et al. Electron, Lett., 1986, 22(21):1129

光通信传输速度的进展引人注目

0 引言

每年一次盛大的美国光通信会议不仅吸引着来自世界各地大批从事这一领域研究和开发的科学家、教授和工程技术人员，而且还招来了一大群厂商。第18届美国光通信会议OFC'95从2月26日～3月3日在美国圣地亚哥召开。会议最突出的主题是如何提高光纤通信的速度。会上发表论文200多篇，另有31篇特邀报告，9个辅导报告(Tutorial)，31个短课，7篇专题讨论报告。

1 光纤

在光纤测量方面，英国EG&G公司Barlow评述了测量偏振模色散(PMD)的不同方法，报道了由纤芯中两个正交的偏振模本征群速差别所引起的PMD效应，还分析了由于光纤在安装中随机弯曲和扭绞所产生的非本征效应。目前，光学时域反射测量仪(OTDR)普遍用于测量光纤的衰减特性，在如何扩大OTDR的动态范围和改善空间分辨率方面已有许多报道。日本NTT公司Sato等人提出了一种称为啁啾OTDR的新方法，即将间隔相同但波长不等的多个探测脉冲进行压缩，再将压缩后的多个脉冲合成为单个探测脉冲，得到的结果是探测脉宽在50ns时，其空间分辨率可达2.5m。

当前光纤制造厂试验光纤涂层材料的物理性能时多数采用薄膜的形式，而不是接近其实际结构的最终形式。尽管这样做是出于方便的原因，但这种薄膜形式是否能很好模拟这种管壁形式，包括它的几何形状和厚度呢？Corning公司Botelho介绍了一种方法可制备最长为25.08cm的合成涂层管。它是将光纤产品放在液氮流中冷却，然后用商品光纤剥离器将其涂层剥落。采用这种新方法可对使用期内任何光纤的任何一点上的涂层进行试验。

掺杂光纤放大器是这次OFC会议的主要议题。AT&T贝尔实验室的Delavaux等报道了环形器形式的多级掺铒光纤放大器(EDFA)。在三级环形器中测得两顺序端之间平均损耗和隔离分别为1.5dB和70dB。采用环形器设计将允许把上一级剩余的泵浦能量用于下一级，并同时藉引入级间滤波器以减少放大的自发辐射。掺铒光纤放大器现阶段另一个担忧的问题是多波长系统的增益不相等。日本住友电气工业公司的Kashiwada等人解决这一问题的方法是采用不同掺杂的EDFA作为级联段来进行补偿，以Al共掺杂的光纤段其增益在较长的波段(1550～1560nm)提升，而以Al和P共掺杂的光纤段的增益在较短的波段(1535～1550nm)增大。把这两种光纤段交替连接并调节不同的长度，他们在1543～1558nm获得了平坦的增益轮廓线。

对于1550nm窗口的商品级光纤可用掺铒得到最佳的光纤放大器。同样，对于1300nm可

以通过掺镨得到。是否可采用某种方法能在所谓第一窗口，即 850nm 附近建造光纤放大器？回答是肯定的。英国 HP 公司的 Dye 等报道了已全部工程化的掺铥光纤放大器(简称 TDFA)，其发射光谱中心频率为 806nm。这种器件能获得 20dB 增益，最大输出功率达 1110dB。采用的泵浦波长为 780nm，这是现有商品化的 GaAlAs 激光二极管很普通的波段。即使是塑料光纤也已做成了光纤放大器。日本 Keio 大学的 Tagaya 等用有机染料掺杂聚合物渐变折射率光纤放大器(POFA)在短的放大段上得到非常高的功率增益。这些器件固有的高输出功率和高增益(比稀土元素钕和镨掺杂的石英玻璃光纤要高 10^4 倍)，是由于大多数有机染料有大的发射和吸收截面以及塑料光纤的直径比石英光纤大。采用 Rhodamine B 作为掺杂剂、调 Q 倍频的 Nd:YAG激光器为泵浦和工作在 590nm 附近的染料激光器为信号源，他们从 1kW 的泵浦脉冲得到高达 620W(峰值功率)的放大输出，从 560～600nm 可得到 20dB 的增益。可能的应用，包括聚合物光纤局域网和低功率激光二极管的放大器等。

2 光电子器件

掺镨光纤放大器之所以受到重视，一个很重要的原因是它能使世界上现有的大部分在 1.3μm 窗口工作的单模光纤网得到升级。但要做到这一点，必须开发 1.02μm 波长、性能优良的泵浦源。日本 NTT 光电子实验室 Nishiya 等成功地开发了高功率廉价的 1.02μm 应变 InCaAs 量子阱激光二极管。制造这种激光器采用 5.08cm(2in)圆片工艺，激光管发散角相当小，在 18°～9°能很方便地与单模光纤耦合封装，已得到耦合光纤的输出功率为 200mW 量级。即使对掺铒光纤放大器，关于到底采用 980nm 激光管还是 1480nm 激光管作为泵浦源的争论仍在继续中，特别是对现在考虑用于大容量波分复用多级放大系统。美国 AT&T 贝尔实验室的 Lucero 等报告了 980nm 和 1480nm 泵浦的放大器链(每链含 4 个放大器，带有 21.7dB 或 26.7dB 的级际衰减器，以分别模拟 106km 和 131km 的光纤段)的对比，实验结果与数值计算符合良好。对 1480nm 泵浦在光纤段损耗为 26.7dB 情况下，信噪比在整个波长范围比较平坦；而对 980nm 泵浦在光纤段损耗为 21.7dB 时，较为平坦。另一种用 1480nm 泵浦 EDFA 的方法是采用铥掺杂的 ZBLAN 光纤放大器，由二极管泵浦的 Nd:YAG 或 Nd:YLE 激光器进行泵浦。Amoco 激光公司的 Smart 等对由 Tm:ZBLAN 光纤制成的 1480nm 光纤环和光纤光栅激光器作了比较，在 2W 泵浦功率下两类器件都能产生几百毫瓦的 1480nm 输出功率，耦合技术在这里具有重要作用。光纤光栅线性腔具有比环激光器具有较窄的线宽，这使后者更适合于远距离泵浦 EDFA 的应用。

关于 WDM 系统光源的选择当前主要权衡高速还是单片结构。如果需要高速，目前不得不选用分立的激光器，对于多个单片激光二极管(即在同一芯片上有两个或更多不同发射波长的二极管)的调制速率已达到每秒数百兆位。英国 Bath 大学的 Asghari 等报道的一种单片集成多通道激光器的新设计，采用了多通道光栅腔，已达到两个通道的间隔为 2nm，每通道的输出功率为 4.3mW，串扰小于 30dB，调制速率超过 1Gb/s。

会上有许多篇报告涉及传输速率方面。Illionis 大学的 Chang 等人介绍了一种全集成离子注入的 GaAs MESFET 探测器，截止频率为 30GHz MESFET 的结构尺寸为 0.6μm×100μm。全集成的光电子集成电路包含一个 MSM 探测器结构，一个互阻抗放大器以及两个限制的后放大器，具有 2.5Gb/s 的性能。美国圣巴巴拉加州大学的 Petersen 等开发的行波光前端探测器具有

从 30.003～30GHz 的宽带性能。器件基于背面发光的 InGaAs/InP p－i－n 光电子二极管与 50Ω 的共面传输线相集成，可以获得优化的最高量子效率，具有带宽为 45GHz。

日本 NTT 大规模集成电路实验室的 Sano 等在特邀报告中评述了 10Gb/s 和更高速率的光波通信所需的集成电路的格式。对于任何一种高速集成电路，无论是线路、器件或是封装都必须有调和统一的最高性能。在 10Gb/s 和以上的光波通信 IC 的世界中，包括复杂的反馈设计，低抖动的时间电路，多芯片金属封装以及类似的技术。近两年，不同的研究组（包括驱动电路、多路复用器、放大器、解码器及其他）已在 20GHz 和更高的硅双极、GaAs MESFET、HBT 和 HEMT 技术方面有着不同性能的“里程碑”。同时他们指出，超过 10GHz 的全集成器件至今尚未实现。当接近 40GHz 时，信号的互连变得十分重要，因为即使在同一芯片内，三维分布效应和寄生影响还是较严重。即使如此，40Gb/s 的光波通信 IC 期望能在不久的将来实现。

高速发送器也是另一个感兴趣的领域。AT&T 贝尔实验室的 Johnson 等报道了应用分布反馈（DFB）激光器与电吸收调制器单片集成实现了 10Gb/s 信号的发送。尽管 10Gb/s 信号在超过 50km 的标准光纤上发送已有几个小组报道过，但他们多是采用分立的或半分立的器件，或应用聚酰亚胺介质减少寄生电容。而 AT&T 这一研究是在连续的有源层上应用选择性区域生长技术制成调制和激光器集成器件，减少了工艺过程，改善了可靠性。

对塑料光纤系统也不断提高传输速率。日本 JEC 光电子研究所的 Miyasaka 报道了用于高速、渐变折射率塑料光纤数据连接的高速低阈值电流量子阱 AlGaInP 二极管激光器和高速 GaAs/AlGaInP p－i－n 光电二极管探测器，工作速率已达 4Gb/s。这表明用于数据传输的 650nm AlGaInP 激光器有重大的改进。

3 系统和网络

以往的几届 OFC 会议较多注重于光纤和发送/接收器件，而较少涉及整个系统的方面，特别是高速系统。这是很自然的，因为建立系统以前必须开发出所需的元器件。这次会议的重心已明显移向系统和网络。

但是仍有一种器件需要进一步开发。如何把特定的数据流输入运行的网络，如何将数据流取出？如果此网络是 WDM 网络，那么某一种可调滤波器就成为必不可少的。会上有 6 篇这方面的报告。最普遍的一种可调滤波器是声光可调滤波器（AOTF），突出优点是能多波长同时工作，这比其他几种途径具有更大的灵活性。美国 Bellcor 和联合技术研究中心等单位的 Jackel 等人介绍了一个多波长 AOTF 运行中出现的问题及其解决方法。问题是当多路通道同时调节时其开关效率有所下降，并且通道偏移互相靠近在一起，更严重的是偏移的度数是通道间隔的函数，把它们减到最小的方法是采用两个独立的、声光波导通道并在通带中开关每一个其他的波长，然后在声道间的空隙内放置声吸收器，能有效使通常平坦化。

IBM 公司的 Li 等人介绍了另一种用于分组交换 WDMA 网络快速可调波长选择器的形式。他们指出过去报道的许多系统，包括：光栅型波长分离器、光电二极管列阵、确定那一个列阵单元受光照的选择电路及探测电路等。大多数分组交换协议要求波长通道能快速重构，在 1.2Gb/s 系统中要低于 100ns。他们还指出在目前用于多光电探测器的功率消耗对性能有限制（因为光电二极管元件数增加时不得不减小反馈电阻以保持固定的带宽）。

有些系统，例如通过长距离单模光纤连接的高速 LAN 既要求时分复用，又要求波分复用。

美国南加州大学的 Norte 等验证了用半导体光放大器构成的全光 TDM 对 WDM 的数据格式转换器，把 1571nm 的 800Mb/s TDM 的数据流转换成在 1550nm 附近的 4 个 200Mb/s WDM 通道。一种类似的两级器件能把 800Mb/s 的数据流分级成单个 200Mb/s 数据流。

在今天很难确切地说出未来的光纤网到底怎样，但可预想其构造可能相当复杂。英国 BT 实验室的 Cotter 等企图构想未来的 100Gb/s 网络：没有连接，由低功能复杂度的基本元件组成、能自组织，采用超高速光逻辑门和较高水平的电子处理器，具有最低限度的软件基础，有较高的智能化的外部设备。这一设想中的网络不仅与今天已有的网络非常不同，也与目前正在开发的初级光路不一样。达到这些速度的关键将取决于每个分组头部的路由信息的单功能节点操作的全光再生。

4 传输速度的竞争

MIT 林肯实验室、AT&T 贝尔实验室和数据设备公司的联合研究组报道了在波士顿市区有 90km 距离的 20 通道二级无源波长路由的全光试验床。20 通道载波的频率间隔为 50MHz，各通道用光端机进出以提供 155Mb/s，1.244 和 2.488Gb/s 电路交换的“A”服务或每通道以 1.344Gb/s 合计速率提供时分复用/频分复用同时用的“B”服务。试验床能处理多种协议，诸如光纤分布数据界面，以太网、同步光纤网、异步传输模式以及数字和有线电视等。

AT&T 贝尔实验室的研究组用 8 个 5Gb/s 不回零通道在 8000km 长的距离上传输，总容量为 40Gb/s，此传输实验应用一个 8 波长的发送器和一个 1000km 掺铒光纤放大器链环形群。8 个分布反馈激光器使波长范围扩展到从 1556～1959.7nm，通道间隔为 0.53μm，采用一组定向耦合器提供多路复用。放大器链用 20 段 45km 长的负色散光纤和两段正色散光纤。在光谱中没有观察到有回波混频效应出现。8 通道的平均位误码率优于 2×10^{-10}，Q 因子的范围从 16.3～18.1dB。AT&T 贝尔实验室还有一批工程师正致力于在 2400km 的环路上以 2.5Gb/s 速率进行 8 通道的波分多路复用传输，即使在 8 个通道上都不用频导滤波器，无误码（小于 10^{-10}）的传输距离可超过 10Mm，有些通道的无误码传输距离达 12～14.4Mm。

日本 NTT 光网络系统和光电子两个实验室联合研究了在 1000km 的光纤线路上应用色散补偿光纤和带有增益均衡器的 980nm 泵浦掺铒光纤放大器，传输 16 通道 10Gb/s 频分复用的光信号。各通道的波长范围 1549.53～1559.13nm。

另一个 AT&T 贝尔实验室的研究组报告了 8 个 20Gb/s 的通道用常规光纤传输了 232km 距离，中心波长在 1555nm，各通道的间隔为 1.6nm，系统每隔 80km 用一个放大器。

NTT 光网络系统实验室宣布他们达到了最高的传输速率。研究者报告了单通道单偏振、200Gb/s，时分复用的光传输实验用了 2.1ps 光脉冲。他通过预定标的时钟直接从 200Gb/s 信号驱动一个全光的复用解调器。100km 的传输线包含两段 40km 和一段 20km 的色散位移光纤，在 40km 和 80km 处装有在线放大器。全部光纤的零色散波长是 1.558μm。

以网络和应用为导向的光纤通信

1996年的光纤通信会议(OFC'96)于2月25日~3月1日在美国加里福尼亚州的圣荷西市召开。会上发表论文约300篇。此外,还开展了专题讨论、短课、辅导课、论坛等交流活动。本次会议与历年的相比,其内容更突出了光纤的网络和应用导向。OFC'96的技术主席加拿大通信研究中心的K.Hill指出:"我们正致力于使会议向更多的应用方面转移"。大会报告有两个。Netscape通信公司的总裁J.Barksdale分析了当前邮电工业面临的需求急剧增长的形势,特别是Internet用户迅速扩展,使国家电话网的负荷大大加重,这就要求光纤系统能提供更大的带宽。AT&T海底系统公司的总裁W.Cartor介绍了海底光纤系统在全球信息基础设施中的重要作用。海底光纤网络的世界市场十分巨大,同时对这种长距离的光纤网提出了更高的性能指标,要求采用更新的技术。

1 美国光纤网

近年来,美国先进研究项目局(ARPA)在光网络和有关技术,包括波分多路复用(WDM)和时分多路复用(TDM)的研究开发方面已有重要部署。第一阶段从1993年开始安排了两个WDM联合试验项目:AON (All-Optical Network)和ONTC (Optical Networks Technology Consortium)。AON以MIT的林肯实验室、AT&T及Digital设备公司为首;ONTC由Bellcore、Nortel、哥伦比亚大学、休斯公司、Lawrence Livermore国家实验室、Rockwell公司和联合技术公司组成。在1995年的OFC会议上,ONTC成功地演示了四结点的全光网络的许多功能。有一项ONTC的关键技术(网络的接入组件)就在今年的会上演示,它的功能有WDM多路复用和多路解调,以及插分(add/drop)复接功能。虽然第一期的努力成功地显示了网络的技术可行性,但没有提出价格上的竞争性,包括器件制造和网络管理两方面,也没考虑如何在商业网络与军用网络之间相互沟通。

1995年ARPA开始了光网络研究的第二期计划,试图为上述问题找出答案。为此组成了4个联合体,MONET、NTON、WEST和ICON。MONET (Ultiwavelength Optical Network)以Bellcore和AT&T为首,并有Bell Atlantic、Bell South、Pacific Telesis、国家安全局和海军研究实验室参加。NTON (National Tramsparent Optical Network)主要由ONTC原来的成员组成,新增加了Sprint和Pacific Bell,但Bellcore不再参加。WEST的队伍由Rockwell公司、Ortel公司、加州理工大学以及在洛杉矶、圣地亚哥和圣巴巴拉的三所加州大学分校所组成。ICON由IBM和Corning公司组成。这些联合体在各前沿如器件技术,区域和干线网络结构,网络管理和控制,以及现场试验的准备等方面,都取得了不同程度的进展。

在器件方面,研制成一种新型液晶波长路由开关,并作了演示;传输OC-48的八通道激光器列阵的制造工艺有了改进;波长转换的不同方法正在研究中。在区域网交换方面,从光纤色

散的考虑提出了一种分级星形系统结构，为了模拟网络的管理和控制，正在开发一种端到端的模拟软件包。现场试验方面，在会议展览厅中，演示了第一代 OC-48WDM 光网络。这个光网络的关键技术有：4λ/8λ 波分多路复用，每波长的传输码率为 2.5Gb/s；谱宽为 1546～1560nm，相邻波长的间隔为 2/4nm；OC-48SONET 为主干，声光可调滤波器作为波长切换开关；集成的 8λ 的接入组件，副载波多路利用终端等。该试验床网络将改建成联结伯克利加州大学、Pacific Bell(在 San Ramon 市)、Lawrence Livermore 国家实验室以及 Sprint（在 Burlingame)等湾区 4 个地点的交换中心通话(hub-spoke)网络。并计划在 1997 年 6 月再在 Berkeley 和 Burlingame 间增加一条光缆，可构成各种 WDM 的环形和网形网络。

MONET(多波长光网络)联合体，首先由美国的 AT&T 和 Bellcore 等五家公司组成，后来国家安全局(NSA)和海军实验室(NRL)也相继参加。MONET 计划的目标是要把网络结构、器件技术、网络管理和商业应用进一步开发、证实和集成，以实现大容量、高性能、价格合理和可靠的多波长光网络在全国或全球范围提供商业和政府应用方面的服务。这一光网络可以支持现有的和未来的各种通信标准，将区域交换网、长途干线网和专用网联结起来，能够以光信号直接进网，实现宽带，与格式、码率和协议无关。

2 欧洲光纤网

ACTS(先进通信技术及服务)计划是欧洲委员会在 1995～1998 年期间支持邮电通信领域的预研和技术开发的主要计划。其主要内容与早期的 RACE 计划相结合，在 ACTS 计划中进行现场试验，可以利用欧洲共同体中各国的国家级设施。计划在技术方面攻克 6 个关键领域：①互相有关的数字多媒体；②光子技术；③高速网络；④移动和个人通信网络；⑤智能网络和服务工程；⑥通信服务和系统的质量、保密和安全。

在第二个光子技术领域，共列出 22 个研究项目，有 120 个不同的组织参与，总预算达 173MECU(其中 81MECU 由共同体提供)。内容十分广泛，从光电子器件的开发，通过光互连和分组交换，到全光网络试验。高速传输项目的范围从 10～40Gb/s，采用电子的 TDM、OTDM 和孤子技术。用户网络项目中包括宽带 PON(无源光网络 Passive Optical Network)、光纤和同轴电缆以及光纤和绞对电缆等混合系统。采用 WDM 路由技术的光网络光传输层的主要研究项目，其中重要的子系统有插分多路复用和光交叉连接(Optical Cross-Connects，OCC)。

参加该计划的成员分别参加 PHOTON、OPEN、METON、WOTAN 和 COBET 等 5 个项目组。PHOTON (Pan Eruopean Photonic Transport Overlay Network)项目将跨过奥地利和德国的边界，将维也纳、Passau 和慕尼黑在一星形网络中采用 WDM OCC，在传输距离达 500km 的网络中采用的码率可达 10Gb/s，有 8 个波长复用，每通道间隔约 400GHz，参加此项目的有西门子，BBC 和菲力浦等 12 个公司和组织。

OPEN(Optical Pan European Network)项目将在两个现场试验段上提供透明的传输服务，第一个系统将连接挪威的 Oslo 和丹麦的 Hjorring 和 Thisted，以验证类树状拓扑结构的路由。第二个系统用来研究联结巴黎和布鲁塞尔的四结点拓扑结构，以评估光透明性的极限，采用四通道的 WDMOCC 将传输 STM-16 的数据。在实验中将验证在 WDM 和 OTDM 系统之间的一个全光接口。此项目由 14 个组织参加，包括两所大学。

METON(Metropolitan Optical Network)项目利用斯德哥尔摩的 Gigabit 网来连结。通过一个

中央 ATM 交换点连接起来的两个 WDM 环上的 5 个远程 ATM 多路复用器站，将采用 4 个波长复用，间隔为 4nm，在远程站中由光插分复用器提供波长的选择。在中央交换点有一个 OCC 对整个网络进行波长路由的控制，服务层、ATM/SDH 路径层以及光传输层将得到验证。有 11 个组织参加此项目。

WOTAN(Wavelength-agile Optical Transport and Access Network)项目将利用英国连接几个城镇，包括剑桥和 Norwich 等的 East Anglican 实验网。为了验证用于干线网互连中主要的 ATM 交换、用户网和干线网(例如 LAN 与 LAN 之间的连接)中用户到用户的光连接，还将试验一个波长灵活的 TDMA 宽带 PON 系统，通道间隔为 1.6nm 左右。另外，各 PON 之间全光 WDMA/TDMA 分布交换也将在实验中验证，通道间隔为 0.4nm。

COBNET(Corporate Optical Backbone Networrk)项目将利用与上述 WOTAN 相同的试验网设备，通过公共网中的"清洁通道"连接不同地点的两个用户建筑物网 CPN(Customer Premises Network)。每个 CPN 将包含一个 WDM 环供距离超过 2km 的结点用和一个 SCM 环供比较靠近的结点用。

3 海底光纤网

这次会议对海底光纤网络相当重视。由于国际邮电通信业务急剧扩大，自 1985 年开始敷设海底光纤通信系统以来，在过去的十年中全世界建造的海底光纤网已达 20 万 km。这些数字网已对全球通信的容量和质量产生了深刻的影响。

1995 年开始，采用最新的光子技术和网络技术，在世界范围内安装新一代的海底光纤网的工程将先后在本世纪末和下世纪初完成，这将使国际通信网变为真正的全球海底通信网。美国 AT&T 贝尔实验室在会上报道了 6 个正在建设中的海底光纤网，有的已接近完成。这些网的通信容量十分大而且在安装运行后还可不断升级扩展，例如 TAT-12/13 网将在 1996 年年内完工。该网络启用时的容量将是 1988 年第一条跨大西洋光纤系统 TAT-8 的 32 倍，在安装好的一段上进行试验已表明这新的光纤网能采用 WDM 技术来升级，因此实际的通信容量最少还可增加 1 倍。

这些网络中有许多还将具有自动变更路由的能力，以提供不中断的服务，例如总长为 24000km 的光缆环围绕太平洋的 TPC-5 网将能自动改变通信路由以确保通信不受阻碍。如果有船在抛锚时将美国到日本之间的光缆损坏，通信将自动地改由经过夏威夷和关岛的光缆代替。这一路由切换过程的全部时间不超过 300ms，丝毫不妨碍用户的正常通话。

其他的网如 12000km 的亚太光缆网 APCN(Asian Pacific Cable Network)，环球光纤网 FLAG(Fiber Link Around the Globe)，非洲一号网 AONEN(Africa-ONE Network)以及泛美光缆网 PACN(Pan American Cable Network)等，将比以前更多的国家和地区通过海底光纤网提供全世界范围的通信服务。并且，由于采用最新的纤维光学技术，例如 WDM 等，这些海底网络将能进一步扩大容量，具有更强的适应性和更高的服务质量。

表 1 列出了太平洋区域已建和将建成的海底光缆系统，从建成的年代、系统的长途和采用的主要技术，可以清楚的看出海底光缆发展的趋向和前景。

表 1 太平洋区域的海底光缆系统

光缆系统	建成年份	系统总长/km	主要技术
TPC-3	1989	9 060	280(420)Mb/s
HJK	1990	4 570	$\lambda = 1.3\mu m$
NPC	1991	9 450	光/电再生系统
TPC-4	1992	9 840	560Mb/s
ASEAN	1991/2	4 500	$\lambda = 1.5\mu m$
APC	1993	7 500	光/电再生系统
TPC-5	1995/6	24 500	5.3Gb/s
APCN	1996	12 080	$\lambda = 1.5\mu m$
FLAG	1997	27 300	光放大系统

4 宽带用户接入网

用户接入网指的是以交换局或远端站到用户的连接，一般这种用户线的距离不超过 1km，不仅要求保证高的通信质量和可靠性，还要价格便宜。当前构成接入网的方式有光纤、电缆(对绞线或同轴线)和无线，在技术上和经济上各有利弊。早在十年前，人们就希望能把光纤直接通到用户(即 FTTH)，但是这种接入网的价格高。最近几年比较一致的看法是把光纤通到路边(即 FTTC)，最后的几百米由同轴电缆通至用户，称为光纤同轴混合网(HFC)，具有优化的技术和经济指标。

早期提出的无源光纤网(PON)是广播结构。不久，鉴于容量的限制和网络的运行考虑，开始提出波分复用无源光纤网(WDMPON)。但是，由于这种系统所采用的元器件十分昂贵，使这种方案无法实用。近几年来在元器件上已有新的突破，诸如采用"波导光栅路由器"(WGR)构成无源 WDM 分波器、WDM 接收器和 WDM 光源等。以 WDM 激光器为例，研究包括各种宽带可调的原理和机制，采用可调布拉格光栅、不同谐振波长的多腔共用、固定波长光源的集成阵列等，都有实验验证。这就可以使系统的性能提高，价格降低，付诸实用。

光纤/同轴混合网(HFC)因兼有模拟光波馈送技术和廉价宽带同轴电缆的优点现已在宽带用户接入网中普遍使用。但这种传统的 HFC 网主要是为分布广播服务而设计的。通过采用双信件传器(diplexers)和分离上行(upstream)放大器先提供传统的 5 ~ 40MHz 的上行频带来实现双向传输。但是，这种窄的上行带宽，以及在此频段上入口噪声所引起的通道减少，限制了它能提供的服务，而且乐观的系统设计已指出总的系统带宽可能达到 1GHz，实际上受到同轴放大器(550MHz 或 750MHz)的限制。为了解决那些上行限制和把传统的 HFC 升级已满足高容量的要求，美国 AT&T 的 Bell 实验室提出了一种小光纤站的光纤/同轴混合网(mFNSHFC)方案并进行了验证。这种小光纤站(mFN)只需采用价格便宜的非冷却的 F-P 激光管和 PIN 光电管，因而不需大量增加投资。把光纤更深入地引进同轴分布网，不仅可克服上述各种的限制，而且利用频带下行(downstream)的传输使系统的带宽升级，不再受同轴放大器的限制，可达到 700 ~ 1000MHz。

5 传输速率的新突破

有两篇文章报道了光波在光纤中的传输速率的新突破。

日本富士通实验室在150km，1.3μm零色散单模光纤上实现了1.1Tb/s的WDM传输。WDM有55个波长，范围从1531.70(195.725THz；通道1)～1564.07nm(191.675THz；通道55)，相邻通道间隔为0.6nm(75GHz)。全部通道采用以20Gb/s的NRZ信号由$LiNbO_3$ Mach-Zehnder调制器进行外调制，然后用1.48nm泵浦的高铝共掺杂的掺铒光纤放大器作为后置放大器。线路的中继器采用普通的前置放大器。这些EDFA都在非饱和区运行以获得较宽的波长范围。0.5dB的带宽是19nm，总发射功率为+13dBm，全部信号传输通过150km的单模光纤(平均色散：+15.2ps/nm/km，在1545nm的平均色散斜率+0.064ps/nm^2/km)中继距离为50km。在接受端，全部信号由普通的前置放大器进行放大，其多色色散由色散补偿光纤进行补偿，而不需对所有通道一个个分别地作色散调整。

美国的AT&T Reseach和Bell实验室在55km的非零色散光纤上有50个通道，每通道以20Gb/s进行光信号传输实验，这50个通道是由25个波长的偏振复用来实现的。总传输码率为1Tb/s。25个激光器的输出通过星形耦合器和波导光栅路由器进行复合，波长范围从1542nm(通道1)～1561.2nm(通道25)，通道间隔为100GHz，除了通道16应用DFB激光二极管外，其他所有通道光源都是外腔激光器。每个激光器的输出端接有偏振分束器对全部偏振状态单独控制。复合的波长经放大后通过偏振分束器对全部偏振光束进行排列，25对共偏振波长由3dB耦合器分开，分别由$LiNbO_3$ Mach-Zehnder调制器调制后再经过一偏振分束器合成正交偏振。调制器具有18GHz的小信号带宽和内装偏振器。20Gb/sNRZ的驱动信号由2个10Gb/s电子多路复用产生。50通道20Gb/s信号传输通过具有零色散波长为1513nm和色散斜率为0.07ps/nm^2/km的非零色散光纤。因为采用的放大器是专门为MONET项目开发的平坦增益放大器，不需采用预加强。接受的信号由6GHz低通电滤波器进行滤波。实验结果表明所有50通道的灵敏度为10^{-9}。当全部通道工作时传输质量的下降由两种因素产生，信噪比下降1dB左右和偏振串扰。没有发现相邻通道的串扰和四光子混频引起的品质下降问题。

集成光学三十年回顾及展望

1 前言[1,2]

集成光学(integrated optics)早期的基本构想首先是把光波约束在有限空间的光波导(optical waveguide)内传播;利用光波导可以构成包括光源、探测器、调制器、分束器、透镜和反射镜等一系列有源和无源光波导器件;采用类似集成电路的微加工技术,将这些光波导器件集成在同一基板上,从而构成具有特定功能的光路系统。这种集成光学系统与传统的光学体器件系统相比,具有体积小、质量轻、坚固紧凑、无需人工进行对准、适宜于平面工艺大批量生产和成本降低等优点。集成光学的出现是近代光学发展史上又一里程碑。

集成光学的名称是由 Miller 在 1969 首先提出(Bell System Tech. J., 1969, 48: 2059)。如果要考察光波导的最早的研究则应追溯到 1910 年,由 Hondros 和 Debye 关于电介质棒的研究工作。后来的研究主要针对微波工程应用的介质波导。直到激光问世,才真正开始了光波导的研究。在 60 年代中对光波导理论、光波导特性的测量、光波导耦合方法进行了比较全面和深入的研究。在 70 年代中开发出各种光波导器件,当时研究的光波导材料和器件比较集中于玻璃和晶体等电介质材料。由于这些材料不能构成激光器和探测器,因而开发的集成光学系统以混合集成为主,如射频频谱仪、模数转换器等。进入 80 年代,更多转向半导体光波导器件的研究,特别是对激光器集成化的研究,并将光电子集成(OEIC)和光子集成(PIC)提上日程。进入 90 年代,为了满足实际应用的要求,光集成系统在原来的二维平面波导结构的基础上进一步向空间和三维结构以及一维全光纤结构发展,并进一步演变为光-电-机综合集成系统。

2 材料及加工技术

微电子能取得如今辉煌成就,综合性能优异的硅材料有很大功劳。集成光学没有如此好运,至今尚未找到一种能适应集成光学要求的全能材料,也许这种理想化的单一材料并不存在。与集成电路相比,集成光学器件的种类要繁杂得多。根据是否需要驱动,集成光学器件可分为有源(active device)和无源(passive device)两大类。有源器件包括激光器、探测器、光放大器、光调制器、光开关等;无源器件有透镜、棱镜、反射镜、耦合器、隔离器和滤波器等,不同器件对材料性能的要求有很大区别。例如激光器要求材料具有高受激辐射的能级体系,光调制器要求材料具有高电光系数等。

目前集成光学器件通常应用的材料有半导体(主要是 GaAs 和 InP),晶体(主要是 $LiNbO_3$),石英,玻璃和硅基片(平面光回路),以及聚合物(主要是 PMMA)。集成光学在早期研究阶段,根据能否在同一基片上集成全部光器件把集成光路分为单片集成和混合集成两类。有代表性

的集成光学频谱分析器如图 1 所示，是典型的混合集成光路。平面波导透镜和产生表面声波的换能器是以 $LiNbO_3$ 晶体为基片，通过钛扩散或质子变换形成铌酸锂光波导，GaAs 激光二极管及 Si 探测器线阵在平面波导两端面直接耦合。

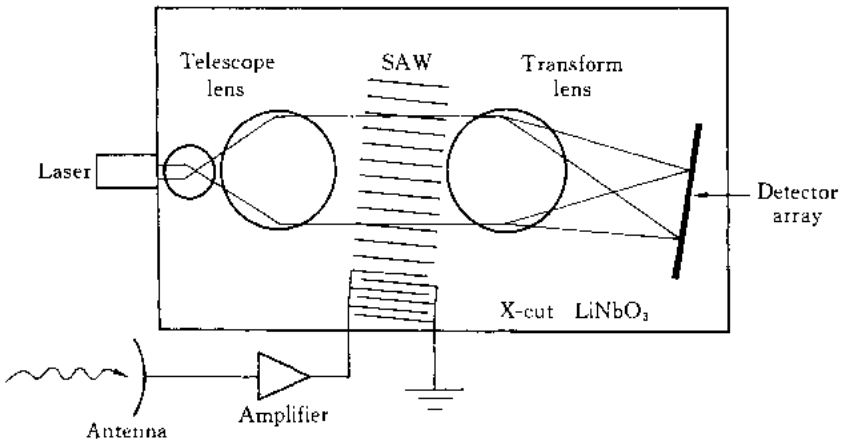


图 1　集成光学频谱分析仪

集成光学器件的制造工艺来源于集成电路的微加工技术。但由于材料和器件结构有很大差异，除了如掩膜板制作、光刻、真空镀膜、溅射沉积和干法刻蚀等常规工艺外，已开发了许多适合于集成光学器件的微加工工艺。例如为了便于集成，DFB 半导体激光器的反射镜面加工成光栅结构，代替一般激光二极管的解理端面，制造 $LiNbO_3$ 光波导器件可以采用质子交换技术等。

近年来迅速发展了一种硅基片上的平面光波回路(Planar Lightwave Circuit, PLC)[3]。这种光波导器件的优点是可以加工成精细结构的光波导，传输损耗低，与光纤匹配好，可以利用其稳定的热光效应制成调制器、开关等有源器件，也能作为混合集成的基片。图 2 给出了采用火焰淀积(Flame Hydrolysis Deposition, FHD)技术制作氧化硅波导的工艺流程。可以看出，这一工艺将光纤制造技术和集成电路的制造技术巧妙地结合在一起。利用这工艺制成的阵列波导光栅(Arrayed-Waveguide Grating, AWG)多路复用器，如图 3 所示。目前在实验室中已完成 64 通道的器件，相邻通道间隔为 0.4nm(50GHz)，AWG 的路径差为 63μm，中央通道的在片损耗3.1dB，通道间串扰 < −27dB.

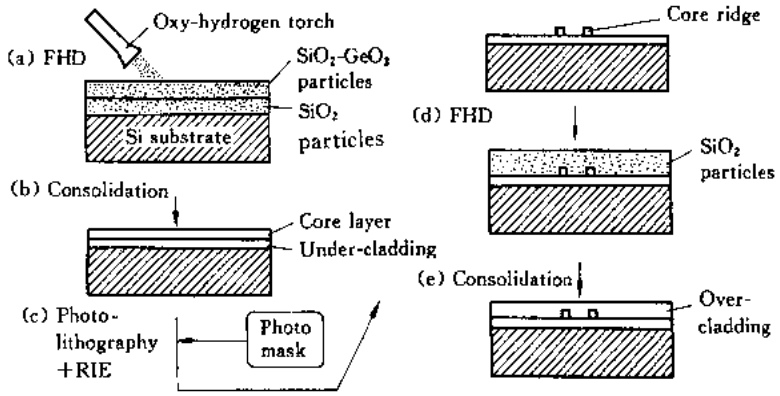


图 2　平面氧化硅波导制造工艺流程

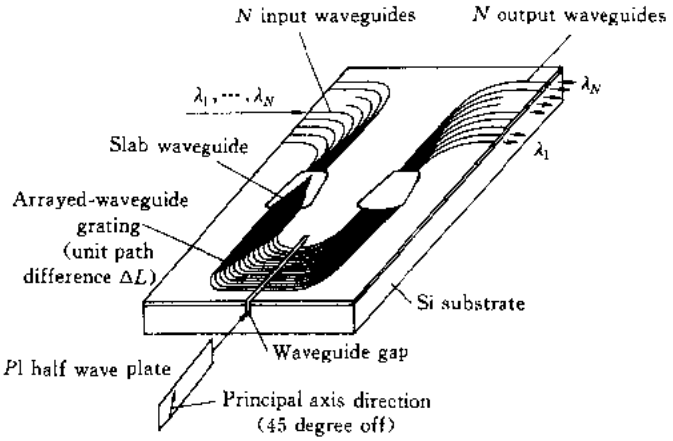


图 3　阵列波导光栅多路复用器

3　光电子集成(OEIC)和光子集成(PIC)

完整的集成光学系统是由多种有源和无源器件构成。有源光器件如激光器、放大器和调制器等需要有电流或电压的驱动、控制、调节、光电转换和放大才能工作。因此，理想的集成光学系统不仅是把光器件集成，还需要将光器件与电子器件集成在一起，这称为光电子集成(Optoelectronic Integrated Circuit，OEIC 或 Photoelectronic Integrated Circuit，PEIC)，如图 4 所示。OEIC 或 PEIC 不仅可以进一步缩小体积，结构紧凑和增加可靠性，而且能提高性能，例如可以缩短接线、减少寄生电容、提高响应频率、少受外界电磁干扰，特别当器件的集成度提高后，不

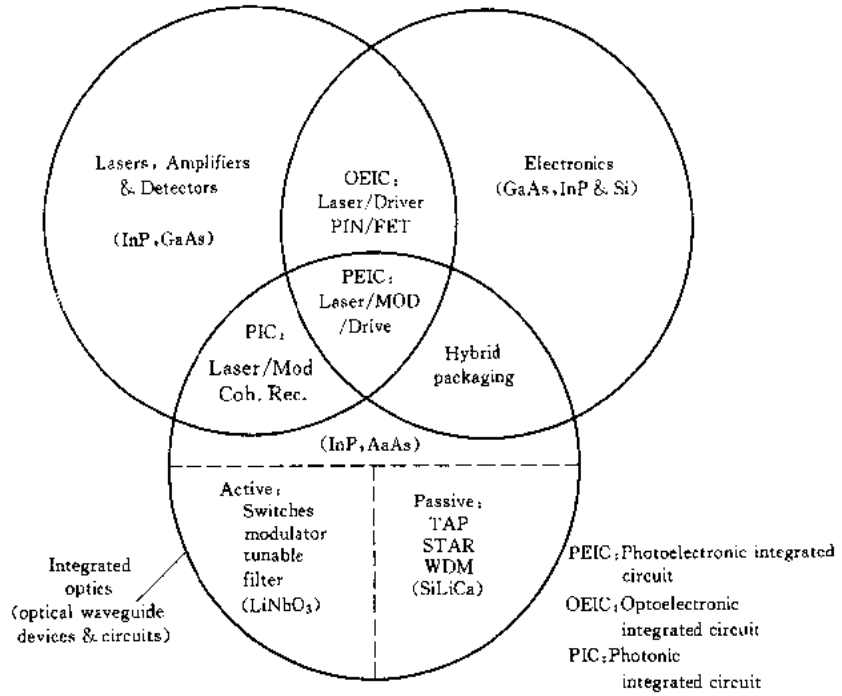


图 4　OEIC、PEIC 和 PIC 的构成框图

依靠光电子集成，器件的功能将无法实现。从应用的角度看，现阶段 OEIC 的发展很大程度上是由于高速光纤通信的推动，主要集中在光发射机和光接收机两方面。

1978 年美国加州理工学院制成了世界上第一个短波长 OEIC 发射机。早期的 OEIC 发射机由一个激光器和一个电子器件组成。现在已有包括发射、驱动、功率监控等功能较完备的单片集成芯片。例如美国 AT&T 贝尔实验室 1993 年报道了 InGaAsP/InP 系 BH-LD(掩埋异质结激光器)和 HBT(异质结双极型晶体管)的 OEIC 芯片。其中 BH-LD 用 3 次低压 MOCVD(金属有机化学气相淀积)生长，HBT 采用 MBE(分子束外延)和选择生长。所得光发射机的调制频率为 3GHz。OEIC 光发射机一般由 BH-LD + MESFET(场效应晶体管)或 HBT；QW-LD(量子阱激光管) + MESFET、HEMT(高电子迁移率晶体管)或 HBT，以及 MQB-DFB-LD(多量子阱分布反馈激光管) + HEMT 或 HBT 构成。图 5 表示近期发展的由垂直腔表面发射激光器(VCSEL)与 HBT 集成的 OEIC 光发射机。

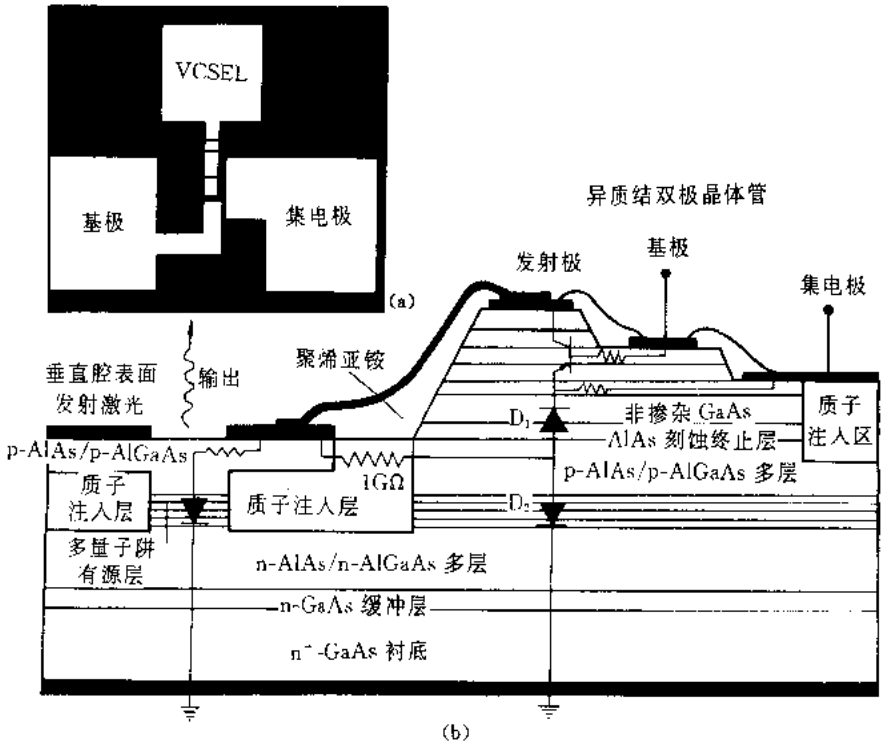


图 5　VCSEL 与 HBT 的 OEIC 器件

(a) 平面版图；(b) 截面结构

由于 OEIC 光发射机中激光器和电子器件在性能上、构造上差异大，对材料、工艺和器件结构的要求也各不相同，因此研制进展较慢，实用产品不多。OEIC 光接收机其结构比较简单，且在 GaAs 基片上制作 MESFET 的工艺相当成熟，光子器件与电子器件容易兼容，其发展速度超过 OEIC 光发射机。现已能制出有多级放大电路、实现多路复用和功能较完备的单片集成光接收机。OEIC 光接收机的研究重点是高速、低噪声和高灵敏度。采用量子阱超薄层结构，发展了二维电子气 HEMT 和高速 HBT。光电探测器除传统的 pin 结构外，近年又发展了 MSM(金属-半导体-金属)结构，其寄生电容小，本身是平面结构，易于与 MESFET 和 HEMT 兼容。例如在 1997 年美国光纤通信会议上最新报道的 OEIC 光接收机由 InP 基 pin 和 InAlAs/InGaAs

HBT集成的16通道光电接收器列阵，每通道的平均带宽达11GHz，高频下通道间最大的串扰低于 -40dB[4]。

光子集成(Photonic Integrated Circuits，PIC)是把同类的或/和不同功能的光器件包括有源光器件、无源光器件和光波导等在同一基片上集成为集成光学系统或子系统。这对高速大容量光纤通讯、光神经网、光计算和光信息处理等应用领域十分必要。例如在50路光频信号的WDM系统中要有50个发射回路，每个回路都包含有其不同波长的激光器，如何将50回路传输光束有效地耦合到一根光纤中，没有PIC技术是难以实现的。

PIC技术在现阶段首先设法将基本单元集成在一起，诸如激光器LD、探测器PD、调制器MD、光放大器OA、光栅GR和波导WG等。例如LD+MD的PIC近年来有大量研究报道，因为这是超高速光传输中所必须的。虽然量子阱激光器响应带宽可高达30GHz，但是在直接调制下运行，激光器的模噪声(chirp)会严重影响系统传输质量，因而必须采用外调制技术。目前有两种外调制技术可供选择：采用 $LiNbO_3$ 调制器技术比较成熟，带宽大，但不能实现单片集成；另一种选择是把LD+MD单片集成，其结构如图6所示。激光器与光放大器(LD+OA)的PIC对较大功率的模拟信号光传输十分必要，尤其是对高保真度的CATV。同时，在波分复用系统中对多路光信号进行实时放大时，LD+OA的集成器件也是必不可少的。光放大器与探测器(OA+PD)的集成可以提高探测器接收的光信号幅值，从而有利于克服接收机前置电子放大器的热噪声限制。此外还有：LD+WG，PD+WG，LD+GR，PD+GR以及将更多的光子器件集成在一起。例如光计算等应用领域所需的三维光集成系统中，二维激光器阵列、二维探测器列阵、二维空间光调制器列阵等都是由大量同类光器件构成的PIC。而且这一系列器件最终还需要电子器件辅助。因此，可以说PIC是实现完整的OEIC的重要组成部分。PIC的概念与集成光学早期提出的单片集成光学的目标是一致的，只是当时技术上一时尚难以实现，而且应用也没有提上议事日程。

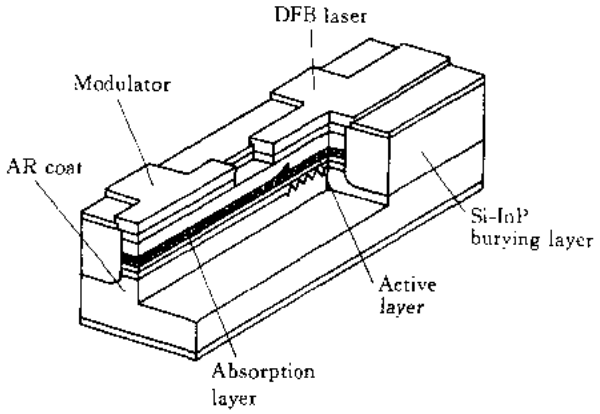


图6 激光器与电吸收光调制器的PIC

4 非线性集成光学

波导非线性光学效应与体光学器件相比具有两个明显的特点，即光强高和非线性互作用距离长。在体光学情况下，为了获得高光强通常需要把激光束聚焦。但是在体介质中，光束聚焦愈细，这种聚焦束被保持的距离愈短。因此非线性作用的效率不得不在提高光强与增加作用长度之间权衡。在波导中，光束在一维(平面波导)或二维(通道波导或光纤)受限制，其几何尺寸为光波长的量级，波导的长度主要由传输损耗决定。在集成光学中波导的长度为数毫米或数厘米，对光纤来说可用米或公里计数。由于波导或光纤中光束的截面小，因而即使在不大的光功率下也能获得很高的光强。将来也有可能应用非相干光在波导中产生非线性效应。因此可见，光波导是实现各种光学非线性效应的理想结构。从应用的角度来看，光学非线性效应

的实际应用首先将通过集成光学和光纤器件来实现。

包括平面波在内的大多数非线性光学的互作用可采用相位和幅值的渐变近似来分析。对光波导的处理方式可用耦合模理论。它给出信号光束复数幅值的增长率：

$$\frac{\mathrm{d}}{\mathrm{d}z}a^{(m,n)}(z) = \mathrm{i}\frac{k^2(\omega)}{2\varepsilon_0\beta^{(m,n)}}\frac{\int_{-\infty}^{\infty}P_{\mathrm{i}}^{\mathrm{NL}}(x,y)f_{\mathrm{i}}^{(m,n)}(x,y)\mathrm{d}x\mathrm{d}y}{\int_{-\infty}^{\infty}|f^{(m,n)}(x,y)|^2\mathrm{d}x\mathrm{d}y}\exp[\mathrm{i}(\beta^{(m,n)}-\beta_1)^2]$$

此式基本上表示取决于输出导波场轮廓的非线性极化在横坐标上投影的平均值，这一沿横坐标的平均值将引出交叠积分。

已有多种二阶光学非线性 $\chi^{(2)}$ 现象在平面波导形式中实现，这包括二次谐波发生、差频发生、光参量放大及光参振荡等，其中二次谐波发生(SHG)是研究得最为广泛和深入。特别是近几年的研究进展比过去二十年大大地加速了。一个重要的推动因素是可以使 GaAs 做的激光二极管的红外波段的光束经倍频获得紫外光束以适应数据存储和复印技术发展的需要，并可能成为最早商品化的光学非线性器件。

三阶光学非线性 $\chi^{(3)}$ 现象是一种全光互作用，即形成的非线性极化由三个光场的乘绩所决定。对二阶 $\chi^{(2)}$ 现象来说，其虚部一般是不能利用的，因为它表示损耗。与此相比，$\chi^{(3)}$ 的实部和虚部两者都能导致有重要意义的现象。例如，其虚部可表征受激 Raman 和 Brillouin 散射，而实部对应于折射率受光强的变化、参量混和、简并四波混频等。这些三阶现象最早都是从光纤开始研究的，其中简并四波混频、相干反 Stockes 散射以及各种与折射率随光强改变的现象已应用于对集成光学器件的研究。

大约在十年前，人们提出在普通的集成光学(包括光纤)器件中引进折射率随光强而变的波导介质，能实现全光模式的操作。这些器件可以用于全光信息处理，其速度仅限制于非线性和材料体系的开启和关断时间，具有亚皮秒级的响应。现有的集成光学器件，如耦合器、调制器和开关等，其工作原理是通过耦合场间的波矢匹配，或基于其位相受外界调制(典型情况如光电调制)的两个波导之间的干涉效应。每一种线性器件基本上都能通过在波导区采用非线性材料而转化成全光器件。迄今为止，已研究过的非线性集成光学器件有光栅和棱镜耦合器、光栅反射器、定向耦合器、M-Z 干涉器等。

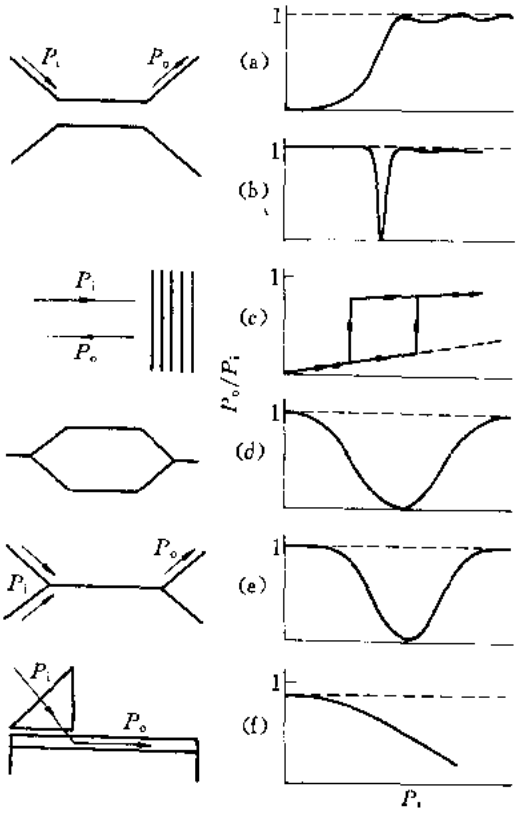


图 7　非线性和线性集成光学器件输出功率对输入功率的响应特性

(a) 半拍长定向耦合器；(b) 全拍长定向耦合器；
(c) 分布反馈光栅；(d) M-Z 干涉器；
(e) 模式分选器；(f) 棱镜耦合器

一组典型的全光波导器件，如定向耦合器、分布反馈光栅、M-Z 干涉器、模式分选器及棱镜耦合器对光功率变化的响应示于图 7。为对比起见，图中也列出了各种对应的线性器件的特性。开关特性最明显的是前两种器件，即非线性定向耦合器和非线性布拉格反射器。非线性定向耦合器的响应是两同向传播模之间非线性耦合这类器件所具有的典型特性，通常的结构是两弱耦合的平行通道；非线性分布反馈光栅的响应是相反光栅的响应是相反方向传播的波振幅相耦合的典型情况。

5 三维集成光学[6,7]

以上所讨论的各种集成光路，包括 OEIC 和 PIC 在内，其结构多数情况下是二维平面型的。一般来说，它只能处理零维(点)或一维(线)空间的光信号。但是客观上在大容量数据流、图像处理、机器视觉、神经网系统、光学平行逻辑运算和集成电路芯片的光互连等许多应用领域实际需要的是二维空间的光信息处理。

从物理本质讲，电子是费米子，光子是玻色子，光子不像电子那样带有电荷。电子之间通过电磁场相互作用，导致了电子信号很容易自身串扰或受外界干扰。光子之间很难相互作用，因此光信号可沿各自通道传播，不论其通道相互平行或交叉都不会相互干扰，这就造成了光学的固有平行性。对一普通的透镜来说，在 $1cm^2$ 上可以毫不困难地分解成 1000×1000 个像素，这充分说明了光子的高度平行性($>10^6$)。这种含有巨量平行光通道的空间三维信息的传播和处理恰恰是光子比电子最突出的优点之一。集成光学器件要适应这种应用领域，这就需要提出赋有新内涵的三维集成光学概念。

作为一个应用例子，图 8[8]表示光学平行数字处理系统中的三维光集成器件的功能要求。大体上说，设计和开发中的三维集成光学器件较多的是属于层状结构。每层中都含有各种功能的光子或电子器件，它们也可以分别集成在不同的层中。实验表明，在三维集成光路中没有必要限定所有的光器件都是波导型的。例如现在已经出现的二维折射型或二元光学微透镜阵列，就是一种有广泛应用前景的非光波导型集成光器件。在三维集成光路中，光束可以在层($X-Y$ 平面)内从一个光器件进入另一个光器件，也可以在层与层之间(Z 方向)传播。后面将提到的微光台(MOB)就是一种空间集成光学系统的例子。光束无论在层内或垂直于层传播

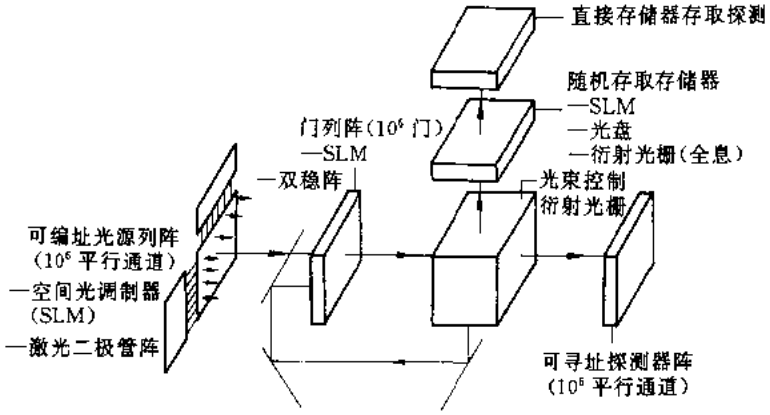


图 8 全光平行数字处理机的系统结构

时，既可以限制在波导内，也可以通过自由空间。光子可以在介质中，也可以在真空中传播，而且很容易通过真空和介质的界面。这一特性与电子严格被限制在导体中传导完全不同，三维集成光路的设计可以充分利用这一特点。

美国 OptiComp 公司开发的高速低能耗数字光计算实验样机是典型的三维混合光电集成系统[9]。DOC Ⅱ(第二代数字光计算机)的系统结构是一种光学波尔矢量-矩阵乘法器，如图 9 所示。此系统能实现 $\boldsymbol{f}=\boldsymbol{Ax}$ 运算，其中 $\boldsymbol{x}=(x_i)$，$i=1,2,3,\cdots,I$；$\boldsymbol{f}=(f_j)$，$j=1,2,3,\cdots,J$，分别表示输入和输出矢量；$\boldsymbol{A}=(a_{ij})$是控制逻辑矩阵，这是一个二维空间光调制器，由其构成单元控制通光或不通光。DOC Ⅱ运算的数字有 32 位长，采用双轨格式，由 64 个独立驱动的脉冲激光器作为输入光。一维探测器列阵表示出输出矢量 $\boldsymbol{f}$，每一个探测器对入射光具有阈值和倒转功能，阈值设置对应于 0 和 1 位的光功率，因此每个探测器是一个 NOR 门。

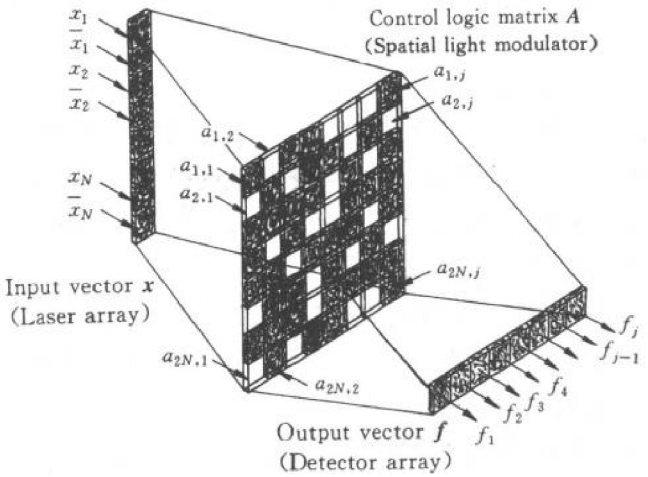


图 9　DOC Ⅱ采用的光学波尔矢量-矩阵乘法器系统结构

HPOC(高性能光计算机)是在 DOC Ⅱ基础上为了获得较大的 GIBP(互联门带宽积)和进一步小型化而设计的，采用二进制张量-矩阵乘法器系统结构和集成化器件，如图 10 所示。与 DOC Ⅱ相似，HPOC 能实现 $\boldsymbol{F}=\boldsymbol{AX}$ 运算。其中 $\boldsymbol{X}=(x_{ij})$是输入矩阵，由 $N\times N$ 二维 VCSEL(垂直腔表面发射激光器)列阵作为数据输入；$\boldsymbol{A}$ 是控制张量，其硬件是 DOIE(衍射光互联元件)列阵，提供互联图形的编码。$\boldsymbol{F}=(f_{kl})$是输出矩阵，是二维探测器矩阵。与 DOC Ⅱ不同，这里每个探测器不仅实现阈值和倒转操作(NOR 功能)，还将电信号转换成光信号作为输出，称 DANE(探测、放大、负值和发射)单元。HPOC 集成系统的截面图如图 11 所示，其

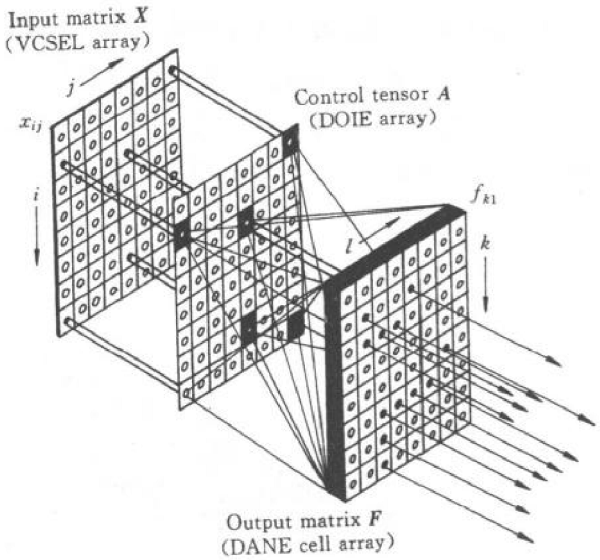


图 10　HPOC 采用的光学波尔矩阵-张量乘法器的系统结构

总体积仅 4cm^3，而 DOC Ⅱ的面积为 1.0cm^2。

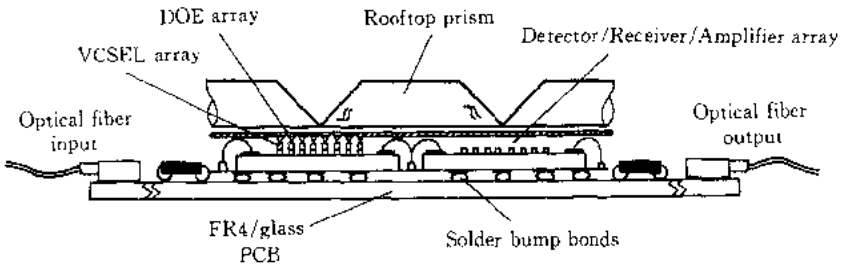


图 11　HPOC 组件的截面结构图

6　光纤(一维)集成光学

在光纤传输中需要大量光连接器、光耦合器、光分路器等无源光器件。虽然这些器件都可采用平面光波导来实现，但由于其耦合技术和插入损耗等原因，实际上大量应用的是光纤型无源器件。这是由于其光场匹配、易于耦合、插入损耗小，而且制造工艺简便、可靠性好、成本低。

参铒光纤放大器的发明是光纤技术自低损耗石英光纤问世后又一个里程碑，也是光纤器件从无源器件发展到有源器件的重大突破。光纤放大器的工作原理简单地说是利用掺入光纤的活性离子在泵光作用下的粒子数反转从而对入射光信号提供光增益。虽然早在 1964 年就有人提出掺 Nd^{3+} 光纤放大器的设想，但直到 1985 年低损耗掺杂石英光纤研制成后，光纤放大器才成为现实。许多稀土离子都被用作掺杂剂进行实验，研究得最多的是掺 Nd^{3+}、Pr^{3+}（用于 1.3μm 波长）和掺 Er^{3+}（用于 1.55μm 波长）光纤放大器，其中尤以掺 Er^{3+} 光纤放大器（EDFA）最为成熟，而且在 1990 年实现商品化。采用 980nm 或 1480nm 两波长的泵浦光，对 EDFA 的泵浦效率可达 11dB/mW，因此几毫瓦的泵浦功率，就可以获得 30～40dB 的光增益。

光纤放大器对光纤通信的巨大贡献是系统中可以不再使用传统的光-电-光转换的中继器，这种无光电转换的光纤传输可以在受系统性色散和放大器自发辐射积累效应限制前达到最大的距离。由于光纤放大器的出现，使人们对相干光通信系统研究的热情大大地减退了。

对光纤器件发展有重大影响的另一项技术是 1989 年前后光折变光纤光栅的发明。当掺锗光纤中通过激光时，光纤的折率将随光强的空间分布发生相应的变化，这称为光折变效应。折射率变化的大小与光强成非线性关系。如用激光干涉条纹（全息照相）或通过位相掩膜辐照掺锗光纤，就会形成光纤光栅。利用光纤光栅不仅可以构成光纤激光器，包括光纤 DFB 激光器，还可做成光纤滤波器、模式转换器和波分复用器等光纤光栅型的无源光栅器件。

除利用石英光纤和掺杂石英光纤开发了一系列有源和无源光纤器外，其他材料和用于波谱更宽的光纤和光纤器件有含氟光纤、电光晶体光纤，有机材料和聚合物光纤以及非线性光纤器件等都取得了不同程度的进展。

基于以上实验成果，人们自然会想到将不同特性的光纤器件集成为具有特定功能的光纤集成系统或子系统，有人称之为光纤集成光学。与传统的集成光学（二维）和三维集成光学相比，属于一维集成光学。这种系统的特点是易于与光纤连接或耦合，甚至可以将光波传输与其他光信号处理功能合为一体，提高响应速度和效率，并能缩小体积，增加可靠性和降低费用。在光纤通信、光互连、光神经网及光纤传感器等许多领域有广泛应用。

图 12(a)表示在 1997 年光纤通讯会议(OFC'97)报道的五个波长 DFB 光纤激光器集成系统[10]。每个激光器由 5cm 长的掺铒石英光纤做成,掺铒浓度为 $1.5\times10^{25}m^{-3}$,光纤芯径为 4μm,数值孔径为 0.27。制造布拉格光栅是由 KrF 准分子激光器为光源,波长为 248nm。五个不同波长的光纤激光器刻好光栅后熔接在一起,以 1480nm 半导体激光器为泵源,功率为 60mW。这样集成光纤激光器输出波形如图 12(b)所示,相邻波长的间隔为(1±0.1)nm。这初步实验结果所显示出的激光功率输出的差别主要原因是泵浦功率较小以及光纤激光器之间熔接不够理想。这种方案的集成光纤激光器不难做成八波长光源,有可能应用于 WDM 光纤通信系统。

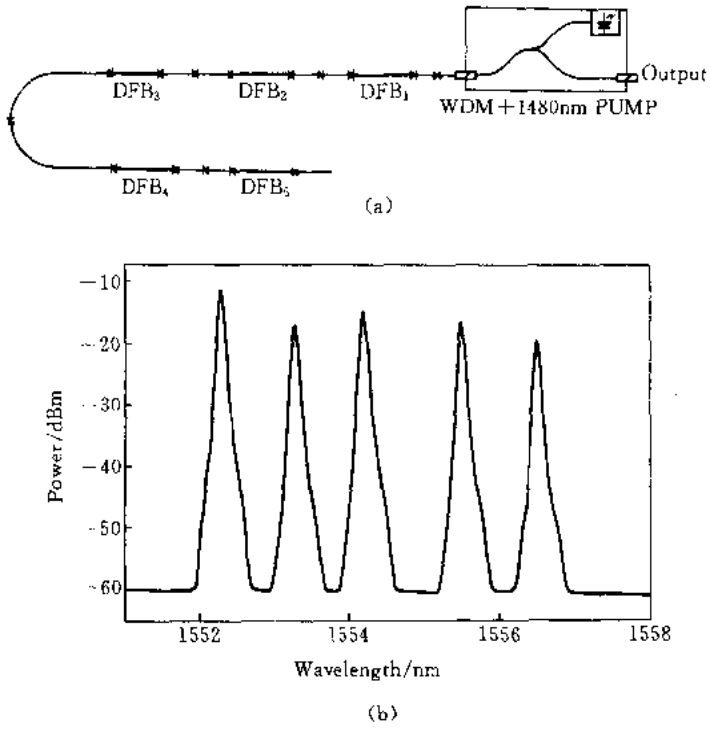


图 12 五波长集成 DFB 光纤激光器
(a) 实验系统; (b) 输出波形特性曲线

7 微光-电-机集成系统(MOEMS)[11~13]

早在 60 年代,几乎正当微加工技术用于发展集成光学、微光学和光电子器件的同时,也已有人设想微加工技术用来制造微机械结构。由于硅不仅是理想的半导体材料,而且也具有良好的力学和机械性能,因而采用制造半导体集成电路的微加工技术,可以用硅材料制成微型机械部件和装置,例如微传感器和微执行器等。而且这种微机械装置与微电子器件尺寸和材料相兼容并集成在共同的硅基片上,现在这已被大家称为微电(子)机(械)系统(Micro-Electro-Mechanics Systems,MEMS)。

微光-电-机系统(Micro-Opto-Electro-Mechanics Systems,MOEMS)就是把光子器件(集成光学和微光学)、微电子器件和微机械结构采用相兼容的基板材料及微加工技术集成在一起成为完整的体系(见图 13)。这种 MOEMS 能充分地体现这类器件相结合的综合性能,不仅能使系统

结构进一步小型化，而且可导致新一代器件和装置的诞生，如三维集成器件等。

如前所述，作为信息载体的光子与电子相比，不仅由于其速度快，信息容量大，更由于其固有平行性，这是电子所无法比拟的。但是现有的光子和光电子集成器件与集成电路相似的平面结构，难于实现光子所固有的空间平行性的特点。因此，目前光波导器件、光电子器件、平面微光器件列阵和全息器件、集成电路芯片和芯片间的光互联，以及光纤和光纤器件与其他光器件的对准和耦合基本上还要靠人工装配和调整。采用 MOEMS 就可能由精密的微机械结构来实现和保证，并且可以免除人工装配操作，大大减少时，提高质量和成品率，降低成本。

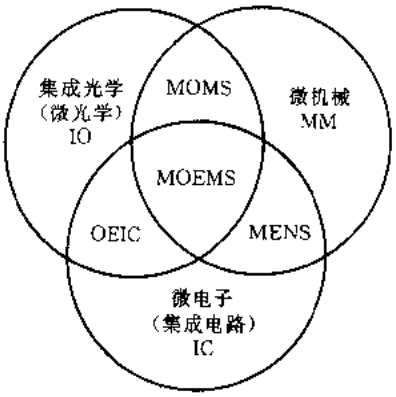


图 13　微光-电-机系统(MEOMS)

MEOMS 器件和系统是集成光学发展的又一新阶段，有着极为广泛的应用领域。开发较早的数字微镜器件(Digital Micromirror Devices，DMD)应用于大屏幕投影显示，现已成为商品投放市场。DMD 由 2048 x 1152 可动微反射镜列阵构成，具有很高分辨率，可满足高清晰度电视的要求。DMD 中每一个可动微镜分别由置于其下面的 COMS 电路驱动，从而对入射到其表面上的光束进行调制。从镜面反射的光束通过投影透镜在大屏幕上产生图像，目前屏幕的对角线可达 4.88m。

最近有人提出了一项称为自由空间微光台(Free-Space Micro-Optical Bench，FS-MOB)的三维空间集成光学系统。它是利用 MOEM 技术将光学元件、定位器、执行器和其他机械结构元件集成在单一硅基板上。由于光束是平行于 FS-MOB 而传播，因此可以把许多垂直于基板而立的三维微光学元件，包括 Fresnel 微透镜、折射微透镜、反射镜、分光镜、光栅及滤波器等采用 MOEM 技术加工并实现空间集成，如图 14 所示[14]。FS-MOB 系统可以大大地减少体积、质量和制作成本，它可以应用在光互连、光开关、光扫描、印刷、显示及数据记录等系统上。

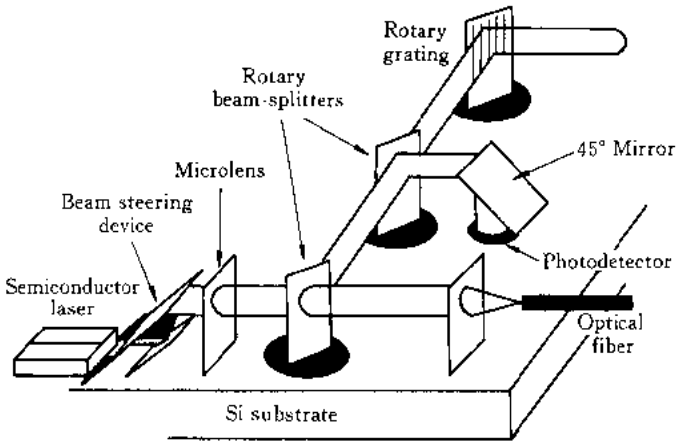


图 14　自由空间微光台(FS-MOB)示意图

又如过去曾经是十分昂贵和复杂庞大的自适应光学系统，采用 MOEM 技术就可以将传感器、处理器和执行器集成在同一基片上。当通过有湍流的大气或媒质观察物体时，由于快速涨落，使波前畸变，造成图像模糊不清。自适应光学的功能就是设法补偿湍动的影响，使模糊的

图像清晰。这种光学系统在许多领域有重要应用，例如在军事上、光通信、生物医学和激光焊接等。自适应光学系统的基本工作原理是采用波前传感器测出畸变，由处理器计算出校正量，最后由执行器根据计算结果进行补偿。图 15 表示这系统工作原理的框图。采用常规的技术构成的自适应光学系统非常复杂，可靠性差，成本很高，阻碍了其广泛应用。最近已提出可将微传感器、微处理器和微执行器（变形微镜列阵）整个系统集成在同一芯片上，如图 16 所示[15]。在此系统中，干涉型波前传感器用来探测波前畸变；并行处理的模拟电路用于波前重构；并行操作的放大器是为了驱动变形反射微镜列阵。模拟折射和衍射微光学元件采用新设计的模拟掩膜板一步法制作。执行器和驱动电路采用标准的 CMOS 硅工艺加工。执行器的响应频率实验结果为 15kHz，最大可用行程为 4μm。整个自适应光学系统的带宽超过 10kHz。

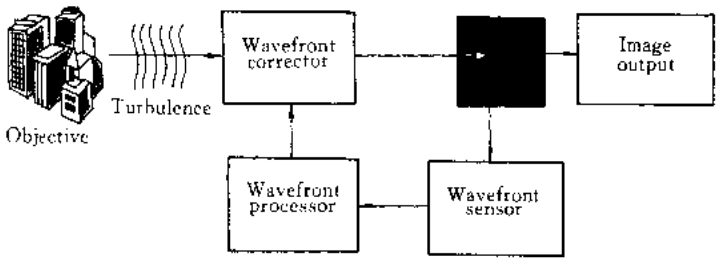


图 15　自适应光学系统框图

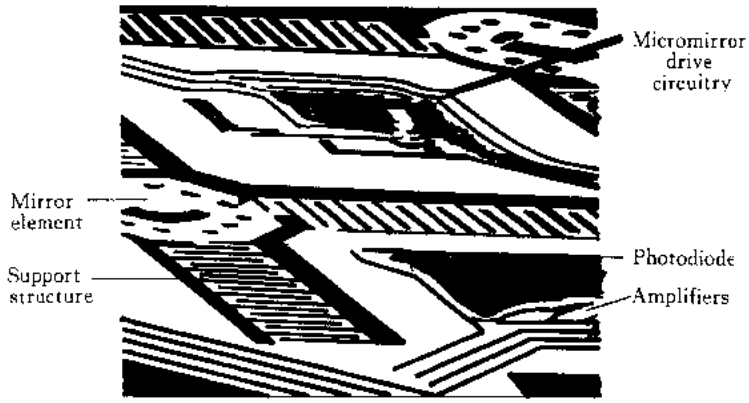


图 16　变形微镜自适应光学芯片扫描电镜照片

8　结语

集成光学从提出到今天已将近有三十年。随着光纤通信、光盘和光存储、光显示、光信息处理和光计算等光波技术应用的迅速发展，集成光学器件和系统得到了广泛的研究、开发和部分应用。现在，集成光学的概念已有了很大的扩展，它的理论和技术也得到了极大的发展。人们对集成光学的认识更为深刻和全面，不再有人期望有朝一日集成光学可完全代替集成电路，也不再有人消极地认为集成光学好比海市蜃楼，可见不可及。的确，集成光学器件的实用化和商品化要比人们原先预期的来得慢，因为人们往往将集成光学与集成电路的发展作简单的类比。

从半导体晶体管的发明到超大规模集成电路技术的成熟，大致经历了三十年历程，成为20世纪中影响最为深广的一项科学技术。光子作为信息的载体与电子相比在物理学上有许多重要特性，例如，光子不易发生交互作用，因此光束可以交叉通过而互不影响，光束传播具有固有的并行性；光子在真空中的速度为光速，不像电子在导线中受 *RC* 时间常数限制等。目前电子器件及其系统的响应时间最快达到 10^{-9}s，即 ns 量级，几乎到了其固有的物理极限。而光子的响应时间可达到 10^{-15}s，即 fs 量级，目前已能获得十几个 fs 的光脉冲。因此，光子学从本质上讲能克服和解决电子学遇到的种种困境和障碍。光子学要做电子学做不到的事，并创更加光耀夺目的技术世纪。但这将是比现有电子技术难度更高的技术，无疑就需要更长的孕育和准备时间。另一方面，正如本文前面已指出的，集成光学器件也离不开电子器件的辅助，有时甚至需要精密的微机械结构作为自装配或执行器。这也就更增加了集成光学的技术难度。

有人将光子比作是现代技术皇冠上的夜明珠。光子在光纤通信和 CD 光盘领域已显示出的优势，只能看作是光子技术初露锋芒。光子技术的潜力目前尚难于估量。集成电路是建造宏伟电子技术大厦的砖，同样，光子技术的成功决离不开集成光学器件的发展。集成光学无论在材料，器件，系统结构和加工技术各不断提出大量的新课题，集成光学在未来的21世纪必将有更大的发展。

参 考 文 献

[1] T. Tamir. Integrated optics. second edition. Springer-Verlag, 1979

[2] 陈益新.集成光学-理论和技术.上海：上海交通大学出版社，1985

[3] M. Kawachi. Integrated silica waveguide technologies. Technical Digest of OFC'96, ThT1, San Jose, 1996

[4] K. C. Syao, et al. 16-channel monolithically integrated InP-based Pin/HBT photoreceiver array with 11GHz channel bandwidth and low cross talk. Technical Digest of OFC'97, TuD5, Dallas, Texas, 1997

[5] Y. X. Chen. Nonlinear integrated optics (Invited). SPIE Proceedings, 1994, 2364: 174 ~ 187

[6] 陈益新.三维集成光学新概念(特邀报告)，吉林大学自然科学学报，1990，(特刊)：1

[7] Y. X. Chen. A new concept of 3-dimensional integrated optics. Optoelectronics Devices and Technologies, 1990, 5 (1): 109

[8] 陈益新.光计算.上海：上海交通大学出版社，1988

[9] P. S. Guilfoyle, D. S. Mccallum. High-speed low-energy digital optical processors. Optical Engineering, 1996, 35 (2): 436 ~ 442

[10] P. Varming, et al. Five-wavelength DFB fiber laser source. Technical Digest of OFC'97, TuH4, Dallas, Texas, 1997

[11] 陈益新.微光电机械系统(MOEMS)的技术和应用.光子学报，1997，26(21)：45 ~ 54

[12] Y. X. Chen, Thin film technologies for micro-opto-electro-mechanical system applications (Invited). SPIE Proceedings, to be Published in 1997

[13] Y. X. Chen, Technologies and applications for micro-opto-electro-mech-anical (MOEM) systems (Invited). Technical Digest of 2nd Optoelectronics and Communications Conference (OECC'97), 10D1-2, Seoul, Korea, 1997

[14] M. C. Wu. An overview of micromachining for optical communications. Technical Digest of OFC '97, WB4 Invited, Dallas, Texas, 1997

[15] R. L. Clark, et al. Micro-opto-electro-mechnical (MOEM) adaptive optic system. SPIE Proceedings, 1997, 3008: 12 ~ 24

二维空间光调制器的研究和应用

0 引言

空间光调制器(Spatial Light Modulator, SLM)的作用是改变空间光分布的相位、偏振、幅值或光强,甚至也可以是波长。这种分布函数可能是由某种电信息所产生,也可能是另一种分布的光信息所决定。前者称为电寻址空间光调制器(E-SLM);后者则为光寻址空间光调制器(O-SLM)。这种将信息转换成一维或二维光数据场的器件,在光信息处理和光计算系统中能够充分利用光的快速、并行处理和高度互连的能力。图 1 给出了设想中的全光处理机的一般框图,在此系统中,SLM 可以完成多种功能。因而这种器件,特别是二维空间光调制器(2-D SLM),在近 20 年中引起了极大的重视和广泛的研究。

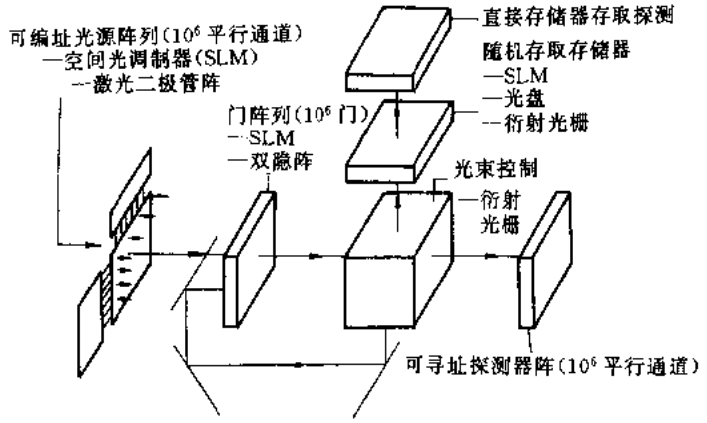


图 1 全光多处理器系统

SLM 早期应用目的是企图作为电视或计算机的平面显示、印刷中的编辑和排版及二维光数据存储等。激光的出现,不仅使以上的应用可能变成现实,并使它在相干光信号处理领域的应用得到了迅速的发展,这包括在光信息处理、光数据处理和光计算中的应用。应用 1-D SLM 可完成频谱分析、卷积、相关、适配滤波等处理。2-D SLM 的应用则更为广泛,例如:增强匹配滤波器、非相关的图形识别、白光彩色图像处理偏微分和积分方程的求解等。特别是近年来光学处理被重视,大量的模拟和数字光处理和光计算机系统结构被提出,其中大多数都能利用高性能的 SLM 来实现。当前研究中的一重要趋向是充分利用光独有的三维几何能力在二维的处理平面之间提供能任意排列的、高密度的并能重构的互连。利用 SLM 在系统处理器、印刷电路板,甚至集成电路芯片之间实现这种高度平行光互连,能发展成一种高性能的光电子混合处理器。在最新提出的许多高度平行的光计算系统的结构方案中可利用 SLM 构成:细胞式自动机、通用并行有限态计算机、关联神经网处理器等。这些新概念有希望解决多传感器处理

器、图像识别、语音识别、高级符号处理及各种人工智能等复杂问题。

1 结构型式和工作原理

各种文献报道的 2-D SLM 已有数十种，择主要的列于表 1。

1.1 光寻址 2-D SLM

光寻址 2-D SLM 有以下几种主要结构型式：

(1) 夹层(三明治)结构是 2-D SLM 最早采用的一种基本型式，如图 2 所示。工作原理是写入光作用于光导电层引起的偏压加于电光调制层，控制读出光的相位。幅值(光强)或振幅。如图所示的反射型结构，其中央层是反射镜同时起着光阻挡作用，液晶光阀(LCLV)就属于此类。

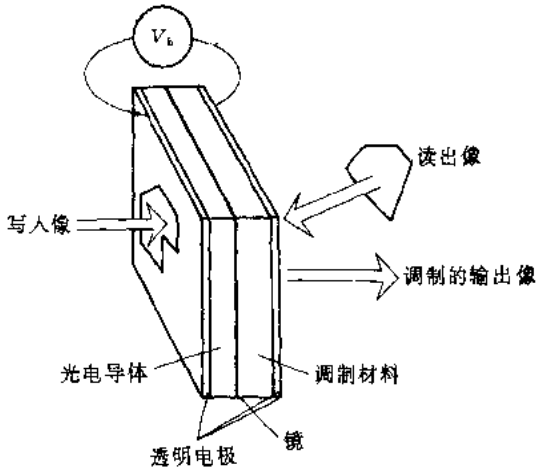


图 2 夹层结构 SLM

(2) 如果电光调制材料同时具有光电导或光吸收，这时可以省去单独的光敏层，如只读光调制器(PROM)，简单地由一层光折变材料，如 BSU 或 BGO 晶体的两个表面上加上透明电极构成。

(3) 改进的夹层结构，如变形膜调制器示于图 3。光敏层产生的偏压加到能受电场作用而变形的膜片上，使反射的读出光束的相位发生变化。这类中有代表性的是变形镜器件(DMD)，见图 4。这是在有光电二极管阵列的 Si 衬底上沉积一层较厚的多晶硅，再经腐蚀构成悬臂结构，在电场作用下产生变形。

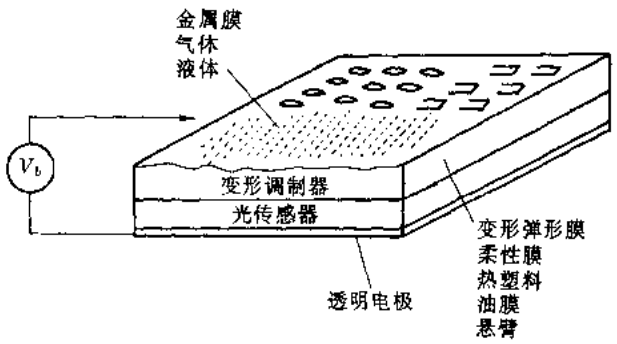


图 3 变形膜 SLM

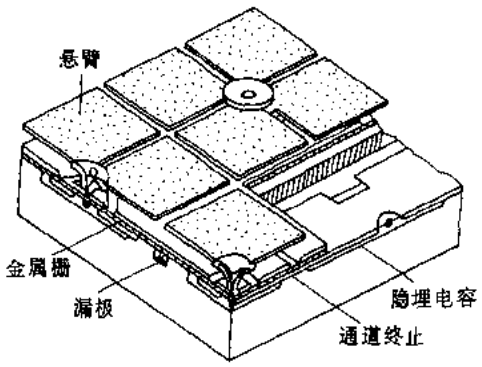


图 4 DMD 原理结构

(4) 具有光发射和微通道板的调制器，其中光阴极将写入图像转变为图像，再经微通道板(MCP)放大。MSLM 调制器就是一例，其结构见图 5。这种器件可使输入图像以正电荷和负电荷分布存储在调制器材料上。

(5) 一种不同于传统的夹层结构的调制器，是将光敏区和调制器相并列，最早由 UCSD 提出的 PLZT 调制器，其每单元的一半面积上淀积 Si 做成光电探测器和放大器电路，另一半为光调制器，其原理和结构如图 6 所示。

(6) 近年来光双稳和光非线性器件构成的 SLM 引起了普遍的重视，其结构通常采用光标

表 1

序号	名称/型号	调制材料	寻址方式		分辨率 1p/mm (No. Pixels)	灵敏度/ ($\mu J \cdot cm^{-2}$)	响应时间			研究单位
			光传感	电传感			写/ms	读/ms	存储	
1(O)	LCLV	向列液晶	CdS	—	30	6	10	15	15ms	Hughes
2(E)	Titus	KD_2PO_4	—	电子束	20	—	30	5	h	Sodern
3(O)	TP	热塑料	PVK 薄膜	—	200 ~ 1600	5	10	100	年	NRC 等
4(O)	PROM	BSO 或 BGO	BSO 或 BCO	—	6	5	< 0.1	< 0.1	< 2h	Sumitomo
5(O)	MSLM	$LiNbO_3$, KDP	光阴极和 MCP	—	10	3 × 10 量子极限	10	20	天或月 1 ~ 30d	Optron 等
6(E)	Talaris	油膜	—	电子束	1023 × 1023	—	—	—		GE
7(O)	Librascope	Smectic 液晶	液晶 (热吸收)	—	40	10	0.005	0.001	30d	Singer
8(E)	LIGHT - MOD/SIGHT - MOD	YIG(磁光)	—	矩阵电极	128 × 128	—	10	10	年	Litton, Semetex
9(E)	TIR	$LiNbO_3$	—	Si 电路	5000 × 1	—	< 0.001	< 0.001	< 1μs	Xerox
10(O)	Phototitus	KD_2PO_4	Si 二极管	—	10	2	2	< 0.5	5s	Lockheed
11(O)	PLZT	PLZT	Si 光晶体管	—	10	10	< 0.01	< 0.01	s	UCSD
12(O)	RUTICON	变形弹性膜	非晶 Se	—	40 ~ 120	30	5	4	15min	Xerox
13(O)	DMD	变形膜	Si 光晶体管	—	128 × 128	2	0.025	0.04	200ms	TI
14(O)	PEMLM	变形膜	光阴极和 MCP	—	40	3 × 10 量子极限	0.01	0.01	天或月 1 ~ 30d	NRL
15(E)	VO_2	VO_2	—	电子束	20	—	10	10	年	USC
16(E)	Optical Tunnel Array(OTA)	悬片	—	矩阵电极	16 × 16	—	1	1	15ms	VARAD

注：序号 1 ~ 9 已商品化；10 ~ 16 有可能 5 ~ 10 年实现；(O)光寻址；(E)电寻址

准具型式，例如 Fabry-Perot 腔或多层干涉滤波器等。

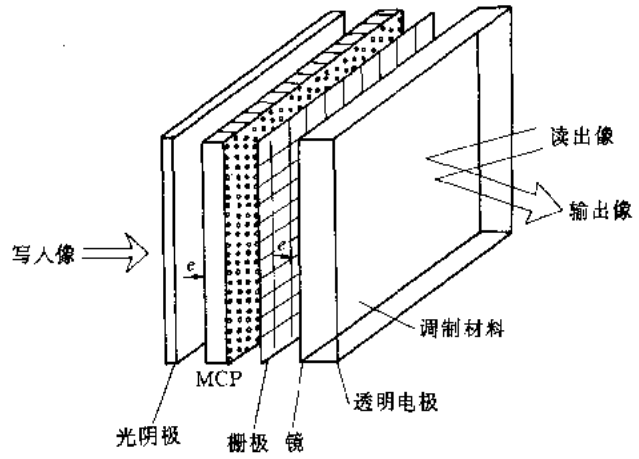
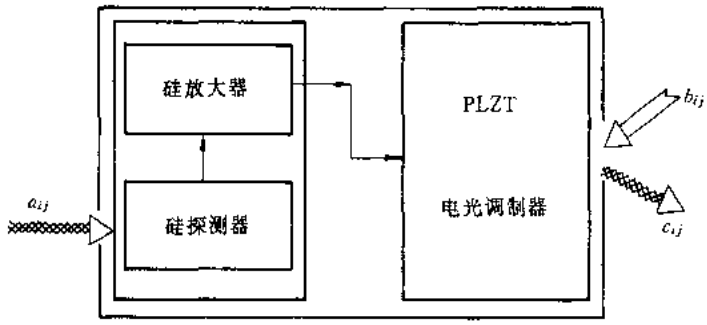
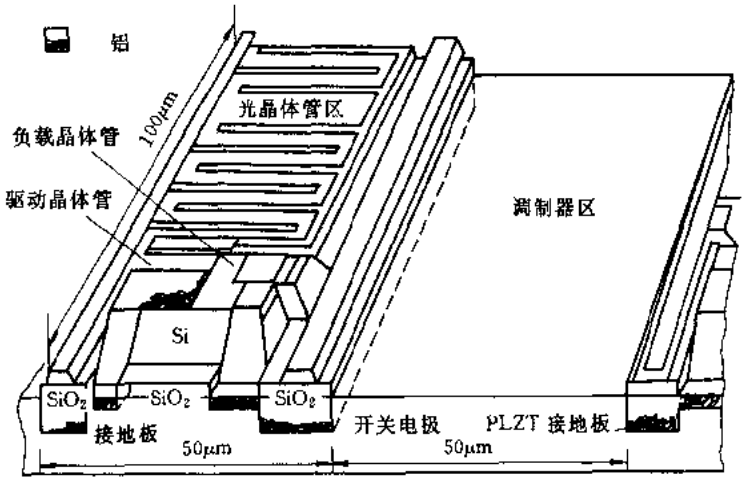


图 5　MSLM 原理结构



(a)原理



(b)结构

图 6　PLZT SLM 原理和结构图

1.2　电寻址 2-D SLM

用于光寻址的 2-D SLM 的结构型式和光调制器材料，大都能同时作为电寻址的 SLM，所不同的是它不需光敏感器件，而由电子器件代替。电寻址的方式主要有如下几种：

(1) 电子束寻址 SLM。光调制材料置于真空管内由扫描电子束写入与阴极射线管的萤光屏很相似。由于没有光传感器要求将读出光与之隔离,因而读出可以做成反射式,也可采用透射式。表 1 中序号 2(E)Titus 就是这一类。另外,MSLMP EMLM,RUTICON 等都可做成电寻址的器件。

(2) 矩阵电极寻址 SLM。这种器件的寻址通过两组垂直交叉的线阵电极来实现,它们可以分别置于调制材料的两面,也可以放在同一面上。例如在 SIGHT MOD LIGHT MOD 器件中的磁光材料 YIG 上有二维阵列的小台面($76\mu m \times 76\mu m$)。每个台面是一单磁畴,可以取两种相反的方向,两种方向使读出光束的偏振有不同的 Faraday 旋转,从而产生两种二进制的振幅调制电平。

(3) 半导体寻址 SLM。简单的矩阵电极寻址器件一般受到调制机制的限制,由于它缺乏阈值或非线性的驱动特性,不能对从两组交叉电极中任一电极的驱动信号加以区分。解决的途径是在每个交叉点上加进有源器件,例如薄膜晶体管。这类半导体寻址器件结构复杂,制造困难而且昂贵,但由于能获得高性能 SLM,所以最近在研究开发中的器件多数属于这种类型,在 LCLV、PLZT、DMD 及片式 SLM 中都可采用。

(4) 其他的电寻址方式。还有利用多通的声光 Bragg 衍射器件等。PLZT 调制器是在调制的每单元中。

2 功能和应用

SLM 的功能和应用有如下几方面:

(1) 图像乘法运算。从图 2 不难看出,光寻址 2-D SLM 的输出信号像应是写入像和读出像的乘积,这种器件的基本功能是二维信息图形的实时平行乘法器。采用适当的光路,这种操作可用于矢量和矩阵的模拟数值乘法、匹配滤波器、波前共轭、二维场之间的可变互连以及可编程的模板运算。

(2) 幅值放大。这种三端构造的 SLM 相当于二维光晶体管列阵。当采用光强均匀的读出光束,就可使光强弱的写入像增强,也可用来将放大的输出光束与产生弱写入光束的光学系统去耦合或隔离。当均匀的读出光束具有与写入光束所不同的性质时,这种器件具有非常有用的转换功能,例如从非相干光转换成相干光,或从一种波长到另一种波长。

(3) 电寻址的 SLM。作为光学和电子处理子系统之间的接口器件具有各种光电转换功能。例如,在混合光电子处理器中 E-SLM 可作为输入元件或数据格式化元件。这种器件在电子处理系统之间实现可变光互连很有用。它具有与 O-SLM 相似的乘法功能,这时输出像将是读出像与有效反射率的乘积,而有效反射率是写入的电子信息的函数。

(4) 信息存储。SLM 的存储功能可使输出像成为以前存储在内的写入像与现在的读出像的乘积。因而 SLM 可用作为存储器件。具有存储功能的 O-SLM 一般可用替代全息片,并能多次应用。对 E-SLM 来说,存储功能的重要性还在于可以避免二维数据流的不断更新,这要求驱动的电子处理器具有很大的信息带宽。对 O-SLM 来说,存储功能具有通过延长时间来增强探测光强很弱的输入像的能力。

(5) 非线性操作。这是另一个十分重要的功能,SLM 能在每一点上对输入的写入光强产生非线性操作。事实上,大多数调制器其非线性是固有的特性,例如有许多调制器采用交叉偏

振器或干涉器读出，由此产生的读出光束的幅值是写入光强的正弦函数关系。只有在调制特性的 sin $2n$ 点附近，才能得到线性调制。如果将调制点偏置在 $\sin(2n-1)$ 点附近，很容易得到对比倒置的调制。MSLM 及 PEMLM 等 SLM 就能进行这样的操作。另一种操作是对图像作阈值处理，这在级联系统中可用来重新产生二进制光信号，或在光处理器中作决定等。其他的非线性操作包括实现波耳逻辑运算；光像场的模-数转换等。

3 性能及改进

有许多特性参数可以作为 SLM 质量的估价，并用以作相互的比较。其中最重要的有：空间分辨率，一般用每毫米内能分辨多少线对（lp/mm）表示，也可用能分辨点的总数表示；其次是器件的响应时间是很重要的参数；对 O-SLM 来说，光传感器的灵敏度也十分重要；读出信号的动态范围，这包括位相、幅值，或偏振，将对输出像的对比度、信噪比、输入-输出的非线性及空间响应的均匀性等有直接影响；其他如光学质量、加工的困难程度、器件寿命、可靠性等都需考虑。一些主要 SLM 的分辨率，灵敏度和响应时间列于表 1 中以资比较。

应该指出，SLM 目前能达到的性能参数离光信号处理、光计算或其他应用所提出的要求，还有较大的差距，如表 2 所示。但可以相信，这些要求的参数不是不可能达到的。

表 2 SLM 的性能和要求

性能参数	现有水平	实用要求	性能参数	现有水平	实用要求
可分辨单元数	100×100	1000×1000	动态范围（灰级）	5	100
帧速	$10Hz$	$>10Hz$	平整度	几个波长	1/5 波长
灵敏度（O-SLM）	$<50\mu J/cm$	量子极限	空间均匀性	10%	1%
存储时间	1s	>1h			

4 发展趋向

要获得高性能 SLM，主要是高分辨率和高响应速度，这要通过采用新的材料和设计新的器件结构来实现，当然也离不开采用新的微加工技术或应用新的调制机理。不同调制材料的响应速度为：电光晶体 10^{-12}s；CaAs（多量子阱或标准具）10^{-12}s；PLZT（单晶）10^{-9}s；变形膜 0.5×10^{-6}s；悬片 10^{-6}s；磁光开关 10^{-6}s；铁电液晶 10^{-6}s；声光（Bragg 调制）10^{-6}s；PLZT（多晶）10^{-5}s；变形弹性膜 10^{-3}s；向列型液晶 10^{-2}s。

近年来，对用于 SLM 的新材料，除了表 1 中所列的以外，还特别注意了开发多量子阱材料、铁电液晶以及有机晶体和聚合物电光材料。

研究新的调制机理有光致变色（photochromic）、光二向色（photodichroir）、电子致色（cathodochromic）、电泳、电毛细管（electrocapillary）等效应。

新的器件结构包括对夹层结构的各种变种型式、二维阵列激光二极管、光集成调制器的叠

合等。

在加工方面尽可能采用超大规模集成电路已开发成熟的微加工技术，包括图形的发生和转移，薄膜淀积和生长，以及微细图形和结构的刻蚀等。

参考文献

[1] A. D. Fisher, J. N. Lee. The current status of two-dimensional spatial light modulator technology. SPIE, 1986, 634: 352 ~ 371

[2] John A. Neff. Major initiatives for optical computing. Optical Engineering, 1987, 26(1): 2 ~ 9

[3] Spatial light modulators and applications II. SPIE Proceedings, 1987, 825

二维 Si/PLZT 混合集成空间光调制器

1 引言

二维空间光调制器(2D-SLM)是一种能改变二维空间分布光束的幅值、相位、偏振以及波长的一种列阵器件。这种分布函数可能是由某种电信息所产生,也可能是某种分布的光信息所决定,前者称为电寻址空间光调制器,后者称为光寻址空间光调制器。这种将信息转换成二维光数据场的列阵器件,在光通信、光信息处理和光计算等系统中能够充分利用光的快速、并行处理和高度互连的能力,在光交换、光互连、光逻辑运算、光存储和光输入输出等方面有非常广泛的应用。因而,近年来 2D-SLM 的研制在国际上获得了越来越多的重视。目前,国际上已提出的 2D-SLM 结构方案已有数十种,但能商品化的产品为数不多,大部分还正在研究开发中。我国研究和开发的空间光调制器只有液晶光阀(LCLV)和只读光调制器(PROM),其性能远远不能满足上述广泛应用,特别是光计算的要求。

美国 UCSD 大学于 1986 年首次提出了 Si/PLZT SLM 的结构方案,该结构以 Si 作为光敏材料实现光寻址功能,以 PLZT($Pb_{1-x}La_x(Zr_yTi_z)_{1-\frac{x}{4}}O_3$)透明铁电陶瓷作为光调制材料实现空间光束的调制功能。Si 是性能极为优越且制备工艺最为成熟的光敏材料,PLZT 是迄今能投入使用的无机物中电光系数最大的材料。因而,Si/PLZT 结构充分利用了两者独特的优点,在材料选择上有较大的优势,我们认为,这种结构的 SLM 有可能较早达到光计算和光信息处理实际应用的要求。

2 Si/PLZT SLM 的基本工作原理

二维 Si/PLZT SLM 由完全相同的功能单元二维排列构成,每个单元均由一个 Si Darlington 电路和一个 PLZT 电光调制器构成,单元工作原理如图 1 所示。工作时,二维写入光像场照射到 SLM 上,每个光像素被对应的 Si 光晶体管接收并进行光电转换,因而产生了与写入光强成正比的二维电流场,每个单元的硅放大器对该单元的硅光晶体管输出电流进行放大,输到该单元 PLZT 调制器电极间的负载电阻上,使得调制电极上产生一个压降,由于电光效应,该压降在 PLZT 材料内产生的电场改变了材料的折射率,这样就可对读出光像素进行调制得到输出光像素。如果读出光的偏振方向与电极平行,则输出光是相位调制的,如果读出光的偏振方向与电极方向成 45°,则输出偏振调制光,加检偏器即可成

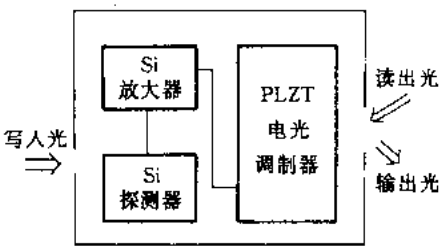


图 1 Si/PLZT SLM 单元工作原理图

为强度调制光。

3 混合集成 Si/PLZT SLM 器件结构设计及制作

以 Si 电路作为光敏元件和驱动部分，以 PLZT 作为电光调制材料，理论上应采用如图 2 所示的结构形式。其中的 Si 电路既可以体硅器件，也可以是蓝宝石上的薄膜硅器件，后者便于集成。PLZT 既可用体块材料，也可以用薄膜材料，薄膜材料在响应时间和空间分辨率等方面的性能优于块材，我们原拟采用的技术路线是：在外延于蓝宝石的单晶硅衬底的部分面积上制作光电晶体管及驱动电路，再在其余面积上用射频磁控溅射法制作 PLZT 薄膜，并做上电极，成为如图 3 所示的结构。但由于蓝宝石上外延 PLZT 难度极高，且垂直通光要求薄膜很厚（~10μm），制作很困难，因此采用了体硅电路与块材 PLZT 电光调制器混合集成的方案，即先分别在硅片上制作光晶体管及放大电路列阵，在 PLZT 材料上制作电光调制器列阵，然后将硅器件列阵划开粘接到 PLZT 调制器上的对应部分，构成混合集成 Si/PLZT SLM。

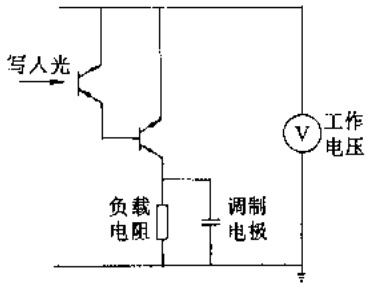


图 2 Si/PLZT 等效电路图

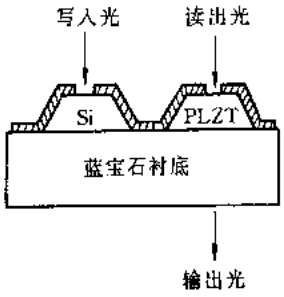


图 3 一种 Si/PLZT SLM 单片集成方案

硅器件列阵每个单元结构如图 4 所示，由一个硅光晶体管和一个普通晶体管构成光学 Darlington 器件，该器件具有较大的增益，可将写入光转换成足以驱动 PLZT 电光调制器的电信号。

PLZT 电光调制器列阵是在具有大电光系数的透明铁电陶瓷 PLZT 上制作叉指电极列阵构成的，可对输入光进行偏振调制和强度调制。电极版图如图 5 所示。

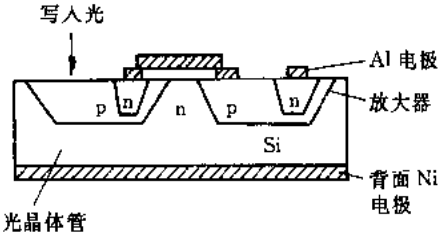


图 4 Si Darlington 电路结构图

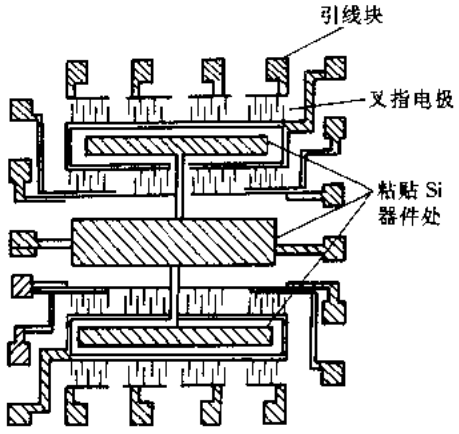


图 5 PLZT 调制器列阵电极版图

采用精密切割工艺将硅器件列阵划开，对准粘接到 PLZT 电光调制器每个单元的对应位置上，然后将两者的对应电极引线连接，即可构成 Si/PLZT 混合集成 SLM 器件。目前，由于混合集成的工艺条件的限制，我们是在 4×4 单元的 PLZT 电光调制器列阵上粘接 16 个 Si 探测器，构成 4×4 像素的 Si/PLZT 混合集成光寻址空间光调制器，单元尺寸为 600μm×600μm。

在器件制作中，需对材料及器件结构参数等进行不断的优化，包括 Si 材料的电阻率；PLZT 材料的掺 La 量；光敏区及调制区的面积；放大电路的负载；叉指对数和电极间距等。Si 的电阻率直接影响探测器的暗电流和击穿电压，前者决定器件的消光特性，后者决定器件所能达到的工作电压，从而影响器件的动态范围，我们选用了 2～8Ω·cm 的 Si 材料。PLZT 的电光特性取决于 La 的含量，为了得到尽量高的电光系数和尽量小的电滞回线，选用的 PLZT 材料成分为 9/65/35，其电光系数为 $(0.2\sim1.0)\times10^{-16}m^2/V^2$。每个单元的调制区域面积越大，器件插入损耗越小，而光敏面积越大，灵敏度越高，需统筹设计。我们选择的尺寸是，每个单元调制区面积为 250μm×500μm，光敏面积为 $(100\times100)\mu m^2$。放大电路的负载电阻影响 Si 电路的输出电压幅度和器件的响应时间，在线性范围内，电阻越大，输出电压越高，但响应时间越慢。我们选择的负载电阻为 10 kΩ。理论上刻槽电极和双面电极性能优于平面电极，由于工艺条件限制，我们采用平面电极，但仍需对其电极形状进行优化设计，包括叉指对数和电极间距，前者影响调制电容，从而影响器件的响应时间，后者影响调制均匀性和半波电压值。间距越大，均匀性越好，但半波电压越高。我们采用的叉指对数是 7，电极间距为 20μm。

4 器件性能及测试结果

首先分别对 Si 电路和 PLZT 电光调制器的性能进行测试。Si 晶体管的击穿电压 $V_{ceo}=90V$，放大倍数 $\beta=60\sim80$，电路响应时间 15μs（10kΩ 负载），响应带宽 $DC=50kHz$。

PLZT 电光调制器性能测试采用如图 6 所示的装置（虚线框部分不要）。测试结果为：半波电压 90V，响应时间 12～15μs，调制带宽 $DC=20kHz$。

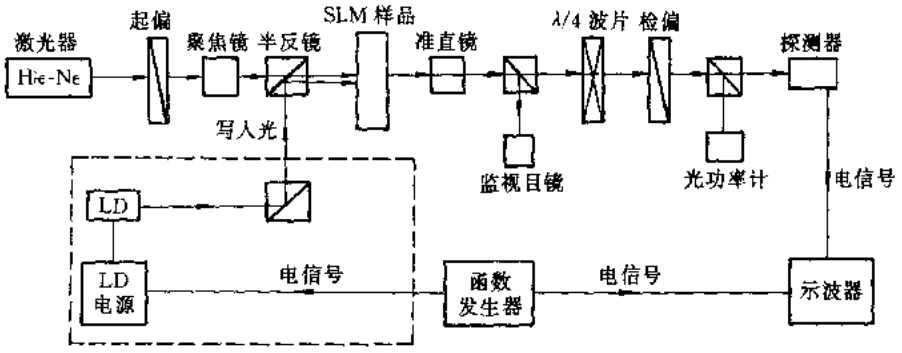


图 6 SLM 性能测试光路图

反映空间光调制器的性能指标除空间分辨率（与单元数有关）外，主要有：(1)帧速。反映器件对读入图像的幅处理速度；(2)动态范围。反映输出光信号的幅值变化范围；(3)灵敏度。使得输出光信号达到预定动态范围所需的写入光能量。上述三个参数均可用图 6 所示的装置进行测试。

(1) 帧速。输出的交变光信号大小随写入光信号频率的增大而减小，下面是测试数据（以

下测试均是在 50V 的工作电压下进行)：

	写入光信号频率/kHz									
	1	2	3	4	5	6	7	8	9	10
输出电信号相对值/mV	4.5	4.5	4.5	4.1	3.9	3.6	3.3	3.2	3.0	2.8

以电信号幅度下降 3dB 时的写入光信号频率作为帧速，可得帧速为 8kHz。图 7 是在 5kHz 频率下，SLM 对不同写入光信号的响应波形图。

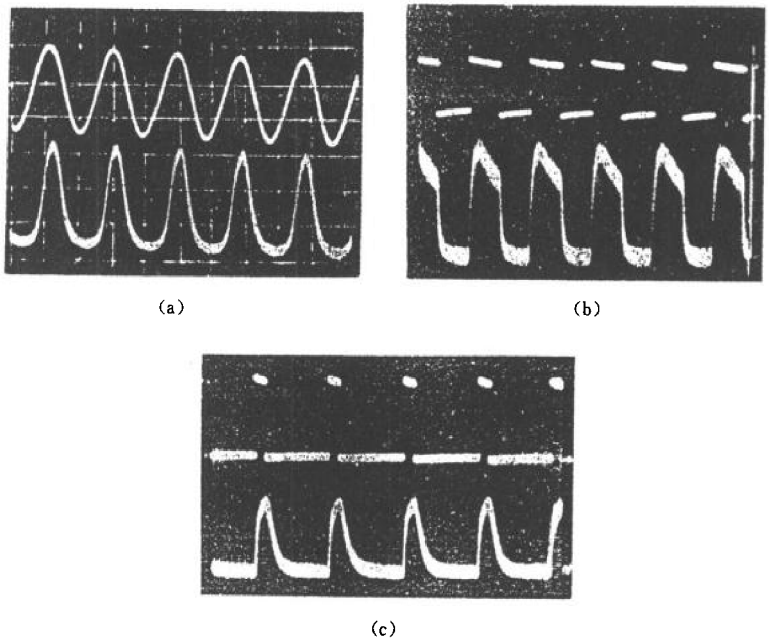


图 7　SLM 对不同写入光信号的响应波形

(2) 动态范围。以一恒定的读出光打到 SLM 上，在不同的写入光强下，将得到不同的输出光强，下面是测试数据(读出光强为 0.5mW)：

	写入光强/mW									
	0	0.057	0.244	0.483	0.702	0.935	1.190	1.45	1.67	1.96
输出光强/nW	15	160	165	200	240	262	293	321	360	375

该器件的最大动态范围为：

$$\gamma = I_{max}/I_{min} = 375\text{nW}/15\text{nW} = 25:1$$

(3) 灵敏度。对于 SLM，达到一定的动态范围所需的写入光能量就是灵敏度。即

$$\text{灵敏度} = \frac{\text{响应时间} \times \text{写入光功率}}{\text{写入光光斑面积}}$$

测量在方波写入光信号下输出光信号的方波前沿得到器件的响应时间为 20μs，在动态范围测量中，得知达到预定动态范围(这里设定是 10:1)所需的写入光功率是 0.244mW，写入光斑面积是 4.6mm^2，则灵敏度 = 20 × 0.244/4.6 = 0.106μJ/cm^2。由上述测量可知，研制的 SLM 样品性能指标达到：动态范围 25:1，帧速 8kHz，灵敏度 0.106μJ/cm^2。

目前，有关 Si/PLZT SLM 的二维特性的研究工作，大单元数 Si/PLZT SLM 以及单片集成 Si/PLZT SLM 等研制工作正在进行中。

参 考 文 献

[1] A. D. Fisher, J. N. Lee. SPIE, 1986, 634: 352
[2] Sadic. C. Esener. Opt. Eng., 1986, 25(2): 250
[3] S. H. Lee. Opt. Eng., 1987, 26(5): 406

未来计算系统中的光互连

光束由于其扩展的传输容量及巨量的平行处理能力已被作为富有极大潜力的新的信息载体。在未来的计算系统中,光学不仅为此提供新的器件技术,同样也将开拓新的算法和系统结构,最终目的是要通过巨量的平行处理和分布处理实现与人类意识活动相仿的柔性信息处理功能。在未来的计算系统中,光束的应用在以下三方面已引起极大的重视,即光互连(optical interconnection)、光神经系统(optical neural systems)以及光数字系统(optical digital systems)。光神经系统是通过光束连接各神经元实现实时学习以及图像和其他分布数据的并联处理;光数字系统的功能是应用逻辑计算原理通过光束实现高计算精度的巨量平行处理;发展光互连技术的目标是要克服电子系统目前所面临的"连线极限",同时光互连也是实现光神经系统及光数字系统的关键技术。

光互连把以超大规模集成电路 VLSI 为代表的先进电子技术与光通信技术相结合以消除电子系统中信息传输中遇到的问题,诸如传输时延、线路之间的串音、连线与安装上的空间限制以及消耗功率大等。光互连技术旨在突破上述极限以提供一种高速度、大容量和柔性的信息传输。

为了实现上述光互连的目标,需要发展相应的光互连器件和光互连网络。

光互连器件包括光电子有源和无源器件,将这些器件组合以实现高速度、大容量和可重构的互连网络,并在时间、空间和波长领域采用高密度的多路复用技术。

极高速光互连器件其目标是利用亚皮秒高速器件采用时分多路复用(TDM)以实现大容量互连网络。

空间平行或功能性光互连器件;其目的是利用空分多路复用(SDM)技术并与其各种电子-光子功能相结合以实现能重构的高流通量的光互连。

波长平行或功能性光互连器件,是利用波分多路复用(WDM)技术以达到大容量光互连的目标。用 WDM 技术实现高容量信号母线。

无源光互连元件包括激光和衍射元件,其目标是开发具有稳定和高密度光互连两者优点的光元件。

在上述器件以后,作为下一代的紧凑小型化的器件将是由不同材料体系和功能所组合成的先进的光电子集成器件和回路。

光互连网络的系统结构和设计的目标是要获得坚固、高速和多通道的光互连系统,具有能重构和自定路线的柔性功能。

芯片间和芯片内的光互连目的是开发在处理元件之间、处理器与存储器之间,或存储器与存储器之间的柔性光互连网络。

进一步的研究是光电子器件和无源光器件的组件化,实现光互连器件的集成化和小型化,解决需要高精度对准问题。

光互连技术

1 光互连的意义

巨型计算机(supercomputer)是现代科学技术,特别是国防尖端技术和高技术所迫切需要的,如核武器设计、空间技术、气体动力学、长期天气预报、石油勘探、粒子束模拟计算、实时图像识别和人工智能等。随着集成电路技术的发展,依靠提高主频来提高系统功能,难度越来越大,主频 1ns 被公认是一个工程极限,目前已达到 2～4ns,要再提高就十分困难[1],目前比较一致的看法是,巨型机的发展趋势是大规模并行机(massive parallel machine)[2]。这不仅在技术上成为可能,而且在经济上也是可行的。

目前,大规模并行机从连接性质来分有两种[3]:一是处理器和处理器连接。每个处理器都与少量其他处理器连接,其网络结构可以是网格状或超立方体等。这种类型也可是所谓分布内存机,每个处理器都有一个大容量的本机内存,如 Connection Machine、InteliPSC/860、NCUBE2 和 Touchstone 都是一些商品化的产品。另一种是处理器和内存连接,每一个处理器都可以通过多级互连网络访问大容量共享内存的任意部分,处理器之间的通信是通过对共享内存的读和写进行的,这也就是所谓共享内存机,Alliant Fx/280、BBNTC 2000、Conourrent 3280E、Encore 91、FPS System 500 及 Seguent Symmetry 是一些已商品化的产品。无论哪一种,这其中的关键是需要高效、快速、大容量的互连网络,以实现处理器和处理器或处理器和内存之间中间结果的信息交换或者说通信。互连网络是巨型机的核心,决定着性能价格比。

另外,在智能计算机方面,能像人脑一样进行学习和判断的神经计算机,它的核心并不是逻辑运算,而是互连网络的构成。就电子器件的开关速度而言,比人脑神经细胞的反应速度至少快 1 万倍[4]。作为生物开关的神经细胞,其响应时间一般为毫秒级或更大,根本的区别在于人脑神经系统具有巨量并行处理单元和互连。据现有研究表明,一个神经网络大约有一千亿个神经细胞,每个神经细胞有高达一万个实触(synapse)与周围相连接。这充分说明人脑的智能不是由神经细胞的反应速度,而是由高度互连的巨量神经细胞的并行处理能力所决定的。近年来,人工神经网络已成为世界范围的研究热点,这不仅为智能计算机的发展开辟了新的途径,对其他学科也将带来巨大影响。

在传统的电互连网络中,由于任何连接导线都不可避免地存在着一定的电阻 R 和电容 C,因而存在着 RC 延时,而且随着器件工作频率的提高,RC 延时往往超过了晶体管的开关时间[4],这就造成了互连带宽的提高受限于连接导线,并且电信号的传输速率只有光速的千分之几,产生时钟扭歪现象(clock skew)。另一方面,由于 VLSI 尺寸的缩小,还会造成连接导线拥挤,没有足够的空间来排列,彼此之间的耦合、干扰也增大,当线度减小到所预计的极限量级时(0.1μm),这个问题将变得更严重。总之,伴随着集成电路技术中的器件尺寸减小、芯

片尺寸增大、频率的提高，以及系统结构方面的大规模并行技术的采用而来的是数据流量的巨增，而电互连由于在带宽、互连密度、时钟扭歪、能耗、抗干扰性等方面的限制，无法解决这一矛盾。

因而，用光互连来解决这个计算机中通信的"瓶颈"在80年代初被提了出来[5]。光互连是以光子取代电子作为信息载体来实现功能单元之间的信息交换。光互连的分类从层次来看，可分为：芯片内之间的互连；芯片之间的互连；电路板之间的互连；计算机之间的互连。从互连所采用的信道来看，可分为：光纤互连；波导互连；自由空间互连。

2 实现光互连的物理依据

光学信息通道的信息流量大。从物理本性上看[6]，光子不具有静质量，既可以在真空中传播，也可以在介质中传播，并且很容易通过真空和介质的界面，无论是在光波导中还是自由空间中，光信号都以该介质中的光速传播而与接受信号的元件数目无关，传输信息的速度高。光频与微波波段相比高10^4，所以它的频带很宽，即使考虑电-光、光-电转换及信道的限制，目前技术也可提供几个GHz的带宽[7]。而且光在传播过程中能量损耗很小(对计算机大小而言，损耗可忽略)[8]。

光学固有的并行性。光子不像电子那样带有电荷，电子之间通过电磁场而相互作用，导致电子信号很容易相互干扰或受外界影响，当频率很高时更为严重[9]。光子之间则很难相互作用，因此光波导可以相互穿越，只要交叉角大于10°左右[10]就不会有明显的交叉耦合；自由空间光束可以在三维空间互相穿越而没有明显的互作用。光互连不受平面或准平面的限制，而且不仅可以芯片边进行I/O连接，还可以在芯片内进行I/O连接。限制光互连密度主要有两个因素[10]：对自由空间而言是可分辨的光点尺寸；对波导而言是所要求的波导尺寸。芯径为5μm的单模光纤目前已很普遍，做到尺寸为1μm的光波导是很可能的。自由空间的传输主要受成像系统衍射的光点尺寸限制。无论是导波或非导波系统，为了避免串话，必须把光点分开几个波长的距离。即使有这种限制，从理论上讲也可制作出$(4\sim5)\times10^4/mm^2$的光波导或自由空间通道，或$5000/mm^2$的光纤通道。由于光学的并行性，系统的复杂度与系统的大小关系不大，即系统扩展时，复杂度增加不大。

光互连的扇出数主要受到可以用于探测器的功率限制。考虑电-光和光-电转换效率、传输长度、频带等因素，可知光互连的扇出数比电互连的大许多倍[7,10,11]。对采用并行处理技术的巨型机，即使有限的广播能力，即一点到多点的连接能力，也就是扇出能力，也可以显著地改变许多方面应用的性能，再考虑到能耗，即广播是以较低的数据传输速率进行的，性能也会有很大改进[3]。

光学易实现可重构的互连。因为光互连并不一定需要实际的"硬"连接(这一点对移动物体的连接也是很有利的)，原则上，可以把互连图形信息写入到可重构的光互连部件，从而实现动态互连。而动态电互连的实现则比较困难，存在许多难以克服的困难。

用光子作为信息载体在传输方面的优越性已在光纤通信上充分显示，现在，光纤系统不仅非常成功地用于通信干线和地区网络中，而且正在发展将光纤直接通到每个用户的技术；另外，在光存储和激光打印技术方面光子的作用也已显示出来。可以预见，随着光电子和集成光学器件(如光源列阵、空间光调制器和探测器列阵等)的发展、光互连结构理论和技术的深入研

究，在90年代，光互连的总体性能将会超过电互连，特别是全息光互连和使用空间光调制器的光互连结构将会取得重大突破，达到电互连无法达到的性能指标，充分显示光学在互连上的优势。

3 光互连方案及相关器件

目前，光互连方案从信道来分主要有：光纤互连；波导互连；自由空间互连三种。光纤互连适用于电路板之间或计算机之间这个层次上的连接，可以利用光纤通信的成功经验，比较容易实现，已经进行了好几种互连方案的实验工作[12~14]。与电互连相比，其优点是长度-带宽积高、扇出量大、系统功耗低等，采用分立的光源和探测器。波导互连可以提供高密度互连通道，适用于芯片内或芯片之间这个层次上的互连，采用集成光源和探测器，由集成光路来连接，这种互连目前还不很成熟[15~17]。自由空间光互连适用于芯片之间或电路板之间这个层次上的连接，可以使互连密度接近光的衍射极限，不存在信道对带宽的限制，易于实现重构互连。主要有使用全息光学元件(HOE)、空间光调制器(SLM)、透镜和反射镜的几种，是目前的一个研究热点。下面例举一些使用全息光学元件和空间光调制器的互连方案。

3.1 法国 ONERA-CERT 的 MILOID 方案[19~21]

采用 N^2-并行矩阵-矢量内积结构，如图1所示。光源采用1-D LD 列阵，波长 0.33μm，尾纤为 100~140μm 阶跃光纤；SLM 采用 Hughes IR LCLV，中心波长在 800~850nm，电子束导址

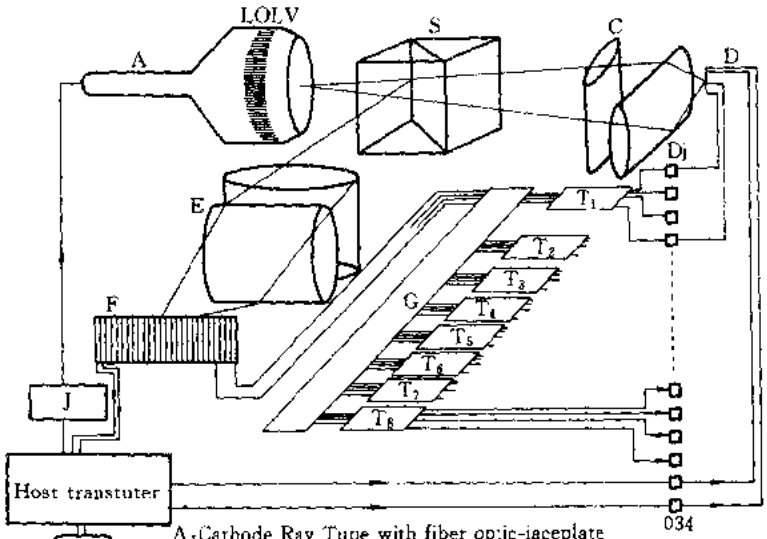


图1

(CRT),探测器用 UDT 光电二极管列阵。该工作从 1986 年开始,到 1990 年完成了 transputer 的两个单向通信连接和串话实验,估计连接数目小于 100 ~ 150,重构时间约 200ms,数据传输速率小于 100MHz。

3.2 美国 NUSC 的 AO 光子开关[22]

采用 N^2-并行内积 AO 偏转结构,如图 2 所示。开关复杂度为 $O(N)$。可以买到的最大 1-D 输入列阵数为 128,1-D 输出列阵数受 AO 器件的时间-带宽积决定,目前可达 100 多。声光器件用氧化铅玻璃做成,中心频率 70MHz,带宽 40MHz,632nm 的衍射效率的典型值为 80%,插入损耗 2 ~ 6dB,最坏情况的信号-串音比大于 30dB,重构时间 1ms 量级。

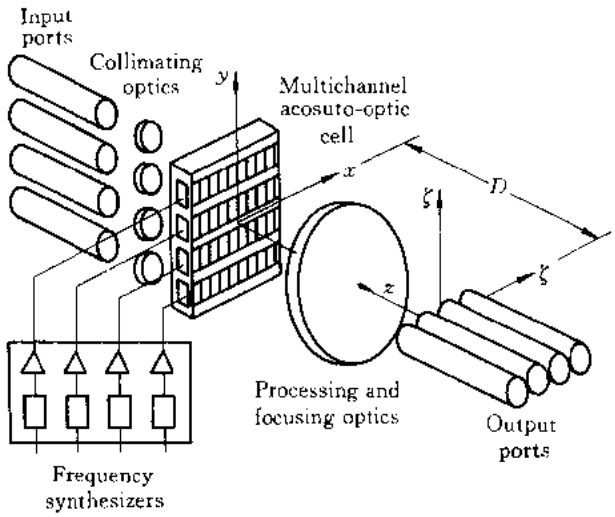


图 2

3.3 美国 TII-CRL 的 DMD 光学交叉开关[23]

采用 N^2-并行矩阵-矢量内积结构,如图 3 所示。SLM 用的是变形镜器件(DMD),电寻址,预计传输速率达 GHz,128 ~ 1000 条通道。

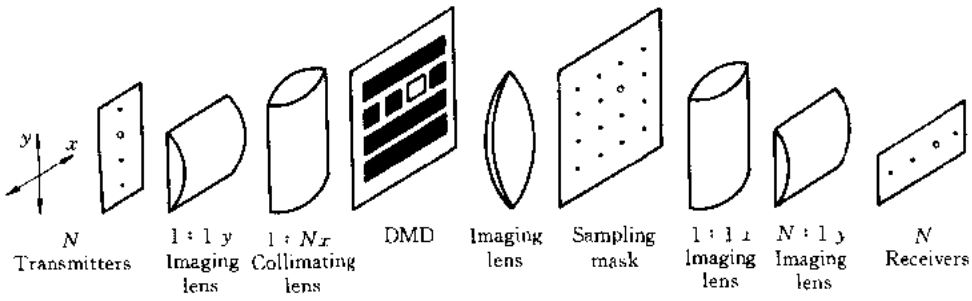


图 3

3.4 美国 UC-COCS 和 DU-CSD 的全息互连网络[24]

该方案使用了 SLM 和多路全息图,如图 4 所示。SLM 用来对 1-D 光场进行编码,动态控制全息路径。SLM 是 4×4 FLC SLM,帧速 10kHz,消光比 25:1,光折变材料用 $LiNbO_3$,必需事先做

上角度多路全息图。与使用 SLM 的光学交叉开关互连网络比不需要 N^2 个开关(只需 $\sqrt{2N}$ 个),与使用体全息图的相比,可以克服光折变材料响应慢的缺点,重构时间主要决定于 SLM 的响应时间,预计使用Ⅲ-V(Inp/InGaAs)MOW SLM 重构速率可达 GHz。实验了 4×4 的互连和串音情况,多路全息图可做到 20 幅。

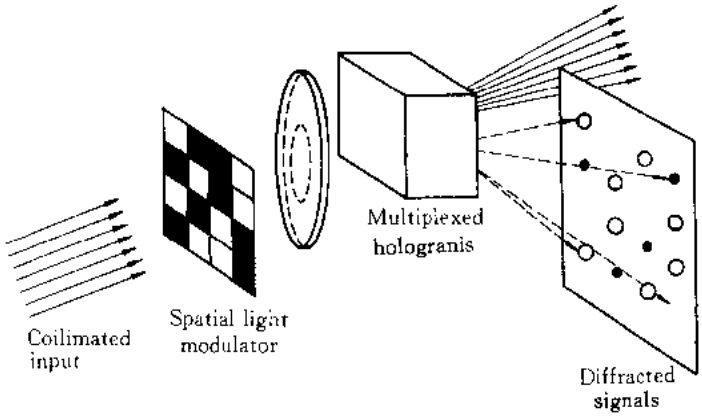
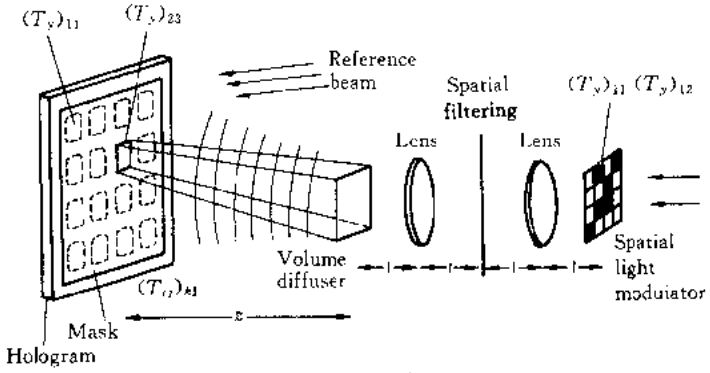


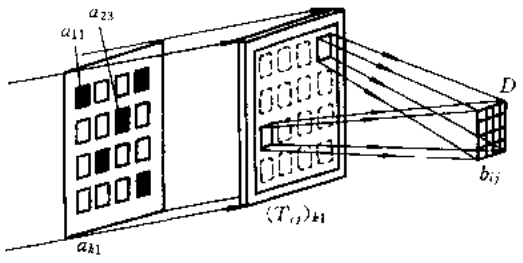
图 4

3.5 美国 POC-RD 的无透镜全息光互连[25]

具有 N^4-并行度,数学表达式为:$b_{ij}=\sum (T_{ij})_{kl}a_{kl}$,如图 5 所示。图 5(a)为全息记录过程,(b)为重现过程。N 可达 256,光可保持垂直入射方式(适合于 SLM),不必以不同的入射角来产生多路全息图。可以考虑用扩束和光栅偏转列阵来提高光效率。



(a)



(b)

图 5

(a) 全息记录过程; (b) 重现过程

3.6 美国 RISC[26~28]

这是一种新颖的利用光折变晶体的方案，N^2-并行矩阵-矢量内积结构，原理是光折变晶体中两波混合非互易能量转换，记录和读出同时进行，如图 6 所示。光折变晶体用 $BaTiO_3$（$BaTiO_3$ 和 SBN 是目前效率最高的实时光折变晶体），光强 $1W/cm^2$ 时建立全息图需 1ms 时间量级，缺点是光强要求太高，最大的优点是能量效率高。

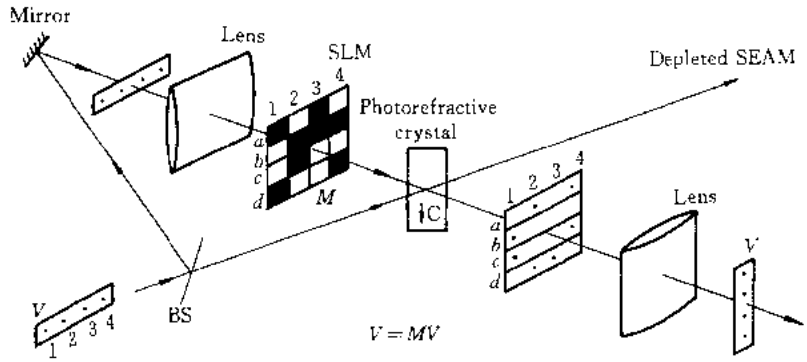


图 6

目前，已提出的互连方案还有许多种[29,30]，这方面的工作仍处于实验、摸索阶段，究竟能否有一种方法具有绝对优势，还是各种方案各有其优势和应用场合，还有待于进一步的发展来证明。但使用光互连来取代（或部分取代）电互连已成为共识，需要光、电和计算机的科研工作者通力合作，才能取得丰硕的成果。

4 光学交叉开关互连网络分析

普通的交叉开关概念是指能够实现没有信息争用的 N 个输入，N 个输出的任意点到点的互连开关网络。广义的交叉开关网络不仅可以实现点到点的连接，而且还具有广播（broadcasting）和汇集（funnelling）能力，具有可重构性和非阻塞性，这正是并行处理中所要求的互连网络。然而随着处理器数目的增加，电子交叉开关网络的硬件费用、体积和管脚数目（VLSI 芯片的管脚数目不可能超过几百）成为限制网络规模的主要因素，另外还存在能耗、时钟分配、控制、可靠性和延迟等问题。一般认为，交叉开关阵列的设备量是 $O(N^2)$，当 N 很大时，其成本可能超过全部 $2N$ 台处理器、存储器和 I/O 设备的成本，比较合理的规模是 $N<32$[31]。IBM 也做有 64×64 的交叉开关用于 Logic Simulation Machine（LSM）和 256×256bit 的交叉开关用于 Yorktown Simulation Engine（YSE）[32]。但一是费用太高；二是存在长周期的可靠性问题。光学交叉开关的提出正是基于以上的原因，利用光学固有的并行性等特点，在 N 比较大时实现交叉开关的互连功能。

使用空间光调制器（SLM）的光学交叉开关互连网络如图 7 所示，其数学模型是 N^2-并行矩阵-矢量内积，$\boldsymbol{C}=\boldsymbol{A}\cdot\boldsymbol{B}$。其中：$\boldsymbol{A}$、$\boldsymbol{C}$ 为 N 维行向量；$\boldsymbol{B}$ 为 $N\times N$ 矩阵[33]。$\boldsymbol{A}$ 为输入光源，可以是 N 个 LED 或 LD 列阵，二进制信号 1 代表某一光强，0 代表不发光；$\boldsymbol{B}$ 是空间光调制器（光控或电控），其中的 0 代表关态即光不能通过，1 代表开态即光能通过；$\boldsymbol{C}$ 为输出的 N 个实时探测器。这种 N^2-并行矩阵-矢量内积光互连结构，就是所说的广义光学交叉开关互连网络，不

仅是具有可重构性和非阻塞性的单级网络，而且空间光调制器的开关设置和传输数据可以是异步方式，非常灵活、方便，是功能最强的互连网络。

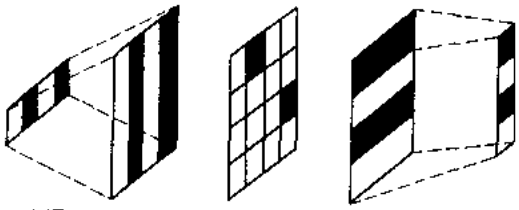


图 7

互连系统的性能除了与互连结构有关外，还取决于光源、光学系统、开关器件和探测器等的性能，下面分析该光学交叉开关互连网络的三个主要参数：互连数目、重构时间和数据传输速率。

(1) 互连数目。互连数目是互连网络所能实现的最大连接数目。当 i 光源与 j 探测器连接时：

$$V_j(t) = \frac{T_{op} G \gamma_{ON}}{N}\left[E_i(t) + \frac{1}{C}\sum_{k=j,k\neq i}^{K} E_k(t)\right] + Be(t)$$

其中：$V_j(t)$ 是第 j 个探测器 t 时刻的输出电压；T_{op} 是光学系统的总能量传输系数；G 是探测器的光电增益；γ_{ON} 是 SLM 为开态时的能量传输系数；N 是互连数目；$E_i(t)$ 和 $E_k(t)$ 分别是第 i、k 个光源 t 时刻的输出光能；$C = \dfrac{\gamma_{ON}}{\gamma_{OFF}}$，是 SLM 的消光比，$\gamma_{OFF}$ 是 SLM 为关态时的能量传输系数；$Be(t)$ 系统引入的噪声。

当 $V_i(t_d) > V_d$ 时，输出逻辑 1；当 $V_i(t_d) < V_d$ 时，输出逻辑 0。其中：t_d 是判断时刻；V_d 是阈值。

假定在 t_d 时，有 l 个光源输出逻辑 1，$k-l$ 个光源输出逻辑 0，则其概率为：

$$BER_{k,l} = C_k^l p^l q^{k-l} \qquad k = N-1$$

其中：p 是光源输出为逻辑 1 的概率；q 是光源输出为逻辑 0 的概率。因此：

$$BER_0^k = \sum_{l=0}^{K} C_k^l p^l q^{k-l} p_0^{k,l}, \quad V_L + \frac{k-l}{C}V_L + \frac{l}{C}V_H + Be > V_d$$

其中：BER_0^k 是总的把逻辑 0 误为逻辑 1 的概率；$p_0^{k,l}$ 是把逻辑 0 误为逻辑 1 的概率；$V_L = (T_{op}/N) G\gamma_{ON} E_L$，$V_k = (T_{op}/N) G\gamma_{ON} E_H$，$E_L$ 和 E_H 分别是光源输出为逻辑 0 和逻辑 1 时的光能。

同样，把逻辑 1 误为逻辑 0 的总概率 BER_1^k 为：

$$BER_1^k = \sum_{l=0}^{K} C_k^l p^l q^{k-l} p_1^{k,l}, \quad V_H + \frac{k-1}{C}V_L + \frac{1}{C}V_H + Be < V_d$$

其中：$p_1^{k,l}$ 是把逻辑 1 误为逻辑 0 的概率。

假定 Be 是偏差为 σ 的高斯白噪声；在任何时刻有相等数目的光源 $(K/2)$ 处于逻辑 1 和 0 态因而可以把阈值设置为：

$$V_d = V_L + \beta(V_H - V_L) + \frac{K}{2C}(V_H + V_L) \qquad 0 < \beta < 1$$

$$p = q = 1/2, \quad \sqrt{SN} = (V_H - V_L)/2\sigma$$

则可以得到互连数目 N 与误码率 BER、信噪比 SN(或平均光功率)、消光比 C 以及阈值设置 β 的关系：

$$BER(SN) = \left(\frac{1}{2}\right)^{k+1}\sum_{l=0}^{K} C_k^l\left\{G\left[\left(2\beta + \frac{k-2l}{C}\right)\sqrt{SN}\right] + G\left[\left(2 - 2\beta - \frac{k-2l}{C}\right)\sqrt{SN}\right]\right\}$$

$$G(Z) = \frac{1}{\sqrt{2\pi}}\int_{z}^{\infty} e^{-t^2/2}dt$$

计算机间的数据传输，误码率要求低于 10^{-3}，而计算机内的数据传输，误码率要求低于 10^{-12}。在保证误码率的前提下，提高互连数目，一方面可以通过改进空间光调制器的性能增大消光比 C；另一方面可以提高系统的信噪比，如降低光电探测器的噪声系数；但更实际一些的是采用纠错码技术，降低误码率，从而提高互连数目。纠错码一般适用于功率受限而带宽不太受限的信道，这恰好和光通信的特点相一致。采用纠错码技术后的误码率 BER′ 为：

$$BER' = BER - \left[\sum_{k=1}^{r_1} C_{r_1}^{k} - BER^{k}(1 - BER)^{r_1-k}k\right]/n$$

其中：n 是码长；k 是纠错个数。

从计算结果可知，N 可以大大提高，N 可以大于 C。$N = 500$ 是可以实现的，采用多级网络的话，$N = 1000 \sim 20000$ 是可能的。

(2) 重构时间。重构时间 T 是改变全部开关的时间，即建立一个完全不同的互连路径的时间。原则上，重构时间应小于 1 个数据位的持续时间，但在保证数据不丢失的前提下，且每一种互连路径下传输的数据量比较大时，重构时间大于一个数据位的持续时间也是很有意义的。有四种控制开关列阵的方式：顺序、行并行、随机和并行。只有光寻址 SLM 才能实现（N 相当大时）并行控制。重构时间 T 决定于列阵的控制方式、互连数目 N 及每一个开关的开关时间 t_s。顺序控制时，$T = N^2 t_s$；行并行或随机时，$T = Nt_s$；并行控制时，$T = t_s$。目前可用于光学交叉开关互连网络的声光、电光、磁光 SLM 的开关时间在 $1 \sim 0.1\mu s$，所以重构时间比较大，随着 SLM 性能的改进，1ns 的开关时间是可以达到的，从而对中、大规模的互连网络来说，重构时间在 ns（并行）或 μs（行并行）量级。

(3) 数据传输速率。数据传输速率是在给定误码率的互连路径中的数据位速率。对于我们所讨论的自由空间被动光学交叉开关互连网络，网络内不存在光信号的探测和再生，数据传输速率决定于光源和探测器的带宽，以及探测器所能接收到的平均光功率（与光源功率和信道损耗有关）。分立 LD 和探测器的带宽分别可达 20GHz 和 40ps，再考虑到光源功率、N/C 比值、光学系统损耗和探测器的灵敏度等因素，实现 1Gb/s 的数据传输速率是很容易的；一维列阵光源和探测器的该值小一些，大约为 50Mb/s，这也比典型的电子系统的 10Mb/s 大（虽然电子开关的速度达 50～100ns）。目前，实验室水平的光开关的开关时间已达到达 0.1ps，光窄脉冲达 30fs，可见光学系统的数据传输速率的潜力很大，而电子系统的数据传输速率的极限为 1Gb/s。

我们拟采取如图 8 所示的光路实现 N 个输入接口与 N 个输出接口的任意互连（即任意第 i 个输入端可以与任意一个或几输出端连接），这种互连是动态的，可以通过 SLM 的控制信号（寻址信号）任意改变。首先使用现有的 LCLV 作为开关器件的 SLM，在此基础上再采用我们自己研制的 Si/PLZT SLM 代替 LCLV，以期大幅度的减小互连网络的重构时间（可达到 $1\mu s$ 左右）。

我们采取的研究方法及步骤为：

首先，采用分立的光源（带尾纤）、探测器及固定掩版搭出 4×4 的互连网络（具体光路见图 8），光源为四个带尾纤的波长在 $0.8 \sim 0.85\mu m$ 的 GaAs/GaAlAs LD，探测器与此相匹配，光纤调节架采用 Si 片上的 V 型槽，通过搭建该光路取得下列参数或经验。

(1) 采用图 8 所示的光路，通过示波器上波形图验证该光路的互连能力。

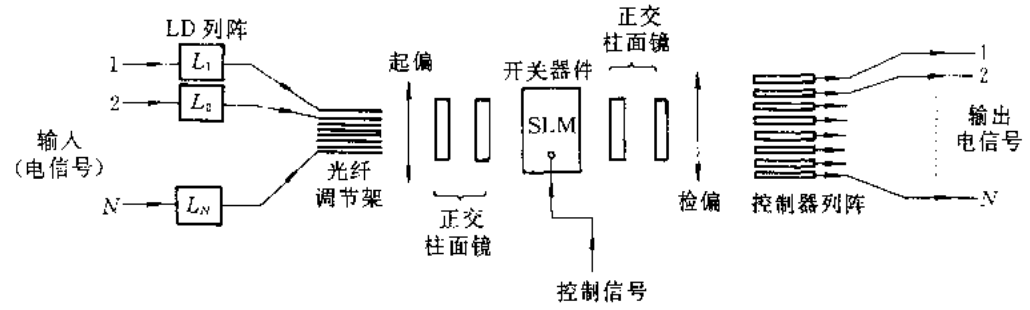


图 8

(2) 采用图 9 所示的光路测试该互连网络的传输速率和串话。通过串话测试,一方面可取得光路对准的经验,另一方面可检验影响互连数目 N 的因素(除 SLM 的对比度外)。

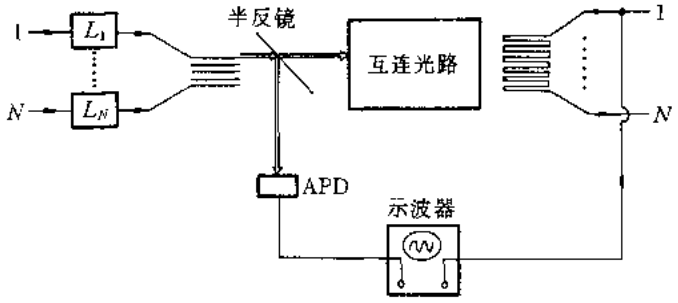


图 9

(3) 采用不同的 V 型槽(槽宽和槽间距)、不同的掩版(方框尺寸)、不同的探测器(光敏面尺寸)以及不同的柱面镜(焦距和孔径),摸索各个元件之间尺寸的最佳匹配,以达到最佳的互连效果。并可据此结果设计 Si/PLZT SLM 的像素尺寸以及由已知的 LCLV 像素尺寸设计出其他元件的尺寸。

第二,根据上述经验设计出一维列阵的 LD、PIN 探测器的尺寸参数,搭建出采用一维列阵的光源。探测器以及电寻址的 LCLV 的 4 × 4 互连网络,并对其进行性能测试,重构时间采用图 10 所示光路进行测试。

在此基础上以 Si/PLZT SLM 代替 LCLV,希望在重构时间上有数量级的提高。

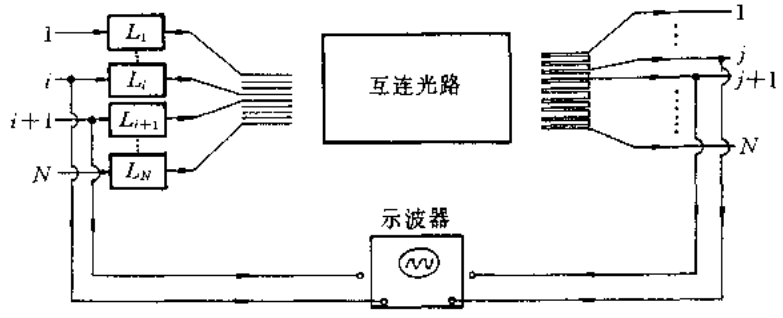


图 10

我们正在研制的 Si/PLZT SLM 单元结果如图 11 所示,探测器列阵对写入光进行光电转换和放大,提供调制器工作电压。PLZT 电光调制器采用 F-P 腔结果可大幅度降低调制电压和功耗,两者的混合集成采用精密抛光面间的直接压紧、固定的方式或采用倒装焊接技术。这种结

构空间分辨率高、功耗低、灵敏度高、工艺易实现，因而实用性很强。这种方案在国际尚属首次提出。制作时先分别制成 PLZT 和 Si 探测器列阵，再将两者混合集成。

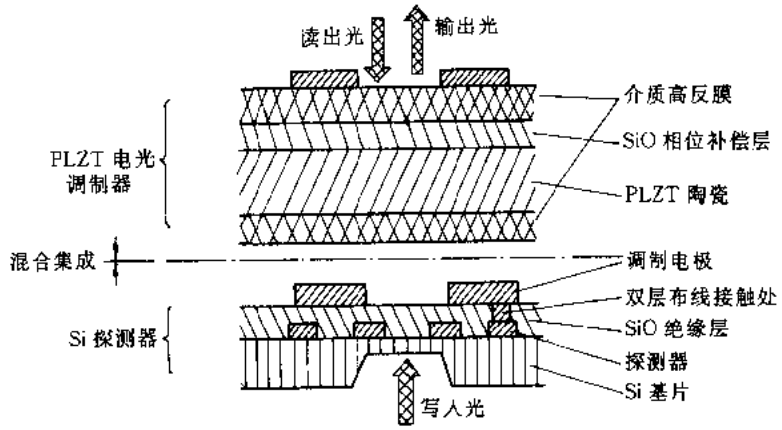


图 11

上述为光寻址 SLM，如光互连网络中的控制信号为电信号，则需制作电寻址的 SLM，其结构与图 10 类似，可将控制信号直接加到放大寻址电路上，代替光探测器。

参 考 文 献

[1] 陆绍福.90 年代的热潮——超高速计算转向并行处理.计算机世界,1992,(5):2

[2] 李 凯.国际计算机发展趋势.计算机世界,1992,(5):20

[3] L.Rudolph. A Role for Optics in Future Parallel Processing. SPIE,1991,15:175

[4] 陈益新等.光计算.上海:上海交通大学出版社,1990

[5] J.W.Goodman, et al.Optical Interconnections for VLSI Systems.Proc.IEEE,1984,72(7):850

[6] 陈益新.光计算技术展望.中国计算机报,1992,(5):12

[7] J.A.Neff. Optical Interconnects.MAPLE PRESS,1988.166

[8] J.E.Midwinter, et al. Optoelectronic Interconnects in VLSI; The Reality of Digital Optical Computing? IEEE LCS Magzine,1990.40

[9] R.T.Chen.Optical Interconnects: A Solution to Very High Speed Integrated Circuits and Systems. SPIE,1990,1374:182

[10] P.R.Haugen,et al.Optical Interconnets for High Speed Computing.Opt.Eng.,1986,25(10):1076

[11] J.W.Goodman.Fan-in and Fan-out with Optical Interconnections.OPTICA ACTA,1985,32(12):1489

[12] A.R.Dias,et al.Fiber-optic Crossbar Switch with Broadcast Capability. SPIE,1987,825:170

[13] K.P.Jackson.Optical Fiber Interoonnection for the Computer Center.SPIE,1988,994:50

[14] M.K.Kilcoyne,et al.Optical Fiber Crossbar Switch.SPIE,1990,1215:17

[15] F.Lin,et al.Optical Muitiplanar VLSI Interconnects Based on Multipiexec Waveguide Holograms.Appl.Opt.,1990,29(8):1126

[16] Y.Yameda,et al.Guide-wave Chip-to-chip Optical Interconnections.SPIE,1988,881:164

[17] Tatsuo Izawa. Plastic Plannar Waveguides for Optical Interconnects

[18] R.T.Chen,et al.Optical Interconnection Using Polymer Microstructure Waveguides.Opt.Eng.,1991,30(5):622

[19] P.Churoux,et al.Optical Crossbar Network Analysis.SPIE,1987,382:42

[20] M.Fraces,et al.A Multiprocessor Based on an Optical Crossbar. Network:the MILORD Project.SPIE,1988,963:223

[21] M.Fraces,et al.The Optical Crossbar Network MILORD Machine:Last development and Results.SPIE,1990,1281:66

[22] D.O.Harris, et al. Acousto-Optic Photonic Szitch: An Optical Crossbar Architecture. SPIE, 1989, 1178:93
[23] R.W.Cohn. Link Analysis of a Deformable Mirror Device Based Optical Crossbar Switch. SPIE, 1987, 325:178
[24] E.S.Maniloff, et al. Hoiographic Routing Network for Parallel Processing Machines. SPIE, 1989, 1136:283
[25] F.Lin. Practical Realizations of N^4 Optical Interconnects. Appl. Opt., 1990, 29(35):5226
[26] P.Yeh, et al. Optical Interconnection Using Photorefractive Dynamic Holograms. Appl. Opt., 1988 27(11):2093
[27] A.Chiou, et al. Energy Effeciency of Optical Interconnections Using Photorefractive Holograms. Appl. Opt., 1990, 29(8):1111
[28] A.Chiou, et al. Reconfigurable Optical Interconnection Using Photorefractive Holograms. SPIE, 1151:24
[29] Toshikazu Sakano, et al. Design and Performance of A Moltiprocessor System Employing Board-to-board Free-space Optical Interconnections: COSINE-1. Appl. Opt., 1991, 30(17):2334
[30] T.J.Cloonan, et al. Self-Routing Crossbar Packet Switch Employing Free-space Optics for Chip-to-chip Interconnections. Appl. Opt., 1991, 30(26):3721
[31] 刘重庆.并行处理机与应用.上海:上海交通大学出版社,1990
[32] A.A.Sawchuk, et al. Optical Crossbar Networks. IEEE. Computer, 1987, 50
[33] A.A.Sawchuk, et al. Dynamic Optioal Interoonnetions for Parallel Processors. SPIE, 1986, 625:143

磁记录技术的现状和趋向

1 引言

磁记录是当代信息存储的一项主要技术。无论作为计算机的外存储设备，或是在录音和录像等消费类电子产品领域，虽然有半导体和光存储等技术的激烈竞争，但磁记录仍有绝对的优势。这不仅因为磁记录具有优异的记录性能、应用灵活、价格便宜，而且在技术上仍具有相当大的发展潜力。可以预期，到本世纪末，磁记录技术在信息存储领域中的主导地位不会被取代。

根据所记录的信号类型，记录介质的物理型式或信号编码方式等的区别，磁记录应用可采用不同的方法进行分类。目前，通常将磁记录应用分为数字磁记录和模拟磁记录两大类，前者主要包括用于计算机数据存储的硬磁盘、软磁盘和磁带，它们采用二进制的数字信号；后者主要用于图像、声音和仪器信号的存储，它们采用频率调制，线性模拟记录，也可用数字编码，但是其输出必定是模拟输入信号的真实再现。

数字磁记录和模拟磁记录的要求主要取决于系统功能上的区别。对数字磁记录，最主要的要求是数据可靠性高，记录数据的快速存取以及记录每位信息的成本低。对模拟磁记录的基本要求往往是信噪比高、失真度低以及播放时单位时间的成本低。

一般来说，每个磁记录系统都由机械驱动的记录介质和相对应的磁头这两主要组成部分构成，记录介质往往带有适当的包装，以保证其方便和可靠地使用。同时，还必须有信号处理的电子线路将输入信号转换成一定波形的电流送至记录的磁头。与此相对应，另一种信号处理是将从磁头读出的重放信号无畸变地回复到原始信号供输出。

自从 1940 年后期以来，磁记录系统长期沿用着涂敷磁带、环形磁头以及纵向（水平）记录方式。最先应用的主要产品是记录磁带，直到现在继续广泛使用，如图 1 所示，采用交流偏置的线性模拟记录方法。许多仪器上的应用也是交流偏置，它们要求在增加记录密度的同时确保高的信噪比，这促使录音和仪器的应用在高密度磁记录技术中一直处于领先地位。美国 IBM 公司 1981 年（Harris, et al[1]）设法将磁带记录用于数字数据，并采用无偏磁的非线性记录。这种磁带驱动器要求高速运行以获得快的数据存取。

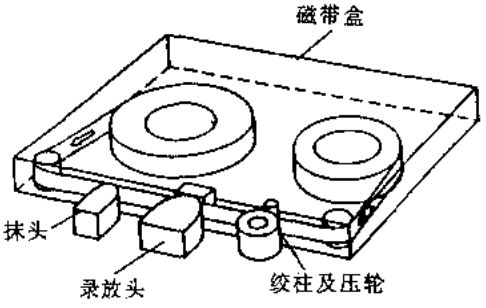


图 1 盒带录音

磁带记录的另一个重大发展是用于录像。由于录像要求非常高的频率，磁头静止的磁带机无法实现如此的高频记录。1956 年 Ampex 公司提出了采用磁头沿磁带横向扫描的方法使磁带录像取得突破，这时所用的磁带宽为 5.0cm，磁头横向扫描速度高达 40m/s，磁带纵向走带速度也有 38cm/s。到

1976年，随着技术的进一步改进，日本的公司(Kihara, et al[2], Shiraishi, et al[3])提出了采用磁头螺线扫描的方法，使系统的结构坚固紧凑，成本下降，从而引进到消费电子产品，成为现在已十分普及的家用录像机，其系统装置原理见图2。

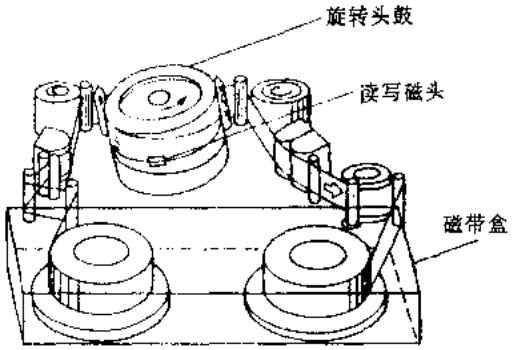


图2 螺旋扫描录像装置

磁记录在另一方面的重要发明是IBM公司于1957年提出了用于数字数据存储的旋转式硬磁盘装置(Lesser and Haanstra[4], 1957年)，如图3所示。利用空气轴承使读写头起飞，从而可大大提高头-盘间的相对速度。这不仅证明具有高度可靠性和数据传输速率，而且磁头沿盘片径向运动作快速随机存取。因而使这种硬磁盘数字记录装置成为计算机普遍采用的在线外存储器。同时，接触式的软盘记录装置由于驱动器使用方便和成本低廉，于1974年(Noble[5])问世以来已大量应用了微型计算机和个人计算机系统，其装置见图4。

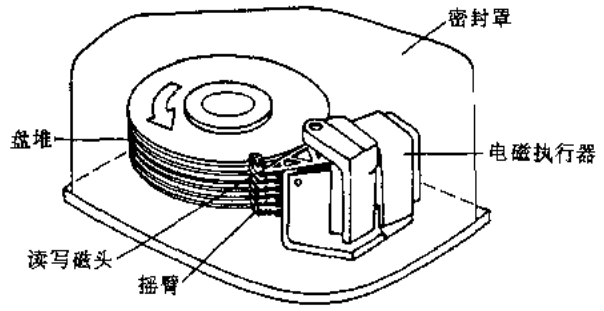


图3 硬磁盘驱动器结构原理

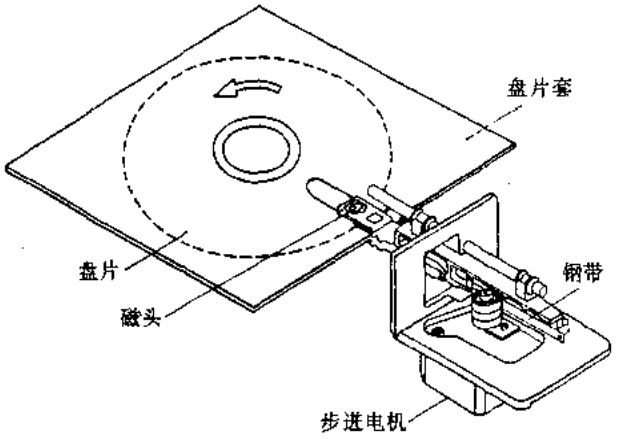


图4 软磁盘驱动器结构示意图

用激光束在磁介质上进行读写的磁光记录(Kryder[6], 1985)是另一项值得重视的技术。它主要的特点是头和介质完全不接触，而且两者间隔相当大，并能实现可擦除的记录，提供非常高的道密度，从而获得很高的记录面密度，其装置原理如图5所示。磁光记录的原理是当激光束局部加热垂直磁化膜的微区时，只要在较低的外磁场作用下就可使磁化反转，这样就可写入信息。这是利用磁化的温度效应，信息的擦除也可以利用这一效应。这种写入和擦除方式有

两类,即磁化反转在居里温度或在补偿温度发生。记录信息的读出将利用磁光效应——克尔磁光效应或法拉第磁光效应,即当线偏振激光束在磁化记录介质表面反射或透过时,光束的偏振面发生旋转,旋转的方向和角度大小与介质磁化的方向和强度有关。由于信息的读出没有热效应,因此读出所需要的激光功率比写入时低。激光束的聚焦光斑与波长有关,采用波长较短的激光可以增加记录密度。

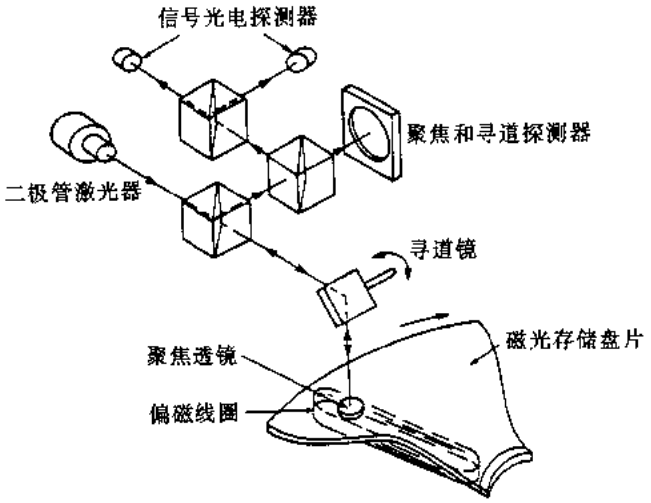


图5 磁光记录装置原理

由于磁记录产品成功地应用于数据存储、录音、录像和仪器记录等广泛领域,这就大大推动了磁头、介质和有关电子器件的性能不断提高。例如,磁头和记录介质所用材料的磁性能和机械性能的改进使记录面密度不断增加。同时随着机械装置和电子线路等不断改进,道密度也得到相应提高。记录面密度为线密度和道密度的乘积。磁记录技术在最近三十年的发展中,面密度已提高约六个数量级。硬盘,软盘和磁带录像(VTR)记录密度的发展现状见图6。

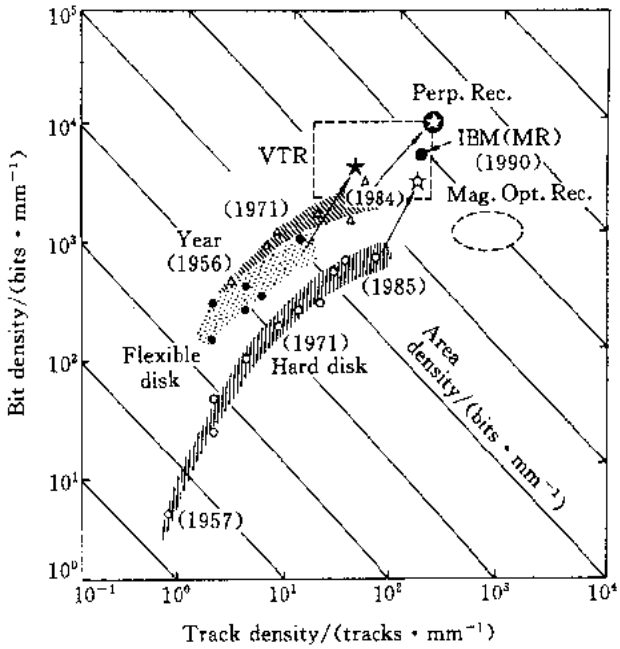


图6 磁记录密度的发展

由于篇幅原因，下面仅就数据存储和薄膜磁头技术发展的趋向作一概要评述。

2 硬盘磁记录

硬盘磁记录的装置如图3所示。记录系统包括一组表面覆有记录介质的盘片装在主轴上构成盘堆，它由一个主轴电机带动在密闭的壳体内高速旋转，信息的写入和读出是通过每个盘片表面可以精确定位的磁头来实现，这些磁头固定在能加载一定力的簧片上并与一排平行的取数臂相固定而构成头堆。在头与盘片表面之间通过相对运动产生的空气动力以保持一定的间隙，这称为飞行高度。取数臂由电磁执行机构驱动并使磁头可以精确定位在所要求的信息道上实现随机存取，如图7所示。这种设计已普遍采用于从35.6～6.3cm的各种硬盘系统，它们的容量已从数兆字节增加到数千兆字节。

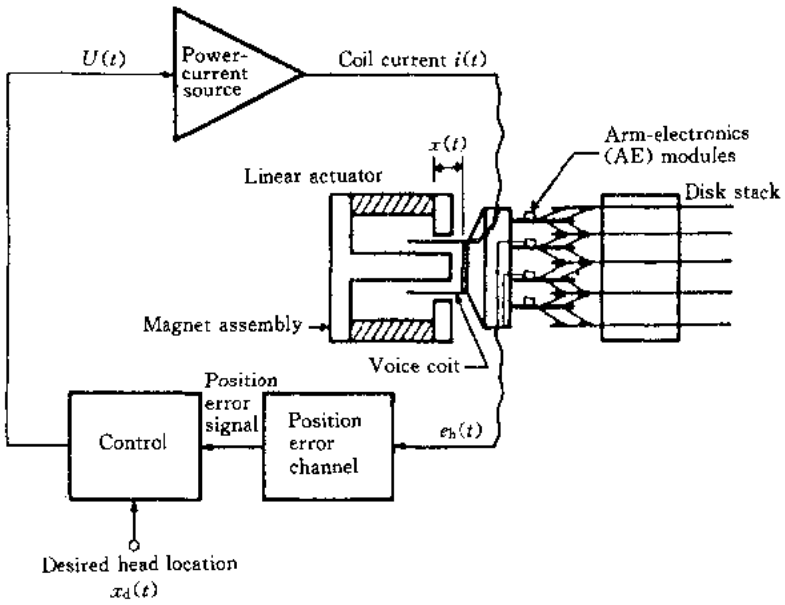


图7 典型的磁头定位伺服系统框图

但是随着对硬盘记录装置要求的进一步提高，上面介绍的那种系统的设计已面临着新技术的严峻的挑战，这些涌现出的新技术包括：(1)双元件组合的薄膜磁头；(2)连续薄膜记录介质；(3)独立控制的埋入伺服；(4)空气轴承的执行机构；(5)高级的信号处理和编码；(6)垂直磁记录模式。

1989年，美国IBM公司报道了最新的实验结果，13.3cm(5.25in)的硬盘驱动器的面密度已提高到1.8Mb/mm^2，已超过了光记录的面密度，它采用了感应和磁阻组合的双元件薄膜磁头，合金薄膜介质，埋入伺服和"部分响应最大可能"PRML(Partial Reponse Maximum Likehood)编码等新技术。

3 数据磁带记录

近年来数据磁带记录技术有了很大的发展，特别是盒带技术。目前数据记录采用的磁带

有开盘式、单盘盒式、双盘盒式等，磁带宽度有 12.7mm(0.5in)、6.3mm(0.25in)、8mm 和 4mm 等。在高性能的系统中，磁带驱动机构必须保证在 508cm/s(200in/s)的高速下使磁带精确地导引并与多道读写磁头良好接触，对在线应用的磁带机必须能快速起停以减少搜寻时间，因而需采用伺服控制的低惯性电机，并应用真空柱使磁带从读写磁头区域导柱上的机械解耦。最近采用了数据流的电子线路缓冲技术使对高速起停的要求有所减轻。使磁带的加速和减速得以减少而不致损失磁带上的记录面积。另外，缓冲器的应用也有助于不同驱动器之间磁带互换，以保证多道磁头与磁道间的重新对准。这样，数据磁带机的设计可以简化如图 8 所示。这是采用静止多道读写头的单轴盒带的驱动器，这种驱动器能与在线数据磁带机数据率相匹配。

12.7mm(0.5in)数据磁带记录密度的提高可通过改进氧化物磁性介质，降低记录介质的缺陷数以及优异的头和介质界面设计，使头和界面之间保持均匀的很小的间隙，从而磨损减少。最有代表性的 12.7mm(0.5in)磁带机是 IBM 公司于 1984 年推出的 IBM-3480 多路并行读写盒带存储子系统，它用于大型和超大型计算机系统中，已逐步取代原来 1.27mm 开盘式磁带机 3420。这是因为 3480 具有许多明显的优点：体型小，机构紧凑，记录介质有盒保护，操作时不与人手接触，采用 18 路并行薄膜磁阻磁头及功能强的纠错码，可靠性提高等，现在美国和日本许多磁带机的公司都开发了各种 4380 的兼容磁带机。

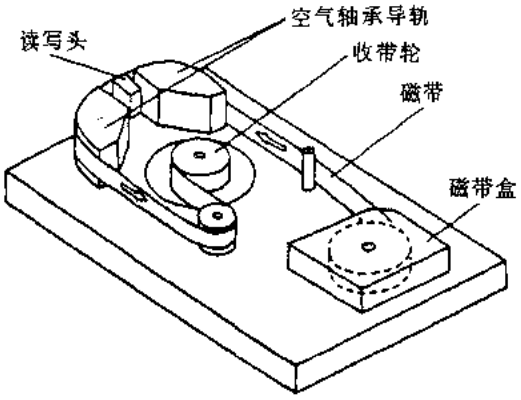


图 8　用于数据记录的盒带驱动器

在微机和工作站系统中，6.3mm(0.25 in)盒带驱动器 QIC 得到了飞速的发展，其单机容量达 525MB，并逐步增加到 1.35GB、2.7GB、和 5.4GB，甚至会超过 10GB。

根据现有报道，各种数字磁、光、半导体存储技术最新发展的比较列于表 1[7]。可以肯定地说，磁记录仍是一个富有活力的技术和产业，1990 年的年销售额已超过 500 亿美元，而且能期待有进一步的发展。现在在美国、日本和欧洲，许多大学拥有相当规模的教授和研究生，正在从事这一领域的研究。

4　硬盘磁阻头的实用化

磁阻头从磁记录介质读出信息，是利用某些金属的电阻由于作用磁场的变化而改变的特性，这称为磁阻(magnetoresistive)效应。这种磁阻头首先出现在计算机用的磁带机中。由于磁阻头具有比普通感应式磁头更高的灵敏度和输出信号，能与高矫顽力介质之间相配合实现高密度记录，其输出信号幅值与头和介质之间相对速度无关等一系列优点，近年来，许多公司都在竞相研究，比谁能最先把磁阻头用于硬盘驱动器中。

1990 年 7 月在磁记录国际会议上提出硬盘磁阻头报告的公司有美国的 IBM 和 AMC(Applied Magnetics Corporation)、日本的 NEC，美国的 Eastman Kodak Company 提出了一种高密度磁带机磁阻头。在硬盘磁阻头方面，从产品性能和研究工作的深度看，美国 IBM 公司仍居于领先地位。从大学来看，Carnegie Mellon 大学在磁阻头的基础研究方面是很引人注目的，但由于设

表 1　数字磁、光和半导体存储技术的比较

		容量 GB/单元[a]	位密度 位/mm	道密度 道/mm	面密度 /(kb·mm^{-2})	堆密度 面/mm	体密度 /(kb·mm^{-3})[b]	数据率 /(Mb·s^{-1})[c]	取数时间/s[d]
硬磁盘机	IBM 3390(1989)	3.8	1 100	88	96	0.08	8	34	0.02
	Seagate Elite(1989)133mm(5.25in)	1.2	1 320	74	100	0.3	30	24	0.018
	IBM 实验样机(1989)	–	6 200	290	1 800	–	–	28	–
数字磁带机	IBM 3480(1984)	0.2	900	1.5	1.3	40	52	24	15
	QIC 525(1989)	0.525	800	5	4	120	480	2.0	40
	Exabyte 8mm(1987)	2.3	2 125	32	60	80	4 800	1.9	100
	R-DAT (1987)	1.3	2 400	74	180	80	14 400	1.5	15
	Matsushita 数字 VCR 样机(1989)	36	2 800	150	420	100	42 000	27	~ 120
光盘机	CD(1981)	0.64	1 600	600	1 000	0.5	500	1.5	0.6
	Sony 磁光机(1989)	0.64	950	740	700	0.4	280	0.7	0.1
半导体	4Mb 动态 RAM(1989)	0.5	–	–	1 200	–	–	12	80ns

注：a)每单元是指每头盘组件、或每磁带盒、或每光盘；b)对磁带和光盘来说是离机的；c)每个头，除 IBM 3480(18 头)和数字 VCR 样机以外，其中有两个同时有效的；d)随机寻道的等待和平均存取时间

备条件的限制，美国的任何大学都不能进行薄膜磁头和磁阻头的产品研制。

由于磁阻头的局限性只能读出，不能写入，因而实用的磁头必须设计成双元件的，即采用感应头作为写入和擦除，磁阻头供读出。IBM 公司研制的双元件磁阻头能达到的记录密度为 1.8Mb/mm^2，在道误码率低于 10^{-9}。在此以前，IBM 公司 3390 硬盘驱动器具有最大记录密度为 0.1Mb/mm^2。IBM 公司这种磁阻双元件磁头记录的信号，其道与道之间距离为 3.4μm，每位的长度(即位波长)为 160nm。具有屏蔽结构的磁阻元件与盘片间的飞行高度约为 60nm。这种磁头的磁阻元件应用镍铁合金(坡莫合金)薄膜，主要特点是具有很高的磁阻系数，当磁场强度与电流之间的夹角改变时，电阻的变化最大可达 2.5%。

双元件磁阻头与单元件感应式薄膜头相比，MR 头其主要优点为除了具有很高的灵敏度外，并能获得比较理想的输出波形，易于均衡，能用矫顽力极高的记录介质，因而可达到很高的位密度。在道宽大大减小情况下，不会产生严重的不稳定性问题。与 MIG 头相比，MR 头的主要优点是更适用狭磁道和高频记录。MR 头与所有各种感应头相比，其突出优点是输出幅值高，特别在盘片速度大大减少的情况下仍不受影响，还可用比写入头的磁极宽度更狭的读出头，使偏道的问题得到很好地克服。

但是，要使 MR 头实际应用于商品驱动器产品，还有许多工作要做，由于磁阻头必须做成既有 MR 元件又有感应元件的双元件磁头，这就造成结构和加工都比普通的薄膜磁头更为复杂，以至成品率降低、成本提高。虽然已证实这样高的记录密度对硬盘驱动器来说是可行的，但在工程实际上，能够确保在道密度 300 道/mm 和飞行高度 60nm 条件下稳定工作，尚需不断改进。因为这时磁头在道上的定位公差只有 0.25μm，这比现有产品要小 1/2 ~ 2/3，这意味着其寻道时间和定位的精度还要提高许多。这不仅在机械上和取数臂结构上要进一步改进，而且还必须采用数字信号处理及扇区伺服等新技术。

5 薄膜磁头的改进和扩大应用

薄膜磁头的实际应用已有 10 年以上的历史，其性能有了很大提高。最早的产品磁道宽度达 50μm，现在已减少到 5.5μm，甚至更小；线圈匝数已从 8 圈增加到 45 圈或更多。最初的薄膜磁头采用干法工艺即掩模蒸镀，其后又改进采用溅射淀积及溅射及离子束刻蚀。后来发展了电镀工艺制造磁极和线圈，得到广泛应用。目前大多数感应式薄膜磁头的制造采用了干法和湿法的混合工艺，即厚度为数十 nm 的电镀底层、磁极间隙介质层以及最上面的覆盖层采用溅射方法，线圈和磁极采用掩模电镀；而线圈之间以及与上下磁极间的绝缘是经高温烘焙的抗蚀剂。

随着记录密度的提高，要求磁极尖的宽度减少，在线圈方面既要减少面积，而且线圈的电阻和电感不应增加，反而要求更低。这些都要通过提高电镀的分辨率来实现。可以指出，电镀的分辨率主要由光学制版的分辨率决定。根据现有光学制版的设备和技术达到这些要求应该不成问题。与大规模集成电路相比，对抗蚀剂材料的要求在这里将更为重要。

虽然磁阻头显示出了巨大的优越性，但由于它只能读出的局限，人们对感应薄膜磁头的进一步改进仍寄予极大希望。其中一项重要的研究是采用新的磁极材料来替代现有的 NiFe 薄膜。当采用高矫顽力的金属薄膜记录介质和磁头的尺寸不断减少时，限制边缘磁畴，改进磁头的稳定性，提高磁轭和极尖的磁导率和饱和磁化强度等就显得尤为迫切，如果这些材料能形成

多层结构则更为理想。从湿发工艺来说，现已找到采用单个电镀槽可以制成 NiFeCu/Cu/NiFe-Cu 多层膜。另外，CoFe 和 CoNiFe 及 NiFeB 的多层结构也能从单一的溶液中获得。如果采用干法溅射工艺，在多层结构材料选用方面具有更大灵活性。

例如在溅射 Fe 或 NiFe 时通进一部分 N 形成 Fe-N 或 NiFe-N 非磁性层，就可以大大改善多层膜的性能。IBM 公司研究了这种多层结构，总厚度从 0.7～3.0μm 变化，做成磁极的几何形状后不出现闭合磁畴。膜的磁致伸缩接近于零，易轴的 $H_c \leqslant 32A/m$，难轴的 $H_c = 16A/m$，$H_K = 320 \sim 480A/m$ 以及 $4\pi M_s \approx 20kG$。

薄膜磁头的另一重大努力方向是提高生产率和降低成本。一种富有成效的尝试是用硅片作为基片。为了避免目前这种薄膜磁头由于精密机械加工而使成品率降低和成本提高，这里采用了不同的结构形式：将线圈平面和磁头滑块的飞行面相平行，而且即使表面光洁度要求极高的飞行面也不再经过研磨而成，全部可以通过类似于 VLSI 加工的平面工艺来完成。这种头称为 IC 薄膜头，在 100mm 的硅片上可以完成 1000 个 IC 头，200mm 硅片可做 4000 个 IC 头，从而使生产效率提高，成本将低于传统的铁氧体线绕头。

这种磁极和线圈都由薄膜技术加工成的薄膜磁头，另一个发展动向就是从硬盘驱动器向其他磁记录装置扩展。首先是磁带机用薄膜磁头。磁带记录包括模拟和数字记录，采用薄膜磁头的优点是可以做成道宽很狭的多磁道磁头，适合大量生产，降低制造成本。因此实际上，当开始研究薄膜磁头时，人们就已注意到了在磁带记录中的应用。与硬盘驱动器相比，磁带头由于工作时是与记录介质紧密接触的，因此存在严重的头和带之间的机械界面问题需合理解决。另外，头和带之间的相对速度远比磁盘中低，因而使用 MR 头作为读出元件尤为必要。

最早实际应用的薄膜磁带头是用于高性能的计算机磁带机中，共有 18 道，头-带机械界面问题是通过将薄膜头技术与传统的铁氧体头技术相结合来解决的。这种头现在有几家制造厂生产，每年产量几万只。数字化专业录音的发展，要求提供多道数字录音设备。例如高性能的数字录音机(DASH 格式化)应用了 48 道的薄膜头。

6 磁头设计和特性研究更多采用计算机辅助

由于记录密度的提高，磁道越来越窄，磁极尖的宽度减少，因而现在采用二维空间的近似分析方法计算磁头场必须用三维空间模型来代替。由于磁头场分布十分复杂，因而对其性能影响的因子繁多，所以这种模拟的计算量极大，往往要化若干天的时间，有时甚至要数月。Carnegie Mellon 大学提出了记录头的三维有限元分析方法，其中几何形状的产生是采用专用的固化的模型指令，优化的有限元网格是通过 Delaunay 算法而自动产生的。这种算法与误差分析结合，从而提供了自适应的网格细化，还采用高次正切矢量有限元，获得高精度解。采用此法设计磁头，可以决定磁头的饱和特性、边缘磁场分布、以及一些主要的读写性能。这种方法还能在生产过程中遇到几何形状和材料参数变化时，对性能的影响可以很快地确定。

要能精确地模拟薄膜磁头的读写性能，至少要取 36 个参数，其中 19 个是磁头的几何尺寸和材料参数，17 个是盘片、前置放大器、写电流、滤波器及系统参数。Read-Rite 公司报道了薄膜磁头读写性能的 Monte-Carlo 模拟。研究了 10 个磁头参数，包括：磁极厚度、间隙长度、道宽、喉部高度、线圈电阻、电感、绝缘厚度、磁轭厚度、磁导率及飞行高度等变化对磁头电性能的影响。

参 考 文 献

[1] J.P.Harris, W.B.Philips, J.F.Wells, et al. Innovations in the Design of Magnetic Tape Subsystem. IBM J.Res. Dev., 1981,25:691

[2] N.Kihara,F.Kohno, Y.Ishigaki. Development of a New System of Cassette Tape Consumer VTR. IEEE Trans. Consum. Electron., 1976, CE-22:26

[3] Y.Shirashi, A.Hirota. Viedo Cassette Recorder development for Consumers. IEEE Consum. Electron., 1978, CE-24:468

[4] M.L.Lesser, J.W.Haanstra. The Random-Access Memory Accounting Machine.1. System Organization of the IBM 305. IBM J.Res.Dev.,1957,(1):62

[5] D.L.Noble. Some Design Consideration for an Interchangeable Disk File. IEEE Trans. Magn., 1974,MAG-10:571

[6] M.Kryder Magneto-Optic Recording Technology. J.Appl. Phys., 1985,57:3913

[7] R.W.Wood. IEEE Spectrum,1990,27(5):32

垂直磁记录的新进展

1 概述

磁记录是当今世界极富生命力的一个产业，包括计算机系统和消费用的全部产品在内的全部销售额已超过500亿美元，并且增长趋势不减。磁记录的面密度以每2年或3年增加1倍的速度，已保持了三、四十年，而且随着技术的不断改进，这种增长速度还将继续下去。与其他各种记录技术相比，磁记录仍占着绝对优势。

垂直磁记录作为磁记录的一项新技术经过了10年的研究开发，目前正逐步走向实用化。1989年在日本召开的第一届垂直磁记录国际会议可以作为一个里程牌，它总结了10年来在垂直磁记录的理论上和技术上所取得的巨大成就。从物理上说，垂直磁记录可以提供更高的记录密度；从技术来说，它与现有的纵向磁记录具有高度兼容性。本文就垂直磁记录在软盘、硬盘和磁带磁记录装置上的应用，以及垂直磁记录的磁头和记录介质的发展作一评述。

2 垂直磁记录的应用

2.1 软盘

理论和实验证明，垂直磁记录的性能与头盘间隙的关系比纵向记录尤为明显。一般认为头盘间隙保持在小于0.1～0.15μm，才能使垂直磁记录的优越性得到充分利用。因而将垂直磁记录首先考虑应用于接触式记录装置如软盘和磁带。对13.3cm(5.25in)和0.9cm(3.5in)的软盘来说，采用垂直记录时，其位密度的第一步目标是达到50～100KBPI。这种装置的头盘组合方式有两种：一是环形头和单层介质；另一是单极头和双层介质。为了提高单极头读出灵敏度，设计成W型结构[1]，在主磁极两旁增加铁氧体辅助磁极。

垂直记录软盘介质主要有两类，一类是以Co-Cr为代表的金属膜；另一类是钡铁氧体粉末的涂胶盘。对金属膜软盘，头盘界面的耐久性和可靠性是严重问题，为解决这一问题已进行了许多研究[2]，其中包括有一些特殊的磁头滑块的设计[3]。另一方面，对钡铁氧体涂胶盘，由于它与常规的软盘片工艺几乎完全兼容，头盘界面也没有严重问题，适用于大量生产。日本东芝公司在1985年首先推出的0.9mm(3.5in)的垂直记录软盘非格式化容量为4Gb，其位密度为35KBPI，道密度为135TPI[4]。随后通过采用新的扇区伺服技术，已将容量提高到16Gb[5]，扇区伺服图形如图1所示。主要参数如下：双面总容量：16Gb(非格式化)；道密度：540TPI；位密度：35KBPI(MFM)；扇区数：64；传输速率：3GB/s；转速：901r/min；平均寻道时间：50ms；盘片介质：0.9cm(3.5in)，钡铁氧体(H_c = 60kA/m)。

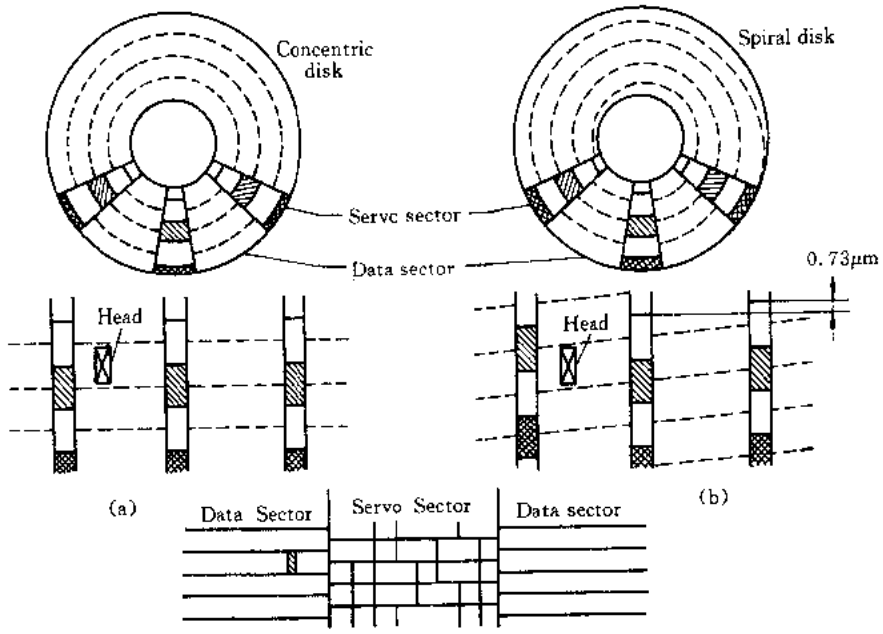


图 1　东芝公司 0.9cm,16Gb 钡铁氧体软盘扇区伺服图形[5]
(a) 同心道盘片；(b) 螺线道盘片

2.2　硬盘

硬盘垂直磁记录的研究不断取得进步[6~8]。采用单极头和双层金属膜的头盘组合系统，当飞行高度在 0.1 ~ 0.15μm 时，记录位密度 D_{50} = 40 ~ 50KFCI。垂直磁记录与纵向磁记录的记录密度与头盘间隙关系的比较见图 2[7]。美国的 Censtor 和 Northern Telecom 曾联合设计制造了供计算机系统应用的 20.3cm(8in)硬盘驱动器，其主要性能如下[9]：结构：20.3cm(8in)，全高，8 盘片，16 数据面；容量：2.26Tb(非格式化)；位密度：16.9KFCI，(2.7)码，25.3KBPI；道密度：2270TPI；道伺服系统：埋入式伺服，旋转式取数；数据传输率：3.0Gb/s；转速：3656r/min。

随着纵向磁记录硬盘驱动器的不断提高，这对垂直磁记录硬盘提出了严峻的挑战。现在普遍认为，要使硬盘垂直磁记录性能明显优于传统的纵向记录，其头盘间隙必须保持在 0.1μm 以下，甚至达 0.05μm。为此，在减小头盘间隙确保其耐久性方面已进行了大量的研究。这不仅需要减小滑块的尺寸，设计新的滑块，例如负压式滑块等，还要采用更有效的保护层和润滑剂[10]。不过，在改善头盘界面特性方面，不论是垂直磁记录或是纵向磁记录，其要求和目标几

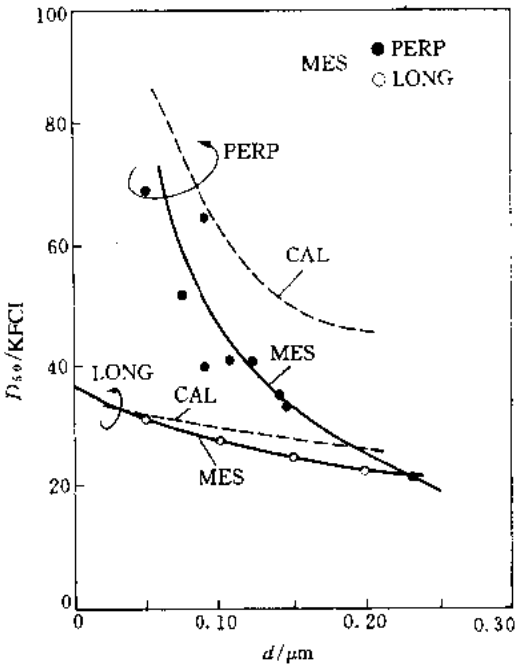


图 2　垂直磁记录与纵向磁记录的记录密度与头盘间隙关系的比较[7]

乎是完全一致的。

2.3 录音、录像和其他应用

垂直磁记录用于录音和录像已进行了大量的实验研究。结果表明，由于接触式垂直磁记录可以获得非常高的记录密度，因而在发展数字录音(DAT)和数字录像(DVT)应用上有很大竞争力。日本 NHK 的一项数字表明，采用窄道单极头，主磁极为 Fe-Si-Al-N 薄膜，宽 2.6μm，厚 0.23μm，CoCr/NiFe 双层膜作为记录介质，在记录密度为 50KFPI 下的标称输出幅值为 75($nV_{o-p}\mu m \cdot T \cdot m \cdot s$)，以及 $C/N = 51dB$，这时每位的记录面积小于 $1\mu m^2$。

由于 Co-Ni 蒸镀金属膜带(ME 带)已成功地应用于高频带 8mm 的 VTP 系统。这证明金属膜带现在已达到实用化要求。因而许多日本公司把其推向市场。三种不同方法真空蒸镀 Co-Cr 各向异性膜垂直磁记录特性的比较见图 3[12]。

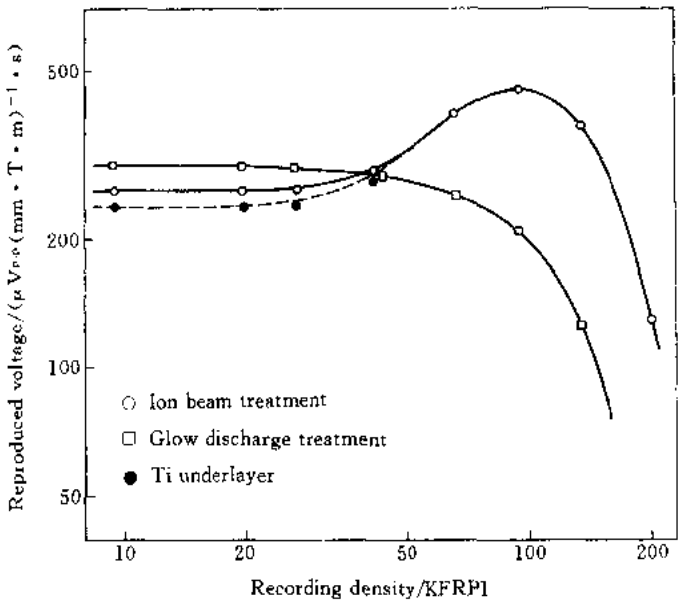


图 3 三种不同方法真空蒸镀 Co-Cr 各向异性膜垂直磁记录特性的比较[12]

采用钡铁氧体的涂胶带作为录音、录像和接触复制的实验也有大量的报道，由于这种记录介质的加工工艺及接触界面问题与普通磁带没有根本差别，有可能首先用于 DAT 系统的复制带[13,14]。不同矫顽力的钡铁氧体复制带的输出特性与复制时偏流的关系如图 4[14]所示。

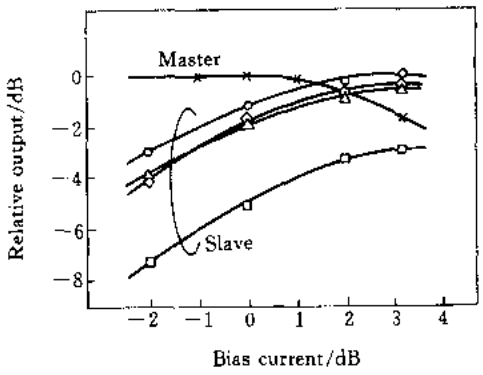


图 4 不同矫顽力的钡铁氧体复制带的输出特性与复制时偏流的关系[14]
$H_c/(kA \cdot m^{-1})$ ○:48.0; △:51.6; ◇:54.8; □:64.0

3 垂直记录磁头

3.1 结构

与水平记录不同，垂直记录的磁头最重要的特

性是能获得很强的垂直磁场分量和高的磁场梯度。虽然常规的环形头也具有磁场的垂直分量,因而至今仍有人利用环行头和单层介质系统来进行垂直记录,但由于其磁场的垂直分量和梯度都不够高,不能充分发挥垂直记录的特点。因而,近年来集中研究的是探极头或称单极头(probe type single pole type)。这种磁头有许多不同的形式。

最早是由日本东北大学研制的将主磁极与辅助极分别置于记录介质两面的称为辅助极驱动的单极头(APD-SPT),线圈是绕在辅助磁极上的。这种单极头成功地在实验室中用来进行软盘和磁带记录,但无法用于硬盘记录。这种分布在介质两面的磁头结构虽有不少改进,但根本的缺点是无法实用化。

后来发展的垂直记录磁头结构都趋向于将主磁极与辅助极结合成一体置于介质的同一面,用于软盘的垂直记录头主要有 Sony 公司开发的 W 型单极头(WSP)和东北大学开发的井字型头。用于硬盘的垂直记录头主要有 Censtor 公司的单极头(SPH)和与富士通公司开发的薄膜单极头 TFP。

此外,常规的环形头仍在作进一步改进,以提高其读写特性。与此同时,已开始利用 MR 头来读出垂直记录的信息,其特点不仅是灵敏度高,而且读出的信号与头盘间相对速度无关。

美国 Censtor 公司 1989 年在日本召开的第一次国际垂直磁记录会议(PMRC'89)上发表的温盘垂直记录磁头的结构,完全不同于 1985 年在美国明尼苏达 Intermag 会议上发表的。其结构示意图如图 5 如示。

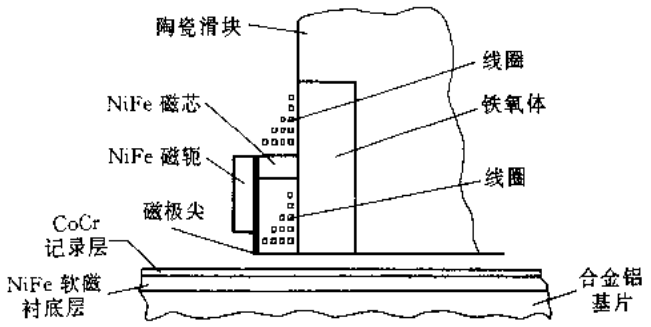


图 5 美国 Censtor 公司的温盘垂直记录磁头结构示意图[6]

磁头的滑块是陶瓷,嵌入铁氧体作磁通回路。主磁极材料为 NiFe,厚 0.7μm,道宽 8μm,线速为 25m/s 时的飞高为 0.15μm,读写线圈是四层,无中心抽头,共 50 圈。介质采用 CoCr 和高磁导率软磁双层膜,所获得的主要记录特性如下:孤立脉冲波幅值:700μV_{P-P};孤立脉冲宽度 PW50:1.65μm;Taa(20KFCI):400μV_{P-P};分辨率(20KFCI):75%;Taa(16KFCI):550μV_{P-P};分辨率(16KFCI):83%;Window Marging(10^{-10} BER):25%(2.7) Window on MTI;相邻道间串扰:< -50dB.

3.2 材料

对磁头性能影响最大的是磁极材料。垂直头设计的一个重要关键要求是得到分辨率和灵敏度(效率)的综合平衡。薄的磁极可以获得高分辨率,但不易得到较高的记录磁场,这是由于磁极的饱和使磁阻增大。因而主磁极材料不仅要求有高磁导率,还应具有高饱和磁通密度。

目前,实验用磁头有许多采用 CoZr 为基的非晶磁膜,其饱和磁通密度可大于 12T。其他在研究中的较好的磁极材料为铁基磁膜、钴基晶体磁膜和非晶磁膜,但至今这些材料尚存在不同

的缺点，例如热稳定性差，影响了它们的实用化。比较成功的几种材料是：加入少量的 Re 和 Pd，使具有高导磁导的 Co 基非晶磁膜的 B_s 增大。也可以减少 Zr 或 Hf 的含量而仍保持非晶态。另一途径采用 B_s 很高的富 Fe 合金，但其导磁率并不十分突出。例如 FeSi 膜加入 Re，其 B_s 可达 1.8T。日本东北大学建议的多层 SiFe 膜其 B_s 可超过 1.8T，并且其导磁率也可大于 1000。也有人发现，氧化铁的 B_s 可达 2.4T。但是，这些材料是否能与磁头的制作工艺相适配，报道尚不多。

日本 NHK 研究实验室采用 Fr-Si-Al-N 薄膜制作了窄磁道单极录像磁头，所获得的记录特性优于 Co-Zd-Nb 薄膜。这种薄膜的 M_s = 1080emu/cc，H_c = 2.4A/m，μ_1 = 2500，玻璃烧结温度不影响其磁性能。采用磁极宽度 2.6μm，厚度为 0.23μm，输出信号幅值 75($nV_{0-p}\mu m \cdot T \cdot m \cdot s$)。

滑块材料对磁头的性能，特别是其力学特性和使用寿命有重要意义。对垂直记录磁头来说，根据其结构不同，曾采用过不同的材料，如镍锌或锰锌铁氧体、陶瓷和玻璃，也有采用 TiC 为基板材料。由于盘片基板和记录介质的不同，考虑到其硬度、耐磨性、粘滞系数等的要求，不能指望用一种材料满足各种盘片的要求。对于 CoCr 记录介质的硬盘片来说，应用铁氧体一般来说是合宜的，不论是探极头，或是环形头，都可用陶瓷材料作滑块。

3.3 摩擦学(Tribology)

这主要是指磁头和介质之间的磨损、润滑和头盘撞损几率(俗称擦盘)的问题。从原则上讲，擦盘是很难完全消除的，只能是尽量减少其几率。因而，今年来不论是水平记录还是垂直记录，都对头盘摩擦学问题有大量研究。除了设计优化的滑块的几何形状及改善滑块轨道表面的研磨质量外，往往在介质表面附加保护层和润滑层。

用于 CoCr 或 CoCr/NiFe 介质的保护层，多数采用溅射淀积 C 膜，厚度为数 nm。也有采用溅射或 CVD 方法淀积 SiO_2 和 CoO 等膜。作为润滑层，可以采用液体或固体膜。但对温盘，普遍采用固态润滑膜，例如石墨、MoS_2、BN 和聚四氧乙烯等。一般润滑层较薄，不超过 1nm。

对垂直磁记录技术来说，只有当头盘问题愈来愈小时，才能充分显示出它的优越性。因而头盘摩擦学更为重要。现在许多制造厂努力的目标是要争取把头盘间隙减小到 0.05μm。最近，提出了一种新的观念，当间隙小到一定程度成为一种准接触状态。实际上，即使在软盘和磁带的情况下，也不能认为有完全的接触，即头盘间隙不可能绝对为零。东北大学发表了"接触式硬盘"的实验结果，引起许多学者的兴趣和不同的评论。

由于磁极材料和滑块材料的磨损速率往往是不同的，而一般情况下磁极材料更易磨损，结果造成主磁极凹陷的现象，等于使头盘间隙增大，这对高密度记录是十分不利的。因而选用主磁极材料与滑块覆盖材料的恰当组合是很重要的。

3.4 噪声

整个记录系统中的噪声主要来自三方面：电子电路、记录介质和磁头。电路噪声只有频率很低时才有实际影响。目前情况下，系统总噪声以介质噪声为主，双层介质中衬底是产生噪声的一个重要来源。但当采用狭磁道来增加记录面密度，以及数据传输速率愈来愈高时，磁头噪声最终将成为数字磁记录系统中的主要角色，因而现在对磁头噪声的研究已引起了极大重视。

日立研究所研究了不同匝数线圈的薄膜磁头的噪声和阻抗。采用 Nyquist 理论，对于线圈少的磁头可通过它的电阻来估计其噪声。为了提高信号的读出幅值需要将线圈圈数增加，这

时磁头的电感就不能忽略。在计算中必须计及磁头阻抗的频率特性，从计算和实验数据的一致性可认为磁头的噪声由读出电路阻抗的实部所决定，这表明磁头的噪声主要是由磁极的损耗所引起的。

4 垂直记录介质

目前垂直磁记录介质的研究仍然是世界各国十分关注的一个领域，研究的方向大致可分为两个方面：一是 Co-Cr 记录介质性能的进一步改善，主要集中在进一步了解薄膜微观结构与制备工艺和磁性能的关系；二是研究新的记录介质材料，新材料的研究以钡铁氧体作为新的垂直记录介质已引起人们的热切关注。如 TOSHIBA 公司已研制出以钡铁氧体作为记录介质的 4MB 垂直记录软磁盘驱动器，正在研究 16MB 的软磁盘驱动器。

4.1 薄膜成分、微观结构

应用 TEM 等手段，研究 Co-Cr 膜磁畴时看到，磁畴的结构随膜厚和成分及磁化过程而变。膜厚 20nm 时，磁畴花样为羽毛状；10nm 厚时，磁畴的花样为封闭的羽毛状；100nm 厚时，磁畴为圆点花样。当 Cr 的含量发生改变时，磁畴的形状也发生改变，4.9%Cr 为 Block 花样；19.8% Cr 为波纹花样；21.8%Cr 为交叉结花样(Cross-tie wall)。同时指出，不同的磁畴花样均由 c 轴位向引起的。应用化学浸蚀的方法，对 Co-18%Cr 的 CoCr 介质的研究发现，RF 溅射制备的 Co-Cr 膜内晶粒内有成分偏析，晶粒内有一富 Cr 的核，核的周围有一富 Co 的环，这种成分偏析与相分离原理有关，这种偏析主要发生在膜的长大过程。有文章指出，这种横向偏析将强烈地影响记录介质的矫顽力和各向异性。主要受沉积过程中基板温度的影响，通过控制基板温度，改变基板和采用 Ge 层(软盘)作为衬底层来加以改善。基板温度和 Ar 气压力对 Co-Cr 膜柱状形态也有较大的影响，当基板温度小于 100℃，Ar 气压力小于 0.26Pa 时，Co-Cr 膜有良好的结晶体和致密的柱状形态，然而当基板温度大于 100℃，Ar 气压力为 0.16Pa 时，薄膜中将出现柱状的粒子；当 Ar 气压力降低到 13.3mPa 时，将看不到柱状形态。并已证实，化学成分的偏析和垂直各向异性与柱状结构的形态无关。

为了提高 Co-Cr 膜的抗腐蚀性能，在 Co-Cr 膜中加入 Mo 的研究表明：CoCr-2.9%Mo 有最大的垂直矫顽力和磁饱和强度，$H_{cl}=115.7$kA/m，$M_r/M_s=0.3$，$\Delta\theta=50°\sim8.6°$，Mo 的加入降低了 Cr 的偏聚和改善了膜表面的钝化，从而改善了 Co-Cr 膜的抗腐蚀性能，从 CoCrMo 腐蚀特性看，腐蚀过程仍为一个电化学腐蚀过程，同时，旋转基板比静止基板有更好的 c 轴取向。

4.2 衬底层

在垂直磁记录介质中，通常采用 Ni-Fe 层作为衬底(underlayer)。研究表明，在数据读出过程中，衬底层会产生高幅度的噪声脉冲，对于一般的硬盘驱动器(速度 20m/s，飞行高度 150nm)，衬底层的噪声位于较低的频率中，可以通过利用降低整个频带宽度 5%～10%的滤波器加以解决。然而对于厚的高导磁率(2μm)的衬底层，还存在一个与大脉冲噪声有关的擦去效应，不能通过电子线路加以解决，这个问题的解决方法是一个折中的方案——即适当地擦去数据或者应用低效率的磁头。同时已证明，衬底层的噪声不是巴克豪森跳跃所致，而是由磁头的杂散磁通引起的。有人指出，当 NiFe 膜厚大于 2μm 时，会出现一个大的噪声峰，增加了介质

的噪声，该噪声峰是由沿磁盘半径方向上衬底层中的条纹磁畴引起的。对于厚的 NiFe 膜，控制磁畴结构是降低介质噪声的重要方面。

为了得到高饱和磁化强度、低矫顽力的 NiFe 膜，制备 Ni-Fe 膜的方法应加以改善。在 Ni-Fe 靶平面上增加一 4kA/m 的垂直磁场，其作用可限制 v 电子和在靶邻近区域产生高电离区。对于 0.5μm 厚的 Ni-Fe 膜获得饱和磁化强度 9.5kG，矫顽力 40A/m。此外，外磁场的应用，也大大增加了靶面积的利用率。当 H_e 由 0 增加到 8kA/m 时，靶面积的利用率从 50% 增加到 98%。在溅射 Ni-Fe 膜时，应适当控制氧的分压，否则将使溅射功率增加。H_2O 和 O^+ 峰增加，O_2 分压的增加，H_c 逐渐增加，所以应将氧分压控制在小于 0.42mPa，可以得到矫顽力在 8 ~ 240A/m。也有文章指出，以 Ti 为基，加其他元素可形成 Ti 合金的衬底层。如加入 Cr、Ta、Nb、Co、Pt、Pd 等元素。使 Co-Cr 膜的 c 轴取向更好地垂直于膜面。如 10% Cr + 18% Ta + 18% Pt + Ti 作为 Co-Cr 膜的衬底层。$\Delta\theta_{50} = 5°$，$M_r/M_s = 0.1$，$H_{cl} = 70$kA/m。

4.3 新材料和新工艺

从目前研究的结果看，Co-Cr 记录介质具有良好的磁学性能。但是其机械性能，如磨擦特性、成分偏聚引起的抗腐蚀性、挠性(flexible)不甚理想。溅射制备薄膜方法，虽然仍是目前制备薄膜的主要手段，但是其沉积速率太低，尤其于大批量生产的工业应用中，使成本增加。据此，人们开始寻找新的材料和新的工艺来代替或改善垂直磁记录介质和特性。

新的垂直记录介质有钡铁氧体、Co-CoO 及 Fe-Ti 等。钡铁氧体介质的研究是采用不同的方法降低钡铁氧体的矫顽力，如用 Mn-Zn 层与钡铁氧体构成双层膜，通过控制 Mn-Zn 层的厚度来调节钡铁氧体的矫顽力，当两者的比例为 5:2 时，其矫顽力可以在 0.16 ~ 4.8kA/m 进行调节，当钡铁氧体为 2.5nm，Mn-Zn 膜厚为 1.0nm 时，$M_s = 450$emu/cc，$H_{cl} = 4.8$kA/m，是一适合于垂直磁记录介质的材料，有数据表明：$D_{50} \approx 88$KFRPI，$SN = 40$dB，此外，该材料制成的介质，其摩擦性能优于 Co-Cr 膜。

利用相对靶溅射(facing-targets-sputtering)制得成分为 18% ~ 20%Ti 的 Fe-Ti 膜也有良好的磁学性能，垂直各向异性 $H_k = 240$kA/m，饱和磁化强度 600emu/cc，垂直矫顽力 $H_{cl} = 7.2$kA/m。

由于 Co-Cr 膜中存在 Cr 的偏聚，使 Co-Cr 膜的微观成分不均匀，从而影响其磁性能。有些文章指出：利用氧与 Co 发生局部反应形成部分 CoO 来取代 C_l 构成 Co-CoO 的记录介质。当 $m(O_2): m(N_2) = 10:90$ 时，得到 $H_k = 300$kA/m，$H_{cl} = 89.6$kA/m，$M_s = 570$emu/cc，$D_{50} = 105$KFRPI。此外，热处理也可以进一步改善其读/写性能。

在众多的文献中均提及制备薄膜的方法采用相对靶溅射。其工作原理如下：在溅射过程中，等离子体区产生大量的金属离子，在基板电位位于地电位时，将不能直接沉积在基板上，当基板电位由零电位降低时，通过栅极将离子由等离子体抽出沉积在基板上，基板电位越负，沉积速率越大，当 V_{sub} 电位达到 −60V 时，最大的沉积速率可达 0.8nm/min。该系统的优点是，在极低的 Ar 气压力下(0.133Pa)条件下，得到很高的沉积速率，薄膜的晶体结构和磁性能也随基板电位不同而改变，但是目前尚无确切的解释。

磁盘表面结构(texture)的制备对磁盘性能有很大的影响。目前采用的方法均为机械结构法。有人提出一种新的方法——化学织构(chemical texturing)，制备方法如下：将已抛光的 Al-Mg 合金基板浸在 3% 草酸中，形成 10μm 厚的阳极氧化膜，膜中的气孔尺寸和细胞尺寸(cell size)分别为 30 ~ 40nm 和 100 ~ 120nm，用干净的空气吹干，再用合适的有机脂(如聚酰亚胺、环

氧、有机硅等)涂布在氧化膜的表面,再将盘片放入炉内加热固化涂层,再将盘片抛光,最后将盘片浸入 35ml/L 磷酸和 20g/L 铬酸溶液中溶解表面氧化膜。这样,盘片表面就形成由沉积金属或有机脂组成的 10～60nm 厚的表面凸出部,用化学结构法制备的盘片沉积法制备的盘片沉积 CoCr/FeNi 记录介质,$D_{50} \approx 100$KBPPI。

5 结语

现有的垂直磁记录头和介质两种主要组合的特性已充分显示了比纵向记录有明显提高的记录密度。在实用化方面,钡铁氧体介质用于垂直磁记录软盘和磁带已不成问题,在金属膜介质方面,头和介质的界面问题仍是最引起注意的。但随着进一步的研究和纵向磁记录中金属膜小型硬盘及录音和录像带应用的实际经验的积累,界面问题不是不能解决的。可以期望在今后的高密度磁记录系统中,首先是在接触式记录系统中,垂直磁记录将会占重要地位。

参 考 文 献

[1] J. Hokkyo, K. Hayakawa, I. Saito and K. Shirane. IEEE Trans. Magn., 1984, MAG-20:72

[2] S. Kadokura, K. Kamei, K. Teranishi and S. Sobajima. IEEE Trans. Magn., 1987, MAG-23:2404

[3] A. Mishima, K. Hayakawa and J. Hokkyo. Proceeding of PMRC'89, 1989, 29P:28

[4] M. Imamura, T. Ito, M. Fujiki and H. Kubota. Toshiba Review, 1985, 40:1115

[5] T. Yamada, Y. Sakai, T. Muraoka and T. Suhaya. IEEE Trans. Magn., 1987, Mag-23:2680

[6] R. Tsui, H. Hamrilton, R. anderson, C. Baldwin, P. Simon. Digests of the Intermag'85, 1985, GA-4

[7] S. Tanabe and T. Ozeki. Digests of the Intemag'87, 1987, AB-02

[8] Y. X. Chen, J. L. Zhang, X. L. Zhao and W. L. Tang. Journal of the Magnetics Society of Japan, 1989, 13(SI):343

[9] E. Katz and P. Schreiber. Journal of the Magnetics Society of Japan, vol. 13, 1989, 13(SI):253

[10] T. Miyamoto, I. Sato and Y. Ando. IEEE Trans. Magn., 1987, MAG-23:2386

[11] N. Hayashi, S. Yosahimura and J. Numazawa. Journal of the Magnetics Society of Japan, 1989, 13(Supplement. S1): 545

[12] R. Sugita, K. Tohma, et al. IEEE Trans. Magn., 1989, MAG-25:4183

[13] Y. Okazaki, M. Noda, K. Hara, K. Ogisu. IEEE Trans. Magn., 1989, MAG-25:4057

[14] T. Suzuki, T. Ito, M. Isshiki, N. Saito. IEEE Trans. Magn., 1989, MAG-25:4060

A New Concept of Magnetic Thin-Film Heads with Superconductors

1 Introduction

It is generally understood that in order to increase the density of magnetic recording one of the key points is the optimum design of the magnetic heads. The requirements of these heads are not only very high resolution but also very good sensitivity. Many new schemes of recording heads have been proposed to match the requirements of wider bandwidths and higher recording densities[1,2].

A tremendous breakthrough has been made in high T_c superconductor materials during the last two years[3,4]. It strongly encourages scientists to search for some new superconductor materials which appear to have a higher T_c and can even be operated at room temperature. On the other hand, this situation also promotes the applications of the superconductor to different technologies. In this paper, we consider the use of the perfect diamagnetism of a superconductor film inside or outside two magnetic poles of an inductive thin-film ring head to confine the leakage and divergence of magnetic flux. The finite-element method[5] is used to calculate the field distribution, reproducing pulse and other characteristics for the model of the recording heads which contain superconductor films in several different structures. The calculated results show that such new thin-film heads have some obvious advantages. The experiments of a largescale model of this recording head will be reported in another paper.

2 Models and Calculating Methods

The models of thin-film heads with a superconductor calculated in this paper are shown in Fig. 1, with (a) superconductors only in the gap of poles, (b) superconductors inside the gap and outside two poles, (c) multilayer structure (nonsuperconductor-superconductor-nonsupercon-ductor) in the gap, and (d) a superconductor layer under the recording medium. The dimensions of the models are shown in the same figure. The coil of the head is 16 turns, and the current is 20 mA. The permeability μ of the magnetic poles is 1000 and permeabilities of the superconductor are 0.6, 0.2, and 0.01 in different cases Fig. 1.

For finite-element method calculation a longitudinal section of the head is divided into meshes. In order to get high accuracy, the meshes at the top of the poles are much finer than the other parts. There are a total of 2336 nodes and 2118 elements in our meshes. We input the data of node coordinates and elements, and then the field distribution can be obtained.

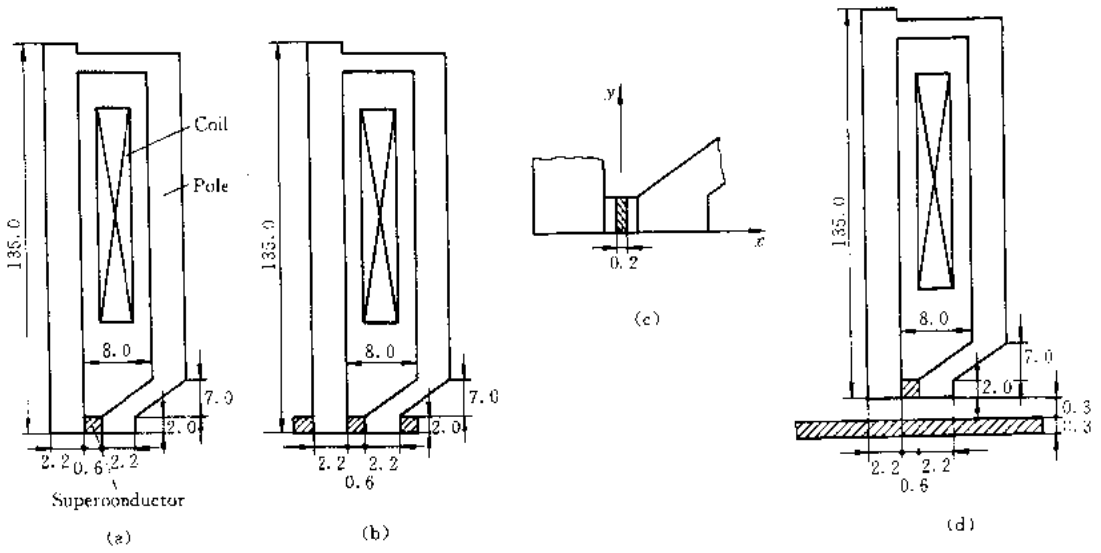


Fig. 1 Structure and size of various head models, (a) using diamagnet material in the gap, (b) using diamagnet material both in the gap and outside of two magnetic poles, (c) using a multilayer of nondiamagnet material and diamagnet material in the gap, and (d) using diamagnet material as the backlayer of the recording medium

By using the reciprocity theorem, we also calculated the isolated pulse output from the magnetic field distribution resulted from the FEM. Assuming the y component of remanence $M_y = 0$, and the x component remanence M_x is constant along the thickness of magnetic recording medium, we have the formula for isolated pulse output:

$$e(x) = \frac{N\mu vW}{i} \int_{-\infty}^{+\infty} \int_{d_m}^{d_m + t_m} H_x(x) \frac{\partial M_x(x - \bar{x})}{\partial x} \mathrm{d}y\mathrm{d}x$$

where $M_x(x) = 2/\pi M_r \arctan \pi x/a$, a is the magnetic transition length parameter which is determined by coercive H_c, remanence M_r, and thickness t_m of the magnetic recording medium. N is the turns of the head coil, μ is the permeability of the magnetic recording medium, v the head medium relative velocity, W the track width, and i the current. All the parameters concerned in the calculation of the paper are determined in accordance with the thin-film ring heads and magnetic recording medium available in our laboratory. The $\partial M_x(x)/\partial x$ which resulted from the assumed function is an even function, and the $H_x(x)$ calculated using FEM is discrete, so the integral can be regarded as the discrete conclusion of $H_x(x)$ and $\partial M_x(x)/\partial x$, therefore, the fast Fourier transform method can be used.

3 Calculated Results

We calculated the longitudinal field components (H_x) in the distance to the pole tip $y = 0.0375 \sim 0.45\mu$m when the permeability of the medium material in the thin-film head gap is 1.0, 0.6, 0.2, and 0.01, respectively. Fig. 2 shows the distributions of H_x in the x direction with the permeability of the gap materials as a parameter as $y = 0.0375\mu$m. The amplitude of H_x obviously increases and the distribution plot of H_x in the x direction becomes sharp as the μ reduces. The relations of the field gradient ($\mathrm{d}H_x/\mathrm{d}x$) to permeability with y as a parameter are shown in Fig. 3. The field gradient decreases

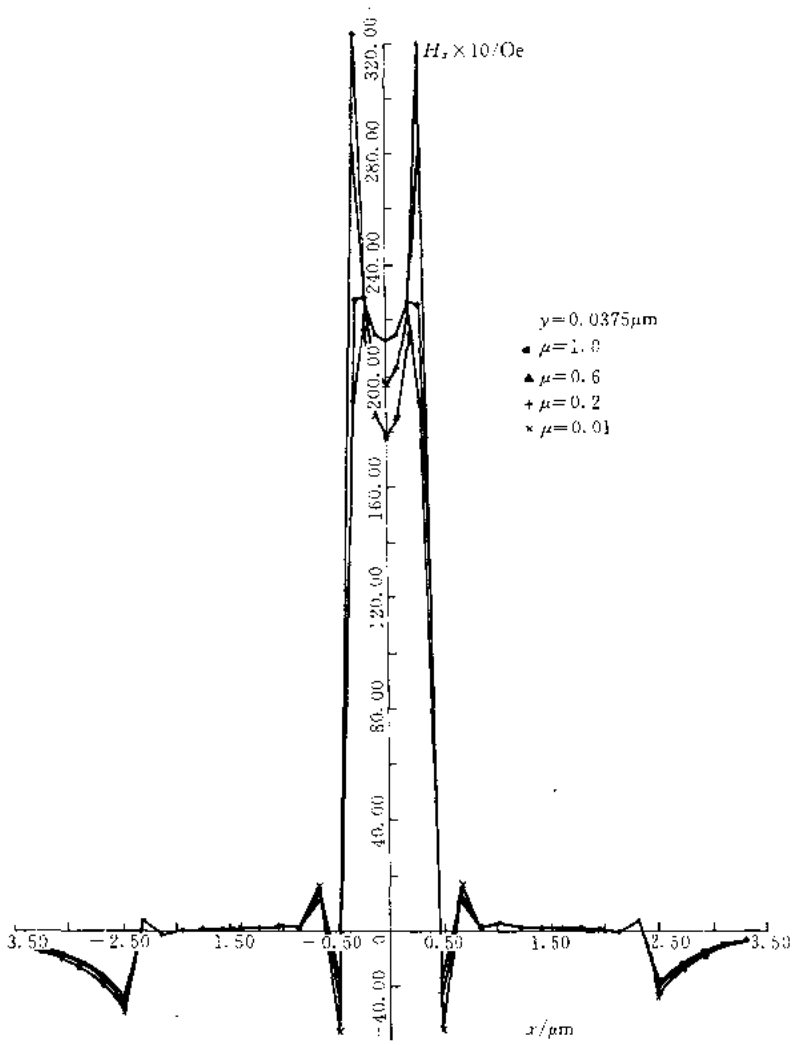


Fig. 2 Distribution of longitudinal field component H_x in the x direction with μ as a parameter ($y = 0.0375\mu m$)

sharply with the permeability increasing and the plot of the field gradient versus the permeability becomes flat when the y increases. Fig. 4 shows the relation of the peak value of longitudinal field component (H_{xp}) to the distance y with the permeability μ as a parameter. Obviously, the peak value of the longitudinal field component using the diamagnet material in the gap is larger than the peak value without that. The H_{xp} decreases while the y increases. The difference between $H_{xp}(\mu = 0.2)$ and $H_{xp}(\mu = 1)$ is apparent near the pole tip, and it becomes small at a distance far away from the pole tip. Therefore, using the diamagnet materials in the gap gives a significant improvement only

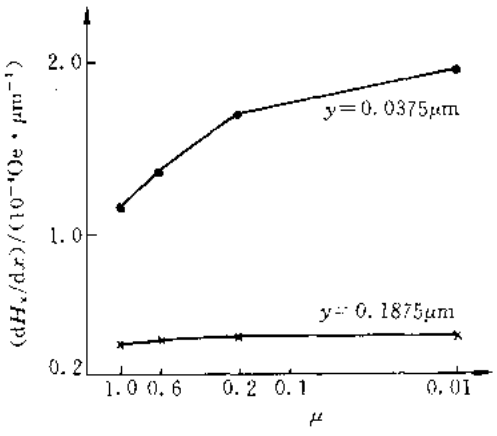


Fig. 3 Relations of dH_x/dx to μ with y as a parameter

when the permeability in the gap is low and when the position to the pole tip is very close to the recording head.

Fig.5 shows the relations H_{xp} to throat height h with the permeability μ as a parameter when y is 0.1875μm. For $\mu = 1$, H_{xp} obviously decreases while the h increases, but there is no significant change in H_{xp} for $\mu = 0.2$. Hence, the influence of the throat height on the field amplitude is reduced by using a diamagnet material in the gap. It indicates that by using a diamagnet material in the gap, the recording sensitivity for high-density recording can be improved, and the yield will increase if the throat height control is not as critical as usual.

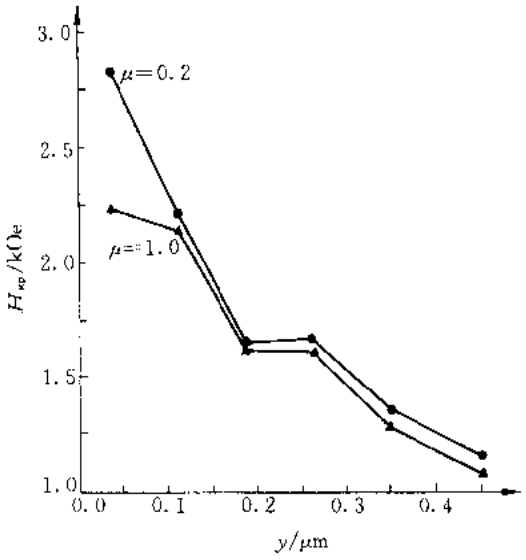


Fig.4 Relations of H_{xp} to y with μ as a parameter

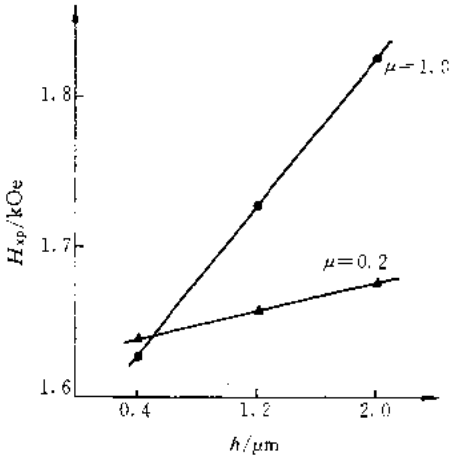


Fig.5 Relation of H_{xp} to throat height h with μ as a parameter

Fig.6 shows the relations of isolated pulse versus various structures when the distance between the pole tip and recording medium D is 0.225μm. The amplitude of isolated output pulse increases by using a diamagnet material as the backlayer of the recording medium. With a multilayer of nondiamagnet material and diamagnet material in the gap the amplitude of the isolated pulse increases a little and is less than the amplitude of an isolated pole using a single-layer diamagnet material in the gap. However, the negative peak of field distribution can be reduced.

Calculated results show that there is no obvious improvement in the recording characteristics by using a diamagnet material on the outside of two magnetic poles.

4 Conclusion

(1) When the superconductor film as a perfect diamagnet is inserted between the magnetic poles of a recording head, the field gradient dH_x/dx at the top of poles will increase, and larger amplitudes of the reproducing poles will be expected as well. However, these effects will reduce very rapidly with distance from the poles.

(2) The dependence of a reproducing signal on the throat height of the poles becomes less critical

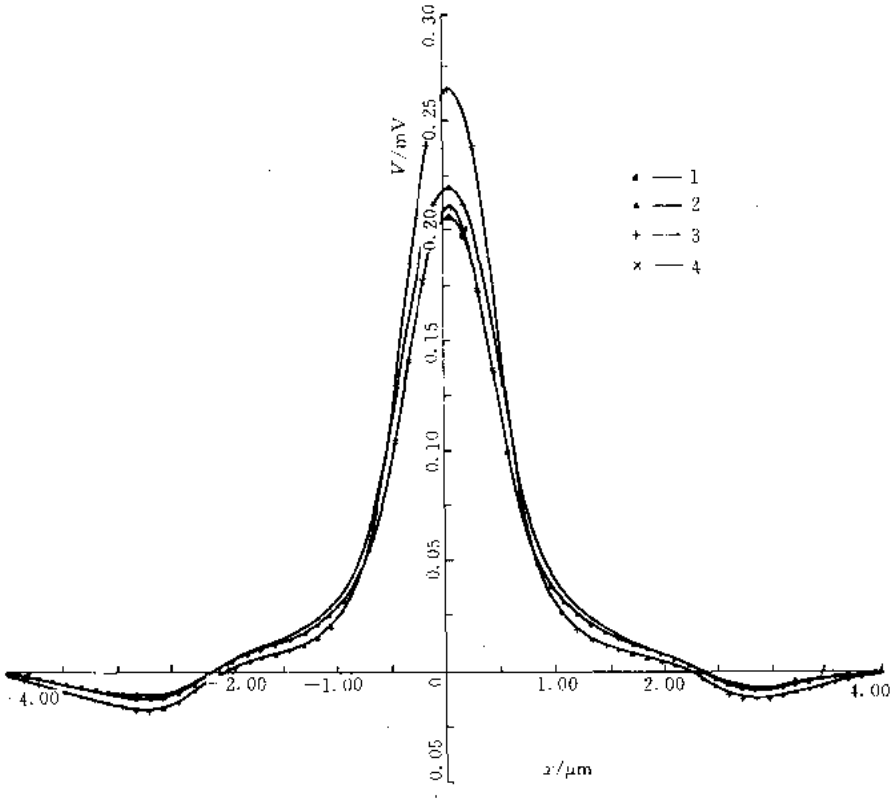


Fig.6 Isolated pulses vs various situations: $D = 0.225\mu m$. (1) $\mu = 1$; (2) $\mu = 0.2$; (3) $\mu = 0.2$ and with $\mu = 0.2$ backlayer; (4) $\mu = 0.2$ multilayer

when the gap is filled with a superconductor film.

(3) There are no obvious improvements on the head performances if the superconductor film is put outside the two poles.

(4) The reproducing amplitude increases when there is a superconductor film under the recording medium.

(5) The bigger the permeability of diamagnets and the more the flying height decreases, the more the performance increases.

References

[1] R. E. Jones Jr., C. D. Mee. Magnetic Recording, Volume 1:Technology. New York: McGraw-Hill, 1987, Chap.4

[2] A. Eiling. J. Appl. Phys., 1987, 62:2163

[3] B. C. Liao, et al. Chin. J. Low Temp. Phys., 1987, 9:183

[4] W. J. Yie, et al. Chin. J. Low Temp. Phys., 1987, 9:279

[5] P. Silvester, M. V. K. Chari. IEEE Trans. Power Apparat. Syst., 1970, PAS-89:1642

[6] R. I. Potter. J. Appl. Phys., 1970, 41:1647

Characterization of the Film Heads for Perpendicular Magnetic Recording

1 Introduction

Because of the physical limitation the longitudinal magnetic recording can no longer meets the demands of the tremandously increasing of the significant approaches to continuonsly enhance the performences of the magnetic recording hereafter. A lot of research works in the laboratories have demonstrated that the perpendicular recording presented much potential in the past years[1,2]. It is greatly expected that the perpendicular recording can be commercialized in the near future. A new thin film head which can be practically used in a perpendicular recording rigid disk drive is reported in this paper. This head has the same slider, suspension and flying height as the conventional longitudinal head and its output signal waveform is quite similar with that of the longitudinal one as well. Therefor, such head appears good compatibility with existed drive systems. We have put such heads togather with perpendicular magnetic recording disks into a commercial product Maxtor-4380 drive instead of the longitudinal heads and disks. This exchanged drive can be operated with a personal computer normally. There is no doubt that the drives will be much better if it can be newly designed to fit the optimum conditions of perpendicular recording heads and disks.

2 Structure and Fabrication

Two schemes of thin film recording head with single pole type are designed as shown in Fig. 1(a)

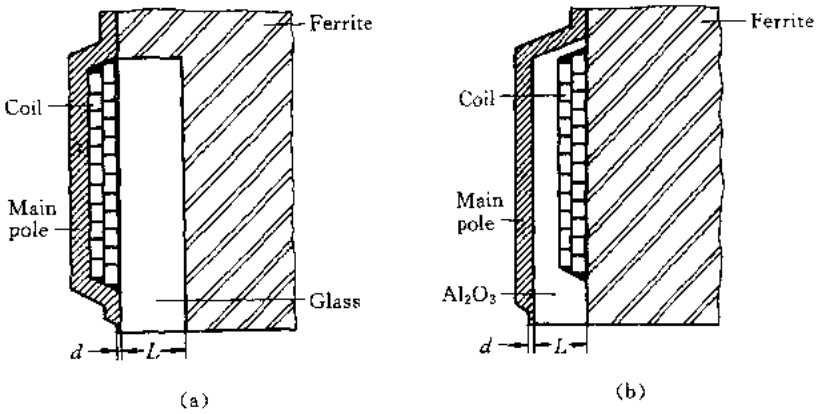


Fig. 1 The structure of the recording heads

and (b). Both the two heads structures are using the ferrite substrate as an auxiliary pole and the soft magnetic alloy film as a main pole or probe. The isolation between the probe and auxiliary pole is non-magnetic medium which in the structure (a) is glass forming by grooving the substrate first and then to fill it up with glass fiber sintered. It is Al_2O_3 in the structure (b) which is formed by sputtering it above the two layers of coil. The advantage of the structure (b) is that the geometry and the magnetic properties of the pole tip are easily to be controlled, because here the main pole is on the flate surface of the Al_2O_3 layer. We designed two different copper coils with 16 and 23 turns repectively. The insulating layers of the coils are formed by high temperature solidfication of the AZ photoresistor. Thethickness of the magnetic pole is 2 to 3 μm, the width of the pole tip is 16 μm, its thickness is 0.4 μm and height is about 2 μm.

3 Results of the Experiments

The resistance, inductance and resonance frequency of the heads are measured by the HP4194A Impedence Gain Phase Analyzer. The typical value of those are 16.5 , 280nH, and > 50MHz repectively. The normal flying height is 0.15 μm, which is measure by using the MeGill 50Z fly height tester. The flying status are 0.10 μm for pitch and 0.02 μm for roll. The read-write dynamic characteristics are measured by using the Adelphi RD-009E tester and the disk in CoCr/NiFe double-layers structure with a coercivity of 67.2kA/m. The results are low frequency output LFTAA 0.912mV, high frequency output HFTAA 0.768mV, Half-height pulse width $PW_{-50} = 93.75$ns, overwrite O/W − 37dB, resolution RES 91.3 and signal to noise ratio SNR 30.6dB, which $LF = 1.25$MHz, $HF = 3.33$MHz and the write current 27.5 mA. The frequency characteristics of the signal output at the inner track is shown in Fig.2. From which 9.4 MHz can be obtained for the D_{50}, it is equal to 41.6 KFCI.

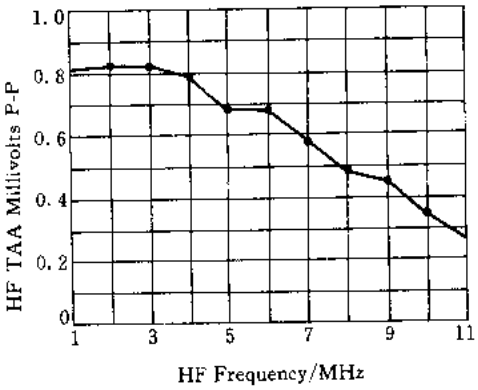


Fig.2 The frequency charateristics of the output signal

References

[1] R. Tsui, H. Hamilton, el al. Digest of the Intermag Conference, GA-4, 1985

[2] J. Toda, T. Yamamoto, H. Iwama, K. Kobayashi. IEEE Translation Journal on Magnetics in Japan, 1985, TJMJ-1 (4):450

The Thin Film Magnetic Recording

1 Introduction

Now the magnetic recording is the main technology in the information storage area facing the strong competition of the other technologies in this area such as the semiconductor storage and optical storage, it still keeps its dominant position whatever as the peripherals of the computers or as the consumer products used in video and audio systems owing to their excellent recording characteristics, flexibility, cheap price and great potentiality of performance.

The two elementary parts in every sorts of magnetic recording installations, are magnetic recording heads and magnetic recording media. The magnetic recording media could be classified to the magnetic tapes, floppy disks and hard disks according to their different uses. Correspondingly, there are several types of magnetic heads to match the different formats of media of their signal writing on the media and their signal reading from the media. The longitudinal recording and the perendicular recording are the two main recording machanisms based on the magnetized directions of the magnetic media while they are recording. Since the demagnetizing is with less influence in the perpendicular recording than it in longitudinal recording while recording in short waves, the recording density could be highly increased. So the perpendicular recording is the important option in the development of the magnetic recording technologies. Generally, the perpendicular recording is usually adopted in the magneto-optical recording.

It is a very effective way to use the thin film technology in making the magnetic heads and the magnetic media, which brings in the better recording function and higher recording density. Now the thin film format is successfully introduced in the making both of the magnetic heads and the magnetic media no matter they are longitudinal recording, perpendicular recording or magneto-optical recording.

This article is the summary based on the research in Shanghai Jiaotong University of the thin film technology used in magnetic recording heads and magnetic recording media.

2 Thin Films for Recording Medium

2.1 The property regirements for the magnetic recording media

Different from particle media, thin film recording media have higher coercivity H_c and higher recording density and output due to the fact that they are totally magnetic materials. As early as 1960's, thin film media used for magnetic recording were investigated, but they did not be used as the hard disk com-

mercially until the early 1980's, and they were earlist, majorly used media.

According to magnetic recording theory, the half-width W_{50} of the signal pulse in digital recording is

$$W_{50} = 2\left[\left(d + \frac{a}{\pi}\right)^2 + \left(\frac{g}{2}\right)^2\right]^{\frac{1}{2}} \tag{1}$$

where d is the height between head and medium, g is the gap of ring head, a is magnetization distribution constant of the media, which is decided by medium thickness t_m, residual magnetization B_r and coercivity H_c, a usually can be expressed as follows:

$$a = 0.6 t_m \frac{B_r}{H_c} \tag{2}$$

From the above formulas, it can be shown that in order to increase the recording density, W_{50} must be reduced, which means that the gap g of head and the height d between head and medium (it is called flying height in hard disc recording) must be reduced. From the point view of media, we should increase the coercivity H_c and reduce the film thickness t_m, the reductin of residual magnetization B_r could increase recording density, but this results in the reduction the amplitud of signal output. In order to ensure the required S/N ratio, B_r should be a medium value.

2.2 Co-NiCr longitudinal recording media

The longitudinal recording medium of Co-30%Ni-7.5%Cr alloy film has been studied[1]. The experiments show that when the films sputtered on the substrates coated with Ni-P, CoNiCr crystal with hexagonal-close-packed lattice structure is produced, it grows along the close-packed plane (0001), the easy axis [0001] of recording layer is nearly perpendicular to disk surface, that means, the hard axis is nearly parallel to the disk surface. In this case, the coercivity of medium is relatively low, the B-H hysteresis loop of the medium is shown on Fig.1(a). When Cr is used as an underlayer, CoNiCr/Cr double layer film is formed. The crystal of Cr layer is bbc structure, it grows along the close-packed plane (110). Because of the atomic area densities of the plane (1011) of CoNiCr and the plane (110) of Cr are very close. According to the energy minimum principle, the most probable arrange is that the plane (1011) of CoNiCr grows epitaxially on the plane (110) of Cr. CoNiCr film is also in the disk surface. This greatly increases the longitudinal coercivity, as shown in Fig.1 (b). As a rersult, the texture structure of the erystal is improved and the coercivity of double layer film increases with the increasing of Cr layer depth, as shown in Fig.2.

The corrosion of the film is a very important factor that affects its applications. The accelerating degradation expriment under the conditions of relative humidity $R.H.$ 88% and temperature 88℃ shows that[2], within 50 hours, the saturation magnetization of CoNiCr/Cr double layer thin film is almost without change, however under the same conditions, the properties of CoNi/Cr double layer degenerate rapidly as shown in Fig.3. The main reason of degeneration

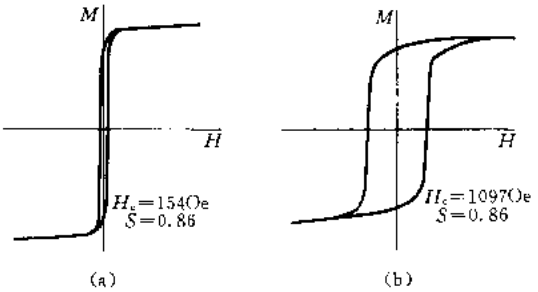


Fig.1 In-plane M-H hysteresis loops of
(a) CoNiCr single layer and
(b) CoNiCr/Cr double layer this films

of magnetic properties is chemical corrosion. When corrosion occurs, the Co atoms in the film continuously diffuse to the surface and get oxidized, this procedure changes the chemical composition of the film, thus results in the reduction of magnetic properties. But when Cr layeris added, it can effectively impede the inter-diffusion of Co and O atoms and the oxidization of Co. In such way, the degeneration of the magnetic properties can be delayed.

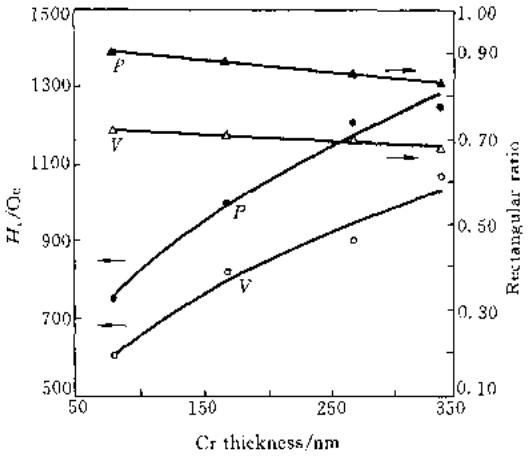


Fig.2 Effect of the thickness of Cr underlayer on the magnetic properties of CoNiCr/Cr double layer films

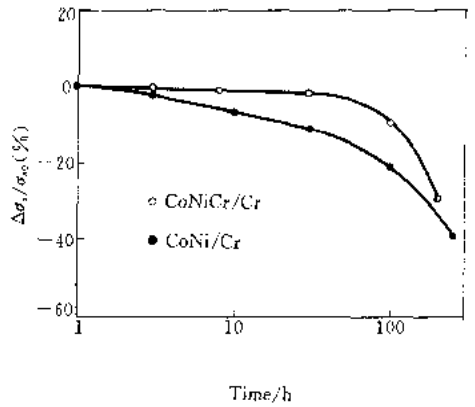


Fig.3 The corrosion behavior of CoNiCr/Cr and CoNi/Cr films. Conditions: 88℃ and R. H. 88%

2.3 CoCr perpendicular magnetic reconding media

As the perpendicular recording media, the thin films are expected to have an anisotropy field that is perpendicular to the film as high aspossible. The thin films of $Co_{80}Cr_{20}$ have been investigated in our laboratory[3]. The anisotropy field of these films mainly come from their hexagonal structure. The *c* axes of the grains in CoCr film orient along the vertical direction of the surface, this makes it possible to be used for perpendicular recording. By choosing the optimum sputtering conditions, the CoCr films with perpendicular anisotropy can be deposited. X-ray diffraction is used to characterize the texture of the film and the spread in the *c*-axis ($\Delta\theta_{50}$). The experiments show that $\Delta\theta_{50}$ from the rocking curve can be as small as 2.1°, as shown in Fig. 4. The anisotropy field H_k of the films with thickness about 0.3 ~ 0.4 μm are 400 ~ 480A/m. It indicates that the CoCr films have very good perpendicular orientation growth characteristics. To form a magnetic flux close circuit, a soft magnetic thin film with high permeability is required to be an underlayer of the recording medium. In our case, the sputtered NiFe film is used to form CoCr/NiFe double layer medium for perpendicular recording.

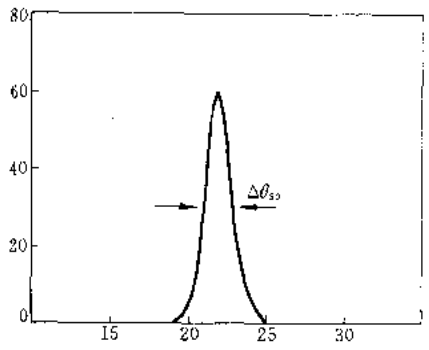


Fig.4 Texture characteristics of the CoCr film, X-ray diffraction spread in the *c*-axis ($\Delta\theta_{50}$)

3 Magnetic Recording Heads

3.1 Property requirement for the magnetic thin films

Thin film magnetic recording heads are construeted by multilayer of thin fierns. Besides magnetic films used for head poles, there are conductor films used for coils and other films used for insulation, gap and overcoat. The properties of these films, are sensitively and directly related to the producing methods. Among these, the properties of magnetic films are most important.

The basic requirements for the magnetic films are as follows:

(1) Permeability μ. It is one of the critical parameters that affect write and read characteristics. It is very sensitive to purity, thermal history, cold mechanical processing and wear environment.

(2) Saturation magnetic induction B_s. It determines the maximum magnetic flux density which can be generated in the head poles. This is very important for the writing function.

(3) Coercivity H_c. It is the magnetic field necessary to reverse the magnetization and decrease the magnetic induction to zero. It illustrates the ease of changing the magnetization and determines the minimum output of the reading signal.

(4) The magnetostriction coefficient λ_s relates the magnitude and sign of strain, along a given crystalline direction or magnetic anistropy axis, to a change in magnetic propertics. The value of λ_s may be positive or negative. As magnetic film to be used as poles, λ_s is required as small as possible, zero magnetostriction materials are generally preferred. The reduction of λ_s in materials results in the reduction of noise.

(5) For the magnetoresistance read heads, magnetoresistance coefficient $\Delta\rho/\rho$ is an important parameter, it represents the fractional change of resistance associated with changing magnetization from parallel to perpendicular to current flow.

According to the above requirements, the materials for thin film head poles by using of permalloy film have been for long time, which composition is 80 wt% Ni and 20 wt% Fe. NiFe film can be produced by sputtering or electroplatting, the latter method has higher depositing rate and is commonly used in thin film head manufacturing. Whichever method is adopted, the properties of NiFe film are sensitive to many factors, even its shape and size.

3.2 Electro-plated NiFe films

By means of the method using an improved inductance sensor, the initial permeabilities and the domain struetures of the NiFe films (thickness from 0.2 ~ 1.2μm) prepared by electroplating on a glass substrate in a magnetic field of 16 kA/m were investigated[4]. By using the photolithography and ion beam or chemical etching process, the NiFe film was cut into stripes with different width. Experments reveal that, first, NiFe films show a loss of its permeability at all frequencies (between 0.1 ~ 15MHz) after having been etched into stripes, as shown in Fig.5. Second, the etched stripes with different width present different permeability characteristrics, namely, the permeability along the stripe increases with the width decreasing when the easy axis is parallel to the stripe, and the result is reversed if the hard axis is

parallel to the stripe. In order to analyse the permeability characters, the domain patterns were observed by Bitter method. These patterns show that there are many closure domains at the edge of the stripe and that the density of the closure domain varies with the width of the stripe. In hard axis pattern the density increases when the width decreases, however, the easy axis pattern shows the opposite results.

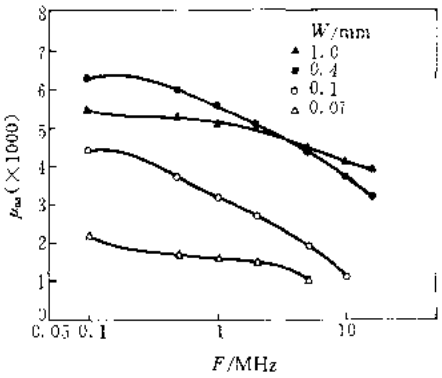


Fig.5 Frequency characteristics of permeability along hard axis for the plated NiFe flim stripes with different width

3.3 Sputtered Fe-Si-Al films

In order to fulfill the requirements of high density recording heads, various new materials with better properties have been studied. The alloy FeSiAl (named Sendust) and Co-base amorphous materials such as CoZr film are among the most interested materials. In comparison with NiFe, FeSiAl alloy has higher magnetization and permeability, higher resistance coefficient and therefore better high-frequency characteristics. It also has high hardness and wearability. But the magnetic properties of this multielement alloy are very sensitive to the change of its chemical composition, this makes it difficult to be produced reliably. We studied the preparation and characteristics of sputtering Fe-SiAl thin film[5]. In order to deposite the films with different compositions, the slices of Si and Al are put evenly on the surface of Fe target. The experiments reveal that when the compositions are 9.3wt% Si, 5.6wt% Al, the remain Fe and 5.6wt% Si, 2.7wt% Al, the remain Fe, both films have good soft magnetic properties. The relation of H_c between films and their composition is shown in Fig.6. The experiments also reveal that in order to prepare the films with good properties, high input sputtering power, appropriate Ar pressure and appropriate annealing temperature and time are required.

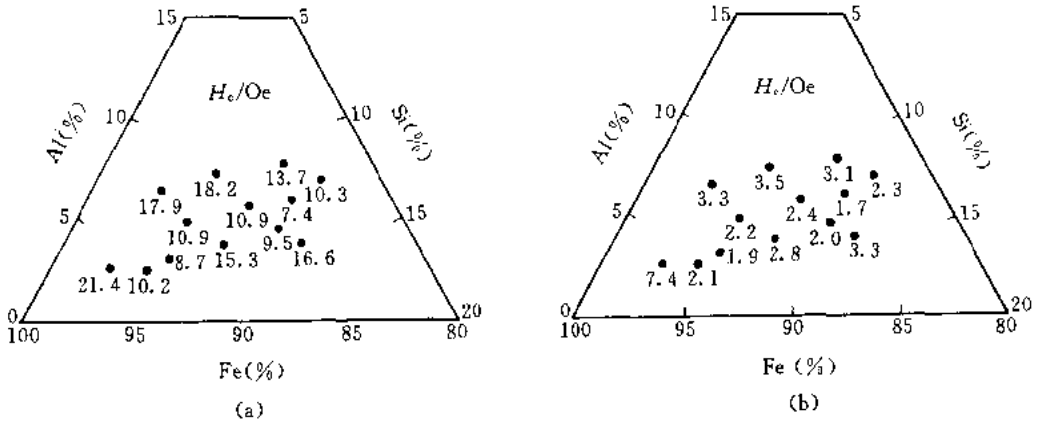


Fig.6 Coercivity H_c vs composition of Fe-Si-Al sputtered films
(a) before annealing and (b) after annealing

3.4 Sputtered Co-Zr-Nb films

Because of its non-crystal structure, amorphous magnetic films present no anisotropy and have high

permeability μ and low coercivity H_c. The composition of the films which the magnetostriction coefficient λ_s is zero can be easily choosen. They also have the advantages of high hardness and high corrosion resistance. We have investigated the effects of various processing conditions to the properties of CoZrNb thin films[6]. In the experiments, an alloy target of $Co_{88.7}Zr_{4.0}Nb_{7.3}$ is used. The experiments show that apparent diffraction peak appears on the X-ray diffraction pattern of the films when the sputtering input power is small (800W). With the increasing of the sputtering power, the diffraction peak drops, until it completely vanishes when the power is high enough (1600W). The measurements of magnetic properties of films show that when the film structure changes from crystal to amorphous, the coercivity H_c decreases sharply, while the saturation magnetic induction $4\pi M_3$ increases a little as shown in Fig.7.

3.5 Magnetoresistive films

The new magnetic head technologies are being driven by the need for higher recording densities. As the size of the recorded signal bits shrinks, these small signals are easier for magnetoresistive than conventional inductive heads to read. The voltage output from an inductive head drops linearly as the bit area gets smaller. But for a magnetoresistive head its output drops much less as the bit area shrinks. The magnetoresistive head makes use of the slight changes in resistance that occur as the magnetized data bits passing beneath it change the angle of magnetization in its magnetoresistive element. In its simplest form the head consists of a narrow stripe mounted in a place perpendicular to the recording media and connected to leads at each end carrying a sense current I_s shown in Fig.8. As a result of the magnetoresistive effect, the resistivity of each portion of this stripe will depend on the angle θ between the direction of magnetization and the current density vector:

$$\rho = \rho_0 + \Delta\rho\cos^2\theta \tag{3}$$

For most materials of interest the increment of resistivity $\Delta\rho$, is in the order of 2 ~ 6 percent of the base resistivity ρ_0. We studied the effect of sputtering conditions of films to its magnetoresistance characteristics[7]. The target is alloy of $Ni_{81}Fe_{19}$. The experiments show that by choosing appropriate sputtering con-

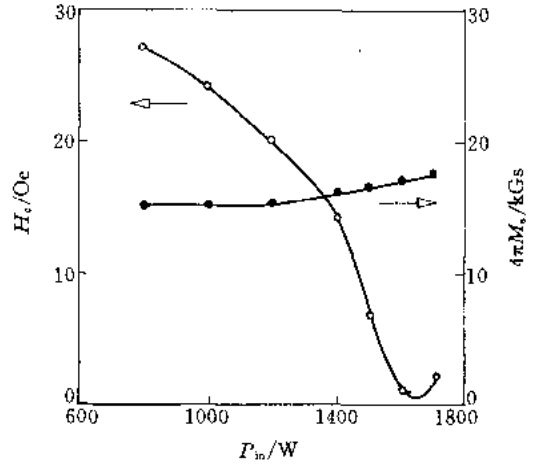


Fig.7 Effect of sputtering power on H_c and M_s of Co-Zr-Nb amorphous films

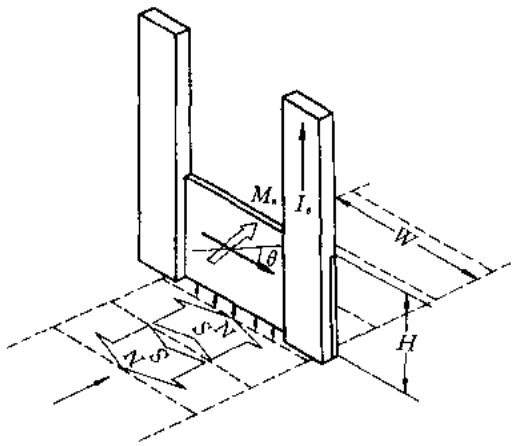


Fig.8 Magnetoresistive head element geometry

ditions, the films which have maximum magnetoresistance coefficient $\Delta\rho/\rho = 5\%$ with thickness 1 μm are prepared. Under the same sputtering conditions, the resistance increases when the thickness of film reduces, while $\Delta\rho$ almost without change, therefore, the magnetoresistance coefficient decreases with the decreasing of film thickness.

3.6 Magnetic thin film heads with superconductors

A new concept[8] of the thin film heads for magnetic recording is to use the superconductive film as a perfect diamagnet to be the gap medium between the two magnetic poles of a thin film ring head for the longitudinal magnetic recording. By using the finite element method, the magnetic field and its distribution for the new designed ring head with a single layer superconductive film or multilayer structure as the medium between the poles was calculated. The results show that the diamagnetic film in the gap not only increases the amplitude of the magnetic field in front of the magnetic poles, but also makes the distribution of the magnetic field much sharper and the amplitude of the field is less influenced by the throat height of the pole tip as well. The bigger the permeability of diamagnets and the more the flying height decreases, the more the performace increases.

The experiment results[9] of the large scale model of a thin film head with superconductors mentioned above show that the values of both the longitudinal and perpendicular components of the useful recording flux density of the thin film head with Y-Ba-Cu oxide superconductive gap between its magnetic poles are obviously larger than that of the conventional head with a non-magnetic gap. It matched very well with the calculation and indicated that the effect of the superconductor diamagnet is of enchancing the head sensitivity. Once the superconductor materials can be operated at the room temperature, this type of thin film head with the perfect diamagnet will present a great potential of application.

3.7 Perpendicular magnetic recording thin film heads

Two schemes of thin film perpendicular magnetic recording head with single pole type were designed as show in Fig.9(a) and (b)[10]. Both the two head structures are using the ferrite substrate as an auxil-

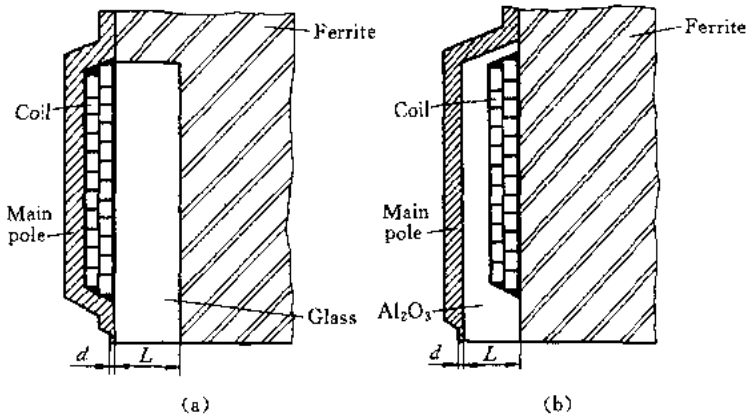


Fig.9 The structures of two type of perpendicular magnetic recording single-pole thin film head

iary pole and the soft magnetic alloy films as a main pole or probe. The thickness of the magnetic pole is $2 \sim 3\ \mu m$, the width of the pole tip is $16\mu m$, thickness is $0.4\mu m$ and the throat height is about $2\mu m$. The read-write dynamic characteristics were measured with the medium in CoCr/NiFe double-layer structure with a coercivity of 67.2kA/m as mentioned before. The experiment results show that the linear recording density can reach 41.6 KFCI.

4 Conclusion

Despite the success of magnetic recording system designs to continuously inerease recording densities in the past, the need for higher magnetization media has steadily increased in order to provide adequate signals in inductive recording applications. The magnetic alloy thin films have the desired high magnetic moment and correspondingly high coercivities. On the other hand, thin film heads with finite pole-tips have been designed for high-bandwidth and high-resolution read-write elements for rigid disk files; and multitrack thin film heads with separate read and write element for data tape recording, where they have a major application today and in the future.

References

[1] S. B. Zhang, Z. S. Zhang, D. Xu, C. S. Yang and B. C. Cai. Study on the CoNiCr/Cr Thin Films as Longitudinal Magnetic Recording Media. Journal of Thin Film Science and Technology (in Chinese), to be pubilished

[2] C. S. Yang, Z. Mei, D. Xu, B. P. Shi and B. C. Cai. The Study of Corrosion Behavior of C/Co NiCr/Cr Thin Film Magnetic Recording Media. Journal of the Magnetics Society of Japan, 1989,13(SI):451

[3] B. C. Cai, et al. Research Report on Thin Film Disks Information Storage Research Centre. Shanghai Jiao Tong University, 1987

[4] W. Pan, Y. F. Zheng and Y. X. Chen. Research on the Permeability and Domian Structures of Etched Permalloy Thin Film Magnetic Pole. Journal of the Magnetics Society of Japan, 1989,13(SI):557

[5] K. Y. Cheng. The Study of Fe – Si – Al Megnetic Thin Films by Magnetron Sputtering. Dissertation of Shanghai Jiao Tong University,1990

[6] S. X. Zhu. The Study of Amorphous CoZrNb Magnetic Thin Films. Dissertation of Shanghai Jiao Tong University, 1989

[7] W. N. Li. Material and Structure Studies of MR Heads. Dissertation of Shanghai Jiao Tong University,1990

[8] Y. X. Chen, D. R. Yan, Y. Xie and M. L. Zhang. A New Concept of Magnetic Thin Film Heads with Superconductors. J. Appl. Phys., 1988,64(10):6026

[9] Y. Xie, D. R. Yan, M. L. Zhang and Y. X. Chen. Experiment Research of a Magnetic Thin Film Head with Superconductor. Journal of the Magnetics Society of Japan, 1989,13 (SI):489

[10] Y. X. Chen, J. L. Zhang, X. L.Zhao and W. L. Tang. Characterization of the Film Heads for Perpendicular Magnetic Recording. ibid,1989.343

新世纪的纳米科学技术

1 前言

纳米科学技术将成为21世纪的核心基础科技。20世纪后半世纪中微米科技在电子学-微电子学(microelectronics)或集成电路方面取得辉煌的成就,对社会带来的进步和对人们生活产生的影响,与电力和蒸汽机相比有过之而无不及,已成为当代高新技术发展的支柱。没有先进的VLSI和ULSI器件,现代的计算机、通信、自控、航天、航空、生物、医学、仪器和设备等高新科技都将是海市蜃楼。在世纪交替之际,微电子学即将走到它的极限(0.1μm),迎来的将是以新的量子效应主宰的纳电子学(nanoelectronics),这将可能成为纳米科技走上舞台的第一场高潮。相继微电子学而崛起的集成光学(微光学)和微机械,在当前已初露端睨。从近几年的发展趋势来看,这两方面在下世纪初必将有更大的突破,特别是微电子学、微光学和微机械三方面的综合集成将会导致一场新的技术革命,它们所造成的影响目前难以估量。

纳(米)科学技术是在纳米尺度上(0.1~100nm),研究物质(包括分子、原子、电子及光子)和器件的结构特性效应机理及其应用的多学科交叉的边缘新学科。其最终目标是在纳米尺度上设计和制造直到基于操纵和控制单个原子的具有特定功能的材料器件和系统。其研究的内容将涉及对纳米尺度的材料器件和材料器件系统及直到分子、原子、电子和光子的行为和功能进行观察、测量、分析、计算、设计、加工、装配和实际应用中所需要的各种技术的原理方法、工具、仪器和设备。

人们期望21世纪成为信息时代。21世纪的信息科学技术将会继续向高速率、大容量、多媒体和广用户的宏伟目标奋进,概括地说:要实现3T(T为Tera的缩写(即10^{12})),所谓3T是THz的信号带宽、Tb/s的传输和运算速率以及TBytes的单盘存储容量,发展的基础仍离不开新材料、新工艺和新器件的不断开拓。单靠现有的微米技术,已满足不了这一要求,还需要发展纳材料(nano-materials)、纳加工(nanofabrication)和纳器件(nano-devices)。

许多学者也预言,21世纪将是生物医学世纪。随着科学技术的高度发展,人类迫切希望揭开自身存在的奥秘,生命科学得到了极为迅速的发展,生命科学在解决人类发展和进步的各种重大问题中的作用越来越重要。现代生命科学的突破,也离不开纳米技术的渗透。纳生物学(nanobiology)是在纳米尺度上研究生物反应机理、生物大分子细胞器的结构功能和动态生物过程,并能对分子和细胞进行修复、复制和调控等操纵及改性。生命过程所必需的能量代谢、物质代谢及其他各种生物生理过程,都是在细胞这一层次进行的,通过纳米技术可以获得在细胞膜和细胞器表面的结构信息,为细胞工程、蛋白质工程和酶工程的发展提供依据和手段。

现代工业消耗大量矿物燃料,释放二氧化碳进入大气,造成了严重的环境问题。而森林和植物则在生产有用的产品的同时,从大气中吸收二氧化碳。植物叶子是利用分子电子器件,如

叶绿素分子和光合作用中心的太阳能收集器。因此,国外有些科学家正在构想利用分子纳米技术(molecular nanotechnology)制造出与绿色植物功能相仿的所谓分子机(molecular machinery),从而能大量生产廉价的太阳能收集器、品质优良的大型结构或成品,又能吸收大气中的二氧化碳。这种分子纳米技术(或称分子工程(molecular engineering))预示着制造方法和制造装备方面将出现一场根本性的变革。通过实现分子水平上的精确控制——模仿在生命组织中发现的控制方法,可以作为未来加工制造的基础。与现有的工业相比,它的生产过程更清洁、产品更丰富、效率更高。如同绿色植物一样,它能在消耗二氧化碳的同时,制造太阳能收集器和其他有用的产品,从日常的衣食住行用品直到分子制造设备自身。

由此可见,纳米科学技术涉及的学科领域十分广阔,从目前研究的范畴来划分大致有:纳物理学(nanophysics)、纳化学(nanochemistry)、纳生物学、纳医学(nanomedicine)、纳材料学(nanomaterial science)、纳电子学、纳光子学(nanophotonics)、纳机械学(nanomechanics)、纳摩擦学(nanotribology)、纳显微学(nanoscopy)、纳计量学(nanometrology)、纳加工(nanofabrication)和分子纳技术等。

综上所述,纳米科学技术将是未来发展信息、能源、材料、生物、医学、环境、工业、农业和国防等许多领域的基础和支柱。不仅创建高新科技学科需要纳米科学技术,即使是传统学科的更新、拓展和改造也与纳米科学技术有密切关系。杨振宁教授在论及21世纪的科技发展时认为,纳米科学技术和生物生命科学将成为最重要的和发挥巨大作用的领域。这是极有科学远见的论断。本文将就纳米科学技术几个主要领域的发展现状及趋向作简要的综述和讨论。

2 敢于想小的人——进入原子分子世界

每当人们论及纳米技术的先驱者时,无例外地都提到因创立量子电动力学而获得诺贝尔物理奖的物理学家费曼(Richard P. Feynman),他在1959年12月一次美国物理学会上所作的题为"底部有充裕的空间"(There's plenty of room at the bottom)具有历史意义的演讲。这篇演讲后来刊登在1960年2月的美国加州理工学院的"Engineering and Science"杂志上。随着纳米科学技术的飞速发展,30年前费曼在他演讲中所作的那些科学预言,正在不断地成为现实。有人把费曼称为爱因斯坦后的第二个天才不是没有道理的。1991年"科学"杂志一期有关"建造小世界:从原子操纵到微加工"的专辑中,重新引录了这篇演讲的有关内容。即使在今天,这篇演讲仍有着指导意义。现将其主要内容摘译如下,仍有其现实意义。

底部有充分的空间

我想描述的一个领域至今尚很少涉及,但是在原则上,那里是大有可为的。这一领域与其他的不一样,他不会告诉我们许多基础物理学,从"什么是新奇粒子"的意义上说,在某种程度上它有些像固态物理学,它可以告诉我们有很大兴趣的关于在复杂情况下出现的新奇现象。特别要指出的最重要的一点是它将会有大量的技术上的应用。

我要说的是关于任何在小尺度上操纵和控制物体的问题。

当我提到这一点时,人们立刻会告诉我关于小型化,以及当前已取得怎样大的进展,他们告诉我关于马达只有小指甲那么大。还告诉我市场出现了一种装置,用它可将主祷文写在大头针的头上。这算得了什么!这是最原始的,是在我要讨论的方向上左右徘徊。下面我要说

的是令人惊愕的小世界。

如何把英国的百科全书写在大头针的头上

有一种方法可能是(虽然我不能确定是否一定行得通):我们采用光,使光学显微镜反过来操作,将它焦聚到非常小的光电屏上。通过电子显微镜的透镜将这些电子束聚焦缩小尺寸,并直接轰击到材料的表面。我不知道如果电子束经过足够长的距离后能否将材料刻蚀掉?如果这对金属表面不行,无论如何应该可以找到某种表面,将它涂在原先的大头针上,这样,在电子轰击后产生的变化可以提供识别。

甚至可以做得更小

当然,这事实——巨量的信息能够承载在极小的空间……是生物学家所熟知的,并解答了存在的奥秘,虽然我们还并不清楚了解如何将组织复杂的生物,例如我们自己的全部信息存储在极小的细胞中……

这种在小尺度上写入信息的生物学例子,使我想到某些事情是可能的。生物不是单单写入信息,还做某些其他的事。生物学系统可能非常小。许多细胞虽然极小,但它们十分有能力;它们制造不同的物质;它们在附近行走;它们扭动;它们做所有各种奇妙的事——全都在极小的尺度上。它们也存储信息。让我们考虑这样的可能性,我们也能做极小的事,那些正是我们所要的——能操纵在那样的水平上制造出物体。

把东西做得极小甚至可能是有经济意义的事情。让我来提醒你们某些关于计算机的问题。

使计算机小型化

我不知道如何把计算机尺寸做小的实际方法,但我确实知道现在的计算机很大,它们占满整个房间。为什么我们不能把它做得很小,把导线一点一点地缩小。我这里所说的"小",例如导线的直径应该是10个或100个原子,电路应该是数百纳米的截面。凡了解计算机逻辑理论的每一个人都会得出结论,计算机可能做的事是十分有趣的——如果它们的复杂程度能提高几个量级,如果它们有成百万倍的元件,它们就能作出判断。

最终,欲使我们的计算机变得越来越快,越来越复杂,我们必须把它们做得越来越小,但是这里有充分的余地让它们做得更小。从物理学定律上我看不出谁能说计算机的元件不能比现在的大大地缩小。实际上,可以有许多优点。

我们怎样能够做出这种器件?我们将应用何种制造工艺?我们可以考虑的一种可能性是前面谈到有关用原子的某种排列来写入,先是蒸发材料,而后在它上面蒸发绝缘体。作为下一层,蒸发另一导线、另一绝缘体,这样继续下去。如此简单地蒸发直到你得到一块材料,它包含了尺寸极细小的各种元件——线圈和电容器、晶体管及其他种种。

将原子重新排列

我并不害怕考虑那最后的问题,最终—在好久的将来(in the great future)—是否能按我们所要的方式排列原子,就是从头到底的那些原子。如果能按我们需要的方式一个一个地排列原子,那将会出现怎样的奇迹?直到现在,我们满足于挖掘地下来开发矿藏,我们将它们加热,

以大尺度对他们处理，我们希望得到纯洁的物质，但往往存在如此多的杂质，等等。而且我们经常必须接受自然给予我们的某种原子排列。至今我们尚未得到任何东西，比方说具有棋盘般的原子排列，或将杂质原子的排列恰好分开 100nm。

我们怎么能做出具有正确层数的层状结构？如果确实做到了按我们需要的方式排列原子，将会得到怎样特性的材料？这对理论上的研究是十分有趣的。我不能确切地看到将发生什么，但我毫不怀疑，当我们对在小尺度上排列物体有所控制时，我们将有广阔的天地使物质获得可能的特性以及我们可做各种不同的事。

小世界中的原子

当我们来到了非常非常小的世界时—比如说 7 个原子的电路—我们将遇到许许多多新事物，这表示对设计来说是全新的机会。原子在小尺度上的行为与大尺度时不相同，因为它们满足量子力学定律。所以当我们走下去并在那里不停地拨弄原子时，我们在不同的定律下工作，我们能期待去作不同的事。我们能以不同的方法制造，我们不仅能应用电路，也可以是包括量子化能级，或量子化自旋的相互作用的某些系统，等等……

在原子水平上，我们有各种新的力、各种新的可能性和各种新的效应。材料制备和生产的问题将有很大的不同。正如我说过我是受到生物现象的鼓舞，在那里化学力以反复的方式用来产生不可思议的效应(其中之一是作者)。

费曼这篇经典性的演讲距今已近 40 年，在这段时期，科学技术得到了空前的发展。他在这演讲中的许多预言不断地成为现实，人类能够控制和操纵单个原子的“好久的将来”也真的到来了。人们重读这篇演讲时，会感到非常亲切。可是在 1959 年 12 月，听众中许多人对这演讲并不理解。有的为之惊奇，有的认为是开开玩笑。当然更多的人并不知道这演讲的存在。但是不论信与不信，知与不知，事物的发展确实是沿着费曼指出的方向在稳步前进。正如 IBM 阿曼登研究中心的 Don Eigler，他在 1989 年用 STM 操纵原子，并将 35 个 Xe 原子排列成 IBM 三字母，他在参加操纵原子的研究工作前，并没有读过费曼的演讲，当他最后看了费曼的报告后十分风趣地说：“当我读到它时，好像费曼的灵魂在我背后说：‘看！我 30 年前就想到这些事了。’”

3 零的突破——扫描隧道显微镜是观察、操纵单个原子的有力工具

能够直接对原子作观察乃至操纵，这已是包括费曼在内的许多科学家一直梦寐以求的事。直到扫描隧道显微镜(STM)问世，终于使许多年来的愿望成为现实。STM 是由在苏黎世的 IBM 实验室的物理学家 Binning 和 Rochrer 为研究表面现象而发明的。1979 年他们申请了 STM 的第一个专利。1982 年他们利用 STM 获得了硅晶体表面图像，清楚地显示出单个硅原子的排列。具有讽刺性的是，他们这项工作的重要性没有很快得到认识。Rohrer 和 Binning 关于这新工具的科学论文竟被拒绝出版，评语是“不够兴趣”。1986 年，Binning 和 Rochrer 因发明 STM 而获得了诺贝尔奖，因发明某种新工具而得诺贝尔奖是很少见的，足可见此项工作的重大意义。这可以说是人类进入原子世界的零的突破。今天，有关 STM 家族的会议已吸引了来自世界各方数以千计的科学家。

STM 的工作原理是基于量子力学中的隧道效应。把两块导体或半导体电极靠得相当近但

不接触，保持一个很小的距离 d，按照经典理论，电子不能穿越其间的势垒而形成电流。根据量子理论，由于电子固有的波动性，固体表面的电子密度并不突然减小为零，而是在很短的距离内（约数 0.1nm）按指数式迅速衰减。因而“表面”对电子来说是一个模糊的界面，它只是表示最外层原子的轮廓。如果两金属电极接近到几 0.1nm 之内，两者外围电子云发生交叠，在一定的外加电压作用下，电极间有电流通过，此电流称为隧道电流。按照量子力学转移哈密尔顿理论，隧道电流 J 取决于两电极表面电子波函数的交叠，因而与电极间距 d 有密切关系，可表示为：

$$J \propto V\exp(-Kd)$$

式中：V 为外加电压；K 是常数与金属表面电子功函数有关。一般情况下，d 变化为 0.1nm 时，隧道电流变化近一个数量级。这说明隧道电流对距离的变化十分敏感，STM 就是根据这一特性而制成的。

STM 的基本结构与工作原理如图 1 所示[2]。两个电极做成探针和平板（即样品）。隧道电流的大部分将集中在探针的尖端附近，因而样品上的隧道电流主要也将限制在正对针尖附近的小区内。这表示隧道电流仅与针尖相对的样品表面很小区域内的形貌和电子态有关。若探针的尖端有单原子突出，则隧道电流的有效作用范围最小，可达到单原子尺度。若让针尖沿样品表面移动，隧道电流的变化可反映出针尖所到达的样品表面的形貌和表面电子态。因而，利用 STM 可以获得原子级分辨率的表面信息。

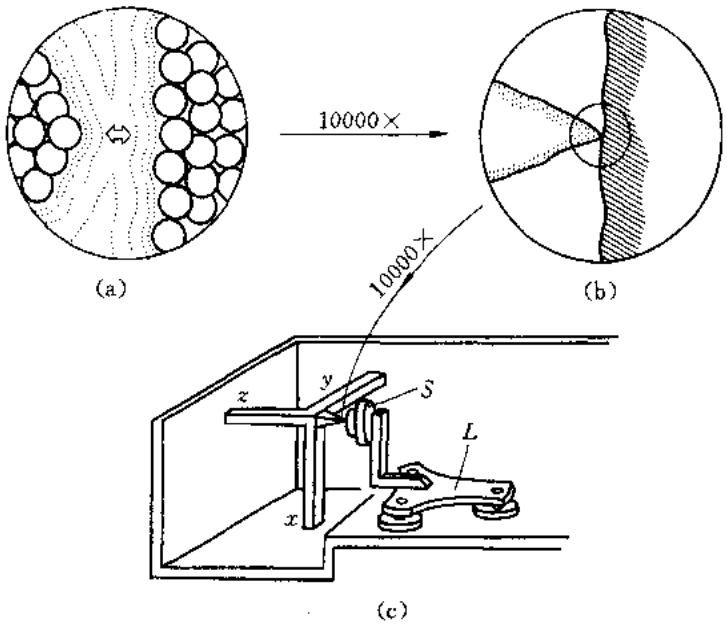


图 1　STM 基本结构及工作原理

STM 的探针用 x、y、z、三个方向的压电陶瓷驱动。首先将针尖沿 z 方向驱动接近样品表面，然后在 x 和 y 方向驱动的压电陶瓷上加上扫描电压，使针尖沿样品表面作 xy 二维扫描，同时记录隧道电流的变化。整个 STM 装置，必须有良好的机械防振措施，它可以在大气、液体中工作，也可安放在超高真空室中，与其他表面分析仪器相配合使用。

STM 主要有两种不同的操作模式：恒定电流模式和恒定高度模式。前者是使隧道电流在探针扫描时保持恒值，则针尖的上下起伏反映出样品表面的轮廓和形貌；后者是保持针尖的高

度不变，记录隧道电流随样品表面位置的变化，经计算机处理后成为样品的形貌图像。这两种模式各有利弊。在恒定电流模式中，针尖始终与样品保持一定距离，允许样品表面有较大起伏，但探针高度的跟踪需要精确的反馈系统，并伴随一定的延时，故扫描速度受限制。而恒定高度模式则不同，扫描速度可以提高，但要求样品的表面不能起伏太大；否则，表面低凹区的形貌细部不易分辨，凸出部分太高也易损坏针尖。

更重要的是，STM 不仅可用于观察，还可用来进行材料表面的局部改性，进行纳米量级的加工，甚至可以操纵单个原子和分子。恰好在 Feynman 发表上述讲话 30 年后，1989 年美国加州 IBM 研究中心的科学家们搬动 35 个 Xe 原子拼排成 IBM 三个字母，如图 2 所示。它宣告了人类直接操纵原子时代的开始。

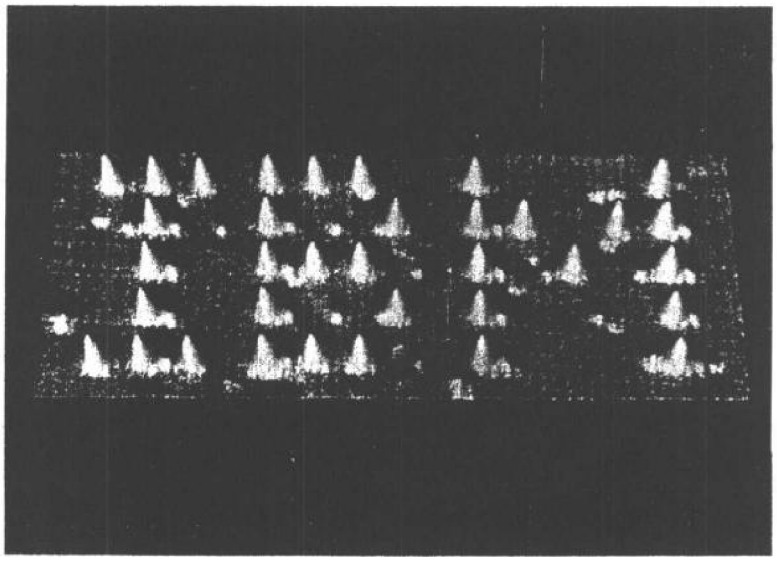


图 2　35 个 Xe 原排列成 IBM 字母的形貌

利用 STM 操纵原子的方法和过程可以有许多种，大致上可分为平行操作和垂直操作两类[3]。在平行操作中，被吸附的原子或分子沿着表面移动；垂直操作中，原子或分子将从表面向 STM 的探针尖转移或反之。不论哪种过程，其目标都是要将原子有目的地重新排列。

首先讨论平行操作。吸附着的原子或分子在表面的再分布可以通过不同的过程或机理实现，例如：场辅扩散(field-assisted diffusion)或移滑(sliding)等过程。在这类过程中，被操纵的原子与衬底表面之间的键永不断开，也就是原子始终处于吸附势阱中保持着良好的吸附状态。在这些过程中所需要的能量是克服沿表面扩散的势垒。能量的数值范围大致为吸附能的 1/10 ~ 1/3，可以从很弱的物理吸附的数十毫伏(在密堆金属表面的原子：如 Pt(111)上的 Xe)，到较强的化学吸附大约从 0.1 ~ 1.0eV。

在 STM 正常的成像过程中，针尖和样品表面间有强电场存在，当隧道间隙为 0.5nm 和电势差为 1 ~ 10V 时，电场强度的范围是 2 ~ 20V/nm。此电场不是均匀的，集中在针尖附近，模拟计算的电场分布如图 3(a)所示，其中针尖的直径为 10nm，离开半无限的金属表面的高度为 0.5nm，电势差为 3V。这样强的电场可以使原子产生场电离和解吸附。原子的场辅操纵可以在较低的电场下进行，空间不均匀电场与吸附原子的偶极矩间能产生势能梯度或沿表面的力，结果造成原子的场辅定向扩散。这种效应不仅是一种操纵技术，而且可用来测量原子的偶极矩及极化率。STM 的场辅定向扩散由 Whitman 等首先对 GaAs 和 InSb(110)表面上的 Cs 原子观

察到[4]。当这些衬底处在负极性(-2~-3V)时,这些表面上的 Cs 原子的 STM 图像具有稳定结构,当 Cs 密度低时呈现一维锯齿状排列,如果将试样的极性反转为正时,产生明显的 Cs 原子扩散流向针尖下的强电场区,Cs 原子迁移的数量与试样上正极性维持的时间有关。

STM 探针对表面上吸附的原子或分子通常都有力的作用。其中一个分量是由于原子间的势能,那是吸附体与探针最外层原子间的化学束缚力,通过调节探针的位置可以改变

探针作用于吸附体上力的大小和方向。因此,就可能用探针拉着吸附原子沿表面运动,这称为移滑操纵,这种移滑的过程可用图 3(b)表示。开始时被移动的吸附原子放在 STM 成像的位置 a,而后提高设定的隧穿电流,从而由反馈回路使探针下降到新高度 b,这时探针和吸附原子间作用力增加。如果在使电流保持恒定的条件下,将探针沿表面作横向移动 c,探针就能拖着吸附原子沿着其轨迹到达预定的终点 d,最后回到起始的设定电流,探针升高,STM 恢复到原来的成像模式 e。用移滑过程操纵吸附原子的特征参数是探针的阈值高度,当超过此高度时,探针与吸附体之间作用力太弱,不再能操纵。在阈值高度时,探针与吸附体间的作用力恰好强到足以使被拖着沿着表面运动。由于 STM 探针在表面上的绝对高度不是直接测量值,而是由隧道结的电阻表示,此电阻不仅与探针高度有单--的关系,而且可以精确控制。

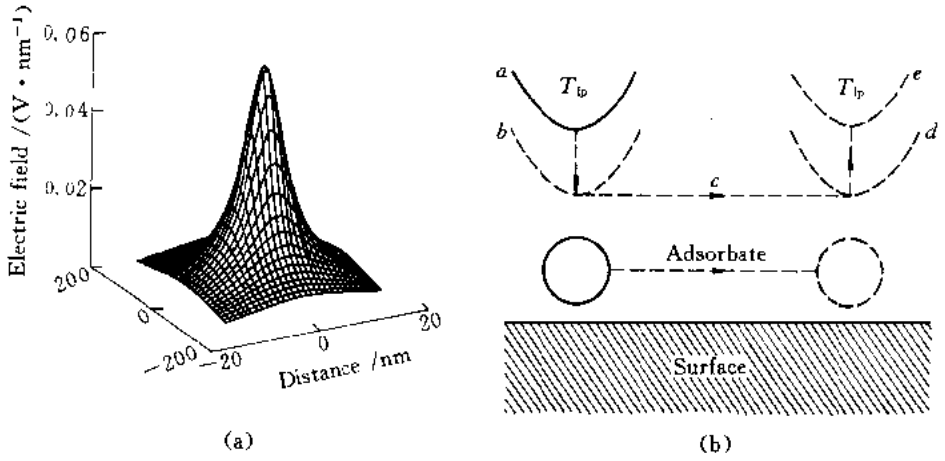


图 3 (a) STM 探针附近电场的空间分布 (b) 移滑操纵过程示意图

下面讨论垂直过程,即 STM 作原子操纵时产生原子、分子或原子团从样品表面转移到探针,或反过来。这与上面讨论的平行操纵中被移动的原子自始至终不脱离衬底表面不同。为简单起见,下面仅讨论原子从表面向探针转移的情况。这种过程需要的能量是原子为了从表面到探针必须克服的势垒高度。这势垒高度取决于探针与表面分开的距离,当探针-表面间距大时,这能量趋于极限值即吸附能,当探针与原子靠得很近时,这能量减小接近于零。因而调节探针高度可获得适合我们要求的势垒高度。根据已有的研究报告,垂直过程可以由不同的模式实现,如接触或接近转移,场蒸发和电迁移等。

从概念上说接触转移是最简单的原子操纵过程。将探针向原子靠近直到隧道结二边探针和表面吸附阱相结合,即分开两阱的势垒消失,这时吸附原子可看成同时为探针和表面束缚,而后将探针带着吸附的原子返回。这过程成败的关键是当探针撤回时吸附原子与表面的键必须断开。我们可以设想当探针返回时,吸附原子在探针和表面之间有一个选择,其决定何者具有最高的束缚能。实验表明,单个的 Xe 原子在接触转移过程中大多能可靠地从 Pt(111)或 Ni(110)表面平台向探针转移。对 Xe 原子来说,当外加电势在 ±0.05V 范围内对此过程没有

任何影响。

如果探针和样品间的距离略为增加，探针与表面原子的吸附阱可靠得相当近，使中间的势垒明显地降低，但仍保留一有限数值，这样，就可以由热运动产生原子转移，这过程称为接近转移(transfer-near-contact)。此过程的速率与 $v\exp(-Q/kT)$ 成正比。其中：v 是频率因子；Q 是探针及样品间较少的势垒；k 是波尔兹曼常数；T 是绝对温度。当势垒减少到 ~0.75eV 时，如果 $v\approx10^{13}s^{-1}$ 和 $T=300K$，达到的转移速率为 $1\,s^{-1}$。这种接触或接近转移最简单的方式是完全不施加任何电场、电势差或电流。但是在某些情况下也应该可以通过接触时在结上施加偏置以确定转移的方向。

场蒸发(field evaporation)是当探针和表面间由于外加电压脉冲而形成原子转移，这种过程可以看作离子的热运动克服"肖脱基"势垒而蒸发，而外加电场使导体外的势能降低。比较典型的研究报道是 Si 原子在超高真空和室温条件下能可逆地在 Si 表面和 STM 的 W 探针之间转移[5]。这项实验中在 Si 表面施加了 +3V 脉冲电压，结果在针尖相对的表面上是形成凸起的小丘，其周围凹下成沟形。STM 所具有场蒸发效应与场离子显微镜(FIM)的主要区别是阈值低得多。例如，FIM 针尖使 Au 和 Si 蒸发的电场强度在 30 ~ 50V/nm，而 STM 的阈值仅 4 ~ 10 V/nm。另外，STM 可造成负离子场蒸发，这在 FIM 中从未观察到。研究表明，这两者的区别是不同电极结构所决定。

基于上述 Si 原子可在 Si 表面和 W 探针之间可逆转移的实验，Lyo 等[6]提出了用 STM 可以实现原子开关。由于隧道结的电导与电极间距离具有指数式关系，在隧道电流区内原子任何少量的重排很容易引起电导可测量出变化。如果隧道结电极间一个原子的可逆运动可以唯一地用外施电信号进行控制，就能实现原子尺度的双稳态电子开关。

4 纳电子科学技术将成为下一世纪信息时代的中心

数十年来，从事微电子技术的人习惯于将电子看成粒子。现在随着半导体器件的结构越来越小，人们注意到电子量子力学的两重性的另一面——电子的波动性已不能被忽视。包括材料科学家的贡献在内，现在已能造出尺寸如此小的材料的器件使电子波的传输受到约束，迫使它们具有特定的波长和能量。当结构尺寸小于 100nm 以下时，现有半导体晶体管的工作的物理原理已不再适用，需要利用新的物理效应。器件物理学家称这为"量子区"，要求对电子的行为有不同的数学描述。

在量子区电子最重要的量子现象是隧穿(tunneling)，即电子能穿透势垒，与电子作为经典粒子只能越过势垒不同，如图 4 所示[7]。另一种是量子尺寸效应(quantum size effect)或约束效应(confinement effect)。

在量子阱结构中，电子(或空穴)的运动沿量子阱生长方受到约束，其能量发生量子化，形成一系列分立的能极，称量子能极。这种由于载流子被限制在势阱中而引起能量的量子化，即为量子约束效应。它对量子阱或超晶格材料和器件的电学和光学性质产生极为重要的影响。

图 5 是一个双势垒半导体结构的能带图和伏安特性。它是说明隧穿效应和约束效应的典型例子。电子在势阱中形成的约束能级，最低的能级 E_0 高于导带底能级，E_0 是量子阱中的基态，第一激发态为 E_1，还可能有 $E_2, E_3, \cdots$ 等。基态和激发态的分立能量值和量子阱中子能级的多少都由量子阱的深度和宽度所决定，而阱深和阱宽与组成量子阱结构的材料有关。双势

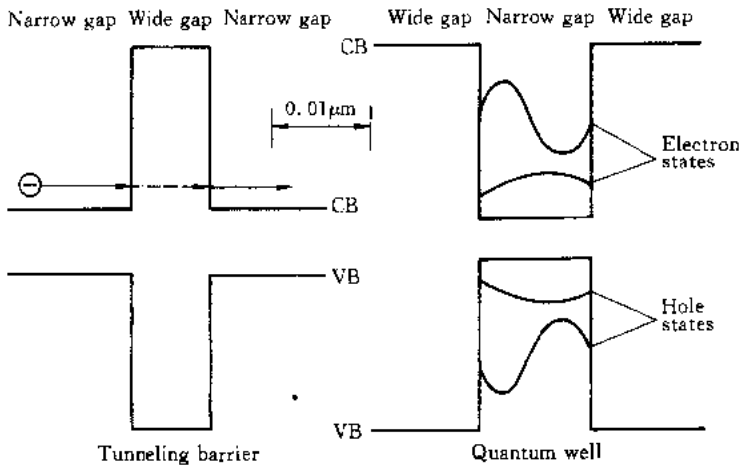


图 4　量子异质结构:电子波

垒两边是重掺杂区,当施加偏压时,形成隧穿电流。当不加偏压时,能带不发生倾斜,没有隧穿电流。如图 5 中 A。外施偏压将引起双势垒结构的能带倾斜,如图 5 中 B 和 C。当负电极区的费米能级与势阱中的势态能级高度相等时,重掺杂区的电子可以有最大的几率隧穿进入 E_0 能极,然后再隧穿到双势垒另一边的空态能级中去,形成大的隧穿电流。当费米能级与势阱能级不重合时,由于势垒边界上的量子力学反射效应,电子的隧穿几率减小。因而,图中伏安特性中电流的峰值表示费米能级与势阱中能级相重合。如果势阱中有多个能级存在,则伏安特性会有多个峰值出现。伏安特性中呈现的负微分电阻,使这种双重势垒结构可用来制造重要的量子器件,例如共振隧穿二极管和晶体管。由于隧穿过程的响应时间极短,用它可以制造高速器件和电路。

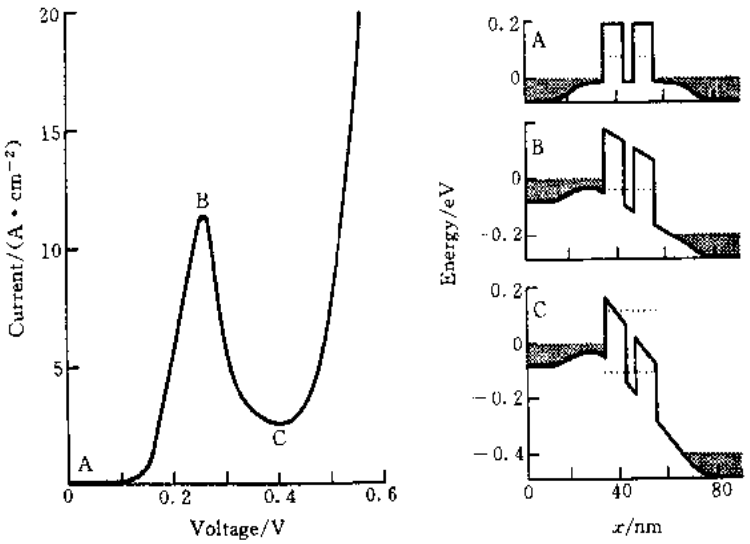


图 5　典型谐振隧穿二级管在不同工作点的伏安特性及相应的导带轮廓

这种纳米尺度的量子结构,最简单的情形就是上述量子阱,由两个异质结势垒构成。电子在量子阱中被量子化的影响可定量地用电子态密度来描述。态密度函数是电子在固定能量范围内所允许的量子态数量的度量。在非量子约束的半导体导带中三维电子气的态密度通常随

高出导带底的能量的平方根而增加。而对于上述量子阱中二维电子气的态密度则形成阶梯函数，如图 6 所示。

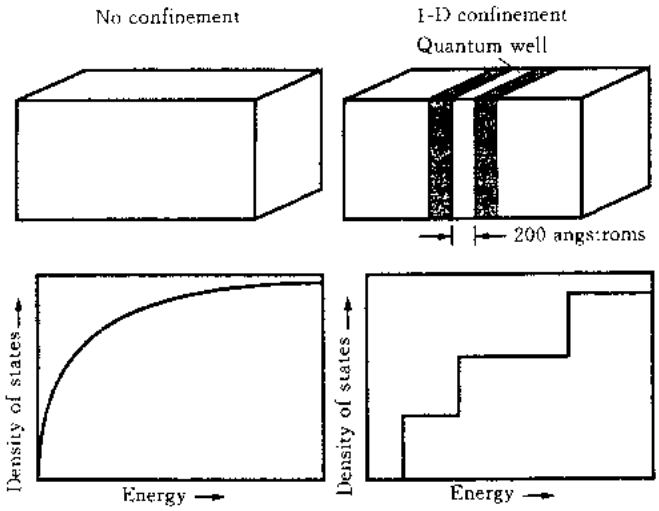


图 6　3-D 和 2-D 电子气的态密度

如果对量子阱结构的横向再加一维限制，比如说是 y 方向。这样电子在 y 方向受约束，也将被量子化，得到的是一维电子气。这种结构称为量子线。这时电子的态密度在每子带的边缘成为无限个不连续函数，而随能量平方根的倒数下降。以此类推，再进一步在另一维也加限制，比如说在 x 方向，得到的量子结构称为量子点。如果每维度的尺寸足够小，电子三维自由度都消失，形成束缚态，也可看成零维电子气。这时态密度成为 δ 函数，在有束缚态存在的能量处出现，其他能量处都为零。量子线和量子点的结构及其态密度函数如图 7 所示。这种从常规半导体的宽能带向量子线、量子点的狭窄能级转化可使电路和光器件具有高得多的速度和效率。

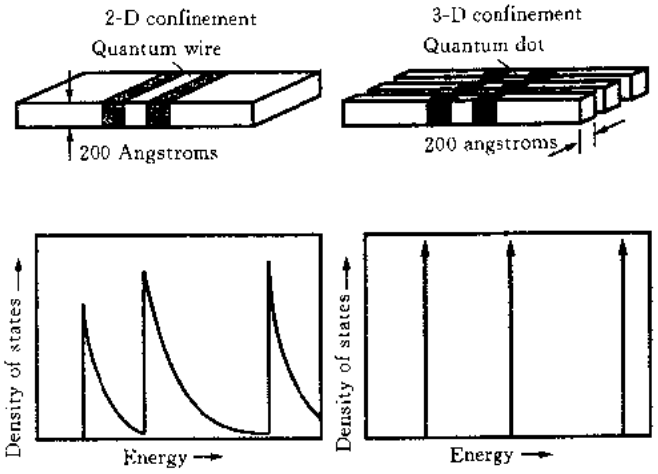


图 7　量子线、量子点结构，1-D 和 2-D 电子气的态密度

例如，从现有半导体激光器激发一束光需要大量电功率，因为材料的宽能带中仅仅只有少数能级可有效地产生激光。但是，量子结构耗费电功率很少，因为大部分或全部能态集中在所要求的能级上。因而少量的功率可以获得同样强的激光。另一有利因素是量子线、量子点的尺度往往只有数 10nm 或更小，这意味着制成的电路和激光器将比现有的缩小几个量级。

但并不是所有这些都是未来的事。有一种量子结构，即量子阱器件已在许多方面获得实际应用，例如卫星微波接收器应用的晶体管，光纤通讯系统和光盘放录机中的激光器等。可以乐观地说，量子力学已开始进入了人们的日常生活。

但是，纳电子学所有的极小尺寸只相当于几百原子宽或数 10nm，造成了制造更为复杂的量子结构的巨大障碍。加工这些系统面临着严重的挑战。人们正努力用尽半导体工艺传统工具的极限——用光刻产生电路的图形，刻蚀工艺用来雕刻材料，淀积技术用于添加细丝或薄层新材料。同时更鼓励研究发明某些新方法，例如用化学方法合成精细结构——一个分子一个分子地将它们建造起来，代替从大块材料上把它们刻出来。

即使作出了极大的努力，目前能得到的量子结构并不都如理论上所期望的那样足够小、足够尖锐和足够均匀。并且一旦做成了单量子结构，他们还面临更难克服的挑战，要将单个器件互联成超密集的计算机芯片或超大容量系统的单元。科学家认为完成这种组装在理论上是可行的，但需要的时间恐怕会比人们想象的更长。不过正如 IBM 的科学家 Armstrong 所指出的："我相信纳科学和纳技术将成为下一世纪信息时代的核心，它将与本世纪 70 年代以来微米尺度对科学和技术带来的革命影响同样巨大。"

5 纳生物学——为人类认识生命奥秘打开大门

现代生命科学已从描述性和实验性阶段向更高的定量和分析性阶段发展。其中分子生物学已成为主要基础，研究的核心问题是生物大分子的结构、功能及包含的各种生命信息。蛋白质和核酸是生物大分子中最基本的，一切生命现象都与这两种大分子有密切的联系，因而成为当今分子生物学研究的焦点。蛋白质和核酸分子其结构尺寸一般在数 10nm 以下，因而光学显微镜无法直接观察到分子的结构和形貌。以往对纳米尺度的生物分子结构研究只能借助于电子显微镜和 X-射线衍射，但都有很大的局限性。扫描探针显微镜(SPM)，包括 STM、AFM(原子力显微镜)以及 NSOM(近场扫描光学显微镜)等的出现，为分子生理学的研究提供了强有力的手段。它们可在接近自然的大气或液体条件下成像，样品不易变形或收损，分辨率高，具有高度的直观性，能提供三维表面信息。1989 年国际上第一张 DNA 双螺旋分子直观图像，就是用 STM 得到的。可见，纳米技术与生物学的结合，将会对生物学和生命科学的发展有巨大的推动作用。

应用纳米技术，可以动态地研究生物分子生理条件下的结构，特别是可以获得细胞内的诸多信息，它们是生命信息的重要基础。生命过程所必需的能量代谢、物质代谢及其他各种生物生理过程，都是细胞这个基本单元中进行的。利用 SPM 可得到细胞膜、细胞器表面的结构信息，以及在不同环境条件下的变化。另外，如果用尖端直径极小的纳传感器，插到活细胞内可以不严重干扰细胞的正常生理过程，这样就能获取活细胞内足够的动态信息，得到对机体生理及病理过程的认识，为临床疾病提供诊断和治疗的客观依据。纳生物学的发展，将为细胞工程、蛋白质工程、酶工程以及基因工程大大增加活力。

纳生物学研究的另一个重要方面是利用已掌握的生物和生命信息，来创造出新的生物器件和生物机器而造福于人类。

人的大脑是由蛋白质构造的，纳米尺度的分子，具有多种多样的功能，这种活器官可作为超高速开关、化学反应的容器、图像识别、电池、数据库、装配工人以及其他许多。全世界有数以百计的科学家正探索在纳技术和生物学的前沿，希望有朝一天，也能造出像大脑一样的分子

器件。当然，这是一条漫长和艰苦的路。今天，科学家们虽然尚未做到这点，但是在人造和自然之间已迈出了一大步。例如设想把生物分子用于新一代计算机部件、医疗诊断的工具以及化学和生物传感器等。

由于生物分子不仅尺度小、结构复杂、十分娇嫩，往往移出了它的生长地就易失效，因而科学家要克服重重障碍。这需要不同学科的专家协同工作，采用不同的工具和技术，才能取得成效，利用基因工程可以裁制蛋白质构造并赋予特定功能。有机化学为之提供新材料用于固定住和保存改变了的蛋白质。电气工程可以为获得蛋白质内部活动的信号解决探测技术等等。当所有这些都一切就绪，便可能将生物分子的简单组件或甚至是单个生物分子转变成用户专用的器件或机器。

美国 Bath 大学的教授和学生们正利用蛋白质作为化学处理工厂。他们集中致力于铁红质(ferritin)的研究，这是从人的肝脏和其他许多器官中发现的蛋白质，它形成 8nm 宽的笼子结构，对氧化铁具有很强的亲合力。当我们体内的自由铁遇到氧结合成铁锈时，这过程可在铁红质结构内部发生，它能将有毒化合物安全地关闭在笼内。他们研究了这种天然蛋白质的结构和特性，一旦有所发现，就有可能利用这种蛋白质作为反应容器对其他材料的颗粒尺寸加以控制。他们发现除了氧化铁外，这种蛋白质能笼住其他几种化合物，包括氧化锰和硫化铁。下一步是设法改变蛋白质本身的化学亲合力，使它能俘虏其他更多的化合物。他们还验证了包含在特殊矿化产品的蛋白质内有两个关键部位，从原理上说，就可能将蛋白质加工适合于我们所需要的矿化产品。这种可能性引起了半导体研究者的巨大兴趣，他们正在努力于构造具有量子力学特性的精细结构，即前面提到的量子点等。更有甚者，这种被封闭的颗粒可以不受限制地在人体内畅游，有可能用它来诊断或治疗疾病。

其他许多科学家设想生物分子可发挥更多的效用。他们希望利用生物分子对环境特别高的敏感性来制造高性能的生物传感器。现在医疗诊断中应用的常规传感器通常都是通过探测固定在目标分子上的酶所引发的化学变化。而这种新的更灵敏的传感器不同，将是从单个生物分子的内部活动直接提取信号。美国 Wayne 州立大学医学院的生物传感器研究组利用一种称为细菌视紫红质(bacteriorhododopsin)的光敏细菌蛋白质。这种生物分子在光照下将释放出质子，并可转变成可测的微弱电子信号。实验表明，蛋白质分子的质子释放活动及其产生的信号与分子的化学环境有关，而且这种特定的化学灵敏度随光的波长而变。例如，当暴露在一种波长下，分子发射的信号随媒质酸度的增加而减小，而在另一波长下，分子的灵敏度仅仅与氯离子的浓度有关。利用这种结果可以制成双功能生物传感器，只要通过改变光的波长就能在测量酸度和氯浓度之间选择。

另一些科学家正在研究利用某些蛋白质的灵敏度来制造生物分子开关，并考虑在半导体技术中可以作为存储器。例如美国斯坦福大学的一位化学教授把希望寄托在另一种细菌蛋白质，它对光的响应是释放出电子。就自然过程来说，由这种“光合作用中心”转移的电子会产生化学反应，为器官的生命过程提供能量。这位教授设法使作用中心把电子传递给其他的蛋白质或直接给金属电极。因为这种光驱动的电子传递可用第二束光或外电场加以开关。从理论上看，这种蛋白质在电子或光电子回路中可以作为类晶体管元件用。这位教授的信念是，既然现在许多人都在忙于把常规半导体做的微电路不断缩小尺寸，为什么不可以用蛋白质来制造晶体管呢？他相信总有一天可以找到一种方法将生物分子元件在大范围内自己相互连接起来。用分子生物制造元件的真正好处在于它们能自组装(self-assemble)。他在做出自联线的生

物晶体管以前，目前正致力于将蛋白质粘结到电极和其他表面，通过遗传方法改变细菌，使其产生备有小分子挂钩的改性蛋白质。他指出一旦能在生物分子与电子元件之间能建造接口，就可以有许多事可做。

在 Syracuse 大学也有人在研究开发生物分子的计算机元件，但他并不想等待接口的出现。为避开对接口和联线的需要，他利用上述用来做传感器的细菌视紫红质。这种蛋白质当光照时除了发生电信号以外，也改变其光学性质。利用这一特性可做成全光计算机存储器。将细菌视紫红质在玻璃盘上形成薄膜，当用聚焦得很细的激光束照射时，就可在薄膜上将信号写入。读出是应用第二束激光确定哪一个光斑是被光学改变了的。原型样机实验表明其速度与最快的半导体存储器一样。他们估计当商品化后，其价格只有 1/10 还不到。

以上这些研究，在概念上是要使蛋白质具有目前常规计算机中非生物元件同样的功能。但是 Wayne 州立大学的计算机科学家和生物学家认为未来的生物分子计算机具有完全不同的工作原理。当然某些生物分子能处理电的或光的信号，但是蛋白质确是最善于识别和对相互的形式起反应，这就是抗体和酶如何发现它们的目标分子的原因。作为全"神经分子计算机"的第一步，他们建议把形状探测能力放入一种"瓶中计算机"(computer-in-a-jar)，使它能识别图形——这是计算机长期来的挑战。研究的方案是一种未辨别出的图形或图像，将决定释放到容器中的蛋白质混合体。例如一支铅笔的颜色，长而细的固体外形将释放一组特定的蛋白质，而电话机的不同外形将释放非常不同的另一组，这些蛋白质的每一组将自组装成特定的镶嵌结构。为了辨别这种镶嵌结构——也就是原始的图像，器件将监视一组各种各样的酶的活性，每一组酶倾向于固定在特定的镶嵌结构上。因此酶的活性将呈现出原始的图形：一支铅笔、一架电话机或是其他某种东西。虽然他们在实验室中已试验了一些原理，目前这种图形识别机仅仅是常规计算机上的模拟。

生物分子"纳机器人"像神经分子计算机一样目前还是一种设想，它的长期目标是要能通到人体各部分对受损的组织进行复杂的修复。美国亚历桑那大学一位教授花了二十多年的时间从事微管(microtubule)的研究，这是一种细长圆柱型的蛋白质分子，神经原和许多其他细胞都有这种结构。在各种化学或其他作用的影响下，蛋白质单元可以增加或失落，从而使微管生长、缩短或弯曲。当一微管的抽动或弯曲影响到相邻的微管时，将出现一种信号传遍细胞的微管网。他相信通过这种途径，微管提供一种细胞内部的智能，它能监视细胞的环境及触发其许多响应。他认为微管是天然的计算机，如果我们能懂得其编码和信息处理原理，并用某种遗传输入使它们存取，我们就能控制它们的行为。现在取得的成果是已能在计算机上对通过微管网络的信息流建模，并对部分模型进行测试。采用一具双针尖的 STM，其中一针尖用来对微管网络作电激励，另一针尖探测信号如何传播。一旦搞清了不同信号影响，他们尝试通过遗传改变组成的蛋白质和使网络接受特定的信号序列来对微管网络进行编程。

如果纳机器人需要推进器，美国 Utab 大学生物学教授研究了一种 25nm 宽的分子马达，它由细菌类似螺旋桨的许多鞭毛作为动力，最高转速达 18000r/min，这马达可推动平均尺寸为 30μm 的细胞。为了列出马达所有零件的清单，他已设法改变马达中蛋白质基因的编码并研究每种改变造成的影响，不过迄今唯一能识别的零件是"燃料喷射器"——能提供马达能源的质子通道。他希望能进一步弄清楚定子、转子、机壳及传动等所对应的零件。

为了将分子马达装到较大的结构上，或为了把开关、反应室以及附件进行联合装配，作为装置完善的生物纳工程必须要有适当的支架，有人认为 DNA 可充当此角色。将 DNA 多股分

子进行交织可以形成格子或网络结构。通过改变分子中的关键序列得到了 DNA 分枝的集合体并可折成立方体。可以相信还可获得更复杂的结构。下一步的工作希望把其他分子能加入到 DNA 中来,这样就能使具有电子性能的分子依靠 DNA 领先作支架而形成电路。

现在可以毫不怀疑,通过纳生物工程,即使要制造出自然中最精巧的复杂的产物——人脑的复制品也不是不可能的,这如同当时提出登上月球一样,似乎不可思议,但终究有一天能付诸现实。

6 分子机——纳技术在21世纪的奋斗目标[9~12]

从上面的论述可以看出,纳米尺度的材料和器件的制造有两种相反的途径,即从大到小或自上而下(top down)和从小到大或自下而上(bottom up)。所谓自上而下,就是在已有的微加工技术的基础上,不断改进,使适合于更小尺度的加工,从微米量级到亚微米量级再到纳米量级。自下而上的方法则完全不同,是要从操纵和控制原子和分子出发,把一个一个分子集合成宏观的材料、器件,甚至是系统,这是一种与现有的制造方法截然不同的新观念。本文一开始提到的 Feynman 就是代表这一新观念的先驱者之一。其他如 MIT 资深的教授 Hippel,他在 1959 年出版了《分子科学与分子工程》一书,1962 年在《科学》杂志发表了"材料的分子设计"专文,都是这一领域的经典著作。

80 年代中期以来活跃在这一前沿领域的学者中,在 MIT 获得分子纳技术领域博士学位的 Drexler 付出了巨大的努力,同时他在这方面的著作和观点也引起了极大的争议和轰动。Drexler 作为一个技术的理想家喜欢考虑的一个不容忽视的问题是世界将不能负担。他提出在几十年内,由于蛋白质尺度的机器可以一个一个分子制造出工程师在不违背自然规律的前提下设计的任何东西,从而使人类文化结构发生剧烈的重新安排。1991 年他在 36 岁时获得 MIT 博士学位,在这以前,他的第一部著作《创造机:即将来临的纳技术时代》(Engins of Creation)出版于 1986 年[9]。书中他设想了蛋白质和器官尺寸的机器,带来的变化将与工业革命、抗菌素和核武器是同样的巨大突破。在 Drexler 微细的新时代产品中有延长寿命的分子机,它能巡视人体的细胞、修复损坏的蛋白质;太阳能为动力的纳工厂,它能将暖室中的气体如二氧化碳从空气中排除,吐出氧气,并将碳原子仍返回它们从那出来的储煤及储油库中。5 年后,他又合作出版了《无限将来:纳技术革命》一书[10],以通俗的语言撰编了一个分子装配机的故事,这机器从原子和分子的配料开始,而后将它们变成这样的东西,例如,坚固的住房装备有完善的厨房设施和空调机,在将来,红十字会可以把它们运送到灾区作救济。不仅如此,他还发表许多技术论文以及包含各种学科的文章,在大学中培育纳技术研究小组,周游许多国家给以不计其数的讲演,甚至直接向国会提交谘文,因而在公众中获得了"纳技术先生"的称号。

与此同时,Drexler 也遭到了不少非议,甚至是激烈的谴责。认为此人的想法纯属科学幻想,认为他既没有发明出什么,也没有具体的验证,没有严格的科学依据。甚至有人说他是个科学骗子。英国权威的科学杂志《自然》的编辑认为 Drexler 所以在其他的同行中遭到轻视和贬低,是由于他几乎普遍地忽视了他所要操纵的物体的化学和物理性质。

如果说 Drexler 前两本书主要是一般定性的描述,而 1992 年出版的《纳系统:分子机,制造及计算》是一本 500 多页包括详尽的化学、物理基础和计算分析的技术教科书,是他纳技术充满血肉的蓝图,同时也回答了人们向他提出的一连串问题。正如他在前言中所说,这是他 15

年来研究分子纳技术的总结。此书为美国出版商协会(AAP)列为 1992 年最杰出的计算机科学图书。加州理工学院化学及应用物理教授 Goddard 评价说,Drexler 用这本书建立了分子纳技术领域。详细的分析给量子化学家和合成化学家表明了如何运用他们的化学键和分子知识来开发纳技术的制造系统,为物理学家和工程师显示了如何把他们宏观系统的概念降低到分子级水平。麻省理工学院的电气工程和计算机科学教授 Minsky 的评论是:器件较以前极度地缩小,将使工程、化学、医学和计算机技术重新造型。我们如何来理解机器可以这样小?《纳系统》一书概括了一切:动力和强度、摩擦及磨损、热噪声及量子不确定性。这本书是下世纪工程的开始。

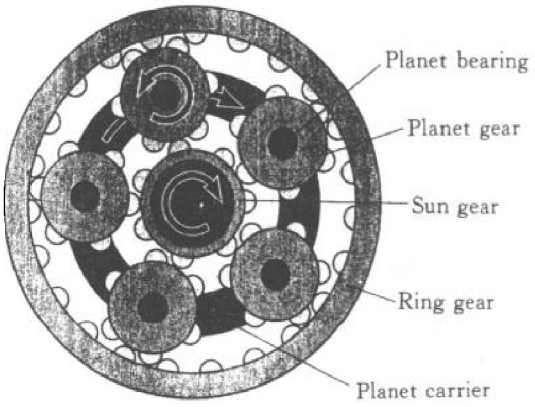


图 8　行星齿轮的机构图

《纳系统》全书共分为三部分,第一部分是物理原理,第二部分是元件和系统,第三部分是实施战略。书中对所讨论的问题都有理论的分析、定量的计算并给出清楚的图示。例如设计分子行星齿轮系统,图 8 表示这一系统的机构,中央高速转动的太阳齿轮驱动转动较慢的行星齿轮载体,如果当外围的环齿轮固定时。一般来说不论是太阳齿轮、环齿轮或是行星载体都可以被固定,用来约束其他两个元件的相对运动。图 9 是第一次完成的行星齿轮的原子设计,共有 3557

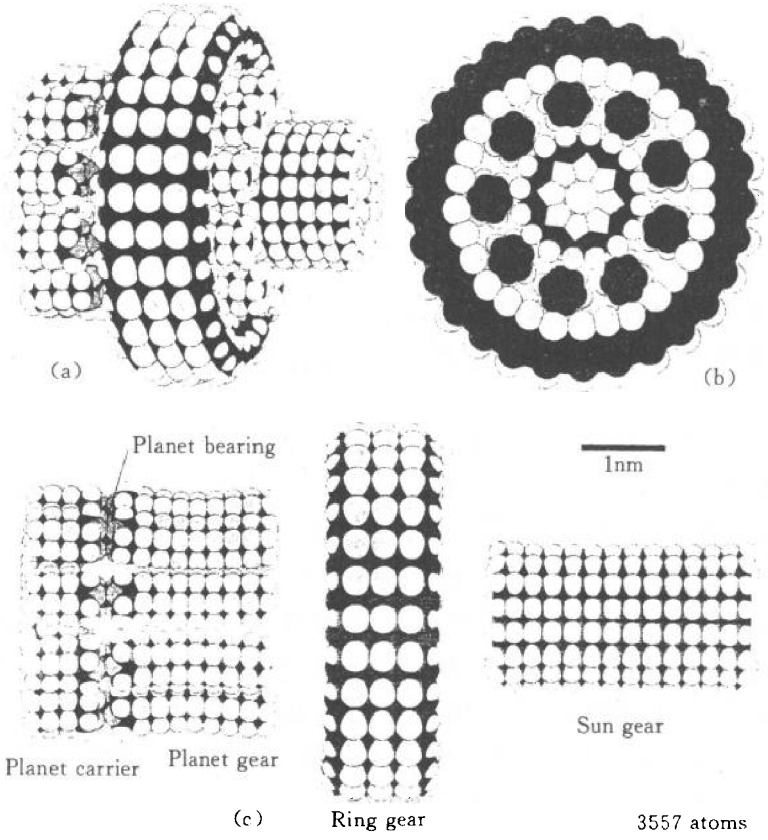


图 9　行星齿轮的原子结构

(a) 立体图; (b) 侧视图; (c) 分解图

个原子构成。早期的几个设计都不稳定，由于能量的最小化使齿轮从系统中挤出。这一设计是与 Xerox 研究中心的 Merkle 合作完成的，应用的软件是 Polygraf 分子造型软件。

又如提出的纳机械计算系统利用的逻辑器件是由滑杆构成的，其开关时间仅 0.1ns，每门消耗的能量$\ll KT_{300}$。寄存器单元的能量消耗可接近理论最小值 $1n(2)KT$。逻辑杆和寄存器可连接起来构成寄存器—寄存器联合逻辑系统，四个寄存器—寄存器的传递时间为 1.2ns，这一性能允许纳机械 RISC 机的时钟速度可达 1GHz，执行指令速率达 1000MIPS。包含 10^6 类似于晶体管的联锁(interlock)的 CPU 系统占有尺度为 400nm 的立方体，对时钟、功率供应和冷却都作了描述和分析，对于 1GHz 的 CPU 系统，其功耗为 1nW，每瓦每秒执行的指令大于 10^{16}。图 10 和图 11 表示这种滑杆逻辑门原理。图 10 是或门逻辑，任意一输入杆的位移都可带动输出杆的位移而不影响另一输入杆。与门的工作原理示于图 11，其中只有当二输入杆同时作用时才能使输出杆位移。

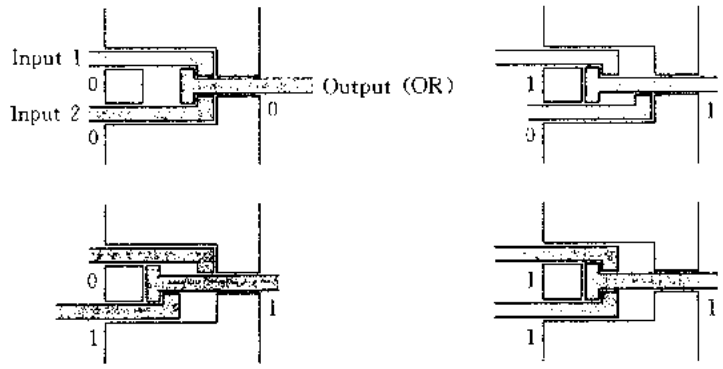


图 10　或门操作原理

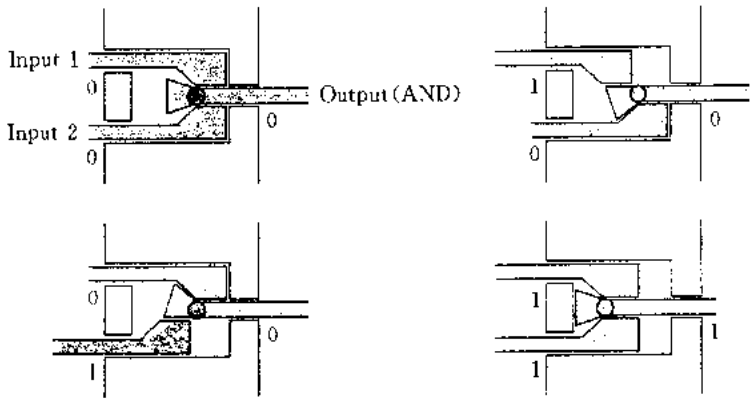


图 11　与门操作原理

在讨论分子制造系统时，首先分别研究了分子的筛选、运输和装配的机构，而后进行系统集成。图 12 表示一种转动式的分子筛选装置，在转子上有调节接收器，并通过表面来控制，把外面容器内的分子筛选到里面的容器中，而不能返回过去。图 13 表示从液相中取出分子的机构，并将它们共价地键合到移动带上。图中省略了接受器的调节性质。图 14 表示装配器的外形及其运动的范围。这种装配器呈空管形，可以弯曲内部的空间用于驱动机构。运动需要有连接器，还要有屈从器。套筒式的连接器并带有刻有螺纹的接口可连续地调节长度，倾斜的连接器用来使臂旋转一定角度。它具有最大的伸展长度为 100nm，直径为 30nm，

典型的壁厚 3.5nm，不包括基座在内，大约需要 4×10^7 个原子构成，图 15 是它的截面结构及其运动范围。

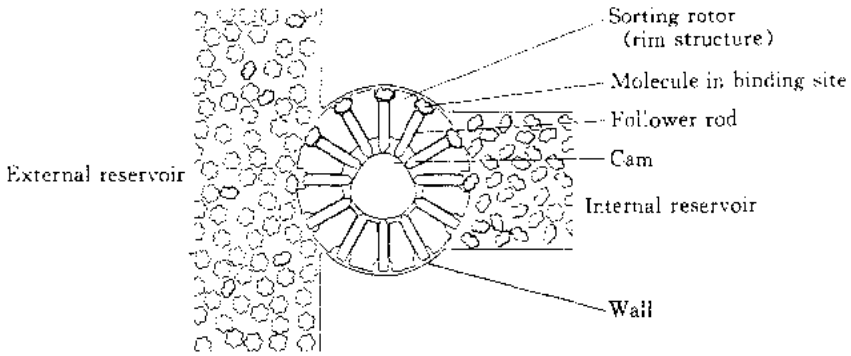


图 12　带有调节接收器的筛选器

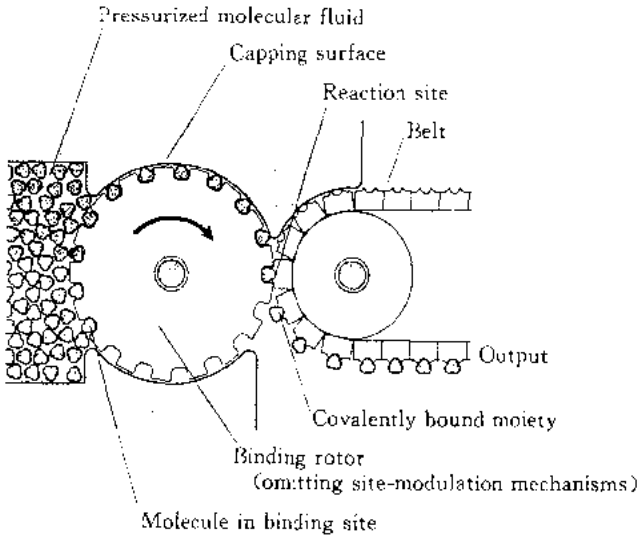


图 13　传送带式的卸载和输送装置

作为一个分子制造系统的实例，它可以用来制造出宏观的物体，子系统的容量是这样设计的，使包含有机小分子的原料液在 1h 内转变成 1kg，尺寸为 0.2m 的物体。材料的供应需通过分子筛选器，定向排列，几个阶段的装配等。

1993 年 11 月 Rice 大学宣告一项“纳米技术创始”，基本思想是在休斯顿的校园里建造一幢新大楼，将各领域和系的纳学科的专家集中在一起。这里的研究人员将创造新王国——未来的分子、结构和纳机器的基础工作、这个创始的主要发起人是 C_{60} 的共同发现人 Smalley。他公开宣称他是 Drexler 迷。

他认为基于纳米尺度的科学和技术很可能成为 21 世纪最重要的技术之一。甚至可能是最最重要的。经过在整个大学层次上深入地研究发现它在不同系和学科之间有着明显协同方面，事实上已漂亮地跨越了工程-科学的壁垒。并且因为我们发现这是有效的重整旗鼓，把我们的思想引导到为将来吸引新成员，建立仪器中心，最终，我们希望引入大学的教育中。为什么应该教育学生成为古老技术的科学家和工程师？他们应该成为将来的一部分。

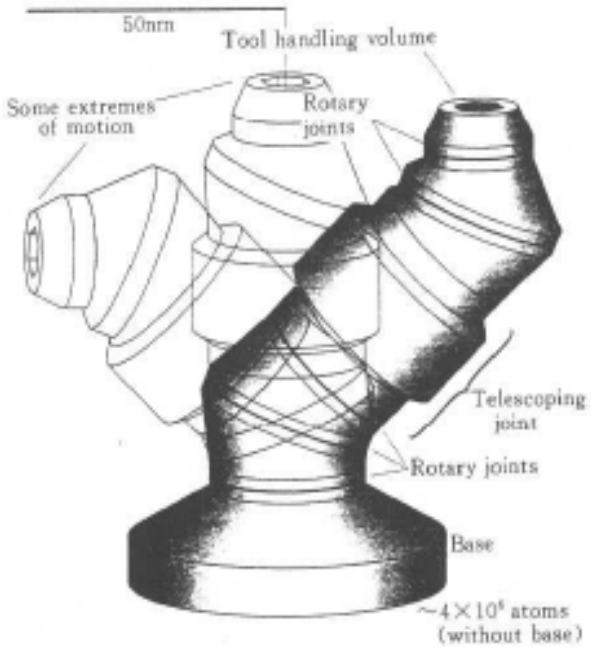


图 14　装配器的外形图

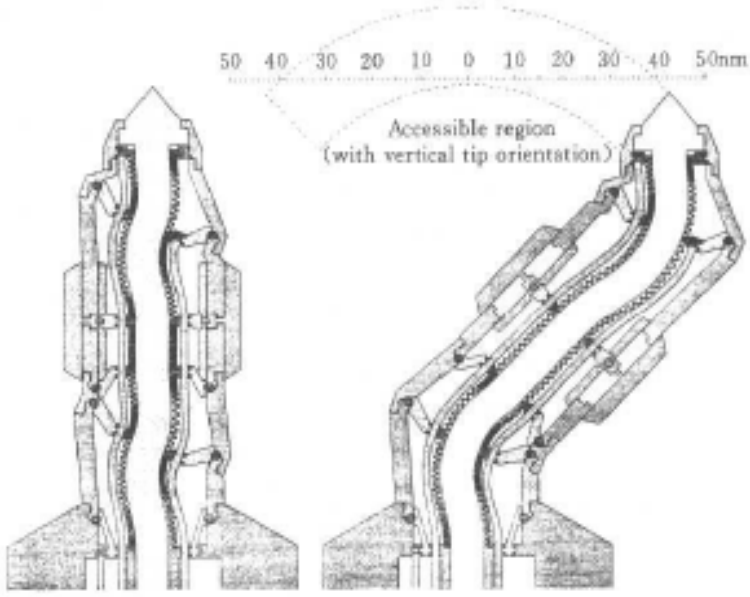


图 15　装配器的截面结构

参 考 文 献

[1] T. Appenzeller. The Man Who Dare to Think Small. Science, 1991,254(5036):1300

[2] G. Binning, H. Rohrer. Scanning Tunneling Microscopy. Surface Science,1985,152/153:17~26

[3] J.A. Stroscio, D.M. Eigler. Atomic and Molecular Manipulation with the Scanning Tunneling Microscope. Science, 1991,254:1319

[4] L.T. Whitman, J.A. Stroscio, R.A. Dragoset, R.J. Celotta, Science,1991,251:1206

[5] I.W. Lyo, P. Avouris. Science, 1991,253:173
[6] D.M. Eigler, C.P. Lutz, W.E. Rudge. An Atomic Switch Realized with the Scanning Tunneling Microscope. Nature,1991,352:600~603
[7] R.T. Bate. Nanoelectronics. Nanotechnology,1990,1:1
[8] D.H. Freedman. Exploiting the Nanotechnology of Life. Science,1991,254:1308
[9] K.E. Drexler. Engines of Creation. New York: Doubleday,1986
[10] K.E. Drexler, C. Peterson, G. Pergamit. Undounding the Future: The Nanotechlonogy Revolution. New York: William Morrow,1991
[11] K.E. Drexler. Nanosystems: Molecular Machinery, Manufacturing, and Computation. New York: John Wiley and Sons,1992
[12] E. Regis. Nano: the Emerging Science of Nanotechnology. Little: Brown and Company,1995

近场扫描光学显微镜中探针优化设计的考虑

0 引言

最近几年来，由于超高分辨率近场扫描光学显微镜(NSOM)的出现，在生物化学和半导体器件制造方面已取得了令人瞩目的进展。由于 NSOM 具有光学显微镜和扫描隧道显微镜(STM)[1]的诸多优点，从而引起国际上许多研究人员的足够重视。这种 NSOM 的分辨率突破了衍射极限，它的分辨率主要取决于其中所采用的探针尺度[2](即孔径尺度)。因此加工理想的扫描探针是提高 NSOM 分辨率的关键。

目前，扫描探针一般都用光纤制成，主要采用以下两种方法：一种是采用腐蚀法[3,4]；另一种是采用熔融拉锥法[5]。对于一个理想的扫描探针，它应具备以下几个特点：(1)从原始光纤传输到探针顶点的光能损耗要尽可能地小，最好是绝热(adiabatic)[6]探针；(2)探针具有最好的集光本领；(3)使用上要比较方便。为此，本文将详细讨论具体问题，并提出研制 NSOM 探针的一种新方法。

利用腐蚀法研制 NSOM 探针时，对光纤的要求是除了须选用单模光纤之外，还应考虑到所采用光纤的制作工艺。NSOM 探针的形状主要是依赖于 $NH_4F \cdot H_2O$ 混合液体对光纤芯内不同掺锗浓度的腐蚀速率。一般来说，掺锗浓度越高，腐蚀速率越低；反之，就越高。当单模光纤中掺锗浓度如图 1(a)分布时，才能形成尺度较小的 NSOM 探针。这种光纤采用 OVD 方法[7]。如采用 MCVD 方法[8]制造的光纤，由于工艺上的原因，在光纤芯内或多或少地存在着中心区锗离子的降低(见图 1(b))，因而单单用腐蚀法就很难得到几十纳米数量级的 NSOM 探针。另外需要指出的是，只用腐蚀法研制成的大渐变度探针，在使用中很可能导致探针的边缘与被测样品发生接触[9]，这样既影响测量精度，又可能损坏探针。用熔拉法制作的探针已在 NSOM 中获得了较成功的应用，但是由于这种锥形探头大都用高压(1～2 kV)电弧火花或 CO_2 激光器来熔拉，因此，这种光纤探针的特点是光纤包层渐变度极大，一般在 10°～20°，而纤芯的渐变度也可达 0.69°～1.38°。这些值都已远远超过了抑制高阶模激发的渐变条件，从而导致激发出许多高阶模，如 LP_{11}、LP_{02}、LP_{21}，使得光纤中基模与它们发生强烈的耦合，引起损耗剧烈上升。图 2 是一个渐变区长度约为 310μm 光纤的输出功率谱响应，其顶点的尺度约为 1μm 左右。从中可以看出，在测量波长的范围内已产生大于 20dB 的损耗。可想而知，当光纤顶点尺度达到几十纳米数量级时，产生的损耗将会大于 50dB[5]。为此，我们将建议一种新的方法，即先采用微小火焰进行熔融拉

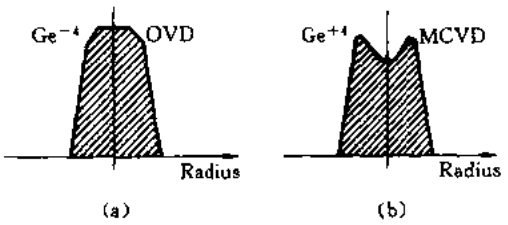


图 1 TiHe Ge^{-4} distribution in fiber core in which(a) is manufactrued by OVD method and (b) is by MCVD method

锥，然后再进行腐蚀方法。用这种方法，可以避免上述两种方法各自所导致的不足之处。

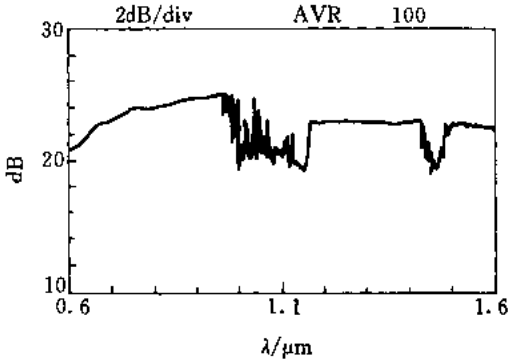


图 2 Shows spectral loss a fiber probe with the length of 300 μm

1 微米区渐变光纤的特性

1.1 理论分析

根据上面所提及的不足，我们建议一种研制光纤探针的新方法，即“先熔拉后腐蚀”的方法。用这种方法所研制的探针一般具有如图 3 所示的外形。这里可以把它分成三个区域进行讨论：(1)未拉锥区(即原始光纤)；(2)微米级尺度锥区；(3)纳米级尺度锥区。在此，重点分析第二和第三区域中光的传输及其接收光子的能力。

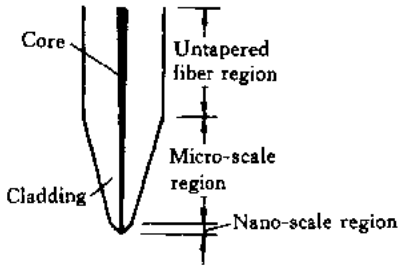


图 3 The sketch of the purposed optical fiber probe

为了抑制渐变区高阶模的激发，必须确保渐变度足够小，但是不可能把探针做得很长，这样会影响 NSOM 的稳定性。为此，专门加工一个微小火焰(microburner)，其最小加热区可达 1mm 左右。利用这样一个微小火焰制作出的锥形光纤，在渐变区只发生 LP_{01} 和 LP_{11} 模之间的相互耦合，其他模式的作用可以忽略不计。基于 W. J. Stewart & J. D. Love 的绝热锥形光纤的判据[6]，适当控制微米尺度光纤的锥度和长度，就可获得具有较小传输损耗的渐变光纤。A. W. Snyder 已经证明了在渐变的阶跃光纤中，LP_{01} 和 LP_{11} 模之间的耦合系数可表示为[10]

$$K = \frac{\lambda}{\rho(z)} \frac{\mathrm{d}\rho(z)}{\mathrm{d}z} \tag{1}$$

式中：$\rho(z)$ 为光纤的本地半径；z 是沿锥形光纤的传输距离。而且渐变区中 LP_{01} 模和 LP_{11} 模的传输常数差可近似用下式表示：

$$\Delta\beta = \lambda \frac{(5.520^2 - 2.405^2)}{4\pi} \cdot \frac{\exp(-2/V)}{n_{\mathrm{clad}}\rho^2} \tag{2}$$

式中：$V = (2\pi\rho/\lambda)(n_{\mathrm{clad}}^2 - n_{\mathrm{ext}}^2)^{1/2}$ 是归一化频率，n_{clad}、n_{ext} 分别是光纤包层的折射率和其外部折射率(一般为空气，$n_{\mathrm{ext}} = 1$)，可得到 LP_{01} 模在渐变光纤中所传输的光功率为

$$P_{LP_{01}} = 1 - K/\sqrt{\beta^2 + K^2} \cdot \sin^2(z \cdot \sqrt{\beta^2 + K^2}) \tag{3}$$

假定所研制的渐变光纤的本地半径满足

$$\rho(z)=0.5[(\rho_1+\rho_2)-(\rho_2-\rho_1)\cos(\pi z/L)] \tag{4}$$

则其形状如图4所示。在忽略泄漏损耗的情况下，LP_{01}模的传输功率如图5所示。从图中可以看出两个特点，一是锥形光纤的渐变度越大，LP_{01}和LP_{11}模之间的最大耦合功率就越大；二是$\rho(z)$越小，LP_{01}和LP_{11}模之间的耦合捭长就越小。这与文献[11]的结果是一致的。

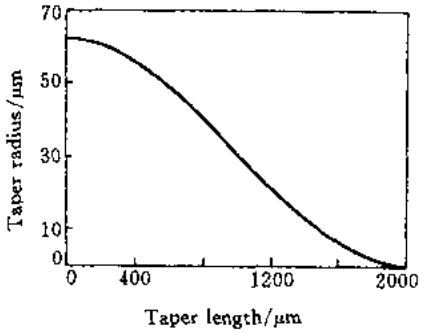


图4 The outline of taper fiber to be fabricated

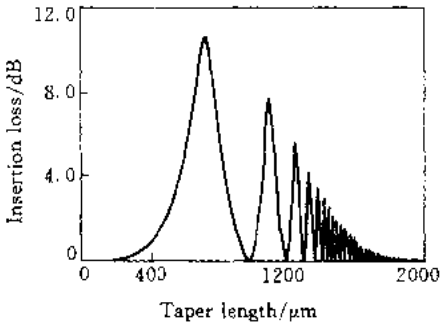


图5 Optical power propagating in the region of taper fiber

根据赛尔梅耶公式[12]，得到了给定锥形光纤头的波长特性，如图6所示。从图中可看出，随着波长的增长，LP_{01}和LP_{11}模之间的耦合逐步增强，其波长灵敏度不仅与锥形光纤的渐变度有关，而且与锥形光纤的长度有关。因此，在制作锥形光纤时要特别考虑其工作波长，以使得锥形光纤头具有最大的输出功率，即要做好图3中微米尺度区的光纤。

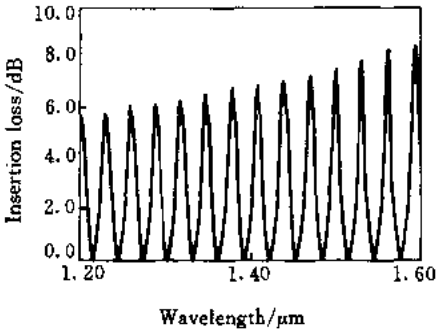


图6 Spectral loss of LP_{01} mode

1.2 实验结果与讨论

根据上述理论的分析和计算，首先设计了一个加热温区范围可变的紫源，其最小加热宽度达到1mm左右。以丁烷为燃料，氧气为助燃剂，在重力作用下进行对称拉锥，并用pin管探测器对光纤的输出功率进行监视，以保证在双锥光纤腰部割断后具有最大的输出功率，其实验简图如图7所示。在实验中，分别采用约为2mm，1.5mm，1mm的火焰宽度做了试验，并进行了比较。火焰宽度大时，在锥形光纤中，LP_{01}和LP_{11}的耦合较弱，渐变区较长。火焰宽度达到1mm时，LP_{01}，LP_{11}模之间的耦合非常强，最大耦合功率达85%。图8是一个实际渐变光纤的波长响应。使用火焰宽度约1mm，单锥长只有2.5mm，尖端的尺度达到1μm以下。从图中可清楚地看到，即使它尺度小于1μm，但是以特定的波长而言，LP_{01}模的传输损耗居然不足1dB。这与用高压电弧火花拉锥得到的结果相比，的确大相径庭。因此，只要在NSOM的工作波长处监测渐变光纤在拉锥过程中输出功率的变化，并适当控制拉锥长度，那么在微米尺度区一定能获得低损耗的渐变光纤。

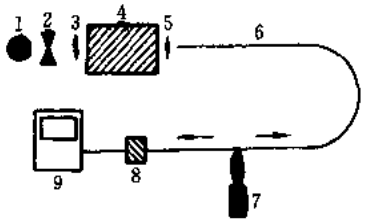


图7 The set-up of fabricating tapered optical fiber
1. white-light source, 2. Modulator
3. Lens－1, 4. Monochromer
5. Lens－2, 6. Optical fiber
7. Micro-burner, 8. Detector, 9. mV meter

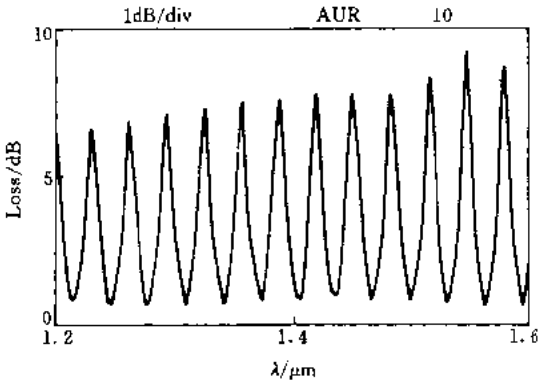


图 8 The spectral characteristics of a practical tapered fiber with length of 2.5mm

2 纳米尺度锥区的光纤特性

这一区域的形状主要是通过腐蚀法获得的，其特点是总长度一般不超过几个纳米。因此在这一区域可以忽略由于高阶模耦合所引起的损耗，其能量损耗主要是以辐射损耗的形式表现出来的。从直径 1μm 降到 50nm 的过程中，损耗一下子增加 2～3 个数量级，从探针顶点输出的能量基本上与其面积成正比。因此，我们对这一区域光的传输特性不作详细研究，而主要讨论它的集光本领。

2.1 探针的集光本领

目前着重于这方面研究发表的论文较少，主要着眼于研制探针。由于探针研制的工艺不同，在计算其集光本领时所建立的模型也有所不同。为了计算这种探针的集光本领，建立了如图 9 所示的模型。图中：E_1 为单色光场；n_1，n_2，n_3 分别为不同区域的折射率。为简化起见，假设 n_1 与 n_2 的界面是绝对光滑，而且光场以 φ_1 角到达 n_1 和 n_2 的界面时，满足内反射条件。则由于光子隧道效应而被锥角为 θ 的探针所收集到的光场 E_3 与 E_1 的强度之比为

$$R(x)=\left|\frac{E_3}{E_1}\right|^2=\frac{4\gamma^2n^2\cos\varphi_1}{4\gamma^2n^2\cos^2\varphi_1-(1-n^2)^2\sinh(2\pi\gamma D/\lambda)} \tag{5}$$

式中：假定 $n_1=n_3=n=1.458$；$n_2=1$；G 为工作波长；$\gamma=[(n\sin(\varphi_1))^2-1]^2$。需要指出的是，假设 E_1，E_2，E_3 的波矢 $\boldsymbol{K}$ 是不随 θ 变化的。因此，探针从消逝场中所接收到的隧道光子总功率为

$$P=\frac{n\varepsilon\boldsymbol{K}}{4\pi}\mid E_1\mid^2\int_0^{2\pi}\cos\zeta d\zeta\int_0^{\rho}R(x)\mathrm{d}x \tag{6}$$

式中：ε 为介电常数；ρ 为纳米区探针的最大半径(见图 9)。通过计算式(6)，得到具有不同锥角探针接收隧道光子的相对强度，见图 10。从中可以发现：(1)锥角越小，接收功率越小，反之亦然；(2)探针与样品之间距离较小时，接收功率较大；(3)锥角较大时，存在一个最佳接收距离；(4)探针与样品之间距离小于工作波长时，才会出现隧道光子；(5)根据不同探针接收隧道光子的特点，可以认为锥角越小，它的垂直分辨率越低，反之就越高。这些结论对于具体研制纳米尺度区探针具有重要的参考价值。

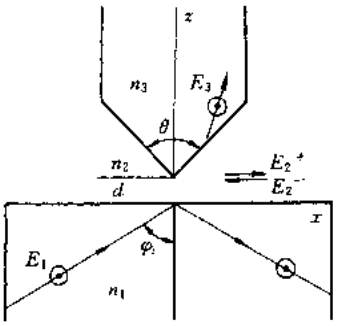


图 9 Model of the estimntiom pick-up efficiency of the probe with the cone

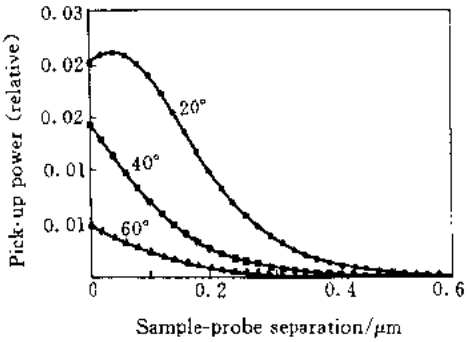


图 10 Pick-up power versus sampleprobe separation

2.2 纳米尺度区光纤锥度的控制

纳米尺度区渐变光纤的研制是极为重要的，它决定了 NSOM 的垂直和水平方向的分辨率。通常选用氢氟酸(HF)作为腐蚀剂。为了便于调整锥角，选用 NH_4F, HF 和 H_2O 的混合液。有文献报道[3]，假定三种液体(NH_4F, HF, H_2O)的组分比为 $x:1:1$，当 $x=3$ 时，获得锥角 $\theta=45°$；$x=5$ 时，$\theta=25°$；$x=10$ 时，$\theta=20°$，误差仅为的 0.5°。因此利用这一方法是完全可行的。一方面可提高纳米尺度锥形区的导光能力，提高系统接收的信噪比(SNR)；另一方面可根据测量样品的要求，通过探针制作工艺的第二步(腐蚀)来调节探针顶点的锥角(cone)，从而调整 NSOM 的垂直及水平方向的分辨率[9]。

3 结论

对 NSOM 探针的性能进行了较细致的分析，提出了 NSOM 探针的三段模型，并利用锥形光纤的绝热判据分析了微米尺度区域锥形光纤的传输特性，另外还利用隧道光子模型讨论了探针的集光本领，得到了一些具有重要参考价值的结论。同时，我们还建议了一种加工 NSOM 探针的新方法，可望探针的传输损耗能降低 1～2 个数量级，这对提高 NSOM 的分辨率有着重要的实际意义。

参 考 文 献

[1] G. Bining, et al. Phys. Rev. Lett., 1982, 57～60
[2] Pat Moyer, et al. Laser Focus World, 1993, 105～107
[3] T. Pangaribuan, et al. Electron. Lett., 1993, 1978～1979
[4] Shudong Jiang, et al. Jan. J. Appl. Phys., 1991, 2107～2111
[5] E. Betzig, et al. Science, 1991, 1468～1470
[6] W. J. Steward, et al. Proc. ECOC'85, 1985, 559～562
[7] Ting Ye Li, Fiber Fabrication, ACADEMIC PRE. INC., 1985
[8] MacChesney, et al. Proc. IEEE, 1974, 1278～1281
[9] Shudongliang, et al. Jan. J. Appl. Phys. (Part 1), 1992, 2282～2287

[10] A. W. Snyder. IEEE Trans. Microwave Theory Tech., 1970, 383 ~ 388
[11] D. Marcuse, J. Lightwave Tech., 1987, 125 ~ 133
[12] 范崇澄, 等. 导波光学. 北京: 北京理工大学出版社, 1988, 185 ~ 187

微光电机械系统的技术和应用

0 引言

光子作为信息载体在光通信、光存储和平面显示等方面所产生的巨大效果离不开光子和光电子器件，制造这些器件又离不开微加工(microfabrication)技术。微加工技术的出现和发展原是由于微电子器件的需求，但是这种加工技术上的创新给人类科学技术的进步带来了难以估量的影响，不仅在今天，而且完全可以预计，当进入21世纪后，它仍将具有强大的生命力。

早在60年代，几乎正当微加工技术用于发展集成光学和光电子器件的同时，把微加工技术用于制造微机械结构的设想已被提出来了。由于硅不仅是理想的半导体材料，而且也具有良好的力学和机械性能，因而可以采用制造半导体集成电路的微加工技术，同样可以用硅材料制成微型的机械另件或装置，例如传感器和执行器等。而且这种尺寸相兼容的微机械装置可以与微电子器件在共同的硅基板上集成在一起，进行对信号的接收和放大或对机构的驱动和调节。这称为微电(子)机(械)系统(Micro-Electro-Mechanics Systems, MEMS)。现在已有大量MEMS集成器件被开发，如微齿轮泵、汽动涡轮、微马达等，有些已得到实际应用，如微加速度传感器和微压力传感器等。

微光电机械系统(Micro-Opto-Electro-Mechanics Systems, MOEMS)就是把光子器件、微电子器件和微机械结构或装置采用相兼容的基板材料及微加工技术集成在一起。光子、电子和机械三种元器件相互集成以形成一个完整的体系，如图1所示。这种把微光子微电子和微机械三者有机集成在一起，能最充分地显示这三类器件的综合性能，不仅能使系统结构进一步小型化，而且可能导致新一代器件和装置的诞生，如三维集成器件等。

作为信息载体的光子与电子相比，不仅由于其速度快、信息容量大，更由于其固有的平行性，这是电子所无法比拟的。但现有的光子和光电子集成器件与集成电路相似基本上是平面结构，无法实现光子所固有的空间平行性的特点。因此，目前光纤、光波导器件、光电子器件、平面光栅列阵器件和全息器件以及集成电路芯片上和芯片间的光对准、耦合和互连基本上还要靠人工装配和调整。采用MOEMS就可能由精密的微机械结构来保证，免除人工操作，并且可以大大减少工时、提高成品率、保证质量和降低费用。

本文将讨论MOEMS所应用的主要微加工技术，并列举MOEMS几种典型器件和装置的研究和应用的最新进展。

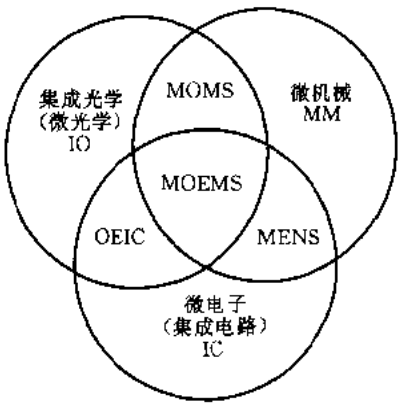


图1 微光电机械系统(MOEMS)

1 硅微机械制造技术

用于 MOEMS 最基本的硅微机械制造(micromahining)可分为体微机械制造(bulkmicromach - ining)和表面微机械制造(surface micromachining)两类[1],前者是对硅的体材料(单晶基片)采用各向异性腐蚀剂,通过湿法腐蚀而获得所要的几何图形和尺寸的体结构,利用这种方法做成的 MEMS 器件如微压力传感器、微加速度计等。实际上,大量用于光纤与平面波导集成光学器件以及其他光子、光电子器件对准耦合装置的 V-槽,就是采用这种体微机械制造方法制作的。

表面微机械制造是在硅基板上通过蒸镀、溅射或化学汽相淀积(CVD)及多次光刻形成多层膜图形,然后把作为中间支撑的牺牲层(sacrificial layer)材料除去而保留所需要结构。这种技术可以不需要经过装配将多种部件一次制成,包括具有可动部件复杂的微结构。例如微绞链的制造需应用这种方法,其加工步骤如图 2 所示[2],包括对两层多晶硅及氧化物牺牲层共三次光刻。首先是在基片表面上淀积一层磷硅玻璃(PSG)作为牺牲层,而后淀积第一层非掺杂的多晶硅(Poly-1),经第一次光刻和刻蚀后,再淀积一层掺杂的 PSG,见图 2(a)。所有薄膜都采用低压化学汽相淀积(LPCVD)并通过退火以减少多晶硅层的应力。作为限制绞链销绞的搭扣需要固定在基片上,对两层氧化层光刻和刻蚀形成接触孔后,再淀积第二层多晶硅在开孔处与基片相连,然后经第三次光刻和刻蚀形成搭扣如图 2(b)所示。最后,在氢氟酸中把牺牲层全腐蚀掉,绞链销就能在搭扣中自由转动,如图 2(c)和(d)所示。

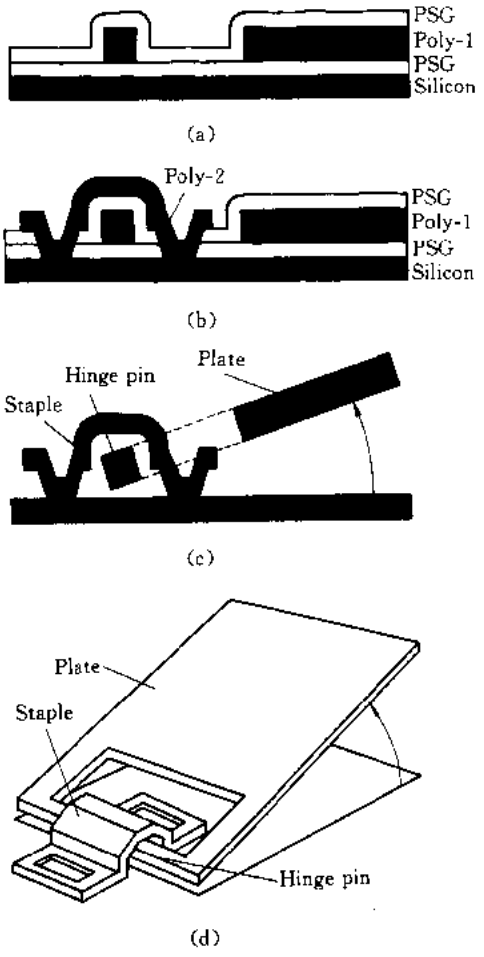


图 2 微绞链的表面微机械加工
(a)首先在硅基片上沉积磷硅玻璃,然后淀积第一层多晶硅并光刻出图形,最后第二次淀积磷硅玻璃;(b)将接触处的二层 PSG 刻蚀去,而后淀积第二层多晶硅并光刻出图形;(c)在牺牲层刻蚀时将全部氧化物除去,第一层多晶硅可自由地从基片平面上转出;(d)释放后绞链平板的立体图

2 LIGA 制造技术[3]

LIGA 是 Lithograpic Galvanoformung 和 Abformung 三个词的简写,是 X-射线深度光刻、电铸和塑铸三种工艺的组合。80 年代初由德国卡尔斯鲁厄(Karlsuhe)核研究中心发明并取得专利。

LIGA 技术的特点是利用 X-射线的深穿透能力制作出高度(深度)和高宽比很大的精细结构,其主要工艺过程如图 3 所示。首先由 X-射线通过 X-射线掩膜板对抗蚀剂曝光,经显影后形成抗蚀剂结构,通过导电的基片对抗蚀剂结构进行电铸而获得金属结构,去除抗蚀剂后得到金属模。利用这金属模进行微塑铸又可得到塑料器件或塑料模,利用塑料模可进行第二次微

电铸而大批量制造。

用 LIGA 技术可以制造出由许多种金属、塑料和陶瓷微器件，目前其厚度已可达 1mm，高宽比可超过 100，且垂直度好，侧壁光滑。如果与上面介绍的牺牲层工艺相结合，也可获得活动的微结构，这是常规的微加工技术难以实现的。

推广 LIGA 技术遇到的困难是必须利用昂贵的同步辐射装置作为 X-射线源，另外 X-射线掩模板也很难制作。对于厚度不太高的结构可以采用常规的紫外光源代替同步辐射 X-射线源，并配合使用特种抗蚀剂，目前已能获得厚度达 150μm 的微结构。被称为准 LIGA 技术，可以大大降低设备投资和产品成本。

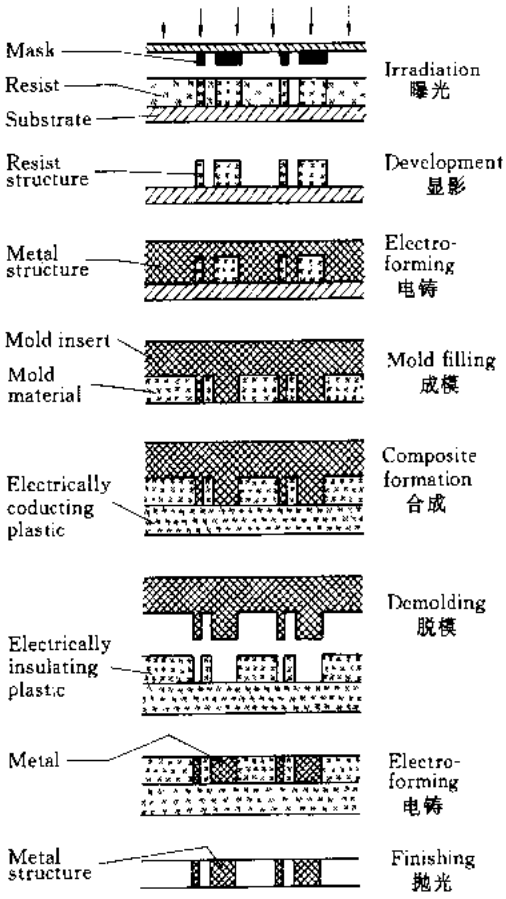


图 3　LIGA 技术的工艺过程

LIGA 技术在制造微光器件、集成光学和 MOEMS 中有广阔的应用前景。图 4 所示的光波导多路复用器件就是 LIGA 技术应用于集成光路的一个实例[4]。有机玻璃聚合物 PMMA 在可见光到近红外波段范围具有优良的光学性质。采用 LIGA 技术可以构成非常光滑和精确的侧面，因而可用于制造许多集成光路的微光器件。应用折射率不同的三层 PMMA 构成平面光波导，在光输入端的另一方面用 LIGA 技术制成自聚焦反射光栅。对于不同波长，反射光束有不同的偏转角，与不同波长的接受光栅相对应，从而构成多路复用器或多路解调器。这种微型频谱分析器可用于光纤通信中的 WDM 系统以及其他测量和系统分析。

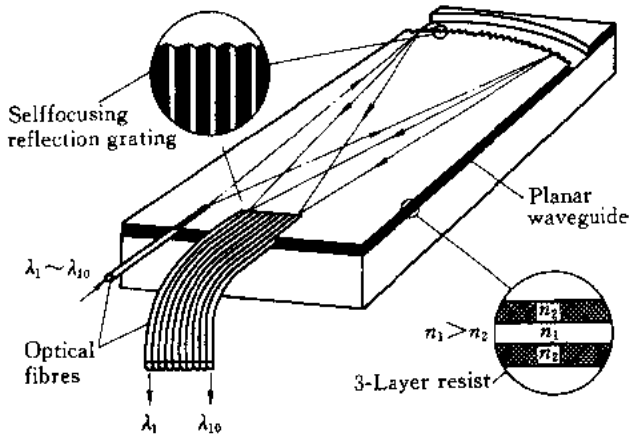


图 4　用 LIGA 技术制作的 PMMA 平面光波导型多路复用器件

3 微机械光调制器

用微机械结构对光束调制的先驱研究开始于60年代后期[5,6]。调制的形式包括改变光束的位相、光束扫描、干涉效应等，实现方法可通过变形膜、偏转镜或改变反射率等。这类器件可以做成分立的单元件结构，也可以做成线阵或平面列阵结构，例如后面将讨论的数字微镜器件(DMD)是一种高密度集成的二维微机械空间光调制器。这类光调制器在光纤通信、光信息处理、全息显示、高清晰度电视及大屏幕投影等领域有广阔应用前景。

美国AT&T贝尔实验室近年来开发了称为MARS(Mechanical Anti-Reflection Switch)器件[7]，它的用途之一是在用户光纤网(FTTL)中的光调制器。在无源光网络系统PONS(Passive Optical Network System)中，只有交换局安装激光器，当信息从交换局向用户传送时，可以对激光器发射的光束直接调制，当用户端向交换局向用户传送时，可以对激光器发送来的没有调制的光束调制后返回交换局。这样一根光纤和一个激光器能完成双向通信。在大多数场合下，这种返回信息的传输速率只需几Mb/s。采用MARS调制器可以降低FTTL的造价，而且这种器件还具有稳定性好、插入损耗低、高对比度和对偏振不敏感等优点。

图5为MARS的工作原理图。使基板上方膜的厚度恰好为入射光束波长λ的1/4，当膜与基板间的空隙为$m\lambda/4$，其中m为奇数时，这一结构呈现高反射状态。当膜由静电吸引向基板移近了$\lambda/4$(即m变为偶数)时，器件将变为透射状态。多数情况$m=1$，膜的移动从$\lambda/4$到与基板相接触。

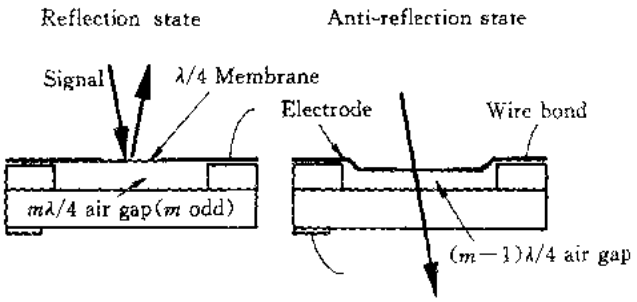


图5 MARS的工作原理

MARS的典型结构如图6所示。机械作用区包含中央平板及支持臂，光作用区由中央平板上的窗口确定。器件可以采用不同的材料体系，制作工艺随材料而异。例如其中有一种材料体系是由Al膜为牺牲层，在上面用PECVD生长一层SiN_x为反射膜，电极是钛和金双层膜并用剥离法(lift-off)做图形，SiN_x用SF和CCl F进行RIE刻蚀，最后用常规的Al腐蚀液除去牺牲层，使膜的作用区脱开。

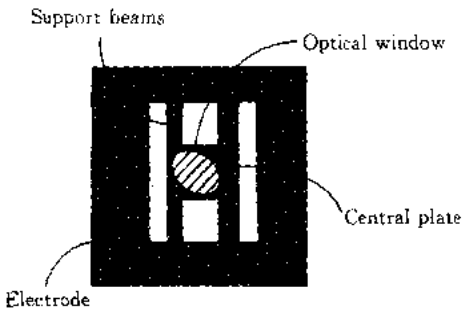


图6 MARS的结构

器件性能的测量表明上升和下降时间小于100ns，驱动电压为50V，从−50～90℃范围内光强对比度大于10dB。在500kHz运行2个月没有任何性能退化。当误码率为10^{-9}时，该器件可以在1.5Mb/s

工作。

采用 Fabry-Perot 腔的光调制器，其原理和结构如图 7 所示[8]。器件的制造采用表面微机械制造技术，全部工序用五套掩膜板。开始先在 Si 基板上淀积钝化抗反射涂层包括湿氧二氧化硅和 LPCVD 氮化硅。对抗反层刻蚀形成基板接触孔，而后用 LPCVD 淀积厚度为 278nm 的多晶硅并用等离子体刻蚀出图形，此多晶硅导电层作为 F-P 腔的第一镜面并充当底电极。下一步用 CVD 淀积 1.6μm 厚的掺磷的二氧化硅(PSG)作为牺牲层，并用氢氟酸刻蚀图形。接着是淀积第二层 LPCVD 多晶硅，为了保证足够的机械强度，厚度为 464nm。为了对多晶硅膜膜掺杂，在其上面再淀积一层 PSG 作为掺杂源，在 1050℃扩散和退火可以消除层间应力。经氢氟酸将牺牲层腐蚀使 F-P 腔的第二多晶硅镜膜释放，最后蒸镀 2μm 厚 Al 膜，并刻蚀图形作为连接线。对这一器件性能测试结果为：消光比 9dB，插入损耗 16dB，开关时间 180ns，最高调制速率 2.8Mb/s，对偏振不敏感，驱动电压不超过 90V。

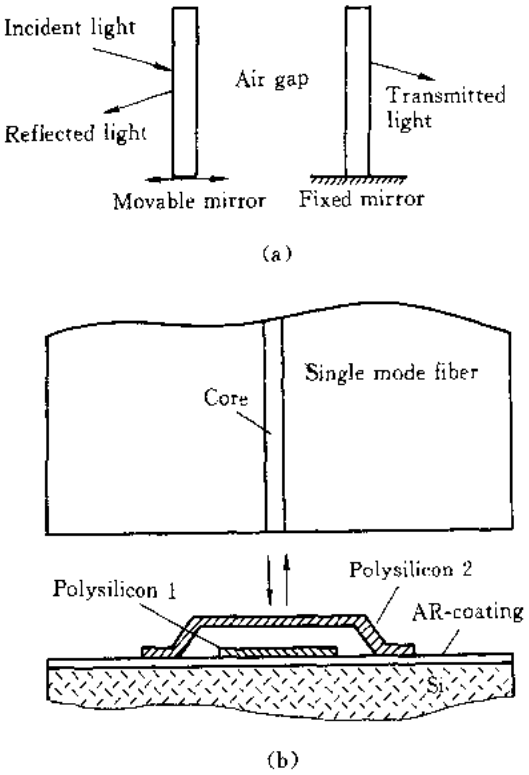


图 7

(a)由一固定镜面和一可动镜面构成的 F-P 腔，具有高反射低吸收；(b)微机械 F-P 光调制器的结构及与单模光纤的耦合

4 数字微镜器(DMD)

数字微镜器(Digital Micromirror Device，DMD)是一种通过微镜列阵反射而成像的二维空间光调制器(SLM)，由 Texas 仪器公司(TI)的工程师在该公司长期研究变形镜光阀的基础上于 1987 年发明的。DMD 中的每一个可动微镜分别覆盖在一个 CMOS 静态 RAM 存储单元上，由存储单元中数据决定的静电力使微镜扭转 +10°(开)或 -10°(关)，从而对入射到表面上的光束进行调制。从镜面反射的光束通过投影透镜在大屏幕上产生图像，目前屏幕的对角线已达 4.88m。微镜列阵结构用表面微机械制造方法加工，单元微镜及控制电极的结构及器件的扫描电镜照片如图 8 所示[9]。每存储单元上面有用来操纵微镜的二个地址电极和二个“着陆台”(landing pad)，在这些电极上面是铝合金的镜，通过扭转绞链由两个支柱支撑。微镜有三个状态，第三个状态即处于水平位置。实际上，在镜面与地址电极之间形成电容。+5V(数字 1)加在一个地址电极上，0V(数字 0)加在另一个地址电极上，负偏压加在镜面。在静电荷作用下，镜面转向 +5V 电极，当镜面碰到了“着陆台”时停下来，由于“着陆台”的限制，使镜面的偏转角度保持一定值，使 DMD 在整个面积上具有良好的均匀性。

DMD 的制造过程如下：首先采用 CMOS 标准工艺完成静态 RAM 和偏转地址电极。而后在整个硅片涂上一层聚合物，其厚度将决定反射镜离开硅片表面的高度。在安置支柱的地方将聚合物层刻蚀开孔形成接触面。接着为加工较薄的绞链铝层和较厚的镜面铝层进行淀积，光

刻和刻蚀出所需的图形和结构。最后通过等离子体刻蚀把聚合物层(牺牲层)全部除去。使镜面通过绞链片和支柱架空在硅片上,这高度刚好使镜面扭转并与硅片上"着陆台"接触时形成 +10°的偏角。

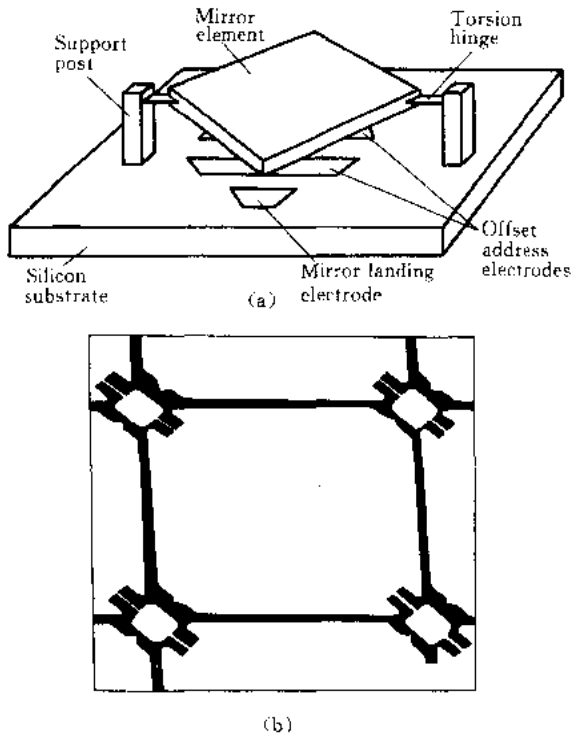


图 8

(a) DMD 单元镜面的结构; (b) 器件表面的扫描电镜照片

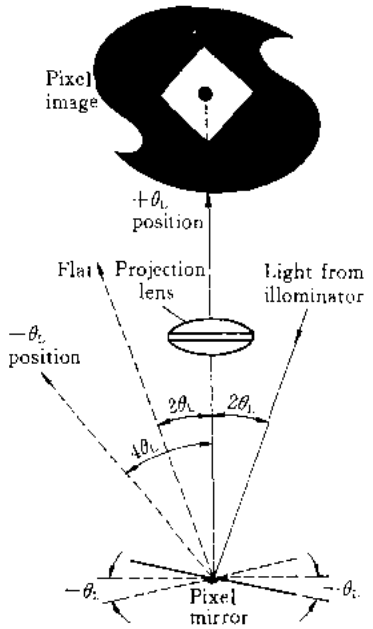


图 9 单元微镜在 +θ 角度将光束反射通过投影透镜成像原理

DMD 投影成像的原理如图 9 所示,当微镜单元在 +θ 角时,入射的光束通过镜面反射经投影透镜到达屏幕成像,当微镜单元在水平或 −θ 位置时反射光束偏离投影透镜。目前,每个镜面的尺寸为 16μm × 16μm,总面积为 37mm × 22mm,像素为 2048 × 1152,具有很高分辨率,可满足高清晰度电视的要求。

5 自由空间集成光学

自由空间集成光学与导波集成光学相比,其优点是具有高的空间带宽,无干涉的光路由二维空间互连及光信息处理(如傅里叶光学)的能力。但是大多数自由空间光元件的制造限于基片表面,难于在单一基片上进行空间光元件的集成。利用硅基片的微机械结构,可以实现自由空间光元件的集成。这硅基片充当"微光学平台"(Micro-Optical Bench, MOB)[10]作用,光路中的微透镜、反射镜、光栅及其他微光元件可以在 MOB 的掩膜设计时考虑其予对准,进一步的精密调节可通过 MOB 上的微调节器、微定位器、微转台来完成。如果有源光器件也集成在硅基片上,则可构成完整的光路系统,图 10 表示硅基片上的自由空间微机械集成光路系统。

Si 基片上的三维集成光路系统采用的表面微机械制造,首先在 Si 基片上淀积 2μm 厚磷硅玻璃(PSG-1)作为牺牲层。接着再淀积 2μm 厚多晶硅层(Poly-1),在这层上通过光刻和干法刻

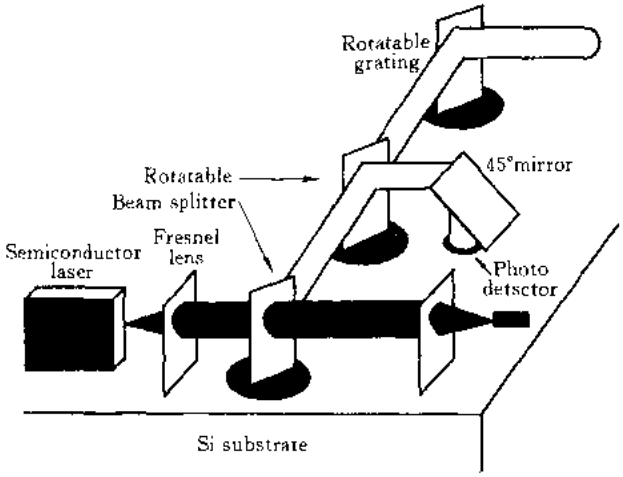


图 10　硅基片上的自由空间三维微机械集成光路系统

蚀形成微光元件图形，诸如 Fresnel 透镜、反射镜、分束器及光栅等。转动的绞链销也在此层中形成。在做好图形的 Poly-1 层上再淀积 0.5μm 厚的牺牲层 PSG-2。一些支架结构诸如搭扣(stape)、弹簧锁片(spring-latch)及扭力弹簧(torsion spring)等做在第二层多晶硅层(Poly-2)上，搭扣和扭力弹簧的底部固定在 Si 基片上，因此在淀积 Poly-2 前需穿过 PSG-1 和 PSG-2 形成接触孔。Poly-2 层上的结构与 Poly-1 层连结只需穿过 PSG-2 形成接触孔。采用氢氟酸将 PSG 层腐蚀除去后，置有微光元件的多晶硅小板能从基片上脱出，通过绞链销能在搭扣中自由转动。当小板立起后弹簧锁片的顶部滑进小板的槽内，并扣进槽的狭小部分，这就保证小板固定不动。三维微光元件装配后，在立直的多晶硅表面电镀金层。对二进位 Fresnel 透镜或反射镜，需镀较厚的金层以完全阻挡光束通过暗区或构成充分反射的镜面，而较薄的金层要求能部分透光或作为分束器。

图 11 是边缘发射激光器与微 Fresnel 透镜的自对准混合集成装置。微 Fresnel 透镜直径为 280μm，光轴离基片高 254μm，与集成的激光器发光点的高度相一致。为了保持透镜板与基片精确的角度，专门设计两片带有 V 槽的多晶硅精确定位夹，使透镜板的两侧嵌入 V 槽，增加了结构的机械强度和稳定度。图中表示的激光管是将侧边安装在 Si 基片上，因为激光的基片厚度是固定的，而侧边的宽度可精确切割，与透镜的光轴高相配合。激光管与透镜的自对准是通

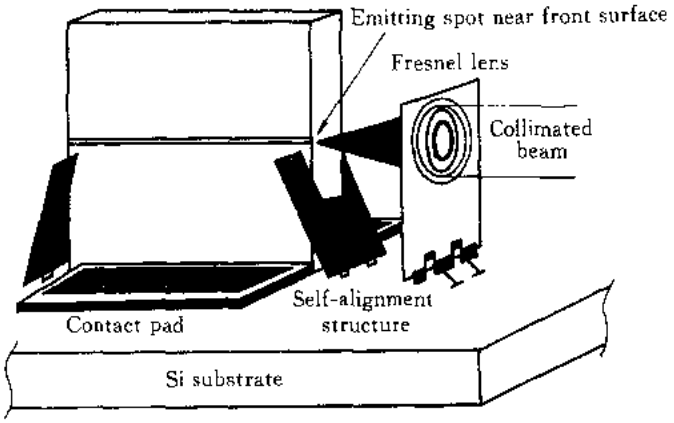


图 11　边缘发射激光器与微 Fresnsl 透镜的自对准混合集成系统

过多晶硅上光刻形成装配支架实现的，这包括由左右两片接触电极形成的凹槽和前后两片带槽的定位夹。

采用类似于微马达的结构[11]，可以在 Si 基片上形成微转台和微位移器。图 12 是安装在微转台上的微光栅扫描电镜照片。转台做在第一层多晶硅上，轴和轮毂做在第二层多晶硅上，光元件也做在第二层多晶硅上，对转台的成型不受影响，Poly-2 弹簧锁片及搭扣的底部在这结构中将与 Poly-1 转台相连接。带光元件的 Poly-2 板的下端连接到由 Poly-1 和 Poly-2 通过开孔构成的微绞链。除去牺牲层材料后，Poly-1 板可在 Si 基片上自由转动。基片上刻有 36 个记号用来指示转台的角度。

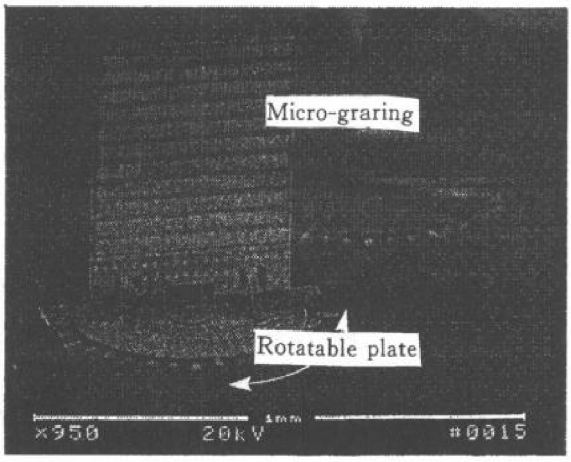


图 12　安装在转台上的微光栅 SEM 照片

垂直腔表面发射激光器(VCSEL)适合做成平面列阵，在光与互连、二维光信息处理、光计算等方面有重要应用。VCSEL 列阵与微透镜列阵及其他微光元件列阵的集成采用这种三维微机械结构具有紧凑、精确、便于封装、价格低等许多优点。图 13 表示 8 × 1VCSEL 列阵与微 Fresnel 透镜列阵的自对准混合集成结构。VCSEL 列阵的尺寸为宽 2mm，高 350μm，厚 125μm，每个激光器间隔为 250μm，与透镜列阵的间距相一致。在制造透镜列阵的过程中，VCSEL 的接触电极及对准块同时集成在 Si 基片上，使 VCSEL 安装时能自动对准并保证足够的精度。

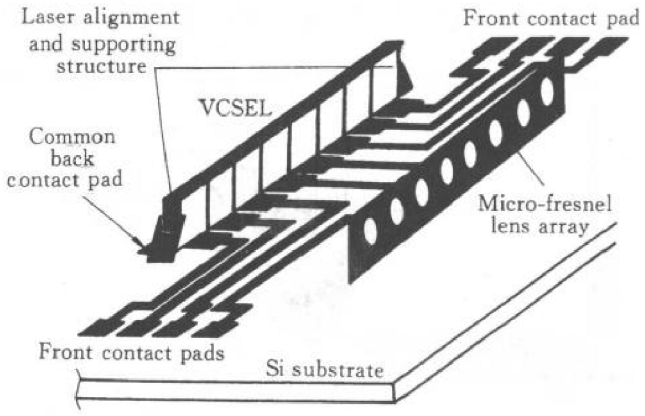


图 13　8 × 1 垂直腔表面发射激光器列阵与微 Frenel 透镜列阵的自对准混合集成系统

采用相似的设计方法，几乎可将现有常规光学平台上布置的各种光路都能用 MOB 结构来实现。例如可制成三维可调 F-P 标准具[12]，用于条码扫描器的表面微机械制造的静电梳驱动的扫描微镜[13]等。

6 结语

上面所介绍的一些 MOEMS 结构仅仅是近年来报道的研究成果的很小一部分。当代的工业、农业、交通、通信、医疗、环保、国防和科学研究等各方面所应用的生产设备，测试仪器不少已发展到光学、电子和机械三种系统的综合利用。微电子技术的先驱发展所取得的辉煌成果，充分显示了光子、电子和机械系统的综合运用和小型化、微型化，是 21 世纪技术发展的必然趋势。从这一观点出发，那么现在涌现的 MOEMS 系统只是如同刚出地平线的初升太阳。

参考文献

[1] Peterson K.E. Silicon as Amechanical Material. Proc. IEEE, 1982, 70: 420 ~ 457

[2] Pister K.S.J., Judy M.W., Burgett S.R., et al. Microfabricated Hinges. Sensors and Actuators, 1992, A33: 249 ~ 256

[3] Becker E.W., Ehrfeld W., Hagmann P., et al. Fabrication of Microstructures with High Aspect Ratios and Great Structural Heights by Synchrotron Radiation Litthography, Galvanoforming, and Plastic Moulding(LIGA Process). Microelectronic Engineering, 1986, 4: 35 ~ 36

[4] Menz W., Becher W., Harmening M., et al. The LIGA Technique-a Novel Concept for Microstructrues and the Combination with Si-Technologies by Iniection Molding. Proceedings of IEEE Micro Electro Mechanical Systems, 1991, 69 ~ 73

[5] Preston K.Jr. An Array Optical Spatial Phase Modulator. Proceedings of IEEE International Solid State Circuits Conference, New York, 1968, P 100

[6] Van Raalte J.A. A New Schlieren Light Valve for Television Projection. Appl. Opt., 1970, 2225 ~ 2230

[7] Walker J.A., Goossen K.W., Amey S.C., Frigo N.J., Iannnone P.P. A Silicon Optical Modulator with 1.5MHz Operation for Fiber-in-the-Loop Applications. Proceedings of 8th International Conference on Solid-State Sensors and Actuators, and Eurosensors IX, Sweden, 1995, 285 ~ 288

[8] Marxer C., Gretillat M.A., Jaeckin V.P., Baetting R., Anthamattten O., Vogel P., de Rooij N.F. Mhzopto-a-Chanical Modulattor. Proceedings of 8th International Conference on Solid-State Sensors and Actuators, and Eurosensors IX, Sweden, 1995, 289 ~ 292

[9] Younse J.M. Mirrors on a Chip. IEEE Spectrum, 1993, Nov, 27 ~ 31

[10] Lin L.Y., Lee S.S., Wu M.C., Pistter K.S.J. Micromachined Integrated Optics for Free-Space Interconections. Procedings IEEE Micro Electro Mechanical Systems Asterdam, 1995, 77 ~ 82

[11] Fan L.S., Tai Y.C., Muller R.S. IC-Processed Electrostattic Micromotors. Sensors and Actuattors, 1989, 20: 41 ~ 476

[12] Lin L.Y., Lee S.S., Wu M.M.C. Fabrication of Novel Three Dimensional Tunable Fanry-Perot Etalon by Surface-Micromachining. Intenational Conferece on Integrated Optics and Optical Fiber Communication, Hong, Kong, 1995, Tu D2 - 6

[13] Kiang M.H., Sollgaard O., Muller R.S., Lau K.Y. Surface-Micromachined Electrostic-Comb Driven Scanning Micromirrors for Barcode Scanners. Proceedings IEEE Micro Electro Mechanical Systems, 1996, 192 ~ 197

Thin Film Technologies for Micro-Opto-Electro-Mechanical System Applications

1 Introduction

Micro-optics and integrated optics has many interesting applications in optical fiber communication, optical signal processing, optical storage and optical sensors. However, fabrication of micro-optic devices and circuits in existence involved individual assembly with the aocociated reliability problems and overall large sizes. Marriage of integrated optics with microelectronics so called OEIC has resulted in many important features such as immunity to electromagnetic inference, ultra-high frequency response, parallel processing and increased simplicity. The potential combination of micro-optics and integrated optics with microelectronics and micromechanics to create a broader class of micro devices and system(MOEMS) further as shown in Fig.1[1] and lead to the demonstration of commercial devices such as the digital micromirror device(DMD) for projection displays, optical bench on chip, microrobotic, scanning probes, optical shutters choppers, tunable optical filters, optical switches、three dimensional integrated optics, compact optical processors and system packaging. The constituent technologies in MOEMS also allow for batch processing and productional low coat. This so called 3M combination well result in high performance devices and system that are lighter, more reliable more efficient, easier to assemble and package, and less expensive than conventional products.

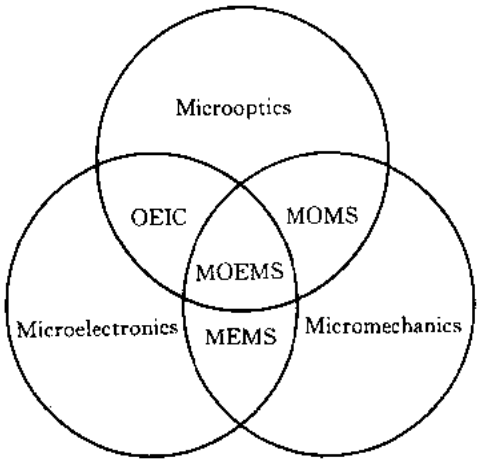


Fig.1 MOEM system-3M combination

In this paper, this elementary structure of system, various thin films and related technologies such as the bulk and surface micromachining and LIGA processing for the devices and systems of micro-opto-electro-mechanics will be reviewed and discussed.

2 Elementary Structure of Moems

MEOMS is a functional system which is constructed by the elements of microoptics, microelectronics and micromechanics combining together interactively. According to function of the component parts, usu-

ally a complete MOEMS consists of microsensors, microactuators and microprocessors which are physical formed by microoptics, microelectronics or micromechanics devices as shown Fig. 2[2]. Microsensors receive the information from outside(such as force, light, voice, gas…) and transform them into certain forms of signal(optical, electrical or mechanical…). Then the signals transmit into microprocessors. Finally, the processed results from the microprocessors input to the microactuators and are converted into actions or other information. However, the functions and the complexity of the systems are different, the architecture and the components of the MOEMS are also different, or it is only the sub-system of MOEMS, for example microsensor system or microactuator system etc. It is no room here to discuss all the elements and components in MOEMS mentioned above. However, in the following of this section, some interesting recent progresses are reviewed.

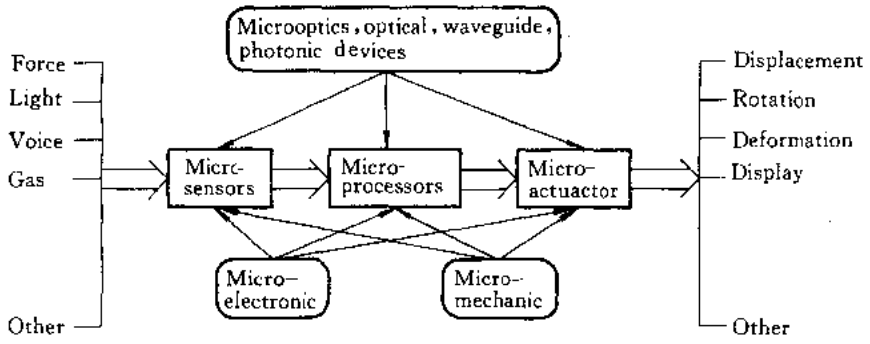


Fig. 2 Principle diagram for MEOM microsystems

A prerequisite for the development of many MOEMS is a mature microactuator technology. Toward this goal, research on design, fabrication, and characterization of new microactuators has increased significantly in the last few years. A kind of the main microactuactors is micromotor which can provide much larger motion comparing with the deformable microstructures. Unfortunately, for the most electrostatic micromotors, their drive moment is small(10nm), operating voltage is high(100V), and rotating energy is difficult to output. In Shanghai Jiao Tong University, a new design for electromagnetic drived micromtor was developed[3]. By using DC brushless motor structure, this new design can take the advantage of simple structure as AC motor, easy to produce without spark due to changing electrodes, high operating efficiency and tunable rotative velocity. In order to increasing the moment, a stator is constructed by three phase multilayer planar windings on substrate and a rotor consisting of permanent magnetic film poles with alternative polarization float above the stator to provide large interactive area as shown in Fig. 3. The film magnetoresistance positioning sensors which ensure the control accuracy of current direction converting are pre-fabricated on the stator substrate before making the windings. As the preliminary measurement, the drive moment is about $1.5\mu N \cdot m$, when operating current is 120mA, the tunable range of rotative velocity is 100 ~ 500r/min.

In MEOMS, the micro optics and waveguide optics elements are such as movable micromirrors, microlenses and array, binary optics and holograms elements, optical waveguide devices, integrated lighted light sources and photo detectors, etc. For the microelectronics elements, there are various drive circuits, control circuits, detector and amplification circuits, resistance, capacitance and inductance sensing-ele-

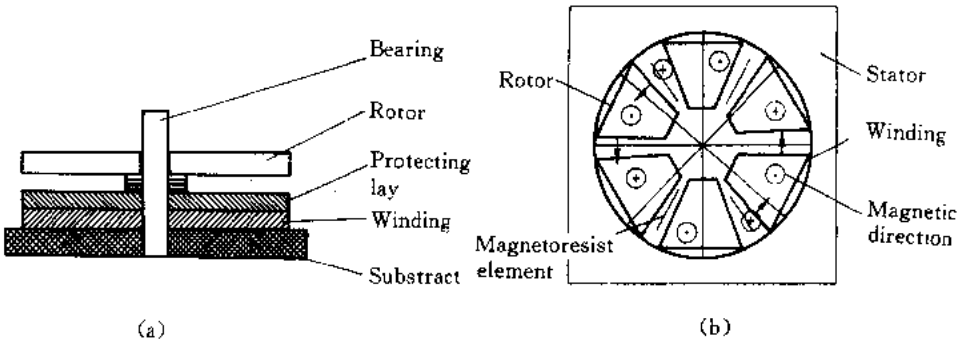


Fig.3 (a)cross-sectional schematic drawing, and (b)electromagnetic interaction between rotor and stator of electromagnetic micromotor

ments, and the sensors for gas, chemicals and biologics besides the conventional integrated circuits (ICs). The micromechanical elements generally include deformable membranes, movable plates, cantilevers, rods and sliders, gears, micro hinges, micro-joints, rotatable and displaceable stages.

One of the most important optical element in MOEMS is diffractive and binary optics. Their typical profiles shown in Fig.3 can be classified to refractive profiles, diffractive profiles, zeroth order profiles and combined profiles according to their main optical effect[4]. Diffractive profiles in this classifications are periodic or non periodic pattern like gratings and holograms. Binary, multilevel and continuous profiles are possible. The feature size of the periods a in the range of about $1 \sim 10\lambda$ order. For the zeroth order profiles, by decreasing the feature size of diffraction profiles down to the λ or sub-wavelength region birefringency, polarization and resonance properties become dominant. Combined profiles can fit the request for miniaturization and improvement of the optical functionality. Possible examples are refractive lens and diffractive lens, or diffractive grating and sub-wavelength polarizing grating, etc.

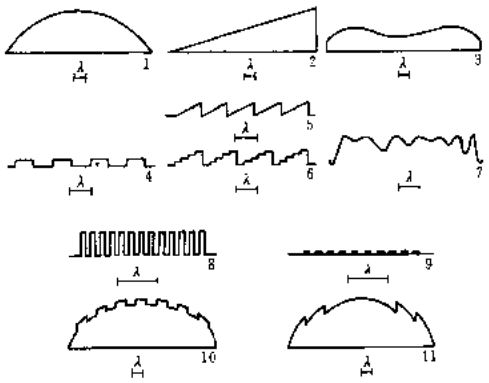


Fig.4 Typical diffractive and binary optical profiles:
refractive structures(1 ~ 3)
diffractive structures(4 ~ 7)
zeroth order structures(8,9)
combined structures(10,11)

A new micromechanical element for MOEMS is microjoints. The basic concept of microjoints is the use microfabrication to produce scaled down versions of well known macro joints such as: dovetails joints, dado joints, end joints and pin joints, etc. two examples of the fundamental micro-opto-mechanical components are *xyz* positioning microstages and $1 \times n$ fiber optic switch[5]. The *xyz* microstage is predicated on the use of self constrained dovetail joints as illustrated in Fig.5. The construction of the dovetail slide is based on the anisotropic etching geometries of single crystal(100) silicon, in principle, the technique could be equally applied to virtually any anisotropic etched crystalline material. GaAs, InP, Quartz, etc. A $1 \times n$ fibers optic switch may also be realized by using the dovetail microjoint fabrication process as

Fig.6. An input optical fiber is mounted on a micromachined linear translational stage with n corresponding output fibers mounted opposite the input fiber to receive the optical energy. A thin film of ferromagnetic material was deposited on the input fiber stage and a magnetic field was applied to either side to actuate the input fiber. For the preliminary tests, the measured insertion loss between fibers was approximately-1dB and the cross talk was less than -60dB. The calculations indicate insertion losses can be reduced to about 0.4dB, the versatility of this technique allows for the assembly of complete optical microsystems composed of a diversity of optical components on a single micromachined "optical bench".

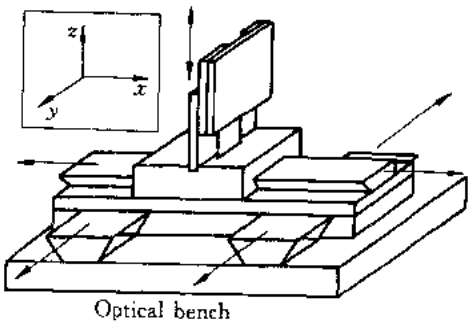


Fig.5 Diagram of an *xyz* positional stage with dovetail slides

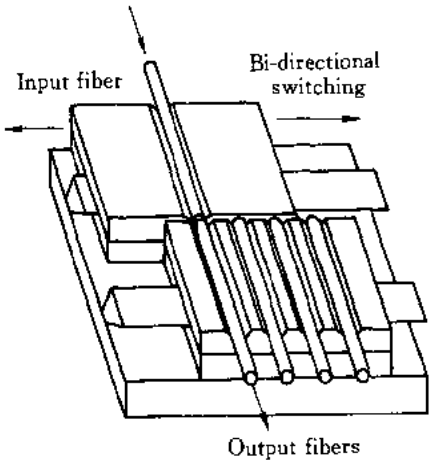


Fig.6 Diagram of 1 x 4 fiber optic switch. The input fiber is magnetical actuated between the output fibers with mechanical detenles to align the input to output fibers

3 Microfabrication and Micromachining

Silicon microfabrication technology, and more specifically silicon micromaching, and microfabrication or micromaching technologies expanded to various materials such as glass, quartz, ceramics and GaAs, etc., has been a key factor for raped progress of microsensors, microactuators, MEMS and now MOEMS. Silicon micromachining, which refers to forming microscopic mechanical parts on a silicon substrate or out of a silicon subtract, has emerged as an extension of IC fabrication technology. Micromachining is used to fabrication a variety of mechanical microstructures of great diversity including beams, diaphragms, grooves, orifices, cavities, pyramids, needles, springs, suspensions, hinges, gears, linkages and micromotors. Bulk and surface micromachining, as well as bonding of substract and electroforming in conjunction with high-aspect-ratio lithography, a integral components of silicon micromachining and other micromachining.

Buck micromachining uses wet and dry etching techniques in conjunction with etch masks and etch stops to form micromechanical elements and devices from the substrate. There are two key capabilities that make bulk micromachining a feasible and diversified technology. First, anisotropic etching, taking silicon as a sample, anisotropic etchants of silicon such as ethylene-diamine and pyrocatecol (EDP),

potassium hydroxide(KOH), and hydrazine are available which preferentially etch single crystal silicon along given crystal planes. Second, etch masks and etch-stop techniques are available which can be used in conjunction with silicon anitropic etchants to selectively prevent regions of silicon from being etched. As a result, it is possible to fabricate microstructures in a silicon subtract by appropriated combining etch masks and etch-stop patterns with anisotropic etchants. The basic fabrication process for bulk micromachining of a dovetail microjoint is outlined in Fig.7.

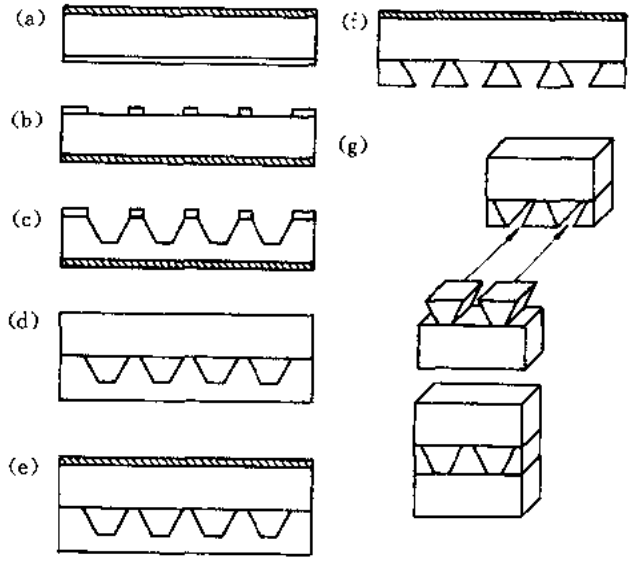


Fig.7 Process flow for the fabrication of dovetail microjoint(typical bulk micromachining processes)
(a) an LPCVD silicon nitride etch mask is deposited on (100) silicon wafers; (b) nitride patterned on one side; (c) wafer anisotropically etched to the desired depth; (d) nitride stripped, and wafer bonded to a second wafer; (e) a second layer of silicon nitride is deposited on the bonded wafers, selectively removed on one side; (f) exposed silicon thinned in KOH to define the dovetails; (g) wafer diced and assembled

Micromachining, bonding techniques are used to assemble individually macromachined parts to form a complete system. Usually entire wafers or individual dies are bonded together. When used in conjunction of three dimensional structures which are thicker than a single wafer. This is important for micropumps and microvalves where more than one cavity is needed in series in the thickness direction. Several processes have been developed for bonding a pair of silicon wafer together directly face to face. Anodic bonding is used silicon to glass. As a sample, in the process for fabrication of a dovetail microjoint Fig. 7(d), an anisotropically etched water is bonded to second wafer by using silicon fusion bonding technique.

Surface micromachining demonstrated initially by Nathanson, et al[6] in building a free standing metal-gate field-effect transistor, relies on encasing the structural parts of the device in layers of a sacrificial material during the fabrication process. In other words, there are two primary components in a surface micromachining process. One is structural layers, of which the final microstructures are made, and another is sacrificial layer, which separate the structural layers and are dissolved in the final stage of device fabri-

cation. Fig.8 illustrates a typical surface micromachining process for a cantilever structure[7]. The merits of suface micromachining are first itself in contrast to buck micromachining and bonding, the bulk of the silicon wafer itself is not etched. Therefore, wafers undergoing a surface micromachining process may utilize an IC fabrication facility in "normal" condition, without disturbing existing IC fabrication processes. Second, more important, the layered fabrication nature of surface micromachining provides for significant flexibility in design of micromechanical devices. For example, the fabrication of a rotor on a center bearing, or in general mechanisms, is not possible in buck micromachining and would be much more complicated by bonding. Extension of the basic processes described above to incorporate additional structure and sacrificial layers will provide even more flexibility in the design of micromechanical system. However, surface micromachining has an inherent limitation. It is, indeed, a planar fabrication process, and is limiting for mechanical design. For many microactuactor and micropackaging applications, thick, high-aspect-ratio(HAR)devices offer the possibility of production high torque or force because of larger interaction areas. HAR processes generally consist of two steps, plating molds fabricating and plating or electroforming. Currently, three major methods being investigated for making HAR plating molds.

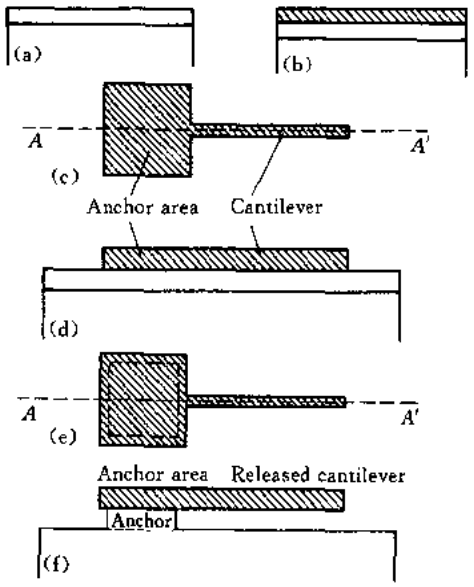


Fig.8 Schematics demonstrating surface micromachining for a cantilever
(a)sacrificial layer deposited, (b)structure layer deposited, (c)top plan view after pattering of the structural layer, (d)cross section AA' of(c), (e)top play view after release, and (f)cross section AA' of(e)

LIGA[8] is probably the best-known technique of fabricating HAR structures. LIGA is an acronym derived from the German word Lithografie, Galvanik, Abformung, which means lithography, electroforming, and injection molding. In this process, high-intensity, low-divergence, hard X-rays are used as the exposure source for the lithography. PMMA(polymethylmethacrylate) is used as the X-ray resist. LIGA processes shown in Fig.9. Thicknesses of several hundreds of microns and aspect-ratio of more than 100 have been achieved. However, a synchrotron is necessary for LIGA, and since only a few exist in the world, use of this process is limited.

Photolithography using near-UV light sources and commercially available positive photoresists[9] or photosensitive polyimides[10] are also can fabricate HAR plating molds. Although in comparison to LIGA, this technique, so-called LIGA-like process, is limited in terms of thickness and aspect ratio, but it can provide a simple means of fabrication HAR plating molds with conventional photolithography equipment and reasonably cost effective. Generally HAR can be obtained, for example, a HAR comb finger pattern (20μm thick resist, and 2μm gaps and 4μm fingers) was fabricated using conventional photo lithogra-

phy[9], and the maximum aspect ratio of 8:1(for about 10μm wide lines) can be achieved in the photosensitive polyimides process.

The third method to make a HAR mold by dry etching of polyimides has been reported. In these processes, some modifications of traditional RIE systems are necessary to achieve HAR. For example, magnetically-controlled dry etching of fluorinated polyimides[11] with a Ti mask was used for deep etching with excellent mask selectivity and smooth sidewalls. Circular cylinder 15μm in diameter and 100μm in height have been obtained using these methods.

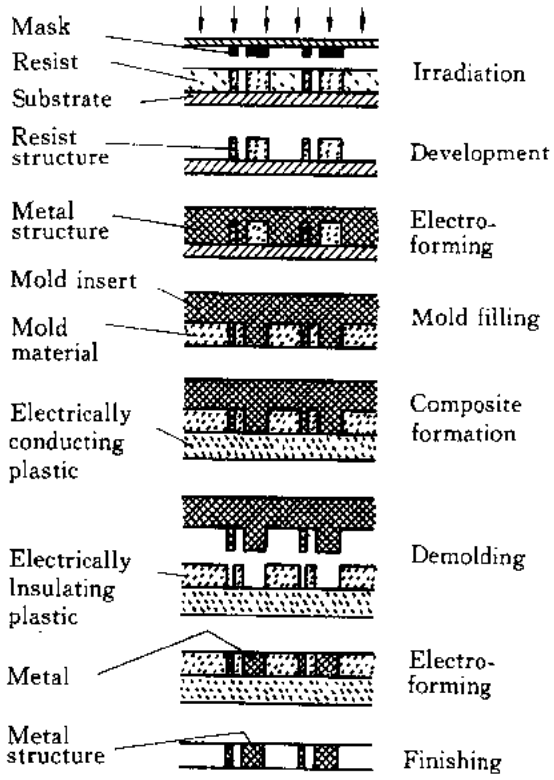


Fig.9 Process steps of LIGA technique

A critical part of the HAR processes is plating to form the metallic micromechanical parts in the mold. In plating, metal is deposited from ions in a solution following the shape of the plating mold. This process has many attractive features:(1)it is additive process; (2)the thickness of the plated metal can be large since the plating rate is high;(3) a variety of metals can deposited or co-deposited, such as Ni, Cu, Au or alloys like Ni-Co, Ni-Fe, and(4) this process can achieve smooth reflective metal surfaces for optical application. Some new potential applications, such as using Ni-Fe for magnetic actuators, conductive materials micro coils(such as for electromagnetic micromotor mentioned above), and Ni-Si for solar cells, as well as electrostatic actuators or mechanical and parts, can be produced by plating.

Reference

[1] Y.X.Chen. Integrated Optics and Optoelectronics with Micro-Opto-Electro-Mechanical-System(MOEMS). Invited Paper, First National Conference on Photonics, Shenzhen, China, Oct.1996, 10 ~ 14

[2] Y.X.Chen. Technologies and Applications for Micro-Opto-Electro-Mechanical System(MEOMS). National Symposium on Micro-Systems, Chongqing, China, April.1997, 10 ~ 14

[3] Zhang X.L., Zhao S.B.Zhang, B.C.Cai. The Research for Electromagnetic Micromotor with Diameter of 2mm and Its Control Circuit. Micro and Nanometer Science & Technology, 1995,(1):40

[4] E.B.Kley, B.Schnabel, U.D.Zeitner. E-bean Lithography-an Efficient Tool for the Fabrication of Diffractive and Micro Optical Elements. SPIE Proceedings, San Jose, USA, February, 1997, 3008:222 ~ 232

[5] C. Gonzalez, R.J.Welty, R.L.Smith, S.D.Collins. Microjoinery for Optomechanical Systems. SPIE Proceedings, San Jose, USA, February, 1997, 3008:171 ~ 178

[6] H.C.Nathanson, R.A.Wickstrom. A Resonant Gate Silicon Surface Transistor with High Q Band Pass Properties. Appl. Phys. Lett., 1965,(7):84

[7] T.Hirano, T.Furuhata, K.J.Gabriel, H.Fujita. Operation of Sub-Micron Gap Electrostatic Comb-Drive Actuators. Technical Digest, 6th Int. Conf. Solid State-Sensors and Actuators, San Francisco, 1991.873~876

[8] E.W.Becker, W.Ehrfeld, P.Hapmann, A.Maner, D.Munchmeyer. Fabrication of Microstructures with High Aspect Ratios and Great Structure Height by Synchrotron Radiation Lithography, Galvanoforming and Plastic Moulding (LIGA) Process. Microelectronics Engineering, 1986,4:35~56

[9] S.Furukawa, H.Miyajima, M.Mehregany, C.C.Liu. Electroless Plating of Metals for Micromechanical Structures. Technical Digest of the 7th Int. Conf. Solid-State Sensors and Actuators: Transducer'93, Yokohama, Japan, 1993, 66~69

[10] M.Allen. Polyimide-Based Processes for the Fabrication of Thick Electroplated Microstructures. Technical Digest of the 7th Int. Conf. Solid-State Sensors and Actuactors: Transducer'93, Yokohama, Japan, 1993,60~65

[11] F.Shimokura, A. Furuya, S.Matsui. Fast and Extremely Selective Ployimide Etching with a Magnetically Controlled Reactive Ion Etching System. Proceedings of IEEE Micro Electro Mechanical System Workshop, Nara, Japan, 1991, 192~197