

并行计算

——结构•算法•编程

主讲教师：谢磊

并行计算——结构·算法·编程

- * 第一篇 并行计算的基础
 - * 第一章 并行计算机系统及其结构模型
 - * 第二章 当代并行机系统：SMP、MPP和Cluster
 - * 第三章 并行计算性能评测
- * 第二篇 并行算法的设计
 - * 第四章 并行算法的设计基础
 - * 第五章 并行算法的一般设计方法
 - * 第六章 并行算法的基本设计技术
 - * 第七章 并行算法的一般设计过程

并行计算——结构·算法·编程

- * 第三篇 并行数值算法
 - * 第八章 基本通信操作
 - * 第九章 稠密矩阵运算
 - * 第十章 线性方程组的求解
 - * 第十一章 快速傅里叶变换
- * 第四篇 并行程序设计
 - * 第十二章 并行程序设计基础
 - * 第十三章 并行程序设计模型和共享存储系统编程
 - * 第十四章 分布存储系统并行编程
 - * 第十五章 并行程序设计环境与工具

第一章 并行计算机系统及结构模型

- * 1.1 并行计算
 - * 1.1.1 并行计算与计算科学
 - * 1.1.2 当代科学与工程问题的计算需求
- * 1.2 并行计算机系统互连
 - * 1.2.1 系统互连
 - * 1.2.2 静态互联网络
 - * 1.2.3 动态互连网络
 - * 1.2.4 标准互联网络
- * 1.3 并行计算机系统结构
 - * 1.3.1 并行计算机结构模型
 - * 1.3.2 并行计算机访存模型

并行计算

- * 并行计算
 - * 并行机上所作的计算，又称高性能计算或超级计算。
 - * **Parallel Computing** : Multiple processes cooperating to solve a single problem.
 - * **A Parallel Computer** is a “collection of processing elements that communicate and cooperate to solve large problem fast” [David E. Culler]
- * 三大科学
 - * 计算科学、理论科学与实验科学
- * 计算科学
 - * 计算物理、计算化学、计算生物等

并行计算

- * 科学与工程问题的需求
 - * 气象预报、油藏模拟、核武器数值模拟、航天器设计、基因测序等。
- * 需求类型
 - * 计算密集、数据密集、网络密集。
- * 美国HPCC计划：重大挑战性课题，3T性能
- * 美国Petaflops研究项目：Pflop/s。
- * 美国ASCI计划：核武器数值模拟。

高性能计算机

Period	Supercomputer	Peak speed	Location
1946–1956	U. of Pennsylvania ENIAC	50 kFLOPS	Aberdeen Proving Ground, Maryland, USA
1956–1958	MIT TX-0	83 kFLOPS	Massachusetts Inst. of Technology, Lexington, Massachusetts, USA
1958–1960	IBM SAGE	400 kFLOPS	U.S. Air Force, USA
1960–1961	UNIVAC LARC	500 kFLOPS	Lawrence Livermore National Laboratory, California, USA
1961–1964	IBM 7030 "Stretch"	1.2 MFLOPS	Los Alamos National Laboratory, New Mexico, USA
1964–1969	CDC 6600	3 MFLOPS	Lawrence Livermore National Laboratory, California, USA
1969–1974	CDC 7600	36 MFLOPS	Lawrence Livermore National Laboratory, California, USA
1974–1975	CDC Star-100	100 MFLOPS	Lawrence Livermore National Laboratory, California, USA
1975–1976	Burroughs ILLIAC IV	150 MFLOPS	NASA Ames Research Center, California, USA
1976–1981	Cray-1	250 MFLOPS	Los Alamos National Laboratory, New Mexico, USA (80+ sold worldwide)

高性能计算机

1981–1983	CDC Cyber 205	400 MFLOPS	(numerous sites worldwide)
1983–1984	Cray X-MP/4	941 MFLOPS	Los Alamos & Lawrence Livermore Nat. Laboratories , Battelle , Boeing
1984–1985	M-13	2.4 GFLOPS	Scientific Research Institute of Computer Complexes , Moscow , USSR
1985–1989	Cray-2/8	3.9 GFLOPS	Lawrence Livermore National Laboratory , California , USA
1989–1993	ETA10-G/8	10.3 GFLOPS	Florida State University , Florida , USA
1993–1994	Thinking Machines CM-5	37.5 GFLOPS	Los Alamos National Laboratory , California , USA
1994–1995	Fujitsu Numerical Wind Tunnel II	236 GFLOPS	National Aerospace Lab , Japan
1995–2000	Intel ASCI Red	2.15 TFLOPS	Sandia National Laboratories , New Mexico , USA
2000–2002	IBM ASCI White	9.216 TFLOPS	Lawrence Livermore National Laboratory , California , USA
2002.6–2004	NEC Earth Simulator	35.86 TFLOPS	Yokohama Institute for Earth Sciences , Japan

高性能计算机

<u>2004.11–2005.6</u>	IBM Blue Gene/L prototype	74 TFLOPS	IBM, Rochester, Minnesota, USA
<u>2005.6–2005.11</u>	IBM Blue Gene/L prototype	135.5 TFLOPS	IBM, Rochester, Minnesota, USA
<u>2005.11–2007.6</u>	IBM Blue Gene/L prototype	280.6 TFLOPS	IBM, Rochester, Minnesota, USA
2007.11	IBM Blue Gene/L prototype	478.2 TFLOPS	IBM, Rochester, Minnesota, USA
2008.6	IBM Roadrunner	1.026 PFLOPS	IBM, Los Alamos, USA
<u>2008.11–2009.6</u>	IBM Roadrunner	1.105 PFLOPS	IBM, Los Alamos, USA
<u>2009.11–2010.6</u>	Cray Jaguar	1.759 PFLOPS	Cray, Oak Ridge National Laboratory, USA
2013 2017 2020	天河2号 神威·太湖之光 Fugaku富岳	54.9PFlops 93.015PFlops 415.530 PFlops	国防科技大学, 中国 国家并行计算机工程技术研究中心, 中国 日本理化学研究所(RIKEN)与富士通公司

全球十大最快超级计算机 中国位居第一（2013年）

- * 2013年，美联储科技杂志对全球超级计算机进行了排名，选出了其中最快的十台，其中中国有两台超级计算机入榜，并且“天河二号”凭借着双精度浮点运算峰值速度达到每秒5.49亿亿次问鼎该宝座。
- * 这也是中国超级计算机时隔两年半后运算速度重返世界之巅。此前的2010年11月，“天河一号”曾以每秒4.7千万亿次的峰值速度，首次登上超级计算领域顶峰。

全球十大最快超级计算机 中国位居第一（2013年）

- * “天河二号”由国防科大研制的天河二号超级计算机系统，以峰值计算速度每秒5.49亿亿次、持续计算速度每秒3.39亿亿次双精度浮点运算的优异性能位居榜首，成为全球最快超级计算机。
- * 此外本次全球超级计算机排行前十名分别是天河二号、泰坦、红杉超级计算机、K Computer、米拉、Stampede、Juqueen、vulcan、SuperMUC、天河一号。

神威·太湖之光超级计算机（2016年）

- * 2016年6月20日，TOP500组织在法兰克福世界超算大会（ISC）上，“神威·太湖之光”超级计算机系统登顶榜单之首，成为世界上首台运算速度超过十亿亿次的超级计算机。而“中国芯”“申威26010”的问世，也成为中国自主研发打破30年技术封锁的一柄利器。
- * 峰值性能125.436PFlops，世界第一；持续性能93.015PFlops，世界第一；性能功耗比6051MFlops/W，还是世界第一。

神威·太湖之光超级计算机（2016年）

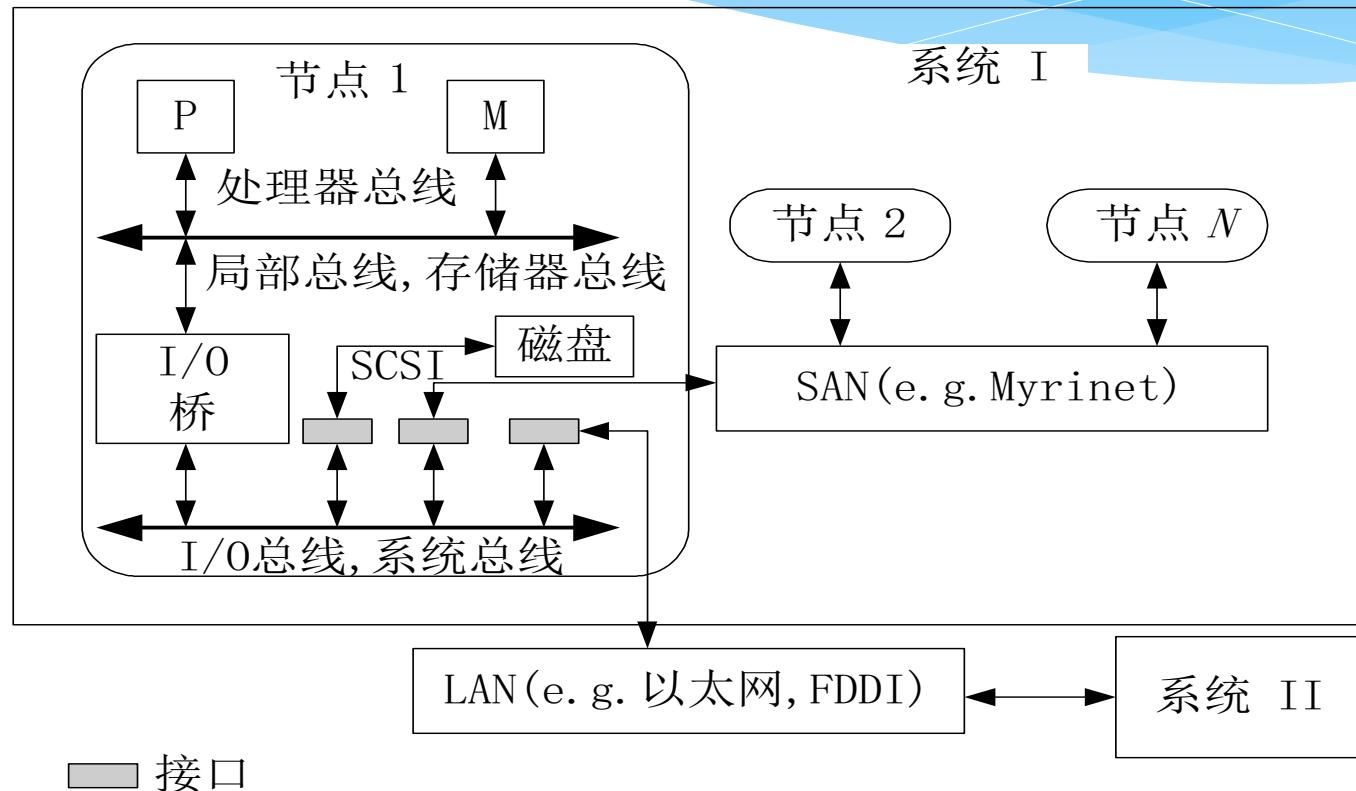
* 神威·太湖之光超级计算机由40个运算机柜和8个网络机柜组成。每个运算机柜比家用的双门冰箱略大，打开柜门，4块由32块运算插件组成的超节点分布其中。每个插件由4个运算节点板组成，一个运算节点板又含2块“申威26010”高性能处理器。一台机柜就有1024块处理器，整台“神威·太湖之光”共有40960块处理器。



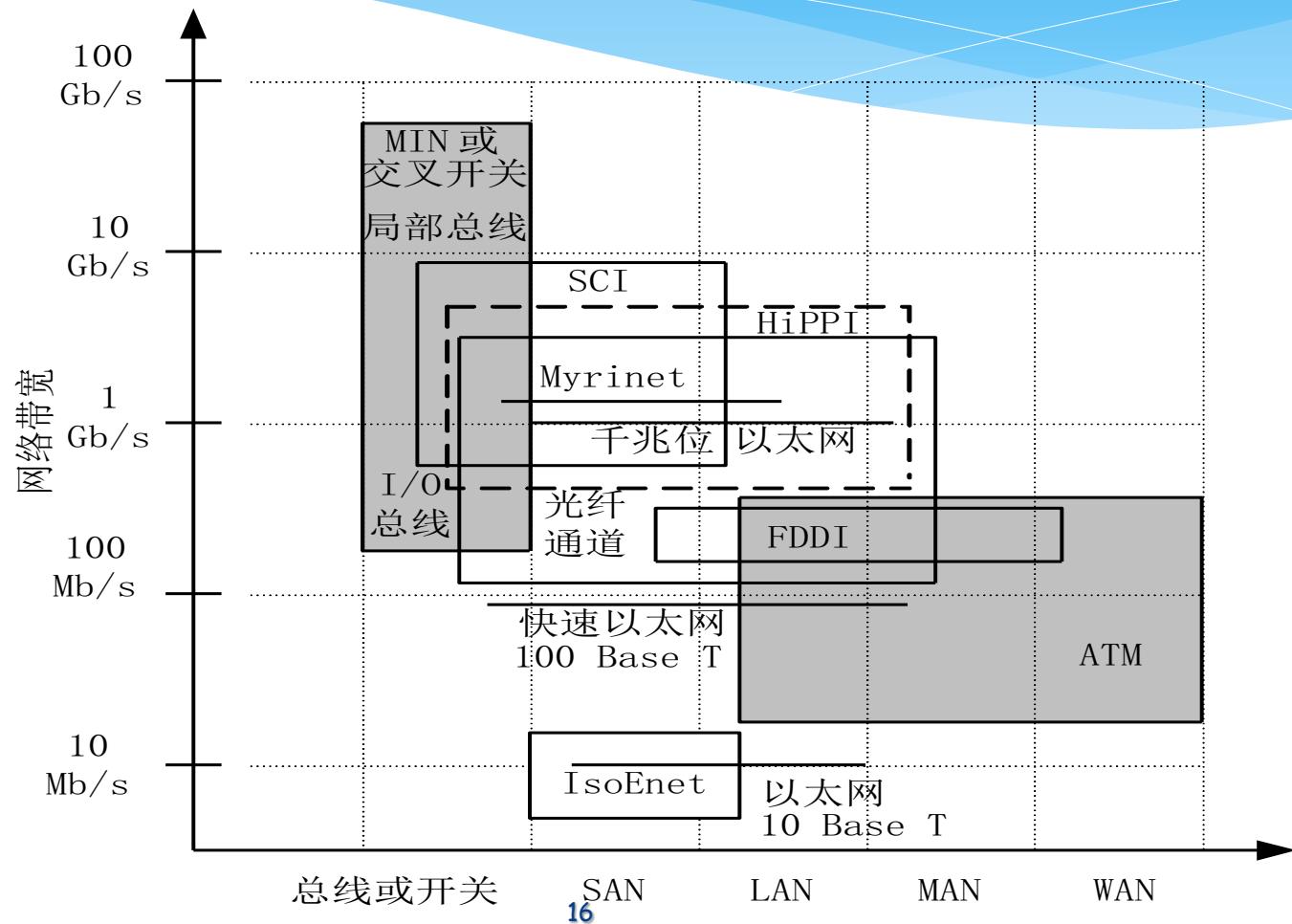
系统互连

- * 不同带宽与距离的互连技术
- * 总线
 - * 处理器总线、局部总线（存储器总线）、I/O总线（系统总线）
- * SAN：系统域网络
- * LAN：局域网络
- * MAN：都域网
- * WAN：广域网

局部总线、I/O总线、SAN和LAN



系统互连



网络性能指标

- * 节点度 (Node Degree) : 射入或射出一个节点的边数。在单向网络中，入射和出射边之和称为节点度。
- * 网络直径 (Network Diameter) : 网络中任何两个节点之间的最长距离，即最大路径数。
- * 对剖宽度 (Bisection Width) : 对分网络各半所必须移去的最少边数
- * 对剖带宽 (Bisection Bandwidth) : 每秒钟内，在最小的对剖平面上通过所有连线的最大信息位（或字节）数
- * 如果从任一节点观看网络都一样，则称网络为对称的 (Symmetry)

静态互连网络与动态互连网络

- * 静态互连网络

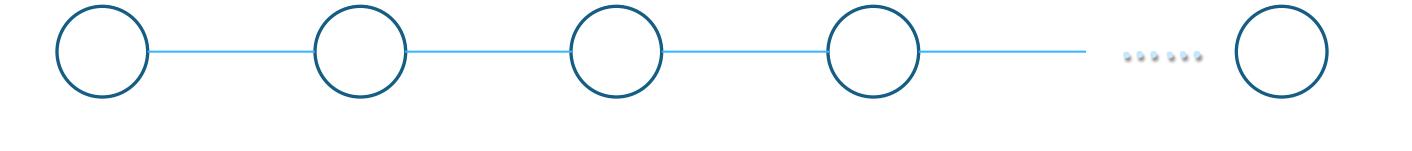
- * 处理单元间有着固定连接的一类网络，在程序执行期间，这种点到点的链接保持不变
- * 典型的静态网络有一维线性阵列、二维网孔、树连接、超立方网络、立方环、洗牌交换网、蝶形网络等

- * 动态网络

- * 用交换开关构成的，可按应用程序的要求动态地改变连接组态
- * 典型的动态网络包括总线、交叉开关和多级互连网络等。

静态互连网络 (1)

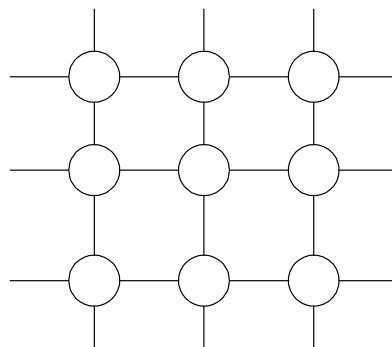
- * 一维线性阵列 (1-D Linear Array) :
 - * 并行机中最简单、最基本的互连方式，
 - * 每个节点只与其左、右近邻相连，也叫二近邻连接，
 - * N 个节点用 $N-1$ 条边串接之，内节点度为2，直径为 $N-1$ ，对剖宽度为1
 - * 当首、尾节点相连时可构成循环移位器，在拓扑结构上等同于环，环可以是单向的或双向的，其节点度恒为2，直径或为 $\lfloor N/2 \rfloor$ （双向环）或为 $N-1$ （单向环），对剖宽度为2



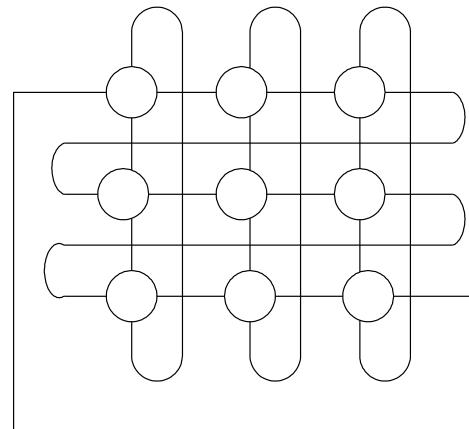
静态互连网络 (2)

* $\sqrt{N} \times \sqrt{N}$ 二维网孔 (2-D Mesh) :

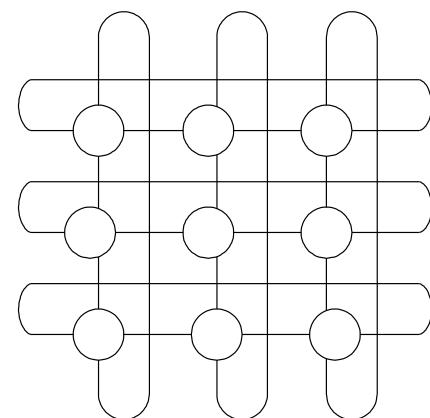
- * 每个节点只与其上、下、左、右的近邻相连（边界节点除外），节点度为4，网络直径为 $2(\sqrt{N}-1)$ ，对剖宽度为 \sqrt{N}
- * 在垂直方向上带环绕，水平方向呈蛇状，就变成 Illiac 网孔了，节点度恒为4，网络直径为 $\sqrt{N}-1$ ，而对剖宽度为 $2\sqrt{N}$
- * 垂直和水平方向均带环绕，则变成了2-D环绕 (2-D Torus)，节点度恒为4，网络直径为 $2\lfloor\sqrt{N}/2\rfloor$ ，对剖宽度为 $2\sqrt{N}$



(a) 2-D网孔



(b) Illiac网孔
20

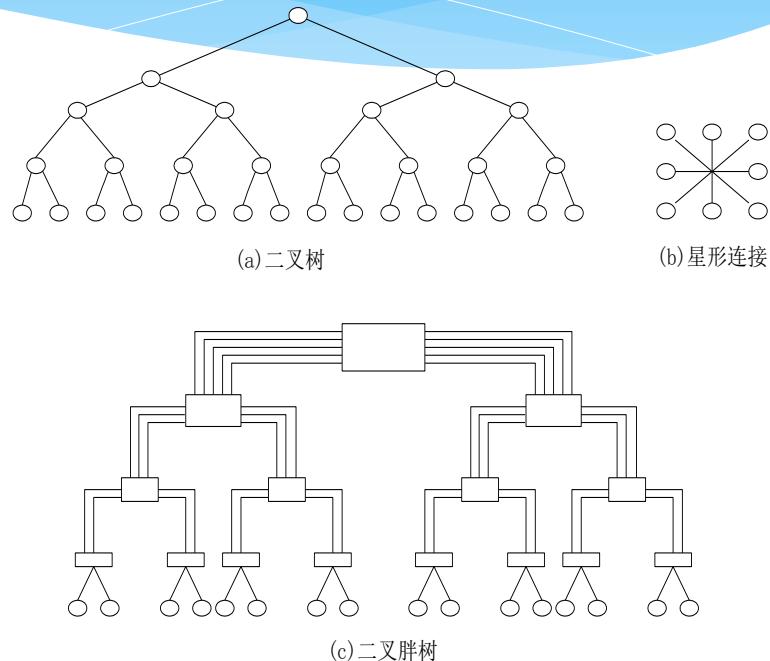


(c) 2-D环绕

静态互连网络 (3)

* 二叉树:

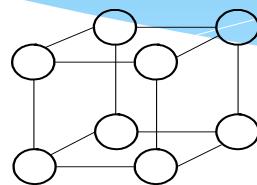
- * 除了根、叶节点，每个内节点只与其父节点和两个子节点相连。
- * 节点度为3，对剖宽度为1，而树的直径为 $2(\lceil \log N \rceil - 1)$
- * 如果尽量增大节点度为 $N-1$ ，则直径缩小为2，此时就变成了星形网络，其对剖宽度为 $\lfloor N/2 \rfloor$
- * 传统二叉树的主要问题是根易成为通信瓶颈。胖树节点间的通路自叶向根逐渐变宽。



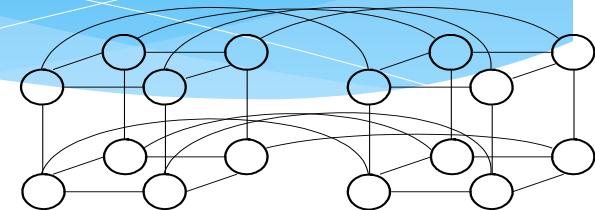
静态互连网络 (4)

* 超立方：

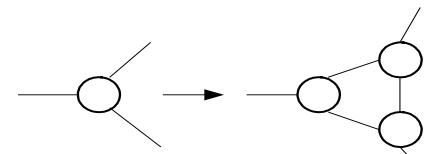
- * 一个 n -立方由 $N = 2^n$ 个顶点组成，3-立方如图(a)所示；4-立方如图(b)所示，由两个3-立方的对应顶点连接而成。
- * n -立方的节点度为 n ，网络直径也是 n ，而对剖宽度为 $N/2$ 。
- * 如果将3-立方的每个顶点代之以一个环就构成了如图(d)所示的3-立方环，此时每个顶点的度为3，而不像超立方那样节点度为 n 。



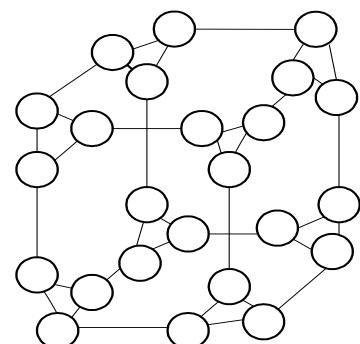
(a) 3-立方



(b) 4-立方



(c) 顶点代之以环

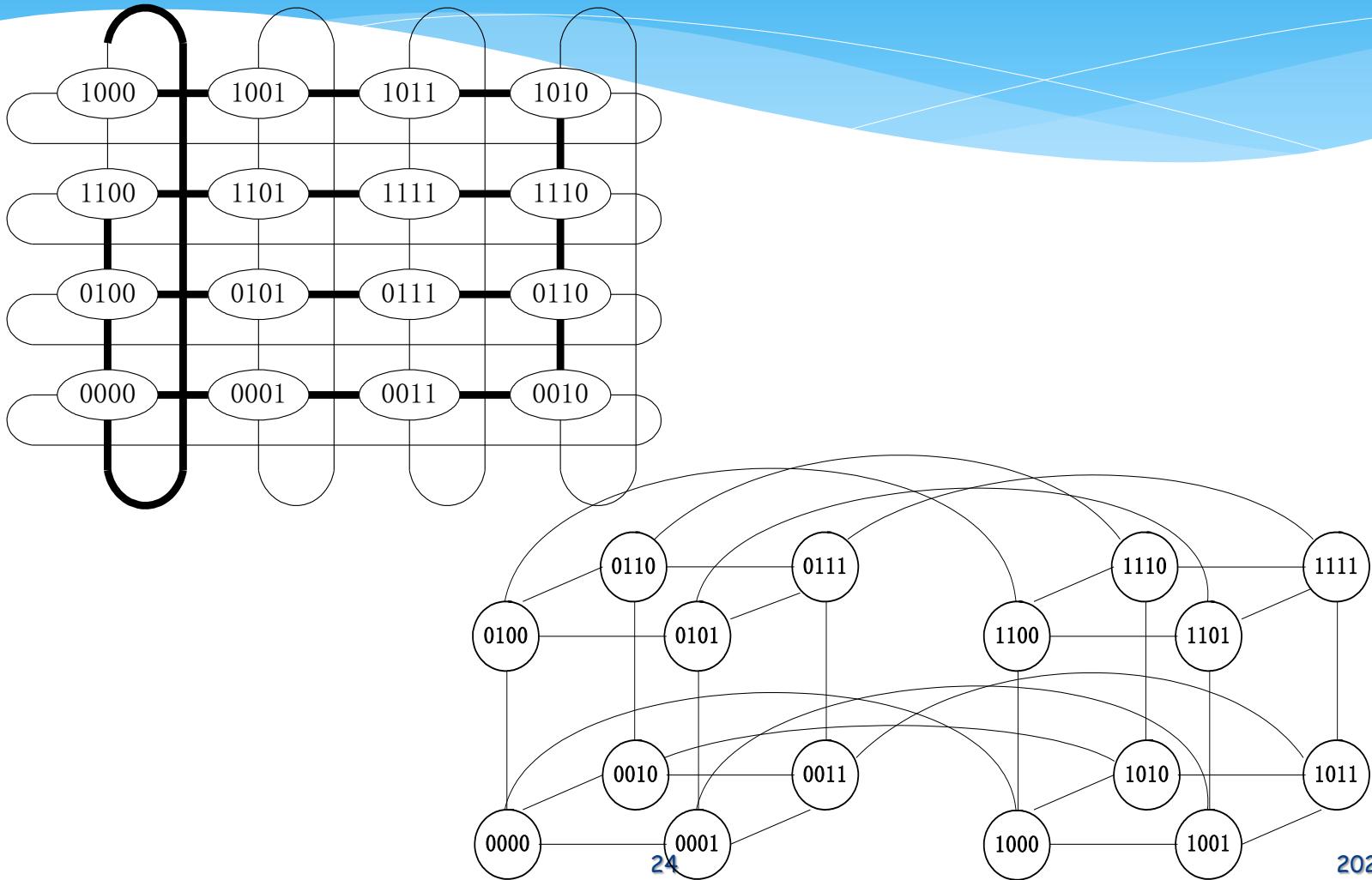


(d) 3-立方环

嵌入

- * 将网络中的各节点映射到另一个网络中去
- * 用膨胀 (Dilation) 系数来描述嵌入的质量，它是指被嵌入网络中的一条链路在所要嵌入的网络中对应所需的最大链路数
- * 如果该系数为1，则称为完美嵌入。
- * 环网可完美嵌入到2-D环绕网中
- * 超立方网可完美嵌入到2-D环绕网中

嵌入



静态互连网络特性比较

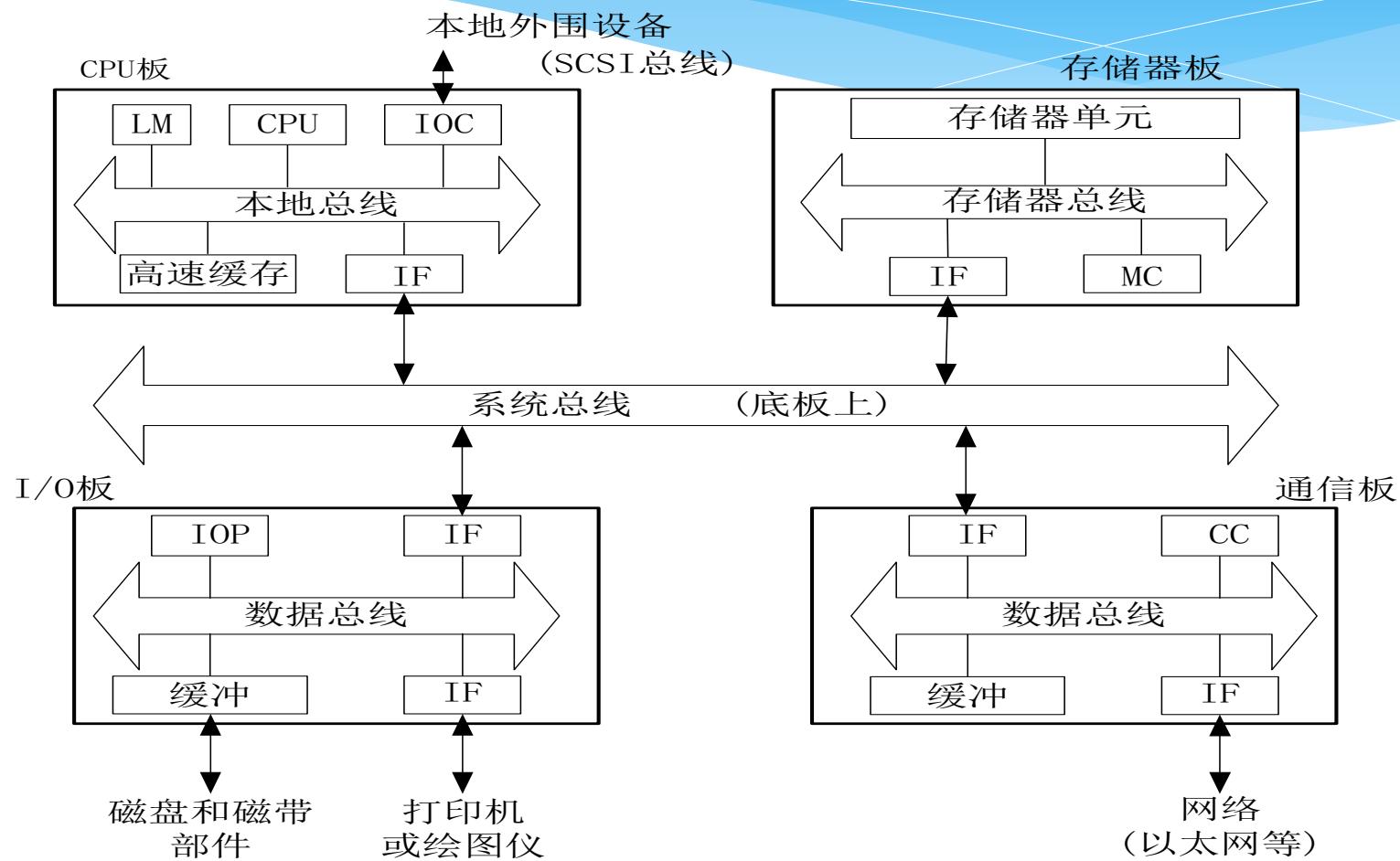
网络名称	网络规模	节点度	网络直径	对剖宽度	对称	链路数
线性阵列	N	2	$N - 1$	1	非	$N - 1$
环形	N	2	$\lfloor N/2 \rfloor$ (双向)	2	是	N
2-D网孔	$(\sqrt{N} \times \sqrt{N})$	4	$2(\sqrt{N} - 1)$	\sqrt{N}	非	$2(N - \sqrt{N})$
Illiac网孔	$(\sqrt{N} \times \sqrt{N})$	4	$\sqrt{N} - 1$	$2\sqrt{N}$	非	$2N$
2-D环绕	$(\sqrt{N} \times \sqrt{N})$	4	$2\lfloor \sqrt{N}/2 \rfloor$	$2\sqrt{N}$	是	$2N$
二叉树	N	3	$2(\lceil \log N \rceil - 1)$	1	非	$N - 1$
星形	N	$N - 1$	2	$\lfloor N/2 \rfloor$	非	$N - 1$
超立方	$N = 2^n$	n	n	$N/2$	是	$nN/2$
立方环	$N = k \cdot 2^k$	3	$2k - 1 + \lfloor k/2 \rfloor$ 25	$N/(2k)$	是	$3N/2$

动态互连网络(1)

- * 总线：

- * 总线实际上是一组导线和插座，连接处理器、存储模块和I/O外围设备等。
- * 总线系统用以主设备（如处理器）和从设备（如存储器）之间的数据传输。
- * 目前已有很多总线标准：PCI、VME、Multics、Sbus、MicroChannel。
- * 多处理机总线系统的主要问题包括总线仲裁、中断处理、协议转换、快速同步、高速缓存一致性协议、分事务、总线桥和层次总线扩展等。

动态互连网络(2)

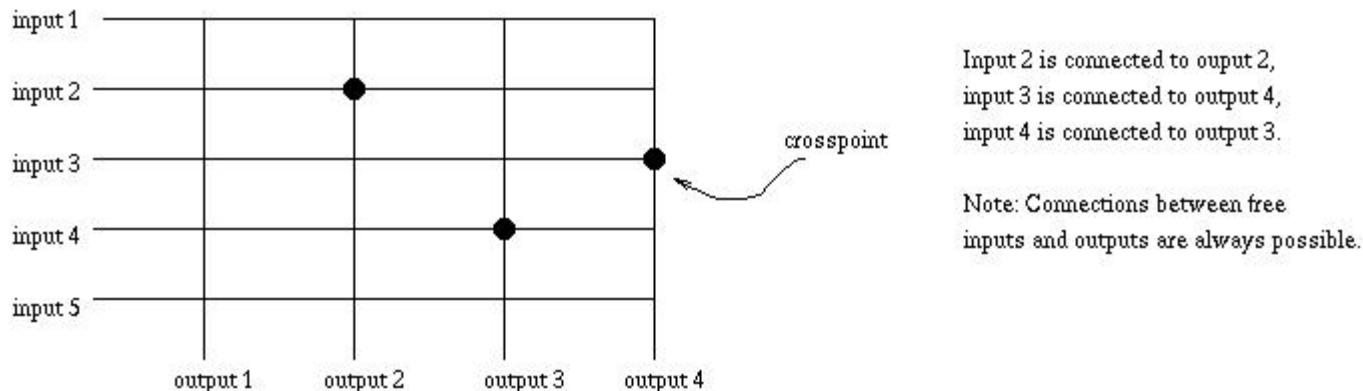


开关成本最高
需要 n^2 个开关
早期超算使用

动态互连网络 (3)

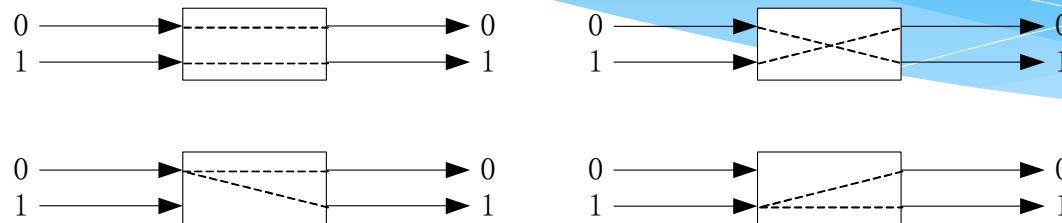
■ 交叉开关 (Crossbar) :

- 单级交换网络，可为每个端口提供更高的带宽。象电话交换机一样，交叉点开关可由程序控制动态设置其处于“开”或“关”状态，而能提供所有（源、目的）对之间的动态连接。
- 交叉开关一般有两种使用方式：一种是用于对称的多处理机或多计算机机群中的处理器间的通信；另一种是用于SMP服务器或向量超级计算机中处理器和存储器之间的存取。

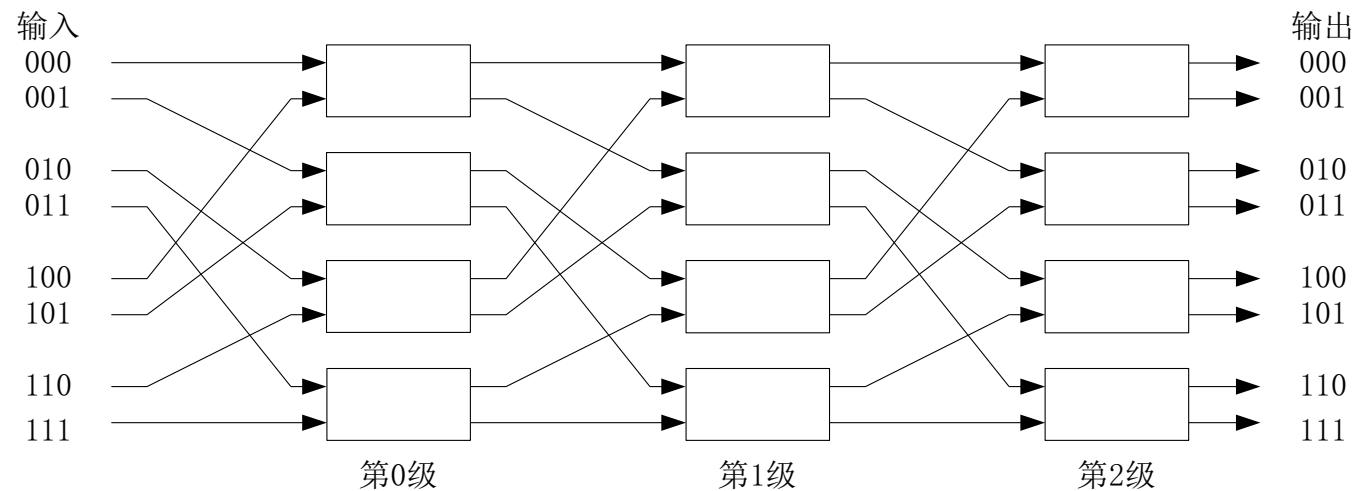


动态互连网络 (4)

* 单级交叉开关级联起来形成多级互连网络MIN
(Multistage Interconnection Network)



(a) 4种可能的开关连接



(b) 一种8输入的Omega网络

整体：butterfly结构

从单一输入来看：二叉树结构，中间状态合并，用状态切换适配输入输出要求

数量级： $O(n \log n)$

最大并发数量级： $O(n)$

动态互连网络 (5)

- * 交换开关模块：
 - * 一个交换开关模块有n个输入和n个输出，每个输入可连接到任意输出端口，但只允许一对一或一对多的映射，不允许许多对一的映射，因为这将发生输出冲突
- * 级间互连 (Interstage Connection)：
 - * 均匀洗牌、蝶网、多路均匀洗牌、交叉开关、立方连接
 - * n输入的Ω网络需要 $\log_2 n$ 级 2×2 开关，在Illinois大学的Cedar[2]多处理机系统中采用了Ω网络
 - * Cray Y/MP多级网络，该网络用来支持8个向量处理器和256个存储器模块之间的数据传输。网络能够避免8个处理器同时进行存储器存取时的冲突。

动态互连网络比较

动态互连网络的复杂度和带宽性能一览表

网络特性	总线系统	多级互连网络	交叉开关
开关复杂度	$O(n)$	$O(n \log_k n)$	$O(n^2)$
每个处理器带宽	$O(w/n) \sim O(w)$	$O(w) \sim O(nw)$	$O(w) \sim O(nw)$
报道的聚集带宽	SunFire服务器中的Gigaplane 总线: 2.67GB/s	IBM SP2中的 512节点的HPS: 10.24GB/s	Digital的千兆开关: 3.4GB/s

* n , 节点规模 w , 数据宽度

标准互联网络 (1)

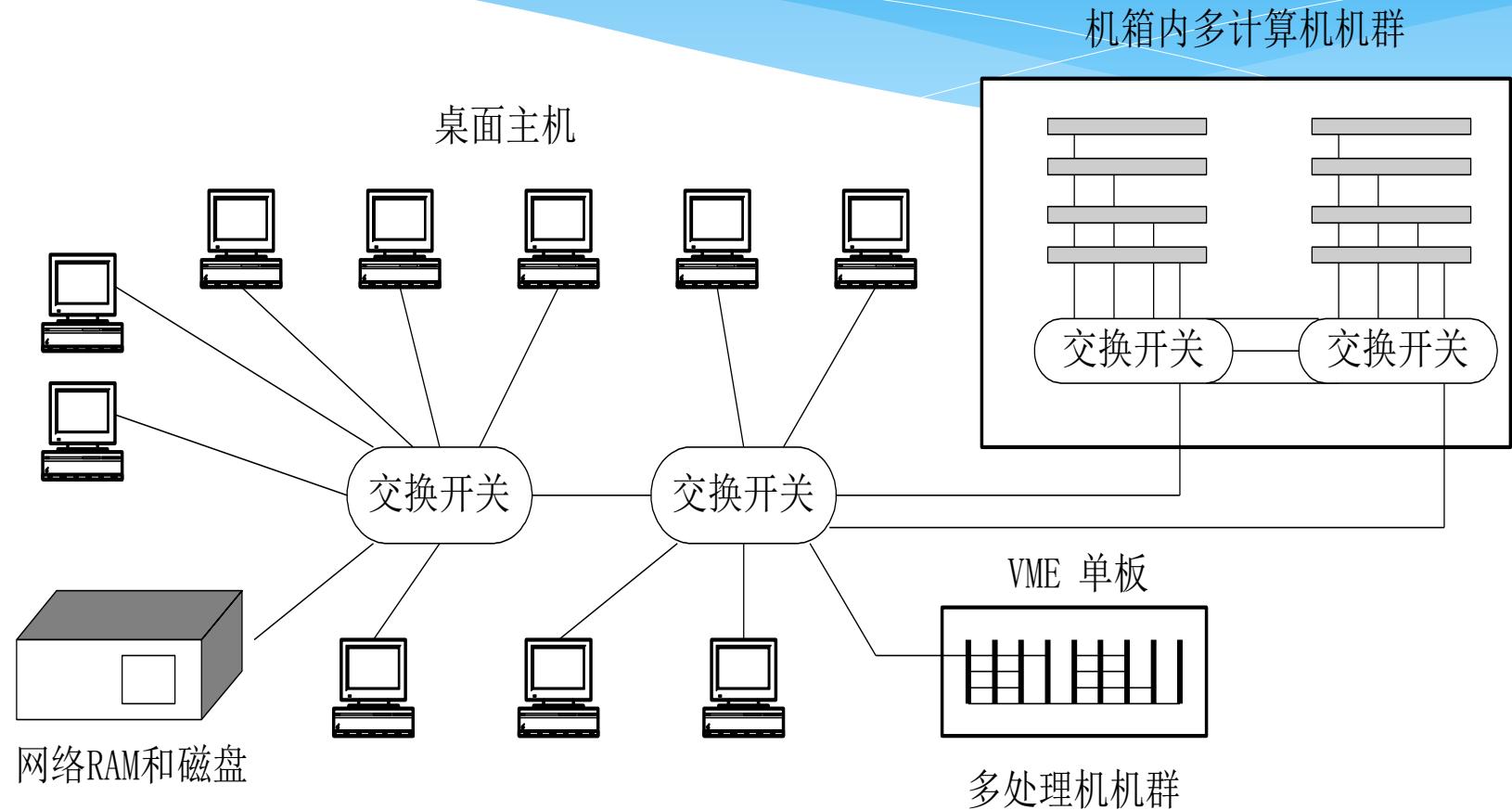
- * Myrinet:

- * Myrinet是由Myricom公司设计的千兆位包交换网络，其目的是为了构筑计算机机群，使系统互连成为一种商业产品。
- * Myrinet是基于加州理工学院开发的多计算机和VLSI技术以及在南加州大学开发的ATOMIC/LAN技术。Myrinet能假设任意拓扑结构，不必限定为开关网孔或任何规则的结构。
- * Myrinet在数据链路层具有可变长的包格式，对每条链路施行流控制和错误控制，并使用切通选路法以及定制的可编程的主机接口。在物理层上，Myrinet网使用全双工SAN链路，最长可达3米，峰值速率为 $(1.28+1.28)$ Gbps（目前有 $2.56+2.56$ ）

标准互联网络 (1)

- * Myrinet:
 - * Myrinet 交换开关 :8,12,16 端口
 - * Myrinet 主机接口 :32位的称作 LANai 芯片的用户定制的 VLSI 处理器，它带有 Myrinet 接口、包接口、DMA 引擎和快速静态随机存取存储器 SRAM。
 - * 140 of the November 2002 TOP500 use Myrinet, including 15 of the top 100

Myrinet 连接的 LAN/Cluster



标准互连网络 (2)

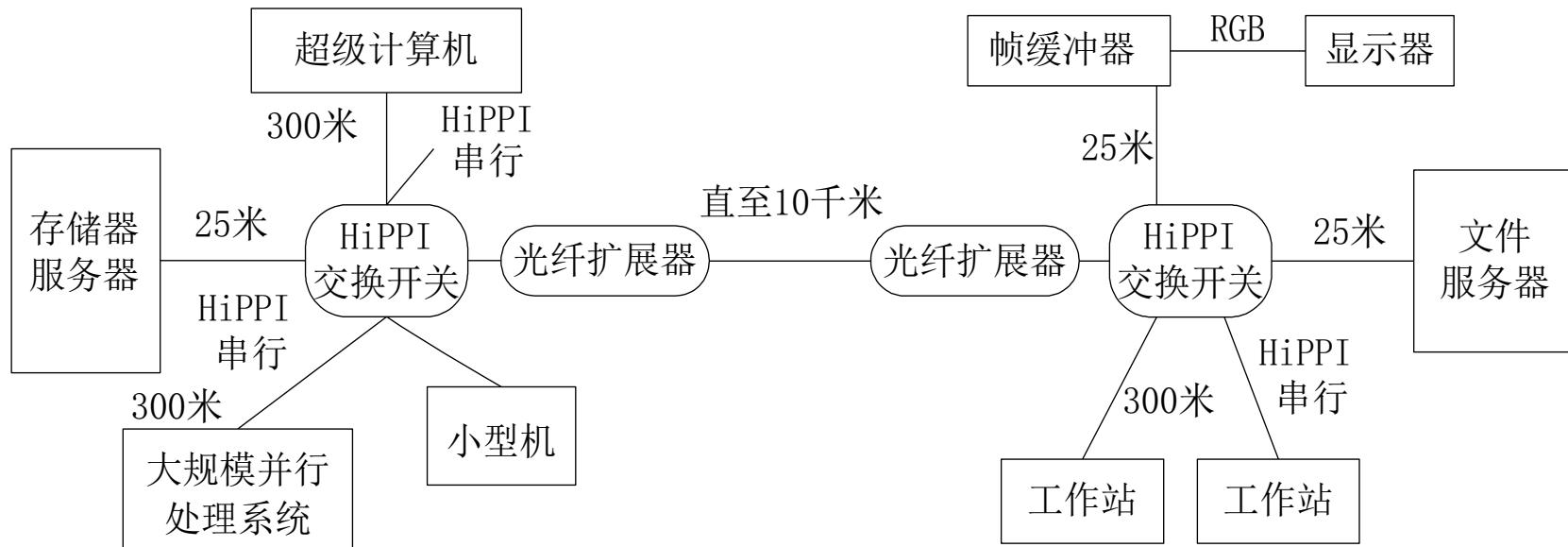
- * 高性能并行接口 (HiPPI)

- * Los Alamos国家实验室于1987年提出的一个标准，其目的是试图统一来自不同生产商生产的所有大型机和超级计算机的接口。在大型机和超级计算机工业界，HiPPI作为短距离的系统到系统以及系统到外设连接的高速I/O通道。
- * 1993年，ANSI X3T9.3委员会认可了HiPPI标准，它覆盖了物理和数据链路层，但在这两层之上的任何规定却取决于用户。
- * HiPPI是个单工的点到点的数据传输接口，其速率可达800Mbps到1.6Gbps。

标准互连网络 (2)

- * 高性能并行接口 (HiPPI)
 - * 开发成功了一种能提供潜在的6.4Gbps速率，比HiPPI快8倍且有很低时延的超级HiPPI技术，
 - * SGI公司和Los Alamos国家实验室都开发了用来构筑速率高达25.6Gbps的HiPPI交换开关的HiPPI技术。
 - * HiPPI通道和HiPPI交换开关被用在SGI Power Challenge服务器、IBM 390主机、Cray Y/MP、C90和T3D/T3E等系统

使用HiPPI通道和开关构筑的LAN 主干网



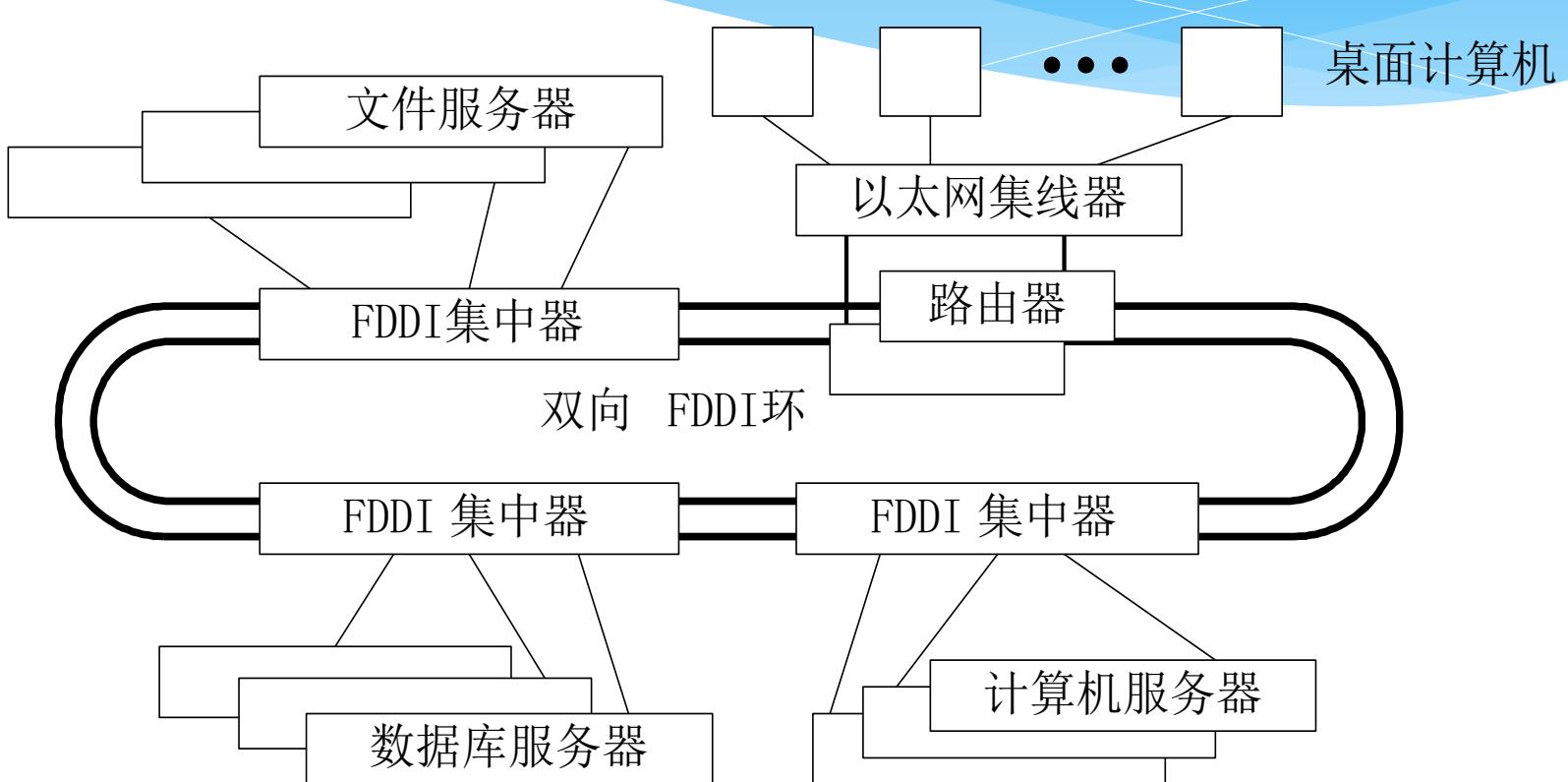
标准互连网络 (3)

- * 光纤通道FC (Fiber Channel)：
 - * 通道和网络标准的集成
 - * 光纤通道既可以是共享介质，也可以是一种交换技术
 - * 光纤通道操作速度范围可从100到133、200、400和800Mbps。FCSI厂商也正在推出未来具有更高速度（1、2或4Gbps）的光纤通道
 - * 光纤通道的价值已被现在的某些千兆位局域网所证实，这些局域网就是基于光纤通道技术的
 - * 连网拓扑结构的灵活性是光纤通道的主要财富，它支持点到点、仲裁环及交换光纤连接

标准互连网络 (3)

- * FDDI：
 - * 光纤分布式数据接口FDDI（Fiber Distributed Data Interface）
 - * FDDI采用双向光纤令牌环可提供100-200Mbps数据传输速率
 - * FDDI具有互连大量设备的能力
 - * 传统的FDDI仅以异步方式操作

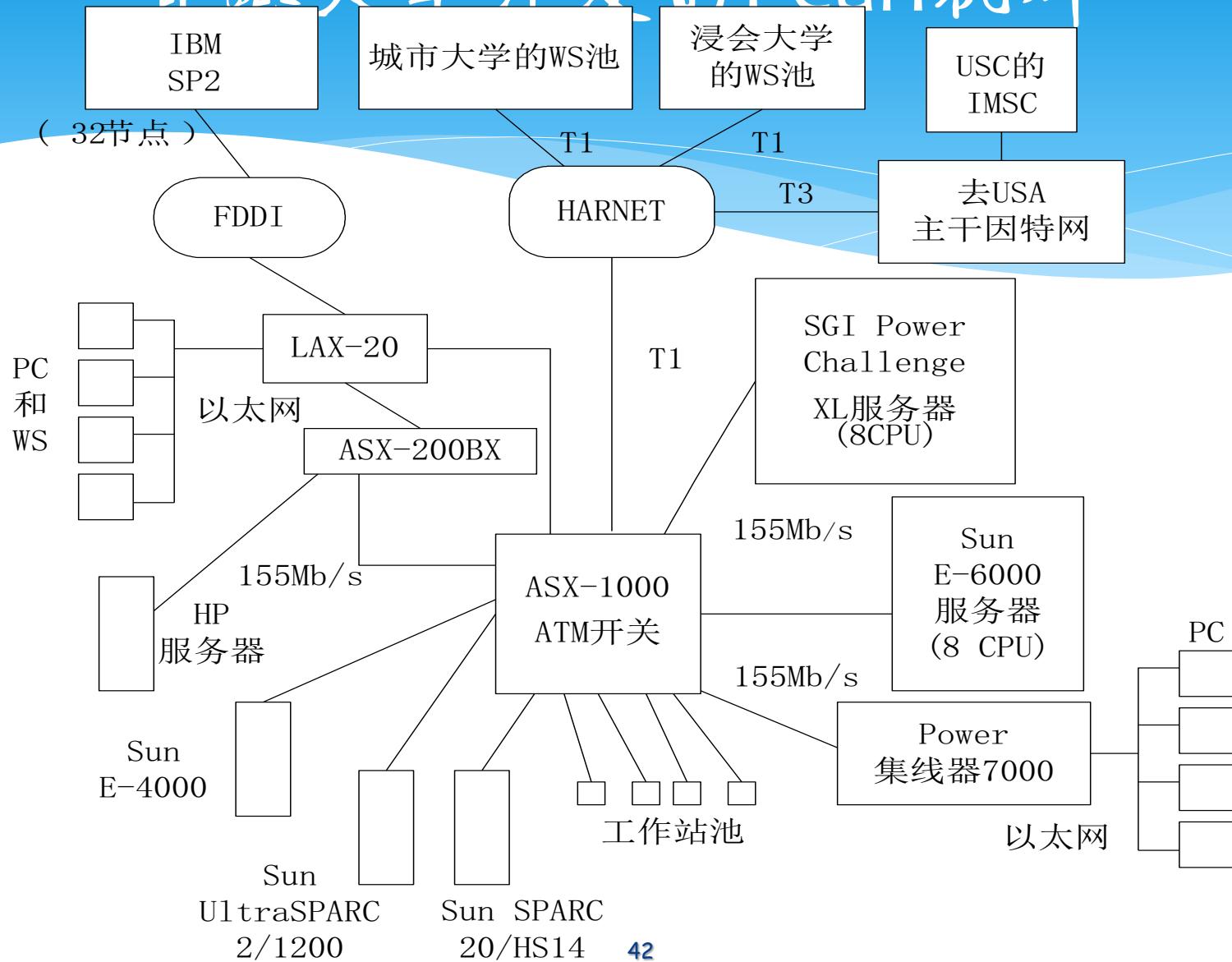
双向FDDI环作为主干网



标准互联网络（4）

- * ATM (Asynchronous Transfer Mode) :
 - * 由成立于1991年的ATM论坛和ITU标准定义。
 - * ATM是一种独立于介质的消息传输协议，它将消息段变成更短的固定长度为53字节的报元进行传输。
 - * 这种技术是基于报元交换机制。ATM的目的是将实时和突发数据的传输合并成单一的网络技术。
 - * ATM网络支持从25到51、155和622Mbps不同的速率，其速率越低ATM交换器和使用的链路价格越低。

香港大学开发的Pearl机群



标准互联网络（4）

* 以太网

- * 以太网（Ethernet）是一种计算机局域网组网技术。IEEE制定的IEEE 802.3标准给出了以太网的技术标准。它规定了包括物理层的连线、电信号和介质访问层协议的内容。以太网是当前应用最普遍的局域网技术。它很大程度上取代了其他局域网标准，如令牌环网（token ring）、FDDI和ARCNET。
- * 以太网的标准拓扑结构为总线型拓扑，但目前的快速以太网（100BASE-T、1000BASE-T标准）为了最大程度的减少冲突，最大程度的提高网络速度和使用效率，使用交换机（Switch hub）来进行网络连接和组织，这样，以太网的拓扑结构就成了星型，但在逻辑上，以太网仍然使用总线型拓扑和CSMA/CD（Carrier Sense Multiple Access/Collision Detect 即带冲突检测的载波监听多路访问）的总线争用技术。

标准互连网络 (5)

代别 类型	以太网 10BaseT	快速以太网 100BaseT	千兆位以太网 1GB
引入年代	1982	1994	1997
速度 (带宽)	10Mb/s	100Mb/s	1Gb/s
最大距离	UTR (非屏蔽双扭对)	100m	25—100m
	STP (屏蔽双扭对) 同轴电缆	500m	25—100m
	多模光纤	2Km 412m (半双工) 2Km (全双工)	500m
	单模光纤	25Km	20Km 3Km
主要应用领域	文件共享, 打印机共享	COW计算, C/S结构, 大型数据库存取等	大型图像文件, 多媒体, 因特网, 内部网, 数据仓库等

SISD、MIMD、SIMD、MISD

- * 1966年，Michael Flynn根据指令和数据流的概念对计算机的体系结构进行了分类，这就是所谓的Flynn分类法。Flynn将计算机划分为四种基本类型，即SISD、MIMD、 SIMD、 MISD。
- * 传统的顺序执行的计算机在同一时刻只能执行一条指令（即只有一个控制流）、处理一个数据（即只有一个数据流），因此被称为单指令流单数据流计算机（Single Instruction Single Data, SISD）。

根据指令和数据区分
Instruction Data Single/Multiple

SISD、MIMD、SIMD、MISD

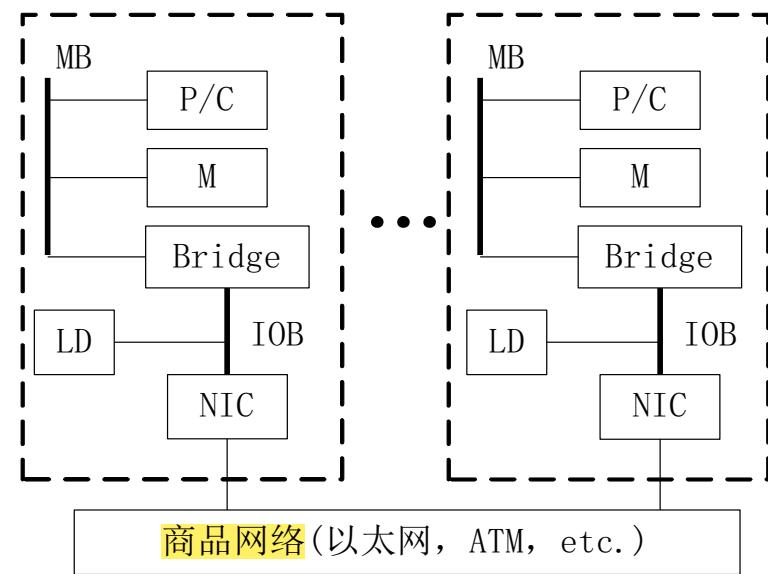
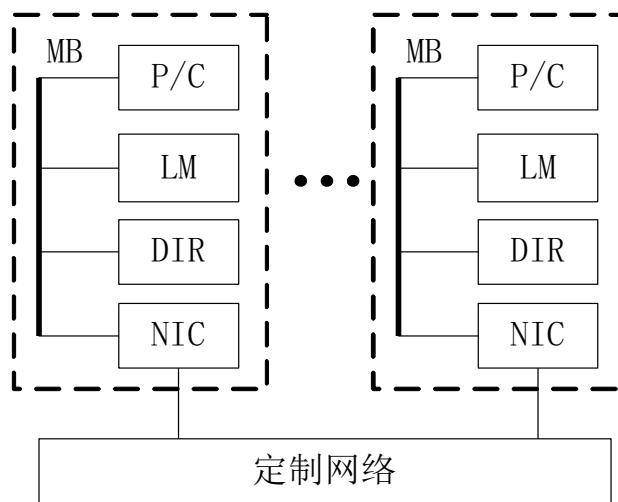
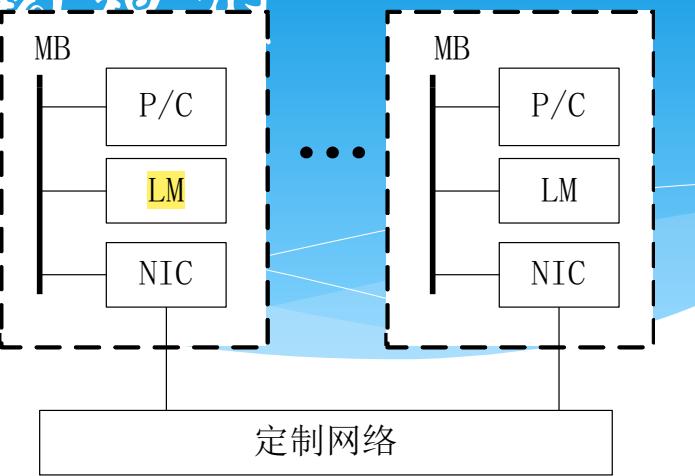
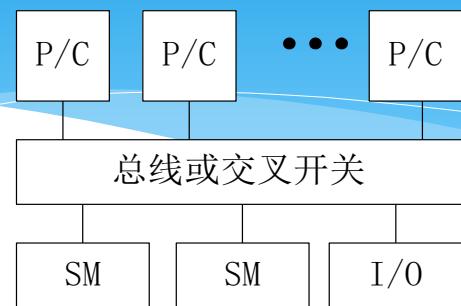
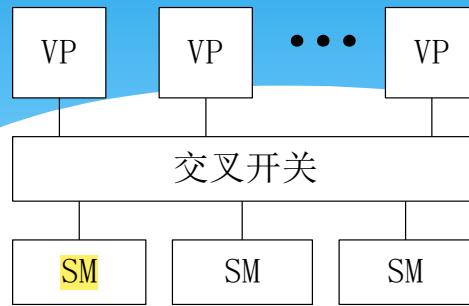
- * 而对于大多数并行计算机而言，多个处理单元都是根据不同的控制流程执行不同的操作，处理不同的数据，因此，它们被称作是多指令流多数据流计算机，即MIMD（Multiple Instruction Multiple Data,MIMD）计算机。
- * 曾经在很长一段时间内成为超级并行计算机主流的向量计算机除了标量处理单元之外，最重要的是具有能进行向量计算的硬件单元。在执行向量操作时，一条指令可以同时对多个数据（组成一个向量）进行运算，这就是单指令流多数据流（Single Instruction Multiple Data,SIMD）的概念。因此，我们将向量计算机称为SIMD₄₆计算机。

SISD、MIMD、SIMD、MISD

- * 第四种类型即所谓的多指令流单数据（Multiple Instruction Single Data, MISD）计算机。在这种计算机中，各个处理单元组成一个线性阵列，分别执行不同的指令流，而同一个数据流则顺次通过这个阵列中的各个处理单元。这种系统结构只适用于某些特定的算法。
- * 相对而言， SIMD和MISD模型更适合于专用计算。在商用并行计算机中， MIMD模型最为通用， SIMD次之， 而MISD最少用。 PII的MMX指令采用的是SISD， 高性能服务器与超级计算机大多属于MIMD。

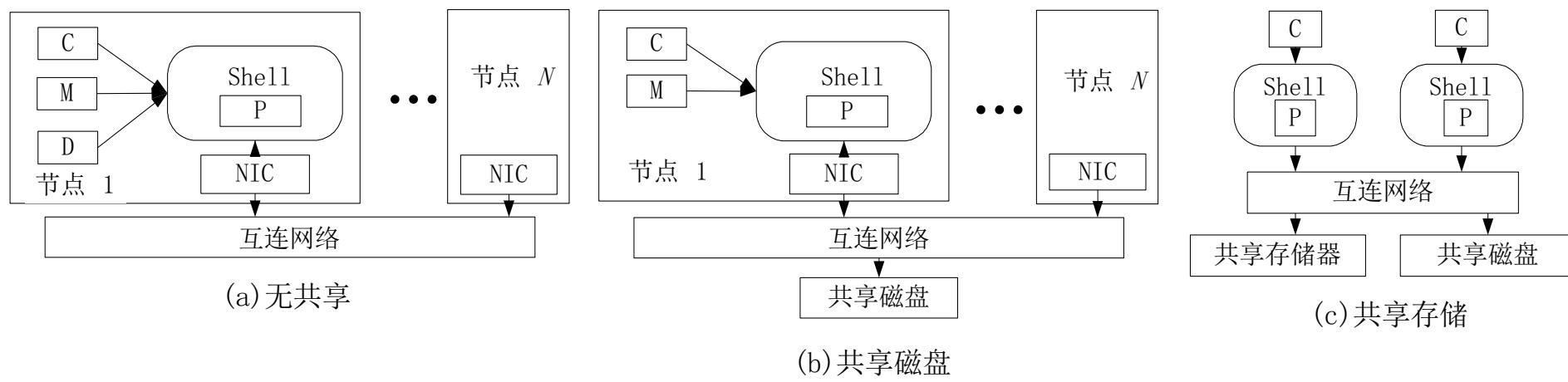
*

并行计算机结构模型



并行计算机体系合一结构

- * SMP、MPP、DSM和COW并行结构渐趋一致。
- * 大量的节点通过高速网络互连起来
- * 节点遵循Shell结构：用专门定制的Shell电路将商用微处理器和节点的其它部分（包括板级Cache、局存、NIC和DISK）连接起来。优点是CPU升级只需要更换Shell。



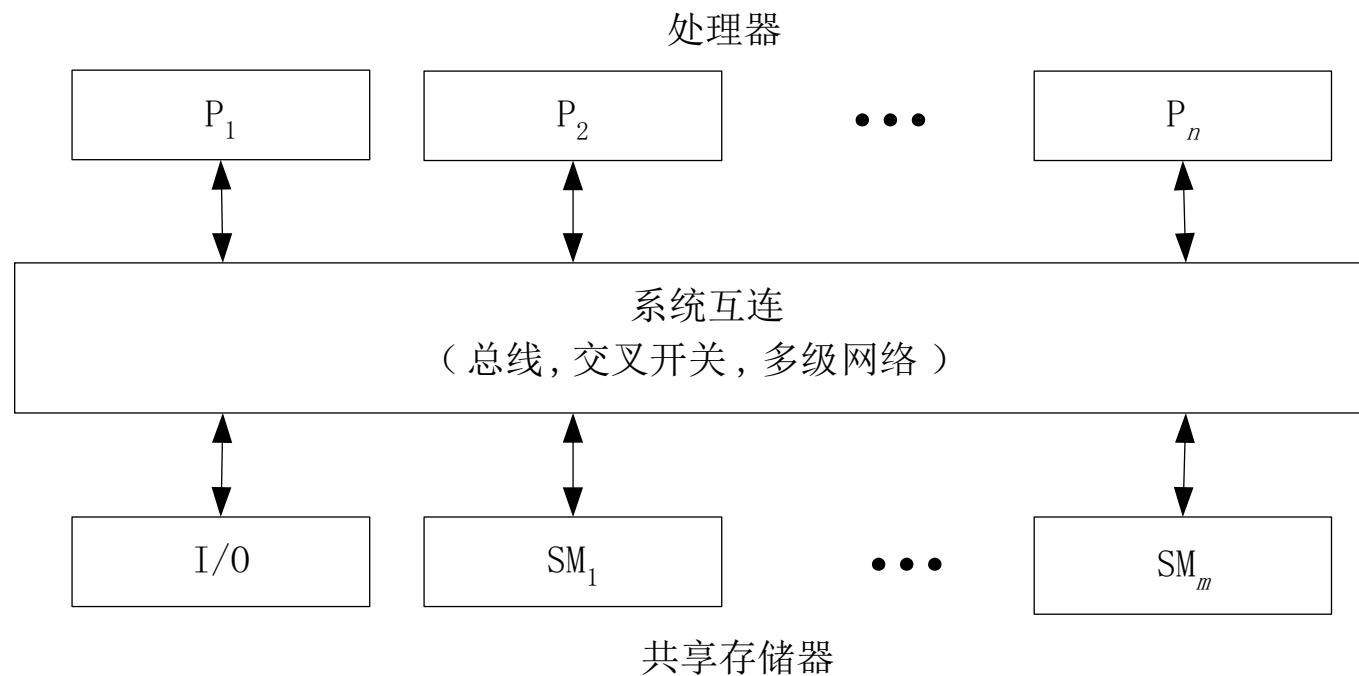
五种结构特性一览表

属性	PVP	SMP	MPP	DSM	COW
结构类型	MIMD	MIMD	MIMD	MIMD	MIMD
处理器类型	专用定制	商用	商用	商用	商用
互连网络	定制交叉开关	总线、交叉开关	定制网络	定制网络	商用网络（以太ATM）
通信机制	共享变量	共享变量	消息传递	共享变量	消息传递
地址空间	单地址空间	单地址空间	多地址空间	单地址空间	多地址空间
系统存储器	集中共享	集中共享	分布非共享	分布共享	分布非共享
访存模型	UMA	UMA	NORMA	NUMA	NORMA
代表机器	Cray C-90, Cray T-90, 银河1号	IBM R50, SGI Power Challenge,	Intel Paragon, IBMSP2,	Stanford DASH, Cray T 3D	Berkeley NOW, Alpha Farm

并行计算机访存模型（1）

- * UMA (Uniform Memory Access) 模型是均匀存储访问模型的简称。其特点是：
 - * 物理存储器被所有处理器均匀共享；
 - * 所有处理器访问任何存储字取相同的时间；
 - * 每台处理器可带私有高速缓存；
 - * 外围设备也可以一定形式共享。

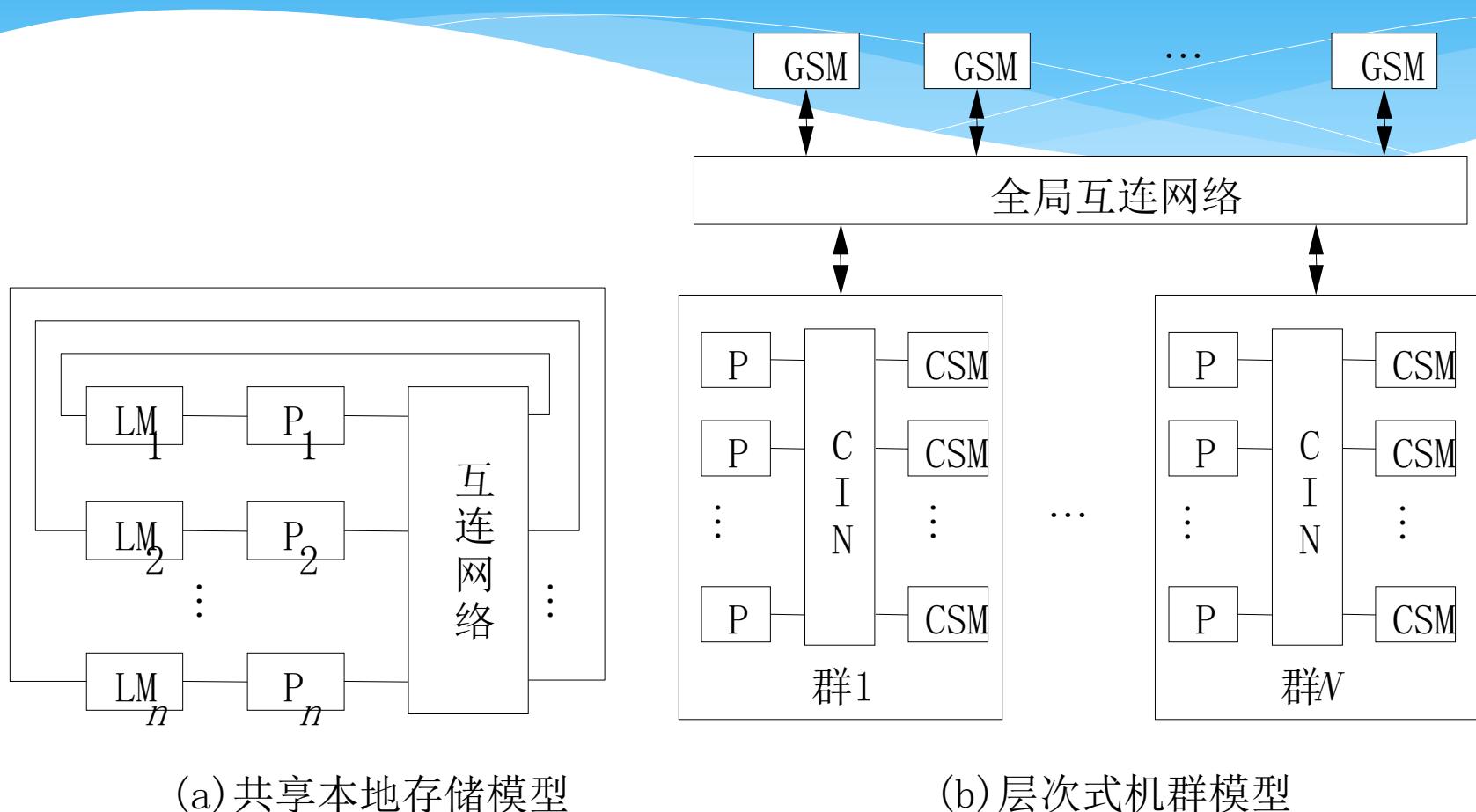
并行计算机访存模型 (1)



并行计算机访存模型 (2)

- * NUMA(Nonuniform Memory Access)模型是非均匀存储访问模型的简称。特点是：
 - * 被共享的存储器在物理上是分布在所有的处理器中的，其所有本地存储器的集合就组成了全局地址空间；
 - * 处理器访问存储器的时间是不一样的；访问本地存储器LM或群内共享存储器CSM较快，而访问外地的存储器或全局共享存储器GSM较慢(此即非均匀存储访问名称的由来)；
 - * 每台处理器照例可带私有高速缓存，外设也可以某种形式共享。

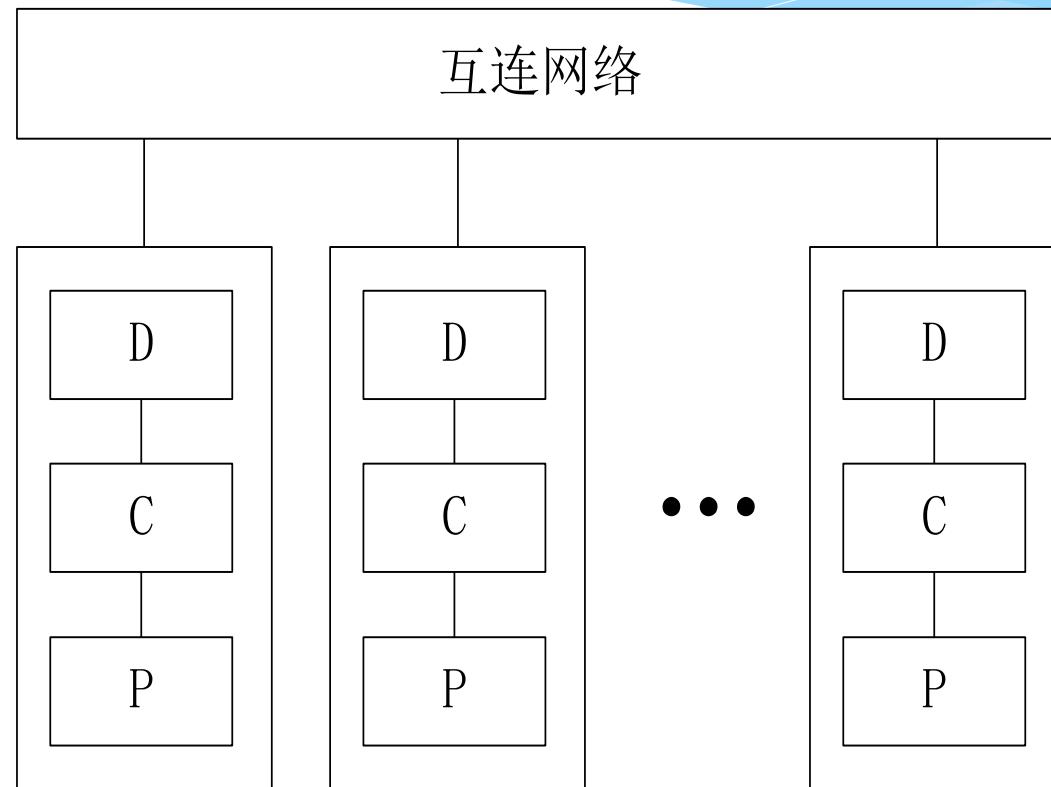
并行计算机访存模型 (2)



并行计算机访存模型 (3)

- * COMA(Cache-Only Memory Access)模型是全高速缓存存储访问的简称。其特点是：
 - * 各处理器节点中没有存储层次结构，全部高速缓存组成了全局地址空间；
 - * 利用分布的高速缓存目录D进行远程高速缓存的访问；
 - * COMA中的高速缓存容量一般都大于2 级高速缓存容量；
 - * 使用COMA时，数据开始时可任意分配，因为在运行时它最终会被迁移到要用到它们的地方。

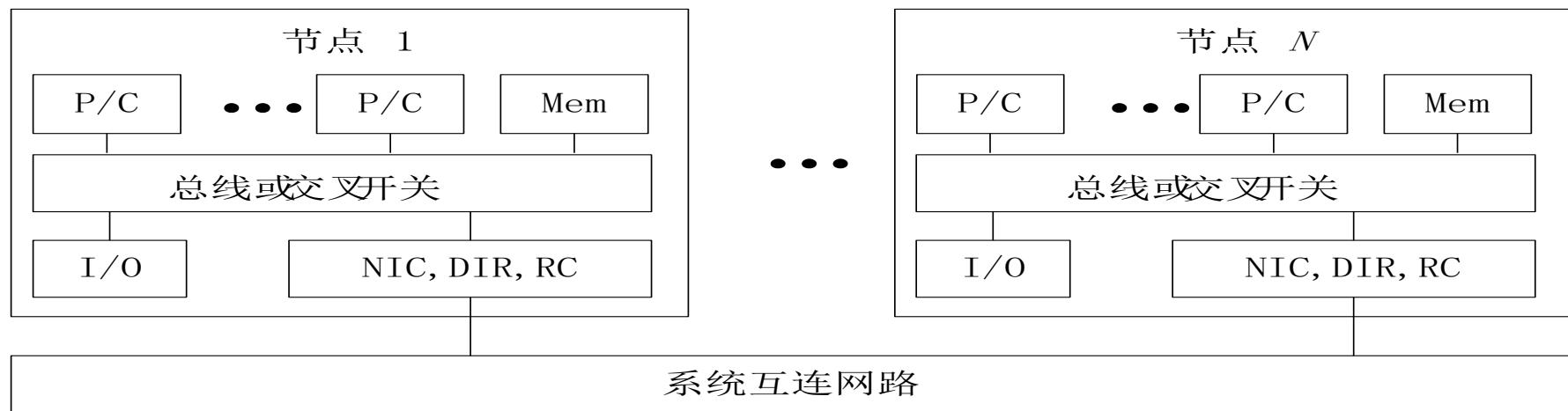
并行计算机访存模型 (3)



并行计算机访存模型（4）

- * CC-NUMA (Coherent-Cache Nonuniform Memory Access) 模型是高速缓存一致性非均匀存储访问模型的简称。其特点是：
 - * 大多数使用基于目录的高速缓存一致性协议；
 - * 保留SMP结构易于编程的优点，也改善常规SMP的可扩放性；
 - * CC-NUMA实际上是一个分布共享存储的DSM多处理机系统；
 - * 它最显著的优点是程序员无需明确地在节点上分配数据，系统的硬件和软件开始时自动在各节点分配数据，在运行期间，高速缓存一致性硬件会自动地将数据迁移至要用到它的地方。

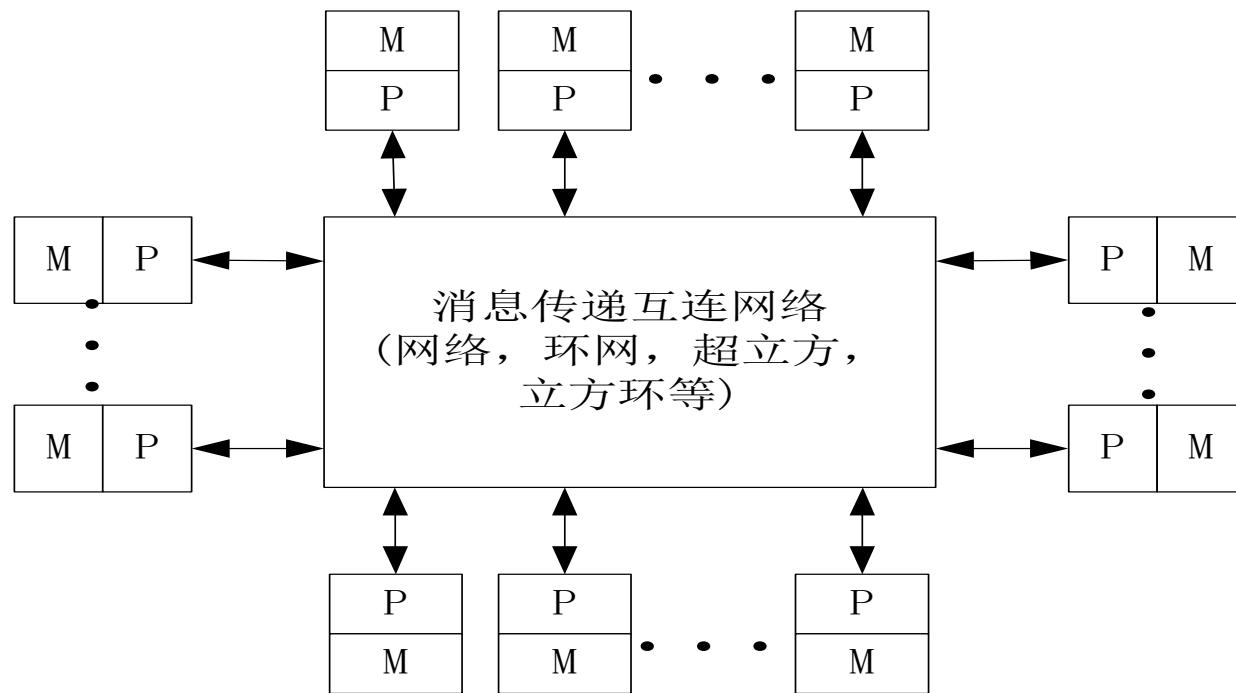
并行计算机访存模型 (4)



并行计算机访存模型（5）

- NORMA (No-Remote Memory Access) 模型是非远程存储访问模型的简称。NORMA的特点是：
 - 所有存储器是私有的；
 - 绝大数NORMA都不支持远程存储器的访问；
 - 在DSM中，NORMA就消失了。

并行计算机访存模型 (5)



构筑并行机系统的不同存储结构

