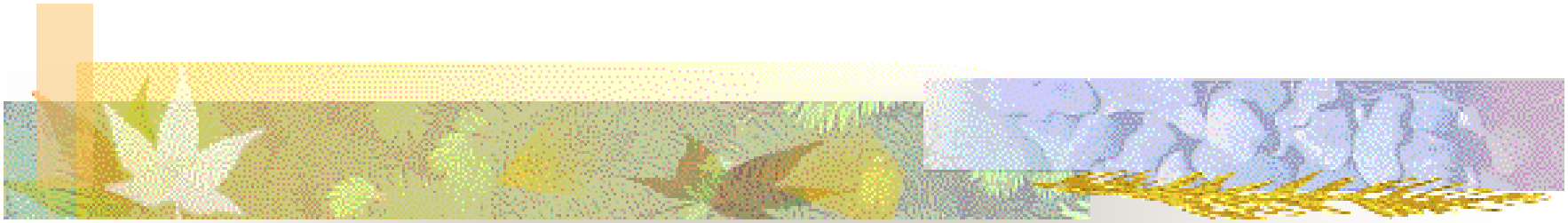
第六章 多处理机



内容

- 1 多处理机的概念
- 2 多处理机结构
- 3 多处理机系统的存储器结构
- 4 多处理机的软件
- 5 超级计算机并行体系结构

1 多处理机的概念

- **1.1 多处理机的定义**
- **1.2 多重处理对处理机特性的要求**

1.1 多处理机系统的定义

P.H.Enslow对多处理机作了下列定义：

- 包含两个或两个以上功能大致相同的处理器
- 所有处理器共享一个公共内存
- 所有处理器共享I/O通道、控制器和外围设备
- 整个系统由统一的操作系统控制，在处理器和程序之间实现作业、任务、程序段、数组和数组元素等各级的全面并行

多处理机属于多指令流多数据流计算机(MIMD)

多处理机与并行处理机对比：5方面

- 结构：

  多处理机：通用，PE数少，高速灵活通信

  并行处理机：专用，PE数多，固定有限通信

- 程序的并行性：

  多处理机的并行性存在于指令外部，在多个任务之间，容易识别；

  并行处理机的并行性存在于指令内部，不易识别。

并行任务派生

- 并行处理机把同种操作集中，由指令直接启动各PE同时工作
- 多处理机用专门的指令（FORK, JOIN）表示并发关系，一个任务开始执行时能够派生出与它同时执行的一些任务。
- 如果任务数多于处理机数，多余的任务进入排队器等待。

- 进程同步
  - 并行处理机仅有一个CU，自然是同步的。
  - 多处理机中，各个处理机执行不同的指令，工作进度不会也不必保持相同。要采取同步措施来保持程序要求的正确顺序。

- 资源分配和进程调度
  - 并行处理机的PE是固定的，用屏蔽来改变实际参加操作的PE数目
  - 多处理机执行并发任务，需用处理机的数目不固定，各处理机进出任务的时刻不相同，所需共享资源的品种、数量随时变化
  - 资源分配和进程调度问题，对整个系统的效率有很大的影响

多处理机的优点

- 很高的性能价格比： 单处理机的性能价格比随其规模的增大而下降
- 很高的可靠性：冗余度大、可维护性、可用性
- 很高的处理速度：多个处理器并行运算
- 很好的模块性：大量重复设置，结构灵活性、可扩充性、可重构性

1.2 多重处理对处理机特性的要求

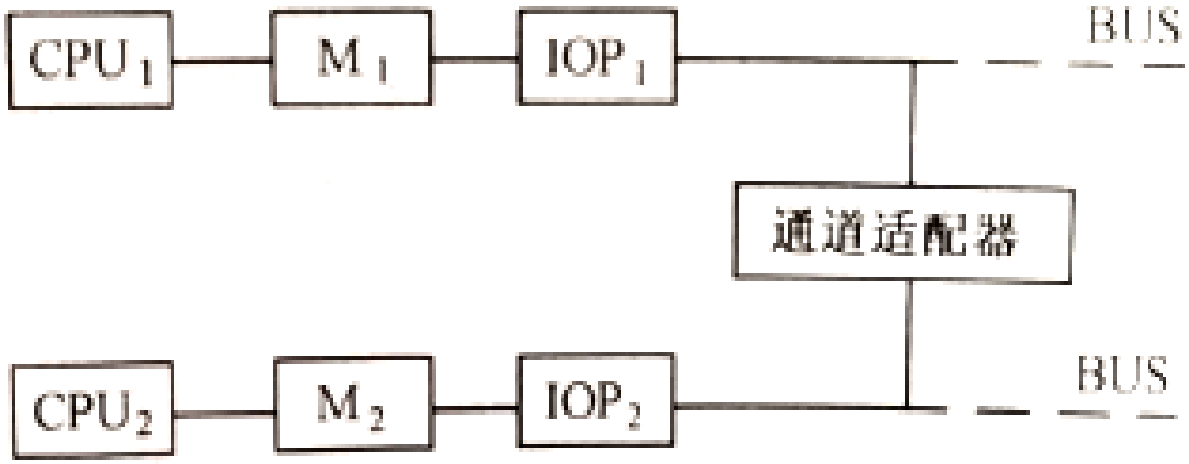
- 进程恢复能力
- 有效的现场切换
- 大的物理地址空间和虚拟地址空间
- 高效率的同步原语
- 处理机之间有高效率的通信机构
- 指令系统（应能支持过程级并发）

2 多处理机结构

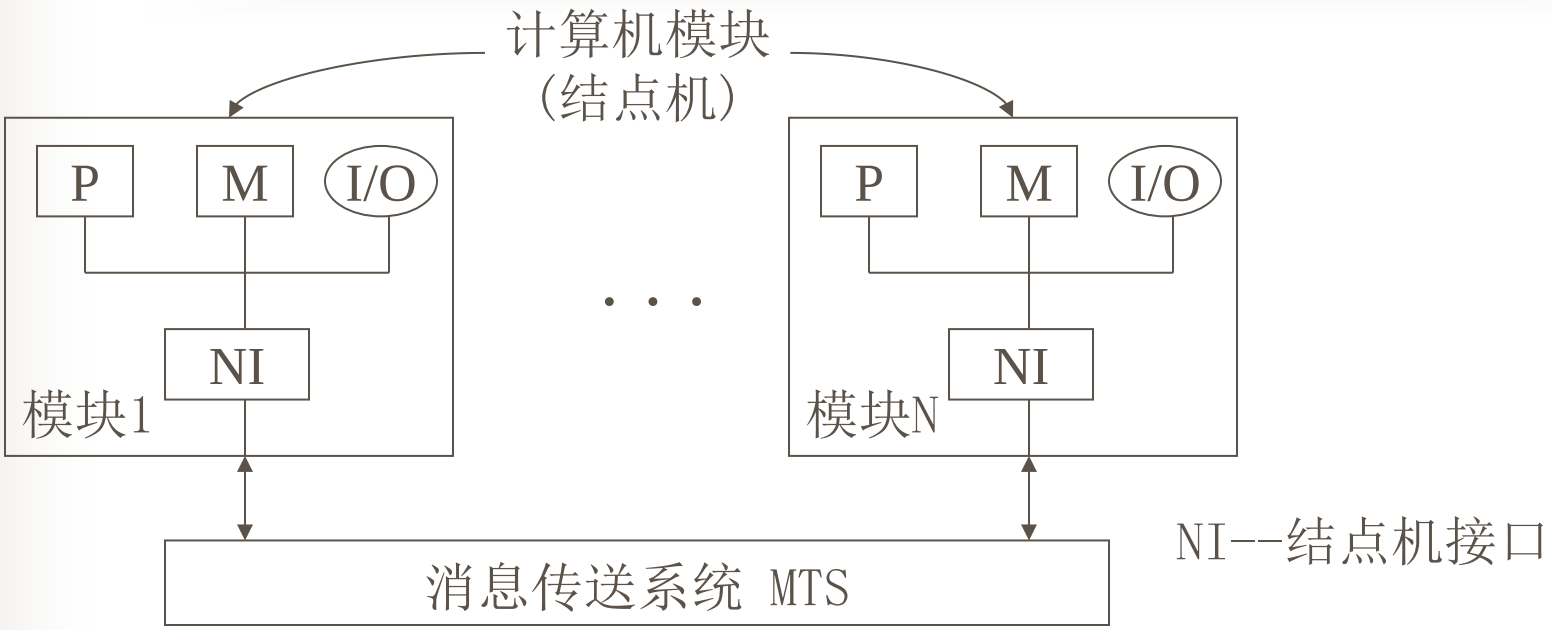
- **2.1 松耦合**
- **2.2 紧耦合**
- **2.3 多处理机的互联结构**

2.1 松耦合系统(LCS)

a） 通道连接：每台计算机独立，通过通道适配器连接。通信时，发送方可把接收方认作是自己的一个I/O设备，能完成两个主存储器之间的数据传送。



b）通信线路连接：



特点：通过消息传送系统实现机间通讯；每个模块是独立的处理机，整个系统可看成是一个分布系统。

互连网络：MTS有总线、环形、多级网络等种类；

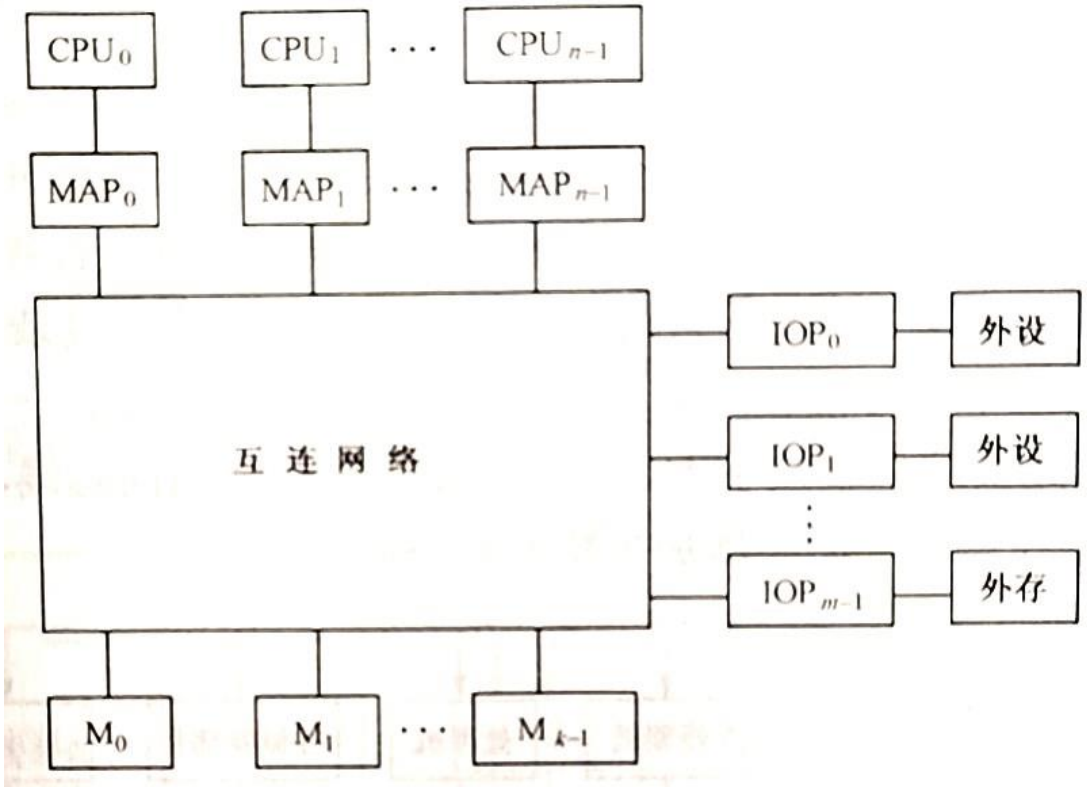
结构：有层次和非层次两种结构。

与计算机网络区别：

单一的系统物理地址空间；

每个PE的存储器均可被其它PE访问，通过通道仲裁开关CAS实现。

2.2 紧耦合系统(TCS)



特点：通过共享主存实现机间通讯。

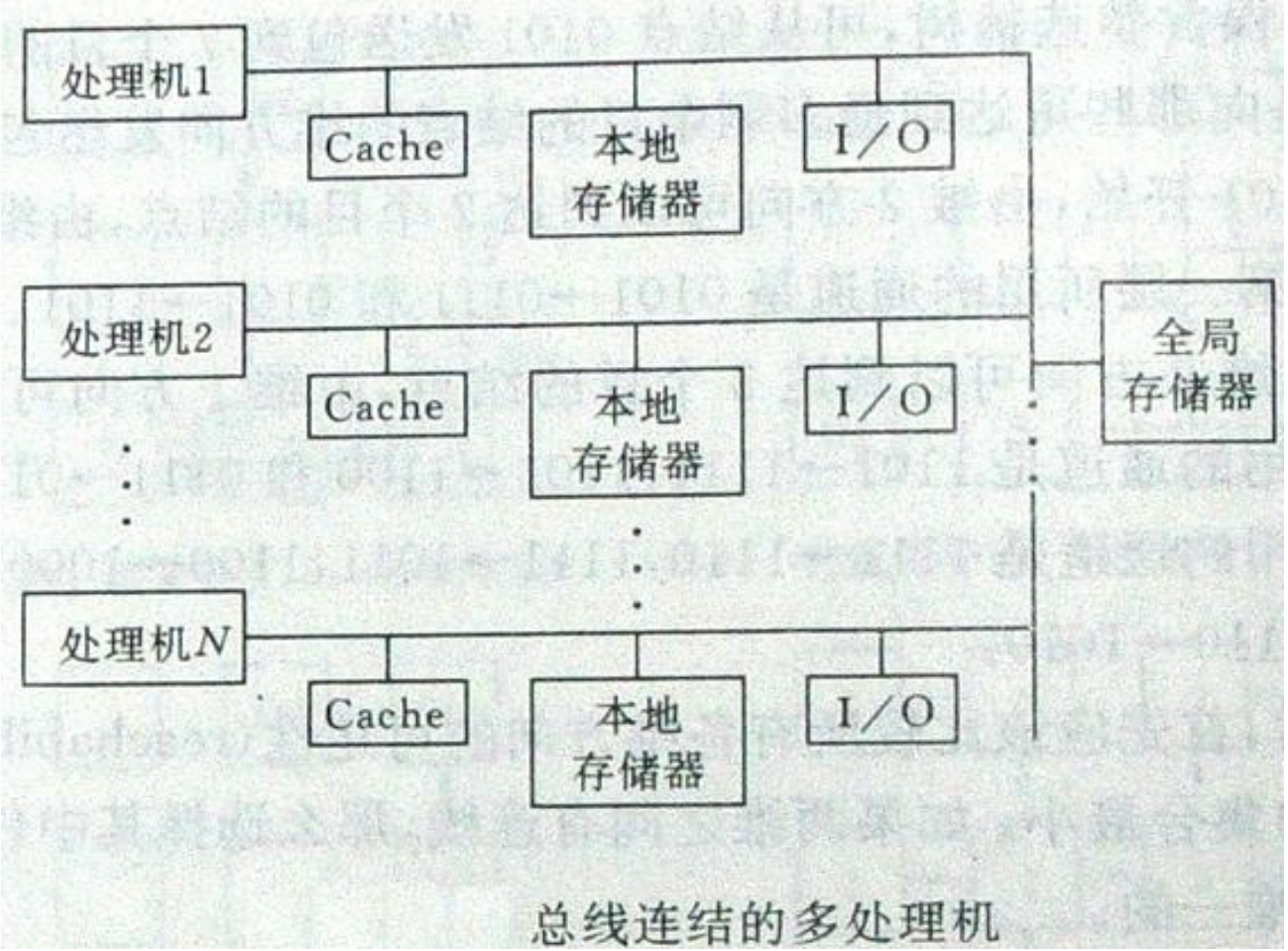
互连网络：高速总线或高速开关

- 当某个处理机要访问主存储器，只需通过它的存储映象部件（MAP），就可以把全局的逻辑地址变换成局部的物理地址（即某一存储模块内的物理地址）。
- 互连网络不仅要提供高速的传输通路，而且具有选择有效路径、仲裁访问冲突等功能。对于输入输出设备的访问也与访问存储器一样，只是它们的界面通过输入输出处理机（IOP）来进行。

2.3 多处理机的互连网络

- 多处理机的主要特点是各处理机共享一组存储器和I/O设备。
- 这种共享是通过两个互连网络实现的：一个是处理机和存储器模块之间的互联网络；另一个是处理机和I/O子系统（I/O接口和I/O设备）之间的互联网络。
- 互连网络可以采用不同的物理形式，一般可有四种基本结构。

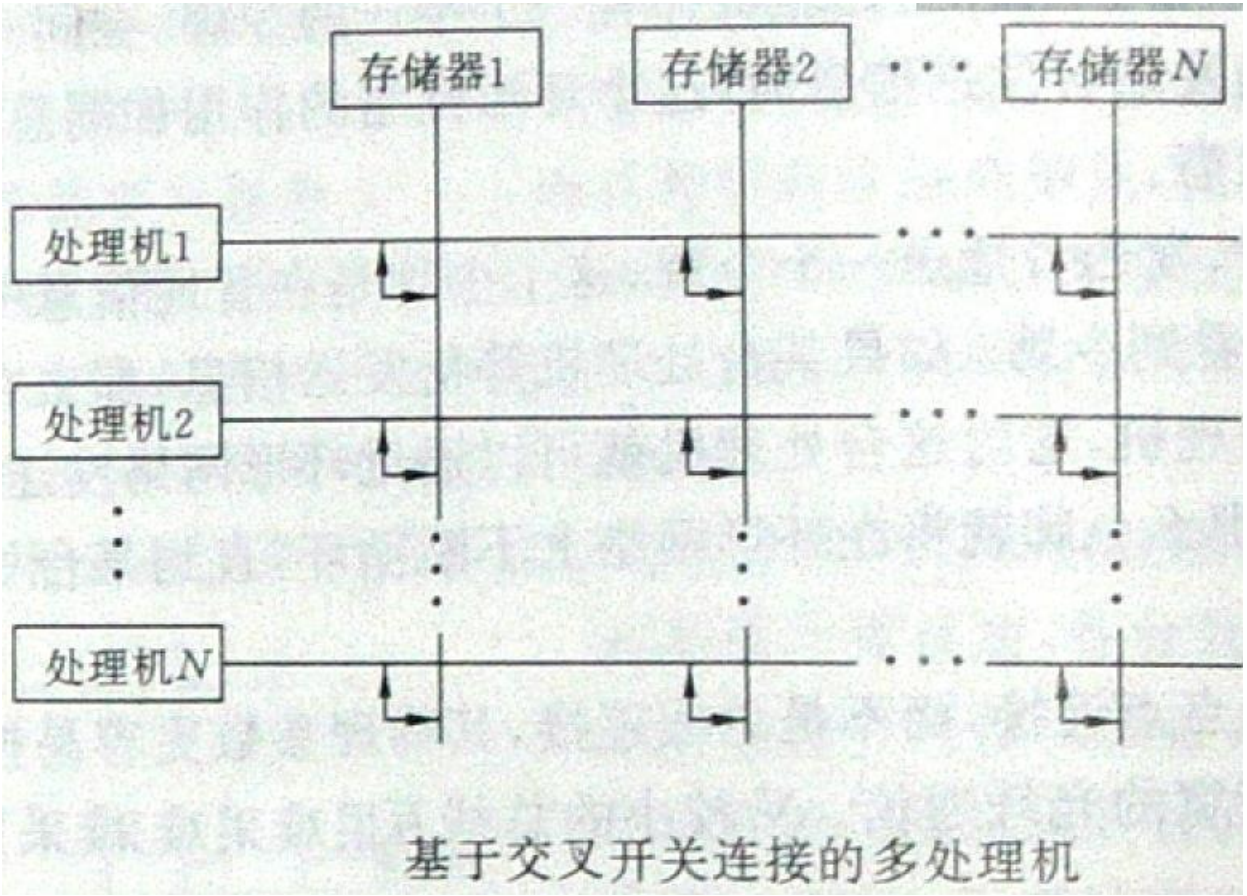
（1）总线结构



总线连结的多处理机

- 把所有功能模块（或部件）连接到一条公共通信通路上。
- 公共通信通路也称为时分或公共总线。这种总线结构的特点是简单、容易实现，也容易扩展（重构）。
- 总线是一个无源部件，通信完全由发送和接收的总线接口控制。
- 由于总线是共享资源，所以必须有总线请求和仲裁的机构，以避免发生总线冲突。

（2）交叉开关



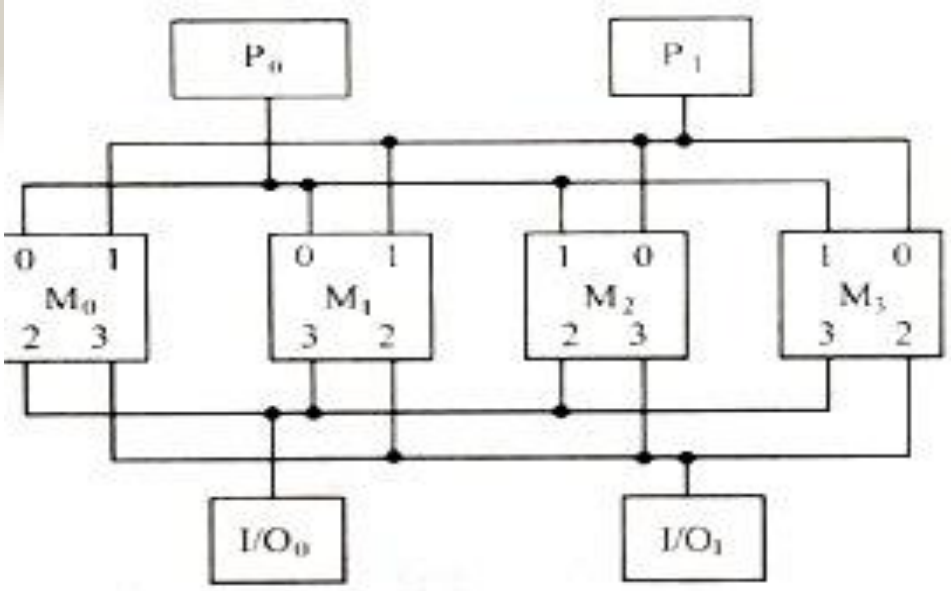
基于交叉开关连接的多处理机

- 当不断增加总线数目，使每个存储器模块有它自己单独可用的通路形成的互联网络称为无阻塞交叉开关。
- 它的特点是开关和功能部件的接口非常简单，而且支持所有存储器模块同时通信。每个交叉点不仅能切换并行传播，而且必须能解决在同一存储器周期内访问同一个存储器模块的多个请求之间的冲突。通常用预设的优先级来处理冲突。

（3）多端口存储器

- 如果把分布在交叉开关矩阵网络上的控制、转接、优先级仲裁等逻辑功能转移到存储器模块的接口上，就形成了多端口存储器系统

- 对于访问冲突，常用的解决方法是每个存储器端口分配一个永久优先级，而各个主控模块相对于某个存储器模块有一个优先级别序列。



具有优先级的多端口存储器系统

- 例如对于M0而言，其能接收主控模块的访问优先次序为P_0、P_1、I/O_0、I/O_1；对于M1而言，则为P_0、P_1、I/O_1、I/O_0；对于M3而言，则为P_1、P_0、I/O_1、I/O_0；对于M3而言，则为P_1、P_0、I/O_1、I/O_0。

（4）多处理机的多级网络

- 由于开关过于复杂，对于大规模交叉开关用多个小规模交叉开关“串联”和“并联”，组成多级交叉开关网络，以取代单级的大规模交叉开关。
- **单级互联网络不能实现任意处理器之间的互联。为此：**
  - **循环互联网络**
  - **多级互联网络。**

目标：完成某结点与其它任一结点的连接；
同时完成多对结点的连接。

方法：从时间性和空间性方面开发。

循环互连网络（时间性）

组成：DTR_{in}、DTR_{out}、MUX、IN。

结构：

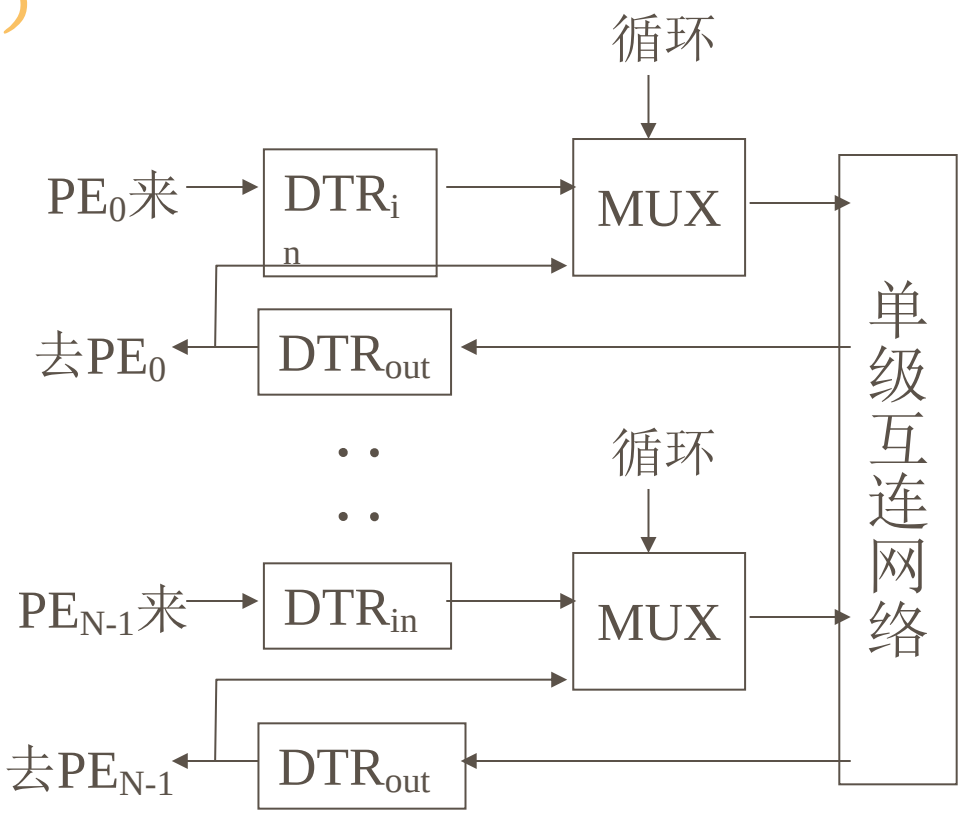
一个单级ICN+MUX

特点：

节省了设备，

增加了时间，

每个MUX可单独控制



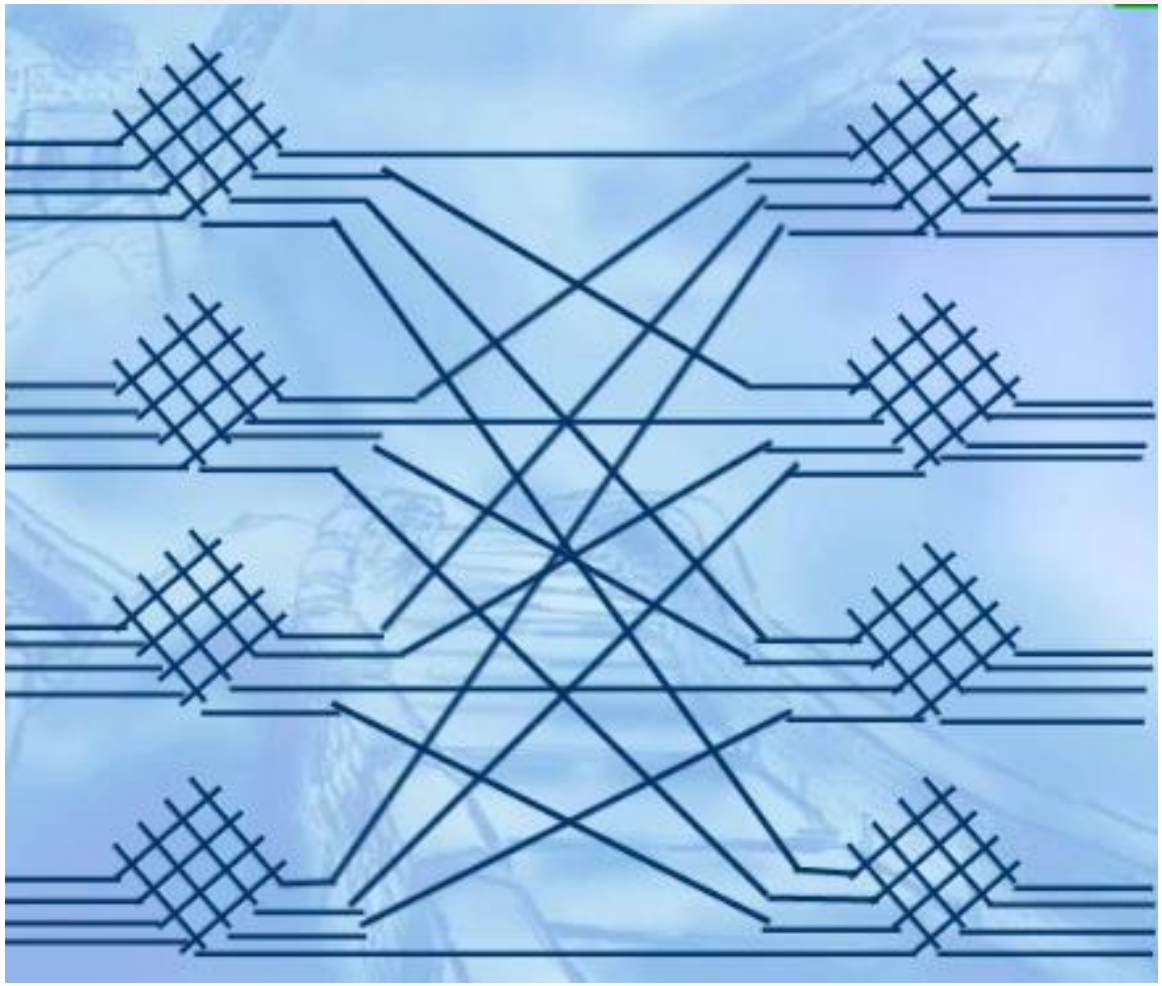
多级互连网络（空间性）

（参见第5章的互联网络）

基本的多级互联网络有：

- 多级立方体网络
- 多级混洗交换网络
- 多级PM2I网络根据拓扑结构进行分类

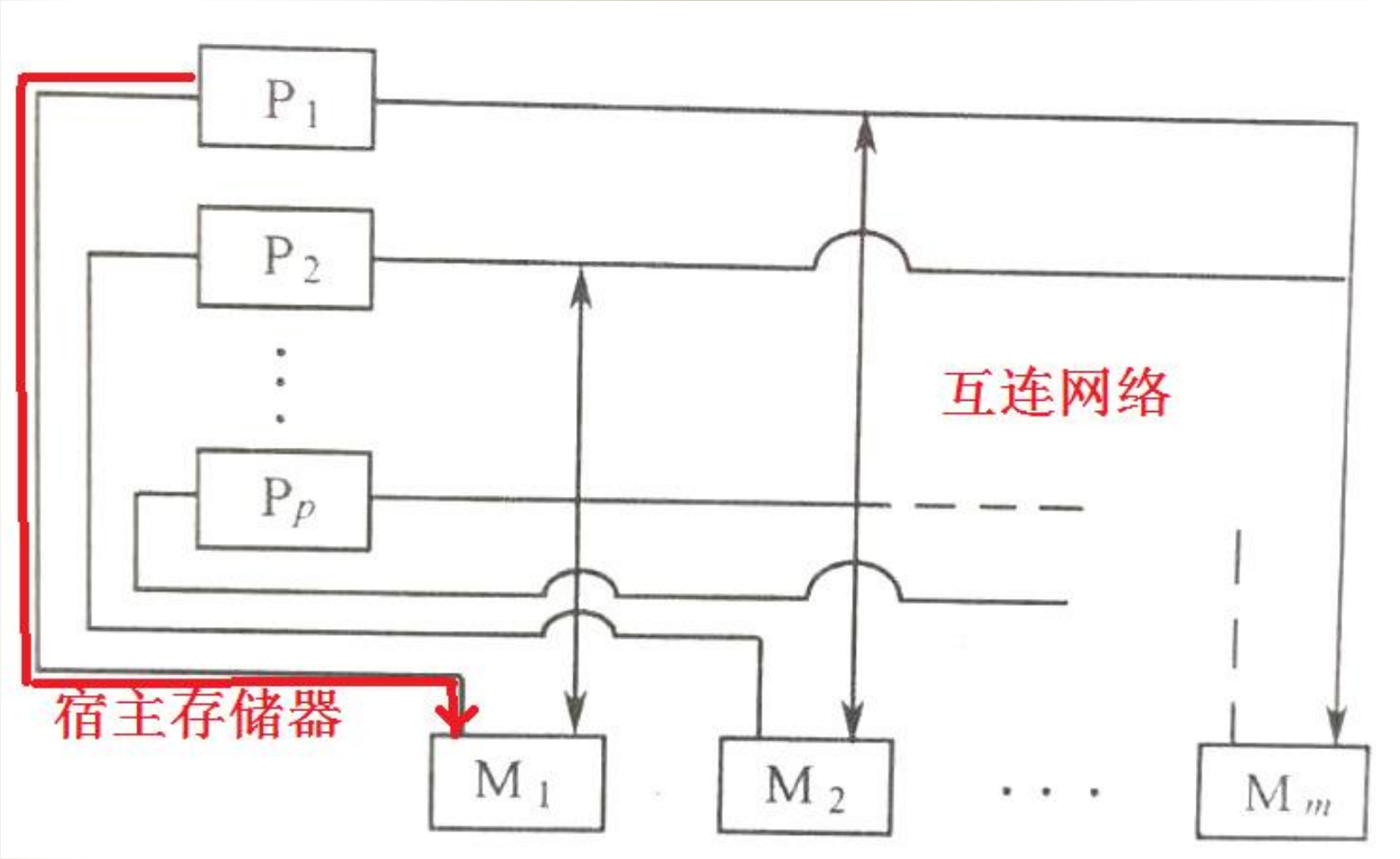
Delta网络：用a×b交叉开发模块构成的$a^n \times b^n$的交叉开关网络，其中指数n为级数。



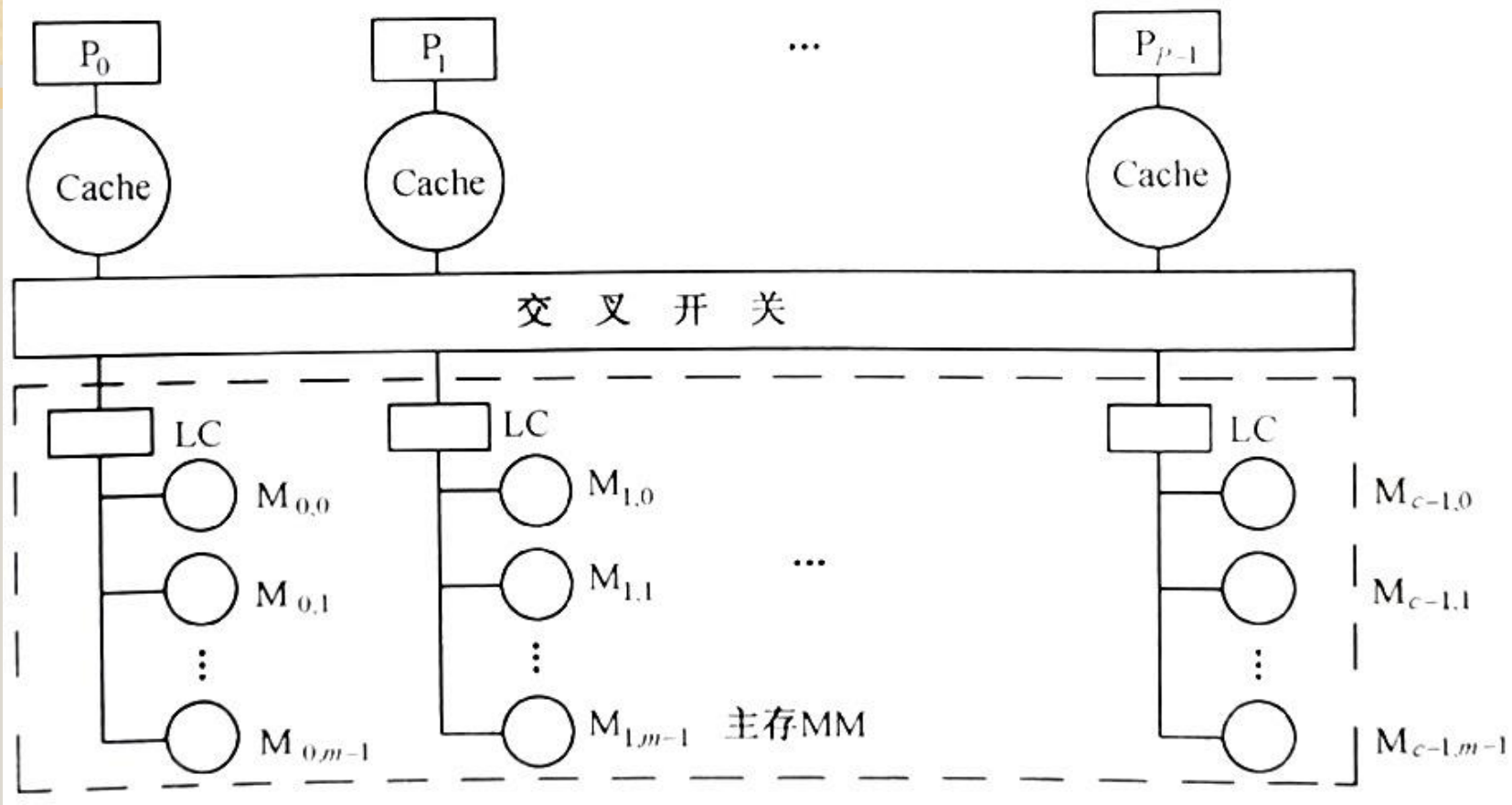
$4^2 \times 4^2$的Delta网络

3 多处理机系统的存储器结构

- 在多处理机系统中，为了减少访存冲突，主存采用并行存储器结构。
- 多个存储模块可采用低位交叉编址技术，也可采用高位交叉编址技术。
- 能为某处理机进程放置大多数页面的存储器模块称为该处理机宿主存储器。
- 如果该处理器的现行进程全部活动页面在宿主存储器内，而且该存储器不包含其他处理机的页面，则处理机不会遇到存储冲突。



宿主存储器结构



二维存储器结构

- 各列的存储模块之间按高位交叉编址，而列内各模块按低位交叉编址，每列有一个列控制器连到互联网络。

多处理机系统的特点

① 结构灵活性
- 结构灵活，实现各种复杂的机间互联模式。

② 程序并行性
- 多任务之间，挖掘各种潜在的并行性。

③ 并行任务派生
- 多处理机是多指令流操作方式

④ 进程同步

⑤ 资源分配和任务调度
- 多处理机执行并发任务

4 多处理机的软件

- 4.1 算术表达式的并行算法
- 4.2 程序并行性分析
- 4.3 并行程序语言

4.1算术表达式的并行算法

- 并行性的开发在于算法。
- 顺序处理机采用的循环及迭代算法不适用于多处理机。
- 采用直解法，揭示更多的并行性。
- 例如多项式

$$E1=a+bx+cx^2+dx^3$$

$$E1= a+x\{b+x[c+x(d)]\}$$

- 哪种表达具有更好的并行性？

4.2 程序并行性分析

一、并行性开发

多个程序段，如果不相关，则可并行。

如果存在，则影响执行次序：

- 数据相关

  Pi　A=B+D

  Pj　C=A*E

- 数据反相关

  Pi　C=A*E

  Pj　A=B+D

- 数据输出相关

  Pi　A=B+D

  Pj　A=A+C

注：三种相关情况之外，则是Pi和Pj数据不相关。

这三种数据相关情况对程序并行性的影响将转化为下列几种可能的执行次序。

（1）写-读串行次序

P1和P2服从交换律时，虽仍需串行执行，但允许P1和P2执行的次序对换，这称为可交换串行

P1　A=2*A

P2　A=3*A

为可交换串行；但

P1　A=B+1

P2　A=A+1

就是不可交换串行的。

（2）读-写次序

如果二程序段之间只包含第二种数据相关情形：

- 只须保持在先的语句先读，在后的语句后写的次序，
- 既允许串行，也允许并行执行，不允许交换次序。如：

P1　A=B+D

P2　B=C+E

（3）可并行次序。

如果程序的两段之间不存在任一种数据相关情况，即无共同变量

P1　A=B+C

P2　D=B+E

（4）必并行次序

如果两程序段的输入变量互为输出变量，则而这必须并行执行，而不允许串行执行。如

P1　A=B

P2　B=A

两语句的左右变量互相交换，必须并行执行，且需保持读写完全同步。

4.3 并行程序语言

源程序和目标程序中需要有专门的语句和指令，以描述并行关系。

（1）高级语言中描述并行性的语句

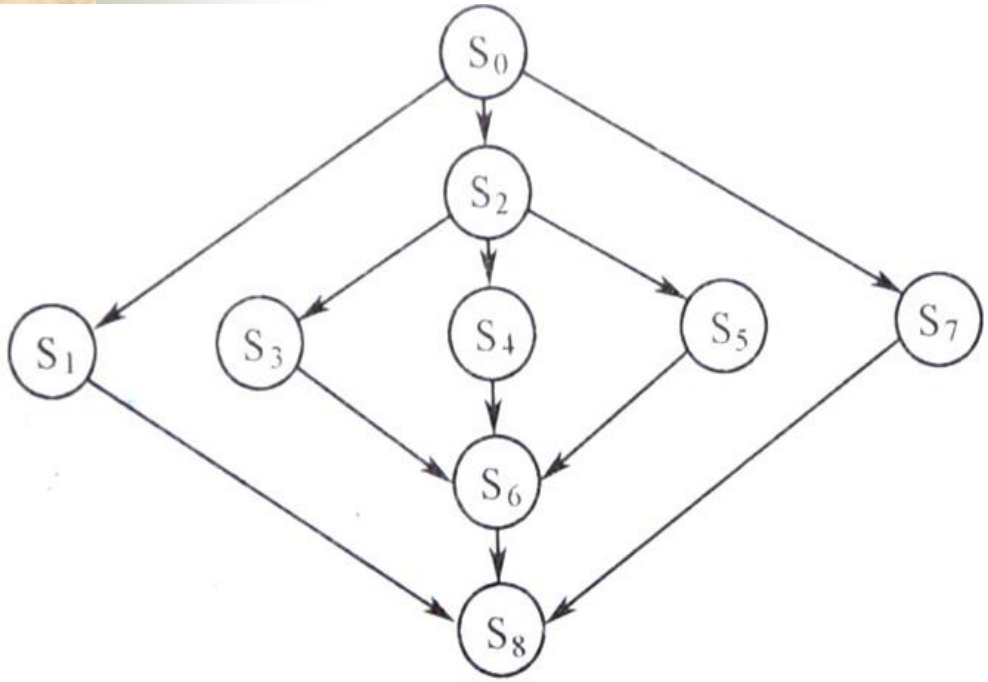
- 例如，E.W.Dijkstra的语言方案是：

Cobegin（Parbegin）

- 可并行执行的语句或进程S1S2...Sn

Coend（Parend）

并行语句也可以嵌套。例如：




```
begin
S0;
Cobegin
  S1;
  begin
      S2;
      Cobegin S3;S4;S5; Coend
      S6;
  end
  S7;
Coend
S8;
end
```

（2）并行编译

依靠并行编译程序，直接从算术表达式产生能并行执行的目标程序。

例： Z=E+A*B*C/D+F

利用普通串行编译算法，产生三元指令组为

1 *AB
2 *1C
3 /2D
4 +3E
5 +4F
6 =5Z

指令间均相关，需5级运算。

如采用并行编译算法可得

1	*AB	2	/CD
3	*12	4	+EF
5	+34		
6	=5Z		

1、2为第一级；

3、4为第二级；

5、6为第三级；

分配给两个处理机，只需三级运算。

（3）描述程序并行性的指令

并行程序的执行是一个不断地进行并行性任务的派生和汇合的过程。

在机器语言中，描述派生和汇合关系的并行控制指令常用：

FORK A ；派生出标记符A所对应的新进程

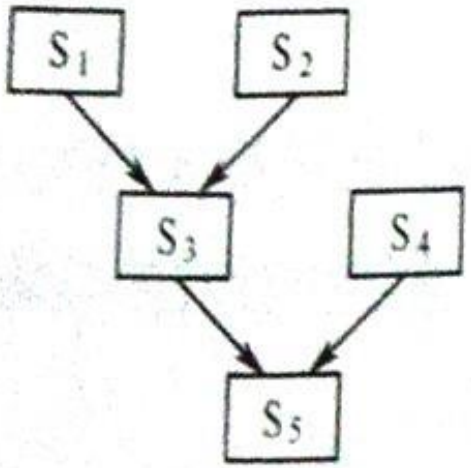
JOIN N ；JOIN指令有一个计数器，执行时会加1，并与N比较，若<N则等待

利用FORK和JOIN反映并行关系如下：

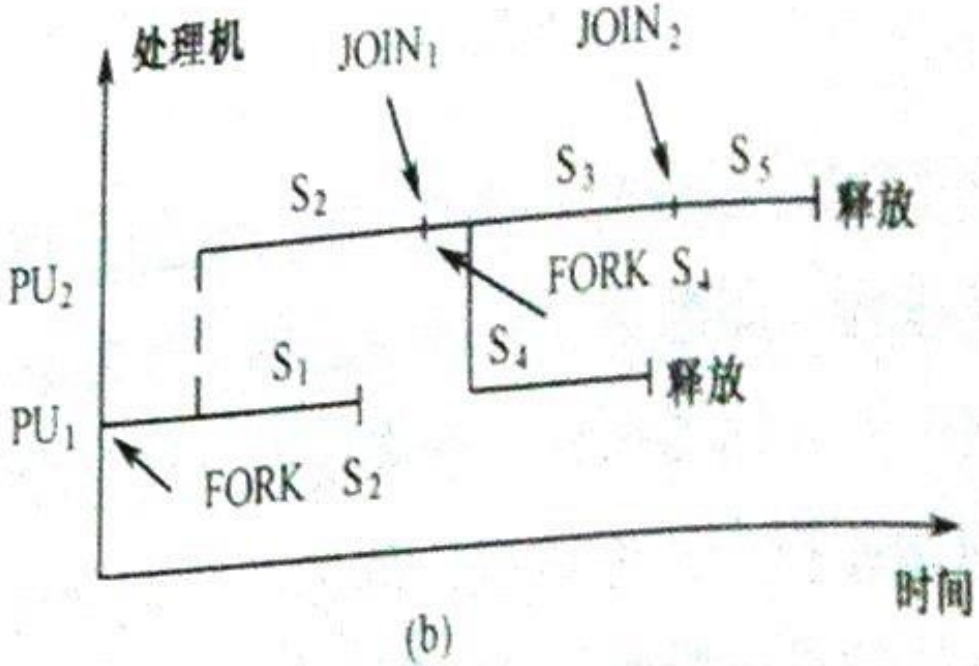
如：Z=E+A*B*C/D+F

S1	G=A*B
S2	H=C/D
S3	I=G*H
S4	J=E+F
S5	Z=I+J

```
      FORK S2
S1    G=A*B
      JOIN  2
      GOTO  S3
S2    H=C/D
      JOIN  2
S3    FORK  S4
      I=G*H
      JOIN  2
      GOTO  S5
S4    J=E+F
      JOIN  2
S5    Z=I+J
```


S1
S2
S3
S4
S5

(a)

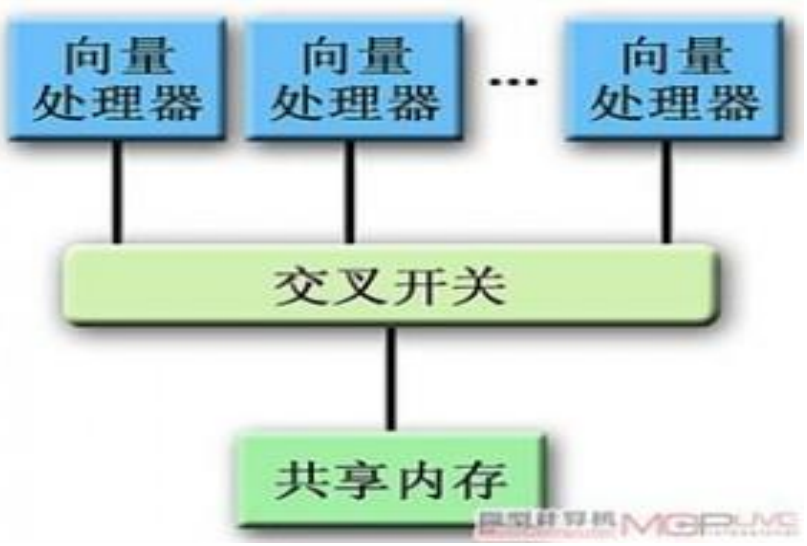

处理机
JOIN1
JOIN2
S2
S3
S5
释放
PU2
FORK S4
S4
S1
释放
PU1
FORK S2
时间

(b)

5种超级计算机并行体系结构

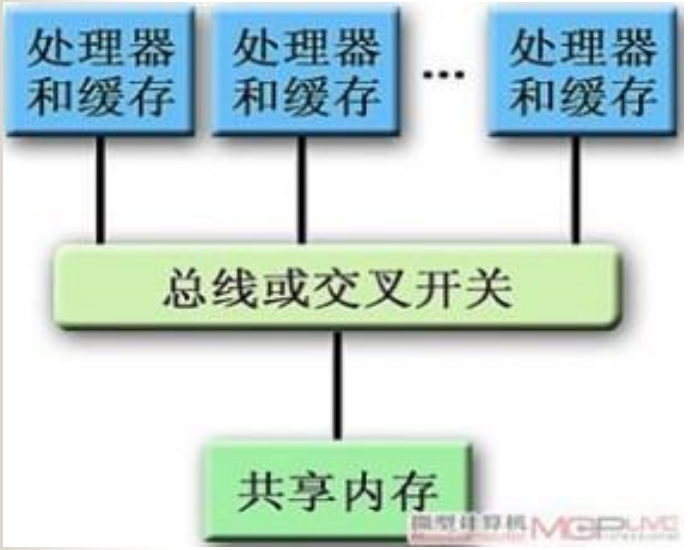
- 并行向量处理
  (Parallel Vector Processing，PVP)系统
- 对称式多处理
  (Symmetric Multi Processing，SMP)系统
- 分布式共享内存
  (DistributedShared Memory，DSM)系统
- 大规模并行处理
  (Massive Parallel Processing，MPP)系统
- 集机群式超级计算机系统(Cluster)

并行向量处理系统PVP



- 采用一定数量的、并行运行的向量处理器和共享式内存(Shared Memory，SM)结构的计算机系统。
- 代表机型有Cray XMP、Cray YNP、NEC SX2、我国的银河一号和二号等。

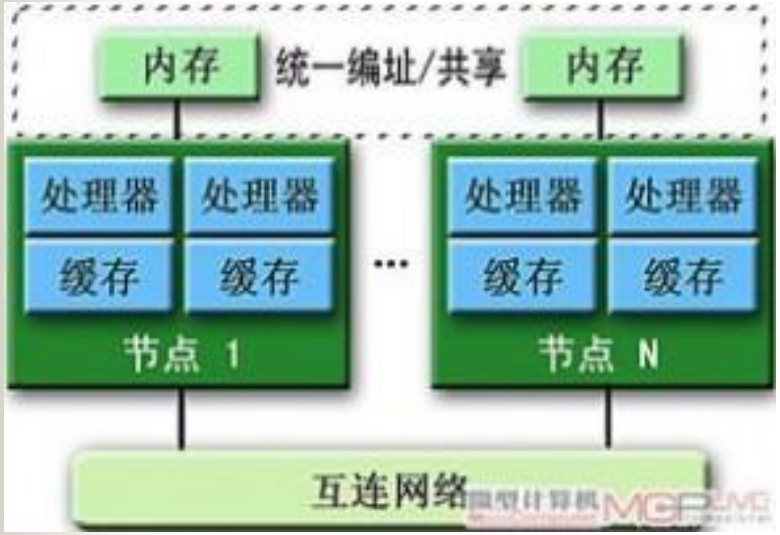
对称式多处理系统SMP



- 采用一定数量、并行运行的微处理器和共享式内存(SM)结构的计算机系统。各处理器通过系统总线或交叉开关连接共享的内存模块，可均等或对称地共享内存和其它系统资源并由同一操作系统管理，提高整个系统的数据处理能力，因此SMP属于一致性内存访问(Uniform Memory Access，UMA)方式

SMP的代表机型有IBM R50、SGI Power Challenge、Sun SPARC Center 2000、曙光一号等

分布式共享内存系统DSM



- 并行处理节点:完整的计算单元(配置有处理器和内存模块)，各节点通过高速网络互连，系统由单一操作系统管理，分布于各个节点的全部内存被统一编址，可由所有用户共享

- 与SMP不同，**DSM对内存资源的共享是非对称的**，每个节点访问本地内存与远程节点内存时的延迟和带宽是不同的
- DSM系统属于“非一致性内存访问”(Non-Uniform Memory Access，NUMA)方式
- 代表机型有SGI Origin 2000/3000、Sequent NUMA-Q、HP/Convex SPP 1600、银河三号和神威一号等。

大规模并行处理系统MPP

- 成百上千计算节点组成的并行处理计算机系统，每个计算节点配置一个或多个处理器，各个节点相对独立，有各自独立的内存模块和操作系统。
- MPP特点是可以获得很高的峰值运算速度，且由于系统的内存分布于各个节点，所以MPP属于分布式内存(Distributed Memory (DM)结构，具有易扩展性。
- 代表机型主要有IBM SP2、Intel Paragon、CRAY T3E、曙光1000等。

集群式超级计算机系统

- 代表机型有洛斯阿拉莫斯国家实验室的Avalon Cluster、ASCI Blue Mountain、深腾1800/6800和曙光2000/3000等

- 上世纪90年代中后期，随着Intel芯片等造价低廉的微型计算机组件的出现和网络技术的迅速发展，使采用普通微型机或工作站作为计算节点并采用高速网络互连的并行计算系统成为了可能，超级计算机体系结构由此开始迈入集群时代
- 从内存访问方式上看，机群系统采用了与MPP相同的分布式内存(DM)结构，因而具有很高的可扩展性。

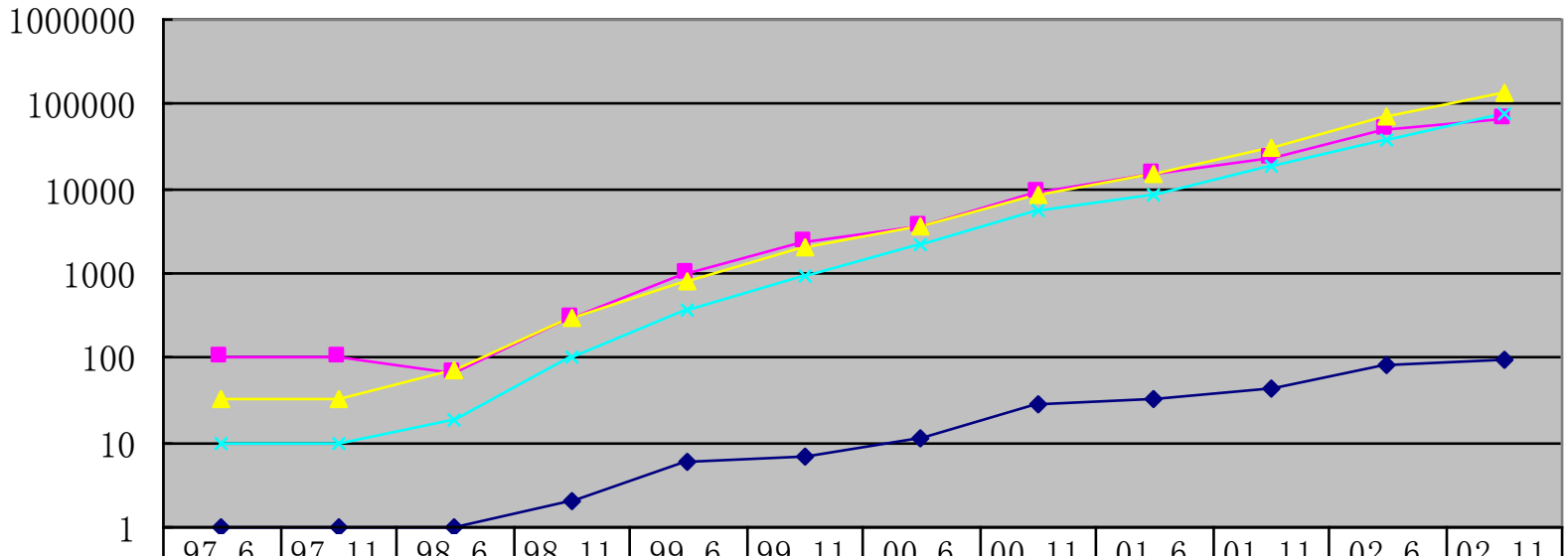
集群式计算机产生的原因

- PVP, MPP, NCC-NUMA, CC-NUMA, SMP 的共性：
- 从底层到顶层实现整个计算机， 技术难度大，硬件研制周期长
- 软件研制周期长：缺乏通用系统软件和应用软件，
- 增加研制成本，降低竞争力

集群式计算机产生的原因

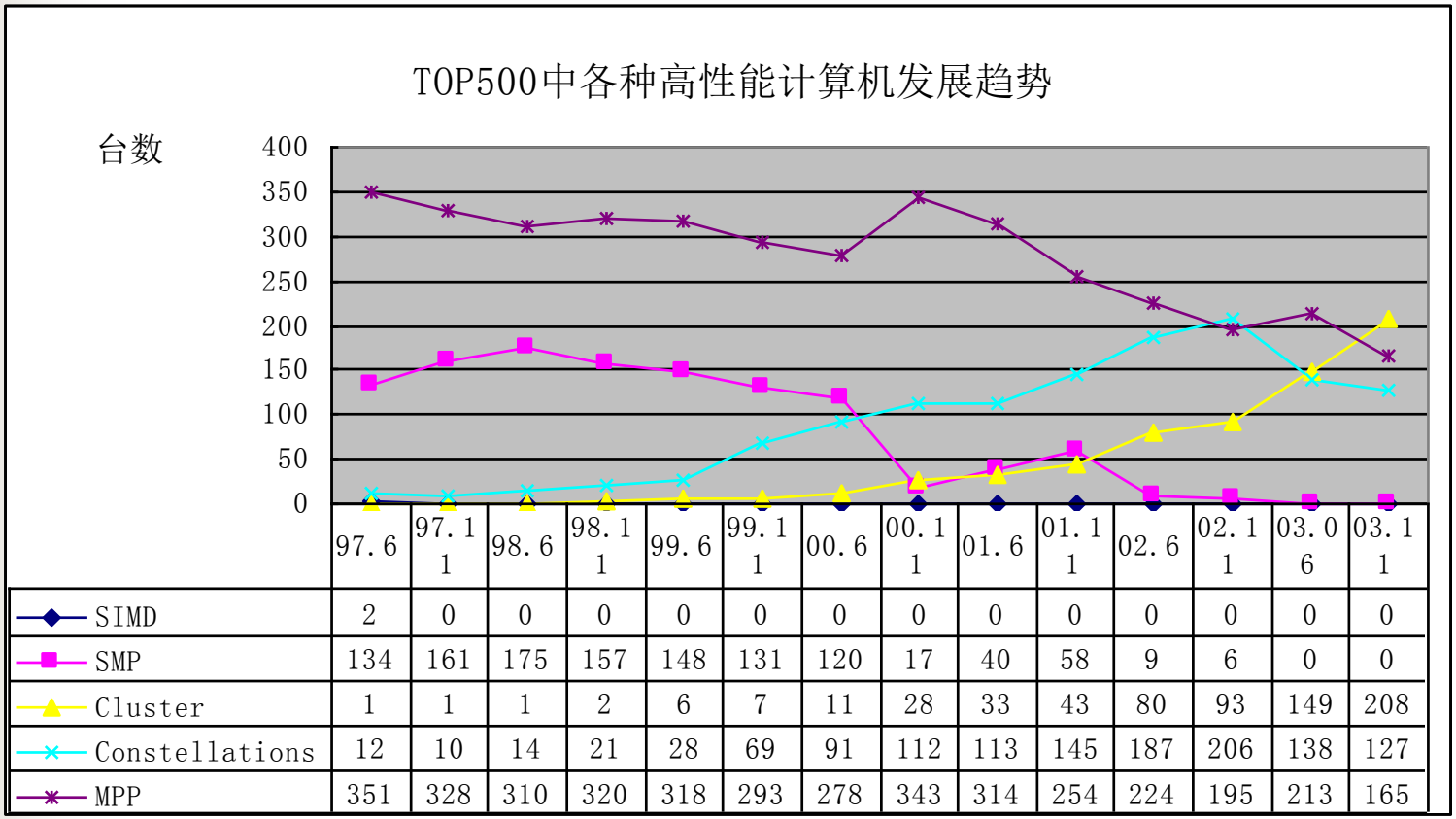
- 集群（Cluster）能以较短的研制周期、集成最新技术、汇集多台计算机的性能呢，达到较高的性能价格比
- 结构灵活、通用性强、安全性高、易于扩展、高可用性和高性价比等诸多优点

TOP500中Cluster的数量和性能



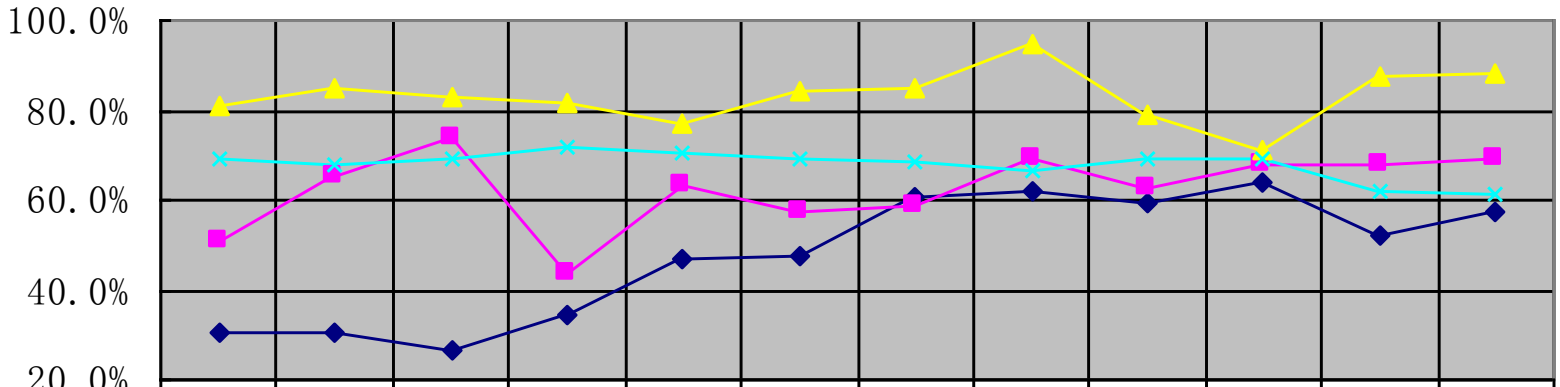
	97.6	97.11	98.6	98.11	99.6	99.11	00.6	00.11	01.6	01.11	02.6	02.11
数量	1	1	1	2	6	7	11	28	33	43	80	93
cpu总数	100	100	68	290	1026	2412	3668	9174	15373	22526	50181	66614
峰值速度[GF/S]	33	33	72	299	789	2009	3720	8717	14689	29523	69774	136048
LINPACK[GF/S]	10	10	19	102	372	961	2254	5409	8745	18831	36569	78052

TOP500的全球超级计算机500强排名中，集群式系统所占比率连年上升，现已达到83%以上。



	97.6	97.11	98.6	98.11	99.6	99.11	00.6	00.11	01.6	01.11	02.6	02.11	03.06	03.11
SIMD	2	0	0	0	0	0	0	0	0	0	0	0	0	0
SMP	134	161	175	157	148	131	120	17	40	58	9	6	0	0
Cluster	1	1	1	2	6	7	11	28	33	43	80	93	149	208
Constellations	12	10	14	21	28	69	91	112	113	145	187	206	138	127
MPP	351	328	310	320	318	293	278	343	314	254	224	195	213	165

TOP500中高性能计算机的效率



	97.6	97.11	98.6	98.11	99.6	99.11	00.6	00.11	01.6	01.11	02.6	02.11
Cluster	30.3%	30.3%	26.4%	34.1%	47.1%	47.8%	60.6%	62.1%	59.5%	63.8%	52.4%	57.4%
Constellations	50.5%	65.3%	73.8%	43.9%	63.3%	57.4%	58.5%	69.1%	62.7%	67.8%	68.2%	69.2%
SMP	81.1%	84.7%	83.2%	81.4%	77.1%	84.1%	84.9%	94.5%	79.3%	71.0%	87.4%	88.2%
MPP	69.2%	68.1%	69.5%	71.6%	70.3%	69.3%	68.2%	66.2%	69.1%	69.0%	62.1%	61.5%

集群的定义

- 集群系统是利用高速通信网络将一组计算节点连接起来；
- 在并行程序设计和集成开发环境支撑下统一调度、协调处理；
- 集群系统中的主机和网络可以同构或异构；
- 利用消息传递方式实现机间的通信
- 建立在操作系统上的并行编程环境完成系统的资源管理及相互协作

集群系统的基本组成

- 节点：微型机、工作站或SMP并行机
- 互联技术：高带宽的以太网、异步传输模式ATM、InfiniBand互联技术
- 软件技术：
  - 节点操作系统
  - 中间件实现单一系统影像：单机式的管理控制，单一的地址空间和单一的文件系统：
  - 并行编程环境

集群系统的基本组成


并行应用
串行应用
并行编程环境
用户接口
作业管理等
SSI、HA等机群系统中间件
操作系统
操作系统
操作系统
节点 1 计算机
节点 2 计算机
...
节点 N 计算机
网络接口
网络接口
网络接口
高速互连网络

性能指标

- 峰值速度：计算得出的理论值
  - 理论峰值速度（亿次）=节点机每个CPU主频(MHz)*CPU每个时钟周期执行浮点运算的次数*CPU总数目/10^8
- 实测速度：国际通用的超级计算机基准评测软件Linpack，采用求解 线性方程组合特征值问题的方法综合评价浮点运算性能
- 运行效率
  - 实测速度与峰值速度的比率

本章重点

- 多处理机与并行处理机的区别
- 多处理机结构
- 并行编程