



高性能计算技术

第三讲 并行计算机体系结构（2）

kjhe@scut.edu.cn

华南理工大学计算机学院

复习

- 一个网络规模（节点数）为 N 的三维环绕（**3D Torus**）静态网络，其对剖宽度是多少？
- 网络规模（节点数）相同的超立方和二维环绕（**2D Torus**）相比，当网络数据交换较多时，用哪种网络比较好？为什么？
- 采用 $k \times k$ 开关单元的 $N \times N$ 的多级互联网络与交叉开关相比在硬件复杂度降低了多少量级？
- 列举几个标准互联网络。

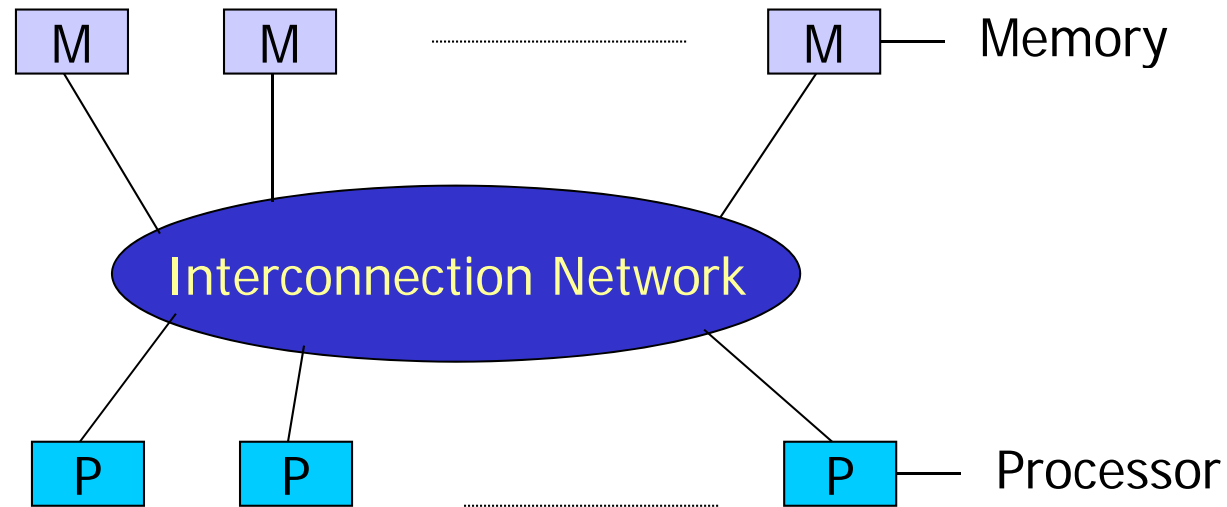
主要内容

- 并行计算机访存模型
- 并行计算机存储组织
- 并行计算机系统

MIMD结构

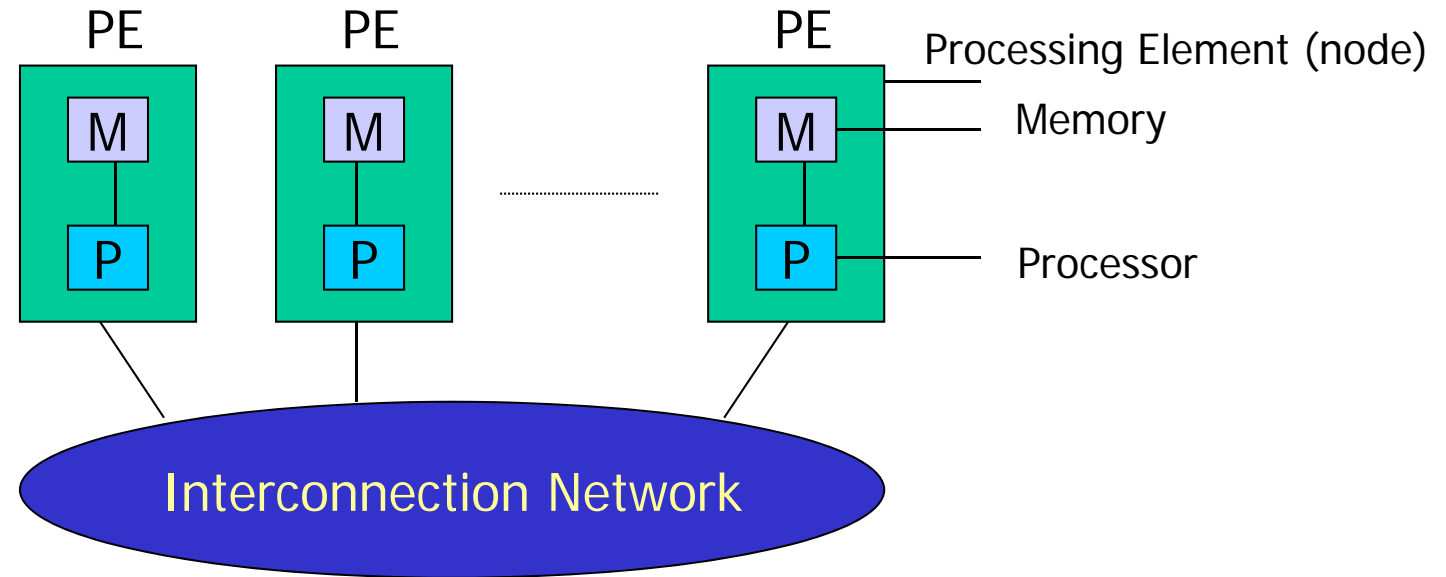
- 共享存储（Shared Memory） MIMD / Multiprocessor
- 分布式存储（Distributed Memory） MIMD / 消息传递（Message Passing） MIMD / Multicomputer

共享存储



- 单一地址空间（Single address space）：存储模块定义了一个可在处理器间共享的单一地址空间
- 任何处理器可以通过互联网存取任何存储模块
- 例子： SGI Origin, Sun E10000

分布式存储

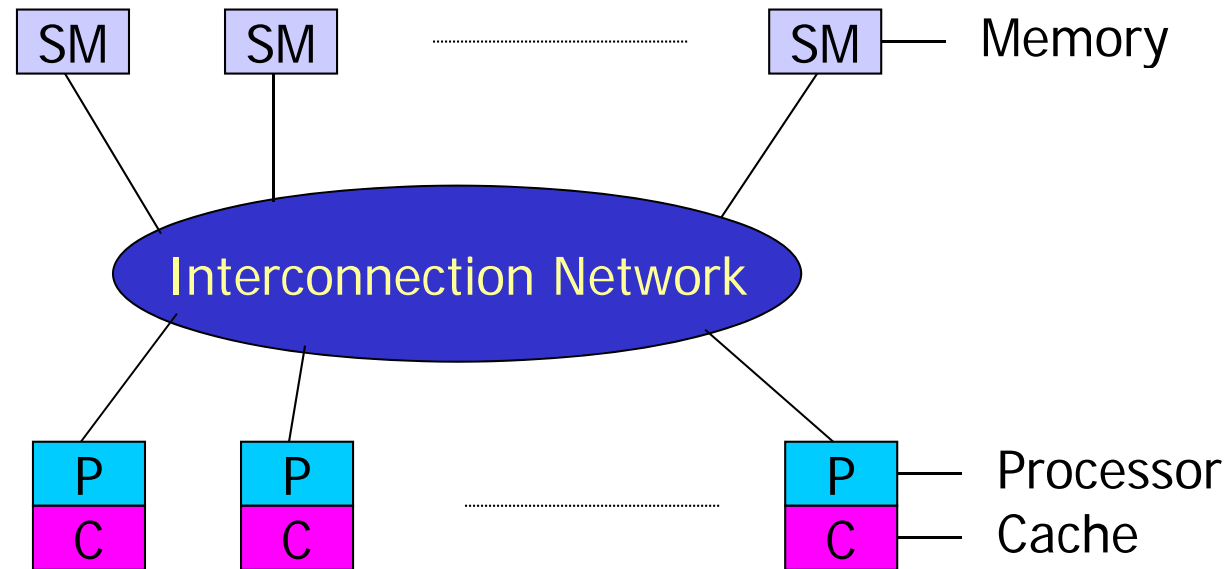


- 处理器单元（PE）独立工作，每个处理器有自己本地存储
- 通过消息传递（message passing）来交互。PE不能直接存取其他PE的内存，必须通过消息传递来交换处理器之间的数据
- 例子： CRAY T3E, IBM SP, 集群（Cluster）

存储器结构分类

- 集中式存储器
 - UMA (Uniform Memory Access)
- 分布式存储器
 - NUMA (Non-Uniform Memory Access)
 - NCC-NUMA (Non-Cache Coherent NUMA)
 - COMA (Cache Only Memory Architecture)
 - CC-NUMA (Cache Coherent NUMA)
 - NORMA (No-Remote memory Access)

UMA



- UMA (Uniform Memory Access)
- 例子: Pentium Pro Quad, Sun Starfire

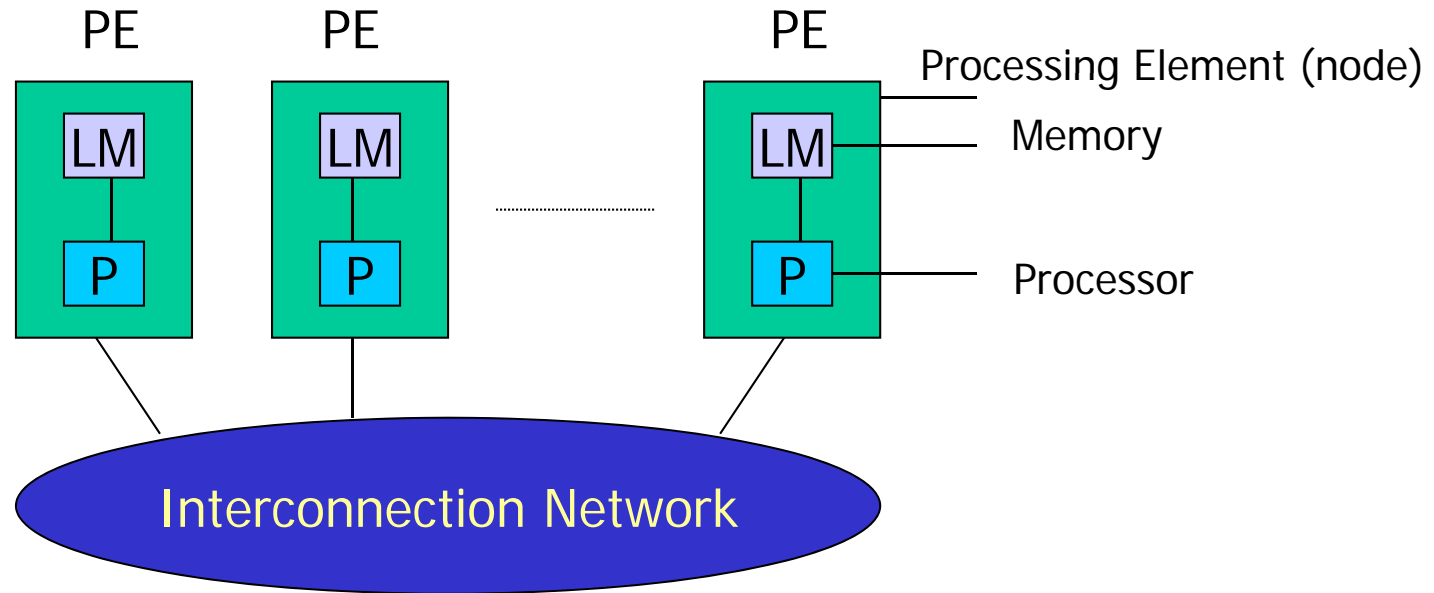
UMA

- UMA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

可扩展的共享存储

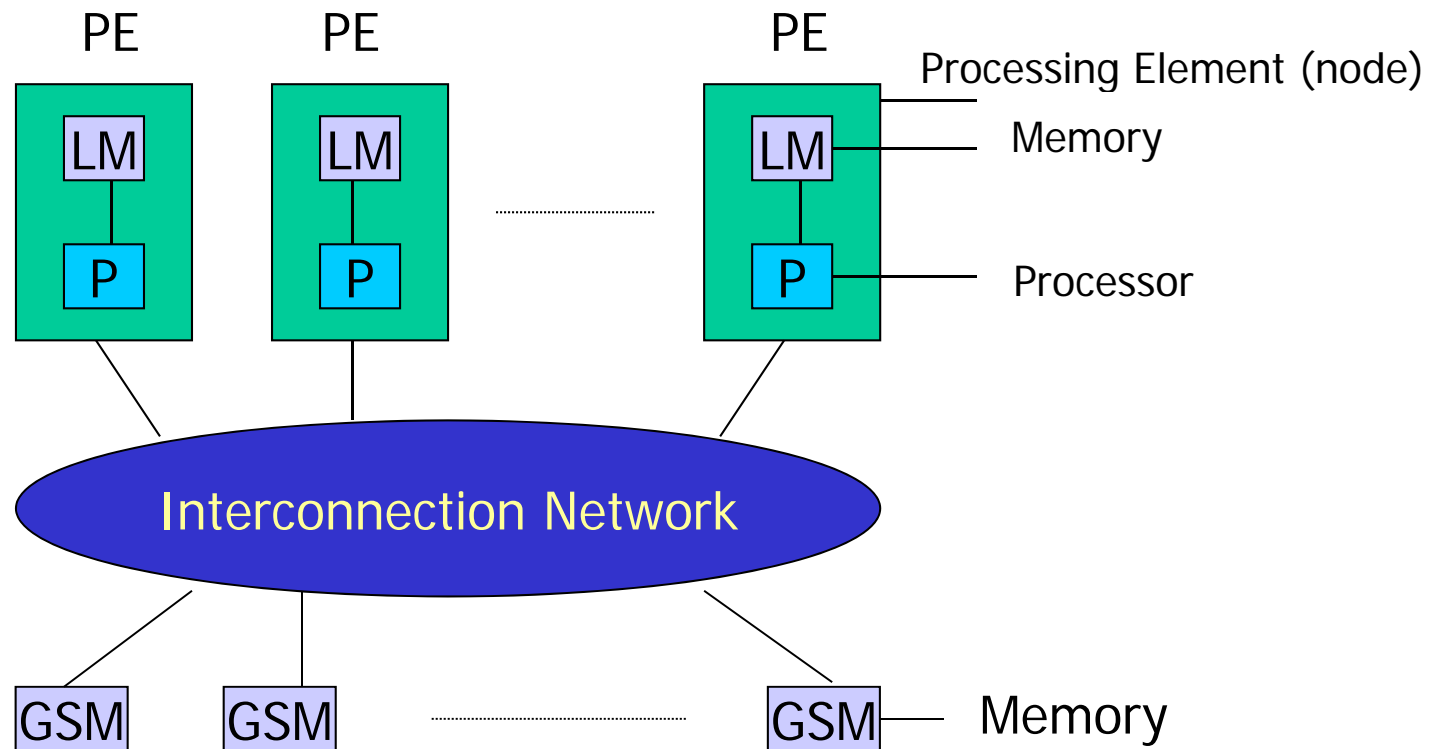
- 通过高吞吐量（High-throughput），低时延（low-latency）的互联网络
- 每个节点有一个本地缓存或存储
 - 存储缓存一致性问题（Cache coherence problem）
- 逻辑共享的存储可以通过一系列本地存储来实现
 - 分布式共享存储MIMD: NUMA

分布式共享存储（DSM）



- **分布式共享存储（Distributed Shared Memory）**和分布式存储（**Distributed Memory**）：物理结构是一样的
- 分布式共享存储：本地存储是全局地址空间的组成部分，任何处理器可直接存取其他处理器的本地存储
- 分布式存储：本地存储有独立的地址空间，不能直接存取远程处理器的存储空间

NUMA

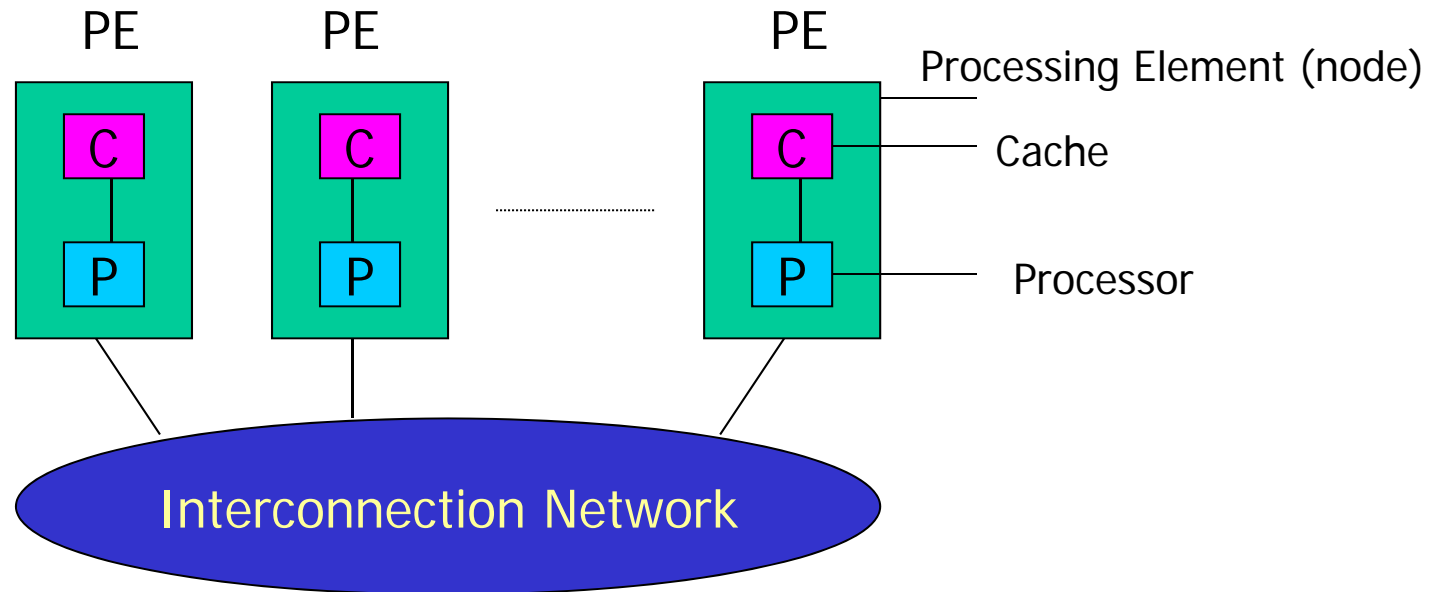


- NUMA (Non-Uniform Memory Access)
- 例子: Cray T3E, SGI Origin

NUMA

- NUMA（NCC-NUMA）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）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）
 - 每台处理器可带私有高速缓存，外设也可以某种形式共享

COMA

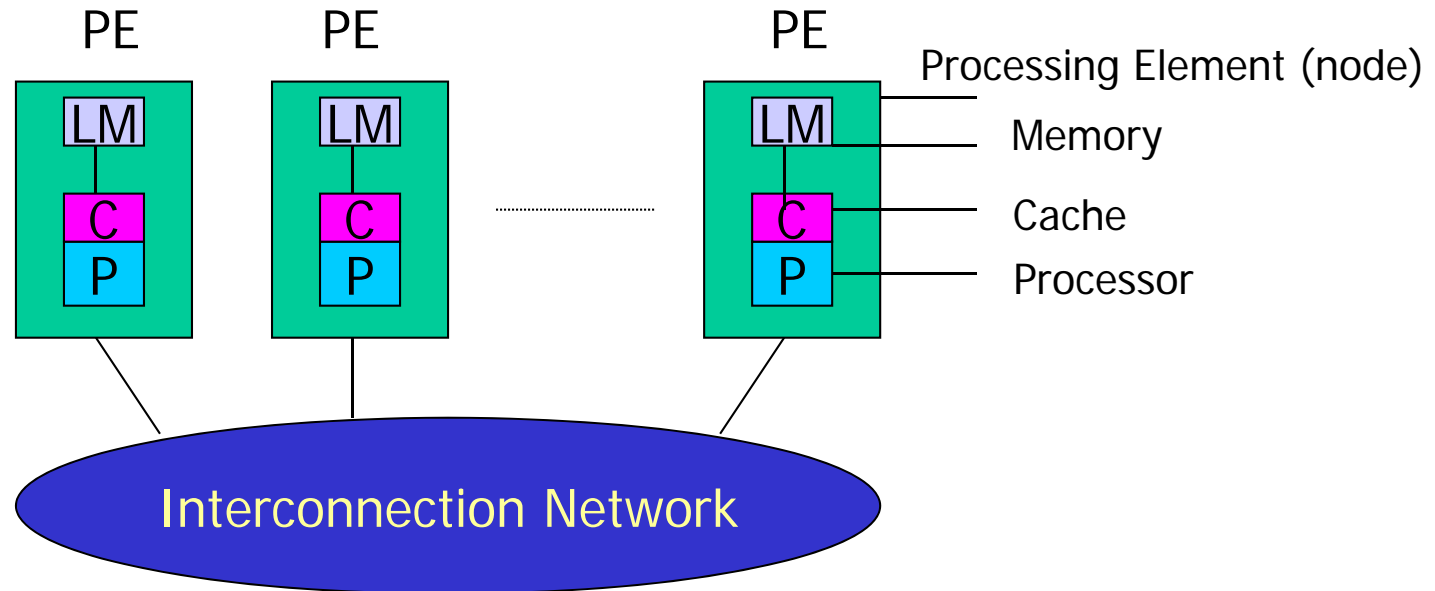


- COMA (Cache Only Memory Architecture)
- 例子： Kendall Square. Research公司的KSR - 1

COMA

- COMA模型是高速缓存存储访问模型的简称，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COMA中没有物理地址，数据可动态迁移
 - 经过“预热”，数据将被“吸引”到处理节点附近

CC - NUMA

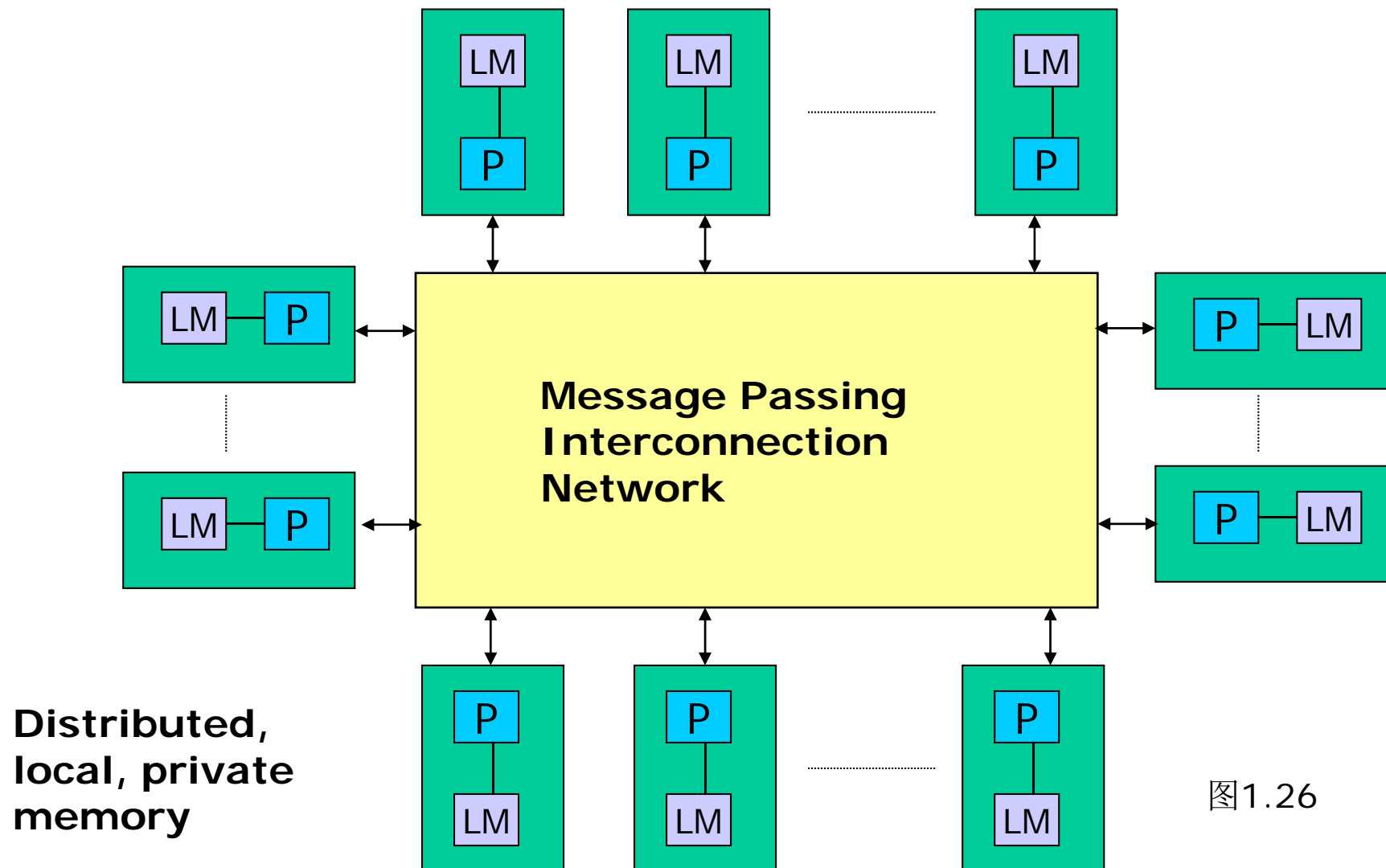


- CC-NUMA (Cache Coherent Non Uniform Memory Access)
- 例子: SGI Origin, Stanford DASH, Sequent NUMA-Q

CC-NUMA

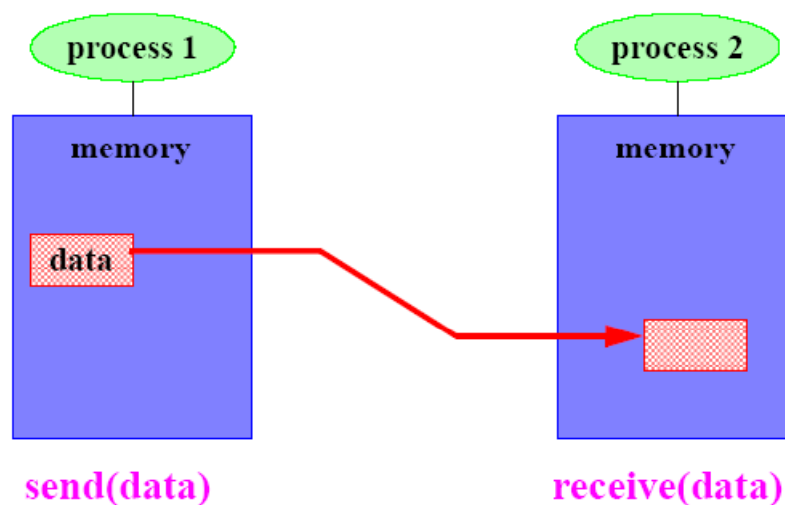
- CC-NUMA模型是高速缓存一致性非均匀存储访问模型的简称，其特点是：
 - 大多数使用基于目录的高速缓存一致性协议；
 - 保留SMP结构易于编程的优点，也改善常规SMP的可扩放性；
 - CC-NUMA实际上是一个分布共享存储的DSM多处理机系统；
 - 对高速缓存一致性提供硬件支持
 - 它最显著的优点是程序员无需明确地在节点上分配数据，系统的硬件和软件开始时自动在各节点分配数据，在运行期间，高速缓存一致性硬件会自动地将数据迁移至要用到它的地方。

NORMA(No-Remote memory Access)



NORMA

- NORMA模型是**非远程存储访问**模型的简称，其特点是：
 - 所有存储器是私有的
 - 节点不能访问远程存储器，而必须通过**消息传递**方式



构筑并行机系统的不同存储结构

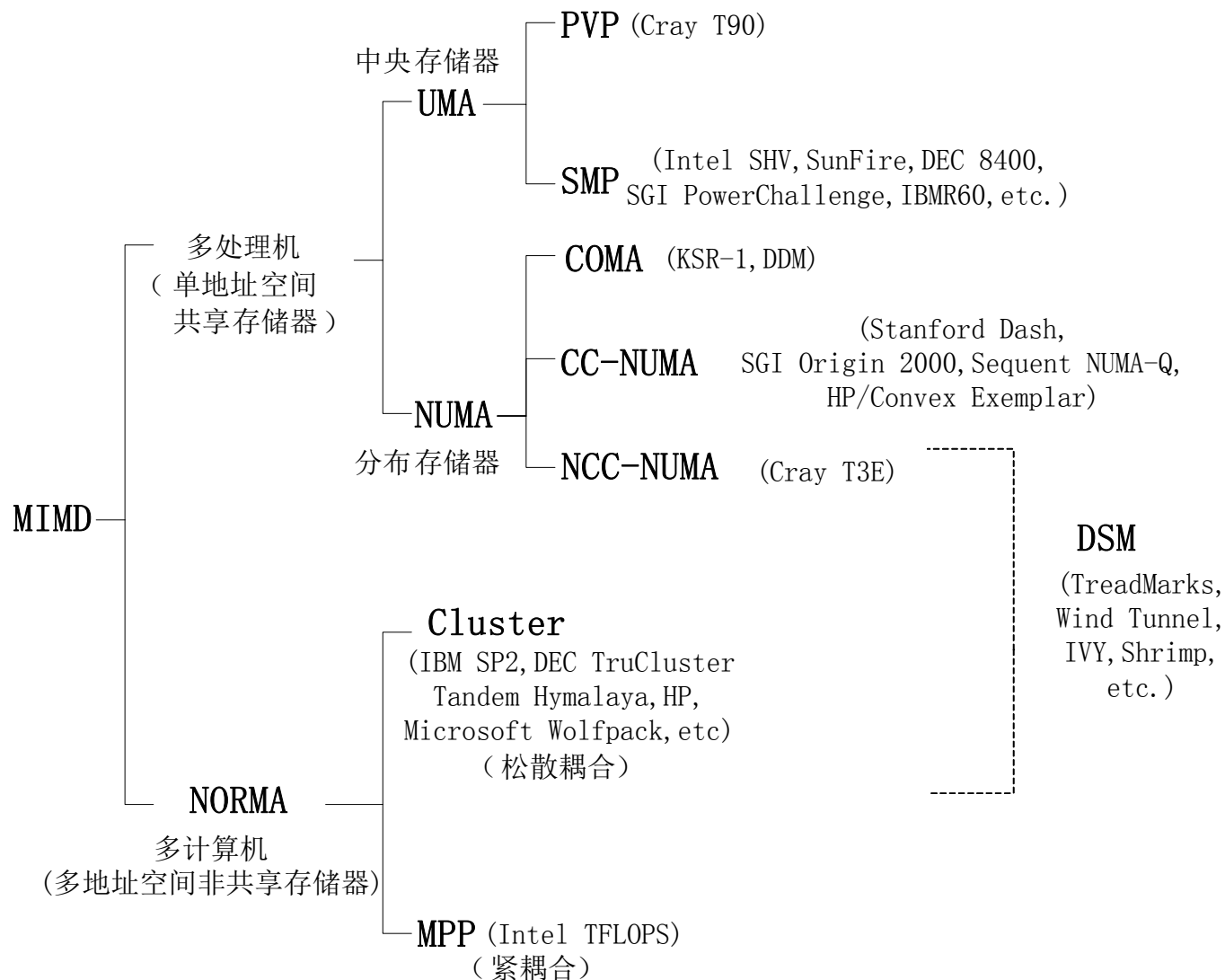


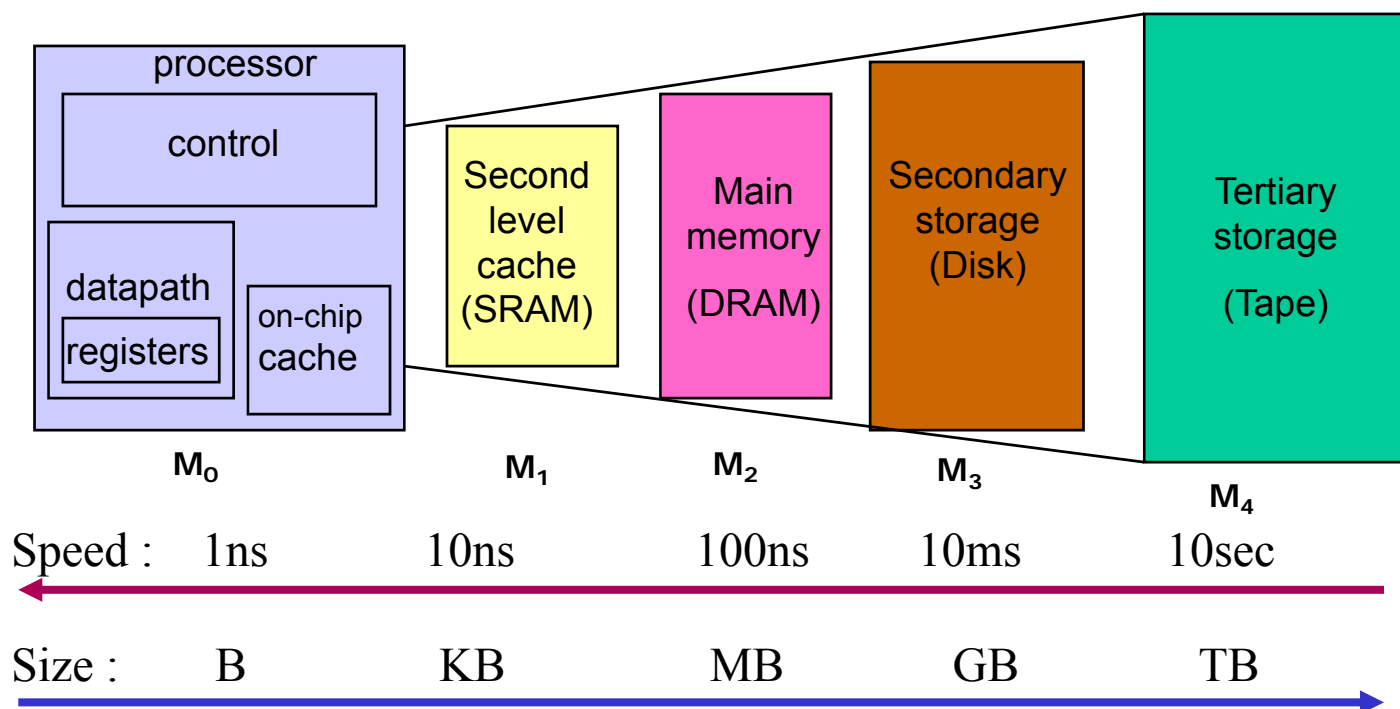
图1.27

主要内容

- 并行计算机访存模型
- 并行计算机存储组织
- 并行计算机系统

存储层次结构

- 弥补**CPU**与主存间的速度差异
- 各个层次间的访问速度和容量差别



- 应用程序如何利用存储层次结构？

存储层次结构的参数

- 参数:

- 容量 (Capacity) : C
- 时延 (Delay) : L
- 带宽 (Bandwidth) : B

远程存储 (**Remote memory**) : 通过互联网访问

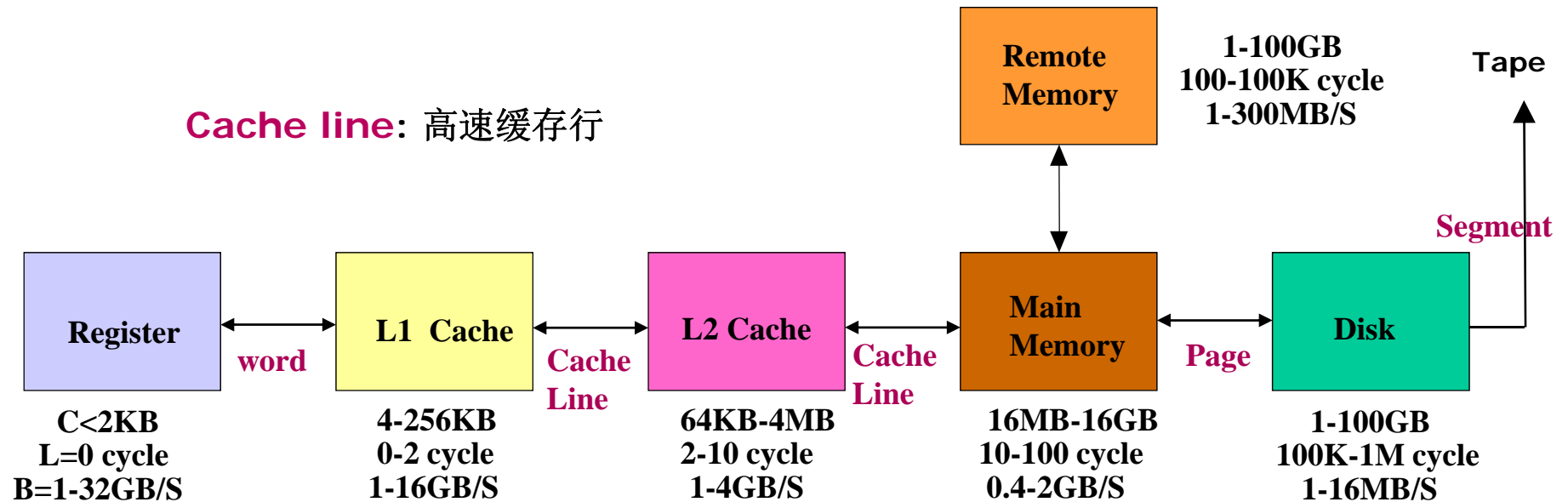
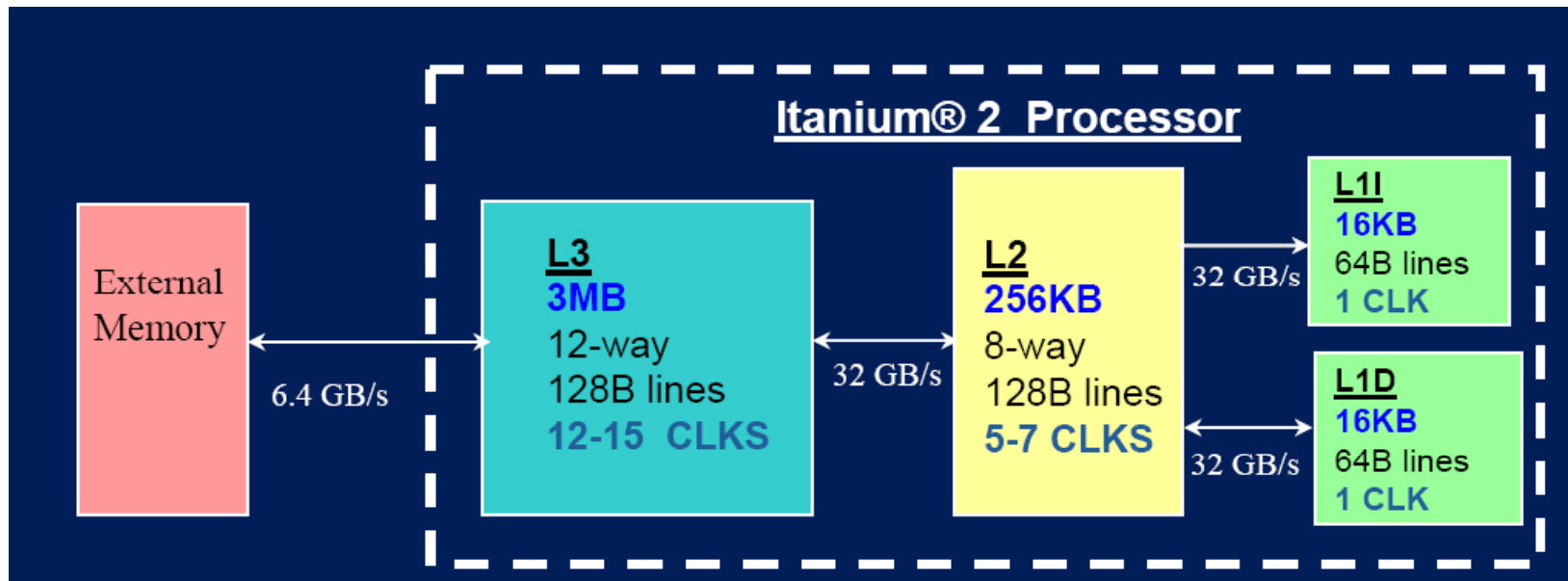


图3.1

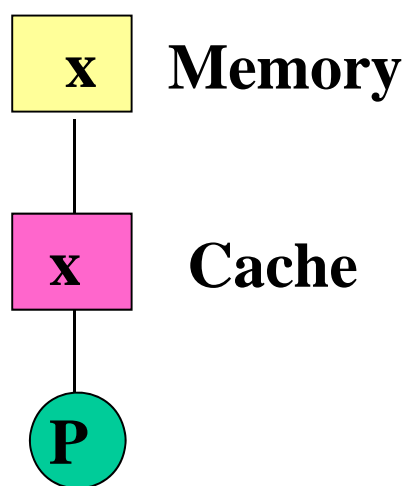
例子：Intel安腾2（Itanium2）

- McKinley核心Itanium 2处理器主频：1GHZ、900MHZ
- 32KB L1缓存，256KB L2缓存和3MB L3缓存，可以提供高达6.4GB/s的数据传输带宽

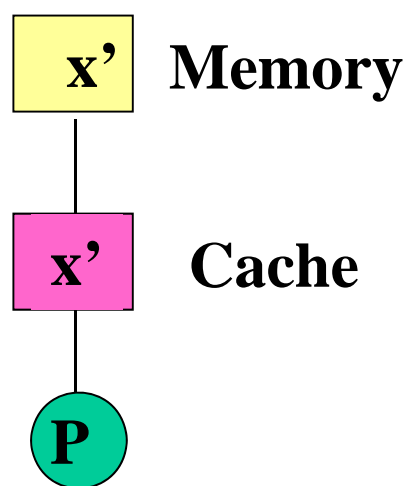


L1D: Data Cache L1I: Instruction Cache

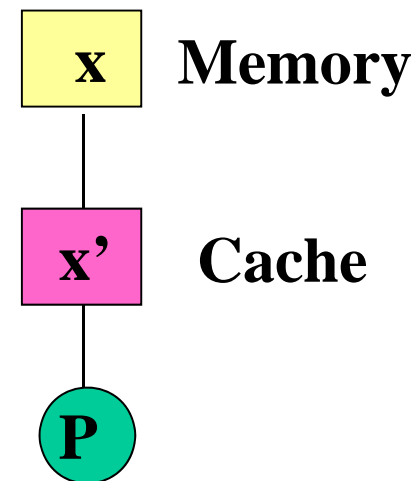
高速缓存一致性 (Cache Coherence)



之前



写直达WT
(Write Through)




写回WB
(Write Back)

商业系统通常采用write-back.

缓存不一致的问题

Time	Event	Cache for A	Cache for B	Memory for x
0				1
1	CPU A reads X	1		1
2	CPU B reads X	1	1	1
3	CPU A writes 0 to X	0	1	0

invalid value

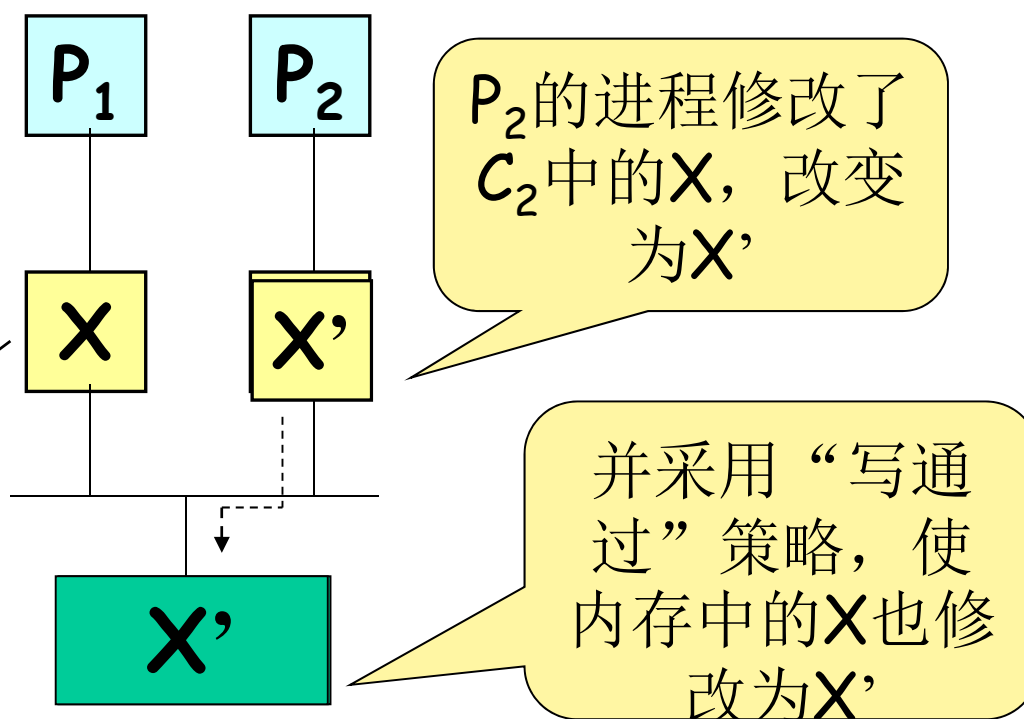
- 出现**Cache**不一致的原因主要有以下三个方面
 - 共享可写数据
 - 多处理机的进程迁移（图1.38）
 - 绕过**Cache**的I/O操作（图1.39）

进程迁移引起的不一致性

情况一WT: P_1 的 C_1 和 P_2 的 C_2 中都有共享数据 X 的拷贝

由于某种原因该
进程迁移到 P_1 上

此时 P_1 的 C_1 中仍然是 X ，而不是它先
修改过的 X' 。



- P_1 P_2 指的是处理器
- C_1 C_2 指的是 P_1 P_2 对应的高速缓冲存储器

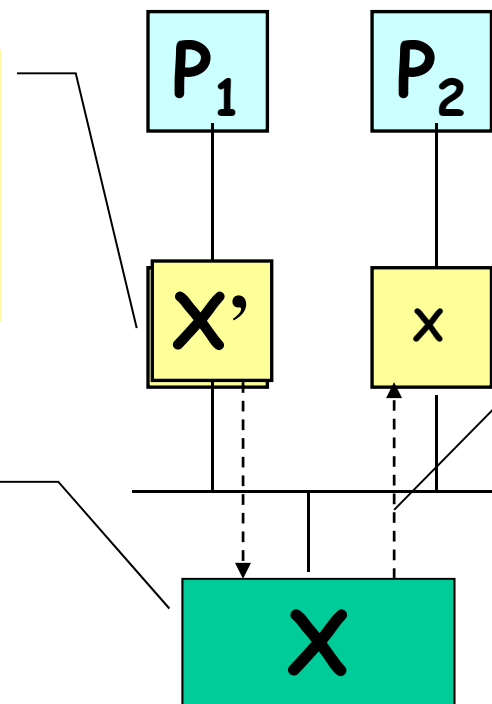
进程迁移引起的不一致性

情况二WB:

P_1 的 C_1 中有共享数据 X 的拷贝，而
 P_2 的 C_2 中没有该共享数据

若 P_1 的进程对 X
进行了修改，
使之变为 X'

采用“写回”策略，
暂时没有对内存中的
 X 进行修改。



由于某种原因，该进程
迁移到了 P_2 上运行

P_2 上的该进程
运行时将从内存中读
取 X 并将 X 调入
 C_2

那么，这个迁移了的进程此时读取的是
 X ，而不是它先前修改过的 X' 。

解决缓存不一致的方法

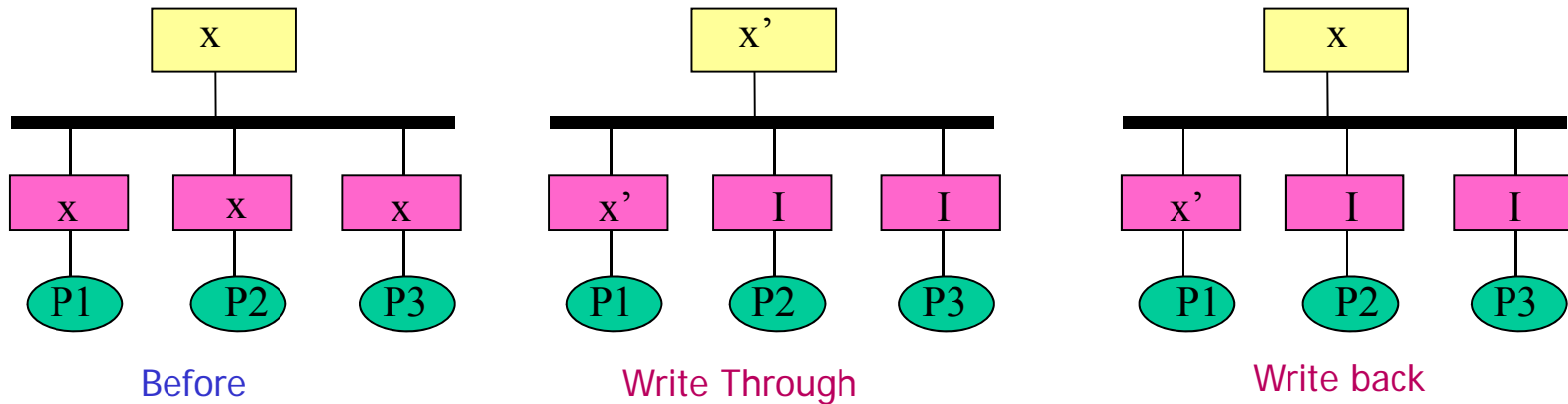
- 总线侦听（Bus snooping）
 - 所有处理器监听共享总线获知写操作，以修改本地缓存
- 基于目录（Directory-based）的协议
 - 在主存中设置一个目录表，记录所有高速缓存的位置和状态
 - 发送消息使得远程缓存无效或者被更新

监听协议

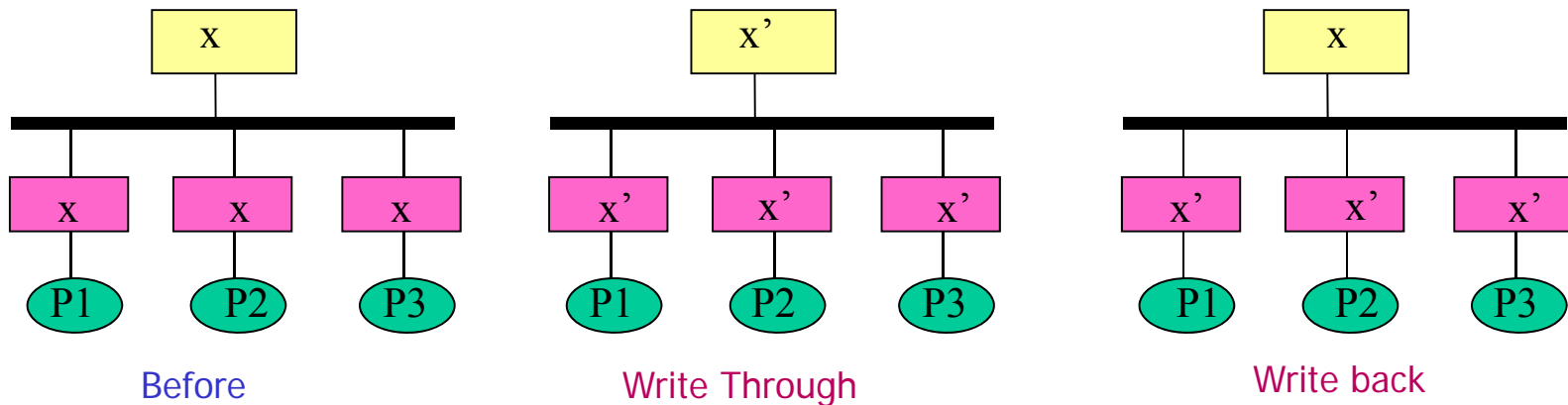
- 基于总线的方法。缓存控制器（cache controllers）监听总线的操作，以更新或无效缓存块。
- 两种更新策略
 - 写无效（Write invalidate）
 - 写更新（Write-update）
- 商业系统通常采用 write-invalidate
 - 来节省带宽

两种类型的写协议

- 写无效（Write-invalidate）

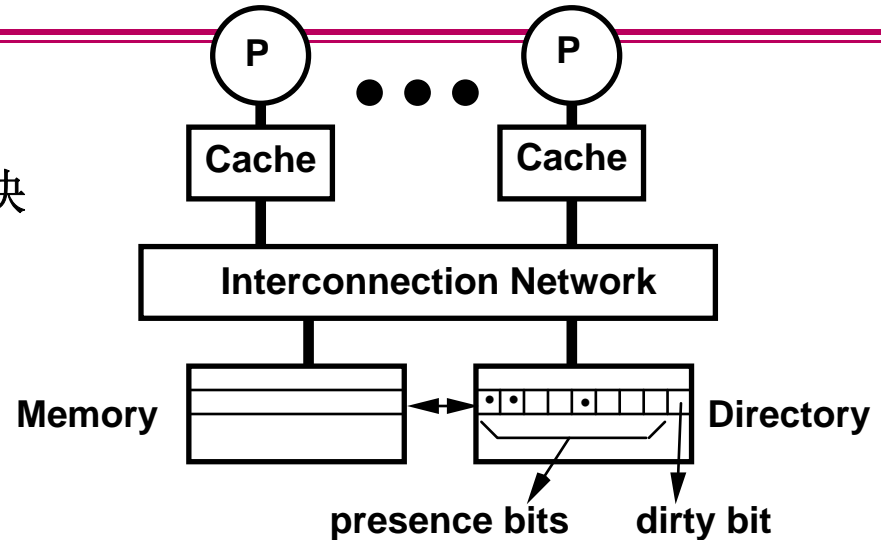


- 写更新（Write-update）



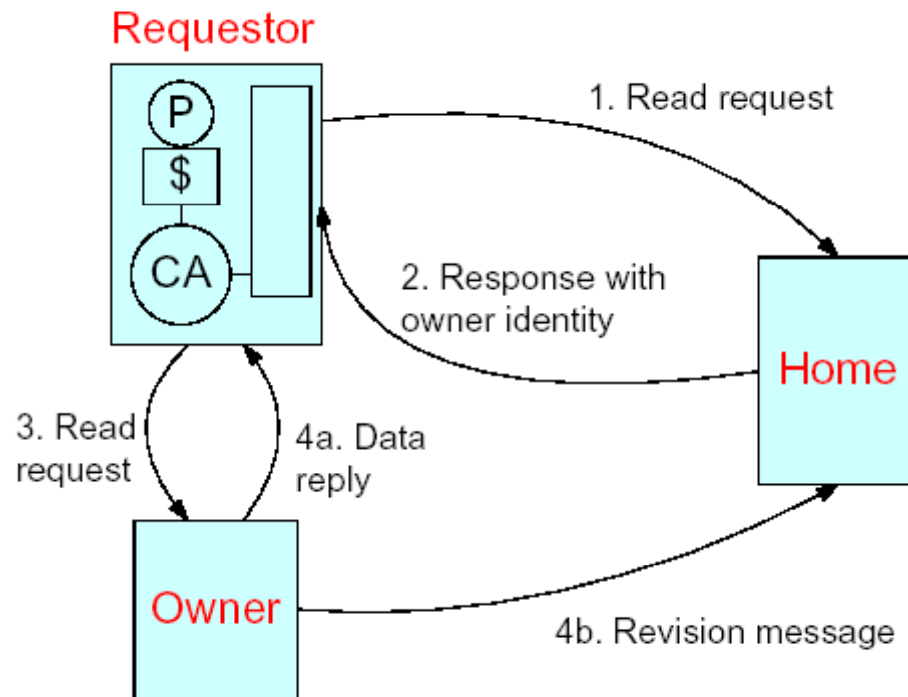
基于目录的协议

- **k**个处理器
- 在主存中维护一个目录表，对每个缓存块，有**k**个**presence-bits**（是否存在），1个**dirty-bit**

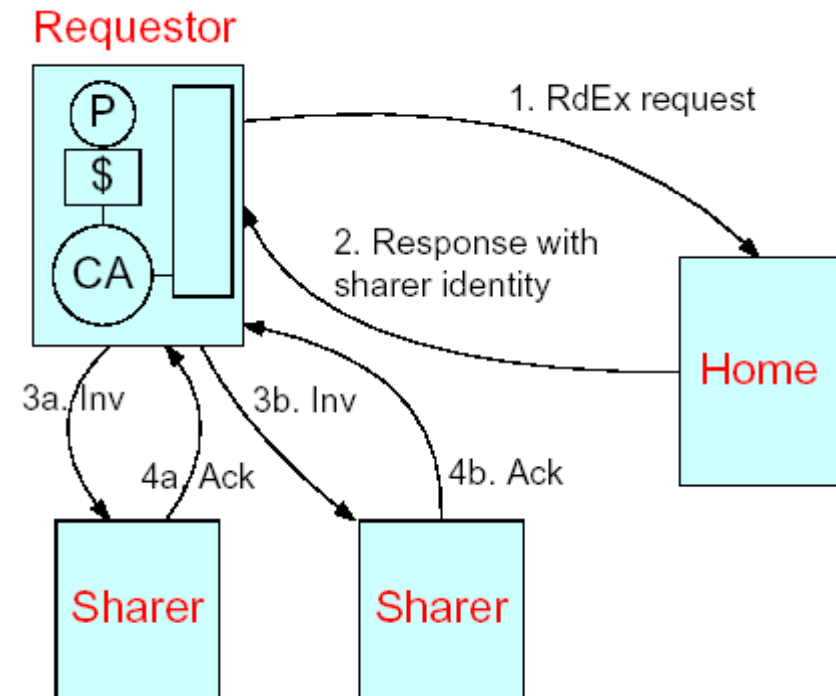


- 第*i*个处理器从主存读取数据：
 - 如果主存的 dirty-bit是OFF，则直接从主存读取，并令 $p[i]$ 为ON;
 - 如果主存的 dirty-bit是ON，则 只有一个处理器的存在位为“1”，从该处理器读取数据到主存，将dirty-bit设为OFF;并令 $p[i]$ 为ON; 将数据返回给处理器*i*;
- 第*i*个处理器写数据到主存：
 - 如果主存的 dirty-bit是OFF，则使得包含该数据的所有缓存块为无效，清除所有 $p[k]$;将dirty-bit设为ON;并令 $p[i]$ 为ON...;
 - 如果主存的 dirty-bit是ON?

基本目录事务



Read Miss



Write Miss

- 主节点 (**Home node**) : 包含主存储区的节点
- 所有者节点 (**Owner node**) : 要提供数据的节点

讨论

- 多处理器（Multiprocessor）和多计算机（Muticomputer）
 - 体系结构
 - 存储
 - 通讯
- 请列出以下存储模型和体系结构的独特之处
 - Parallel system architecture (PVP,SMP,MPP,DSM,COW)
 - Memory models (UMA,NUMA,CC-NUMA,NCC-NUMA,COMA,NORMA)

总结 (1)

Features	PVP	SMP	MPP	DSM	COW
Architecture	MIMD	MIMD	MIMD	MIMD	MIMD
Processor Type	Customer-Designed	Commercial	Commercial	Commercial	Commercial
Interconnection Network	Customer-Designed Crossbar Switcher	Bus, Crossbar Switcher	Customer-Designed	Customer-Designed	Commercial Network (eg. Ethernet)
Comunication	Shared Variable	Shared Variable	Message Passing	Shared Variable	Message Passing
Address Space	Single	Single	Multiple	Single	Multiple
Memory Access	Shared	Shared	Distributed	Distributed Shared	Distributed
Memory Model	UMA	UMA	NORMA	NUMA	NORMA
Example Machine	Cray C-90, Cray T-90, 银河1号	IBM R50, SGI Power, Sun Starfire, 曙光1号	Intel Paragon, IBM Option White, 曙光 1000/2000	Standford DASH, Cray T3D	Berkeley NOW

总结（2）

- 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의 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。

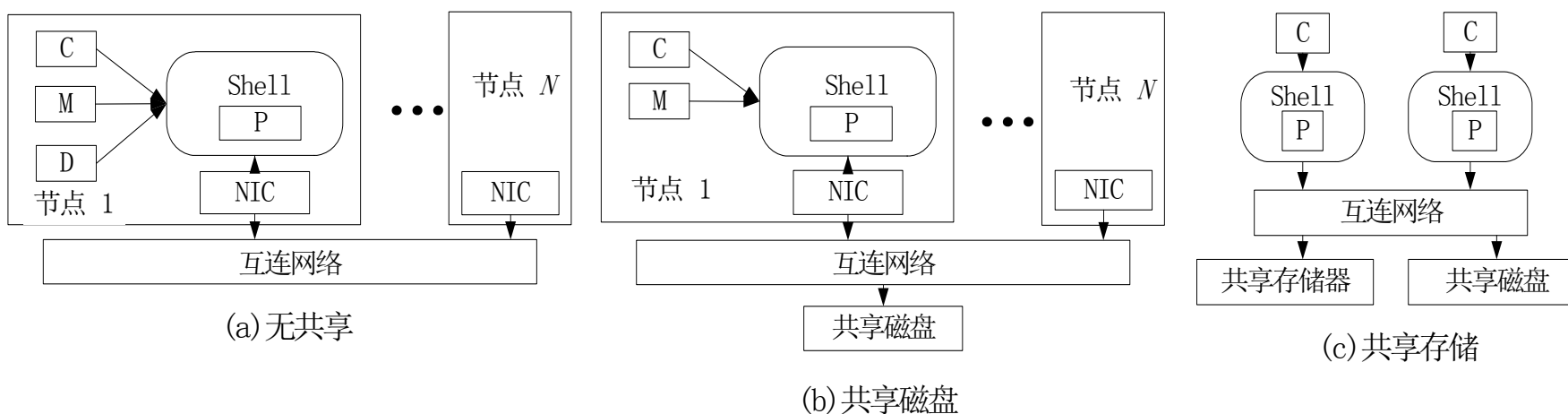


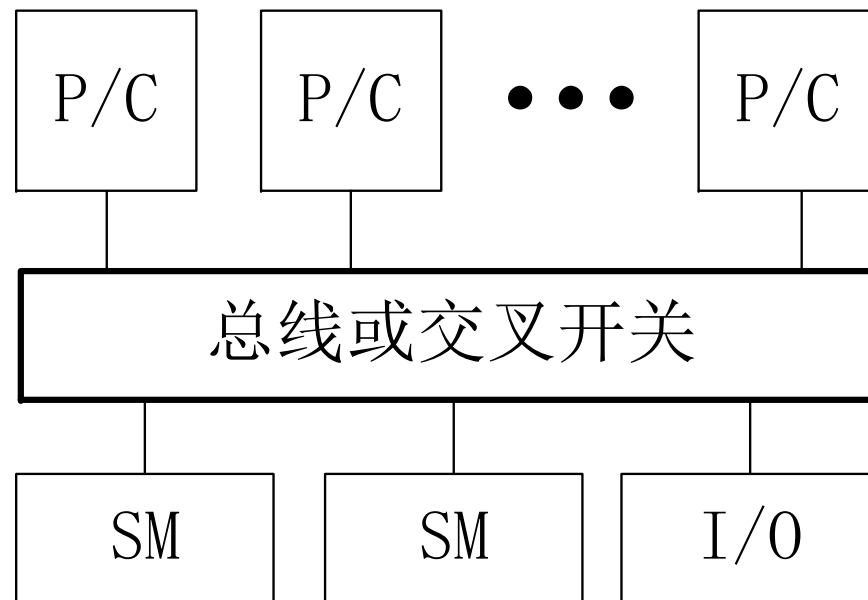
图1.21

主要内容

- 并行计算机访存模型
- 并行计算机存储组织
- 并行计算机系统
 - SMP
 - MPP
 - Cluster

对称多处理器SMP

- **SMP (Symmetric Multiprocessing)** : 采用商用微处理器, 通常有片上和片外Cache, 基于总线连接, 集中式共享存储, **UMA**结构
- 例子: **SGI Power Challenge, DEC Alpha Server, 曙光 (Dawning) 1, HP SuperDome, Sun SunFire, IBM Regatta**



现代商业 SMP 系统

System Features	HP 9000 Superdome	Sun Fire 15K	IBM P690 Regatta
CPUs	128	106	32
CPU Types	PA-8800 1.0GHZ	UltraSPARC 3 1.2GHz	POWER4+ 1.90GHz
Partitions	16/128	18	32
Maximum memory per partition	1TB	1/2 TB	1TB
Total hot-swap PCI-X I/O slots	96/192 slots	72 slots	160 slots
Interconnect Network	Crossbar 64GB/s (Peak)	150 MHz Sun Fireplane	Crossbar
I/O bandwidth	32GB/s (Peak)	21.6 GB/s (Sustained)	44GB/s (Aggregated)
Memory bandwidth (Peak)	256GB/s	172.8GB/s	205GB/s

SMP的优缺点

- 优点
 - 对称性
 - 单地址空间，易编程性，动态负载平衡，无需显示数据分配
 - 高速缓存及其一致性，数据局部性，硬件维持一致性
 - 低通信延迟，Load/Store完成
- 问题
 - 欠可靠，BUS，OS，SM
 - 通信延迟（相对于CPU），竞争加剧
 - 慢速增加的带宽（MB double/3年，IOB更慢）
 - 不可扩放性---> CC-NUMA

CC-NUMA

- CC: Cache Coherent

NUMA: Non-Uniform Memory Access

- Memory是物理分布的
- Memory是全局可存取的
- 本地存储快于远程存储的（Non-Uniform Memory Access = NUMA）
- 不同节点的本地存储不互相影响

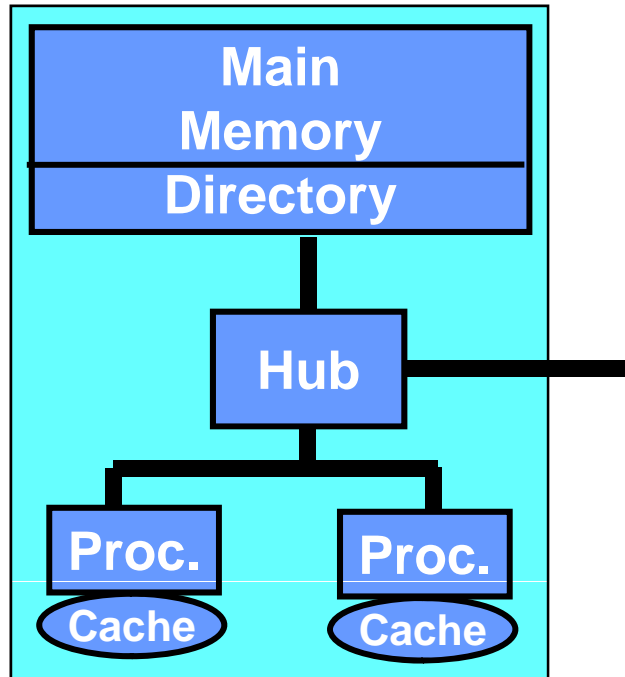
- 例子：

- Stanford Dash：第一个使用基于目录的缓存一致性解决方案
- SGI Origin 2000（Silicon Graphics Inc.）：可支持多至1024个处理器，曾经占据95%的cc-NUMA型机器的市场

SGI Origin 2000

- O2K的体系结构
 - CC-NUMA
 - 节点板：双CPU
 - 互联：超立方（Hypercube）
 - 缓存一致和虚拟内存
 - 分布式存储构成单一的（逻辑的）存储空间，任何处理器都可以访问
 - 可扩展型：时延和带宽

Origin 2000 节点板



节点板 (Node Board)

- **MIPS R12000 CPU**
 - 64-bit RISC design, 400MHz
 - 5 fully-pipelined execution units
 - 8MB L2 cache
 - 32KB 2-way set-associative instruction and data caches

路由和互联结构

- 6路的非阻塞交叉开关（6-way non-blocking crossbar）：9.3 GBytes/sec
 - 每个端口1.56 Gbyte/sec（双工）
- 虫蚀选路（wormhole routing）技术
- 源同步驱动器SSD（Source Synchronous Driver）和源同步接收器SSR（Source Synchronous Receiver）
 - 将390MHz的外部信号转换为核心频率97.5MHz
 - 64位/16位转换
- 链路级协议LLP（Link-Level Protocol）
 - 确保消息的可靠传输

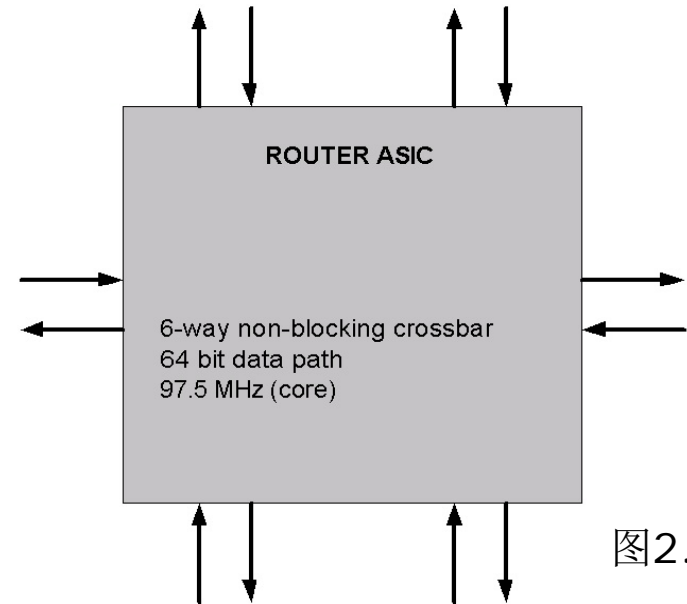
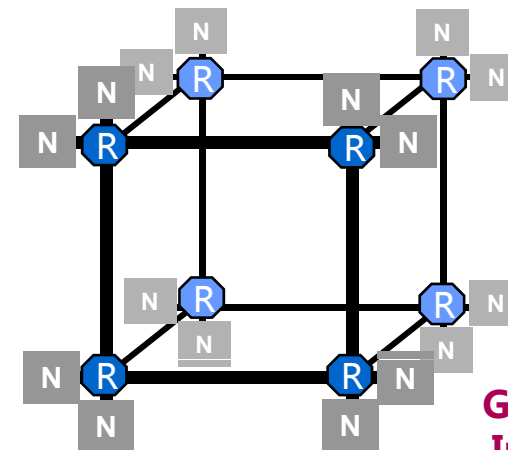
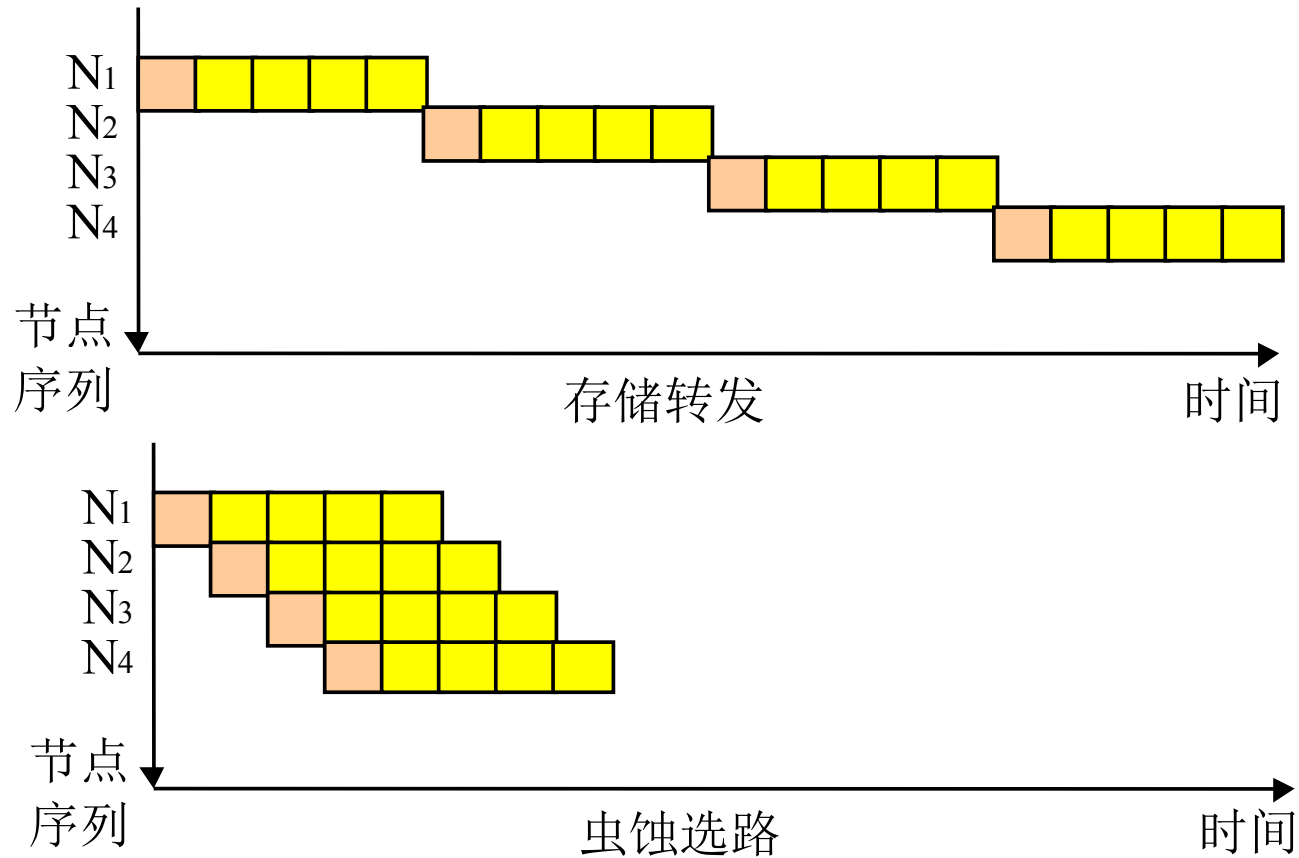


图2.5



Global Switch Interconnect

多处理机系统数据通信方式



- 所有包经过同样的路由，没有路由信息
- 强制按序发送，消除包序列开销
- 校验信息加在消息级，消除包校验开销

例子

- 设多处理器计算机中两个结点之间的距离为**10**，一个处理器发出的消息包含**100**个片段（**flit**），假设每个时钟周期可以在连路上传递一个片段，问在存储转发和虫孔寻径两种方式下消息的传递最快分别需要花费多少时间？

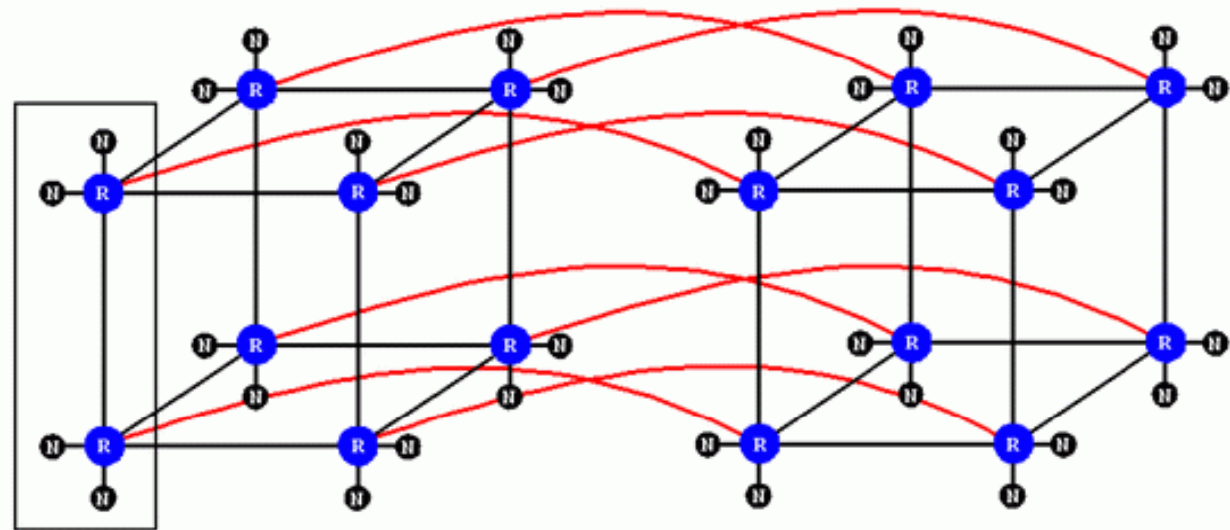
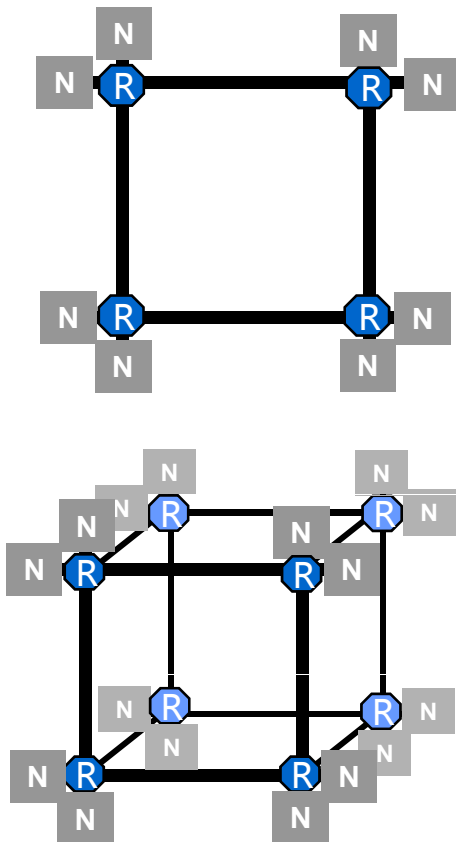
存储转发方式，消息传递时间为

$$10 * 100 = 1000 \text{ 个时钟周期}$$

虫蚀选路方式，消息的第一个片段在网络上的传递时间为**10**个时钟周期，后**99**个片段增加**99**个时钟周期，共**109**个时钟周期。

互联拓扑

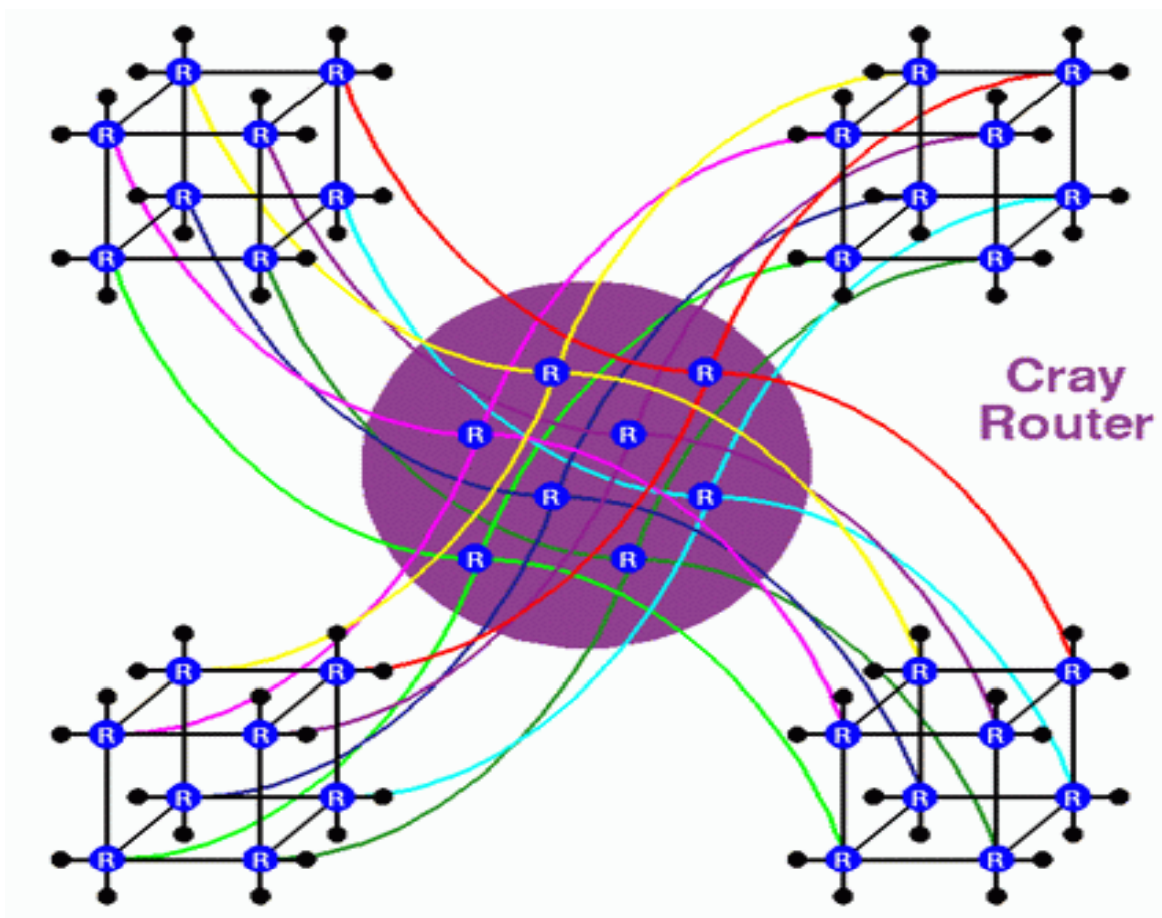
- 互联拓扑: “**fat hypercube**” (胖超立方)
- 每个路由器 (router) 连接一对节点
- 路由器通过超立方连接



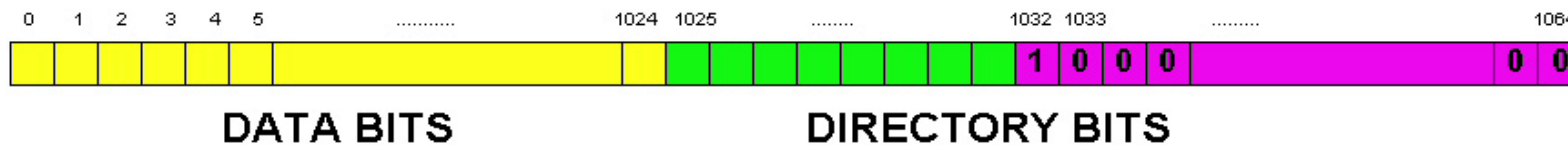
Directly connect two 32-node systems via Craylink cables
using the one free link on each router

元路由器

- 64个节点



基于目录的缓存一致性

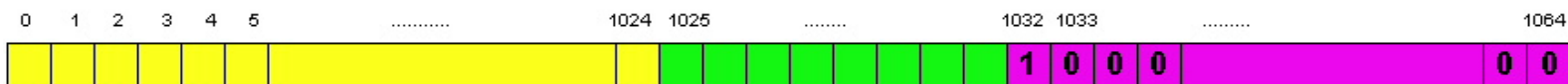


目录比特位由以下两部分组成:

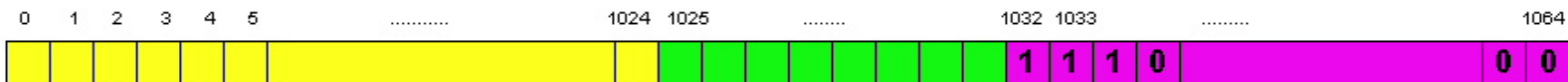
- **8-bits integer** (绿颜色): 表示拥有数据的节点
- **Bit map** (紫颜色): 表示有数据拷贝的节点

缓存管理的例子

1. CPU 0的线程将数据读入缓存



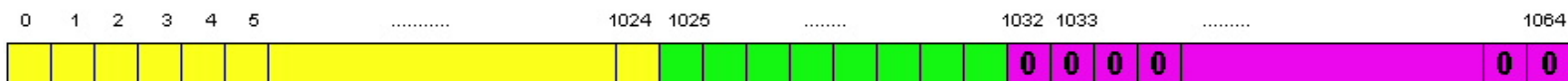
2. CPU 1和2的线程将数据读入缓存



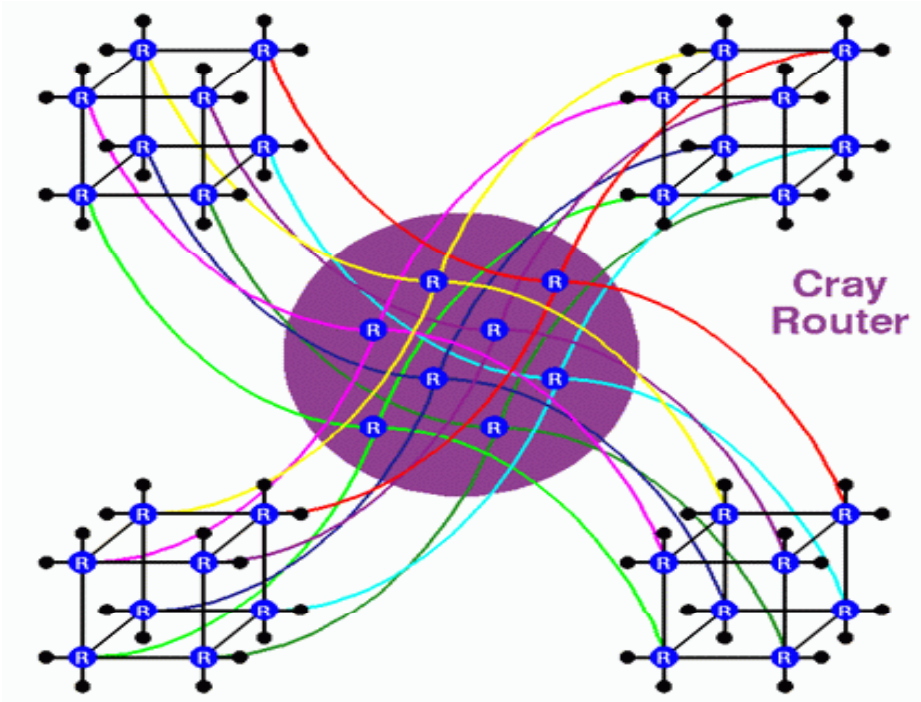
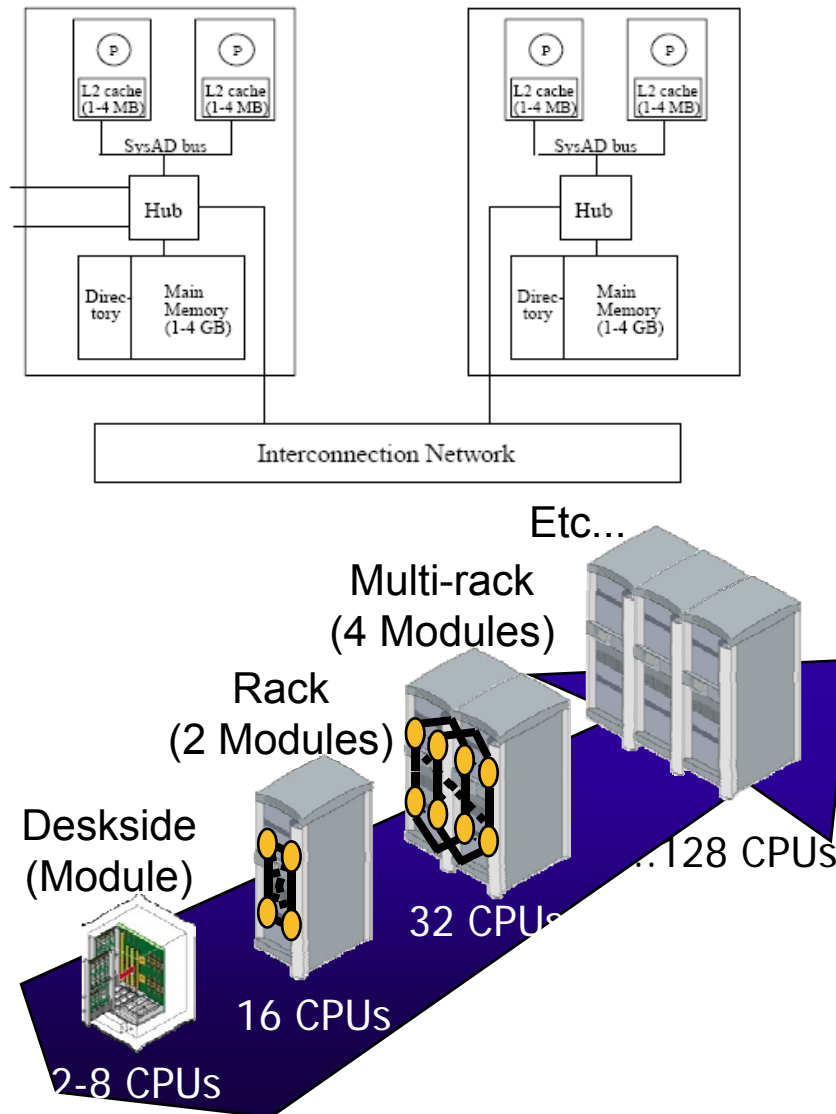
3. CPU 2的线程修改了缓存的数据



4. 将所有缓存块的数据都设为无效



Origin 2000 – 128 Processors



O2K存储时延

数据位置	读时延（时钟数）
寄存器	0
L1 Cache（片上）	1-3
L2 Cache（片外）	10
本地存储器	61
远程存储器 （1个路由器以远）	117
远程存储器 （2个路由器以远）	137
远程存储器 （3个路由器以远）	157
远程存储器	97+20*(number of router hops)

表2.3

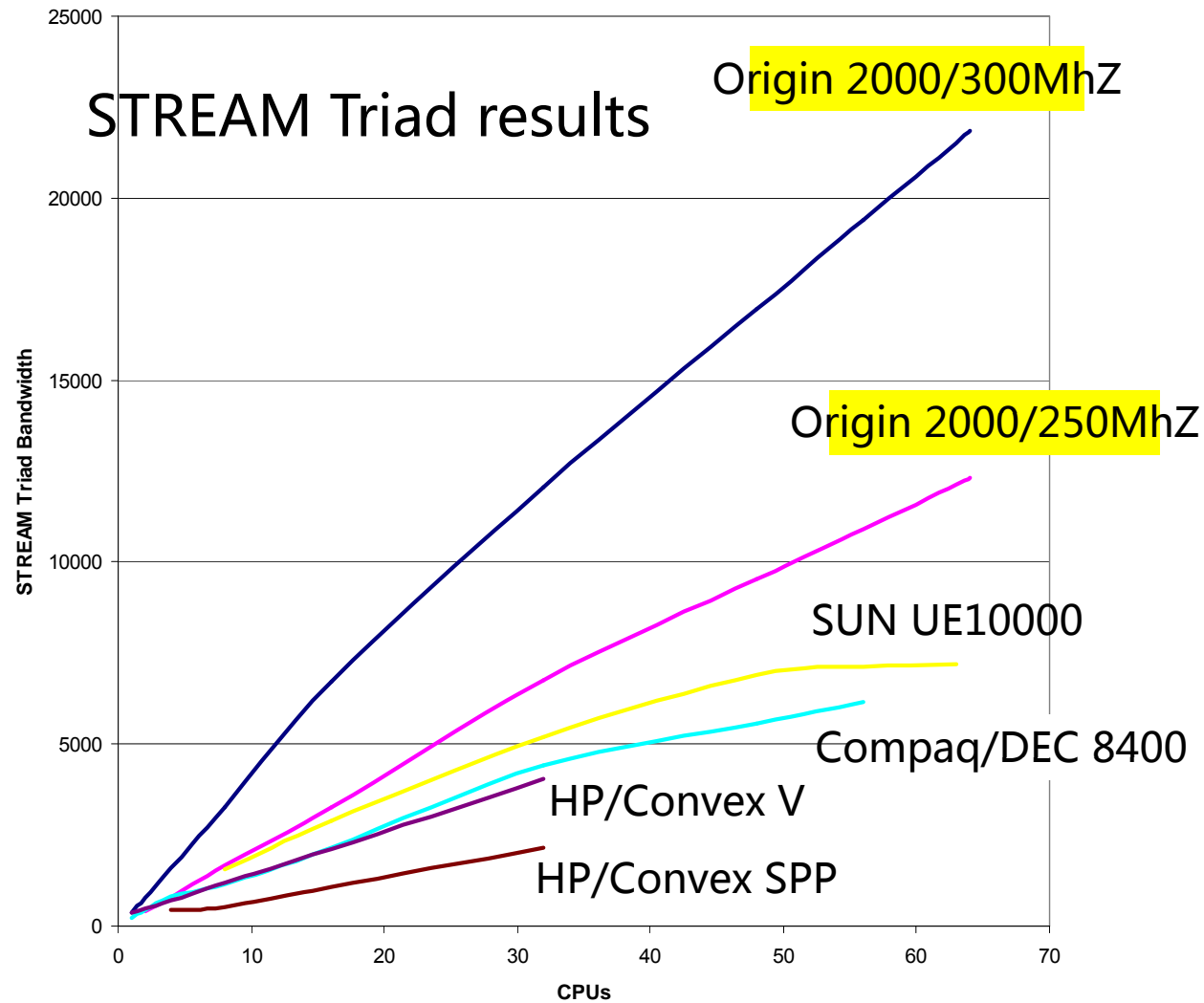
带宽 (GB/s)和延迟 (ns)

System Configuration I/O:Node:CPU	Peak Memory BW (total)	Peak I/O BW (total)	Bisection BW (total)	Router Hops (avg/max)	Memory Latency (avg)
1:1:2	0.78	1.56	-	-	343
2:2:4	1.56	3.12	-	-	441
2:4:8	3.12	6.24	1.56	0.75/1	623
4:8:16	6.24	12.48	3.12	1.63/2	691
8:16:32	12.48	24.96	6.24	2.19/3	674
16:32:64	24.96	51.20	12.48	2.97/5	851
32:64:128	49.92	99.84	24.96	3.98/6	959

表2.4

Bisection BW/CPU = ?

Origin 2000的扩展性



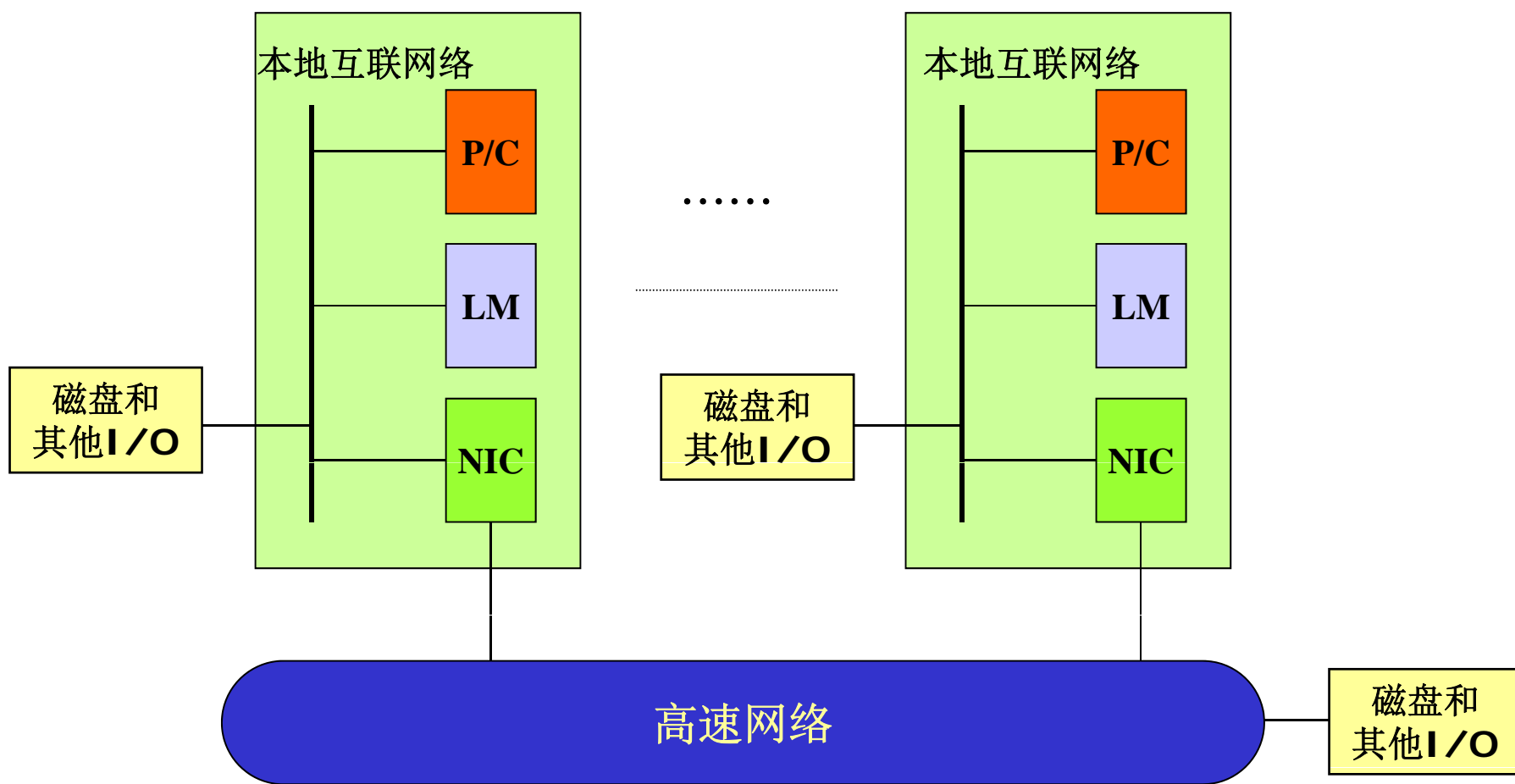
主要内容

- 并行计算机访存模型
- 并行计算机存储组织
- 并行计算机系统
 - SMP
 - MPP
 - Cluster

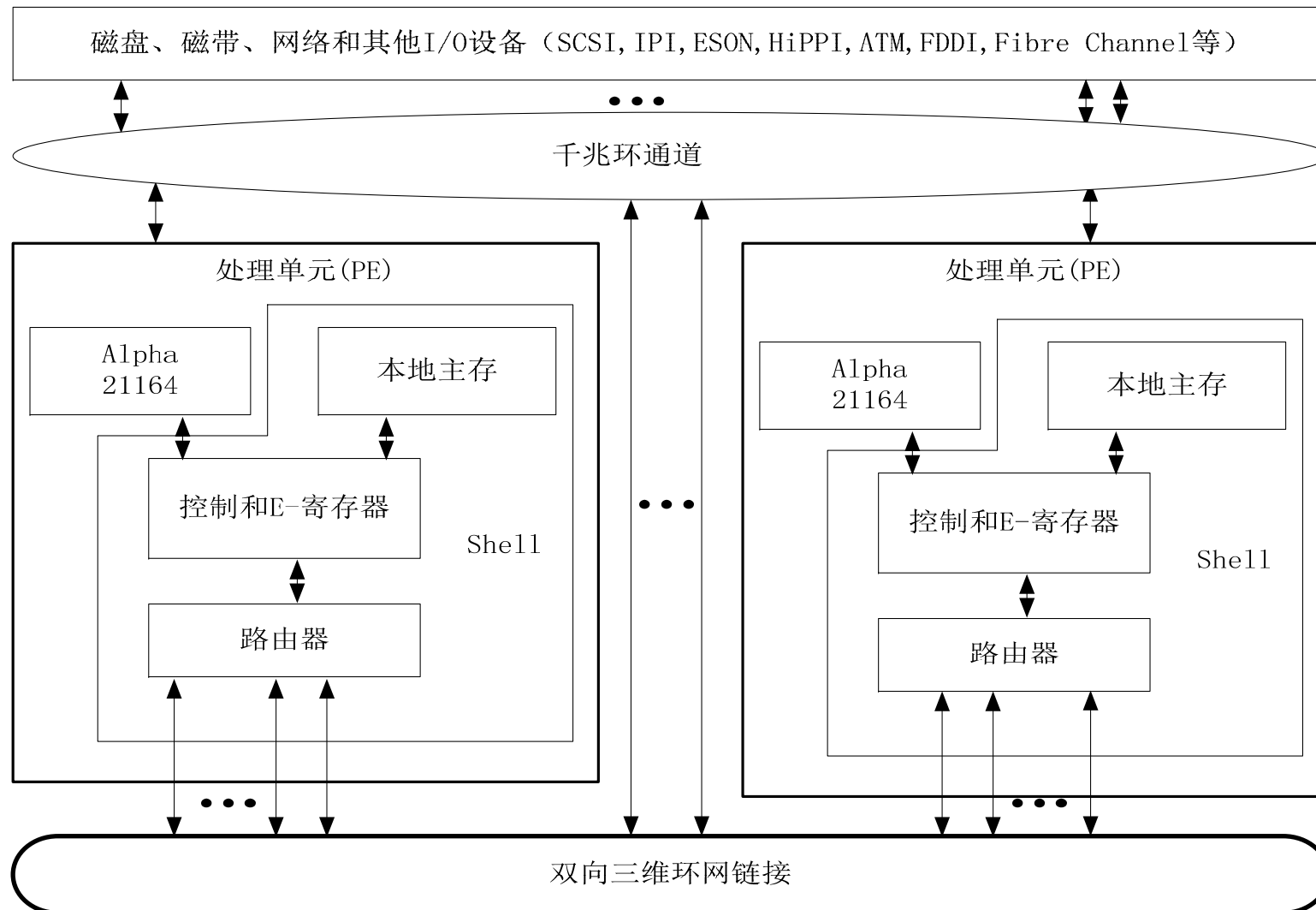
大规模并行处理机MPP概述

- 大规模并行处理机MPP（Massively Parallel Processor）通常是指具有下列特点的大规模的计算机系统：
 - 节点中使用商品化微处理器，且每个节点有一个或多个微处理器
 - 节点内使用物理上分布的存储器
 - 具有高通信带宽和低延迟的互连网络，节点间紧耦合
 - 能扩展成具有成百上千个处理器
- 两种实现途径
 - NCC-NUMA体系结构，Cray T3E
 - NORMA体系结构，Intel/Sandia ASCI Option Red

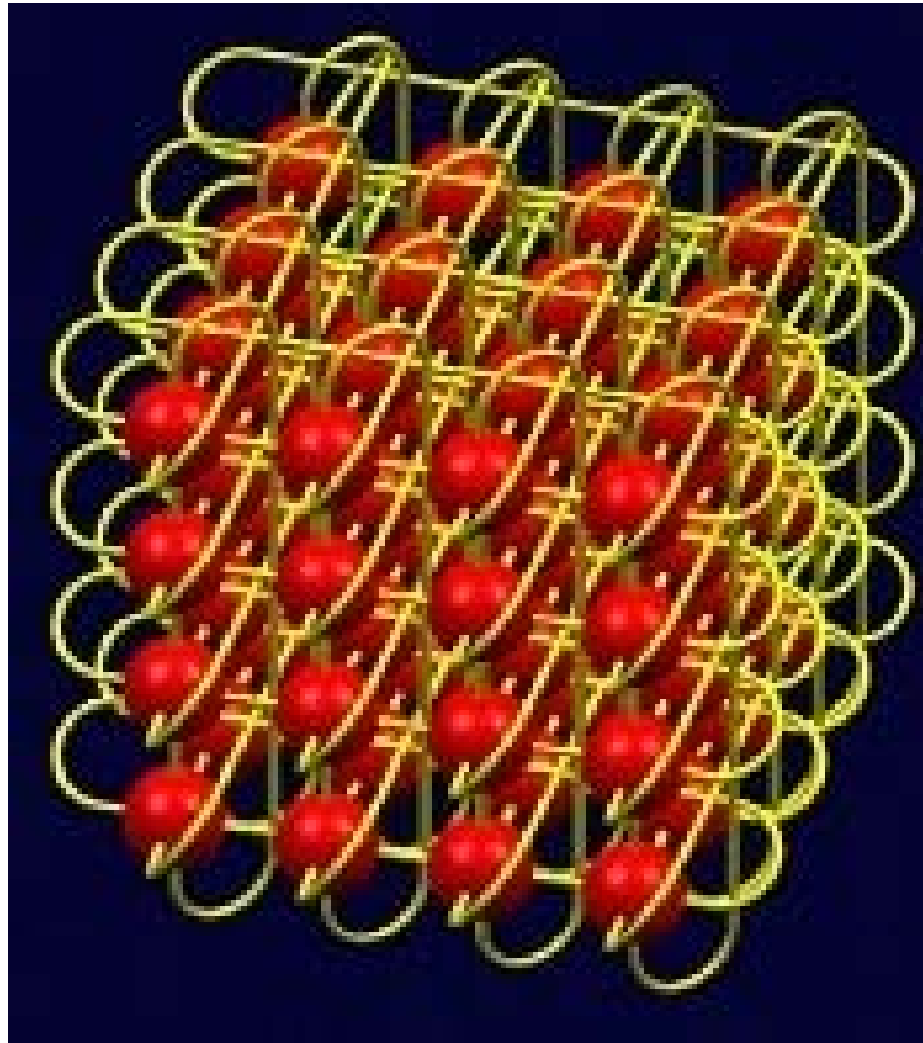
MPP的结构图



实例分析1: Cray T3E的体系结构



三维环网 (3D Torus)



性能特点

- 分布式共享主存（NCC-NUMA）的多处理机
 - 多个处理单元PE（Processing Element）通过一个三维双向环网互连
 - 由一些千兆环通道提供与I/O设备的连接
- T3E-900是1996年底发布的T3E增强型。

属性	T3E	T3E-900
处理器时钟频率(MHz)	300	450
峰值处理器速度(Mflops)	600	900
处理器数量	6~2048	6~2048
系统峰值速度(Gflops)	3.6~1228	5.4~1843
物理主存容量(GB)	1~4096	1~4096
总峰值主存带宽(GB/s)	7.2~2450	7.2~2450
I/O通道最大数目	1~128	1~128
总峰值I/O带宽(GB/s)	1~128	1~128
峰值三维环网链接带宽(MB/s)	600	600

实例分析2: Intel/Sandia ASCI Option Red

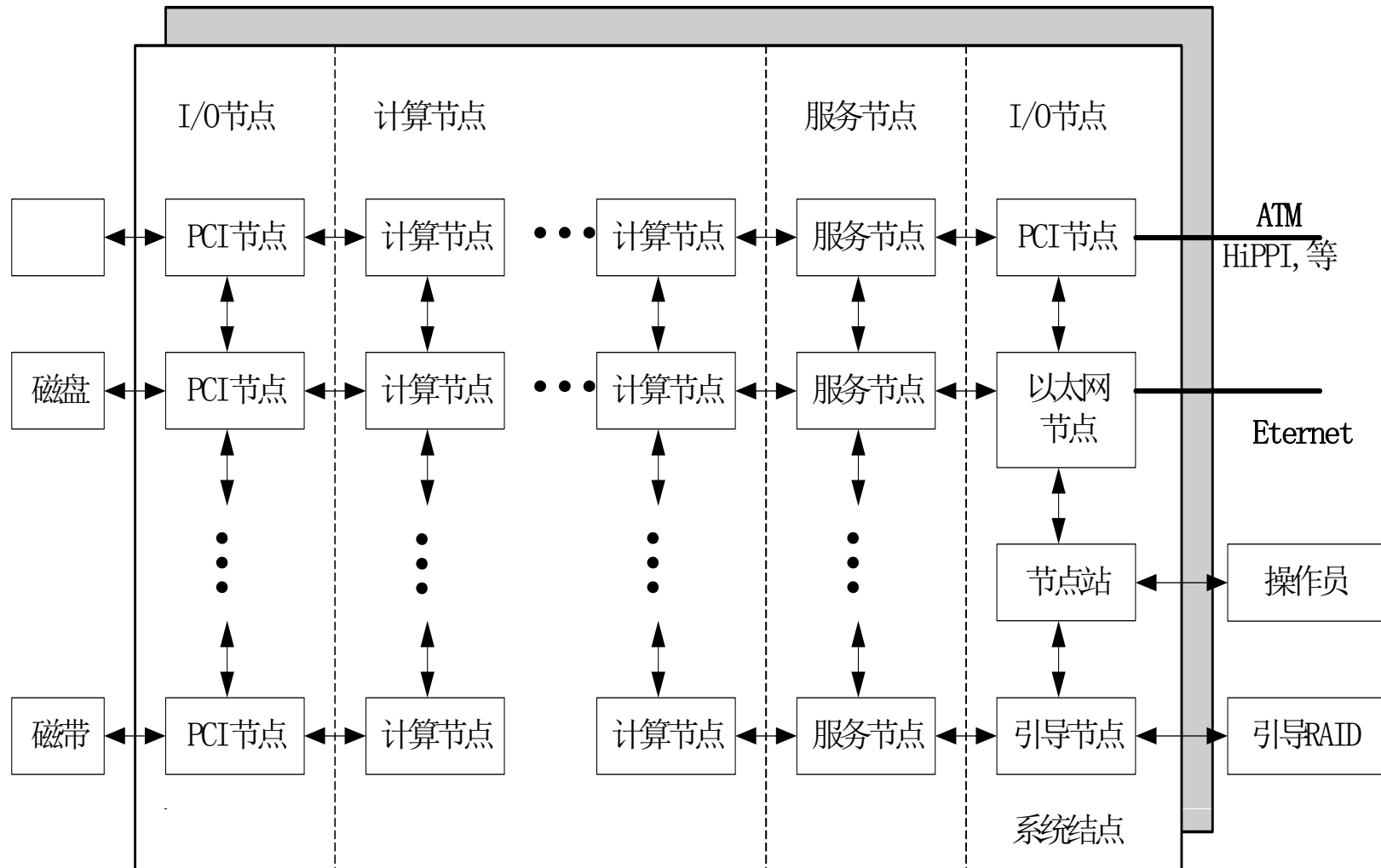


图2.9

ASCI Option Red



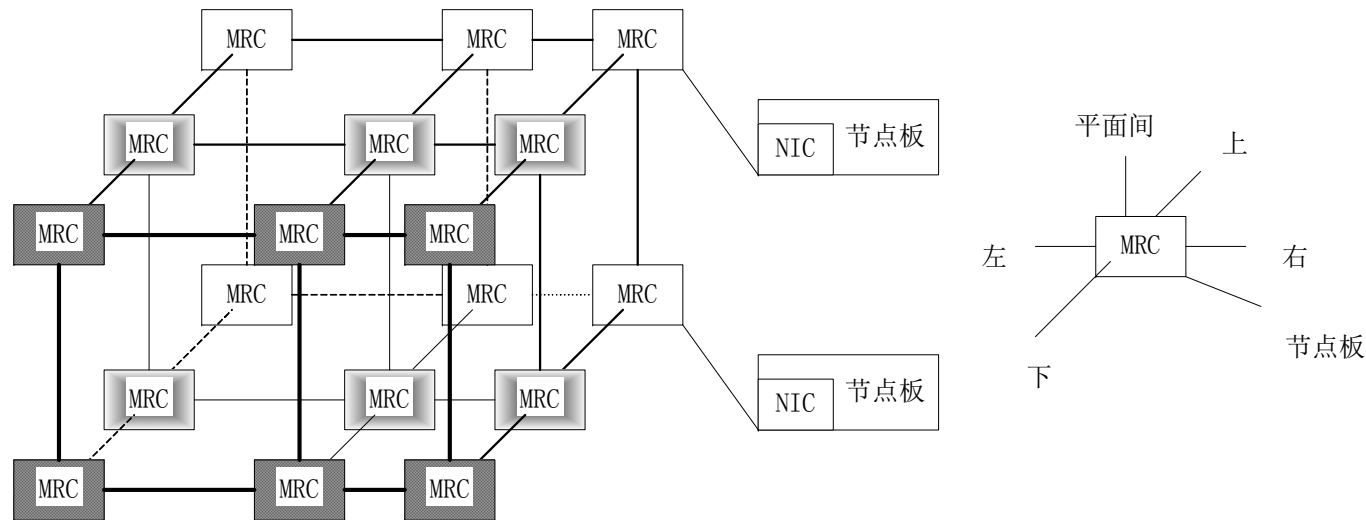
Option Red的体系结构

- 世界第一台Teraflops级超级计算机（98年2月），峰值：1.8Tflop/s，放在美国Sandia国家实验室（1997.6）
- 共有4608个节点（每个节点有两个200MHz Pentium Pro处理器）和594GB的主存，其峰值速度为1.8Tflop/s、峰值截面（Cross-Section）带宽为51 GB/s
 - 计算节点（Compute Node）4536个，执行并行计算
 - 服务节点（Service Node）32个，用于支持登录、软件开发及其它交互操作
 - I/O节点（I/O Node）24个，用于存取磁盘、磁带、网络(以太网、FDDI、ATM等)和其它I/O设备
 - 系统节点（System Node）2个，用于支持系统RAS（reliability, availability, serviceability）能力：其中引导节点（Boot Node）负责初始系统引导及提供服务；节点站（Node Station）用于单一系统映象支持
 - 备份节点。
- 1540个供给电源，616个互连底板和640个磁盘（大于1TB的容量）

<http://www.sandia.gov/ASCI/Red/>

系统互连

- 节点由一个内部互连设备ICF相连
 - ICF使用了双平面（Two-Plane）网孔拓扑
 - 每个节点主板通过主板上的NIC网孔选路部件MRC（Mesh Routing Component）。MRC有六个双向端口，每个能以400 MB/s的单向峰值速度传送数据，全双工时为800 MB/s，4个端口用于平面内左、右、上、下的网孔互连，还有一个端口用于平面间互连
 - 从任意节点发出的消息借助虫蚀选路通过任一平面送至另一节点，这将降低时延，从而提高了系统可用性



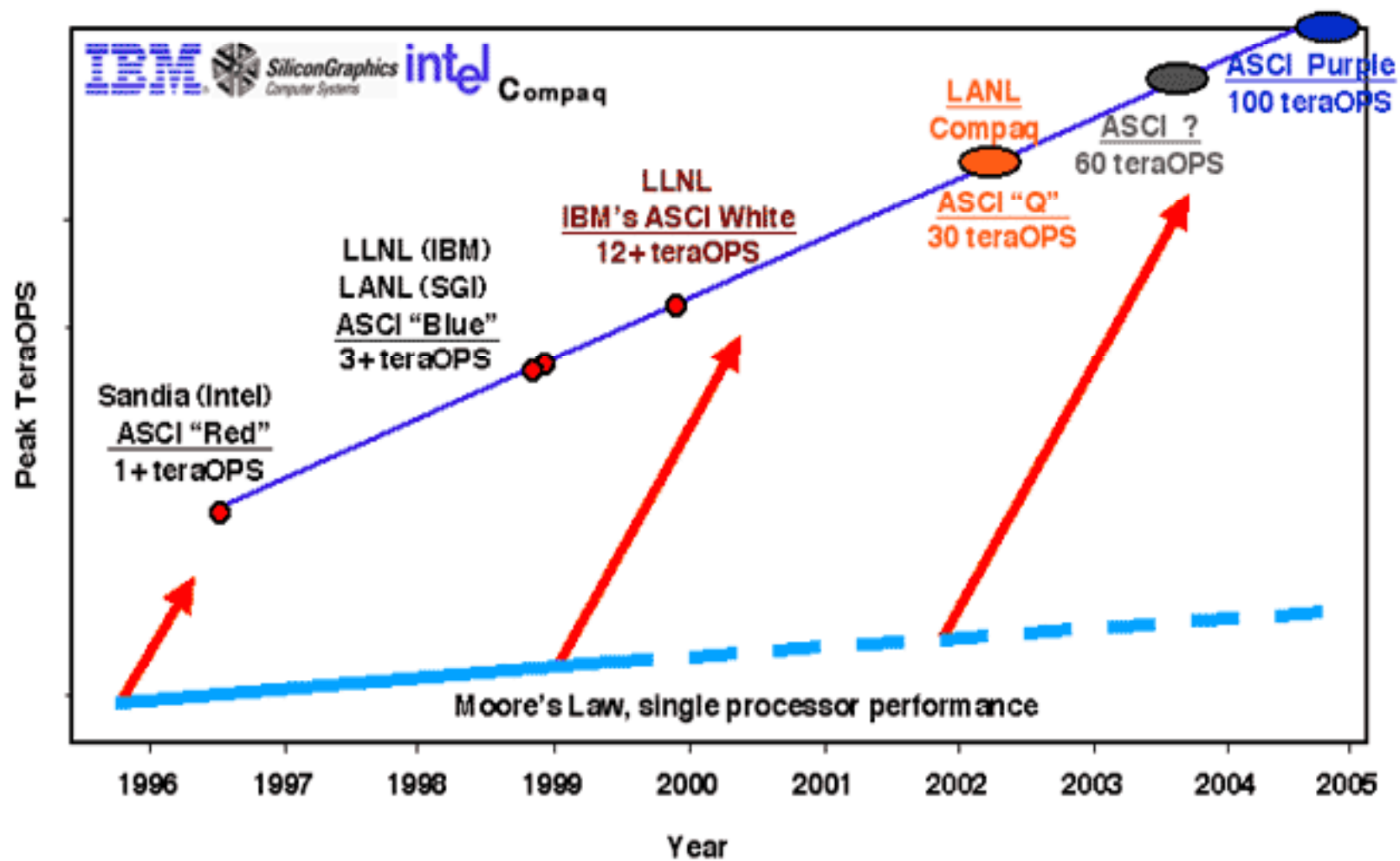
Option Red的系统软件

- 系统、服务和I/O节点都运行Paragon操作系统，它是一个基于OSF的分布式Unix系统
- 计算节点运行一个称为Cougar的轻量级内核LWK（Light-Weight Kernel）（图2.12）
 - LWK设计更强调性能，它能有效支持多达几千个节点的MPP，只提供并行计算所需的功能，而不是一般的操作系统服务
- 同时提供了对这两个系统间接口的支持，包括高速通信、Unix编程接口和一个并行文件系统

ASCI可扩充设计策略

- 加速战略计划创新ASCI（Accelerated Strategic Computing Initiative）
 - 1994年DOE
 - 1996年 1Tflop/s系统，2000年 10至30Tflop/s系统，2004年 100Tflop/s系统，且这些系统应该成本相近。
 - 不仅瞄准峰值速度，而且总的系统持续的应用性能要 10^5 倍于1994年
- 平衡的可扩充设计
 - 着重用于科学计算应用的高端平台，而非大批量市场平台和市场热点应用；
 - 使用尽可能多的商品化市售（COTS）硬件和软件部件，着重开发主流计算机公司未有效提供的关键技术；
 - 使用大规模并行体系结构，着重于缩放和集成技术，将数千个COTS节点纳入一个有单一系统映象的高效平台

ASCI 平台性能发展图



平衡设计策略

- 平衡的可扩展硬件
 - **平衡设计准则**：1Gflop/s峰值速度应与1GB主存、50GB磁盘、10 TB档案存储器、16 GB/s高速缓存带宽、3GB/s主存带宽、0.1GB/s I/O磁盘带宽以及1MB/s档案存储器带宽相匹配；
- 平衡的可扩展软件
 - **ASCI**认为新的软件开发将使性能改进10到100倍

属性	1996	1997	1998	2003
应用性能(倍数)	1		1000	100,000
峰值计算速度(Gflops)	100	1000	10,000	100,000
主存容量(TB)	0.05	0.5	5	50
磁盘容量(TB)	0.1~1	1~10	10~100	100~1000
档案存储容量(PB)	0.13	1.3	13	130
I/O速度(GB/s)	5	50	500	5000
网络速度(GB/s)	0.13	1.3	13	130

ASCI平台



ASCI / MPP平台

- Option Red、Blue Pacific、Blue Mountain和Option White等MPP系统已被安装在3个国家实验室
- Intel Option Red 是典型的MPP系统
- SGI Blue Mountain系统由48个节点的机群所组成，其中每一个节点是一个有128个处理器的Origin 2000 CC-NUMA系统。节点内的互连为胖超立方体。48个Origin 2000系统用4兆位HiPPI—800交换开关连成一个机群，其中每个链路的双向峰值带宽为1.6Gb/s
- 2个IBM系统均为高端SP系统
- 其他：HP ASCI Q, ASCI Red Storm, ASCI Purple, IBM BlueGene/L/P

四个ASCI比较

特性	Option Red	Option Blue		Option White
		Blue Pacific	Blue Mountain	
制造商	Intel	IBM	SGI	IBM
安装场所	Sandia	Livermore	Los Alamos	Livermore
完成日期	1997年6月	1998年12月	1998年12月	2000年12月
成本(百万美元)	55	94	<110	85
所选用处理器	Pentium Pro 200MHz 200Mflop/s	PowerPC 604 332MHz 664Mflop/s	MIPS 10000 250MHz 500Mflop/s	POWER3 311MHz 1244Mflop/s
系统体系结构	NORMA-MPP	SMP机群 4CPU/ 节点 1464节点	CC-NUMA机群 128CPU/节点 48节点	SMP机群 16CPU/节点 512节点
节点内连接	总线	交叉开关	胖超立方体	交叉开关
节点间连接	分离2D网孔	Omega开关	千兆位开关	Omega开关
处理器数量	9216	5856	6144	8192
峰值速度	1.8Tflop/s	3.888Tflop/s	3.072Tflop/s	10.2Tflop/s
主存容量	594GB	2.5TB	1.5TB	4TB
磁盘容量	1TB	75TB	75TB	150TB

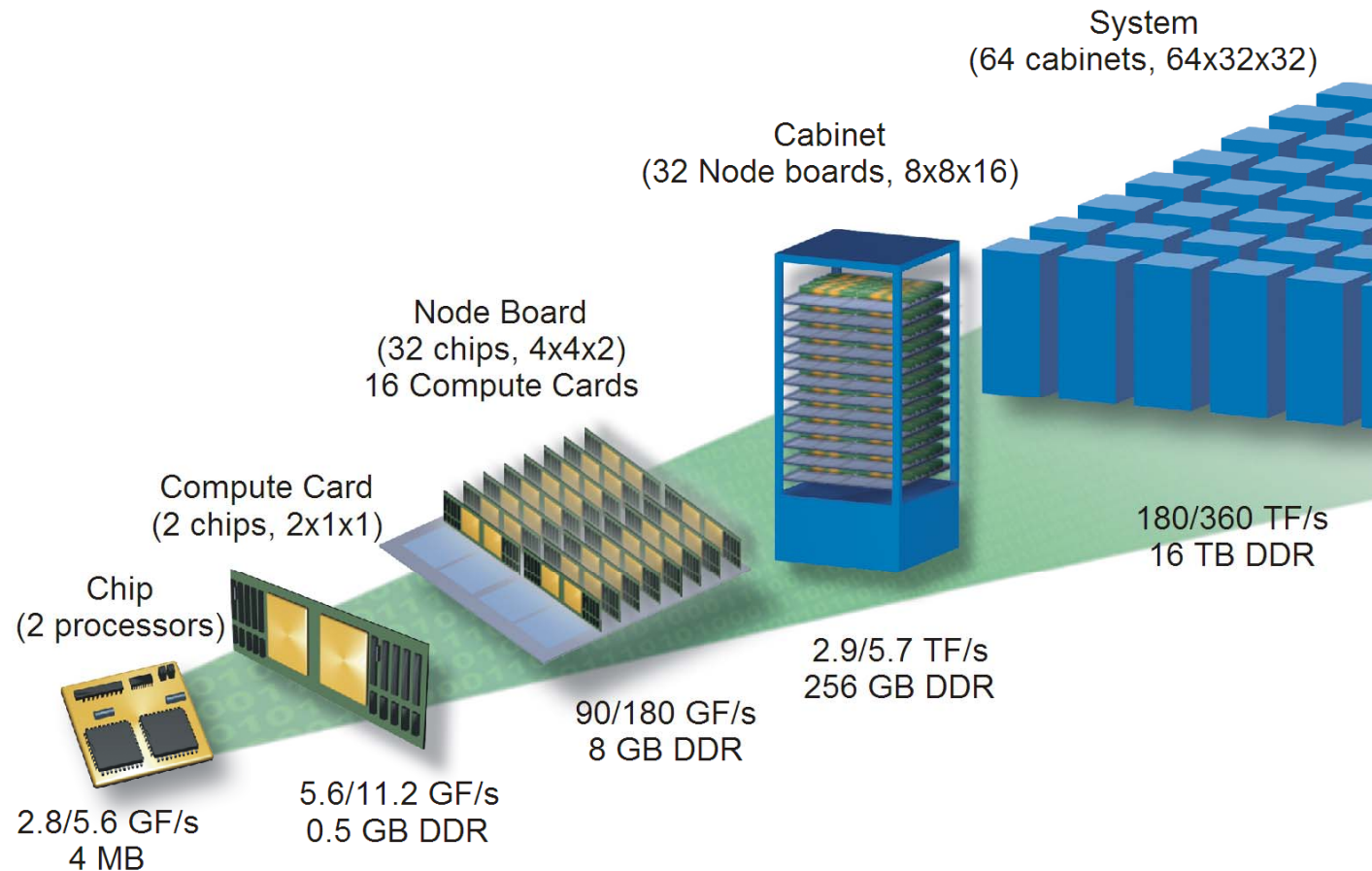
实例分析3：蓝色基因（Blue Gene）

- IBM BG/L是IBM Blue Gene系列的第一台机器。其最初的设计目标是满足蛋白质折叠的需求，以及更广泛的商业应用
- IBM BG/L的体系结构特性使得它具备了功耗低、造价低、体积小的特点，同时性能也达到了每秒百万亿次的级别。目前全世界已经有28台BG系统
- 安装在美国国家实验室LLNL的最大的BG/L（安装时间：2004年第二季度到2005年第三季度）
 - 65,000+ 计算节点 (131,000+ 个处理器)，放在 64个机柜中
 - 每个节点有2个低功耗的700-MHz PowerPC处理器
 - 互联：三维环网（3D torus）+树（tree）
 - 峰值： 367 Tflop/s； Linpack测试： 280 Tflop/s（2007.11世界第一）

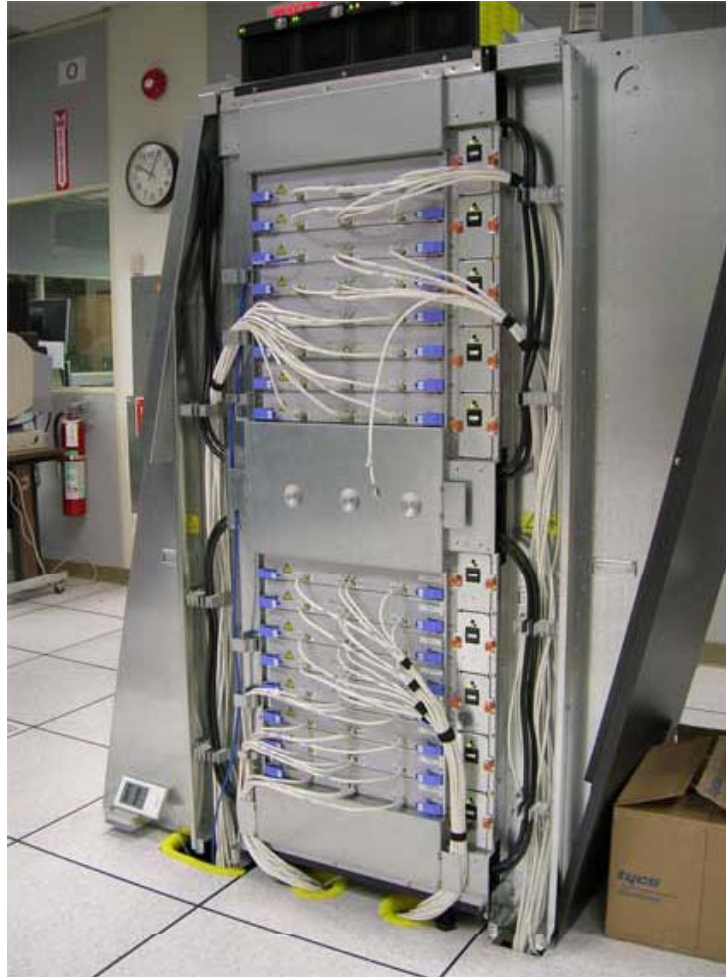
TOP 500中的Blue Gene

Site	Cores	Rmax (TFLOPS)	Rpeak (TFLOPS)
#5: Forschungszentrum Juelich (FZJ), Germany, 2009 (#3)	294912	825.50	1002.70
#8: DOE/NNSA/LLNL, USA, 2007 (#5)	212992	478.20	596.38
#9: Argonne National Laboratory, USA, 2007 (#7)	163840	458.61	557.06
#12: DOE/NNSA/LLNL, USA, 2009 (#9)	147456	415.70	501.35
#23: King Abdullah University of Science and Technology, Saudia Arabia, 2009 (#14)	65536	185.17	222.82
#38: IDRIS, France, 2008 (#24)	40960	116.01	139.26

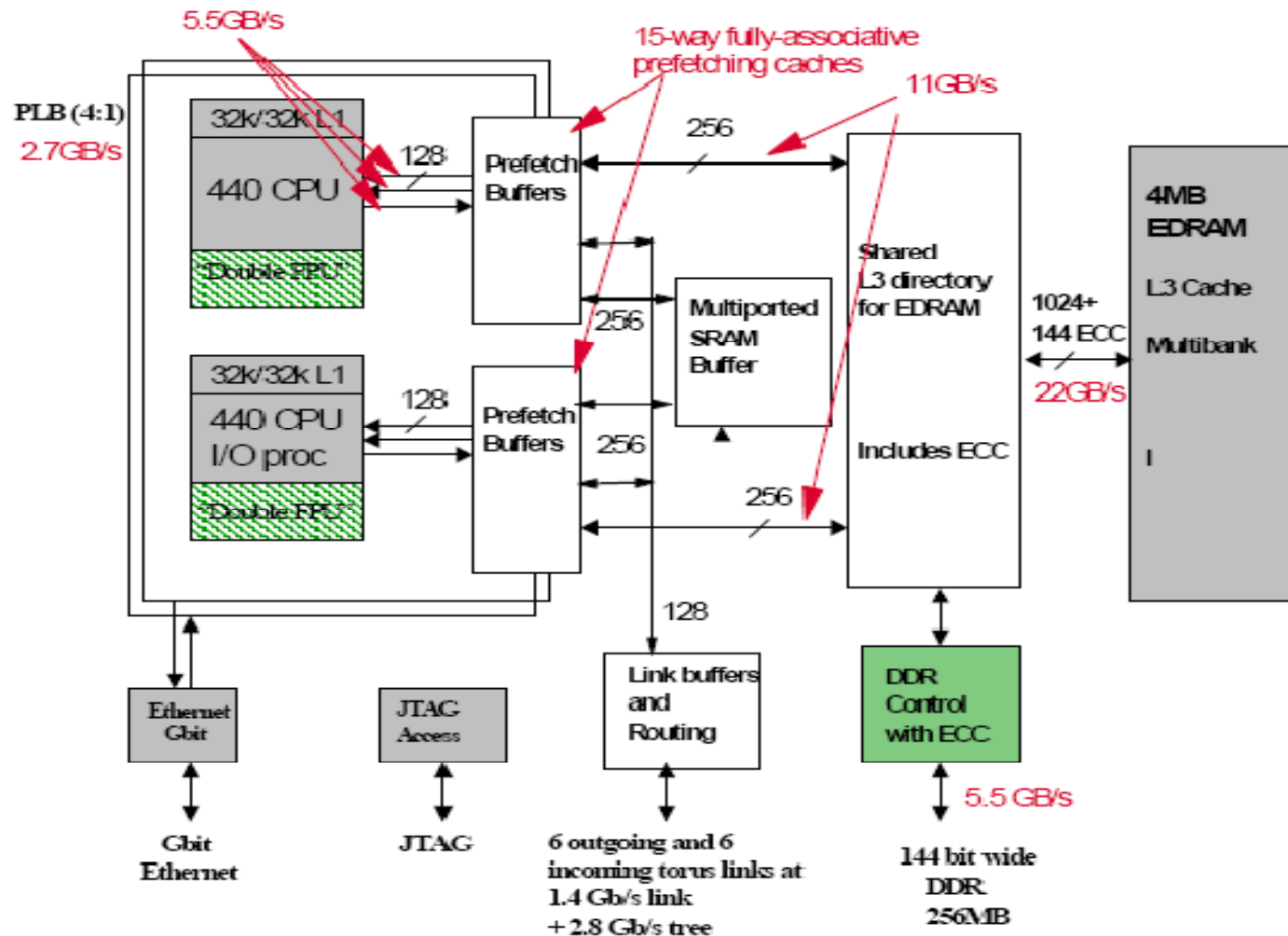
从芯片到机架



网络连接：Gigabit Ethernet



BG/L处理器芯片



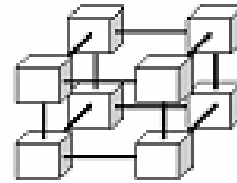
BG/L处理器芯片

(System-on-a-chip: SoC)

- SoC芯片，每个芯片里面集成了多处理器、内存、通信逻辑。就通过复制这样一个个芯片来搭建整个系统。
- 2个700-MHz **PowerPC 440 处理器**
 - 各有两个浮点计算单元（floating-point units）
 - 各有32-kB L1数据缓存（data caches）
 - 4 flops/proc-clock peak (=2.8 Gflop/s per processor)
 - 2 8-B loads or stores / proc-clock peak in L1 (=11.2 GB/s per processor)
- 共享的 2-kB **L2 cache**
- 共享的 4-MB **L3 cache**
- 片外（off-chip）存储控制器：512 MB 或1 GB 的**共享存储**
 - 每个节点的峰值存储带宽为5.6 GB/s

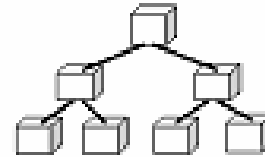
5个网络控制器

- 三维环网络（3D Torus）：可以实现任意两个计算结点之间的点对点传输。在每个三维环的单个链路方向上的硬件带宽是175MB每秒。
- 两个全局的树型网络（Tree）
 - 一个用于实现全局广播操作
 - 另一个用于实现全局的中断响应
- 两个千兆以太网（Gigabit Ethernet）
 - 一个用于实现计算节点与文件系统以及主机之间的通信
 - JTAG（Joint Test Action Group）：用于全局的控制和监视



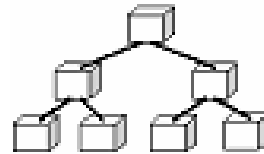
3 Dimensional Torus

- Point-to-point



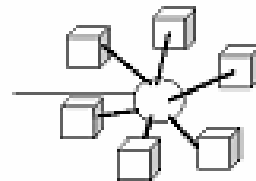
Global Tree

- Global Operations



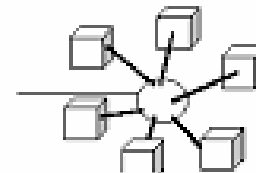
Global Barriers and Interrupts

- Low Latency Barriers and Interrupts



Gbit Ethernet

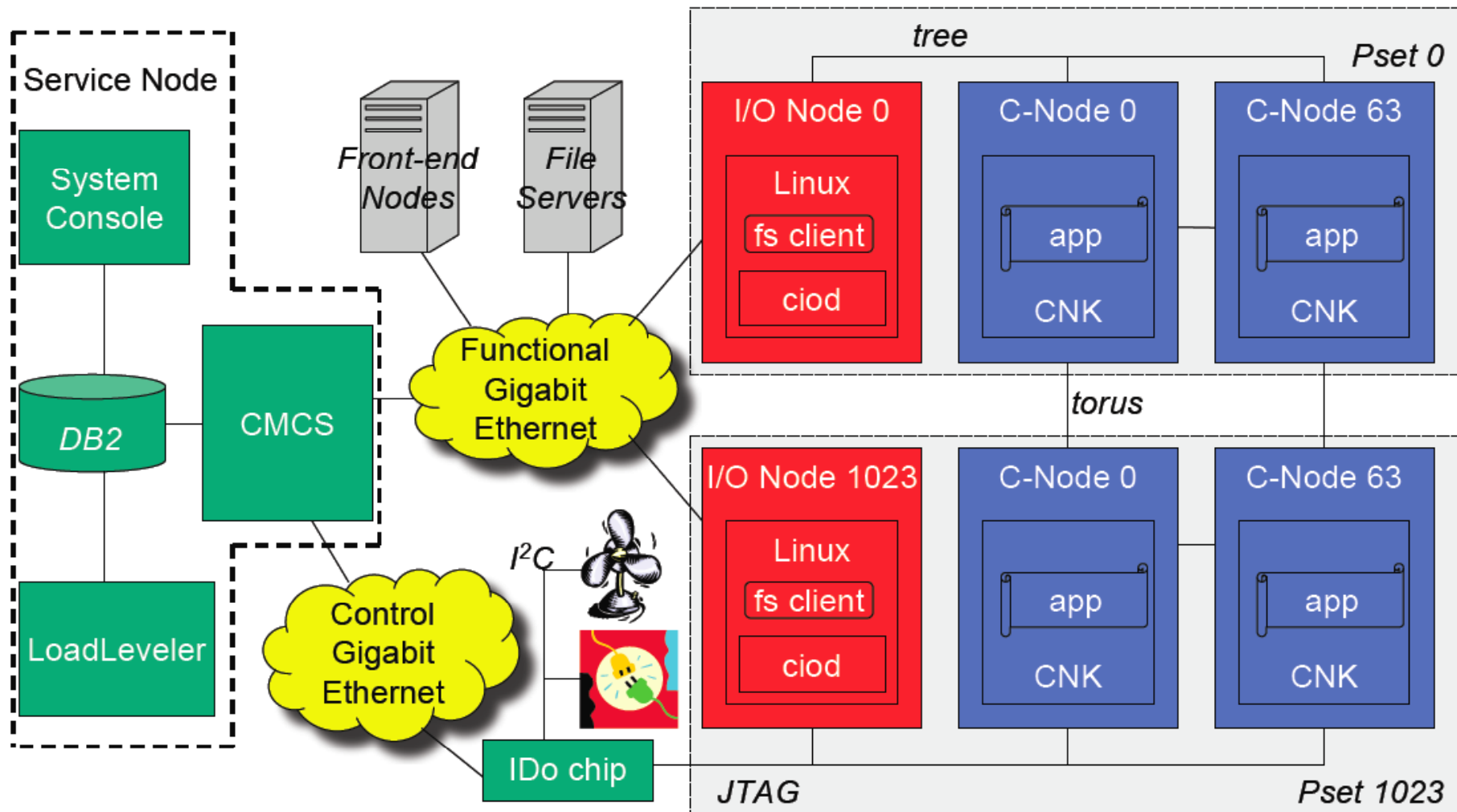
- File I/O and Host Interface



Control Network

- Boot, Monitoring and Diagnostics

集成的BG系统和软件



BG操作系统和功能

- 计算节点：运行CNK（Compute Node Kernel）
 - 每个节点一次运行一个任务
 - CNK只占用很少的内存
- I/O节点（I/O nodes）：运行嵌入式Linux
 - 运行后台程序CIOD来管理计算节点
 - 进行文件I/O
 - 运行并行文件系统（GPFS）
- 前端节点（Front-end nodes）：运行SuSE Linux
 - 支持用户登陆
 - 运行交叉编译和连接
 - 提交作业和管理作业
- 服务节点（Service node）：运行SuSE Linux
 - 使用DB2来管理4个系统数据可
 - 运行系统管理软件

MPP小结

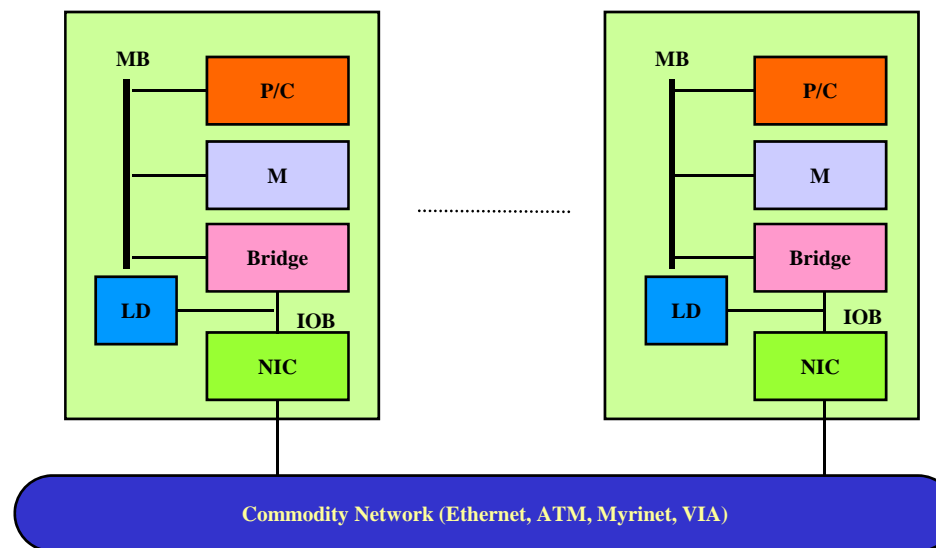
- 八十年代后期及九十年代中前期迅速发展
 - Thinking Machine公司的CM5, Intel公司的Paragon, IBM公司的SP2, 以及Cray公司的T3D
 - 主要被用于科学计算
- 九十年代后期, 随着一些专门生产并行机的公司的倒闭或被兼并, MPP系统慢慢从主流的并行处理市场退出
 - 由于消息传递系统相对共享存储系统比较容易实现, 它仍成为实现超大规模并行处理的重要手段, 不过由于价格和应用领域的原因, 基于消息传递的MPP系统的研制逐渐成为了政府和大公司的行为
 - 新涌现的高性能计算系统绝大多数都将是由可扩放的高速互连网络连接的基于商用微处理器的对称多处理机 (SMP) 集群

主要内容

- 并行计算机访存模型
- 并行计算机存储组织
- 并行计算机系统
 - SMP
 - MPP
 - Cluster

工作站集群COW

- 分布式存储，MIMD，工作站+商用互连网络，每个节点是一个完整的计算机，有自己的磁盘和操作系统，而MPP中只有微内核
- 优点：
 - 投资风险小
 - 系统结构灵活
 - 性能/价格比高
 - 能充分利用分散的计算资源
 - 可扩充性好
- 问题
 - 通信性能
 - 并行编程环境
- 例子：Berkeley NOW，Alpha Farm，FXCOW

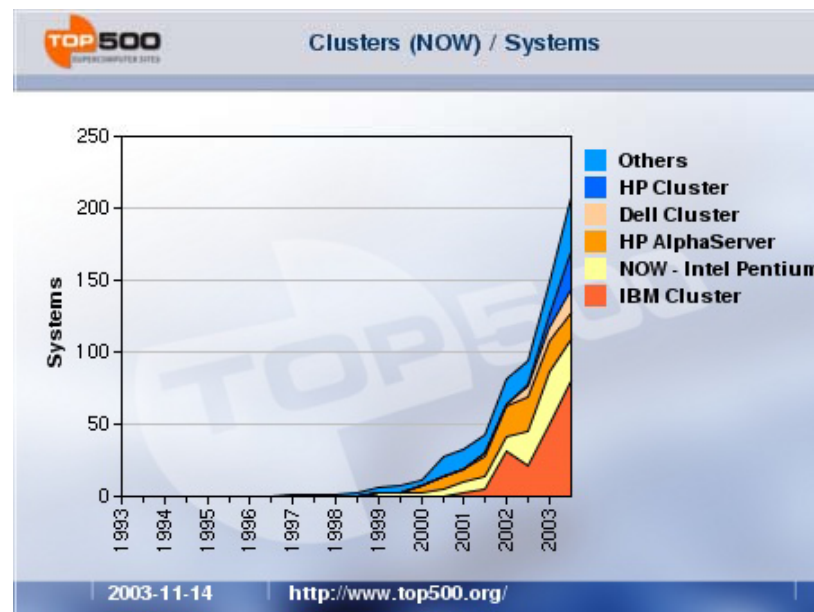


基本概念

- 集群（机群）是一组独立的计算机（节点）的集合体，通常有以下特征：
 - 各节点都是一个完整的系统：工作站，PC机或SMP机器；
 - 互连网络通常使用商品化网络，如以太网、FDDI、ATM等；
 - 网络接口与节点的I/O总线松耦合相连；
 - 各节点通常有一个本地磁盘；
 - 各节点有自己的完整的操作系统。
 - 各节点除了可以作为一个单一的计算资源供交互式用户使用外，还可以协同工作并表现为一个单一的、集中的计算资源供并行计算任务使用。
- 集群与分布式系统的区别：
 - 集群继承了分布式系统的大部分知识
 - 分布式系统通常是一个计算机的动物园，具有许多不同种类的计算机
 - 集群通常是同构，耦合度较紧密，节点间互为信任关系

TOP500中的集群

- 集群系统在高性能计算机中所占比例迅速增
 - TOP500中最普通的并行机体系结构
 - TOP500中目前有400个集群系统
 - 导致了高性能计算机的“平民化”——如2003年11月排名第3的**系统X**（System X）是由美国弗吉尼亚工学院（Virginia Polytechnic Institute and State University, Virginia Tech）的一群师生采用商用部件花了4个月时间制造出来。



集群系统的迅速发展的原因

- 集群价格便宜并且易于构建； 穷人的解决方案，
Gordon Bell奖
- 作为集群节点的工作站系统的处理性能越来越强大
- 局域网上新的网络技术和新的通信协议的引入，高带宽
低延迟的节点间通信
- 集群系统比传统的并行计算机更易于融合到已有的网络
系统中去
- 集群上的开发工具更成熟，传统并行计算机上缺乏一个
统一的标准
- 集群的可扩展性良好

五个关键问题

- 可用性（Availability）：如何充分利用集群中的冗余资源，使系统在尽可能长的时间内为用户服务，检查点、故障接管、错误恢复以及所有节点上的容错支持等
- 单一系统映像SSI（Single System Image）：利用表现为一个单一的系统，提供对系统资源的统一访问
- 作业管理（Job Management）：利用需要获得较高的系统使用率，利用上的作业管理软件需要提供批处理、负载平衡、并行处理等功能
- 并行文件系统（Parallel File System）：利用上的许多并行应用要处理大量数据，需进行大量的I/O操作，必须要有一个高性能的并行文件系统支持。
- 高效通信：松耦合、较长的链接、延迟大、标准协议开销、低级协议

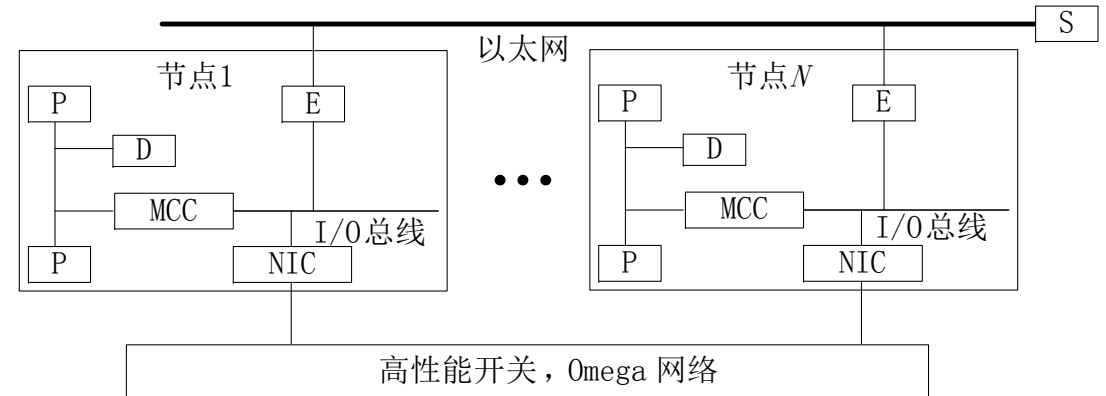
集群型大规模并行机SP2

- 设计策略:

- 集群体系结构
- 标准环境
- 标准编程模型
- 系统可用性
- 精选的单一系统映像

- 系统结构:

- 由完全独立的IBM RS6000 workstations组建
- 高性能开关 HPS 多级 Ω 网络
- 宽节点、窄节点和窄节点2



Beowulf集群

- 通过普通计算机硬件构件的并行结构环境
- 节点：普通PC或服务器
- 1994年夏季，Thomas Sterling和Don Becker用16个节点和以太网组成了一个计算机集群系统，称为Beowulf集群
- Beowulf集群提供了一种使用COTS（Commodity off the shelf）硬件构造集群系统以满足特殊的计算需求的方法。COTS是指象PC和以太网这种广为应用的标准设备，它们通常可以由多家厂商提供，所以通常有很高的性价比。



**32 processor Xeon
Beowulf cluster**

Top500中的Linux集群

Name/Org	CPUs	Interconnect	Cores	Rmax (TFLOPS)	Rpeak (TFLOPS)
#2 Dawning Cluster, China	TC3600 Blade System, Intel X5650, NVidia Tesla C2050 GPU	Infiniband	120640	1271	2984.3
#3: IBM Roadrunner, USA	PowerXCell 8i 3.2GHZ/ Opteron DC 1.8GHZ	Infiniband	129600	1105.00	1456.70
#7: NUDT TH-1 Cluster, China	Xeon E5540/E5450, ATI Radeon HD 4870	Infiniband	71680	563.1	1206.19
#10: SUN Blade System, USA	Sun Blade x6275, Xeon X55xx 2.93 Ghz	Infiniband	42440	433.50	497.396

2010年6月

实例分析：Roadrunner



- **1.6 PetaFlop 峰值浮点计算 (Peak DP Floating point)**
- **360 个服务器机架，约 12,000平方英尺 (三个足球场大)**
- **混合系统结构： Opteron X64 AMD processors (System x3755 servers) and Cell BE (Cell Broadband Engine) 刀片服务器 (Blade Servers) 通过高速网络互联**
- **互联网络： InfiniBand**

SMP、MPP、集群的比较

系统特征	SMP	MPP	集群
节点数量(N)	$\leq O(10)$	$O(100)-O(1000)$	$\leq O(100)$
节点复杂度	中粒度或细粒度	细粒度或中粒度	中粒度或粗粒度
节点间通信	共享存储器	消息传递 或共享变量（有DSM时）	消息传递
节点操作系统	1	N(微内核) 和1个主机OS(单一)	N(希望为同构)
支持单一系统映像	永远	部分	希望
地址空间	单一	多或单一（有DSM时）	多个
作业调度	单一运行队列	主机上单一运行队列	协作多队列
网络协议	非标准	非标准	标准或非标准
可用性	通常较低	低到中	高可用或容错
性能/价格比	一般	一般	高
互连网络	总线/交叉开关	定制	商用

课程小结

- Parallel Computer System Architectures
 - PVP : Parallel Vector Processors
 - SMP : Symmetric Multiprocessors
 - MPP : Massively Parallel Processors
 - DSM : Distributed Shared Memory
 - COW : Cluster Of Workstations
- Parallel Computer Memory Access Models
 - UMA : Uniform Memory Access
 - NUMA : Non-Uniform Memory Access
 - COMA : Cache-Only Memory Access
 - NORMA : NO-Remote Memory Access
- 典型并行计算机系统
 - Origin 2000
 - Cray T3E, ASIC MPP, BG/L
 - SP2, Beowulf Cluster, Roadrunner

推荐网站和读物

- 《并行计算—结构、算法、编程》
 - 第1章：并行计算机系统及其结构模型
 - 第2章：当代并行计算机系统介绍
- 可扩展并行计算（“Scalable Parallel Computing”）
 - 第4章：微处理器构件
 - 第5章：分布式存储器和时延容忍
 - 第6章：系统互连和千兆位网络
 - 第8章：对称多处理器和CC-NUMA多处理机
 - 第10章：服务器和工作站群
 - 第11章：MPP的体系结构和性能
- 陈国良等，《并行计算机体系结构》，高等教育出版社，2002
- The BlueGene/L Team, “An Overview of the BlueGene/L Supercomputer”, Supercomputing 2002 Technical Papers

下一讲

- 并行计算性能评测
 - 《并行计算—结构、算法、编程》第3章