# 第7章 多处理机与多计算机

郑宏 副教授 计算机学院 北京理工大学



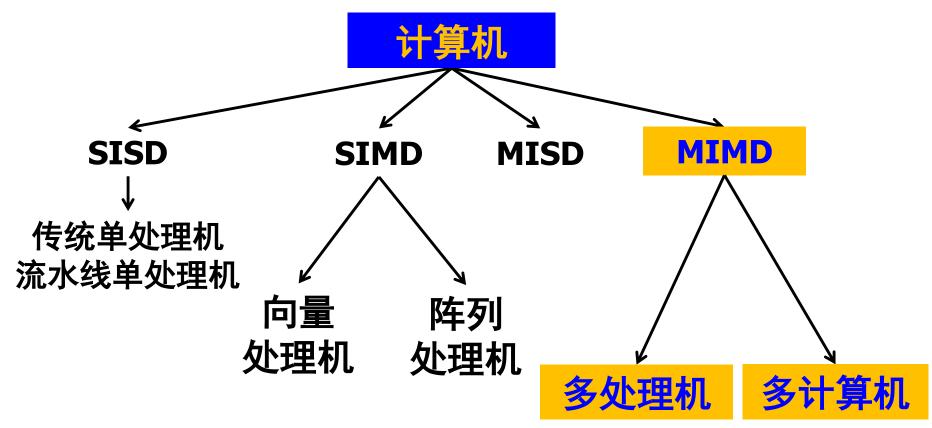
# 学习内容

- **■7.1** 多处理机概念
- ■7.2 多处理机结构
- **■7.3** 多核处理器
- ■7.4 多处理机的多Cache一致性
- ■7.5 多处理机的机间互连形式
- 7.6 程序并行性
- **■7.6** 多处理机性能
- ■7.7 多处理机操作系统



# 7.1 多处理机概念

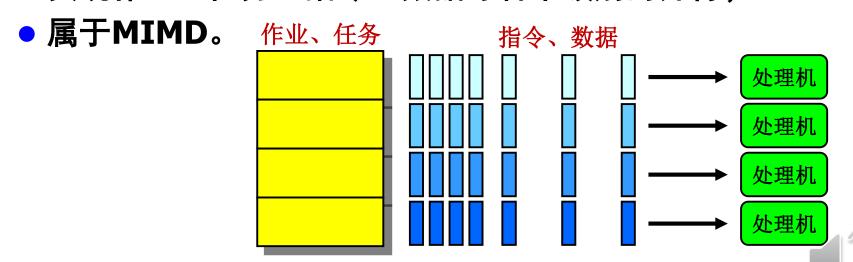
■弗林分类法:并行处理能力



# 7.1.1 多处理机定义

#### ■ 多处理机:

- 由两台及以上处理机组成的计算机系统;
- 各处理机有自己的控制部件、局部存储器,能执行各自的程序,可以共享公共主存和所有外设;
- 各处理机通过某种形式互连,相互通信;
- 实现作业、任务、指令、数据等各个级别的并行;



# 7.1.1 多处理机定义

- ■多处理:
  - 在多个处理机上运行同一道程序或作业的不同任务。可以串行,也可以并行。
- 并行处理: 并行计算, 高性能计算
  - 同时执行两个或更多个处理的一种计算方法。

 大任务
 分解
 多个子任务

 小型
 小同处理

 快速求解
 不同处理单元

5

# 7.1.1 多处理机定义

- ■多处理机优点:
  - •提高性能;
  - ●提高可靠性;
  - •减少及其功耗;
  - •提高效费比;
  - ●等。



- ■多种分类方法
- ■根据实现并行技术途径,分为3类:
  - 同构型: 基于资源重复
  - •异构型:基于时间重叠
  - •分布式:基于资源共享



#### 同构型

- 基于资源重复
- 处理机类型/功 能相同
- 并行处理或执 行任务

#### 异构型

- 基于时间重叠
- 处理机类型/功 能不同
- 重叠处理各任 务

#### 分布式

- 基于资源共享
- 处理机类型/功能相同或不同
- 并行/协同完成 任务处理
- 由统一操作系统统一管理资源

- ■耦合度:
  - 反映多处理机之间物理连接的紧密程度和交叉作用的能力的强弱
- ■根据耦合度,分为2类:
  - 紧耦合多处理机
  - 松耦合多处理机



- 紧耦合多处理机
  - 直接耦合系统
  - 通过公共硬件(如存储器, I/O系统等)通信
  - 典型: SMP, 多核处理器等
  - 一般将紧耦合多处理机称为多处理机
- 松耦合多处理机
  - 间接耦合系统
  - 通过通道、通信线路或网络、消息传递系统通信
  - 典型: 机群, 计算机网络等
  - 一般将松耦合多处理机称为多计算机



### 7.1.3 多处理机特点和主要技术问题

- ■多处理机特点:
  - 结构灵活:满足多种并行计算的不同需求。
  - •程序并行:实现作业、任务之间的并行。
  - 并行任务派生:可并行执行任务的识别、派生与 分配。
  - 进程同步:解决数据相关和控制依赖问题。
  - 资源调度与分配:动态分配资源和调度任务,以 获得更好性能和更高效率。

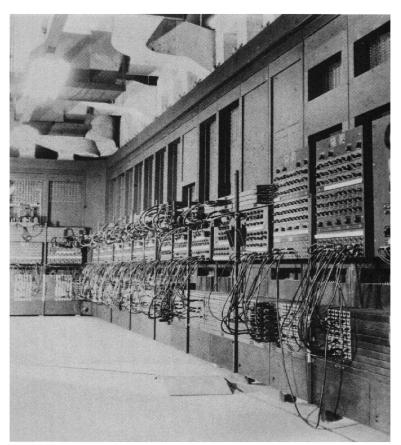


### 7.1.3 多处理机特点和主要技术问题

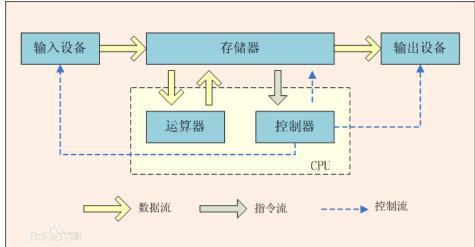
- 多处理机主要技术问题:
  - 结构灵活性与通用性:以适应、满足多种并行计算算法不同的需求
  - 进程通信方法: 共享内存、互连网络、消息传递机制等
  - 运行模型: 数据并行 + 处理并行
    - ◆ 基于共享内存的运行模型
    - ◆ 基于消息传递机制的运行模型
  - 并行性表达:
    - ◆ 编译程序自动发现:如 并行C或Fortran
    - ◆ 并行程序设计语言: 如 HPF(High Performance Fortran)
    - ◆ 运行时库:如 LAPACK
  - 并行算法

12

■ 1946 年,世界上第一台计算机ENIAC 诞生



用途: 弹道计算和氢弹研制





约翰·冯·诺伊曼 John von Neumann 1903 - 1957

匈牙利人、美国人 数学、计算机科学、 经济学

■ 1972年,第一台并行计算机ILLIAC IV(美国伊利诺依大学)





- 60 年代末开始建造,72 年建成
- 74 年运行第一个完整程序,76年运行第一个应用程序
- 64 个处理器, SIMD, 是当时性能最高的 CDC7600 机器的2-6倍
- 公认的1981年前最快,1982年退役
- 可扩展性好,但可编程性差

14

■ 1976 年,超级计算元年: Cray-1 向量机







1925-1996 电子工程学学士 应用数学硕士 超级计算之父 Cray研究公司的 创始人

- 一般将Cray-1 投入运行的1976 年称为 "超级计算元年"
- 编程方便,但可扩展性差
- 以Cray 为代表的向量机称雄超级计算机界十几载

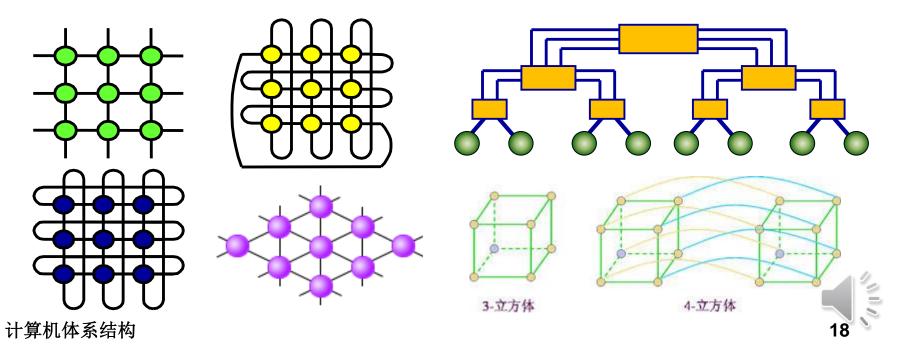
15

- 80 年代初期:以MIMD 并行计算机的研制为主
  - Denelcor HEP (1982年): 第一台商用MIMD 并行计算机
  - IBM 3090: 80 年代普遍为银行所采用
  - Cray X-MP: 1982, Cray研究公司发布了第一台MIMD 并行计算机。



- 80 年代初期: 共享存储器多处理机(Shared-Memory Multi-Processor)
  - SMP (Symmetrical Multi-Processing): 在一个计算机 上汇集一组处理器,各处理器对称共享内存及计算机的其 他资源,由单一操作系统管理,极大提高整个系统的数据 处理能力

- 80 年代后期: 具有强大计算能力的并行机
  - 通过二维Mesh连接的Meiko(Sun)系统
  - 超立方体连接的MIMD 并行机: nCUBE-2(27Gflops)、
     Intel iPSC/80(7Gflops)
  - 共享存储向量多处理机Cray Y-MP 等



- 90 年代初期: MIMD 类型占据绝对主导位置
  - MPP (Massively Parallel Processing): 大规模并行处理结构
    - ◆ 每个结点相对独立,有一个或多个微处理器
    - ◆ 每个结点有自己的操作系统和独立内存,避免内存访问瓶颈
    - ◆ 各个结点只能访问自己的内存模块
    - ◆ 通过高速互联网络构成
    - ◆ 扩展性较好



#### First computer to defeat a world champion!!



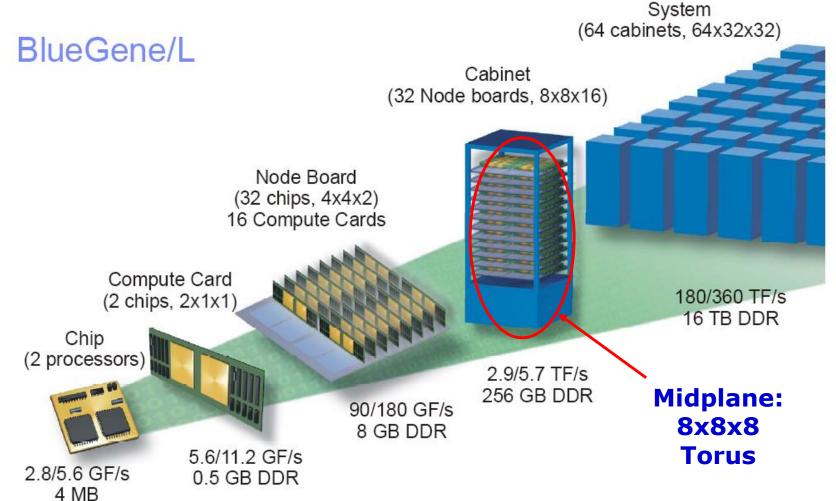
**Garry Kasparov** 



**Deep Blue** 

In February 1996, IBM's Deep Blue defeated grandmaster Garry Kasparov. It was then assigned to predict the weather in Atlanta, Georgia, during the 1996 Summer Olympic Games

# Blue Gene/L: An Example





- 90 年代初期: MIMD 类型占据绝对主导位置
  - 分布式存储 MPP 并行机主要有:
    - ◆ Intel Touchstone Delta: 含512 个i860 微处理芯片,二维 Mesh连接,峰值性能为32Gflops,8GB 内存。
    - ◆ CM-5E: 含16-1K 个标量RISC SPARC 处理器(含四个向量部件,峰值性能128Mflops,32MB 内存),胖树数据网、二叉控制网及四叉诊断网连接各个处理器。
    - ◆ CRAY T3D: 16-1K 个结点,每结点含2 个处理器(64 位 RISCDECchip 21064,峰值150Mflops),局部内存64MB, 最大存储规模128GB,结点间双向三维Torus 连接。

- 90 年代初期: MIMD 类型占据绝对主导位置。
  - 共享存储并行机:
    - ◆ CRAY Y-MP C90: 由16台向量机构成的并行机 (峰值性能 16Gflops)
    - ◆ 国产YH-2 (4 台向量机,峰值性能1Gflops)
  - 分布共享存储并行机:
    - ◆ Kendall Square KSR-1: 1991,它提供给用户透明的共享存储结构,每个环含32个结点,多个环以层次结构相互连接,可扩展到1024个结点,峰值性能为15Gflops。



- 90 年代中后期: 高性能微处理器+高性能网络。体系结构框架趋于统一 (DSM、SMP、NOW)
  - DSM (Distributed Shared Memory): 分布式共享存储
    - ◆ 以结点为单位,每个结点有一个或多个CPU
    - ◆ 专用的高性能互联网络连接(Myrinet, Infiniband, ...)
    - ◆ 分布式存储:内存模块局部在每个结点中
    - ◆ 单一的操作系统,单一的内存地址空间
    - ◆ 可扩展到上百个结点,支持消息传递、共享存储并行程序设计



- 90 年代中后期:高性能微处理器+高性能网络。 体系结构框架趋于统一(DSM、SMP、NOW)
  - DSM (Distributed Shared Memory): 分布式共享存储
    - ◆ NUMA 结构:在原来分布式存储并行机的基础上,增加局部内存的全局共享功能,提供共享存储并行编程环境。代表:典型代表为Cray T3E 和Cray T3E-1200,最多可扩展到2048个CPU,采用了当时最先进的微处理芯片(主频600MHz),峰值性能达到2.5Tflops。
    - ◆ CC-NUMA结构: CC-NUMA 结构具有比SMP 更好的可扩展性,并且能保持SMP 的共享存储特性。代表: SGI Origin 系列超级服务器

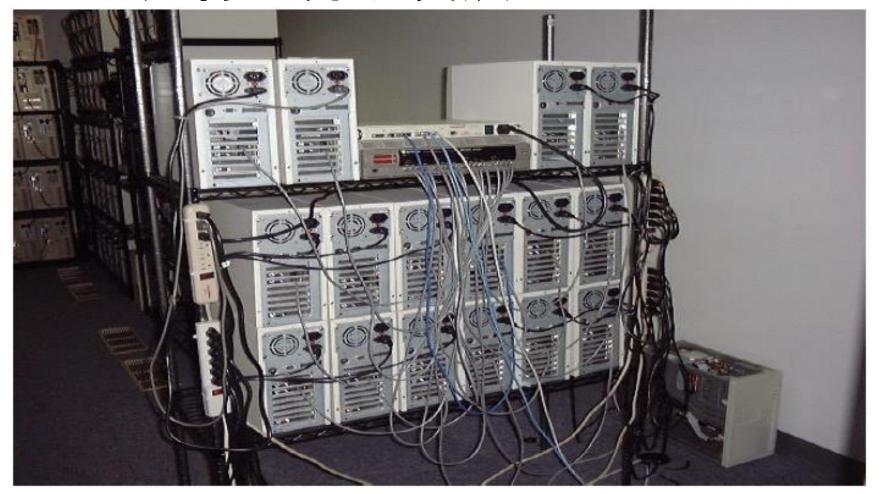
25

- 90 年代中后期: 高性能微处理器+高性能网络。体系结构框架趋于统一 (DSM、SMP、NOW)
  - 以高性能微处理芯片和互连网络通信技术为基础,共享存储对称多处理机(SMP)系统得到了迅速发展
    - SUN Ultra E10000
    - ♦ SGI Power Challenge R10000
    - ◆ HP C-240
    - DEC Alphaserver 400C



- 90 年代中后期: 高性能微处理器+高性能网络。体系结构框架趋于统一 (DSM、SMP、NOW)
  - NOW (Network of Workstations) 工作站机群
    - ◆ 每个结点都是一个完整的工作站,有独立的硬盘与UNIX系统
    - ◆ 结点间通过低成本的网络(如千兆以太网)连接
    - ◆ 每个结点安装消息传递并行程序设计软件,实现通信、负载平衡等。被大量中小型计算用户和科研院校所采用
  - 也称为COW(Cluster of Workstations)
  - NOW(COW)与MPP之间的界线越来越模糊
  - 1994年夏, Thomas Sterling和Don Becker用16台PC 和以太网组装了一个计算机集群系统,命名为 "Beowulf"。

27



世界上第一台Beowulf 集群,1994年

28

- 90 年代中后期: 高性能微处理器+高性能网络。体系结构框架趋于统一 (DSM、SMP、NOW)
  - 第三代Beowulf集群Loki和Hyglac分别在LANL和加州理工学院建成,两者均采用了16个Pentium Pro处理器和100Mb/s的快速以太网,不同之处在于使用了不同的网络结构。
  - 两台机器被联合起来用于求解NBody问题,系统 总性能达到2GFLOPS,获得了1997年的戈登● 贝尔性价比奖



52结点的贝奥 武夫机群Borg, 在麦吉尔大学 被用来搜寻脉 冲星双星系统。



- **2000**年之后: 迅猛发展
  - Cluster 机群/集群
    - ◆ 每个结点含多个商用处理器,结点内部共享存储
    - ◆ 采用商用机群交换机通过前端总线连接结点,结点分布存储
    - ◆ 各个结点采用Linux 操作系统、GNU编译系统和作业管理系统

#### Constellation 星群

- ◆ 每个结点是一台子并行机
- ◆ 采用商用机群交换机通过前端总线连接结点,结点分布存储
- ◆ 各个结点运行专用的结点操作系统、编译系统和作业管理系统

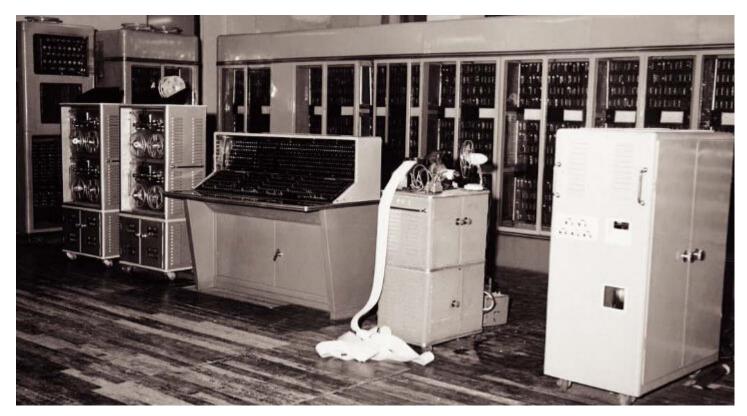
#### MPP

◆ 专用高性能网络,大多为政府直接支持

31

# 7.1.5 中国并行计算机发展

■ 1958 年第一台国产计算机诞生 — — 103型计算机



运行速度1500次/每秒



# 7.1.5 中国并行计算机发展

■我国的超级计算机系列

• 曙光:中科院计算所,曙光公司,上海超算中心

•深腾:联想集团

•银河:国防科大

• 神威: 国家并行计算机工程技术研究中心

•天河:国防科大



# 银河

■ 银河一号: 1983年12月, 我国第一台每秒运算达1 亿次以上的计算机





- 1992年,银河-II 问世,每秒运算达10亿次
- 1997年,银河-III 问世,每秒运算达130亿次

34

# 曙光

- 1993年10月,国家智能计算机研究开发中心(后成立北京市曙光计算机公司)研制成功"曙光一号"SMP多处理机
- 2000年推出每秒3000 亿次的曙光3000 超级服务器
- 2004年6月,推出11万亿次的曙光4000A超级计算机,排名全球第十
- 2008年6月,曙光5000A发布,实际速度超160万亿次,排名世界第十



曙光5000A



# 神威

■ 1999年9月,由国家并行计算机工程技术研究中心 牵头研制成功的"神威" 计算机系统投入运行,峰 值运算速度达每秒3840亿次。



神威采用国产CPU

# 深腾

- 2002年,联想发布深腾1800计算机,排名全球第43位,成为首家正式进入排行榜前100的中国企业
- 2003年,深腾6800计算机发布,列全世界TOP500第14位,其78.5%的整机效率列世界通用高端计算机第一名
- 2008年12月,联想发布百万亿次超级计算机深腾7000,排名全球19



实现了大规模异构体系结构、系统均衡设计方法、 商效能结点机、大规模无 高效能结点及千核级应用 技术突破。整机系统软件 实现对大规模异构资源的 统一管理和高效使用。

深腾7000



## 天河

- 2009年10,中国首台千万亿次超级计算机"天河一号"诞生,使中国成为继美国之后世界上第二个能够研制千万亿次超级计算机的国家。
- 2010年10月,升级后的天河一号排名世界第一
- 2013年6月,天河二号再次问鼎世界第一,功耗达24兆瓦,也是目前 TOP500里功耗最大的





# 学习内容

- **■7.1** 多处理机概念
- 7.2 多处理机结构
- **■7.3** 多核处理器
- ■7.4 多处理机的多Cache一致性
- ■7.5 多处理机的机间互连形式
- 7.6 程序并行性
- **■7.6** 多处理机性能
- ■7.7 多处理机操作系统



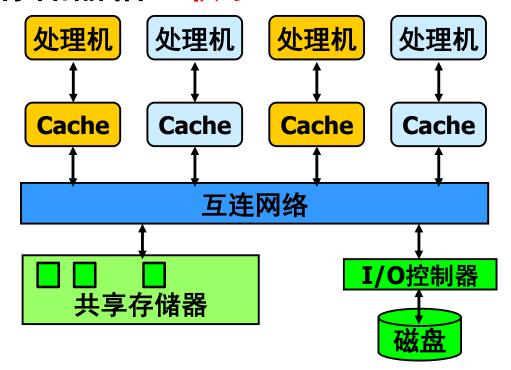
# 7.2 多处理机结构

- ■从存储器的分布和使用上看,多处理机 系统分为两种结构:
  - 共享存储器结构
  - 分布式存储器结构



### 7.2.1 共享存储器结构

■ 各处理机通过互连网络共享存储器和I/O设备,并通过共享存储器相互联系。



共享存储器结构的多处理机系统



### 7.2.1 共享存储器结构

- ■特点:
  - 各处理机共享存储器,并通过对共享存储器 读/写实现相互通信;
  - •对存储单元的任何修改对其他处理机都是可 见的:
  - ●延迟低,但扩展性差;

### 共享数据进入Cache产生了一个新的问题:

Cache的一致性问题

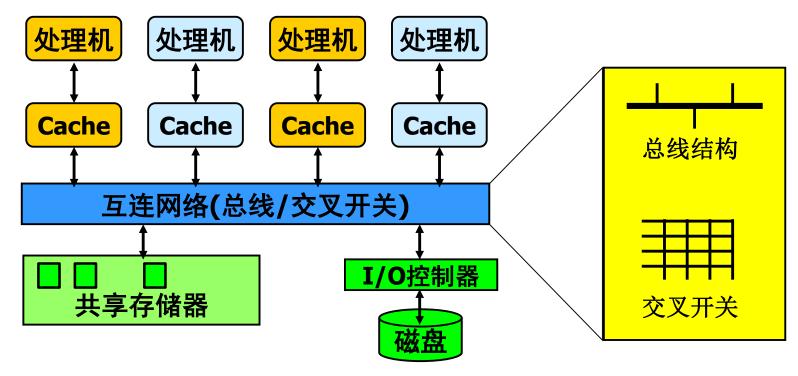
# 7.2.1 共享存储器结构

- ■根据共享存储器实现方式不同又细分为:
  - UMA结构
  - NUMA结构
    - ccNUMA
    - NCC-NUMA
  - COMA结构



- **UMA = Uniform Memory Access**
- ■均衡存储器访问结构
- 或 集中式共享存储器 (Centralized Shared-Memory) 结构
- ■特点:
  - 各处理机对存储器的访问时间、访问功能相同
- 这种结构的处理机称为对称多处理机 (SMP
  - = Symmetric Multiprocessors)





UMA结构的多处理机系统

- 互连网络可以是总线、交叉开关或多级交换网络。
- 大多数的对称多处理机采用总线连接。

#### ■ 优点:

- (1)性能提高
- (2) 高可用性
- (3) 增量式增长
- (4) 可扩展性好
- (5)透明

#### ■ 缺点:

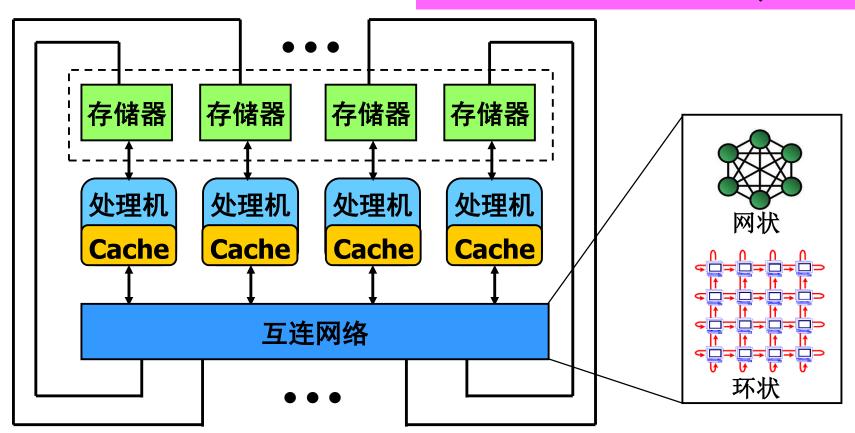
- 所有处理机对共享存储器的访问都要经过互连网络,当规模较大时,访问延迟较大。因此通常增设大容量本地 Cache
- 适合小规模时采用



- NUMA = Non Uniform Memory Access
- 非均衡存储器访问结构
- 也称为分布式共享存储器 (DSM=Distributed Shared-Memory) 结构
- 特点:
  - 分布于各个处理机的存储器被统一编址,可由所有处理机 共享;
  - 根据存储器位置的不同,各处理机对存储器的访问时间不相等。处理机访问本地存储器的速度较快,通过互连网络访问其他处理机上的远地存储器相对较慢。

47

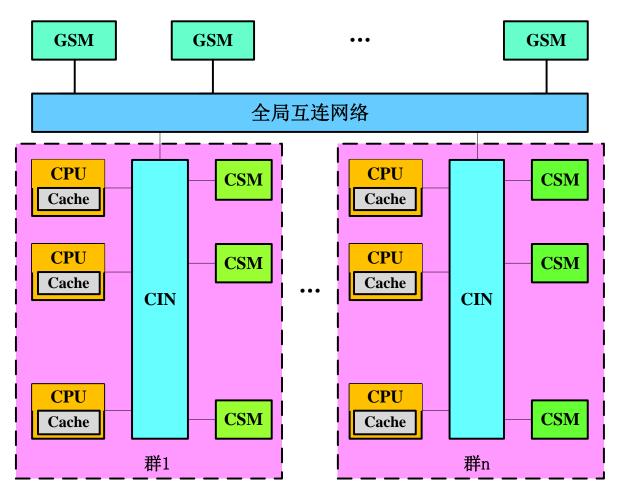
- ☑ 各处理机拥有自己的本地存储器, 可以独立工作;
- ☑ 各处理机借助互连网络、通过消息传递机制相互通信;



NUMA结构的多处理机系统 - 共享本地存储器



允许各处理机拥有自己独立结构,甚至可以是SMP。



NUMA结构的多处理机系统 - 层次式机群模型



### ■ 优点:

- •比SMP扩展性好,并行程度更高,性能更好;
- 采用与SMP相同的编程模型,为SMP编写的程序 仅需少量修改即可移植运行;
- 每个处理机都可以访问较大的存储空间,因此可以更高效地运行大程序;
- 实现数据共享时不需要移动数据;
- •传递包含指针的数据结构比较容易;
- 系统构建成本较低,利用成熟技术搭建系统。

### ■缺点:

- 如果过多地访问远程存储器,则性能会下降;
- 对存储器的访问不透明,需要处理分页(例如哪个页面在哪个存储器中)、进程分配等,对软件设计要求较高

### 3. ccNUMA

- ccNUMA = Cache-coherent Non Uniform Memory Access
- ■高速缓存一致性非均匀存储访问
- ■在NUMA多处理机中,逻辑上共享的存储器在物理上是分布的。如果各处理机 Cache内容一致,则将这种NUMA称为 ccNUMA

### 3. ccNUMA

- ccNUMA注重开拓数据的局部性和增强系统的可扩展性。在实际应用中,大多数的数据访问都可在本结点内完成,网络上传输的主要是高速缓存无效性信息而不是数据。
- ccNUMA和COMA的共同特点是它们都对高速缓存一致性提供硬件支持,而在NCC-NUMA (Non-Cache Coherent Non-Uniform Memory Access) 中,则没有对高速缓存的一致性提供硬件支持。
- 绝大多数商用ccNUMA多处理机系统使用基于目录的高速缓存一致性协议。

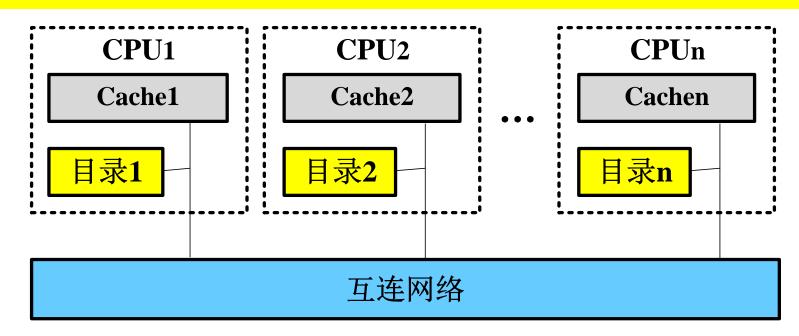
### 4. COMA

- COMA = Cache-Only Memory Architecture
- ■仅用高速缓存存储器结构
- NUMA的一个特例,只是将NUMA中的分布 存储器换成了Cache
  - 各处理机结点上没有主存储器,没有存储层次结构,仅有Cache
  - 所有的高速缓存构成了全局地址空间,全部 Cache组成了全局虚拟地址空间
  - 对远程Cache的访问通过分布式Cache目录进行

54

### 4. COMA

- 各处理机结点上没有主存储器,没有存储层次结构,仅有Cache
- 所有的高速缓存构成了全局地址空间,全部Cache组成了全局虚 拟地址空间
- 对远程Cache的访问通过分布式Cache目录进行



#### COMA多处理机结构

55

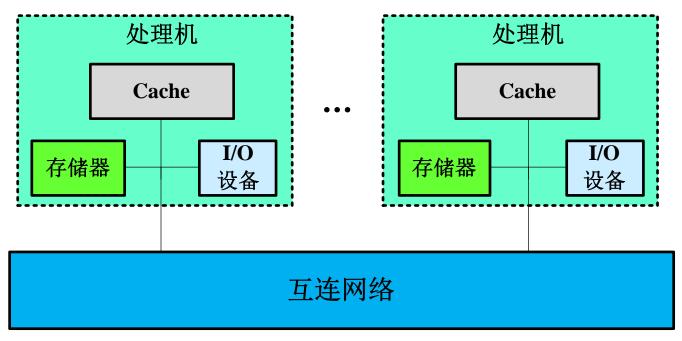
# 7.2.2 分布式存储器结构

- 也被称为非远程存储访问 (NORMA, No-Remote Memory Access) 模型
- 各处理机拥有自己的本地存储器,在本地操作系统 控制下独立工作。
- 各处理机的本地存储器是私有的,不能被其他处理机访问。
- 各处理机借助互连网络、通过消息传递机制相互通信,实现数据共享。
- 大规模并行处理机(MPP)、机群(cluster)等采 用了这种结构。

56

## 7.2.2 分布式存储器结构

●各处理机借助互连网络、通过消息传递机制相互通信, 实现数据共享



分布式存储器多处理机

## 7.2.2 分布式存储器结构

### ■优点:

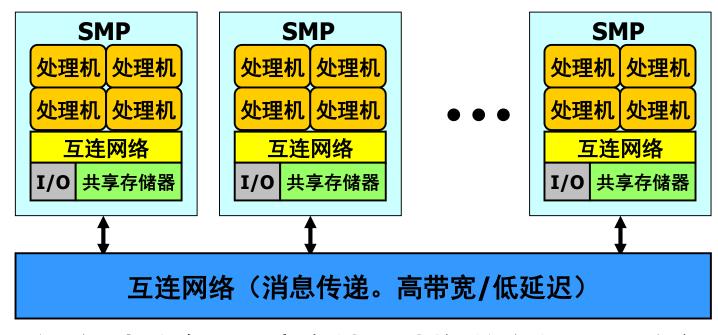
•结构灵活,扩展性较好

### ■缺点:

- 任务传输以及任务分配算法复杂,通常要设计专有算法
- 处理机之间的访问延迟较大
- ●需要高带宽的互连



## 分布-共享存储器结构



混合型 分布—共享存储器结构 的多处理机系统

- ☑ 每个节点为共享存储器结构的SMP系统;
- ☑ 各SMP系统借助互连网络、通过消息传递机制相互通信;
- ☑ 是目前大规模并行处理(MPP)等系统普遍采用的结构。

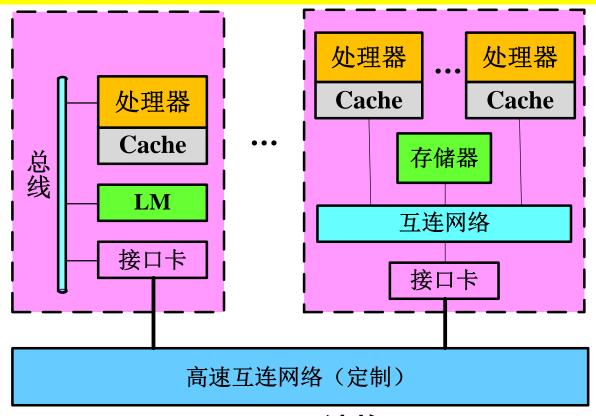
## 7.2.3 大规模并行处理机

- MPP = Massively Parallel Processor
- 由几百或几千台高性能、低成本处理机组成的 大规模并行计算机系统
  - 每个处理机都有自己的私有资源,如内存、网络接口等
  - 每个处理机能直接访问的只有本地存储器,但不能直接访问其它处理机的存储器
  - 处理机之间以定制的高带宽、低延迟的高速互连 网络互连



## 7.2.3 大规模并行处理机

● MPP系统大多采用分布式存储结构。所有存储器在物理上是分布的,而且都是私有的。每个处理机能直接访问的只有本地存储器,不能直接访问其它处理机的存储器。



MPP结构



## 7.2.3 大规模并行处理机

### ■特点:

- 处理结点采用商用处理机
- 系统中有物理上的分布式存储器
- 采用高通信带宽和低延迟的互连网络(专门设计和定制的)
- 能扩放至成百上千乃至上万个处理机
- 它是一种异步的MIMD机器
  - ◆程序系由多个进程组成,每个都有其私有地址空间, 进程间采用传递消息相互作用



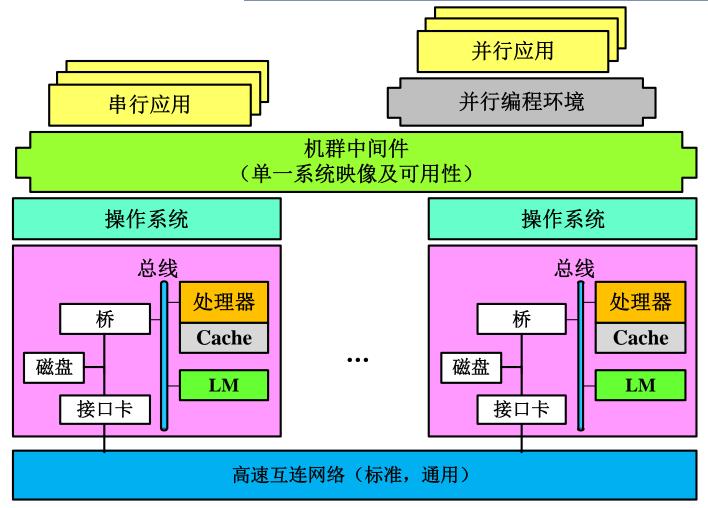
### 7.2.4 机群

- ■也被称为集群
- COW = Cluster of Workstations
- NOW = Network of Workstations
- 通过一组松散耦合的计算机软件和/或硬件连接起来、 高度紧密地协作完成计算工作的计算机系统
  - 结点: 单独运行的商品化计算机
  - 网络: 高速通用网络, 如局域网或互连网络连接
  - •操作系统:并行编程环境,系统资源管理和相互协作
  - 集群中间件:屏蔽差异,在异构系统上提供统一运行环境和服务



### 7.2.4 机群

从结构和结点之间的通信方式来看,机 群属于非均匀存储访问的MIMD型的分 布式存储并行计算机



机群系统结构

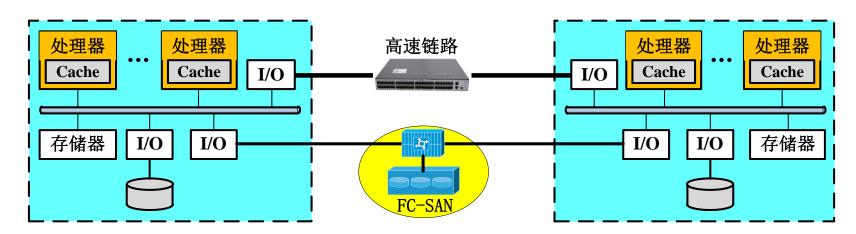


## 结点间连接模型

### 两种连接模型:无共享(零共享)模型和共享磁盘模型



无共享(零共享)模型



共享磁盘模型



## 机群分类

高可用性机群 High-Availability Clusters

关键任务应用

提供冗余

避免单点故障

负载均衡机群 Load Balancing Clusters

> 将负载分发到 可用结点

减轻处理负载

获得高性能和 高可用性 高性能计算机群 High-Performance Clusters

将计算任务分配到 不同计算结点

提高计算能力

高吞吐计算

分布式计算

**Beowulf** 

# 机群特点

- ■每一个结点都是一个完整的工作站。一个结点 也可以是一台PC或SMP;
- 各结点通过某种低成本的商品(标准)网络互 连;
- ■各结点内总是有本地磁盘;
- 结点内的网络接口松散耦合到I/O总线上;
- ■每个结点中驻留有完整的操作系统



# 机群关键技术

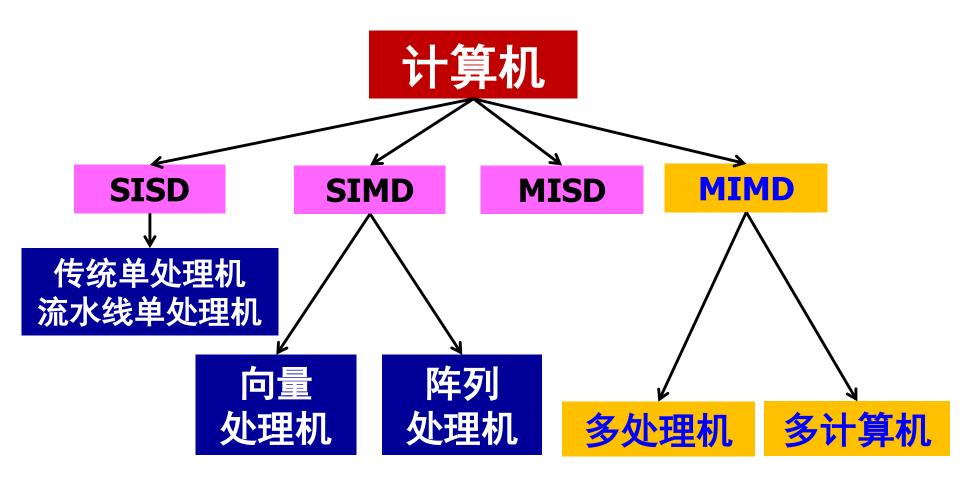
- ■高效的通信系统;
  - 如交换式千兆位以太网,万兆位以太网
- ■并行程序设计环境;
  - o如 PVM, MPI, OpenMP, HPF, Express等
- ■并行程序设计语言
  - ●如 FORTRAN、C和C++
- 全局资源管理与利用
  - •操作系统,COW管理,作业管理等



### PVM、SMP、DSM、MPP和COW的比较

属性	PVP	SMP	DSM	MPP	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 (以太网、 ATM)
通信机制	共享变量	共享变量	共享变量	消息传递	消息传递
地址空间	单地址空间	单地址空间	单地址空间	多地址空间	多地址空间
系统存储 器	集中共享	集中共享	分布共享	分布非共享	分布非共享
访存模型	UMA	UMA	NUMA	NORMA	NORMA
代表机器	Cray C-90 Cray T-90 银河1号	IBM R50 SGI Power Challenge 曙光1号	Stanford DASH Cray T 3D	Inter Paragon IBM Option White 曙光- 1000/2000	Berkeley NOW Alpha Farm IBM SP2 曙光 3000/4000

## 弗林分类法一并行体系结构





## MIMD计算机分类

