高性能计算与云计算

第二讲 并行计算体系结构（1）

何克晶
kejinghe@gmail.com
华南理工大学
计算机科学与工程学院

華南理工大學
South China University of Technology

目录

- Flynn分类
- 并行计算机的内存体系
- 并行计算机的处理器

并行计算的分类

- 指令流/数据流分类法，即费林-Flynn分类法。
  - （1）指令流（Instruction Stream）：机器执行的指令序列。
  - （2）数据流（Data Stream）：指令调用的数据序列，包括输入数据和中间结果。
  - （3）多倍性（Multiplicity）：在系统性能瓶颈部件上同时处于同一执行阶段的指令或数据的最大可能个数

Flynn分类

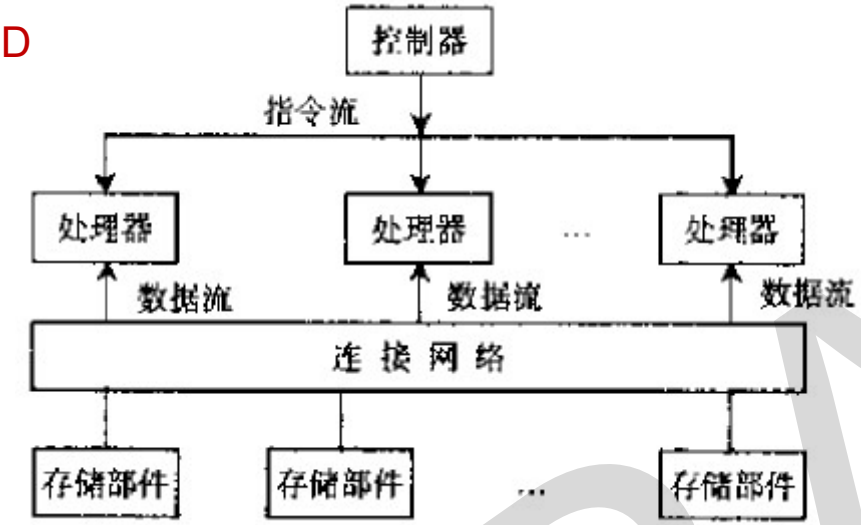
- 两个独立的维度——指令流和数据流的不同组织方式，将计算机系统分为四类：
  - 单指令单数据流（Single Instruction stream and Single Data stream , SISD）系统
  - 单指令多数据流（Single Instruction stream and Multiple-Data stream , SIMD）系统
  - 多指令单数据流（Multiple-Instruction stream and Single Data stream , MISD）系统
  - 多指令多数据流（Multiple-Instruction stream and Multiple-Data stream , MIMD）系统

Flynn分类

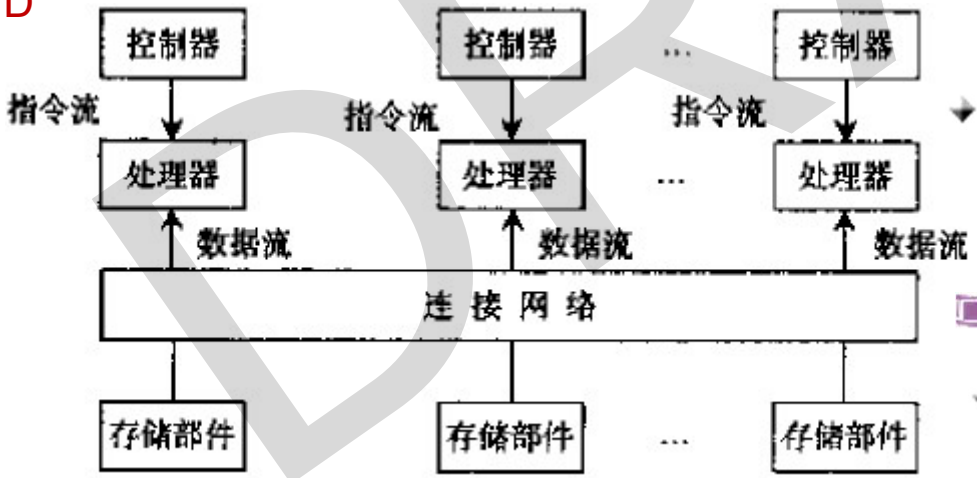
- SISD：SISD系统是一种传统的顺序执行的单处理器计算机，它的硬件不支持任何形式的并行计算，所有的指令都是串行执行。在任何一个时钟周期内，CPU只能处理一个数据流，因此这种机器被称作单指令流单数据流机器
- SIMD：SIMD系统有多个处理单元，由单一的指令部件控制，按照同一指令流的要求为它们分配各不相同的数据流并进行处理。系统结构由一个控制器、多个处理器、多个存贮模块和一个互连总线（网络）组成
- MISD：MISD系统有多个处理单元，每个处理单元按照多条不同的指令要求同时对同一数据流及其处理输出的结果进行不同的处理，把一个单元的输出作为另一个单元的输入
- MIMD ：MIMD系统又称为多处理机系统，是指能实现指令、数据作业、任务等各级全面并行计算的多机处理系统，可以将一个主任务分解为众多子任务并行执行以缩短工作时间

SIMD vs MIMD

SIMD 计算机系统结构



MIMD 计算机系统结构



	SIMD	MIMD
控制器	一个	多个
处理器	多个	多个
存储部件	多个	多个
连接网络	有	有

区别1:

- SIMD的各个处理器同步运行，即分别在来自存储器的不同数据流上并行执行相同的指令流，也就是说有一个指令流和多个数据流。
- MIMD的各个处理器异步运行，即在各自的数据流上执行自己的指令流，也就是说有多个指令流和多个数据流。

区别2:

- SIMD的各个处理器同步使用连接网络，而MIMD的异步使用连接网络。

并行计算机基本属于MIMD的范畴

目录

- Flynn分类
- 并行计算机的内存体系
- 并行计算机的处理器

并行计算机体系结构

组成要素

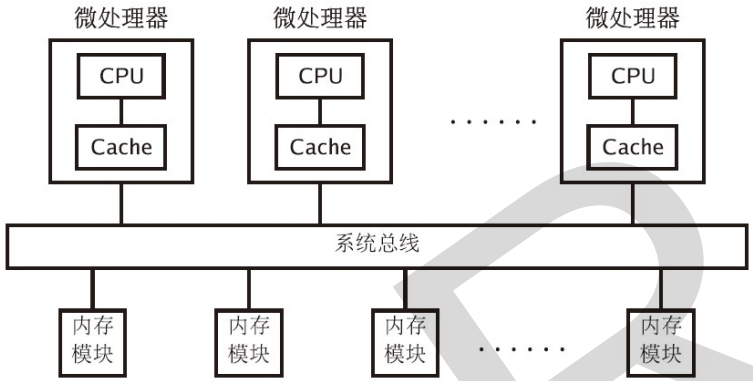
- 处理器（processor）：计算单元
- 互联网络（interconnetct network）：连接
- 内存（memory）：多个存储模块组成

并行机的基本特征是具备多个计算单元和存储模块，各个模块通过互联网络耦合

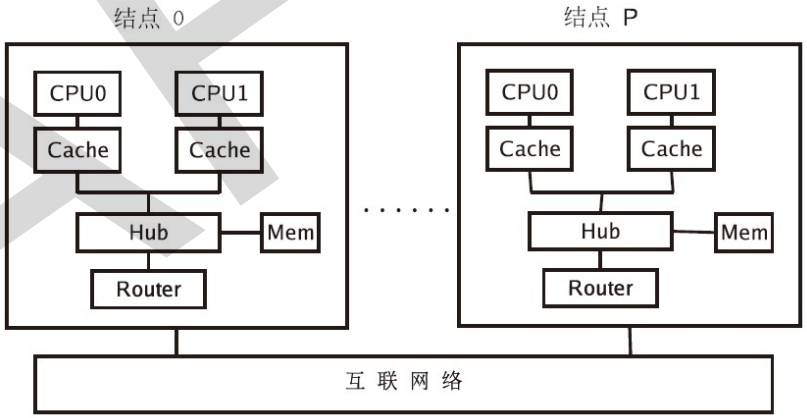
- 根据耦合的紧密程度可分为紧耦合和松耦合。
- 不同的并行计算机，其各模块耦合的松紧程度可以有区别

共享内存 vs 分布式内存

- 共享内存的多处理机（multiprocessors）体系



- 分布式内存的多计算机（multicomputers）体系



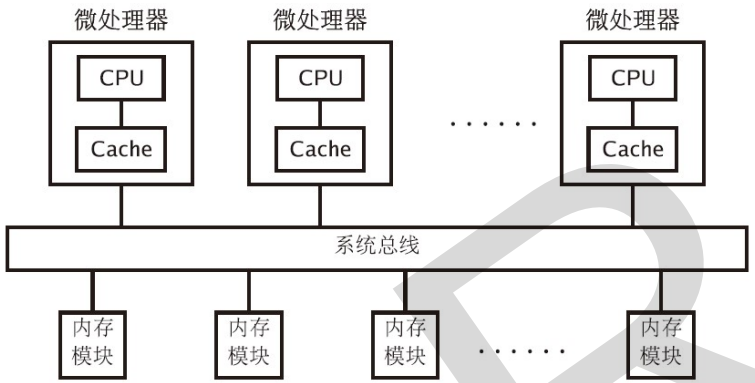
- 欠可靠、可扩展性较差。

并行计算机访存模型

- UMA（Uniform Memory Access）模型：均匀存储访问模型。
- NUMA（Non-Uniform Memory Access）模型：非均匀存储访问模型。
- COMA(Cache-Only Memory Access)模型：全高速缓存存储访问。
- CC-NUMA（Coherent-Cache Nonuniform Memory Access）模型：高速缓存一致性非均匀存储访问模型。
- NORMA（No-Remote Memory Access）模型：非远程存储访问模型。

共享内存

- 共享内存

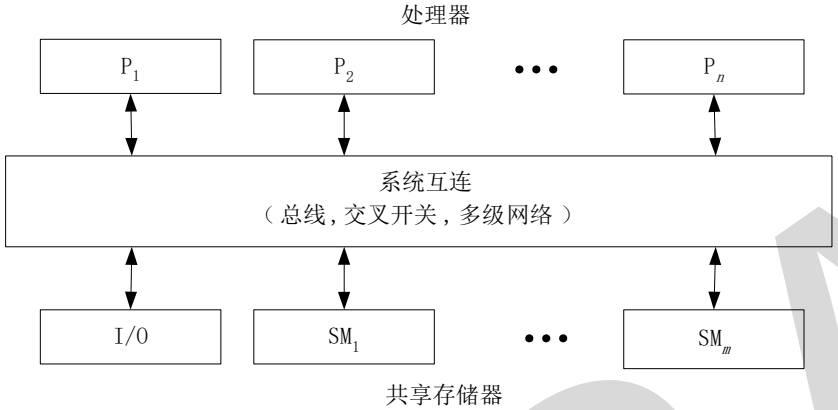


- 通常也称为紧密耦合多处理机，它具有一个所有处理器都可以访问的全局物理内存。共享内存系统的特性有：
- （1）对称性：系统中任何处理器都可以访问任何的内存单元和I/O设备
- （2）单地址空间：内存中每一个位置在整个的内存地址范围内有一个唯一的地址
- （3）低通信延迟：处理器间的通信可以利用共享内存来进行数据交换
- （4）高速缓存及其一致性：多级高速缓存可以支持数据的局部性，而其一致性可由硬件来增强

- 欠可靠、可扩展性较差。

UMA（Uniform Memory Access）

- 均匀存储访问（UMA）



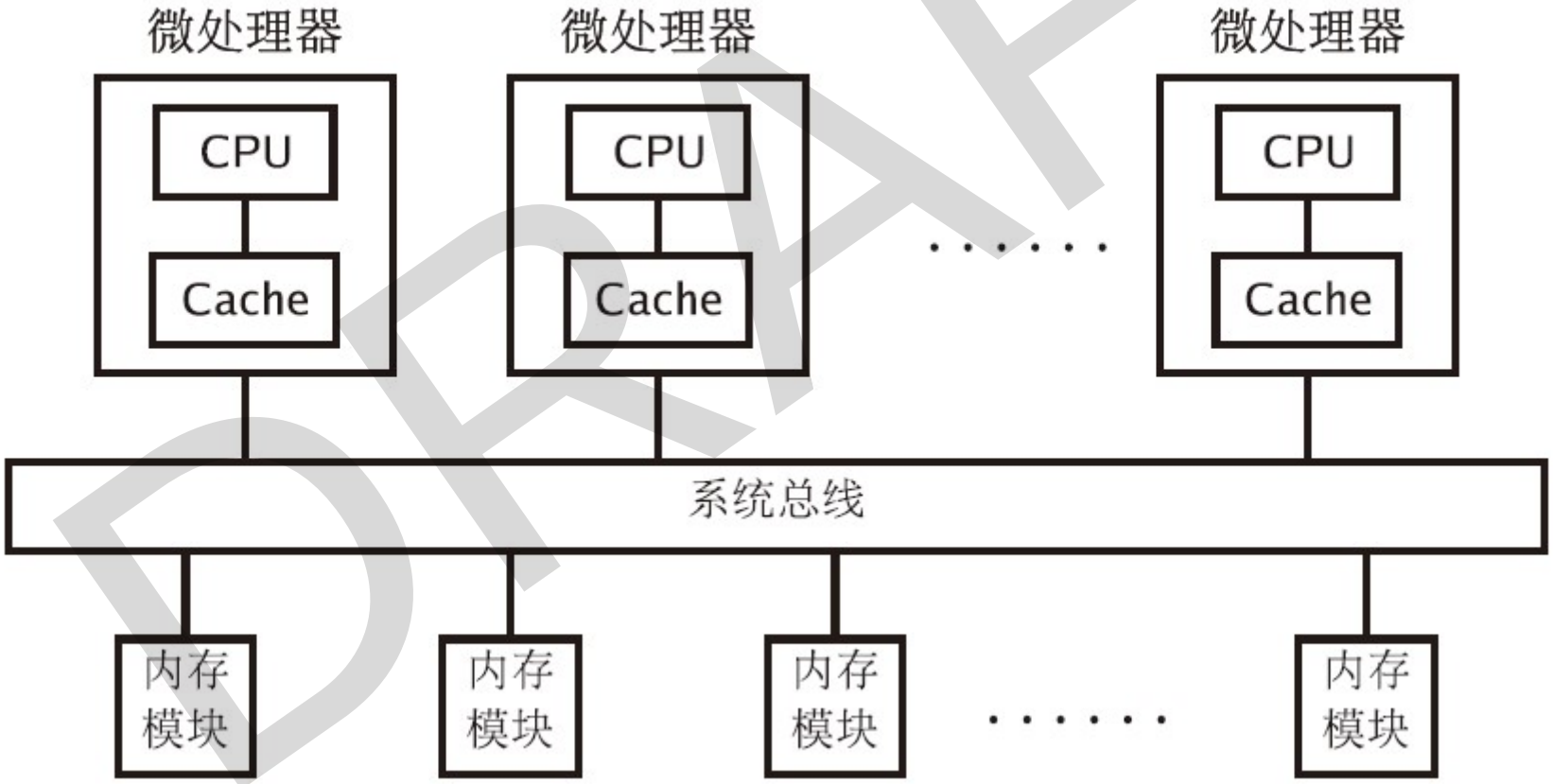
  - 物理存储器被所有处理器均匀共享；
  - 所有处理器访问任何存储字取相同的时间；
  - 每台处理器可带私有高速缓存；
  - 外围设备(I/O)也可以一定形式共享。
  - 发生访存竞争时，仲裁策略平等对待每个结点，即每个结点机会均等；

- 对称多处理（Symmetric Multiprocessing , SMP）
  - SMP使用的是微处理器和高速缓存

- 并行向量处理机（Parallel Vector Processoe, PVP）
  - PVP使用的是高性能专门设计定制的向量处理器（Vector Processor, VP

SMP (Symmetric Multiprocessing)

- 内存模块和处理器对称地分布在互联网络的两侧;
- 内存访问属典型的均匀访问模型。



SMP的特点

- 对称共享存储
  - 系统中任何处理器均可直接访问任何存储模块中的存储单元和I/O 模块，且访问的延迟、带宽和访问成功的概率是一致的。所有内存单元统一编址。各个处理器之间的地位等价，不存在任何特权处理器。操作系统可在任意处理器上运行。
- 单一的操作系统映像
  - 全系统只有一个操作系统驻留在共享存储器中，它根据各个处理器的负载情况，动态地分配各个进程到各个处理器，并保持各处理器间的负载平衡。
- 局部高速缓存cache 及其数据一致性
  - 每个处理器均配备局部cache，它们可以拥有独立的局部数据，但是这些数据必须与存储器中的数据保持一致。

SMP的特点（2）

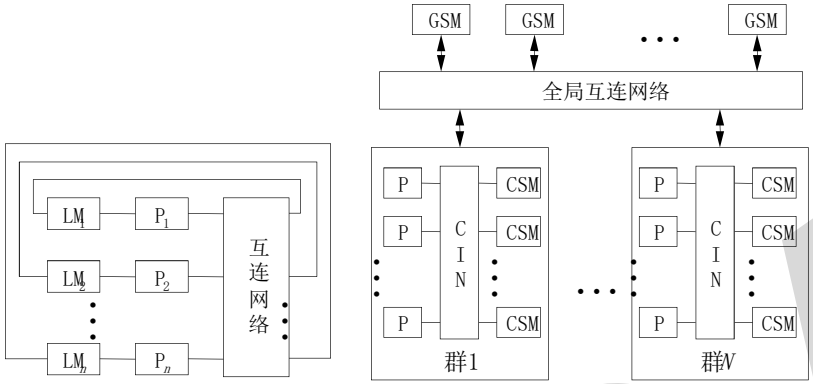
- 低通信延迟
  - 个进程通过读/写操作系统提供的共享数据缓存区来完成处理器间的通信，其延迟通常小于网络通信的延迟。
- 共享总线带宽
  - 所有处理器共享总线的带宽，完成对内存模块和I/O 模块的访问。
- 支持消息传递、共享存储并行程序设计。

SMP的缺点

- 欠可靠
  - 总线、存储器或操作系统失效可导致系统崩溃。
- 可扩展性（scalability）较差
  - 由于所有处理器共享总线带宽，而总线带宽每3 年才增加2 倍，跟不上处理器速度和内存容量的增加步伐，因此，SMP 并行机的处理器个数一般少于32 个，且只能提供每秒数百亿次的浮点运算性能。

NUMA（Non-Uniform Memory Access）

- 非均匀存储访问（NUMA）



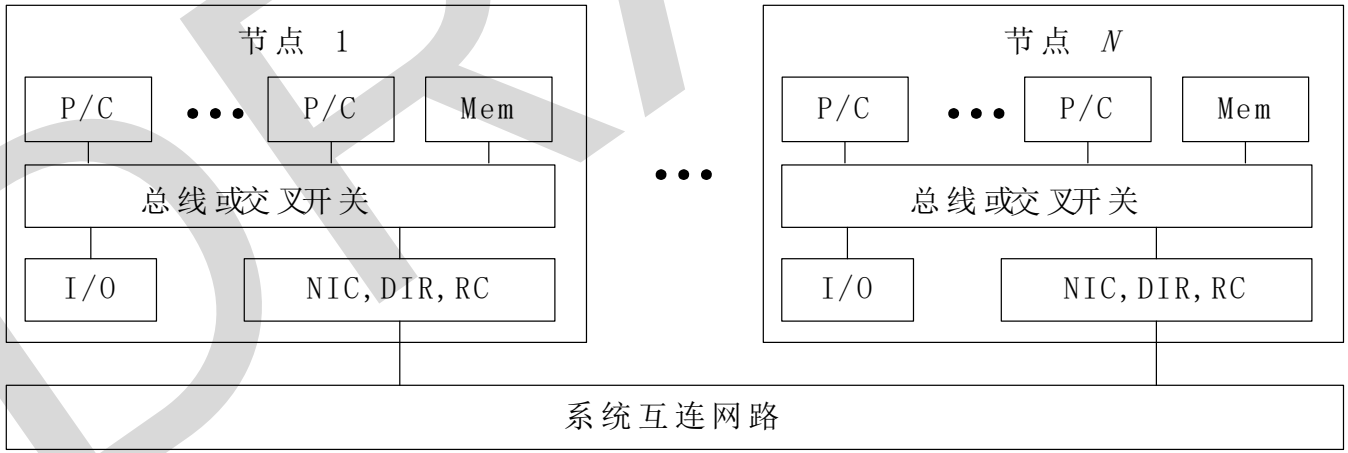
(a) 共享本地存储模型　　(b) 层次式机群模型

- 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；
- 处理器访问存储器时间是不一样的；
- 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。

- 如果缓存一致性能够得到维护，那么也可以称之为高速缓存一致性NUMA（Cache Coherent NUMA，CC-NUMA），反之可称为高速缓存非一致性NUMA（Non- Cache Coherent NUMA , NCC-NUMA）
- 特例：全高速缓存存储访问（Cache-Only Memory Access, COMA）模型，COMA各个处理器节点没有存储层次结构，所有节点的高速缓存构成了全局地址空间

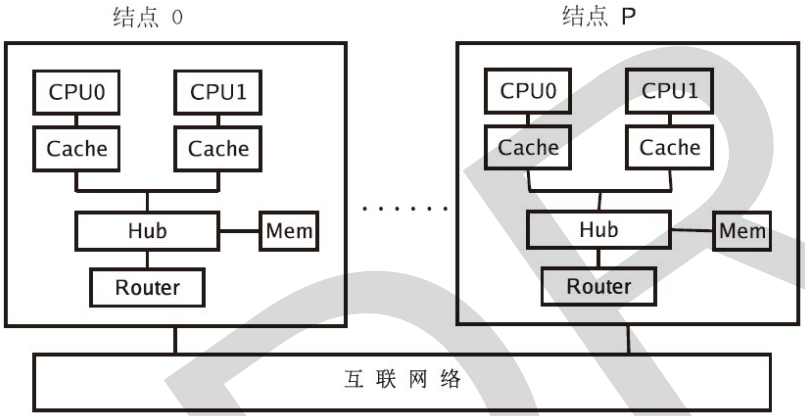
CC-NUMA（高速缓存一致性非均匀存储访问模型）

- 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。

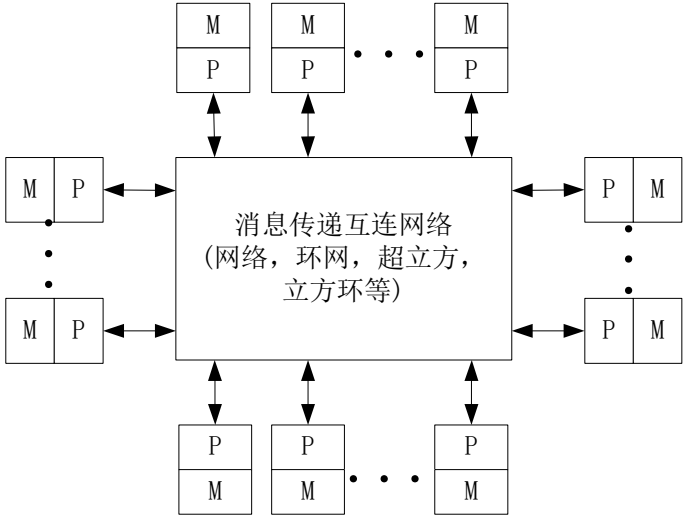


分布式内存

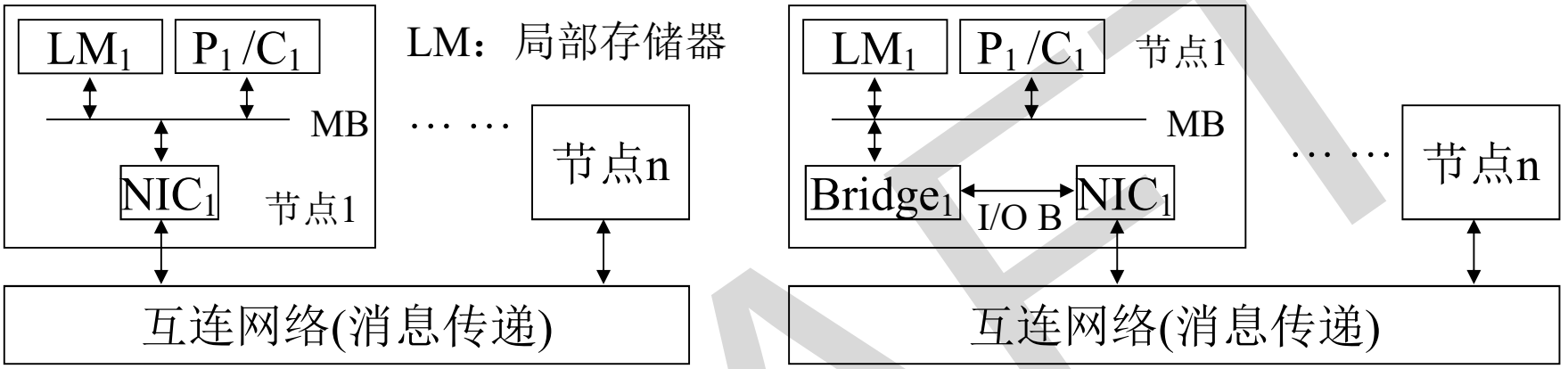
- 分布式内存的多计算机（multicomputers）体系



- 分布式内存系统中处理器都有各自的内部寄存器，一个核内的内存地址对其他核不可见，只能由该处理器所访问，对于所有CPU都没有单一全局地址空间的概念，这类的分布式计算机系统称为非远程存储访问（No-Remote Memory Access, NORMA）



NORMA（非远程存储访问）



优点是：

（1）内存可以随着CPU的数量进行扩展，增加处理器数量将使内存的大小等比例增加。

（2）各个处理器可以无冲突地快速访问自己的内存，也不存在维护缓存一致性的开销。

（3）成本效益上，可以使用商用、现成的处理器和网络。

局限性：

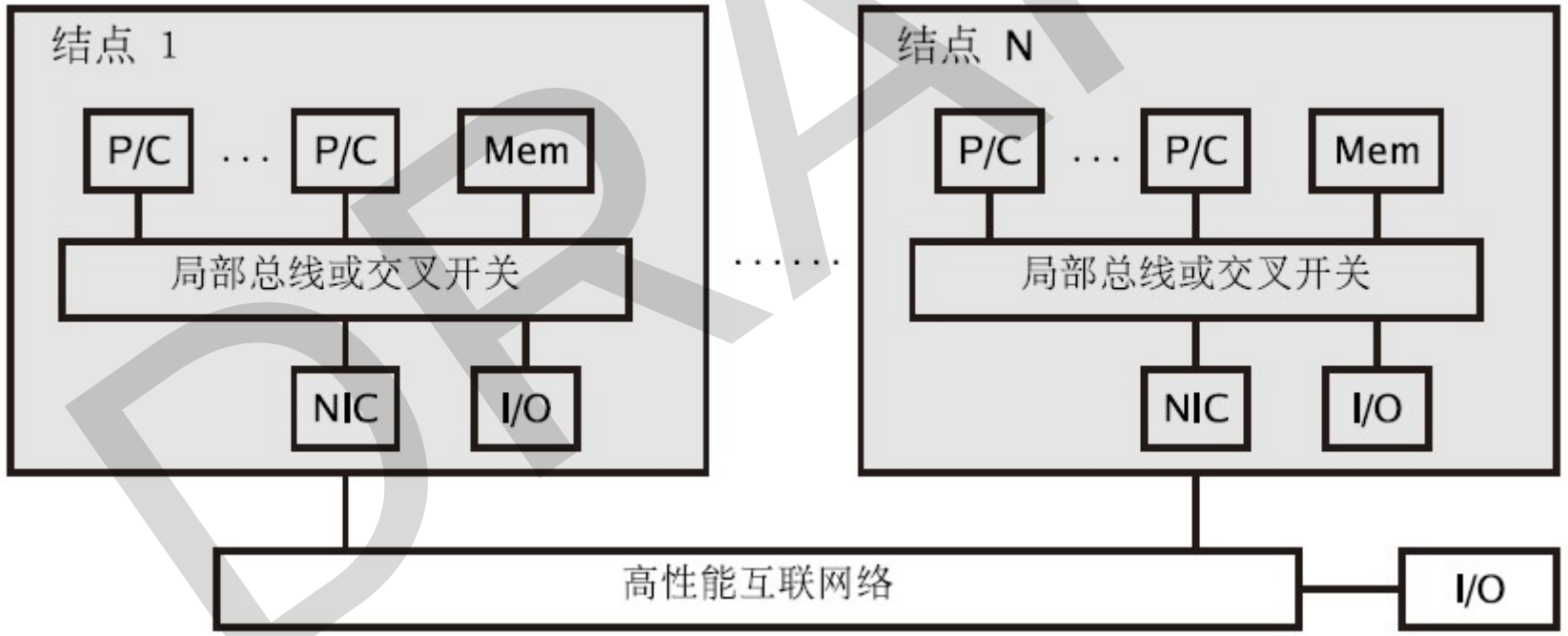
（1）程序员将要负责所有处理器间数据通信相关的细节问题。

（2）很难从基于全局内存空间的数据结构上建立起到分布式内存管理的映射。

（3）非一致的内存访问时间使得驻留在远程节点上的数据比节点本地数据的访问需要更长的时间。

大规模并行处理机（Massively Parallel Processor , MPP）

- 大规模并行机系统是典型的分布存储系统
- 内存访问属非远程存储访问模型（NORMA）。

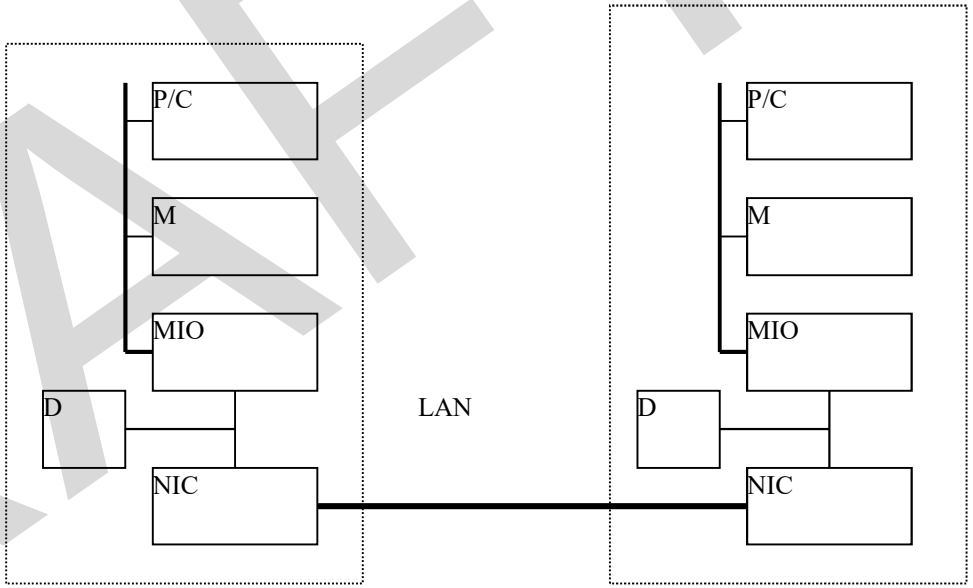


MPP的优点

- 由数百个乃至数千个计算结点和I/O 结点组成，每个结点相对独立，并拥有一个或多个微处理器。
  - 这些结点配备有局部cache，并通过局部总线或互联网络与局部内存模块和I/O 设备相连接。
  - 这些结点由局部高性能网卡(NIC) 通过高性能互联网络相互连接。
  - 各个结点均拥有不同的操作系统映像。
    - 一般情况下，用户可以将作业提交给作业管理系统，由它负责调度当前最空闲、最有效的计算结点来执行该作业。但是，MPP也允许用户登录到某个特定的结点，或在某些特定的结点上运行作业。
  - 各个结点间的内存模块相互独立，且不存在全局内存单元的统一硬件编址。
  - 仅支持消息传递或者高性能Fortran 并行程序设计，不支持全局共享的OpenMP 并行程序设计模式。

集群/机群（Cluster）工作站机群COW

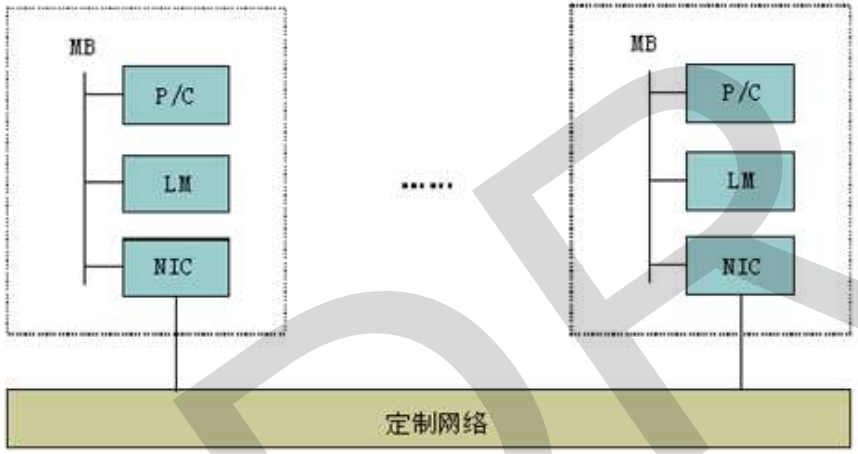
- 分布式存储，MIMD，工作站+商用互连网络，每个节点是一个完整的计算机，有自己的磁盘和操作系统
- 优点：
  - 投资风险小
  - 系统结构灵活
  - 性能/价格比高
  - 能充分利用分散的计算资源
  - 可扩放性好
- 问题
  - 通信性能
  - 并行编程环境



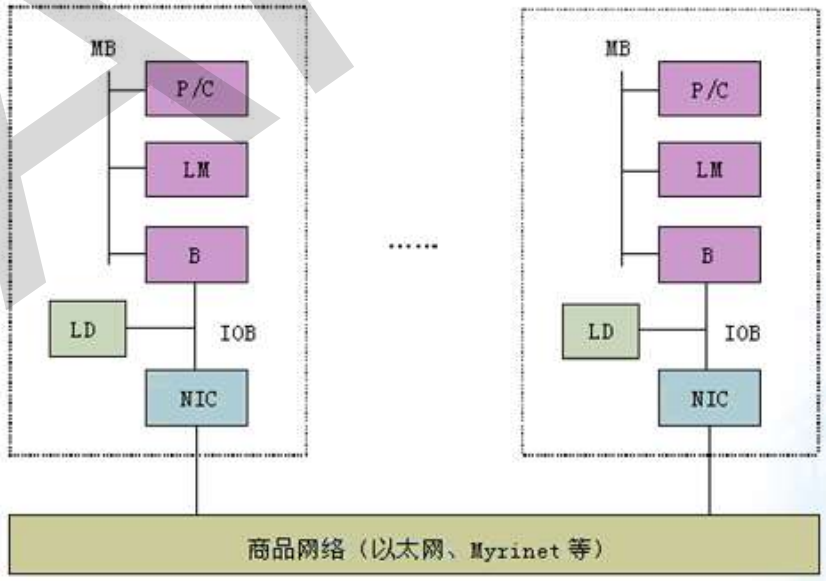
- 例子：Berkeley NOW，Alpha Farm, FXCOW

MPP vs COW

- COW的每个结点都是一台完整的计算机（可能没有鼠标、显示器等外设）。COW的每个结点上都有完整的操作系统，而MPP的每个结点上通常只有操作系统的微核。
- COW的每个结点内有本地磁盘，而MPP的结点内没有。
- COW各结点的网络接口是连接到I/O总线上的（松耦合），而MPP各结点的网络接口是连接到存储总线上的（紧耦合）。



其中MB：存储器总线，P/C：处理器和高速缓存，NIC：网络接口电路，LM：表示本地存储器。



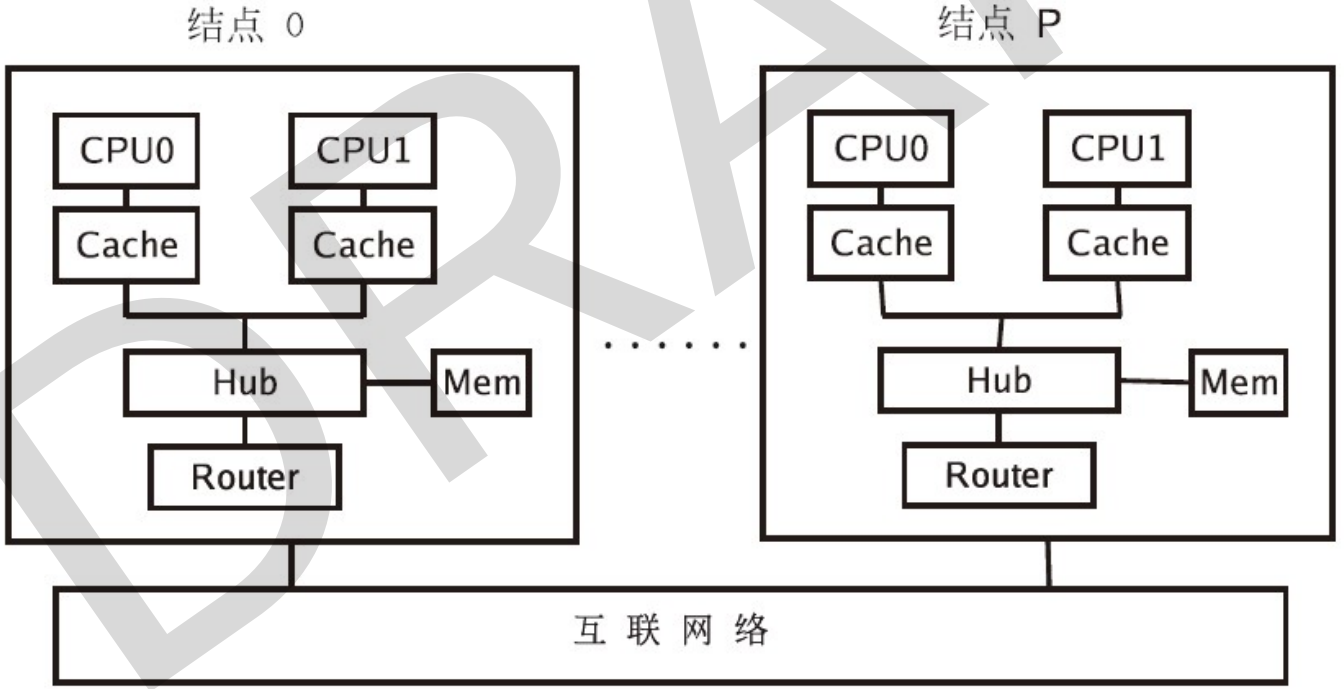
其中LD：本地磁盘，B：存储总线与I/O总线的接口，IOB：I/O总线。

SMP\MPP\机群比较

系统特征	SMP	MPP	机群
节点数量(N)	≤O(10)	O(100)-O(1000)	≤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

DSM （Distributed Shared Memory, DSM）

- 内存模块局部在各个结点内部，并被所有结点共享。这样，可以较好地改善对称多处理共享存储并行机的可扩展能力
- 内存访问属非一致内存访问模型（NUMA）。



DSM的特点（与SMP的相同点）

- 单一的操作系统映像。同SMP
- 基于cache 的数据一致性。同SMP
- 低通信延迟（同SMP）与高通信带宽
  - 专用的高性能互联网络使得结点间的延迟很小
  - 通信带宽可以扩展。
- 支持消息传递、共享存储并行程序设计。同SMP

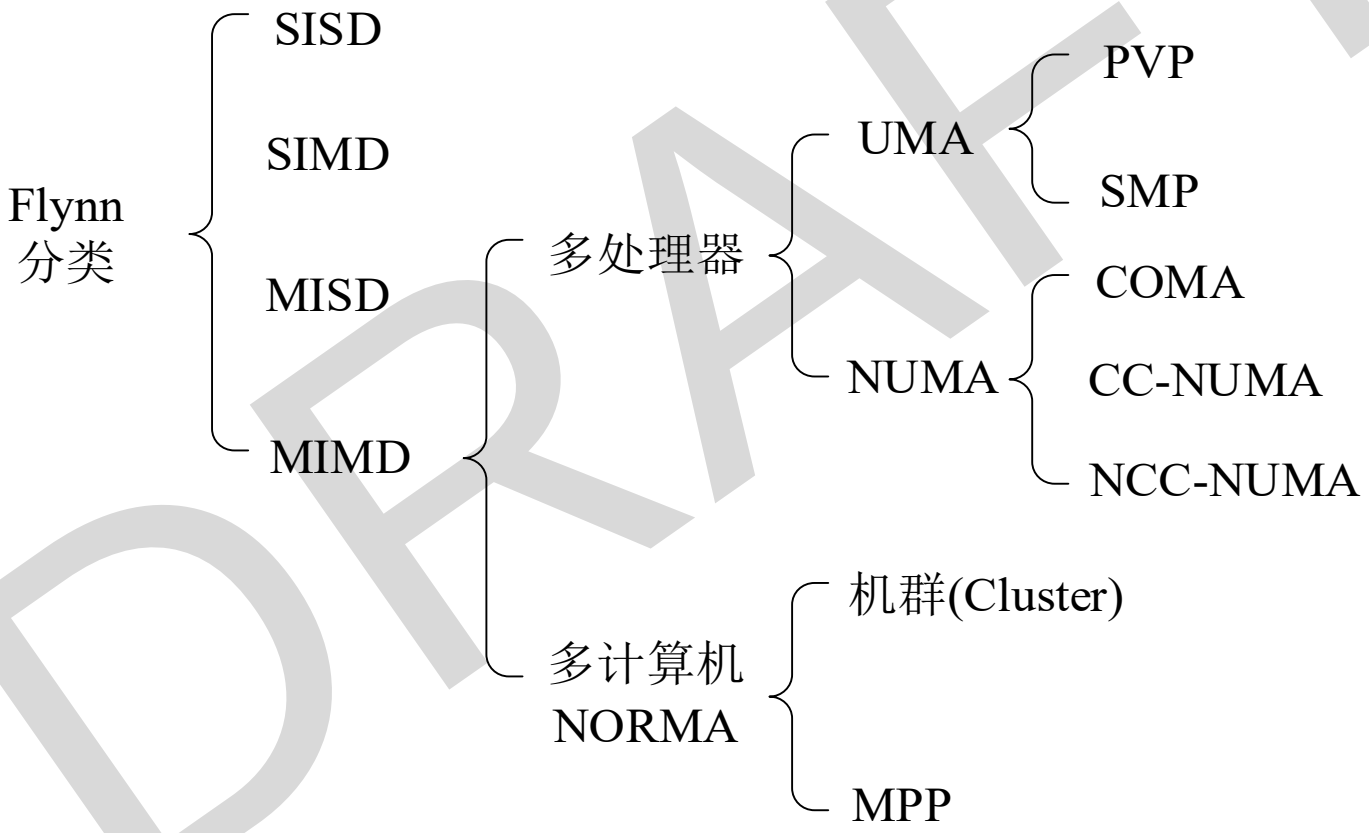
DSM的特点（2）

- 并行机以结点为单位，每个结点包含一个或多个CPU，每个CPU 拥有自己的局部cache，并共享局部存储器和I/O设备，所有结点通过高性能互联网络相互连接
- 物理上分布存储
  - 内存模块分布在各结点中，并通过高性能互联网络相互连接，避免了SMP 访存总线的带宽瓶颈，增强了并行机的可扩展能力。
- 单一的内存地址空间
  - 尽管内存模块分布在各个结点，但是，所有这些内存模块都由硬件进行统一编址，并通过互联网络连接形成了并行机的共享存储器。各个结点既可以直接访问本地局部内存单元，又可以直接访问其他结点的局部内存单元。

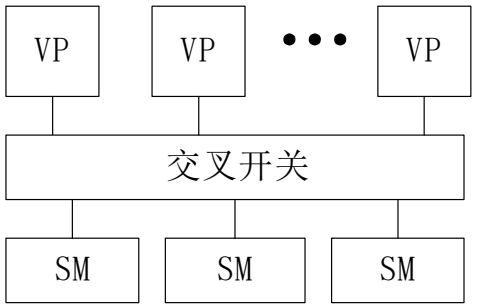
DSM的特点（3）

- 非一致内存访问（NUMA）模式
  - 由于远端访问必须通过高性能互联网络，而本地访问只需直接访问局部内存模块，因此，远端访问的延迟一般是本地访问延迟的3倍以上。
- 可扩展性强
  - 可扩展到数百个结点，能提供每秒数千亿次的浮点运算性能。
  - 但是，由于受cache 一致性要求和互联网络性能的限制，当结点数目进一步增加时，DSM 并行机的性能也将下降。

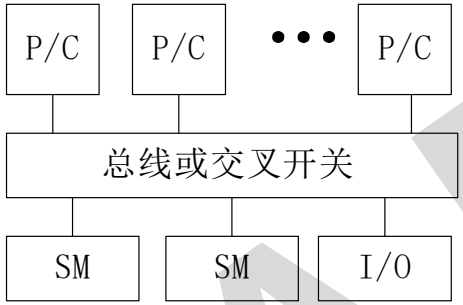
内存体系


Flynn
分类
SISD
SIMD
MISD
MIMD
多处理器
UMA
PVP
SMP
NUMA
COMA
CC-NUMA
NCC-NUMA
多计算机
NORMA
机群(Cluster)
MPP

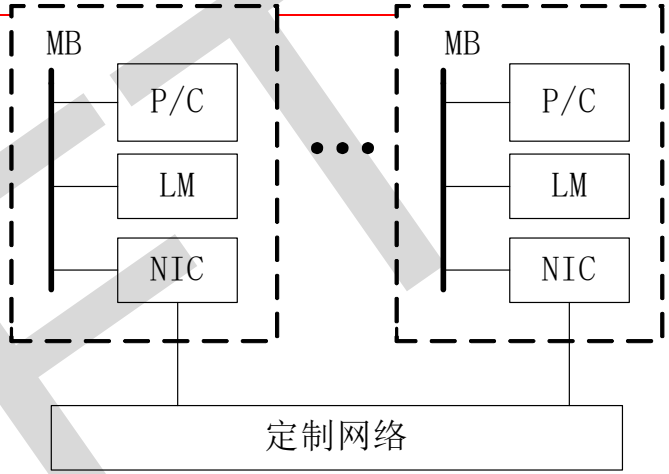
并行计算机结构模型



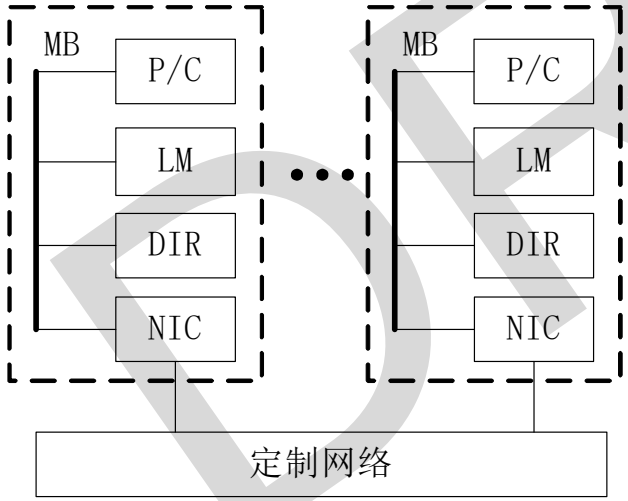
(a) PVP



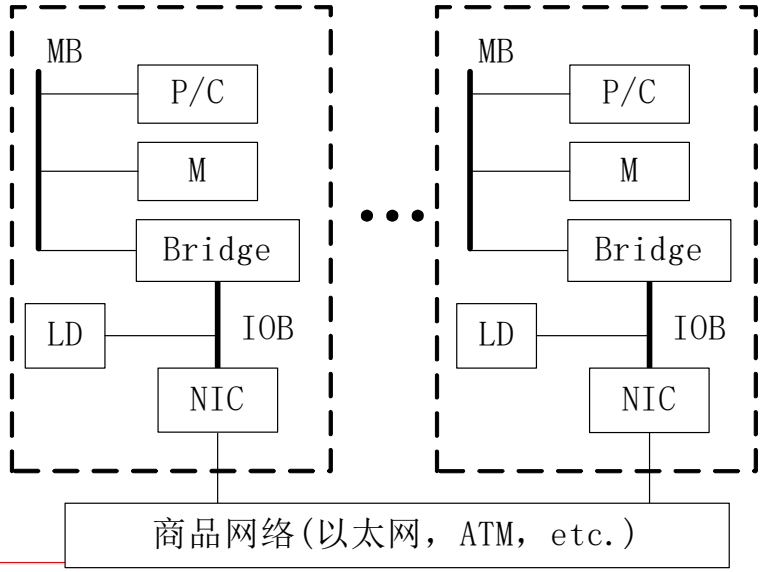
(b) SMP



(c) MPP



(d) DSM



(e) COW

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，曙光1号	Intel Paragon，IBMSP2，曙光1000/2000	Stanford DASH，Cray T 3D	Berkeley NOW，Alpha Farm

目录

- Flynn分类
- 并行计算机的内存体系
- 并行计算机的处理器

处理器

- 传统的单核处理器对计算机系统性能提升产生瓶颈
  - 必然导致多核架构的普及并且逐步成为主流
- 多核处理器
  - 通过集成多个核来提高性能，每个核的主频可以适当降低，这样可以减少功耗
  - 多核处理器中的内核采用了小型的处理器，功能相对简单，这使得多核处理器的设计和验证的周期比较短，开发风险和成本能够有效地降低，而且便于优化和重新设计。
  - 在多核处理器中，如果一个程序采用了线程级并行编程，那么这个程序在运行时可以把并行的线程同时交付给两个核心分别处理，因而程序运行速度得到极大提高。这些程序往往可以不作任何改动就直接运行在双核电脑上。多核处理器与传统的对称更多处理器系统相近，所以线程级的应用可以比较容易地从传统的单处理器或对称多处理器平台移植到多核环境中。
  - 当在多核处理器上同时运行多个单线程程序的时候，操作系统会把多个程序的指令分别发送给多个内核，从而使得同时完成多个程序的速度大大加快

协处理器

- GPGPU
- MIC
- FPGA