并行计算

——结构•算法•编程

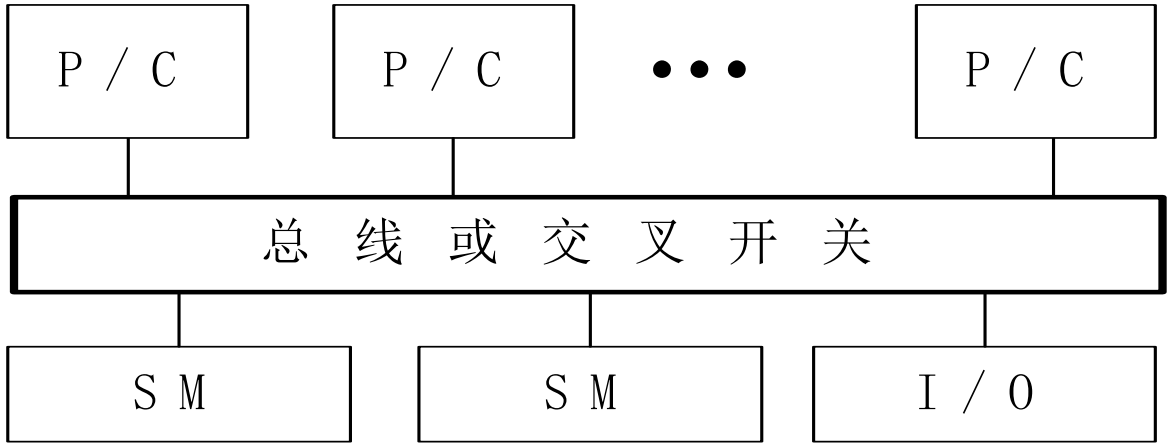
主讲教师：谢磊

第二章 当代并行机系统

* 2.1 共享存储多处理机系统
  * 2.1.1 对称多处理机SMP结构特性
* 2.2 分布存储多计算机系统
  * 2.2.1 大规模并行机MPP结构特性
* 2.3 机群系统
  * 2.3.1 大规模并行处理系统MPP机群SP2
  * 2.3.2 工作站机群COW

对称多处理机SMP(1)

* SMP: 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, Dawning 1



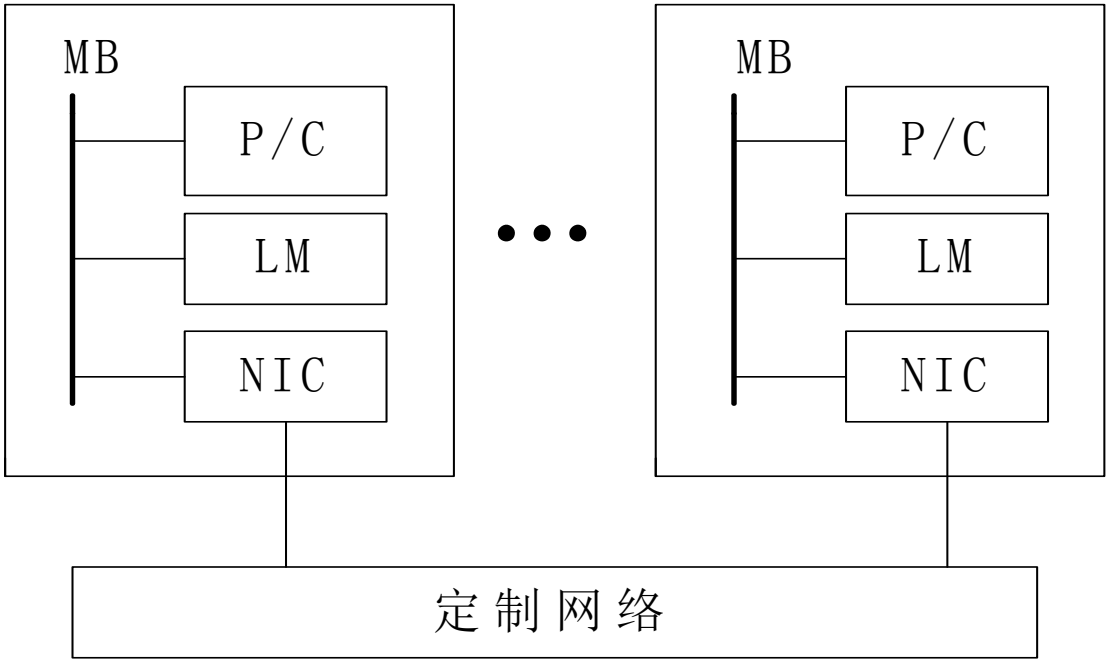
对称多处理机SMP(2)

* 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

大规模并行机MPP

* 成百上千个处理器组成的大规模计算机系统，规模是变化的。基于NORMA结构，高带宽低延迟定制互连。
* 可扩放性
  * Mem, I/O,平衡设计
* 系统成本
  * 商用处理器，相对稳定的结构，SMP,分布
* 通用性和可用性
  * 不同的应用，PVM,MPI,交互，批处理，互连对用户透明，单一系统映象，故障
* 通信要求
* 存储器和I/O能力
* 例子：Intel Option Red，IBM SP2 Dawning 1000

大规模并行机MPP



典型MPP系统特性比较

MPP模型	Intel/Sandia ASCI Option Red	IBM SP2	SGI/Cray Origin2000
一个大型样机的配置	9072个处理器，1.8Tflop/s(NSL)	400个处理器，100Gflop/s(MHPCC)	128个处理器，51Gflop/s(NCSA)
问世日期	1996年12月	1994年9月	1996年10月
处理器类型	200MHz，200Mflop/s Pentium Pro	67MHz，267Mflop/s POWER2	200MHz，400Mflop/s MIPS R10000
节点体系结构和数据存储器	2个处理器，32到256MB主存，共享磁盘	1个处理器，64MB到2GB本地主存，1GB到14.5GB本地磁盘	2个处理器，64MB到256MB分布共享主存和共享磁盘
互连网络和主存模型	分离两维网孔，NORMA	多级网络，NORMA	胖超立方体网络，CC-NUMA
节点操作系统	轻量级内核（LWK）	完全AIX（IBM UNIX）	微内核Cellular IRIX
自然编程机制	基于PUMA Portals的MPI	MPI和PVM	Power C， Power Fortran
其他编程模型	Nx，PVM，HPF	HPF，Linda	MPI，PVM

MPP所用的高性能CPU特性比较

属性	Pentium Pro	PowerPC 602	Alpha 21164A	Ultra SPARC II	MIPS R10000
工艺	BiCMOS	CMOS	CMOS	CMOS	CMOS
晶体管数	5.5M/15.5M	7M	9.6M	5.4M	6.8M
时钟频率	150MHz	133MHz	417MHz	200MHz	200MHz
电压	2.9V	3.3V	2.2V	2.5V	3.3V
功率	20W	30W	20W	28W	30W
字长	32位	64位	64位	64位	64位
I/O高速缓存	8KB/8KB	32KB/32KB	8KB/8KB	16KB/16KB	32KB/32KB
2级高速缓存	256KB (多芯片模块)	1~128MB (片外)	96KB (片上)	16MB (片外)	16MB (片外)
执行单元	5个单元	6个单元	4个单元	9个单元	5个单元
超标量	3路(Way)	4路	4路	4路	4路
流水线深度	14级	4~8级	7~9级	9级	5~7级
SPECint 92	366	225	>500	350	300
SPECfp 92	283	300	>750	550	600
SPECint 95	8.09	225	>11	N/A	7.4
SPECfp 95	6.70	300	>17	N/A	15
其它特性	CISC/RISC混合	短流水线长L1高速缓存	最高时钟频率最大片上2级高速缓存	多媒体和图形指令	MP机群总线可支持4个CPU

机群系统(1)

* **计算机机群(Cluster)**简称**集群**是一种计算机系统，它通过一组松散集成的计算机软件和/或硬件连接起来高度紧密地协作完成计算工作。
* 在某种意义上，他们可以被看作是一台计算机。
* 集群系统中的单个计算机通常称为节点，通常通过局域网连接，但也有其它的可能连接方式。
* 集群计算机通常用来改进单个计算机的计算速度和/或可靠性。一般情况下集群计算机比单个计算机，比如工作站或超级计算机性能价格比要高得多。

机群系统(2)

* 集群分为同构与异构两种，它们的区别在于：组成集群系统的计算机之间的体系结构是否相同。集群计算机按功能和结构可以分成以下几类:
  * 高性能计算集群 High-performance (HPC) clusters
  * 负载均衡集群 Load balancing clusters
  * 高可用性集群 High-availability (HA) clusters
  * 网格计算 Grid computing

机群系统(3)

* 高性能计算集群

  * 高性能计算集群采用将计算任务分配到集群的不同计算节点而提高计算能力，因而主要应用在科学计算领域。比较流行的HPC采用Linux操作系统和其它一些免费软件来完成并行运算。这一集群配置通常被称为Beowulf集群。这类集群通常运行特定的程序以发挥HPC cluster的并行能力。这类程序一般应用特定的运行库，比如专为科学计算设计的MPI库。
  * HPC集群特别适合于在计算中各计算节点之间发生大量数据通讯的计算作业，比如一个节点的中间结果或影响到其它节点计算结果的情况。

机群系统(4)

* 负载均衡集群
  * 负载均衡集群运行时，一般通过一个或者多个前端负载均衡器，将工作负载分发到后端的一组服务器上，从而达到整个系统的高性能和高可用性。这样的计算机集群有时也被称为服务器群（Server Farm）。一般高可用性集群和负载均衡集群会使用类似的技术，或同时具有高可用性与负载均衡的特点。
  * Linux虚拟服务器（LVS）项目在Linux操作系统上提供了最常用的负载均衡软件。

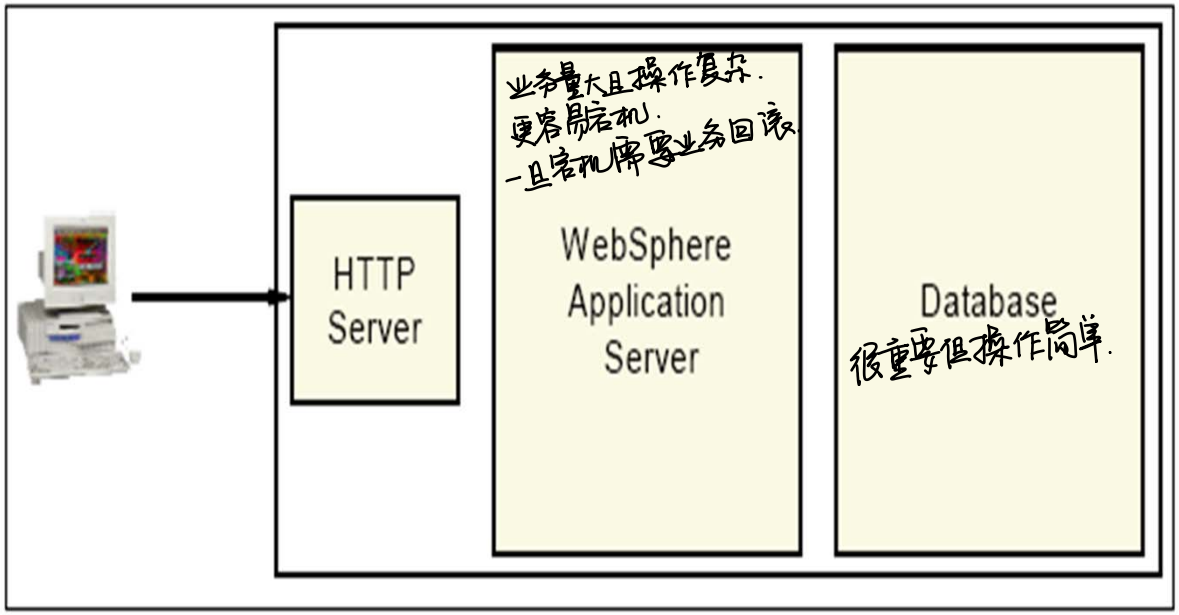
机群系统(5)

* 高可用性集群
  * 一般是指当集群中有某个节点失效的情况下，其上的任务会自动转移到其他正常的节点上。还指可以将集群中的某节点进行离线维护再上线，该过程并不影响整个集群的运行。

切换→重启
宕机成本较高.

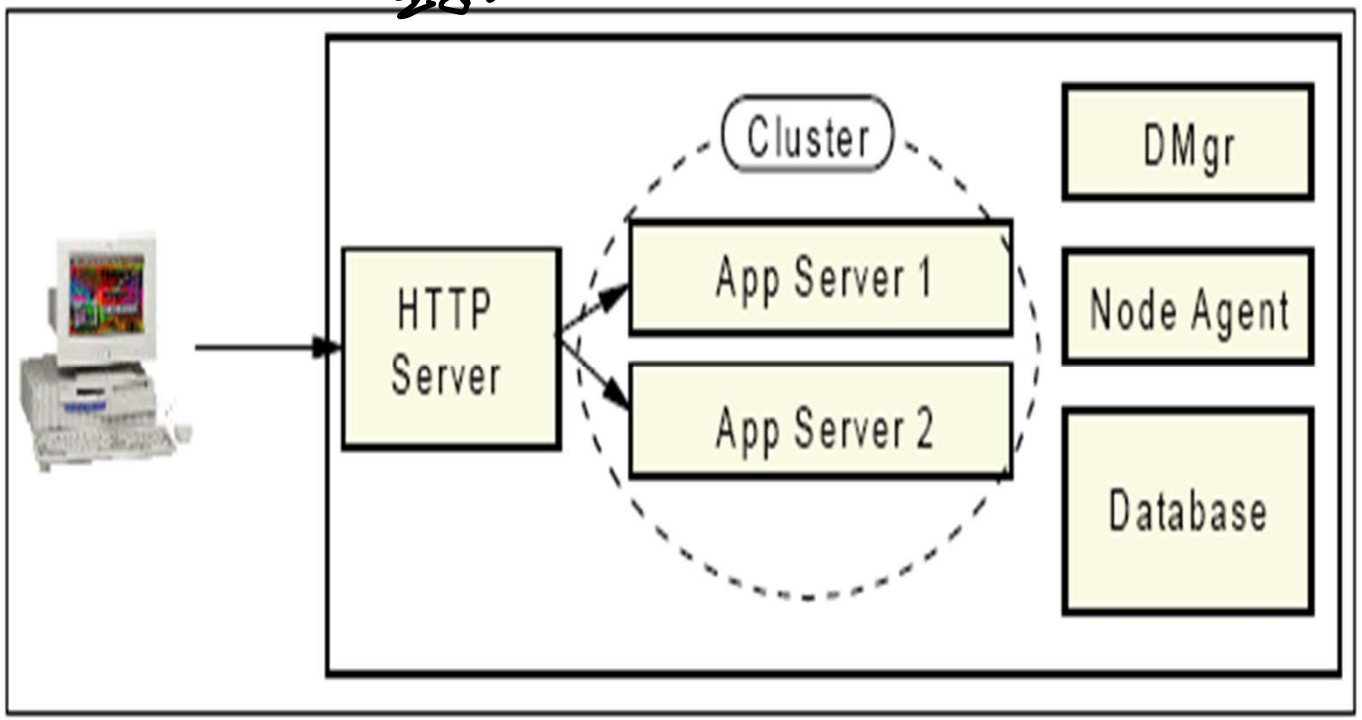
IBM WAS HA 高可用性. High availability.

* Levels of WebSphere system availability
  * WebSphere system HA level 1-level 5
* WebSphere system HA level 1

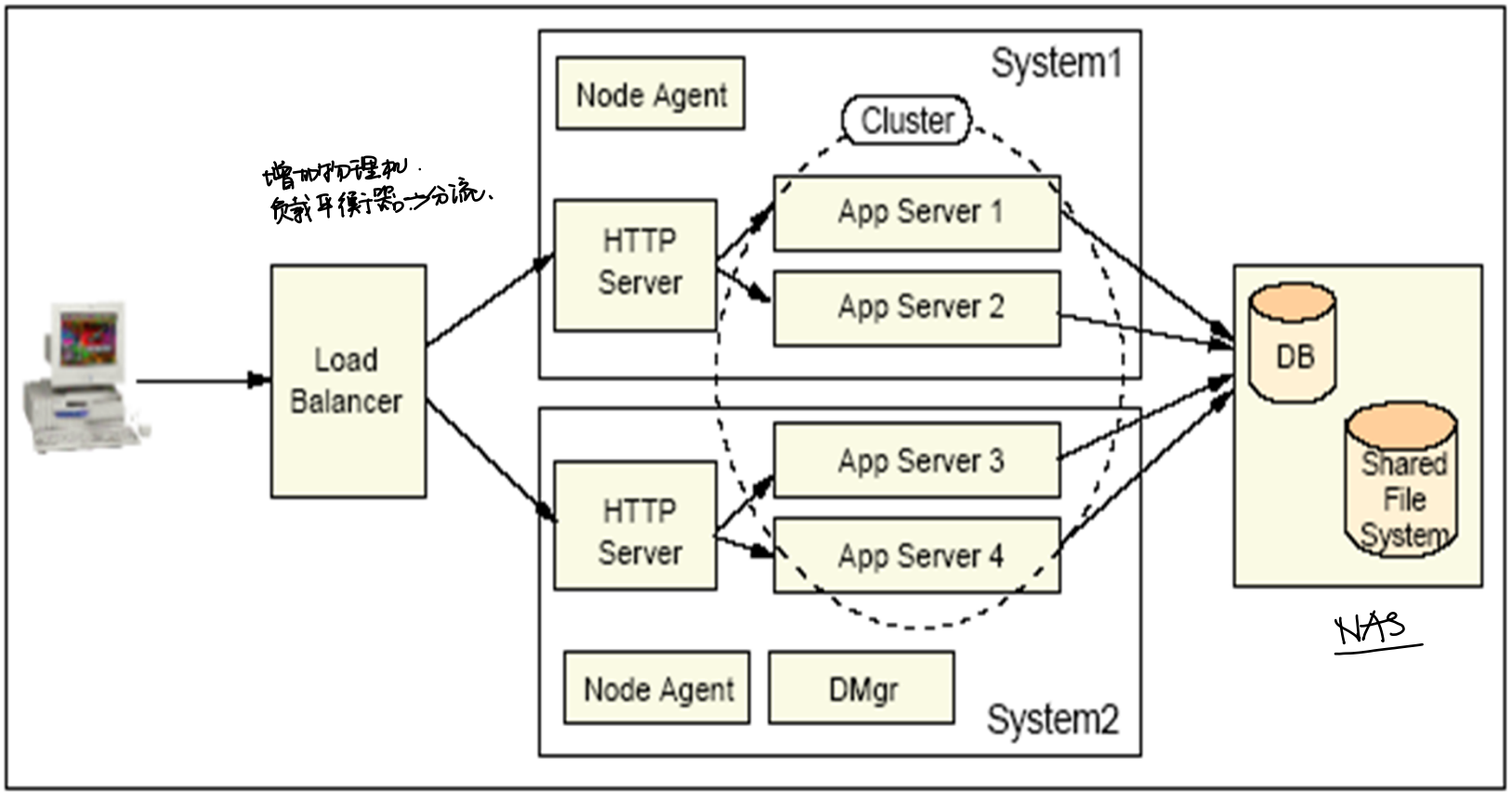


WAS HA-level 2

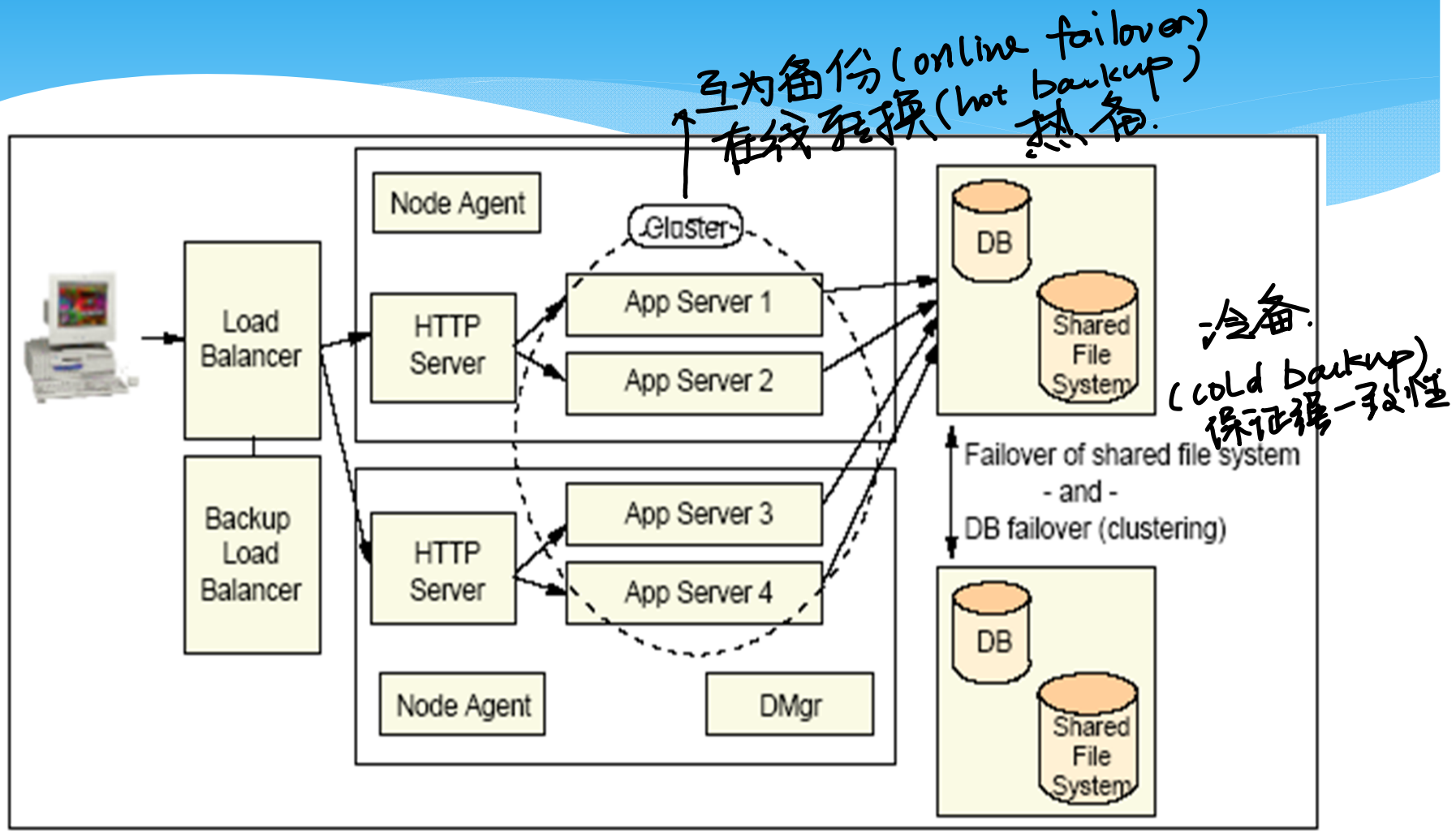
主要缺点：共有单个物理机(OS)



WAS HA-level 3



WAS HA-level 4

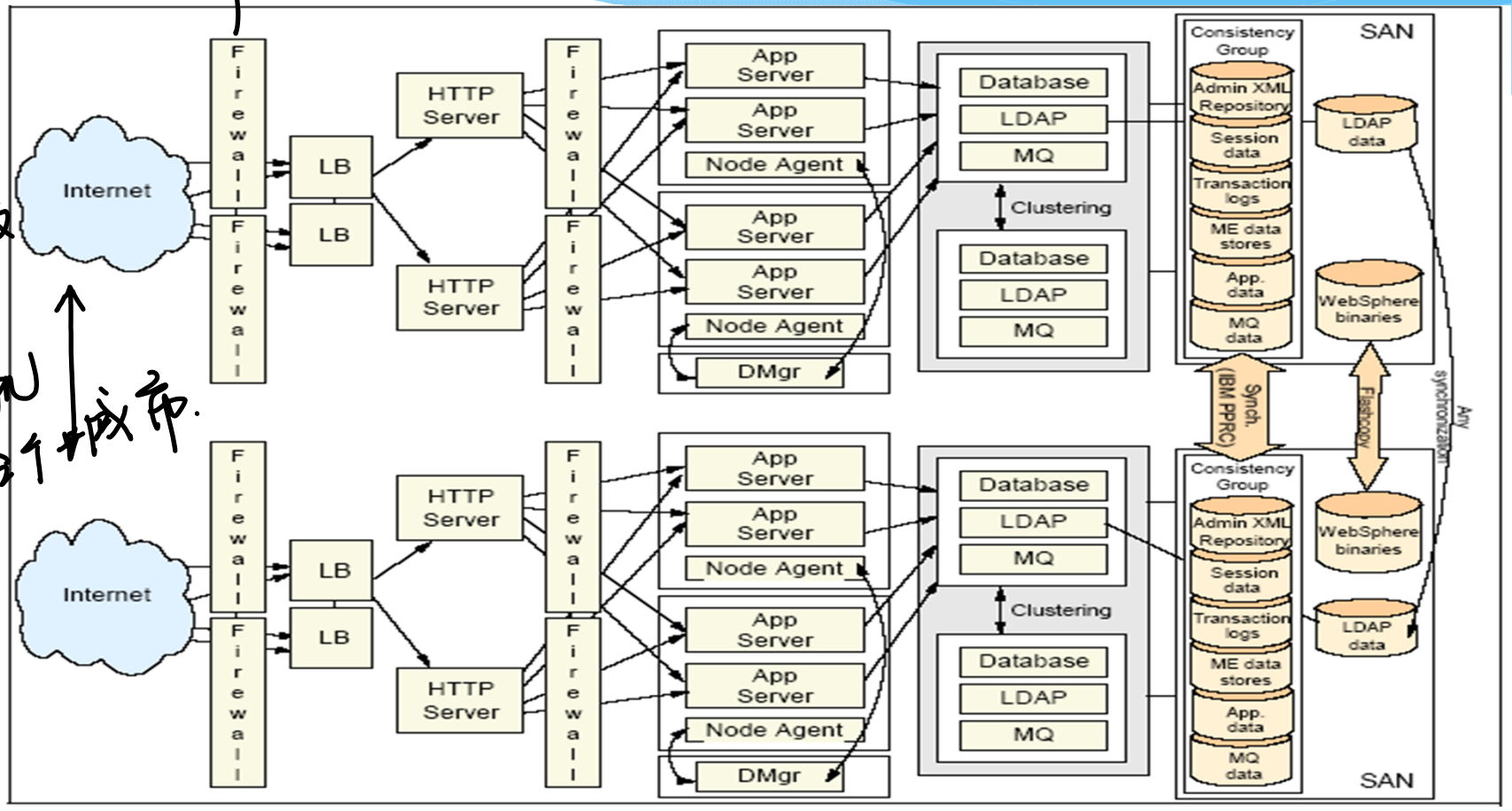


WAS HA-level 5

安全考虑

地域级备份.

将物理机分散到各个城市.



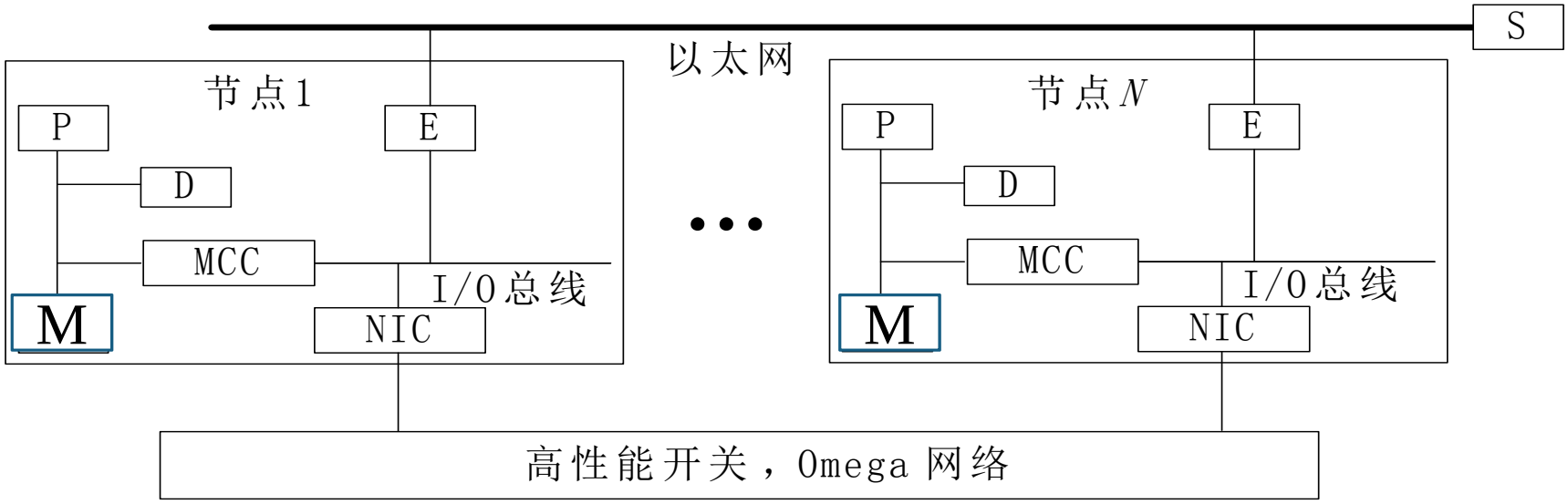
机群系统(6)

* 网格计算
  * 网格计算或网格集群是一种与集群计算非常相关的技术。网格与传统集群的主要差别是网格是连接一组相关并不信任的计算机，它的运作更像一个计算公共设施而不是一个独立的计算机。还有，网格通常比集群支持更多不同类型的计算机集合。
  * 网格计算是针对有许多独立作业的工作任务作优化，在计算过程中作业间无需共享数据。网格主要服务于管理在独立执行工作的计算机间的作业分配。资源如存储可以被所有结点共享，但作业的中间结果不会影响在其他网格结点上作业的进展。

机群型大规模并行机SP2

* 设计策略：
  * 机群体系结构
  * 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：
  * 高性能开关 HPS 多级Ω网络
  * 宽节点、窄节点和窄节点2

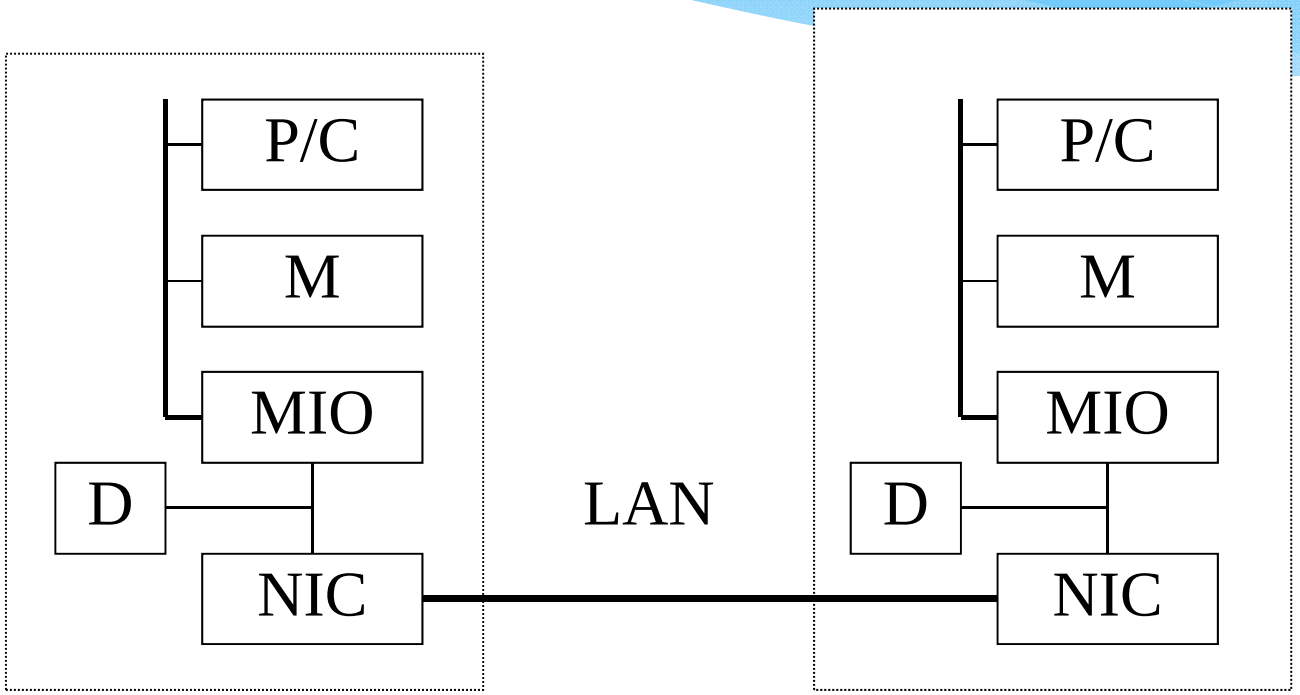
机群型大规模并行机SP2



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，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
  * 可扩放性好
* 问题
  * 通信性能
  * 并行编程环境
* 例子：Berkeley NOW，Alpha Farm, FXCOW

工作站机群COW



典型的机群系统

典型的机群系统特点一览表

名称	系统特点
Princeton:SHRIMP	PC商用组件，通过专用网络接口达到共享虚拟存储，支持有效通信
Karsruhe:Parastation	用于分布并行处理的有效通信网络和软件开发
Rice:TreadMarks	软件实现分布共享存储的工作站机群
Wisconsin:Wind Tunnel	在经由商用网络互连的工作站机群上实现分布共享存储
Chica、Maryl、Penns:NSCP	国家可扩放机群计划：在通过因特网互连的3个本地机群系统上进行元计算
Argonne:Globus	在由ATM连接的北美17个站点的WAN上开发元计算平台和软件
Syracuse:WWVM	使用因特网和HPCC技术，在世界范围的虚拟机上进行高性能计算
HKU:Pearl Cluster	研究机群在分布式多媒体和金融数字库方面的应用
Virgina:Legion	在国家虚拟计算机设施上开发元计算软件

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

并行计算机存储组织

* 层次存储技术
  * 自上而下，速度由快到慢，容量由小到大
* 高速缓存的一致性
  * 高速缓存的一致性问题
  * 高速缓存写策略
    * 写通过（Write-Through）
    * 写回（Write-Back）
  * 高速缓存不一致的原因
    * 由共享可写数据造成的不一致
    * 由进程迁移造成的不一致
    * 由绕过高速缓存的I/O操作所造成的不一致

并行计算机存储组织

* 保证高速缓存的一致性
  * 监听总线协议
    * 写无效
    * 写更新
  * 基于目录的协议
    * 使用一个目录来记录共享数据的所有高速缓存行的位置和状态