上海第二工业大学

并行计算
Parallel Computing

主讲人　陈林

Sep,2021

并行计算——结构·算法·编程

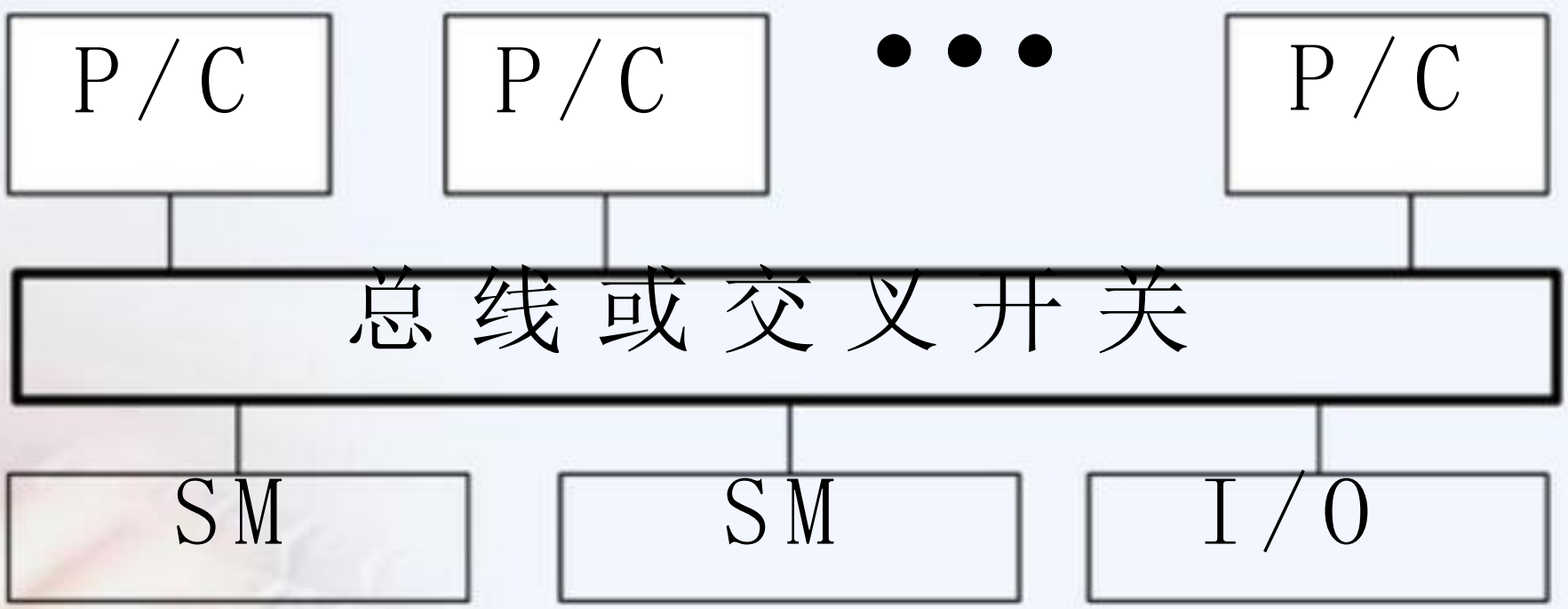
- 第一篇 并行计算的基础
  - 第一章 并行计算与并行计算机结构模型
  - 第二章 并行计算机系统互连与基本通信操作
  - 第三章典型并行计算机系统介绍
  - 第四章 并行计算性能评测

第三章 典型并行计算机系统介绍

- 3.1 共享存储多处理机系统
  - 3.1.1 对称多处理机SMP结构特性
- 3.2 分布存储多计算机系统
  - 3.2.1 大规模并行机MPP结构特性
- 3.3 分布共享存储多计算机系统
  - 3.3.1 分布共享存储计算机系统特性
- 3.4 机群系统
  - 3.4.1 大规模并行处理系统MPP机群SP2
  - 3.4.2 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：任何处理器均可访问任何存储单元和I/O设备
  - 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：支持数据的局部性，数据一致性由硬件维持
  - 低通信延迟：可由简单的Load/Store指令完成
- 问题
  - 欠可靠：BUS,OS,SM失效均会造成系统的崩溃
  - 可观的通信延迟（相对于CPU）：竞争会加剧延迟
  - 慢速增加的带宽：MB double/3 year，IOB更慢
  - 不可扩放性（用总线连接）。为此，或改用交叉开关连接，或改用CC-NUMA，或改用Cluster

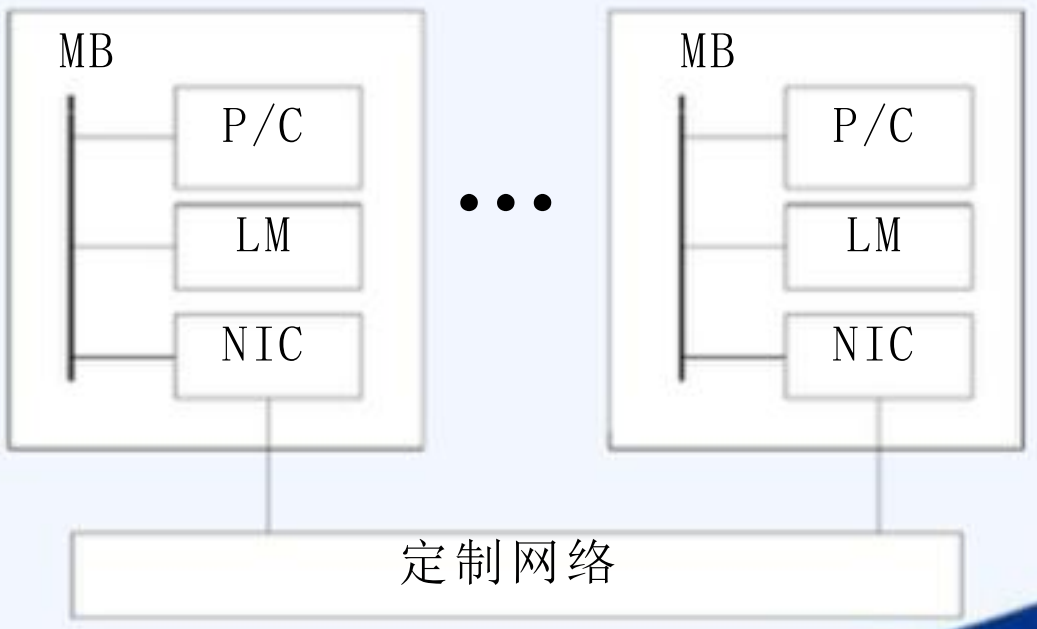
第三章 典型并行计算机系统介绍

- 3.1 共享存储多处理机系统
  - 3.1.1 对称多处理机SMP结构特性
- 3.2 分布存储多计算机系统
  - 3.2.1 大规模并行机MPP结构特性
- 3.3 分布共享存储多计算机系统
  - 3.3.1 分布共享存储计算机系统特性
- 3.4 机群系统
  - 3.4.1 大规模并行处理系统MPP机群SP2
  - 3.4.2 工作站机群COW

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。
- NORMA结构，高总计带宽，相对低延迟，定制互连。
- 可扩放性：Processors, Memory, Bandwidth, I/O, 平衡设计
- 系统成本：商用处理器，相对稳定的结构，SMP节点，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
- 通信要求：高于标准的LAN
- 较大存储器和I/O能力
- 现在MPP与Cluster难以区别
- 例子：Intel Option Red

  IBM SP2, Dawning 1000



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	8KB/8KB	32KB/32KB	8KB/8KB	16KB/16KB	32KB/32KB
高速缓存					
2级高速缓存	256KB (多芯片模块5个单) 元	1~128MB (片外) 6个单元	96KB (片上) 4个单元	16MB (片外) 9个单元	16MB (片外) 5个单元
执行单元					
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

第三章 典型并行计算机系统介绍

- 3.1 共享存储多处理机系统
  - 3.1.1 对称多处理机SMP结构特性
- 3.2 分布存储多计算机系统
  - 3.2.1 大规模并行机MPP结构特性
- 3.3 分布共享存储多计算机系统
  - 3.3.1 分布共享存储计算机系统特性
- 3.4 机群系统
  - 3.4.1 大规模并行处理系统MPP机群SP2
  - 3.4.2 工作站机群COW

DSM计算机系统特性

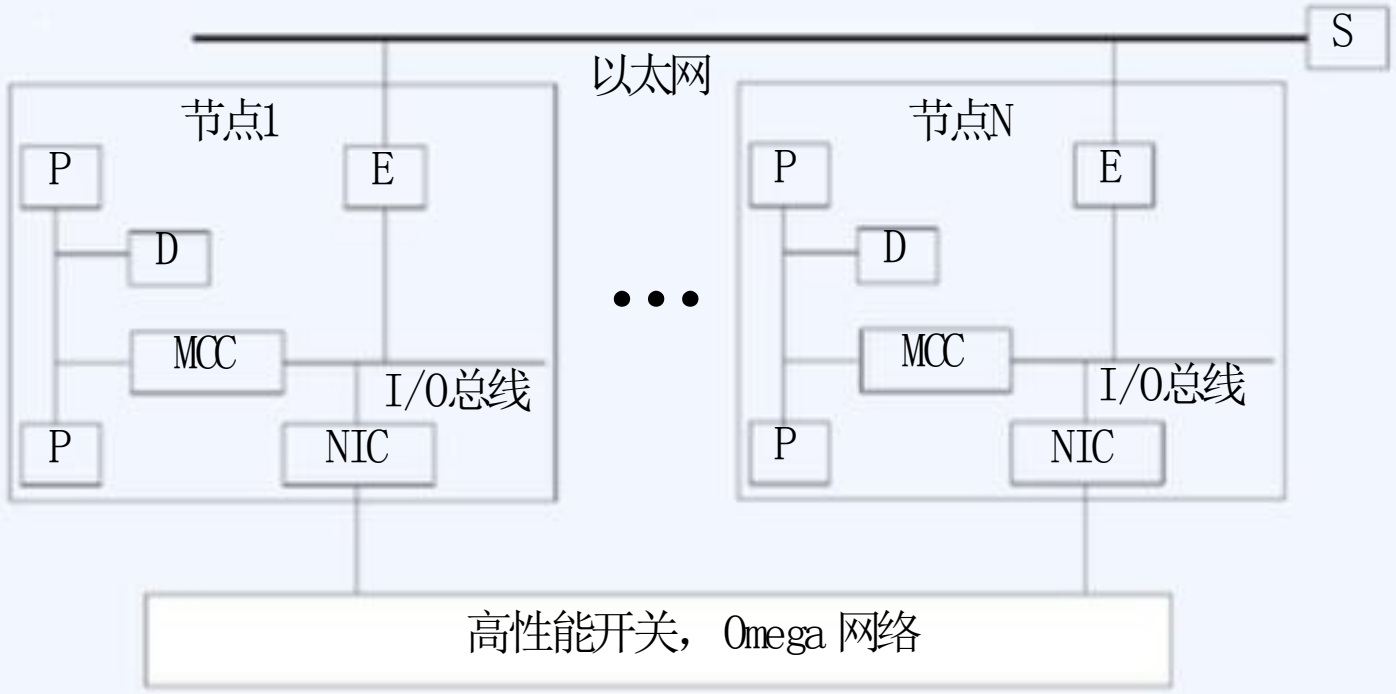
- DSM结构特性
  - 共享存储系统采用分布共享，减少集中共享的冲突
  - 采用高速缓存来缓和由共享引起的冲突和分布存储引起的长延迟
  - 保持了共享编程的方便性和软件的可移植性
- 存储一致性问题
  - 非均匀存储访问和高速缓存一致性问题
  - 影响了一些技术的应用和系统的可扩放性
- DSM系统分类
  - 硬件实现的共享存储：CC-NUMA、NCC-NUMA、COMA
  - 软件实现的共享存储：共享虚拟存储(SVM)
- 典型机器：SGI Origin 2000

第三章 典型并行计算机系统介绍

- 3.1 共享存储多处理机系统
  - 3.1.1 对称多处理机SMP结构特性
- 3.2 分布存储多计算机系统
  - 3.2.1 大规模并行机MPP结构特性
- 3.3 分布共享存储多计算机系统
  - 3.3.1 分布共享存储计算机系统特性
- 3.4 机群系统
  - 3.4.1 大规模并行处理系统MPP机群SP2
  - 3.4.2 工作站机群COW

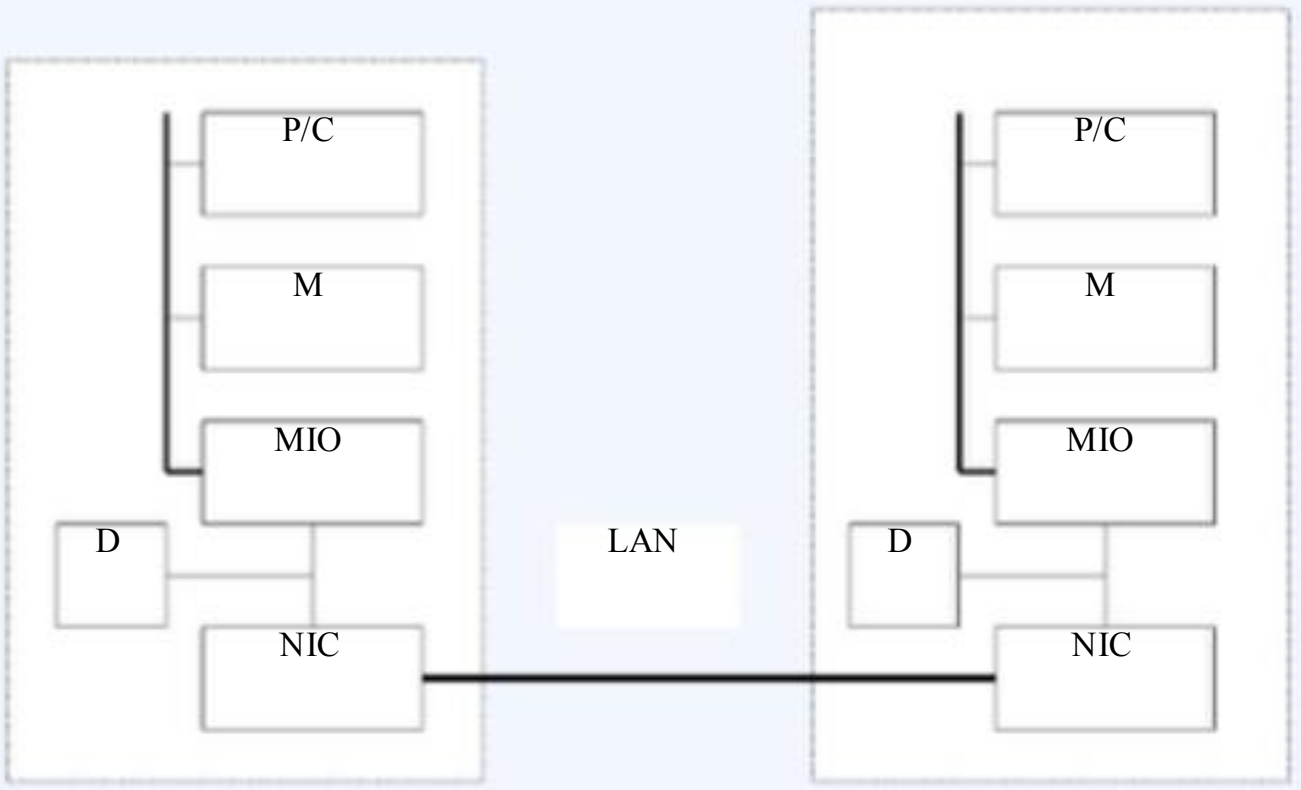
机群型大规模并行机SP2

- IBM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（多级的Ω网络）
  - 宽节点、窄节点和窄节点1



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

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

Activity

- 文献阅读1:
  从术语变化看高性能计算机的发展（孙凝晖等）。

- 文献阅读2:
  计算思维（陈国良）。

- 文献阅读3:
  查阅国产处理器(方舟、龙芯、寒武纪...)发展和现状的资料。