§2.1 并行计算机系统结构模型（第二部分）

徐悦甦(Yueshen Xu)

ysxu@xidian.edu.cn

软件工程系

西安电子科技大学

上机安排

西安电子科技大学
XIDIAN UNIVERSITY

- 上机时间（四次，16学时）
  - 按周数：第11、12、14、17周的星期三
  - 按日期：5月16号，5月23号，6月6号，6月27号
  - 全部为星期三晚上，18：30至21：30，每次三个小时
  - 地点：G楼346、348机房
  - 内容（视课程情况可能微调）
    - 第一次（5.16）：Java并发
    - 第二次（5.23）：Java并行
    - 第三次（6.6）：Python并发与并行
    - 第四次（6.27）：MapReduce单机并行

上机安排

西安电子科技大学 XIDIAN UNIVERSITY

- 星期三晚上6点半开始，至9点半结束

3		学期	G346, 348					上 学 期20180316																
4		月	2月	3月				4月				5月					6月				7月				8
5		周	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
6		日	26	5	12	19	26	2	9	16	23	30	7	14	21	28	4	11	18	25	2	9	16	23	30
7	星	课期	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|
8	期	次	4	11	18	25	1/4	8	15	22	29	6/5	13	20	27	3/6	10	17	24	1/7	8	15	22	29	5/8
9		上										五							端						
10	一	下										一							午						
11		晚										节							节						
12		上										五	中												
13	二	下										一	考												
14		晚										节	试												
15		上									运		中												
16	三	下									动		考												
17		晚									会		试	并	并		并			并					
18		上						清			运														
19	四	下						明			动														
20		晚						节			会														

本节提纲

西安电子科技大学 XIDIAN UNIVERSITY

- 并行计算机访存模型
  - 均匀存储访问与非均匀存储访问
  - 高速缓存存储访问
  - 非远程存储访问模型等
  - 体系结构与访存模型的关系

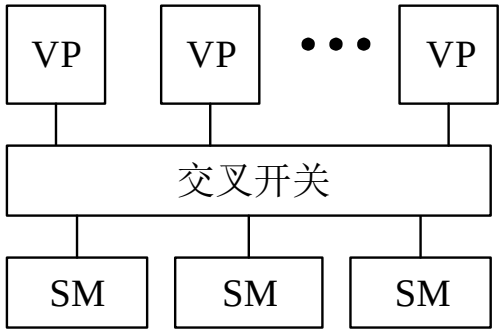
从存储器的角度

- 并行计算机性能测评
  - 并行计算机的基本性能指标
  - 加速比性能测评
  - 可扩展性性能测评

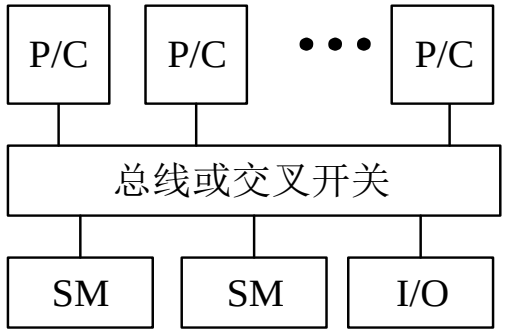
讨论一部分

上节课回顾
➔并行计算机体系结构模型汇总

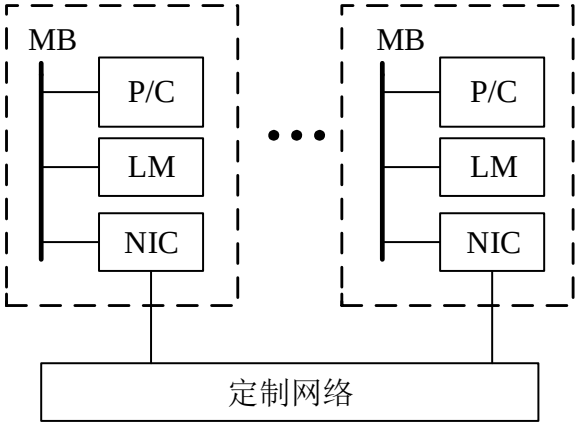
- 各实现形式比对



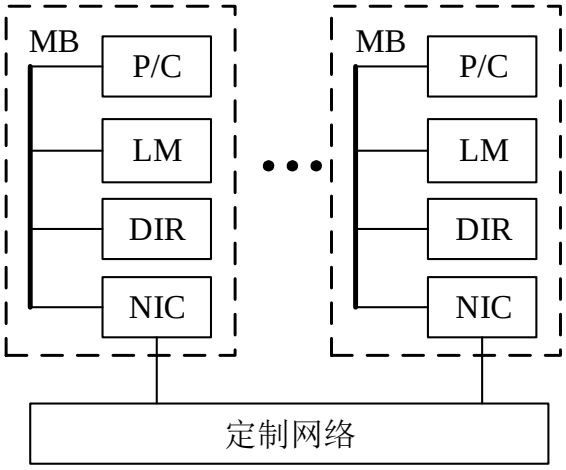
PVP



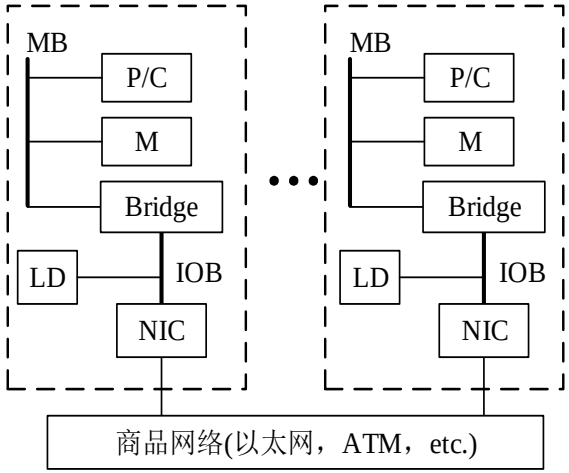
SMP



MPP



DSM



COW

并行计算机体系结构汇总

西安电子科技大学
XIDIAN UNIVERSITY

- 各并行计算机结构特性（比较）

属性	SIMD	PVP	SMP	MPP	DSM	COW
同构性	SIMD	MIMD	MIMD	MIMD	MIMD	MIMD
同步性	指令级同步	异步或弱同步	异步或弱同步	异步或弱同步	异步或弱同步	异步或弱同步
通信机制	数据并行	共享变量	共享变量	消息传递	共享变量	消息传递
地址空间	单空间	单空间	单空间	多空间	单空间	多空间
访存模型	UMA	UMA	UMA	NORMA	NUMA	NORMA
互连网络	定制	交叉开关	总线或交叉开关	定制	定制	商用

并行计算机访存模型 vs. 体系结构

西安电子科技大学 XIDIAN UNIVERSITY

- 常见的并行计算机访存模型

五种常见的并行计算机访存模型

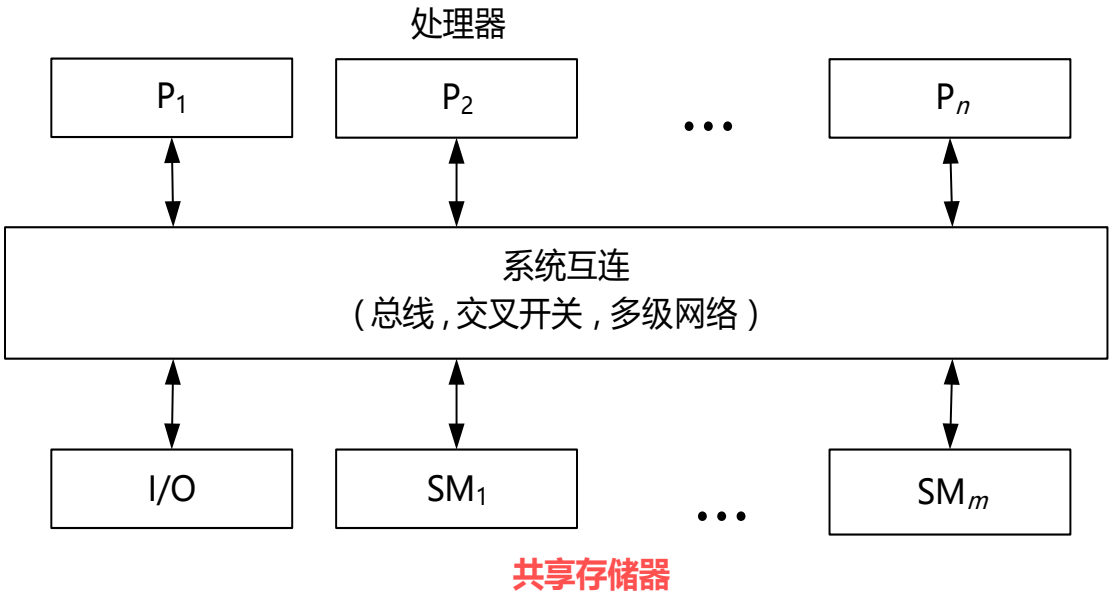
- UMA（Uniform Memory Access），均匀存储访问模型
- NUMA(Non-uniform Memory Access)，非均匀存储访问模型
- COMA(Cache-Only Memory Access)，全高速缓存存储访问
- CC-NUMA（Coherent-Cache Non-uniform Memory Access），高速缓存一致性非均匀存储访问
- NORMA（No-Remote Memory Access），非远程存储访问

结构越来越复杂

并行计算机访存模型

UMA（Uniform Memory Access）

- 均匀存储访问模型



- 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所用时间相同
- 每台处理器可带私有高速缓存

并行计算机访存模型

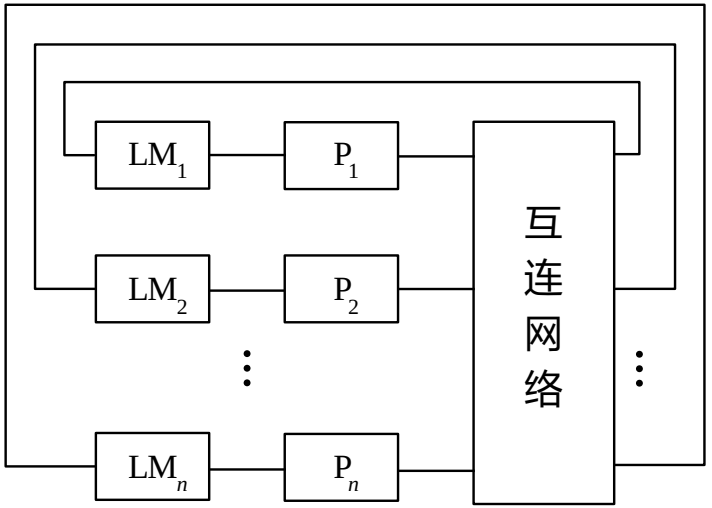
西安电子科技大学 XIDIAN UNIVERSITY

NUMA(Non-uniform Memory Access)

- 非均匀存储访问模型
  - 被共享的存储器在物理上分布在所有的处理器中，所有本地存储器的集合组成了全局地址空间
  - 处理器访问存储器的时间不同
    - 访问本地存储器LM或群内共享存储器CSM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

GSM = Global Shared Memory
CSM = Cluster Shared Memory

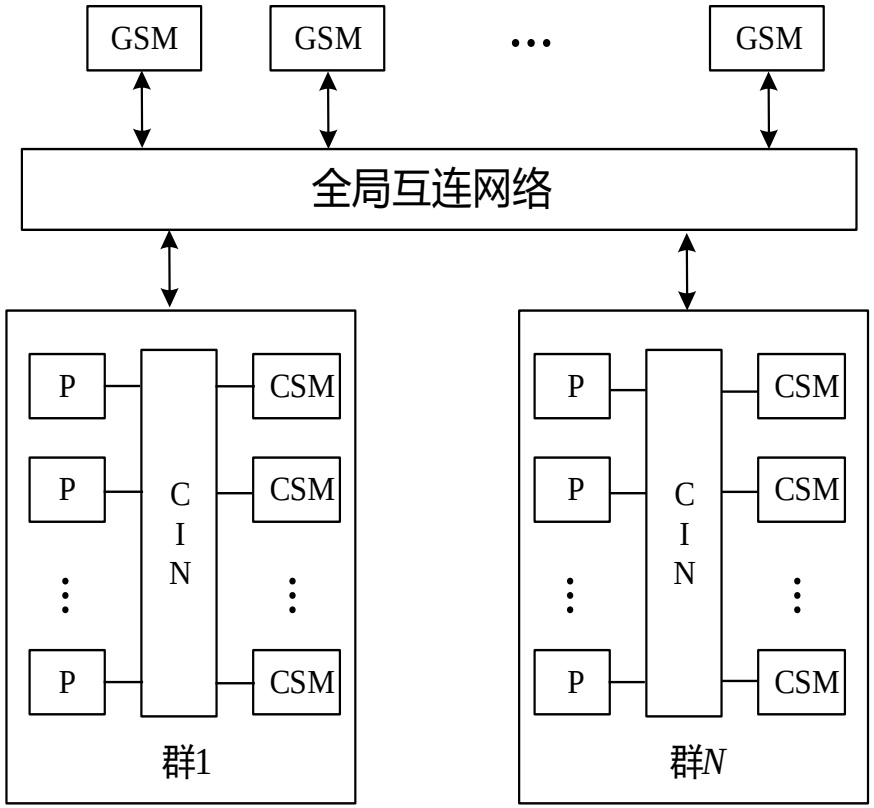
模型一，名词解释 →
LM = Local Memory
P = Processor



共享本地存储模型

并行计算机访存模型

- 非均匀存储访问模型 (NUMA)，模型二



层次式机群模型

模型二，名词解释 →

GSM = Global Shared Memory
CSM = Cluster Shared Memory
CIN = Cluster Interconnection Network
P = Processor

并行计算机访存模型

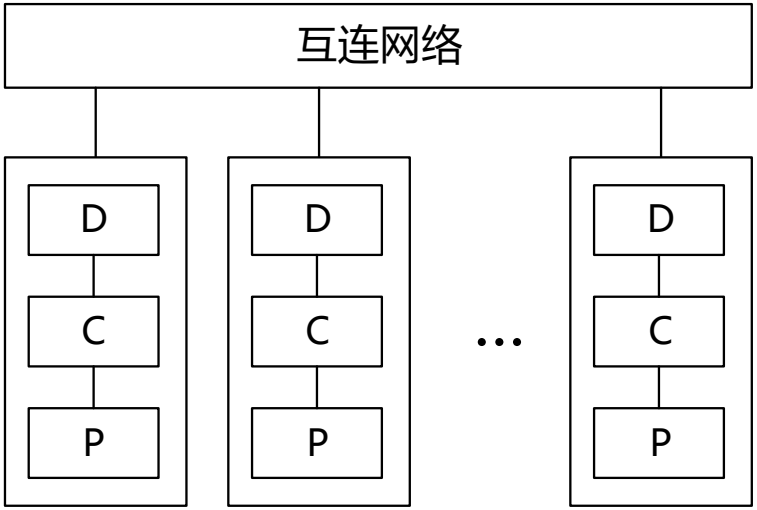
西安电子科技大学 XIDIAN UNIVERSITY

COMA(Cache-Only Memory Access)

- 全高速缓存存储访问
  - 各处理器节点中没有复杂的存储层次结构，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D进行远程高速缓存的访问
  - 使用COMA时，数据开始时可任意分配，因为在运行时它最终会被迁移到需要使用它们的地方

名词解释 →

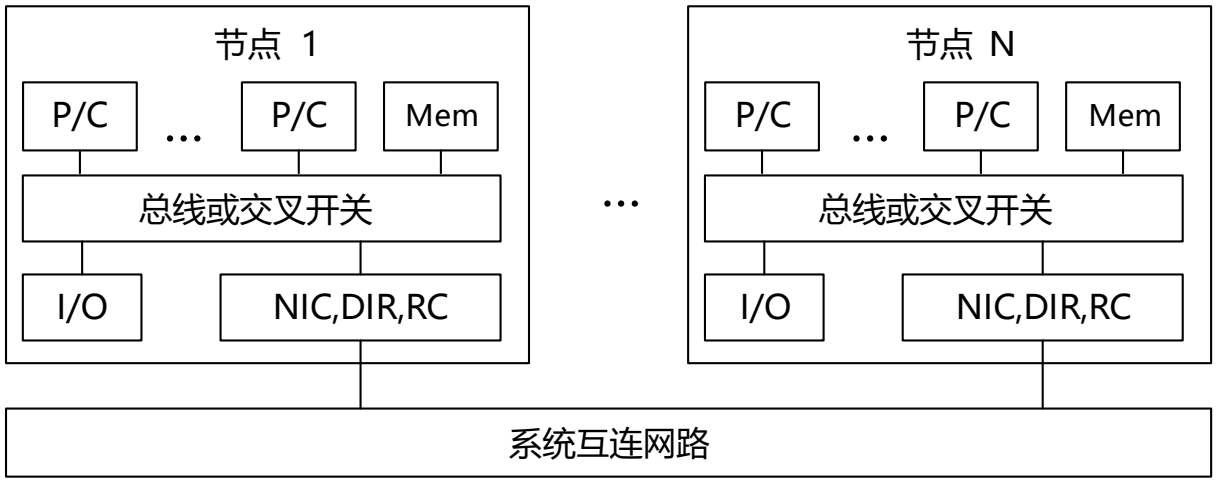
D = Directory
C = Cache
P = Processor



并行计算机访存模型

西安电子科技大学 XIDIAN UNIVERSITY

- CC-NUMA（Coherent-Cache Nonuniform Memory Access）
  - 高速缓存一致性非均匀存储访问模型
    - 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



SMP = Symmetric Multi-processor

DSM = Distributed Shared-Memory

RC = Remote Cache

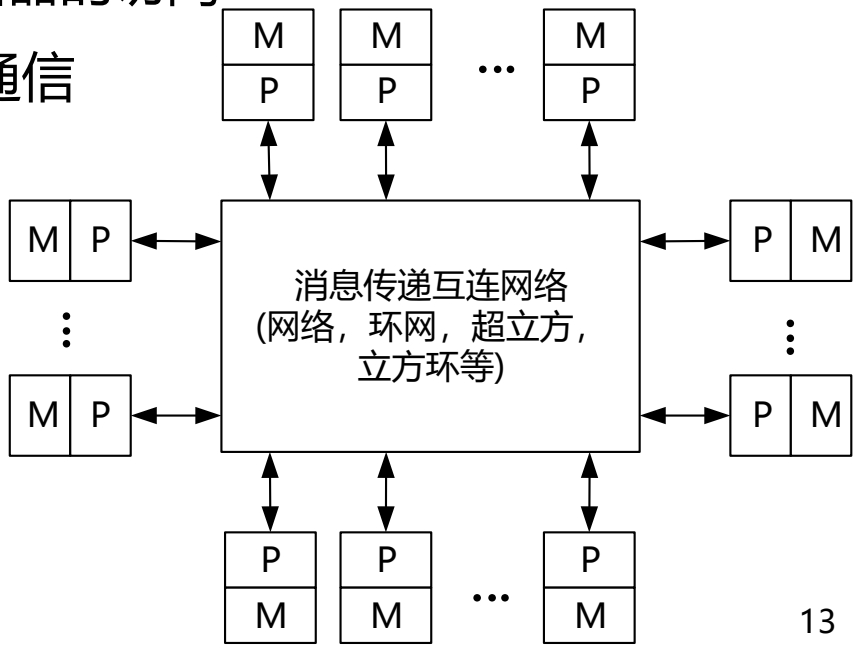
NIC = Network Interface Circuitry

DIR = 高速缓存目录

并行计算机访存模型

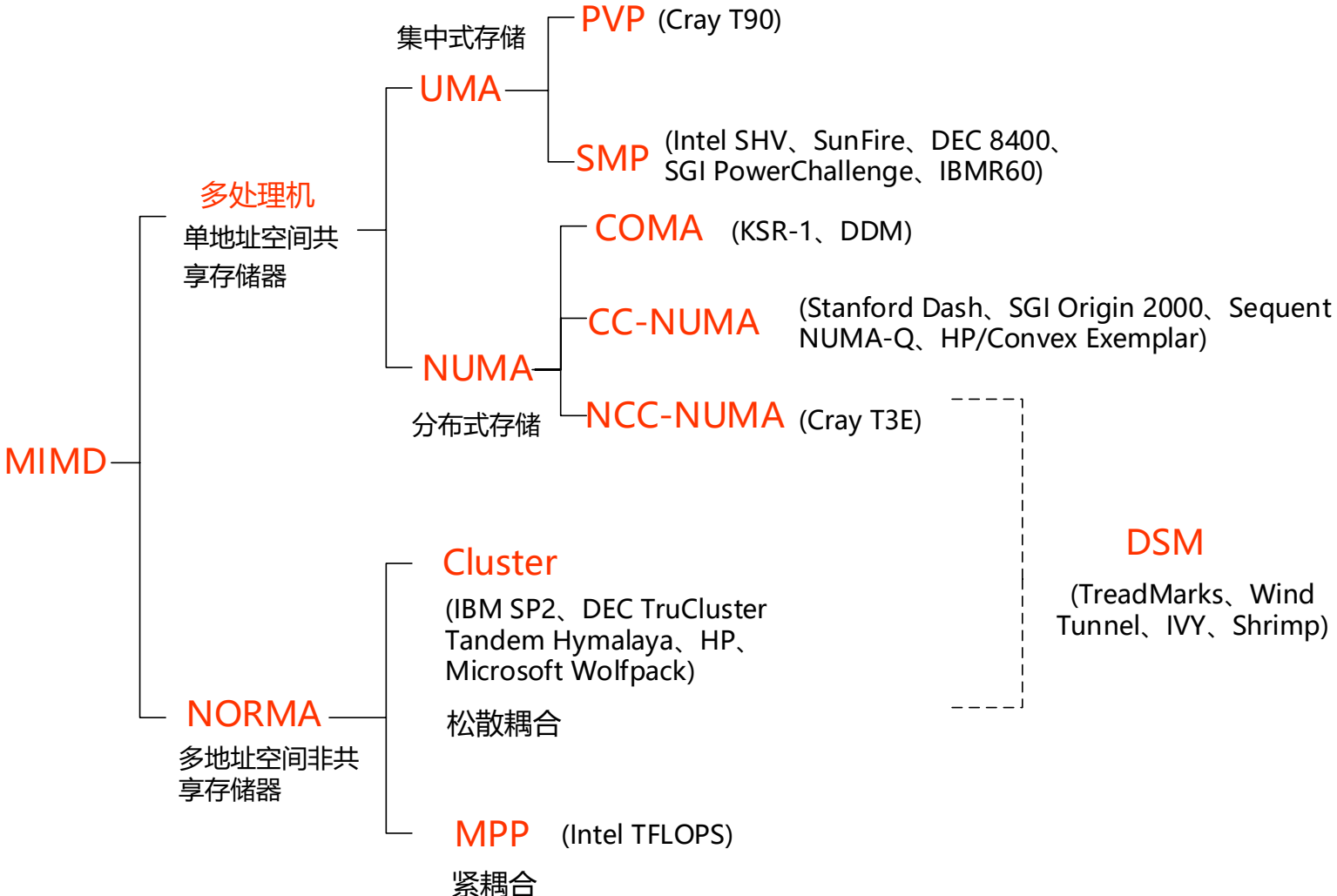
NORMA（No-Remote Memory Access）

- 非远程存储访问模型
  - 所有存储器是私有的，仅能由其所属的处理器访问
    - 对应的典型的体系结构模型：集群
  - 绝大数NORMA都不支持远程存储器的访问
  - 各处理器之间通过消息传递方式通信



并行计算机体系结构概括

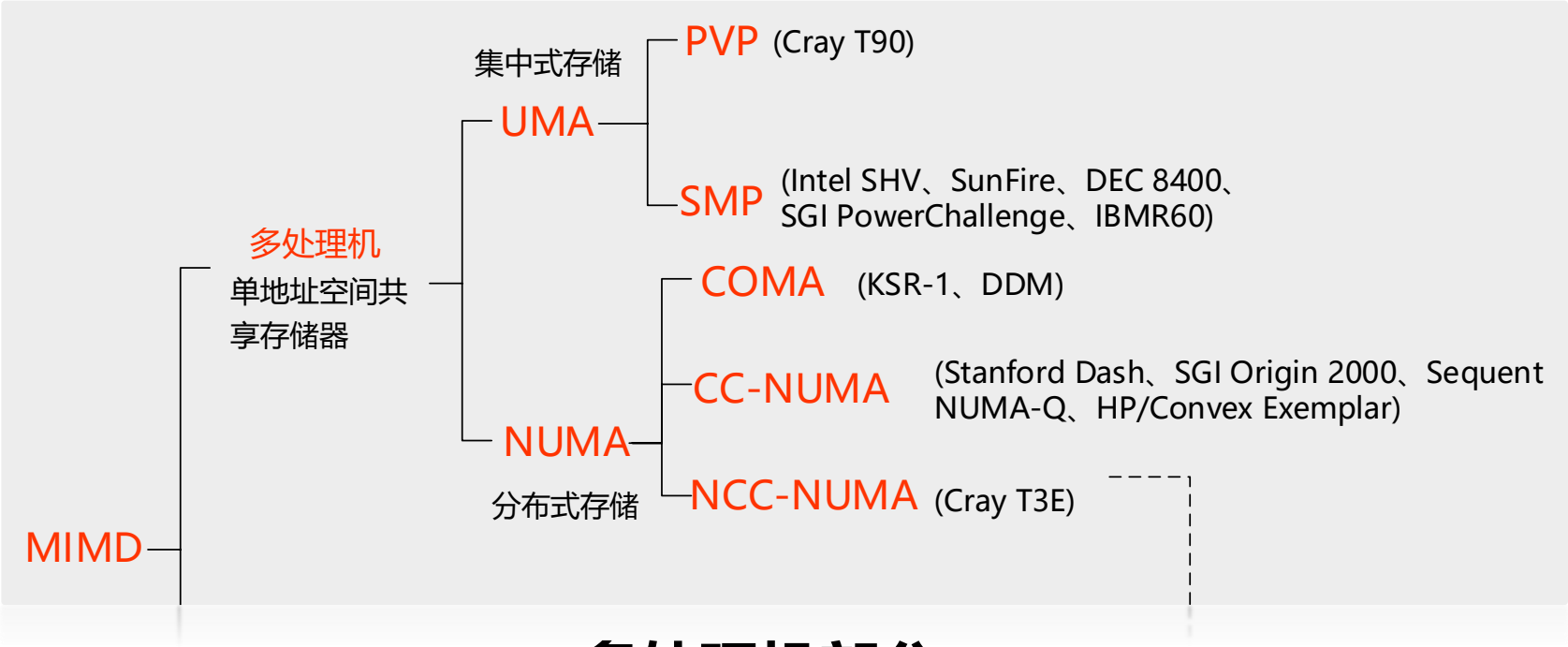
- 并行计算机体系结构与存储结构关系



并行计算机体系结构概括

西安电子科技大学 XIDIAN UNIVERSITY

- 并行计算机体系结构与存储结构关系



多处理机部分

UMA = Uniform Memory Access
NUMA = Non-uniform Memory Access
SMP = Symmetric Multi-processor
PVP = Parallel Vector Processor

并行计算机体系结构概括

- 并行计算机体系结构与存储结构关系

NORMA部分

多地址空间非共享存储 vs. 单地址空间共享存储
NORMA = No-Remote Memory Access
MPP = Massive Parallel processor
DSM = Distributed shared memory

并行计算机体系结构概括

西安电子科技大学 XIDIAN UNIVERSITY

扩展学习资源

- SGI公司的Power Challenge系统
  - 典型的SMP并行计算机
- OptionRed系统
  - Intel与Sandia联合开发，MPP并行计算机
- SGI Origin 2000系统
  - 典型的CC-NUMA结构
- Berkeley的NOW计划
  - 工作站机群的先驱性系统

SGI Power Challenge XL

SGI Origin 2000

§2.2 并行计算机性能测评

并行计算机性能测评

西安电子科技大学 XIDIAN UNIVERSITY

- 并行计算机性能测评概述
  - 硬件层面或机器层面
  - 程序层面或软件层面或算法层面

并行计算机性能测评

西安电子科技大学
XIDIAN UNIVERSITY

- 为什么要进行并行机性能测评？
  - 对管理人员
    - 发挥并行机长处，提高并行机使用效率
  - 对用户
    - 减少用户购机盲目性，降低投资风险
  - 对架构师与设计人员
    - 改进系统结构设计
    - 促进软、硬件结合，合理划分功能
  - 对市场人员
    - 提供客观、公正的评价标准

并行计算机性能测评

- CPU的部分通用性能指标
  - 浮点运算数（FLOPS）
    - 每秒浮点运算次数，floating-point operations per second，FLOPS
    - 原因：
      - 1. 并行计算，特别是科学计算中大量使用浮点数，即小数
      - 2. 浮点数运算比整数运算更花费时间，更有测量的意义与效果
  - 执行时间 *T*
    - T_{comput}：计算时间，T_{paro}：为并行开销时间，T_{comm}：为相互通信时间

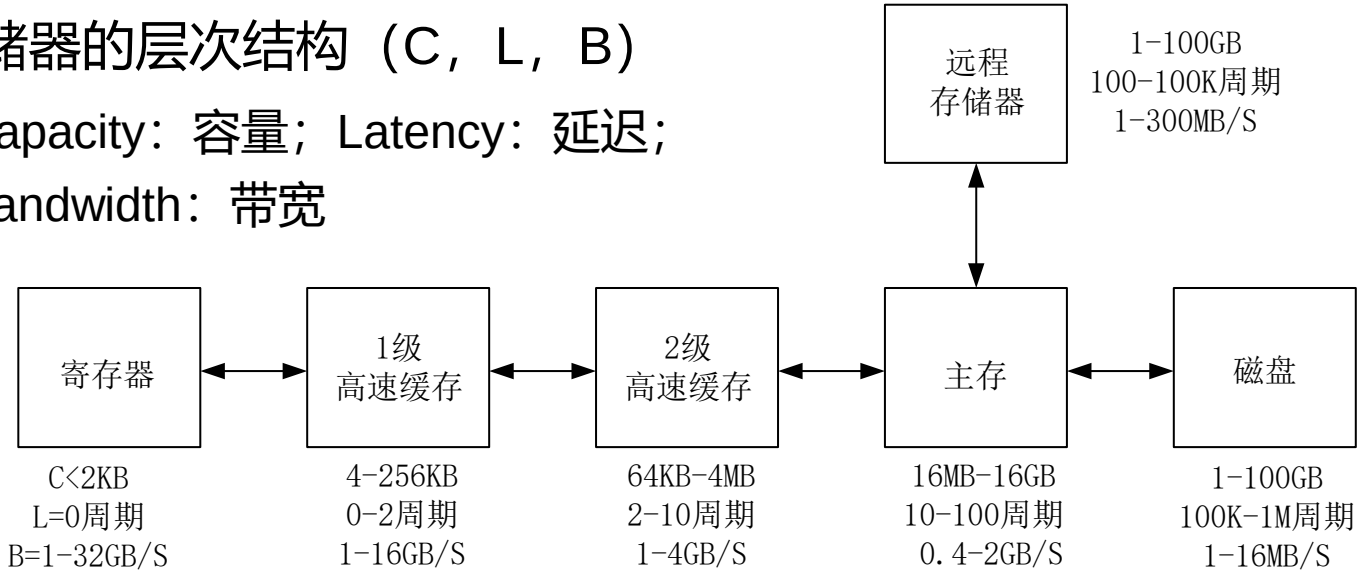
$$T_n = T_{comput} + T_{paro} + T_{comm}$$

并行计算机性能测评

西安电子科技大学 XIDIAN UNIVERSITY

存储器性能

- 存储器的层次结构（C，L，B）
  - Capacity：容量；Latency：延迟；
  - Bandwidth：带宽



- 估计存储器的带宽
  - RISC命令 add r1,r2,r3
  - 8bytes 100MHz
  - B = $3*8*100*10^6$ B/s= 2.4GB/s

注：RISC，Reduced Instruction Set Computing，精简指令集

并行计算机性能测评

西安电子科技大学
XIDIAN UNIVERSITY

- 并行通信开销
  - 多机，多核，并行通信开销相对于计算很大
  - 以PowerPC为例
    - PowerPC
      - 全称：**P**erformance **O**ptimization **W**ith **E**nhanced **R**ISC – **P**erformance **C**omputing，基于精简指令集（RISC）架构的中央处理器
      - 由IBM、Apple等发起，但后被Intel取代；仍为重要的CPU种类
    - 每个周期15ns，执行4flops;
    - 创建一个进程1.4ms，可执行372000flops
  - 开销的测量
    - 乒乓方法（Ping-Pong Scheme）

并行计算机性能测评

- 节点0发送m个字节给节点1；节点1从节点0接收m个字节后，立即将消息发回节点0
- 总的时间除以2，即可得到点到点通信时间，也就是执行单一发送或接收操作的时间

➢ 可一般化为热土豆法（Hot-Potato），也称为救火队法（Fire-Brigade） 0——1 —— 2 —— … —— -n-1 —— 0

■ 热土豆法

➢ Hot-potato routing

➢ 当收到一个分组时，不管它的目的地址如何，一定要以最快的速度转发出去，就像拿到一个烫手的热土豆一样

并行计算机性能测评

西安电子科技大学
XIDIAN UNIVERSITY

```
for i=0 to Runs-1 do
   if  (my _node _id =0)   then            /*发送者*/
            start _time =second()
            send an m-byte message to node 1
            receive an m-byte message from node 1
            end_time = second()
            total_time = end_time – start_time
            communication_time[i] = total_time/2
    else if  (my_node_id = 1)   then      /*接收者*/
            receive an m-byte message from node 0
            send an m-byte message to node 0
    endif
endfor
```

关键字
注释

并行计算机性能测评

西安电子科技大学
XIDIAN UNIVERSITY

- 并行开销的表达式：整体通信
  - 典型的整体通信包括
    - 播送（Broadcasting）
      - 处理器0发送m个字节给所有的n个处理器
    - 收集（Gather）
      - 处理0接收所有n个处理器发来在消息，所以处理器0最终接收了m×n个字节
    - 散射（Scatter）
      - 处理器0发送了m个字节的不同消息给所有n个处理器，因此处理器0最终发送了 $m \times n$ 个字节
    - 全交换（Total Exchange）
      - 每个处理器均彼此相互发送m个字节的不同消息给对方，所以总通信量为 $m \times n^2$ 个字节

并行计算机性能测评

- 循环移位（Circular-shift）
  - 处理器i发送m个字节给处理器i+1，处理器n-1发送m个字节给处理器0，所以通信量为m × n个字节

并行计算在机器上的成本与价格

- 机器的性能/价格比（Performance/Cost Ratio）
  - 指用单位代价（通常以百万美元表示）所获取的性能（通常以MIPS或MFLOPS表示）
  - 两个重要的单位
    - MIPS：每秒百万条指令数，Million Instructions Per Second，侧重衡量整数运算性能
    - MFLOPS：每秒百万条浮点运算次数， megaFLOPS （mega floating point operations per second）
- 利用率（Utilization）：可达到的速度与峰值速度之比

并行计算机性能测评

西安电子科技大学
XIDIAN UNIVERSITY

算法层面/程序层面性能评测

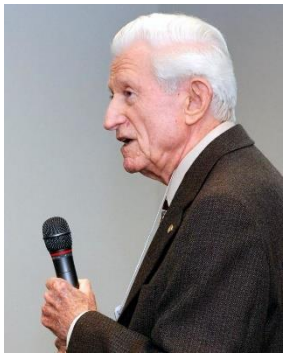
- 开始关注软件，而不是硬件
- 加速比性能评测
  - 并行系统的加速比是指对于一个给定的应用，并行算法（或并行程序）的执行速度相对于串行算法（或串行程序）的执行速度加快了多少倍
  - Amdahl 定律
  - Gustafson定律
  - Sun Ni定律
- 可扩展性评测
  - 等效率度量标准
  - 等速度度量标准
  - 平均延迟度量标准

并行计算机性能测评

西安电子科技大学 XIDIAN UNIVERSITY

算法层面/程序层面性能评测

- 开始关注软件，而不是硬件
- 加速比性能评测
  - 并行系统的加速比是指对于一个给定的应用，并行算法（或并行程序）的执行速度相对于串行算法（或串行程序）的执行速度加快了多少倍
  - Amdahl 定律
  - Gustafson定律
  - Sun Ni定律
- 可扩展性评测
  - 等效率度量标准
  - 等速度度量标准
  - 平均延迟度量标准



Gene Amdahl

John L. Gustafson

并行计算机性能测评

西安电子科技大学
XIDIAN UNIVERSITY

- Amdahl 定律（相等负载）
  - P：处理器数
  - W：问题规模（计算负载、工作负载，给定问题的总计算量等）
  - W_s：应用程序中的串行分量，f是串行分量比例

$$f = W_s / W$$

  - W_p：应用程序中可并行化部分，1- f为并行分量比例

$$W_s + W_p = W$$

  - Amdahl定律的出发点
    - 固定不变的计算负载
    - 固定的计算负载分布在多个处理器上的
    - 增加处理器加快执行速度，从而达到了加速的目的

并行计算机性能测评

西安电子科技大学 XIDIAN UNIVERSITY

Amdahl 定律

- S：加速比 ➔ 如何计算?
- 固定负载的加速公式 $S = \frac{W_s + W_p}{W_s + W_p/p}$ 注：$W_s + W_p = W$，$W_s = W \times f$
- $W_s + W_p$可相应地表示为$W \times (f + (1-f))$

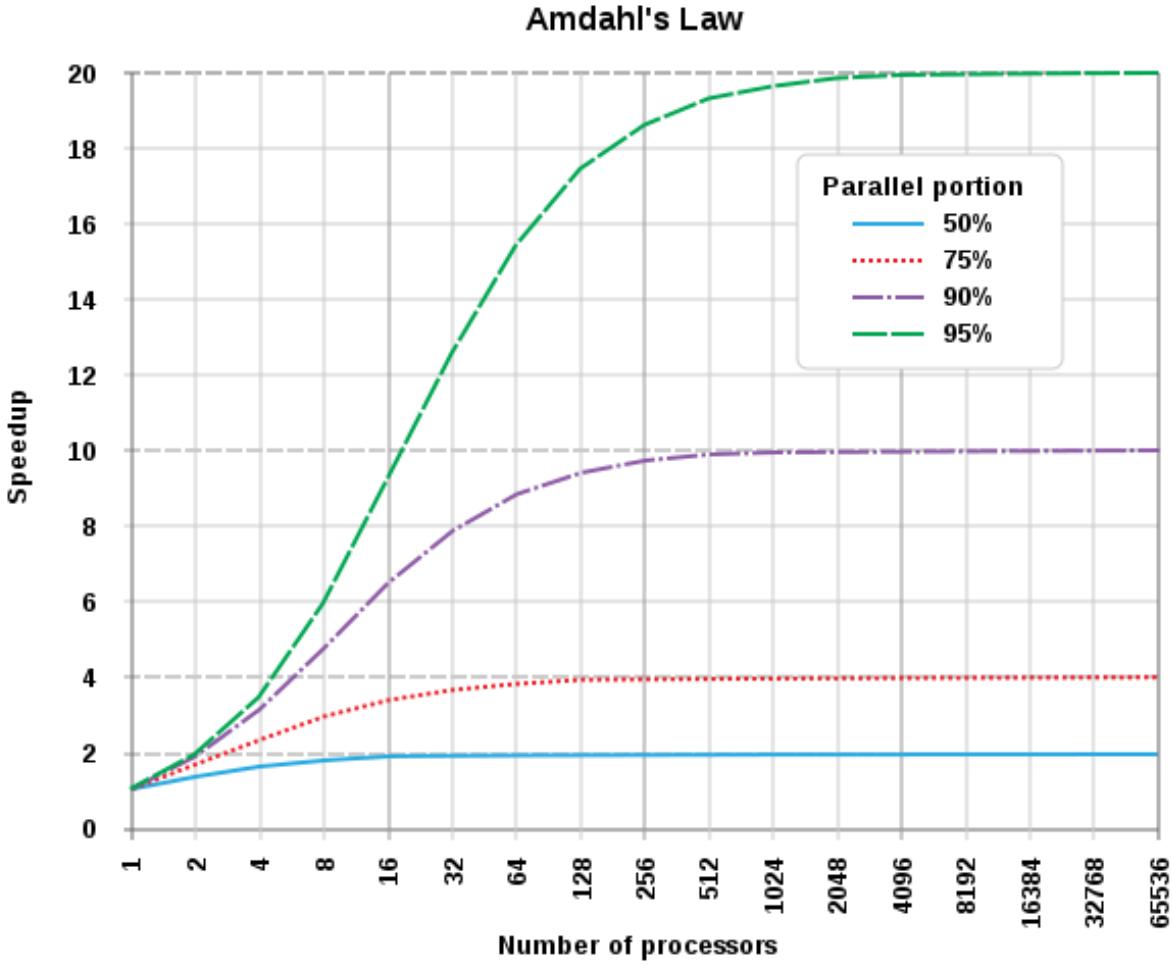
$$S = \frac{f + (1-f)}{f + \frac{1-f}{p}} = \frac{p}{1 + f(p-1)}$$

- $p \to \infty$时，上式极限为 $S = 1/f$
- 这意味着随着处理器数目的无限增大，并行系统所能达到的加速比上限为$= 1/f$，这是不是一个很有利的结论

并行计算机性能测评

西安电子科技大学 XIDIAN UNIVERSITY

- Amdahl 定律



并行计算机性能测评

- 继续考虑并行的额外开销
  - W_o为额外开销

$$S = \frac{W_s + W_p}{W_s + \frac{W_p}{p} + W_O} = \frac{W}{Wf + \frac{W(1-f)}{p} + W_O}$$

$$= \frac{p}{1 + f(p-1) + W_O\, p/W}$$

  - $p \to \infty$时，上式极限为 $S=\frac{1}{f+W_O/W}$
  - 串行分量越大和并行额外开销越大，则加速越小

并行计算机性能测评

西安电子科技大学
XIDIAN UNIVERSITY

- Gustafson定律（相等时间）
  - 出发点
    - 精度要求
      - 对于很多大型计算，精度要求很高，即在此类应用中精度是个关键因素，而计算时间是固定不变的
      - 此时为了提高精度，必须加大计算量，相应地亦必须增多处理器数才能维持时间不变
    - 实际应用
      - 除非学术研究，在实际应用中没有必要固定工作负载而计算程序运行在不同数目的处理器上
      - 增多处理器必须相应地增大问题规模才有实际意义

并行计算机性能测评

Gustafson定律

W_s：应用程序中的串行分量，f是串行分量比例

W_p：应用程序中可并行化部分，$1-f$为并行分量比例

$$S' = \frac{W_s + pW_p}{W_s + p \cdot W_p/p} = \frac{W_s + pW_p}{W_s + W_p}$$

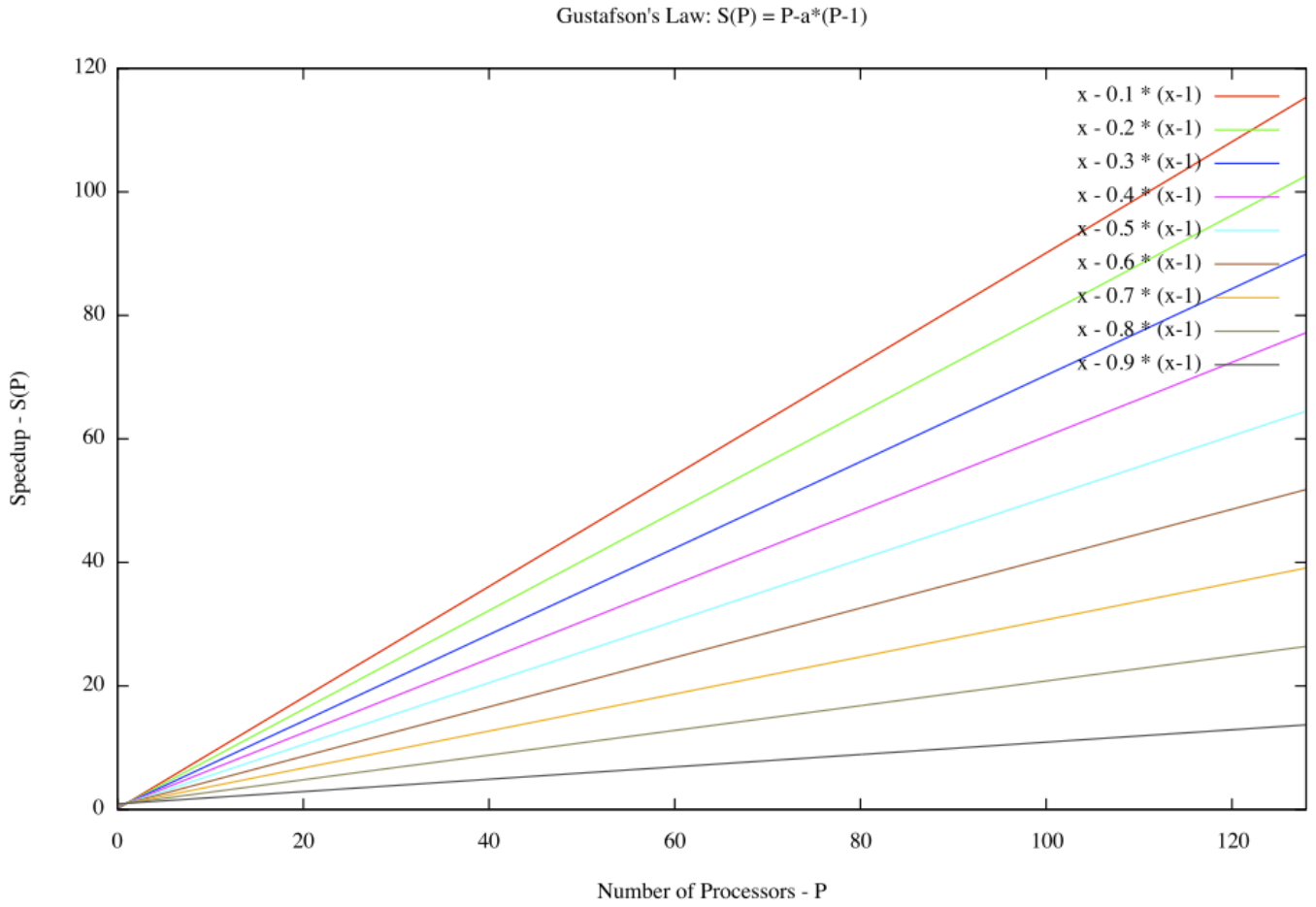
$$S' = f + p(1-f) = p + f(1-p) = p - f(p-1)$$

结论

- 当p充分大时，S'与p几乎成线性关系，其斜率为$1-f$
- 这意味着随着处理器数目的增加，加速几乎与处理器数成比例的线性增加
- 串行比例f不再是程序的瓶颈，这对并行计算来说是一个乐观的结论

并行计算机性能测评

- Gustafson定律



并行计算机性能测评

➔ 并行机的性能指标汇总

西安电子科技大学
XIDIAN UNIVERSITY

名 称	符 号	含 义	单 位
机器规模	n	处理器的数目	无量纲
时钟速率	f	时钟周期长度的倒数	MHZ
工作负载	W	计算操作的数目	Mflop
顺序执行时间	T_1	程序在单处理机上的运行时间	s（秒）
并行执行时间	T_n	程序在并行机上的运行时间	s（秒）
速度	$R_n = W/T_n$	每秒百万次浮点运算	Mflop/s
加速	$S_n=T_1/T_n$	衡量并行机有多快	无量纲
效率	$E_n = S_n/n$	衡量处理器的利用率	无量纲
峰值速度	$R_{peak} = n R'_{peak}$	所有处理器峰值速度之积，R'_{peak}为一个处理器的峰值速度	Mflop/s
利用率	$U = R_n/R_{peak}$	可达速度与峰值速度之比	无量纲
通信延迟	t_o	传送0-字节或单字的时间	Ms
渐近带宽	r_∞	传送长消息通信速率	MB/s

并行计算机性能测评

西安电子科技大学
XIDIAN UNIVERSITY

- 并行计算的可扩展性（Scalability）
  - 影响加速比的因素
    - 处理器数与问题规模
    - 求解问题中的串行分量
    - 并行处理所引起的额外开销
      - 通信、等待、竞争、冗余操作和同步等
  - 可扩展性
    - 在确定的应用背景下，计算机系统（或算法或程序等）性能随处理器数的增加而按比例提高的能力
  - 可扩展性的内涵
    - 增加处理器数会增大额外开销和降低处理器利用率

并行计算机性能测评

西安电子科技大学
XIDIAN UNIVERSITY

- 额外开销的增加可能慢于有效计算的增加
- 算法中的串行分量比例不是固定不变的
  - 串行部分所占的比例随着问题规模的增大而缩小

■ 并行计算中可扩展性的作用

- 确定解决某类问题用何种并行算法与何种并行体系结构的组合
  - 可以有效地利用大量的处理器
- 对固定的问题规模，确定在某类并行机上最优的处理器数与可获得的最大的加速比
- 用于指导改进并行算法和并行机体系结构
  - 以使并行算法尽可能地充分利用可扩充的大量处理器

并行计算机性能测评

西安电子科技大学
XIDIAN UNIVERSITY

- 可扩展性：如何衡量
  - 并行计算要调整的是处理数p和问题规模W
  - 两者可按不同比例进行调整，此比例关系可能是线性的，多项式的或指数的等）就反映了可扩展的程度
- 目前没有一个公认的，或者被普遍接受测评标准
  - 常用的可扩展性测评标准（三个）
    - 等效率度量标准
    - 等速度度量标准
    - 平均延迟度量标准

第一次作业（已布置）

西安电子科技大学
XIDIAN UNIVERSITY

题目

- 题目1. 从可扩展性上比较SMP与MPP两种体系结构
  - 可扩展性的角度包括：内存大小可扩展性、处理器数目可扩展性、磁盘容量可扩展性等，至少选取一个
- 题目2. 试说明（可举例，也可论证）：如何使用SIMD的机器上实现MIMD并行应用?
  - 即假设你身边有的只是SIMD的机器，但任务却是要实现一个并行应用，那是否有办法可完成?

第一次作业（已布置）

西安电子科技大学 XIDIAN UNIVERSITY

- 要求
  - 独立完成，发送电子版，可以是word，也可以pdf
  - 作业提交地址
    - xdsepc2018@163.com
    - 附件命名规则：“学号+姓名+作业一”
    - 时间：两周之内，至第四次上课之前提交
    - 篇幅：没有绝对的页数要求，但是两道题目一般不应少于1页A4纸的2/3
  - 课件地址
    - http://web.xidian.edu.cn/ysxu/teach.html

课件地址

西安电子科技大学
XIDIAN UNIVERSITY
徐悦甦的主页： 个人信息 科学研究 论文成果 荣誉奖励 科研团队 课程教学 招生信息 English
课程教学
1. 承担课程与竞赛：
主讲：《信息检索》
《信息检索》课件：
上机报告、课堂作业与课程大作业说明：/ysxu/files/20170609_152043.pdf
§ 4.1 文本聚类：/ysxu/files/20170427_104024.pdf
§ 4.2 文本分类：/ysxu/files/20170506_135325.pdf
§ 5 推荐系统：/ysxu/files/20170516_183350.pdf
§ 6 语义网：/ysxu/files/20170530_200219.pdf
§ 7 课程回顾与复习：/ysxu/files/20170609_165935.pdf
并行计算课件
§ 1 课程说明与并行计算概述：ppt (点击直接下载)

西安电子科技大学
XIDIAN UNIVERSITY

§2.2 主流并行计算机系统与性能测评

http://web.xidian.edu.cn/ysxu/

结构 | 编程 | 算法 | 应用