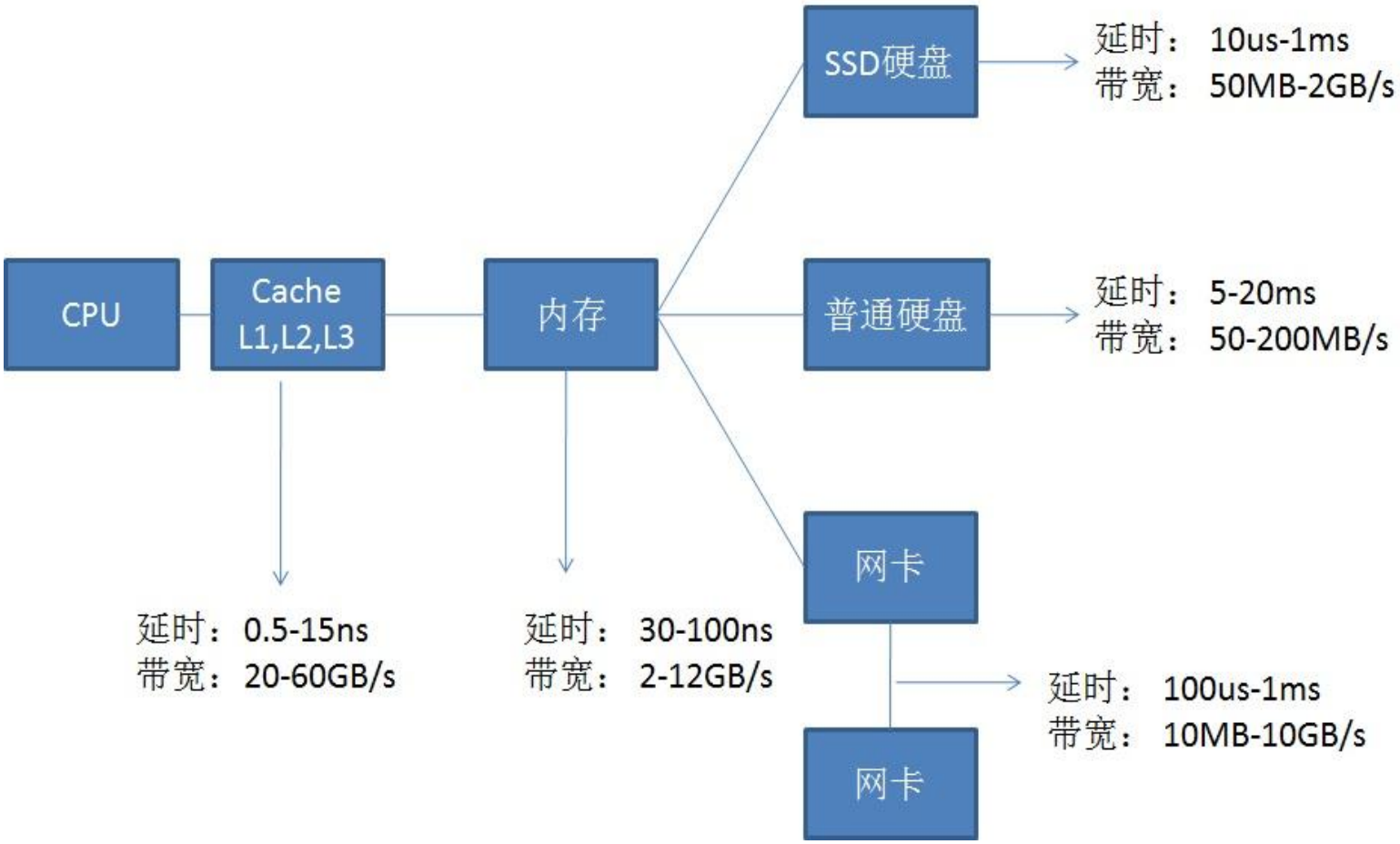
硬件体系结构

——关注硬件分类，性能

何登成
微博：@何_登成

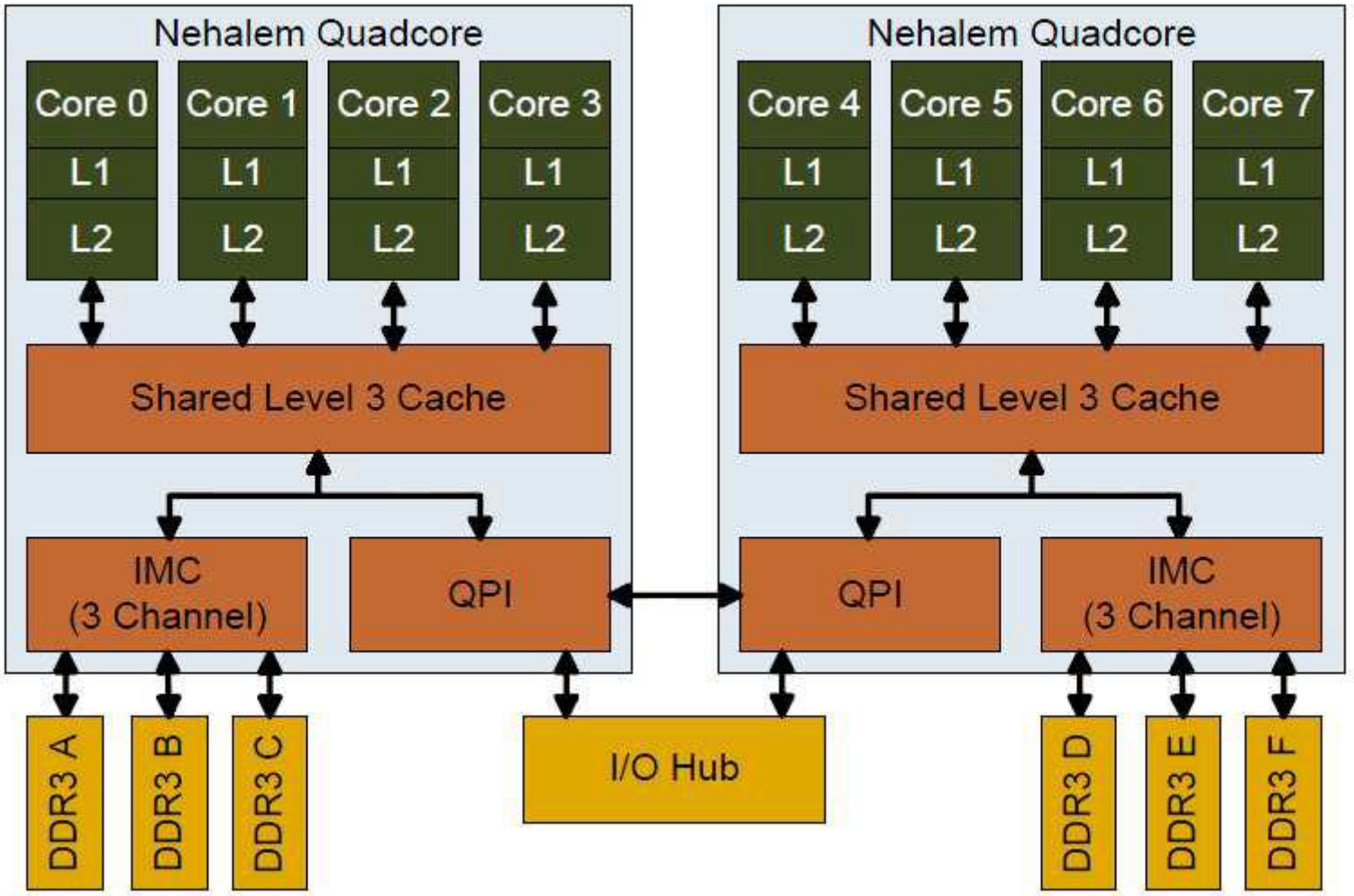
硬件体系结构及性能


CPU
Cache
L1,L2,L3
内存
SSD硬盘
普通硬盘
网卡
网卡
延时： 0.5-15ns
带宽： 20-60GB/s
延时： 30-100ns
带宽： 2-12GB/s
延时： 10us-1ms
带宽： 50MB-2GB/s
延时： 5-20ms
带宽： 50-200MB/s
延时： 100us-1ms
带宽： 10MB-10GB/s

Outline

- CPU
- 内存
- HDD/SSD
- 网卡
- RAID卡
- 虚拟化

CPU-Architecture

- 

CPU-组件与性能指标

- CPU
  - 主频：CPU的时钟频率，内核工作的时钟频率
  - 外频：系统总线的工作频率
  - 倍频：CPU外频与主频相差的倍数
  - 前端总线：将CPU连接到北桥芯片的总线
  - 总线频率：与外频相同或者是外频的倍数
  - 总线数据带宽：(总线频率 * 数据位宽) / 8

- L1,L2,L3 cache (缓存数据与指令)
  - L1,L2: core独占； 带宽：20-80GB/S；延时：1-5ns
  - L3：core之间共享； 带宽：10-20GB/S；延时：10ns
  - Cache line size：64 Bytes

- Interface
  - QPI
    - Intel中，连接一个CPU中的多个处理器(processors)，直接互联
    - QPI带宽：~20GB/s

CPU-程序设计优化

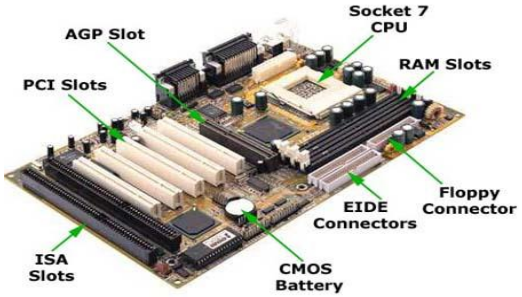
- Cache-conscious
  - 目的：提高L1/L2/L3 cache命中率，降低访问内存的概率，降低系统响应时间
  - 手段：
    - 降低Cache冲突概率
    - 提高Cache利用率——数据结构重整

- False-sharing
  - 多线程程序，当不同的线程同时读写同一cache line上的不同数据时发生；
  - 危害：False-sharing会导致L1/L2 cache miss，增加系统延时
  - 例如：int threadArray[2];

- 系统监控
  - 对程序做性能测试时，善用系统性能监控命令，定位瓶颈
  - 例如：top

Memory

- Memory
  - Cpu与外部沟通的桥梁，计算机所有程序在内存中运行。其作用是用于暂时存放cpu中的运算数据，以及与硬盘等外部存储器交换的数据.(From 百度百科)



- Memory-种类
  - RAM：　随机存取存储器
  - SRAM：　静态随机存储器　CPU Cache
  - DRAM：　动态随机存储器
  - SDRAM：　同步动态随机存储器　与谁同步？——理论上速度可达到与CPU同步
  - DDR SDRAM：　双倍数据传输率SDRAM　DDR DDR2 DDR3

- Memory特性
  - 随机定位；易失存储；容量有限(单块容量小，插槽少)

Memory-性能指标

- Memory-性能指标
  - 核心频率(F1)：内存的工作频率
    - 倍增系数：(预读位宽/2) * 2(上升下降沿均可传输)
  - 时钟频率(F2)：核心频率通过倍频技术得到(倍频，既为预读位宽，基准为2 bits)
    - DDR：F2 = F1；DDR2：F2 = 2F1；DDR3：F2 = 4F1
    - F2 = F1 * 倍增系数 / 2
  - 数据传输频率(F3)：传输数据的频率
    - DDR：F3 = 2F1；DDR2：F3 = 4F1；DDR3：F3 = 8F1
    - F3 = F1 * 倍增系数
  - Throughput
    - Throughput = F1 * (内存总线位数/8) * 倍增系数 = F3 * (内存总线位数/8)
  - Latency
    - 30 – 100 ns
- DDR400 DDR3-800 (single channel, 64-bit)
  - 400 800？
    - 400，800数字，指的是数据传输频率F3；对应的F1为200，100
  - 带宽
    - DDR400：　200 * 64/8 * 2 = 400 * 64/8 = 3.2GB/s
    - DDR3-800：100 * 64/8 * 8 = 800 * 64/8 = 6.4GB/s
  - 双通道
    - 带宽 * 2

Memory-性能指标(续)

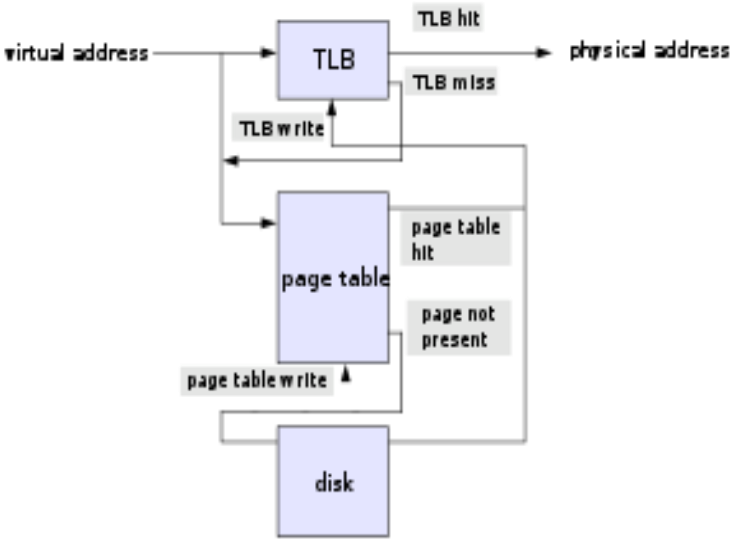
	Read	Write	Copy	Latency
Memory	5174 MB/s	3968 MB/s	3983 MB/s	88.4 ns
L1 Cache	51041 MB/s	50910 MB/s	101816 MB/s	0.9 ns
L2 Cache	21581 MB/s	13098 MB/s	21221 MB/s	6.0 ns
L3 Cache				

CPU Type	DualCore Intel Pentium (Wolfdale-2M, LGA775)		
CPU Clock	3192.0 MHz (original: 3200 MHz)		
CPU FSB	199.5 MHz (original: 200 MHz)		
CPU Multiplier	16x	CPU Stepping	R0
Memory Bus	399.0 MHz	DRAM:FSB Ratio	12:6
Memory Type	Single Channel DDR3-800 SDRAM (6-6-6-15 CR1)		
Chipset	Intel Eaglelake G41		
Motherboard	Unknown		

EVEREST v5.50.2100 / BenchDLL 2.5.292.0 (c) 2003-2010 Lavalys, Inc.

Memory-定位

- 虚拟地址(Virtual Address)
  - 为了区分不同进程的存储空间，每个进程的虚拟地址空间是连续的，但对应的物理地址空间不一定连续
  - 32位系统：32位指针 = 4G Bytes
  - 64位系统：64位指针 = 16E Bytes
- 页表(Page Table)
  - 虚拟地址到物理地址的映射表
- TLB(Translation lookaside buffer)
  - 缓存Virtual address 到Physical address的映射关系，加快查找
- 大页(Huge pages)
  - 降低页表空间消耗
  - 固定内存空间，防止swap



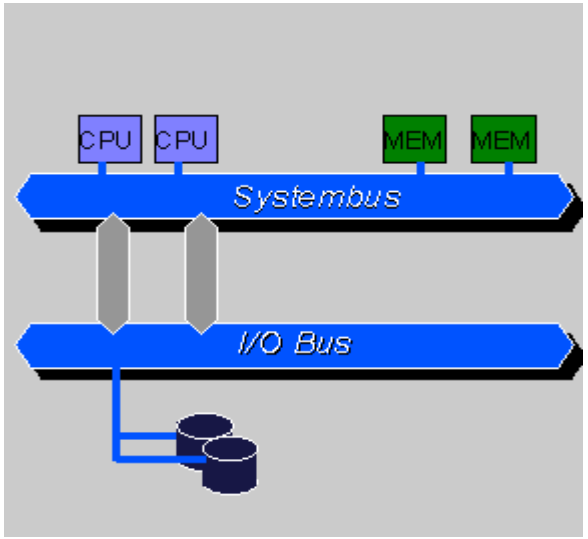
Memory-编程指导

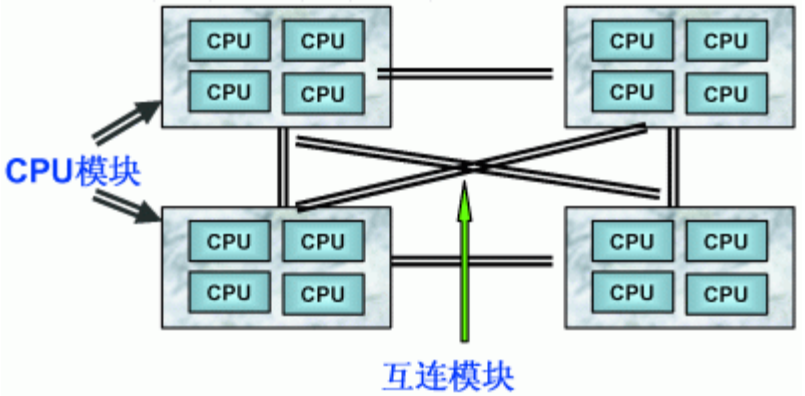
- Memory conscious
  - 所有提供L1/L2/L3 cache命中率的方法，在memory中同样适用(Memory Latency << HDD)
    - 经常访问数据，常驻内存

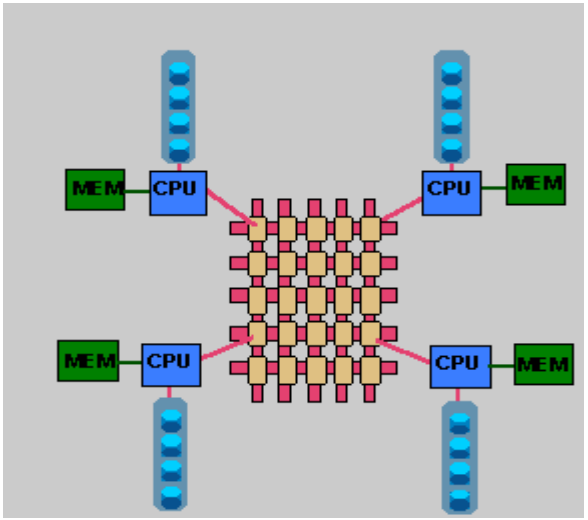
服务器体系结构

- SMP/UMA - Symmetric Multi Processing/Uniform Memory Architecture
  - 服务器中多CPU对称工作，无主次关系。各CPU共享相同的物理内存，访问内存任何地址所需时间相同。程序设计简单。
  - 缺点：Scale-up，难以扩展；内存访问冲突
- NUMA-Non-Uniform Memory Access
  - 多CPU模块，每个CPU模块具有独立的本地内存(快)，但可访问其他CPU内存(慢)，共享存储。代表：HP Superdome，IBM P690
  - 缺点：全局内存访问性能不一致；程序设计需要特殊考虑。
- MPP-Massive Parallel Processing
  - 由多个SMP服务器通过节点互联网络连接而成，每个节点访问本地资源(内存、存储等)，完全无共享(Share-Nothing)。最易扩展，软件层面即可实现。代表：网易DDB
  - 缺点：数据重分布；程序设计复杂；

服务器体系结构(cont.)



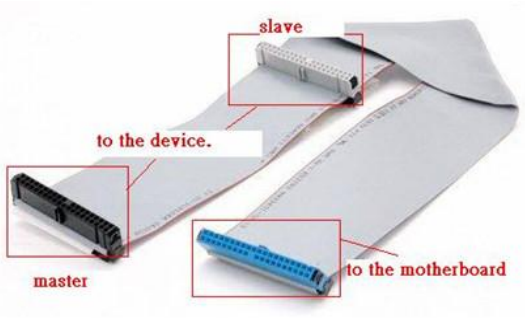


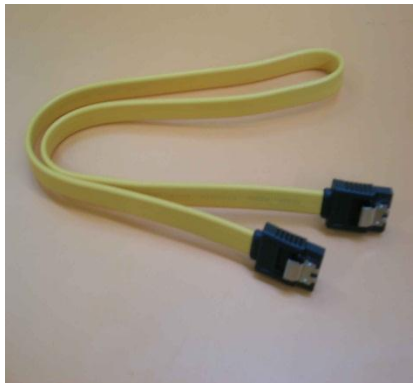


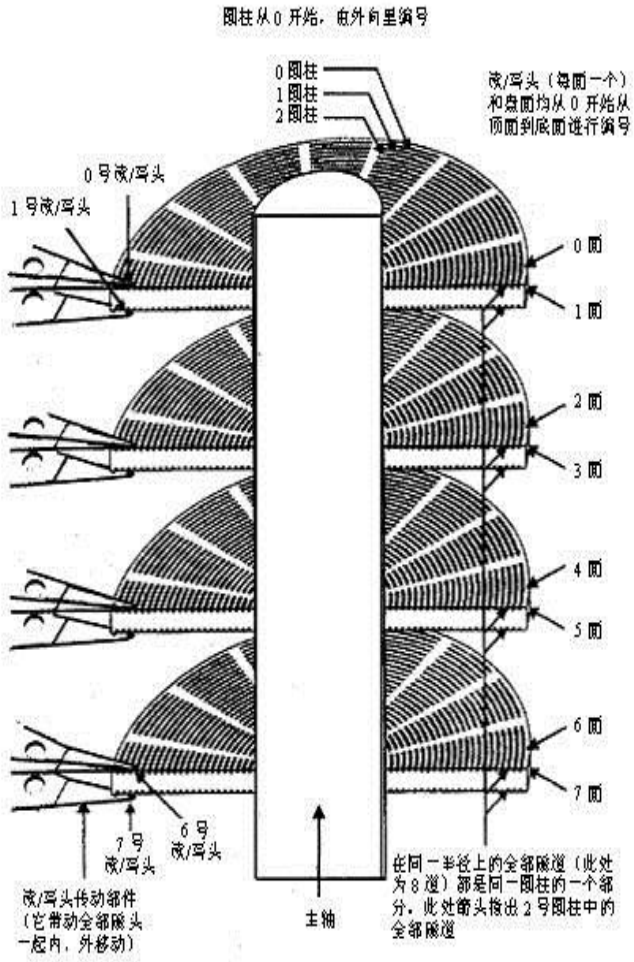

```
# numactl --hardware
available: 2 nodes (0-1)
node 0 size: 32276 MB
node 0 free: 26856 MB
node 1 size: 32320 MB
node 1 free: 26897 MB
node distances:
node   0   1
  0:  10  21
  1:  21  10
```

HDD/SSD-HDD概述

- 硬盘接口
  - ATA
    - IDE(PATA, parallel ATA), SATA(Serial ATA)
  - SCSI
    - SCSI(parallel), SAS(serial)
  - FC(Fibre Channel)
    - FC-SCSI (serial)
- 基本参数
  - 容量
  - 转速
  - 平均访问时间
  - 传输速率
    - 内部传输率(data transfer rate)
    - 外部传输率(sustained transfer rate)



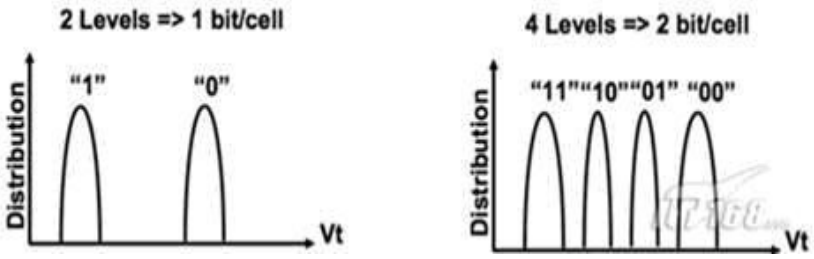




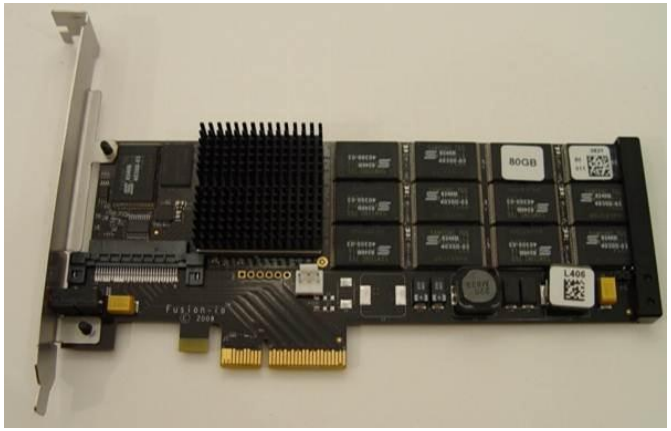
- 数据线：IDE VS SATA

HDD/SSD-SSD概述

- 原理
  - 电学存储介质，对浮动栅极的充电状态标识0或1
- 存储技术
  - NAND， NOR
  - SLC，MLC
    - 电压分区粒度
    - 下图，横坐标为电压，纵坐标为值



- 接口类型
  - SATA，SAS，FC，PCIe(高速接口)
- SSD特性
  - 粒度划分：page 4K； block n*pages (256K)
  - Read/Write， page
  - Erase， block
  - 更新 = read + erase + write



HDD-性能指标

- HDD-SAS 15K RPM
  - 基本性能指标
    - 平均寻道时间(E)：~4ms
    - 旋转延时(L)：2ms
    - 内部传输时间(X)：0.8ms
    - 吞吐率(Throughput)：Read 140MB/s Write 140MB/s
    -
  - 磁盘服务时间
    - Rs = E+L+X = 6.8ms
  - IOPS
    - 147 iops
    -
  - IO响应延时
    - R = Rs / (1 - U) ；I/O利用率越高，响应延时越大
    - U：I/O利用率
- HDD分析
  - 顺序读写快，随机读写慢
  - 随机读写，性能稳定
  - 高容量，价格便宜

HDD-性能指标(续)

RPM Rotations Per Minute	Rotations Per Second	Rotations Per Mili-second	Full Rotation	Rotational Latency (Half Rotation)	Average Seek Time	IO Time	IOPS
(x)	*(x/60)*	*(x/60,000)*	*(1/ [x/60000])*	*(1/ [x/60000]) / 2*			
				Y	Z	(Y+Z)	(1/[Y+Z])*1000
15,000	**15,000/60**	**15,000/ 60,000**	**4ms**	**2ms**	**4ms**	**6ms**	**167**
10,000	**10,000/60**	**10,000/ 60,000**	**6ms**	**3ms**	**5.15ms**	**8.15ms**	**122**
			10ms	**5ms**	**9ms**	**14ms**	**71**
7,200	**7200/60**	**7,200/ 60,000**	**8.4ms**	**4.2ms**	**9.9ms**	**14.1ms**	**71**

SSD-性能指标

- SSD性能指标
  - IOPS
    - 随机读：35000(intel x25-e)；　　120000(Fusion-io ioDrive)
    - 随机写：3300(intel x25-e)；　　90000(Fusion-io)
  - Throughput
    - 连续读：250MB/s(x25-e)；750MB/s(Fusion-io)
    - 连续写：170MB/s(x25-e)；500MB/s(Fusion-io)
  - Latency
    - read：　75us(x25-e)；　　26us(Fusion-io)
    - write：200us(Fusion-io)
    - erase：~2ms
- SSD分析
  - 高IOPS，低IO延时
  - 体积小，省电

SSD-Erase

- Erase影响
  - 更新 = read + erase(block) + write = 写放大
  - erase代价高，latency = 2ms
  - erase影响write性能
    - ssd随机写性能较差
  - erase次数有限(wear-out)
    - SLC：10万次erase；MLC：1万次erase
- SSD层面优化
  - FTL(Flash Translation Layer)：物理逻辑地址映射
  - Reclamation：异步擦除策略，降低延时
  - Wear Leveling：均衡写磨损，提升寿命
  - Spare Area：预留空间，减少写放大

HDD/SSD-编程指导

- HDD
  - 高延迟
    - 热点常驻内存
    - 随机写 ——>连续写：消除 **寻道+旋转** 延时

- SSD
  - 随机写慢
    - 随机写转换为连续写：
      - 数据库：write data ——> write log
      - Fractal Tree：无随机写，转换为顺序写+随机读
    - 缓存，合并写操作：
      - 内存cache，定期回刷，合并期间写操作
  - 写放大
    - 控制每次写入大小

- 性能监控
  - 程序完成，做性能测试过程中，用监控命令定位系统瓶颈
  - 例如：iostat

Raid/Raid卡

- Raid定义
  - 独立磁盘冗余数组(Redundant Array of Independent Disks)
  - 使用廉价磁盘提供快速、可靠的存储。

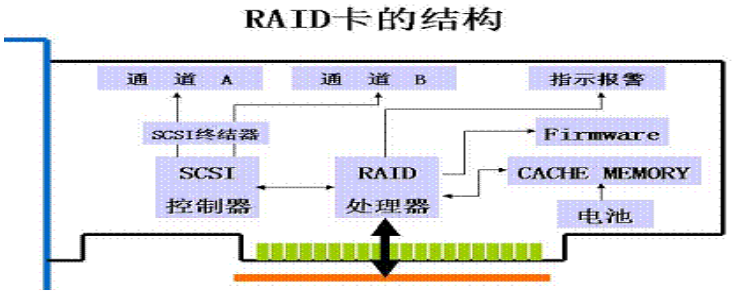
- Raid功能
  - 增强数据集成度
  - 增强容错功能
  - 增加处理量或容量

- Raid分类(常用)
  - Raid0(条带)；raid1(镜像)；raid10；raid5；raid6；raidz；
    - 120块 15K转速的HDD盘，在raid10，raid5下分别能够提供多少iops？
    -

- Raid支持
  - 软件raid
  - 硬件raid：raid卡

Raid/Raid卡(cont.)

- Raid卡
  - 有自己的cpu，cache memory，通过集成或借用主板上的scsi控制器管理硬盘，输出到主机的称之为逻辑单元(logical units)



- Raid卡基本参数
  - 连接主机 host bus adapter(HBA，HBA卡)
  - 后端接口(Back-end interface)：IDE，SATA，SCSI，FC，SAS
  - 前端接口(Front-end interface)：SATA，SCSI，FC，ISCSI …

- Cache Memory
  - 预读：read ahead；pre-fetch
  - 回写：write-back；write-through

  - BBU(Battery Backed Unit)
    - 电池保护的write cache。保护cache memory在断电时数据不丢失。
    - Write-through **vs** write-back + BBU？

网络/网卡

- OSI七层模型
  - 自下而上：物理层；数据链路层；网络层；传输层；会话层；表示层；应用层
- 网络(企业级)
  - 10GbE
    - 10 Gigabit Ethernet
  - Infiniband
    - Infiniband SDR(1X, 4X); Infiniband DDR(1X, 4X);
    - RDMA – Remote direct memory access
  - Fibre Channel

- 延时/带宽/传输距离
  - 10GbE
    - ~15 usec / 10Gbit/s /
  - Infiniband(QDR 4X)
    - ~1-2 usec / 32Gbit/s / 17m(双绞铜线)；数公里(光缆)
  - Fibre Channel

- 网络延时远远小于HDD延时(也小于SSD延时)，使得RAC，Exadata等设计成为可能

网络/网卡

- 网卡
  - 功能
    - 发送：数据封装为帧；接收：接收帧，并将帧重新组合数据。
  - 种类
    - NIC (Network Interface Controller)： Ethernet
    - HBA (Host Bus Adapter)： Fibre Channel
    - HCA (Host Channel Adapter)： Infiniband
- 交换机(switch)
  - 以太网交换机；光纤交换机；Infiniband交换机
- 关键技术
  - 多路径
    - 增加系统的性能，提供网络层面高可用(HA)
  - Zoning(交换机) vs LUN masking(网卡)
    - 网络区域隔离，增加系统的稳定性以及安全性
  - MTU
    - 最大传送单位 (Maximum Transmission Unit)。默认：1500 bytes

网络编程

- AIO
  - 目的：
  - 方法：
    - Select
    - Poll
    - Epoll

网络补充-大融合

- 融合
  - 所有的网络传输协议，都可以运行在其他网络硬件层面之上。
  - 有效融合各网络优势
  - 减少硬件投入成本

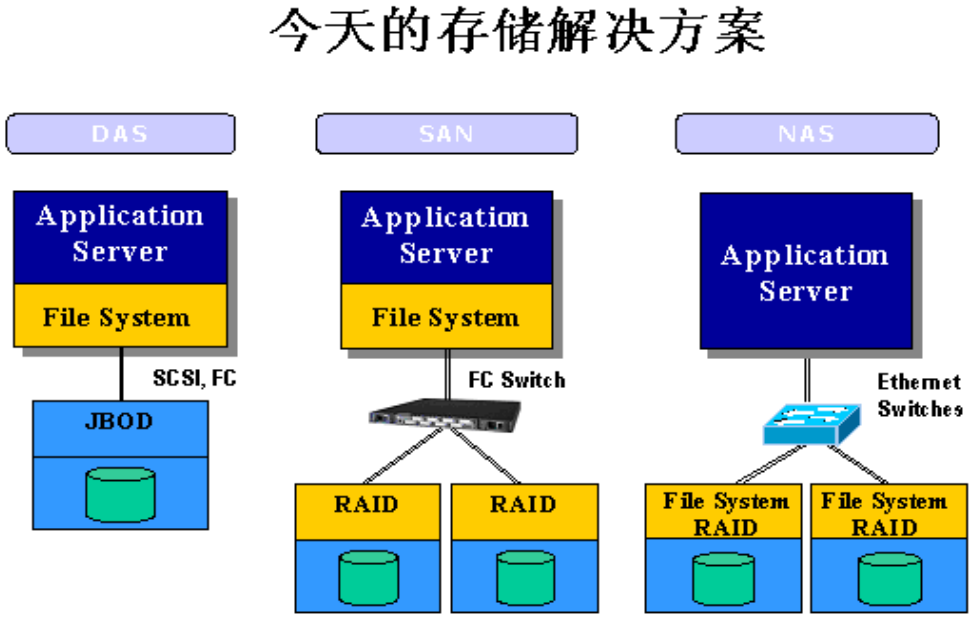
- 分类
  - Ethernet
    - iSCSI(scsi over ethernet); FCOE(Fibre Channel over Ethernet); FCIP; IFCP;…
  - Fibre Channel
    - IPFC(Internet Protocol over Fibre Channel);…
  - Infiniband
    - iSER(iSCSI Extensions for RDMA); FCoIB(Fibre Channel over InfiniBand)…

存储

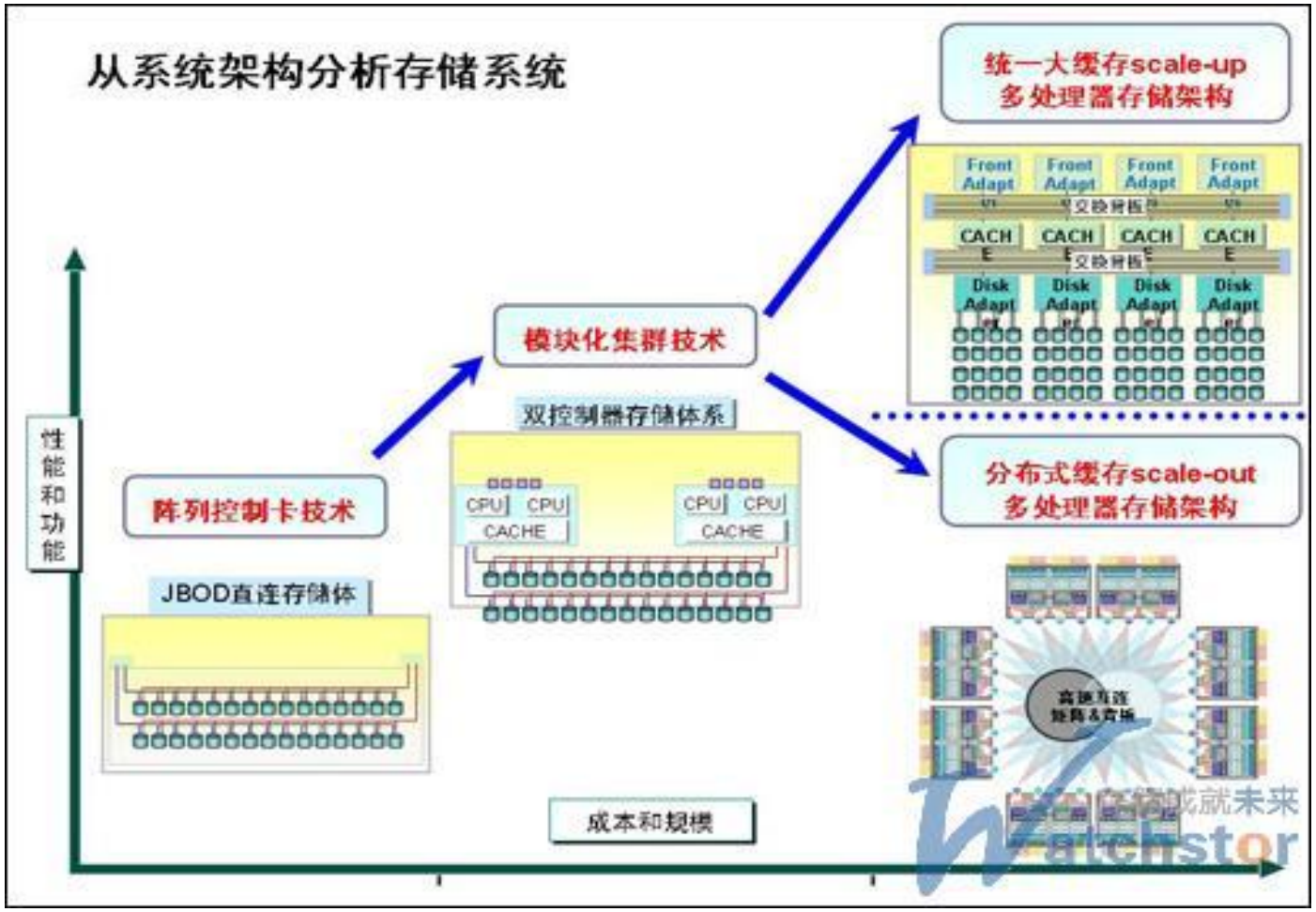
- 盘柜
  - JBOD(Just a Bunch Of Disks)
  - 代表：Dell M1000
- 盘阵
  - 中低端：emc cx series；
  - 高端：emc dmx series；IBM DS series；IBM XIV；HP 3PAR；
  - 区别：更大的缓存；更高的处理性能
- 存储划分
  - 条带化(打散数据，降低单一磁盘冲突，发挥所有磁盘的性能)
    - 条带深度：条带大小，条带单元 (单块磁盘)
    - 条带宽度：一个条带集中的驱动数 (多块磁盘)
    - OLTP：条带宽度 >= IO请求大小/条带深度 (保证一个IO，一块磁盘只服务一次)
- 软硬件一体机
  - Oracle exadata

网络存储

- DAS(Direct Attached Storage)
  - 直接附加存储，将存储通过SCSI接口或光纤通道直接连接到计算机
- NAS(Network Attached Storage)
  - 网络接入存储，采用TCP/IP等技术，网络交换机连接存储系统和主机
- SAN(Storage Area Network)
  - 存储区域网络，采用光纤通道技术，通过光纤交换机连接存储阵列和主机
  - 如图



存储的发展


从系统架构分析存储系统
统一大缓存scale-up
多处理器存储架构
Front Adapt
CACH E
Disk Adapt
模块化集群技术
双控制器存储体系
CPU
CACHE
分布式缓存scale-out
多处理器存储架构
阵列控制卡技术
JBOD直连存储体
性能和功能
成本和规模

Q & A

谢谢大家！