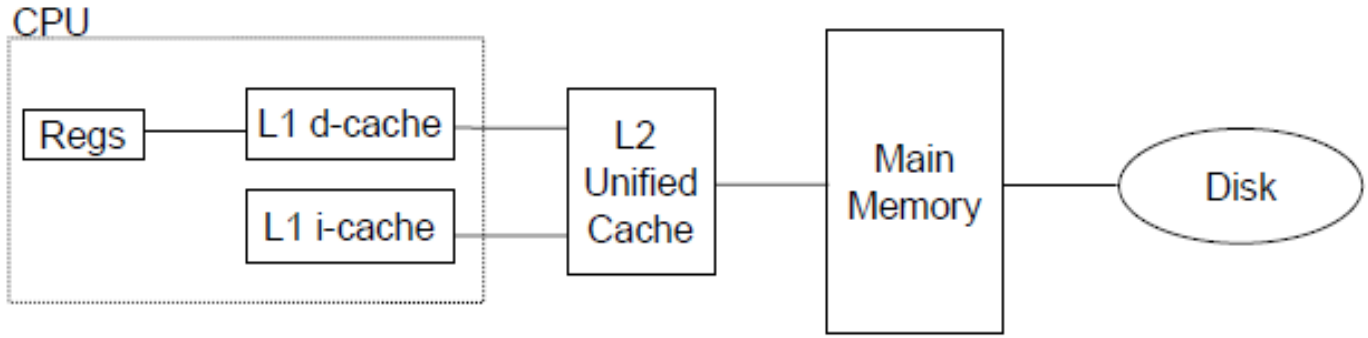
5.3.7 多级Cache

刘 芳 副教授
国防科学技术大学计算机学院

多级Cache系统成为主流：在Cache-Memory系统中使用更多的层次结构，以掩盖CPU访存延迟，提高处理器的执行效率



一个典型的多级cache组织结构

多Cache系统设计的主要考虑因素：

(1) 单级/多级

- 片内(On-chip)Cache：将Cache和CPU作在一个芯片上
- 外部(Off-chip)Cache：不做在CPU内而是独立设置一个Cache
- 单级Cache：只用一个片内Cache
- 多级Cache：同时使用L1 、L2 Cache，有些系统还有L3 Cache

L1 Cache更靠近CPU，其速度比L2快，其容量比L2小

多Cache系统设计的主要考虑因素：

(2) 联合/分立

分立：数据和指令分开存放在各自的数据和指令Cache中

联合：数据和指令都放在一个Cache中

一般L1 Cache都是分立Cache，为什么？

L1 Cache的命中时间比命中率更重要！减少命中时间以获得较短的时钟周期

一般L2 Cache都是联合Cache，为什么？

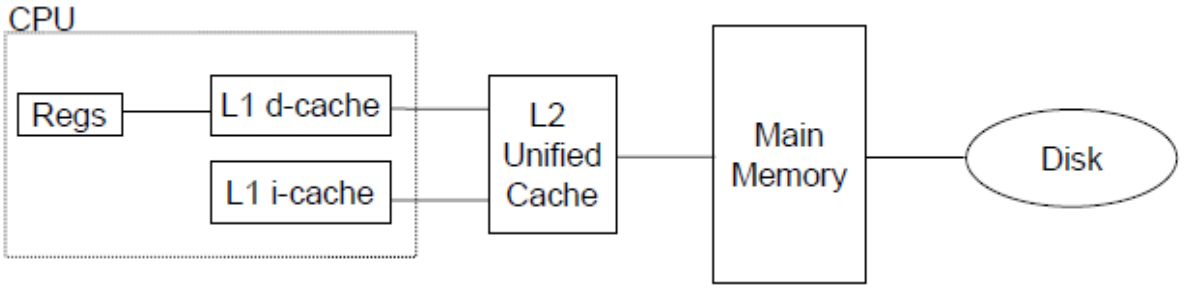
L2 Cache的命中率比命中时间更重要！降低缺失率以减少访问主存缺失损失

两级Cache 系统的缺失损失(Miss Penalty)分析：

(1) 单级/多级

- 若L2 Cache包含所请求信息，则缺失损失为L2 Cache的访问时间
- 否则要访问主存，并同时取到L1 Cache和L2 Cache(缺失损失更大)

L1 Cache更靠近CPU，其速度比L2快，其容量比L2小



5.3.7 多级Cache

例：某处理器的CPI为1（如果所有访问能在L1 Cache命中），时钟频率为5GHz。假设访问一次主存的时间为100ns(包括所有的缺失处理)，设平均每条指令在L1 Cache中的缺失率为2%；如果增加一个L2 Cache，访问时间为5ns，而且容量大到使L2 Cache缺失率减为0.5%，问处理器速率提高了多少？

解：如果只有一级Cache，则缺失只有一种：

L1缺失(访问主存)，其缺失损失为：100ns×5GHz=500个时钟周期

总的CPI = CPI +每条指令中存储器停顿的时钟周期

= 1+ 500×2% = 11.0

5.3.7 多级Cache

例：某处理器的CPI为1（如果所有访问能在L1 Cache命中），时钟频率为5GHz。假设访问一次主存的时间为100ns(包括所有的缺失处理)，设平均每条指令在L1 Cache中的缺失率为2%；如果增加一个L2 Cache，访问时间为5ns，而且容量大到使L2 Cache缺失率减为0.5%，问处理器速率提高了多少？

解：如果只有一级Cache，则缺失只有一种：

L1缺失(访问主存)，其缺失损失为：100ns×5GHz=500个时钟周期

总的CPI = CPI +每条指令中存储器停顿的时钟周期

= 1+ 500×2% = 11.0

如果有二级Cache，则有两种缺失：

L1缺失(访问L2Cache)：5ns×5GHz = 25个时钟周期

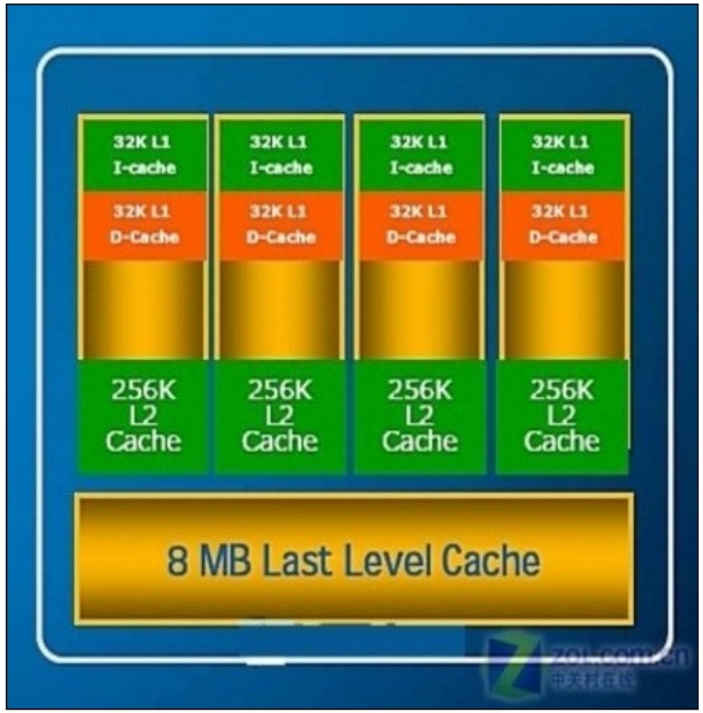
L2缺失(访问主存)：100ns×5GHz=500个时钟周期

总的CPI = CPI + 每条指令的一级停顿时钟周期+二级停顿的时钟周期

= 1+ 25×2% + 500×0.5% = 4.0

因此，二者的性能比为11.0/4.0=2.8倍

5.3.7 多级Cache



Nehalem Core i7处理器缓存结构图

Per core:

-32KB, 4-way L1 $I

-32KB, 8-way L1 $D

-256KB, 8-way L2

Shared

- 8 MB, 16-way L3

缓存技术的应用很广泛

Type	What cached	Where cached	Latency (cycles)	Managed by
CPU registers	4-byte word	On-chip CPU registers	0	Compiler
TLB	Address translations	On-chip TLB	0	Hardware
L1 cache	32-byte block	On-chip L1 cache	1	Hardware
L2 cache	32-byte block	Off-chip L2 cache	10	Hardware
Virtual memory	4-KB page	Main memory	100	Hardware + OS
cache	Parts of files	Main memory	100	OS
Network buffer cache	Parts of files	Local disk	10,000,000	AFS/NFS client
Browser cache	Web pages	Local disk	10,000,000	Web browser
Web cache	Web pages	Remote server disks	1,000,000,000	Web proxy server

缓存技术的基本思想

充分利用程序访问的局部性特点，将大容量、慢速存储器中当前刚用过的局部数据复制或暂存在小容量、快速存储器中，提高计算机系统访问效率