More Caching

Since caching is well studied, there are a lot of variations on cache concepts.

There is also a lot of data with which to evaluate the alternatives.

General principles

The larger the cache, the better (Figure 5.10)

A 64B line, fully associative, LRU unified cache has

10% miss rate at 2KB, but a 1% miss rate at 64KB

Set associativity increases beyond 4-way have almost no advantage, and 4-way's advantage over 2-way is marginal.

Finding Miss Rate from fully associative by "adjustment."

Unadj'd MR	<1%	1-2%	2-6%	6-12%	12-25%
Direct	1.65	1.52	1.38	1.37	1.33
2-Way	1.21	1.17	1.14	1.13	1.11
4-Way	1.06	1.05	1.04	1.03	1.03

Line Replacement

Applies to the "associative" part of the cache.

- LRU -- generally a good approximation to predicting the future.
- FIFO -- replace the line that has been in the cache the longest.
- Random -- select a victim line at random.

Smith finds that FIFO and Random perform about the same -- degrading LRU by amounts ranging from 0.96 to 1.38.

Overall, FIFO amplifies the miss rate of LRU by 1.12 on average.

Set associativity makes the LRU counters "trivial":

- 2-way set associative needs 1 bit, toggling to indicate most recent reference.
- m-way associativity requires $(m-1)\log_2 m$ bits per directory entry.

Recent cache designs

	DEC Alpha	IBM RS6000	MIPS R4000	HP PA7100	Intel Pentium
L1 I-Cache					
On/Off	On	Multichip	On	Off	On
Size	8KB	8KB	8KB	128KB	8KB
Virtual/Real	R	R	V	V	R
Organization	DM	2W	DM	DM	4W
LineSz	32B	64B	64B	32B	32B
L1 D-Cache					
On/Off	On	Multichip	On	Off	On
Size	8KB	64KB	8KB	256KB	8KB
Virtual/Real	R	R	V	V	R
Organization	DM	4W	DM	DM	4W
LineSz	32B	128B	64B	32B	32B

I-Cache and D-Cache

Separating the caches has the primary advantage that they can be accessed simultaneously.

A unified cache gets better performance for a given number of bytes.

Consider an asymmetric split, e.g. 25%-75%.

I-cache 64B lines, fully associative, LRU replacement, achieves 10% MR with 1KB, 1% MR with 32KB.

Prohibiting writing to I-cache resident lines is sensible.

"No duplicates policy" ... must check.

On Chip Caches

Differences for on chip caches:

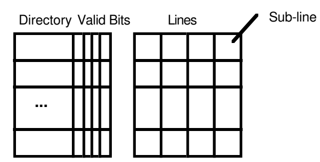
Pins limit the transfer path to/from memory

Cache organization must use area well.

Directory bits cost no more area than data bits, but contain no data.

CAM bits cost 2.2 rbes compared to 0.6 rbes for directory/data bits.

Sectoring allows the cache to fill on demand --

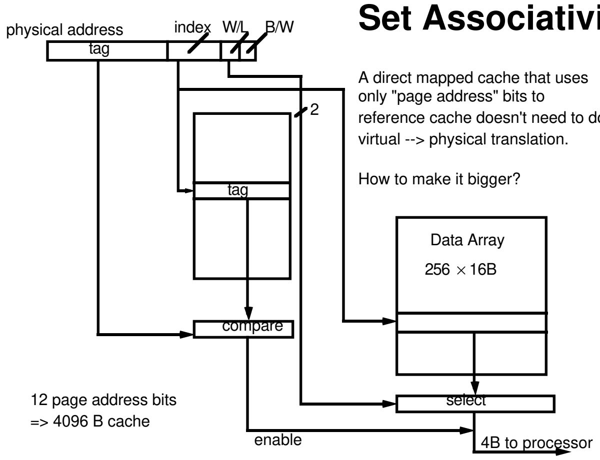


Sub-line = transfer unit from memory

Set Associativity

A direct mapped cache that uses only "page address" bits to reference cache doesn't need to do virtual --> physical translation.

How to make it bigger?



12 page address bits
=> 4096 B cache

Two Level Caches

Two level caching is simply an application of caching to caching. L2 cache should be "compatible" with L1 cache.

L2 should be substantially larger than L1.

L2 line ≥ L1 line.

L2 sets ≥ L1 sets.

L2 associativity ≥ L1 associativity.