Lecture 11: Large Cache Design

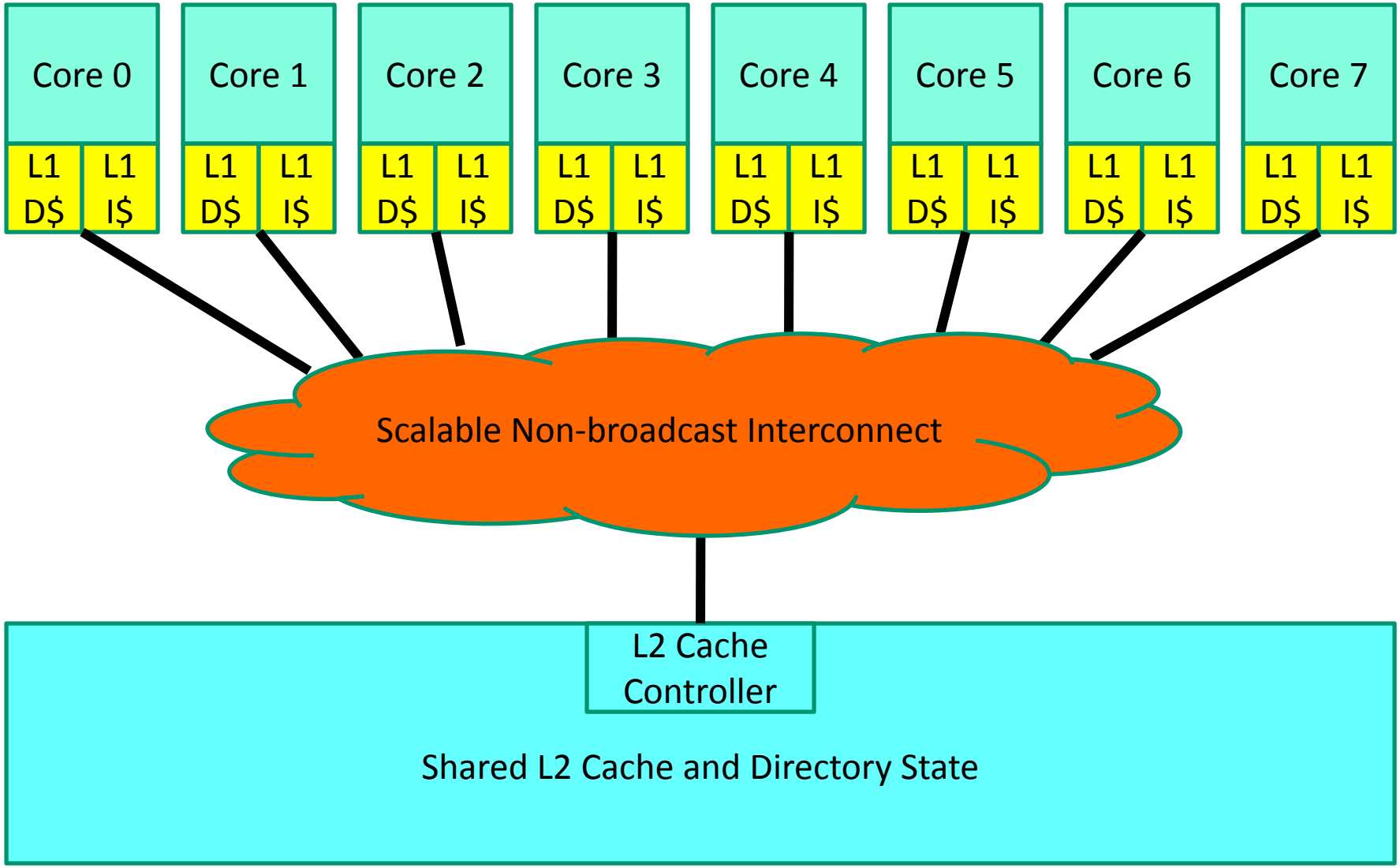
Topics: large cache basics and…

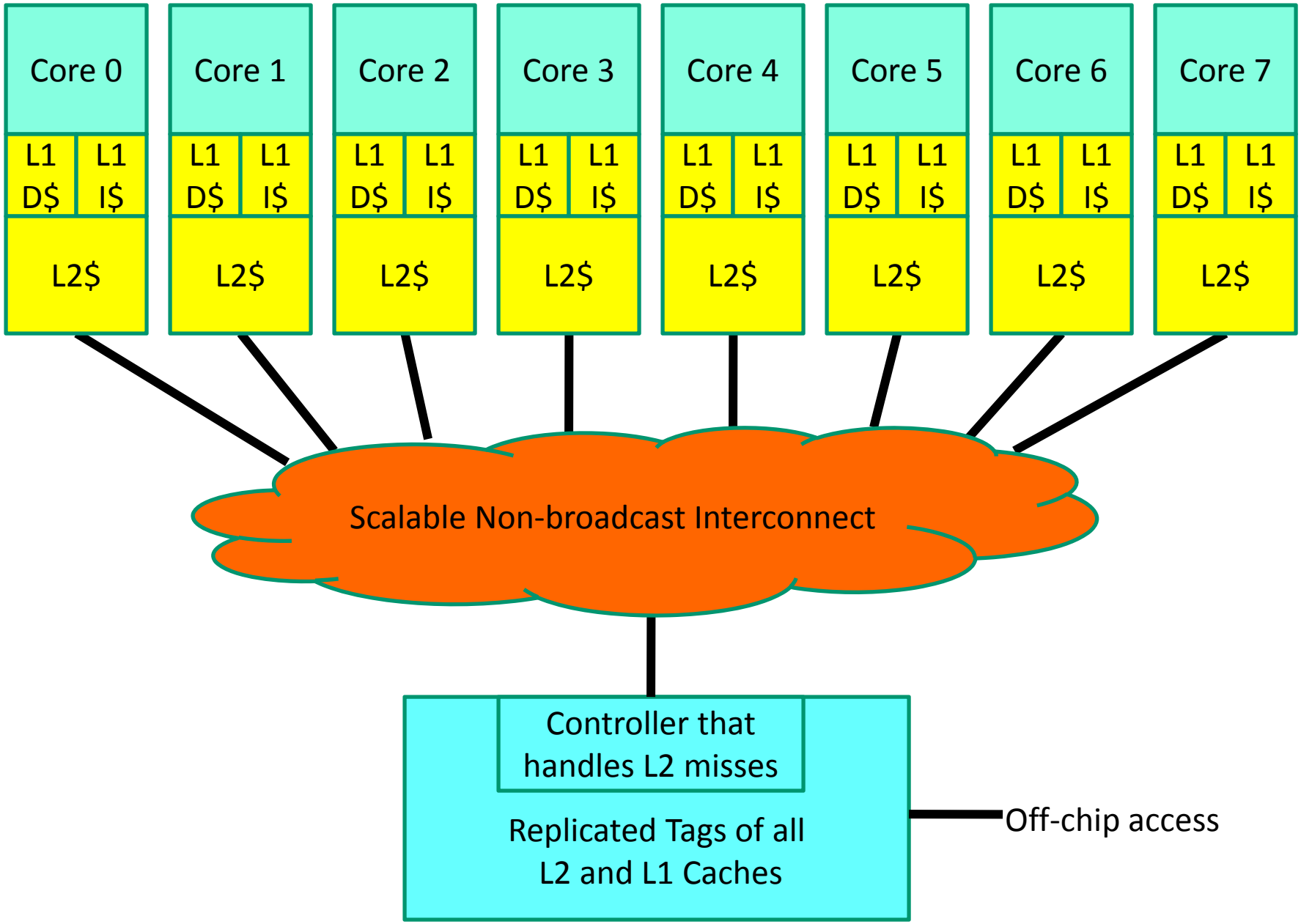
- An Adaptive, Non-Uniform Cache Structure for Wire-Dominated On-Chip Caches, Kim et al., ASPLOS'02
- Distance Associativity for High-Performance Energy-Efficient Non-Uniform Cache Architectures, Chishti et al., MICRO'03
- Managing Wire Delay in Large Chip-Multiprocessor Caches, Beckmann and Wood, MICRO'04
- Managing Distributed, Shared L2 Caches through OS-Level Page Allocation, Cho and Jin, MICRO'06

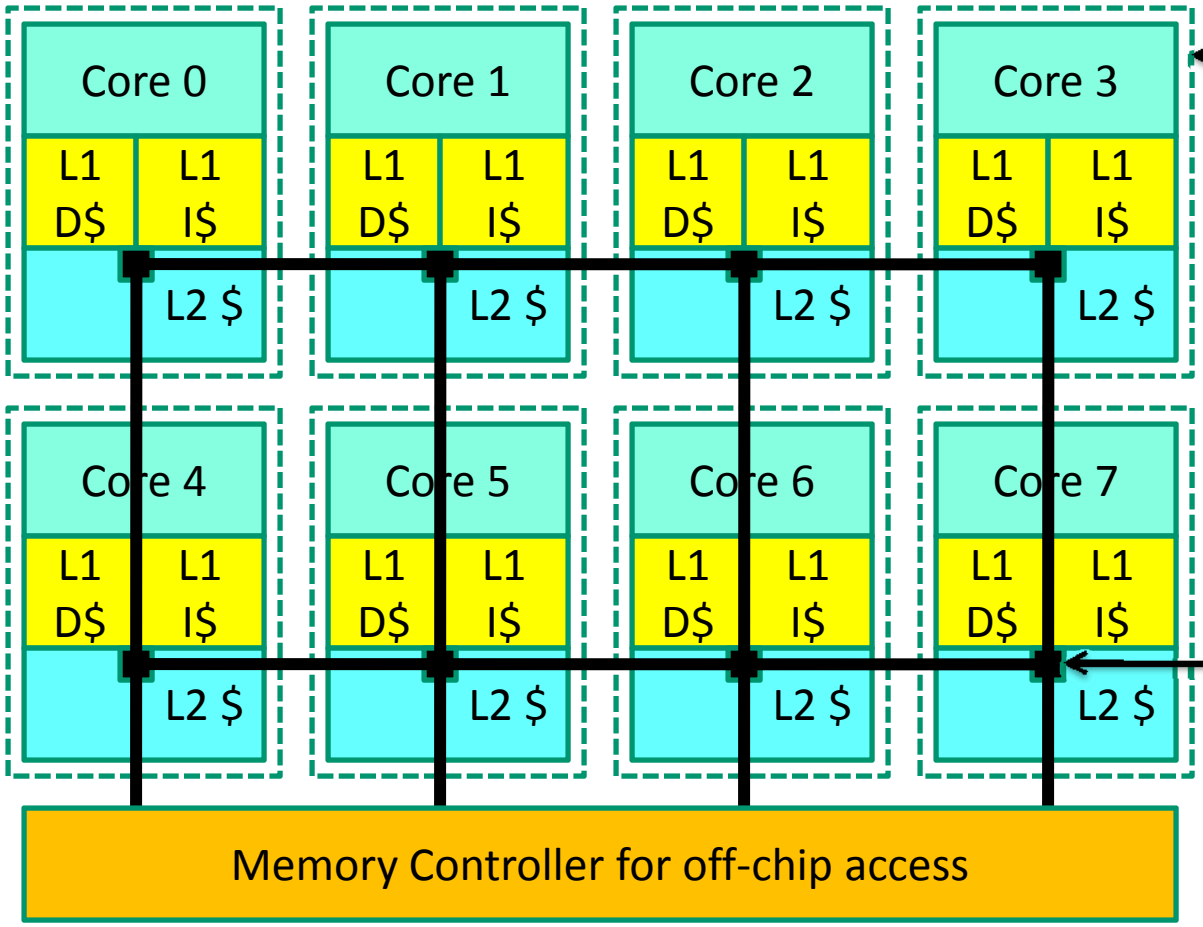
Shared Vs. Private Caches in Multi-Core

- Advantages of a shared cache:
  - Space is dynamically allocated among cores
  - No wastage of space because of replication
  - Potentially faster cache coherence (and easier to locate data on a miss)

- Advantages of a private cache:
  - small L2 → faster access time
  - private bus to L2 → less contention

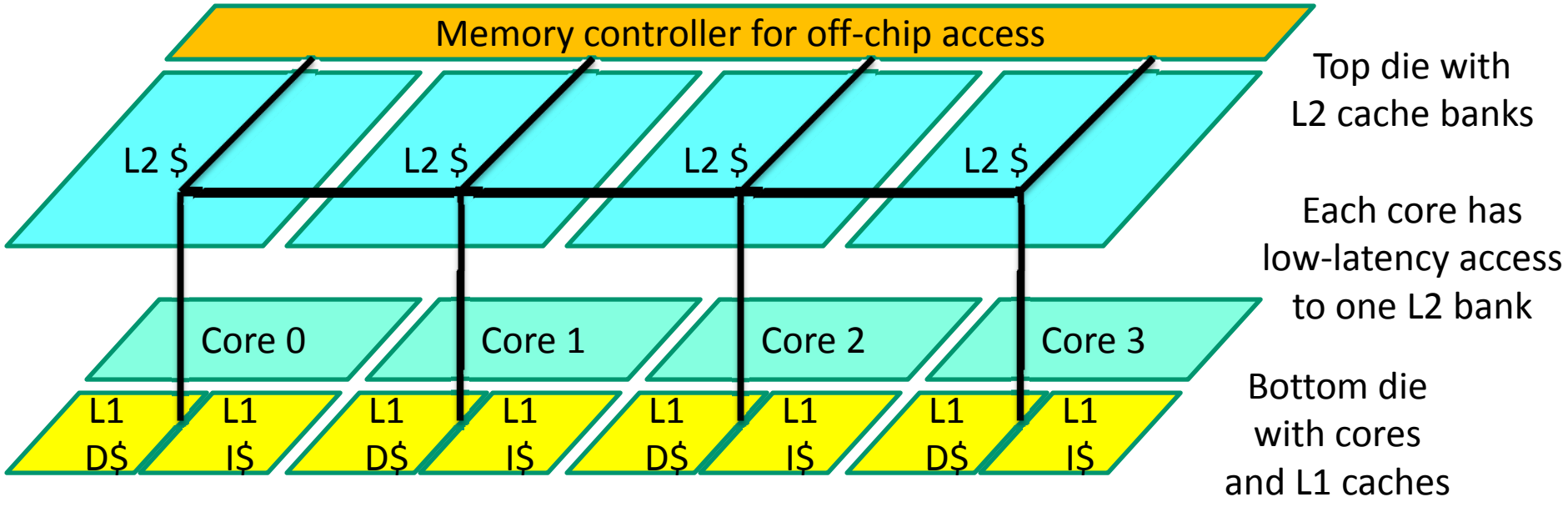

Core 0
L1 D$
L1 I$
Core 1
L1 D$
L1 I$
Core 2
L1 D$
L1 I$
Core 3
L1 D$
L1 I$
Core 4
L1 D$
L1 I$
Core 5
L1 D$
L1 I$
Core 6
L1 D$
L1 I$
Core 7
L1 D$
L1 I$
Scalable Non-broadcast Interconnect
L2 Cache Controller
Shared L2 Cache and Directory State


Core 0
Core 1
Core 2
Core 3
Core 4
Core 5
Core 6
Core 7
L1 D$
L1 I$
L2$
Scalable Non-broadcast Interconnect
Controller that handles L2 misses
Replicated Tags of all L2 and L1 Caches
Off-chip access

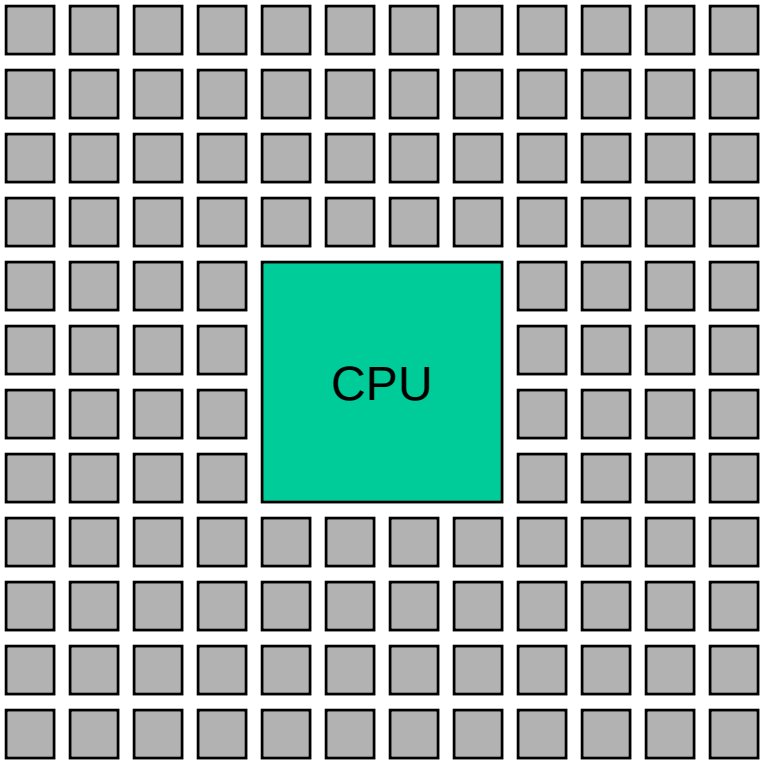


A single tile composed of a core, L1 caches, and a bank (slice) of the shared L2 cache

The cache controller forwards address requests to the appropriate L2 bank and handles coherence operations


Memory controller for off-chip access
L2 $
L2 $
L2 $
L2 $
Core 0
Core 1
Core 2
Core 3
L1 D$
L1 I$
L1 D$
L1 I$
L1 D$
L1 I$
L1 D$
L1 I$
Top die with L2 cache banks
Each core has low-latency access to one L2 bank
Bottom die with cores and L1 caches

Large NUCA



Issues to be addressed for Non-Uniform Cache Access:

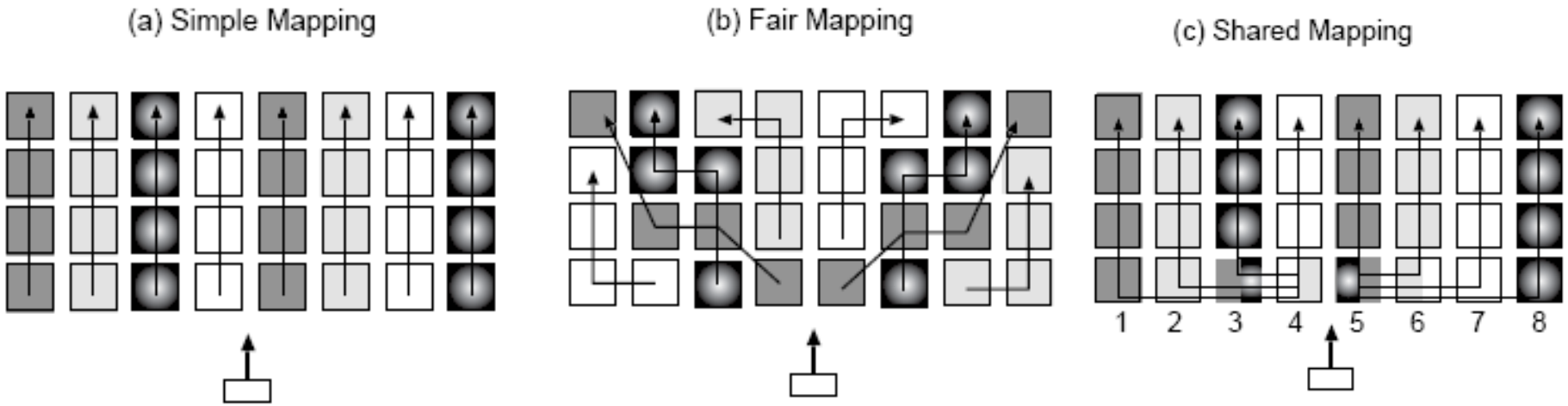
- Mapping
- Migration
- Search
- Replication

Static and Dynamic NUCA

- Static NUCA (S-NUCA)
  - The address index bits determine where the block is placed
  - Page coloring can help here as well to improve locality

- Dynamic NUCA (D-NUCA)
  - Blocks are allowed to move between banks
  - The block can be anywhere: need some search mechanism
  - Each core can maintain a partial tag structure so they have an idea of where the data might be (complex!)
  - Every possible bank is looked up and the search propagates (either in series or in parallel) (complex!)

Kim et al. (ASPLOS'02)



- Search policies:
  - incremental: check each bank before propagating the search
  - multicast: search in parallel
  - smart search: cache controller maintains partial tags that guide search or quickly signal a cache miss
- Movement: Data gradually moves closer as it is accessed
- Placement policy:
  - bring data close or far
  - replaced data is evicted or moved to furthest bank

Results

Average IPC values (16 MB, 50nm technology) :

- UCA cache: 0.26
- Multi-level UCA (L2/L3): 0.64
- Static NUCA: 0.65
- D-NUCA (simple map, multicast, insert at tail, 1-hit/1-bank promotion) 0.71
- D-NUCA with smart search 0.75
- Upper bound (instant L2 miss detection and all hits in first bank) 0.89

Chishti et al. (MICRO'03)

- Decouples the tag and data arrays
- Tag arrays are first examined (serial tag-data access is common and more power-efficient for large caches)
- Only the appropriate bank is then accessed
- Tags are organized conventionally, but within the data arrays, a set may have all its ways concentrated nearby
- The tags maintain forward pointers to data and data blocks maintain reverse pointers to tags

NuRAPID and Distance-Associativity

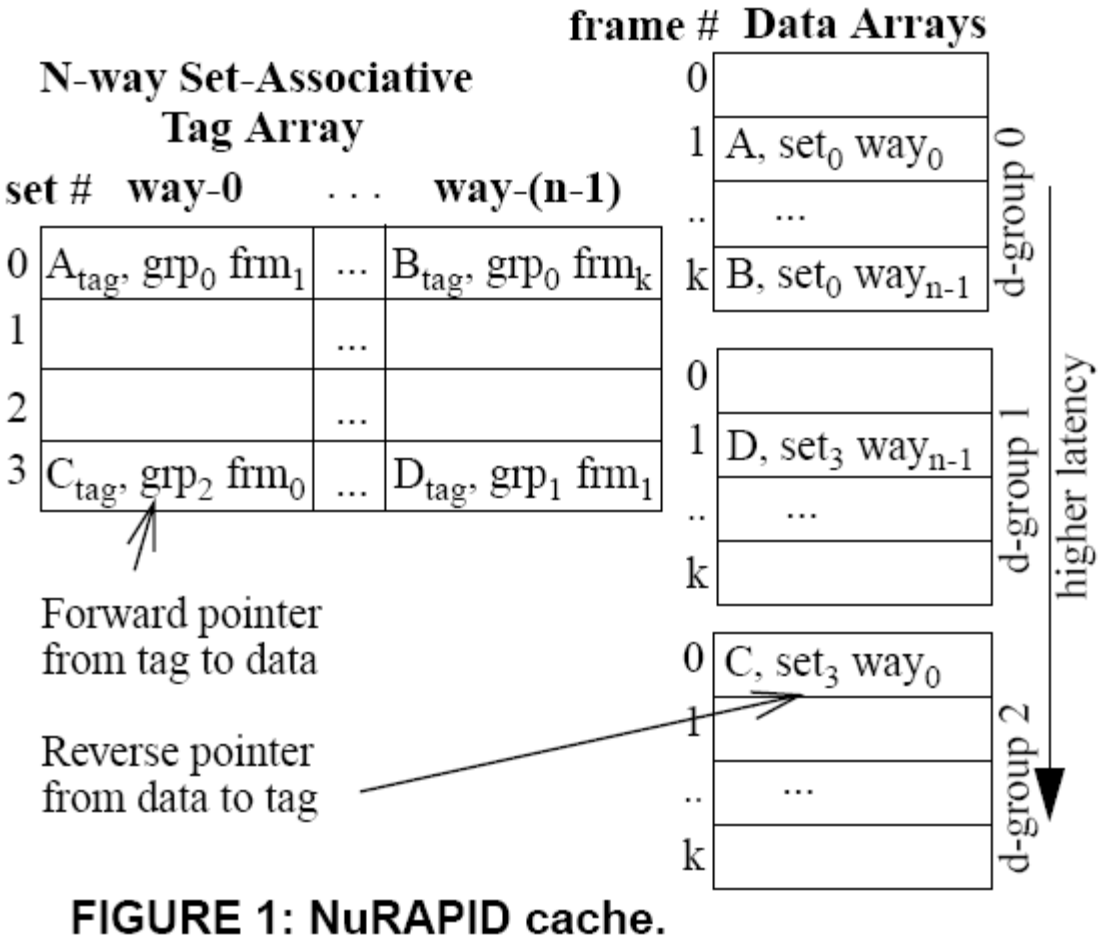
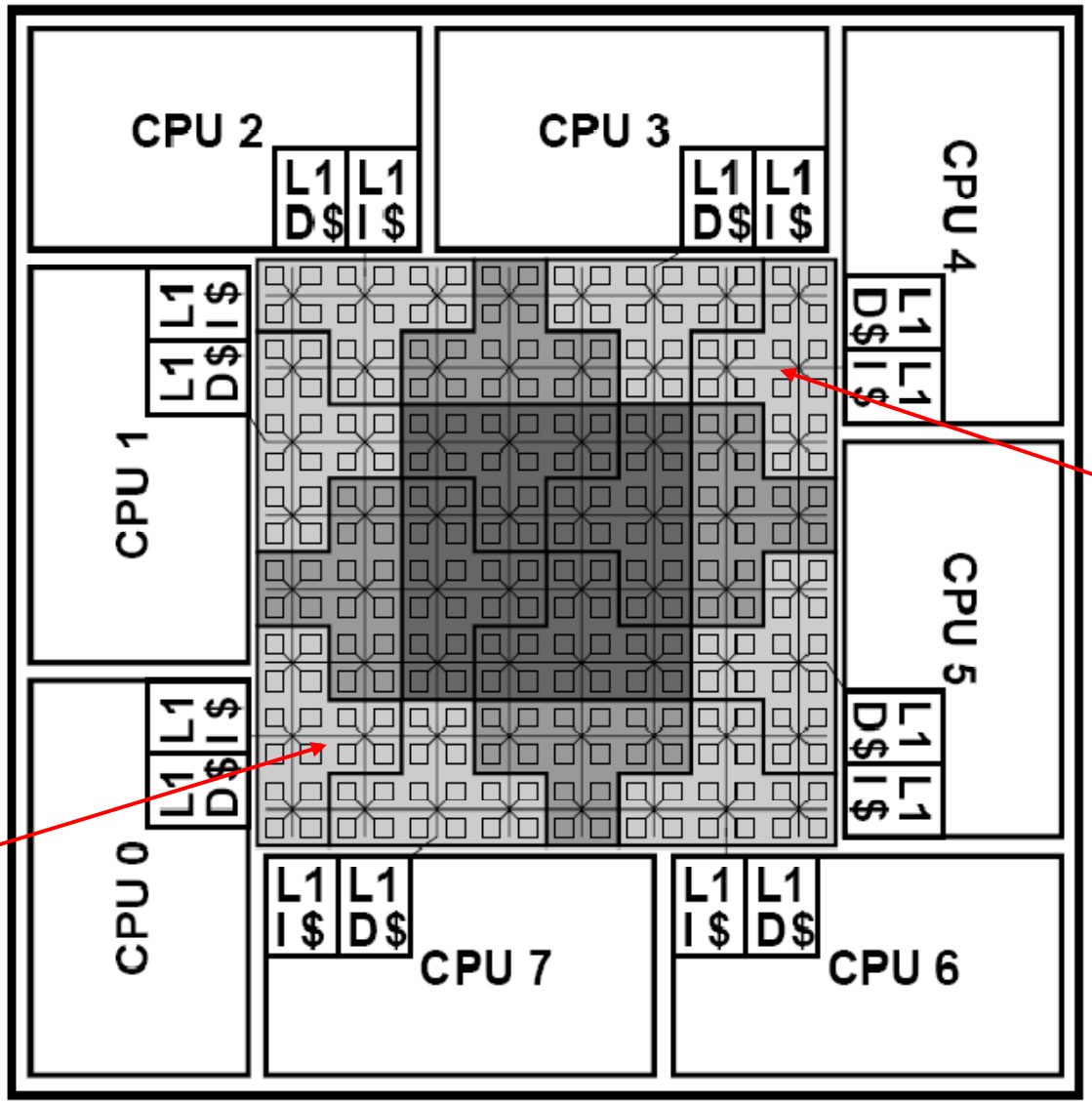


FIGURE 1: NuRAPID cache.

Beckmann and Wood, MICRO'04



Bankcluster Key

- Local
- Inter.
- Center

Data must be placed close to the center-of-gravity of requests

Examples: Frequency of Accesses

Dark → more accesses

←OLTP (on-line transaction processing)

Ocean → (scientific code)

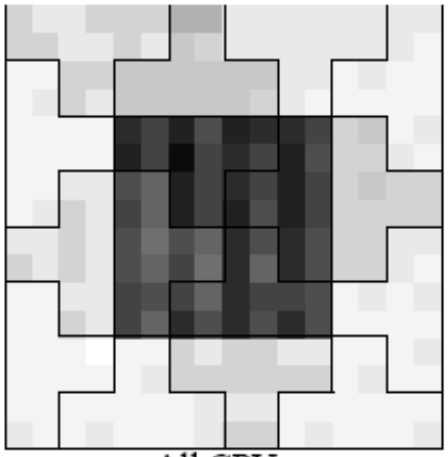


Figure 10. oltp L2 Hit Distribution

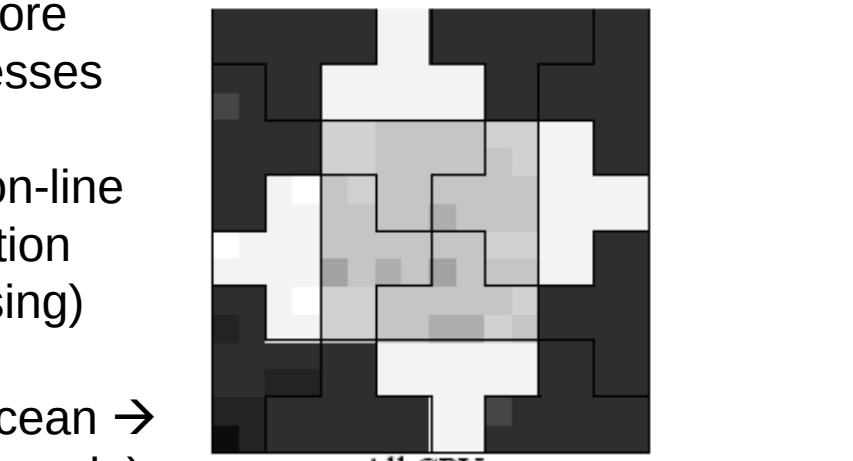
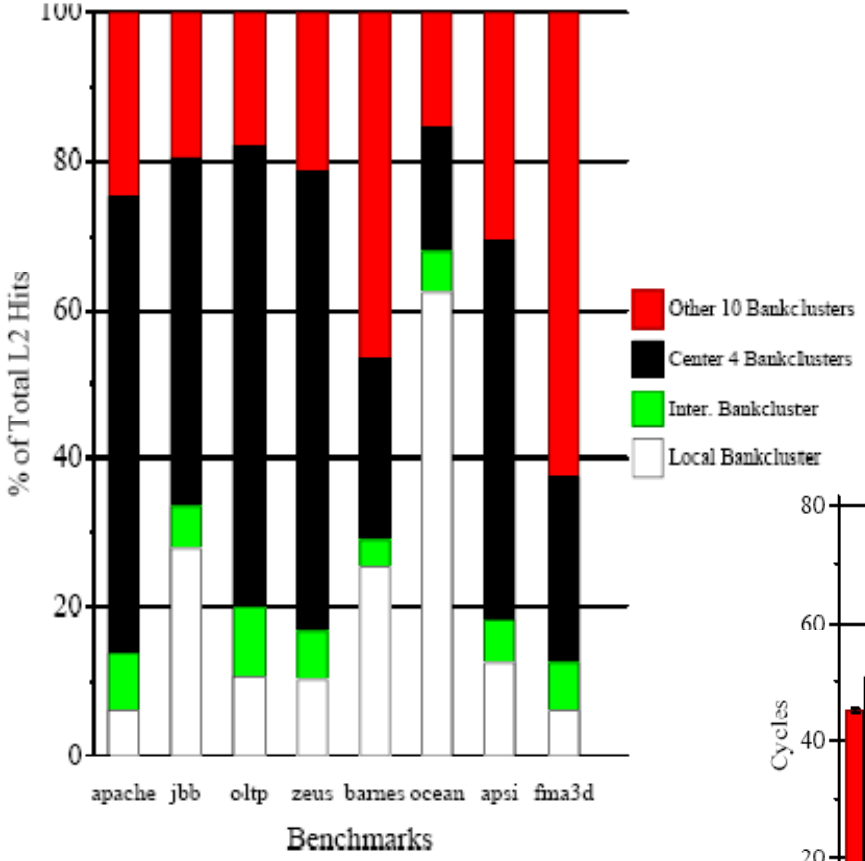


Figure 11. ocean L2 Hit Distribution

Block Migration Results



While block migration reduces avg. distance, it complicates search.

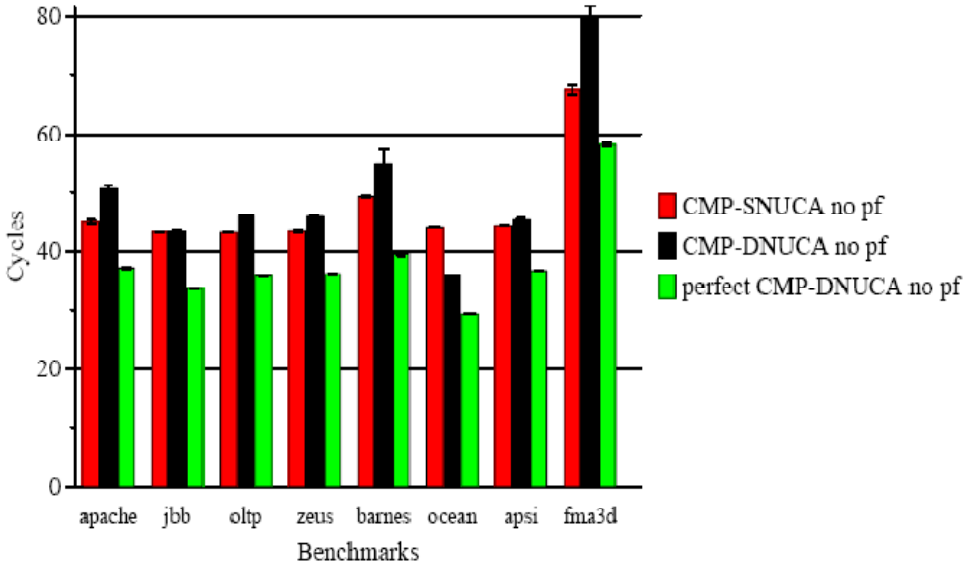
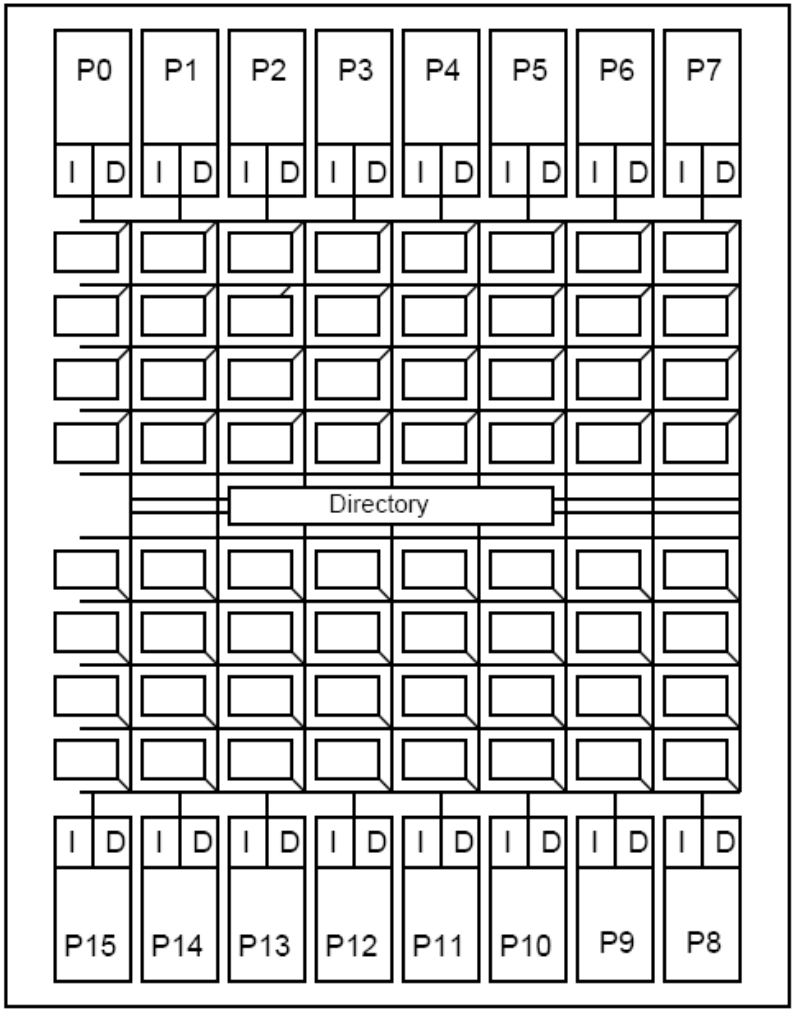


Figure 12. Avg. L2 Hit Latency: No Prefetching

Alternative Layout

(a) CMP Substrate: 16 CPUs 8x8 Banks



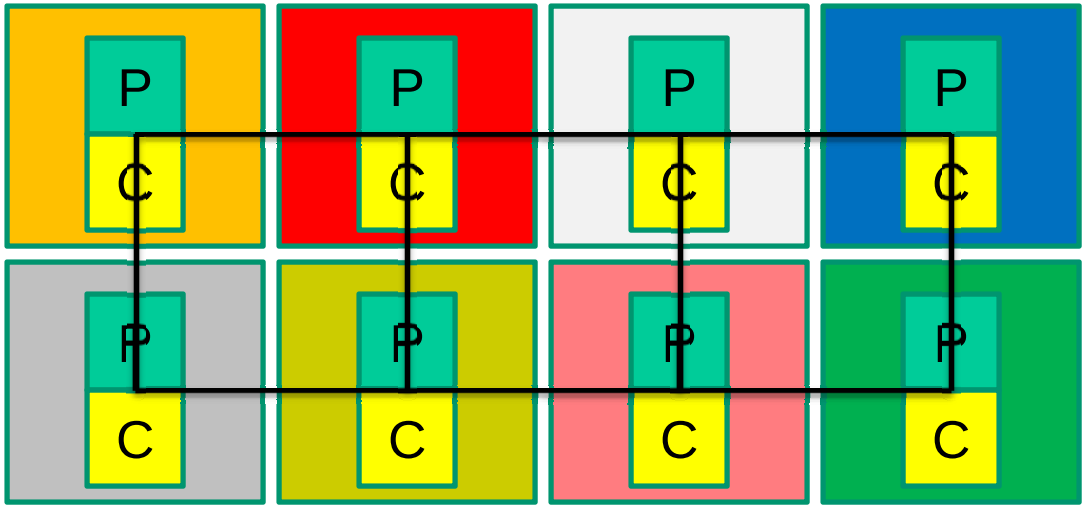
From Huh et al., ICS'05:

- Paper also introduces the notion of sharing degree
- A bank can be shared by any number of cores between N=1 and 16.
- Will need support for L2 coherence as well

Cho and Jin, MICRO'06

- Page coloring to improve proximity of data and computation
- Flexible software policies
- Has the benefits of S-NUCA (each address has a unique location and no search is required)
- Has the benefits of D-NUCA (page re-mapping can help migrate data, although at a page granularity)
- Easily extends to multi-core and can easily mimic the behavior of private caches

Page Coloring Example



- Recent work (Awasthi et al., HPCA'09) proposes a mechanism for hardware-based re-coloring of pages without requiring copies in DRAM memory

Title

- Bullet