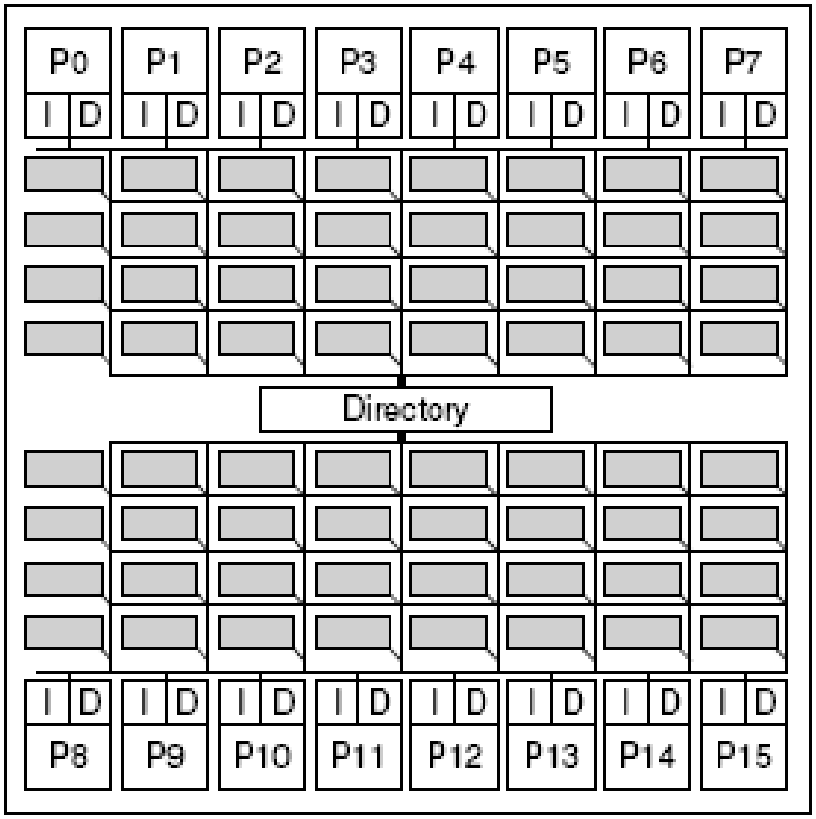
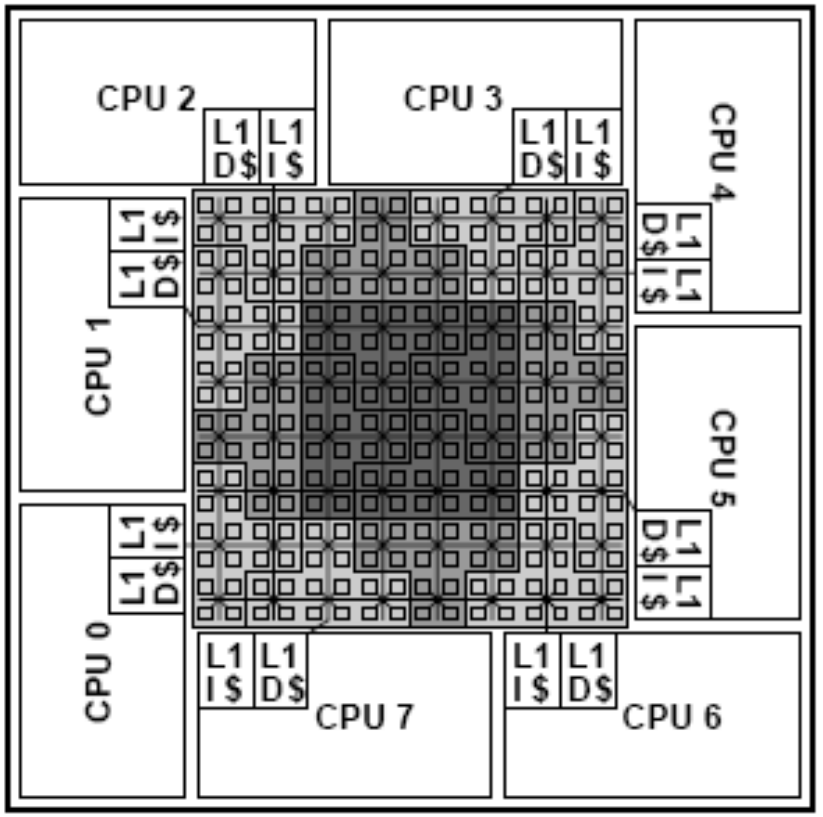
Lecture 19: Networks for Large Cache Design

Papers:

- Interconnect Design Considerations for Large NUCA Caches, Muralimanohar and Balasubramonian, ISCA'07
- Design and Management of 3D Chip Multiprocessors using Network-in-Memory, Li et al., ISCA'06
- A Domain-Specific On-Chip Network Design for Large Scale Cache Systems, Jin et al., HPCA'07
- Nahalal: Cache Organization for Chip Multiprocessors, Guz et al., Comp. Arch. Letters, 2007

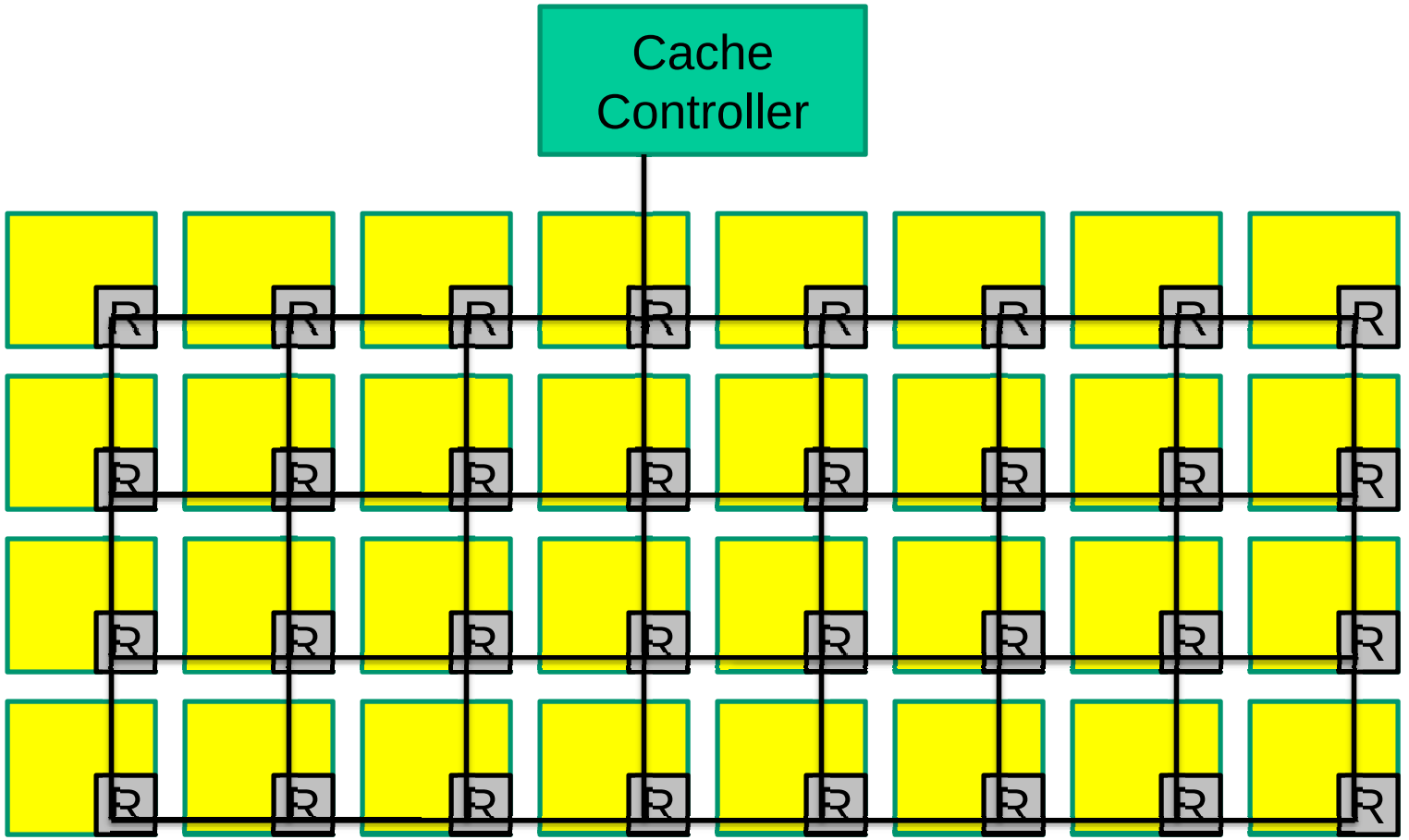
Traditional Networks



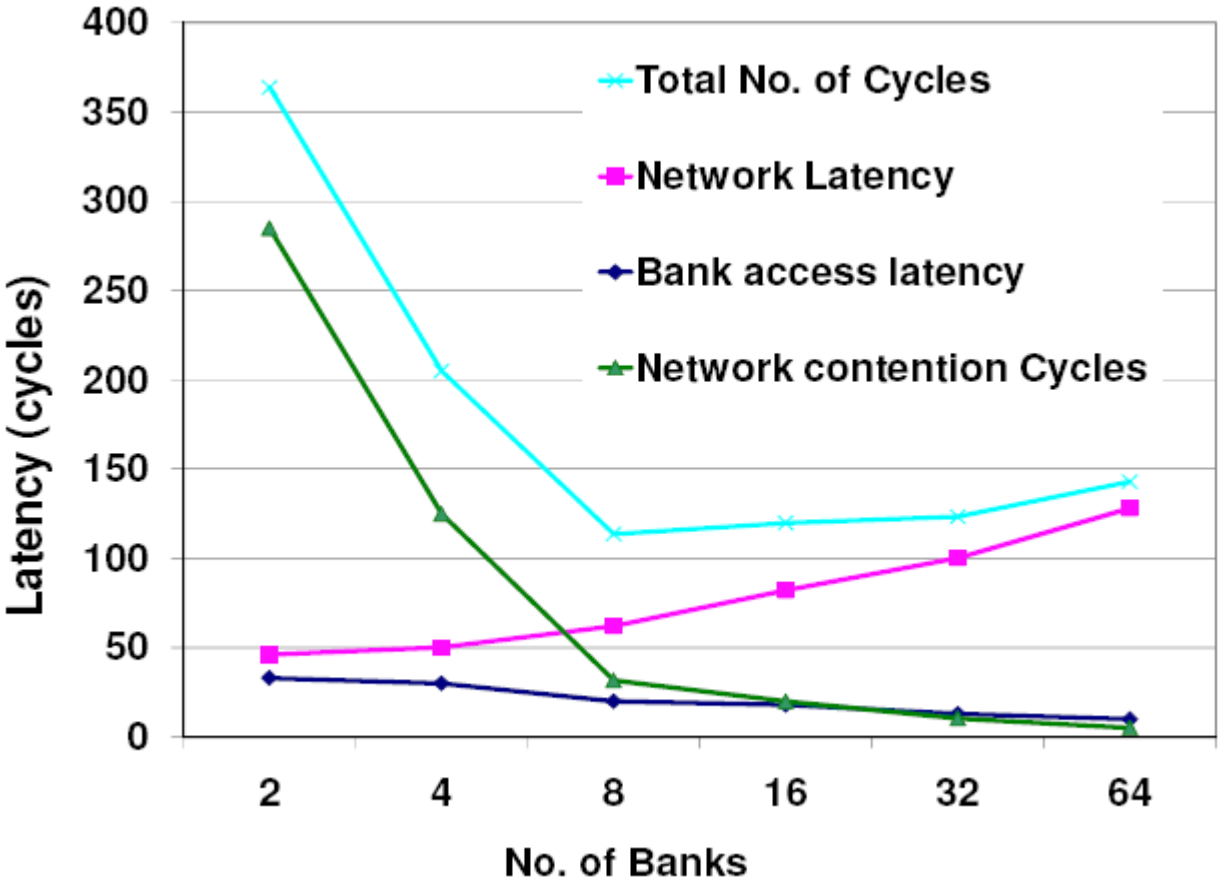


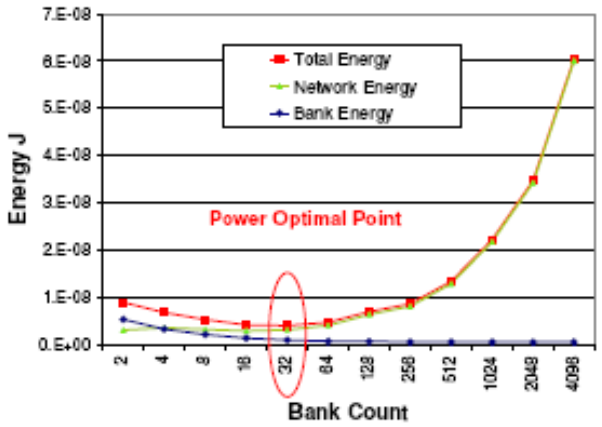
Example designs for contiguous L2 cache regions

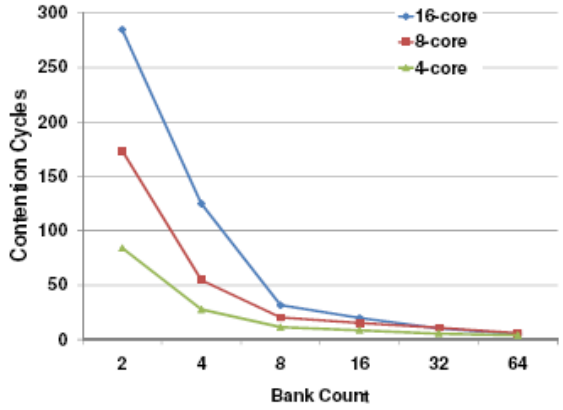
NUCA Delays



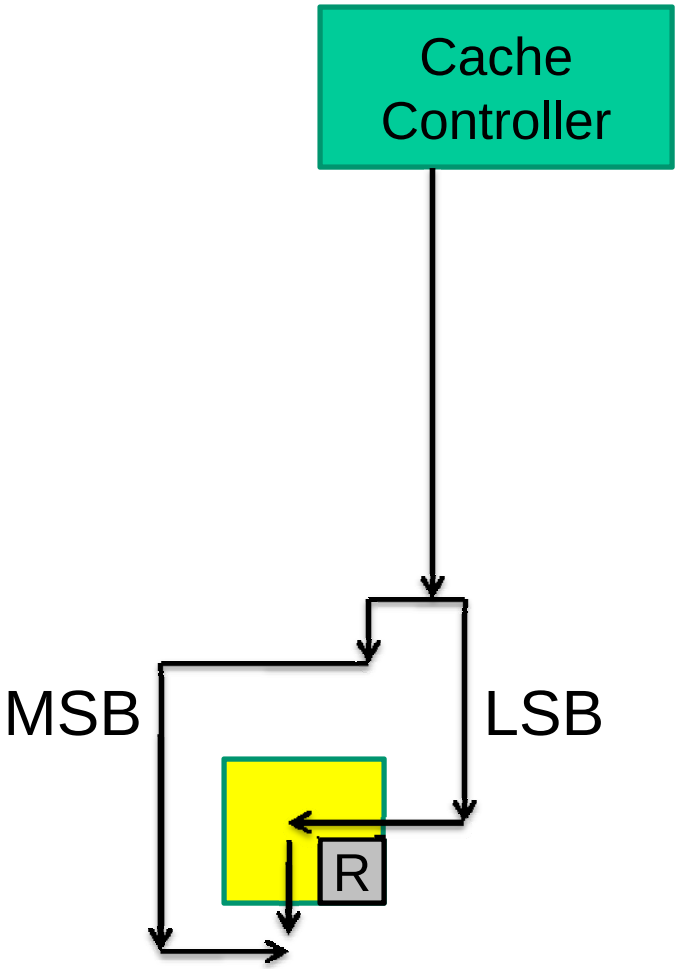
Explorations for Optimality





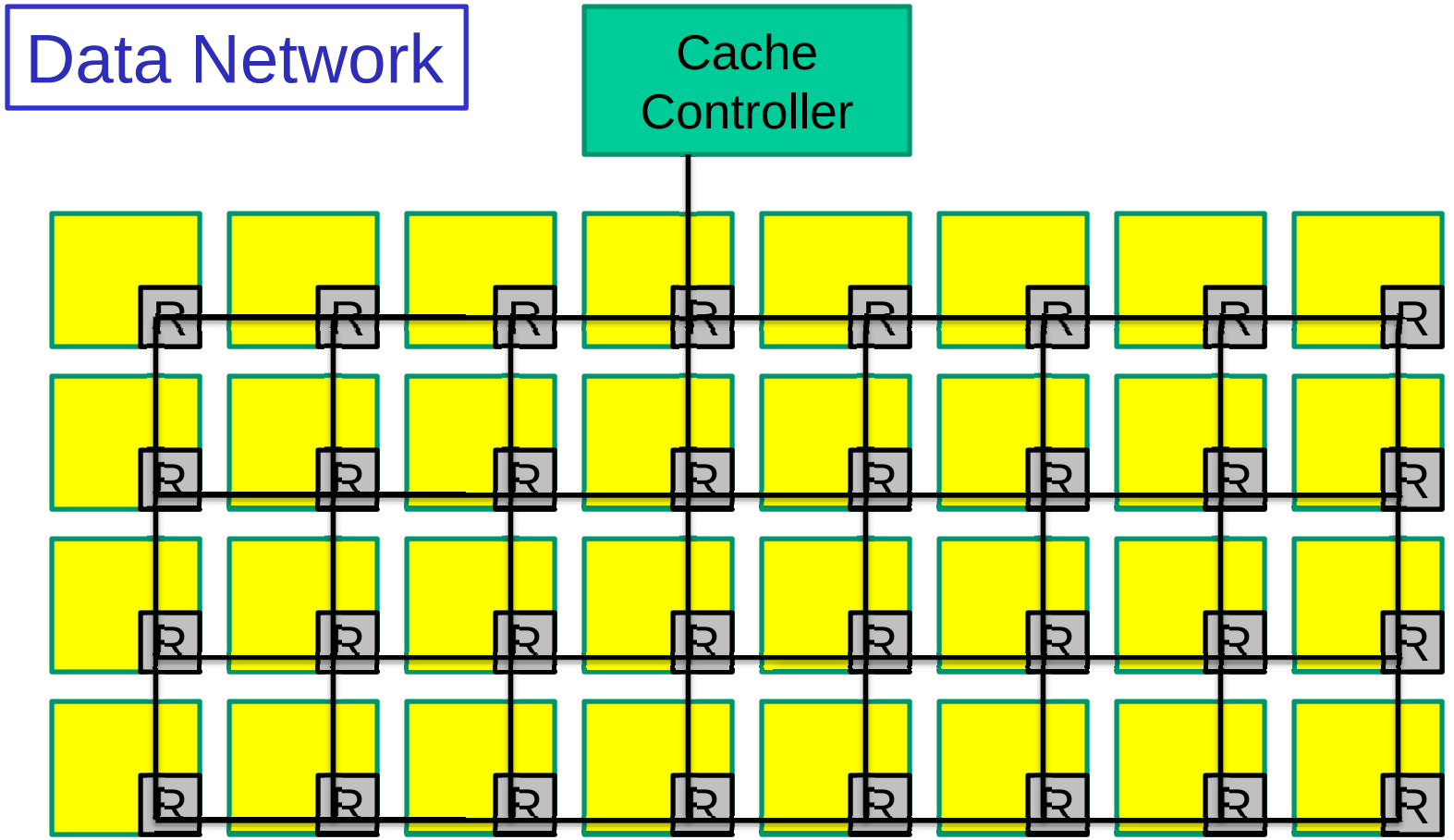


Early and Aggressive Look-Up

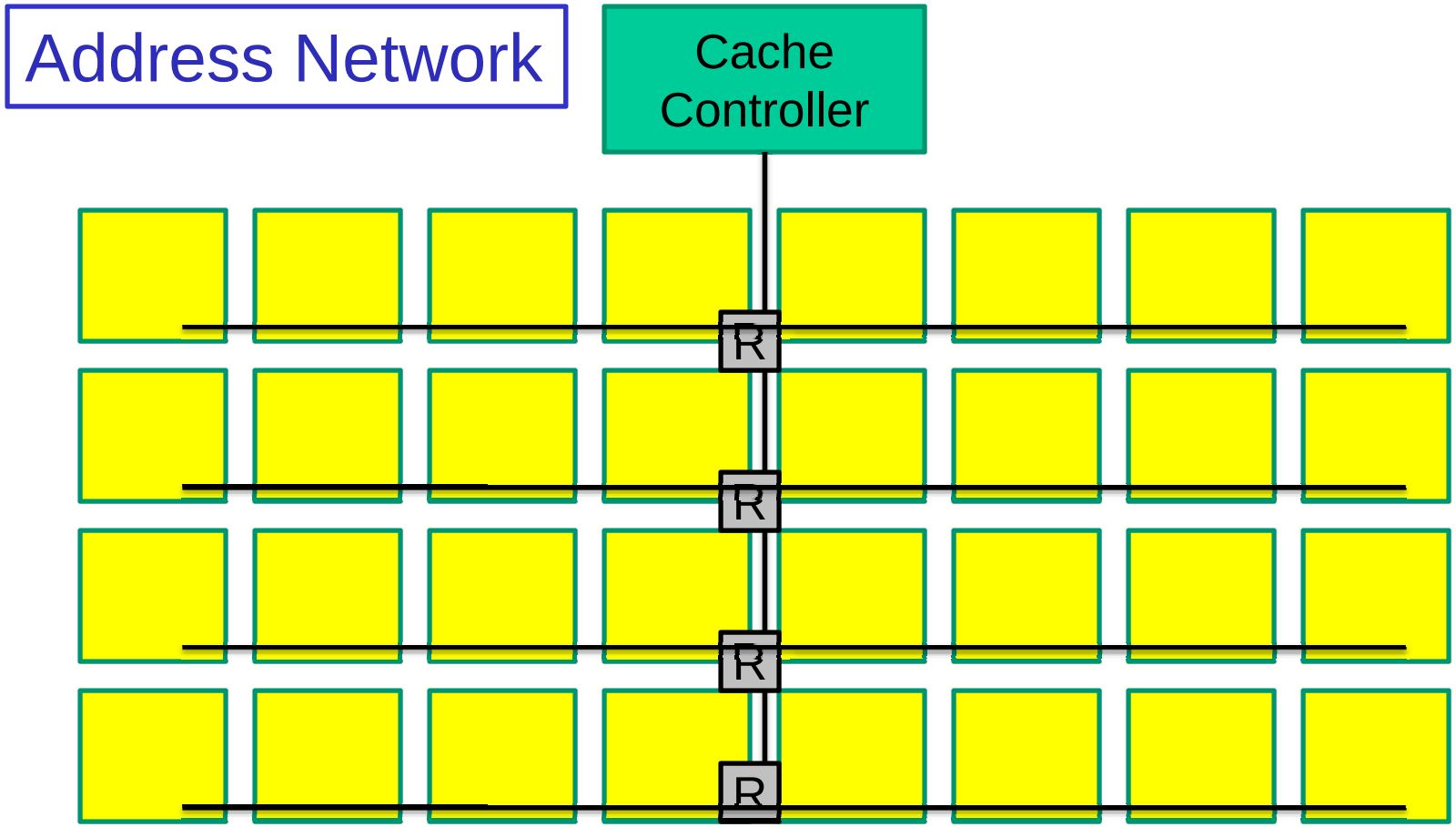


- Address packet can only contain LSB and can use latency-optimized wires (transmission lines / fat wires)
- Data packet also contains tags and can use regular wires
- The on-chip network can now have different types of links for address and data

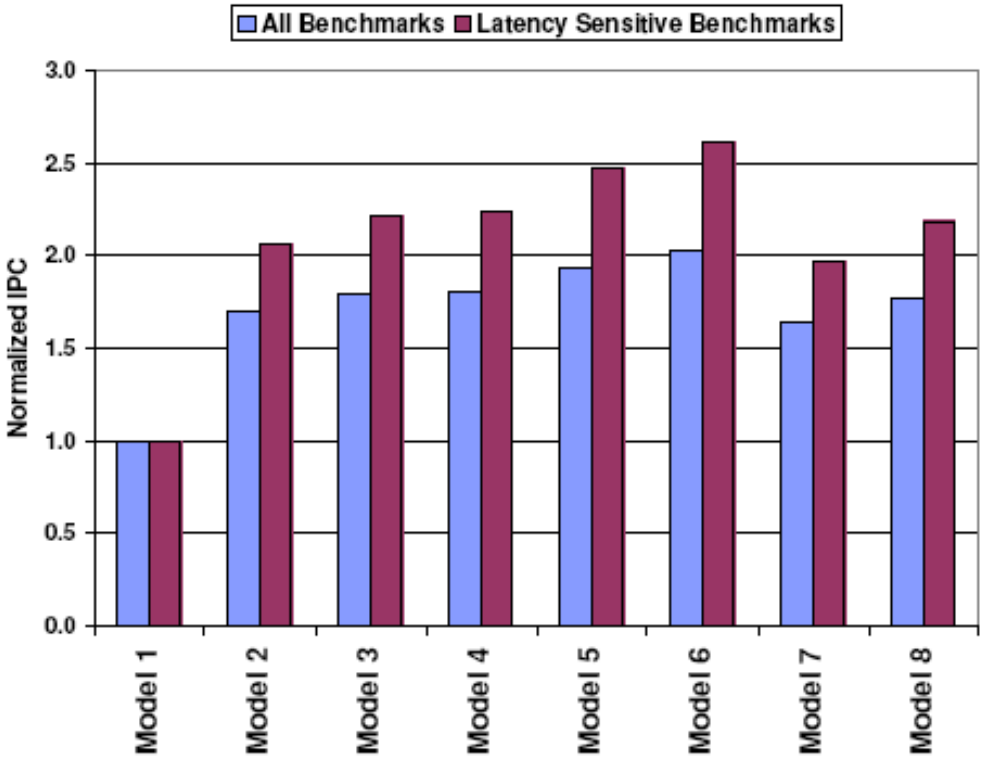
Hybrid Network



Hybrid Network

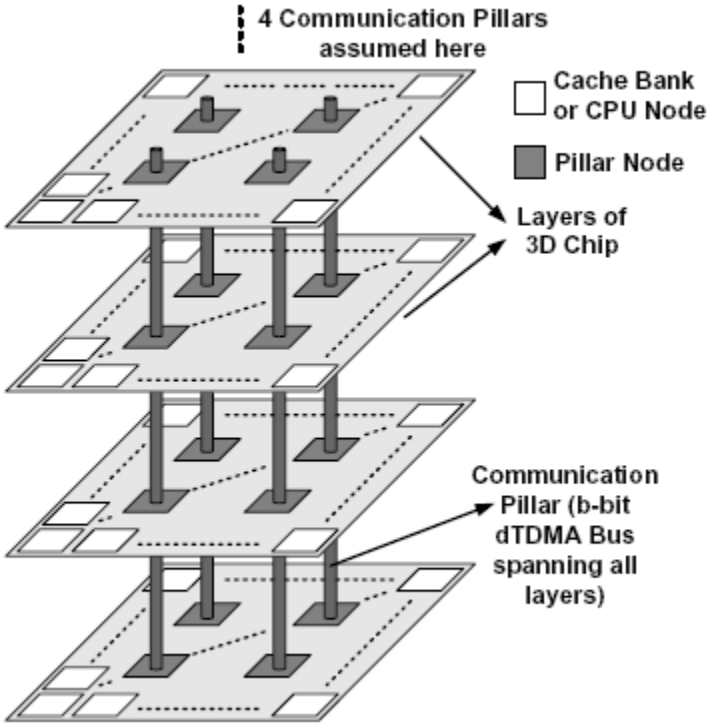


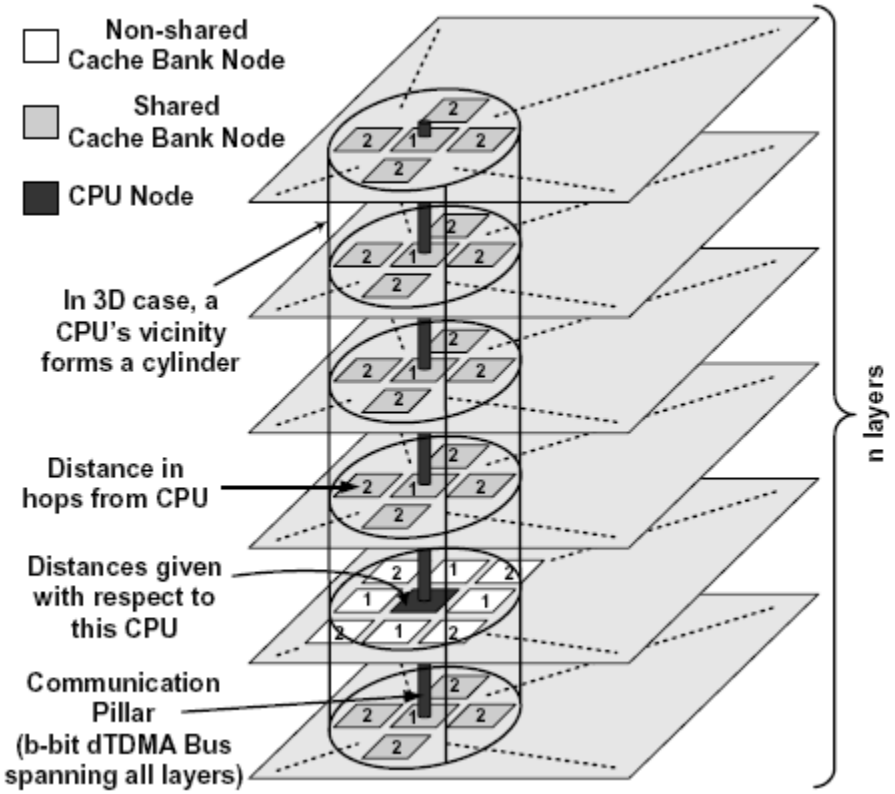
Results



Model	Link latency (vert,horiz)	Bank access time	Bank count	Network link contents	Description
Model 1	1,1	3	512	B-wires (256D, 64A)	Based on prior work
Model 2	4,3	17	16	B-wires (256D, 64A)	Derived from CACTIL2
Model 3	4,3	17	16	B-wires (128D, 64A) & L-wires (16A)	Implements early look-up
Model 4	4,3	17	16	B-wires (128D) & L-wires (24A)	Implements aggressive look-up
Model 5	hybrid	17	16	L-wires (24A) & B-wires (128D)	Latency-bandwidth tradeoff
Model 6	4,3	17	16	B-wires (256D), 1cycle Add	Implements optimistic case
Model 7	1,1	17	16	L-wires (40A/D)	Latency optimized
Model 8	4,3	17	16	B-wires (128D) & L-wires (24A)	Address-L-wires & Data-B-wires

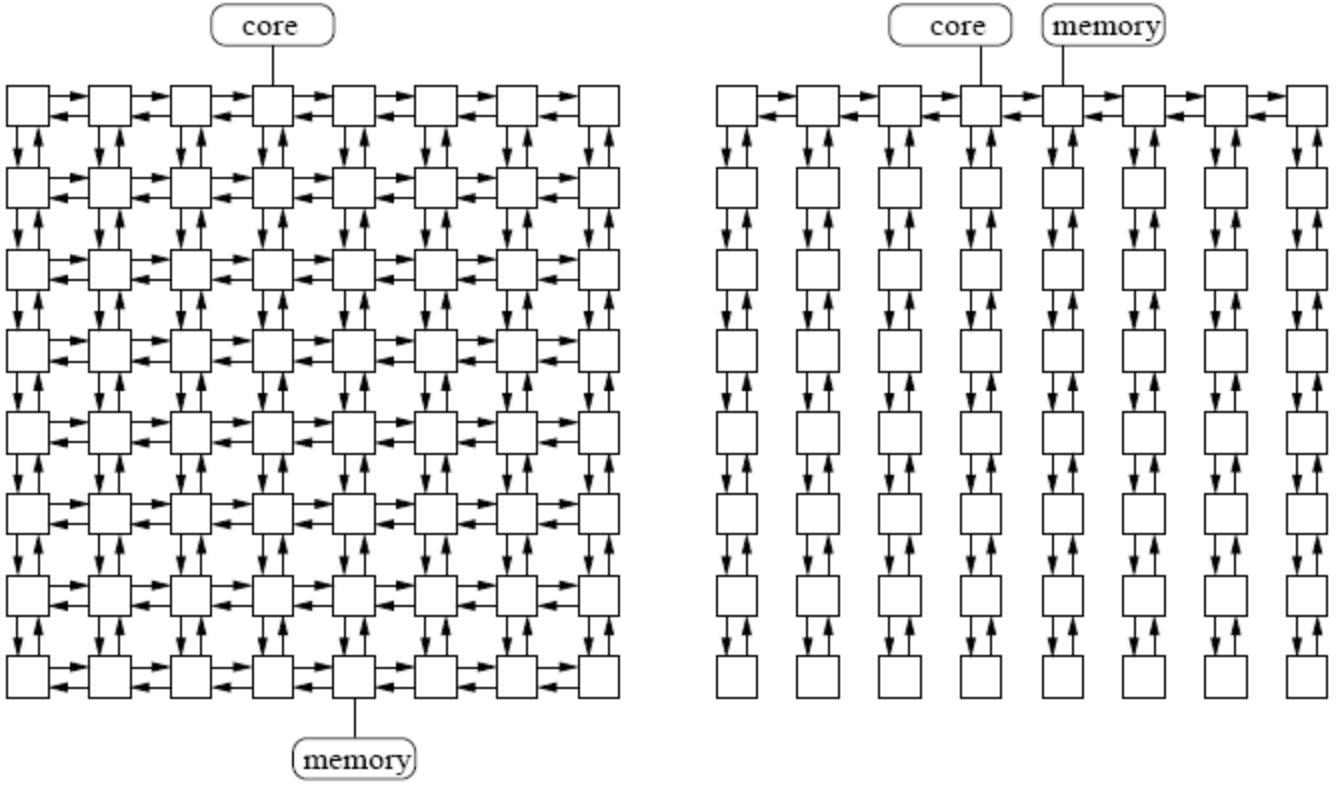
3D Designs, Li et al., ISCA'06





- D-NUCA: first search in cylinder, then multicast search everywhere
- Data is migrated close to requester, but need not jump across layers

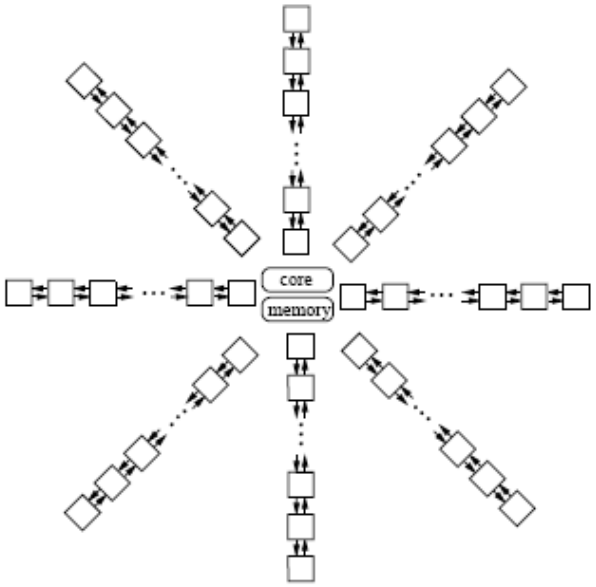
Halo Network, Jin et al., HPCA'07



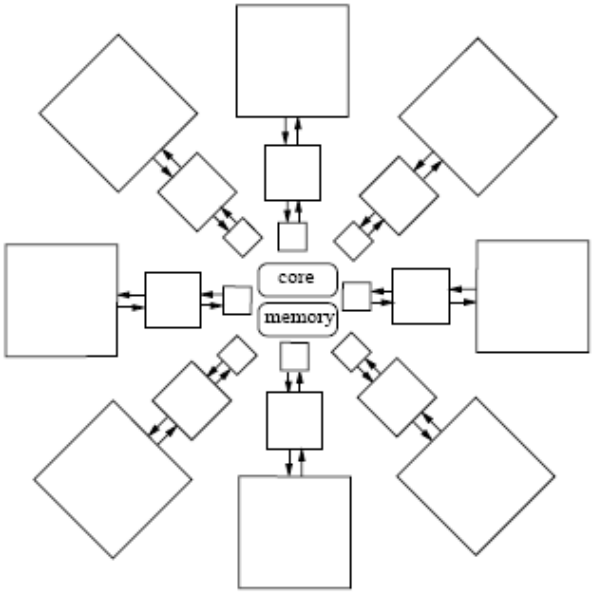
(a) Mesh (b) Simplified Mesh

- D-NUCA: Sets are distributed across columns;
  Ways are distributed across rows

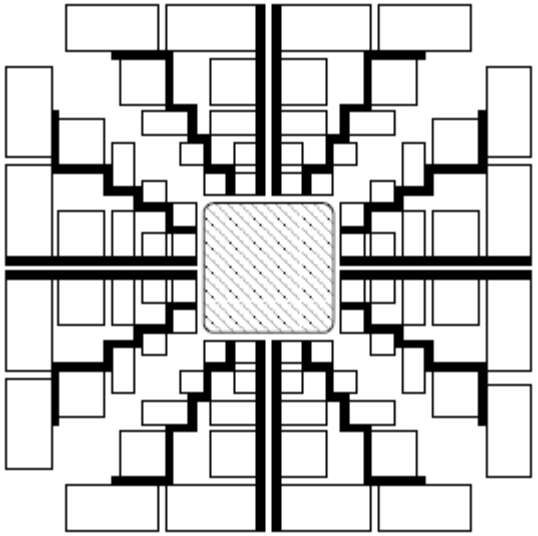
Halo Network



(c) Halo Constructed with Uniform Size Banks

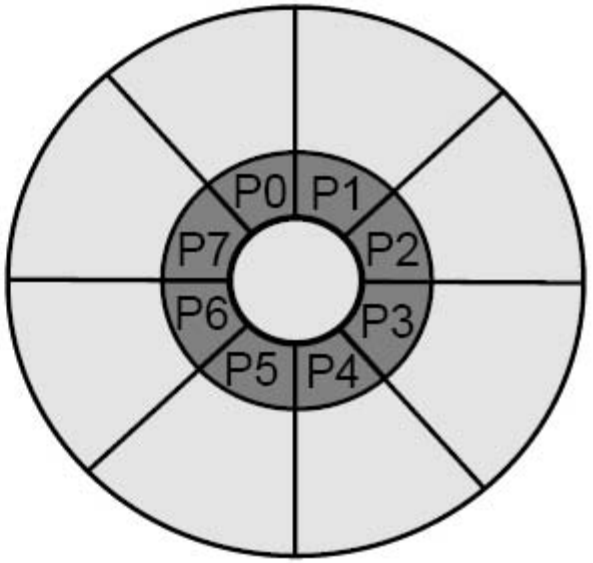


(d) Halo Constructed wit Non-uniform Size Banks



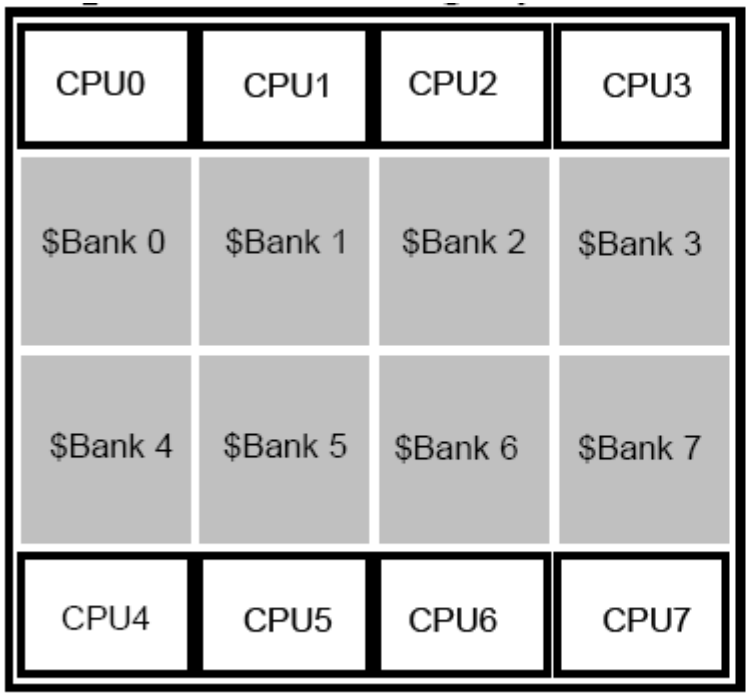
Nahalal, Guz et al., CAL'07

(a) Aerial view of Nahalal Village.

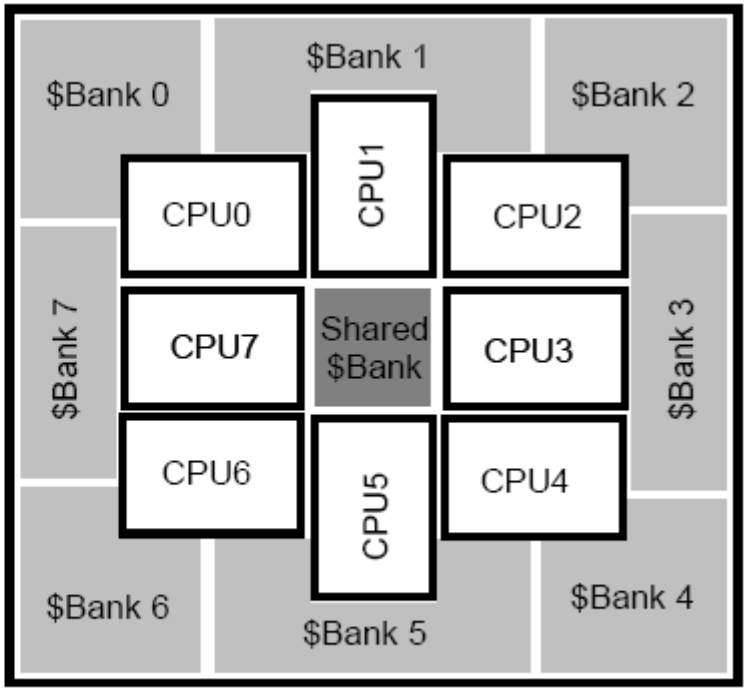


(b) CMP conceptual layout scheme.

Nahalal



(a) CIM layout.



(b) Nahalal layout.

- Block is initially placed in core's private bank and then swapped into the shared bank if frequently accessed by other cores
- Parallel search across all banks

Title

- Bullet