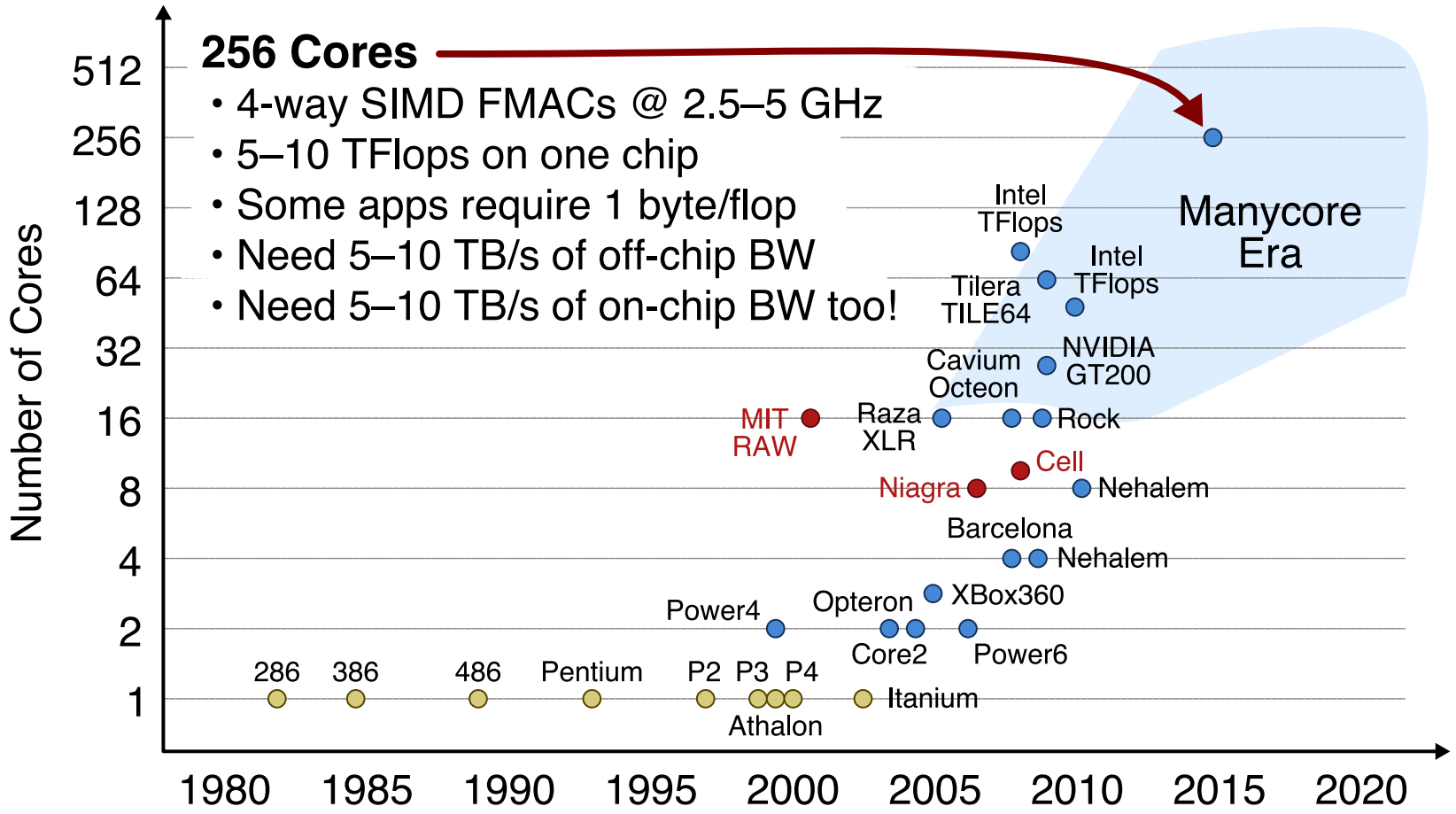
Bandwidth Requirements in Manycore Architectures: What Can 3D Bring?

Olivier Sentieys

INRIA/IRISA
University of Rennes olivier.sentieys@inria.fr

Multicore Bandwidth Requirements

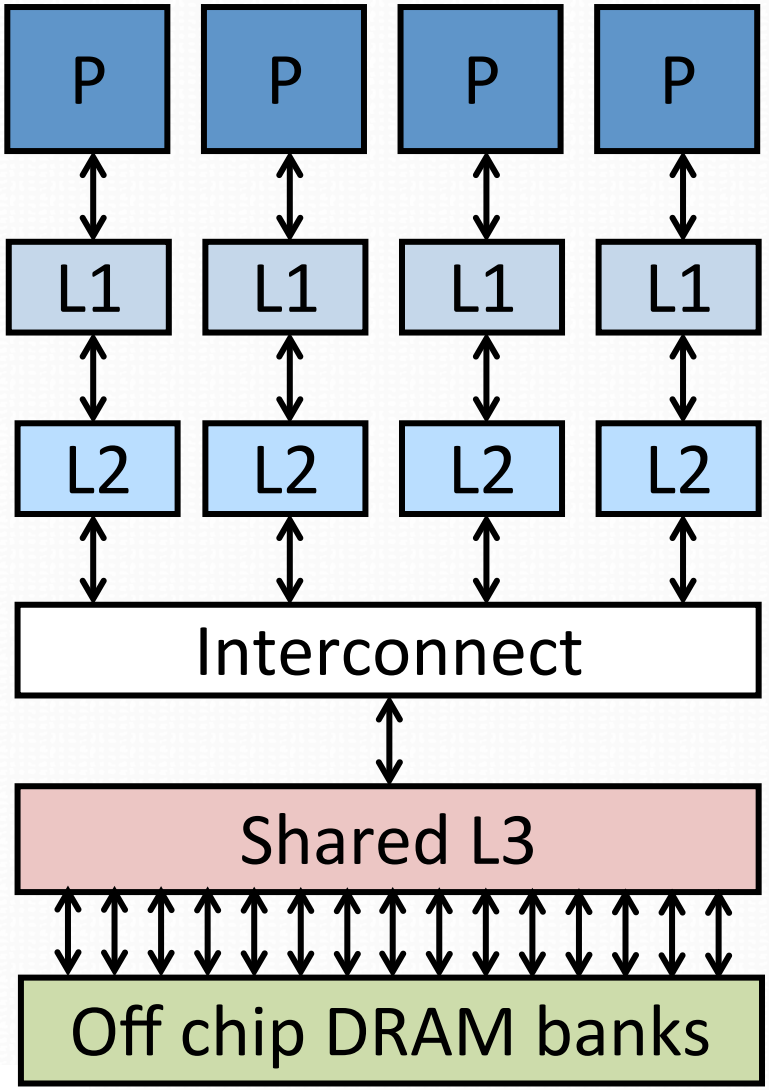


Source: C. Batten, Cornell

Outline

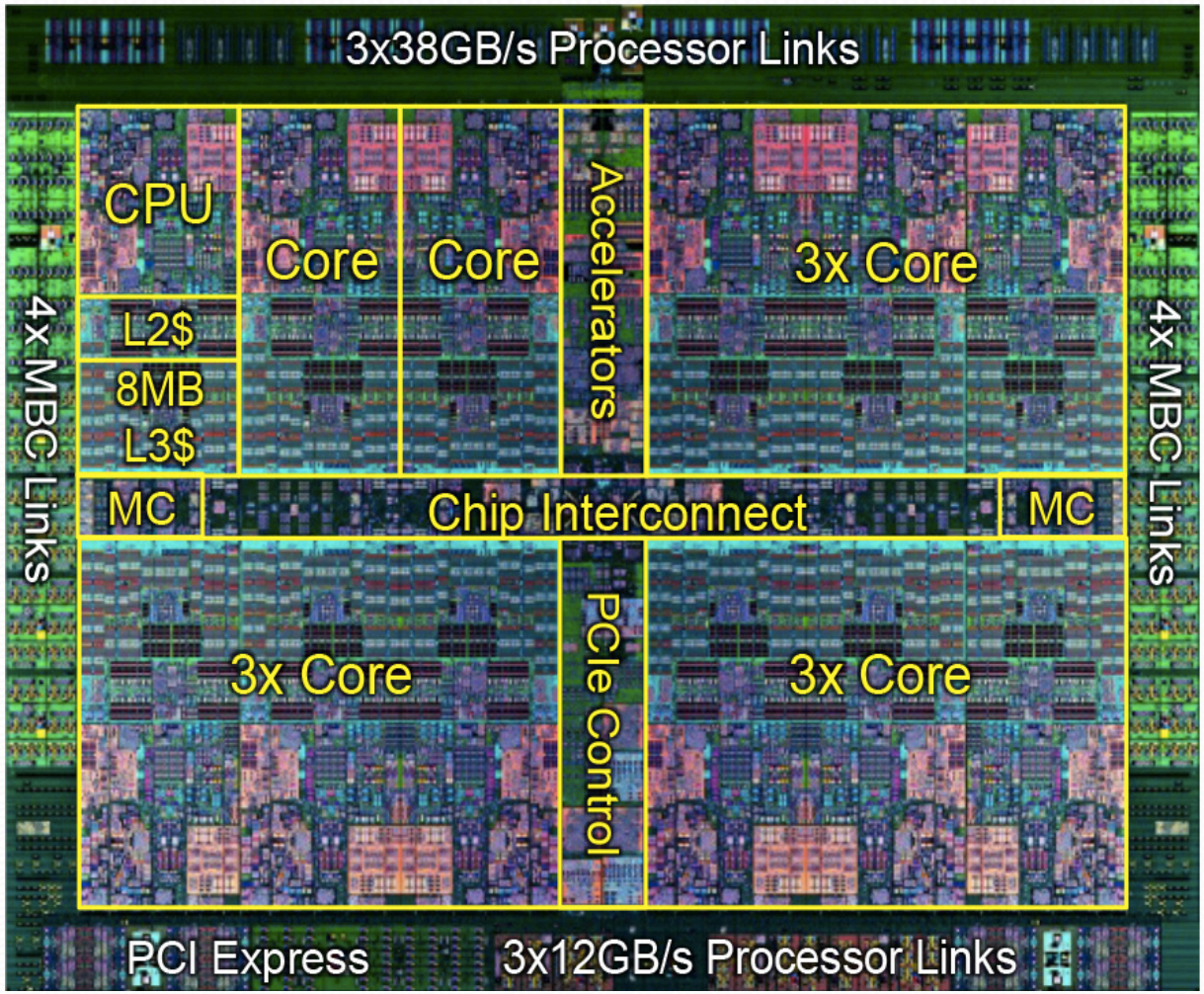
- Homogeneous manycores
  - Distributed shared memory
  - Distributed memory with message passing
- Bandwidth requirements of on-chip communications in manycores
- Can 3D help?
  - 3D memory stacking
  - Towards (3D) heterogeneous manycores

Shared-Memory Multiprocessor



- Processors communicate with shared address space by memory read/ write
- Hardware-managed, implicitly-addressed, coherent caches
- Bandwidth depends on
  - Cache size, associativity
  - Replacement policy, coherence protocol
  - Application requirements

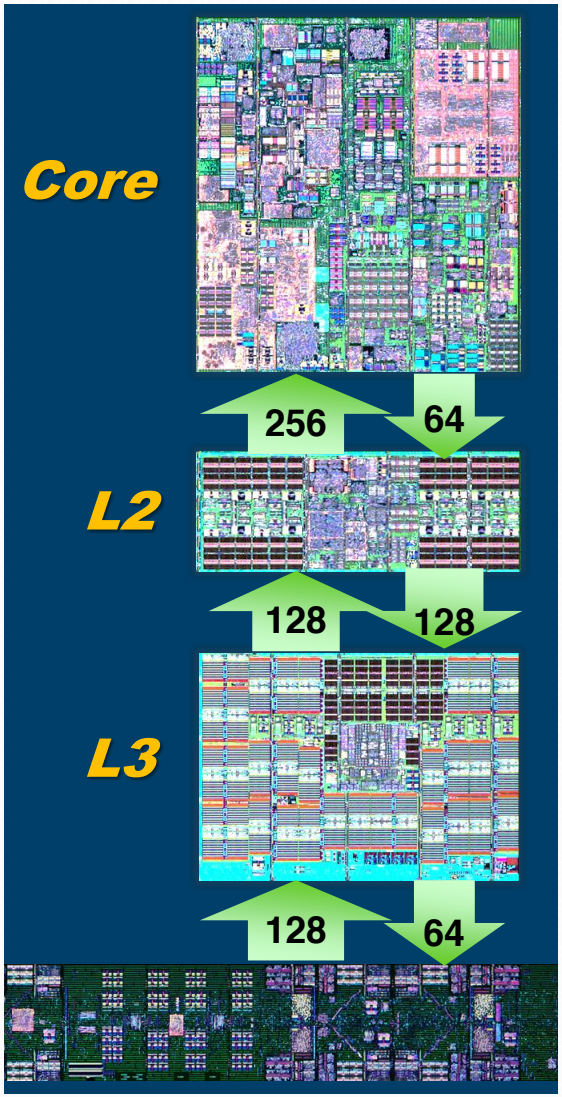
IBM Power 8



- 12 cores (SMT 8)
- 2013 (2015)
- 22nm, 6.5cm^2
- Caches
  - 512 KB SRAM L2 / core
  - 96 MB eDRAM shared L3
  - Up to 128 MB eDRAM L4 (off-chip)

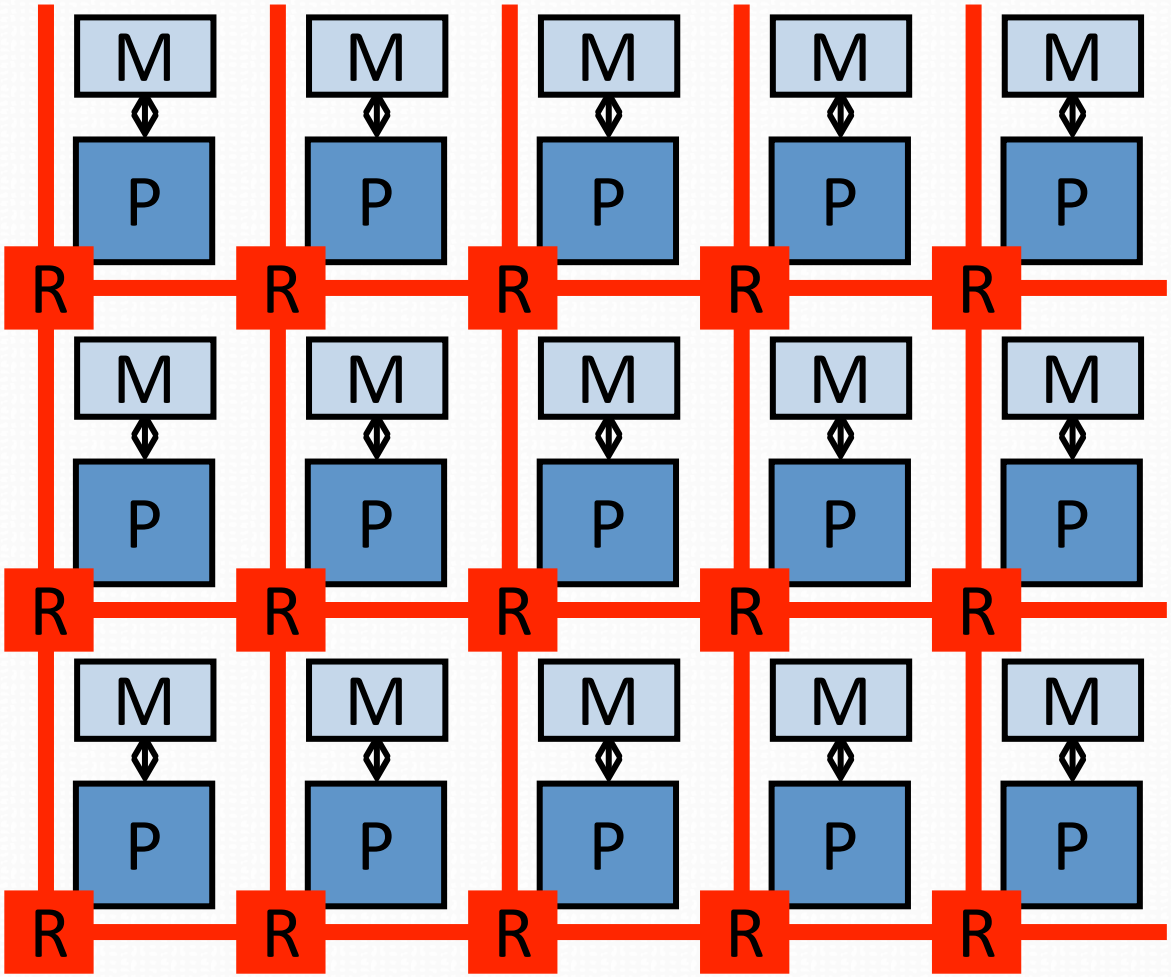
IBM Power 8

- Across 12 core chip
  - 4 TB/sec L2 BW
  - 3 TB/sec L3 BW
- 230 GB/s sustained external memory bandwidth



GB/sec shown assuming 4 GHz

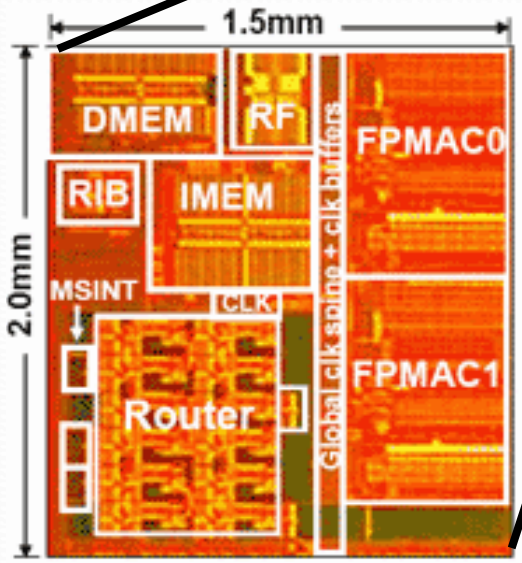
Distributed Memory with Message Passing (streaming memory)

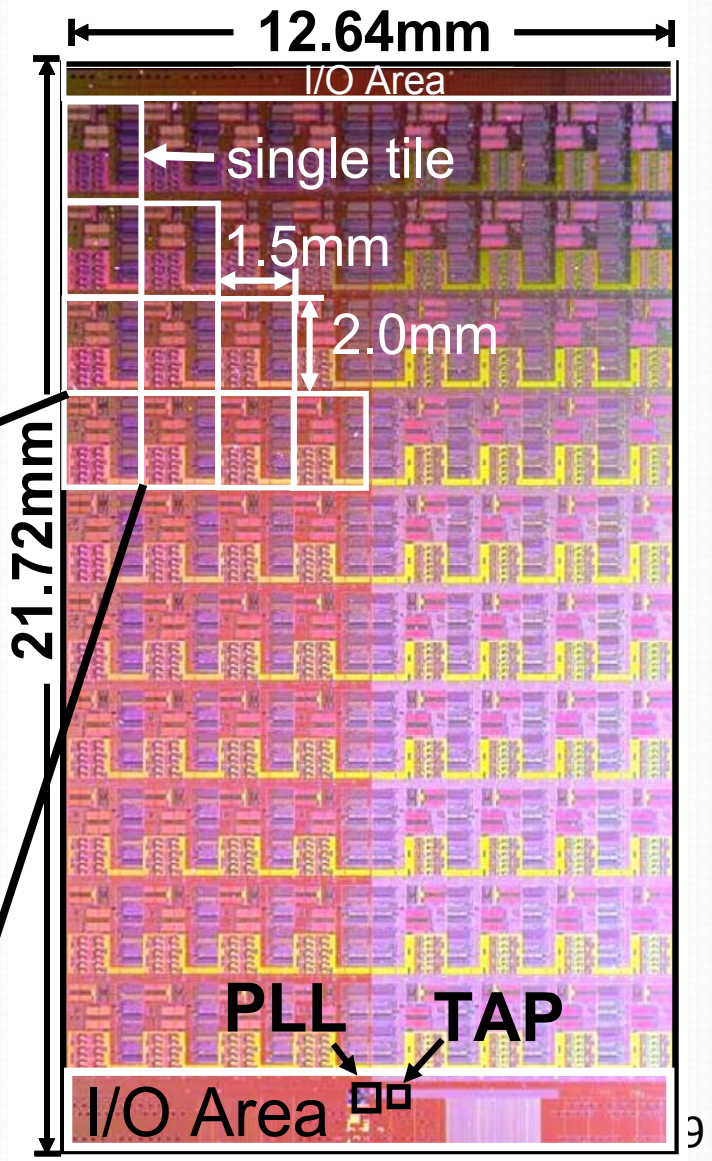


- Separate address space for each processor
- Processors communicate via message passing
- Software-managed, explicitly-addressed, local memories
- Bandwidth
  - Application requirements
  - Software code

Intel's 80 Core Terascale Processor

- 80 cores (2 FMACs)
- 1.6 SP TFOPS @ 5GHz 1.2V
- 320 GB/s bisection router bandwidth





Routers write directly into memory:

any core could write into the memory of any other core with low latency (2 cycles)

Message Passing vs. Shared Memory

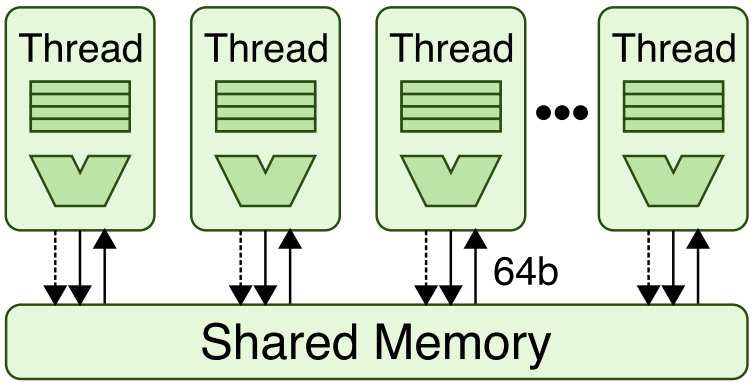
- Simpler and cheaper hardware
- Explicit communication is painful to program but makes programmers aware of costly operations
  - Requires manual optimization

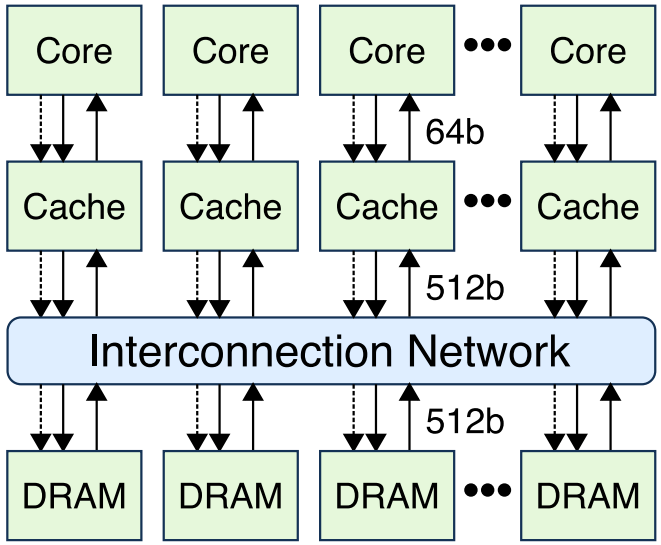
- Easier to write correct initial parallel programs (but often inefficient)
  - Gradually optimize them
- Cache coherence limits scalability
  - Coherence traffic may collide with useful communication

Outline

- Homogeneous manycores
  - Distributed shared memory
  - Distributed memory with message passing
- Bandwidth requirements of on-chip communications in manycores
- Can 3D help?
  - 3D memory stacking
  - Towards (3D) heterogeneous manycores

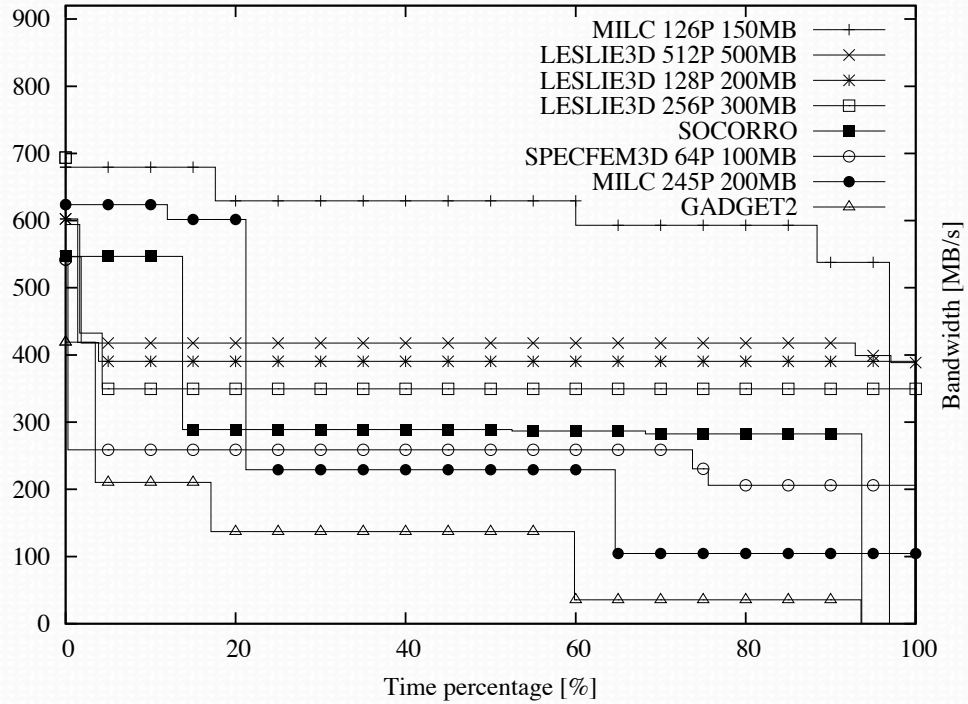
Threaded Application on DSM



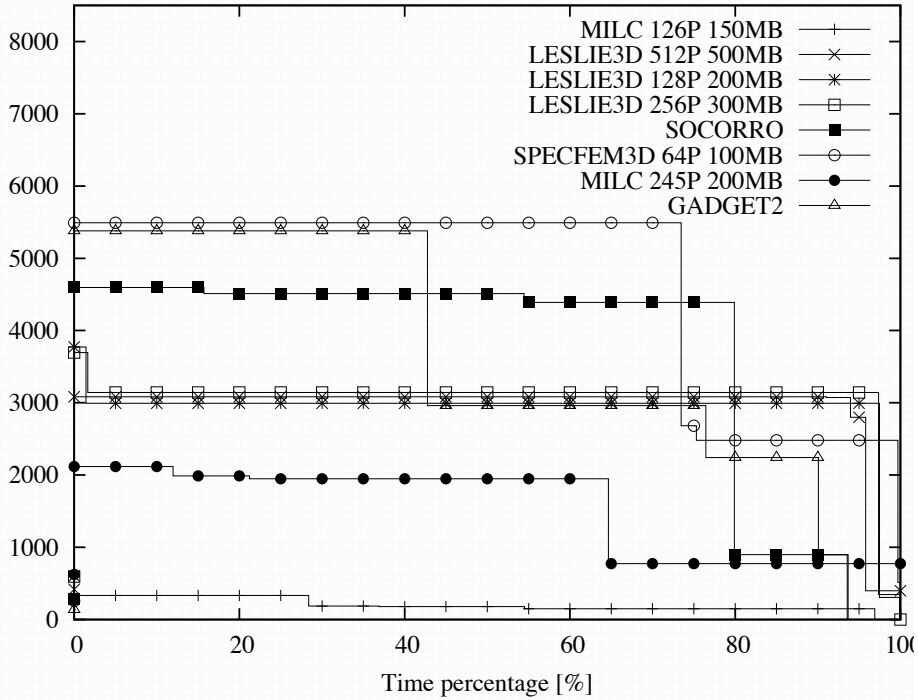


Application Requirements	
Message Size	512b
Message Bandwidth	400MB/s
Peak Bandwidth	8GB/s
Latency	Minimum
Traffic Pattern	Arbitrary

Memory Bandwidth on DSM



per-proc. off-chip memory bandwidth

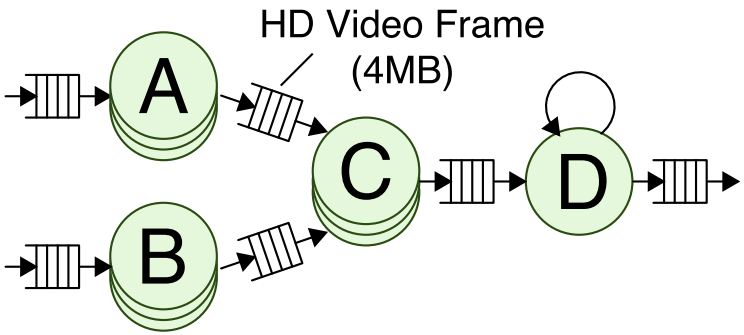


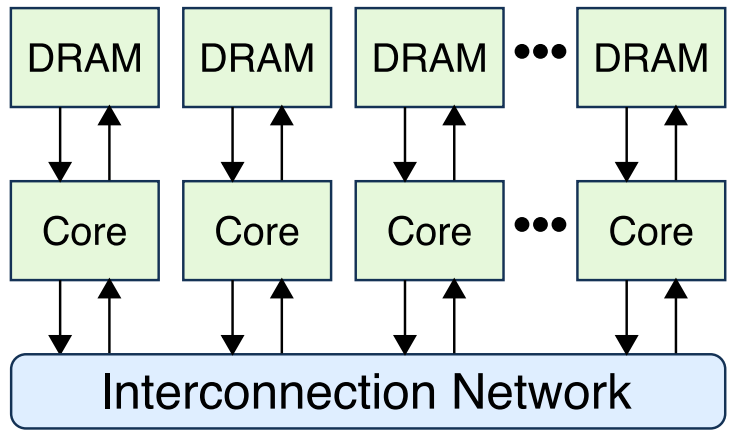
per-processor L2-L1 bandwidth

- Scaling linearly to 100 processors on a chip requires sustained off-chip memory bandwidth of 20-70 GB/s

[Pavlovic et al., ISCA'10]

Streaming Applications on DM



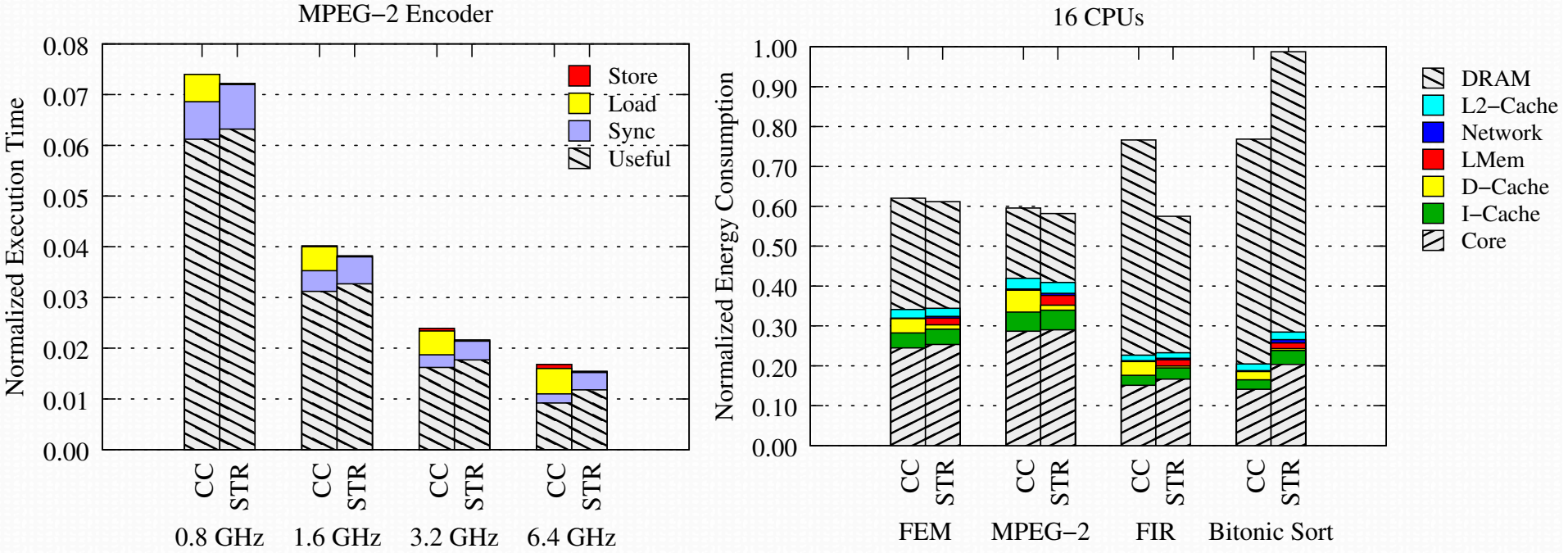


Application Requirements	
Message Size	4MB
Message Bandwidth	120MB/s
Peak Bandwidth	120MB/s
Latency	Tolerant
Traffic Pattern	Streaming

- Strongly dependent on application requirements

Message Passing vs. Shared Memory

- Both models perform and scale equally well
- Execution time and Energy on 16 cores
  - CC: shared memory; STR: distributed memory



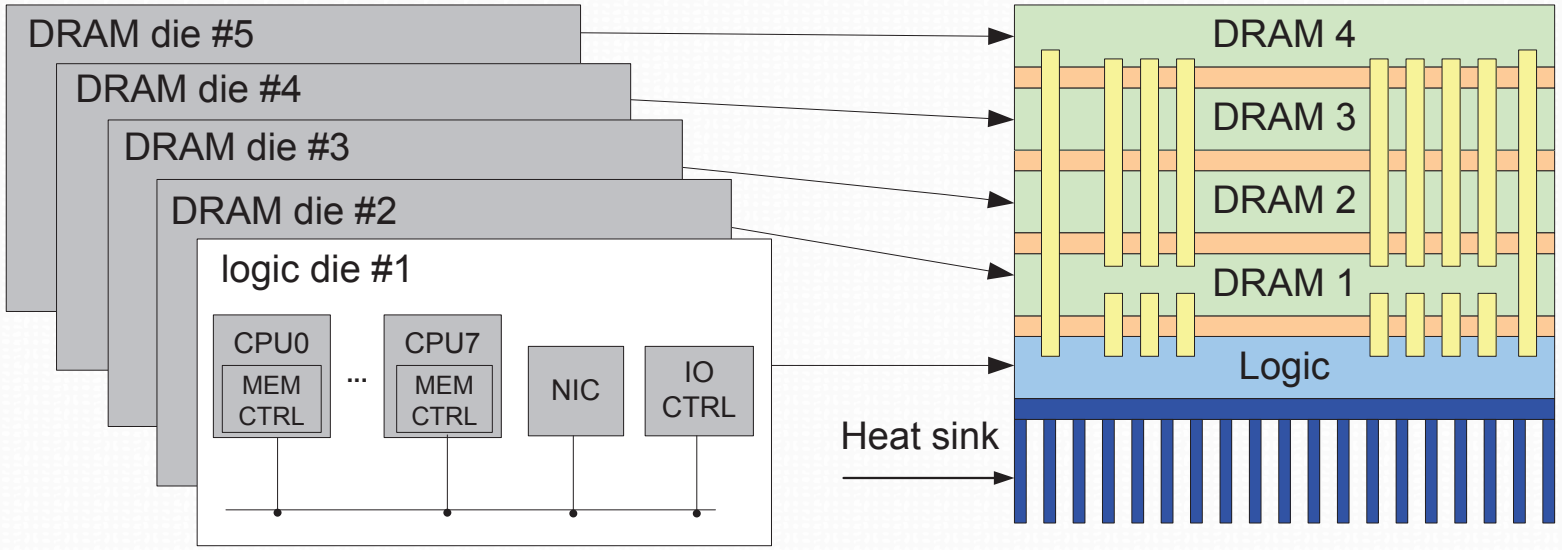
[Leverich et al., ISCA'07]

Outline

- Homogeneous manycores
  - Distributed shared memory
  - Distributed memory with message passing
- Bandwidth requirements and energy of on-chip communications in manycores
- Can 3D help?
  - 3D memory stacking
  - Towards (3D) heterogeneous manycores

PicoServer

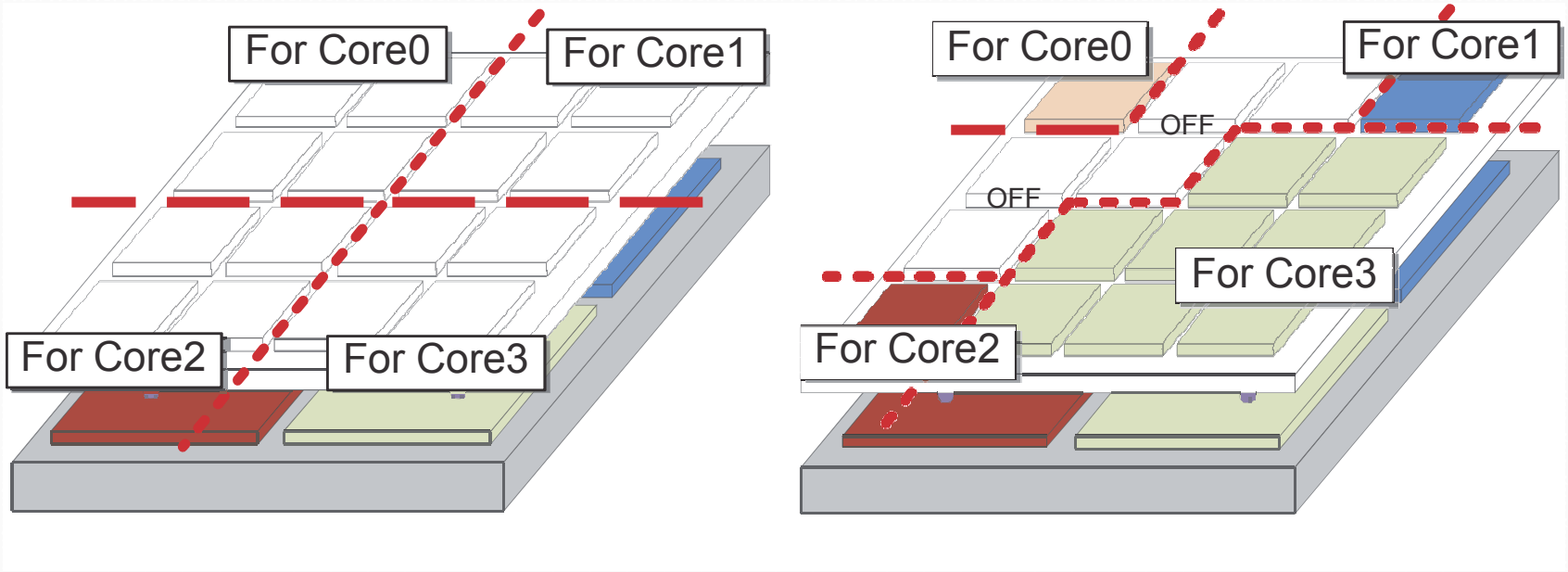
- Energy-efficient multicore architecture with 3D-stacked DRAM



[Kgil et al., SIGOPS'08]

3D Memory Stacks

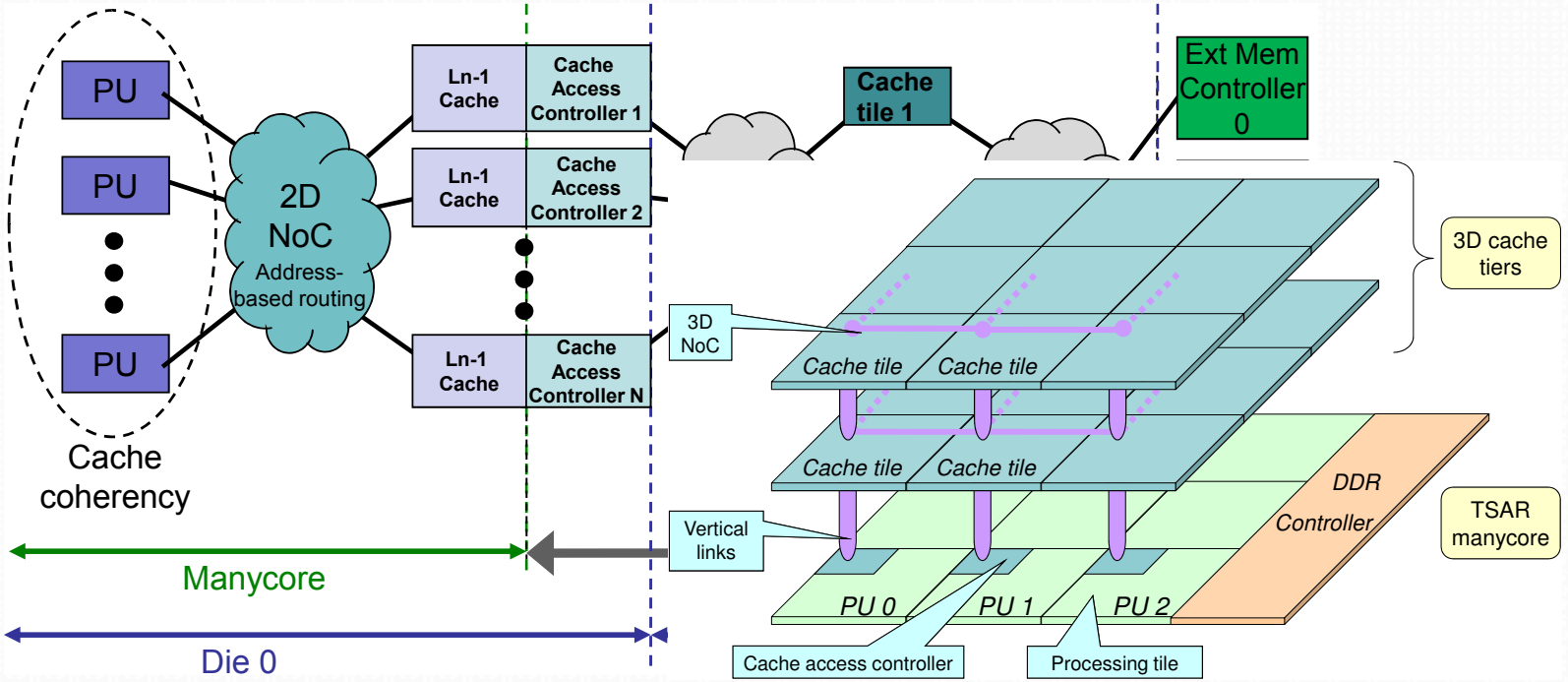
- Moving the compute closer to the data
- Non-Uniform Cache Architecture (NUCA)
  - Dynamic reconfiguration of cache structure



[Jung et al., GLSVLSI'11]

3D Memory Stacks

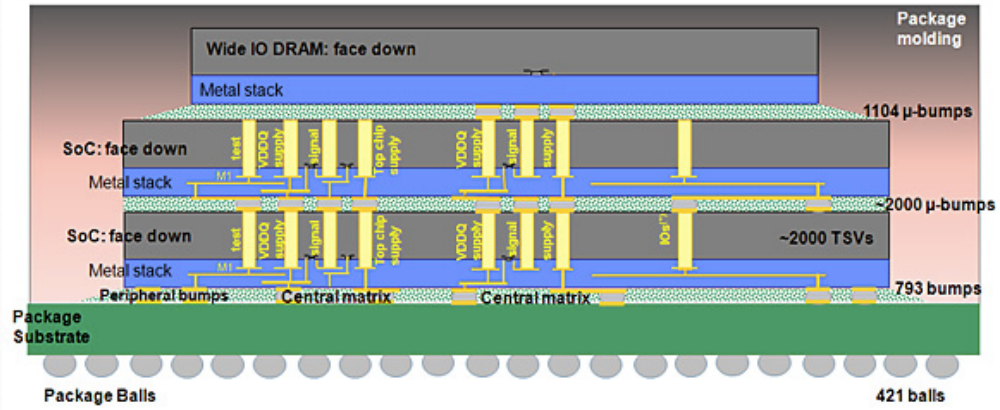
- Moving the compute closer to the data
- Non-Uniform Cache Architecture (NUCA)
  - Dynamic reconfiguration of cache structure



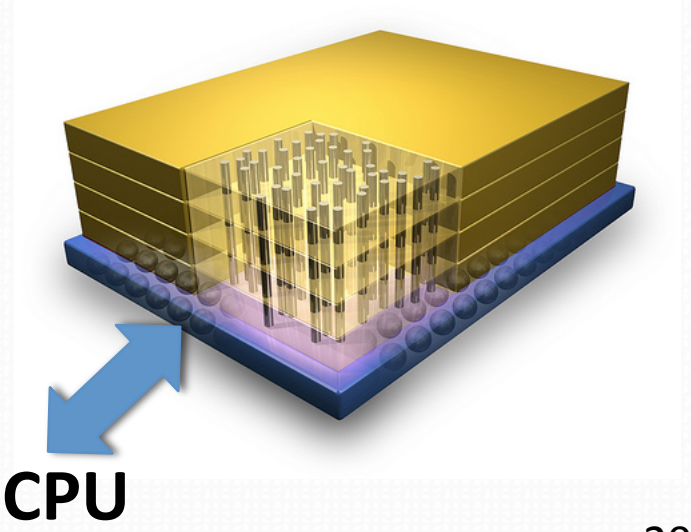
[Dutoit et al., DATE'13]

Wide I/O, Hybrid Memory Cube (HMC)

- Wide I/O memory interface
- 3D die stacks with TSVs
- 2.5D interposer
- 50GB/sec

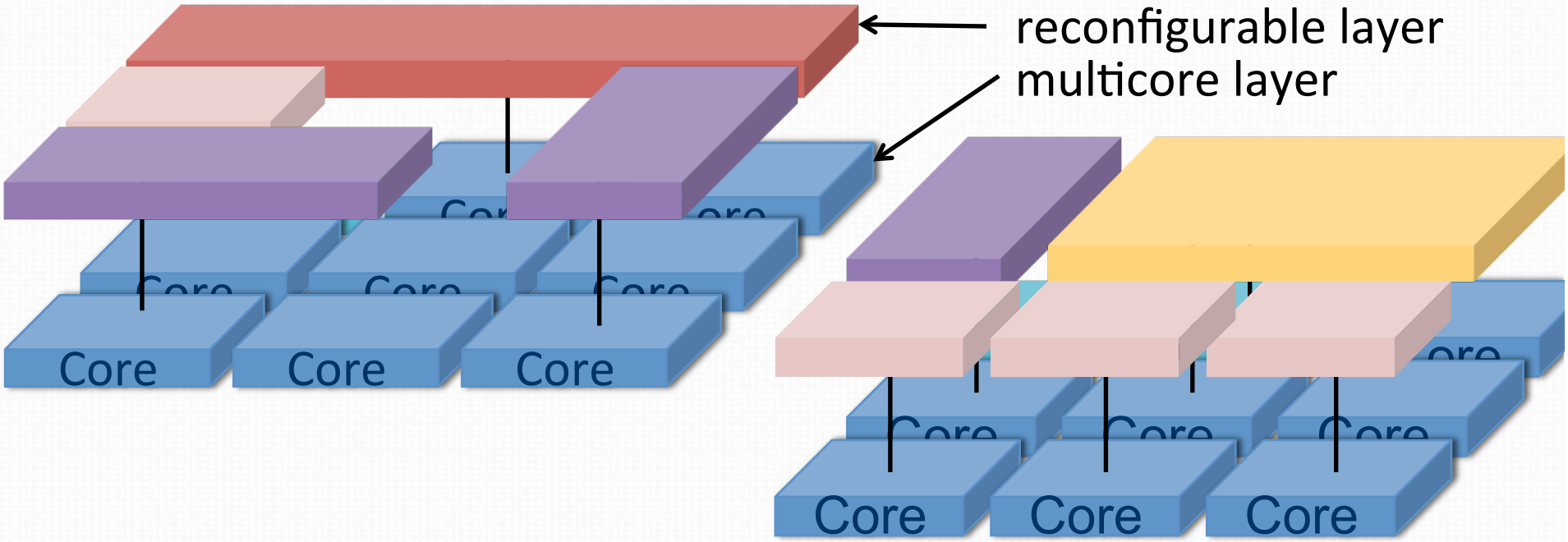


- Micron/Intel's HMC couples a logic layer with 3D-stacked DRAM
- 160GB/sec

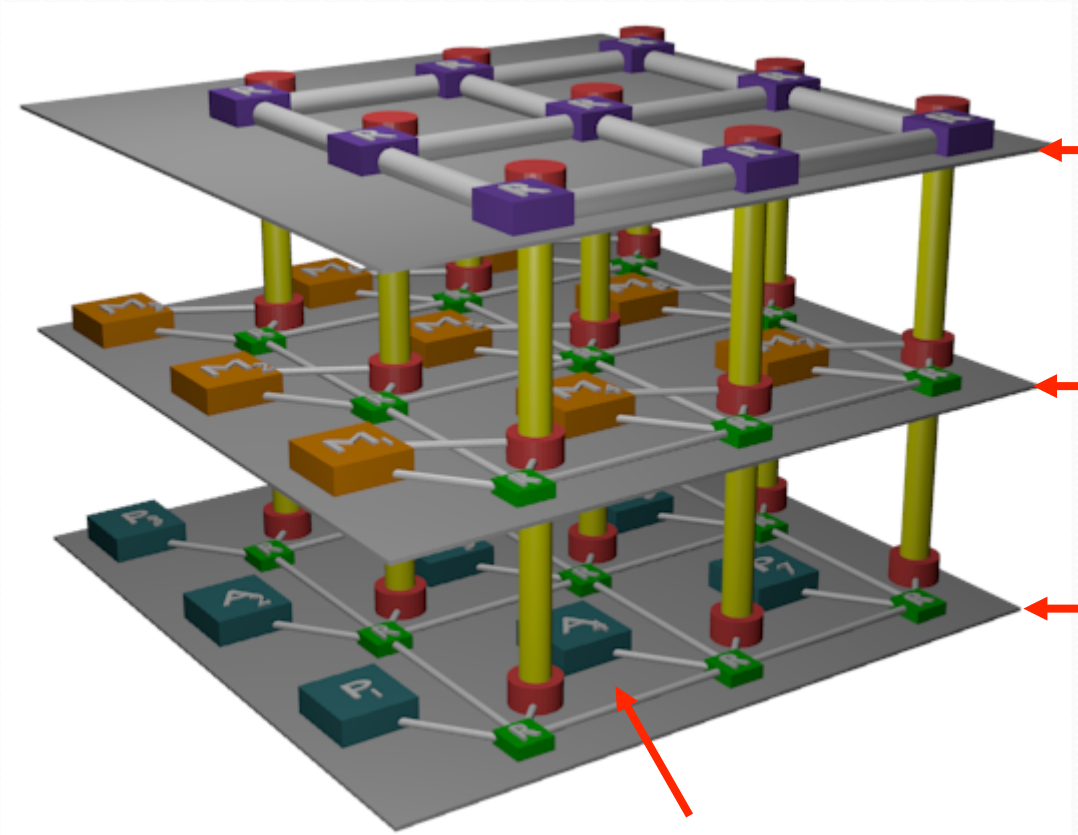


3D-Stacked Reconfigurable Hardware Accelerators

- Improved performance (3D coupling)
- Improved flexibility and resource usage



3D Optical Manycore Project



Conclusions

- On-chip and off-chip bandwidth increases significantly with number of cores and application requirements
  - And big data is coming!
- True for both shared memory and streaming memory models
- Energy of on-chip communications is expected to be the bottleneck

References

1. Jacob Leverich, Hideho Arakida, Alex Solomatnikov, Amin Firoozshahian, Mark Horowitz, and Christos Kozyrakis. Comparing memory systems for chip multiprocessors. In Proceedings of the 34th annual international symposium on Computer architecture (ISCA '07). ACM, New York, NY, USA, 358-368, 2007.
2. Hadi Esmaeilzadeh, Emily Blem, Renee St. Amant, Karthikeyan Sankaralingam, and Doug Burger. 2011. Dark silicon and the end of multicore scaling. SIGARCH Comput. Archit. News 39, 3 (June 2011), 365-376.
3. William J. Dally, James Balfour, David Black-Shaffer, James Chen, R. Curtis Harting, Vishal Parikh, Jongsoo Park, David Sheffield "Efficient Embedded Computing" IEEE Computer, July 2008.
4. William J. Dally, keynote at IPDPS 2011.
5. Jongpil Jung, Kyungsu Kang and Chong-Min Kyung, Design and management of 3d- stacked NUCA cache for chip multiprocessors. *In Proc. of the 21st edition of the great lakes symposium on Great lakes symposium on VLSI*, GLSVLSI '11, pages 91–96, 2011. ISBN 978-1-4503-0667-6.
6. Denis Dutoit, Eric Guthmuller and Ivan Miro-Panades, 3d integration for power- efficient computing. *In Design, Automation Test in Europe Conference Exhibition (DATE), 2013*, mars 2013.
7. Taeho Kgil, Shaun D'Souza, Ali Saidi, Nathan Binkert, Ronald Dreslinski, Trevor Mudge, Steven Reinhardt and Krisztian Flautner, PicoServer: using 3D stacking technology to enable a compact energy efficient chip multiprocessor. *SIGOPS Oper. Syst. Rev.*, 40(5):117–128, octobre 2006. ISSN 0163-5980.
8. Shekhar Borkar, Andrew A. Chien, The Future of Microprocessors, Communications of the ACM, Vol. 54 No. 5, Pages 67-77, 10.1145/1941487.1941507
9. Milan Pavlovic, Yoav Etsion, and Alex Ramirez. Can manycores support the memory requirements of scientific applications?. In Proceedings of the 2010 international conference on Computer Architecture (ISCA'10), Ana Lucia Varbanescu, Anca Molnos, and Rob Nieuwpoort (Eds.). Springer-Verlag, Berlin, Heidelberg, 65-76, 2010.

Challenges for optical interconnects: computer architect's point of view

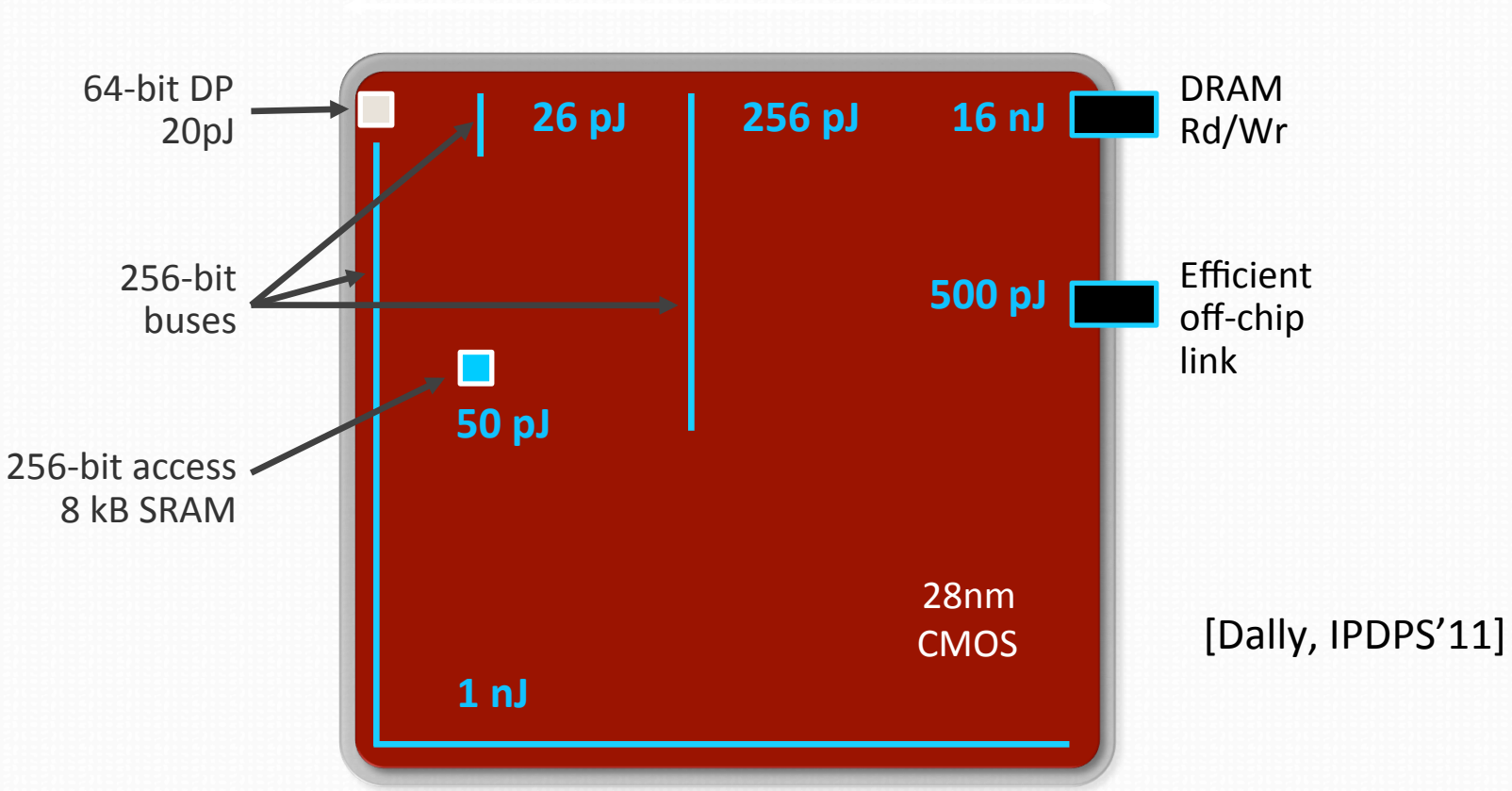
- Are we really bandwidth-limited for on-chip communications?
  - Not really true in current computer architectures
  - But big data will maybe change the game
- What are the big challenges?
  - Amdhal's law
    - Even with 95% of parallel code, speedup *S < 20*

$$S(N) = \frac{1}{(1-P) + \frac{P}{N}}$$

  - Power!
    - Energy efficiency is not scaling along with integration capacity
    - Data movement costs more than computing
- So maybe energy efficiency is the real challenge

The Energy Cost of Data Movement

- Fetching operands costs more than computing



[Dally, IPDPS'11]

The Energy Cost of Data Movement

- Future processor up to 3 Tera-op/sec
- At minimum requires 64b x 9 Tera-operands to be moved each second
- If on average 1mm (10% of die size) then
  - 0.1pJ/bit x 576 Tbits/s
  - consumes 58 Watts!