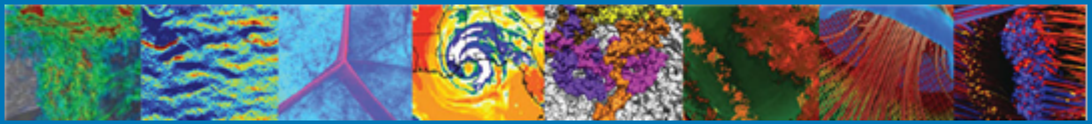


Micro Architecture for Exascale

Argonne Training Program on Extreme-Scale Computing 2013

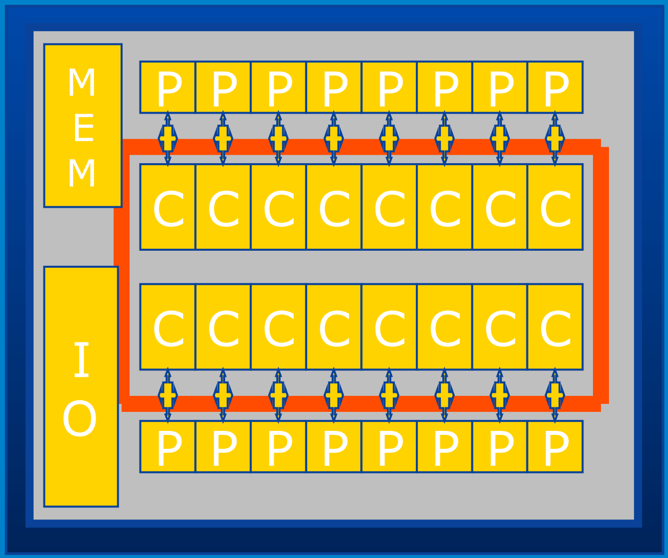
Tryggve Fossum

Computer Architect

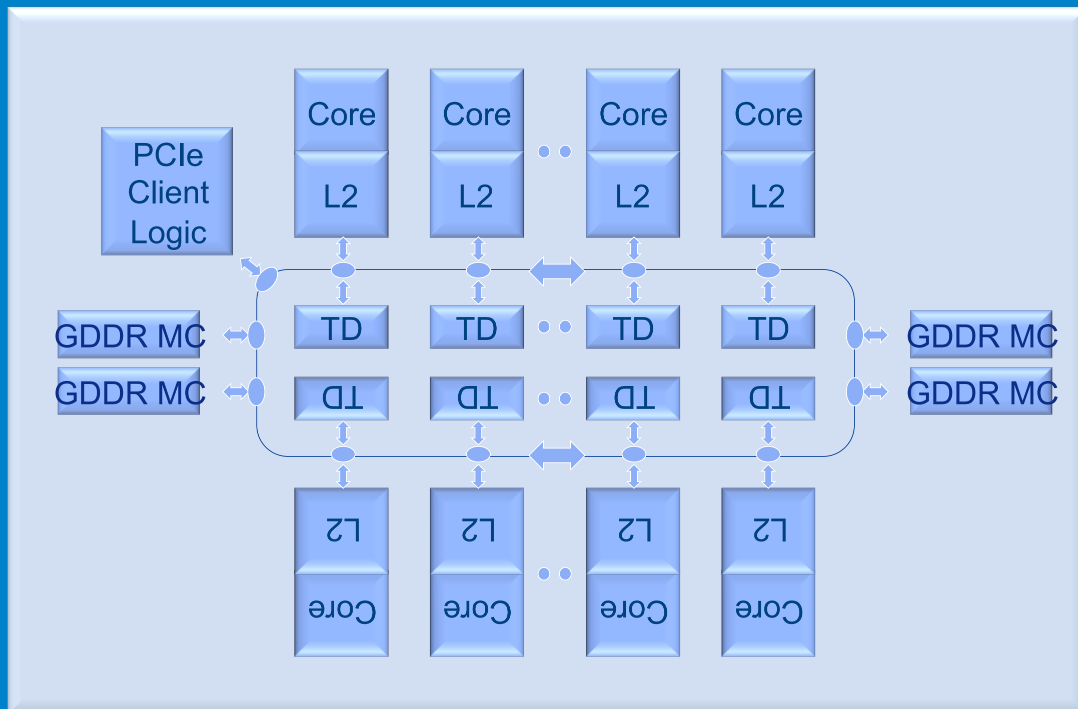
Legal Information

CMP Systems Benefits

- On die interconnect:
  - Higher Bandwidth: TB/s vs. GB/s
  - Shorter Latency: ns vs. 100 ns
  - MilliWatts vs. Watts
- Shared On-die Cache
  - Fast Communication
  - Better hit rate
  - Faster synchronization
  - Less false sharing
- Memory attached to single socket
  - Simplifies System Design
  - Reduces NUMA effects
  - Simplifies application development
  - Simplifies performance tuning
- On-die performance scaling can be almost linear with core count



Knights Corner Micro-architecture



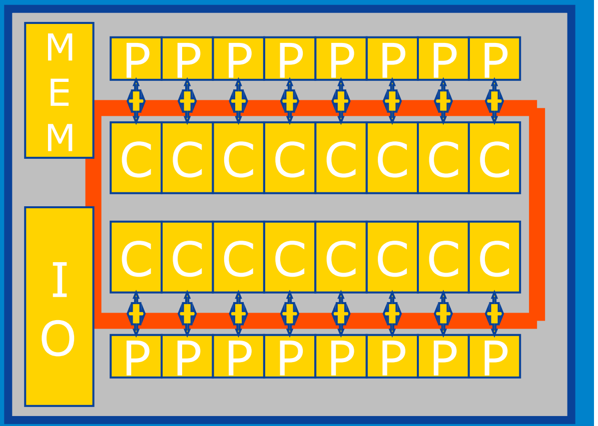
Design Example: A Ring-based On Die Interconnect

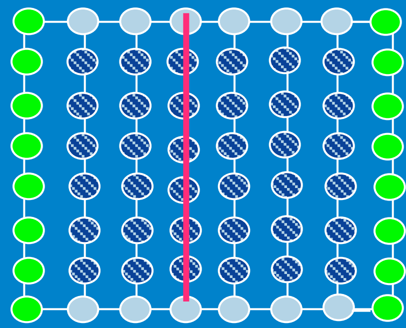
- Topology: Ring shaped Pipelined Bus connects array of processor cores to a distributed, multi-access, cache

Ring BW ~ Frequency x Width x Stops / Distance

  - While a single, unidirectional ring can work, two counter-rotating rings cut average distance in half
  - Adding stops, adds raw bandwidth, canceling out added occupancy
  - Data path wide enough to minimize Serialization effects
- Additional rings can be targeted to shorter, more specialized messages.
- Power encoding to minimize power and di/dt.
- Ring Latency grows with total distance, but is reasonable for a *shared,* higher level cache, close to Manhattan distance wire delay.

Off-die, such a network would be costly. On-die, the regular layout makes it practical.



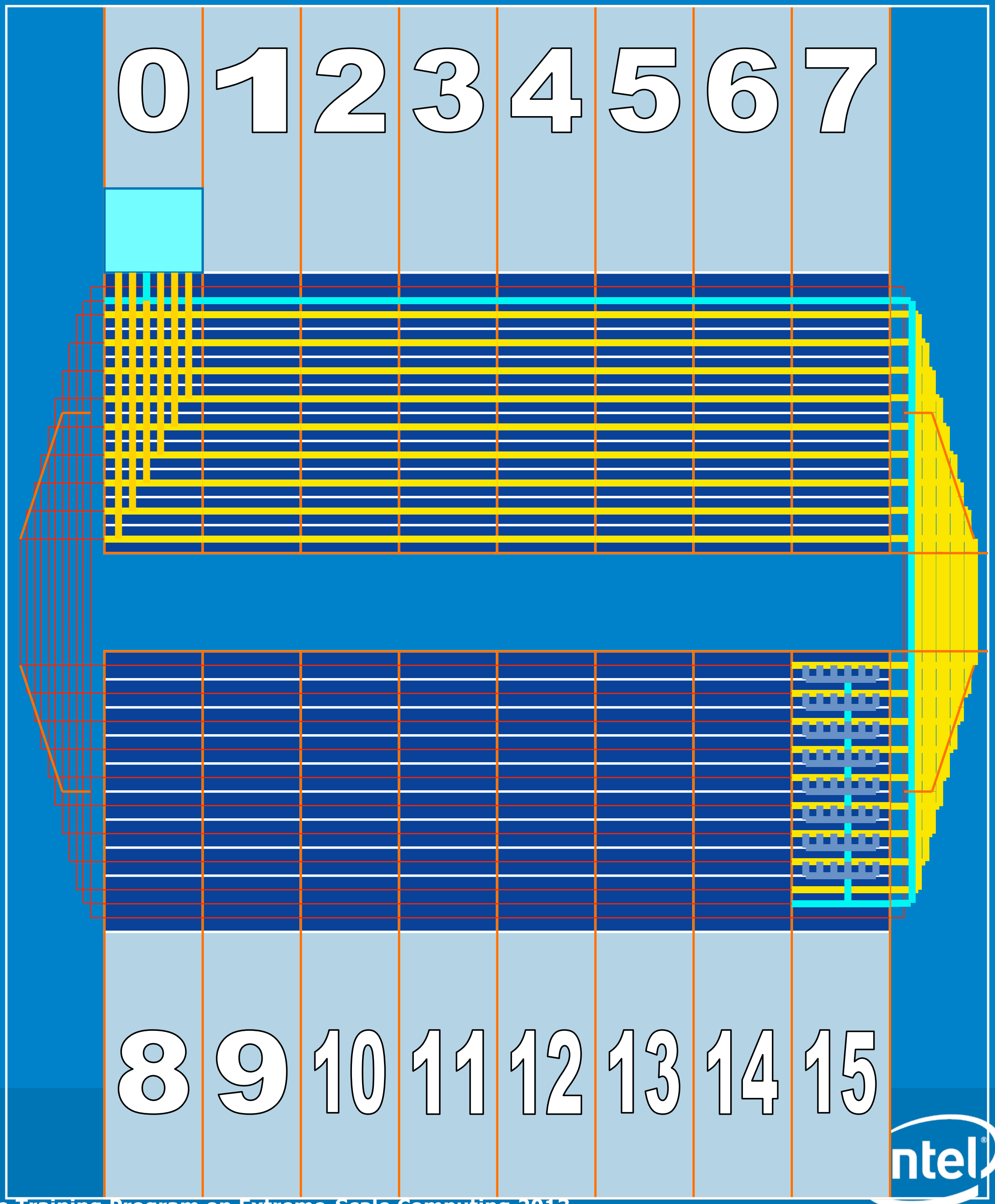


L3 Read Request
(hit)

Request, CORE[0] sends address to Cache[15].

Cache[15] performs cache line read.

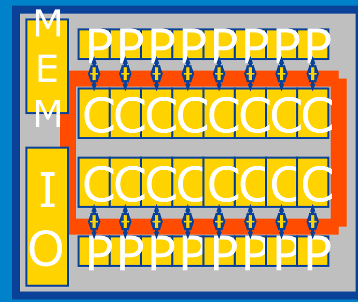
Response, Cache[15] returns data to CORE[0].

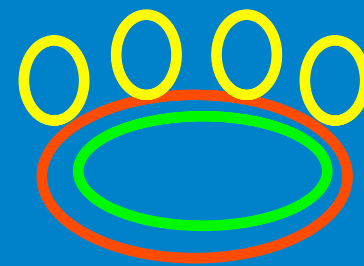


Ring Design Considerations

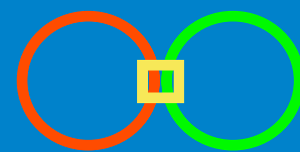
- Flow Control: Rotary rule.
  - Once a packet is on the ring, it can keep going with priority over incoming traffic. No further arbitration. No store and forward.
  - What if destination cannot sink packet?
    - Common network challenge, multiple options
- Routing Algorithm:
  - Greedy – pick ring with shortest distance
    - Minimizes best case latency
    - Not the best for worst case traffic (Tornado example)
    - Hashing algorithms for allocating memory blocks to cache segments eliminate worst case
  - Adaptive:
    - With Rotary rule, congestion shows up as not getting on the ring at some stop, making adaptive routing more beneficial.
    - Select ring that minimizes a x D + b x Q, where D is distance and Q is number of Queue entries.

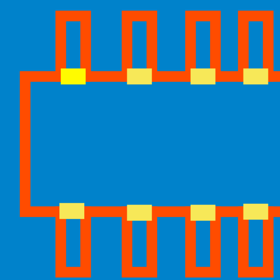
Ring Topologies

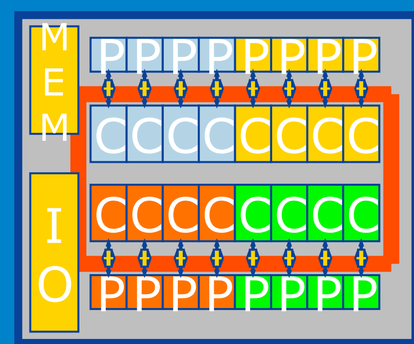




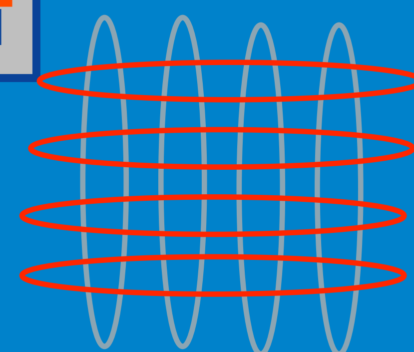
- Floorplan matches CMP layout
- Very simple, but does it scale?
  - Wider, faster
  - Multiple Rings
  - Hierarchical rings
    - May mean extra buffering
    - More complex flow control
    - Locality dependent



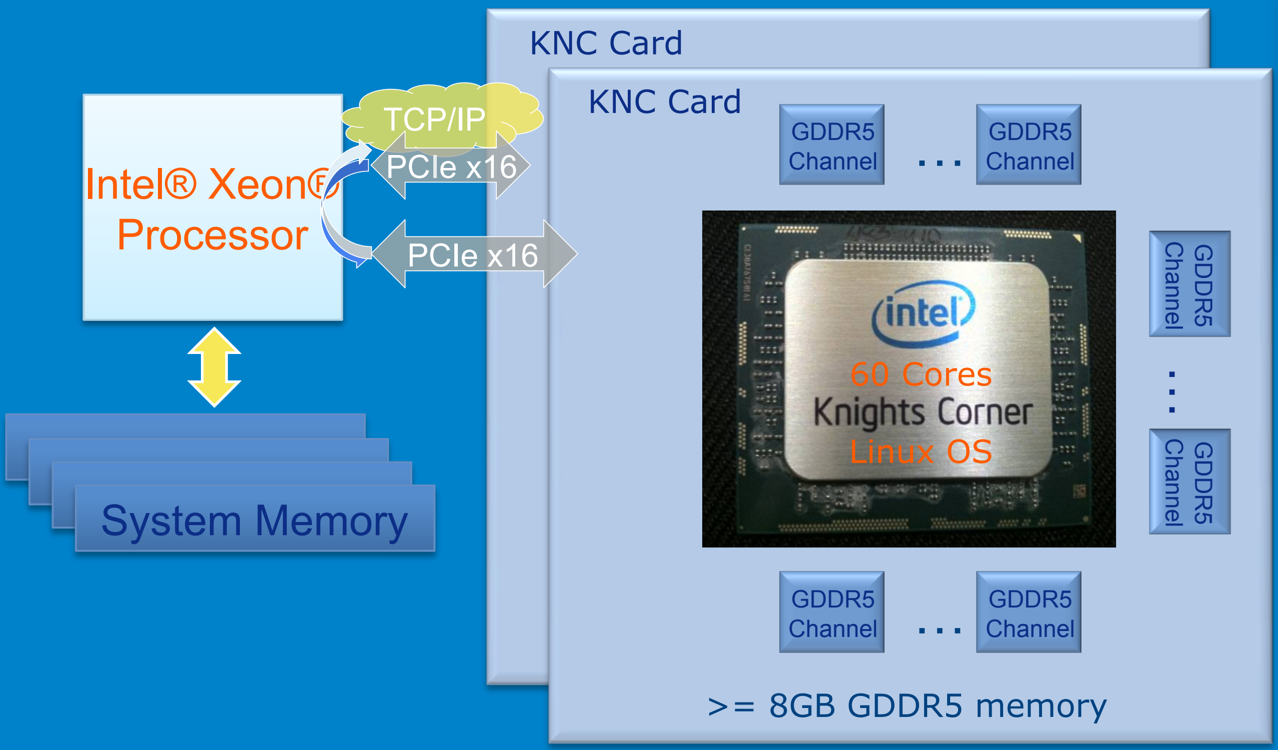


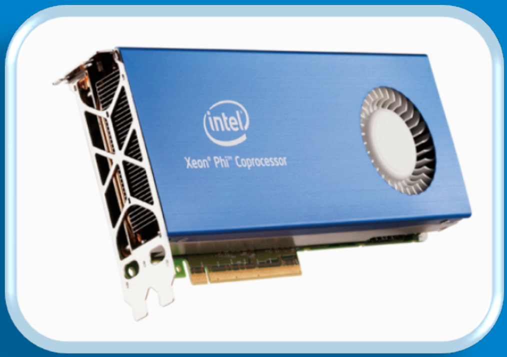


- Possible to increase locality:
  - Affinity aware caching
    - Bypass access to local ring cache
  - Divide ring cache into private segments
    - Most accesses to near-by ring stops
    - Use ring for global snoops and invalidates
- Mesh of Rings
  - Keep benefits of rings
  - Watch corner turns



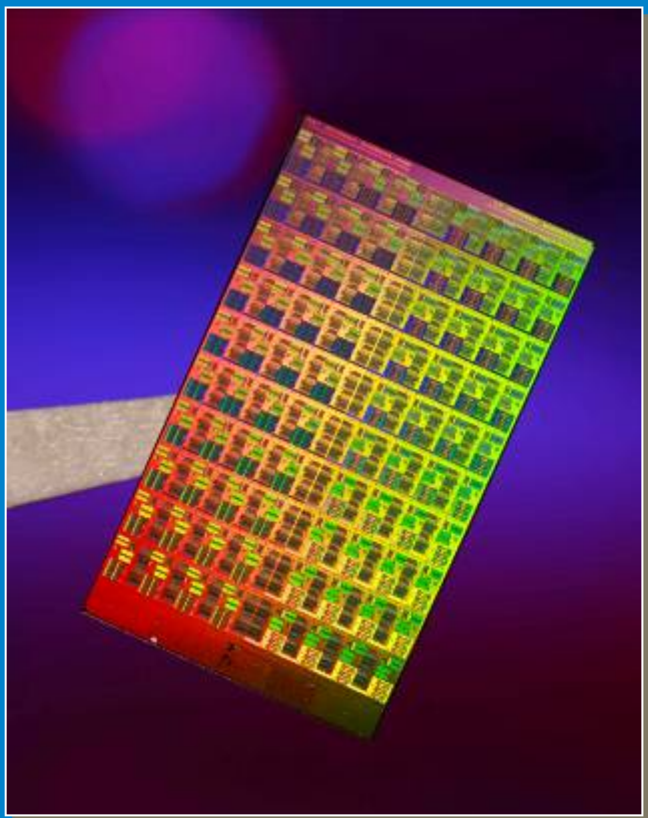
Knights Corner Coprocessor





Teraflops Research Chip

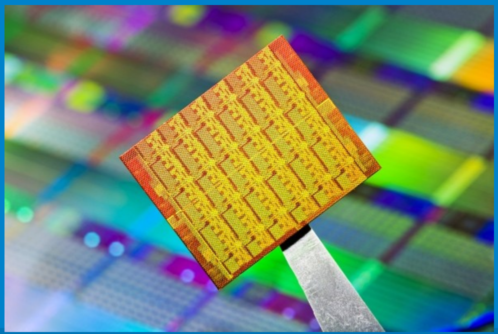
100 Million Transistors ● 80 Tiles ● 275mm^2



First tera-scale programmable silicon:

- Teraflops performance
- Tile design approach
- On-die mesh network
- Novel clocking
- Power-aware capability
- Supports 3D-memory

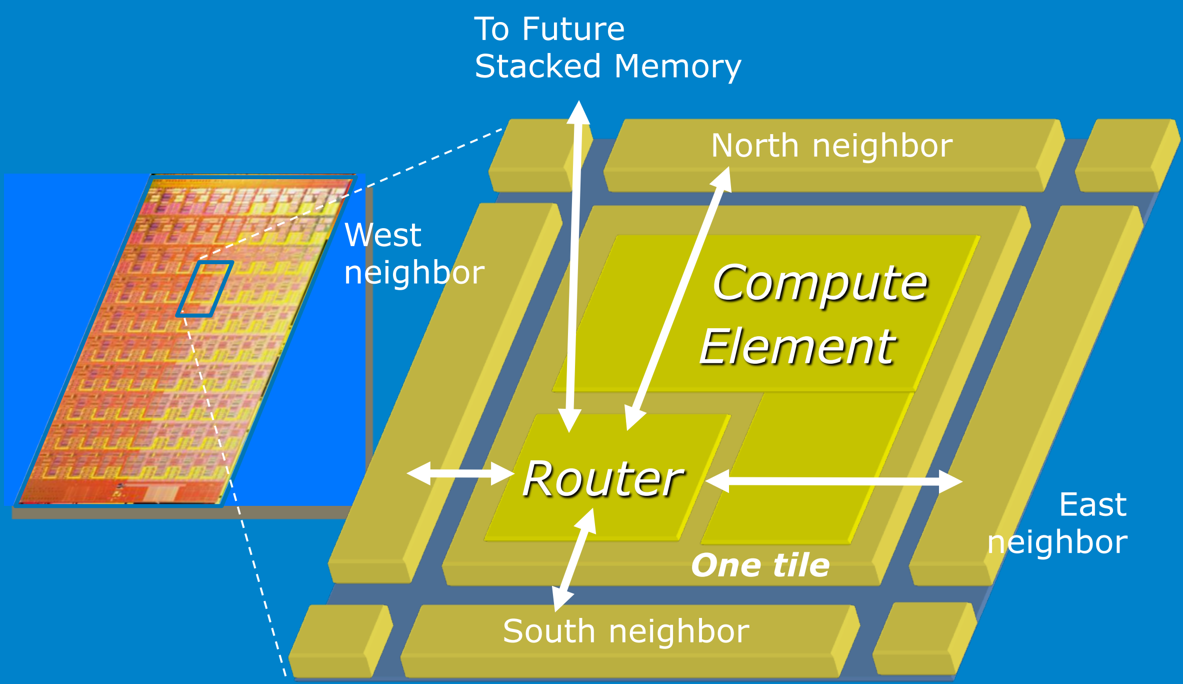
Not designed for IA or product



Follow On: Rock Creek
48 IA cores

[Slides from Intel CTG: Jerry Bautista, et al]

Tiled Design & Mesh Network



Repeated Tile Method:

Compute + router

Modular, scalable

Small design teams

Short design cycle

Mesh Interconnect:

"Network-on-a-Chip"

- Cores networked in a grid allows for super high bandwidth communications in and between cores

5-port, 80GB/s* routers

Low latency (1.25ns*)

* When operating at a nominal speed of 4GHz

Fine Grain Power Management

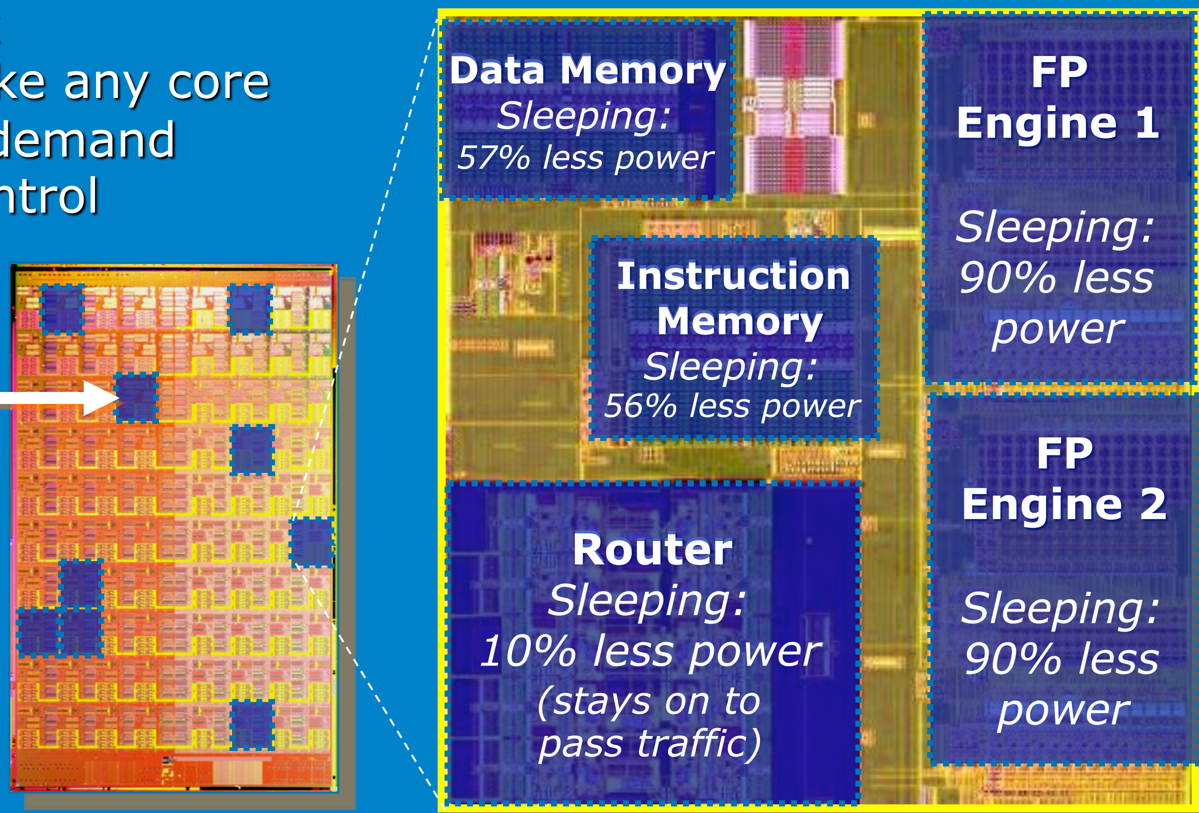
- Novel, modular clocking scheme saves power over global clock
- New instructions to make any core sleep or wake as apps demand
- Chip Voltage & freq. control
  (0.7-1.3V, 0-5.8GHz)

Dynamic sleep

STANDBY:
- Memory retains data
- **50%** less power/tile

FULL SLEEP:
- Memories fully off
- **80%** less power/tile

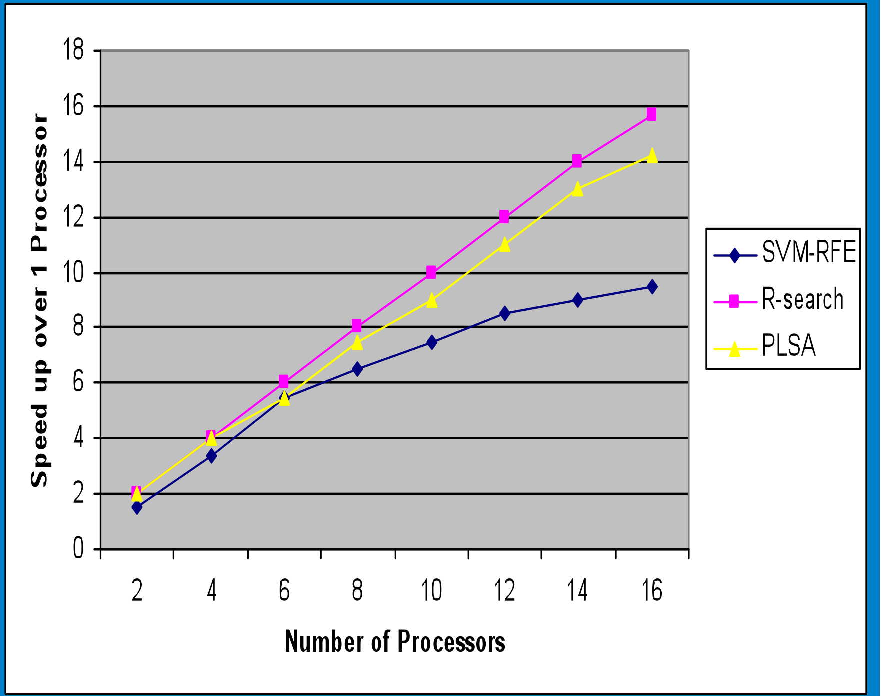
21 sleep regions per tile *(not all shown)*



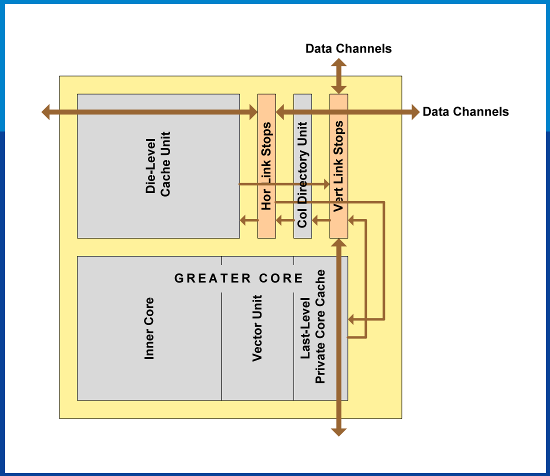
Industry leading energy-efficiency of 16 Gigaflops/Watt

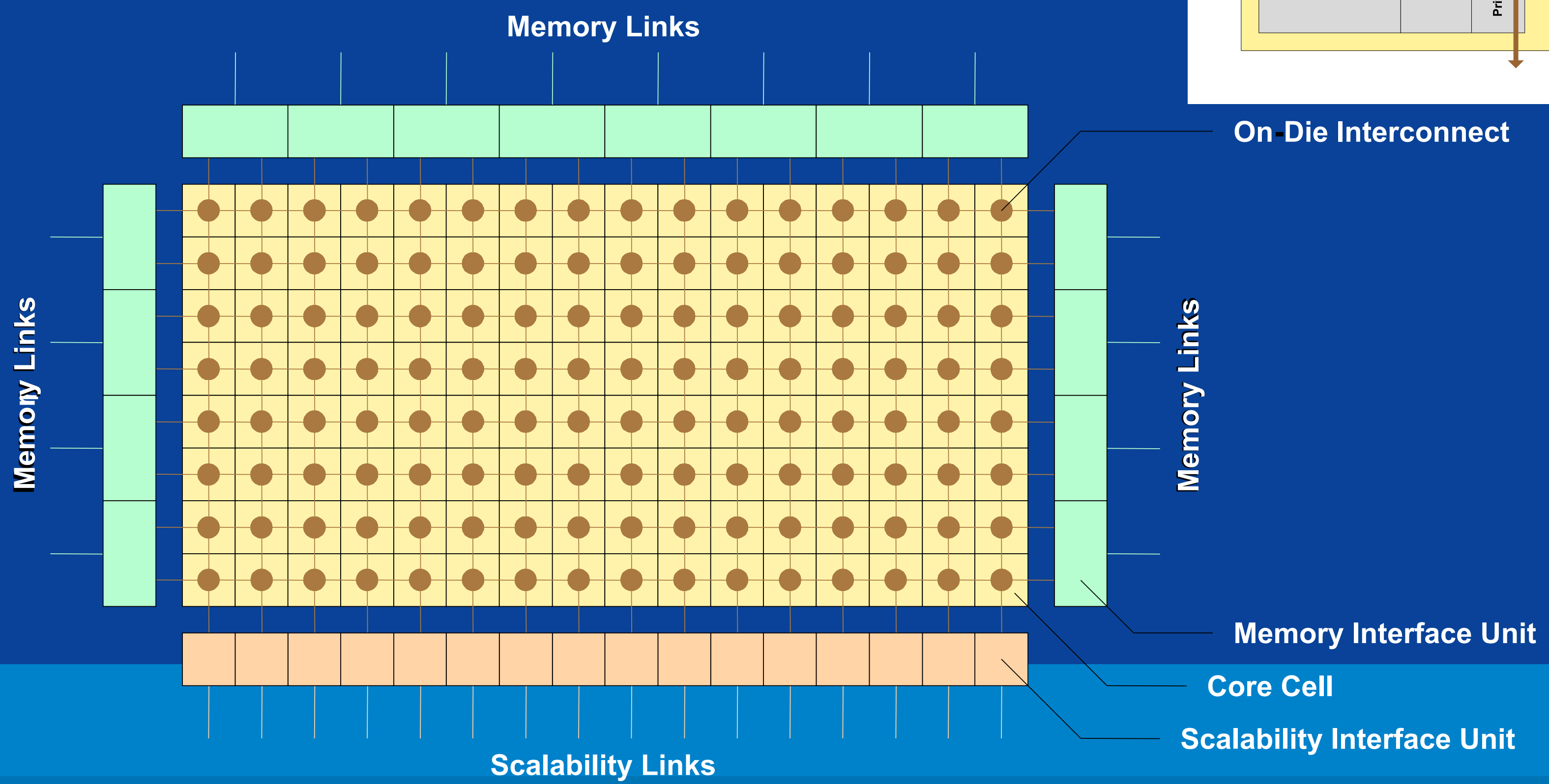
Limits to CMP

- CMP concentrates computation in a single chip, but…
- CMP also concentrates demand for
  - Power and Cooling
  - Memory Bandwidth
  - Memory Capacity
  - Network and IO Bandwidth
  - Cache capacity
- Diminishing returns for large core counts
  - Communication latencies increase
  - NOC Bandwidth per core may drop
- Complications
  - Process variation, Clock skew
  - Visibility, Fault isolation, Test
  - Power management: Current limits, di/dt, hot spots

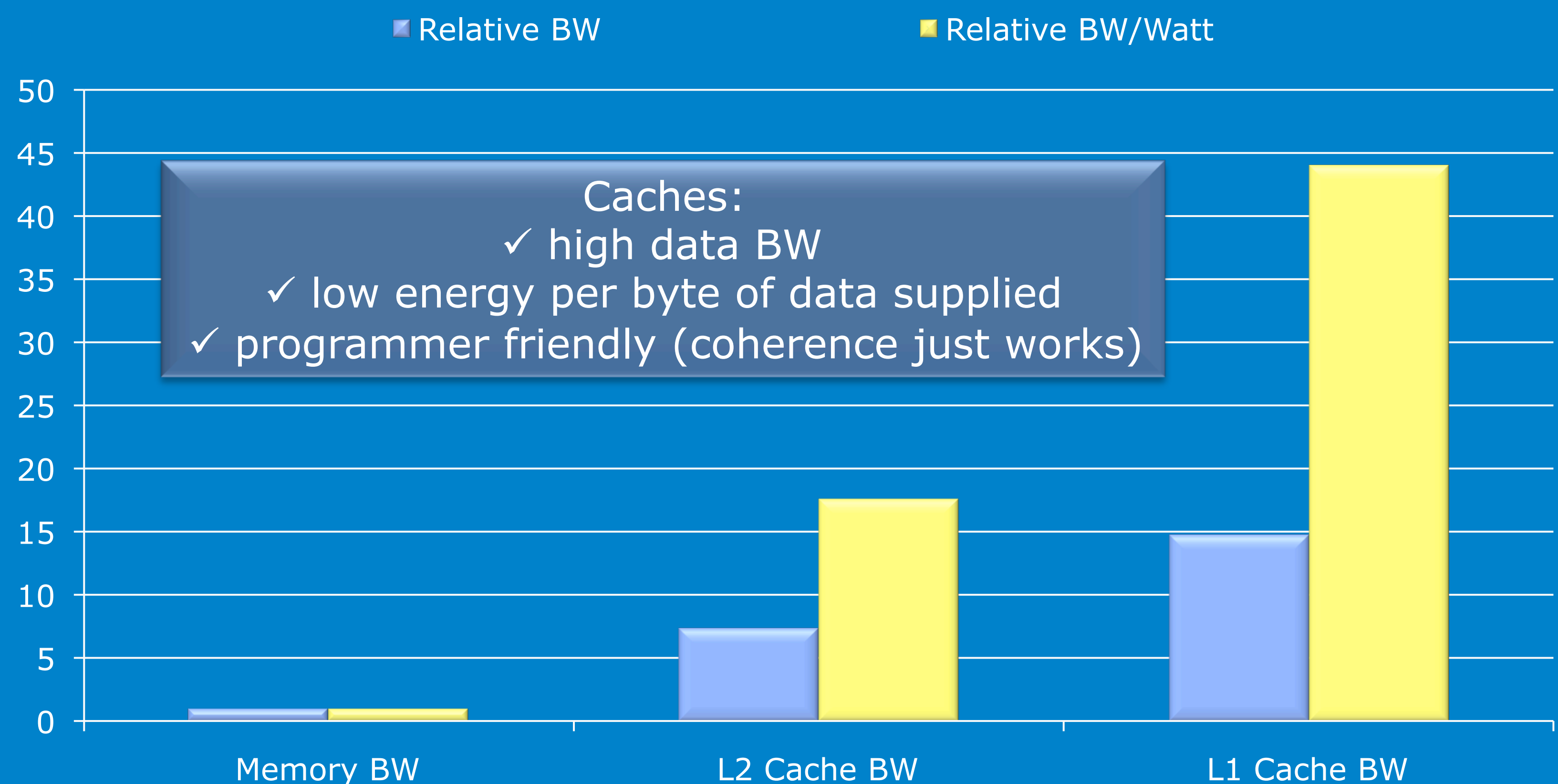


128-Core CMP


Data Channels
Data Channels
Die-Level Cache Unit
Hor Link Stops
Col Directory Unit
Vert Link Stops
GREATER CORE
Inner Core
Vector Unit
Last-Level Private Core Cache


Memory Links
Memory Links
Memory Links
On-Die Interconnect
Memory Interface Unit
Core Cell
Scalability Interface Unit
Scalability Links

Caches – For or Against?



Coherent Caches are a key MIC Architecture Advantage

Results have been simulated and are provided for informational purposes only. Results were derived using simulations run on an architecture simulator or model. Any difference in system hardware or software design or configuration may affect actual performance.

Cache Coherence

- Integration creates gap between integrated and non-integrated functions, such as Main Memory
- The architectural solution has been bigger, better caches
- Which can create coherence bottleneck and design complexity
- Want low (no?) coherence cost when there is no sharing
- Must work well when there is sharing
  - Low power, high bandwidth, low latency
  - When compared to the cost of sharing when there is no cache coherence
- Avoid cache thrashing
  - False sharing, ping pong effects
  - Streaming data wiping out data with high locality
    - Smart allocation

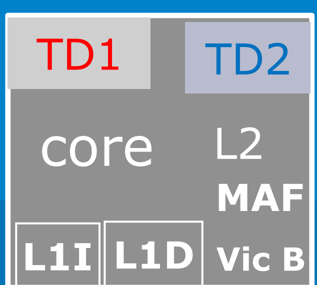
Keeping Track of the data

Full Map – historically default approach. Keeps bit for every cache in the system (with private caches implies a bit for every core in the system). Doesn't scale. Requires extra memory writes to keep up to date.

Linked List of cache entries for same line. Invalidates follow the list. Complex to evict from the middle of the list.

Invalidate Rings – path through all cores/caches that visits each once.

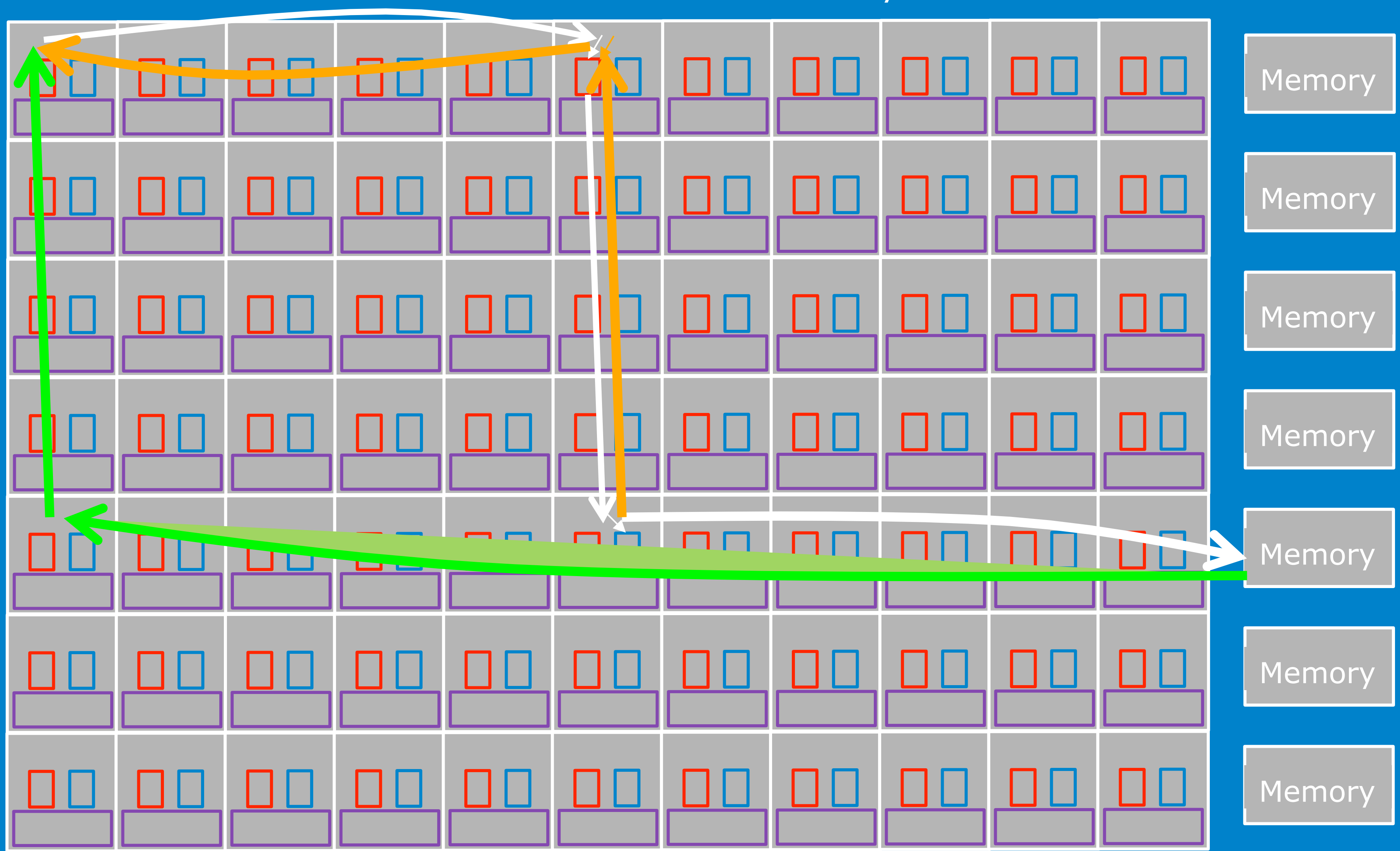
Hierarchical Organization. Use domain bits in second level of tag-directory and then core-valid bits in first level.



Read and service from local neighbor

Two level Directory of Caches in a Mesh

Read Service from other memory



Two level Directory of Caches in a Mesh

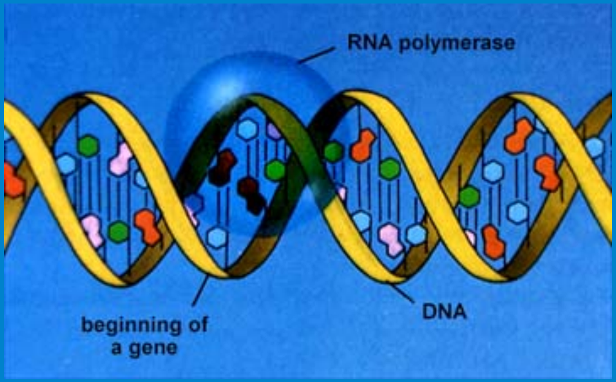
Parallel Bioinformatics Workloads

Structure Learning:

- GeneNet – Hill Climbing, Bayesian network learning
- SNP – Hill Climbing, Bayesian network learning
- SEMPHY – Structural Expectation Maximization algorithm

Optimization:

- PLSA – Dynamic Programming



Recognition:

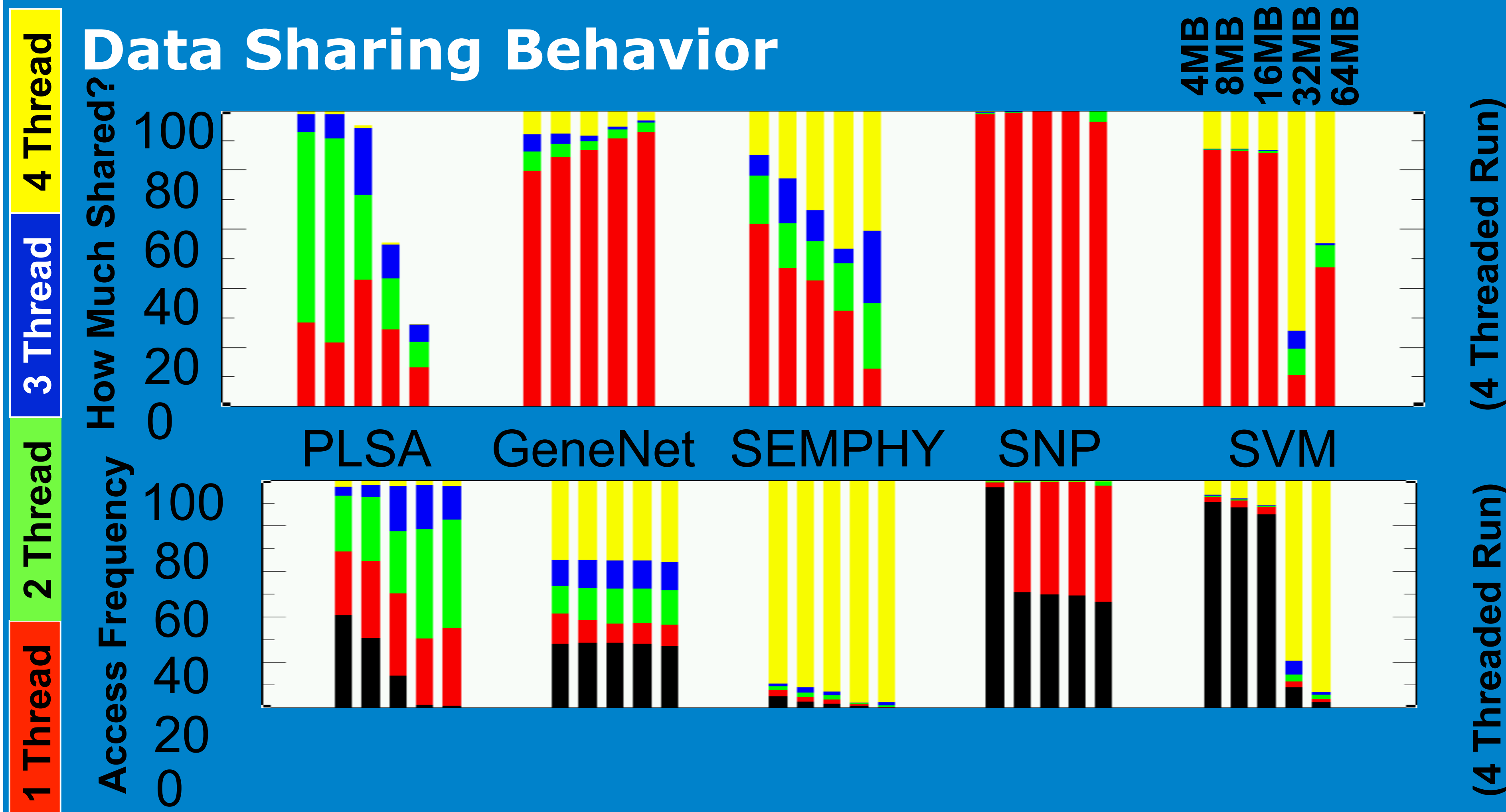
- SVM-RFE – Feature Selection

OpenMP workloads developed by Intel Corporation

- Donated to Northwestern University, NU-MineBench Suite

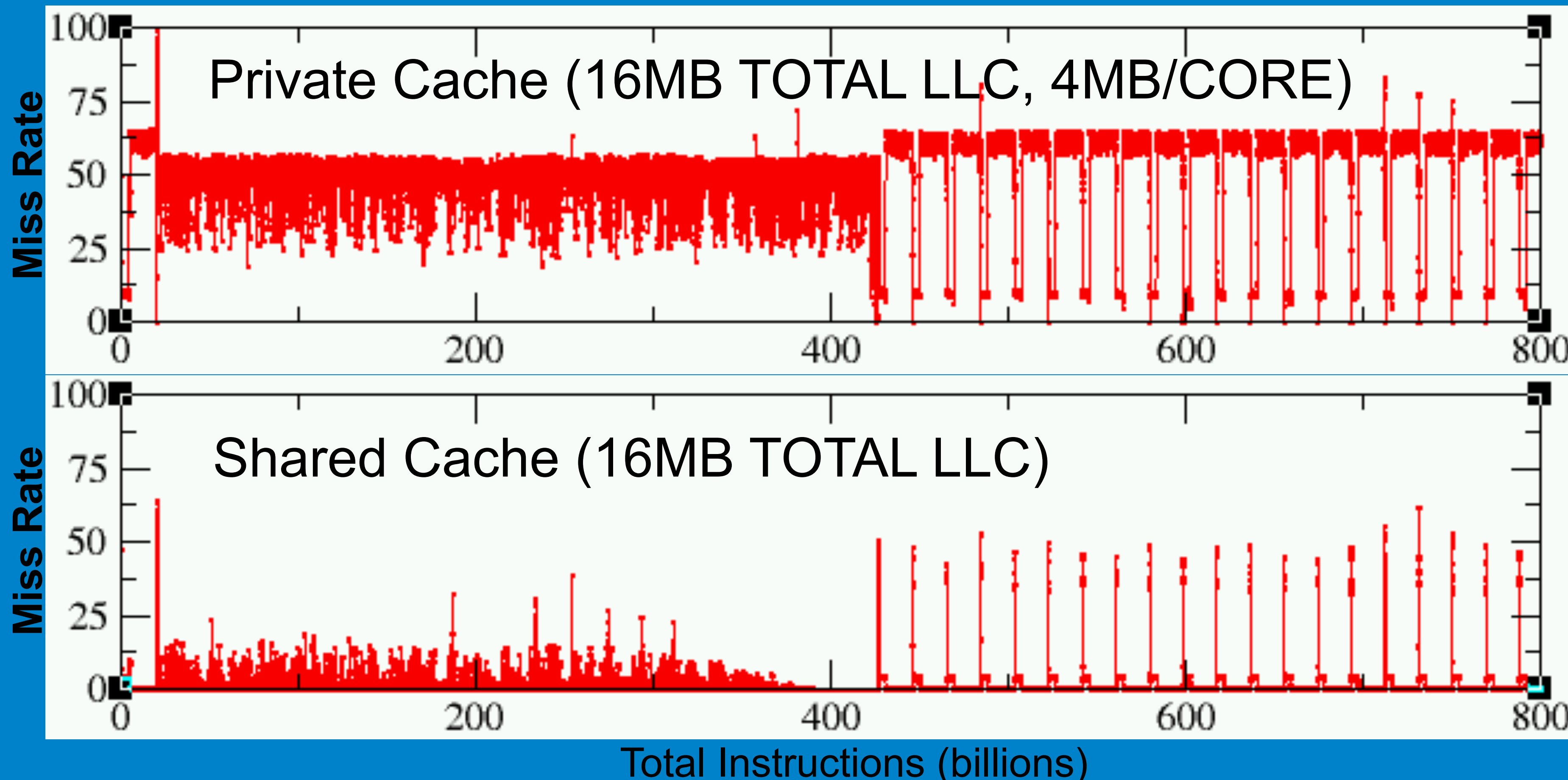
Next four slides from: [Jaleel: HPCA 2006]

Data Sharing Behavior



- Sharing is dependent on algorithm and varies with cache size
- Workloads fully utilize a 64MB LLC
- Reducing cache misses improves data sharing
- Despite size of shared footprint, shared data frequently referenced

Shared/Private Cache – SEMPHY



SEMPHY with 4-threads

Shared cache **out-performs** private caches

Fin

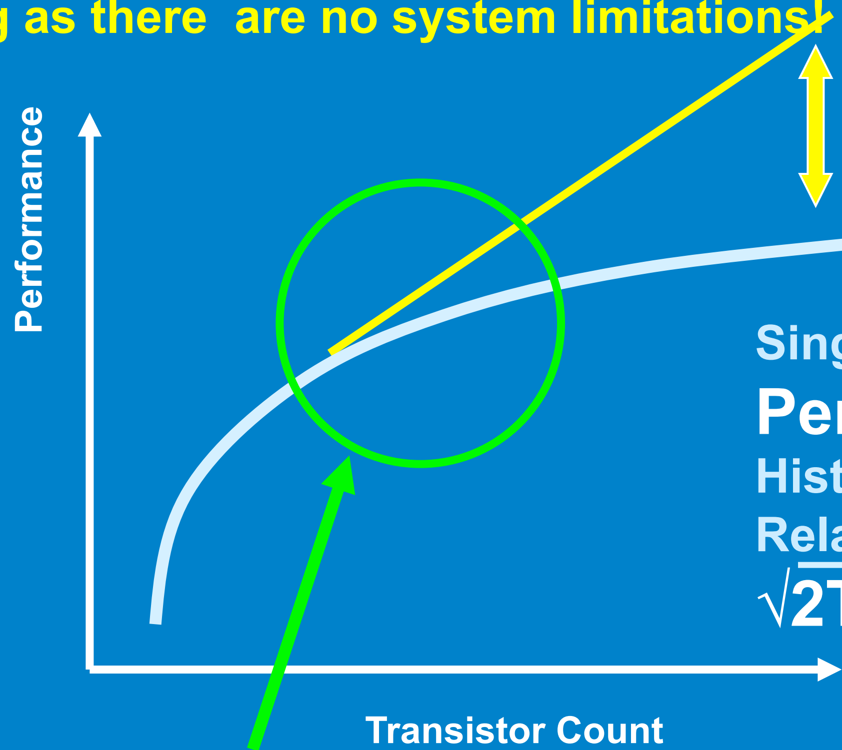
Scaling Example: Alpha EV5

- EV5 was a 4-wide, in order RISC processor
- Scale the process ten times, from 0.5 micron to 16 nm
- EV5 area is 16.5 x 18.1 mm^2 in .5 micron technology, including pin interface and 2^{nd} level cache
  - Moore's law suggests this would reduce by a factor of 2^{10}, to **0.3 mm^2** in **16 nm** technology
- EV5 frequency is 300 MHz
  - Traditional frequency scaling of 1.4 every process generation, leads to a frequency of **9 GHz** in 16 nm technology
- EV5 power is 50W, at 3.3V
  - Assuming a voltage of 0.5V, and changes to C and f cancel out, the dynamic power becomes 50W x $(0.5/3.3)^2$ = **1.1W**
- Core count: 1024 x
- Frequency: 30 x
- Power: > 20 x
- Other growth factors: Interconnects, caches, memory, IO, multi threading, vectors, …

Single Stream, Moore’s Law, and CMP

CMP Performance: Performance ~ x Transistor Count

As long as there are no system limitations!



Gap driving us to Multi-core

Single Stream Performance:

$$\text{Perf} \approx \sqrt{\text{Transistor Count}}$$

Historically fairly accurate

Relates Moore’s Law to Performance:

$$\sqrt{2T} * \sqrt{2} = 2\sqrt{T}$$

Interesting Core Design Area:
Slopes are similar

Transistor speedup due to technology shrink factor of 0.7

Example: Power Management with CMP, optimized for OLTP