Towards Dark Silicon in Servers

ARTIST Summer School 2011, Aix-les-Bains

Babak Falsafi
Director, EcoCloud
ecocloud.ch

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

IT is ever more **indispensable**

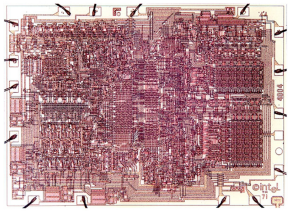
Our life w/o digital data is unimaginable as

- Enterprises
- Scientists
- Governments
- Societies
- Individuals

“He saw your laptop and wants to know if he can check his Hotmail.”

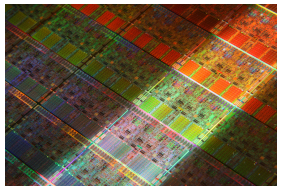
IT: An Exponential Growth

Intel 4004, 1971

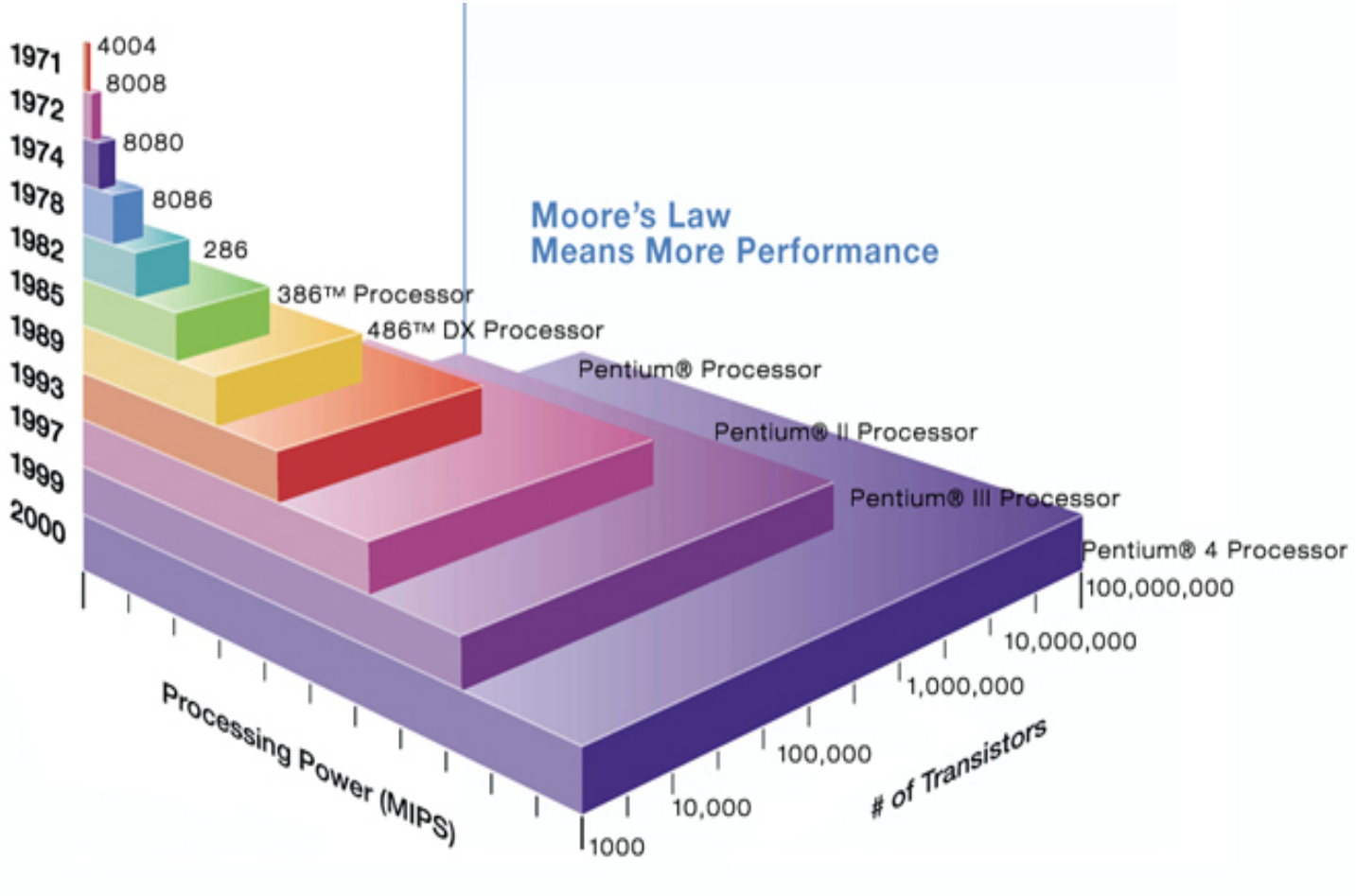


92,000 ops/second

Intel Nehalem, 2009



12,000,000,000 ops/second



Four decades of digital platform proliferation

Exponential increase in density & decrease in cost

A Brief History of IT

Communication Era

Consumer Era

1970s- | 1980s | 1990s | Today+

Mainframes

PC Era

- From scientific instrument to commodity
- From product to service

IT: The Consumer Era

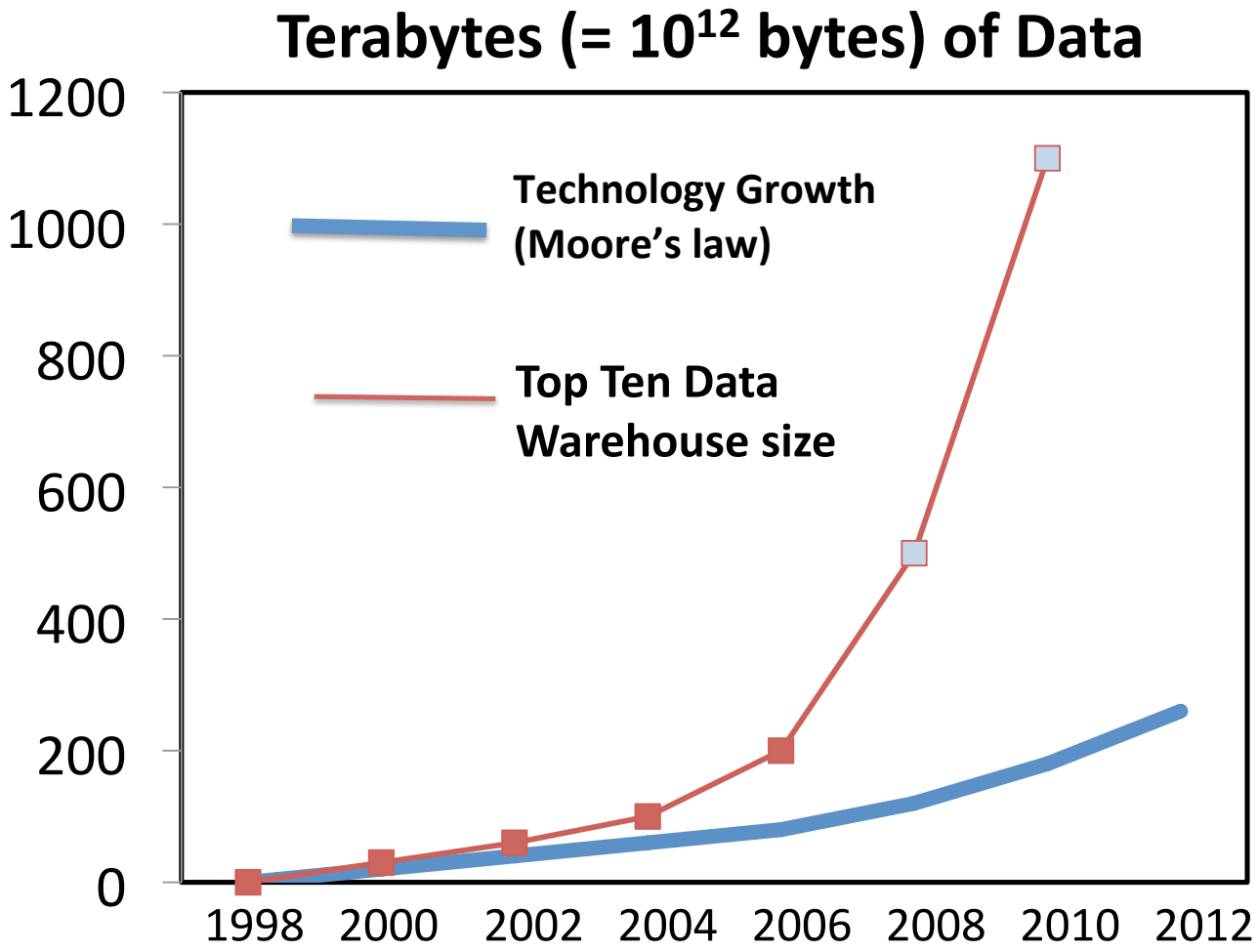
Phenomenal change from decades ago:

- Instant connectivity
- Shopping now online
- Daily interaction > 300 people
- Augmented reality
- Streaming movies
- …..

IT is at core of everyone's life!

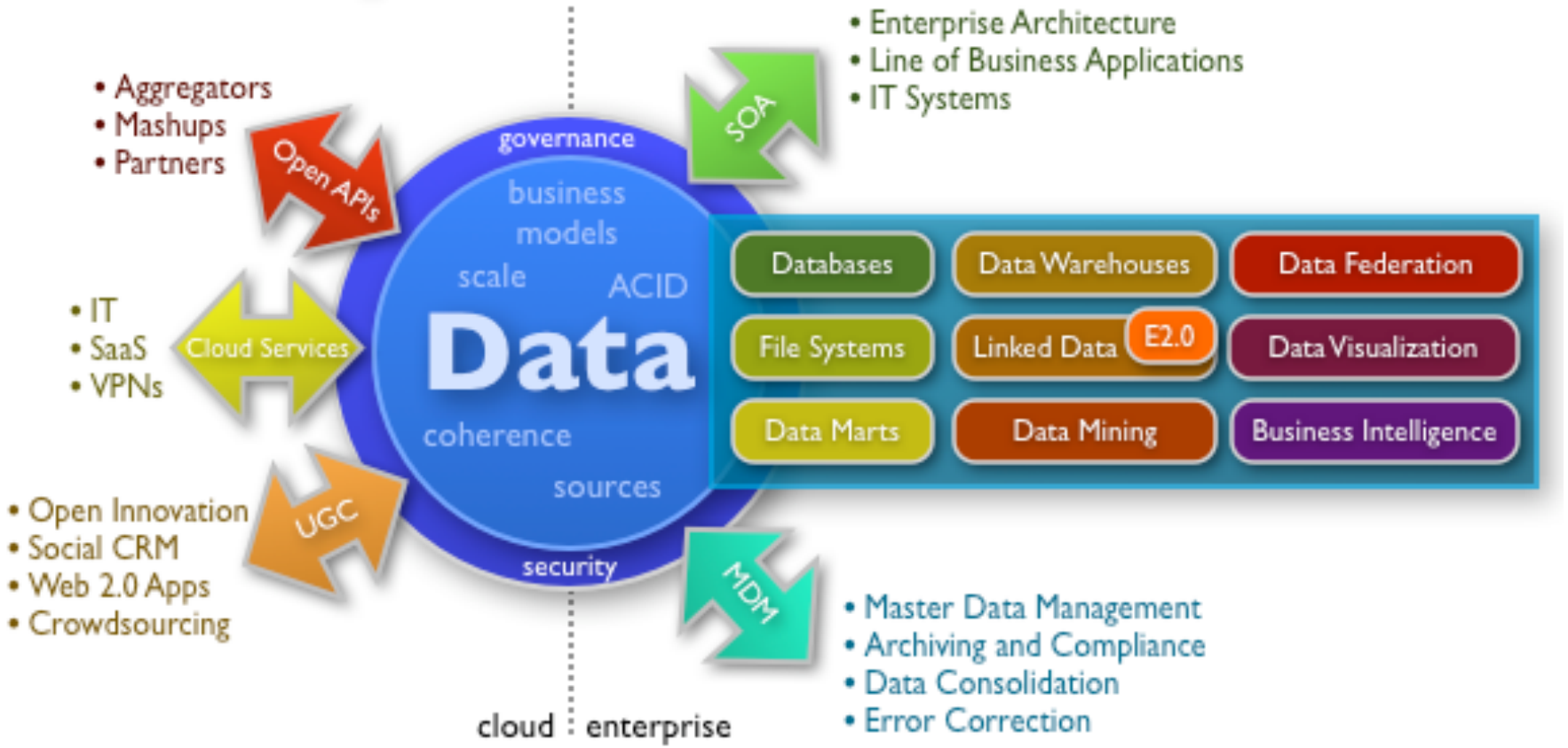
Emergence of Data-Centric IT

- Commerce entirely data-driven
- Science handling massive data
- Companies spending $$$ to collect/analyze data
- Personalized computing



WinterCorp Survey, www.wintercorp.com

Anatomy of a Data-Centric Business

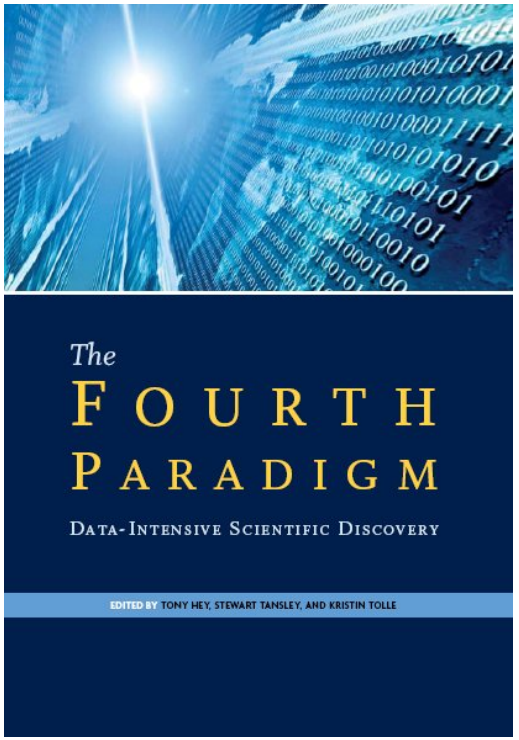


- Era of “knowledge economy”
- 50% of economic value in developed countries
- Dominant supply-chain component of products/services

Data-Centric Science: "The Fourth Paradigm"

Mining data from:

- Archives
- Humans
- Sensors/instruments
- Simulations



Unifying theory, experimentation, simulation, analytics on massive data

It's all about Accessing Data!

Data Centers

Cloud Computing

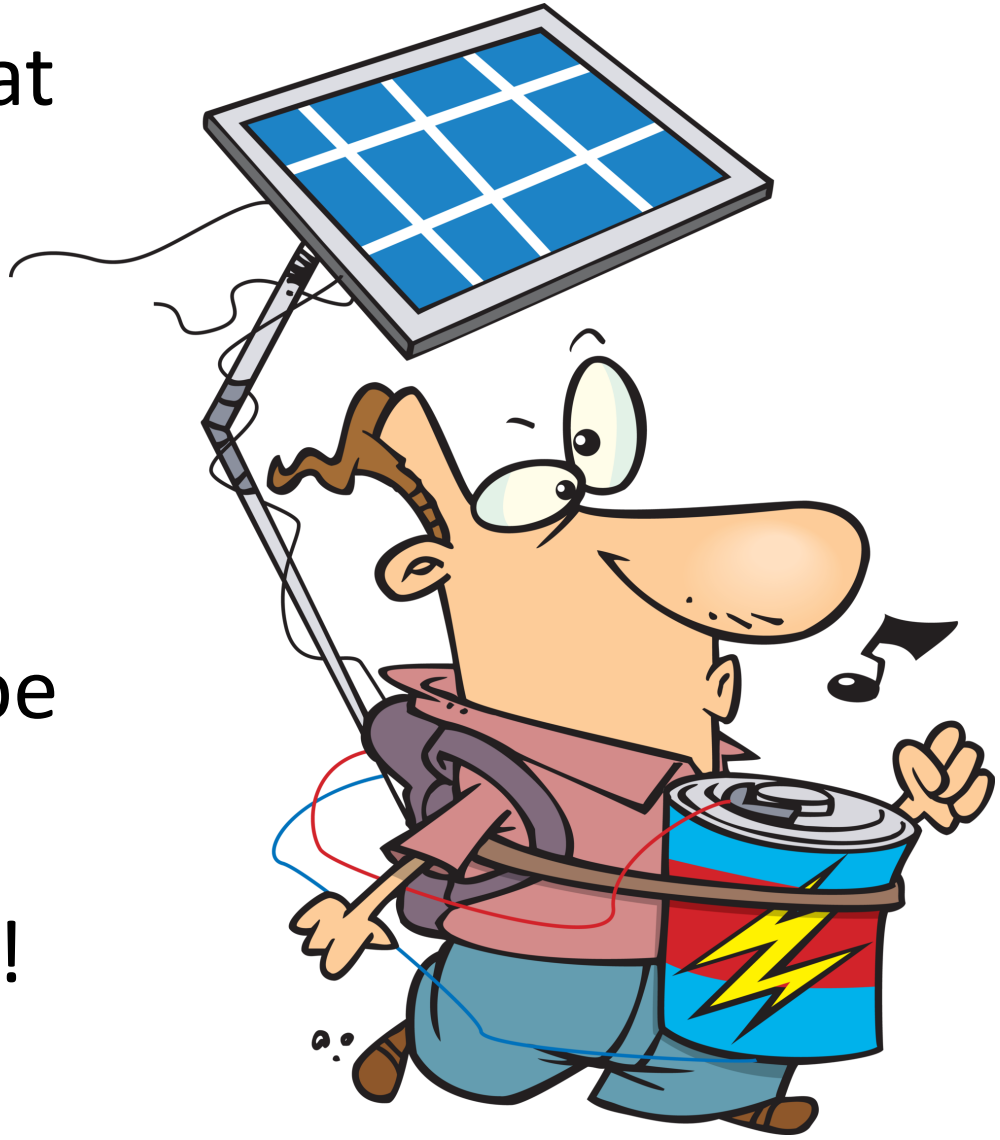
A computing paradigm shift to enable ubiquitous connectivity

But, IT Energy is Shooting Up!

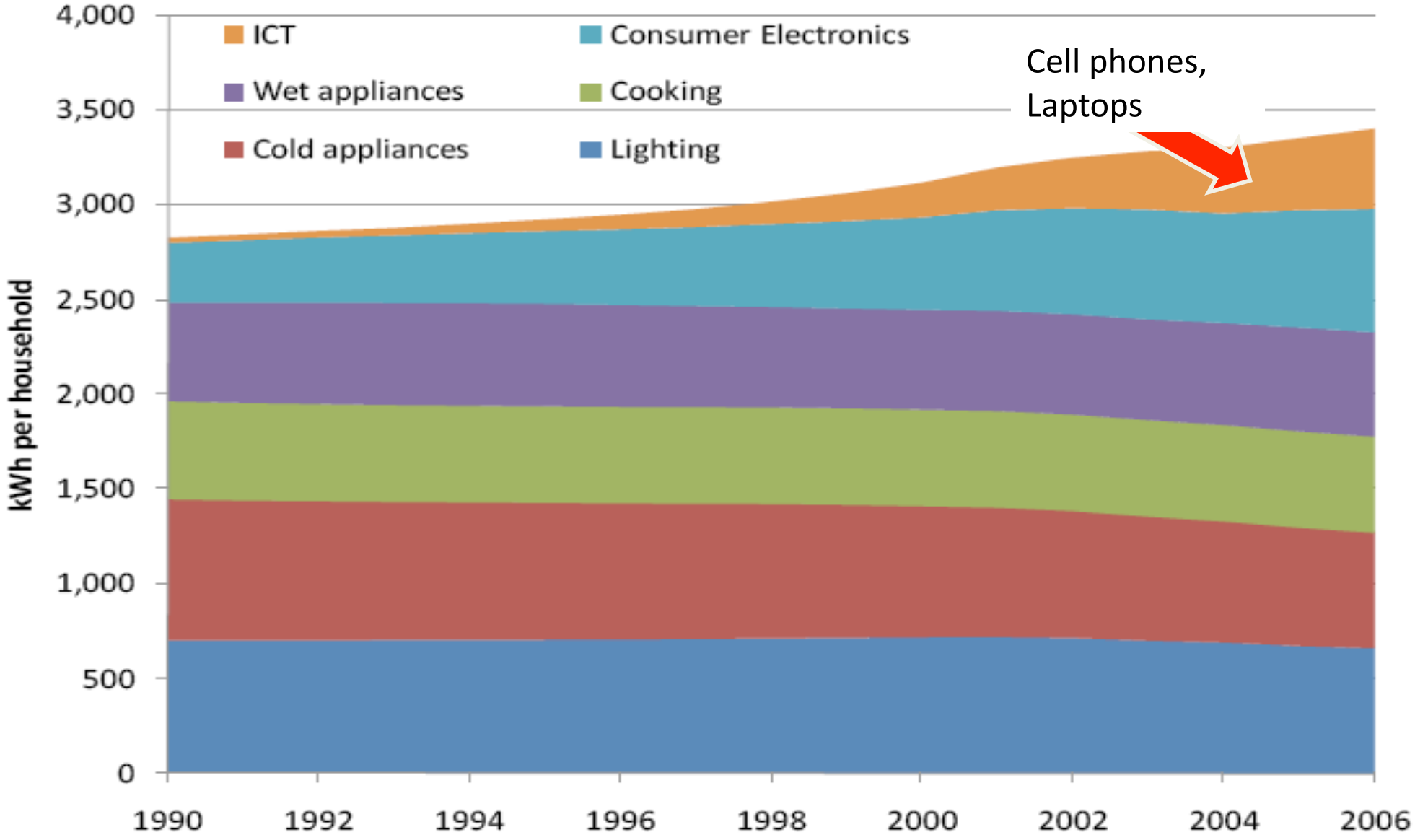
IT riding on technology that was energy-friendly

- Exponentially better performance, density
- Constant power envelope

But, energy is shooting up!



Household Energy in the UK
(UK BERR, 2008)



Household Energy in the US
(NY Times, 2011)

Comparing Energy Use

Comparison of a typical television set-top box configuration with Energy Star-rated appliances and devices.

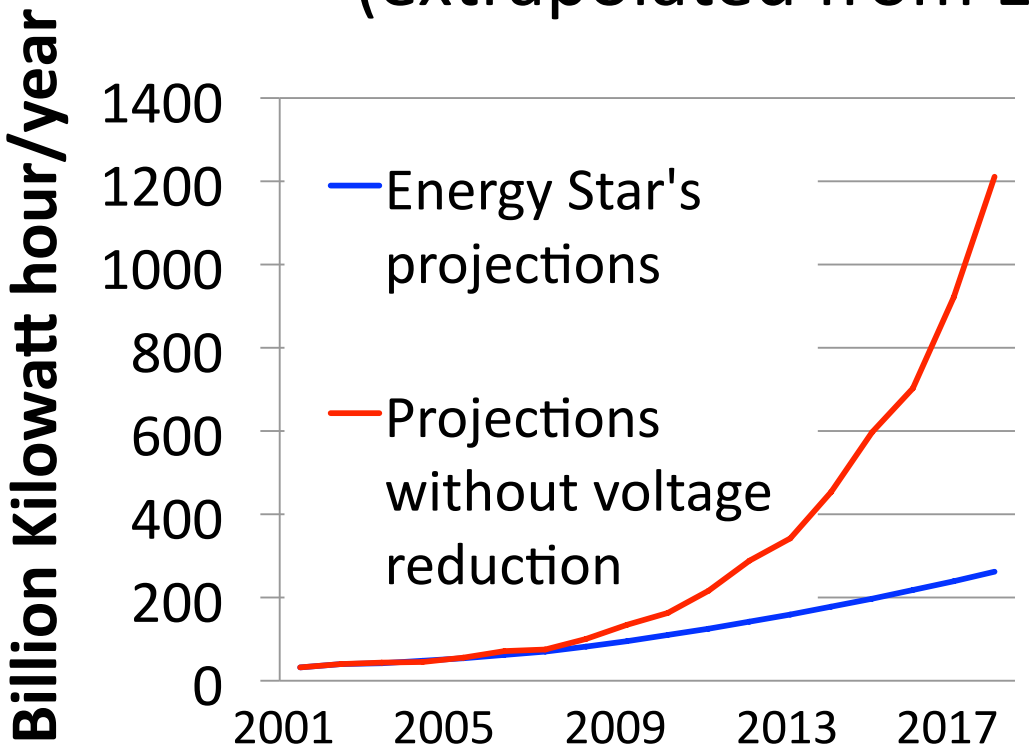
	AVERAGE KILOWATT-HOURS A YEAR	HD SET-TOP BOX	HD DVR	TIME IN USE EACH DAY
Typical HD television set-top box configuration	446	171	275	24 hours
Refrigerator (21-cubic-foot)	415			24 hours
LCD television (42-inch)	181			5 hours
Desktop computer	175			8 hours
Compact fluorescent light bulb (15-watt)	17			3 hours

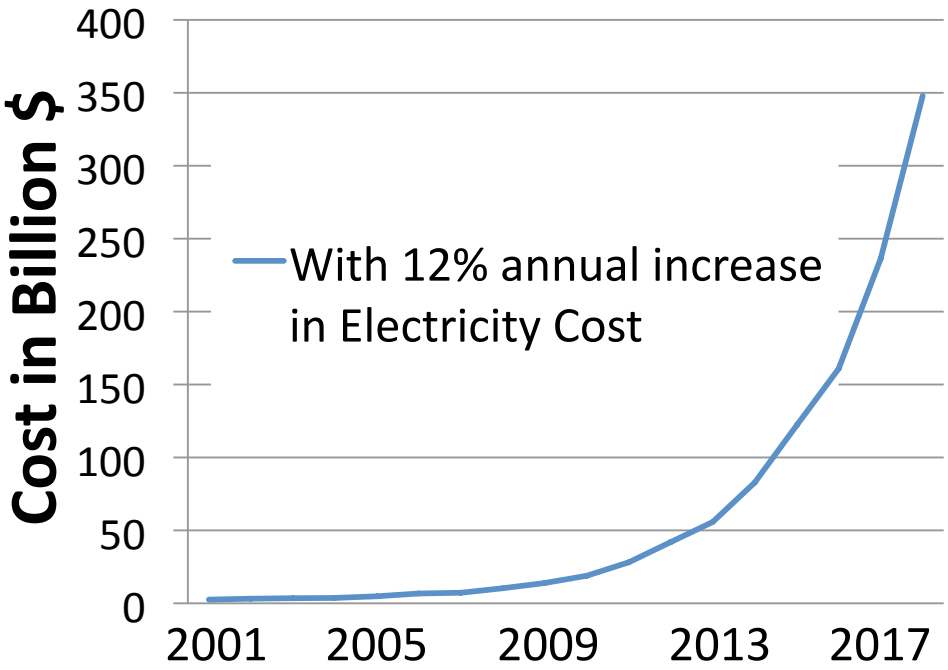
Source: Natural Resources Defense Council

THE NEW YORK TIMES

Data center Energy in the US

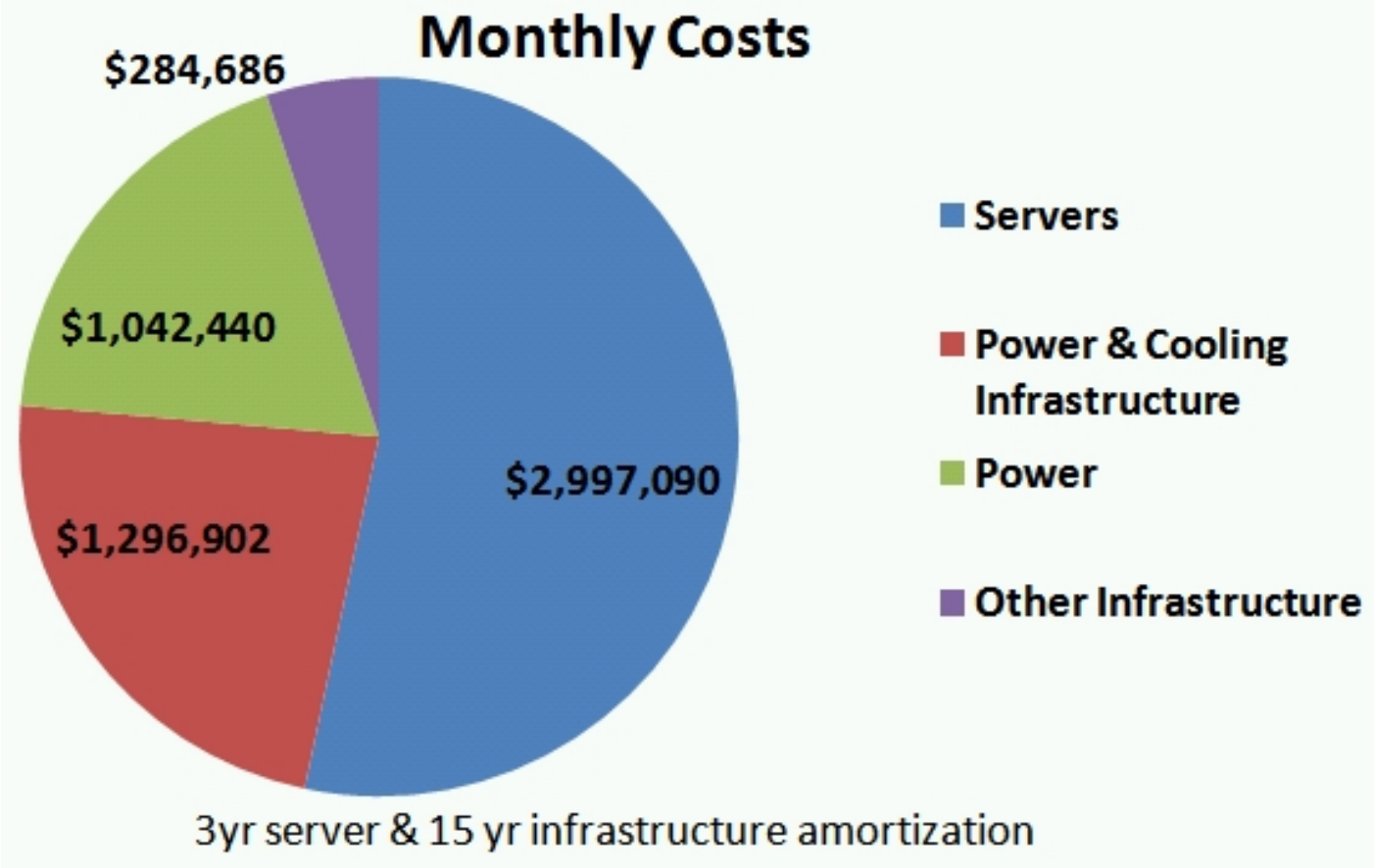
(extrapolated from Energy Star, 2007)





- Exponential costs if not mitigated
- Today, carbon footprint of airline industry

Energy > Capital Cost



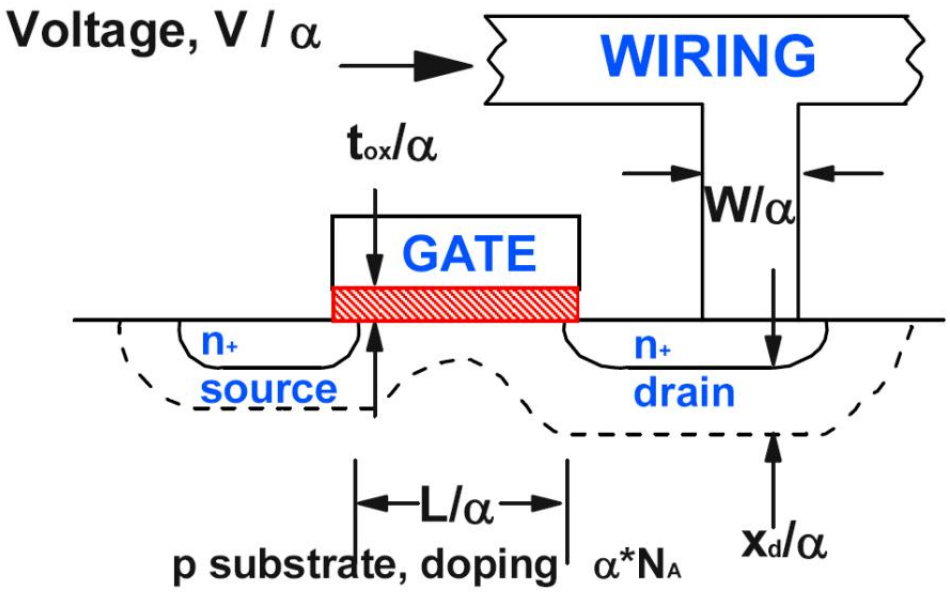
James Hamilton's Blog, mvdirona.com, 2008

- Servers are getting relatively cheaper
- Power is beginning to dominate cost

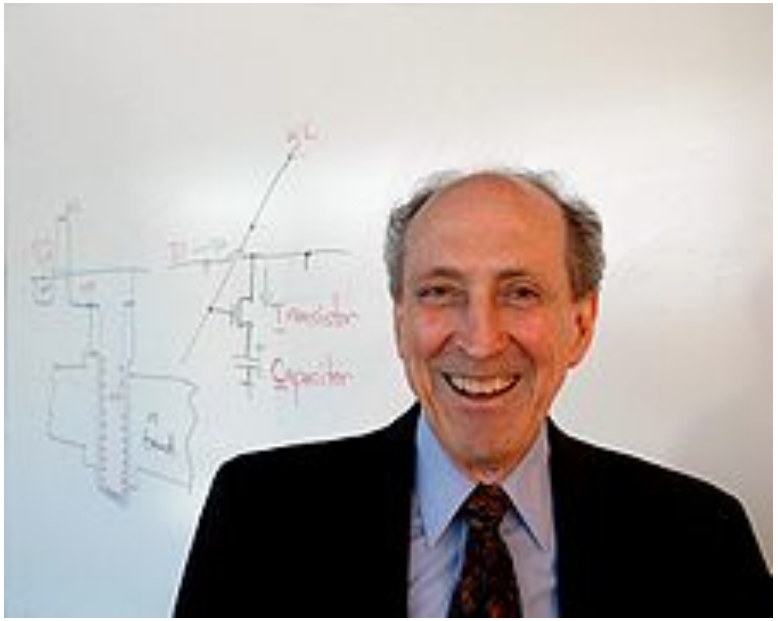
Course Roadmap

- What is up with power (energy)?
  - End of Dennard scaling
  - What does it mean for servers?
- How do we tame power (energy)?
  - What do servers want?
  - What will server chips look like?

Four decades of Dennard Scaling



Dennard et. al., 1974



Robert H. Dennard, picture from Wikipedia

- **$P = C V^2 f$**
- Increase in device count
- Lower supply voltages

➜ Constant power/chip

Leakage Killed Dennard Scaling

Leakage:

- Exponential in inverse of V_{th}
- Exponential in temperature
- Linear in device count

To switch well

- must keep $V_{dd}/V_{th} > 3$

➜V_{dd} can't go down

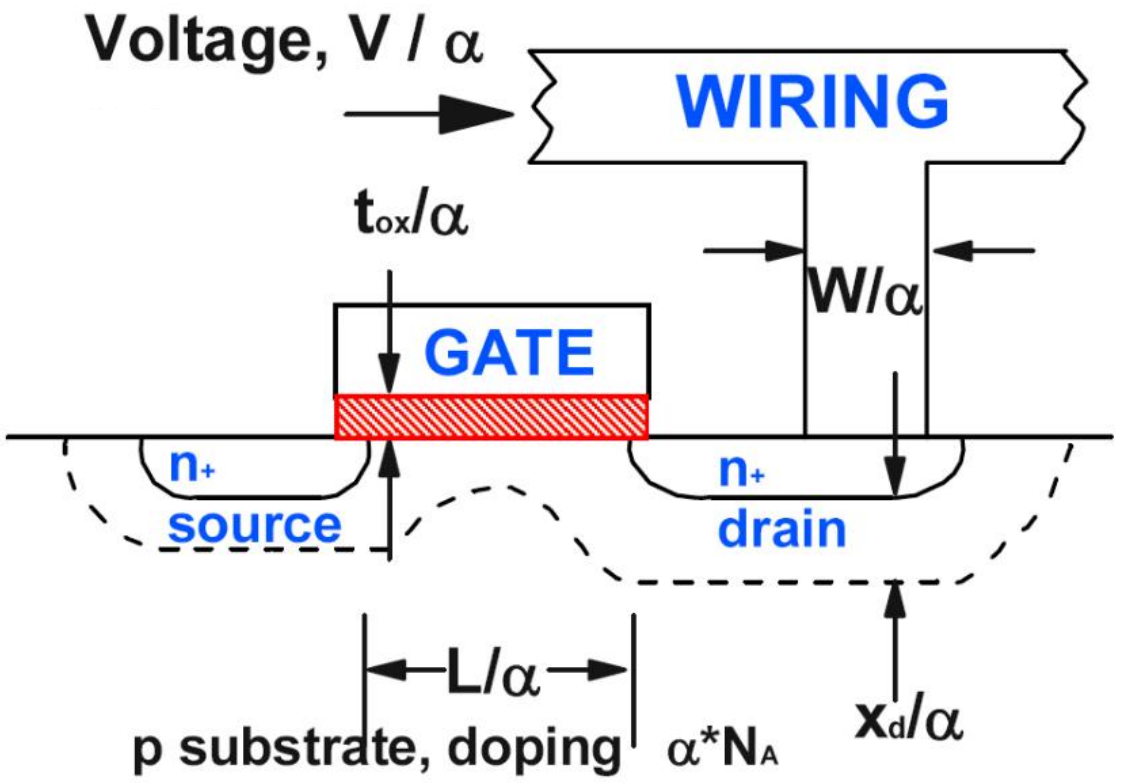
Post-classic CMOS Dennard Scaling

Scaling:

Voltage:	~~V/α~~	V
Oxide:	t_{ox}/α	
Wire width:	W/α	
Gate width:	L/α	
Diffusion:	x_d/α	
Substrate:	αN_A	

Results:

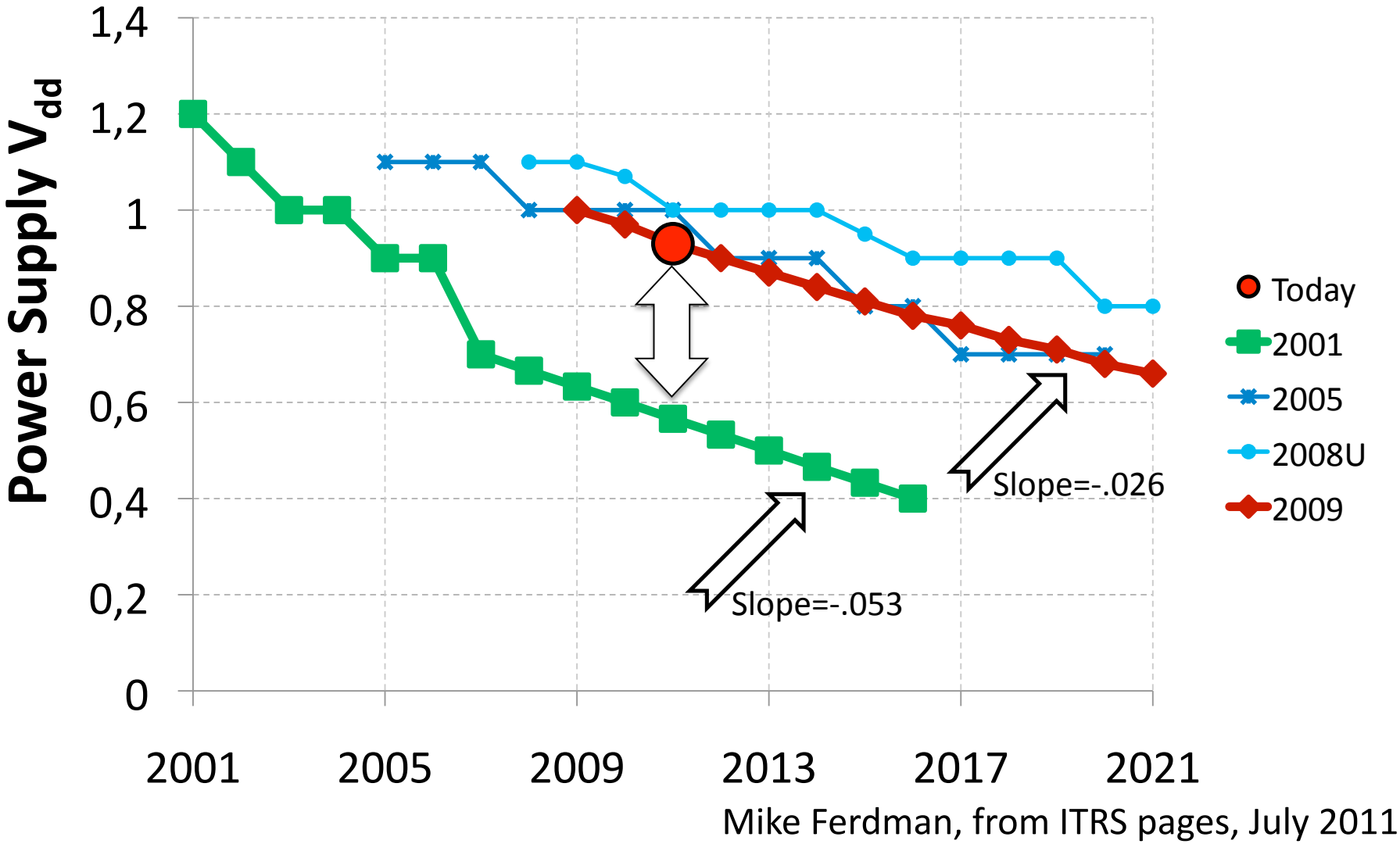
Higher Density:	$\sim\alpha^2$	
Higher Speed:	$\sim\alpha$	
Power/ckt:	~~$1/\alpha^2$~~	1
Power Density:	~~~Constant~~	α^2



Important: ***Diminishing area powered up***

R. H. Dennard *et al.*,
IEEE J. Solid State Circuits, (1974).

End of Dennard Scaling (ITRS)

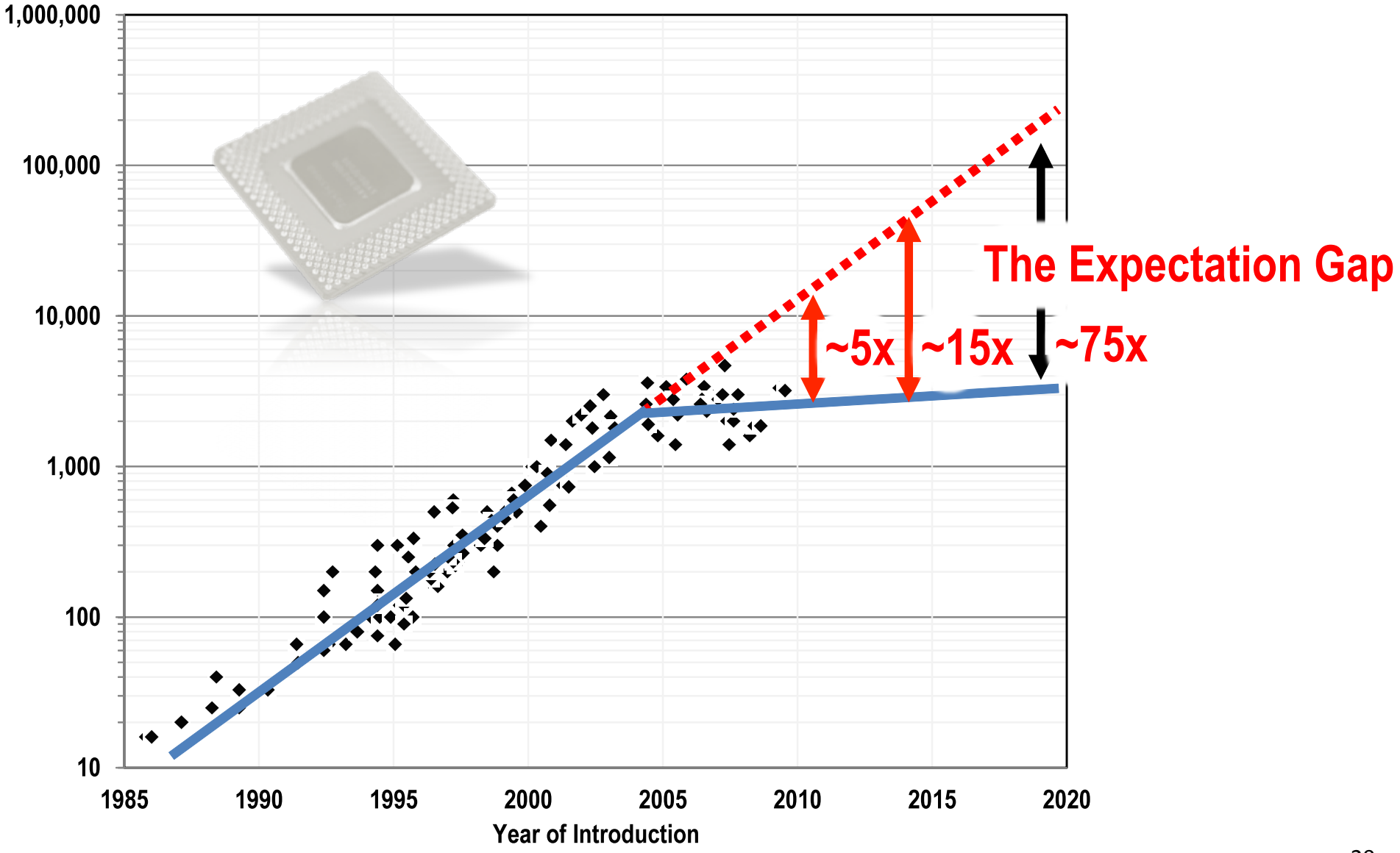


Mike Ferdman, from ITRS pages, July 2011

Supply voltages going down at much lower rate!

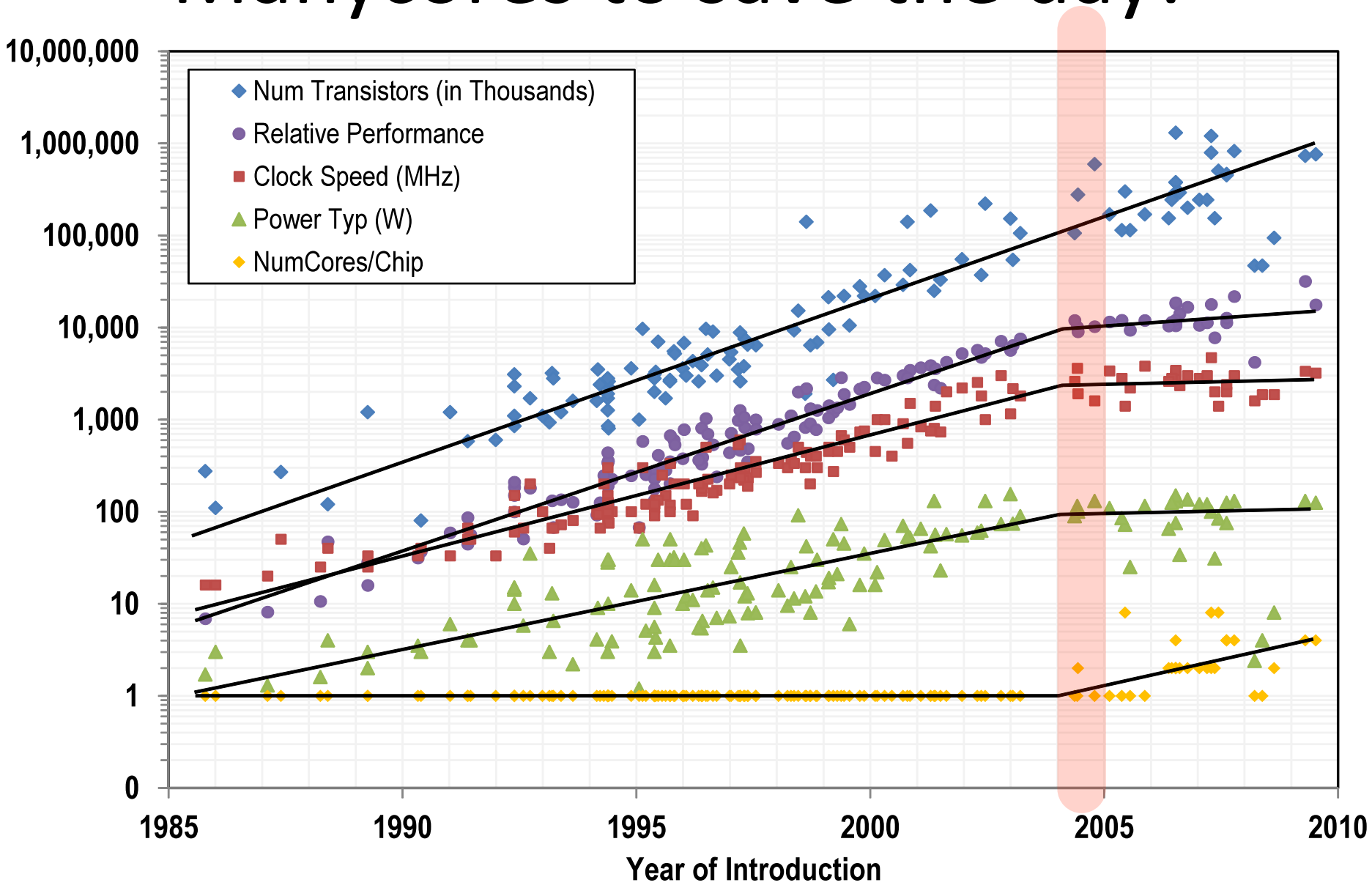
End of Single Core Performance in 2004!

Microprocessor Performance "Expectation Gap" over Time (1985-2020 projected)



National Research Council (NRC) – Computer Science and Telecommunications Board (CSTB.org)

Manycores to save the day!



100 Billion Transistor Chips

Voltages stop:

- More cores
- Simpler cores (increase efficiency)
- Tilera has announced 100-core chips!

But, more cores need

- Parallel software
- Memory bandwidth
- Power (how much simpler?)

A Study of Server Chip Scalability

[Hardavellas, IEEE Micro on Big Chips, 2011]

Actual server workloads today

- Easily parallelizable (performance-scalable)

Actual physical char. of processors/memory

ITRS projections for technology nodes

Modeled power/performance across nodes

For server chips

- Bandwidth is near-term limiter

→ **Power is the ultimate limiter**

A few words about our model

Physical char. modeled after Niagara

Area: cores/caches (72% die)

- scaled across tech. nodes

Power:

- Active: projected V_{dd}/ITRS
  - Core=scaled, cache=f(miss), crossbar=f(hops)
- Leakage: projected V_{th}/ITRS, f(area), 62C

Performance:

- Parameters from real server workloads (DB2, Oracle, Apache, Zeus)
- Cache miss rate model (validated)
- CPI model based on miss rate

Caveat: Parallel Workloads

Workloads are assumed parallel

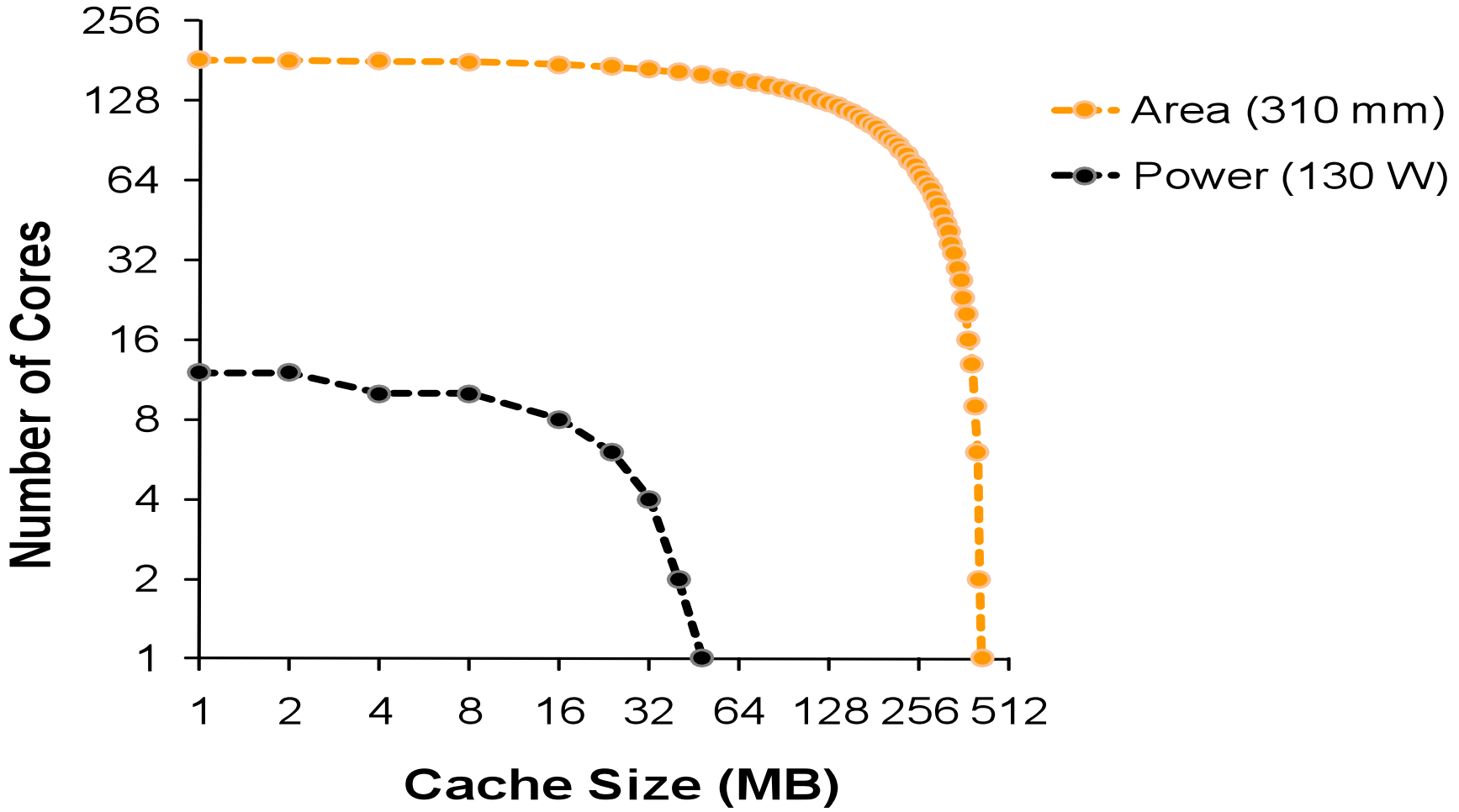
- Scaling server workloads is reasonable

CPI model:

- Works well for workloads with low MLP
- OLTP, Web & DSS are mostly memory-latency dependent

Future servers will run a mix of workloads

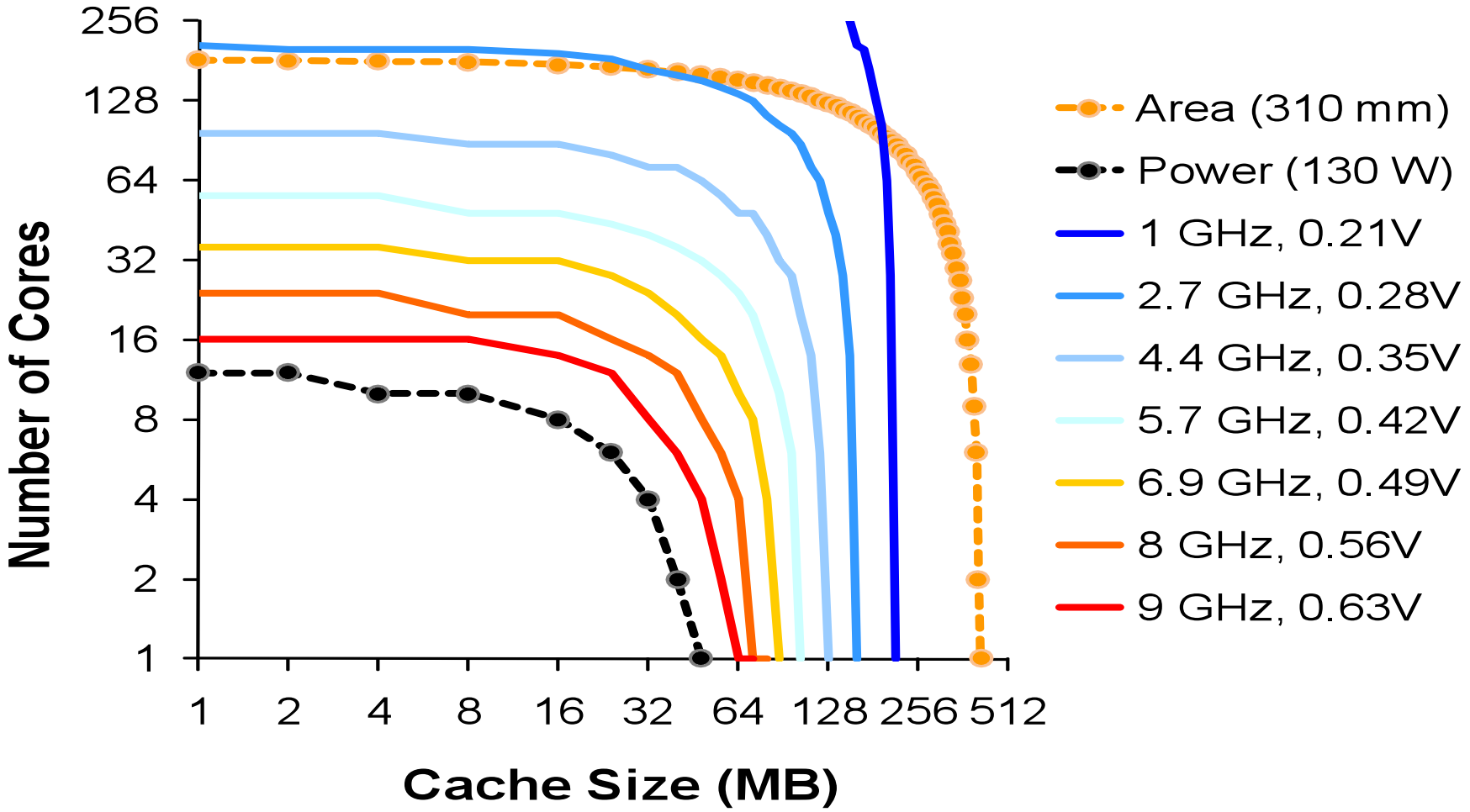
Area vs. Power Envelope (22nm)



✓ Good news: can fit hundreds of cores

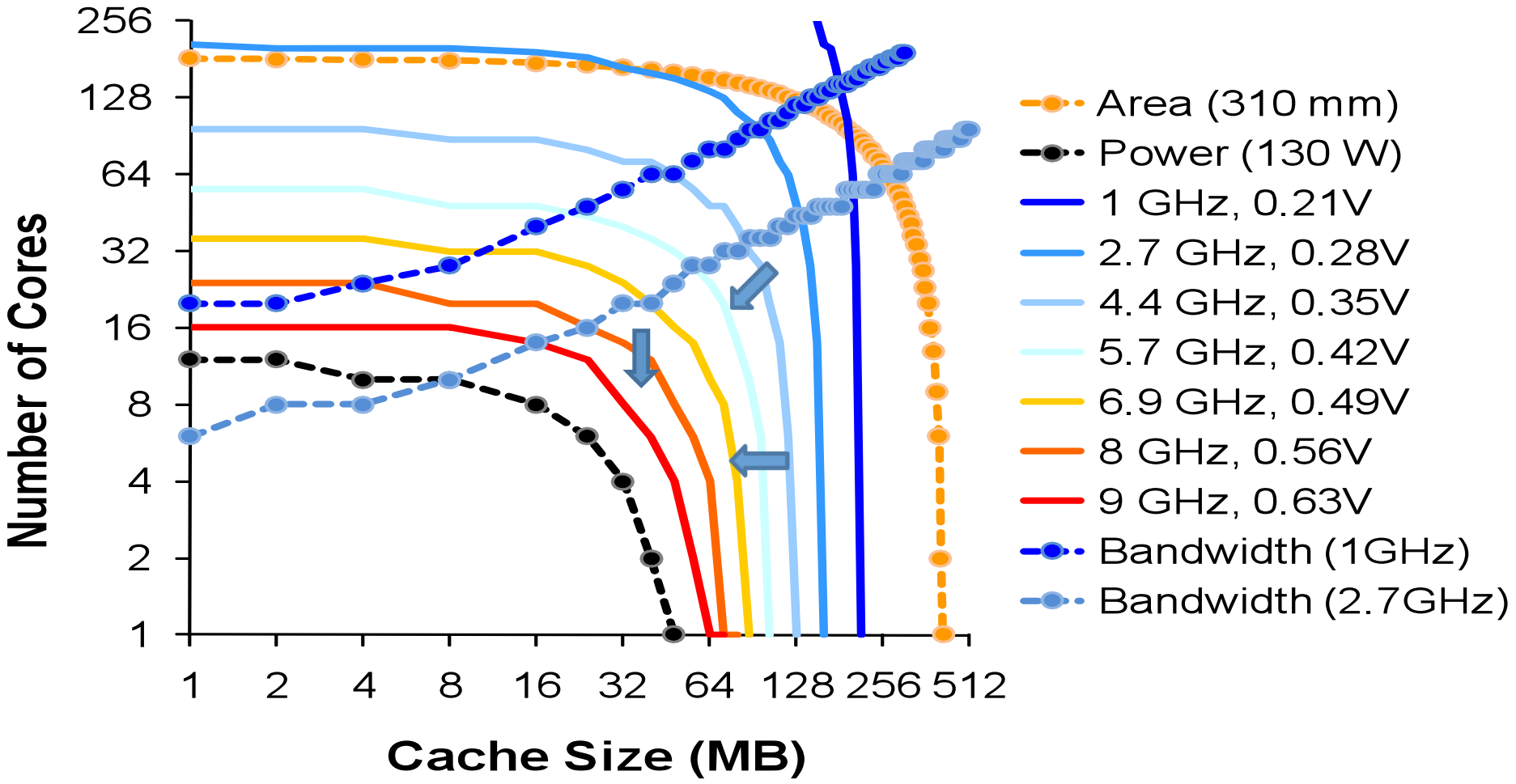
× Can not use them all at highest speed

Pack Slower Cores, Cheaper Cache



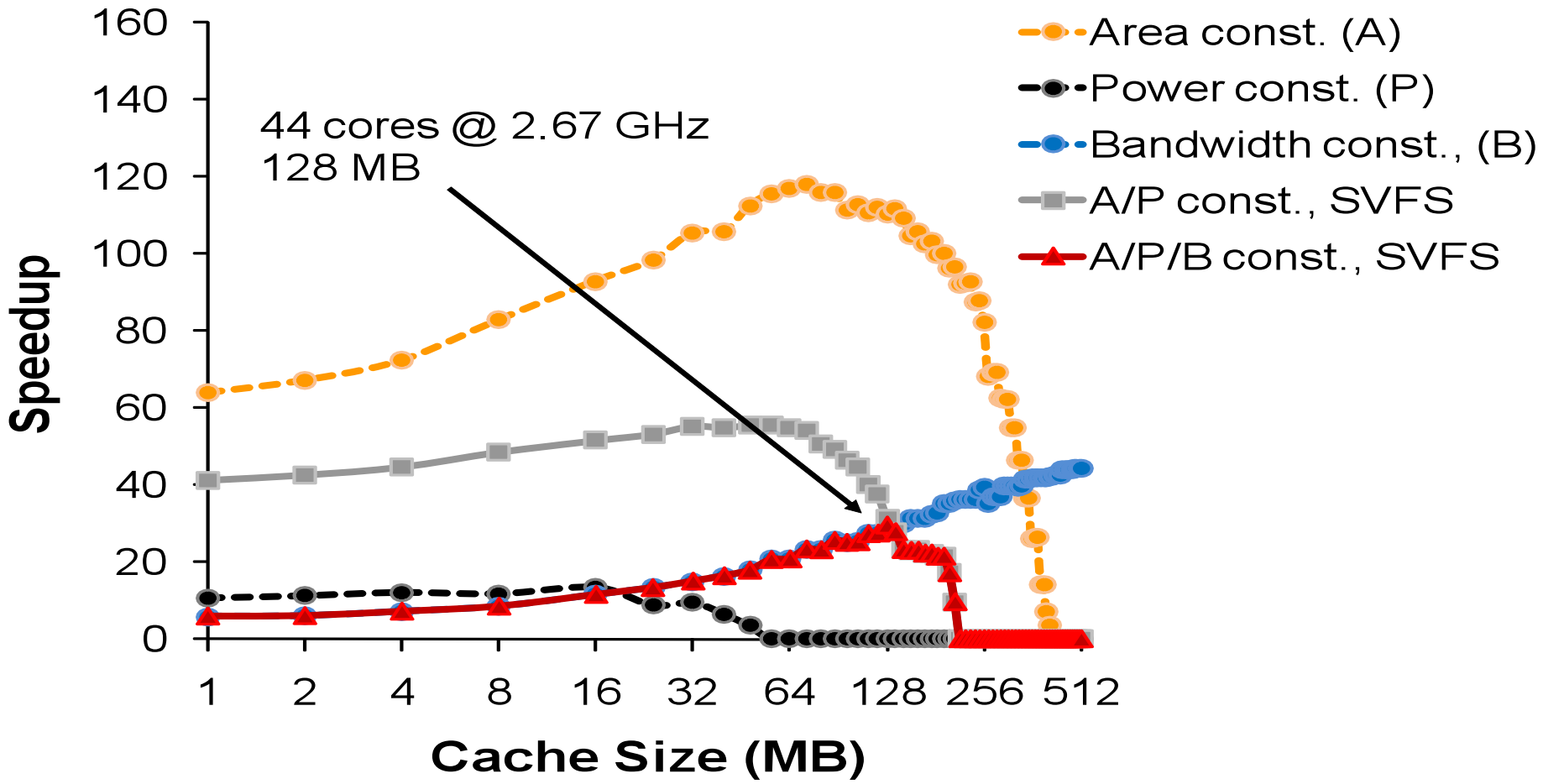
- Result: a performance/power trade-off
- Assuming bandwidth is unlimited

Limited b/w: fewer cores + more cache



- For clarity, only showing two bandwidth lines
- Where would the best performance be?

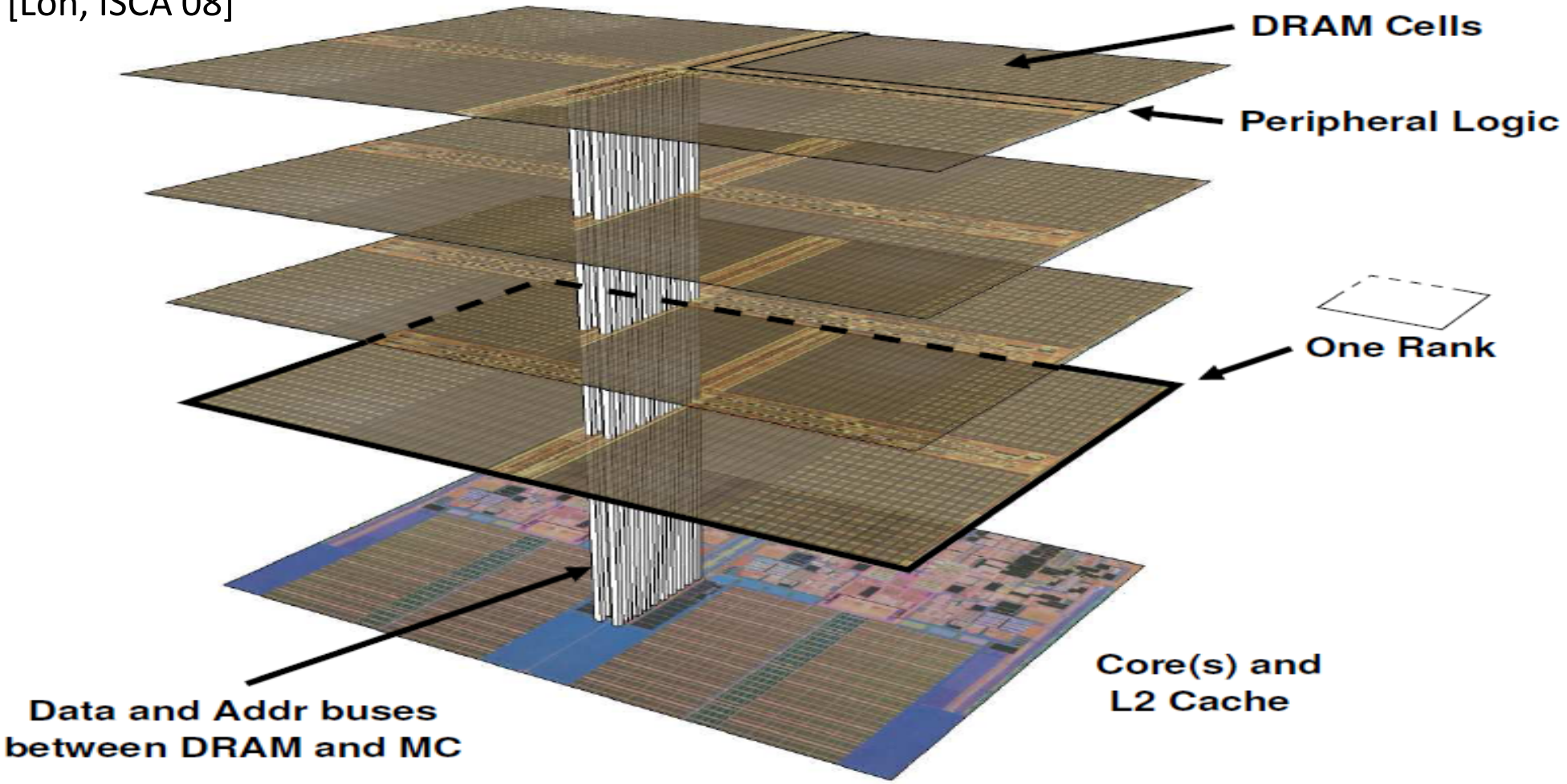
Peak Performance



- B/W constrained, then power constrained
- Fewer slower cores, lots of cache

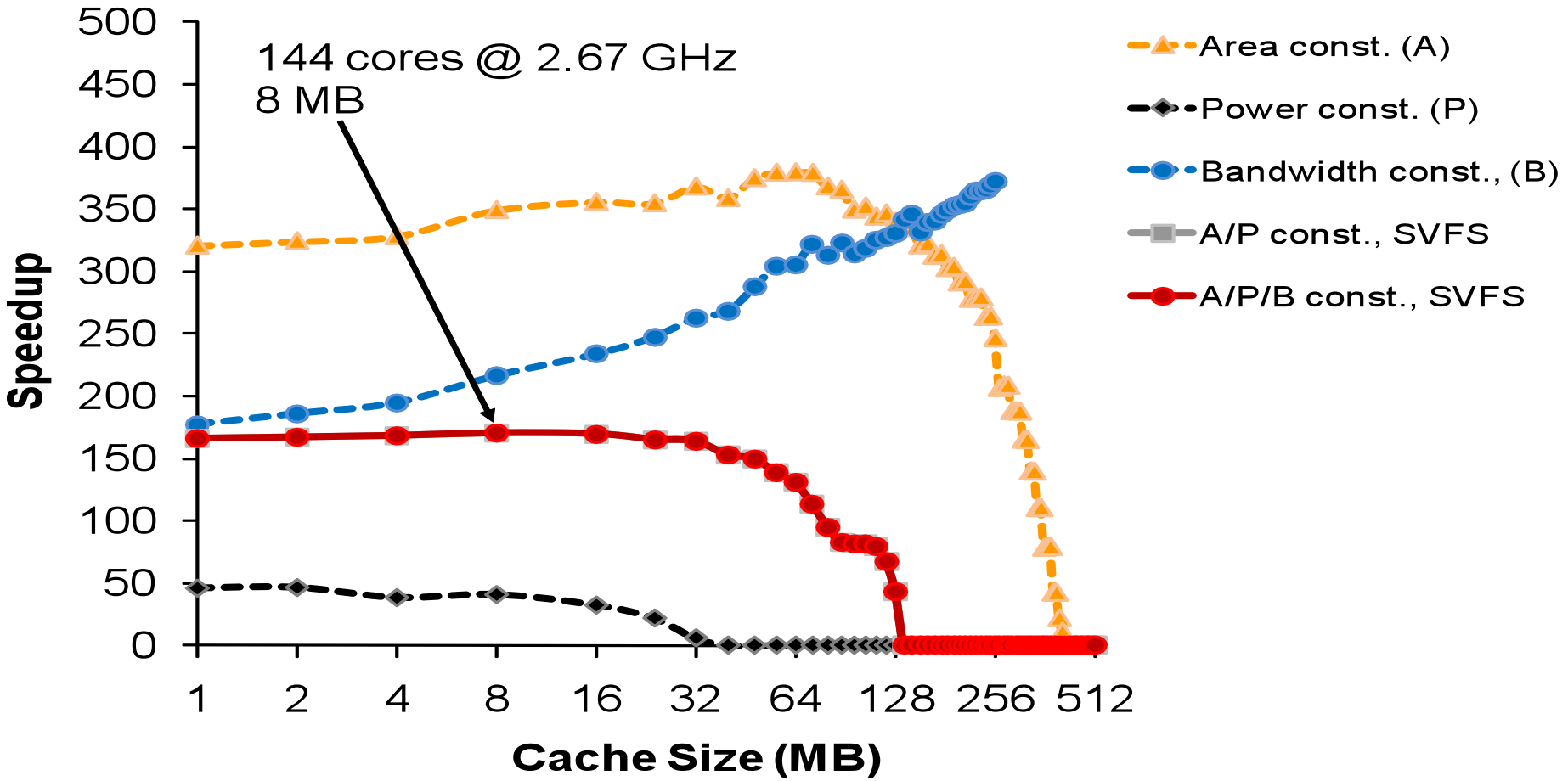
Mitigating B/W with 3D Memory

[Loh, ISCA'08]



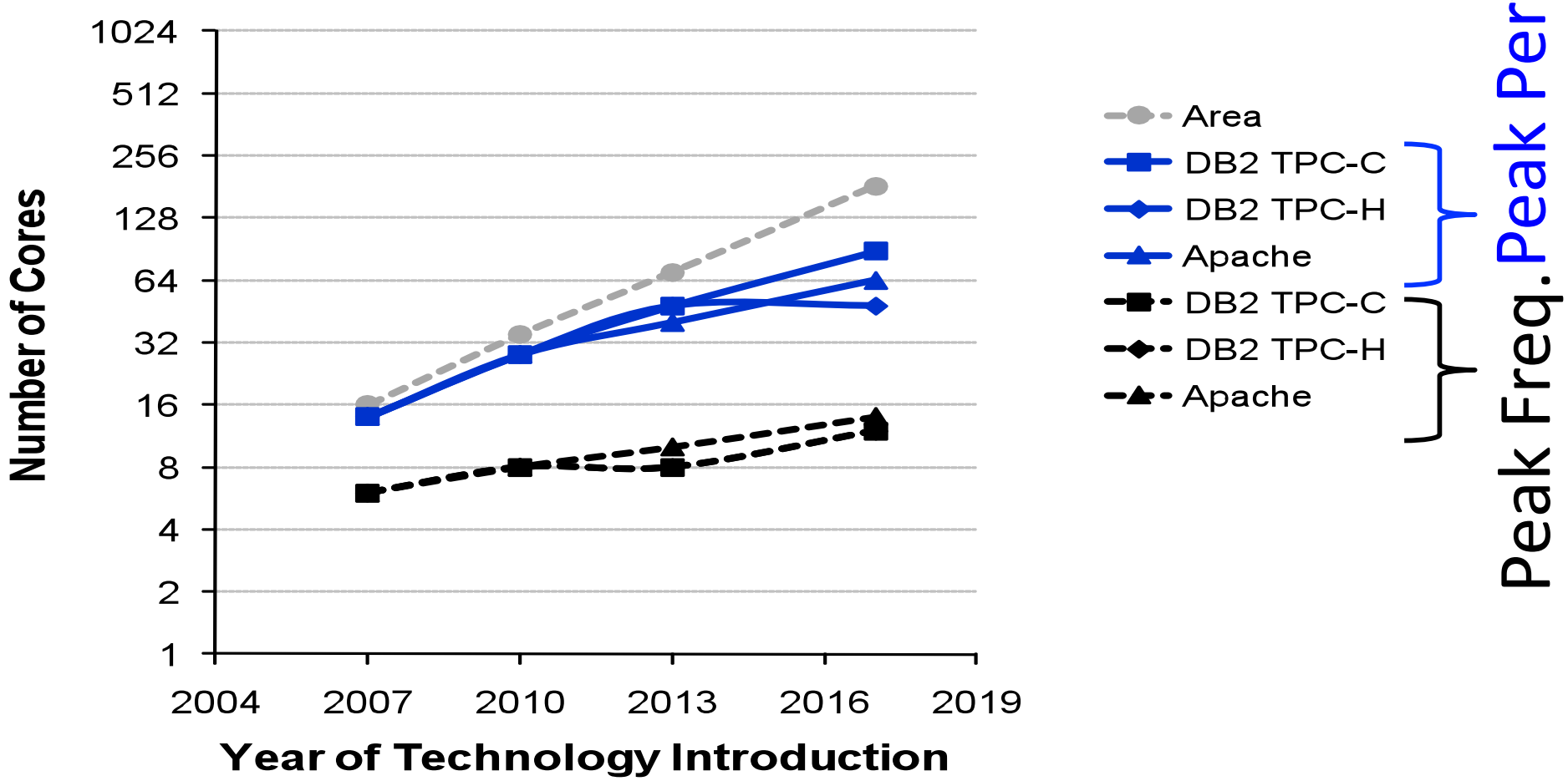
- Delivers TB/sec of bandwidth

Peak Performance with 3D Memory



- Only power-constrained
- Virtually eliminates on-chip cache

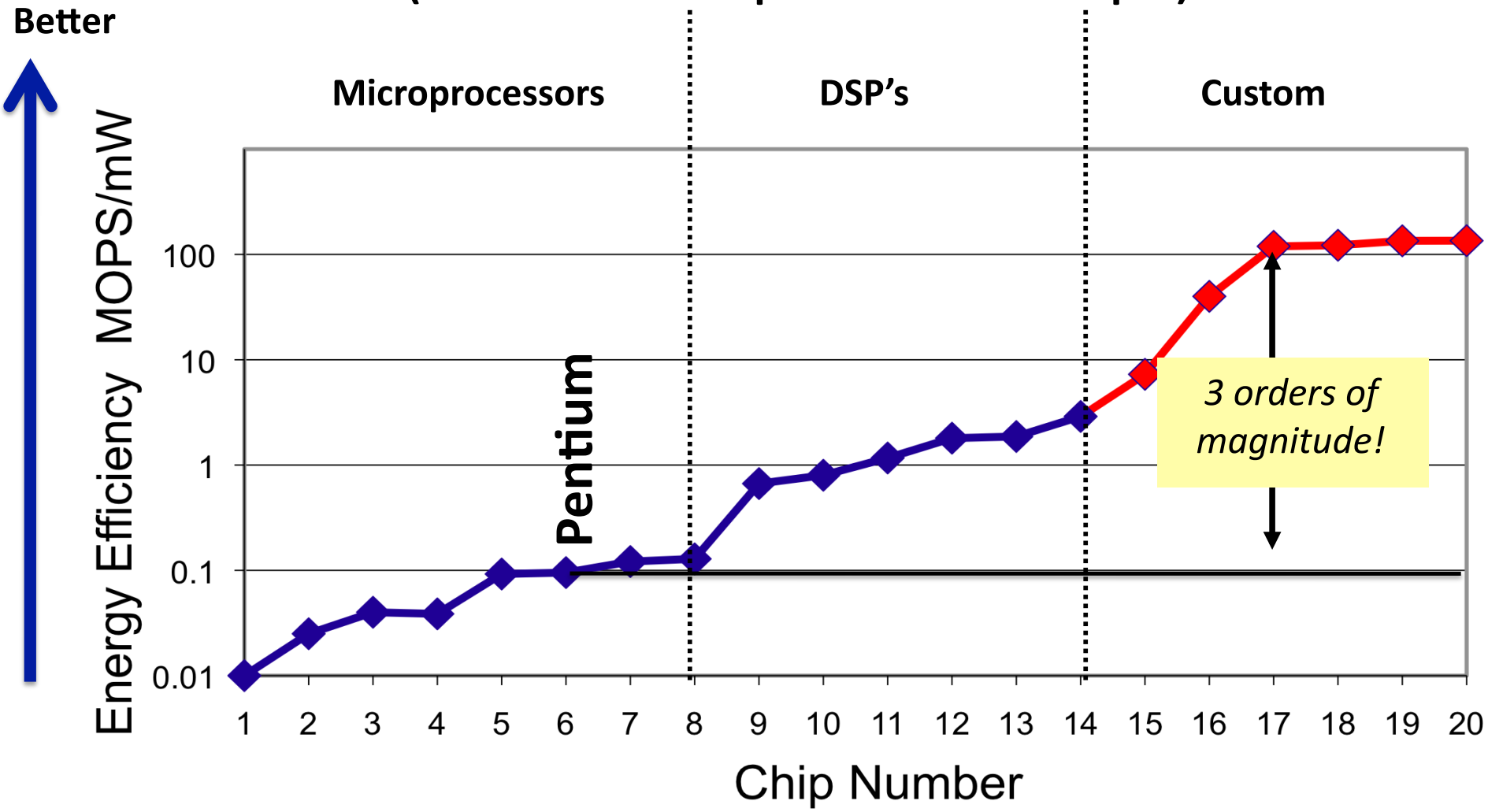
Core Scaling across Technologies



- Assumes a 130-Watt chip envelope
- Pin b/w keeps Niagara from scaling

Specialization can buy 1000x in Energy

(from a sample of 20 chips)



Mihai Budiu, "On the Energy Efficiency of Computation", 2004

Dark Silicon: End of Manycore Scaling

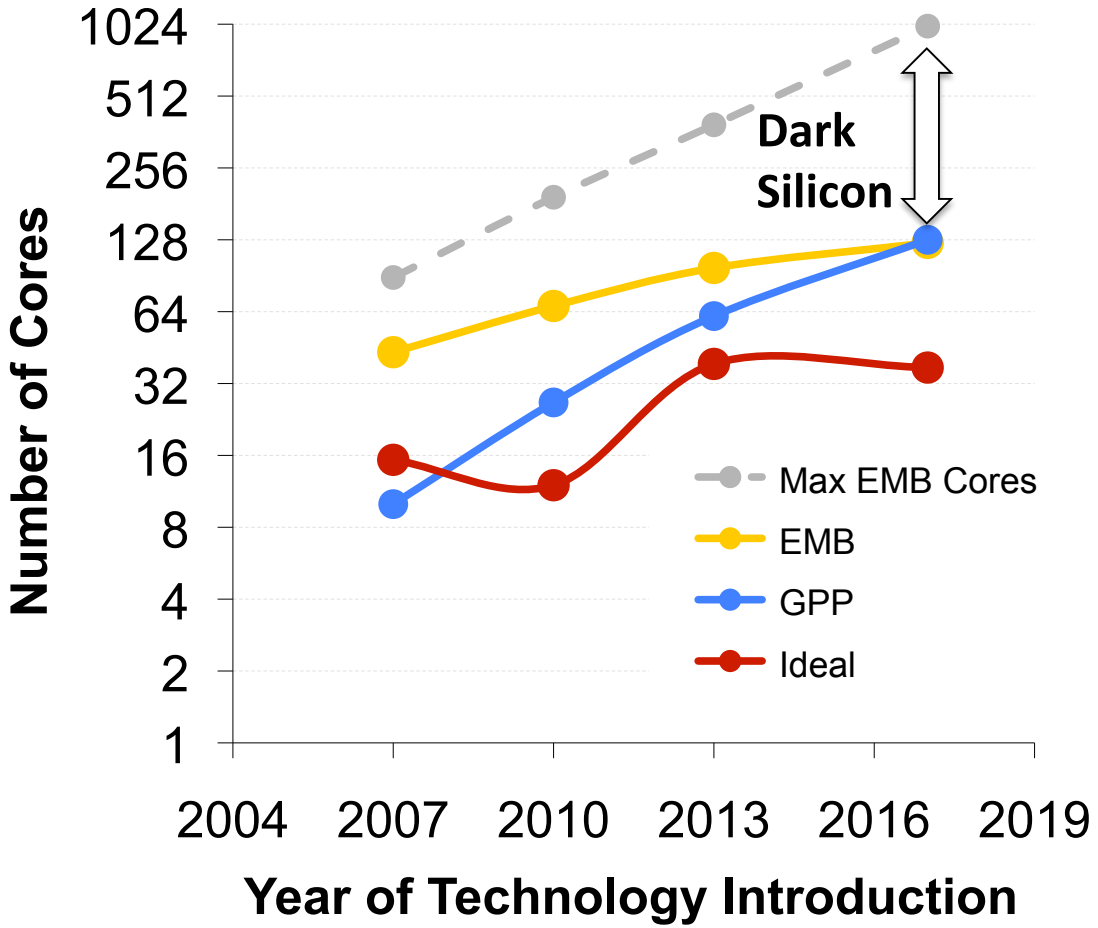
Can not power up chip

Parallelism has limits

even in Servers!

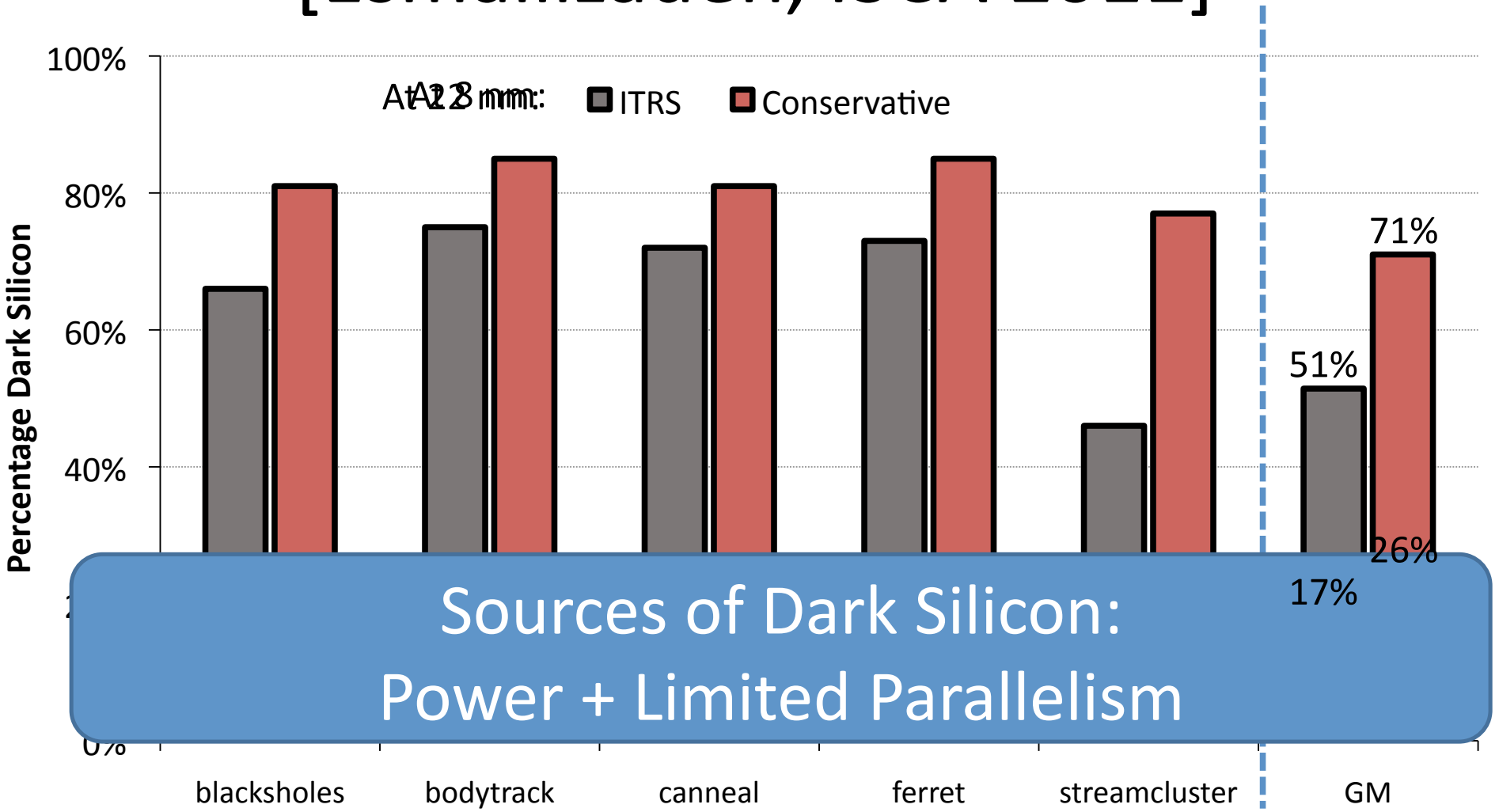
Must:

- specialize
- selectively power up



Hardavellas et. al., "Toward Dark Silicon in Servers", IEEE Micro, July 2011

Dark Silicon in Parallel Desktop Apps [Esmailizadeh, ISCA 2011]



Sources of Dark Silicon:
Power + Limited Parallelism

Where to go from here?

1. Energy-centric IT
   - Minimize joules/work (algo. down to HW)
   - Program for locality + heterogeneity

2. Pray for technology
   - Energy-scalable silicon devices
   - Emerging nanoscale technologies?

3. Infrastructure technology
   - Renewable/carbon-neutral energy
   - Scalable cooling + power delivery

Course Roadmap

- What is up with power (energy)?
  - End of Dennard scaling
  - What does it mean for servers?
- How do we tame power (energy)?
  - What do servers want?
  - What will server chips look like?

CloudSuite:
A Suite of Emerging Cloud Workloads

Soon to come from EPFL (ecocloud.ch)

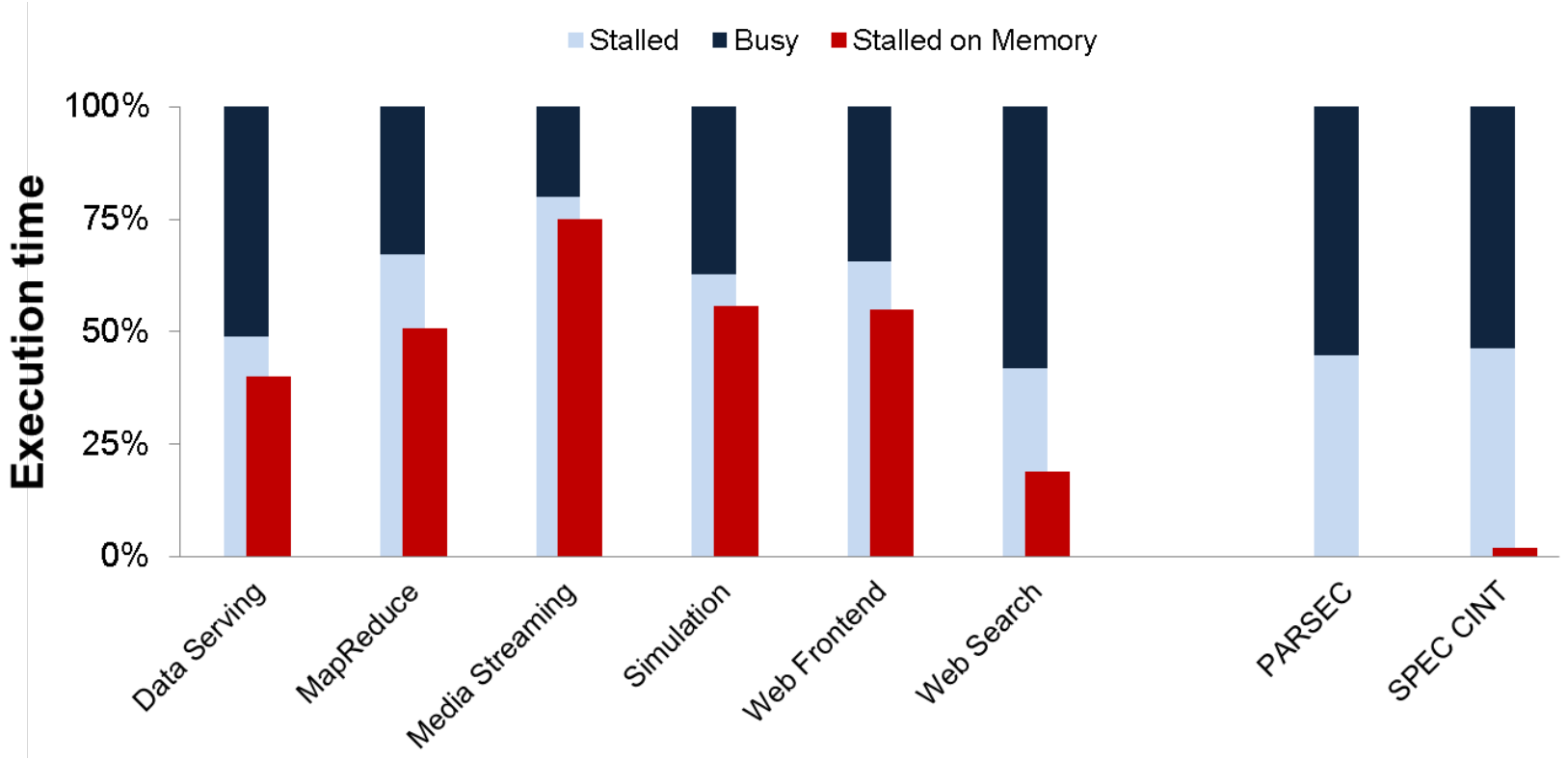
- Data serving
- Classification
- Media streaming
- Simulation
- Web frontend
- Web search

Ran Experiments on Nehalem Blades

	Hardware Specifications
Processor	Intel Xeon 5670, 6 cores, 32nm @2.93GHz
CMP Size	6 OoO cores
Superscalar width	4-wide issue
Reorder buffer	128 entries
Load/Store buffer	48/32 entries
Reservation stations	36 entries
L1 Cache	split I/D, 32KB, 4-cycles access latency
L2 Cache	6-core CMP: 256KB per core, 12-cycles access latency
LLC (L3) cache	12MB, cycles 39-cycles access latency
Memory	24GB, 180/280 cycles access latency local/remote DRAM

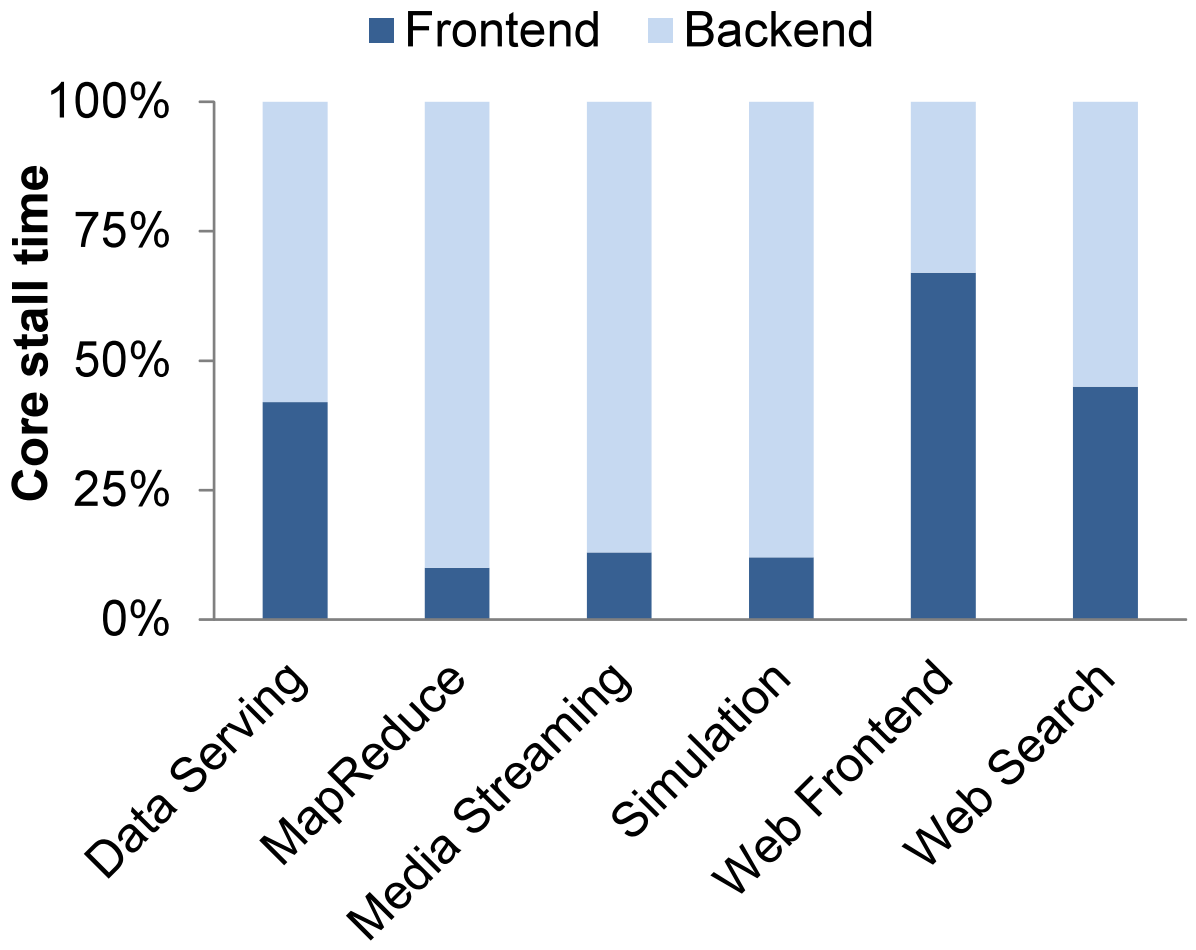
Where does time go?

Execution Breakdown



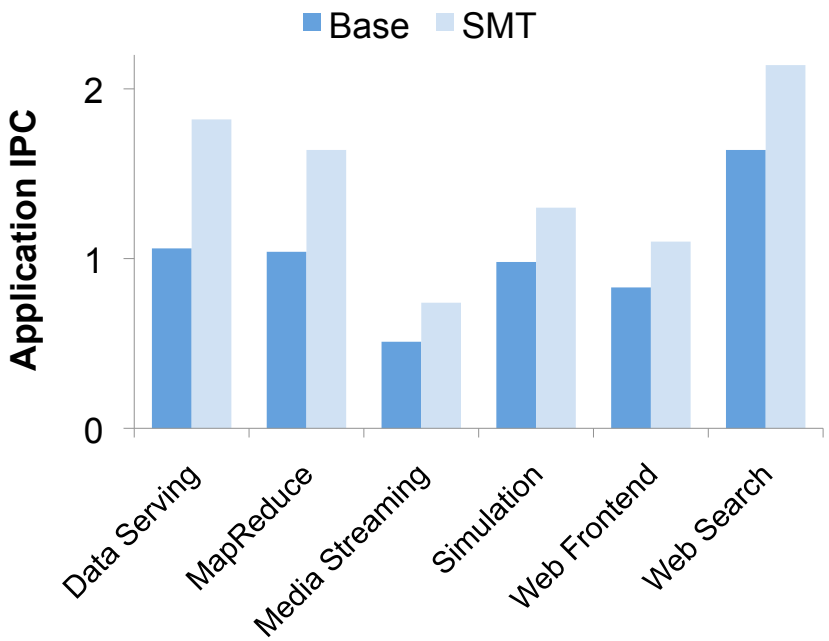
- Unlike desktop/RMS apps, memory stalls dominate
- Design should be centered around memory

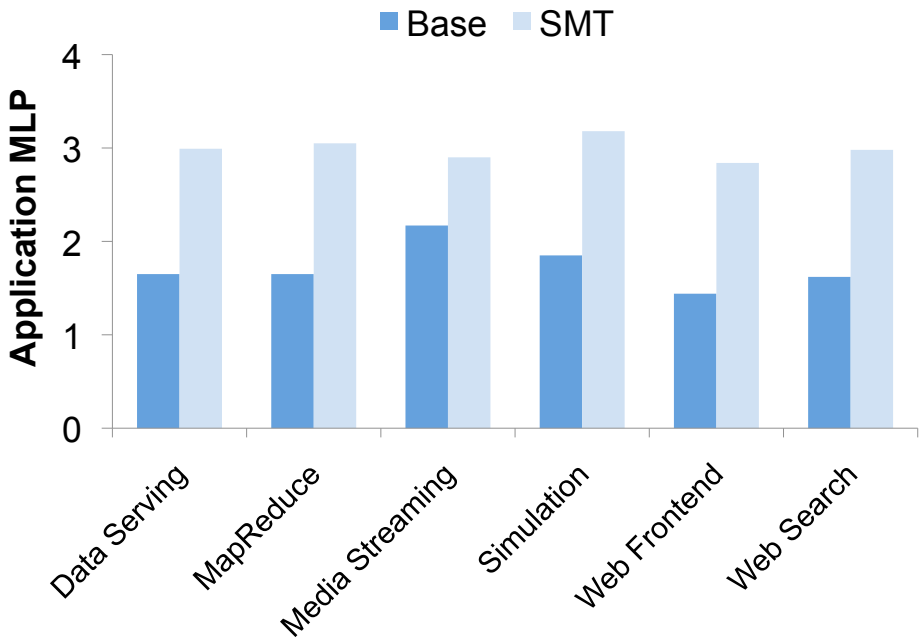
Front-End Inefficiencies



- Instruction fetch: 10-60% of total stalls
- Next-line prefetch. (in the CPU) not efficient

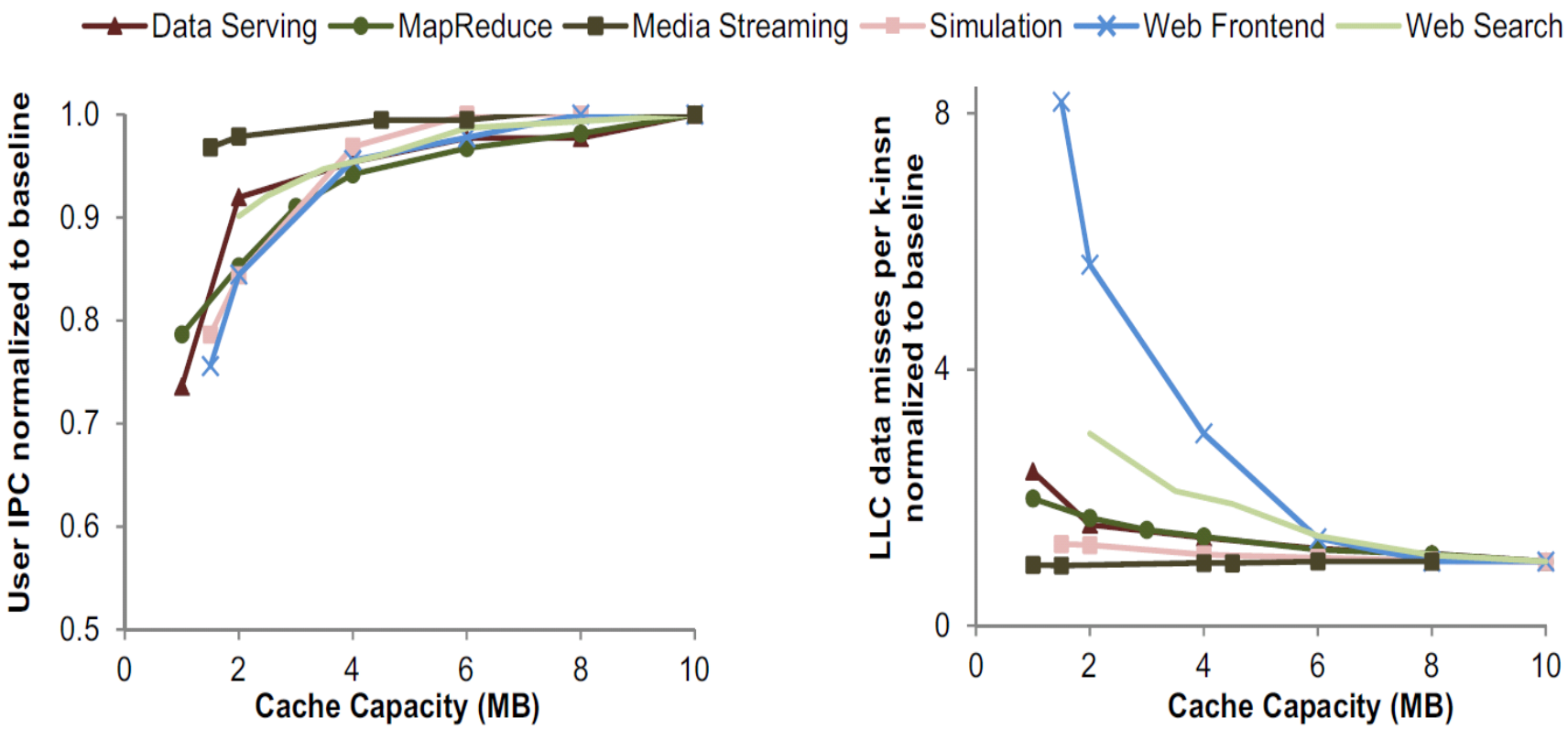
Core Inefficiencies





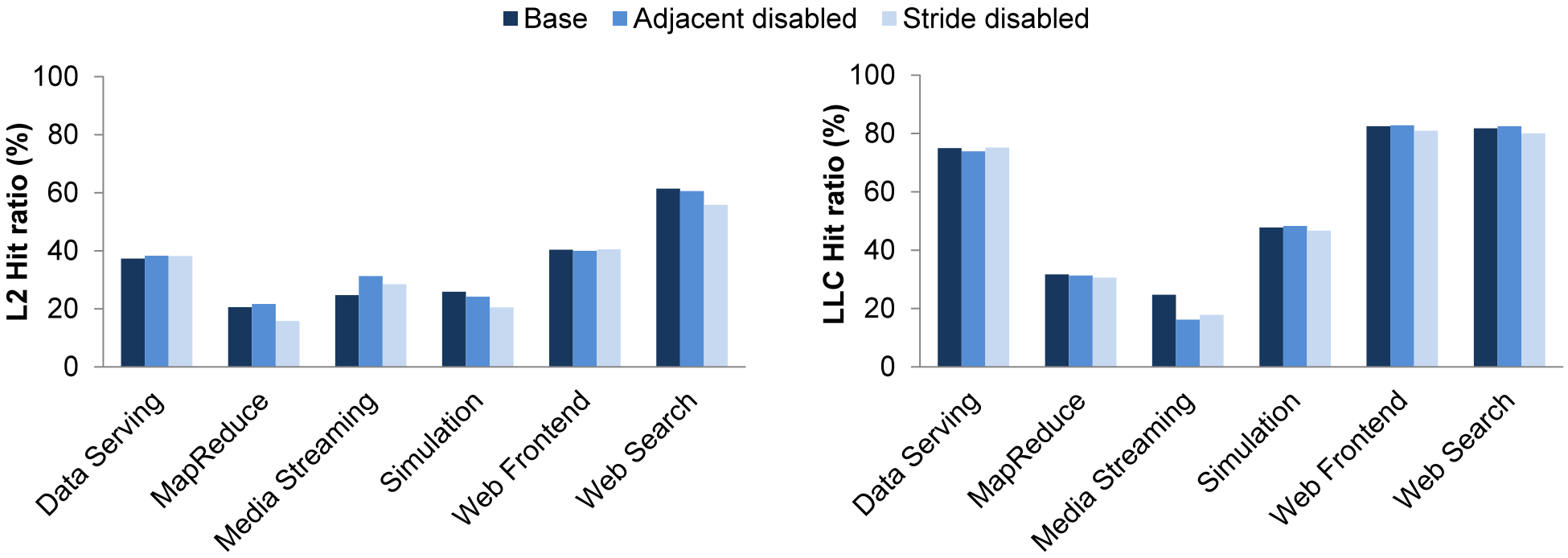
- Low IPC & MLP despite 4-wide OoO core
- Using SMT doubles MLP
- But, SMT achieves only 30% performance gain
  - Threads compete for core resources
  - Intel's SMT fetch not effective

Cache Capacity (LLC) Inefficiencies



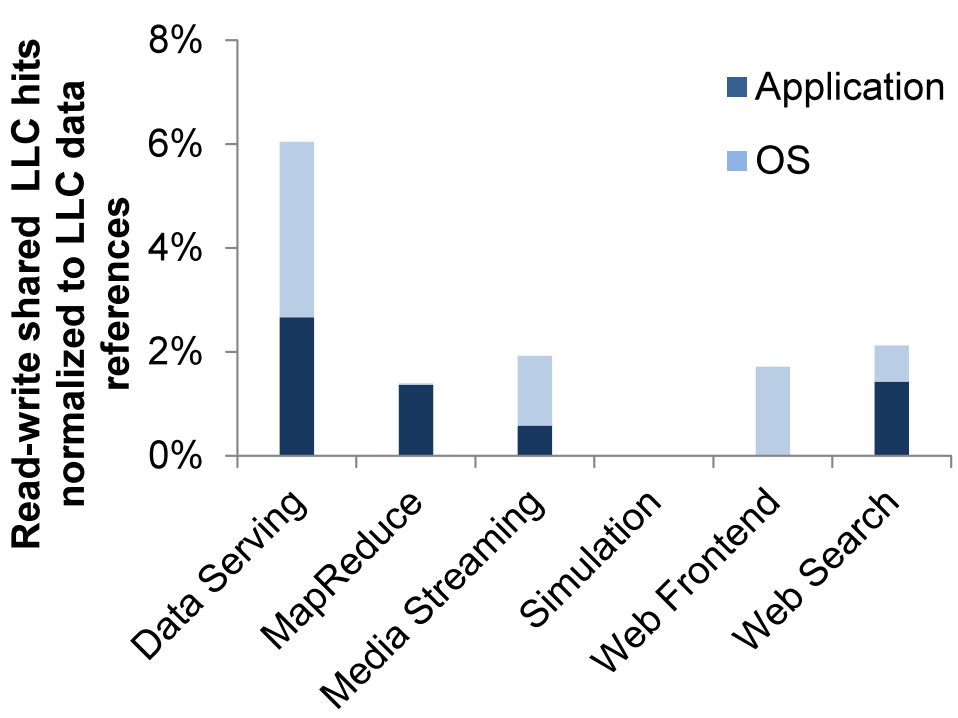
- Large LLC consumes area, but has diminishing returns
- Results (not shown) indicate much LLC is instructions

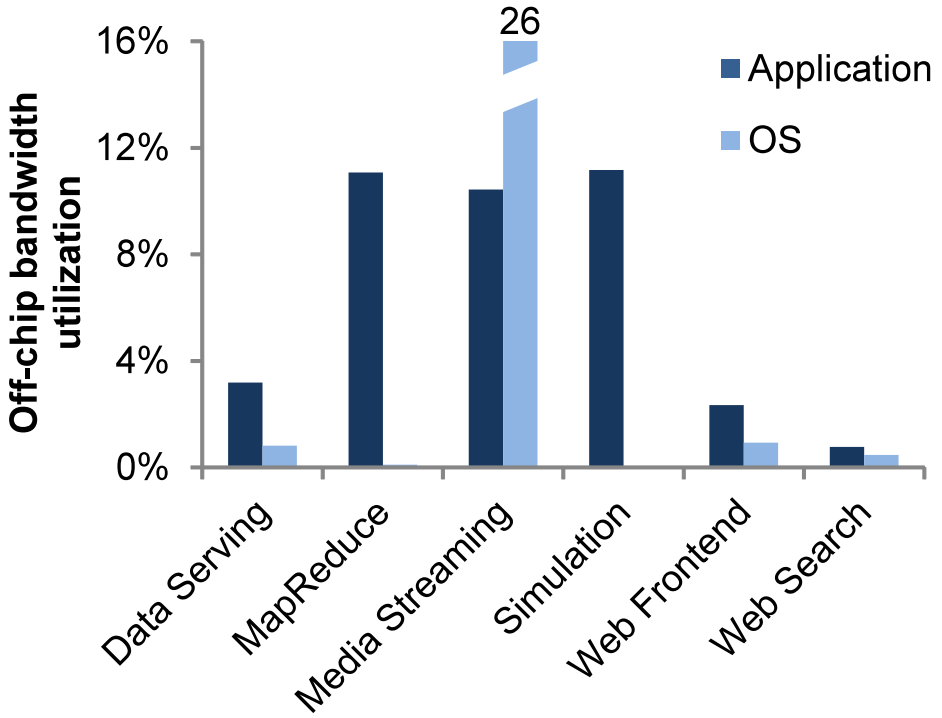
Data Prefetching Inefficiencies



- Existing prefetchers are ineffective
- Pointer-intensive patterns [Wenisch 2005]

Bandwidth Inefficiencies





- Low sharing among working threads
- No need for on-chip shared caches
- Today, pin bandwidth is overprovisioned

CloudSuite Conclusions

Cloud workloads need:

- Simple (multithreaded) cores
- Partitioned caches (no sharing)
- Large on-chip instruction footprints
- Advanced prefetchers

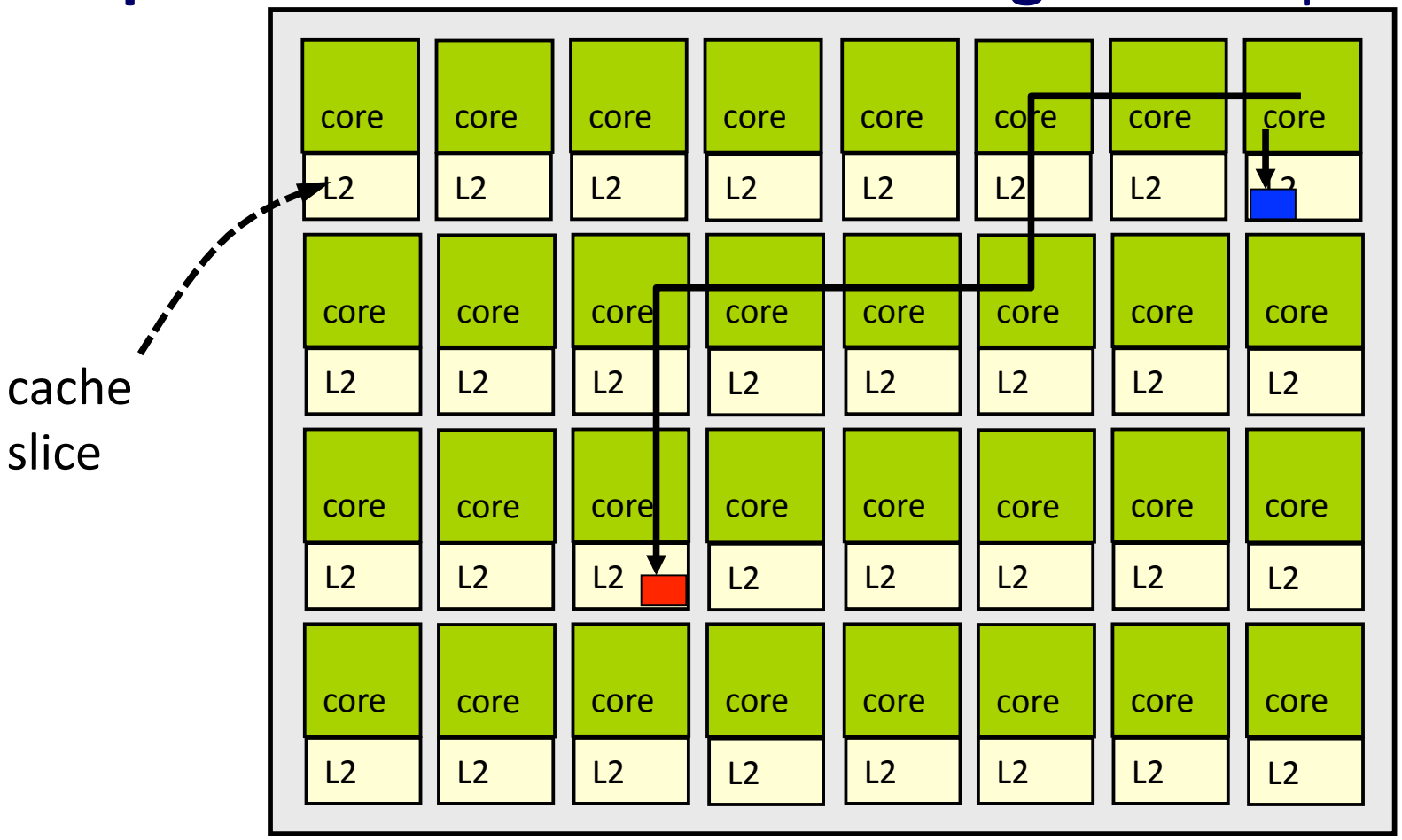
Course Roadmap

- What is up with power (energy)?
  - End of Dennard scaling
  - What does it mean for servers?
- How do we tame power (energy)?
  - What do servers want?
  - What will server chips look like?

Short-term Scaling Implications

- Caches are getting huge
  - Need cache architectures to deal with >> MB
  - E.g., Reactive NUCA [ISCA'09]
- Interconnect + cache hierarchy power
  - Need lean on-chip communication/storage
  - Eurocloud chip: ARM+3D [ACLD'10]
- Dark Silicon
  - Specialized processors
  - Use only parts of the chip at a time

Optimal Data Placement in Large On-chip Caches



- Data placement determines performance
- Goal: place data on chip close to where they are used

Prior Work

- Several proposals for CMP cache management
  - ASR, cooperative caching, victim replication, CMP-NuRapid, D-NUCA

- ...but suffer from shortcomings
  - complex, high-latency lookup/coherence
  - don't scale
  - lower effective cache capacity
  - optimize only for subset of accesses

We need:

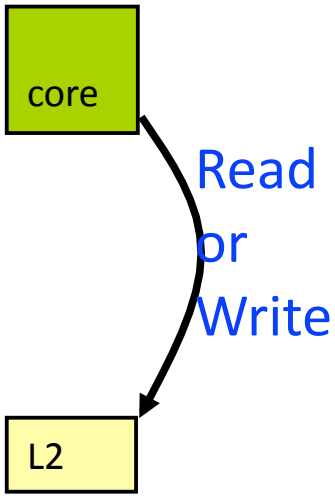
➡ Simple, scalable mechanism for fast access to all data

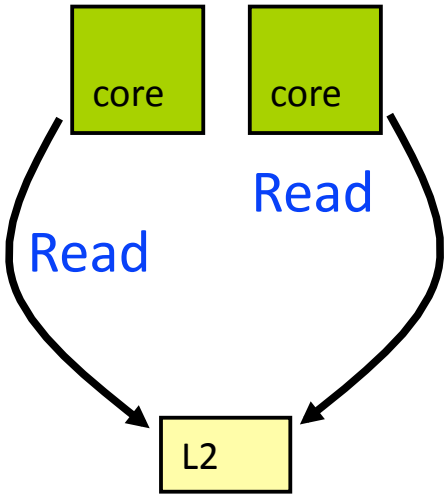
Our Proposal: Reactive NUCA

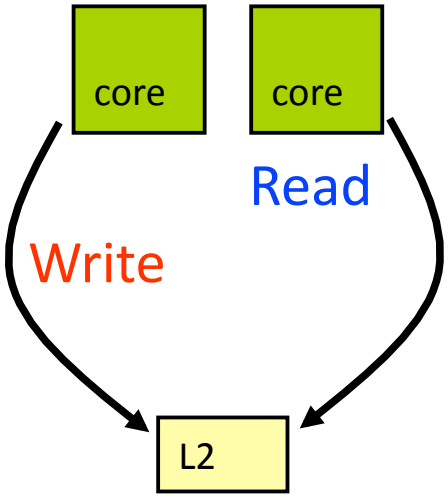
[ISCA'09, IEEE Micro Top Picks '10]

- Cache accesses can be classified at run-time
  - Each class amenable to different placement
- Per-class block placement
  - Simple, scalable, transparent
  - No need for HW coherence mechanisms at LLC
- Speedup
  - Up to 32% speedup
  - -5% on avg. from ideal cache organization

Terminology: Data Types







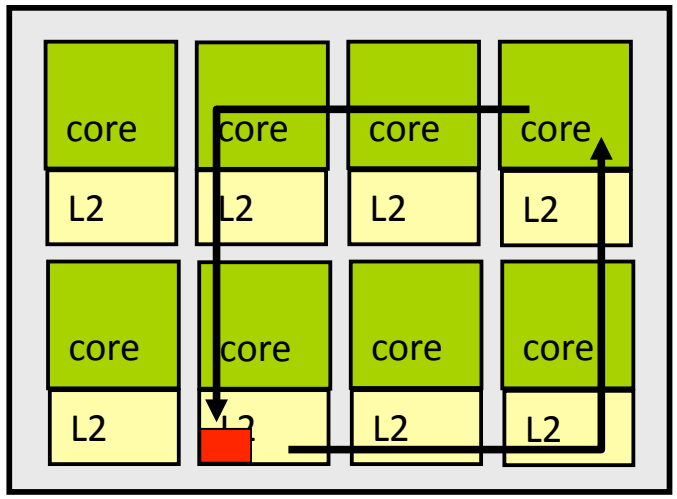
Private

Shared
Read-Only

Shared
Read-Write

Conventional Multicore Caches

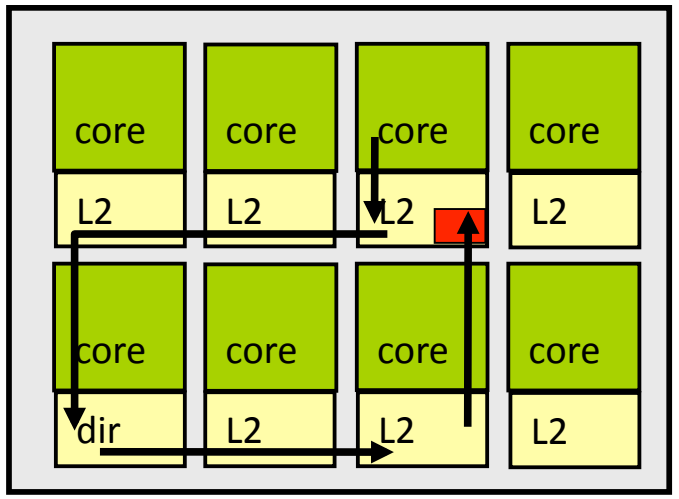
Shared



Addr-interleave blocks

+ High effective capacity
− Slow access

Private



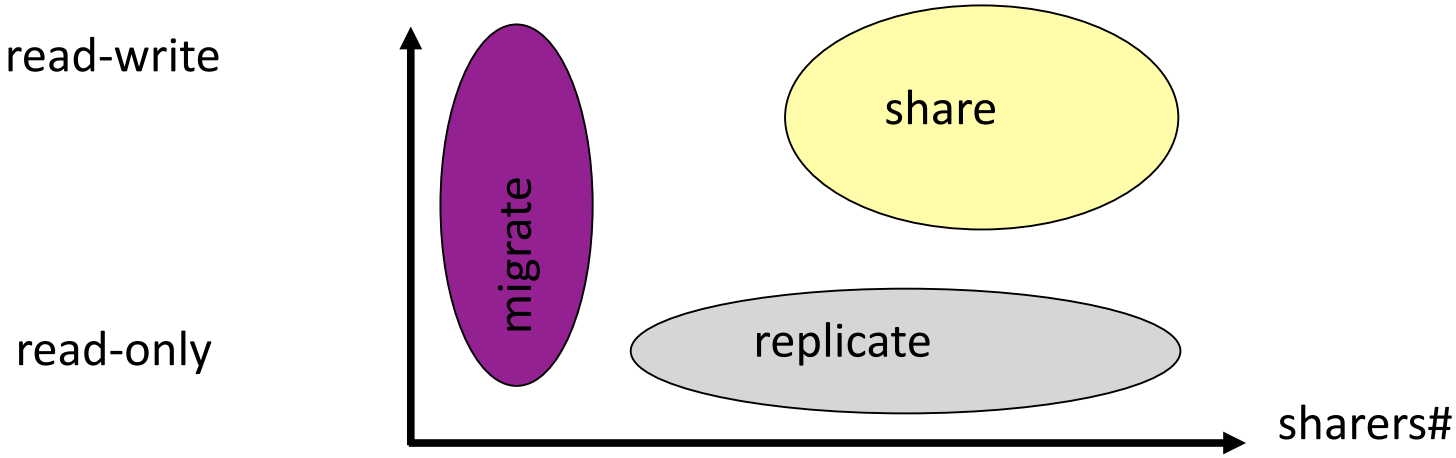
Each block cached locally

+ Fast access (local)
− Low capacity (replicas)
− Coherence: via indirection (distributed directory)

➡ We want: high capacity (shared) + fast access (priv.)

Where to Place the Data?

- Close to where they are used!
- Accessed by single core: migrate locally
- Accessed by many cores: replicate (?)
  - If read-only, replication is OK
  - If read-write, coherence a problem
    - Low reuse: evenly distribute across sharers



Methodology

Flexus: Full-system cycle-accurate timing simulation

Workloads

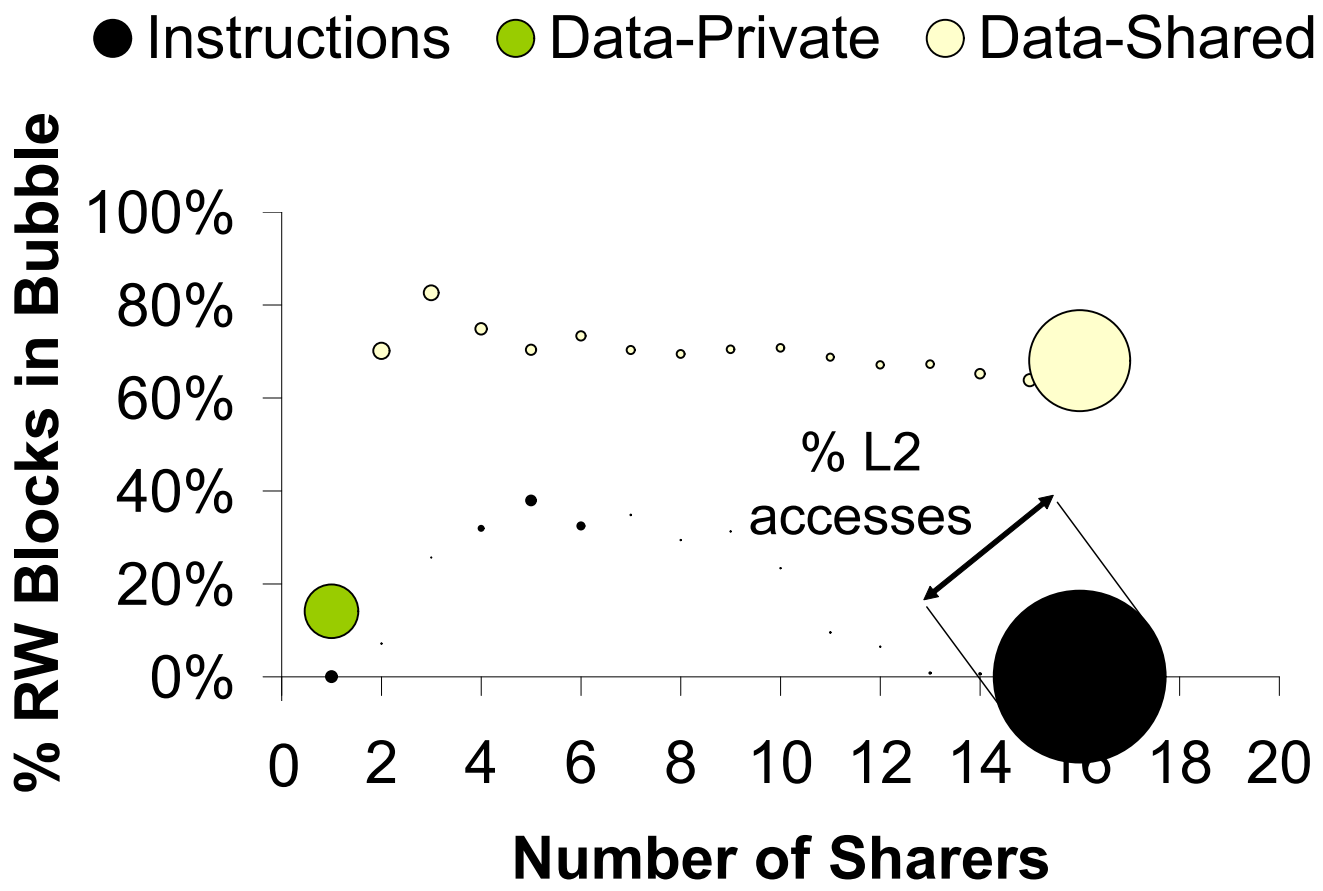
- OLTP: TPC-C 3.0 100 WH
  - IBM DB2 v8
  - Oracle 10g
- DSS: TPC-H Qry 6, 8, 13
  - IBM DB2 v8
- SPECweb99 on Apache 2.0
- Multiprogammed: SPEC2K
- Scientific: em3d

Model Parameters

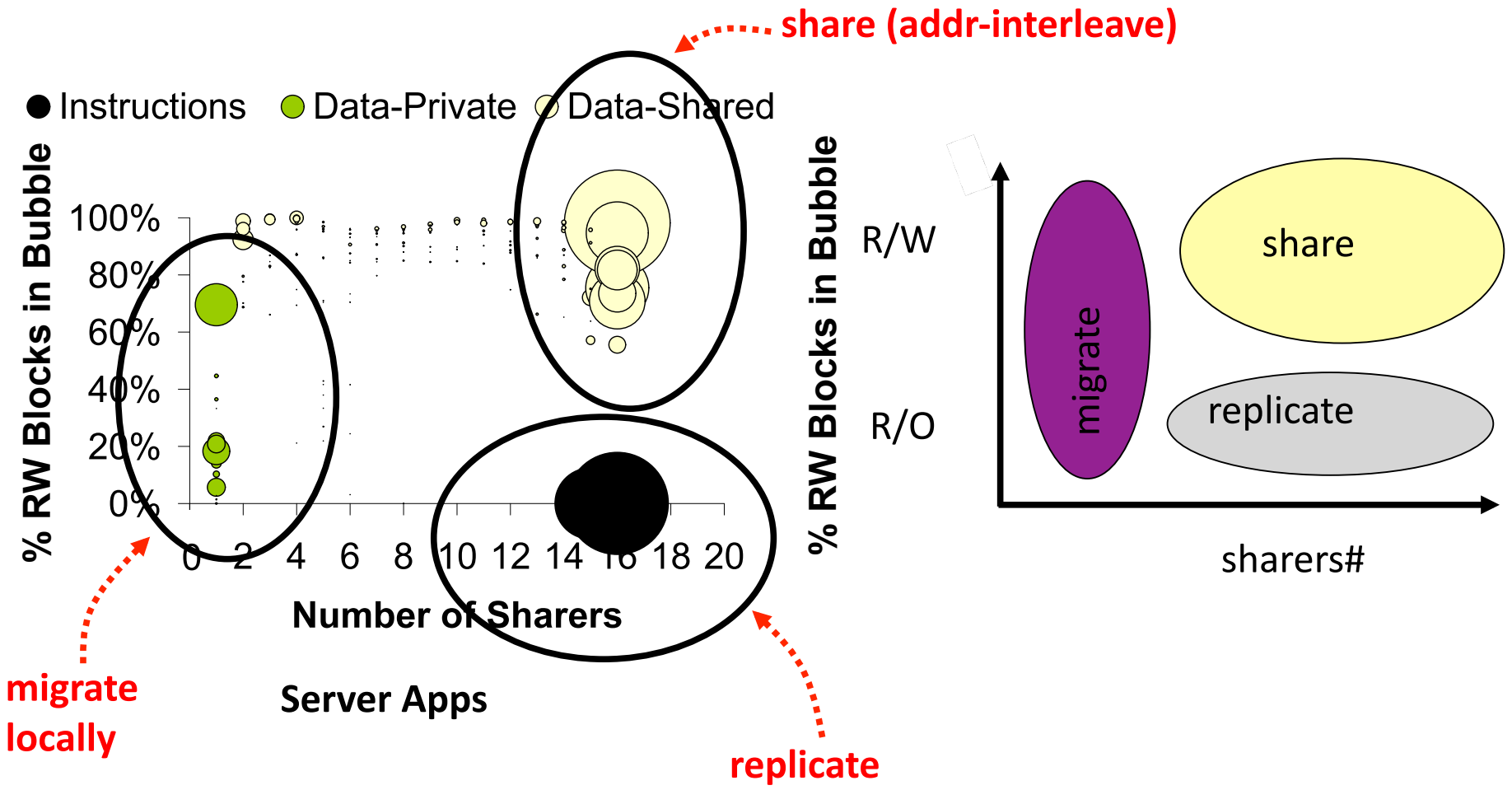
- Tiled, LLC = L2
- 16-cores, 1MB/core
- OoO, 2GHz, 96-entry ROB
- Folded 2D-torus
  - 2-cycle router
  - 1-cycle link
- 45ns memory

Cache Access Classification

- Each bubble: cache blocks shared by x cores
- Size of bubble proportional to % L2 accesses
- y axis: % blocks in bubble that are read-write



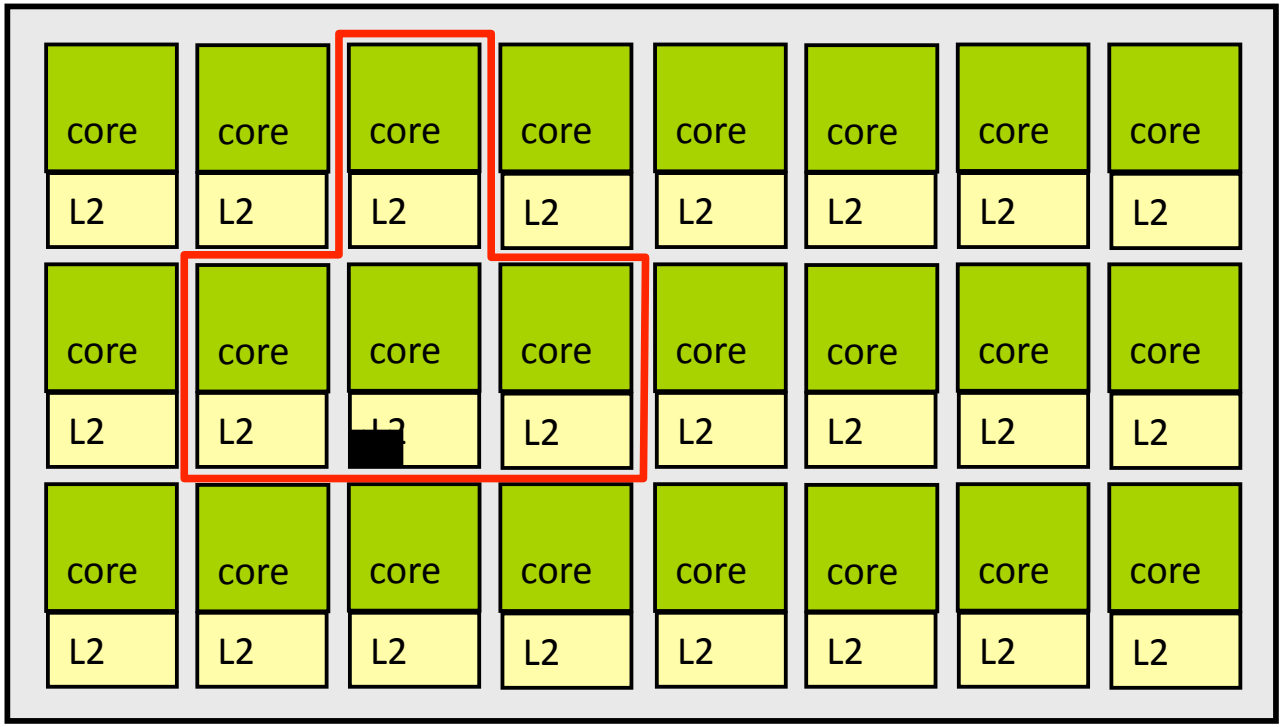
Cache Access Clustering



- Accesses naturally form 3 clusters

Instruction Replication

- Instruction working set too large for one cache slice

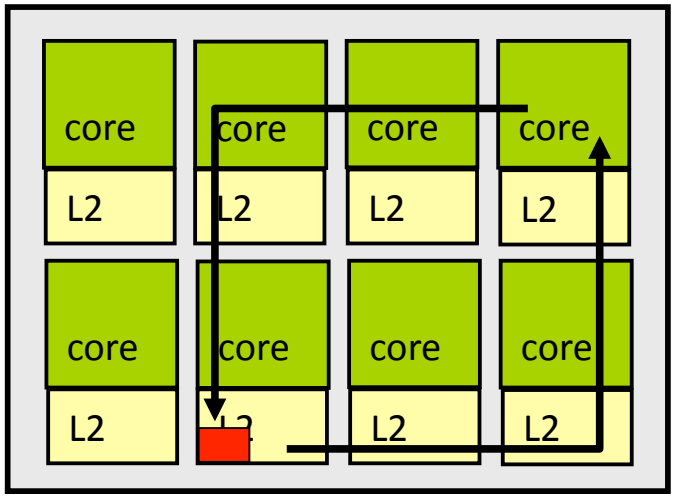


➡ Distribute in cluster of neighbors, replicate across

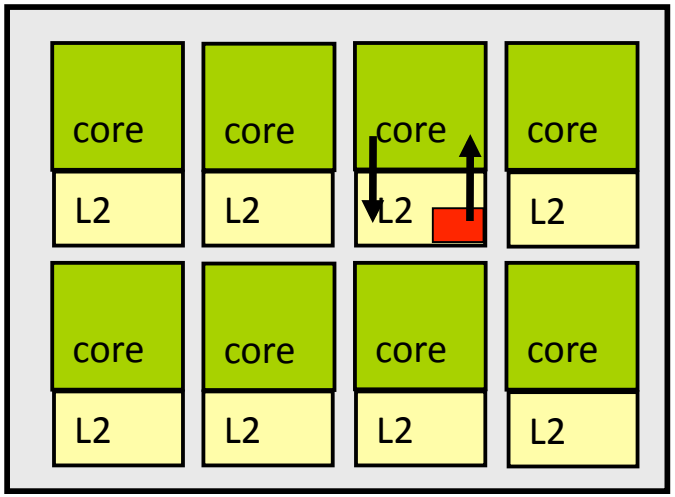
No Need for HW Coherence in LLC

- Reactive NUCA placement guarantee
  - Each R/W datum in unique & known location

Shared data: addr-interleave

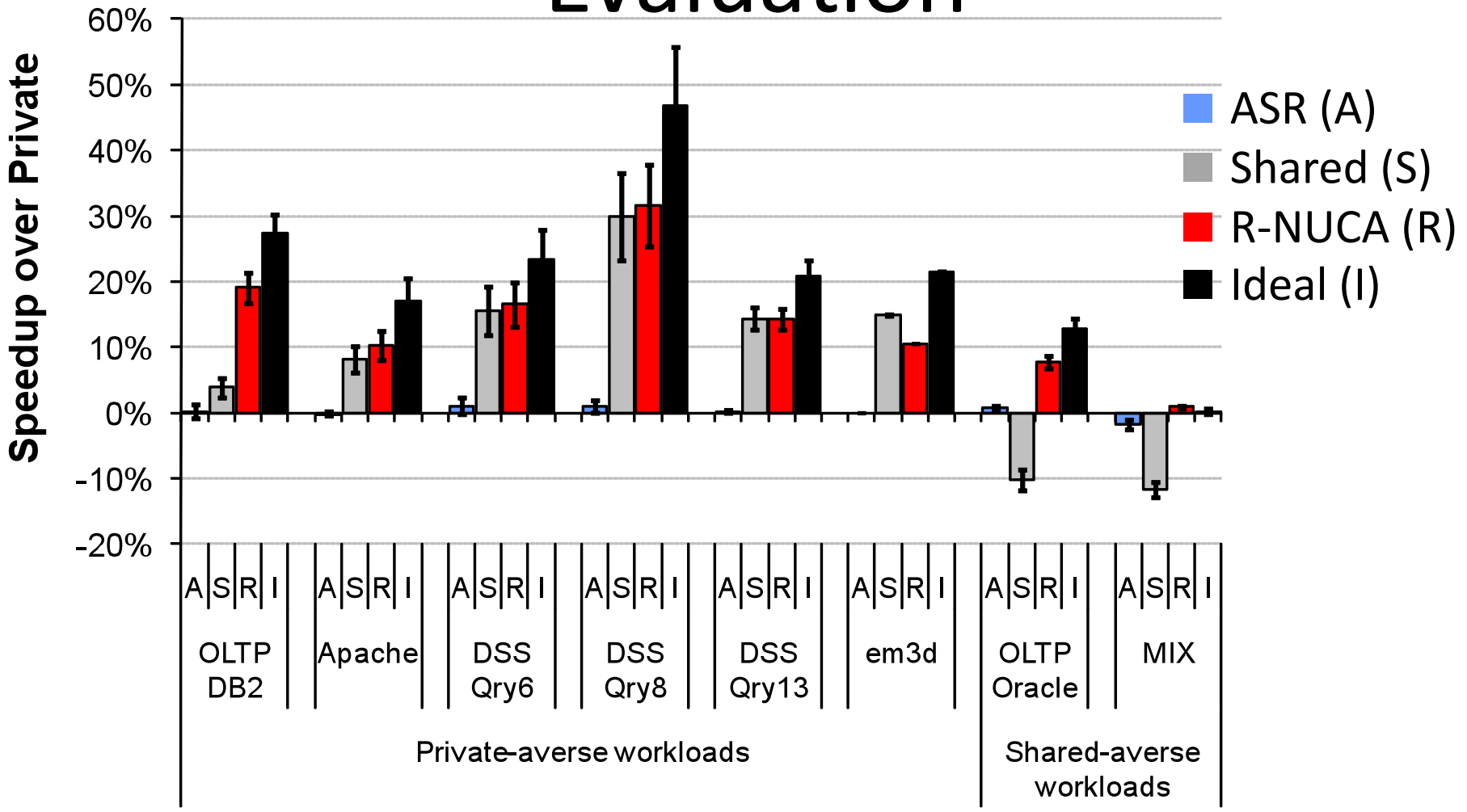


Private data: local slice



➡ Fast access, eliminates HW overhead

Evaluation



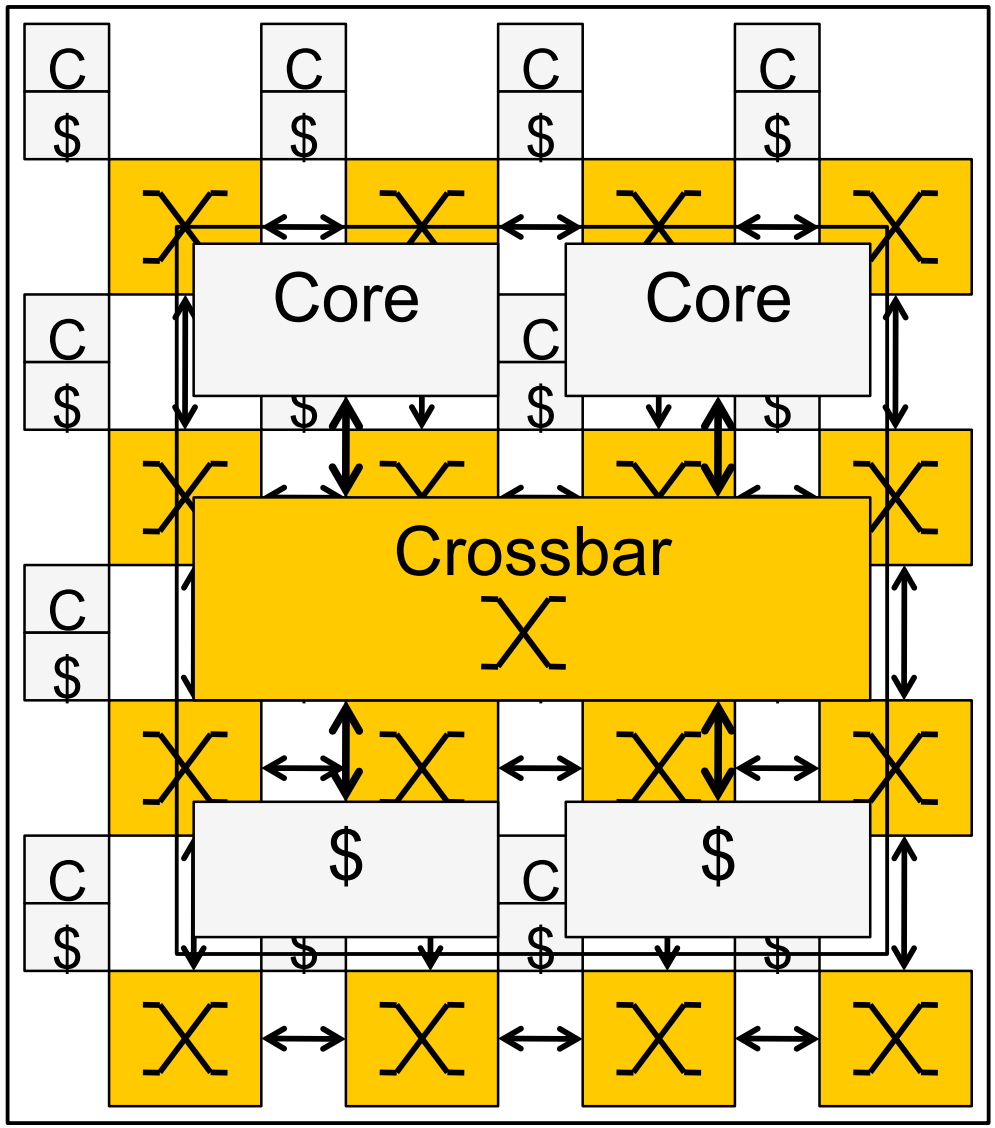
- Delivers robust performance across workloads
  - Shared: same for Web, DSS; 17% for OLTP, MIX
  - Private: 17% for OLTP, Web, DSS; same for MIX

R-NUCA Conclusions

Near-optimal block placement and replication in distributed caches

- Cache accesses can be classified at run-time
  - Each class amenable to different placement
- Reactive NUCA: placement of each class
  - Simple, scalable, low-overhead, transparent
  - Obviates HW coherence mechanisms for LLC
- Robust performance across server workloads
  - Near-optimal placement (-5% avg. from ideal)

NoCs: Major Power Consumer



- Move towards manycore
  - Tiled architectures
- Network-on-Chip (NoC)
  - Significant power consumer
  - 40% MIT RAW
  - 30% Intel Tera-scale
- Cache coherent CMP
  - Server workloads

Proposals to Reduce NoC Power

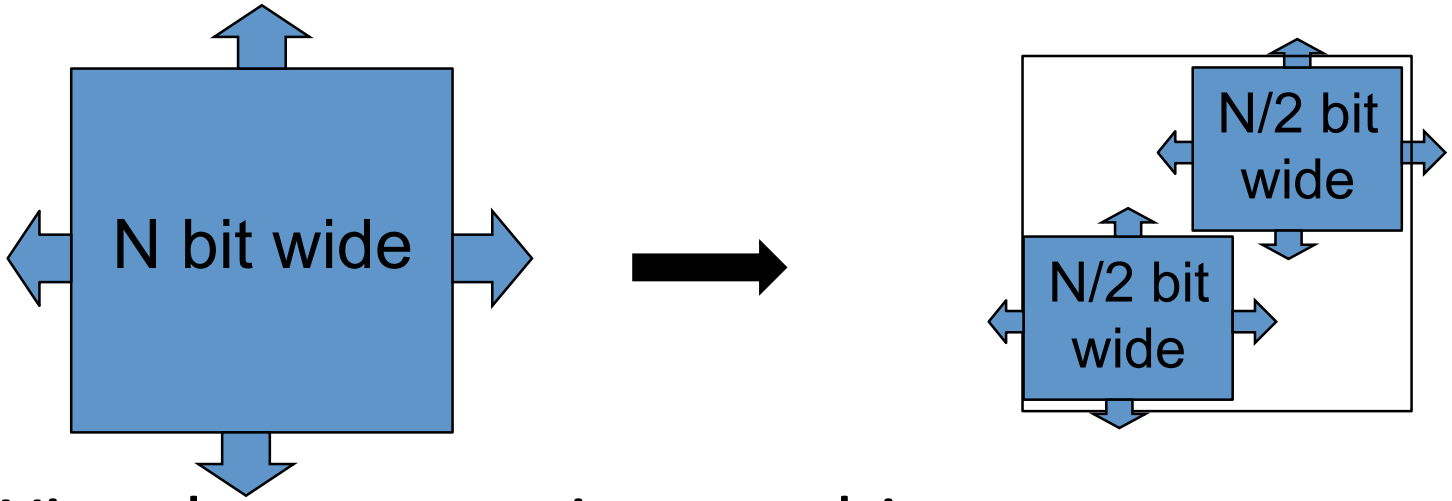
- Multiple networks
  - Better area and power [Balfour & Dally ICS 2006]
- Commercial server workloads
  - Traffic patterns are different
- Run on cache coherent CMPs
  - Strong relation between coherence protocol and NoC
- Not optimized for Server Workload traffic

Our Proposal: CCNoC [WEED 2011]

- Commercial server workloads
  - Optimized for reuse in L1, little sharing
  - Full blown coherence protocol in CMPs
  - Only some transitions are frequent
- Duality in Request/Response message size
- CCNoC
  - Full advantage of heterogeneity
  - Same number of buffers
  - 16% less power same performance as Mesh

Dual Router is More Efficient

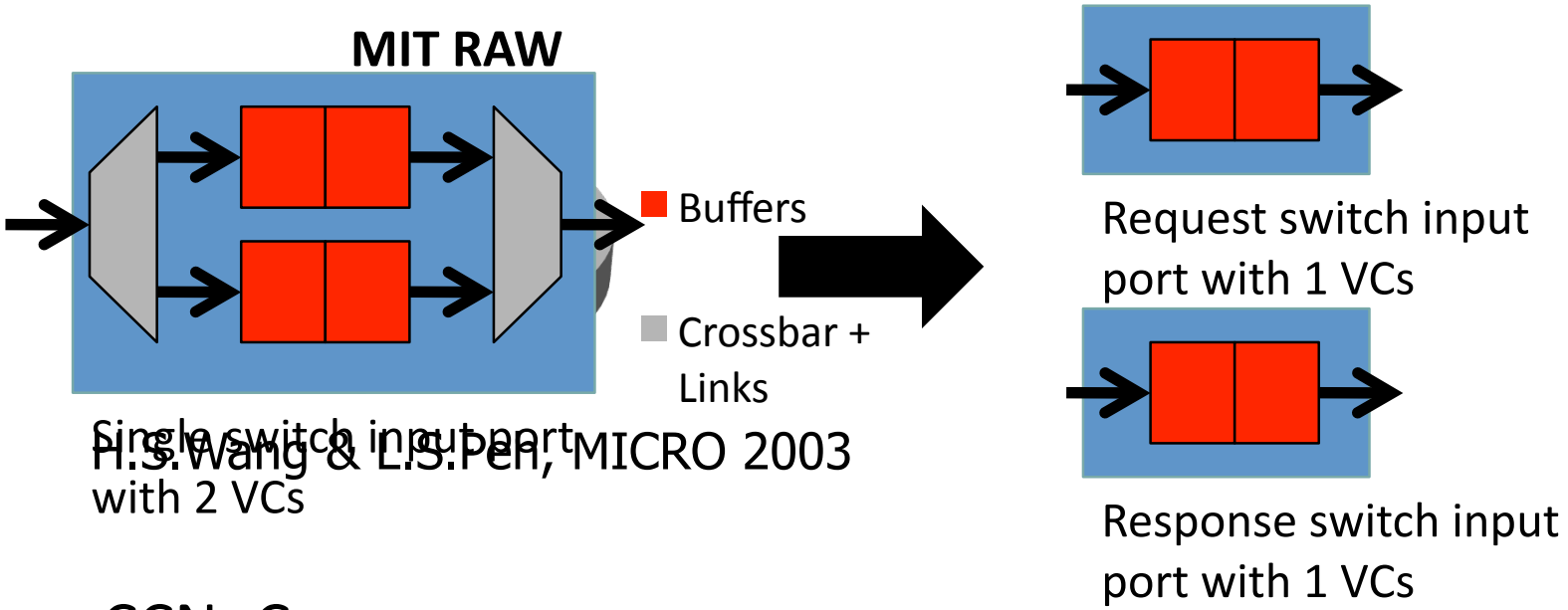
- Dual router
  - Two crossbars per routing node



- Wires less expensive on-chip
  - Use more wires for better performance
- Area and power grows faster than connectivity
  - Balfour & Dally ICS 2006
  - Dual router: better performance, power and area

Right Dual Router Design

- Avoid protocol level deadlock
  - Separate Requests/Responses with Virtual Channels

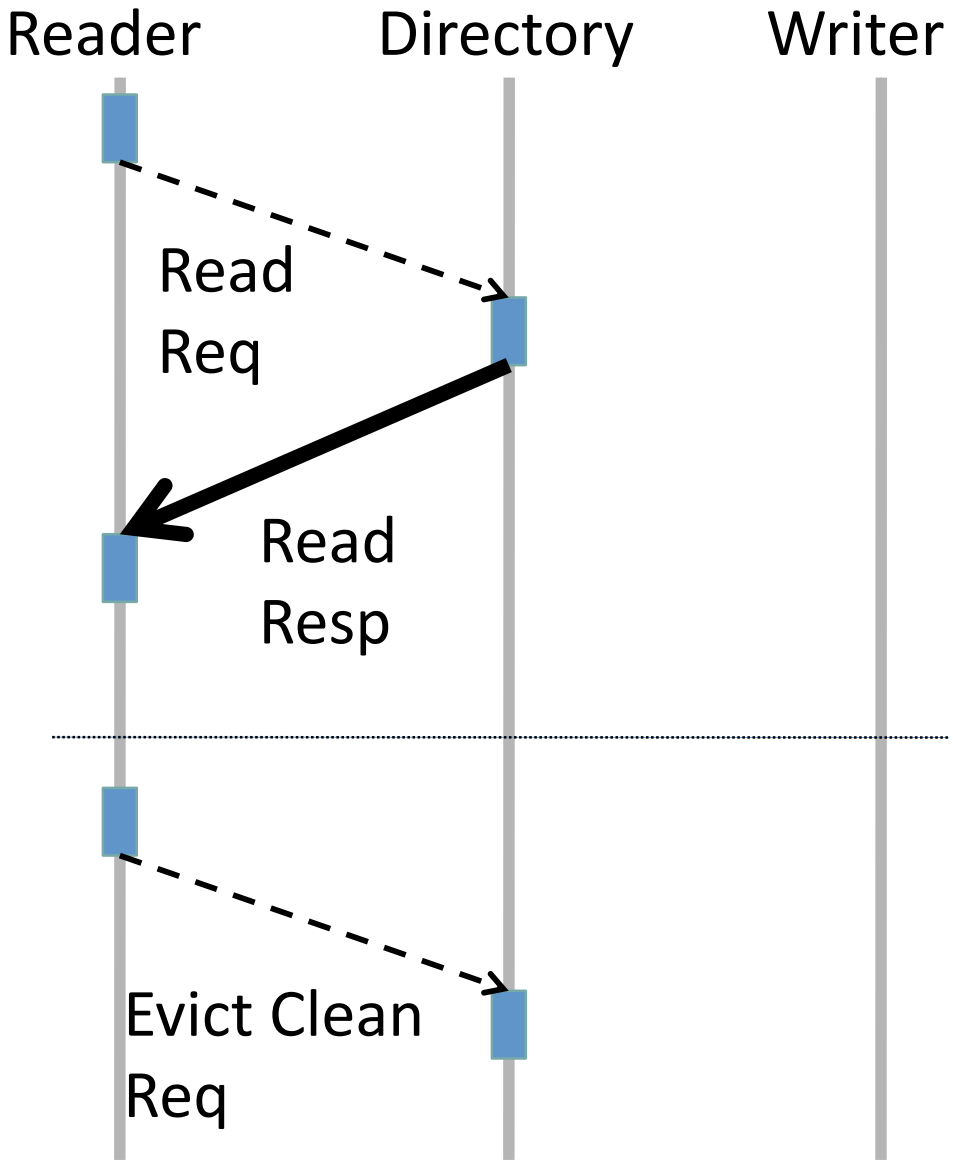


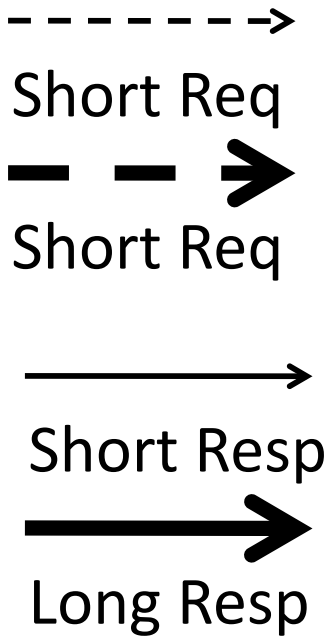
- CCNoC
  - Request / Response sub-networks
  - Same number of buffers

Protocol Activity

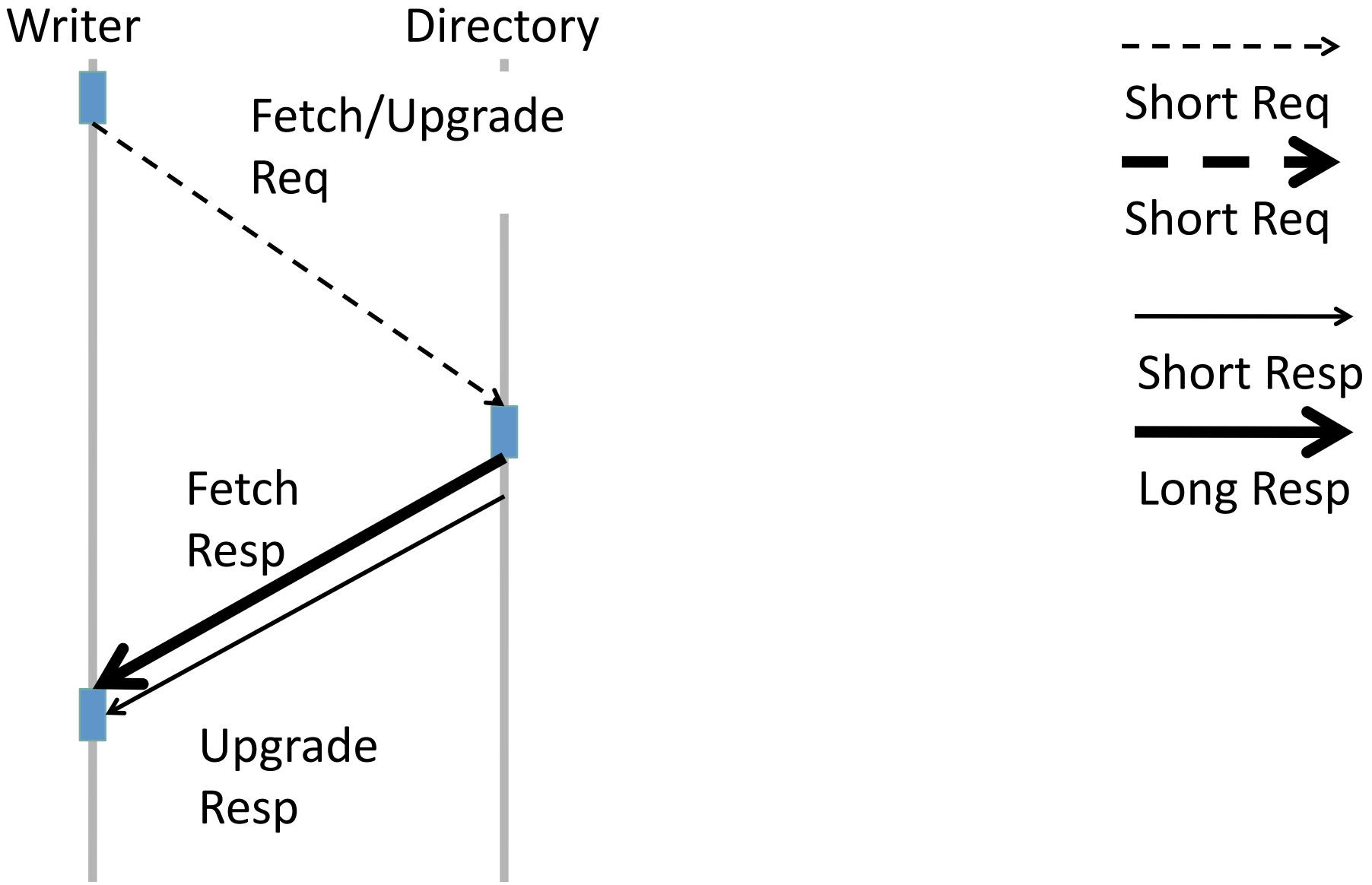
- CMPs implement full blown coherence protocol
  - Some transitions are frequent [Hardavellas ISCA 2009]
    - Read clean block
    - Evict clean block
    - Write to unshared block
  - Other transitions needed for correctness (infrequent)
    - Read dirty block
    - Evict dirty
    - Write to shared block

Frequent Read Protocol Activity

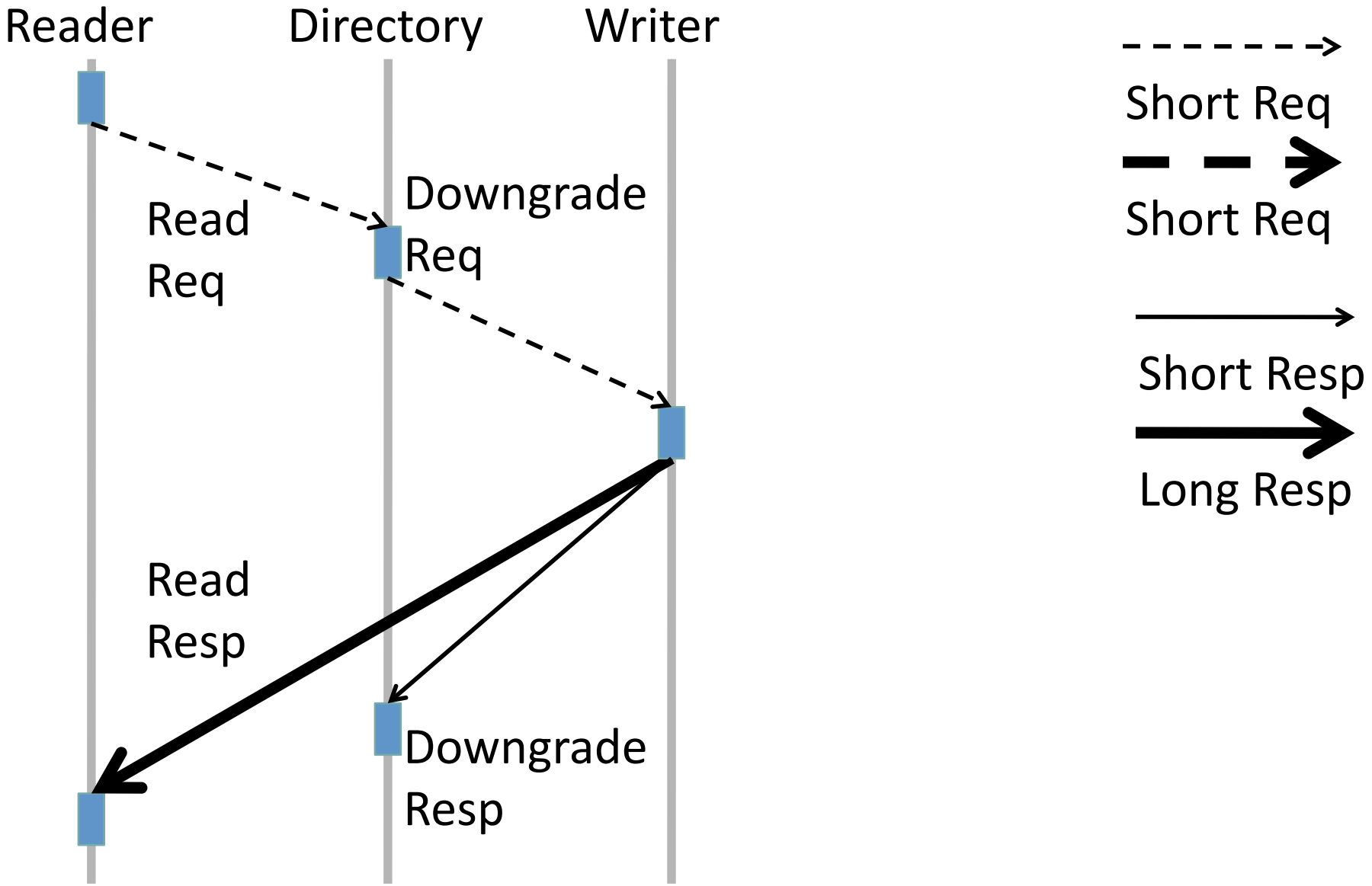




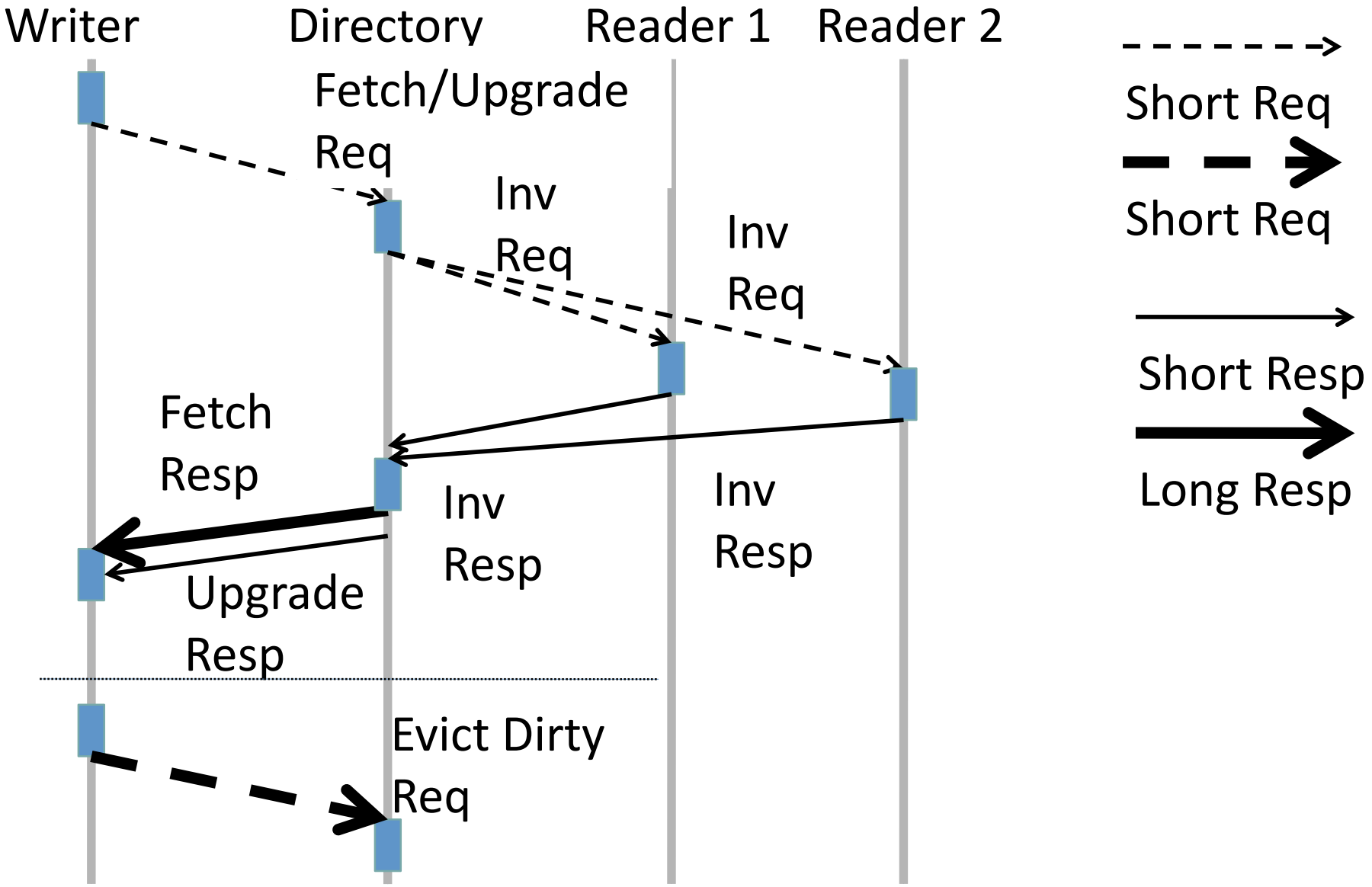
Frequent Write Protocol Activity



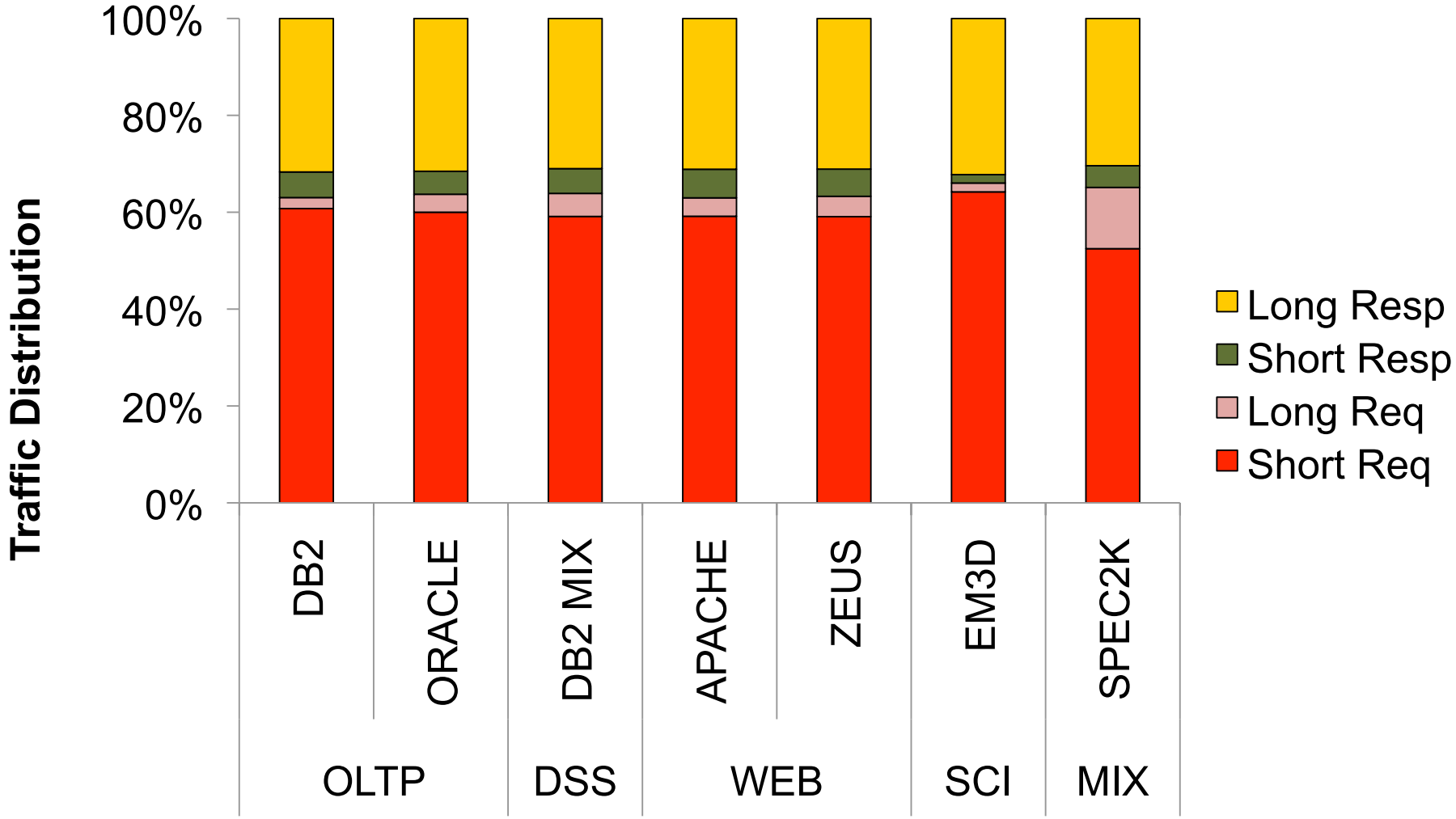
Infrequent Read Protocol Activity



Infrequent Write Protocol Activity

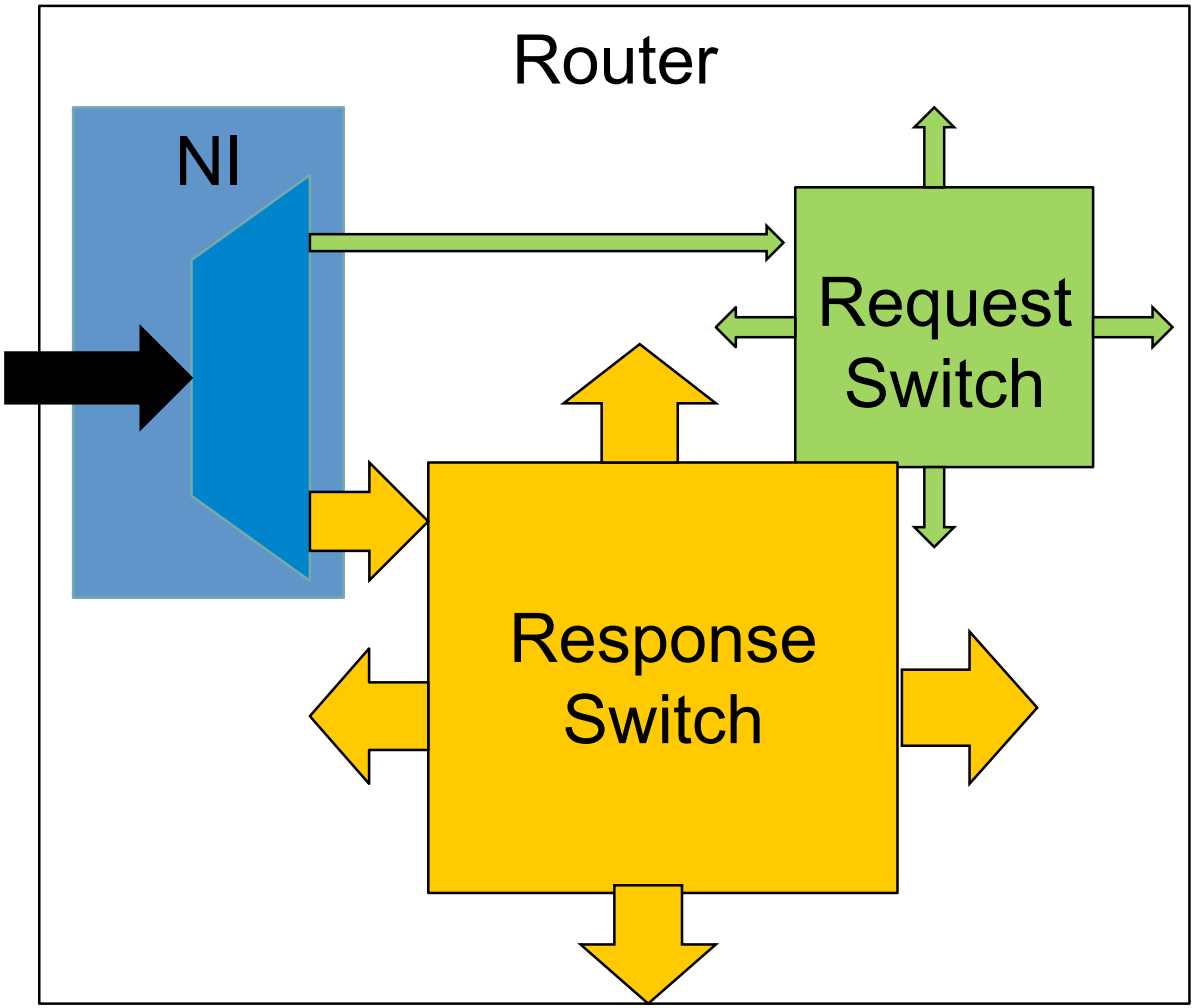


Traffic Analysis



Request: 93% short Response: 86% long

CCNoC Router



Request network narrow: optimized for short messages

Response network wide: optimized for long messages

Previous Work

- Balfour et al. ICS 2006
  - Better than single large router
  - Read/Write traffic
  - Same number of reads and writes
- Yoon et al. DAC 2010
  - Physical channel better than virtual channel
- Not optimized for cache coherent CMP
  - Running commercial server workloads

Evaluation Methodology

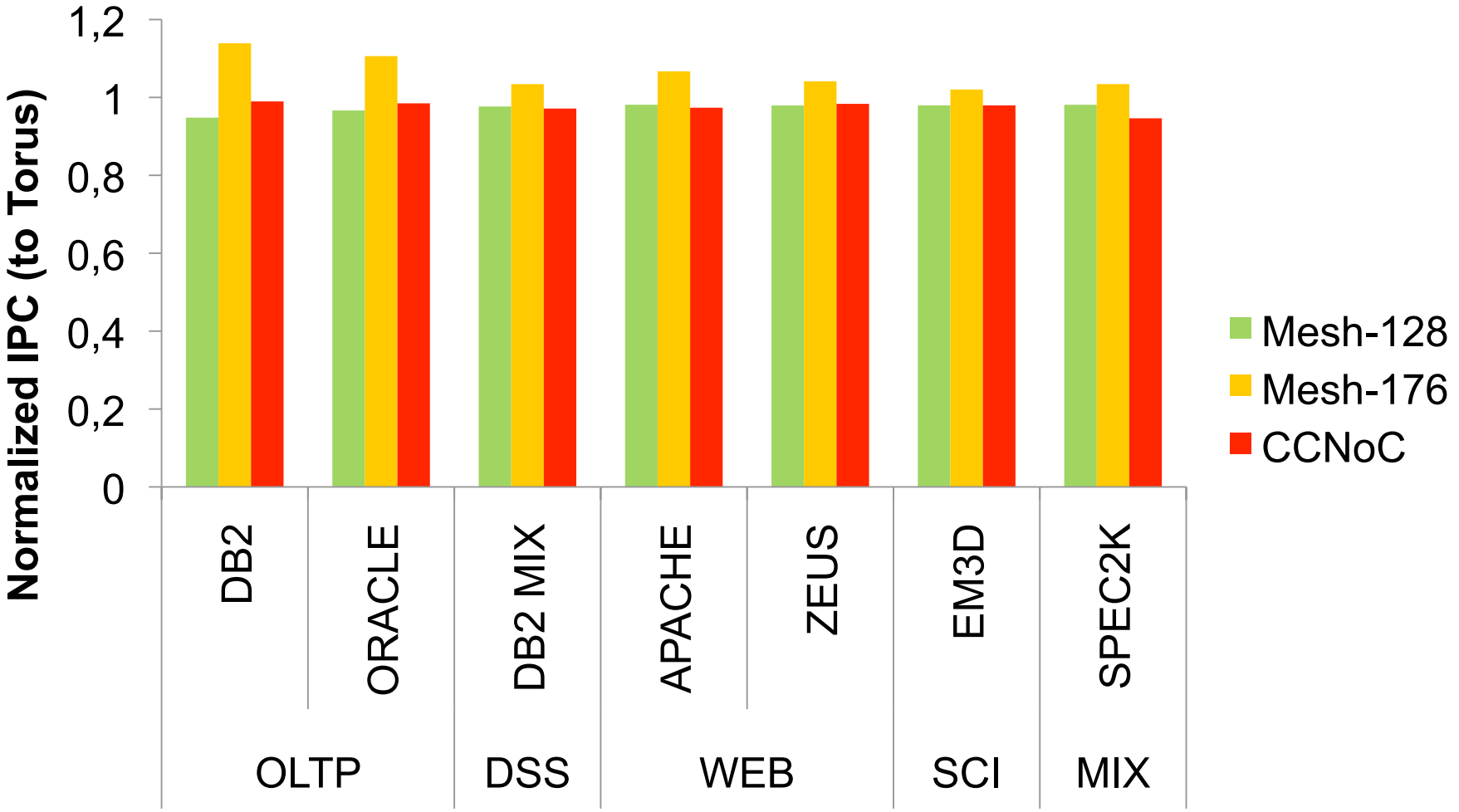
- FLEXUS
  - Full system simulation
  - 16 or 8 UltraSPARC III ISA cores
  - Split I/D, 64KB L1
  - 1 or 2 MB L2
- ORION 2.0
  - power estimation
  - area estimation
- Workloads
  - OLTP: TPC-C
    - IBM DB2 and Oracle
  - DSS: TPC-H
    - IBM DB2
    - Q1, Q6, Q13, Q16
  - Web: SPECweb99
    - Apache and Zeus
  - Scientific: EM3D
  - Multiprogrammed:
    - SPEC2K
    - 2x: gcc, twolf, art, mcf

Evaluation NoCs

- Mesh-128 - *baseline*
  - 128 bit flit width
- Torus - *reference*
  - 128 bit flit width
- Mesh-176 – *high performance*
  - 176 bit flit width
- CCNoC
  - Request: 48 bit flit width
  - Response: 128 bit flit width

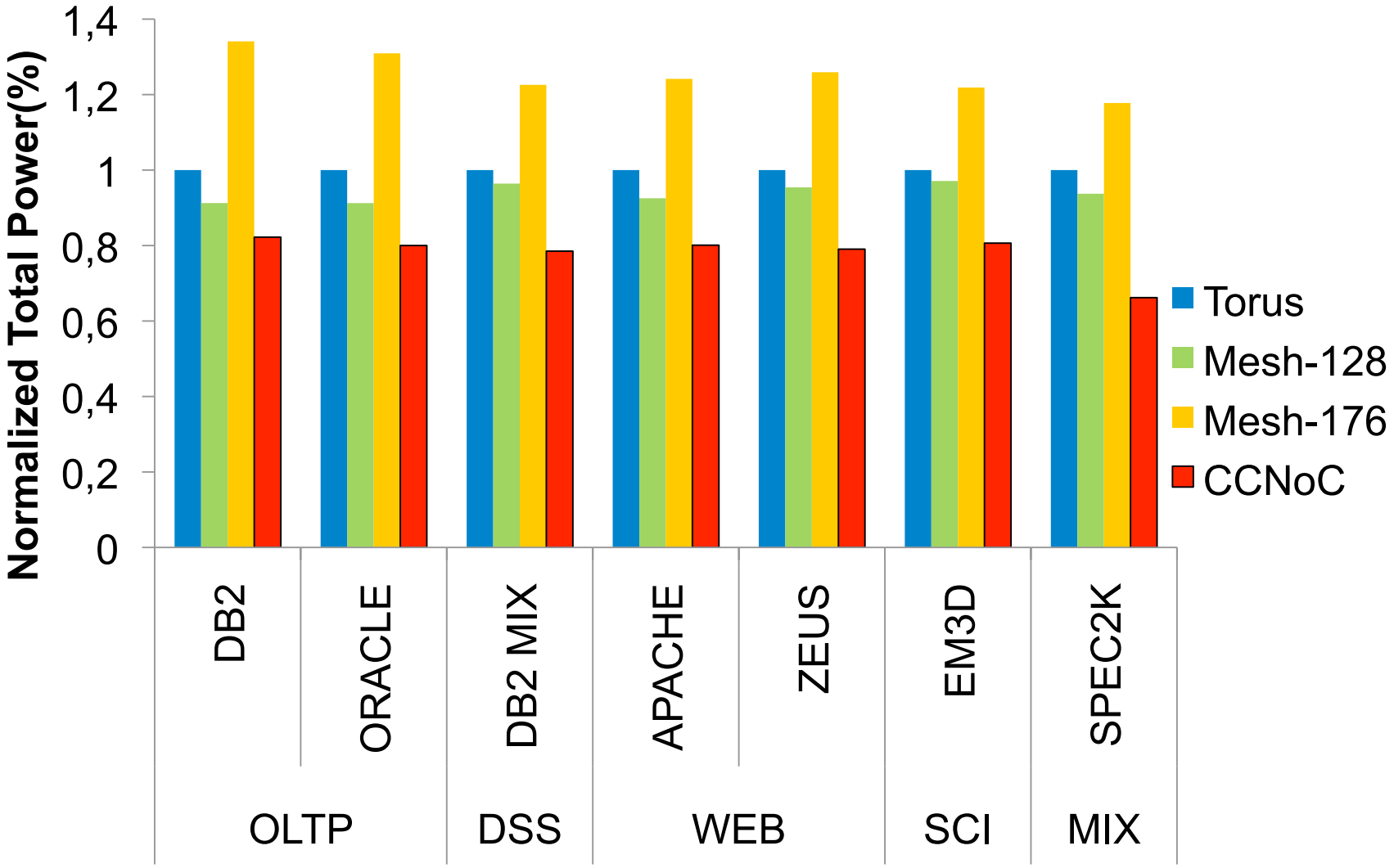
- Switches
  - Wormhole flow control
  - Input queued
  - Transmission protocol
    - On/Off
  - Input buffers
    - 2 entry

Performance



Performance loss: 2% Torus, 8% Mesh-176

Power Savings



Power savings: 16% Mesh-128, 22% Torus, 38% Mesh-176

CCNoC Conclusions

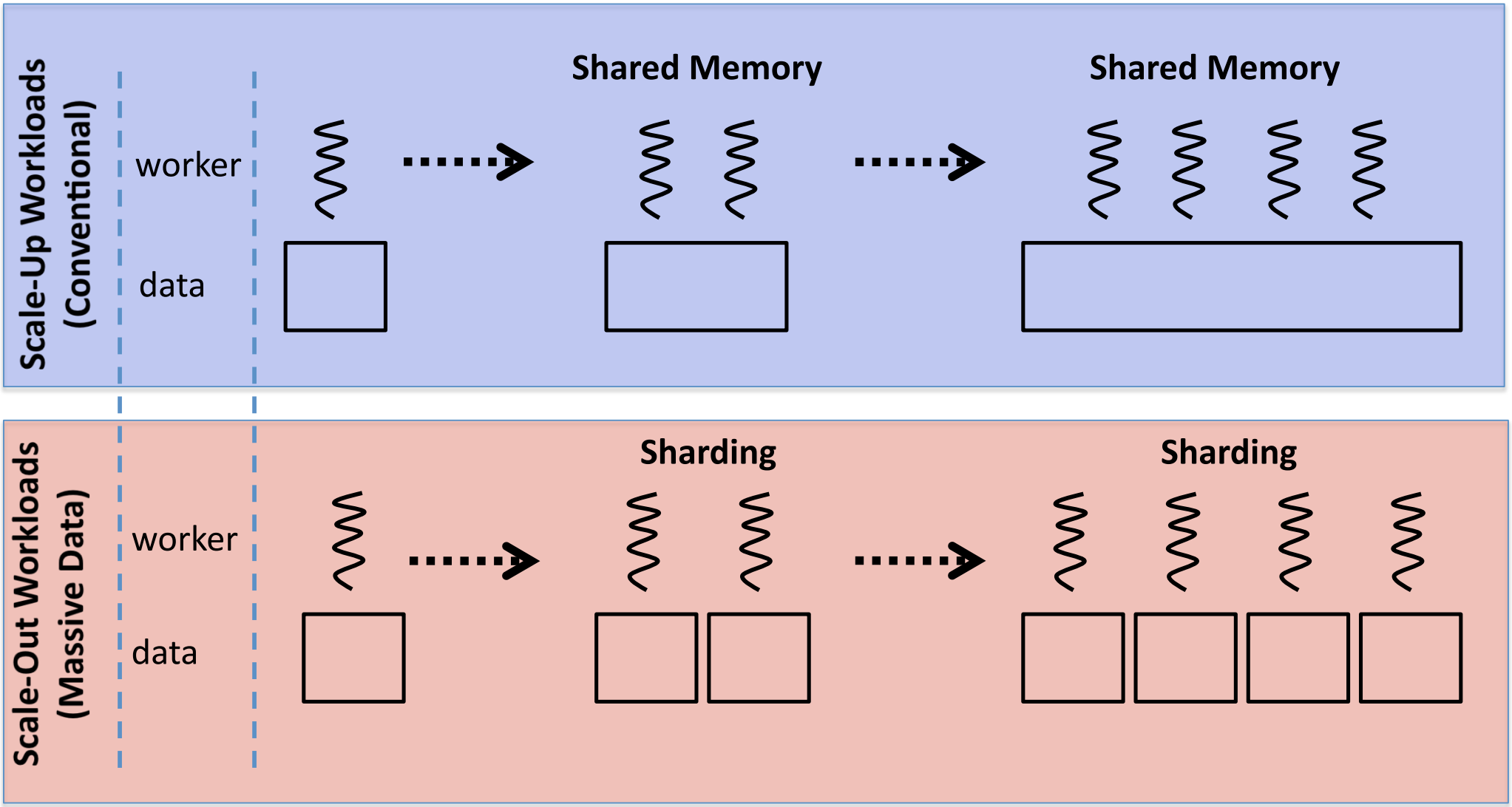
- Duality in Request/Response traffic
  - Request: dominated by short messages
  - Response: dominated by long messages
- Proposed CCNoC
  - Narrow request network
  - Wide response network
- Showed significant power savings
  - 22% against Torus
  - 38% against Mesh-176

What about parallelism?

Much focus on in the community on

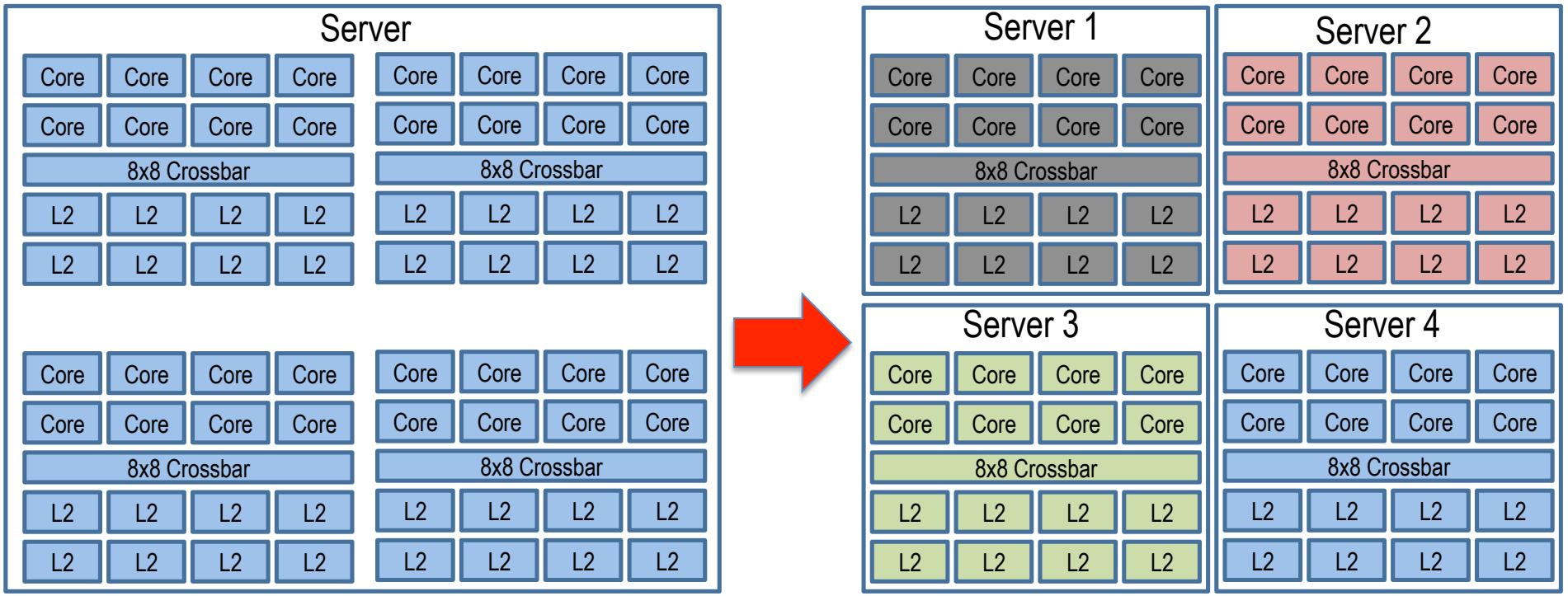
- Parallel programming
- Building monolithic “scale-up” chips
  - i.e., run one parallel app faster with more cores
- Much software out there is “scale-out”
  - i.e., parallelism across nodes/blades
  - Does not scale up!

Scale-Out vs. Scale-Up Workloads



Emerging workloads scale out!

Scale-Out vs. Scale-Up Chips



Scale-Up Chip: Conventional Shared Memory

Scale-Out Chip: Clustered Memory

- Scaling out divides chip among disconnected servers
- Hardware isolation ➔ improved reliability
- Simple hardware (no coherence/NoC)

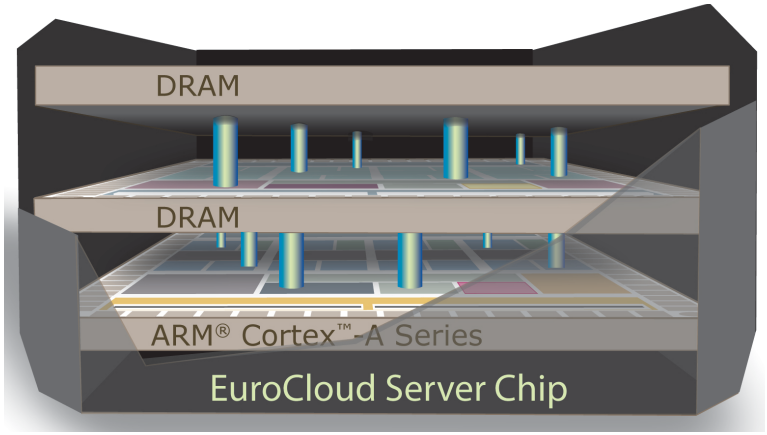
ecocloud

The EuroCloud Server: A Scale-Out Chip for Massive Data

(www.eurocloudserver.com)

3D SoC/DRAM:

- 1000x more connectivity
- 10x less system energy
- Runs off-shelf cloud stack



Your Future 1-Watt Datacenter Chip

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

ARM

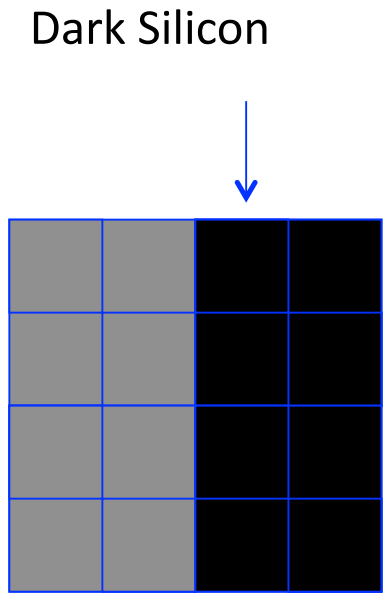
NOKIA Connecting People

imec

UCyprus

What do we do with Dark Silicon? (work from UCSD, slides from Mike Taylor)

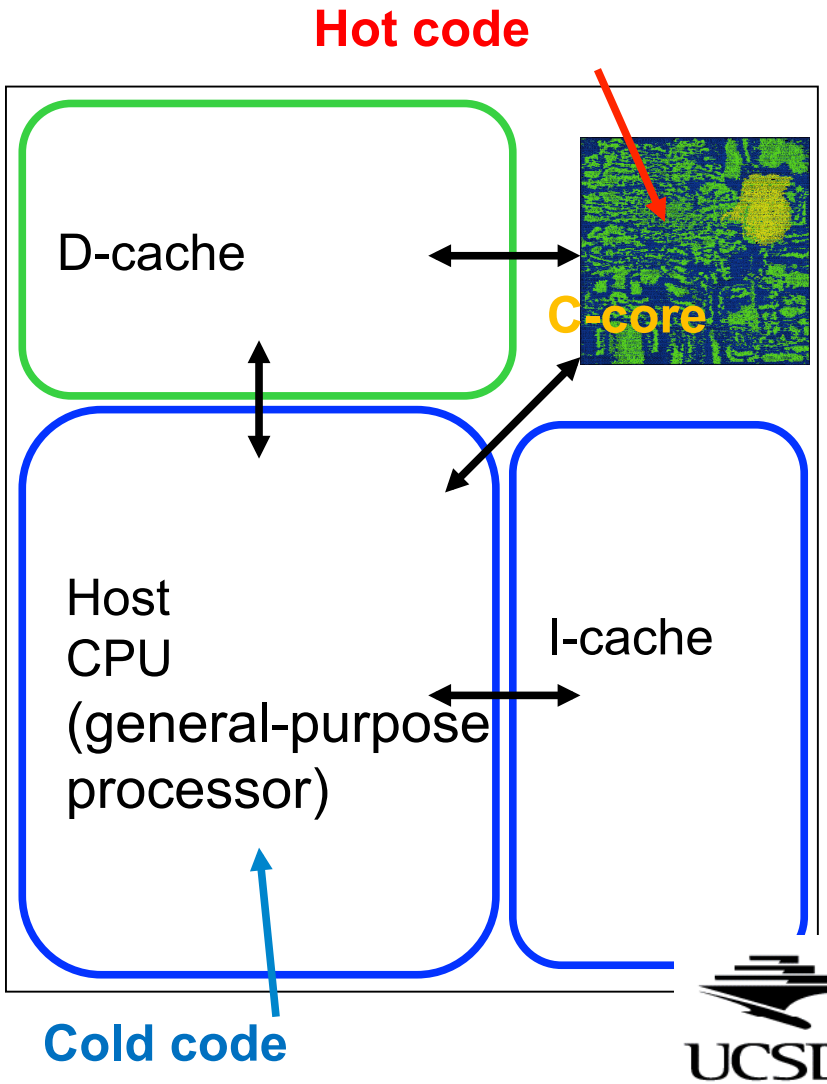
- Idea: Leverage dark silicon to "fight" the utilization wall
- Insights:
  - Power is now more expensive than area
  - Specialized logic has been shown as an effective way to improve energy efficiency (10-1000x)
- UCSD Approach:
  - Fill dark silicon with specialized energy-saving coprocessors that save energy on common apps
  - Only turn on the cores as you need them
  - Power savings can be applied to other programs, increasing throughput
- *Energy saving coprocessors provide an architectural way to trade area for an effective increase in power budget!*



Conservation Cores (C-cores)

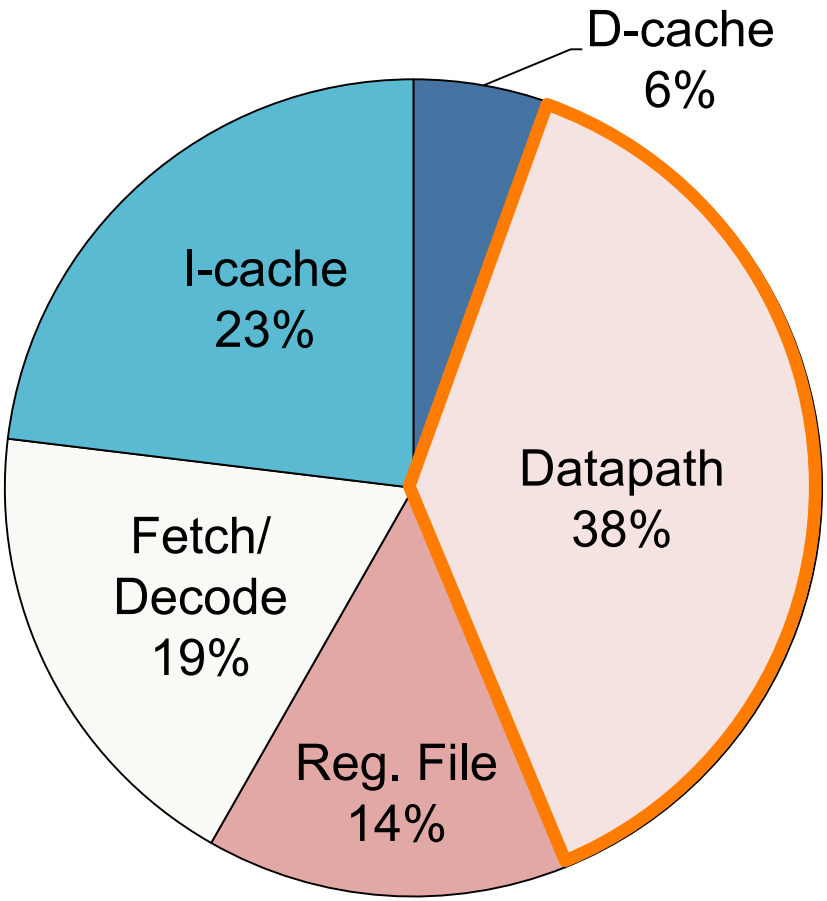
"Conservation Cores: Reducing the Energy of Mature Computations," Venkatesh et al., ASPLOS '10

- Specialized coprocessors for reducing energy in irregular code
  - Hot code implemented by c-cores, cold code runs on host CPU;
  - Shared D-cache
  - ***Patching support in hardware***
- Fully-automated toolchain
  - **No "deep" analysis or transformations required**
  - C-cores automatically generated from hot program regions
  - Drop-in replacements for code
  - Design-time scalable
- Energy-efficient
  - Up to 18x for targeted hot code

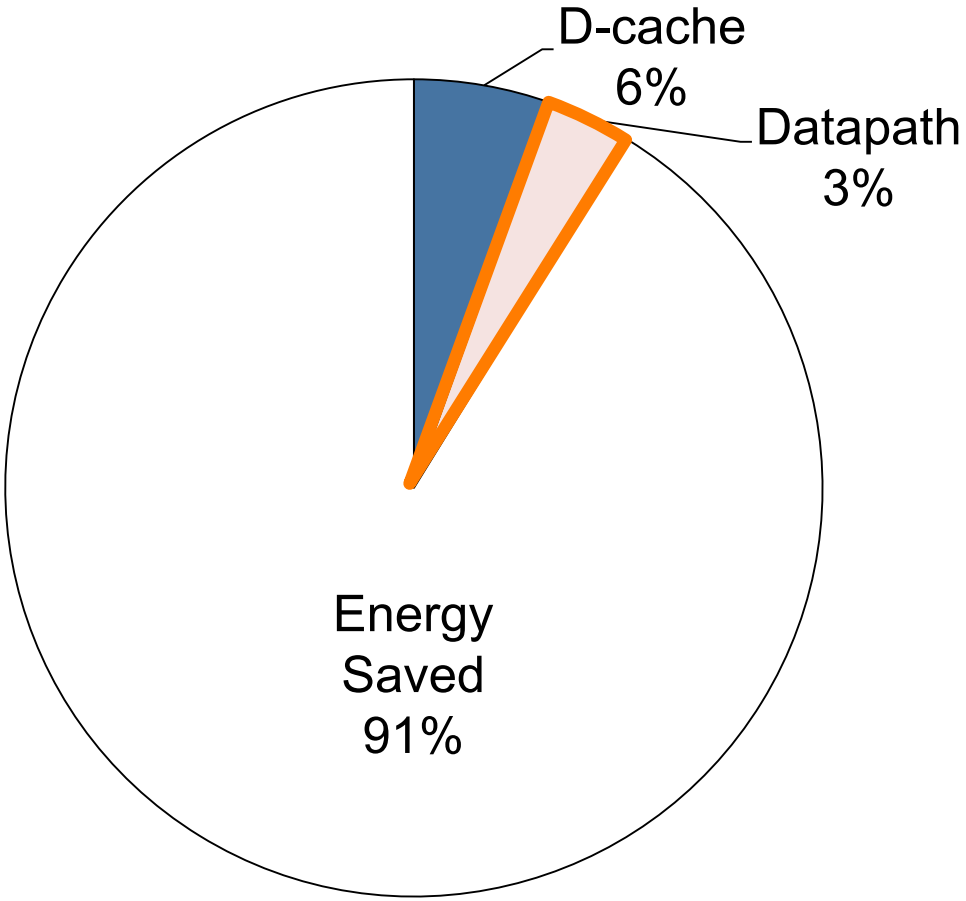


UCSD

Where do energy savings come from?



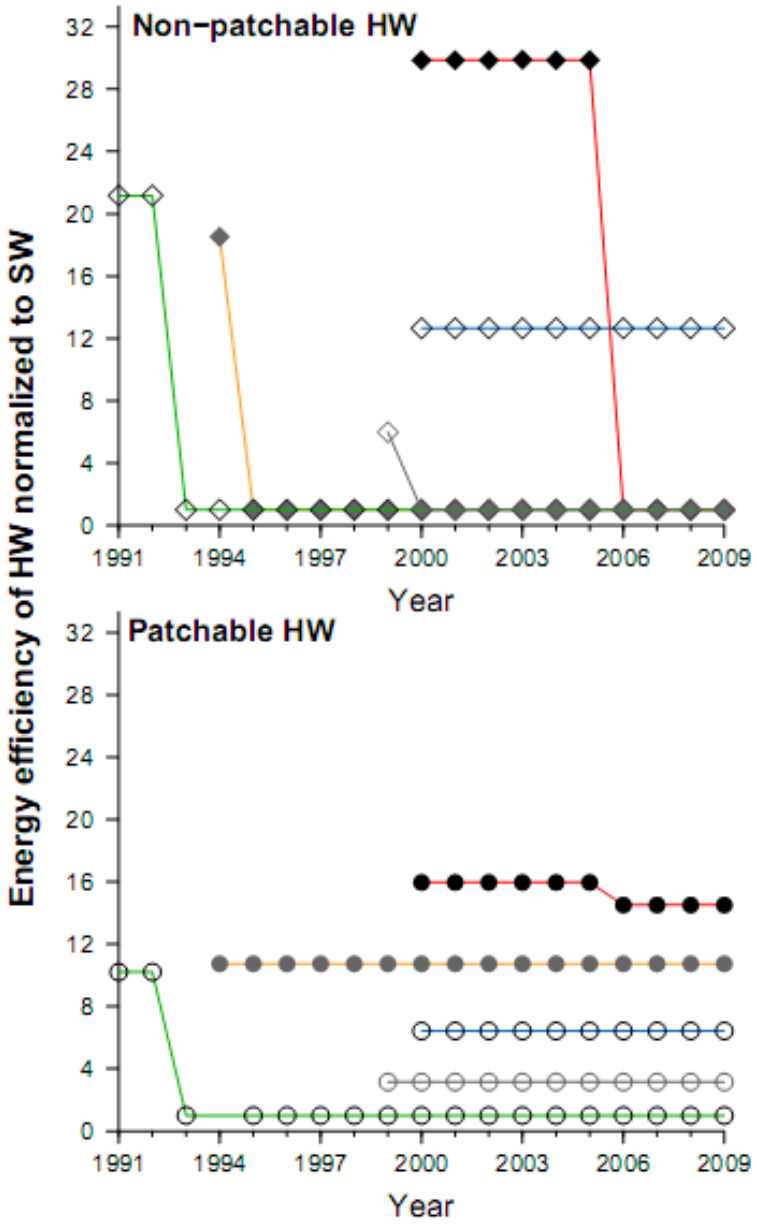
RISC baseline
91 pJ/instr.



C-cores
8 pJ/instr.

Patchability Payoff: Longevity

- Graceful degradation
  - Lower initial efficiency
  - Much longer useful lifetime
- Increased viability
  - With patching, utility lasts ~10 years for 4 out of 5 applications
  - Decreases risks of specialization

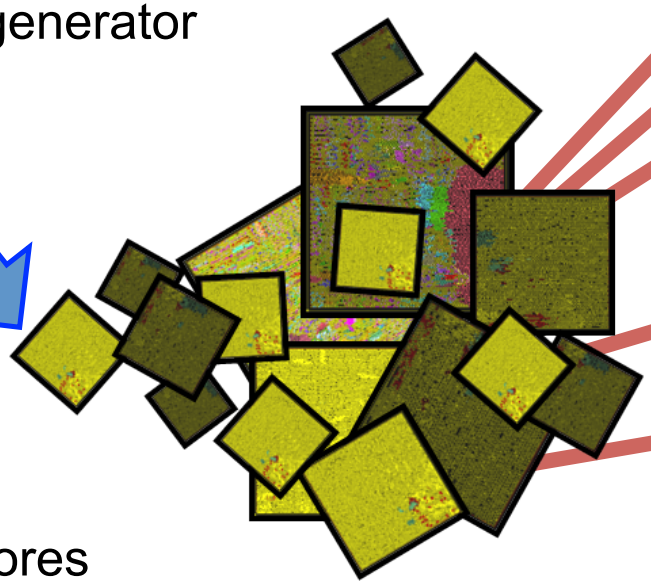


GreenDroid: Using c-cores to reduce energy in mobile processors

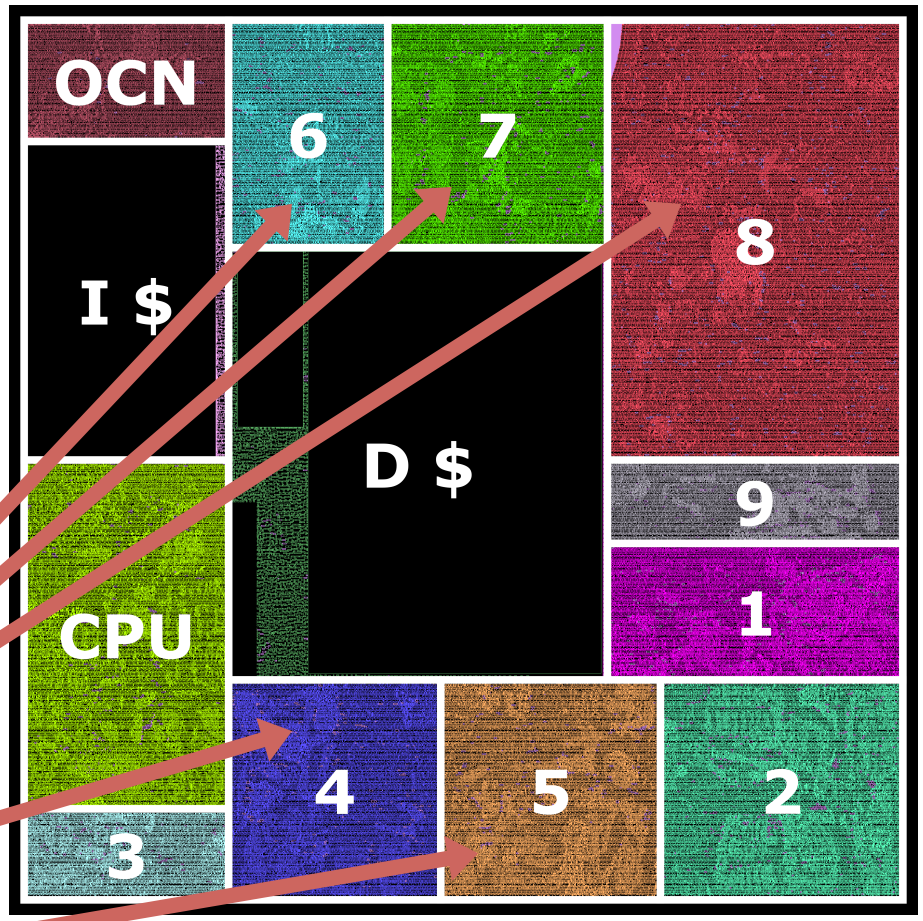
"The GreenDroid Mobile Application Processor: An Architecture for Silicon's Dark Future,"
Goulding-Hotta et al., IEEE Micro Mar./Apr. 2011

Android workload

Automatic c-core generator



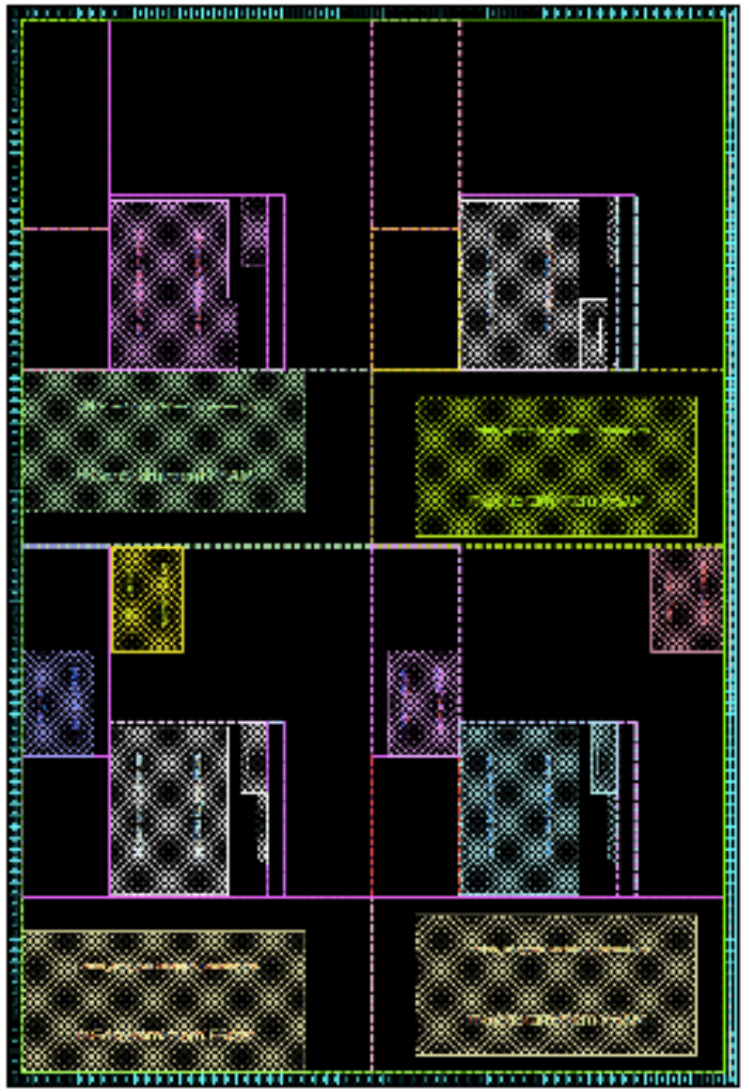
C-cores



Placed-and-routed chip with 9 Android c-cores

UCSD Quad-core GreenDroid Chip

- Scale C-cores via Tiling
- 4 Tiles; each Tile contains
  - 6-10 Android c-cores (~50 total)
  - 32 KB D-cache (shared with CPU)
  - RISC core
  - On-chip network router
- 28-nm Global Foundries
- 1.5-2 GHz
- 2 mm^2



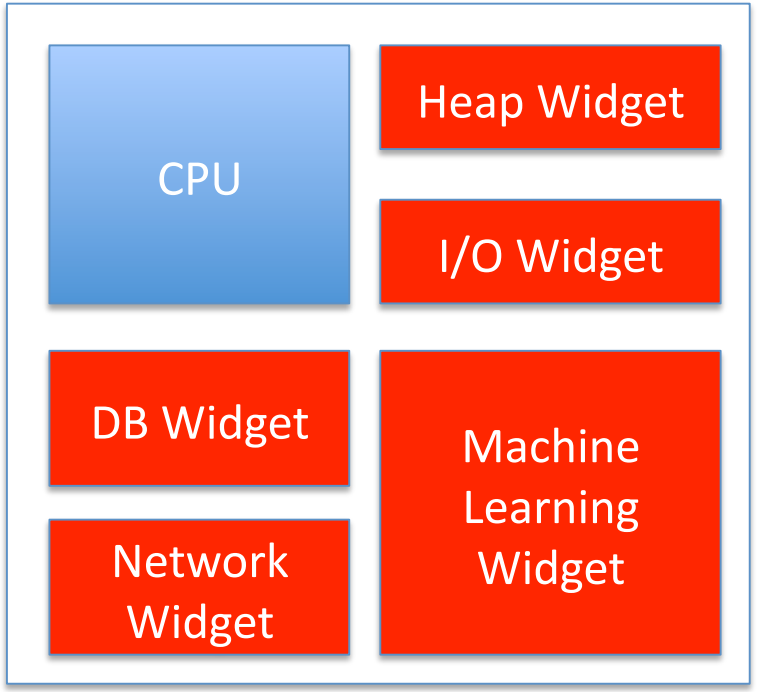
VISA: Vertically-Integrated Servers (parsa.epfl.ch/visa)

Identify energy hogs:

- Services w/ well-defined API
- Specialize
- E.g., Intel's TCP/IP CPU

Power up services on-the-fly

- Exploit higher chip densities
- With design for dark silicon



VISA System-On-Chip

Exact vs. Probabilistic

Much computation is error-resilient:

- Machine learning/analytics
- Image processing/visual computation
- Audio/speech
- Search

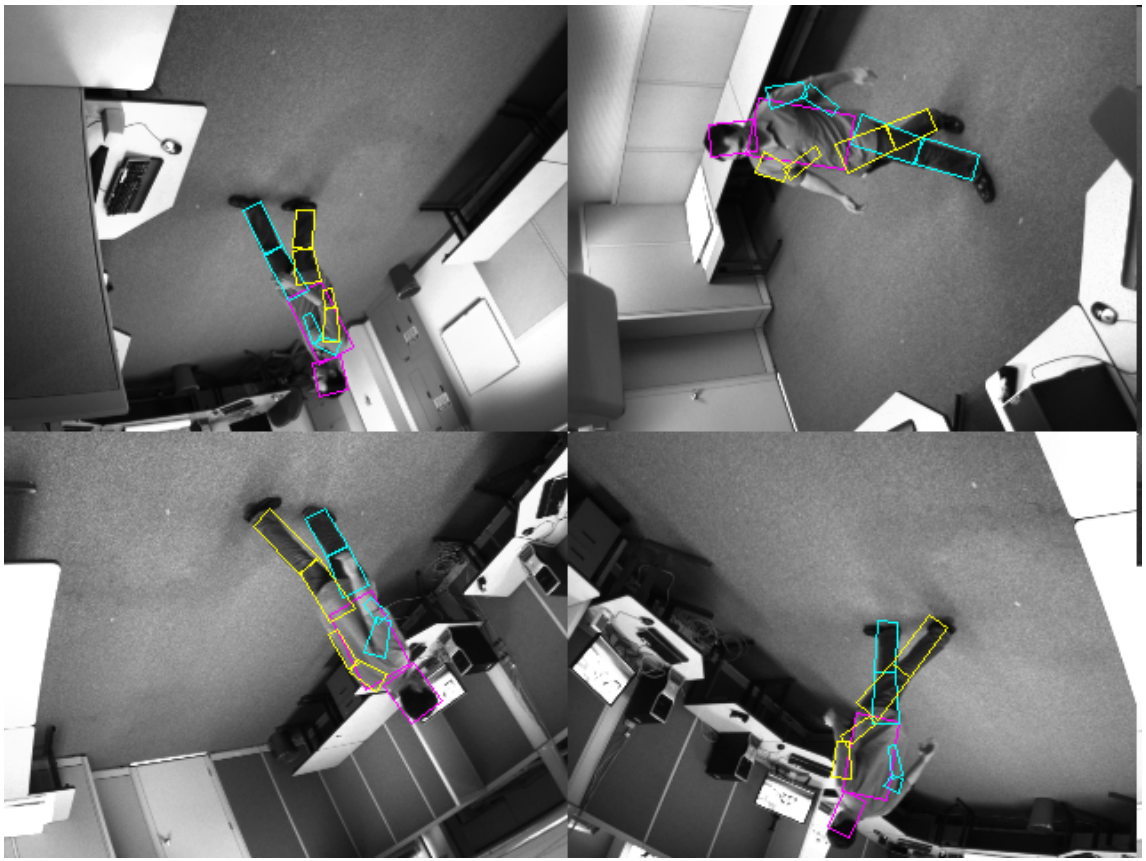
Similarly, two flavors of data

- Exact: affects functionality (pointer address)
- Probabilistic: affects quality (pixels in image)

Perforated (Skipped) Computation

bodytrack benchmark (PARSEC)

- Compiler-driven perforation
- Skip 40% of computation
- Maintains track on head, chest and legs



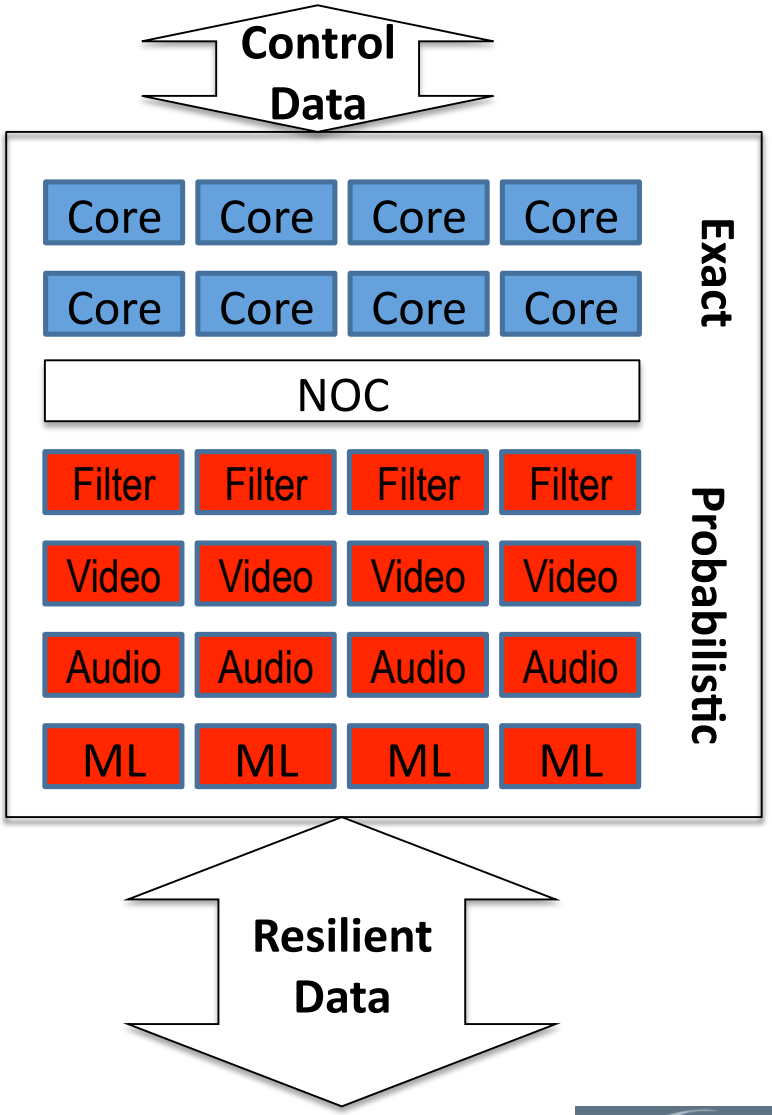
Computation does not have to be exact!

Hoffman et. al., "Using Loop Perforation to Dynamically Adapt Application Behavior to Meet Real-Time Deadlines", 2010

DeSyRe: Probabilistic Computing

Exploit resilience in computation & data

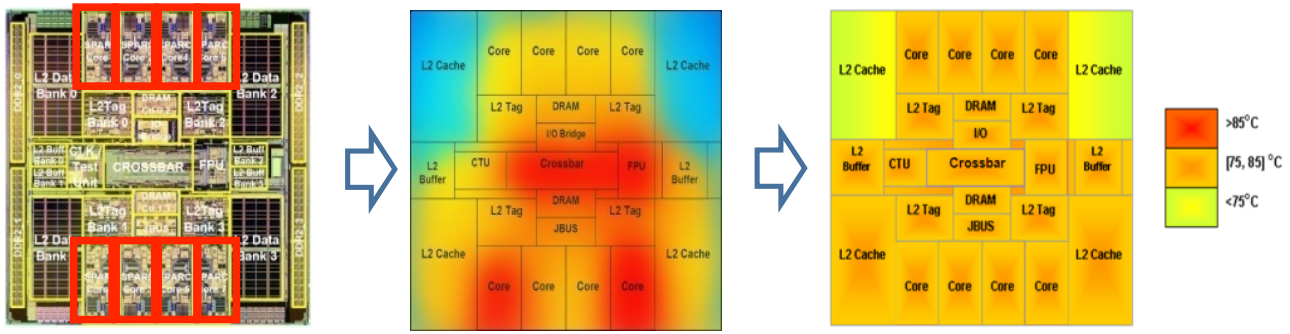
- Reliable + unreliable substrate
- Partition tasks according to resilience
- Maximize throughput with less energy



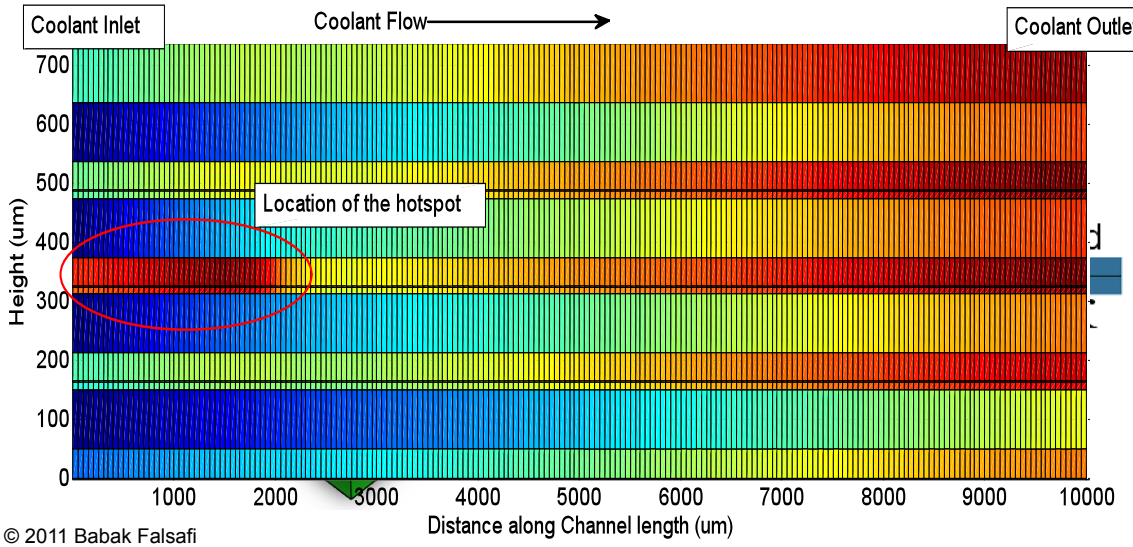
CHALMERS

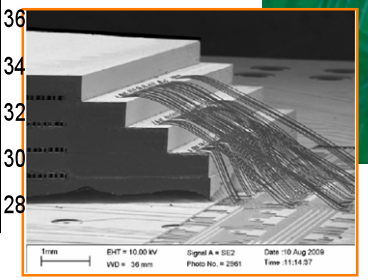
Holistic Integration (Atienza & Thome @ EPFL)

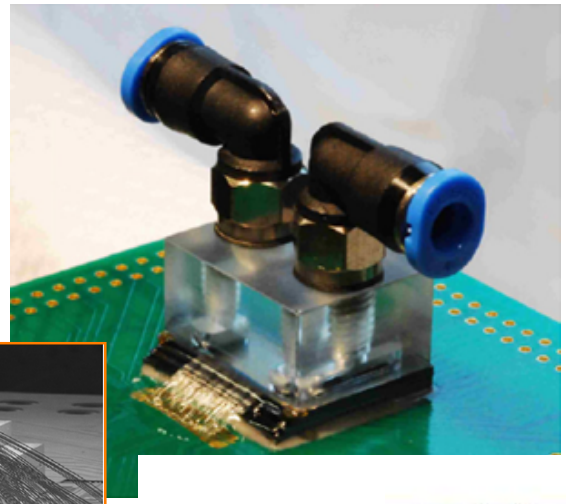
Heat flow prediction in 3D Niagara



Thermal modeling with active cooling







ORACLE IBM

ecocloud

Research Center @ EPFL

Dozen faculty, CSEM
& industrial affiliates

- HP, Intel, IBM, Microsoft, Nokia, Oracle, Credit Suisse, Swisscom,…
- Over several million CHF of annual funding
- Datacenter Observatory (test bed)

Research:

- Energy-minimal cloud computing
- Elastic data bricks and storage
- Scalable cloud applications & services

Making tomorrow's clouds green & sustainable

Bringing it All Together

- IT is changing everything & itself changing
- Future: Plow massive data with minimal energy
- Our way of computing is inefficient & too robust
- Rethink building servers in the Dark Silicon Era

Oh, I can be reliable and for an extra 300 hundred a week can be efficient too.

Thank You!

For more information please visit us at

ecocloud.ch