Towards Dark Silicon in Servers

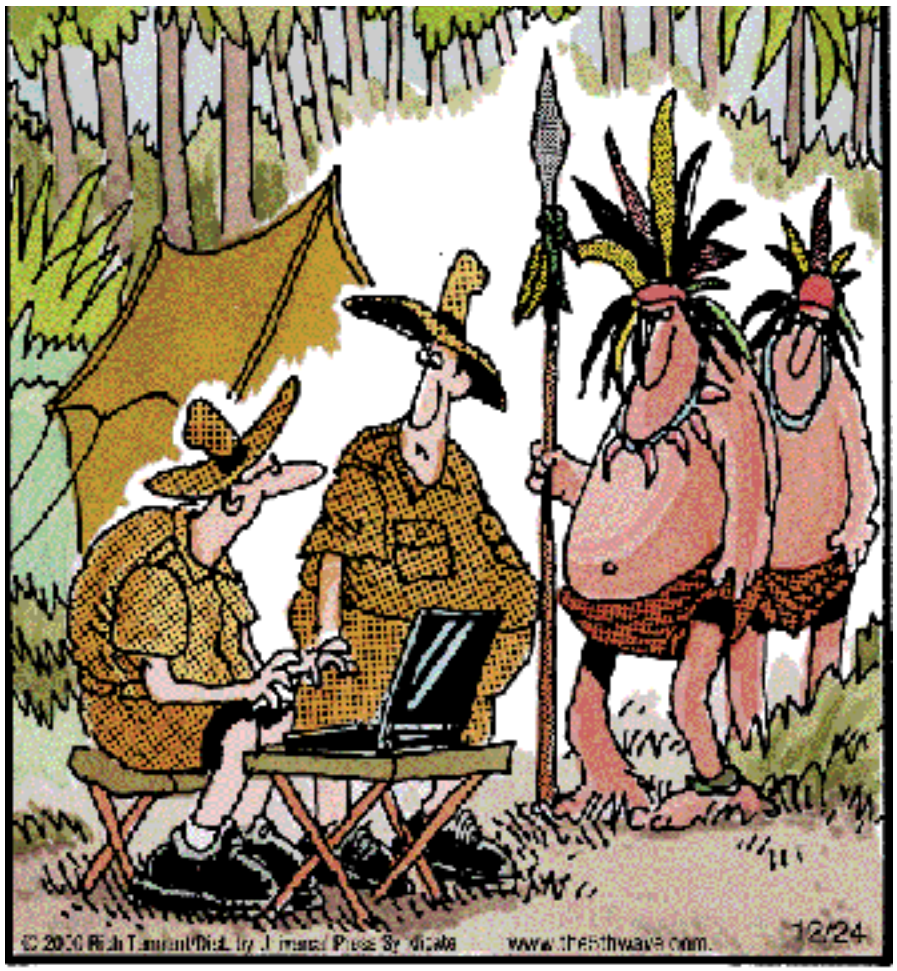
Babak & the team

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

IT is ever more **indispensable**

Our life w/o digital data
is unimaginable as

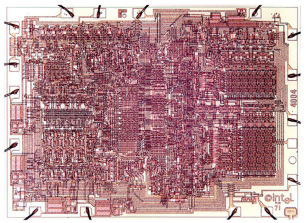
- Individuals
- Governments
- Enterprises
- Research organizations
- Societies



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

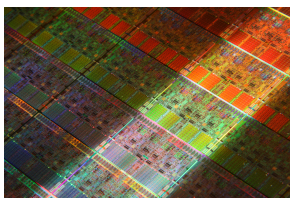
How did we get here? Moore's Law

Intel 4004, 1971

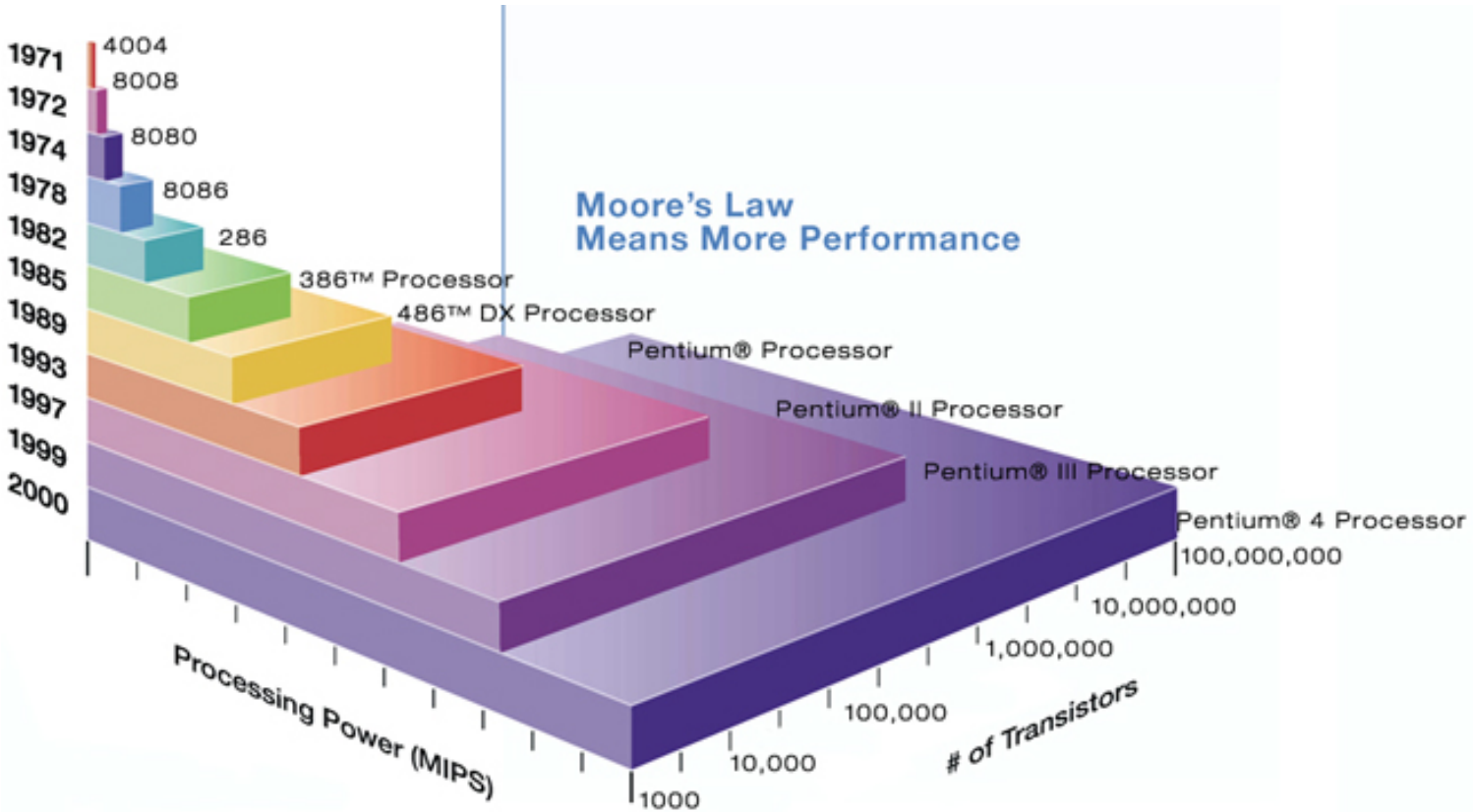


92,000 ops/second

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 computing-centric to data-centric
- Consumer Era: interfacing, connectivity and access

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!

Two Inflection Points Colliding

1. Emergence of Data-Centric Universe
   - IT focus on massive data

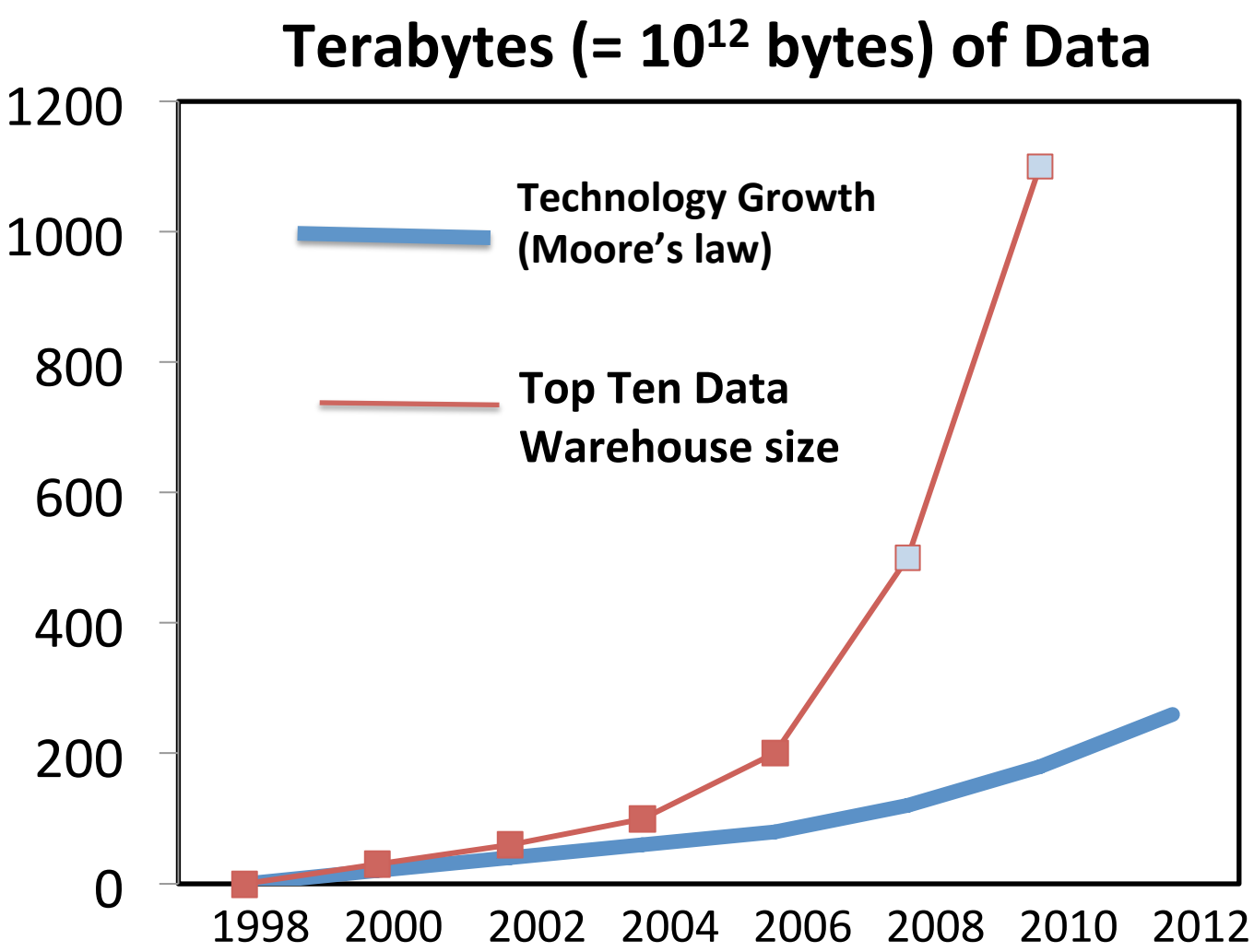
2. End of Dennard Scaling
   - Higher density ➔ higher energy

- Data-Centric Universe meets Energy Wall

What are design implications?

Our Data-Centric Universe: Data Growing faster than Technology

- Companies spending $$$ to collect/analyze data
- Commerce entirely data-driven
- Scientific discovery via massive data
- Personalized computing



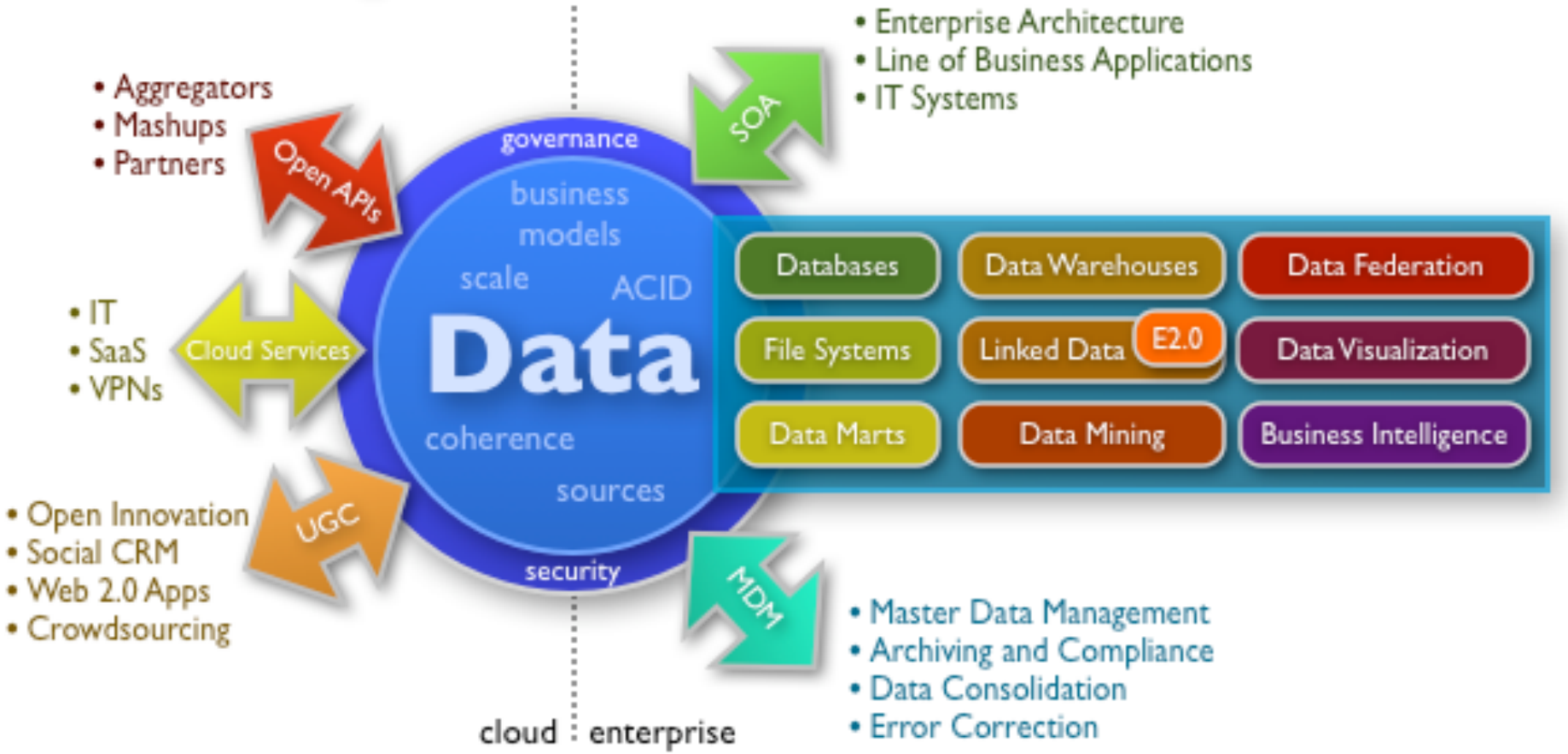
WinterCorp Survey, www.wintercorp.com

Data Deluge: 1.8 Zettabytes in 2011

(source: IDC, picture: Economist)

- Was only 150 Exabytes in 2005
- Supply-chain management, 10x increase in data in a year
- US aerial surveillance models 30x more data in 2011

Anatomy of a Data-Centric Business

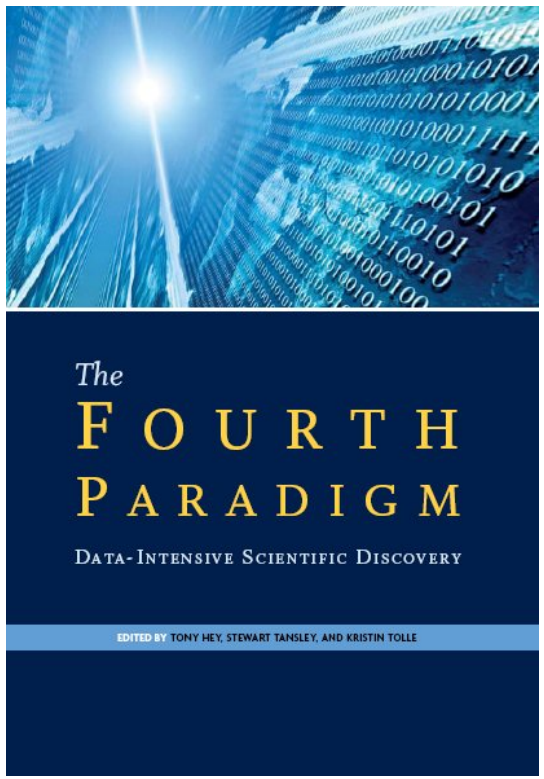


- Era of “knowledge economy”
- 50% of economic value in developed countries
- Dominant supply-chain component of products/services

Data-Centric Science: "The Fourth Paradigm"

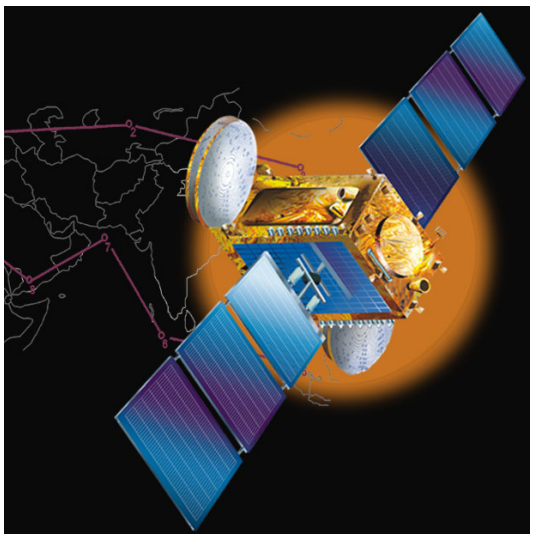
Mining data from:

- Archives
- Humans
- Sensors/instruments
- Simulations



Unifying theory, experimentation, simulation, analytics on massive data

Data Comes in Various Flavors



Satellite

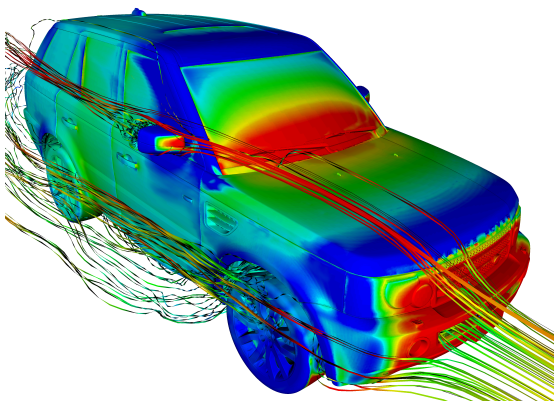
Health

Entertainment

Life

Commerce

Search



Simulation

It's all about Accessing Data!

Cloud Computing

A computing paradigm shift to enable ubiquitous connectivity

How to design for massive data

Two Inflection Points Colliding

1. Emergence of Digital Universe
   - IT focus on massive data

2. End of "Free Energy"
   - Higher density ➔ higher energy

- Data-centric Universe meets Energy Wall

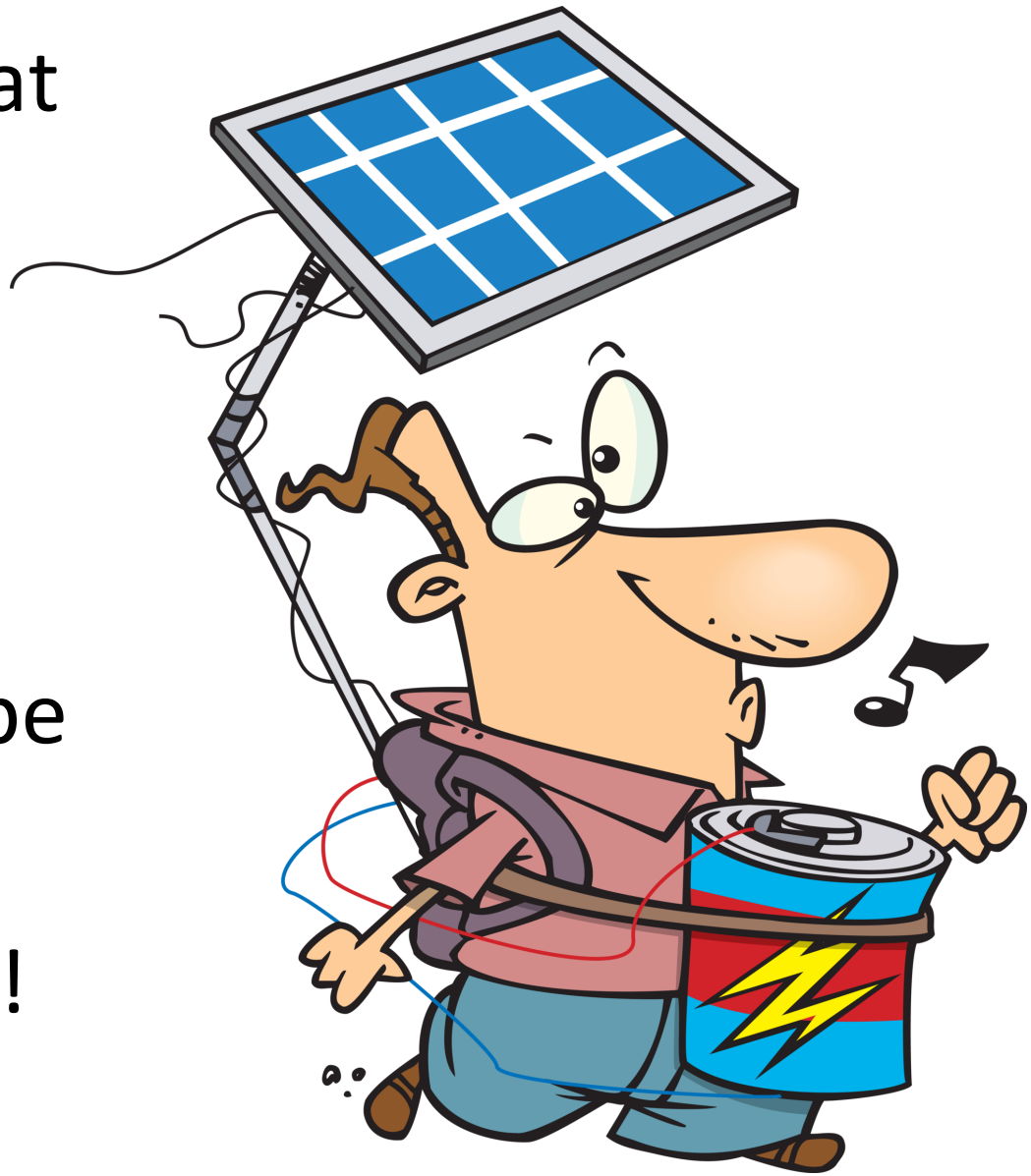
What are design implications?

IT Energy is Shooting Up!

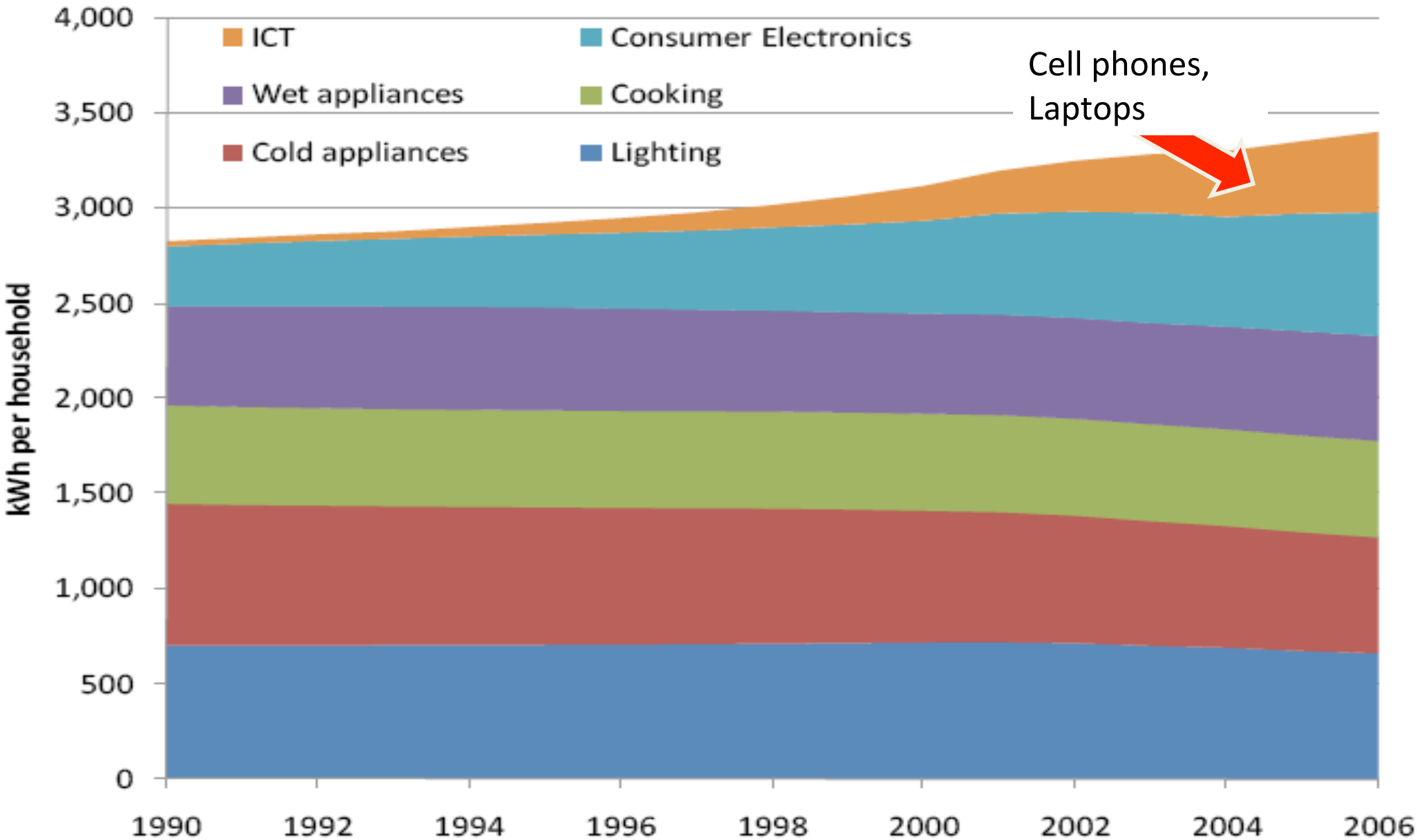
IT riding on technology that was energy-friendly

- Exponentially better performance, density
- Constant power envelope

But, energy is shooting up!

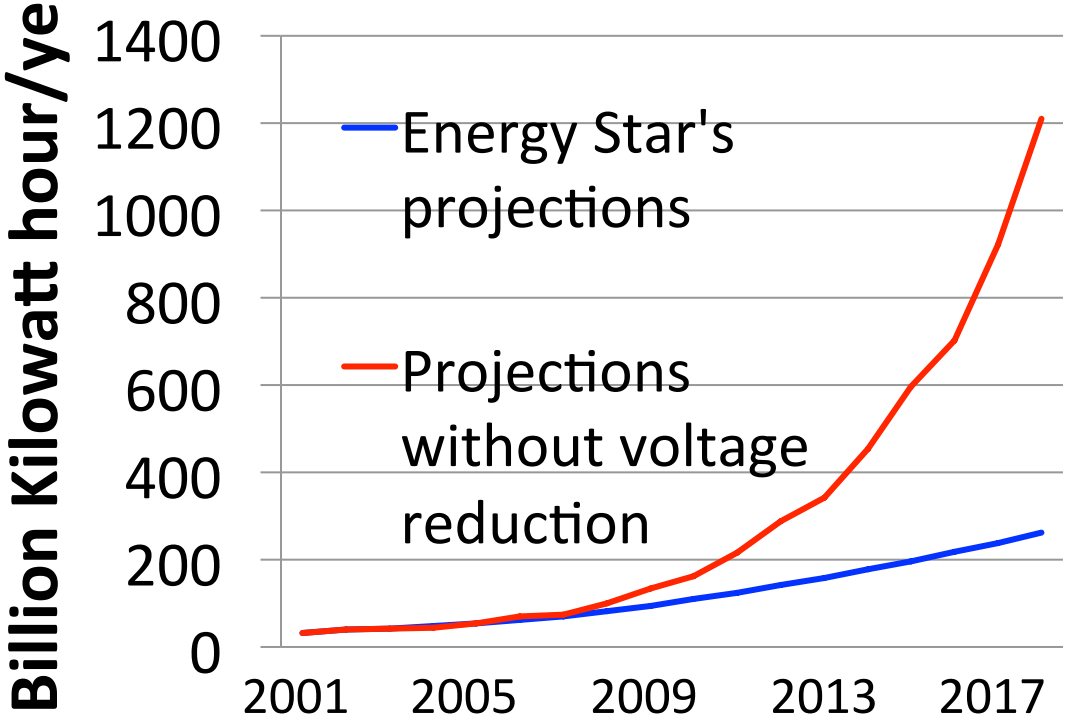


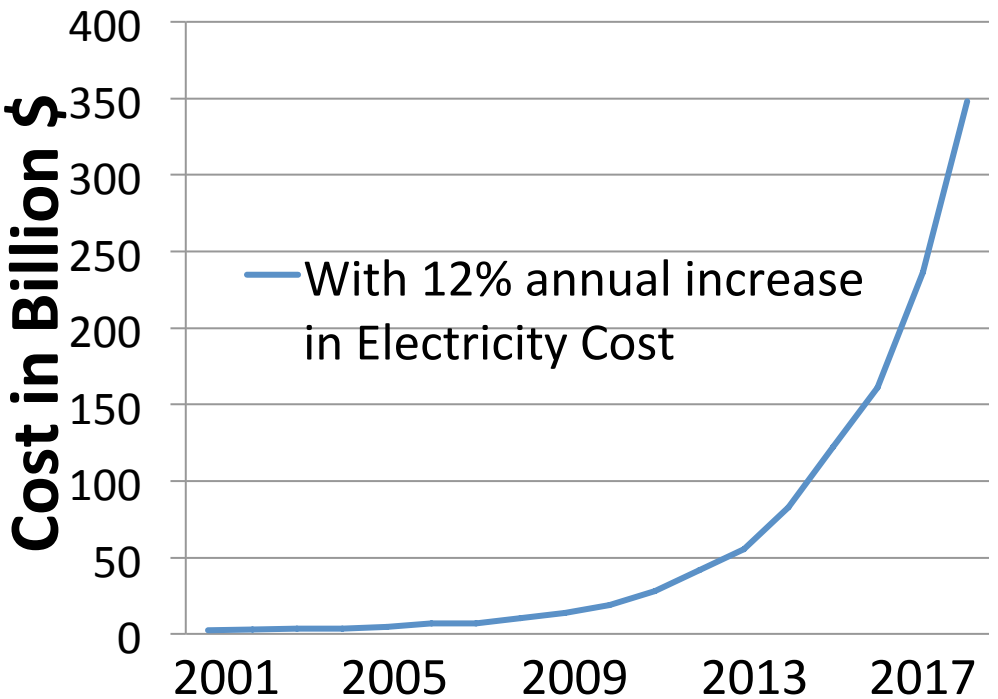
Household Energy in the UK
(UK BERR, 2008)



Data center Energy in the US

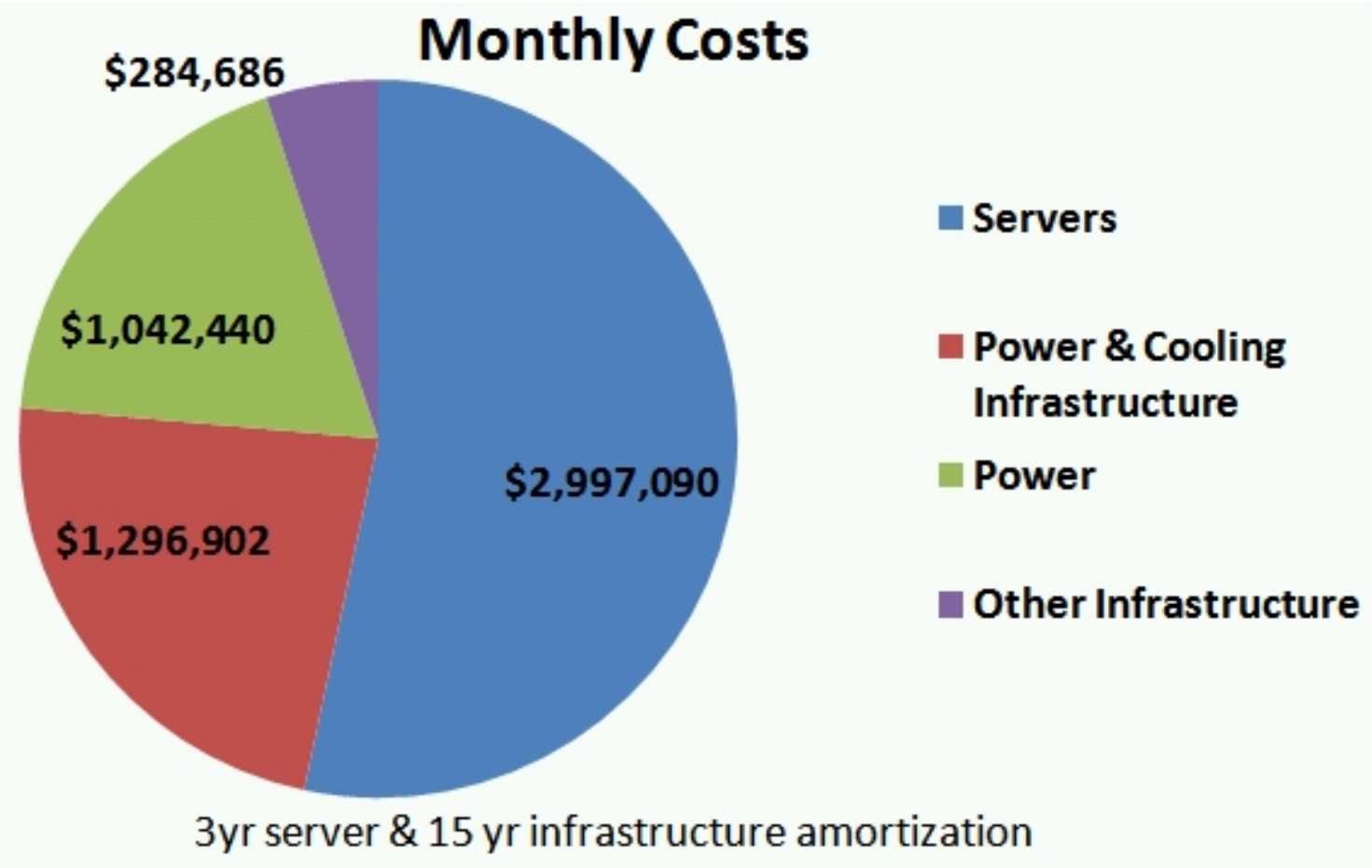
(extrapolated from Energy Star, 2007)





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

Energy ~ Capital Cost?

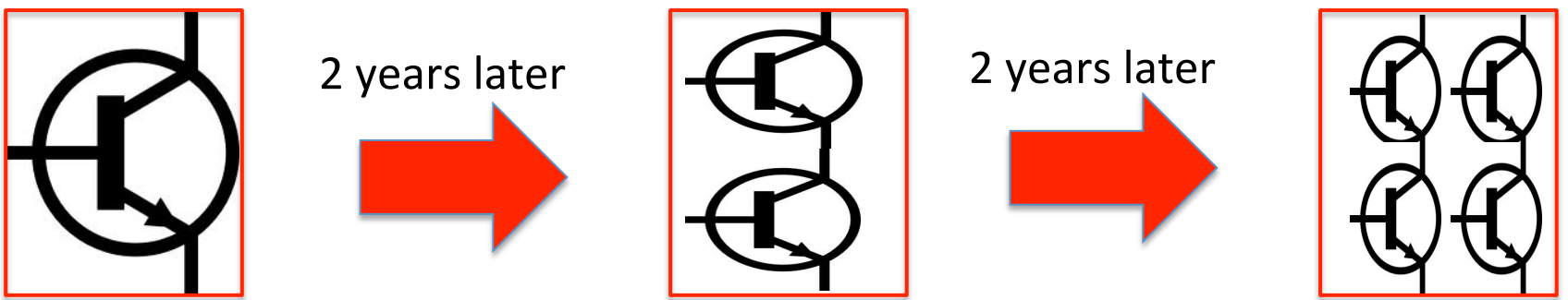


James Hamilton's Blog, mvdirona.com, 2008

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

End of “Free” Energy

1 transistor = 1x energy **2 transistors = 1x energy** **4 transistors = 1x energy**



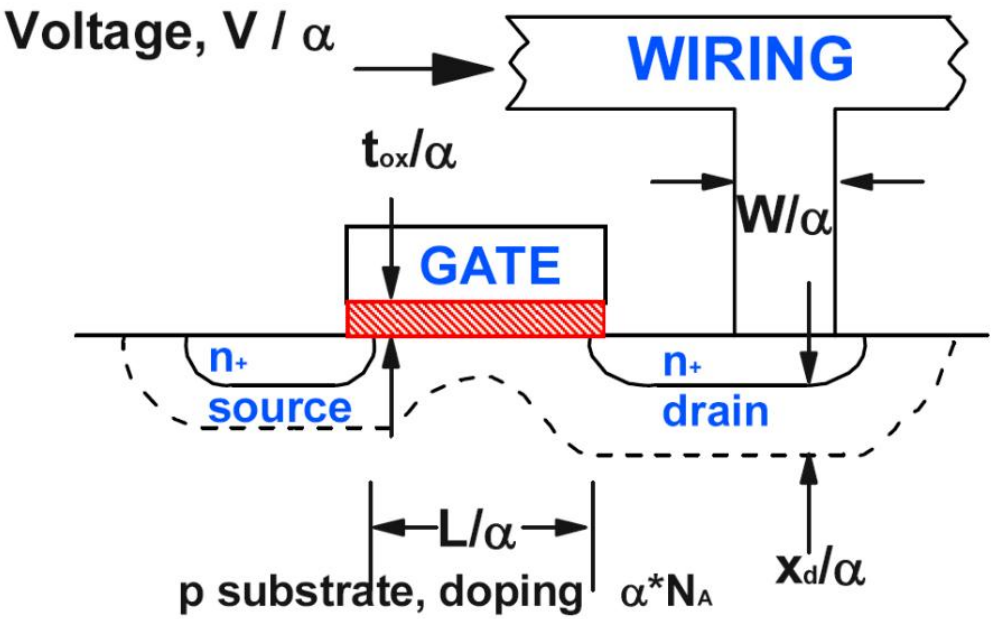
Before (1970~2000):

- Dennard scaling
- Used to make transistors smaller
- Smaller transistors less electricity to operate

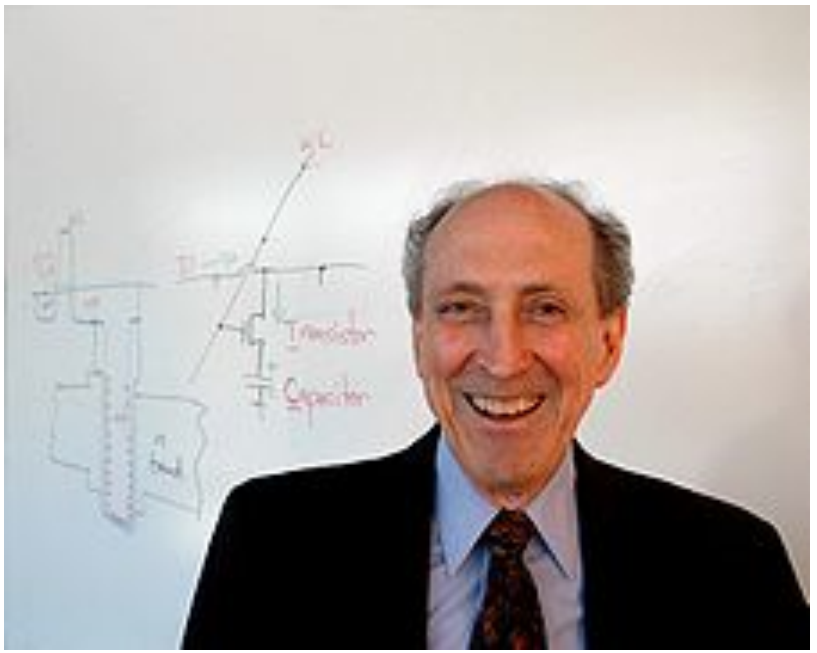
Now (2004-):

- Continue to make transistors smaller
- But, they use similar electricity to operate

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:

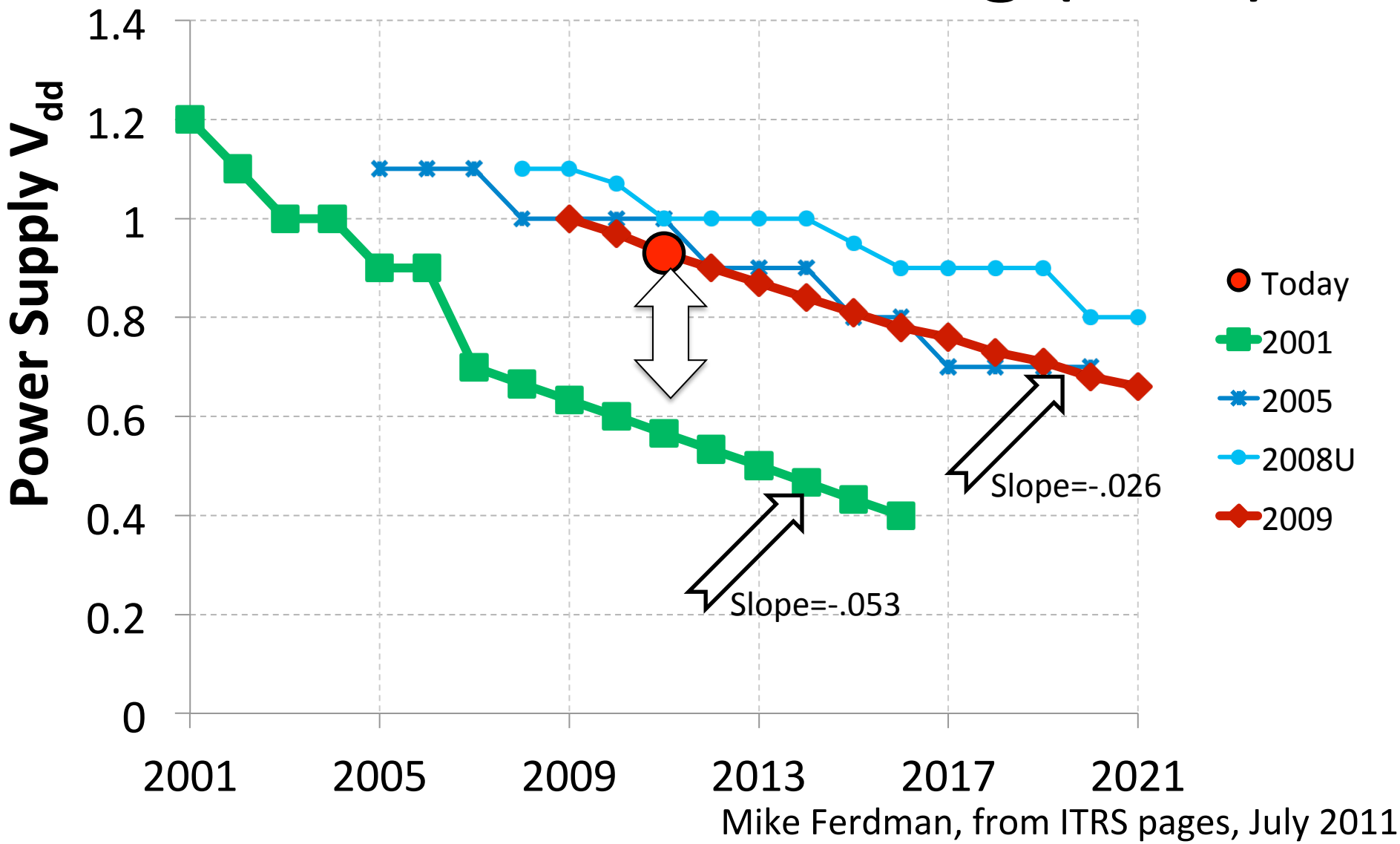
- Exponential in inverse of V_{th}
- Exponential in temperature
- Linear in device count

To switch well

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

➜V_{dd} can't go down

End of Dennard Scaling (ITRS)



Mike Ferdman, from ITRS pages, July 2011

Supply voltages going down at much lower rate!

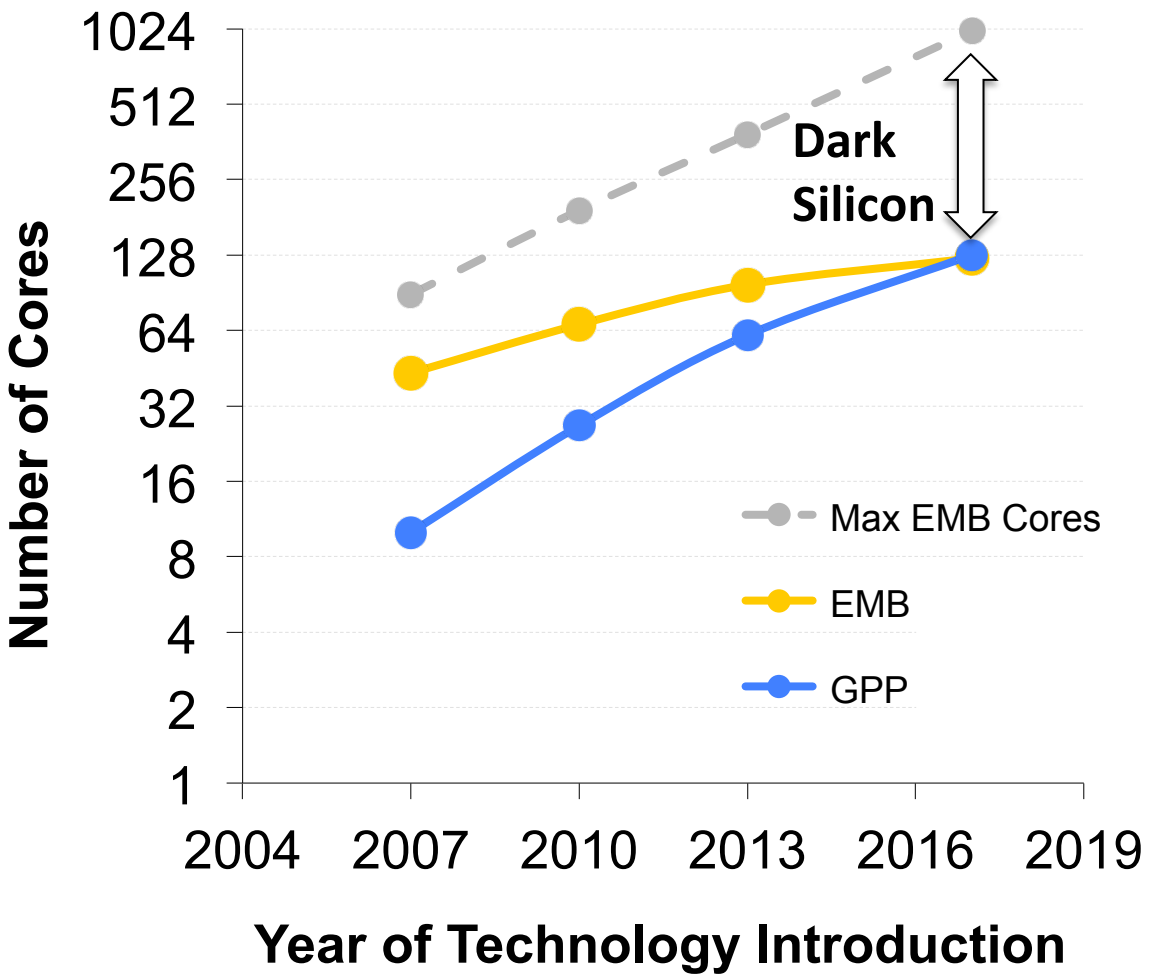
Dark Silicon: End of Multicore Scaling

Can not power up chip for fully parallel SW

Parallelism has limits even in Servers!

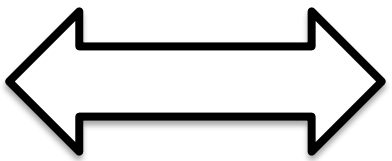
Must:

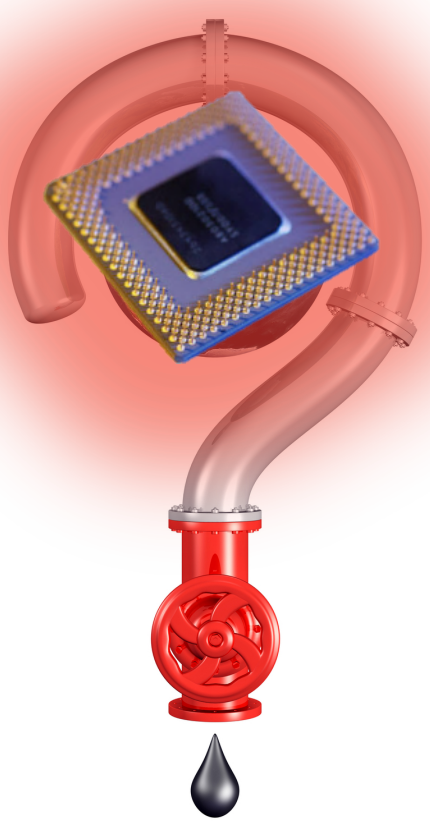
- specialize
- selectively power up



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

Massive Data meets Energy Wall





Need energy-scalable data-centric technologies

1. Short-term: integration (this talk)
2. Long-term: approximation & specialization

ecocloud

Center to bridge Data with Energy

- 16 faculty, 50 researchers
- Roughly $3M/year funding
- Datacenter Observatory (test bed)

Research:

- Technology-scalable Datacenters from programming to cooling
- Scale-out data analytics & management
- Approximation, Integration & Specialization

Microsoft

hp

intel

swisscom

CREDIT SUISSE

NOKIA
Connecting People

ORACLE

My immediate collaborators

Two Inflection Points Colliding

1. Emergence of Digital Universe
2. End of "Free Energy"

- Data-centric Universe meets Energy Wall

  ➔How efficient are today's servers?

  – Scale-Out Processors

Scale-Out Datacenters

Massive Scale

- $100+ M investment
- 5-20 MW power budget

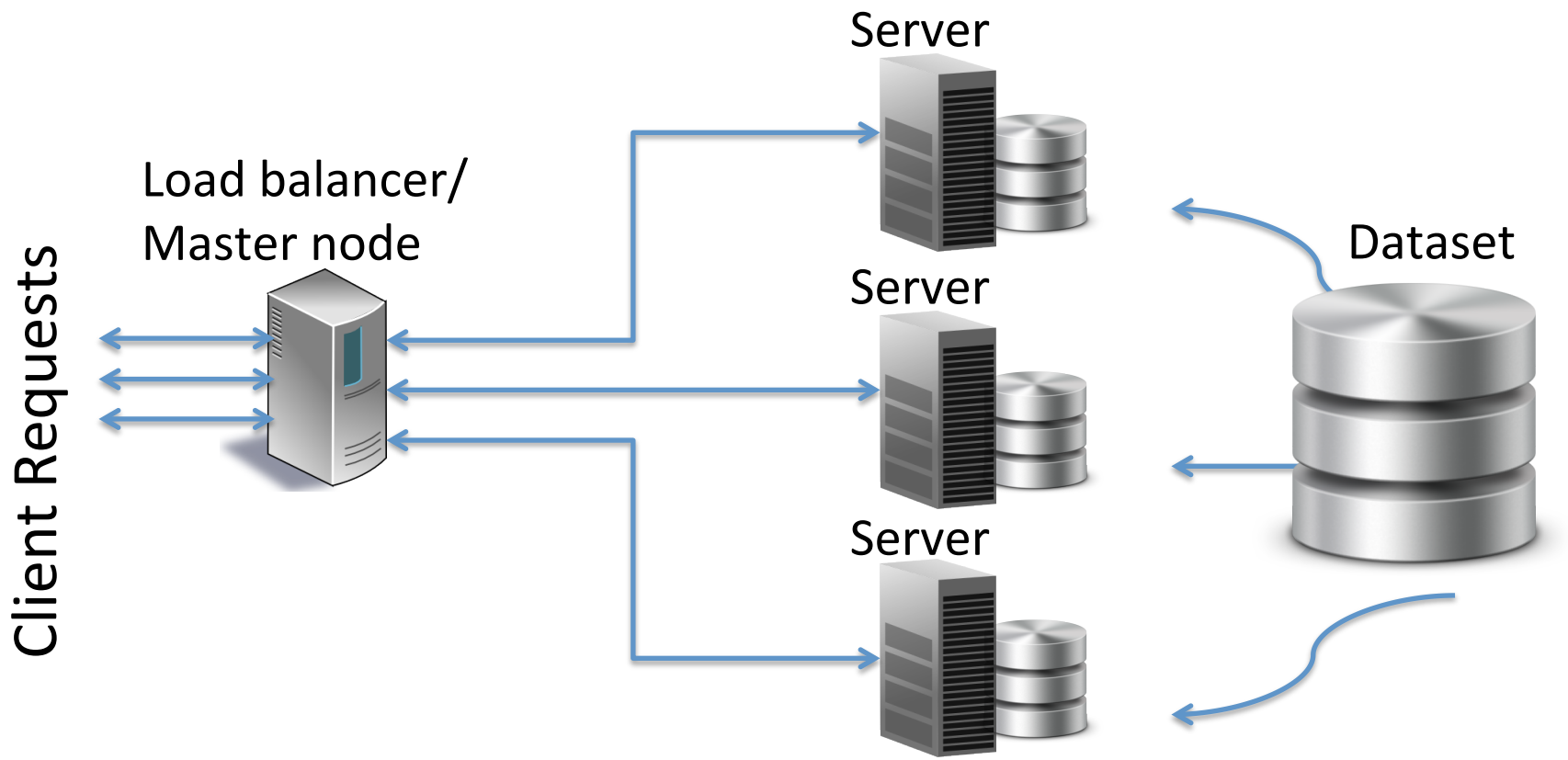
Why?

- Massive data sets, complex queries
- Global demand

Applications

- Web search, media streaming, social connectivity
- Scale-out by design

Characteristics of Scale-Out Apps

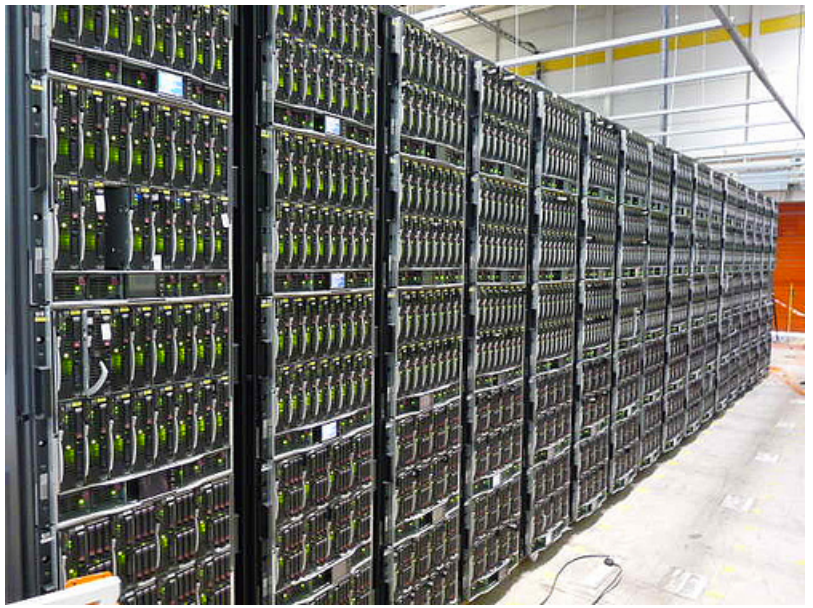


- Many independent requests/tasks
- Huge dataset split into shards
- Minimal communication among servers

Scale-Out Datacenter Internals

Populate a datacenter

with servers

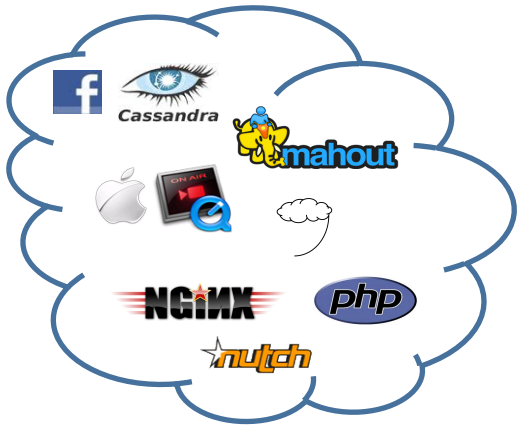


- Tens of thousands of servers
- Large memory per server
- High efficiency = high utilization

Maximize parallelism for efficiency

How Efficient are Today's Servers?

- Created benchmark suite
  - Diverse set of cloud workloads
  - Quantified high-level behavior



- Studied off-the-shelf hardware
  - Used performance counters
  - Identified needs of cloud apps

Modern CPUs don't match needs of cloud apps

Cloud Suite 1.0

(released @ parsa.epfl.ch/cloudsuite, tutorial @ ISCA 2012)

Linux 2.6.32

Data Serving

Cassandra NoSQL

MapReduce

Machine learning on Hadoop

Media Streaming

Apple Quicktime Server

Cloud 9

Symbolic VM constraint solver

Web Frontend

Nginx, PHP server

Web Search

Apache Nutch

Hardware

Dell PowerEdge
M1000e

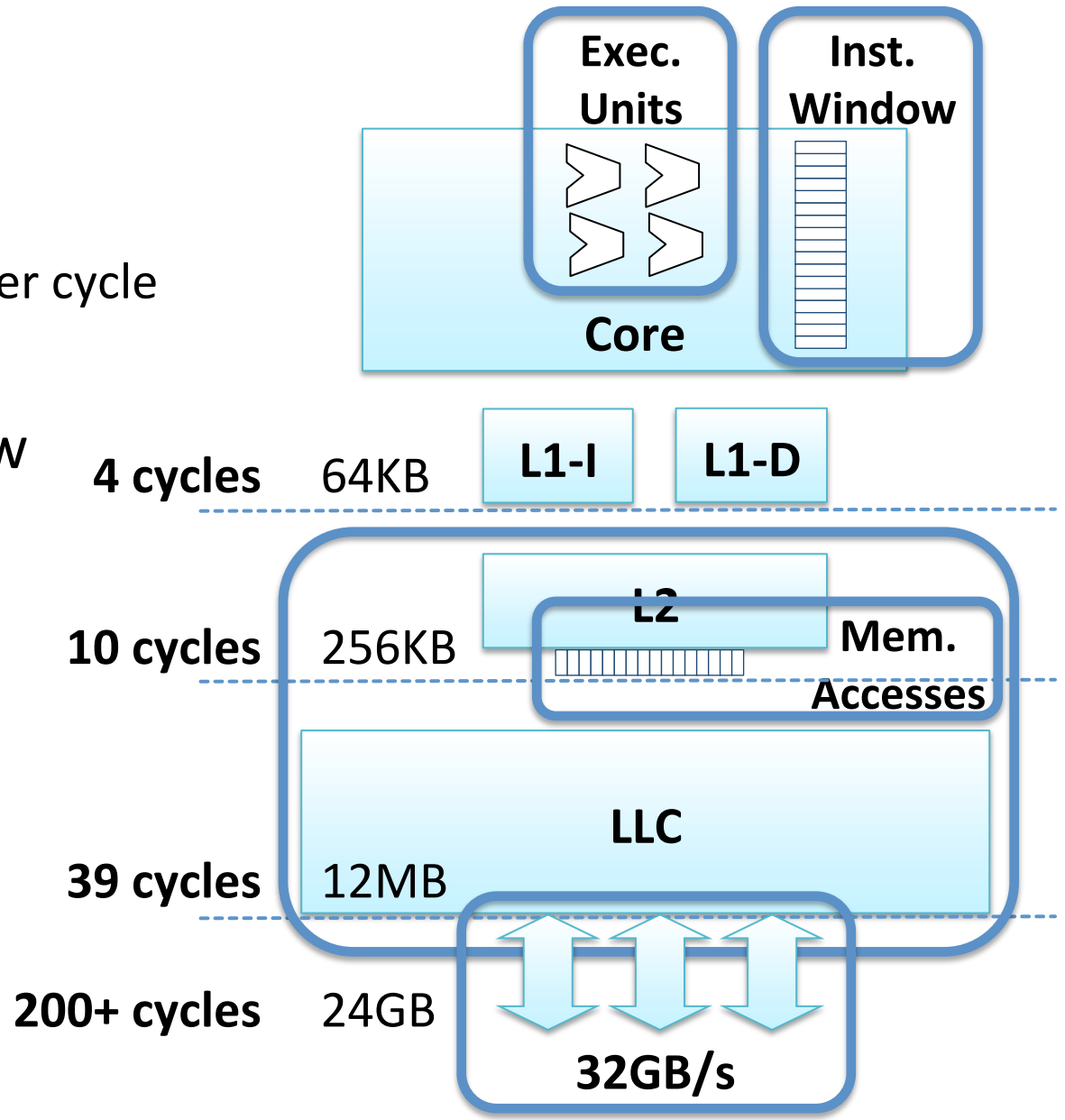
Dell Blades
M610

Two Intel x5670
2.9GHz
6-core, 12MB LLC

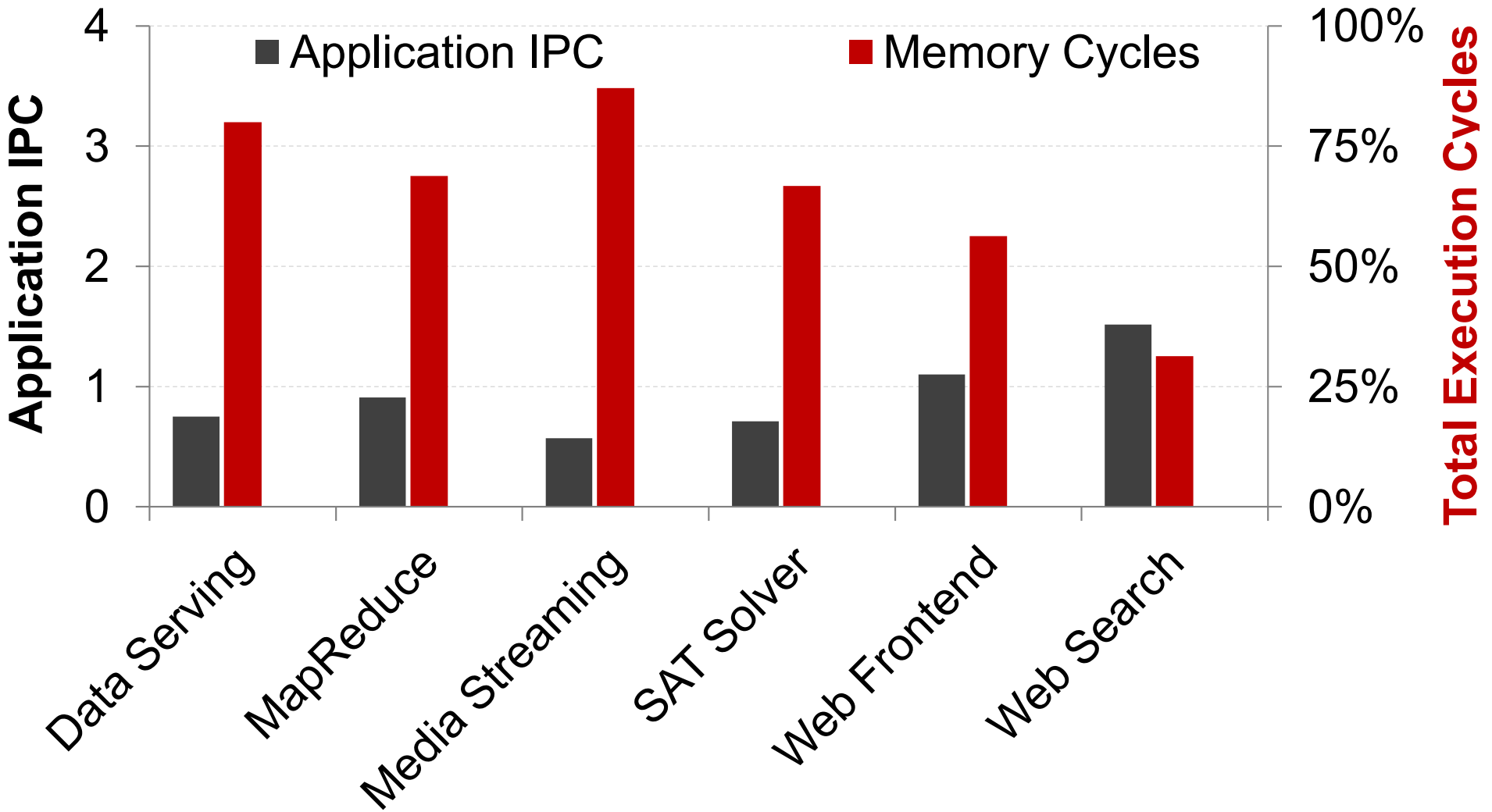
24GB RAM

Methodology: “Server-grade” CPU

- Aggressive OoO cores
  - Run up to 4 instructions per cycle
- Large instruction window
  - 128 instructions in flight
  - 48 loads in flight
- L2 and large LLC caches
- Vast off-chip b/w



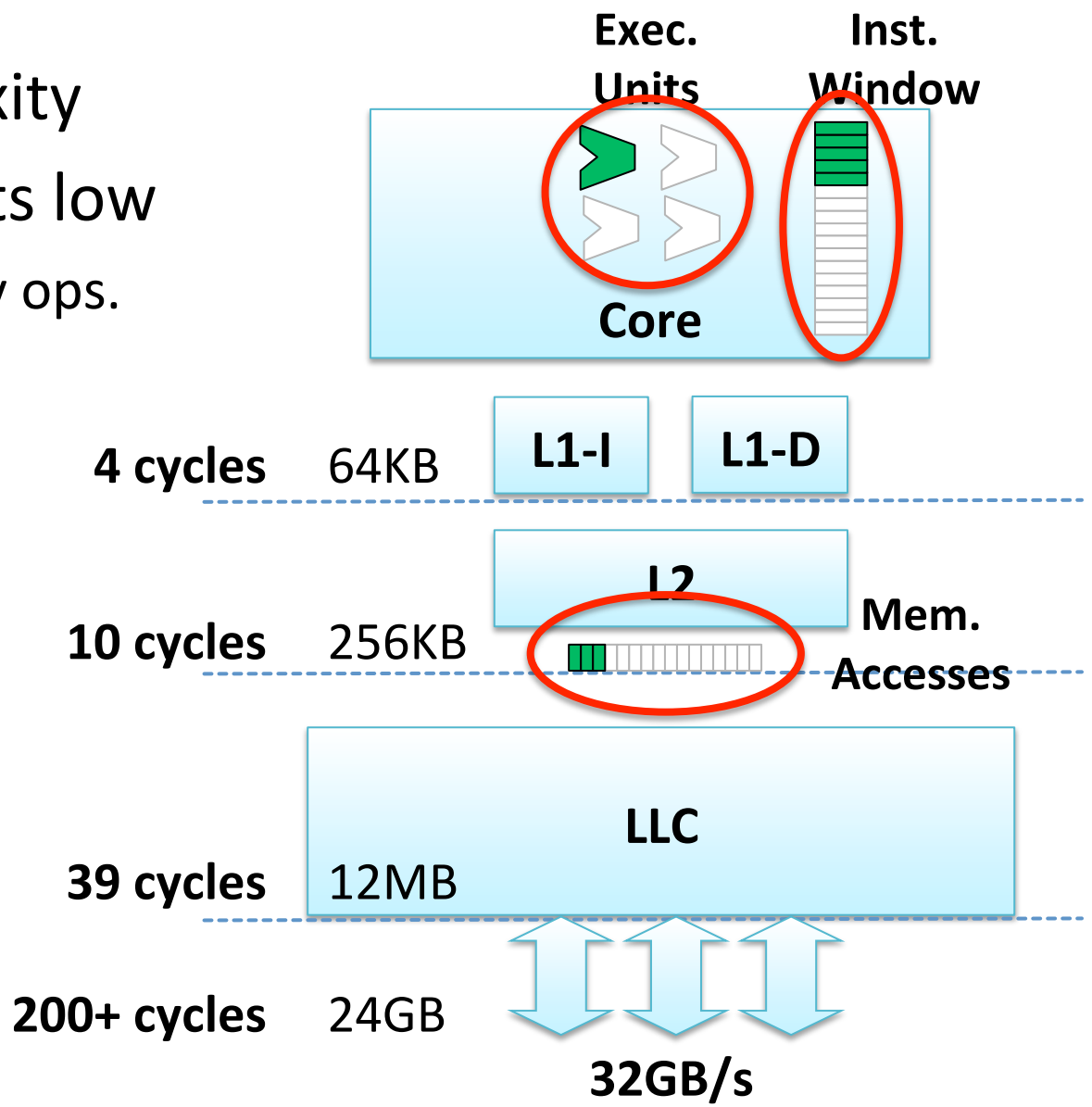
Core Inefficiencies



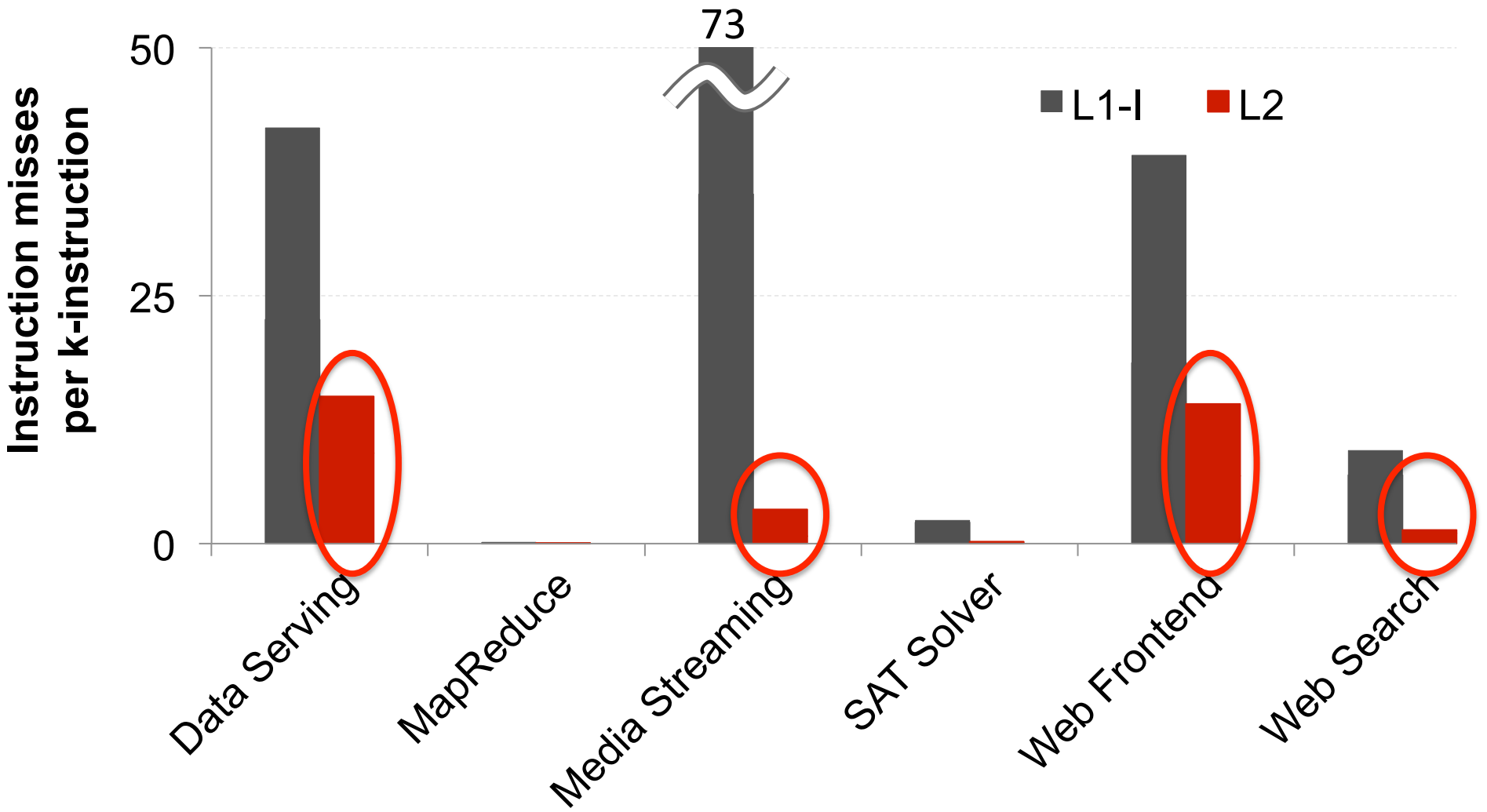
Execute ~1 instruction per cycle

Core Inefficiencies

- Underutilized complexity
- Scale-out requirements low
  - couple parallel memory ops.
  - one execution unit



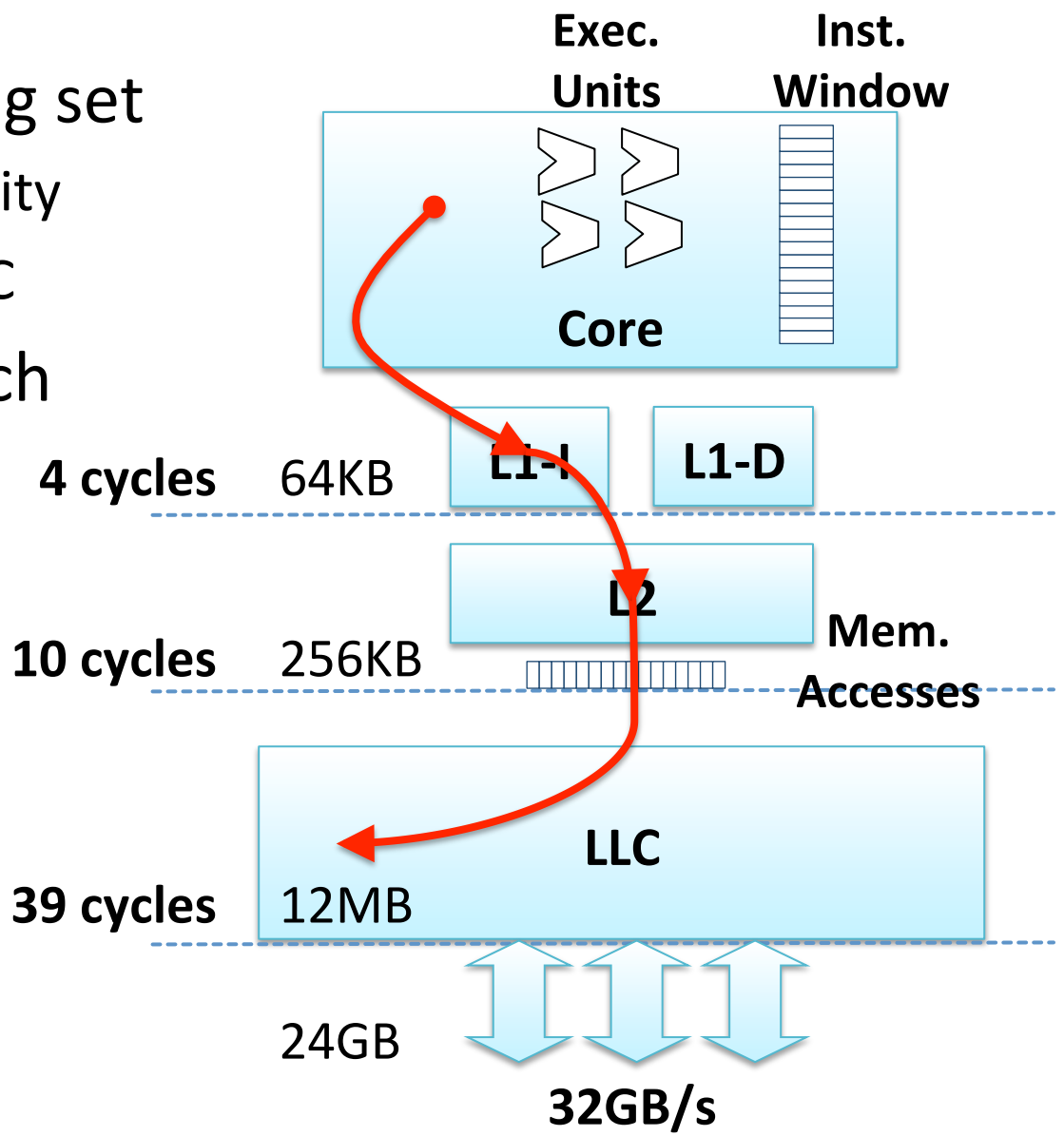
Instruction-Fetch Misses



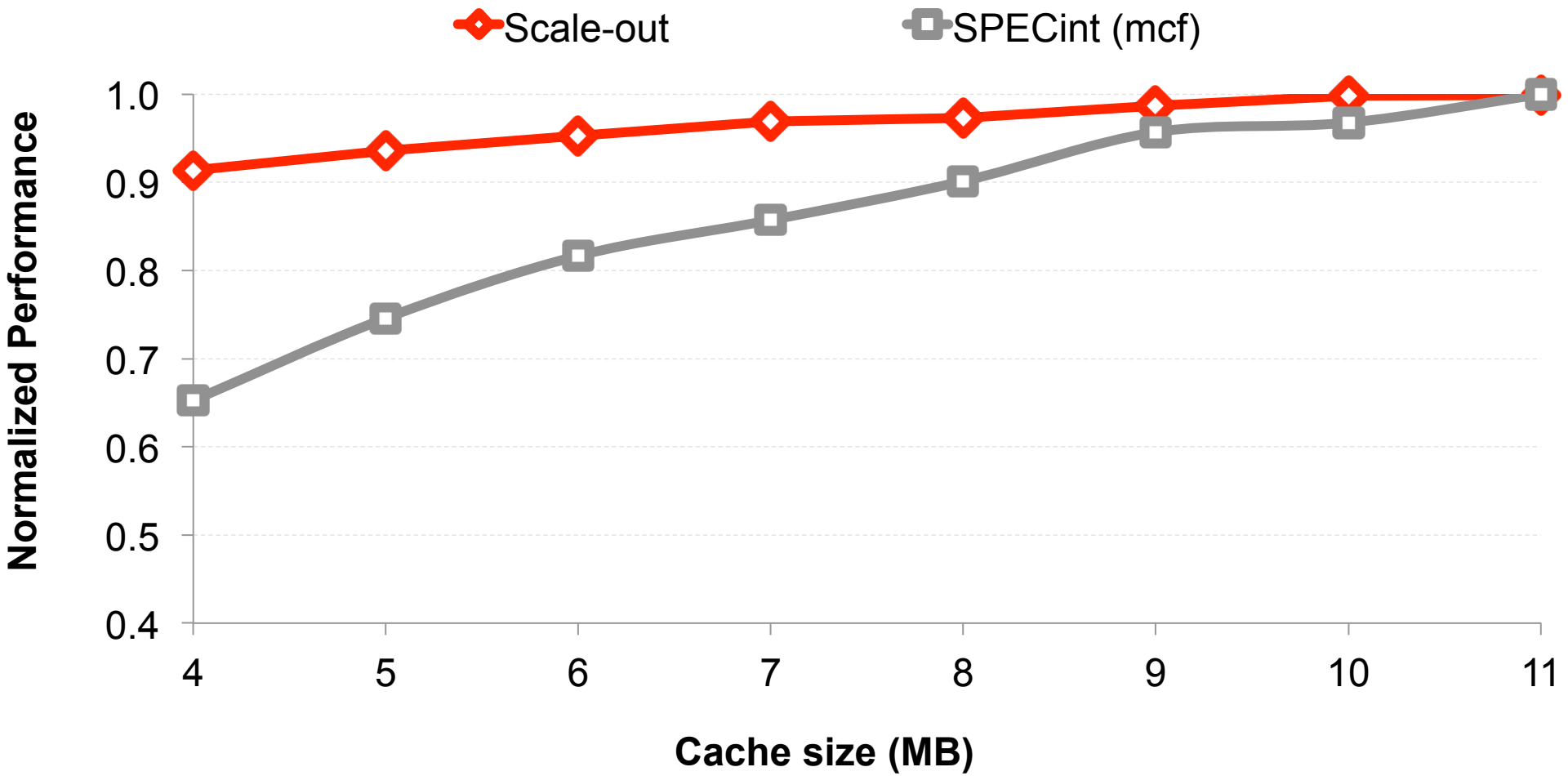
Suffer severe i-cache miss penalties

Instruction-Fetch Inefficiencies

- Large instruction working set
  - Larger than L1 & L2 capacity
  - Instructions read from LLC
- Core stalled during i-fetch

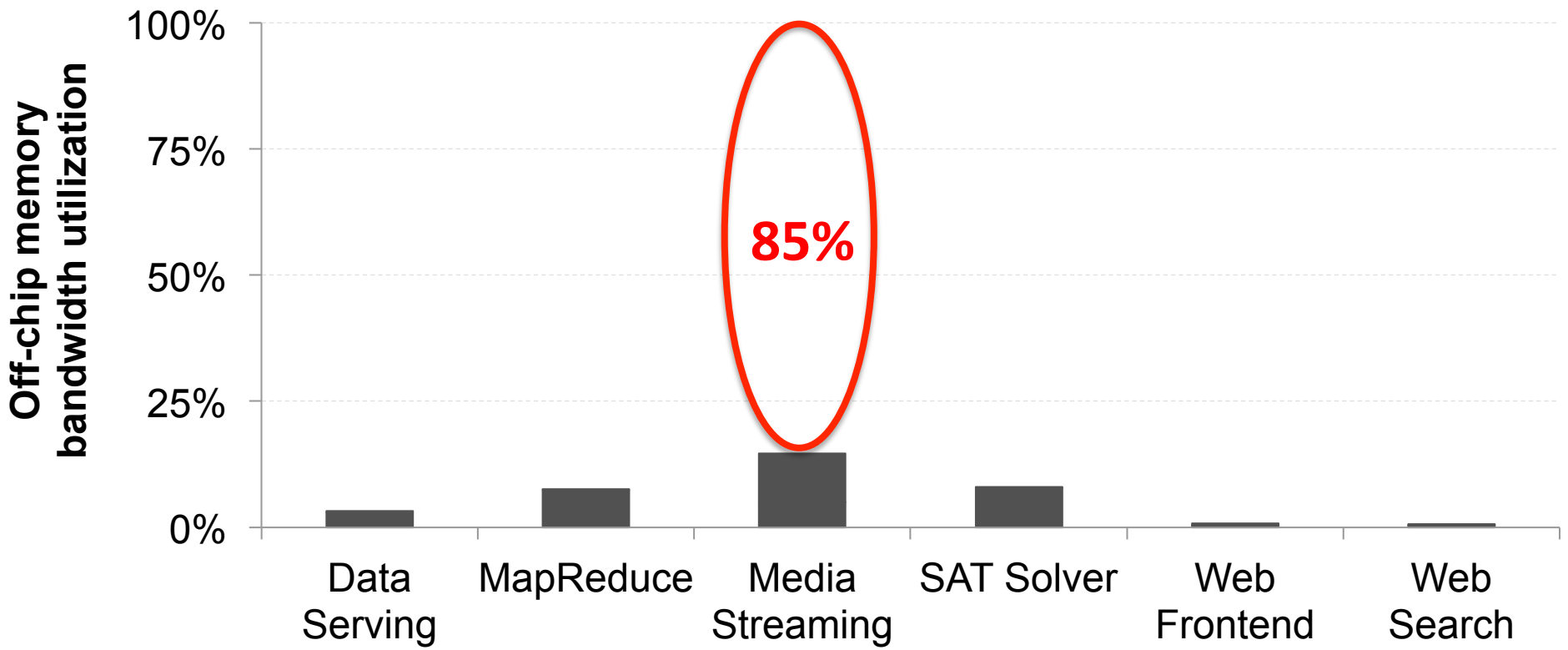


LLC Sensitivity



Minimal performance from large LLC

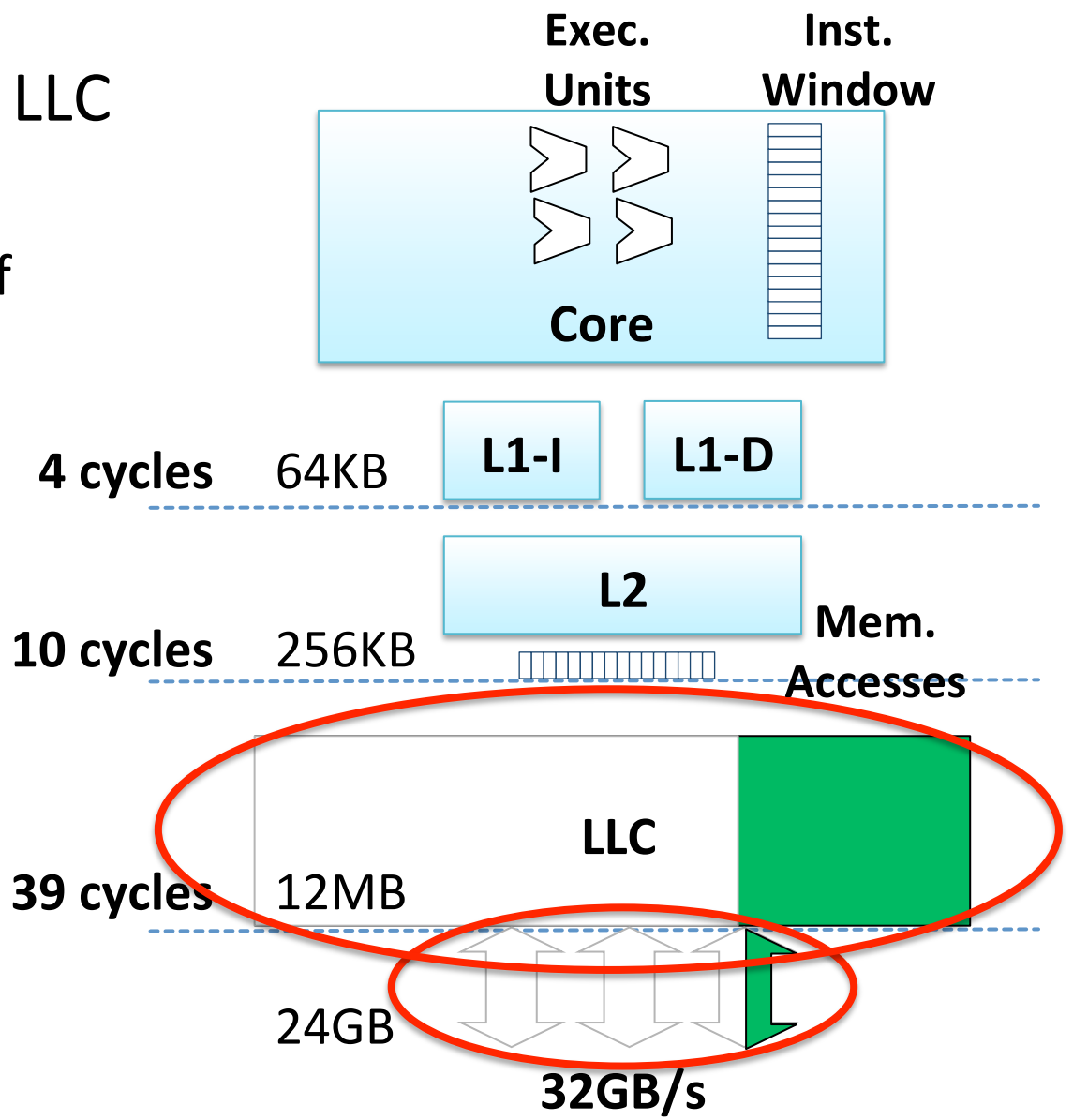
Off-chip Memory Bandwidth



Off-chip BW severely underutilized

LLC and Bandwidth Inefficiencies

- Scale-out needs modest LLC
  - Beyond 3-4MB useless
  - Area & latency w/o payoff
- Low per-core BW needs
  - <15% utilization
  - Too many channels
  - Too high frequency



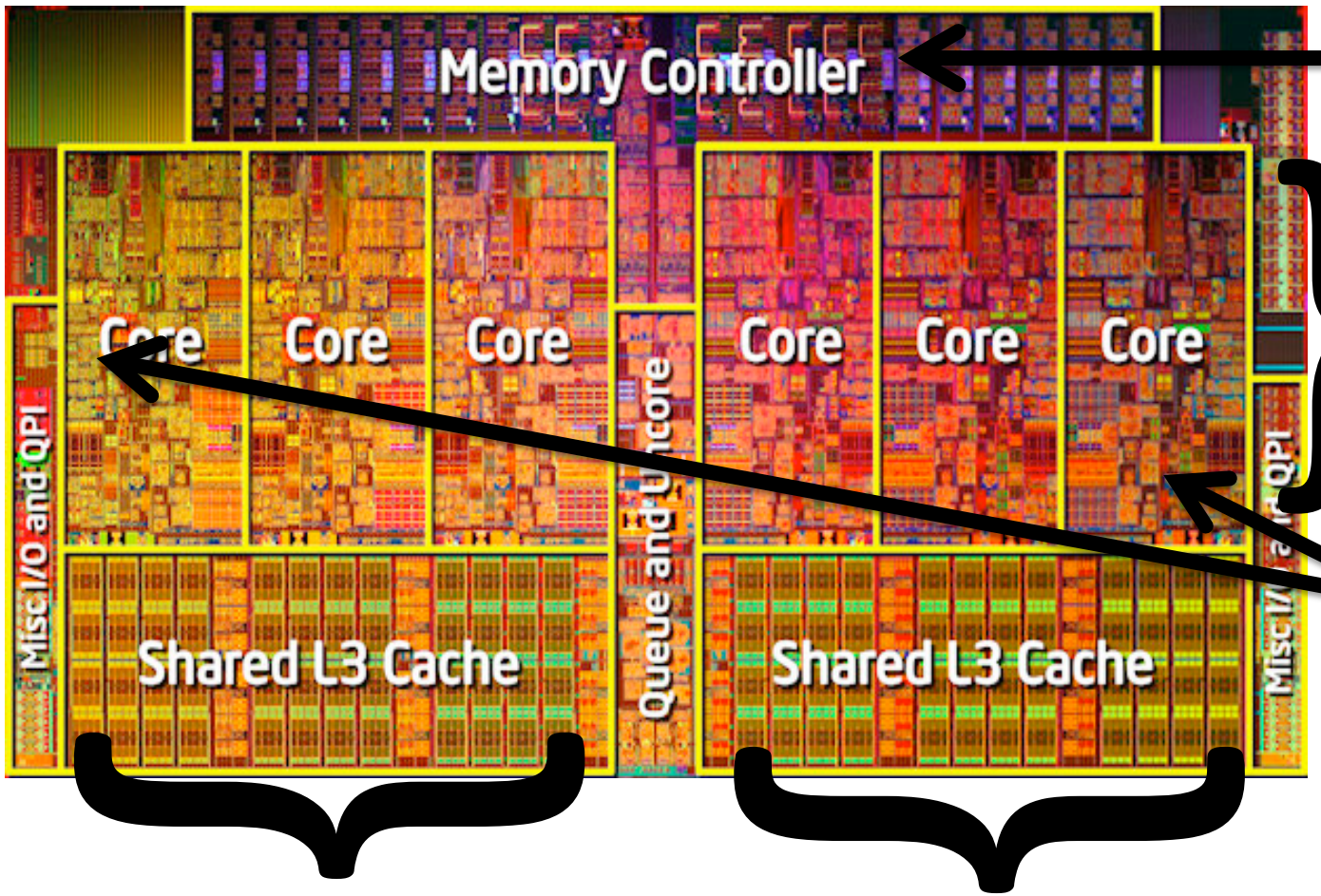
What do Scale-Out Apps Want?

- Lots of simple cores (2-way, low MLP)
- Optimized instruction fetch
  - long instruction fetch latency really hurts
- Smaller, faster LLC
- Minimal on-chip connectivity

Mismatch with both "fat" and "lean" camps

Clearing the Clouds in a nutshell [ASPLOS 2012]

Workload/Server Mismatch



10 MB (80%) waste of space (no reuse)!

Two Inflection Points Colliding

1. Emergence of Digital Universe
2. End of "Free Energy"

- Data-centric Universe meets Energy Wall
  - How efficient are today's servers?

  ➔Scale-Out Processors

Scale-out Workloads & Datacenters

Building scale-out datacenters 101

- Buy as much DRAM as you can per blade
- Populate DRAM with sharded data
- Configure rest of blade to serve data in DRAM

TCO implications:

- DRAM dominates both **costs** & **power** budget
- Maximizing throughput/blade → reduces # of blades
- Fewer blades → better Total Cost of Ownership

Need silicon-optimal processors!

How to measure silicon-optimality?

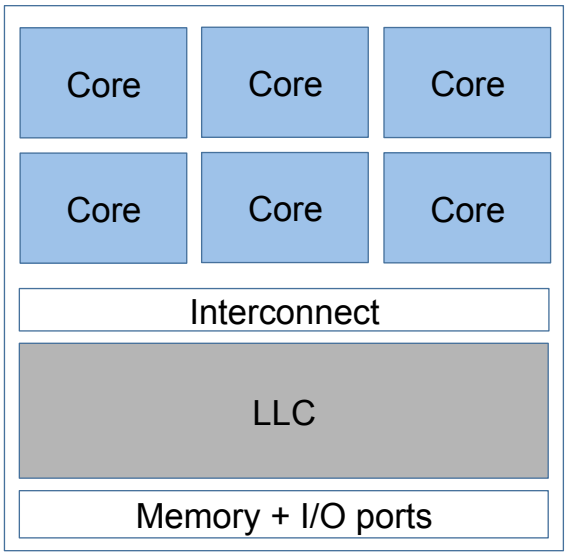
Goal: maximize silicon efficiency

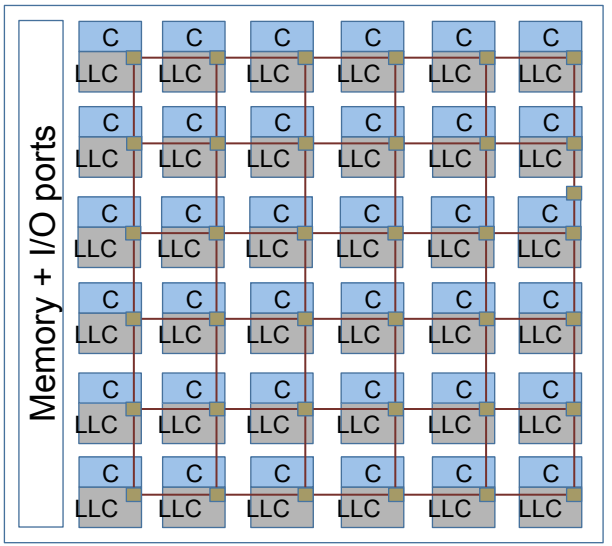
- Best use of area within chip resource budget
- Silicon efficiency ≈ Performance/Unit Area

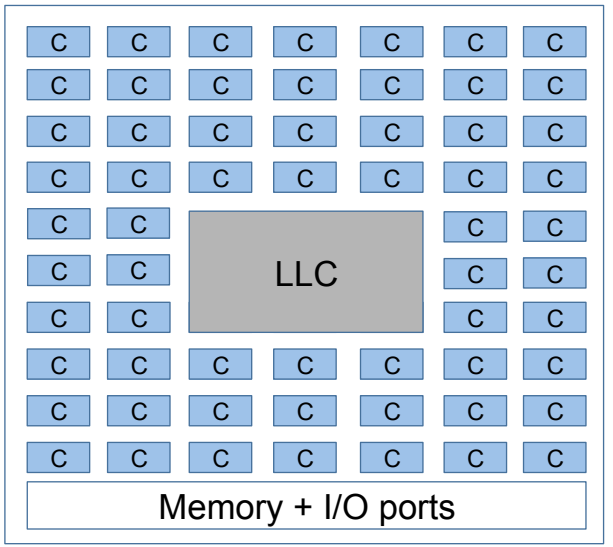
We call this Performance Density (PD)

Maximum PD implies optimal silicon!

Server Processor Landscape







Conventional

- ✗ Few fat cores
- ✗ Large caches
- ✗ Low PD

Intel: Xeon

Tiled

- ✓ Many lean cores
- ✗ Too much cache
- ✗ Long delay
- ✗ Modest PD

Tilera: TileGX-3036

Ideal

- ✓ Many lean cores
- ✓ Minimal caches
- ✓ Short delay
- ✓ High PD

Scale-Out Processors 101 [ISCA 2012]

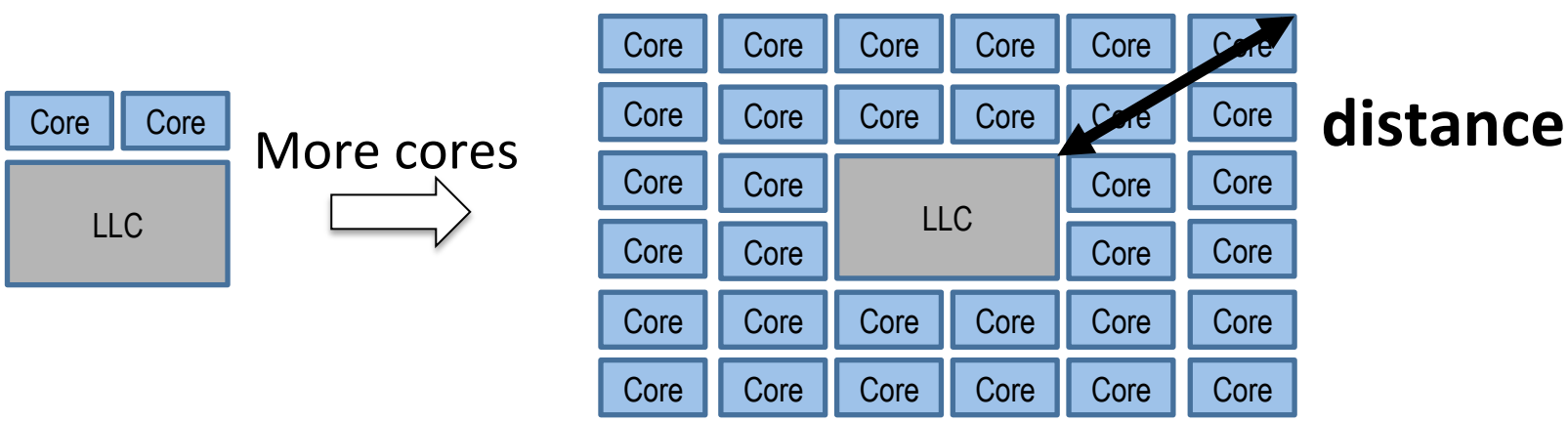
Start with little cache

- Capture instructions and OS data

Surround it with cores

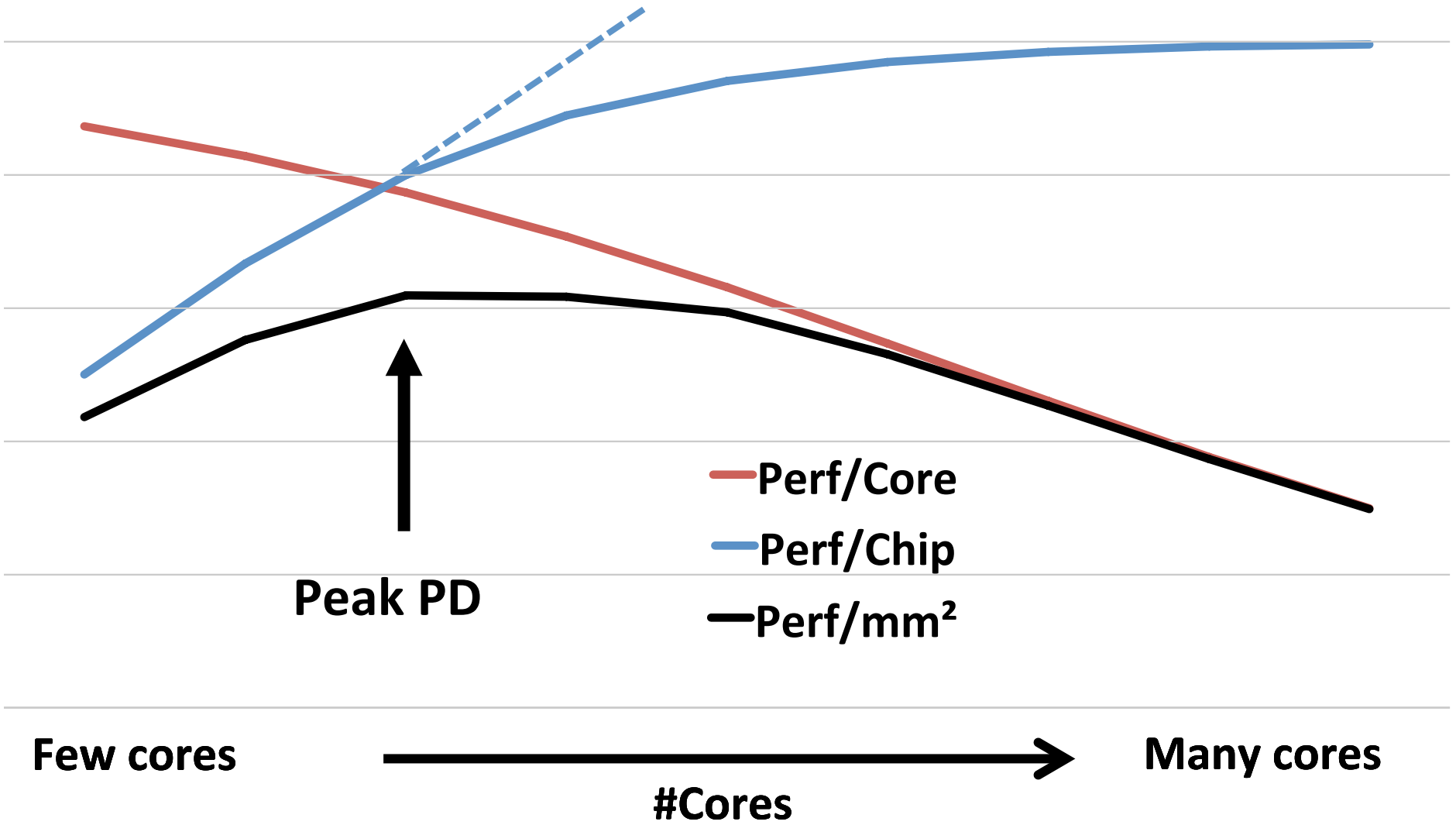
Will reach optimality with tens of cores

- More cores ➔ higher throughput!
- But, more cores ➔ longer distance to cache!



Trade-off: compute ratio and distance

How to Choose the Number of Cores?



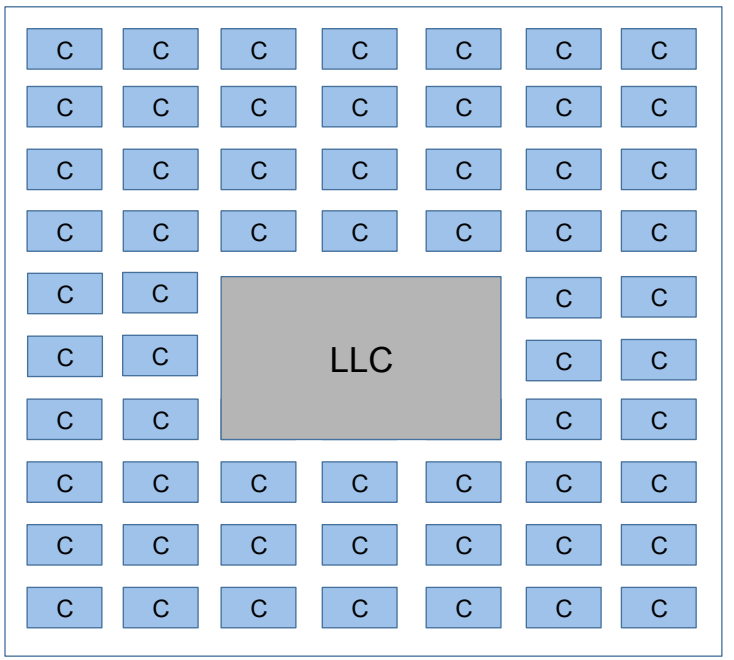
Choose # of cores that maximizes PD

Pod: Optimal Building Block

Silicon unit that maximizes performance density

- Cache capacity: 2MB - 4MB
- Number of cores: 8 ~ 32
  - Function of core area
- Interconnect network: Crossbar
  - Fast, predictable latency

Pod = one server/OS image

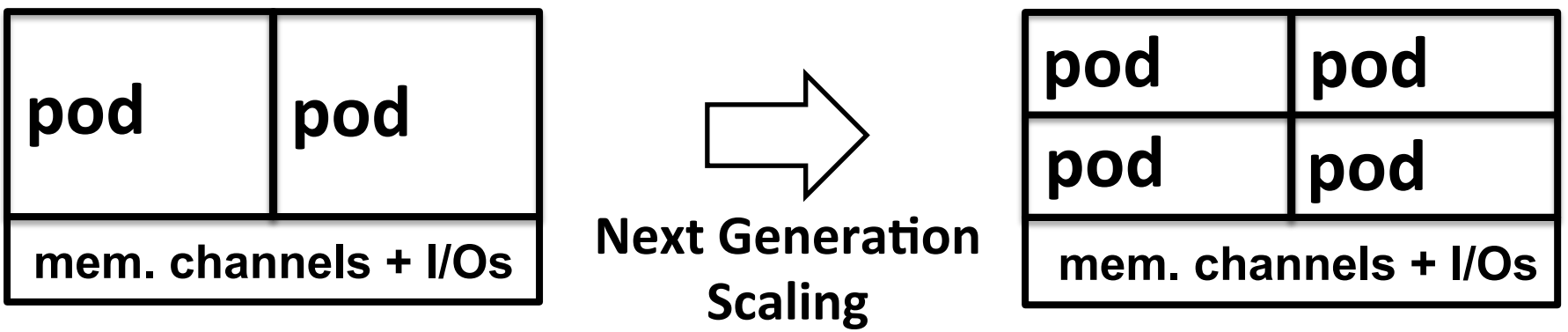


In comparison to Tile:

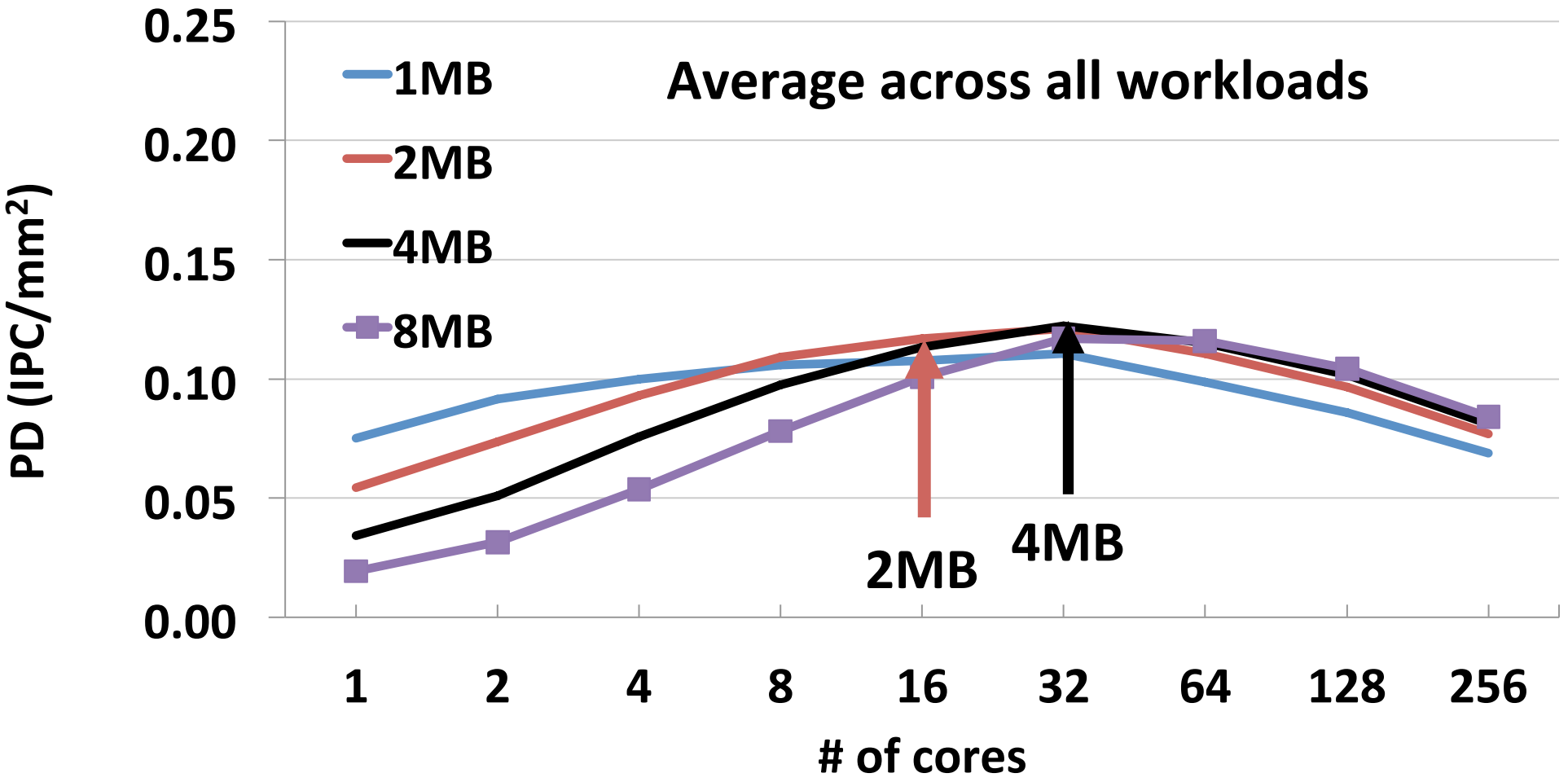
- No connectivity across Pods
- Optimized for
  - shared on-chip (LLC) instruction access
  - independent off-chip data access

Scale-Out Processors

- One or more pods
- Each pod is a server
  - Runs a full software stack
- Pods share memory and I/O interfaces
- Technology scalable
  - Increase # of pods with scaling



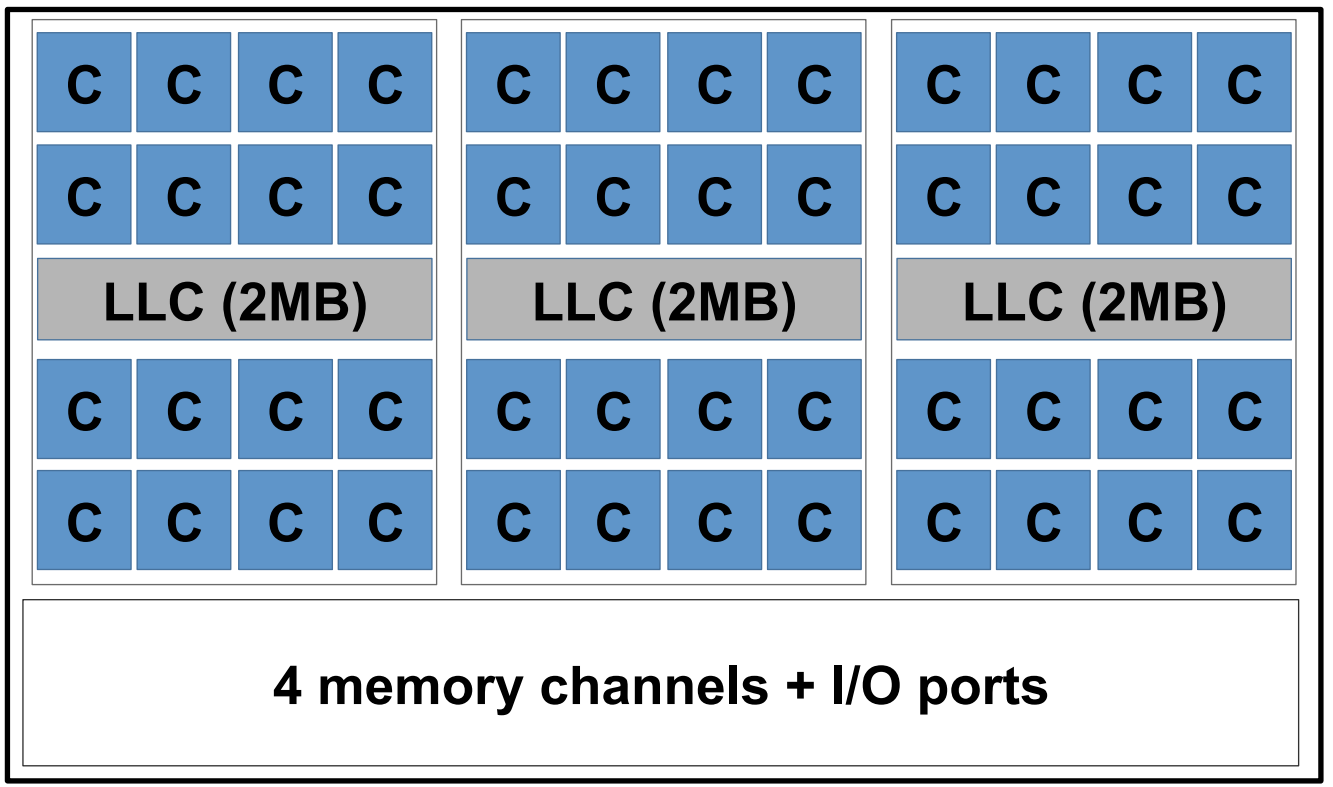
Optimal Pod for ARM Cortex-A15



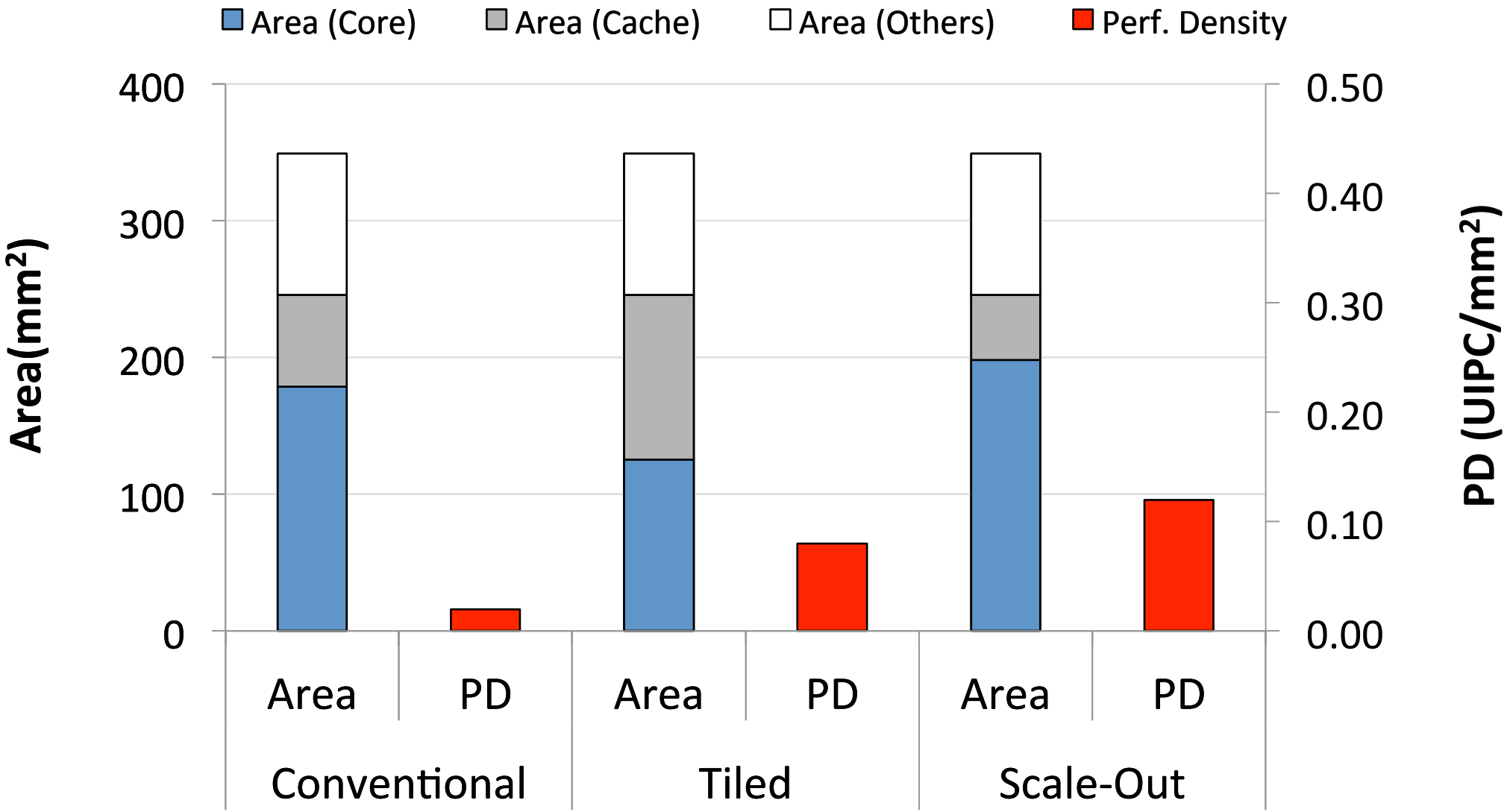
16 cores+2MB : Near-optimal with practical NoC!

Today's Scale-Out Processor

- Die area: 349mm^2
- Power consumption: 81W
- Off-chip traffic: 33GB/s (4 memory channels)
- ~6x better efficiency than Intel Xeon

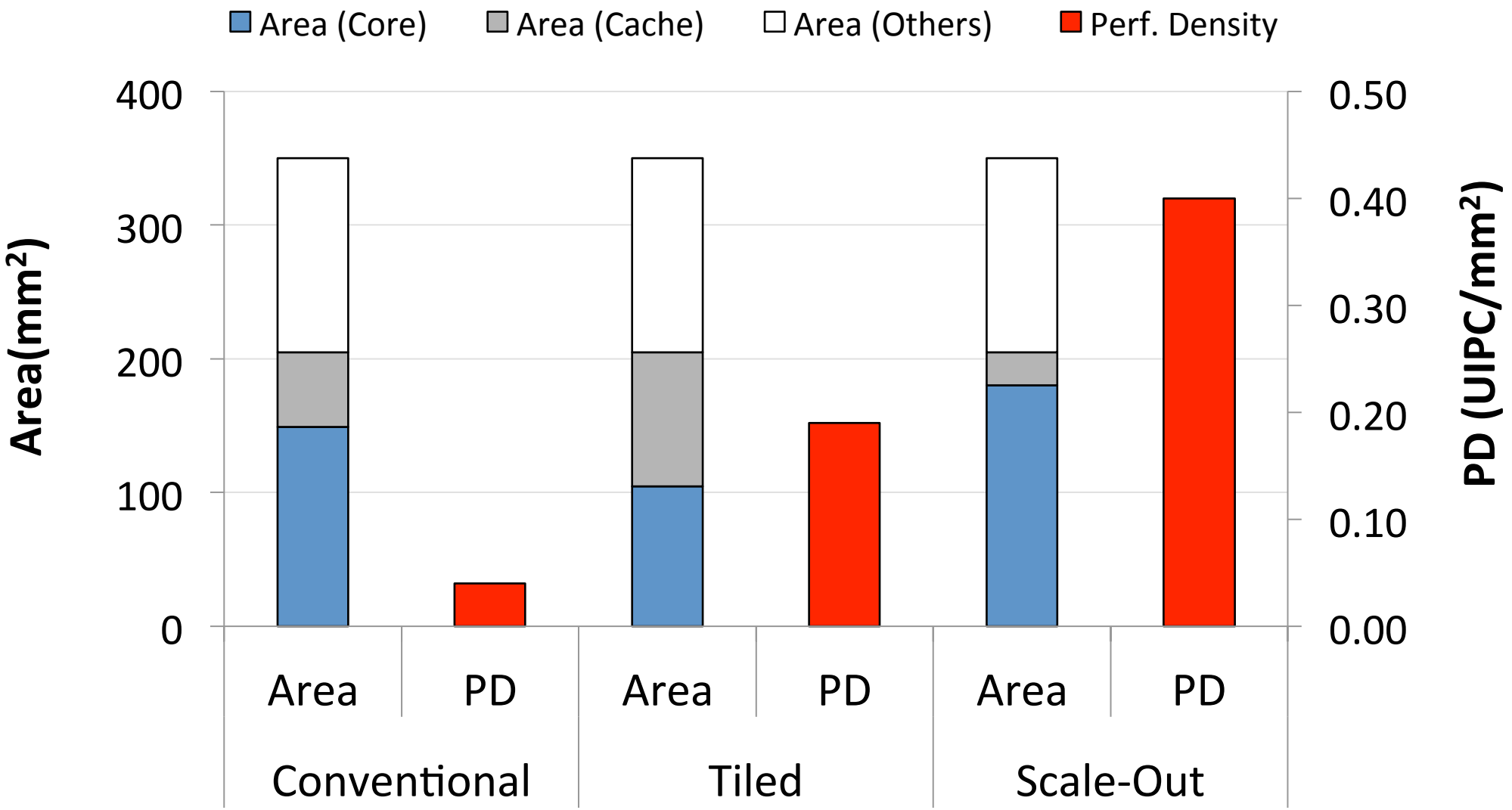


Server Chip Comparison (40nm)



Scale-Out delivers the highest PD!

Server Chip Comparison (20nm)



Scale-Out doubles PD over Tiled!

Scale-Out Processors: Summary

Memory-dominated workloads:

- DRAM dominates costs & power budget
- Maximizing throughput/blade ➔ reduces # of blades
- Fewer blades ➔ better Total Cost of Ownership

Optimize silicon for serving memory:

- Lots of small cores, cache to hold the program
- Run servers on “optimal” pods
- Replicate pods to scale!

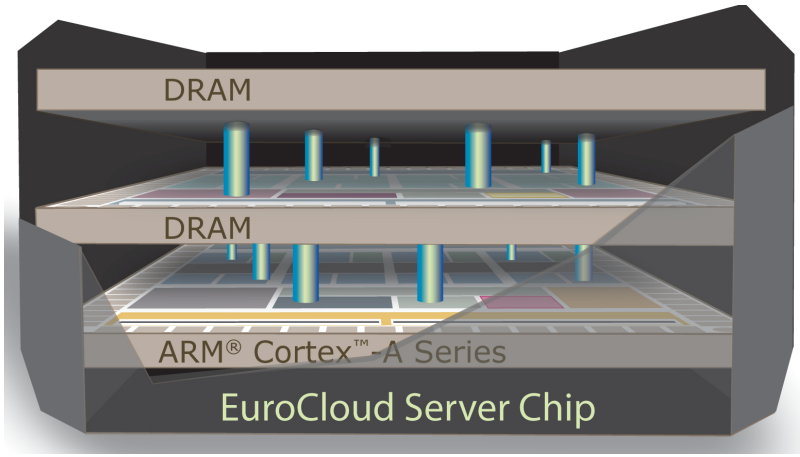
End Result: Technology-Scalable Scale-out Servers!

A Scale-Out Processor for Massive Data

(www.eurocloudserver.com) [ISCA 2012]

EuroCloud Server:

- Swarms of ARM processors
- 3D DRAM (TB/s bandwidth)
- Runs off-the-shelf SW
- Presented to EU parliament (this Fall)



Arriving Soon in a Datacenter near you!

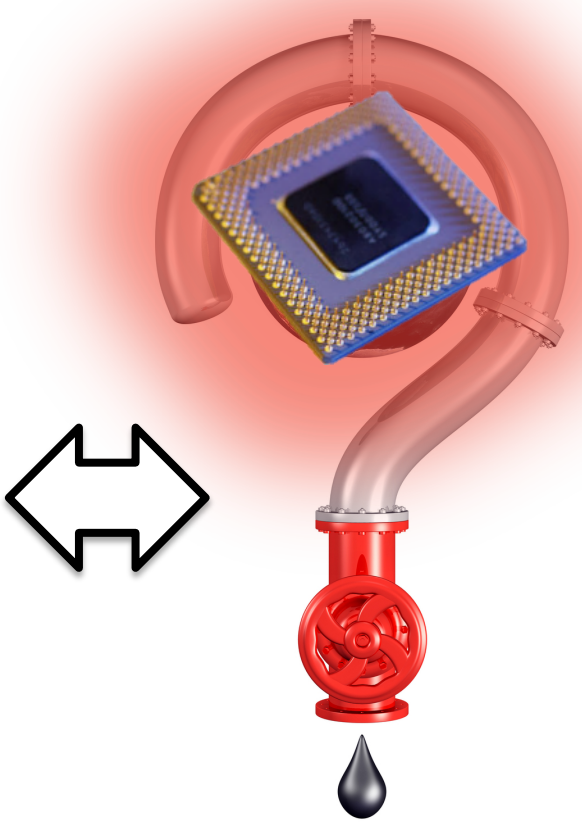
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

SEVENTH FRAMEWORK PROGRAMME

UCyprus

Bringing it All Together

- IT is changing everything & itself changing
- IT systems are inefficient & too robust
- Plow massive data with minimal energy



A new IT revolution is emerging,

we have a great opportunity to lead!

Thank You!

For more information please visit us at

ecocloud.ch