Big Data & Dark Silicon

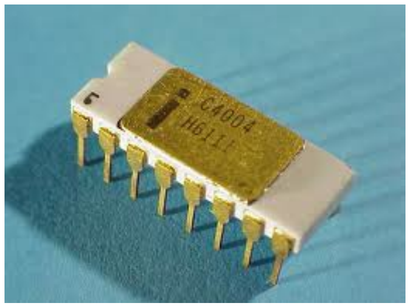
Taming Two IT Trends on a Collision Course

Babak Falsafi
Director, EcoCloud
ecocloud.ch

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Information Technology (IT):
Four Decades of Exponential Growth

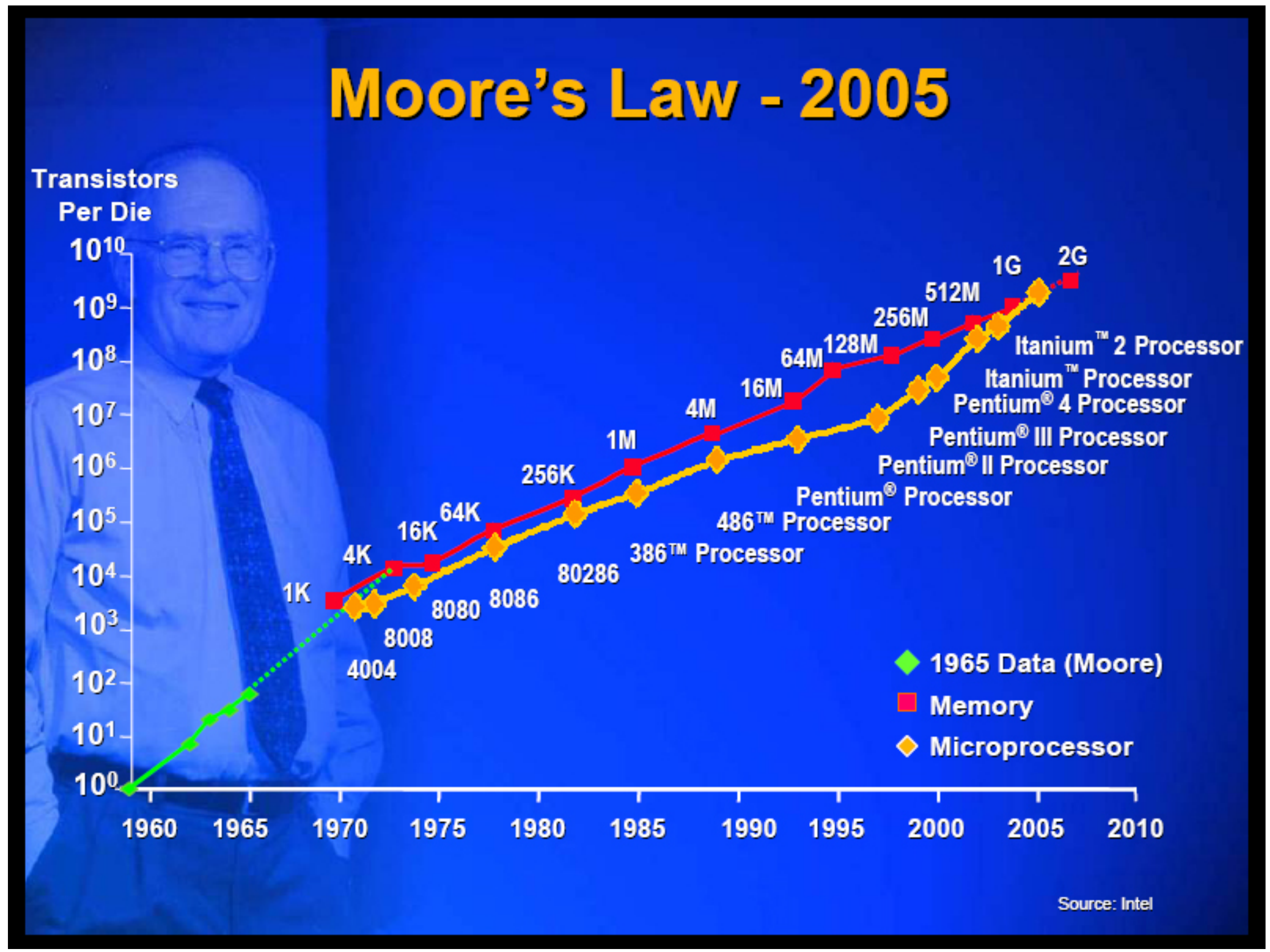
Intel 4004, 1971



92,000 ops/sec

Intel Xeon, 2011

96,000,000,000 ops/sec



IT is at the core everything we do & has become an indispensable pillar for a modern day society!

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

Two IT Trends on a Collision Course

1. Big Data
   - Data grows at unprecedented rates
   - Silicon performance & capacity at 1.5x/year

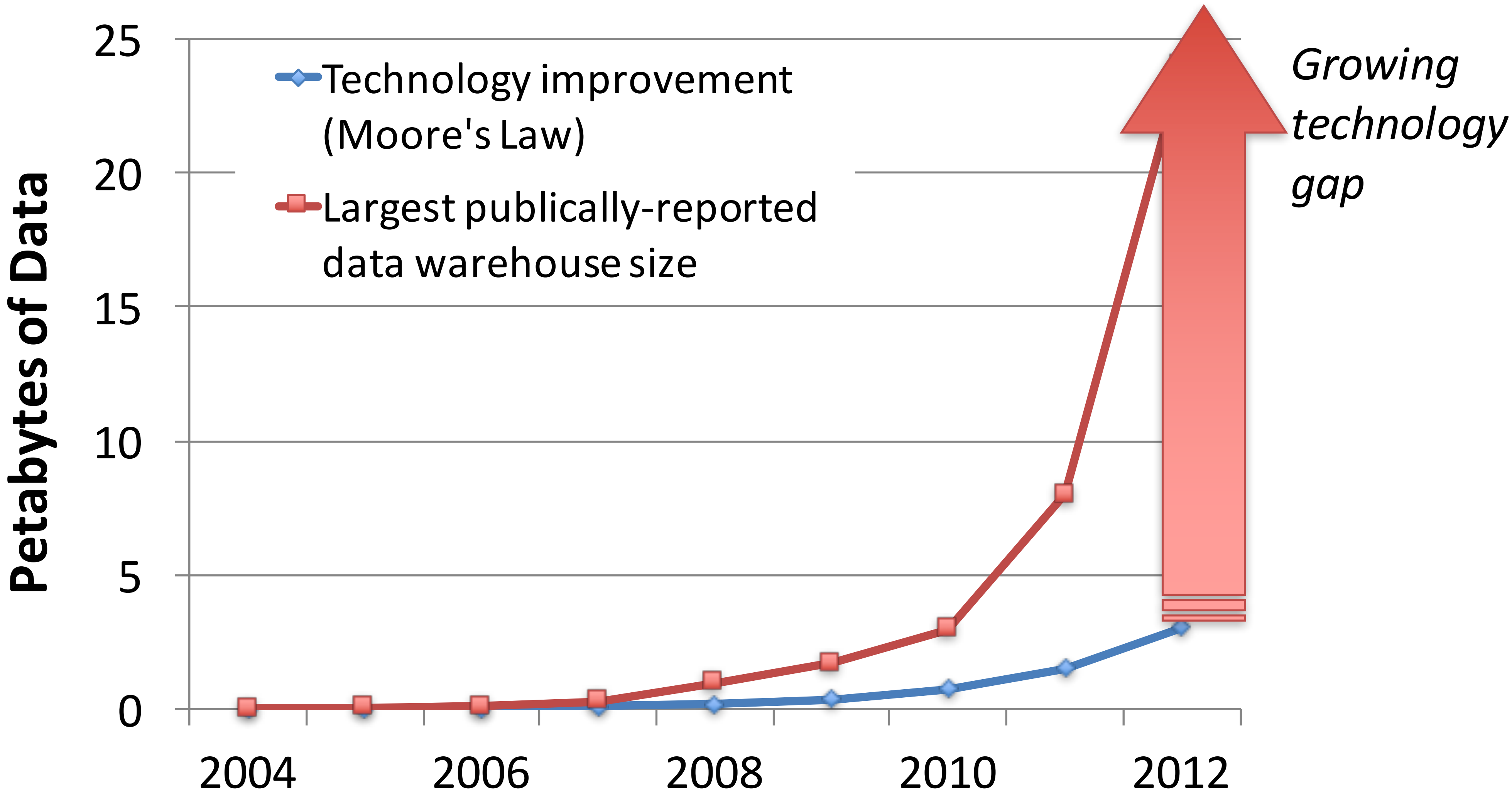
2. Energy
   - Silicon density increase continues
   - But, Silicon efficiency has slowed down/will stop
   - IT energy not sustainable

Inflection Point #1: IT is all about Data

[source: Economist]

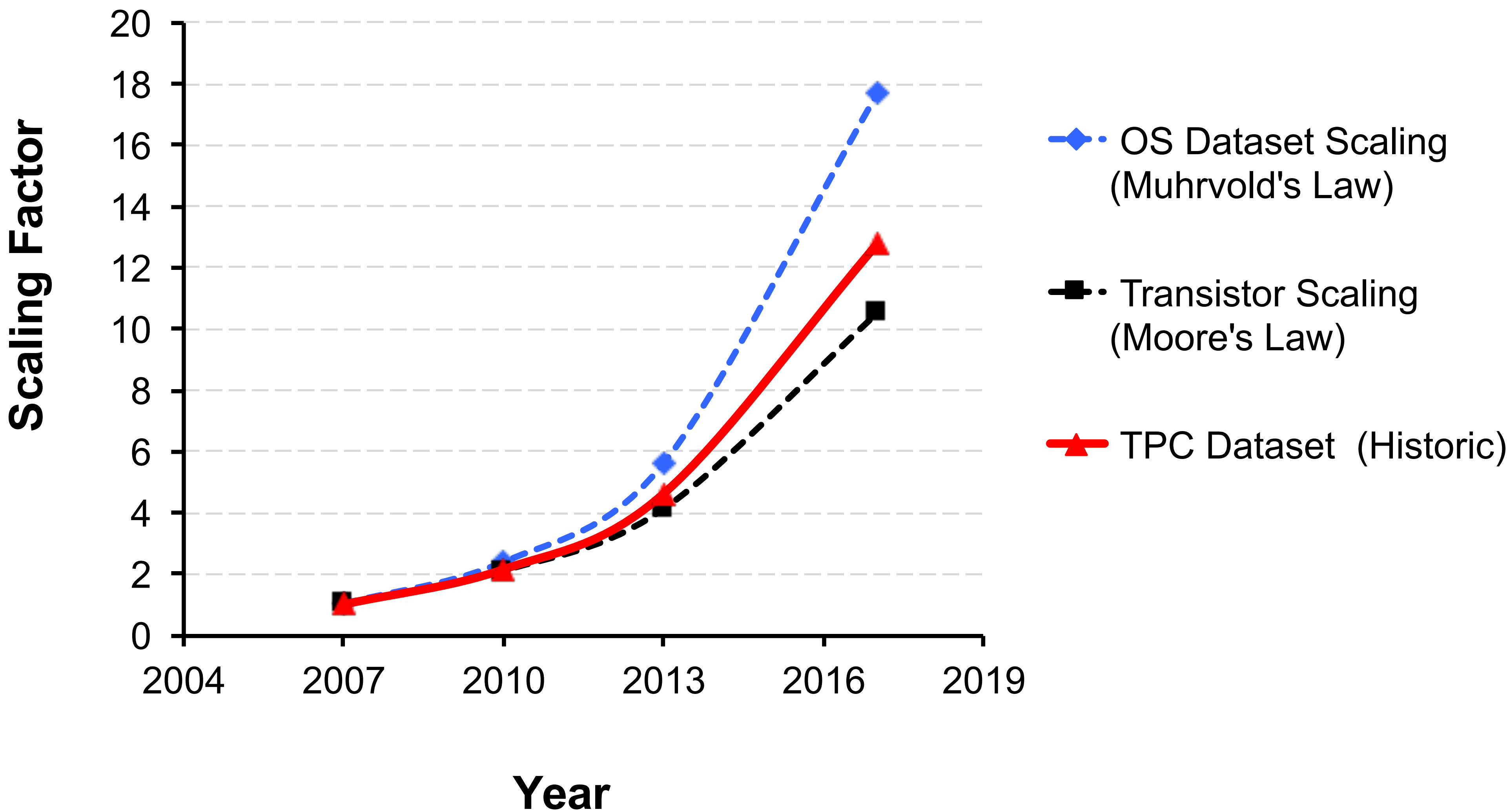
- Data growth (by 2015) = 100x in ten years [IDC 2012]
  - Population growth = 10% in ten years
- Monetizing data for commerce, health, science, services, ….
- Big Data is shaping IT & pretty much whatever we do!

Data Growing Faster than Technology



WinterCorp Survey, www.wintercorp.com

Application/OS Datasets Scaling



Data-Centric IT Growing Fast

Source: James Hamilton, 2012

Daily IT growth in 2012 = IT first five years of business!

Warning! Datacenters are not Supercomputers!

- Run heterogeneous data services at massive scale
- Driven for commercial use
- Fundamentally different design, operation, reliability, TCO
  - Density 10-25KW/rack as compared to 25-90KW/rack
  - Tier 3 (~2 hrs/downtime) vs. Tier 1 (upto 1 day/downtime)
  - ……and lots more

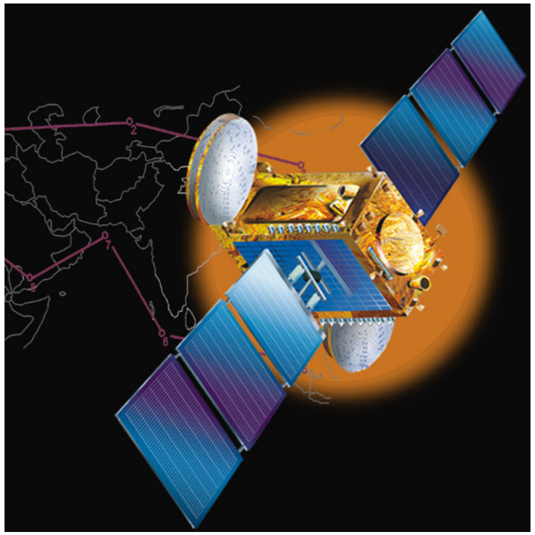
Datacenters are the IT utility plants of the future

Supercomputing

$\neq$

Cloud Computing

Data Comes in Various Flavors



Satellite

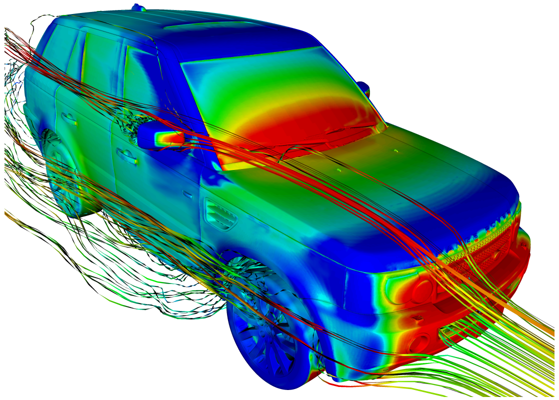
Health

Entertainment

Life

Commerce

Search



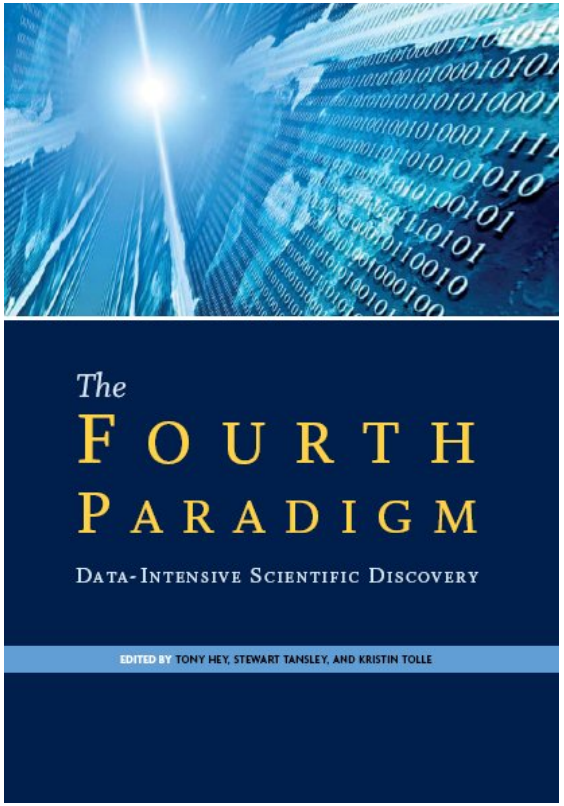
Simulation

Data Shaping All Science & Technology

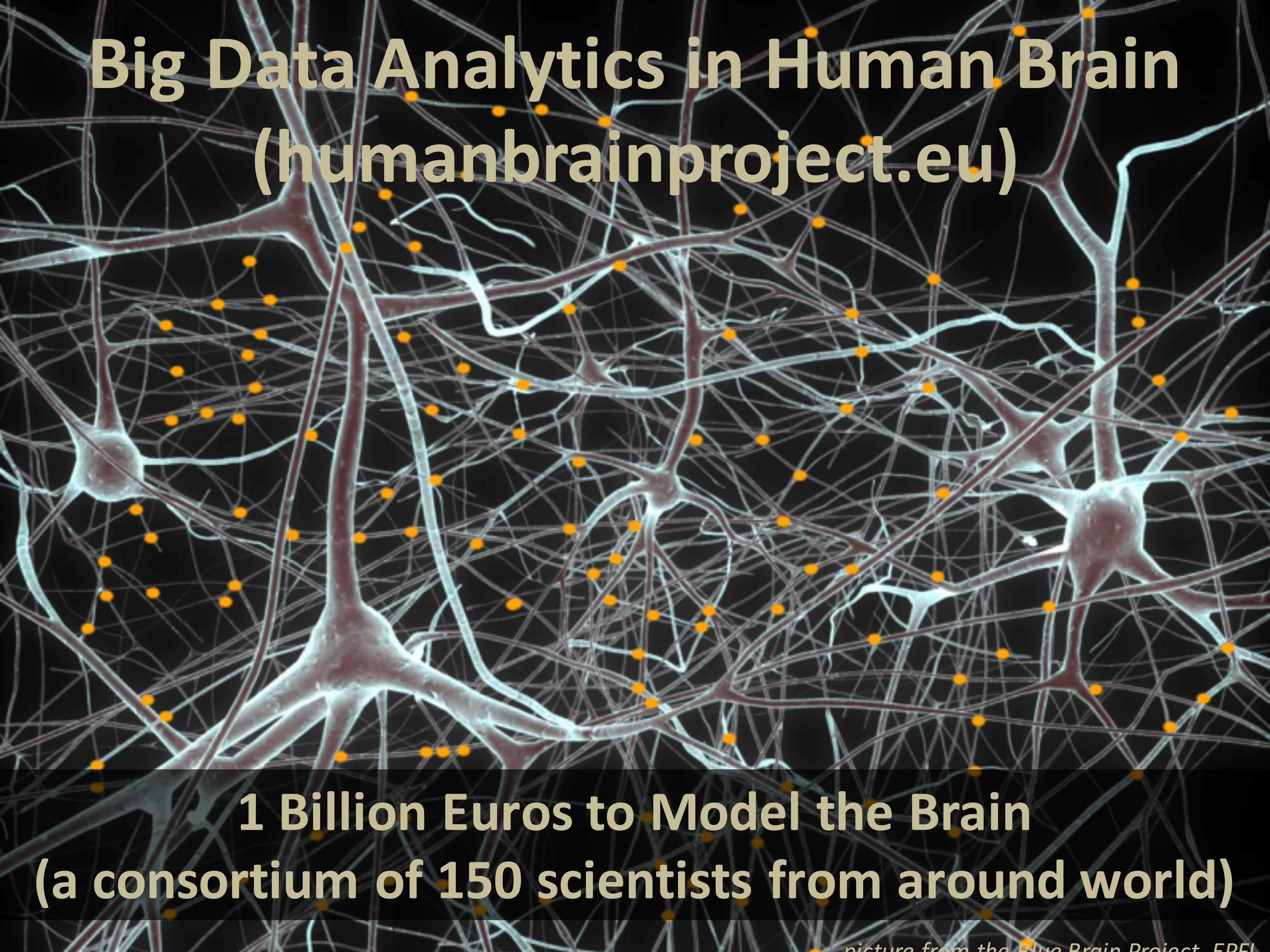
Science entering 4th paradigm

- Analytics using IT on
  - Instrument data
  - Simulation data
  - Sensor data
  - Human data
  - …

Complements theory, empirical science & simulation



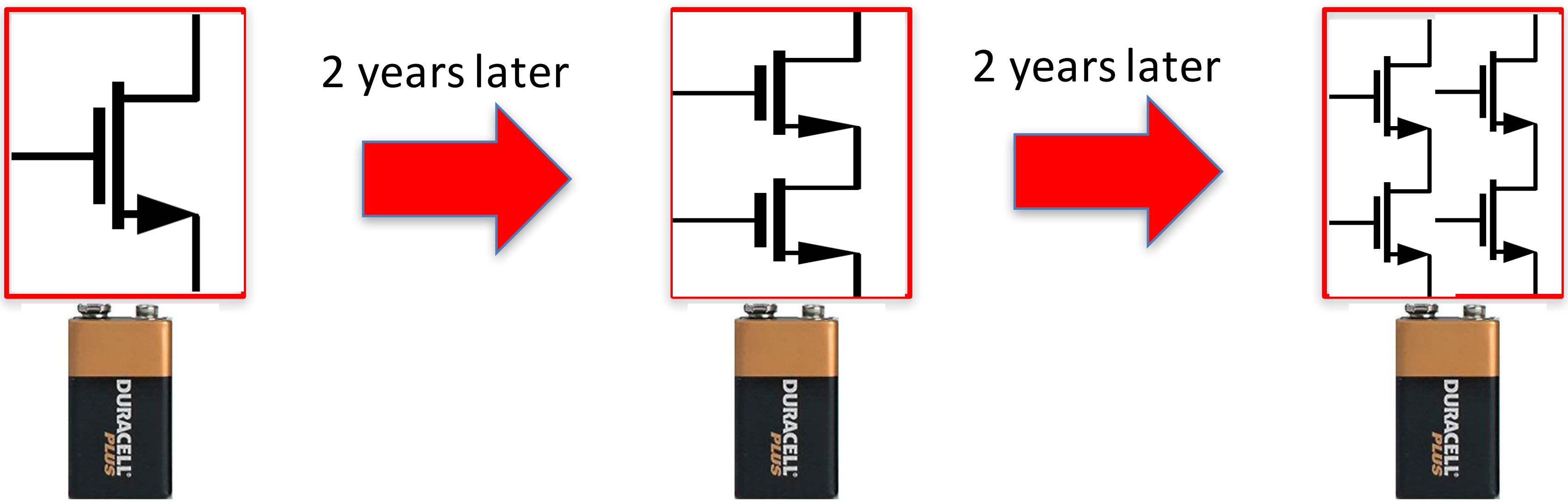
Strategically vital for innovation & tech-based economies!


Big Data Analytics in Human Brain
(humanbrainproject.eu)
1 Billion Euros to Model the Brain
(a consortium of 150 scientists from around world)
picture from the Blue Brain Project, EPFL

Venice Time Machine (vtm.epfl.ch)

Inflection Point #2: Energy used to be "Free"

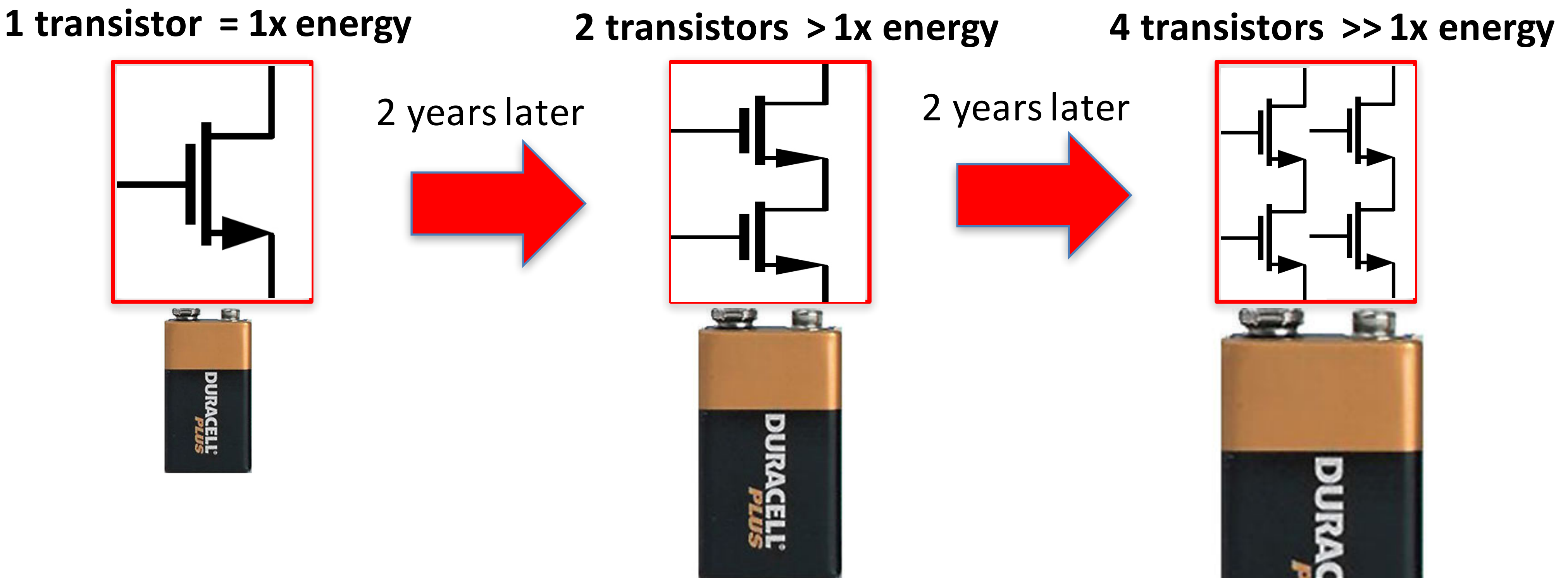
1 transistor = 1x energy **2 transistors = 1x energy** **4 transistors = 1x energy**



Before (1970~2005):

- Used to make transistors smaller
- Smaller transistors less electricity to operate
- Chip energy consumption remained ~ same

No More "Free" Energy

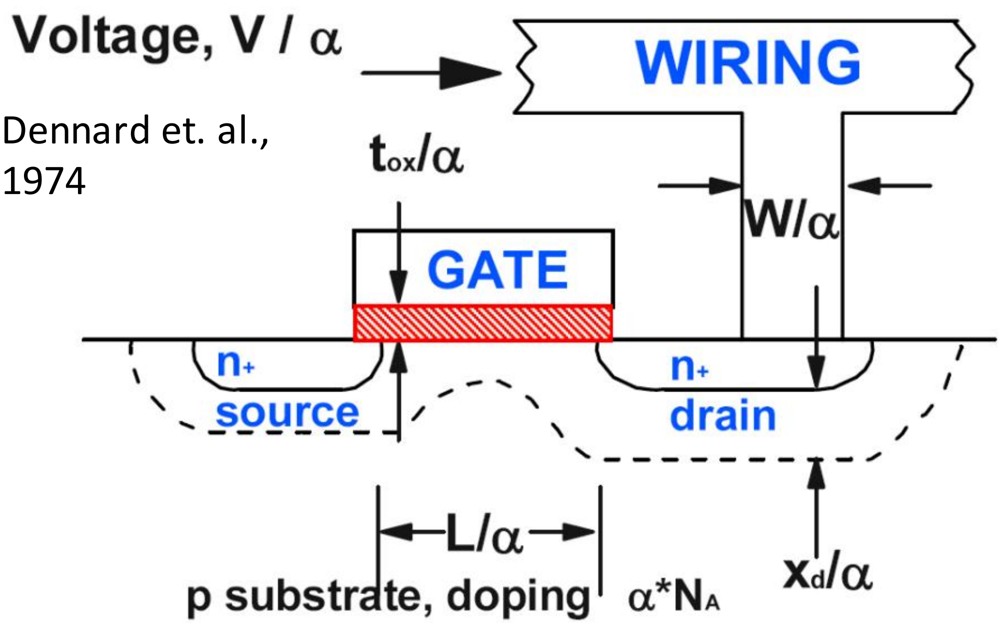


Now (2005-):

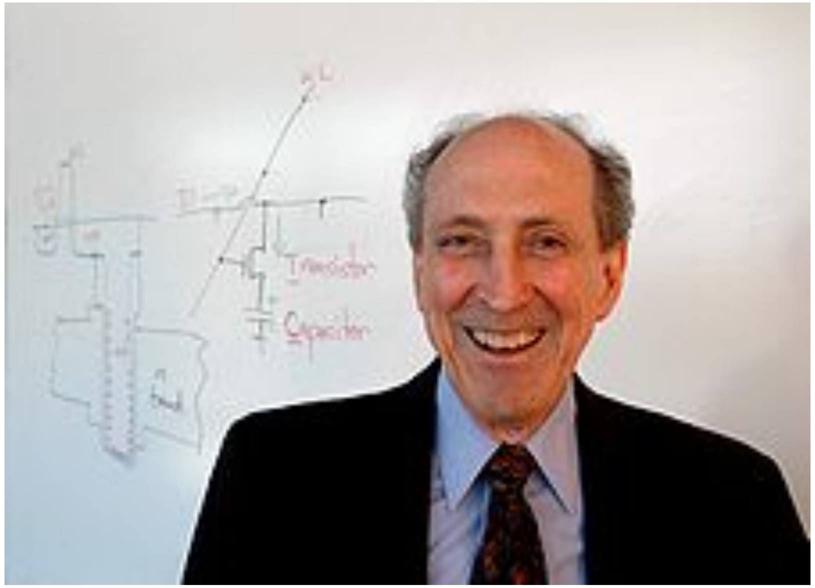
- Continue to make transistors smaller
- But, they use similar electricity to operate
- Chip energy consumption is shooting up

Where did "Free" Energy Go?

Dennard et. al., 1974



Robert H. Dennard, picture from Wikipedia



Four decades of Dennard Scaling (1970~2005):

- **$P = C V^2 f$**
- More transistors
- Lower voltages

➜ Constant power/chip

Leakage Killed Dennard Scaling

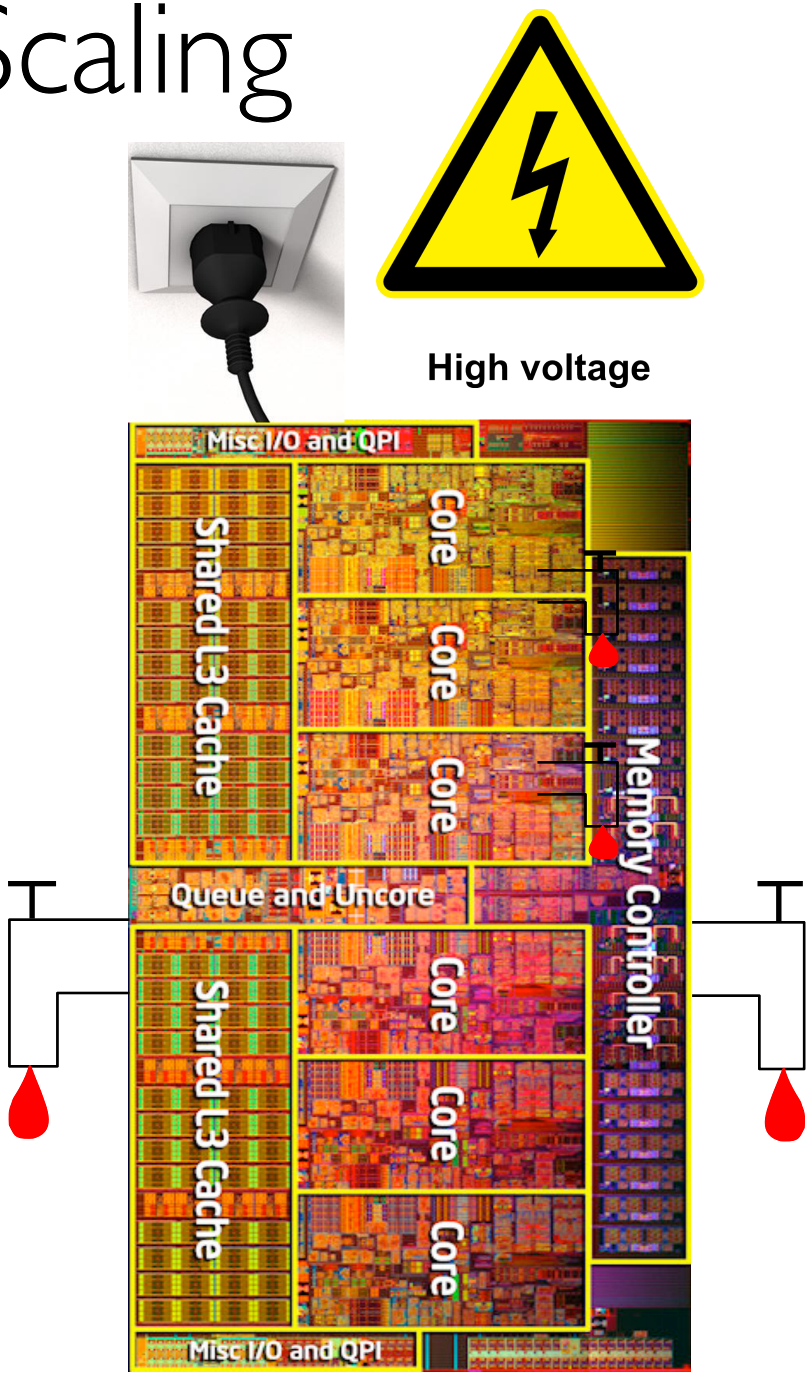
Leakage:

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

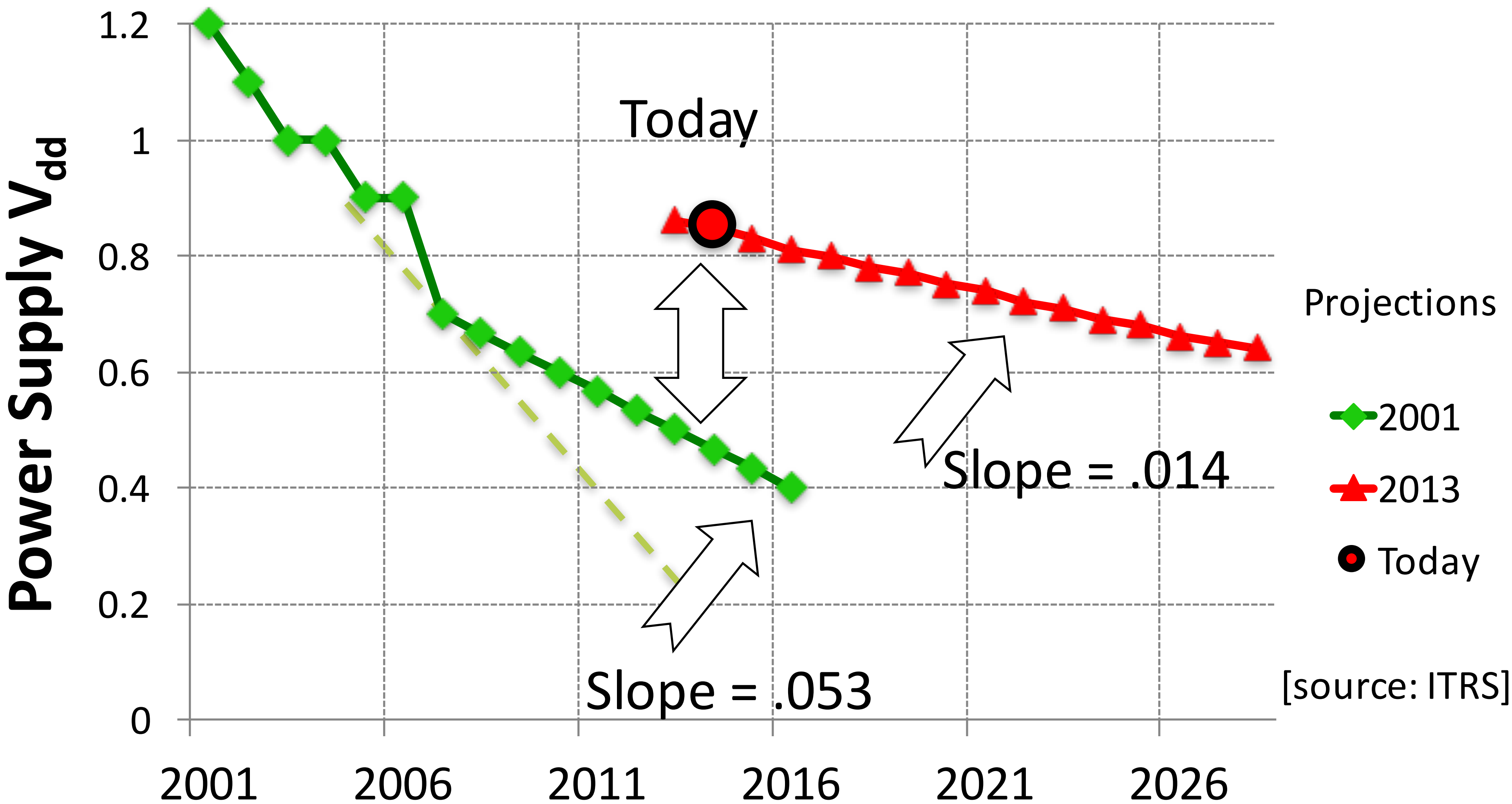
To switch well

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

➜V_{dd} can't go down



End of Dennard Scaling

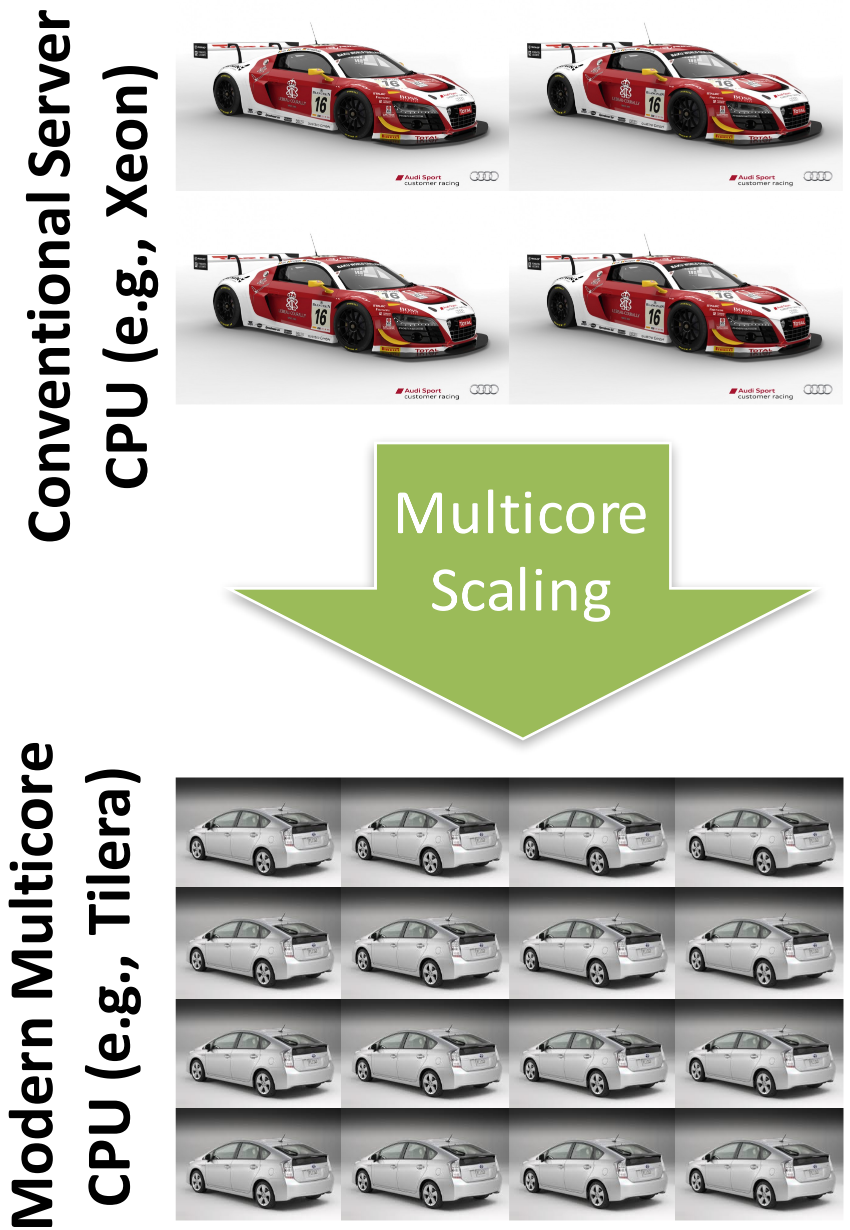


The fundamental energy silver bullet is gone!

The Rise of Parallelism to Save the Day

With voltages leveling:

- Parallelism has emerged as the only silver bullet
- Use simpler cores
  - Prius instead of Audi
- Restructure software
- Each core ➔ fewer joules/op



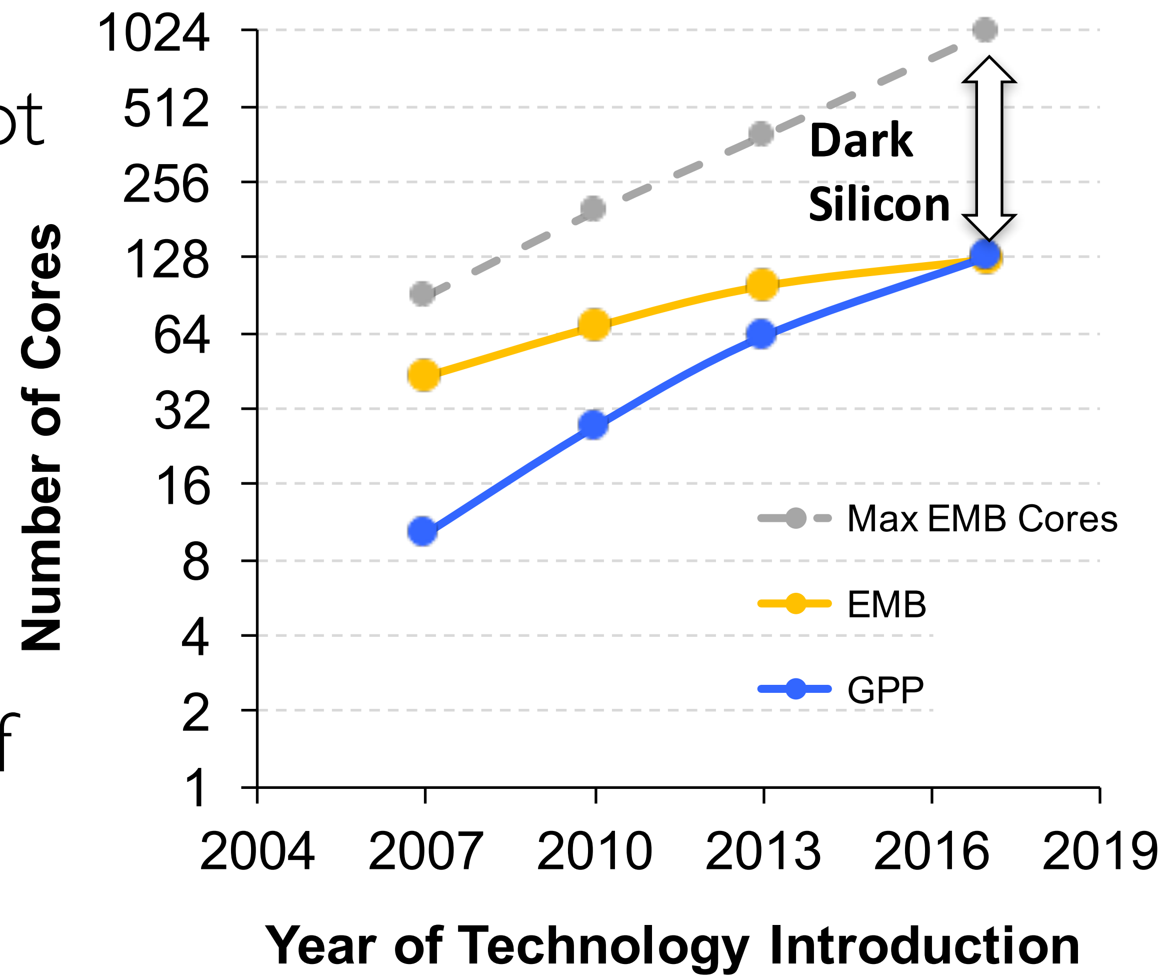
The Rise of Dark Silicon: End of Multicore Scaling

But parallelism alone can not offset leveling voltages

Even in servers with abundant parallelism

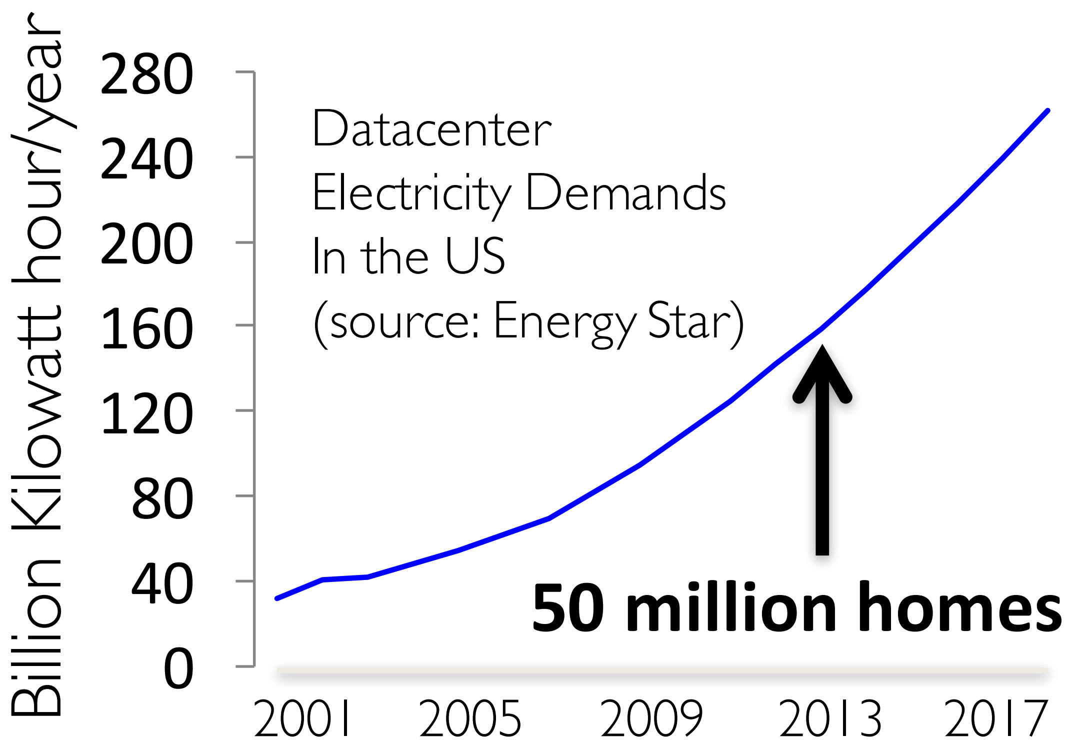
Core complexity leveled off too!

Soon, cannot power all chip



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

Higher Demand + Lower Efficiency: Datacenter Energy Not Sustainable!



- Modern datacenters → 20 MW!
- In modern world, 6% of all electricity, growing at >20%!

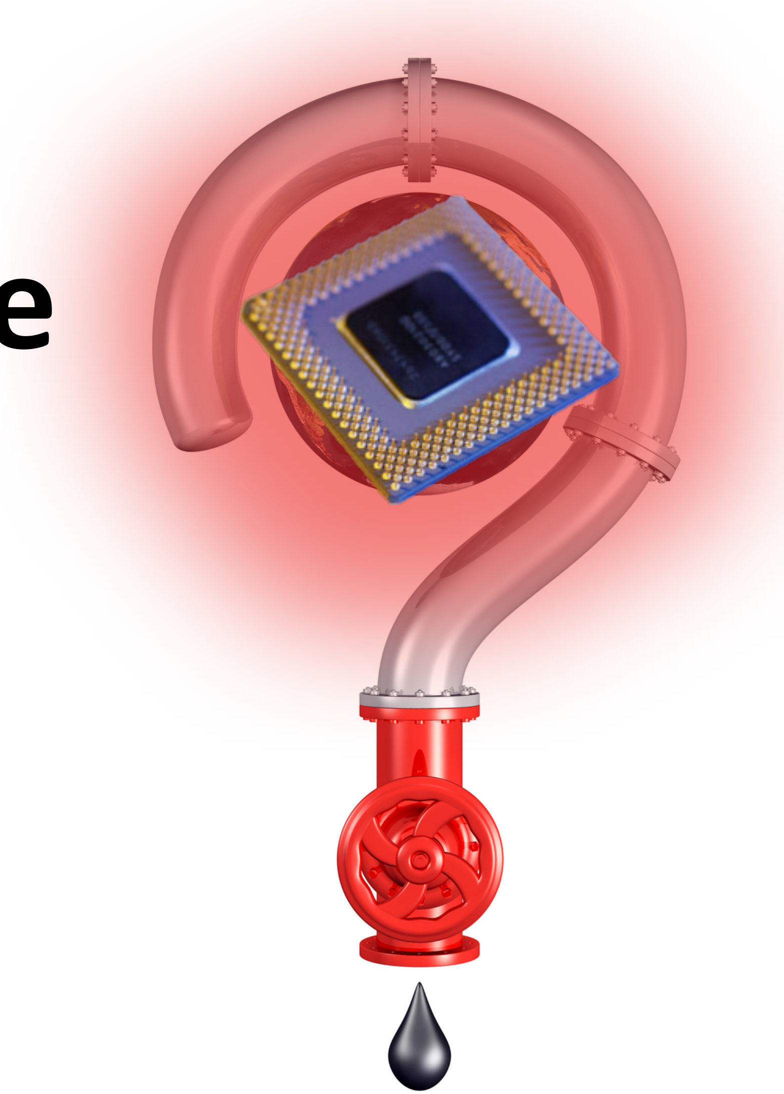
Big Data

IT's Future

Bridging

Technologies

Big Energy



ecocloud

Microsoft
vmware
HUAWEI
EATON Powering Business Worldwide
hp
intel
swisscom
CISCO

Center to bring efficiency to data

- 15 faculty, 50 researchers
- Around $6M/year budget

Mission:

- Energy-efficient data-centric IT
- From algorithms to machine infrastructure
  - E.g., Big Data analytics, integrated computing/cooling,…
- Maximizing Performance/TCO for Big Data

CREDIT SUISSE
ORACLE

EPFL
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

ecocloud.ch

Our Team

Faculty

Aberer

Ailamaki

Argyraki

Atienza

Bugnion

Candea

Cevher

Falsafi



Guerraoui

Koch

Larus

Lenstra

Odersky

Thome

Zwaenepoel

Executive Director


Diallo

Staff


Locca

+50 Researchers

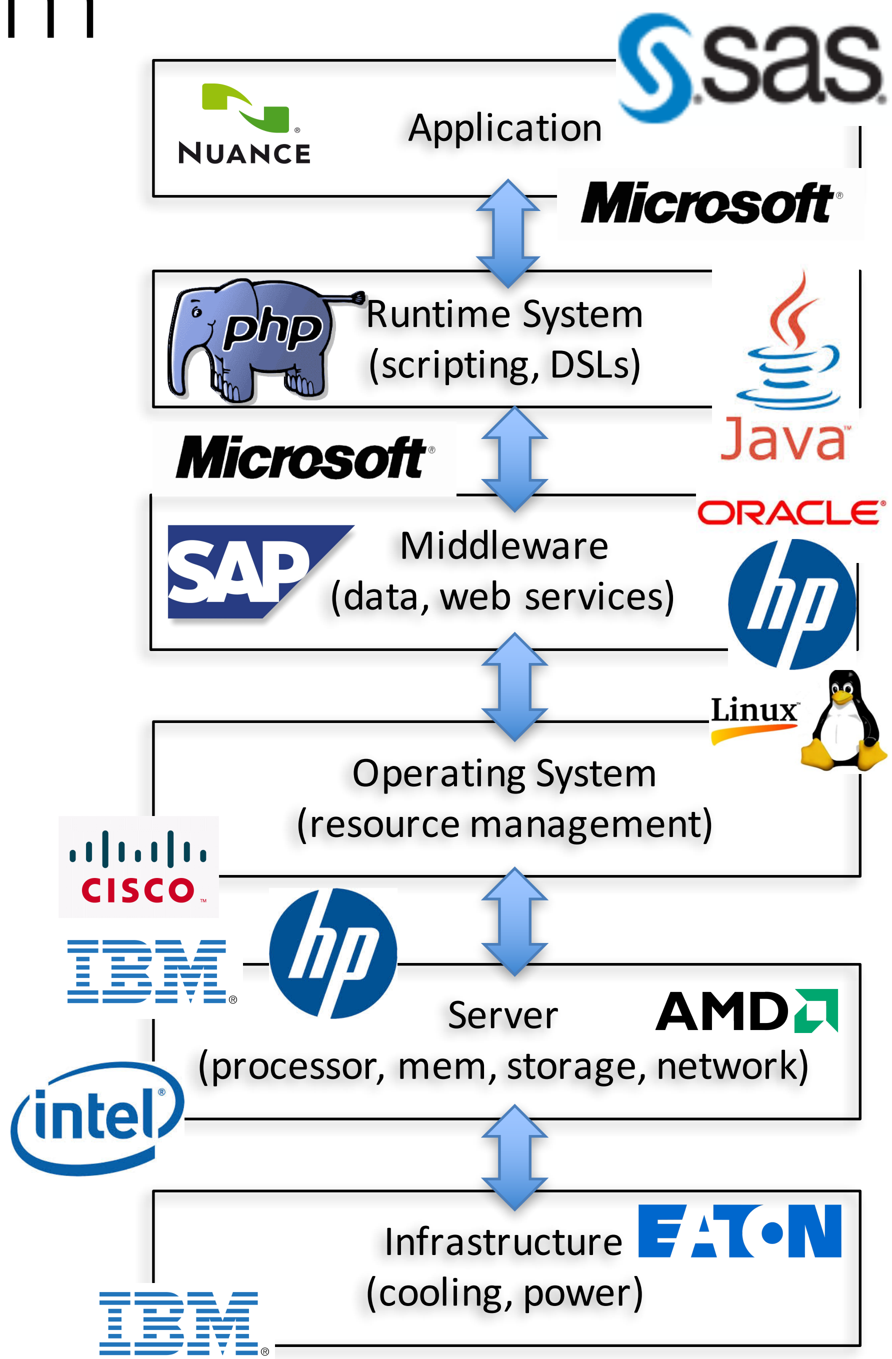
Today's Server Ecosystem

Conventional IT:

- Product based
- Per-vendor layer
- Well-defined interfaces
- Near-neighbor optimization at best

Big vendors (e.g., Amazon, Google)

- Can do cross-layer optimizations
- But,
  - Only limited to services of interest
  - Maybe limited in extent (e.g., software)
  - Proprietary technologies
  - Host all data

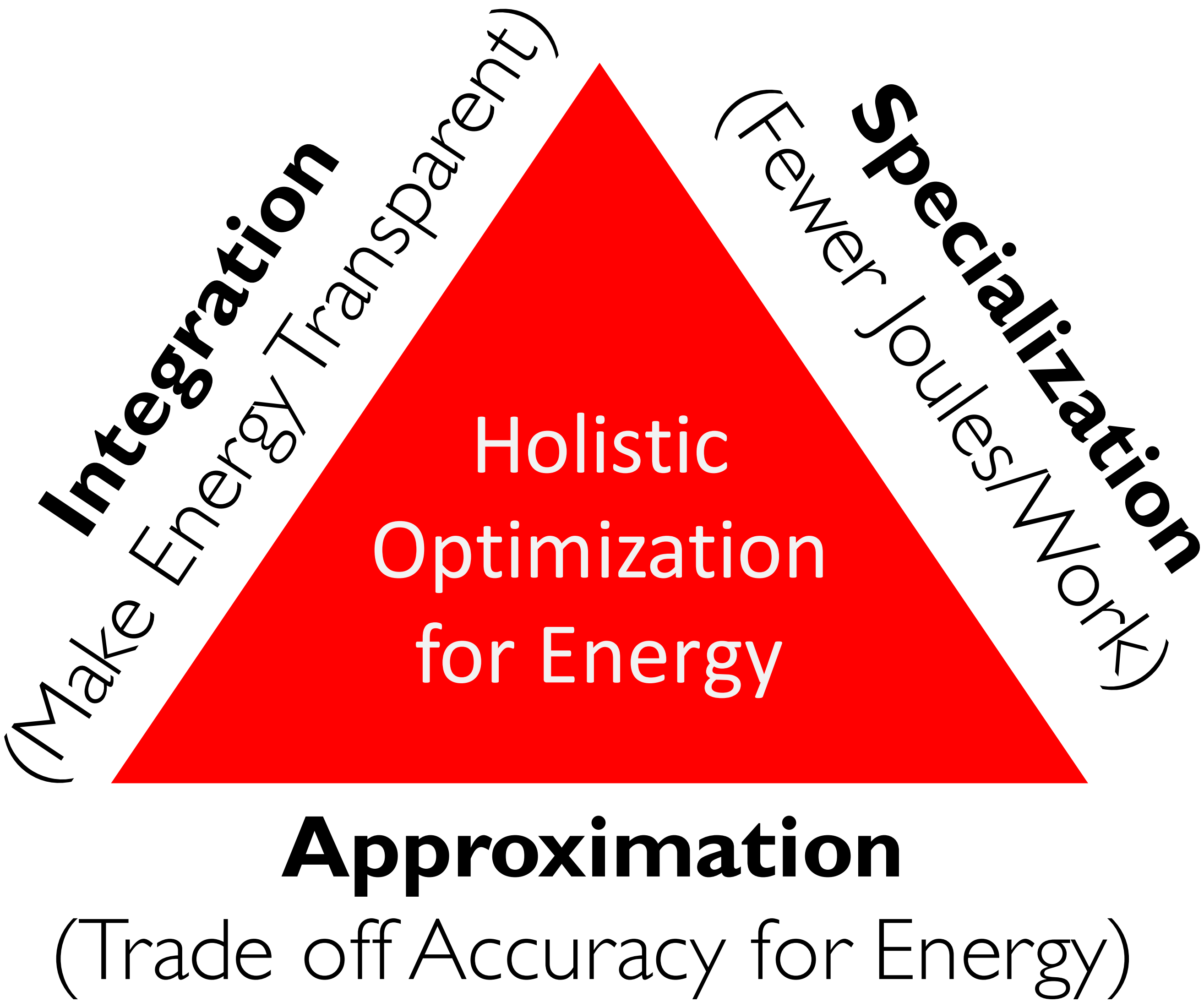


Our Vision:
Holistic Optimization of IT Infrastructure

Holistic optimization

- From Algorithms to Infrastructure
- Novel energy-centric IT paradigms
- Strategic interfaces to monitor, manage & reduce energy as a first class resource

Our Vision:
The ISA Triangle of Efficiency



ecocloud EPFL

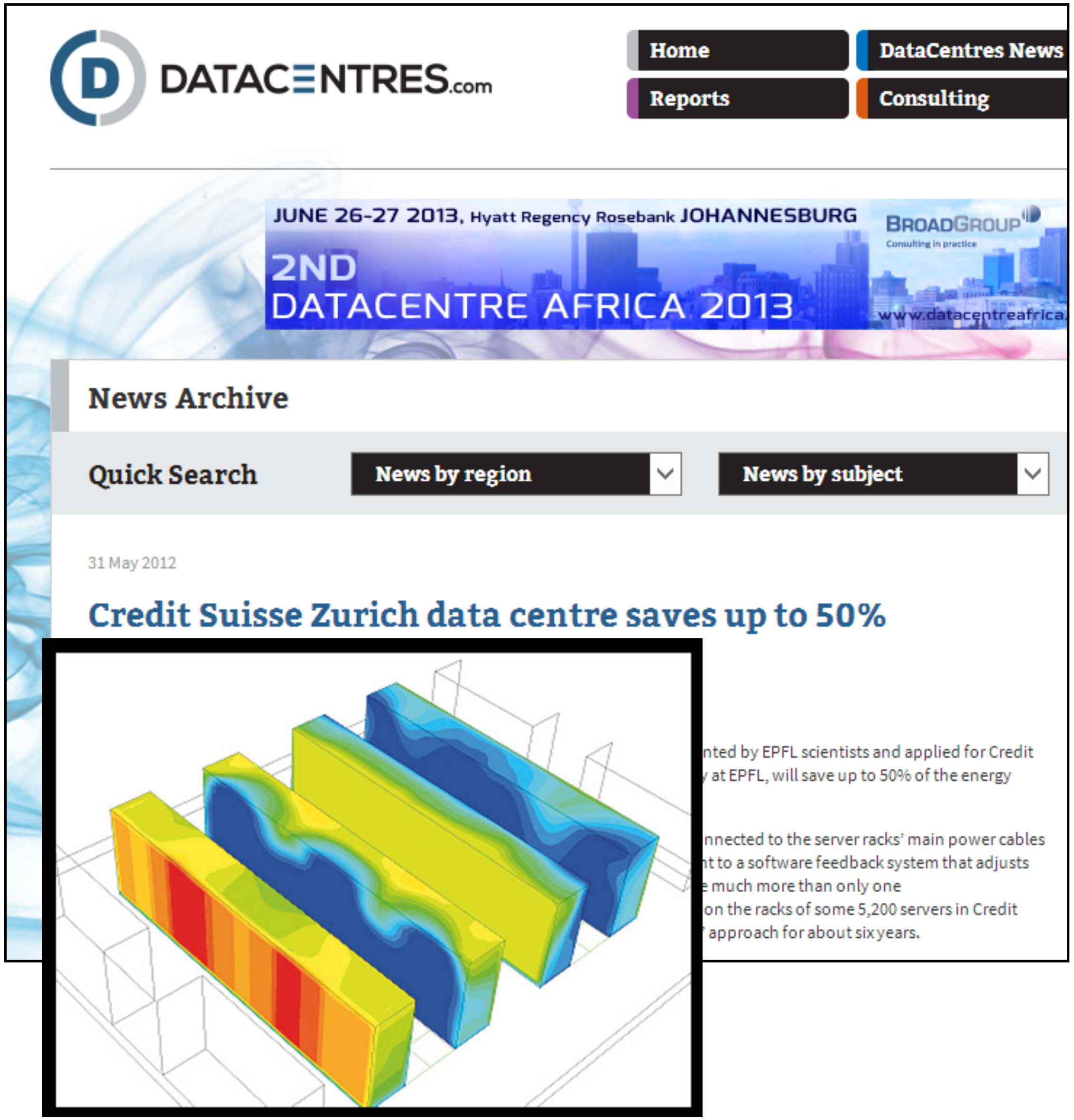
Integrated Thermal & Load Balancing

Project PMSM

- Synergistic IT load/thermal control
- Real-time monitoring of 5K servers
- Fine-grain power/thermal sensors

50% energy savings!

CREDIT SUISSE



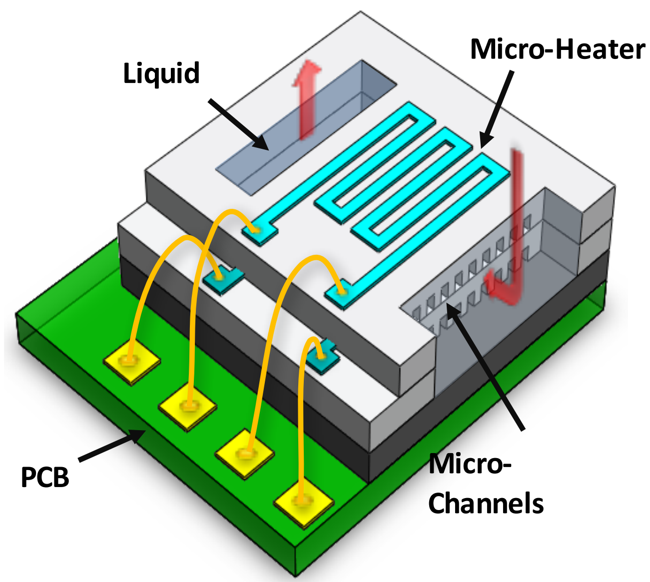
Integrated Cooling: CMOSAIC

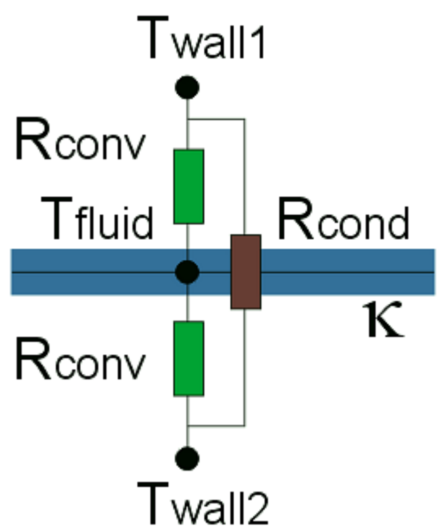
3D server chip

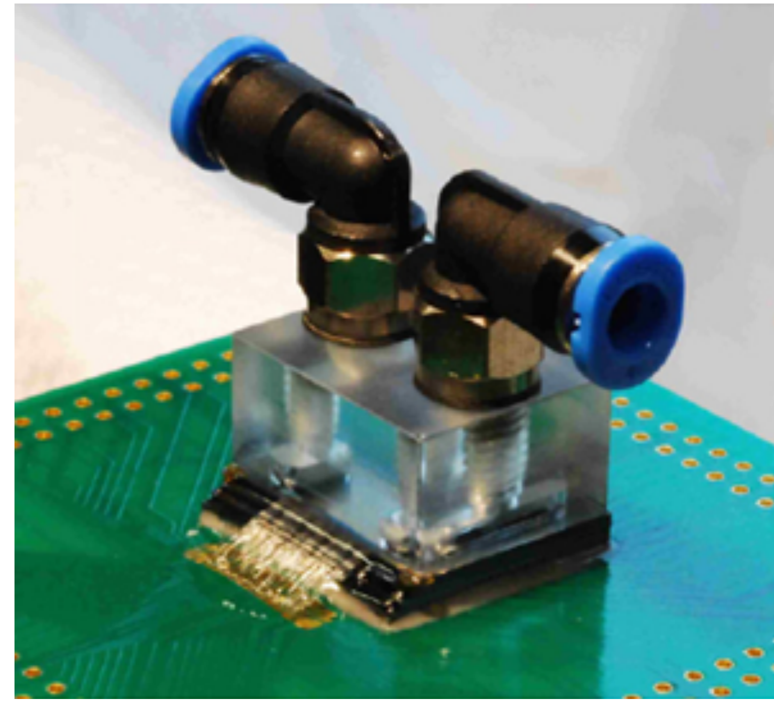
Two-phase liquid cooling

– Enables higher thermals
– Dramatically better heat removal

Prototyped by IBM







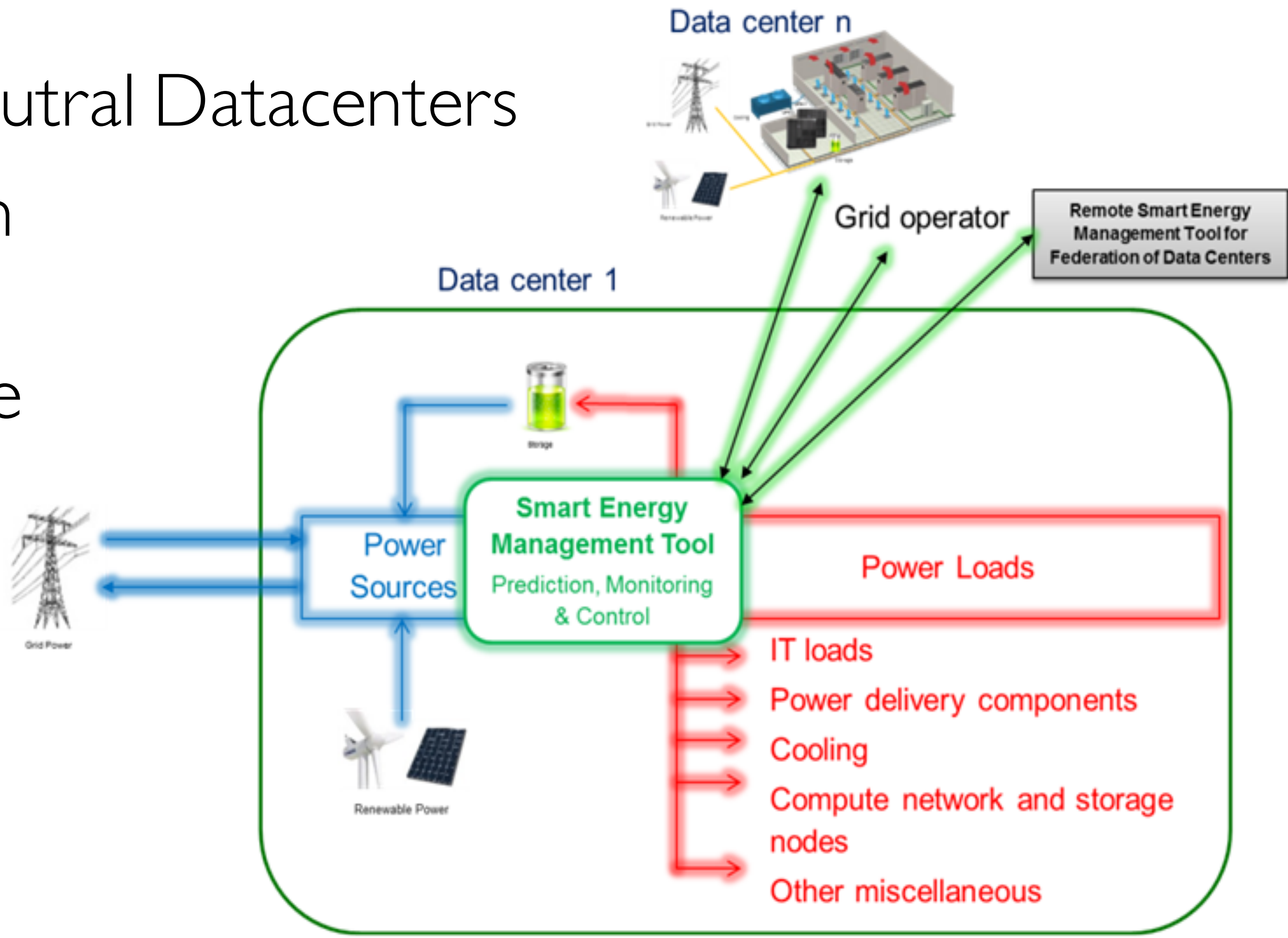
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Integrated Power Subsystem: GreenDataNet

Green Data Net

Towards Energy-Neutral Datacenters

- Power generation + power storage + server resource provisioning
- Federated sites
- Grid load management



ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

NISSAN

SEVENTH FRAMEWORK PROGRAMME

EATON Powering Business Worldwide

CREDIT SUISSE

DE LA RECHERCHE À L'INDUSTRIE cea

Scale-Out Datacenters

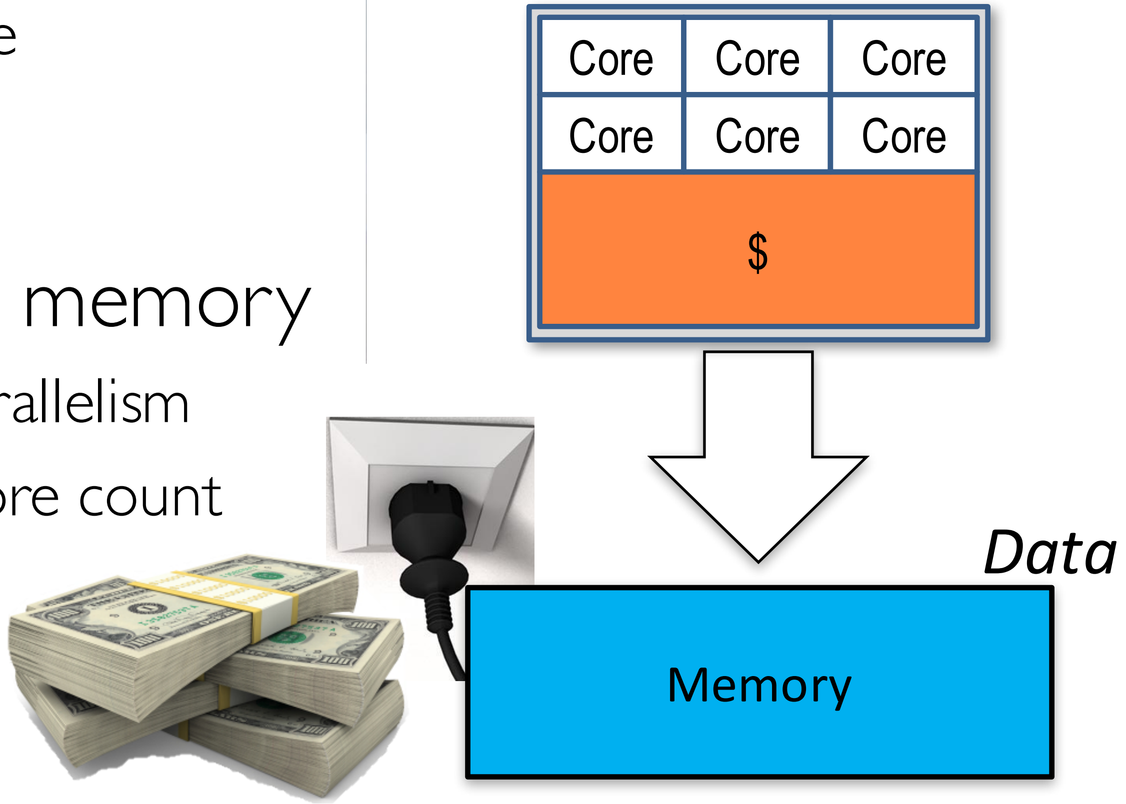
Vast data sharded across servers

Memory-resident workloads

- Necessary for performance
- Major TCO burden

Processors access data in memory

- Abundant request-level parallelism
- Performance scales with core count



Design servers around memory!

How efficient are servers for in-memory apps?

CloudSuite 2.0 (parsa.epfl.ch/cloudsuite)

Data Analytics
Machine learning

Data Caching
Memcached

Data Serving
Cassandra NoSQL

Graph Analytics
TunkRank

Media Streaming
Apple Quicktime Server

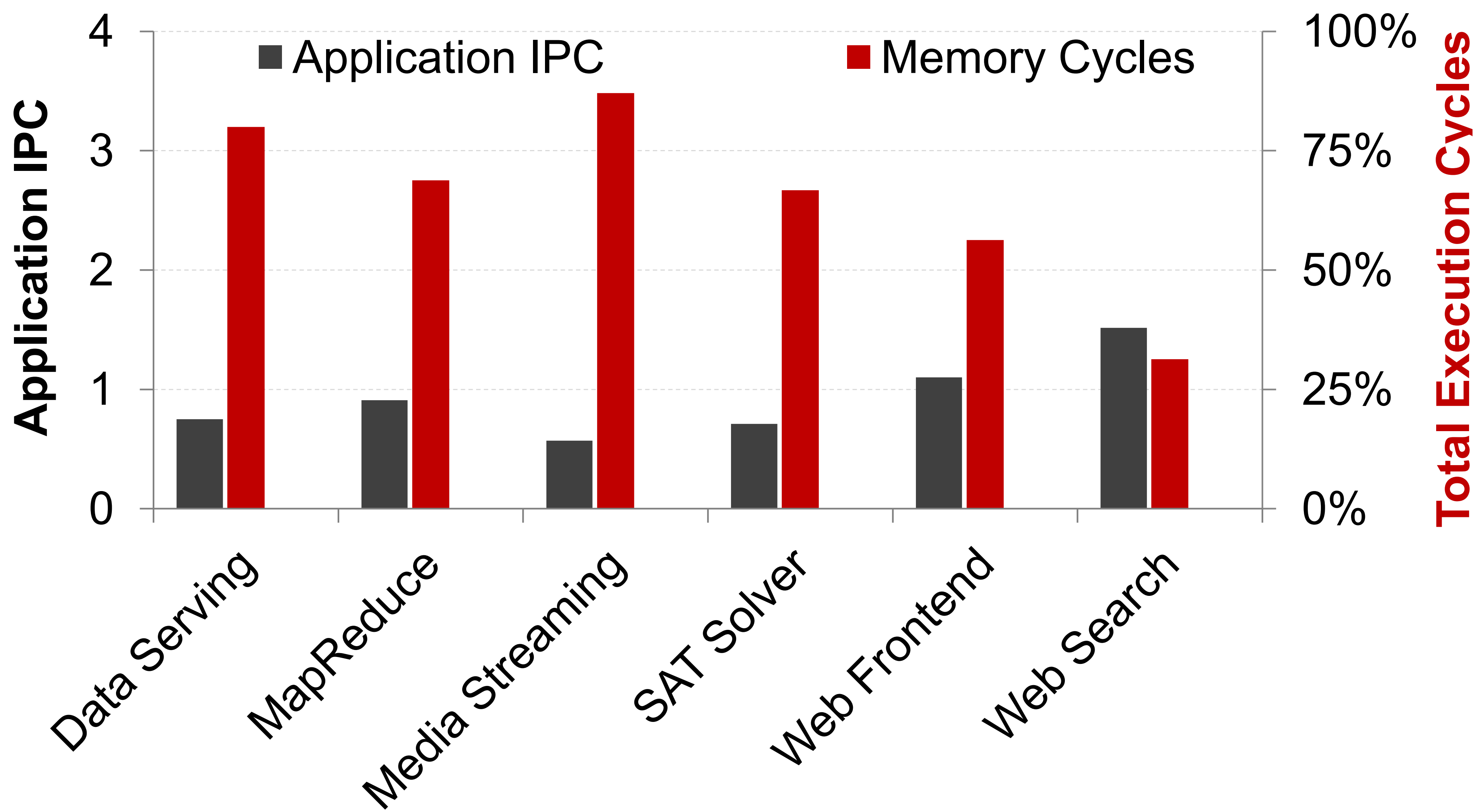
SW Testing as a Service
Symbolic constraint solver

Web Search
Apache Nutch

Web Serving
Nginx, PHP server

In Use by AMD, Huawei, HP, Intel, Google….

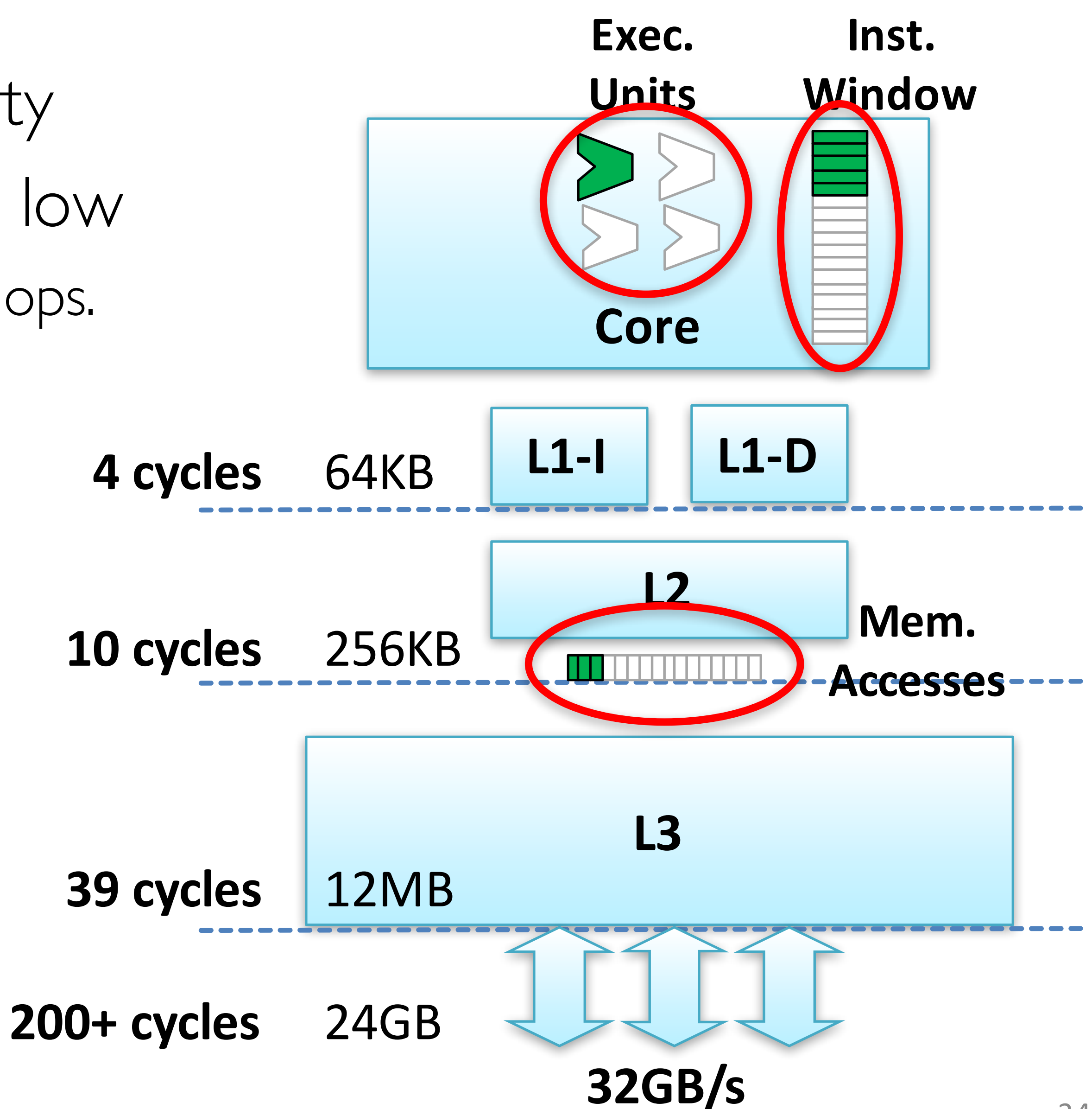
Big Data Workloads Stuck in Memory!



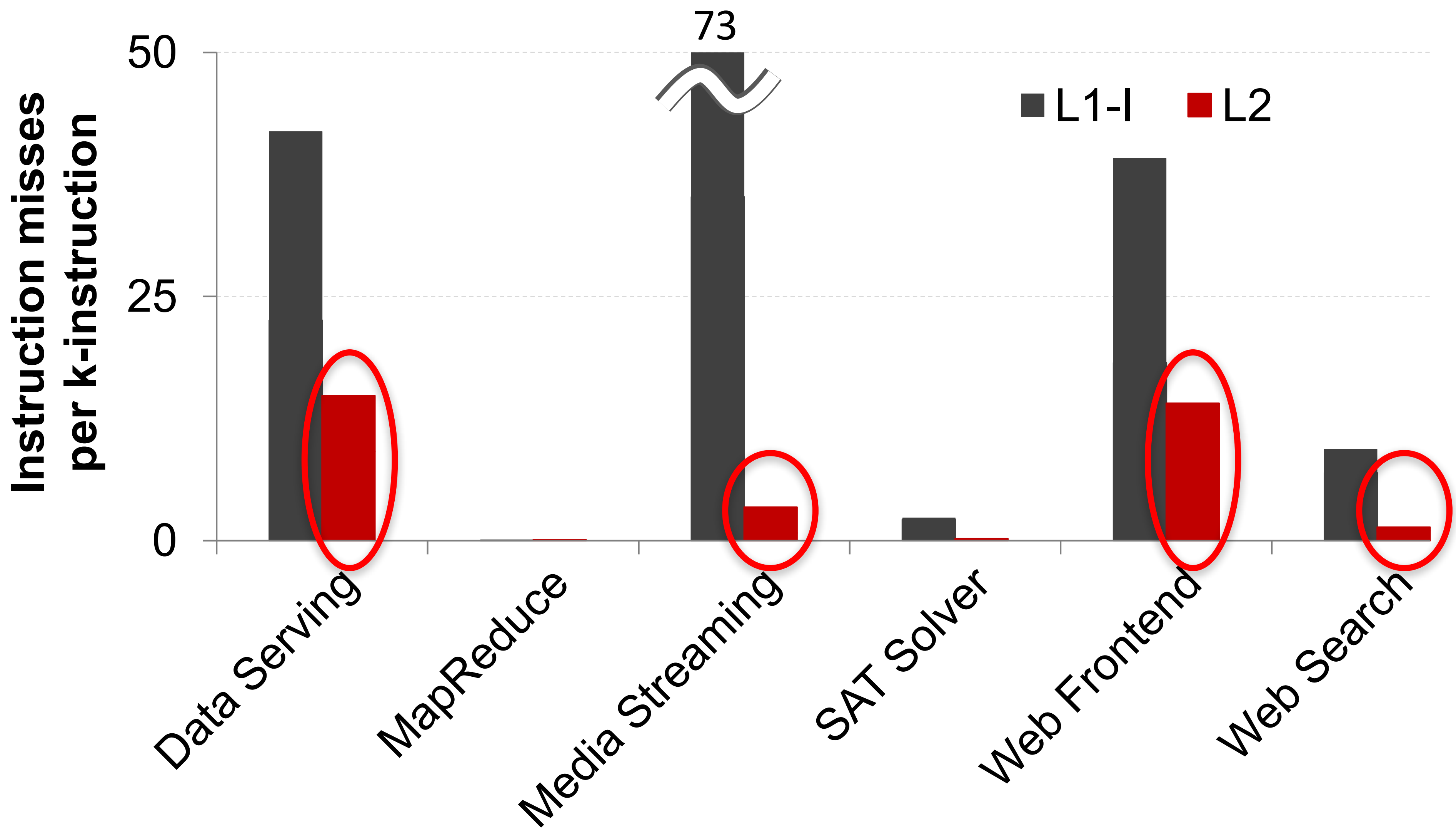
Execute ~1 instruction per cycle

Core Inefficiencies

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



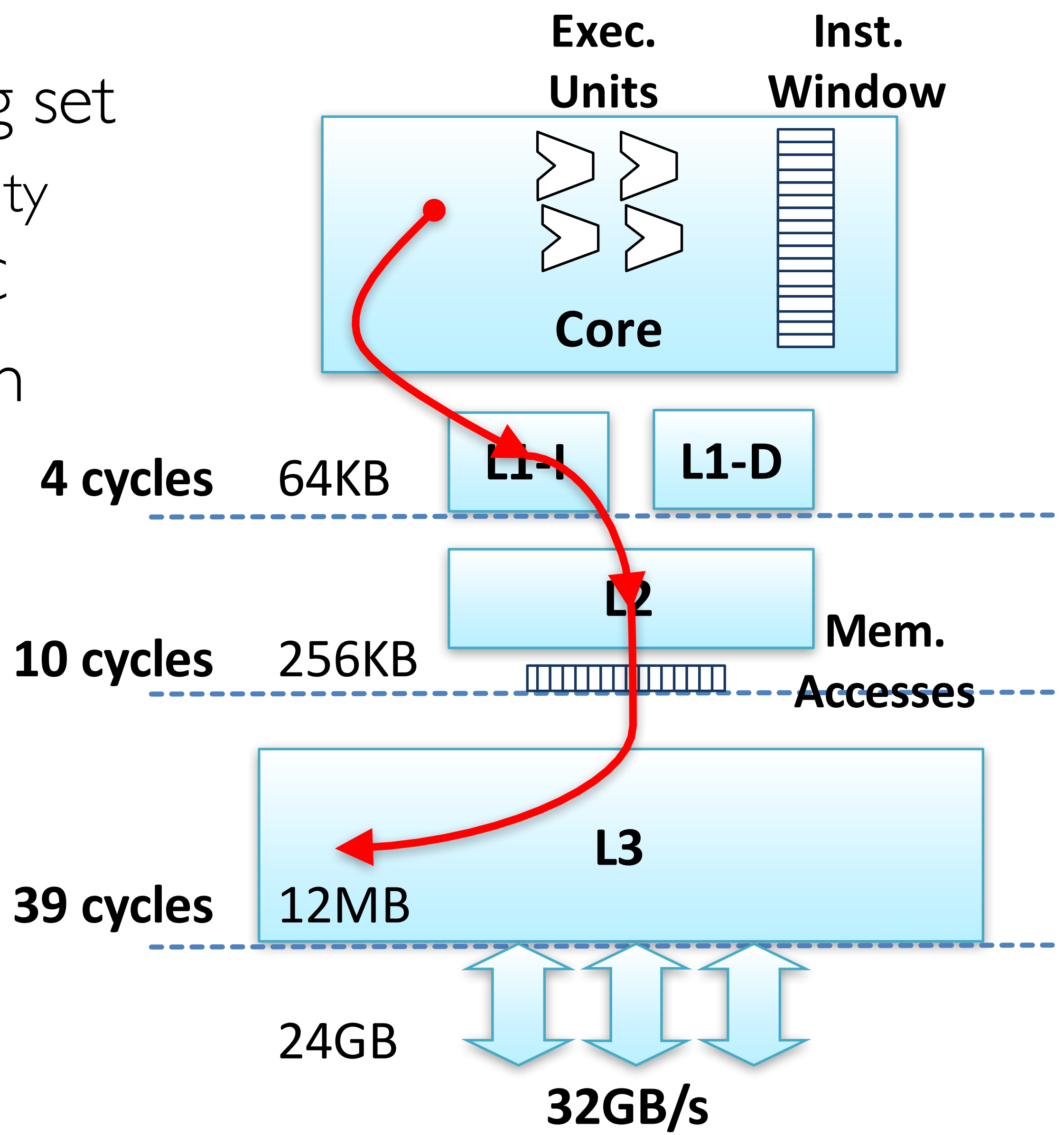
Instruction-Fetch Misses



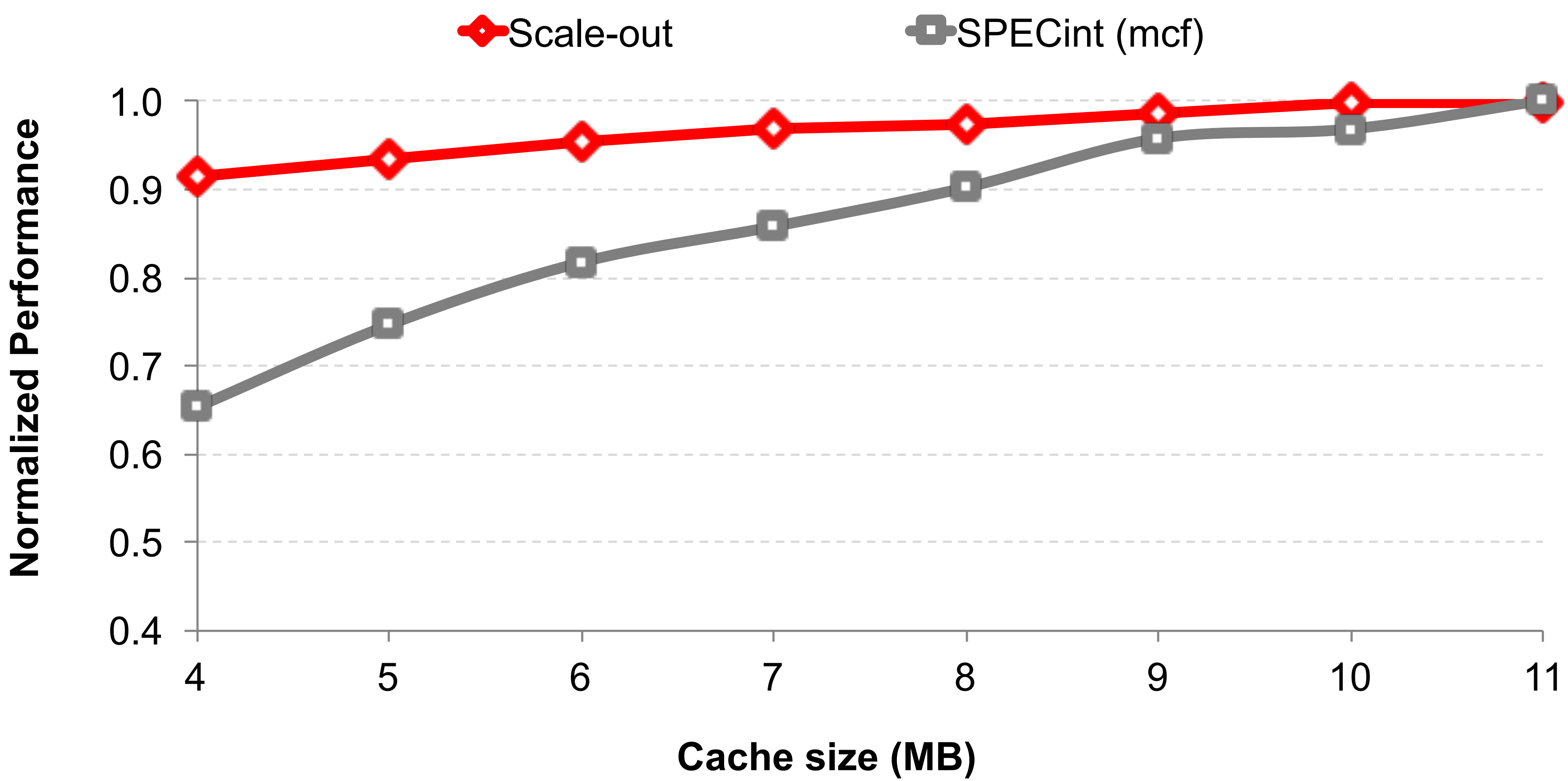
Suffer severe i-cache miss penalties

Instruction-Fetch Inefficiencies

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

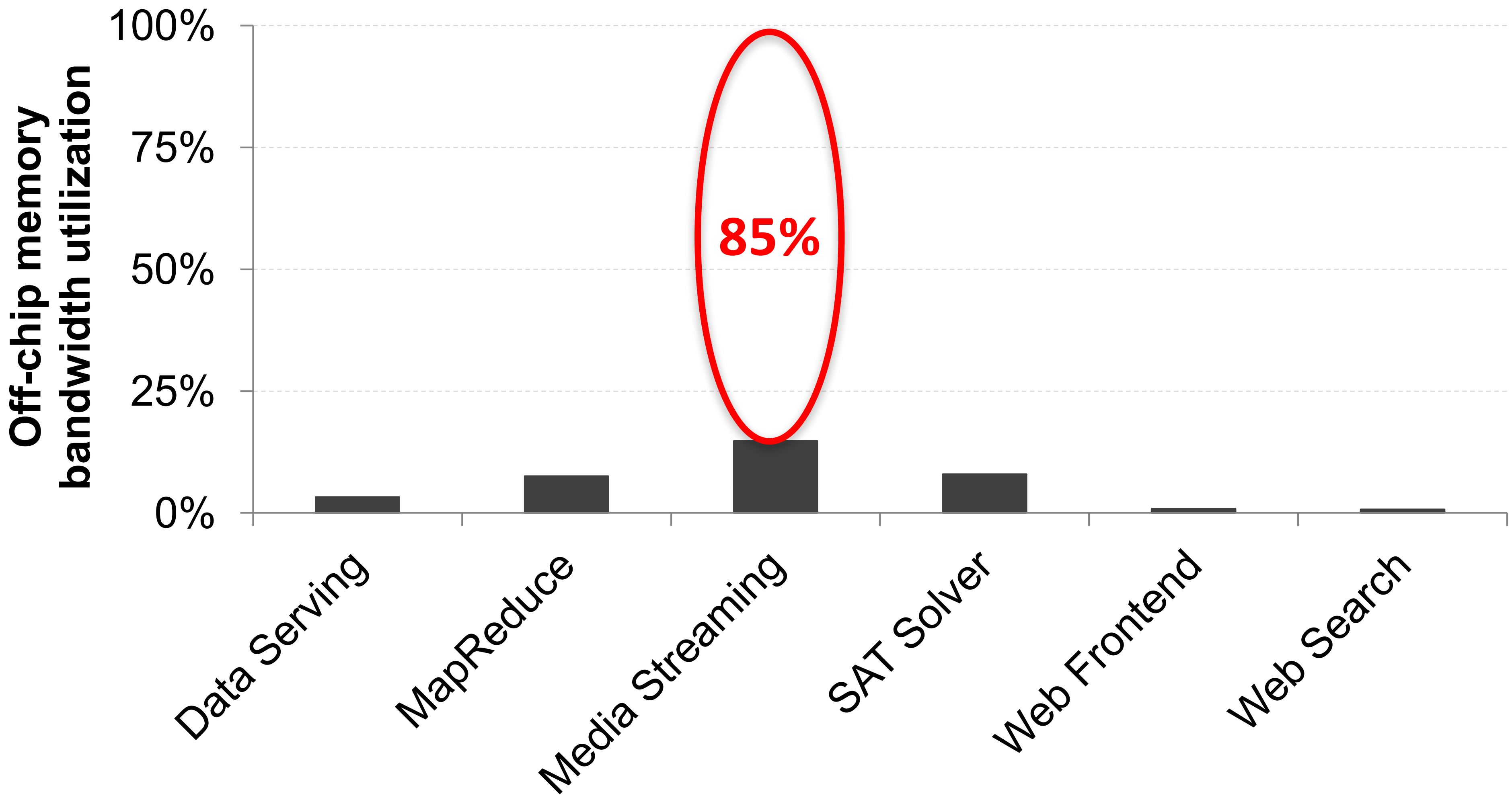


LLC Sensitivity



Minimal performance from large LLC

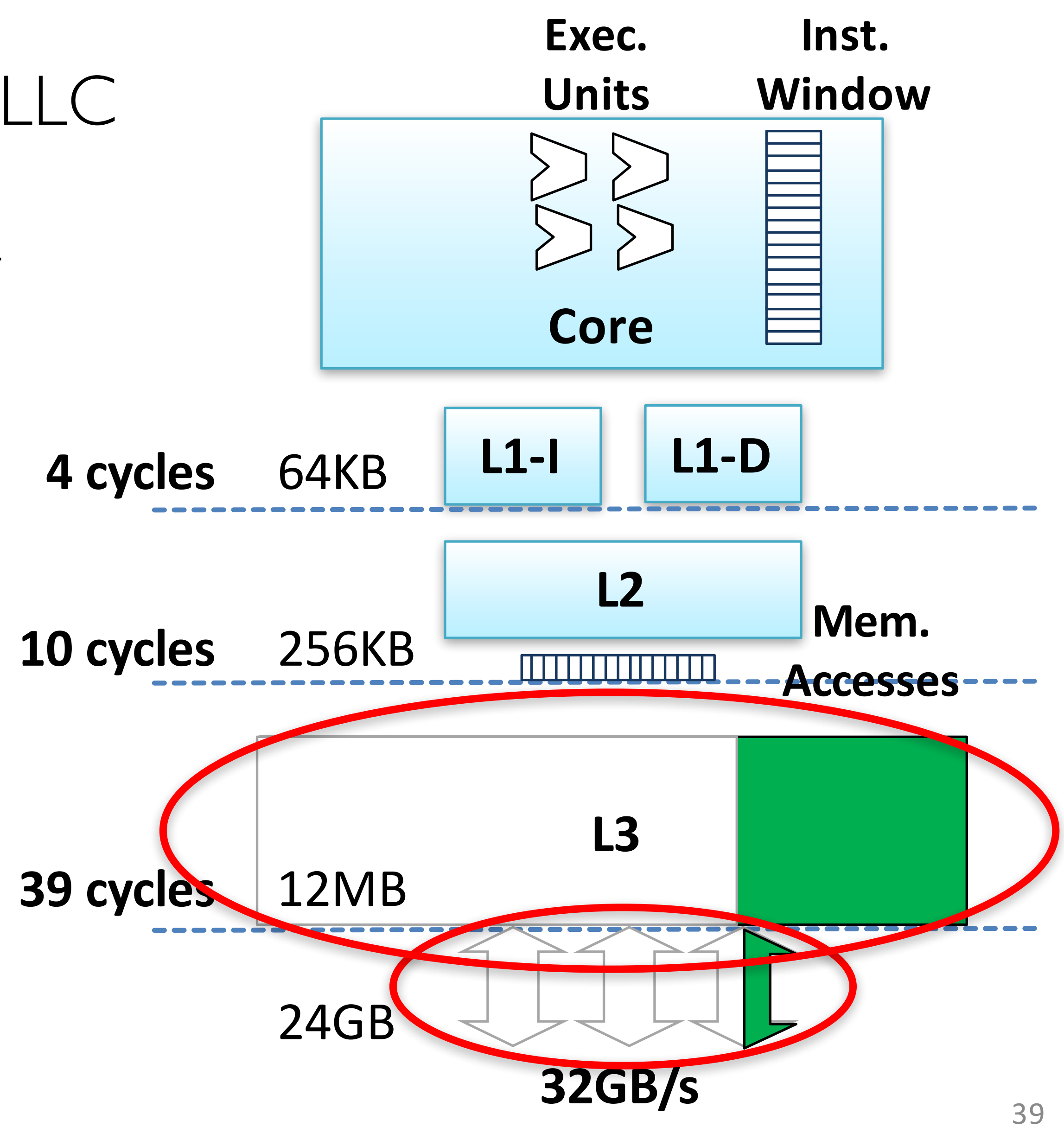
Off-chip Memory Bandwidth



Off-chip BW severely underutilized

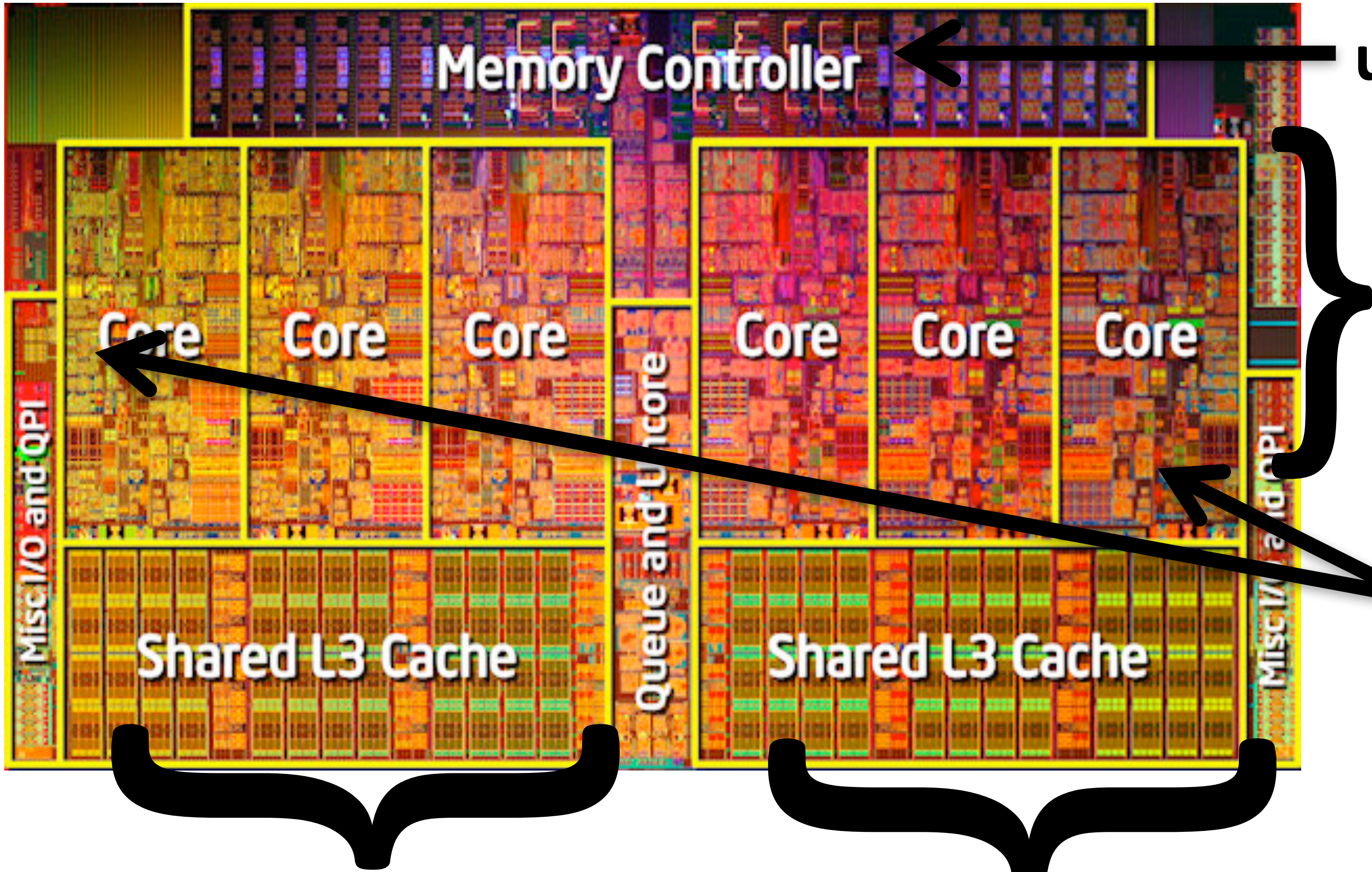
LLC and Bandwidth Inefficiencies

- 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



CloudSuite on Modern Servers [ASPLOS'12, best paper]

Workload/Server Mismatch



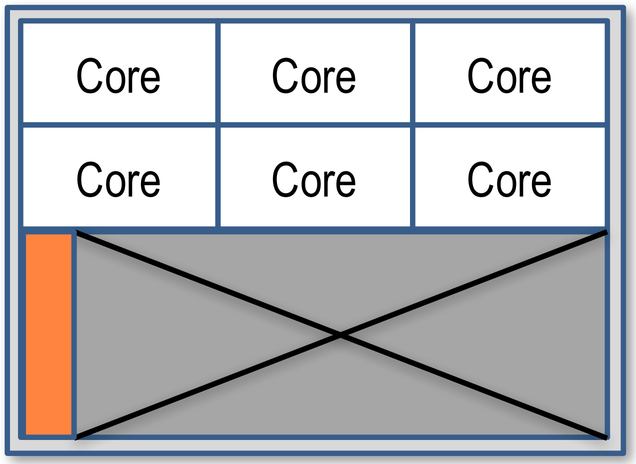
B/W unused!

Cores too fat!

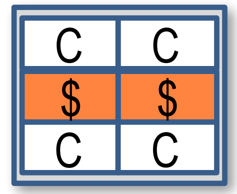
Too few cores!

8 MB (60%) waste of space (no reuse)!

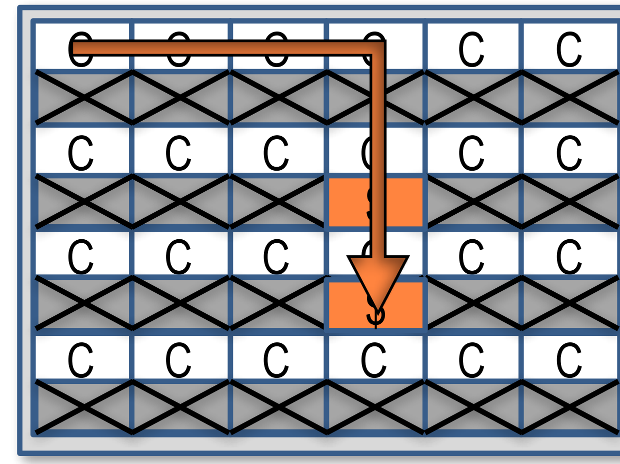
What do Existing Processors Offer?



Intel Xeon (~100 W)



Calxeda (~5W)



Tilera (~30W)

- ✘ Few fat cores
- ✘ Large LLC

- ✘ Few lean cores
- ✓ Compact LLC

- ✓ Many lean cores
- ✘ Large LLC
- ✘ Large distance

Mismatch with workload demands!

Specialized Processors for In-Memory Services: **Scale-Out Processors** [ISCA'12, IEEE Micro'12]

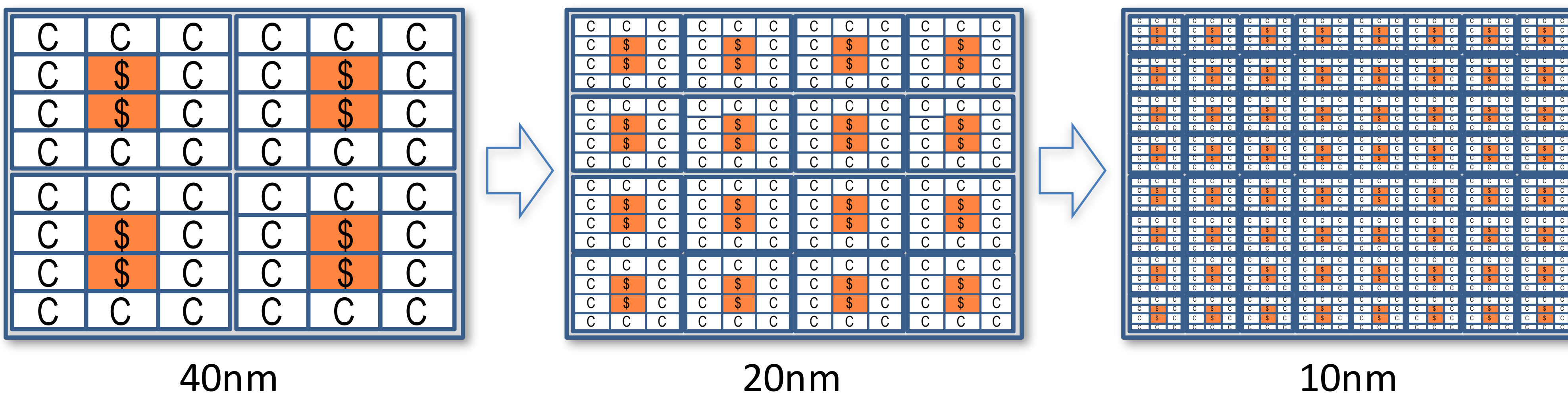
One or more stand-alone (physical) servers

- Runs a full software stack

No inter-pod connectivity or coherence

- Scalability and optimality across generations

Pods can share chip I/O (e.g., memory, network, etc.)



Inherently Software Scalable!

NOC-Out: [Micro'12]

Specialized Network-on-Chip for Pods

Exactly the **opposite** of current NoCs

- Cache coherent
- But, designed for core-to-cache communication
- Not core-to-core!

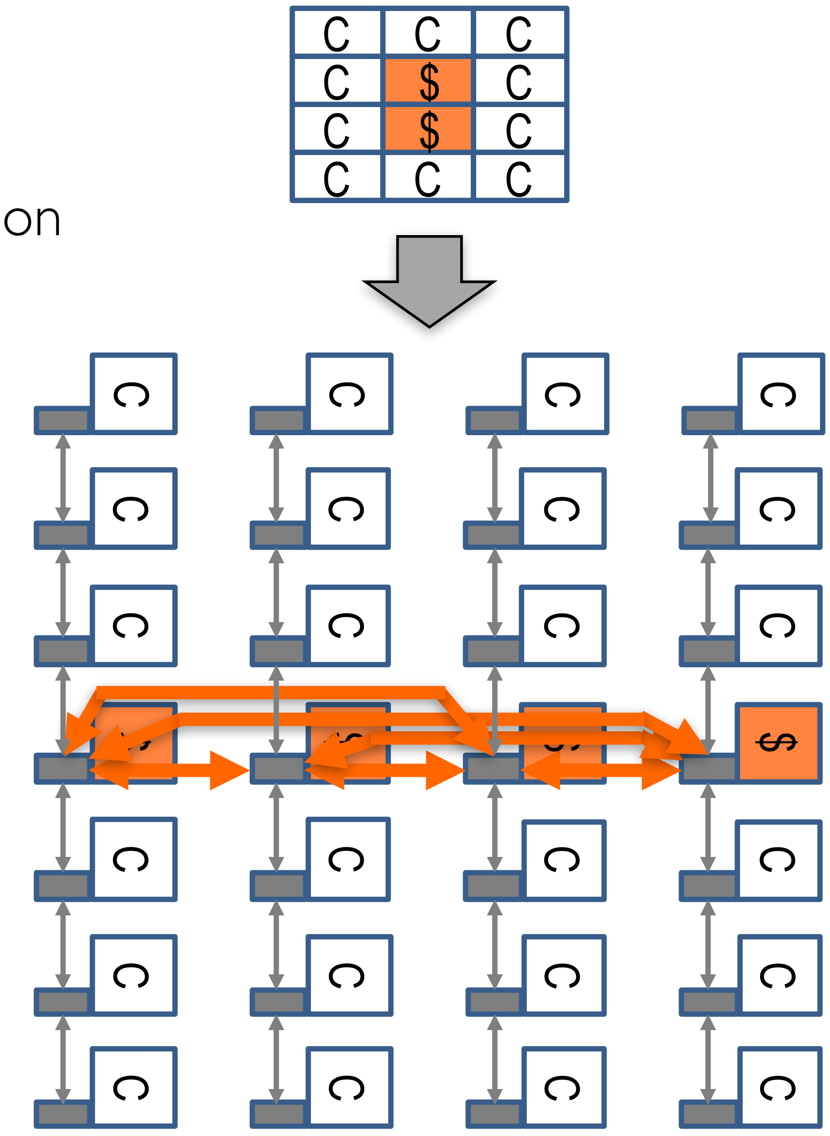
LLC network:

- Flattened Butterfly (FB) topology

Request & *Reply* networks:

- Tree topology
- Limited connectivity for efficiency

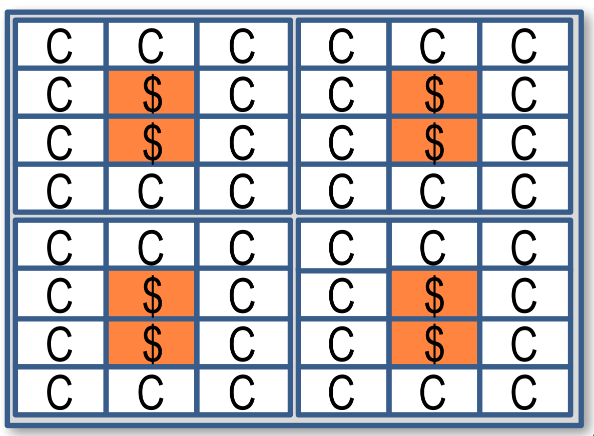
FB's performance at 1/10th cost

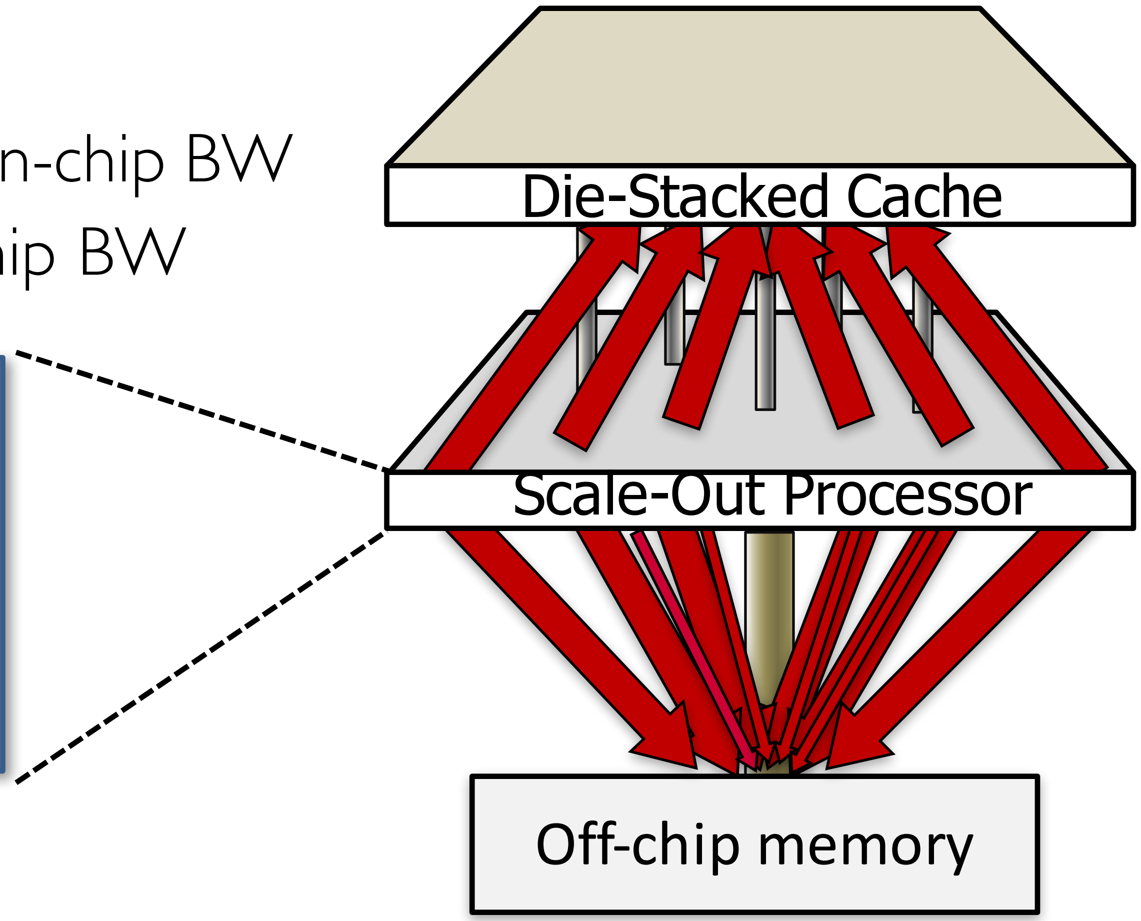


Footprint Cache: [ISCA'13]
Effective Die-Stacked Caching for Pods

Die-Stacked Caching:

- Rich connectivity → High on-chip BW
- High capacity → Low off-chip BW





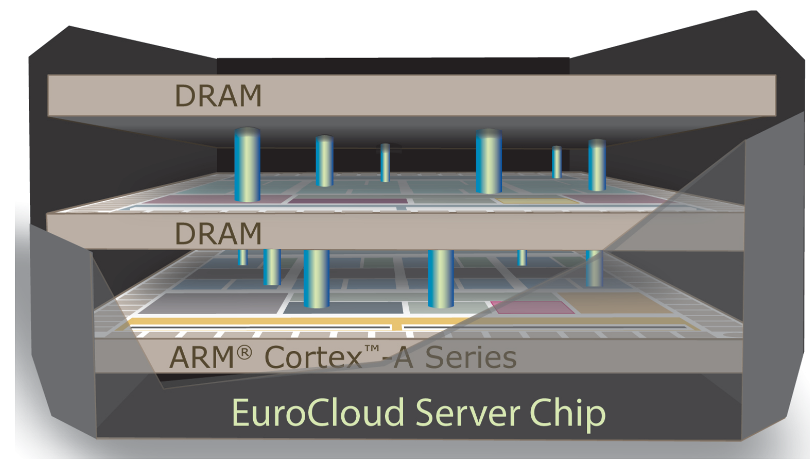
Footprint Cache:

- Allocate tags for pages
- Predict & fetch page's footprint

EuroCloud Server: (eurocloudserver.com)
3D Scale-Out Chip for In-Memory Computing

Mobile efficiency in servers

- Swarms of ARM cores
- 3D memory
- 10x performance/TCO
- Runs Linux LAMP stack



Planned prototype:

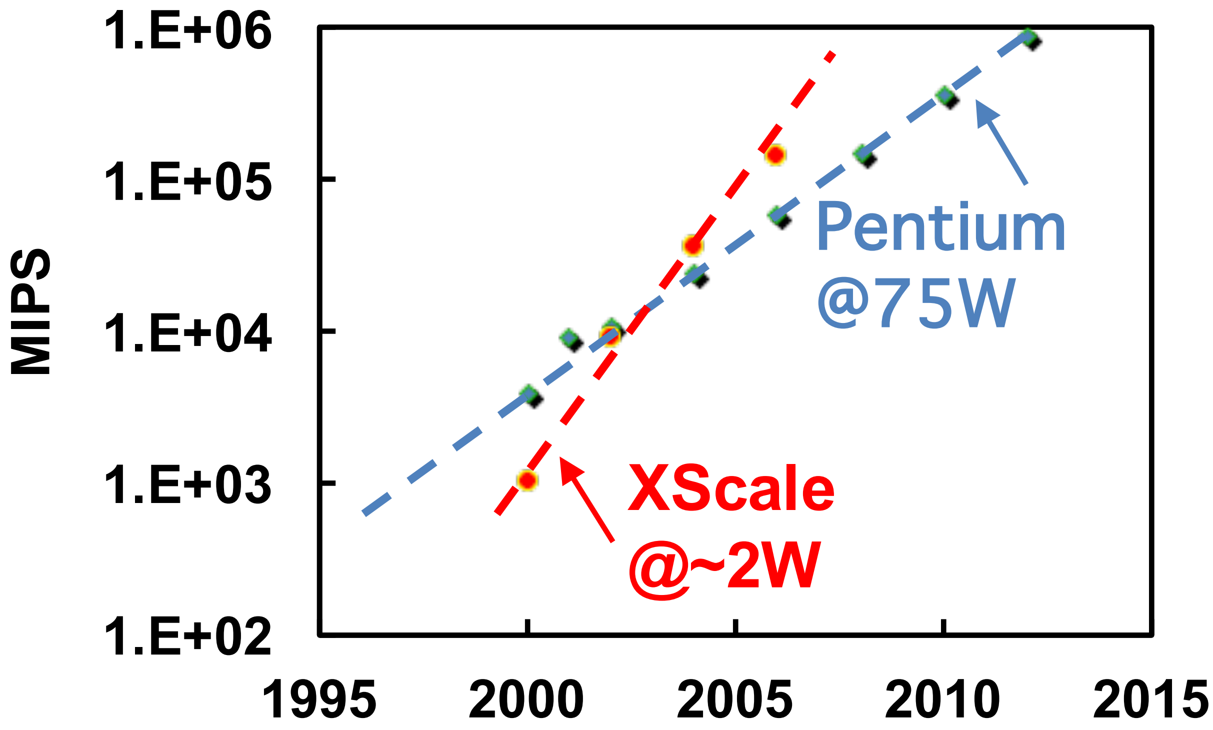
- ARM/ST/cea + Chalmers/FORTH in EuroServer FP7
- Data Processing Unit by Huawei

SEVENTH FRAMEWORK PROGRAMME   ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE   imec   ARM

Flashback 2004:
Shekhar Borkar's (Intel Fellow) Keynote @ Micro

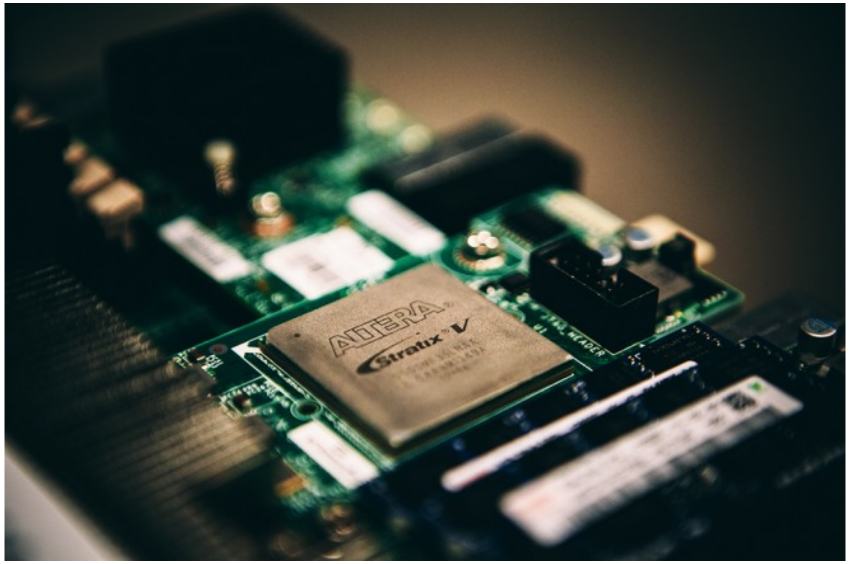
Intel's TCP/IP Processor

2.23 mm X 3.54 mm,
260K transistors



An idea too early for its time?

Specialization: An idea whose time has come



- *One FPGA per blade*
- *All FPGAS connected in half rack*
- *6×8 2-D torus topology*
- *High-end Stratix V FPGAs*
- *Running Bing Kernels for feature extraction and machine learning*

Microsoft Unveils Catapult to Accelerate Bing!

[EcoCloud Annual Event, June 5th, 2014]

Specialized Database Stack: DBToaster

TOASTER

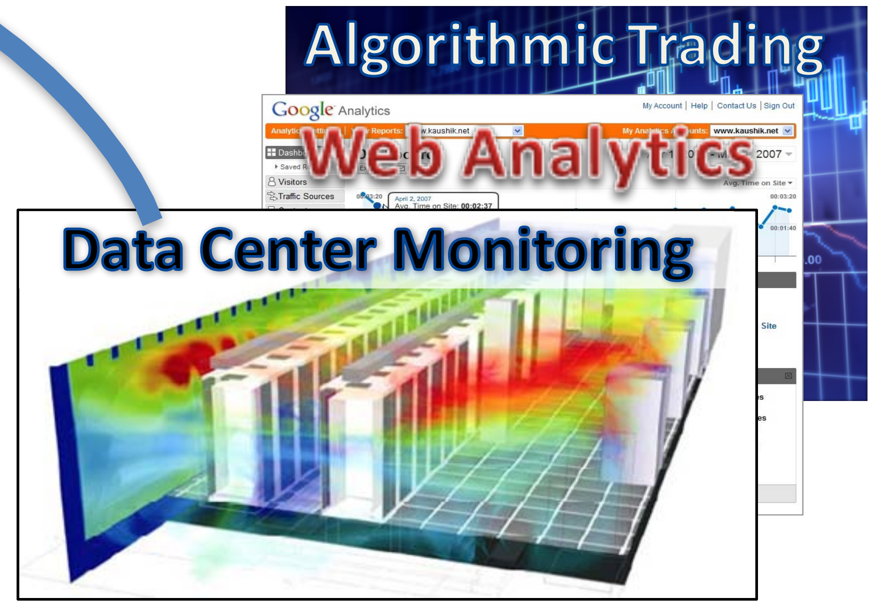
Compiling offline analytics into online/incremental engines

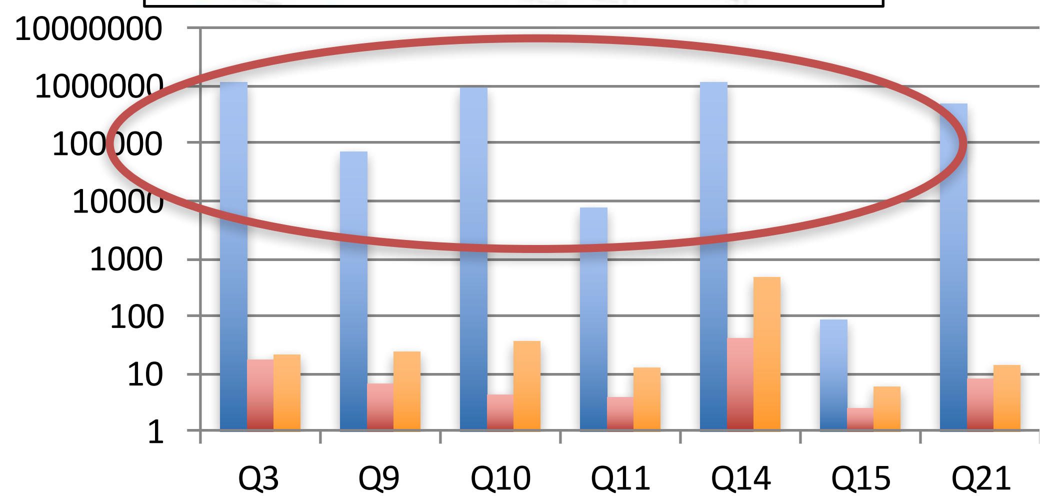
Aggressive code specialization

Data Stream

Low-latency in-memory stream processing

Up to 6 OOM faster than commercial systems

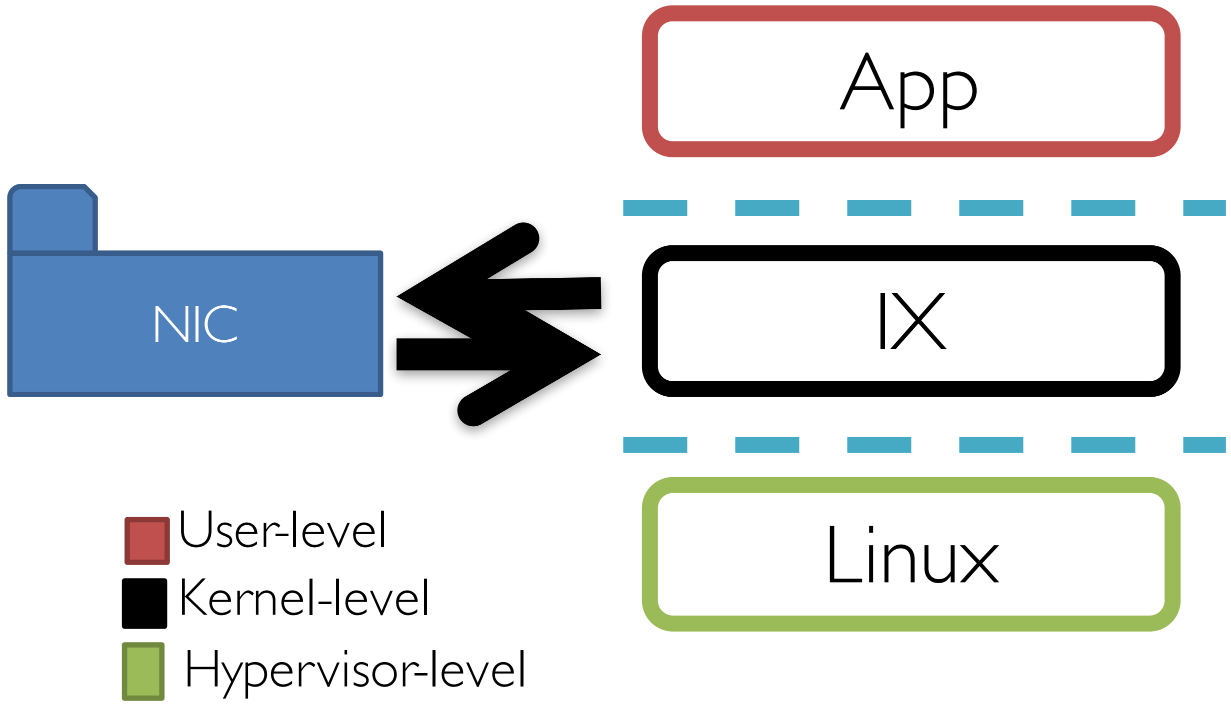




dbtoaster.org

Specialized Network Stack: The IX kernel [Belay'14, OSDI best paper]

- Data plane principles: zero-copy, run-to-completion, coherence free
- Protected operating system with clean-slate API
- Specialized for **in-memory** event-driven applications



3.6x throughput with <50% latency @ 99^{th} percentile

Today's Network Fabrics Bottleneck!

In-Memory Latency critical services

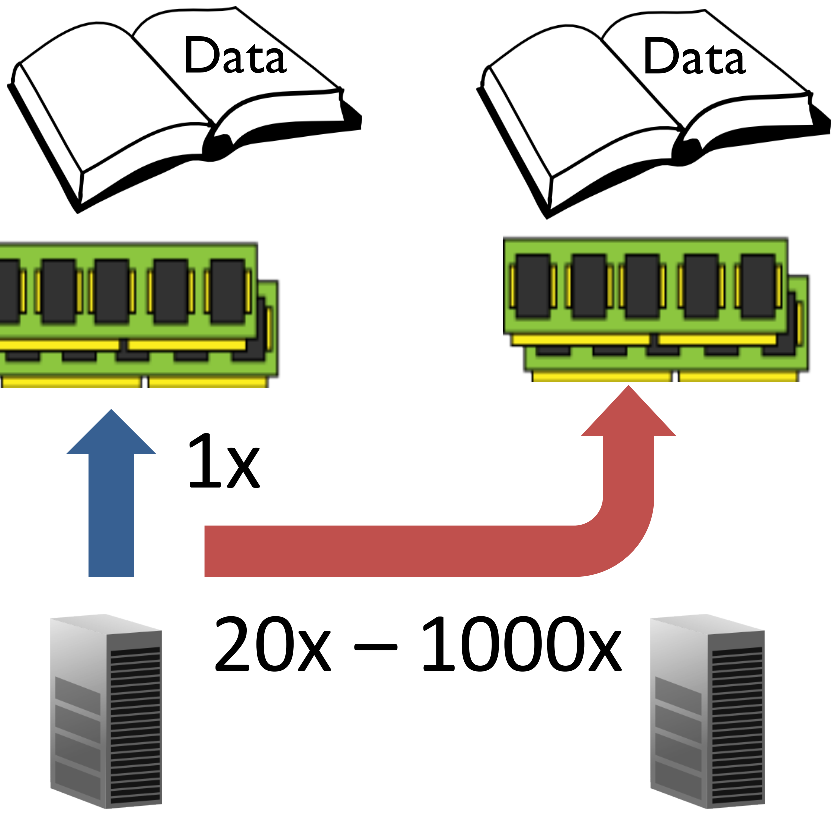
– Graphs, KV, DB

Vast datasets → distribute

– Often within rack

Today's networks:

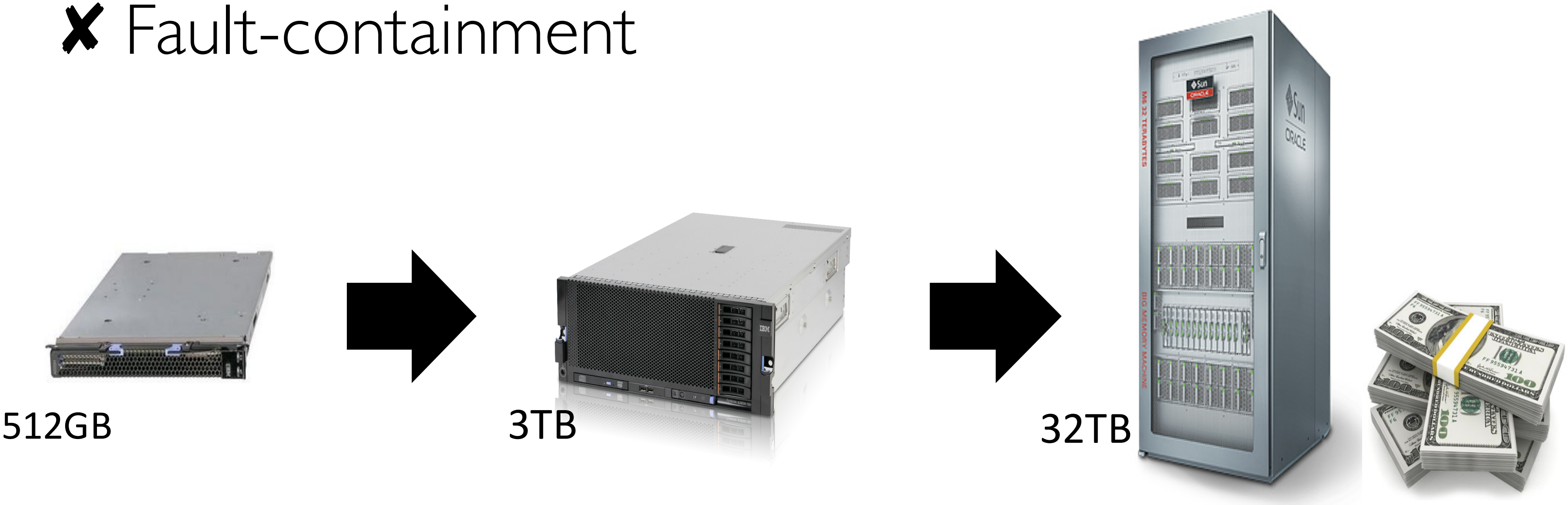
✘ Latency 20x-1000x of DRAM



Remote access latency >> local access latency

Big Data on ccNUMA: Expensive

- ✓ Ultra-low latency
- ✗ Cost and complexity of scaling up
- ✗ Fault-containment



Ultra low-latency but ultra expensive

Big Data on Commodity Fabrics: Slow

✓ Cost-effective rack-scale fabrics of SoCs

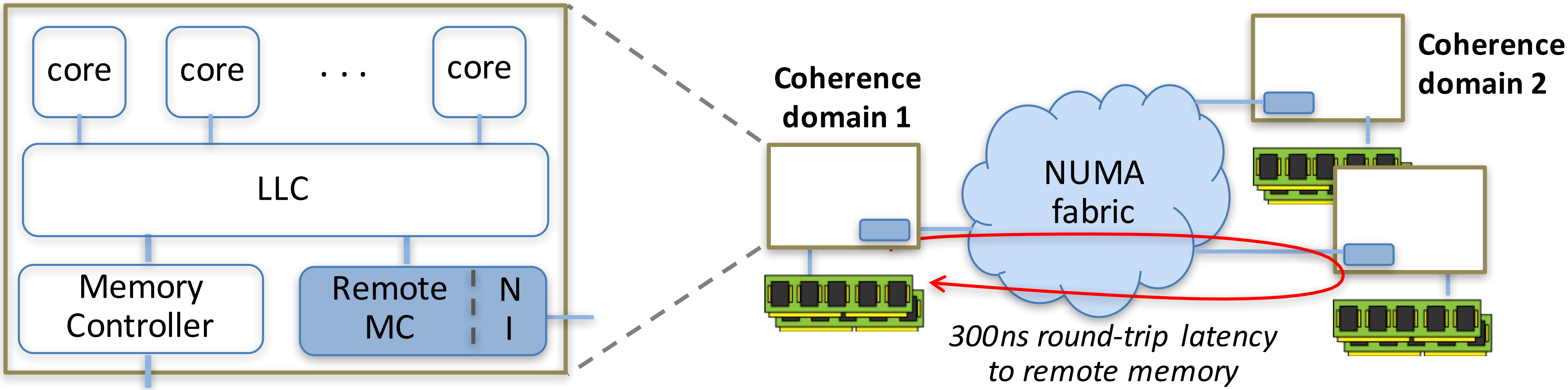
✗ High remote latency (~ >10 us)

AMD's SeaMicro

HP's Moonshot

Need low-latency rack-scale fabric!

Scale-Out NUMA (soNUMA): Rack-scale In-memory Computing [ASPLOS'14]



- Global virtual address space w/o global coherence
- RDMA-inspired programming model
  - Integrated Network Interface (NI)
  - Software Accessible Remote Memory Controller (RMC)
- Lean NUMA fabric
  - Reliable user-level messaging over a minimal protocol

A few words on Approximation

Data services are probabilistic

➔ Yet digital platforms are precise!

Much opportunity at the algorithmic/software level

- Learning algorithms (Cevher et. al.)
- Approximate querying (Koch et. al.)
- Programming (Rinard et. al.)

Architecture?

- Bad: von Neumann not best suited for approximation
  - Control path dominates energy
  - Dual datapath shown (Ceze et. al.) not useful
- Good: support for neural processing
  - Analog (Temam et. al.) or Digital (Esmailizadeh et. al.)

Summary

Two IT trends on a collision course:

- Data growing at ~10x/year
- Nearing end of Dennard & Multicore Scaling
- Need technologies to bring efficiency to data

Moving away from products to services

- Future opportunities are in cross-layer design

Long term:

Integrate + **S**pecialize + **A**pproximate (ISA for Big Data)

Thank You!

For more information please visit us at

ecocloud.ch

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE