Memory-Centric Server Architecture

Babak Falsafi
Director, EcoCloud
ecocloud.ch

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

ecocloud
an EPFL research center

A Brief History of IT

ecocloud
an EPFL research center

Mobile Era

Consumer Era

1970s- 1980s 1990s Today+

Mainframes

PC Era

- From computing-centric to data-centric
- Consumer Era: Internet-of-Things in the Cloud

The future of IT is Data

ecocloud
an EPFL research center

- 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!

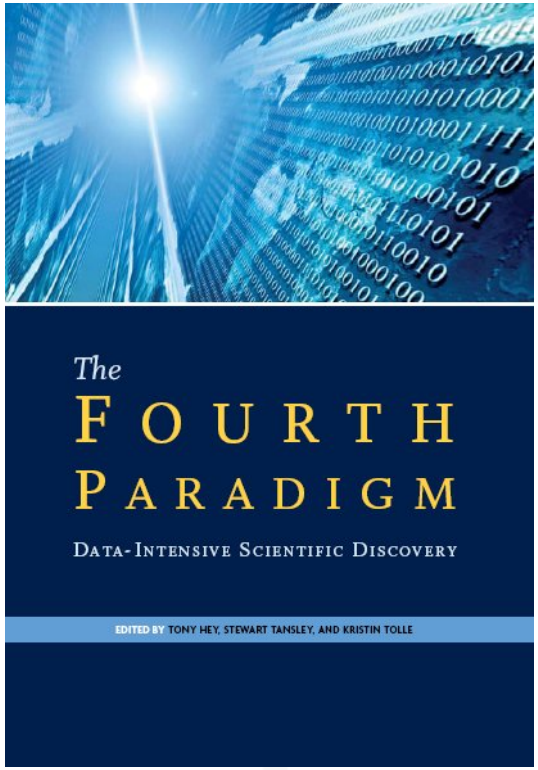
Data Shaping All Science & Technology

ecocloud
an EPFL research center

Science entering 4th paradigm

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

Complements theory, empirical science & simulation

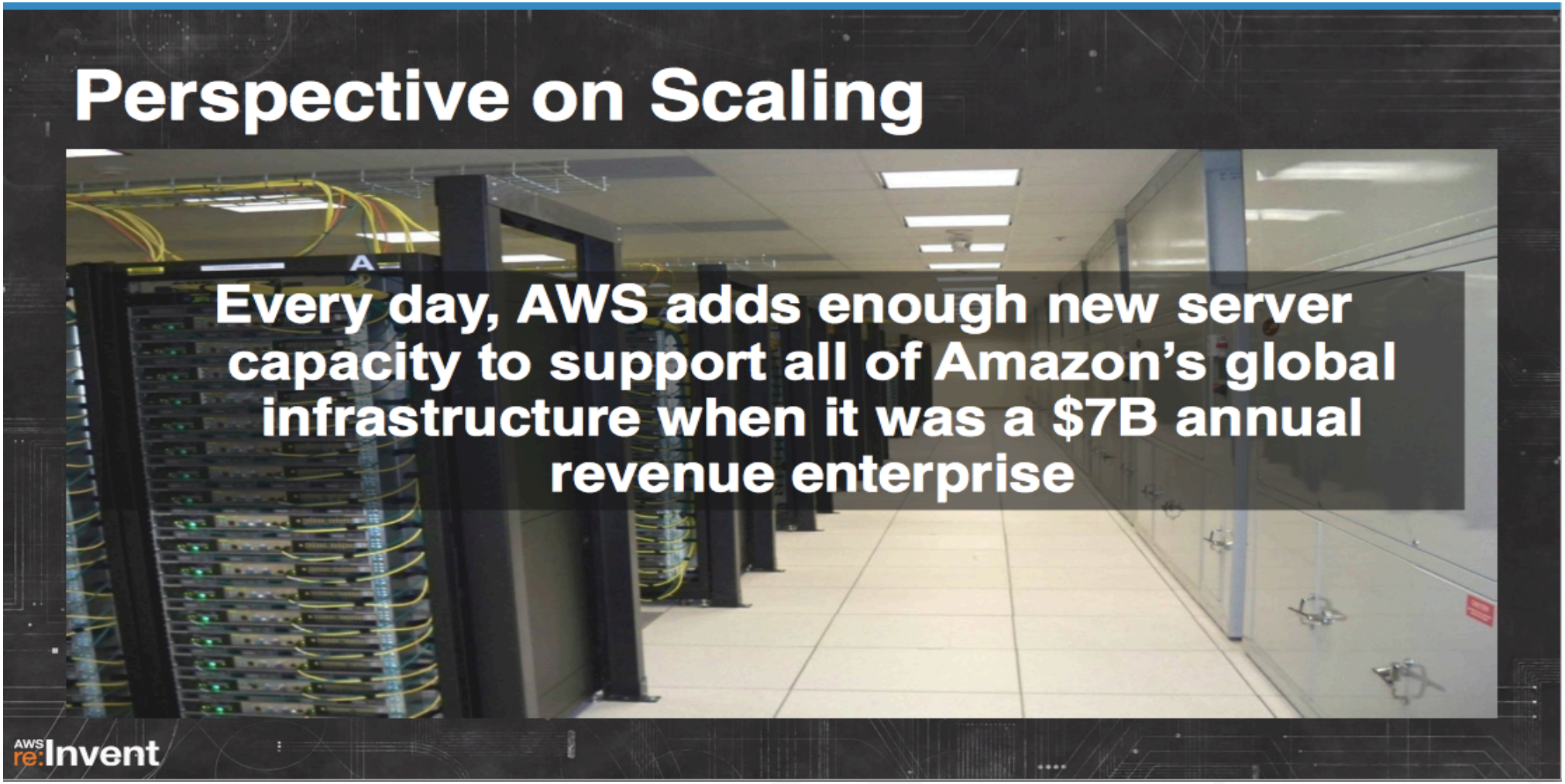


Data-centric science key for innovation-based economies!

Source: James Hamilton, 2014

mvdirona.com/jrh/TalksAndPapers/JamesHamilton_Reinvent20131115.pdf

ecocloud
an EPFL research center



Daily IT growth in 2014 = All of AWS in 2004!

Warning!
Datacenters are not Supercomputers

ecocloud
an EPFL research center

- 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

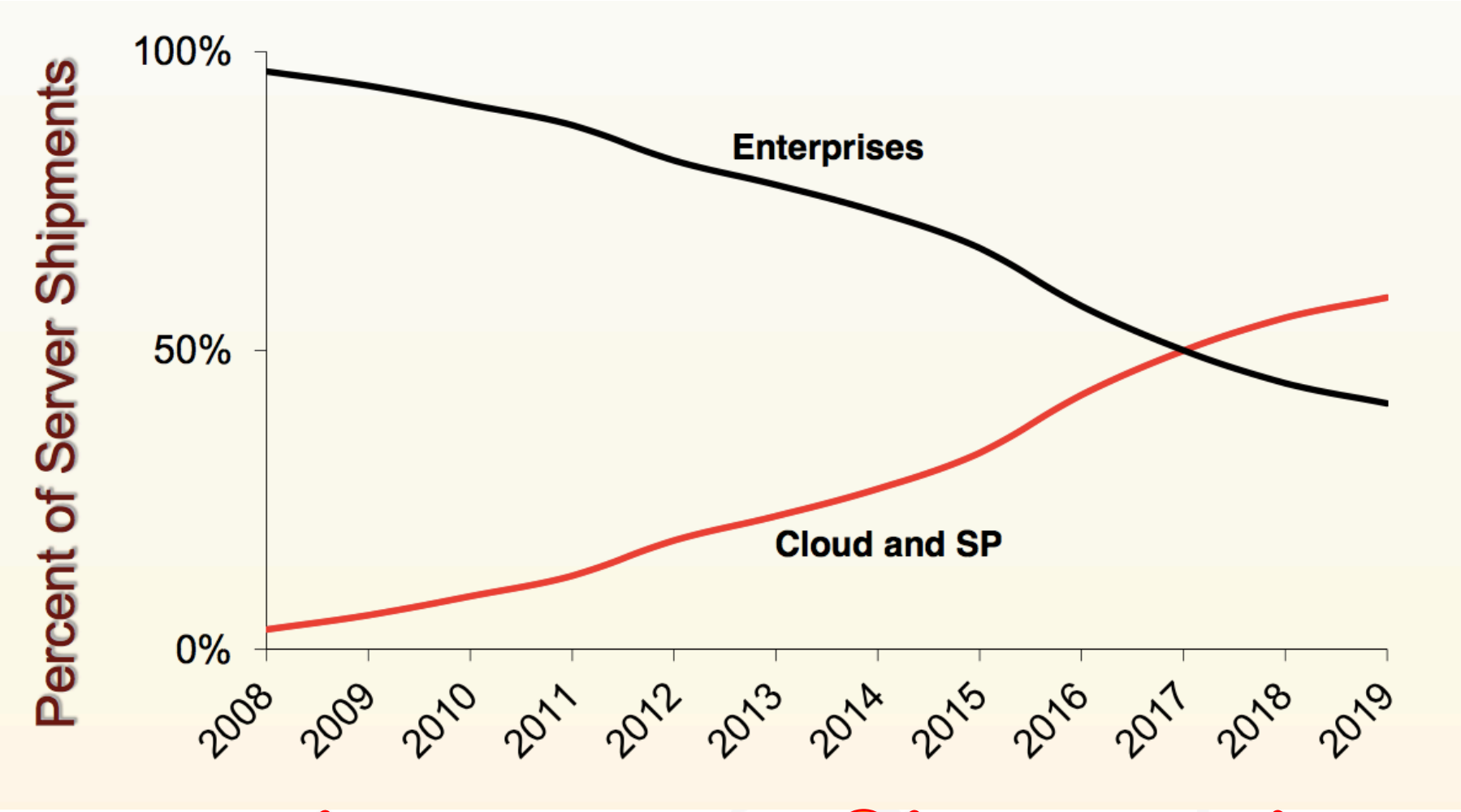
Datacenters are the IT utility plants of the future

Supercomputing

$\neq$

Cloud Computing

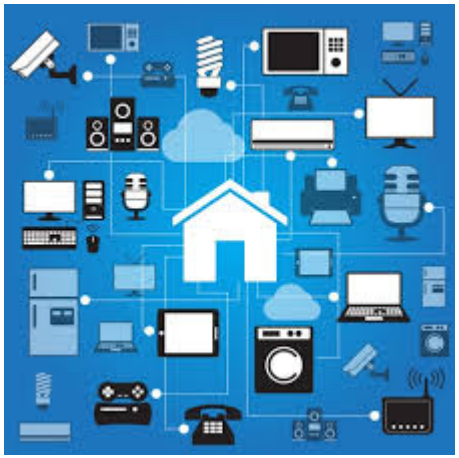
Cloud Taking Over Enterprise



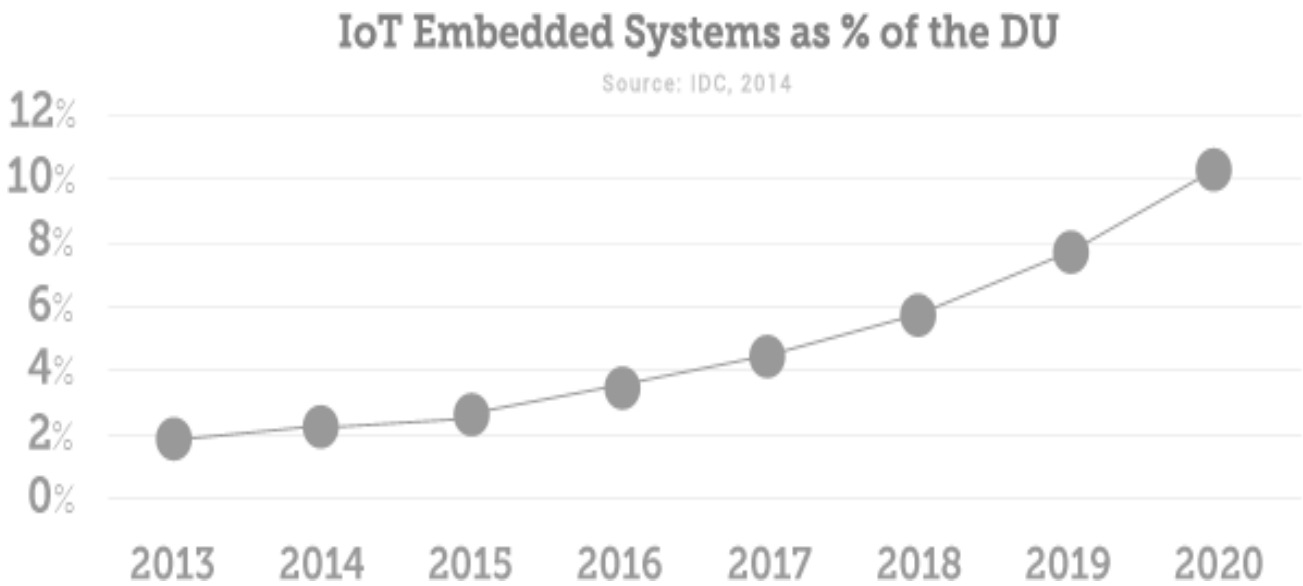
Source: Dell 'Oro 2Q15

Internet-of-Things (IoT) Growing Fast Too

ecocloud
an EPFL research center



20 Billion Connected Devices



4 Zettabytes of Data, 10% of Digital Universe

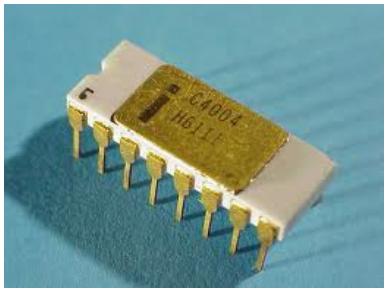
$7 Trillion
Market Revenue

Source: IDC Worldwide and Regional IoT forecast, EMC Digital Universe with Research and Analysis by IDC

Moore's Law: Five Decades of Exponential Growth

ecocloud
an EPFL research center

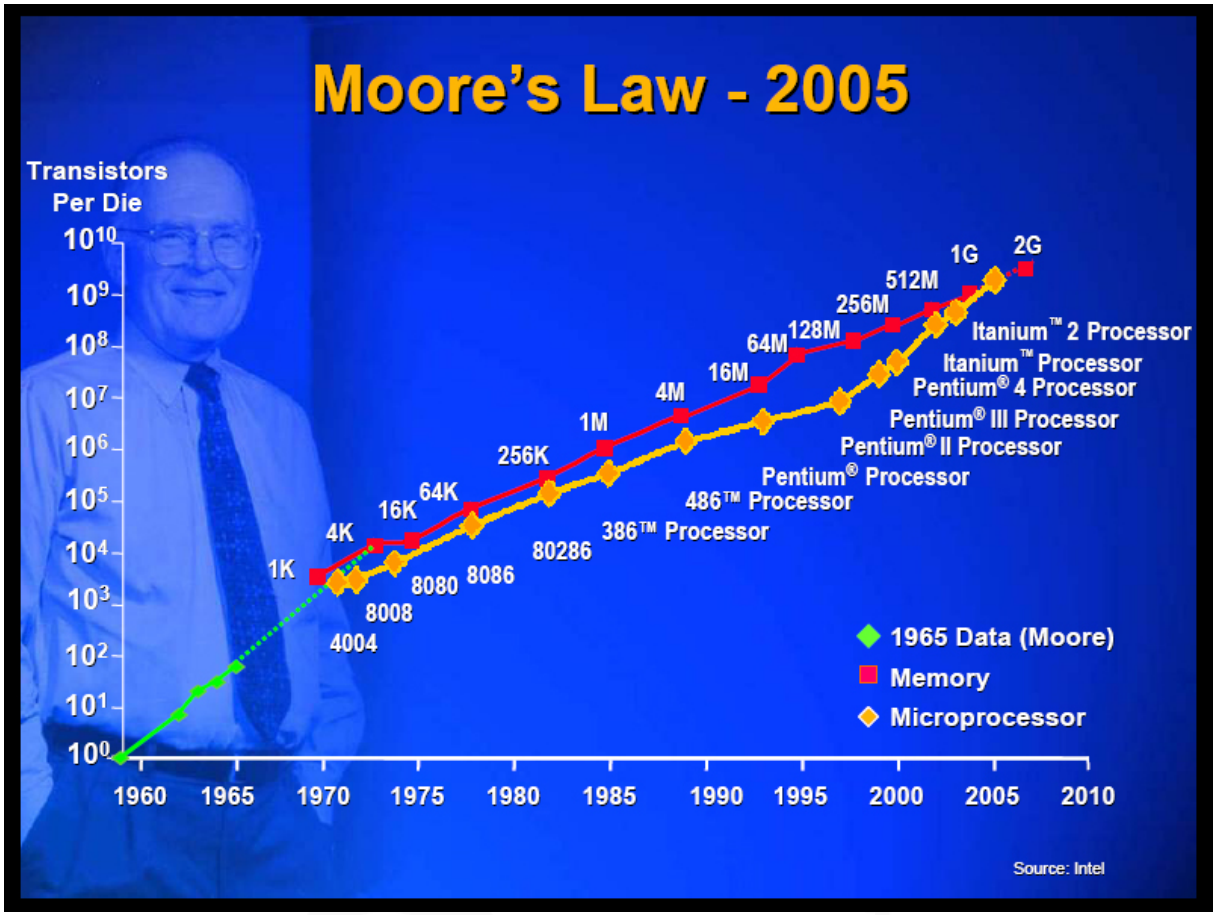
Intel 4004, 1971



92,000 ops/sec

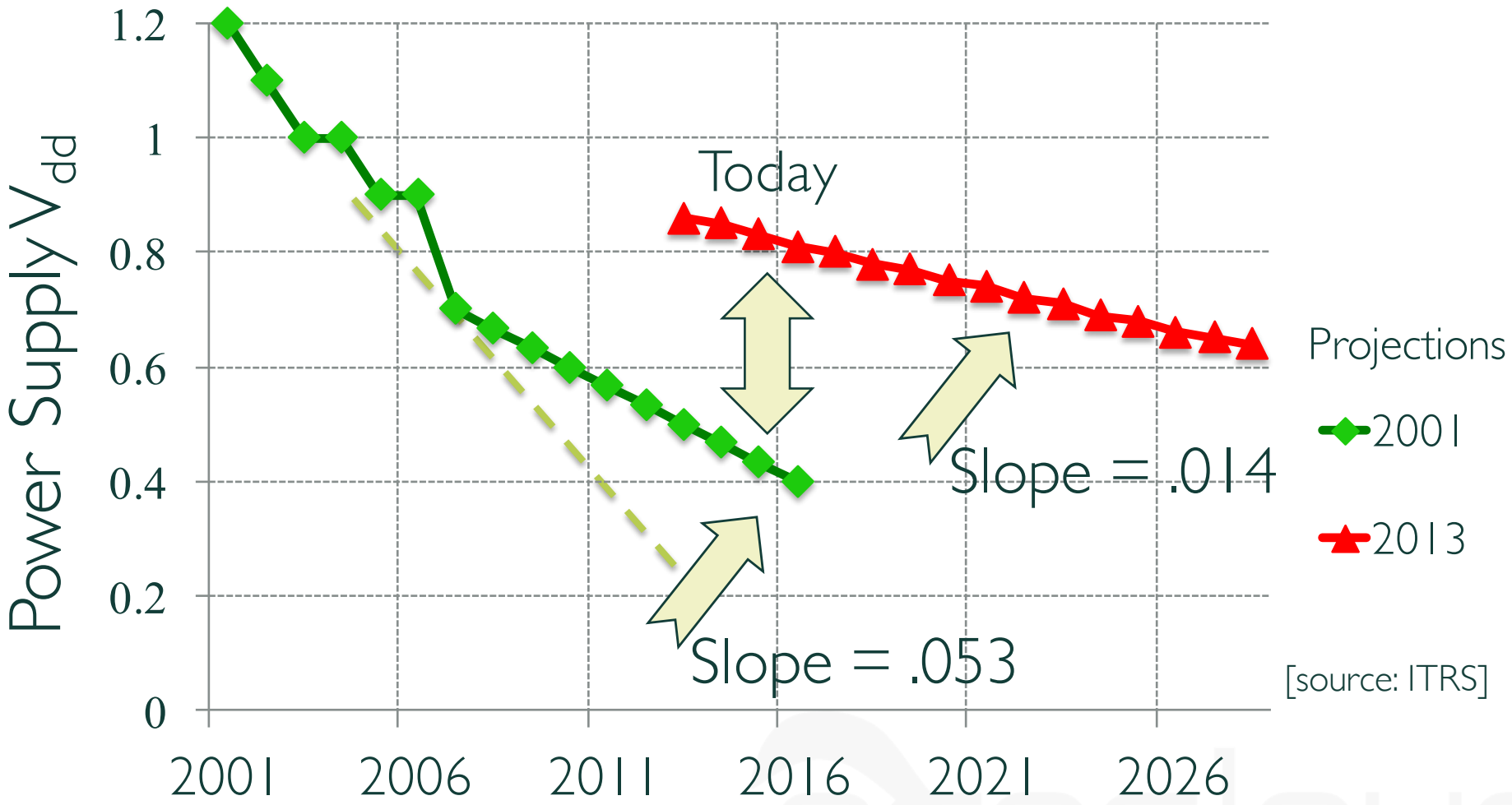
Intel Xeon, 2014

266,000,000,000 ops/sec



Made IT an indispensable pillar of our society!

End of Dennard Scaling



The fundamental energy silver bullet is gone!

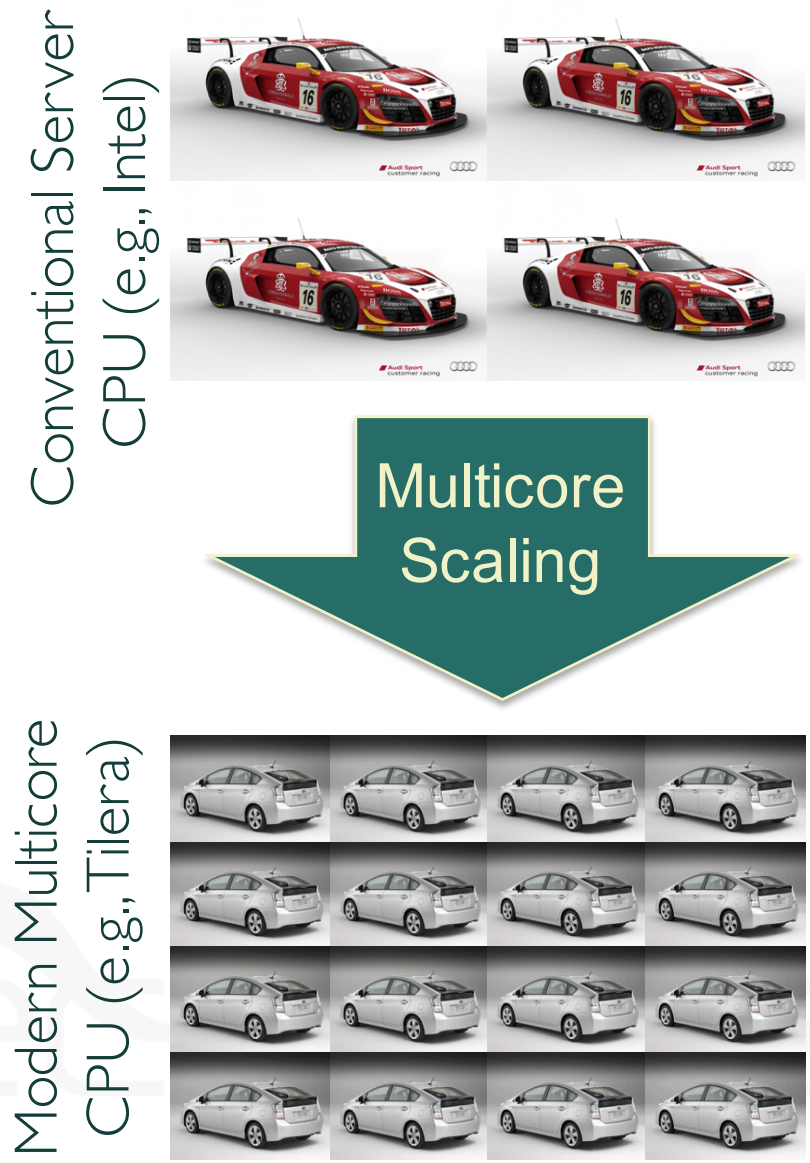
Parallelism is out of steam!

ecocloud
an EPFL research center

With voltages leveling:

- Parallelism has emerged as the only silver bullet
- Use simpler cores
  - Prius instead of race car
- Restructure software
- Each core ➔

  less joules/op



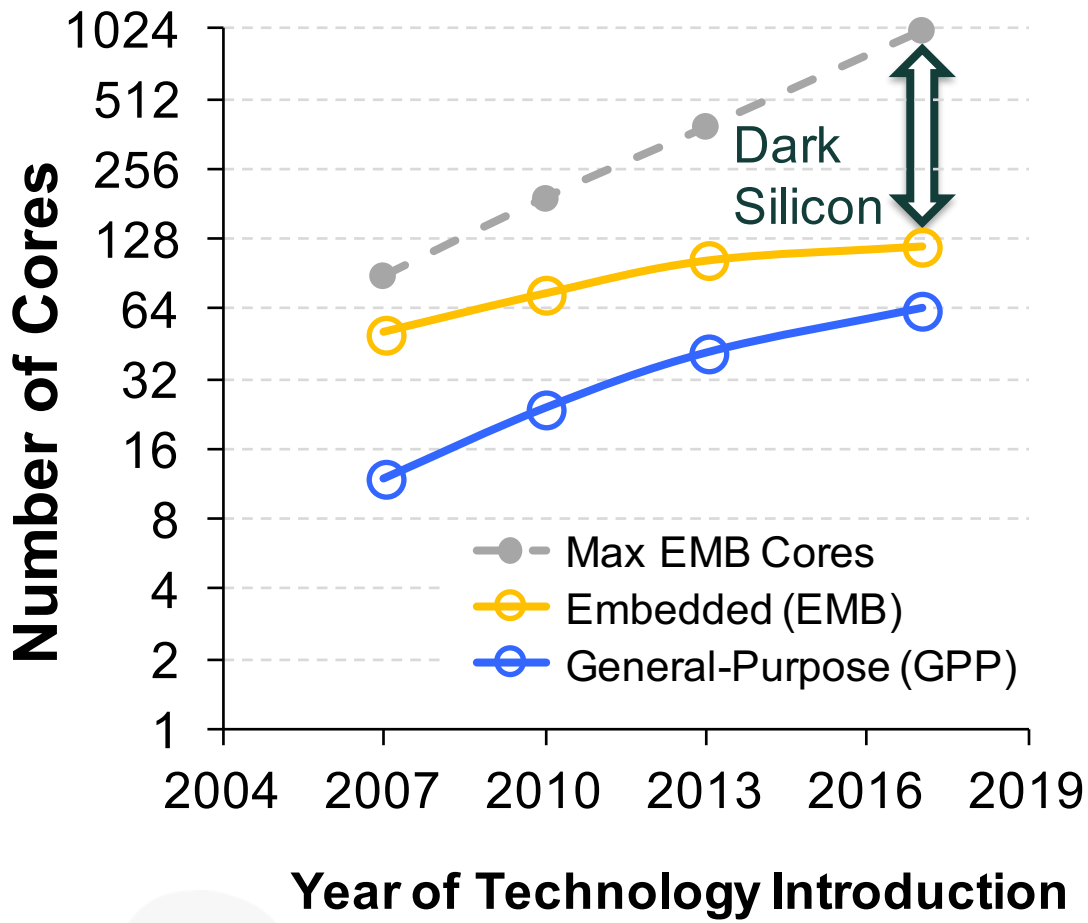
End of Multicore Scaling

ecocloud
an EPFL research center

But parallelism can not offset leveling voltages

Even in servers with abundant parallelism

Need a holistic approach to optimization

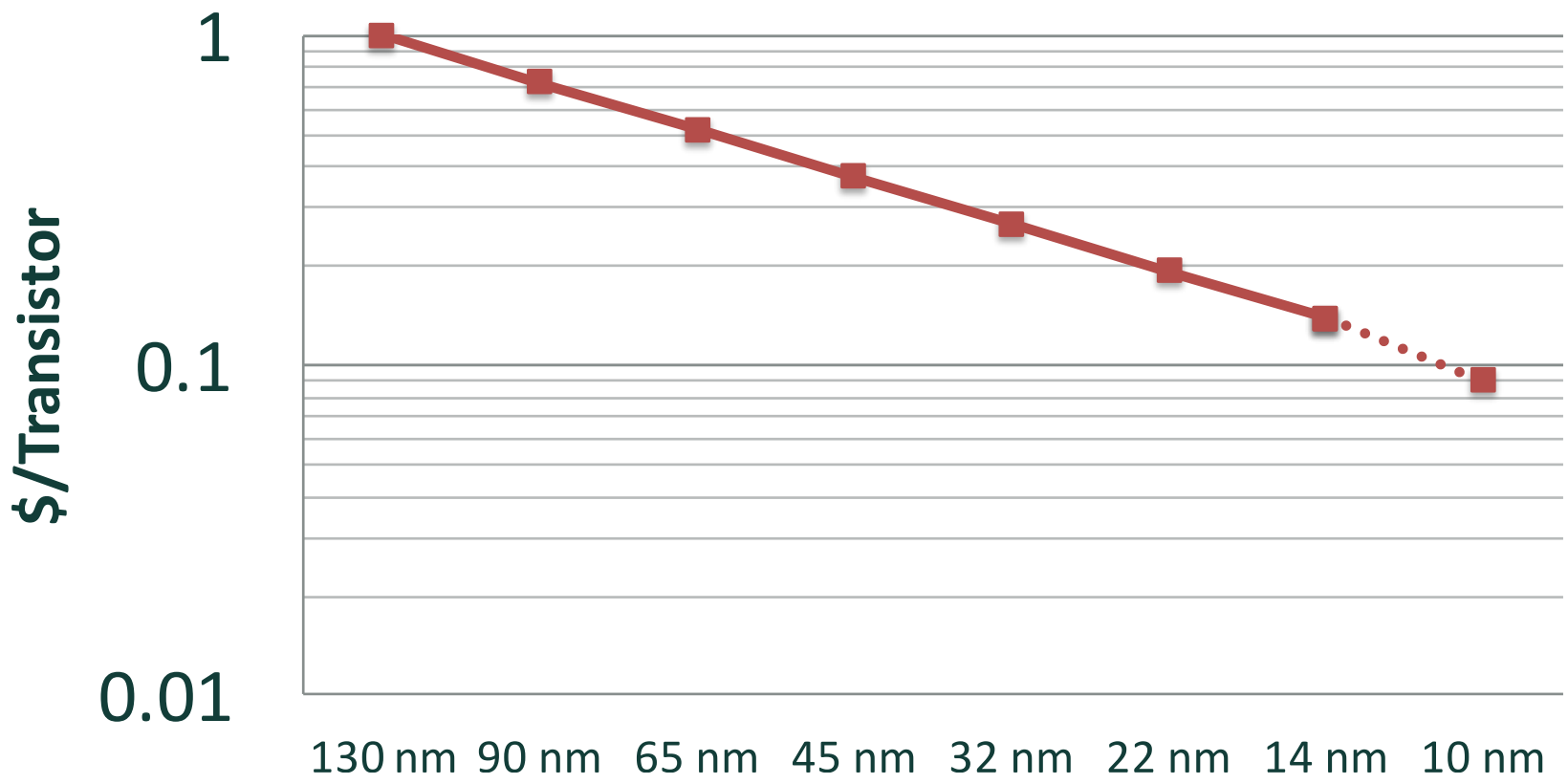


Hardavellas et. al.
"Toward Dark Silicon in Servers"
IEEE Micro, 2011

Slowdown in Moore's Law

ecocloud
an EPFL research center

Mark Bohr's (Intel) Keynote [ISSCC'15]

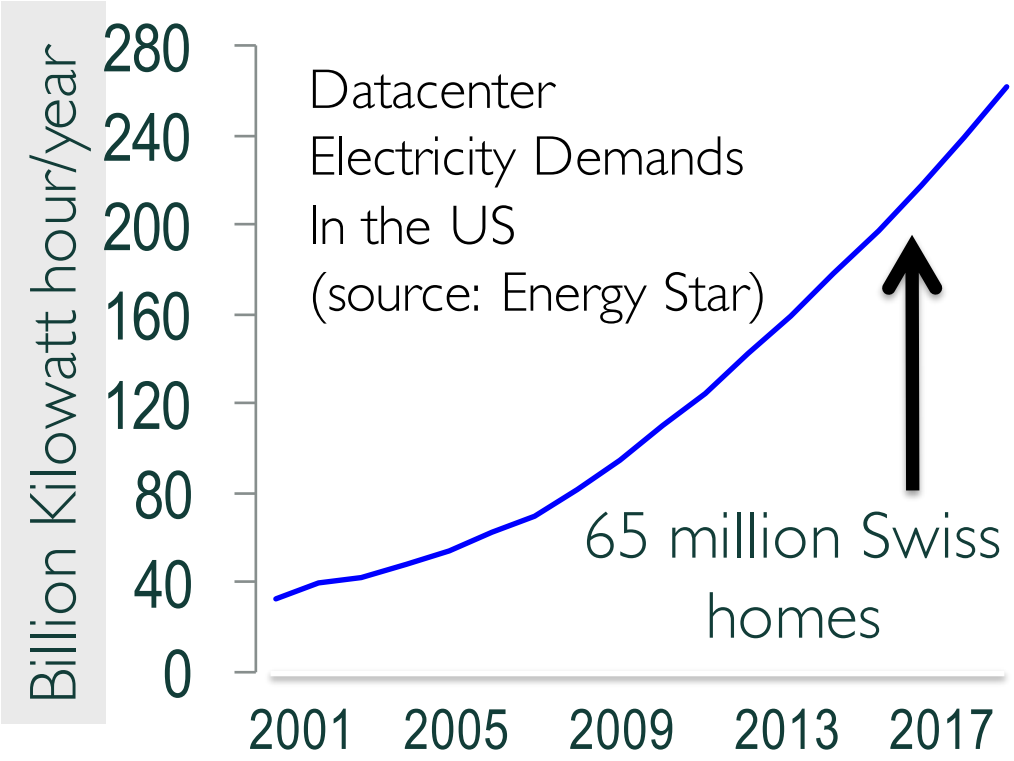


Moore's Law: $/transistor dropping for fifty years

- Intel is pushing for a bit more
- Competitor saw $/transistor go up 2015

Higher Demand + Lower Efficiency: Datacenters at Physical Limits!

ecocloud
an EPFL research center



- Centralization helps exploit economies of scale
- But, platform scaling is a grand challenge

ecocloud

Center at EPFL

- 18 faculty, 50 researchers
- 6M CHF/year external funds

Mission:

- Designing datacenters of future
- From algorithms to infrastructure
- Maximizing value for data

Microsoft
Google
Hewlett Packard Enterprise
HGST
HUAWEI
intel
swisscom
AXA
EATON Powering Business Worldwide
IBM
CISCO
CREDIT SUISSE
ORACLE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

ecocloud.ch

Today's Server Ecosystem

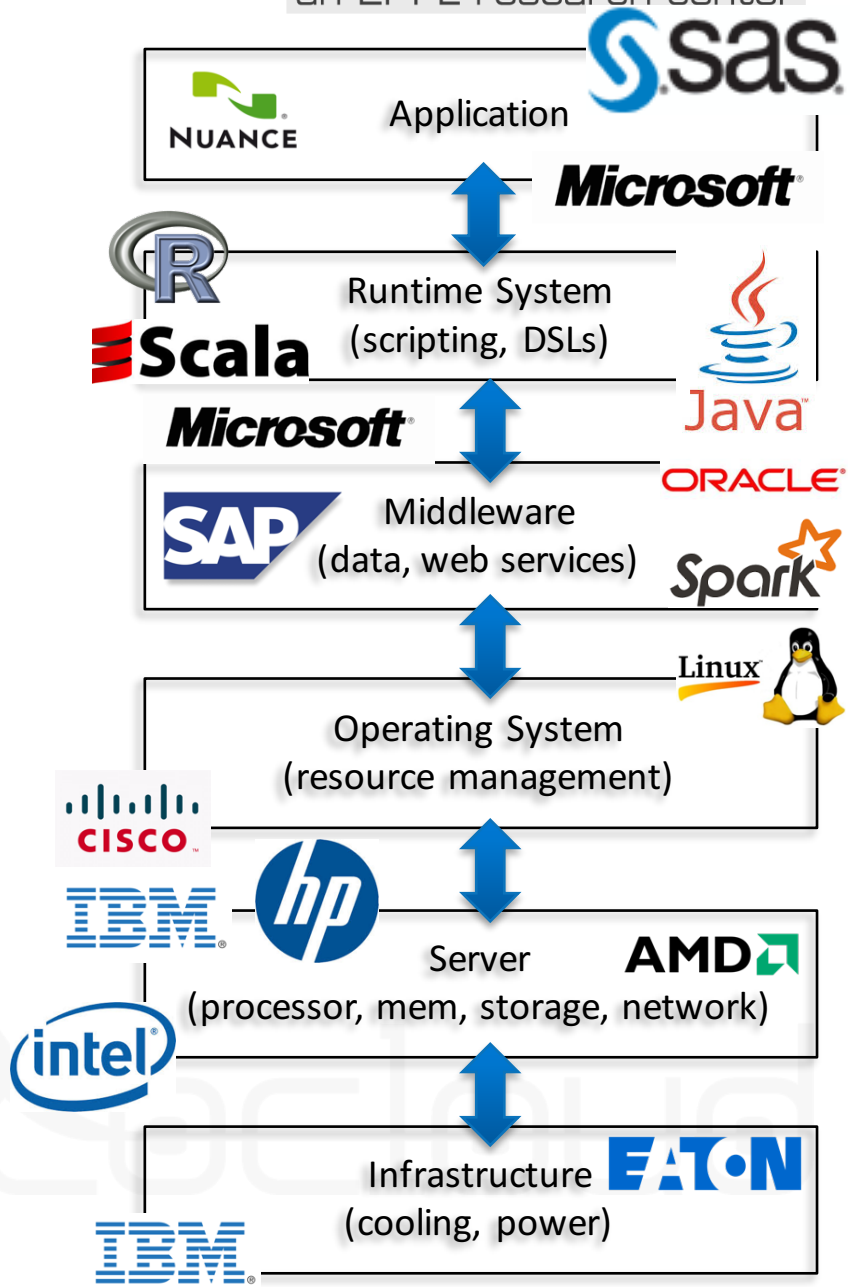
ecocloud
an EPFL research center

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
  - Are limited in extent (e.g., software)
  - Monopolize (closed) technologies



Our Vision:
Holistic Optimization of Datacenters

ecocloud
an EPFL research center

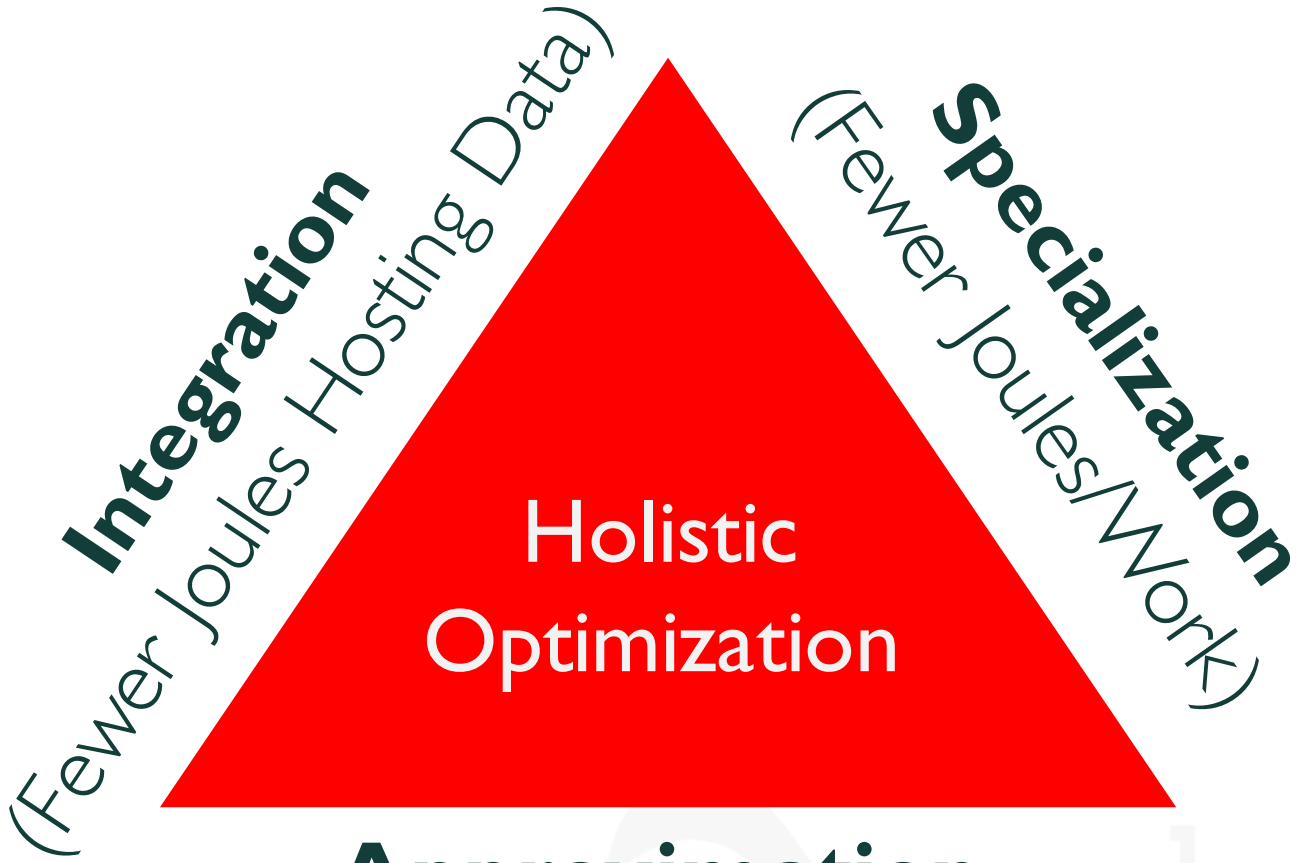
Holistic optimization

- From algorithms to infrastructure
- Cross-layer integration
- IT paradigms to monitor, manage & reduce energy

Open technologies!

Our Vision:
The ISA Triangle of Efficiency

ecocloud
an EPFL research center



Approximation
(Tailor Precision for Fewer Joules)

Outline

- ~~Overview~~
- How efficient are servers today?
- DB Accelerators
- Summary

Scale-Out Datacenters

ecocloud
an EPFL research center

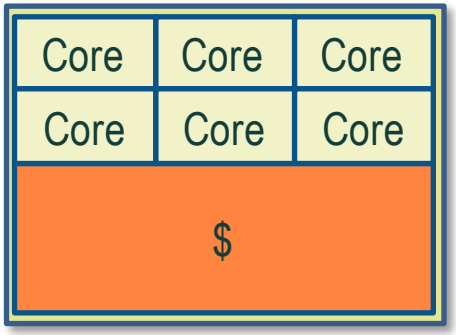
Vast data sharded across servers

Memory-resident workloads

- Necessary for performance
- Major TCO burden

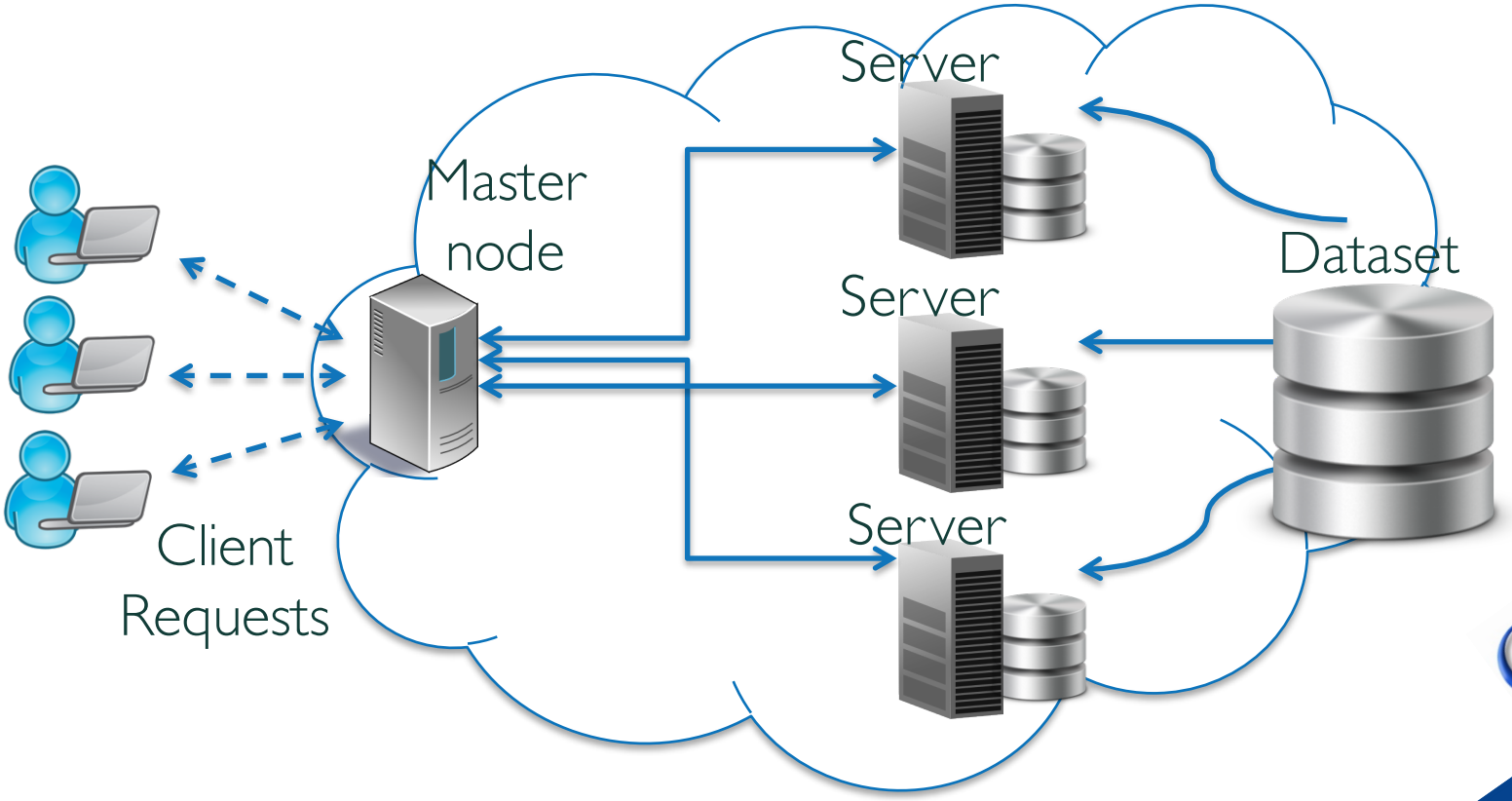
Put memory at the center

- Design system around memory
- Optimize for data services



Servers driven by the DRAM market!

In-Memory Scale-Out Services



- Many independent requests/tasks
- Huge dataset split into shards
- Use aggregate memory over network

How Efficient are Servers Today?
CloudSuite 3.0 (parsa.epfl.ch/cloudsuite)

Data Analytics
Machine learning

Data Caching
Memcached

Data Serving
Cassandra NoSQL

Graph Analytics
GraphX

Media Streaming
Nginx, HTTP Server

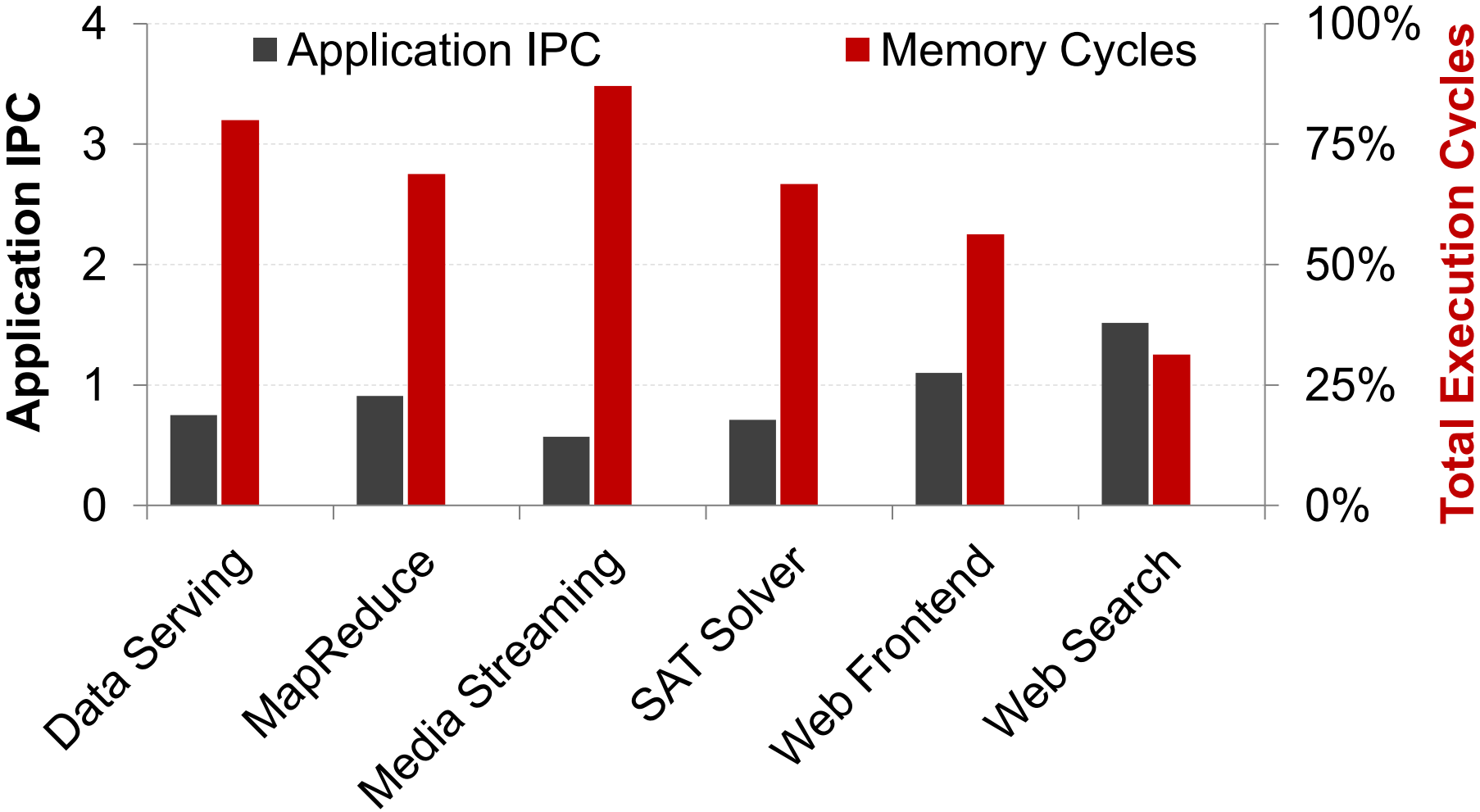
Web Serving
Nginx, PHP server

Web Search
Apache Solr & Nutch

In-Memory Analytics
Recommendation System

Building block for Google PerfKit, EEMBC Big Data!

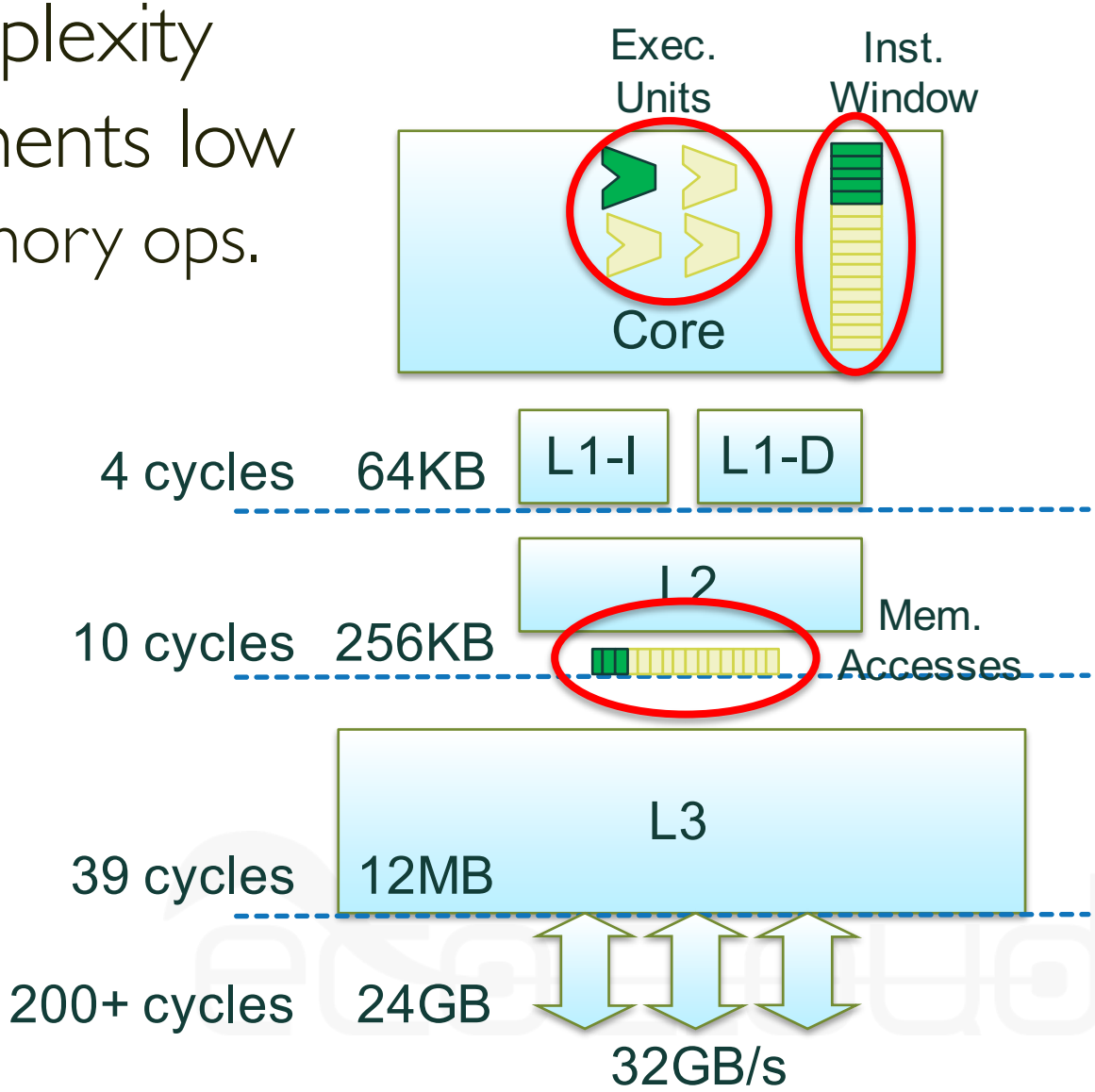
But, Services are Stuck in Memory!



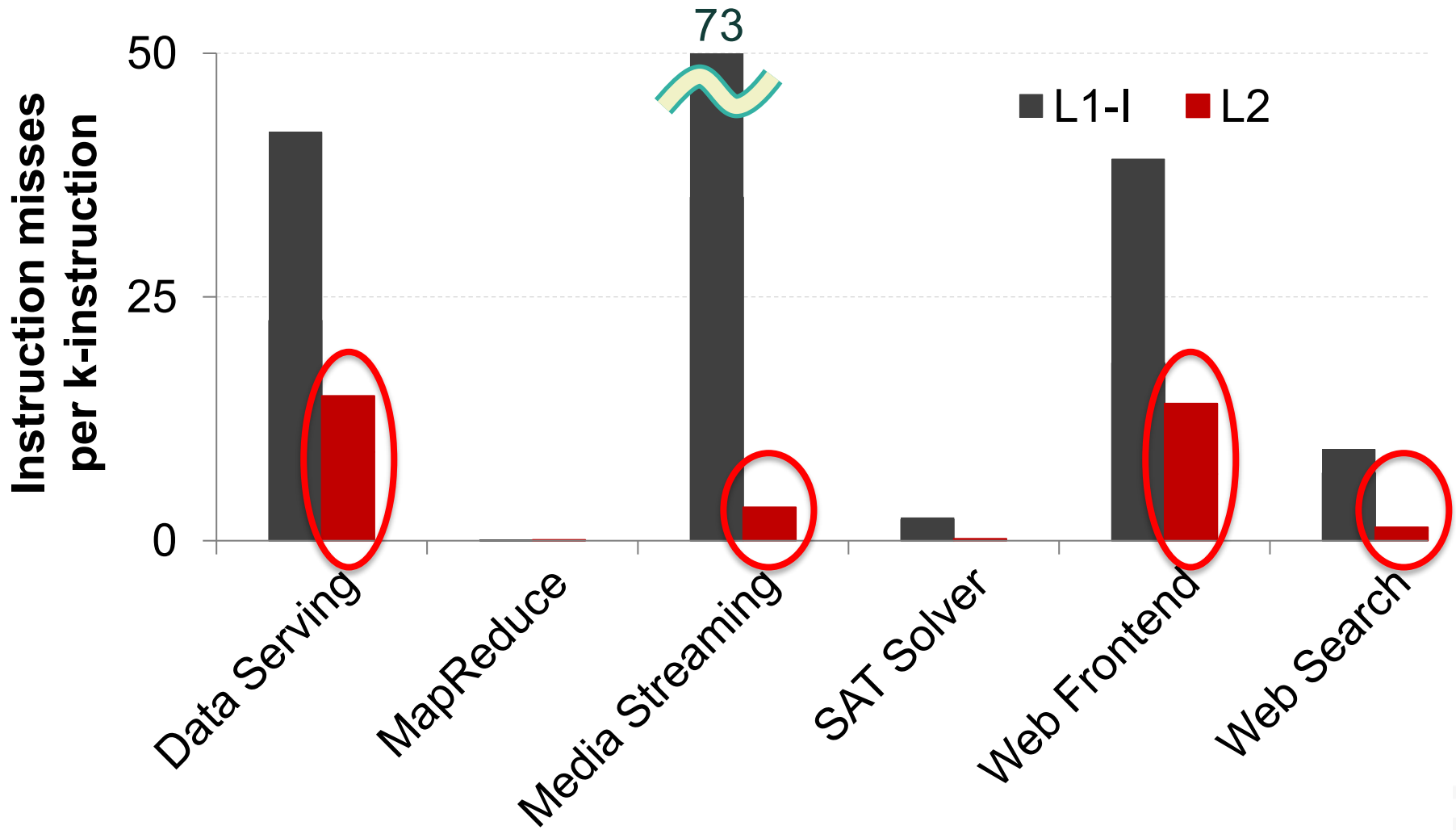
Execute ~1 instruction per cycle

Core Inefficiencies

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



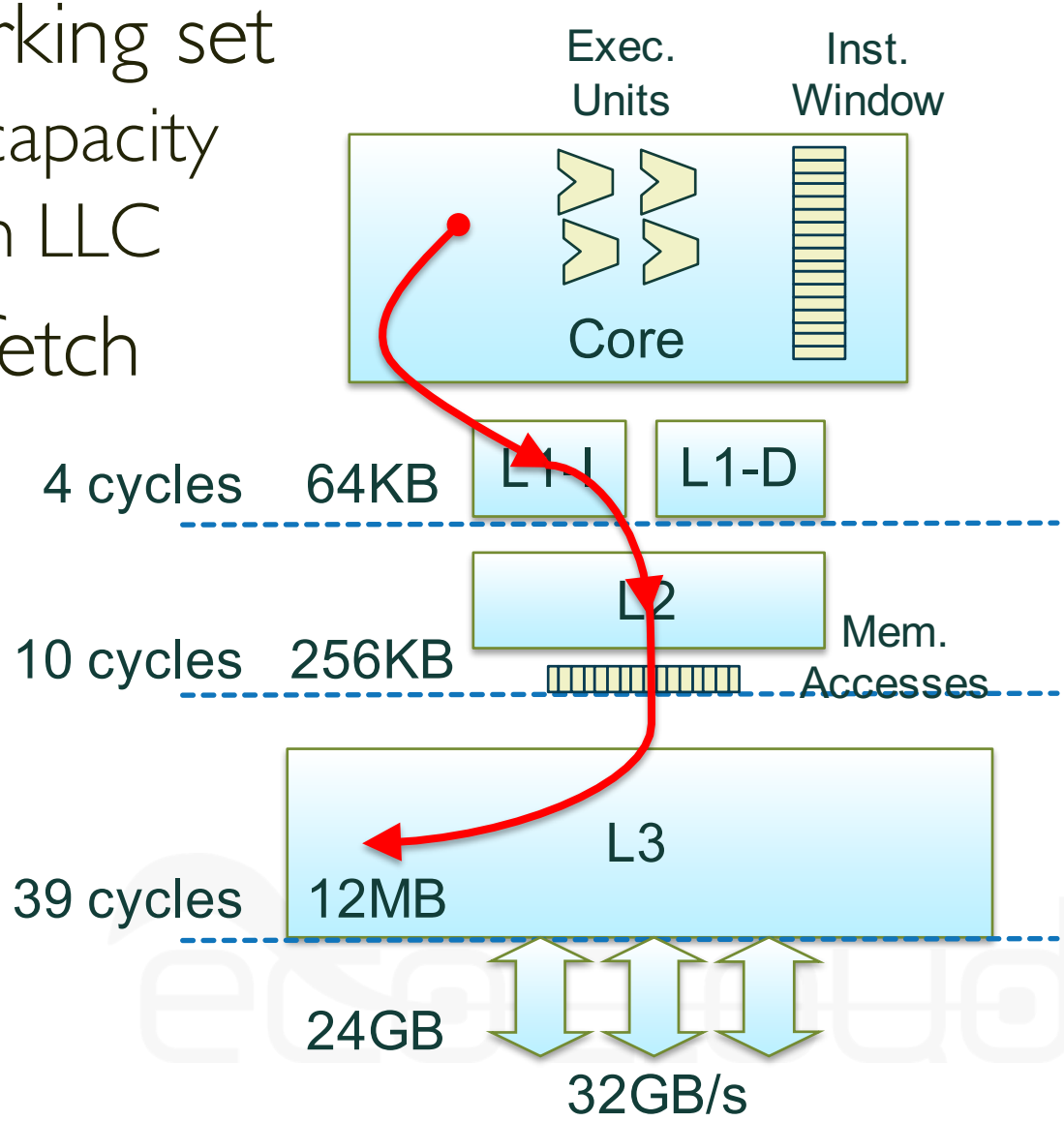
Instruction-Fetch Misses



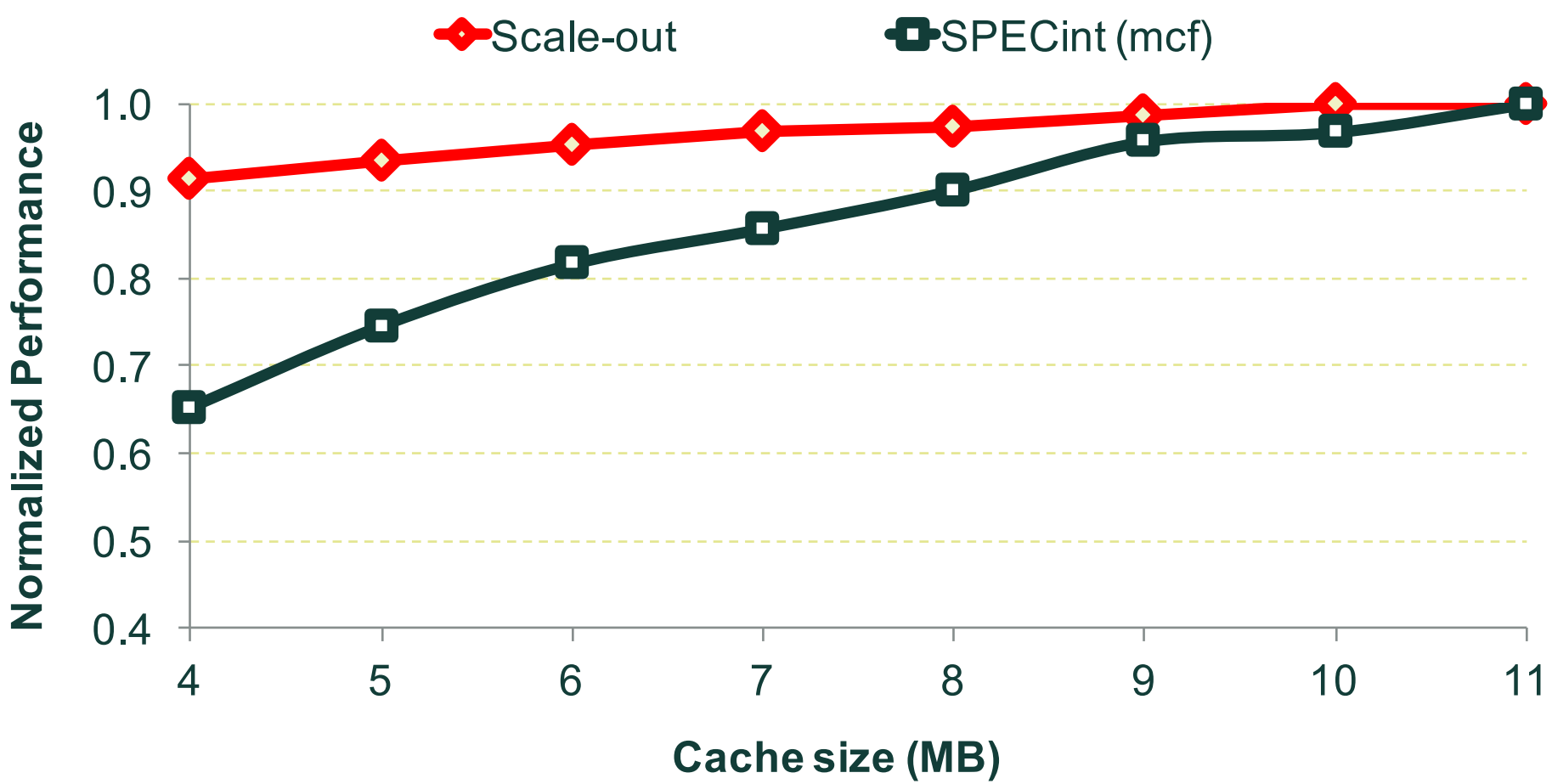
Suffer severe i-cache miss penalties

Instruction-Fetch Inefficiencies

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



LLC Sensitivity

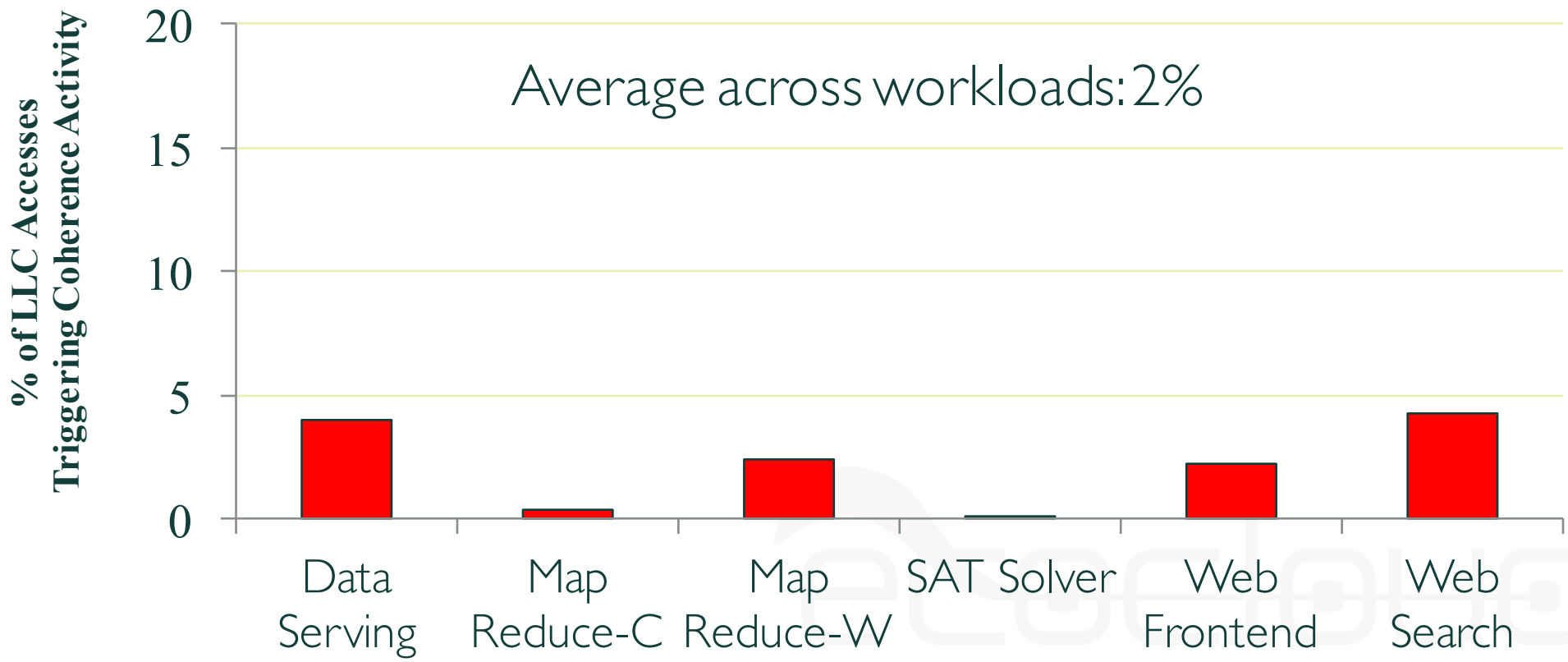


Minimal performance from large LLC

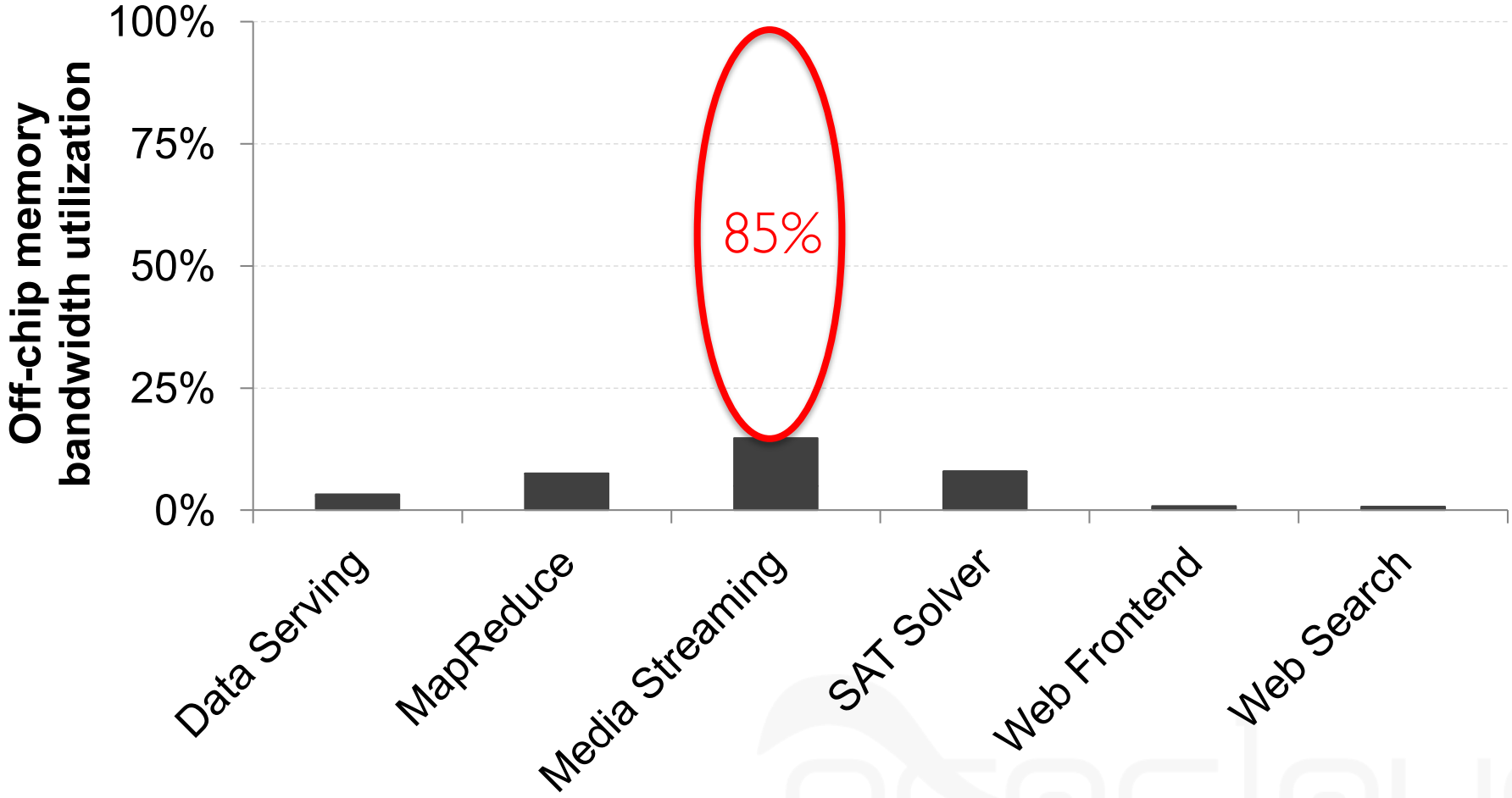
Where do instructions/data come from?

- Instructions: in LLC
- Data: in memory

} Nothing useful in remote L1 caches!



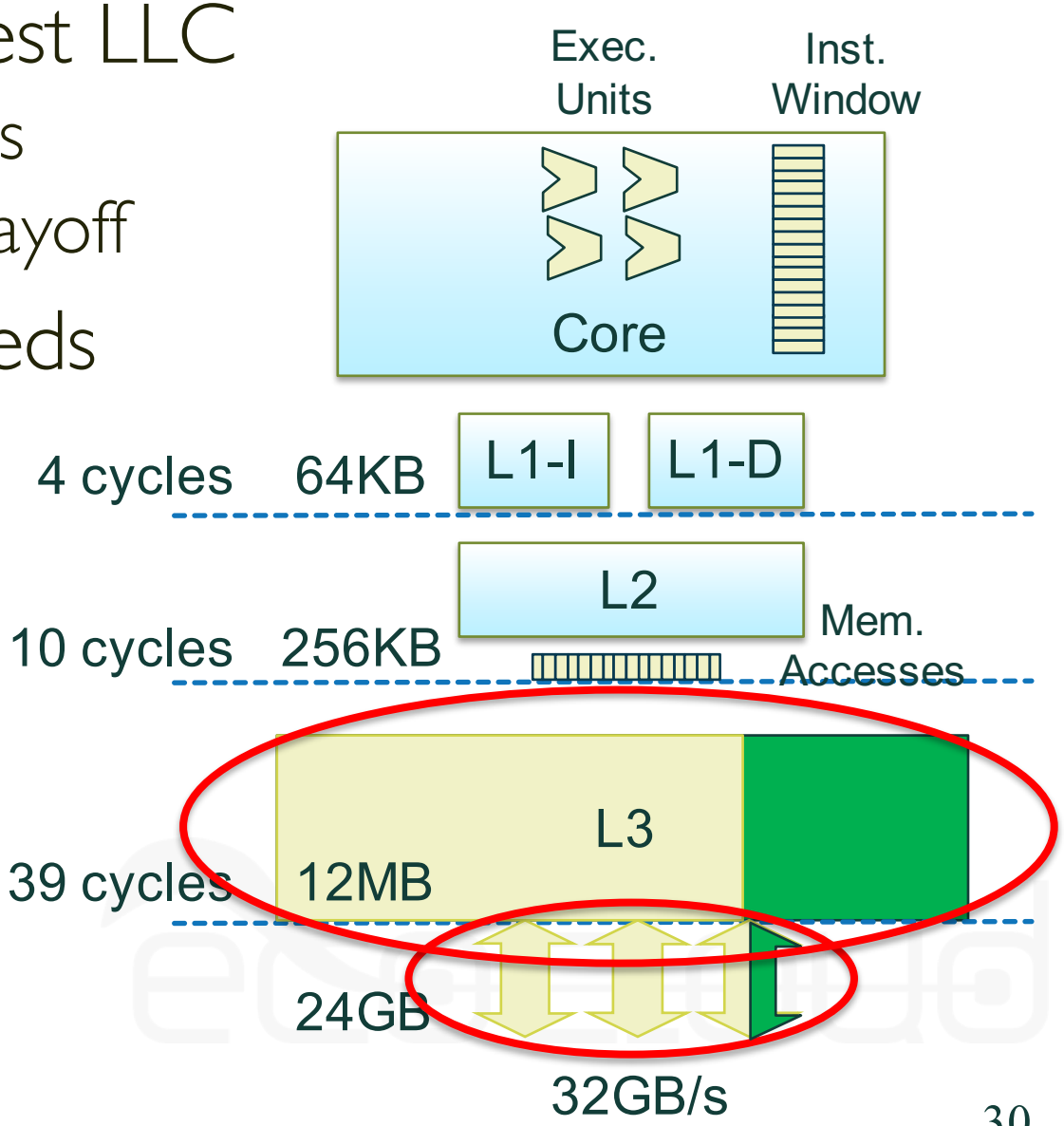
Off-chip Memory Bandwidth



Off-chip BW severely underutilized

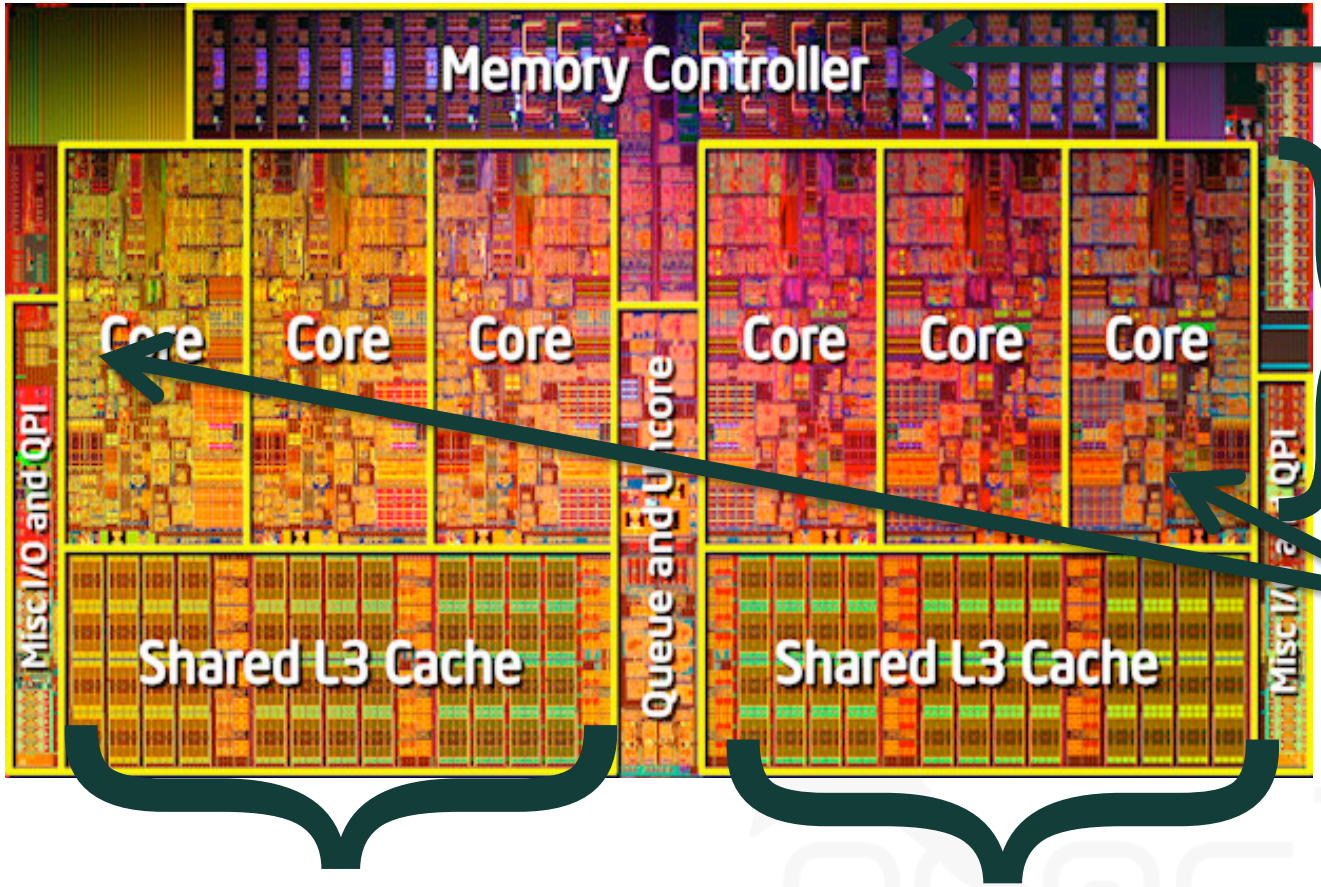
LLC and Bandwidth Inefficiencies

- 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 Scale-Out Services Need?

ecocloud
an EPFL research center

Cores share instructions

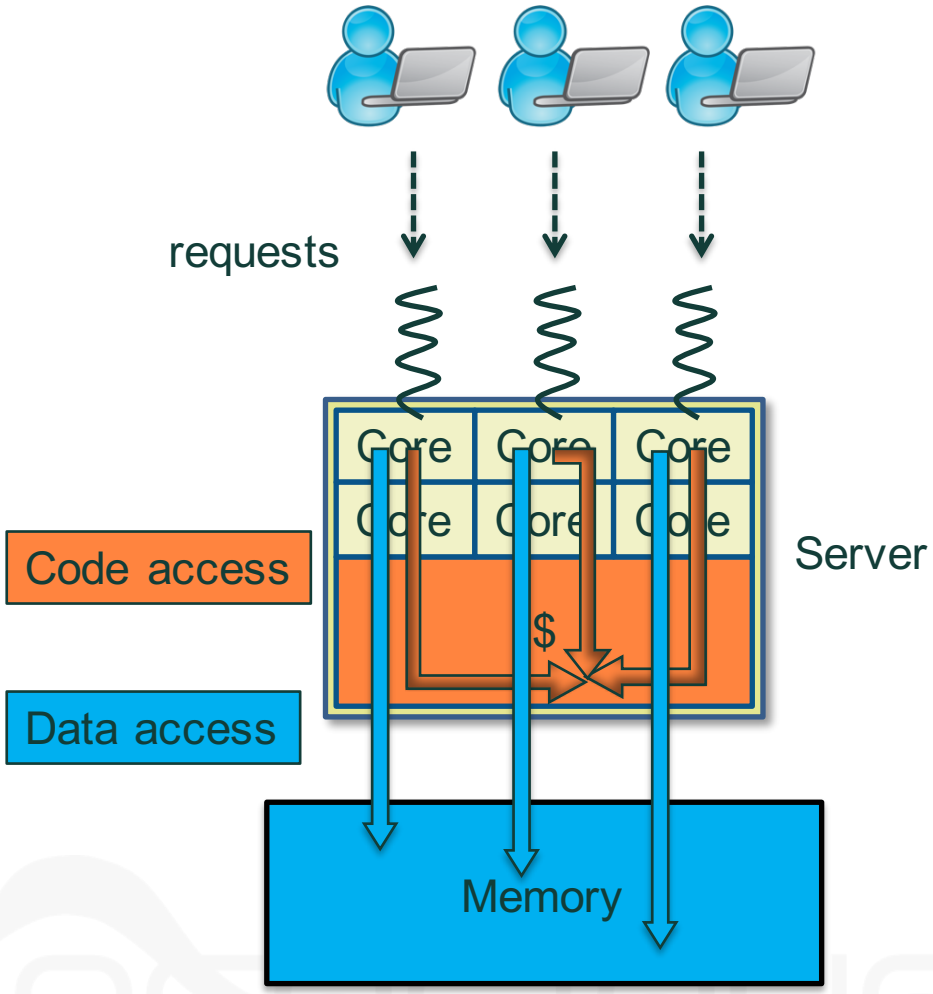
- Large code footprint fits in LLC
- A few MB SRAM for instructions

Data is in memory

- Data footprint dwarfs LLC
- Do not waste SRAM for data

Cores communicate rarely

- Independent requests
- Core-to-cache traffic



Common traits across applications

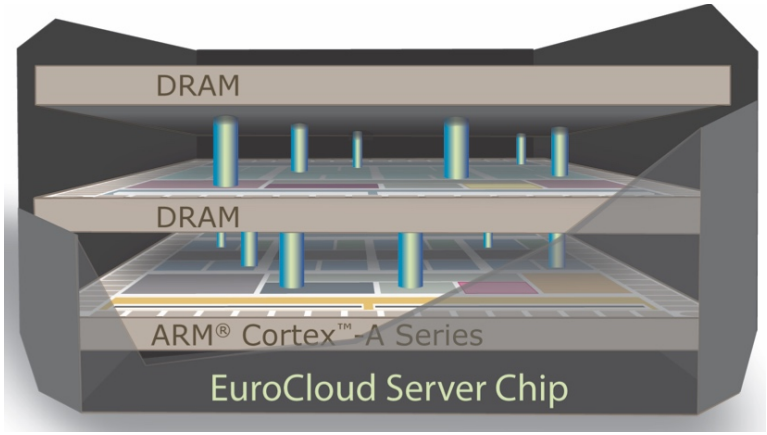
Scale-Out Processors

[ISCA'13,ISCA'12,Micro'12]

ecocloud
an EPFL research center

Server Chip:

- Disconnected cache-coherent pods
- 3D memory
- 10x performance/TCO
- Runs Linux LAMP stack



Processor SoC:

- 64-bit ARM cores
- Custom degree of OoO/MLP
- NoC designed for fast instruction supply
- LLC designed for on-chip instruction working set

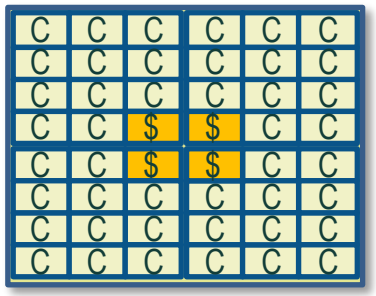
imec
EPFL
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
SEVENTH FRAMEWORK PROGRAMME
UCy
ARM

NOC-Out: [MICRO'12] Specialized Network-on-Chip for Servers

ecocloud
an EPFL research center

Exactly the **opposite** of current NoCs

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

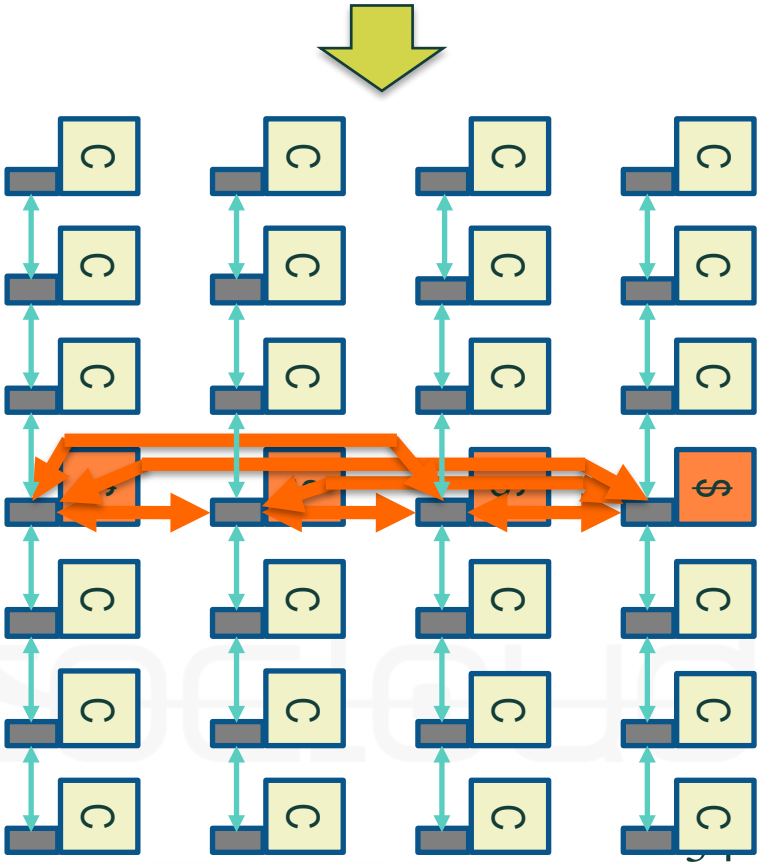


LLC network:

- Flattened Butterfly (FB) topology

Request & *Reply* networks:

- Tree topology
- Limited connectivity for efficiency



FB's performance at 1/10th cost

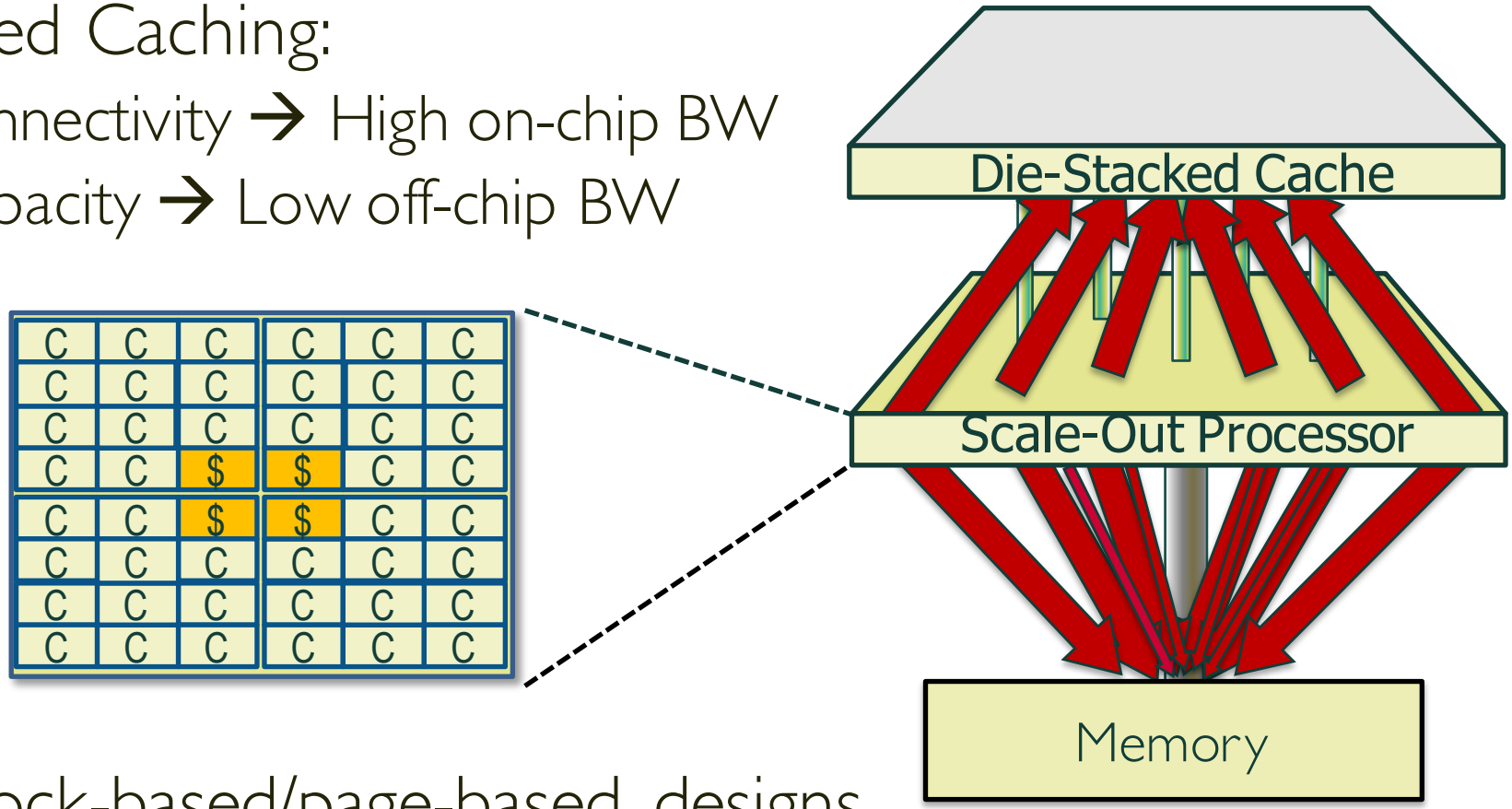
Effective Die-Stacked Caching for Servers

[ISCA'13, MICRO'14, IEEE Micro'16]

ecocloud
an EPFL research center

Die-Stacked Caching:

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



Hybrid block-based/page-based designs

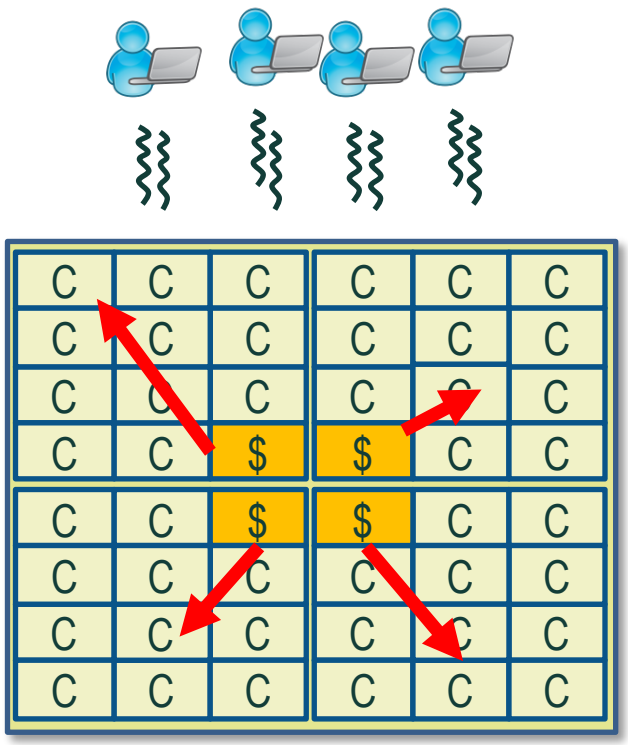
- Embed tags in DRAM
- Predict & fetch page's footprint

Specialized Instruction Supply
[MICRO'08,11,13,15]

Instruction supply highly repetitive

Record & replay instruction streams

- Eliminate 99.7% of all i-cache misses
- Capture discontinuities in control flow
- Embed front-end meta data



Centralized engine

- Stream front-end state

Emerging server processors call for a holistic front-end solution

Cavium ThunderX: A Scale-Out Processor

EE Times Connecting the Global Electronics Community

BREAKING NEWS SLIDESHOW: CES: Bosch Aims to Connect Whole World

designlines WIRELESS & NETWORKING

News & Analysis

Big-Data Benchmark Brewing

EEMBC works on SoC-agnostic spec

Rick Merritt
10/15/2014 08:00 AM EDT

SAN JOSE, Calif. — A new benchmark suite for scaled-out servers is in the works with the first piece of it expected early next year. The processor-agnostic metrics aim to set standards for measuring today's data center workloads.

A new cloud and big-data server working group of the Embedded Microprocessor Benchmark Consortium (EEMBC) hopes to deliver a suite of seven benchmarks. It aims to complete before April three of them -- memory caching, media serving, and graph analysis.

"Typically when we go to a server customer they ask for SpecInt numbers, that's been the traditional benchmarks for servers for a long time, but SpecInt is not a very good metric for distributed data loads or available instruction and memory parallelism," said Bryan Chin, a distinguished engineer from Cavium.

MICROPROCESSOR report

Insightful Analysis of Processor Technology

ThunderX Rattles Server Market

Cavium Develops 48-Core ARM Processor to Challenge Xeon

By Linley Gwennap (June 9, 2014)

48-core 64-bit ARM SoC

[based on "Clearing the Clouds", ASPLOS'12]:

- 3x L1 instruction cache size
- Custom cores for moderate MLP
- Minimal LLC (replaced with cores)
- Crossbar for fast instruction fills

CAVIUM NETWORKS

Integrated Compute in Memory

[IEEE Micro'16]

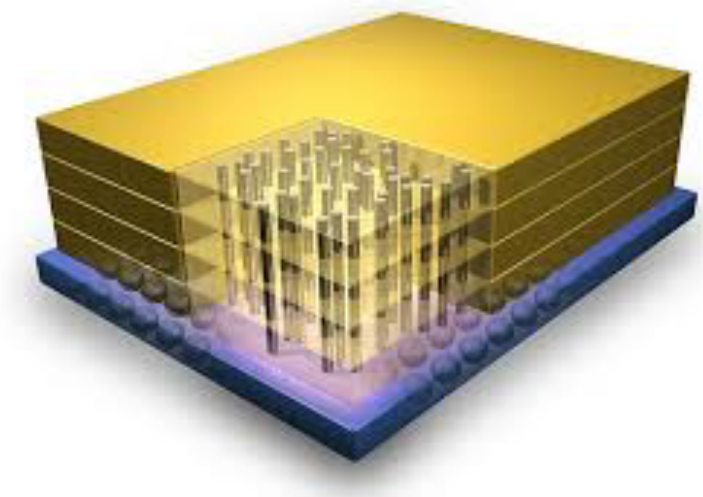
ecocloud
an EPFL research center

Why in-memory?

- Minimize data movement & energy
- Leverage DRAM's massive internal BW

Basic data services:

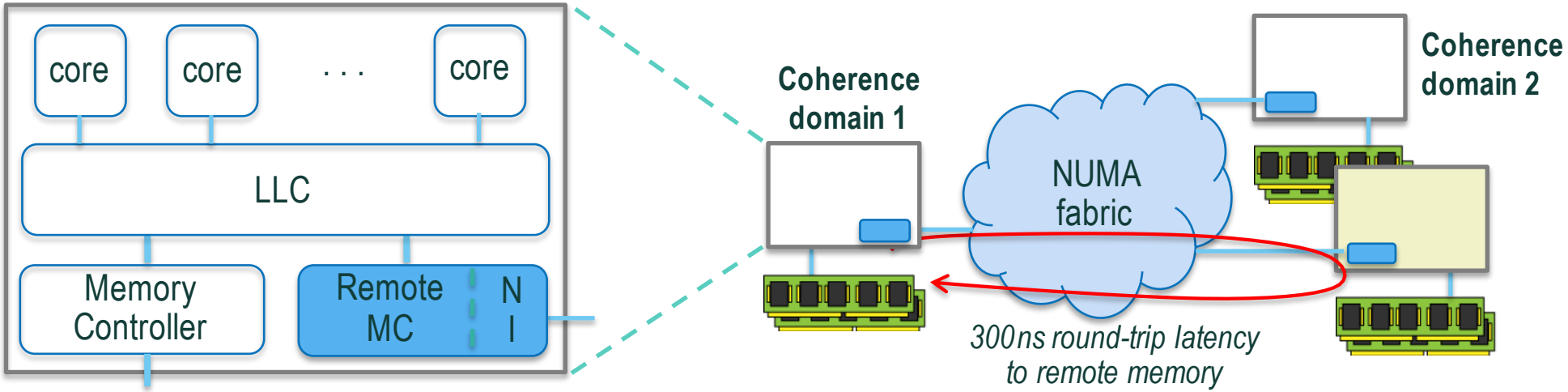
- Scan, Join, GroupBy, Filter
- Best for sequential access
- Accelerators must co-exist with conventional memory semantics



10x better efficiency for a database join operation!

Scale-Out NUMA: In-memory Rack-Scale Computing

[ASPLOS'14 ,ISCA'15, MICRO'16]



Rack-scale networking suffers from

- Network interface on PCI + TCP/IP
- Microseconds of roundtrip latency at best

soNUMA:

- Manycore network interface integrated into NoC
- Protected global memory read/write
- Supports fine-grain & bulk object communication

Outline

- ~~Overview~~
- ~~How efficient are servers today?~~
- DB Accelerators
- Summary

Databases underlie data-intensive apps

ecocloud
an EPFL research center

Most frequent task: find data

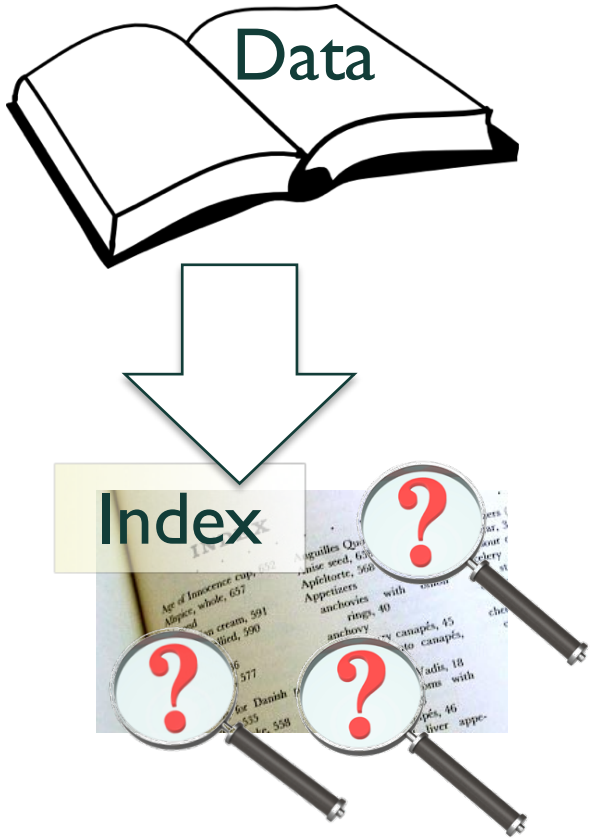
- E.g., build a user's Facebook page

Indexes used for fast data lookup

- Rely on pointer-intensive data structures

Indexing efficiency is critical

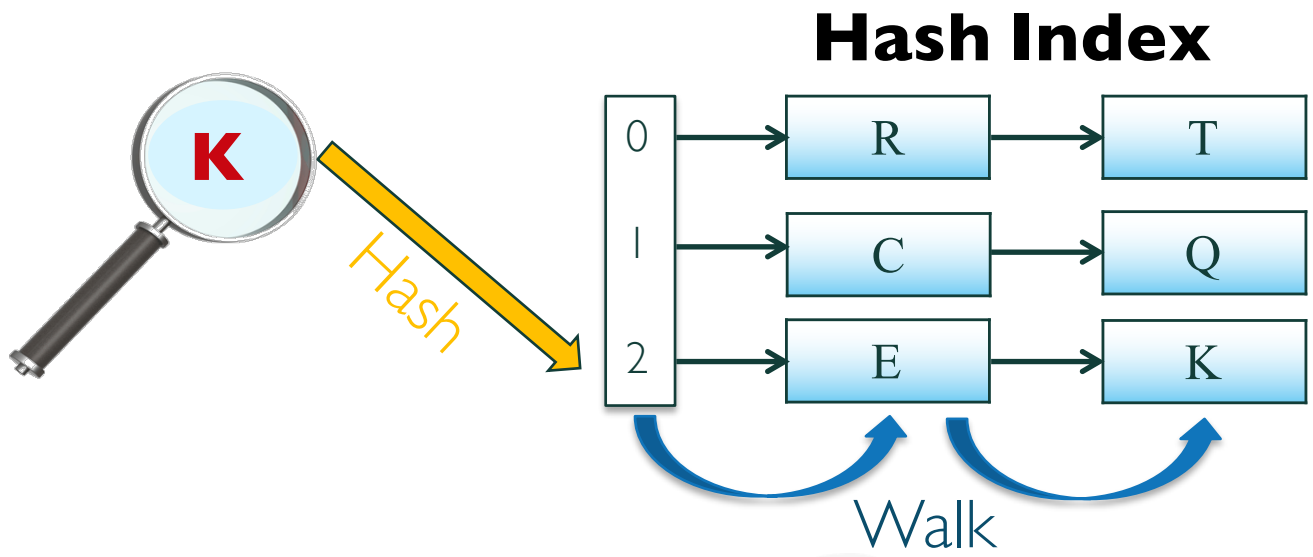
- Many requests, abundant parallelism
- Power-limited hardware



Need high-throughput and energy-efficient index lookups

Indexing Basics

Hash index: fundamental index structure

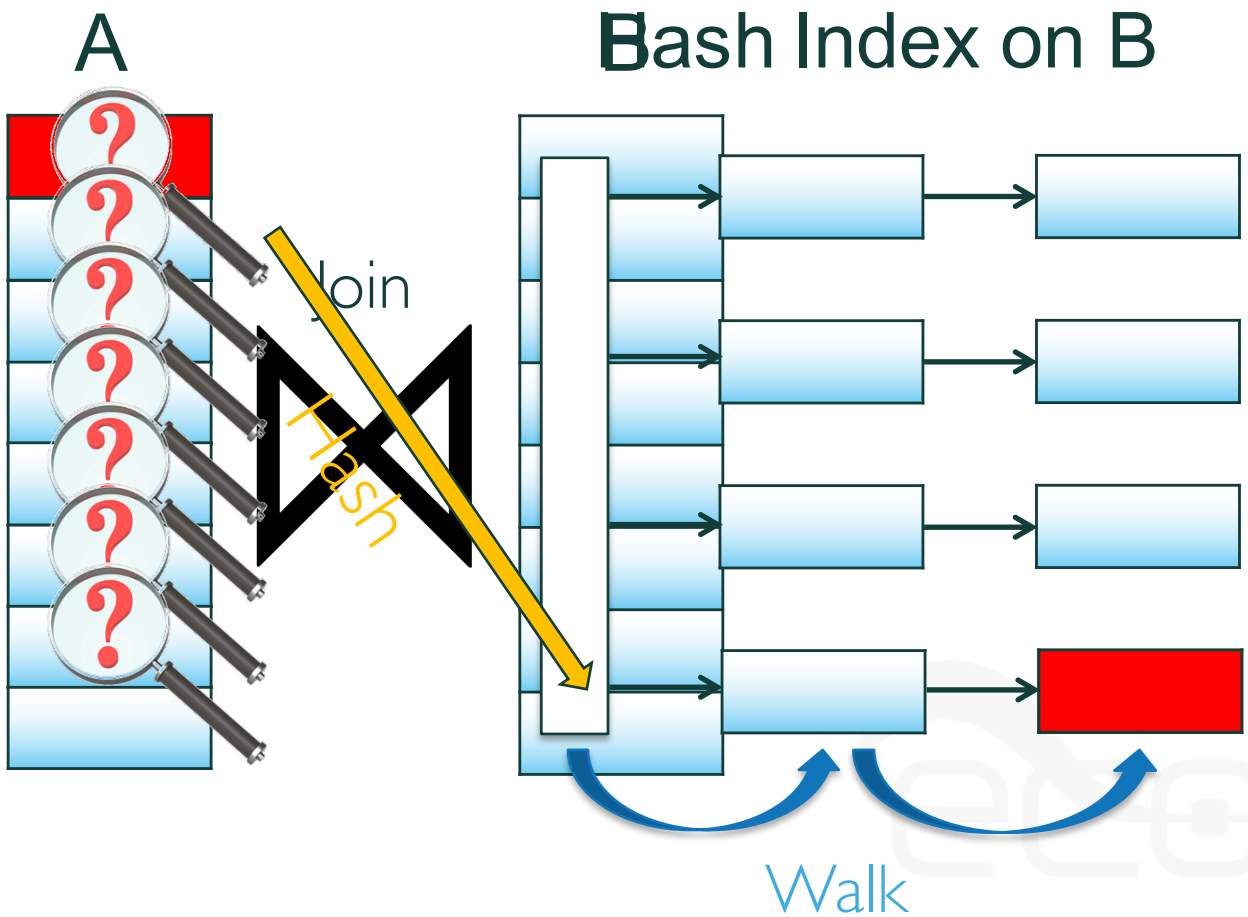


Dominant operation: join via hash index

Join via Hash Index

Lookup on index for every entry in A

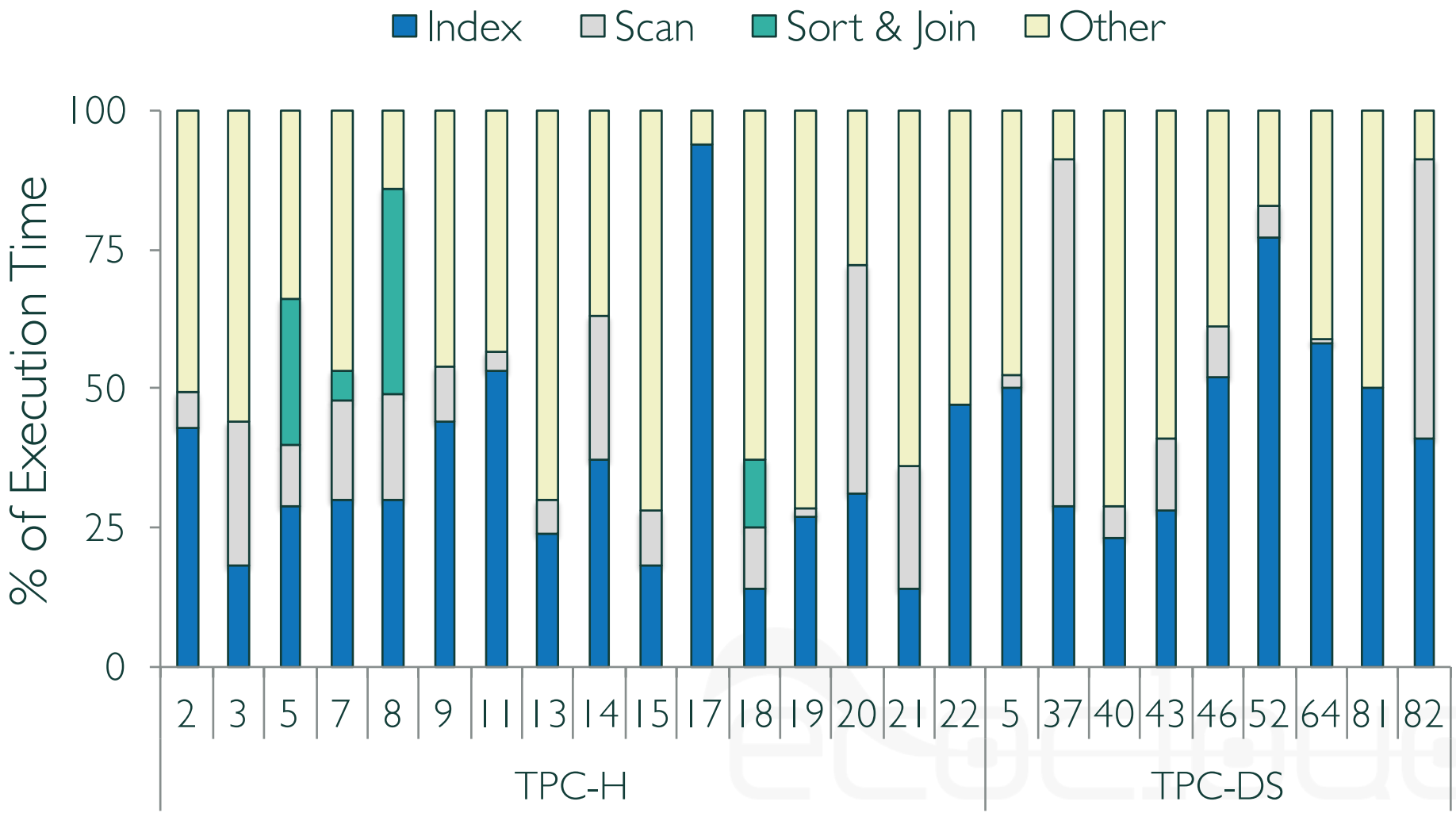
Join: find the matching values in A and B



How Much Time is Spent in Index Lookups?

ecocloud
an EPFL research center

Measurement on Xeon 5670 CPU with 100GB Dataset



Index lookup is the biggest contributor to time

Dissecting Index Lookups

ecocloud
an EPFL research center

Hash: avg. 30% time of each lookup

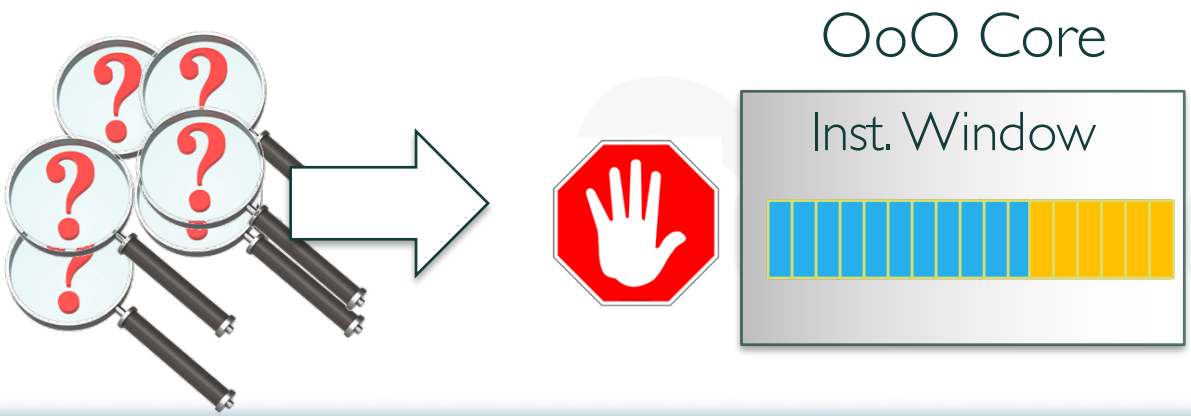
- Computationally intensive, high cache locality

Walk: avg. 70% time of each lookup

- Trivial computation, low cache locality

Next lookup: inherently parallel

- Beyond the instruction window capacity



Index Lookups on General-Purpose Cores

ecocloud
an EPFL research center

Index Lookups

- Data in memory
- Inherent parallelism

OoO Cores

- Pointer-chasing → Low MLP
- Limited OoO inst. window
  - One lookup at a time

OoO cores ill-matched to indexing

Roadmap for Efficient and High-Throughput Index Lookups

ecocloud
an EPFL research center

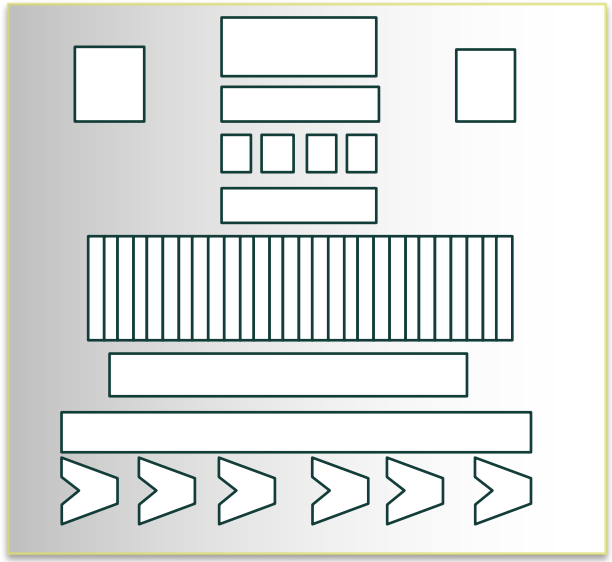
1. Specialize
   - Customize hardware for hashing and walking
2. Parallelize
   - Perform multiple index lookups at a time
3. Generalize
   - Use a programmable building block

Step 1: Specialize

ecocloud
an EPFL research center

Design a dedicated unit for hash and walk

- Hash: compute hash values from a key list
- Walk: access the hash index and follow pointers

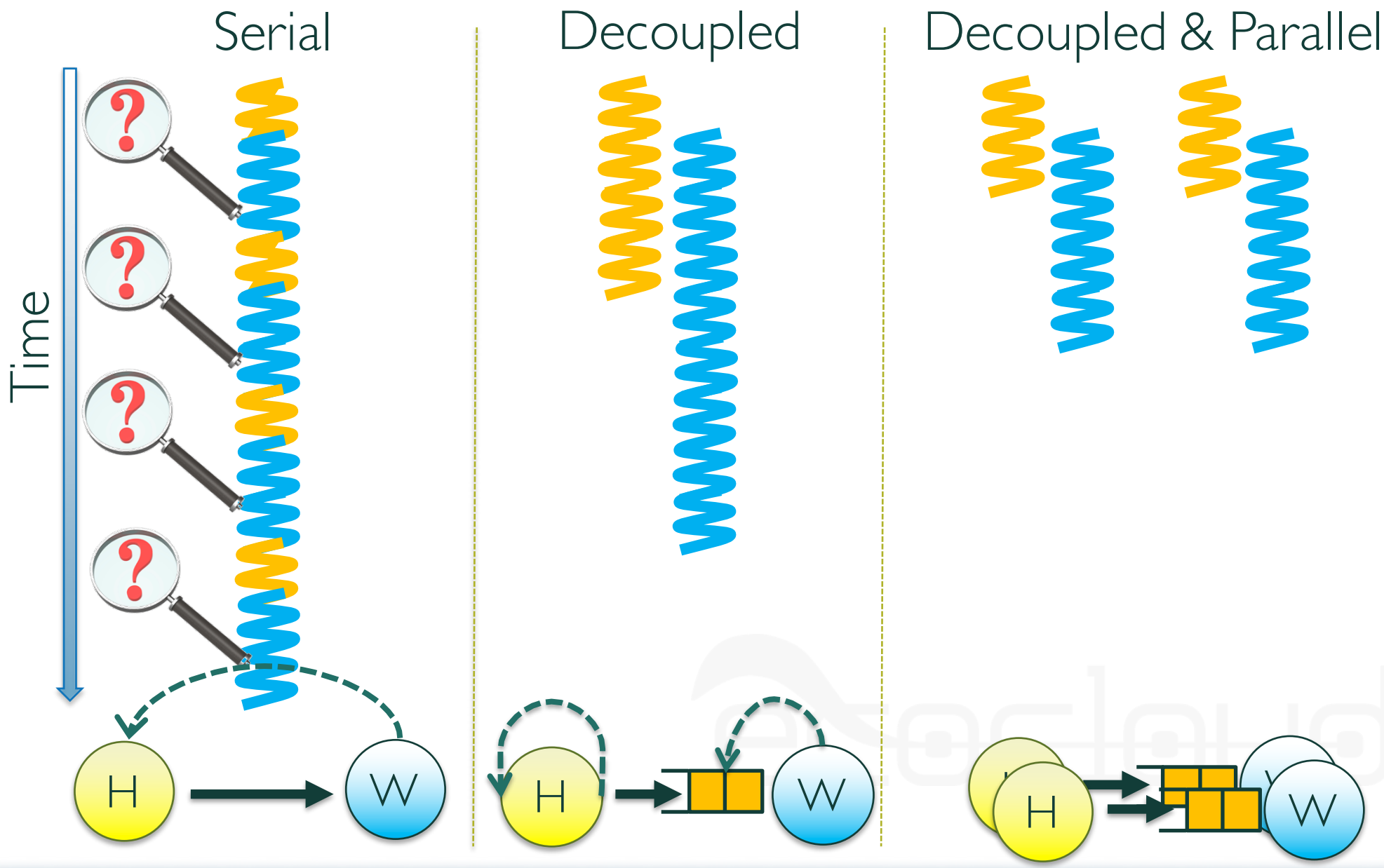


General-purpose OoO

Specialized hash and walk hardware

Step 2: Parallelize

ecocloud
an EPFL research center



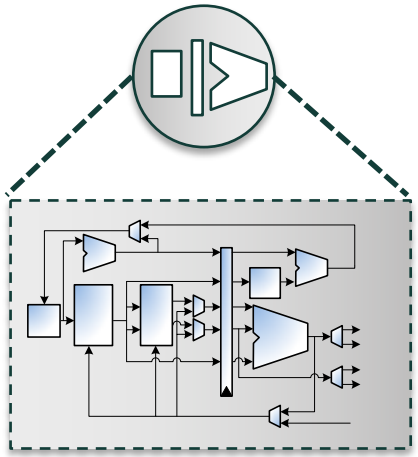
Step 3: Generalize Widx Units

ecocloud
an EPFL research center

Common building block for hash and walk

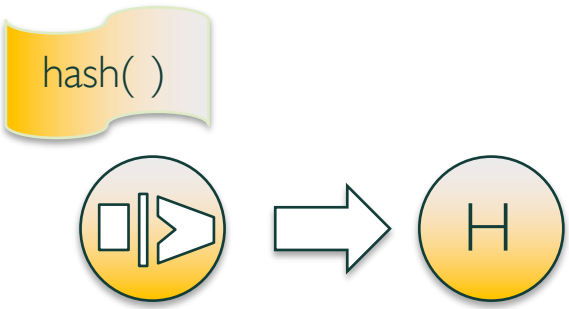
- Two-stage RISC core
- Custom ISA

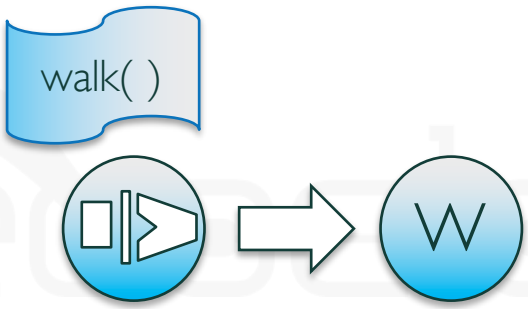
Widx unit



Programmable

- Execute functions written in Widx ISA
- Support limitless number of data structure layouts

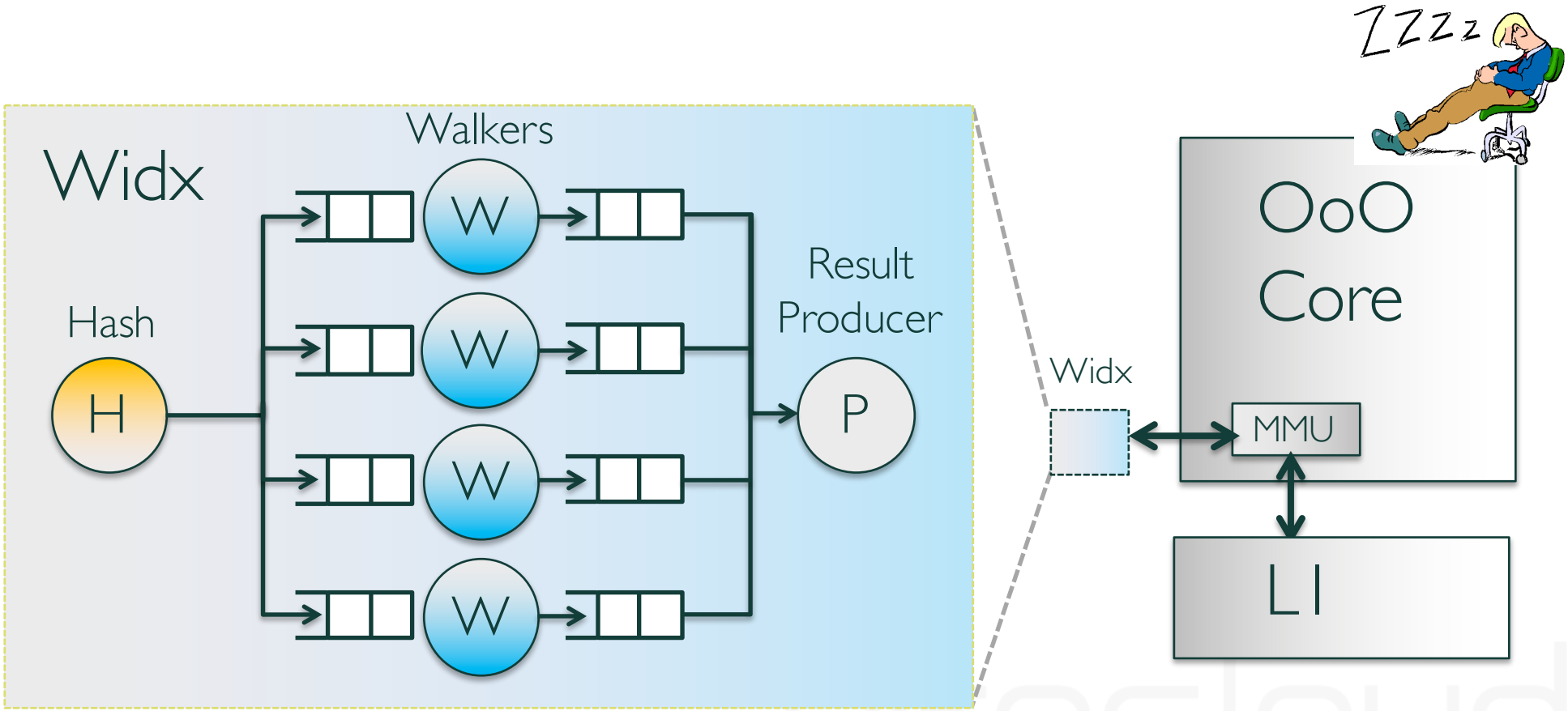




Putting it all together: Widx

ecocloud
an EPFL research center

When Widx runs, core goes idle



Simple, parallel hardware

Methodology

ecocloud
an EPFL research center

Flexus simulation infrastructure [Wenisch '06]

Benchmarks

- TPC-H on MonetDB
- TPC-DS on MonetDB
- Dataset: 100GB

uArch Parameters

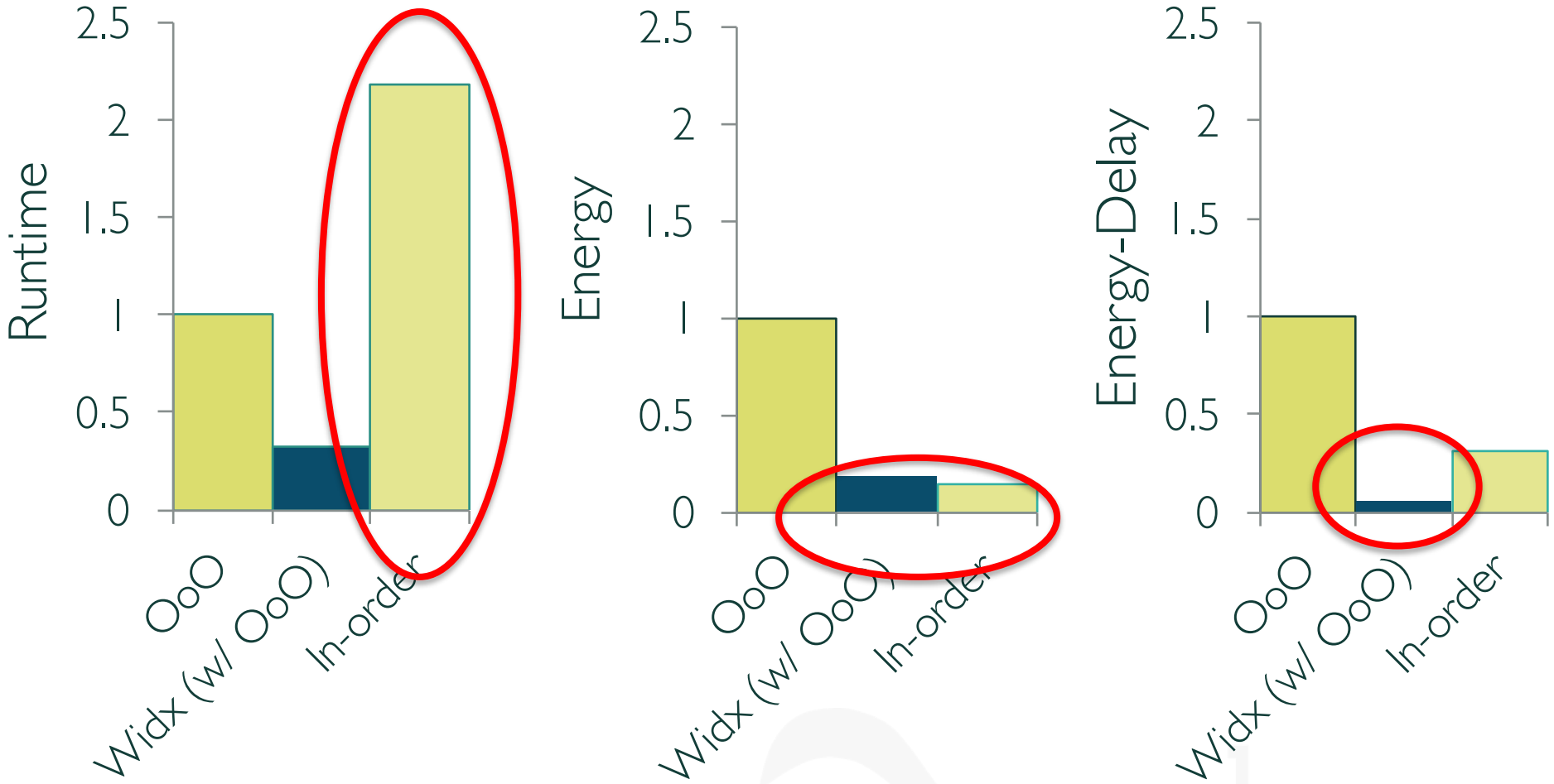
- Core Types
  - OoO: 4-wide, 128-entry ROB
  - In-order: 2-wide
- Frequency: 2GHz
- L1 (I & D): 32KB
- LLC: 4MB

Area and Power

- Synopsys Design Compiler
- Technology node: TSMC 40 nm, std. cell
- Frequency: 2GHz
- Widx Area: 0.24mm^2
- Widx Power: 0.3W

Widx Efficiency

ecocloud
an EPFL research center



5.5x reduction in indexing energy vs. OoO core

Asynchronous Memory Access Chaining [VLDB'16]

ecocloud
an EPFL research center

Use insights to help Xeon servers

- Decouple hash & walk in software
- Create & manage walker queues in software wraparound

2.3x speedup on Xeon

- Unclogs the internal microarchitecture
- Maximizes memory level parallelism

Summary

ecocloud
an EPFL research center

Trends for data & online services:

- Data growing at exponential rate
- Online services are in-memory
- Memory is a big fraction of TCO

Specialize servers around DRAM

- Opportunities abound
- Processors, accelerators, memory, network, system
- E.g., accelerators for database management

Thank You!

ecocloud
an EPFL research center

For more information please visit us at

ecocloud.ch

ecocloud

EPFL
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE