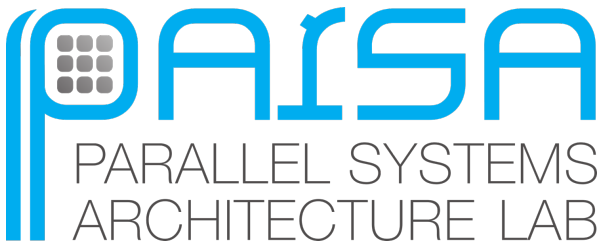
WHAT'S HOT?
POST-MOORE DATACENTER ARCHITECTURE

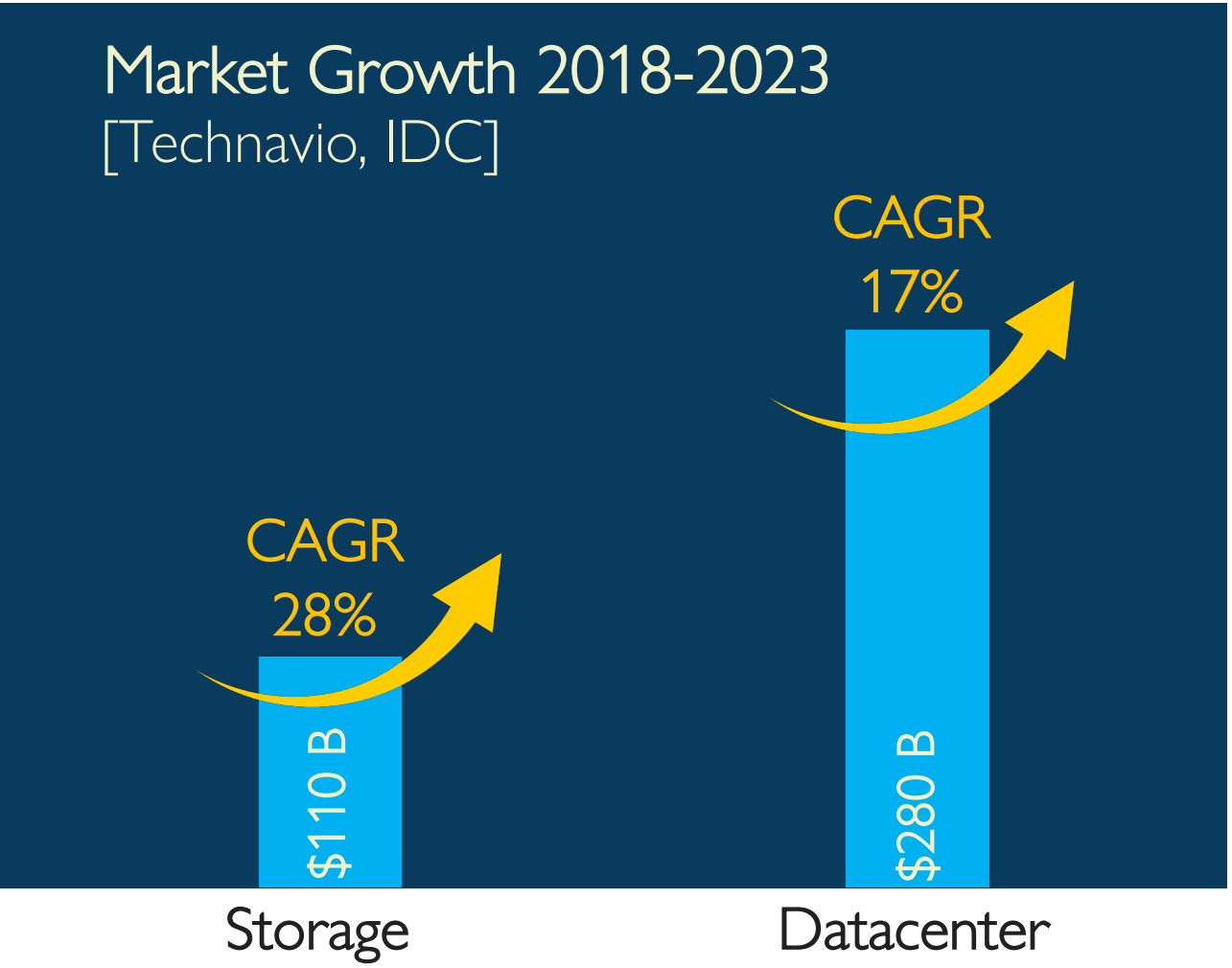
Babak Falsafi



parsa.epfl.ch

DATACENTER GROWTH

PARALLEL SYSTEMS ARCHITECTURE LAB



- Data → fuel for digital economy
- Exponential demand for digital services
- Many apps (e.g., AI) with higher exponential demand

DATACENTERS ARE BACKBONE OF CLOUD

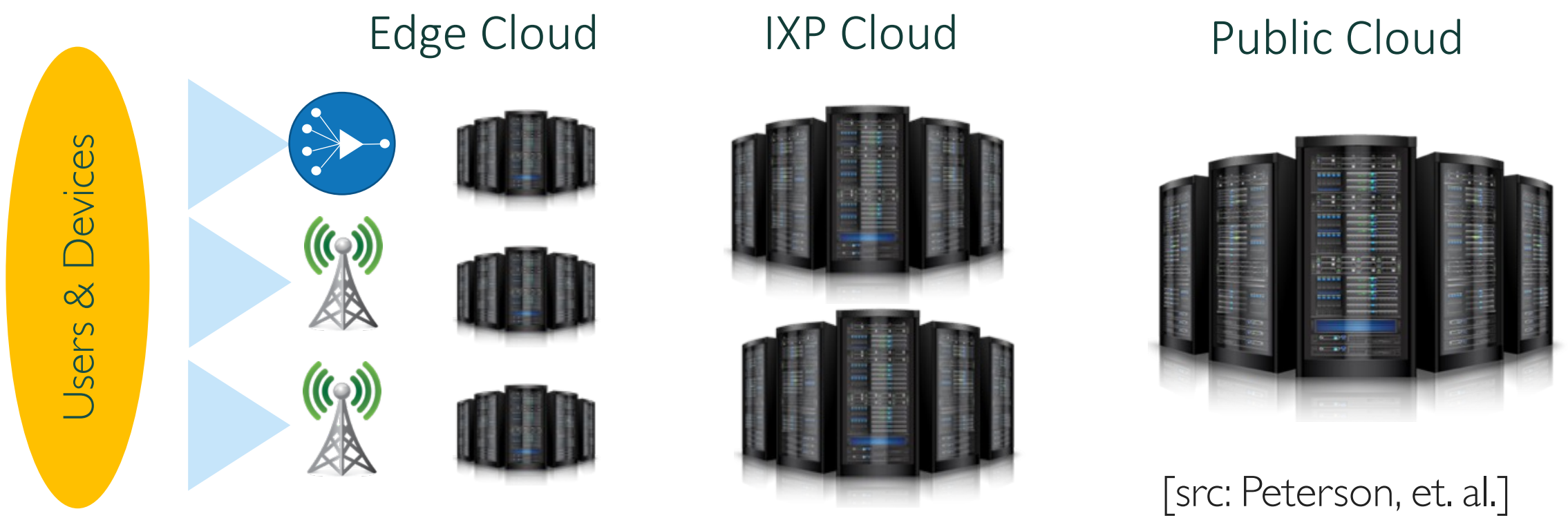
PARALLEL SYSTEMS ARCHITECTURE LAB

- 100s of 1000 of commodity or home-brewed servers
- Centralized to exploit economies of scale
- Network fabric w/ µ-second connectivity
- Often limited by
  - Electricity
  - Network
  - Cooling

350MW, Boydton

CLOUDS AT VARIOUS SCALES

PARALLEL SYSTEMS ARCHITECTURE LAB



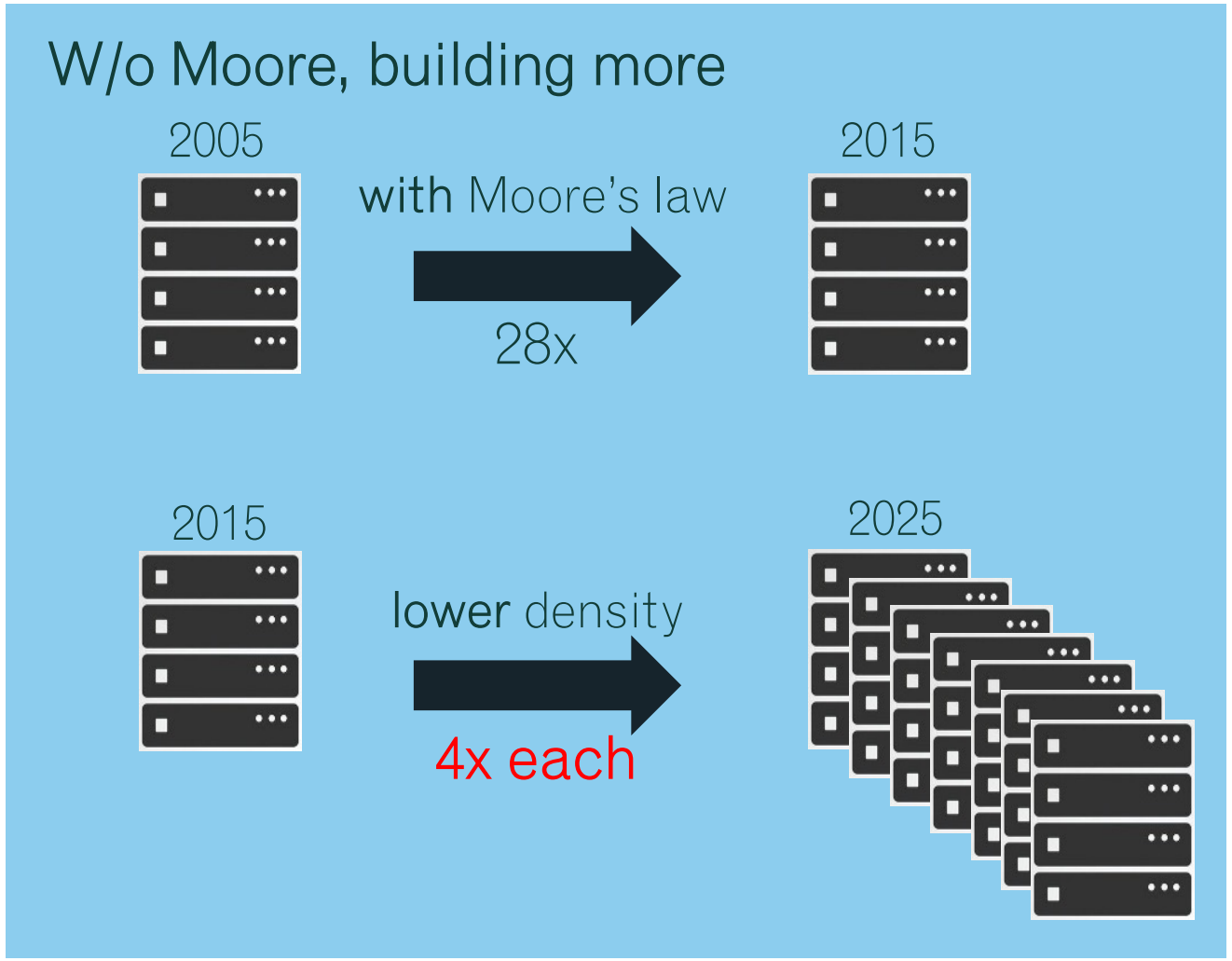
[src: Peterson, et. al.]

Temporal/Sensitive/Local Data

Persistent/Global Data

DATACENTERS NOT GETTING DENSER

PARALLEL SYSTEMS ARCHITECTURE LAB



End of Moore's Law (of Silicon)

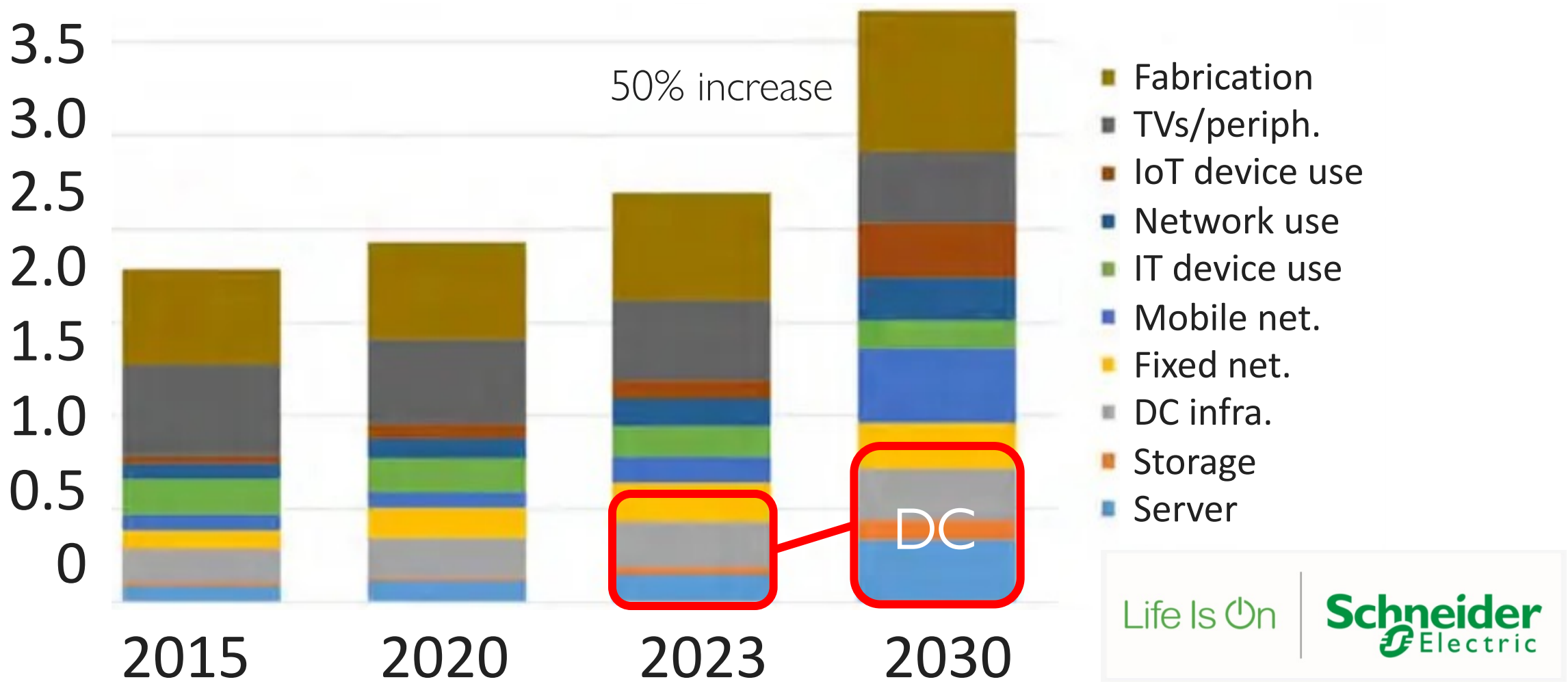
- Five decades of doubling density
- Recent slowdown in density
- Chip density limited by physics

Growth means building more

- 41%/year → 28x in ten years
- At 15%/year → 7x more DCs

ELECTRICITY IN 1000TWH

[Digital economy & climate impact, May 2022]



PARALLEL SYSTEMS ARCHITECTURE LAB

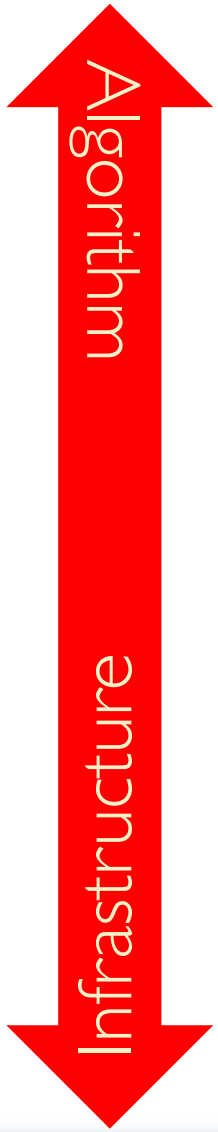
Life Is On Schneider Electric

POST-MOORE DATACENTERS

PARISA
PARALLEL SYSTEMS ARCHITECTURE LAB

Design for "ISA"

- Integration
  - Move data less frequently
  - Move data less distance
- Specialization
  - Customize resources
  - Less work/computation
- Approximation
  - Adjust precision



OUTLINE

PARALLEL SYSTEMS ARCHITECTURE LAB

- ~~Overview~~
- Post-Moore servers
  - Today's servers
  - ISA opportunities
- Datacenter sustainability
- Summary

SCALE-OUT DATACETNERS

PARALLEL SYSTEMS ARCHITECTURE LAB

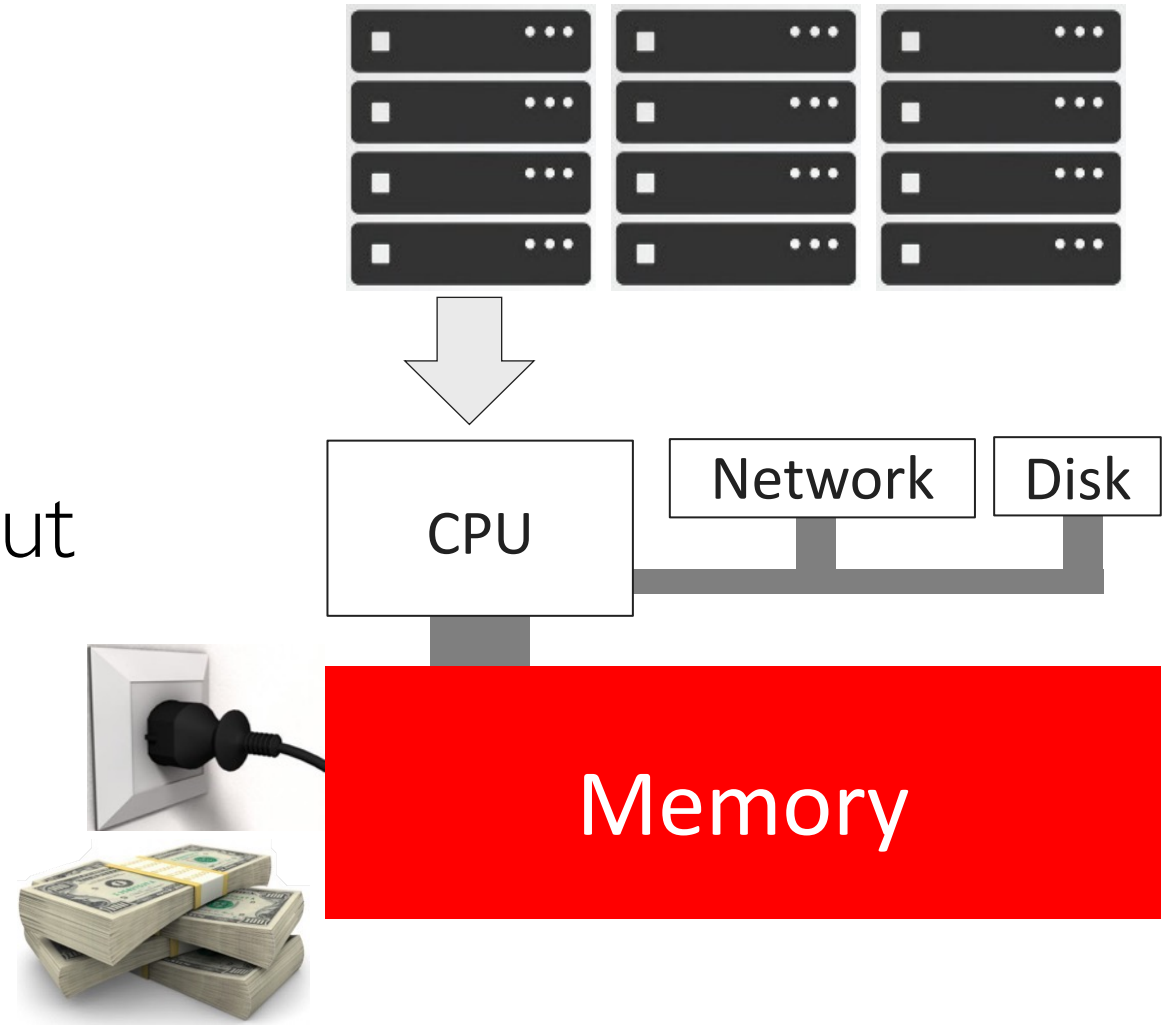
Cost is the primary metric

Online services hosted in memory

Divide data up across servers

Design server for low cost, scale out

☞Memory most precious silicon



TODAY'S SERVERS

PARALLEL SYSTEMS ARCHITECTURE LAB

Today's platforms are PCs of the 80's

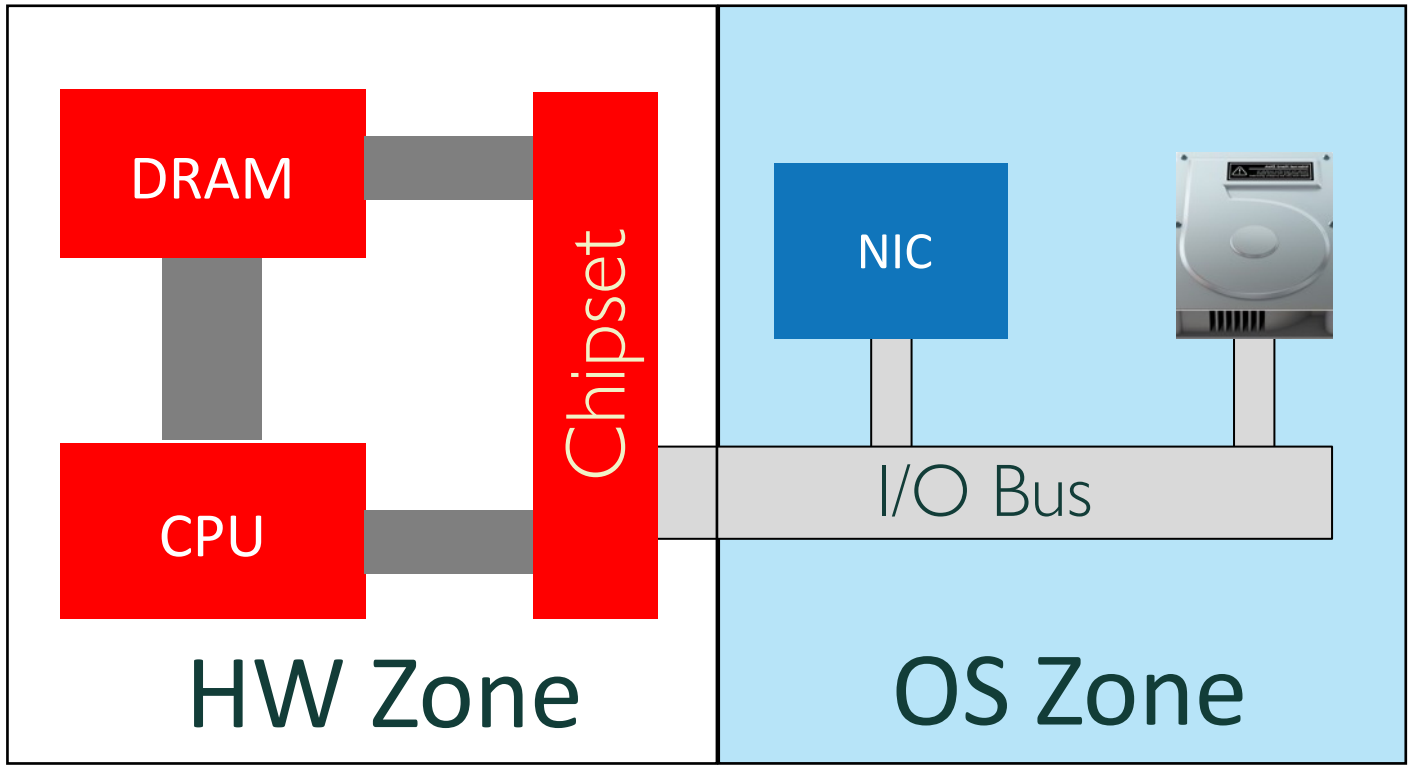
- CPU "owns" and manages memory
- OS moves data back/forth from peripherals
- Legacy interfaces connecting the CPU/mem to outside
- Legacy POSIX abstractions

Fragmented logic/memory:

- Manycore network cards w/ own memory
- Flash controllers with embedded cores and memory
- Discrete accelerators with own memory

80S' DESKTOP

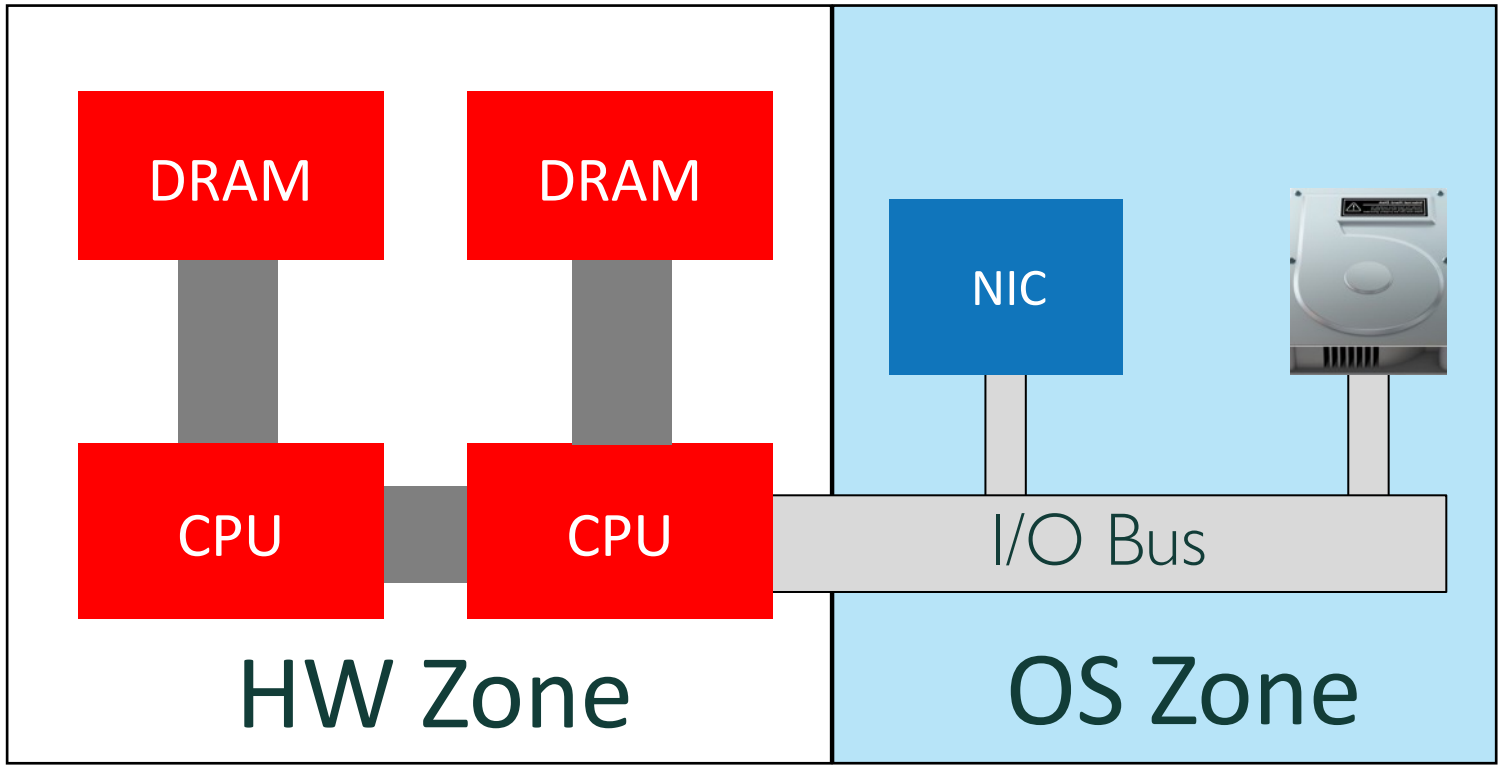
PARALLEL SYSTEMS ARCHITECTURE LAB



- 33 MHz 386 CPU, 250ns DRAM
- OS: Windows, Unix BSD (or various flavors)
- Focus: mlutiprogrammed in-memory compute

TODAY'S SERVER: 80S' DESKTOP

PARISA
PARALLEL SYSTEMS ARCHITECTURE LAB

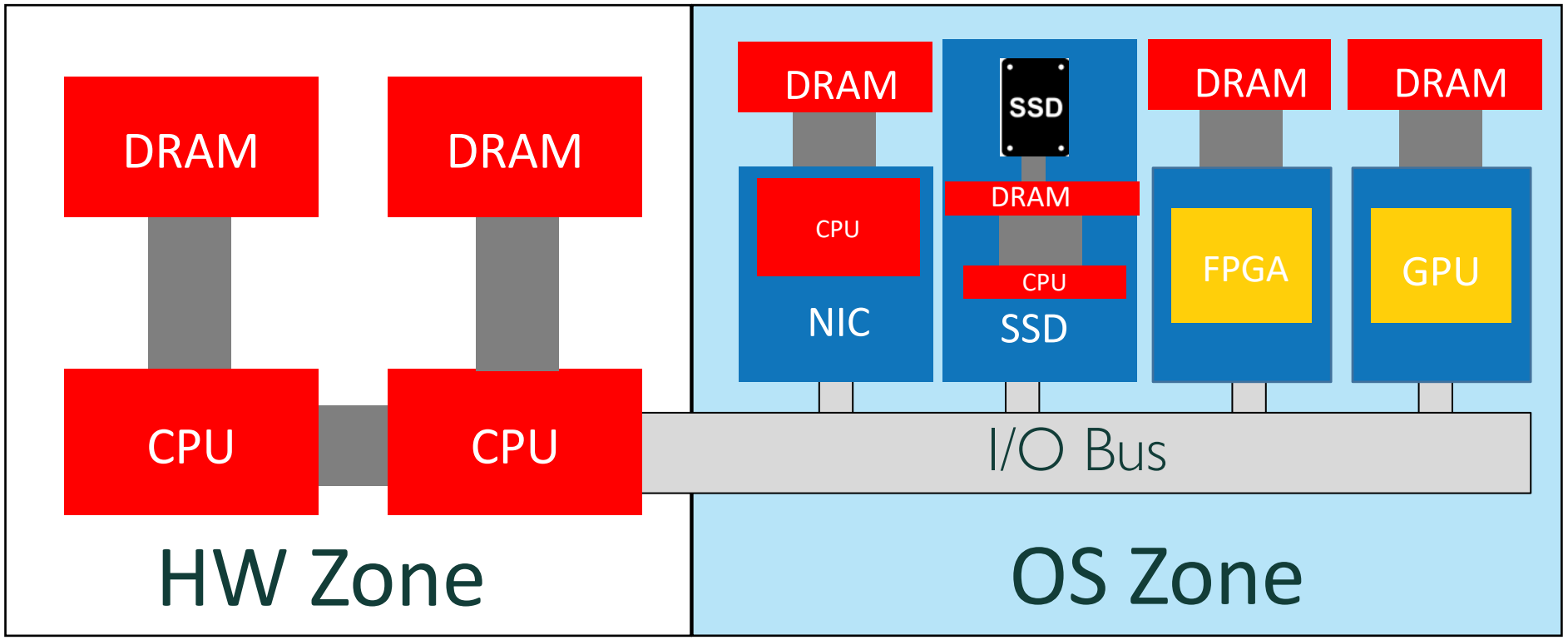


Red Hat

- Dual 2GHz CPU's, 50ns DRAM
- OS: Linux (and various distributions)

TODAY'S SERVER: 80S' DESKTOP

PARISA
PARALLEL SYSTEMS ARCHITECTURE LAB

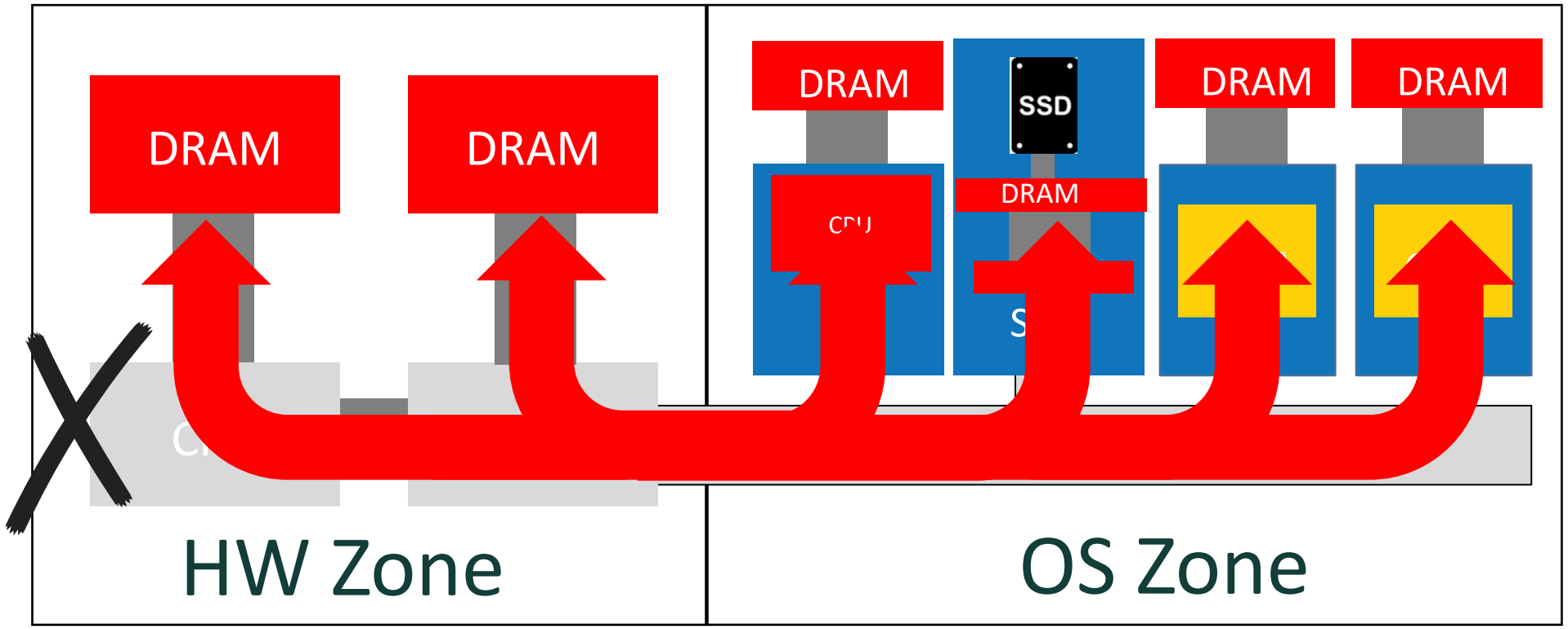


- Dual 2GHz CPU's, 50ns DRAM, Linux
- Bottlenecked by legacy interfaces
- Fragmented silicon

Red Hat

TODAY'S SERVER: 80S' DESKTOP

PARALLEL SYSTEMS ARCHITECTURE LAB

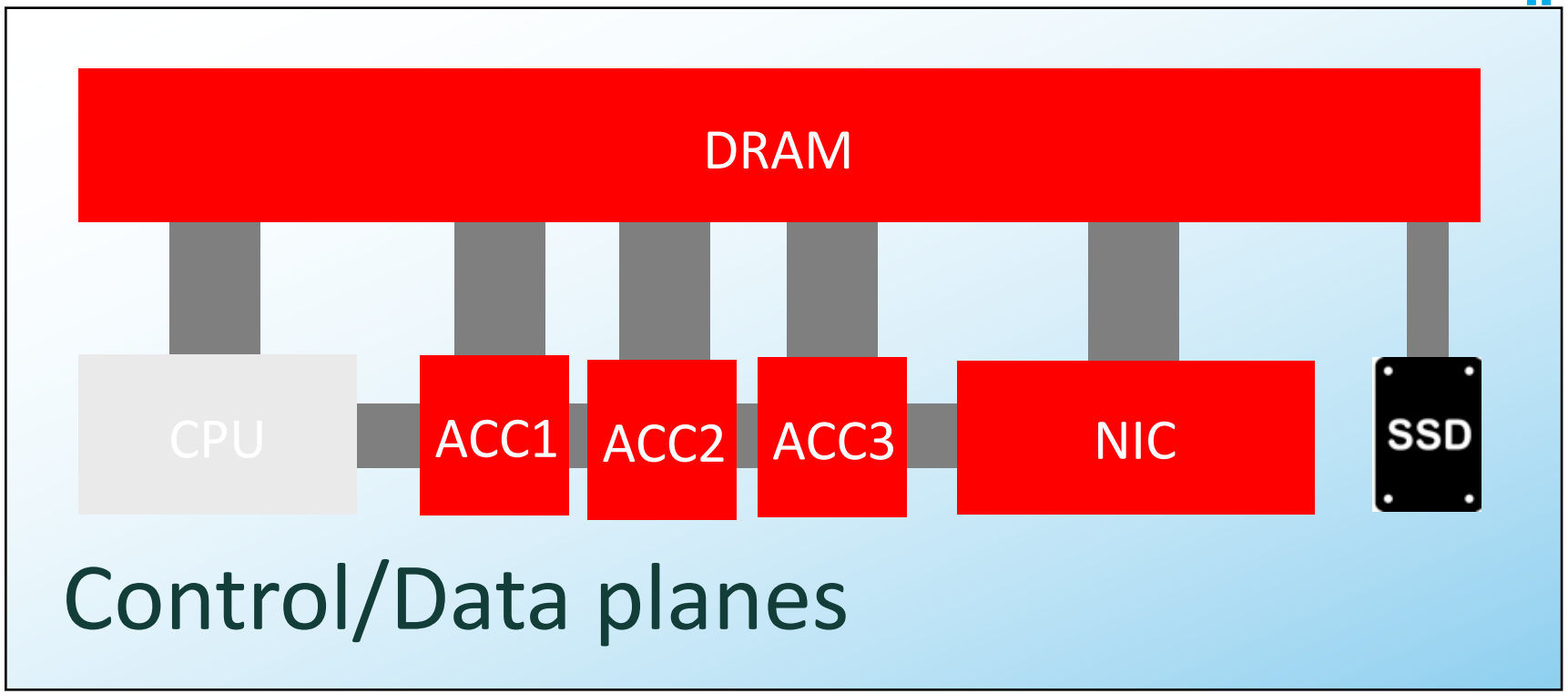


- Dual 2GHz CPU's, 50ns DRAM, Linux
- Bottlenecked by CPU, OS & legacy interfaces
- Fragmented silicon

RedHat

IDEAL POST-MOORE SERVER

PARALLEL SYSTEMS ARCHITECTURE LAB



- Think of the server as a network
- Control plane: set up via CPU & OS
- Data plane: protected access to memory
- Eliminates silicon fragmentation

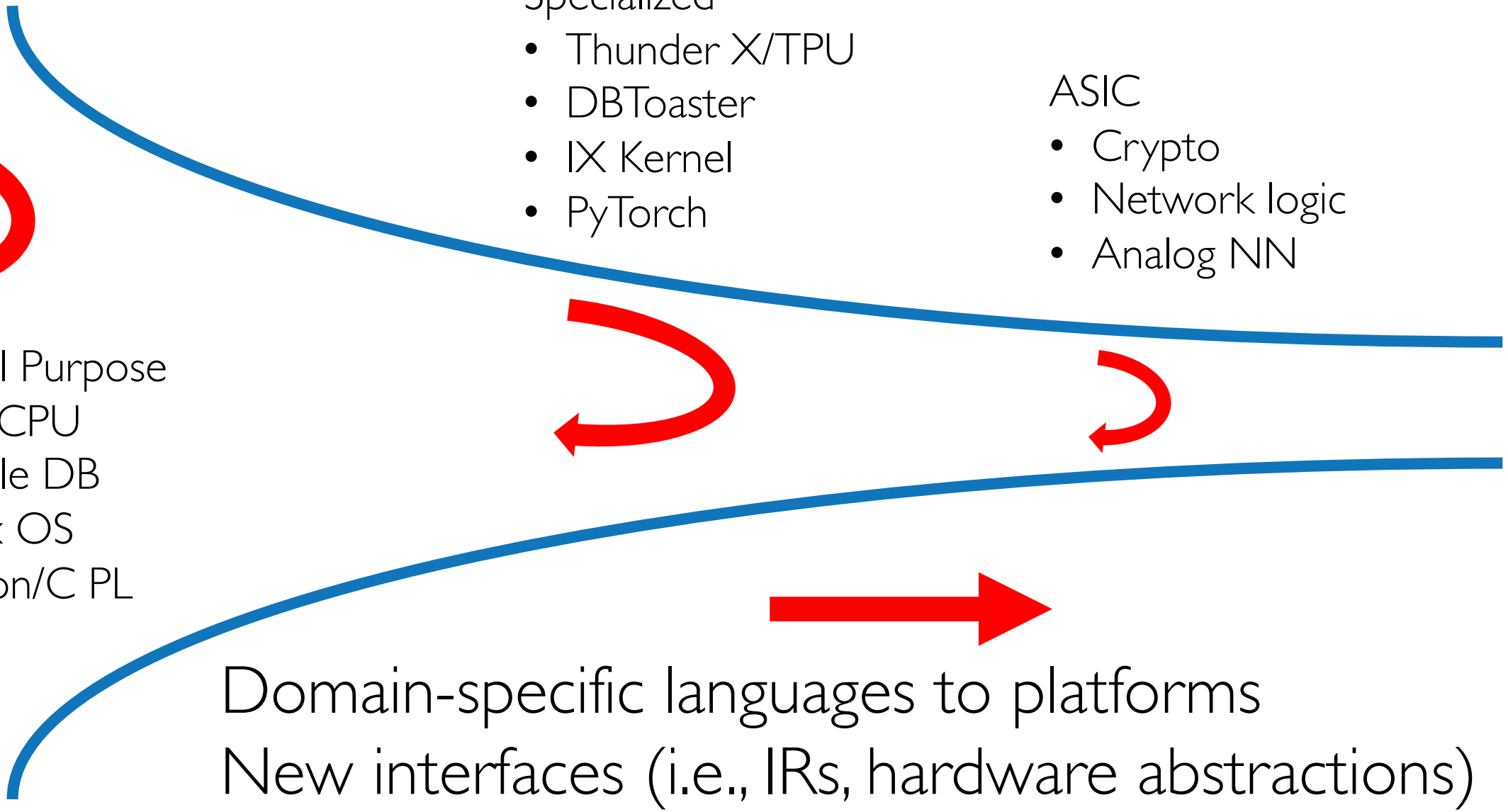
OUTLINE

PARISA
PARALLEL SYSTEMS ARCHITECTURE LAB

- ~~Overview~~
- Post-Moore servers
  - ~~Today's servers~~
  - ISA opportunities
    - CPU/Memory/Storage/Network
    - AI
- Datacenter sustainability
- Summary

THE SPECIALIZATION FUNNEL

PARALLEL SYSTEMS ARCHITECTURE LAB



Specialized
- Thunder X/TPU
- DBToaster
- IX Kernel
- PyTorch

ASIC
- Crypto
- Network logic
- Analog NN

General Purpose
- Intel CPU
- Oracle DB
- Linux OS
- Python/C PL
- …..

Domain-specific languages to platforms
New interfaces (i.e., IRs, hardware abstractions)

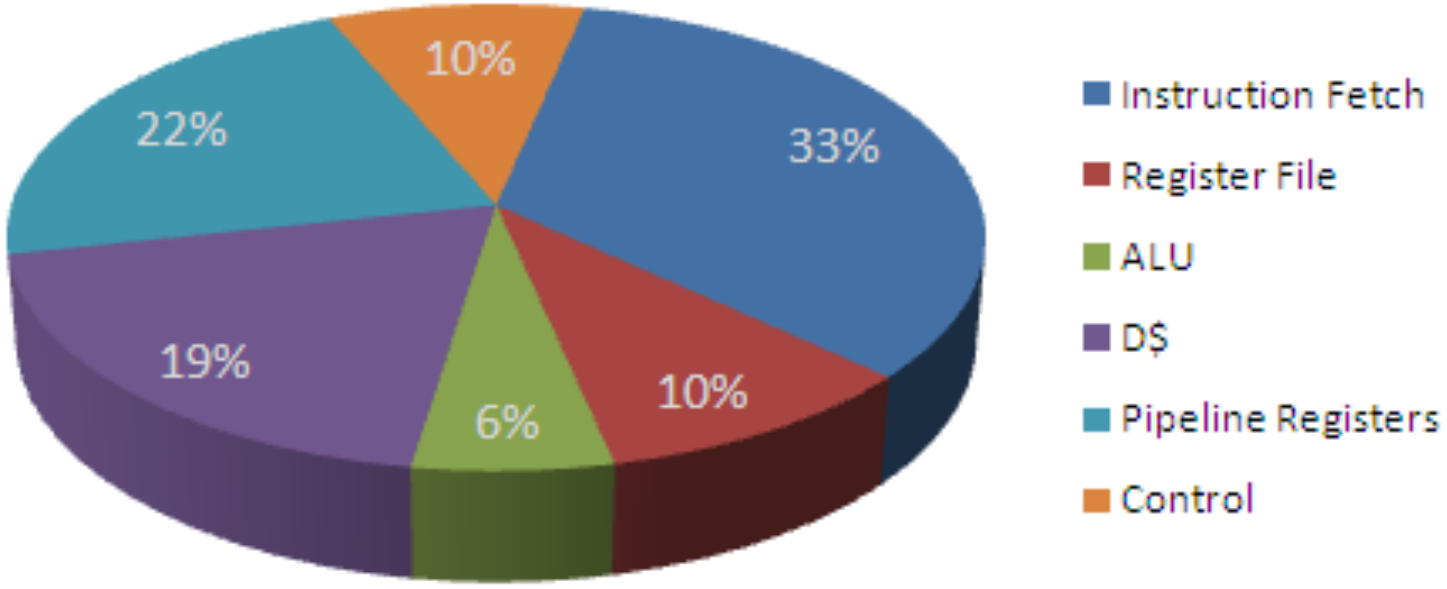
THE LIMITS OF CPUS

PARALLEL SYSTEMS ARCHITECTURE LAB

CPUs follow the von Neumann machine organization

- Machine instructions fetched from memory
- Operands fetched/written to memory
- Referred to as von Neumann bottleneck

Only 6% power in Pentium 4 spent in arithmetic (ALU)



[src: Chen, et. al., IEEE Transactions, 2006]

CPU SPECLIAZATION FOR WORKLOADS

PARALLEL SYSTEMS ARCHITECTURE LAB

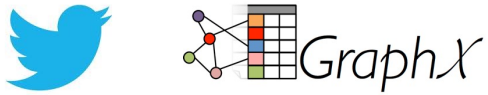
- First-party workloads (e.g., search, retail, media)
  - Data management
  - Analytics
  - Monoliths to microservices
- Third-party workloads (cloud)
  - Containerized
  - Serverless

CloudSuite (4.0 release @ cloudsuite.ch)

PARALLEL SYSTEMS ARCHITECTURE LAB

Data Analytics
Machine learning

Graph Analytics
GraphX



In-Memory Analytics
Recommendation System

Web Search
Apache Solr & Nutch

Media Streaming
Nginx, HTTP Server

Web Serving
Nginx, PHP server

Data Caching
Memcached

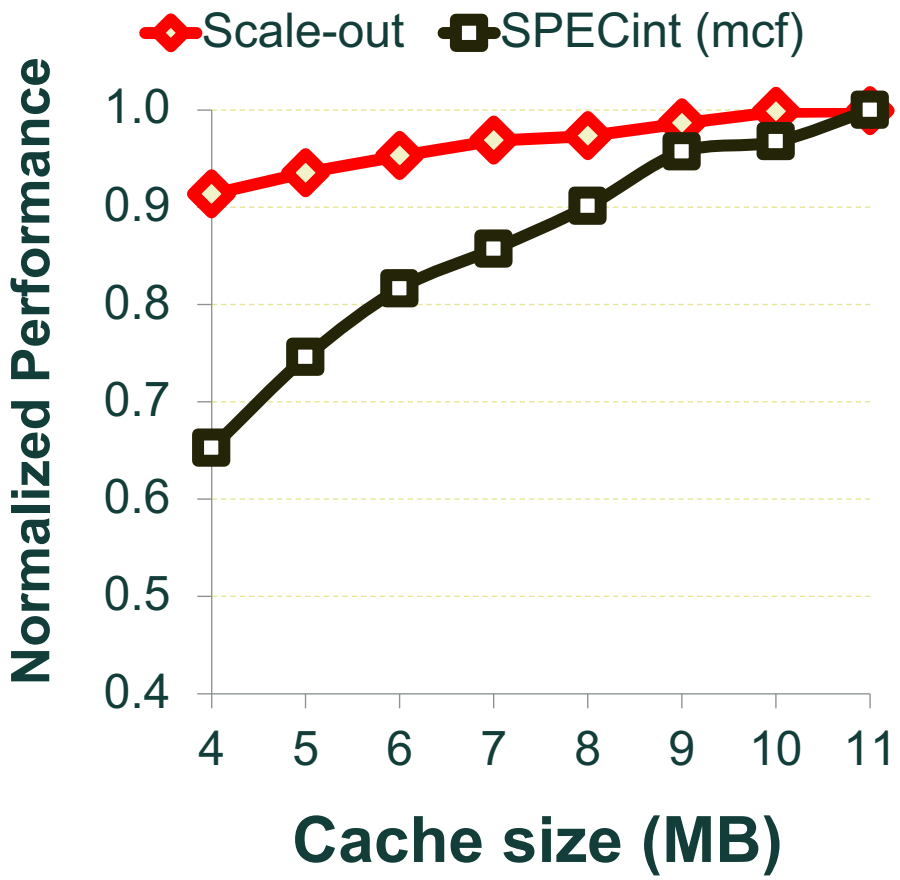
Data Serving
Cassandra NoSQL

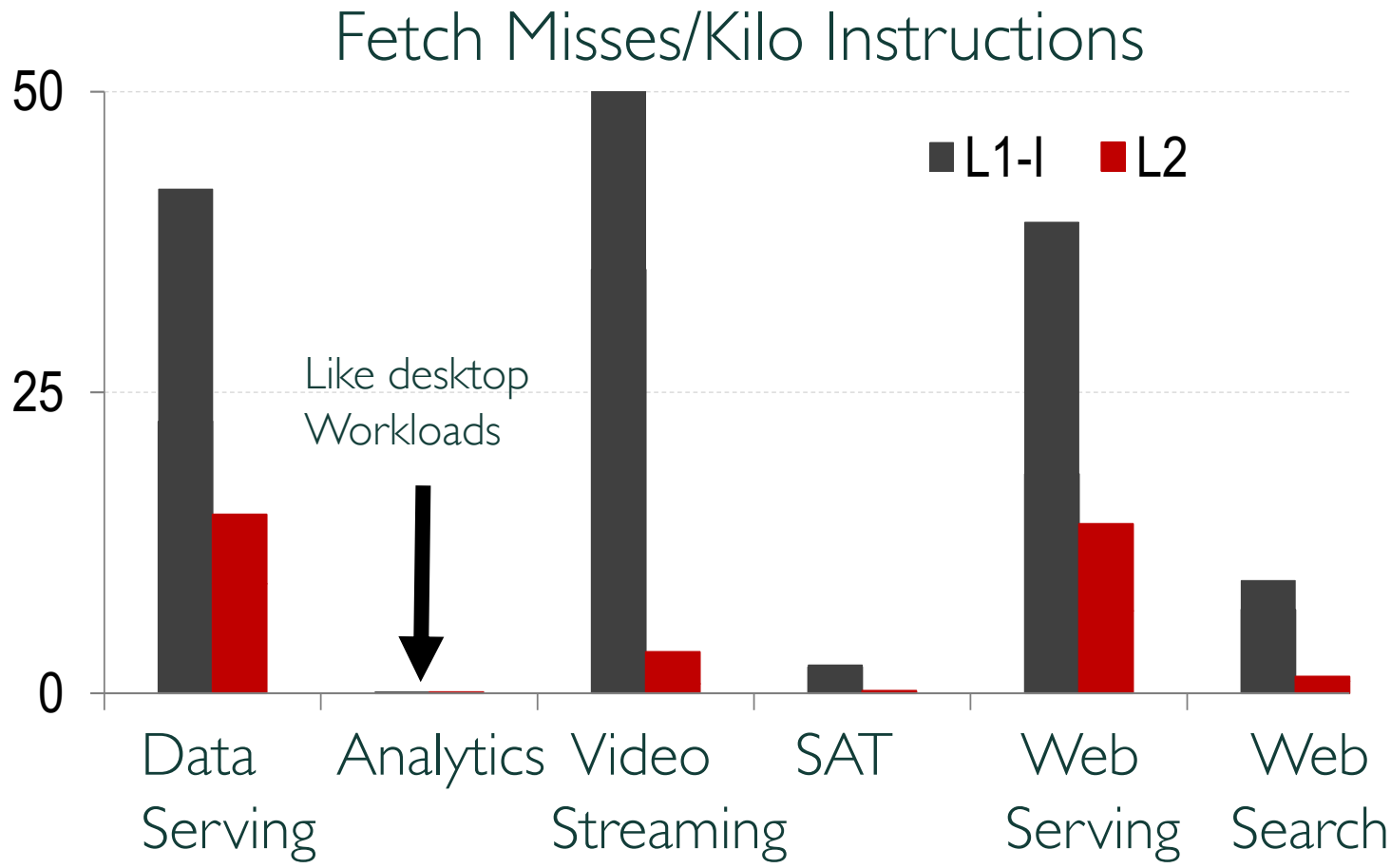
Supports x86, ARM64, RISC-V (coming)

intel

SERVICES STUCK IN MEMORY [ASPLOS'12]

PARSA
PARALLEL SYSTEMS ARCHITECTURE LAB



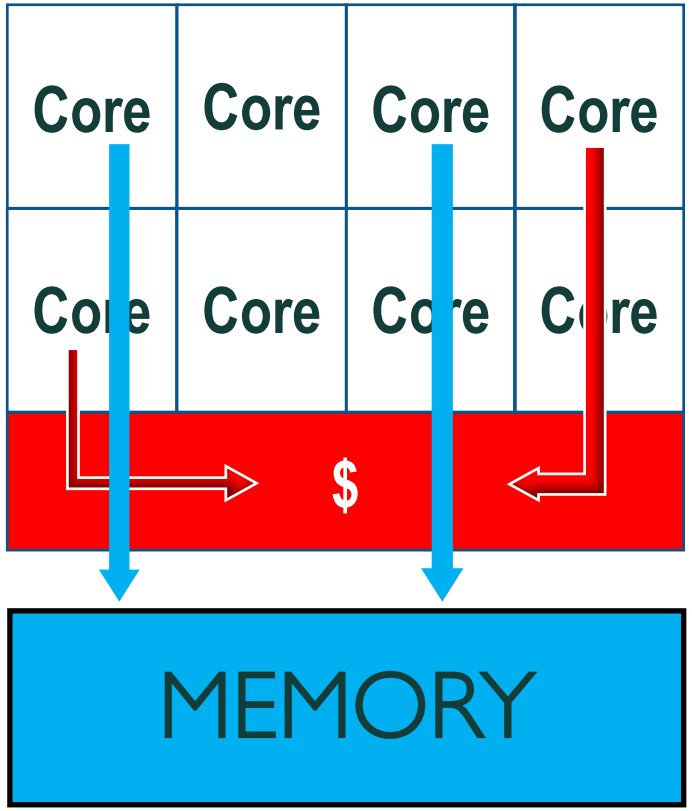


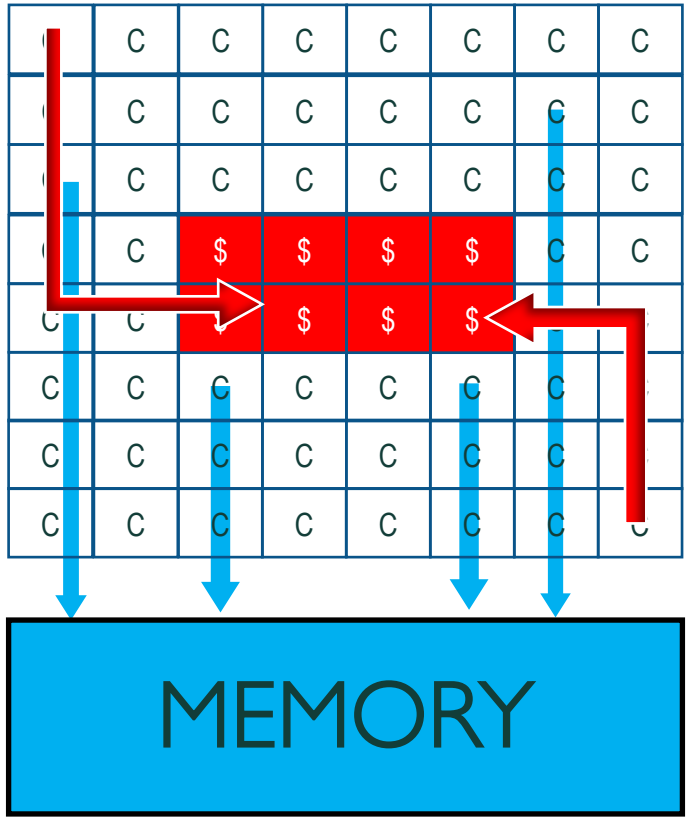
Cache overprovisioned

Instruction supply bottlenecked

SCALE-OUT PROCESSOR (SOP)

PARALLEL SYSTEMS ARCHITECTURE LAB





- General-purpose CPU
- ✗ Logic 60% of silicon
- ✗ 6x bigger cores

- 3-way OoO ARM
- ✓ 85% logic, 7x more cores
- ✓ Faster instruction supply

CLOUD-NATIVE CPU [c.a. 2014]

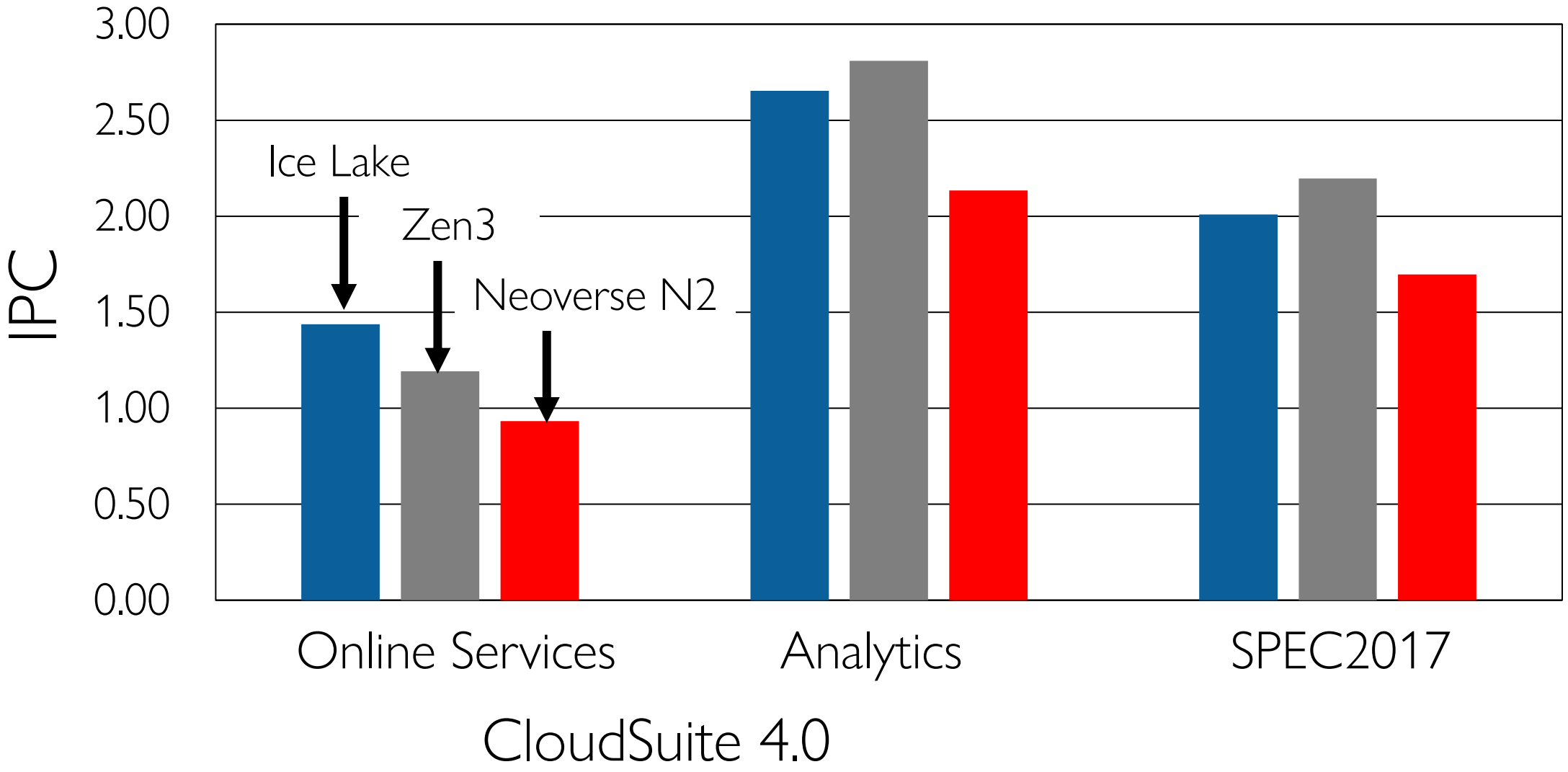
PARISA
PARALLEL SYSTEMS ARCHITECTURE LAB

Thunder X

- Based on SOP blueprint
- Designed to serve data
- 7x more core than cache
- Optimizes instruction supply
- Ran stock software
- 10x throughput over Xeon

X86 VS. ARM SINGLE THREAD

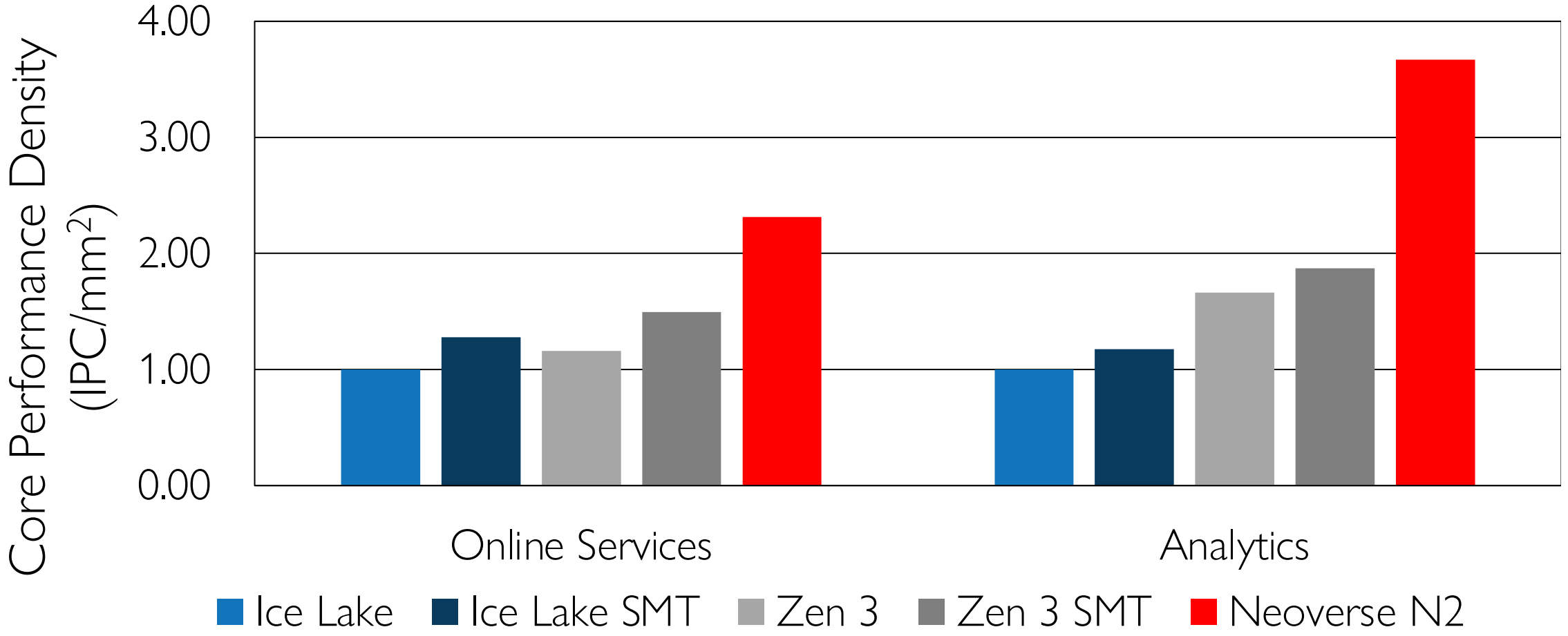
PARALLEL SYSTEMS ARCHITECTURE LAB



CloudSuite 4.0

SILICON EFFICIENCY IN X86 VS. ARM

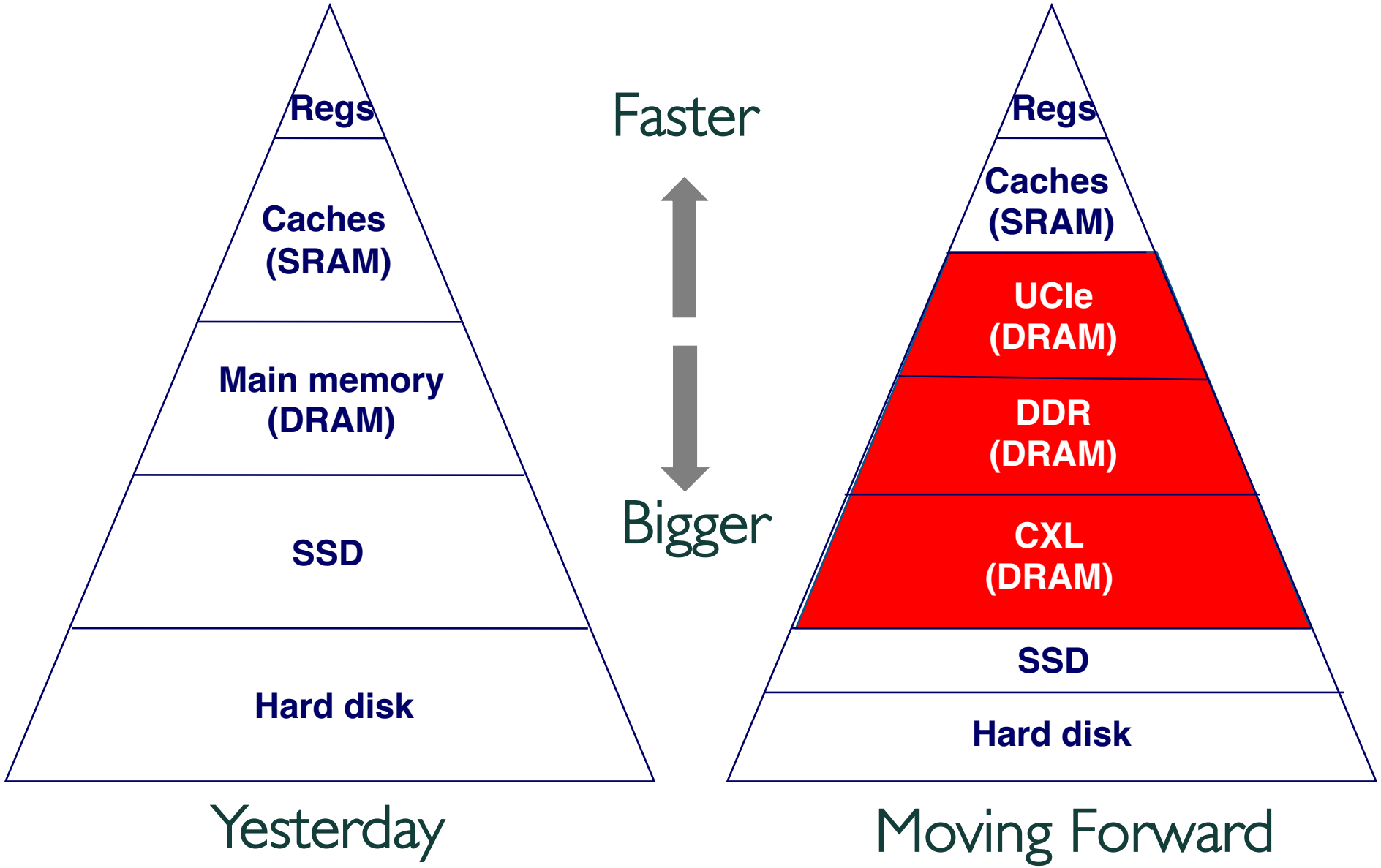
PARALLEL SYSTEMS ARCHITECTURE LAB



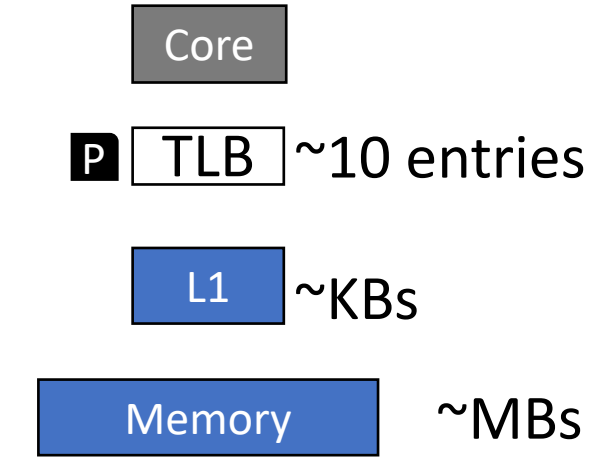
The online services meet end-to-end tail latency requirements

TB-SCALE HIERARCHIES

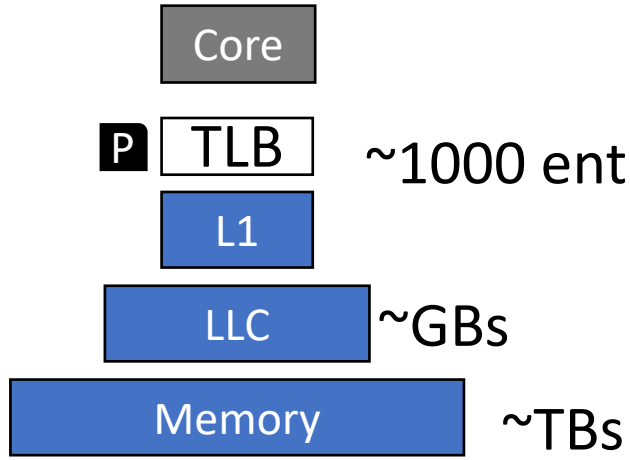
PARALLEL SYSTEMS ARCHITECTURE LAB



THE VM BOTTLENECK: TLBs



Platforms in 90s



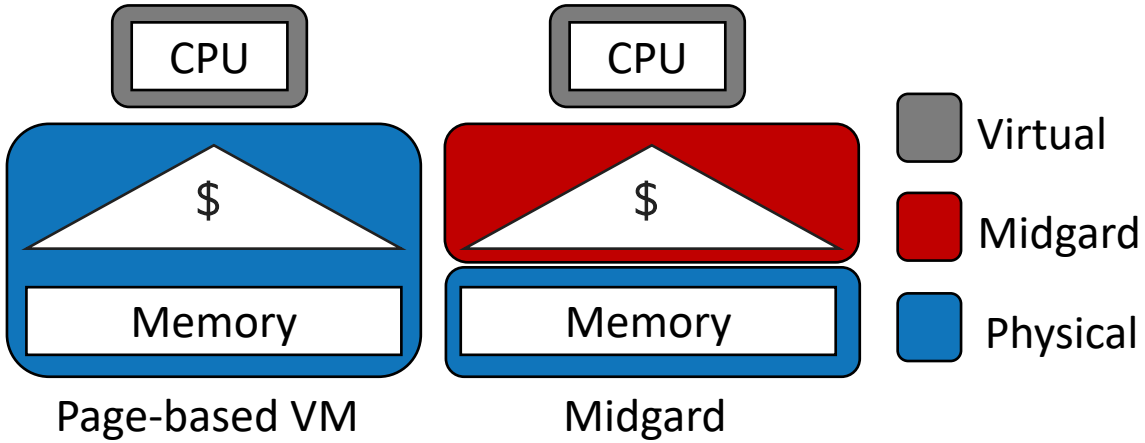
Platforms today

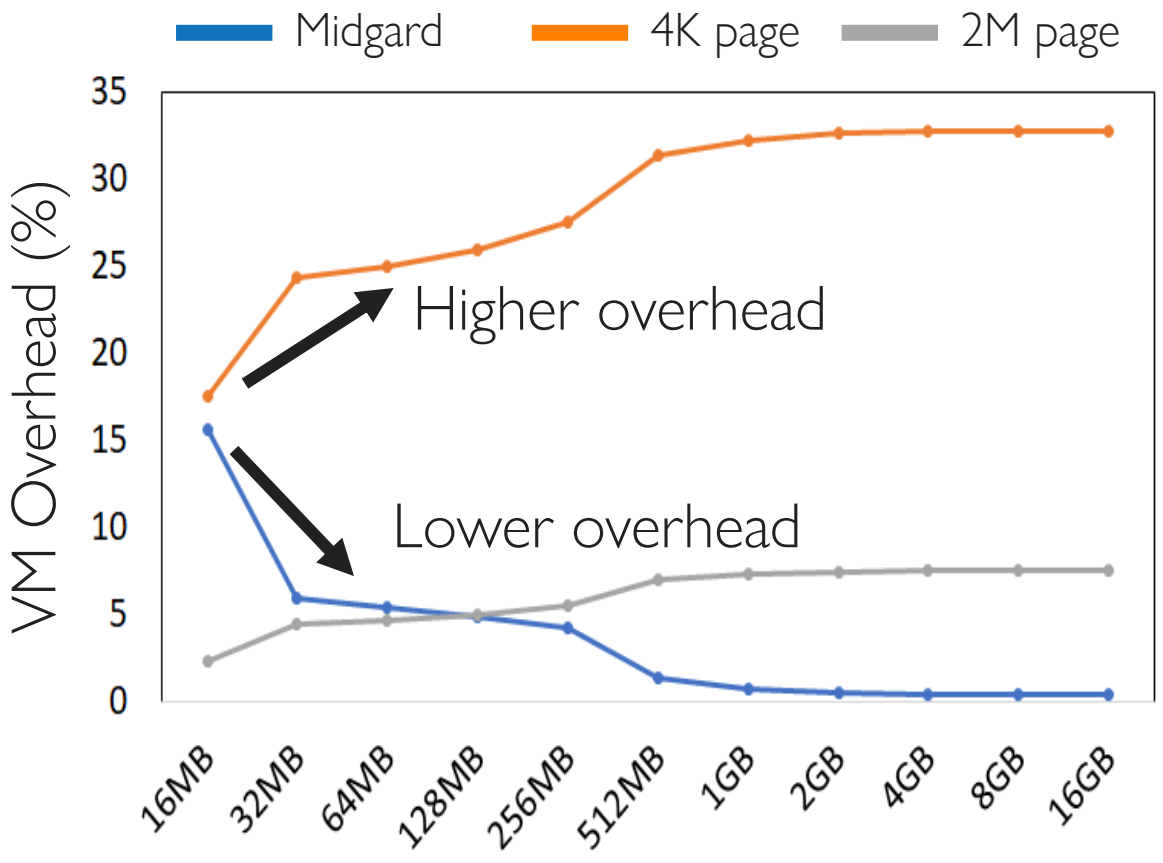
Product	Year	Cores	Cache capacity	TLB entries	Coverage (4KB)
Intel P4	2000	1	256KB SRAM	64	256KB
Intel KabyLake	2016	4	128MB eDRAM	1536	6MB
Apple M1	2020	8 (4+4)	16MB SRAM	3096	12MB (16KB)
AMD Zen3	2021	64 (8x8)	256MB SRAM	2048	8MB
Intel Sapphire Rapids	2022	56 (14x4)	64GB HBM2	?	?

TB-SCALE MEMORY WITHOUT TLB

midgard

midgard.epfl.ch





- Keeps POSIX (VMA) interface to apps
  - Linux, MacOS/iOS, Android
- Eliminates page-based translation in $

✓ Unclogs virtual memory for security, virtualization, accelerators

intel

TB-SCALE MEMORY WITHOUT TLB

midgard

Midgard Roadmap:

- CPU microarchitecture/OS [ISCA'21'23]
- Compartmentalization [IEEE S&P'23]
- Virtualization/Containerization
- Accelerator ecosystem/IO
- Monolith/μservices/serverless
- ….

EPFL

THE UNIVERSITY OF EDINBURGH

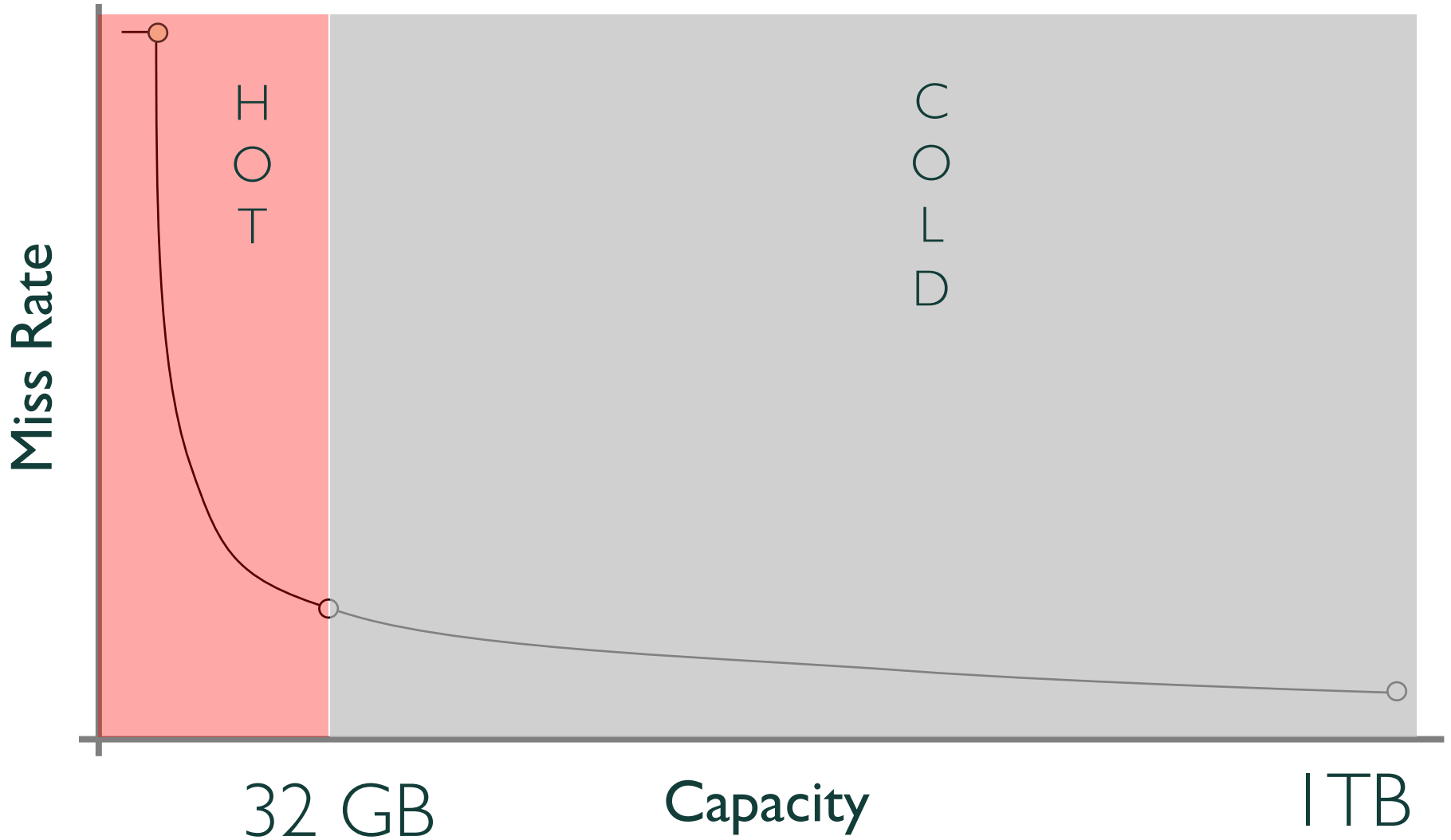
LUX ET VERITAS Yale University

Intel Transformative Server Architecture Center

intel

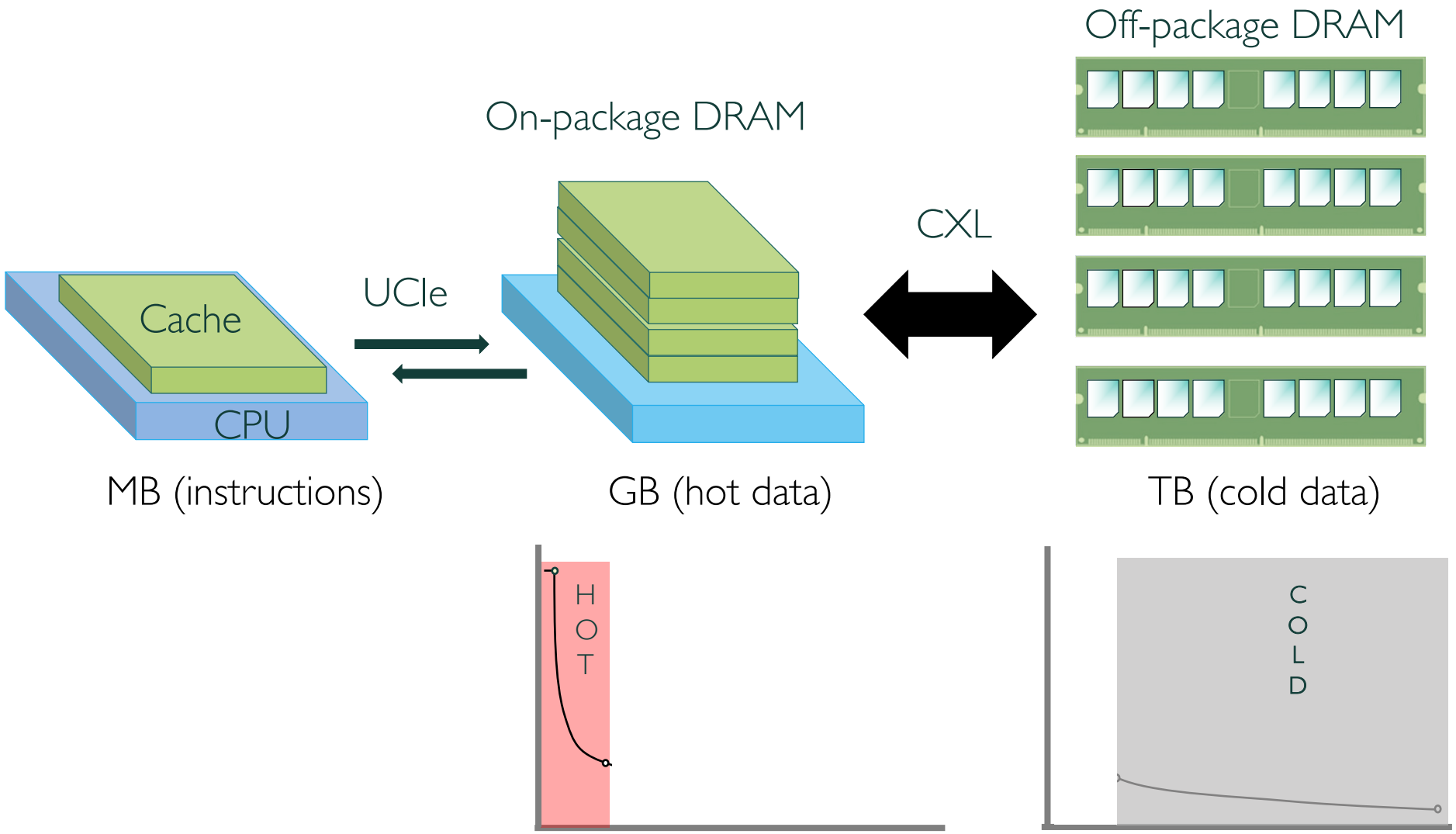
CLOUD-NATIVE MEMORY HIERARCHIES

PARALLEL SYSTEMS ARCHITECTURE LAB

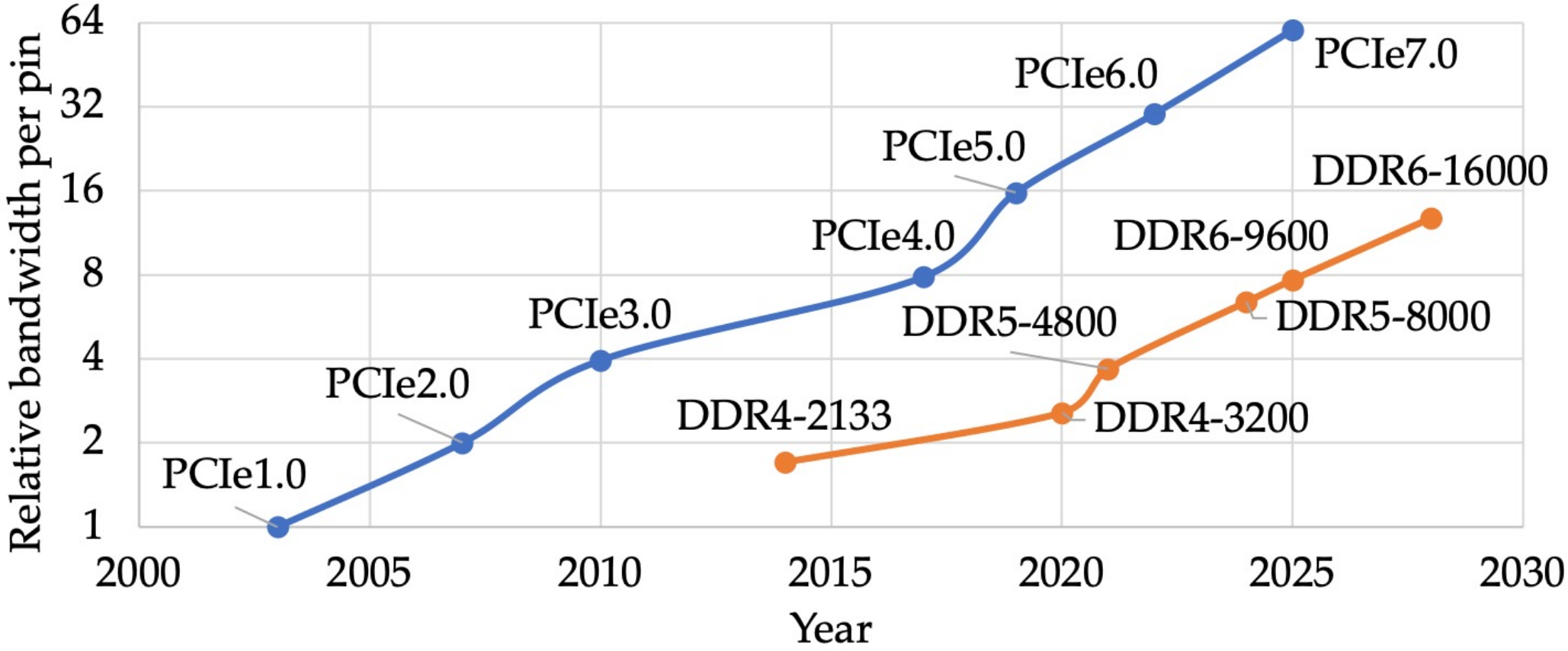


EMERGING MEMORY HIERARCHIES

PARALLEL SYSTEMS ARCHITECTURE LAB



PCIe BANDWIDTH PROJECTIONS



[A Case for CXL-Centric Servers, Cho, et. al.]

- CXL will eventually subsume DDR

CXL OPPORTUNITIES

PARALLEL SYSTEMS ARCHITECTURE LAB

- Memory is currently wasted (50% in containers)
- Much memory is stranded (Pond, ASPLOS'23)
- Pool memory in proximity (e.g., Scale-Out NUMA)
  - Both requests and data partitions are skewed
  - Helps with load balancing
- Memory accounts for much of fabrication emissions
  - Use CXL to keep (old not new) DDR memory

ON-PACKAGE OPPORTUNITIES: ACCELERATORS

PARALLEL SYSTEMS ARCHITECTURE LAB

Data management

- Data copy/move
- Compression/decompression
- Serialization/deserialization
- Encrypt/decrypt
- Scatter/gather

Analytics

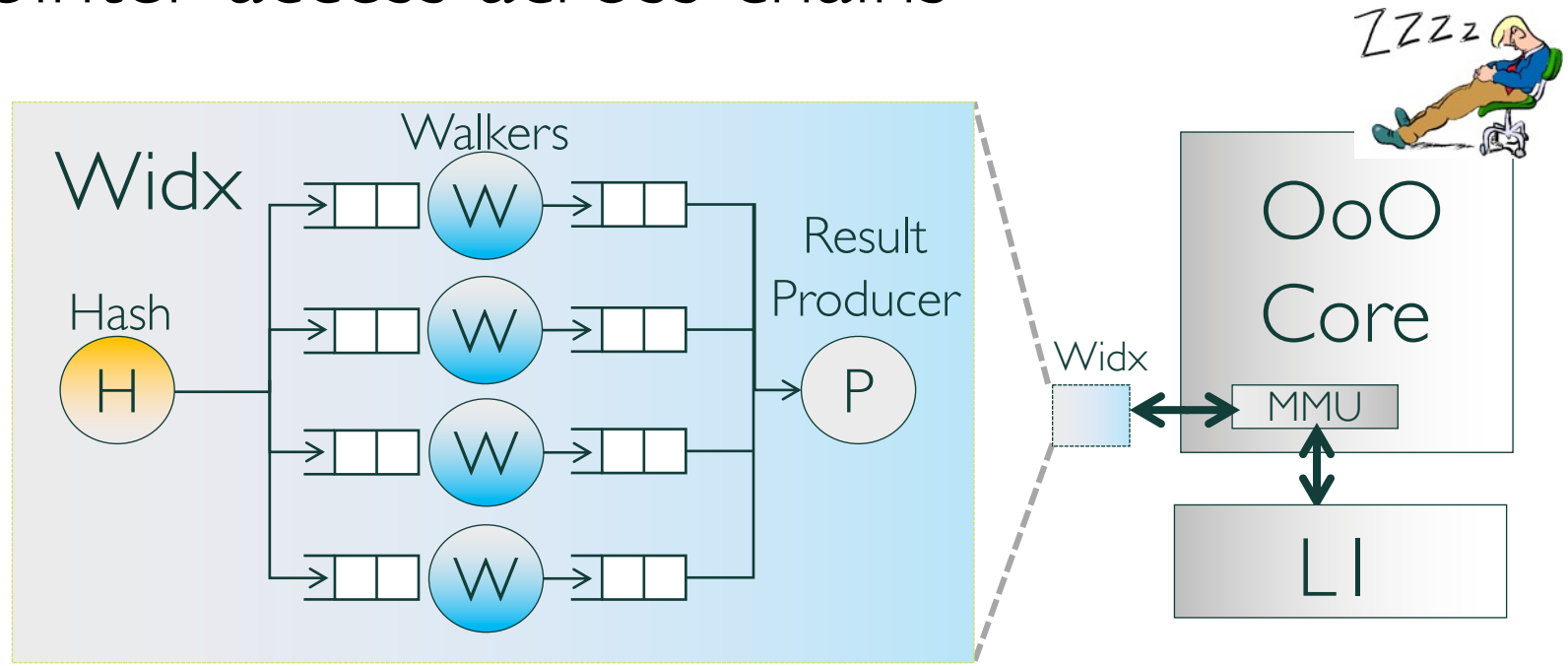
- Spark operations (e.g., scan, filter, groupby)
- SQL operations (e.g., intersect, union, join)
- Matrix operations

Memory views: rows for data management , columns for analytics

CHASING POINTERS W/ WALKERS

PARALLEL SYSTEMS ARCHITECTURE LAB

- Traverse data structures (e.g., hash table, B-tree)
- Parallelize pointer chains
- Overlap pointer access across chains



15x better performance/Watt over Xeon

WALKERS IN SOFTWARE [VLDB'16]

PARALLEL SYSTEMS ARCHITECTURE LAB

Use insights to help CPUs

- Decouple hash & walk(s) in software
- Schedule off-chip pointer access with co-routines

2.3x speedup on Xeon

- Unclogs dependences in microarchitecture
- Maximizes memory level parallelism
- DSL w/ co-routines
- Integrated in SAP HANA [VLDB'18]

NEAR-MEMORY PROCESSING [ISCA'17]

PARALLEL SYSTEMS ARCHITECTURE LAB

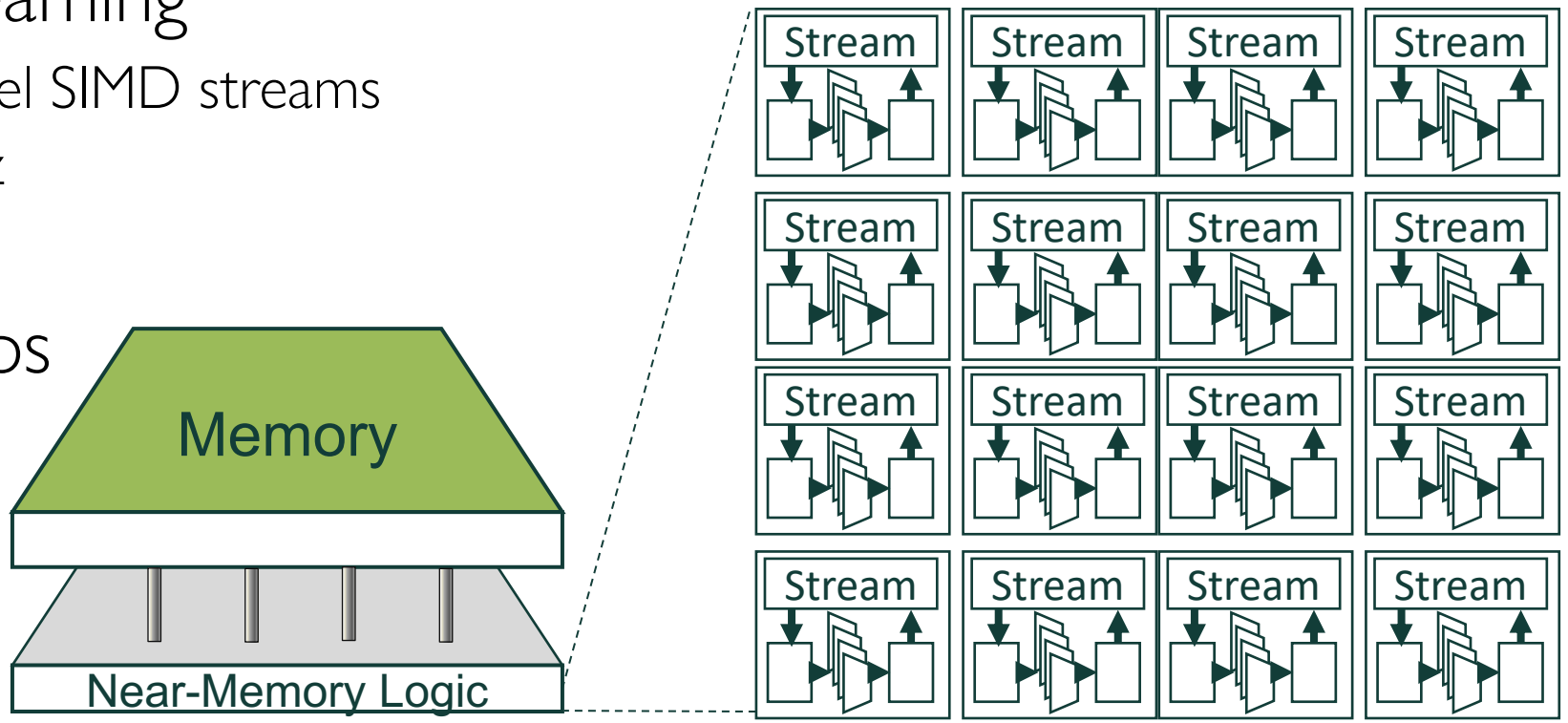
Stream data out of row buffers

SIMD cores + data streaming

- Saturates b/w with parallel SIMD streams
- 1024-bit SIMD @ 1 GHz
- No caches

Runs Spark Analytic Ops

50x over Xeon



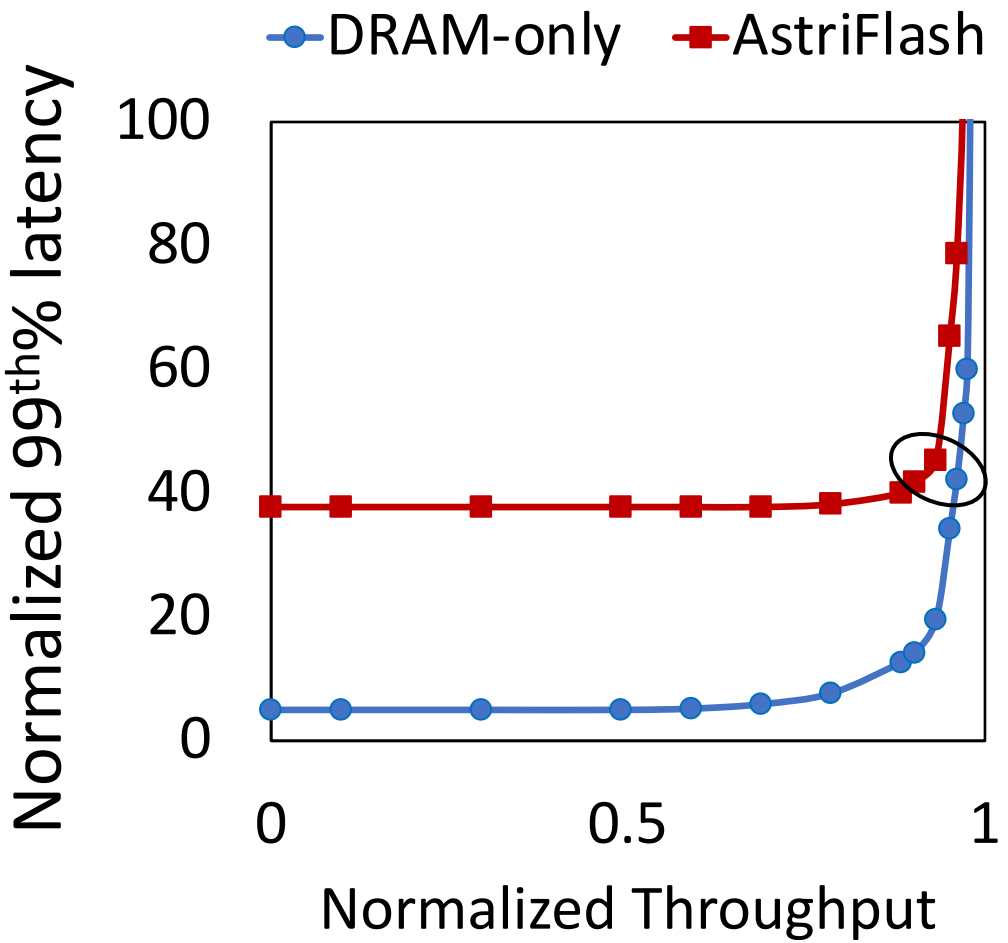
Algorithm/hardware co-design maximize near-memory performance

ONLINE SERVICES IN FLASH [HPCA'23]

PARALLEL SYSTEMS ARCHITECTURE LAB

	Cost	Latency
DRAM	1x	~100 ns
SCM	1/5x	1-10 µs
SSD	1/30x-1/50x	> 50 µs (OS)

- Host & serve data from SSD
- Map SSD as memory, DRAM as cache
- Co-design CPU & OS for
  - paging
  - µs-level thread switch



Maintains tail latency with only 5% lower throughput

NETWORKS

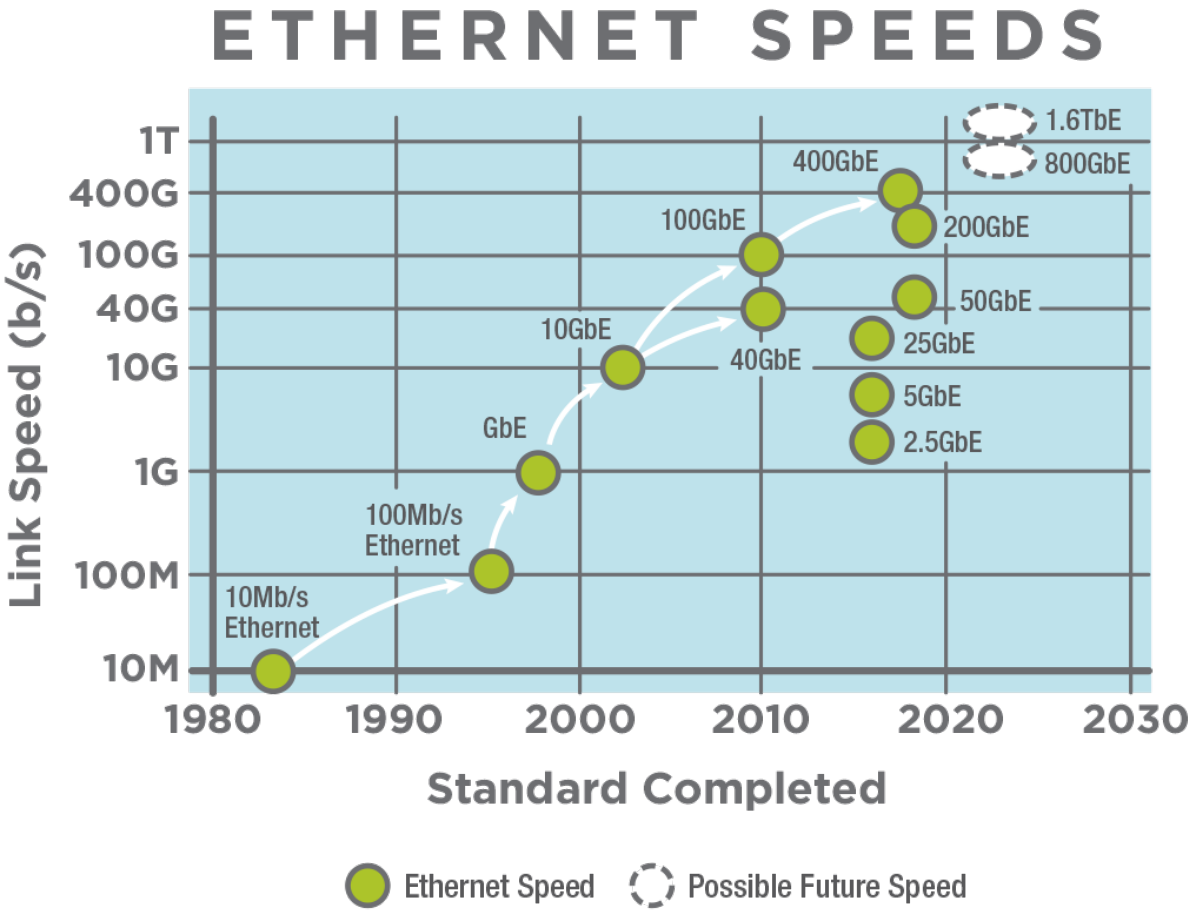
PARALLEL SYSTEMS ARCHITECTURE LAB

Network stack bottleneck:

- B/W growing faster than silicon
- Emerging µServices + serverless
- RPC, orchestration, ….

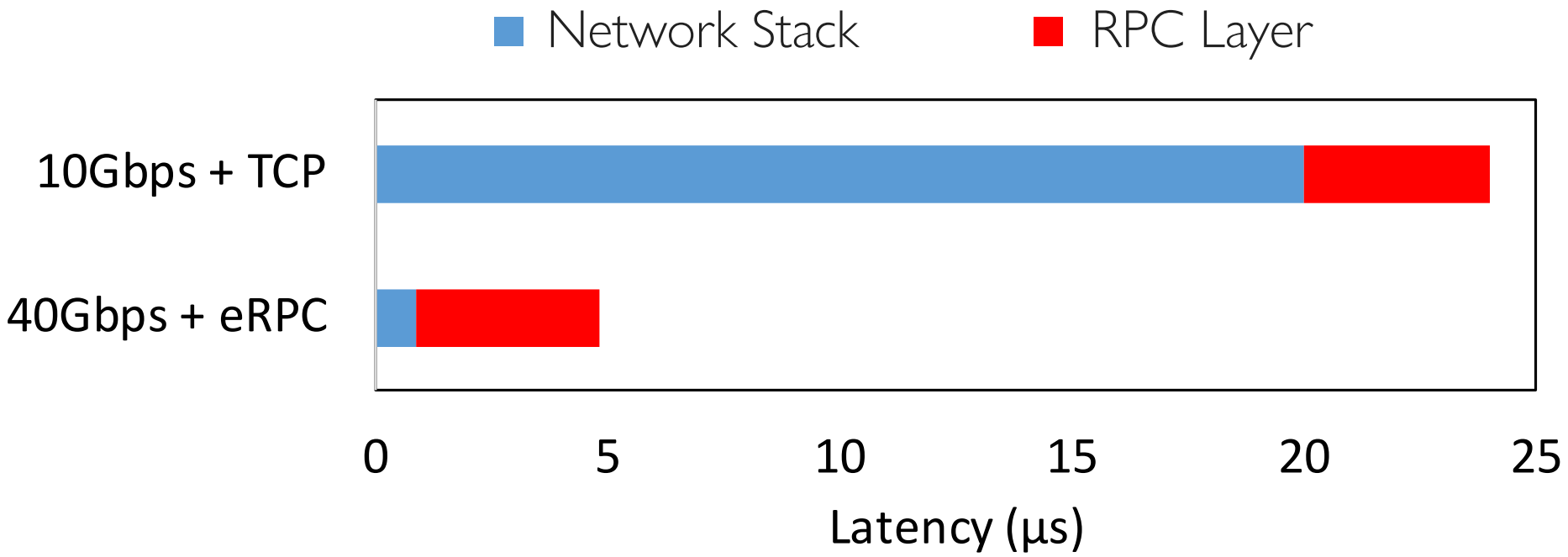
Key challenges:

- New abstractions
- Co-design of network stacks



RPC ACCELERATORS

PARALLEL SYSTEMS ARCHITECTURE LAB



- Wire time and protocol stacks have shrunk
- RPC dominates CPU cycles in µServices
- E.g., data transformation @ ~2.4Gbps w/ Thrift on Xeon

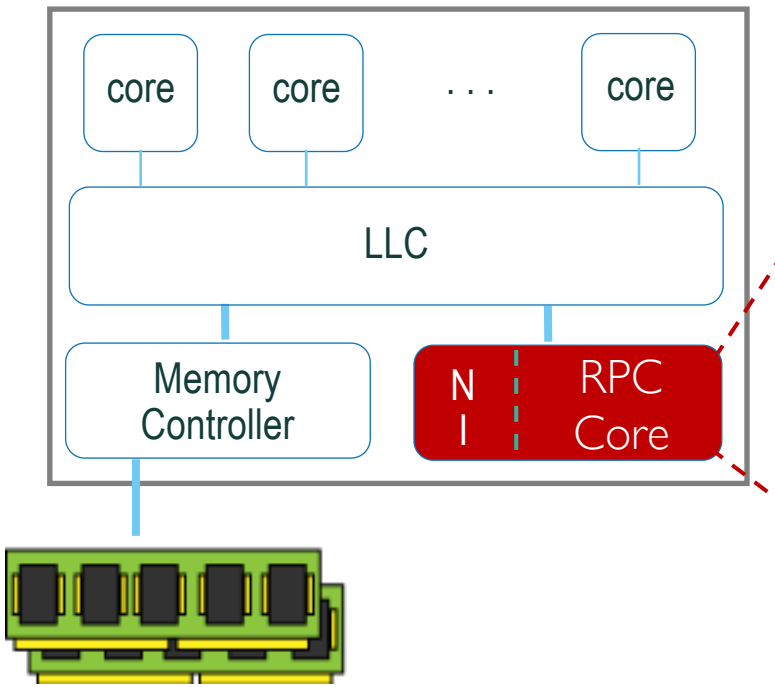
CEREBROS & NEBULA [ASPLOS'20,ISCA'20,MICRO'21]

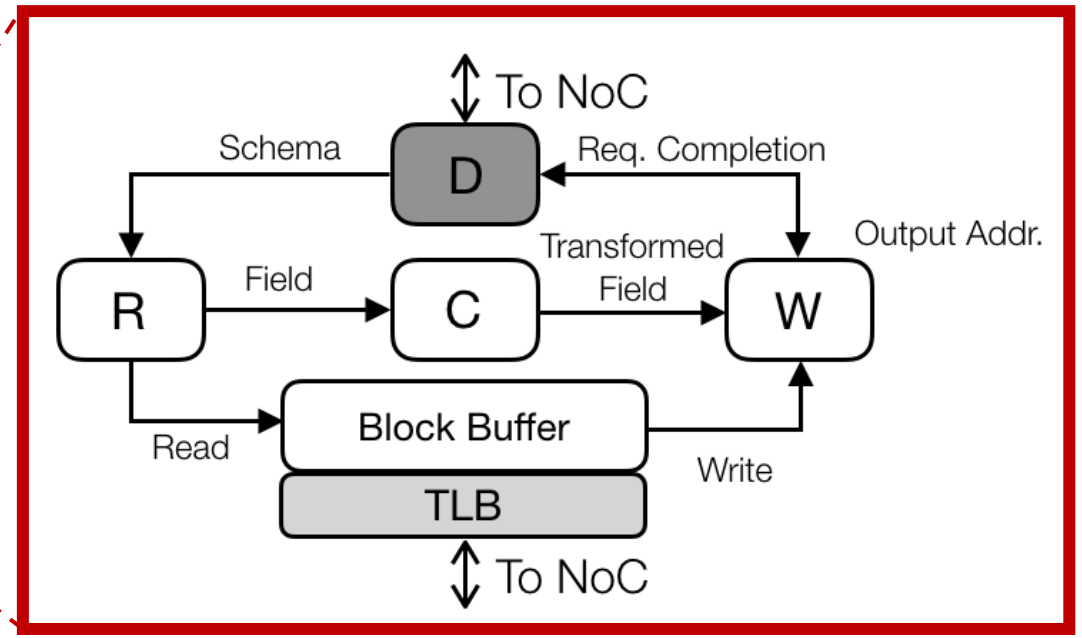
PARALLEL SYSTEMS ARCHITECTURE LAB

Meta

RPC processing at line rate:

- A "schema" (not instructions) interface to an RPC core
- Implements load balancing/affinity scheduling for µServices





OUTLINE

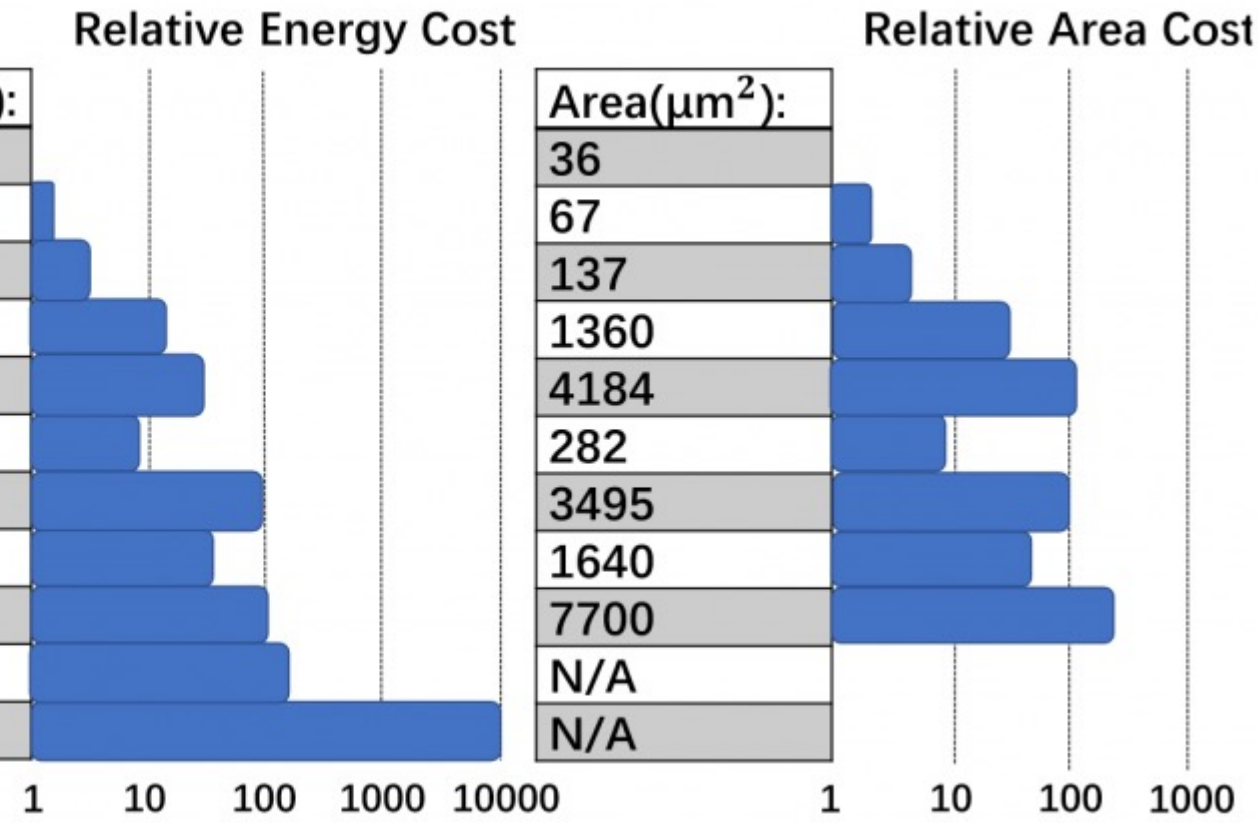
PARALLEL SYSTEMS ARCHITECTURE LAB

- ~~Overview~~
- Post-Moore servers
  - ~~Today's servers~~
  - ISA opportunities
    - ~~CPU/Memory/Storage/Network~~
    - AI
- Datacenter sustainability
- Summary

COST OF LOGIC VS. MEMORY

PARALLEL SYSTEMS ARCHITECTURE LAB

Operation:	Energy(pJ):	Area(μm²):
8b Add	0.03	36
16b Add	0.05	67
32b Add	0.1	137
16b FP Add	0.4	1360
32b FP Add	0.9	4184
8b Mult	0.2	282
32b Mult	3.1	3495
16b FP Mult	1.1	1640
32b FP Mult	3.7	7700
32b SRAM Read (8kb)	5.0	N/A
32b DRAM Read	640	N/A



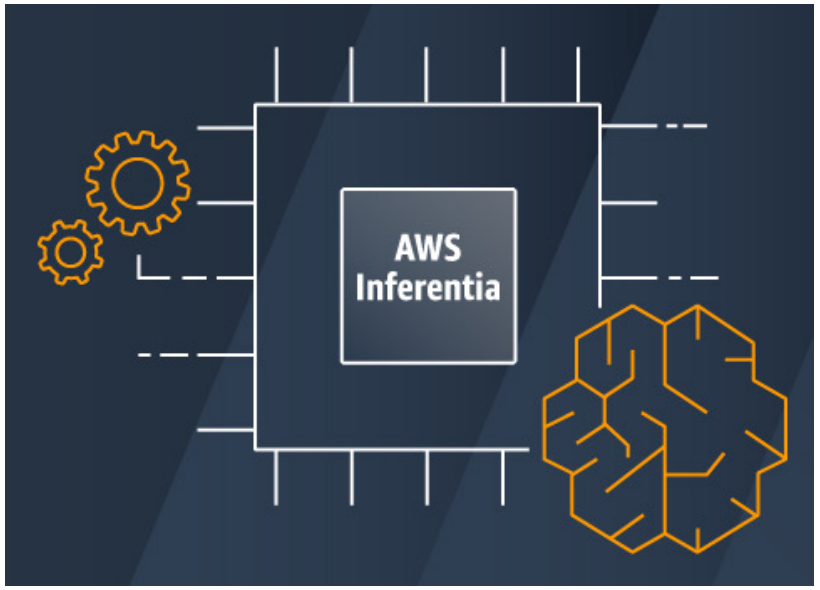
[src: Gholami, et. al.]

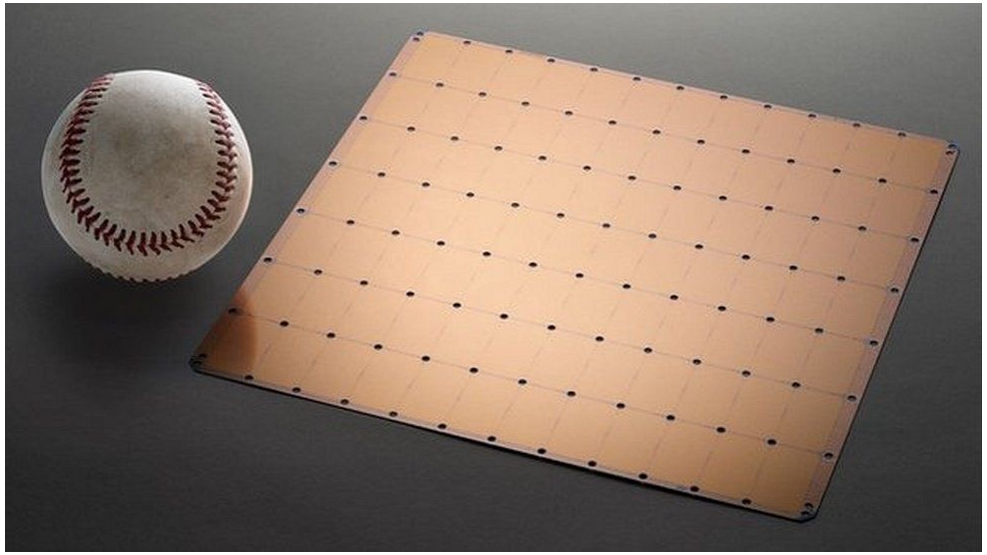
DNN PLATFORM DIVERGENCE

PARALLEL SYSTEMS ARCHITECTURE LAB

Inference platforms:

- Tight latency constraints
- Ubiquitous deployment
- Relies on fixed-point arithmetic





Training platforms:

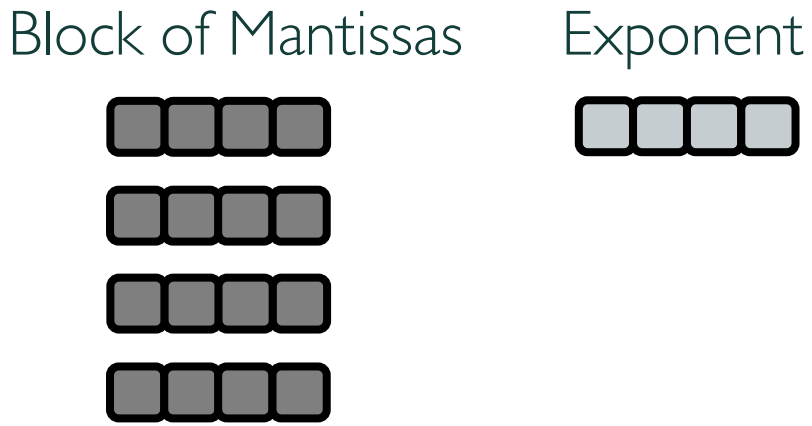
- Throughput optimized
- Server deployment
- Requires floating-point arithmetic

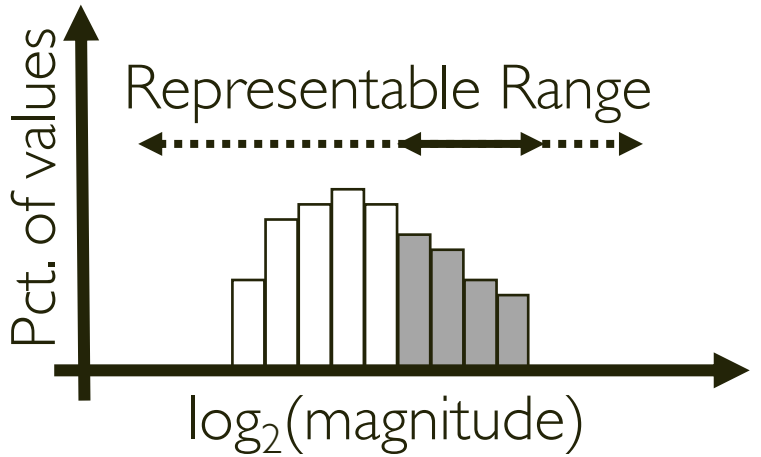
HYBRID BLOCK FLOATING POINT (HBFP)

PARALLEL SYSTEMS ARCHITECTURE LAB

1. Block floating point (BFP): one exponent/tensor
   - Low magnitude variation in tensor products
   - > 90% of all arithmetic operations

2. FP32 for all activations
   - High magnitude variation in gradient updates





Co-Located Training & Inference (CoITraIn)

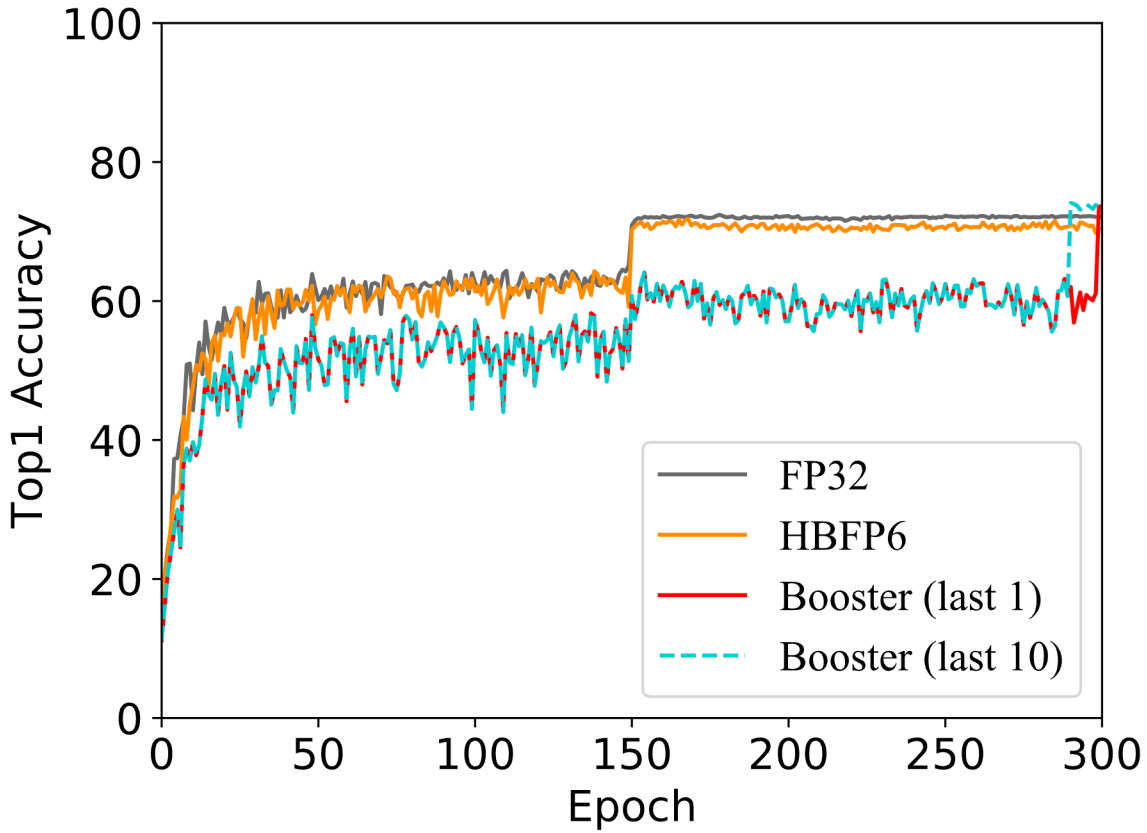
- ✓One accelerator for training and inference
- ✓Eliminates quantization
- ✓Enables online learning

parsa.epfl.ch/coltrain

MIXED-MANTISSA HBFP VS. FP32

PARALLEL SYSTEMS ARCHITECTURE LAB

DenseNet40 on CIFAR100



Transformers

Configuration	BLEU Score
FP32	34.77
HBFP6	34.47
HBFP4	32.64
Booster	**36.08**

FP32 level accuracy while using HBFP4 for majority of operations with 21.3x higher density

OUTLINE

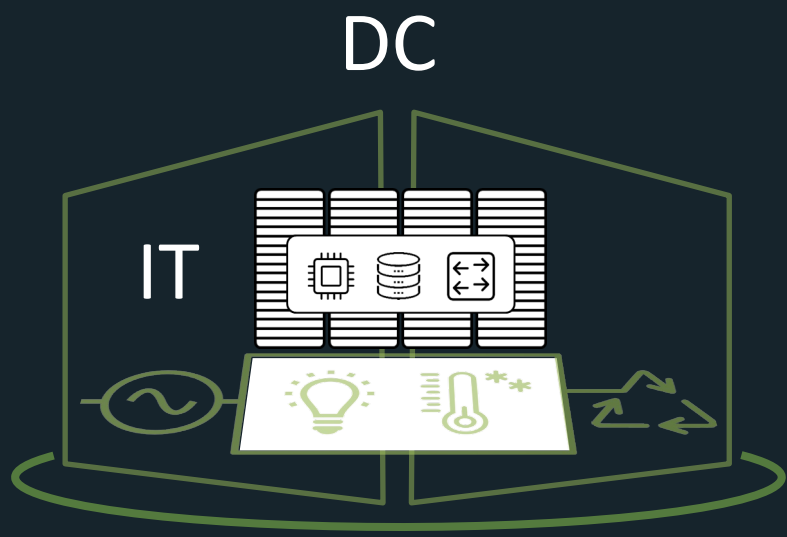
PARALLEL SYSTEMS ARCHITECTURE LAB

- ~~Overview~~
- ~~Post-Moore servers~~
- Datacenter sustainability
- Summary

Today's efficiency metric

power usage efficiency

$$PUE = \frac{\text{Total DC Power}}{\text{IT Power}}$$

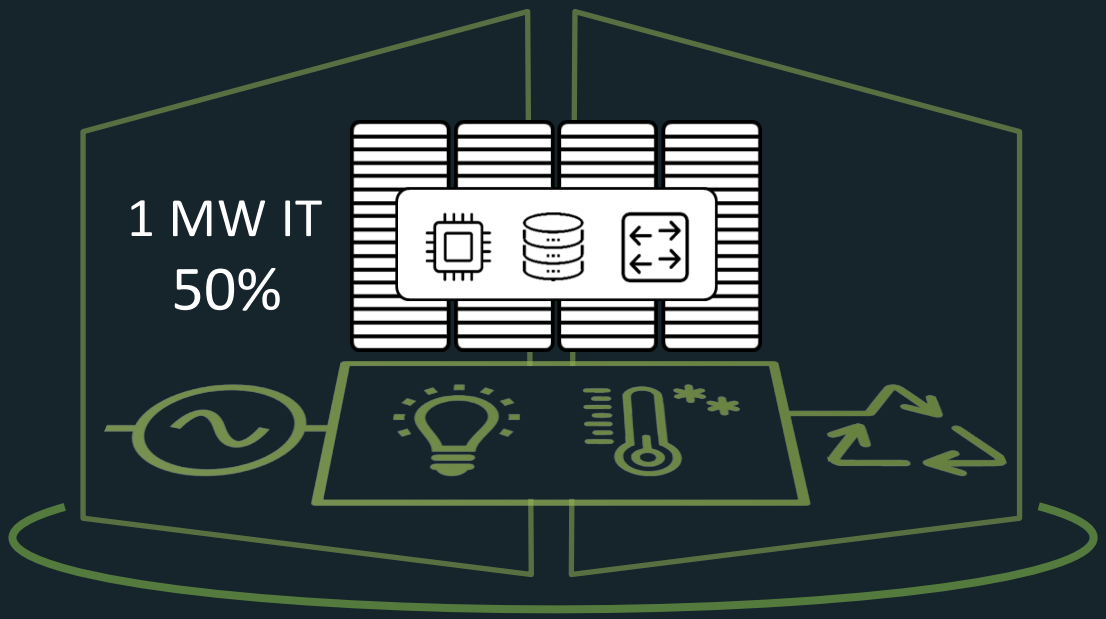


PUE has been around for two decades

What is wrong with PUE?

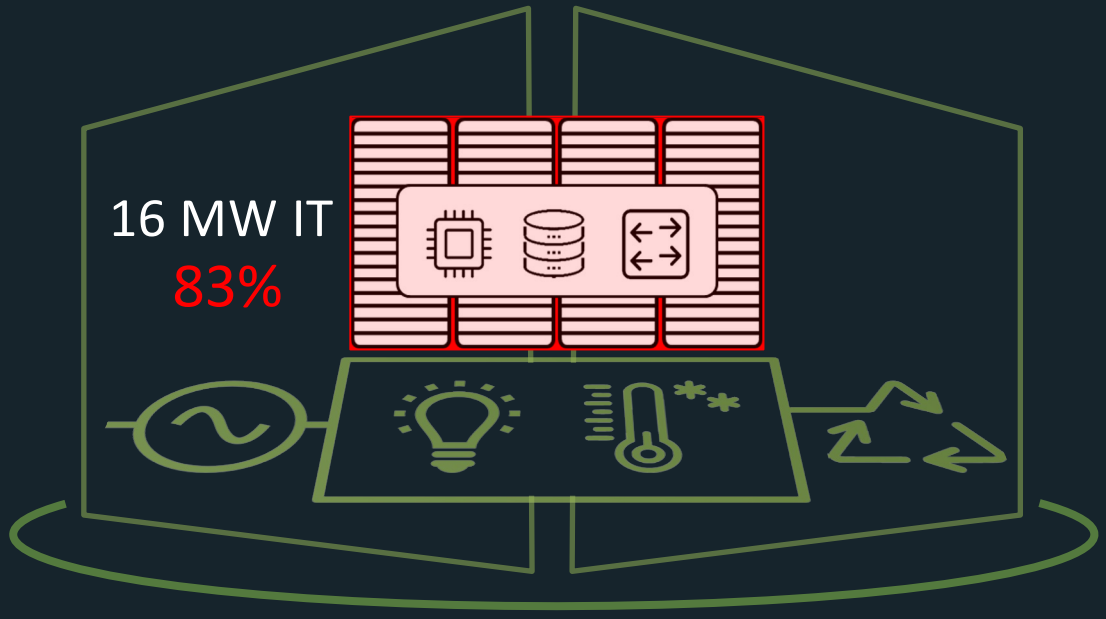
DC electricity goes to IT

2 MW DC in 2010



PUE = 2.0

20 MW DC in 2020



PUE = 1.2

GOODBYE PUE!
HELLO FULL-STACK EFFICIENCY

DC INFRASTRUCTURE EFFICIENCY

- electricity w/ renewables, cooling, heat recycling

IT INFRASTRUCTURE EFFICIENCY

+ compute, storage, network and workloads

DC CARBON FOOTPRINT

+ emissions from input electricity sources

DATA CENTER EFFICIENCY

Visit sdea.ch to find out about our label

EPFL EcoCloud Hewlett Packard Enterprise asut SDCA SWISS DATA CENTER ASSOCIATION energieschweiz Green IT Switzerland HOCHSCHULE LUZERN

DATACENTER BEST PRACTICES

PARALLEL SYSTEMS ARCHITECTURE LAB

Need:

- Metrics for datacenter output
  - E.g., matrix multiply in Python is 10x more work than in C
- Metrics for chip design
  - E.g., speedup not a great metric for accelerators
- Need life cycle from fabrication to recycling
  - Half of the emissions are from fabrication

SUMMARY

PARALLEL SYSTEMS ARCHITECTURE LAB

Post-Moore datacenters:

- Integration + Specialization + Approximation
- Revisit legacy abstractions, SW/HW interfaces
- Holistic algorithm/SW/HW co-design
- Division of control vs. data plane

Datacenter sustainability:

- Best practices
- Metrics

THANK YOU!

PARSA
PARALLEL SYSTEMS ARCHITECTURE LAB

For more information, please visit us at

parsa.epfl.ch

EPFL