

# Integration, Specialization and Approximation the “ISA” of Post-Moore Servers

Babak Falsafi



[parsa.epfl.ch](http://parsa.epfl.ch)

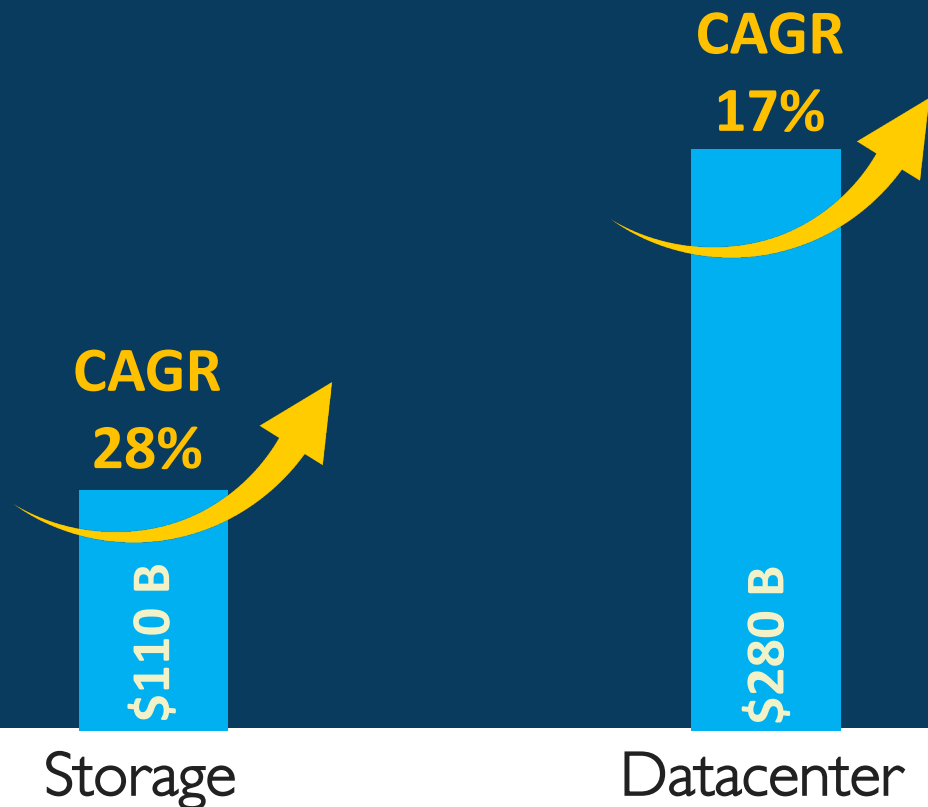
EcoCloud - Copyright 2022

**EPFL**

# DATACENTER GROWTH

## Market Growth 2018-2023

[Technavio, IDC]



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

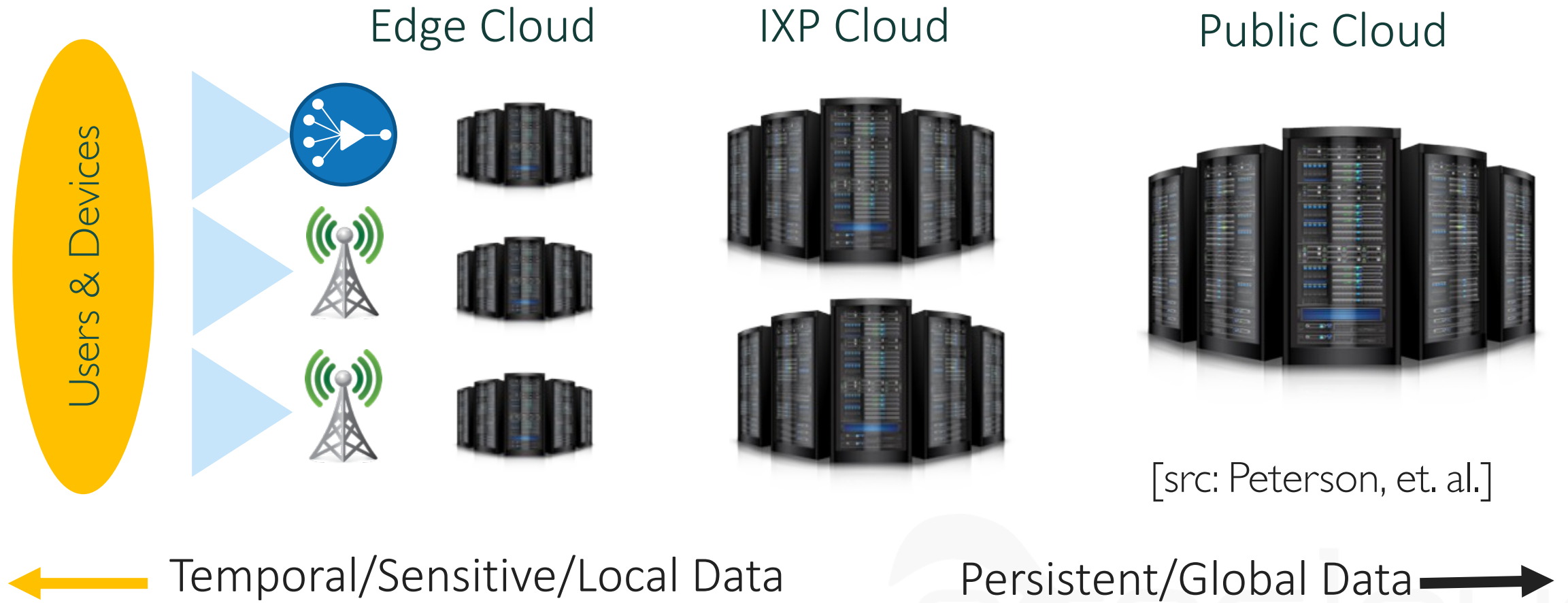
# DATACENTERS ARE BACKBONE OF CLOUD

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



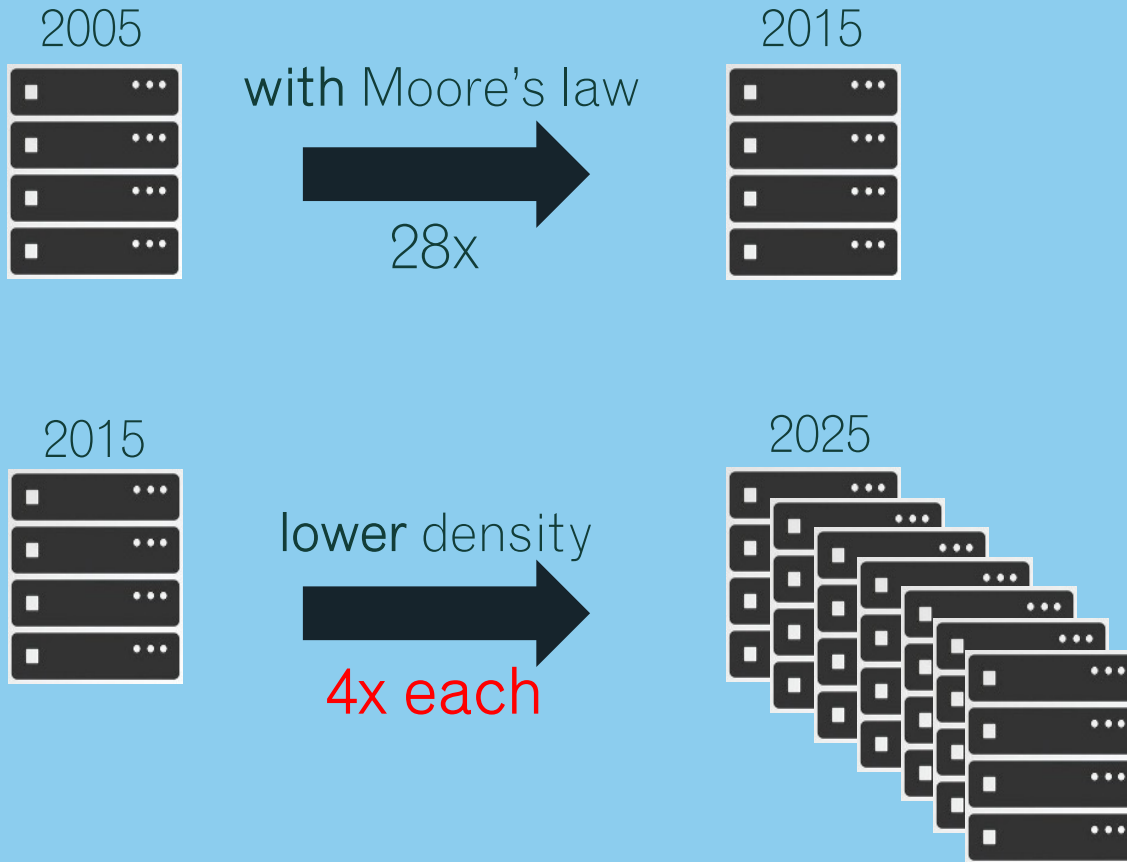
350MW, Bedford

# CLOUDS AT VARIOUS SCALES



# DATACENTERS NOT GETTING DENSER

Without Moore, building more



## 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

# Training a single AI model can emit as much carbon as five cars in their lifetimes

Deep learning has a terrible carbon footprint.

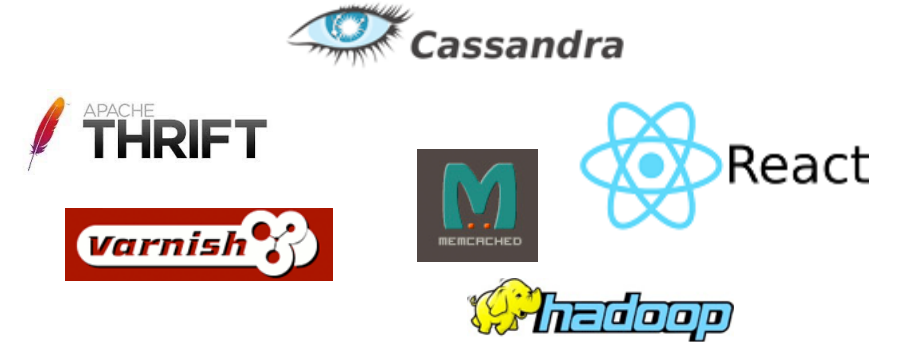
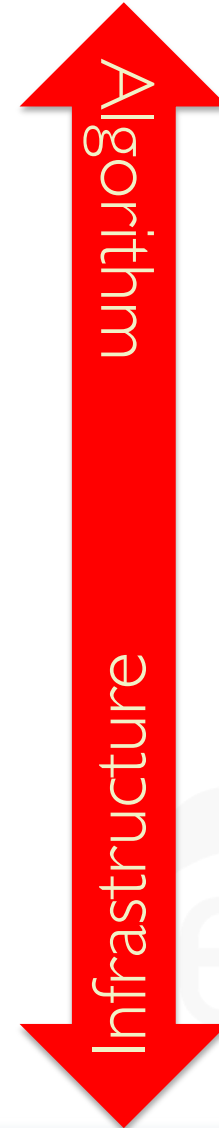
by Karen Hao

Jun 6, 2019

# POST-MOORE DATACENTERS

## Design for “ISA”

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

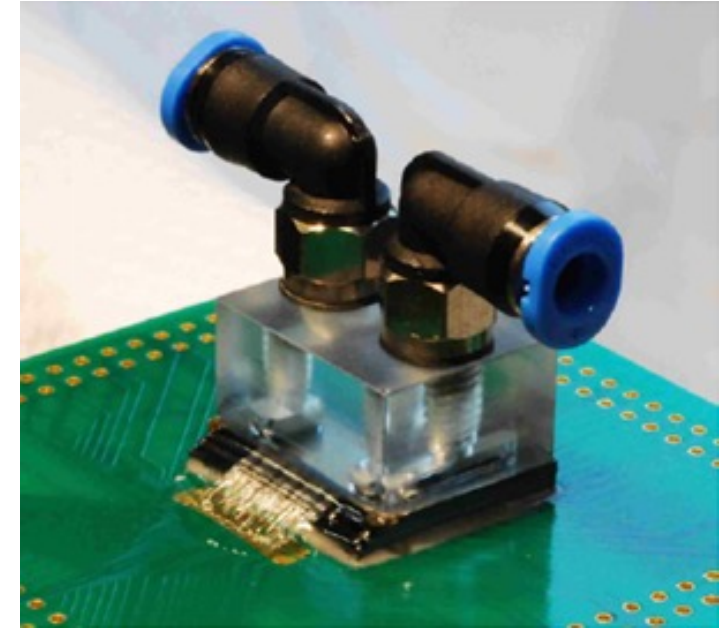
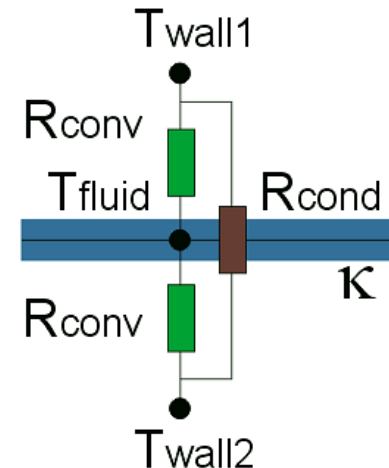
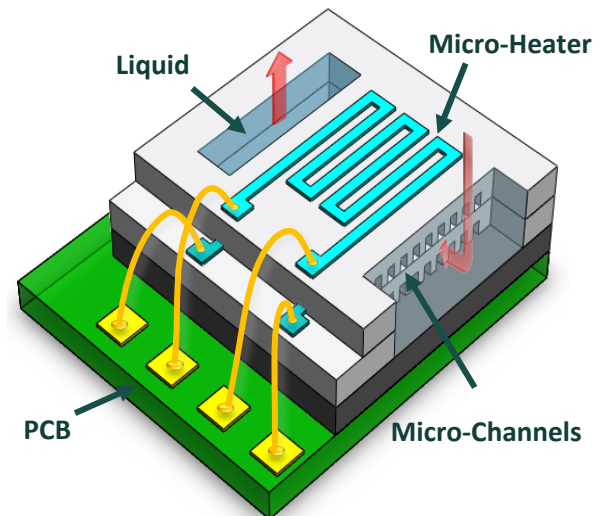


# INTEGRATED COOLING [Thome, Atienza]

3D server chip

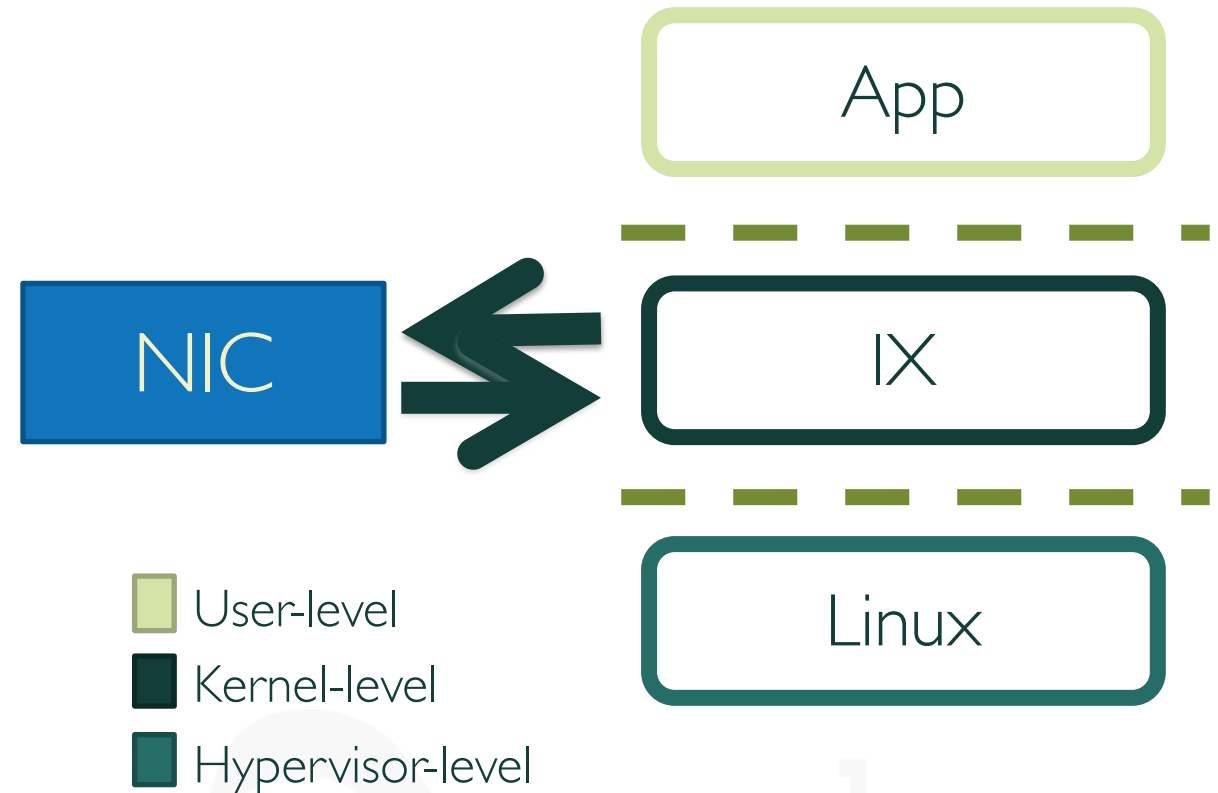
Two-phase liquid cooling

- Uniform higher thermal
- Higher heat removal
- Localized cooling



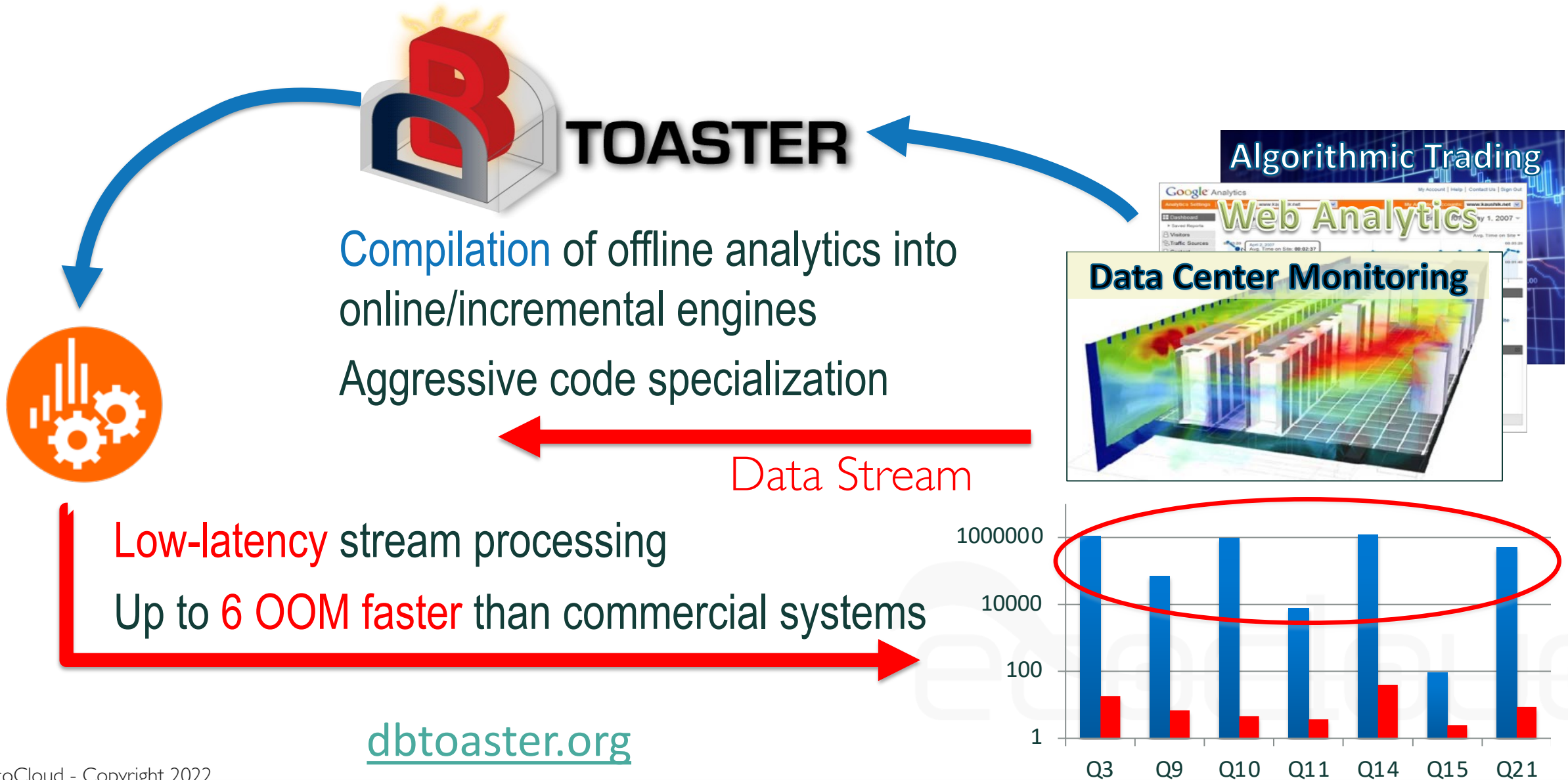
# SPECIALIZED NETWORKS [Bugnion]

- Data plane principles: zero-copy, run-to-completion, coherence free
- Protected operating system with clean-slate API
- Accelerates object sharing in datacenters
- IX Kernel → best paper at OSDI'14
- Follow-on work → SIGOPS'21 dissertation award



3.6x throughput with <50% latency @ 99<sup>th</sup> percentile

# SPECIALIZED DATABASES [Koch]



# QUANTIFYING EFFICIENCY/EMISSIONS BEYOND "PUE" (sdea.ch)

## DC INFRASTRUCTURE EFFICIENCY (PUE+)

- electrical, cooling and heat recycling components

## IT INFRASTRUCTURE EFFICIENCY

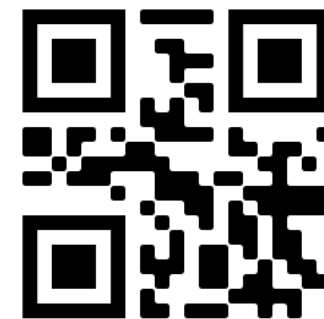
- + compute, storage, network and workloads

## DC CARBON FOOTPRINT

- + emissions from input electricity sources



EFFICIENCY



# OUTLINE

## ■ ~~Overview~~

## ■ Post-Moore Server Architecture

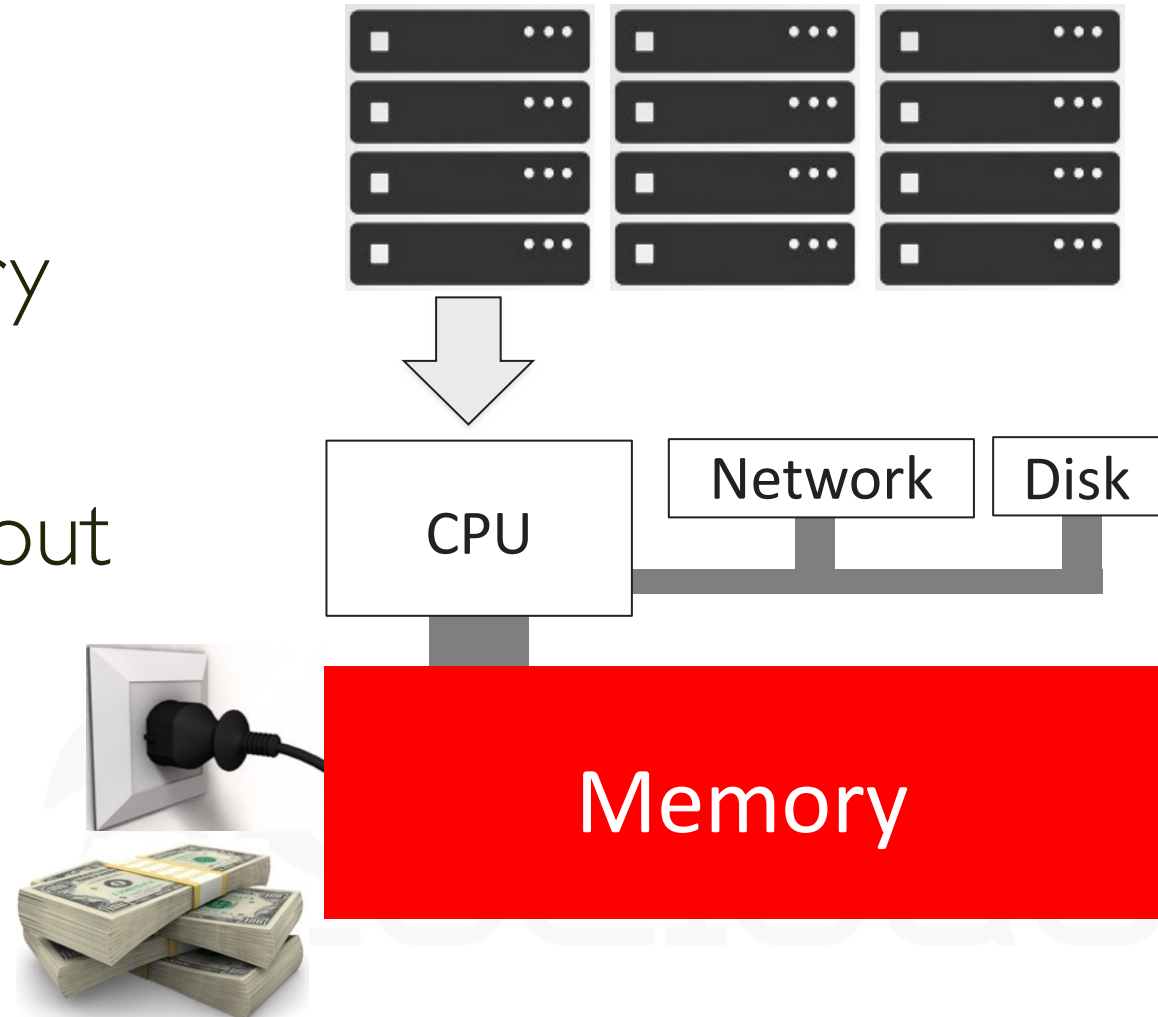
- 80's Desktops
- Specialized CPUs
- Integrated logic/memory
- Integrated networks
- Approximating AI

## ■ Summary



# SCALE-OUT DATACENTERS

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

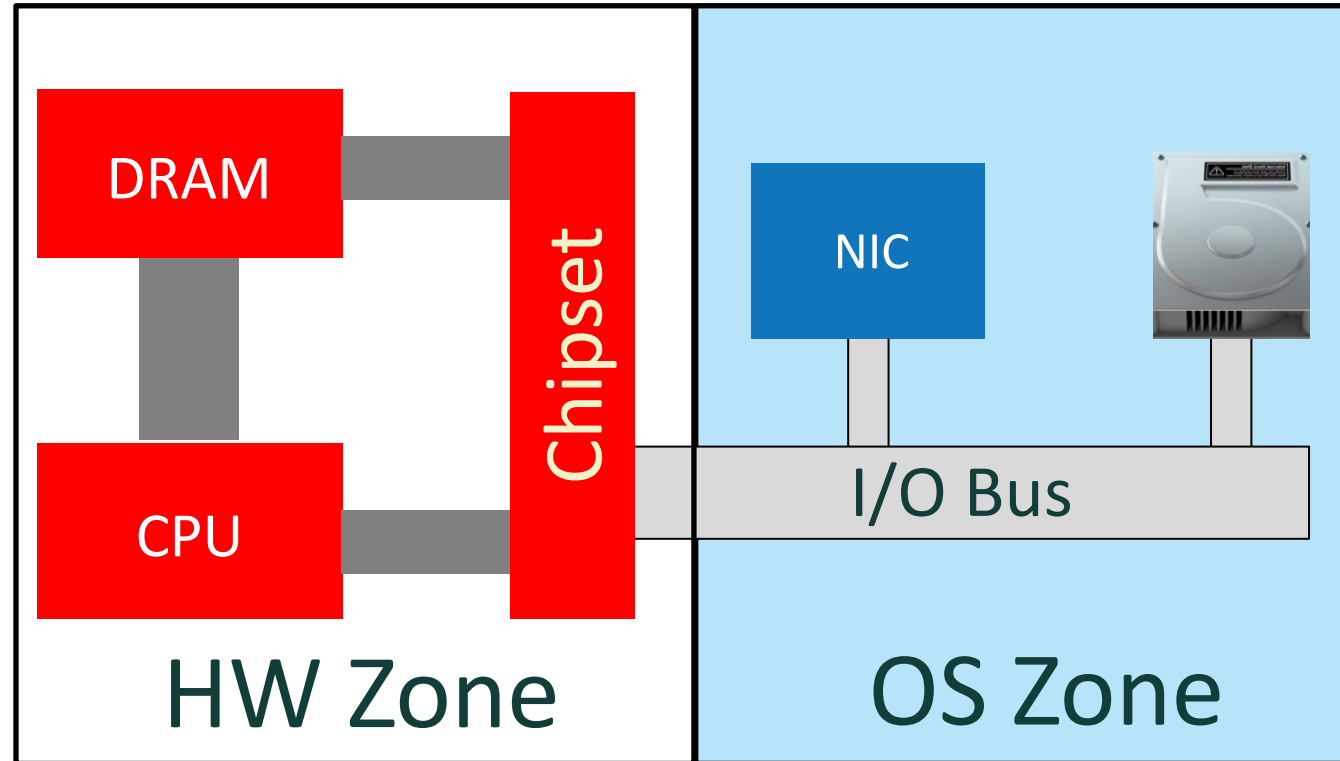
Today's platforms are PC's of the 80's

- 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:

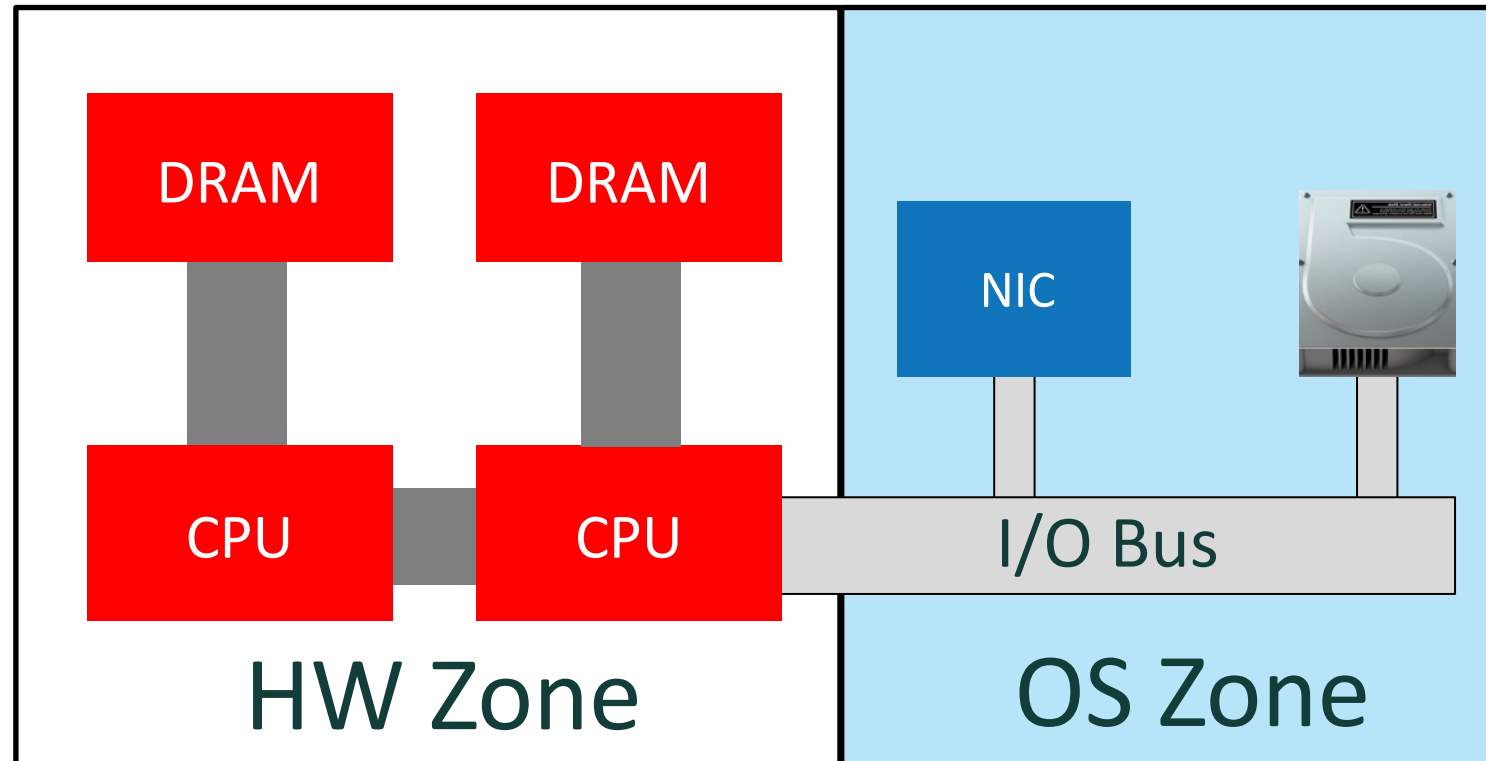
- Manycore network cards w/ own memory
- Flash controllers with embedded cores and memory
- Discrete accelerators with own memory

# 80'S DESKTOP



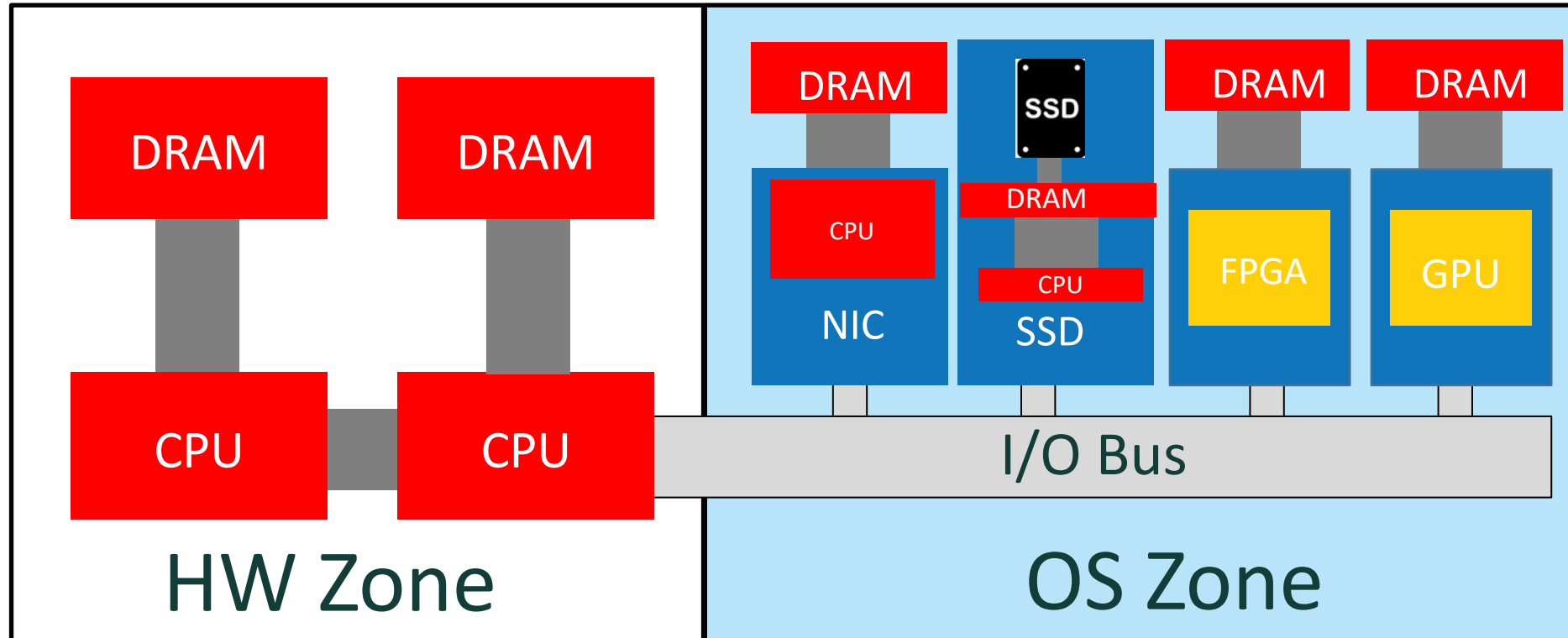
- 33 MHz 386 CPU, 250ns DRAM
- OS: Windows, Unix BSD (or various flavors)
- Focus: multiprogrammed in-memory compute

# TODAY'S SERVER: 80'S DESKTOP



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

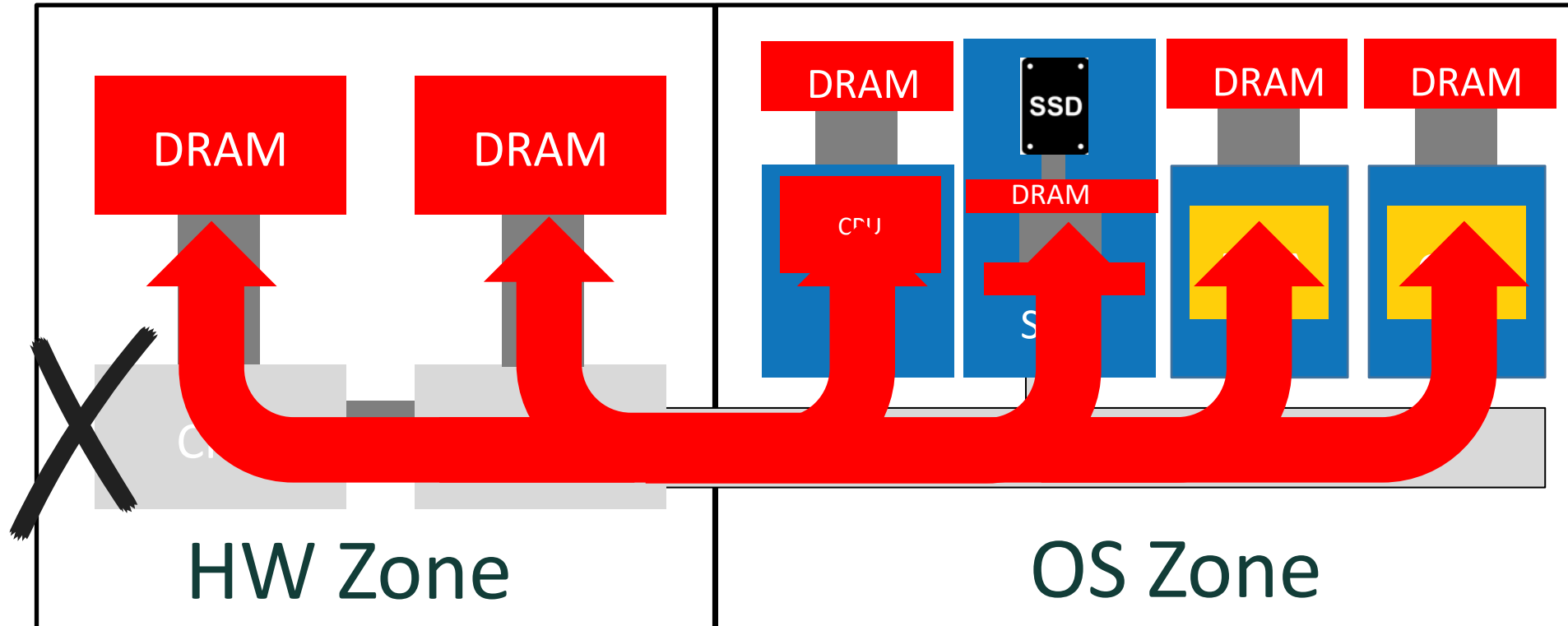
# TODAY'S SERVER: 80'S DESKTOP



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



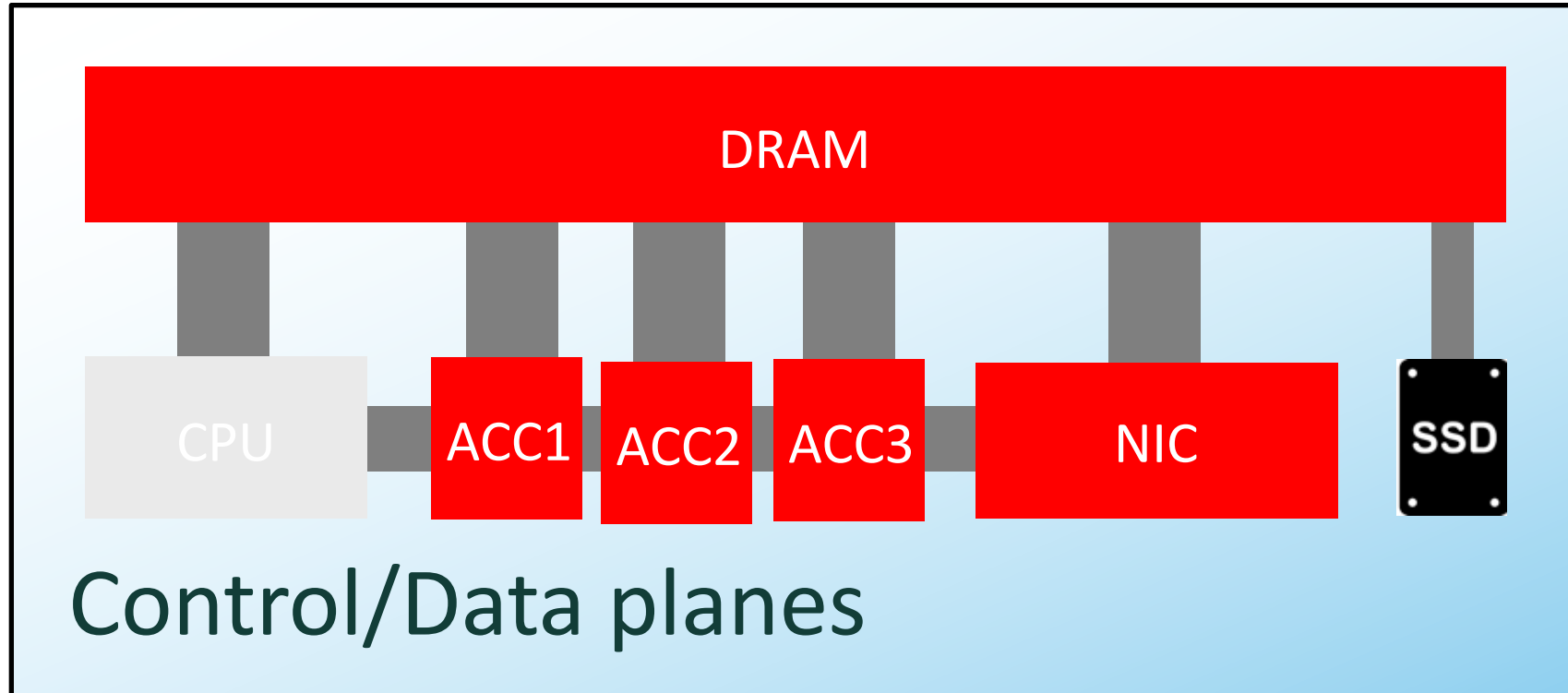
# TODAY'S SERVER: 80'S DESKTOP



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



# IDEAL POST-MOORE SERVER



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

# OUTLINE

- ~~Overview~~

- Post-Moore servers

  - ~~80's Desktops~~

  - Specialized CPUs

  - Integrated logic/memory

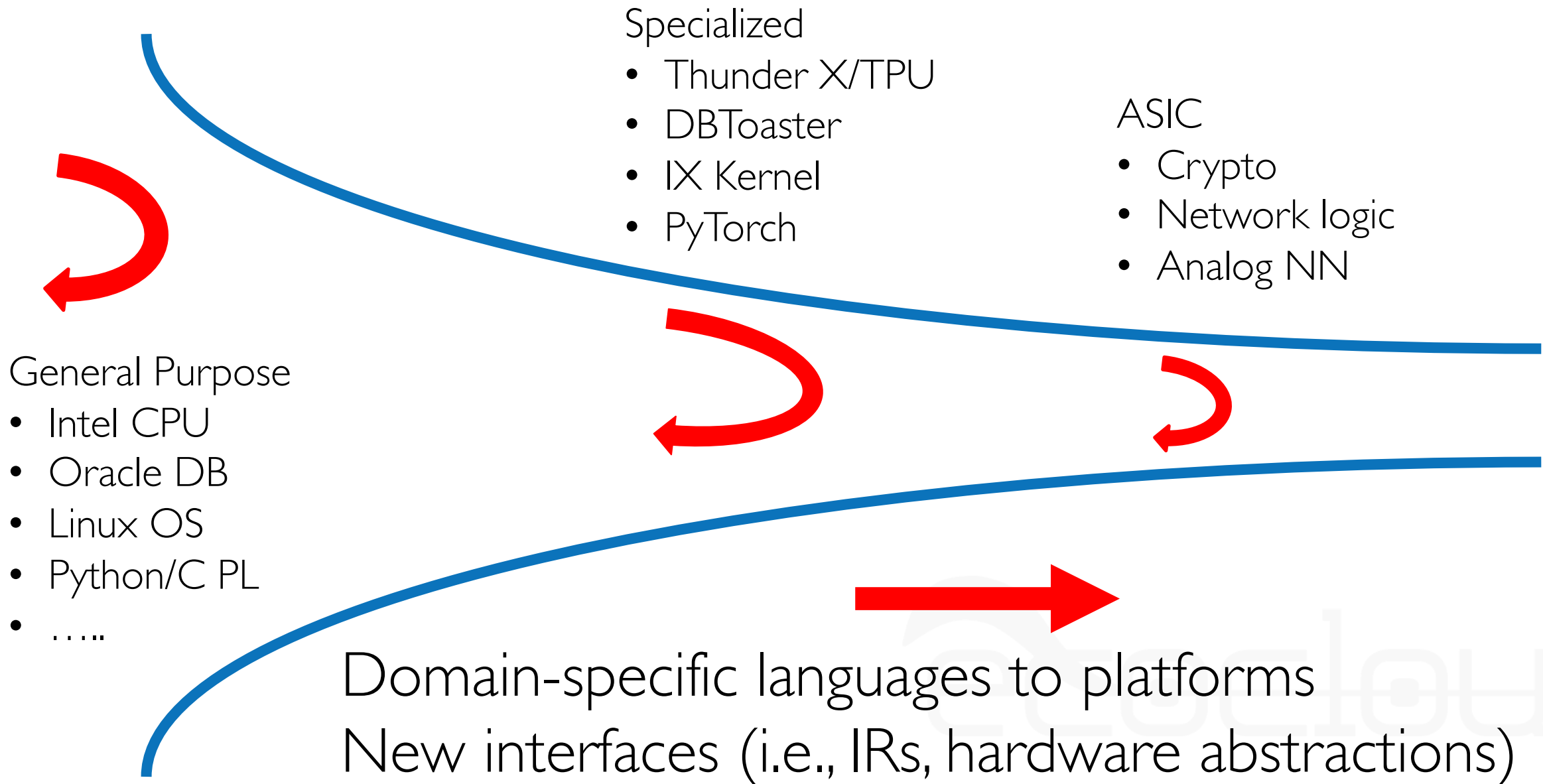
  - Integrated networks

  - Approximating AI

- Summary



# THE SPECIALIZATION FUNNEL

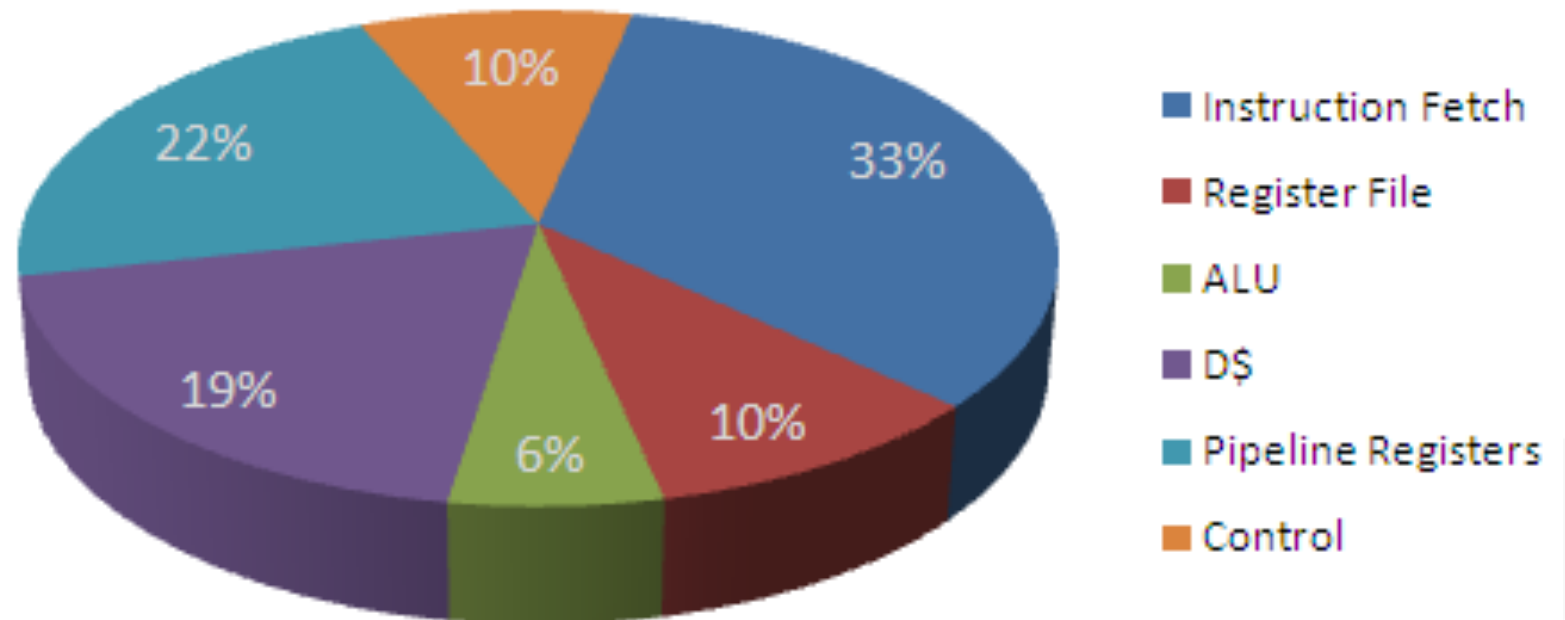


# THE LIMITS OF CPUS

CPU's 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]

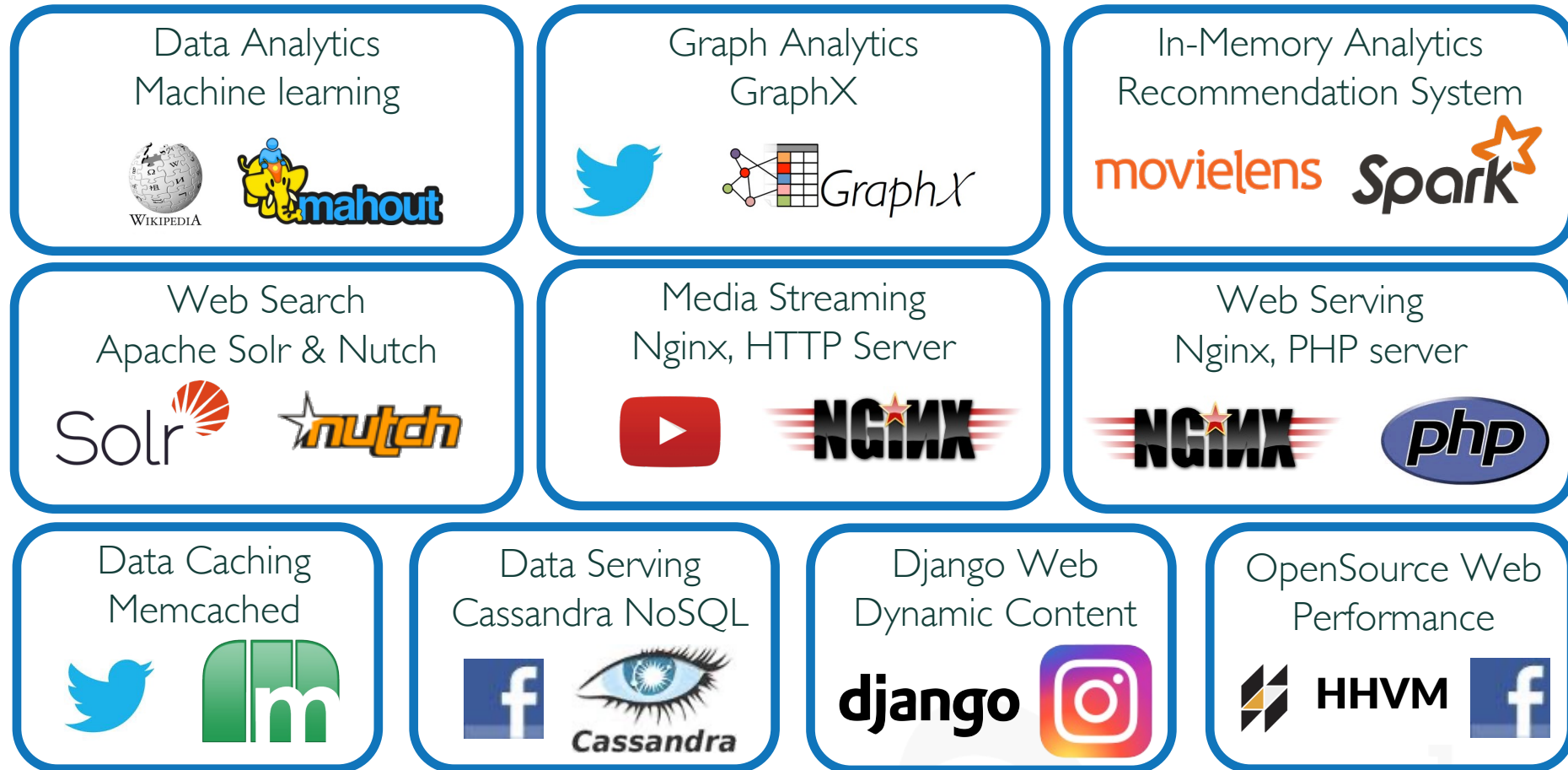
Three classes of workloads in datacenters

- First-party workloads (e.g., search, retail, media)
  1. Data management
  2. Analytics
    - Multi-tier to microservices
- Third-party workloads (cloud)
  3. Containerized
    - Emerging serverless





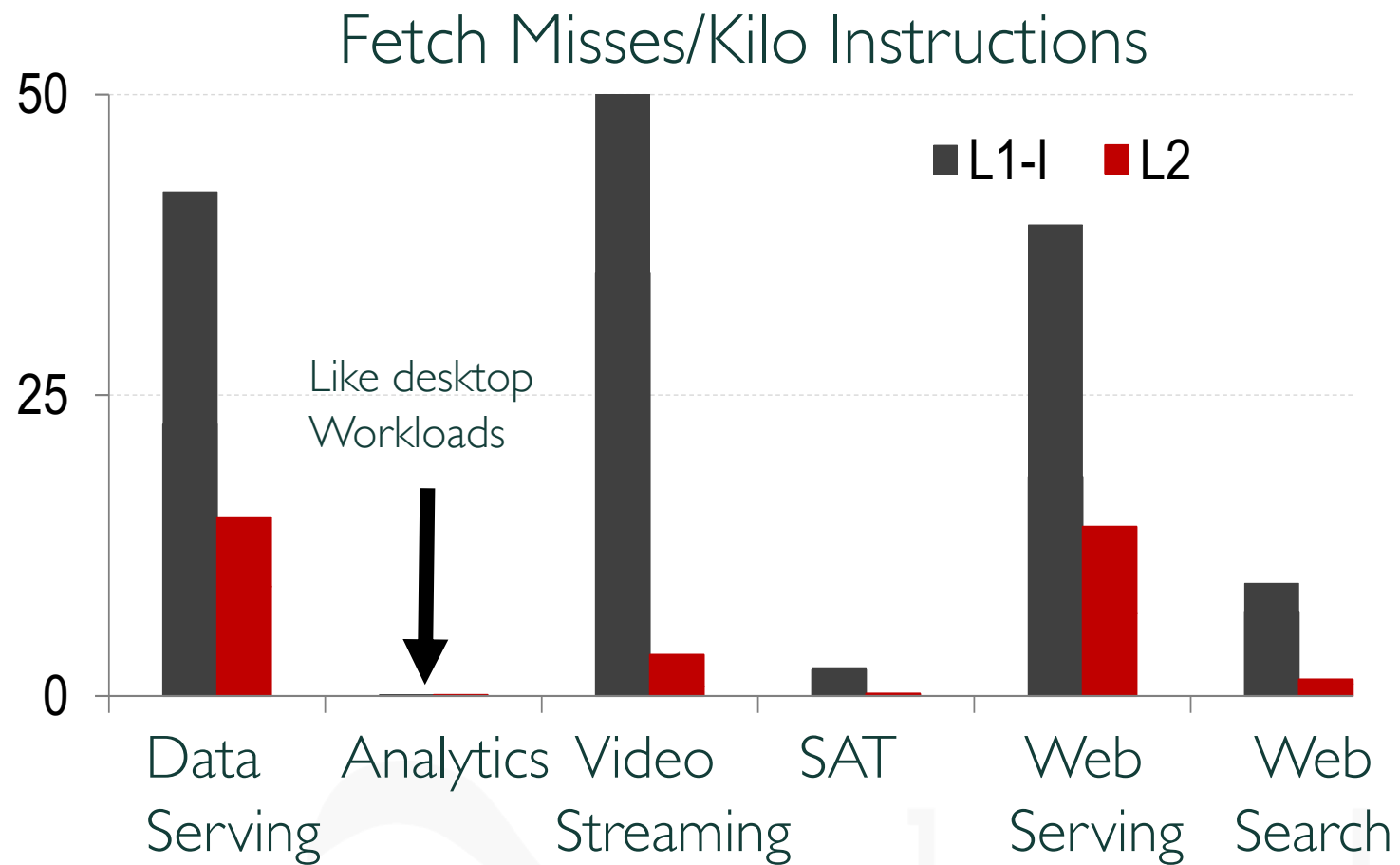
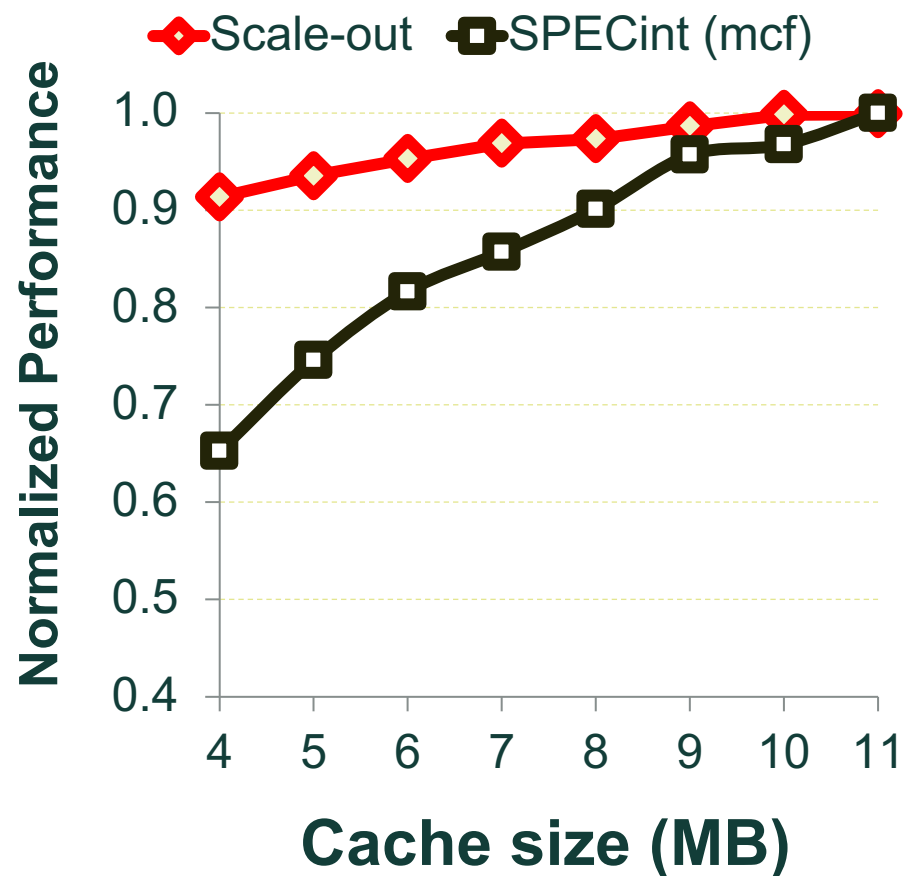
# CloudSuite (cloudsuite.ch, 4.0 coming)



Supports x86, ARM64, RISC-V



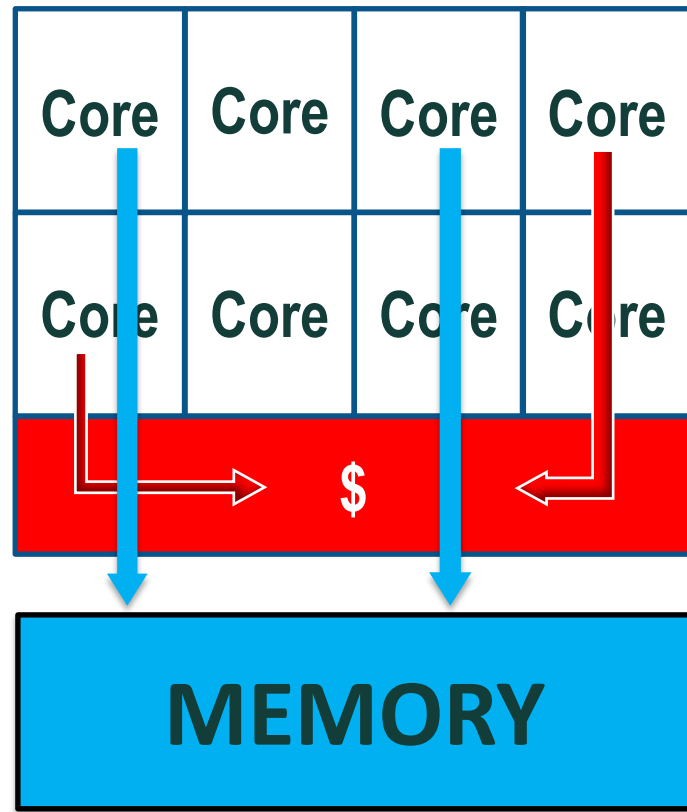
# SERVICES STUCK IN MEMORY [ASPLOS'12]



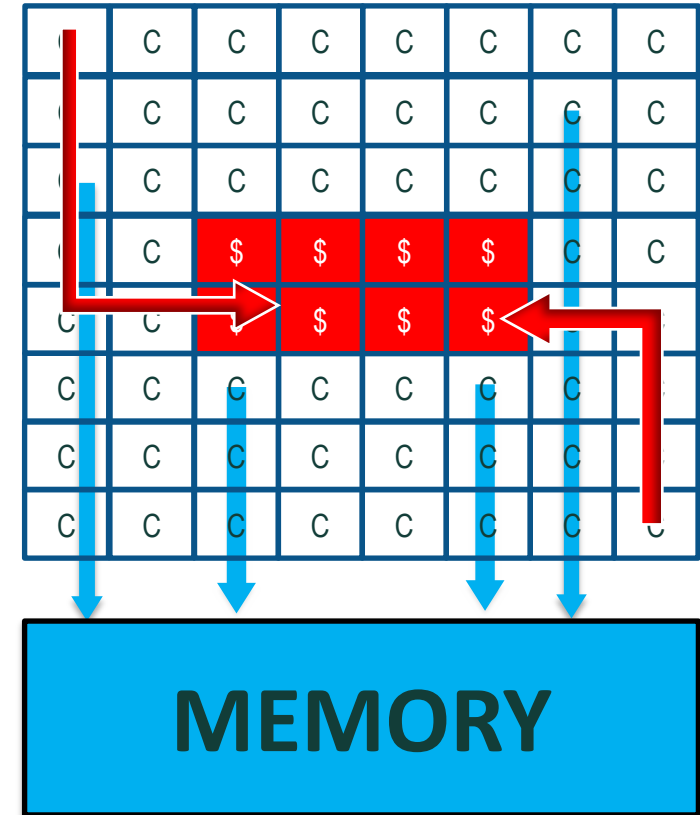
Cache overprovisioned

Instruction supply bottlenecked

# SCALE-OUT PROCESSOR (SOP)

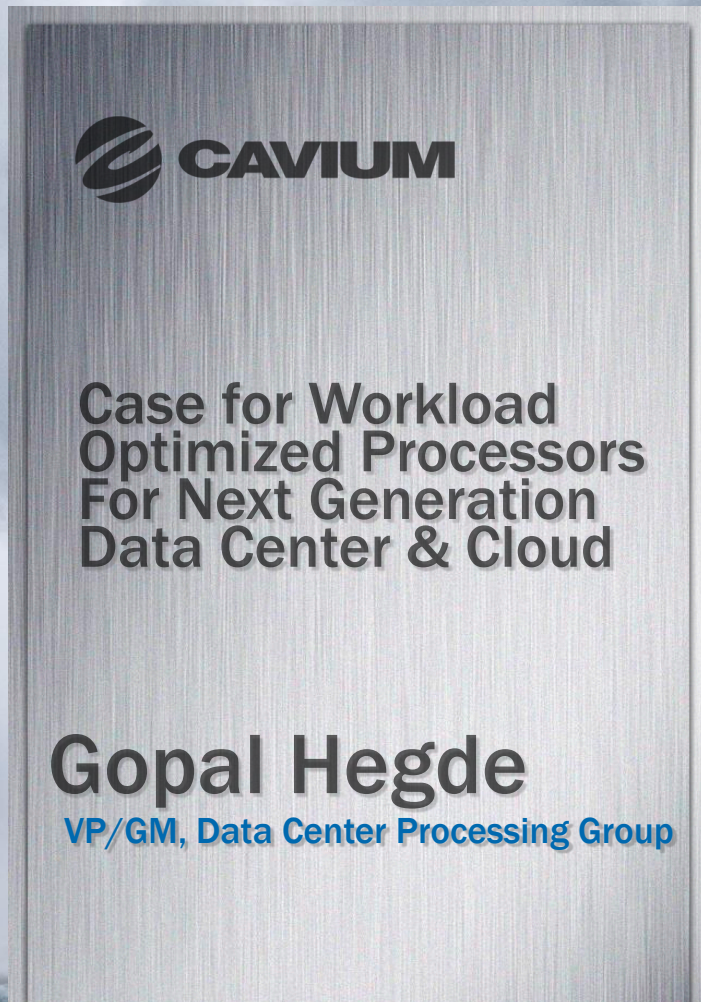


- General-purpose CPU
- ✗ Logic 60% of silicon
- ✗ 6x bigger cores



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

# CUSTOM SERVER CPU [c.a. 2014]



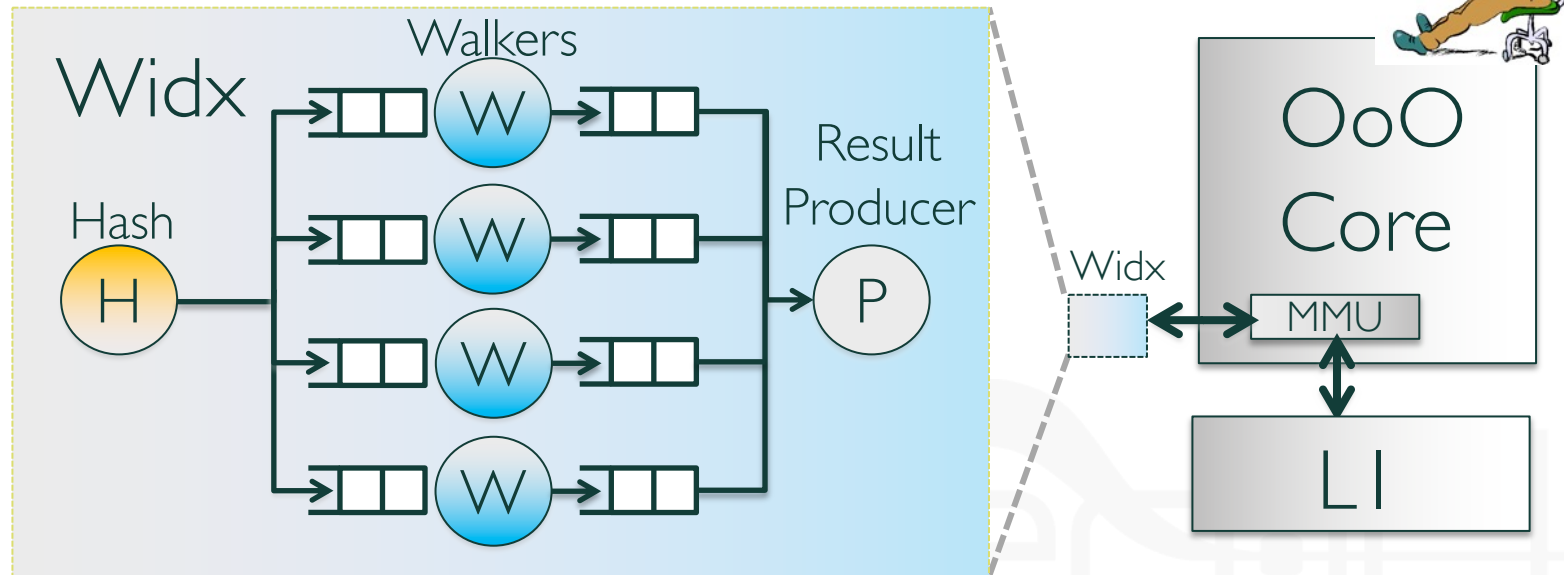
## Thunder X

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

# CHASING POINTERS W/ WALKERS

[MICRO'13]

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



**15x better performance/Watt over Xeon**

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]

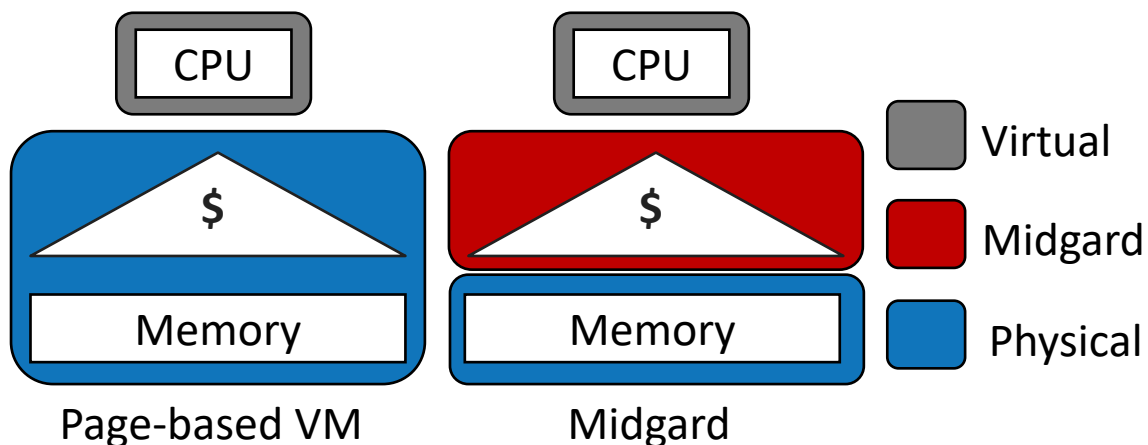


# POST-MOORE VIRTUAL MEMORY [ISCA'21]

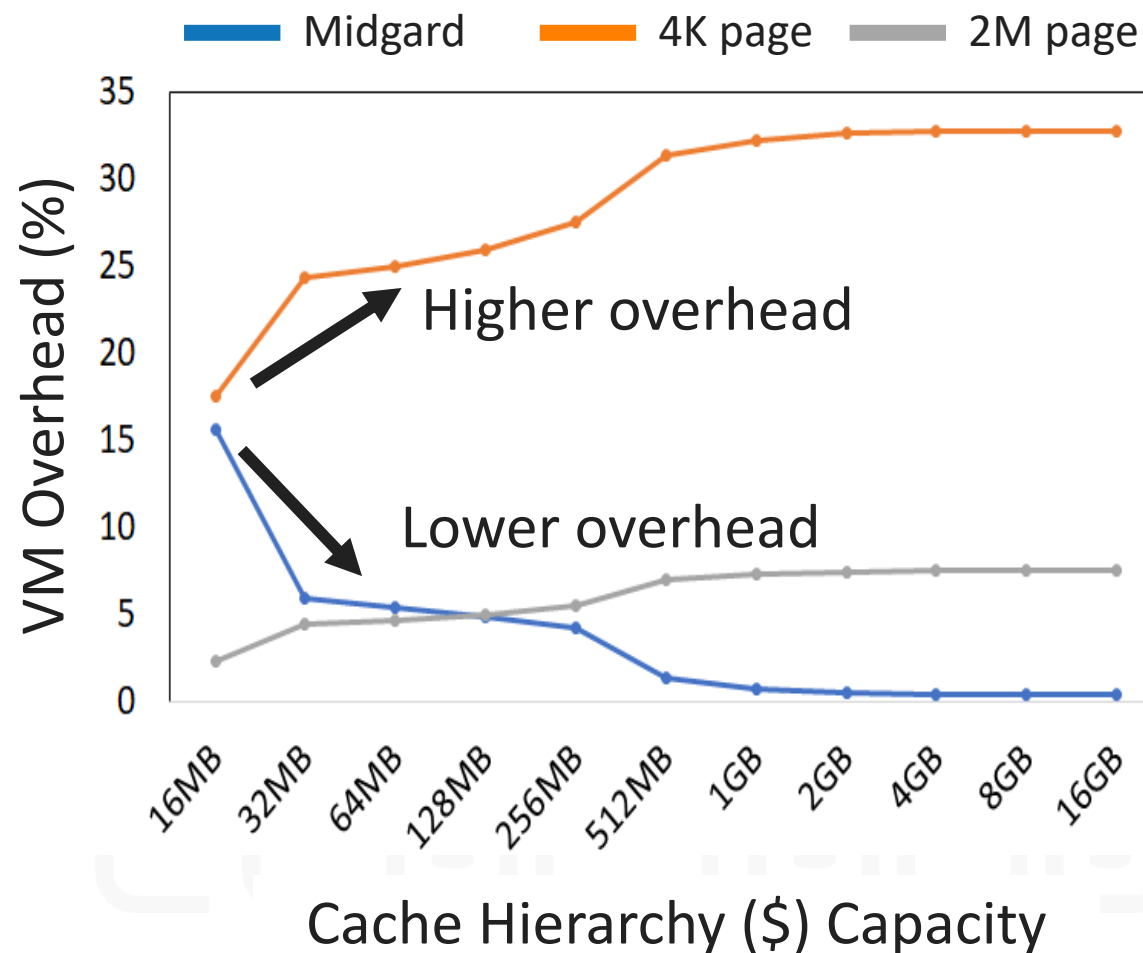
[midgard.epfl.ch](https://midgard.epfl.ch)



midgard



- Keeps POSIX (VMA) interface to apps
  - Linux, MacOS/iOS, Android
- Eliminates page-based translation in \$
- ✓ Unclogs virtual memory for security, virtualization, accelerators



# OUTLINE

- ~~Overview~~

- Post-Moore servers

  - ~~80's Desktops~~

  - ~~Specialized CPUs~~

  - Integrated logic/memory

  - Integrated networks

  - Approximating AI

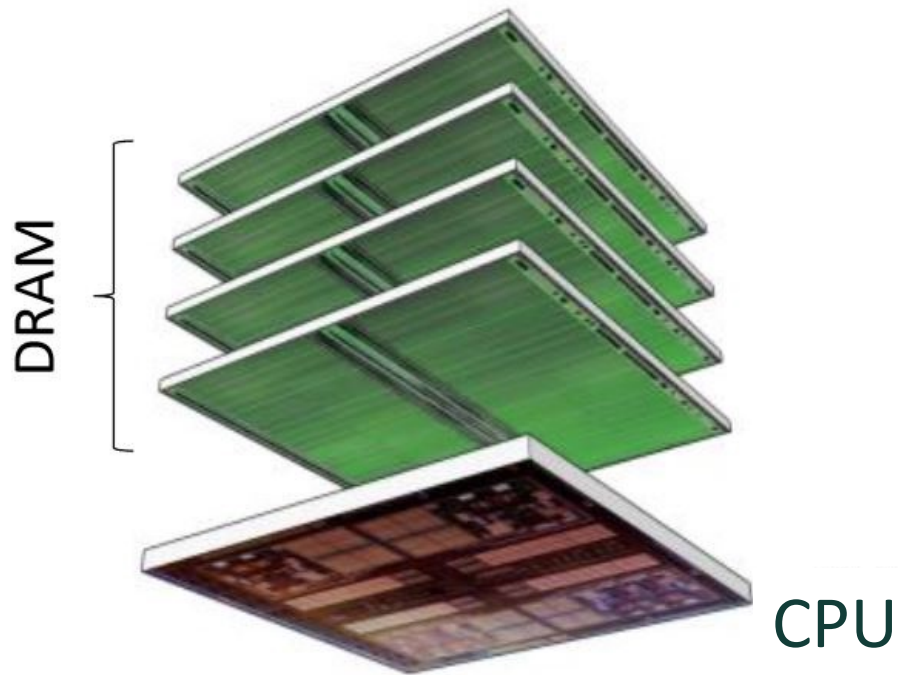
- Summary



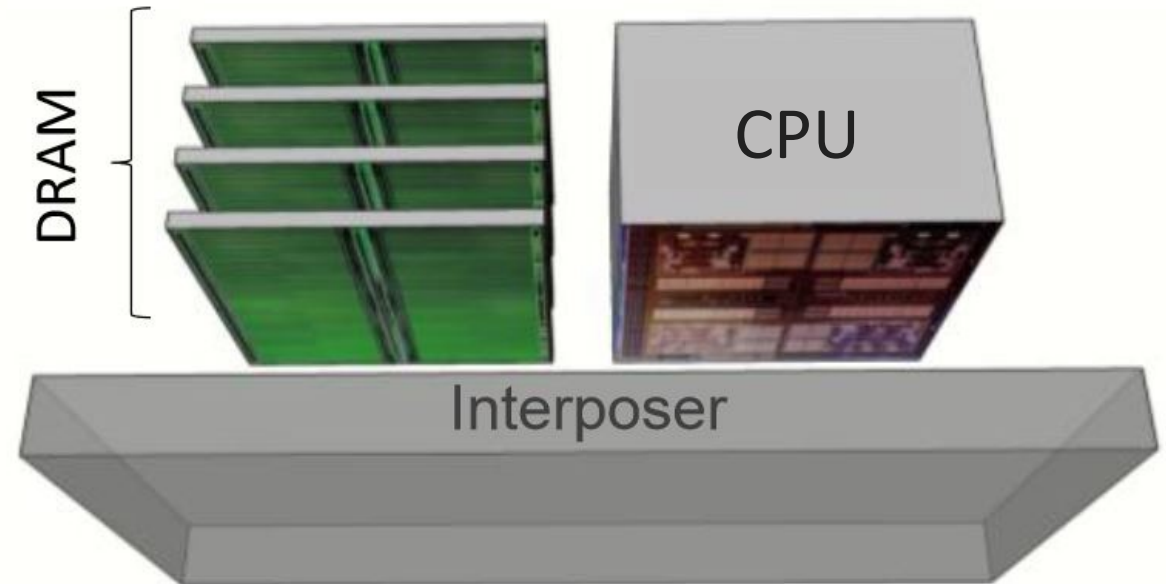
# INTEGRATED LOGIC/MEMORY

Memory chip stack w/ nearby logic

- Minimize data movement
- Massive internal bandwidth

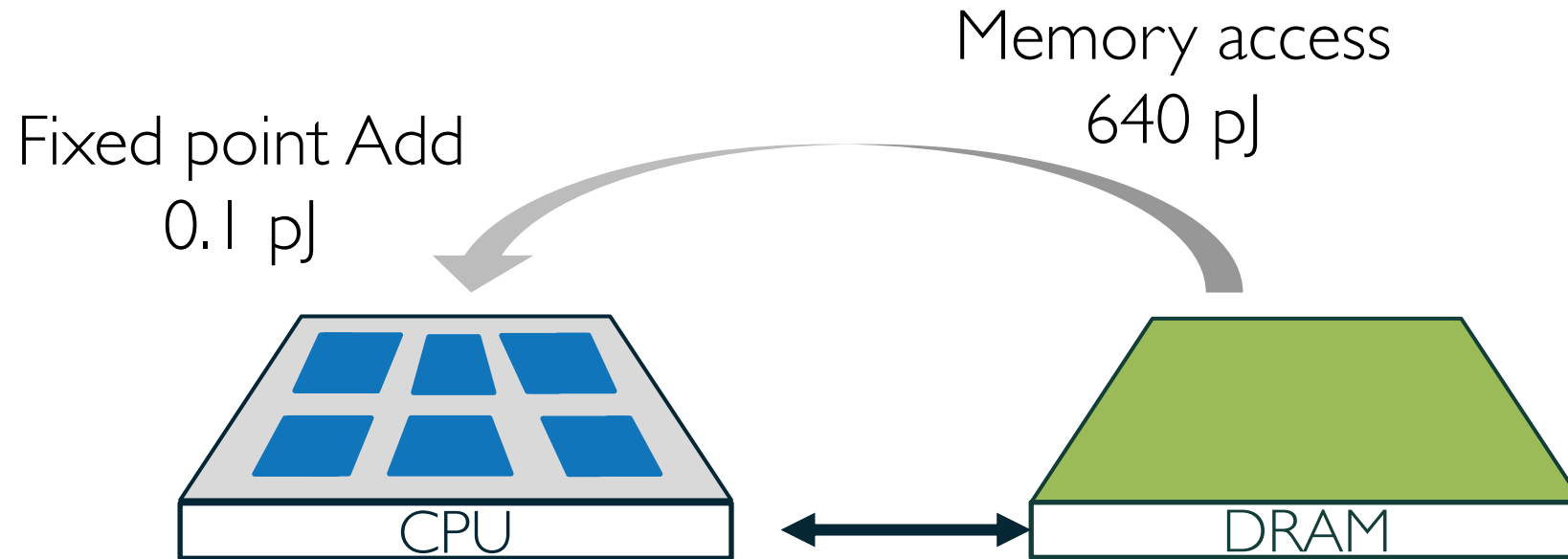


[source: AMD]



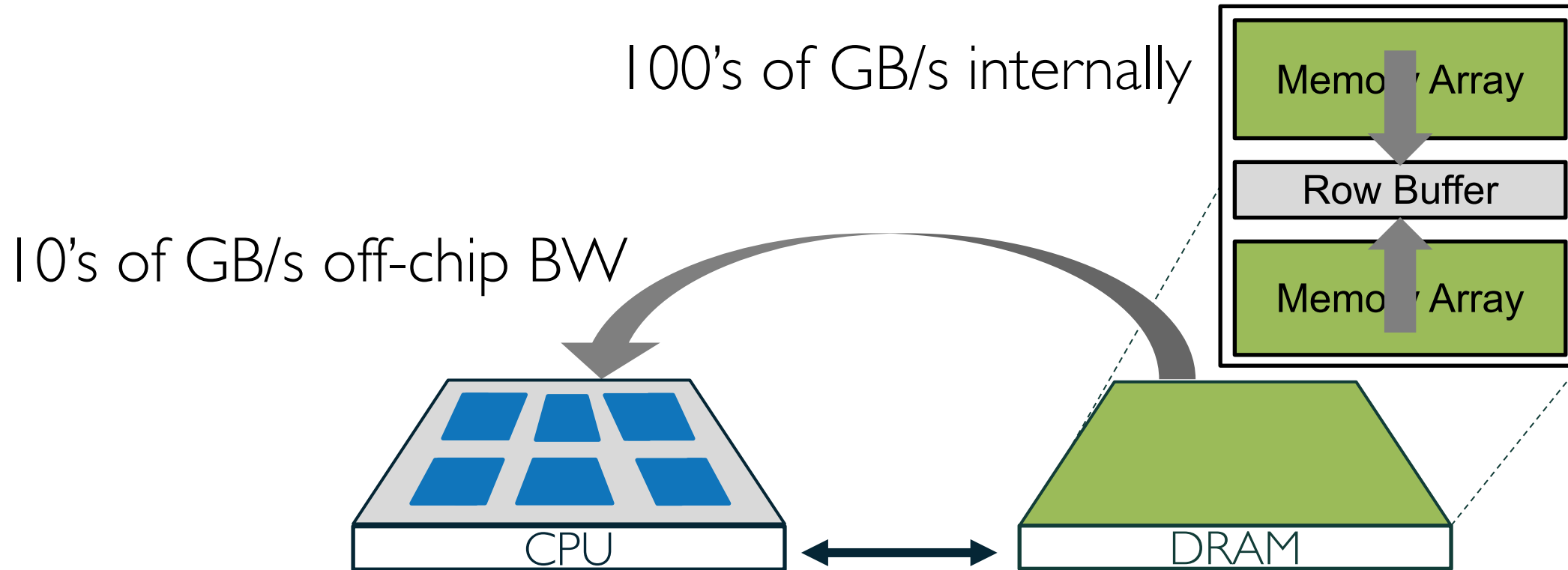
Opportunities for algorithm/hardware co-design

# COST OF MOVING DATA



Data access much more expensive than arithmetic operation

# MEMORY B/W BOTTLENECK



Internal DRAM BW presents big opportunity

# NMP COMMANDMENTS

[IEEE Micro on Big Data'16]

Not (CPU) business as usual

1. DRAM favors streaming over random access
2. DRAM favors parallelism over arithmetic speed
3. NMP DRAM must maintain CPU memory semantics

Co-design algorithm/HW for NMP

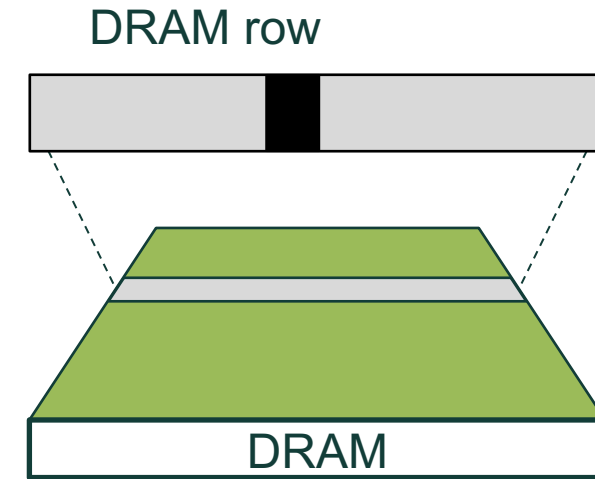
# WHY NOT RANDOM ACCESS?

Internally DRAM is a block device

- Activating a 1KB row
- High latency & energy per row
- Exploit row locality for efficiency

Example:

- For DRAM with 128 GB/s internal bandwidth
- Optimal (parallel) random access only captures ~8 GB/s
- Requires 5x more power



Use algorithms that favor streaming access

## Revisiting Sort join:

- Sort join ( $O(n \log n)$ ) vs. Hash Join ( $O(n)$ )
- Sort tables and then merge join
- Streaming vs. random access

Perform way more work

But, finish faster and use less power!

Trade off algorithm complexity for sequential memory accesses

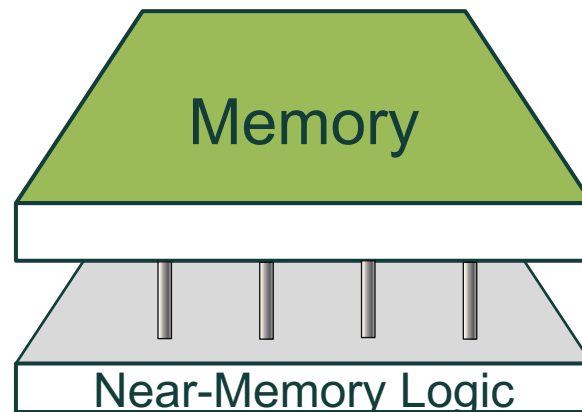
# THE MONDRIAN ENGINE [ISCA'17]

## 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

# OUTLINE

- ~~Overview~~

- Post-Moore servers

  - ~~80's Desktops~~

  - ~~Specialized CPUs~~

  - ~~Integrated logic/memory~~

  - Integrated networks

  - Approximating AI

- Summary

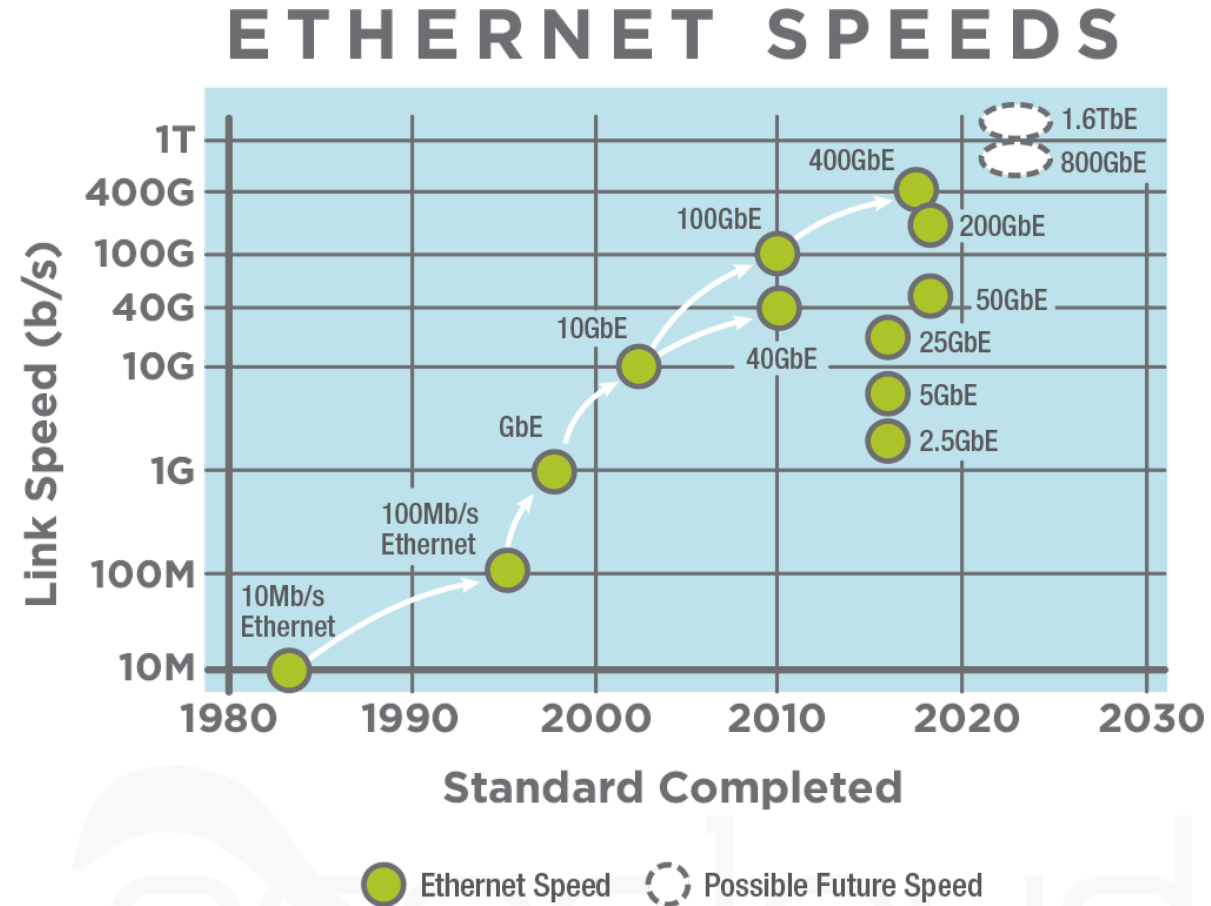


## Network stack bottleneck:

- B/W growing faster than silicon
- Emerging  $\mu$ Services + serverless
- RPC, orchestration, ....

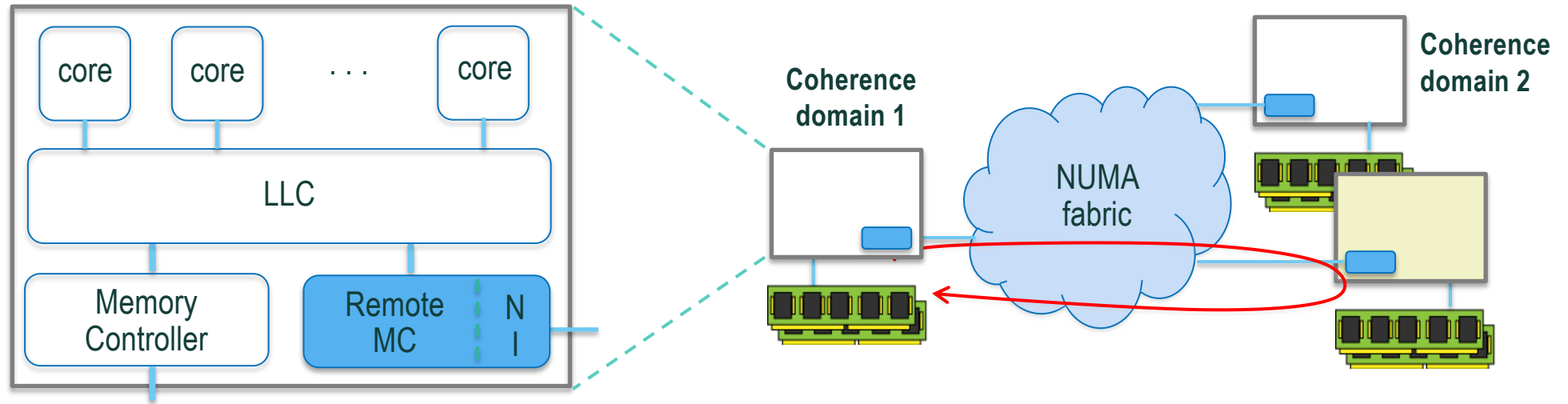
## Key challenges:

- New abstractions
- Co-design of network stacks



# SCALE-OUT NUMA

[ASPLOS'14'19, ISCA'15, MICRO'16]

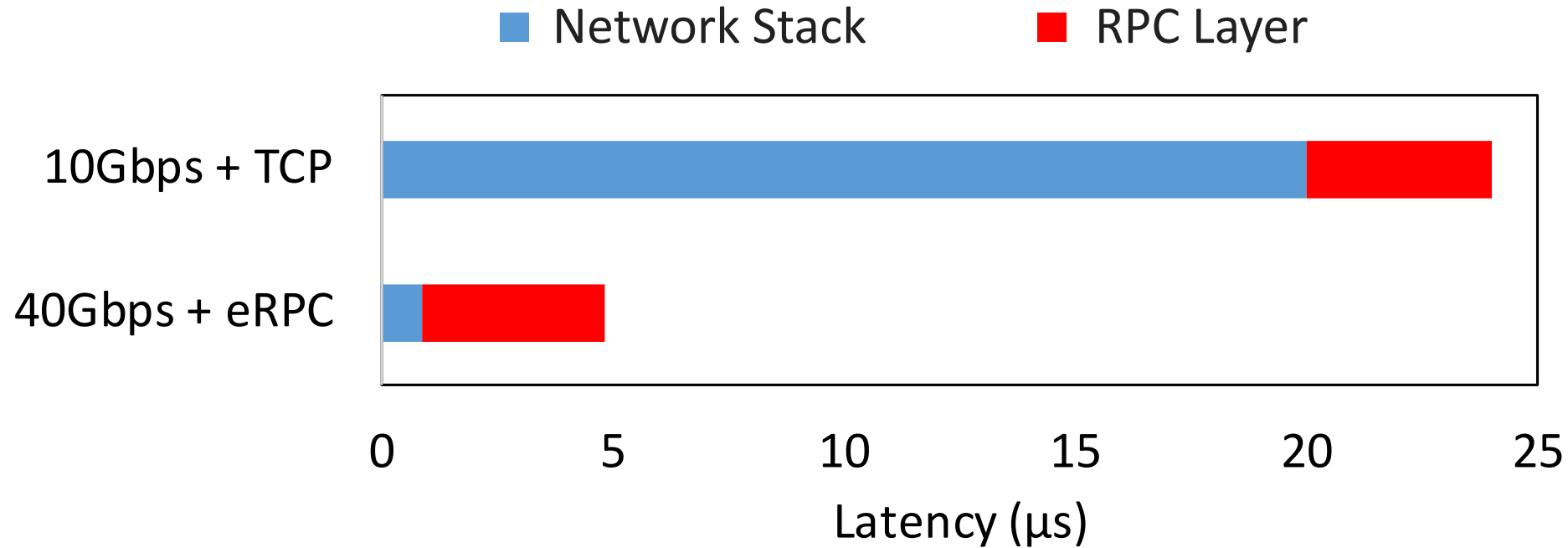


*300ns round-trip latency  
to remote memory*

soNUMA:

- Socket-integrated network interface
- Protected global memory read/write + synch
- Fine-grain (~64B) & bulk objects (~1MB)
- Remote memory ~ 2x local memory latency
- Extensions for messaging & RPC

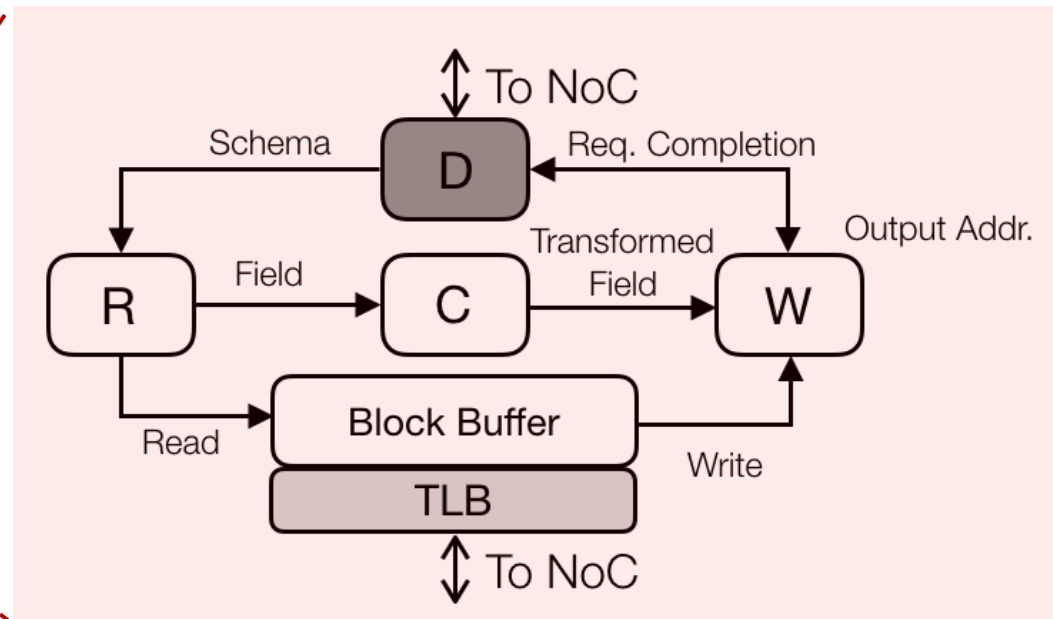
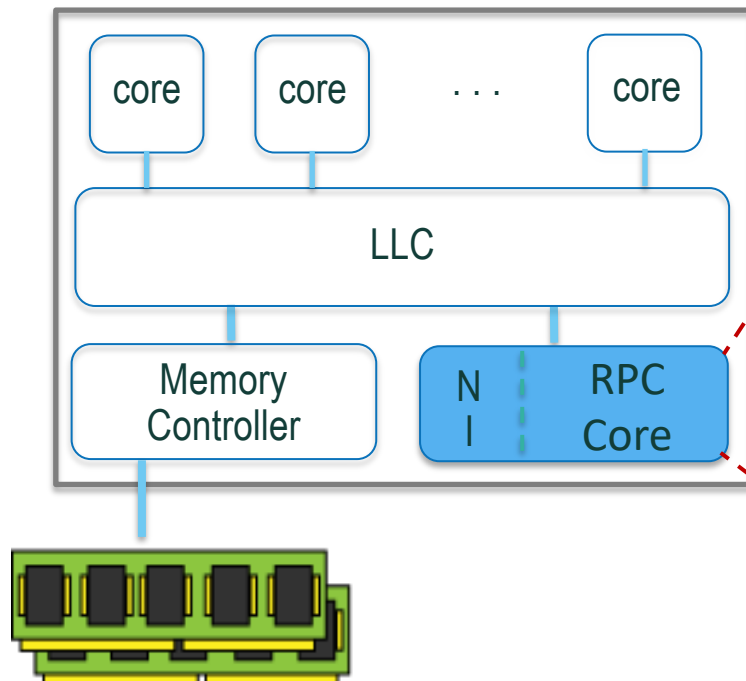
# RPC ACCELERATORS



- Wire time and protocol stacks have shrunk
- RPC dominates CPU cycles in  $\mu$ Services
- E.g., data transformation @  $\sim 2.4$ Gbps w/ Thrift on Xeon

RPC processing at line rate:

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



# OUTLINE

- ~~Overview~~

- Post-Moore servers

  - ~~80's Desktops~~

  - ~~Specialized CPUs~~

  - ~~Integrated logic/memory~~

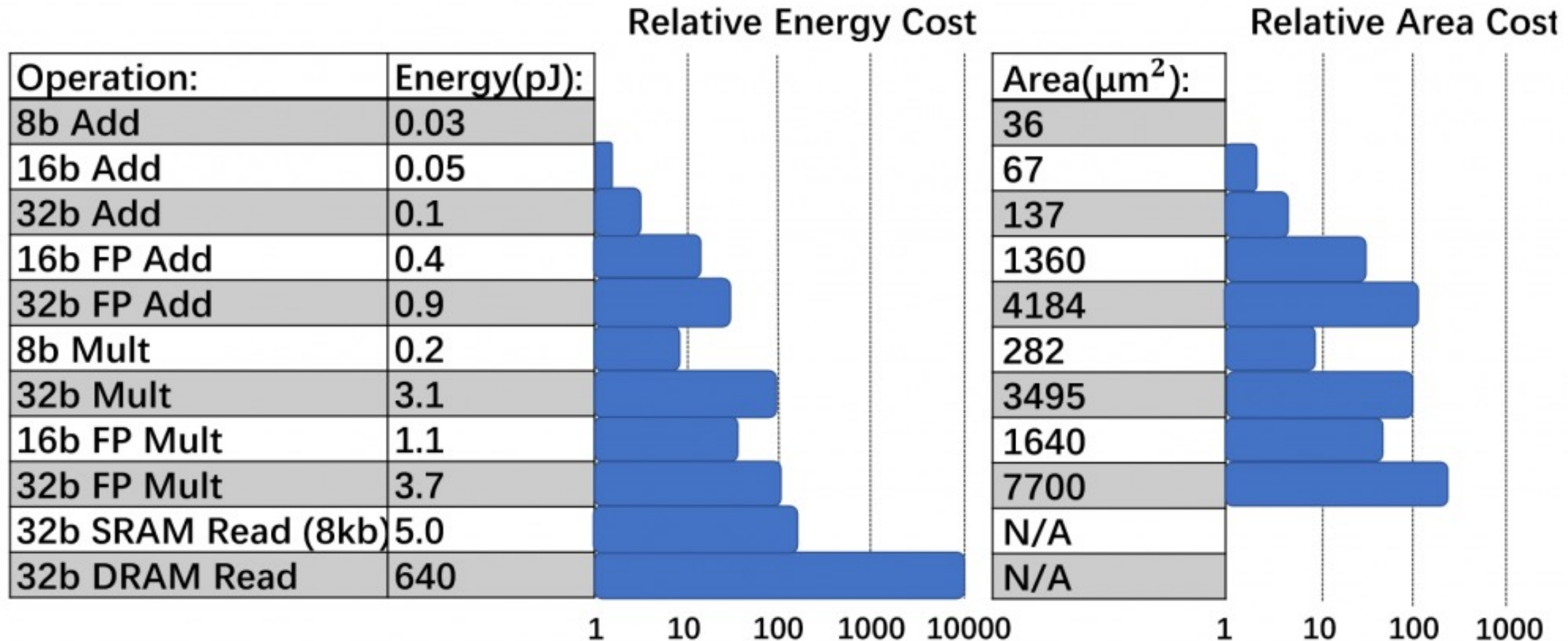
  - ~~Integrated networks~~

  - Approximating AI

- Summary



# COST OF LOGIC VS. MEMORY

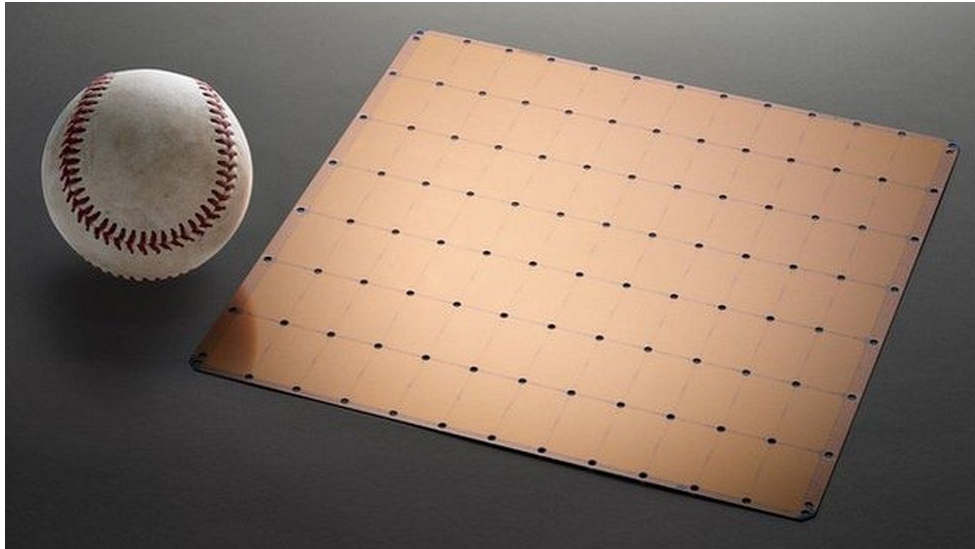
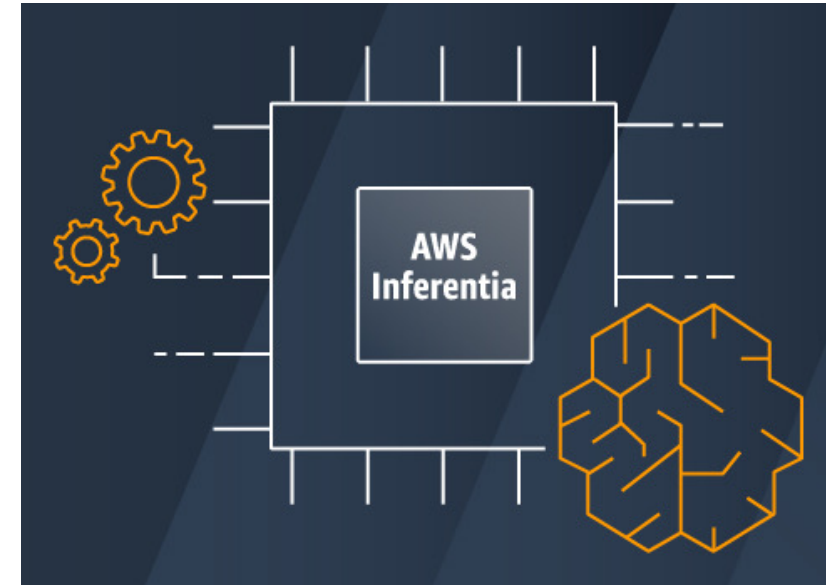


[src: Gholami, et. al.]

# DNN PLATFORM DIVERGENCE

Inference platforms:

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




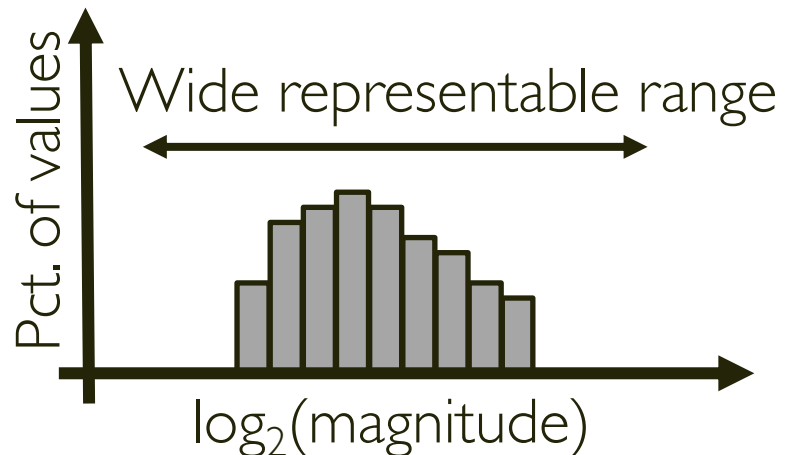
Training platforms:

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

# FLOATING VS. FIXED POINT

## ■ Floating point

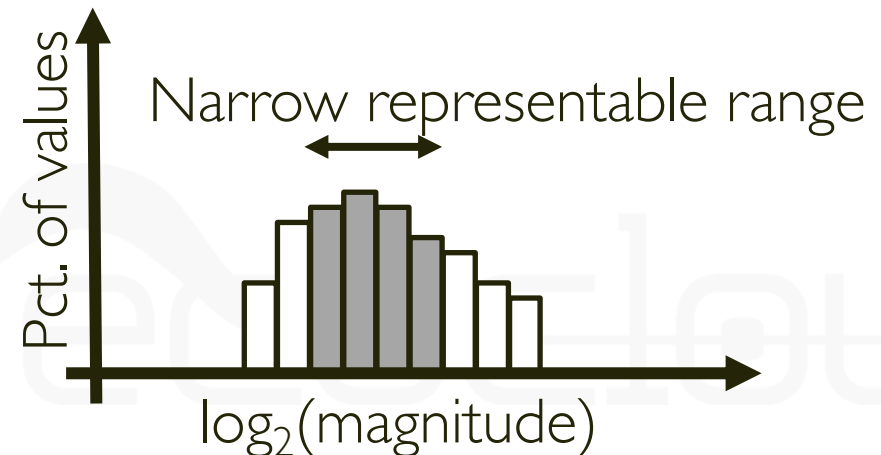
- Mantissa + exponent  

- Wide representable range
- Value has independent range



## ■ Fixed point

- Mantissa  

- Narrow representable range
- Values range pre-determined

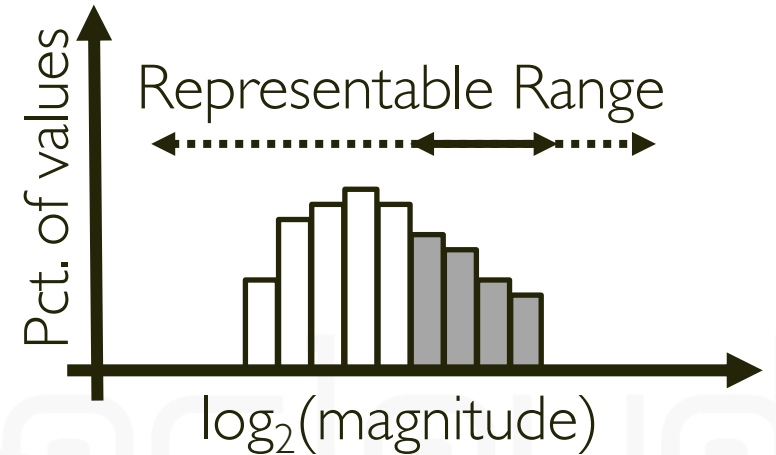
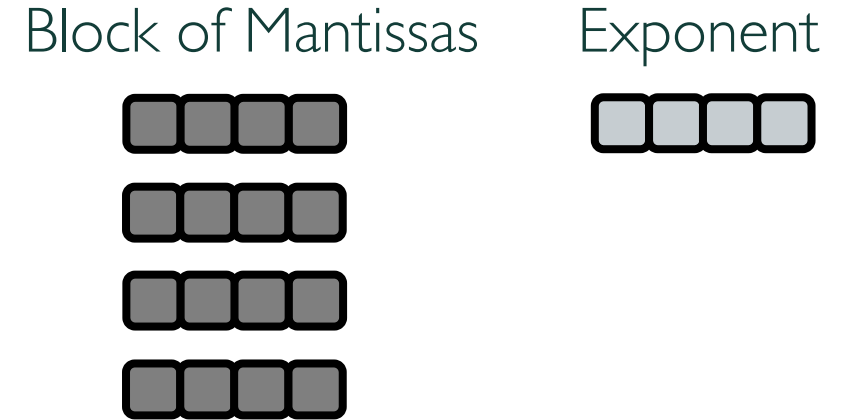


# HYBRID BLOCK FLOATING POINT (HBFP) [NeurIPS'18]

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

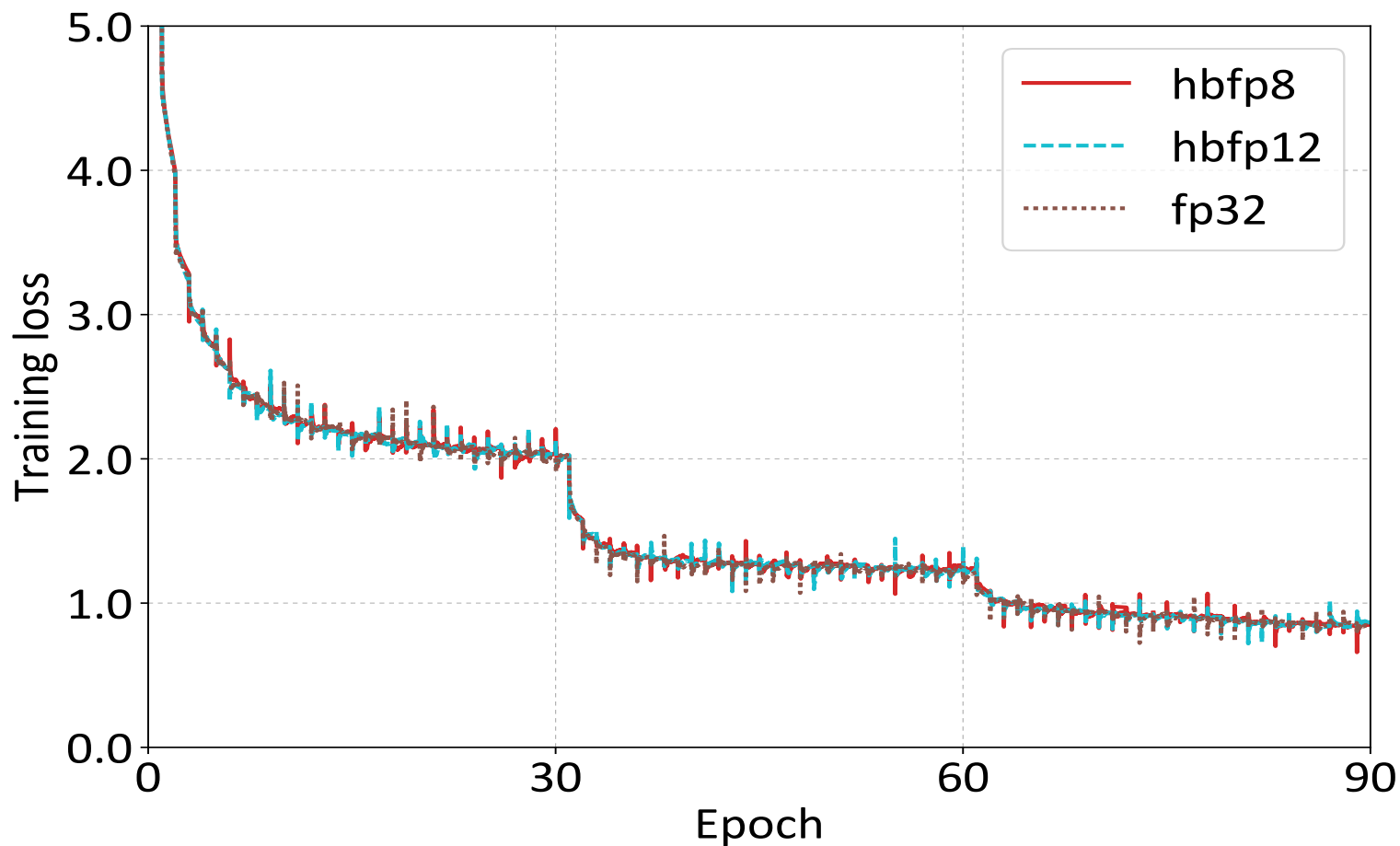
## Co-Located Training & Inference (ColTrain)

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



# HBFP vs. FP32

Resnet-50 on ImageNet



Config.	Top-1 Error (%)
HBFP8	23.88
HBFP12	23.58
FP32	23.64



FP32 performance with 8-bit logic for CNN, LSTM, BERT

## Trends:

- Demand is growing faster than Moore
- Moore's law is slowing down

## Post-Moore servers:

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

Integration + Specialization + Approximation

# THANK YOU!

For more information, please visit us at  
[parsa.epfl.ch](https://parsa.epfl.ch)

# EPFL

