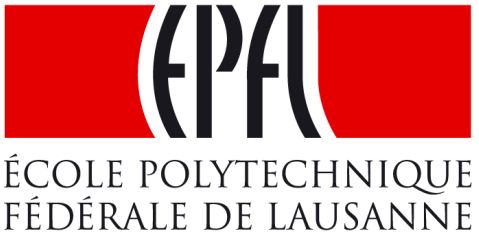
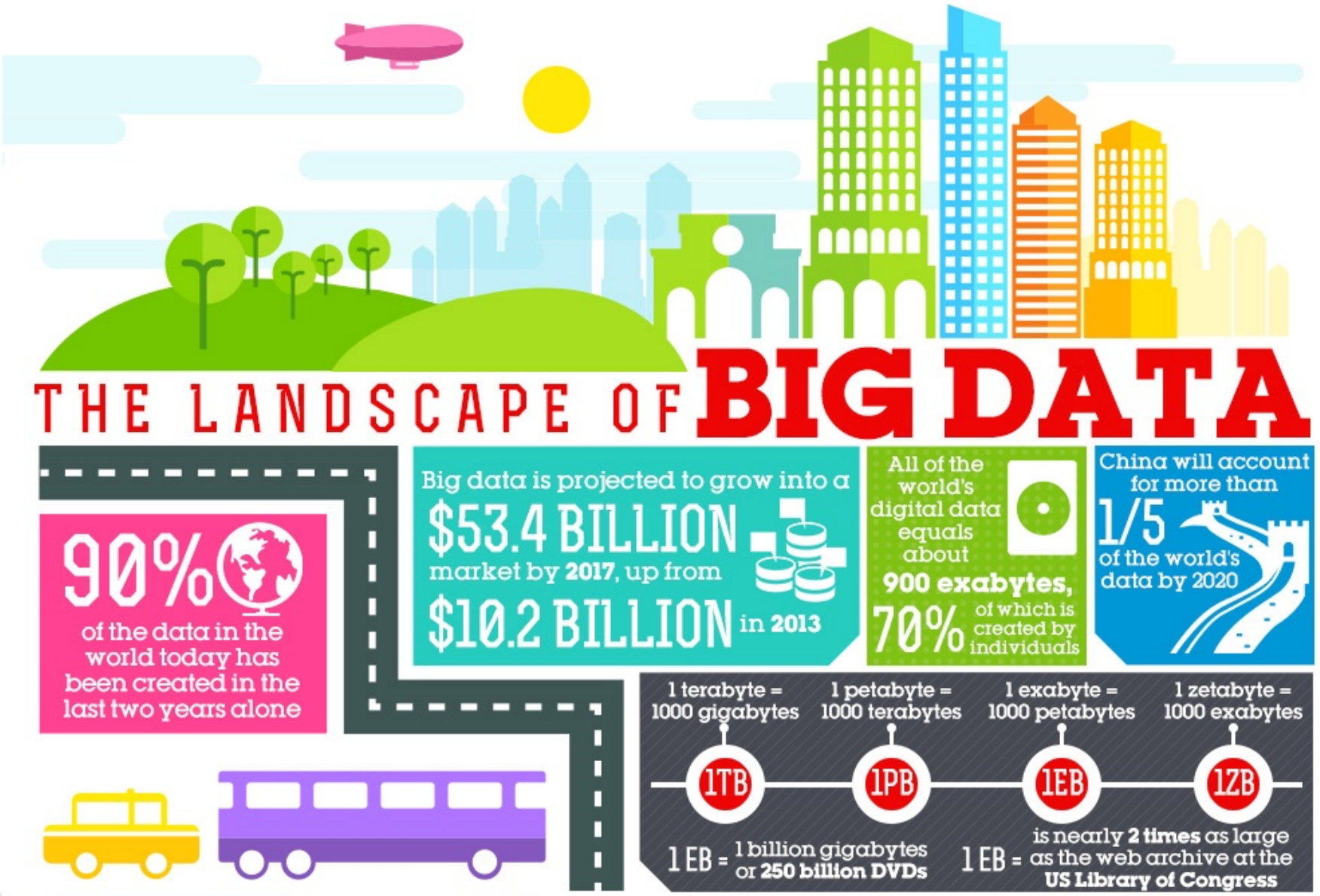
Silicon Heterogeneity in the Cloud

Babak Falsafi
ecocloud.ch



Data Economics

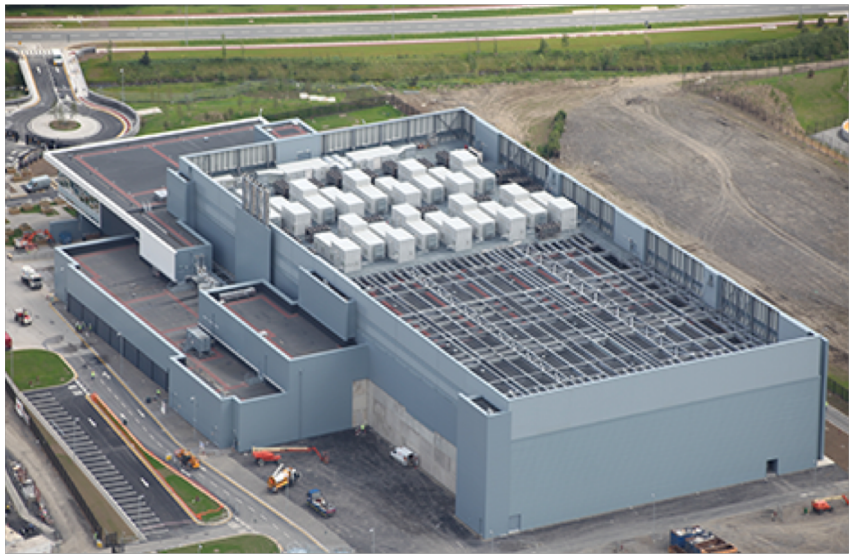
ecocloud
an EPFL research center



Modern Datacenters are Warehouse-Scale Computers

ecocloud
an EPFL research center

- Millions of interconnected home-brewed servers
- Centralization helps exploit economies of scale
- Network fabric provides micro-second connectivity
- At physical limits
- Need sources for
  - Electricity
  - Network
  - Cooling



20MW, 20x Football Field
$3 billion

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

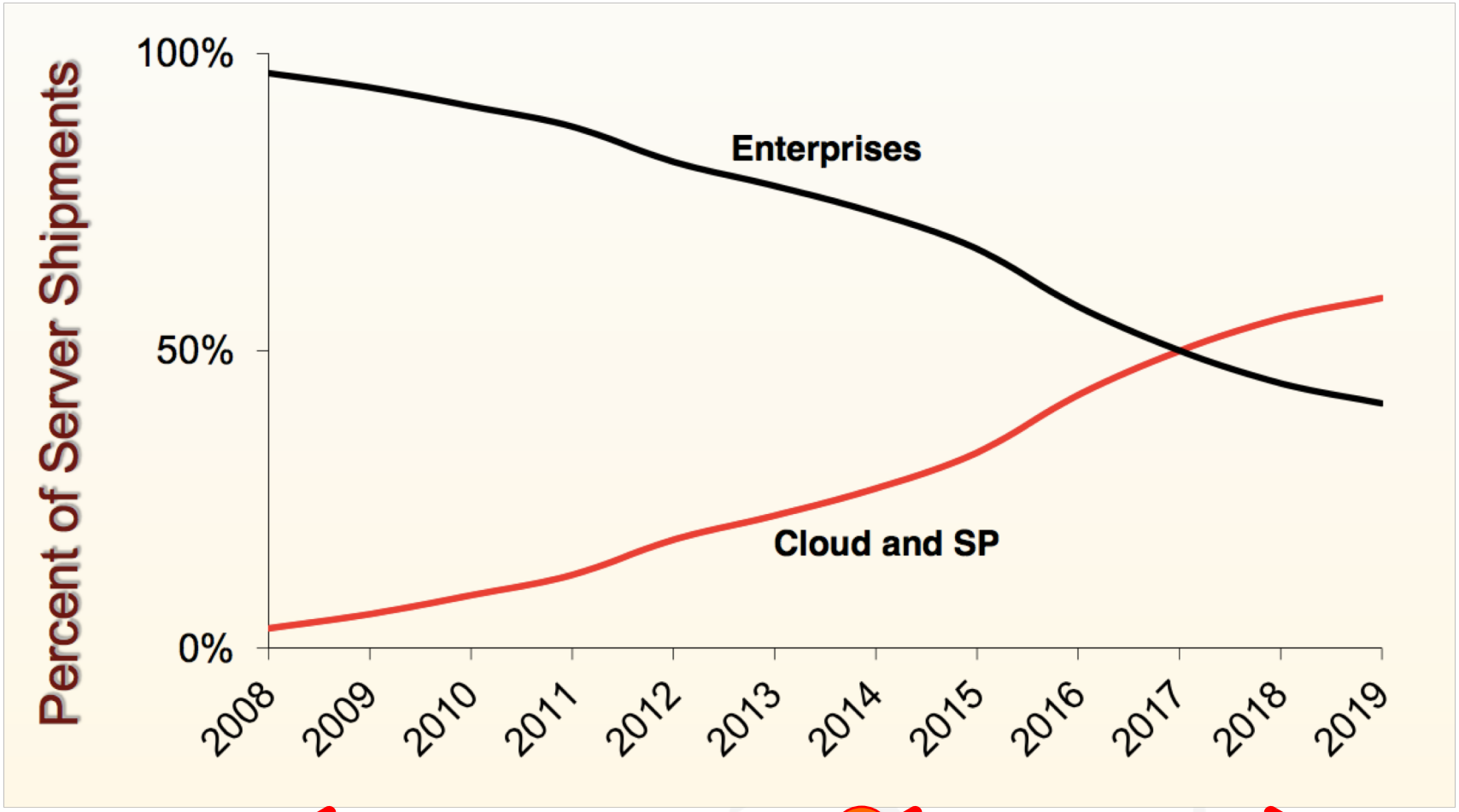
Datacenters are the IT utility plants of the future

Supercomputing

≠

Cloud Computing

Cloud Taking Over Enterprise

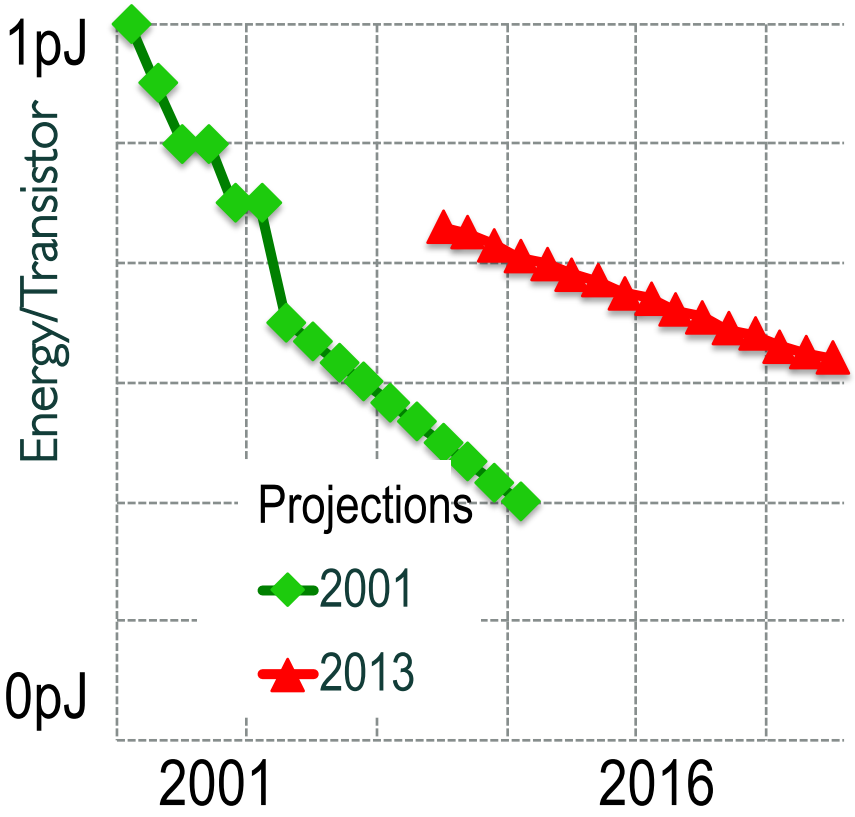


Source: Dell 'Oro 2Q15

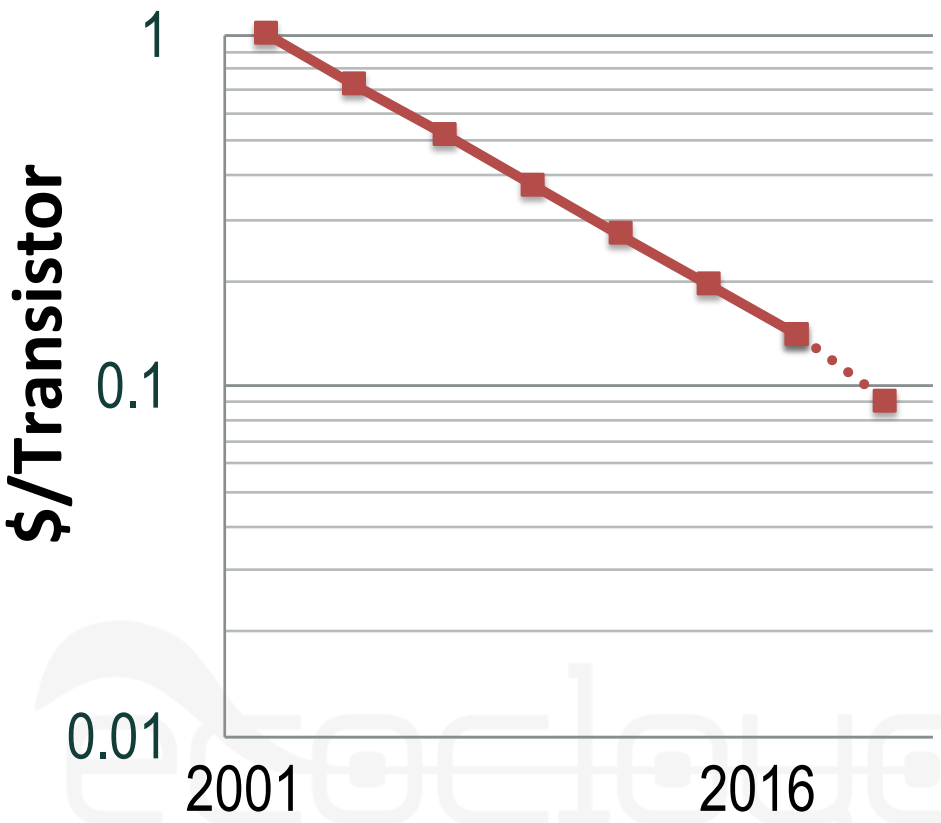
But, Silicon out of steam!

ecocloud
an EPFL research center

Silicon efficiency is dead
(long live efficient silicon)



Moore's Law is Dead too!
[Mark Bohr's Keynote, ISSCC'15]



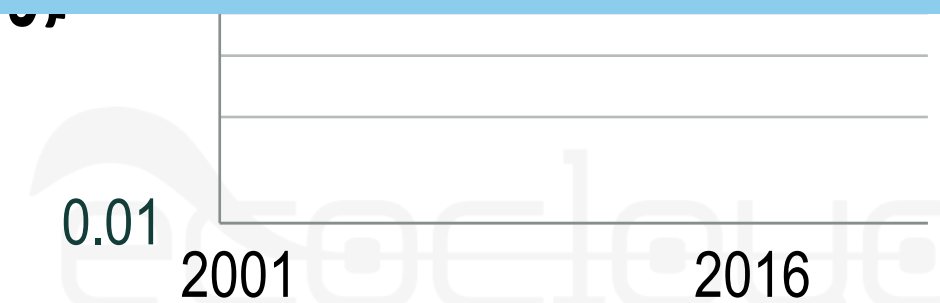
But, Silicon out of steam!

ecocloud
an EPFL research center

Silicon efficiency is dead
(long live efficient silicon)

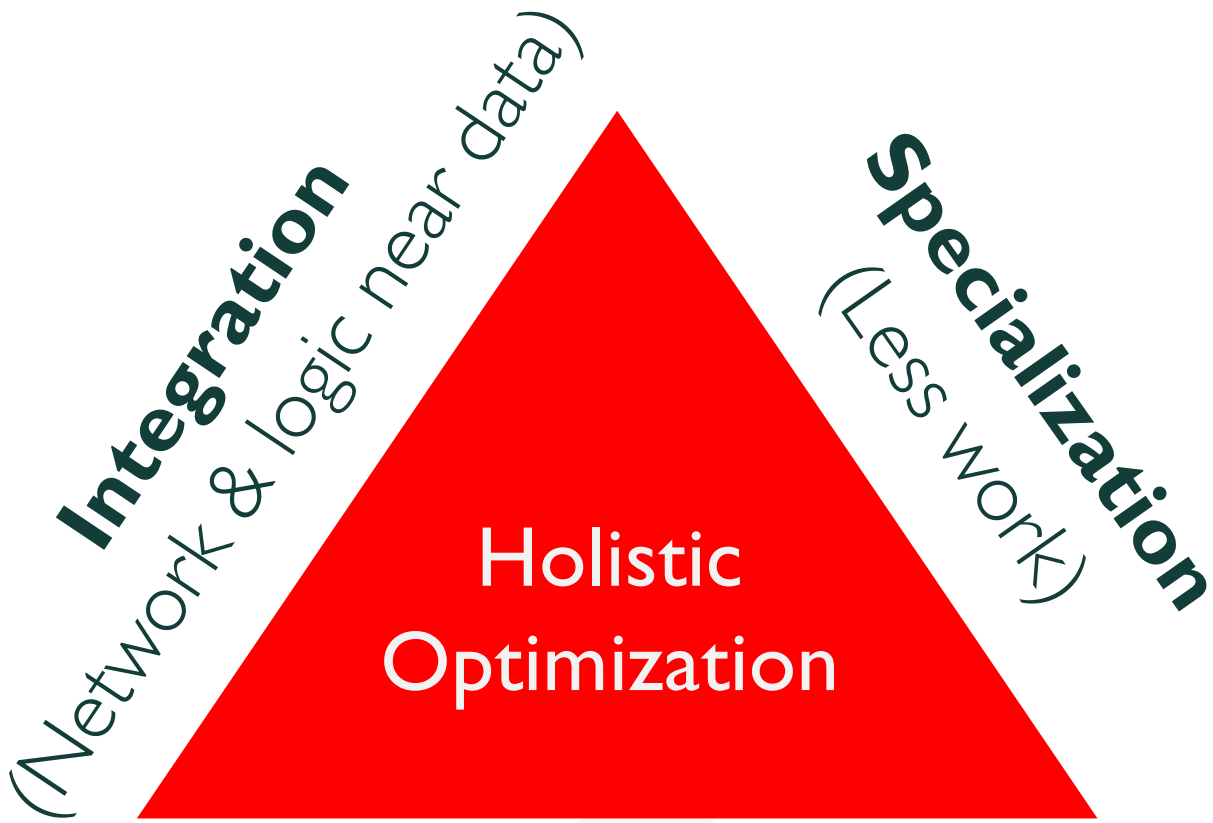
Moore's Law is Dead too!
[Mark Bohr's Keynote, ISSCC'15]

Global Foundries cancelled their 7nm on August 28, 2018!



Optimization Opportunities: The ISA Triangle

ecocloud
an EPFL research center



Scale-Out Datacenters

ecocloud
an EPFL research center

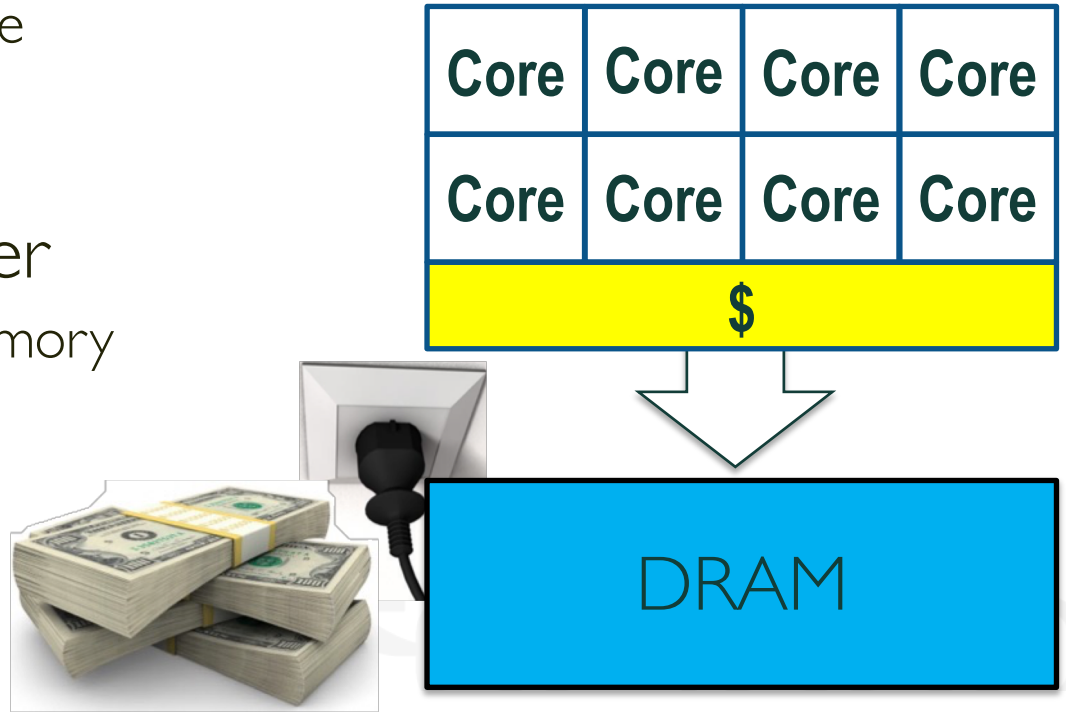
Vast data sharded across servers

Memory-resident workloads

- Necessary for performance
- Major TCO burden

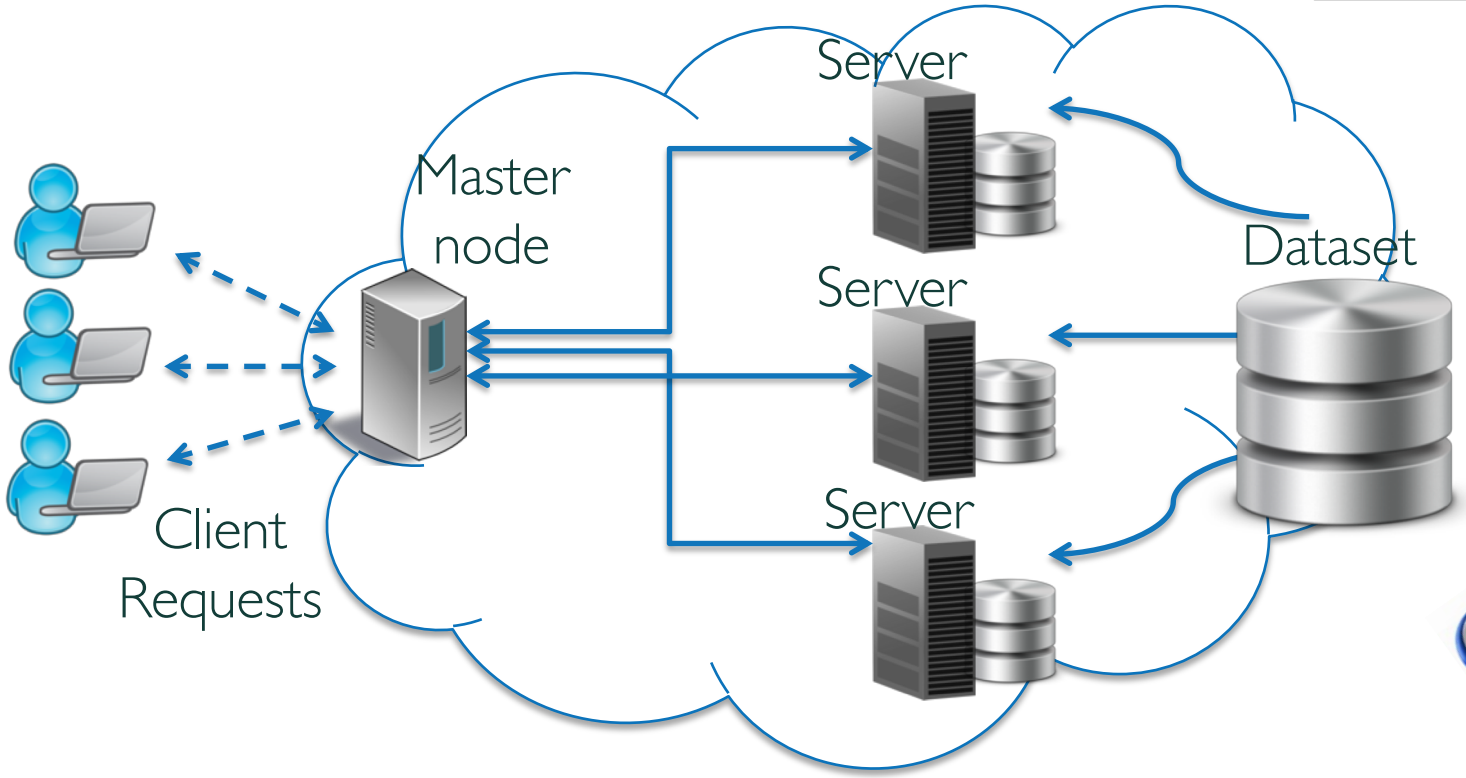
Put memory at the center

- 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

Server Benchmarking with CloudSuite 3.0 (cloudsuite.ch)

ecocloud
an EPFL research center

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!

Scaling CPU's: Manycores

ecocloud
an EPFL research center

- Parallelism has emerged as the only silver bullet
- Use simpler cores
  - Prius instead of Audi R8
- Restructure software

- Each core ➔

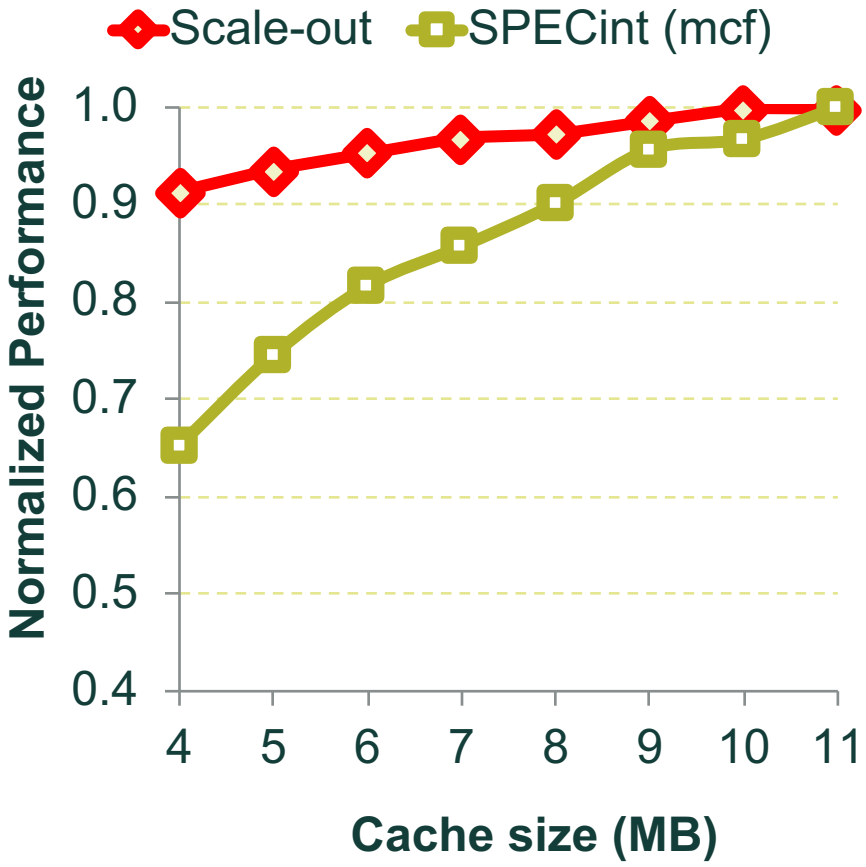
  fewer joules/op

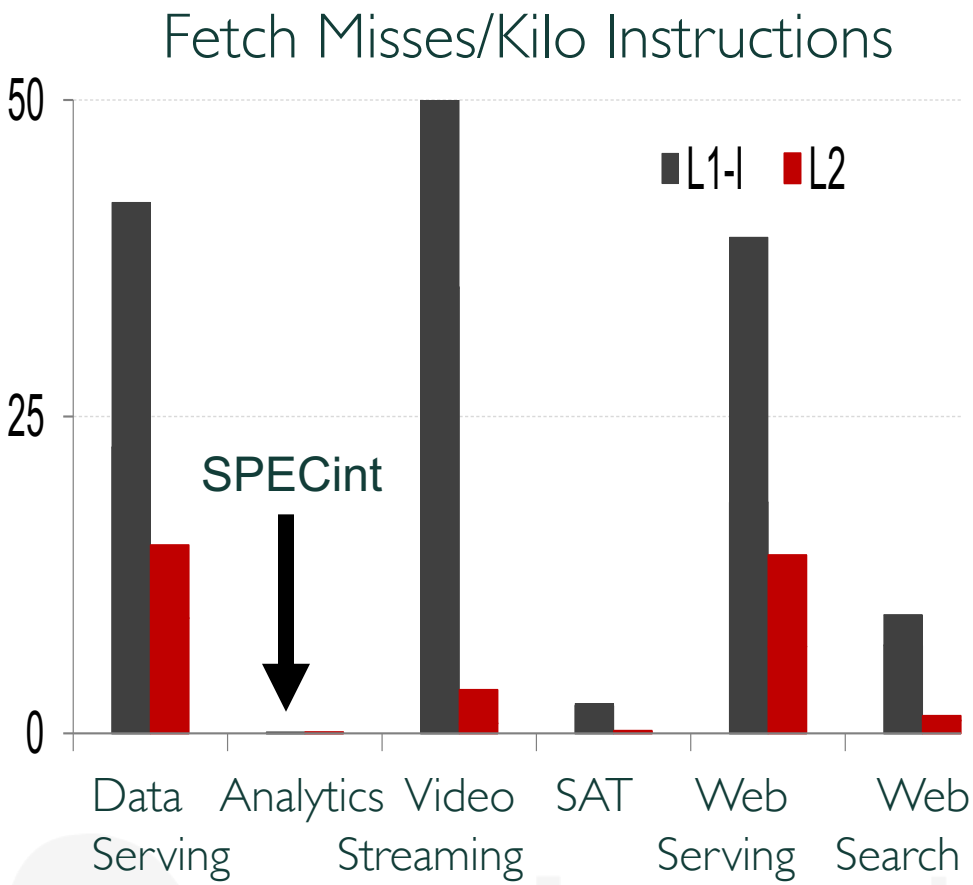
Conventional Server CPU (e.g., Xeon)

Modern Manycore CPU (e.g., Tilera)

But, Services Stuck in Memory (x86 servers) [ASPLOS'12]

ecocloud
an EPFL research center

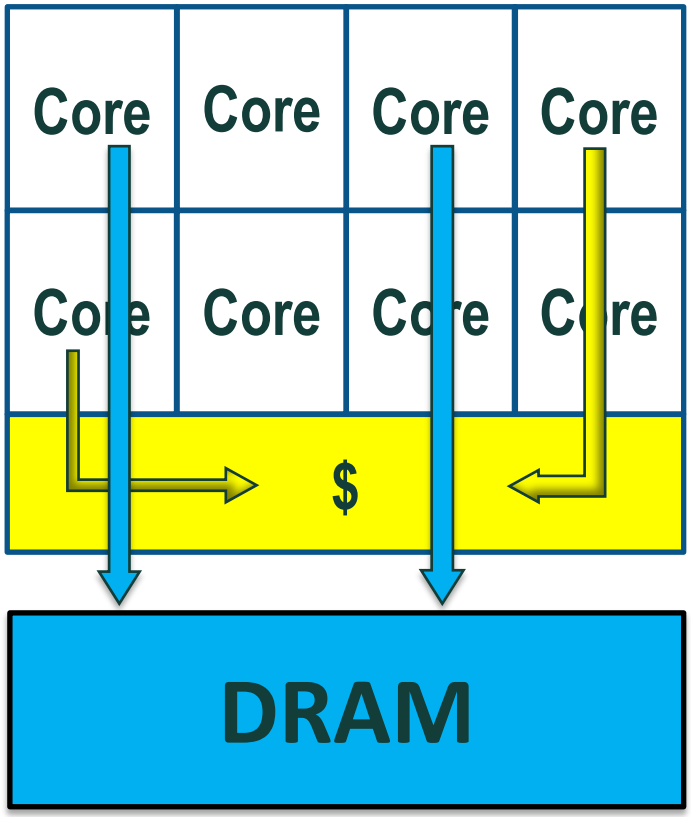


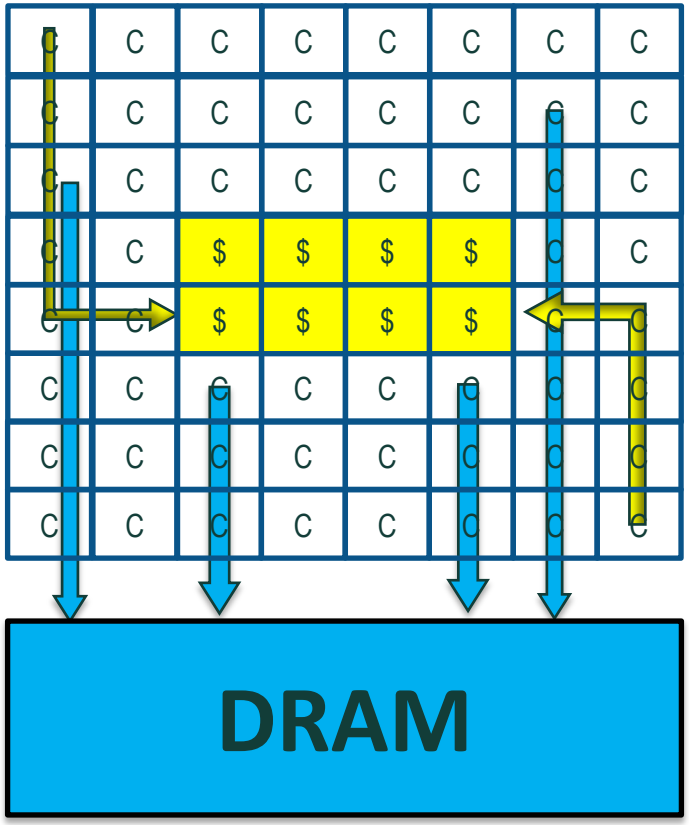


- On-chip memory overprovisioned
- Instruction supply is bottlenecked

Scale-Out Processors (SOP)

ecocloud
an EPFL research center





- x86 server CPU
- ✗ Logic 60% of silicon
- ✗ 6x bigger cores

- 3-way ARM manycore
- ✓ Logic 85% of silicon
- ✓ 7x more parallelism

Innovation in SOP's

[ISCA, MICRO, ASPLOS, MemSys, IEEE Micro'12-18]

ecocloud
an EPFL research center

Instruction supply:

- Core front-end (BP/BTB, Boomerang [Grot])

On-chip networks:

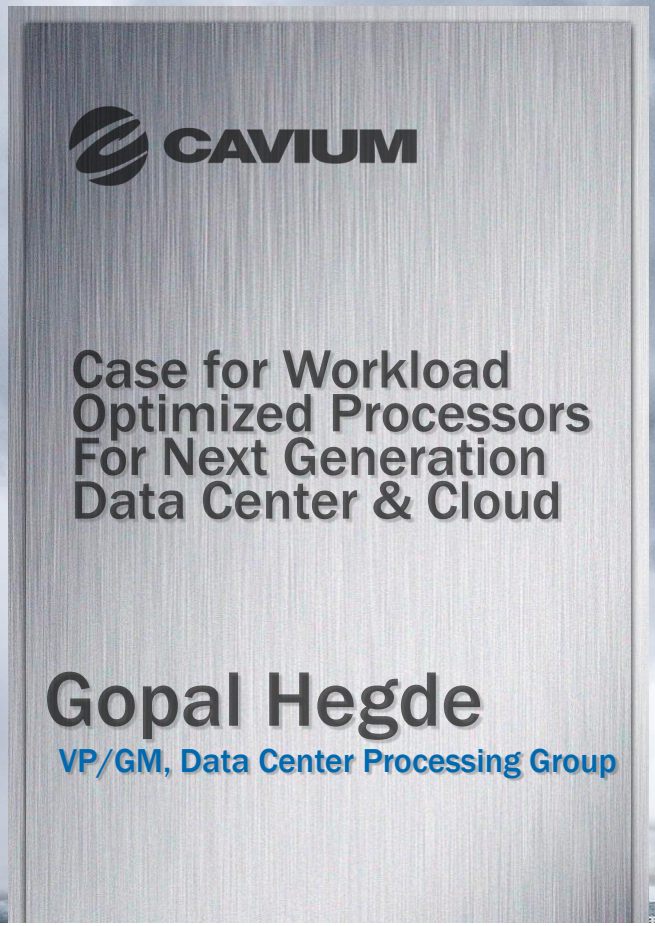
- Core-to-$ rather than core-to-core

Off-chip connectivity:

- HBM, DRAM hierarchy, network

Example SOP

ecocloud
an EPFL research center



Cavium Thunder X

- Based on SOP @ EPFL
- Designed to serve data
- Optimized code supply
- Trade off SRAM for cores
- Runs stock software
- 10x faster than Xeon for CloudSuite

Massively parallel cores

ecocloud
an EPFL research center

- Data parallelism
- Higher memory b/w

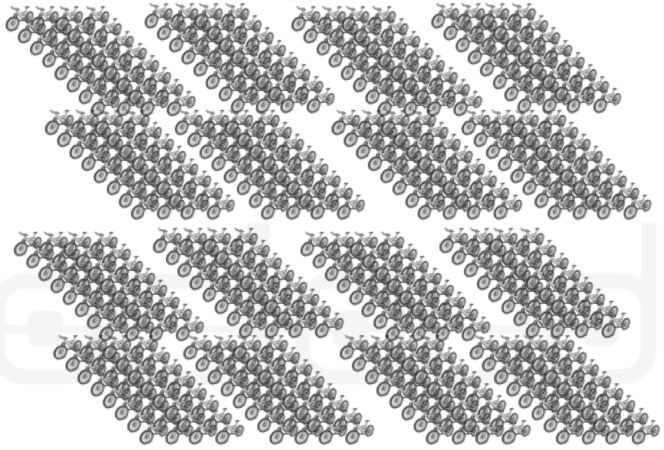
Super simple cores

- Shared front end
- 10x slower clocks

Great for dense parallel computation

Conventional Server CPU (e.g., Xeon)

Modern GPU (e.g., Volta)



Parallelism Alone Can't Help

ecocloud
an EPFL research center

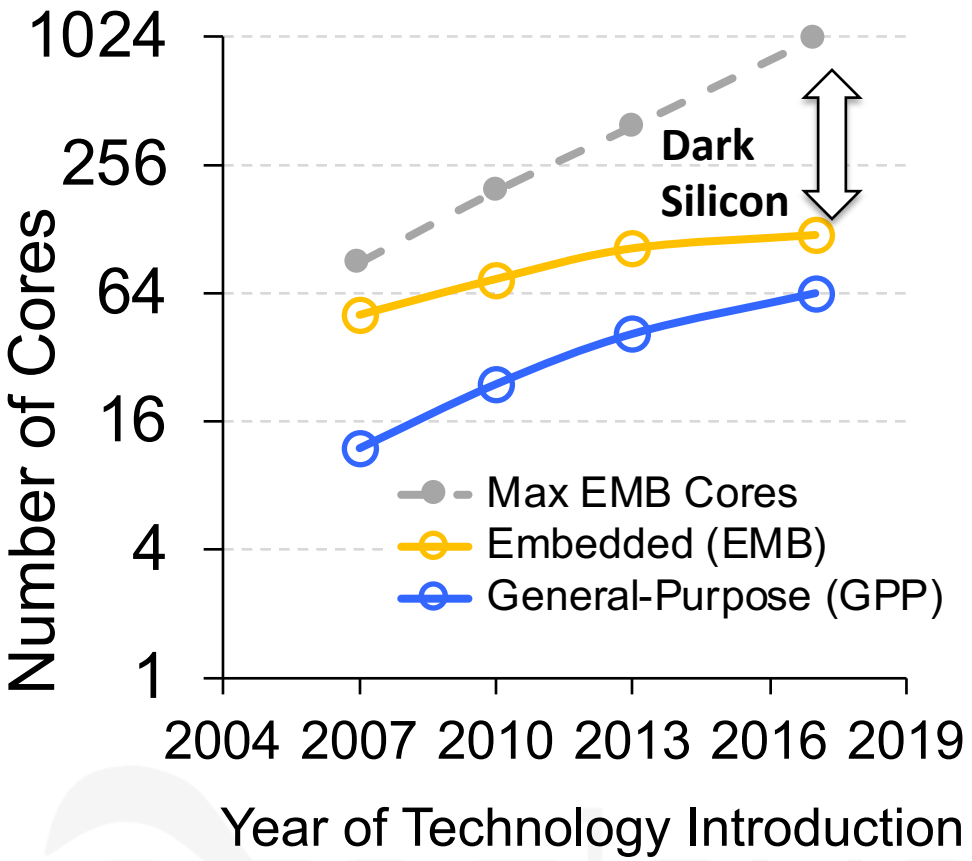
Can populate chips

But, can not operate all

Today's chips are already "dark" (memory)

All future platforms will be heterogeneous

- Selectively activate parts



[source: Hardavellas et. al., "Toward Dark Silicon in Servers", IEEE Micro, 2011]

Custom Computing

[FPGA's vs. GPU's in Data centers, IEEE Micro'17]

ecocloud
an EPFL research center

Reconfigurable

- Best for spatial computing
- Not caching/reuse

Parallel, spatial computing

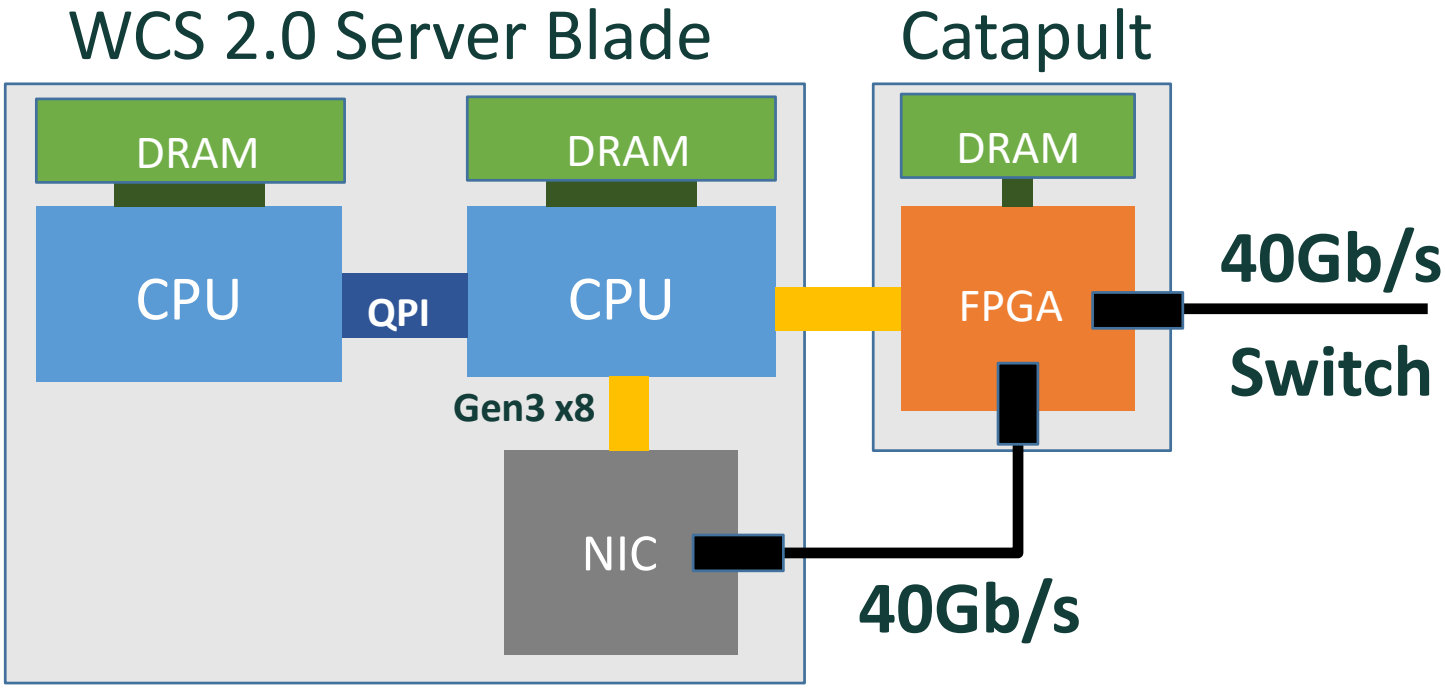
- 10x slower clocks
- Better for sparse arithmetic

Microsoft, Amazon & Intel

Conventional Server
CPU (e.g., Xeon)

FPGA
(e.g., Catapult)

Microsoft's Catapult



FPGA economies of scale:

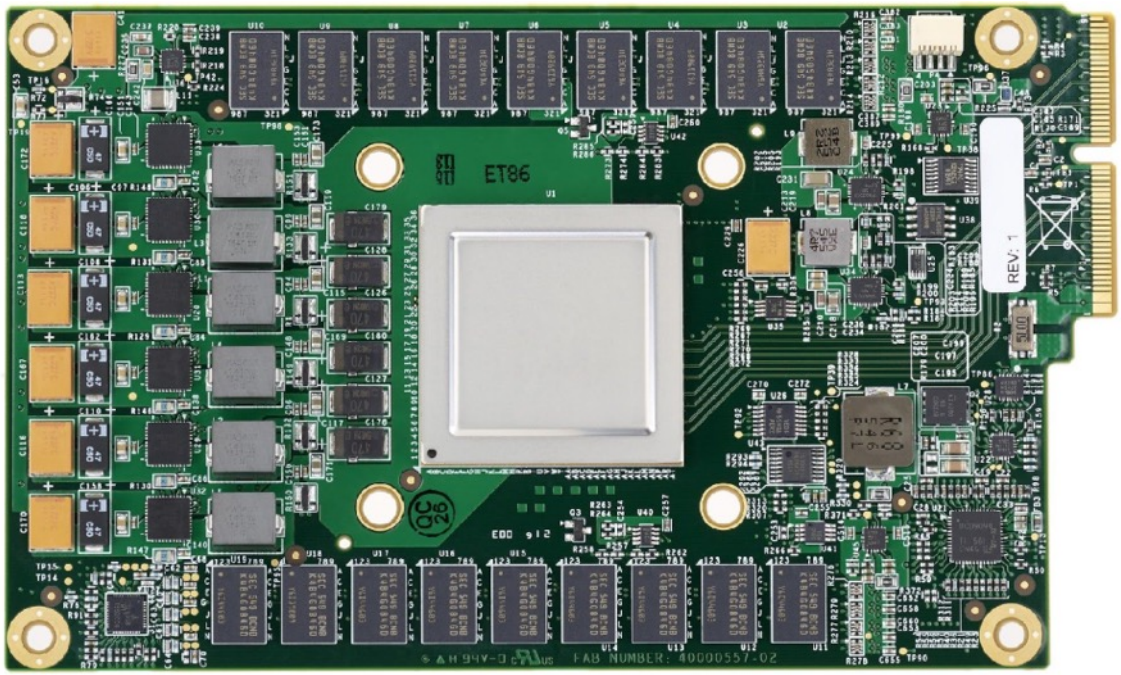
1. Local/remote compute accelerator
2. Network/storage accelerator
3. Configurable cloud

Google's TPU

ecocloud
an EPFL research center

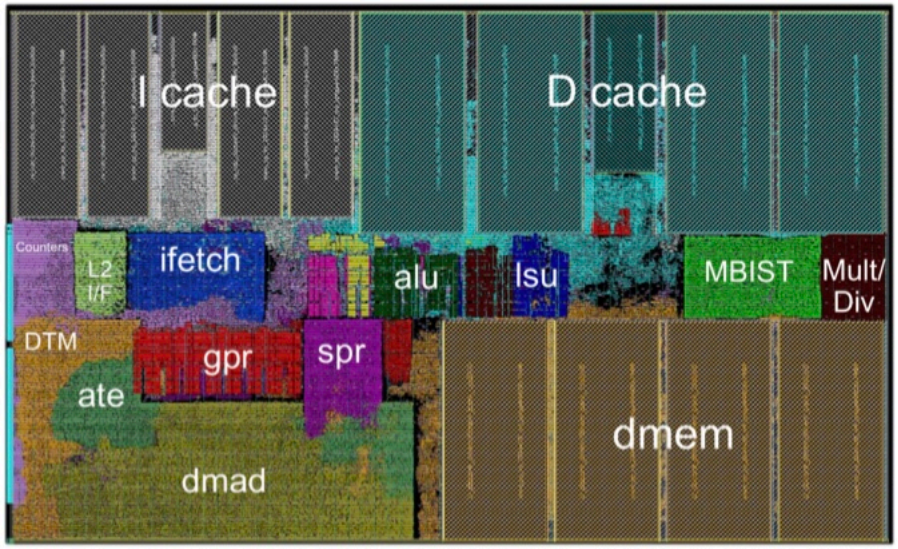
Custom array of arithmetic units:

- Linear algebra for ML/NN
- Currently memory bound
- 10x over GPU
- ML as a service



Oracle's RAPID

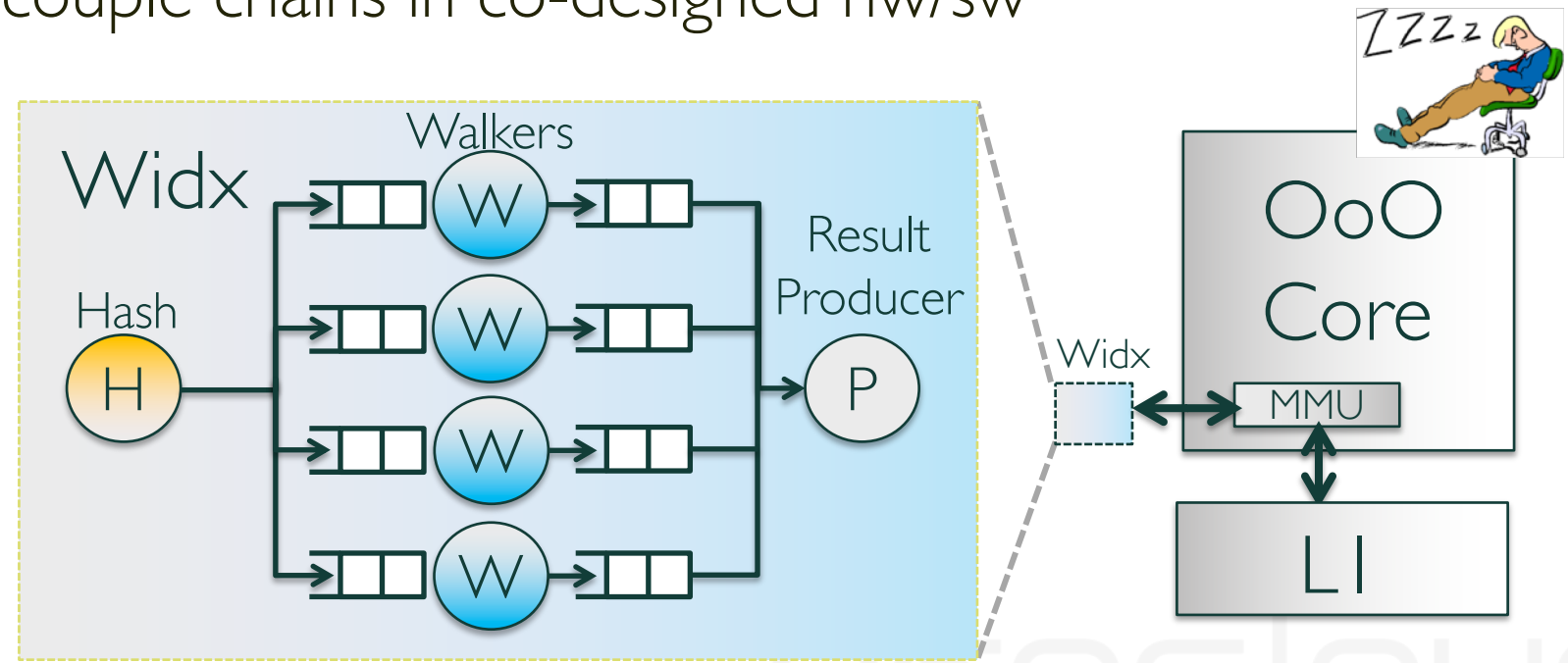
- Accelerator for analytics in SQL
- Data movement engine in hardware
- Custom message passing cores
- Up to 15x better perf/Watt over Xeon



Walkers: CPU-Side Database Accelerators

ecocloud
an EPFL research center

- Pointer-based data structures (e.g., hash table, B-tree)
- Parallel lookups require traversing chains
- Decouple chains in co-designed hw/sw



15x better perf/Watt over Xeon

Walkers in Software [VLDB'16]

ecocloud
an EPFL research center

Use insights to help Xeon

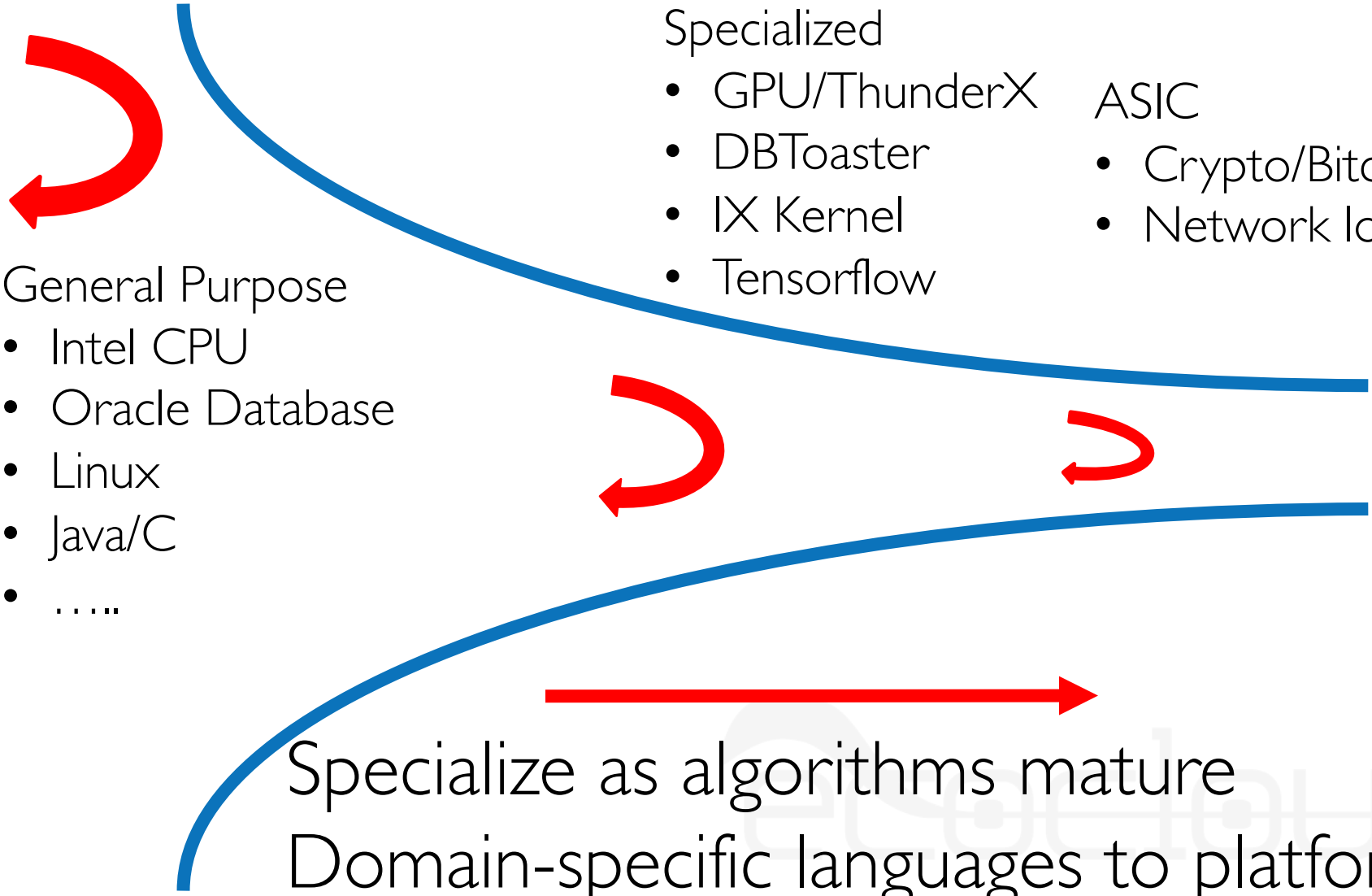
- Decouple hash & walk in software
- Schedule off-chip pointer access with co-routines

2.3x speedup on Xeon

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

Moving Forward: The Specialization Funnel

ecocloud
an EPFL research center



Approximation

ecocloud
an EPFL research center

Modern apps/services are statistical

- Analog input, analog output

Key:

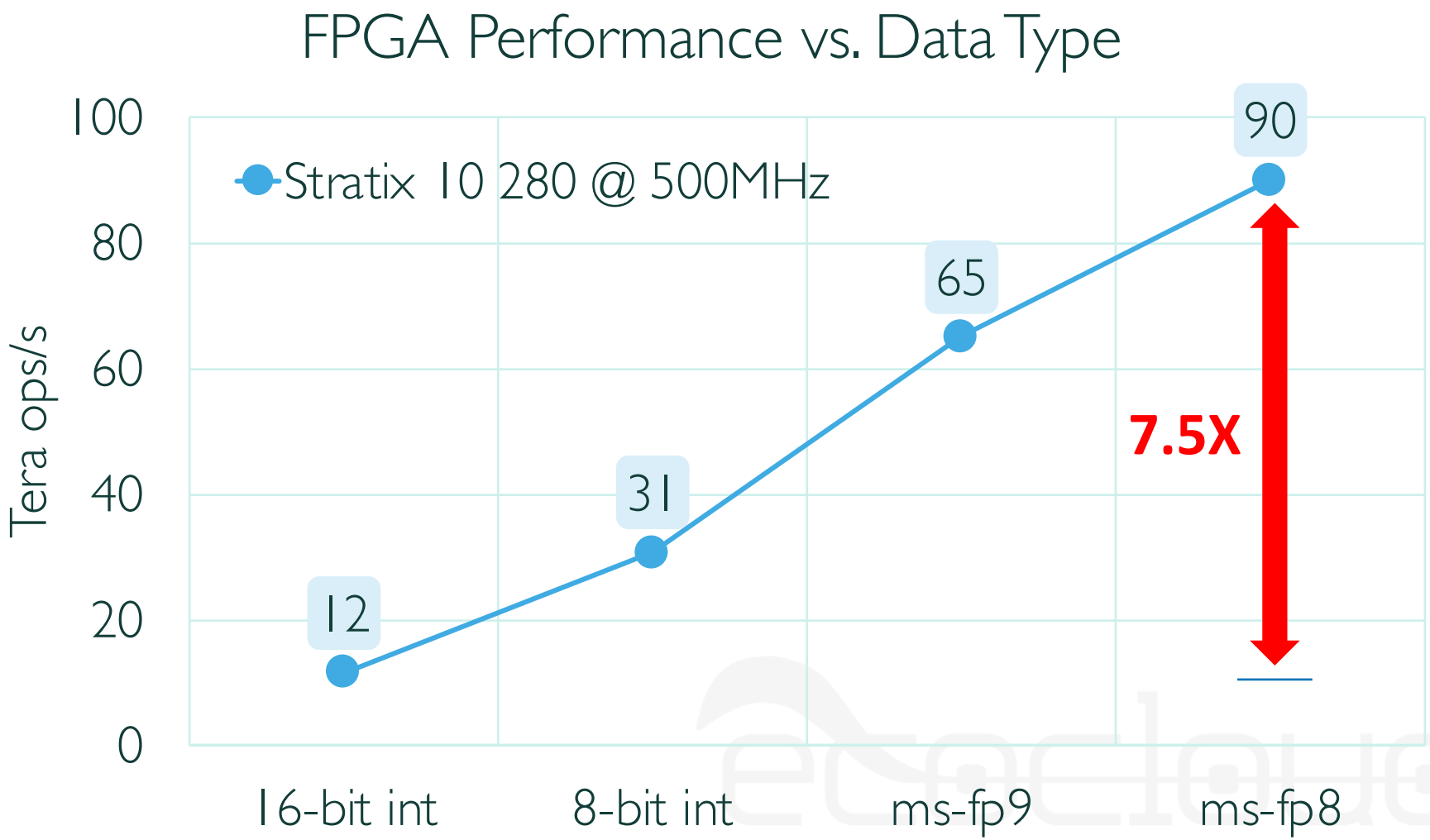
- Much redundancy in data/arithmetic
- Output quality not accuracy or error

Exploit in

- Processing, communication, storage

Arithmetic in Deep Learning (Microsoft Brainwave)

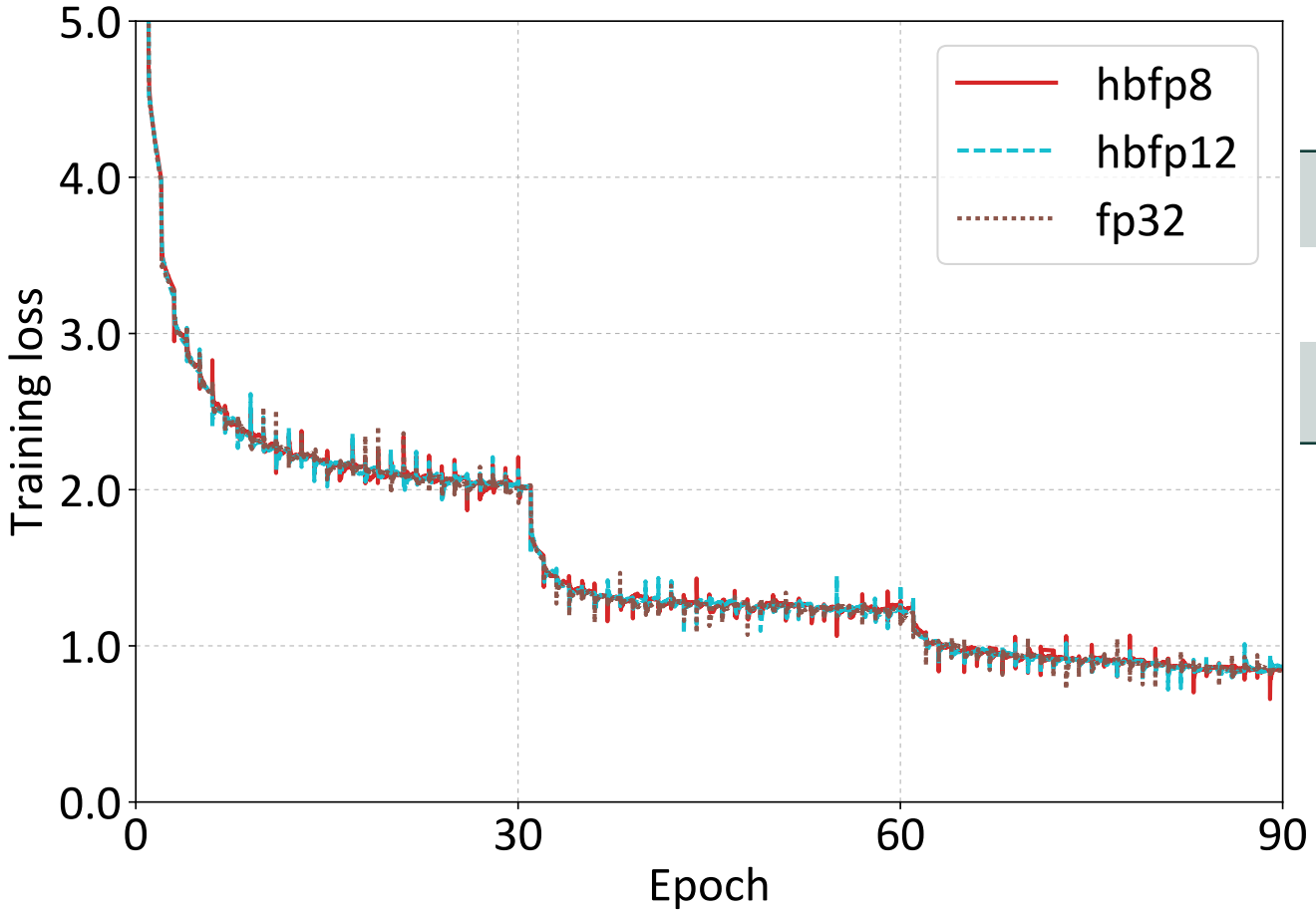
ecocloud
an EPFL research center



HBFP (Block FP) vs. FP32

ecocloud
an EPFL research center

Resnet-50 on ImageNet

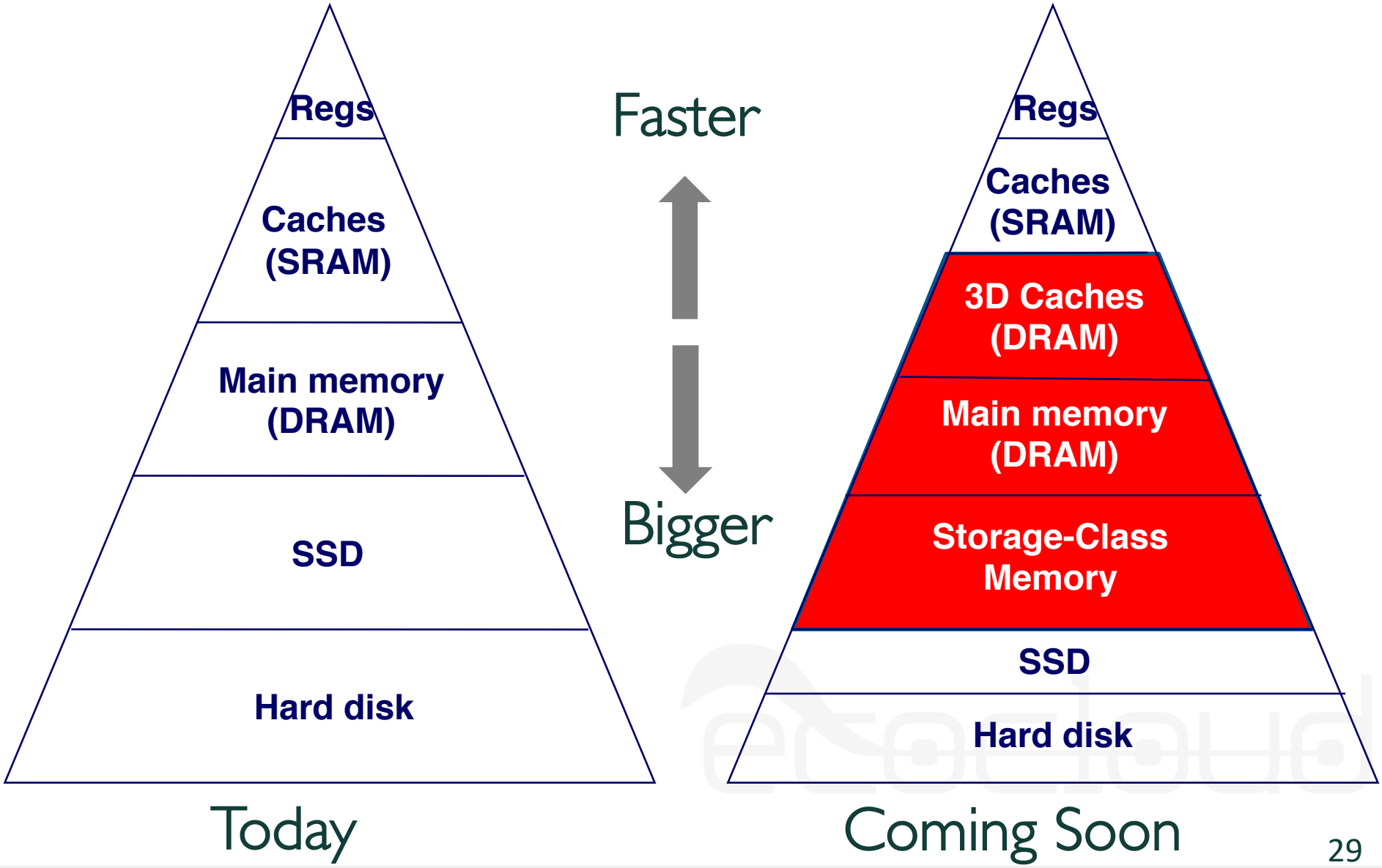


Config.	Top-1 Error (%)
HBFP8	23.88
HBFP12	23.58
FP32	23.64

Medium

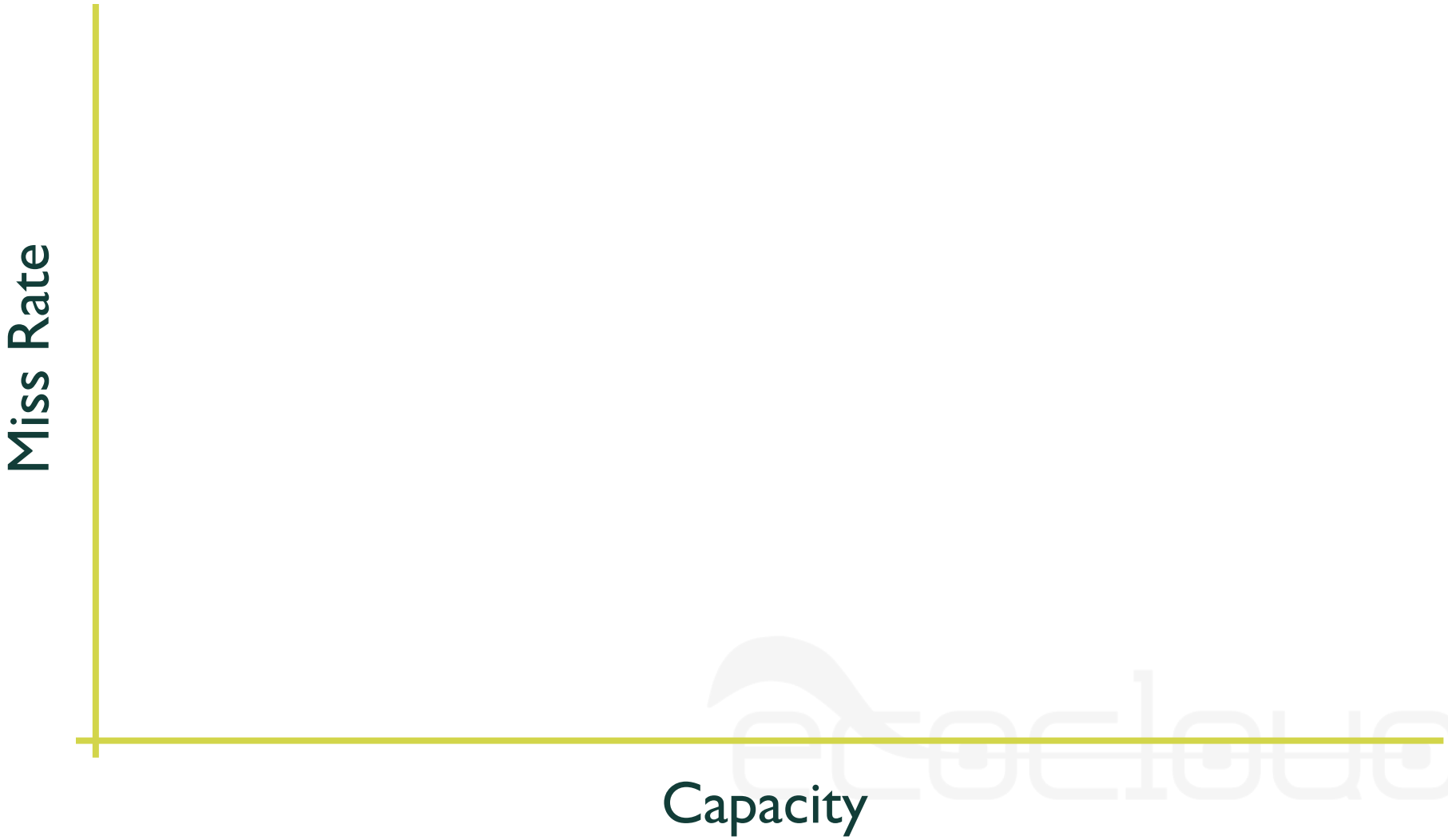
FP32 performance with 8-bit logic [NeurIPS'18]

Memory Hierarchy



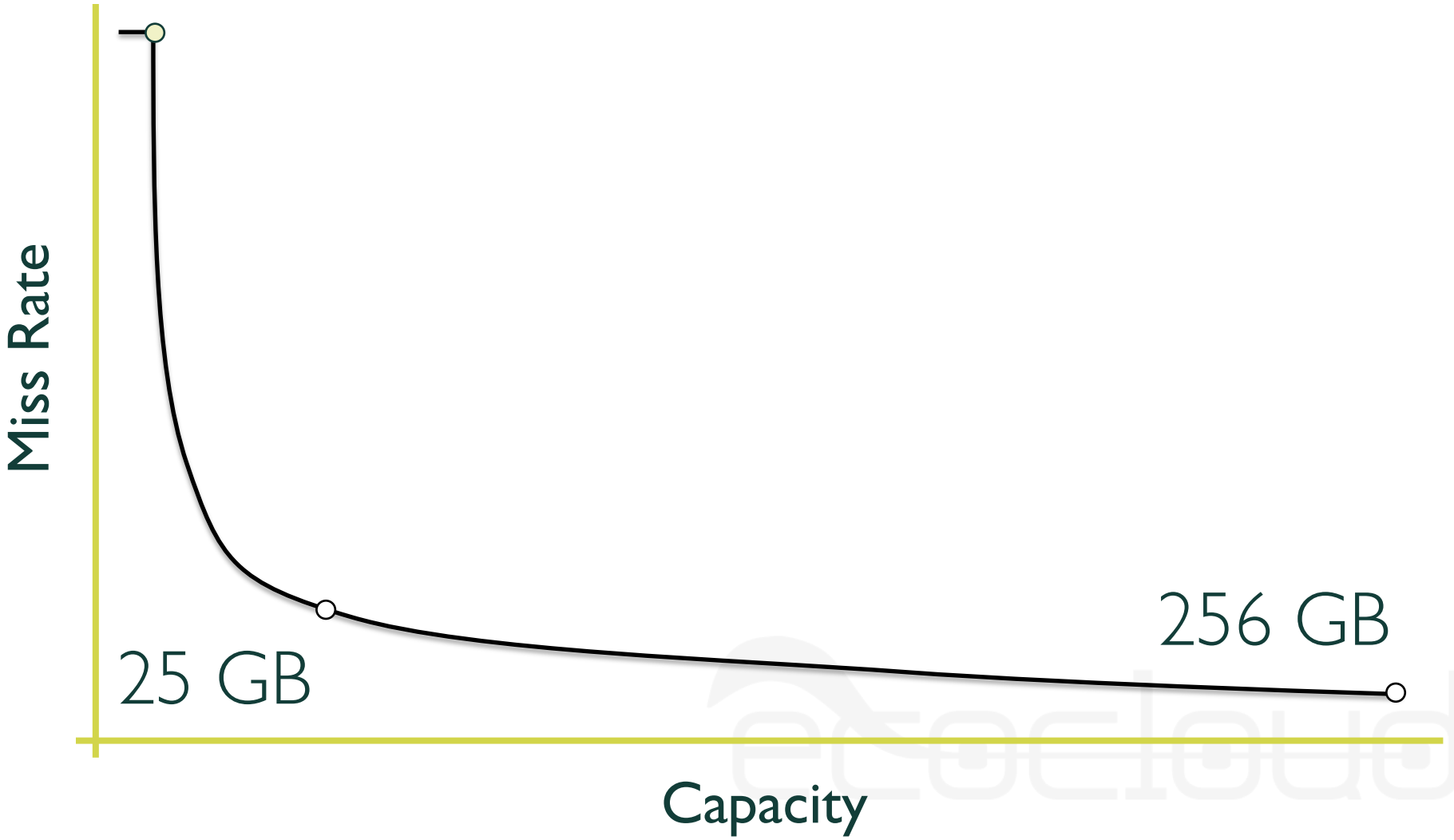
Capacity/Miss Rate 101

ecocloud
an EPFL research center

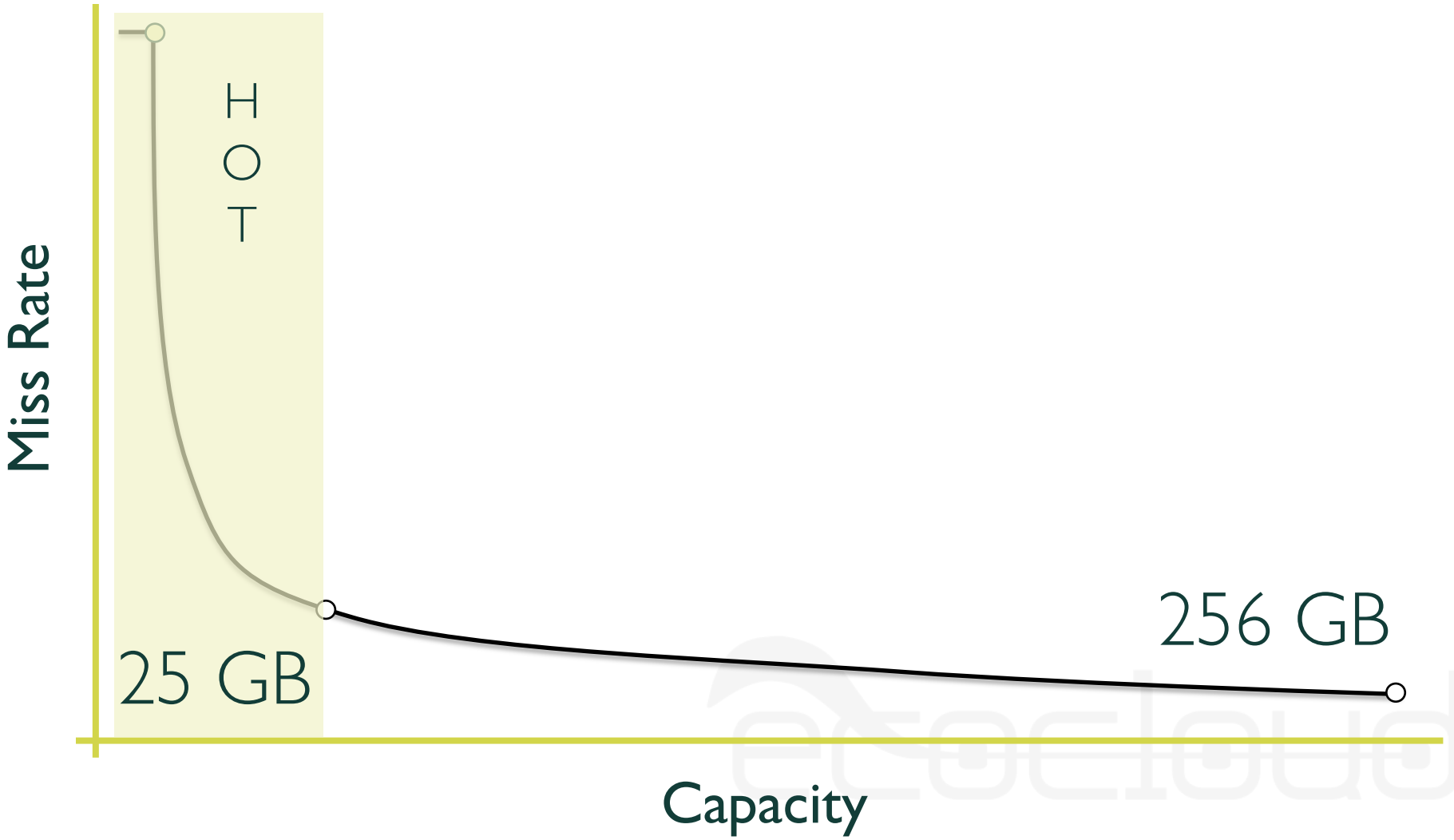


Capacity/Miss Rate 101

ecocloud
an EPFL research center

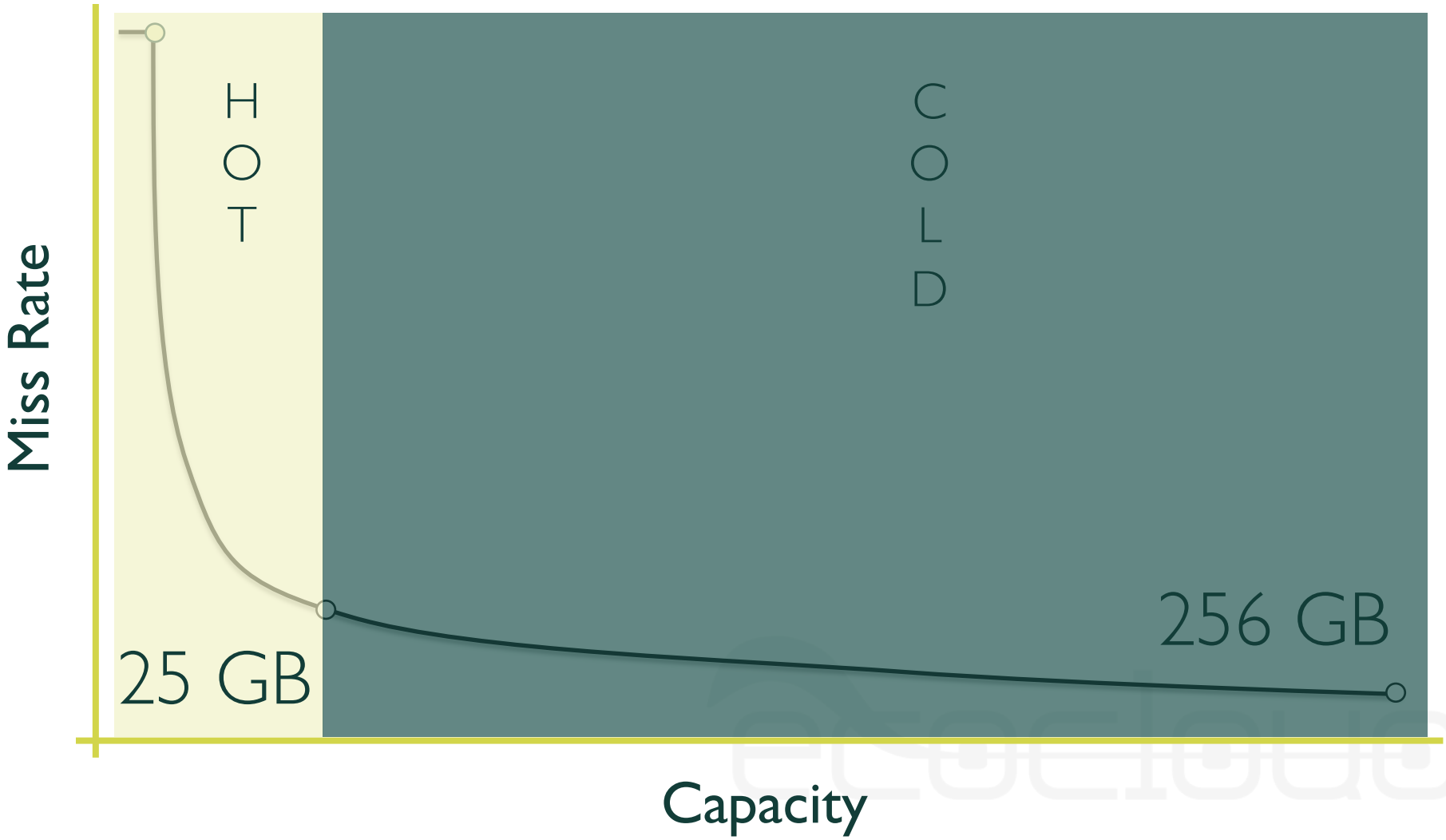


Capacity/Miss Rate 101



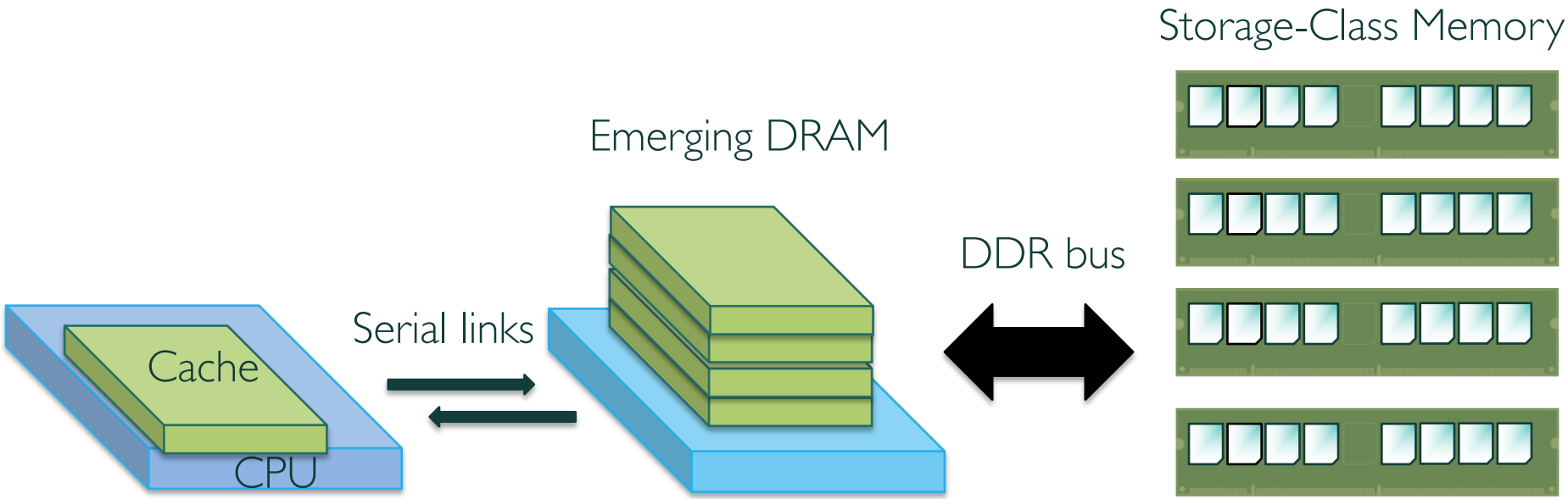
Capacity/Miss Rate 101

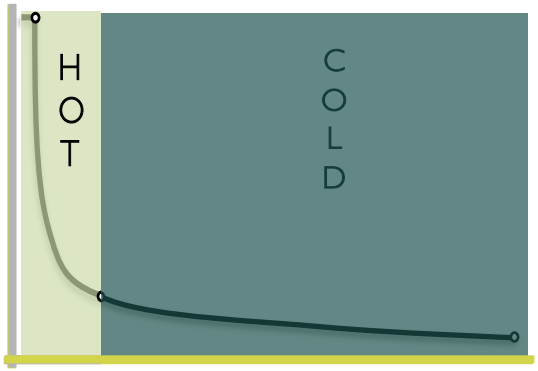
ecocloud
an EPFL research center



Emerging Hierarchies

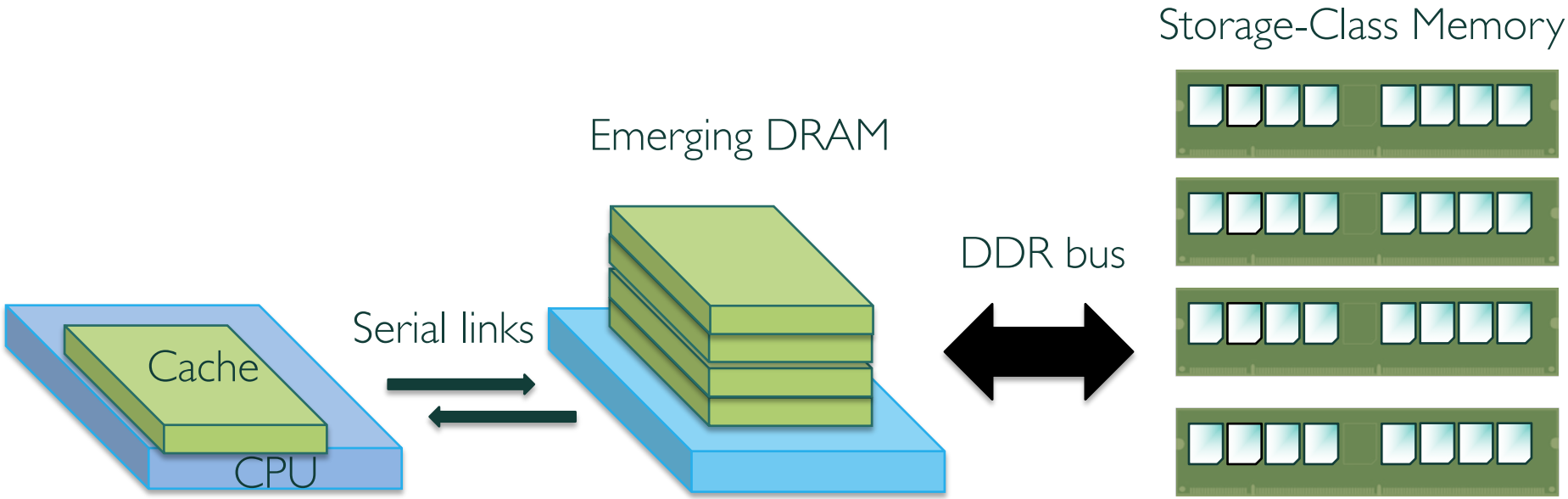
ecocloud
an EPFL research center





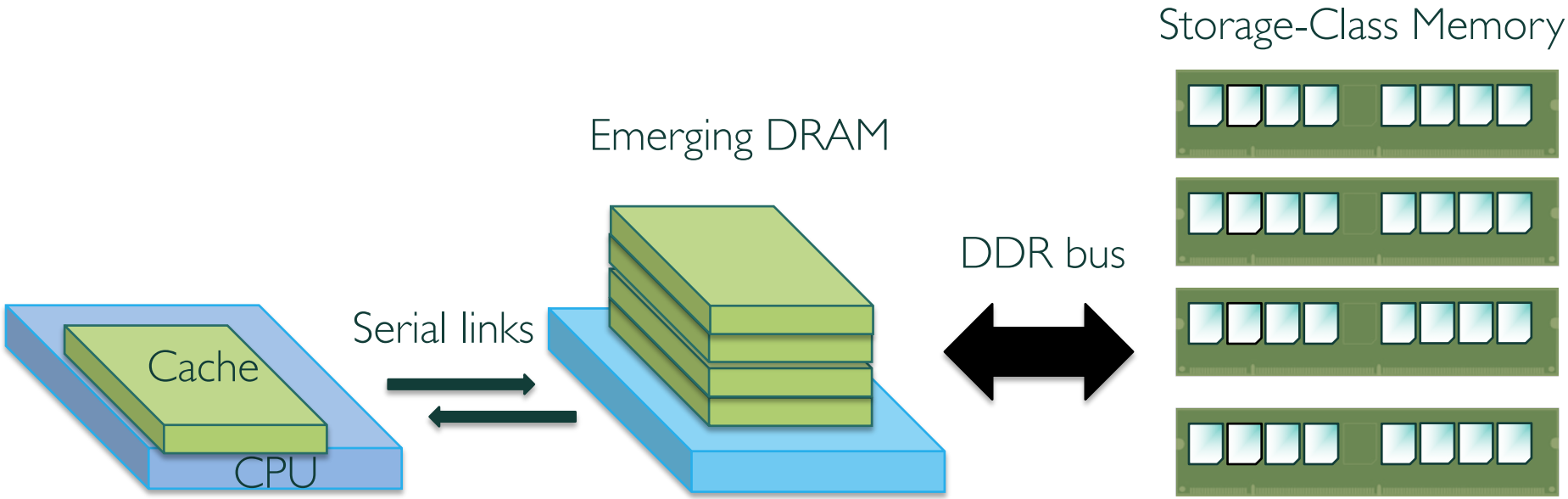
Emerging Hierarchies

ecocloud
an EPFL research center



Emerging Hierarchies

ecocloud
an EPFL research center



Near-Memory Processing

ecocloud
an EPFL research center

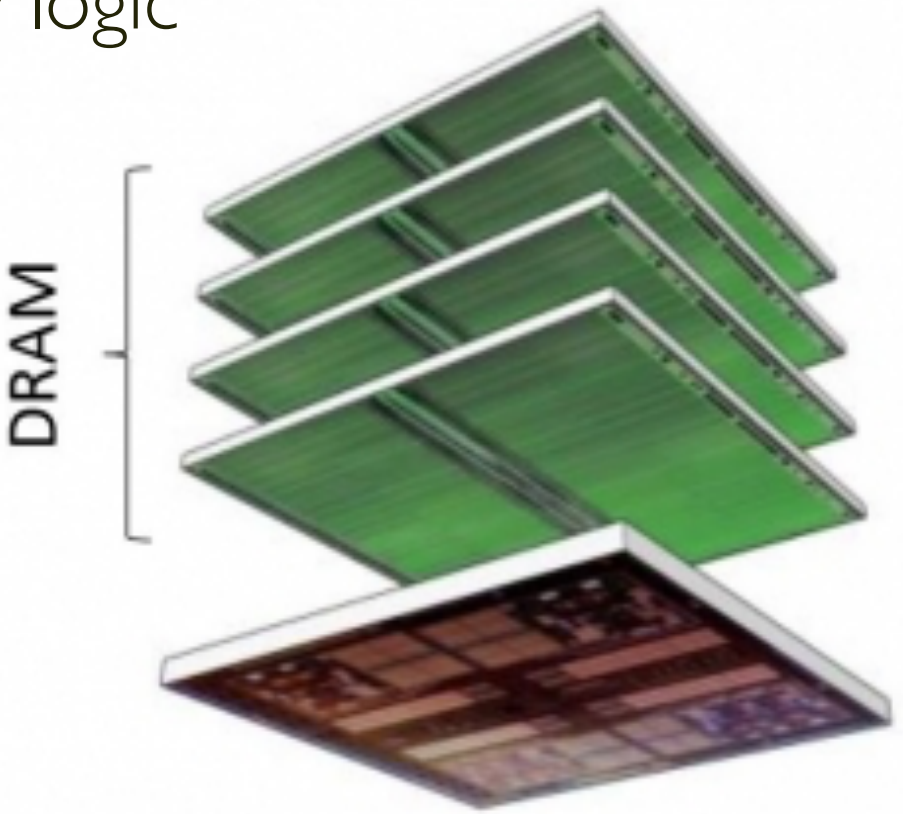
A stack of DRAM w/ nearby logic

- Minimize data movement
- Massive internal bandwidth

Limitations:

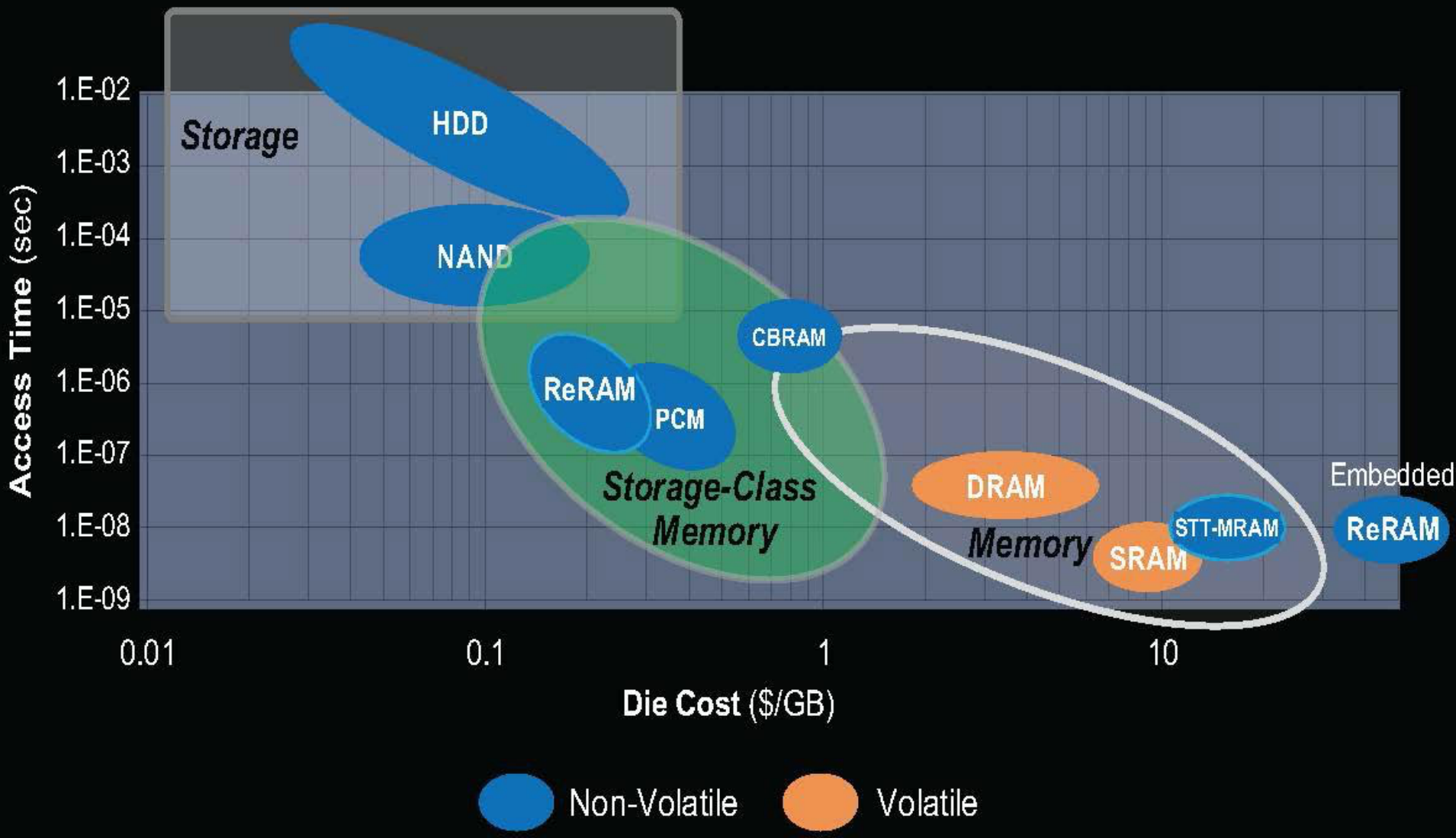
- A few layers of DRAM
- Logic power/thermals (3D)
- Thermals ok for HBM (2.5D)

[source: AMD]



Opportunities for algorithm/hardware co-design

Memory & Storage Hierarchy



Storage-Class Memory

ecocloud
an EPFL research center

Persistence

- 100's of nanosecond vs. microsecond
- Implications for logging & networks

Disparity between reads/writes

- Can read at memory speed
- Writes must be batched/are slow
- Writes consume more power
- DRAM cache can help [MemSys'18]

Online SSD

ecocloud
an EPFL research center

SSD is treated as storage

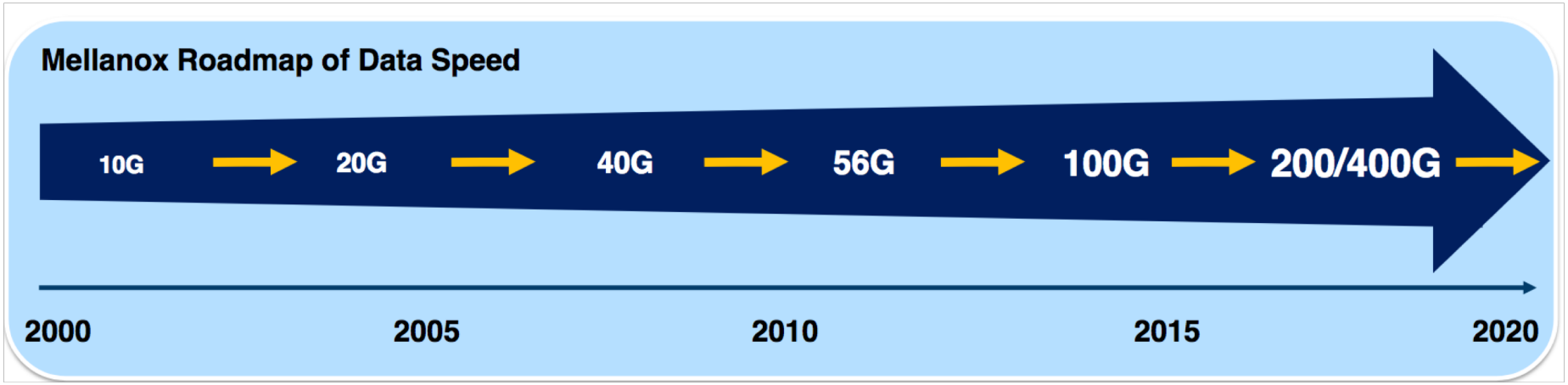
- Online data in DRAM
- But, DRAM costs dominate, slow scalability

Online services:

- Roundtrip tail latency 100's ms
- SSD access is in 50 us (1000x faster)
- SSD is 50x cheaper

Technology to bring SSD online!

Networks



Network stacks/interfaces are a bottleneck:

- Logic growing at 17%/year, network at 20%/year!
- µServices emerging
- RPC stacks, scheduling/dispatch, data transformation, ….

Key challenges:

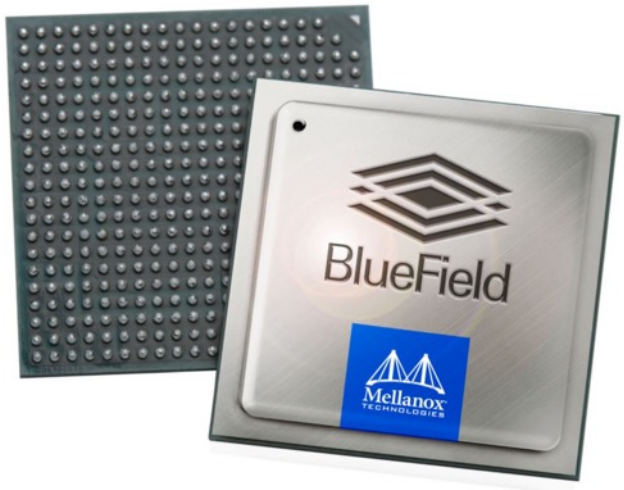
- Abstractions for control/data planes
- Co-design of network stacks

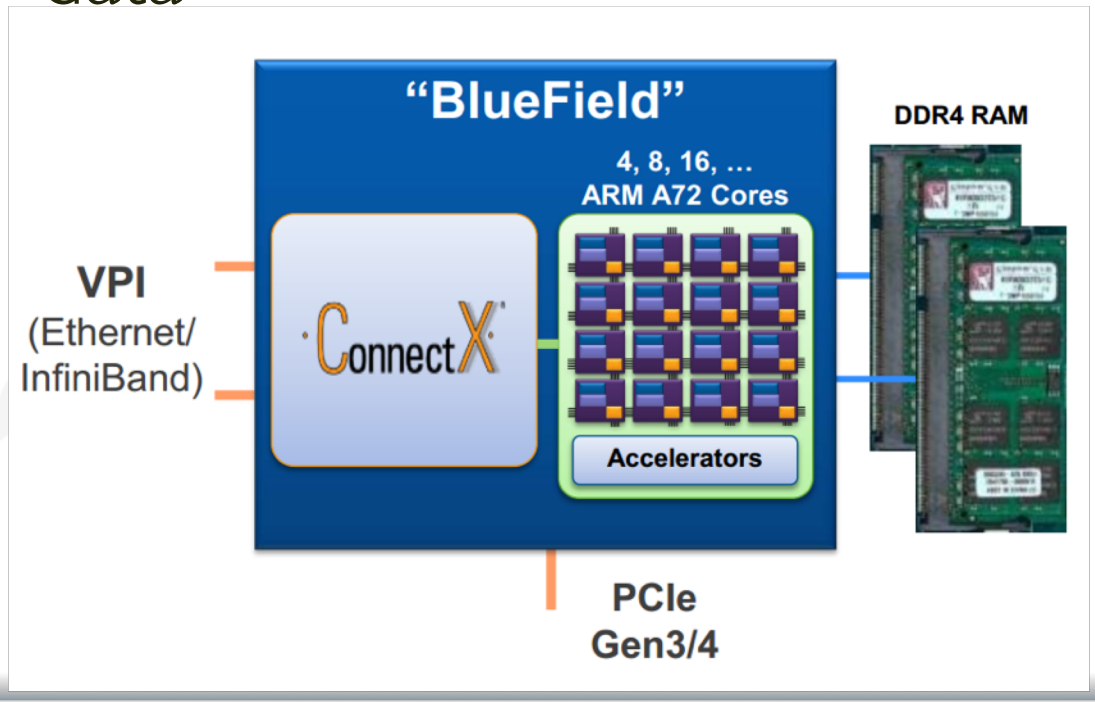
Near-Network Processing: Nvidia BlueField

ecocloud
an EPFL research center

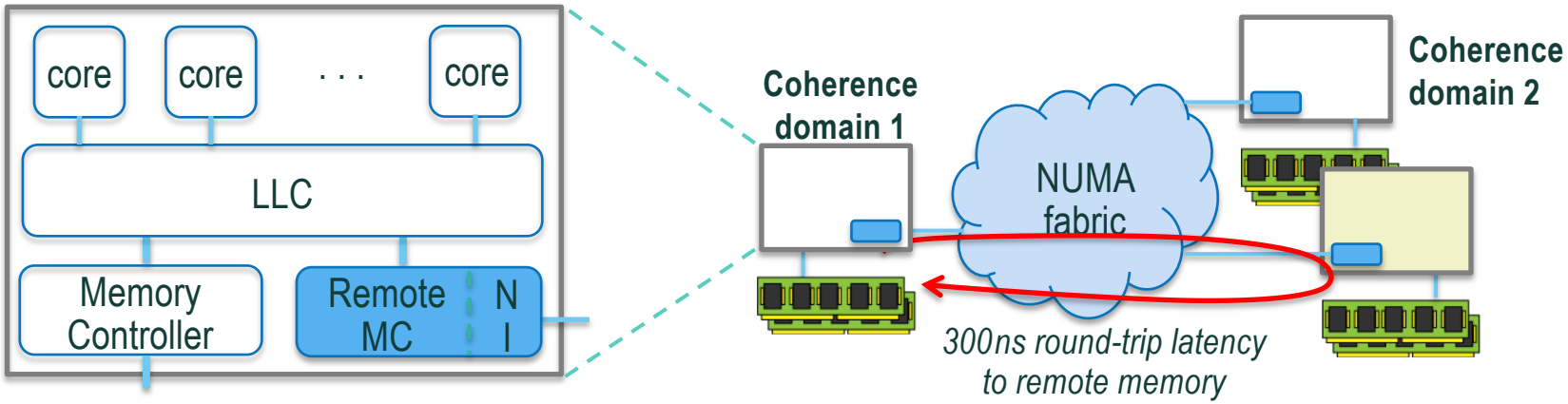
Network-interface integrated manycore

- Up to 32 cores w/ 0.5 TB of DRAM
- Can host an in-memory object store
- RPC over in-memory data





Scale-Out NUMA [ASPLOS'14,ISCA'15, MICRO'16]



Microsoft

HUAWEI

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 [Daglis' thesis]

Summary

ecocloud
an EPFL research center

Server design centered around data

- In-memory services offered over the network

Witnessing end of Moore's Law

- Emerging heterogeneous logic + memory

Future servers nodes:

- Logic & memory with multiple network access points
- Tool chains to go from DSL's ➔ accelerators

Integrate + Specialize + Approximate

Thank You!

ecocloud
an EPFL research center

For more information please visit us at

ecocloud.ch