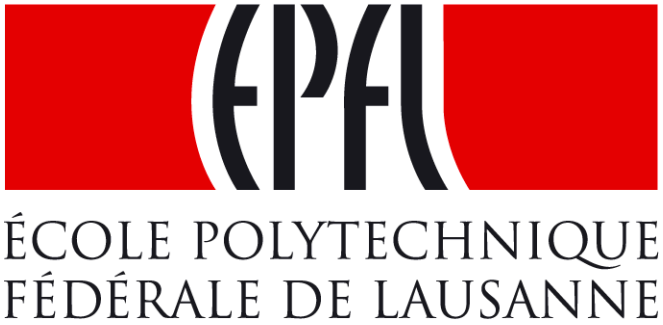
# Server Architecture for the Post-Moore Era

Babak Falsafi

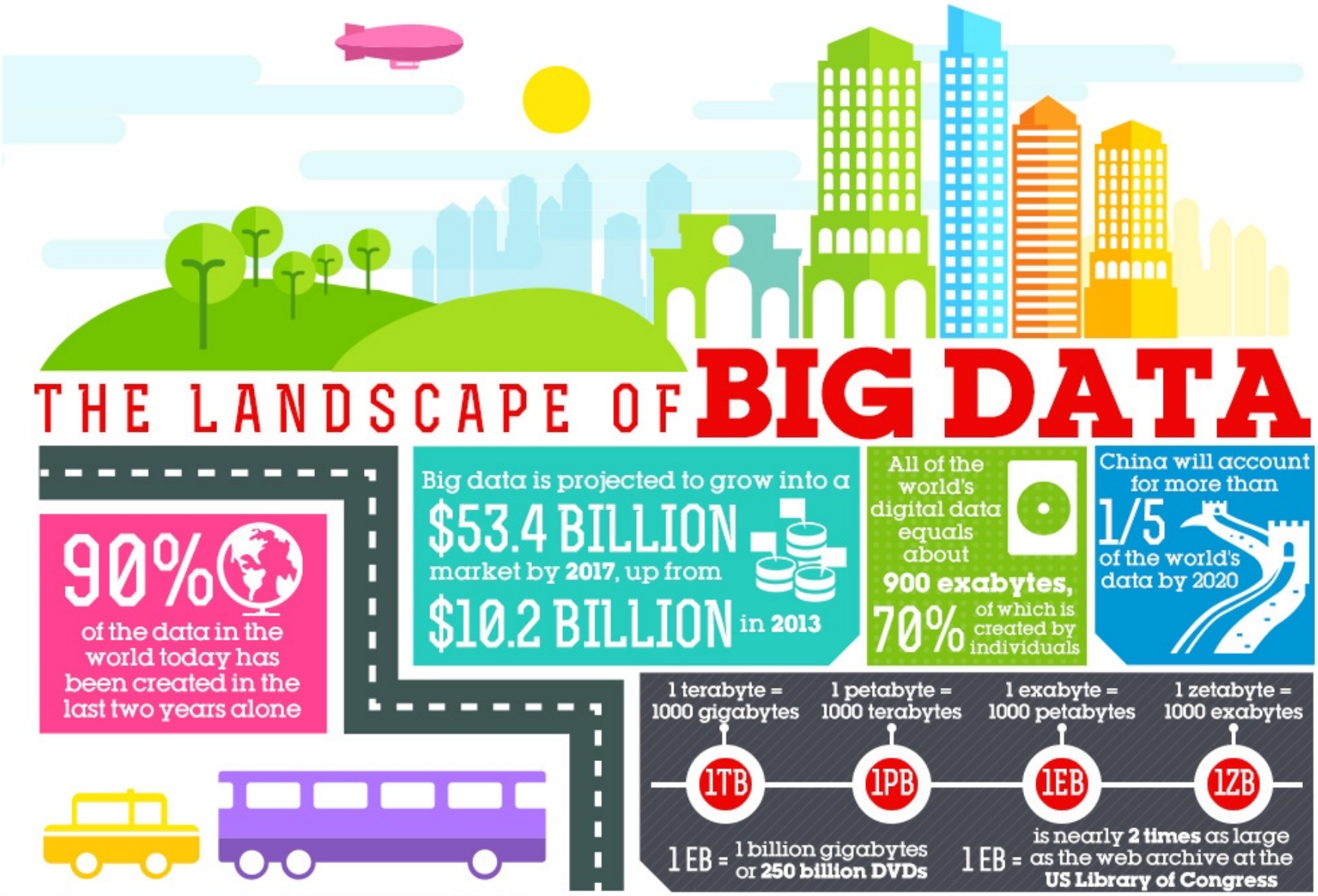
Director, EcoCloud

ecocloud.ch

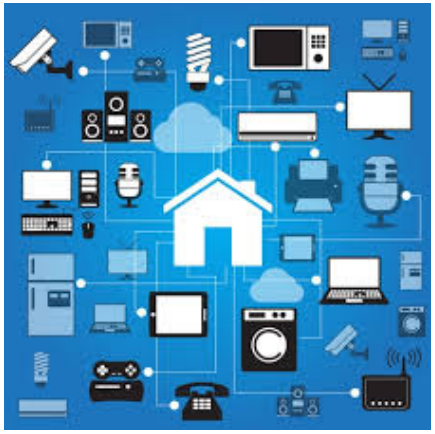



# Data Economics

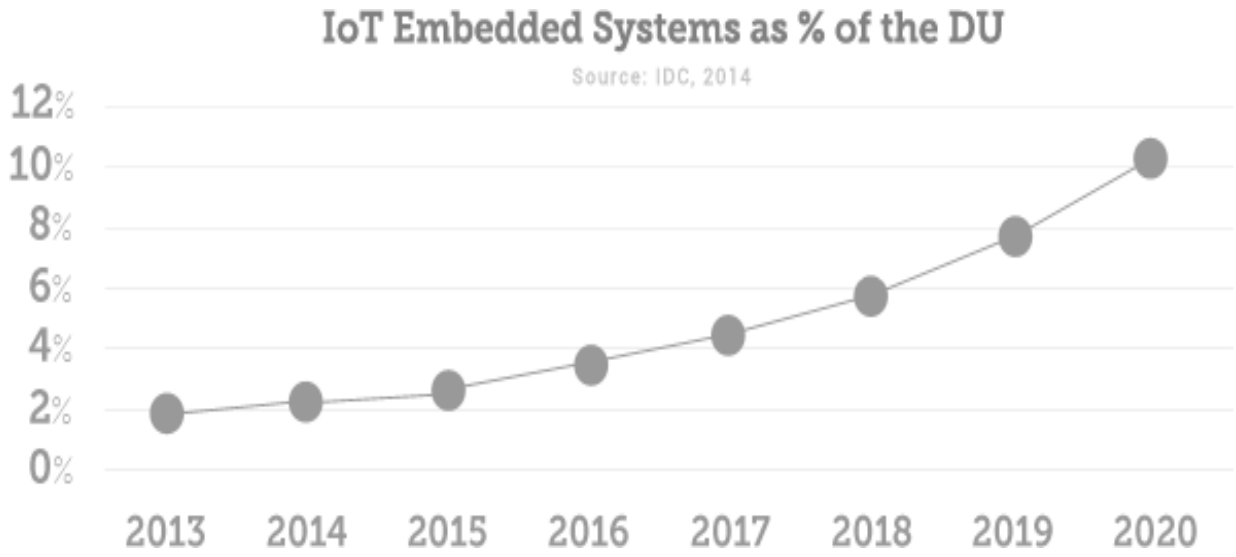
ecocloud
an EPFL research center



# Internet-of-Things (IoT): Data in Flight



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

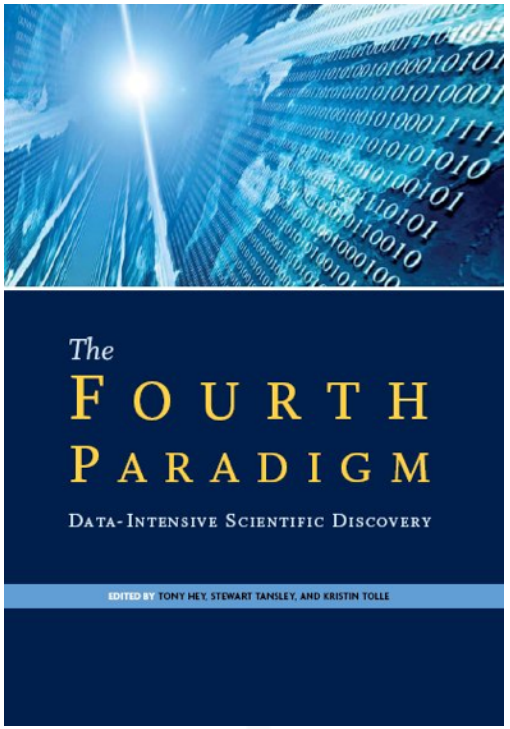
# Data Shaping All Science & Technology

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

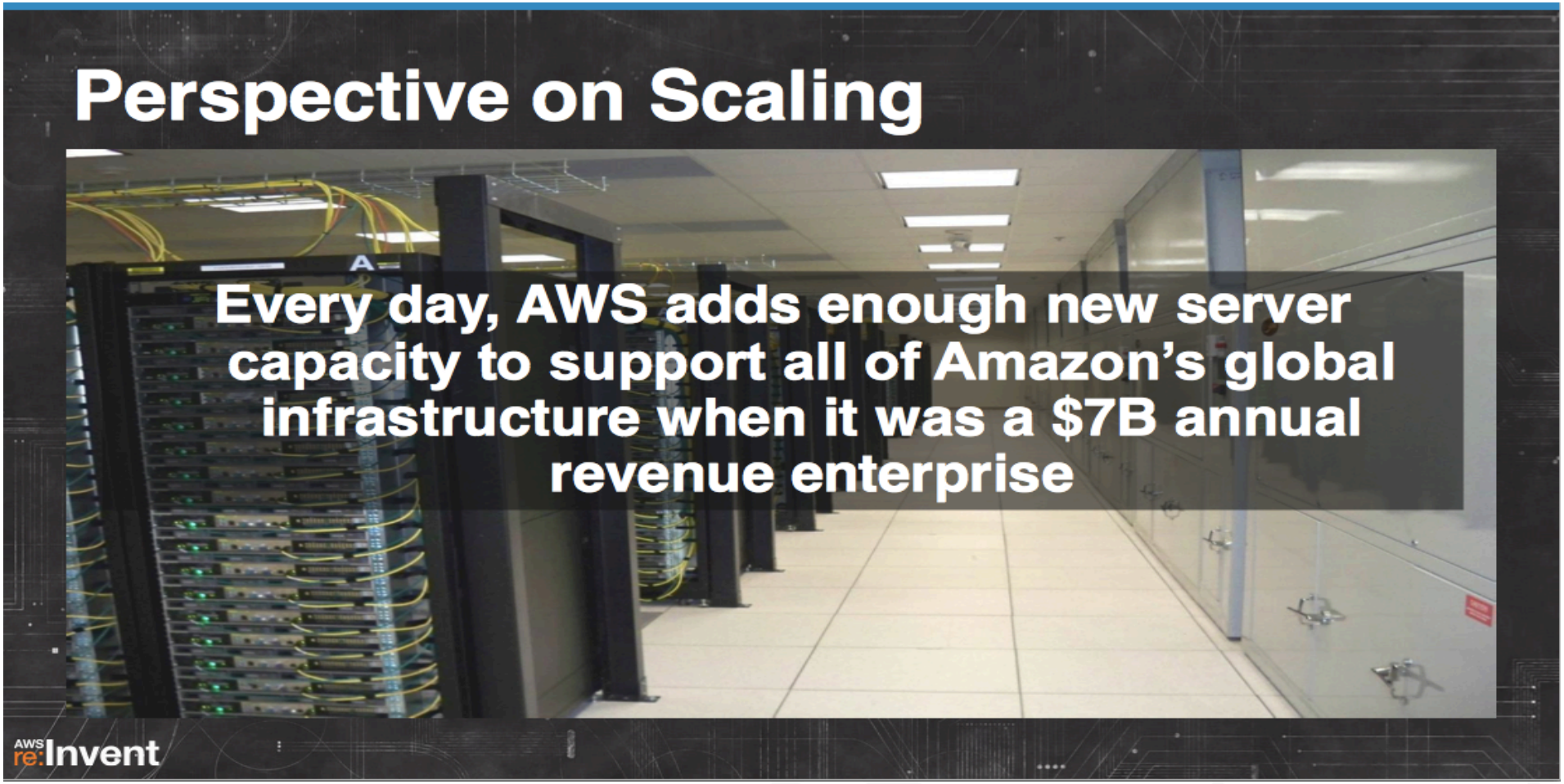



Data-centric science key for innovation-based economies!

*Source: James Hamilton, 2014*

*mvdirona.com/jrh/TalksAndPapers/JamesHamilton_Reinvent20131115.pdf*

ecocloud
an EPFL research center

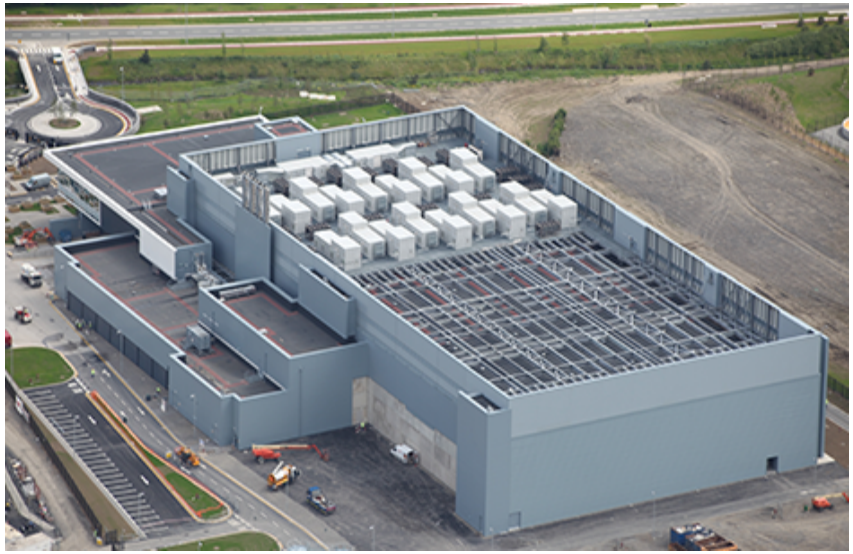



**Daily** IT growth in 2014 = All of AWS in 2004!

# 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**

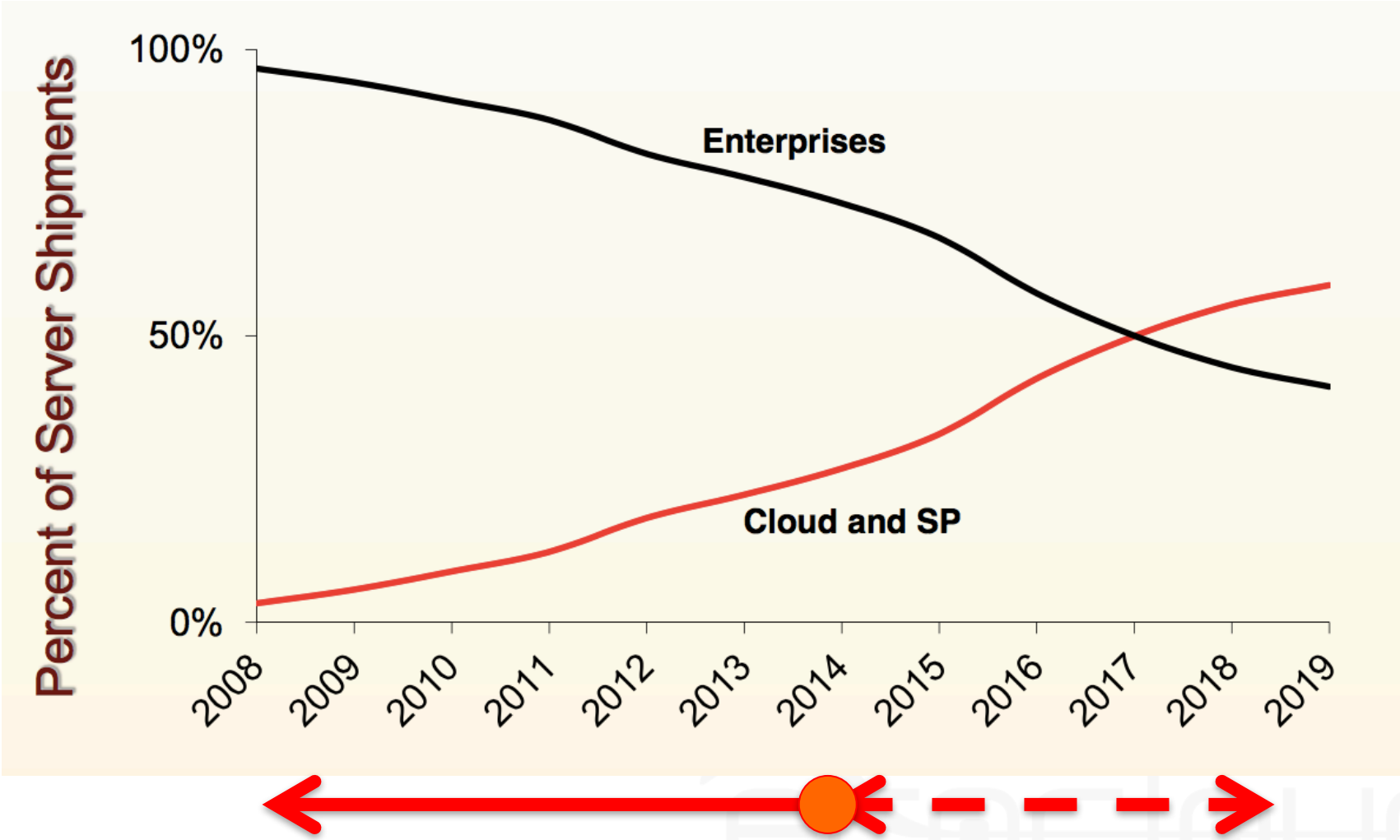

Supercomputing

≠

Cloud Computing

# Cloud Taking Over Enterprise

ecocloud
an EPFL research center

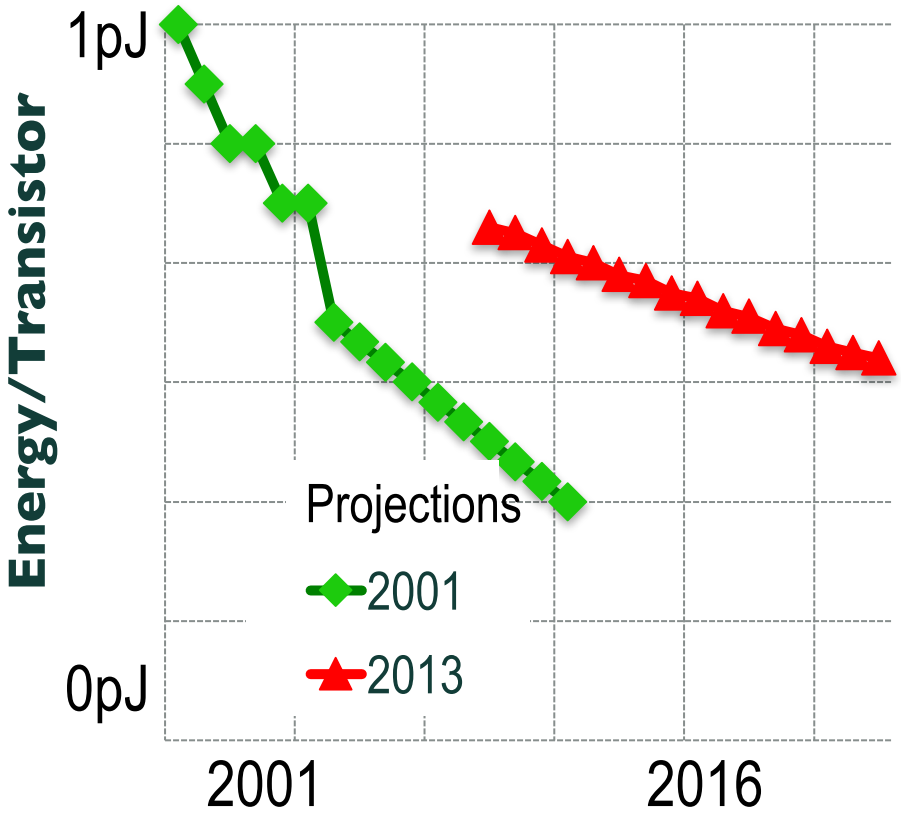



*Source: Dell 'Oro 2Q15*

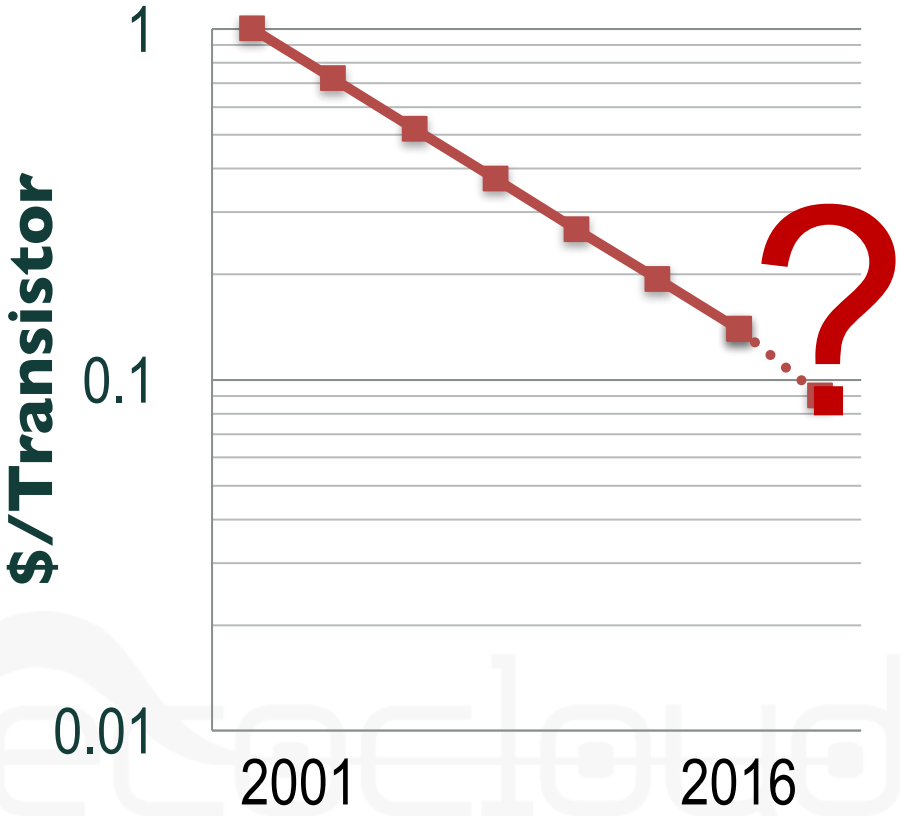
# But…., silicon is running out of steam!

ecocloud
an EPFL research center

## Silicon efficiency is dead (long live)




## Moore's law dying



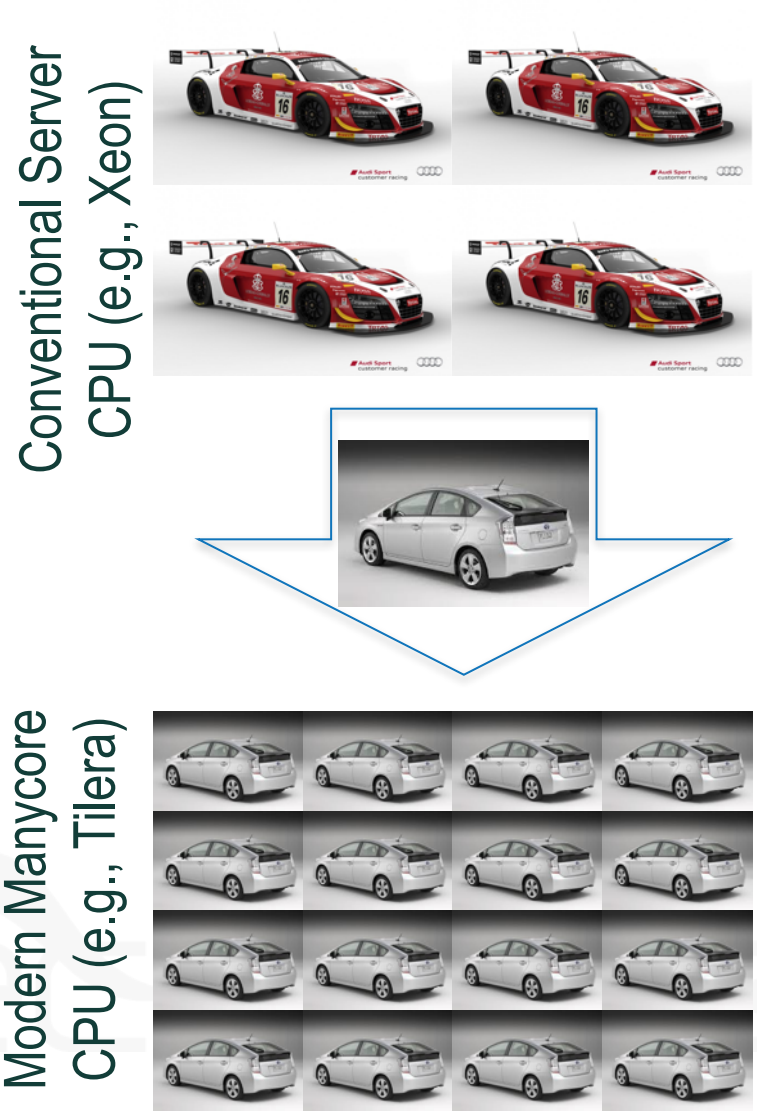
# Manycore Accelerators

ecocloud
an EPFL research center

With voltages leveling:

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

  fewer joules/op



# Massively Parallel Cores

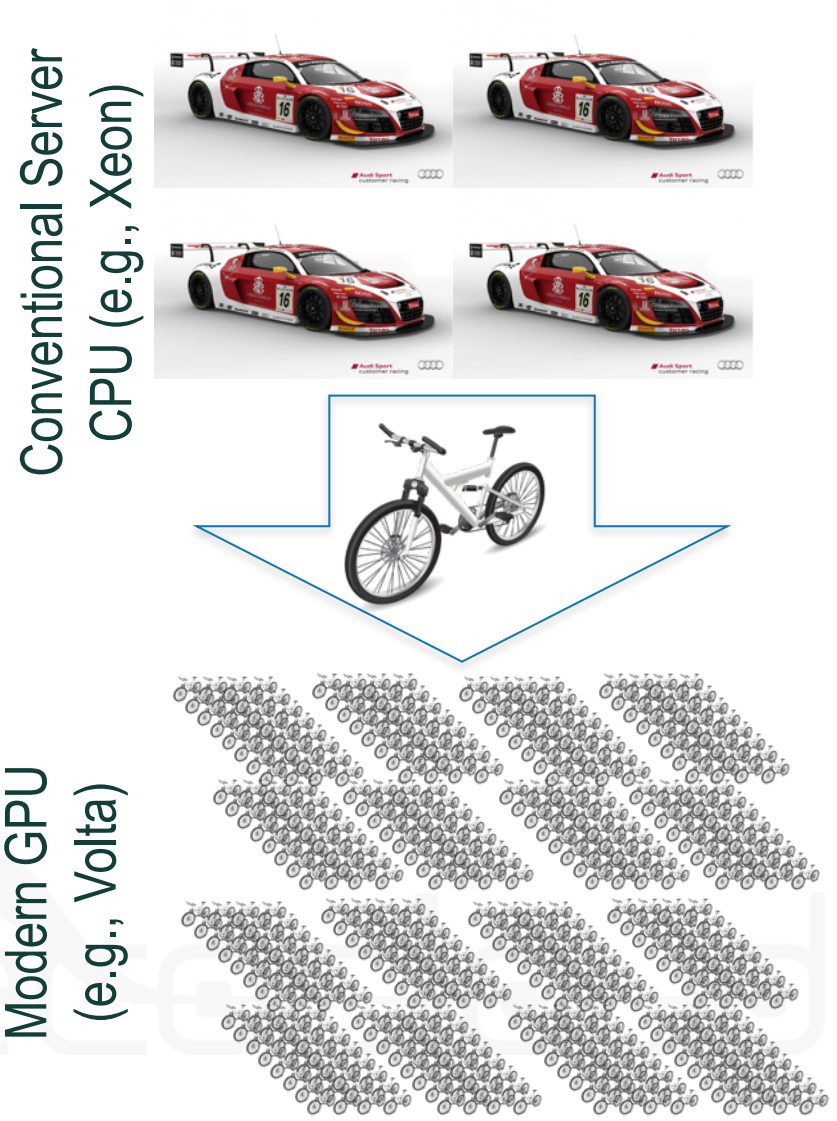
ecocloud
an EPFL research center

- Data parallel
- Higher memory b/w

Super simple cores

- Shared front end
- 10x slower clocks

Great for dense parallel computation

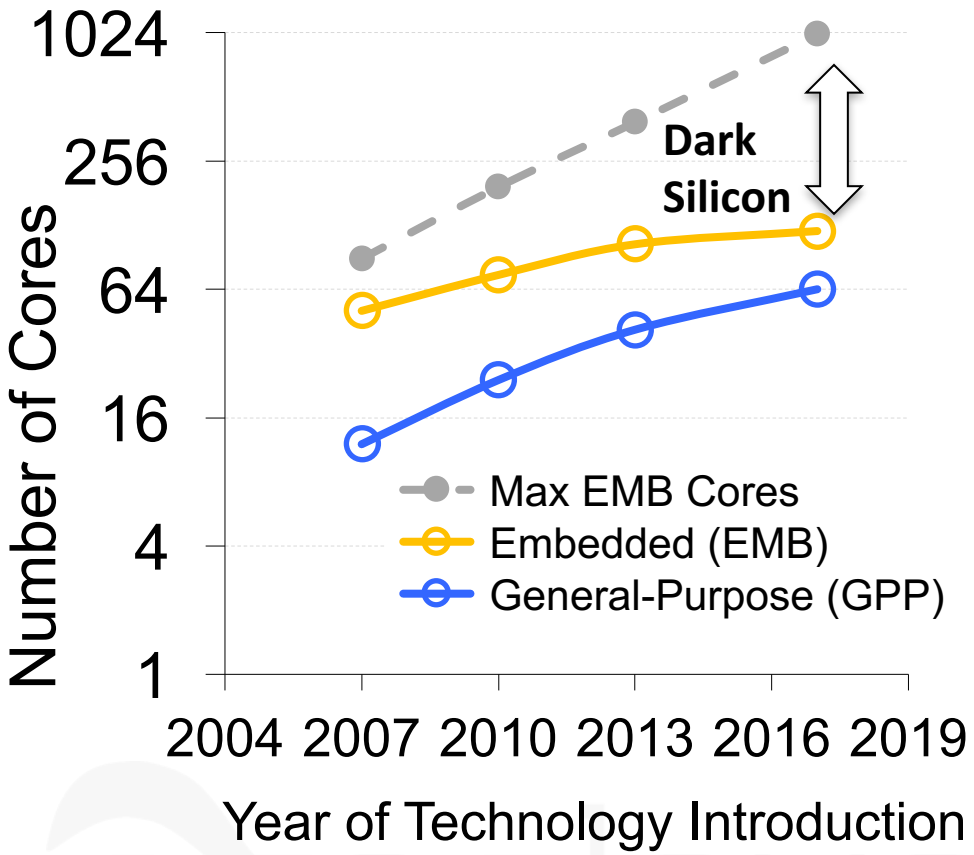


# Parallelism Alone Can't Help

ecocloud
an EPFL research center

**ISA** opportunities

- **I**ntegration
  - Less energy moving
  - Closer to memory
- **S**pecialization
  - Customize work
  - Less work/computation
- **A**pproximation
  - Adjust precision




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

# ecocloud

## Center to bring efficiency to data

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

## Mission:

- Energy-efficient data-centric IT
- From algorithms to infrastructure
- Maximizing value for data

# Our Vision:
# Holistic Optimization of Datacenters

ecocloud
an EPFL research center

From algorithms to infrastructure:

- Cross-layer integration & specialization
- Introspection & resource provisioning

Open technologies!

# Outline

ecocloud
an EPFL research center

- ~~Overview~~
- Memory-Centric Servers
- Near-Memory Processing
- Summary

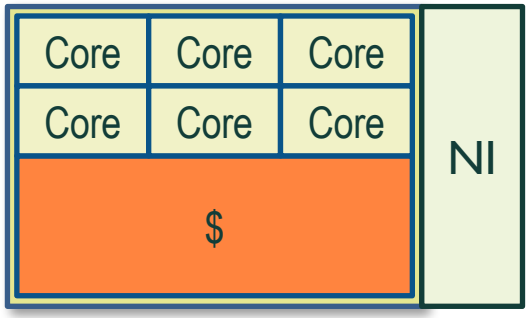
# Scale-Out Datacenters

ecocloud
an EPFL research center

Vast data sharded across servers

Memory-resident workloads

- Necessary for performance
- Major TCO burden



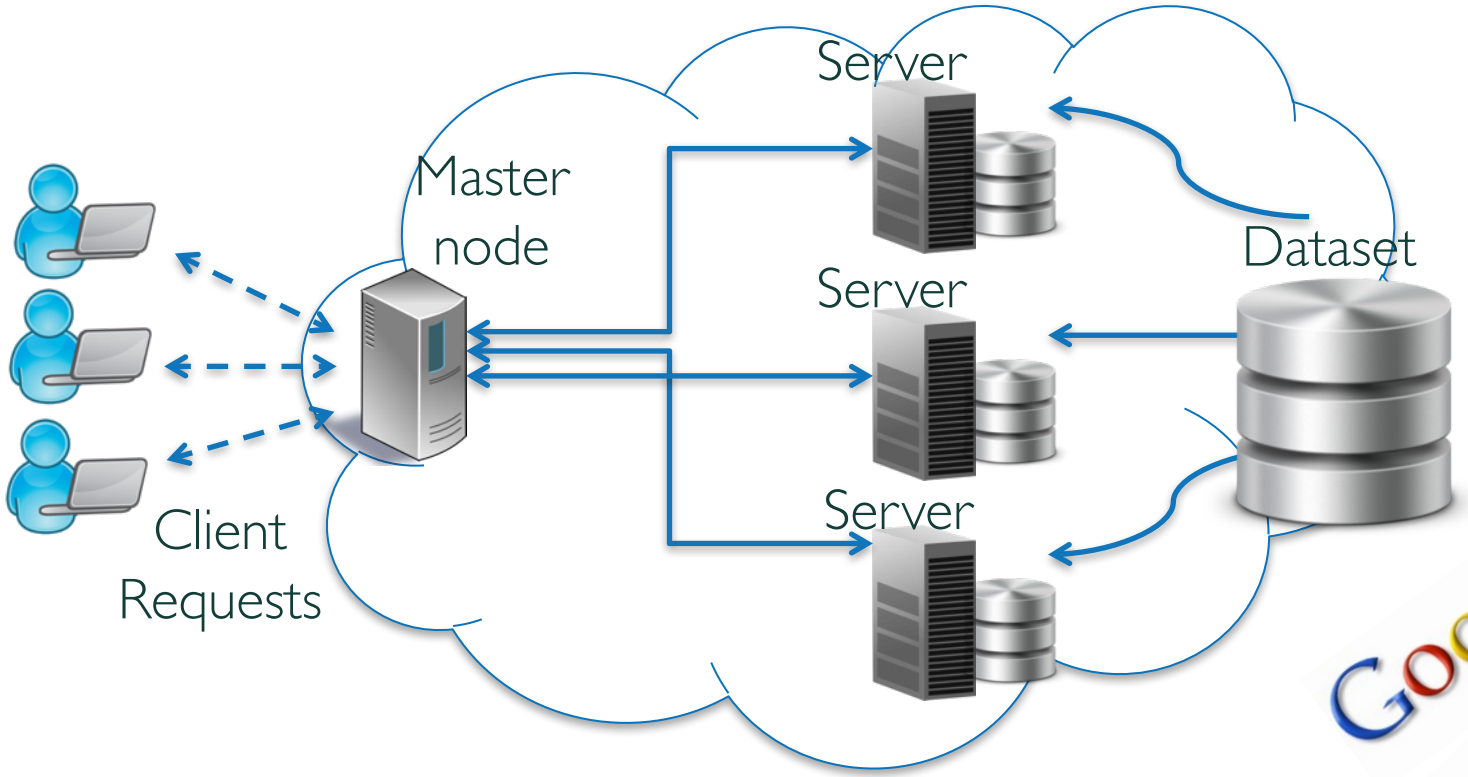

Put memory at the center

- Design system around memory
- Optimize for data services


Servers driven by the DRAM market!

# In-Memory 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

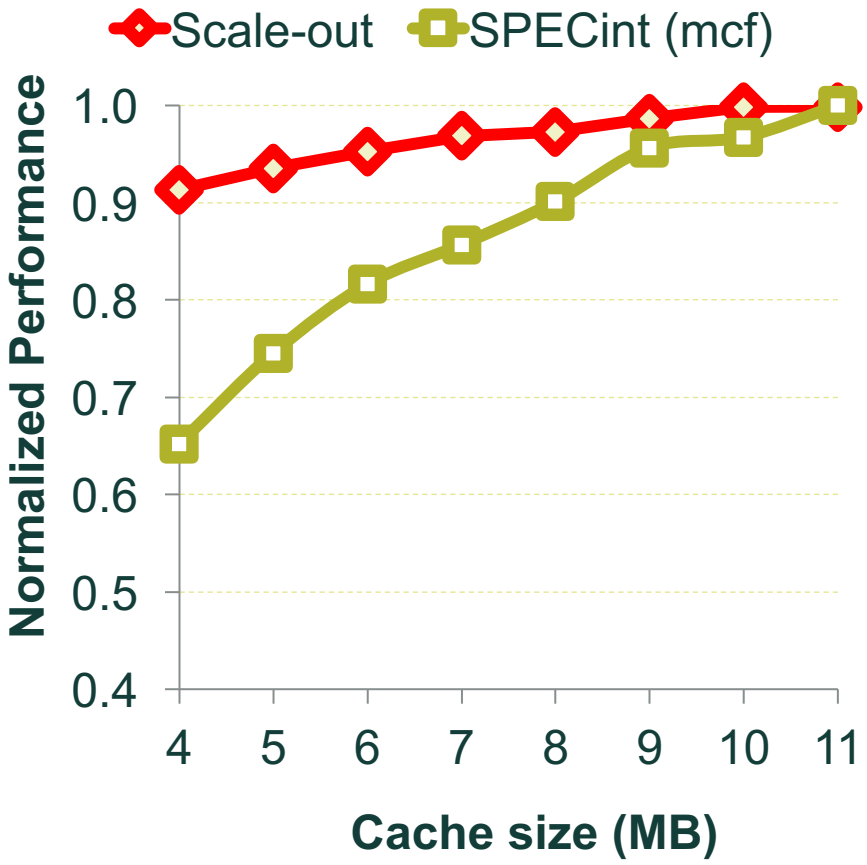

**In-Memory Analytics**
Recommendation System

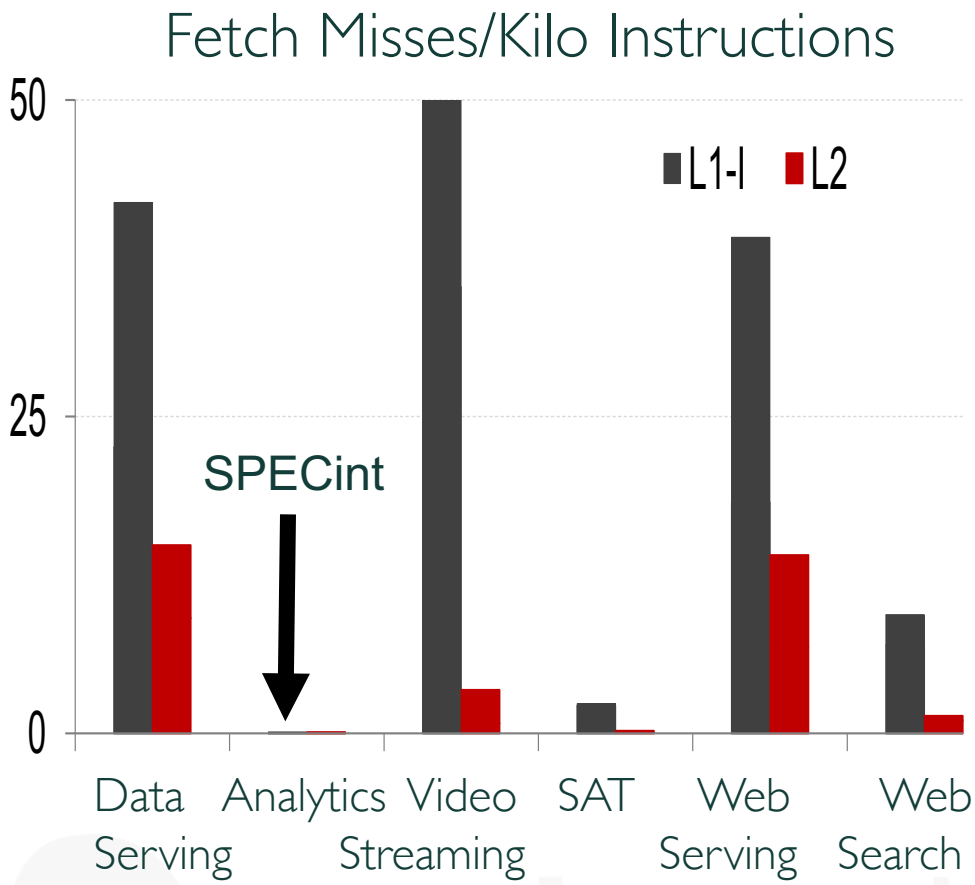
movielens



***Building block for Google PerfKit, EEMBC Big Data!***

# Services are Stuck in Memory

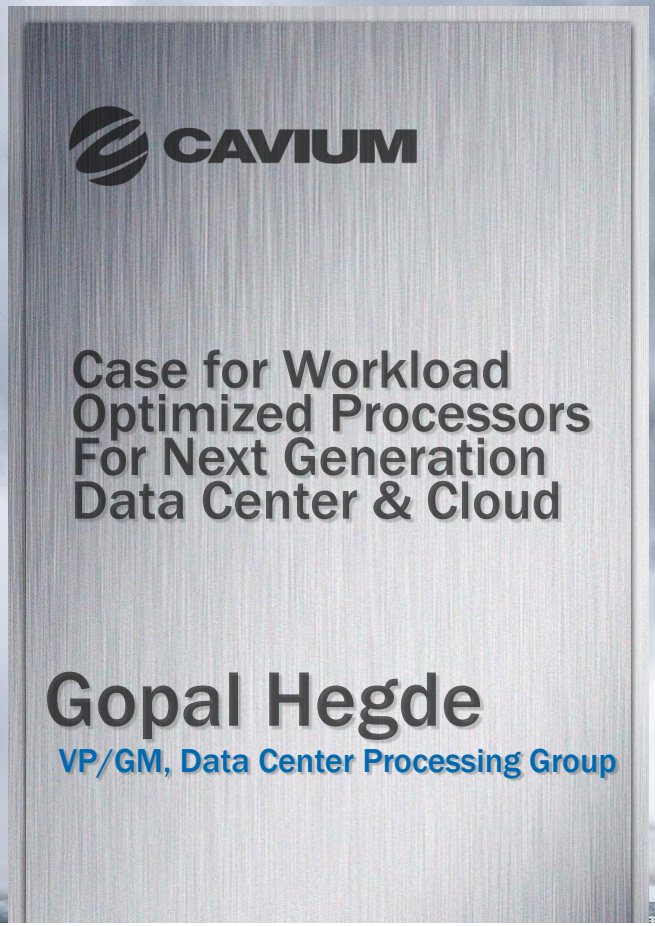



- On-chip memory overprovisioned
- Instruction supply is bottlenecked

# Manycore Accelerator for Data Serving

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

# NOC-Out: NoC for Server Processors

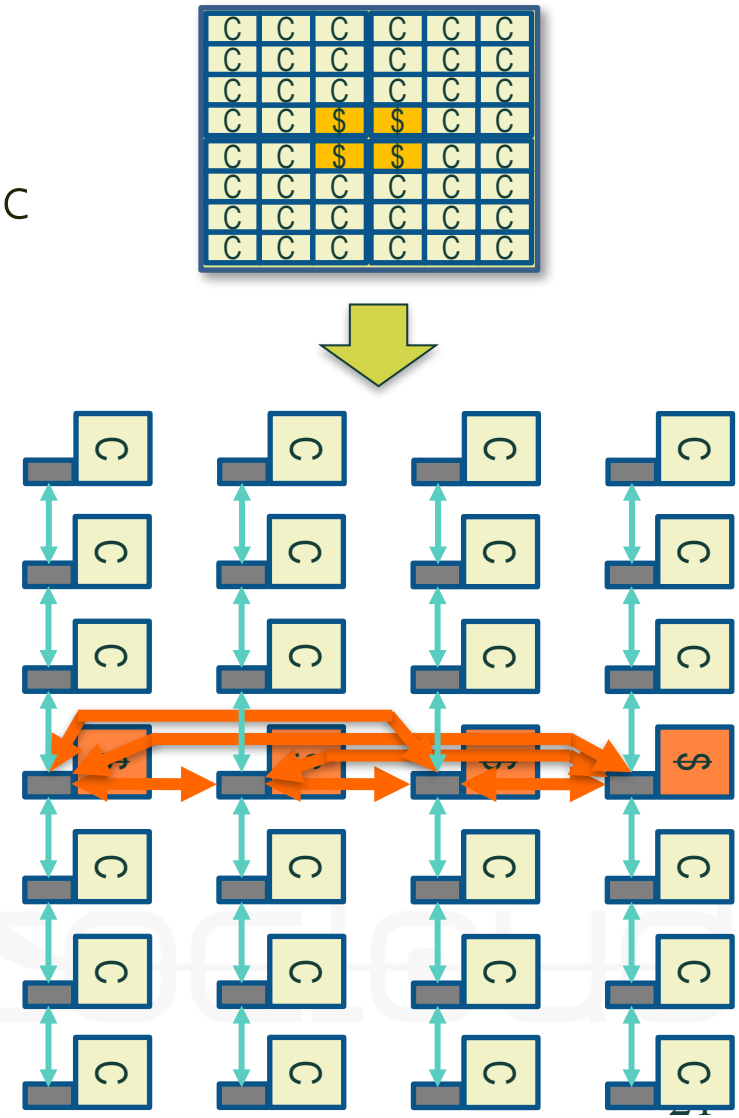
ecocloud
an EPFL research center

Exactly the **opposite** of current NoCs
- Cache coherent
- But, designed for core-to-cache traffic
- Not core-to-core!

*LLC* network:
- Flattened Butterfly (FB) topology

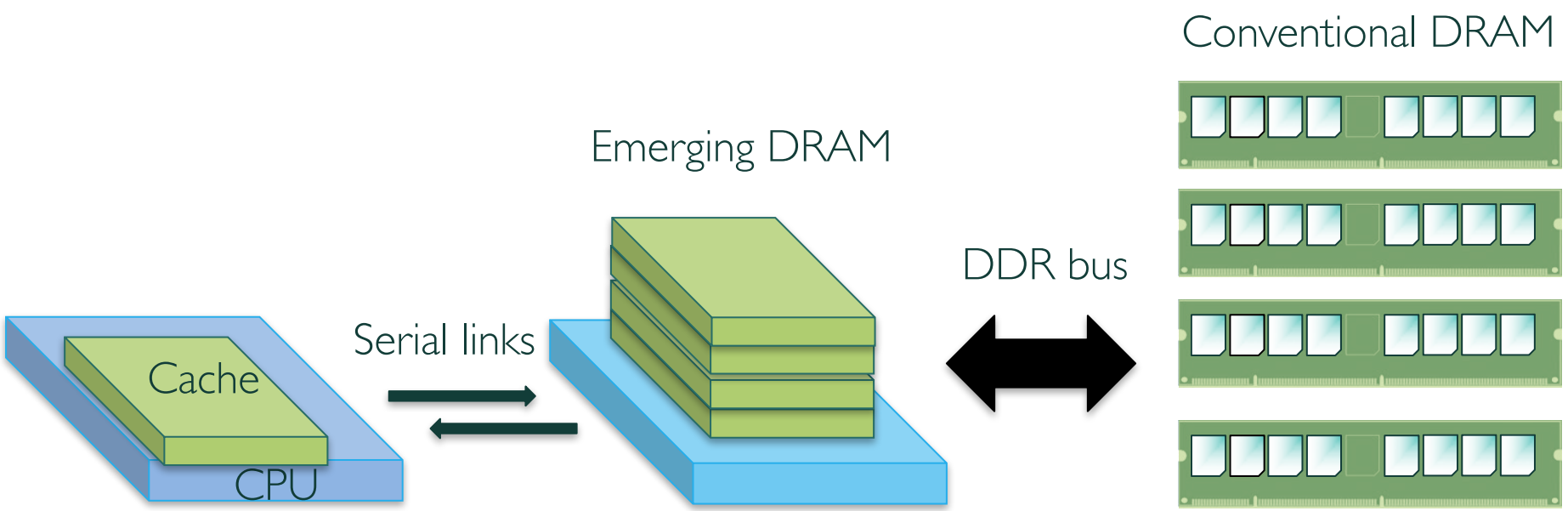
*Request* & *Reply* networks:
- Tree topology
- Limited connectivity for efficiency

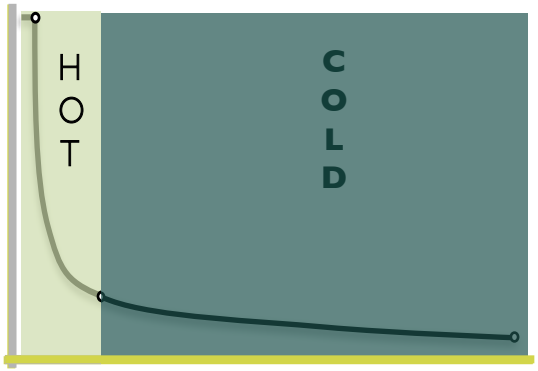
FB's performance at 1/10$^{th}$ cost



# Emerging DRAM as Cache

ecocloud
an EPFL research center



# DRAM Cache with Storage-Class Memory

ecocloud
an EPFL research center

SCM extends DRAM capacity as memory

DRAM cache + SCM:

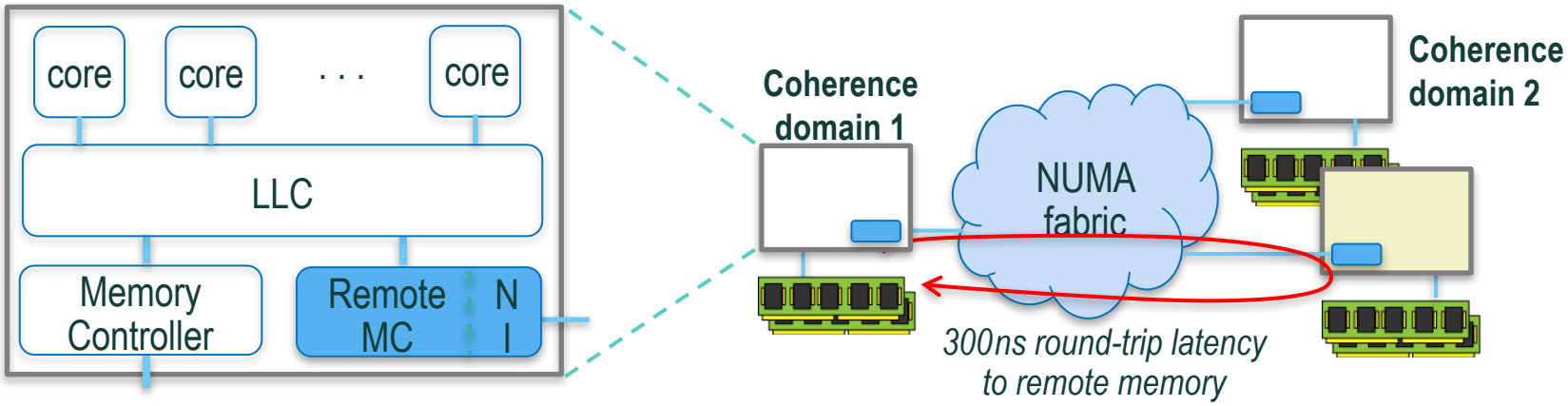
- Provides high b/w access to hot data
- Mitigates the read/write disparity
  - Data read/written in pages (bulk)

DRAM cache + battery:

- Helps mitigate persistent ordering stalls

# Scale-Out NUMA: Rack-Scale Memory Pooling

ecocloud
an EPFL research center




Pool memory over a light fabric:

- Balance load skew in data serving
- Mitigate partition skew in analytics

HUAWEI

Microsoft

soNUMA:

- Socket-integrated network interface (e.g., Sonoma)
- Protected global memory read/write + synch

# 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, dataflow

- 10x slower clocks
- Better for sparse arithmetic

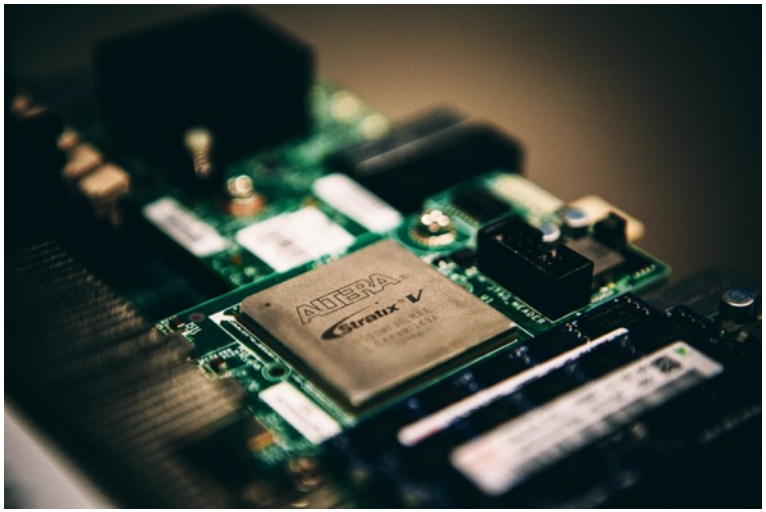
Microsoft, Amazon & Intel

Conventional Server
CPU (e.g., Xeon)

FPGA
(e.g., Catapult)

# FPGA's in Servers






Microsoft Unveils Catapult to Accelerate Bing!

[EcoCloud Annual Event, June 5th, 2014]

Latest version:

- One FPGA per blade
- Sits on the network
- Backend connected to CPU/NI
- Originally to accelerate Bing, Azure
- Now ML service called BrainWave
- Intel's HARP: tighter integration

# Memory-Centric Accelerators Abound

ecocloud
an EPFL research center

## Google's TPU:

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

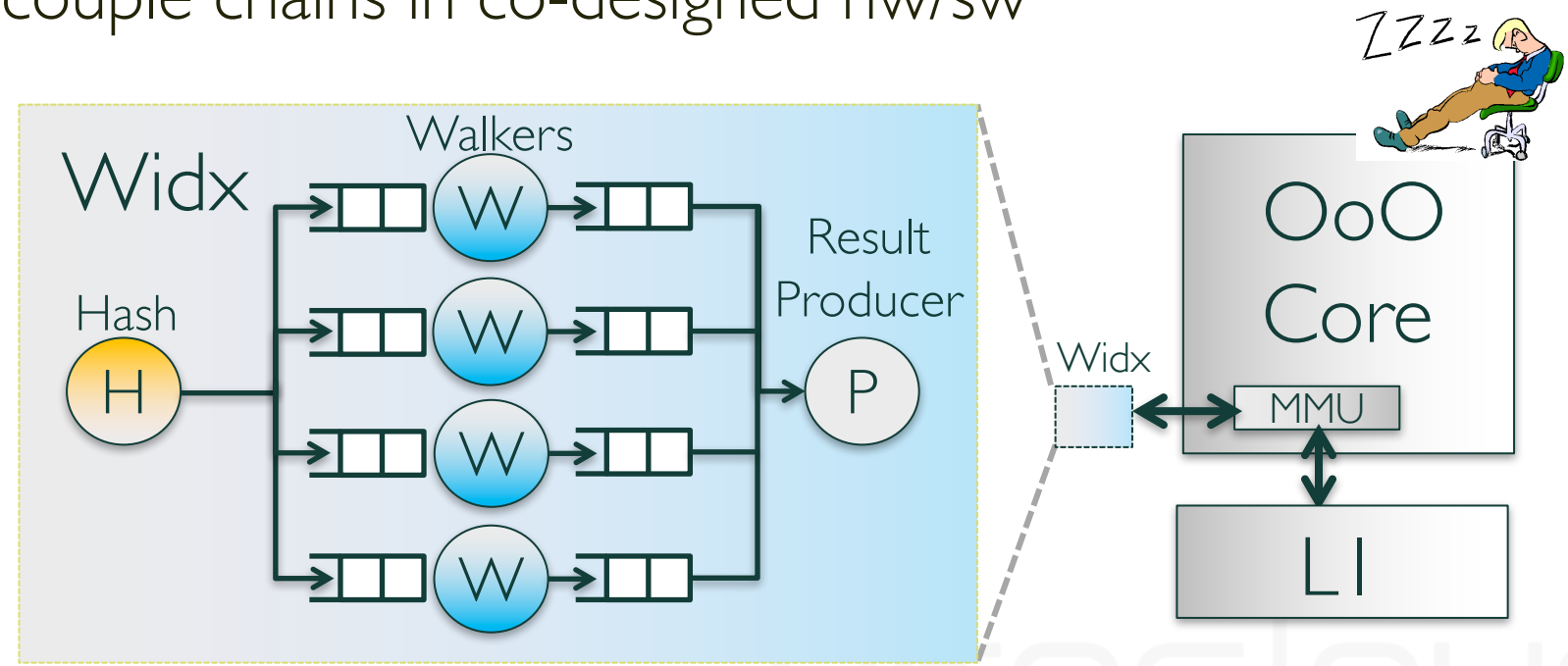
## Oracle's RAPID:

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

# Walkers: Accelerating Data Management

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 performance/Watt over Xeon

# Walkers in Software [VLDB'16]

ecocloud
an EPFL research center

Use insights to help Xeon

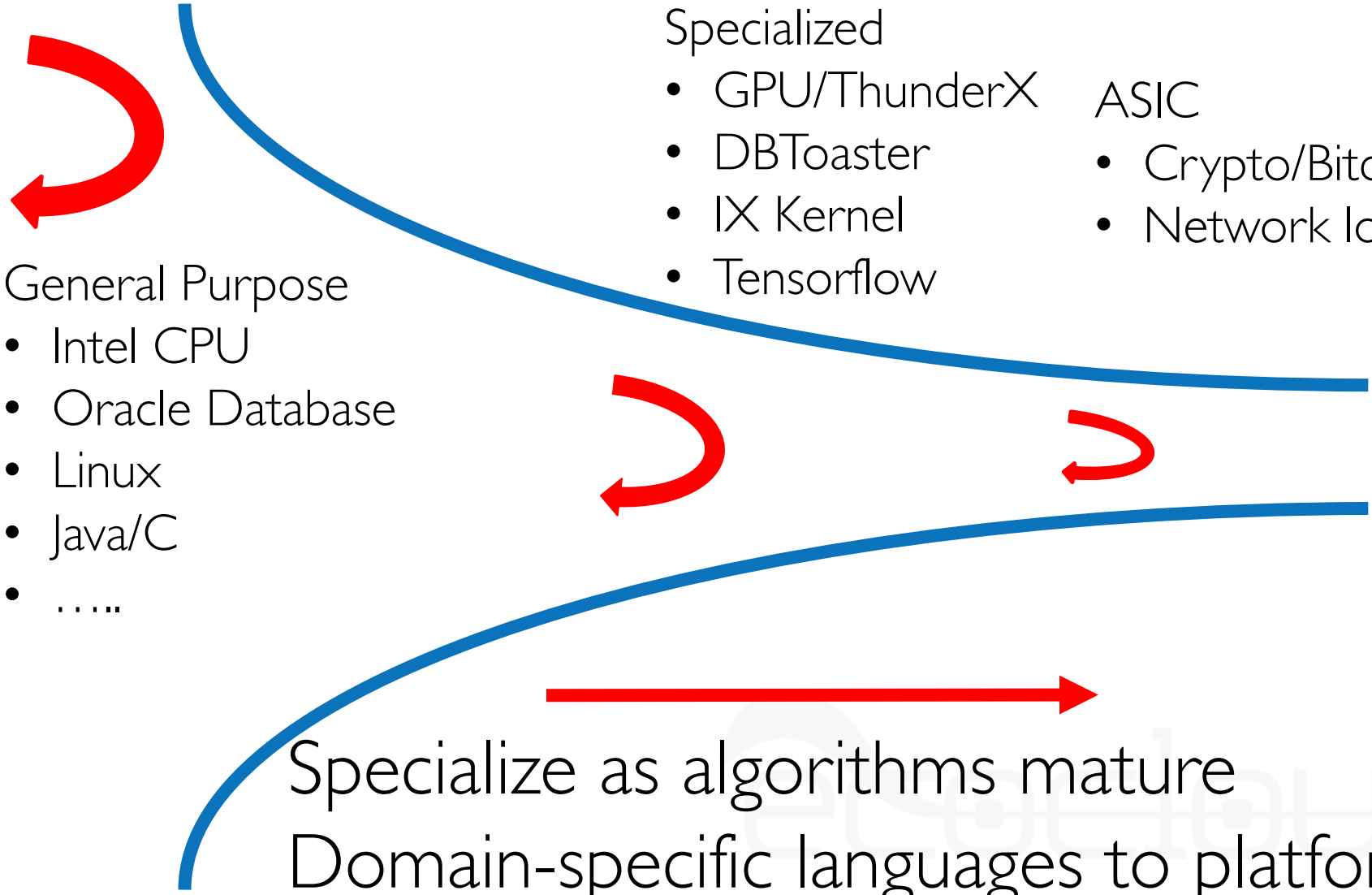
- Decouple hash & walk in software
- Create & manage queues in wraparound code

2.3x speedup on Xeon

- Unclogs dependences in microarchitecture
- Maximizes memory level parallelism
- Under consideration by SAP HANA [VLDB'18]

# The Specialization Funnel

ecocloud
an EPFL research center



General Purpose
- Intel CPU
- Oracle Database
- Linux
- Java/C
- …..

Specialized
- GPU/ThunderX
- DBToaster
- IX Kernel
- Tensorflow

ASIC
- Crypto/Bitcoin
- Network logic

Specialize as algorithms mature
Domain-specific languages to platforms

# 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

# Outline

ecocloud
an EPFL research center

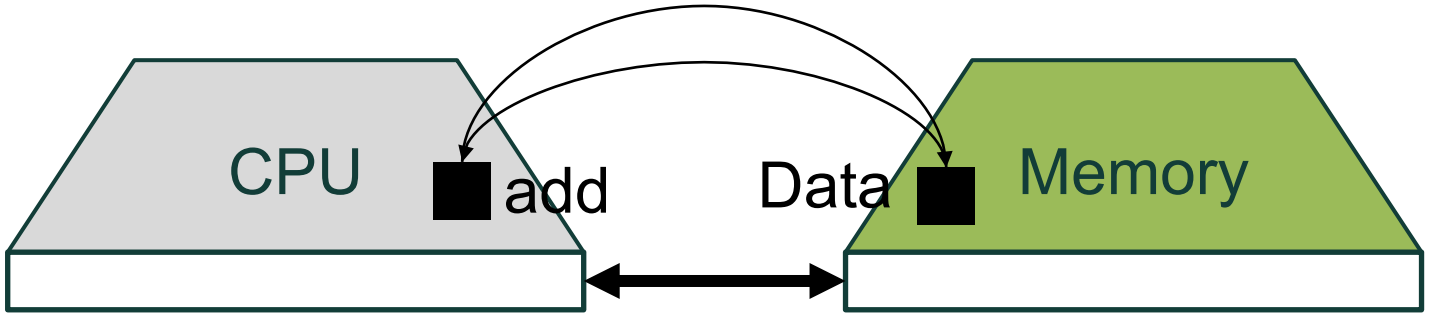
- ~~Overview~~
- ~~Memory-Centric Servers~~
- Near-Memory Processing
- Summary

# What happens on servers?

ecocloud
an EPFL research center

Huge datasets reside in memory

- Fetch data
- Perform minimal computation
- Repeat over dataset



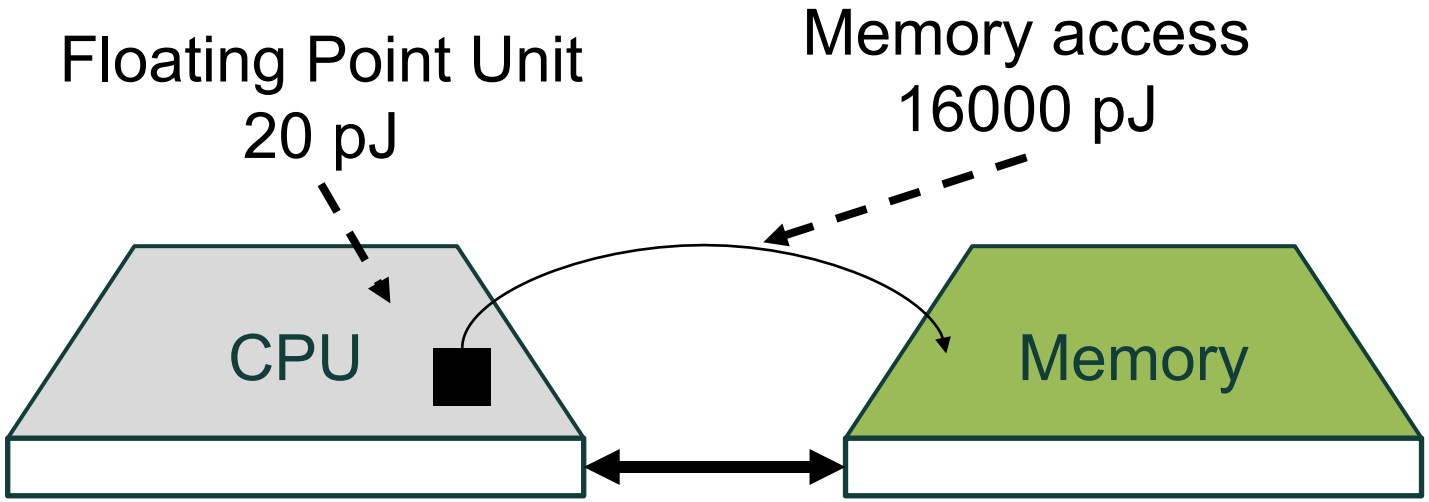

Data-centric services revolve around data movement

# The Cost of Moving Data

ecocloud
an EPFL research center

Arithmetic operations are cheap

Moving data dissipates much energy




[Dally, SC'14 Panel]

Energy is dominated by data movement!

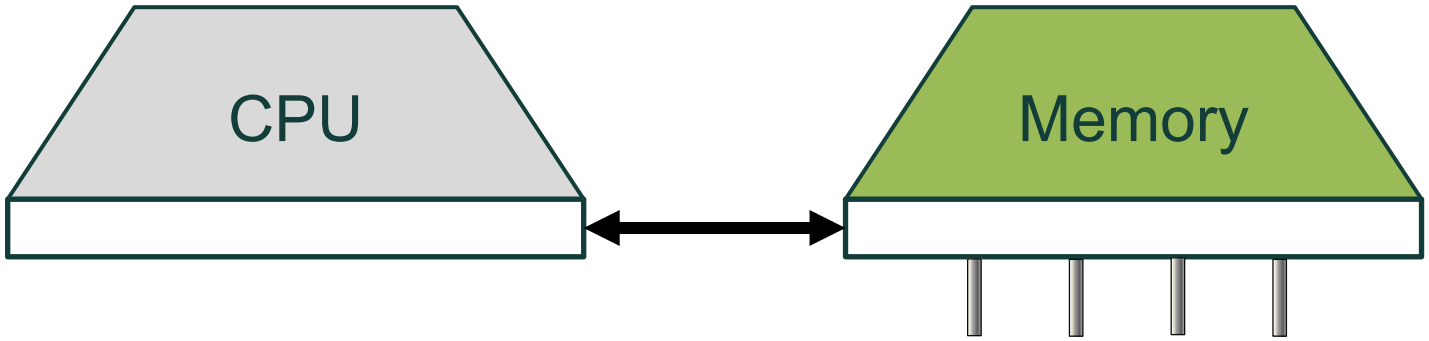
# The Parallelism Bottleneck

ecocloud
an EPFL research center

CPU-DRAM bandwidth (DIMM Channel) ➜ 24 GB/s

- External interfaces either low in b/w or power hungry

But, internal DRAM bandwidth (Chip) ➜ 128 GB/s



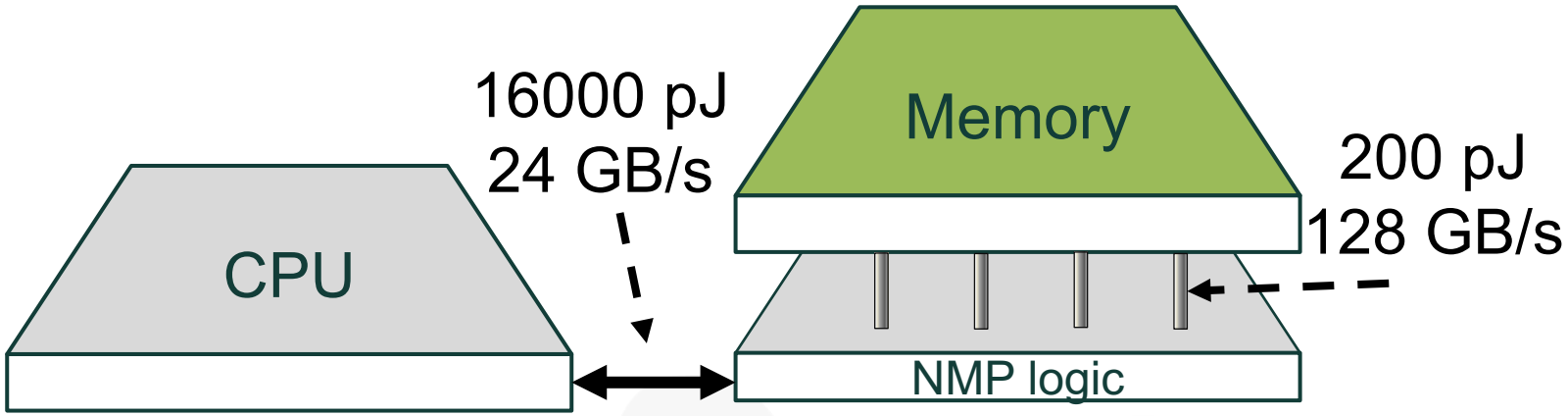

Parallelism is limited by connectivity

# NMP comes to rescue

ecocloud
an EPFL research center

Near-memory processing (NMP):

- A layer of logic placed closer to DRAM
- Die-stacked or on interposer
- Helps exploit parallelism
- Reduces data movement




Redesign algorithms, SW & HW to realize NMP potential

# Why not compute inside DRAM?

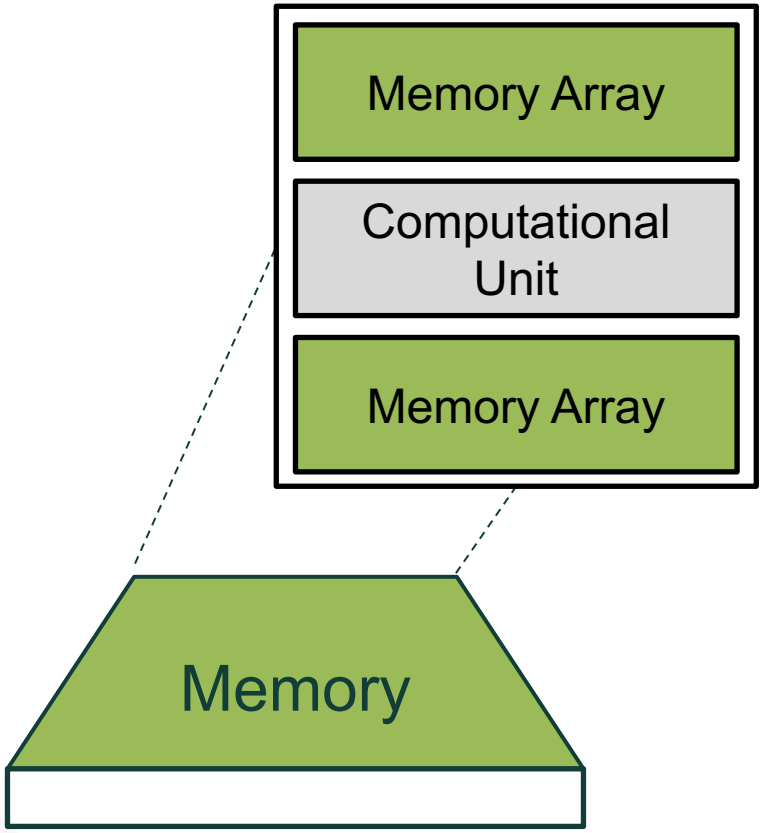
ecocloud
an EPFL research center

## Idea emerged in the 90's

- IRAM/PIM
- Logic & Memory on the same die

## Did not make it

- Lowers DRAM density
- Increases DRAM costs
- DRAM is highly cost-sensitive




Must maintain DRAM cost advantages

# NMP Commandments
## [IEEE Micro issue on Big Data'16]

ecocloud
an EPFL research center

Not (CPU) business as usual

1. DRAM favors streaming vs. random access
   - CPU's leverage reuse & locality in cache hierarchy
2. DRAM favors wide (slow) cores vs. many (fast) cores
   - Stream-level parallelism to match DRAM b/w
3. Memory must maintain semantics relative to CPU
   - Shared address space + coherence between NMP & CPU

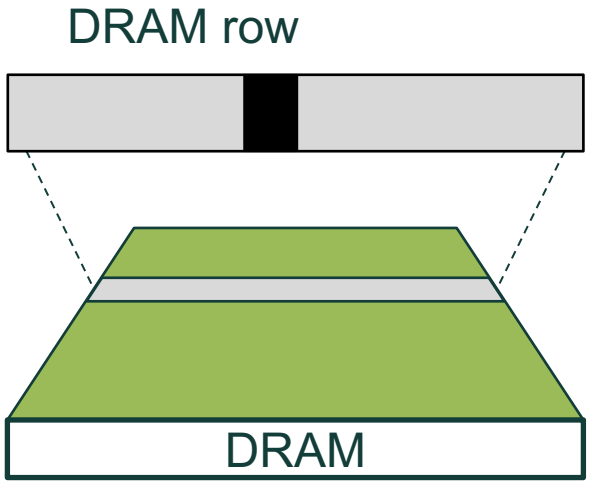
Must co-design algorithm/HW for NMP!

# Why not random access?

ecocloud
an EPFL research center

Internal DRAM structure dictates

- Activating a 1KB row of data
- Dominates access latency & energy




Treat DRAM as a block-oriented device

- Stream data
- Maximize bandwidth & efficiency

Example:

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

Must use algorithms that favor sequential access!

# The Mondrian Data Engine [ISCA'17]

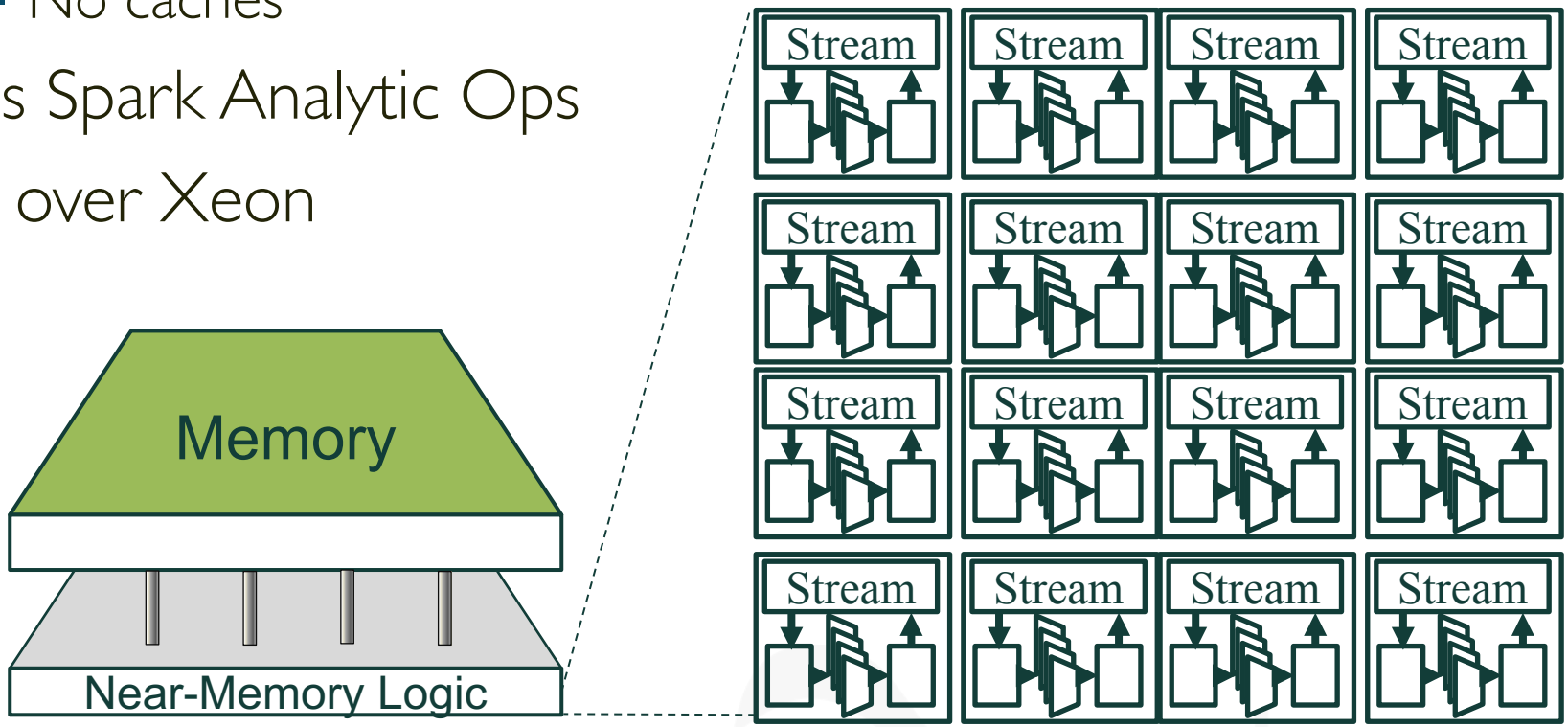
ecocloud
an EPFL research center

SIMD cores + data streaming

- 1024-bit SIMD @ 1 GHz
- No caches

Runs Spark Analytic Ops

50x over Xeon




Algorithm/hardware co-design maximize near-memory performance

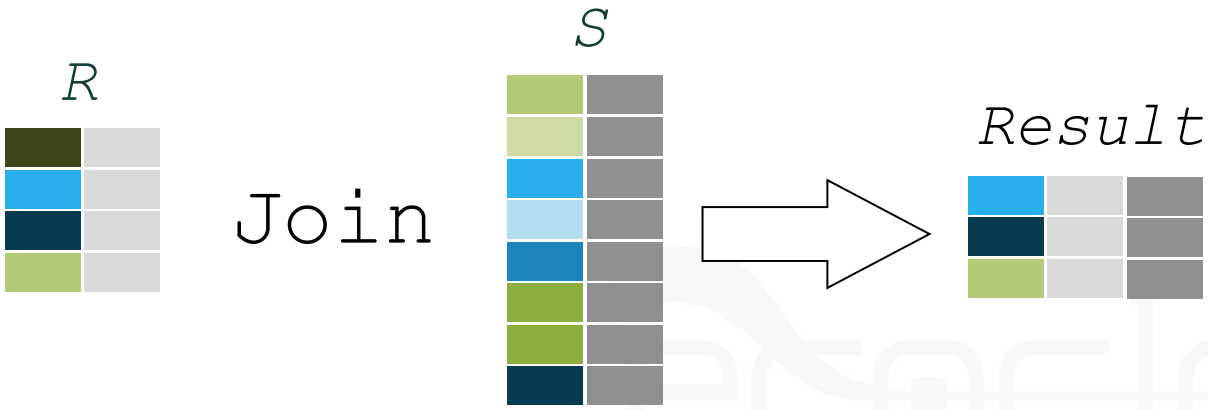
# Case study: Join

Iterates over a pair of tables

Finds the matching keys in two tables

Major operation in data management

```
Q: SELECT ... FROM R, S WHERE R.Key = S.Key
```



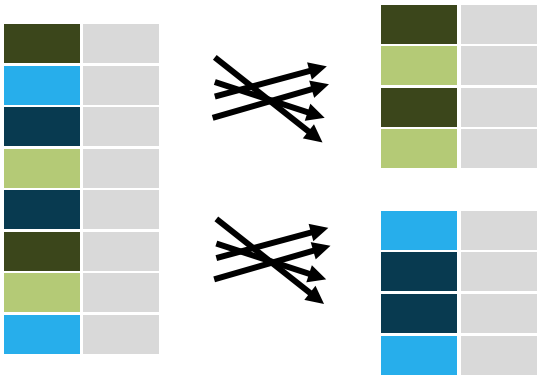
# CPU-centric (Hash) Join

ecocloud
an EPFL research center

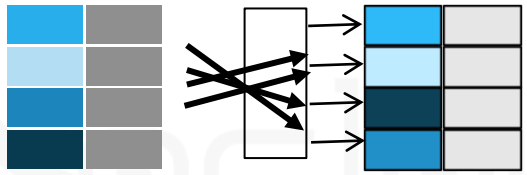
Performed in two phases: Partition & Probe

1. Partition tables based on keys
2. Probe joins partitions
   - Optimized for random accesses to cached data

Partition



Probe



# Access patterns in hash Join

ecocloud
an EPFL research center

| Phases | Hash |
| --- | --- |
| 1. Partitioning | ☹ |
| 2. Build hash table | ☹ |
| 3. Probe hash table | ☹ |

☹: Random access (local or remote)

# Join operation on Mondrian

ecocloud
an EPFL research center

Revisiting Sort join [ASBD'14]:

- Sort join (O(nlogn)) vs. Hash Join (O(n))
- Sort tables and then merge join

Perform more work

But, sort and merge use streaming access

Trade algorithm complexity for streaming memory accesses

# Comparing access patterns

ecocloud
an EPFL research center

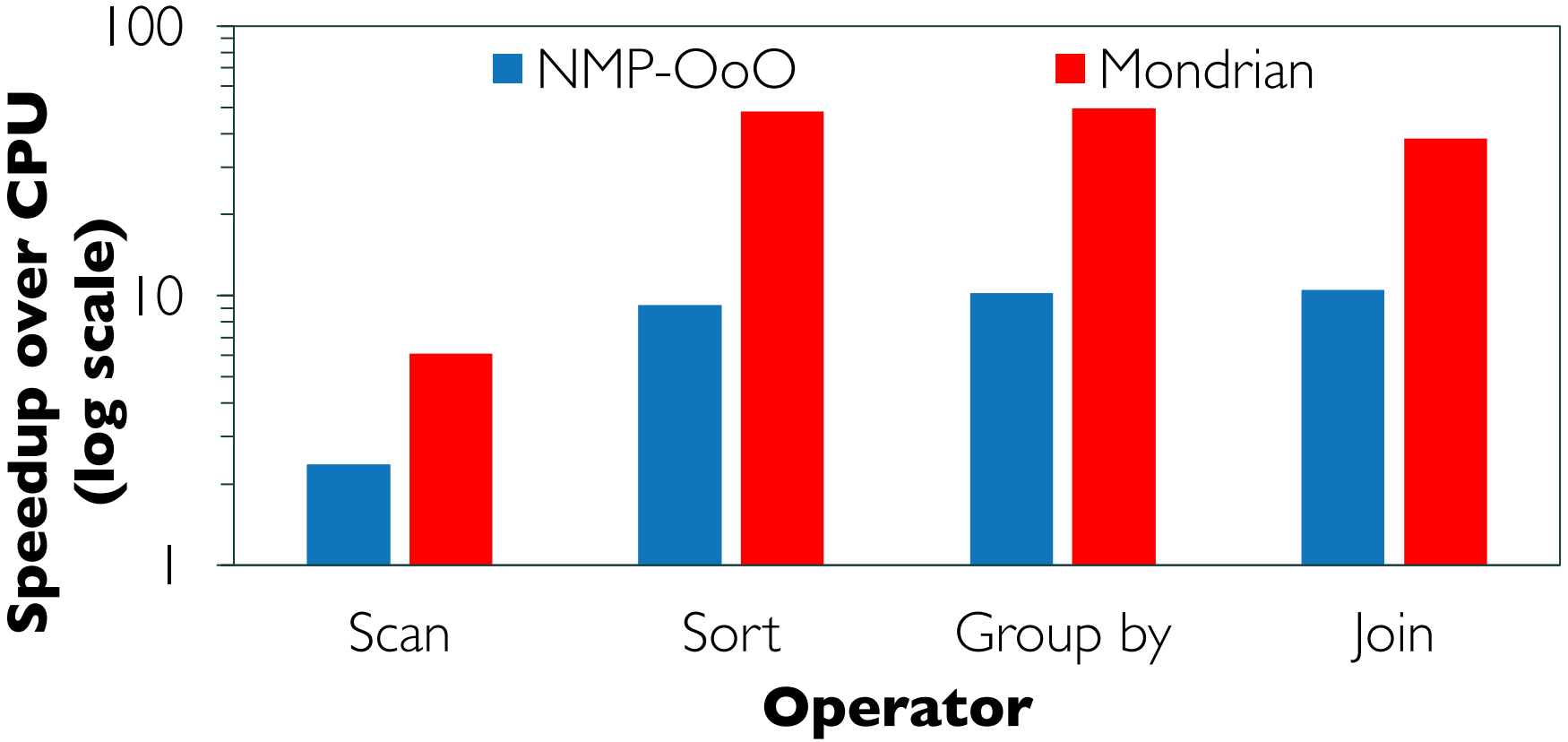
| Phases | Hash | Sort |
|---|---|---|
| **1. Partitioning** | ☹ | ☹/😐 |
| **2. Build / Sort** | ☹ | ☺ |
| **3. Probe / Merge** | ☹ | ☺ |

☹: Random access (local or remote)

😐: Streaming access (remote)

☺: Streaming access (local)

# Performance




- Algorithm alone gets ~10x [ASBD'15]
- Algorithm/hardware co-design gets 50x

# Summary

ecocloud
an EPFL research center

Trends for data & online services:

- Data growing fast
- Online services are in-memory
- Memory is a big fraction of TCO

Post-Moore server designs:

- Opportunities abound
- Processors, accelerators, memory, network, system

Integration + Specialization + Approximation

# Thank You!

ecocloud
an EPFL research center

For more information please visit us at

**ecocloud.ch**



ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE