The Clouds Have Taken Over, But Algorithms are Here to Save the Day

Babak Falsafi
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+

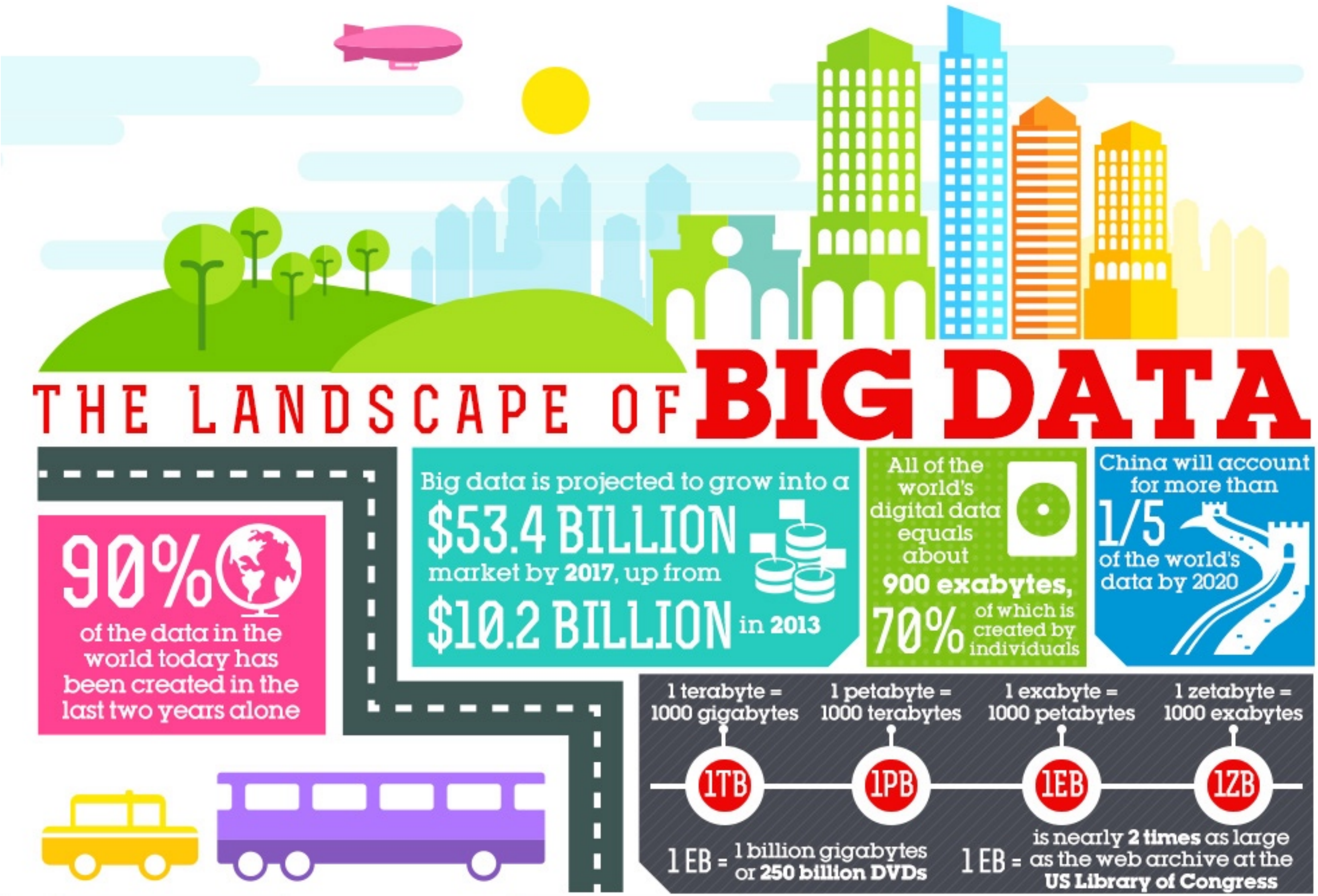
Mainframes

PC Era

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

Data Economics

ecocloud
an EPFL research center

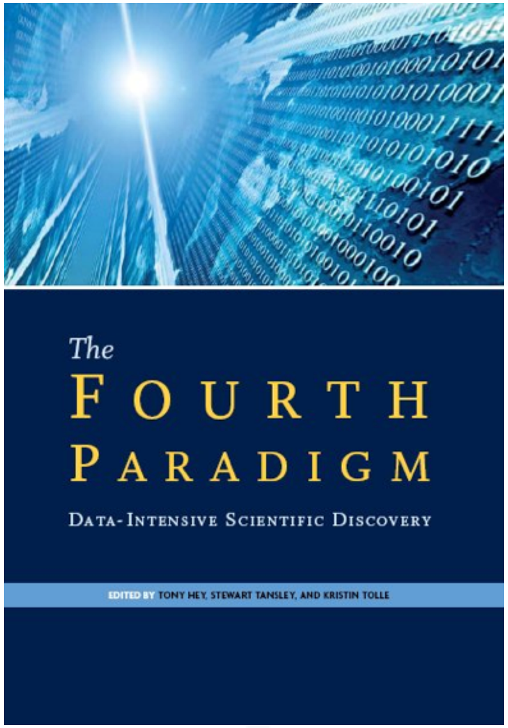


Data Shaping All Science & Technology

Science entering 4th paradigm

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

Complements theory, empirical science & simulation

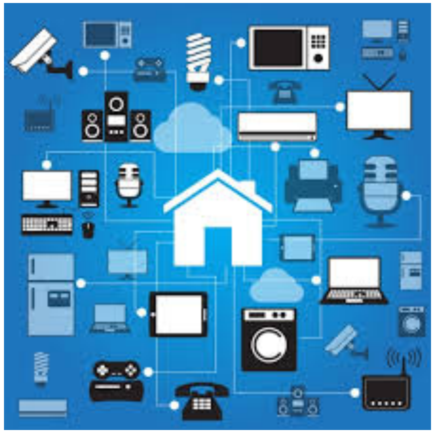


Data-centric science key for innovation-based economies!

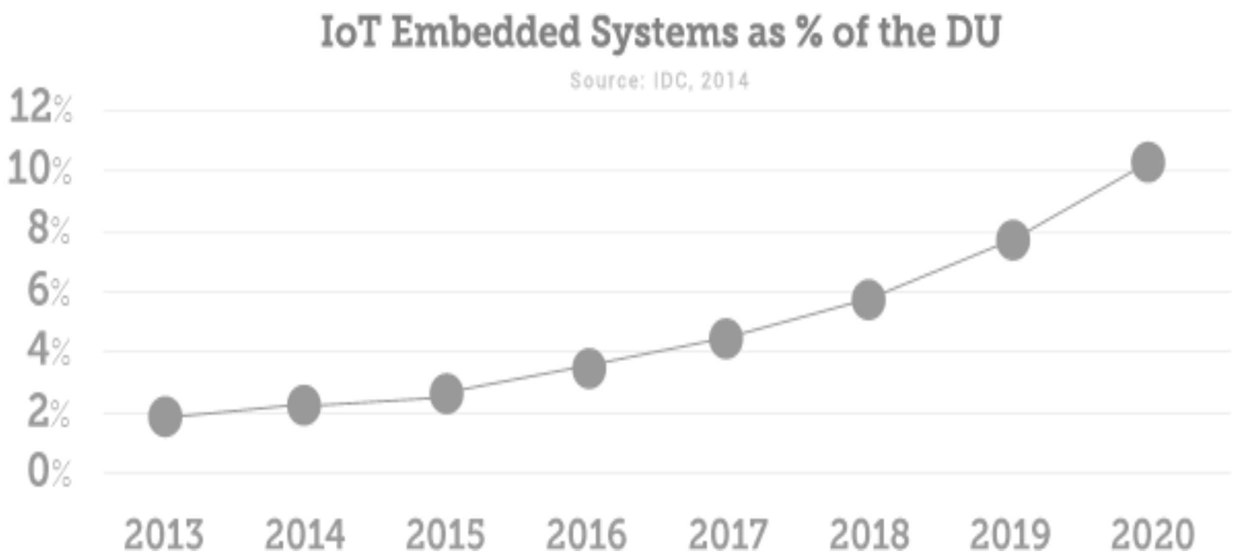
Challenges in Data-Centric Science [Frontiers in Massive Data Analysis, 2013]

- Massive data sets
- Distributed data sources
- Sampling biases & heterogeneity
- Heterogeneous data formats
- Scalable & incremental algorithms
- Algorithms for parallel architecture
- Ensuring data integrity & security
- Enabling data discovery, integration, sharing
- Visualization
- ….

Internet-of-Things (IoT) Growing Fast Too



20 Billion Connected Devices



4 Zettabytes of Data, 10% of Digital Universe

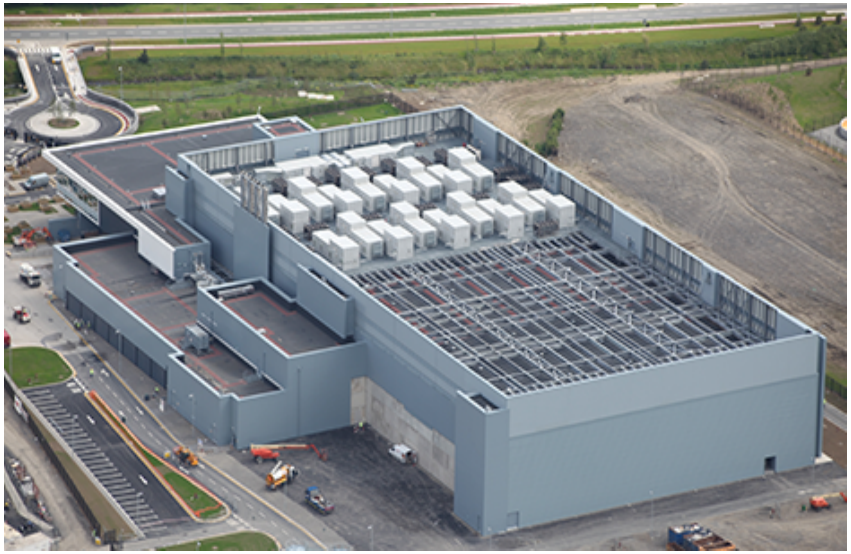
$7 Trillion
Market Revenue

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

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

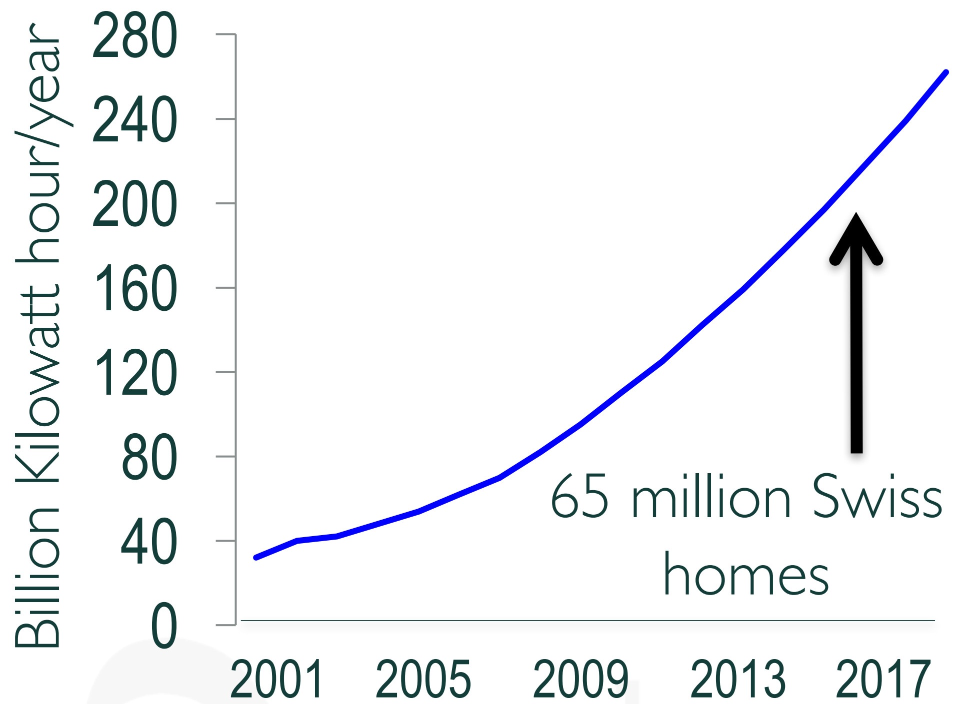


20MW, 20x Football Field
$3 billion

The Ecological Impact of Datacenters

- 1.5% of electricity worldwide
- More in IT-based economies
  - E.g., 6% in London
- Growing ~ 20%

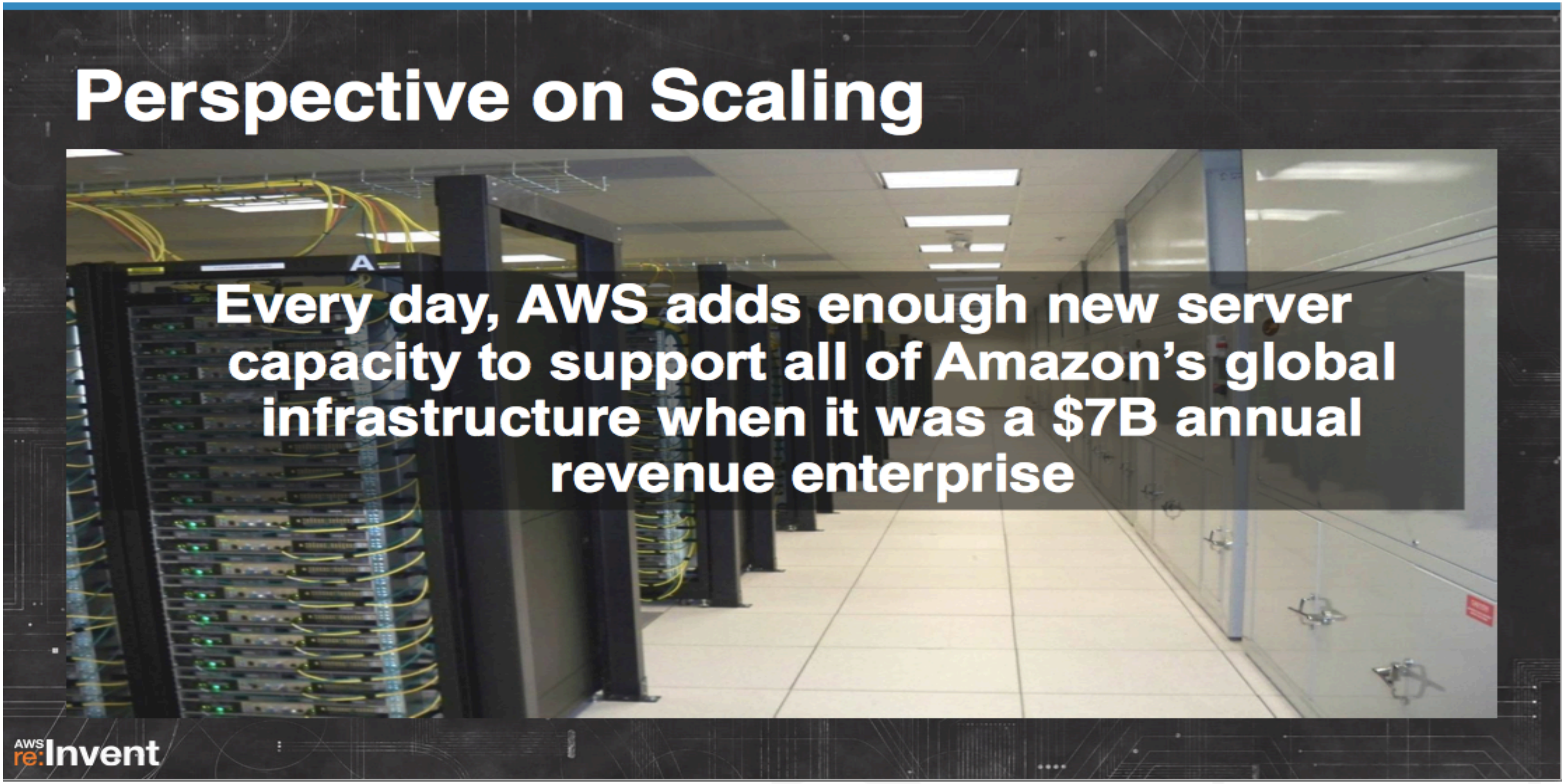
Electricity Demands for Datacenters in the US



[source: Energy Star]

Source: James Hamilton, 2014

mvdirona.com/jrh/TalksAndPapers/JamesHamilton_Reinvent20131115.pdf



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

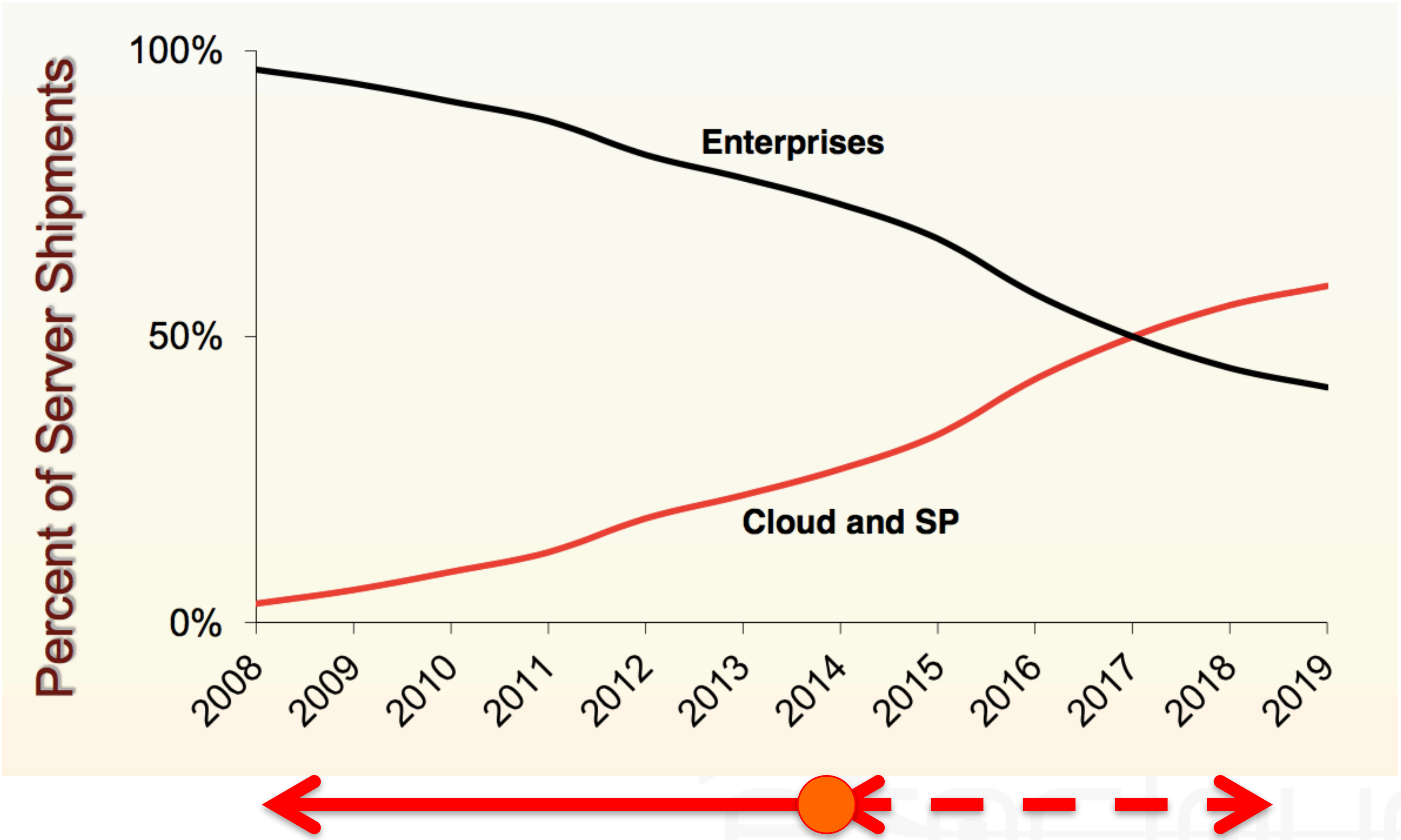
Datacenters are the IT utility plants of the future

Supercomputing

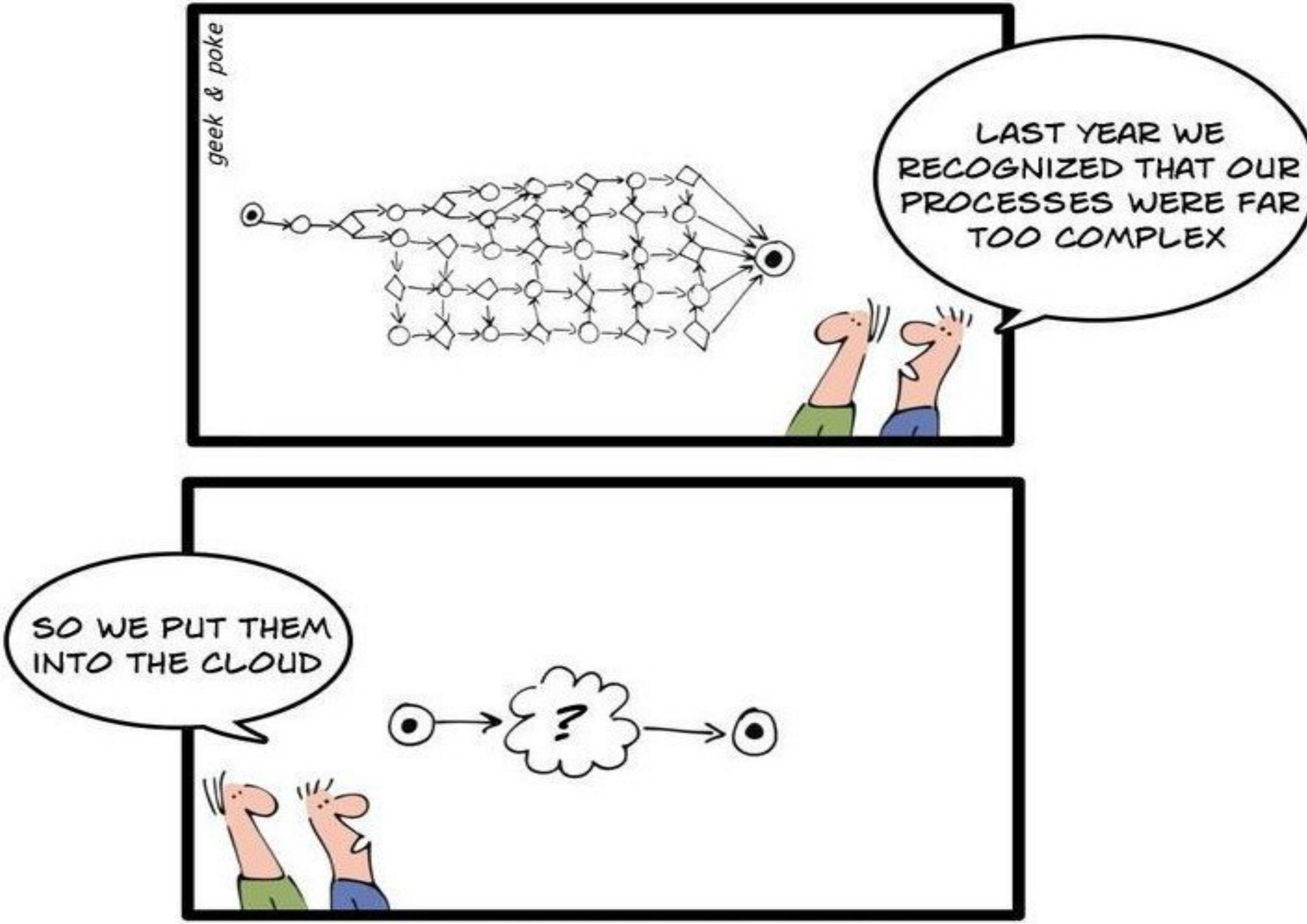
≠

Cloud Computing

Cloud Taking Over Enterprise



Source: Dell 'Oro 2Q15



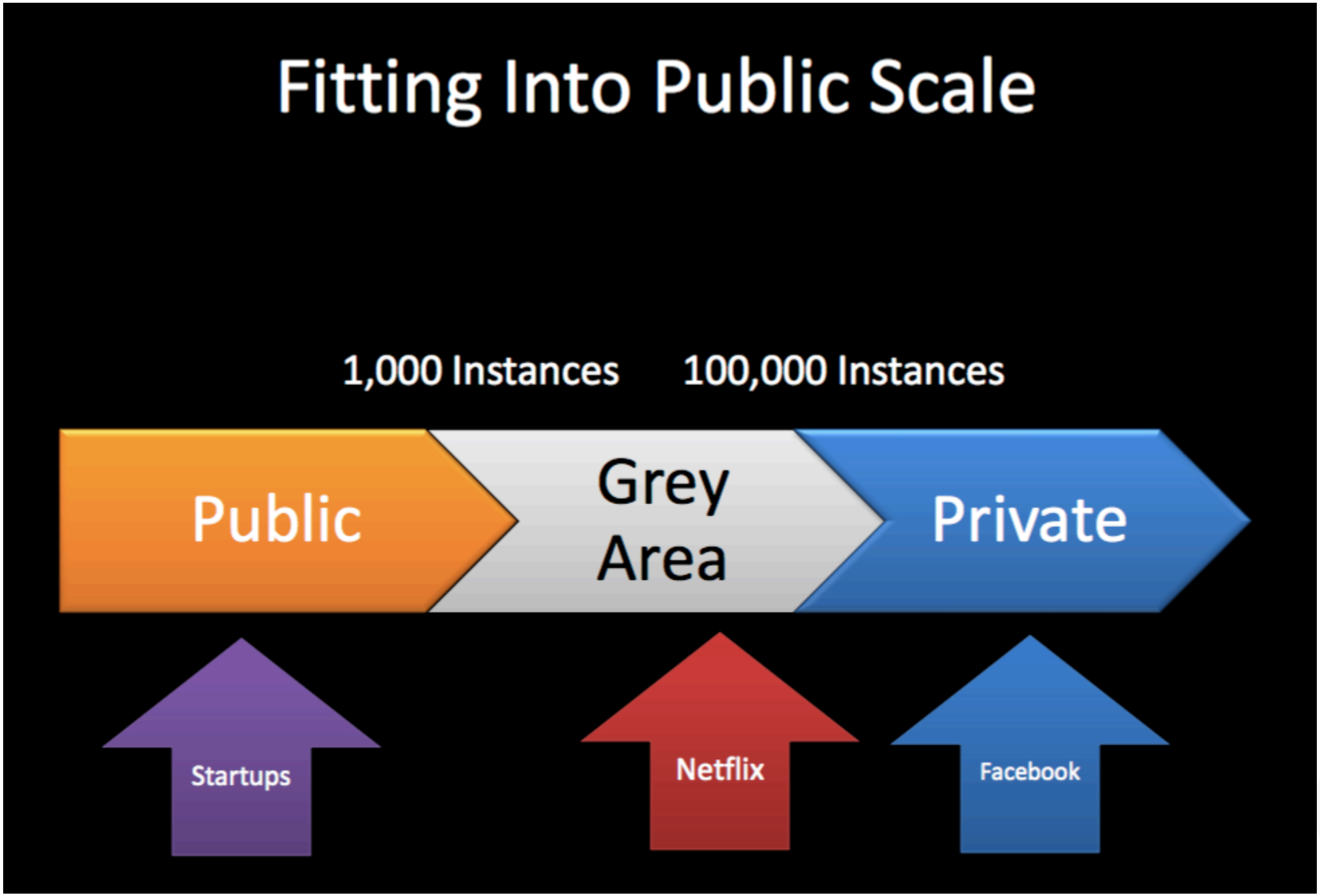
LET THE CLOUDS MAKE YOUR LIFE EASIER

Why the Cloud?

ecocloud
an EPFL research center

- Focus on core business (not IT)
- Massive resources at low cost
  - 1K ➔ 100K nodes TCO/servers drops by 80%
  - At the forefront of technology
- Unprecedented business intelligence
  - Data/operation analytics, enhanced customer view, security,….

Private Clouds Squeezed



Source: Adrian Cockcroft, NetflixOSS, 2013

Applications Abound

juno

airbnb

nest

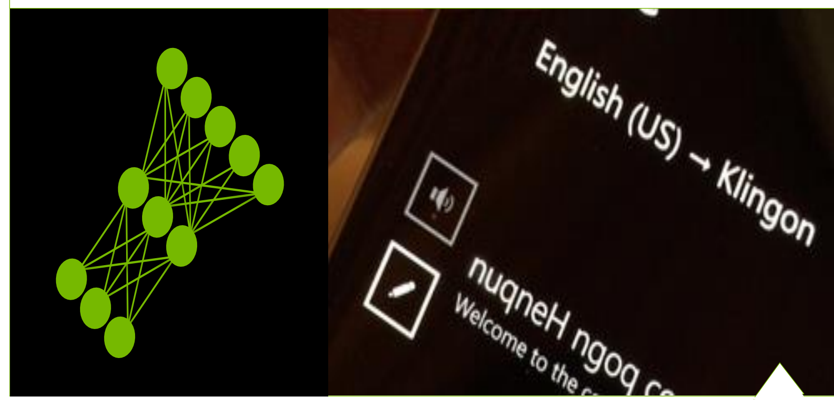
NETFLIX

wealthfront

DEEP LEARNING EVERYWHERE

Image Classification, Object Detection, Localization, Action Recognition

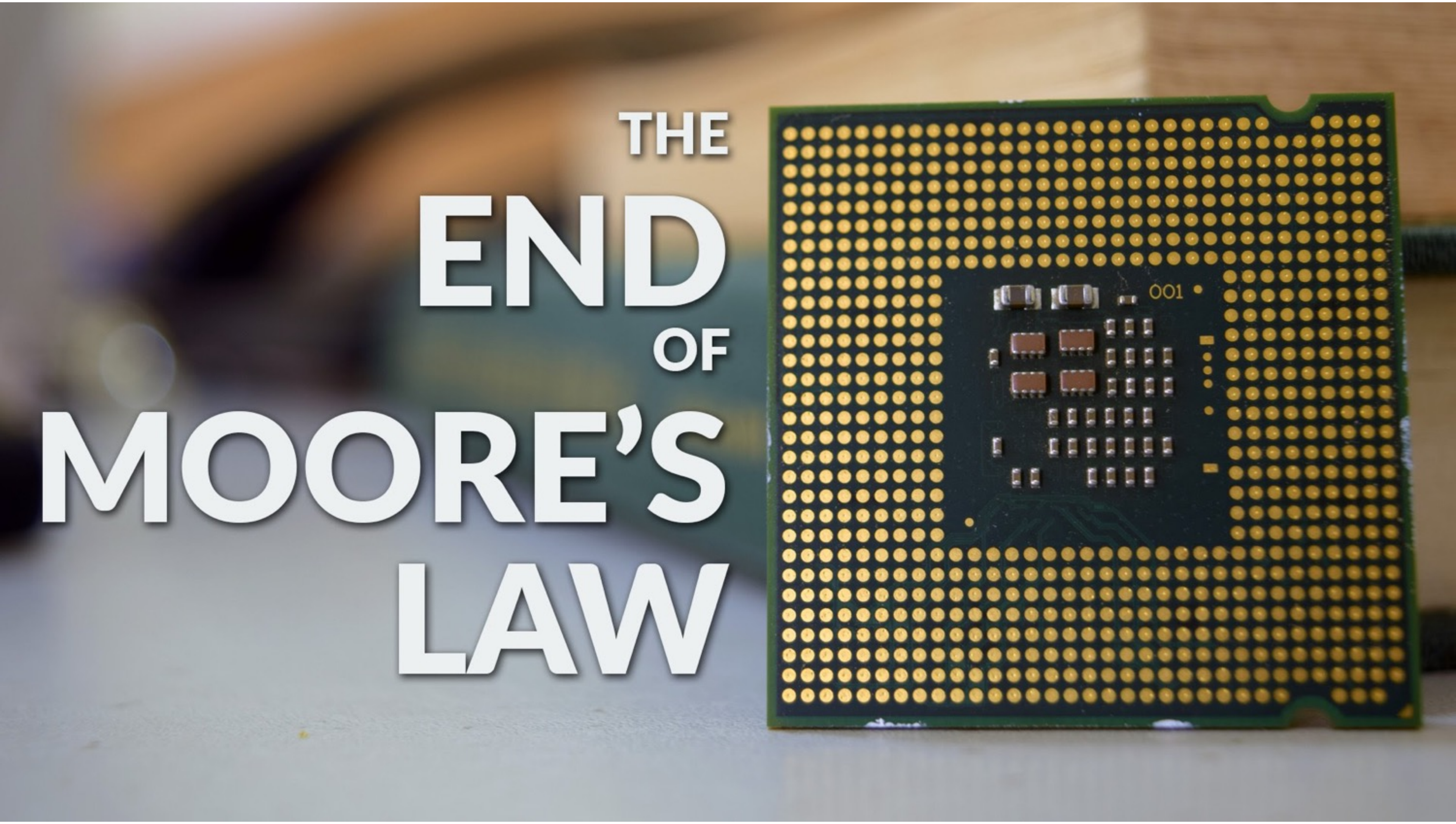
Speech Recognition, Speech Translation, Natural Language Processing



Pedestrian Detection, Lane Detection, Traffic Sign Recognition

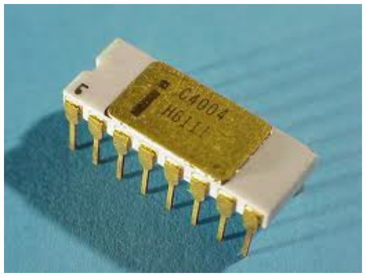
Breast Cancer Cell Mitosis Detection, Volumetric Brain Image Segmentation

Challenges Ahead



Moore's Law:
Five Decades of Exponential Growth

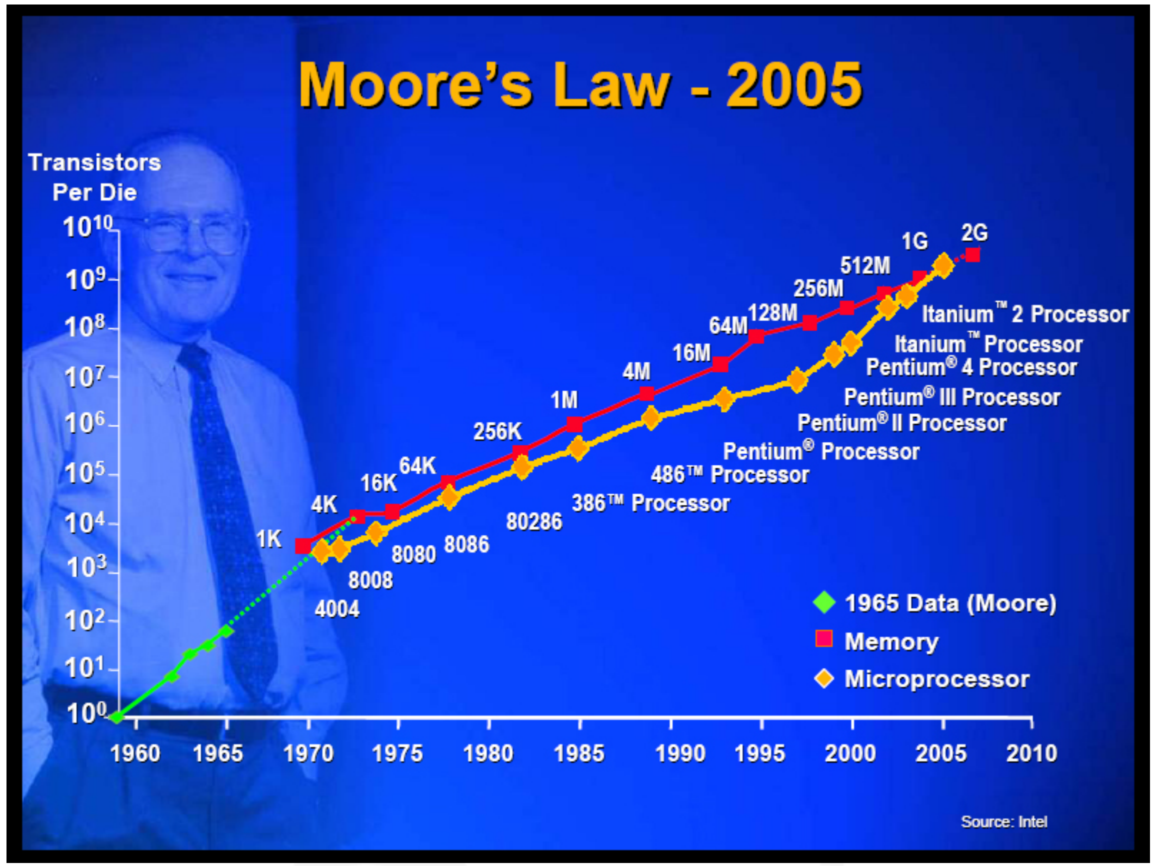
Intel 4004, 1971



92,000 ops/sec

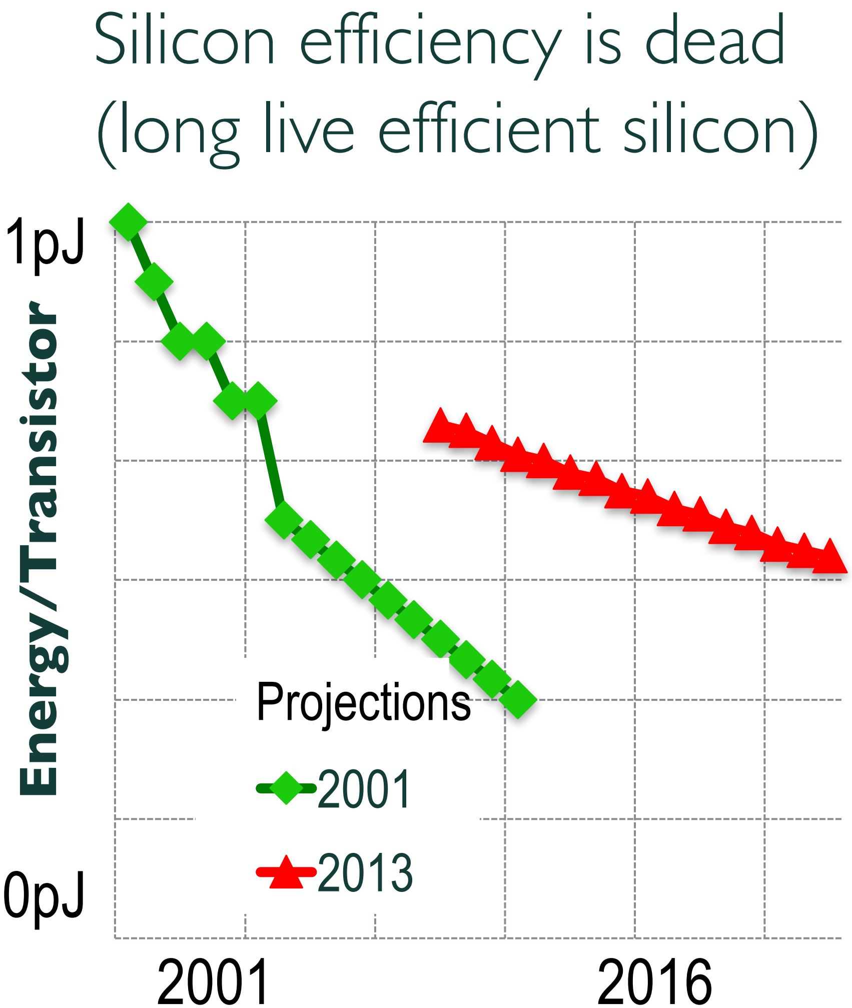
Intel Xeon, 2014

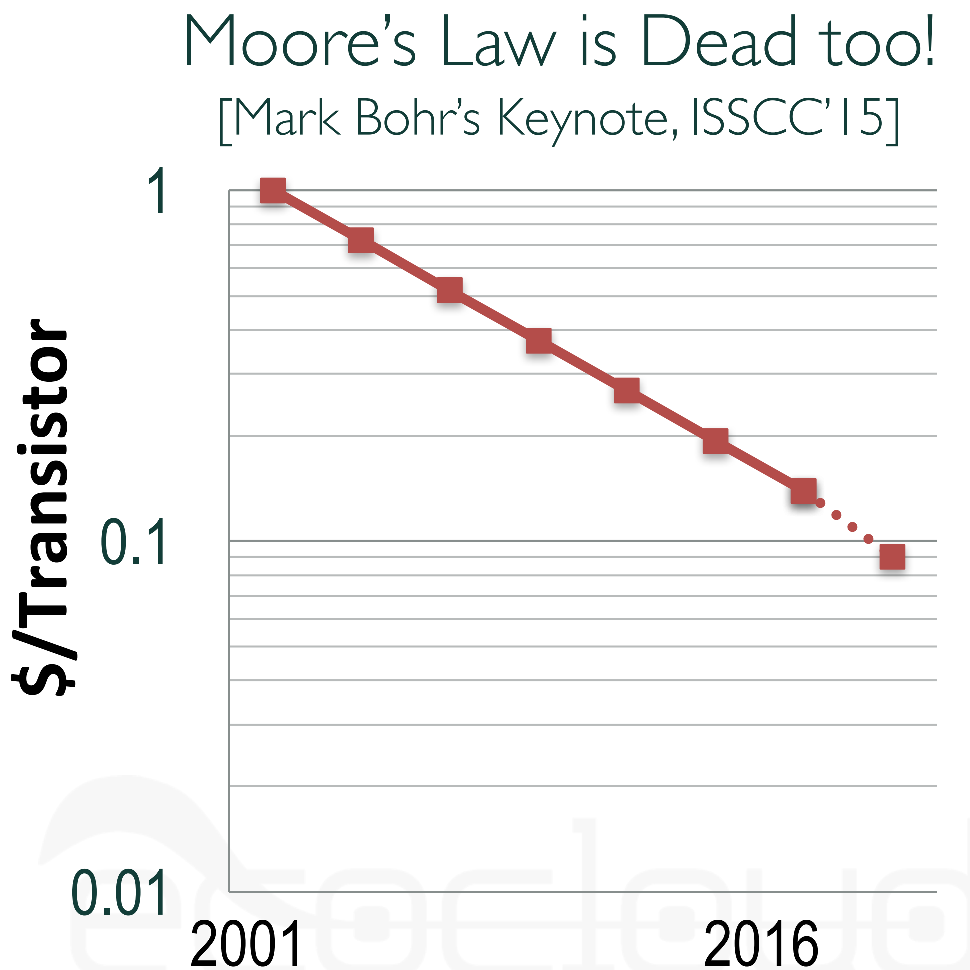
266,000,000,000 ops/sec



Made IT an indispensable pillar of our society!

Silicon is running out of steam!





Recap

- Demand is growing at > 50%/year
- Silicon density was growing at 41%/year
  - Intel chips in 2012 show density growth at 17%

- Where do we go from here?
  - Technologies on the horizon but no silver bullets!
  - Must build platforms ground up
  - But, sustained orders of magnitude can come **only** from algorithms

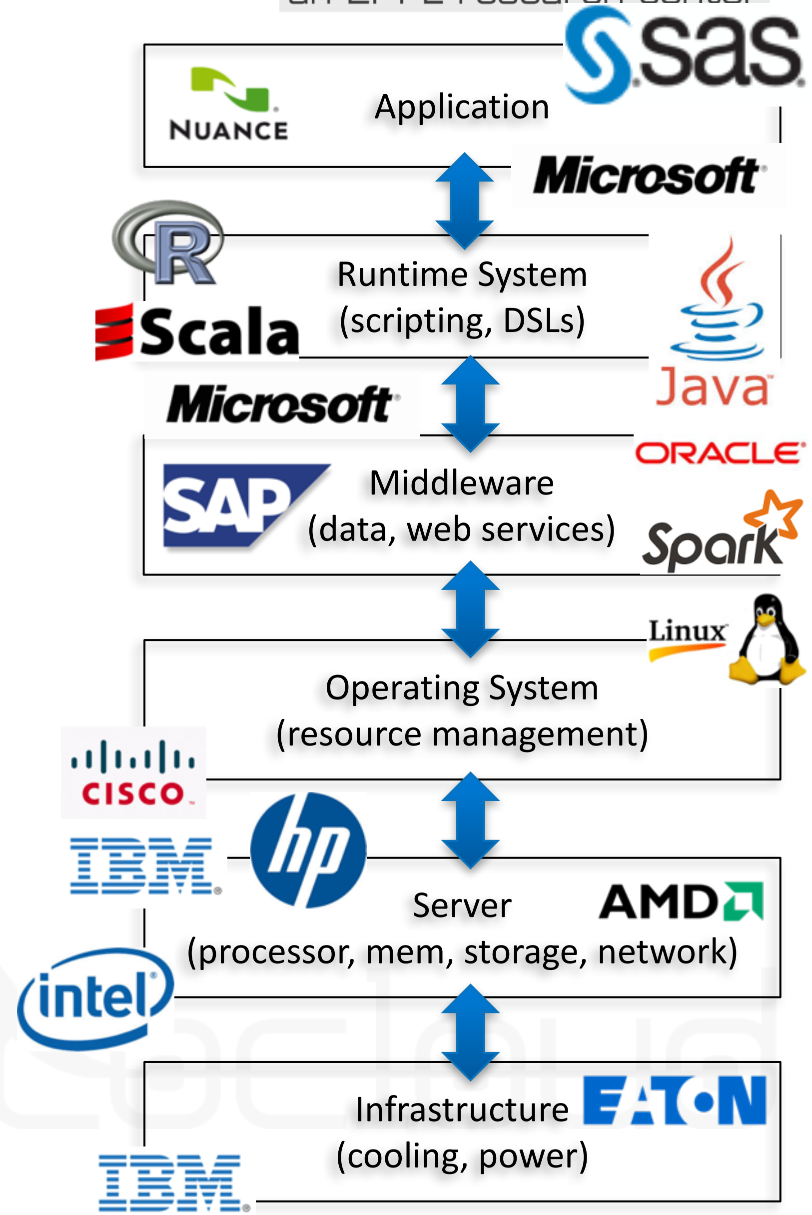
Today's Server Ecosystem

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



Optimizing Server Ecosystem

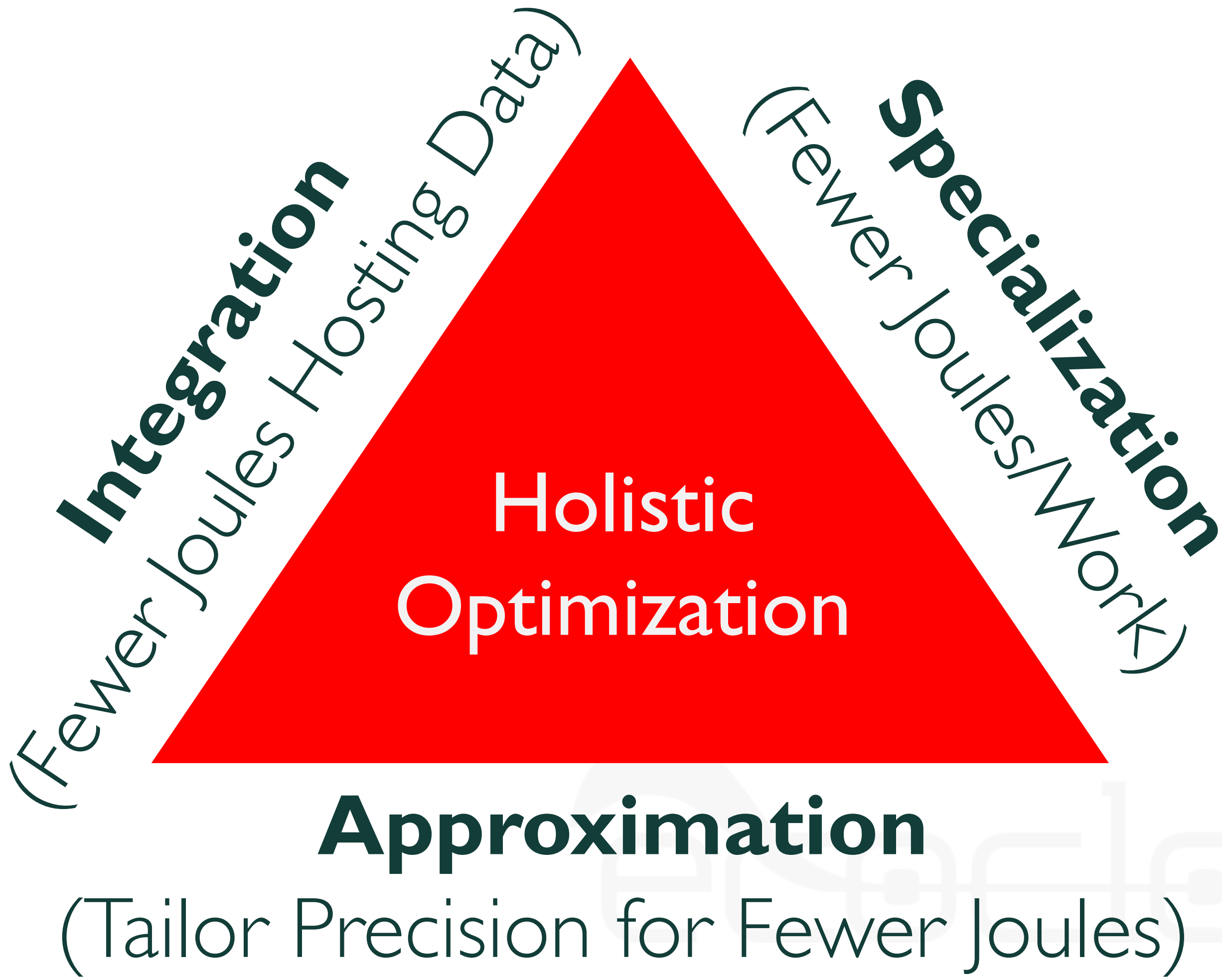
ecocloud
an EPFL research center

Holistic optimization

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

Open technologies!

Optimization Opportunities: The ISA Triangle



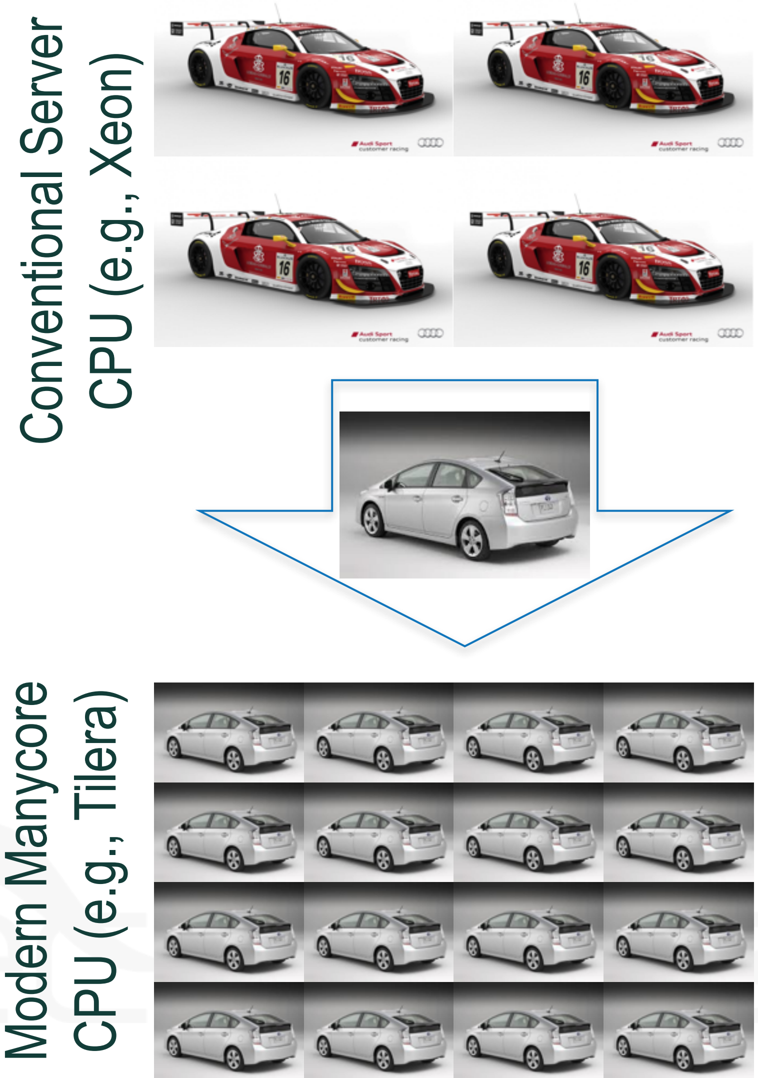
Accelerating Computing: Manycores

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



Server Benchmarking with CloudSuite 3.0 (cloudsuite.ch)

ecocloud
an EPFL research center

Data Analytics
Machine learning
Wikipedia, mahout

Data Caching
Memcached

Data Serving
Cassandra NoSQL
Cassandra

Graph Analytics
GraphX
GraphX

Media Streaming
Nginx, HTTP Server
NGINX

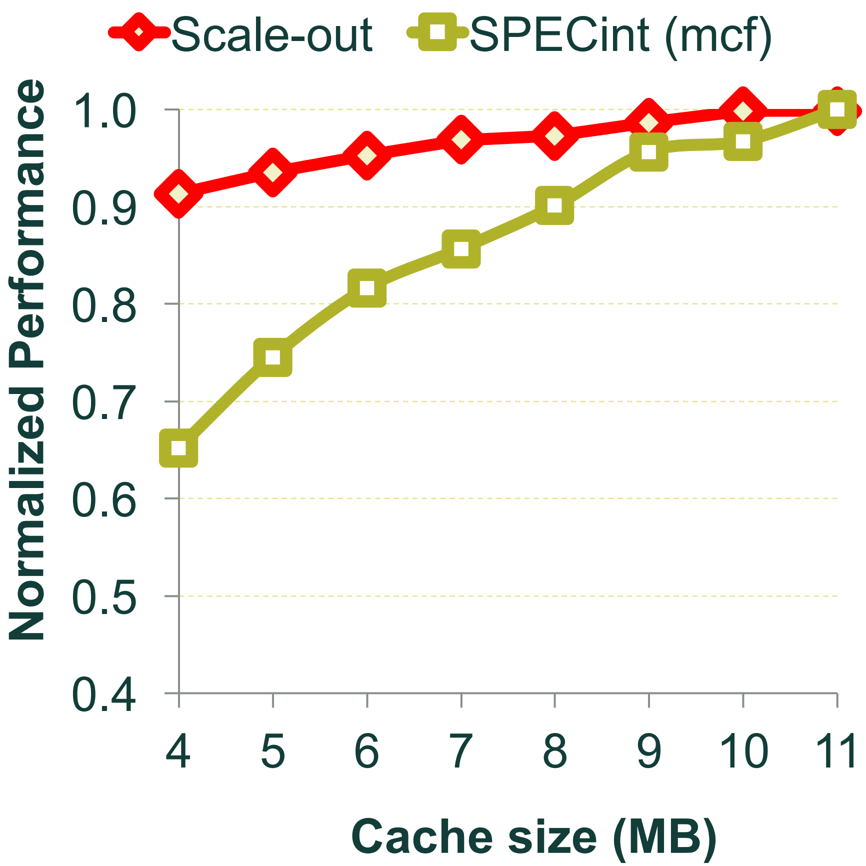
Web Serving
Nginx, PHP server
NGINX, php

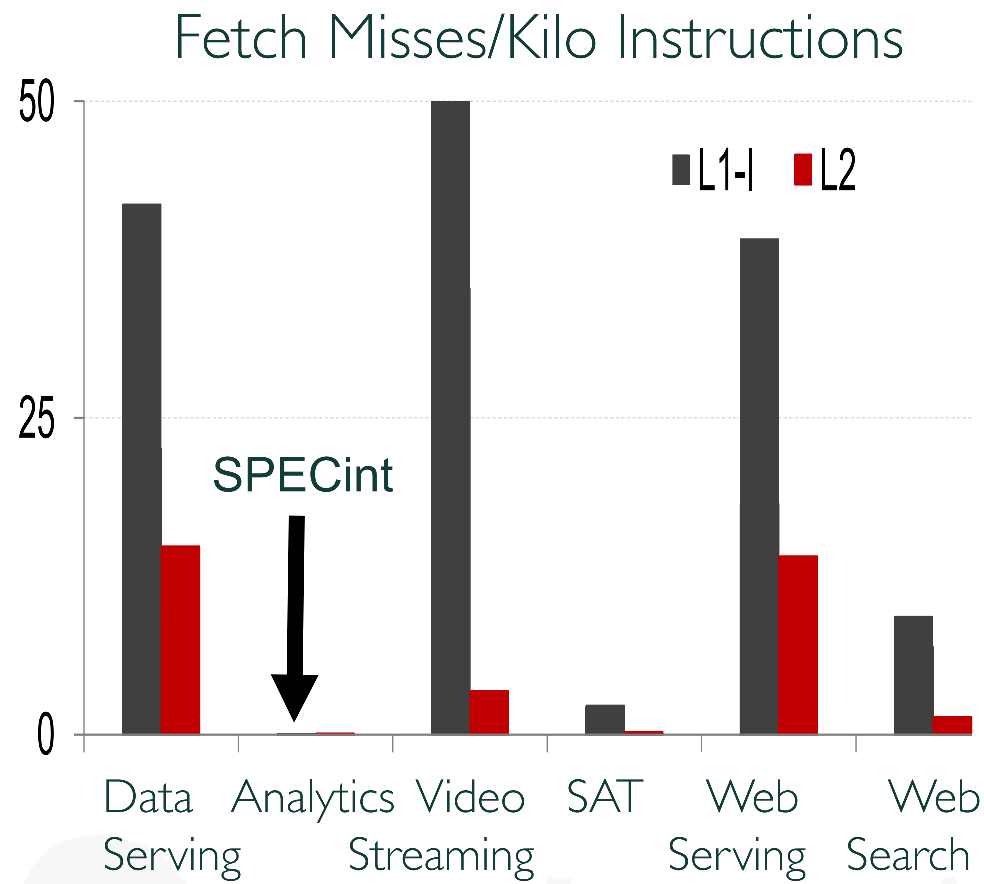
Web Search
Apache Solr & Nutch
Solr, nutch

In-Memory Analytics
Recommendation System
movielens, Spark

Building block for Google PerfKit, EEMBC Big Data!

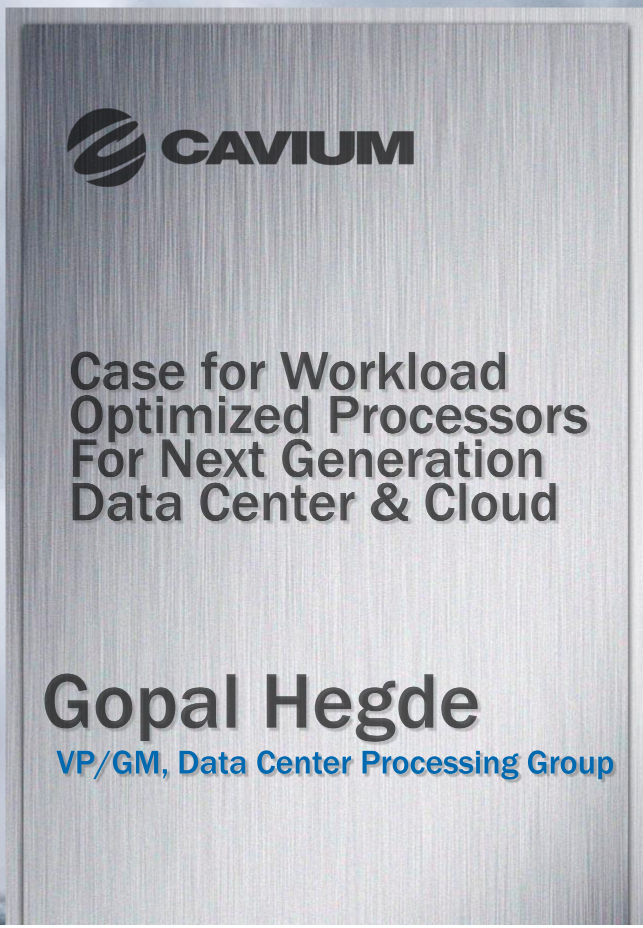
CloudSuite Stuck in Memory [ASPLOS'12]





- On-chip memory overprovisioned
- Instruction supply is bottlenecked

Manycore Accelerator for Data Serving



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

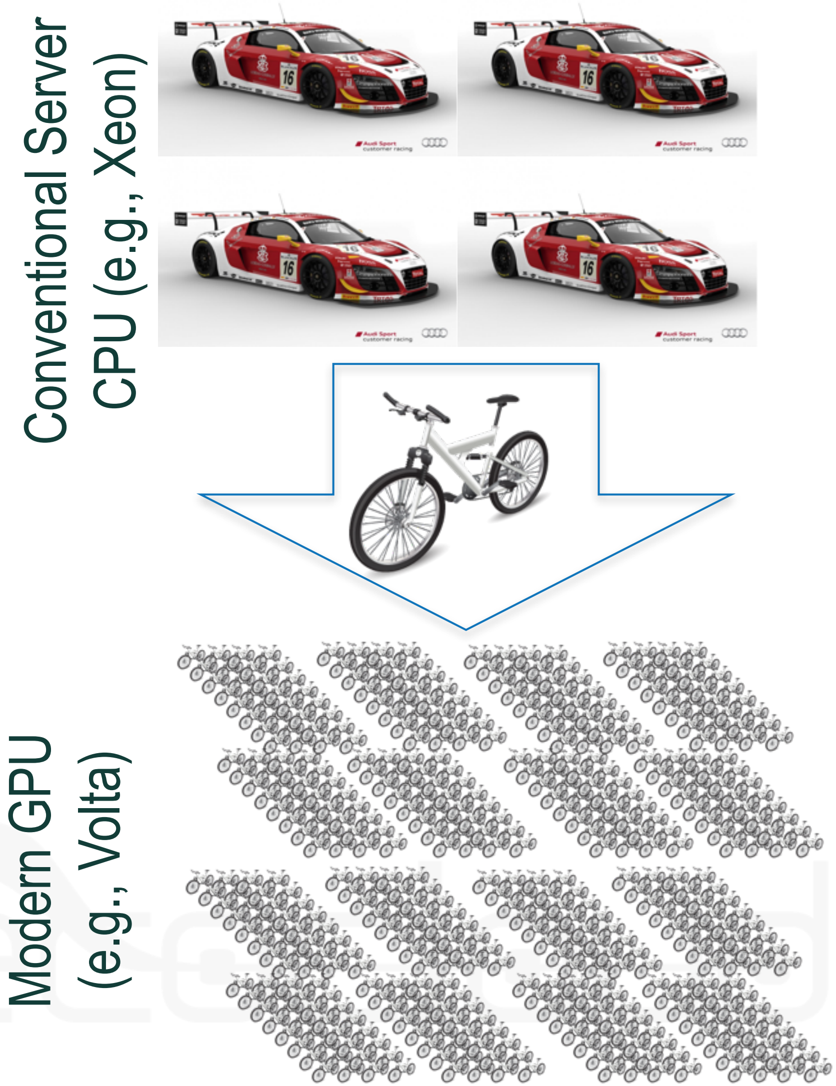
ecocloud
an EPFL research center

- Data parallelism
- Higher memory b/w

Super simple cores

- Shared front end
- 10x slower clocks

Great for dense parallel computation



Parallelism Alone Can't Help

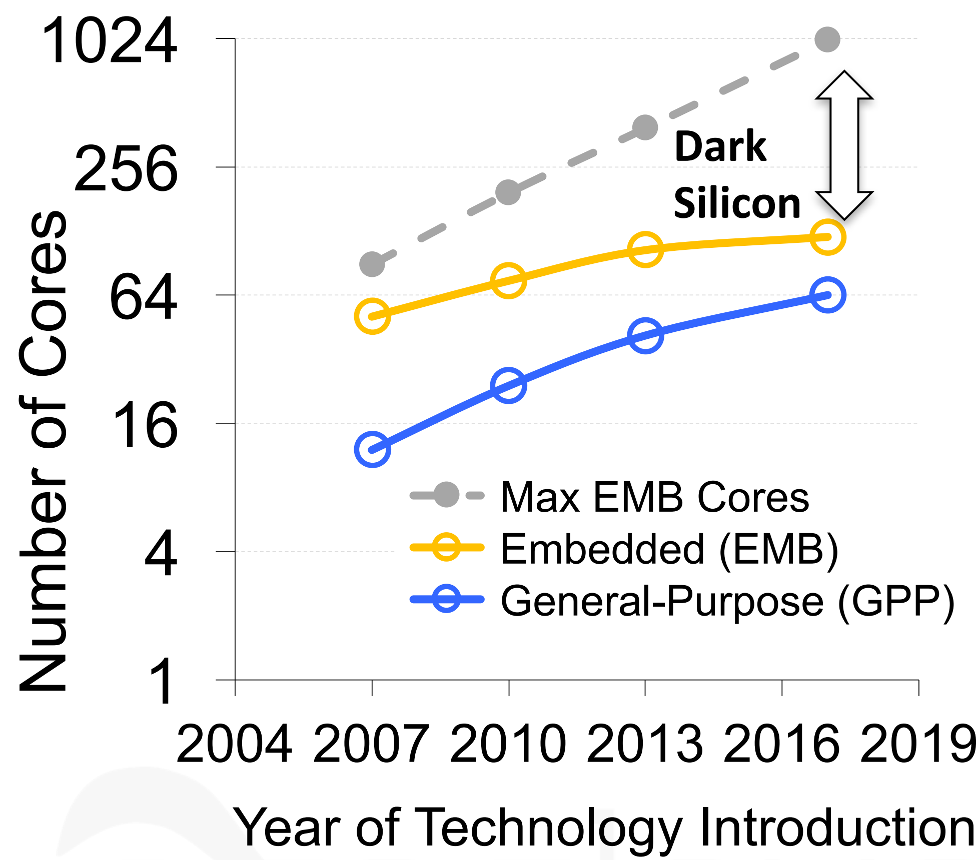
Can populate chips

But, can not operate all

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

All future platforms will be heterogeneous

- Selectively activate parts



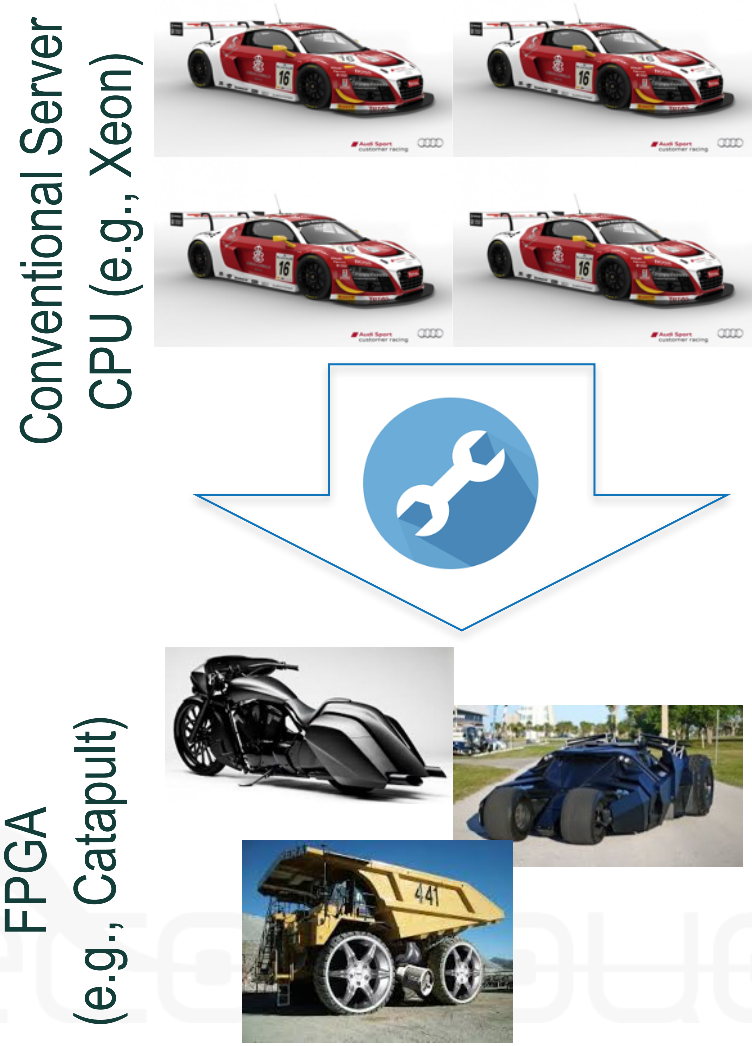
[source: Hardavellas et. al., "Toward Dark Silicon in Servers", IEEE Micro, 2011]

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

Reconfigurable
- Best for spatial computing
- Not caching/reuse

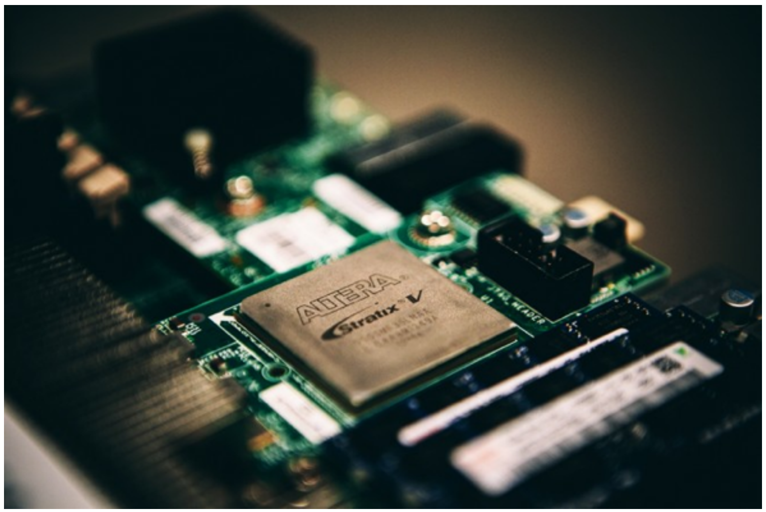
Parallel, dataflow
- 10x slower clocks
- Better for sparse arithmetic

Microsoft, Amazon & Intel



FPGA's in Servers
[MICRO'14]

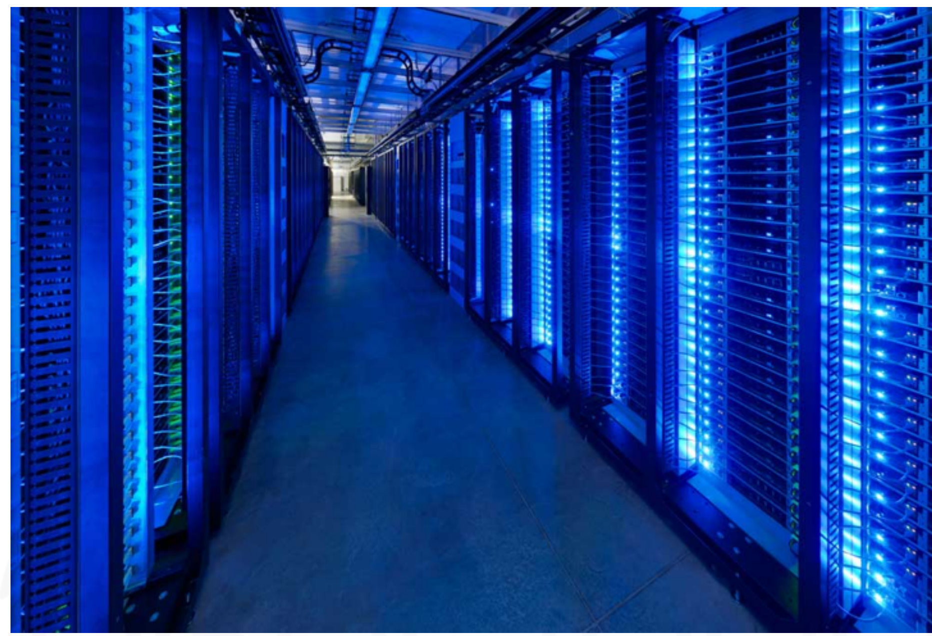
an EPFL research center



Latest version:
- High-end Altera FPGAs
- One FPGA per blade
- Sits on the network
- Backend connected to CPU/NI
- Originally to accelerate Bing, Azure
- Now ML service called BrainWave

Microsoft Unveils Catapult to Accelerate Bing!

[EcoCloud Annual Event, June 5th, 2014]

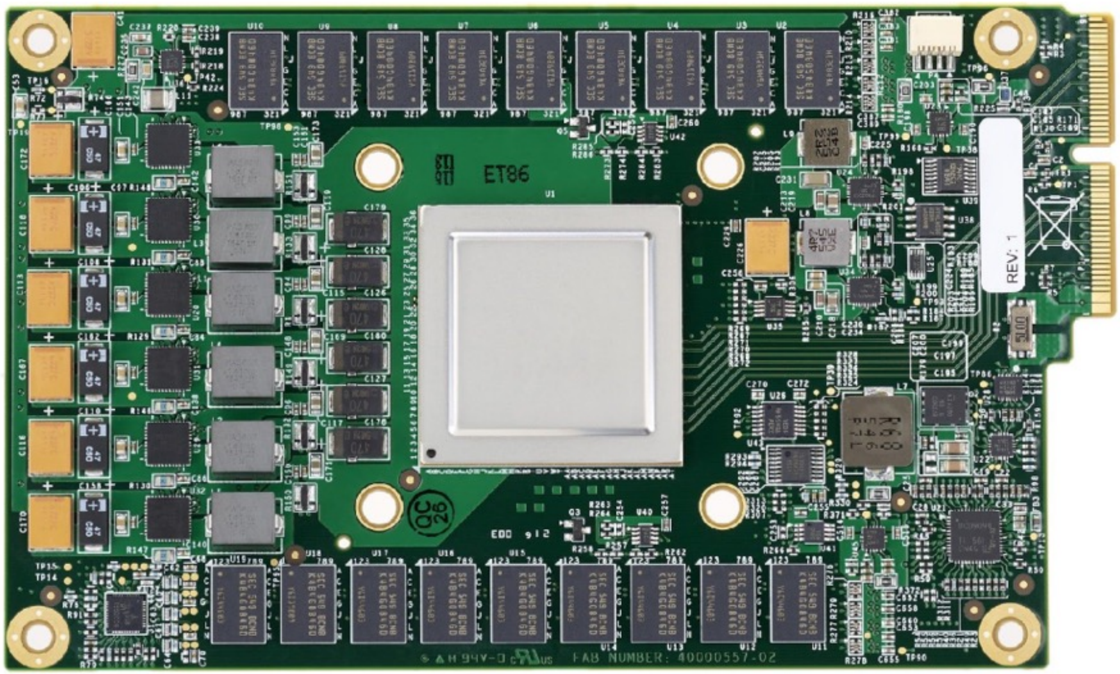


Google's TPU
[ISCA'17]

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
[MICRO'17]

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

Moving Forward: 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

Modern apps/services are statistical

- Analog input, analog output

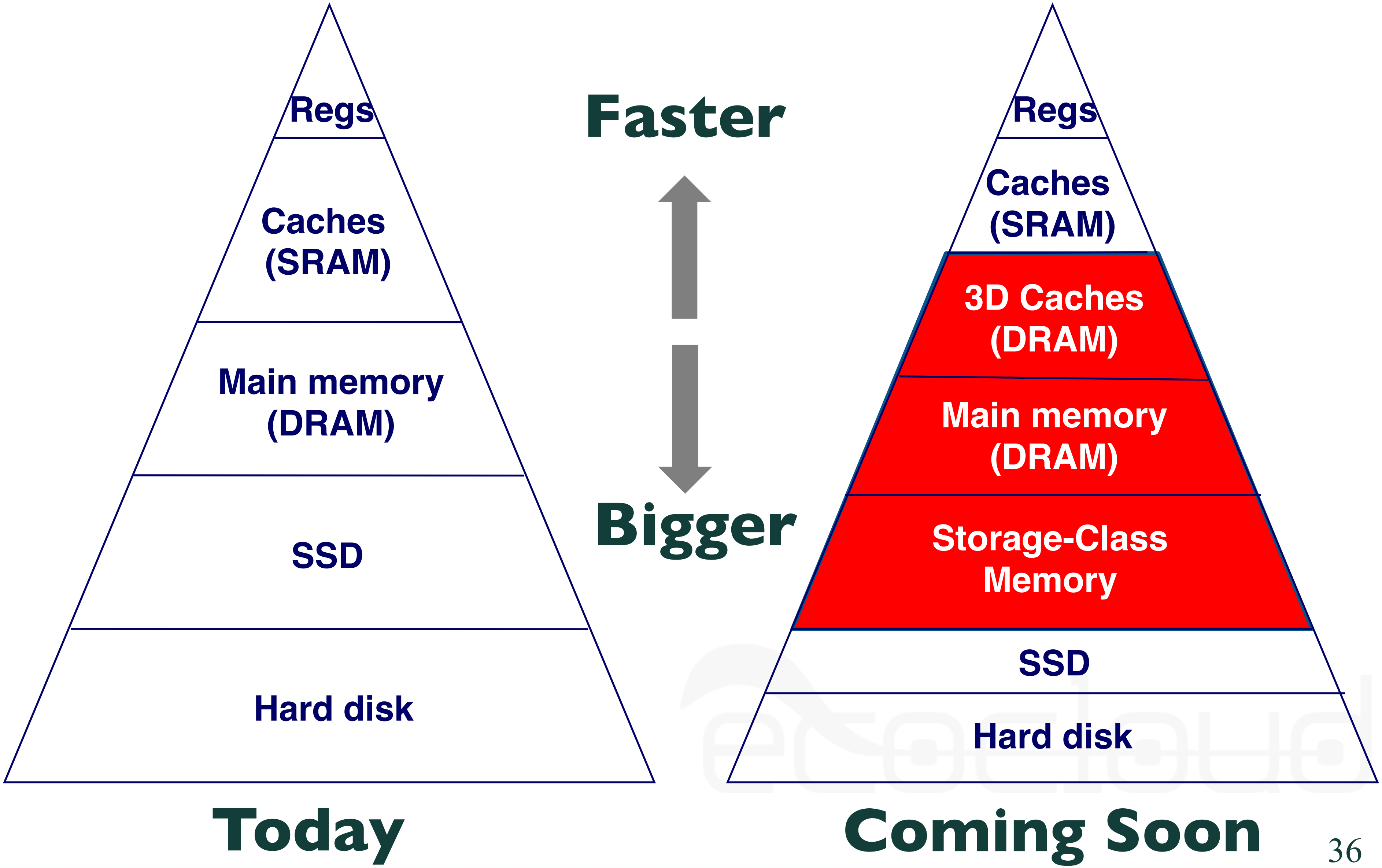
Key:

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

Exploit in

- Processing, communication, storage

Memory Hierarchy



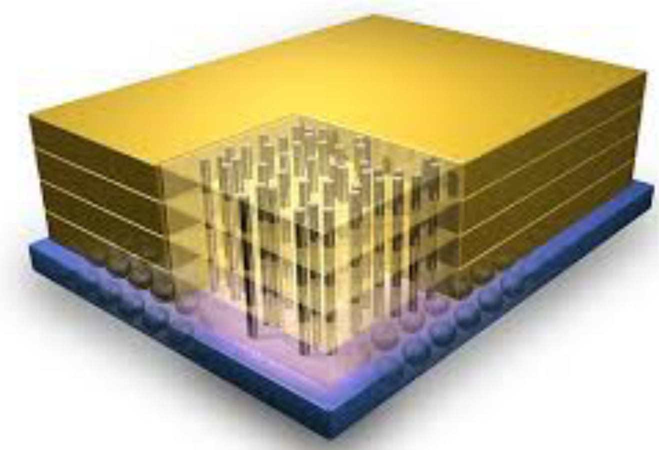
Near-Memory Processing (3D memory) [IEEE Micro'16]

A stack of DRAM with a layer of logic

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

Limitations:

- A few layers of DRAM
- 10x less power in logic
- Uniform thermal envelope



Opportunities for algorithm/hardware co-design

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

ecocloud
an EPFL research center

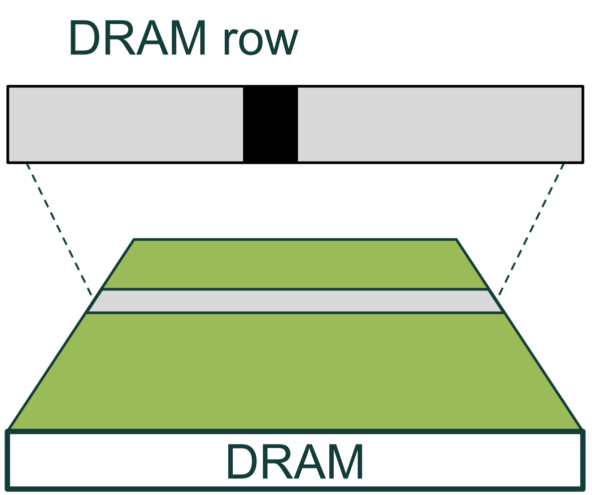
Not (CPU) business as usual

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

Co-design algorithm/HW for NMP!

Why not random access?

Internal DRAM structure dictates
- Activating a 1KB row of data
- Dominates access latency & energy



To exploit bandwidth & efficiency
- Must use most of data in row

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 sequential access!

The Mondrian Data Engine [ISCA'17]

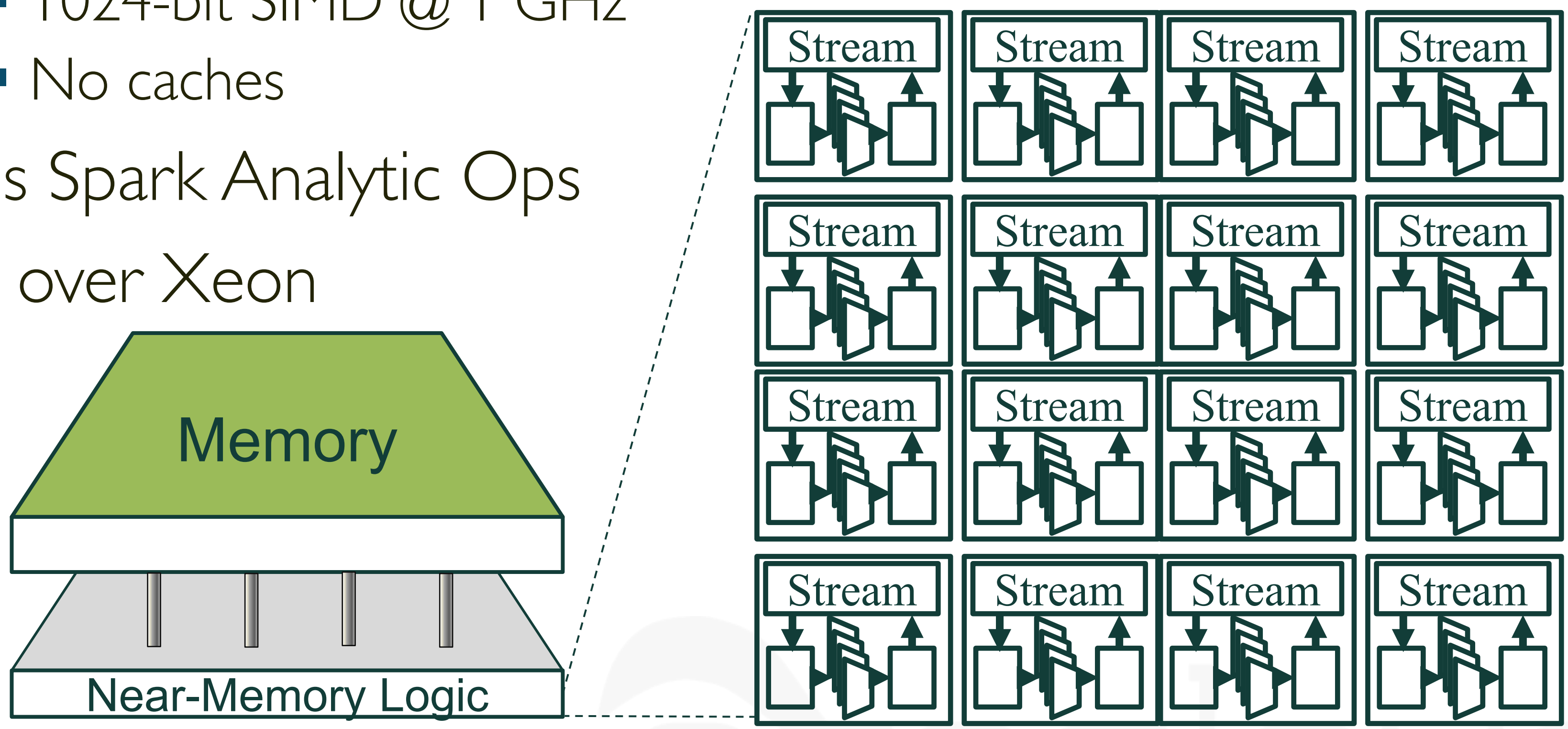
ecocloud
an EPFL research center

SIMD cores + data streaming

- Streams multiple sequential streams
- 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 on Mondrian

Revisiting Sort join [ASBD'14]:

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

Perform way more work

But, finish faster and use less power!

Trade off algorithm complexity for sequential memory accesses

Join 101

Iterates over a pair of tables

Finds the matching keys in two tables

Major operation in data management

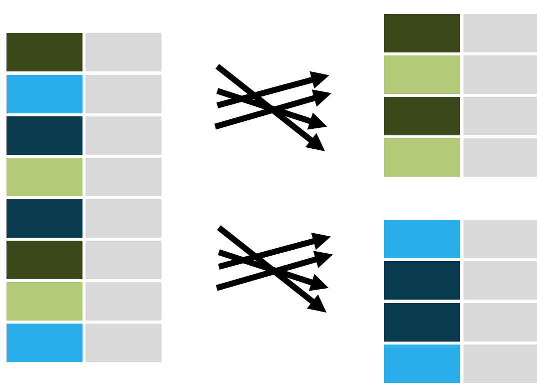
```
Q: SELECT ... FROM R, S WHERE R.Key = S.Key
```

CPU-centric (Hash) Join

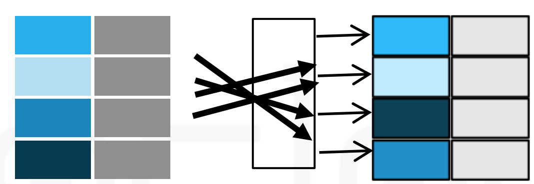
Performed in two phases: Partition & Probe

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

Partition



Probe



Access patterns in hash Join

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

☹: Random access (local or remote)

Comparing access patterns

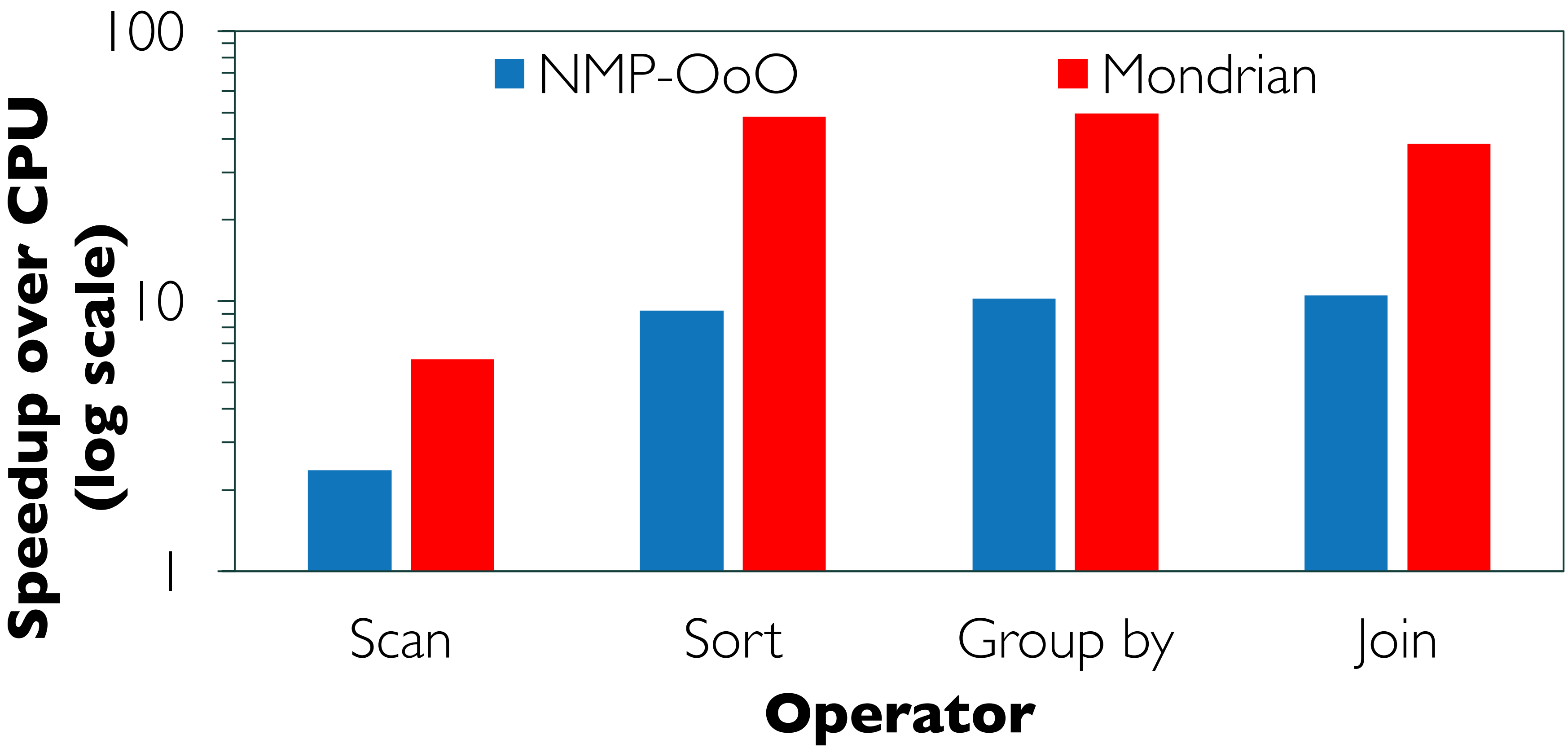
Phases	Hash	Sort
1. Partitioning	☹	☹/😐
2. Build / Sort	☹	☺
3. Probe / Merge	☹	☺

☹: Random access (local or remote)

😐: Sequential access (remote)

☺: Sequential access (local)

Performance



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

X-Stream [SOSP'13]

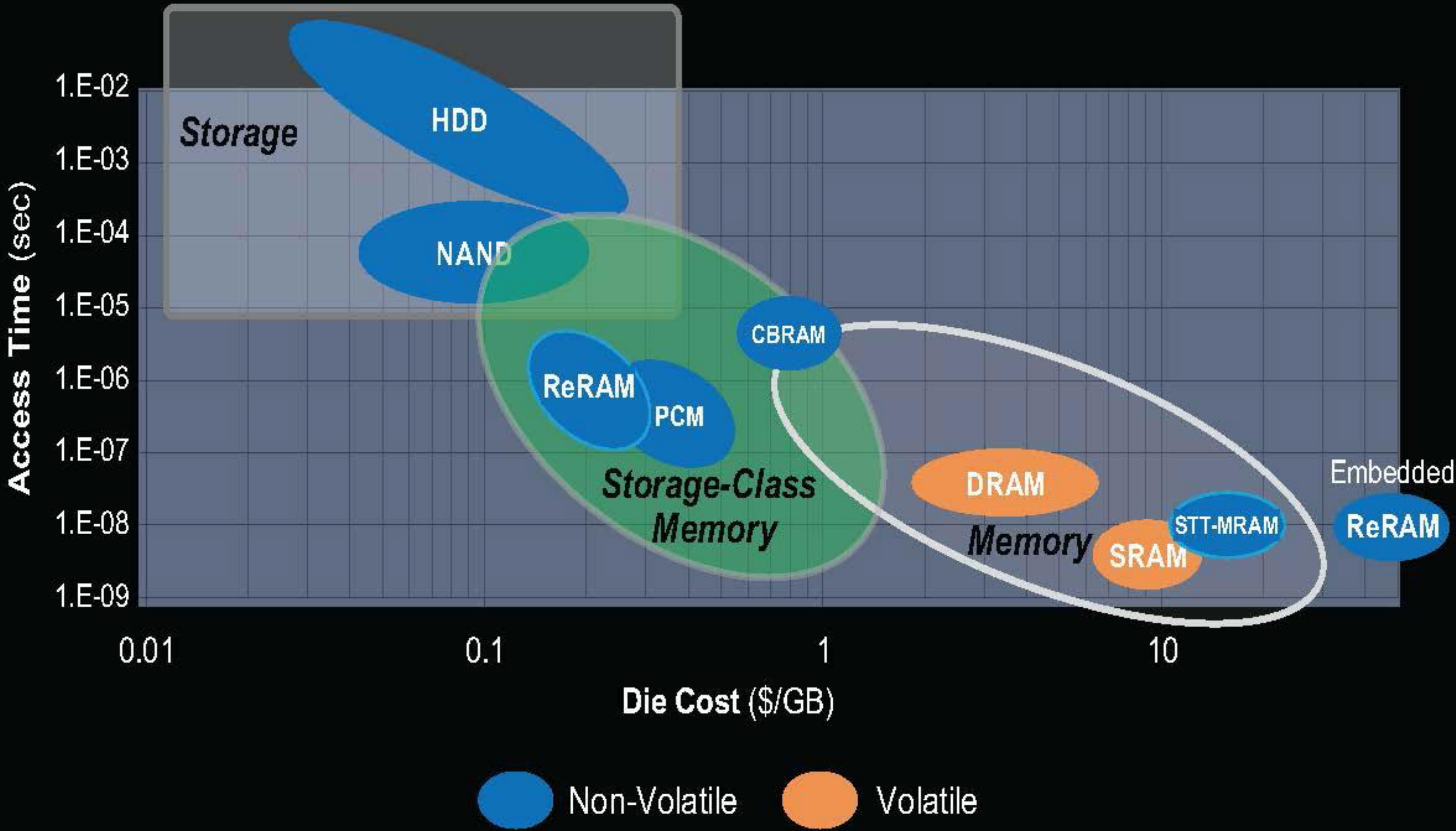
- Graph algorithms without random access
  - Flash, hard disk, …
- Edge-centric rather than vertex-centric
  - Converts random into sequential access

```
for each vertex v
  if v has update
    for each edge e from v
        scatter update along e
```

```
for each edge e
    If e.src has update
        scatter update along e
```

Vertex-Centric ⇒ **Edge-Centric**

Memory & Storage Hierarchy



Storage-Class Memory

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

SCM Algorithms

- Write-efficient databases
  - Favor reads over writes in sorts & join
  - Viglas, et. al., VLDB'14

- (M, ω)-Asymmetric RAM (ARAM)
  - Execute RAM ops on $\Theta(\log n)$-bit words
  - symmetric M words
  - asymmetric unbounded size, write cost ω
  - Gibbons, et. al., SPAA'14'15

Networks

Technology:

- Photonics from racks to boards
- Novel chip-to-chip (wireless)

Abstraction:

- SDN divides control (software) from data (hardware)

Key challenge:

- Inter-cloud exchanges

Challenges Ahead

Digital Sovereignty

ecocloud
an EPFL research center

Microsoft Windows Server System

Yesterday: IT Products

- Bought server & software
- Local usage (in office/building)
- Governed privately
- ✓ Digital Sovereignty

Today+: IT Services

- Cloud services
- Global resources
- Governed by country
- ✗ Loss of Sovereignty

Technologies & legal frameworks to enable transition?

Pros/Cons of Using Cloud

ecocloud
an EPFL research center

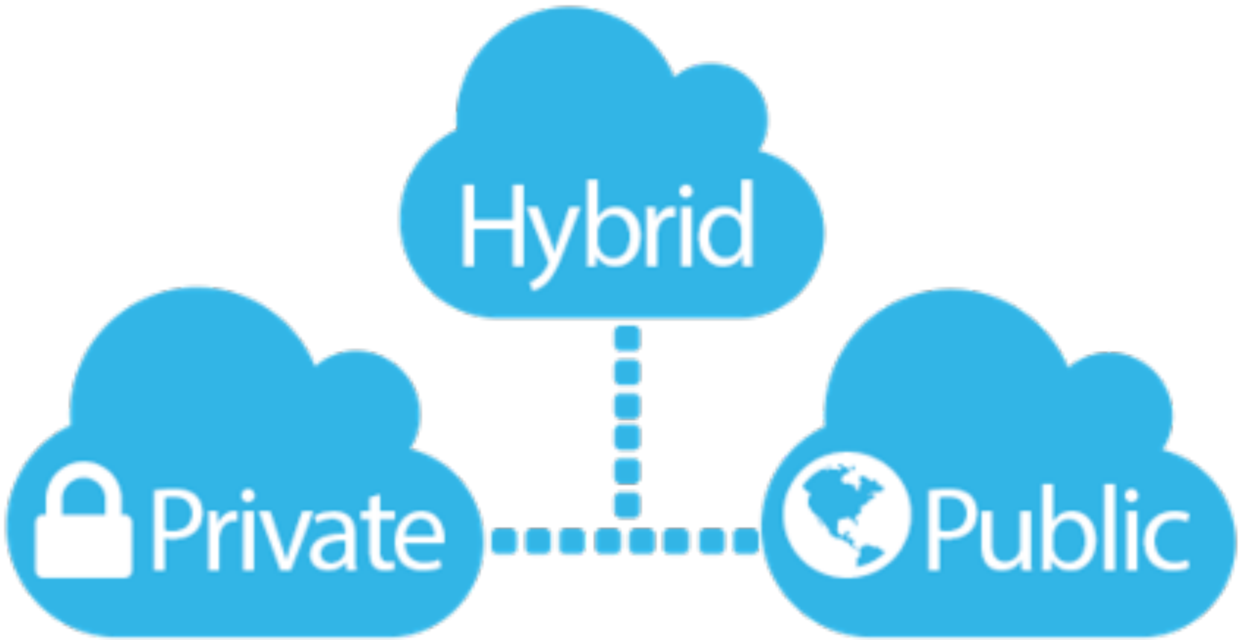
✓Reduced exposure
✓Auditing/testing
✓Automatic management
✓Redundancy
✓Disaster recovery

✗Trusting vendors
✗Accountability
✗Opaque technologies
✗Loss of physical control

source: Peter Mell, Tim Grance, NIST, Information Technology Laboratory

www.nist.gov

Bridging Private/Public Clouds



- Much data is sensitive
- Need algorithms to compute on sensitive data in public
  - E.g., homomorphic analytics, anonymization,..
- Legal frameworks & IT stacks for data hosting services
  - E.g., Government of Luxembourg "Digital Embassy"

Summary

- We live in a Digital Universe
- Clouds are the only path forward
  - Leverage massive data
  - Benefit from economies of scale
- Challenges
  - Scalability no longer comes from technology
  - Need frameworks to guarantee sovereignty
- Future of IT will be about algorithms & data

Thank You!

For more information please visit us at

ecocloud.ch

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE