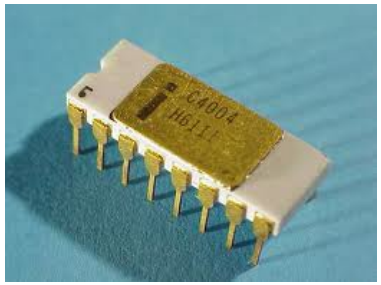
The Future of Data-Centric Computing

Babak Falsafi
ecocloud.ch

Information Technology (IT): Five Decades of Exponential Growth

ecocloud
an EPFL research center

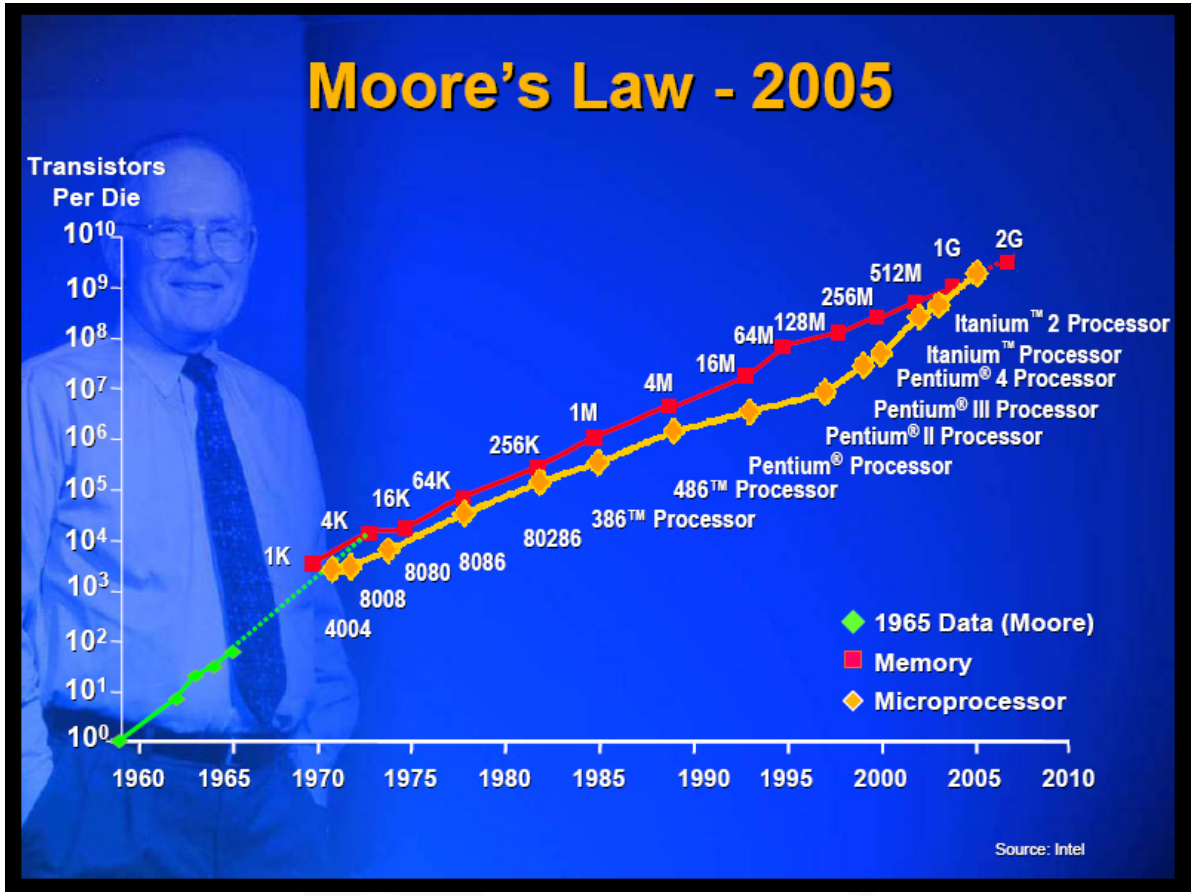
Intel 4004, 1971



92,000 ops/sec

Intel Xeon, 2014

266,000,000,000 ops/sec



IT is at the core everything we do & has become an indispensable pillar for a modern day society!

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

Two Inflection Points for IT

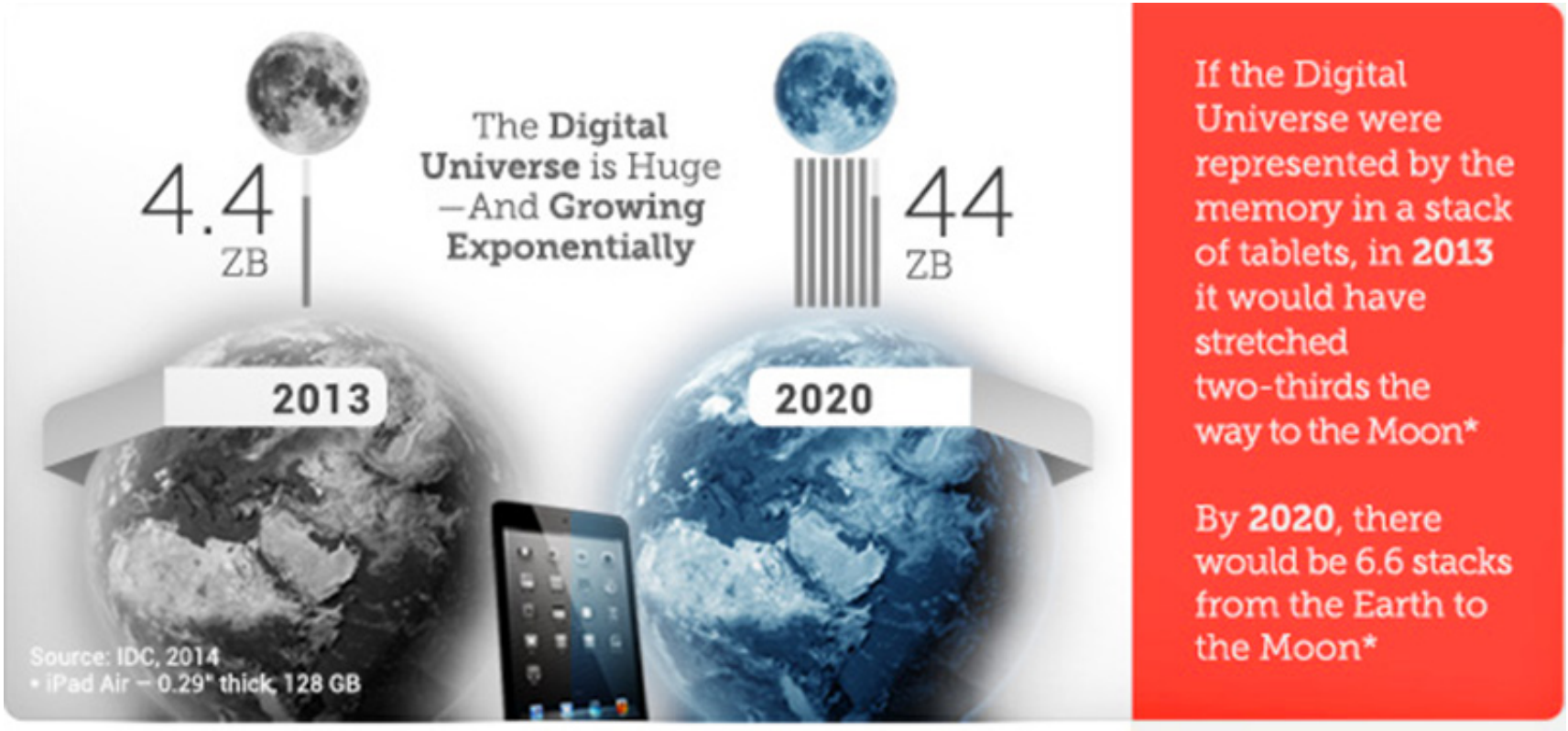
ecocloud
an EPFL research center

1. Big Data
   - Data growth at 1.5x/year
   - Huge demand on scaling platforms

2. Platforms
   - Efficiency scaling stopped ten years ago
   - Silicon density scaling has slowed down
   - IT growth not sustainable

Data is Shaping Future of IT

ecocloud
an EPFL research center



- Data growth (by 2015) = 100x in ten years [IDC 2012]
  - Population growth = 10% in ten years
- Monetizing data for commerce, health, science, services, ….
- Big Data is shaping IT & pretty much whatever we do!

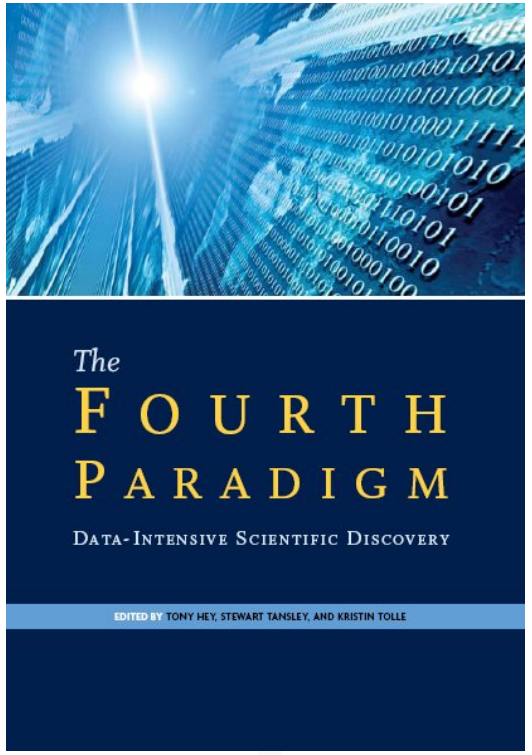
Data Shaping All Science & Technology

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!

Datacenters Growing Fast

ecocloud
an EPFL research center

Source: James Hamilton, 2012

Daily IT growth in 2012 = IT first five years of business!

amazon web services

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 one day/downtime)
  - ……and lots more

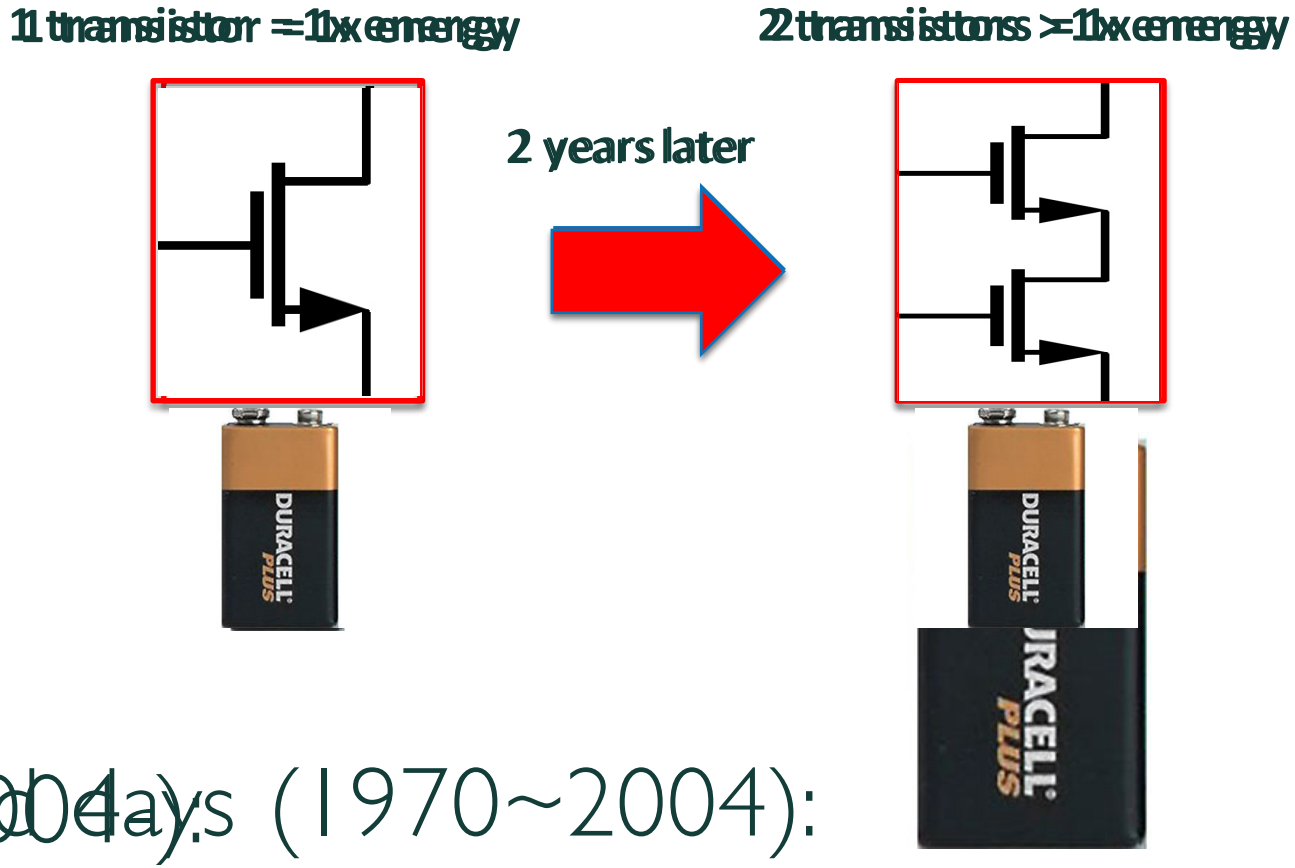
Datacenters are the IT utility plants of the future

Supercomputing

≠

Cloud Computing

But, platforms are not scaling



Now (2004):

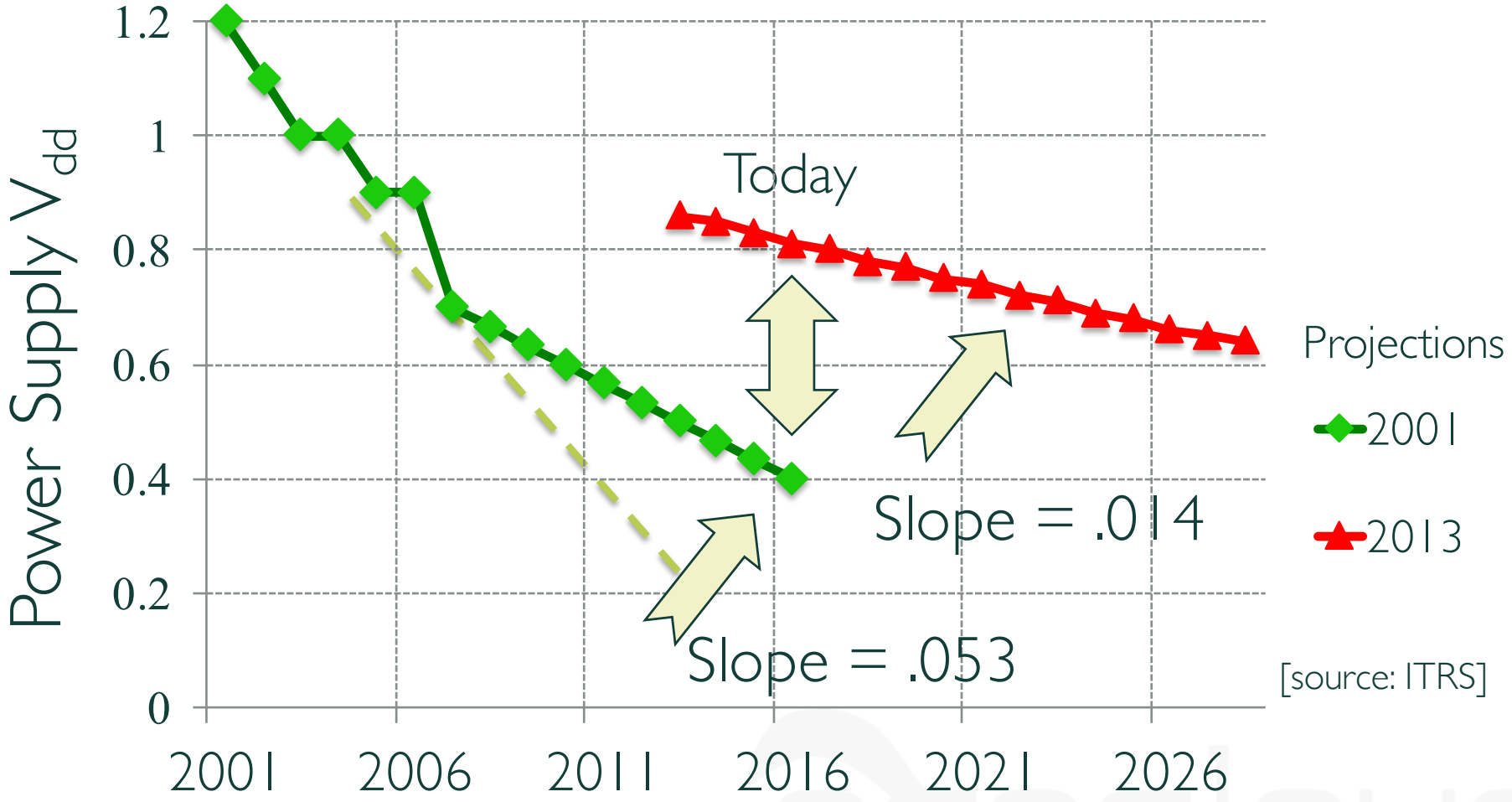
- Continue to make transistors smaller
- But, they use similar electricity to operate
- Chip energy consumption is shooting up

Good old days (1970~2004):

- Used to make transistors smaller
- Smaller transistors less electricity to operate
- Chip energy consumption remained same

Operating voltages leveling

ecocloud
an EPFL research center



The fundamental energy silver bullet is gone!

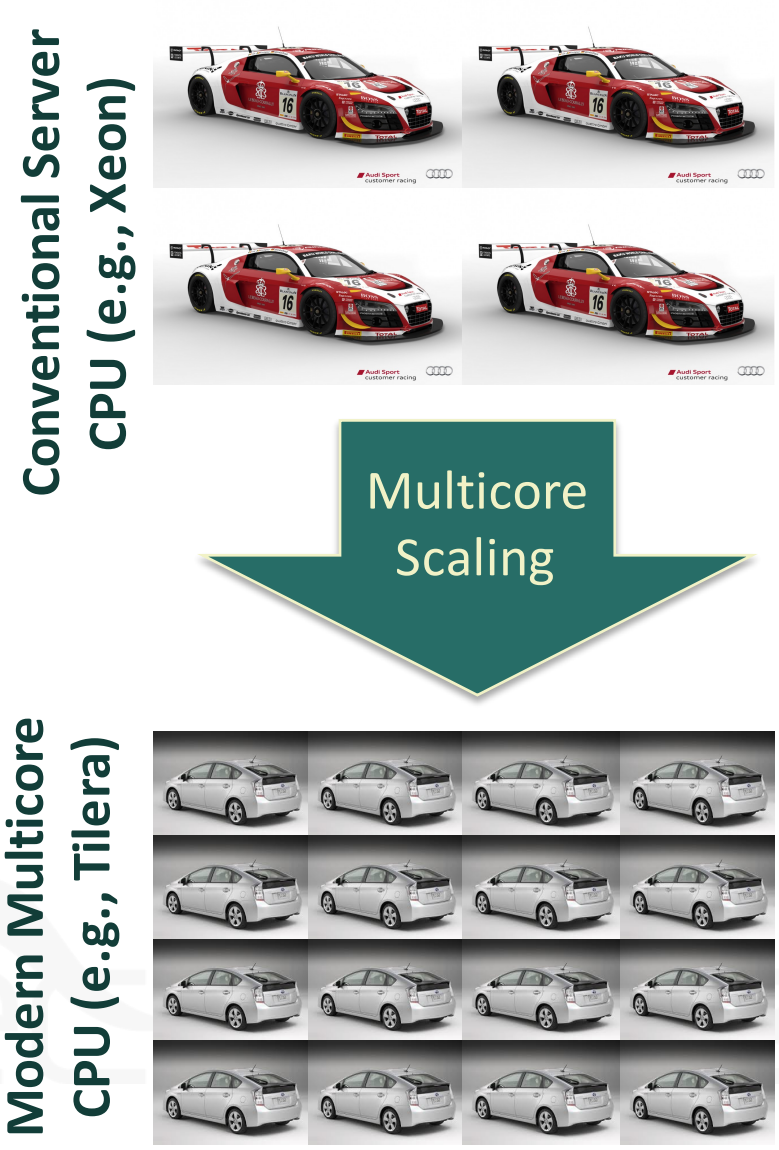
The Rise of Parallelism to Save the Day

ecocloud
an EPFL research center

With voltages leveling:

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

  fewer joules/op



The Rise of Dark Silicon: End of Multicore/GPU Scaling

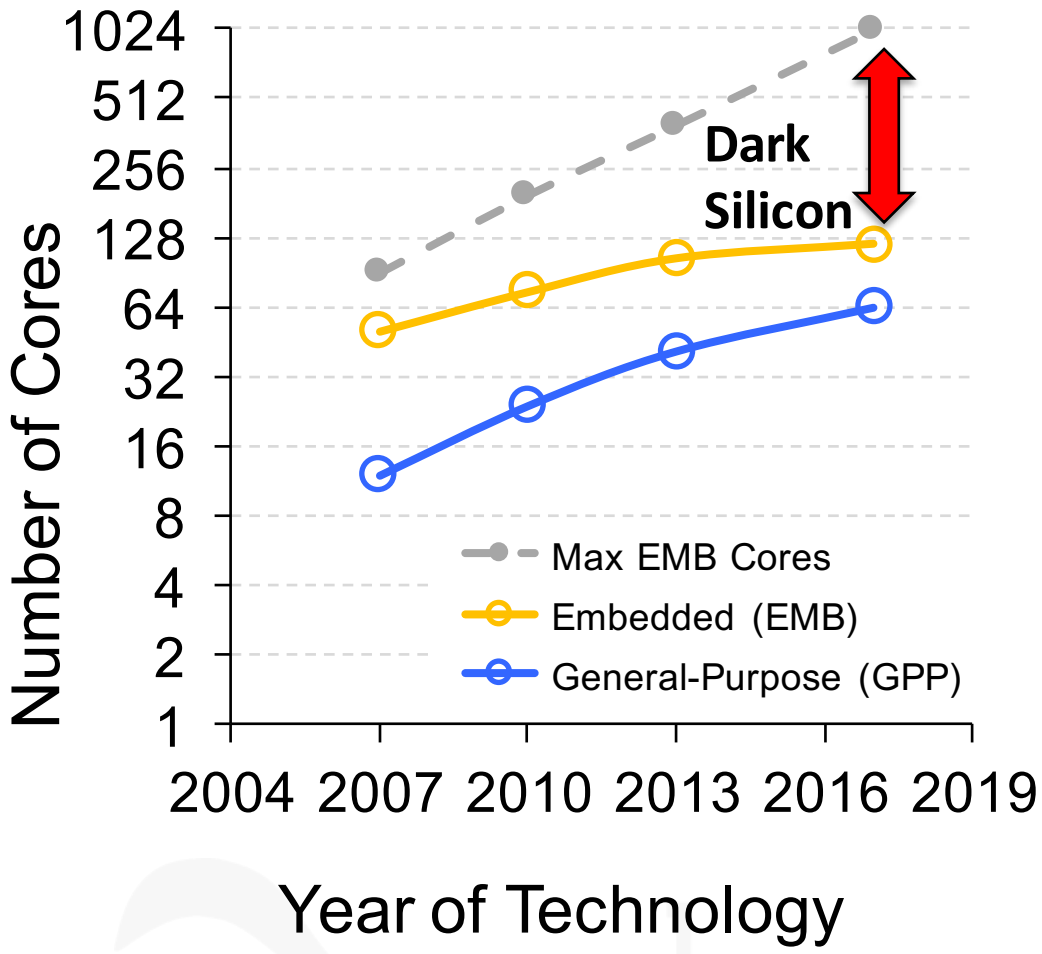
ecocloud
an EPFL research center

But parallelism can not offset leveling voltages

Even in servers with abundant parallelism

Core complexity has leveled off too!

Need a holistic approach



Hardavellas et. al.

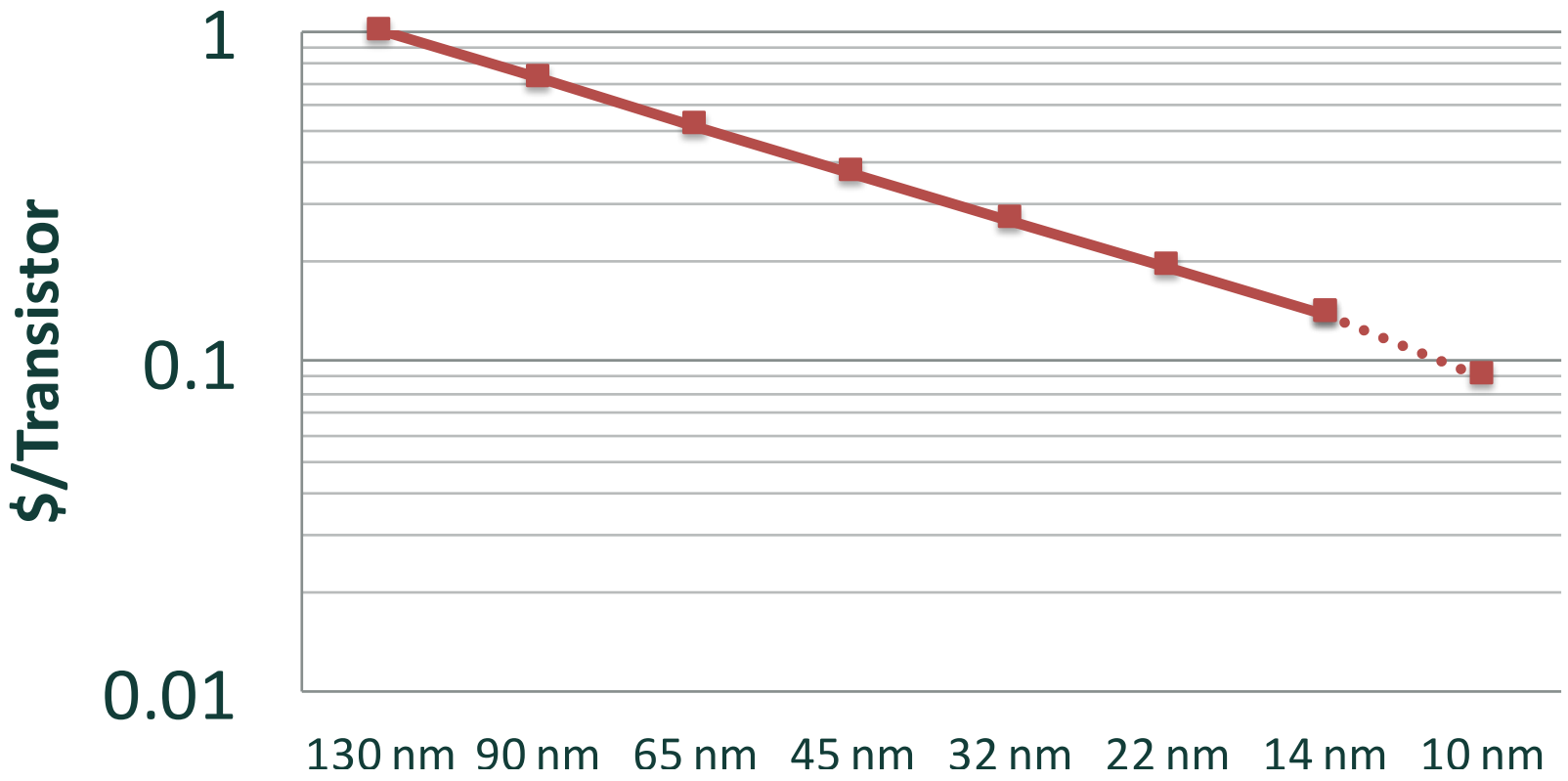
Toward Dark Silicon in Servers

IEEE Micro, 2011

Silicon not getting much denser….

ecocloud
an EPFL research center

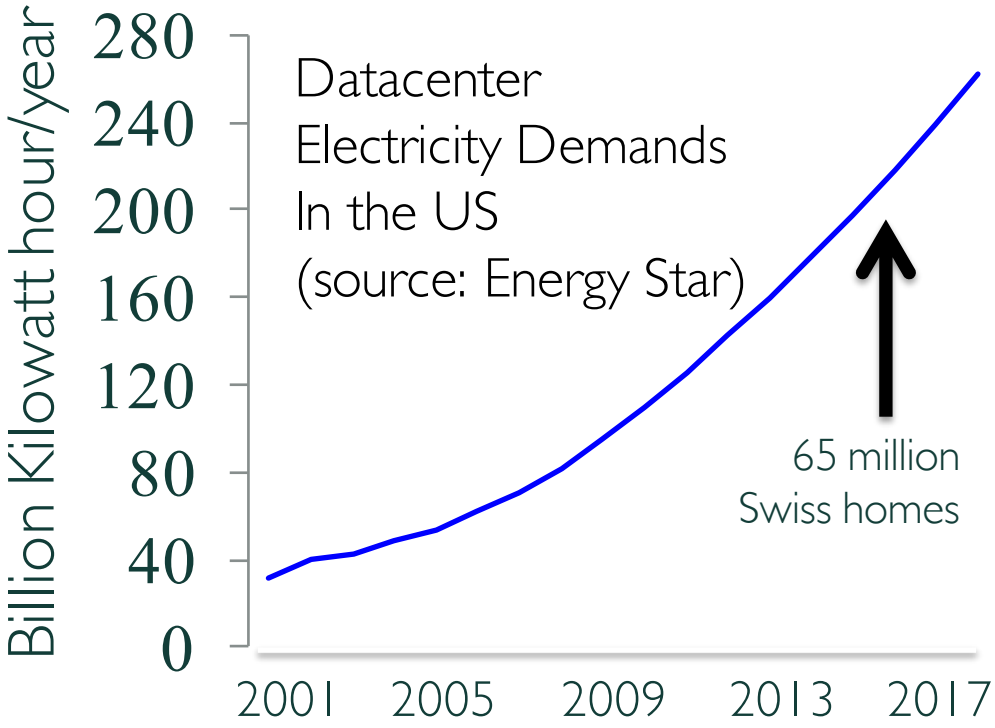
Mark Bohr's (Intel) Keynote [ISSCC'15]

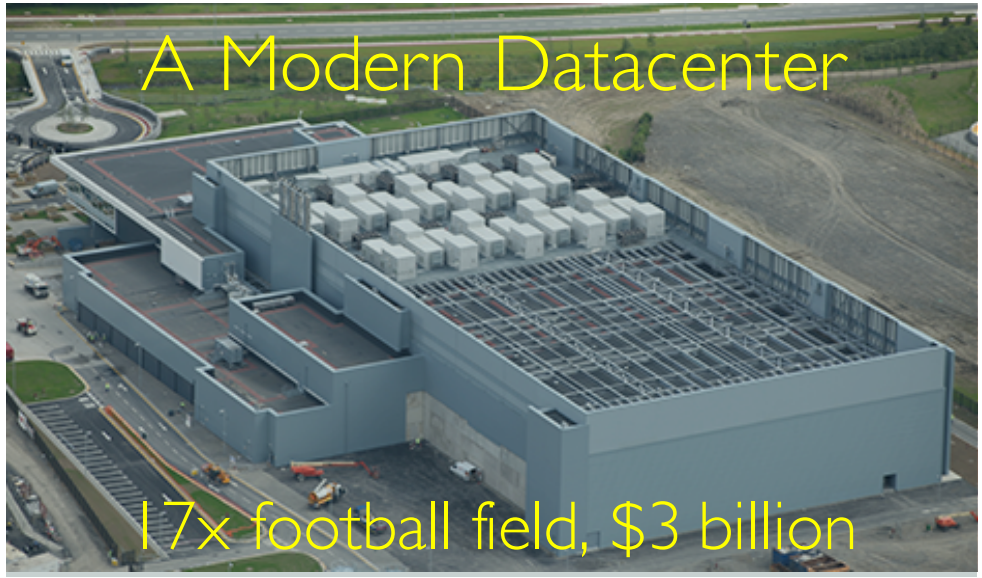


Moore's Law: $/transistor dropping for fifty years

- Intel is pushing for a bit more
- TSMC hinted at prices going up for the first time in 2014

Higher Demand + Lower Efficiency: Datacenter Growth not Sustainable!





- Modern datacenters ➔ 20 MW!
- In modern world, 6% of all electricity and growing fast!

ecocloud
an EPFL research center

Big Data

IT's Future

Bridging Technologies

Big Energy

ecocloud

Center to bring efficiency to data

- 16 faculty, 50 researchers
- 6M CHF/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

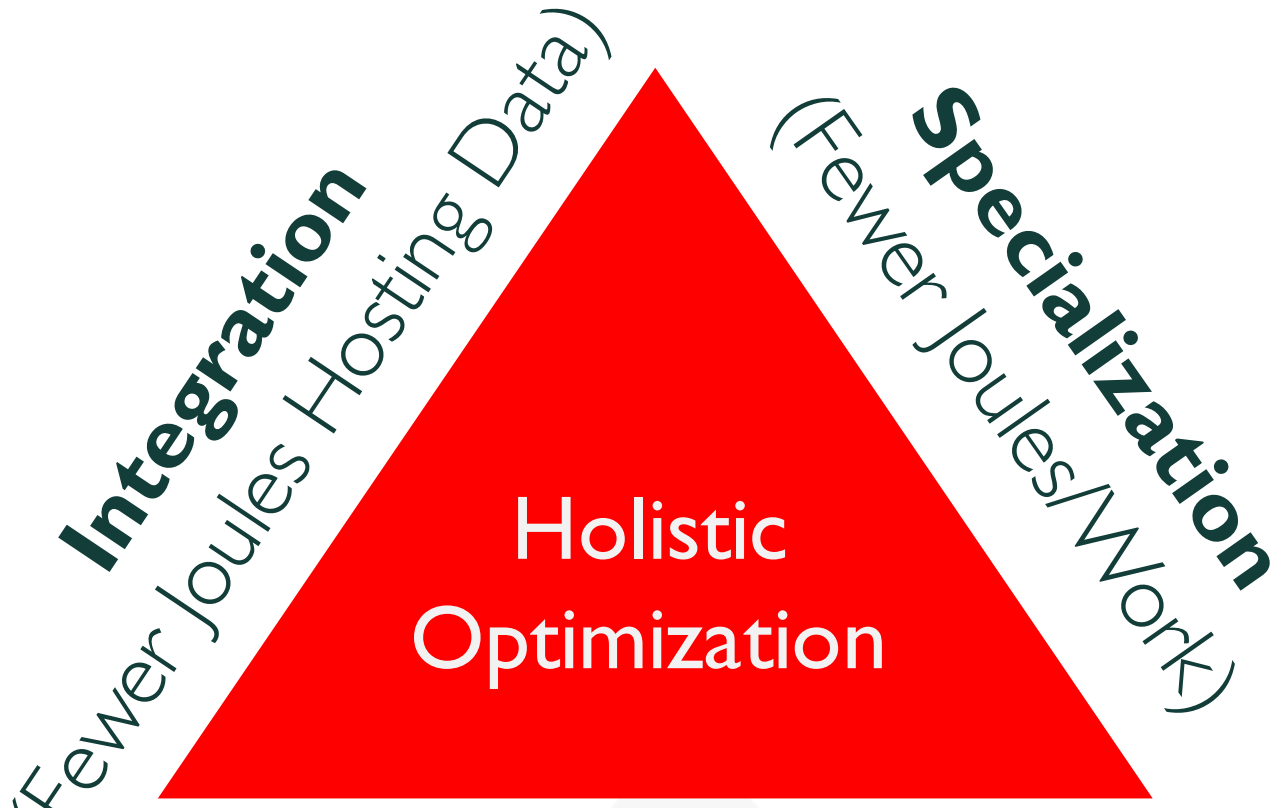
Holistic optimization

- Algorithms to infrastructure
- Cross-layer
- Paradigms to monitor, manage & reduce energy

Open technologies!

Our Vision:
The ISA Triangle of Efficiency

ecocloud
an EPFL research center



Integrated Cooling: CMOSAIC

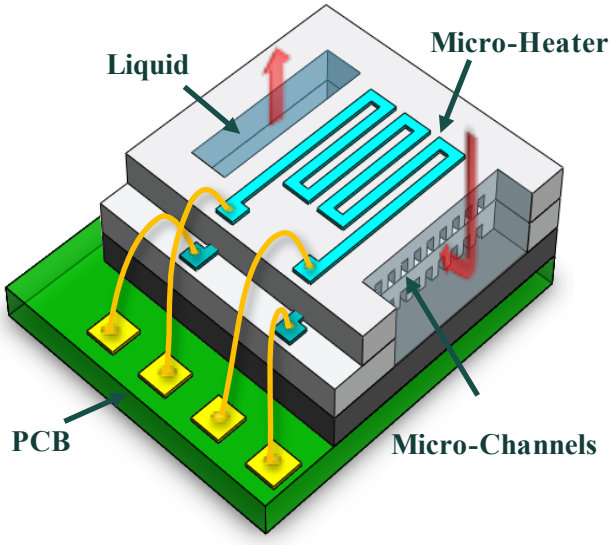
ecocloud
an EPFL research center

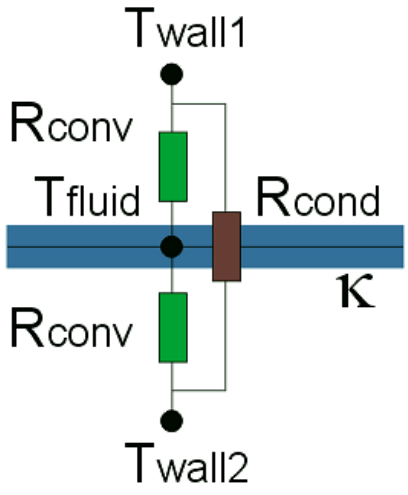
3D server chip

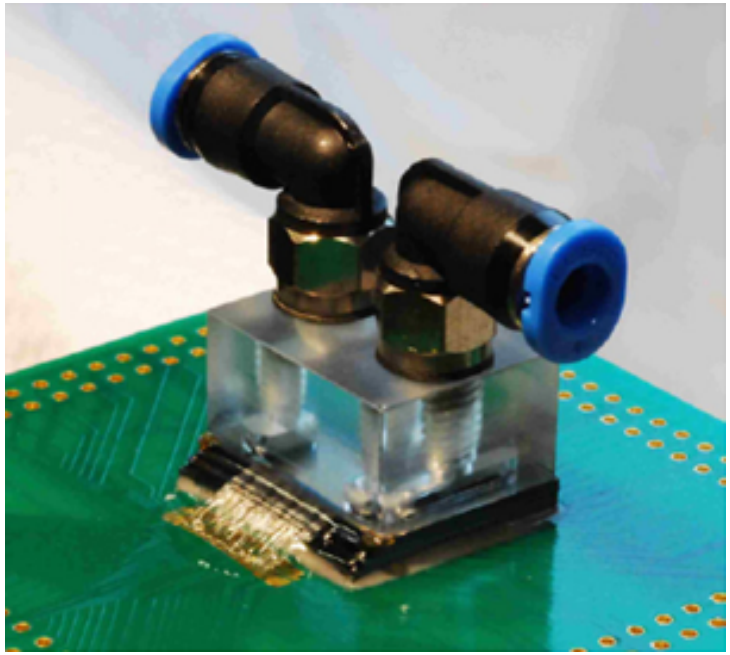
Two-phase liquid cooling

- Enables higher thermals
- Dramatically better heat removal

Prototyped by IBM







IBM

Integrated Compute in Memory

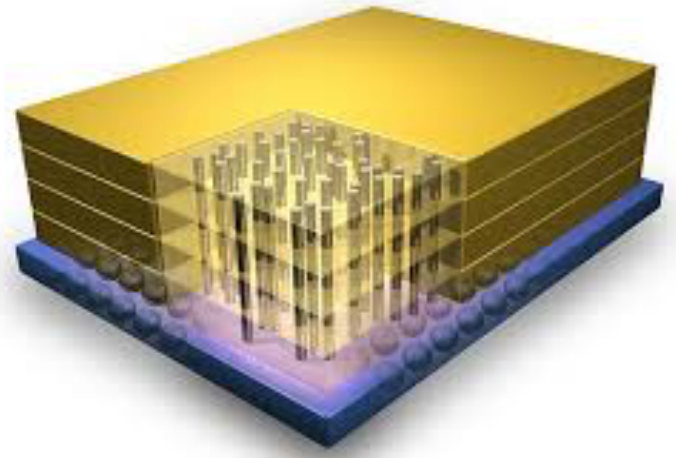
ecocloud
an EPFL research center

Why in-memory?

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

Basic data services:

- Scan, Join, GroupBy, Filter
- Best for sequential access
- Accelerators must co-exist with conventional memory semantics



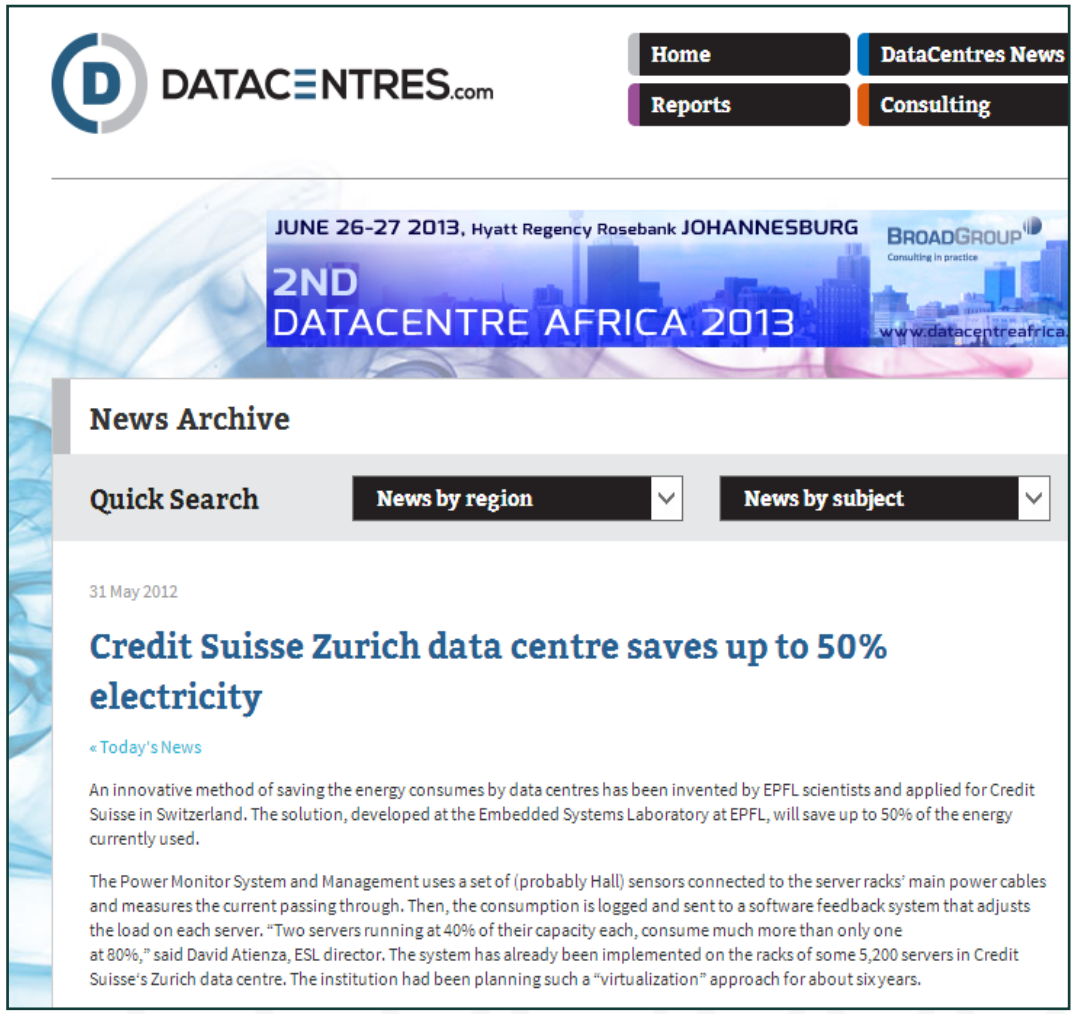
10x better efficiency for a database join operation!

Integrated Thermal Management @ Credit Suisse

ecocloud
an EPFL research center

- Designed fine-grain power/thermal sensors
- Deployed with real-time monitoring of 5K servers
- Software for synergistic load/thermal provisioning

50% better efficiency!



DATACENTRES.com

Home | DataCentres News | Reports | Consulting

JUNE 26-27 2013, Hyatt Regency Rosebank JOHANNESBURG
2ND DATACENTRE AFRICA 2013
BroadGroup Consulting in practice
www.datacentreafrica

News Archive

Quick Search | News by region | News by subject

31 May 2012

Credit Suisse Zurich data centre saves up to 50% electricity

« Today's News

An innovative method of saving the energy consumes by data centres has been invented by EPFL scientists and applied for Credit Suisse in Switzerland. The solution, developed at the Embedded Systems Laboratory at EPFL, will save up to 50% of the energy currently used.

The Power Monitor System and Management uses a set of (probably Hall) sensors connected to the server racks' main power cables and measures the current passing through. Then, the consumption is logged and sent to a software feedback system that adjusts the load on each server. "Two servers running at 40% of their capacity each, consume much more than only one at 80%," said David Atienza, ESL director. The system has already been implemented on the racks of some 5,200 servers in Credit Suisse's Zurich data centre. The institution had been planning such a "virtualization" approach for about six years.

CREDIT SUISSE

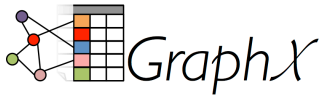
Cloud Benchmarking with CloudSuite 3.0 (parsa.epfl.ch/cloudsuite)

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!

Specialized Servers: Cavium ThunderX

MICROPROCESSOR report

Insightful Analysis of Processor Technology

THUNDERX RATTLES SERVER MARKET

Cavium Develops 48-Core ARM Processor to Challenge Xeon

By Linley Gwennap (June 9, 2014)

EE Times Connecting the Global Electronics Community

BREAKING NEWS SLIDESHOW: CES: Bosch Aims to Connect Whole World

designlines WIRELESS & NETWORKING

News & Analysis

Big-Data Benchmark Brewing

EEMBC works on SoC-agnostic spec

Rick Merritt

10/15/2014 08:00 AM EDT

SAN JOSE, Calif. — A new benchmark suite for scaled-out servers is in the works with the first piece of it expected early next year. The processor-agnostic metrics aim to set standards for measuring today's data center workloads.

A new cloud and big-data server working group of the Embedded Microprocessor Benchmark Consortium (EEMBC) hopes to deliver a suite of seven benchmarks. It aims to complete before April three of them -- memory caching, media serving, and graph analysis.

"Typically when we go to a server customer they ask for SpecInt numbers, that's been the traditional benchmarks for servers for a long time, but SpecInt is not a very good metric for distributed data loads or available instruction and memory parallelism," said Bryan Chin, a distinguished engineer from Cavium.

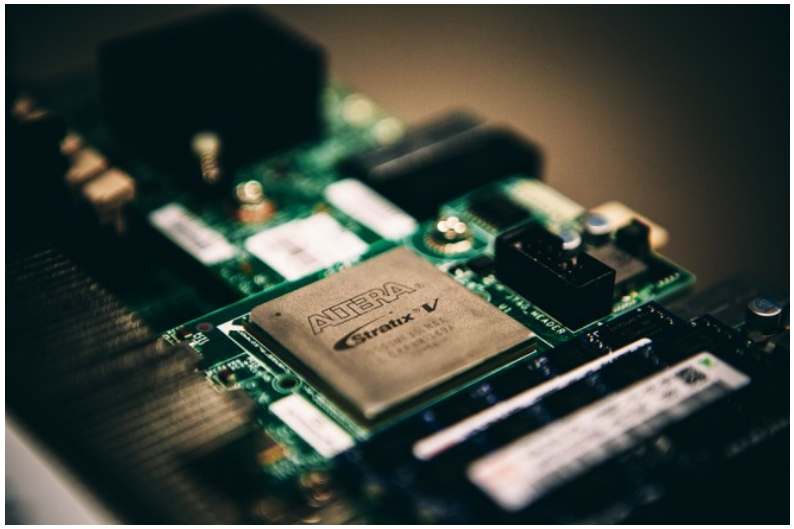
48-core 64-bit ARM SoC

[blueprinted at EPFL]:

- Designed to serve data
- Custom organization
- Runs off-the-shelf software stack
- 10x better faster than Xeon

CAVIUM NETWORKS

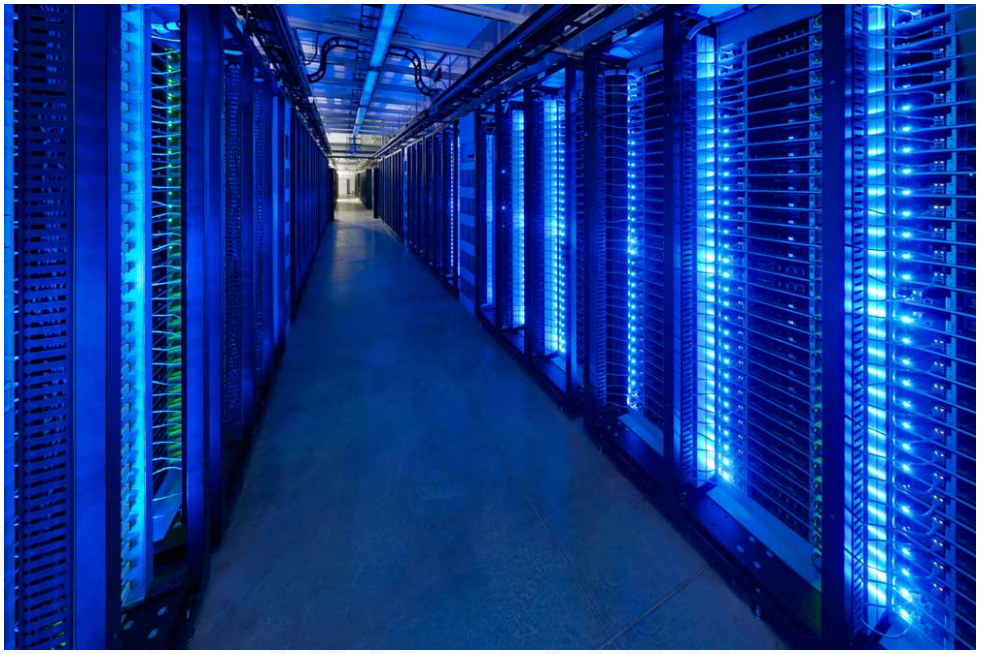
Spatial Computing: An idea whose time has come



- One StratixV FPGA per blade
- DSL's (e.g., Scala, TensorFlow)
- Analytics with spatial computing

Microsoft Unveils Catapult to Accelerate Bing!

[EcoCloud Annual Event, June 5th, 2014]



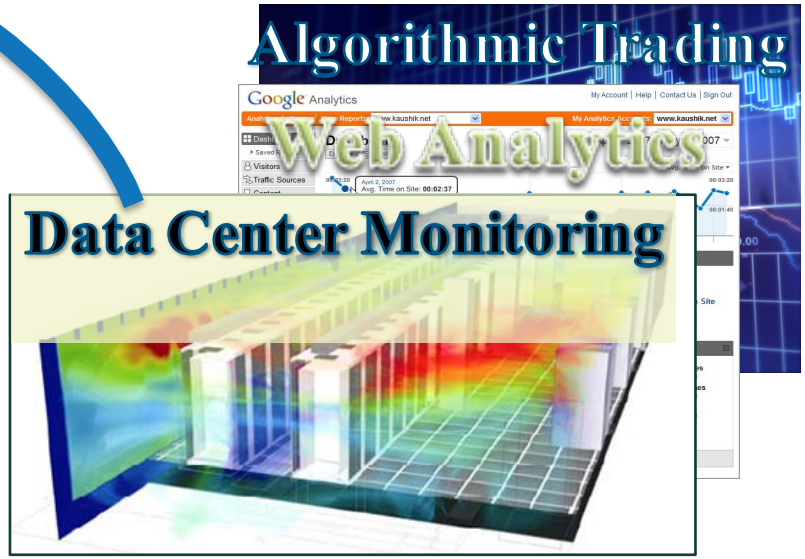
Specialized Databases: Breaking Up Data Jams

ecocloud
an EPFL research center

Compilation of offline analytics into online/incremental engines

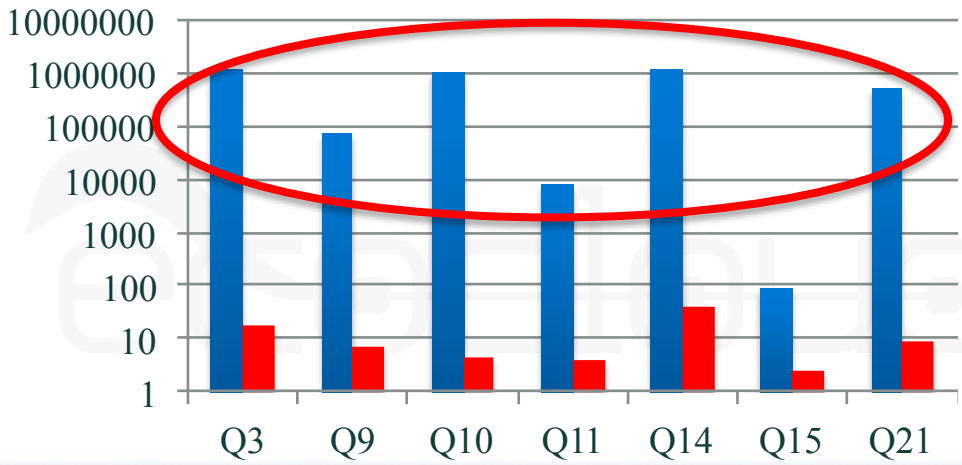
Aggressive code specialization

Data Stream



Low-latency stream processing

Up to 6 OOM faster than commercial systems



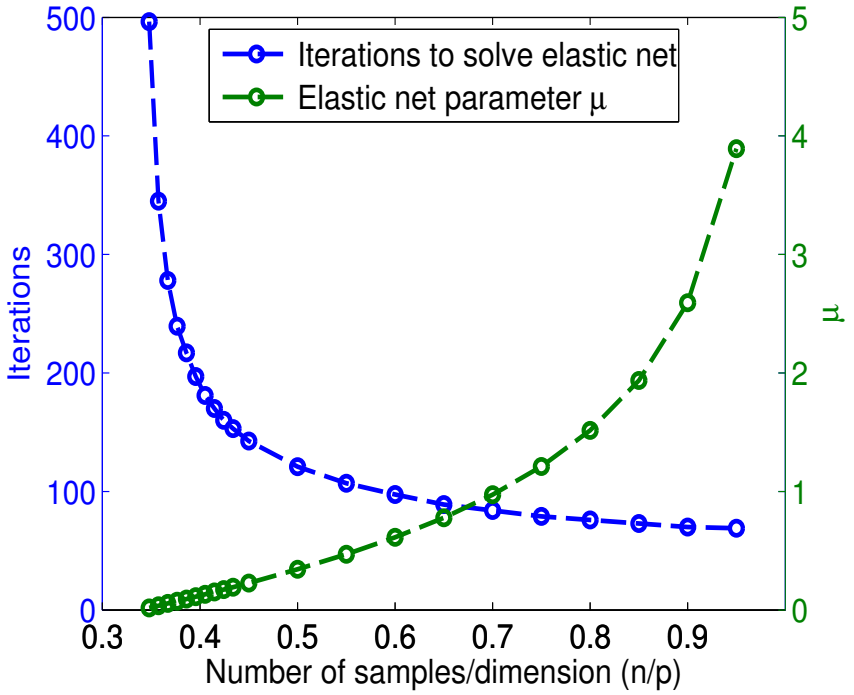
dbtoaster.org

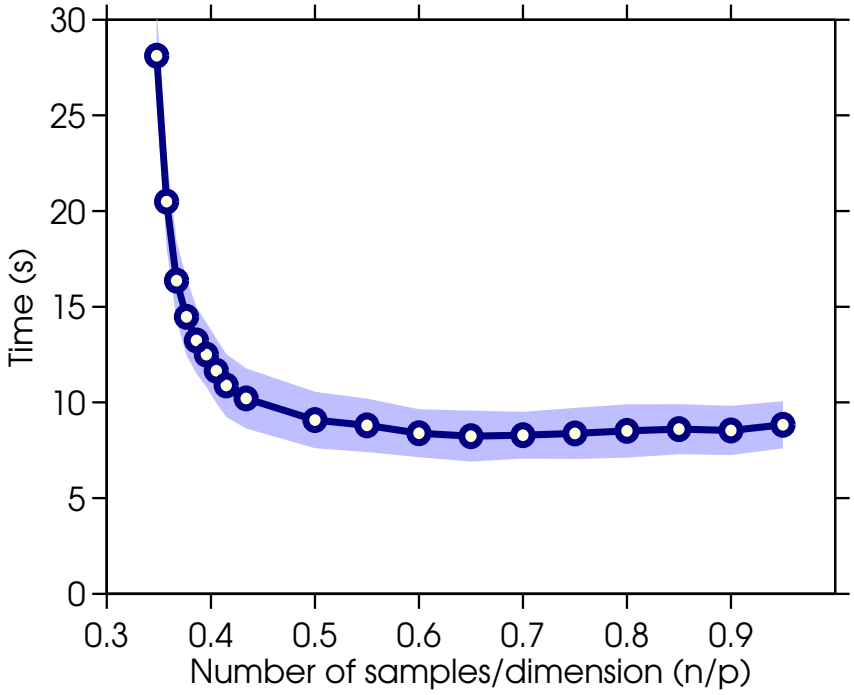
Big Data Analytics

ecocloud
an EPFL research center

Convex machine learning models

- Broad set of applications: sparse SVM's, low-rank matrix completions...
- Bigger data ➔ faster algorithms for the same statistical risk!





Algorithms to simplify problem as data size gets larger!

Bringing it All Together

ecocloud
an EPFL research center

Two inflection points for IT

- Data growing at an exponential rate
- Platform scaling is a grand challenge
- Future IT platforms designed around data

Holistic approaches to datacenter design

Integrate + Specialize + Approximate (ISA)

Thank You!

ecocloud
an EPFL research center

For more information please visit us at

ecocloud.ch

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

7 Giants of Analytics
[National Academies, 2013]

ecocloud
an EPFL research center

- Basic statistics
- *N*-body problem
- Graph theory
- Linear algebra
- Generalized Optimization
- Computing integrals
- Alignment problems