Computing (on Massive Data) with Dark Silicon

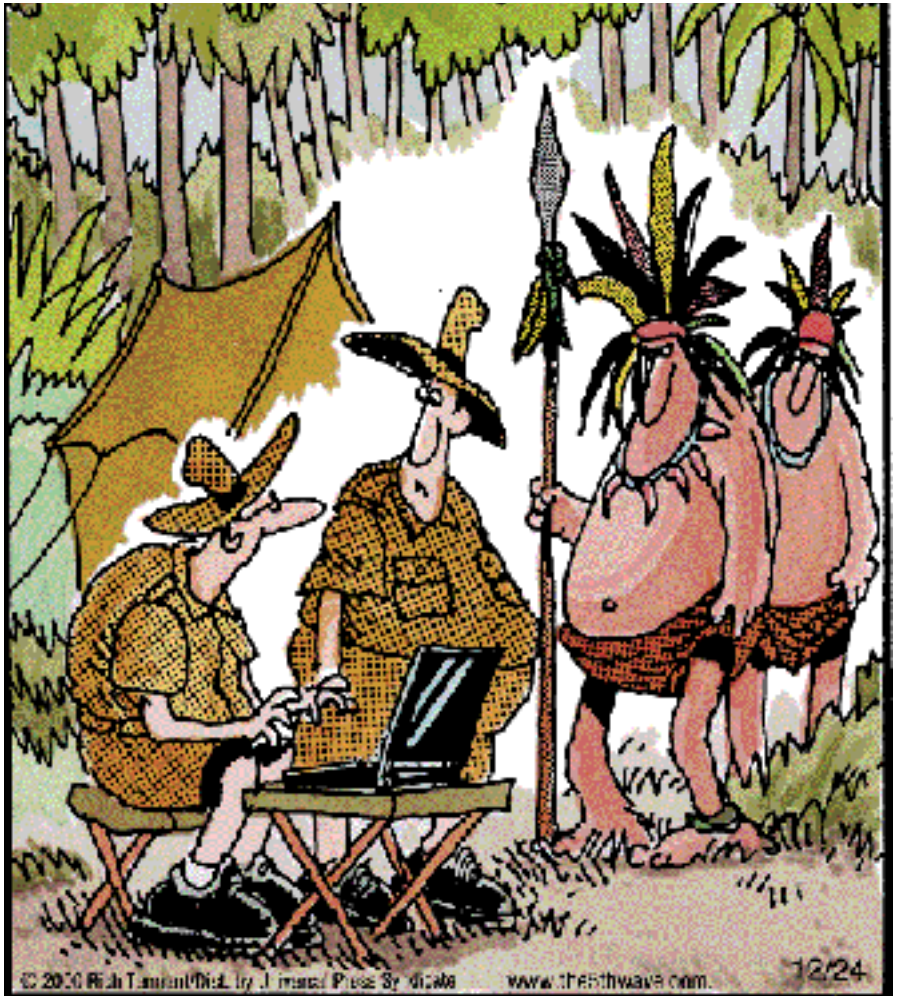
Babak Falsafi
Director, EcoCloud
ecocloud.ch

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

IT is ever more **indispensable**

Our life w/o digital data is unimaginable as

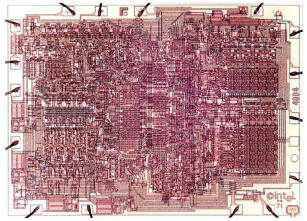
- Enterprises
- Governments
- Societies
- Individuals
- Scientists



“He saw your laptop and wants to know if he can check his Hotmail.”

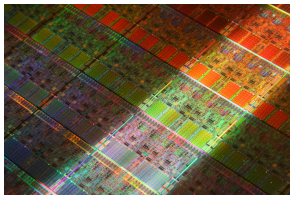
IT: An Exponential Growth

Intel 4004, 1971

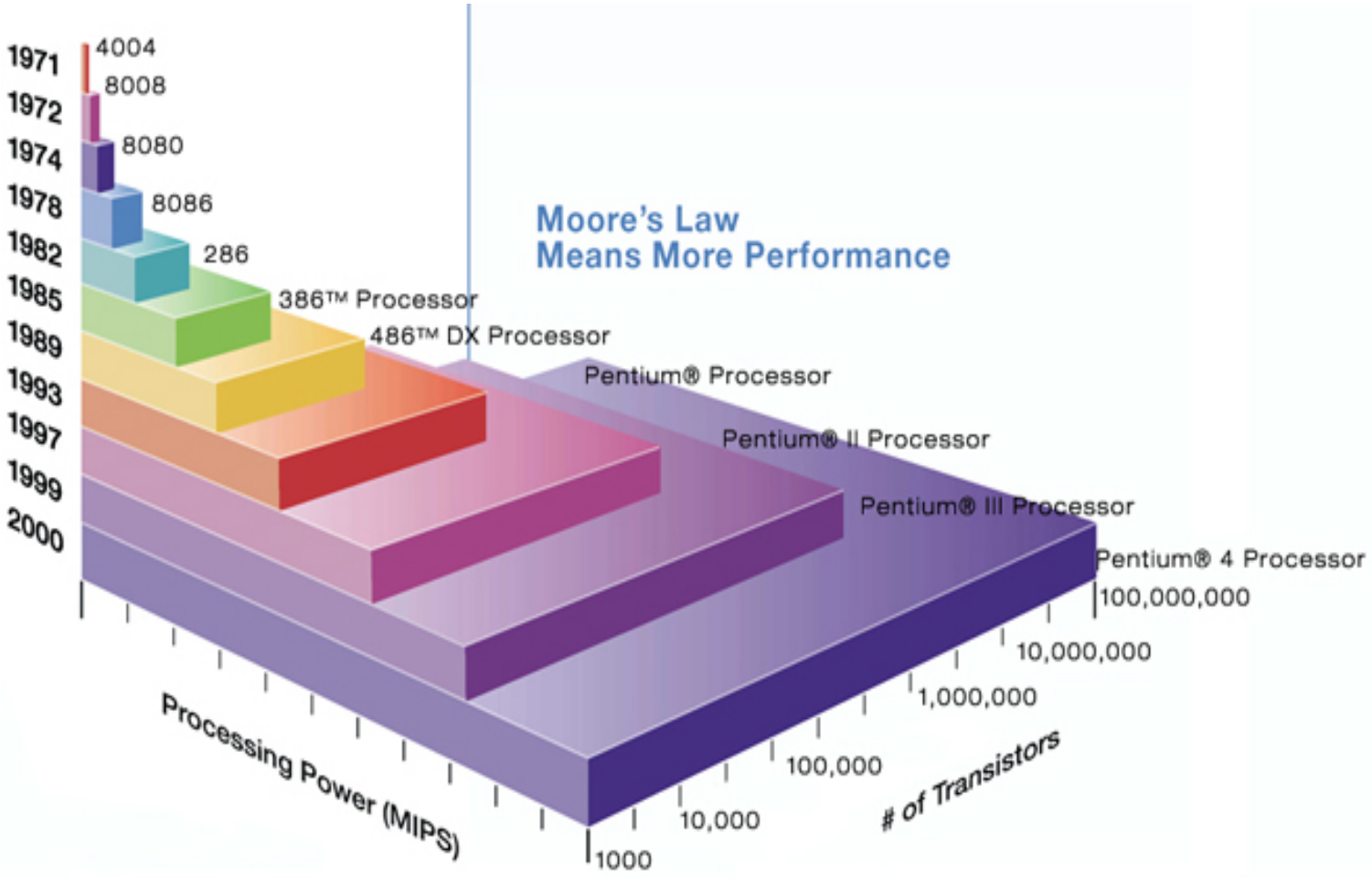


92,000 ops/second

Intel Nehalem, 2009



12,000,000,000 ops/second



Four decades of digital platform proliferation

Exponential increase in density & decrease in cost

A Brief History of IT

Communication Era

Consumer Era

1970s- 1980s 1990s Today+

Mainframes

PC Era

- From scientific instrument to commodity
- From product to service

IT: The Consumer Era

Phenomenal change from decades ago:

- Instant connectivity
- Shopping now online
- Daily interaction > 300 people
- Augmented reality
- Streaming movies
- …..

IT is at core of everyone's life!

Change in IT's Landscape

➔Emergence of Data-Centric Universe
  – IT focus on massive data

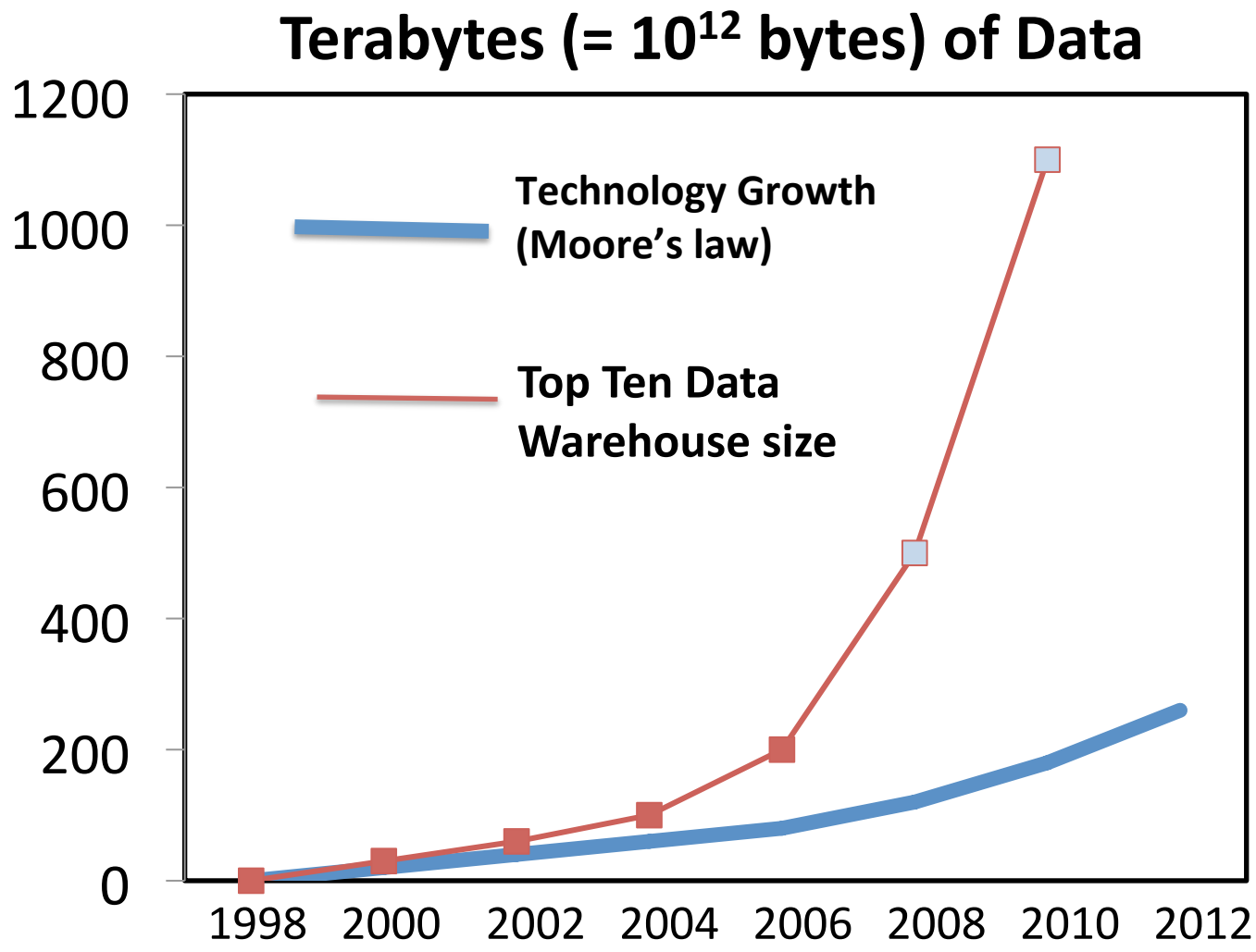
- End of Dennard Scaling
  – Higher density ➔ higher energy

- Data-Centric Universe meets Energy Wall

What are design implications?

Our Data-Centric Universe: Data Growing faster than Technology

- Commerce entirely data-driven
- Science handling massive data
- Companies spending $$$ to collect/analyze data
- Personalized computing



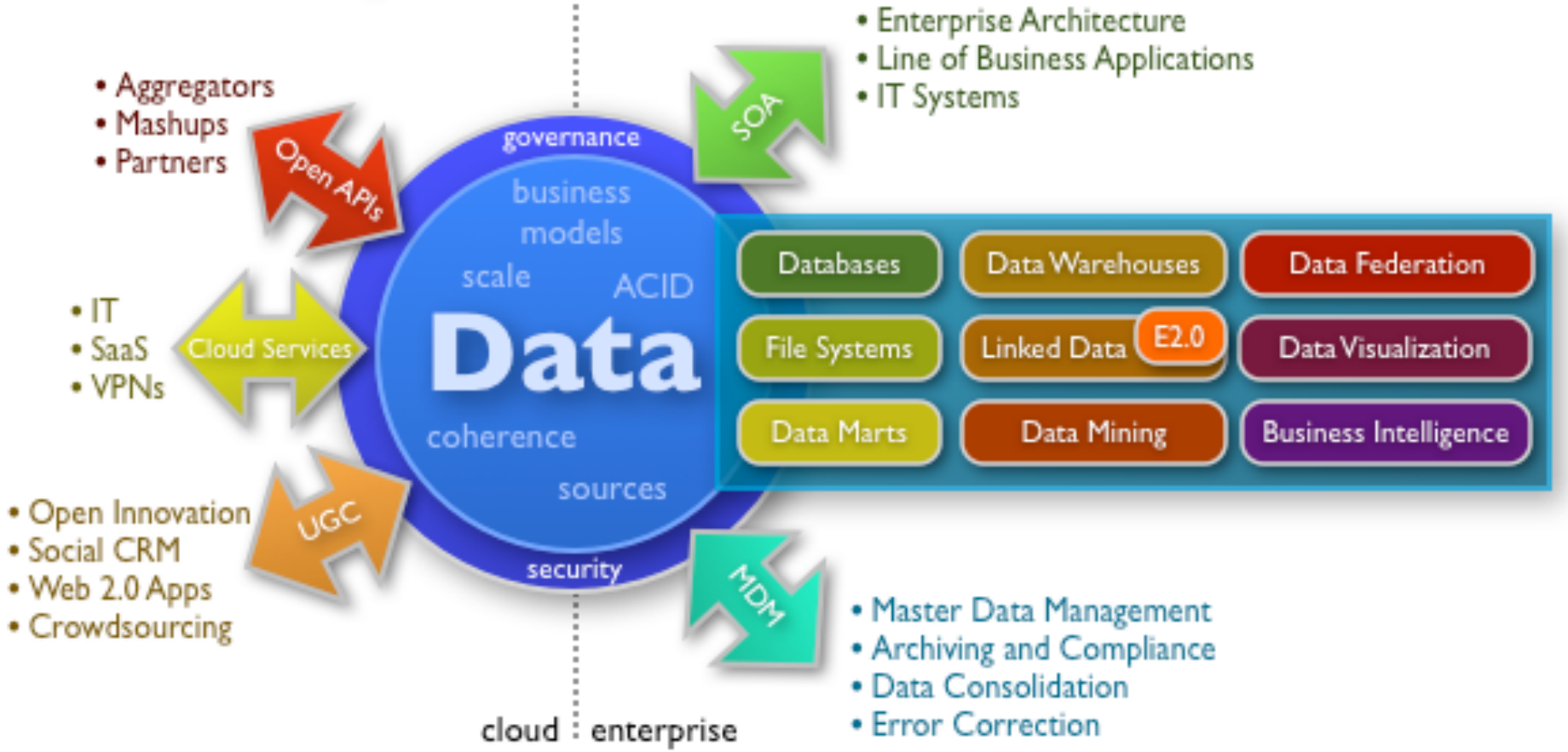
WinterCorp Survey, www.wintercorp.com

Data Deluge: 1200 Exabytes in 2010

(Economist, Feb. 25th 2010)

- Only 150 Exabytes in 2005
- Supply-chain management, 10x increase in data in a year
- US aerial surveillance models 30x more data in 2011

Anatomy of a Data-Centric Business

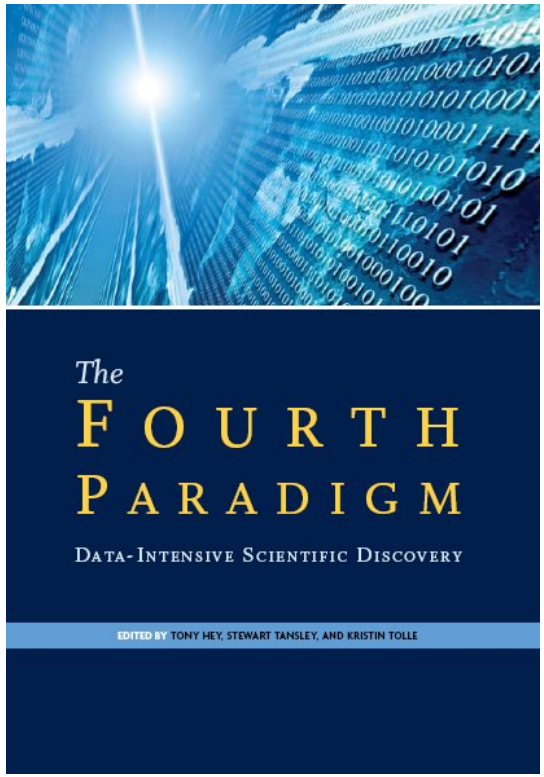


- Era of “knowledge economy”
- 50% of economic value in developed countries
- Dominant supply-chain component of products/services

Data-Centric Science: "The Fourth Paradigm"

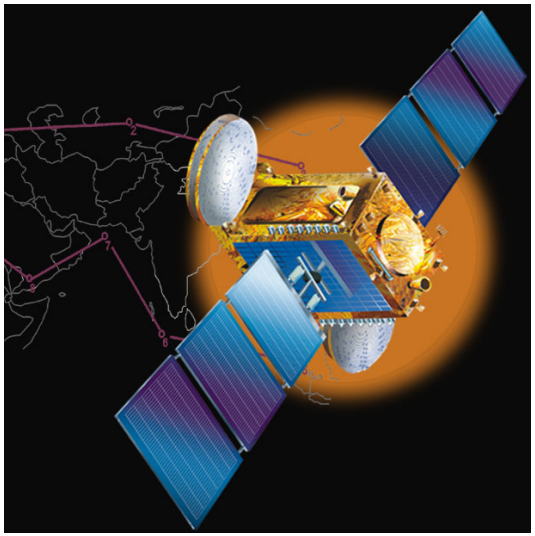
Mining data from:

- Archives
- Humans
- Sensors/instruments
- Simulations



Unifying theory, experimentation, simulation, analytics on massive data

Data Comes in Various Flavors



Satellite

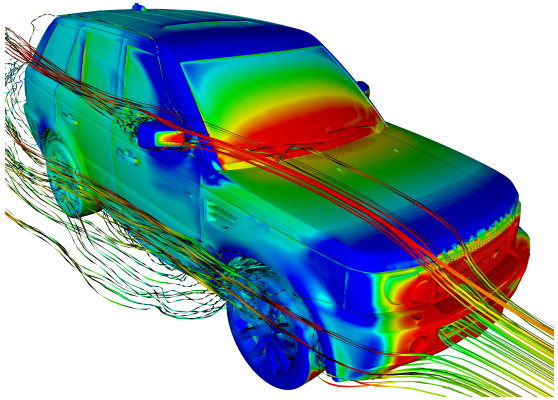
Health

Entertainment

Life

Commerce

Search



Simulation

It's all about Accessing Data!

Data Centers

YouTube

Cloud Computing

A computing paradigm shift to enable ubiquitous connectivity

How to design for massive data

Change in IT's Landscape

- Emergence of Digital Universe
  - IT focus on massive data

➔End of "Free Energy"
  - Higher density ➔ higher energy

- Data-centric Universe meets Energy Wall

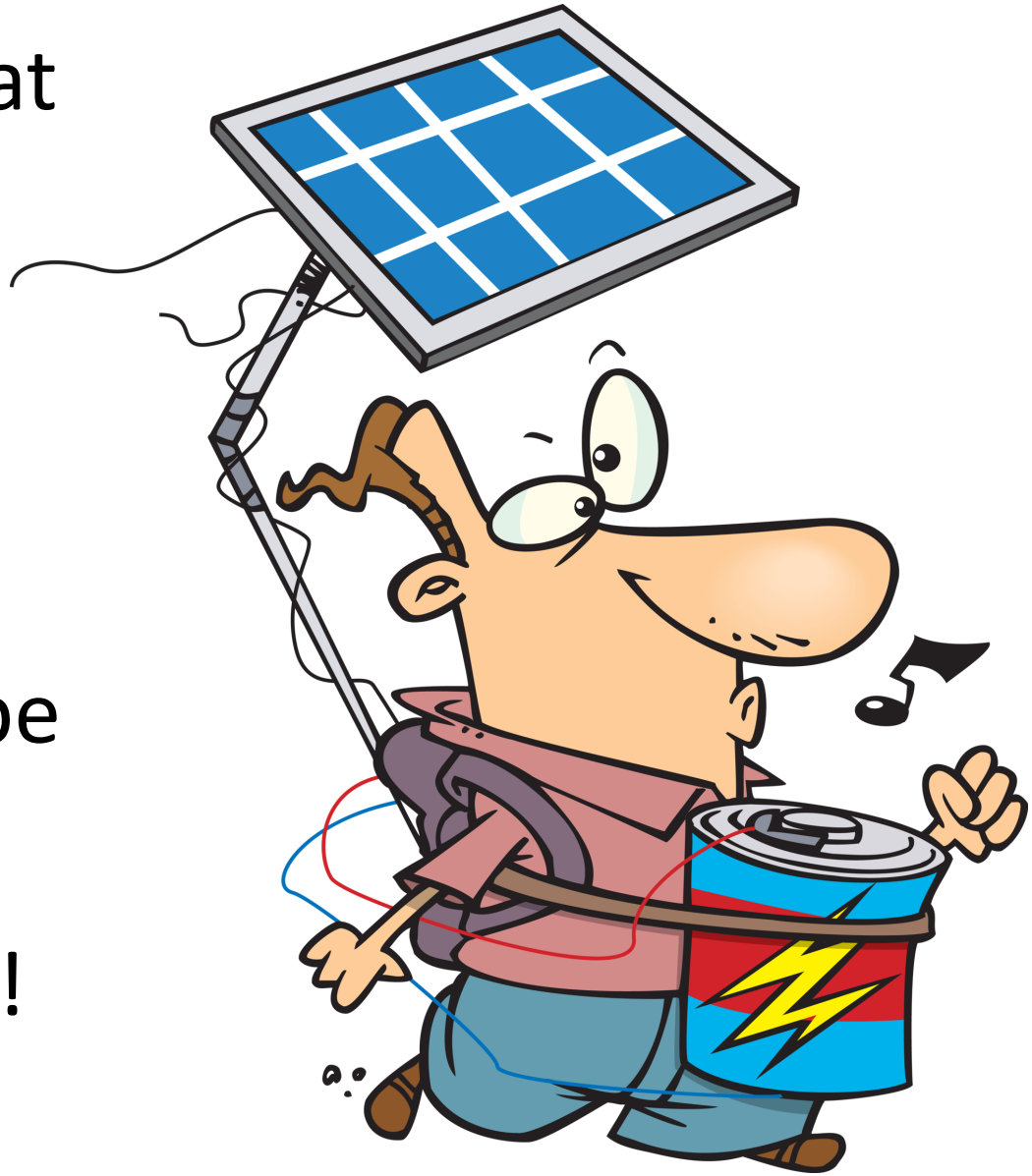
What are design implications?

IT Energy is Shooting Up!

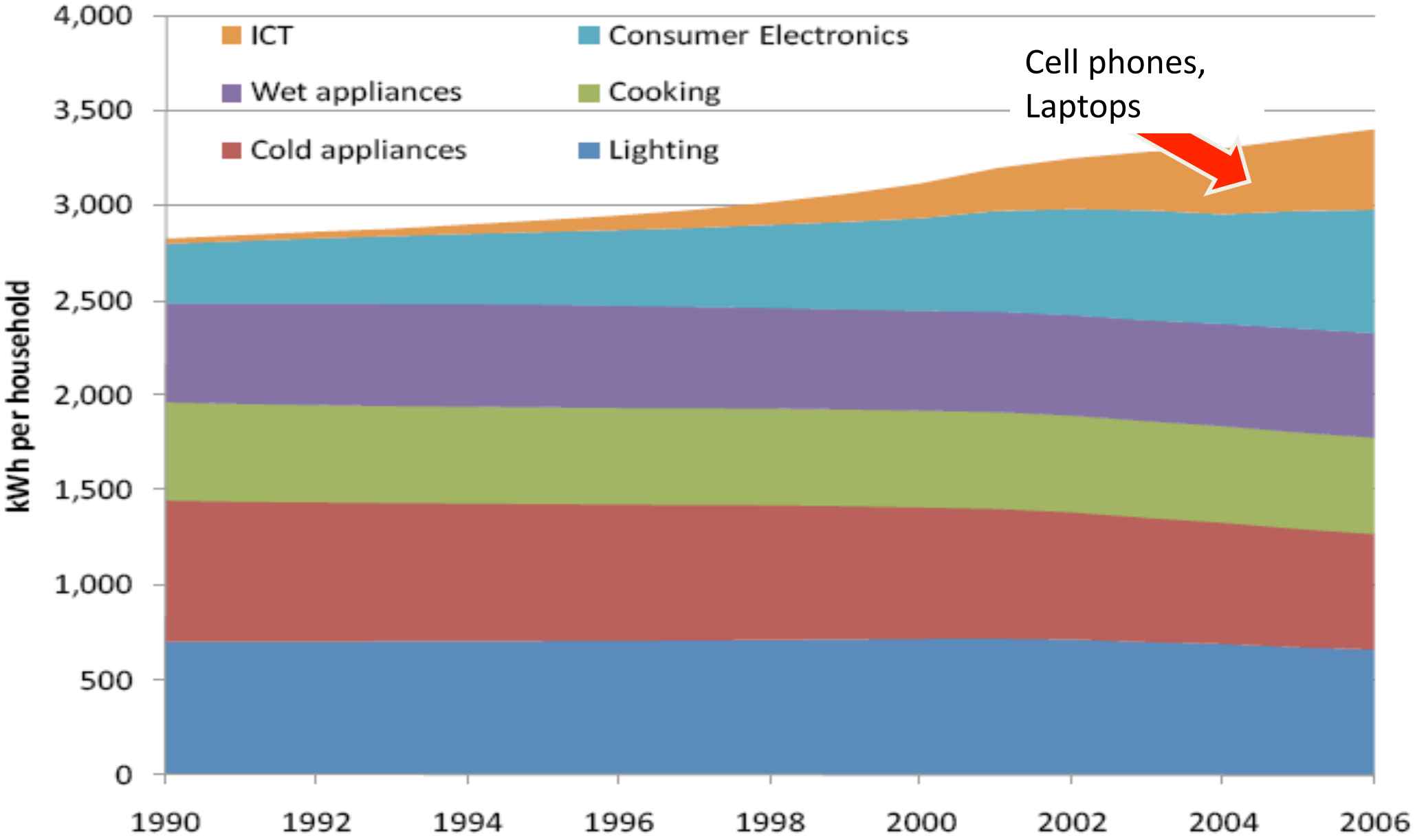
IT riding on technology that was energy-friendly

- Exponentially better performance, density
- Constant power envelope

But, energy is shooting up!



Household Energy in the UK
(UK BERR, 2008)



Household Energy in the US
(NY Times, 2011)

Comparing Energy Use

Comparison of a typical television set-top box configuration with Energy Star-rated appliances and devices.

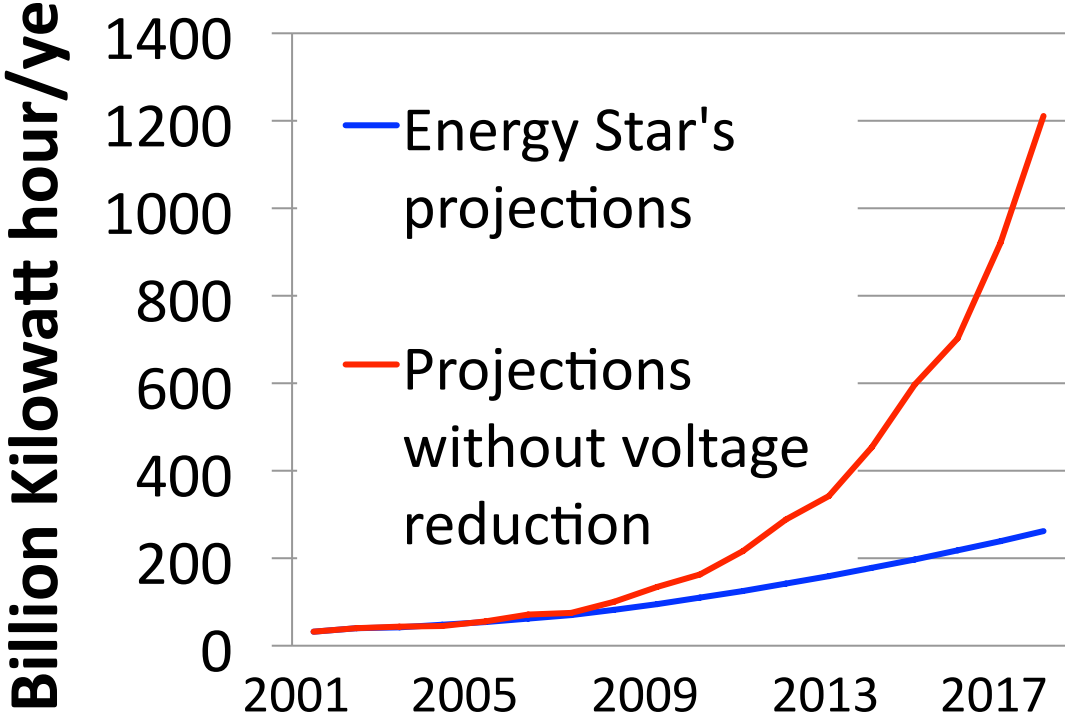
	AVERAGE KILOWATT-HOURS A YEAR	HD SET-TOP BOX	HD DVR	TIME IN USE EACH DAY
Typical HD television set-top box configuration	446	171	275	24 hours
Refrigerator (21-cubic-foot)	415			24 hours
LCD television (42-inch)	181			5 hours
Desktop computer	175			8 hours
Compact fluorescent light bulb (15-watt)	17			3 hours

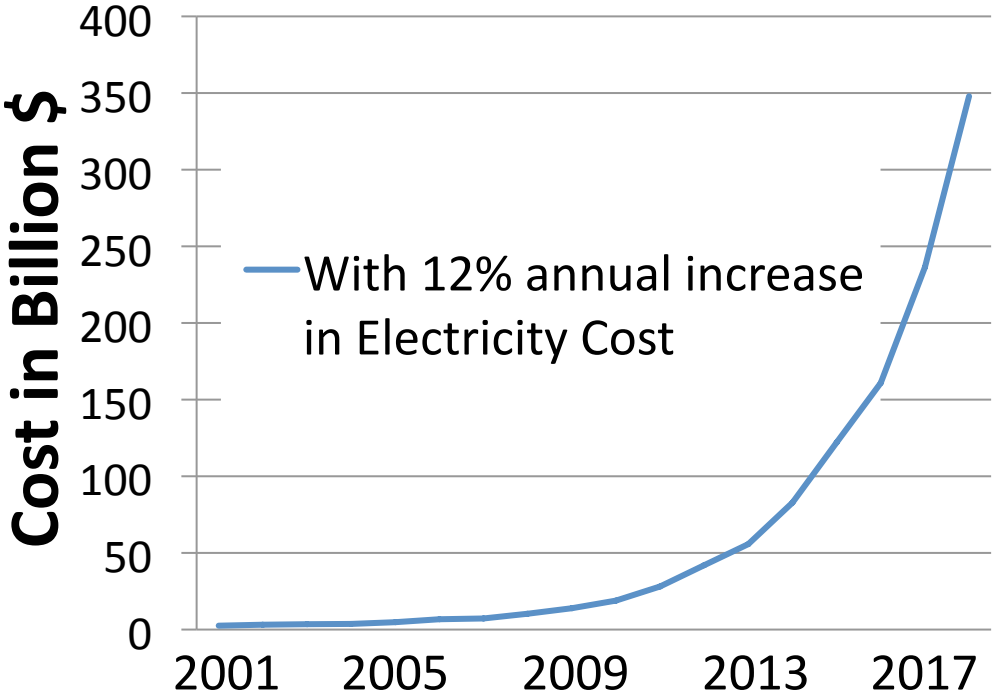
Source: Natural Resources Defense Council

THE NEW YORK TIMES

Data center Energy in the US

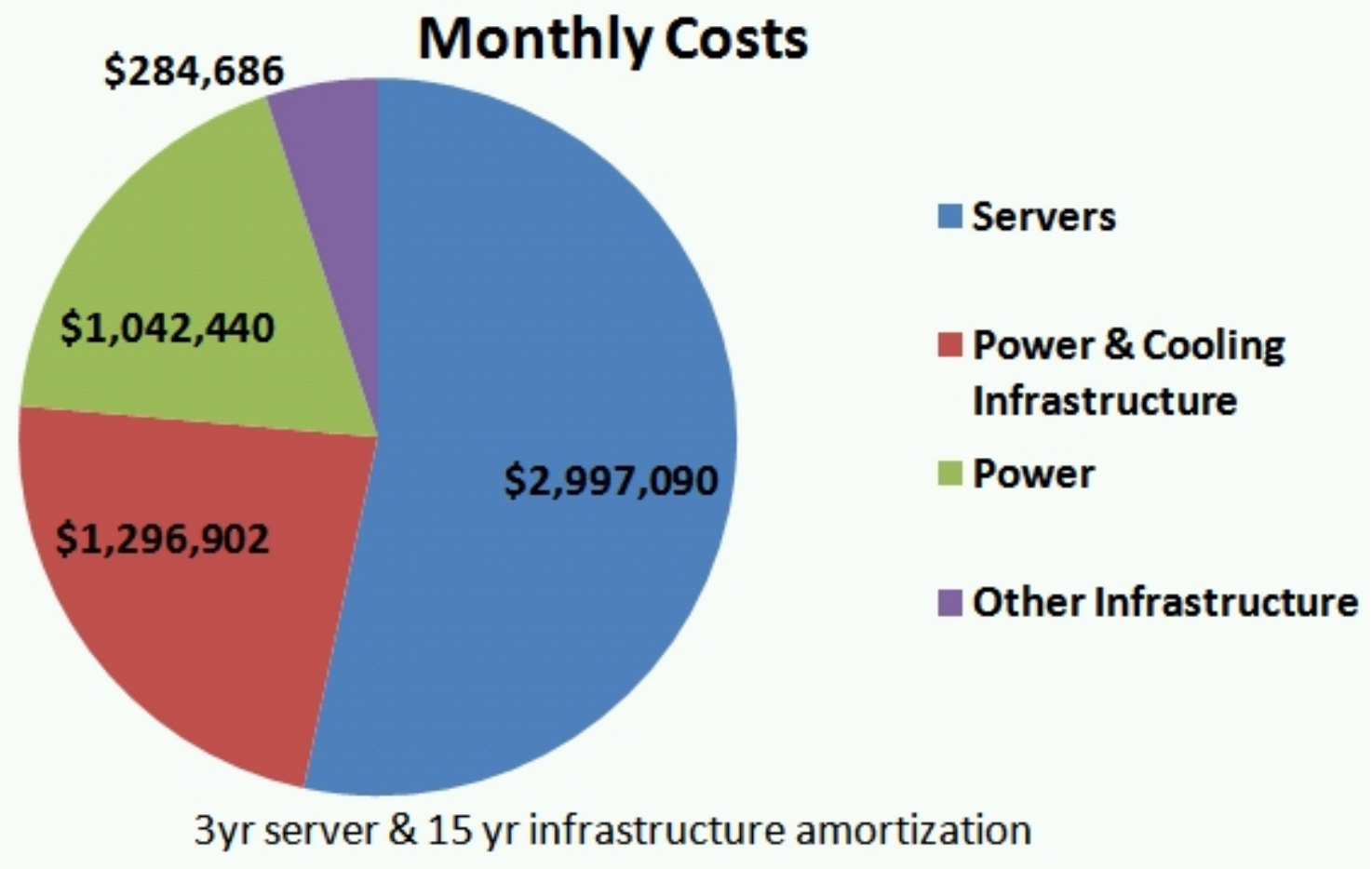
(extrapolated from Energy Star, 2007)





- Exponential costs if not mitigated
- Today, carbon footprint of airline industry

Energy > Capital Cost

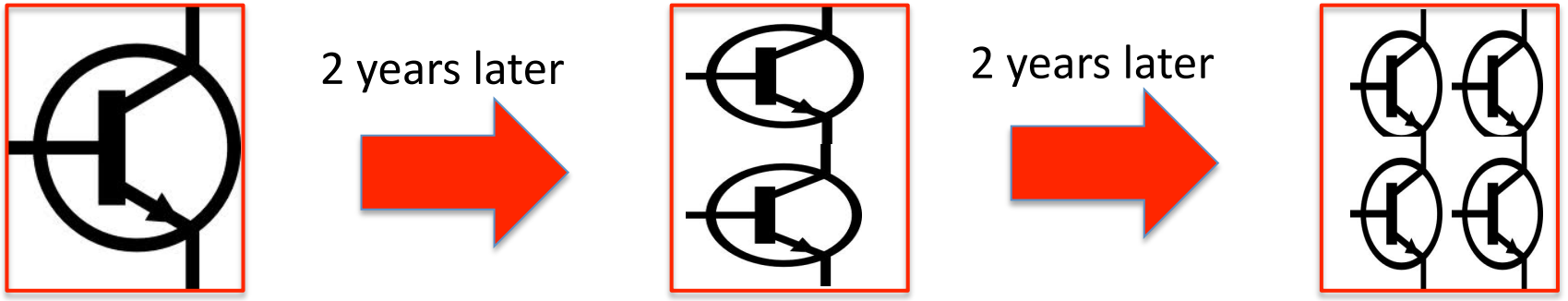


James Hamilton's Blog, mvdirona.com, 2008

- Servers are getting relatively cheaper
- Power is beginning to dominate cost

End of “Free” Energy

1 transistor = 1x energy **2 transistors = 1x energy** **4 transistors = 1x energy**



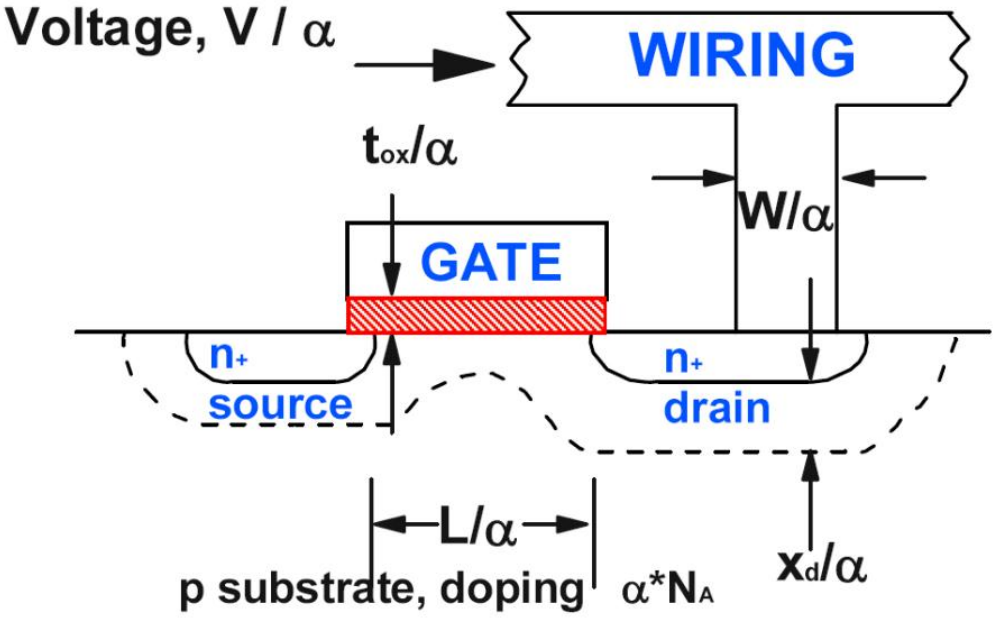
Before (1970~2000):

- Dennard scaling
- Used to make transistors smaller
- Smaller transistors less electricity to operate

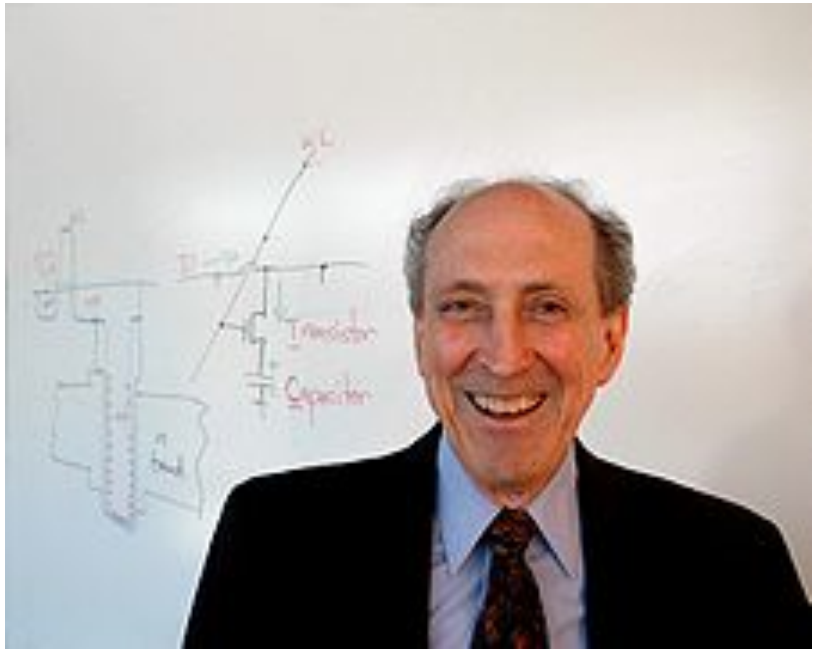
Now (2004-):

- Continue to make transistors smaller
- But, they use similar electricity to operate

Four decades of Dennard Scaling



Dennard et. al., 1974



Robert H. Dennard, picture from Wikipedia

- **$P = C V^2 f$**
- Increase in device count
- Lower supply voltages

➜ Constant power/chip

Leakage Killed Dennard Scaling

Leakage:

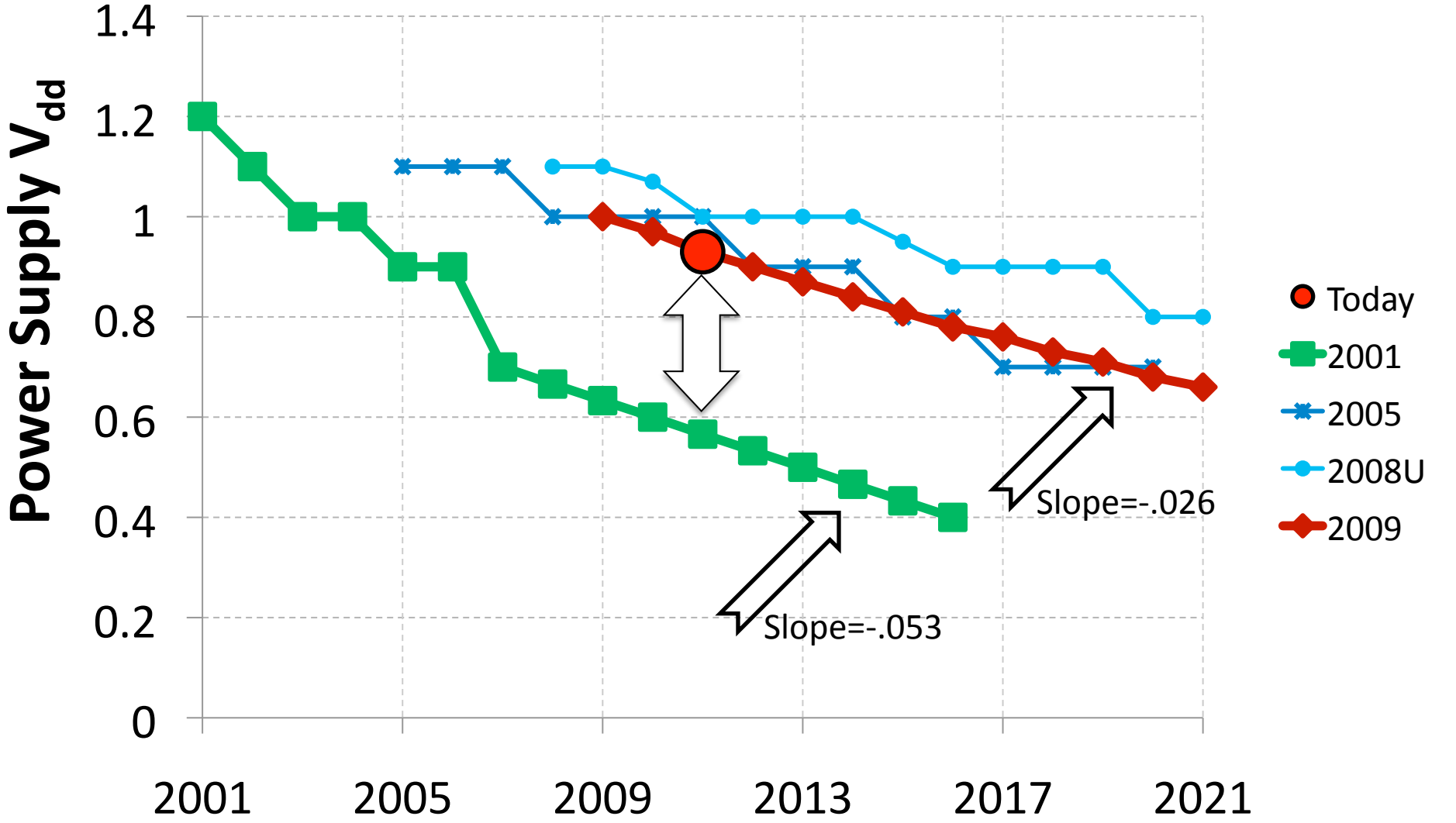
- Exponential in inverse of V_{th}
- Exponential in temperature
- Linear in device count

To switch well

- must keep $V_{dd}/V_{th} > 3$

➜V_{dd} can't go down

End of Dennard Scaling (ITRS)



Mike Ferdman, from ITRS pages, July 2011

Supply voltages going down at much lower rate!

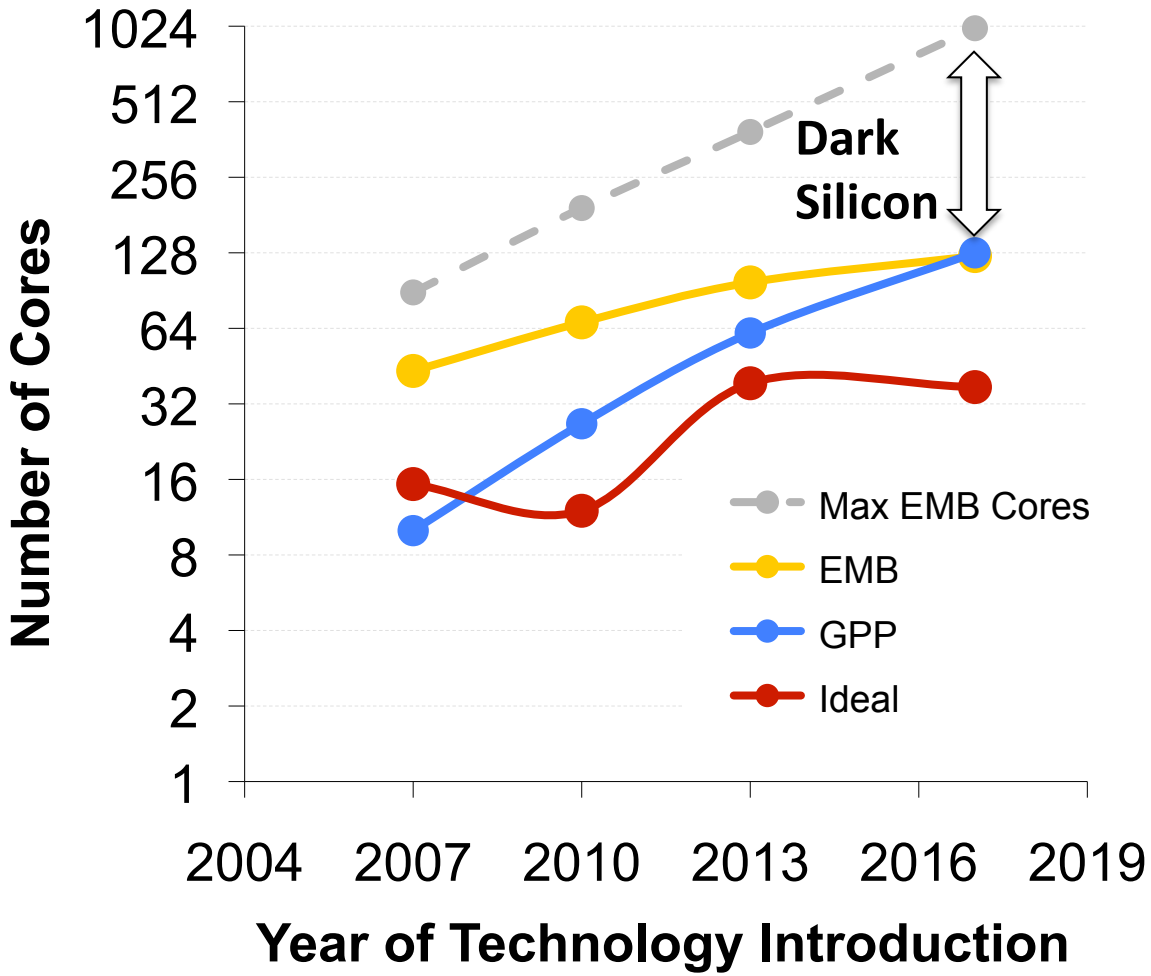
Dark Silicon: End of Multicore Scaling

Can not power up chip for fully parallel SW

Parallelism has limits even in Servers!

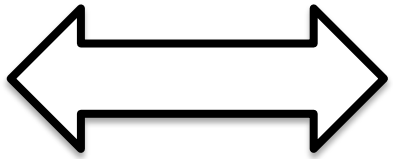
Must:

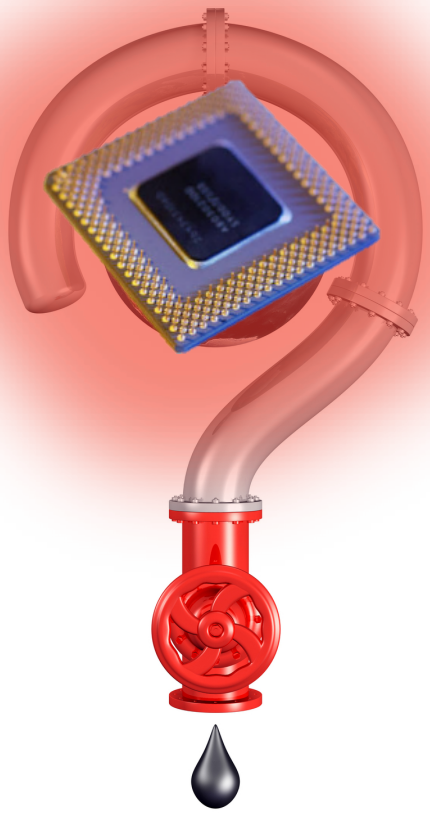
- specialize
- selectively power up



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

Massive Data meets Energy Wall





It's time for Europe to lead the way:

- Existing Industrial Ecosystem (e.g., ARM, ST, SAP)
- Long leadership in energy-efficient design

Change in IT's Landscape

- Emergence of Data-Centric Universe
  - IT focus on massive data
- End of "Free Energy"
  - Higher density ➔ higher energy

➔Data-Centric Universe meets Energy Wall

What are design implications?

What are the design Implications?

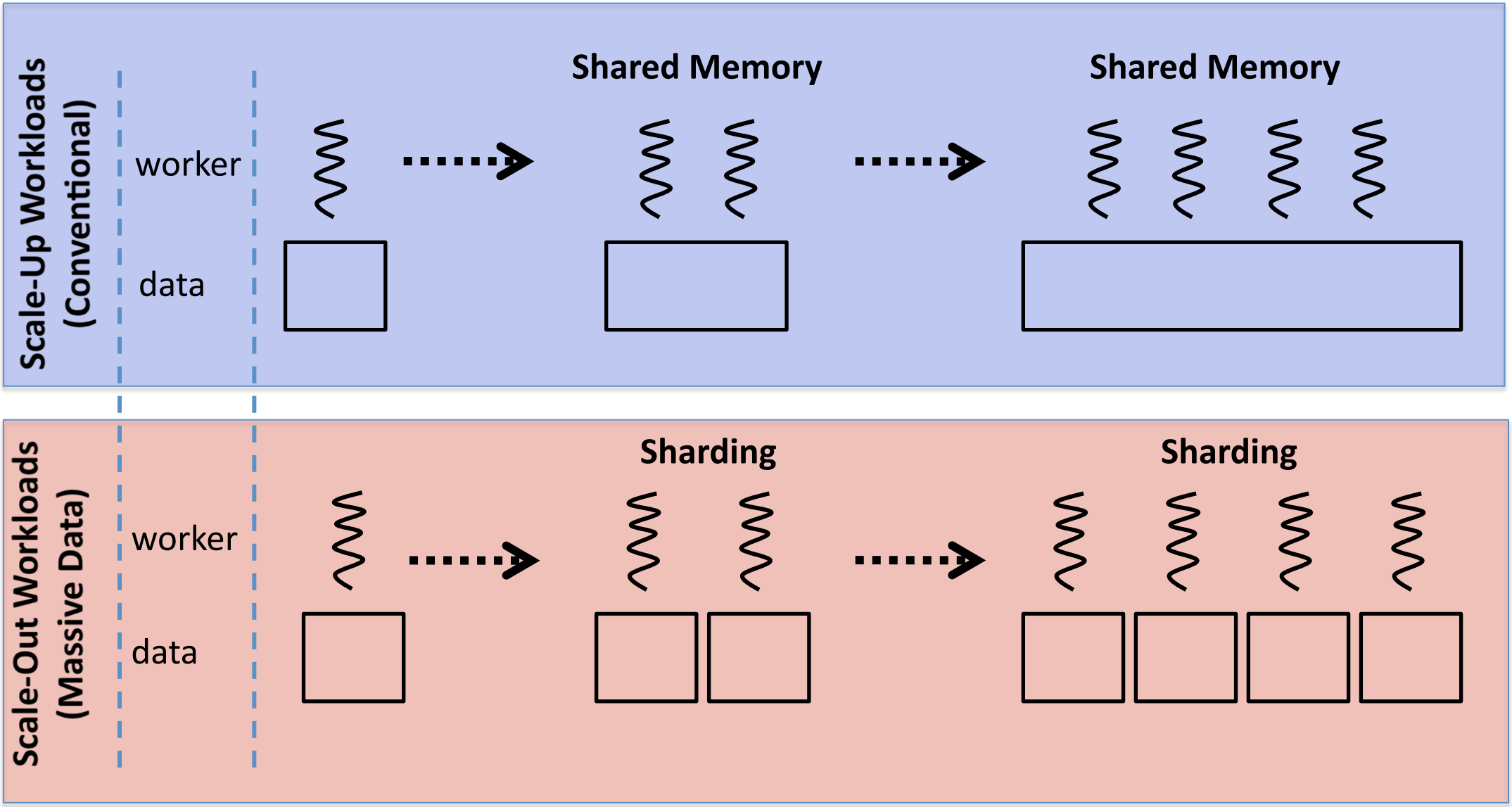
Short term:

- Multicore scaling (parallelism)
- Lower energy + higher data connectivity

Long term:

- Dark Silicon
- Probabilistic computing
- Holistic Integration

Scale-Out vs. Scale-Up Workloads

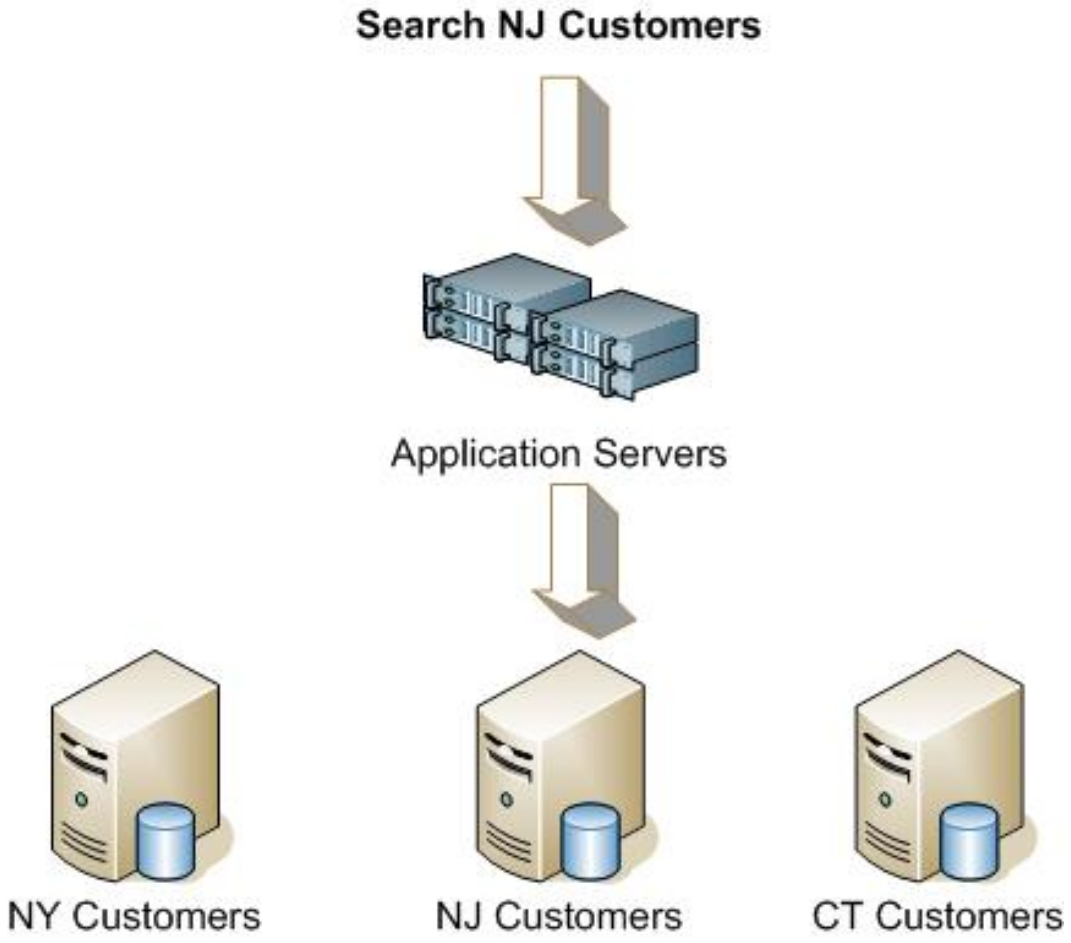


Emerging workloads scale out!

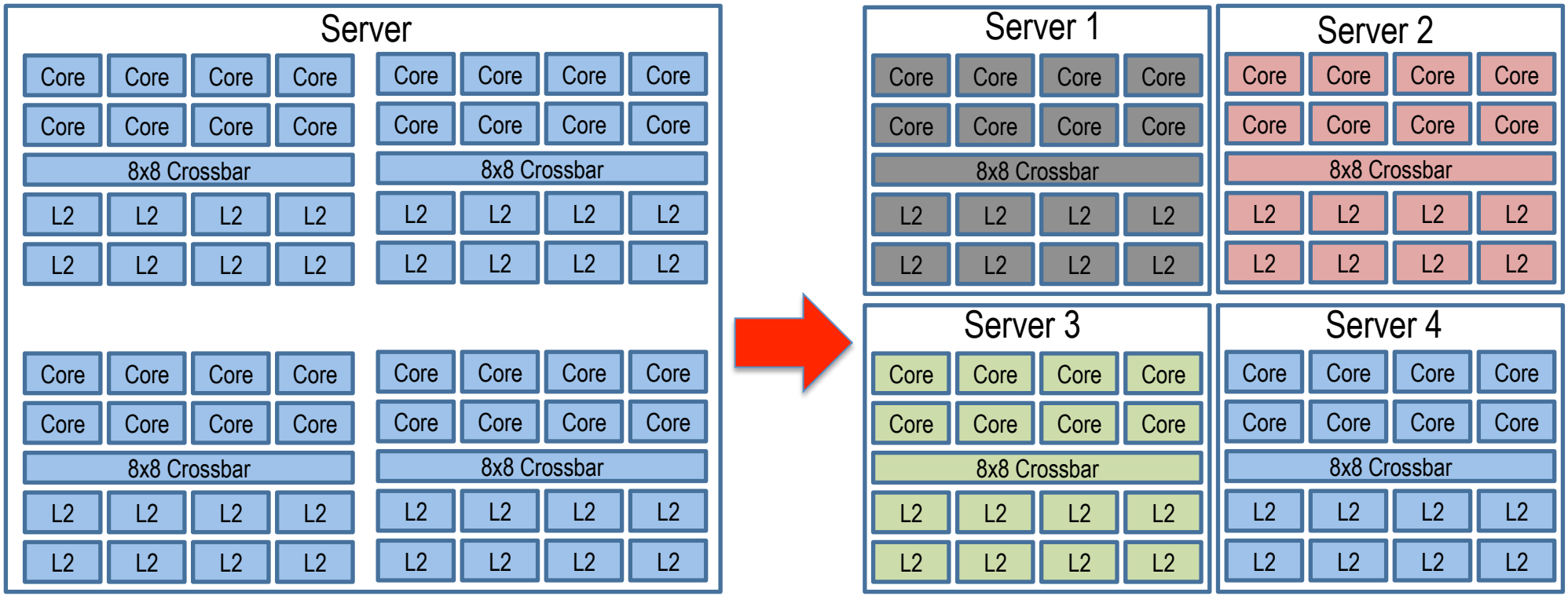
Emerging Workloads are Scale-Out

Examples:

- Data serving (YCSB)
- Streaming
- Search
- Analytics
- Web



Scale-Out vs. Scale-Up Chips



Scale-Up Chip: Conventional Shared Memory

Scale-Out Chip: Clustered Memory

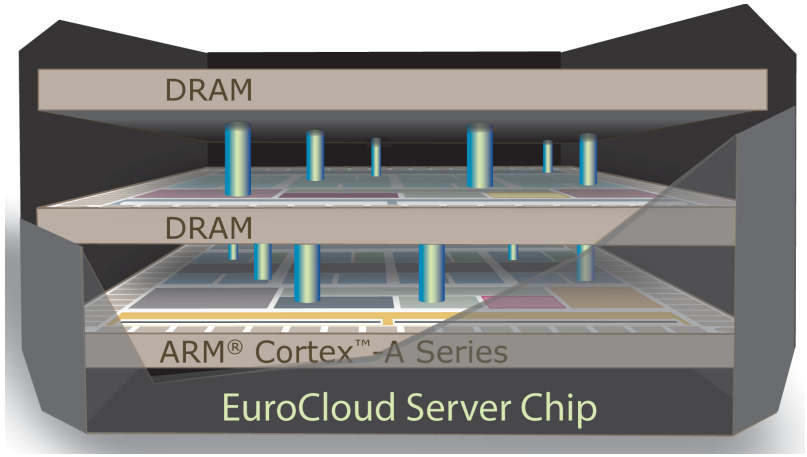
- Scaling out divides chip among disconnected servers
- Maximizes performance density, improved reliability

The EuroCloud Server:
A Scale-Out Chip for Massive Data

(www.eurocloudserver.com)

3D SoC/DRAM:

- 1000x more connectivity
- 10x less system energy
- Runs off-the-shelf SW stack



Your Future 1-Watt Datacenter Chip

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

ARM

SEVENTH FRAMEWORK PROGRAMME

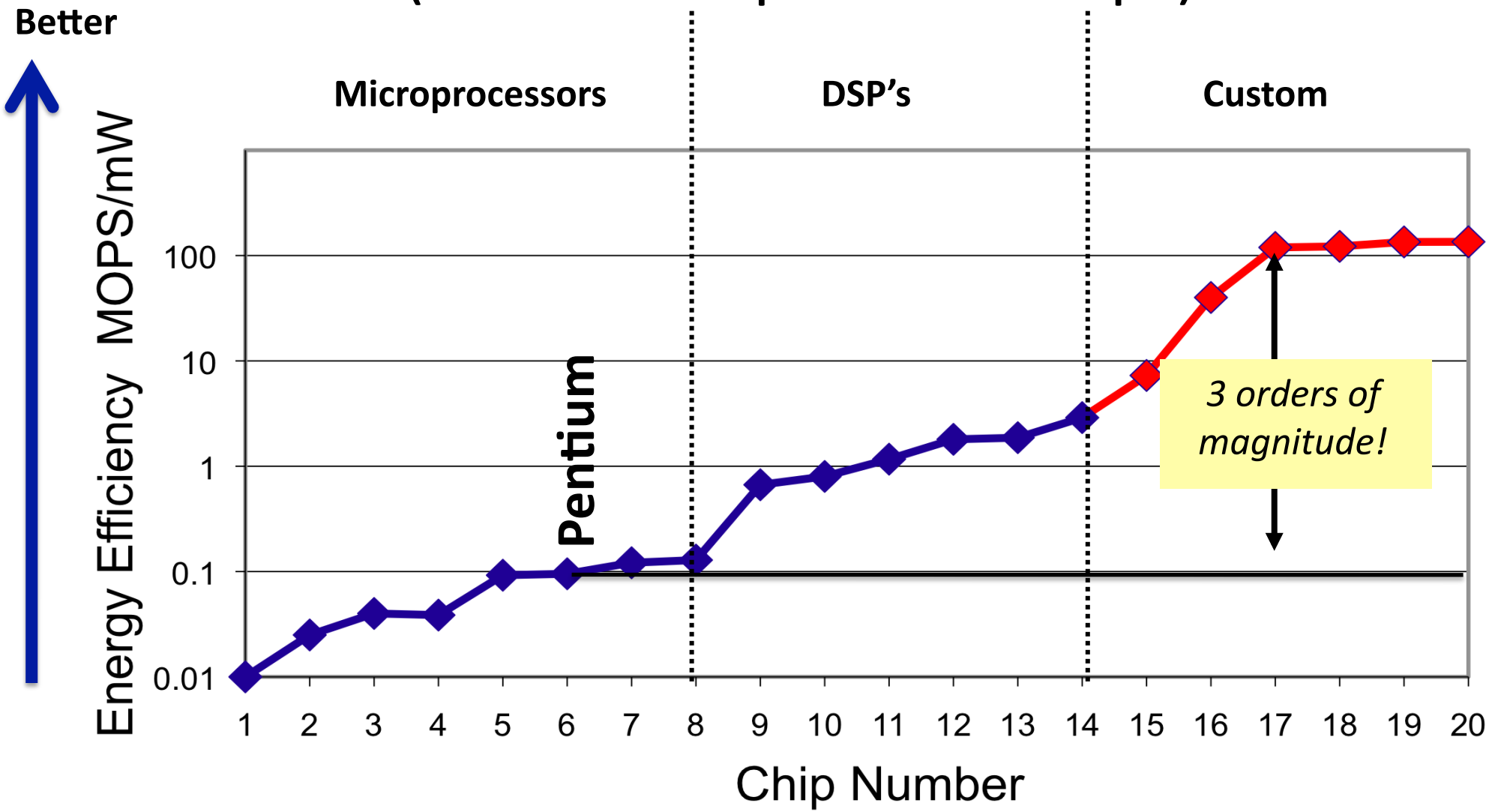
NOKIA Connecting People

imec

UCyprus

Specialization can buy 1000x in Energy

(from a sample of 20 chips)



Mihai Budiu, "On the Energy Efficiency of Computation", 2004

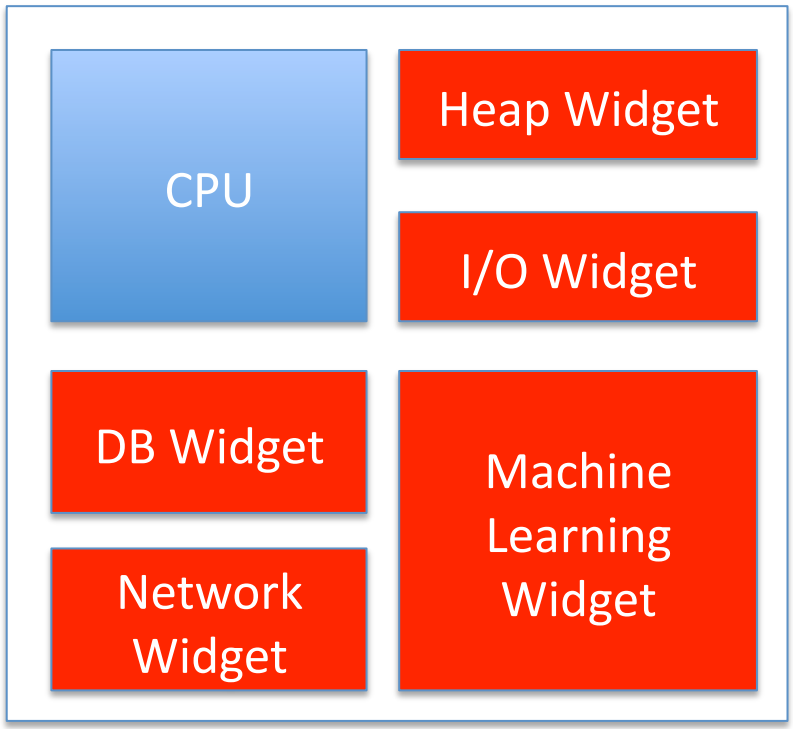
Beyond EuroCloud Server: Vertically-Integrated Server Arch. (VISA)

Identify energy hogs:

- Specialize
- E.g., Intel's TCP/IP CPU

Power up (dark silicon) services on the fly

- Others: Temam @ INRIA, MSR, UCSD



VISA System-On-Chip

Exact vs. Probabilistic

Much computation is error-resilient:

- Machine learning/analytics
- Image processing/visual computation
- Audio/speech
- Search

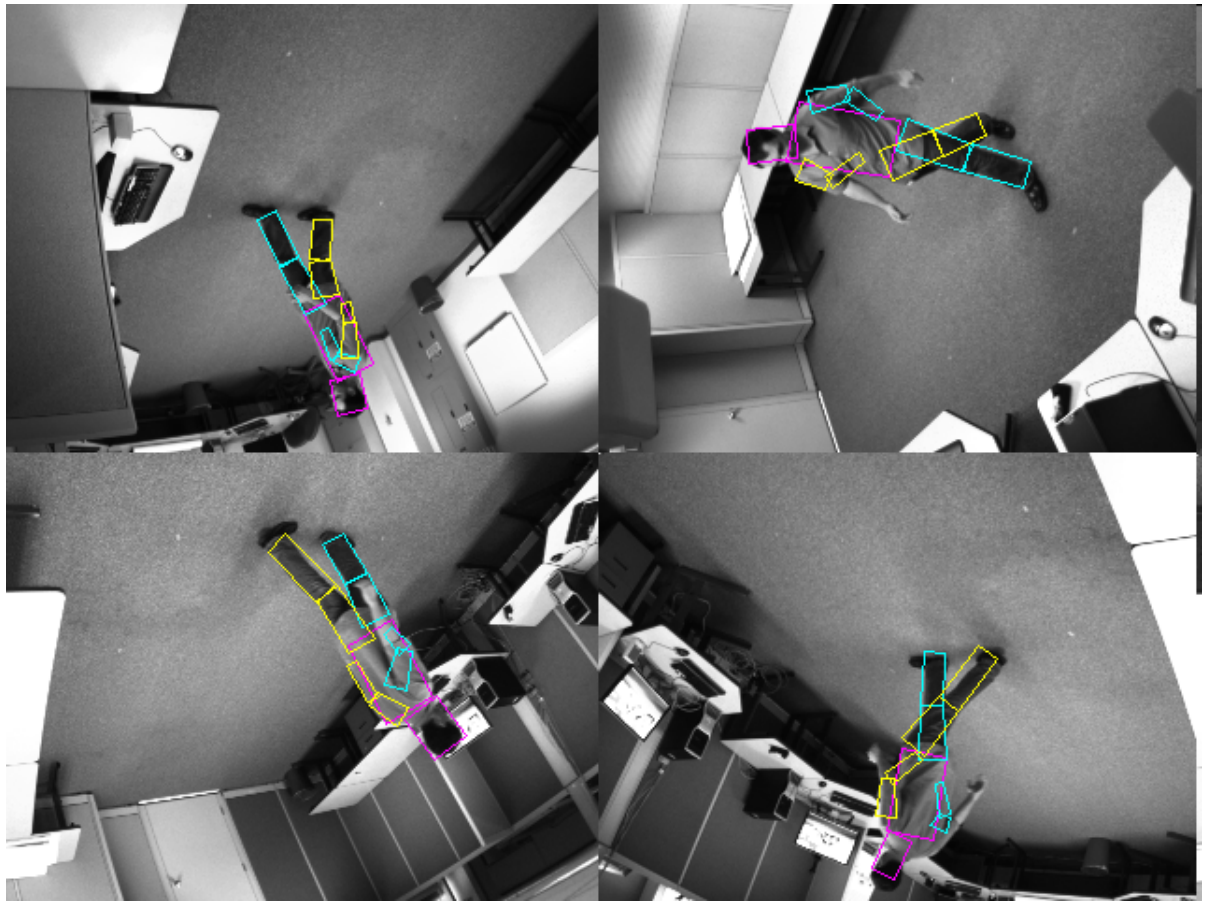
Similarly, two flavors of data

- Exact: affects functionality (pointer address)
- Probabilistic: affects quality (pixels in image)

Perforated (Skipped) Computation

bodytrack benchmark (PARSEC)

- Compiler-driven perforation
- Skip 40% of computation
- Maintains track on head, chest and legs



Computation does not have to be exact!

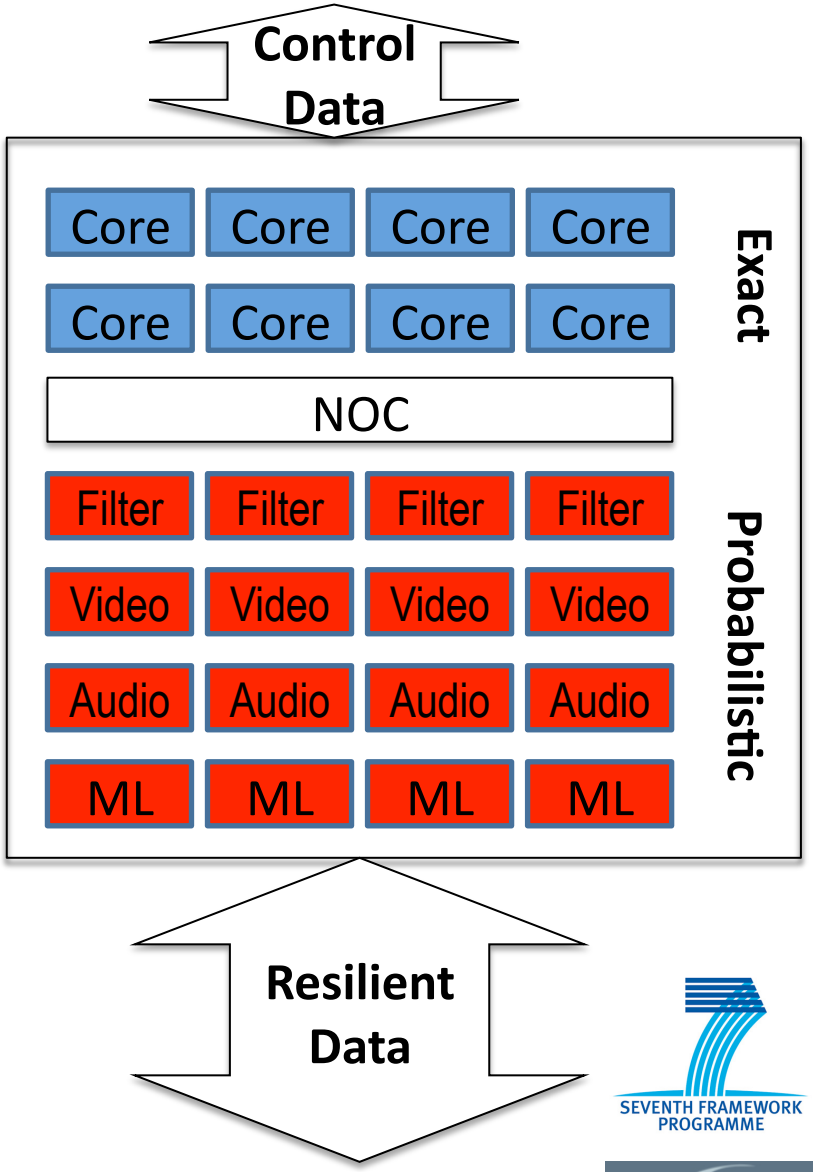
Hoffman et. al., “Using Loop Perforation to Dynamically Adapt Application Behavior to Meet Real-Time Deadlines”, 2010

DeSyRe: Probabilistic Computing

Exploit resilience in

massive data

- Partition according to resilience
- Push voltages down to "unsafe" regions
- Maximize throughput with less energy

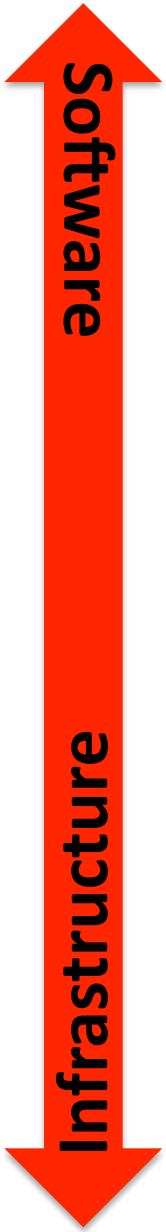


SEVENTH FRAMEWORK PROGRAMME

Holistic Integration Beyond IT

- Need interdisciplinary (sciences + technology)
- Tighter integration enables higher efficiency
- From SW to Energy
- Long-term vision:

→**Energy-neutral IT**



ecocloud

Research Center @ EPFL
ecocloud.ch

Dozen faculty, CSEM
& industrial affiliates

- HP, Intel, IBM, Microsoft, Nokia, Oracle, Credit Suisse, Swisscom,…
- A few million CHF of annual funding
- Datacenter Observatory (test bed)

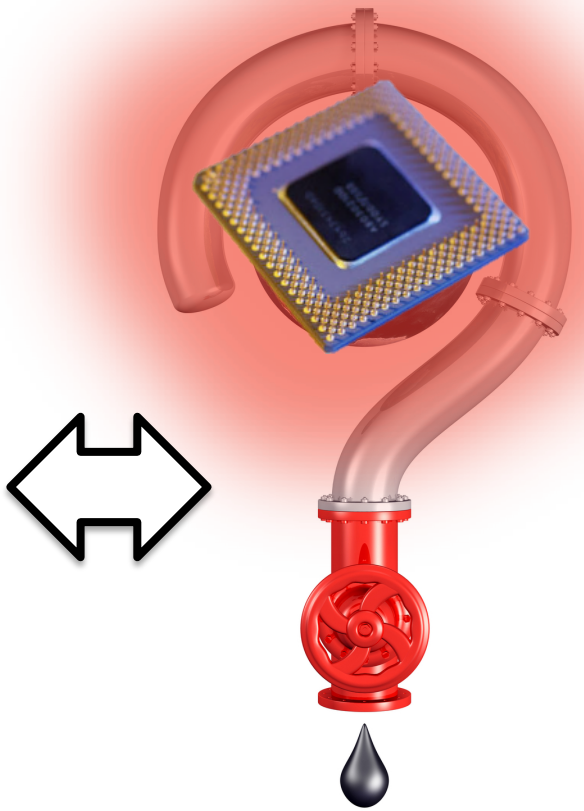
Research:

- Energy-minimal technologies for massive data
- Warehouse-scale data management
- Scalable cloud applications & services

Making tomorrow's clouds green & sustainable

Bringing it All Together

- IT is changing everything & itself changing
- IT systems are inefficient & too robust
- Plow massive data with minimal energy



A new IT revolution is emerging,

we have a great opportunity to lead!

Thank You!

For more information please visit us at

ecocloud.ch