Optimizing for the Cloud:

Tech Trends, Testbeds and Working Together

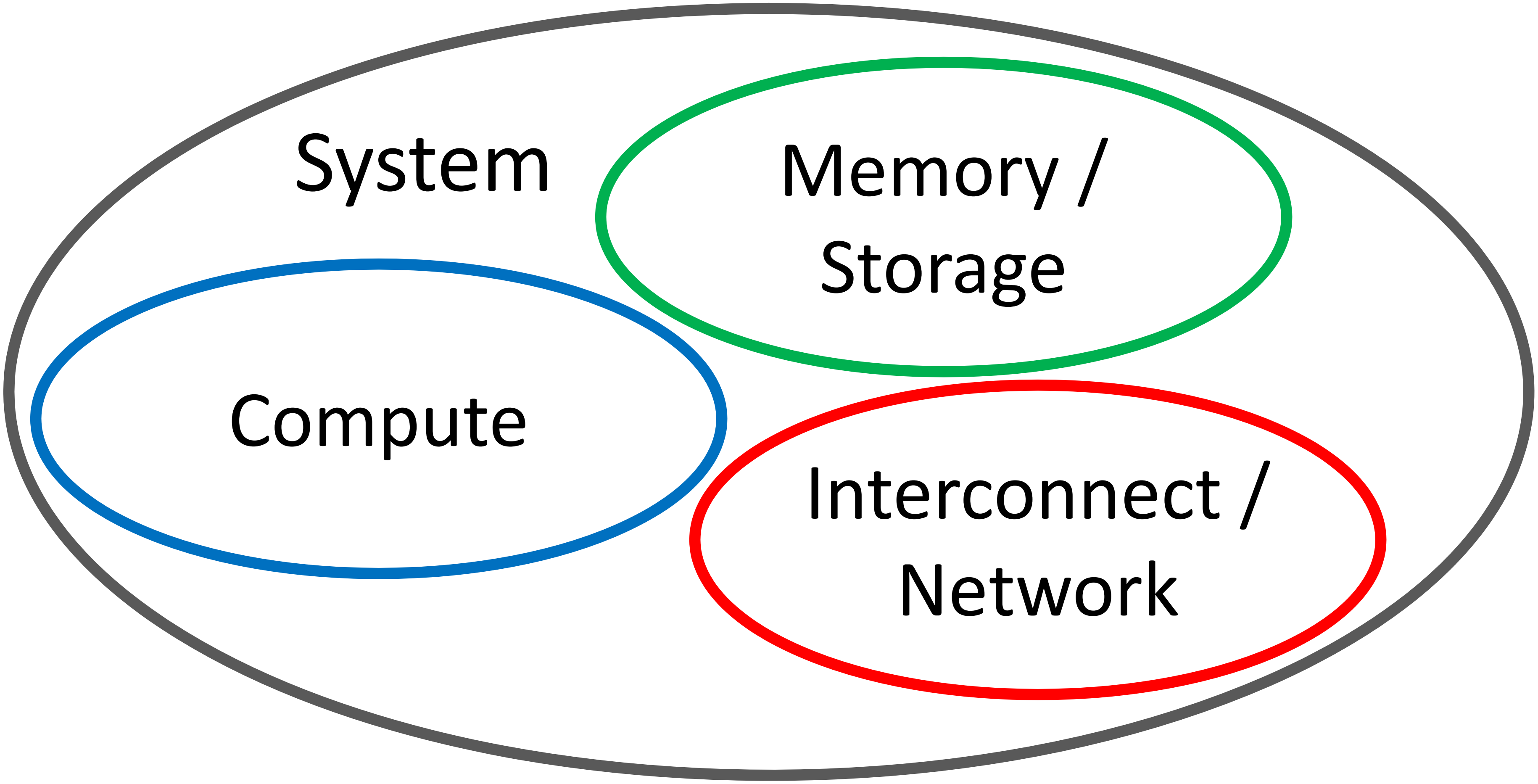
For Cloud ISTC Retreat
8 December 2011

Rich Uhlig
Intel Fellow

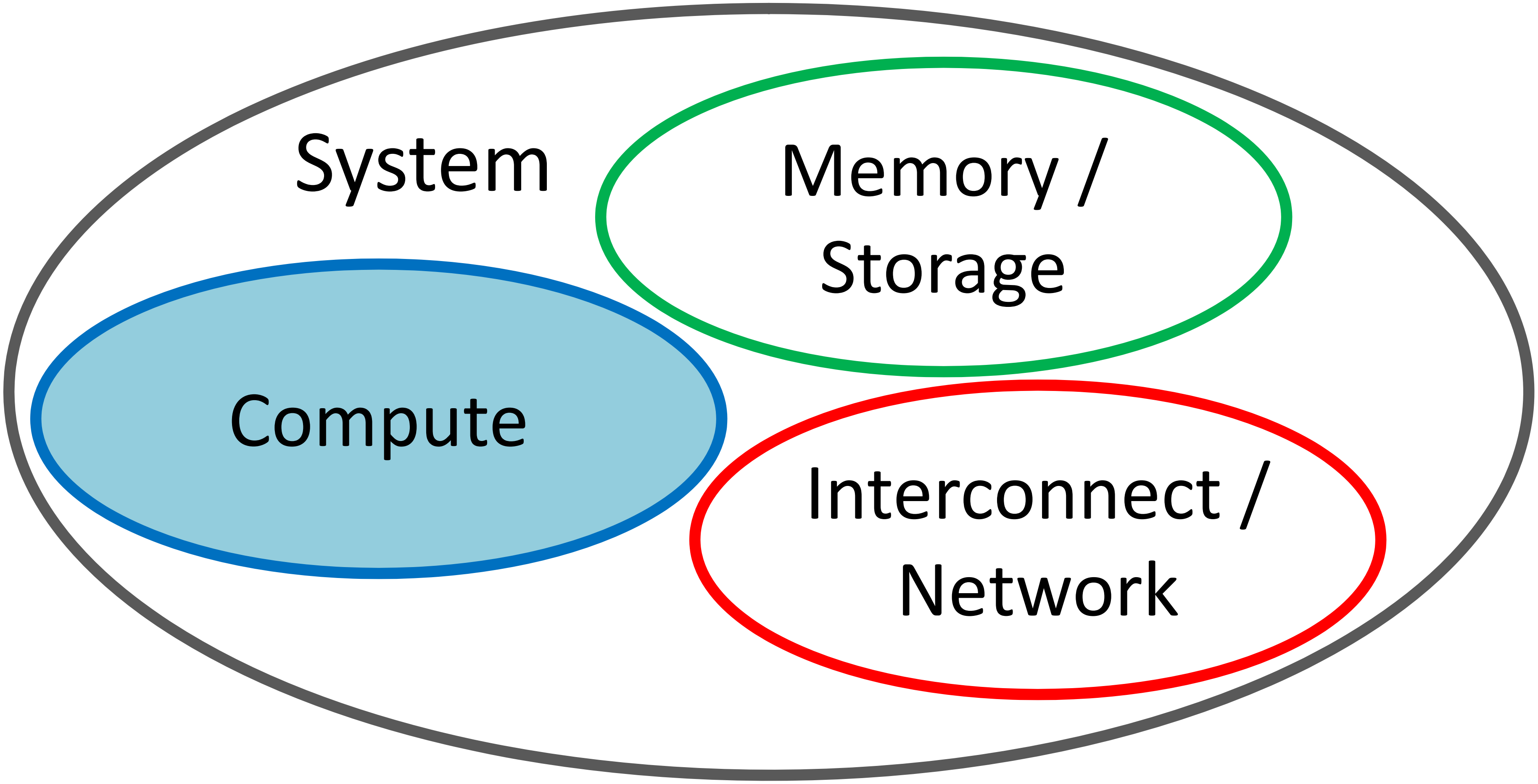
Cloud designs ruled by TCO…

… which requires achieving balance between system resources

Finding Balance



Finding Balance

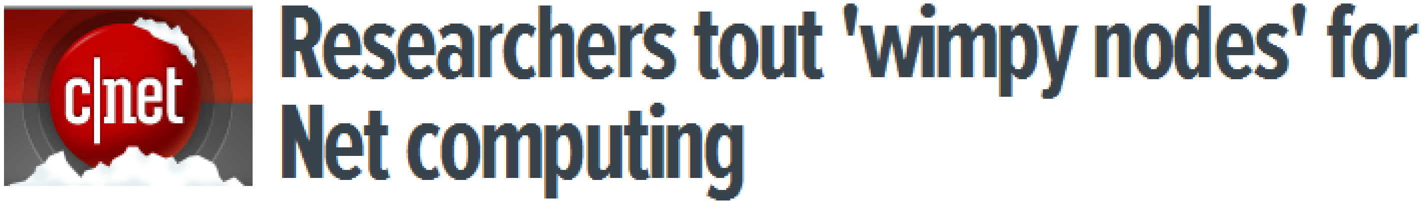

System

Memory / Storage

Compute

Interconnect / Network

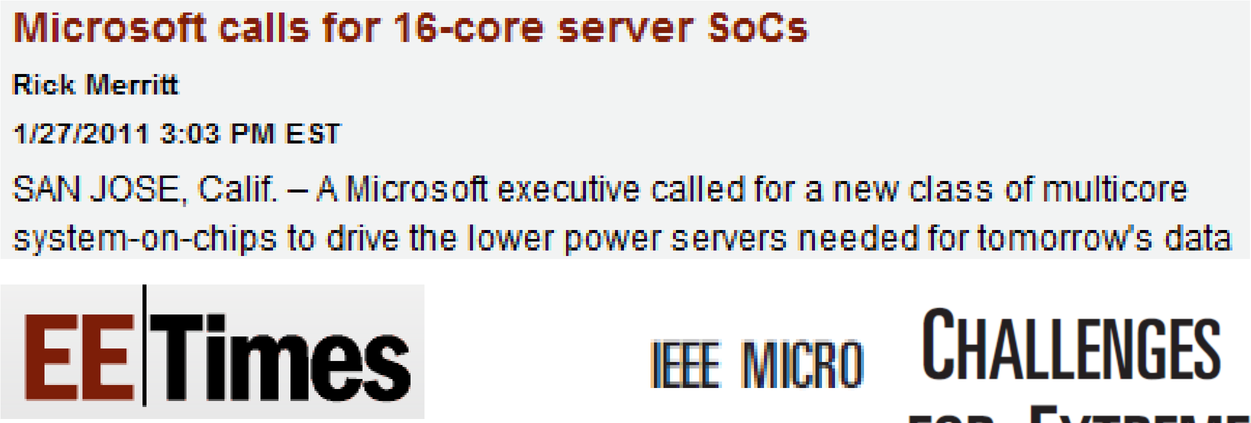
Wimpy versus Brawny?



c|net

Researchers tout 'wimpy nodes' for Net computing

“*Brawny cores still beat wimpy cores, most of the time*” - Urs Hölzle, Google



Microsoft calls for 16-core server SoCs

Rick Merritt

1/27/2011 3:03 PM EST

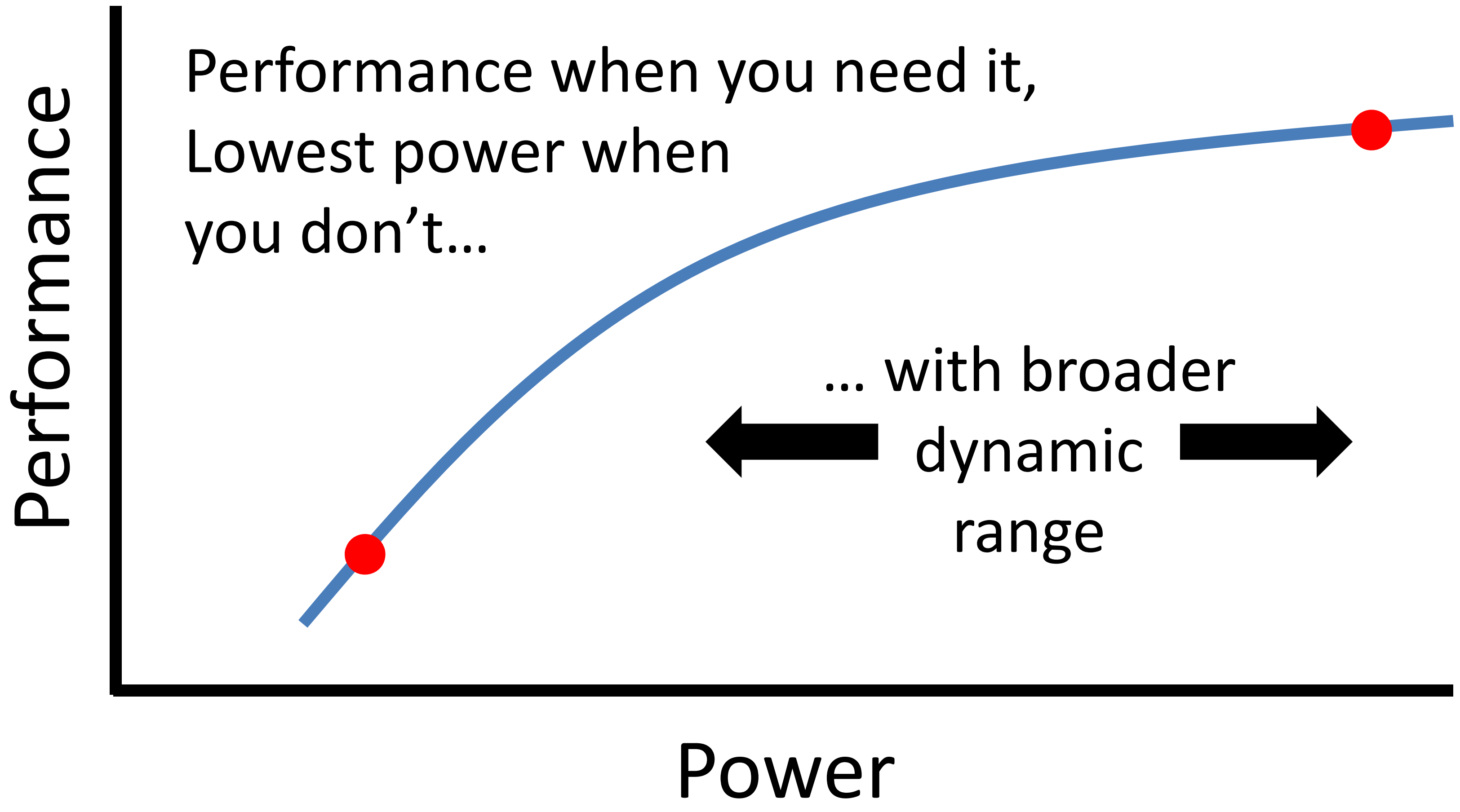
SAN JOSE, Calif. – A Microsoft executive called for a new class of multicore system-on-chips to drive the lower power servers needed for tomorrow's data

EE|Times

IEEE MICRO

CHALLENGES AND OPPORTUNITIES FOR EXTREMELY ENERGY-EFFICIENT PROCESSORS

Getting the Best of Both Worlds

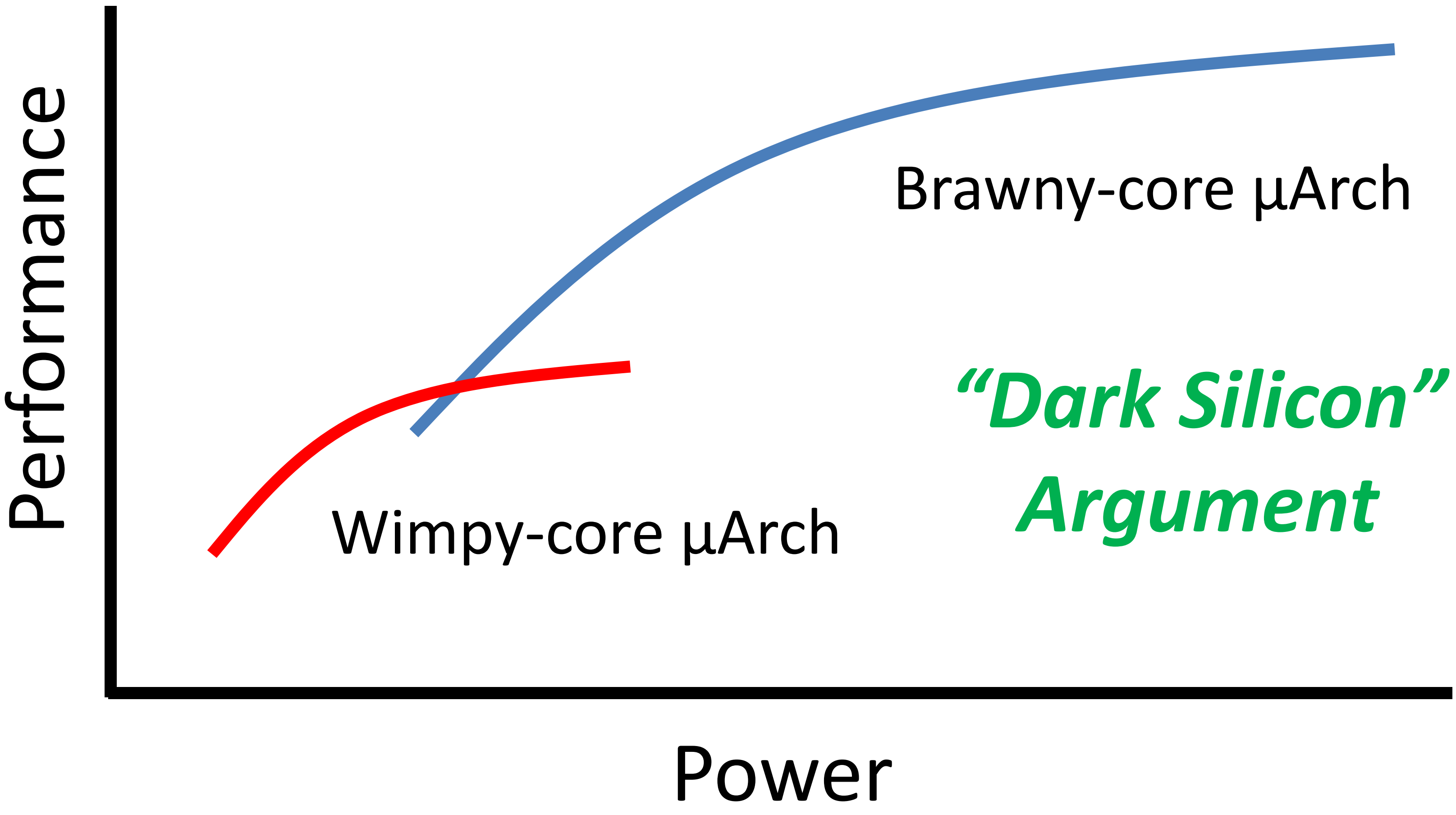


Performance when you need it,
Lowest power when
you don’t…

… with broader
dynamic
range

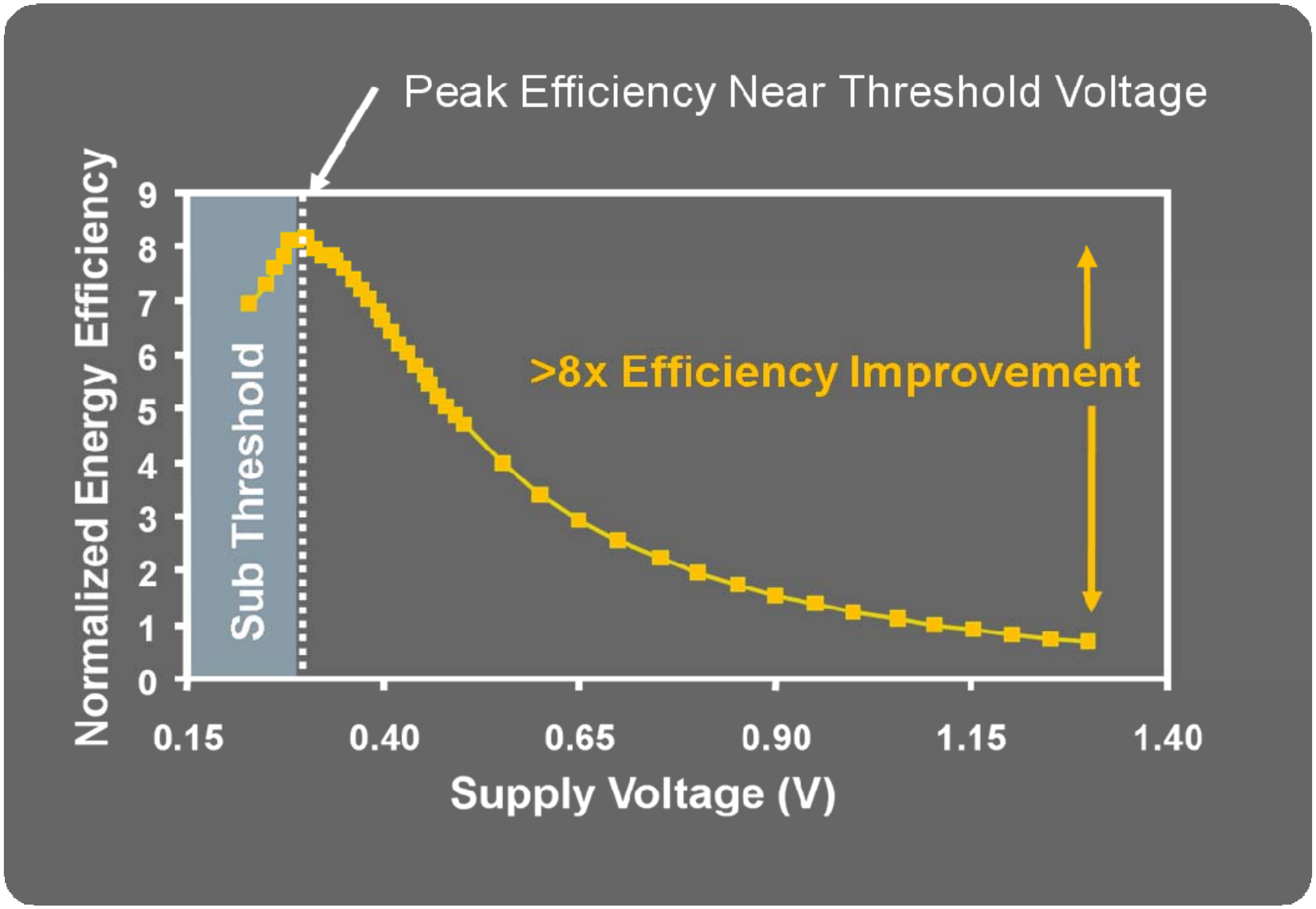
But How?

Stitch together hetero cores?

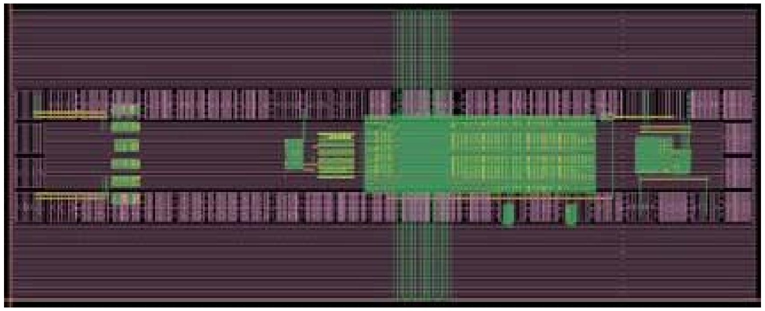


"Dark Silicon" Argument

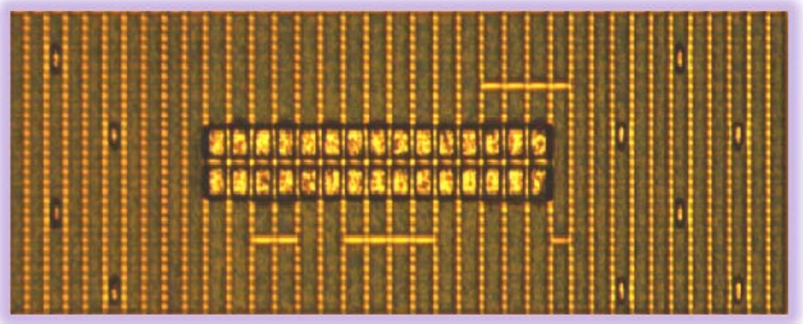
Another Option: NTV Circuits



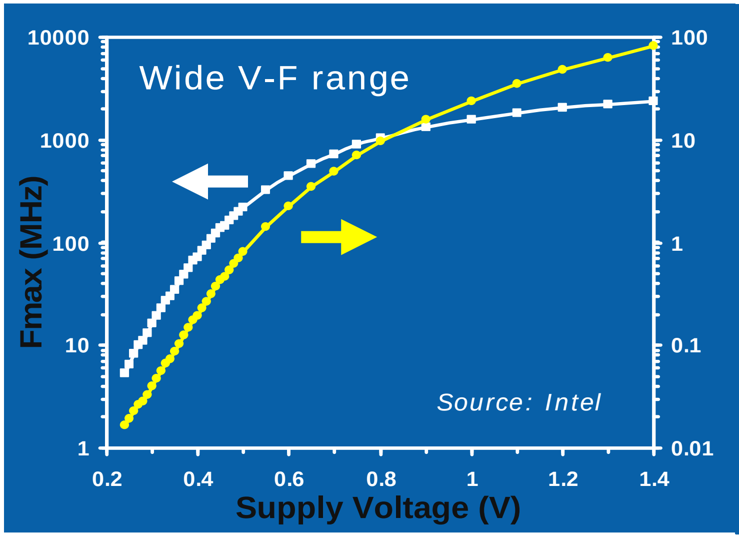
Experimental Intel NTV Test Chips



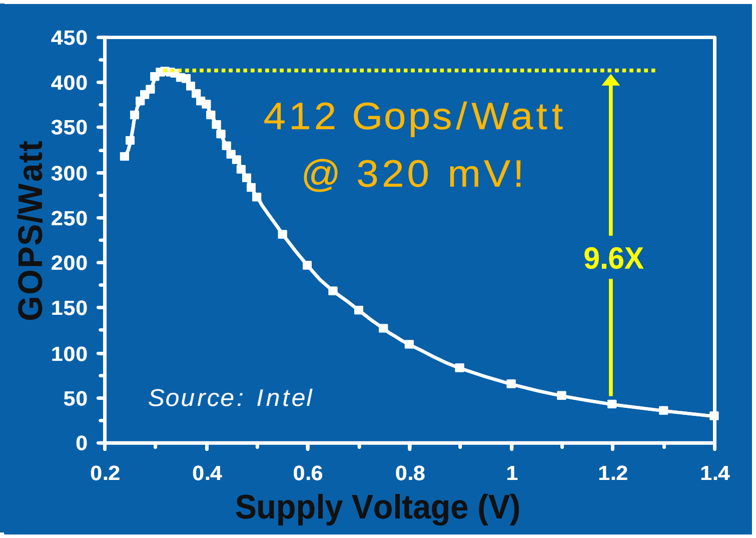
Motion Estimation Engine



AES Encryption Accelerator



Source: Intel



Source: Intel

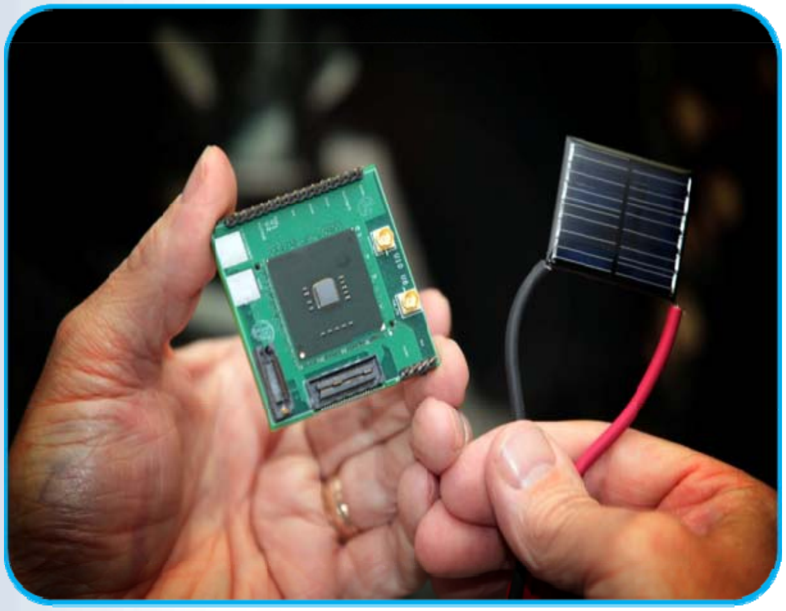
Kaul, et al.
JSSC
Jan 2009

Claremont: A Full NTV Processor

First processor to demonstrate benefits of Near Threshold Voltage circuits

IA concept chip that can ramp from ultra low power (<10mW) to full performance

Scales to over 10X the frequency when running at nominal supply voltage



Boots OS and powered by a solar cell!

Broader dynamic range likely…

… whether through hetero µ-arch, or NTV, or …

So what does this mean for the Cloud?

Cloud workloads tend to be "scale-out" in nature…

… so, operate at the low-power point to increase number of cores/nodes within a given power constraint?

But Cloud workloads also have QoS requirements…

… so, identify the “serial section” and run as fast as possible?

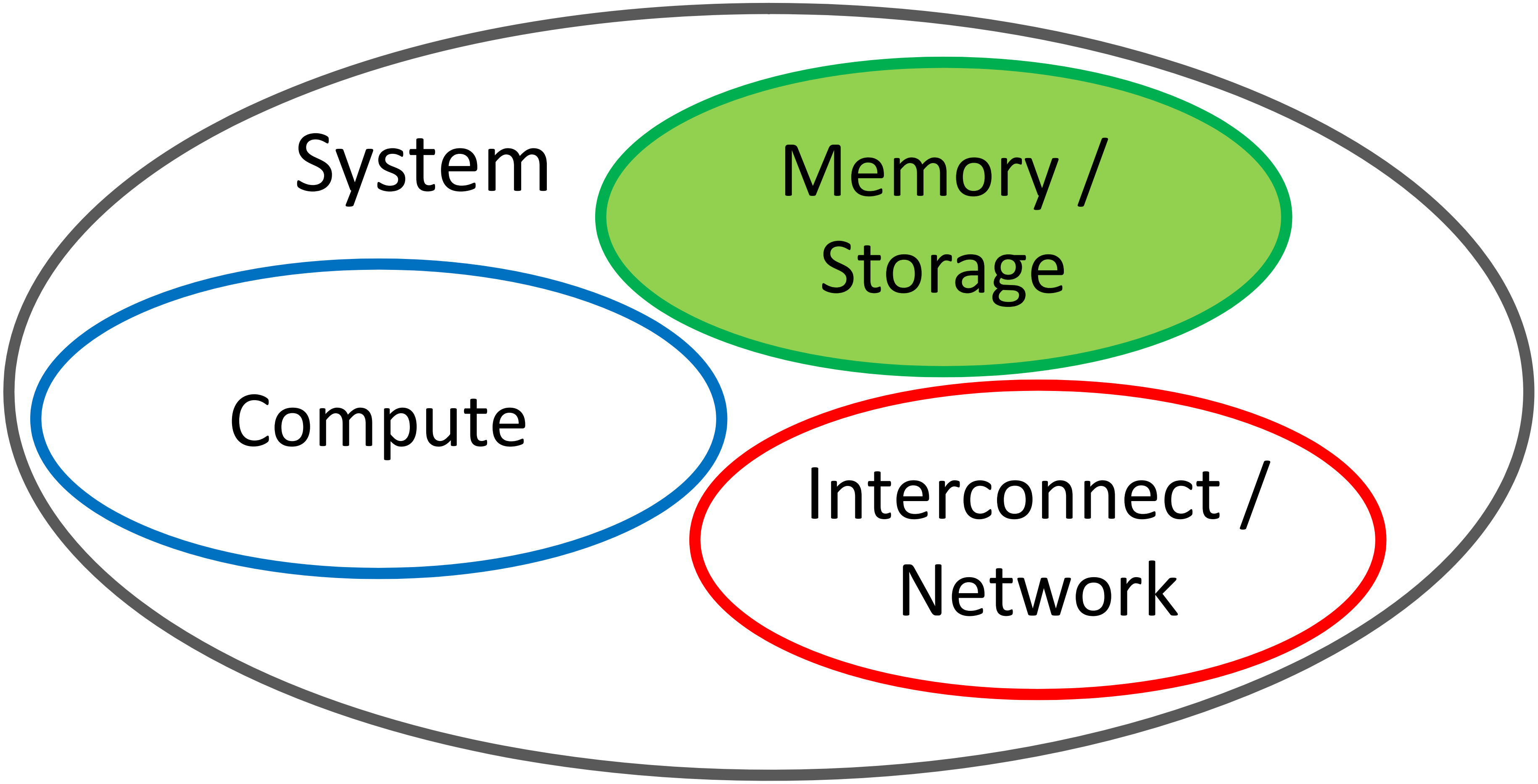
Of course, no single correct answer…

… it all depends on (ever changing) cloud workloads

Shift debate from brawny vs. wimpy to…

… when do workloads need to run at different ends of the efficiency spectrum, and can we determine this dynamically?

Finding Balance



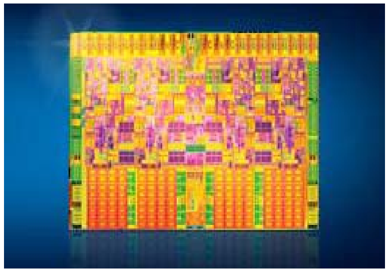
System

Memory / Storage

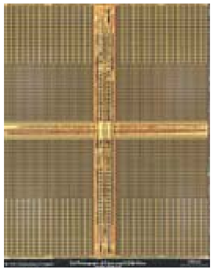
Compute

Interconnect / Network

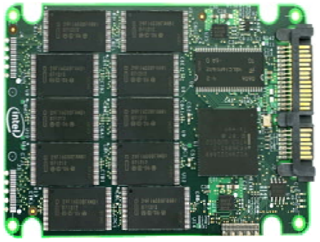
Revisiting the Memory-Storage Hierarchy (again)



On CPU SRAM



DRAM



NVM Storage

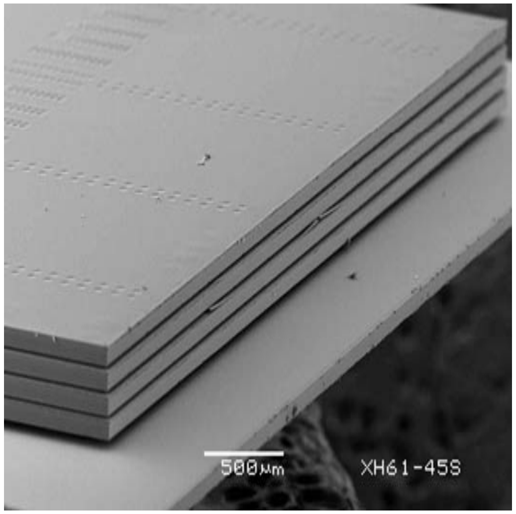
Storage (HDD)

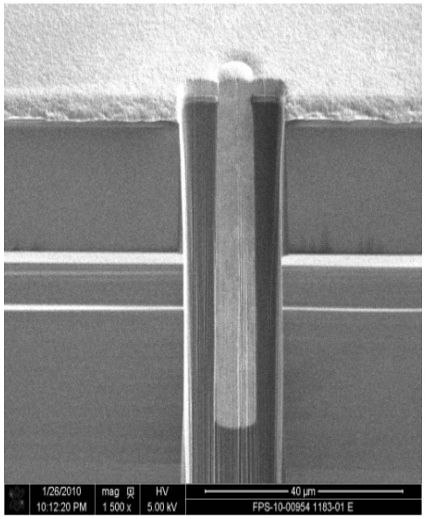
Latency	1x	~=20x	~=20,000x	~=10,000,000x
$$/bit	1x	~=0.05x	~=0.005x	~=0.0005x

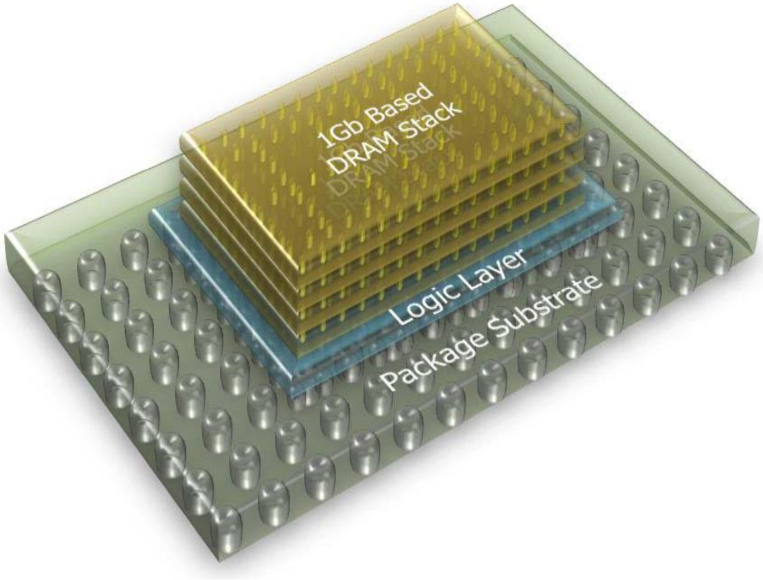
- New Players in the Hierarchy:
- Storage Class Memory & Stacked DRAM

Stacked DRAM

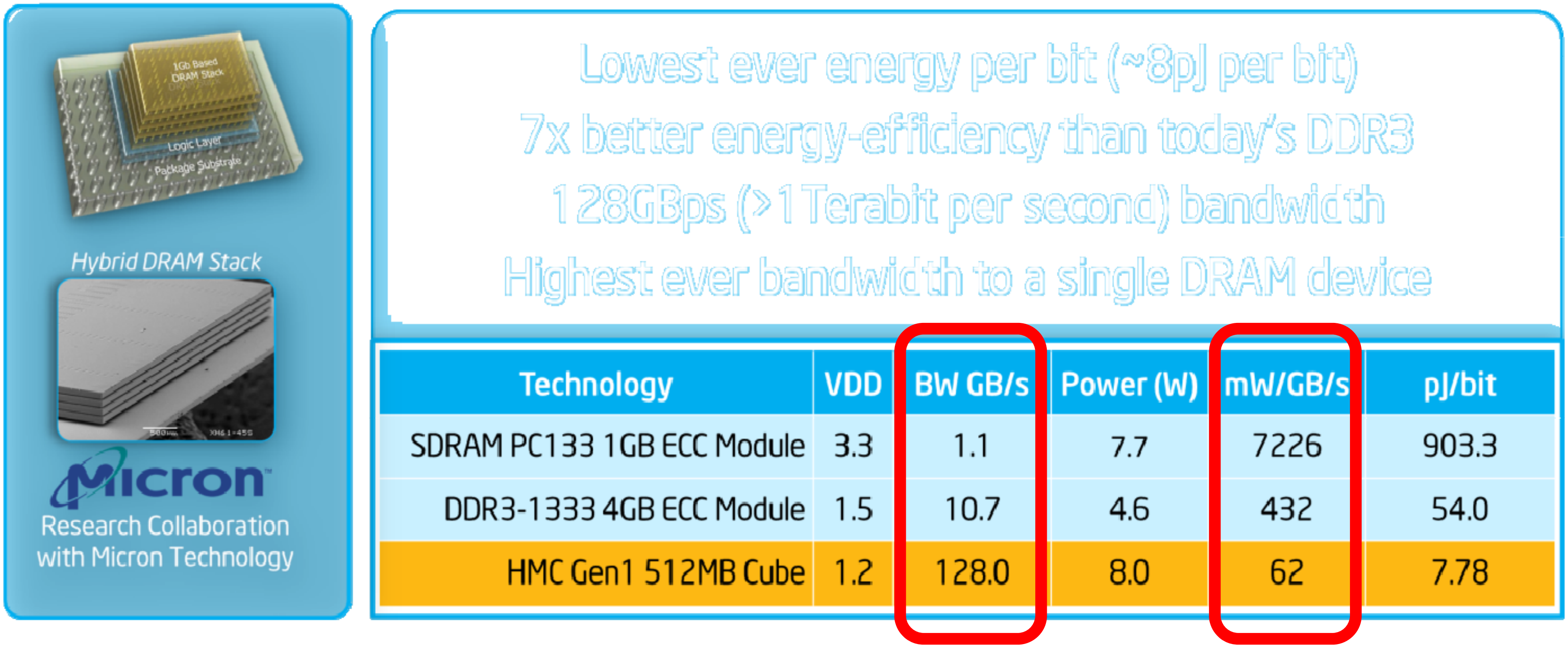
- DRAM die are stacked and interconnected with thousands of TSVs (Through-Silicon Vias)







Experimental Hybrid Memory Cube


1Gb Based DRAM Stack
Logic Layer
Package Substrate

Hybrid DRAM Stack

Micron Research Collaboration with Micron Technology

Lowest ever energy per bit (~8pJ per bit)

7x better energy-efficiency than today's DDR3

128GBps (>1 Terabit per second) bandwidth

Highest ever bandwidth to a single DRAM device

Technology	VDD	BW GB/s	Power (W)	mW/GB/s	pJ/bit
SDRAM PC133 1GB ECC Module	3.3	1.1	7.7	7226	903.3
DDR3-1333 4GB ECC Module	1.5	10.7	4.6	432	54.0
HMC Gen1 512MB Cube	1.2	128.0	8.0	62	7.78

From IDF 2011

Stacked DRAM offers high bandwidth and low latency…

… but capacity still relatively limited, leading to “near” and “far” memory

Storage Class Memories offer new capabilities…

… higher capacity & persistence (relative DRAM)
… lower read & write latencies (relative NAND)

But they also bring new issues…

… writes slower than reads

… and they wear out

Not likely to replace either DRAM or Disks…

… so best viewed as additions to the memory-storage hierarchy

Need a fresh look at OS, VMM, Filesystem and Database Design

How much legacy code could be removed given persistent memory, and what new interfaces and abstractions make sense?

Interface questions…

Block versus Load/Store

Expressing Write Durability

Expressing Update Atomicity

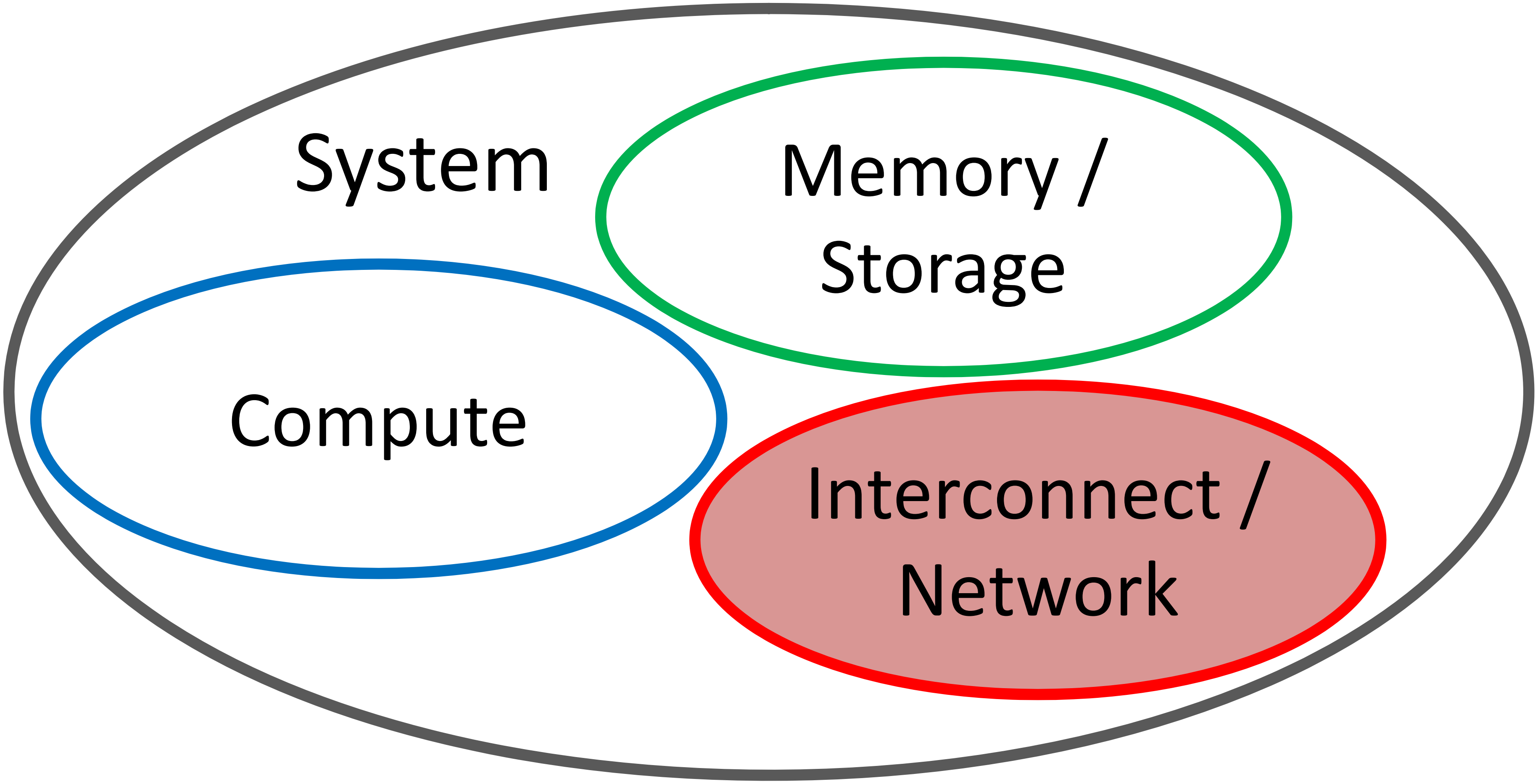
New Abstractions?

Move back to “single level store”?

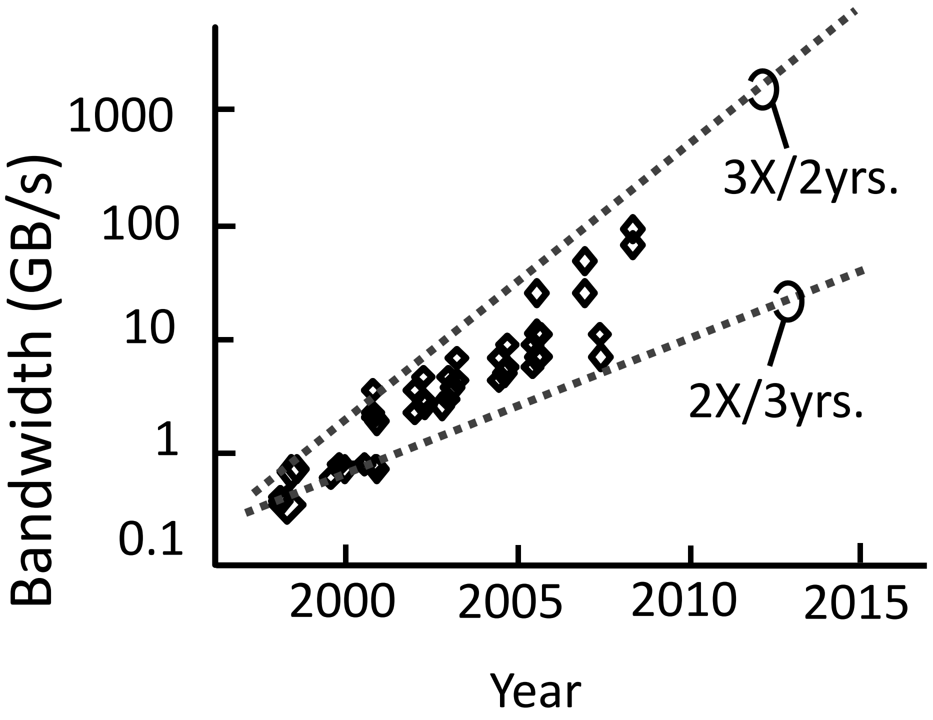
New protection models?

New execution models?

Finding Balance



Interconnect Bandwidth Trends



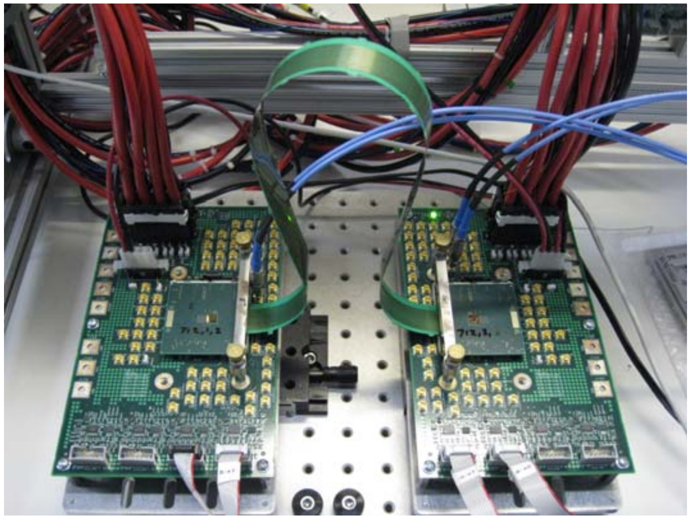
Bandwidth Drivers:

CPU ↔ Memory
CPU ↔ CPU
CPU ↔ Peripheral
CPU ↔ I/O bridge

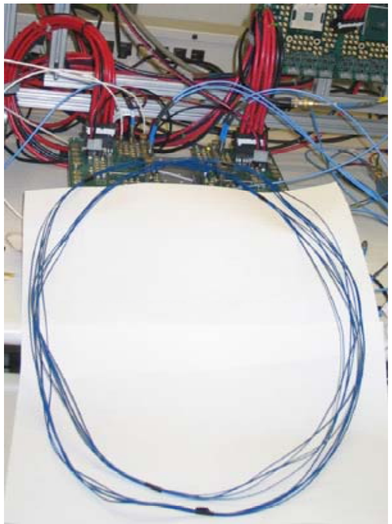
Most apps <1m length

What is the role of electrical & optical interconnects in meeting these bandwidth goals?

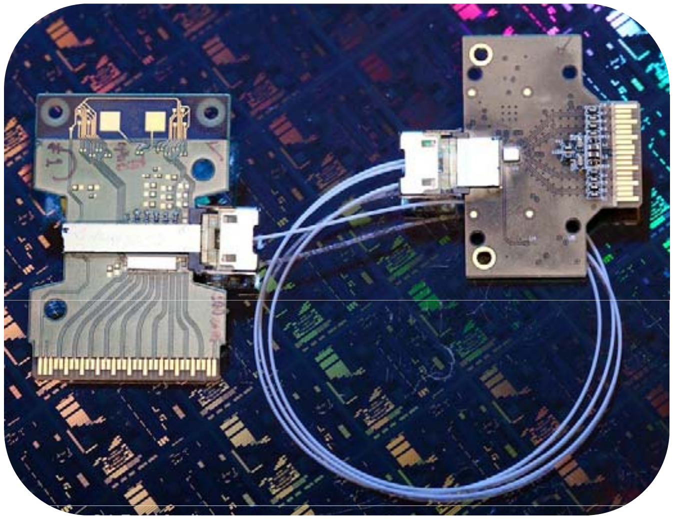
Energy-Efficient Scalable I/O



20” flex interconnect



3m twinax cable



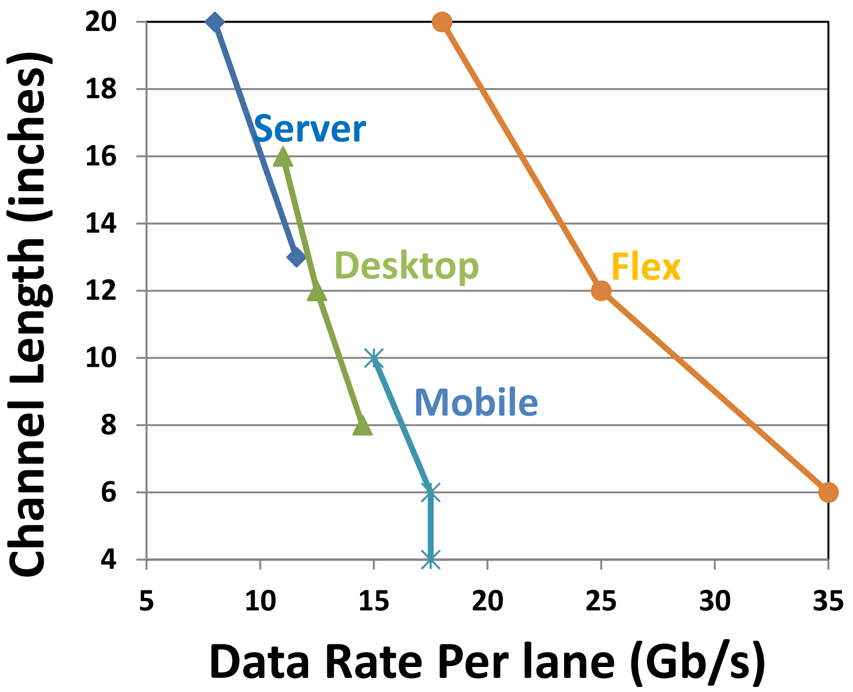
Silicon Photonics

- New cabled interconnects for scaling electrical links
- Silicon photonics for longer-distance optical links

Distance Limits to Bandwidth Scaling

- Traditional interconnects limit electrical signaling to ~10-18 Gb/s

- Top-of-package, cabled interconnects provide scaling to ~35 Gb/s

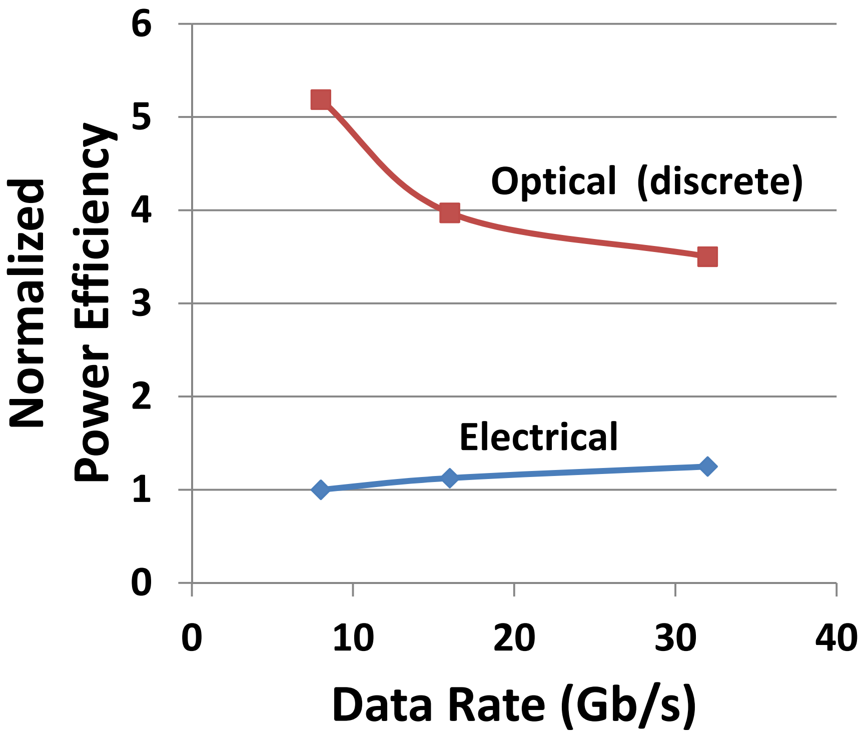


Source: Randy Mooney

Link Energy Efficiency

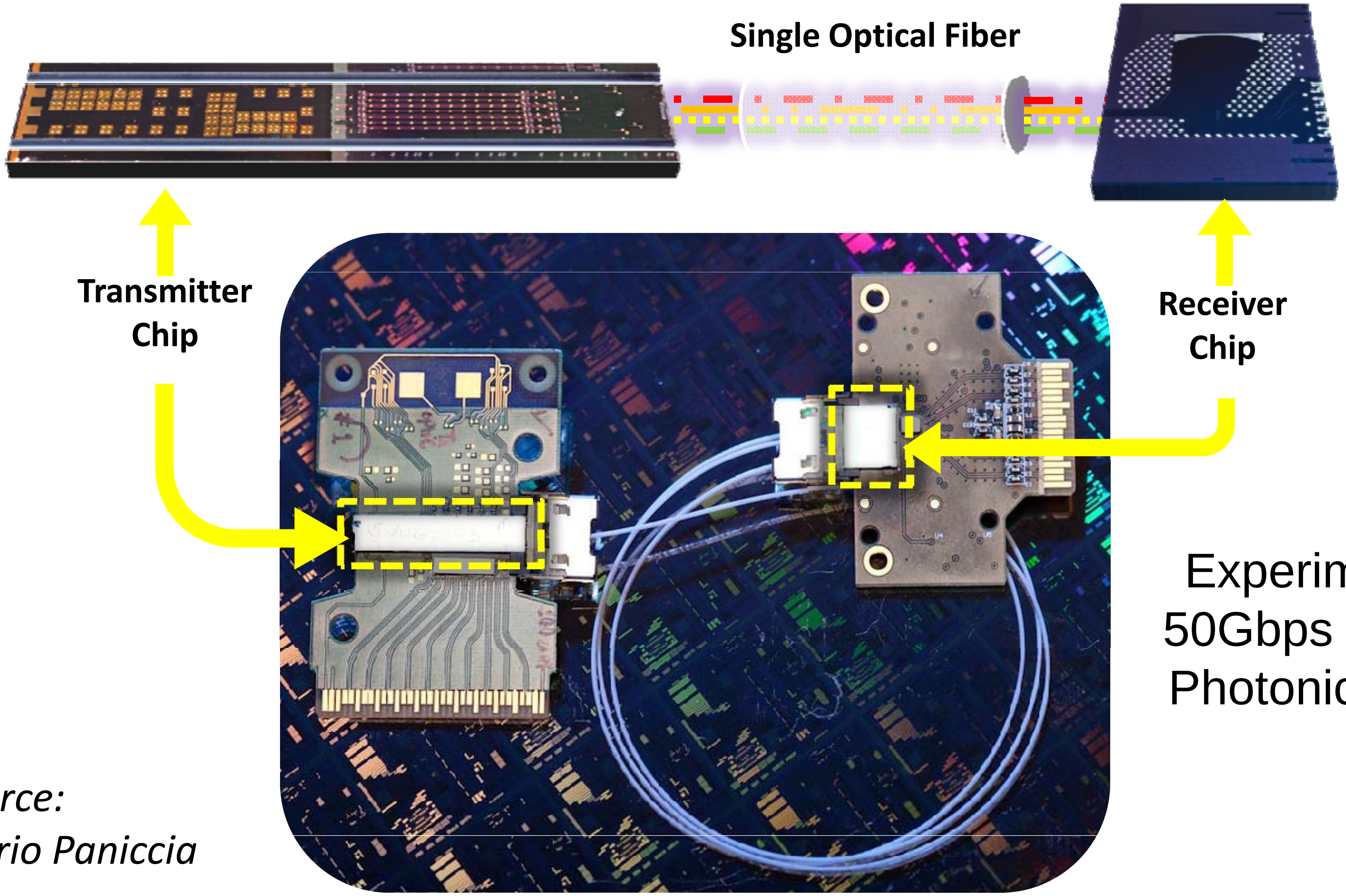
- Electrical links more energy efficient at short distances

- Optical links for higher data rates at distance



Source: Randy Mooney

Integrated Silicon Photonics



Single Optical Fiber

Transmitter Chip

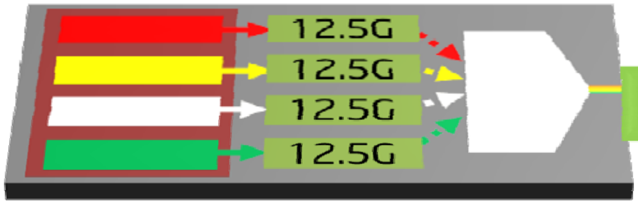
Receiver Chip

Experimental 50Gbps Silicon Photonics Link

Source: Mario Paniccia

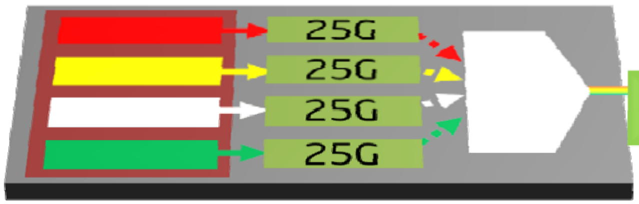
The Path to Tera-scale Data Rates

Today: 12.5 Gbps x 4 = 50Gbps



Scale UP

25 Gbps x 4 = 100Gbps



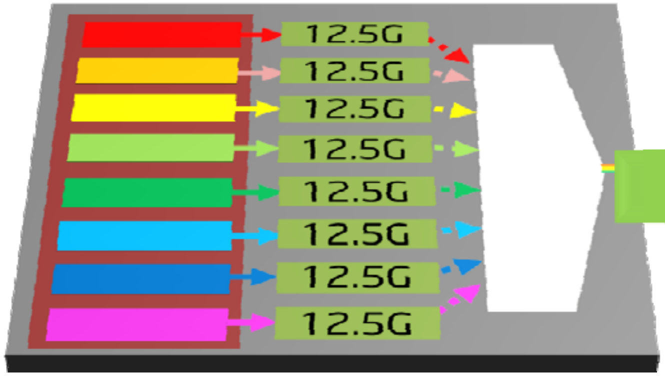
40G, 100G…

Scale OUT

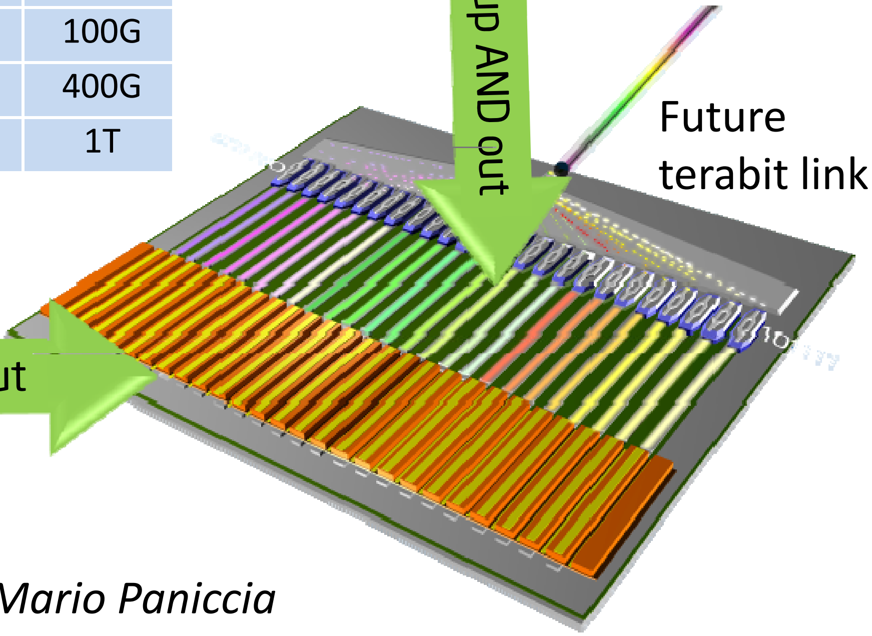
Speed	Width	Rate
12.5	x4	50G
12.5	x8	100G
25	x16	400G
40	x25	1T

Scale up AND out

12.5 Gbps x 8 = 100Gbps



Scale up AND out



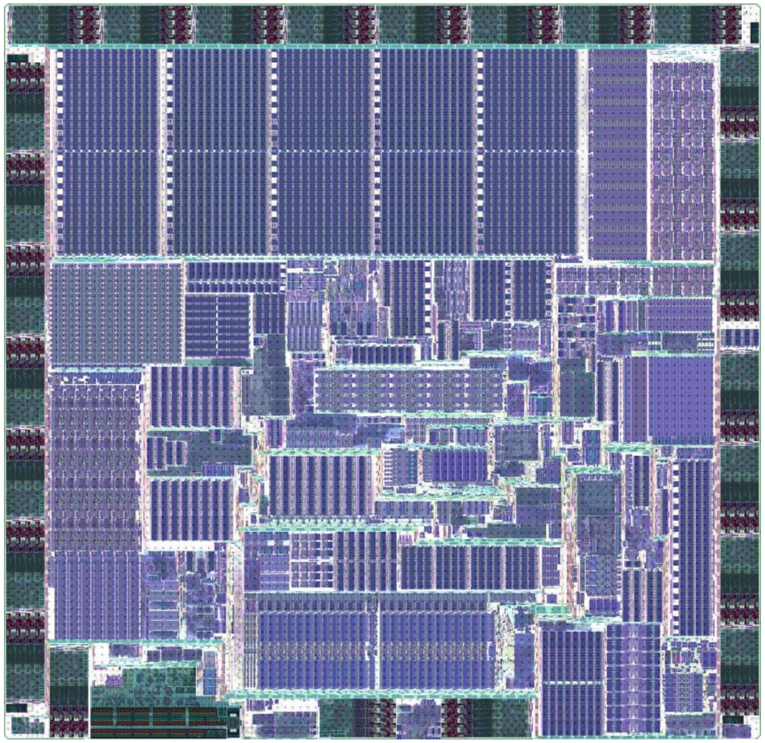
Future terabit link

x16, x32…

Source: Mario Paniccia

Network switch latencies improving…

- Fulcrum Alta Switch
  - 400ns latency
  - 72x10G / 18x40G
- Design
  - Asynchronous Logic
  - Microcode Programmable
  - 15MB Shared Memory

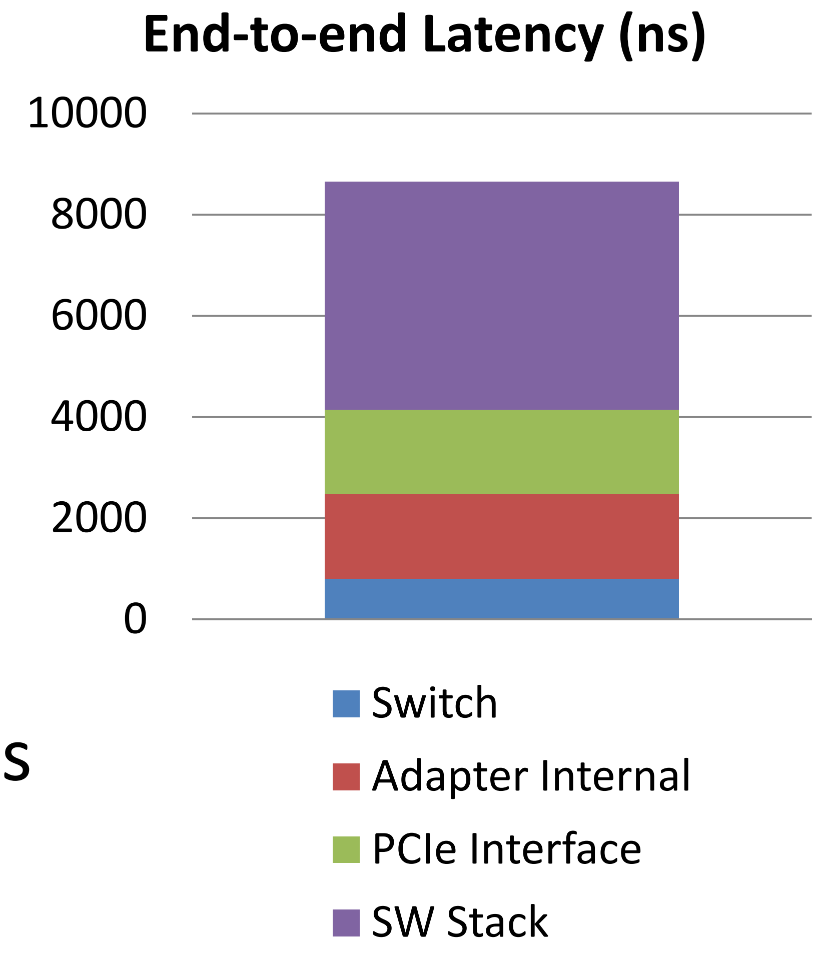


From Hot Chips 2011
-- Mike Davies, Fulcrum

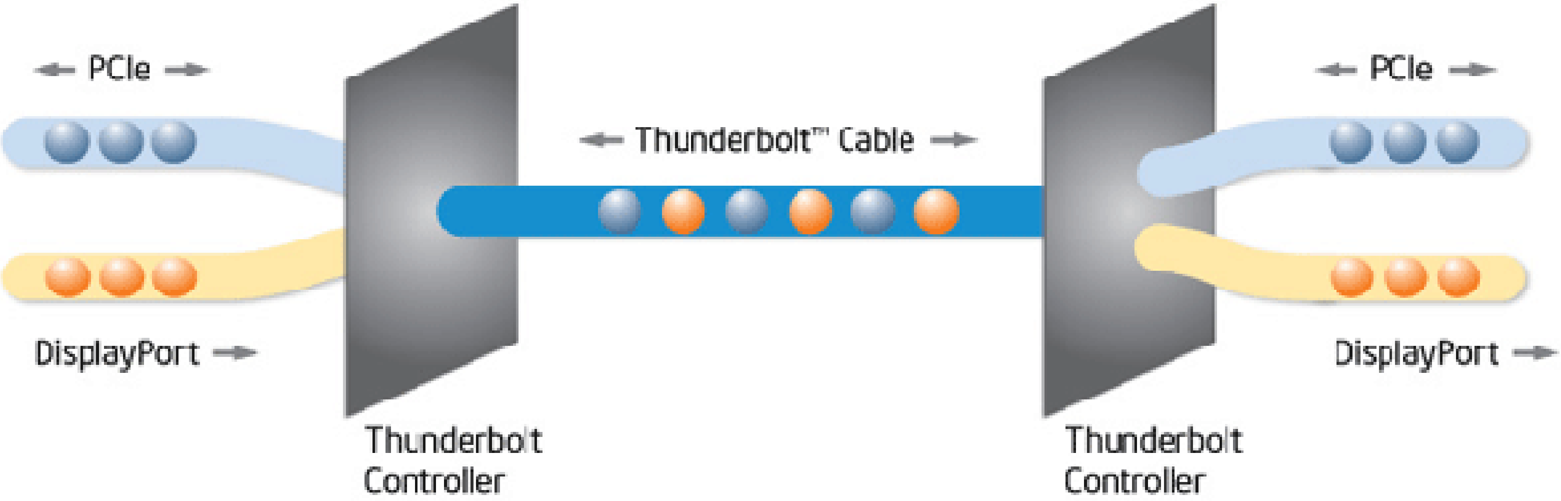
FULCRUM microsystems

… but room to improve the “last inch”

- Ethernet Latencies
  - Niantic NIC
  - Fulcrum Switch
  - Socket Semantics
- Opportunities
  - Integration
  - SW Stack Optimizations
  - New protocols
    (e.g., TCP alternatives)



Converging I/O



Intel® Thunderbolt ™
(a.k.a., Light Peak)

If we can bring 1000s of nodes within < 5 µsec of one another…

… what new cloud applications would that enable?

Is it time to bring HPC-style topologies to Cloud datacenters?

Replace hierarchical tree topologies with direct networks (torus, etc.)?

Is it time to rethink network protocols?

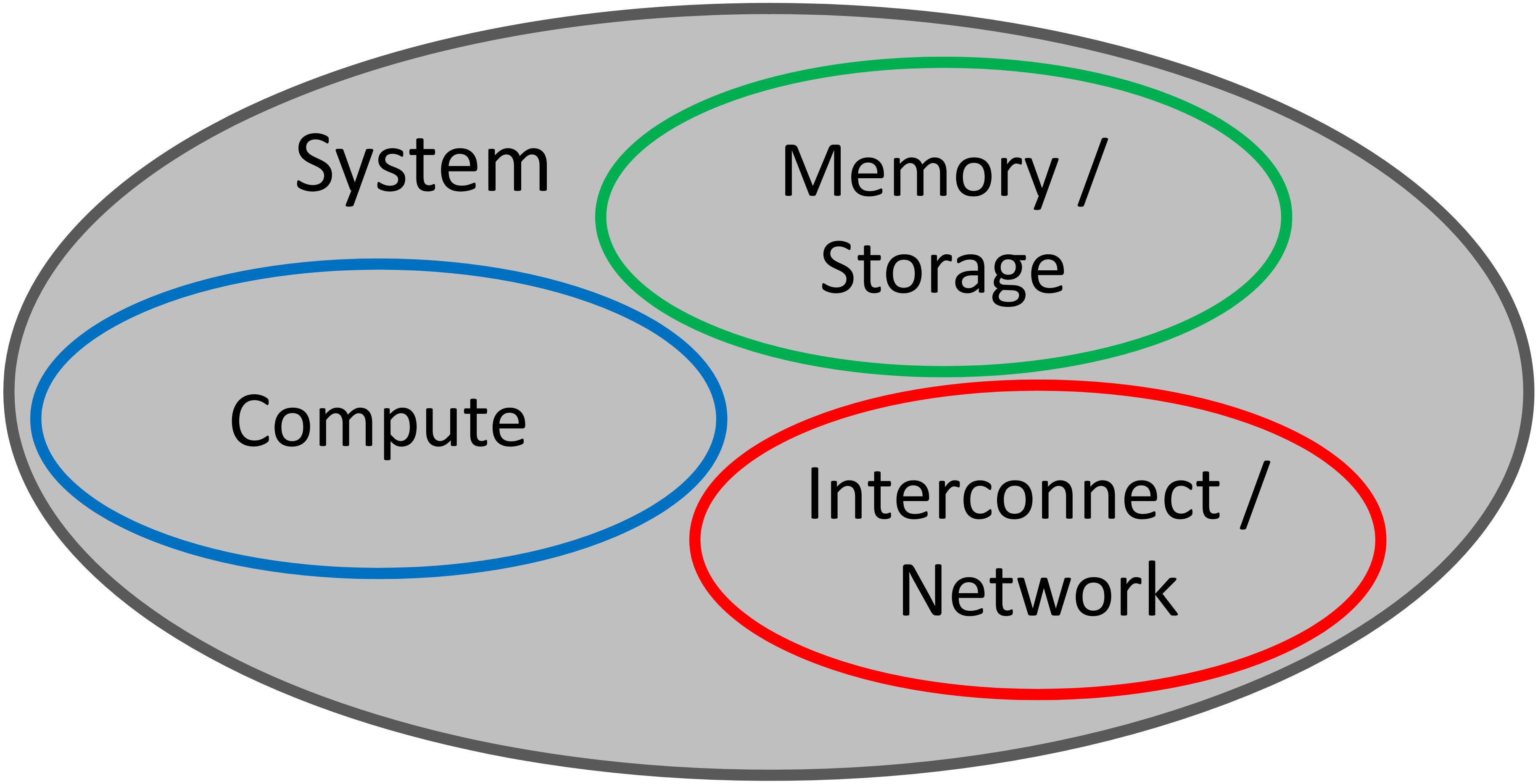
Alternatives to TCP and RDMA?

Ways to access remote NVMs?

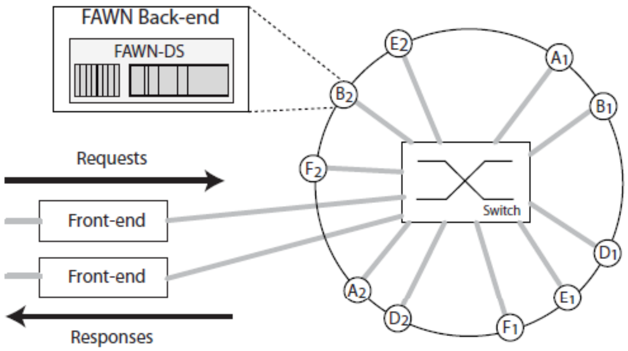
How can we achieve energy proportionality with networks…

… just as with compute nodes?

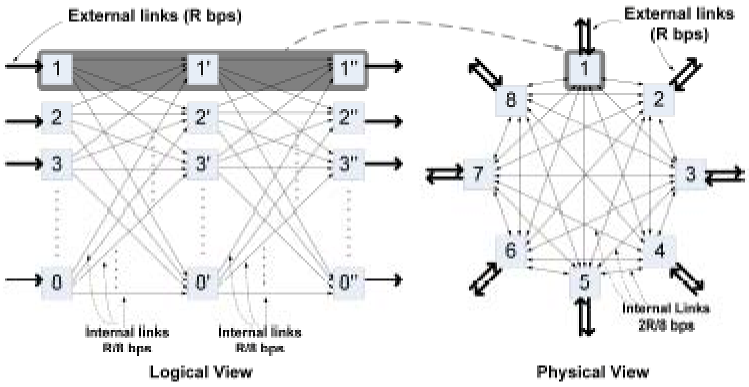
Finding Balance



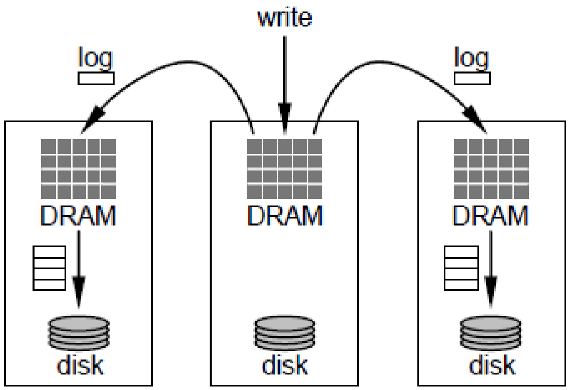
Interesting Recent Systems Projects



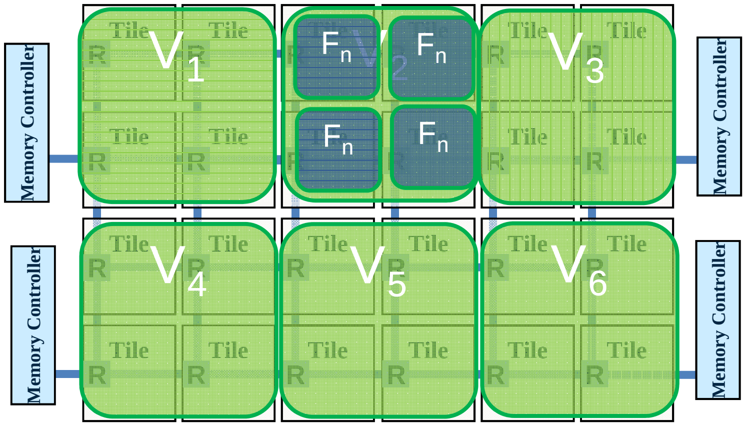
FAWN (CMU)



RouteBricks (Berkeley)



RAMCloud (Stanford)



Single-Chip Cloud Computer (Intel)

Common Themes:
These projects….

(1) Push system to a new “balance point”

(2) Combine innovative hardware and software

(3) Blur edges btw. compute/storage/networking

Can we build a testbed that enables flexible resource rebalancing?

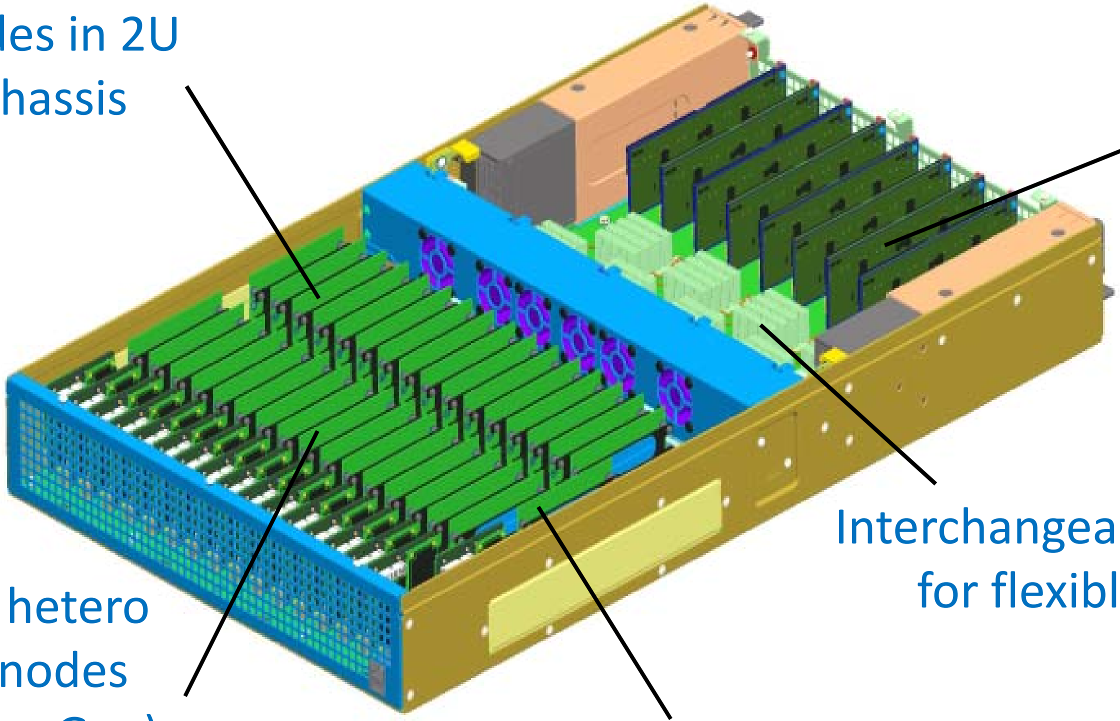
Controlled combinations of compute nodes, memory and storage resources in different proportions and topologies

µCluster Prototype: Goals

- Emphasis on configuration flexibility
  - – Mix and match CPU modules
  - – Different fabric topologies
  - – IO resource balance

- Extensive instrumentation
  - – For power and performance
  - – To understand emerging Cloud workload requirements

µCluster Prototype



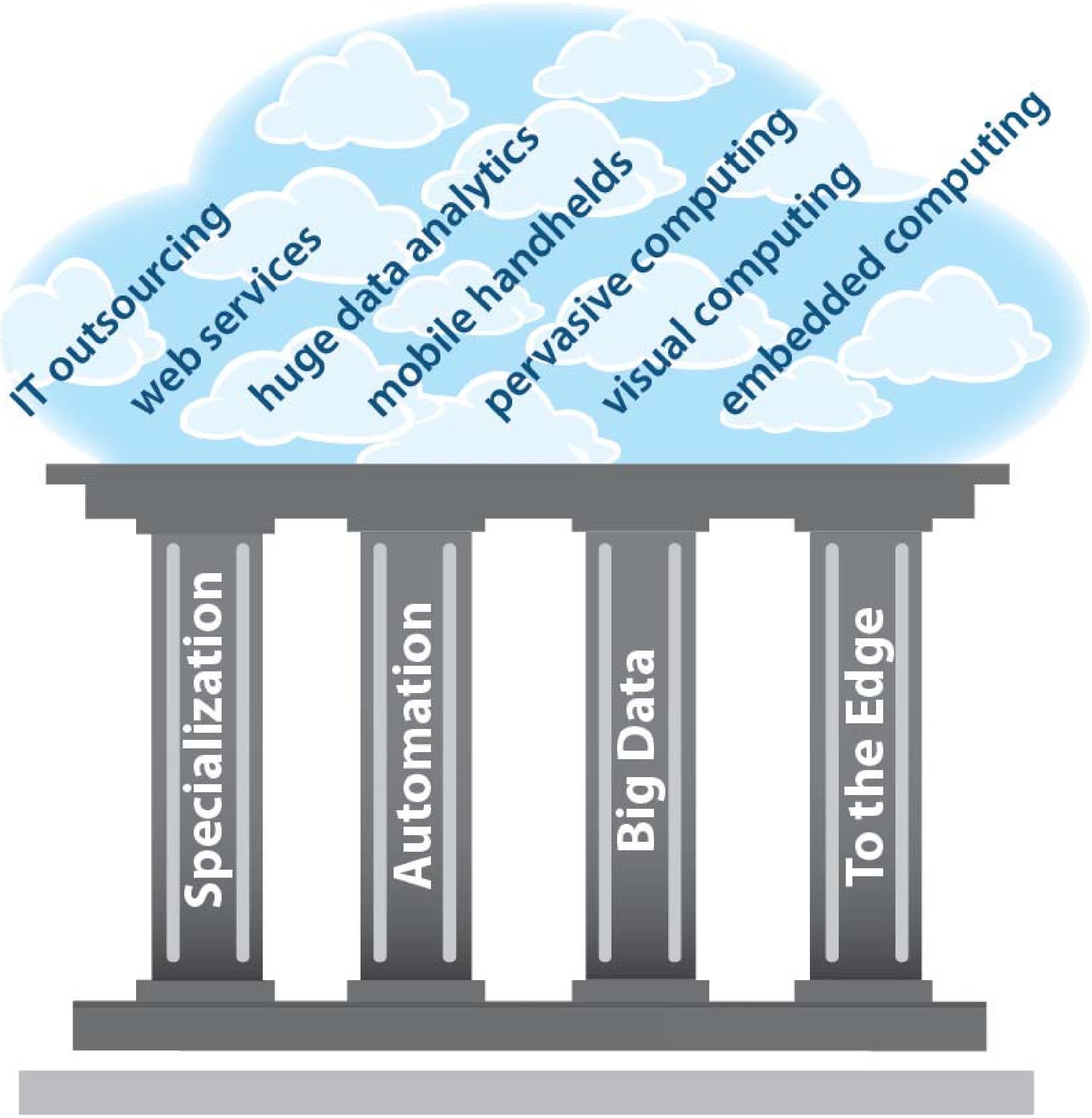
32 compute nodes in 2U chassis

Interconnect based on Light Peak or Ethernet

Interchangeable Fabric Board for flexible topologies

Option for hetero compute nodes (Atom, Xeon, Gen)

Instrumentation for power and perf measurements



IT outsourcing
web services
huge data analytics
mobile handhelds
pervasive computing
visual computing
embedded computing
Specialization
Automation
Big Data
To the Edge

Wish List

- Brainstorm ways to use µCluster
- New Workloads. What's beyond Hadoop, etc.
- New Tools. "Problem Diagnosis" work is great.
- Storage Class Memory Research. Focus on software and ideal interfaces to hardware.
- What emulators would enable such work?
- Take "to the edge" all the way to clients

A standing invitation to visit Intel!

For short, medium or long visits…