Achieving Energy Efficiency by HW/SW Co-design

Shekhar Borkar

Intel Corp.

Oct 28, 2013

This research was, in part, funded by the U.S. Government, DOE and DARPA. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

Outline

Compute roadmap & technology outlook

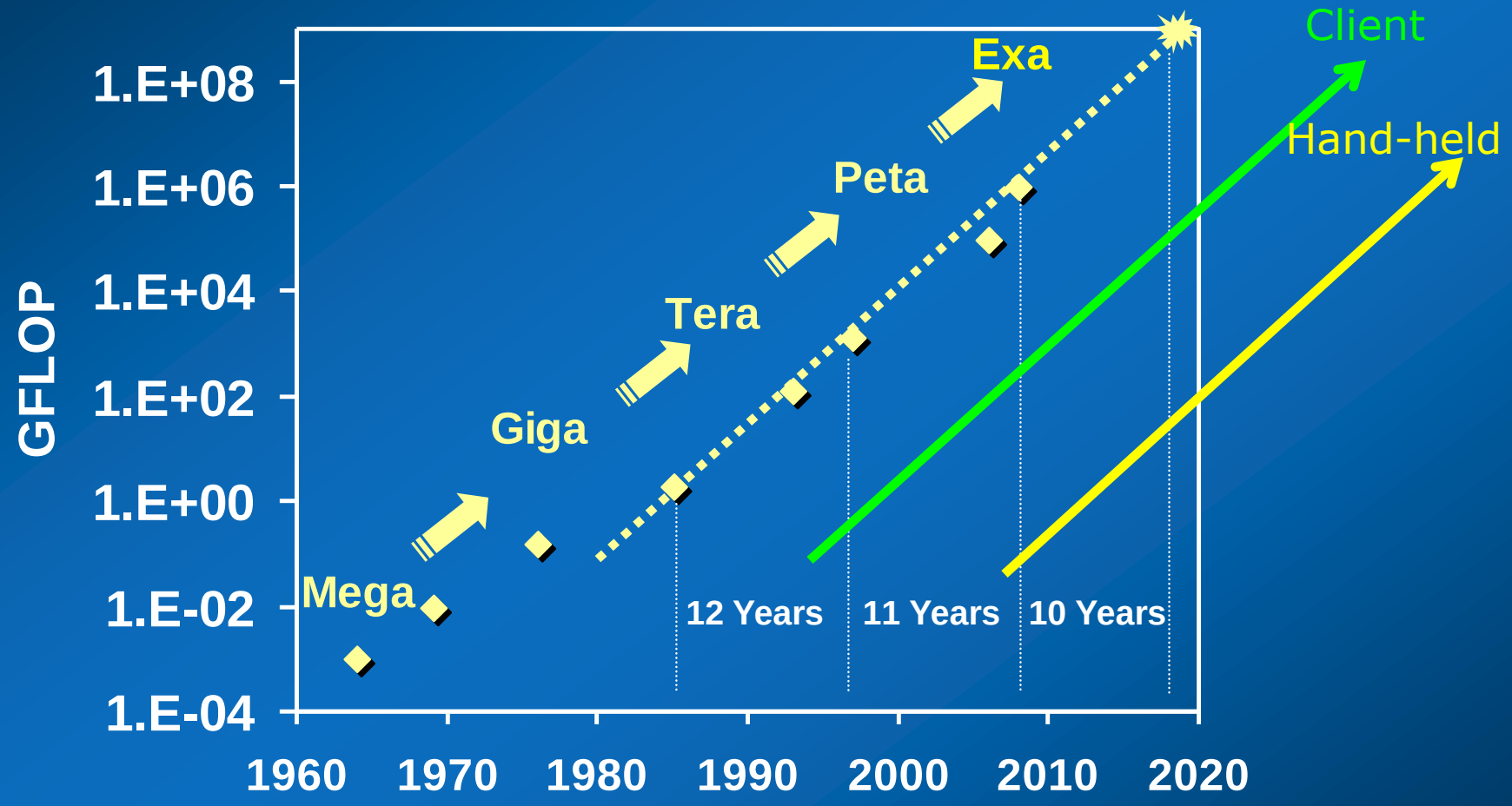
Challenges & solutions for:

- Compute,
- Memory, and
- Interconnect

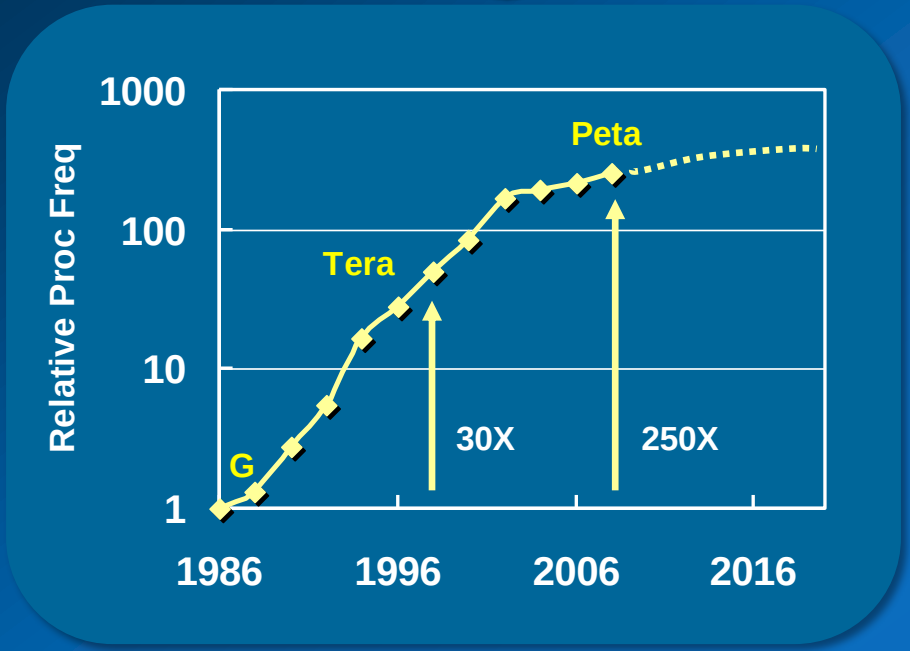
HW/SW Co-design–not just a buzz word!

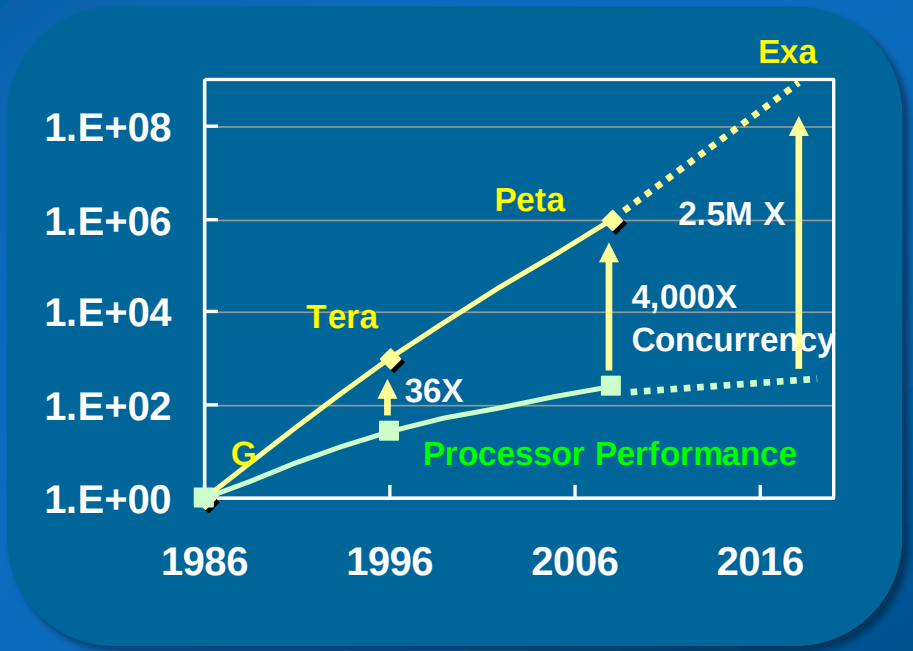
Summary

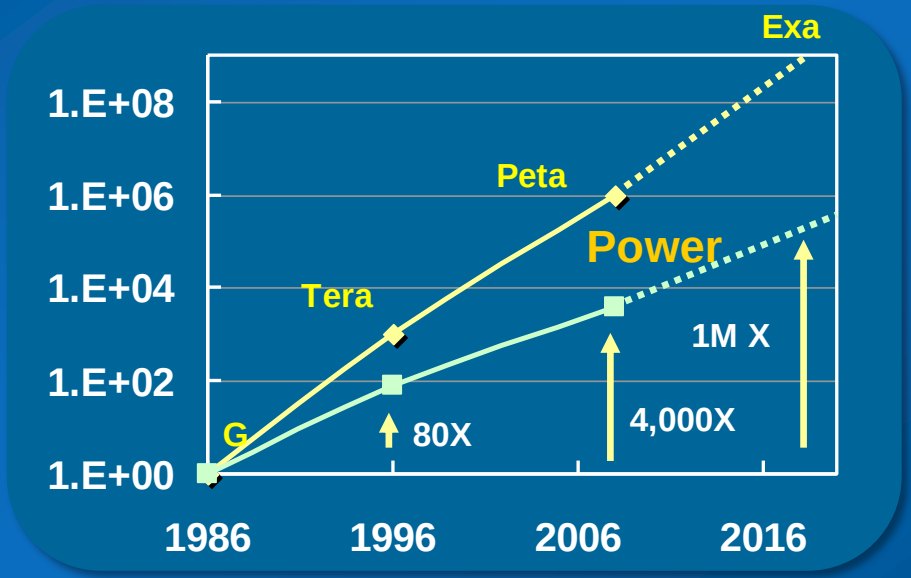
Compute Performance Roadmap



From Giga to Exa, via Tera & Peta







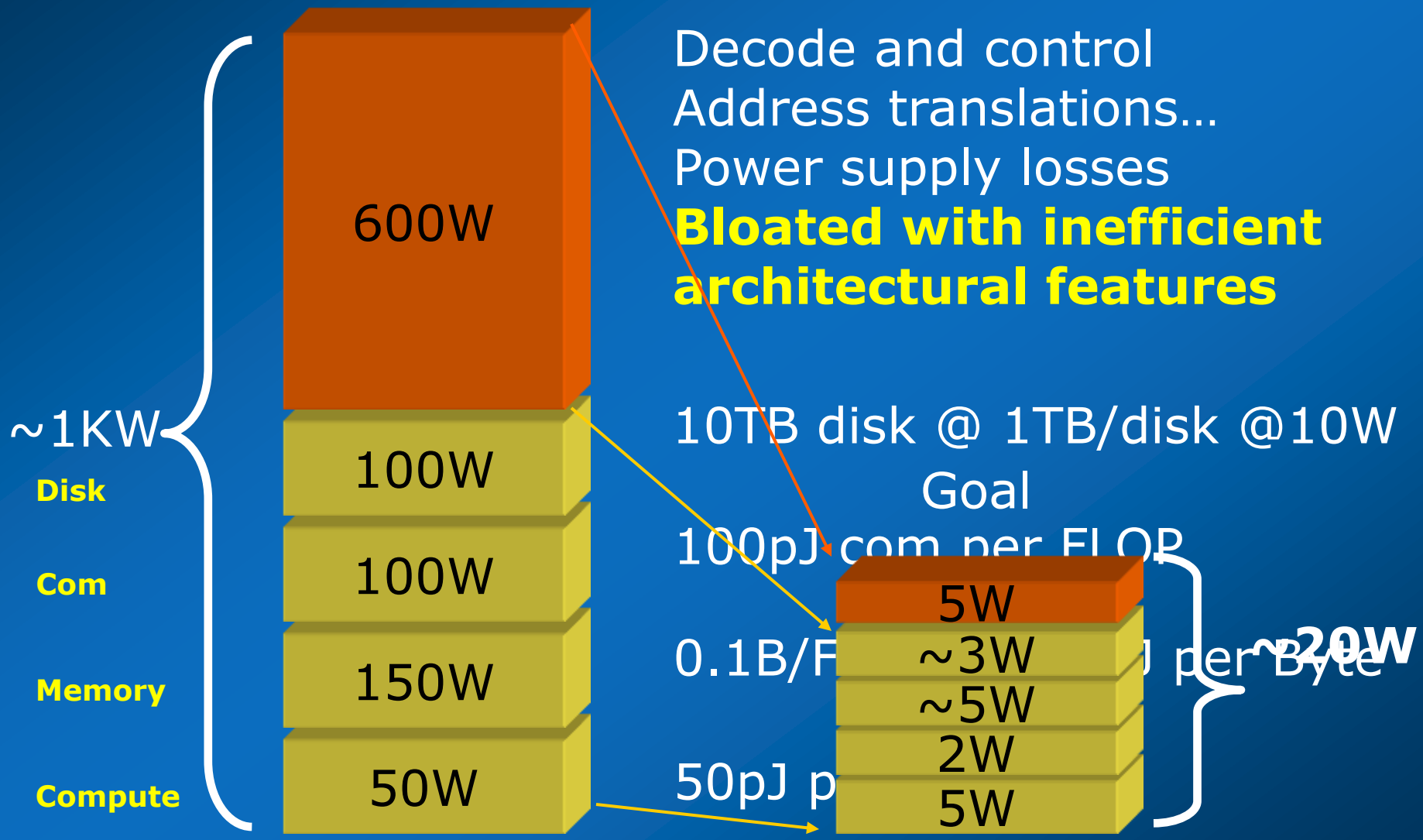
System performance increases faster

Parallelism continues to increase

Power & energy challenge continues

Where is the Energy Consumed?

Teraflop system today



The UHPC* Challenge

**DARPA, Ubiquitous HPC Program*

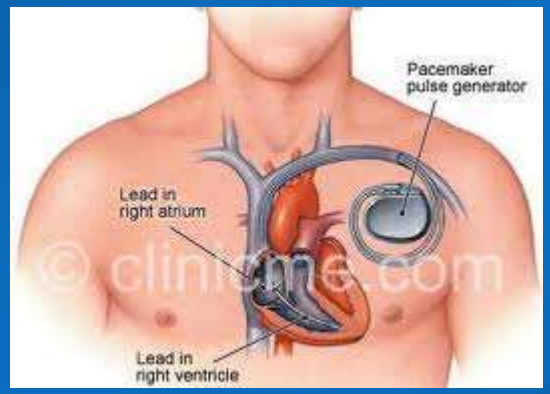
20MW, Exa

20KW, Peta

20W, Tera

20 pJ/Operation

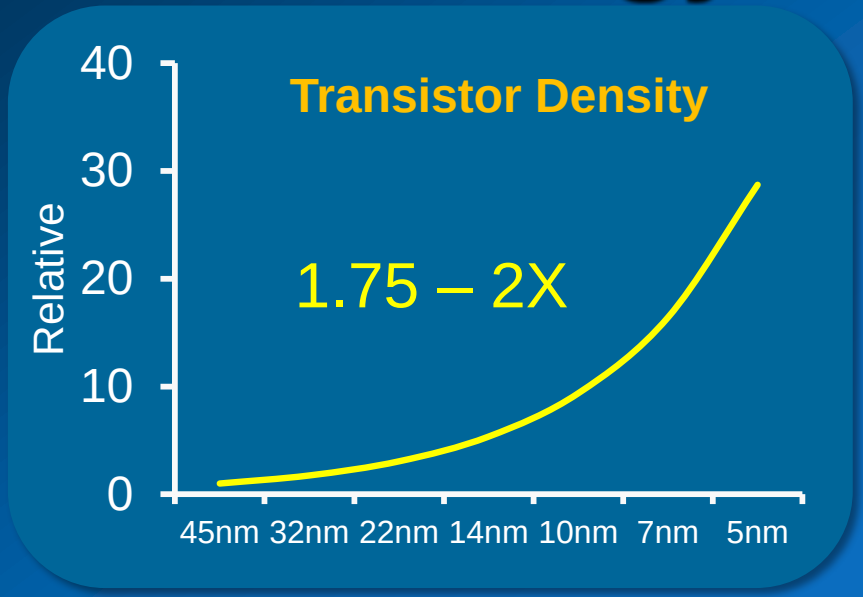
20 μW, Mega

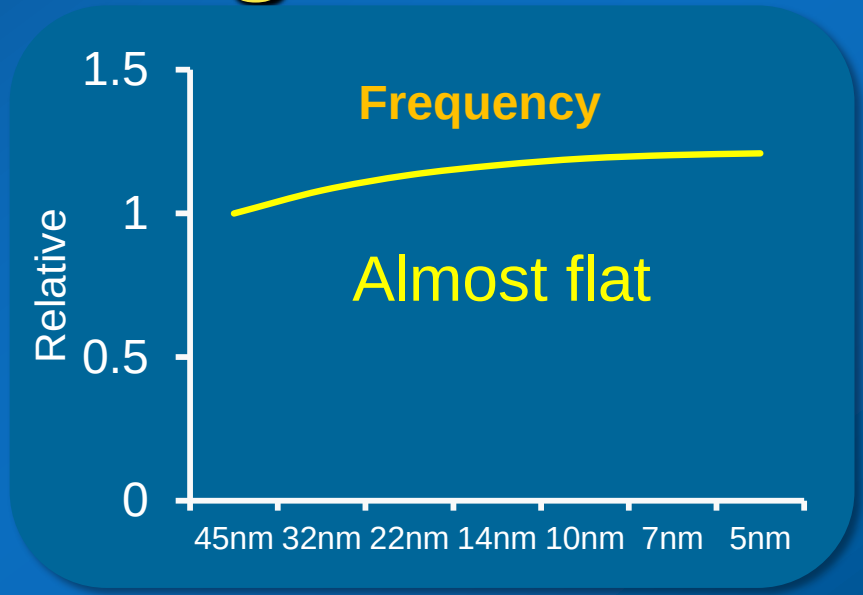


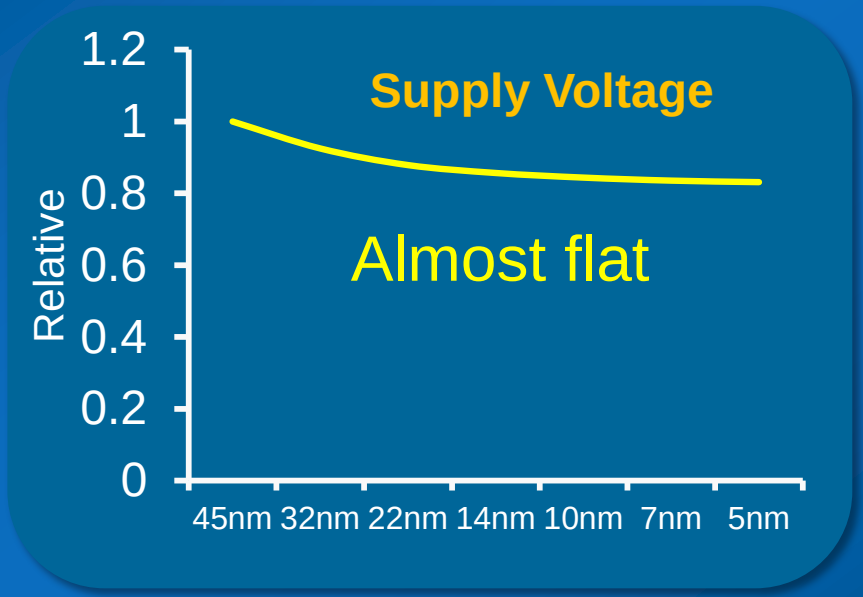
20 mW, Giga

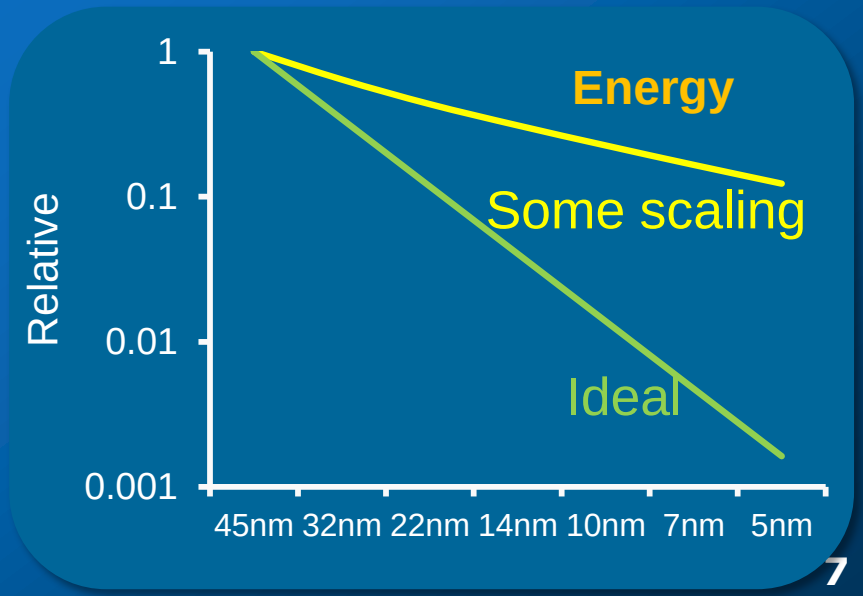
2W, 100 Giga

Technology Scaling Outlook

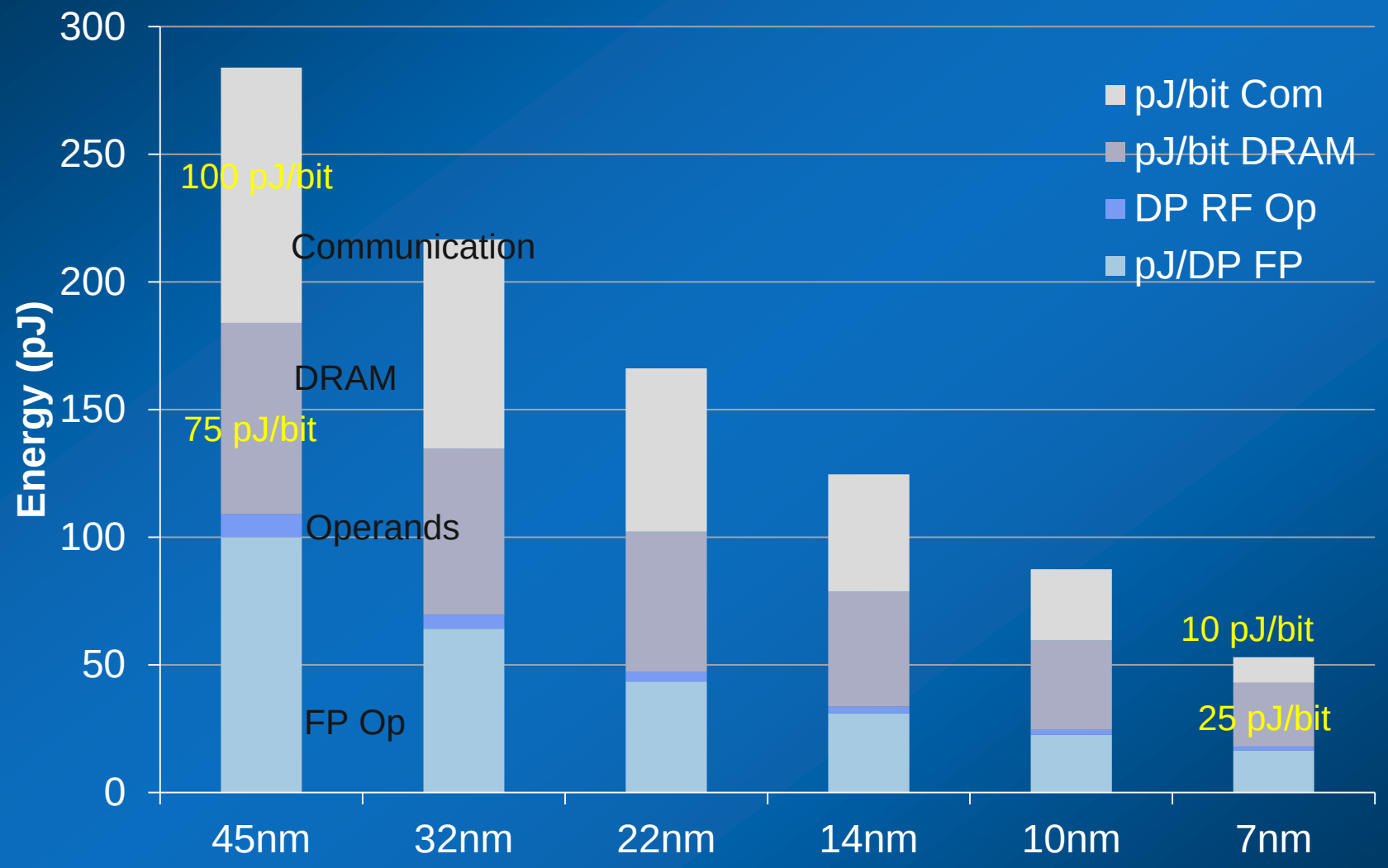








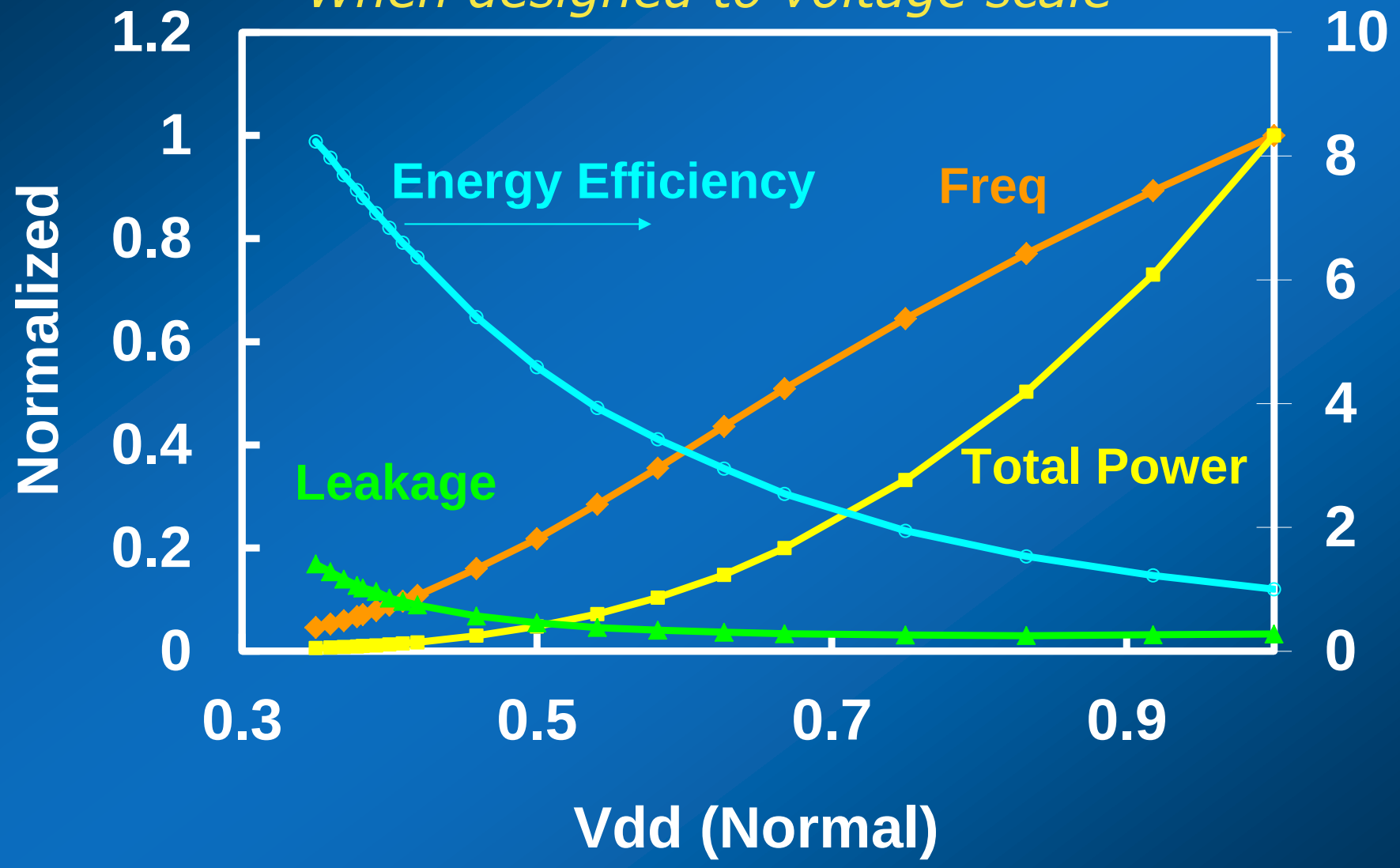
Energy per Compute Operation



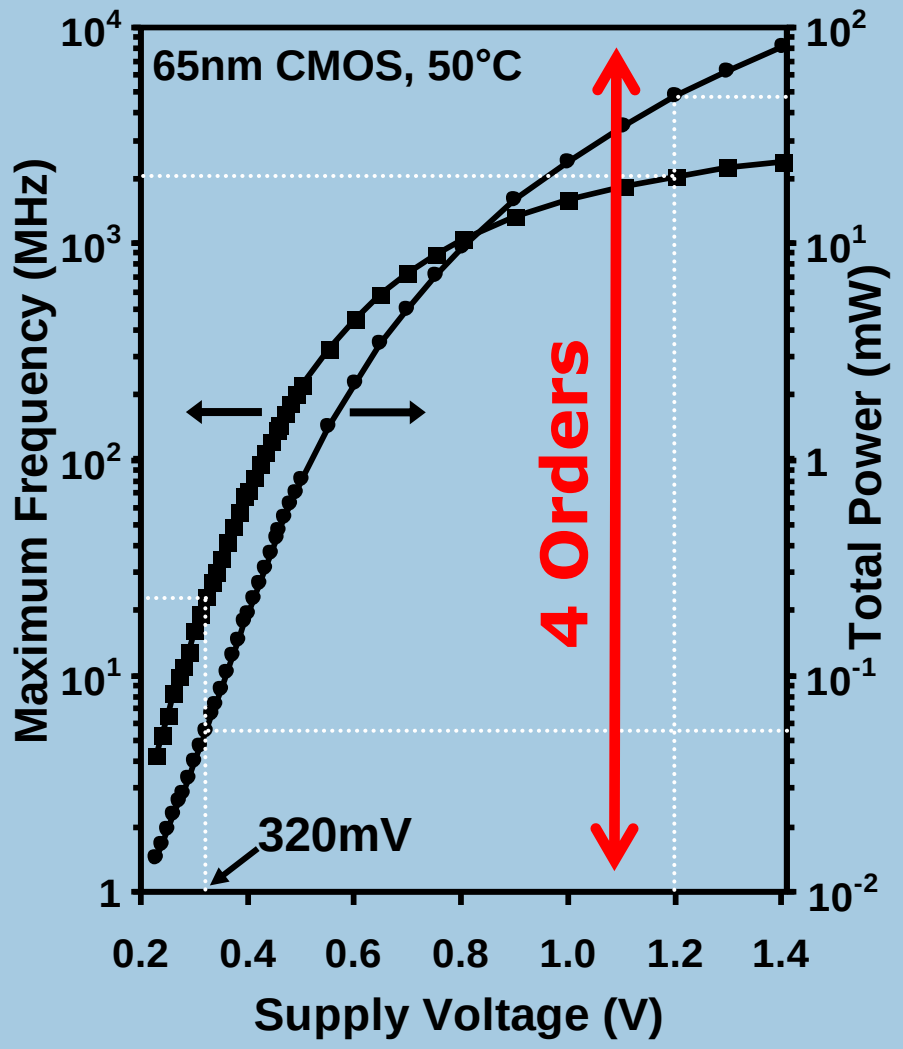
Source: Intel

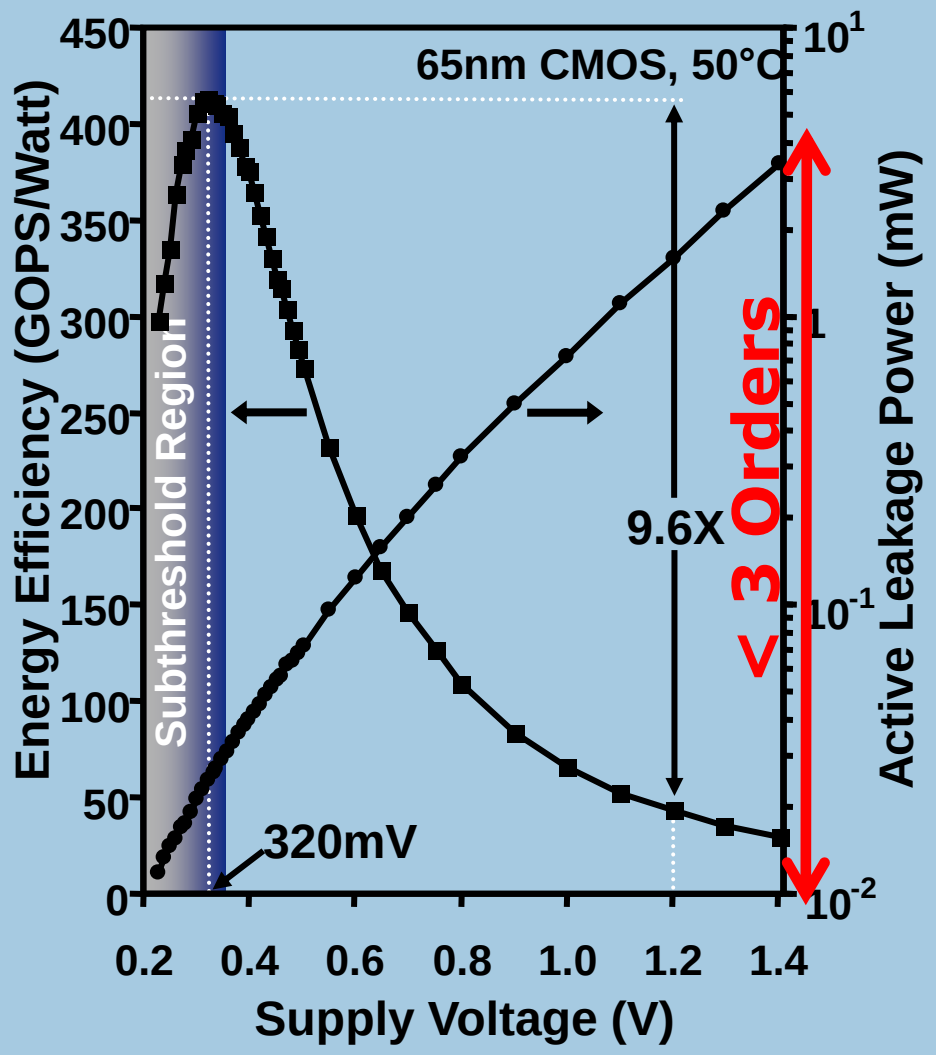
Voltage Scaling

When designed to voltage scale



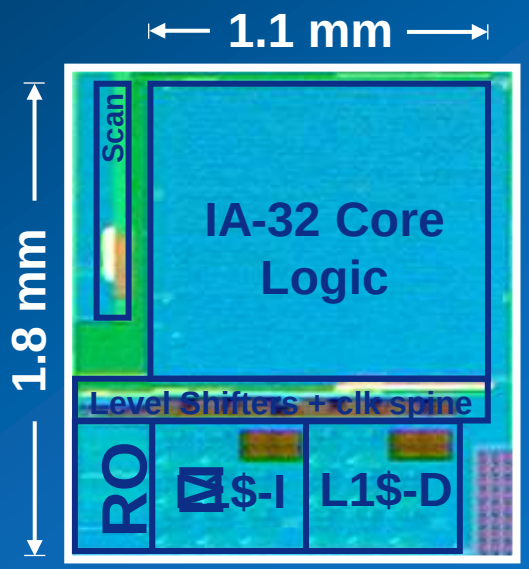
Near Threshold-Voltage (NTV)

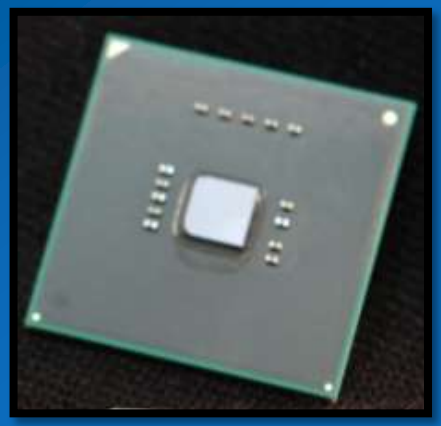




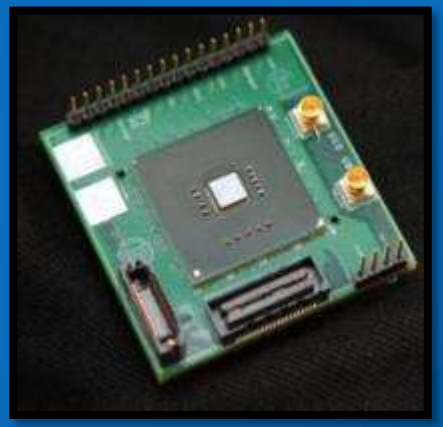
H. Kaul et al, 16.6: ISSCC08

Experimental NTV Processor





951 Pin FCBGA Package



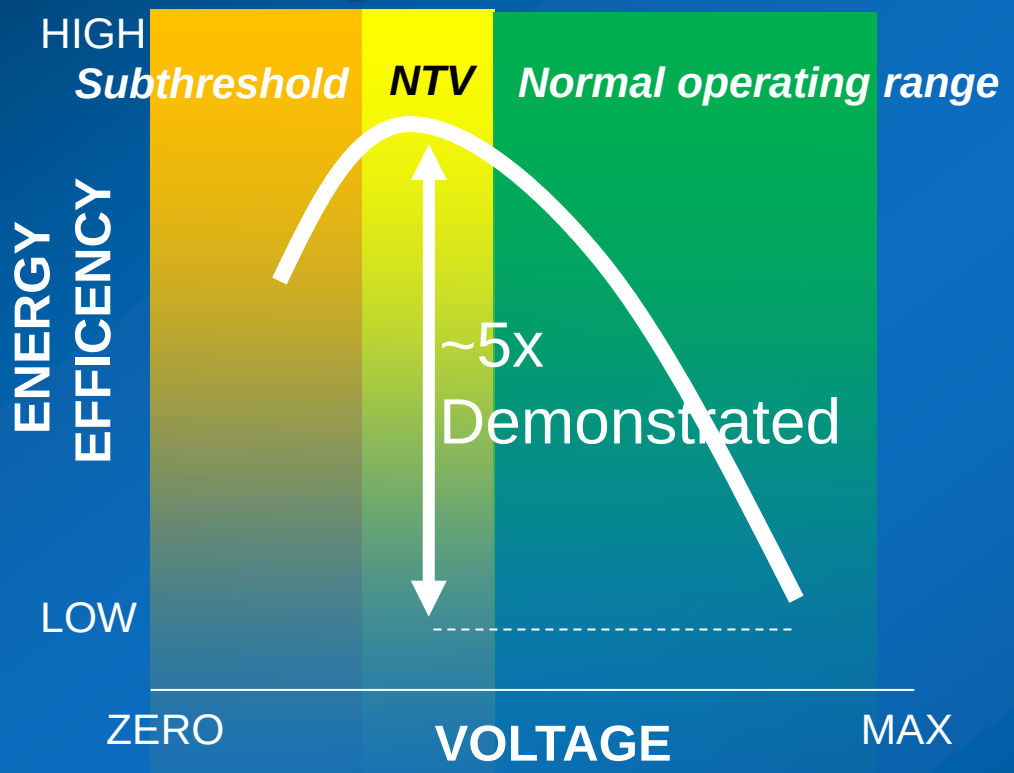
Custom Interposer

Technology	32nm High-K Metal Gate
Interconnect	1 Poly, 9 Metal (Cu)
Transistors	6 Million (Core)
Core Area	$2mm^2$

Legacy Socket-7 Motherboard

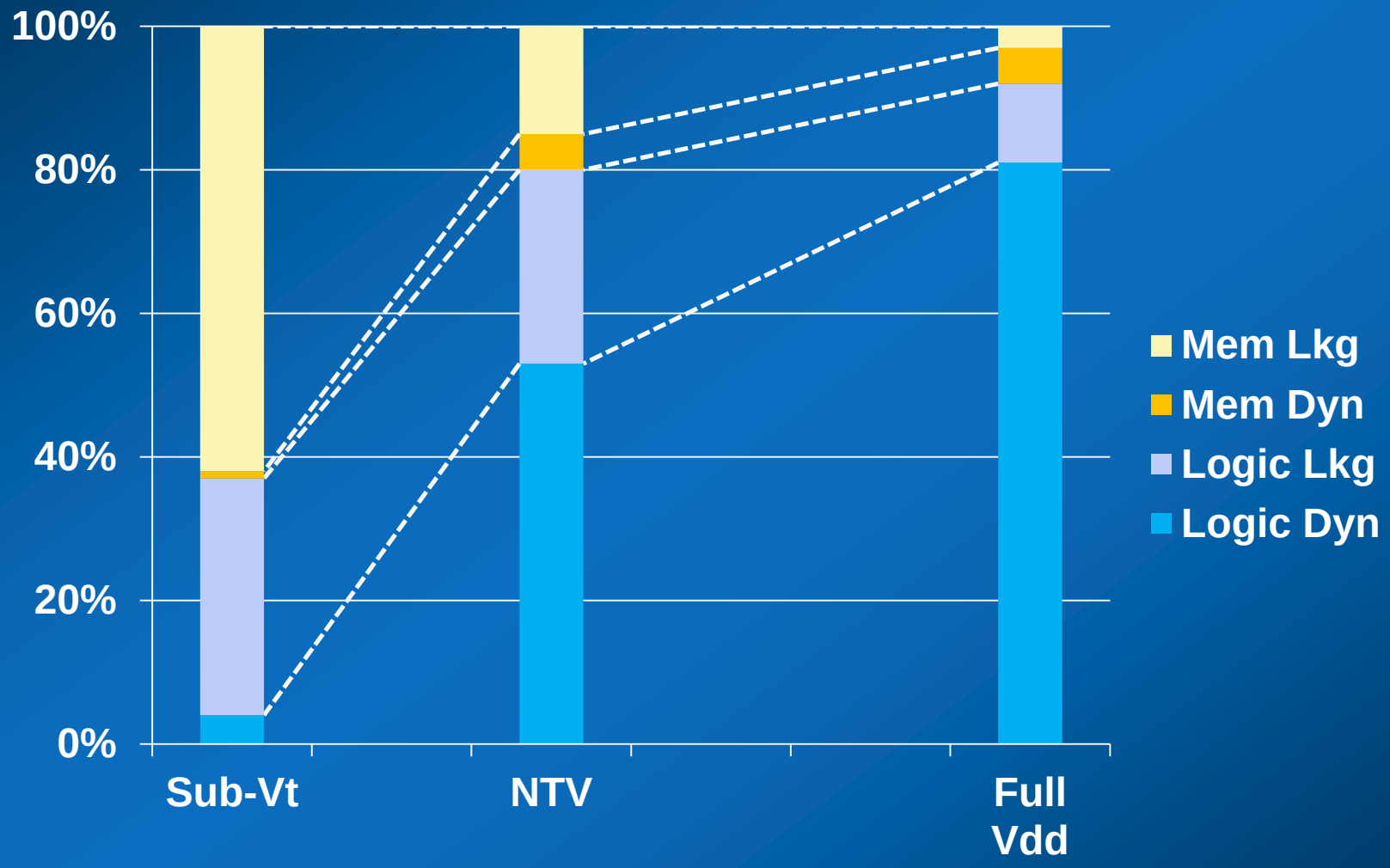
S. Jain, et al, "A 280mV-to-1.2V Wide-Operating-Range IA-32 Processor in 32nm CMOS", ISSCC 2012

Wide Dynamic Range



Ultra-low Power	Energy Efficient	High Performance
280 mV	0.45 V	1.2 V
3 MHz	60 MHz	915 MHz
2 mW	10 mW	737 mW
1500 Mips/W	5830 Mips/W	1240 Mips/W

Observations

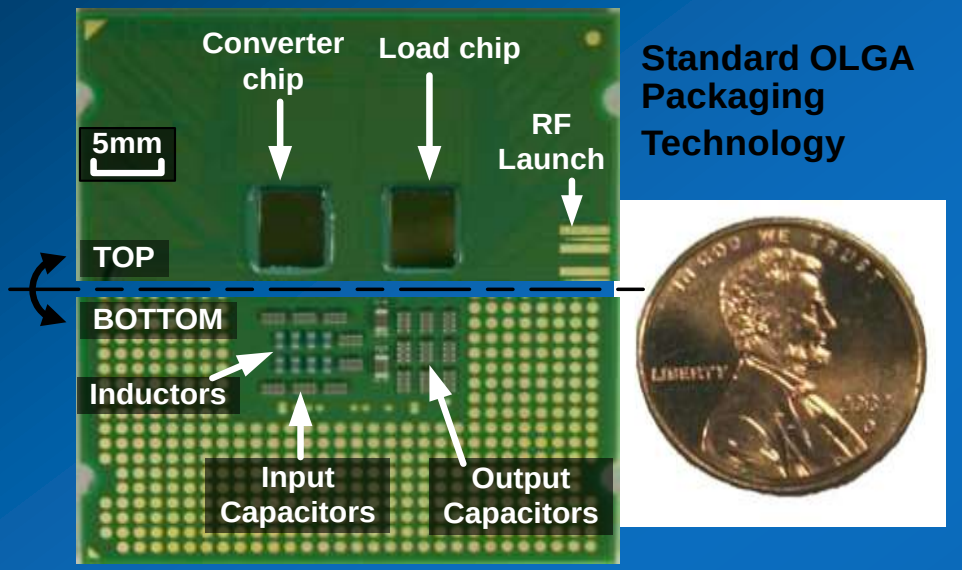


Leakage power dominates
Fine grain leakage power management is required

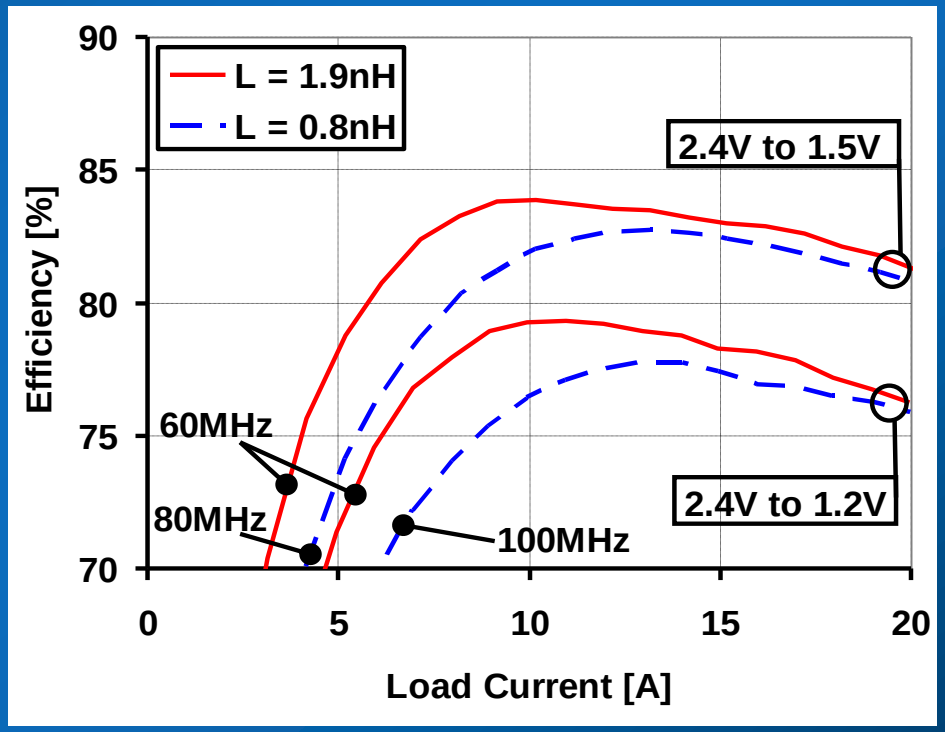
Integration of Power Delivery

For efficiency and management

Integrated Voltage Regulator Testchip



Standard OLGA Packaging Technology



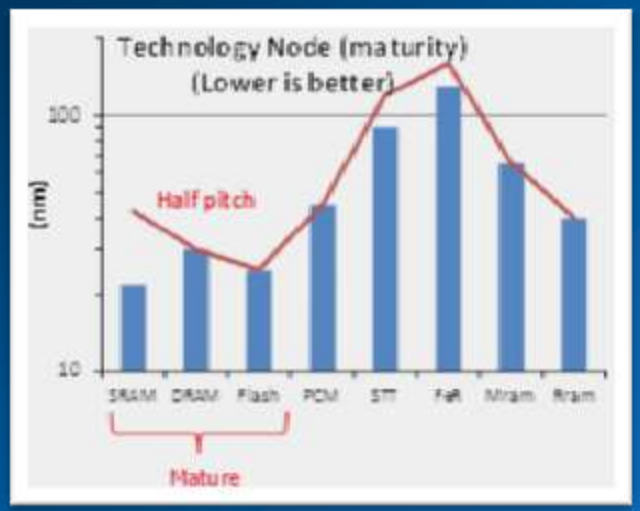
Power delivery closer to the load for

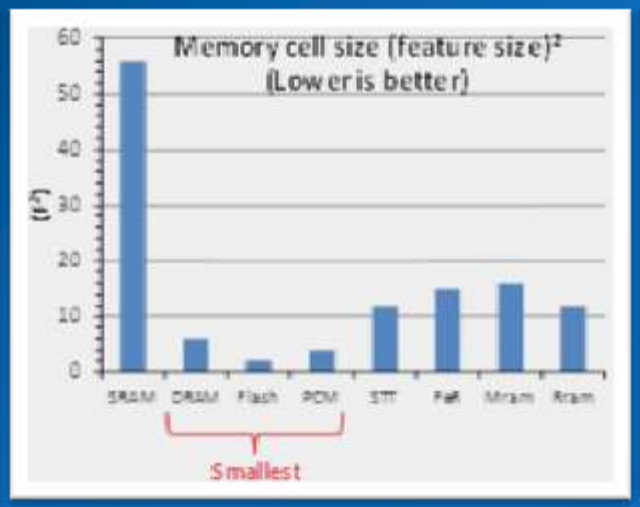
1. Improved efficiency

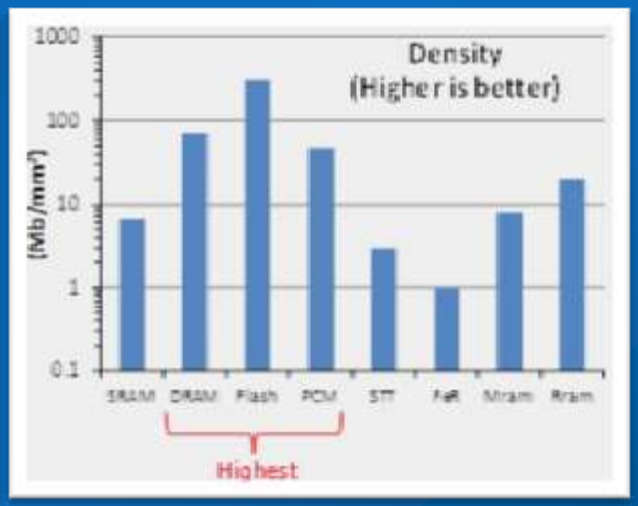
2. Fine grain power management

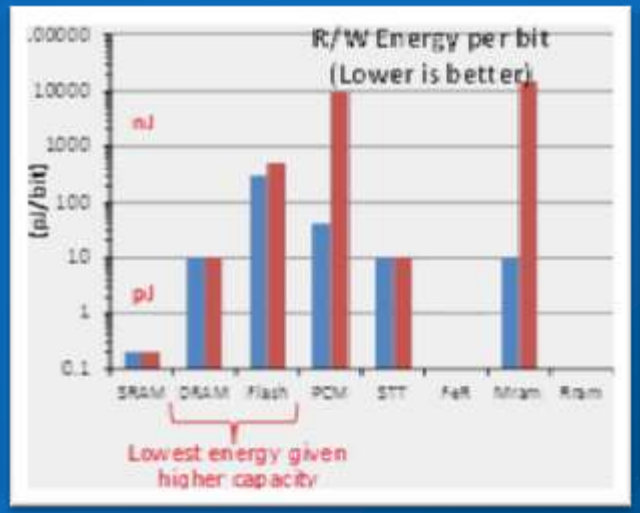
Schrom et al, "A 100MHz 8-Phase Buck Converter Delivering 12A in 25mm2 Using Air-Core Inductors", APEC 2007

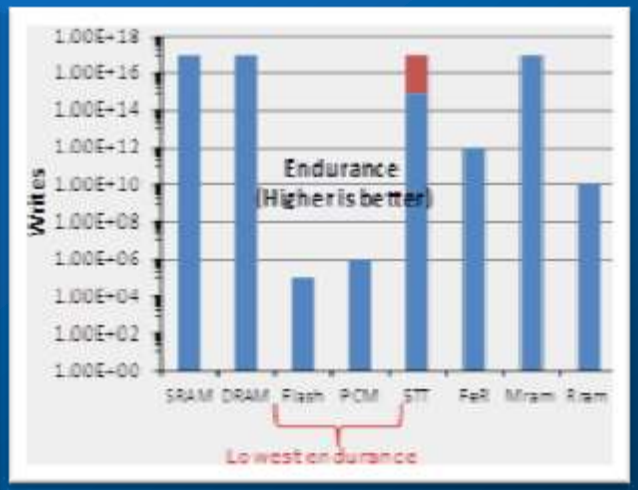
Compare Memory Technologies









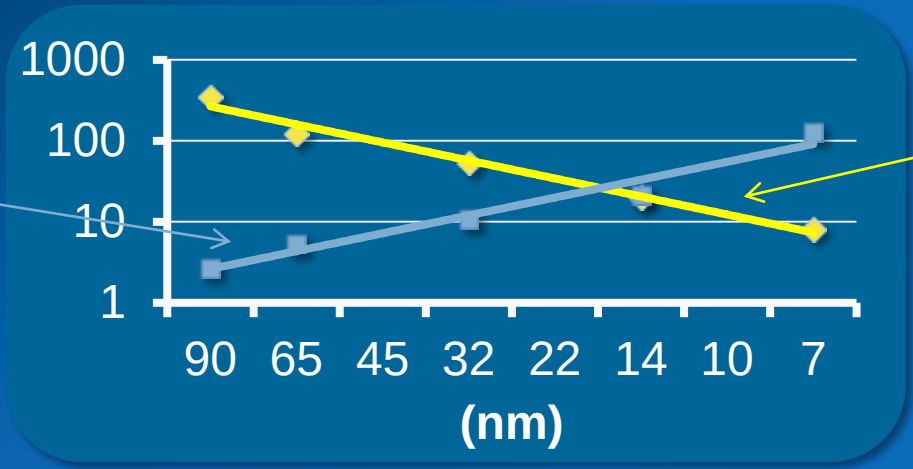


Source: Intel

DRAM for first level capacity memory
NAND/PCM for next level storage

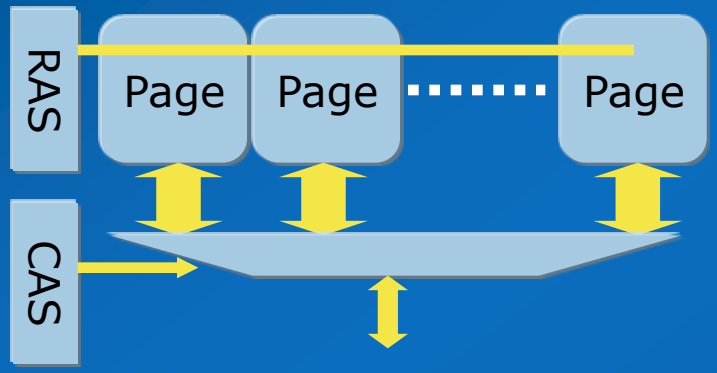
Revise DRAM Architecture

❶ Need exponentially increasing BW (GB/sec)



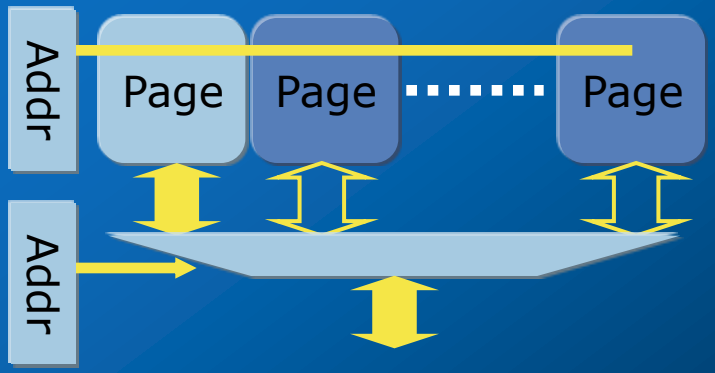
❷ Need exponentially decreasing energy (pJ/bit)

Traditional DRAM



Activates many pages
Lots of reads and writes (refresh)
Small amount of read data is used
Requires small number of pins

New DRAM architecture



Activates few pages
Read and write (refresh) what is needed
All read data is used
Requires large number of IO's (3D)

3D-Integration of DRAM and Logic

Logic Buffer Chip

Technology optimized for:

High speed signaling

Energy efficient logic circuits

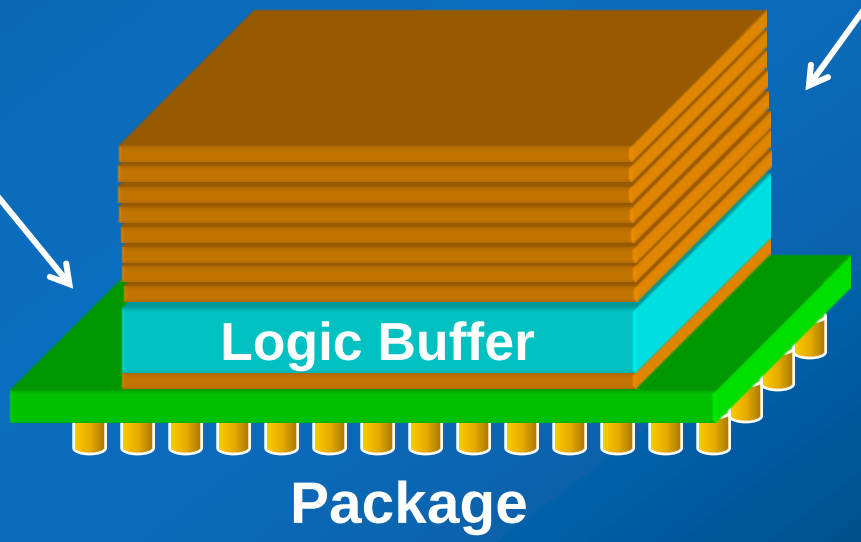
Implement intelligence

DRAM Stack

Technology optimized for:

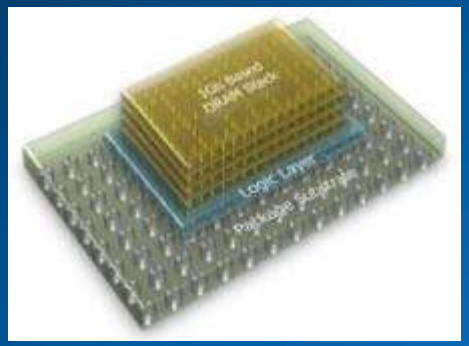
Memory density

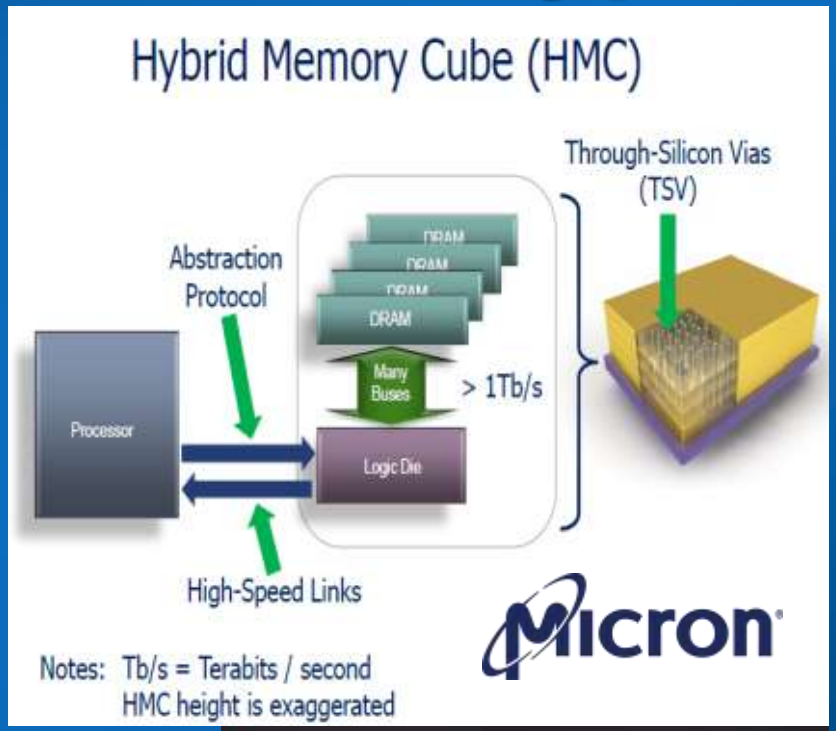
Lower cost



3D Integration provides best of both worlds

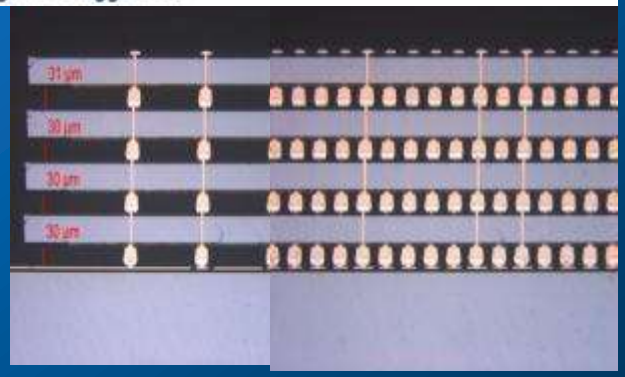
1Tb/s HMC DRAM Prototype





- 3D integration technology
- 1Gb DRAM Array
- 512 MB total DRAM/cube
- 128GB/s Bandwidth
- <10 pj/bit energy

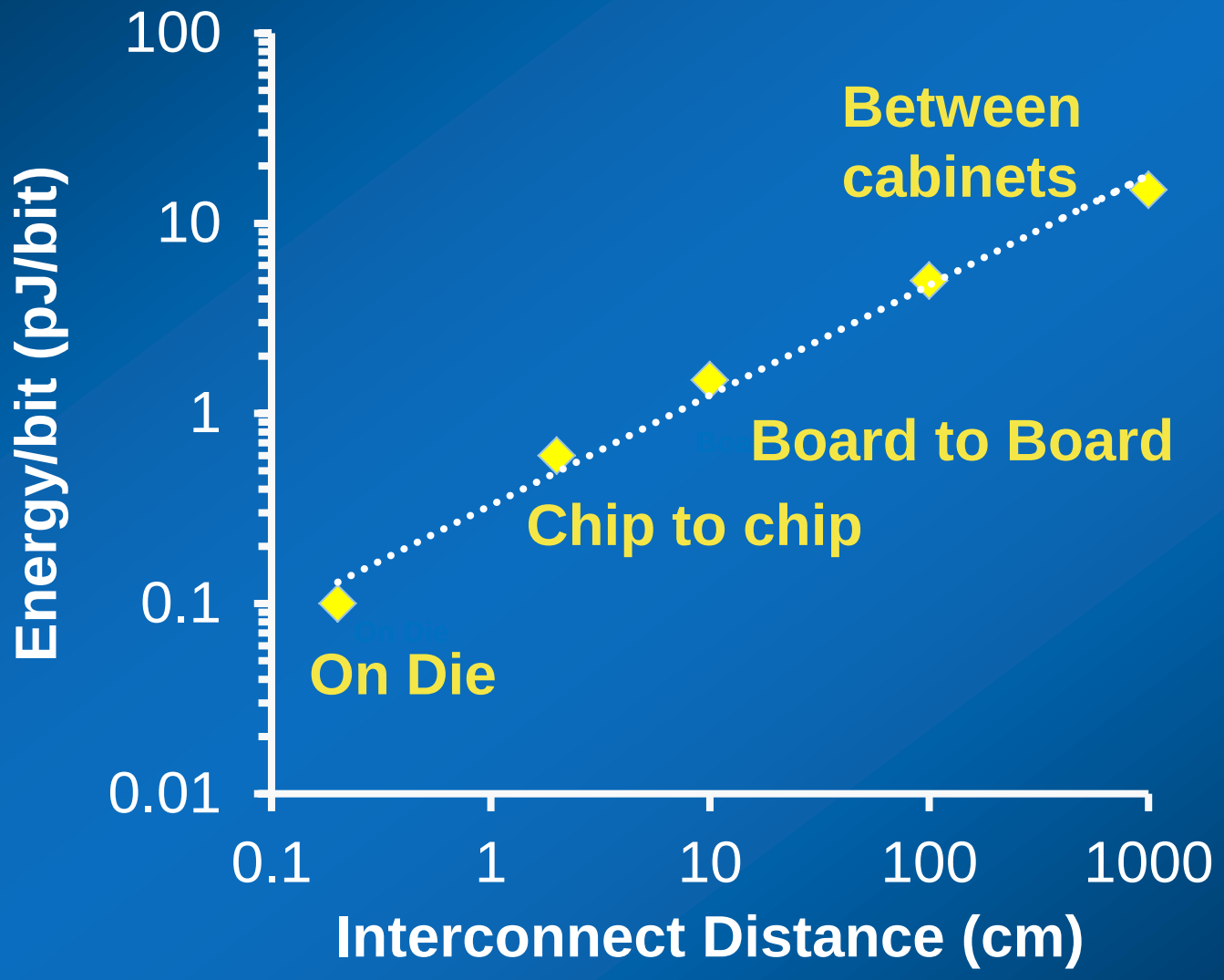
	Bandwidth	Energy Efficiency
DDR-3 (Today)	10.66 GB/Sec	50-75 pJ/bit
Hybrid Memory Cube	128 GB/Sec	**8 pJ/bit**



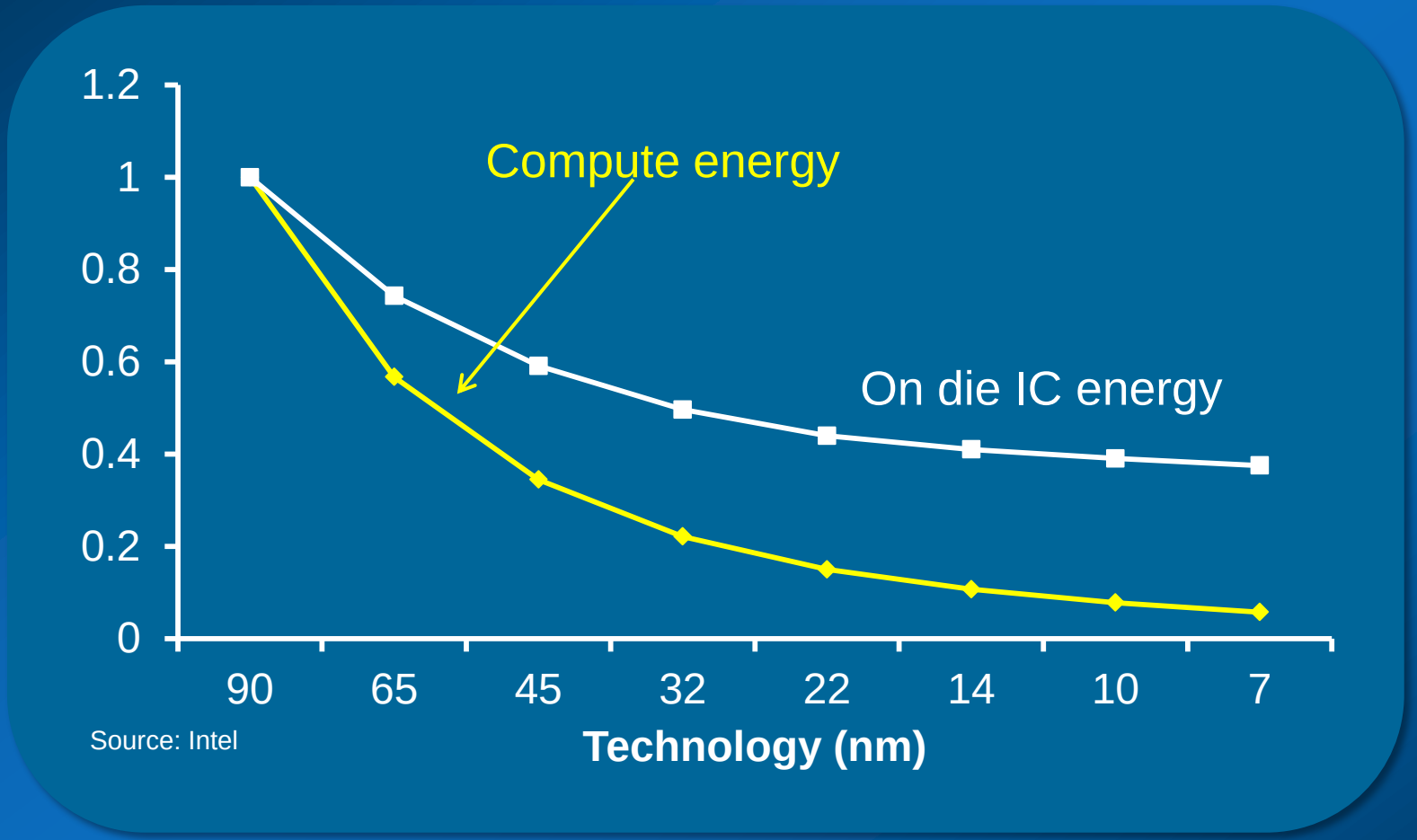
10X higher bandwidth, 10X lower energy

Source: Micron

Communication Energy



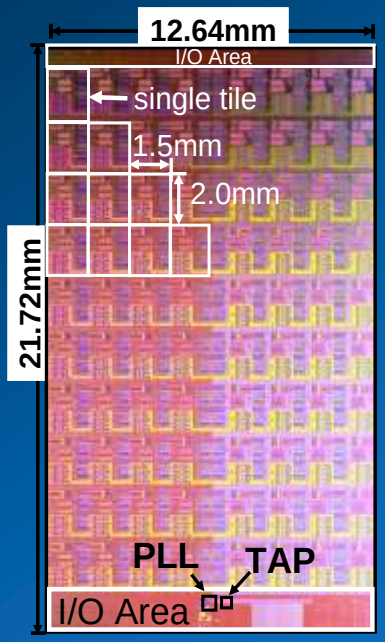
On-die Interconnect

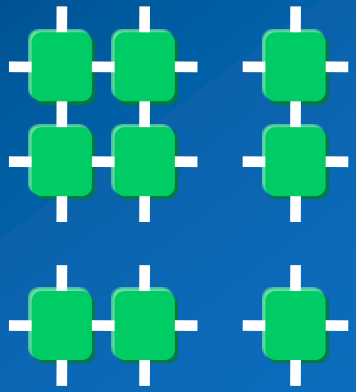


Interconnect energy (per mm) reduces slower than compute
On-die data movement energy will start to dominate

Network On Chip (NoC)

80 Core TFLOP Chip (2006)

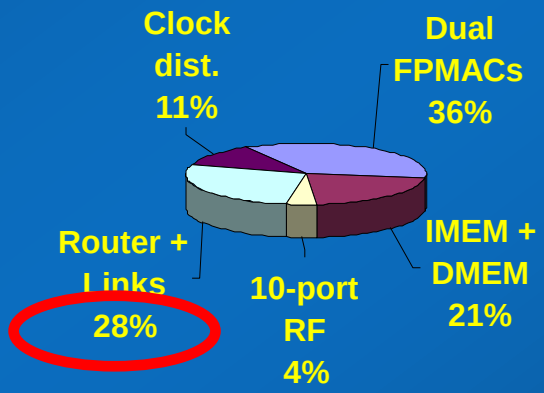




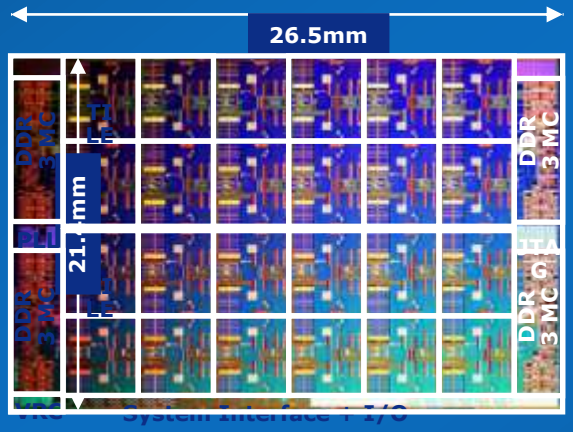
8 X 10 Mesh

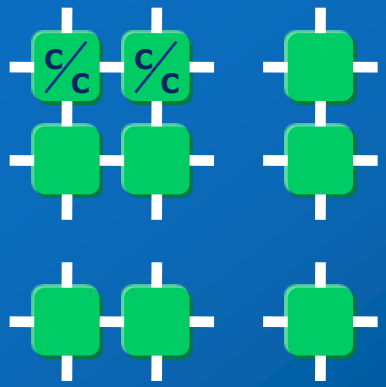
32 bit links

320 GB/sec bisection BW @ 5 GHz



48 Core Single Chip Cloud (2009)



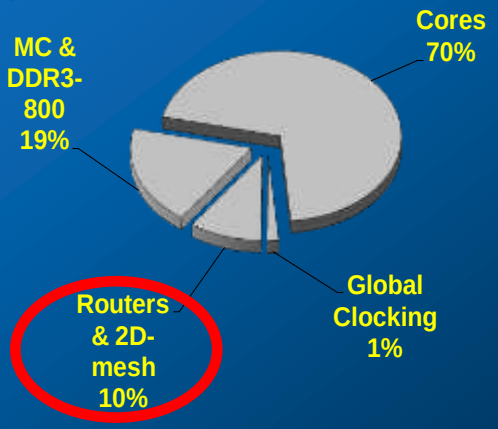


2 Core clusters in 6 X 4 Mesh

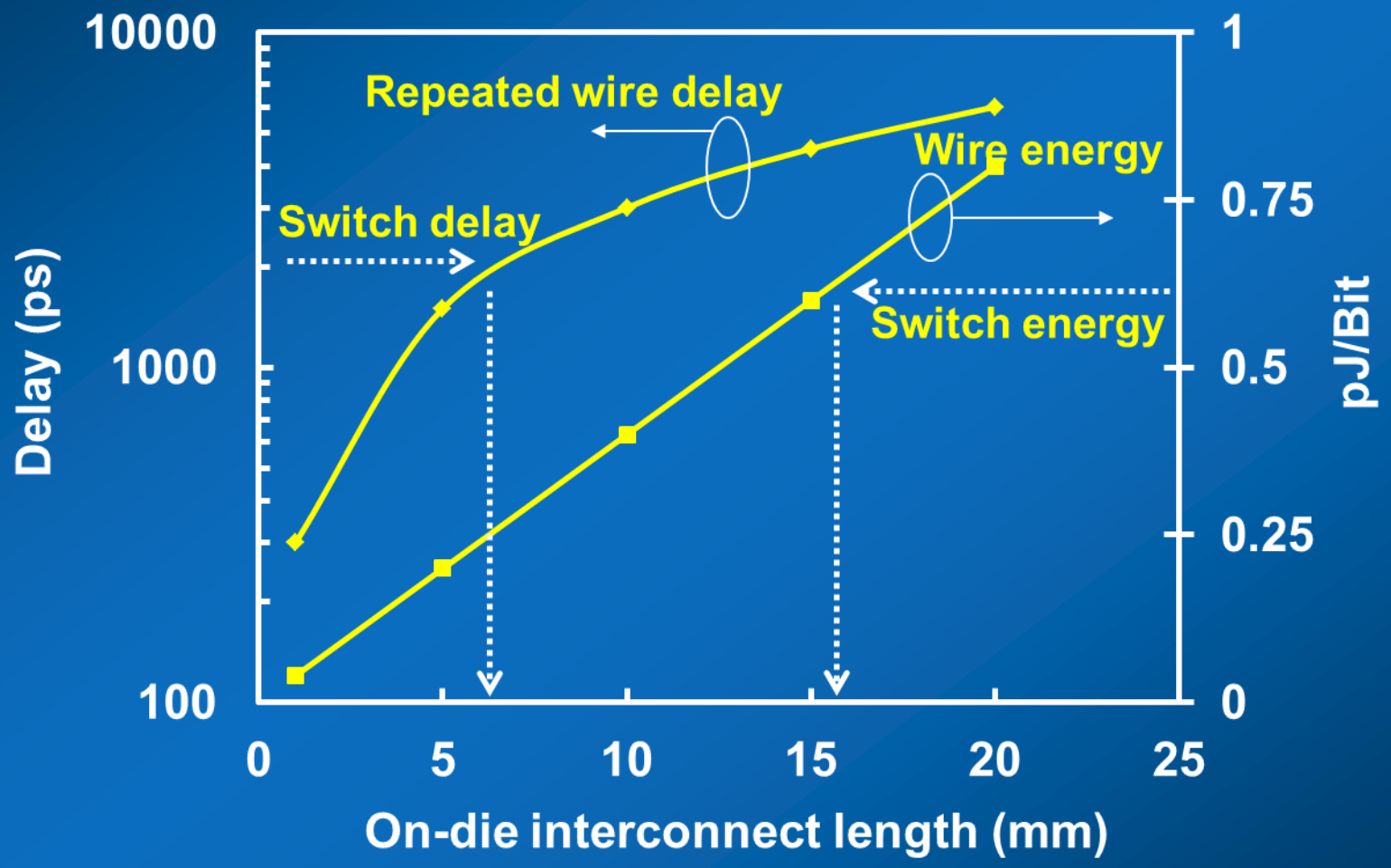
(why not 6 x 8?)

128 bit links

256 GB/sec bisection BW @ 2 GHz

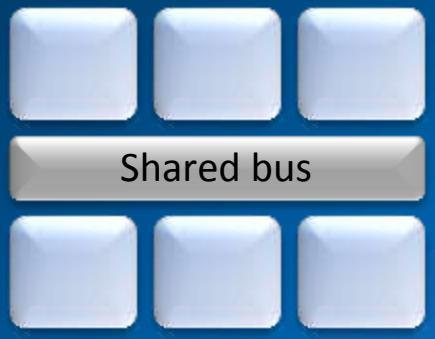


On-chip Interconnect Analysis



Interconnect Structures

Buses over short distance

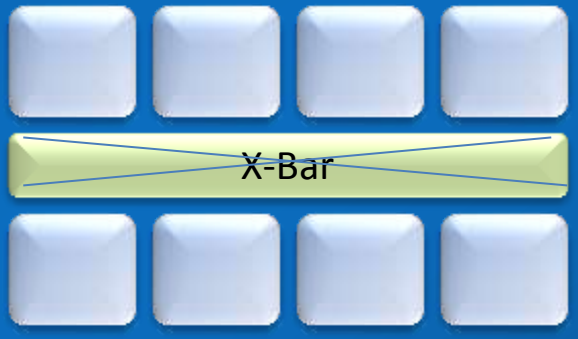


1 to 10 fJ/bit
0 to 5mm
Limited scalability

Shared memory

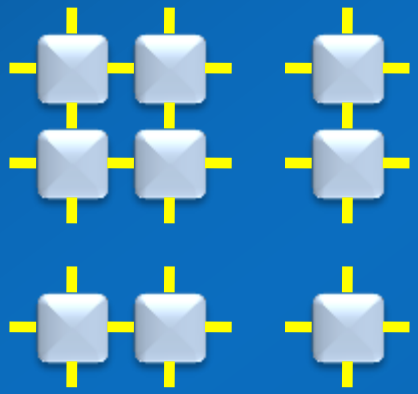
10 to 100 fJ/bit
1 to 5mm
Limited scalability

Cross Bar Switch



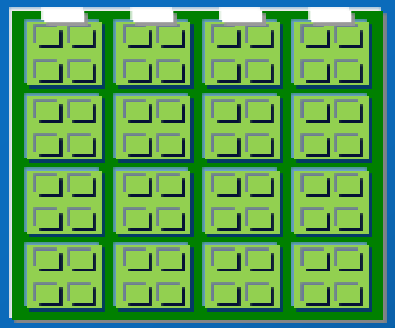
0.1 to 1pJ/bit
2 to 10mm
Moderate scalability

Packet Switched Network

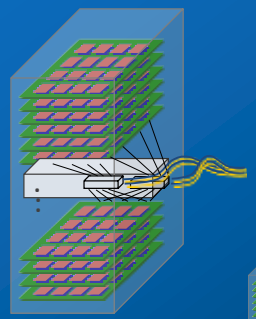


1 to 3pJ/bit
>5 mm, scalable

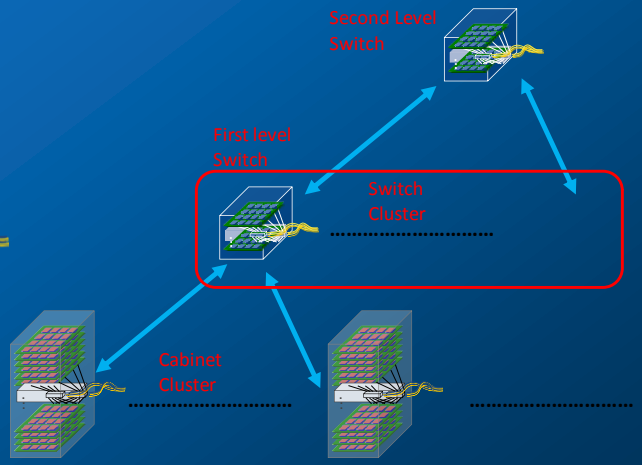
Board



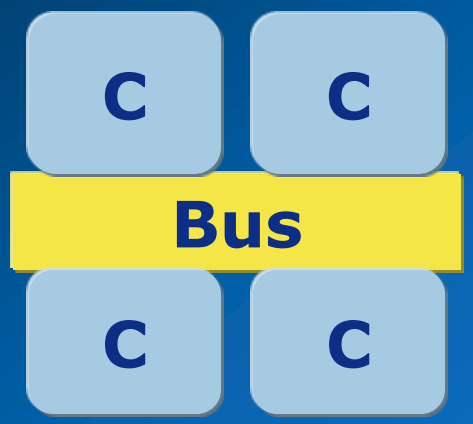
Cabinet



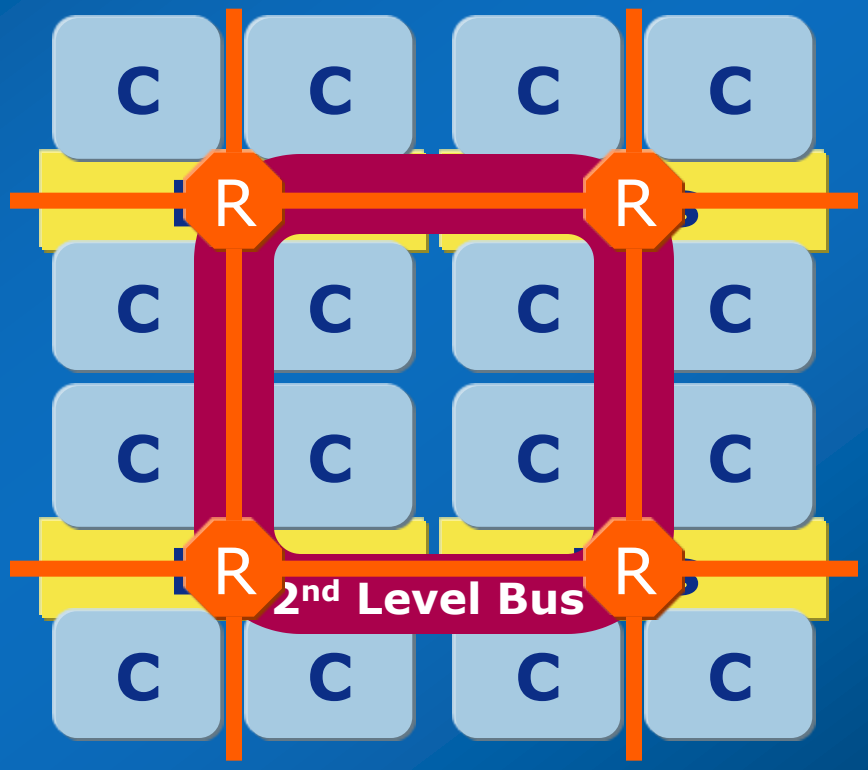
System



Hierarchical & Heterogeneous



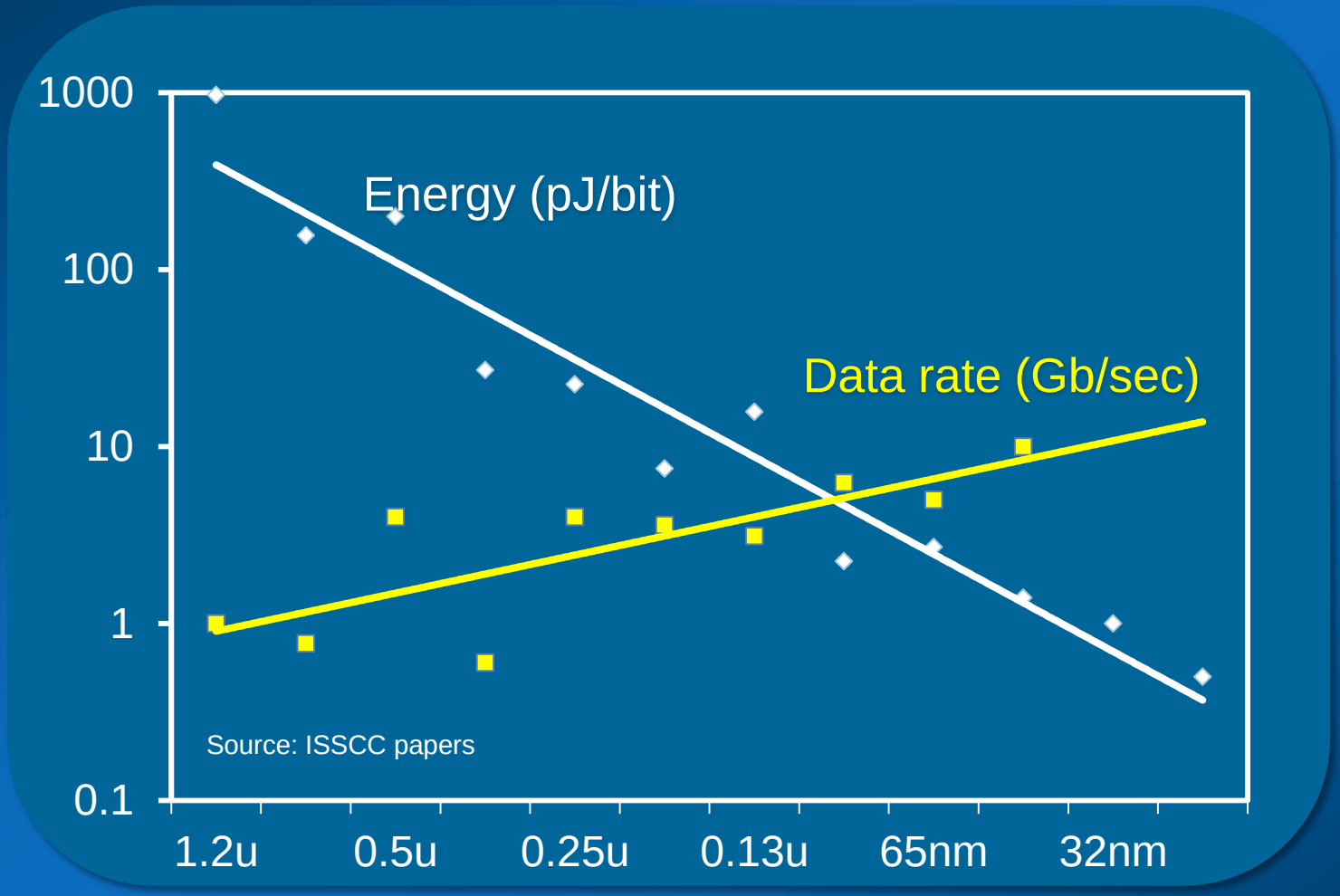
Bus to connect over short distances



Hierarchy of Busses

Or hierarchical circuit and packet switched networks

Electrical Interconnect < 1 Meter



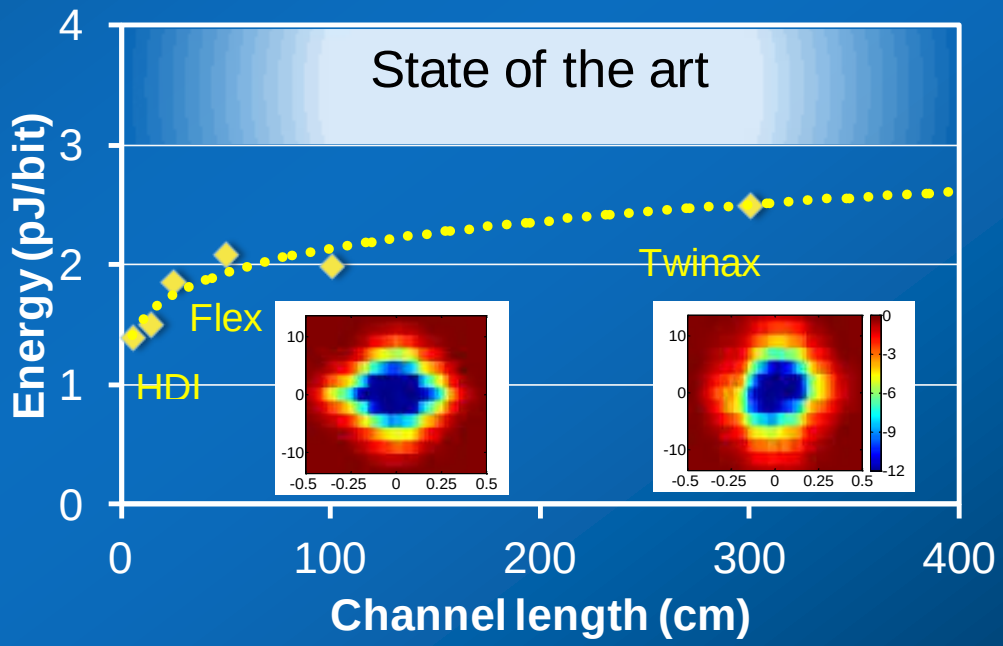
BW and Energy efficiency improves, but not enough

Electrical Interconnect Advances

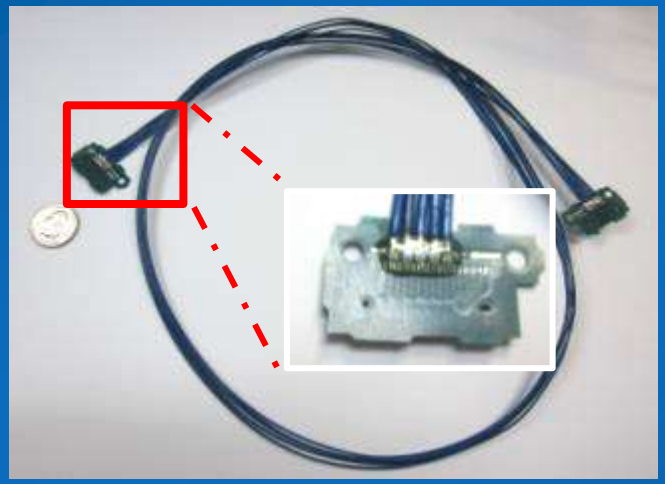
Employ, new, low-loss, non-traditional interconnects

Top of the package connector

Low-loss flex connector



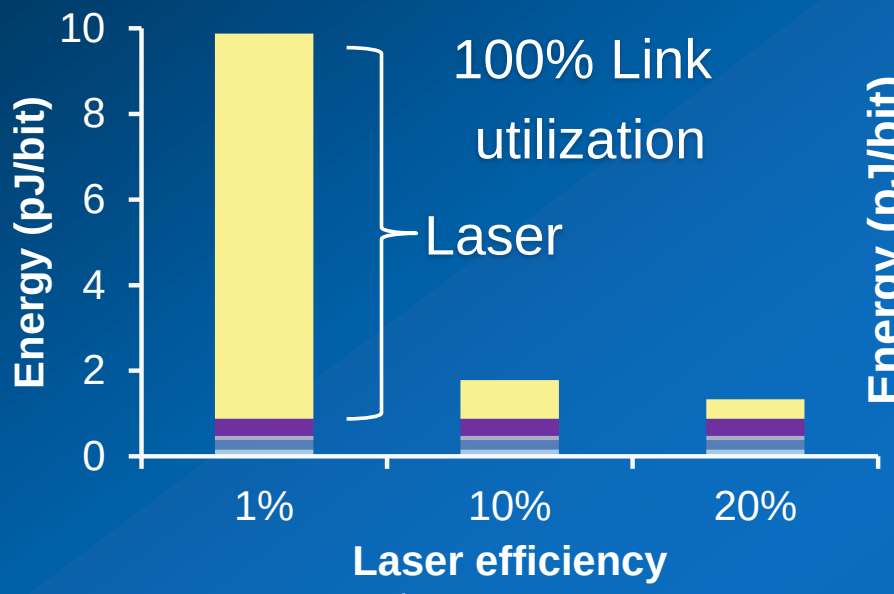
Low-loss twinax



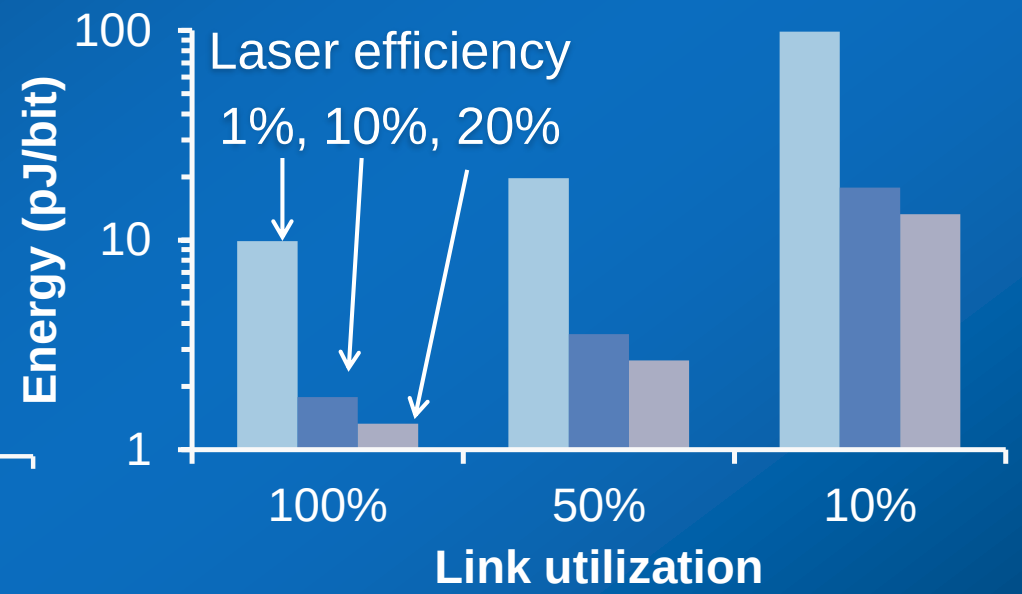
Co-optimization of interconnects and circuits for energy efficiency

O'Mahony et al, "A 47x10Gb/s 1.4mW/(Gb/s) Parallel Interface in 45nm CMOS ", ISSCC 2010; and J. Jaussi, RESS004, IDF 2010

Optical Interconnect > 1 Meter



Source: PETE Study group



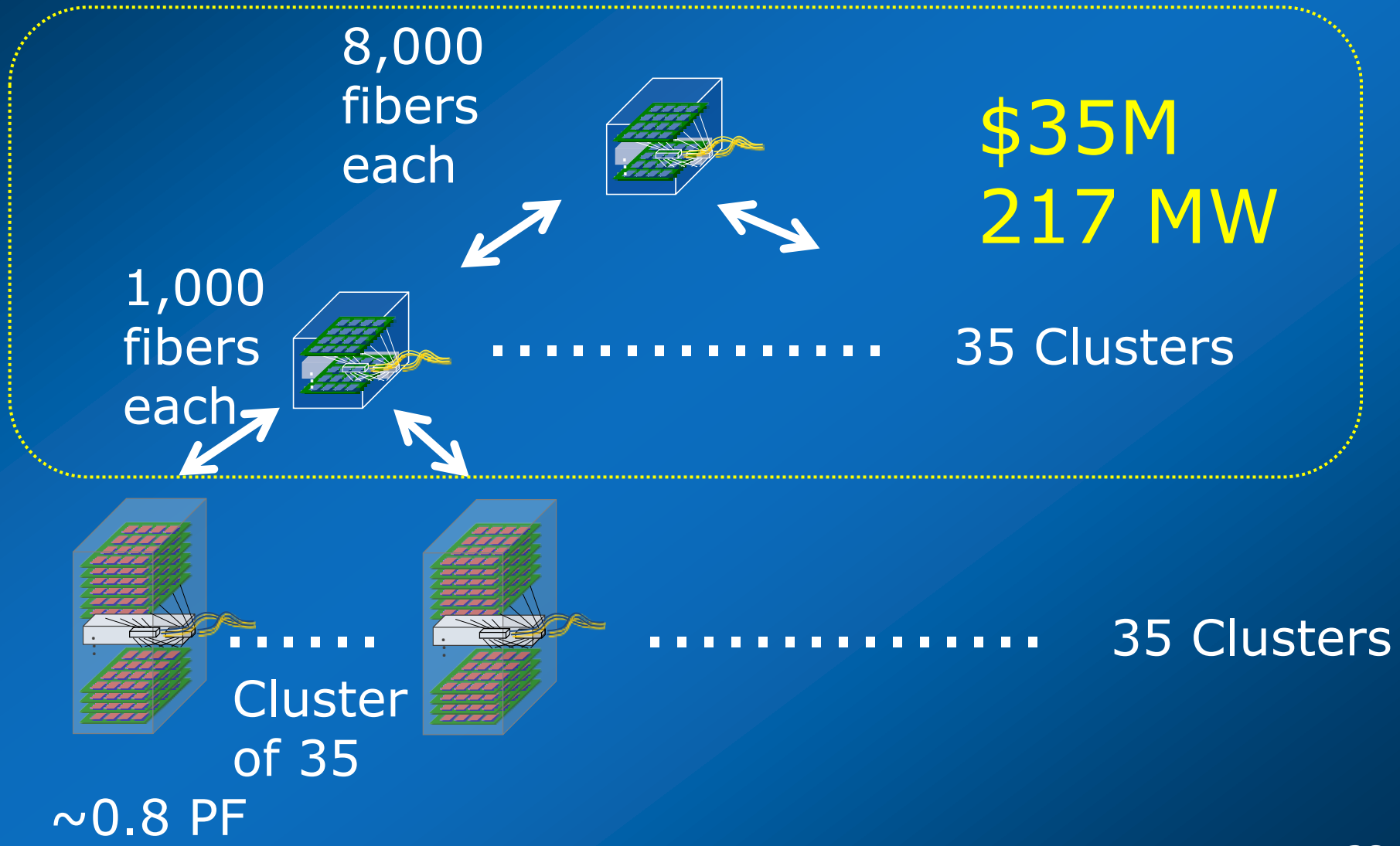
Energy in supporting electronics is very low

Link energy dominated by laser (efficiency)

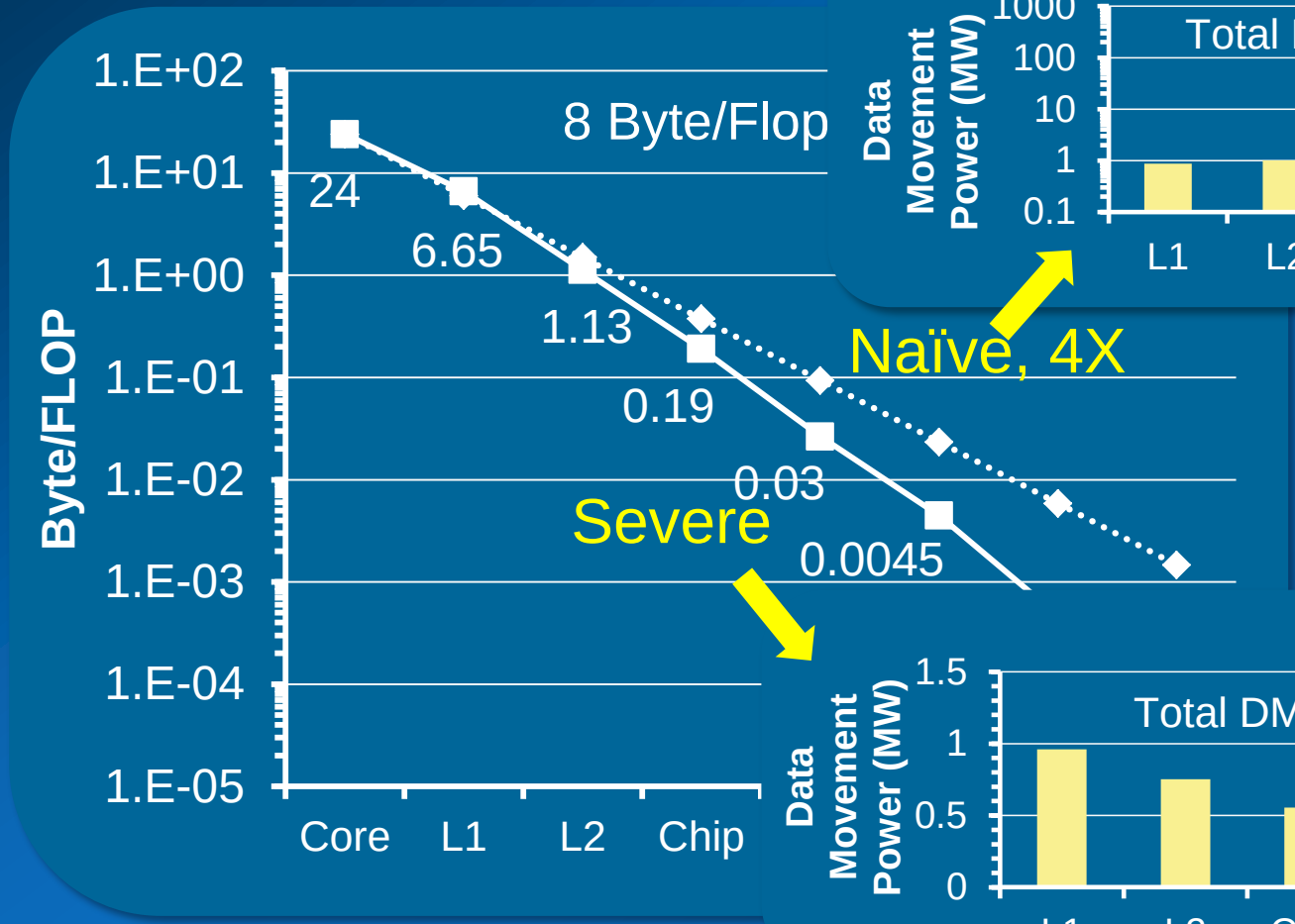
Sustained, high link utilization required

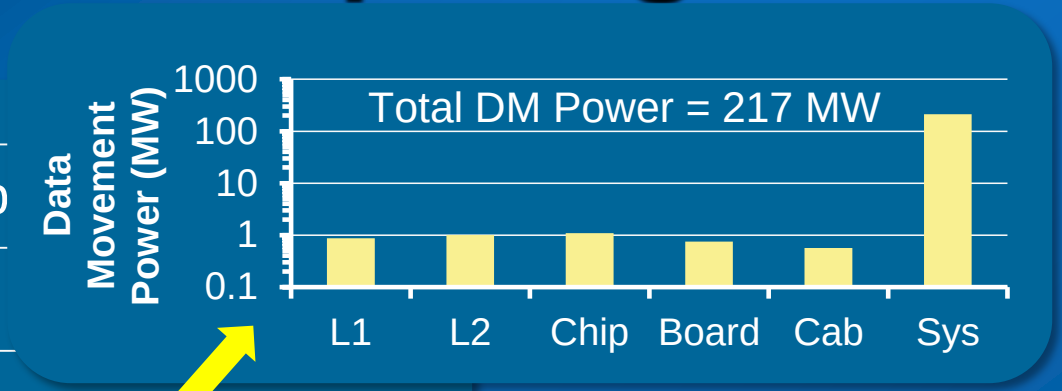
Straw-man Exa— Interconnect

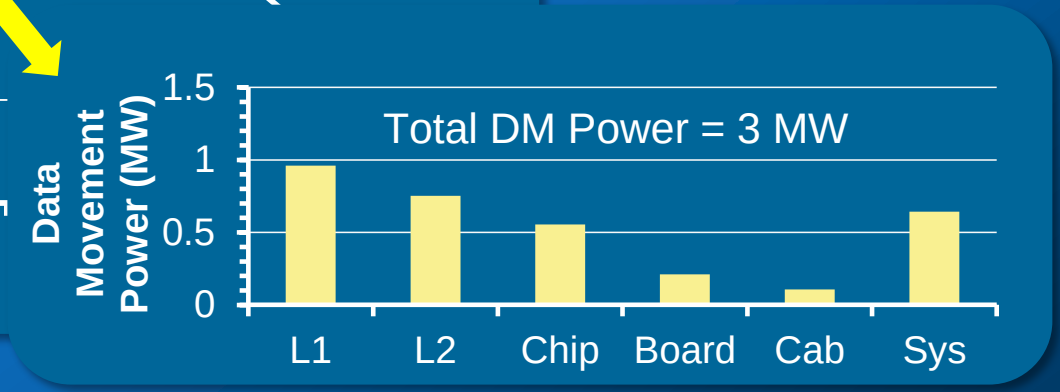
Assume: 40 Gbps, 10 pJ/b, $0.6/Gbps, 8B/FLOP, naïve tapering



Bandwidth Tapering



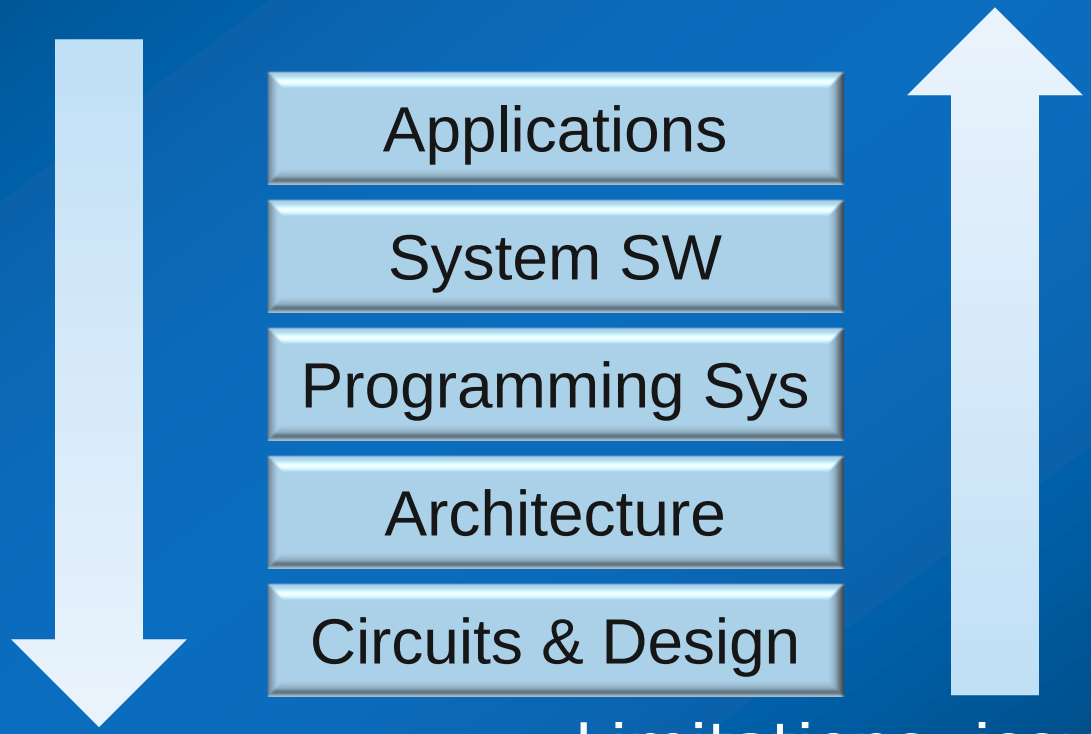




Intelligent BW tapering is necessary

HW-SW Co-design

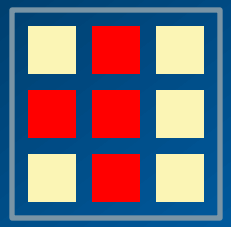
Applications and SW stack provide guidance for efficient system design



Limitations, issues and opportunities to exploit

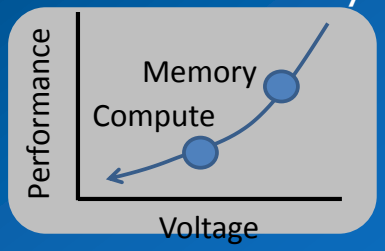
Bottom-up Guidance

1. NTV reduces energy but exacerbates variations



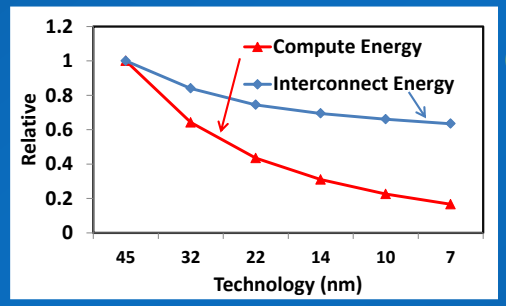
Small & Fast cores
Random distribution
Temp dependent

2. Limited NTV for arrays (memory) due to stability issues



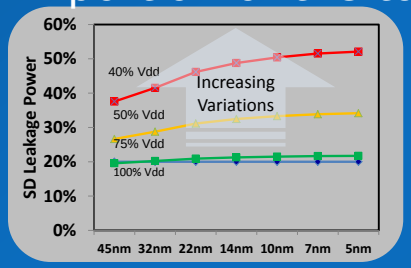
Disproportionate
Memory arrays
can be made
larger

3. On-die Interconnect energy (per mm) does not reduce as much as compute



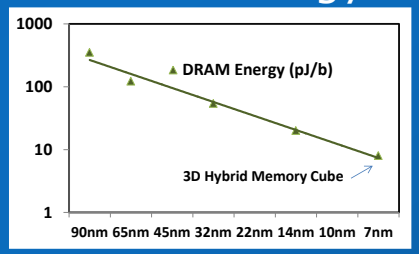
6X compute
1.6X interconnect

4. At NTV, leakage power is substantial portion of the total power



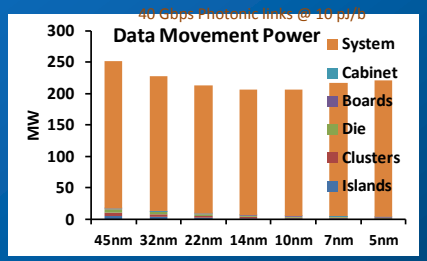
Expect 50%
leakage
Idle hardware
consumes energy

5. DRAM energy scales, but not enough



50 pJ/b today
8 pJ/b
demonstrated
Need < 2pJ/b

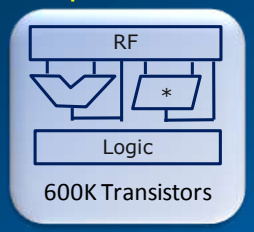
6. System interconnect limited by laser energy and cost



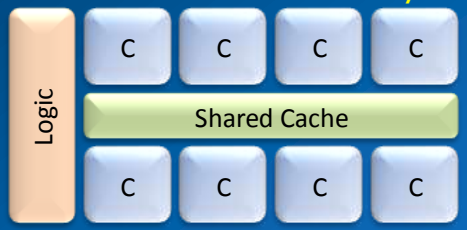
BW tapering and
locality awareness
necessary

Straw-man Architecture at NTV

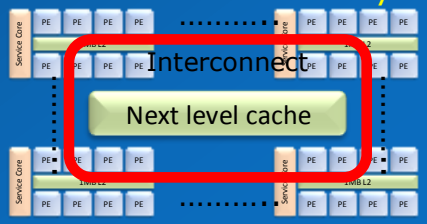
Simplest Core



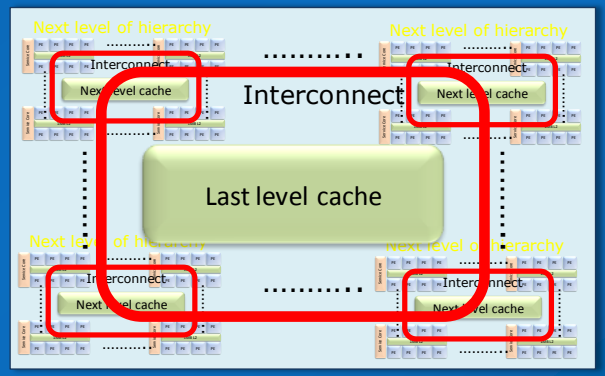
First level of hierarchy



Next level of hierarchy



Processor



	Full Vdd	50% Vdd
Technology	7nm, 2018	
Die area	500 mm2	
Cores	2048	
Frequency	4.2 GHz	600 MHz
TFLOPs	17.2	2.5
Power	600 Watts	37 Watts
E Efficiency	34 pJ/Flop	15 pJ/Flop

Reduced frequency and flops

Reduced power and improved E-efficiency

Compute energy efficiency close to Exascale goal

SW Challenges

Execution model

Programming model

1. Extreme parallelism (1000X due to Exa, additional 4X due to NTV)
2. Data locality—reduce data movement
3. Intelligent scheduling—move thread to data if necessary
4. Fine grain resource management (objective function)
5. Applications and algorithms incorporate paradigm change

Programming & Execution Model

Event driven tasks (EDT)

Dataflow inspired, tiny codelets (self contained)

Non blocking, no preemption

Programming model:

Separation of concerns: Domain specification & HW mapping

Express data locality with hierarchical tiling

Global, shared, non-coherent address space

Optimization and auto generation of EDTs (HW specific)

Execution model:

Dynamic, event-driven scheduling, non-blocking

Dynamic decision to move computation to data

Observation based adaption (self-awareness)

Implemented in the runtime environment

Separation of concerns:

User application, control, and resource management

Over-provisioned Introspectively Resource Managed System

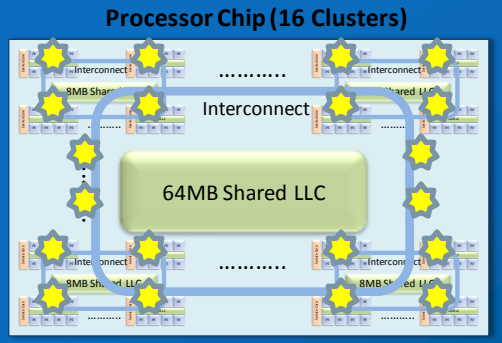
Addressing variations

1. Provide more compute HW
2. Law of large numbers
3. Static profile

Fine grain resource mgmt

Dynamic reconfiguration:
1. Energy efficiency
2. Latency
3. Dynamic resource management

Sensors for introspection



1. Energy consumption
2. Instantaneous power
3. Computations
4. Data movement

1. Schedule threads based on objectives and resources

2. Dynamically control and manage resources

3. Identify sensors, functions in HW for implementation

System SW implements introspective execution model

Summary

Power & energy challenge continues

Opportunistically employ NTV operation

3D integration for DRAM

Communication energy will far exceed computation

Data locality will be paramount

Revolutionary software stack needed

Take HW/SW co-design beyond just a buzz word!