Exascale Computing—a fact or a fiction?

IPDPS 2013

Shekhar Borkar
Intel Corp.
May 21, 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

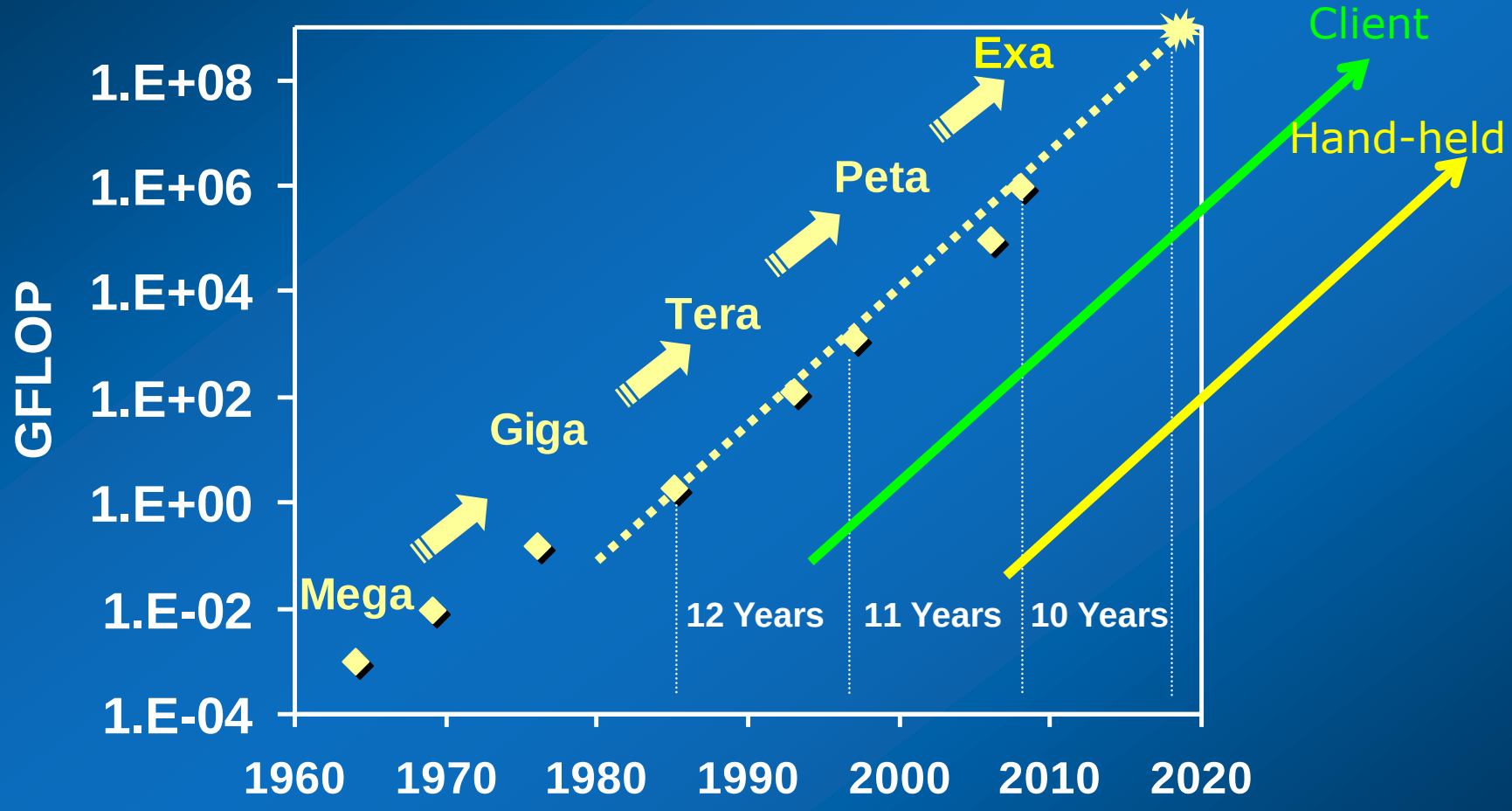
Compute roadmap & technology outlook

Challenges & solutions for:

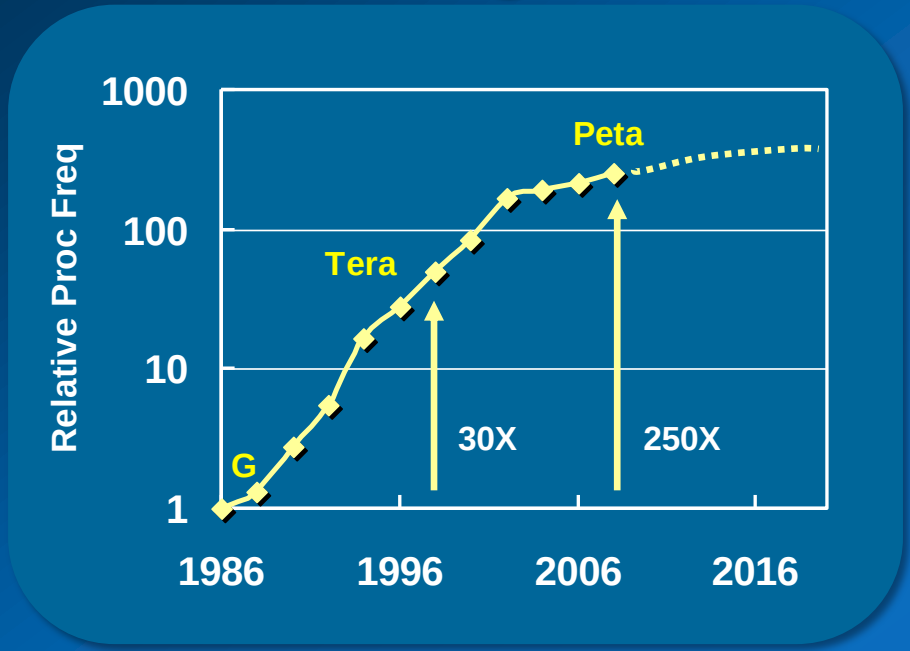
- Compute,
- Memory,
- Interconnect,
- Resiliency, and
- Software stack

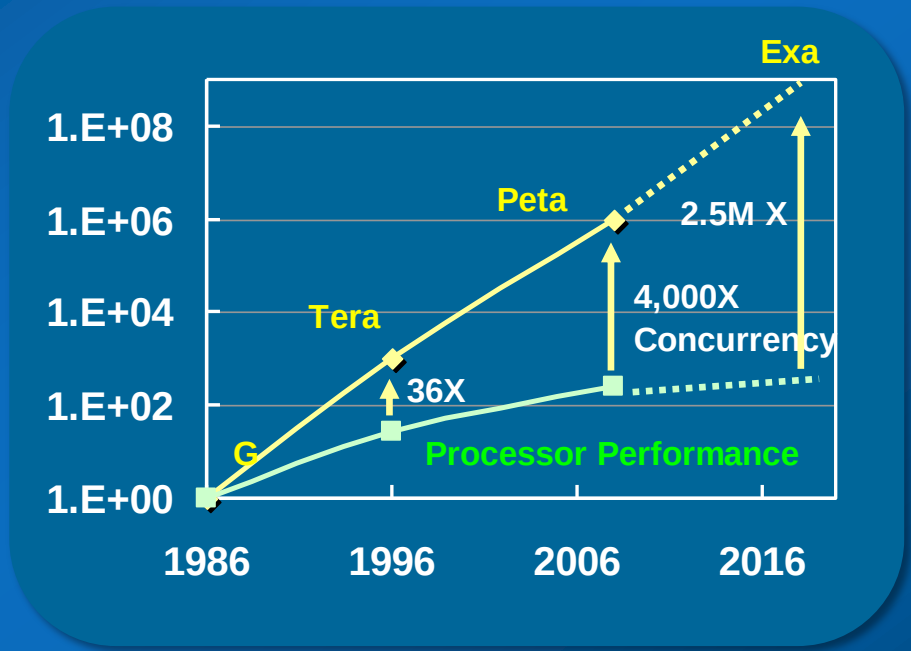
Summary

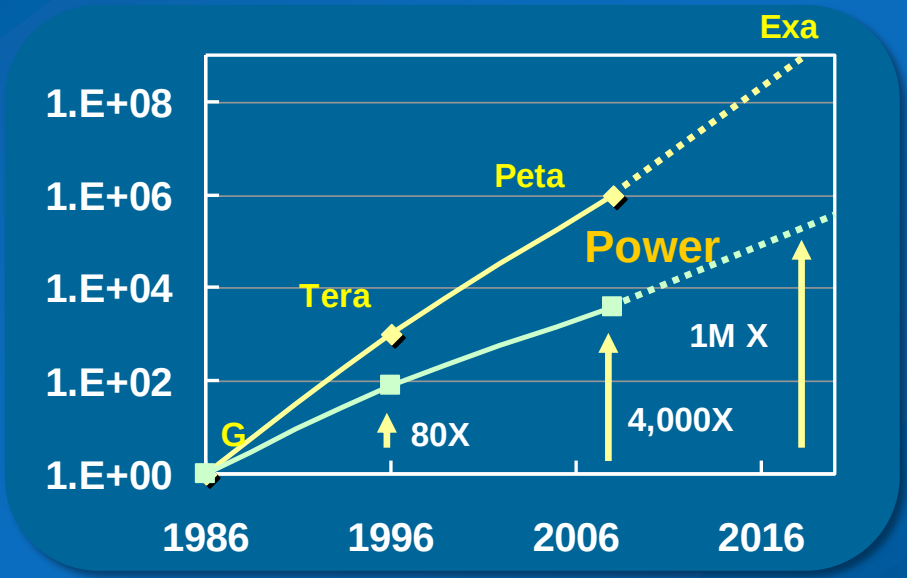
Compute Performance Roadmap



From Giga to Exa, via Tera & Peta







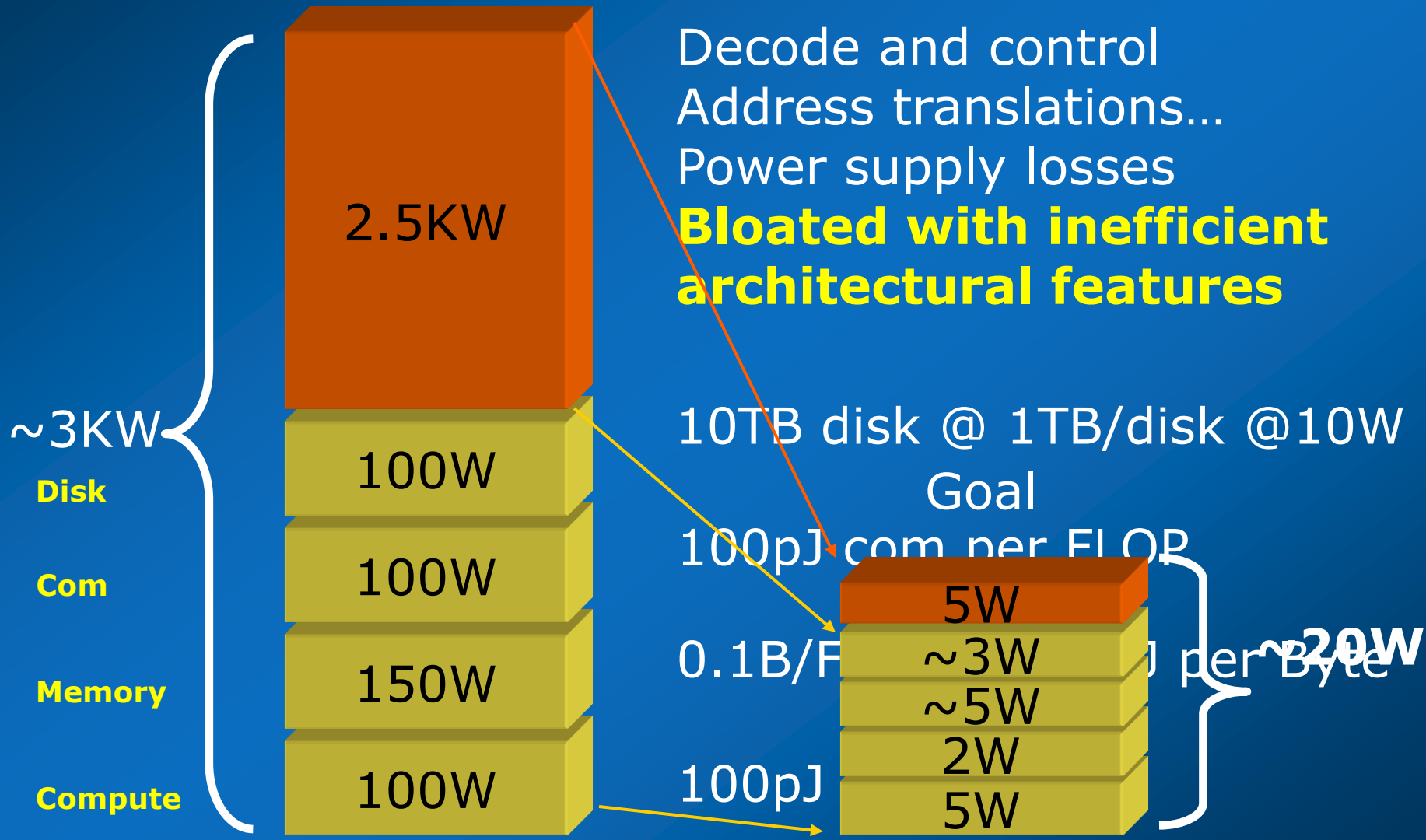
System performance increases faster

Parallelism continues to increase

Power & energy challenge continues

Where is the Energy Consumed?

Teraflop system today



The UHPC* Challenge

**DARPA, Ubiquitous HPC Program*

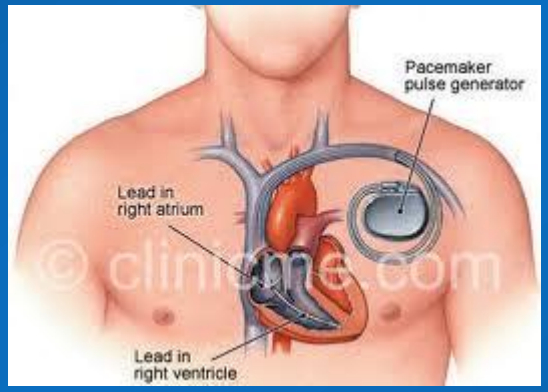
20MW, Exa

20KW, Peta

20W, Tera

20 pJ/Operation

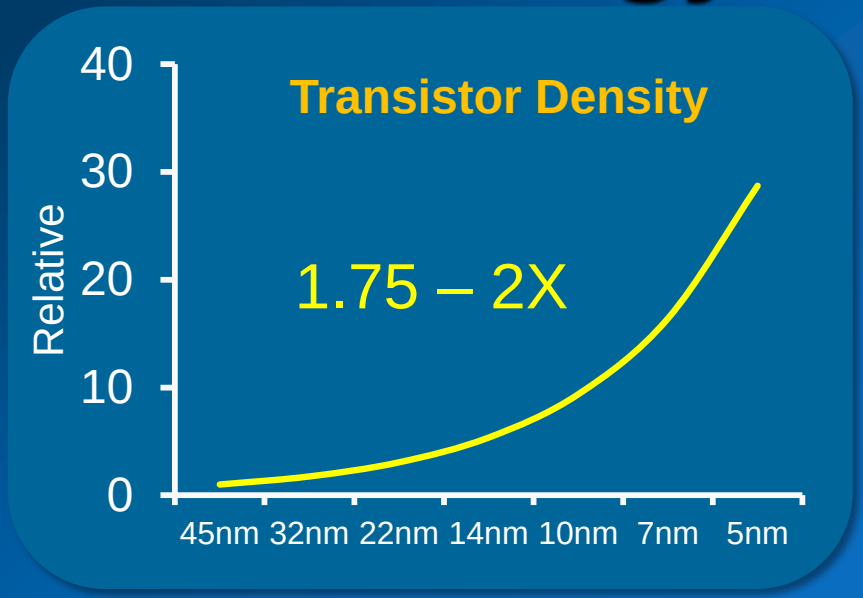
20 μW, Mega

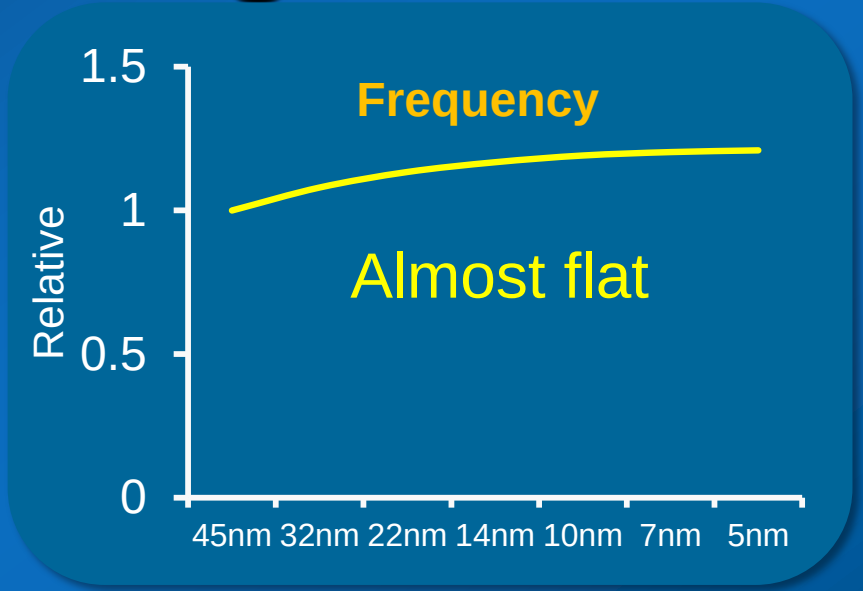


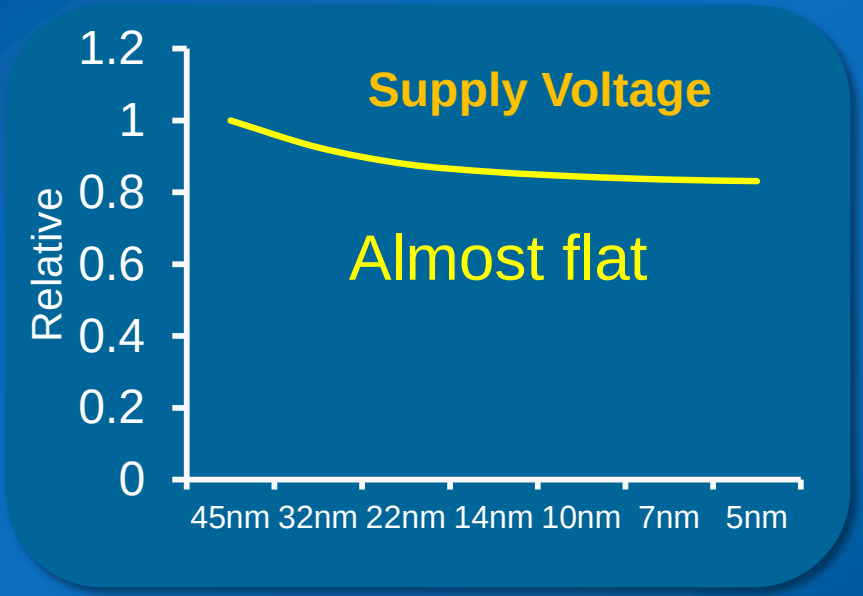
20 mW, Giga

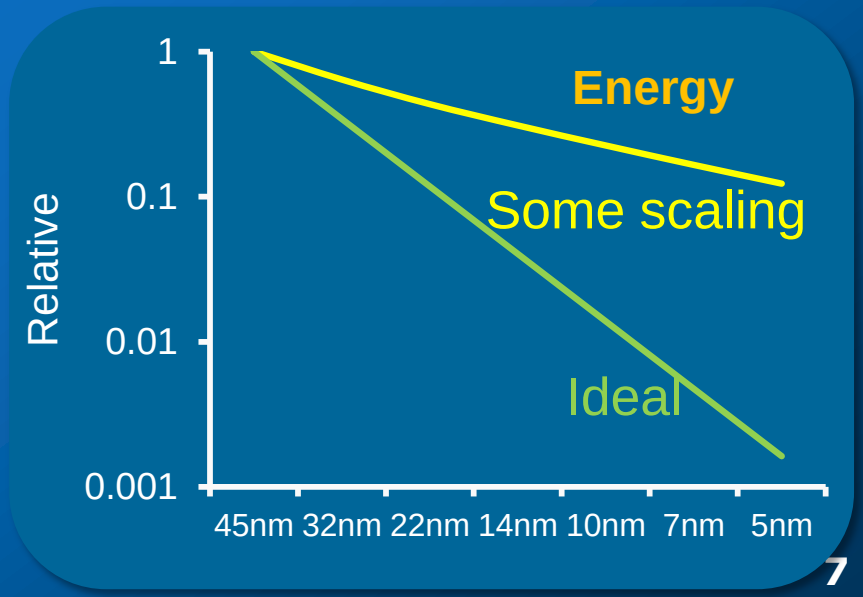
2W, 100 Giga

Technology Scaling Outlook

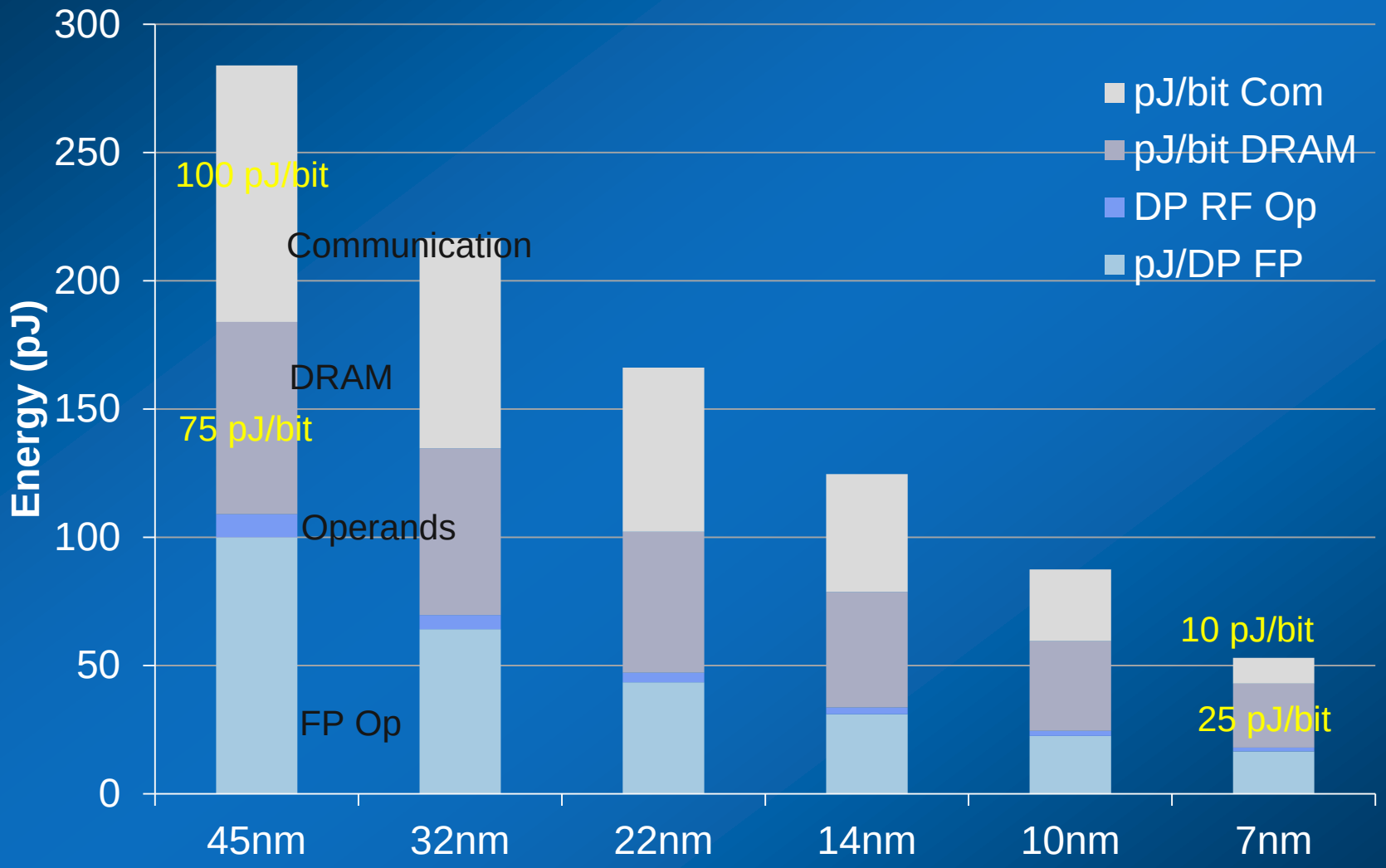








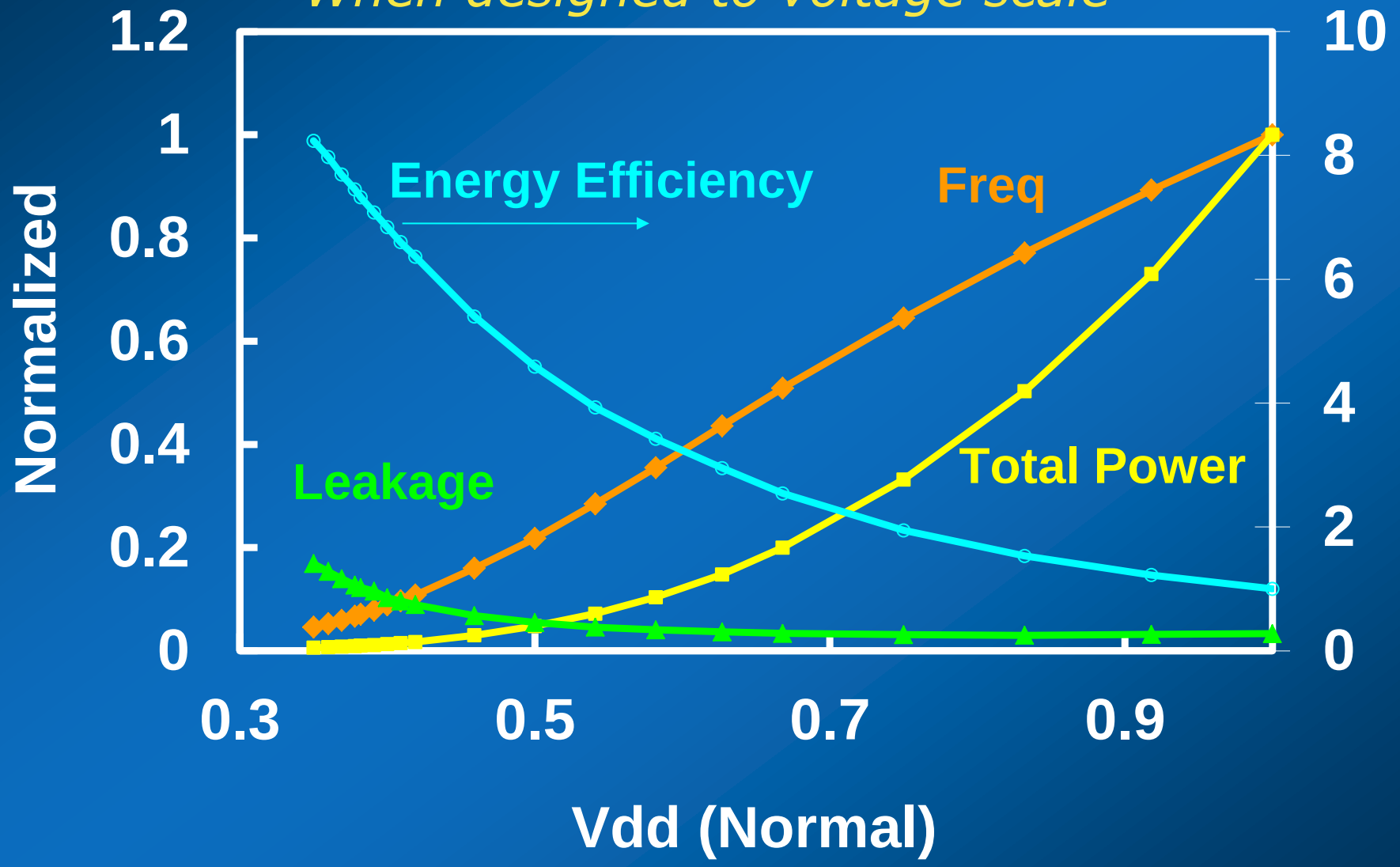
Energy per Compute Operation



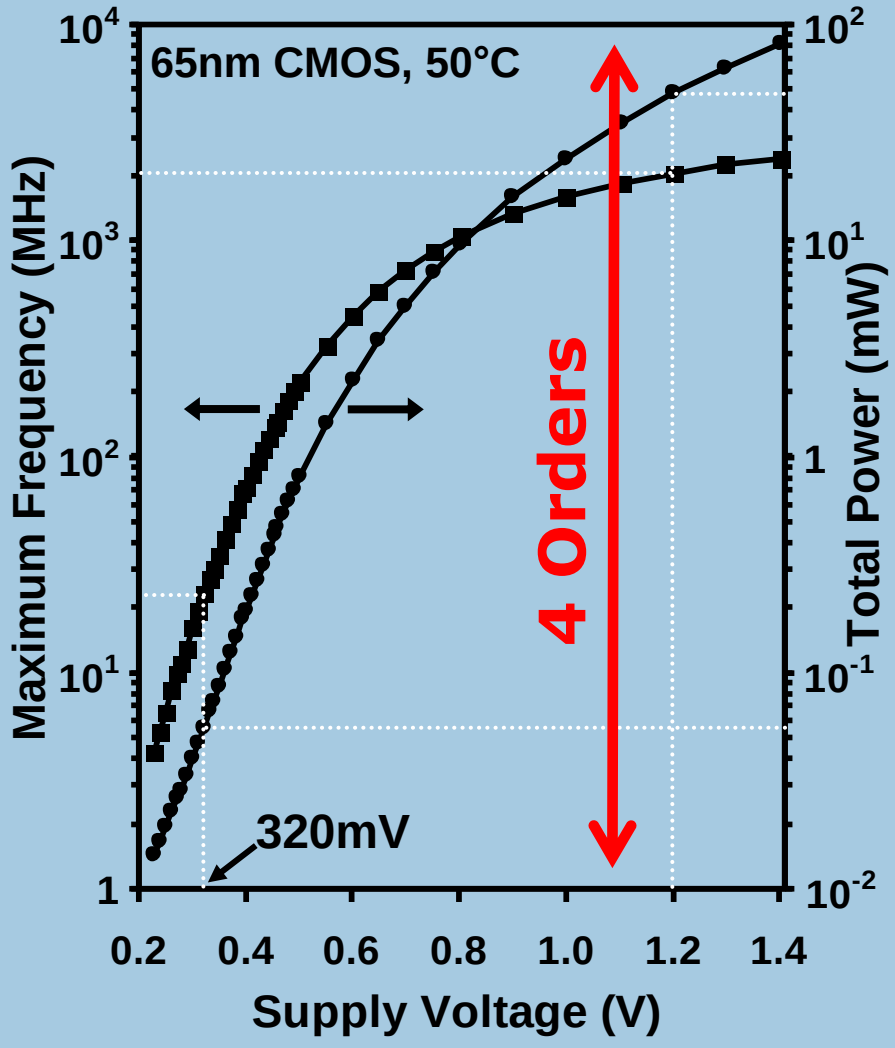
Source: Intel

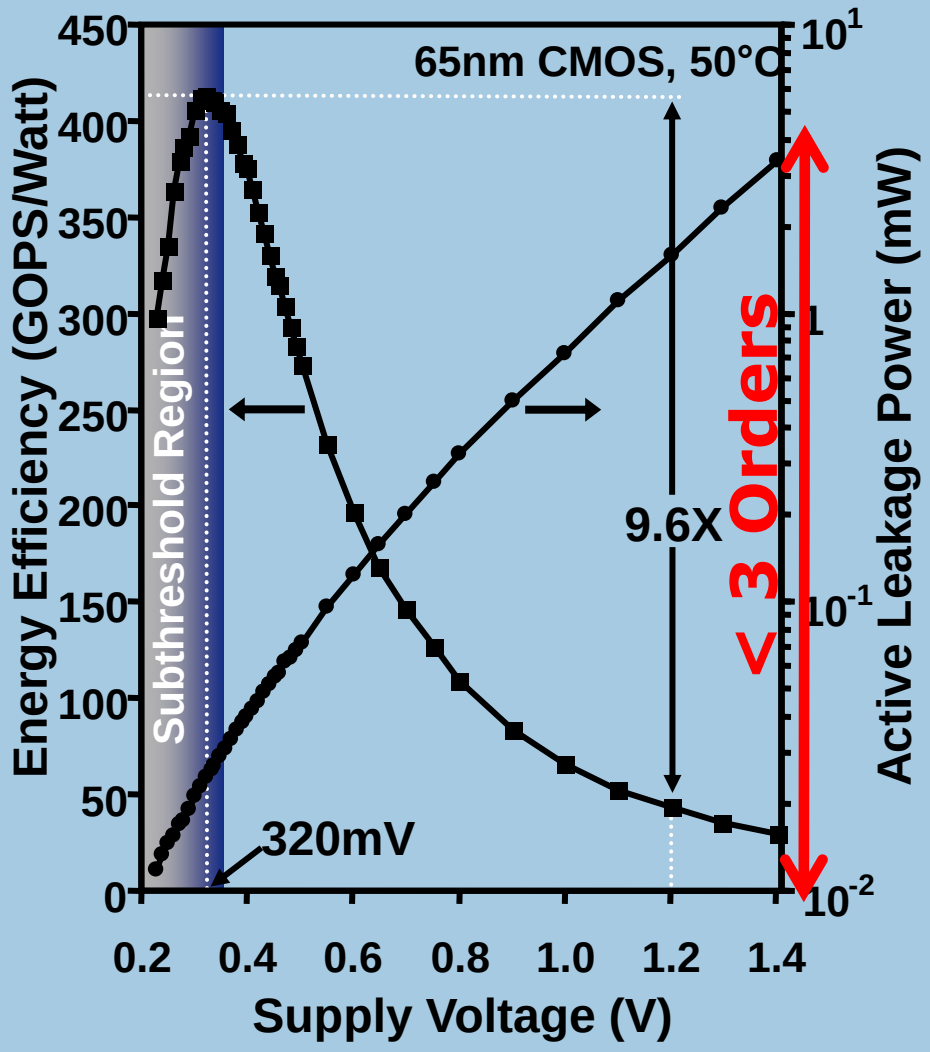
Voltage Scaling

When designed to voltage scale

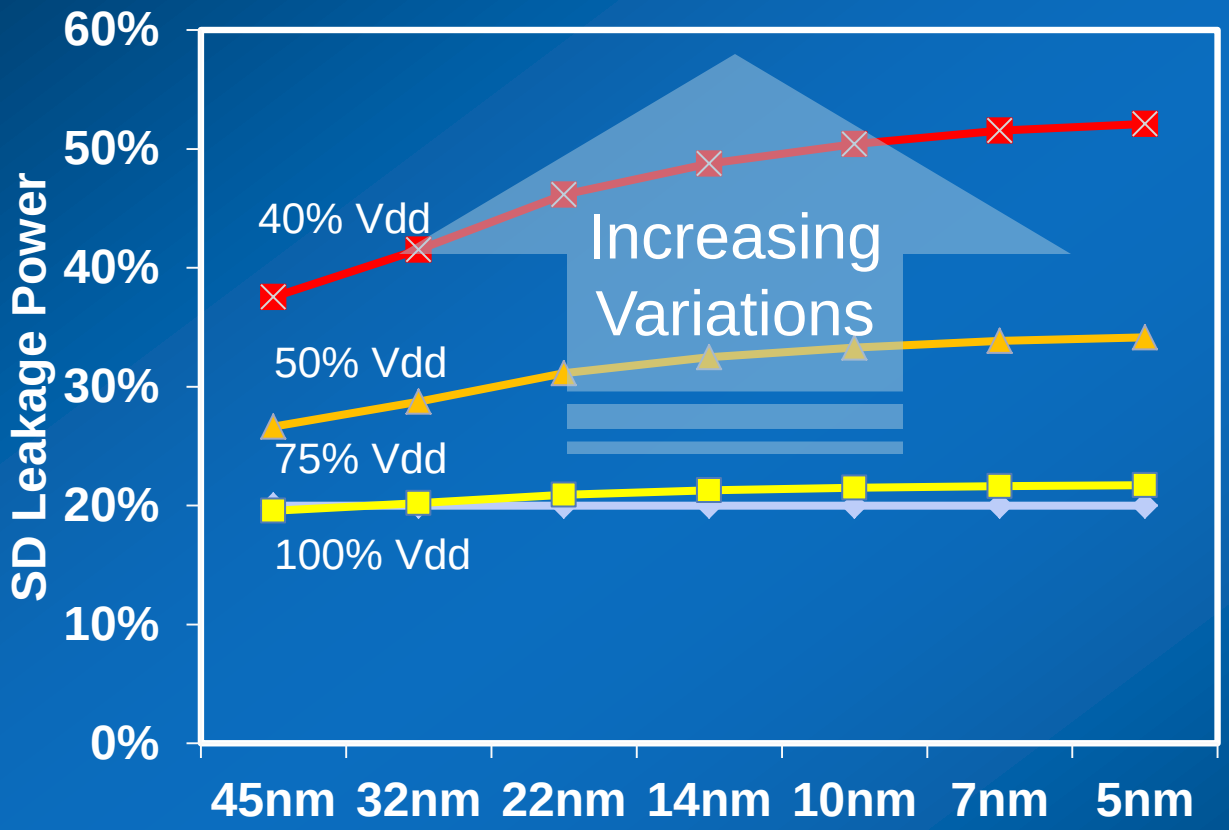


Near Threshold-Voltage (NTV)





Subthreshold Leakage at NTV



NTV operation reduces total power, improves energy efficiency

Subthreshold leakage power is substantial portion of the total

Mitigating Impact of Variation

1. **Variation control with body biasing**

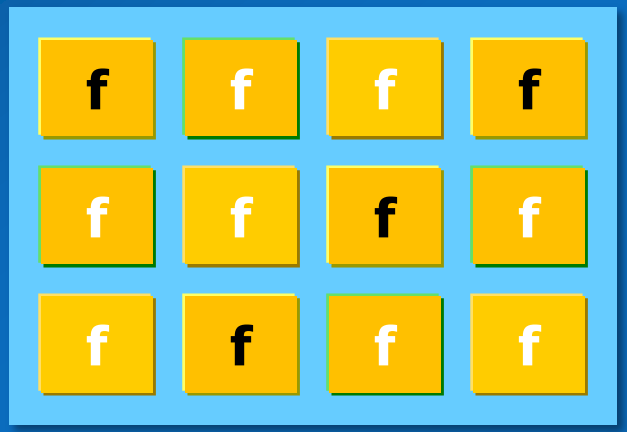
   Body effect is substantially reduced in advanced technologies

   Energy cost of body biasing could become substantial

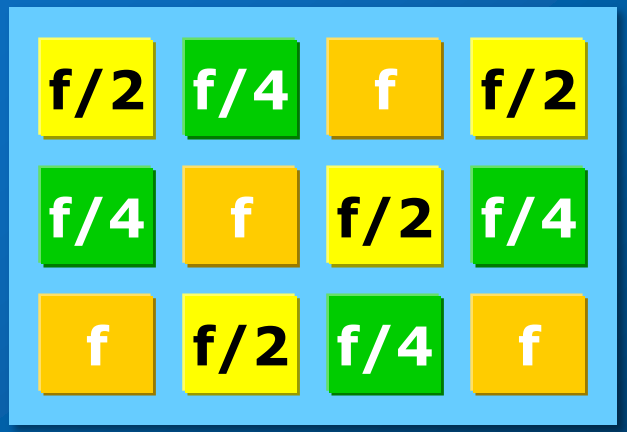
   Fully-depleted transistors have no body left

2. **Variation tolerance at the system level**

Example: Many-core System

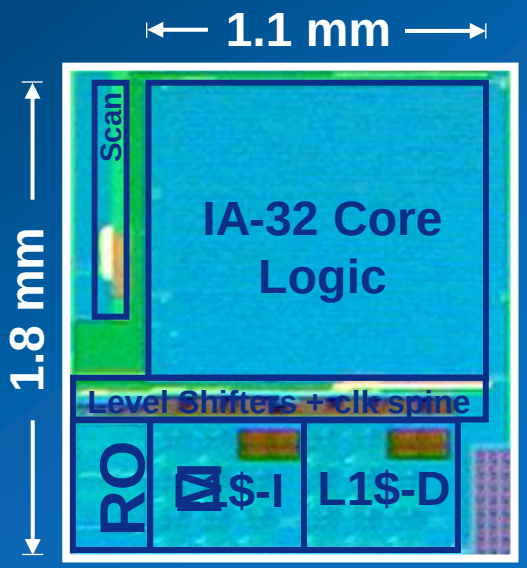


Running all cores at full frequency exceeds energy budget

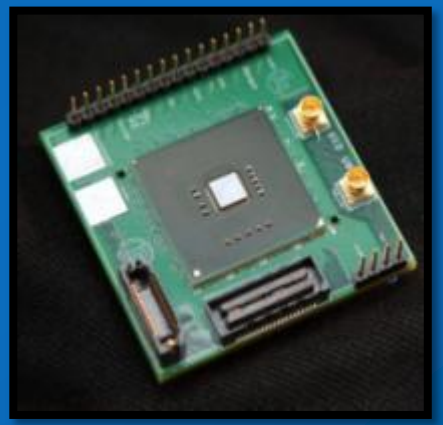


Run cores at the native frequency
Law of large numbers—averaging

Experimental NTV Processor

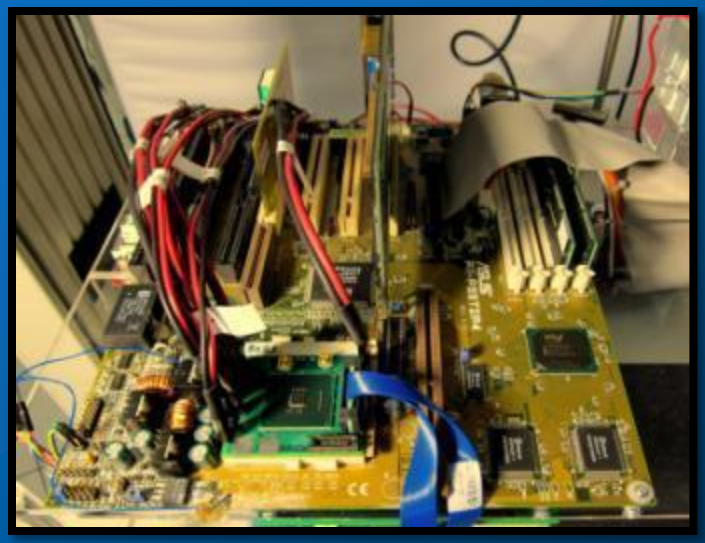


951 Pin FCBGA Package



Custom Interposer

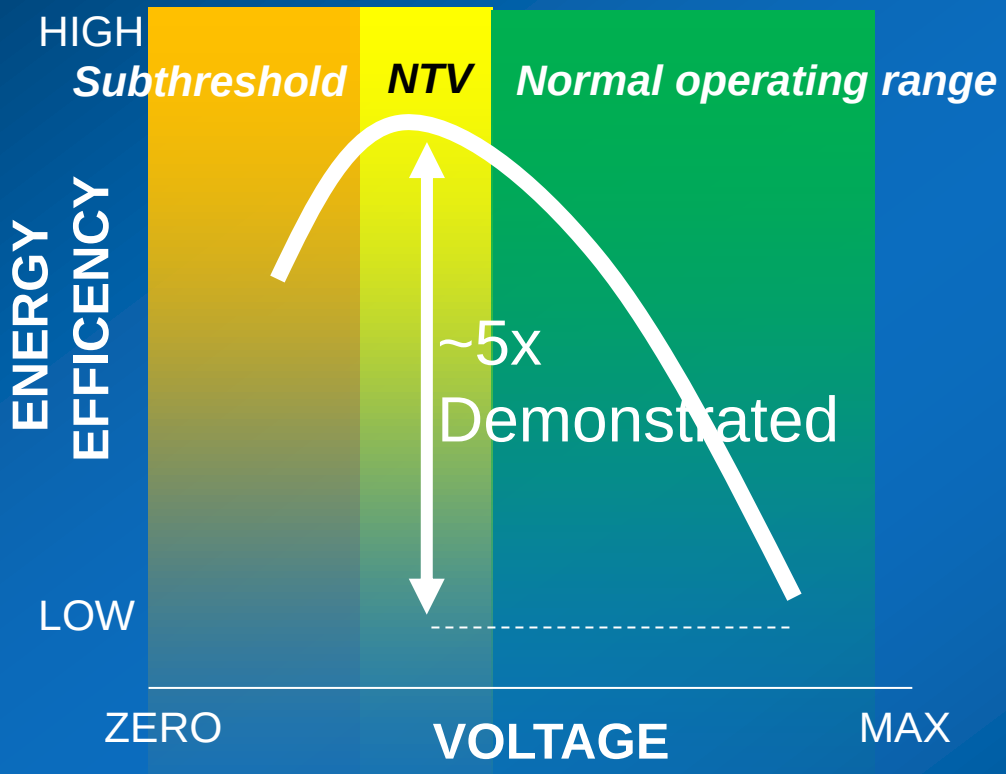
Technology	32nm High-K Metal Gate
Interconnect	1 Poly, 9 Metal (Cu)
Transistors	6 Million (Core)
Core Area	2mm^2



Legacy Socket-7 Motherboard

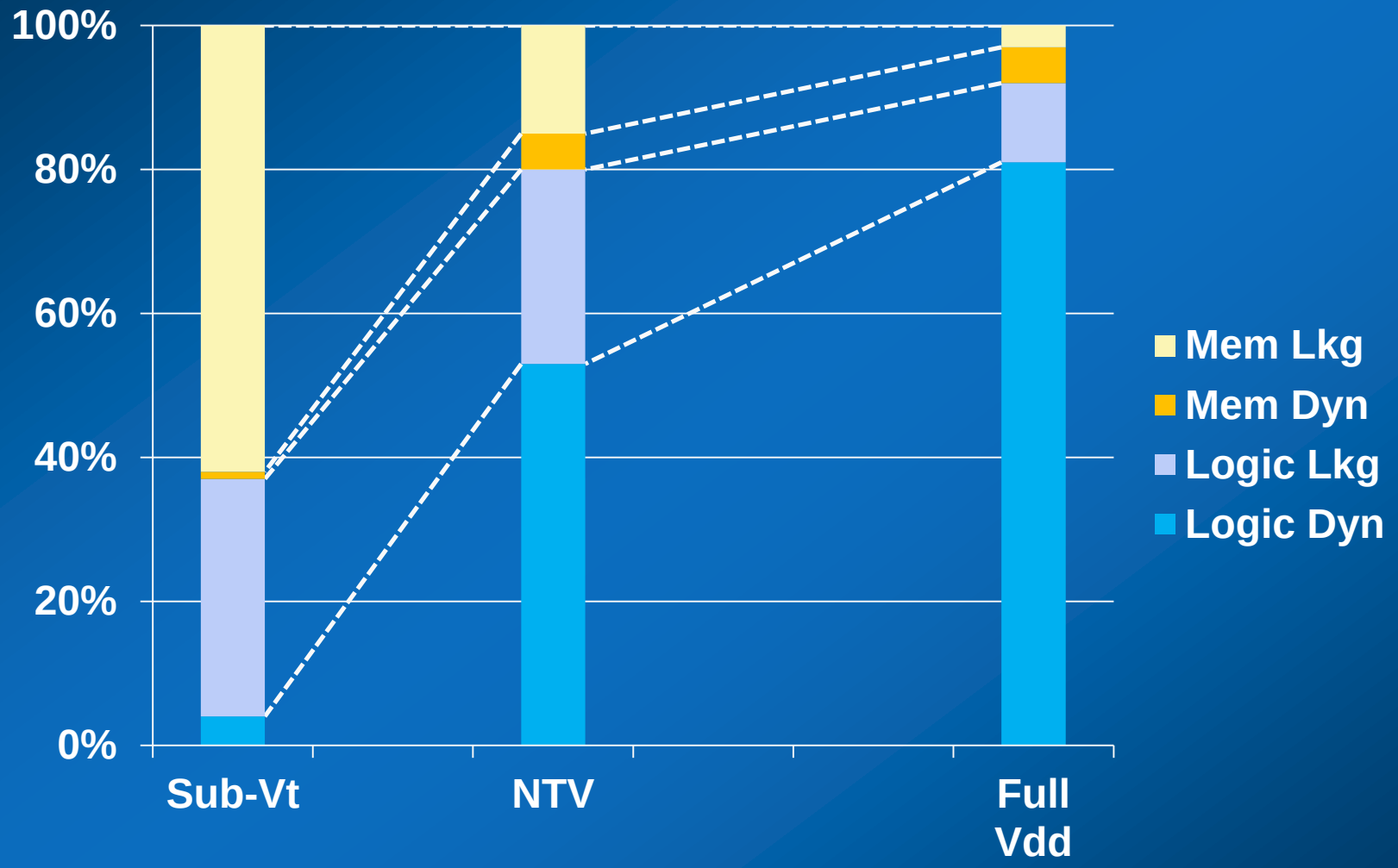
S. Jain, et al, "A 280mV-to-1.2V Wide-Operating-Range IA-32 Processor in 32nm CMOS", ISSCC 2012

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



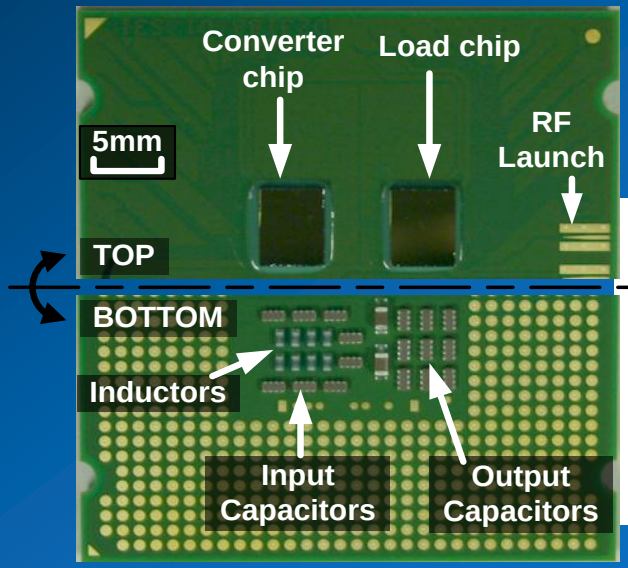
Leakage power dominates

Fine grain leakage power management is required

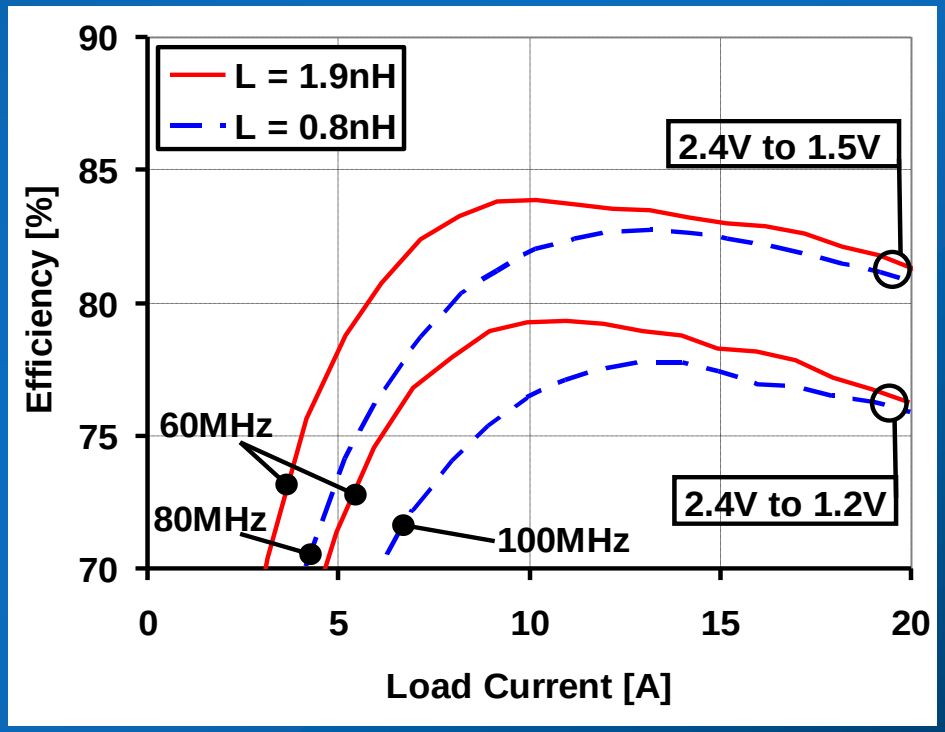
Integration of Power Delivery

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

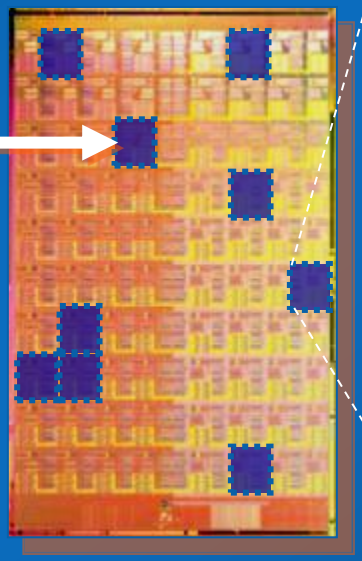
Fine-grain Power Management

Mode		Power Saving	Wake up
Normal	All active	-	-
Standby	Logic off Memory on	50%	Fast
Sleep	Logic and Memory off	80%	Slow

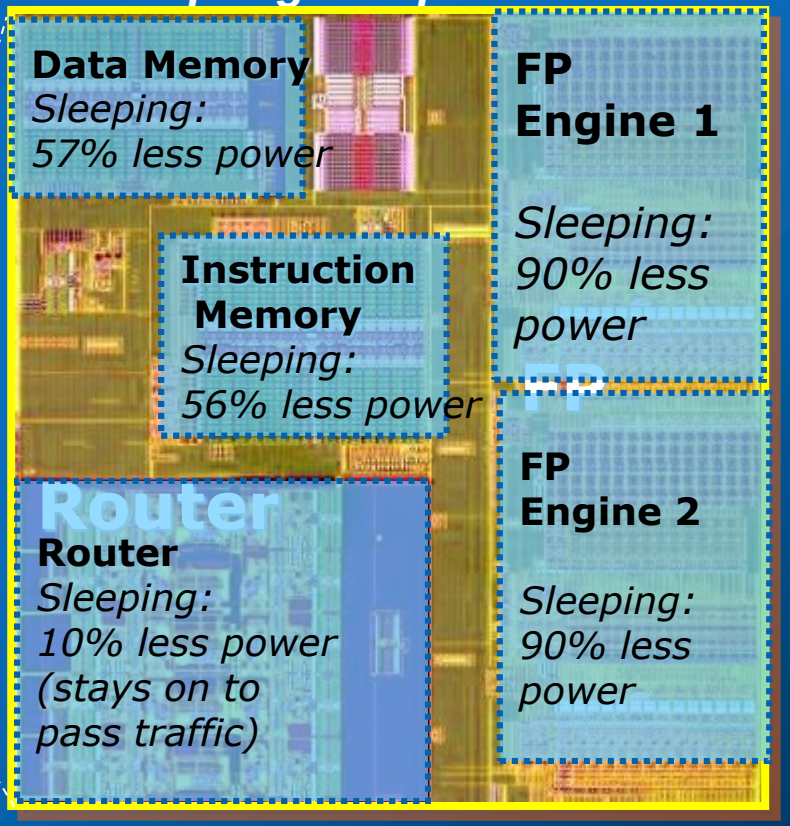
Dynamic Chip Level

STANDBY:
- Memory retains data
- **50%** less power/tile

FULL SLEEP:
- Memories fully off
- **80%** less power/tile



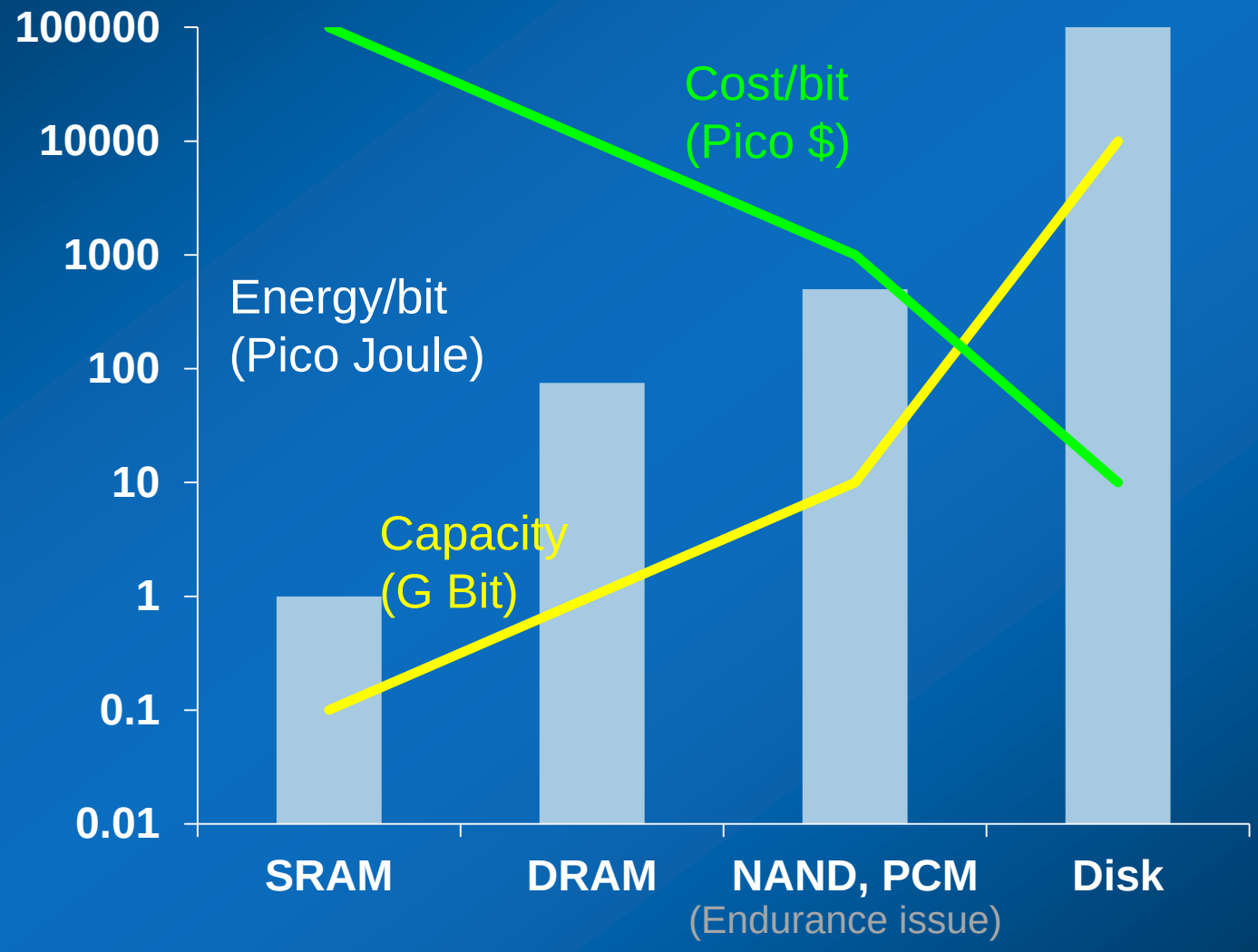
Dynamic, within a core
21 sleep regions per tile



Energy efficiency increases by 60%,

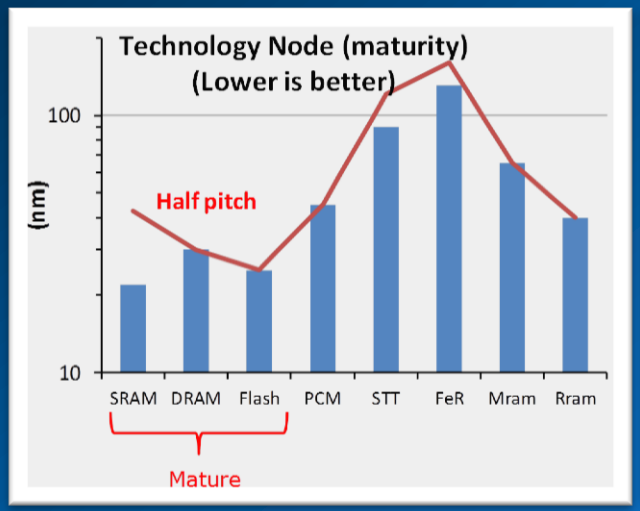
Vangal et al, An 80-Tile Sub-100-W TeraFLOPS Processor in 65-nm CMOS, JSSC, January 2008

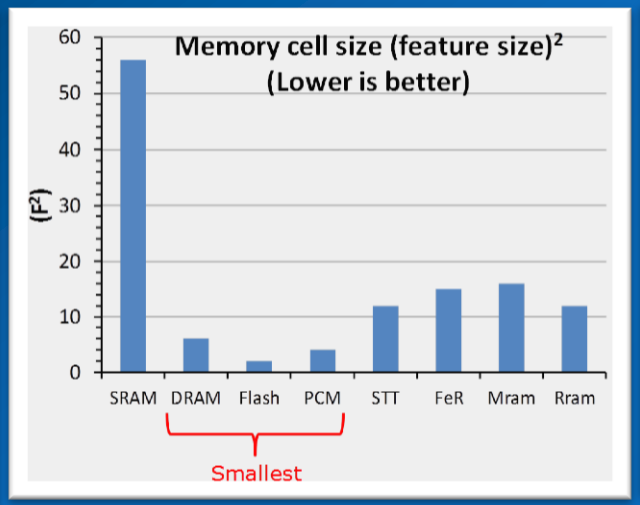
Memory & Storage Technologies

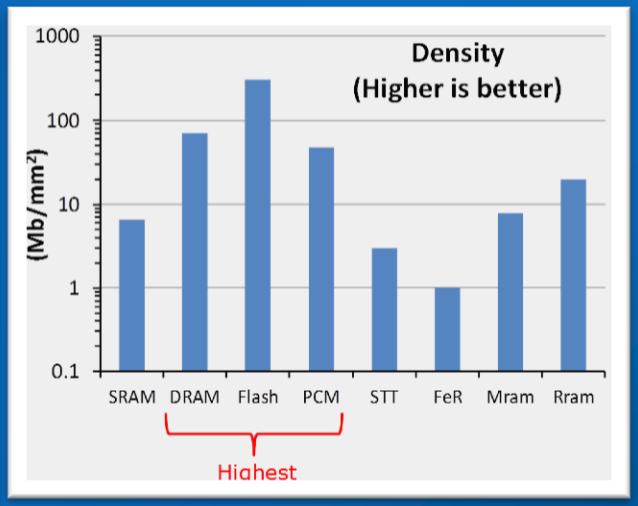


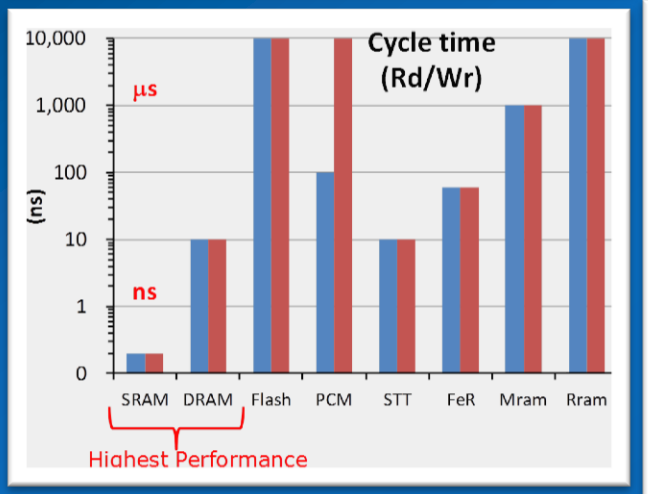
Source: Intel

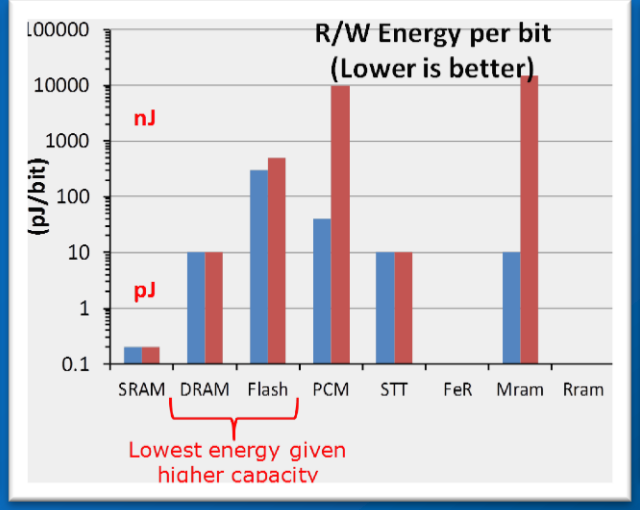
Compare Memory Technologies

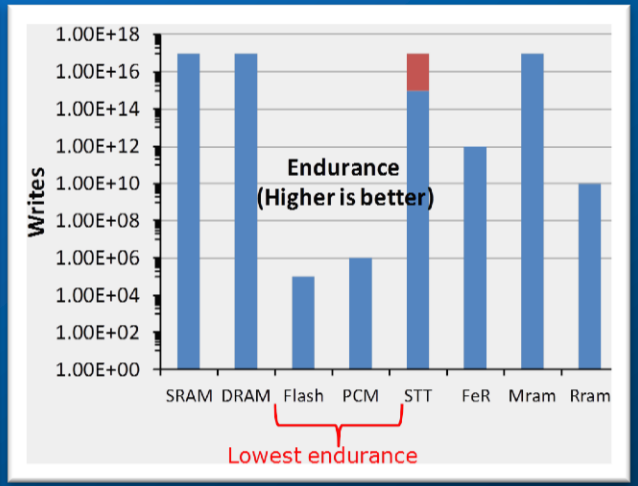










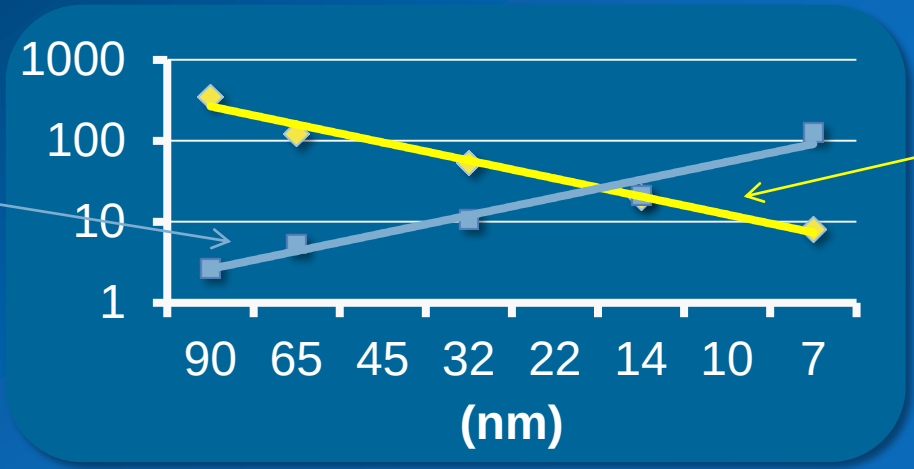


Source: Intel

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

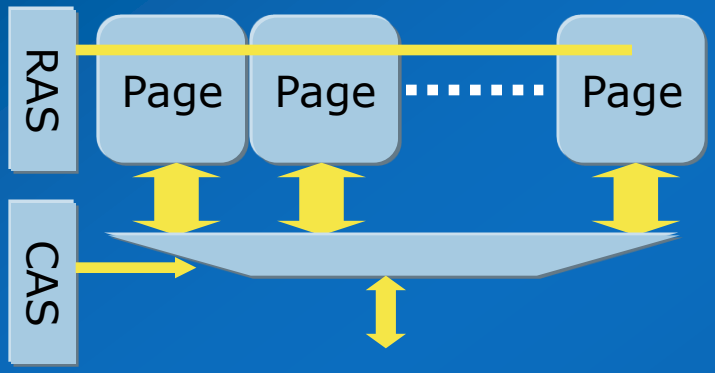
Revise DRAM Architecture

❶ Need exponentially increasing BW (GB/sec)



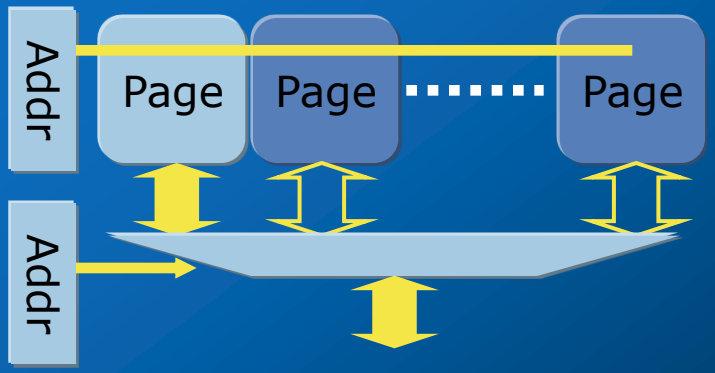
❷ Need exponentially decreasing energy (pJ/bit)

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

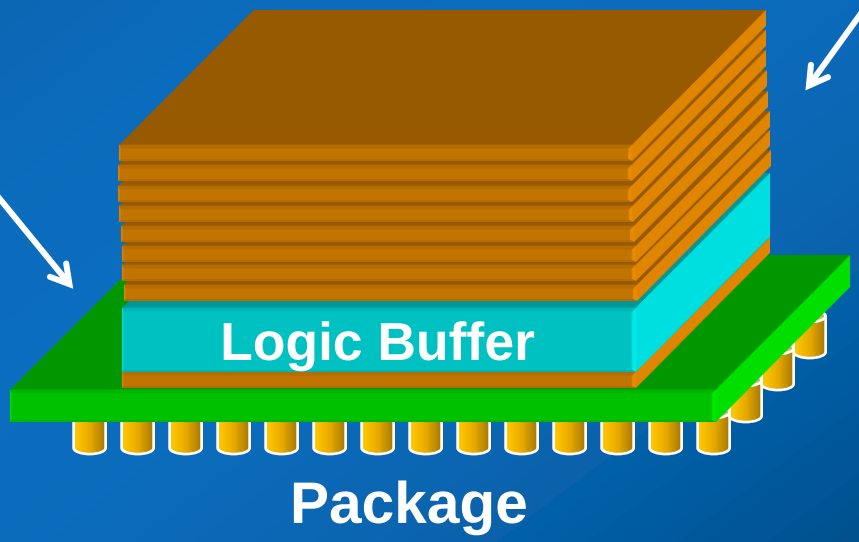
Implement intelligence

DRAM Stack

Technology optimized for:

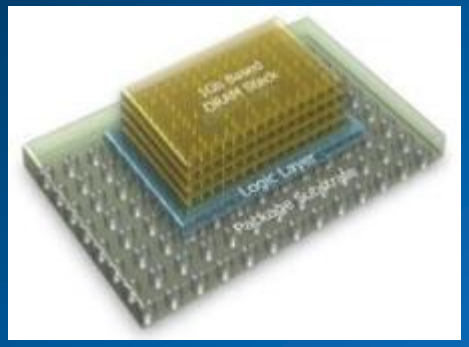
Memory density

Lower cost

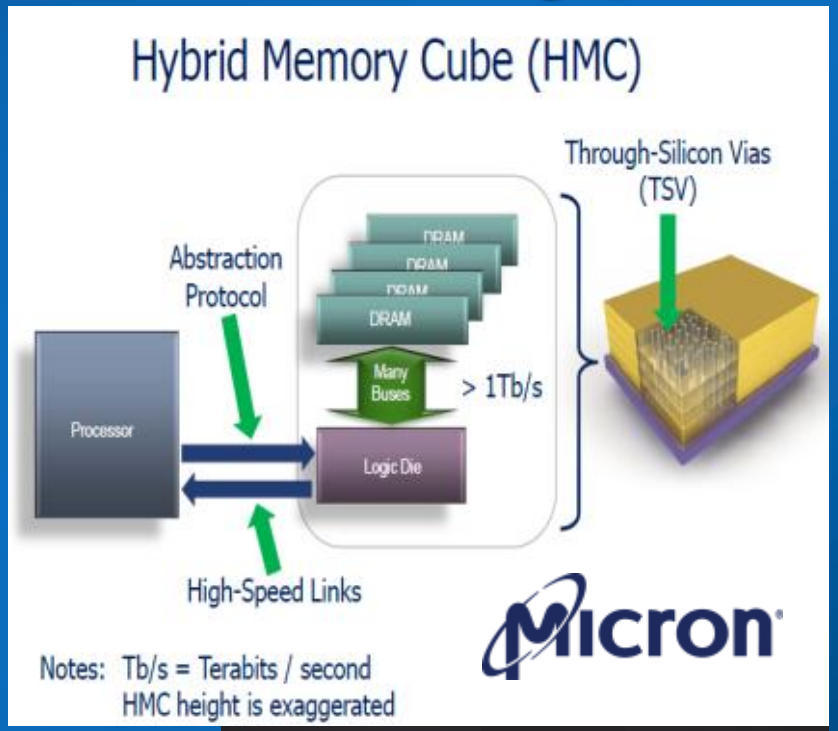


3D Integration provides best of both worlds

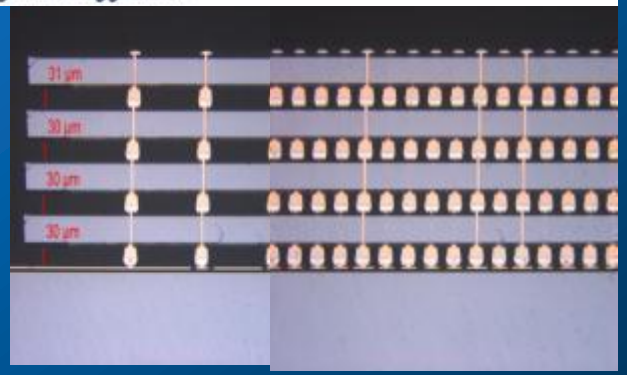
1Tb/s HMC DRAM Prototype



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



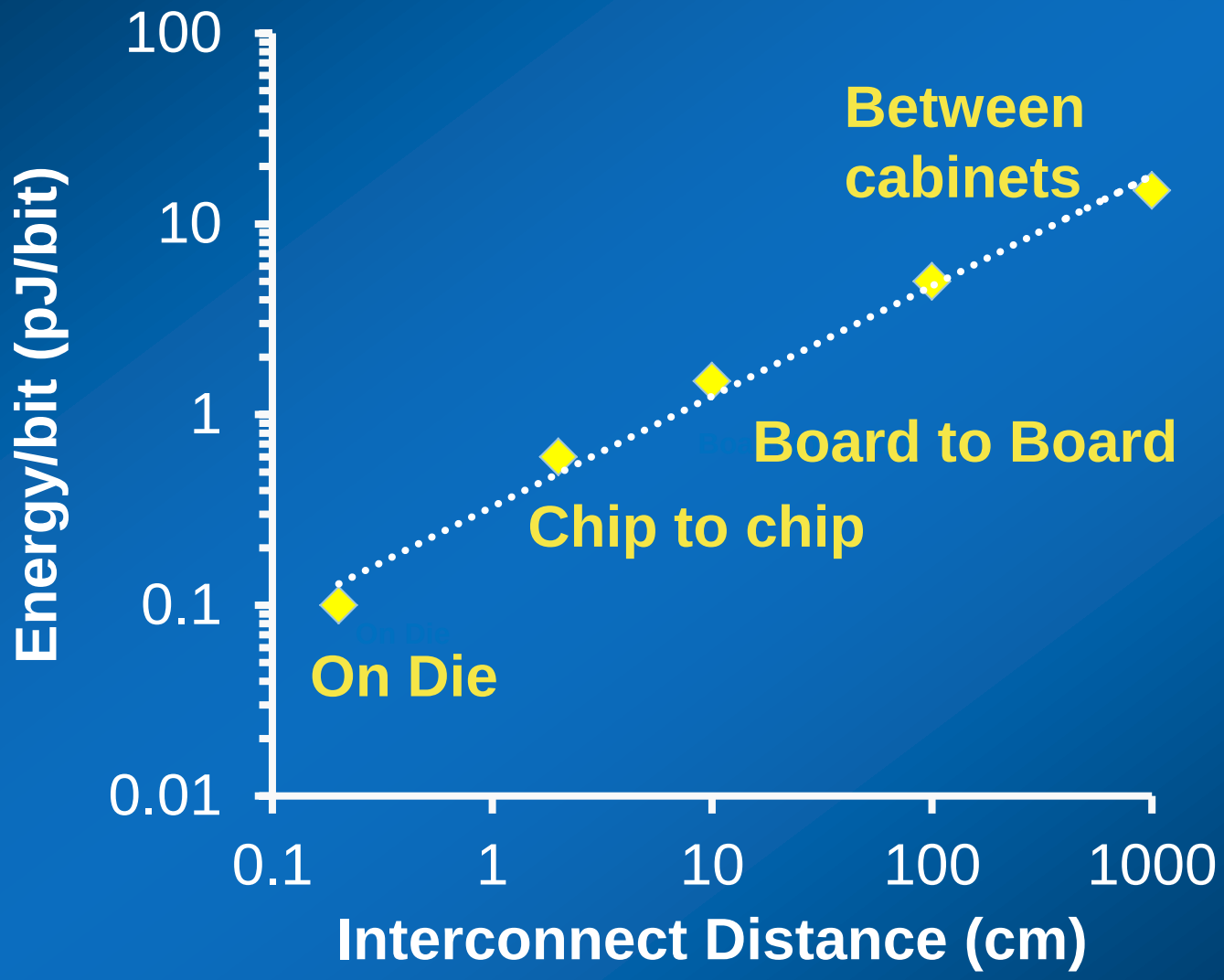
	Bandwidth	Energy Efficiency
DDR-3 (Today)	10.66 GB/Sec	50-75 pJ/bit
Hybrid Memory Cube	128 GB/Sec	**8 pJ/bit**



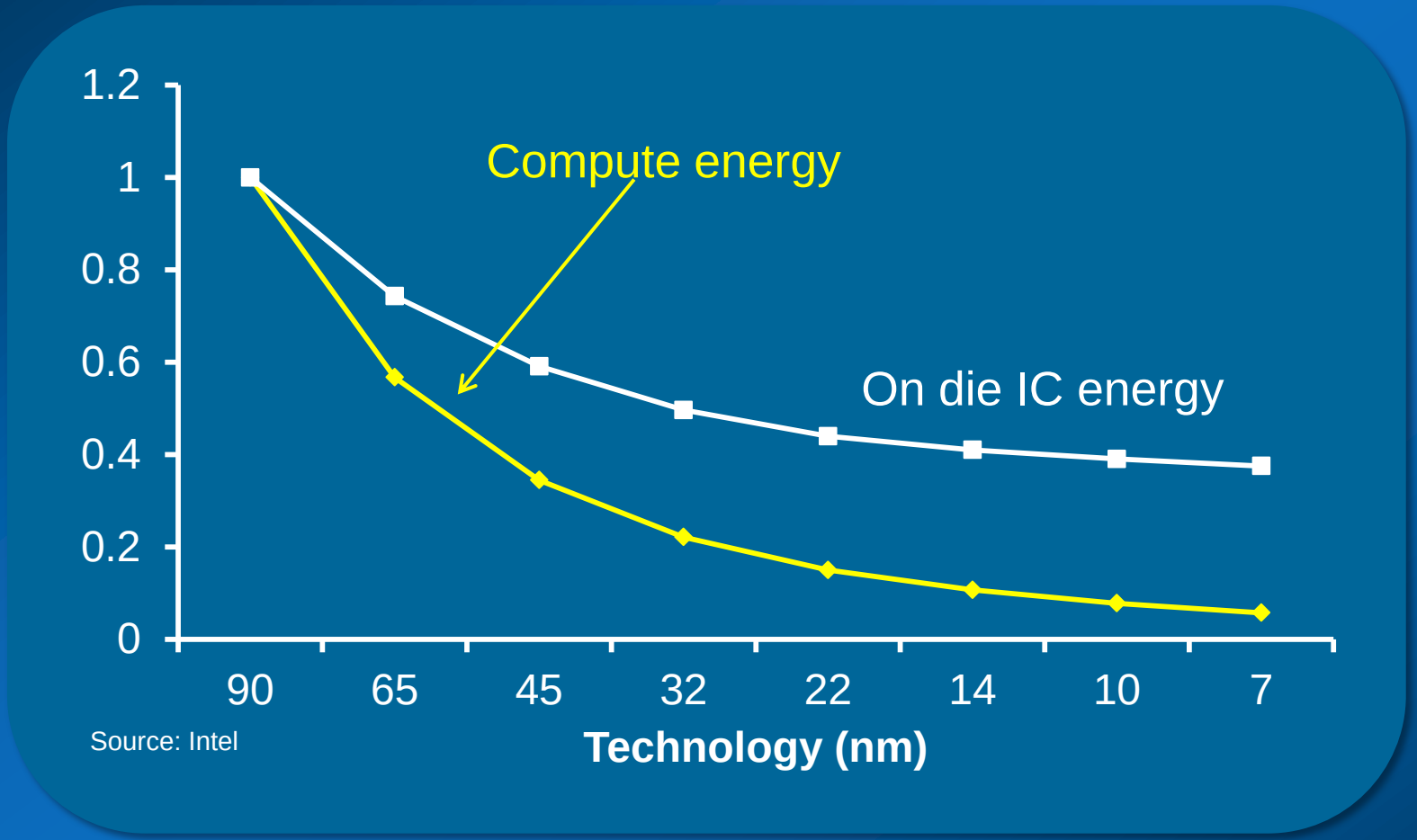
10X higher bandwidth, 10X lower energy

Source: Micron

Communication Energy



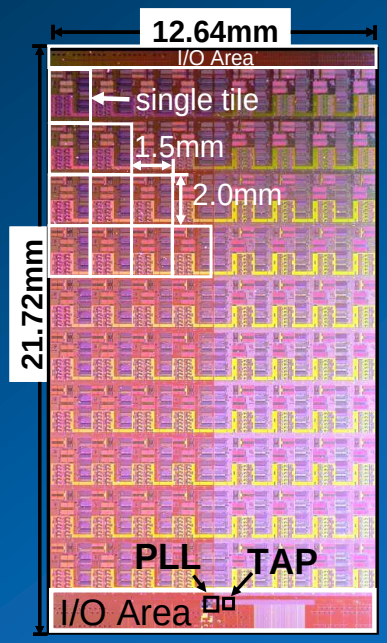
On-die Interconnect

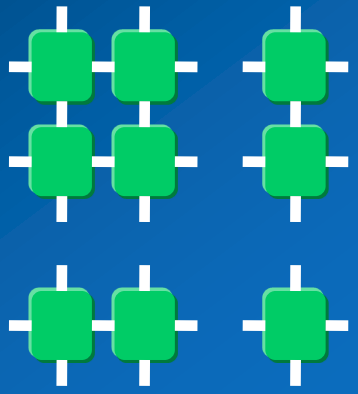


Interconnect energy (per mm) reduces slower than compute
On-die data movement energy will start to dominate

Network On Chip (NoC)

80 Core TFLOP Chip (2006)

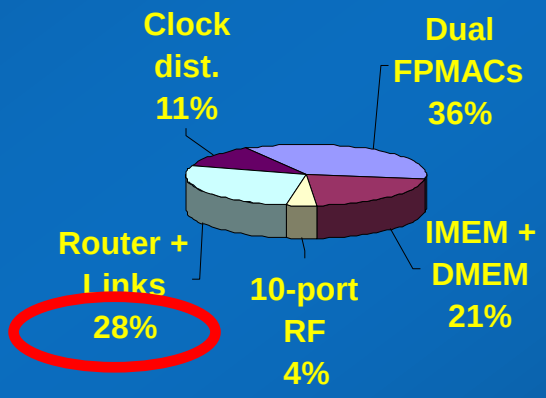




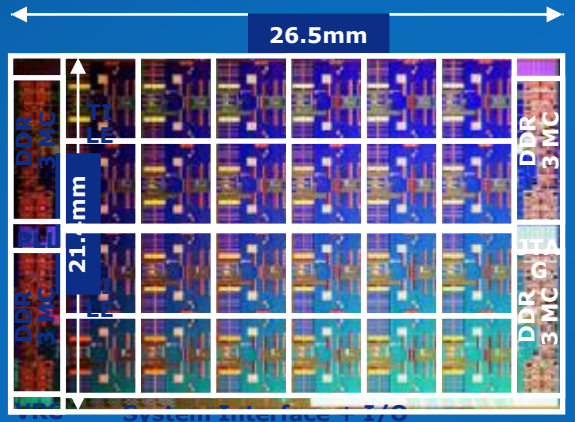
8 X 10 Mesh

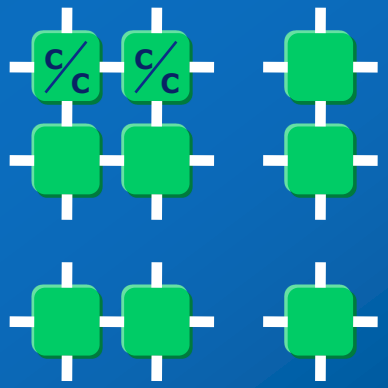
32 bit links

320 GB/sec bisection BW @ 5 GHz



48 Core Single Chip Cloud (2009)



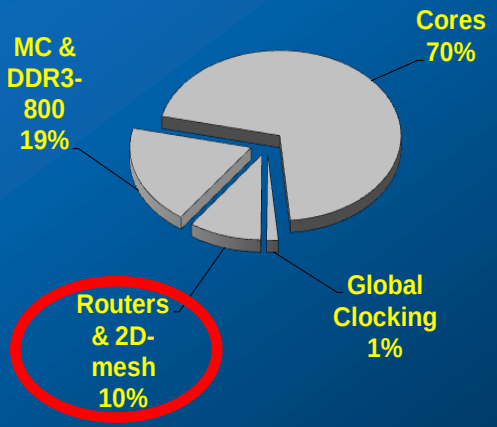


2 Core clusters in 6 X 4 Mesh

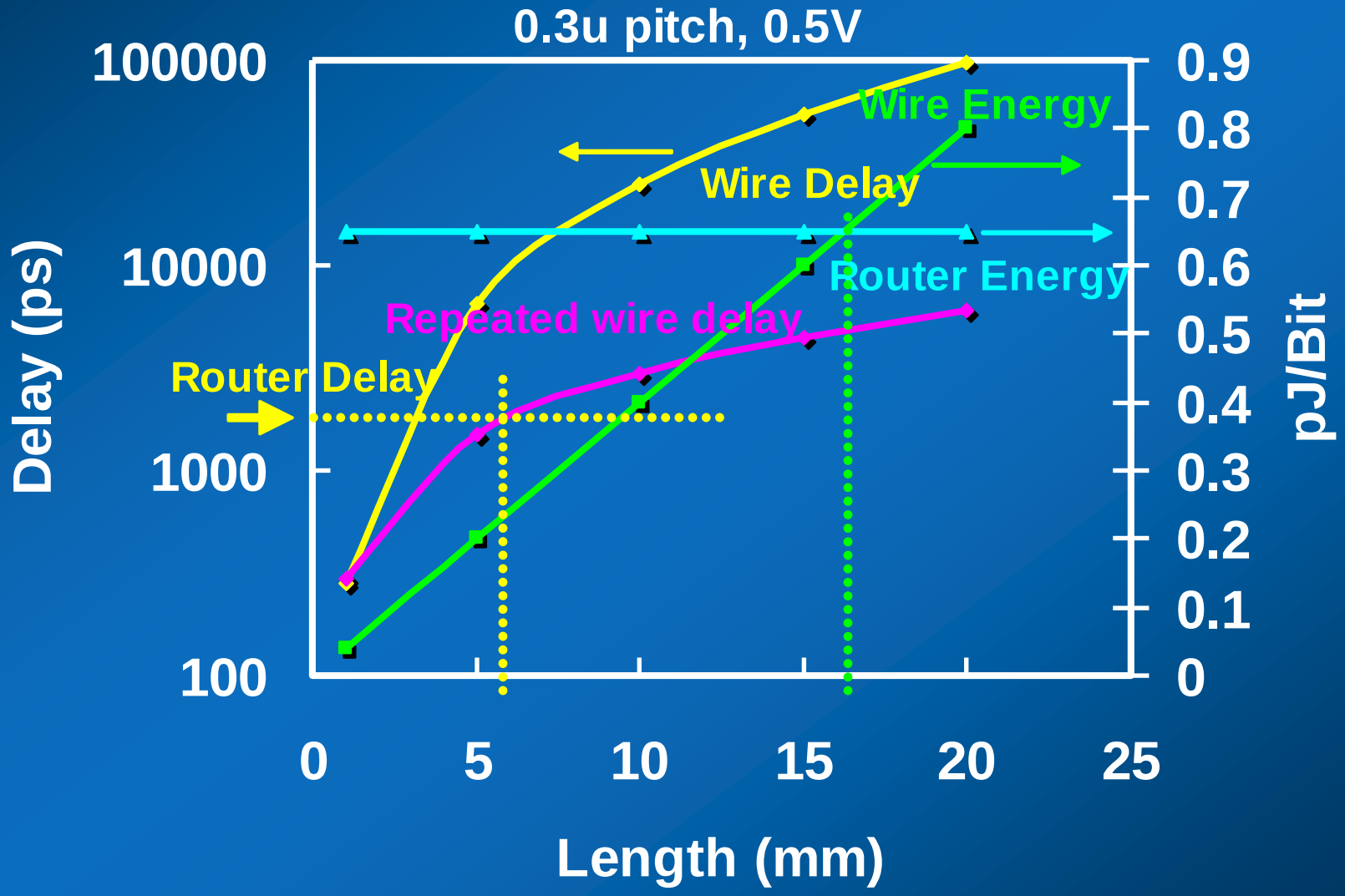
(why not 6 x 8?)

128 bit links

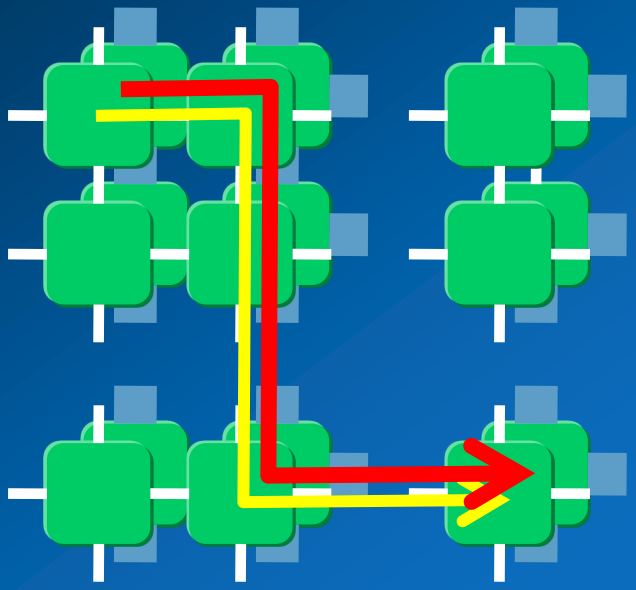
256 GB/sec bisection BW @ 2 GHz

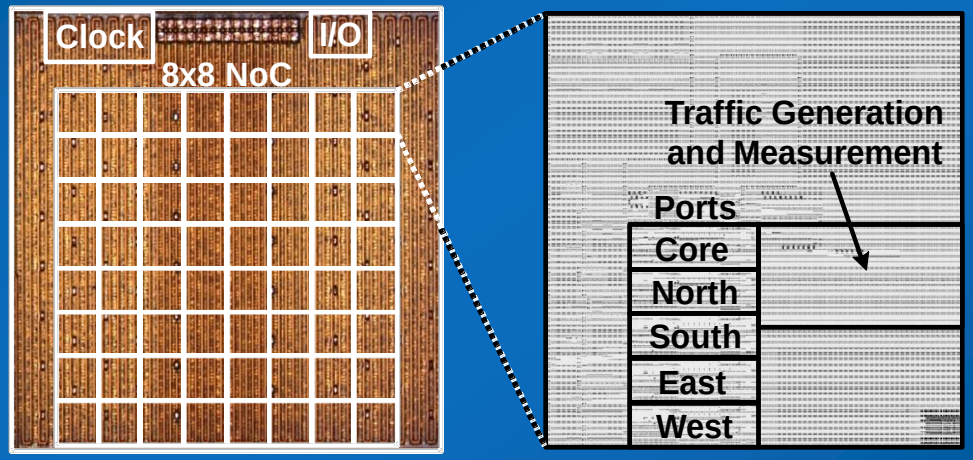


On-chip Interconnects



Circuit Switched NoC





Narrow, high freq packet switched network establishes a circuit

Data is transferred using wide and slower established circuit switched bus

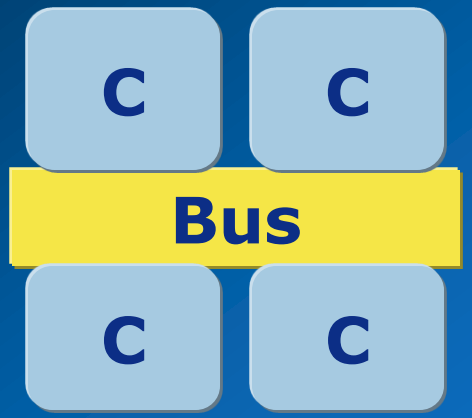
Differential, low swing bus improves energy efficiency

Process	45nm Hi-K/MG CMOS, 9 Metal Cu
Nominal Supply	1.1V
Arbitration and Router Logic	Supports 512b data
Number of Transistors	2.85M
Die Area	6.25mm^2

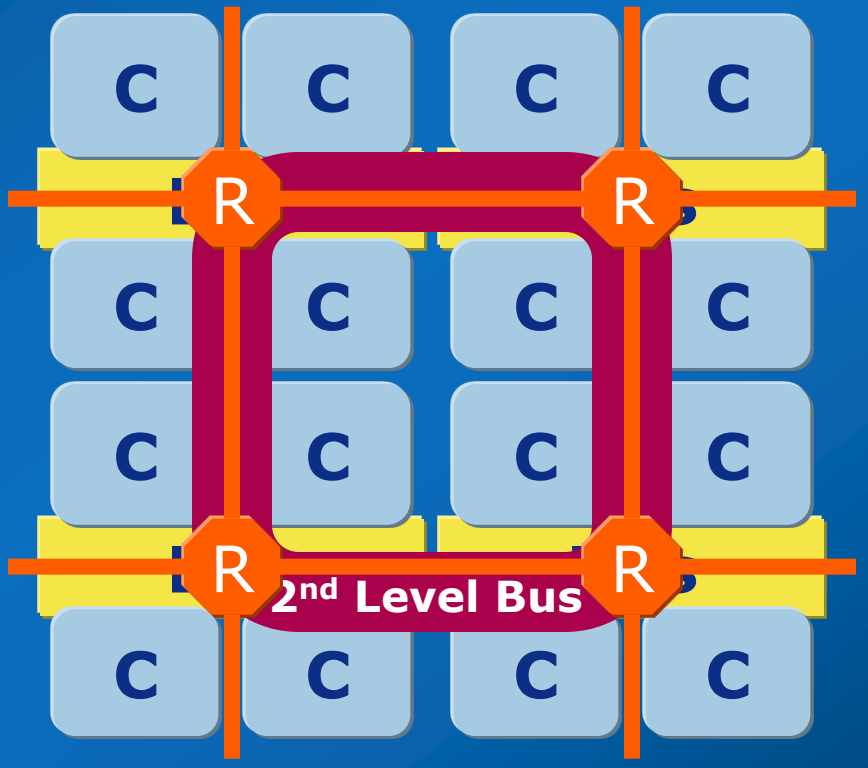
2 to 3X increased energy efficiency over packet switched network

Anders et al, A 4.1Tb/s bisection-bandwidth 560Gb/s/W streaming circuit-switched 8×8 mesh network-on-chip in 45nm CMOS, ISSCC 2010

Hierarchical & Heterogeneous

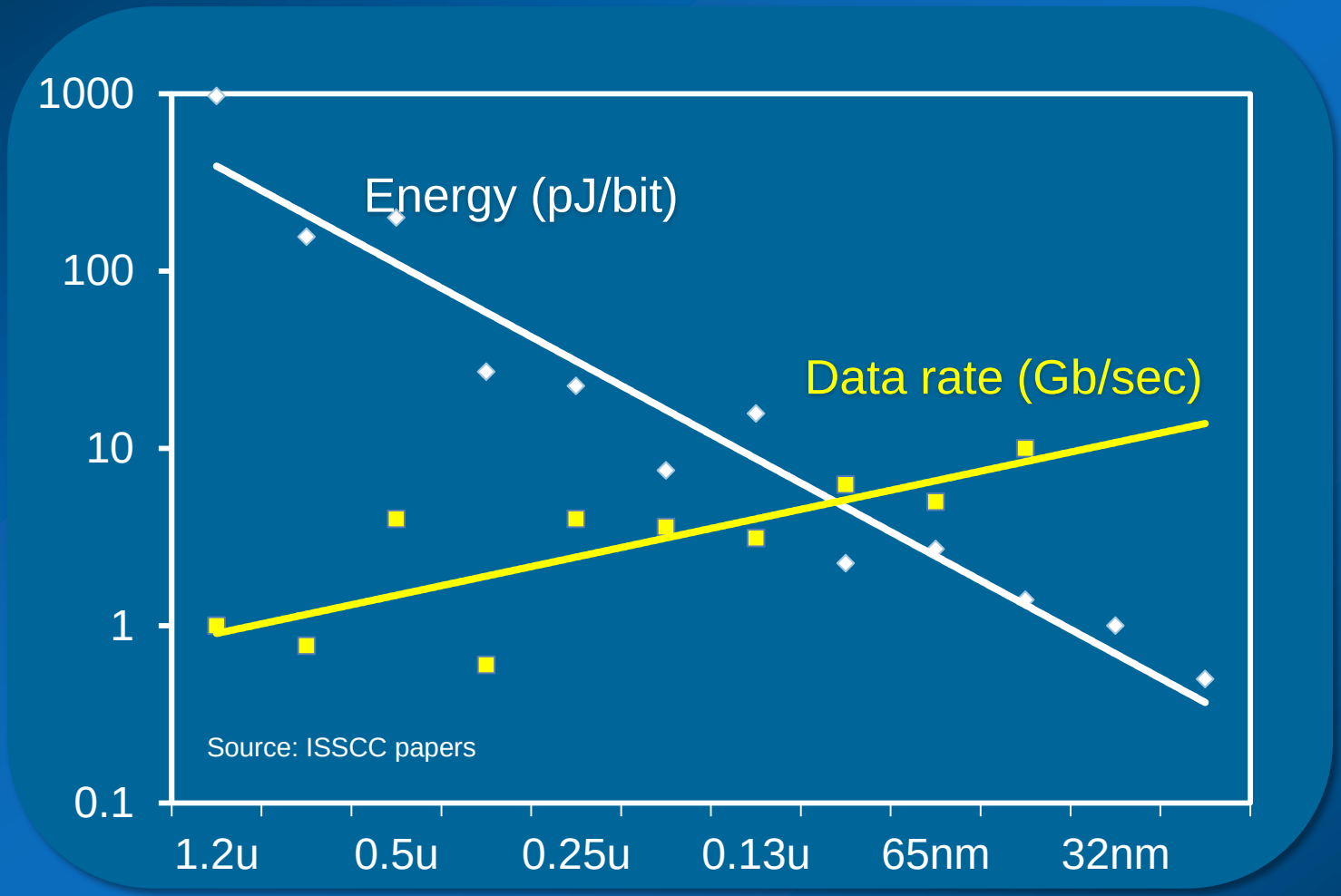


Bus to connect over short distances



Hierarchy of Buses or hierarchical circuit and packet switched networks

Electrical Interconnect < 1 Meter



BW and Energy efficiency improves, but not enough

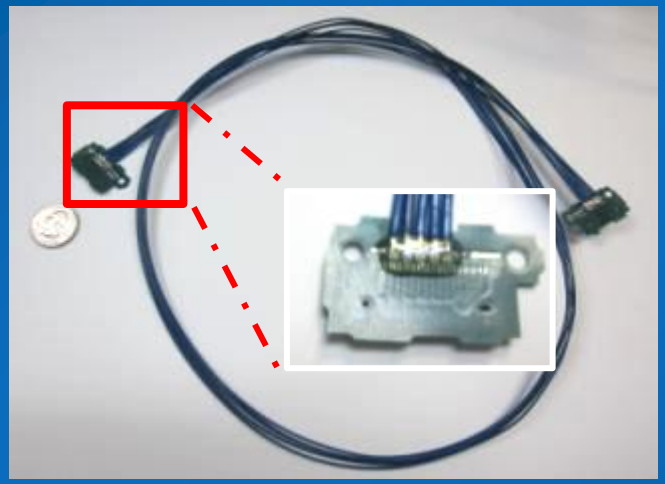
Electrical Interconnect Advances

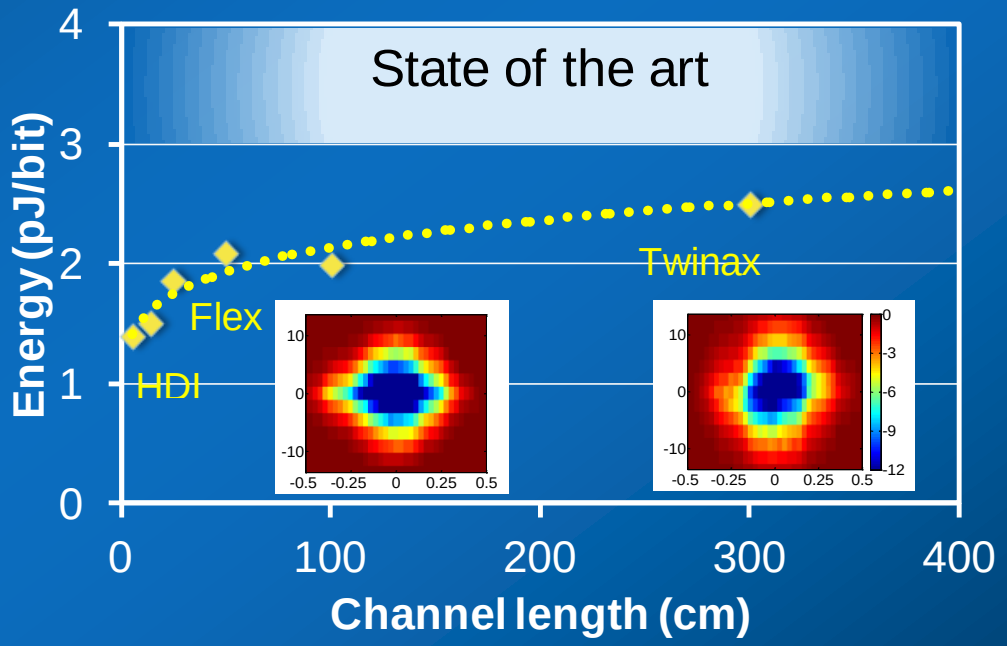
Employ, new, low-loss, non-traditional interconnects

Top of the package connector

Low-loss flex connector

Low-loss twinax

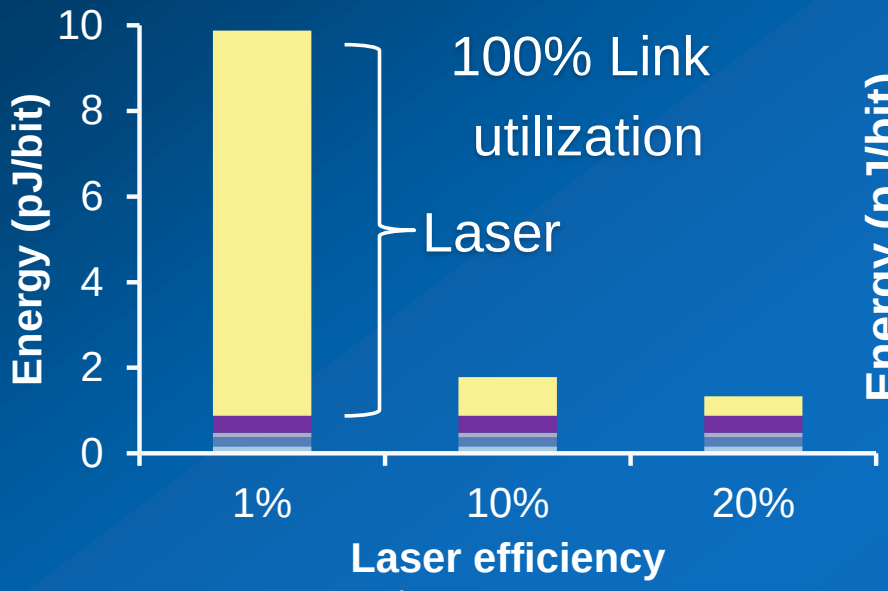


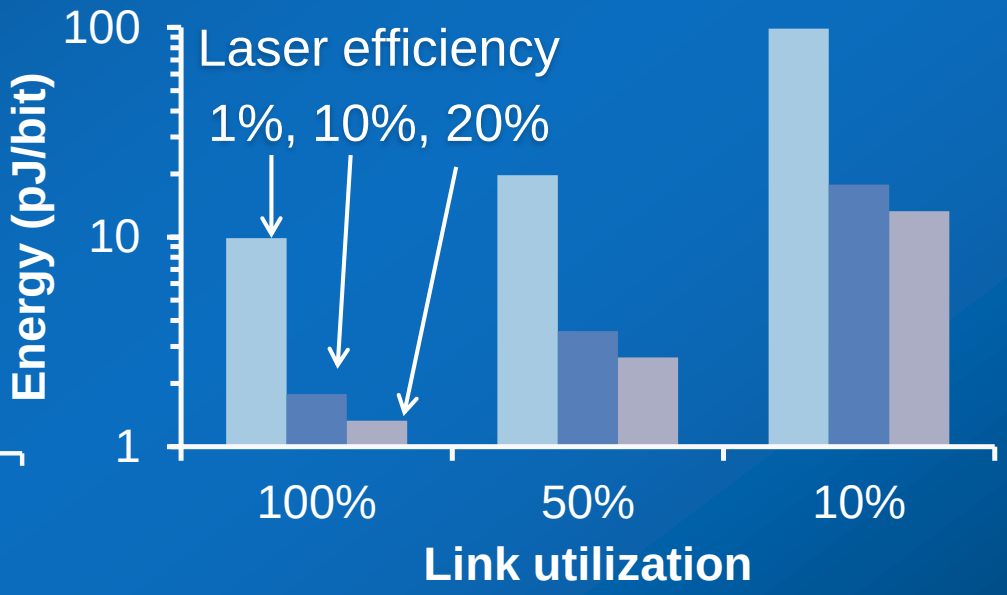


Co-optimization of interconnects and circuits for energy efficiency

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

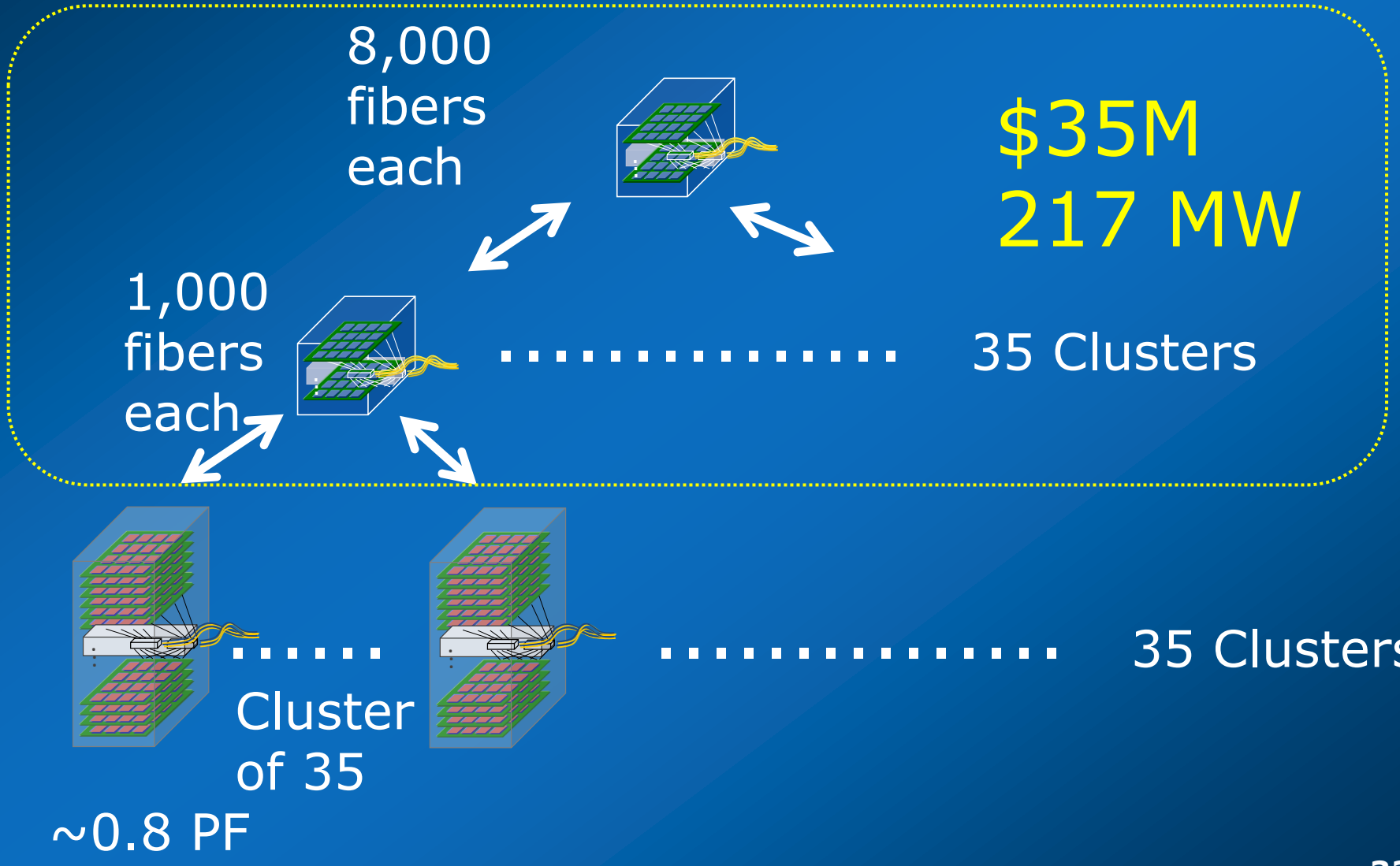
Energy in supporting electronics is very low

Link energy dominated by laser (efficiency)

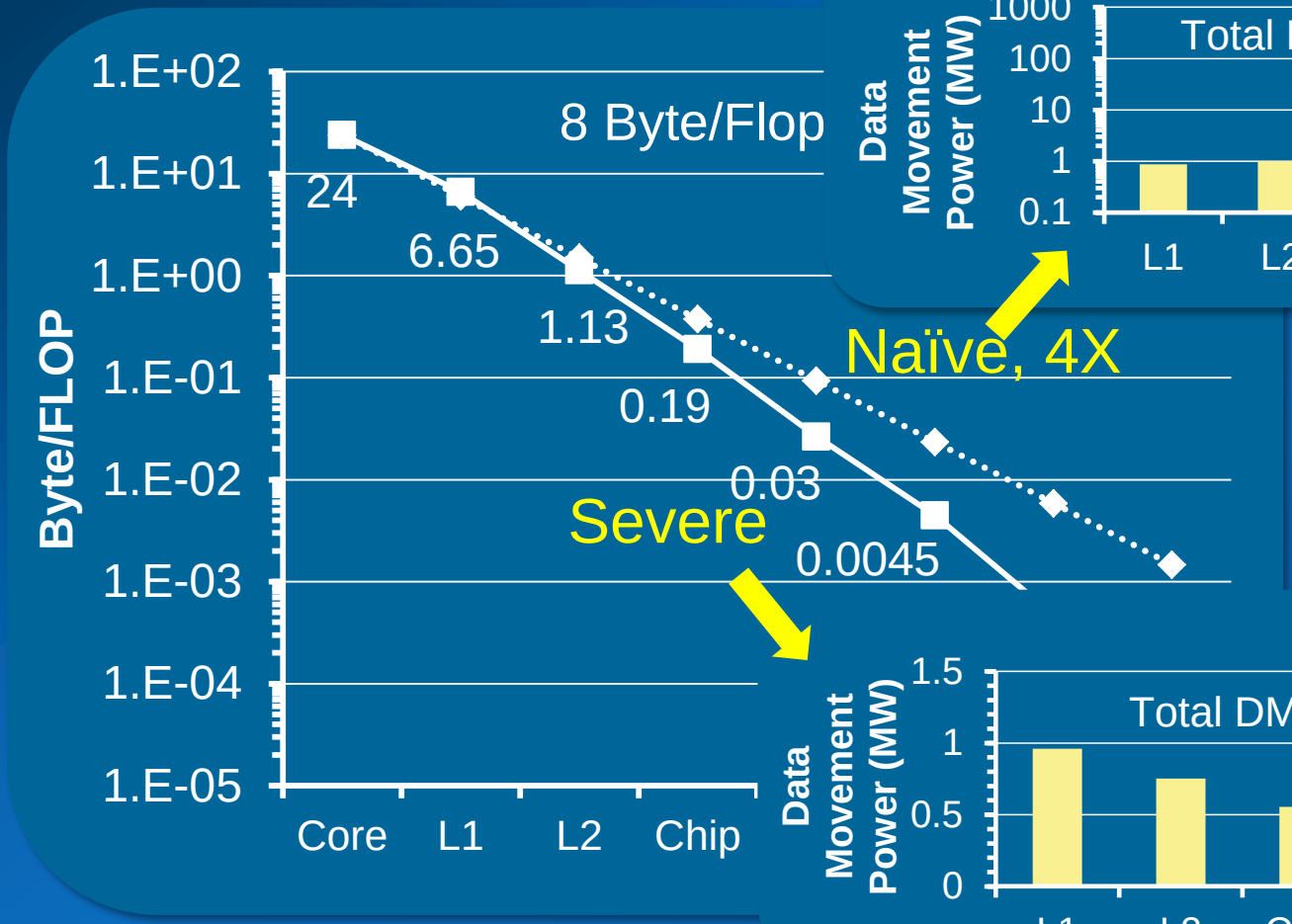
Sustained, high link utilization required

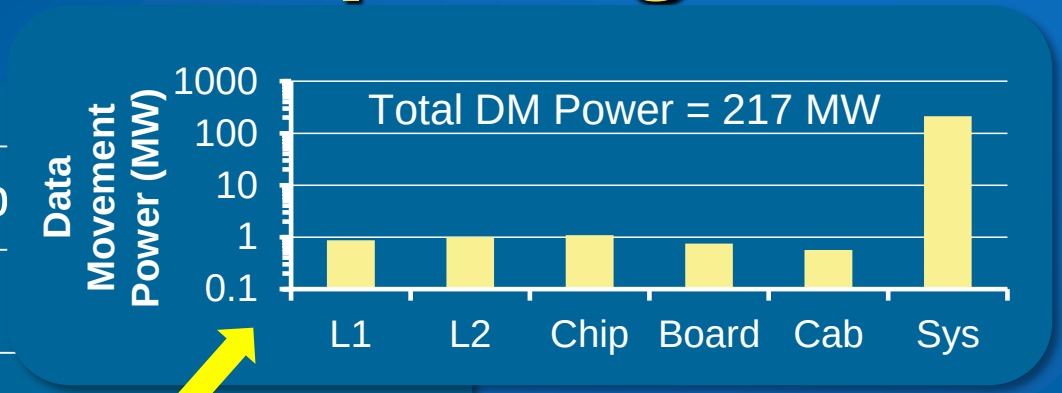
Straw-man System Interconnect

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



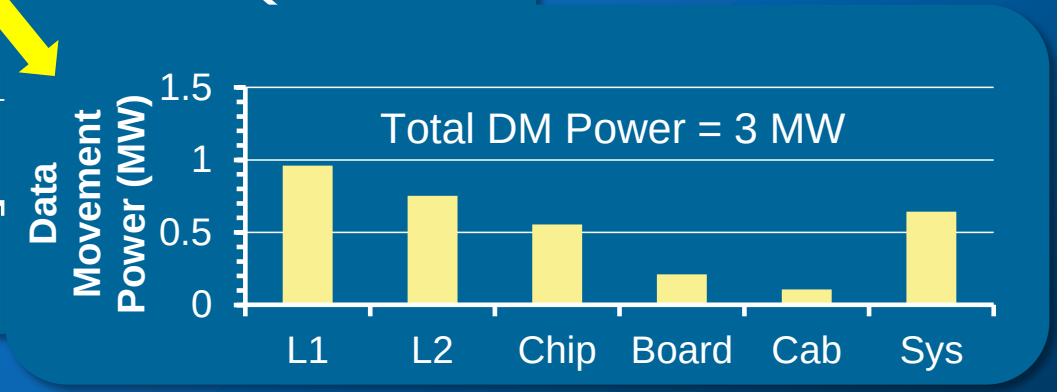
Bandwidth Tapering





Naïve, 4X

Severe



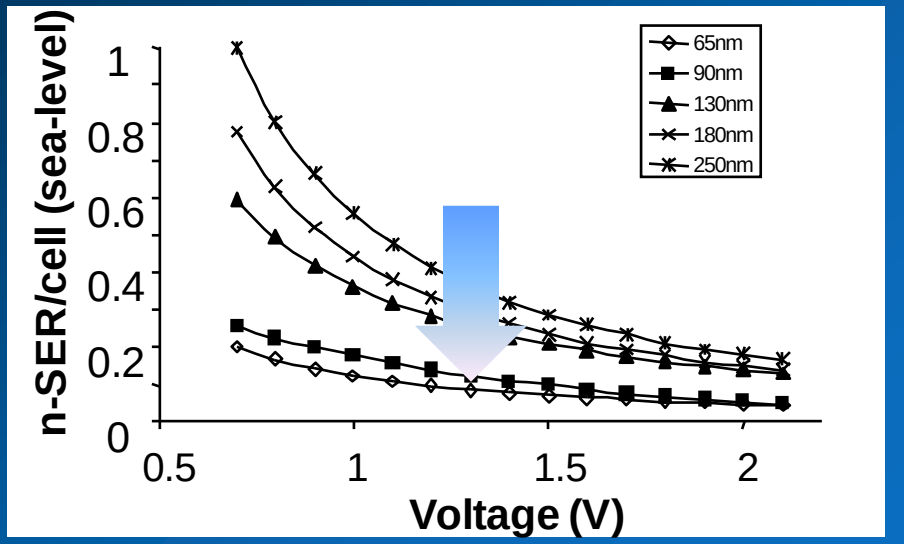
Intelligent BW tapering is necessary

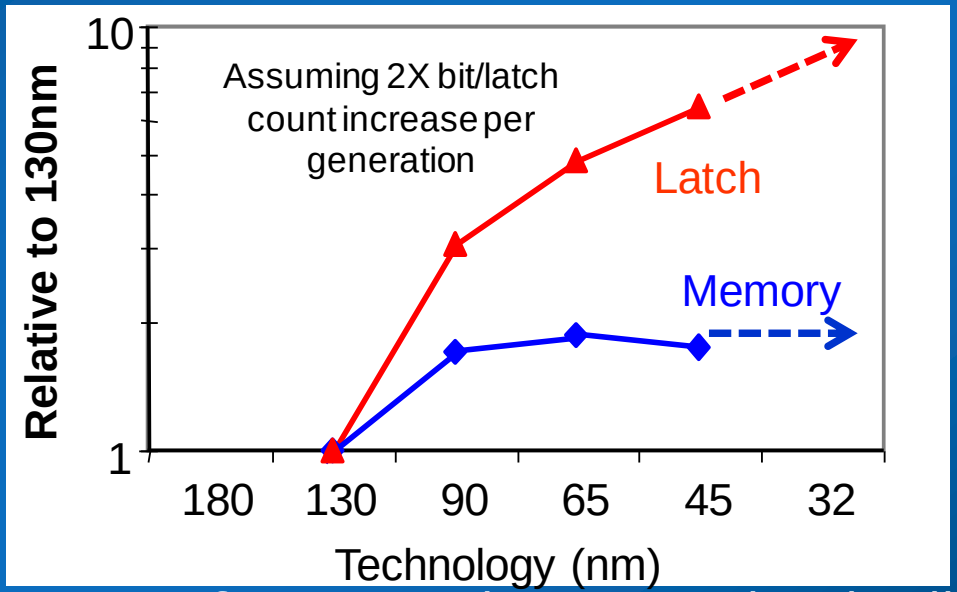
Road to Unreliability?

From Peta to Exa	Reliability Issues
1,000X parallelism	More hardware for something to go wrong >1,000X intermittent faults due to soft errors
Aggressive Vcc scaling to reduce power/energy	Gradual faults due to increased variations More susceptible to Vcc droops (noise) More susceptible to dynamic temp variations Exacerbates intermittent faults—soft errors
Deeply scaled technologies	Aging related faults Lack of burn-in? Variability increases dramatically

Resiliency will be the corner-stone

Soft Errors and Reliability





Soft error/bit reduces each generation
Nominal impact of NTV on soft error rate

Soft error at the system level will continue to increase

Positive impact of NTV on reliability

Low V ➔ lower E fields, low power ➔ lower temperature

Device aging effects will be less of a concern

Lower electromigration related defects

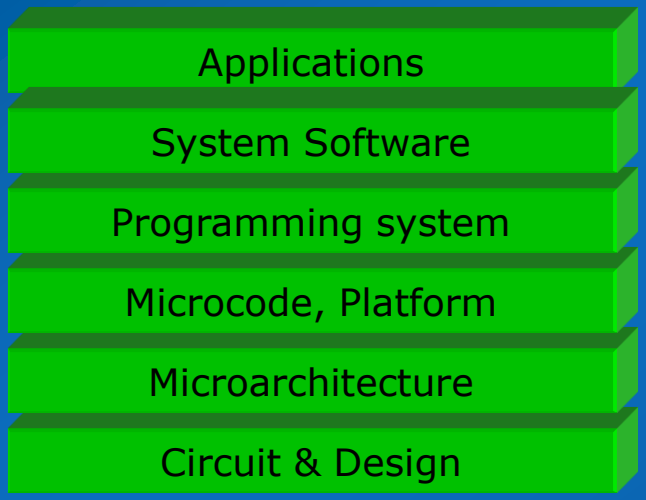
N. Seifert et al, "Radiation-Induced Soft Error Rates of Advanced CMOS Bulk Devices", 44th Reliability Physics Symposium, 2006

Resiliency

Faults	Example
Permanent faults	Stuck-at 0 & 1
Gradual faults	Variability Temperature
Intermittent faults	Soft errors Voltage droops
Aging faults	Degradation

Faults cause errors (data & control)	
Datapath errors	Detected by parity/ECC
Silent data corruption	Need HW hooks
Control errors	Control lost (Blue screen)

Minimal overhead for resiliency



Error detection
Fault isolation
Fault confinement
Reconfiguration
Recovery & Adapt

Architecture needs a Paradigm Shift

Architect's past and present priorities—

Single thread performance	Frequency
Programming productivity	Legacy, compatibility Architecture features for productivity
Constraints	(1) Cost (2) Reasonable Power/Energy

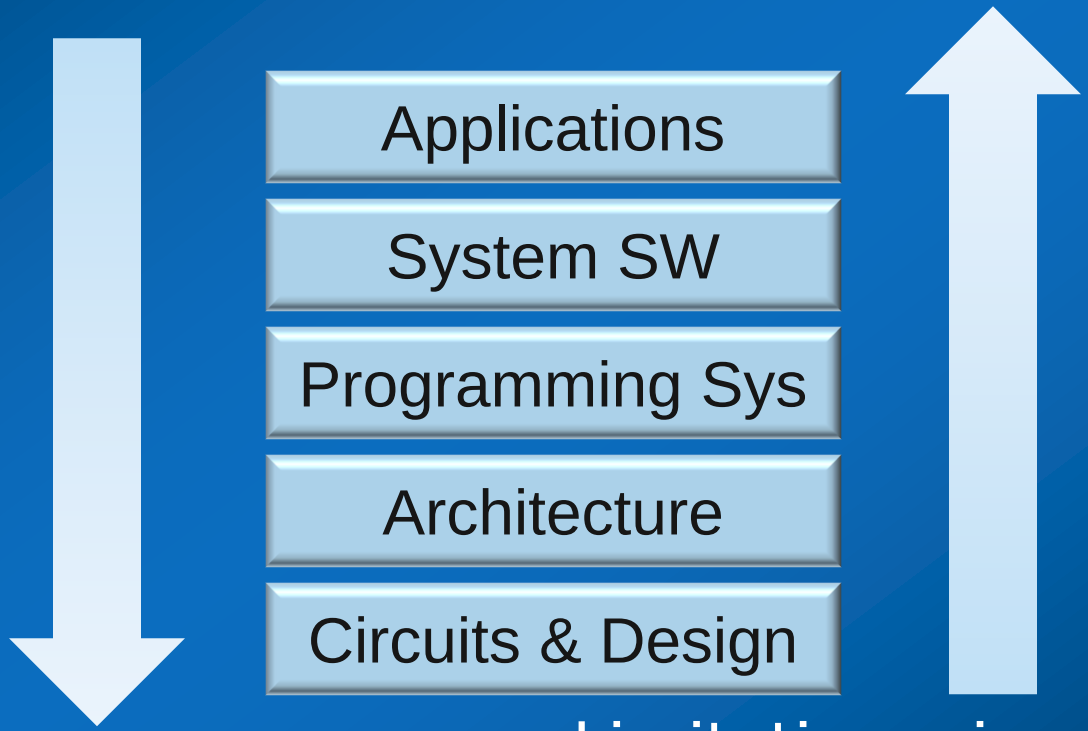
Architect's future priorities should be—

Throughput performance	Parallelism, application specific HW
Power/Energy	Architecture features for energy Simplicity
Constraints	(1) Programming productivity (2) Cost

Must revisit and evaluate each (even legacy) architecture feature

HW-SW Co-design

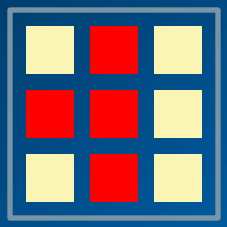
Applications and SW stack provide guidance for efficient system design



Limitations, issues and opportunities to exploit

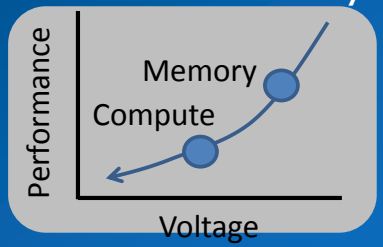
Bottom-up Guidance

1. NTV reduces energy but exacerbates variations



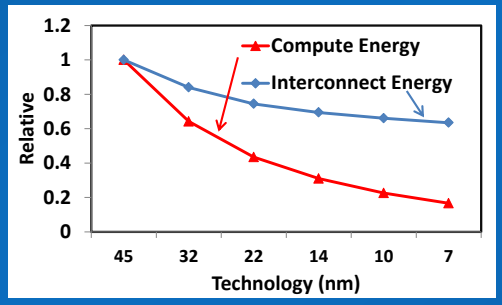
Small & Fast cores
Random distribution
Temp dependent

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



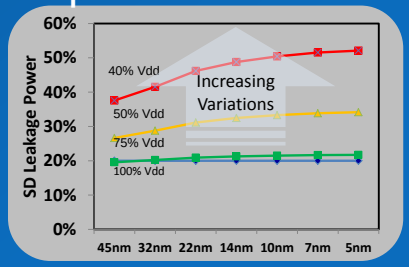
Disproportionate
Memory arrays
can be made
larger

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



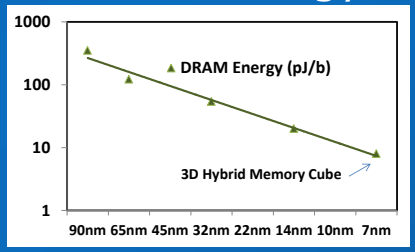
6X compute
1.6X interconnect

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



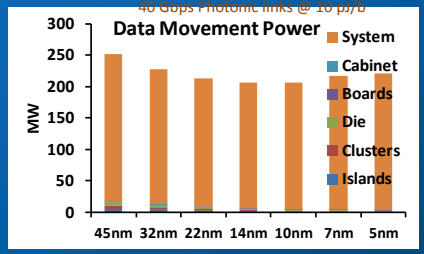
Expect 50%
leakage
Idle hardware
consumes energy

5. DRAM energy scales, but not enough



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

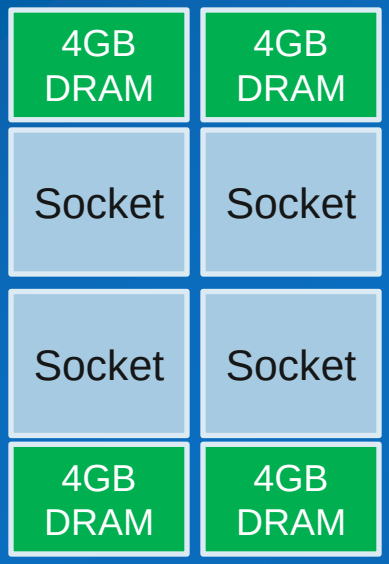
6. System interconnect limited by laser energy and cost



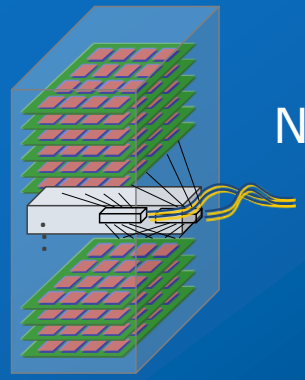
BW tapering and
locality awareness
necessary

Today's HW System Architecture

384 GF/s Peak
260 pJ/F ➲ 260 MW/Exa
55 μB of local memory/F



Coherent domain
1.5 TF Peak
660 pJ/F ➲ 660 MW/Exa
10 mB of DRAM/F

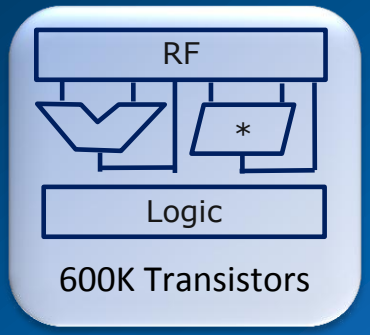


Non-coherent domain

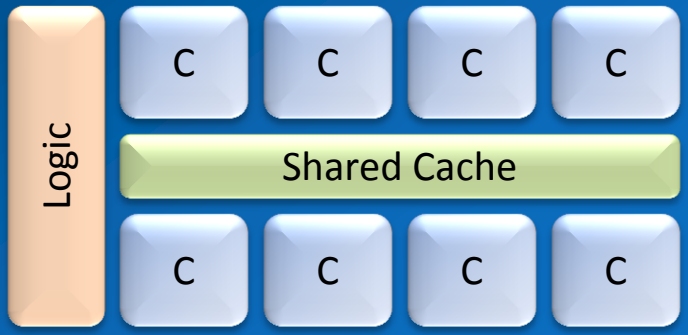
Today's programming model comprehends this system architecture

Straw-man Exascale Processor

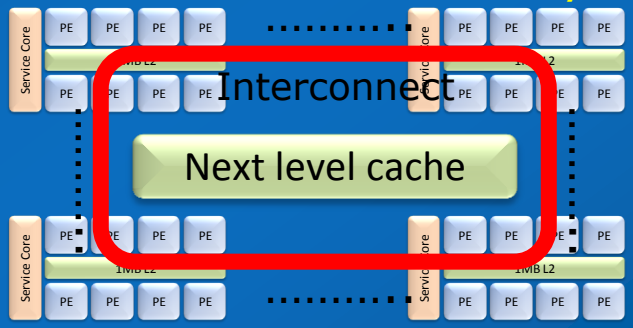
Simplest Core



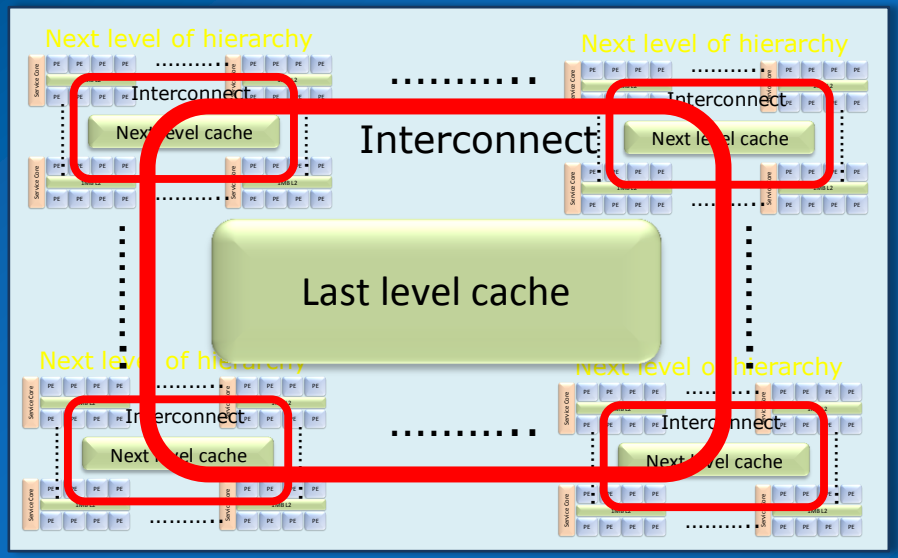
First level of hierarchy



Next level of hierarchy



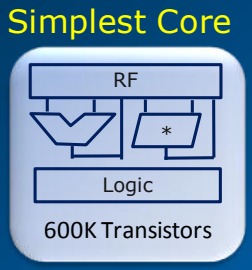
Processor

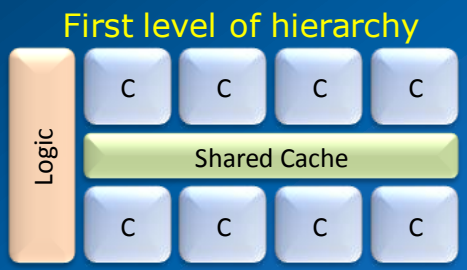


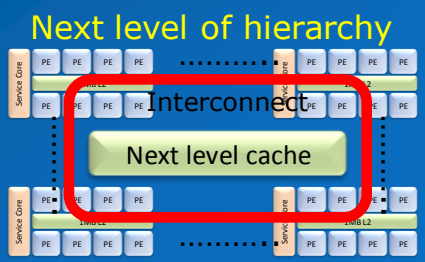
Technology	7nm, 2018
Die area	500 mm2
Cores	2048
Frequency	4.2 GHz
TFLOPs	17.2
Power	600 Watts
E Efficiency	34 pJ/Flop

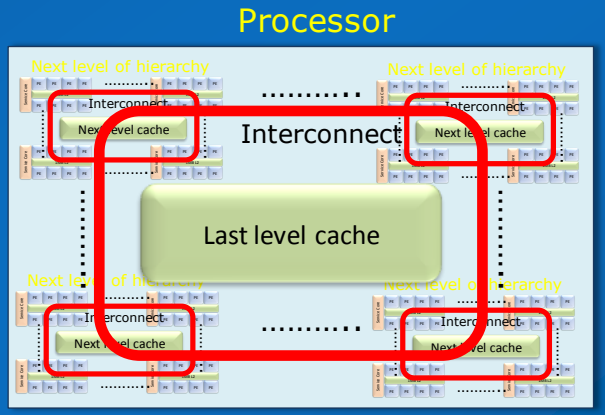
Computations alone consume 34 MW for Exascale

Straw-man Architecture at NTV









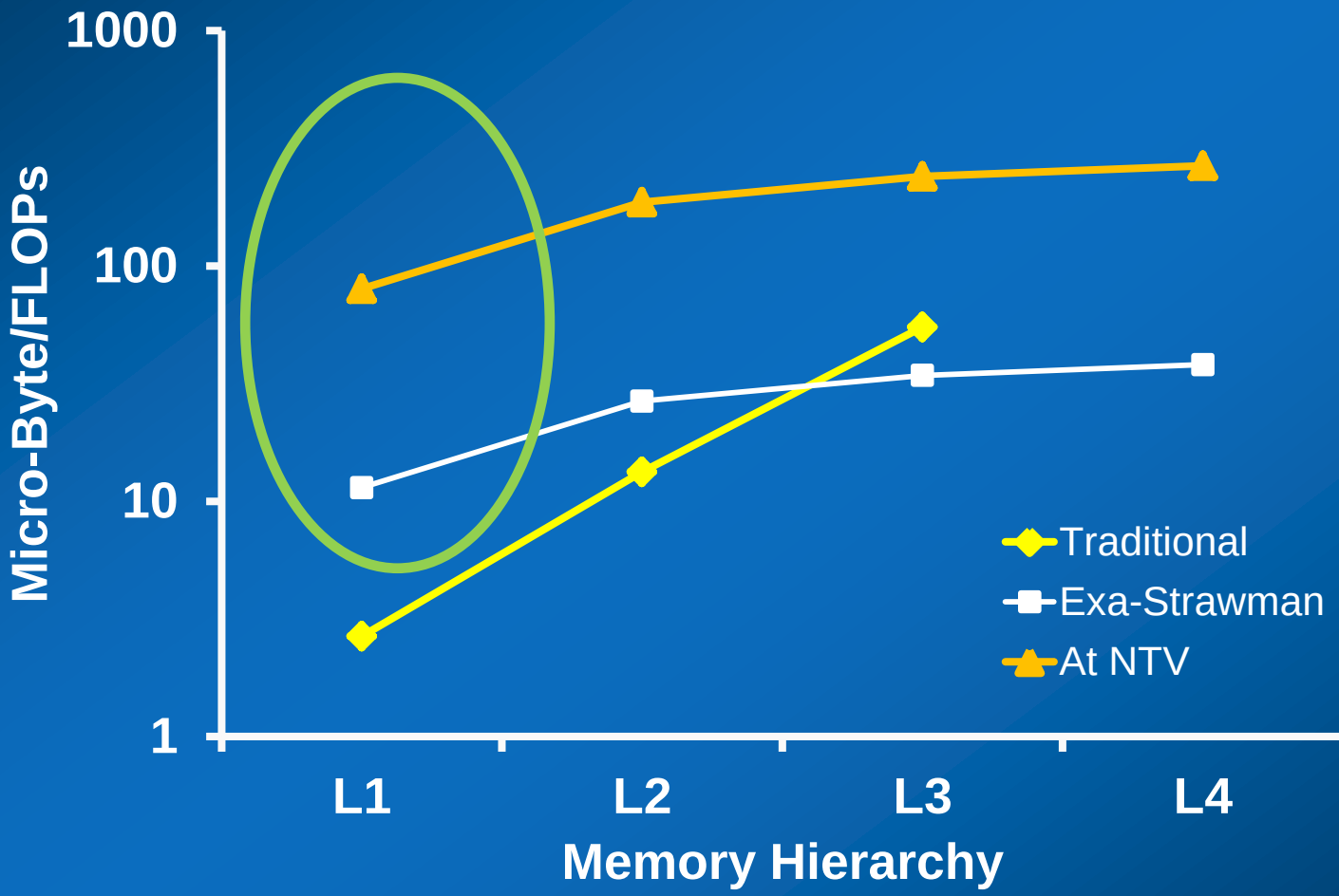
	Full Vdd	50% Vdd
Technology	7nm, 2018	
Die area	500 mm2	
Cores	2048	
Frequency	4.2 GHz	600 MHz
TFLOPs	17.2	2.5
Power	600 Watts	37 Watts
E Efficiency	34 pJ/Flop	15 pJ/Flop

Reduced frequency and flops

Reduced power and improved E-efficiency

Compute energy efficiency close to Exascale goal

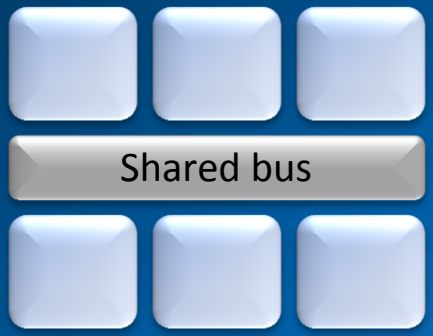
Local Memory Capacity



Higher local memory capacity promotes data locality

Interconnect Structures

Buses over short distance

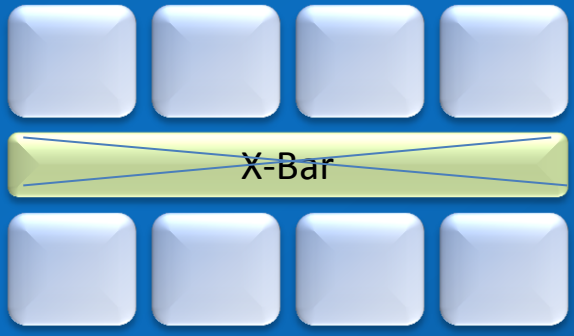


1 to 10 fJ/bit
0 to 5mm
Limited scalability

Shared memory

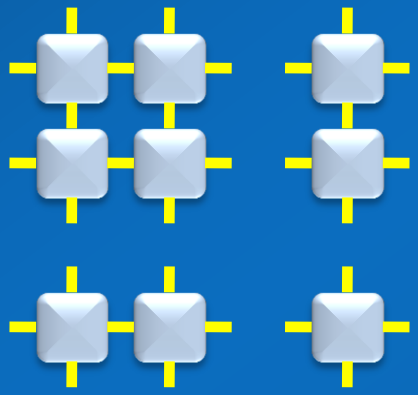
10 to 100 fJ/bit
1 to 5mm
Limited scalability

Cross Bar Switch



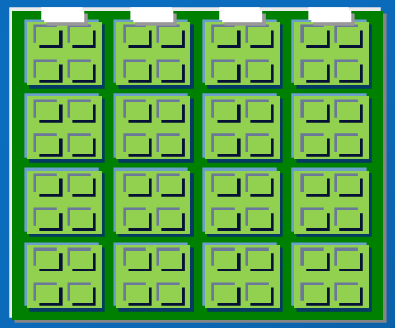
0.1 to 1pJ/bit
2 to 10mm
Moderate scalability

Packet Switched Network

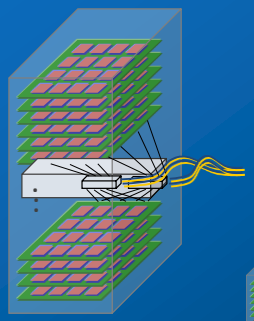


1 to 3pJ/bit
>5 mm, scalable

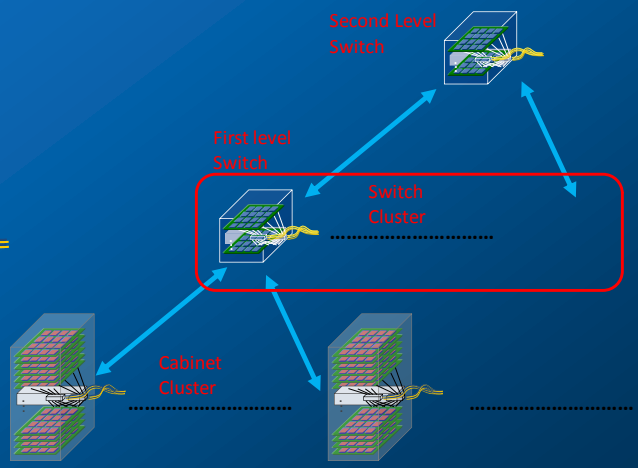
Board



Cabinet



System



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-provisioning, Introspection, Self-awareness

Addressing variations

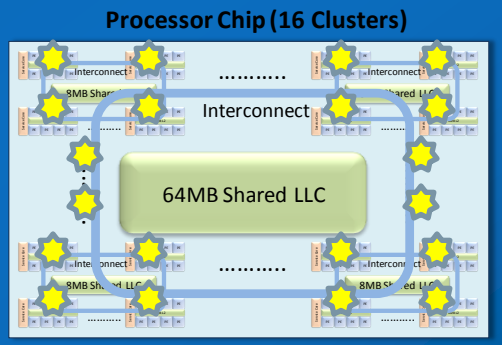
1. Provide more compute HW
2. Law of large numbers
3. Static profile

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

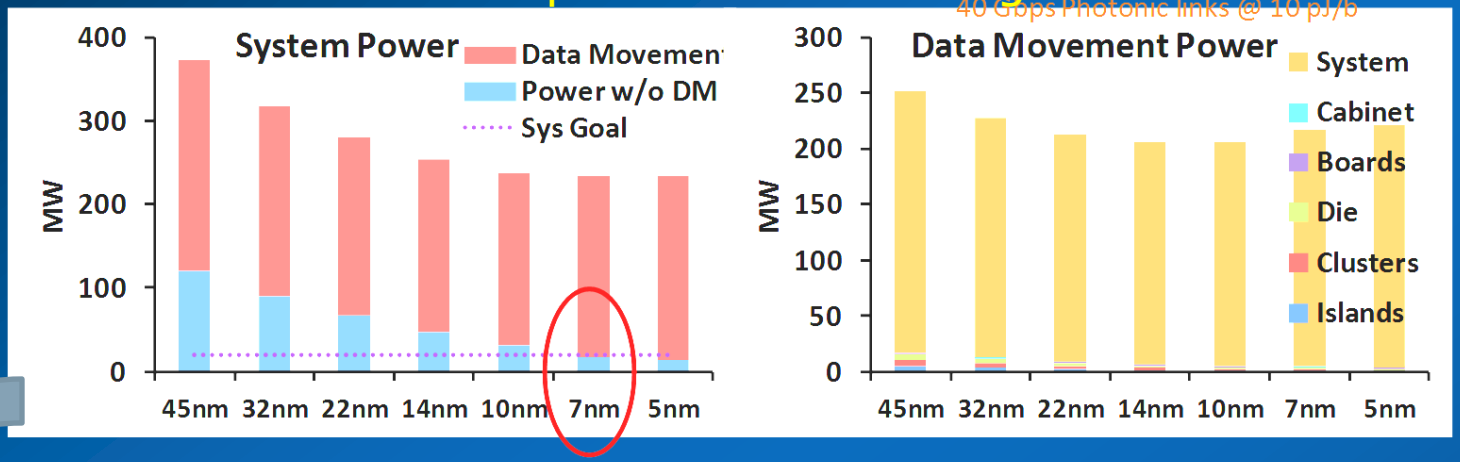
2. Dynamically control and manage resources

3. Identify sensors, functions in HW for implementation

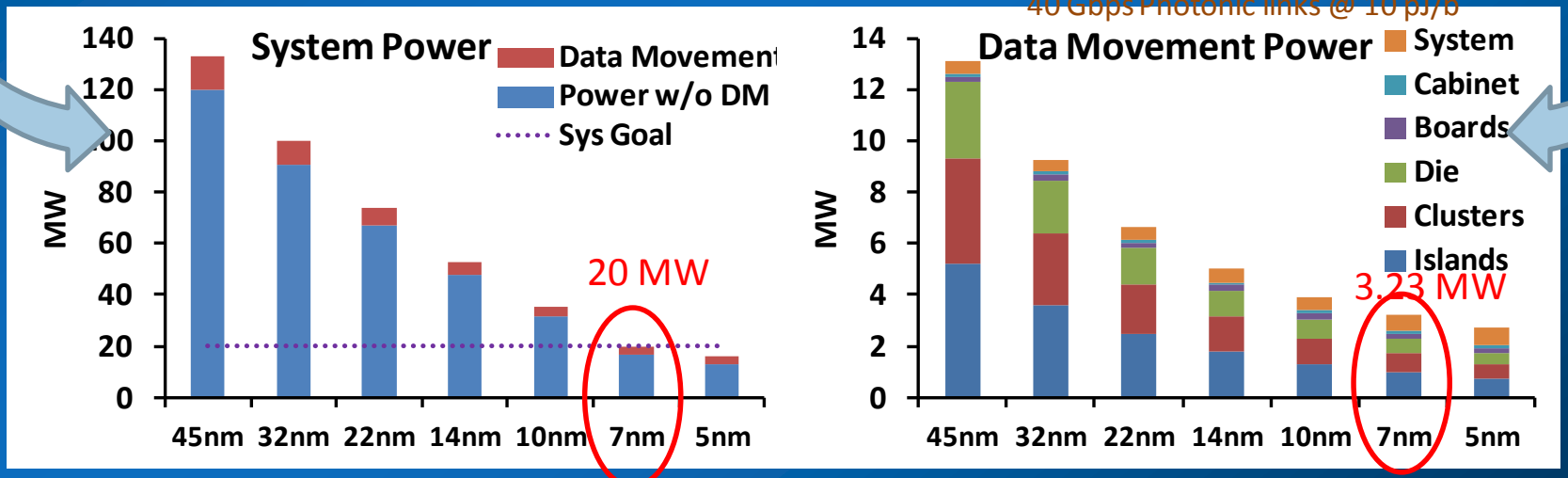
System SW implements introspective execution model

Over-provisioned Introspectively Resource Managed System

Over-provisioned in design



Dynamically tuned for the given objective



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 to make Exascale real