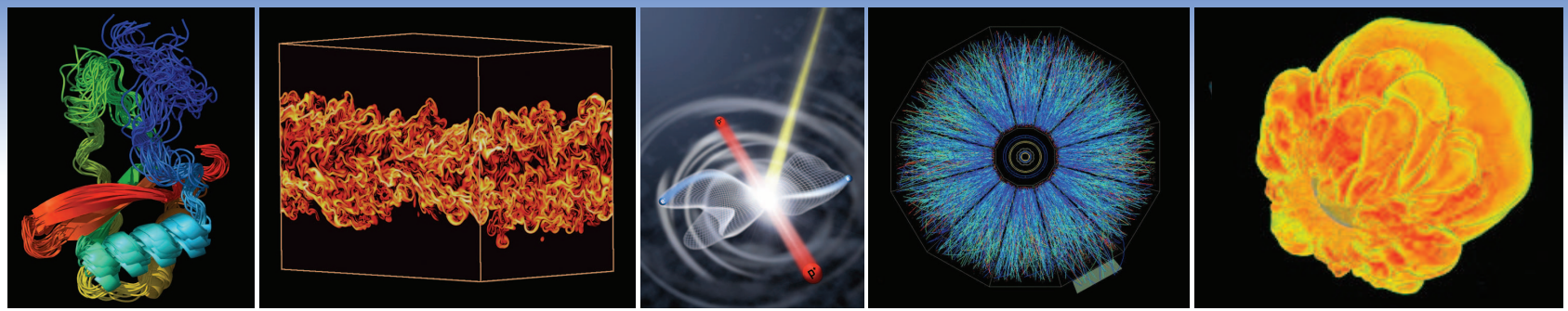


Challenges of Energy Efficient Scientific Computing

John Shalf

National Energy Research Supercomputing Center

Lawrence Berkeley National Laboratory

1st Symposium on Energy Efficient Electronics

Berkeley, June 12, 2009

Office of Science U.S. DEPARTMENT OF ENERGY

NERSC

BERKELEY LAB

NERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Part I

Power Crisis in HPC

Office of Science
U.S. DEPARTMENT OF ENERGY

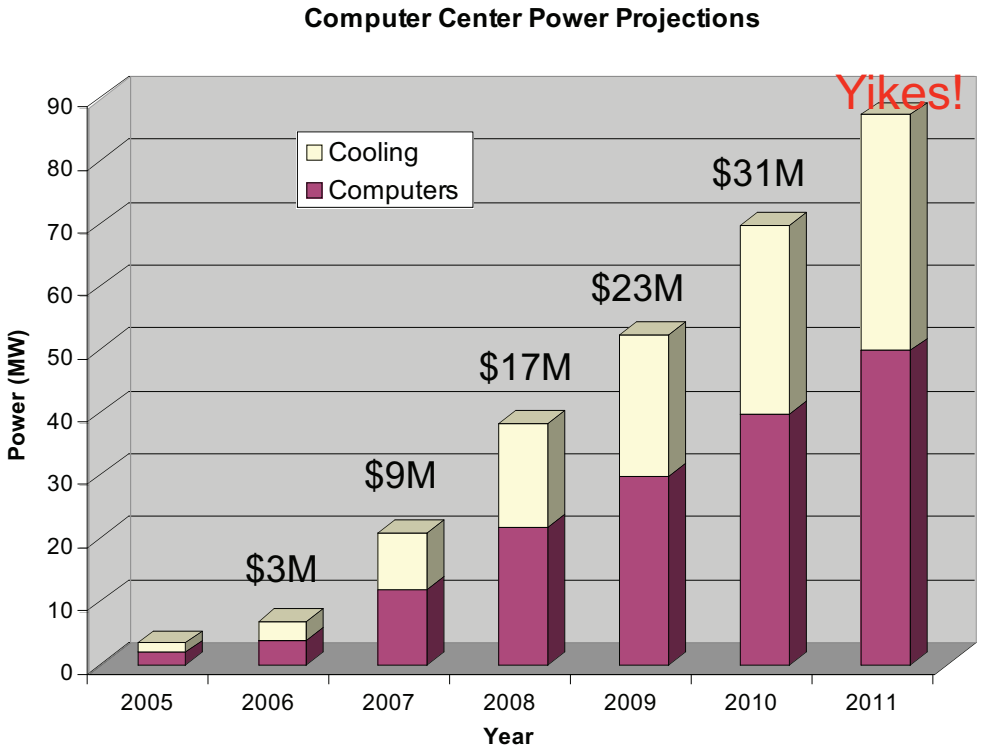
BERKELEY LAB

New Design Constraint: *POWER*

- **Transistors still getting smaller**
  - Moore's Law is alive and well
- **But Dennard scaling is dead!**
  - No power efficiency improvements with smaller transistors
  - No clock frequency scaling with smaller transistors
  - All "magical improvement of silicon goodness" has ended
- **Cannot continue with business as usual**
  - DARPA study extrapolated current design trends and found brick wall at end of exponential curves
  - Can only accelerate existing research prototypes (not "magic" new disruptive technology)!

ORNL Computing Power and Cooling 2006 - 2011

- **Immediate need to add 8 MW to prepare for 2007 installs of new systems**
- **NLCF petascale system could require an additional 10 MW by 2008**
- **Need total of 40-50 MW for projected systems by 2011**
- **Numbers just for computers: add 75% for cooling**
- **Cooling will require 12,000 – 15,000 tons of chiller capacity**



Cost estimates based on $0.05 kW/hr

Annual Average Electrical Power Rates $/MWh

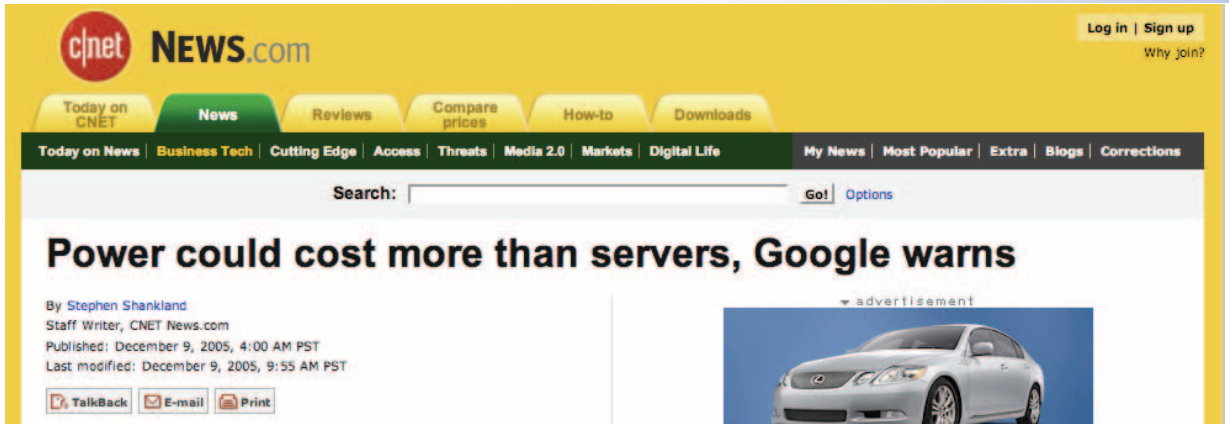
Site	FY 2005	FY 2006	FY 2007	FY 2008	FY 2009	FY 2010	
LBNL	43.70	50.23	53.43	57.51	58.20	56.40	*
ANL	44.92	53.01					
ORNL	46.34	51.33					
PNNL	49.82	N/A					

Data taken from Energy Management System-4 (EMS4). EMS4 is the DOE corporate system for collecting energy information from the sites. EMS4 is a web-based system that collects energy consumption and cost information for all energy sources used at each DOE site. Information is entered into EMS4 by the site and reviewed at Headquarters for accuracy.

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Power is an Industry Wide Problem

(2% of US power consumption and growing)



The New York Times

"Hiding in Plain Sight, Google Seeks More Power", by John Markoff, June 14, 2006

New Google Plant in The Dulles, Oregon, from NYT, June 14, 2006

Relocate to Iceland?

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

HPC Power: It will only get worse

- Recent Baltimore Sun Article on NSA system in Maryland
  - Consuming 75MW and growing up to 15MW/year
  - Not enough power left for city of Baltimore!

- LBNL IJHPCA Study for ~1/5 Exaflop for Climate Science in 2008
  - Extrapolation of Blue Gene and AMD design trends
  - Estimate: **20 MW** for BG and **179 MW** for AMD

- DOE E3 Report
  - Extrapolation of existing design trends to exascale in 2016
  - Estimate: **130 MW**

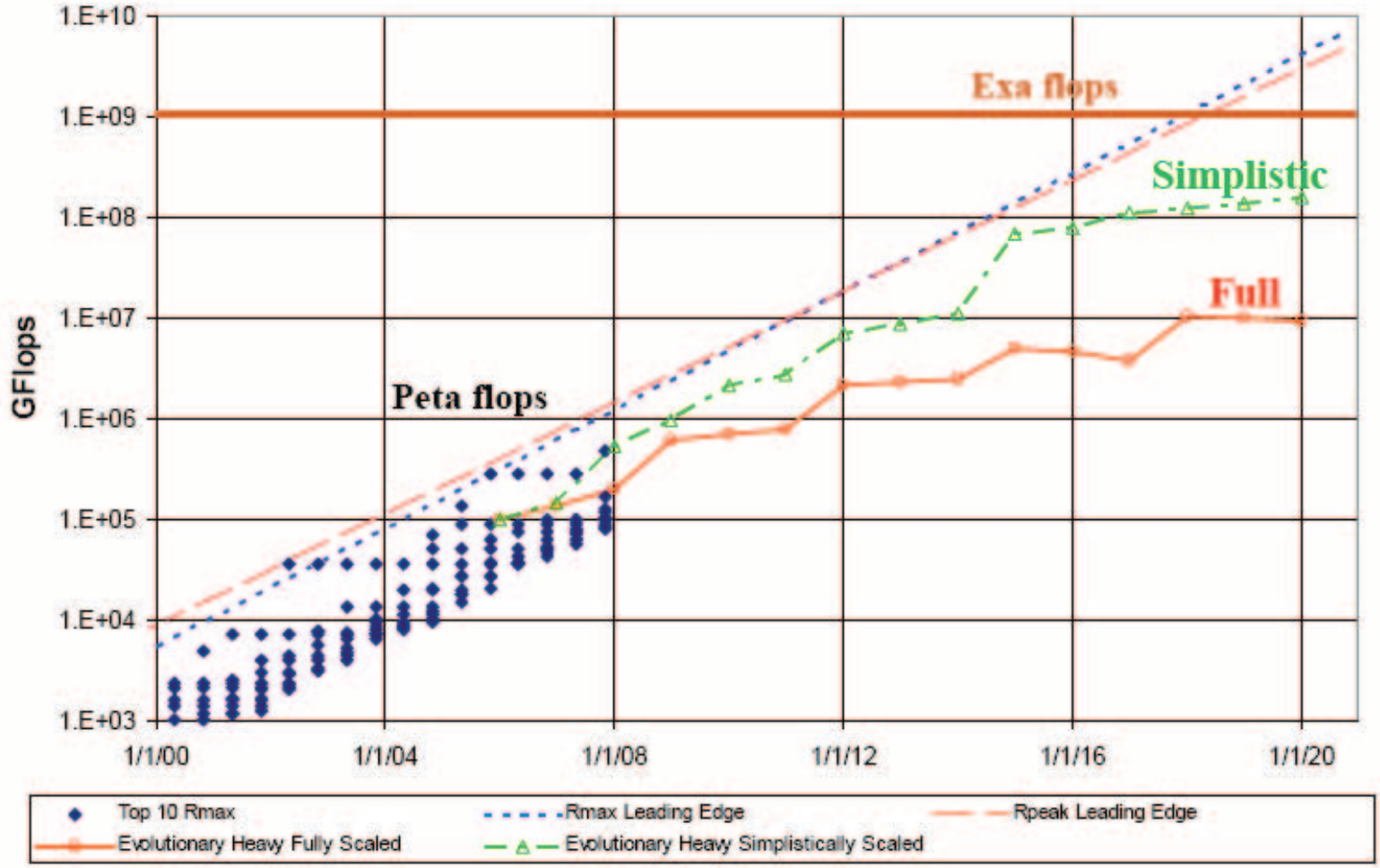
- DARPA Study
  - More detailed assessment of component technologies
  - Estimate: **20 MW** just for memory alone, **60 MW** aggregate extrapolated from current design trends

The current approach is not sustainable!

DARPA Exascale Study

- Commissioned by DARPA to explore the challenges for Exaflop computing
- Two model for future performance growth
  - Simplistic: ITRS roadmap; power for memory grows linear with #of chips; power for interconnect stays constant
  - Fully scaled: same as simplistic, but memory and router power grow with peak flops per chip

We won't reach Exaflops with the current approach



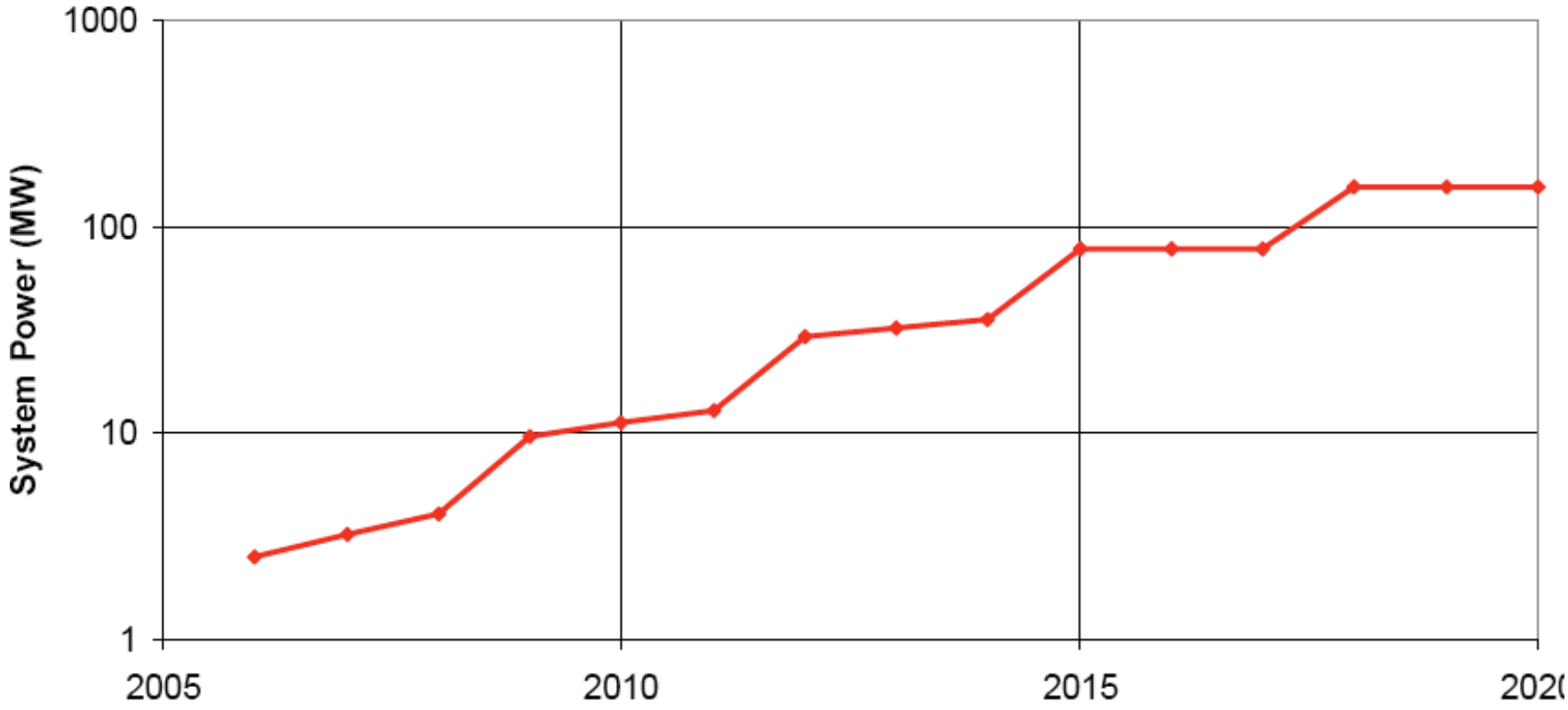
From Peter Kogge, DARPA Exascale Study

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Office of Science U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

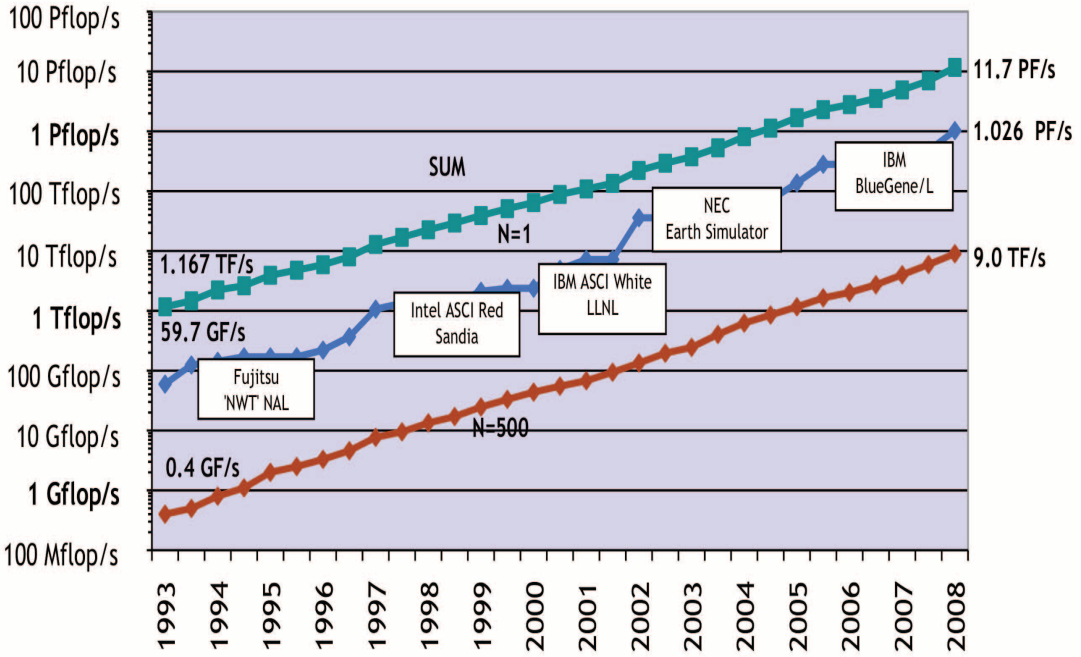
… and the power costs will still be staggering



From Peter Kogge,
DARPA Exascale Study

Primary Design Constraint: *POWER*

- Power Efficiency and clock rates no longer improving at historical rates
- Demand for supercomputing capability is accelerating!
- DOE is targeting an exaflop system for 2016



- Exascale (10^{18} FLOP/s) cannot be built by simply scaling petascale systems
- Power requirements for incremental approach are *profoundly* impractical
- And we have finite $'s for development costs

NERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

The Challenge

Where do we get a 1000x improvement in performance with only a 10x increase in power?

How do you achieve this in 10 years with a finite development budget?

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Where do we get 1000x performance improvement for 10x power?

1. Processor
2. Interconnect
3. Memory
4. Software tuning (auto-tuning)
5. Algorithms
6. Power/Cooling/facilities (ask Bill & Dale)

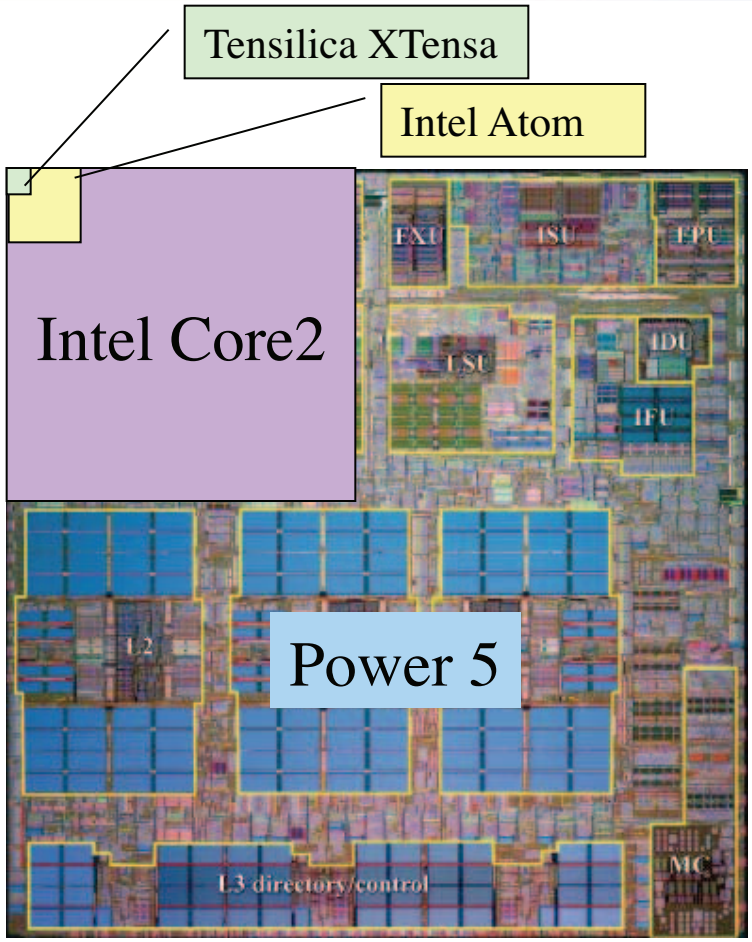
Hardware: What are the problems?

(Lessons from the Berkeley View)

- **Current Hardware/Lithography Constraints**
  - **Power limits leading edge chip designs**
    - Intel Tejas Pentium 4 cancelled due to power issues
  - **Yield on leading edge processes dropping dramatically**
    - IBM quotes yields of 10 – 20% on 8-processor Cell
  - **Design/validation leading edge chip is becoming unmanageable**
    - Verification teams > design teams on leading edge processors

- **Solution: Small Is Beautiful**
  - **Simpler (5- to 9-stage pipelined) CPU cores**
    - Small cores not much slower than large cores
  - **Parallel is energy efficient path to performance:CV^2F**
    - Lower threshold and supply voltages lowers energy per op
  - **Redundant processors can improve chip yield**
    - Cisco Metro 188 CPUs + 4 spares; Sun Niagara sells 6 or 8 CPUs
  - **Small, regular processing elements easier to verify**

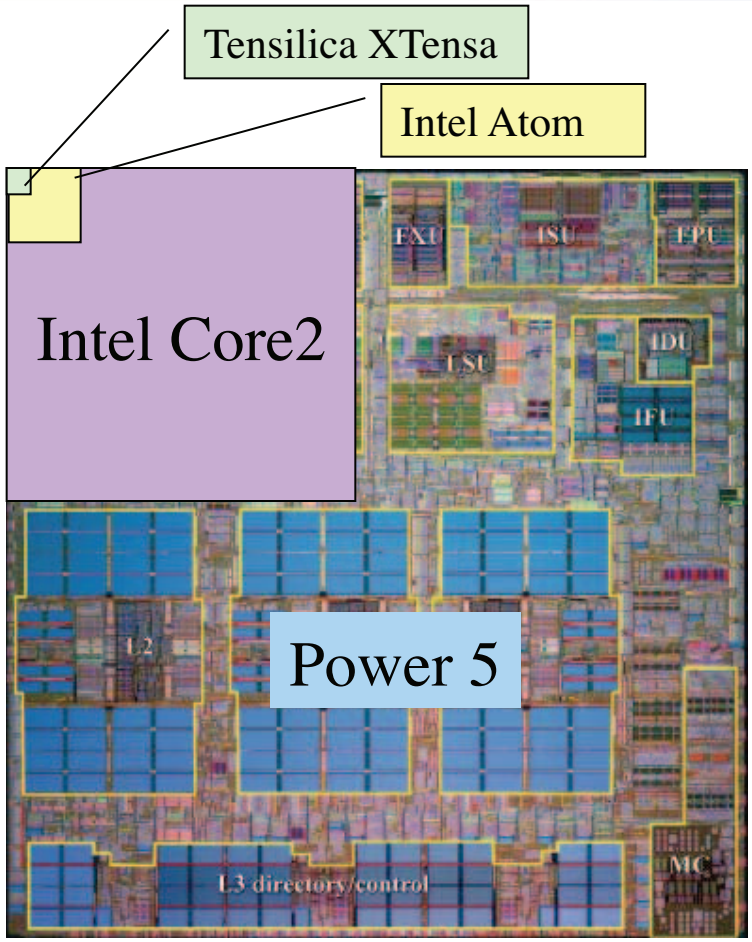
Low-Power Design Principles



- Cubic power improvement with lower clock rate due to V^2F
- Slower clock rates enable use of simpler cores
- Simpler cores use less area (lower leakage) and reduce cost
- Tailor design to application to REDUCE WASTE

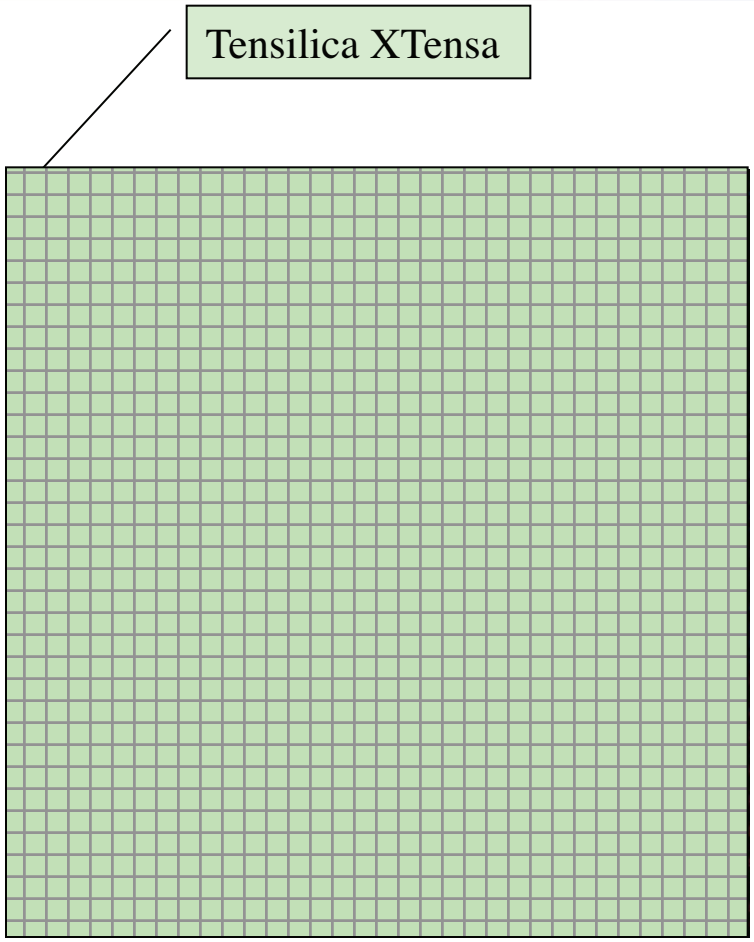
This is how iPhones and MP3 players are designed to maximize battery life and minimize cost

Low-Power Design Principles



- Power5 (server)
  - 120W@1900MHz
  - Baseline
- Intel Core2 sc (laptop) :
  - 15W@1000MHz
  - *4x more FLOPs/watt than baseline*
- Intel Atom (handhelds)
  - 0.625W@800MHz
  - 80x more
- Tensilica XTensa DP (Moto Razor) :
  - 0.09W@600MHz
  - 400x more *(80x-120x sustained)*

Low Power Design Principles



- Power5 (server)
  - 120W@1900MHz
  - Baseline
- Intel Core2 sc (laptop) :
  - 15W@1000MHz
  - *4x more FLOPs/watt than baseline*
- Intel Atom (handhelds)
  - 0.625W@800MHz
  - 80x more
- Tensilica XTensa DP (Moto Razor) :
  - 0.09W@600MHz
  - 400x more (80x-100x sustained)

Even if each simple core is 1/4th as computationally efficient as complex core, you can fit hundreds of them on a single chip and still be 100x more power efficient.

Future HPC Technology Building Blocks

- **Previous Decade**
  - Optimization target: minimize price to buy more hardware
  - COTS: Redirect off-the-shelf components designed for mass market
  - This leveraged "Moore's Law" density improvements

- **Next Decade**
  - Optimization target: minimize power consumed for work performed
  - Specialize and integrate: Embedded + SoC is proven design point
  - This leverages "Bells Law" cost efficiency: *Commodity* not *COTS*

Future HPC Technology Building Blocks

- **Previous Decade**
  - Optimization target: minimize price to buy more hardware
  - COTS: Redirect off-the-shelf components designed for mass market
  - This leveraged "Moore's Law" density improvements

- **Next Decade**
  - Optimization target: minimize power consumed for work performed
  - Specialize and integrate: Embedded + SoC is proven design point
  - This leverages "Bells Law" cost efficiency: *Commodity* not *COTS*

- **Interim solution: Accelerators**
  - Demonstrate huge efficiency potential of manycore
  - Demonstrate we have failed to learn from CM5 (PCIe)
  - Stepping stone to convergence (merge manycore with host memory)
  - But also points to benefits of some specialization

Conclusion

- **Future HPC must move to simpler power-efficient core designs**
  - **Embedded/consumer electronics technology is central to the future of HPC**
  - **Convergence inevitable because it optimizes both cost and power efficiency**

Interconnects

Interconnects: Leading Issues

- Cannot continue to scale fully-connected interconnect topologies
- Cannot continue to scale bandwidth using electrical networks

What technology be applied to address these constraints?

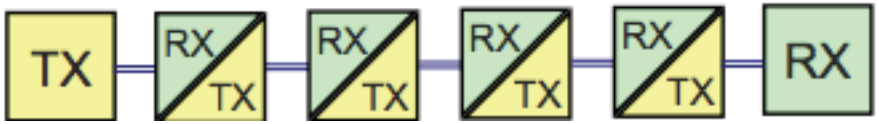
The problem with Wires:

Energy to move data proportional to distance

- **Wire cost to move a bit:**
  - **energy = bitrate * Length^2 / cross-section area**
  - **On-Chip (1cm): ~1pJ/bit, 100Tb/s**
  - **On-Module (5cm): ~2-5pJ/bit, 10Tb/s**
  - **On-Board (20cm): ~10pJ/bit, 1Tb/s**
  - **Intra-rack (1m): ~10-15pJ/bit, 1Tb/s**
  - **Inter-cabinet(2-50m): 15-30pJ/bit, 5-10Tb/s aggregate**
- **To move a bit with optics: target ~1-2pJ/bit for all distance scales(but initial cost high)**

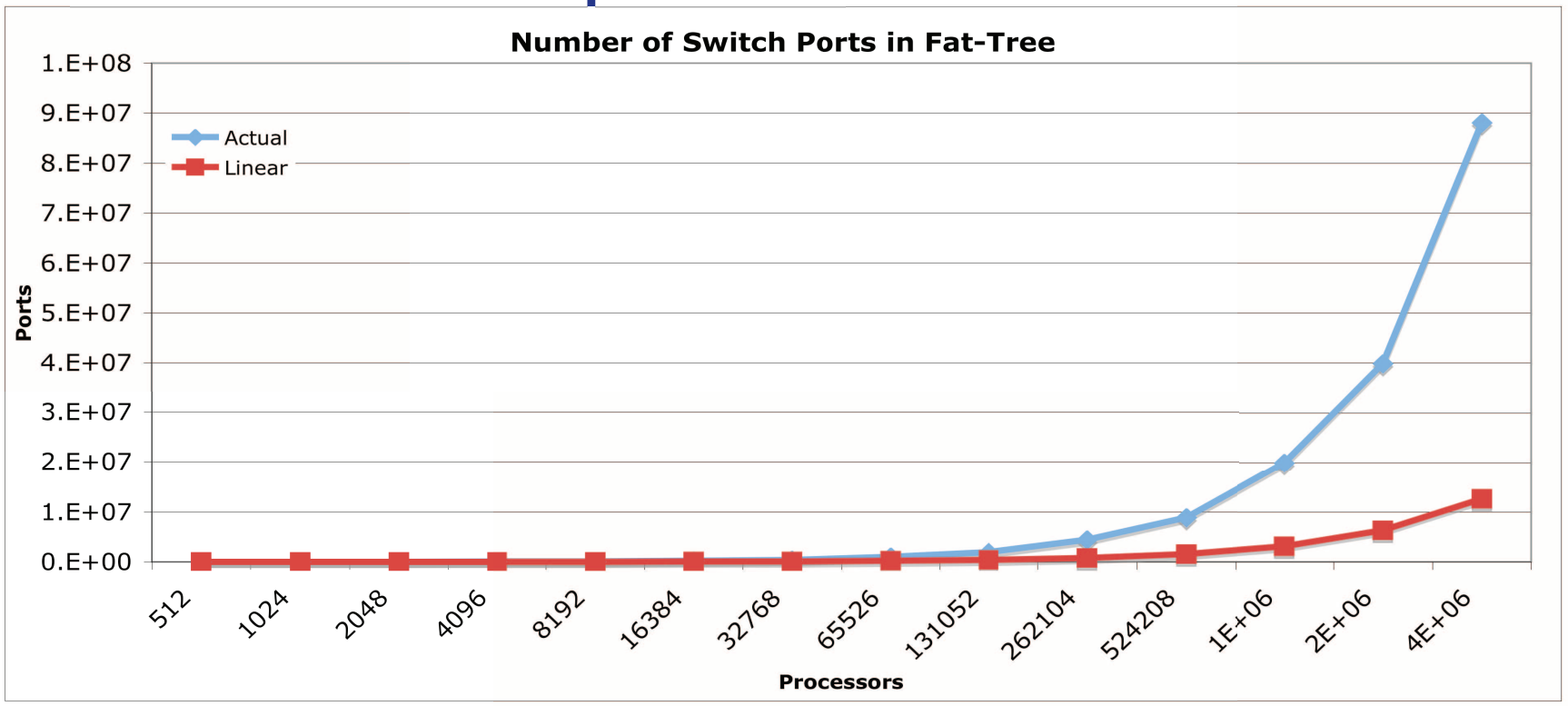
Photonics requires no redrive and passive switch little power

Copper requires to signal amplification even for on-chip connections



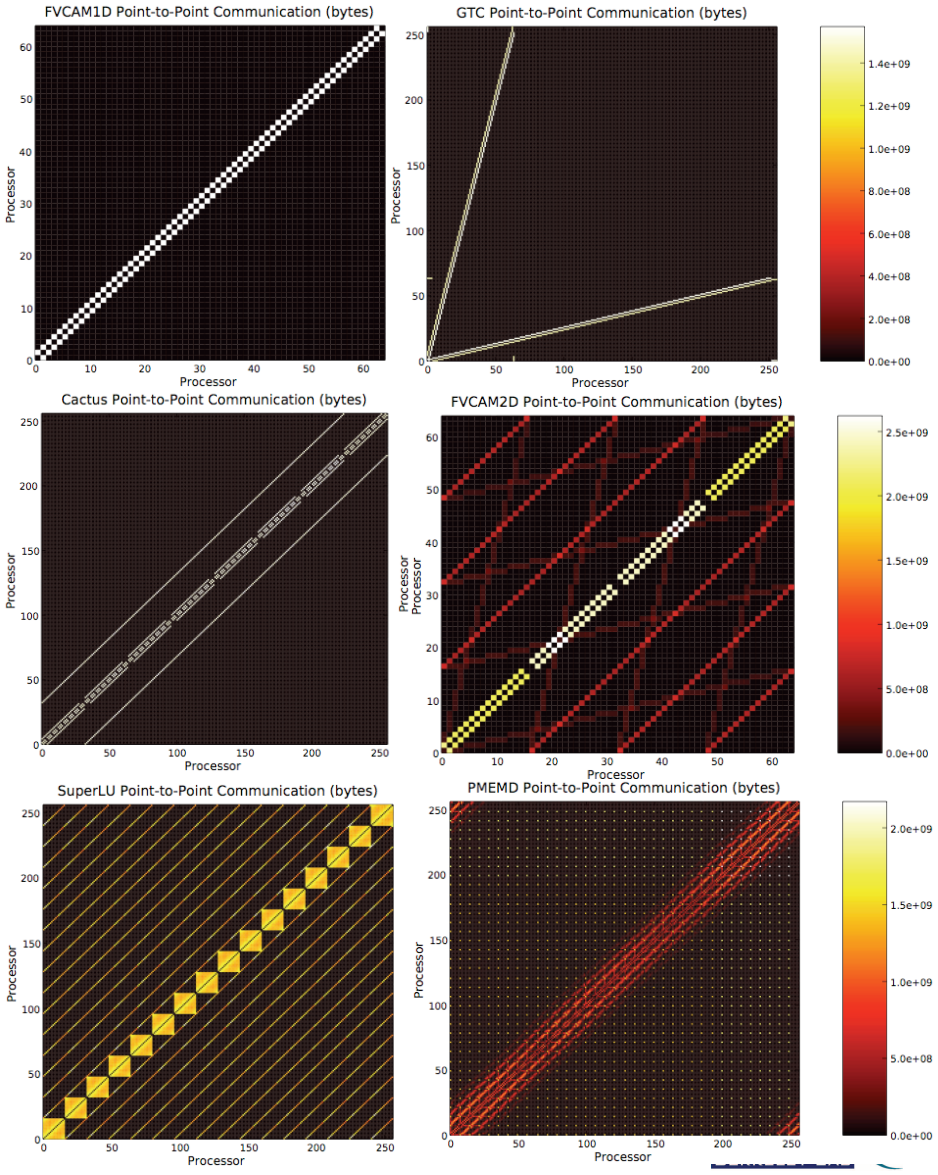
Interconnect Cost
(Scalable Topologies)

- **Fully-connected networks scale superlinearly in cost, but perform the best**
- **Limited-connectivity networks scale linearly in cost, but introduce new problems**

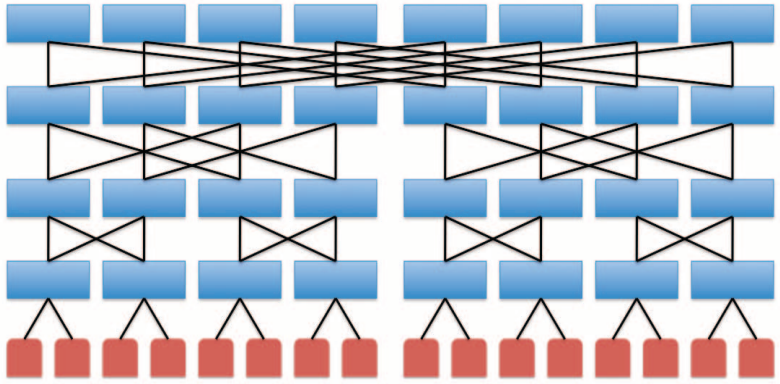


Interconnect Design Considerations for Message Passing Applications

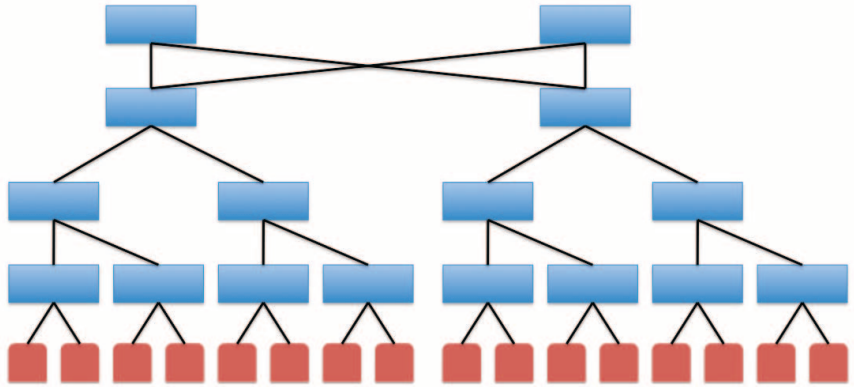
- Application studies provide insight to requirements for Interconnects (both on-chip and off-chip)
  - On-chip interconnect is 2D planar (crossbar won't scale!)
  - Sparse connectivity for most apps.; crossbar is overkill
  - No single best topology
  - Most point-to-point message exhibit sparse topology + often bandwidth bound
  - Collectives tiny and primarily latency bound
- Ultimately, need to be aware of the on-chip interconnect topology in addition to the off-chip topology
  - Adaptive topology interconnects (HFAST)
  - Intelligent task migration?



Using Optical Circuit Switches to Make Fat-Trees into Fit-Trees



A 2-ary 4-tree with 16 nodes.

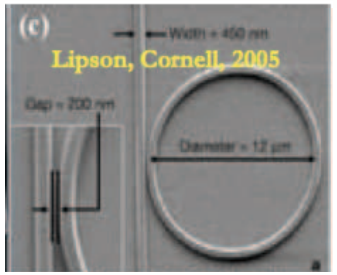


A $(2, 2, 4)$-TL fit-tree with 16 nodes.

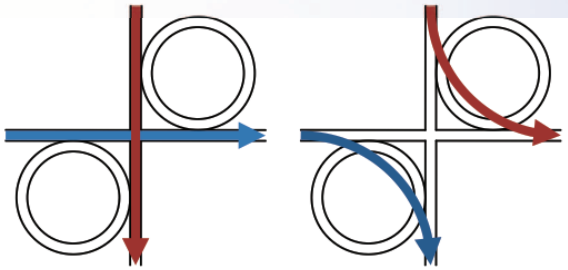
- A Fit-tree uses OCS to prune unused (or infrequently used) connections in a Fat-Tree
- Tailor the interconnect to match application data flows

Silicon Photonics for Energy-Efficient Communication

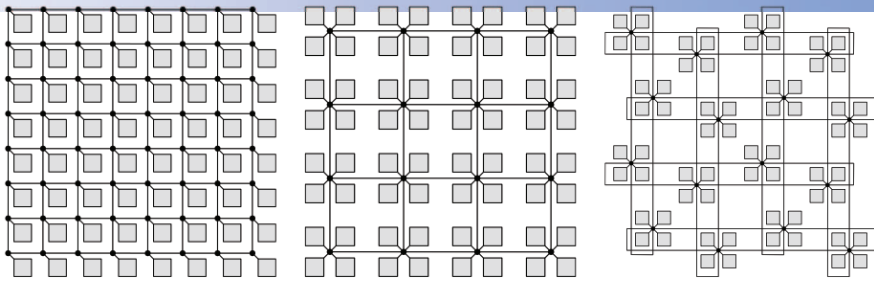
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER



Silicon Photonic Ring Resonator

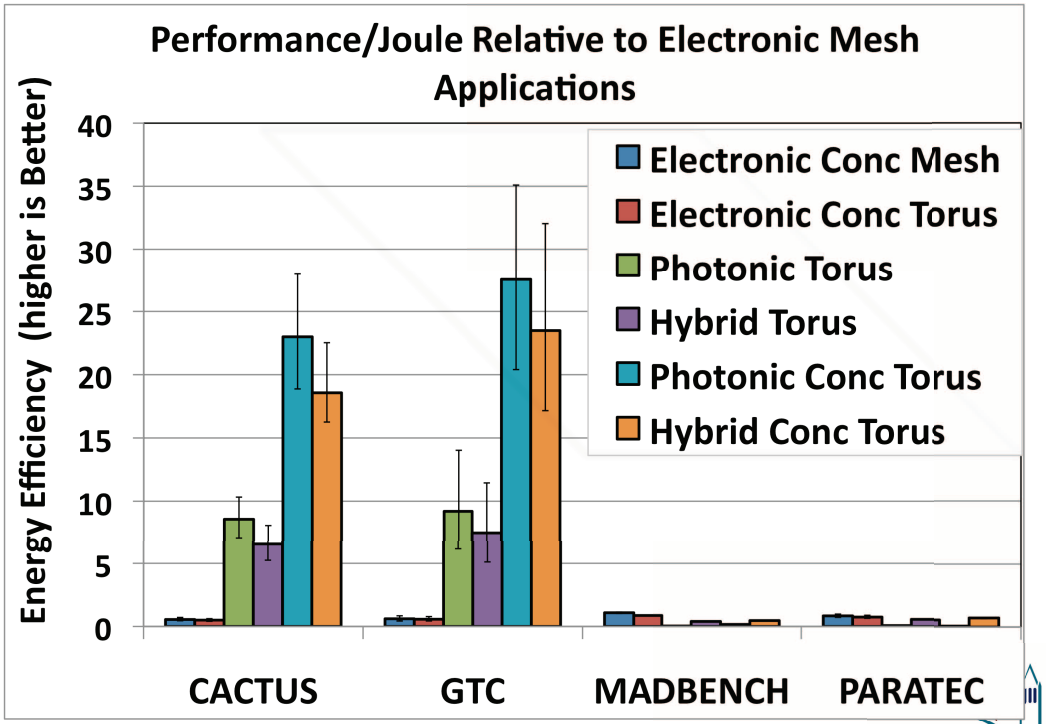


(a) Off state (b) On state



(a) Mesh (b) Concentrated Mesh (c) Concentrated Torus

- **Silicon photonics enables optics to be integrated with conventional CMOS**
- **Enables up to 27x improvement in communication energy efficiency!**



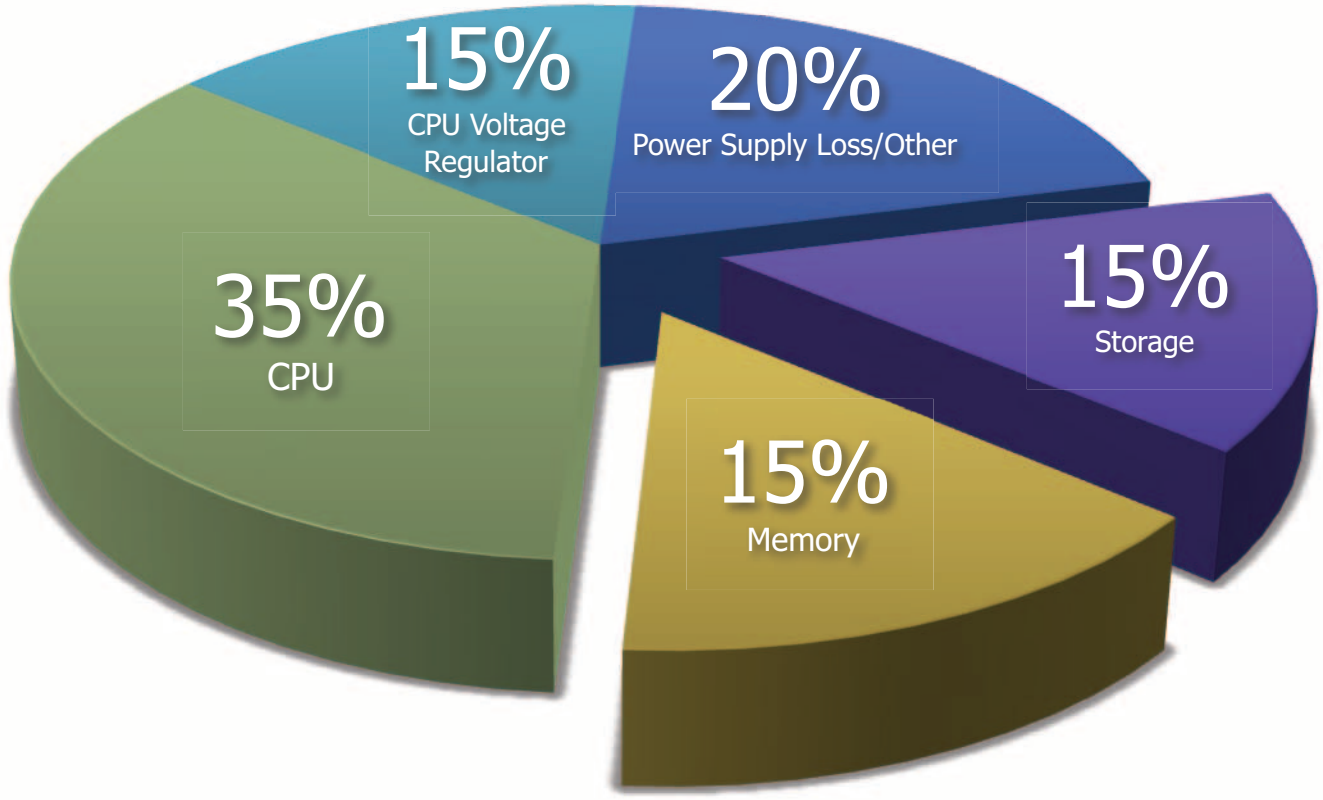
Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Memory

Servers: Recognizing Memory Power Consumption

According to the Environmental Protection Agency (EPA), data centers consumed about 60 billion kilowatt-hours (kWh) in 2006, roughly 1.5 percent of total U.S. electricity consumption.

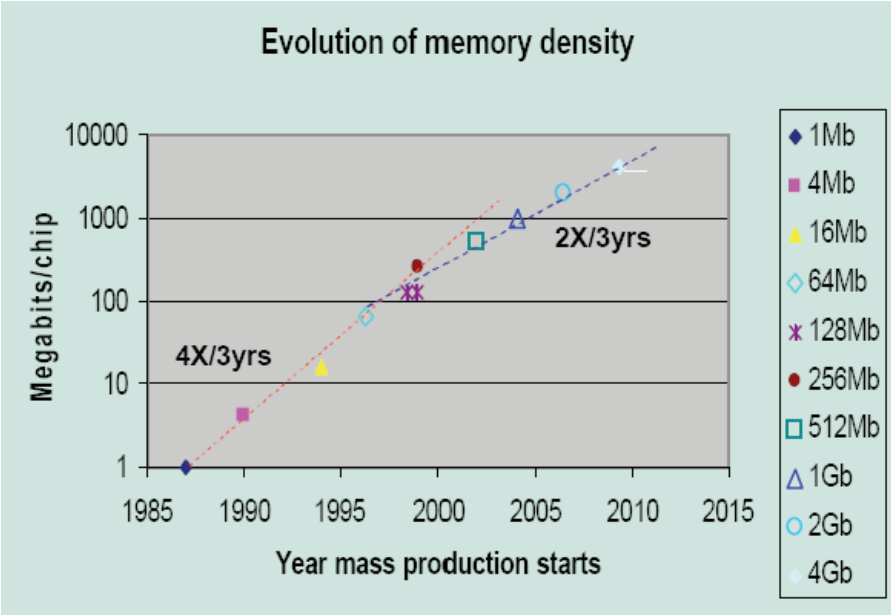


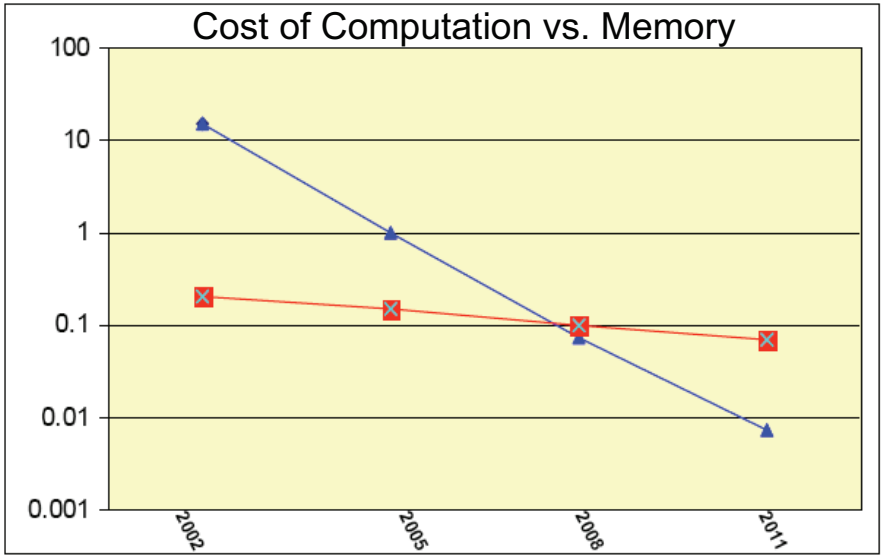
Slide from Dean Klein (Micron Technology)

Technology Challenge

Our ability to sense, collect, generate and calculate on data is growing faster than our ability to access, manage and even "store" that data

- Memory density is doubling every three years; processor logic is every two
- Storage costs (dollars/Mbyte) are dropping gradually compared to logic costs



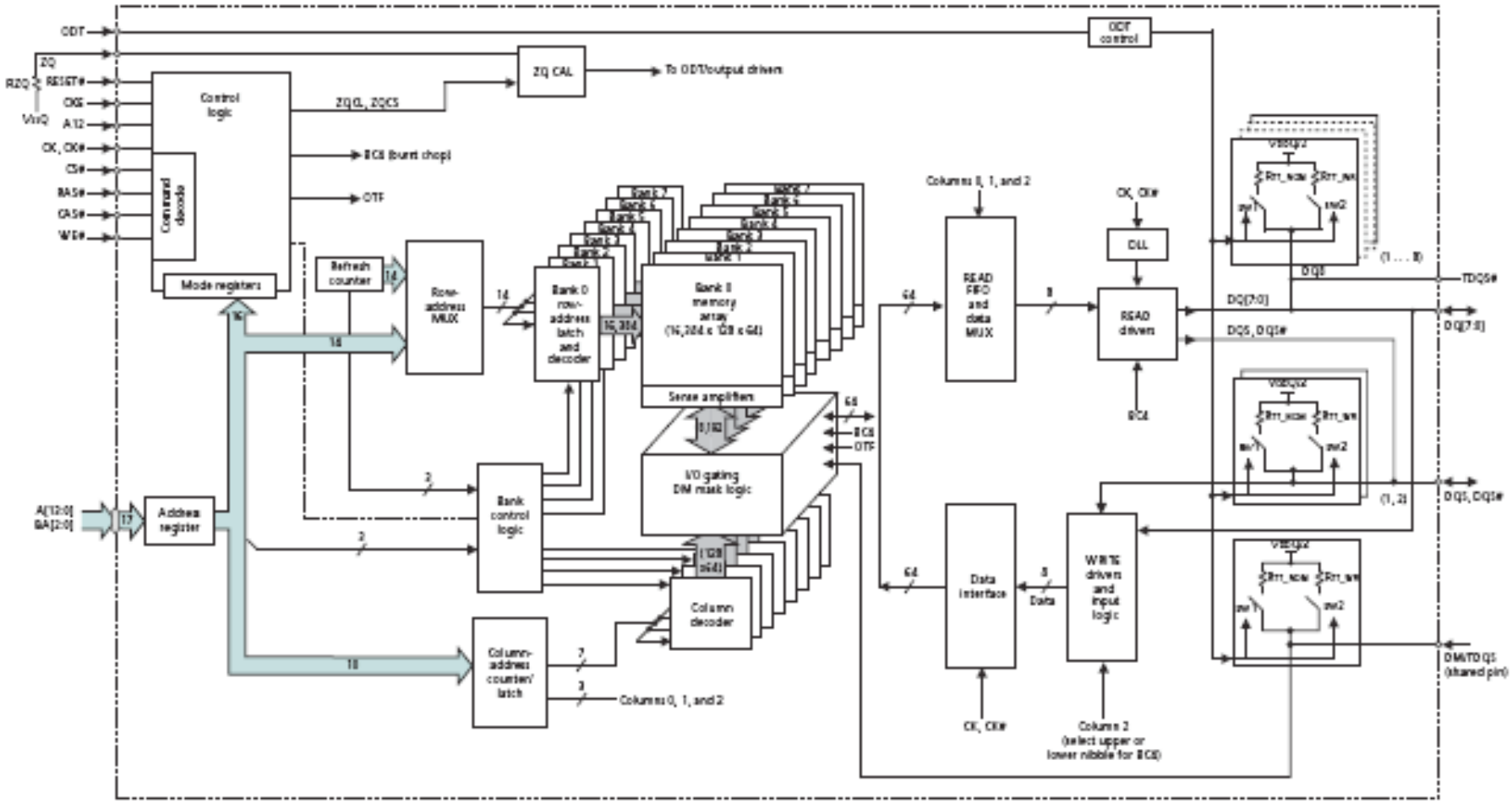


The cost to sense, collect, generate and calculate data is declining much faster than the cost to access, manage and store it

Source: David Turek, IBM

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

1Gbit DDR3 Architecture

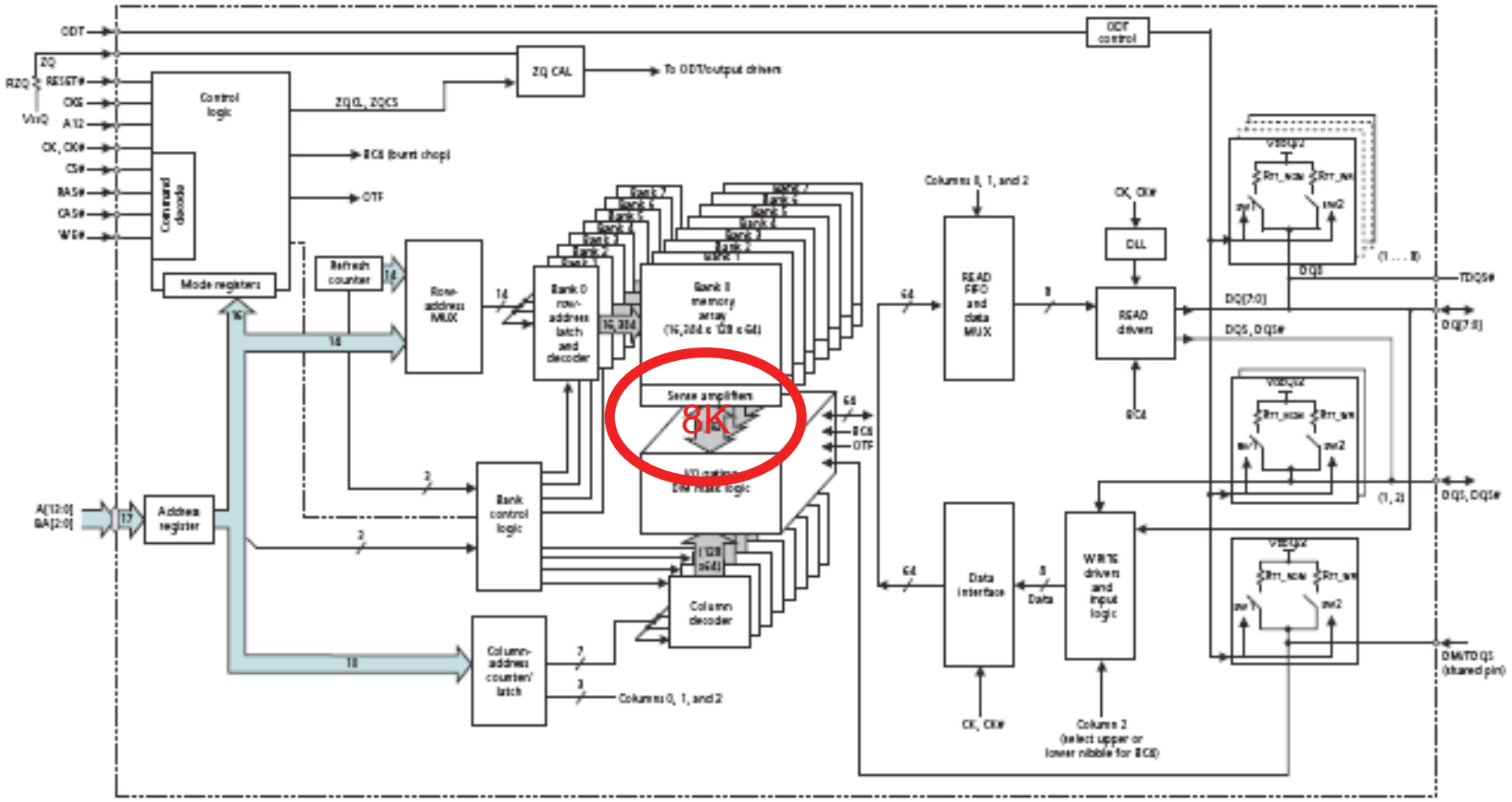


Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

1Gbit DDR3 Architecture



Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Exascale: The Power of Memory

Assumptions

Cell Voltage	1.2	V
Cell Capacitance	25	fF
Bitline Capacitance	75	fF
Memory System Bandwidth	1	EB/sec

Simplified Results:

Energy/bit	36	fJ
Total Memory Cell Power	288	KW
With Bitline	1150	KW
With 512X Over-Fetch	590	MW

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Office of Science U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Conclusions

- **Memory technology requires major reorganization (if industry stays alive)**
  - More ranks/banks, Less over-fetch, new drivers
  - Chip stacking or optical memory interfaces

- **We will have to live with less memory / computational performance**
- **We will have lower memory bandwidth/ computational performance (< 0.001 bytes/ flop)**

NERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

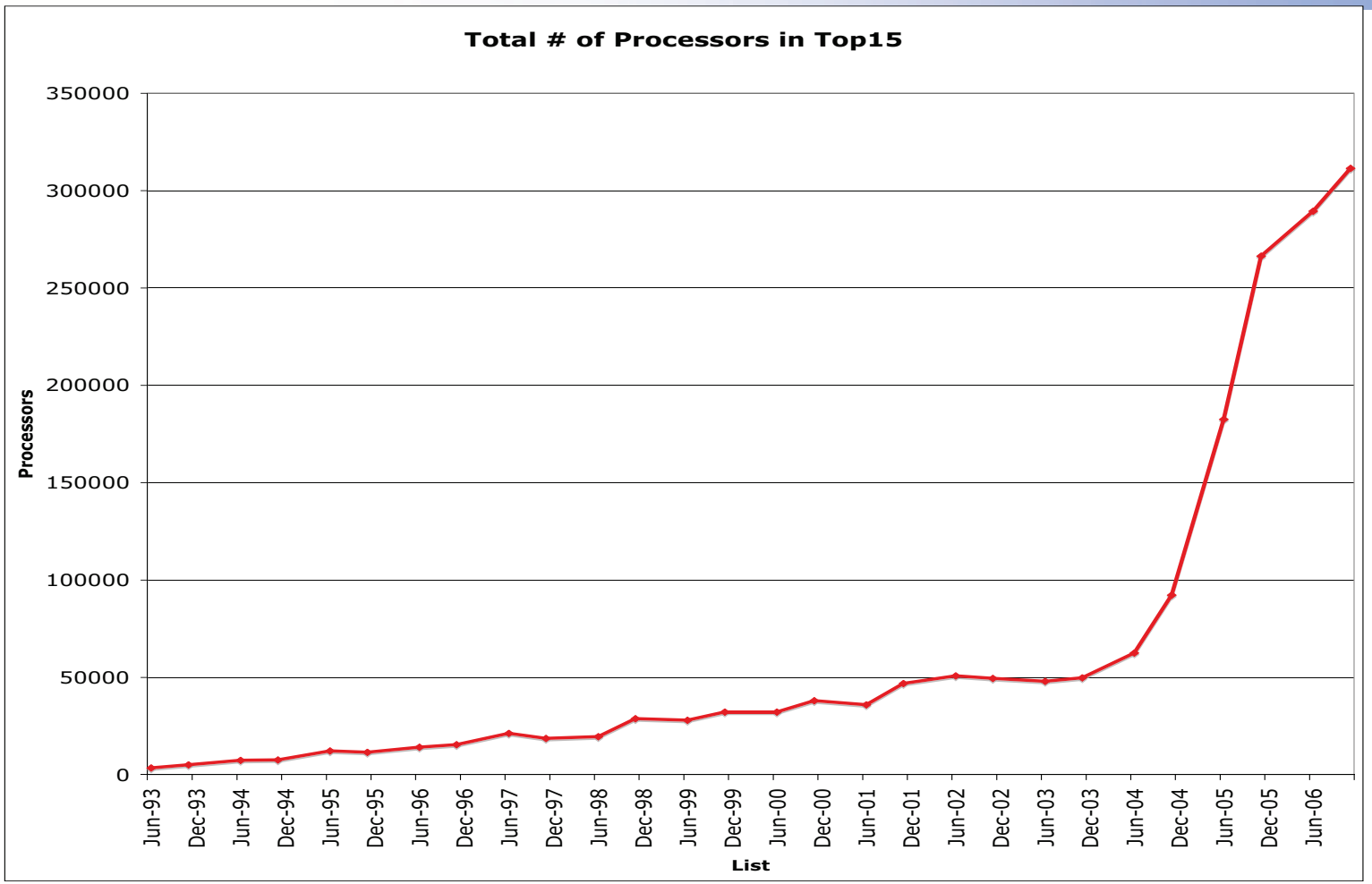
Algorithms

Scaling to Billion-way Parallelism

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

The Future of HPC System Concurrency



Must ride exponential wave of increasing concurrency for forseeable future!

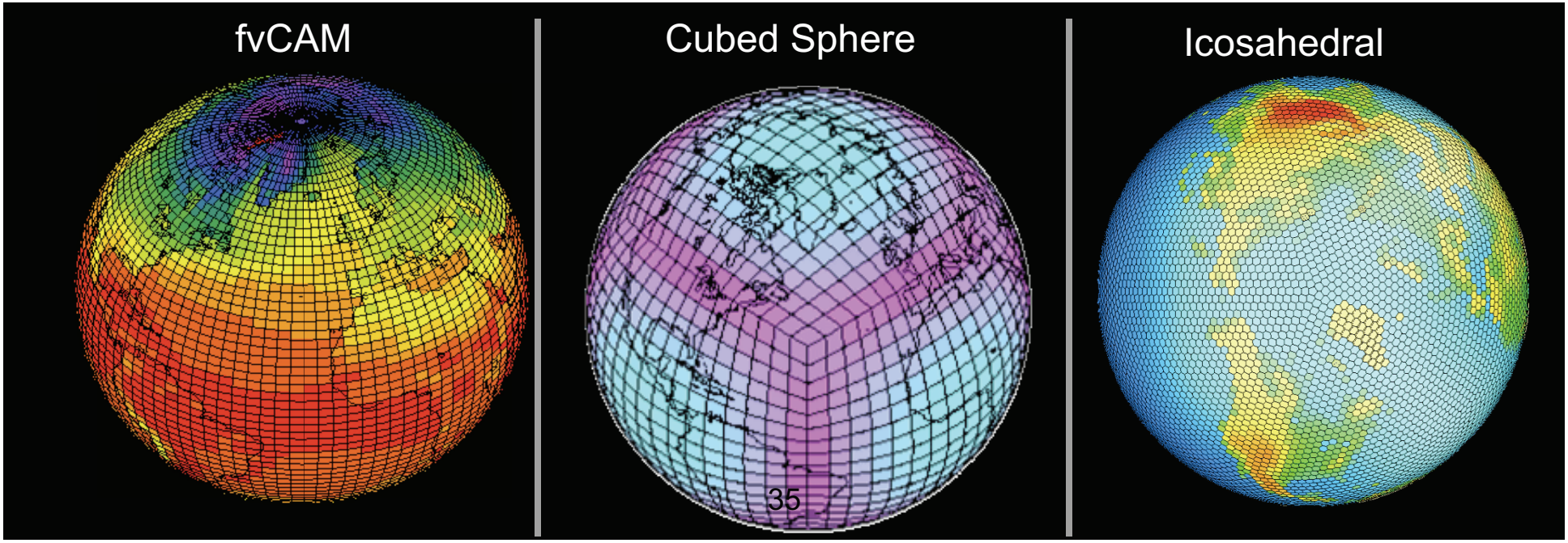
Fortunately, most of the concurrency growth is within a single socket

Climate Model

New Approaches for Massive Parallelism

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

- **Existing Latitude-longitude based algorithm advection algorithm breaks down significantly before 1km scale!**
  - **Grid cell aspect ratio at the pole is 10000!**
  - **Advection time step is problematic at this scale**
- **Ultimately requires new discretization for atmosphere model**
  - **Must expose sufficient parallelism to exploit power-efficient design**
  - **Partner with CSU/Randall Group to use the Icosahedral Code**
  - **Uniform cell aspect ratio across globe**



Where to Find 12 Orders in 10 years?

Jardin & Keyes

Hardware: 3

Software: 9

- 1.5 orders: increased processor speed and efficiency
- 1.5 orders: increased concurrency
- 1 order: higher-order discretizations
  - Same accuracy can be achieved with many fewer elements
- 1 order: flux-surface following gridding
  - Less resolution required along than across field lines
- 4 orders: adaptive gridding
  - Zones requiring refinement are <1% of ITER volume and resolution requirements away from them are ~10^2 less severe
- 3 orders: implicit solvers
  - Mode growth time 9 orders longer than Alfven-limited CFL

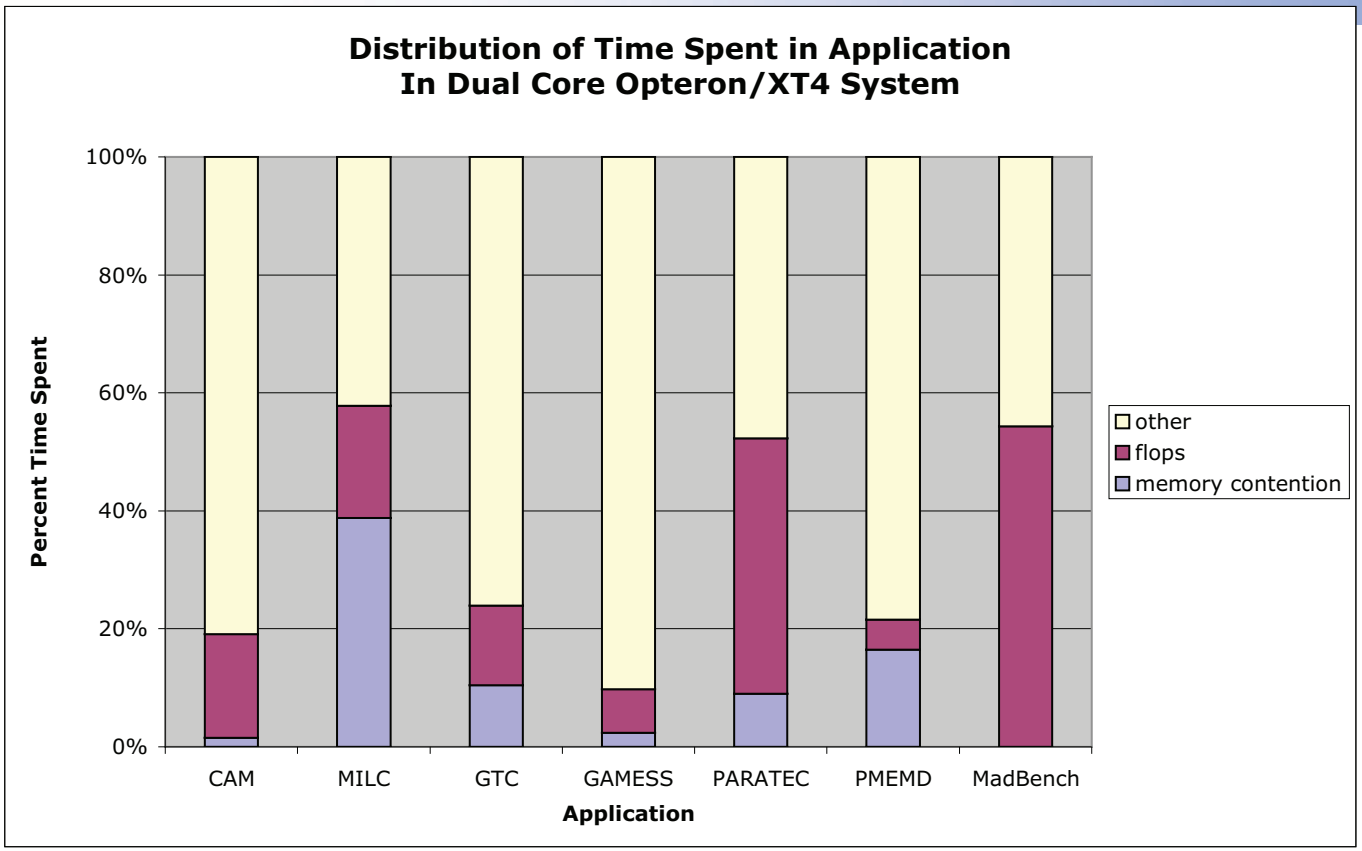
Conclusion

- **Consequence: must find strong-scaling from explicit parallelism**
  - **That's a tall order!**
  - **Used in 1980's to argue against MPPs**

Software Performance

Auto-tuning: Don't depend on a human to do a machine's job.

Performance Profiles
(maintaining system balance)

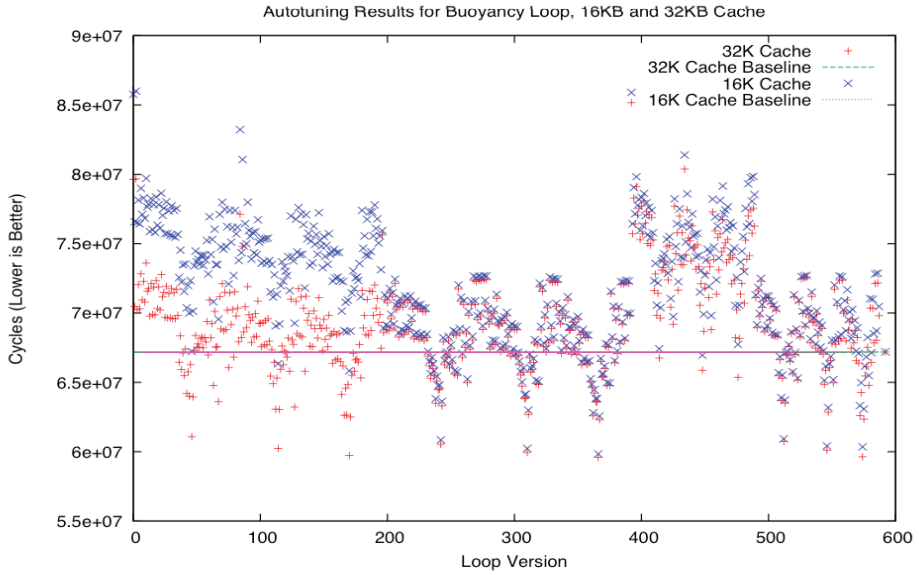


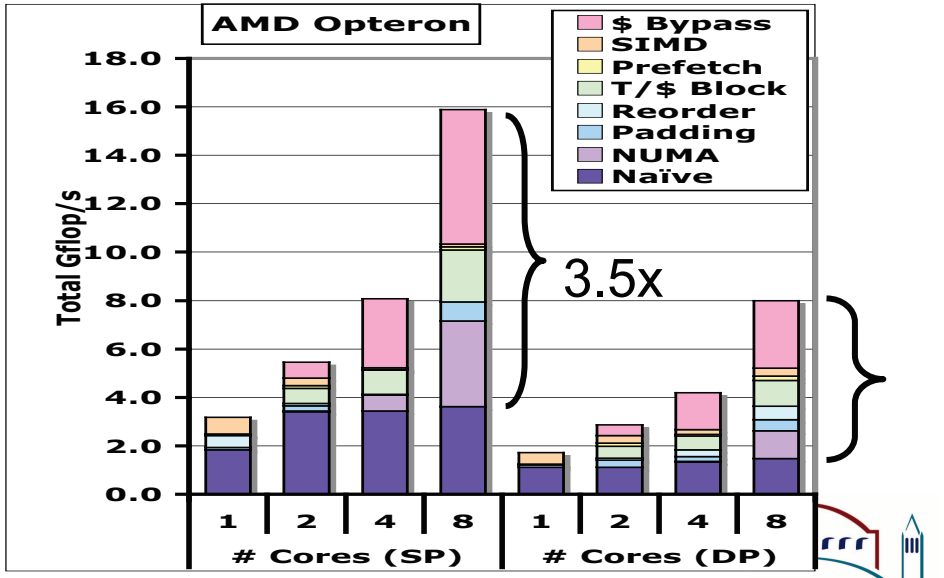
- **Neither memory bandwidth nor FLOPs dominate runtime**
- **The "other" category dominated by memory latency stalls**
- **Points to inadequacies in current CPU core design (inability to tolerate latency)** ***Lets not forget about latency!***

Auto-tuning

- **Problem: want to compare best potential performance of diverse architectures, avoiding**
  - Non-portable code
  - Labor-intensive user optimizations for each specific architecture

- **Our Solution: Auto-tuning**
  - **Automate search across a complex optimization space**
  - **Achieve performance far beyond current compilers**
  - **achieve performance portability for diverse architectures!**





Office of Science
U.S. DEPARTMENT OF ENERGY

NERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

BERKELEY LAB

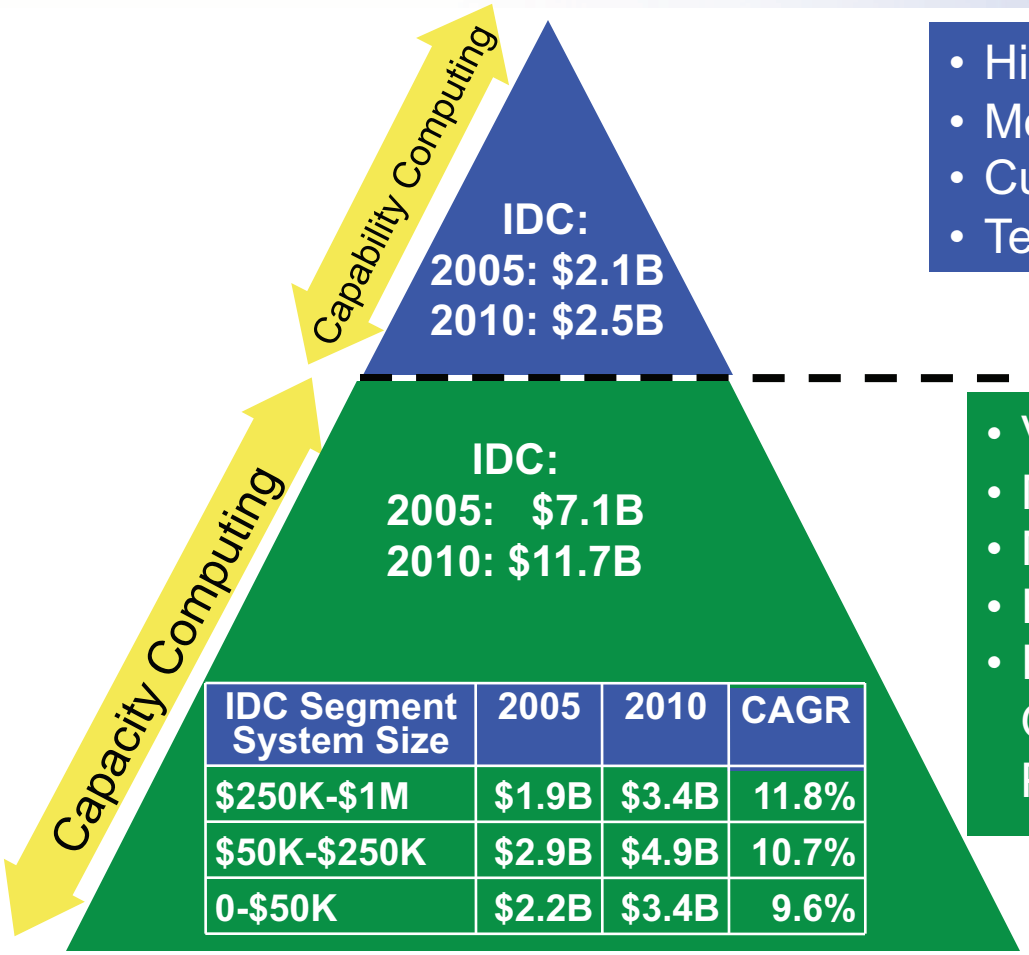
Conclusion

- Huge opportunities for energy-efficiency improvement simply by optimizing code performance
- Compilers cannot achieve this because of insufficient information
  - Assume flat machine model (which is wrong)
  - Cannot exploit domain-specific knowledge
- Auto-tuners
  - Can exploit domain-specific abstraction (motifs)
  - Can automate search of design space for performance portability
- Languages:
  - Need to expose correct machine model (flat model is wrong)
  - Need to express locality

NERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

How Can We Achieve our Goals Cost Effectively?

Intel HPC Market Overview



IDC Segment System Size	2005	2010	CAGR
$250K-$1M	$1.9B	$3.4B	11.8%
$50K-$250K	$2.9B	$4.9B	10.7%
0-$50K	$2.2B	$3.4B	9.6%

- High End Systems (>$1M)
- Most/all Top 500 systems
- Custom SW & ISV apps
- Technology risk takers & early adopte

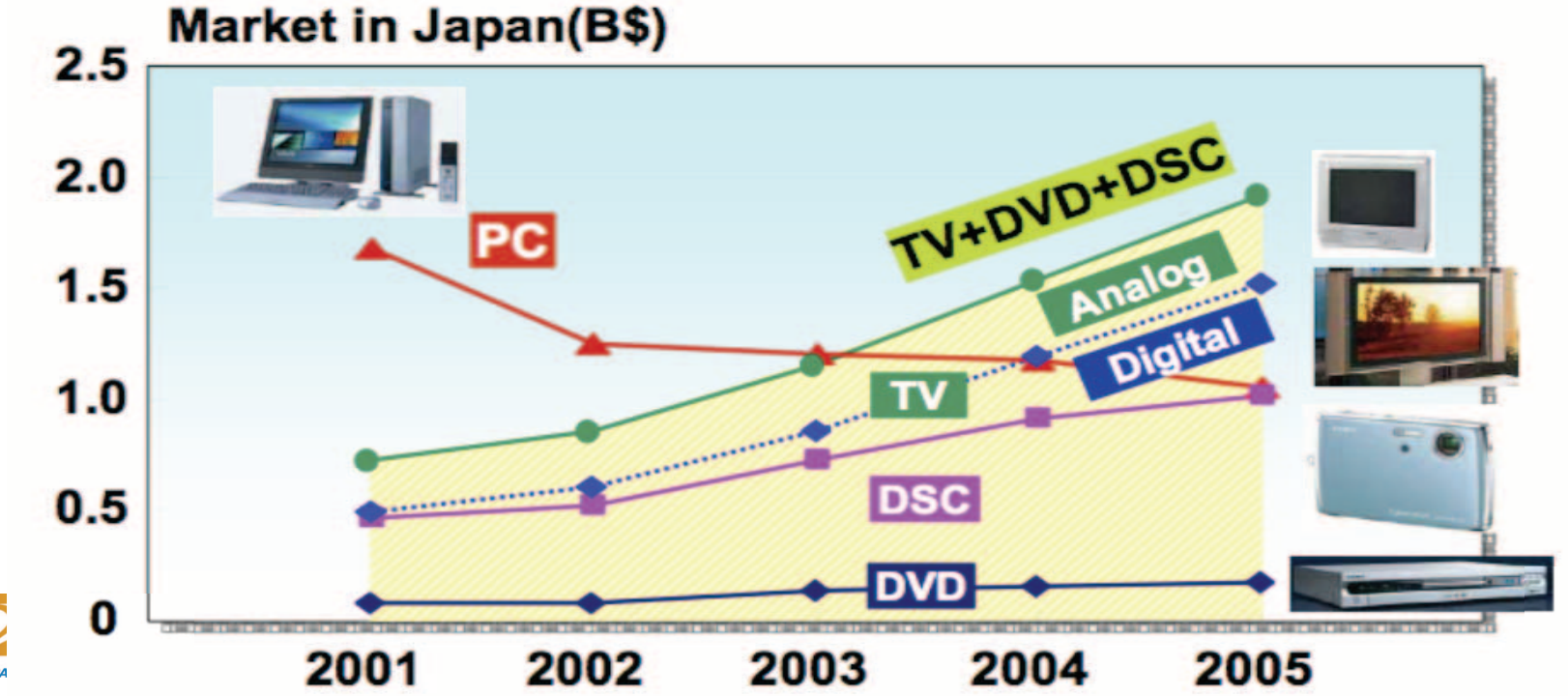
- Volume Market
- Mainly capacity; <~150 nodes
- Mostly clusters; >50% & growing
- Higher % of ISV apps
- Fast growth from commercial HPC; Oil &Gas, Financial services, Pharma, Aerospace, etc.

Total market >$10.0B in 2006
Forecast >$15.5B in 2011

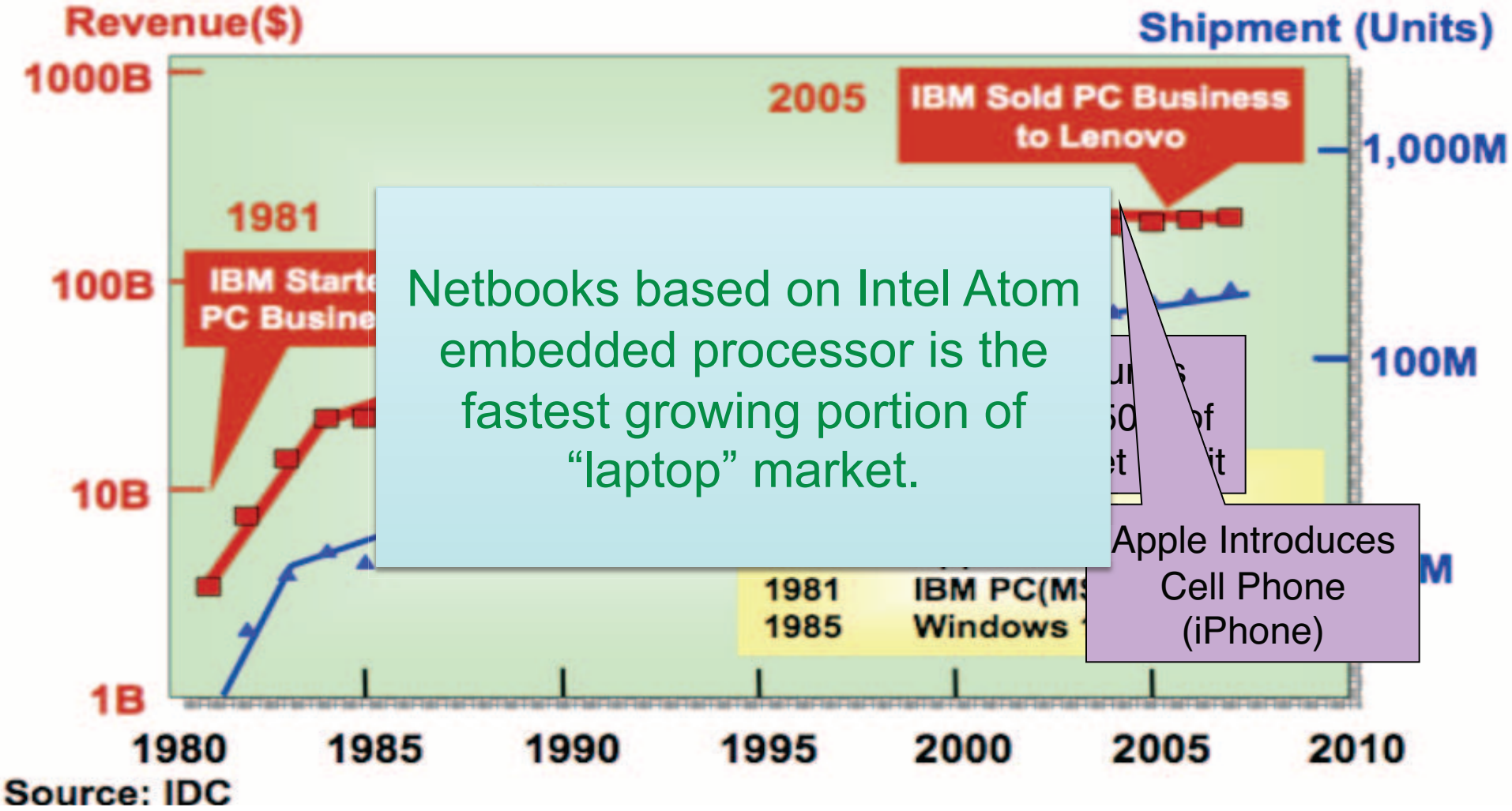
HPC is built with of pyramid investment model

Processor Technology Trend

- 1990s - R&D computing hardware dominated by desktop/COTS
  - Had to learn how to use COTS technology for HPC
- 2010 - R&D investments moving rapidly to consumer electronics/ embedded processing
  - Must learn how to leverage embedded processor technology for future HPC systems



Consumer Electronics has Replaced PCs as the Dominant Market Force in CPU Design!!

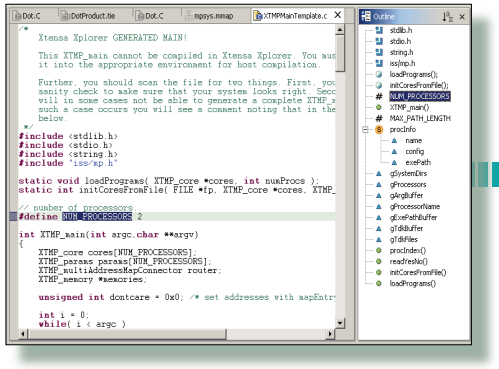


Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

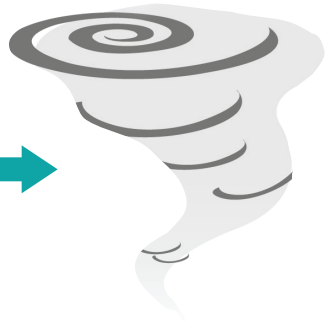
Embedded Design Automation

(Example from Existing Tensilica Design Flow)



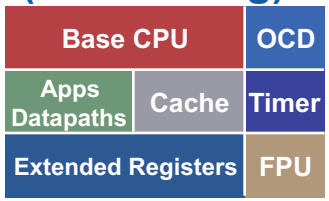
Processor configuration

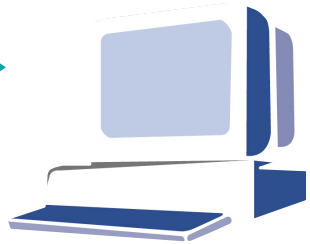
1. Select from menu
2. Automatic instruction discovery (XPRES Compiler)
3. Explicit instruction description (TIE)



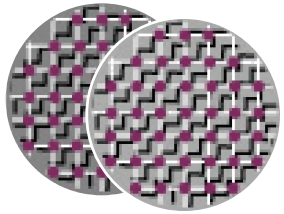
Processor Generator (*Tensilica*)

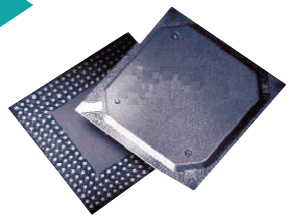
Application-optimized processor implementation (RTL/Verilog)





Tailored SW Tools: Compiler, debugger, simulators, Linux, other OS Ports *(Automatically generated together with the Core)*





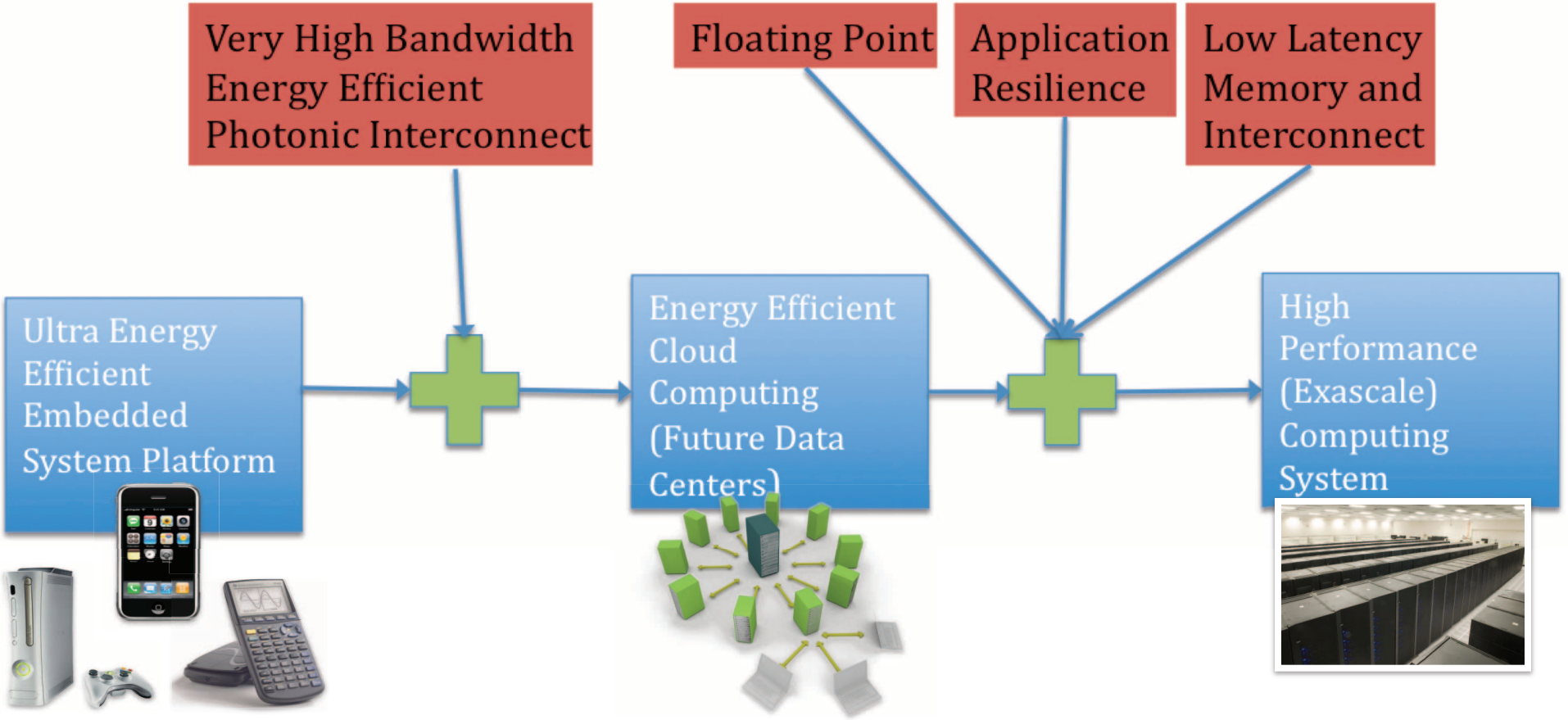
Build with any process in any fab

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Technology Continuity for A Sustainable Hardware Ecosystem



Need building blocks for a compelling environment at all scales

NERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

If this is such a great idea, then why don't you do it?

Eating our own dogfood

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Green Flash Overview

- **Research effort: study feasibility and share insight w/community**

- **Elements of the approach**
  - **Choose the science target first** *(climate for example)*
  - **Design systems for applications** *(rather than the reverse)*
  - **Design hardware, software, scientific algorithms together using hardware emulation and auto-tuning**

- **What is NEW about this approach**
  - **Leverage commodity processes used to design power efficient embedded devices (redirect the tools to benefit scientific computing!)**
  - **Auto-tuning to automate mapping of algorithm to complex hardware**
  - **RAMP: Fast hardware-accelerated emulation of new chip designs**

Applicable to broad range of scientific computing applications

NERSC NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Office of Science U.S. DEPARTMENT OF ENERGY

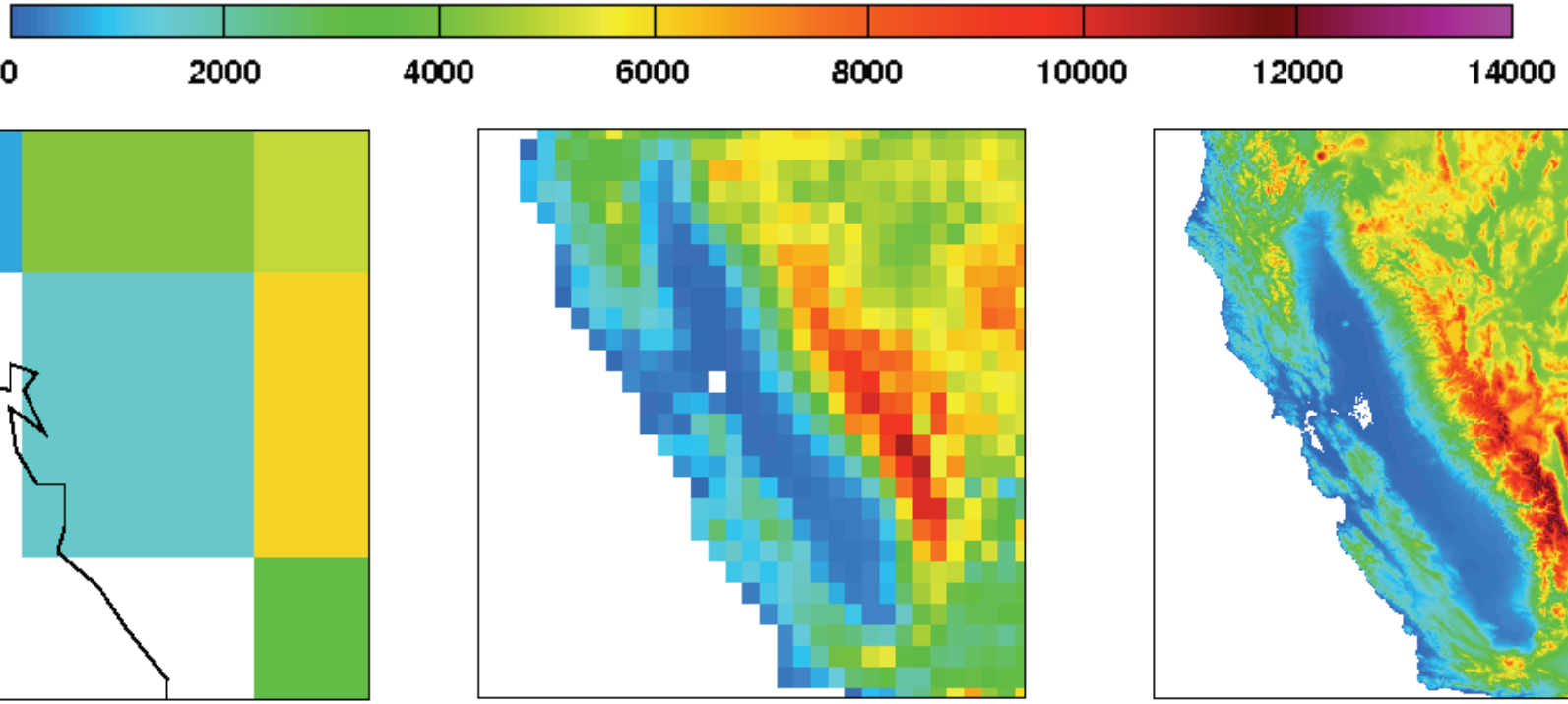
BERKELEY LAB

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Identify Target First!

(Global Cloud Resolving Climate Model)

Surface Altitude (feet)



200km
Typical resolution of IPCC AR4 models

25km
Upper limit of climate models with cloud parameterizations

1km
Cloud system resolving models are a transformational change

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

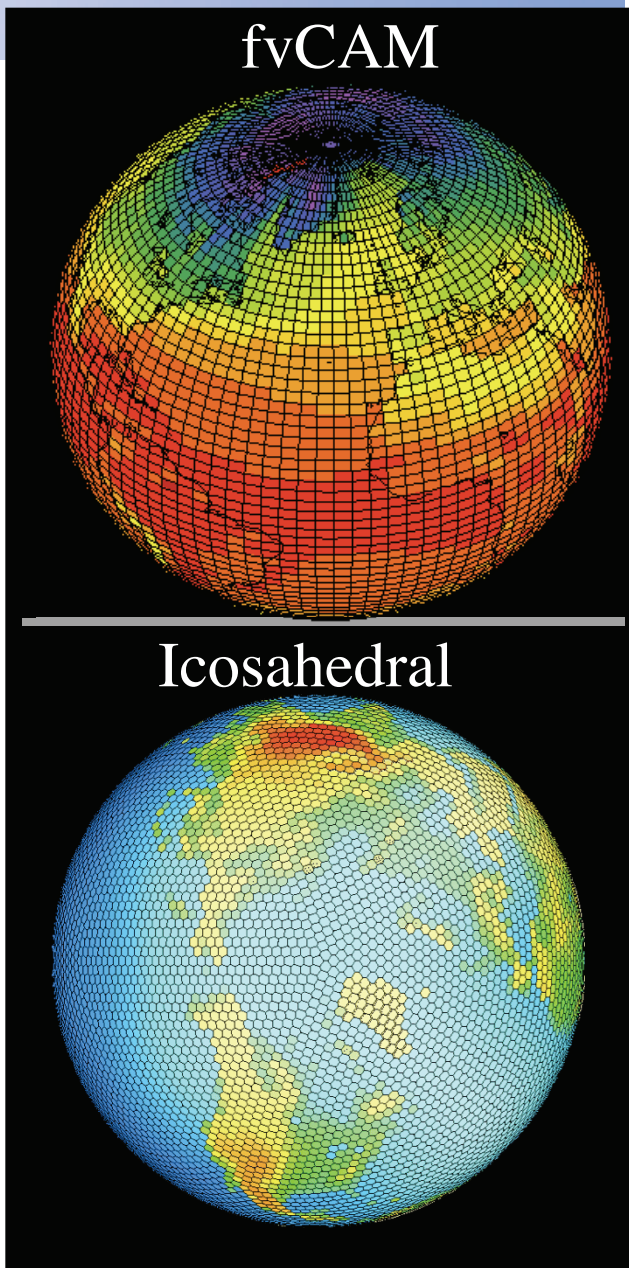
Requirements for 1km Climate Computer

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Must maintain 1000x faster than real time for practical climate simulation

- **~2 million horizontal subdomains**
- **100 Terabytes of Memory**
  - **5MB memory per subdomain**
- **~20 million total subdomains**
  - **20 PF sustained (200PF peak)**
  - **Nearest-neighbor communication**

- ***New discretization for climate model***
  - ***CSU Icosahedral Code***

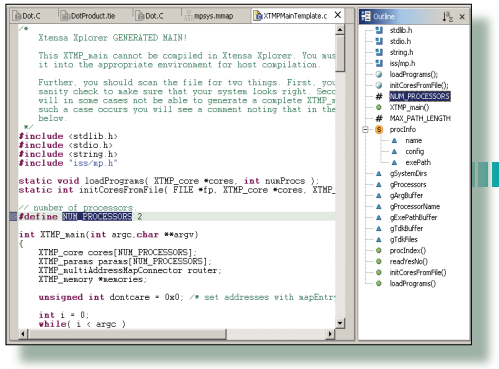


Office of Science
U.S. DEPARTMENT OF ENERGY

Embedded Design Automation

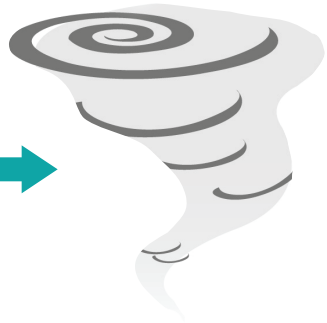
(Example from Existing Tensilica Design Flow)

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER



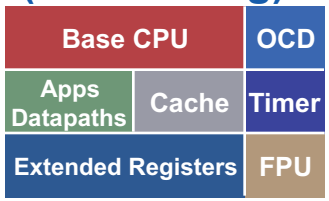
Processor configuration

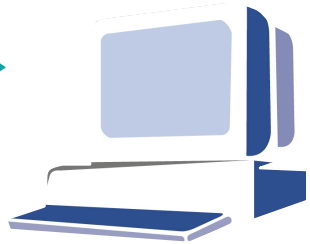
1. Select from menu
2. Automatic instruction discovery (XPRES Compiler)
3. Explicit instruction description (TIE)



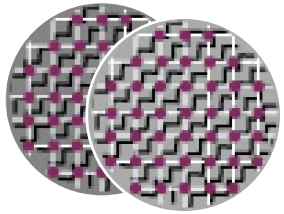
Processor Generator (*Tensilica*)

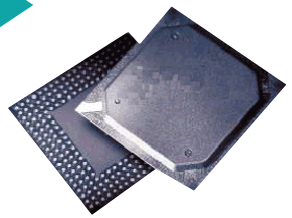
Application-optimized processor implementation (RTL/Verilog)





Tailored SW Tools: Compiler, debugger, simulators, Linux, other OS Ports *(Automatically generated together with the Core)*





Build with any process in any fab

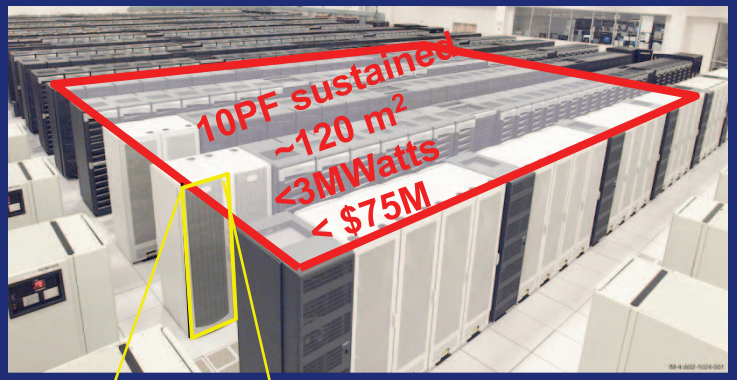
Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Climate System Design Concept

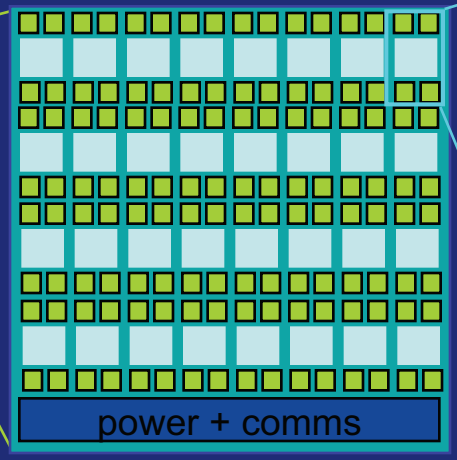
Strawman Design Study

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

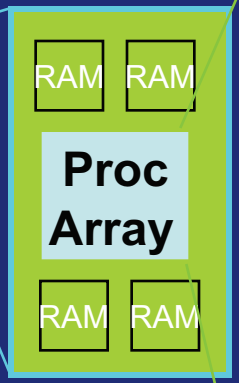


32 boards per rack

100 racks @ ~25KW



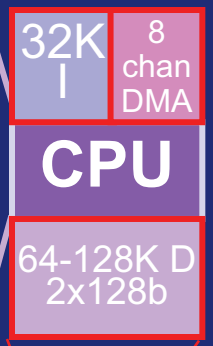
32 chip + memory clusters per board (2.7 TFLOPS @ 700W

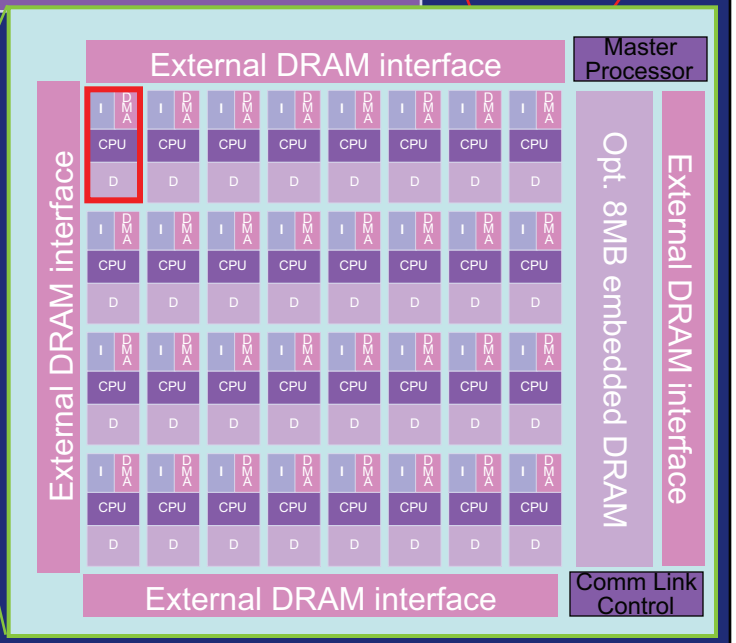


8 DRAM per processor chip: ~50 GB/s

VLIW CPU:

- 128b load-store + 2 DP MUL/ADD + integer op/ DMA per cycle:
- Synthesizable at 650MHz in commodity 65nm
- 1mm² core, 1.8-2.8mm² with inst cache, data cache data RAM, DMA interface, 0.25mW/MHz
- Double precision SIMD FP : 4 ops/cycle (2.7GFLOPs)
- Vectorizing compiler, cycle-accurate simulator, debugger GUI (Existing part of Tensilica Tool Set)
- 8 channel DMA for streaming from on/off chip DRAM
- Nearest neighbor 2D communications grid





32 processors per 65nm chip
83 GFLOPS @ 7W

Green Flash Strawman System Design In 2008

We examined three different approaches:

- **AMD Opteron:** Commodity approach, lower efficiency for scientific applications offset by cost efficiencies of mass market
- **BlueGene:** Generic embedded processor core and customize system-on-chip (SoC) services to improve power efficiency for scientific applications
- **Tensilica XTensa:** Customized embedded CPU w/SoC provides further power efficiency benefits but maintains programmability

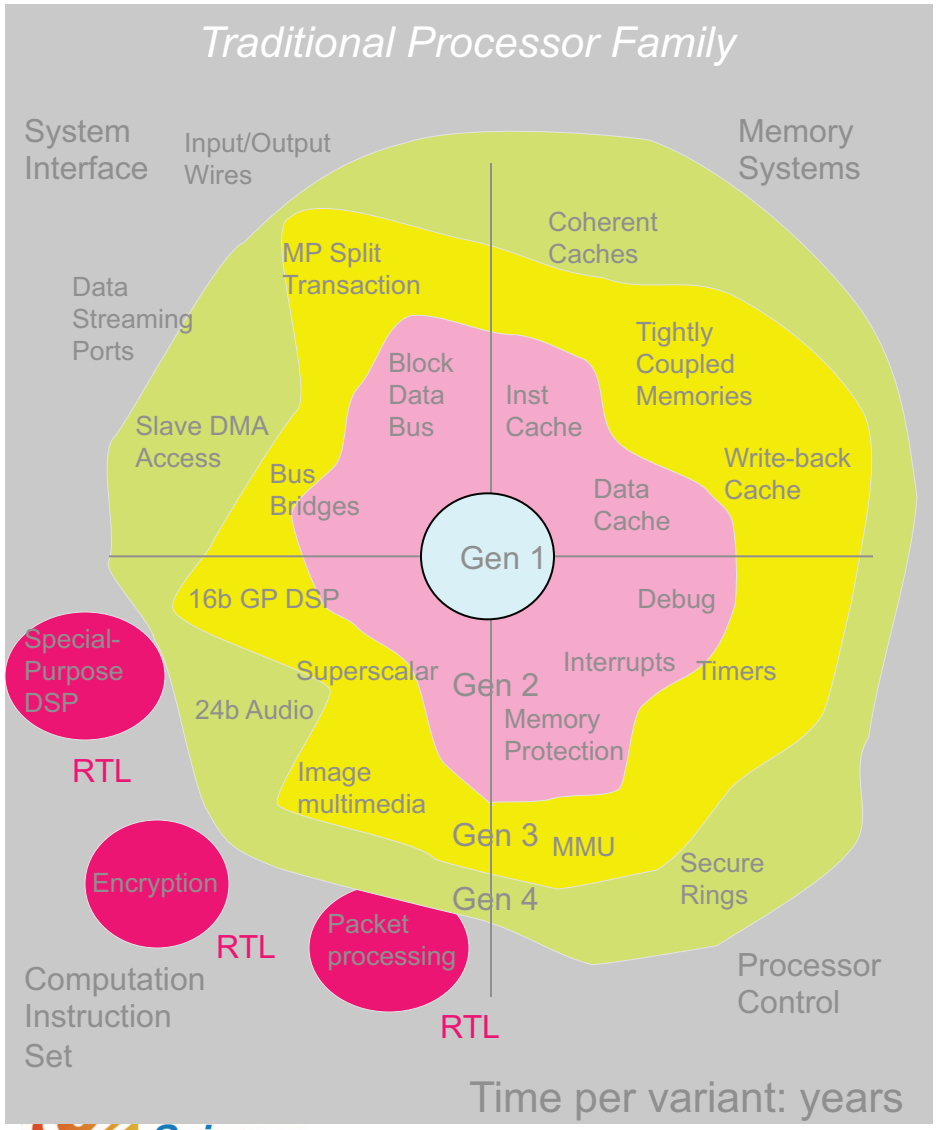
Processor	Clock	Peak/ Core (Gflops)	Cores/ Socket	Sockets	Cores	Power	Cost 2008
AMD Opteron	2.8GHz	5.6	2	890K	1.7M	179 MW	$1B+
IBM BG/P	850MHz	3.4	4	740K	3.0M	20 MW	$1B+
Green Flash / Tensilica XTensa	650MHz	2.7	32	120K	4.0M	3 MW	$75M

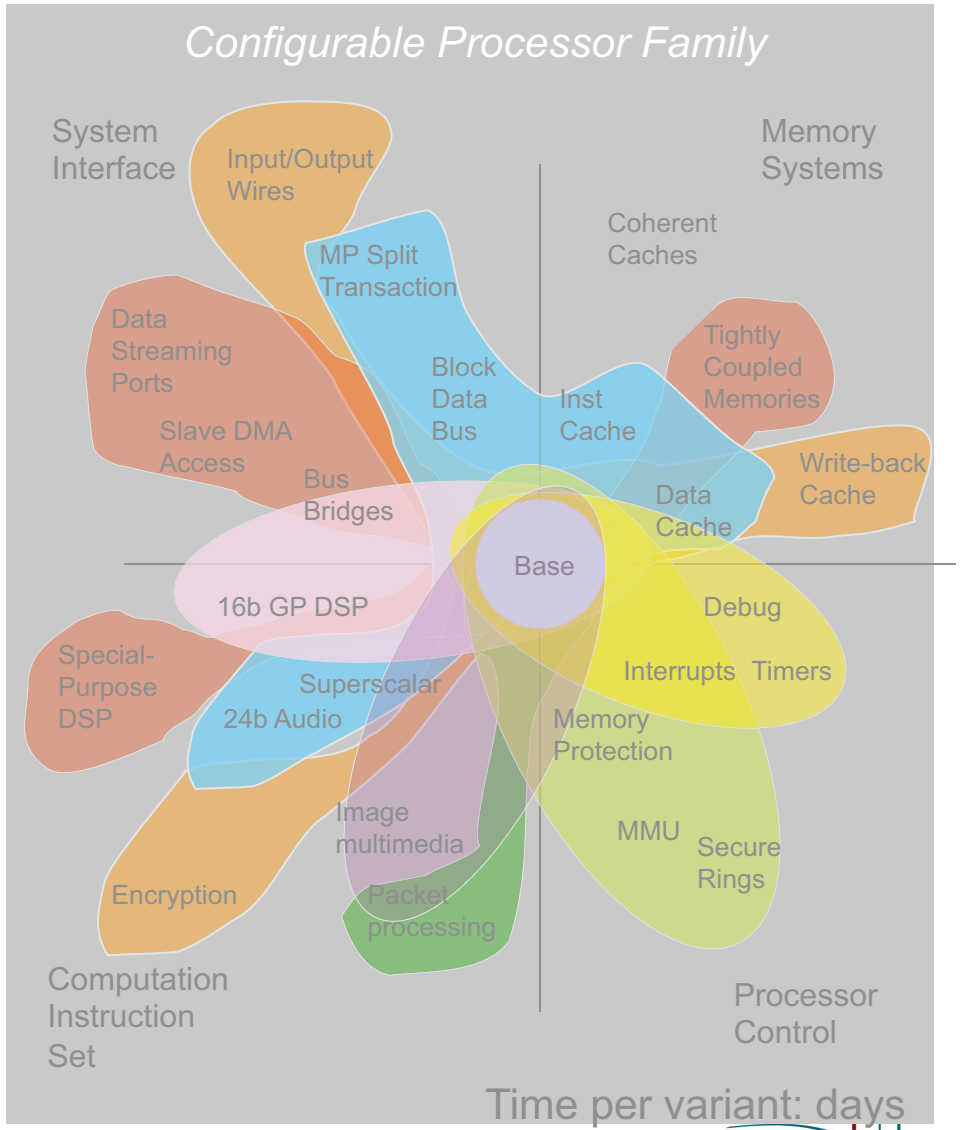
What we have learned from our more detailed design study

Mark Horowitz 2007*: "Years of research in low-power embedded computing have shown only one design technique to reduce power:* ***reduce waste****."*

Seymour Cray 1977: *"Don't put anything in to a supercomputer that isn't necessary."*

Peel Back the Historical Growth of Instruction Sets *(accretion of cruft)*





Area = silicon cost and power

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

A Short List of x86 Opcodes that Science Applications Don't Need!

mnemonic	op1	op2	op3	op4	iext	pf	0f	po	so	o	proc	st	m	rl	x	tested f	modif f	def f	undef f	f values	description, notes
AAA	AL	AH						37								.....a..	o..szapc	.....a.c	o..sz.p.		ASCII Adjust After Addition
AAD	AL	AH						D5	0A								o..szapc	...sz.p.	o....a.c		ASCII Adjust AX Before Division
AAM	AL	AH						D4	0A								o..szapc	...sz.p.	o....a.c		ASCII Adjust AX After Multiply
AAS	AL	AH						3F								.....a..	o..szapc	.....a.c	o..sz.p.		ASCII Adjust AL After Subtraction
ADC	r/m8	r8						10		r					L	c	o..szapc	o..szapc			Add with Carry
ADC	r/m16/32/64	r16/32/64						11		r					L	c	o..szapc	o..szapc			Add with Carry
ADC	r8	r/m8						12		r						c	o..szapc	o..szapc			Add with Carry
ADC	r16/32/64	r/m16/32/64						13		r						c	o..szapc	o..szapc			Add with Carry
ADC	AL	imm8						14								c	o..szapc	o..szapc			Add with Carry
ADC	rAX	imm16/32						15								c	o..szapc	o..szapc			Add with Carry
ADC	r/m8	imm8						80		2					L	c	o..szapc	o..szapc			Add with Carry
ADC	r/m16/32/64	imm16/32						81		2					L	c	o..szapc	o..szapc			Add with Carry
ADC	r/m8	imm8						82		2					L	c	o..szapc	o..szapc			Add with Carry
ADC	r/m16/32/64	imm8						83		2					L	c	o..szapc	o..szapc			Add with Carry
ADD	r/m8	r8						00		r					L		o..szapc	o..szapc			Add
ADD	r/m16/32/64	r16/32/64						01		r					L		o..szapc	o..szapc			Add
ADD	r8	r/m8						02		r							o..szapc	o..szapc			Add
ADD	r16/32/64	r/m16/32/64						03		r							o..szapc	o..szapc			Add
ADD	AL	imm8						04									o..szapc	o..szapc			Add
ADD	rAX	imm16/32						05									o..szapc	o..szapc			Add
ADD	r/m8	imm8						80		0					L		o..szapc	o..szapc			Add
ADD	r/m16/32/64	imm16/32						81		0					L		o..szapc	o..szapc			Add
ADD	r/m8	imm8						82		0					L		o..szapc	o..szapc			Add
ADD	r/m16/32/64	imm8						83		0					L		o..szapc	o..szapc			Add
ADDPD	xmm	xmm/m128			sse2	66	0F	58		r	P4+										Add Packed Double-FP Values
ADDPS	xmm	xmm/m128			sse1		0F	58		r	P3+										Add Packed Single-FP Values
ADDSD	xmm	xmm/m64			sse2	F2	0F	58		r	P4+										Add Scalar Double-FP Values
ADDSS	xmm	xmm/m32			sse1	F3	0F	58		r	P3+										Add Scalar Single-FP Values
ADDSUBPD	xmm	xmm/m128			sse3	66	0F	D0		r	P4++										Packed Double-FP Add/Subtract
ADDSUBPS	xmm	xmm/m128			sse3	F2	0F	D0		r	P4++										Packed Single-FP Add/Subtract
ADX	AL	AH	imm8					D5									o..szapc	...sz.p.	o....a.c		Adjust AX Before Division
ALTER						64					P4+	U[1]									Alternating branch prefix (used only with Jcc instructions)
AMX	AL	AH	imm8					D4									o..szapc	...sz.p.	o....a.c		Adjust AX After Multiply
AND	r/m8	r8						20		r					L		o..szapc	o..sz.pc	.....a..	o......c	Logical AND
AND	r/m16/32/64	r16/32/64						21		r					L		o..szapc	o..sz.pc	.....a..	o......c	Logical AND
AND	r8	r/m8						22		r							o..szapc	o..sz.pc	.....a..	o......c	Logical AND
AND	r16/32/64	r/m16/32/64						23		r							o..szapc	o..sz.pc	.....a..	o......c	Logical AND
AND	AL	imm8						24									o..szapc	o..sz.pc	.....a..	o......c	Logical AND
AND	rAX	imm16/32						25									o..szapc	o..sz.pc	.....a..	o......c	Logical AND
AND	r/m8	imm8						80		4					L		o..szapc	o..sz.pc	.....a..	o......c	Logical AND
AND	r/m16/32/64	imm16/32						81		4					L		o..szapc	o..sz.pc	.....a..	o......c	Logical AND
AND	r/m8	imm8						82		4					L		o..szapc	o..sz.pc	.....a..	o......c	Logical AND
AND	r/m16/32/64	imm8						83		4	03+				L		o..szapc	o..sz.pc	.....a..	o......c	Logical AND
ANDNPD	xmm	xmm/m128			sse2	66	0F	55		r	P4+										Bitwise Logical AND NOT of Packed Double-FP Values
ANDNPS	xmm	xmm/m128			sse1		0F	55		r	P3+										Bitwise Logical AND NOT of Packed Single-FP Values
ANDPD	xmm	xmm/m128			sse2	66	0F	54		r	P4+										Bitwise Logical AND of Packed Double-FP Values
ANDPS	xmm	xmm/m128			sse1		0F	54		r	P3+										Bitwise Logical AND of Packed Single-FP Values

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

More Wasted Opcodes

ARPL	r/m16	r16	
BOUND	r16/32	m16/32&16/32	eFlags
BSF	r16/32/64	r/m16/32/64	
BSR	r16/32/64	r/m16/32/64	
BSWAP	r16/32/64		
BT	r/m16/32/64	r16/32/64	
BT	r/m16/32/64	imm8	
BTC	r/m16/32/64	imm8	
BTC	r/m16/32/64	r16/32/64	
BTR	r/m16/32/64	r16/32/64	
BTR	r/m16/32/64	imm8	
BTS	r/m16/32/64	r16/32/64	
BTS	r/m16/32/64	imm8	
CALL	rel16/32		
CALL	rel32		
CALL	r/m16/32		
CALL	r/m64		
CALLF	ptr16:16/32		
CALLF	m16:16/		
CBW	*AX*		
CBW	*AX*		
CWDE	*EAX*		
CDQE	*RAX*		
CDQ	*EDX*		
CLC			
CLD			
CLFLUSH	m8		
CLI			
CLTS	*CR0*		
CMC			
CMOVB	r16/32/		
CMOVNAE	r16/32/		
CMOVC	r16/32/		
CMOVBE	r16/32/		
CMOVNA	r16/32/		
CMOVL	r16/32/		
CMOVNGE	r16/32/		
CMOVLE	r16/32/		
CMOVNG	r16/32/		
CMOVNB	r16/32/		
CMOVAE	r16/32/		
CMOVNC	r16/32/		
CMOVNBE	r16/32/		
CMOVA	r16/32/		
CMOVNL	r16/32/		
CMOVGE	r16/32/		
CMOVNLE	r16/32/		
CMOVG	r16/32/		

CVTPS2PD	xmm	xmm/m128	
CVTPS2PI	mm	xmm/m64	
CVTSD2SI	r32/64	xmm/m64	
CVTSD2SS	xmm	xmm/m64	
CVTSI2SD	xmm	r/m32/64	
CVTSI2SS	xmm	r/m32/64	
CVTSS2SD	xmm	xmm/m32	
CVTSS2SI	r32/64	xmm/m32	
CVTTPD2DQ	xmm	xmm/m128	
CVTTPD2PI	mm	xmm/m128	
CVTTPS2DQ	xmm	xmm/m128	
CVTTPS2PI	mm	xmm/m64	
CVTTSD2SI	r32/64	xmm/m64	
CVTTSS2SI	r32/64	xmm/m32	
CWD	*DX*	*AX*	
CWD	*DX*	*AX*	
CDQ	*EDX*	*EAX*	
CQO	*RDX*	*RAX*	
CWDE	*EAX*	*AX*	
DAA	*AL*		
DAS	*AL*		

r16/32/64	r/m16/32/64	
r16/32/64	r/m16/32/64	
r16/32/64	r/m16/32/64	
r/m8	r8	
r/m16/32/64	r16/32/64	
r8	r/m8	
r16/32/64	r/m16/32/64	
AL	imm8	
rAX	imm16/32	
r/m8	imm8	
r/m16/32/64	imm16/32	
r/m8	imm8	
r/m16/32/64	imm8	
xmm	xmm/m128	imm8
xmm	xmm/m128	imm8
m8 *m8*	m8 *m8*	
m16	m16	

FXCH4	*ST*	STi
FXCH4	*ST*	STi
FXCH7	*ST*	STi
FXCH7	*ST*	STi
FXRSTOR	*ST*	*ST1*
FXRSTOR	*ST*	*ST1*
FXSAVE	m512	*ST*
FXSAVE	m512	*ST*
FXTRACT	*ST*	
FYL2X	*ST1*	*ST*
FYL2XP1	*ST1*	*ST*
GS	*GS*	
HADDPD	xmm	xmm/m128
HADDPS	xmm	xmm/m128
HLT		
HSUBPD	xmm	xmm/m128

INTO	eFlags	
INVD		
INVLPG	m	

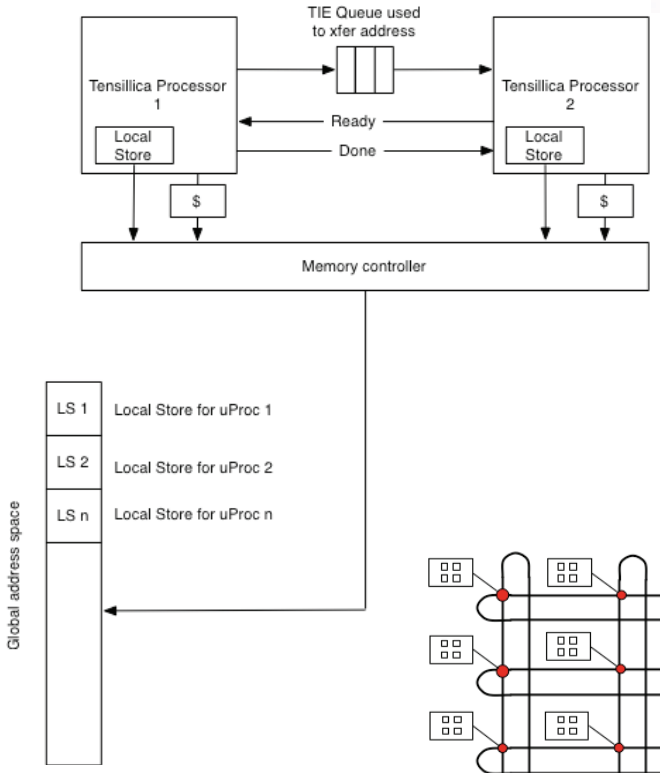
- **We only need 80 out of the nearly 300 ASM instructions in the x86 instruction set!**

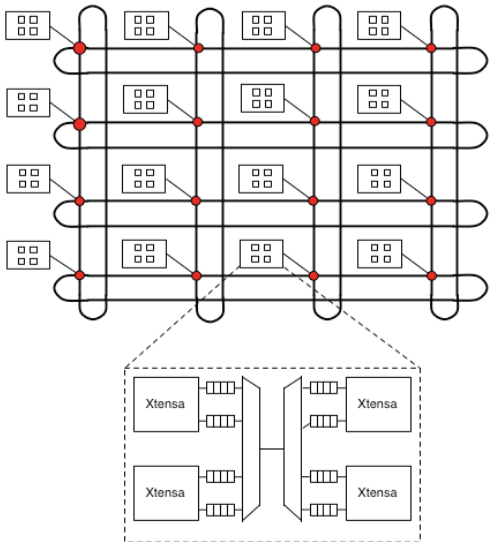
- Still have all of the 8087 and 8088 instructions!
- Wide SIMD Doesn't Make Sense with Small Cores
- Neither does Cache Coherence
- Neither does HW Divide or Sqrt for loops
  - Creates pipeline bubbles
  - Better to unroll it across the loops (like IBM MASS libraries)
- Move TLB to memory interface because its still too huge (but still get precise exceptions from segmented protection on each core)

Office of Science U.S. DEPARTMENT OF ENERGY

Science-Optimized Processor Design

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER



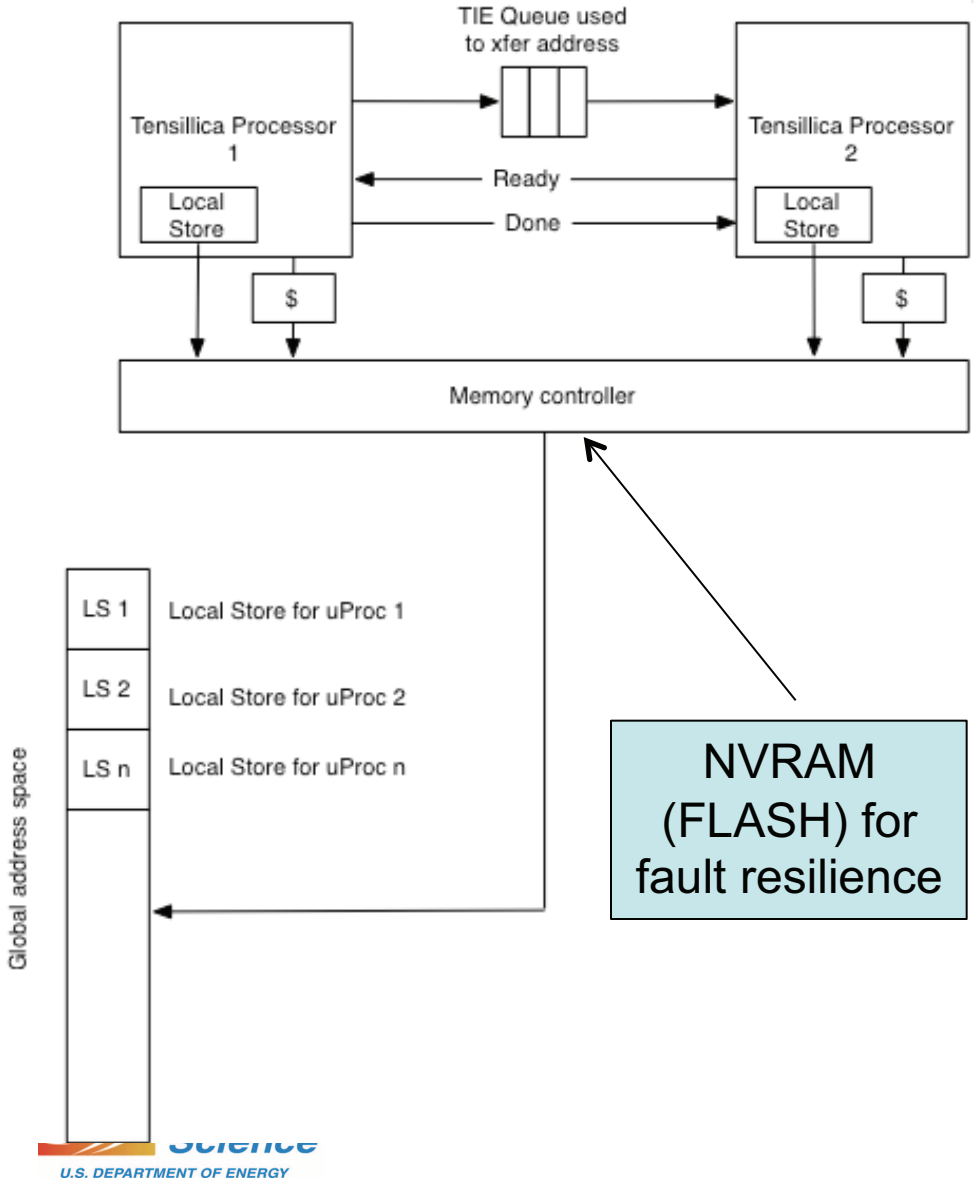


	Intel Core2 (Penryn)	Intel Atom core	Tensilica core w/ 64-bit FP
Die area (mm^2)	53.5	25	5.32
Process	45 nm	45 nm	65 nm
Power	18W	0.625W	0.091W
Freq	2930 MHz	800MHz	370MHz
Flops / Watt	162	1280	***4065***

Office of Science
U.S. DEPARTMENT OF ENERGY

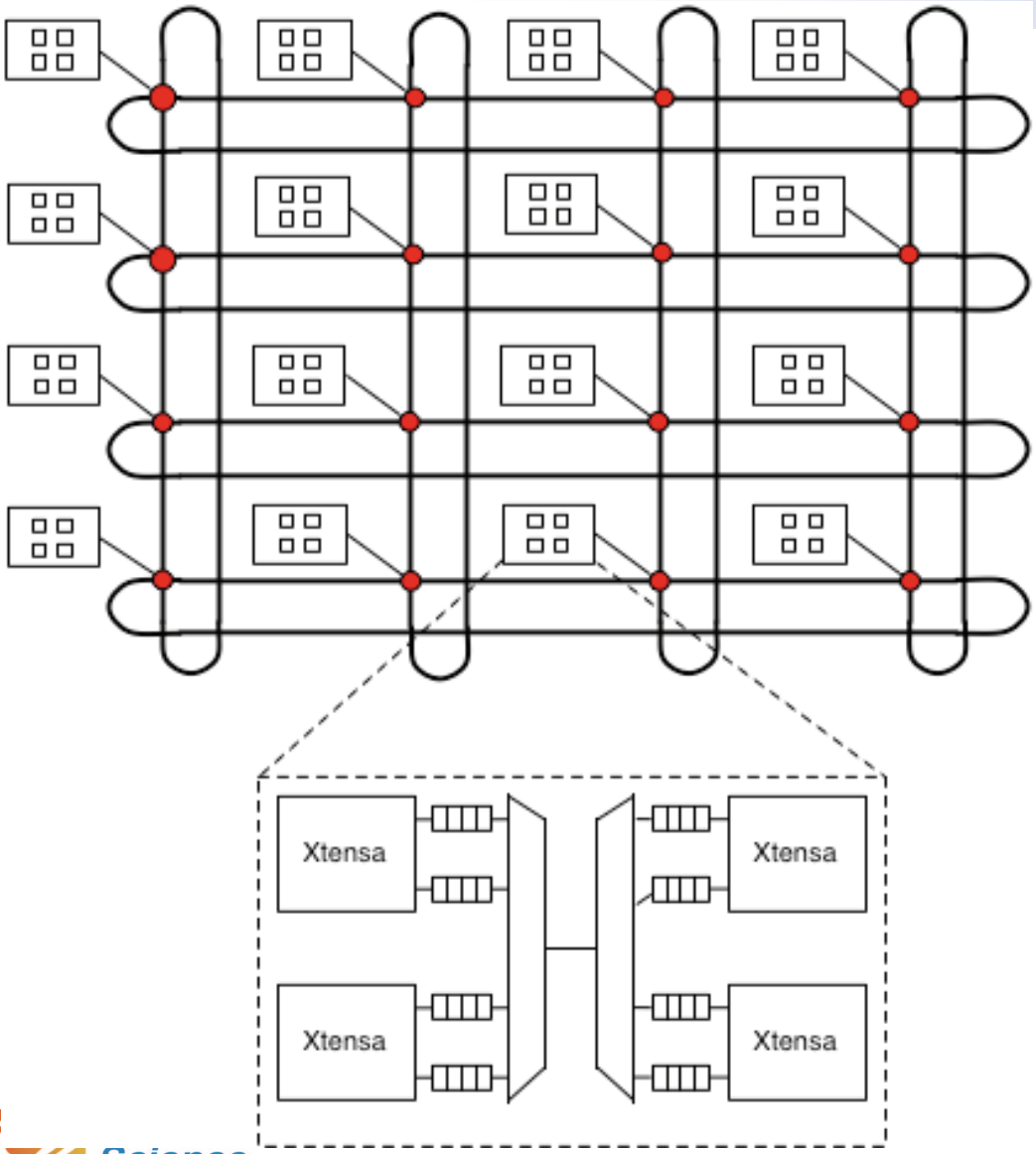
BERKELEY LAB

Architectural Support for PModels
Make hardware easier to program!



- **Logical topology is a full crossbar**
- **Each local store mapped to global address space**
- **To initiate a DMA transfer between processors:**
  - **Processors exchange starting addresses through TIE Queue interface**
    - **Optimized for small transfers**
  - **When ready, copy done directly from LS to LS**
  - **Copy will bypass cache hierarchy**

CMP Architecture - Physical View



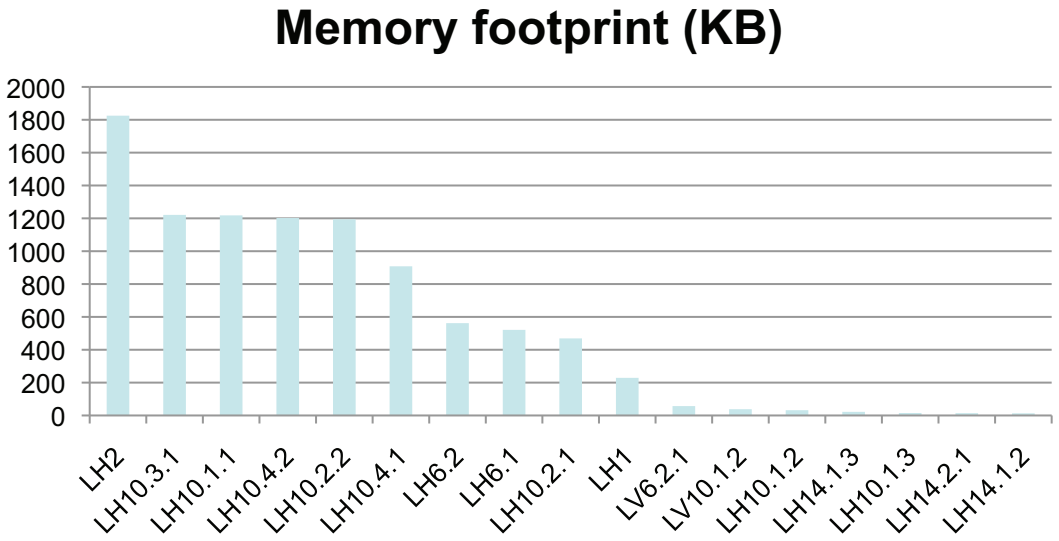
- Concentrated torus
  - Direct connect between 4 processors on a tile
  - Packet switched network connecting tiles
- Between 64 and 128 processors per die

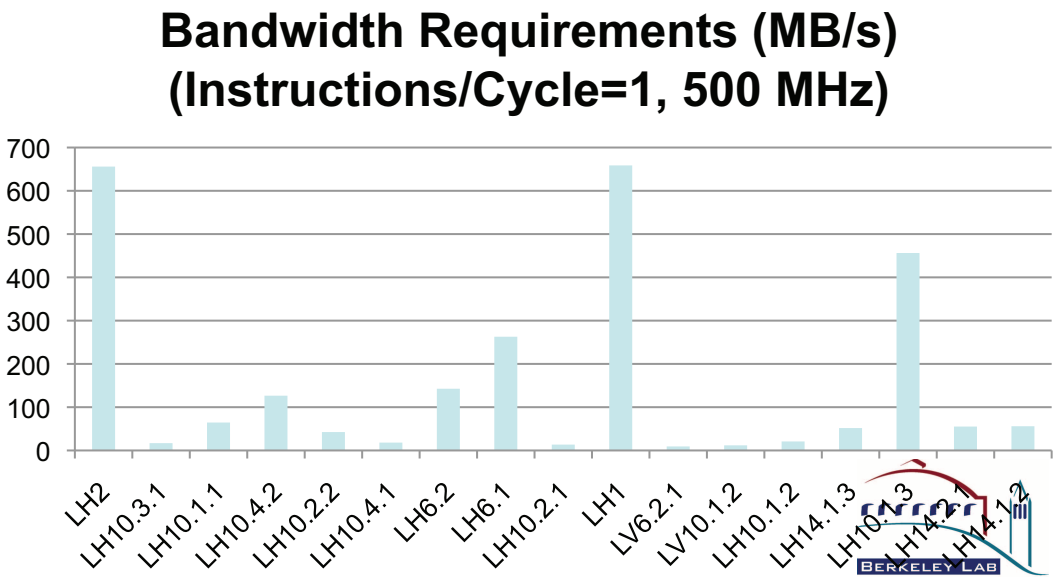
Fault Tolerance/Resilience

- Our Design does not expose unique risks
  - Faults proportional to # sockets (not # cores) and silicon surface area
  - We expose less surface area and fewer sockets with our approach
- Hard Errors
  - Spare cores in design (Cisco Metro)
  - SoC design (fewer components and fewer sockets)
  - Use solder (not sockets)
- Soft Errors
  - ECC for memory and caches
  - On-board NVRAM controller for localized checkpoint
  - Checkpoint to neighbor for rollback

Memory: Perhaps we *don't* need 1 Byte/FLOP *(Scripted Memory Movement)*

- **Trace analysis key to memory requirements**
  - **Actually running the code gives realistic values for memory footprint, temporal reuse, DRAM bandwidth requirements**
- **Memory footprint: unique addresses accessed ➔ size of local store needed**
- **Temporal reuse: maximum number of addresses which will be reused at any time ➔ size of cache needed**
- **DRAM bandwidth**
  - **(instruction throughput) X (memory footprint)/(instruction count)**

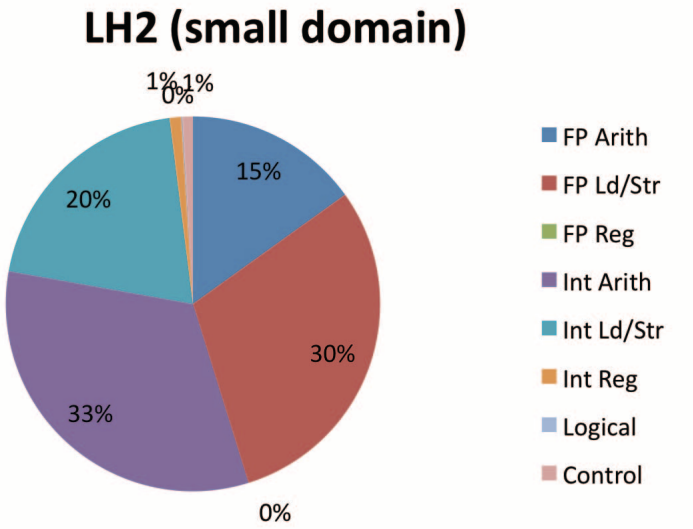


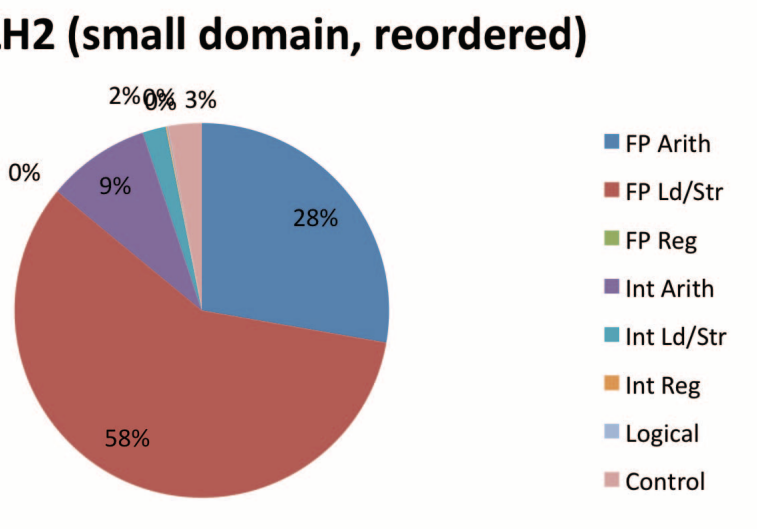


Office of Science
U.S. DEPARTMENT OF ENERGY

NERSC
BERKELEY LAB

Auto-Tuning Can Change Hardware Design Requirements





- Memory footprint: 160 KB
- Cache size requirement: 160 KB
- < 50% instructions are floating-point
  - Huge overhead for address generation
- Although code streams through data, loop ordering was bad ➔ cachelines reused although addresses were not

- Memory footprint: 160 KB
- Cache size requirement: 1 KB
- > 85% instructions are floating-point
  - Good ordering ➔ simpler addressing

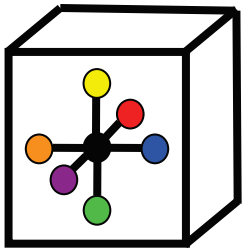
160x reduction in cache size!
2x savings in execution time

Generalized Stencil Auto-Tuning Framework

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

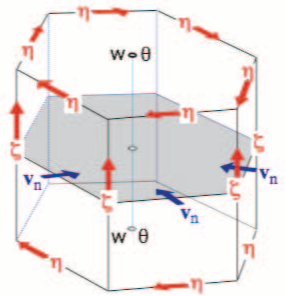
Ability to tune many stencil-like kernels

- No need to write kernel-specific perl scripts
- Uses semantic information from existing Fortran

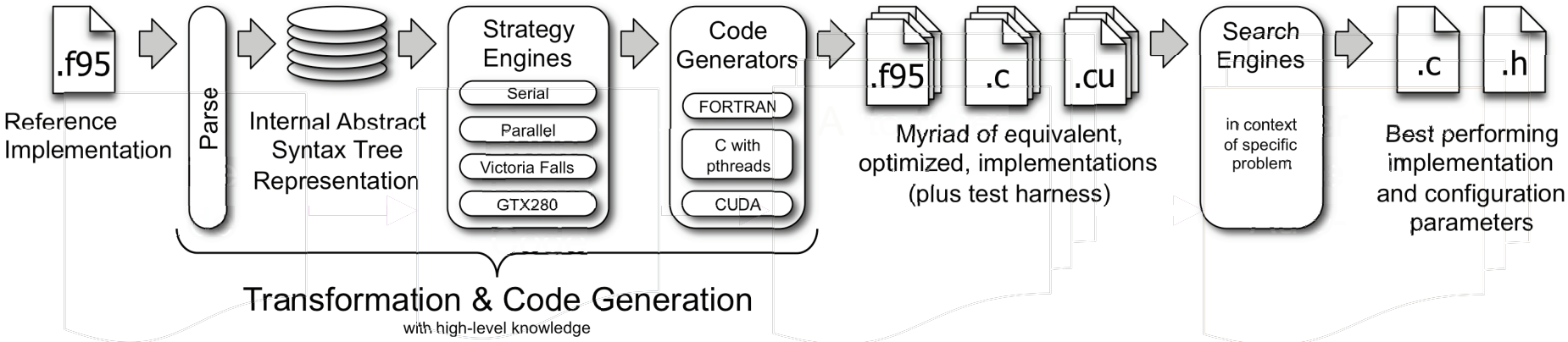


Target multiple architectures

- Search over many optimizations for each architecture
- Currently supports multi/manycore, GPUs



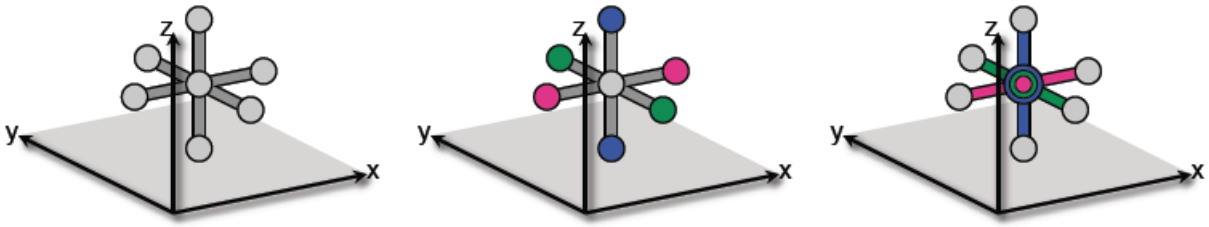
Better performance = Better energy efficiency



Office of Science U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Multi-Targeted Auto-Tuning

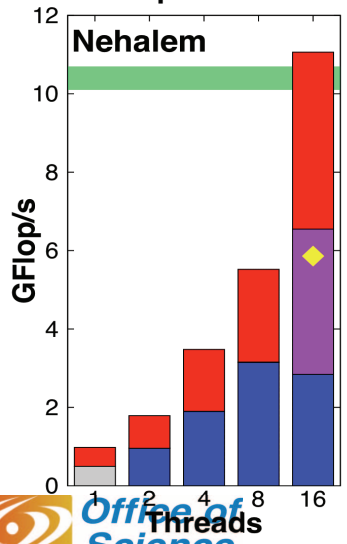



```
do k=2,nz-1,1
do j=2,ny-1,1
do i=2,nx-1,1

 uNext(i,j,k)=
  alpha*u(i,j,k)+
  beta*(u(i+1,j,k)+u(i-1,j,k)+
        u(i,j+1,k)+u(i,j-1,k)+
        u(i,j,k+1)+u(i,j,k-1)
       )
enddo
enddo
enddo
```

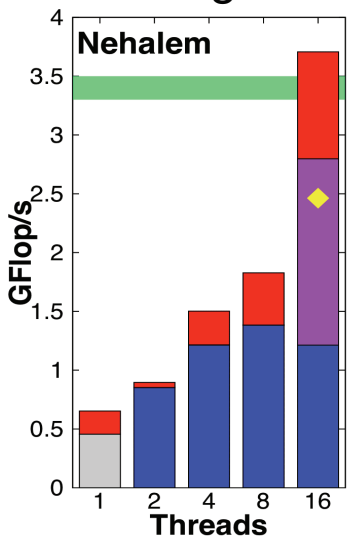
```
do k=2,nz-1,1
do j=2,ny-1,1
do i=2,nx-1,1

 u(i,j,k)=
  alpha*( x(i+1,j,k)-x(i-1,j,k) )+
   beta*( y(i,j+1,k)-y(i,j-1,k) )+
  gamma*( z(i,j,k+1)-z(i,j,k-1) )

enddo
enddo
enddo
```

```
do k=2,nz-1,1
do j=2,ny-1,1
do i=2,nx-1,1

 x(i,j,k)=alpha*( u(i+1,j,k)-u(i-1,j,k) )
 y(i,j,k)= beta*( u(i,j+1,k)-u(i,j-1,k) )
 z(i,j,k)=gamma*( u(i,j,k+1)-u(i,j,k-1) )

enddo
enddo
enddo
```

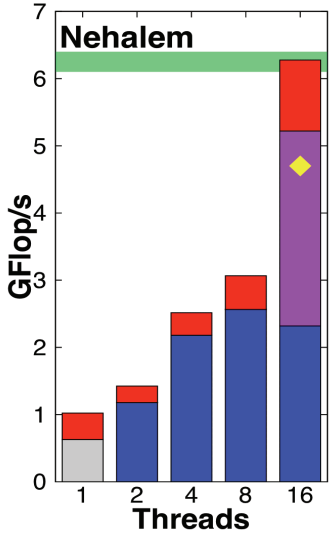
Laplacian



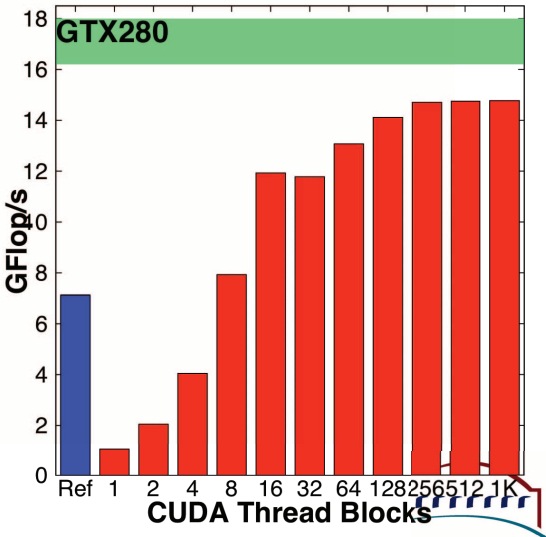
Divergence



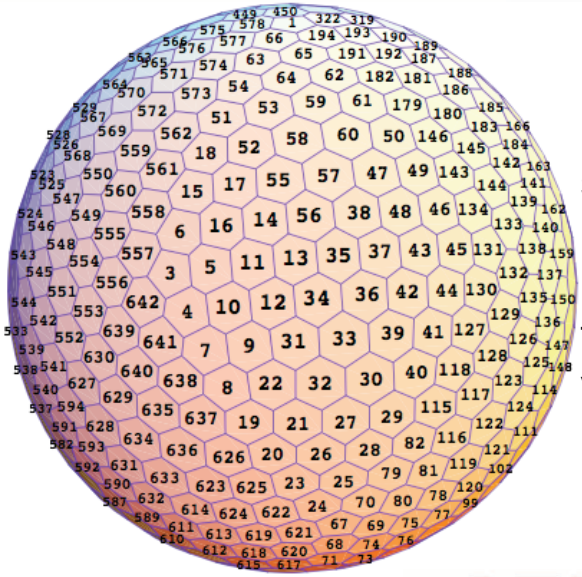
Gradient



Gradient



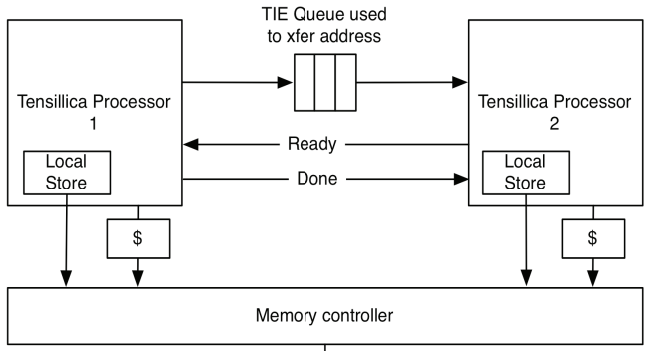
How many processors per chip

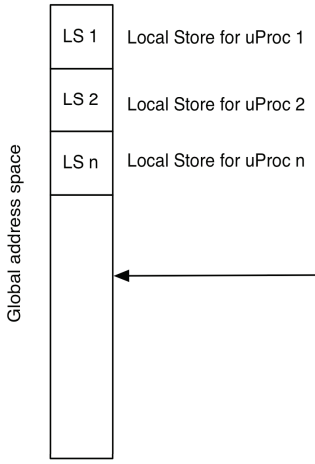


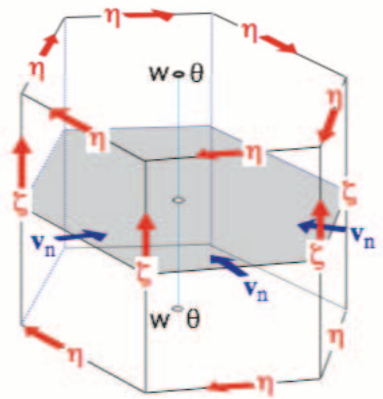
Design Trade-offs

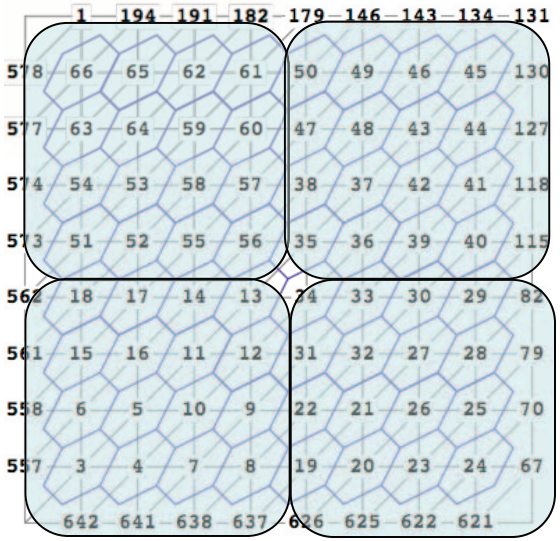
- pack fewer cores in socket to minimize memory bandwidth
- maximize cores in socket to minimize surface-to-volume ratio
- Little's Law latency hiding

Logical View of Processor Network









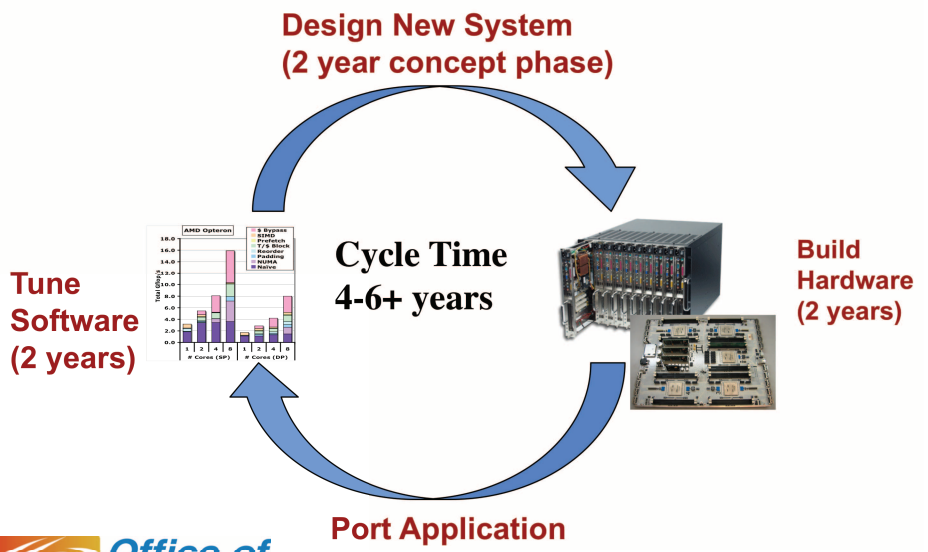
Office of Science
U.S. DEPARTMENT OF ENERGY

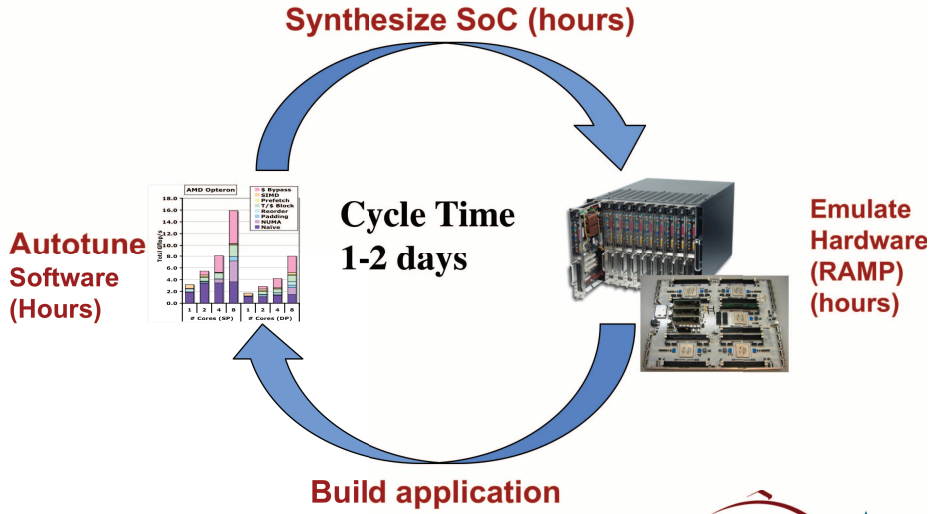
BERKELEY LAB

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Inserting Scientific Apps into the Hardware Development Process

- **Research Accelerator for Multi-Processors (RAMP)**
  - **Simulate hardware *before* it is built!**
  - **Break slow feedback loop for system designs**
  - **Enables tightly coupled hardware/software/science co-design** *(not possible using conventional approach)*



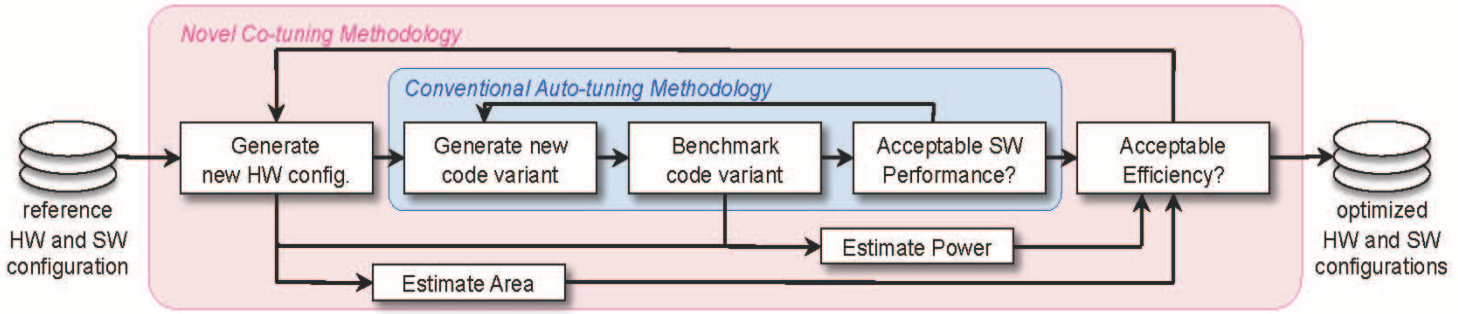


Office of Science
U.S. DEPARTMENT OF ENERGY

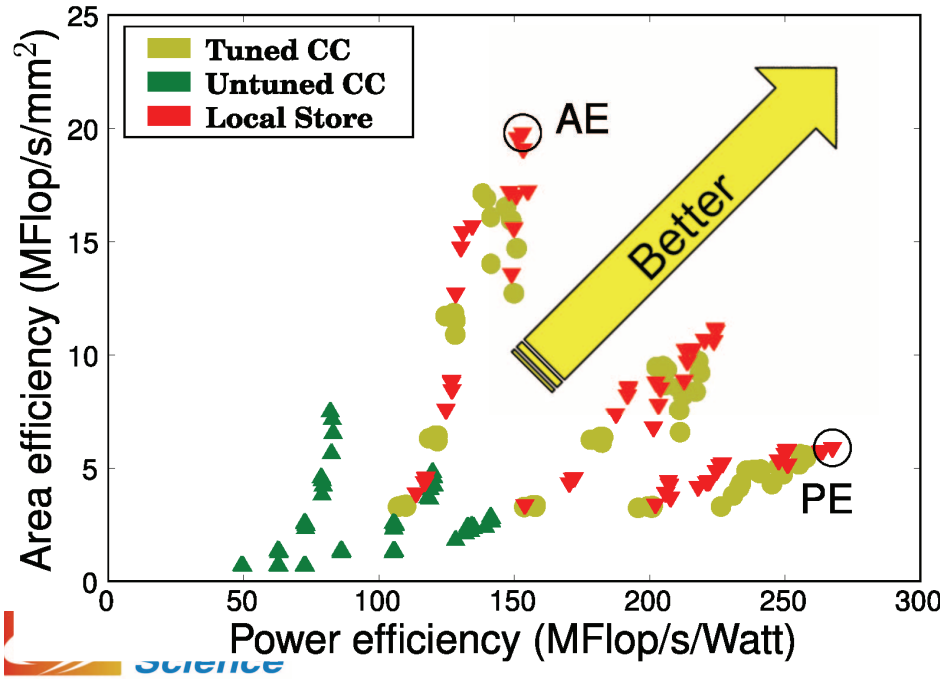
BERKELEY LAB

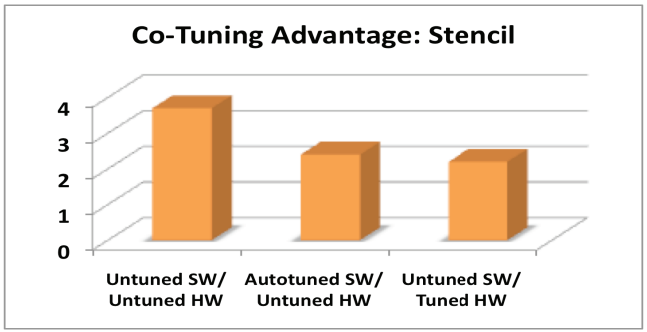
HW/SW Co-Tuning for Energy Efficiency

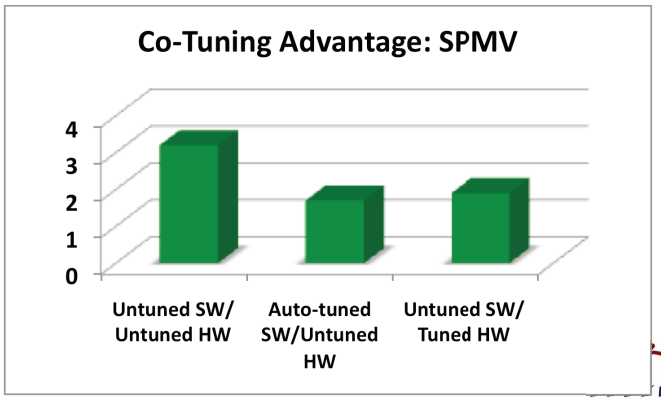
The approach: *Use auto-tuned code when evaluating architecture design points*



Co-Tuning can improve power-efficiency and area-efficiency by ~4x

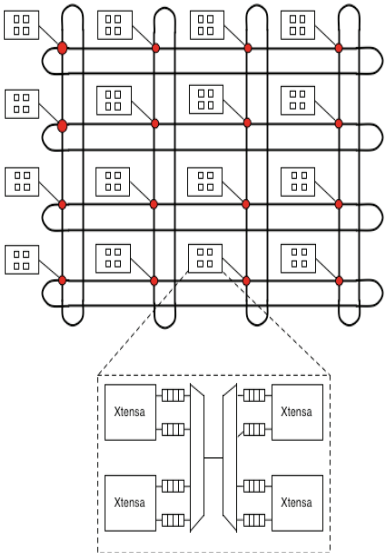


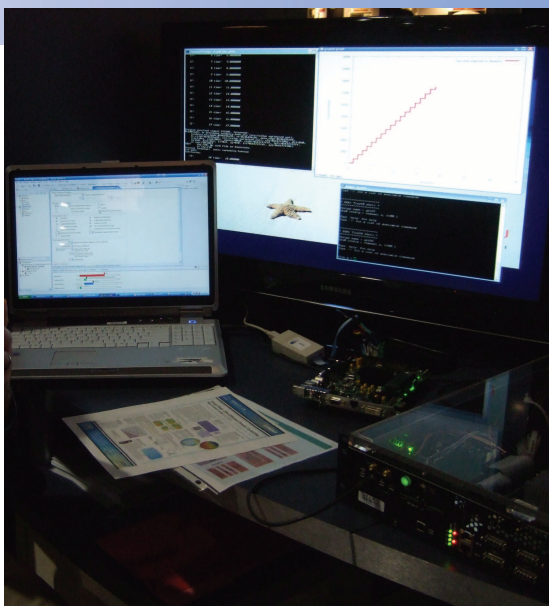




Green Flash Hardware Demo

- Demonstrated during SC '08
- Proof of concept
  - CSU atmospheric model ported to Tensilica Architecture
  - Single Tensilica processor running atmospheric model at 50MHz
- Emulation performance advantage
  - Processor running at 50MHz vs. Functional model at 100 kHz
  - 500x Speedup
- Actual climate code - not representative benchmark





Summary

- **Power is leading design constraint for future HPC**
  - **Future technology driven by handheld space**
  - **Notion of "commodity" moving on-chip**

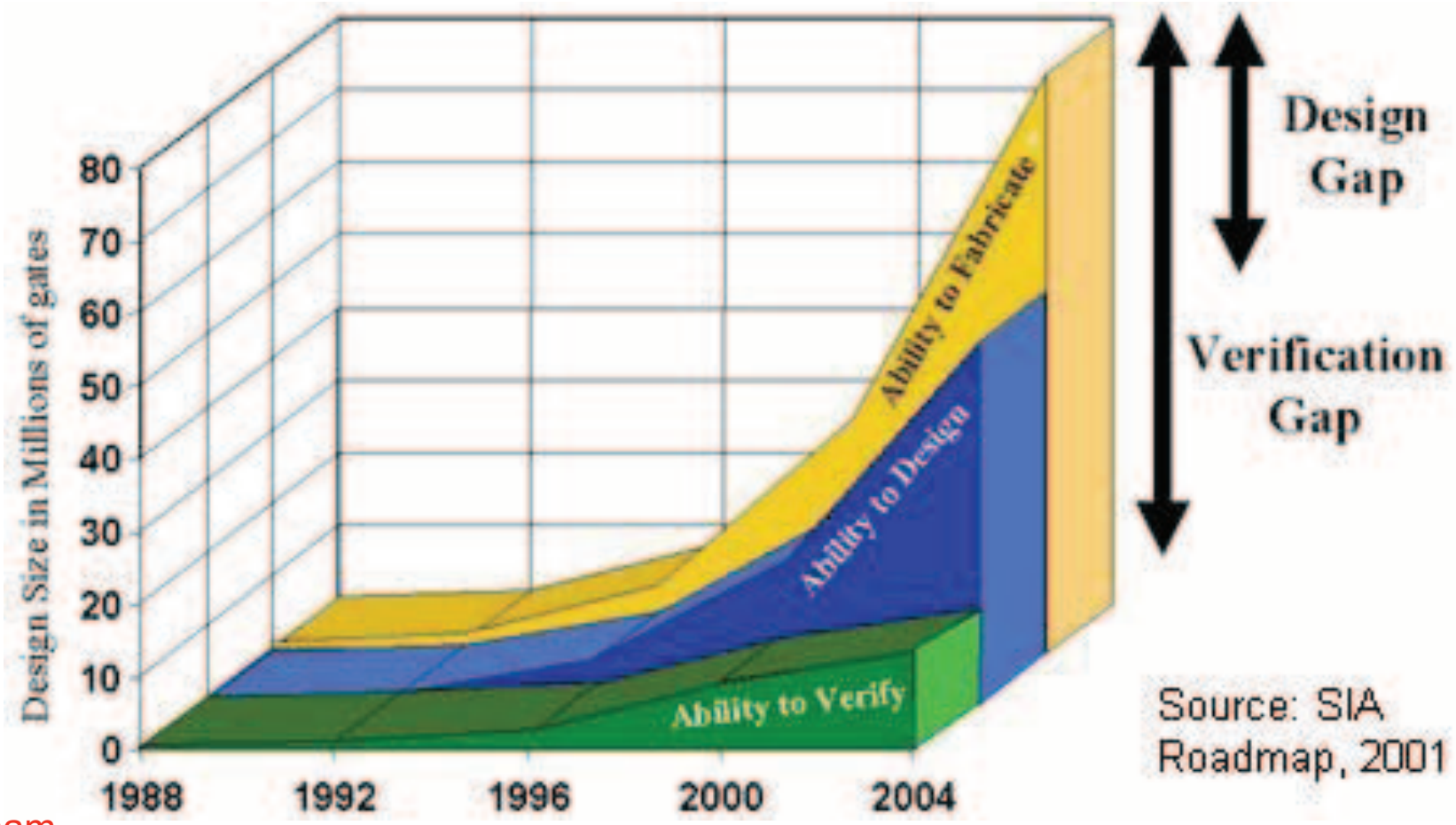
- **Approach for Power Efficient HPC**
  - **Choose the science target first** *(climate in this case)*
  - **Design systems for applications** *(rather than the reverse)*
  - **Design hardware, software, scientific algorithms together using hardware emulation and auto-tuning**
  - **This is the right way to design efficient HPC systems!**

More Info

- **Green Flash**
  - http://www.lbl.gov/CS/html/greenflash.html
  - http://www.lbl.gov/CS/html/greenmeetings.html

- **NERSC Science Driven System Architecture Group**
  - http://www.nersc.gov/projects/SDSA

The Scaling of ASIC Designs

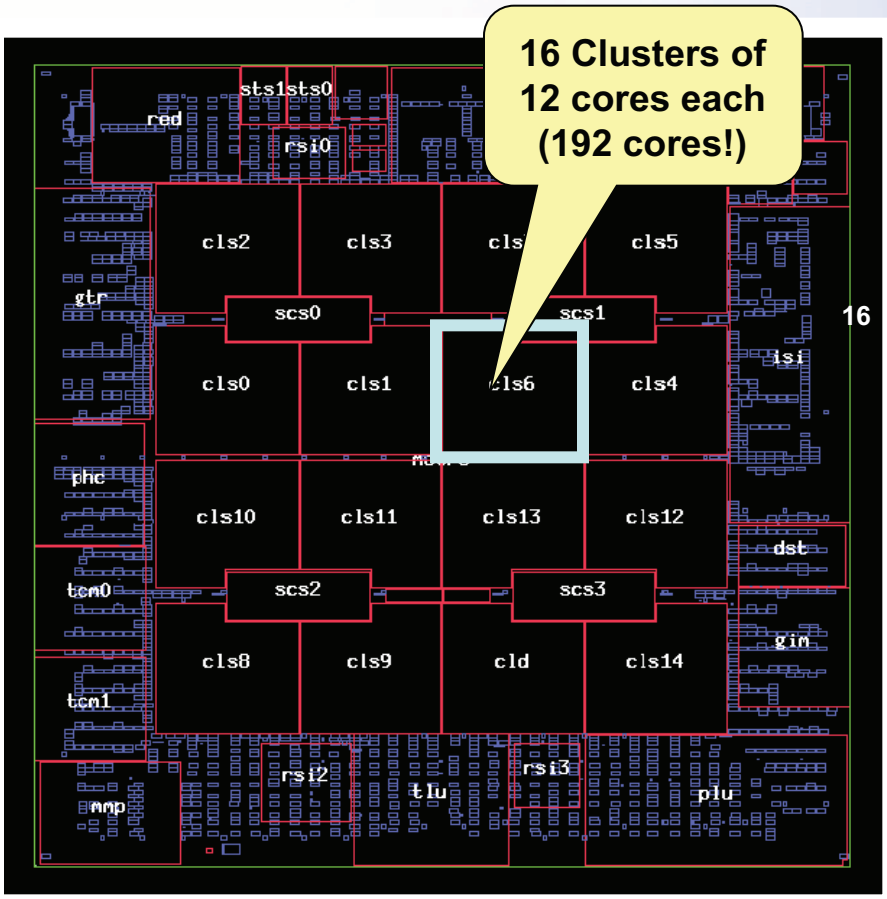


Ofer Sacham
Stanford

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Parallel Computing Everywhere

Cisco CRS-1 Terabit Router



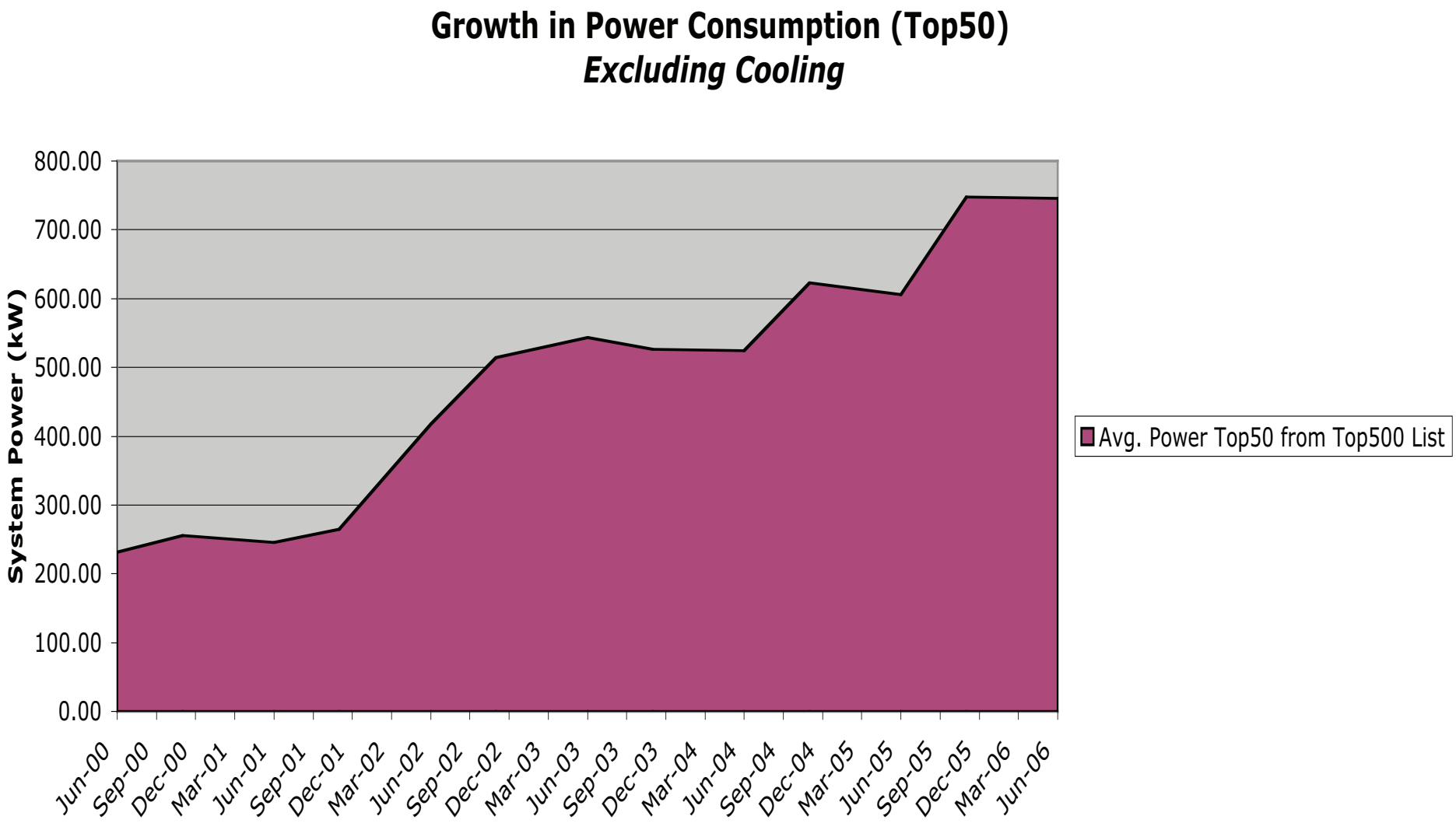
CISCO SYSTEMS

- 188+4 Xtensa general purpose processor cores per Silicon Packet Processor
- Up to 400,000 processors per system
  - (this is not just about HPC!!!)

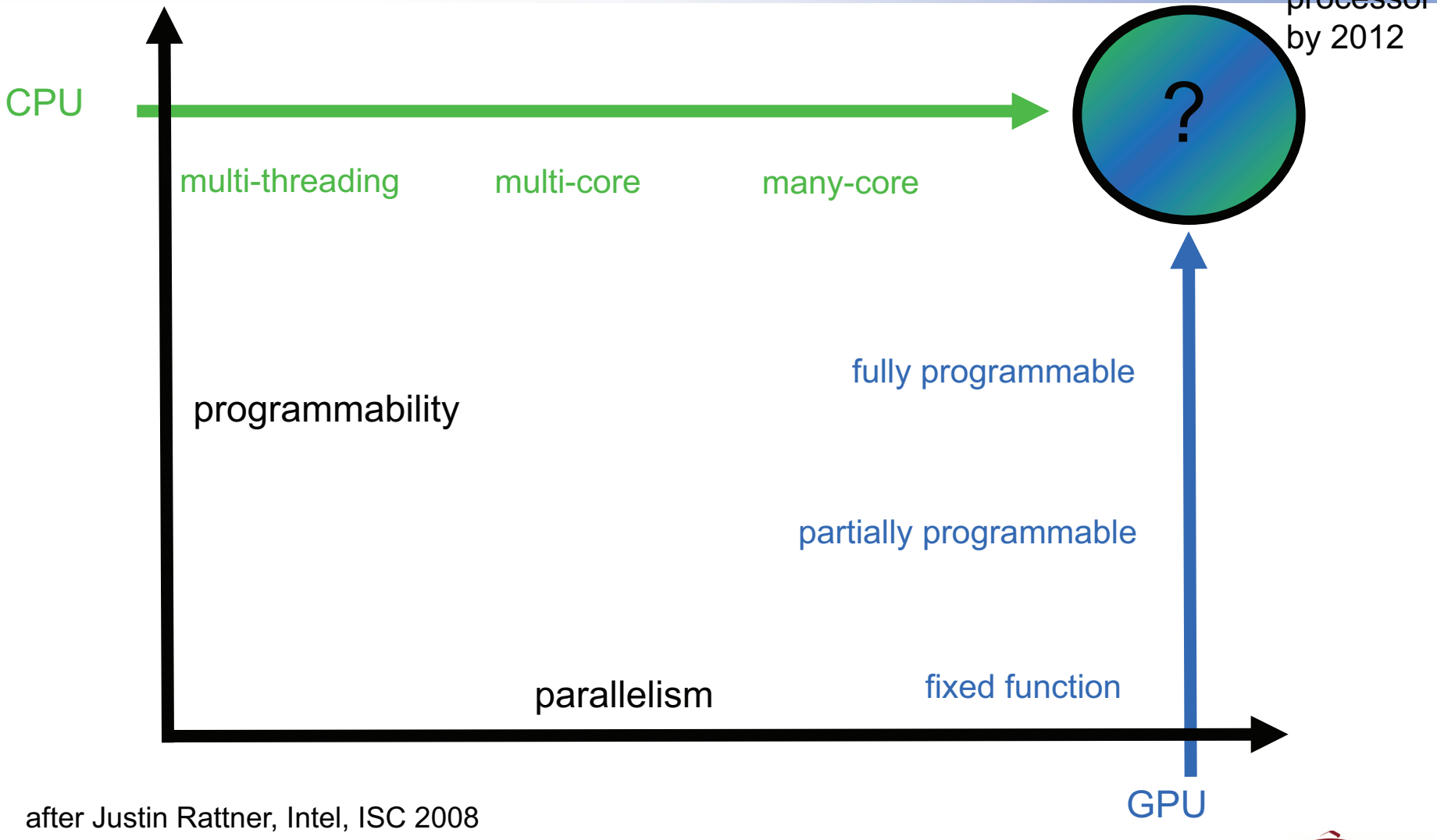
Replaces ASIC using 188 GP cores!
Emulates ASIC at competitive power/performance
Better power/performance than FPGA!
New Definition for “Custom” in SoC

Office of Science
U.S. DEPARTMENT OF ENERGY

Power Consumption by Top500 Systems



A Likely Trajectory - Collision or Convergence?



after Justin Rattner, Intel, ISC 2008

Part III

A Short diversion on Metrics

Metrics: *Can't improve what you don't measure*

- **Collecting Metrics for HPC Power consumption (Green500, Top500, SpecHPC)**
  - **Raise Community Awareness of HPC System Power Efficiency**
  - **Push vendors toward more power efficient solutions (shine a light on inefficiency)**

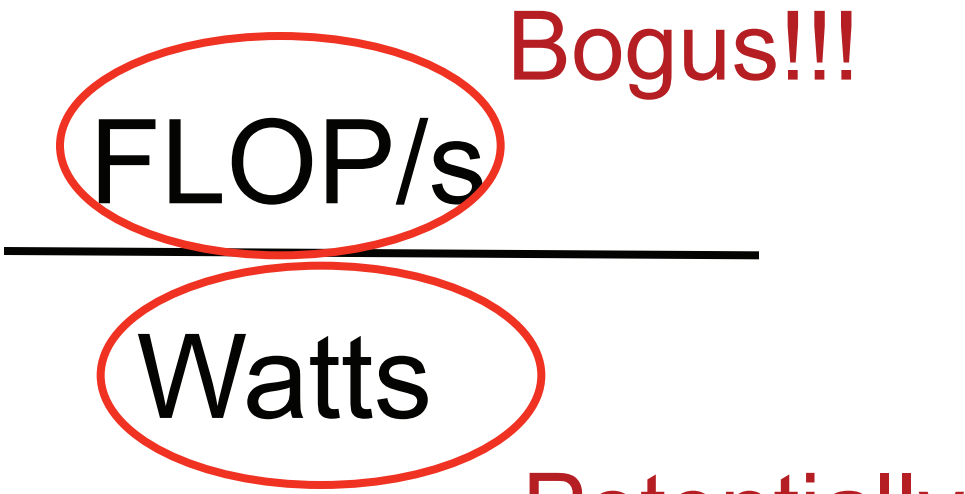
- **Choice of measurement has a dramatic effect on the outcome** ***(Law of unintended consequences)***
  - **Suddenly everything is "green"**
  - **But is anything *really* getting better? (everything looks better on an exponential curve)**

Anatomy of a "Value" Metric

Good Stuff

Bad Stuff

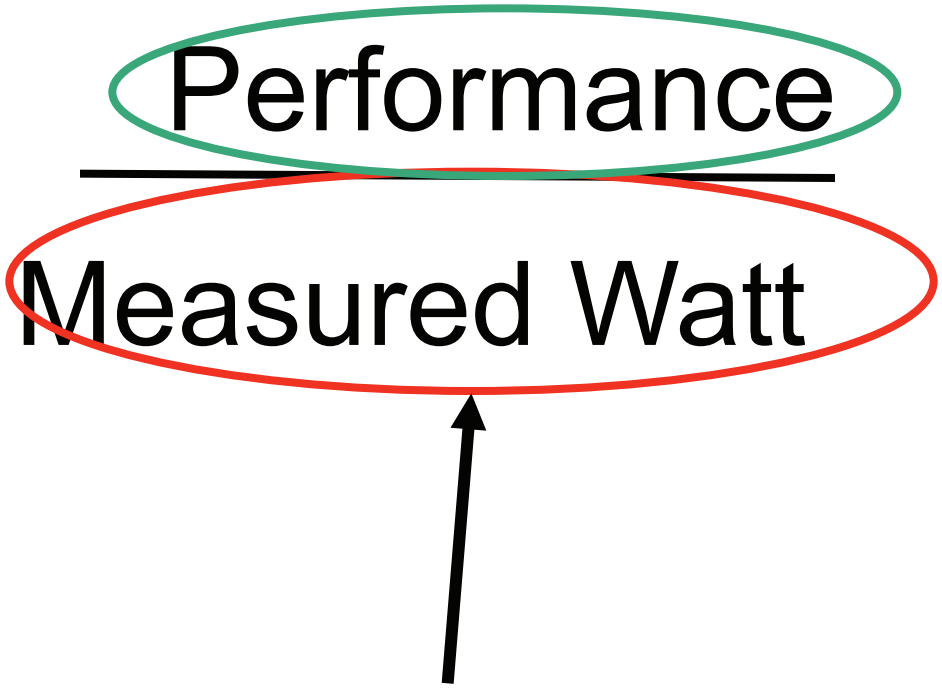
Anatomy of a "Value" Metric



Anatomy of a "Value" Metric

Choose your own metric for performance!

(doesn't need to be HPL, or FLOPS)



Formal process for collecting this data emerging
(Green500, Top500, and eventually SpecPowerHPC)

Are We Really Improving?

Performance/measured_watt

is much more useful than

FLOPs/peak_watt

But, are we getting the desired response?