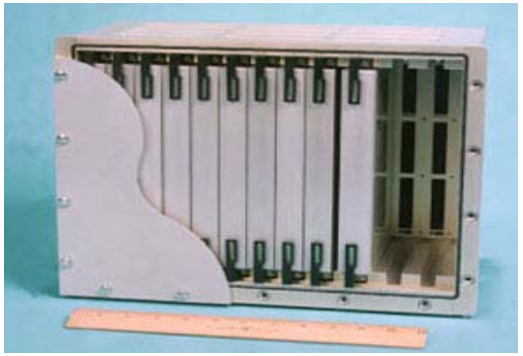
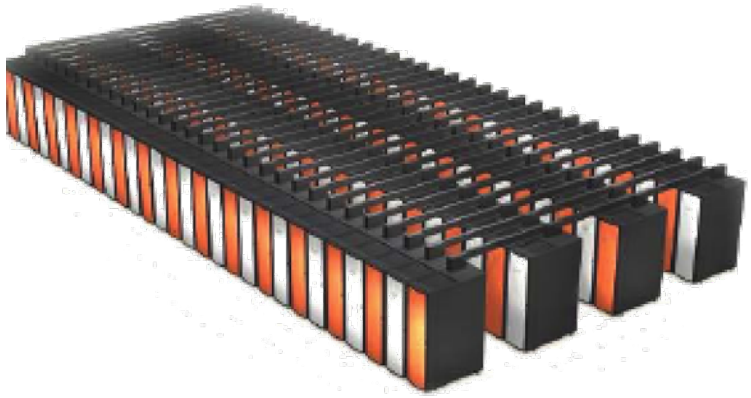
Exascale Computing: Embedded Style



TeraFlop
Embedded

PetaFlop
Departmental



ExaFlop
Data Center

Peter M. Kogge
McCourtney Chair in Computer Science & Engineering
Univ. of Notre Dame
IBM Fellow (retired)
May 5, 2009

Thesis

- **Last 30 years:**
  - "Gigascale" computing first in a single vector processor
  - "Terascale" computing first via several thousand microprocessors
  - "Petascale" computing first via several hundred thousand cores
- **Commercial technology: *to date***
  - Always shrunk prior "XXX" scale to smaller form factor
  - Shrink, with speedup, enabled next "XXX" scale
- **Space/Embedded computing has lagged far behind**
  - Environment forced implementation constraints
  - Power budget limited both clock rate & parallelism
- **"Exascale" now on horizon**
  - But beginning to suffer similar constraints as space
  - And technologies to tackle exa challenges ***very relevant***

Especially Energy/Power

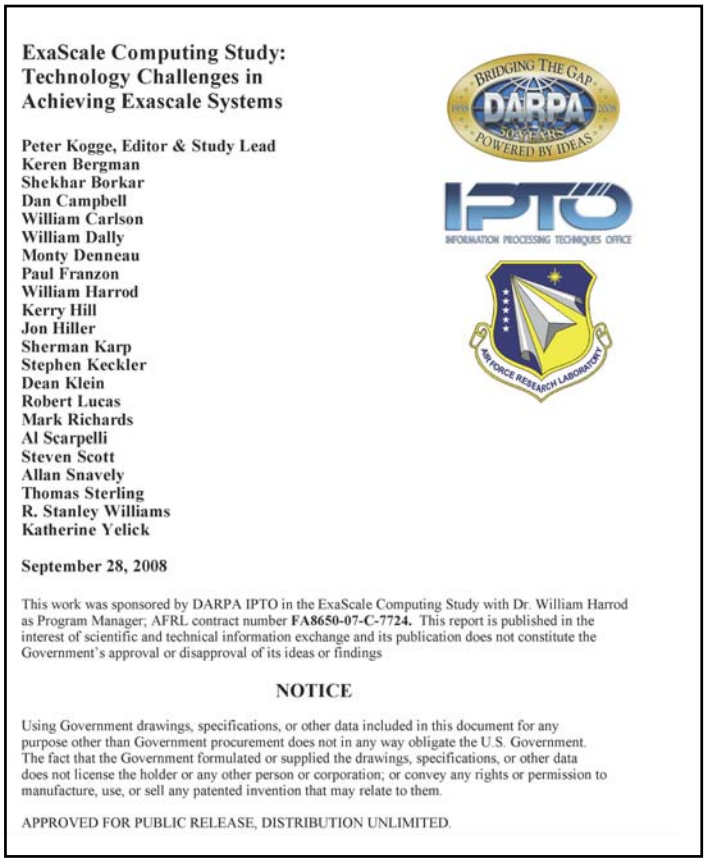
Topics

- The DARPA Exascale Technology Study
- The 3 Strawmen Designs
- A Deep Dive into Operand Access

Disclaimers

This presentation reflects my interpretation of the final report of the Exascale working group only, and not necessarily of the universities, corporations, or other institutions to which the members are affiliated.

Furthermore, the material in this document does not reflect the official views, ideas, opinions and/or findings of DARPA, the Department of Defense, or of the United States government.



ExaScale Computing Study:
Technology Challenges in
Achieving Exascale Systems

Peter Kogge, Editor & Study Lead
Keren Bergman
Shekhar Borkar
Dan Campbell
William Carlson
William Dally
Monty Denneau
Paul Franzon
William Harrod
Kerry Hill
Jon Hiller
Sherman Karp
Stephen Keckler
Dean Klein
Robert Lucas
Mark Richards
Al Scarpelli
Steven Scott
Allan Snavely
Thomas Sterling
R. Stanley Williams
Katherine Yelick

September 28, 2008

This work was sponsored by DARPA IPTO in the ExaScale Computing Study with Dr. William Harrod as Program Manager; AFRL contract number **FA8650-07-C-7724.** This report is published in the interest of scientific and technical information exchange and its publication does not constitute the Government's approval or disapproval of its ideas or findings

NOTICE

http://www.cse.nd.edu/Reports/2008/TR-2008-13.pdf

Note: Separate Exa Studies on Resiliency & Software

The Exascale Study Group

NAME	Affiliation	NAME	Affiliation
Keren Bergman	Columbia	Steve Keckler	UT-Austin
Shekhar Borkar	Intel	Dean Klein	Micron
Dan Campbell	GTRI	Peter Kogge	Notre Dame
Bill Carlson	IDA	Bob Lucas	USC/ISI
Bill Dally	Stanford	Mark Richards	Georgia Tech
Monty Denneau	IBM	Al Scarpeli	AFRL
Paul Franzon	NCSU	Steve Scott	Cray
Bill Harrod	DARPA	Allan Snavely	SDSC
Kerry Hill	AFRL	Thomas Sterling	LSU
Jon Hiller	STA	Stan Williams	HP
Sherman Karp	STA	Kathy Yelick	UC-Berkeley

11 Academic 6 Non-Academic 5 "Government"
+ Special Domain Experts

10+ Study Meetings over 2nd half 2007

The DARPA Exascale Technology Study

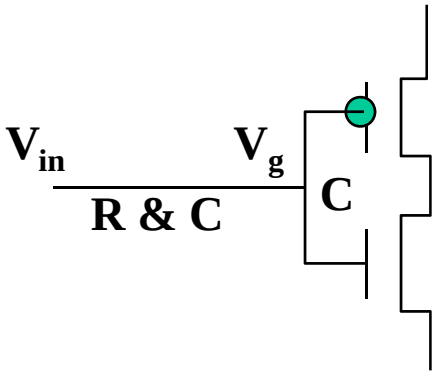
- **Exascale = *1,000X* *capability* of Petascale**
- **Exascale != Exaflops but**
  - Exascale at the data center size => ***Exaflops***
  - Exascale at the "rack" size => ***Petaflops*** for departmental systems
  - Exascale embedded => ***Teraflops*** in a cube
- **Teraflops to Petaflops took 14+ years**
  - 1st Petaflops workshop: 1994
  - Thru NSF studies, HTMT, HPCS …
  - To give us to Peta *now*
- **Study Questions:**
  - Can we ride silicon to Exa By 2015?
  - What will such systems look like?
    - Can we get 1 EF in 20 MW & 500 racks?
  - Where are the Challenges?

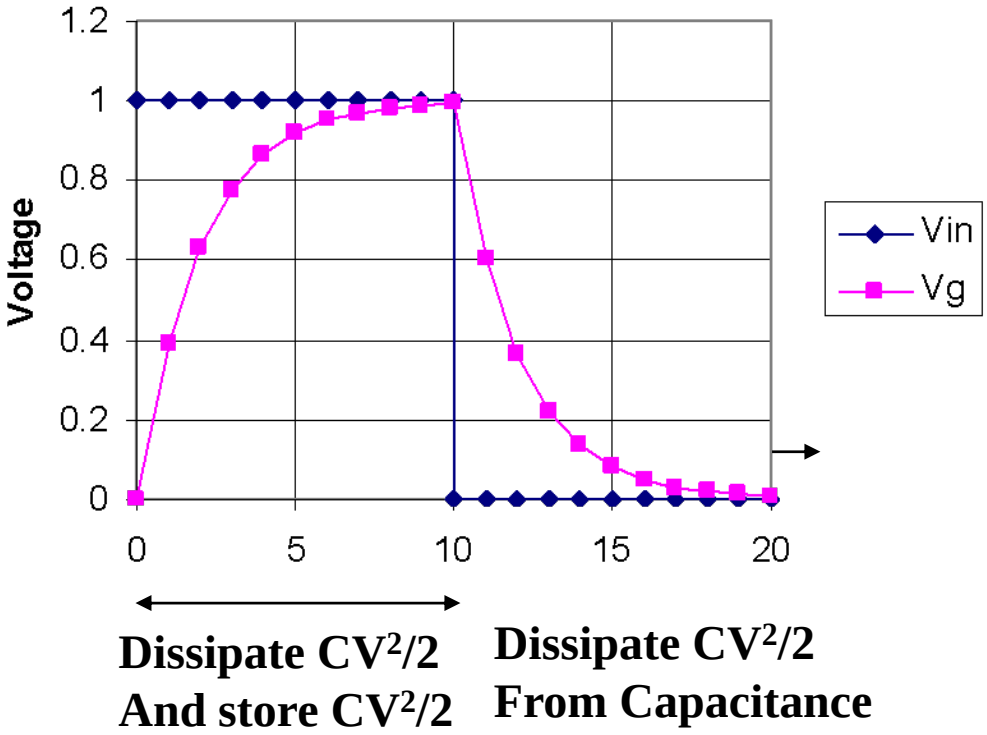
The Study's Approach

- **Baseline today's:**
  - Commodity Technology
  - Architectures
  - Performance (Linpack)
- **Articulate scaling of potential application classes**
- **Extrapolate roadmaps for**
  - "Mainstream" technologies
  - Possible offshoots of mainstream technologies
  - Alternative and emerging technologies
- **Use technology roadmaps to extrapolate use in "strawman" designs**
- **Analyze results & Id "Challenges"**

Context:
Focus on Energy
Not Power

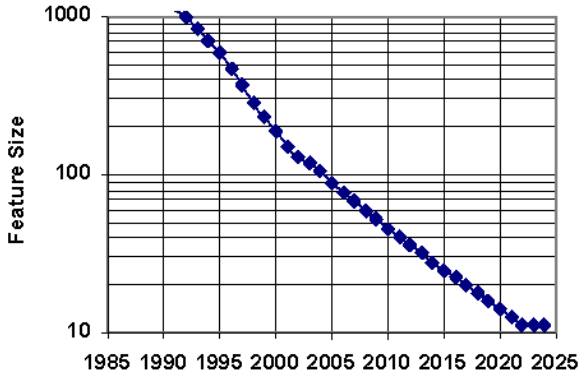
CMOS Energy 101



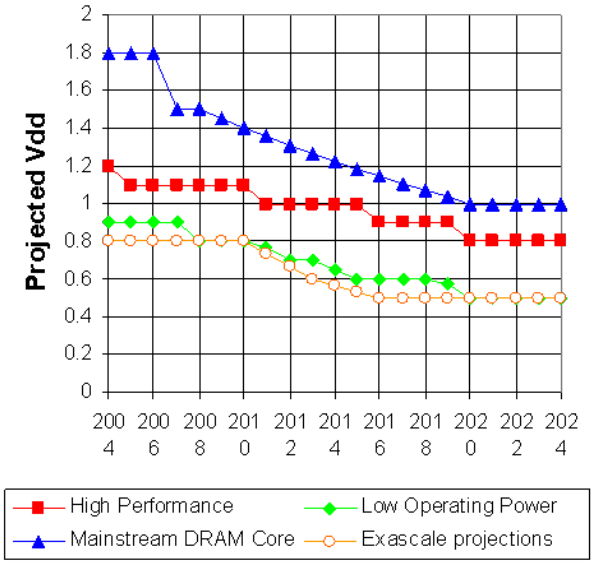


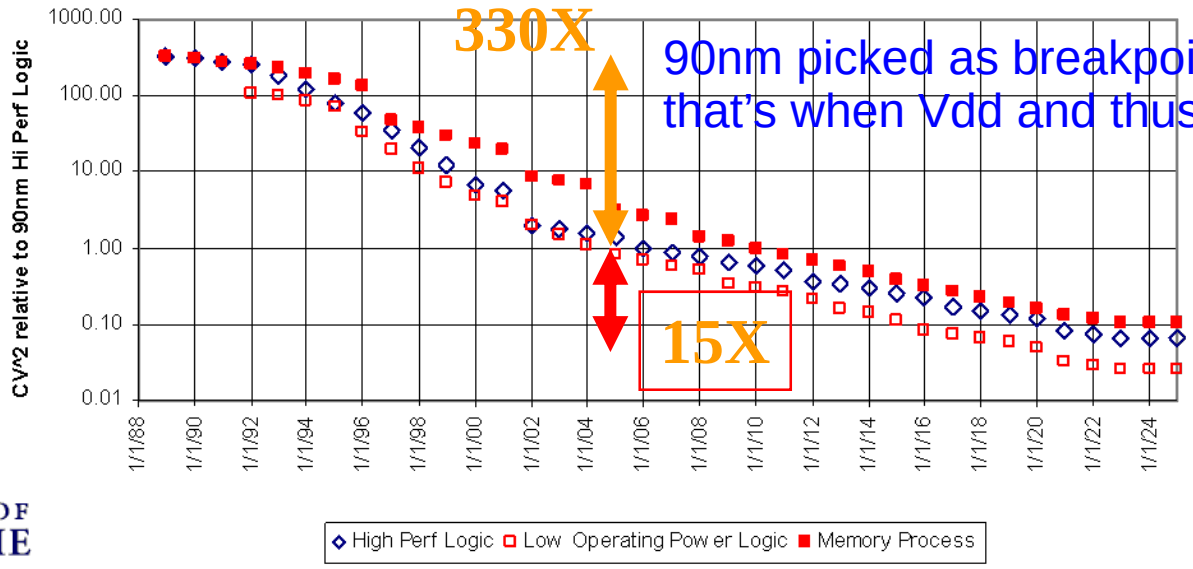
One clock cycle dissipates $C*V^2$

ITRS Projections



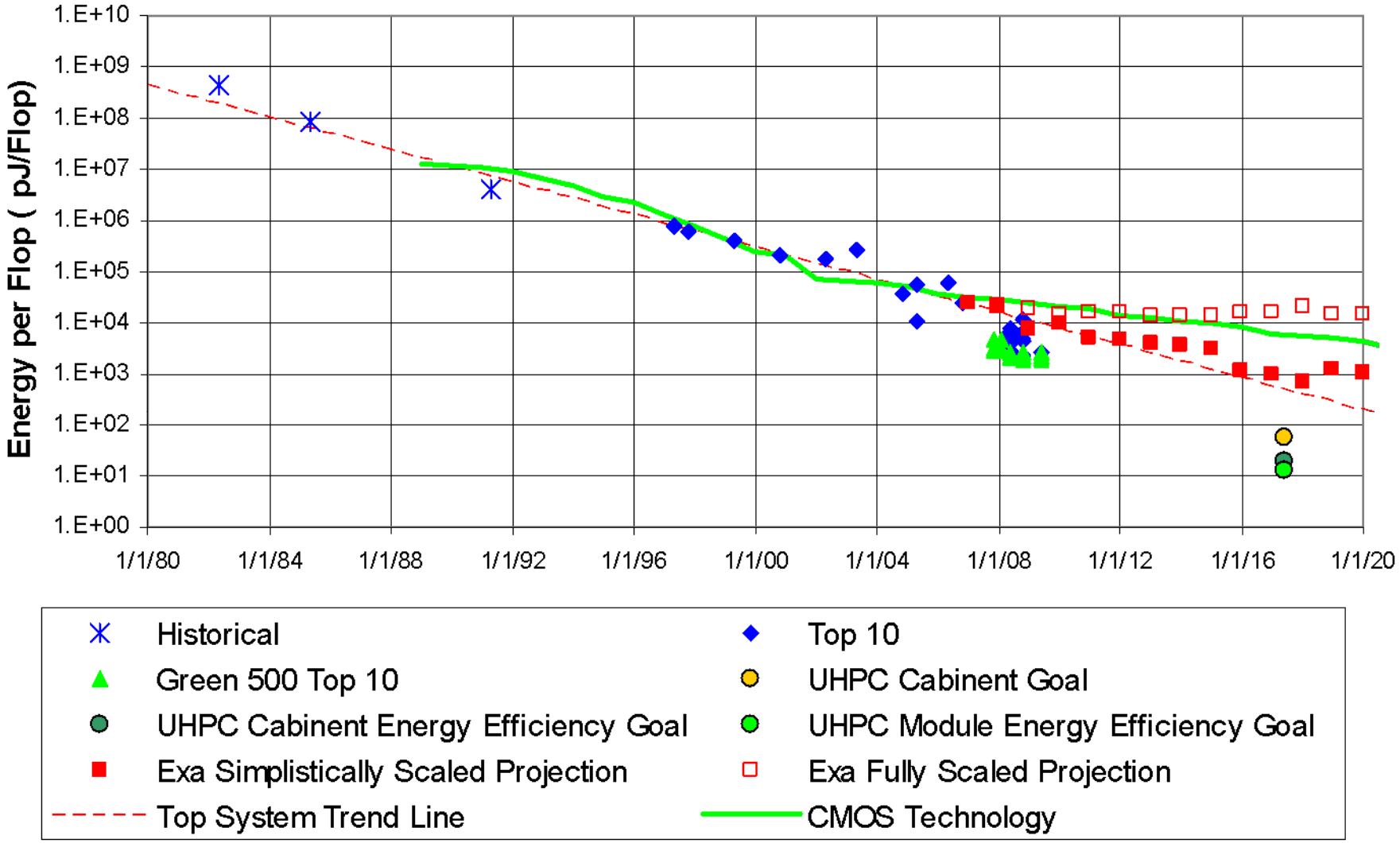
Assume capacitance of a circuit scales as feature size





90nm picked as breakpoint because that's when Vdd and thus clocks flattened

Energy Efficiency

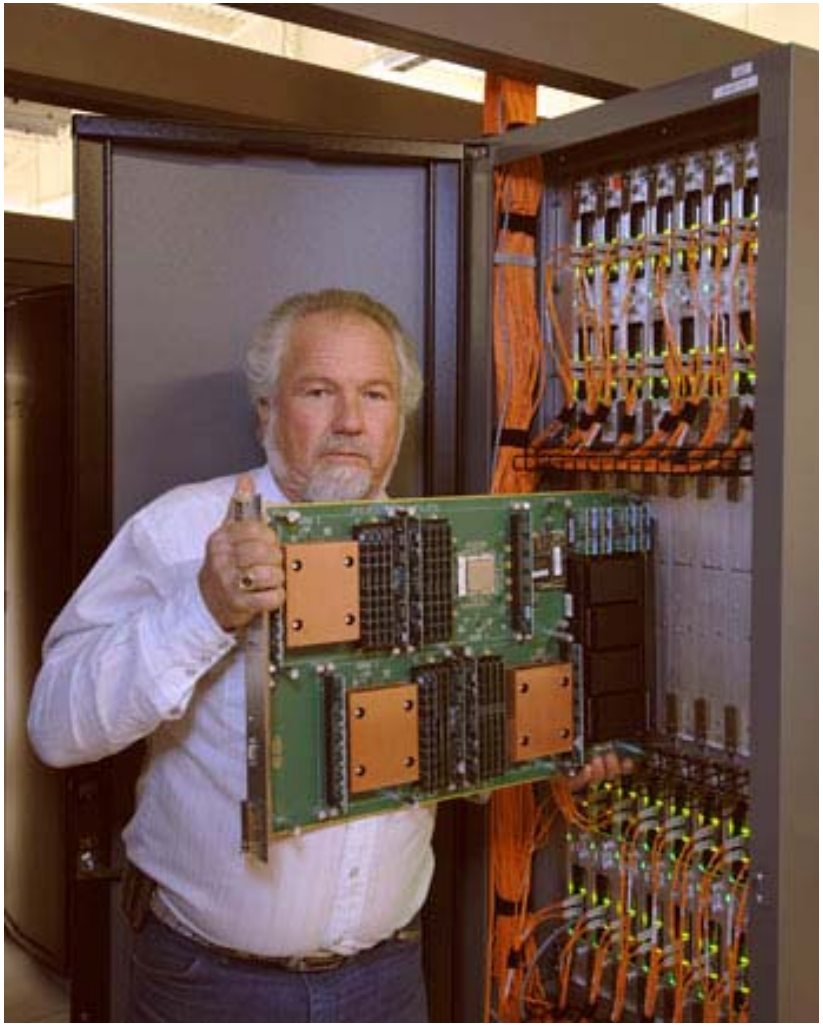


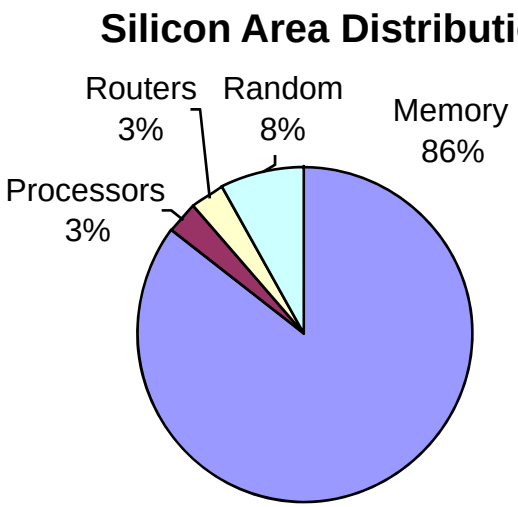
The 3 Study Strawmen

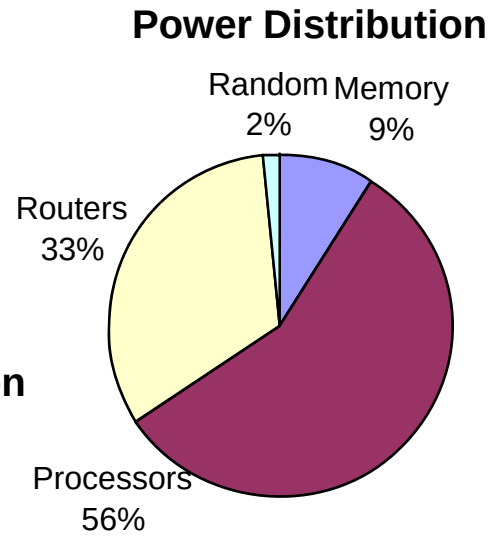
Architectures Considered

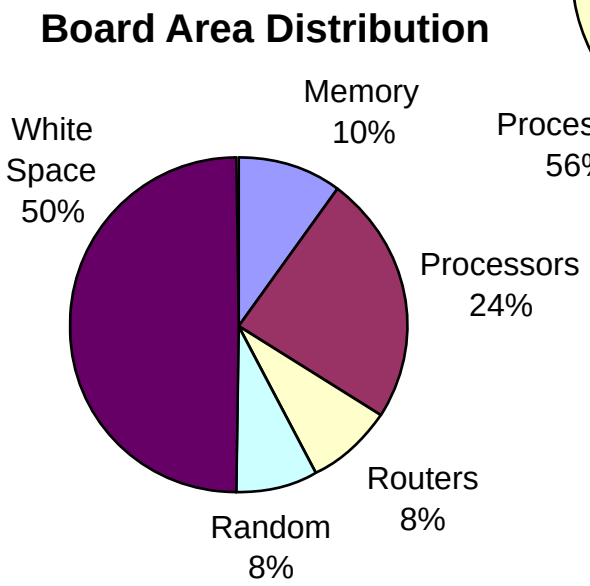
- **Evolutionary Strawmen**
  - **"Heavyweight" Strawman based on commodity-derived microprocessors**
  - **"Lightweight" Strawman based on custom microprocessors**
- **Aggressive Strawman**
  - **"Clean Sheet of Paper" CMOS Silicon**

A Modern HPC System





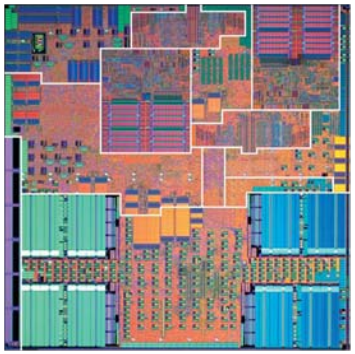




A “Light Weight” Node Alternative

2 Nodes per “Compute Card.” Each node:
- **A low power compute chip**
- **Some memory chips**
- **“Nothing Else”**

System Architecture:
- **Multiple Identical Boards/Rack**
- **Each board holds multiple Compute Cards**
- **“Nothing Else”**



- **2 simple dual issue cores**
- **Each with dual FPUs**
- **Memory controller**
- **Large eDRAM L3**
- **3D message interface**
- **Collective interface**
- **All at subGHz clock**

“Packaging the Blue Gene/L supercomputer,” IBM J. R&D, March/May 2005
“Blue Gene/L compute chip: Synthesis, timing, and physical design,” IBM J. R&D, March/May 2005

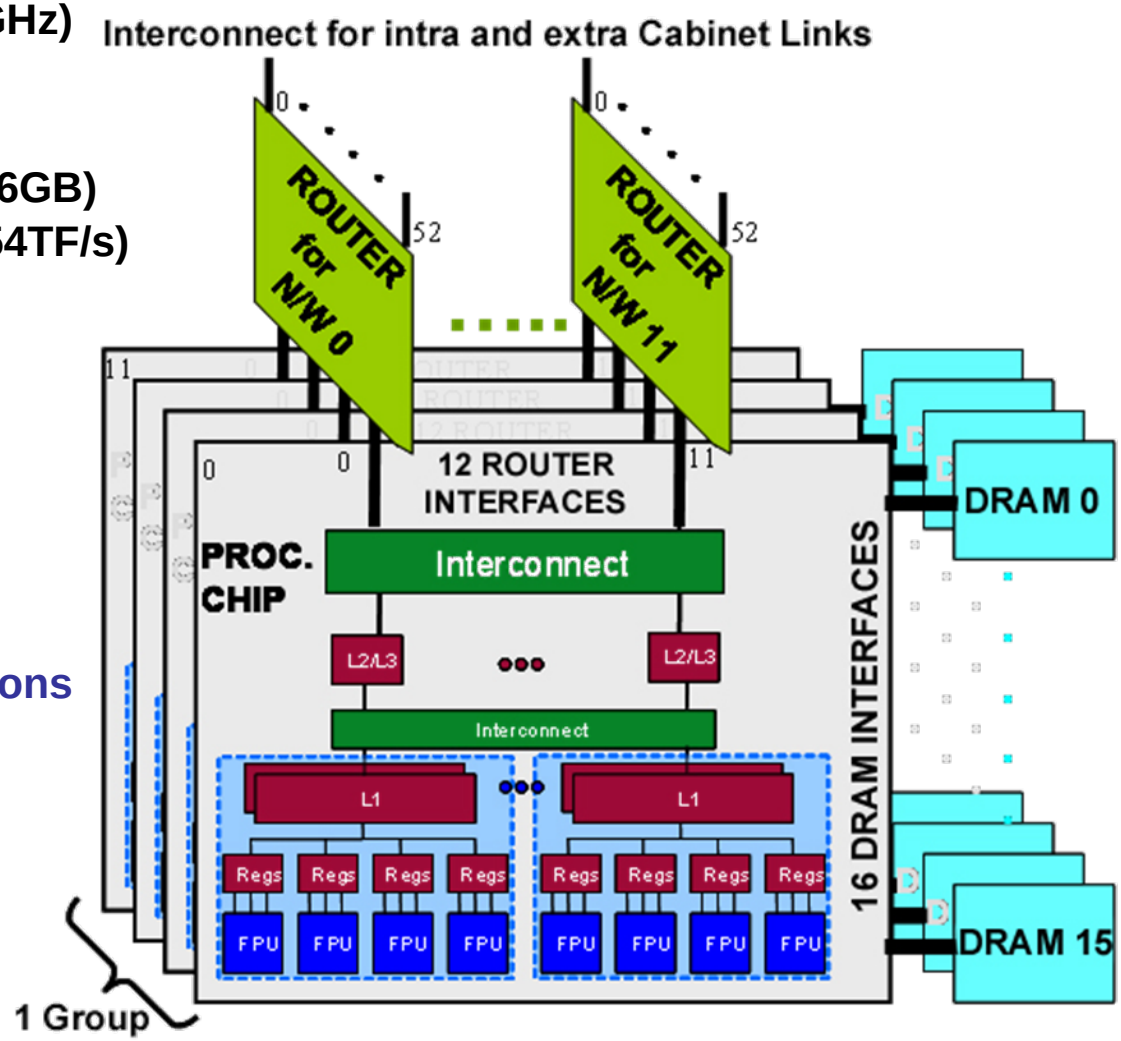
Possible System Power Models: Interconnect Driven

- **Simplistic**: A highly optimistic model
  - Max power per die grows as per ITRS
  - Power for memory grows *only linearly* with # of chips
    - Power per memory chip remains constant
  - Power for routers and common logic remains constant
    - Regardless of obvious need to increase bandwidth
  - True if energy for bit moved/accessed ***decreases*** as fast as "flops per second" increase
- **Fully Scaled**: A pessimistic model
  - Same as Simplistic, except memory & router power grow with peak flops per chip
  - True if energy for bit moved/accessed ***remains constant***
- **Real world: somewhere in between**

1 EFlop/s "Clean Sheet of Paper" Strawman

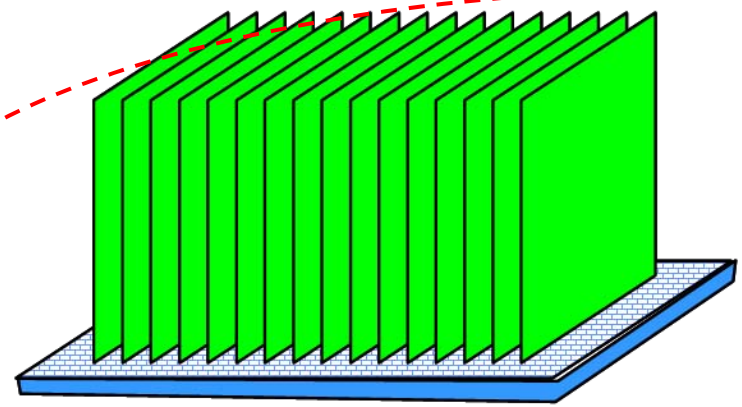
Sizing done by "balancing" power budgets with achievable capabilities

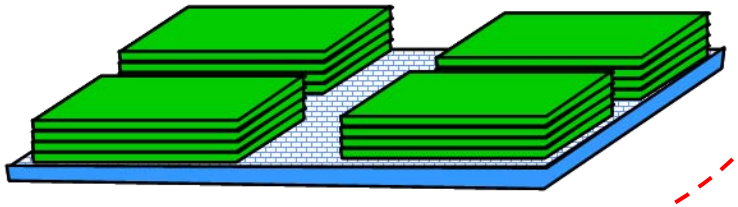
- 4 FPUs+RegFiles/Core (=6 GF @1.5GHz)
- 1 Chip = *742* Cores (=4.5TF/s)
  - 213MB of L1I&D; 93MB of L2
- 1 Node = 1 Proc Chip + 16 DRAMs (16GB)
- 1 Group = 12 Nodes + 12 Routers (=54TF/s)
- 1 Rack = 32 Groups (=1.7 PF/s)
  - 384 nodes / rack
- 3.6EB of Disk Storage included
- 1 System = 583 Racks (=1 EF/s)
  - 166 MILLION cores
  - 680 MILLION FPUs
  - 3.6PB = 0.0036 bytes/flops
  - 68 MW w'aggressive assumptions



Largely due to Bill Dally, Stanford

A Single Node (No Router)

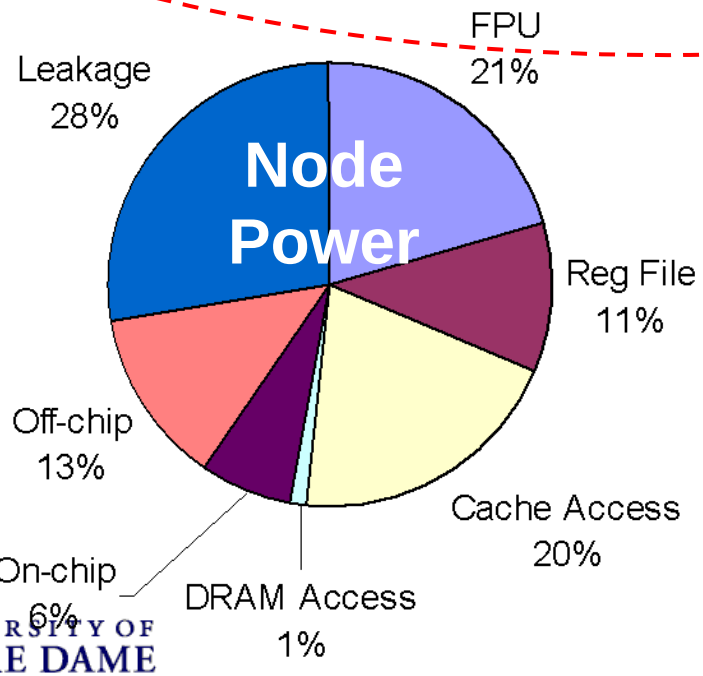




"Stacked" Memory

(a) Quilt Packaging

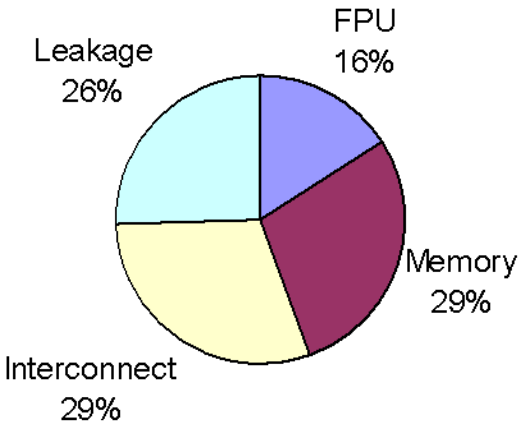
(b) Thru via chip stack



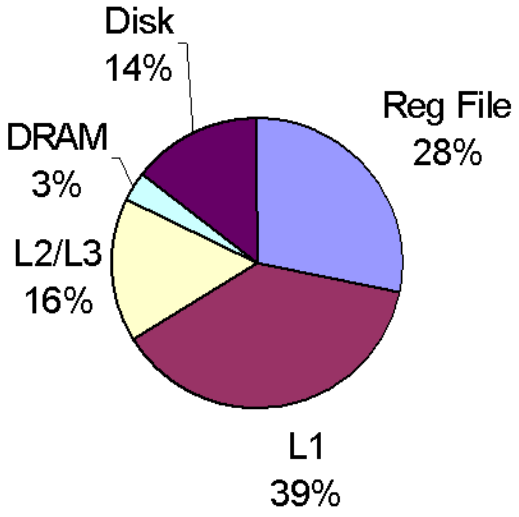
Characteristics:

- **742 Cores; 4 FPUs/core**
- **16 GB DRAM**
- **290 Watts**
- **1.08 TF Peak @ 1.5GHz**
- **~3000 Flops per cycle**

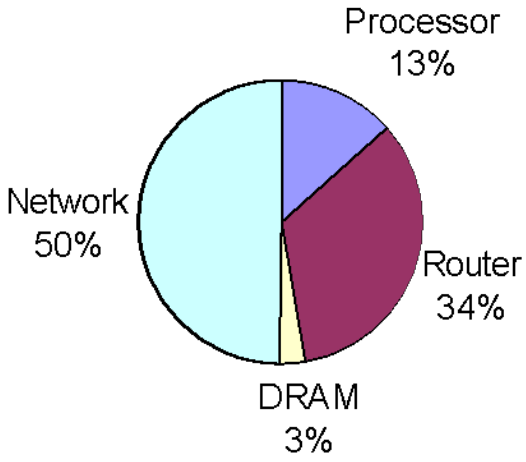
1 Eflops Aggressive Strawman Data Center Power Distribution



(a) Overall System Power



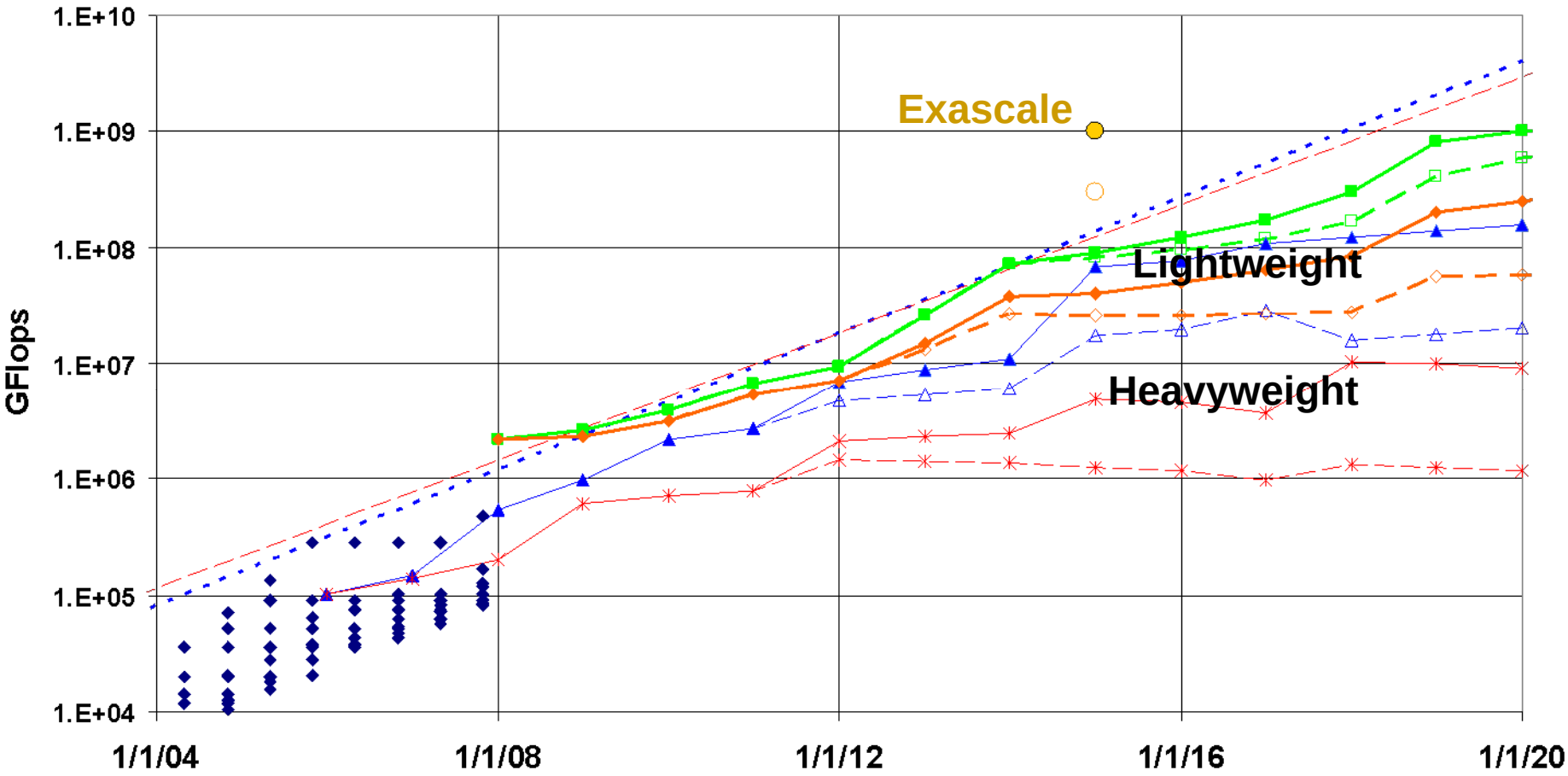
(b) Memory Power



(c) Interconnect Power

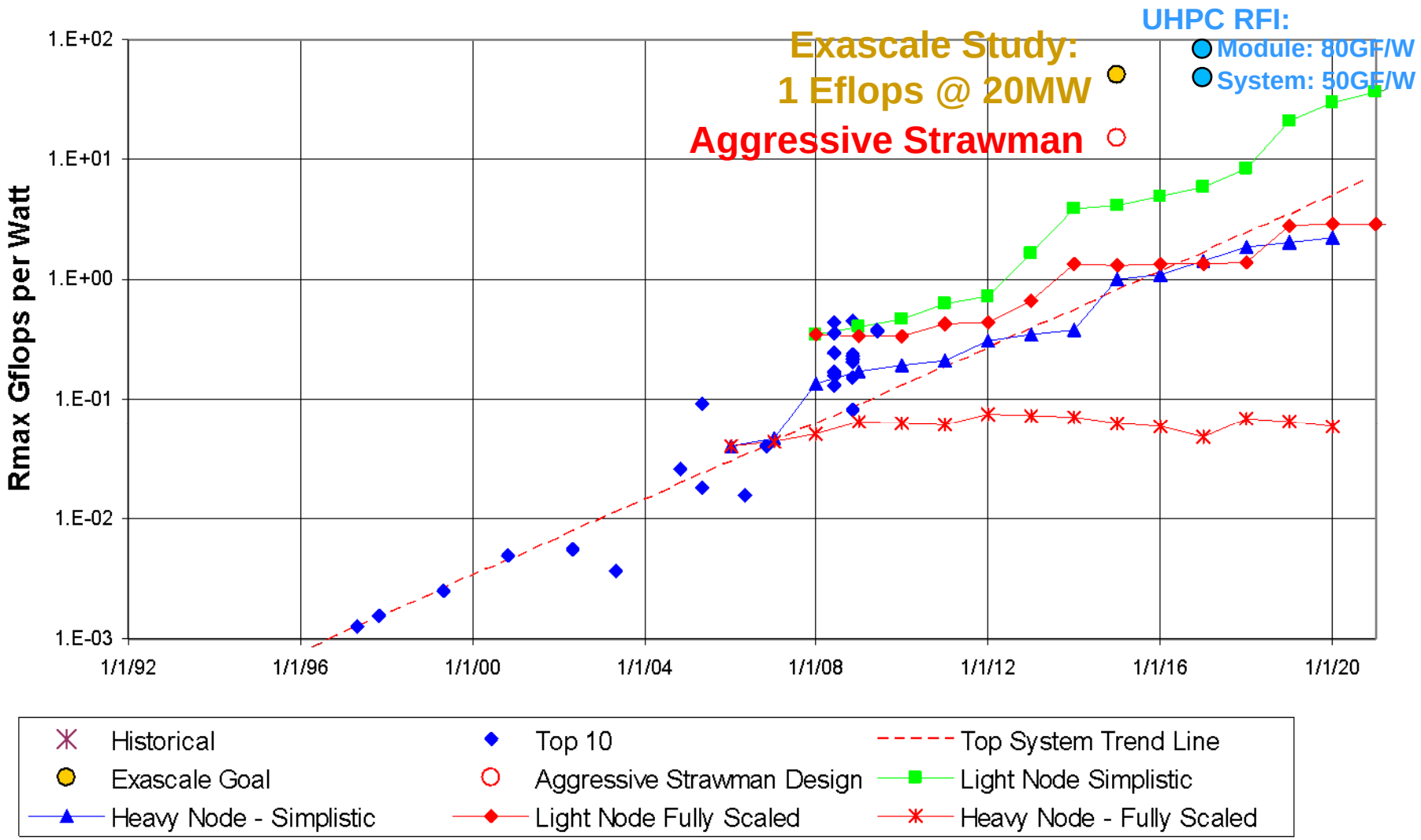
- 12 nodes per group
- 32 groups per rack
- 583 racks
- 1 EFlops/3.6 PB
- ***166 million cores***
- ***67 MWatts***

Data Center Performance Projections

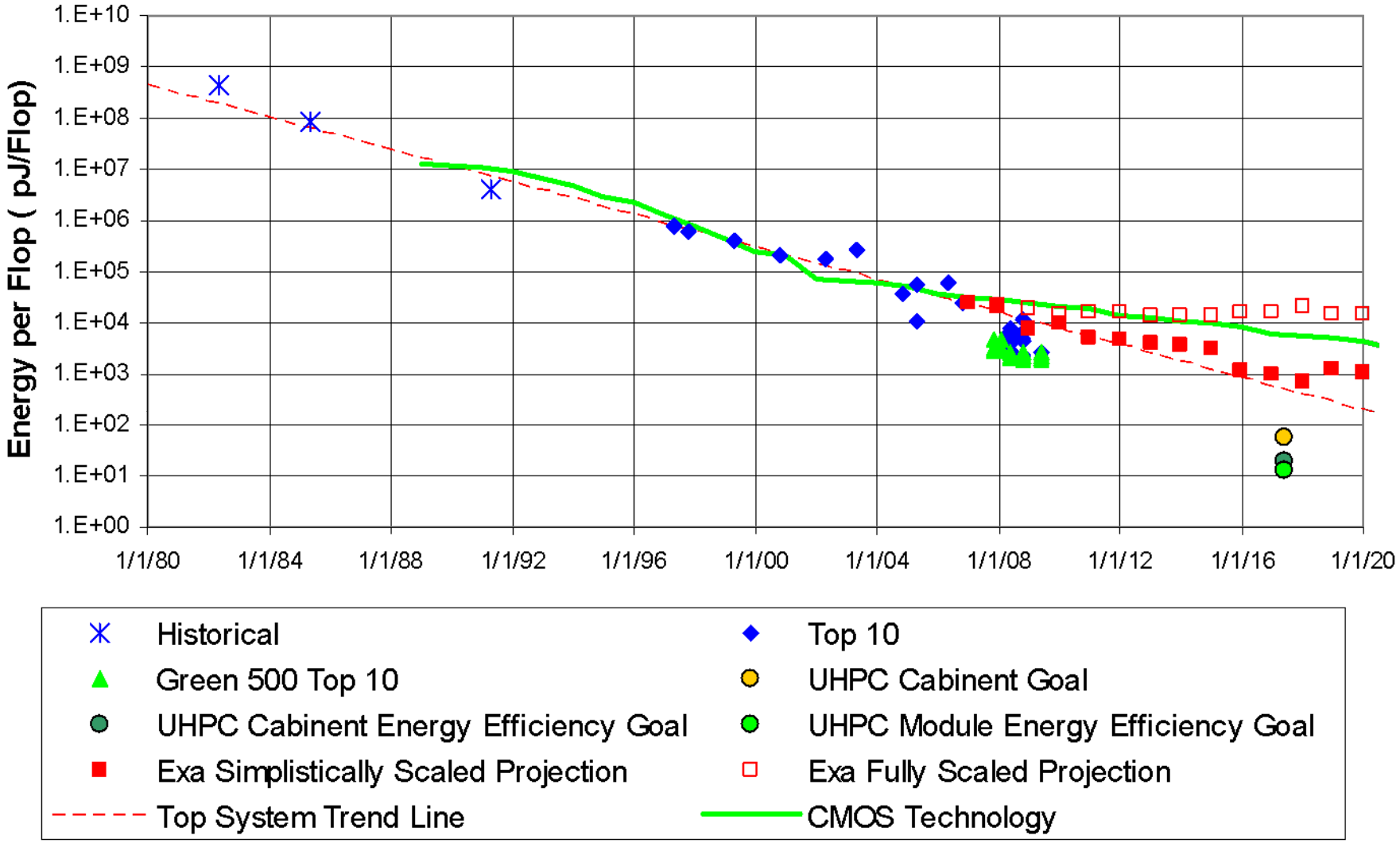


But not at 20 MW!

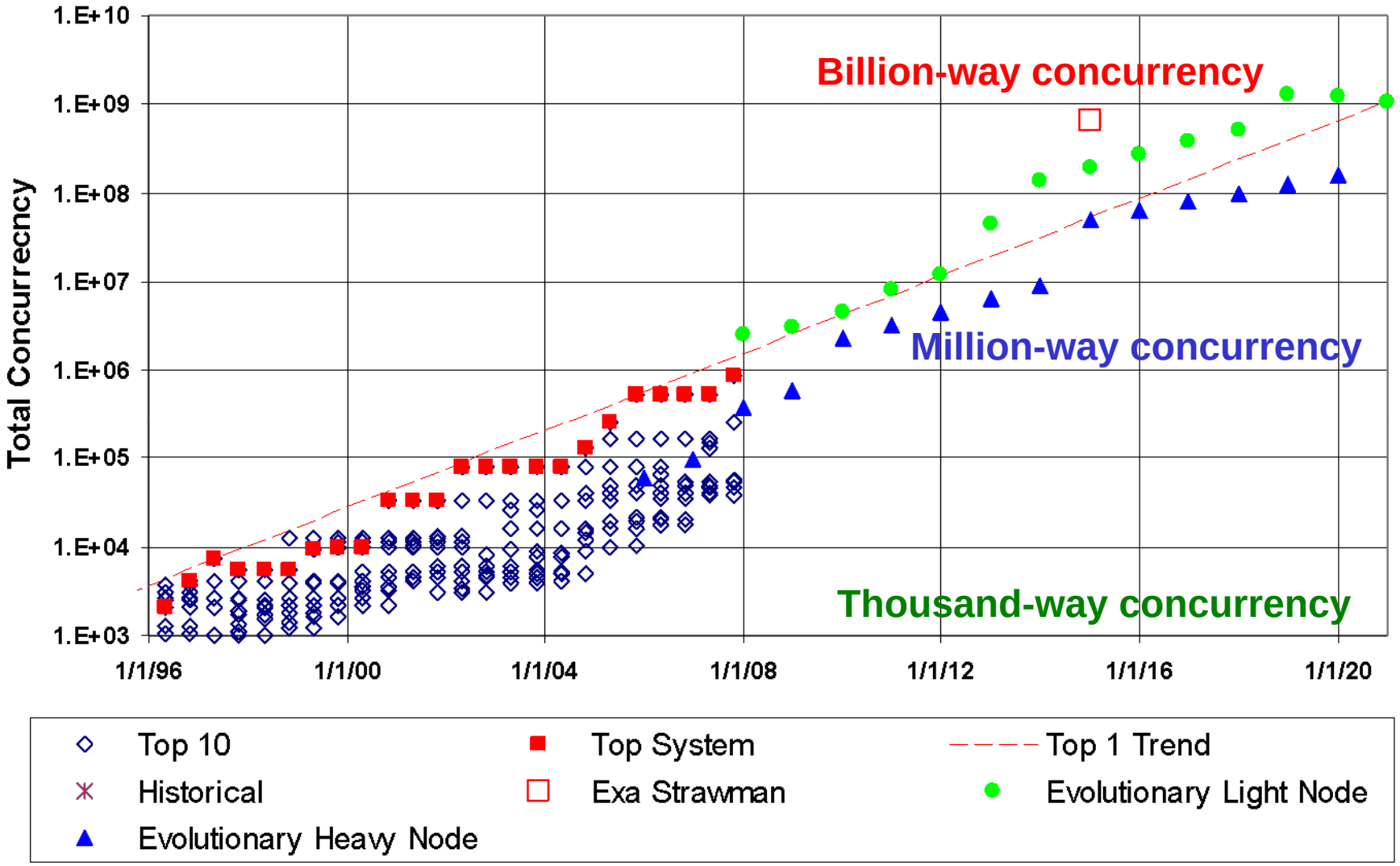
Power Efficiency



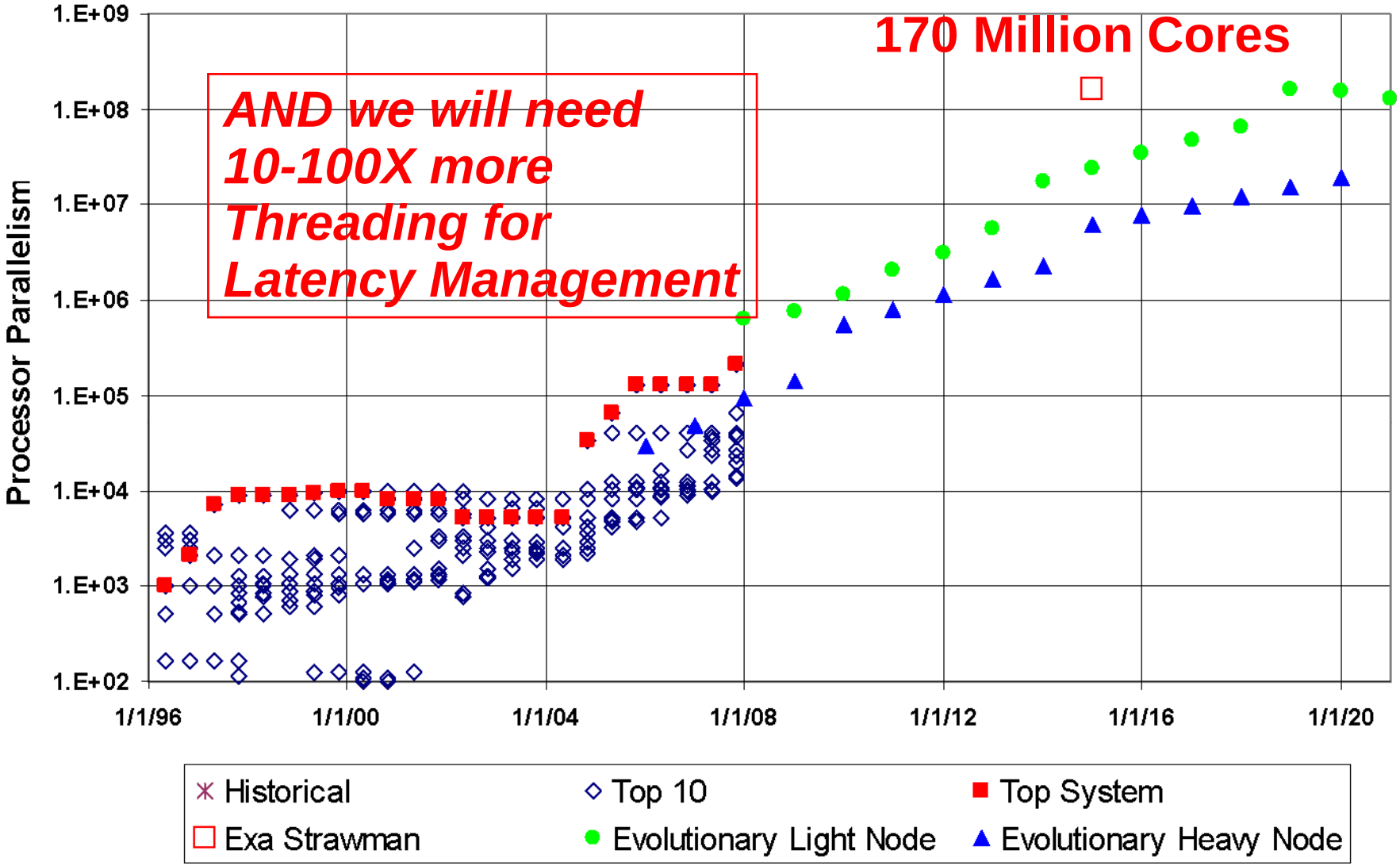
Energy Efficiency



Data Center Total Concurrency



Data Center Core Parallelism



Key Take-Aways

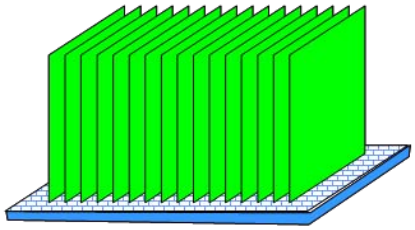
- **Developing Exascale systems really tough**
  - In any time frame, for any of the 3 classes
- **Evolutionary Progression is at best 2020ish**
  - With limited memory
- **4 key challenge areas**
  - **Power:**
  - **Concurrency:**
  - **Memory Capacity**
  - **Resiliency**
- **Requires coordinated, cross-disciplinary efforts**

Embedded Exa: A Deep Dive into Interconnect to Deliver Operands

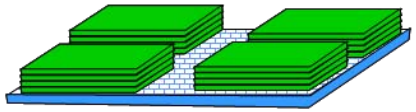
Tapers

- ***Bandwidth Taper***: How effective *bandwidth* of operands being sent to a functional unit varies with location of the operands in memory hierarchy.
  - Units: Gbytes/sec, bytes/clock, *operands per flop time*
- ***Energy Taper***: How *energy cost* of transporting operands to a functional unit varies with location of the operands in the memory hierarchy.
  - Units: Gbytes/Joule, *operands per Joule*
- Ideal tapers: "Flat"–doesn't matter where operands are.
- Real tapers: huge dropoffs

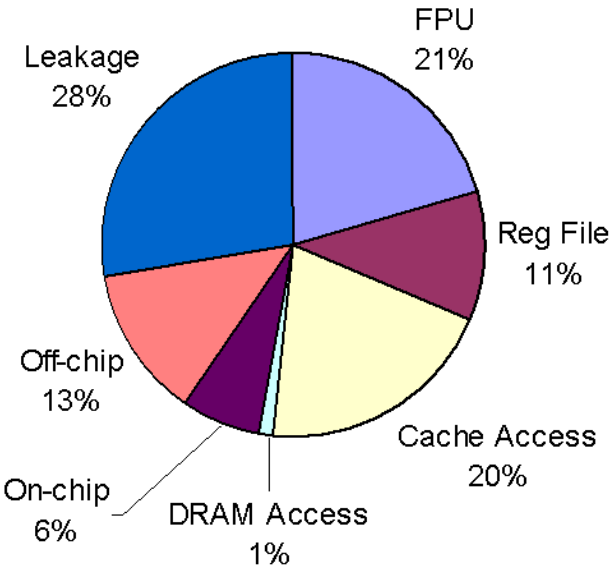
An Exa Single Node for Embedded

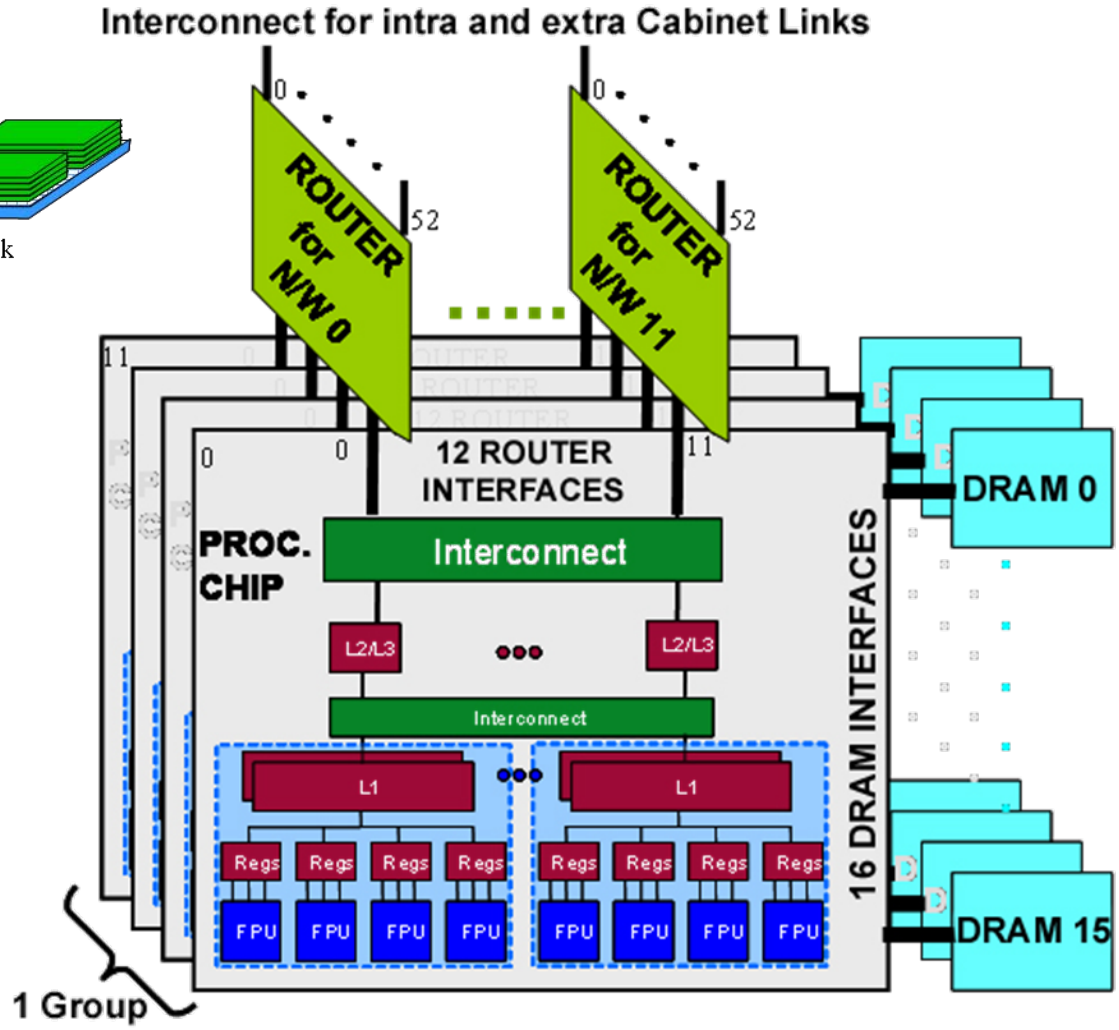


(a) Quilt Packaging

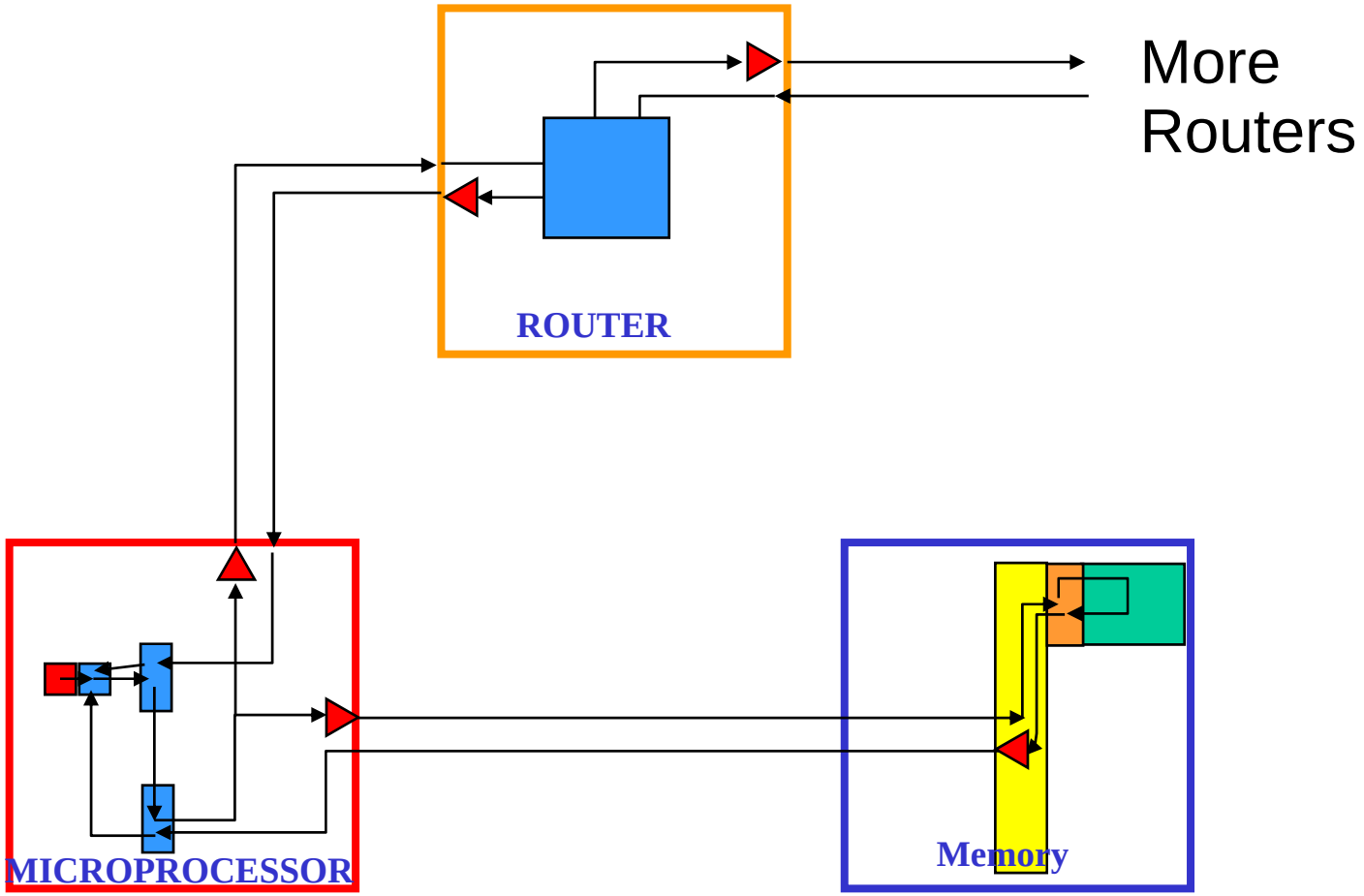


(b) Thru via chip stack





The Access Path: Interconnect-Driven



Some sort of memory structure

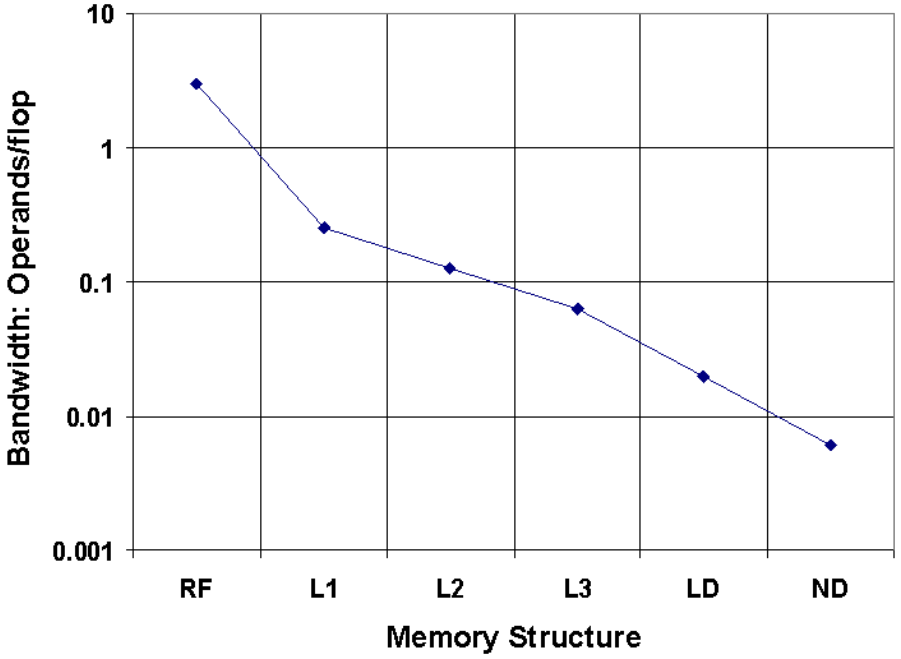
Sample Path – Off Module Access

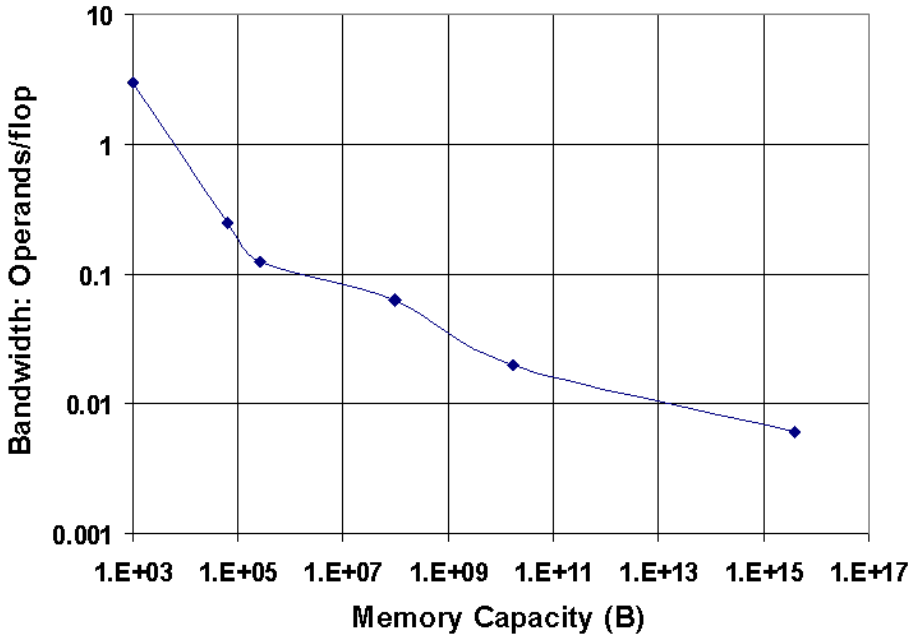
1. Check local L1 (miss)
2. Go thru TLB to remote L3 (miss)
3. Across chip to correct port (thru routing table RAM)
4. Off-chip to router chip
5. 3 times thru router and out
6. Across microprocessor chip to correct DRAM I/F
7. Off-chip to get to correct DRAM chip
8. Cross DRAM chip to correct array block
9. Access DRAM Array
10. Return data to correct I/R
11. Off-cchip to return data to microprocessor
12. Across chip to Routre Table
13. Across microprocessor to correct I/O port
14. Off-chip to correct router chip
15. 3 times thru router and out
16. Across microprocessor to correct core
17. Save in L2, L1 as required
18. Into Register File

Taper Data from Exascale Report

Memory Level	Capacity GB	Bandwidth GB/s	BW Taper Operands/clock	Power mW/flop	Energy Taper Operands/pJ
Register File	1KB	35,960	4	8.2	0.56
L1	64KB	8,992	0.25	5.5	0.068
L2	256KB	4,496	0.125	3.5	0.054
L3	97MB	2,288	0.0625	3.8	0.025
Local DRAM	16GB	712	0.02	10.5	0.0029
Network DRAM	3.8PB	216	0.006	11.5	0.00008

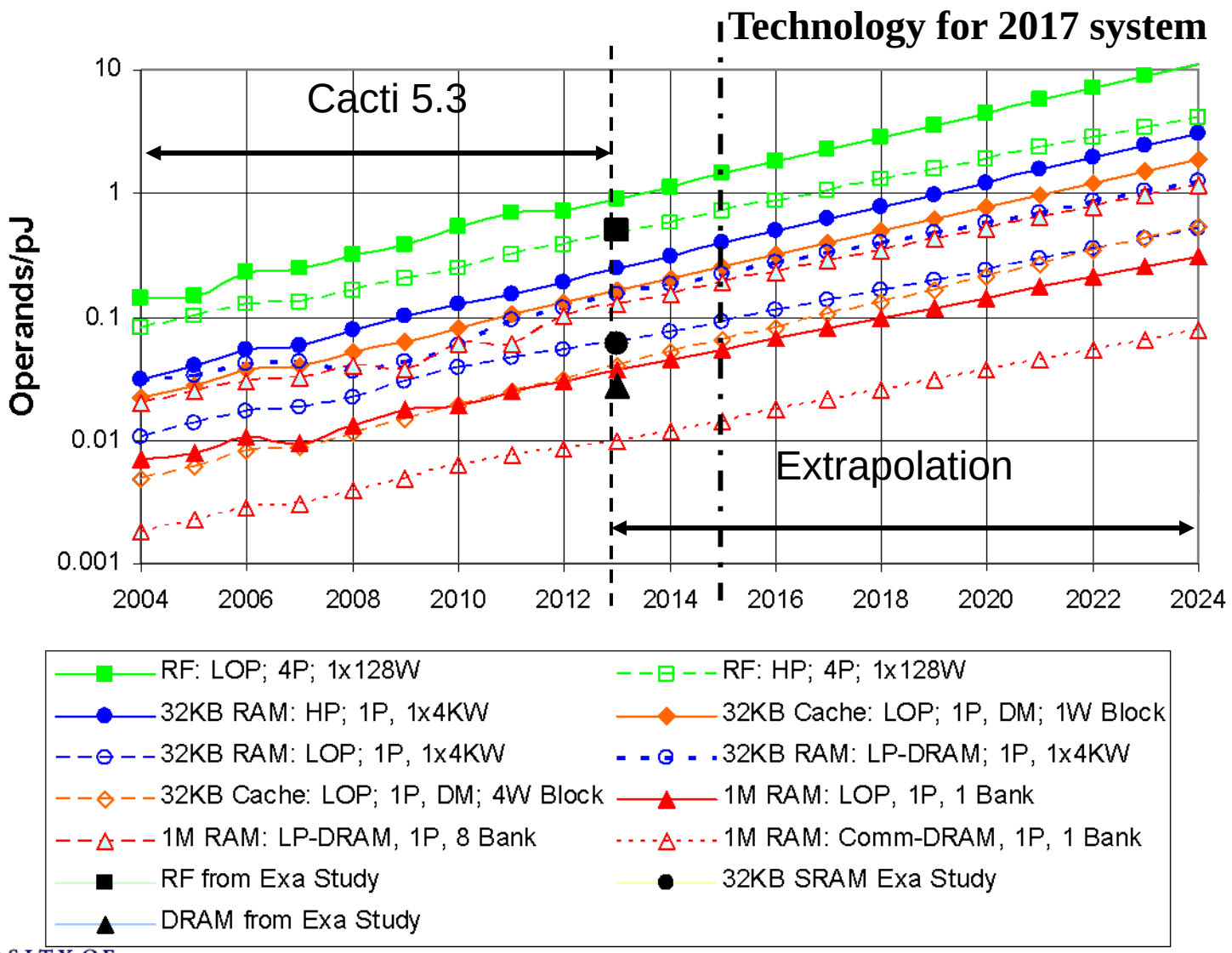
Bandwidth Tapers





1,000X Decrease Across the System!!

Energy Analysis: Possible Storage Components



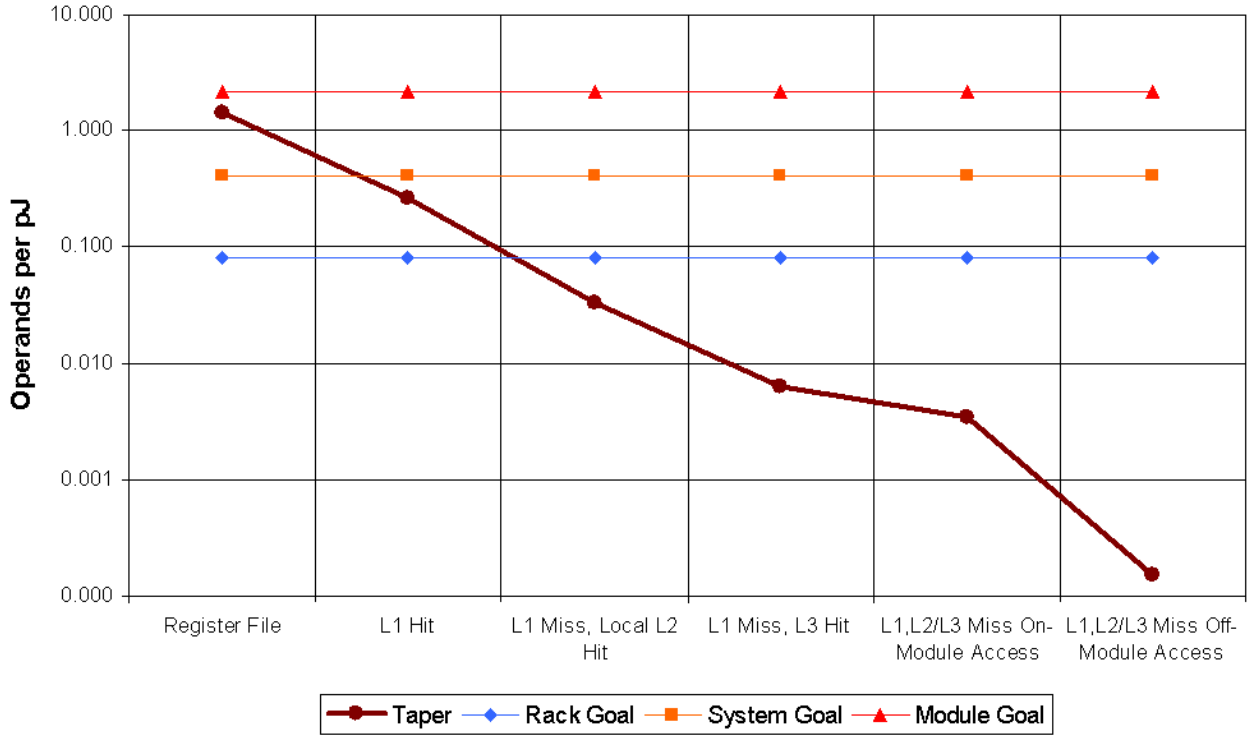
Summary Transport Options

Option	Value	units	Comments
Off-chip drive			
Core to L1	2.1	mm	Scaled to 2015 dimensions
Cross chip	21.3	mm	Unchanged from Exa study
High swing wire	0.107	pJ per bit per mm	
Low swing wire	0.018	pJ per bit per mm	
One word chip-chip - pad	144	pJ per word	Traditional pad
One word chip-chip - TSV	0.8	pJ per word	11 fJ/bit
One word chip-chip - capacitive	144	pJ per word	
One word chip-chip - inductive	10.1	pJ per word	
On-chip Optical			
E/O modulator	7.2	pJ per word	Modulate 1 word onto laser beam
O/E Receiver	7.2	pJ per word	Receive one word back to digital
On-chip Broadband Router	0.5	mW	per router
Power per laser	10.0	mW	per wavelength
Power for 250 lasers	2.5	W	
Off chip Optical			
Laser, per word	96	pJ per word	Uisng 0.3mW per channel @10% activity and 2
Modulator, per word	0.72	pJ per word	
RX + TIA, per word	360	pJ per word	Again at 10% activity
Temperature Control	?	pJ per word	

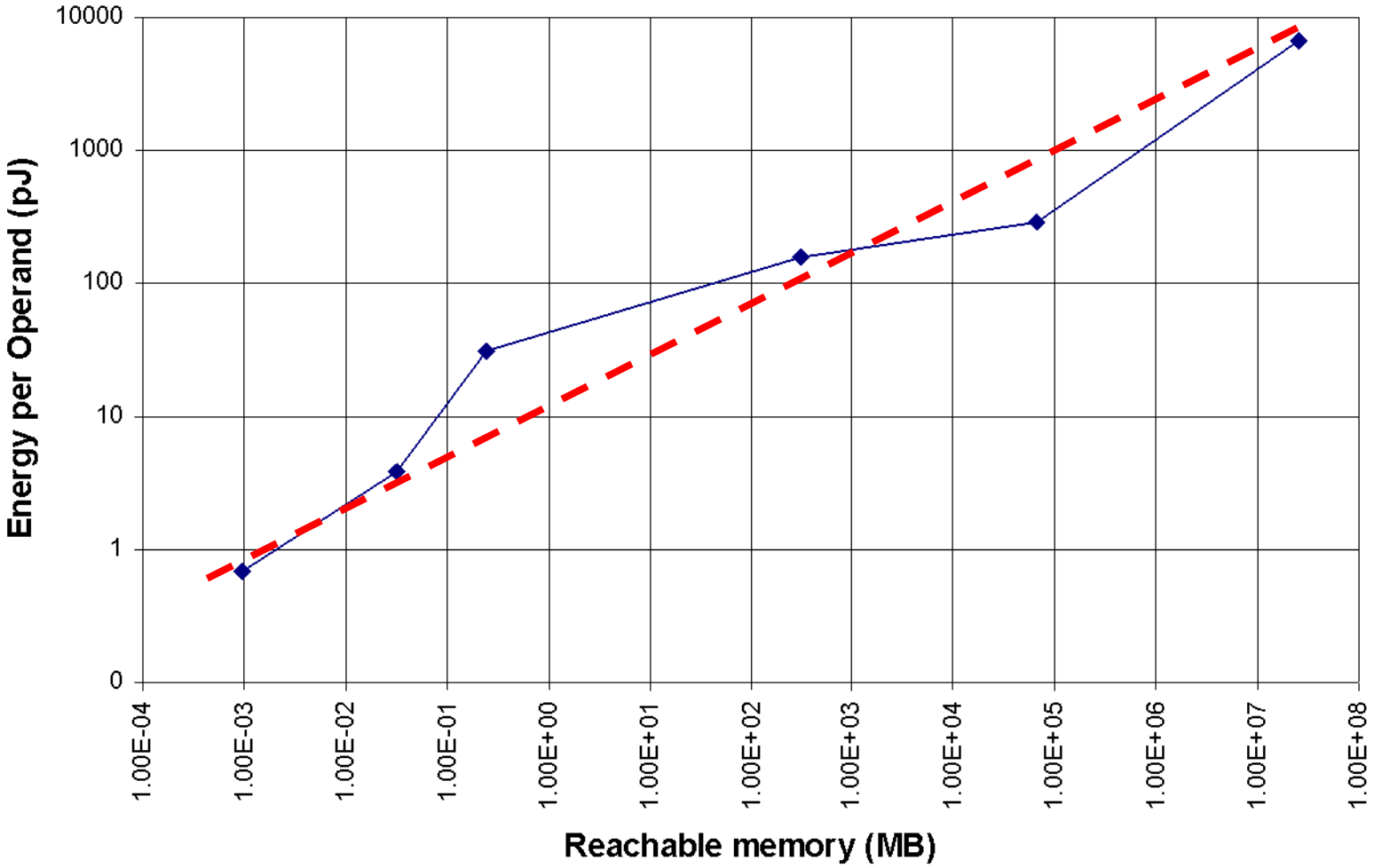
1 Operand/word = 72 bits

Energy Tapers vs Goals

Path	Reachable Capacity(MB)	Energy (pJ) Total	Transport	Access	% Transport	Opernads per pJ
Register File	9.77E-04	0.7	0.0	0.7	0.0%	1.436
L1 Hit	3.13E-02	3.9	0.0	3.9	0.0%	0.260
L1 Miss, Local L2 Hit	2.50E-01	30.6	13.3	17.3	43.4%	0.033
L1 Miss, L3 Hit	3.04E+02	158.3	141.0	17.3	89.1%	0.006
L1,L2/L3 Miss On-Module Access	6.55E+04	292.1	255.0	37.1	87.3%	0.003
L1,L2/L3 Miss Off-Module Access	2.52E+07	6620.8	6607.2	13.7	99.8%	0.000



Reachable Memory vs Energy



What Does This Tell Us?

- **There's a lot more energy sinks than you think**
  - And we have to take *all of them* into consideration
- **Cost of Interconnect *Dominates***
- **Must design for on-board or stacked DRAM, with DRAM blocks *CLOSE***
  - take into account physical placement
- **Reach vs energy per access looks "linear"**
- **For 80GF/W, cannot afford *ANY* memory references**
- **We NEED to consider the entire access path:**
  - Alternative memory technologies – reduce access cost
  - Alternative packaging costs – reduce bit movement cost
  - Alternative transport protocols – reduce # bits moved
  - Alternative execution models – reduce # of movements