Energy at Exaflops

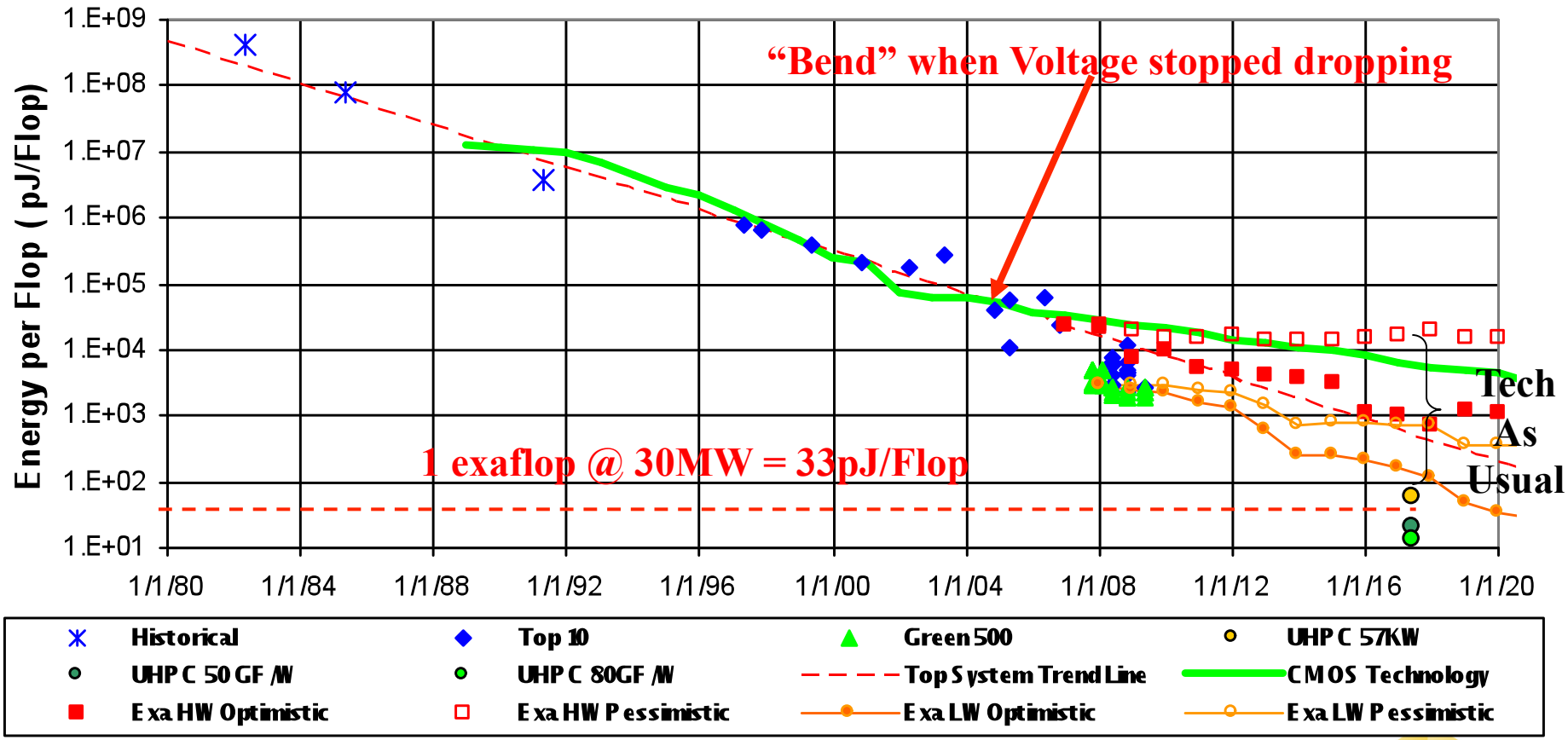
Peter M. Kogge
McCourtney Prof. of CSE
Univ. of Notre Dame
IBM Fellow (retired)

The Key Take Away

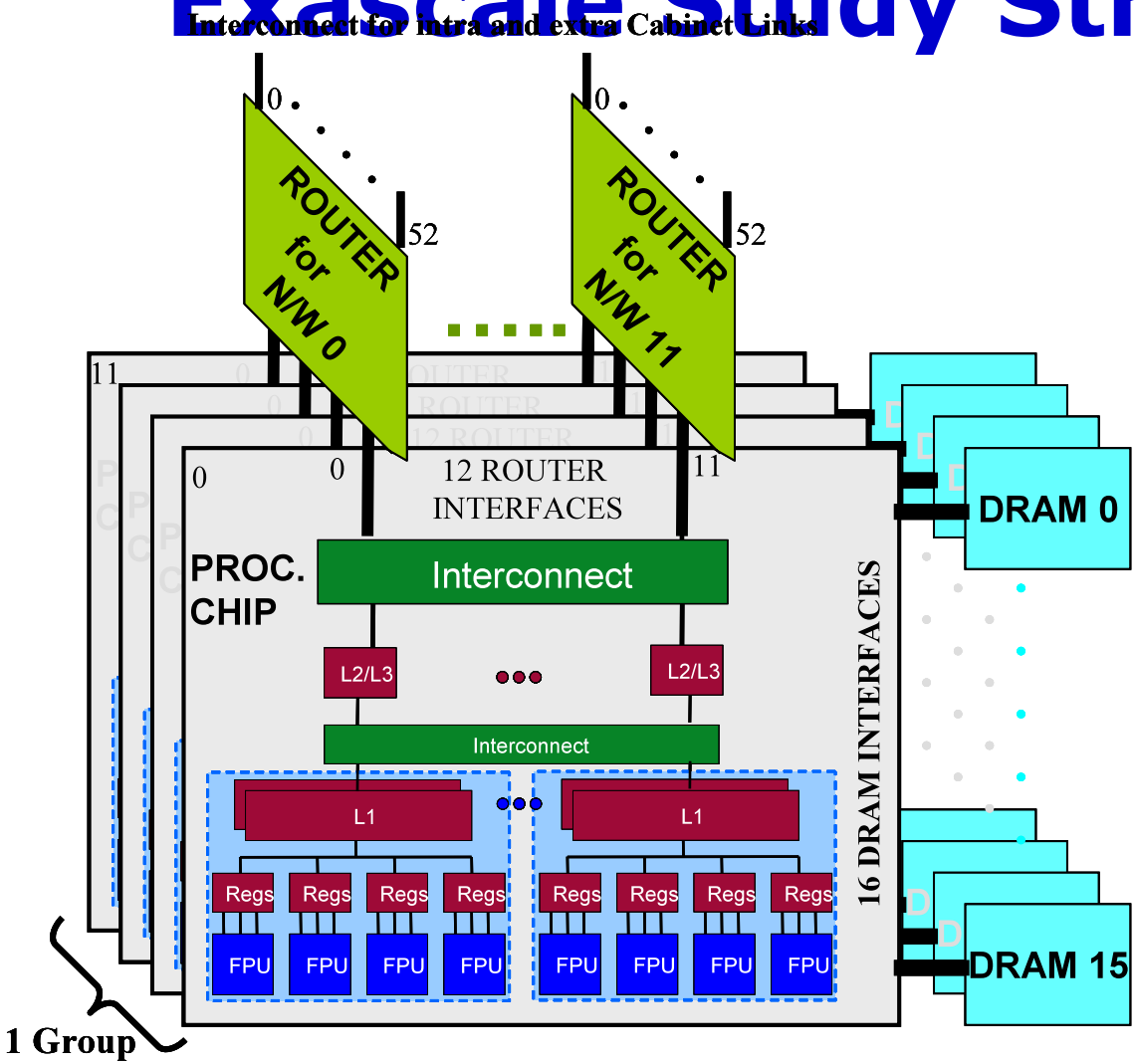
You can hide the latency…
But

You Can't Hide the Energy

"Business as Usual" Energy Projections

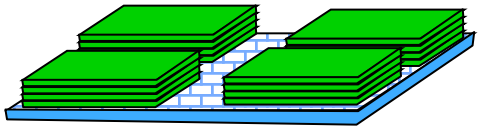


Exascale Study Strawman



Memory Hierarchy

- **Register File**
- **Level 1 Cache**
- **Level 2/3 Cache**
- **On-Module DRAM Memory**
  - **16x1GB**
- **Off-Module Memory**

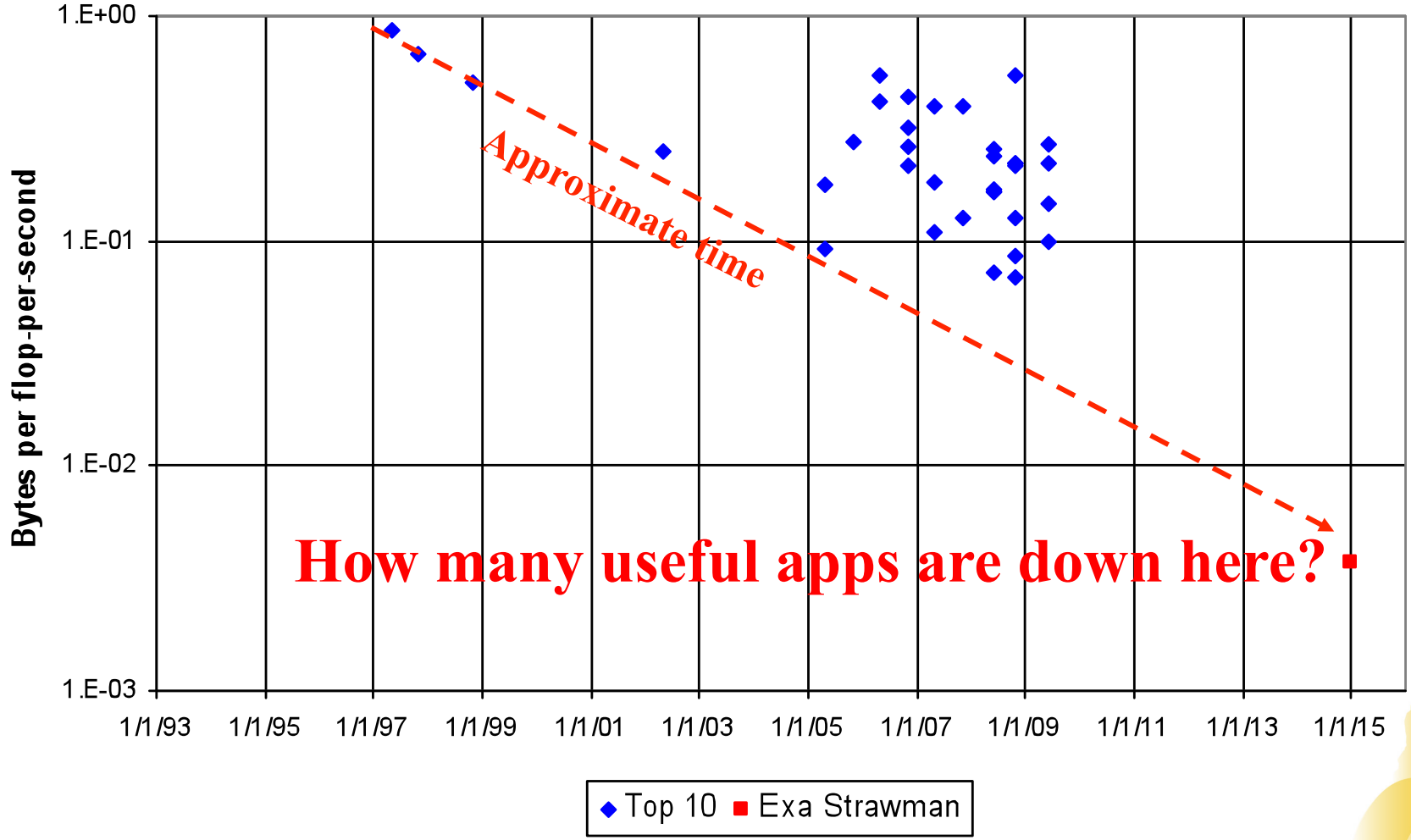


(b) Thru via chip stack

1 Cabinet Contains 32 Groups on 12 Networks

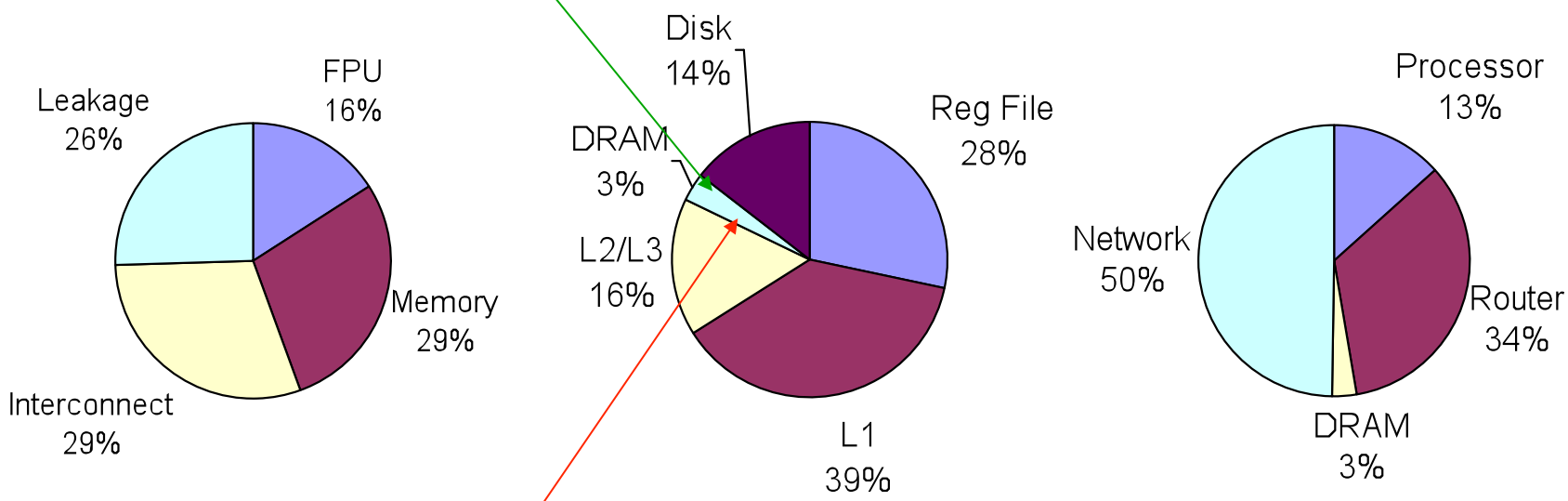
734 "Simple" cores per chip in 2013 technology

But At What Memory Capacity?



Strawman Power Distribution

Assumed a 60X per bit decrease in access energy



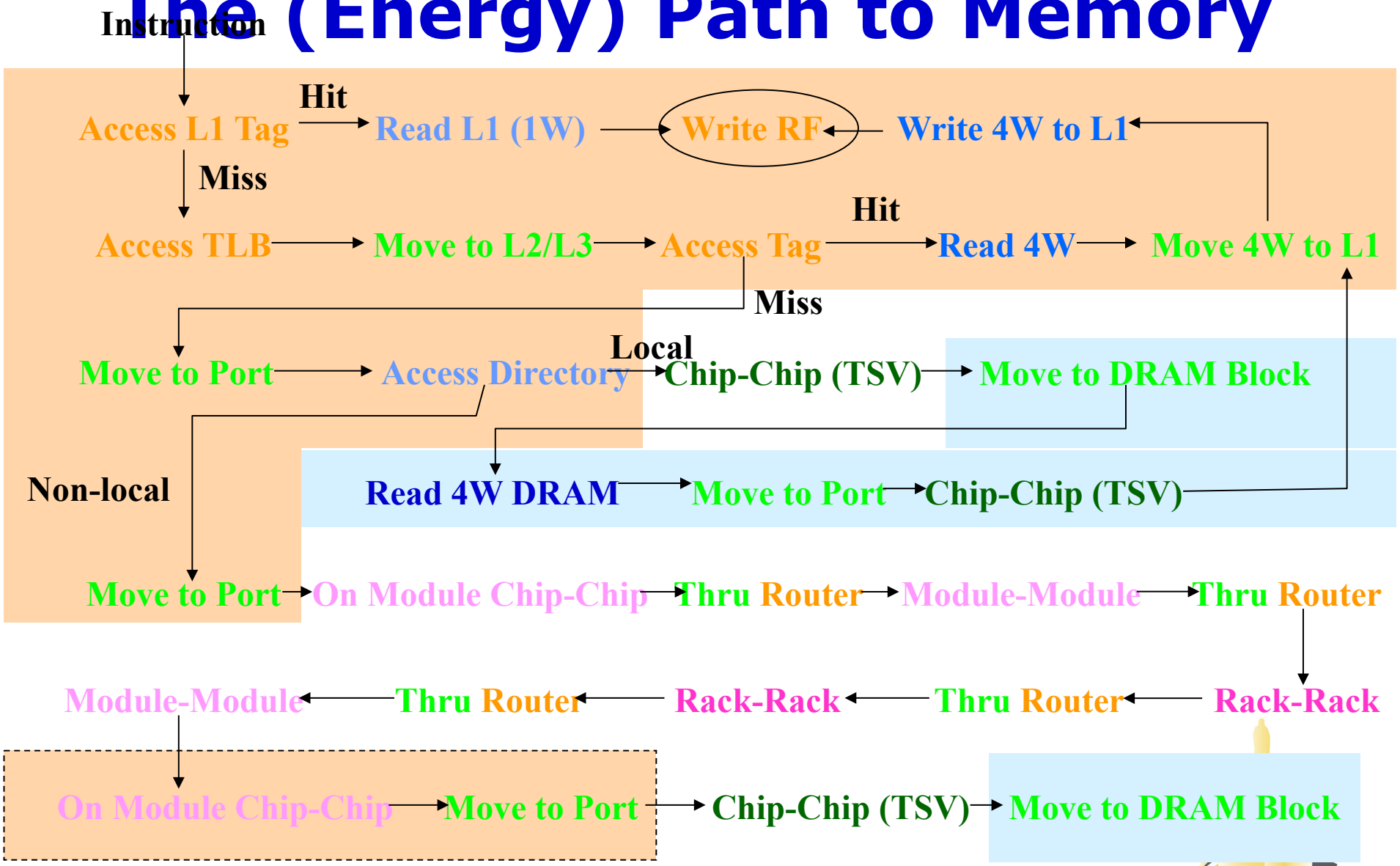
(a) Overall System Power (b) Memory Power (c) Interconnect Power

Growing DRAM capacity 30X at least *doubles* overall Memory Power

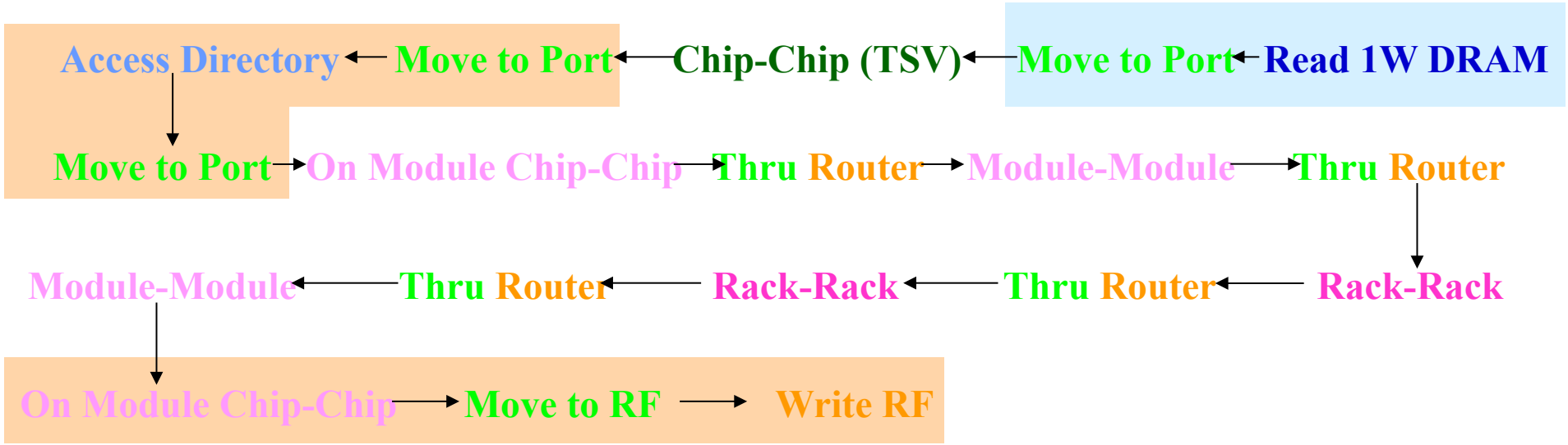
A Budget to Match Panel Goals

- 30MW for 1 Exaflop => 33pJ/Flop
- 20% lost to leakage => ~26pJ/flop
- Using the report's 2013 technology base:
  - ~10.6pJ for FPU operation
  - ~14.6pJ for L1 Instr cache access
  - ~1.8pJ for each Register File access
- Strawman dataflow per flop op: ~19.7pJ
  - 1 instruction fetch amortized over 4 FPU ops
  - Each FPU op takes 2 RF reads & 1 RF write
  - (An FMA would require another RF read)
- This leaves ~6 pJ for memory access/flop
- Moving to 2015 technology reduces per op to ~10pJ
  - Leaving 16pJ for memory

The (Energy) Path to Memory



The Path from (Global) Memory



AND THIS DOESN'T ACCOUNT FOR TLB MISSES!!!

Tracking the Energy

0	Total (pJ)	RF Access	Router Logic	Tag Access	32KB SRAM Access	DRAM Access	On-Chip Transport	TSV	Chip-Board-Chip	Chip-Optical-Chip
L1 Hit	39	30%	0%	28%	42%	0%	0%	0%	0%	0%
L2/L3 Hit	385	7%	0%	6%	34%	0%	53%	0%	0%	0%
Local	1380	1%	0%	2%	2%	68%	26%	1%	0%	0%
Global	13819	0%	24%	0%	0%	1%	13%	0%	10%	52%

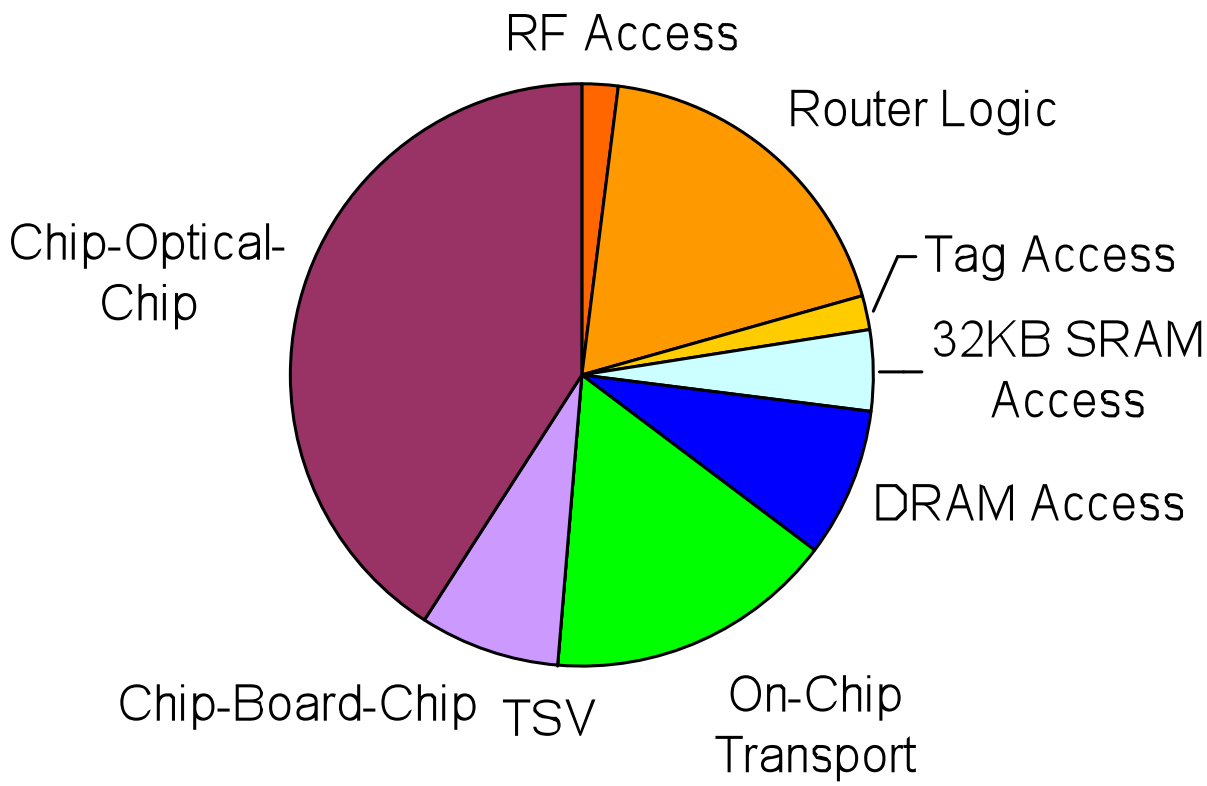
Remember we have only 6-16pJ per flop instruction

What's an "Average" Memory Reference?

706 pJ per access

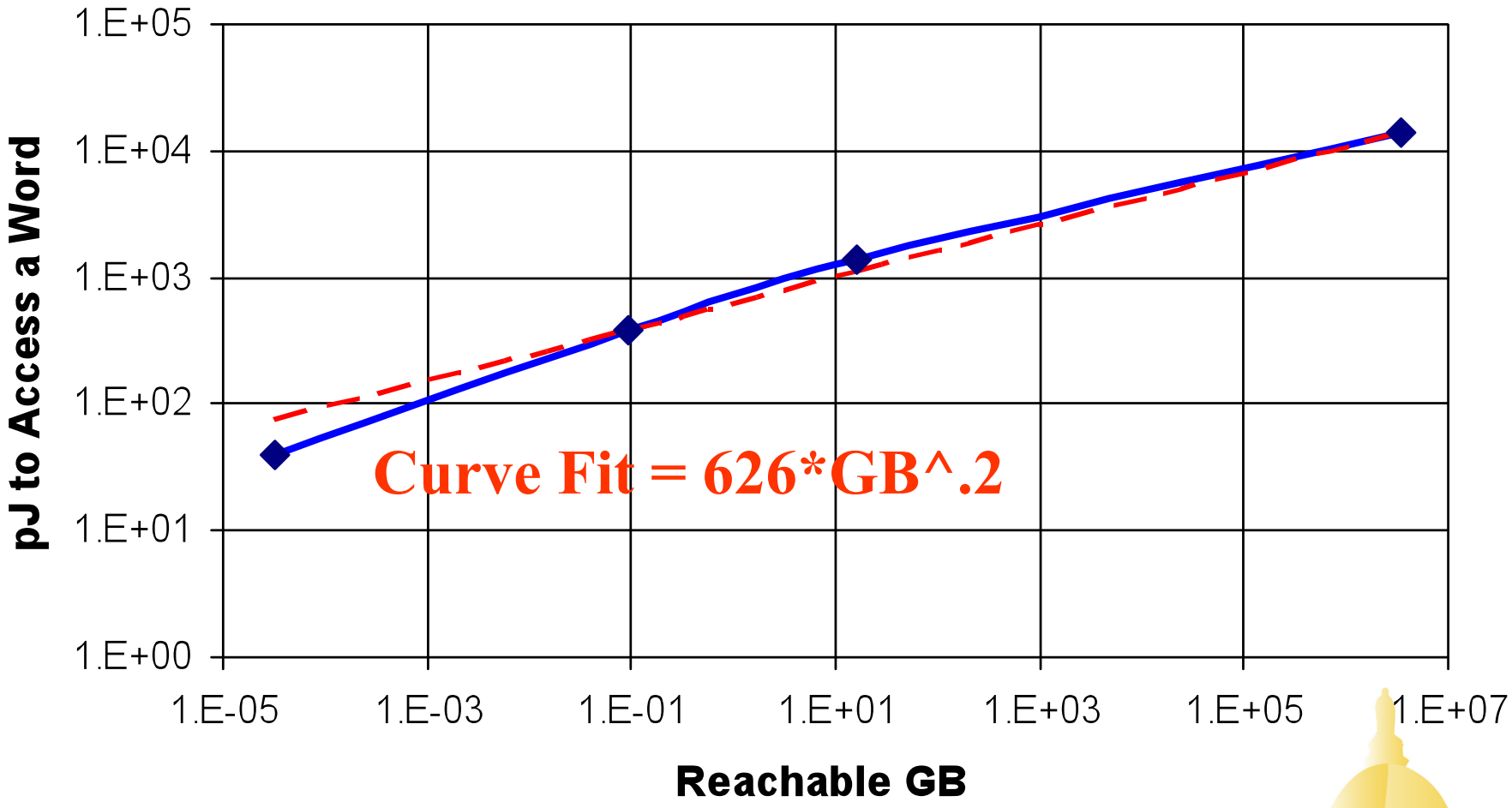
- **80% L1 hit**
- **90% L2 hit**
- **On L2 Miss:**
  - **60% Local**
  - **40% Global**

And Interconnect is 85% of Energy!!!



By Reducing DRAM Power by 60X, It becomes 8% of Average – but still 10X 6pJ budget

Access vs Reach



What Does This Tell Us?

- Cannot afford ***ANY*** memory references
  - One out of every 100+ instructions can access memory
  - Even with 2015 Technology only have a 16pJ budget
    - Make it only 1 out of ever 30+ instructions
- There are lot more energy sinks than you think
- Cost of Interconnect ***Dominates***
- Must design for on-board or stacked DRAM
- We need to redesign 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

Our Study
Wasn't Creative Enough!

And This Means You Can Reference Memory Every ....

- For 100% L1 Hit, once every 6.5 Flops
- For an 80% L1 Hit and 100% L2/L3 Hit, once every 18 Flops
- For an 80% L1 Hit, 90% L2/L3 Hit, and all the rest is local, once every 35 Flops
- For an 80% L1 Hit, 90% L2/L3 Hit, 60% local, and 40% Global, once every 118 Flops

Color Code

- Red: FPU
- Orange: other logic, tag access, RF
- Dark Green: Chip-chip (TSV)
- Light Green: on chip transport
- Purple: rack-rack
- Light Purple: Chip-Chip (non TSV)
- Dark blue: DRAM access
- Light Blue: cache data access