Exascale vs the Exascale Report: Why so long? and Are We Really There Yet?

Peter Kogge

McCourtney Prof. of CSE, Univ. of Notre Dame

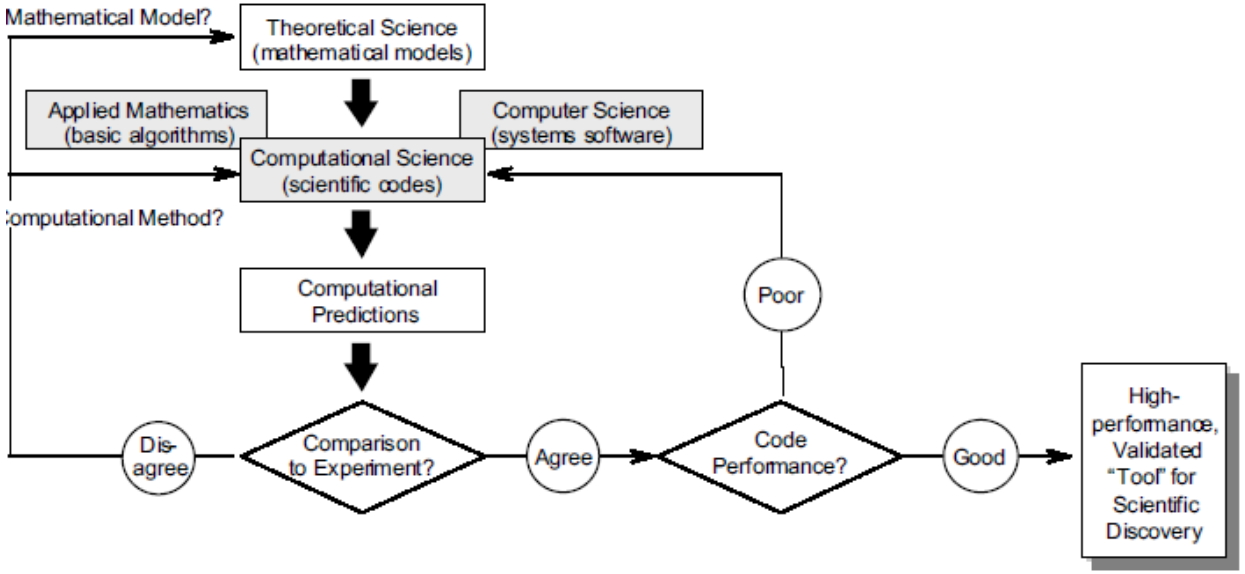
IBM Fellow (retired)

Cray, IEEE Computer Pioneer Awardee

This talk is an update to a talk given at SC'22 by Kogge&Dally

Pre 2008 Performance Projections

- **2000: DOE Report** "Scientific Discovery through Adv. Computing"
  - Beginning of ubiquitous parallelism
  - Challenge: Make full use of terascale
  - **Critical issues**: performance, portability, adaptability
    - Sim codes 5%-10% of peak – and decreases with parallelism



E,g. Computing energy of Iso-octane
- 275M nonlinear equations
- Iterative solution
- 2.5PB between processors
- 15 TB to disk
- 30 PFlops

- **2008 Exascale Study: Need much tighter tie-in to architecture**

The 2008 Exascale Study Report

Report Directive: Exascale by 2018 at 20MW

- What *should* "Exascale" Mean?
- The 2008 state of the art
  - Architectures, Runtimes, Programming, Metrics
- 2008 Application Characteristics
  - Computation vs Memory intensive Apps, Scaling, Concurrency
- Technology Roadmaps
  - Logic: Silicon and Non, Memory, Storage, Interconnect, Packaging, Resiliency, Programming Models

- **Strawman Designs**
  - Subsystem projections, Evolutionary designs (Heavy and lightweight), Aggressive design
- **Challenges & Research Areas**
  - ***Power, power, power, & power*** } **Remain**
  - Memory capacity & bandwidth
  - Programmability } **Practically Solved**
  - Reliability

The 2008 Exascale Eye-Openers

- **Energy/Power** was ***the*** dominant factor
- Only path forward
  - Simpler chips at less than max clock rates but ***massive parallelism*** needed
  - Alternative packaging
  - High degree networks
  - Must deal with greatly reduced memory bandwidth per op

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

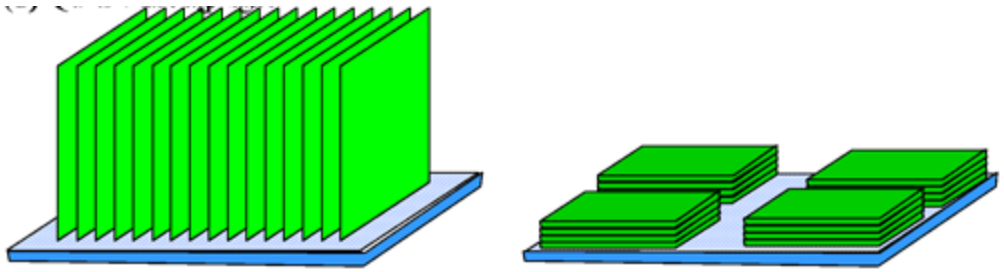
September 28, 2008

NOTICE

APPROVED FOR PUBLIC RELEASE, DISTRIBUTION UNLIMITED

Dongarra, J., Keyes, D. The co-evolution of computational physics and high-performance computing. *Nat Rev Phys* (2024). https://doi.org/10.1038/s42254-024-00750-z

2015 Aggressive Strawman Design (2013 Tech)

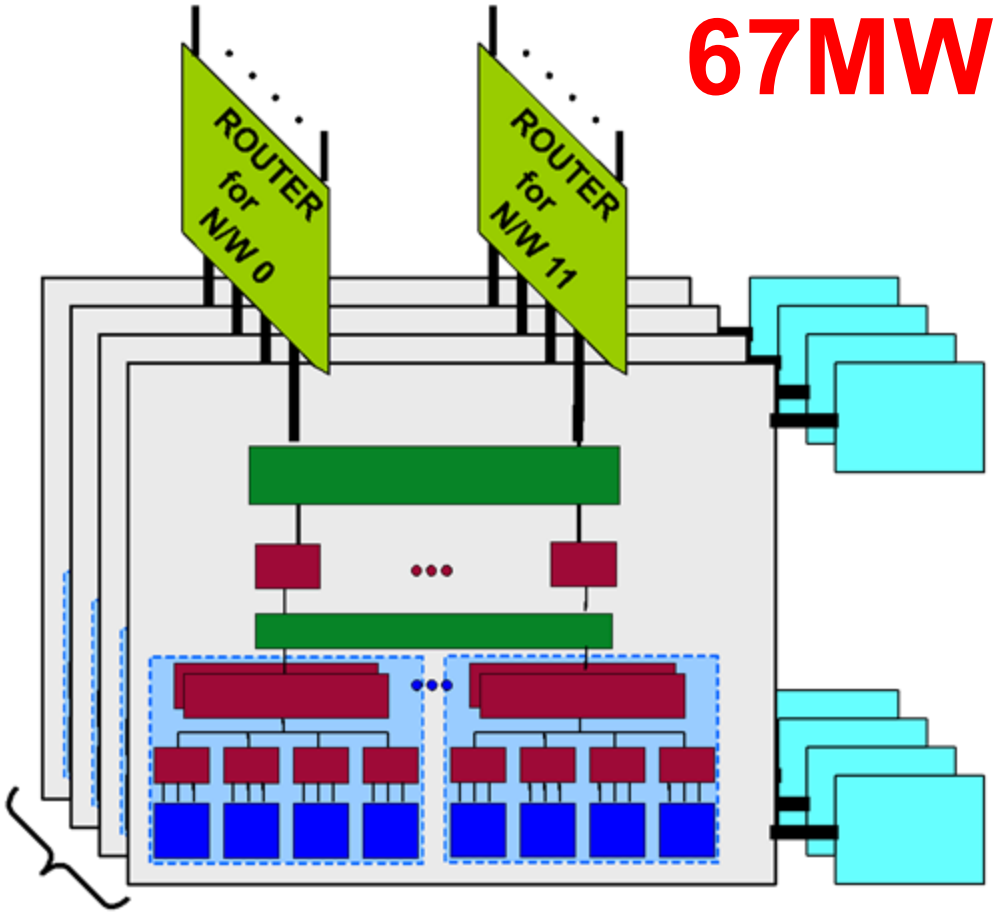


Node: 742 simple cores/chip with 4 FPUs @ 1.5GHz
- 32nm CMOS with 30Gb/s SERDES
- 16 Memory channels: each 1 GB *Stacked* DRAM
- 150 Watts w'o routing chip

Group: 12 nodes with 12 64-radix router chips
- Includes 16 12GB SATA drives for checkpointing

Cabinet: 32 Groups = 384 nodes
- Assumed max power of 120KW

System: 583 Cabinets, 67MW
- 3-hop Dragonfly interconnect (optical)
- 166 million cores with 664 million FPUs

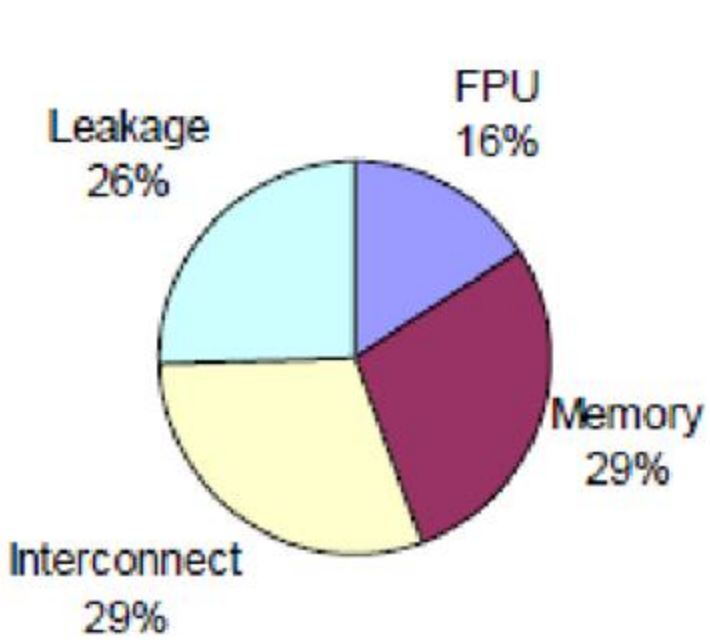


67MW

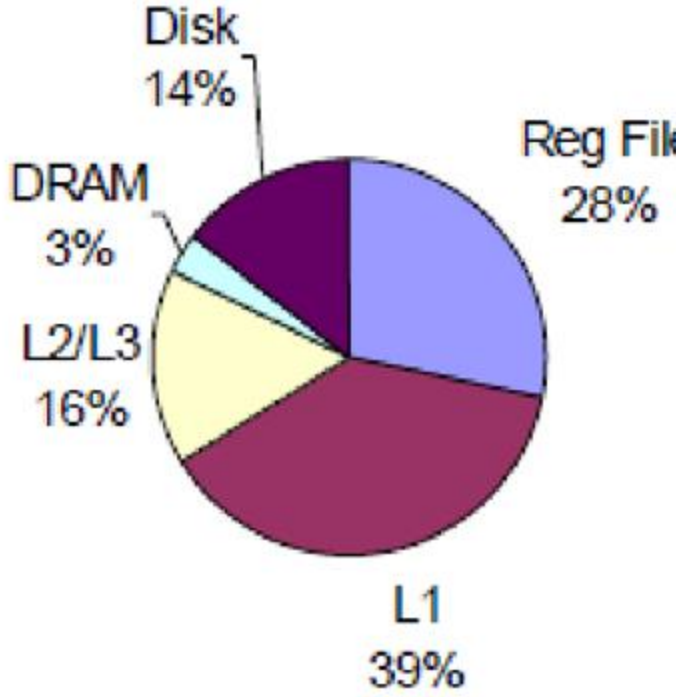
Est. 14.9 GF/W
Or 67 pJ/flop

WD

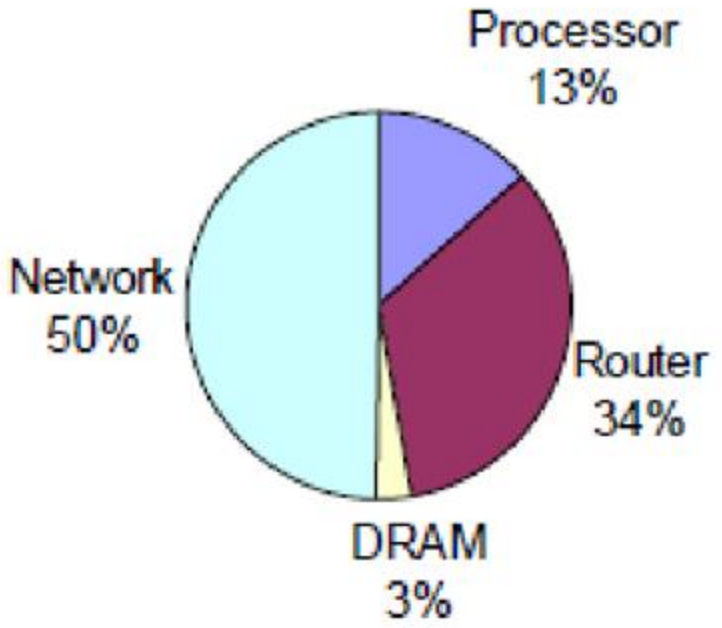
Strawman: Where Did the Energy Go?



(a) Overall System Power



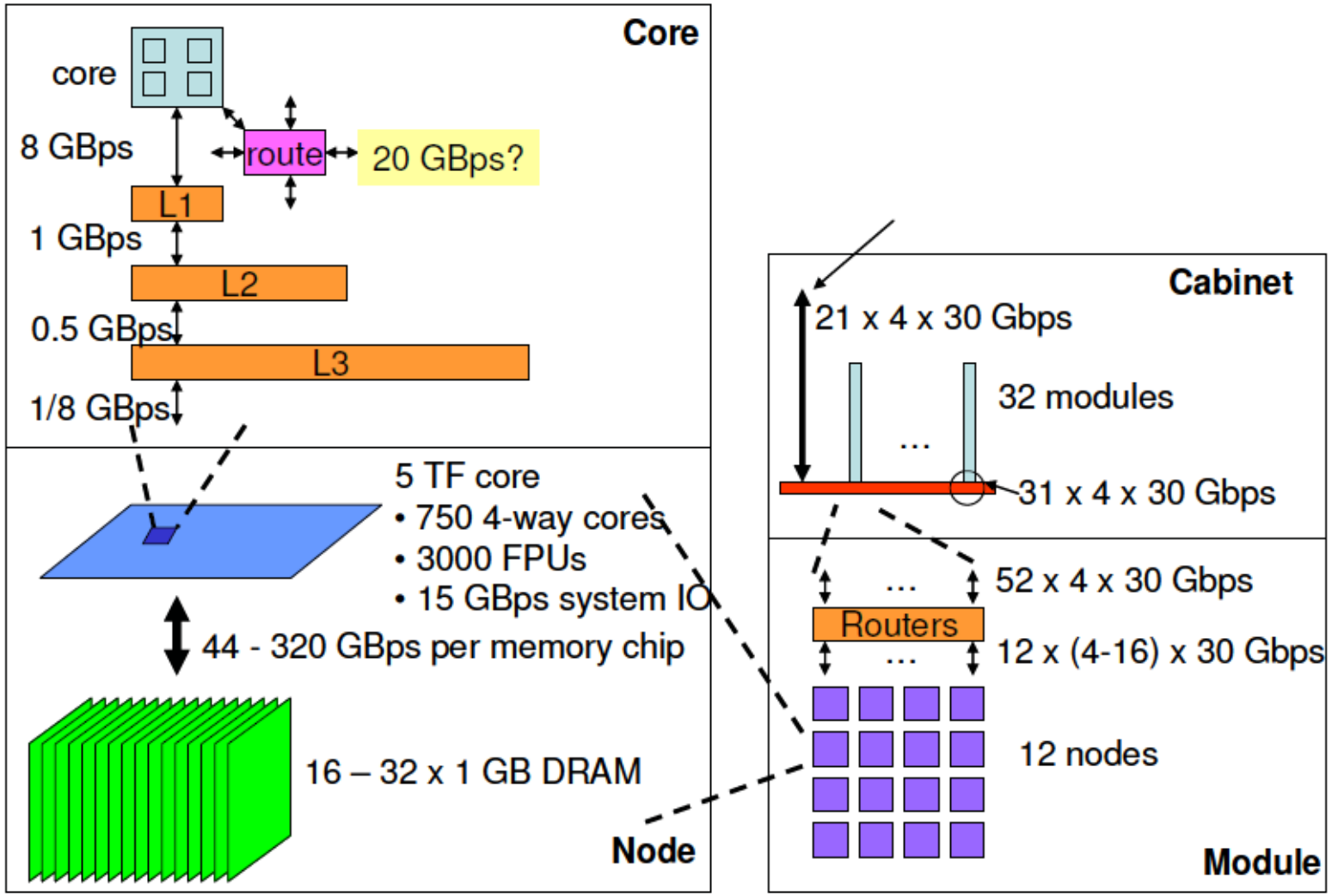
(b) Memory Power

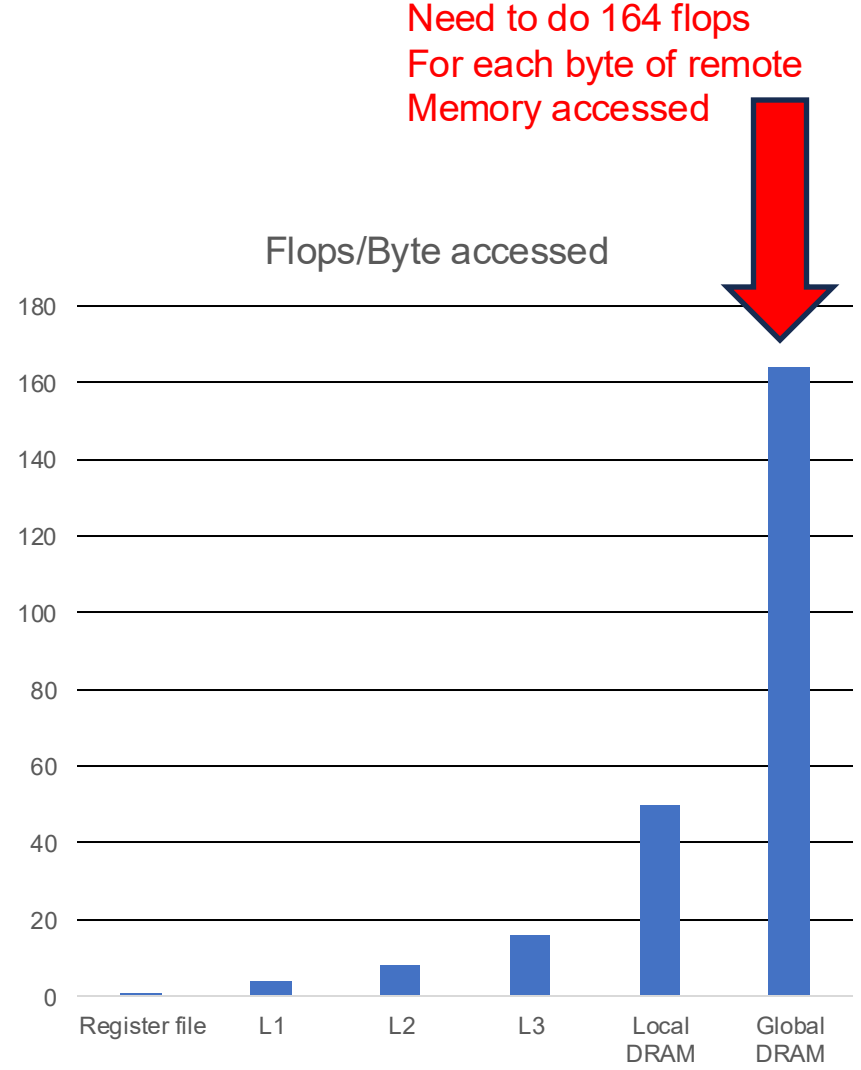


(c) Interconnect Power

Remember: Estimates based on 2013 Tech

Power/Energy Constrained Memory Bandwidth



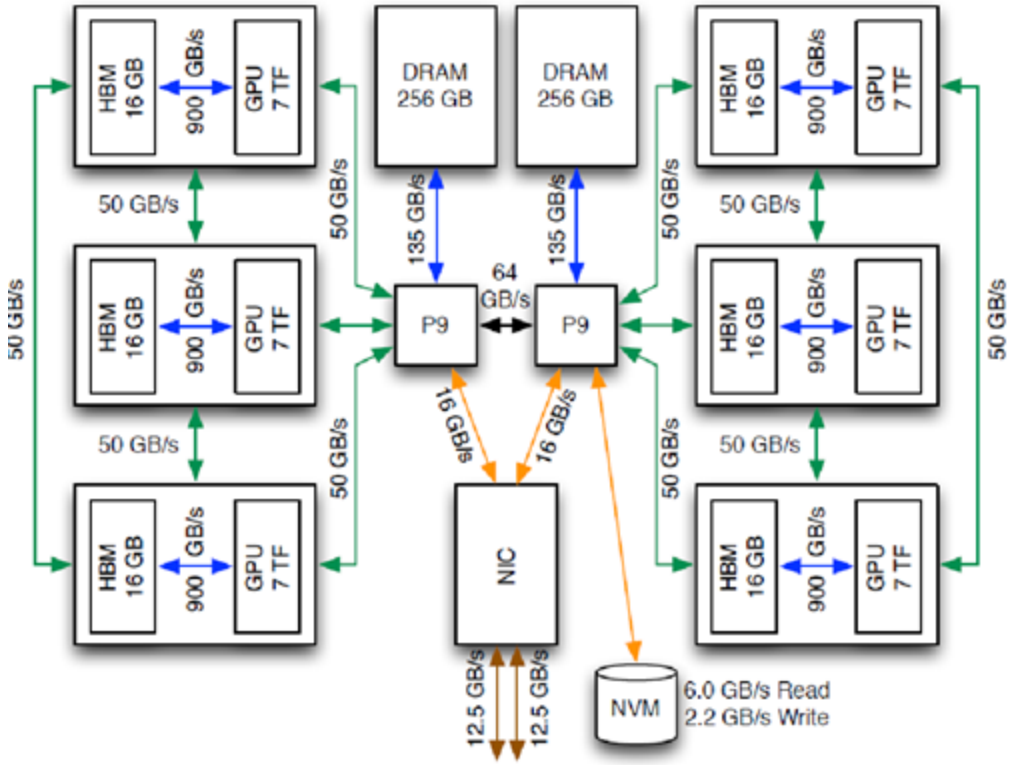


Remember: Estimates based on 2013 Tech

2018: Summit – An Exascale "Could Have Been"

- **Nodes**:
  - Dual 22 core Power 9
  - Hex NVIDIA GV100
  - Mixed DRAM/HBM (Stacked)
- **Cabinet**: 18 Nodes, 55KW
- **System**: 256 compute, 9.8 MW

Interesting Observation:
6.7X expansion of Summit
- **~1+ EF/s sustained**
- **At about 67 MW!**

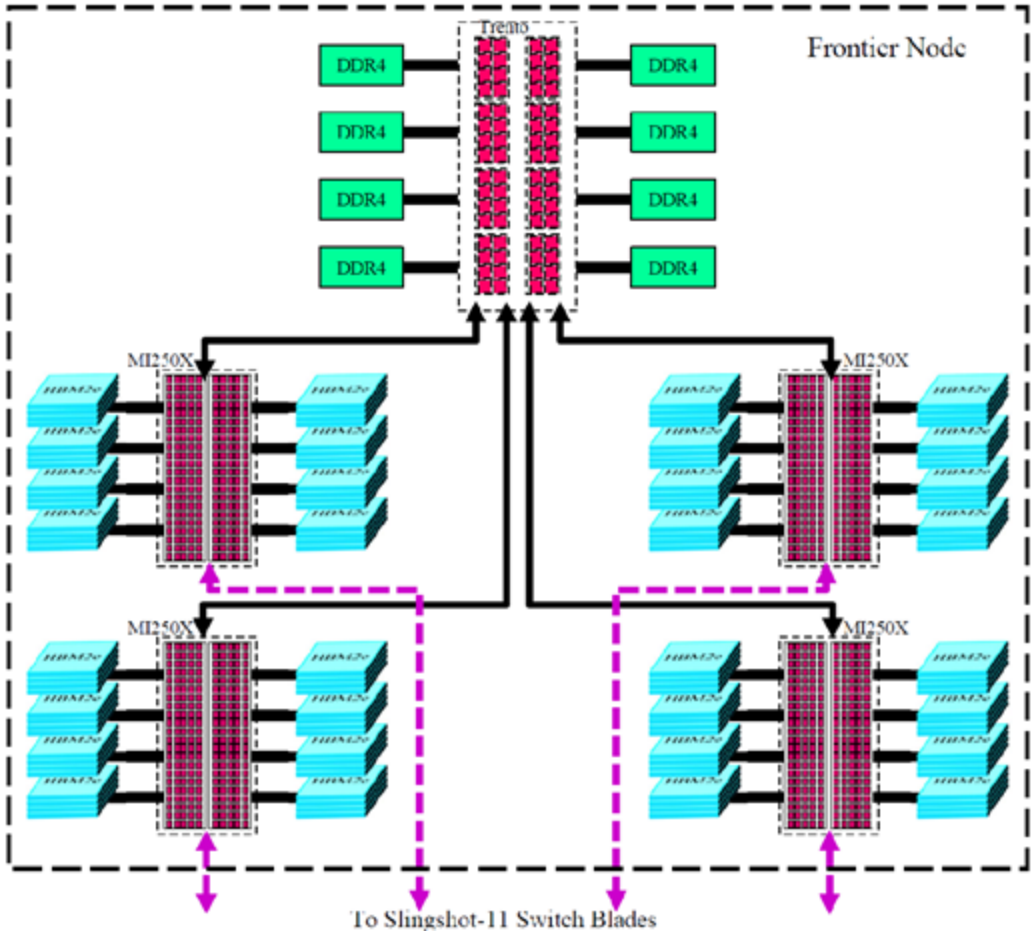


Summit System Overview, T. Papatheodore, 6/1/18

Measured 14.7 GF/W

2022 Frontier Node

- Heterogeneous Processors
  - 64-core 2GHz CPUs
  - Quad GPUs: **closer to Strawman**
    - **But more FPUs/core**
    - **And slightly faster**
- Chiplet design
- Mixed memory hierarchy
  - 8 DDR4 DRAM Channels
  - 8 HBM2e stacks/GPU
- Quad network ports



Measured 52.2 GF/W

~3.3X Strawman

PMK

Frontier vs Strawman

	Road-Runner	2008 Strawman	Frontier
System Counts			
Nodes/Blade	1	12	2
Blades/Chassis	4	1	8
Chassis/Cabinet	3	32	8
Nodes/Cabinet	12	384	128
Total Nodes	3060	223,872	9,408
Cores/Node	40	742	944
MACs/Node	76	2,968	56,832
Total MACs	232K	665M	535M
Memory Metrics			
Total Memory (TB)	36	3,498	9,408
Total Memory BW (TB/s)	378	157,605	125,239
Network Bandwidth Metrics			
Network ports/node	1	12	4
Total Network ports	3,060	2.7M	37,632
Switch Chips/Cabinet		384	64*
Switch Radix	24	64	64
Total Switch Chips	900	223,872	4,736*
Signal Rates (Gb/s)	4	30	56
Inj. B/W/Node (GB/s)	2	180	100
Bisection B/W (TB/s)	0.192	210	540
* Assuming 8 switch cards/chassis			

- Strawman's huge #s of nodes
  - Exploded # of Network ports
  - And thus huge switching costs
- Frontier had fewer, bigger nodes
  - Reduced network ports
- Comparable Memory Bandwidth
  - Use of wide stacked memory
  - But only 3X capacity
- Essentially same N/W topology
  - But 2X better SERDES
  - And 2+X better bisection B/W

WD

Report Card

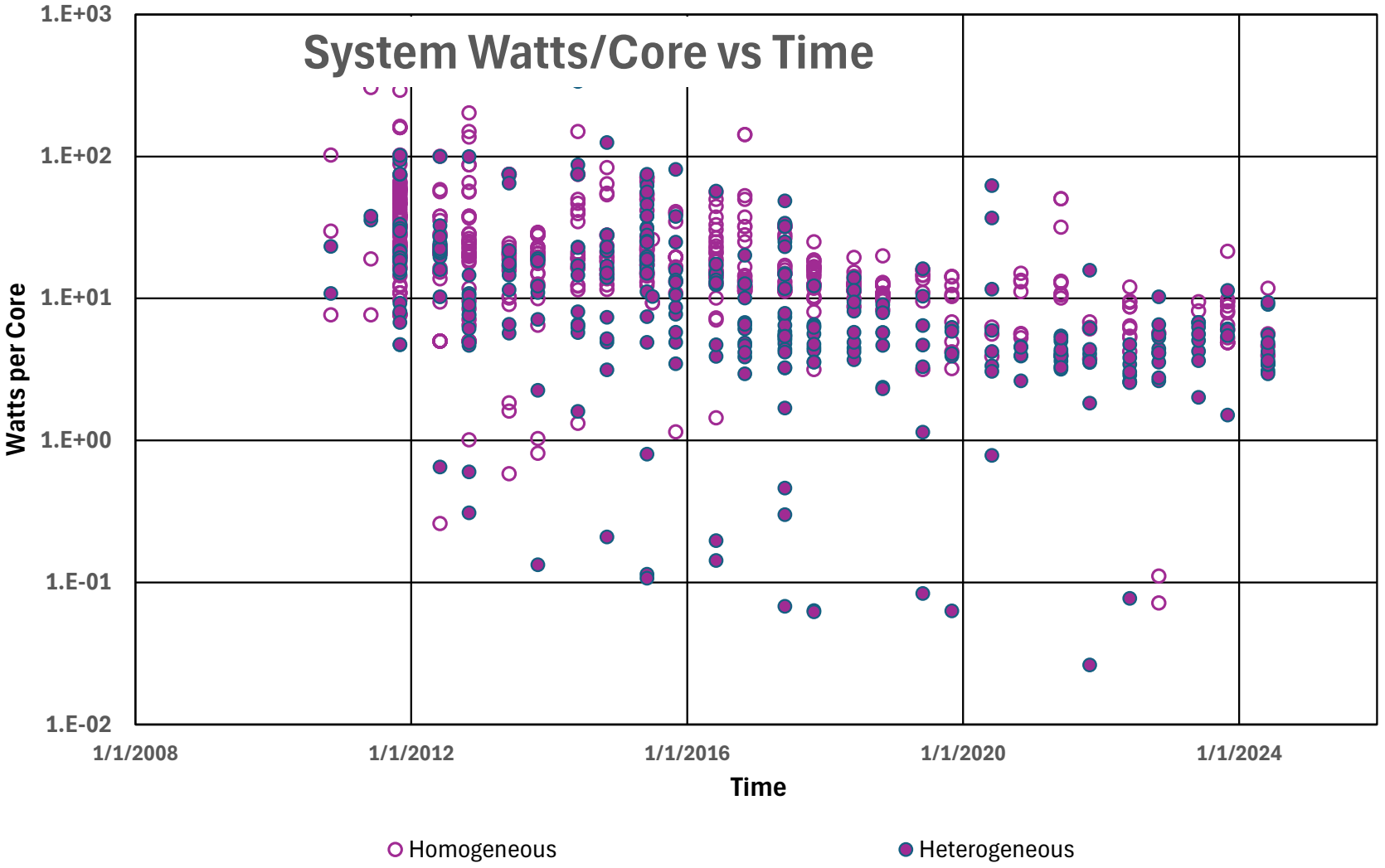
What We Got Right

- CMOS, flat clocks
- Large # of wide simple cores
- Aggressive memory hierarchy
- Stacked memory
- Near reticle-limited dies
- Energy of movement predominates
- Near billion-way concurrency
- Memory concerns were valid
- Dragonfly with hi radix switches
- N/W signaling rate would improve

What We Missed

- Heterogeneous designs
- SIMD width much larger
- Stacked memory: more ports/lower transfer rate
- Machine Learning & short FP
- Massive 500W chips
- Reliability not a show-stopper
- New programming models

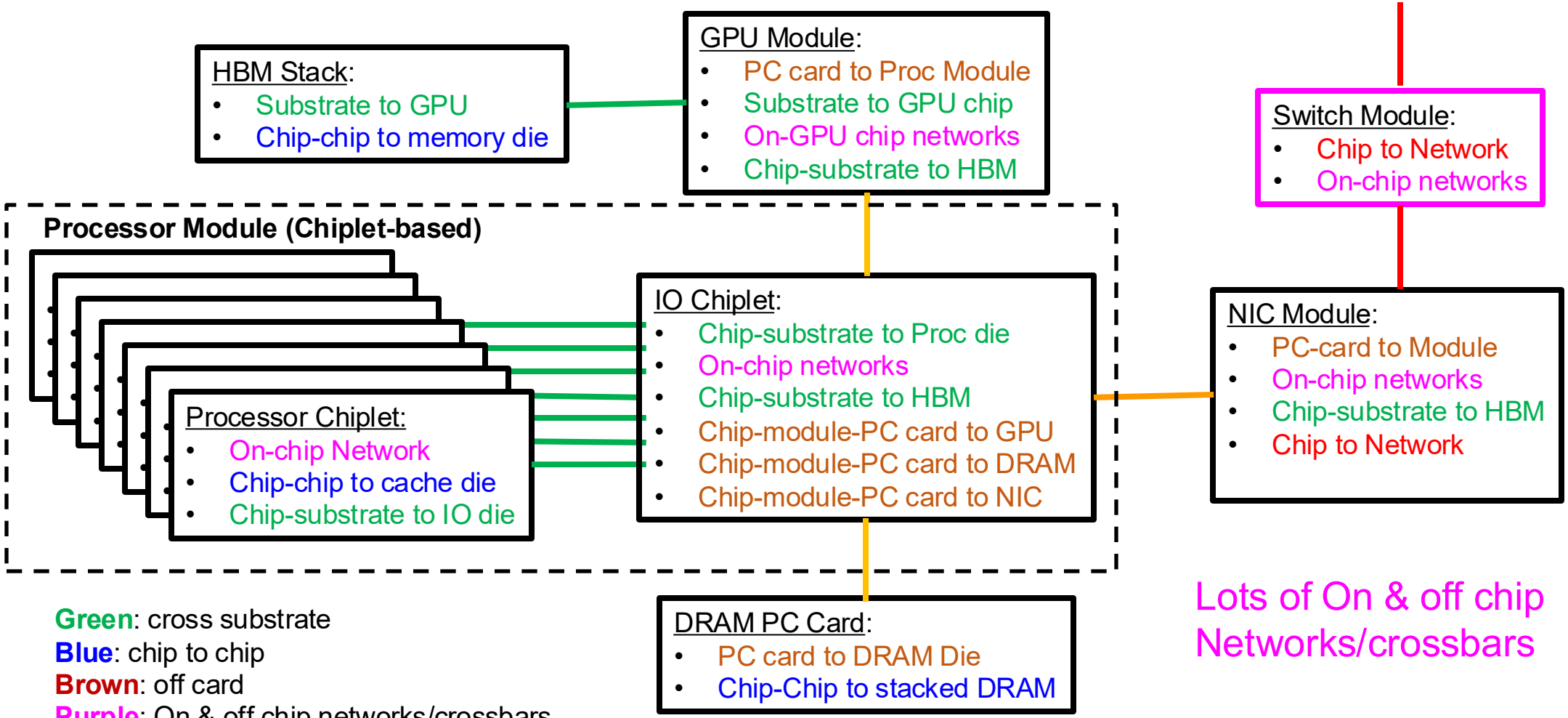
What Has Been Trend of “System Watts/Core”?



- Seems to have been ~ constant since ~2018
- “System” power/core >> chip power per core
- We need to look at “system”

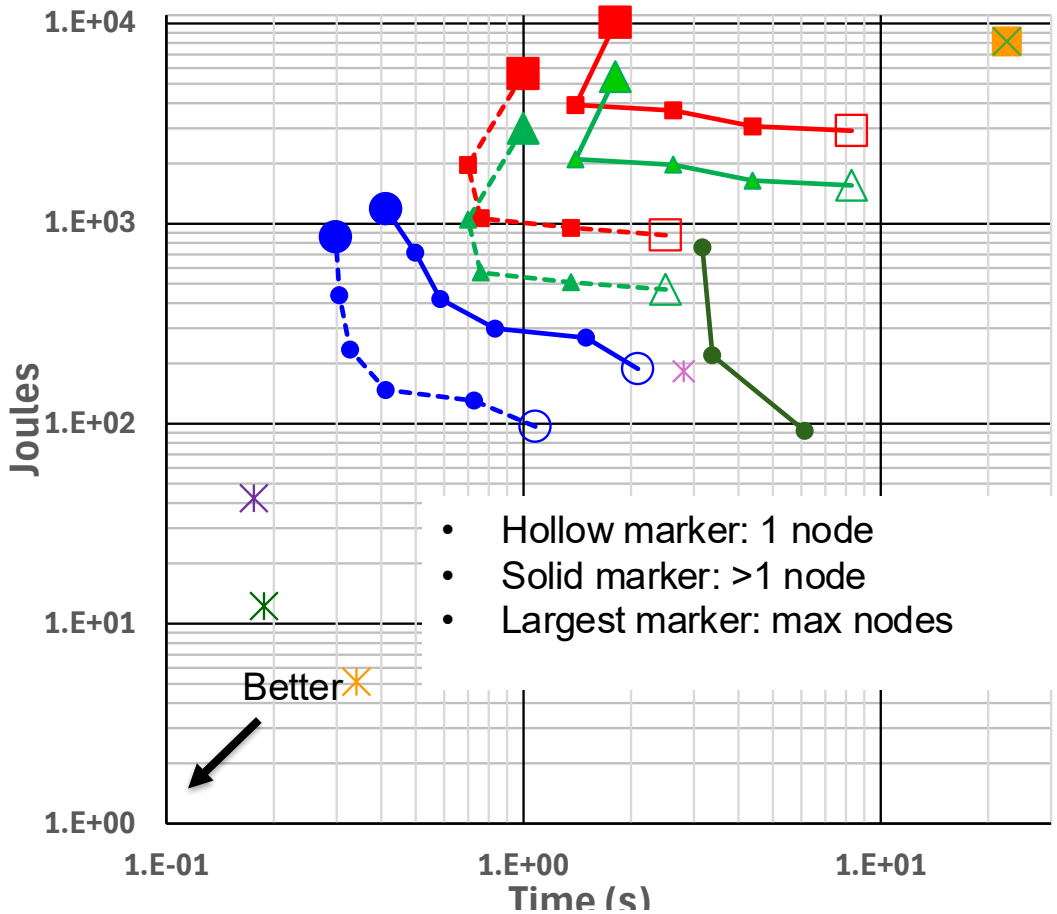
Today's Node Design – Energy of Movement

Same colors = similar functions

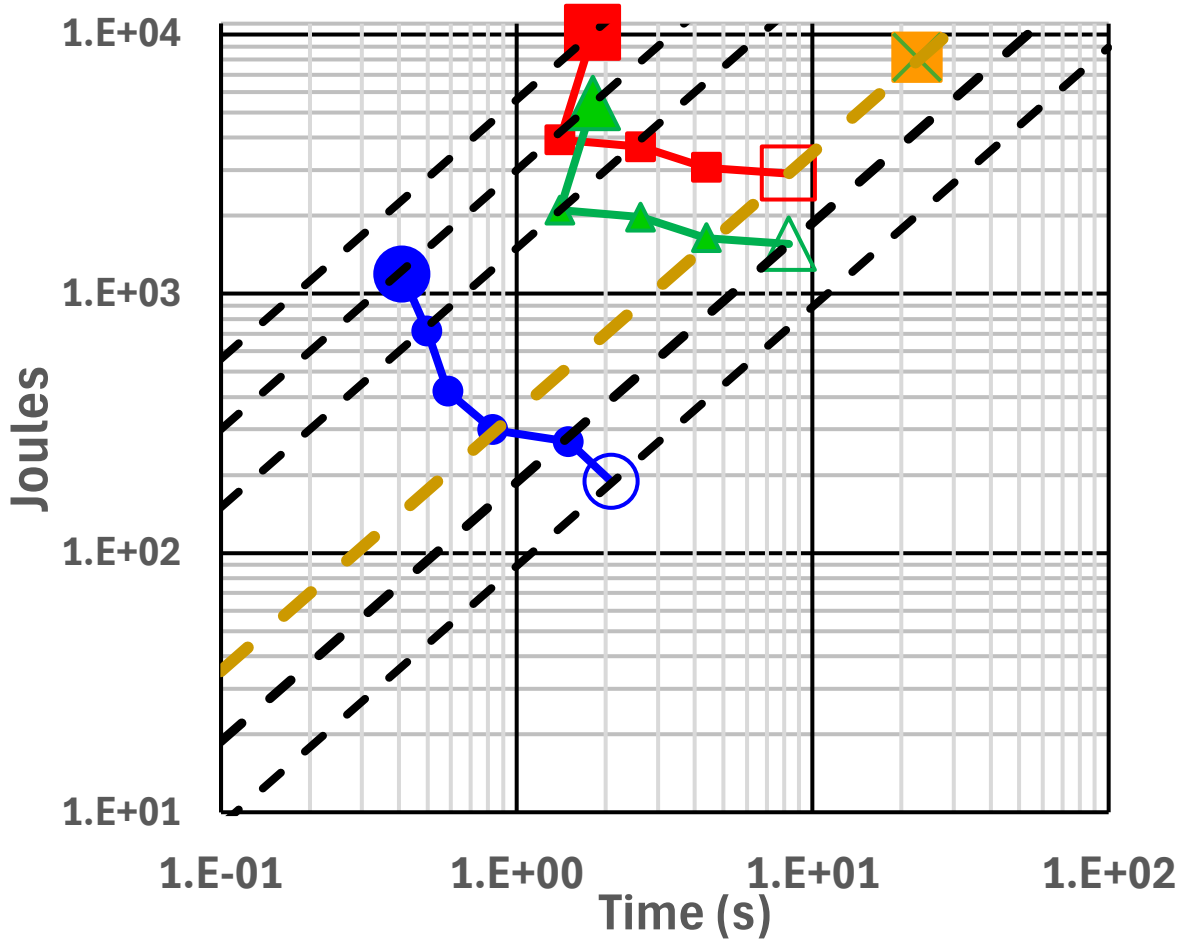


Green: cross substrate
Blue: chip to chip
Brown: off card
Purple: On & off chip networks/crossbars

Energy-Time Trajectories

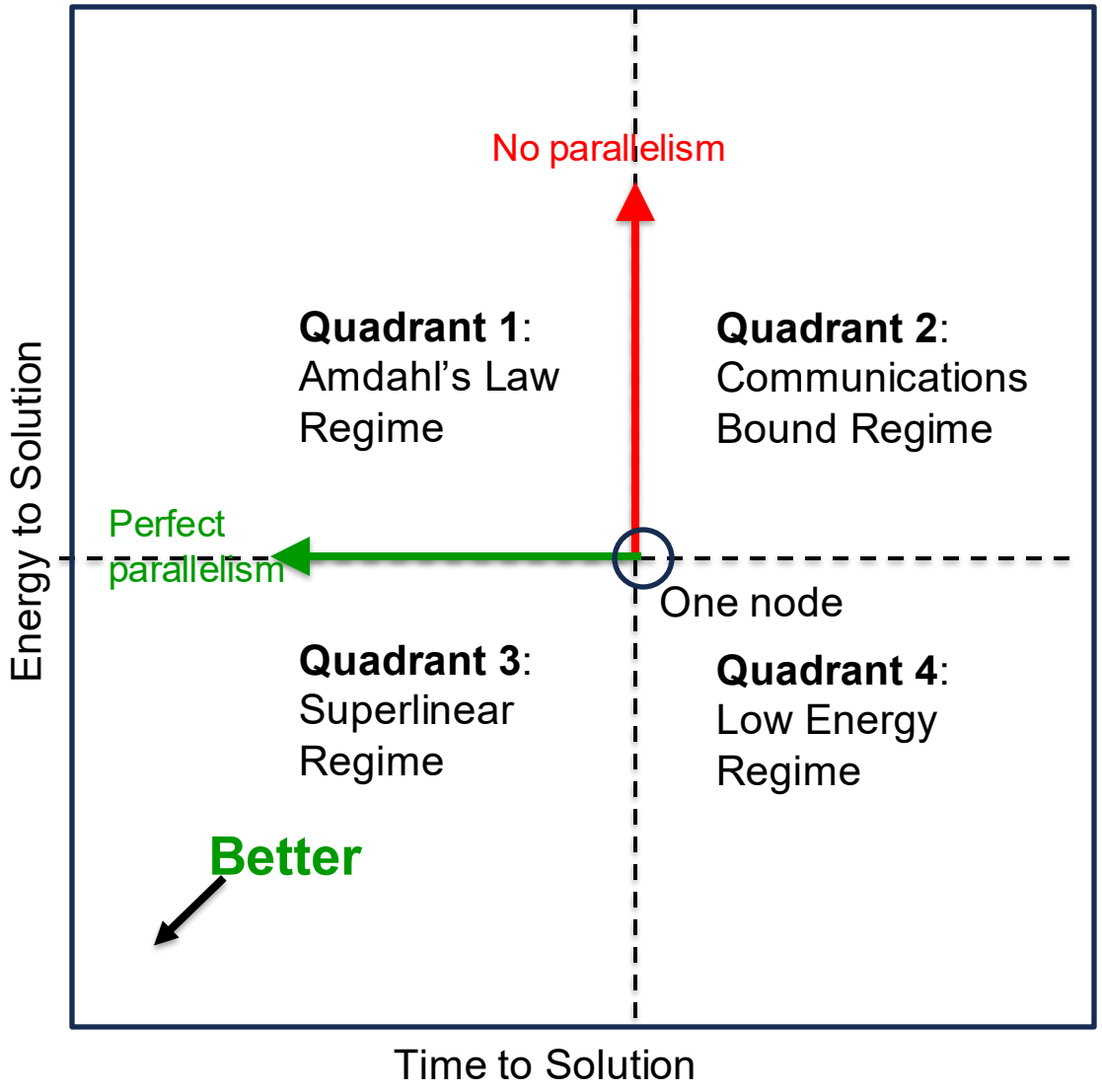


Trajectory: line reflecting using different resources on same problem



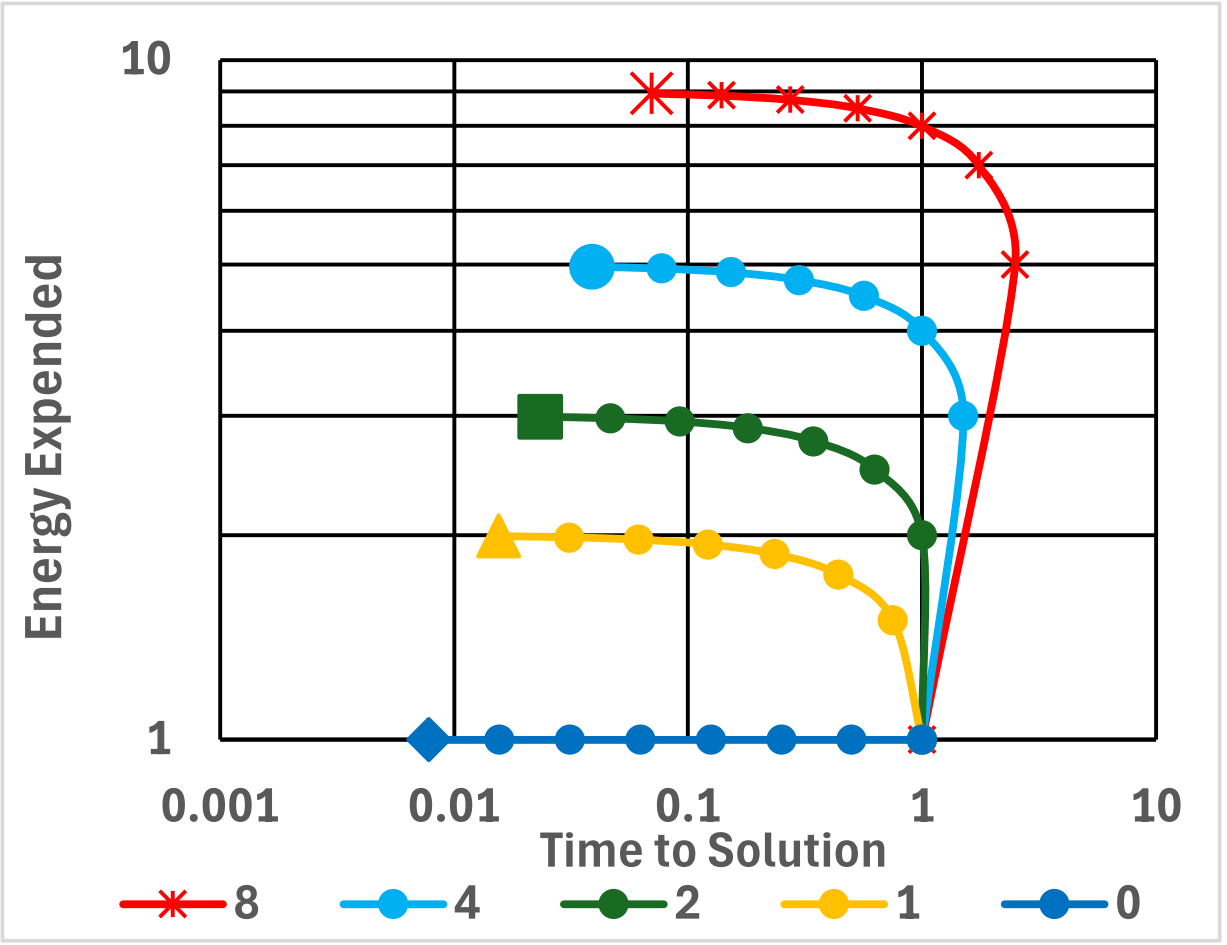
Isopower lines permit comparisons between systems

Quadrants in Energy-Time Space



- **Red Trajectory**: adding more nodes has no effect on time
- **Green Trajectory**: improvement in time matched perfectly by improvement in energy
- **Quadrant 1**: Energy increases as fast as or faster than solution time improves
- **Quadrant 2**: Both Time to solution and energy usage increases
- **Quadrant 3**: Energy and Time decrease
- **Quadrant 4**: Time to solution does not improve, but energy usage does

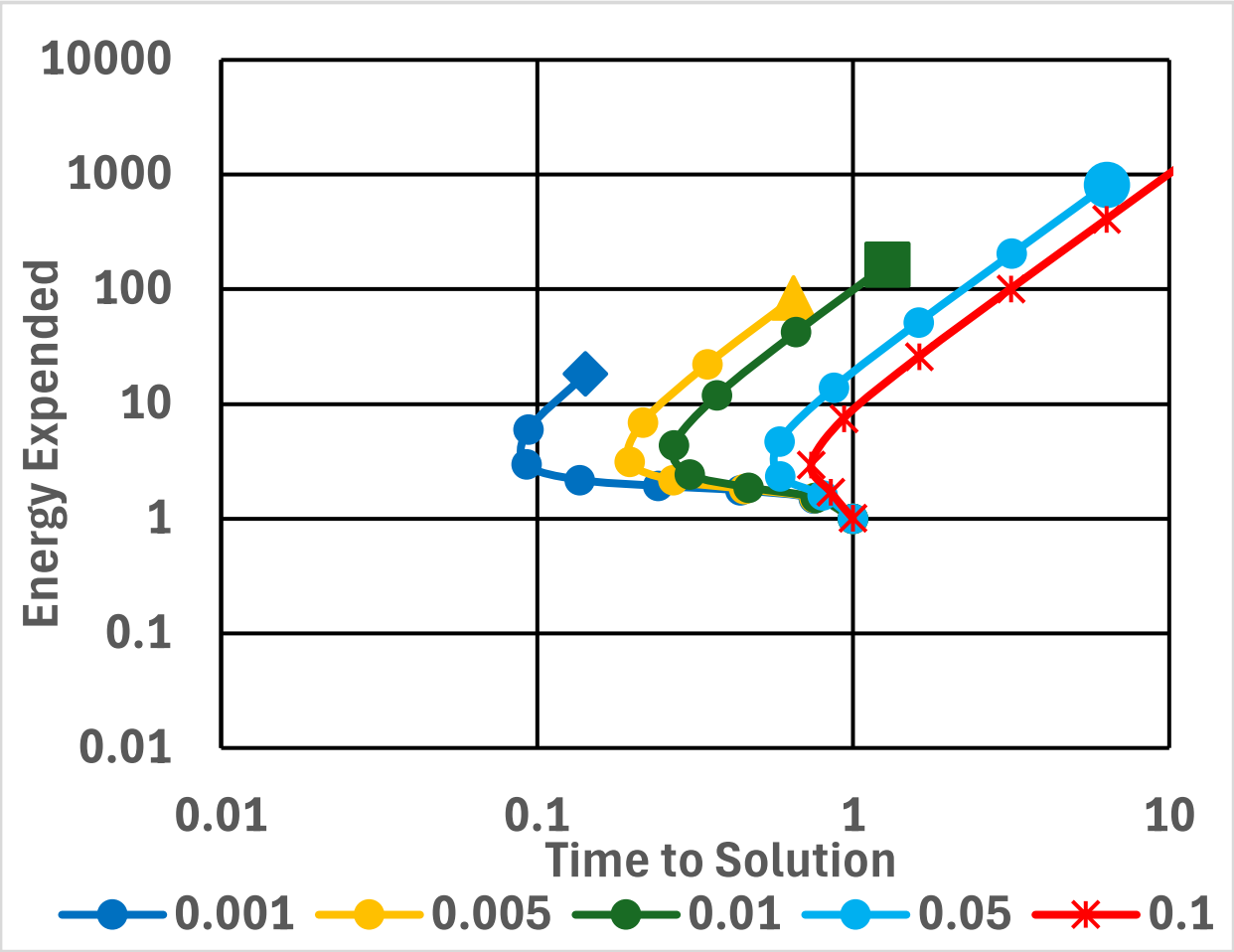
Quadrant 2: A Communication-based Fanout Effect



Different colors are different DC/T ratios

- M = # of “objects”
- D = # of objects “connected to” each object = “fanout”
- T = processing work per object
- T(1) = MT (Amdahl effects later)
- Perfect distribution has M/N objects per node
- Prob. of an object’s neighbor being off-node = (1 – 1/N)
- Cost of one object off-node comm = C
- T(n)/T(1) = (1 + (DC/T)(1-1/N))/n
- Critical Ratio: DC/T
  - Inflection point when DC/T = 2

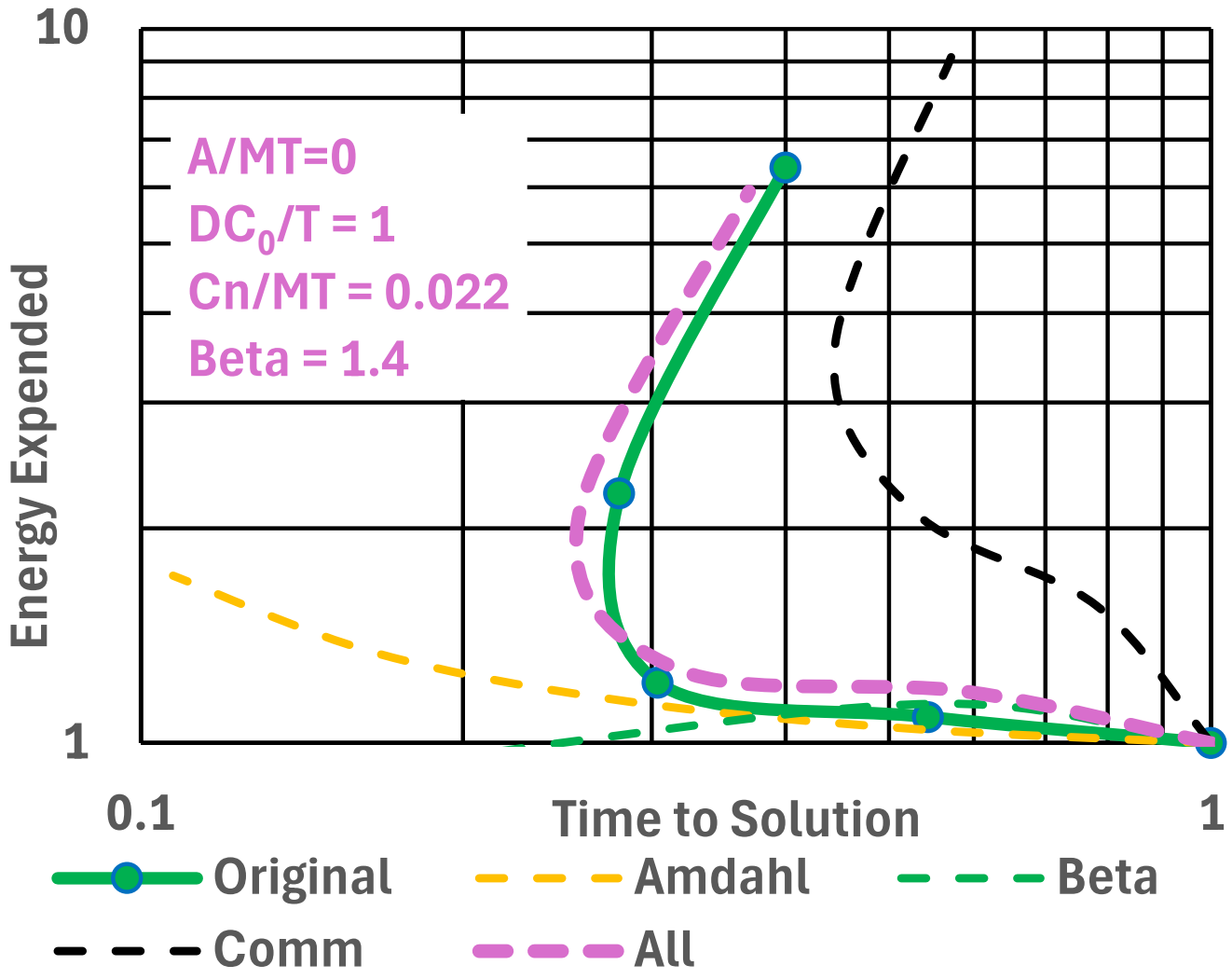
Quadrant 2/3: “Knuckle”: Aggregation-Based Effect



Curves when $DC_0/T=1$ and different values of C_n/MT

- Most modern message-passing systems try to aggregate
- Performance often improves with increasing resources up to a certain point
  - And then get worse!
- $T(n)/T(1) = (1+ (DC_o/T\ )(1-1/n))/n+(C_n/MT)(n-1)$
- Knuckle at $n^2=(1+DC_0/MT)/(C_n/MT)$

Real World Often Requires Multiple Models



Conclusion

- 2008 Strawman was more right than wrong
- Summit in 2018 was actually close!
- Problem is still power, power, and power
- Major issue is in interconnect and memory access
- Energy-Time space helps reveal comm problems
- Aggregation in particular seems to cause “knuckles”
  - Not discussed here but ***alternative architectures*** seem to avoid knuckle effect