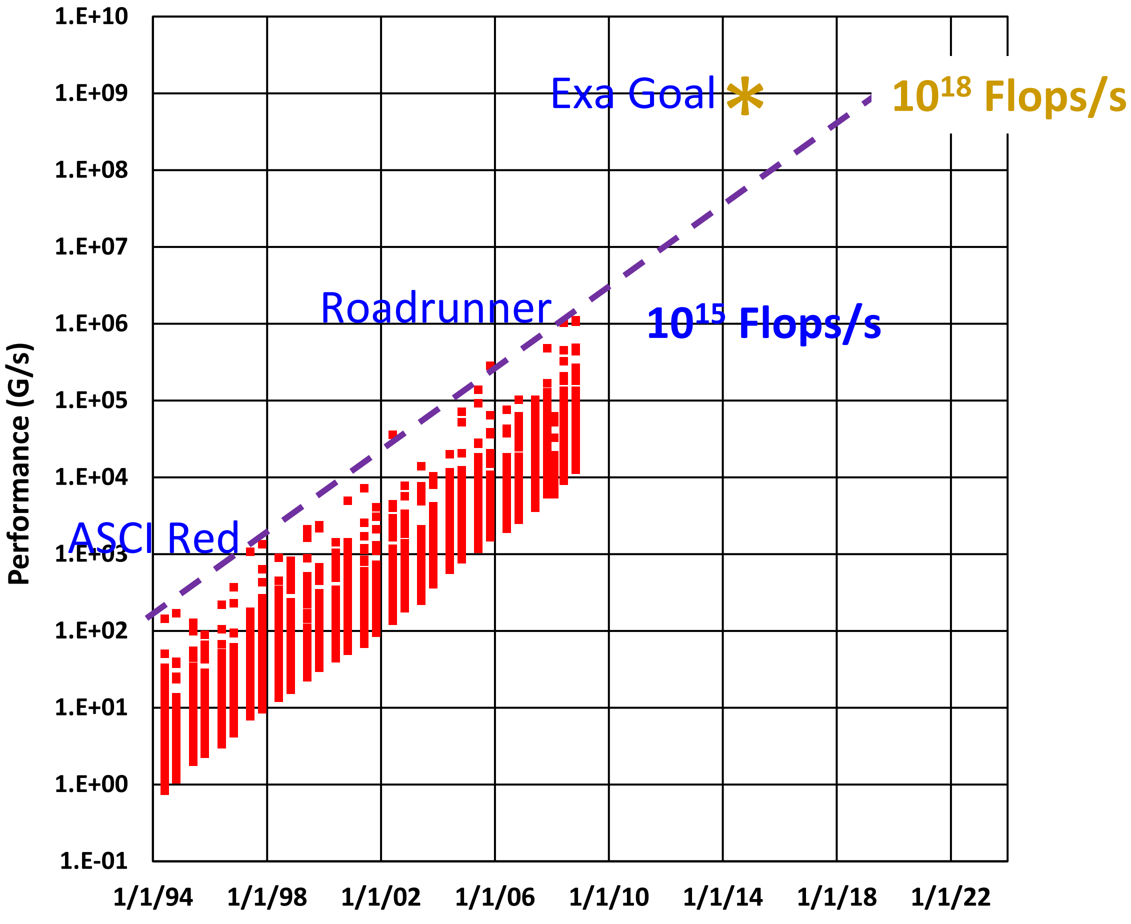
SC22
Dallas, TX | hpc accelerates.

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

Peter M. Kogge
Univ. of Notre Dame

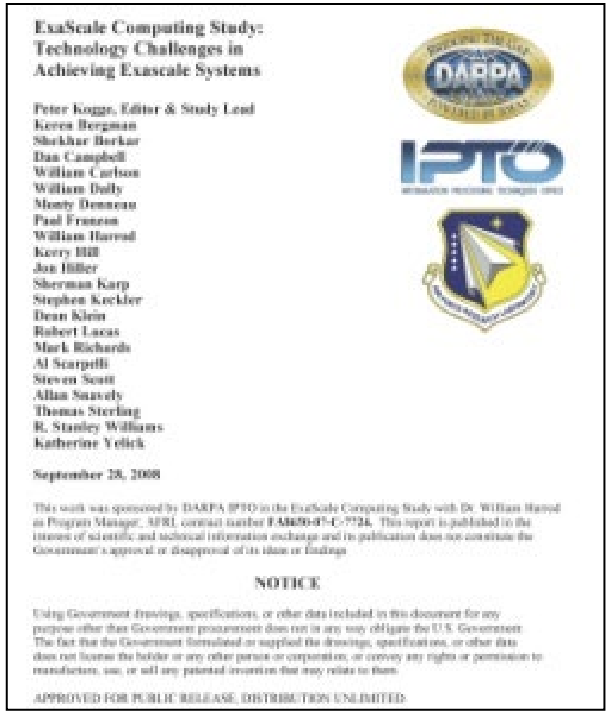
William J. Dally
NVIDIA Corp.

The HPL World in 2008



“First Light” for new TOP500 entries

- Roadrunner: 1+ PF/s
- DARPA (Bill Harrod): Exa by 2015?
- 2008 Exascale Report: Yes, but…



**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

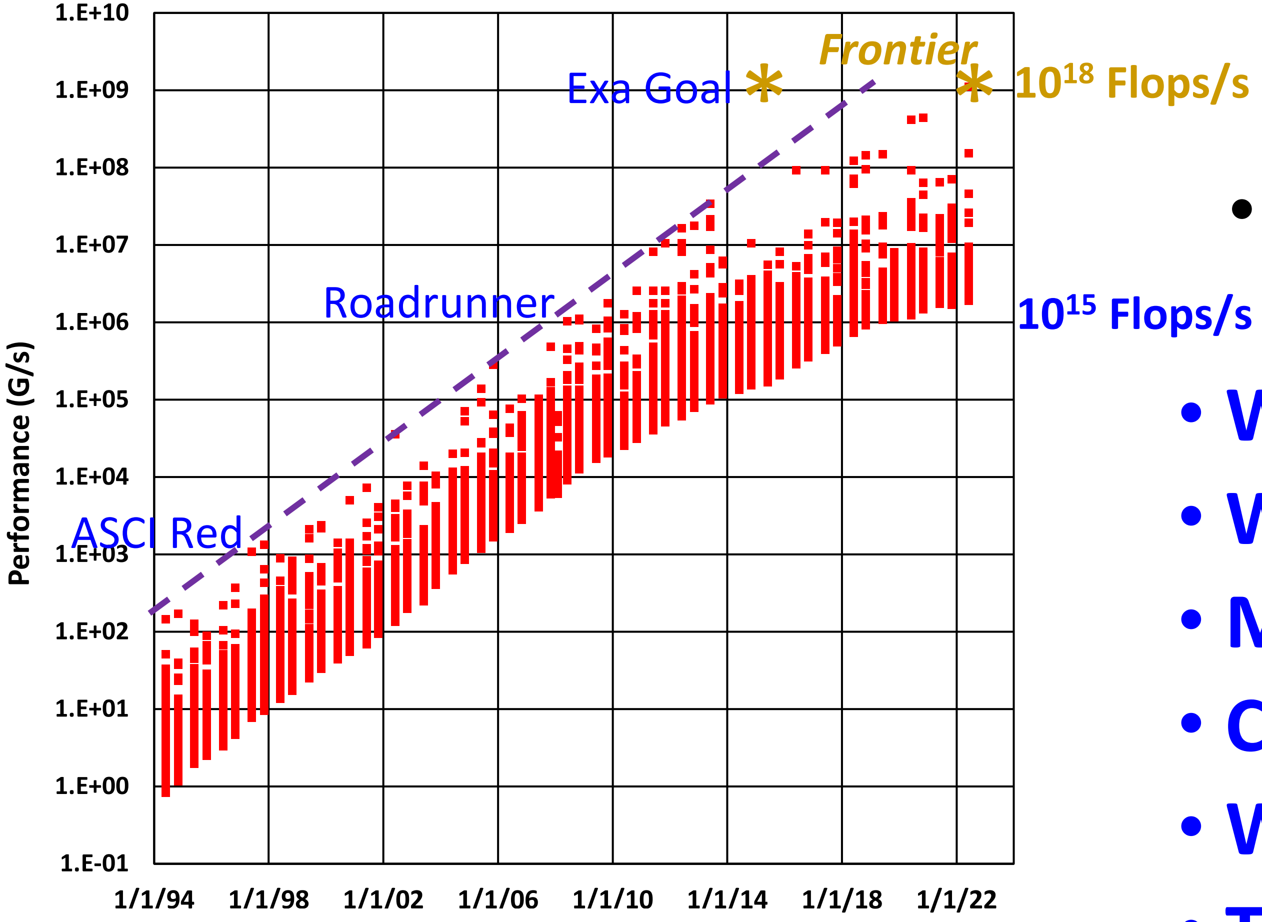
This work was sponsored by DARPA IPTO in the ExaScale Computing Study with Dr. William Harrod as Program Manager; AFRL contract number **FA8650-07-C-7724.** This report is published in the interest of scientific and technical information exchange and its publication does not constitute the Government's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

NOTICE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

APPROVED FOR PUBLIC RELEASE, DISTRIBUTION UNLIMITED.

The HPL World in 2022



"First Light" for new TOP500 entries

- 2022: Frontier Cracks 1EF/s
  - 7 years after Report Goal
  - 4 years after extrapolating curve
- Bounding Curve Changed in 2013

Obvious Questions

- **What Is/Was Exascale?**
- **What Did 2008 Report Predict?**
- **More on the Historical Trail**
- **Comparison to Frontier**
- **What did Report get Right/Wrong?**
- **To Zettascale and Beyond**

The Exascale Study

- **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
  - Memory capacity & bandwidth

Remain

  - Programmability
  - Reliability

Practically Solved

What Was/Is Exascale?

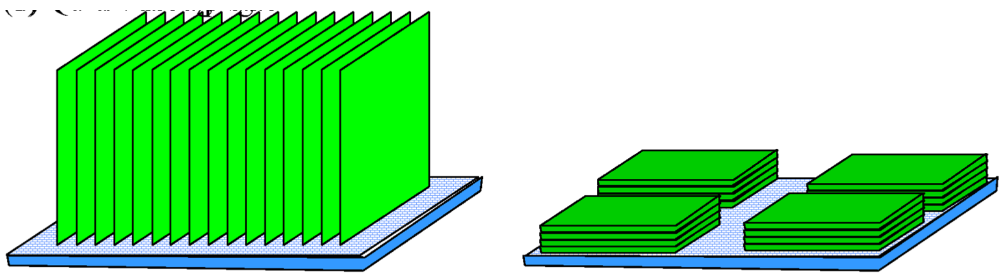
- Report Emphasis: *Try* to change focus from flops
- Goal: ***overall*** 1000x capability over “Petascale” by 2015
  - In Same Footprint for Supercomputer at max 20MW
  - 1000X in a rack (peta scale)
  - 1000X in a module (tera scale)
- Not just flops but
  - Memory
  - Memory Bandwidth
  - Network Bandwidth
  - …
- Plus ability to program massive concurrency

Technologies Investigated

- Logic: power, area, energy, clock
  - CMOS: hi perf/low voltage
  - Options: hybrid, superconducting
  - Voltage scaling
- Main Memory
  - SRAM, DRAM, NAND, Alternatives
  - Reliability, packaging, power
- Storage Memory
  - Disk, Holographical, Archival

- Interconnect: esp. energy
  - On chip
  - DRAM to Processor (Stacking)
  - Intra/inter module
  - Rack to rack
  - Electrical vs optical
- Packaging and Cooling
- Resiliency & Checkpointing
- Programming Models

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

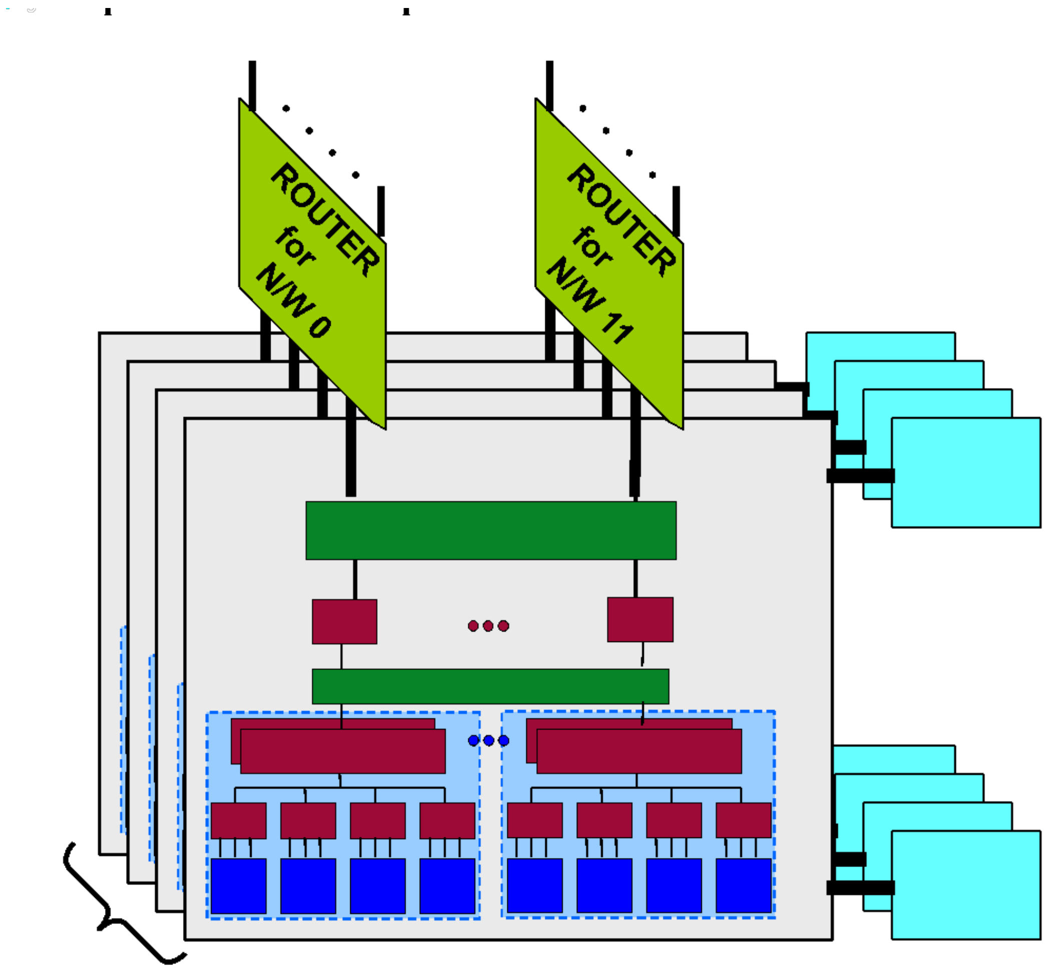
- Includes 16 12GB SATA drives for checkpointing

Cabinet: 32 Groups = 384 nodes

- Assumed max power of 120KW

System: 583 Cabinets, 67MW

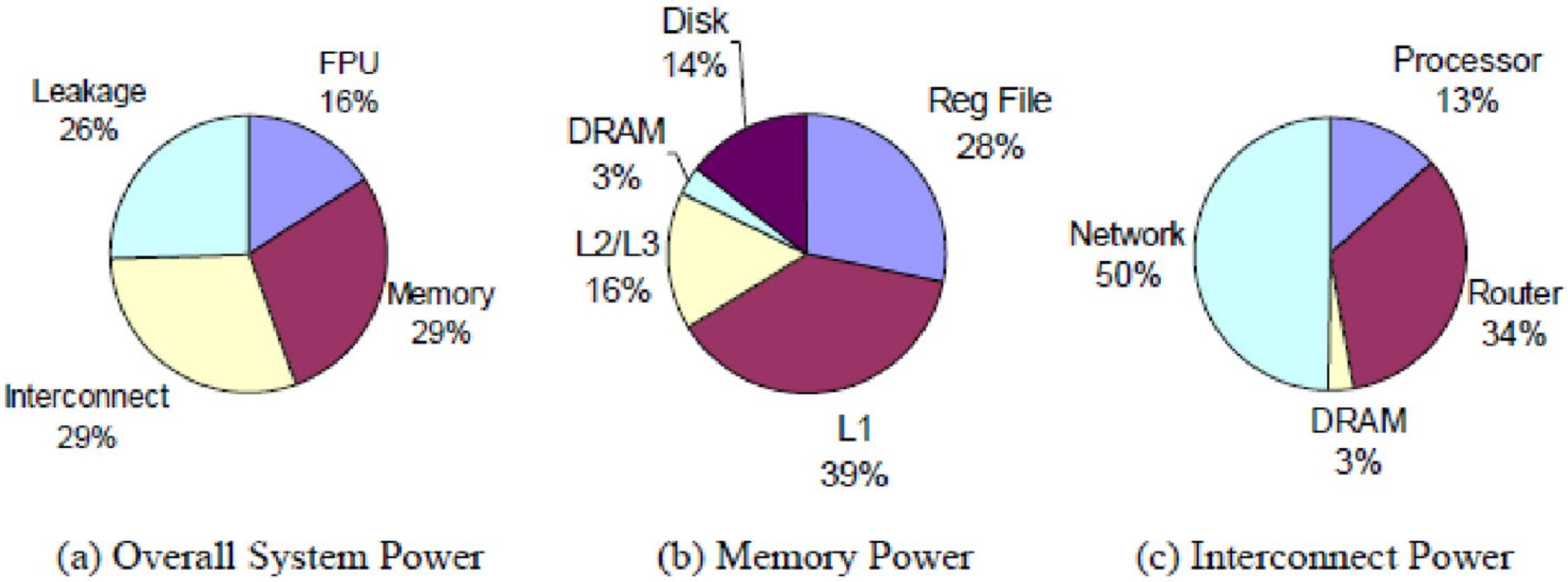
- 3-hop Dragonfly interconnect
- 166 million cores with 664 million FPUs



Est. 14.9 GF/W

WD

Where Did the Energy Go?



(a) Overall System Power

(b) Memory Power

(c) Interconnect Power

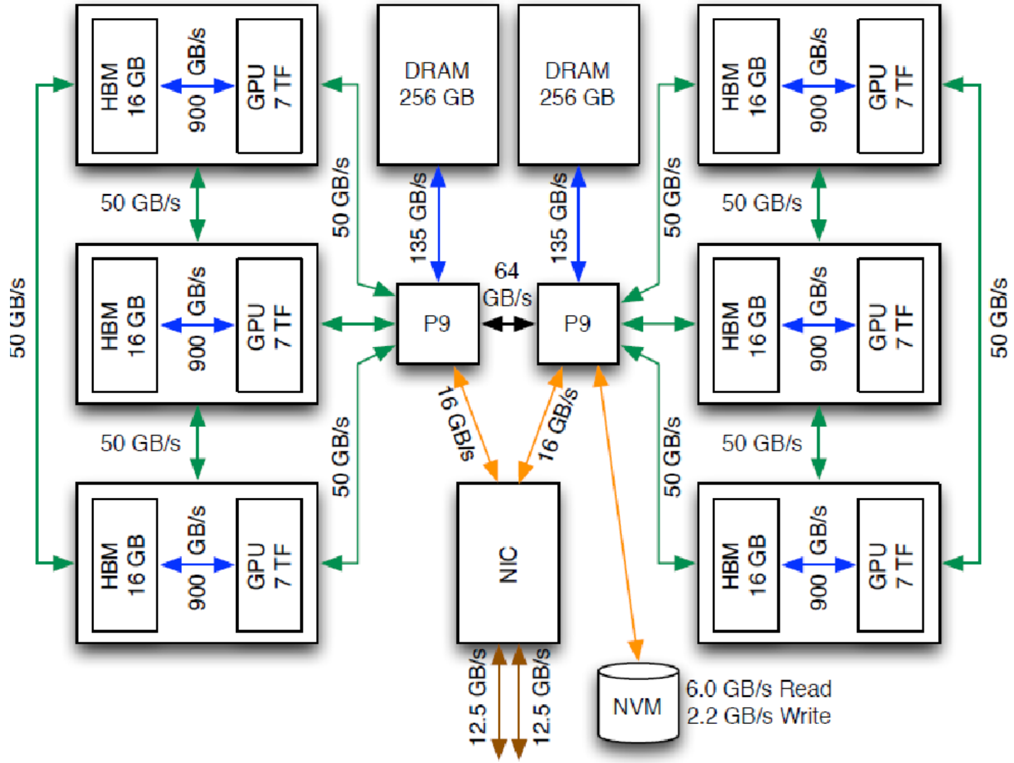
2018: Summit – An Exascale "Could Have Been"

- **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!**



TF	42 TF (6x7 TF)
HBM	96 GB (6x16 GB)
DRAM	512 GB (2x16x16 GB)
NET	25 GB/s (2x12.5 GB/s)
MMsg/s	83

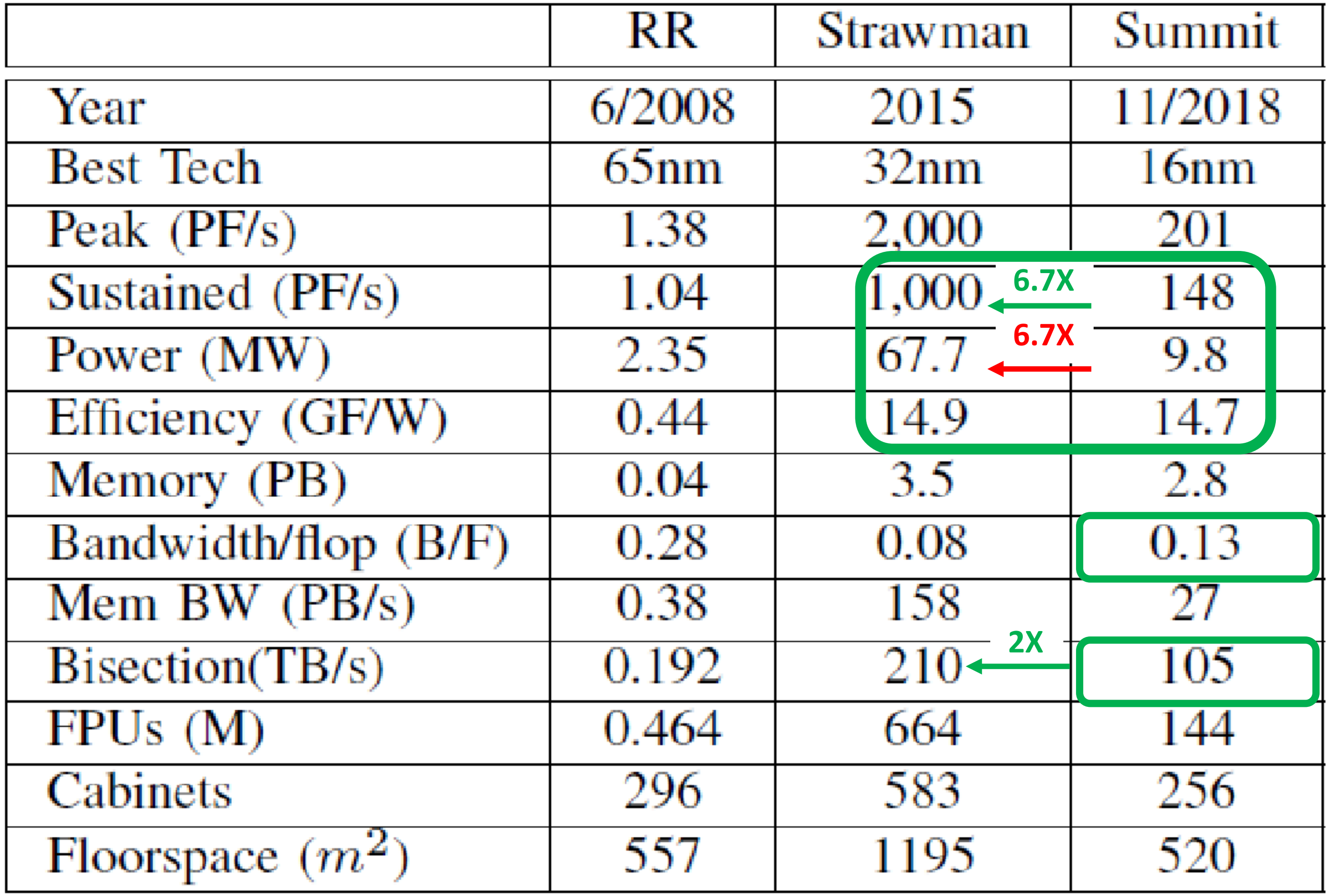
↔ HBM/DRAM Bus (aggregate B/W)
↔ NVLINK
↔ X-Bus (SMP)
↔ PCIe Gen4
↔ EDR IB

HBM & DRAM speeds are aggregate (Read+Write).
All other speeds (X-Bus, NVLink, PCIe, IB) are bi-directional.

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

Measured 14.7 GF/W

Strawman vs. Summit



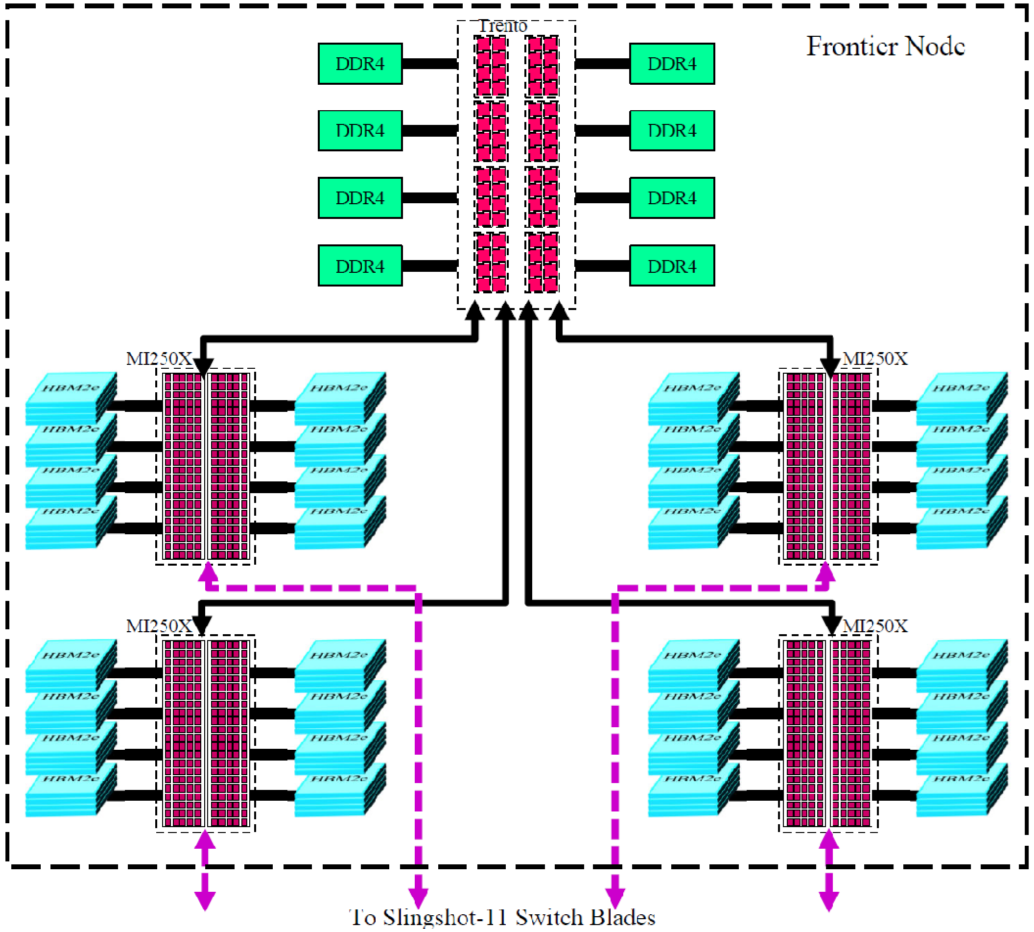
	RR	Strawman	Summit
Year	6/2008	2015	11/2018
Best Tech	65nm	32nm	16nm
Peak (PF/s)	1.38	2,000	201
Sustained (PF/s)	1.04	1,000	148
Power (MW)	2.35	67.7	9.8
Efficiency (GF/W)	0.44	14.9	14.7
Memory (PB)	0.04	3.5	2.8
Bandwidth/flop (B/F)	0.28	0.08	0.13
Mem BW (PB/s)	0.38	158	27
Bisection(TB/s)	0.192	210	105
FPUs (M)	0.464	664	144
Cabinets	296	583	256
Floorspace (m^2)	557	1195	520

63% better

Summit: Could have matched Strawman if scaled up ~6.7X

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

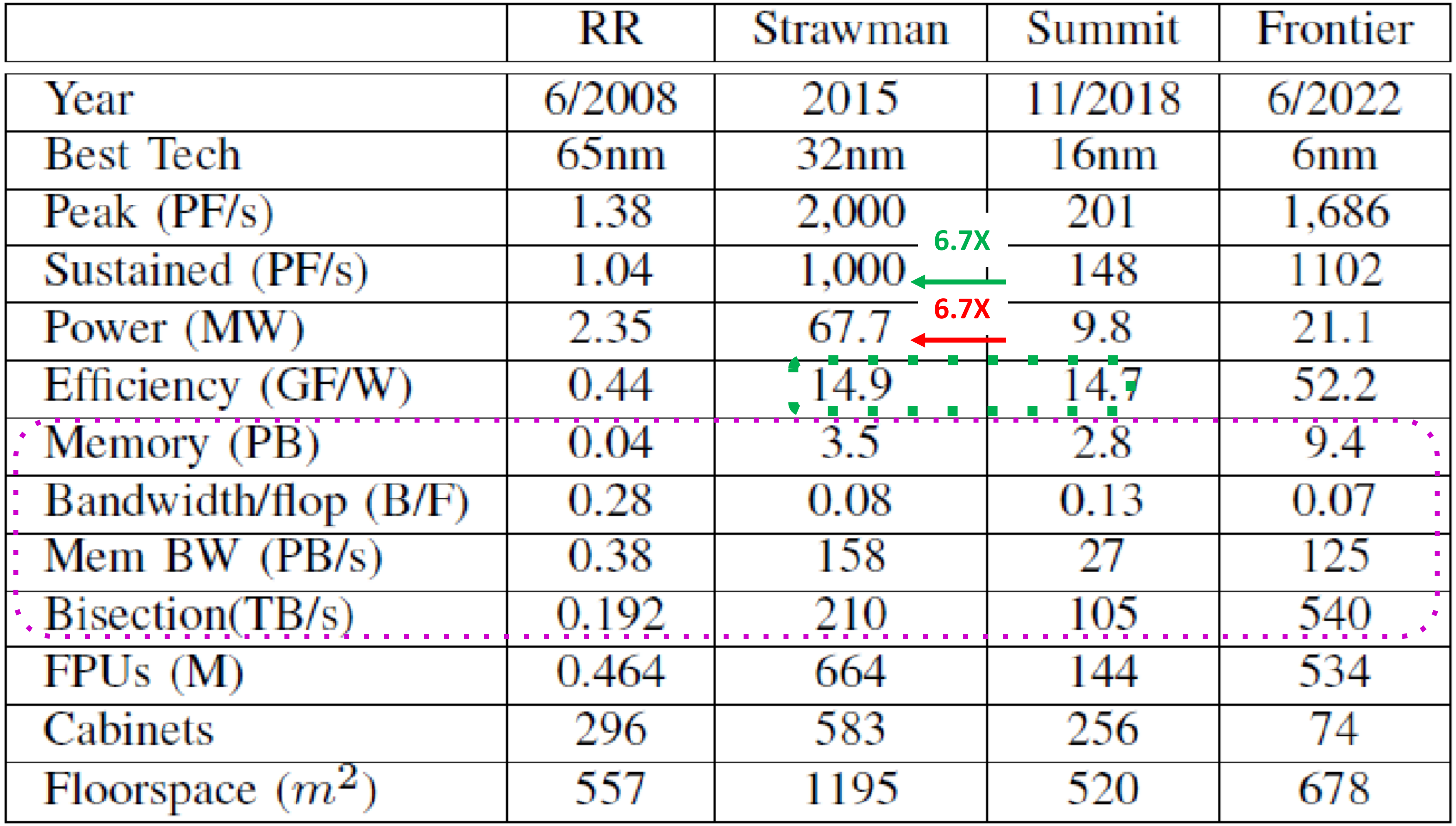
PMK

2022 Frontier System

- **Blade**: 2 nodes
- **Chassis**: 8 Processor Blades
  - With up to 8 Router Blade
  - Arranged perpendicularly
- **Cabinet**: 8 Chassis = 128 nodes
  - Water cooled up to 400KW
  - Over 2X footprint of Strawman
- **System**: 74 compute cabinets
  - With *additional* Cooling Units
  - Again Dragonfly topology

By OLCF at ORNL - https://www.flickr.com/photos/olcf/52117623843/, CC BY 2.0, https://commons.wikimedia.org/w/index.php?curid=119231238

More Detailed Comparison



	RR	Strawman	Summit	Frontier
Year	6/2008	2015	11/2018	6/2022
Best Tech	65nm	32nm	16nm	6nm
Peak (PF/s)	1.38	2,000	201	1,686
Sustained (PF/s)	1.04	1,000	148	1102
Power (MW)	2.35	67.7	9.8	21.1
Efficiency (GF/W)	0.44	14.9	14.7	52.2
Memory (PB)	0.04	3.5	2.8	9.4
Bandwidth/flop (B/F)	0.28	0.08	0.13	0.07
Mem BW (PB/s)	0.38	158	27	125
Bisection(TB/s)	0.192	210	105	540
FPUs (M)	0.464	664	144	534
Cabinets	296	583	256	74
Floorspace (m^2)	557	1195	520	678

6.7X

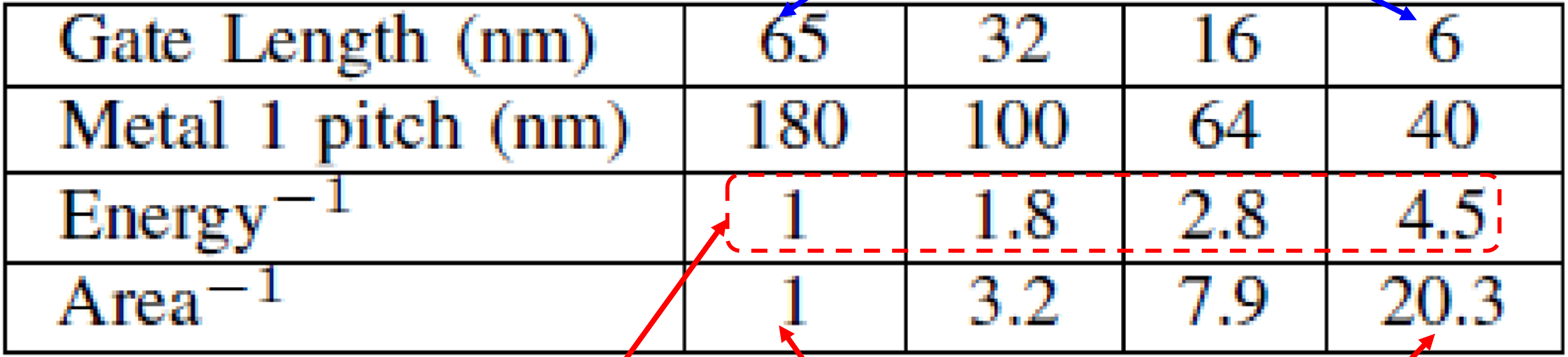
6.7X

Summit: Could have matched Strawman if scaled up ~6.7X

Frontier: not even close to 1000X over Roadrunner in other categories

Technology Changes from Then to Now

>10X

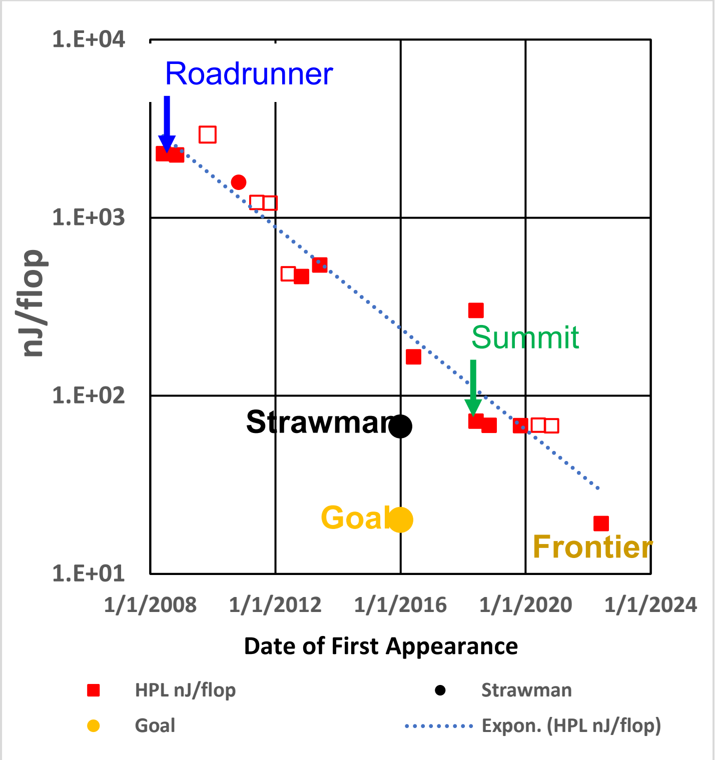


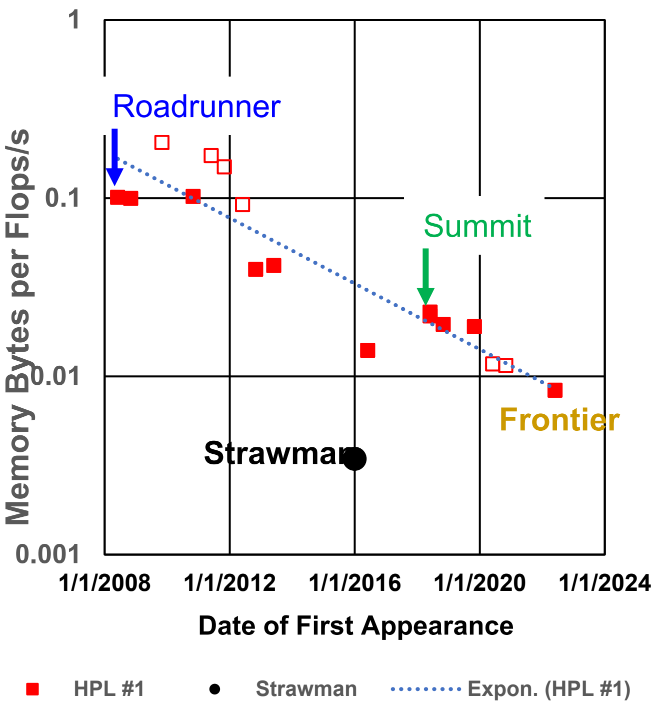
	Roadrunner	Strawman	Summit	Frontier
Gate Length (nm)	65	32	16	6
Metal 1 pitch (nm)	180	100	64	40
Energy^{-1}	1	1.8	2.8	4.5
Area^{-1}	1	3.2	7.9	20.3

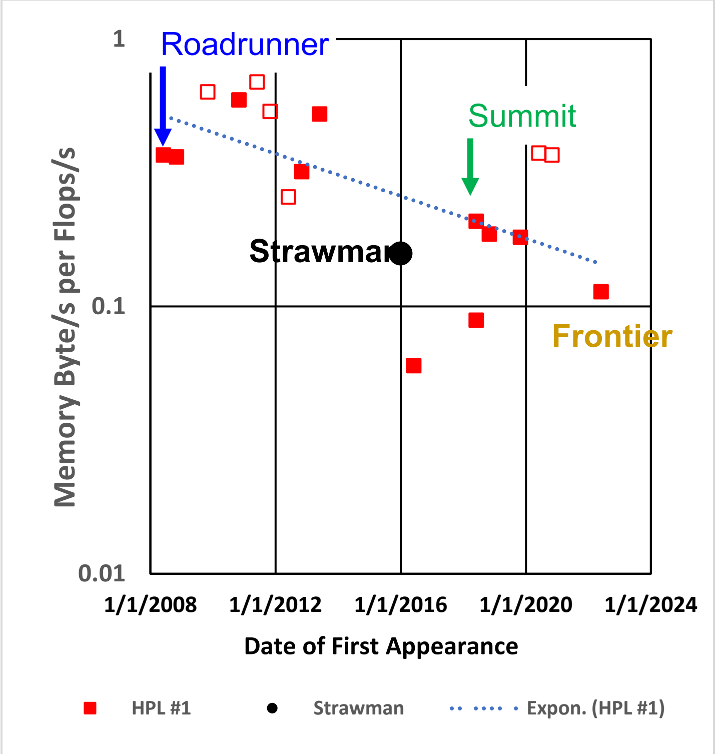
Also nowhere near Dennard Scaling: Voltage scaling has stopped

Nowhere near 100X: Metal 1 pitch dominates density: Not transistors

Changes in System Characteristics







Energy/Flop:

- **Declined >100X** since 2008
- Summit matched Strawman in 2018

Memory Capacity vs Flops/s:

- **Declined >10X** since 2008
- Strawman was even worse

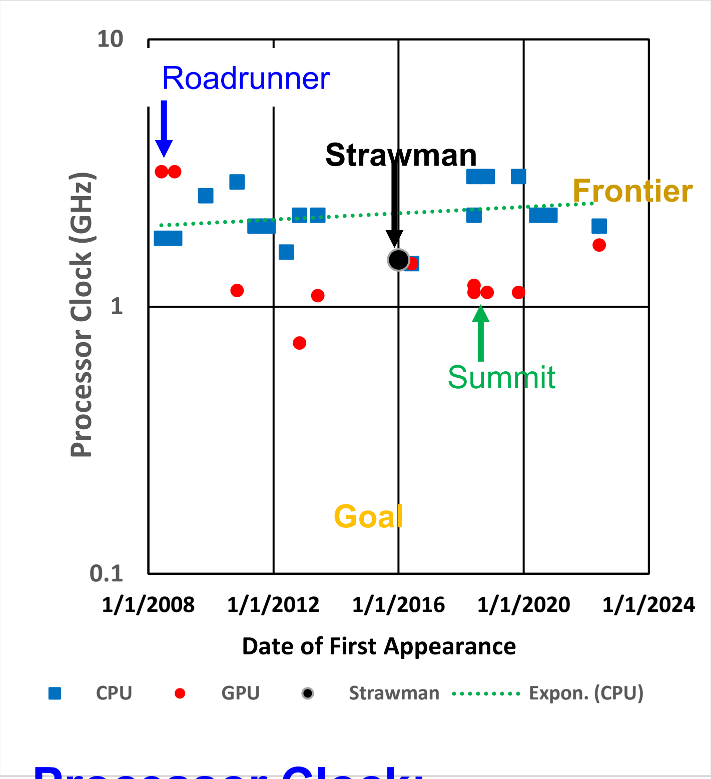
Memory Bandwidth vs Flops/s:

- **Declined >3X** since 2008
- Strawman was down 2X

🟥 Heterogeneous Processor Architecture

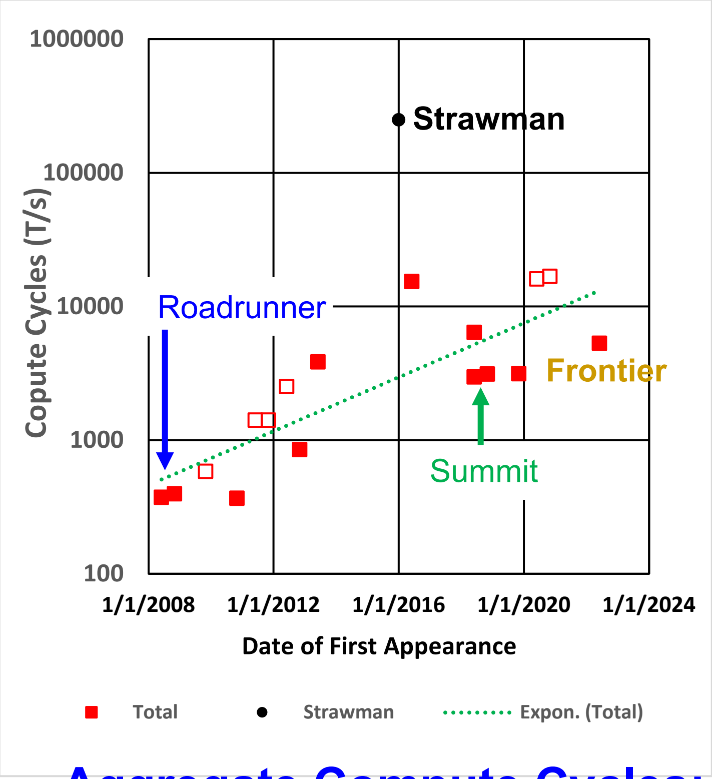
□ Homogeneous Processor Architecture

Changes in Architecture Characteristics



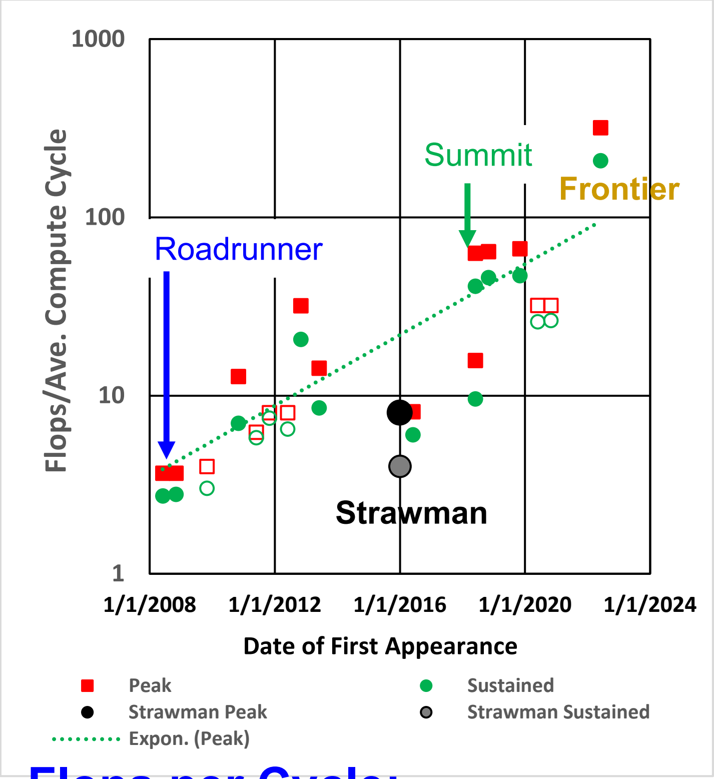
Processor Clock:

- **Essentially Flat since 2008**
- GPUs ran slower than CPUs



Aggregate Compute Cycles:

- **Increased >14X** since 2008
- Strawman had huge # of cores
  - But only 4 FPU wide each



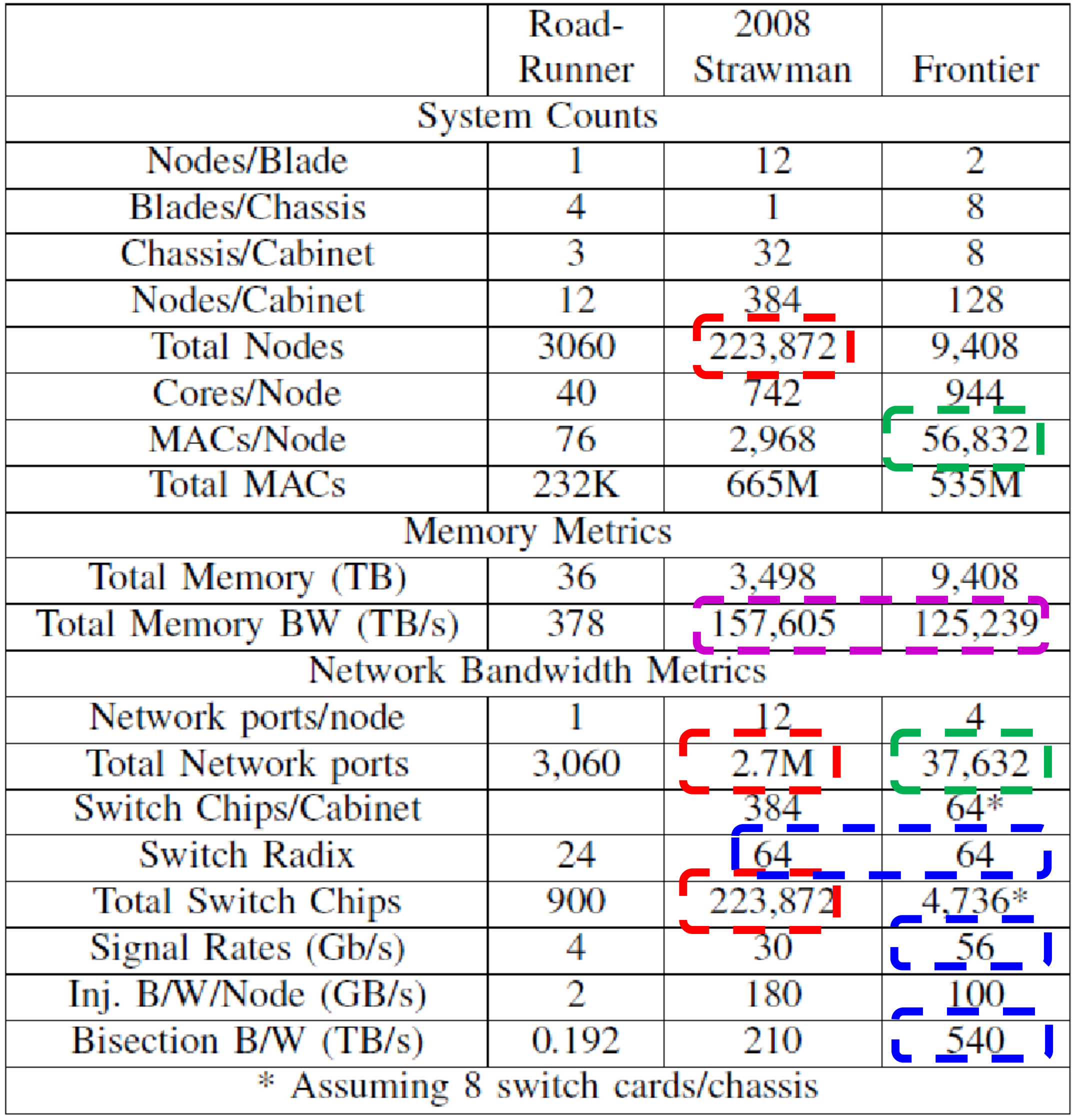
Flops per Cycle:

- **Exploded with Advent of GPU**
- Strawman didn’t go far enough

🟥 Heterogeneous Processor Architecture

□ Homogeneous Processor Architecture

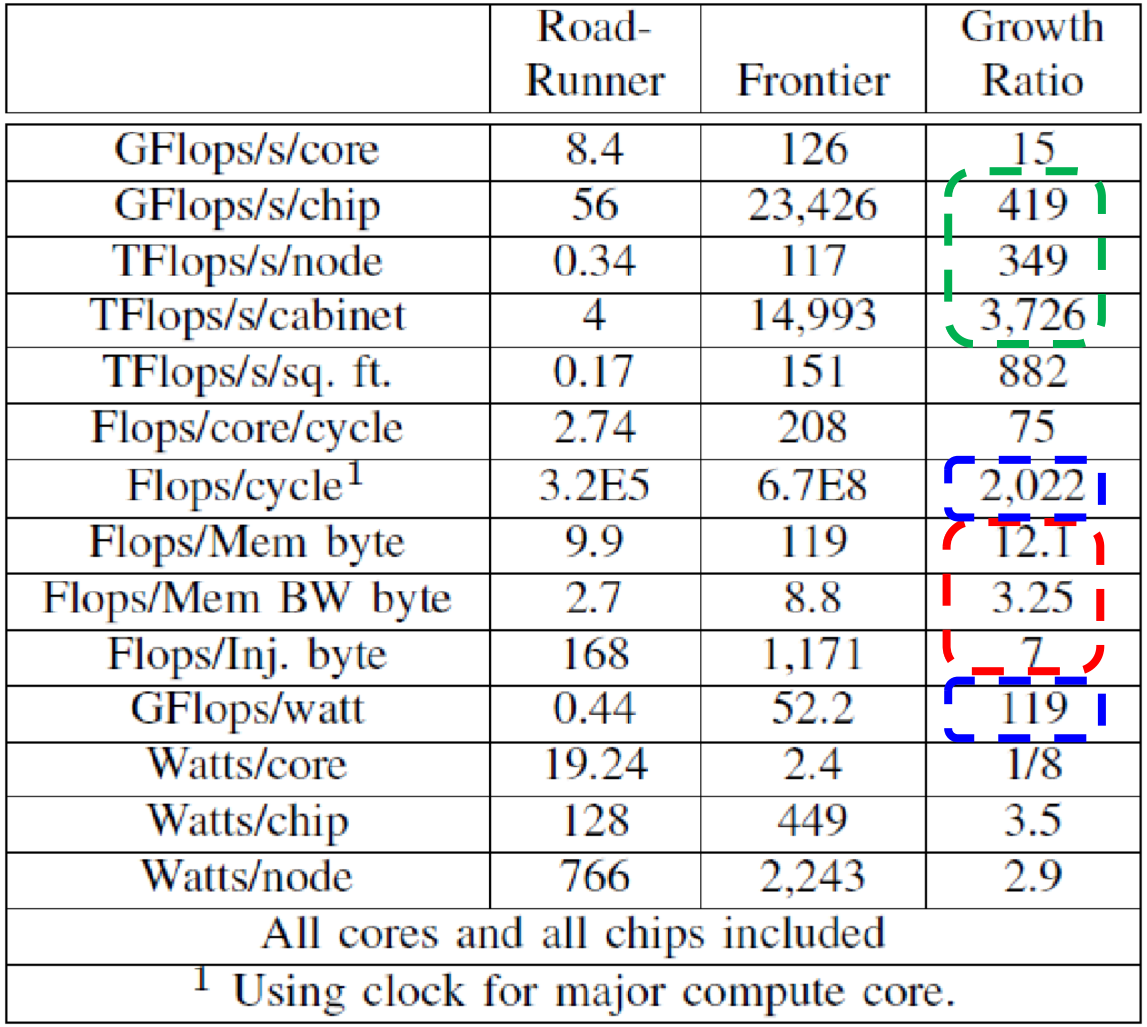
Frontier vs Strawman



	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

Frontier vs Roadrunner: Did We Get 1000X?



	Road-Runner	Frontier	Growth Ratio
GFlops/s/core	8.4	126	15
GFlops/s/chip	56	23,426	419
TFlops/s/node	0.34	117	349
TFlops/s/cabinet	4	14,993	3,726
TFlops/s/sq. ft.	0.17	151	882
Flops/core/cycle	2.74	208	75
Flops/cycle[1]	3.2E5	6.7E8	2,022
Flops/Mem byte	9.9	119	12.1
Flops/Mem BW byte	2.7	8.8	3.25
Flops/Inj. byte	168	1,171	7
GFlops/watt	0.44	52.2	119
Watts/core	19.24	2.4	1/8
Watts/chip	128	449	3.5
Watts/node	766	2,243	2.9
All cores and all chips included			
[1] Using clock for major compute core.			

- Flops/s exceeded 1000X / cabinet
    - But huge cabinets
    - Within 3X for chip & node
- >100X in flops/s per watt
    - And flops/cycle
- Miserable increase in Memory, Memory Bandwidth. N/W Injection Bandwidth

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 coolable
- Reliability not a show-stopper
- New programming models

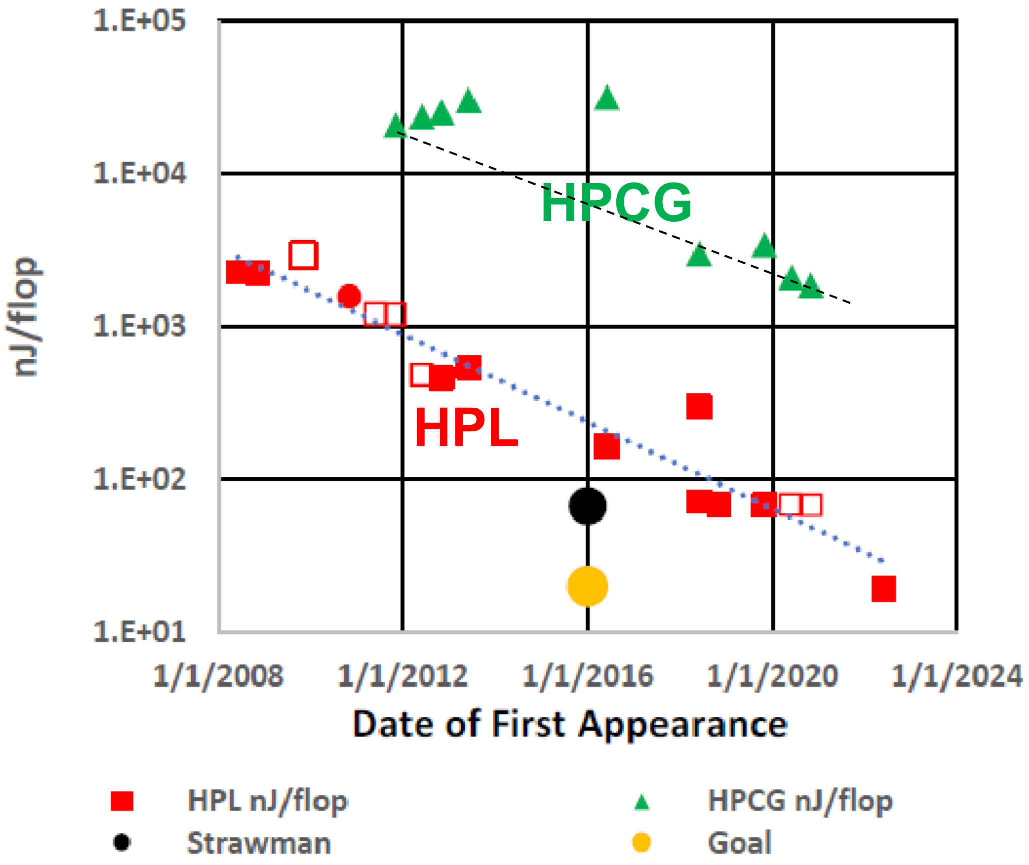
“Zettascale” in 2036?

- Zettascale HPL (10^{21} flops/s) not feasible
  - 64bit FPU might go from today’s 10pJ to 2-3pJ
  - Just math path of ZettaFLOPS HPL machine would consume 2-3GW
- Better: 1000X for today’s critical apps *in same footprint*
  - Multi-physics esp. Climate modeling; Molecular dynamics; …
  - Machine Learning; Bioinformatics; …
- Non-starter: Technology scaling
  - Effective gate lengths may drop 3+X to 1-2nm
  - But metal pitch unlikely to improve significantly
  - 3D stacking might give 8X, but costly & little energy improvement

Bridges to Zettascale

- Efficiency via Specialization
  - Reduced precision & specialized data types & operators
  - Memory system specialized to minimize data movement
    - E.g. 15,000X for bioinformatics accelerator
- Reduce design costs via chiplets
  - Design just the accelerator core, not the whole system
- Growth of AI into Scientific Computation
  - Orders of magnitude improvement on some problems
- Explicit Support for Sparsity
  - Fine grain memory to avoid overfetch
  - Finer-grained transfer on networks for better small-message traffic
  - Efficient scatter/gather, pointer walkers

Example HPCG: Same App as HPL but Sparse Data



- Far less energy efficient
  - H/W resources underutilized
- Insufficient memory B/W
  - Need 8-10 memory bytes/flop
- Rate of improvement not as much

Clearly "Flops at all costs" *not* long term general solution

Conclusions

- 2008 Study nailed need for SIMD many-core, stacked memory, networks based on high radix switches
- But 2013 technology was insufficient
  - Too many endpoints, too much power lost to movement
- Frontier leveraged better technology
  - With wider SIMD, multi-die packaging, better networks & cooling
- More nuanced answer to “Did Frontier achieve *exascale* goals?”
  - Yes if flop-intensive
  - Not if memory or bandwidth-intensive
- Zettascale in 2036?
  - FLOPS on HPL not the question, and not feasible at reasonable energy.
  - 1000x on real applications may be possible
  - Specialization - of operations and memory systems
  - AI for science

Thank You!

Esp. Bill Harrod for all the Exascale studies

And to DOE for pushing to fruition