Intro To Parallel Computing

John Urbanic
Parallel Computing Scientist
Pittsburgh Supercomputing Center

Purpose of this talk

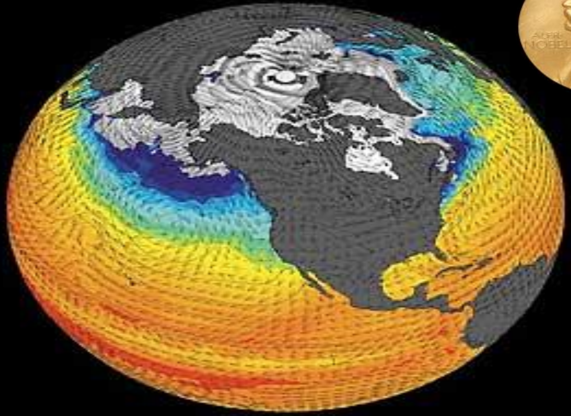
- This is the 50,000 ft. view of the parallel computing landscape. We want to orient you a bit before parachuting you down into the trenches to deal with MPI.

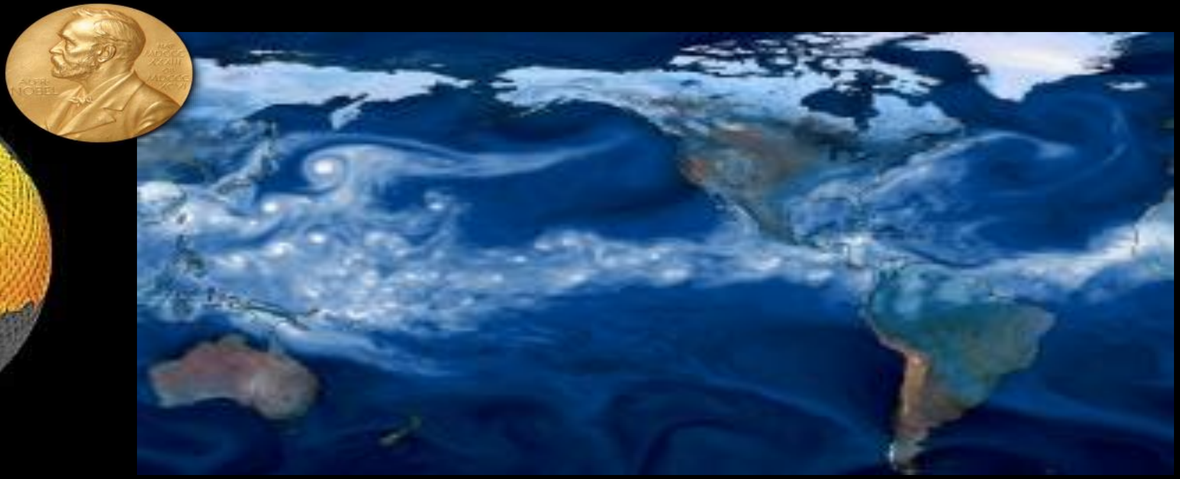
- This talk bookends our technical content along with the Outro to Parallel Computing talk. The Intro has a strong emphasis on hardware, as this dictates the reasons that the software has the form and function that it has. Hopefully our programming constraints will seem less arbitrary.

- The Outro talk can discuss alternative software approaches in a meaningful way because you will then have one base of knowledge against which we can compare and contrast.

- The plan is that you walk away with a knowledge of not just MPI, etc. but where it fits into the world of High Performance Computing.

FLOPS we need: Climate change analysis





Simulations

- Cloud resolution, quantifying uncertainty, understanding tipping points, etc., will drive climate to exascale platforms
- New math, models, and systems support will be needed

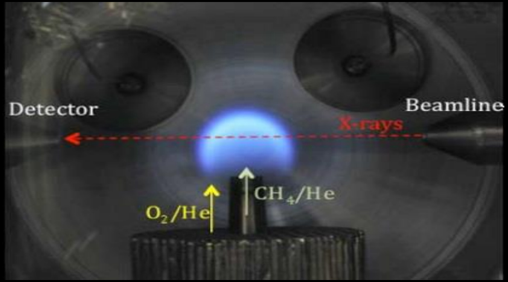
Extreme data

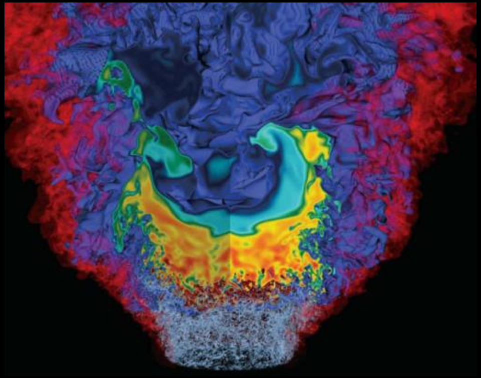
- “Reanalysis” projects need 100× more computing to analyze observations
- Machine learning and other analytics are needed today for petabyte data sets
- Combined simulation/observation will empower policy makers and scientists

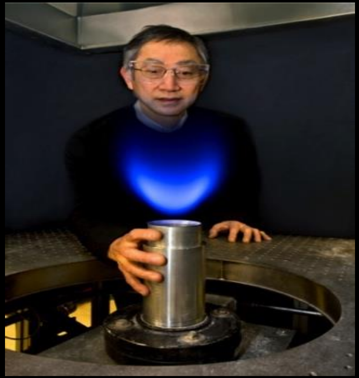
Courtesy Horst Simon, LBNL

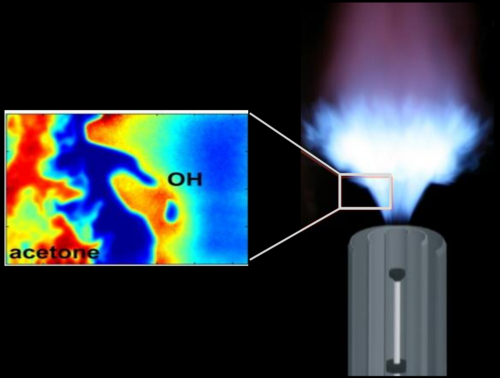
Exascale combustion simulations

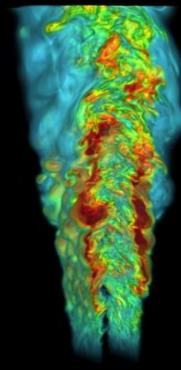
- Goal: 50% improvement in engine efficiency
- Center for Exascale Simulation of Combustion in Turbulence (ExaCT)
  - Combines simulation and experimentation
  - Uses new algorithms, programming models, and computer science







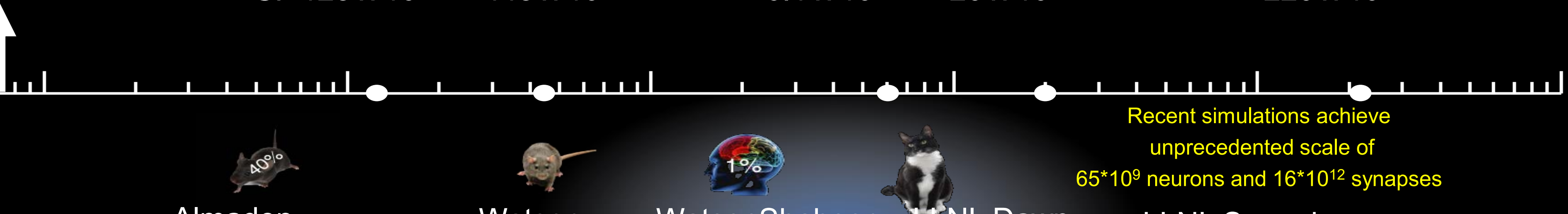




Courtesy Horst Simon, LBNL

Modha Group at IBM Almaden

	Mouse	Rat	Cat	Monkey	Human
N:	16 x 10^6	56 x 10^6	763 x 10^6	2 x 10^9	22 x 10^9
S:	128 x 10^9	448 x 10^9	6.1 x 10^{12}	20 x 10^{12}	220 x 10^{12}



Recent simulations achieve unprecedented scale of $65*10^9$ neurons and $16*10^{12}$ synapses

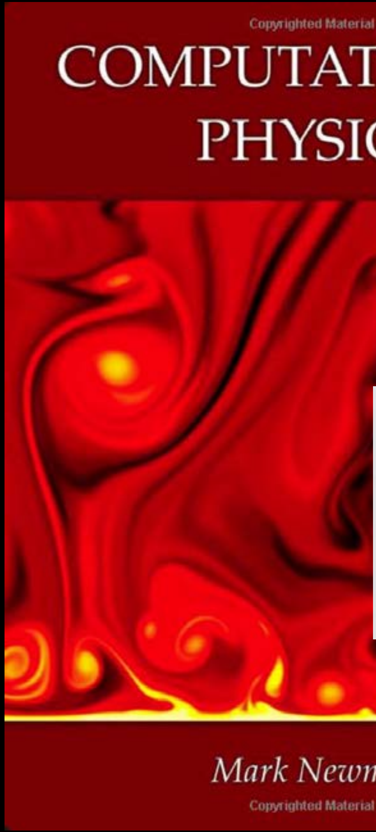
Almaden	Watson	WatsonShaheen	LLNL Dawn	LLNL Sequoia
BG/L	BG/L	BG/P	BG/P	BG/Q
December, 2006	April, 2007	March, 2009	May, 2009	June, 2012

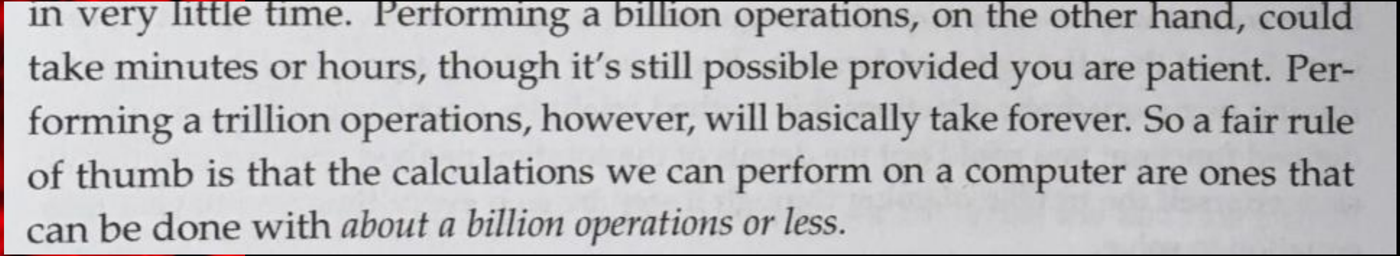
Courtesy Horst Simon, LBNL

'Nuff Said

There is an appendix with many more important exascale challenge applications at the end of our Outro To Parallel Computing talk.

And, many of you doubtless brough your own immediate research concerns. Great!




in very little time. Performing a billion operations, on the other hand, could take minutes or hours, though it's still possible provided you are patient. Performing a trillion operations, however, will basically take forever. So a fair rule of thumb is that the calculations we can perform on a computer are ones that can be done with *about a billion operations or less.*

Welcome to The Year of Exascale!

exa = 10^{18} = 1,000,000,000,000,000,000 = quintillion

64-bit precision floating point operations per second

23,800
Cray Red Storms
2004 (42 Tflops)

133,33
NVIDIA V100
(7.5 Tflops)

Where are those 10 or 12 orders of magnitude?

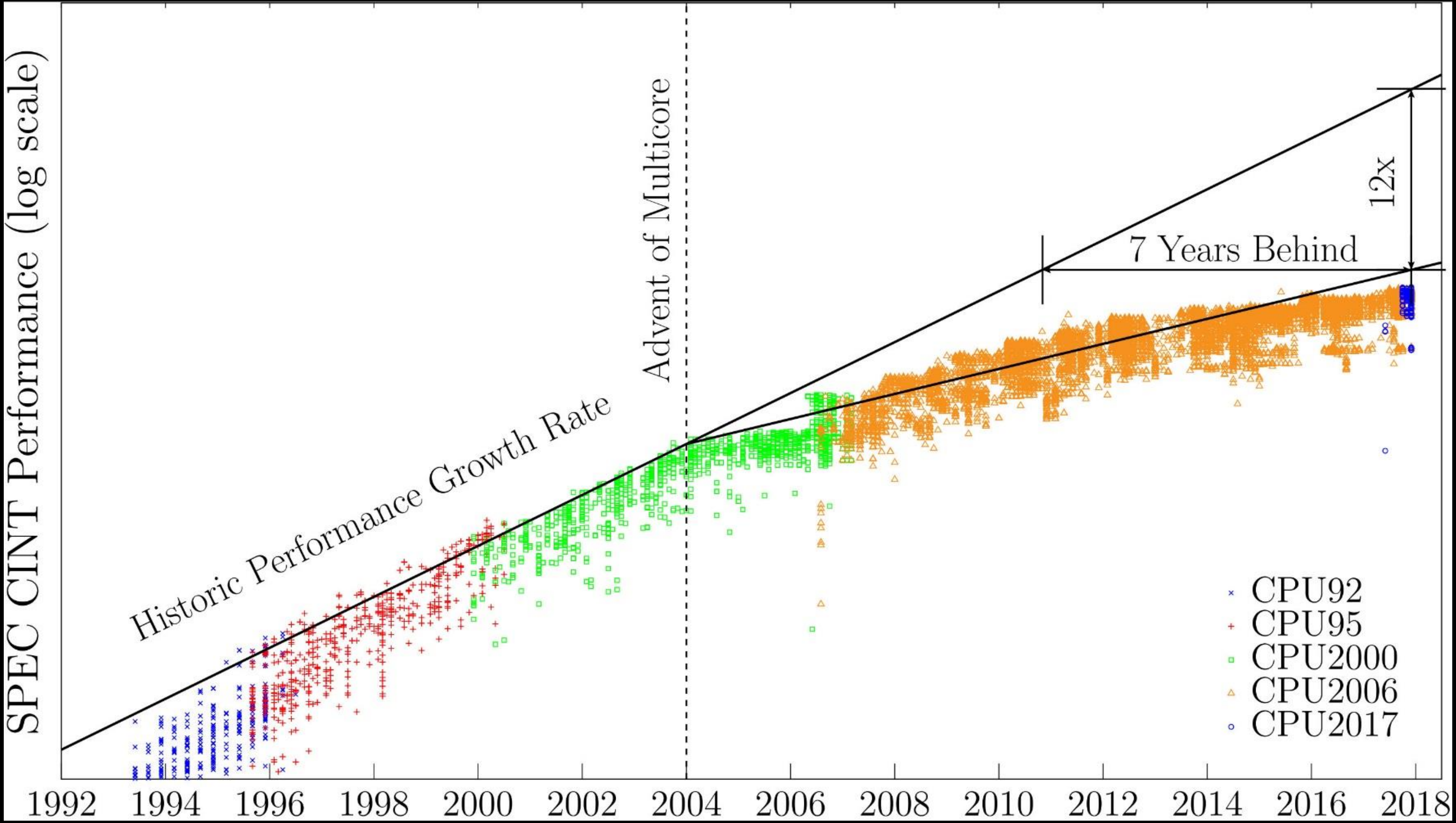
How do we get there from here?

BTW, that's a bigger gap than

VS.

IBM 709
12 kiloflops

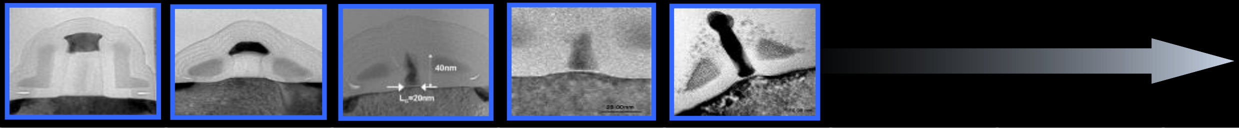
Moore's Law abandoned serial programming around 2004



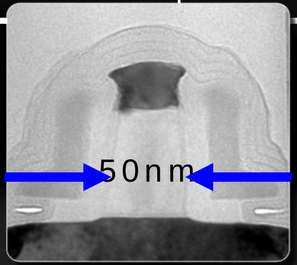
Courtesy Liberty Computer Architecture Research Group

But Moore’s Law is only beginning to stumble now.

Intel process technology capabilities

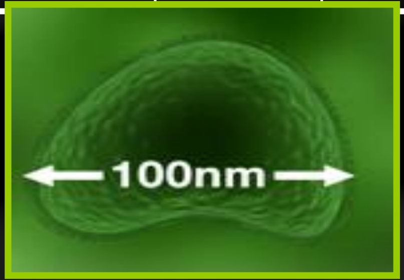


High Volume Manufacturing	2004	2006	2008	2010	2012	2014	2018	2021
Feature Size	90nm	65nm	45nm	32nm	22nm	14nm	10nm	7nm
Integration Capacity (Billions of Transistors)	2	4	8	16	32	64	128	256



Transistor for 90nm Process

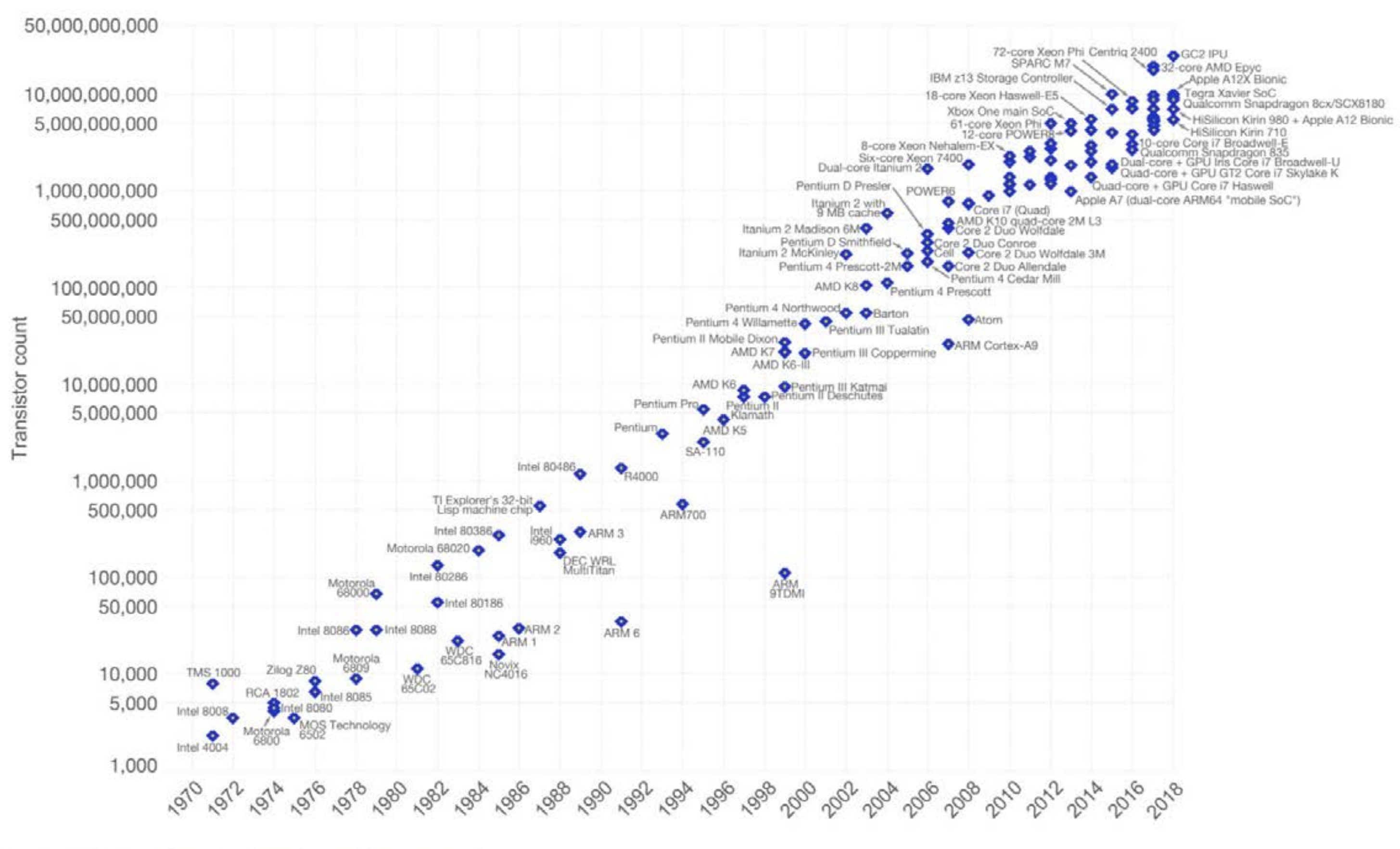
Source: Intel



Influenza Virus

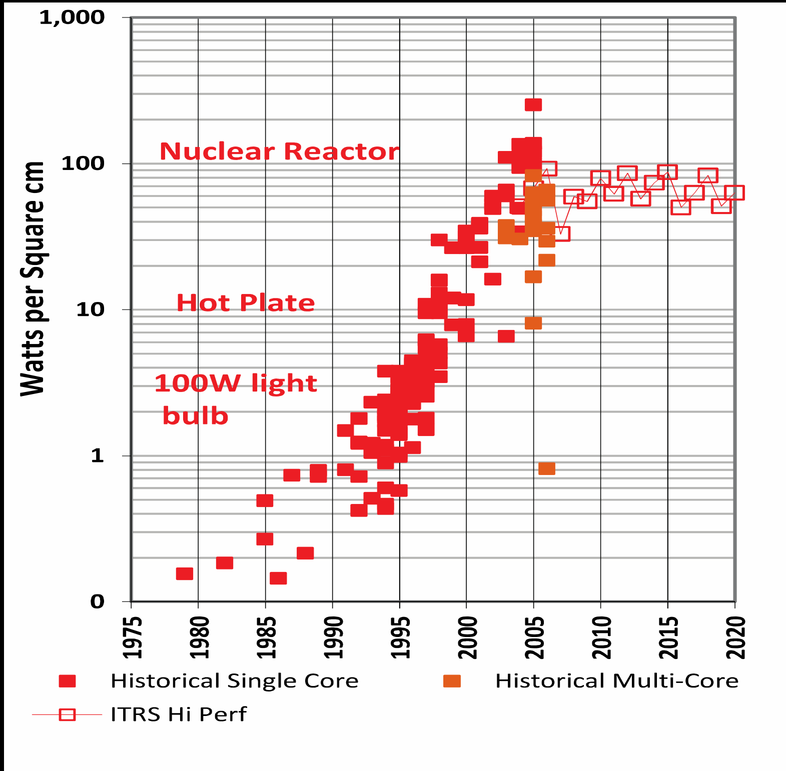
Source: CDC

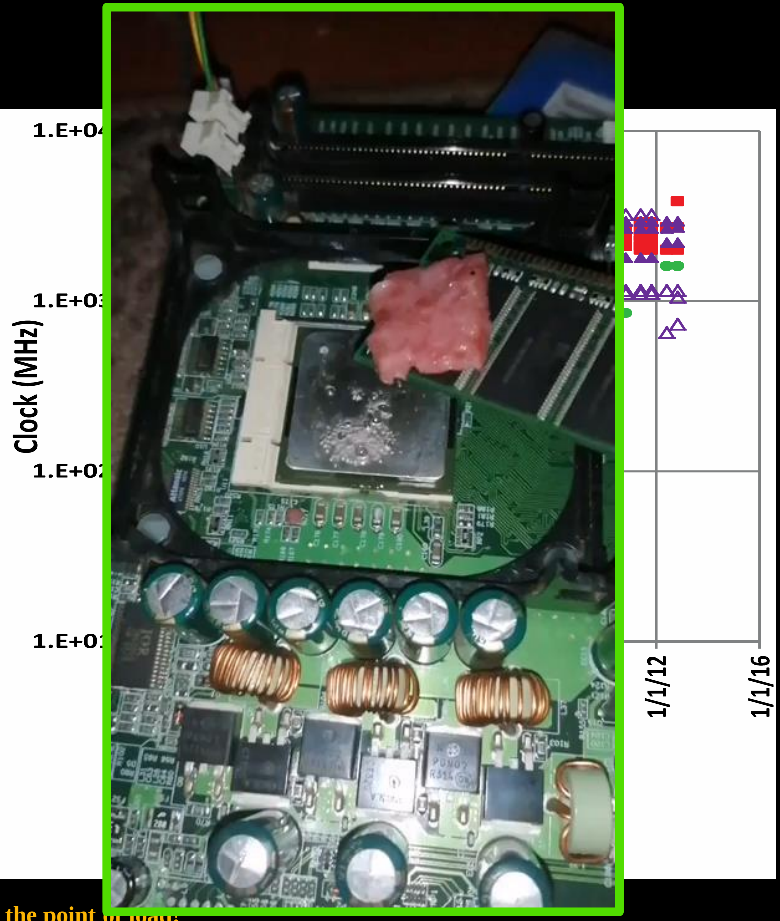
But, at end of day we keep using getting more transistors.



Data source: Wikipedia (https://en.wikipedia.org/wiki/Transistor_count)
The data visualization is available at OurWorldinData.org. There you find more visualizations and research on this topic.

That Power and Clock Inflection Point in 2004… didn't get better.

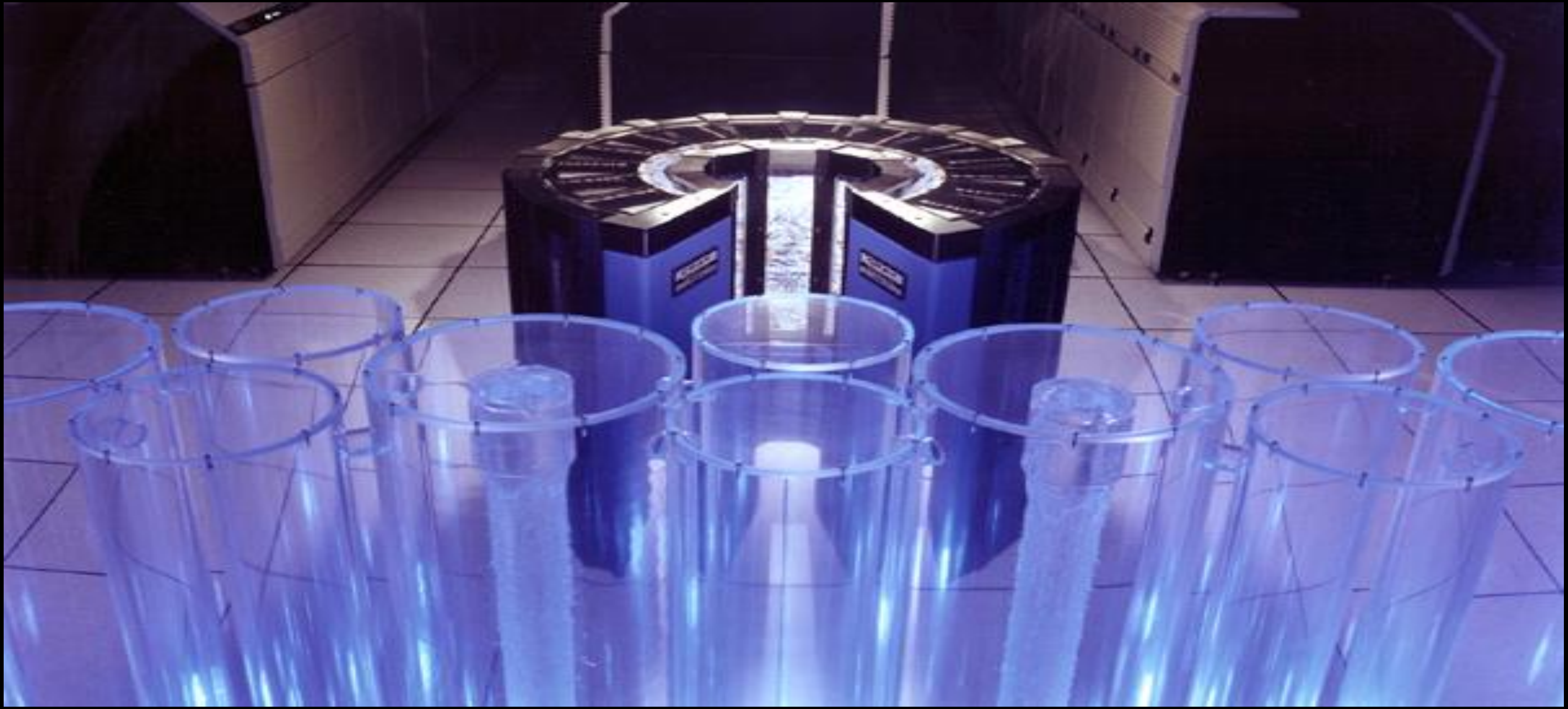




Fun fact: At 100+ Watts and <1V, currents are beginning to exceed 100A at the point of load.

Courtesy Horst Simon, LBNL

Not a new problem, just a new scale…



Cray-2 with cooling tower in foreground, circa 1985

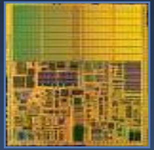
And how to get more performance from more transistors with the same power.

A 15% Reduction In Voltage Yields

RULE OF THUMB

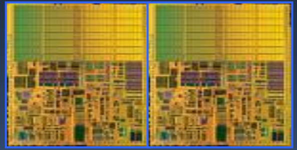
Frequency Reduction	Power Reduction	Performance Reduction
15%	45%	10%

SINGLE CORE



Area	=	1
Voltage	=	1
Freq	=	1
Power	=	1
Perf	=	1

DUAL CORE

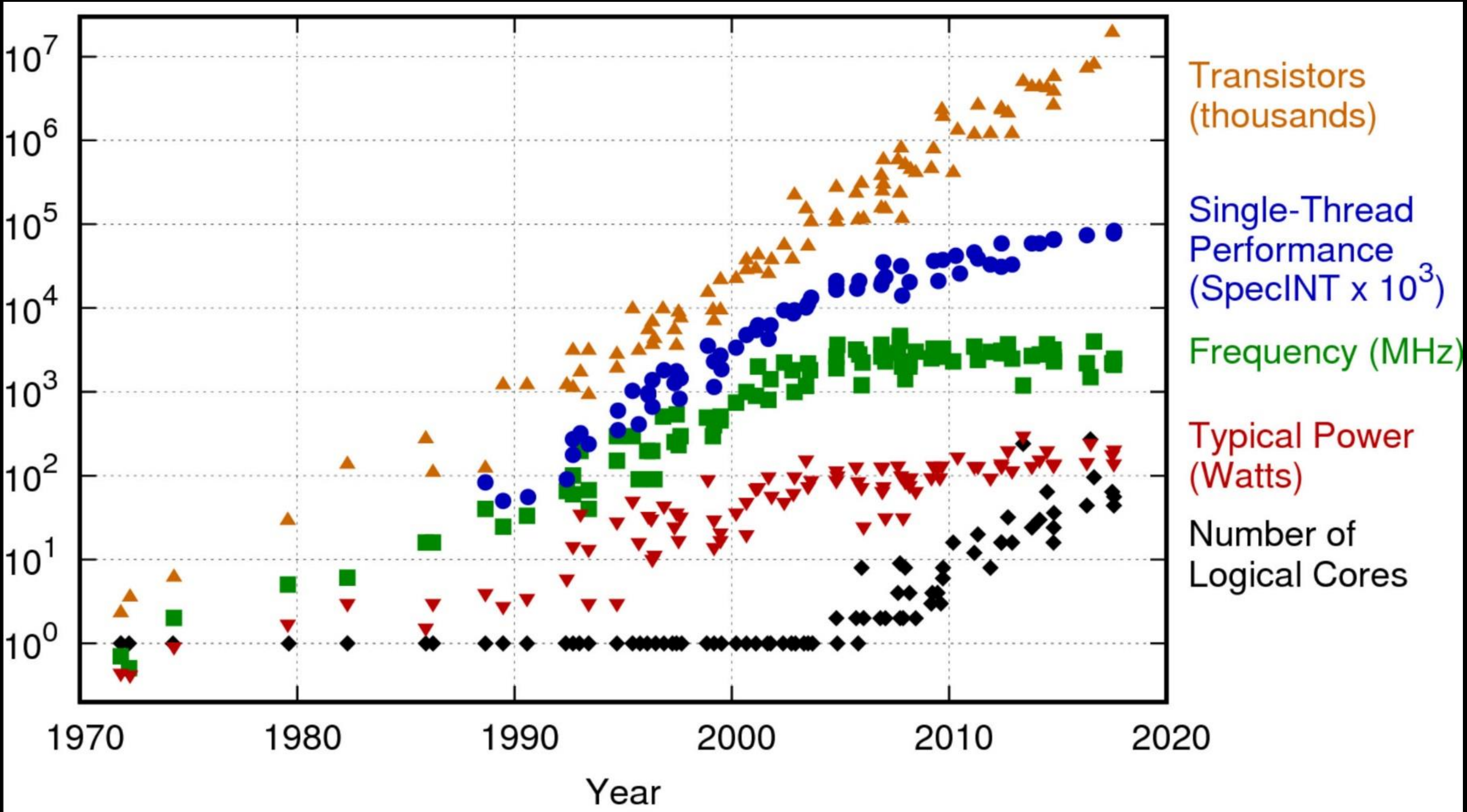


Area	=	2
Voltage	=	0.85
Freq	=	0.85
Power	=	1
Perf	=	~1.8

Single Socket Parallelism

Processor	Year	Vector	Bits	SP FLOPs / core / cycle	Cores	FLOPs/cycle
Pentium III	1999	SSE	128	3	1	3
Pentium IV	2001	SSE2	128	4	1	4
Core	2006	SSE3	128	8	2	16
Nehalem	2008	SSE4	128	8	10	80
Sandybridge	2011	AVX	256	16	12	192
Haswell	2013	AVX2	256	32	18	576
KNC	2012	AVX512	512	32	64	2048
KNL	2016	AVX512	512	64	72	4608
Skylake	2017	AVX512	512	96	28	2688

Putting It All Together



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp

Parallel Computing

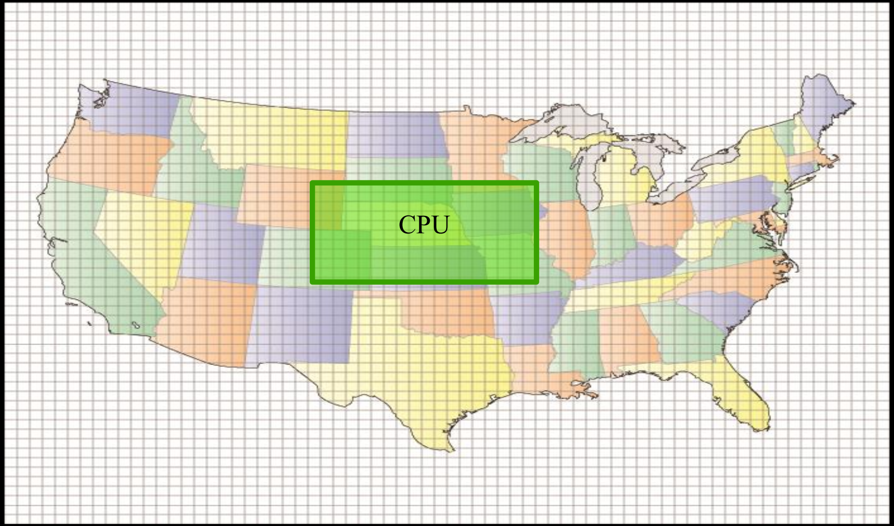
If one woman can make a baby in 9 months...

Can 9 women make a baby in 1 month?

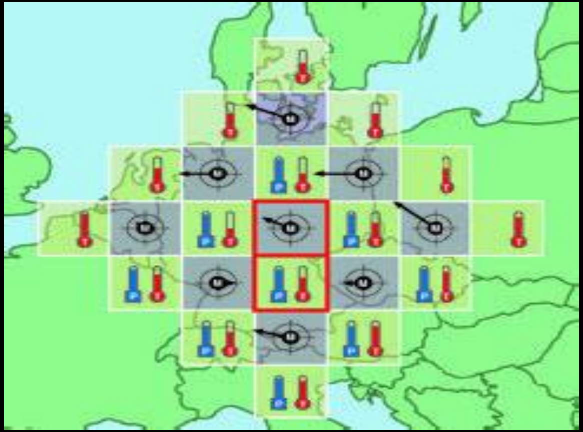
But 9 women can make 9 babies in 9 months.

First two bullets are Brook's Law. From *The Mythical Man-Month.*

Prototypical Application: Serial Weather Model

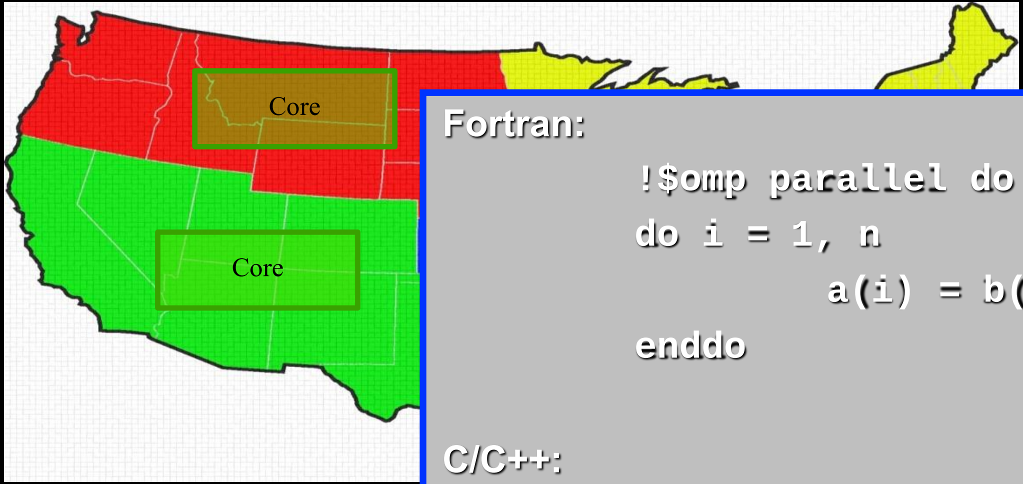


First Parallel Weather Modeling Algorithm: Richardson in 1917



Courtesy John Burkhardt, Virginia Tech

Weather Model: Shared Memory (OpenMP)



Fortran:

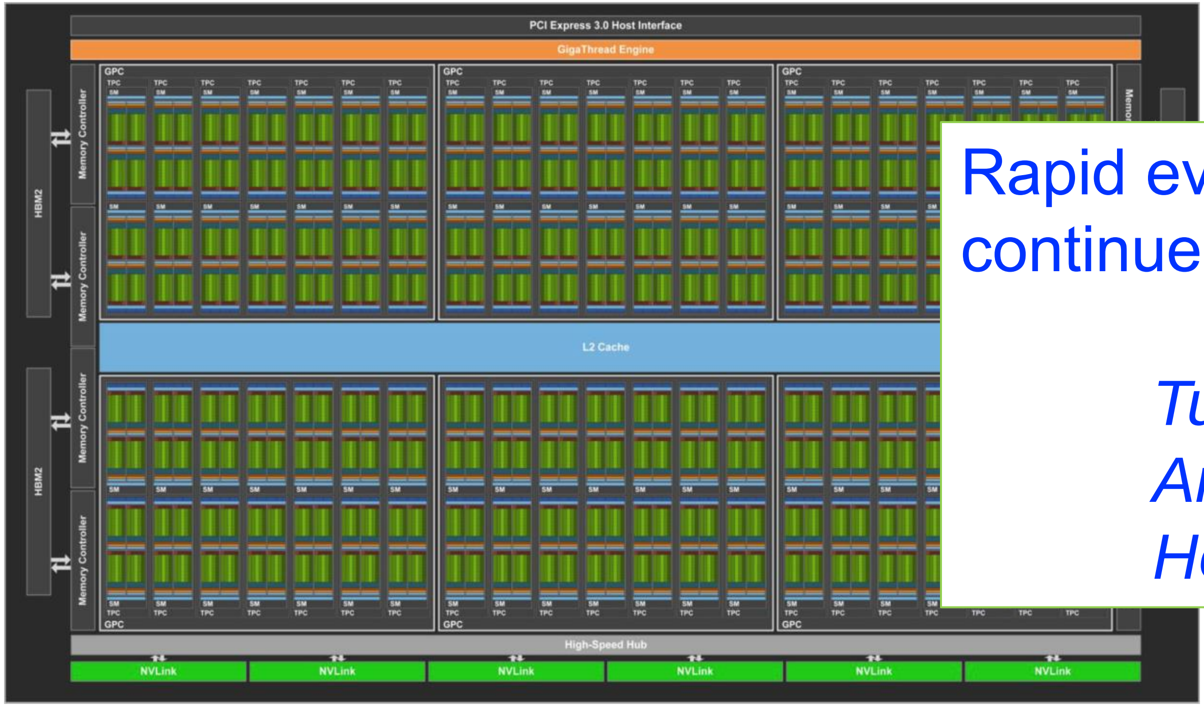
```
        !$omp parallel do
        do i = 1, n
                a(i) = b(i) + c(i)
        enddo
```

C/C++:

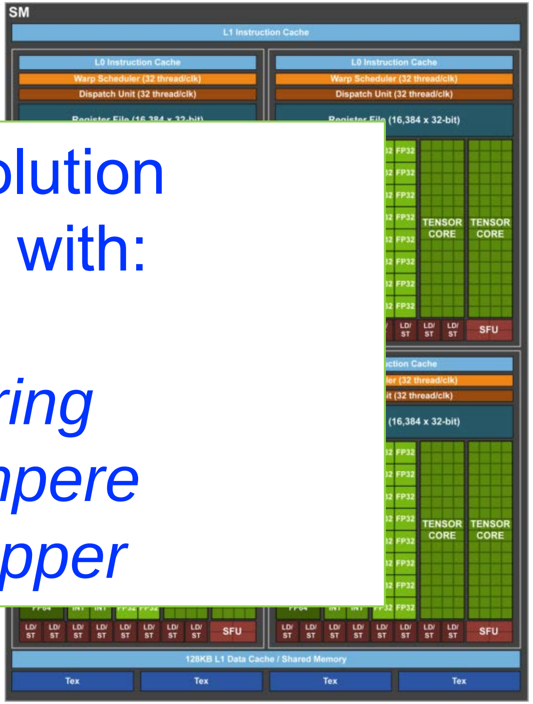
```
      #pragma omp parallel for
      for(i=1; i<=n; i++)
                a[i] = b[i] + c[i];
```

Four meteorologists in th

V100 GPU and SM



Volta GV100 GPU with 85 Streaming Multiprocessor (SM) units

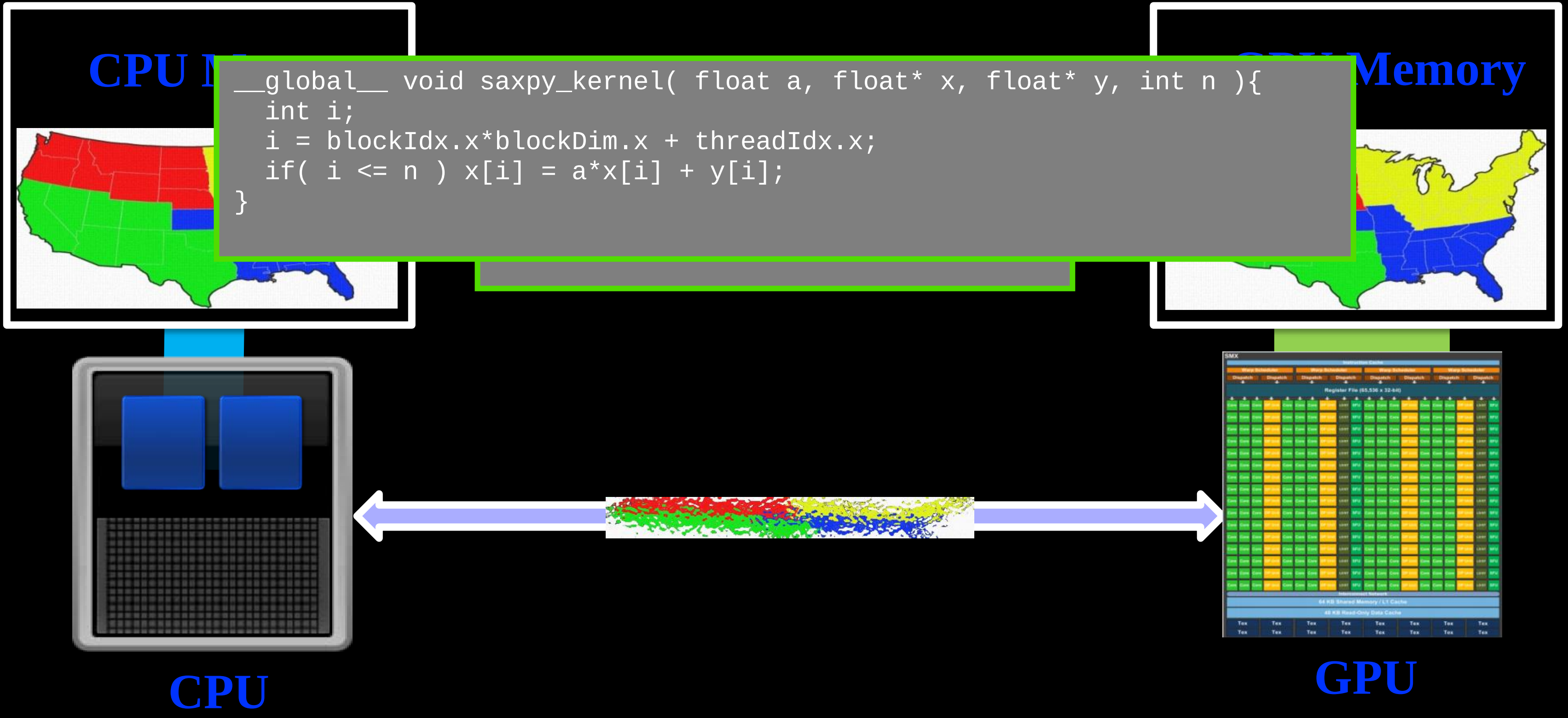

SM
L1 Instruction Cache
L0 Instruction Cache
Warp Scheduler (32 thread/clk)
Dispatch Unit (32 thread/clk)
Register File (16,384 x 32-bit)
FP64
INT
FP32
TENSOR CORE
LD/ST
SFU
128KB L1 Data Cache / Shared Memory
Tex

Volta GV100 SM

Rapid evolution continues with:

Turing
Ampere
Hopper

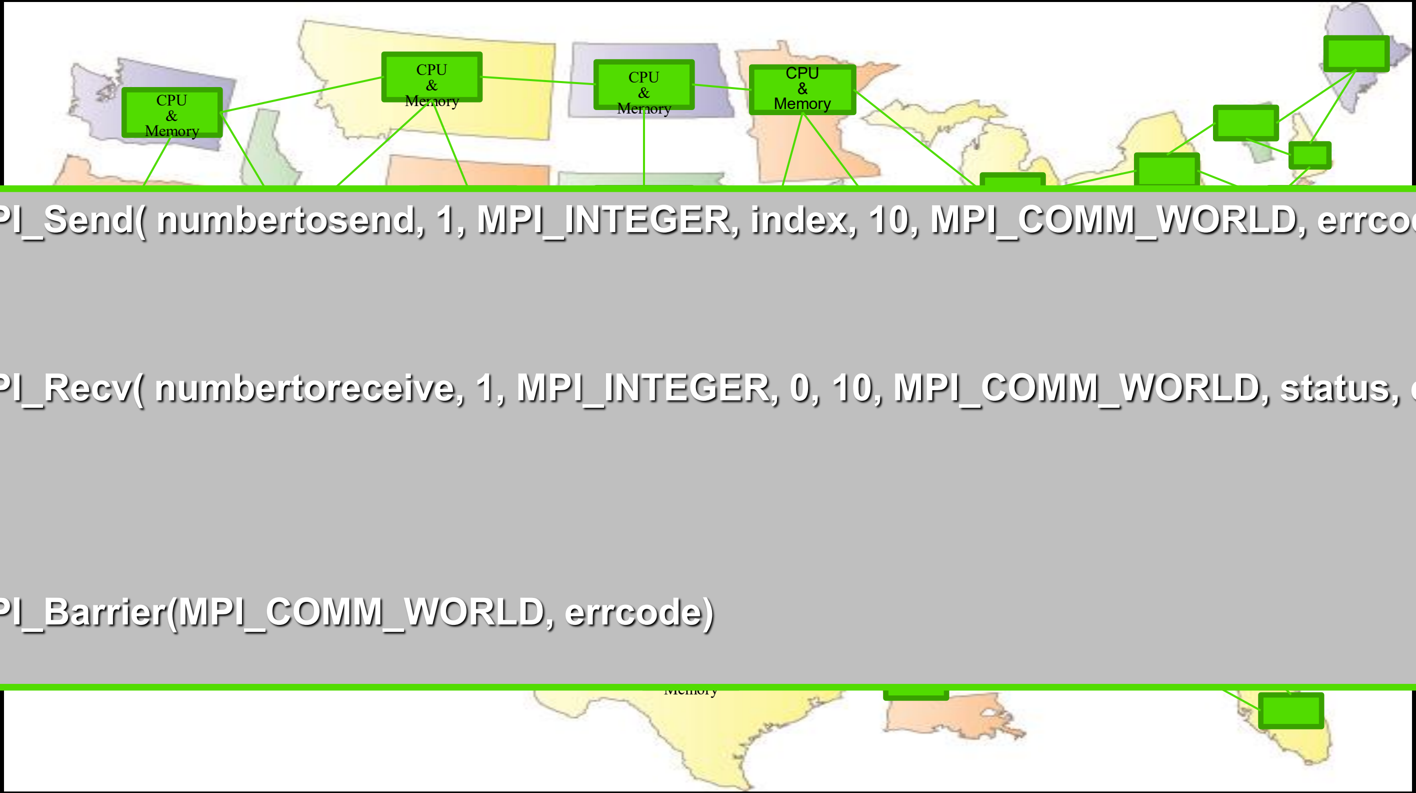
Weather Model: Accelerator (OpenACC)


CPU M
GPU Memory
CPU
GPU
SMX
Register File (65,536 x 32-bit)
64 KB Shared Memory / L1 Cache
48 KB Read-Only Data Cache

```
__global__ void saxpy_kernel( float a, float* x, float* y, int n ){
  int i;
  i = blockIdx.x*blockDim.x + threadIdx.x;
  if( i <= n ) x[i] = a*x[i] + y[i];
}
```

1 meteorologists coordinating 1000 math savants using tin cans and a string.

Weather Model: Distributed Memory (MPI)



```
call MPI_Send( numbertosend, 1, MPI_INTEGER, index, 10, MPI_COMM_WORLD, errcode)
.
.
call MPI_Recv( numbertoreceive, 1, MPI_INTEGER, 0, 10, MPI_COMM_WORLD, status, errcode)
.
.
.
call MPI_Barrier(MPI_COMM_WORLD, errcode)
.
```

50 meteorologists using a telegraph.

MPPs (Massively Parallel Processors)

Distributed memory at largest scale. Shared memory at lower level.

Summit (ORNL)

- 122 PFlops Rmax and 187 PFlops Rpeak
- IBM Power 9, 22 core, 3GHz CPUs
- 2,282,544 cores
- NVIDIA Volta GPUs
- EDR Infiniband

Sunway TaihuLight (NSC, China)

- 93 PFlops Rmax and 125 PFlops Rpeak
- Sunway SW26010 260 core, 1.45GHz CPU
- 10,649,600 cores
- Sunway interconnect

Many Levels and Types of Parallelism

- Vector (SIMD)
- Instruction Level (ILP)
  - Instruction pipelining
  - Superscaler (multiple instruction units)
  - Out-of-order
  - Register renaming
  - Speculative execution
  - Branch prediction

Compiler
(not your problem)

OpenMP 4/5
can help!

OpenMP

- Multi-Core (Threads)
- SMP/Multi-socket

OpenACC

- Accelerators: GPU & MIC

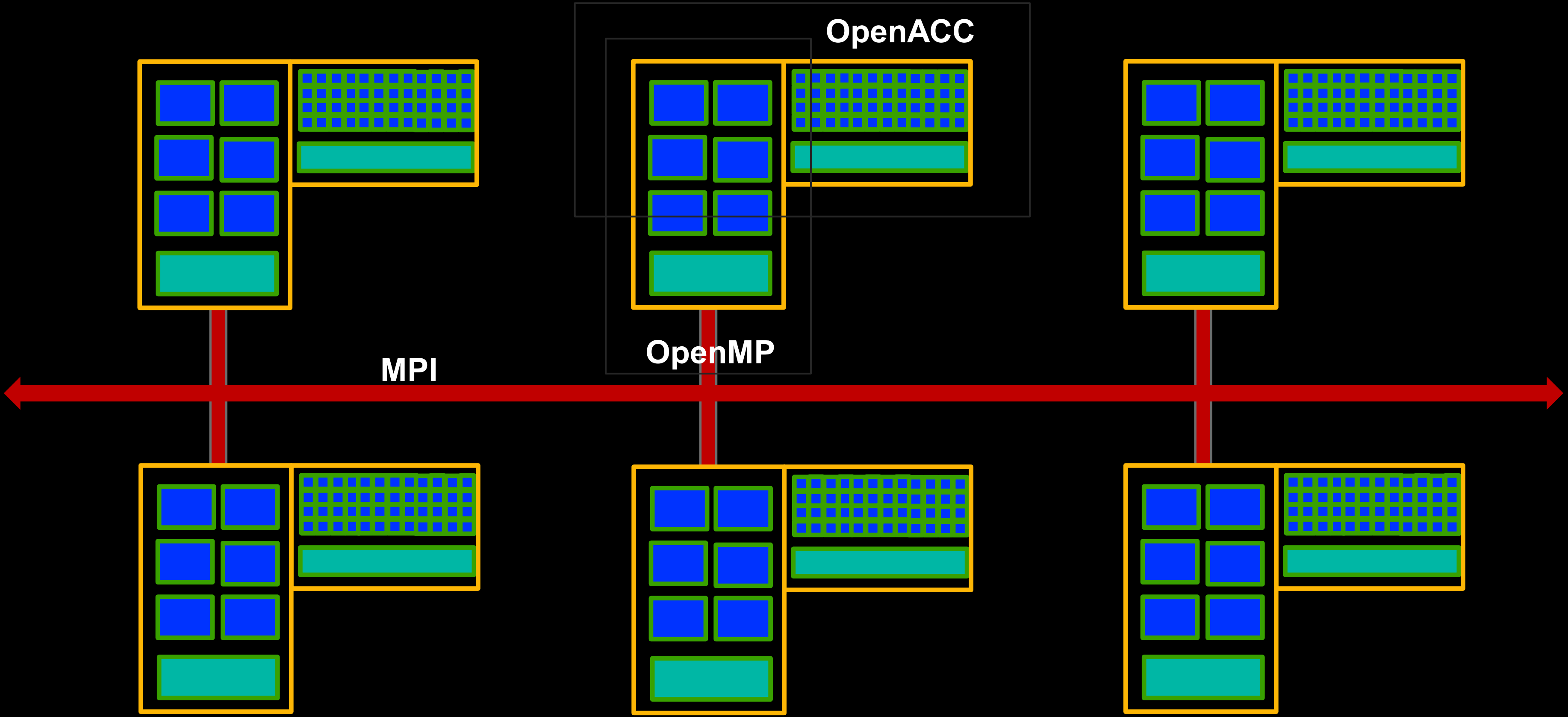
MPI

- Clusters
- MPPs

Also Important

- ASIC/FPGA/DSP
- RAID/IO

The pieces fit like this…



Cores, Nodes, Processors, PEs?

- A "core" can run an independent thread of code. Hence the temptation to refer to it as a processor.

- “Processors” refer to a physical chip. Today these almost always have more than one core.

- “Nodes” is used to refer to an actual physical unit with a network connection; usually a circuit board or "blade" in a cabinet. These often have multiple processors.

- To avoid ambiguity, it is precise to refer to the smallest useful computing device as a Processing Element, or PE. On normal processors this corresponds to a core.

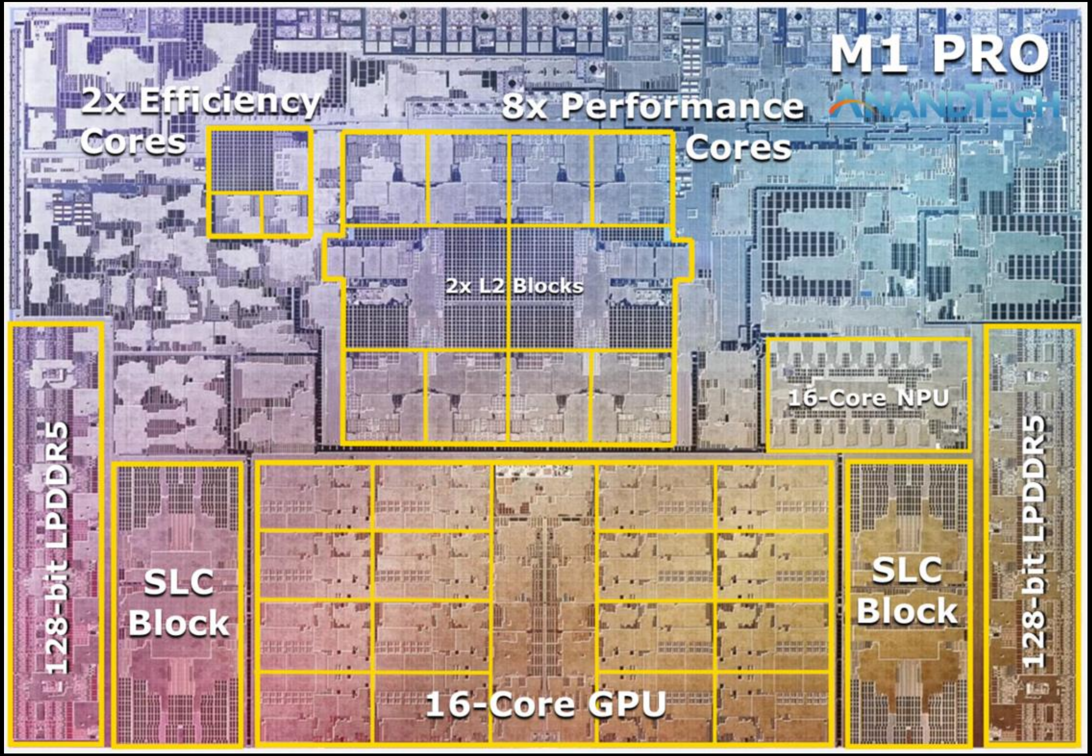
I will try to use the term PE consistently myself, but I may slip up. Get used to it as you will quite often hear all of the above terms used interchangeably where they shouldn’t be. Context usually makes it clear.

Top 10 Systems as of November 2021

#	Site	Manufacturer	Computer	CPU Interconnect [Accelerator]	Cores	Rmax (Tflops)	Rpeak (Tflops)	Power (MW)
1	RIKEN Center for Computational Science **Japan**	Fujitsu	Fugaku	ARM 8.2A+ 48C 2.2GHz Torus Fusion Interconnect	**7,299,072**	**442,010**	**537,212**	**29.8**
2	DOE/SC/ORNL **United States**	IBM	Summit	Power9 22C 3.0 GHz Dual-rail Infiniband EDR NVIDIA V100	**2,414,592**	**148,600**	**200,794**	**10.1**
3	DOE/NNSA/LLNL **United States**	IBM	Sierra	Power9 3.1 GHz 22C Infiniband EDR NVIDIA V100	**1,572,480**	**94,640**	**125,712**	**7.4**
4	National Super Computer Center in Wuxi **China**	NRCPC	Sunway TaihuLight	Sunway SW26010 260C 1.45GHz	**10,649,600**	**93,014**	**125,435**	**15.3**
5	DOE/LBNL/NERSC **United States**	HPE	Perlmutter	EPYC 64C 2.45 GHz Slingshot NVIDIA A100	**761,304**	**70,870**	**93,750**	**2.6**
6	NVIDIA Corp. **United States**	NVIDIA	Selene	EPYC 64C 2.25 GHz Infiniband HDR NVIDIA A100	**555,520**	**63,460**	**79,215**	**2.6**
7	National Super Computer Center in Guangzhou **China**	NUDT	Tianhe-2	Intel Xeon E5-2692 2.2 GHz TH Express-2 Intel Xeon Phi 31S1P	**4,981,760**	**61,444**	**100,678**	**18.4**
8	Forschungszentrum Juelich **Germany**	Bull	Juwels	EPYC 24C 2.8GHz Infiniband HDR NVIDIA A100	**449,280**	**41,120**	**70,980**	**1.8**
9	Eni S.p.A **Italy**	Dell	HPc5	Xeon 24C 2.1 GHz Infiniband HDR NVIDIA V100	**669,760**	**35,450**	**51,720**	**2.2**
10	Microsoft Azure East **United States**	MS Azure	Voyager	EPYC 48C 2.45 GHz InfiniBand HDR NVIDIA A100	**253,440**	**30,050**	**39,531**	

The word is *Heterogeneous*

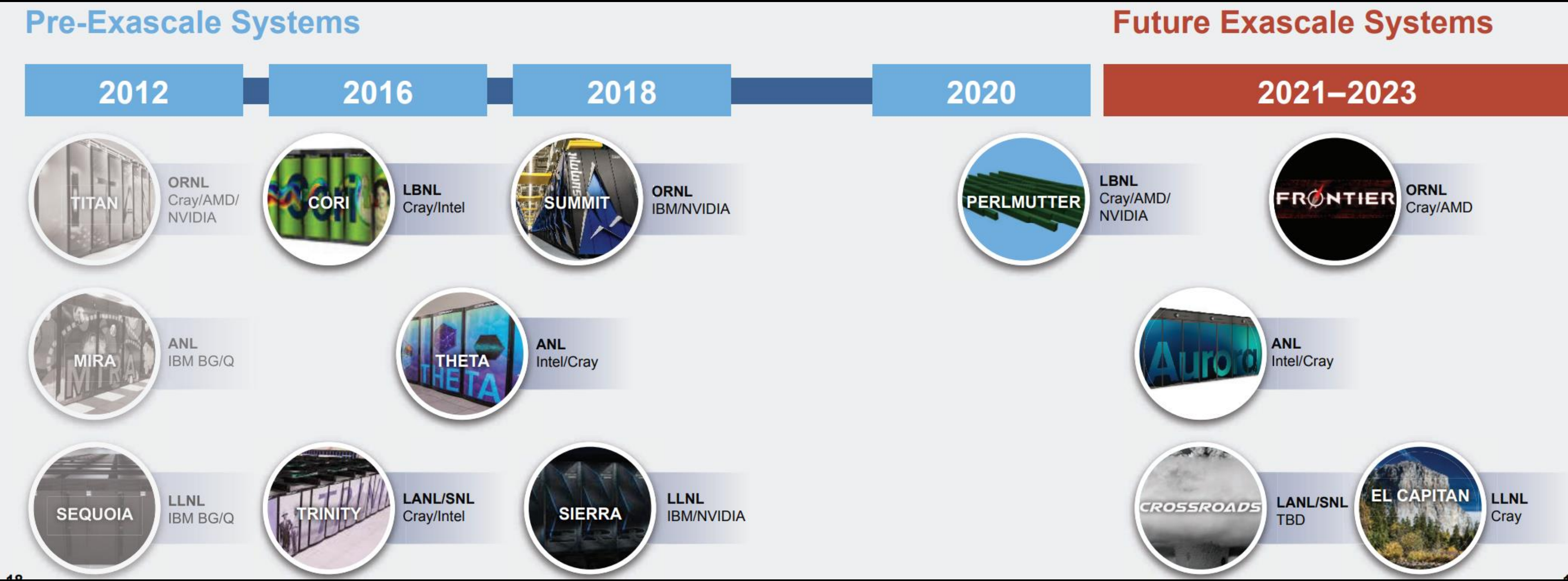
And it's not just supercomputers. It's on your desk, and in your phone.



AnandTech

How much of this can you program?

The Plan



Pre-Exascale Systems

Future Exascale Systems

2012

TITAN
ORNL
Cray/AMD/NVIDIA

MIRA
ANL
IBM BG/Q

SEQUOIA
LLNL
IBM BG/Q

2016

CORI
LBNL
Cray/Intel

THETA
ANL
Intel/Cray

TRINITY
LANL/SNL
Cray/Intel

2018

SUMMIT
ORNL
IBM/NVIDIA

SIERRA
LLNL
IBM/NVIDIA

2020

PERLMUTTER
LBNL
Cray/AMD/NVIDIA

2021–2023

FRONTIER
ORNL
Cray/AMD

Aurora
ANL
Intel/Cray

CROSSROADS
LANL/SNL
TBD

EL CAPITAN
LLNL
Cray

USA: ECP by the Numbers

7
YEARS

$1.7B

A seven-year, $1.7 B R&D effort that launched in 2016

6
CORE DOE
LABS

Six core DOE National Laboratories: Argonne, Lawrence Berkeley, Lawrence Livermore, Oak Ridge, Sandia, Los Alamos

- Staff from most of the 17 DOE national laboratories take part in the project

3
FOCUS
AREAS

Three technical focus areas: Hardware and Integration, Software Technology, Application Development supported by a Project Management Office

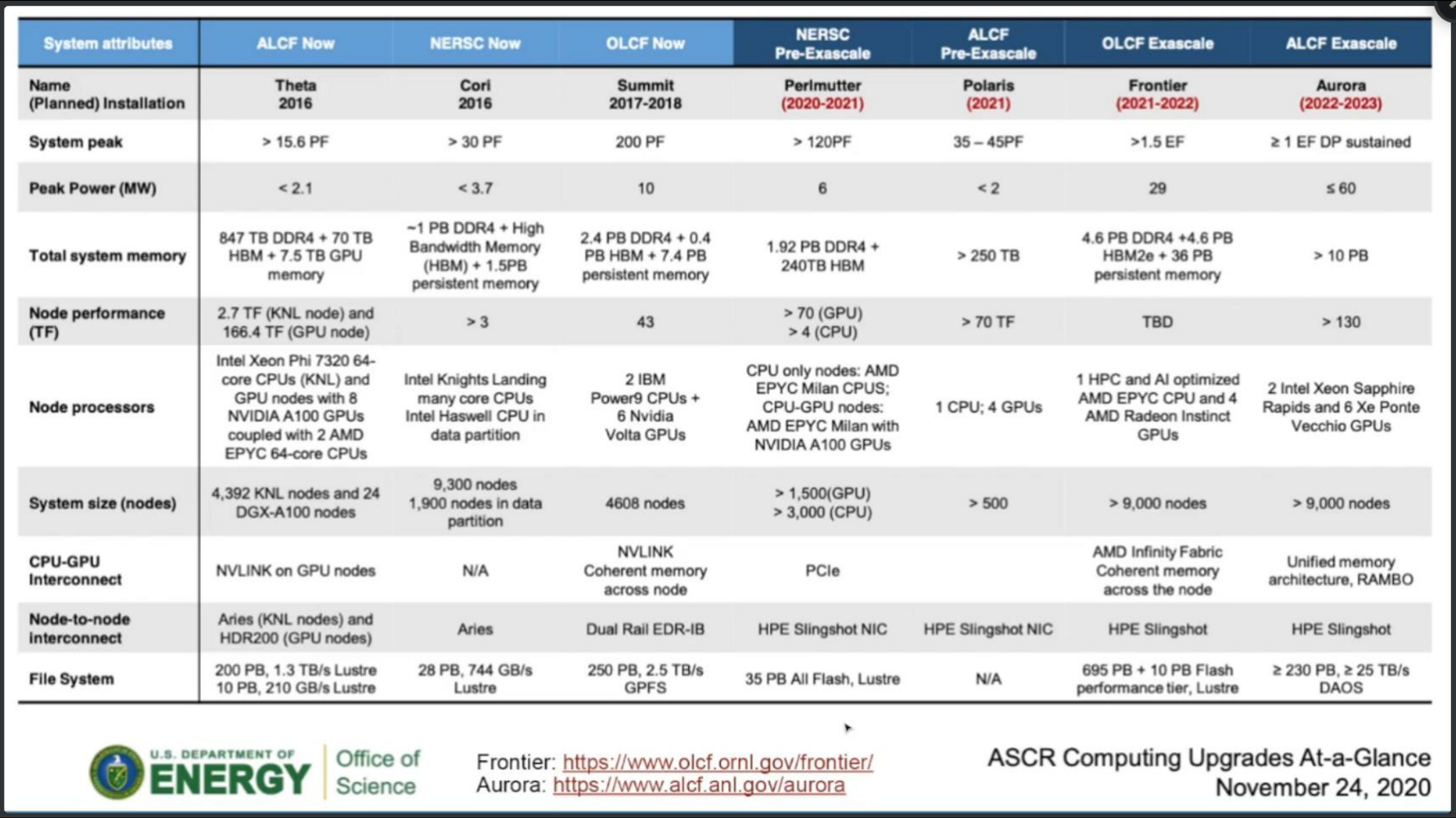
100
R&D TEAMS

1000
RESEARCHERS

More than 100 top-notch R&D teams

Hundreds of consequential milestones delivered on schedule and within budget since project inception

System Designs



System attributes	ALCF Now	NERSC Now	OLCF Now	NERSC Pre-Exascale	ALCF Pre-Exascale	OLCF Exascale	ALCF Exascale
Name (Planned) Installation	Theta 2016	Cori 2016	Summit 2017-2018	Perlmutter (2020-2021)	Polaris (2021)	Frontier (2021-2022)	Aurora (2022-2023)
System peak	> 15.6 PF	> 30 PF	200 PF	> 120PF	35 – 45PF	>1.5 EF	≥ 1 EF DP sustained
Peak Power (MW)	< 2.1	< 3.7	10	6	< 2	29	≤ 60
Total system memory	847 TB DDR4 + 70 TB HBM + 7.5 TB GPU memory	~1 PB DDR4 + High Bandwidth Memory (HBM) + 1.5PB persistent memory	2.4 PB DDR4 + 0.4 PB HBM + 7.4 PB persistent memory	1.92 PB DDR4 + 240TB HBM	> 250 TB	4.6 PB DDR4 +4.6 PB HBM2e + 36 PB persistent memory	> 10 PB
Node performance (TF)	2.7 TF (KNL node) and 166.4 TF (GPU node)	> 3	43	> 70 (GPU) > 4 (CPU)	> 70 TF	TBD	> 130
Node processors	Intel Xeon Phi 7320 64-core CPUs (KNL) and GPU nodes with 8 NVIDIA A100 GPUs coupled with 2 AMD EPYC 64-core CPUs	Intel Knights Landing many core CPUs Intel Haswell CPU in data partition	2 IBM Power9 CPUs + 6 Nvidia Volta GPUs	CPU only nodes: AMD EPYC Milan CPUS; CPU-GPU nodes: AMD EPYC Milan with NVIDIA A100 GPUs	1 CPU; 4 GPUs	1 HPC and AI optimized AMD EPYC CPU and 4 AMD Radeon Instinct GPUs	2 Intel Xeon Sapphire Rapids and 6 Xe Ponte Vecchio GPUs
System size (nodes)	4,392 KNL nodes and 24 DGX-A100 nodes	9,300 nodes 1,900 nodes in data partition	4608 nodes	> 1,500(GPU) > 3,000 (CPU)	> 500	> 9,000 nodes	> 9,000 nodes
CPU-GPU Interconnect	NVLINK on GPU nodes	N/A	NVLINK Coherent memory across node	PCIe		AMD Infinity Fabric Coherent memory across the node	Unified memory architecture, RAMBO
Node-to-node interconnect	Aries (KNL nodes) and HDR200 (GPU nodes)	Aries	Dual Rail EDR-IB	HPE Slingshot NIC	HPE Slingshot NIC	HPE Slingshot	HPE Slingshot
File System	200 PB, 1.3 TB/s Lustre 10 PB, 210 GB/s Lustre	28 PB, 744 GB/s Lustre	250 PB, 2.5 TB/s GPFS	35 PB All Flash, Lustre	N/A	695 PB + 10 PB Flash performance tier, Lustre	≥ 230 PB, ≥ 25 TB/s DAOS

U.S. DEPARTMENT OF ENERGY | Office of Science

Frontier: https://www.olcf.ornl.gov/frontier/
Aurora: https://www.alcf.anl.gov/aurora

ASCR Computing Upgrades At-a-Glance
November 24, 2020

Networks

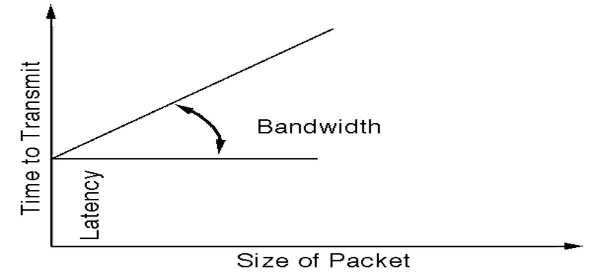
3 characteristics sum up the network:

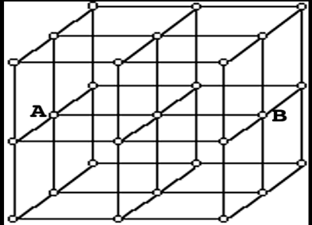
- Latency

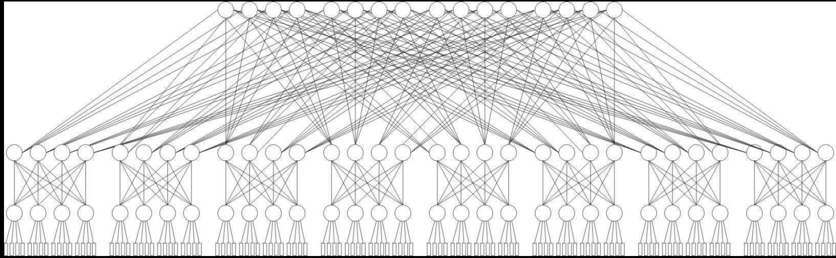
  The time to send a 0 byte packet of data on the network

- Bandwidth

  The rate at which a very large packet of information can be sent



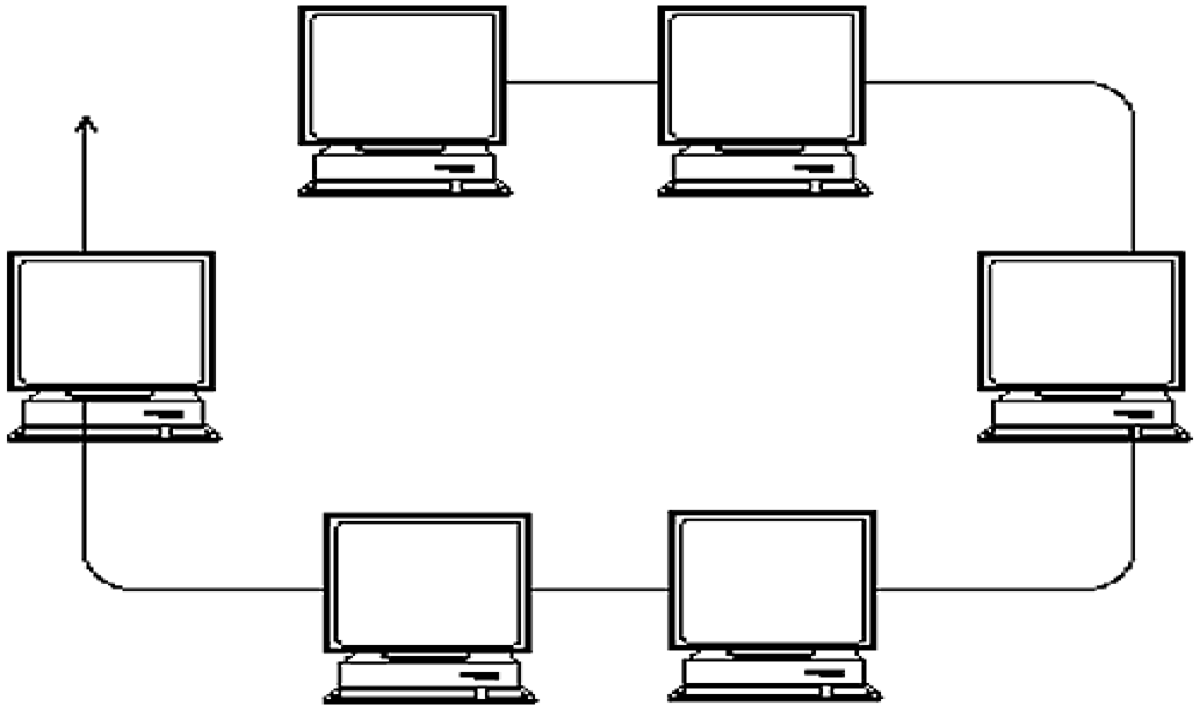




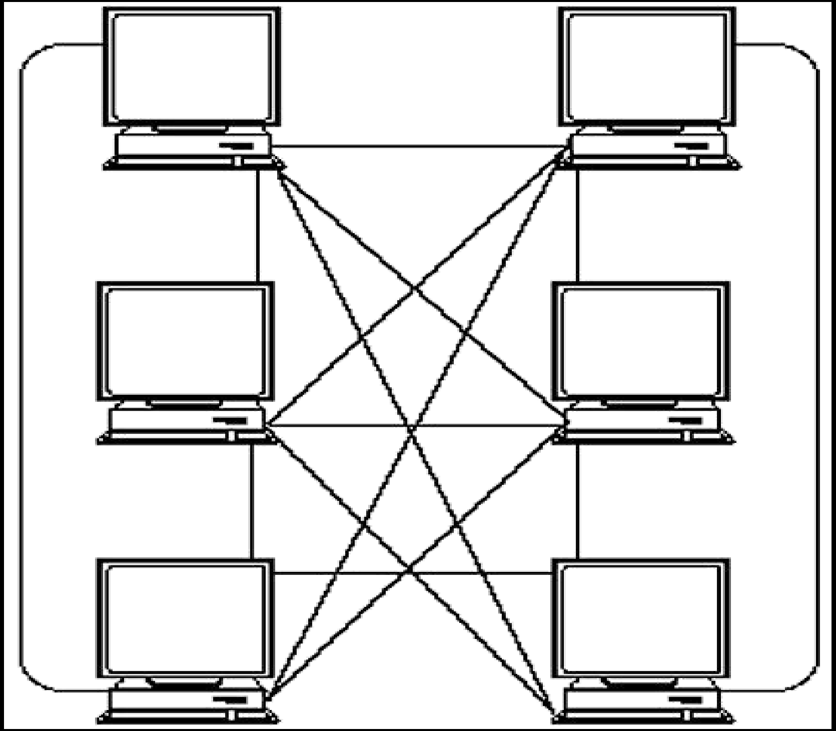
- Topology

  The configuration of the network that determines how processing units are directly connected.

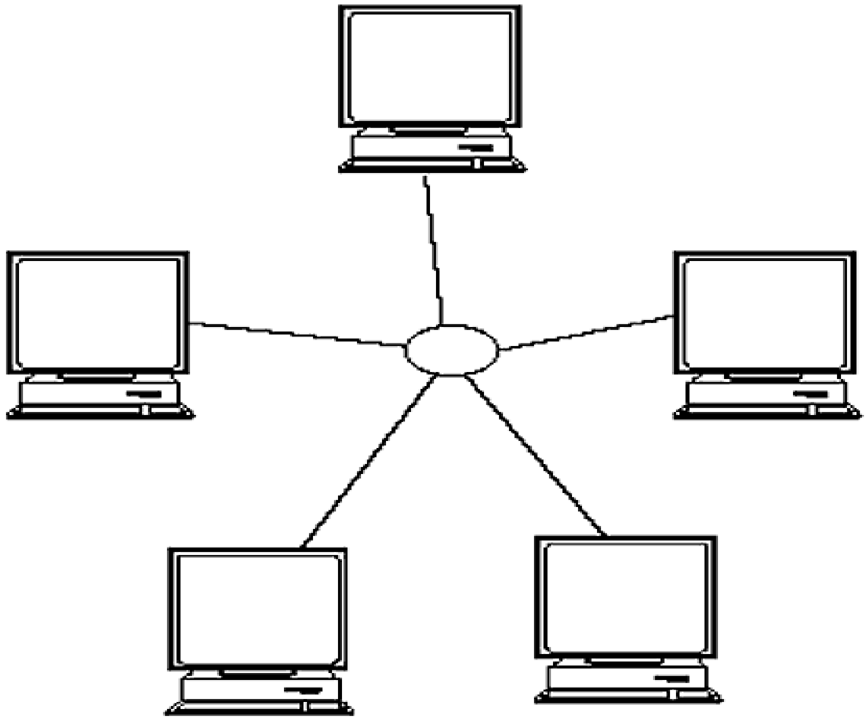
Ethernet with Workstations



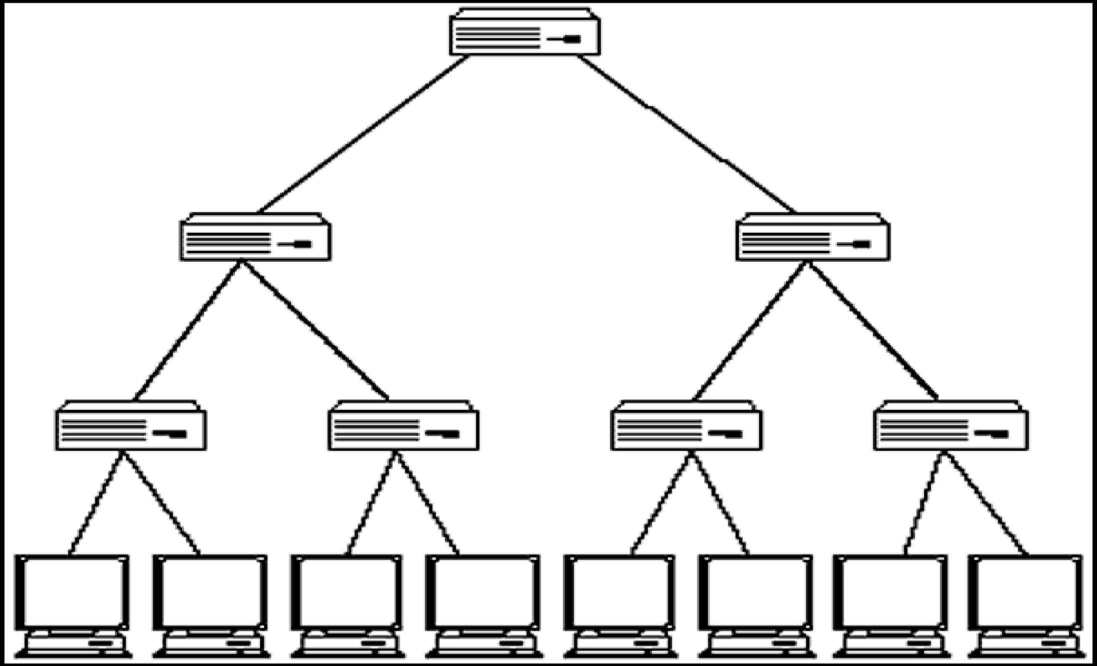
Complete Connectivity



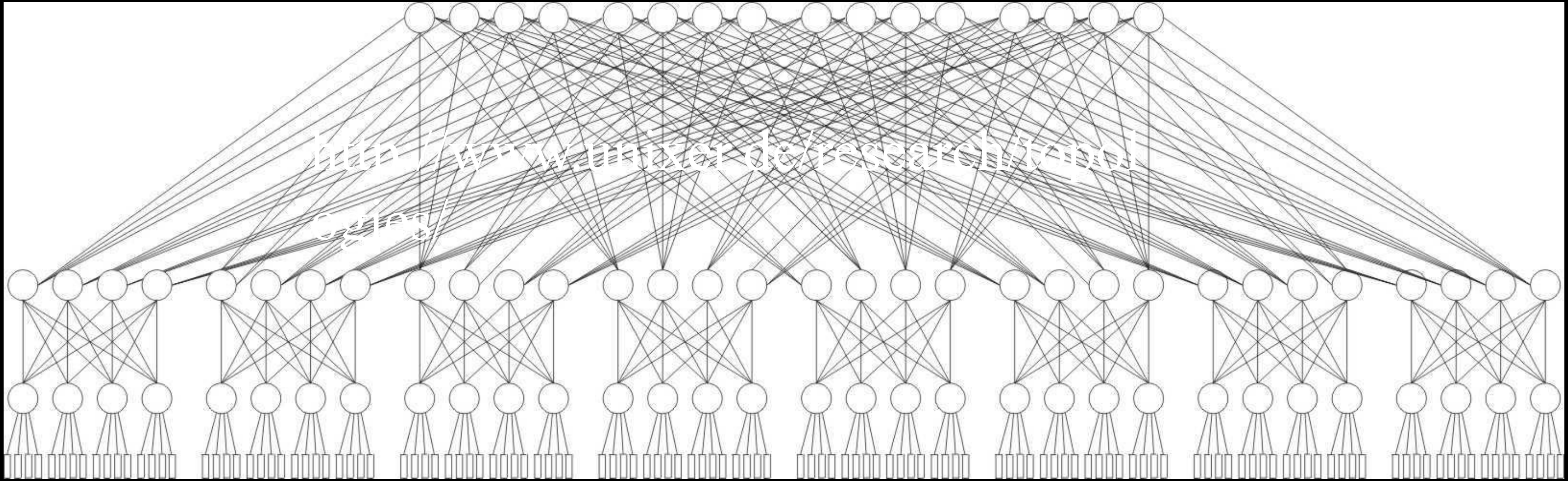
Crossbar



Binary Tree

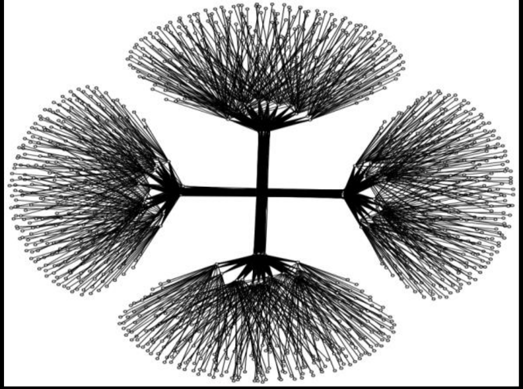


Fat Tree

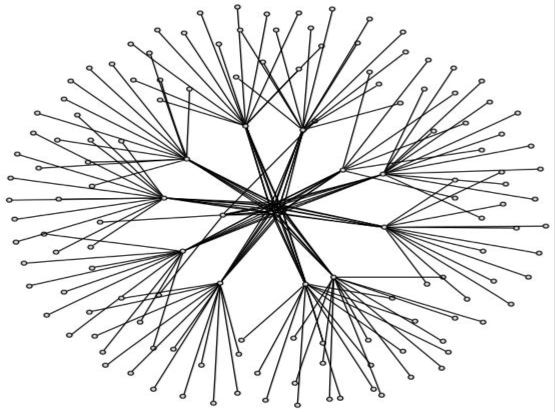


http://www.unixer.de/research/topologies/

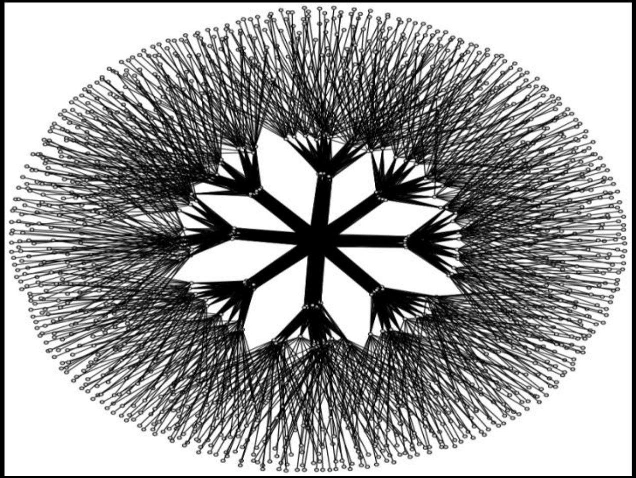
Other Fat Trees



Big Red @ IU

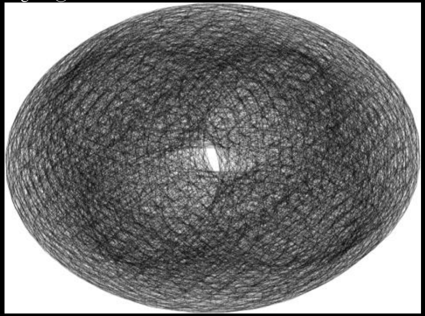


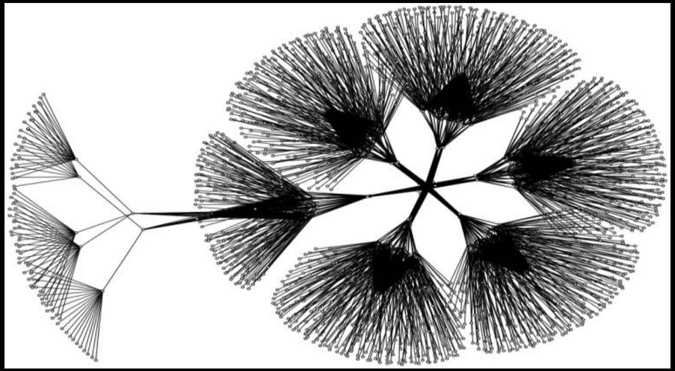
Odin @ IU



Atlas @ LLNL

Jaguar @ ORNL





Tsubame @ Tokyo Inst. of Tech

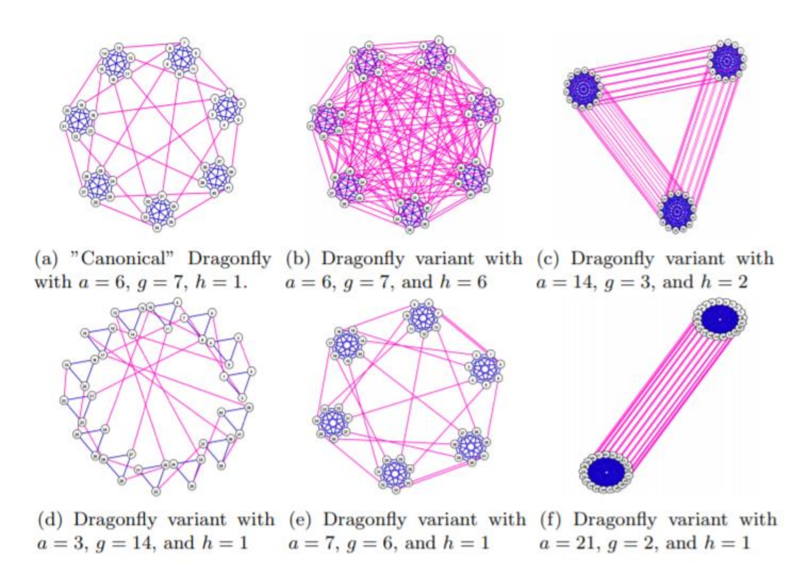
From Torsten Hoefler's Network Topology Repository at
http://www.unixer.de/research/topologies/

Dragonfly

A newer innovation in network design is the dragonfly topology, which benefits from advanced hardware capabilities like:

- High-Radix Switches
- Adaptive Routing
- Optical Links

Various 42 node Dragonfly configurations.



(a) "Canonical" Dragonfly with $a = 6$, $g = 7$, $h = 1$.

(b) Dragonfly variant with $a = 6$, $g = 7$, and $h = 6$

(c) Dragonfly variant with $a = 14$, $g = 3$, and $h = 2$

(d) Dragonfly variant with $a = 3$, $g = 14$, and $h = 1$

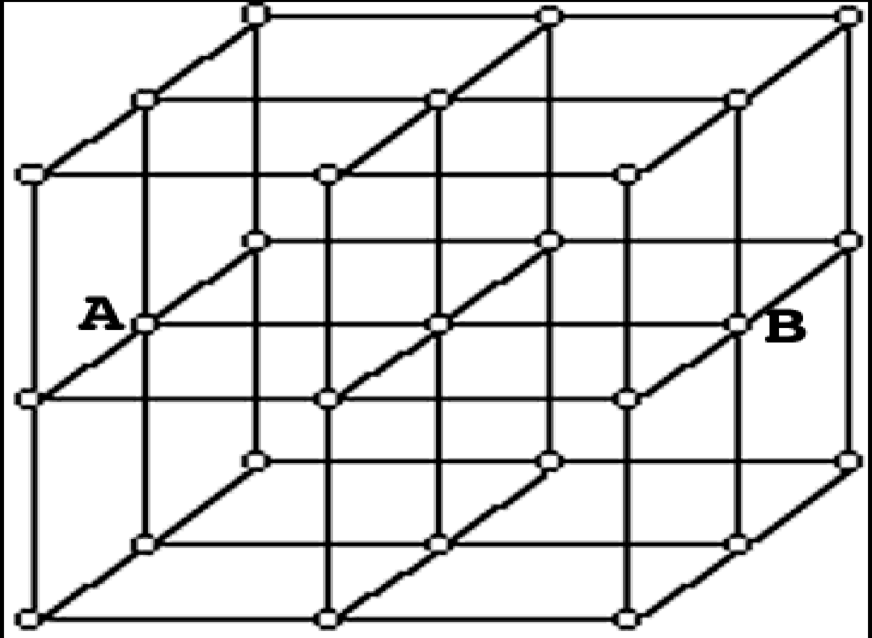
(e) Dragonfly variant with $a = 7$, $g = 6$, and $h = 1$

(f) Dragonfly variant with $a = 21$, $g = 2$, and $h = 1$

Purple links are optical, and blue are electrical.

Graphic from the excellent paper *Design space exploration of the Dragonfly topology* by Yee, Wilke, Bergman and Rumley.

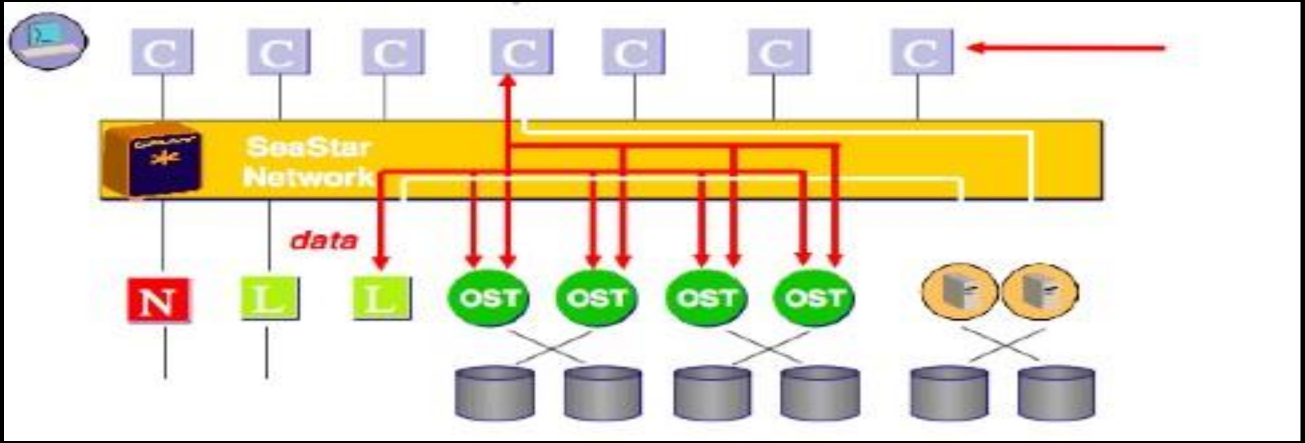
3-D Torus



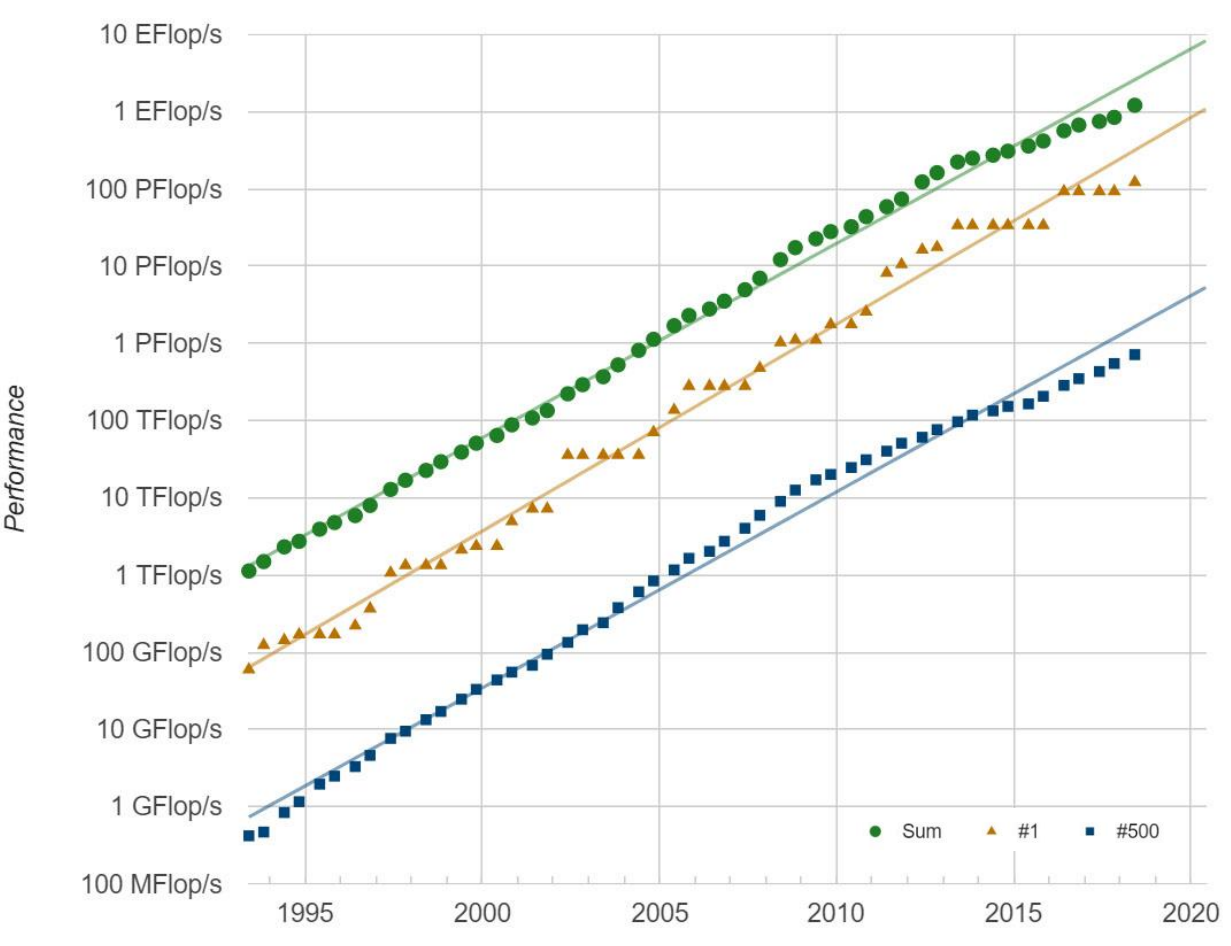
Torus simply means that “ends” are connected. This means A is really connected to B and the cube has no real boundary.

Parallel IO (RAID…)

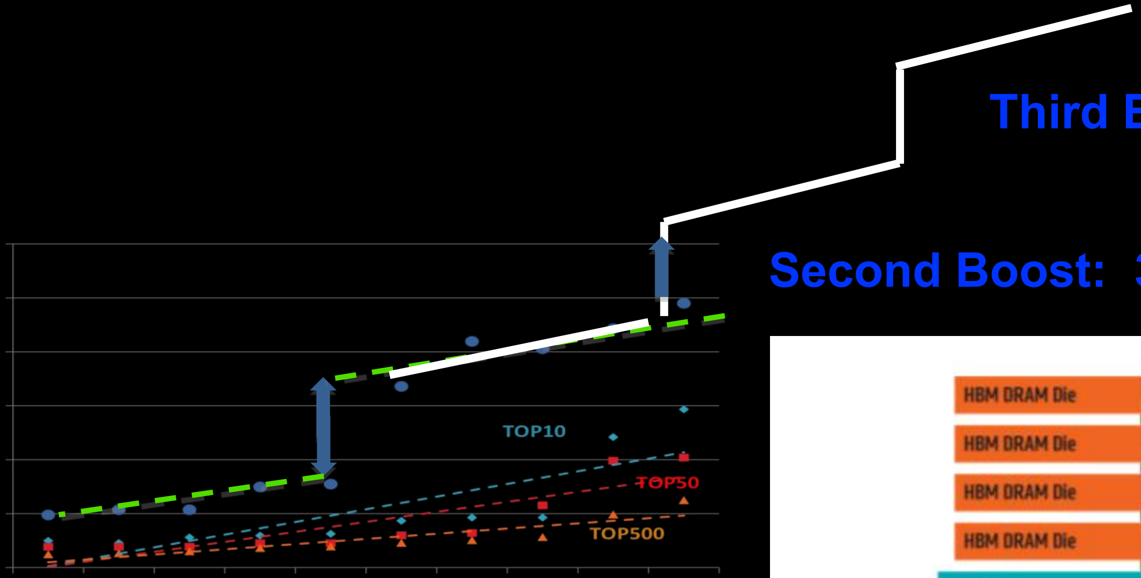
- There are increasing numbers of applications for which many PB of data need to be written.
- Checkpointing is also becoming very important due to MTBF issues (a whole ‘nother talk).
- Build a large, fast, reliable filesystem from a collection of smaller drives.
- Supposed to be transparent to the programmer.
- Increasingly mixing in SSD.



Sustaining Performance Improvements

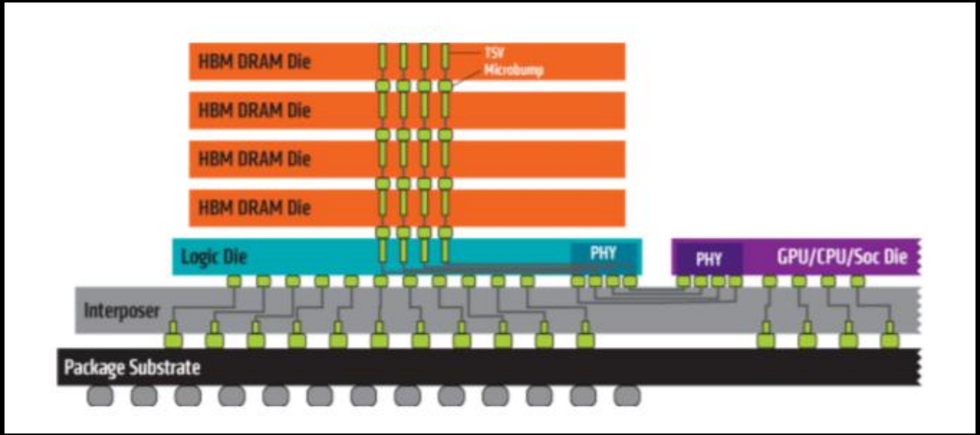


Two Additional Boosts to Improve Flops/Watt and Reach Exascale Target



Third Boost: SiPh (2020 – 2024)

Second Boost: 3D (2016 – 2020)



First boost: many-core/accelerator

Courtesy Horst Simon, LBNL

It is not just “exaflops” – we are changing the whole computational model

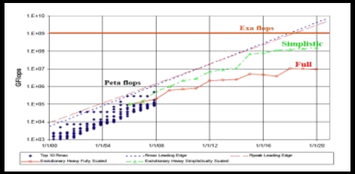
Current programming systems have WRONG optimization targets

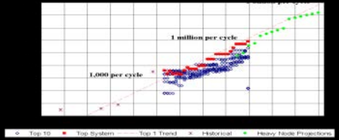
Old Constraints

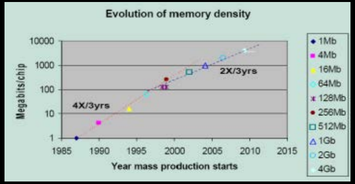
- **Peak clock frequency** ***as primary limiter for performance improvement***
- **Cost: *FLOPs* are biggest cost for system: *optimize for compute***
- **Concurrency: Modest growth of parallelism by adding nodes**
- **Memory scaling*: maintain byte per flop capacity and bandwidth***
- **Locality*: MPI+X model (uniform costs within node & between nodes)***
- **Uniformity: *Assume uniform system performance***
- **Reliability*: It’s the hardware’s problem***

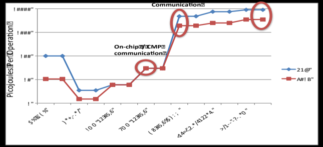
New Constraints

- **Power** *is primary design constraint for future HPC system design*
- **Cost:** *Data movement dominates: optimize to minimize data movement*
- **Concurrency:** *Exponential growth of parallelism within chips*
- **Memory Scaling:** *Compute growing 2x faster than capacity or bandwidth*
- **Locality***: must reason about data locality and possibly topology*
- **Heterogeneity***: Architectural and performance non-uniformity increase*
- **Reliability***: Cannot count on hardware protection alone*









Fundamentally breaks our current programming paradigm and computing ecosystem

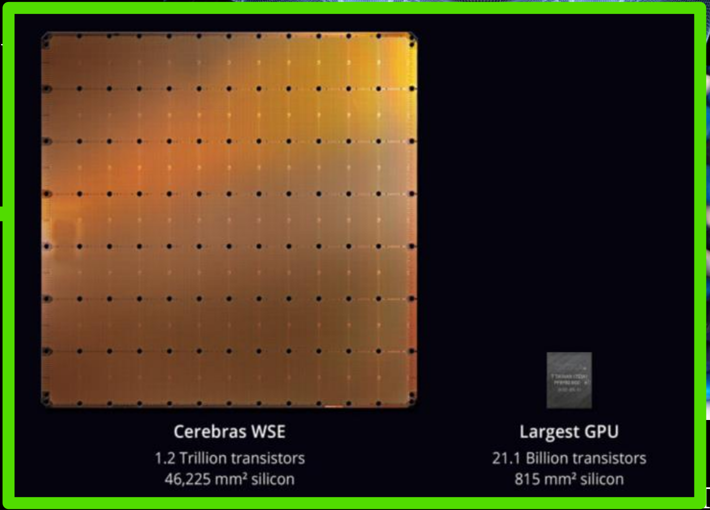
Adapted from John Shalf

End of Moore’s Law Will Lead to New Architectures

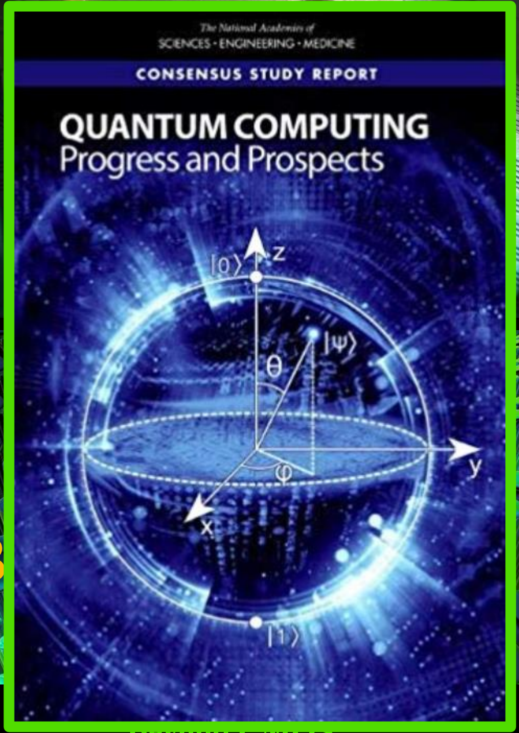
Non-von Neumann

ARCHITECT

von Neu


Cerebras WSE
1.2 Trillion transistors
46,225 mm² silicon

Largest GPU
21.1 Billion transistors
815 mm² silicon


The National Academies of
SCIENCES · ENGINEERING · MEDICINE

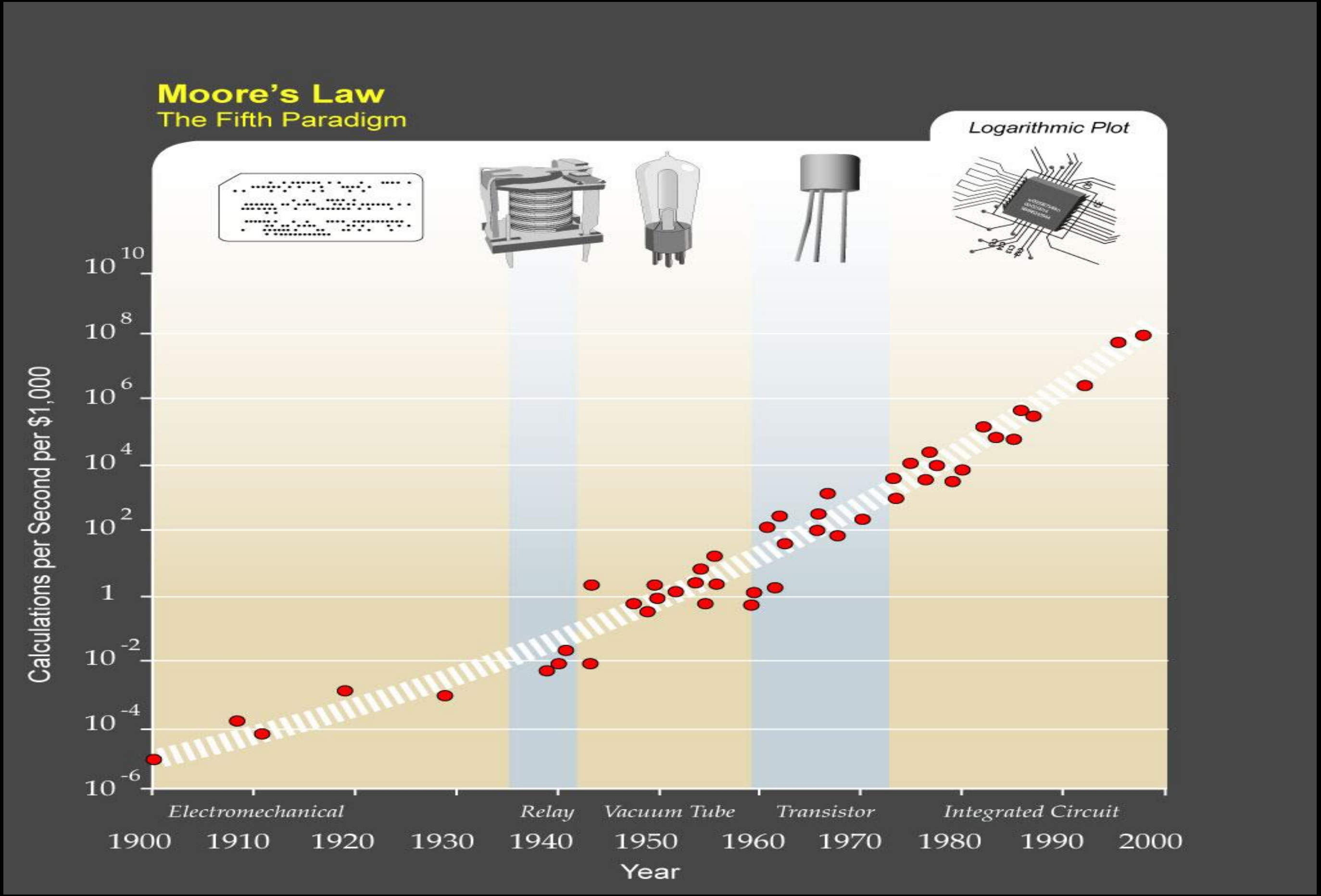
CONSENSUS STUDY REPORT

QUANTUM COMPUTING
Progress and Prospects

TECHNOLOGY

Beyond CMOS

It would only be the 6th paradigm.



We can do better. We have a role model.

- Straight forward extrapolation results in a real-time human brain scale simulation at about 1 - 10 Exaflop/s with 4 PB of memory
- Exascale computers in 2021 will have a power consumption of at 20 - 30 MW
- The human brain takes 20W
- Even under best assumptions in 2021 our brain will still be a million times more power efficient

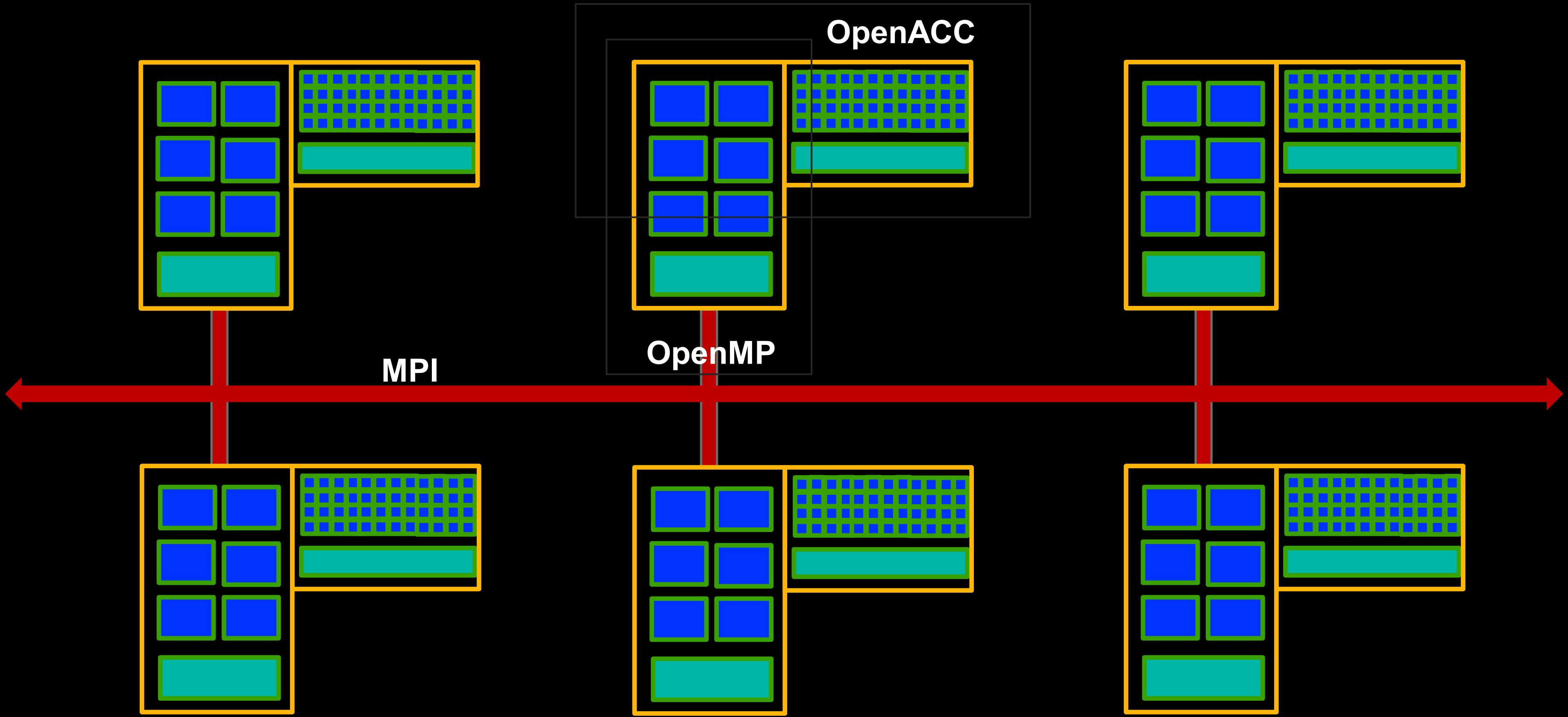
Courtesy Horst Simon, LBNL

Why you should be (extra) motivated.

- This parallel computing thing is no fad.
- The laws of physics are drawing this roadmap.
- If you get on board (the right bus), you can ride this trend for a long, exciting trip.

Let’s learn how to use these things!

In Conclusion…


OpenACC
OpenMP
MPI