

# Intro To Parallel Computing

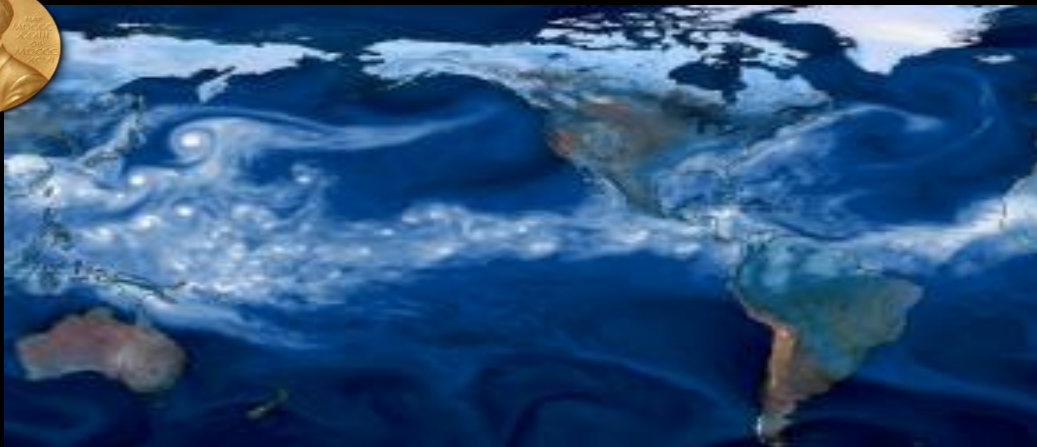
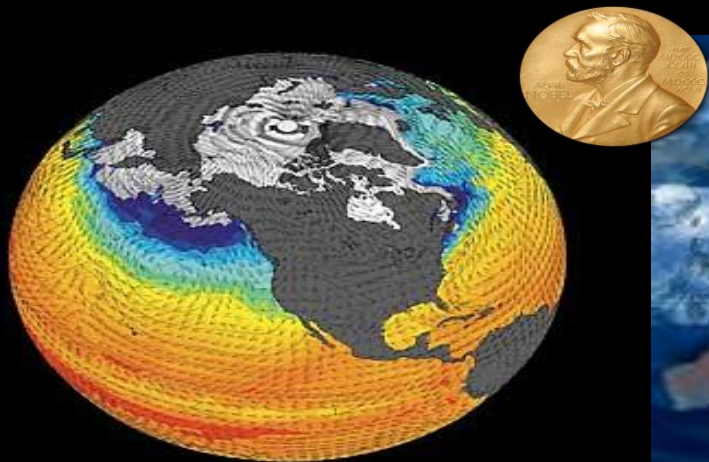
John Urbanic

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.
- 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

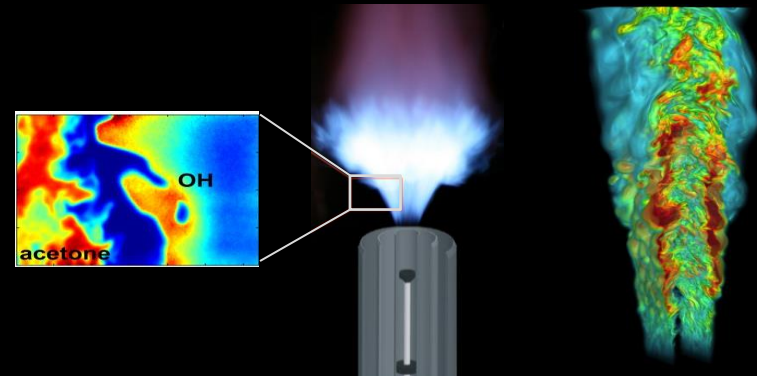
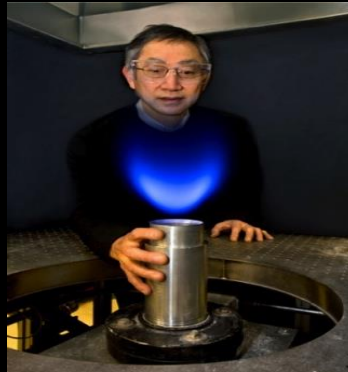
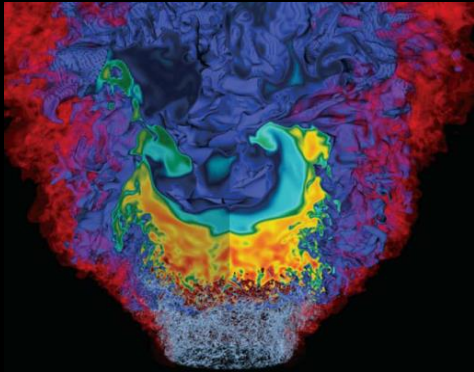
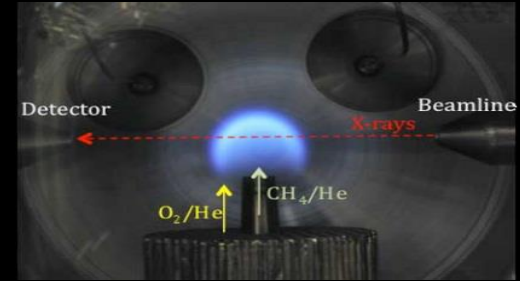
---

## 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
-

# 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









# Modha Group at IBM Almaden



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



				
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

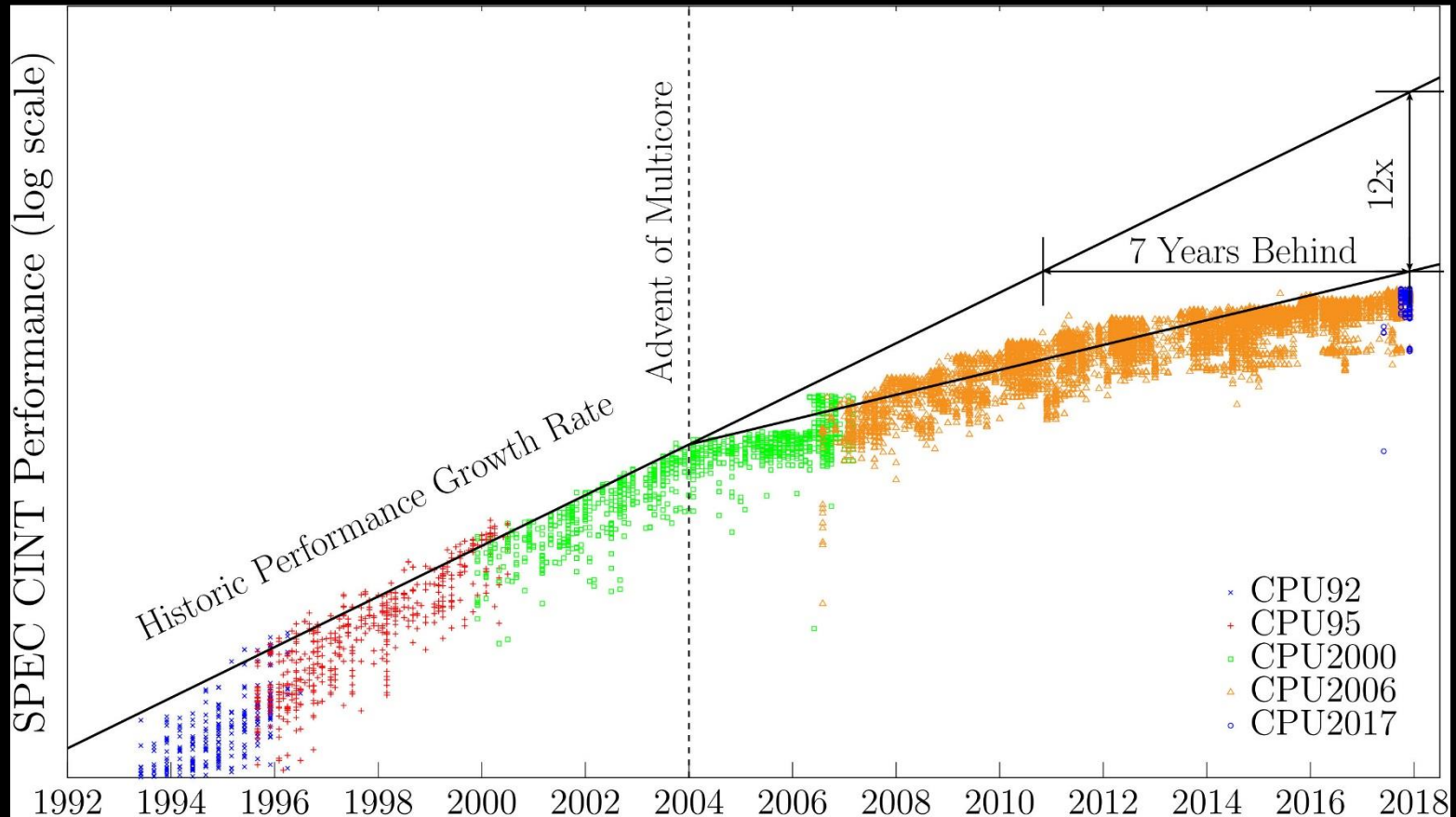
Recent simulations achieve  
unprecedented scale of  
 $65 \times 10^9$  neurons and  $16 \times 10^{12}$  synapses

# 'Nuff Said

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

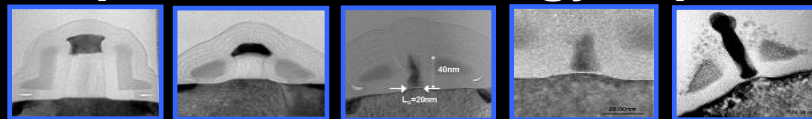
**And, many of you doubtless brought your own immediate research concerns. Great!**

# Moore's Law abandoned serial programming around 2004

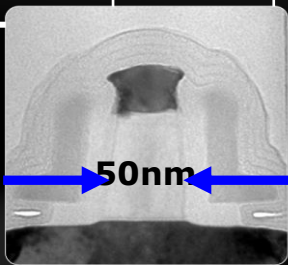


# Moore's Law is not dead yet. Maybe.

## Intel process technology capabilities

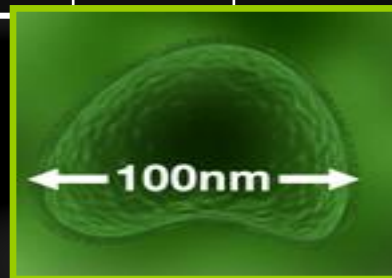


High Volume Manufacturing	2004	2006	2008	2010	2012	2014	2016	2018	2020
Feature Size	90nm	65nm	45nm	32nm	22nm	16nm	14nm	10nm	7nm
Integration Capacity (Billions of Transistors)	2	4	8	16	32	64	128	256	...



**Transistor for  
90nm Process**

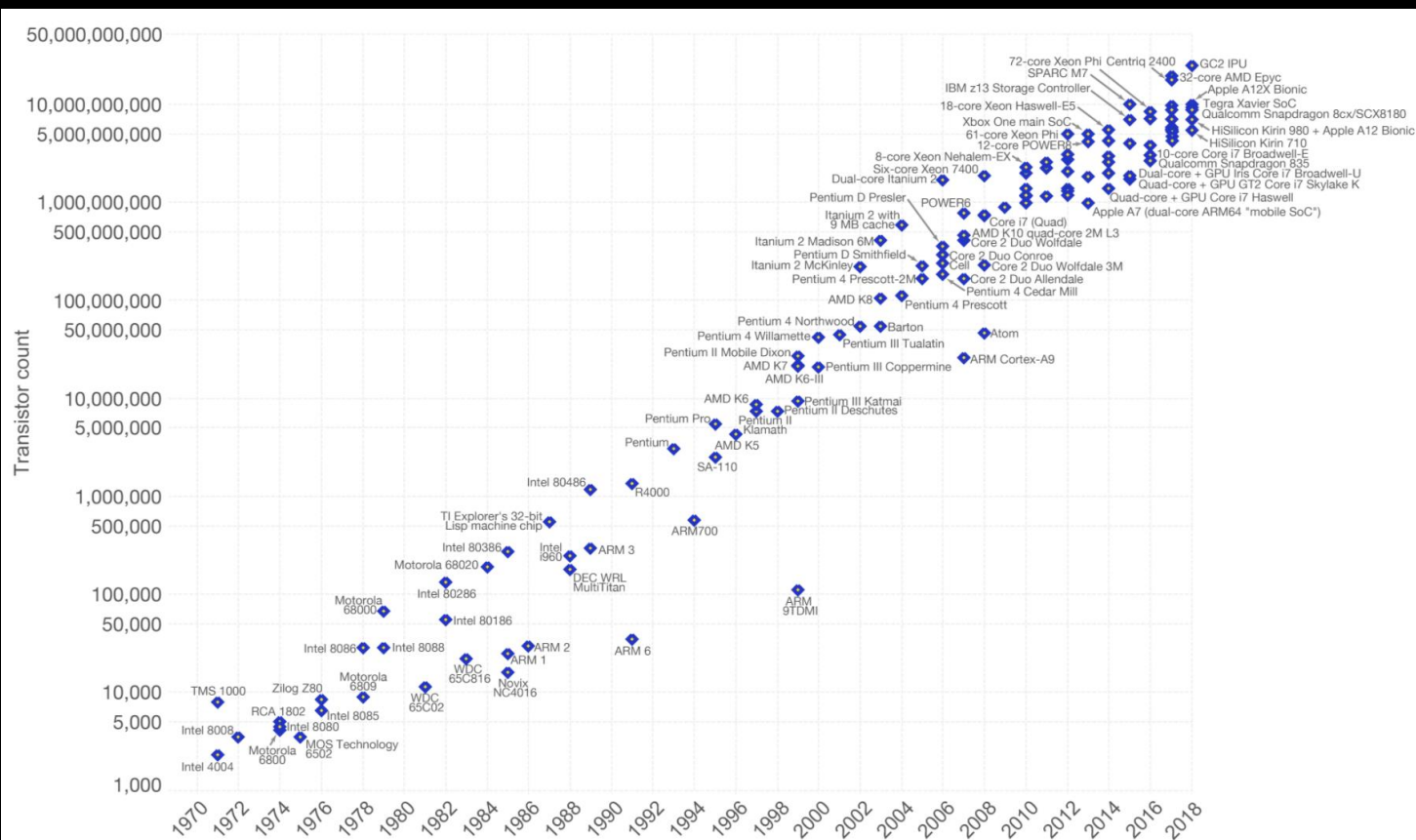
Source: Intel



**Influenza Virus**

Source: CDC

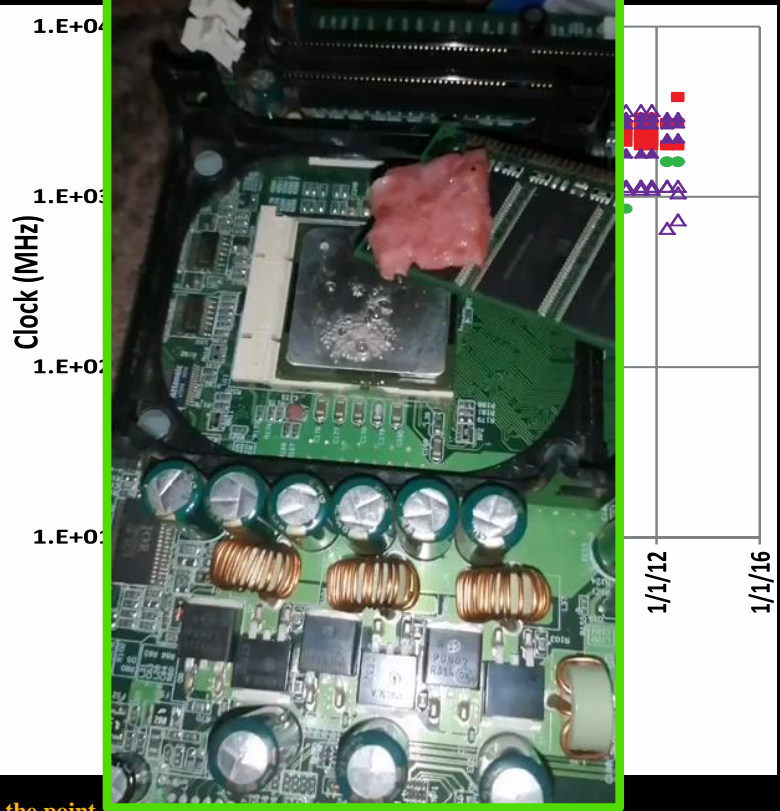
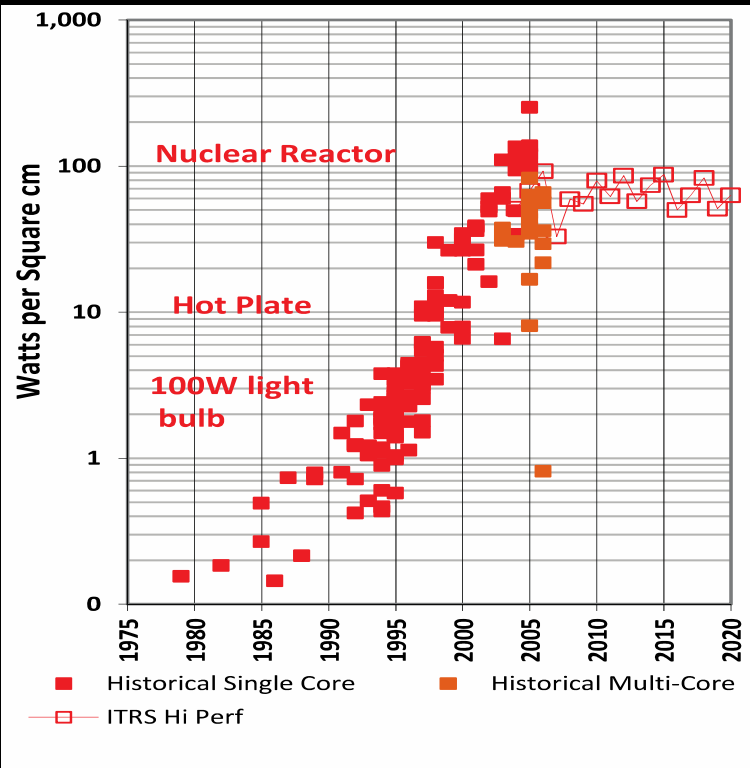
But, at end of day we keep using getting more transistors.



Data source: Wikipedia ([https://en.wikipedia.org/wiki/Transistor\\_count](https://en.wikipedia.org/wiki/Transistor_count))

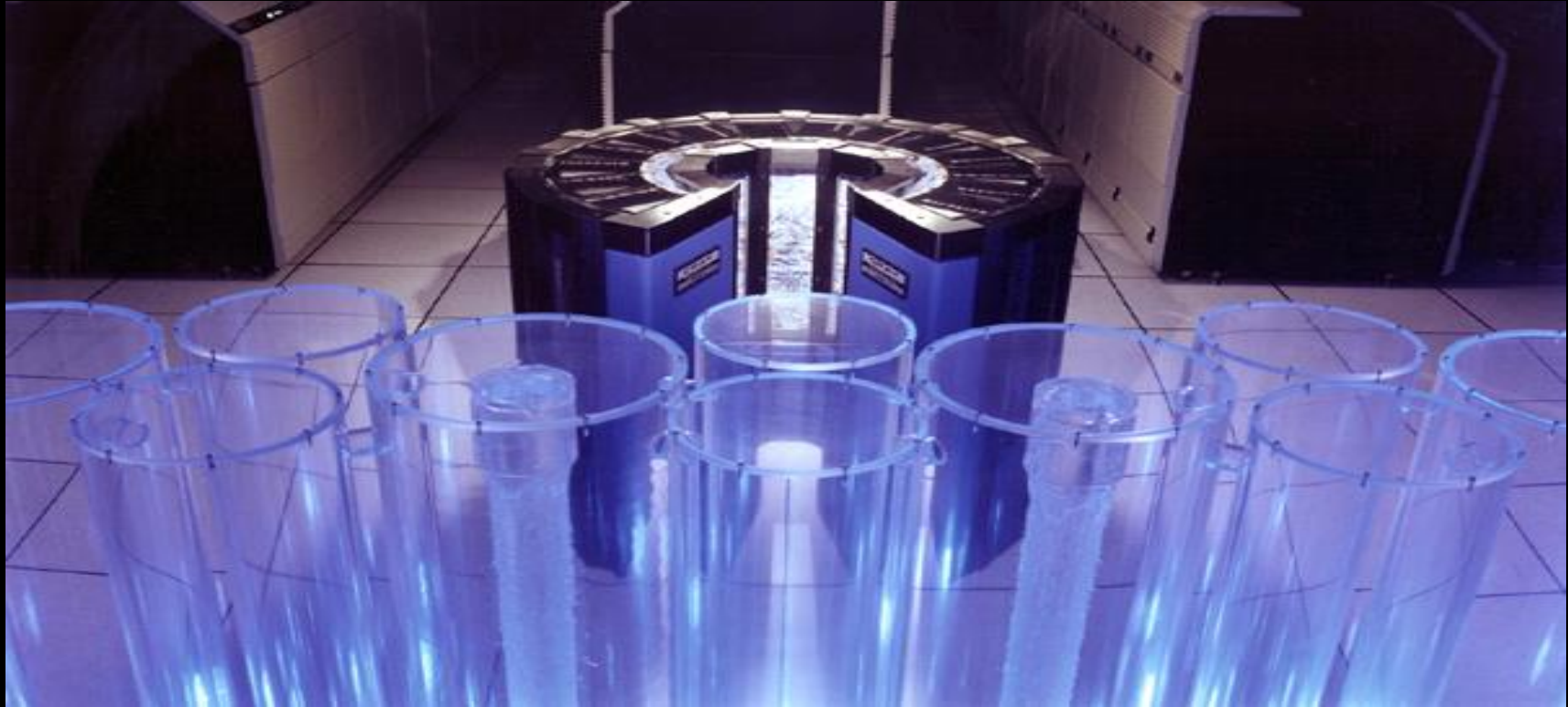
The data visualization is available at [OurWorldinData.org](https://ourworldindata.org). There you find more visualizations and research on this topic.

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.

# Not a new problem, just a new scale...

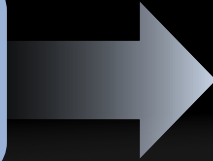


Cray-2 with cooling tower in foreground, circa 1985

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

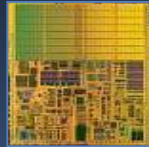
## **RULE OF THUMB**

**A 15%  
Reduction  
In Voltage  
Yields**



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

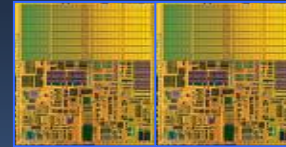
### **SINGLE CORE**



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



### **DUAL CORE**

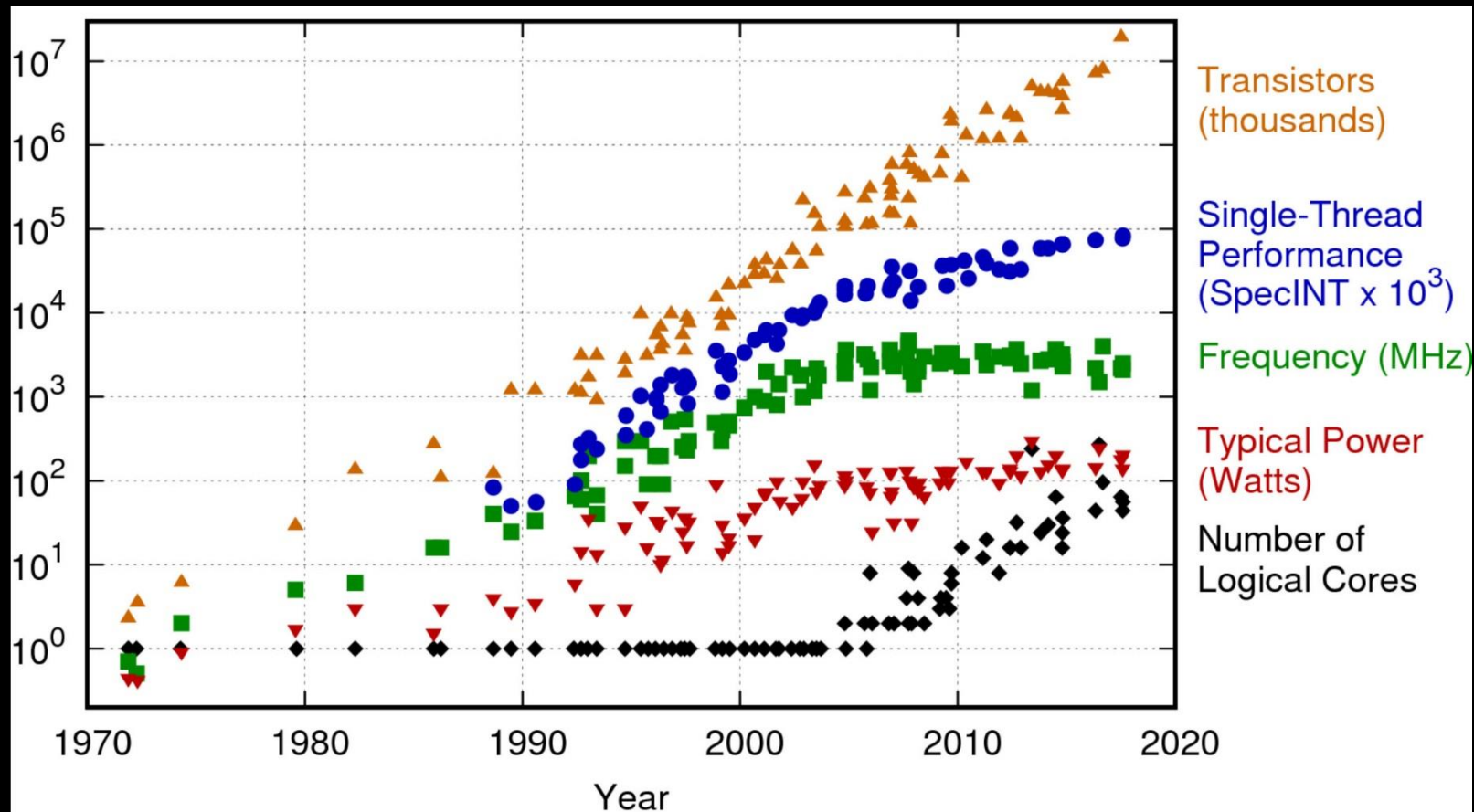


**Area = 2**  
**Voltage = 0.85**  
**Freq = 0.85**  
**Power = 1**  
**Perf = ~1.8**

# 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

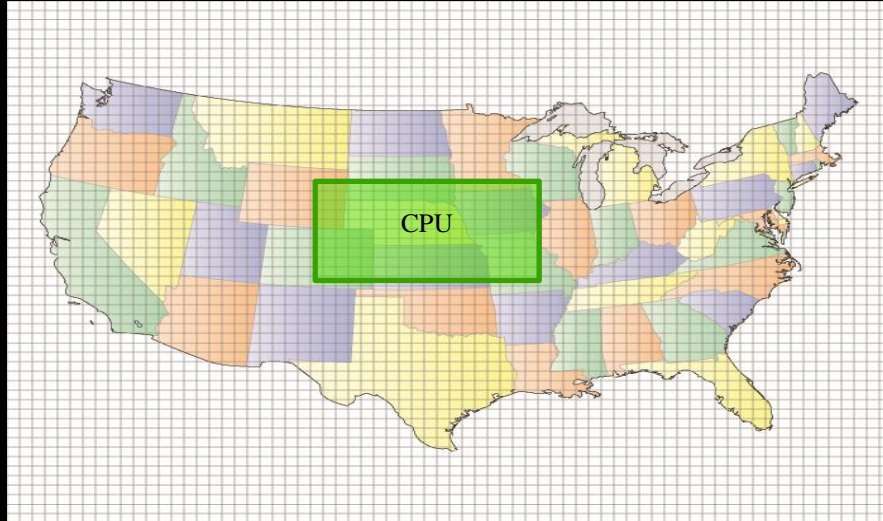
**One woman can make a baby in 9 months.**

**Can 9 women make a baby in 1 month?**

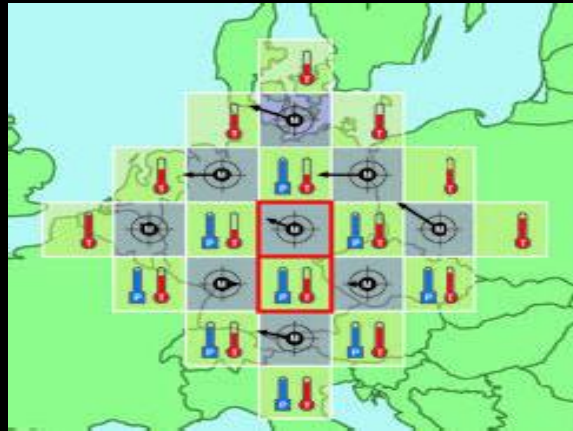
**But 9 women can make 9 babies in 9 months.**

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

# Prototypical Application: Serial Weather Model

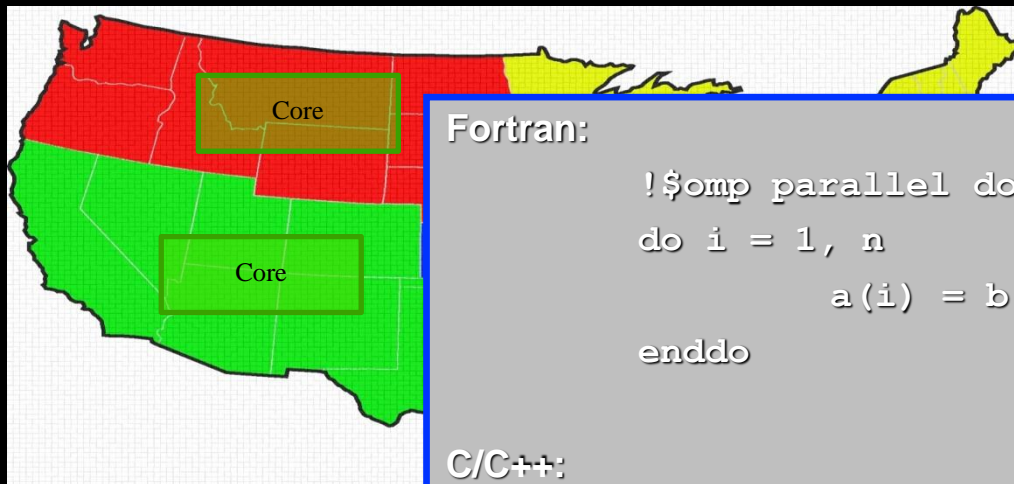


# First Parallel Weather Modeling Algorithm: Richardson in 1917



*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 the*

# V100 GPU and SM



Volta GV100 GPU with 85 Streaming Multiprocessor (SM) units

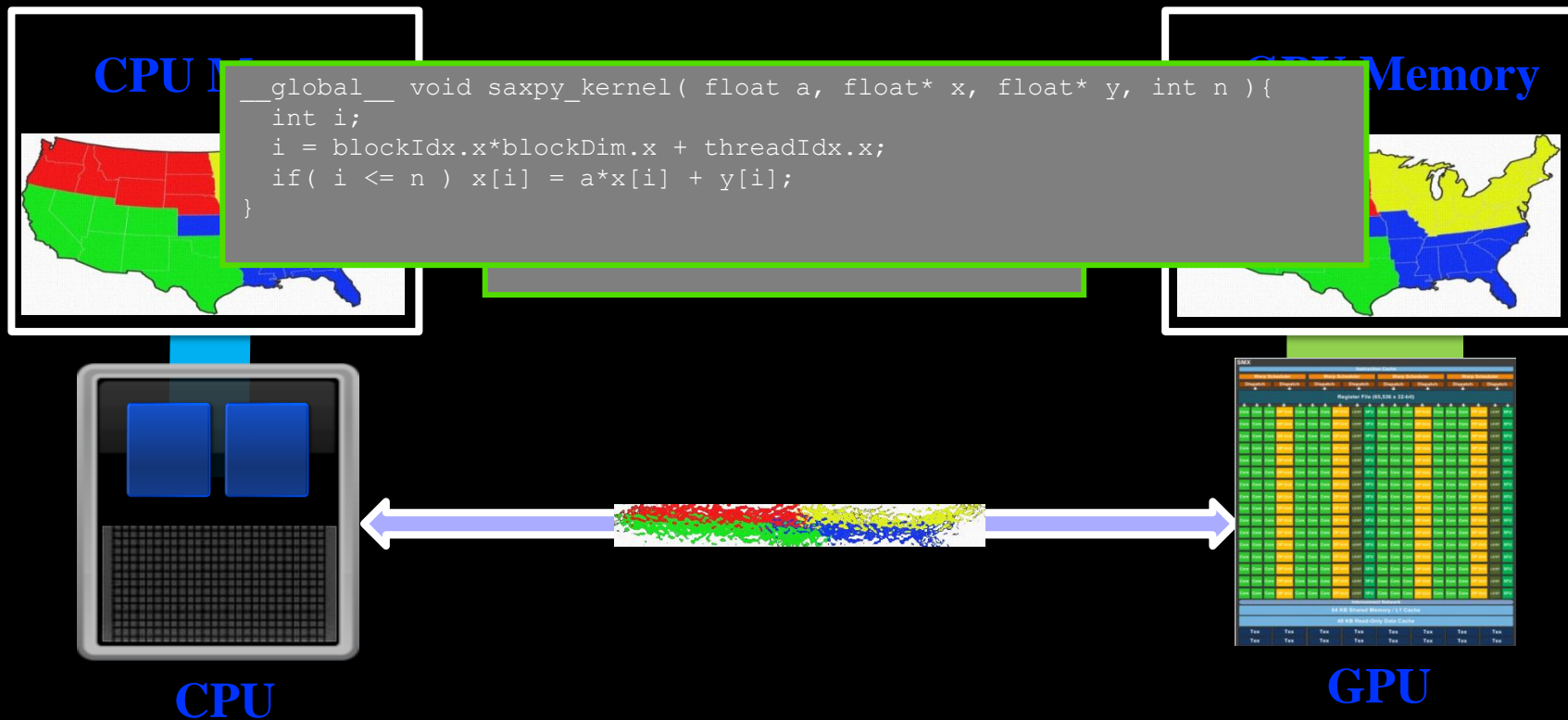


Volta GV100 SM

Rapid evolution  
continues with:

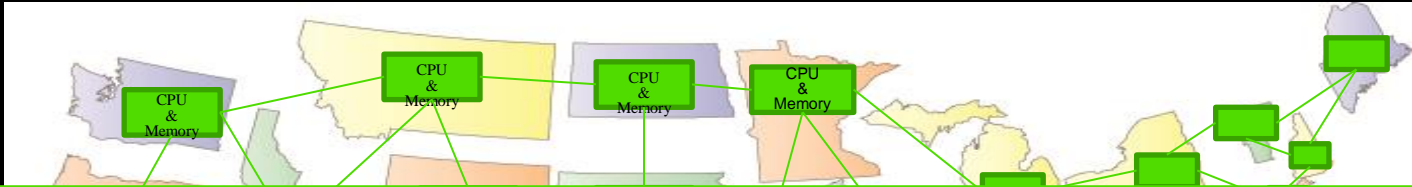
*Turing*  
*Ampere*  
*Hopper*

# Weather Model: Accelerator (OpenACC)



*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)
```

▪  
▪  
▪

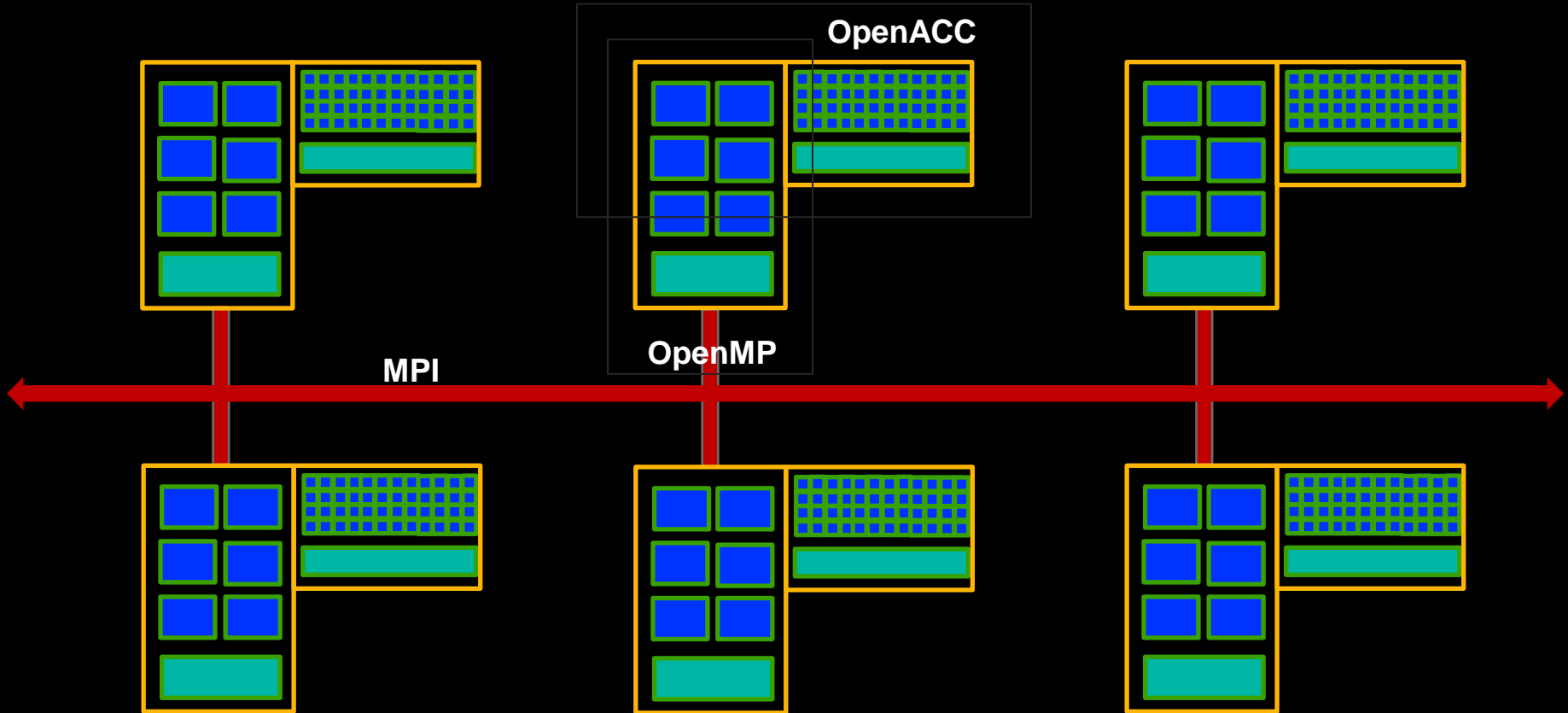
```
call MPI_Barrier(MPI_COMM_WORLD, errcode)
```

▪



*50 meteorologists using a telegraph.*

# 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.*

# 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

OpenACC

MPI

- Multi-Core (Threads)
- SMP/Multi-socket
- Accelerators: GPU & MIC
- Clusters
- MPPs

Also Important

- ASIC/FPGA/DSP
- RAID/IO

# 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



#	Site	Manufacturer	Computer	CPU Interconnect [Accelerator]	Cores	Rmax (Tflops)	Rpeak (Tflops)	Power (MW)
1	RIKEN Center for Computational Science <b>Japan</b>	Fujitsu	Fugaku	ARM 8.2A+ 48C 2.2GHz Torus Fusion Interconnect	7,299,072	415,530	513,854	28.3
2	DOE/SC/ORNL <b>United States</b>	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 <b>United States</b>	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 <b>China</b>	NRCPC	Sunway TaihuLight	Sunway SW26010 260C 1.45GHz	10,649,600	93,014	125,435	15.3
5	National Super Computer Center in Guangzhou <b>China</b>	NUDT	Tianhe-2 (MilkyWay-2)	Intel Xeon E5-2692 2.2 GHz TH Express-2 Intel Xeon Phi 31S1P	4,981,760	61,444	100,678	18.4
6	Eni S.p.A <b>Italy</b>	Dell	HPc5	Xeon 24C 2.1 GHz Infiniband HDR NVIDIA V100	669,760	35,450	51,720	2.2
7	Eni S.p.A <b>Italy</b>	NVIDIA	Selene	EPYC 64C 2.25GHz Infiniband HDR NVIDIA A100	272,800	27,580	34,568	1.3
8	Texas Advanced Computing Center/Univ. of Texas <b>United States</b>	Dell	Frontera	Intel Xeon 8280 28C 2.7 GHz InfiniBand HDR	448,448	23,516	38,745	
9	Cineca <b>Italy</b>	IBM	Marconi100	Power9 16C 3.0 GHz Infiniband EDR NVIDIA V100	347,776	21,640	29,354	1.5
10	Swiss National Supercomputing Centre (CSCS) <b>Switzerland</b>	Cray	Piz Daint Cray XC50	Xeon E5-2690 2.6 GHz Aries NVIDIA P100	387,872	21,230	27,154	2.4

# Networks

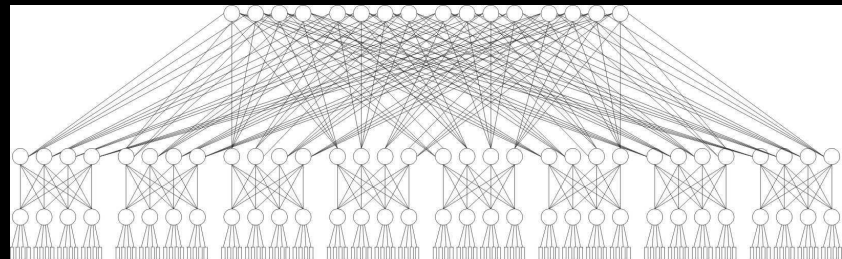
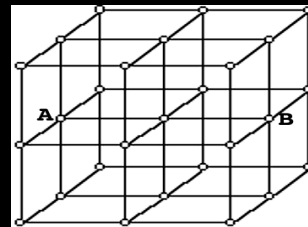
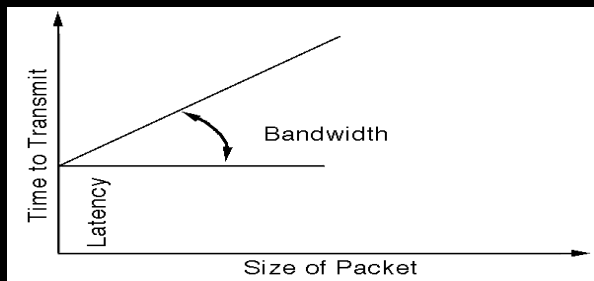
## 3 characteristics sum up the network:

- **Latency**

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

- **Bandwidth**

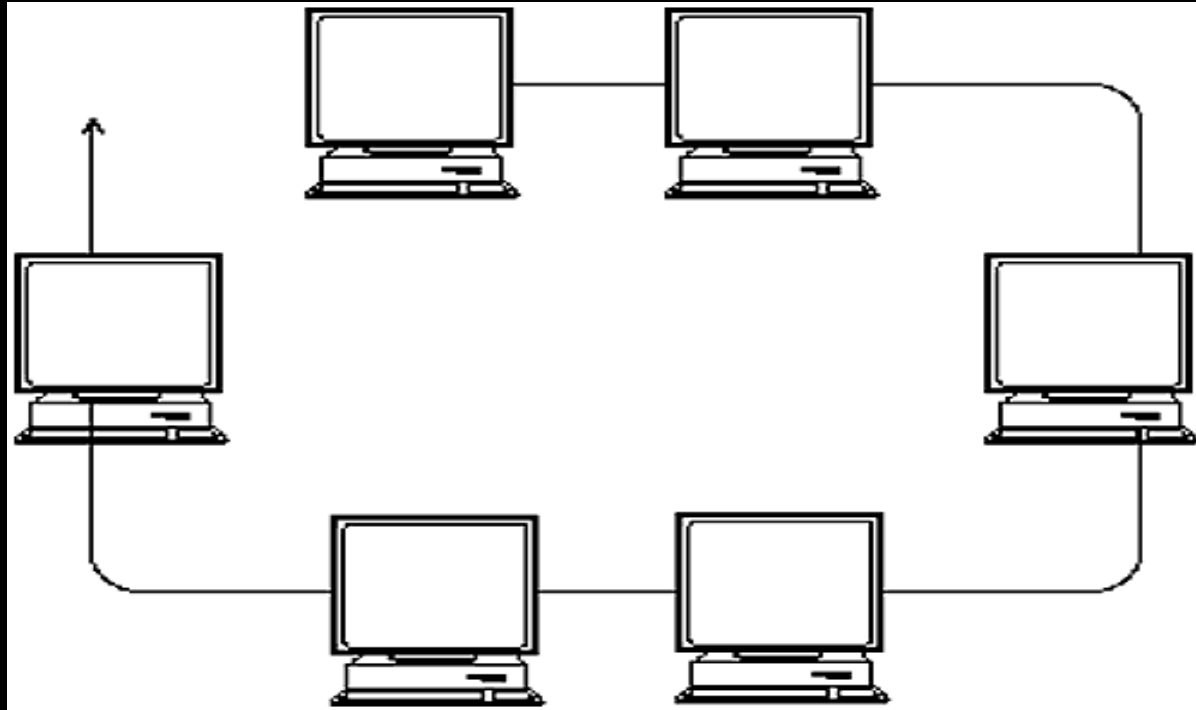
The rate at which a very large packet of information can be sent



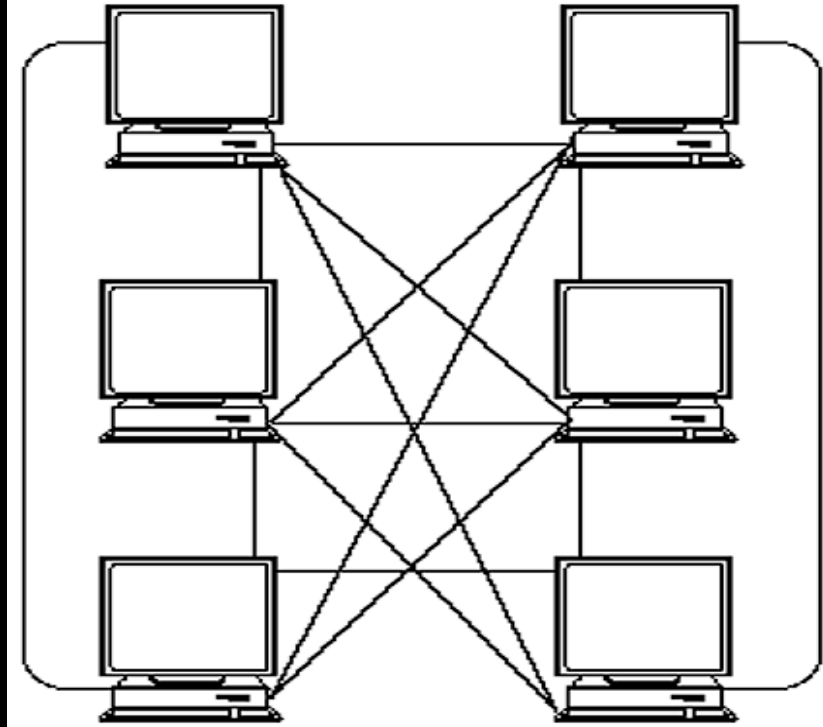
- **Topology**

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

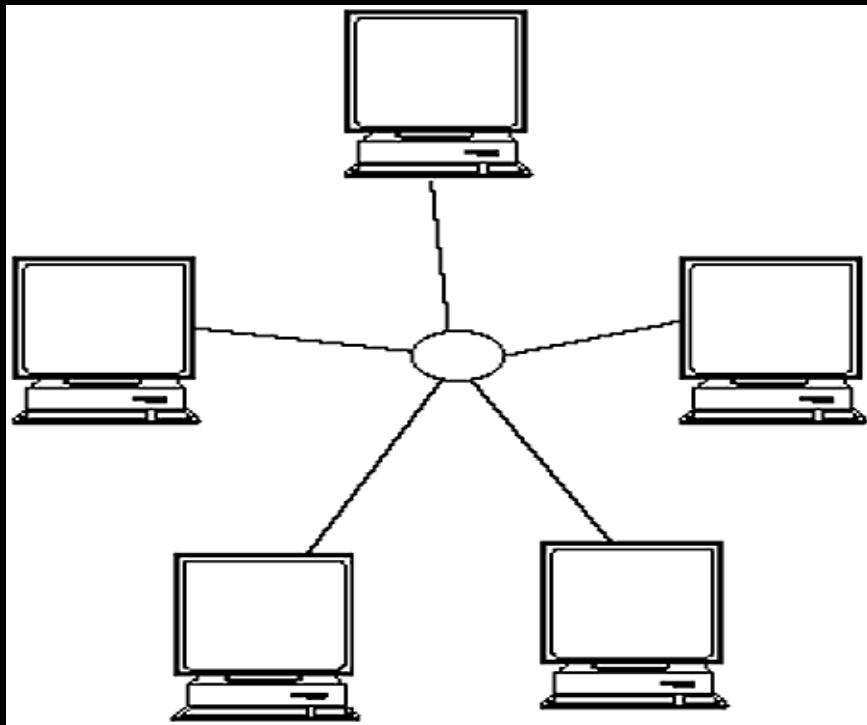
# Ethernet with Workstations



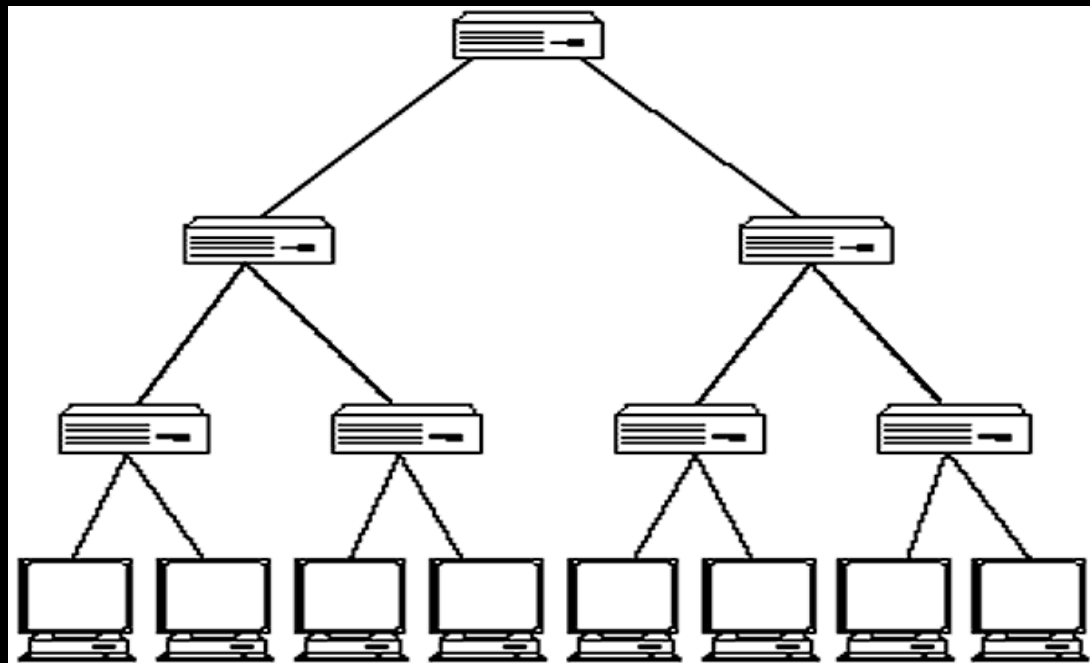
# Complete Connectivity



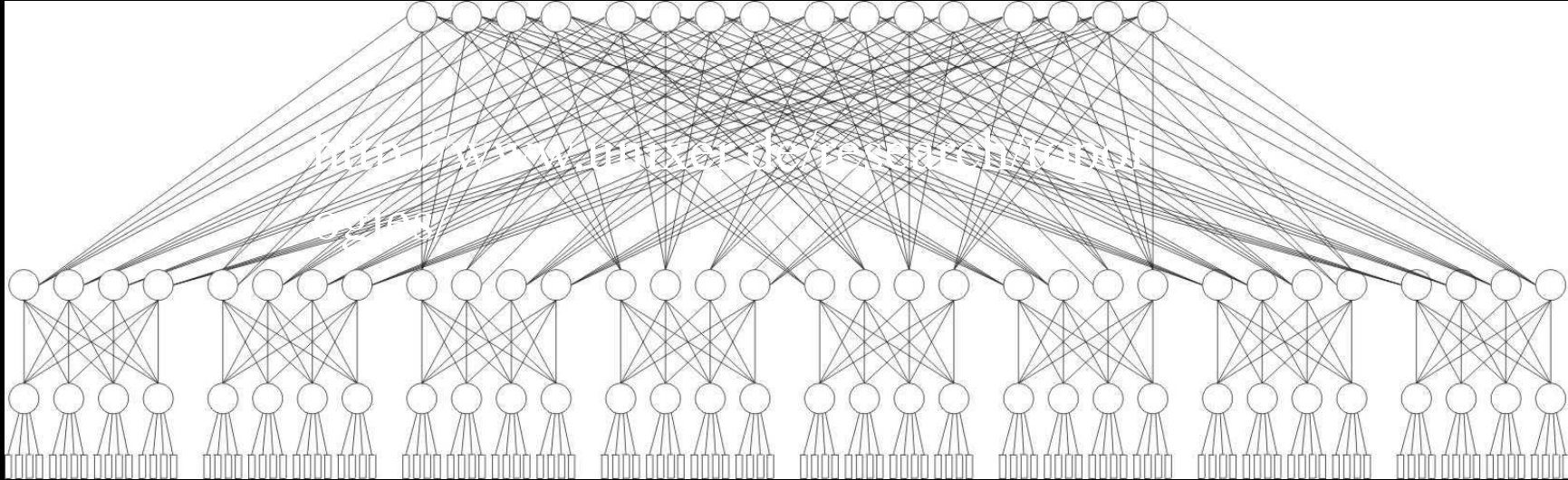
# Crossbar



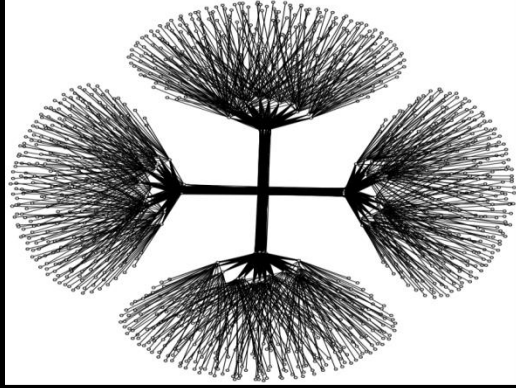
# Binary Tree



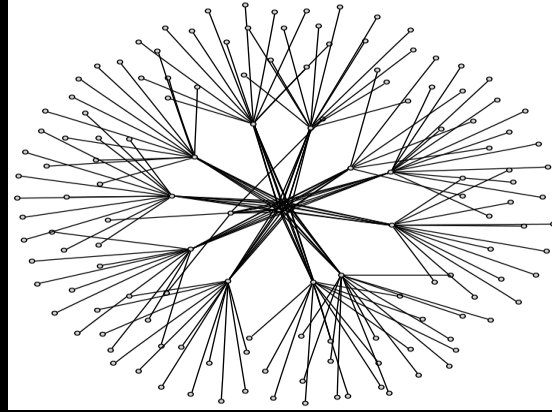
# Fat Tree



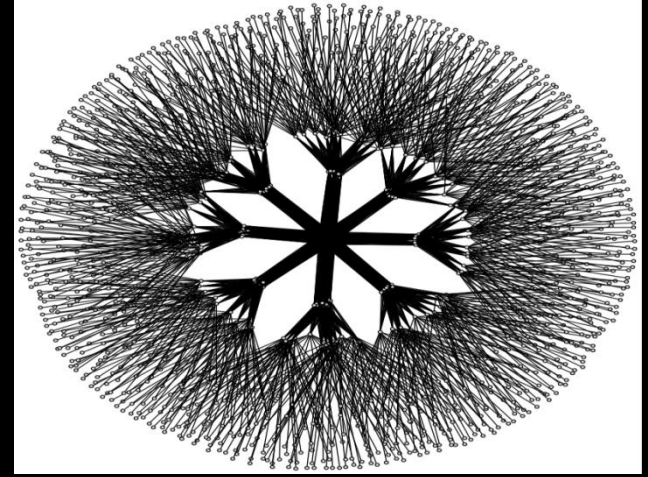
# Other Fat Trees



Big Red @ IU

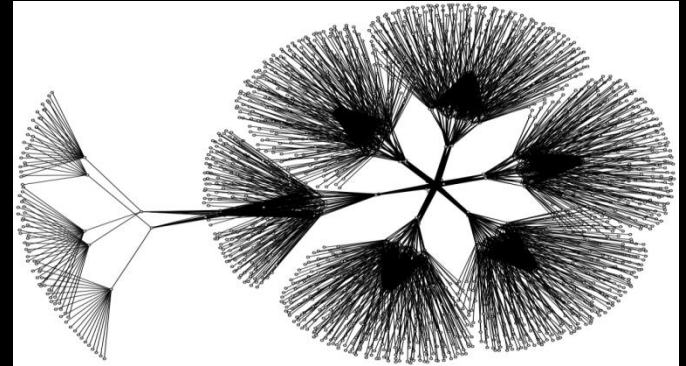
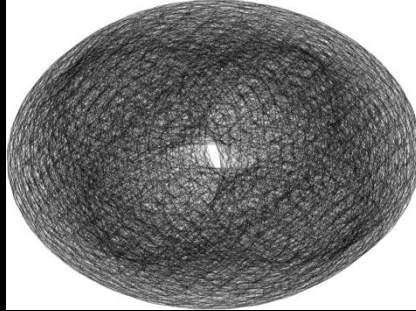


Odin @ IU



Atlas @ LLNL

Jaguar @ ORNL



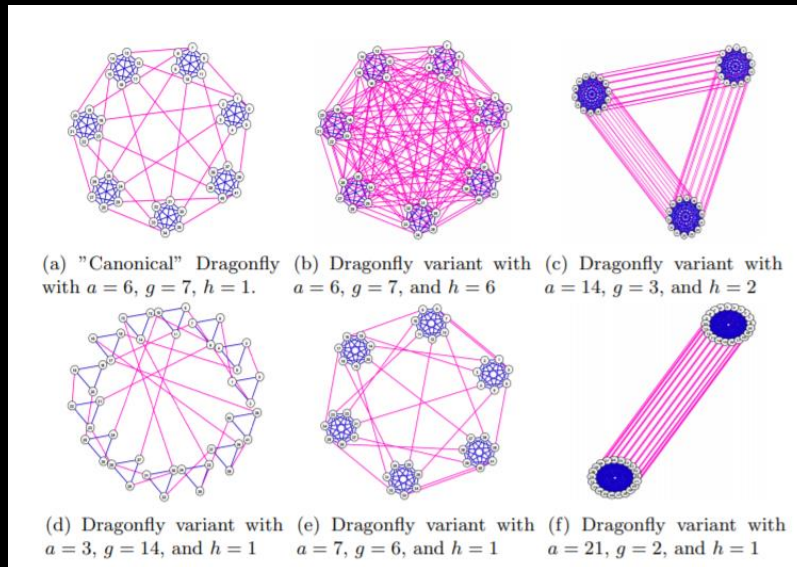
Tsubame @ Tokyo Inst. of Tech

# Dragonfly

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

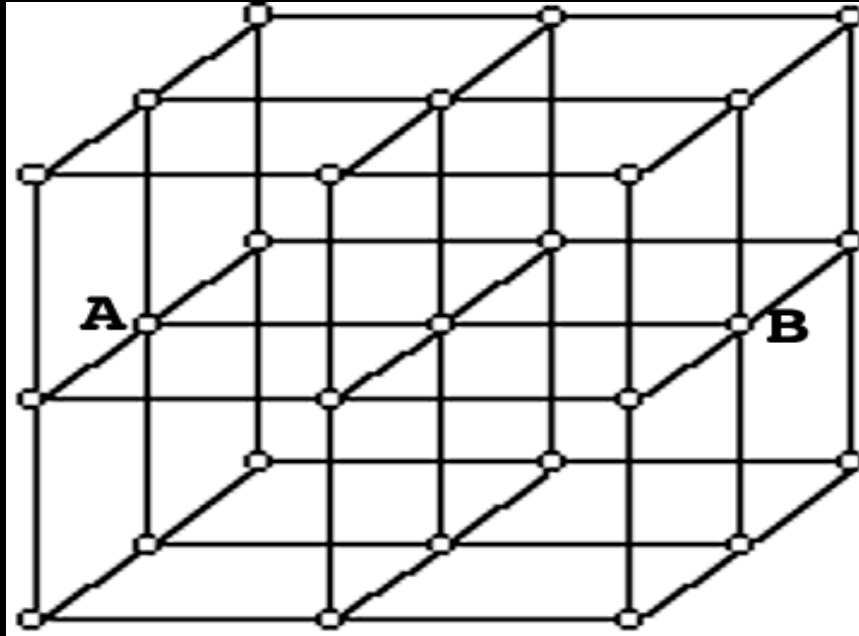
- High-Radix Switches
- Adaptive Routing
- Optical Links

## Various 42 node Dragonfly configurations.



Purple links are optical, and blue are electrical.

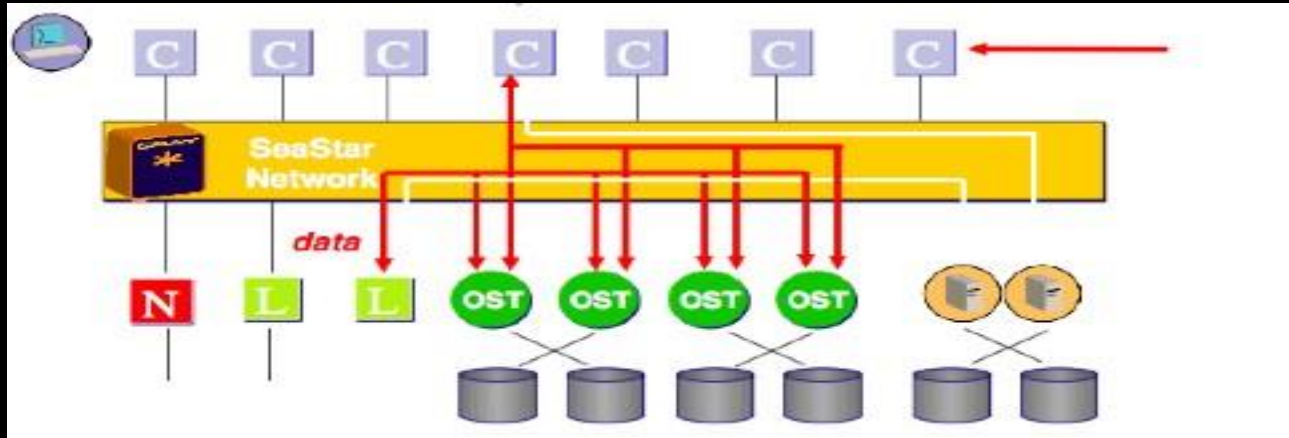
# 3-D Torus



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.



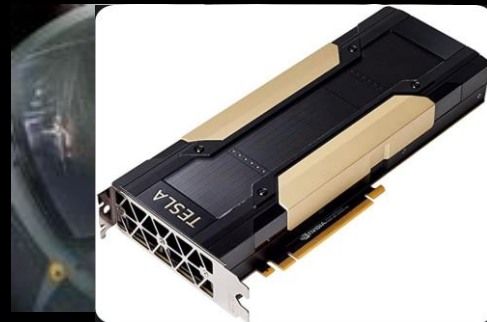
# The Future Is Now!

Exascale Computing and you.

# Welcome to 2021: the year of Exascale!

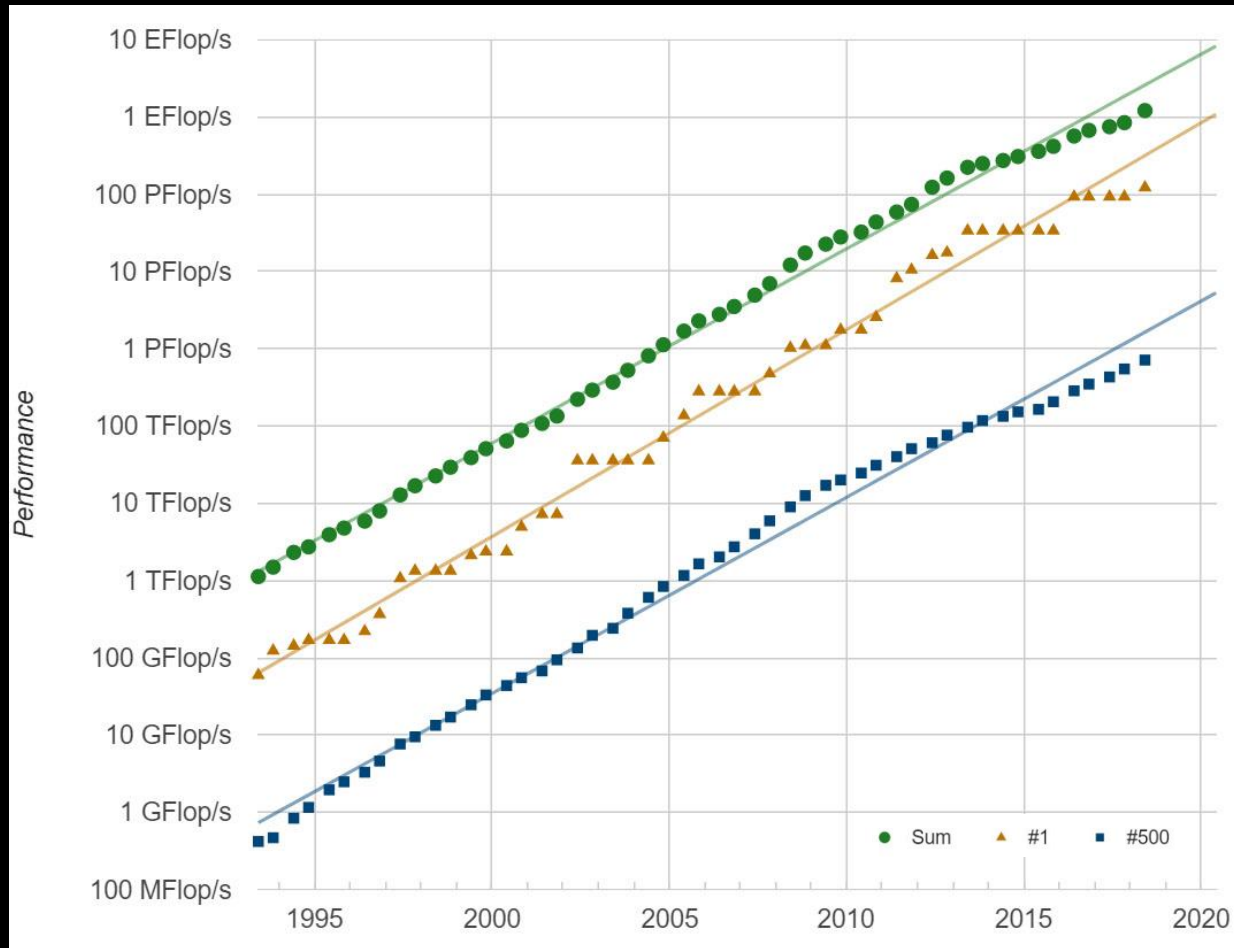
exa =  $10^{18}$  = 1,000,000,000,000,000,000 = quintillion

64-bit precision floating point operations per second



23,800,33  
Cray Red Storms  
NVIDIA V100  
2004 (42 Tflops)

# Sustaining Performance Improvements



# 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

# The Plan

## Pre-Exascale Systems

## Future Exascale Systems

2012



ORNL  
Cray/AMD/  
NVIDIA

2016



LBNL  
Cray/Intel

2018



ORNL  
IBM/NVIDIA

2020



LBNL  
Cray/AMD/  
NVIDIA

2021–2023



ORNL  
Cray/AMD



ANL  
IBM BG/Q



ANL  
Intel/Cray



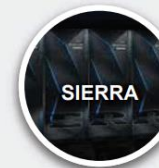
ANL  
Intel/Cray



LLNL  
IBM BG/Q



LANL/SNL  
Cray/Intel



LLNL  
IBM/NVIDIA



EL CAPITAN

LLNL  
Cray



CROSSROADS

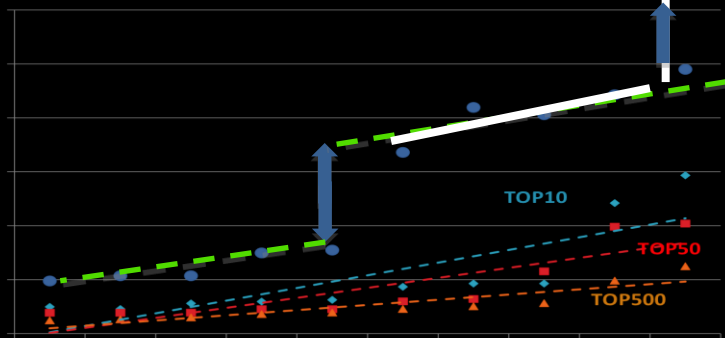
LANL/SNL  
TBD

# System Designs

System	Performance	Power	Interconnect	Node
Aurora (ANL)	> 1 EF		100 GB/s Cray Slingshot Dragonfly	2 Intel Xeon CPU + 6 Intel Xe GPUs
El Capitan (LLNL)	> 1.5 EF	30-40 MW	100 GB/s Cray Slingshot Dragonfly	AMD Epyc CPU + 4 Radeon GPUs
Frontier (ORNL)	> 1.5 EF		100 GB/s Cray Slingshot Dragonfly	AMD Epyc CPU + 4 Radeon GPUs
Perlmutter (LBNL)			Cray Slingshot Dragonfly	2 AMD Epyc CPU + 4 Volta GPUs

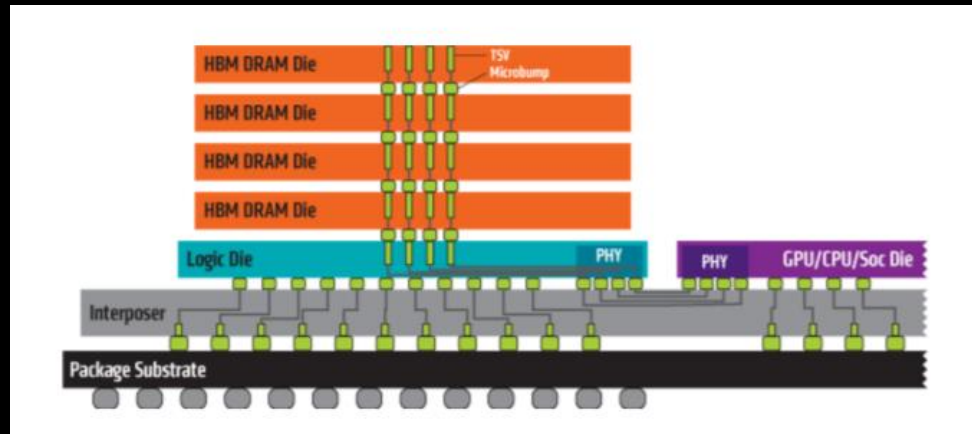
# Two Additional Boosts to Improve Flops/Watt and Reach Exascale Target

First boost: many-core/accelerator



Third Boost: SiPh (2020 – 2024)

Second Boost: 3D (2016 – 2020)



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

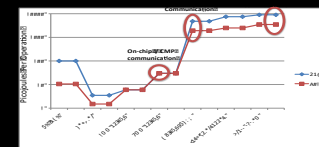
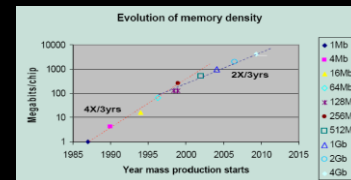
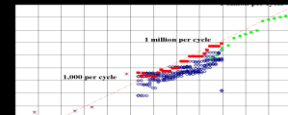
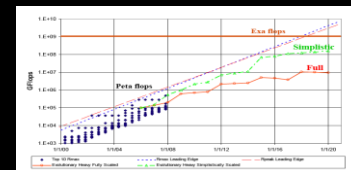
*Current programming systems have WRONG optimization targets*

## 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*

# End of Moore's Law Will Lead to New Architectures

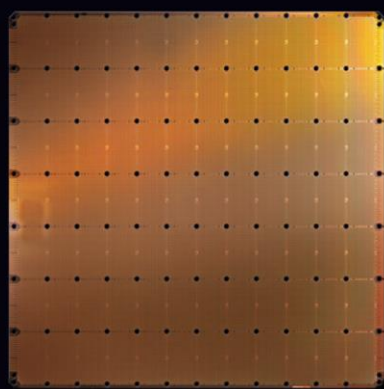


Non-von  
Neumann

NEUROMORPHIC

ARCHITECT

von Neu



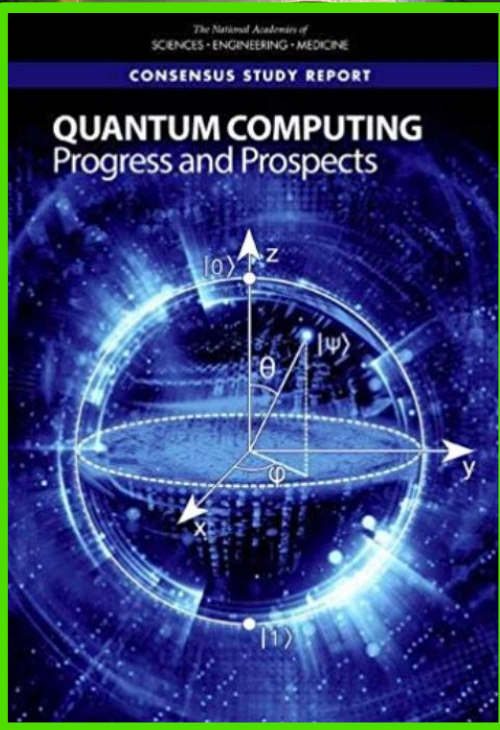
Cerebras WSE  
1.2 Trillion transistors  
46,225 mm<sup>2</sup> silicon



Largest GPU  
21.1 Billion transistors  
815 mm<sup>2</sup> silicon



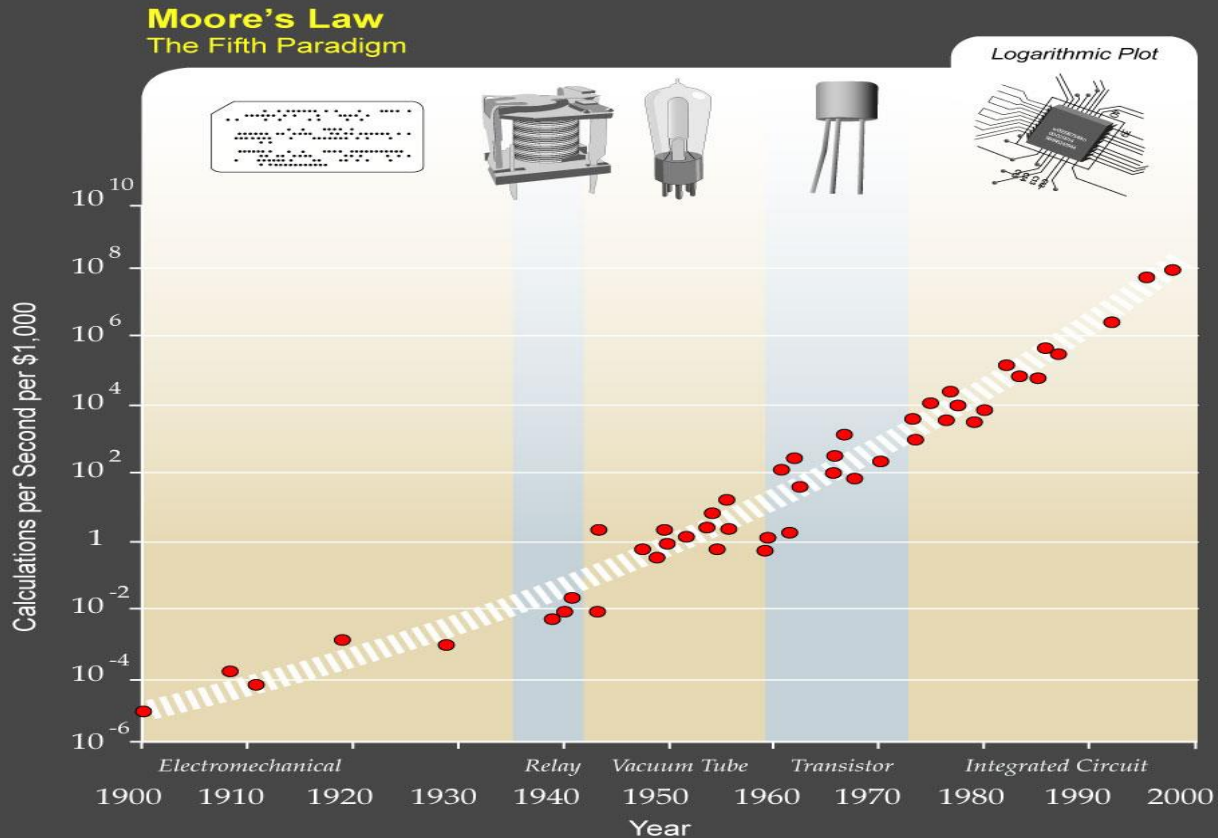
B



Beyond CMOS

TECHNOLOGY

# It would only be the 6<sup>th</sup> 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**
- **Current predictions envision Exascale computers in 2022+ with a power consumption of at best 20 - 30 MW**
- **The human brain takes 20W**
- **Even under best assumptions in 2020 our brain will still be a million times more power efficient**



Copyrighted Material

# COMPUTATIONAL PHYSICS

Revised and expanded

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*.

*Mark Newman*

Copyrighted Material

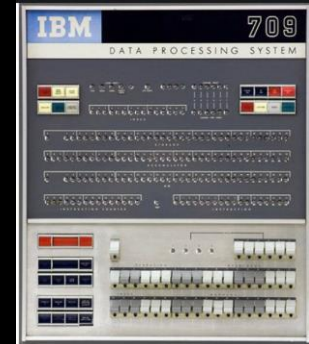
**Where are those 10 or 12 orders of  
magnitude?**

**How do we get there from here?**

**BTW, that's a  
bigger gap than**



**VS.**



**IBM 709  
12 kiloflops**

# 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...

