

Intro To Parallel Computing

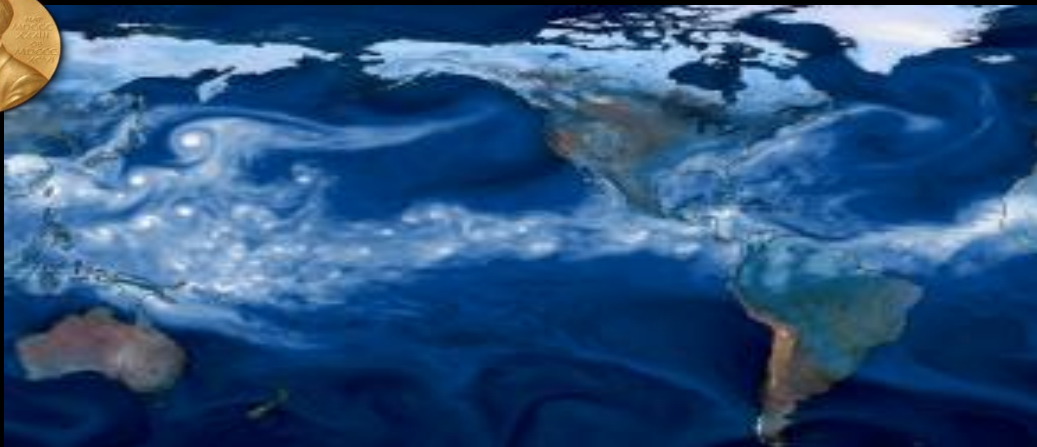
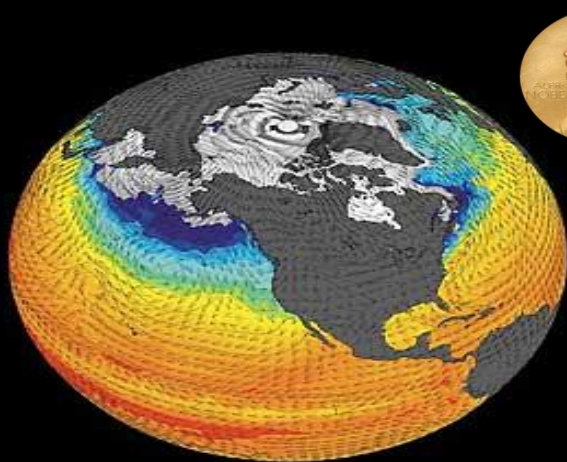
John Urbanic

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.

Compute bound problems abound: Climate change analysis



Simulations

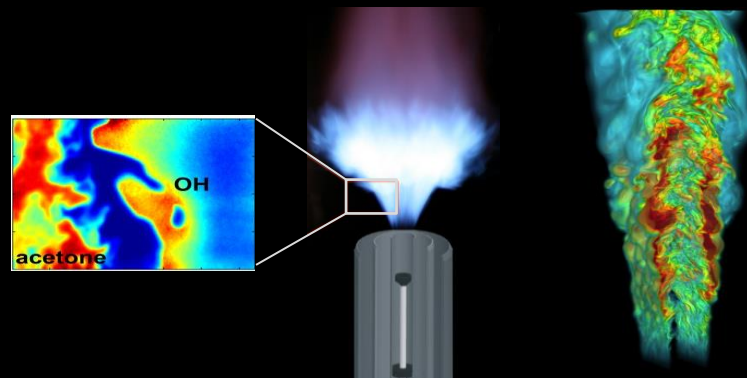
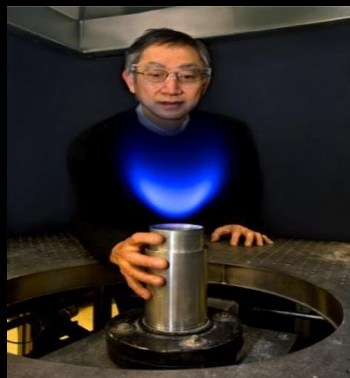
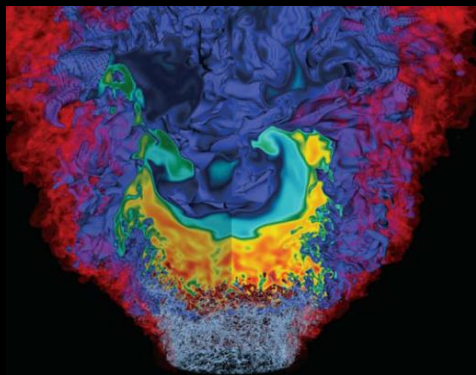
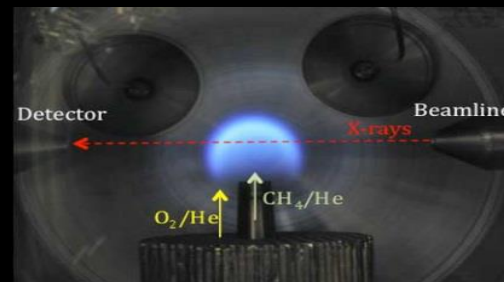
- Cloud resolution, quantifying uncertainty, understanding tipping points, etc., will drive climate to exascale platforms
- New math, models, and systems support will be needed

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
-

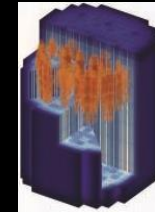
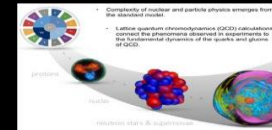
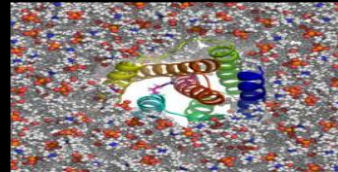
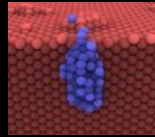
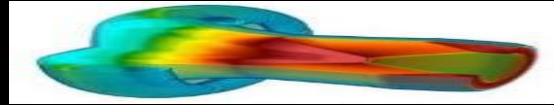
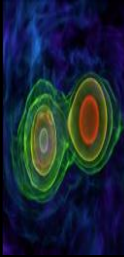
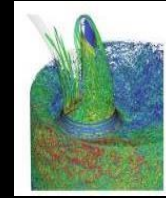
Exascale is needed at all scales: combustion simulations

- 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



The list is long, and growing.

- Molecular-scale Processes: atmospheric aerosol simulations
- AI-Enhanced Science: predicting disruptions in tokamak fusion reactors
- Hypersonic Flight
- Modeling Thermonuclear X-ray Bursts: 3D simulations of a neutron star surface or supernovae
- Quantum Materials Engineering: electrical conductivity photovoltaic and plasmonic devices
- Physics of Fundamental Particles: mass estimates of the bottom quark
- Digital Cells

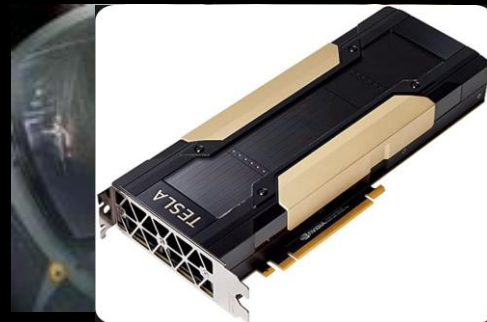


These and others are in an appendix at the end of our Outro To Parallel Computing talk.
And many of you doubtless brought your own immediate research concerns. Great!

Welcome to The Exascale Era!

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

64-bit precision floating point operations per second



23,800,33

Cray RHEL ARM 100

2004 (425 Tflops)

There may also be a Chinese machine, OceanLight, or 3-letter-agency machines on the scene.

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

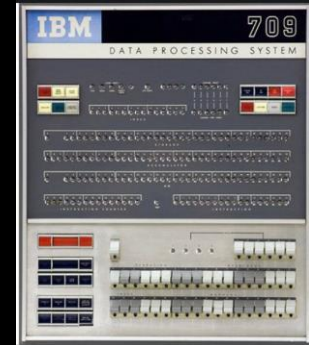
Where are those 10 or 12 orders of magnitude?

How do we get there from here?

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

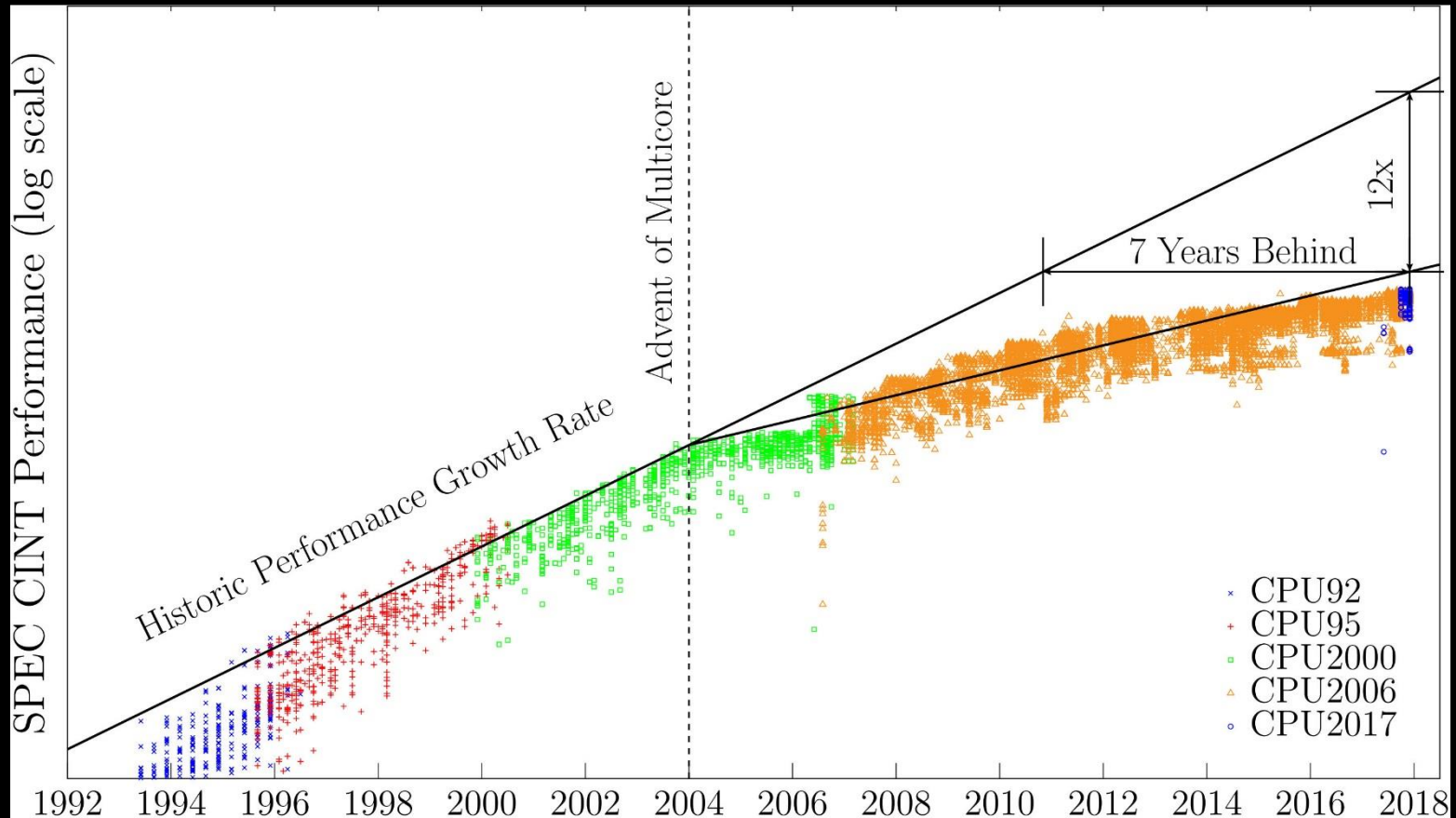


VS.



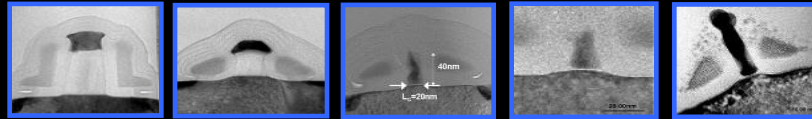
**IBM 709
12 kiloflops**

Moore's Law abandoned serial programming around 2004

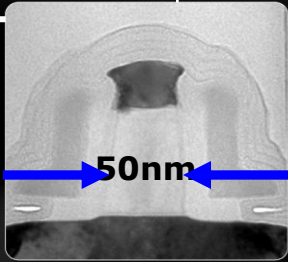


But Moore's Law is only beginning to stumble now.

Intel process technology capabilities

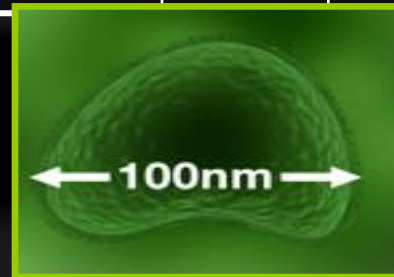


High Volume Manufacturing	2004	2006	2008	2010	2012	2014	2018	2021
Feature Size	90nm	65nm	45nm	32nm	22nm	14nm	10nm	7nm
Integration Capacity (Billions of Transistors)	2	4	8	16	32	64	128	256



**Transistor for
90nm Process**

Source: Intel

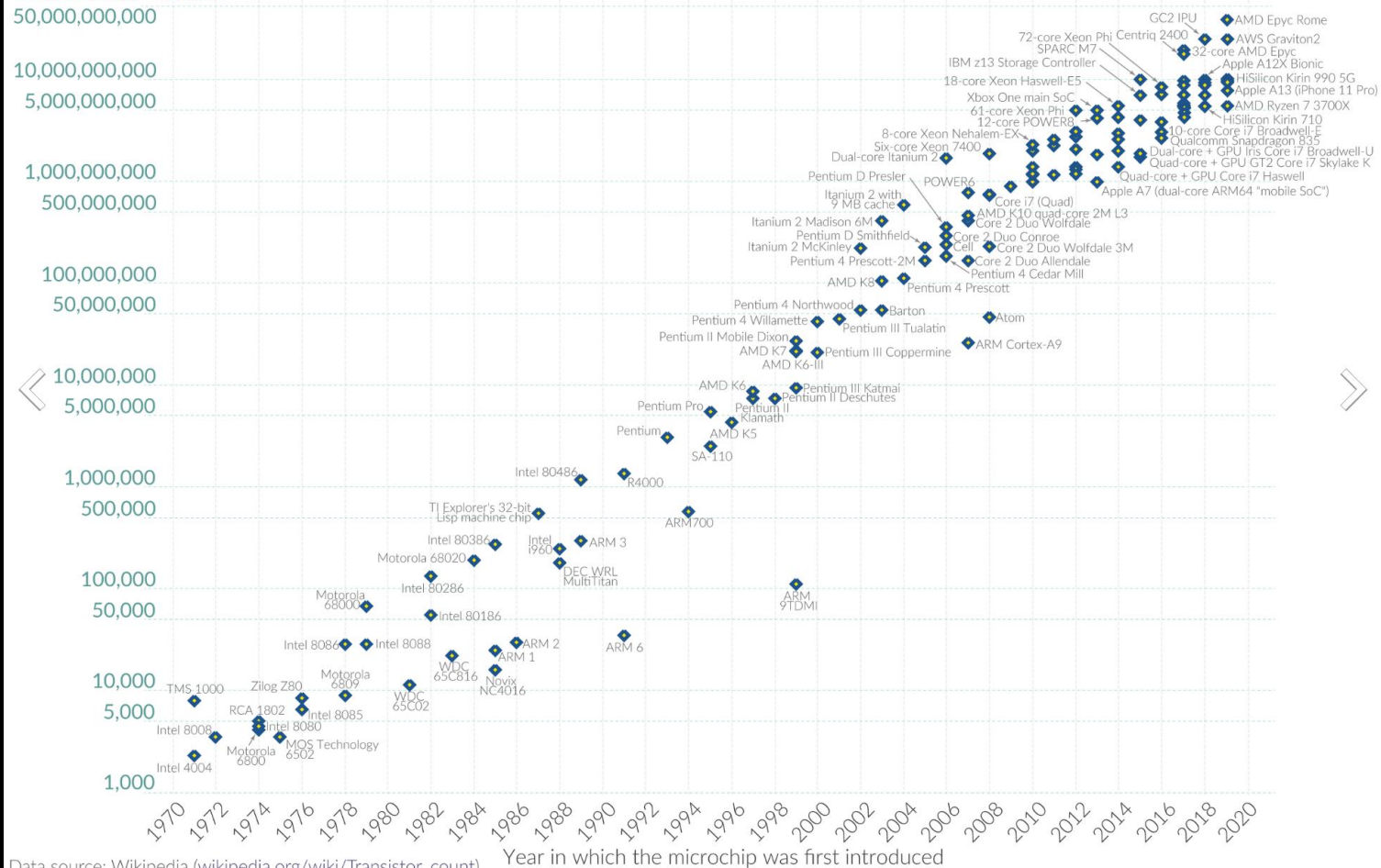


Influenza Virus

Source: CDC

And at end of day we keep using getting more transistors.

Transistor count

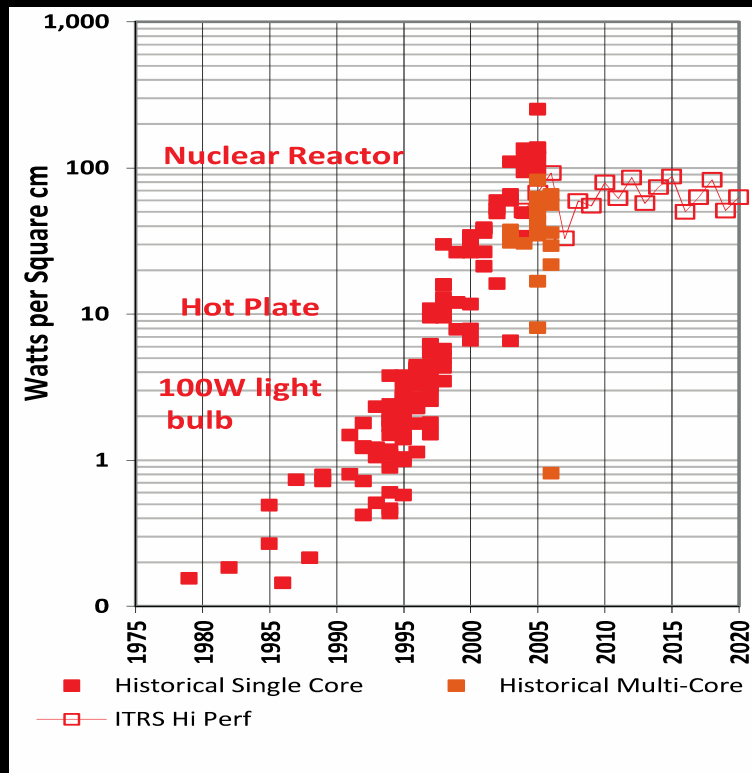


Data source: Wikipedia ([wikipedia.org/wiki/Transistor_count](https://en.wikipedia.org/wiki/Transistor_count))

OurWorldinData.org – Research and data to make progress against the world's largest problems.

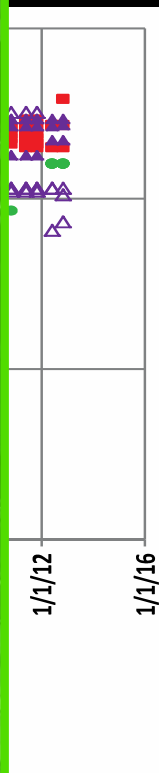
Licensed under CC-BY by the authors Hannah Ritchie and Max Roser.

And run into the real problem. This is the central driver of 21st century computing!



1.E+04
1.E+03
1.E+02
1.E+01

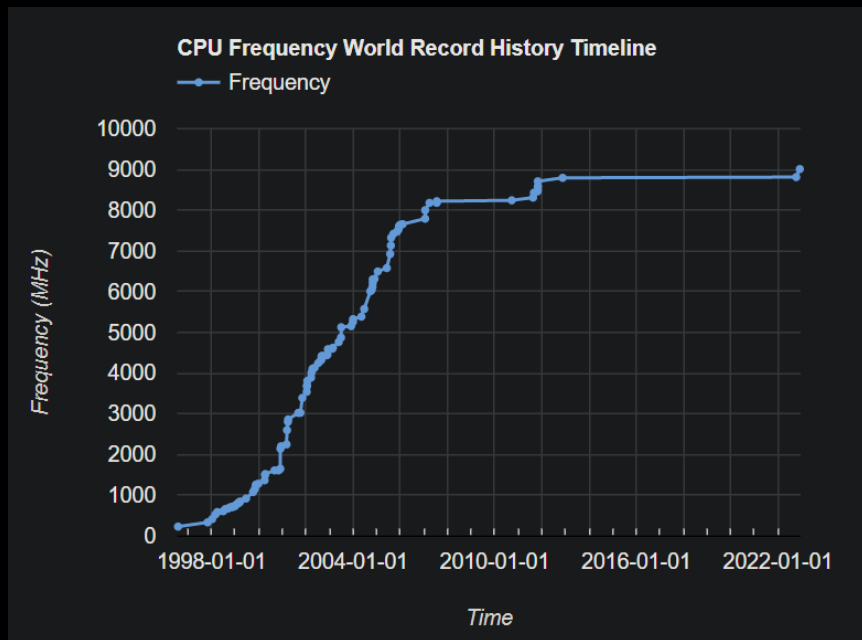
Clock (MHz)



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

Courtesy Horst Simon, LBNL

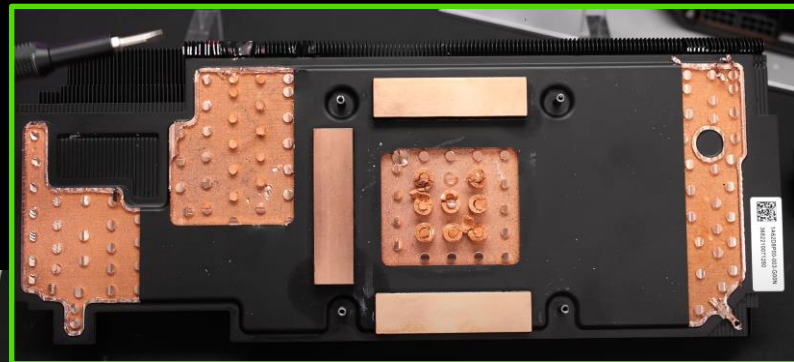
Even when you go extreme...



These are CPUs you can buy.

https://hwbot.org/benchmark/cpu_frequency/halloffame

Complex liquid cooling on a consumer GPU.



For those of you thinking, "Well, at least *my* CPU runs at 4+ GHz."

Previous Generation
without Turbo Boost Technology



ALL Workloads

Intel® Core™ i7-920XM Processor
with Turbo

Highly Threaded
Workload < TDP



2.26GHz for all
4 Cores

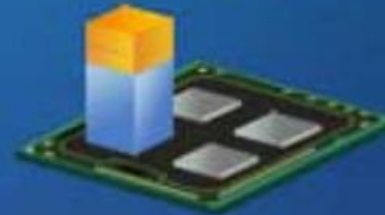
Lightly-Threaded
Workload < TDP



3.06GHz for
2 active cores

Single-Threaded
Workload < TDP

*Near zero power
for inactive cores*



3.20GHz for
1 active core

Maybe sometimes.

Not a new problem...just ubiquitous.



Cray-2 with cooling tower in



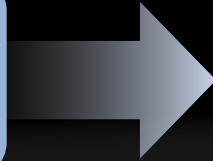
Google's liquid cooled TPU v2 servers deployed in racks. *Source: Google.*

Starting to see 200KW per cabinet in datacenters.

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

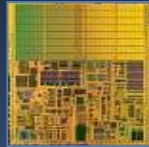
RULE OF THUMB

**A 15%
Reduction
In Voltage
Yields**



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

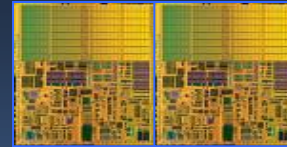
SINGLE CORE



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



DUAL CORE

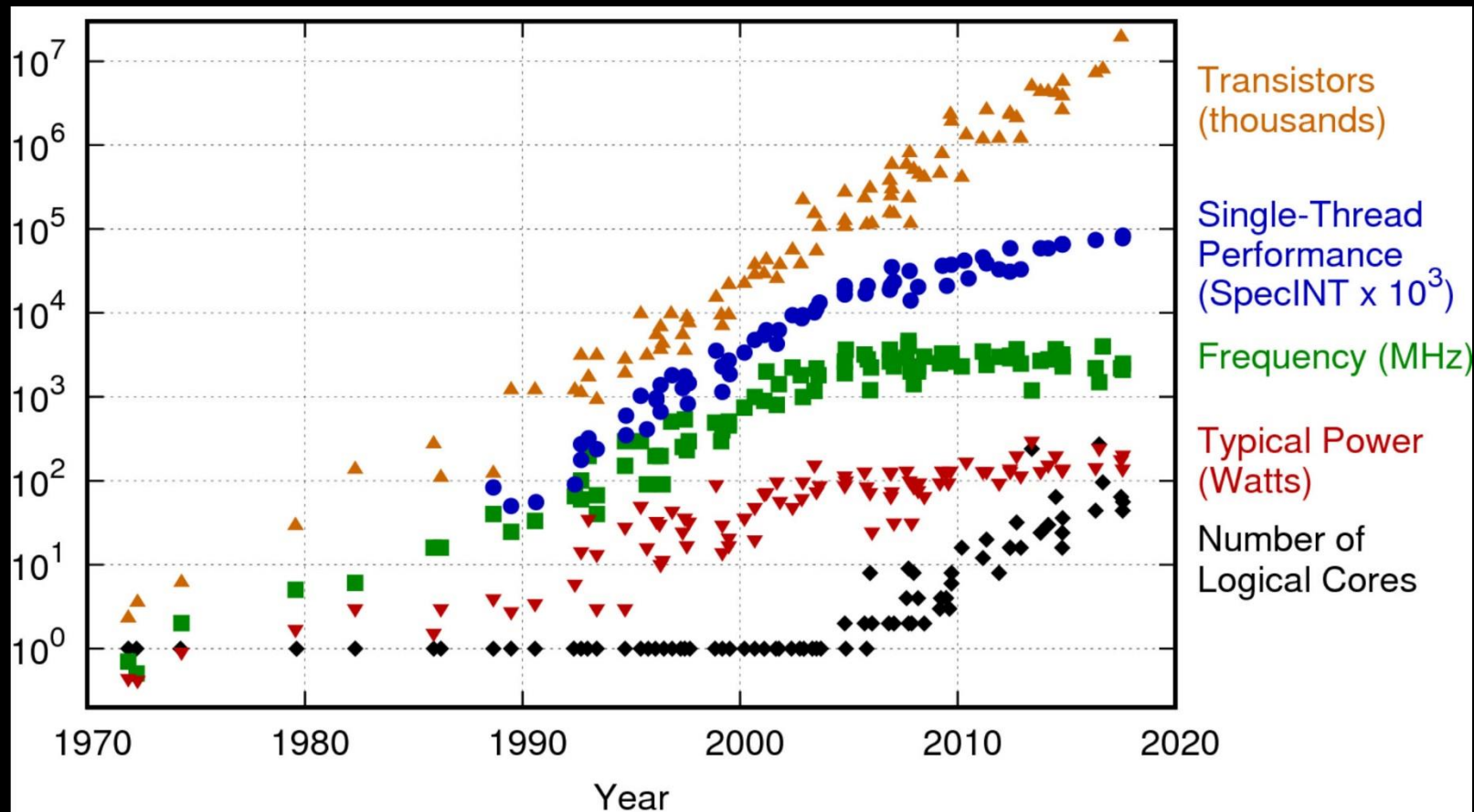


Area = 2
Voltage = 0.85
Freq = 0.85
Power = 1
Perf = ~1.8

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.

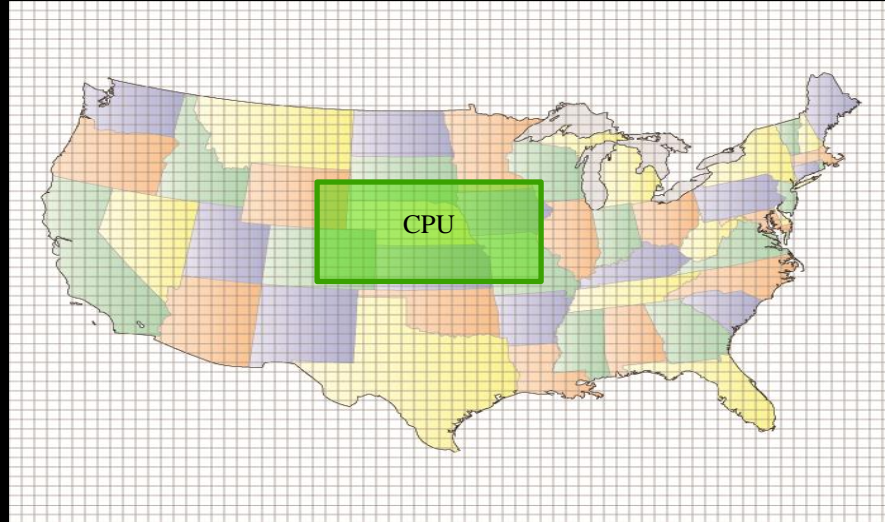
Can 9 women make a baby in 1 month?

But 9 women can make 9 babies in 9 months.

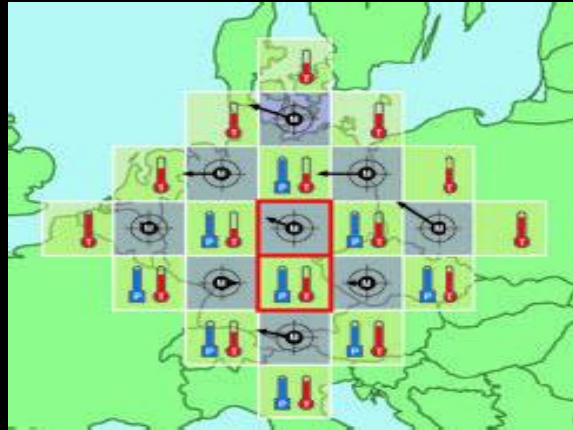
First two bullets are Brook's Law. From *The Mythical Man-Month*.

A must-read for serious project programmers that includes many other classics such as:
"What one programmer can do in one month, two programmers can do in two months."

Prototypical Application: Serial Weather Model

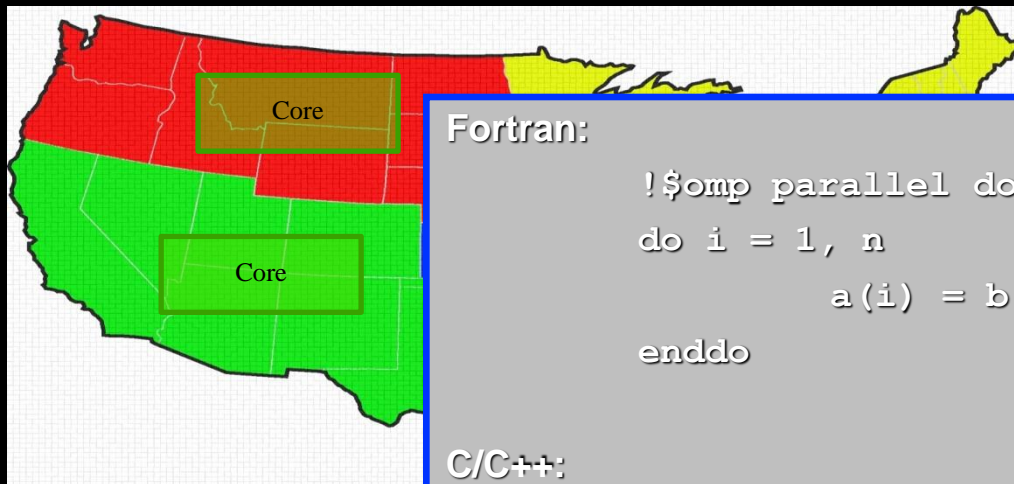


First Parallel Weather Modeling Algorithm: Richardson in 1917



Courtesy John Burkhardt, Virginia Tech

Weather Model: Shared Memory (OpenMP)



Fortran:

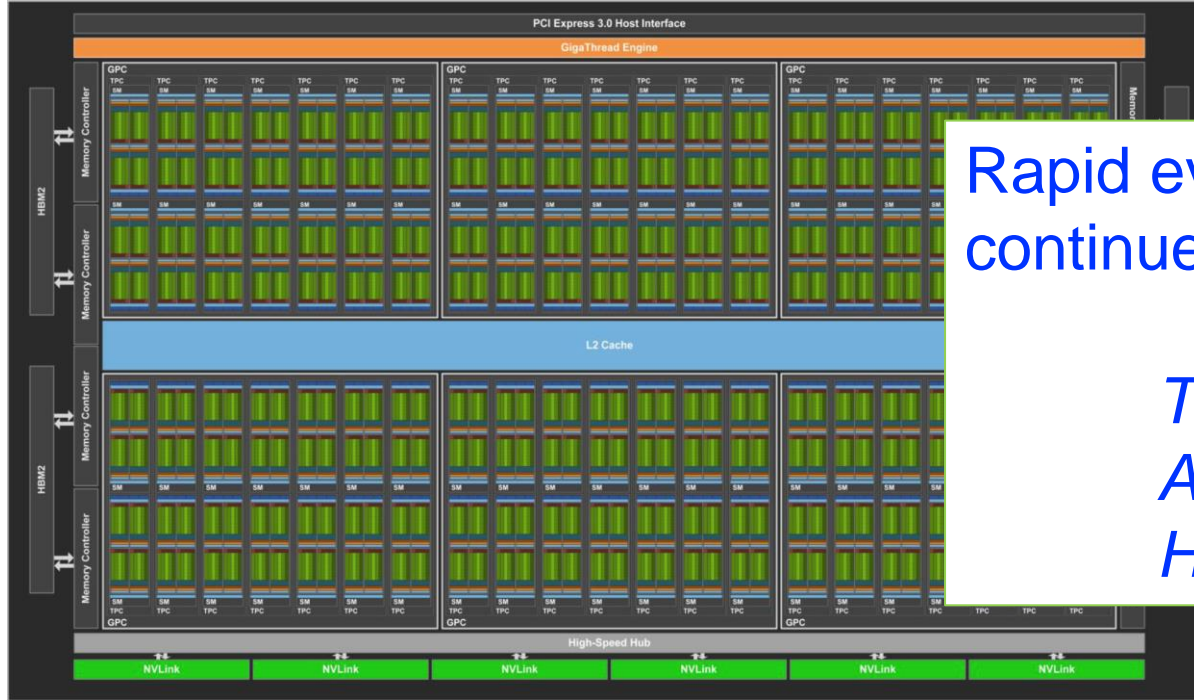
```
!$omp parallel do
do i = 1, n
        a(i) = b(i) + c(i)
enddo
```

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*

Huang's Law

An observation/claim made by Jensen Huang, CEO of Nvidia, at its 2018 GPU Technology Conference.

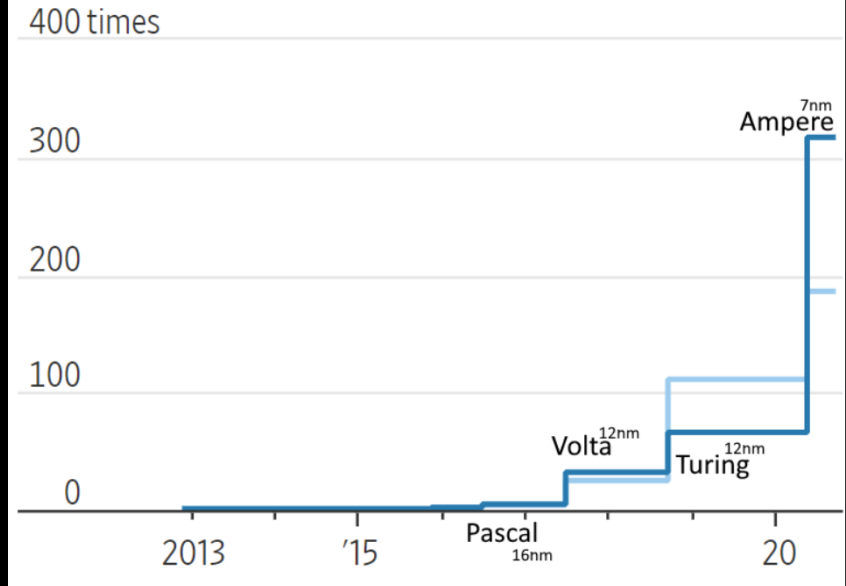
He observed that Nvidia's GPUs were "25 times faster than five years ago" whereas Moore's law would have expected only a ten-fold increase.

In 2006 Nvidia's GPU had a 4x performance advantage over other CPUs. In 2018 the Nvidia GPU was 20 times faster than a comparable CPU node: the GPUs were 1.7x faster each year. Moore's law would predict a doubling every two years, however Nvidia's GPU performance was more than tripled every two years fulfilling Huang's law.

It is a little premature, and there are confounding factors at play, so most people haven't yet elevated this to the status of Moore's Law.

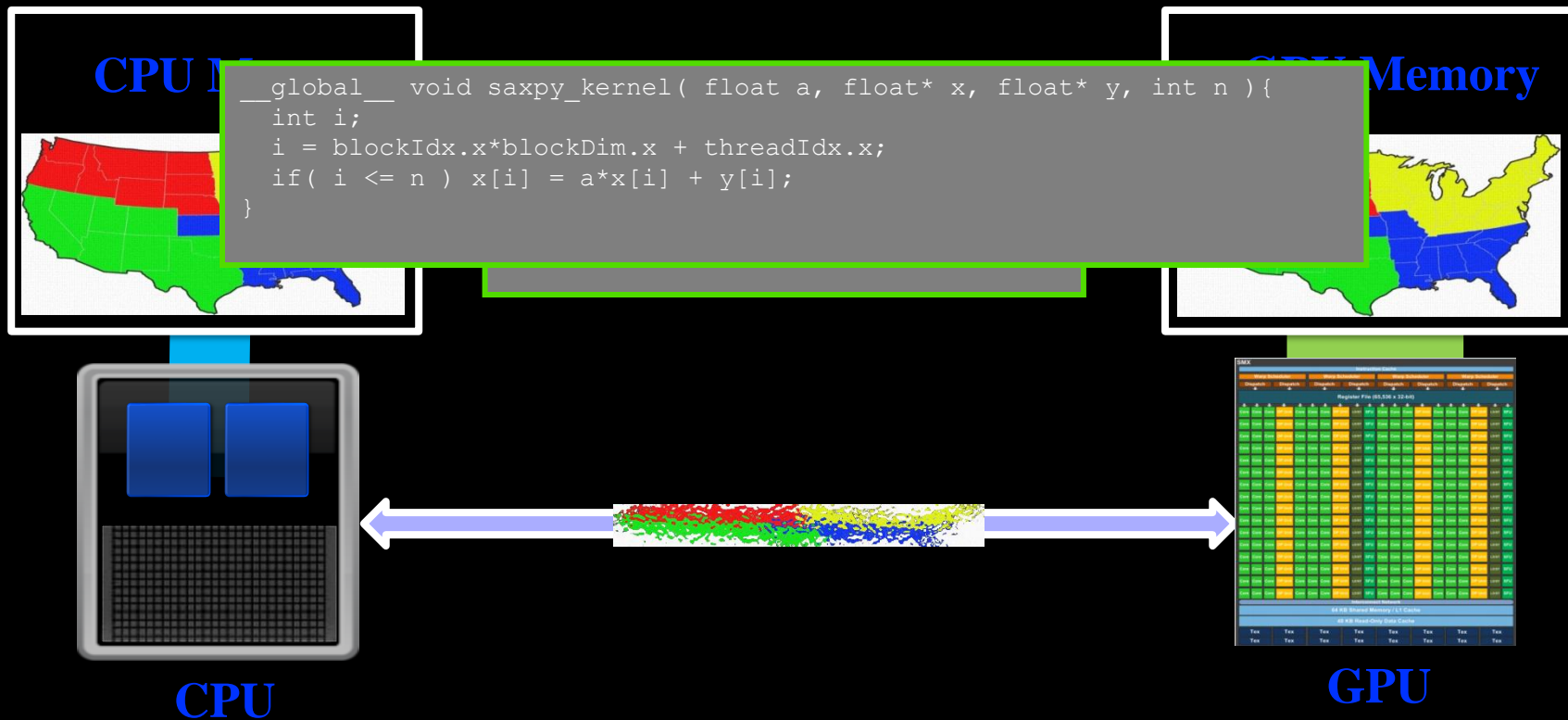
Speed and energy efficiency of Nvidia's chips as a multiple of performance in 2012

- Operations per second
- Operations per second per watt



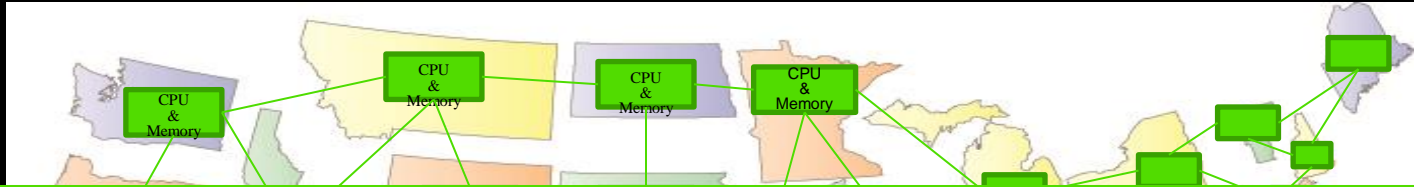
Source: NVIDIA

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)
```

▪
▪
▪

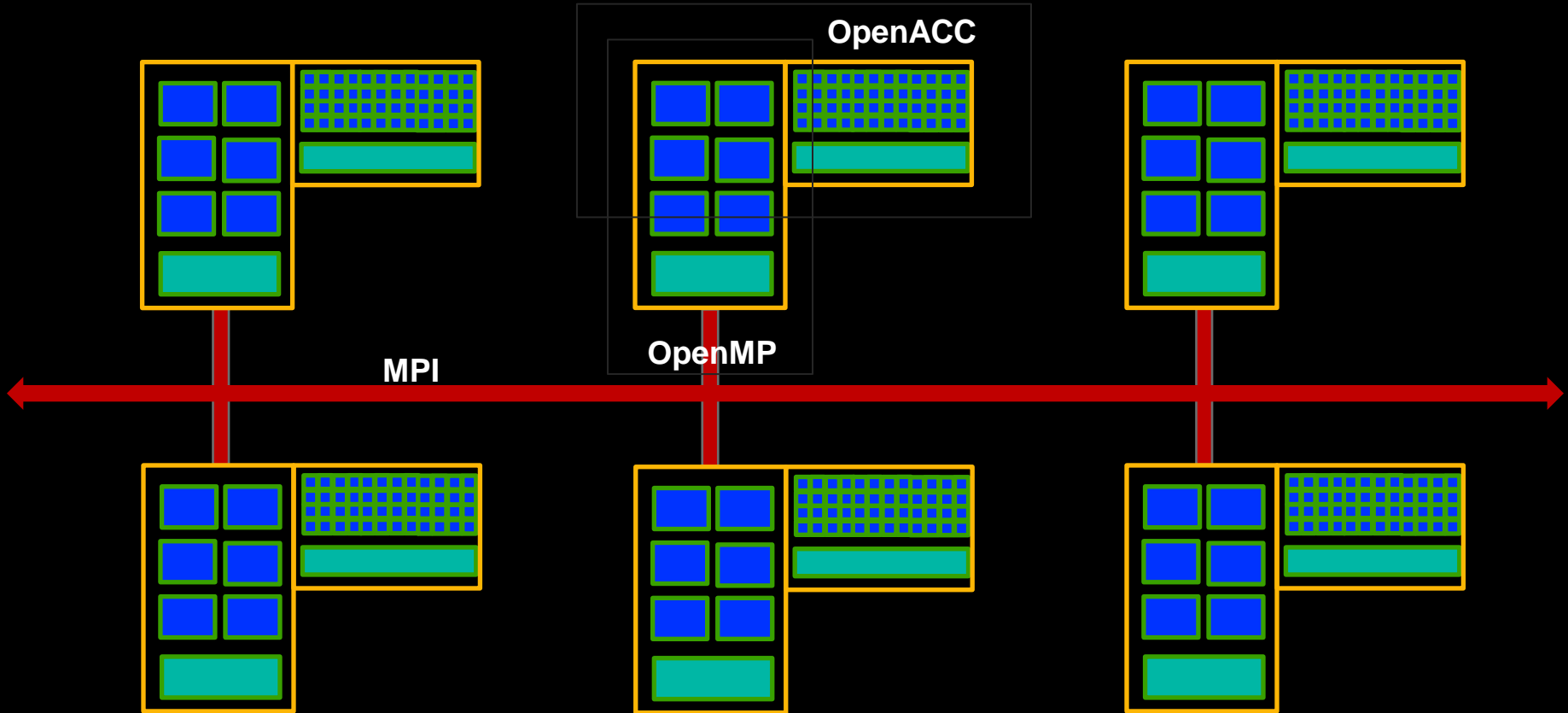
```
call MPI_Barrier(MPI_COMM_WORLD, errcode)
```

▪



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

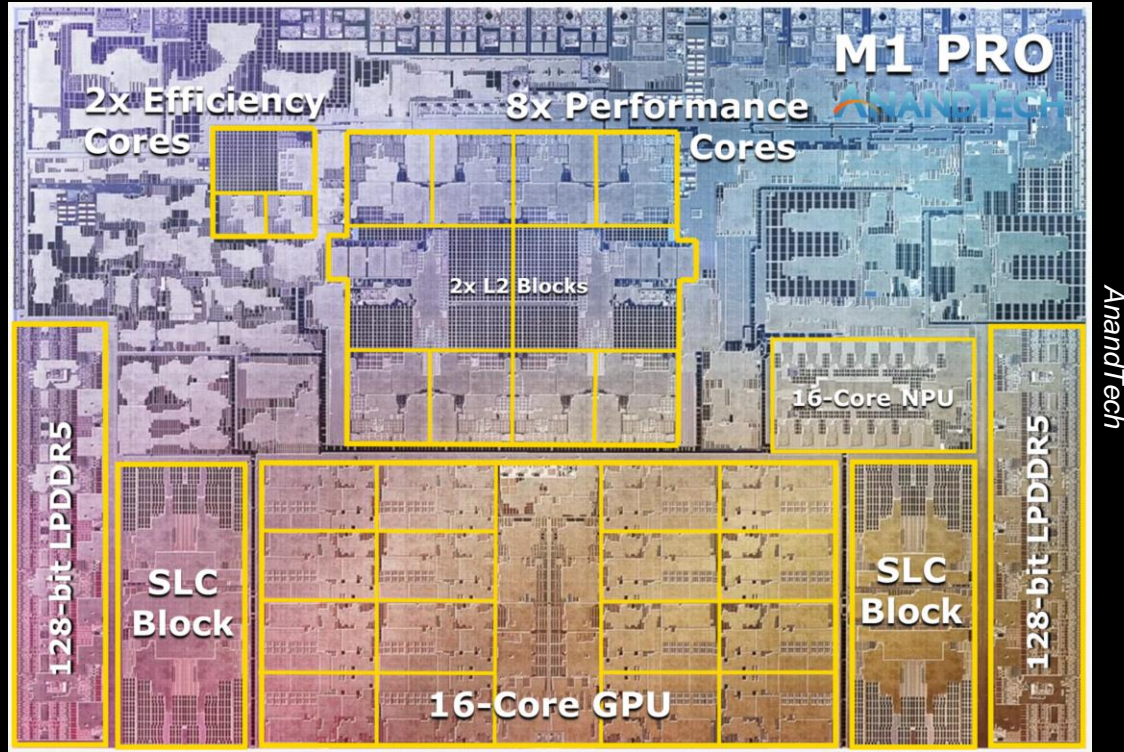
Also Important

- ASIC/FPGA/DSP
- RAID/IO

Top 10 Systems as of November 2024								
#	Computer	Site	Manufacturer	CPU Interconnect [Accelerator]	Cores	Rmax (Pflops)	Rpeak (Pflops)	Power (MW)
1	El Capitan	Lawrence Livermore National Laboratory United States	HPE	AMD EPYC 24C 1.8GHz Slingshot-11 AMD Instinct MI300A	11,039,616	1742	2746	30
2	Frontier	Oak Ridge National Laboratory United States	HPE	AMD EPYC 64C 2GHz Slingshot-11 AMD Instinct MI250X	9,066,176	1353	2055	25
3	Aurora	Argonne National Laboratory United States	HPE	Intel Xeon Max 9470 52C 2.4GHz Slingshot-11 Intel Data Center GPU Max	9,264,128	1012	1980	39
4	Eagle	Microsoft United States	Microsoft	Intel Xeon 8480C 48C 2GHz Infiniband NDR NVIDIA H100	1,123,200	561	846	
5	HPC6	Eni S.p.A. Italy	HPE	AMD EPYC 64C 2GHz Slingshot-11 AMD Instinct MI250X	3,143,520	477	606	8
6	Fugaku	RIKEN Center for Computational Science Japan	Fujitsu	ARM 8.2A+ 48C 2.2GHz Torus Fusion Interconnect	7,630,072	442	537	29
7	Alps	Swiss National Supercomputing Center Switzerland	HPE	NVIDIA Grace 72C 3.1GHz Slingshot-11 NVIDIA GH200	2,121,600	434	574	7
8	LUMI	EuroHPC Finland	HPE	AMD EPYC 64C 2GHz Slingshot-11 AMD Instinct MI250X	2,752,704	379	531	7
9	Leonardo	EuroHPC Italy	500 ThinkSystem SR590, Xeon Gold 5218 16C 2.3GHz, 10G Ethernet, Lenovo Service Provider T	108,800	2.31	4.00	304	7
10	Tuolumne	Lawrence Livermore National Laboratory United States	China				388	2

The word is *Heterogeneous*

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



How much of this can you program?

Networks

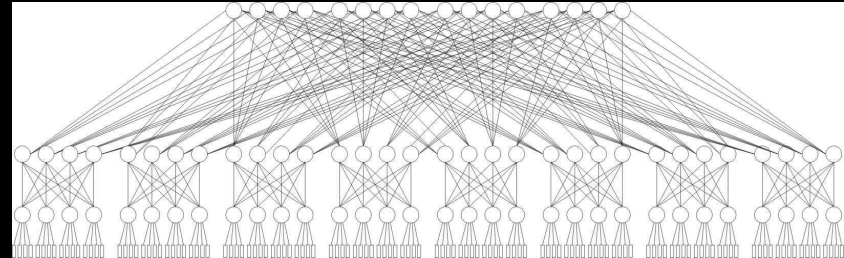
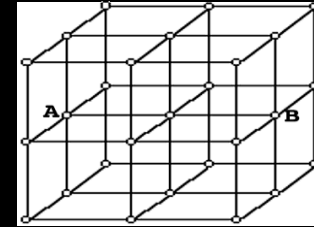
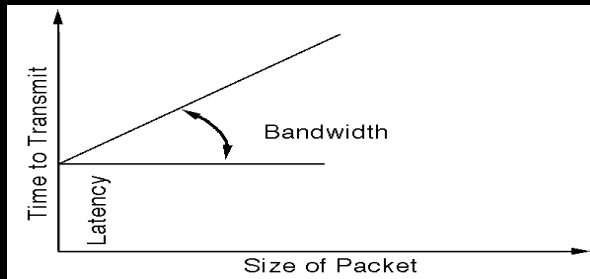
3 characteristics sum up the network:

- **Latency**

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

- **Bandwidth**

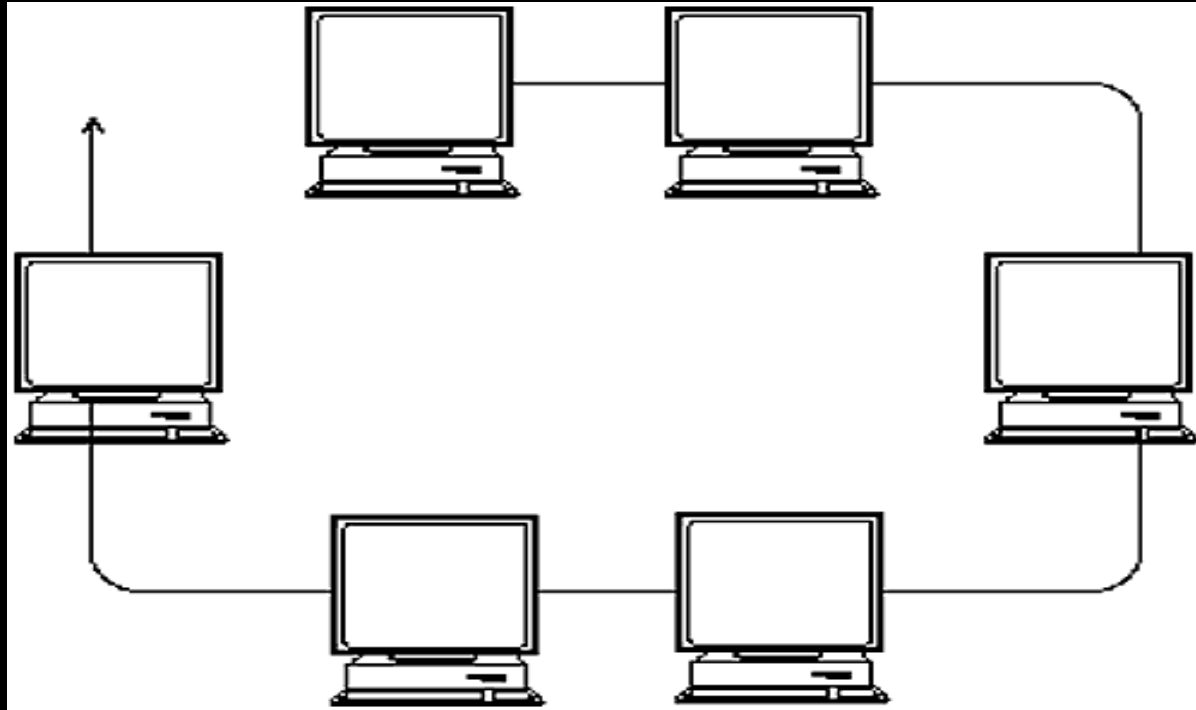
The rate at which a very large packet of information can be sent



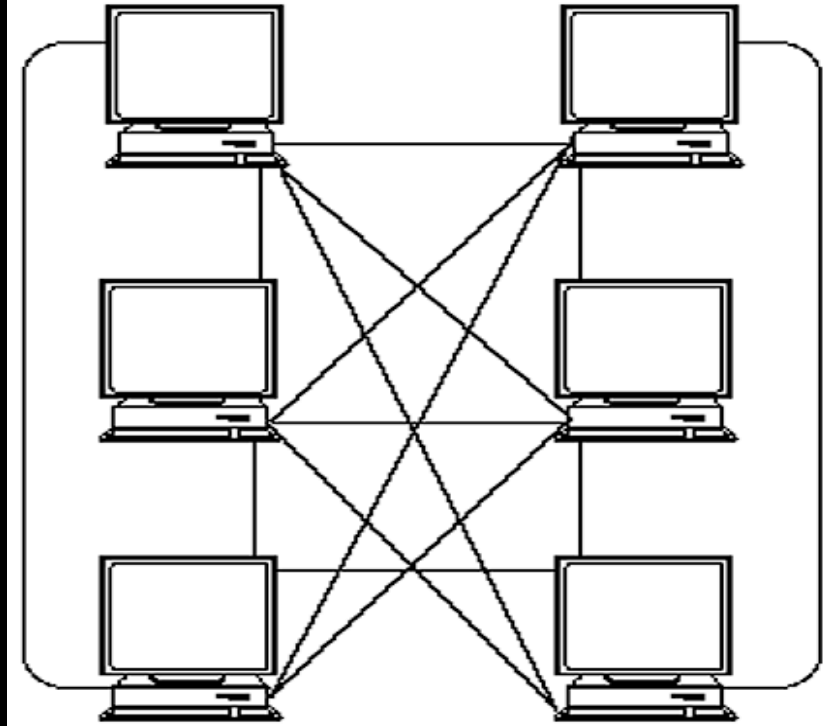
- **Topology**

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

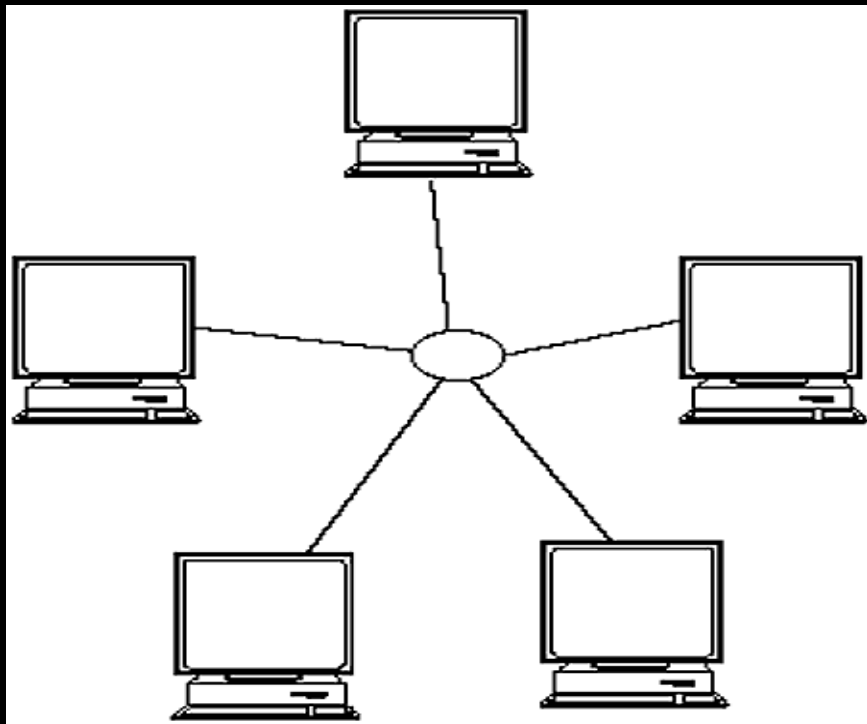
Ethernet with Workstations



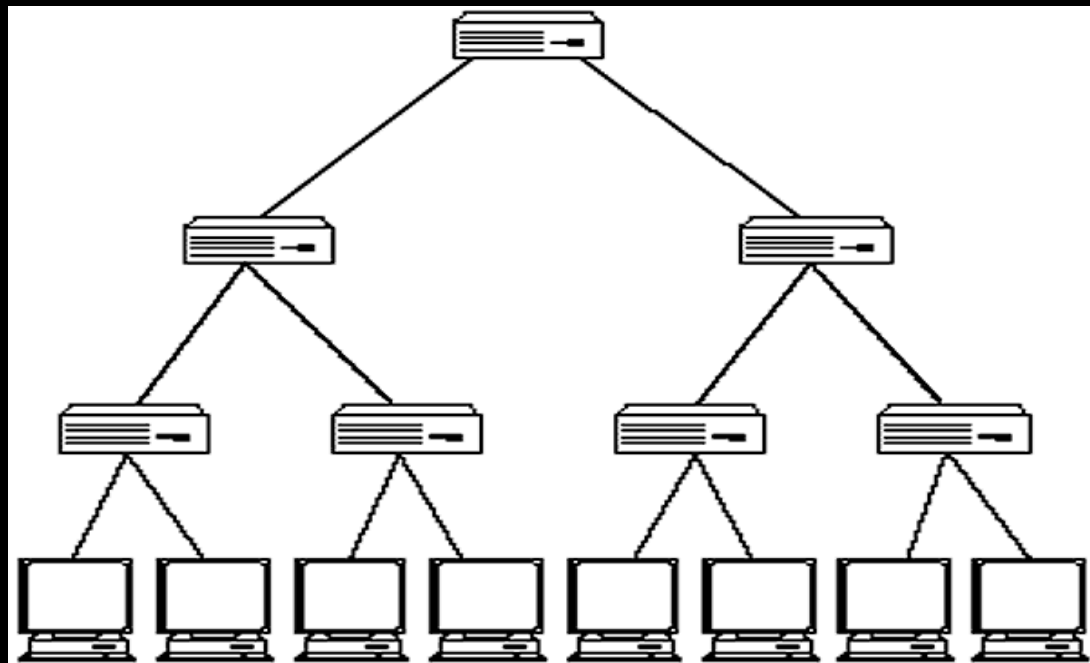
Complete Connectivity



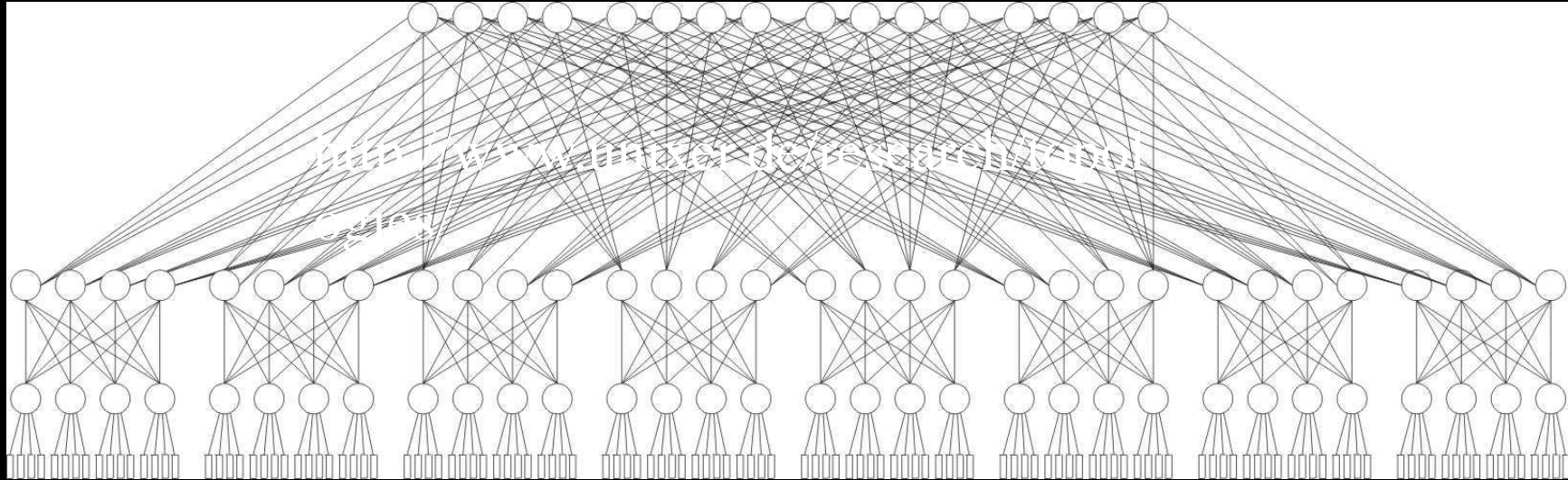
Crossbar



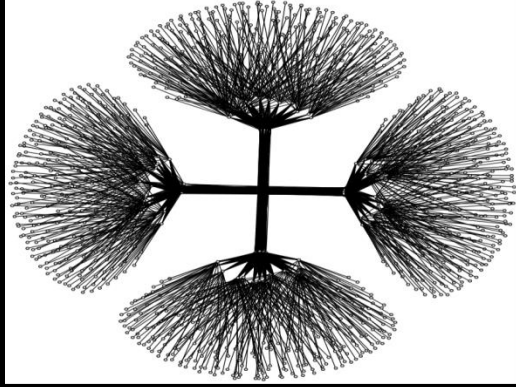
Binary Tree



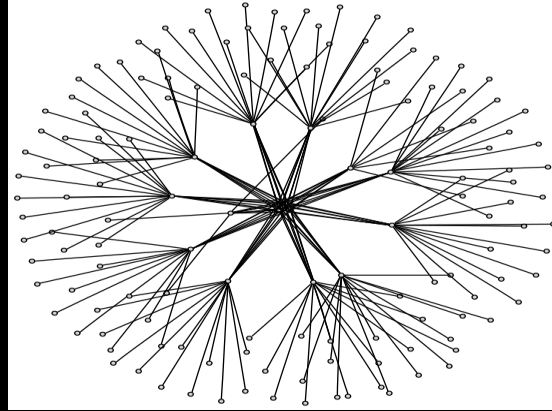
Fat Tree



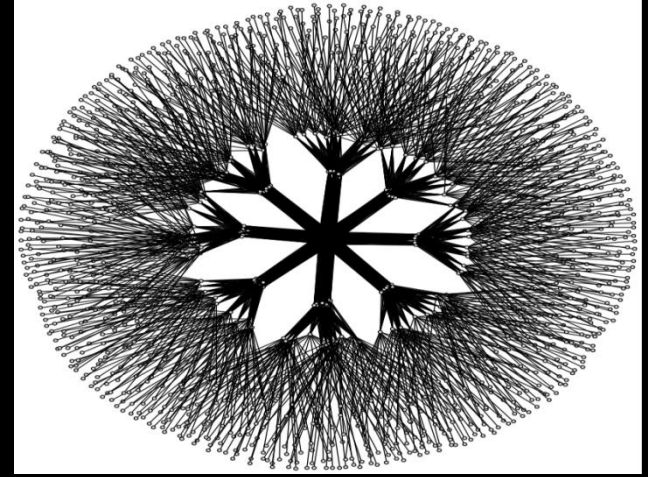
Other Fat Trees



Big Red @ IU

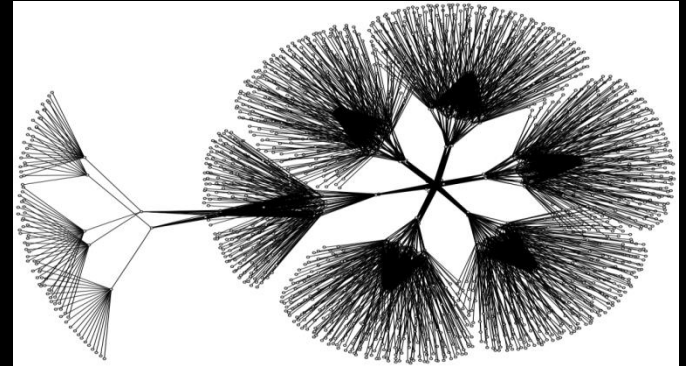
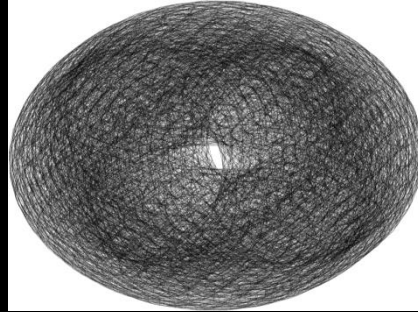


Odin @ IU



Atlas @ LLNL

Jaguar @ ORNL



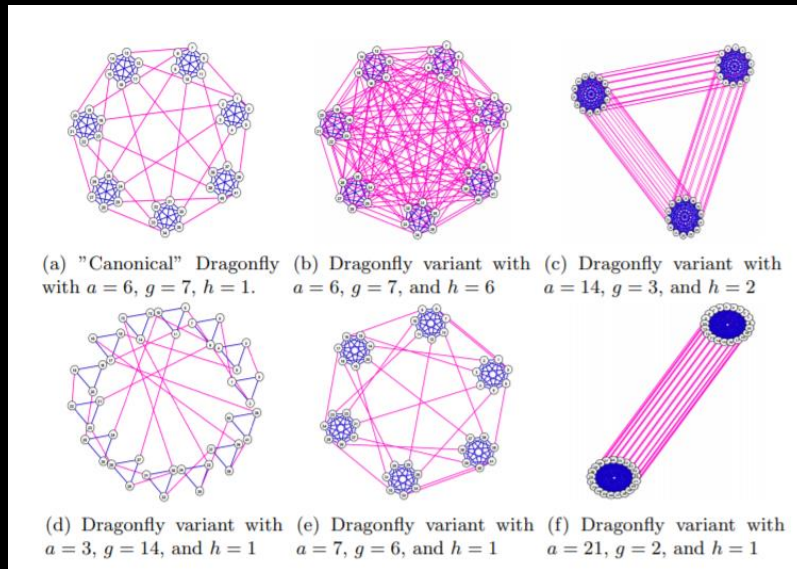
Tsubame @ Tokyo Inst. of Tech

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

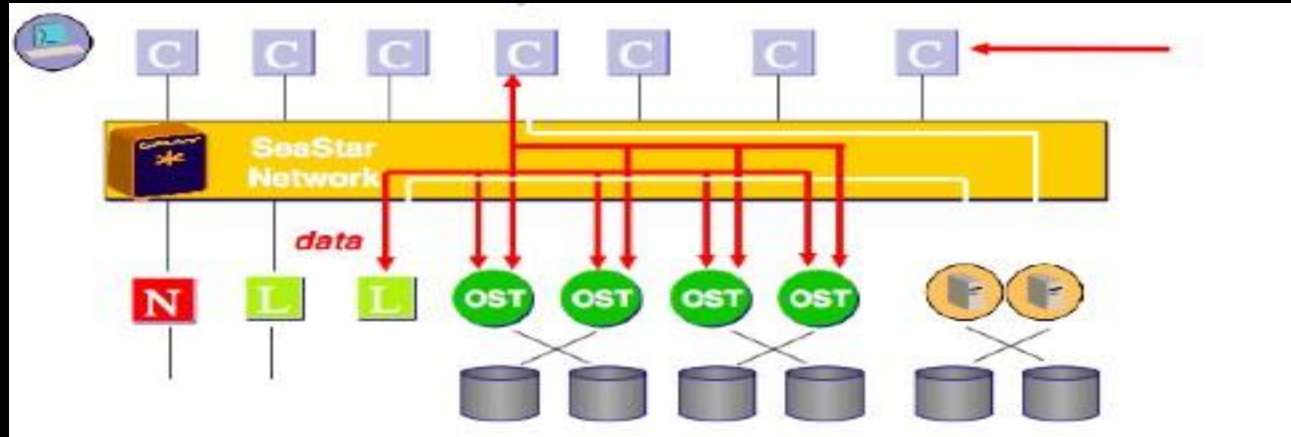
Various 42 node Dragonfly configurations.



Purple links are optical, and blue are electrical.

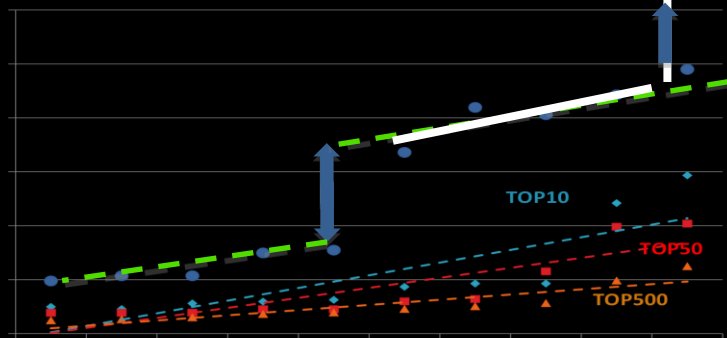
Parallel IO (RAID...)

- 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.



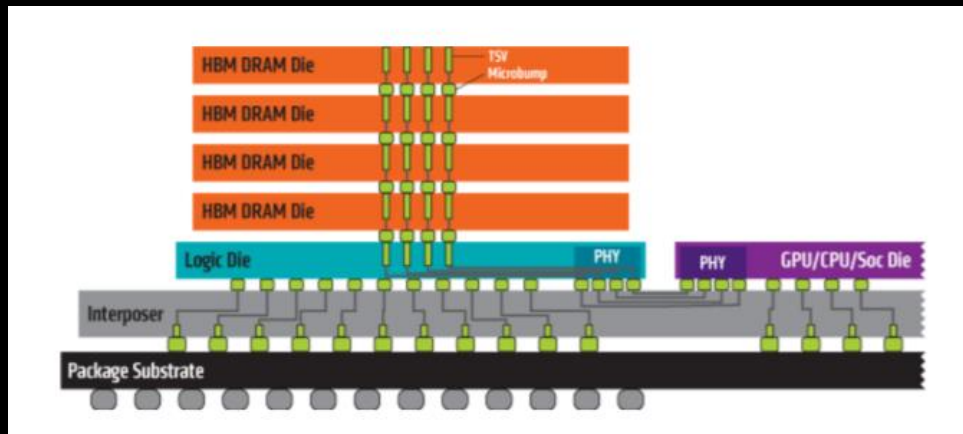
The path to Exascale has not been incremental.

First boost: many-core/accelerator



Second Boost: 3D (2016 – 2023)

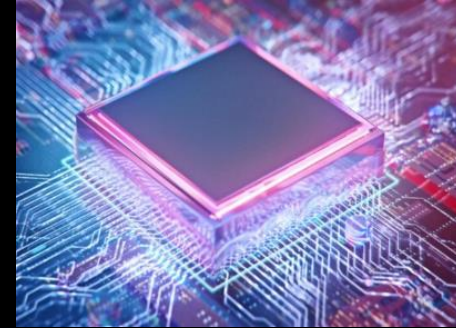
Third Boost: SiPh (2021–)



Is Silicon Photonics a game changer?

Electrically switched networks can operate in “packet switching” mode to lower the effective latency and utilize all the available link bandwidth. The alternative to this mode is “circuit-switching” and it was abandoned by the electronic community long ago. Without practical means to buffer light, process photon headers in-flight, or reverting to switches with expensive optical-electrical-optical conversions, we would have to resort to circuit-switching with all the inherent deficiencies:

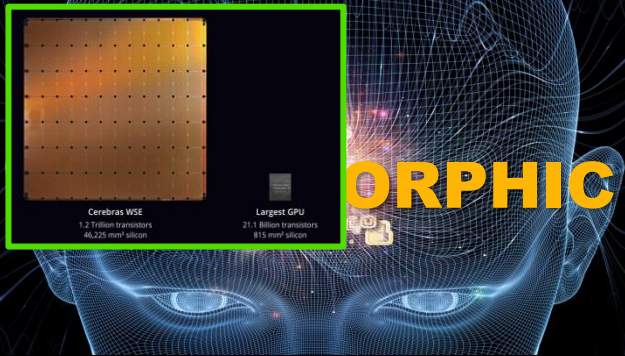
- complex traffic steering calculations
- switching delays
- latency increase due to lack of available paths
- under-utilization of links



Photonics is often cited as an enabler for extensive memory disaggregation, but this yields another challenge, specifically the speed of light. Photons travel at a maximum speed of 3.3 ns/m in fibers. This is equivalent to a level-2 cache access of a modern CPU, not including the disaggregation overhead (such as from the protocol, switching, or optical-electrical conversions at the endpoints). At 3–4 m distance, the photon travel time alone exceeds the first-word access latency of modern DDR memory.

End of Moore's Law Will Lead to New Architectures

Non-von
Neumann



Reversible Computing
Analog computing

**QUANTUM
COMPUTING**



ARCHITECTURE

von Neumann

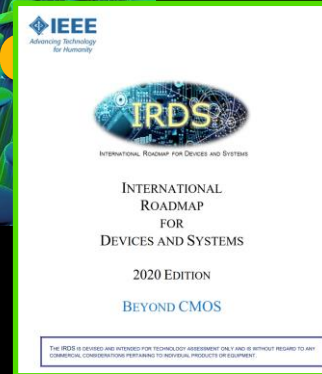


CMOS

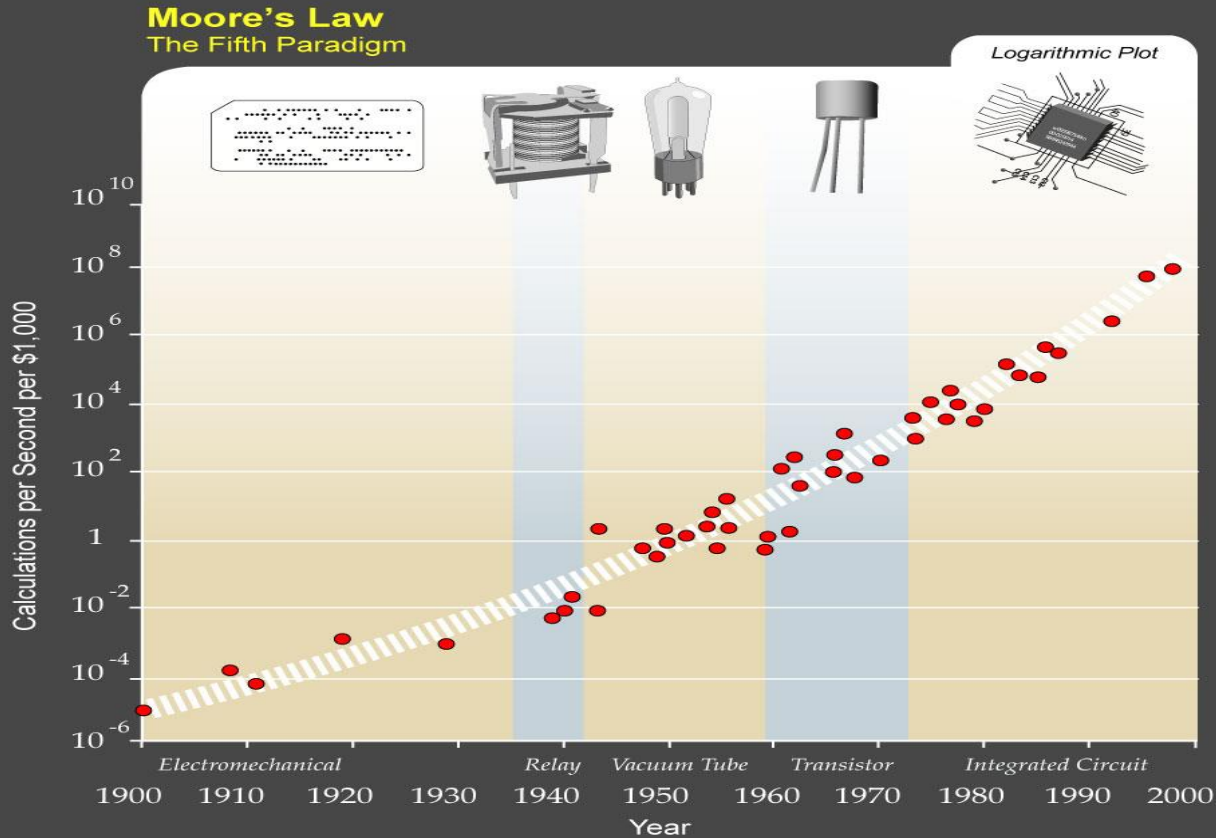


Beyond CMOS

TECHNOLOGY



It would only be the 6th paradigm.



We can do better. We have a role model.

- We hope to "simulate" a human brain in real time on one of these Exascale platforms with about 1 - 10 Exaflop/s and 4 PB of memory
- These newest Exascale computers use 20+ MW
- The human brain runs at 20W
- Our brain is a million times more power efficient!



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...

