Performance & Supercomputers

CPE380, Spring 2023

Hank Dietz

http://aggregate.org/hankd/

University of Kentucky

Performance is about Choices

Airplane	Passengers	Range (mi)	Speed (mph)
Boeing 737-100	101	630	598
Boeing 747	470	4150	610
BAC/Sud Concorde	132	4000	1350
Douglas DC-8-50	146	8720	544

- Execution time for application
- Power / temperature / battery life
- Reliability / availability
- Cost for acceptable functionality
- Size

Response Time vs. Throughput

- Often can trade one for the other:
  Response Time: Time to complete an operation
  Throughput: Jobs completed per unit time

- Performance(X) = 1/ExecutionTime(X)
- X is Performance(X)/Performance(Y) times faster than Y, also: ExecutionTime(Y)/ExecutionTime(X)

For Whom The Clock Ticks

- Posix uses real, user, system time
  - Real "Wall Clock" time always ticks
  - User time while in your code
  - System time while in OS code for you
- Multiplied by #PEs in multiprocessors
- I/O time not reported under Posix
- There are really lots of timing components
  - Processor tick count register
  - OS scheduler in 1-10ms Jiffies

Running What?

- Different program, different performance
- Application (all that really matters!)
- “Toy” program
- Benchmark: representative application
- Micro Benchmark: tests a certain feature
- Synthetic Benchmark: a program written solely to perform like a particular application, but doing nothing useful
- Benchmark Suite: multiple benchmarks

Modeling Time: CPI and IPC

- CPI is clock Cycles / Instruction
- IPC is Instructions / Cycle; i.e., 1/CPI
- Program runtime is:
  (Instructions executed / Program) *
  CPI * (Clock Period)
- Really sum over all instruction types because different instructions have different CPI

An Example

Instruction Type	Execution Count	CPI	Clock Period
A	20	10	10ns
B	10	30	10ns

- This program takes:
  ((20*10)+(10*30))*10ns = 5us
- What can be changed to make it 4us?

What Effects What?

	Instruction Count	CPI	Clock Rate
Program (Algorithm)	Yes!	Indirectly...	No!
Compiler	Yes!	Indirectly...	Power?
ISA	Yes!	Yes!	Indirectly...
Impl. Arch.	uOps?	Yes!	Yes!
VLSI	No!	Indirectly...	Yes!

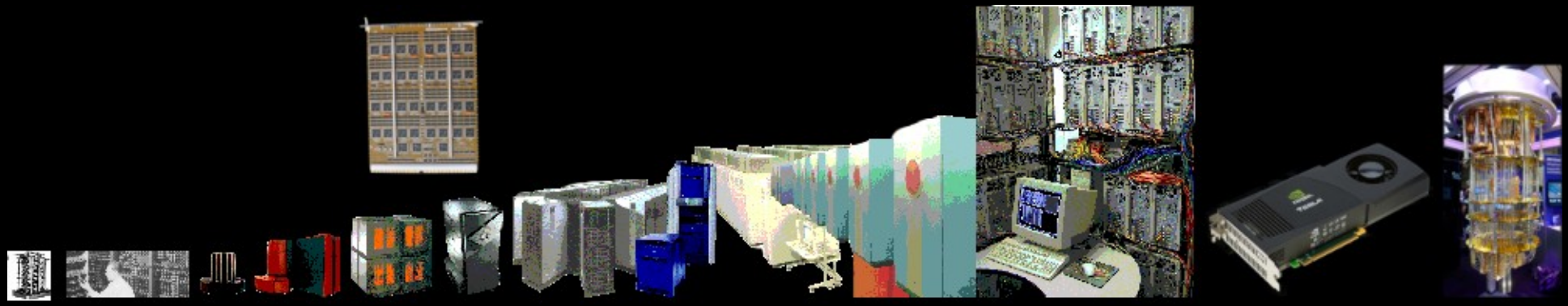
Amdahl's Law

- If 1/N time is not affected by a change, the best possible speedup is only N
- Originally for sequential overhead in parallel code, but applies for any change

Suppose a program spends 80% of its time doing multiplies... you can't get more than a 5X speedup by improving only multiplies!

What Is A Supercomputer?

- Really two key characteristics:
  - Computer that solves big problems…
  - Performance can scale…
- The key is Parallel Processing…
  and modularity brings availability & reliability



The Key Is Parallel Processing

- Process *N* "pieces" simultaneously, get up to factor of *N* speedup
- Modular hardware designs:
  - Relatively easy to scale – add modules
  - Higher availability (if not reliability)

Clusters And Bigger

- Mostly from interchangeable (PC) parts…
  and mostly running some form of Linux
- Cluster or Beowulf is a *parallel supercomputer*
  with tightly coupled, homogeneous, nodes
- Farm is homogeneous, colocated, machines
  with a common purpose (e.g., a render farm)
- Warehouse Scale Computer is a warehouse full
  of racked clusters used for *throughput*
- Grid is many internet-connected machines
- Cloud is *virtualized* grid/WSCs providing *services*

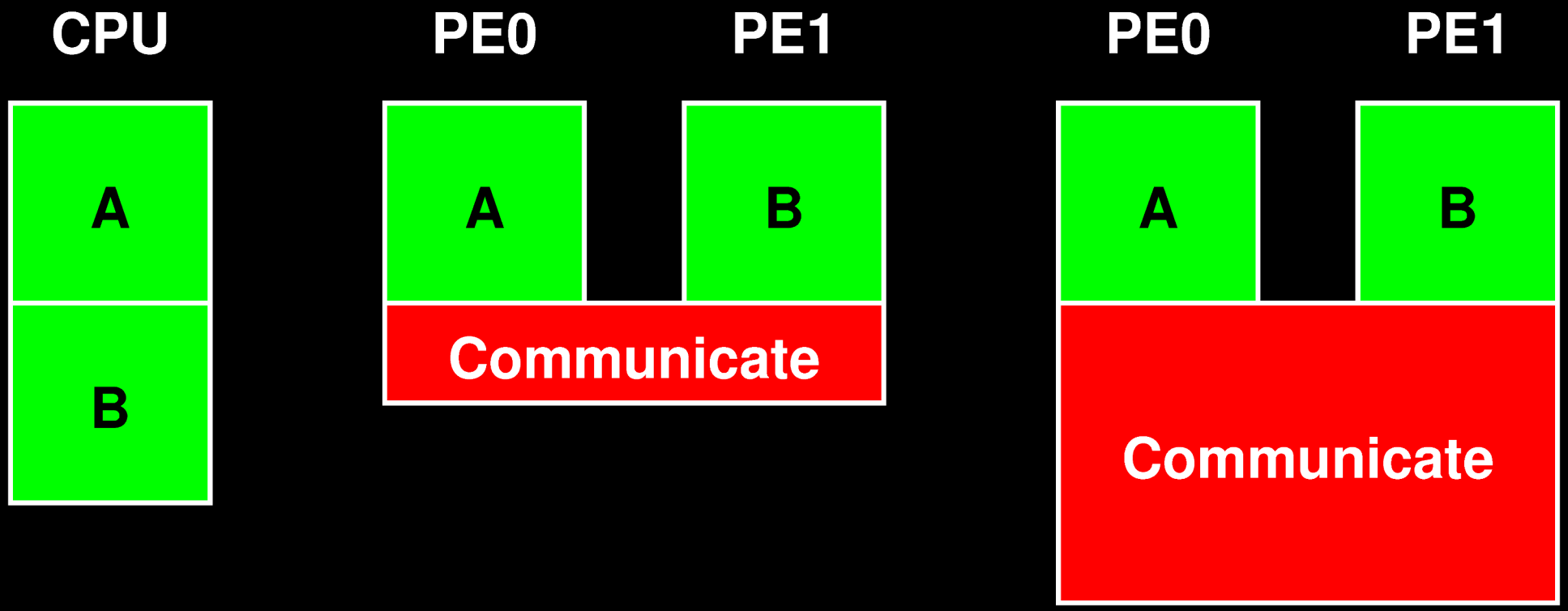
Types Of Hardware Parallelism

- Pipeline
- Superscalar, VLIW, EPIC
- SWAR (SIMD Within A Register)
- SMP (Symmetric MultiProcessor; multi-core)
- GPU (Graphics Processing Unit)
- Cluster
- Farm / Warehouse Scale Computer
- Grid / Cloud

Engineering an Interconnection Network

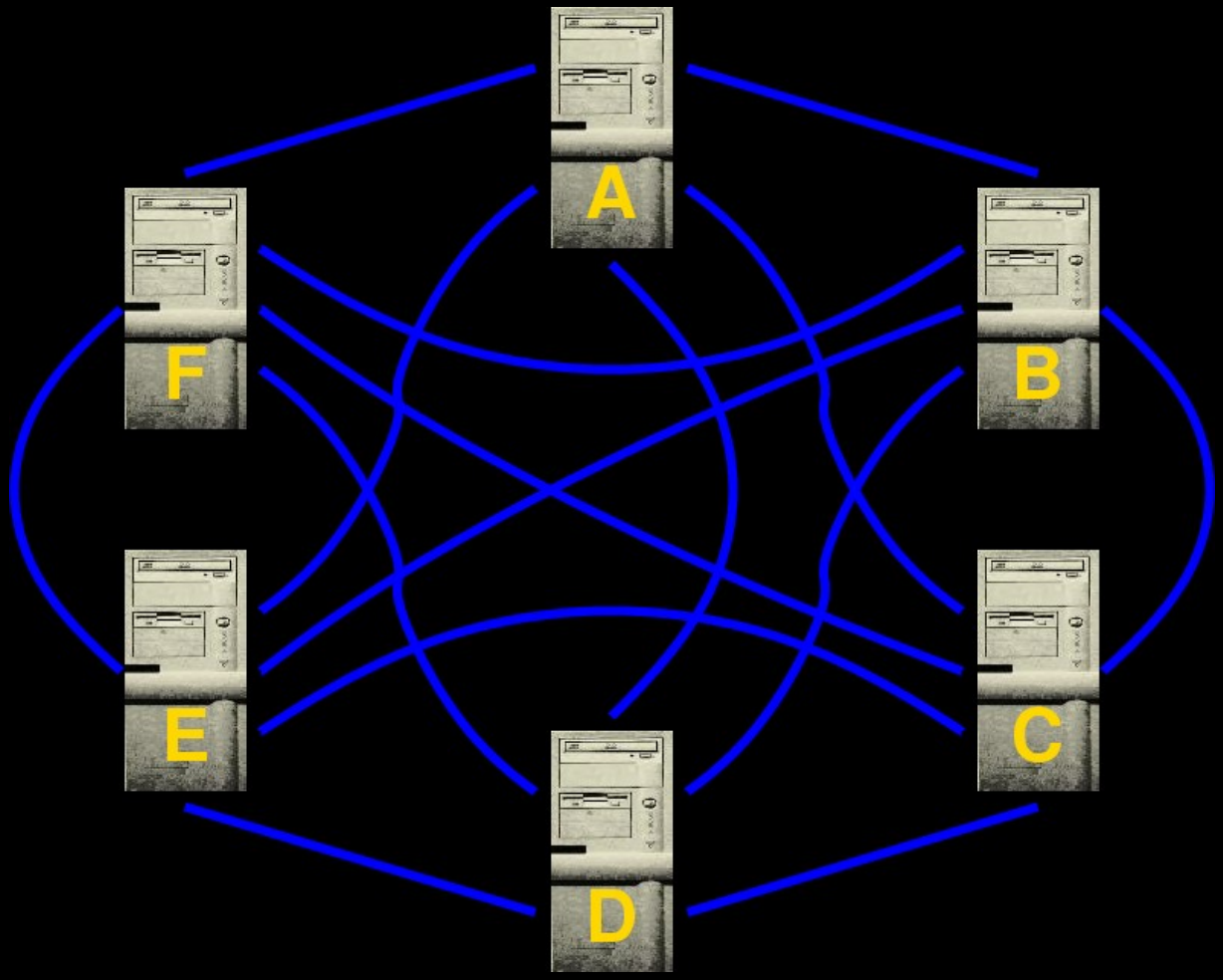
- Parallel supercomputer **nodes** interact
- **Bandwidth**
  - Bits transmitted per second
  - **Bisection Bandwidth** most important
- **Latency**
  - Time to send something here to there
  - Harder to improve than bandwidth
- We'll just consider topology here...

Latency Determines Smallest Useful Parallel Grain Size

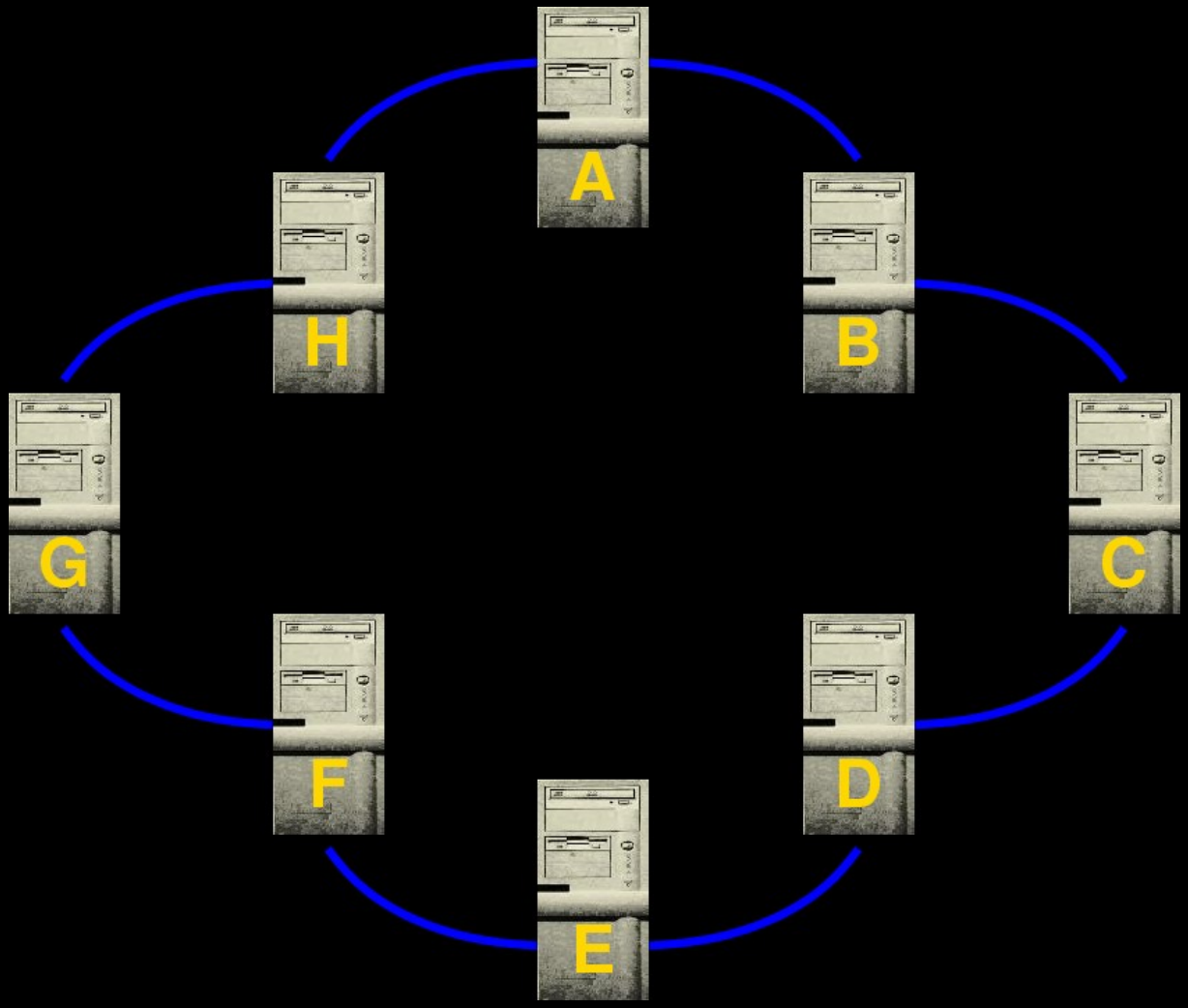


No Network

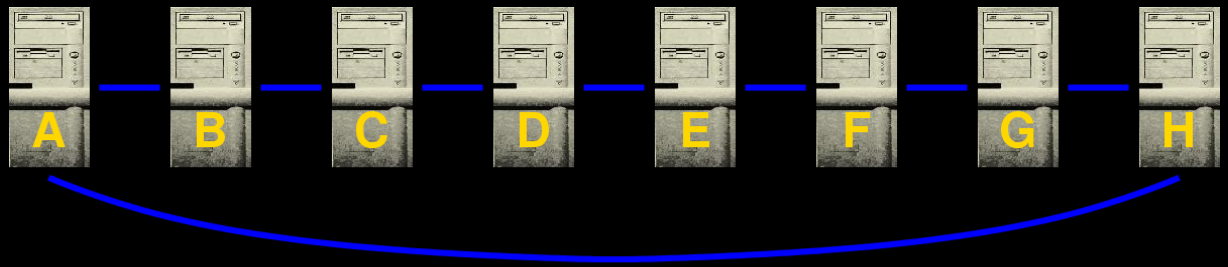
Direct Fully Connected

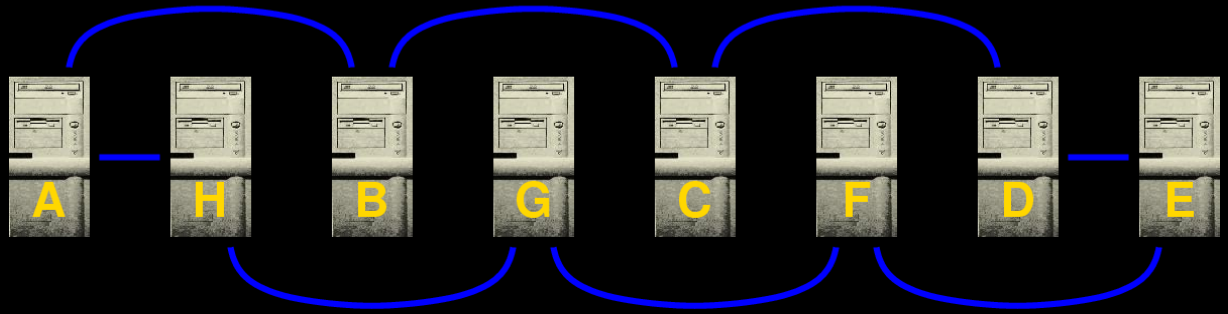


Toroidal 1D Mesh (Ring)

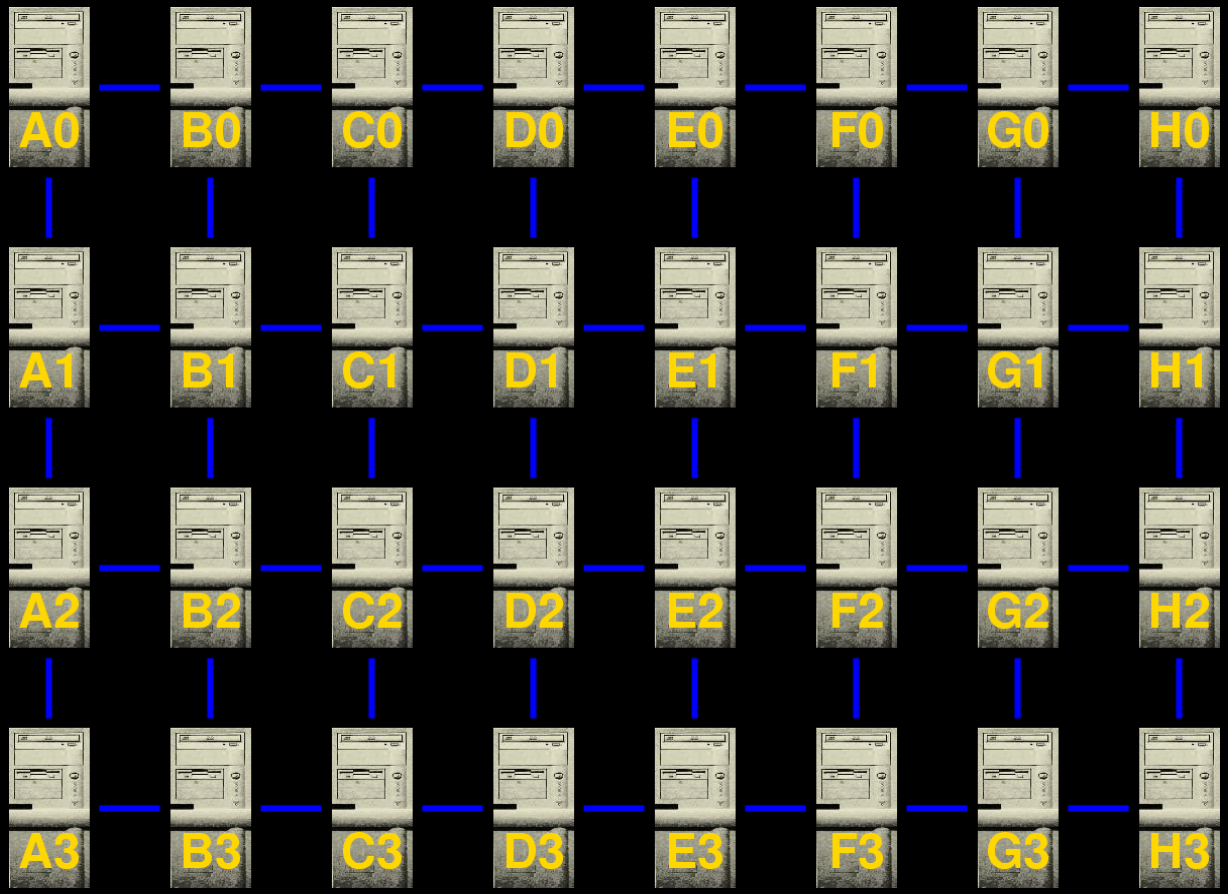


Physical Layout Of Ring

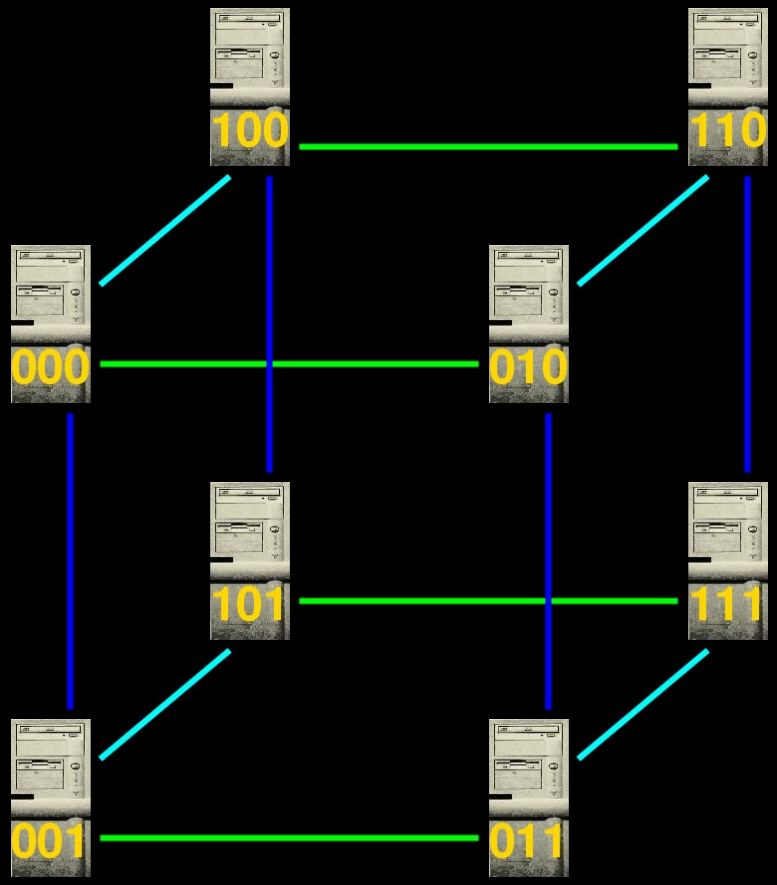




Non-Toroidal 2D Mesh



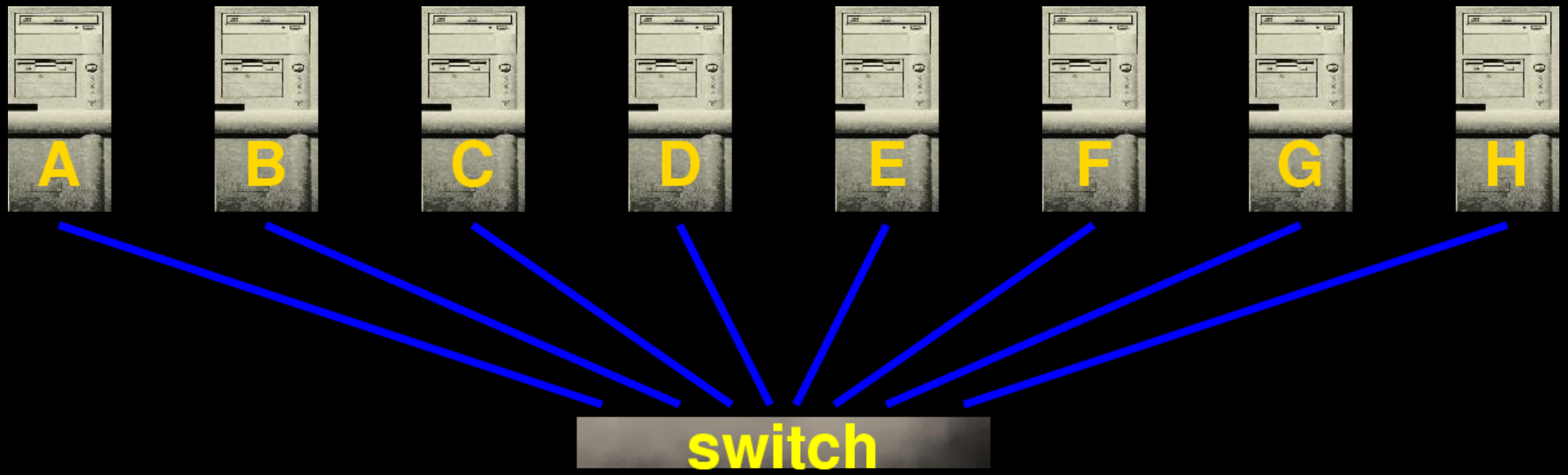
3-Cube (AKA 3D Mesh)



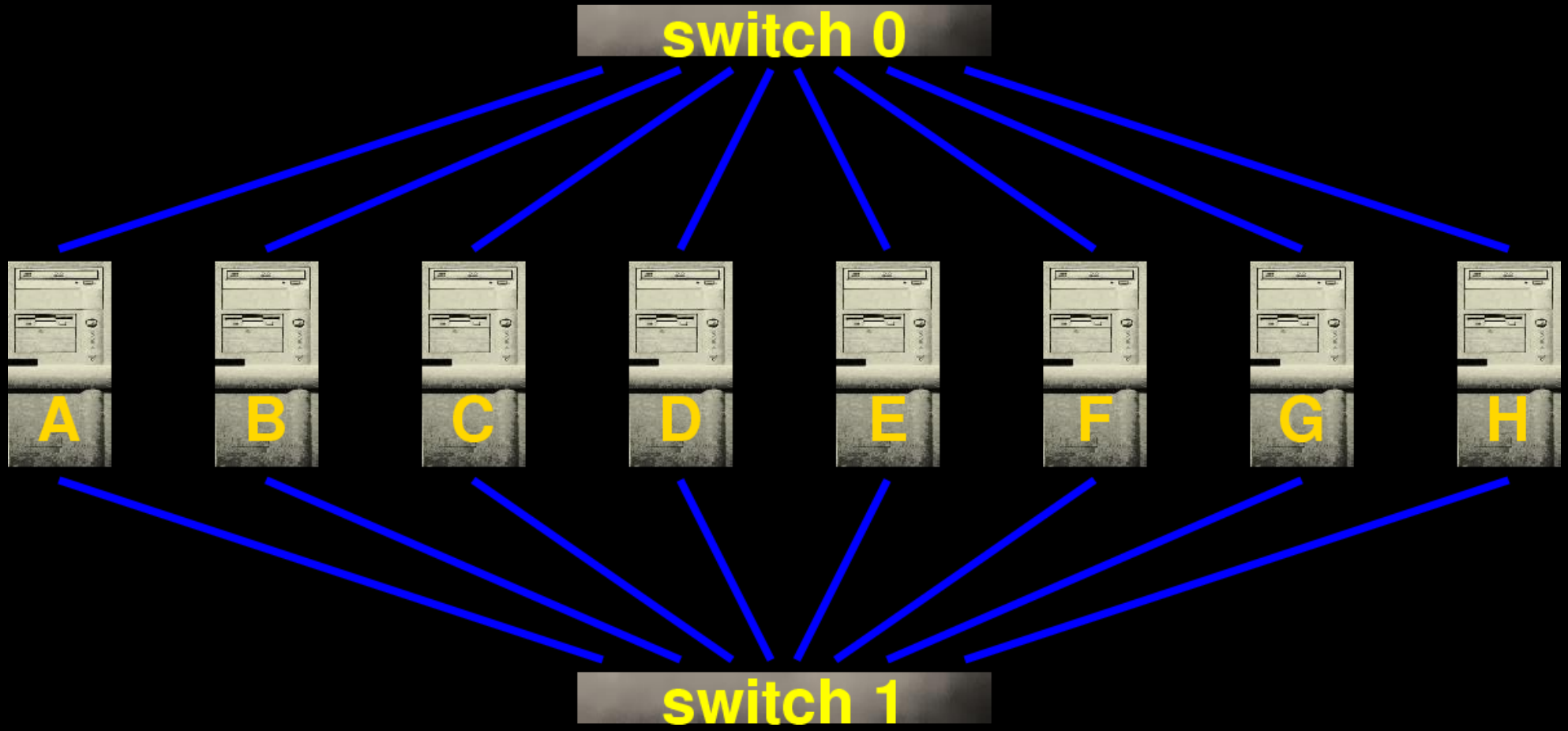
Switch Networks

- Ideal switch connects *N* things such that:
  - Bisection bandwidth = # ports
  - Latency is low (~30us for Ethernet)
- Other switch-like units:
  - Hubs, FDRs (Full Duplex Repeaters)
  - Managed Switches, Routers
- Not enough ports, build a Switch Fabric

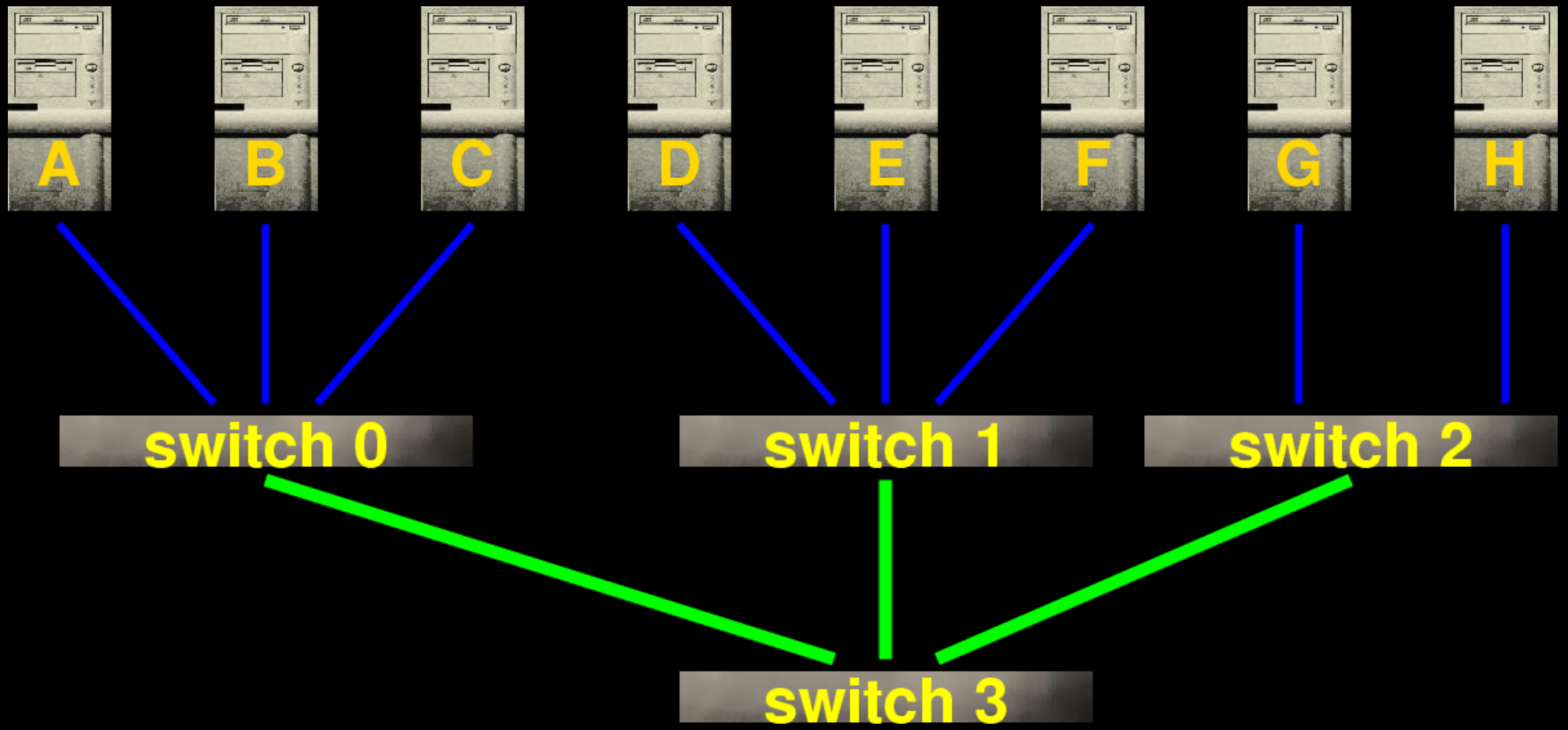
Simple Switch (8-Port)



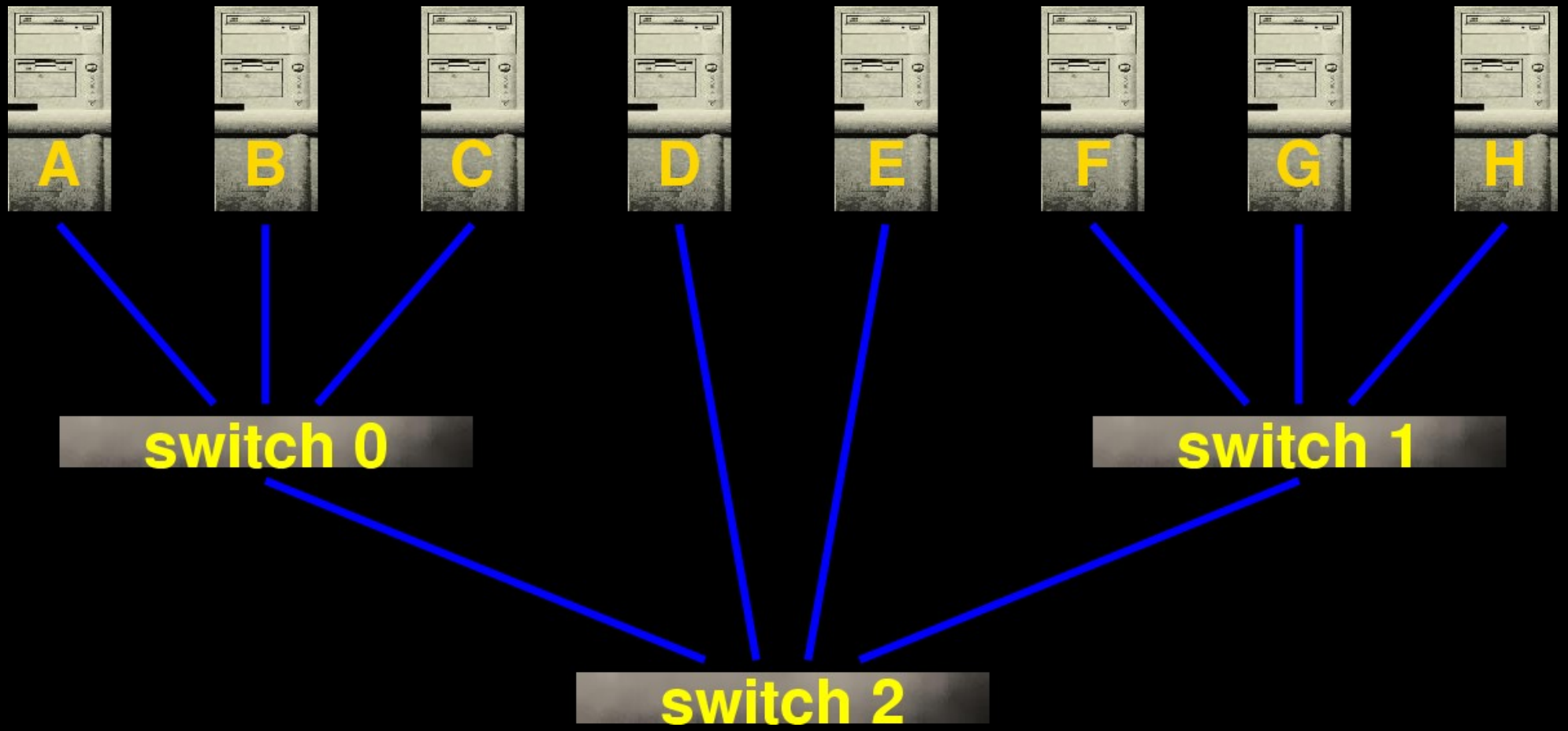
2-Way Channel Bonding



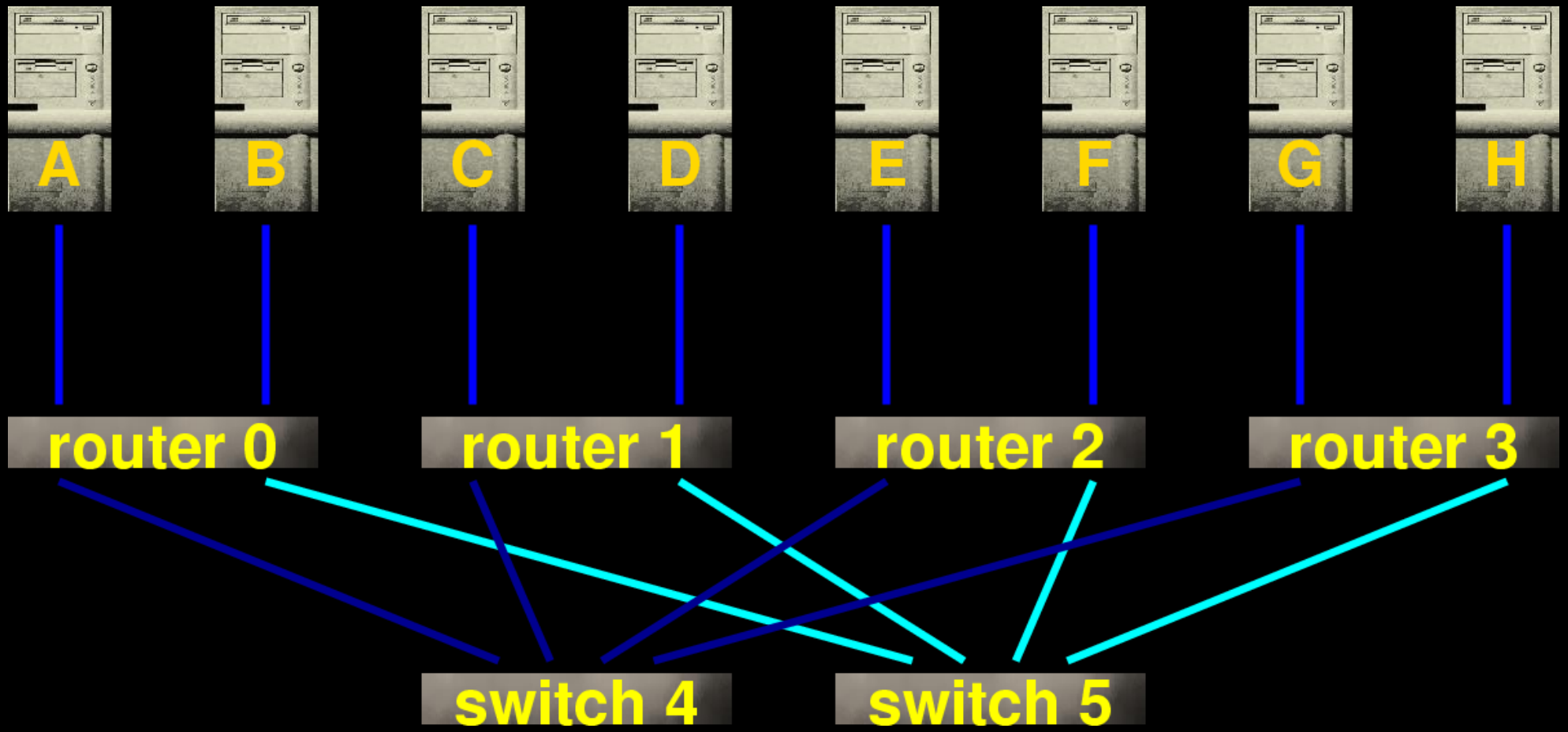
Tree (4-Port Switches)



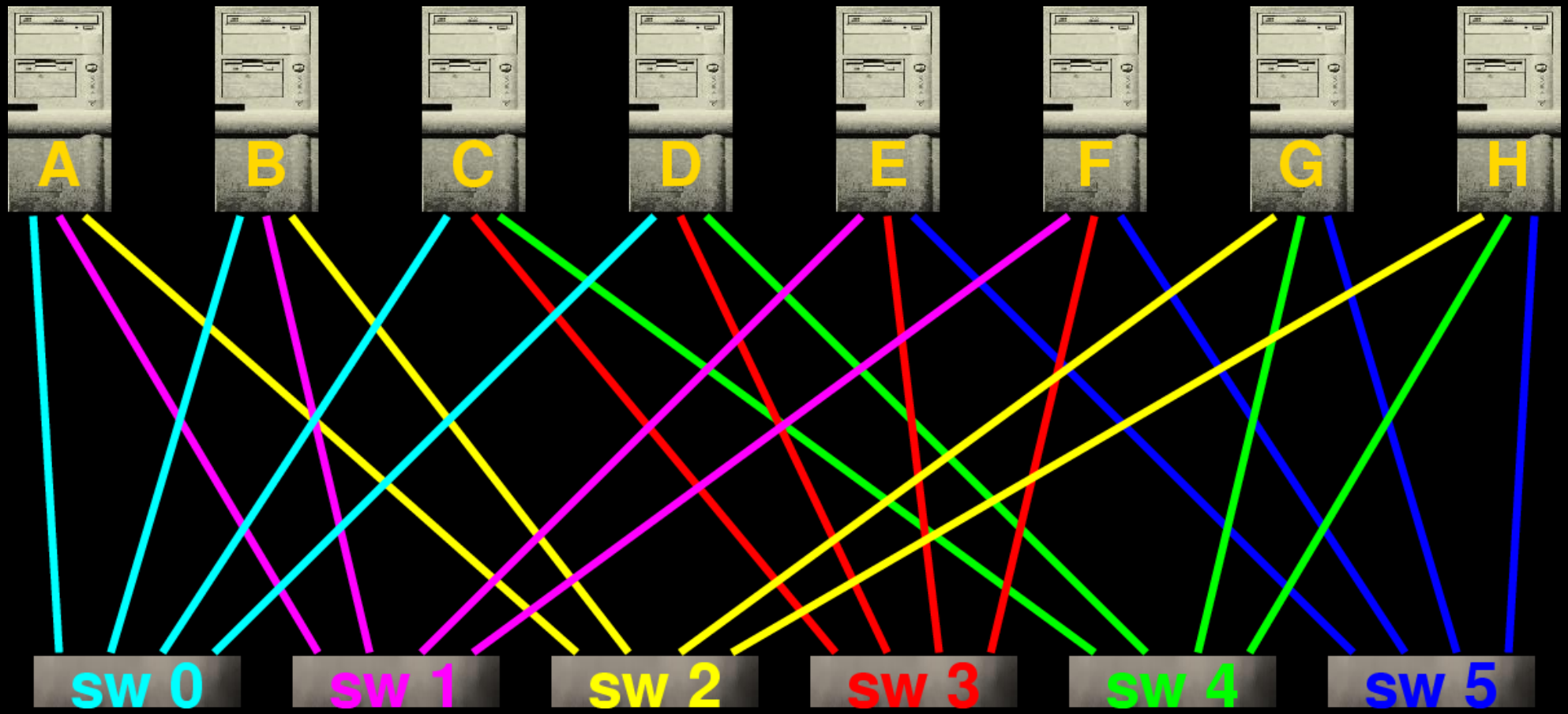
A Better Tree



Fat Tree


A
B
C
D
E
F
G
H
router 0
router 1
router 2
router 3
switch 4
switch 5

Flat Neighborhood Network...
from UK!

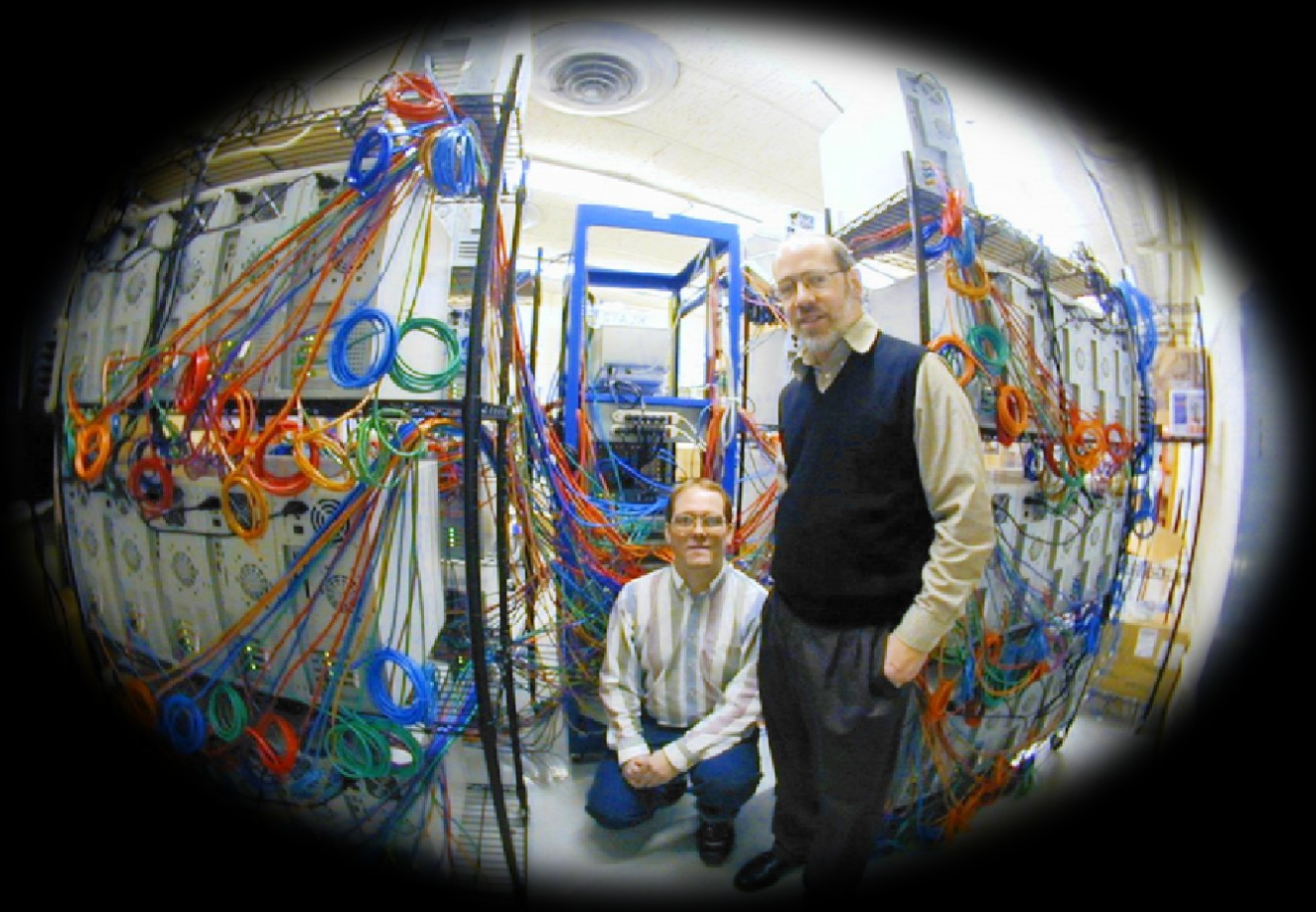


A *Little* Progress…

A GFLOPS is 1 Billion {+,*} per second

1992 MasPar MP1 $1,000,000 / GFLOPS

2000, KLAT2 was first FNN



A *Little* Progress…

A GFLOPS is 1 Billion {+,*} per second

1992	MasPar MP1	$1,000,000 / GFLOPS
2000	KLAT2	$650 / GFLOPS

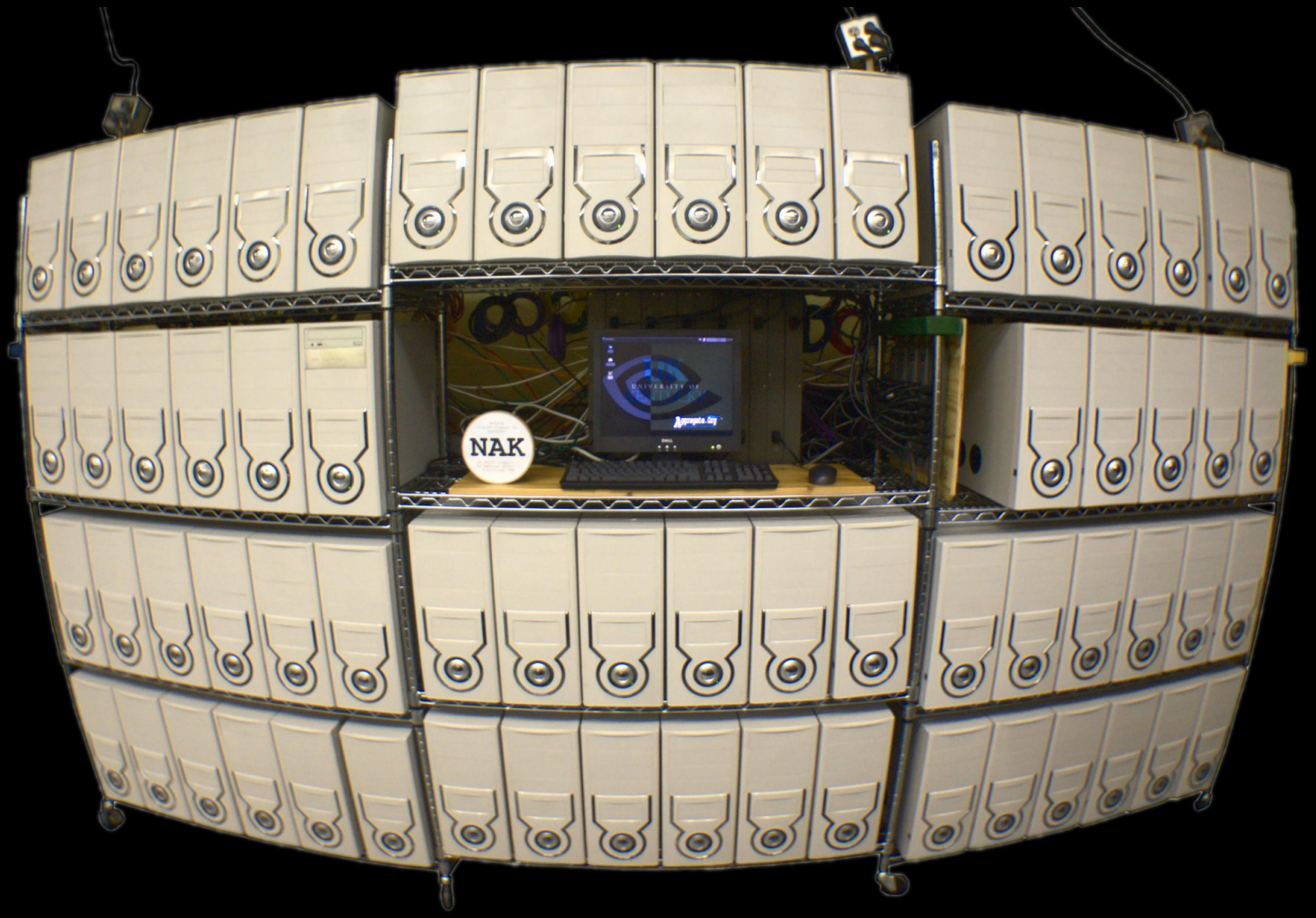
2003, KASY0

A *Little* Progress…

A GFLOPS is 1 Billion {+,*} per second

1992	MasPar MP1	$1,000,000 / GFLOPS
2000	KLAT2	$650 / GFLOPS
2003	KASY0	$84 / GFLOPS

2010, NAK used 1K PE GPUs



A *Little* Progress...

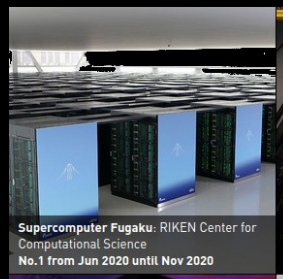
A GFLOPS is 1 Billion {+,*} per second

1992	MasPar MP1	$1,000,000 / GFLOPS
2000	KLAT2	$650 / GFLOPS
2003	KASY0	$84 / GFLOPS
2010	NAK	$0.65 / GFLOPS

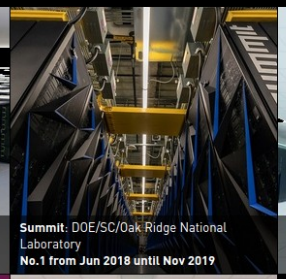
2022 GeForce RTX3090Ti peak is 40 TFLOPS
@ $2K (not counting host)... $0.05 / GFLOPS

A Lesson From http://top500.org

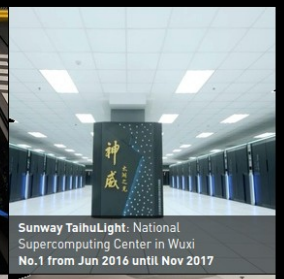
#1 Machines, http://top500.org



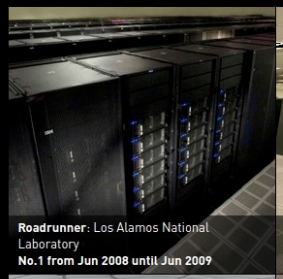
Supercomputer Fugaku: RIKEN Center for Computational Science
No.1 from Jun 2020 until Nov 2020



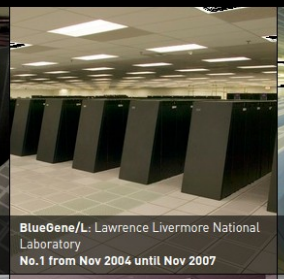
Summit: DOE/SC/Oak Ridge National Laboratory
No.1 from Jun 2018 until Nov 2019



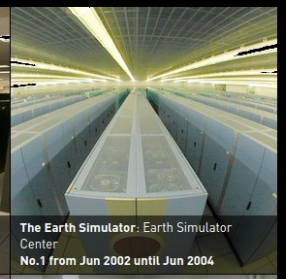
Sunway TaihuLight: National Supercomputing Center in Wuxi
No.1 from Jun 2016 until Nov 2017



Roadrunner: Los Alamos National Laboratory
No.1 from Jun 2008 until Jun 2009

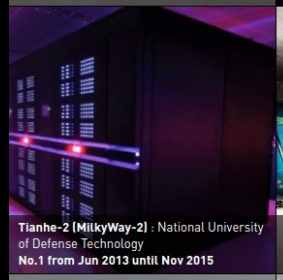


BlueGene/L: Lawrence Livermore National Laboratory
No.1 from Nov 2004 until Nov 2007

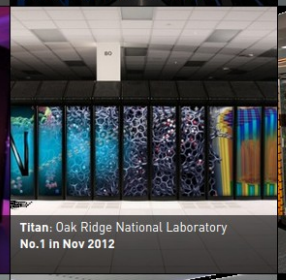


The Earth Simulator: Earth Simulator Center
No.1 from Jun 2002 until Jun 2004

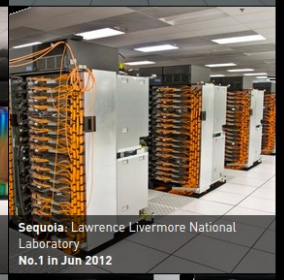
CM-5: Los Alamos National Lab
No.1 in Jun 1993



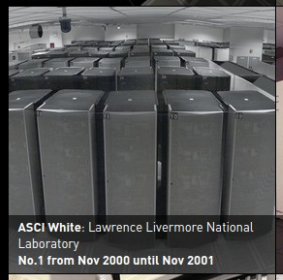
Tianhe-2 (MilkyWay-2) : National University of Defense Technology
No.1 from Jun 2013 until Nov 2015



Titan: Oak Ridge National Laboratory
No.1 in Nov 2012

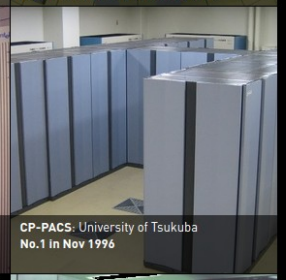


Sequoia: Lawrence Livermore National Laboratory
No.1 in Jun 2012

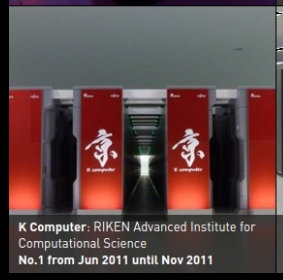


ASCI White: Lawrence Livermore National Laboratory
No.1 from Nov 2000 until Nov 2001

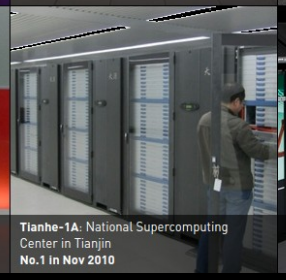
ASCI Red: Sandia National Laboratory
No.1 from Jun 1997 until Jun 2000



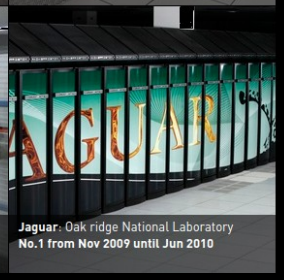
CP-PACS: University of Tsukuba
No.1 in Nov 1996



K Computer: RIKEN Advanced Institute for Computational Science
No.1 from Jun 2011 until Nov 2011

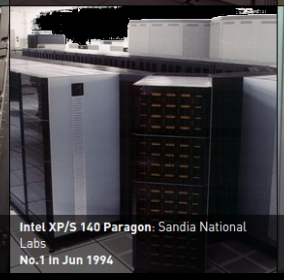


Tianhe-1A: National Supercomputing Center in Tianjin
No.1 in Nov 2010



Jaguar: Oak ridge National Laboratory
No.1 from Nov 2009 until Jun 2010

Hitachi SR2201: University of Tokyo
No.1 in Jun 1996



Intel XP/S 140 Paragon: Sandia National Labs
No.1 in Jun 1994

Numerical Wind Tunnel: National Aerospace Laboratory of Japan
No.1 in Nov 1993

8730112 cores, 1.1EFLOPS, 21MW

The Future

- Design for when product will be released
- Everything is moving down...
  Your cell phone outruns the 1992 MasPar MP1;
  supercomputer today, in your cell phone soon
- More parallelism (and maybe **quantum** too?)
- More heterogeneous (helped by **dark silicon**)
- Everything contains a connected computer
  (e.g., IoT: Internet of Things)… a good thing?