Bird’s-eye view of Computer Systems

Mahendra K. Verma

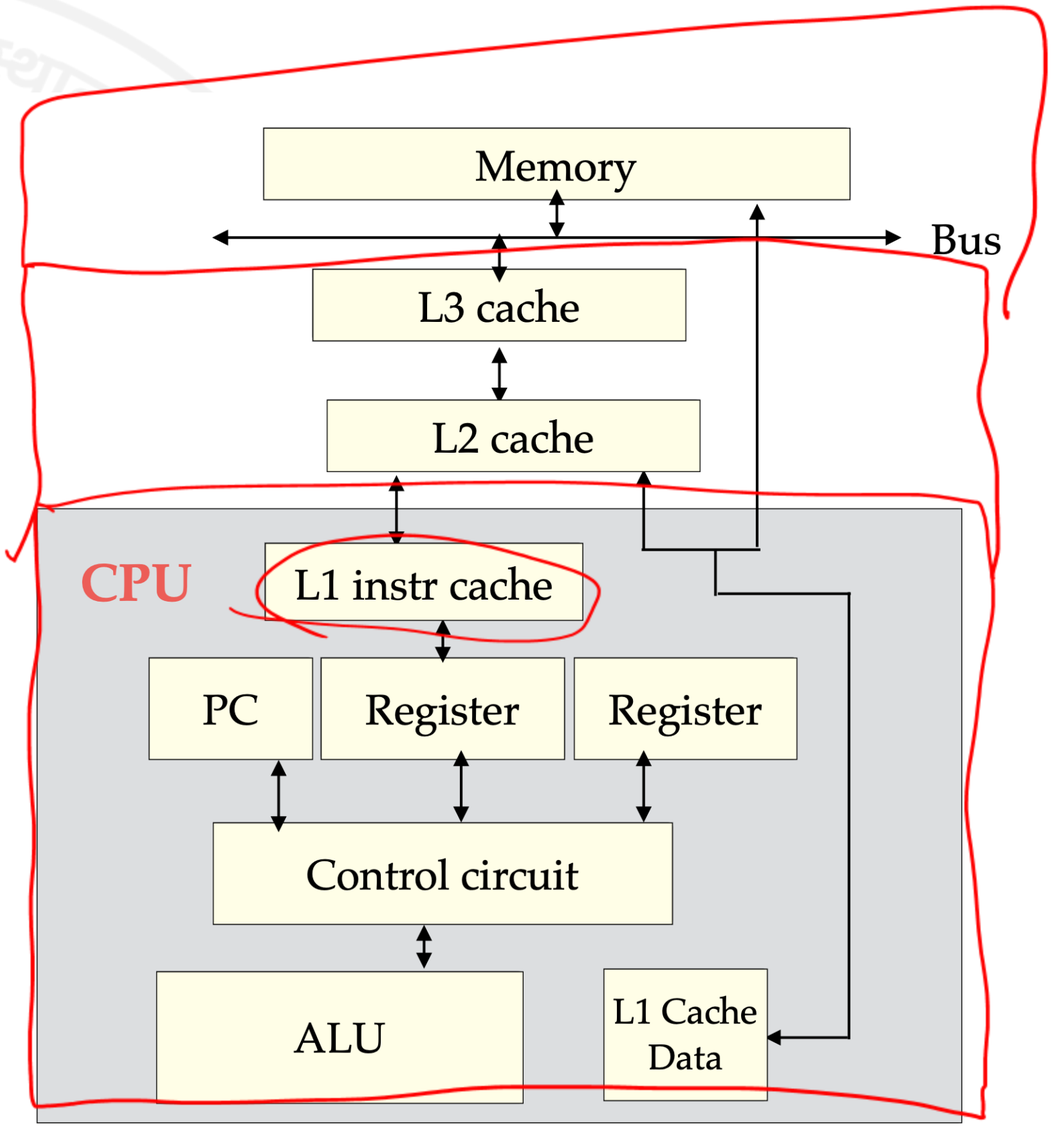
Ref: Hennessy & Patterson: Computer Architecture
& various websites

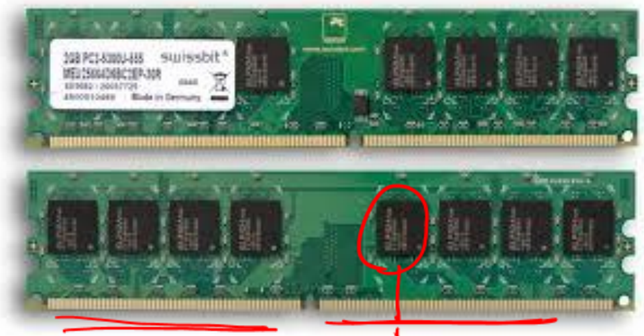
4. Memory

Memory hierarchy

In decreasing order of speed

- Register
- Cache (L1, L2, L3)
- RAM
- Hard disk





DIMM

RAM

SSD

Hard disk

from Wikipedia


Latency
Memory
Access time
Capacity
Speed / mem
CPU Register
0.2 ns
4 KB
500x
L1 cache
1 ns
64 KiB
100x
L2 cache
3-10 ns
1 MiB
25x
L3 cache
10-20 ns
1 MiB
1x
DRAM
50-100 ns
8-4096 GiB
1
Solid state drive (SSD)
25-100 µs
½-10 TB
1/30 x
Hard disk
5-10 ms
1-20 TB
10-5 x
Patterson

Analogy

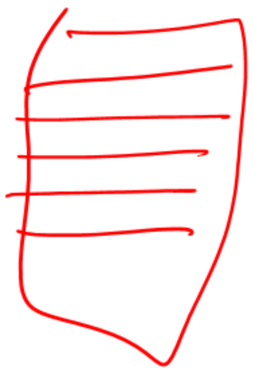
- Books on the table: cache
- Books in a bookshelf: RAM
- Library: HD

Latency

- A delay between data request from the processor and data fetch.
- Also called access time.

- The relevant program and data are brought into the RAM for processing.
- During the execution, parts of the code and data that are needed immediately are sent to the cache from RAM.
- Data access,:first try in L1 cache. If successful, it is called a cache hit. If CPU fails in finding the data, it then looks for the data in L2 cache, and then in L3 cache.
- If the CPU doesn't find the data in any of the memory caches, it attempts to access it from your system memory (RAM). When that happens, it is known as a cache miss.
- From cache or RAM (less often), small chunks of instructions and data arrive at the registers. CPU executes these instructions and operates on the data of the registers (e.g., adding numbers). The output is sent to the cache.
- The cache is refreshed with new set of instructions and data. The output is written back to RAM.
- Large outputs (e.g., arrays) are written on the hard disk, either during the execution or at the end of the program.

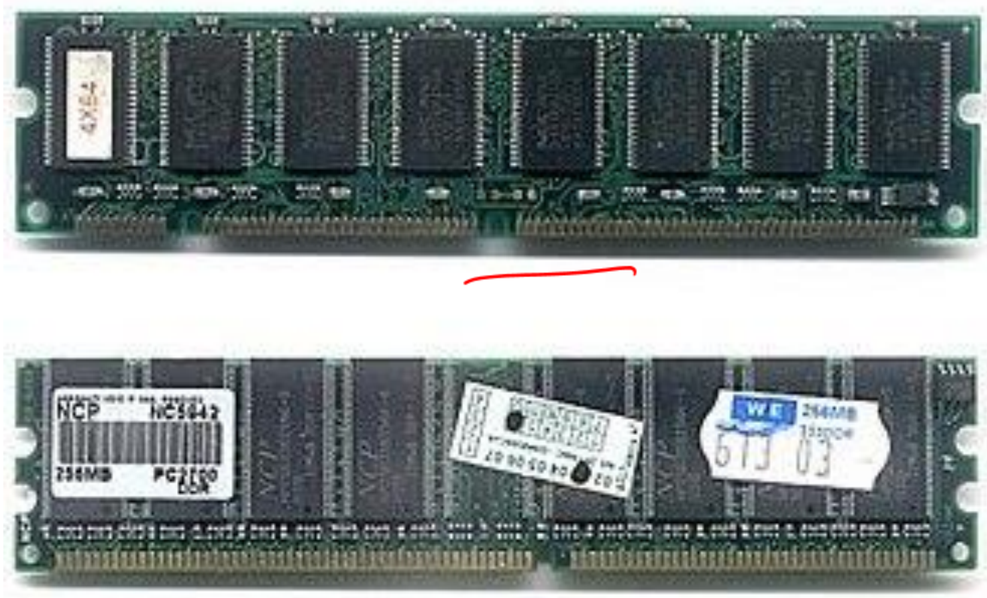
Locality



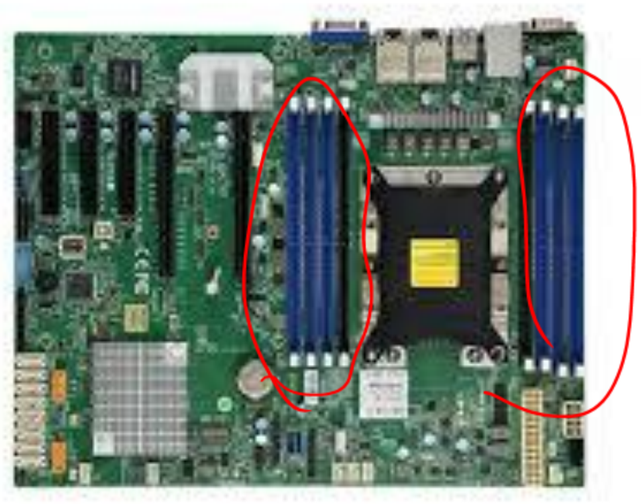
- Spatial locality: Instruction or data elements within relatively close storage locations.
  - Array elements

- Temporal locality: Instruction or data reuse within a relatively small time duration.
  - Variable sum in array sum should be kept in cache till the sum is being compute.

Main memory (RAM)



Wikipedia

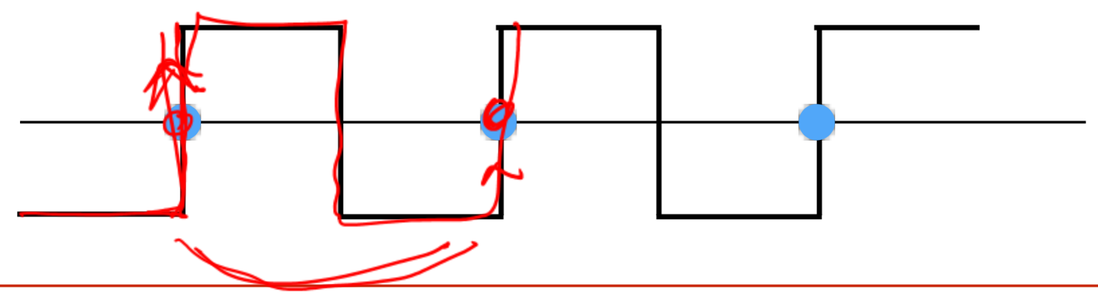


https://www.supermicro.com/en/products/motherboards/

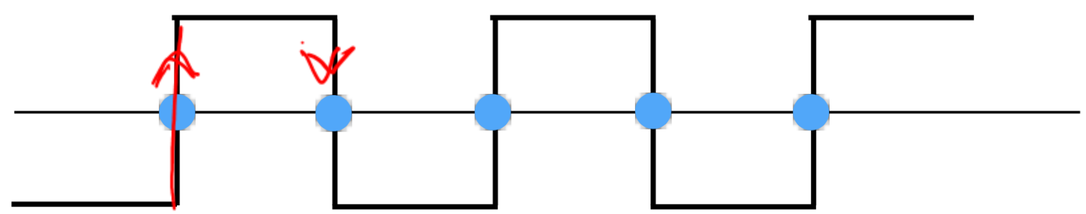
RAM (Random access memory)

- Dual In-Line Memory Module (DIMM)
- DIMM size: 4GB-128GB
- We can fit many DIMMS on a motherboard
- DIMM: 288 pins
- Note: M-1 chip has memory inside the proc.

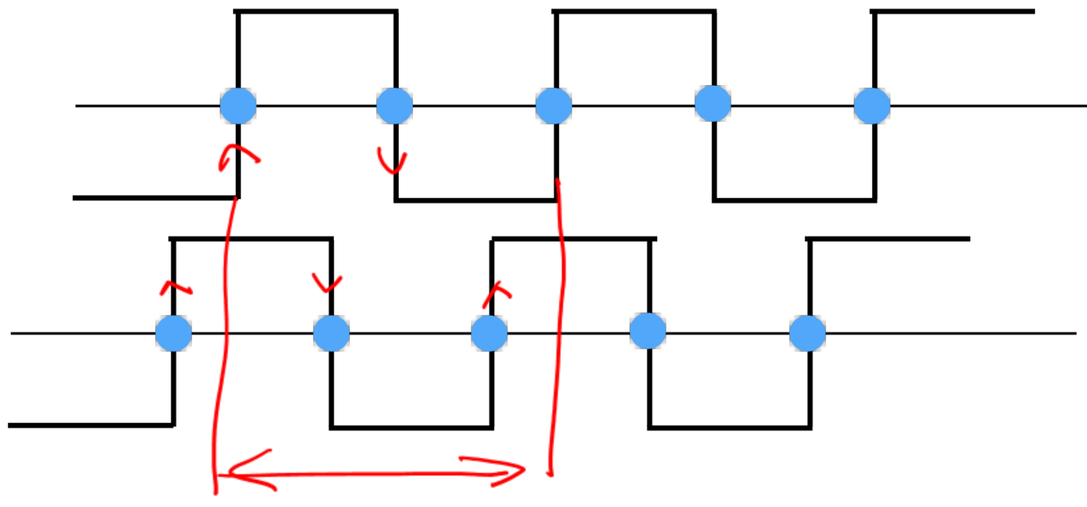
SDR
1 signal/
cycle



DDR
2 signal/
cycle



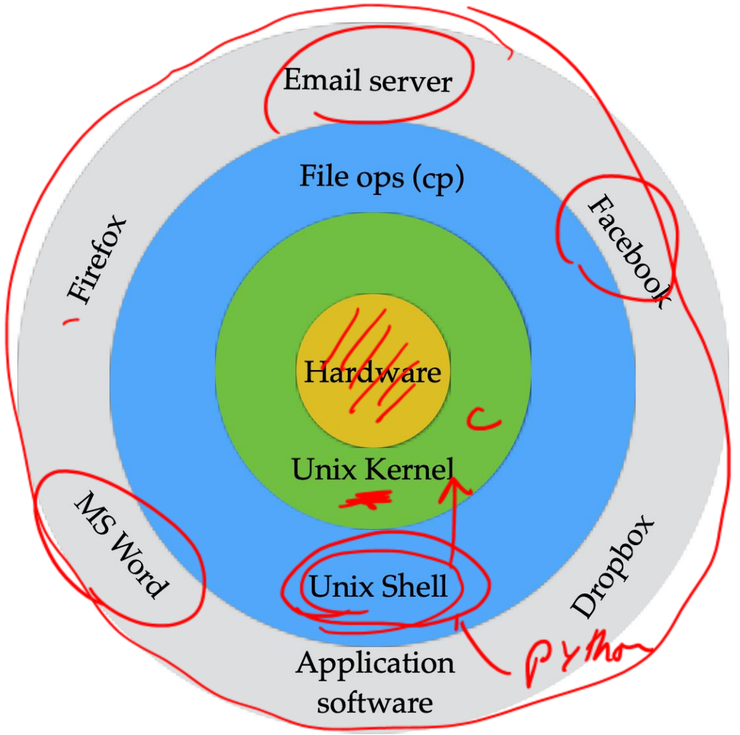
QDR
4 signals/
cycle

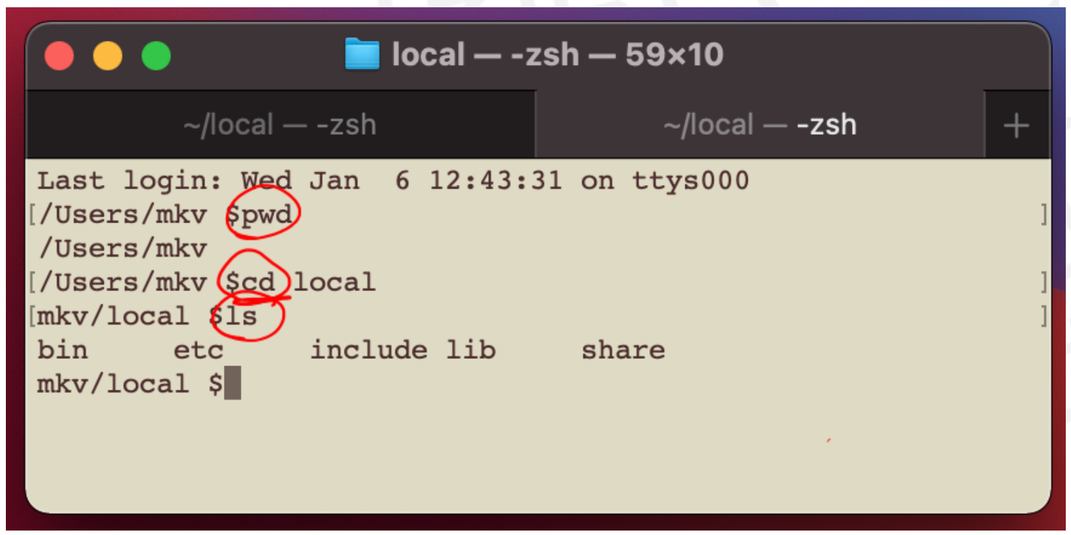


- SRAM: Static RAM uses bistable latching circuit. L2 and L3 cache employ SRAM.
- DRAM: Dynamic RAM (MOSFET technology). DRAM is slower than SRAM.
- SDRAM (synchronous DRAM): Synchronised with the clock speed.
- DDR: Double data rate SDRAM. Here, both upward and downward edges are used for data transmission.
- QDR: Quad Data Rate (QDR) SRAM is a static RAM that can transfer up to four words of data in each clock cycle. It employs upward and downward edges of two signals separated by a phase shift of $\pi/2$.
- GDDR or VRAM: Graphics DDR RAM and Video RAM. GDDRs are faster than DDR RAM.
- HBM: High Bandwidth Memory or 3D stacked SDRAM are packaged on top of a CPU. HBMs have lower latency and higher bandwidth than DDR RAM.

Basics of Operating System

- Unix Kernel manages HW, user, programs, apps, etc.
- Unix shell: Interface to the unix system
- Applications run on top.





Terminal

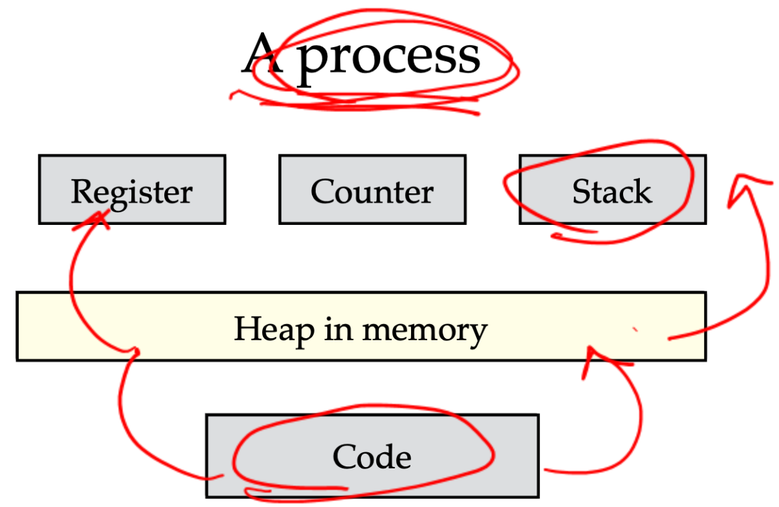
Compilers

Interpreter

- Compilers convert higher-level programs to object codes (1 and 0's).
- These object codes reside in memory.
- CPU executes an object code in sequence.

Program execution

- Register, Program counter, heap, stack reserved.
- First line of the code is loaded into instruction register
- Execute the instruction
- Increment program counter
- Get the next instruction and execu
- Continue till the end
- Release memory & registers at the



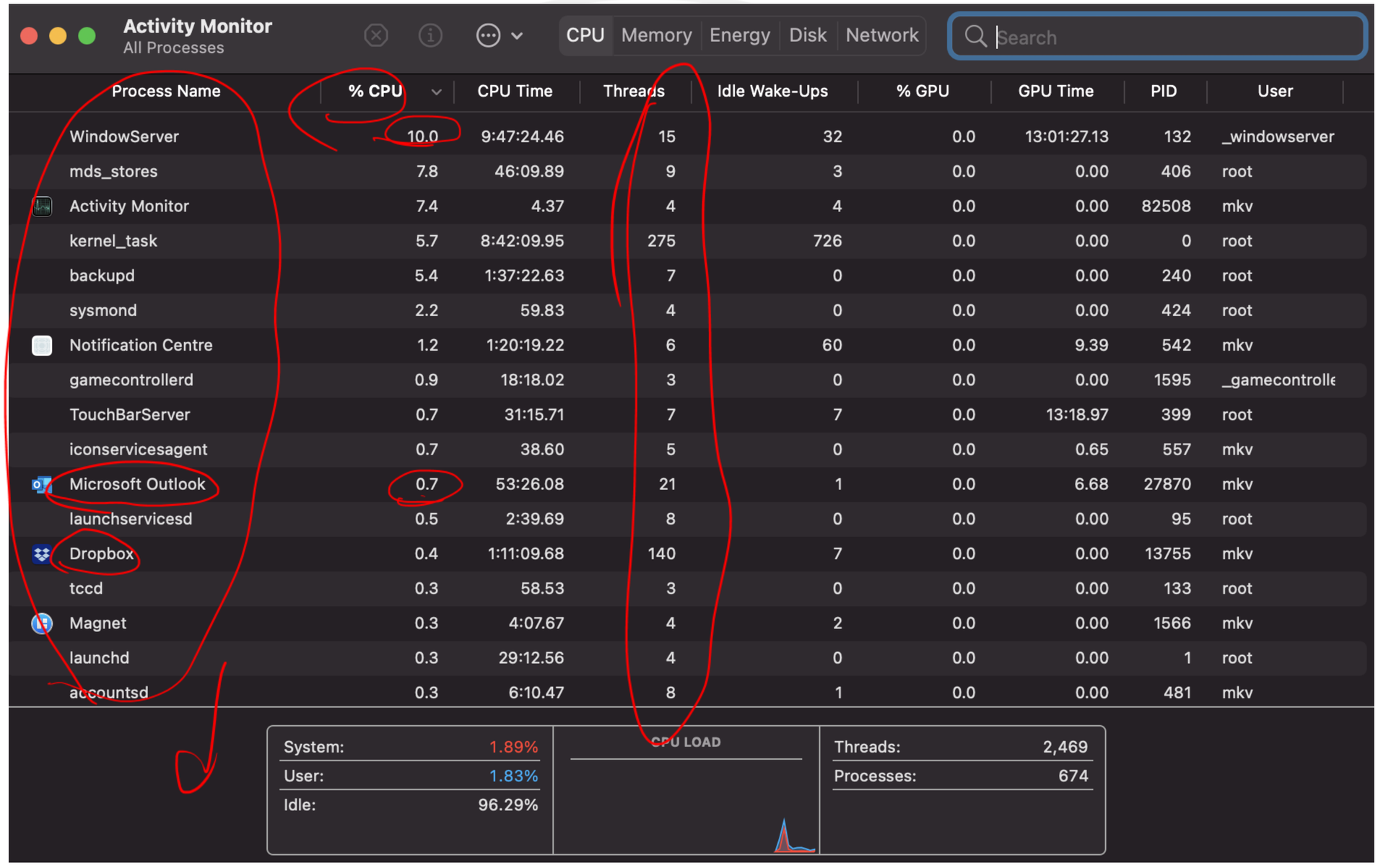
Process

The program along with associated registers, program counter, stack, and heap is called a *process*.

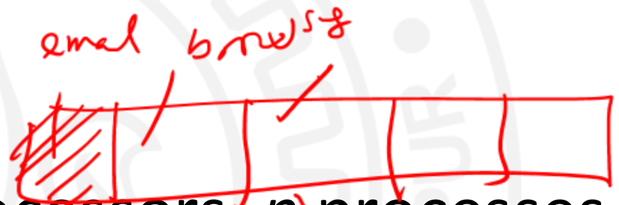
```
[/Users/mkv $ps -A
  PID TTY           TIME CMD
    1 ??         29:17.49 /sbin/launchd
   54 ??          1:04.89 /usr/sbin/syslogd
   55 ??          0:58.46 /usr/libexec/UserEventAgent (System)
```

- PID: Process id
- TIME: Process running time in hour, min, second format
- CMD: Who started and owns the process?

Activity Monitor in Mac



- In a computer with a single-core processor, only a single process can run at a time.
- Multitasking
- HPC: In a multicore system with *n* processors, *n* processes can be executed simultaneously with one process on each core.
- A single program can have several instances. Each instance of the program is a separate process.

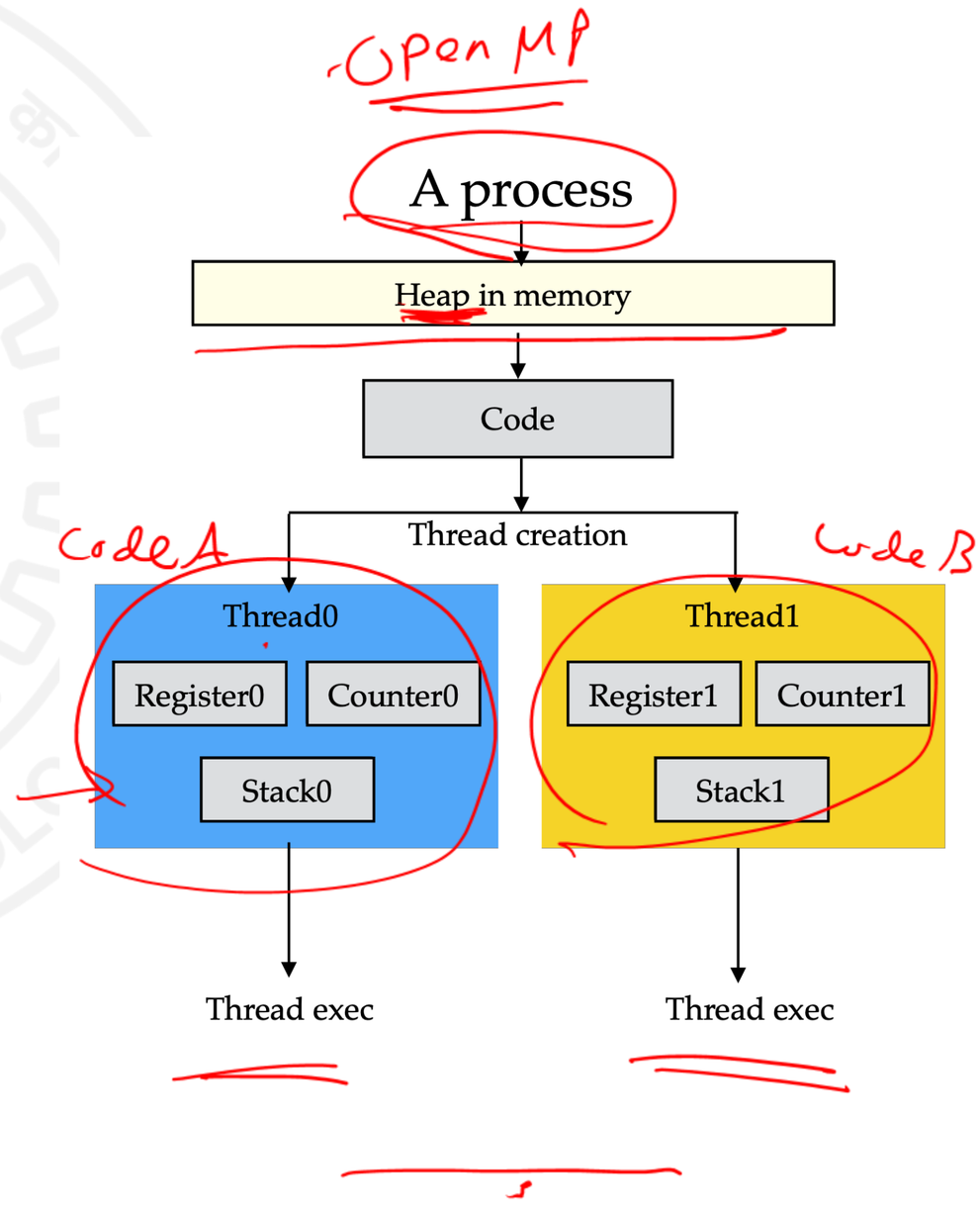


Every process in a computer is independent. That is, a process cannot access the data of another process.

This feature is to ensure that one process does not corrupt the other process.

Thread

- A process can be broken into several units of execution that can run in parallel.
- Such units are called threads.
- Each of these threads has its own registers and stack.
- The threads share the memory (heap) and other resources of the parent process.
- Threads are lightweight. Thread termination is quick.
- SMPs and GPUs employ threads.



Process	Thread
Creation and termination of a process Heavyweight operation	Thread creation and termination is a Lightweight operation
A process has its own memory. One process does not share memory with another process.	Threads share the memory of the parent process that generates the threads.
One core typically runs one process.	One core can create many threads. HPC app: 1 thread/core

DDR: (memory bus clock rate) × 2 (for dual rate)
× 64 (number of bits transferred)
/ 8 (number of bits/byte)

Clock rate
× 2 × 8 × 4

DDR4 = mult by 4

400*2*8*4 = 25600 MB/sec

25.6 GB/sec
↑
bytes

Hard disk (HD) & Solid state drive (SSD)

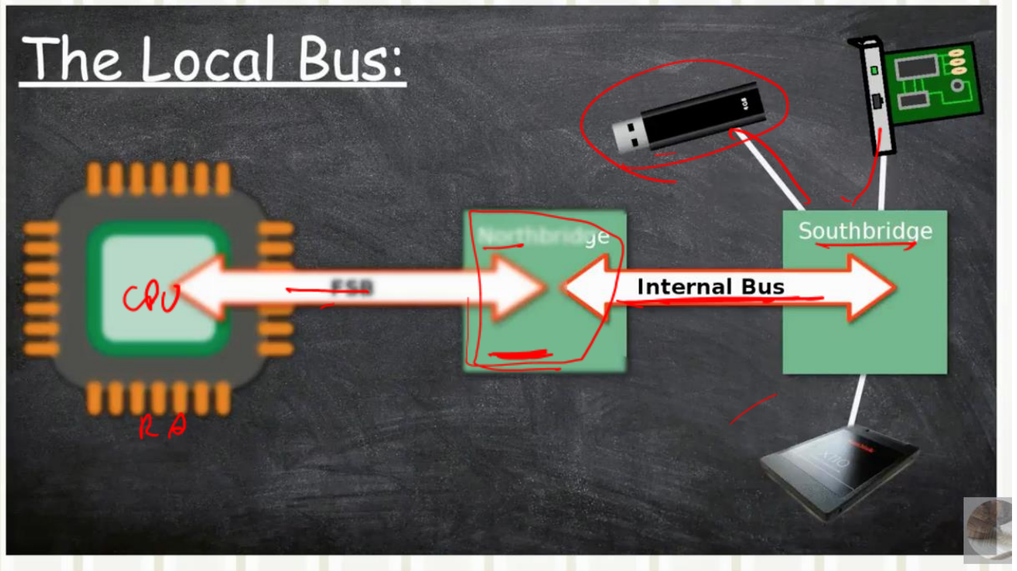
SSD

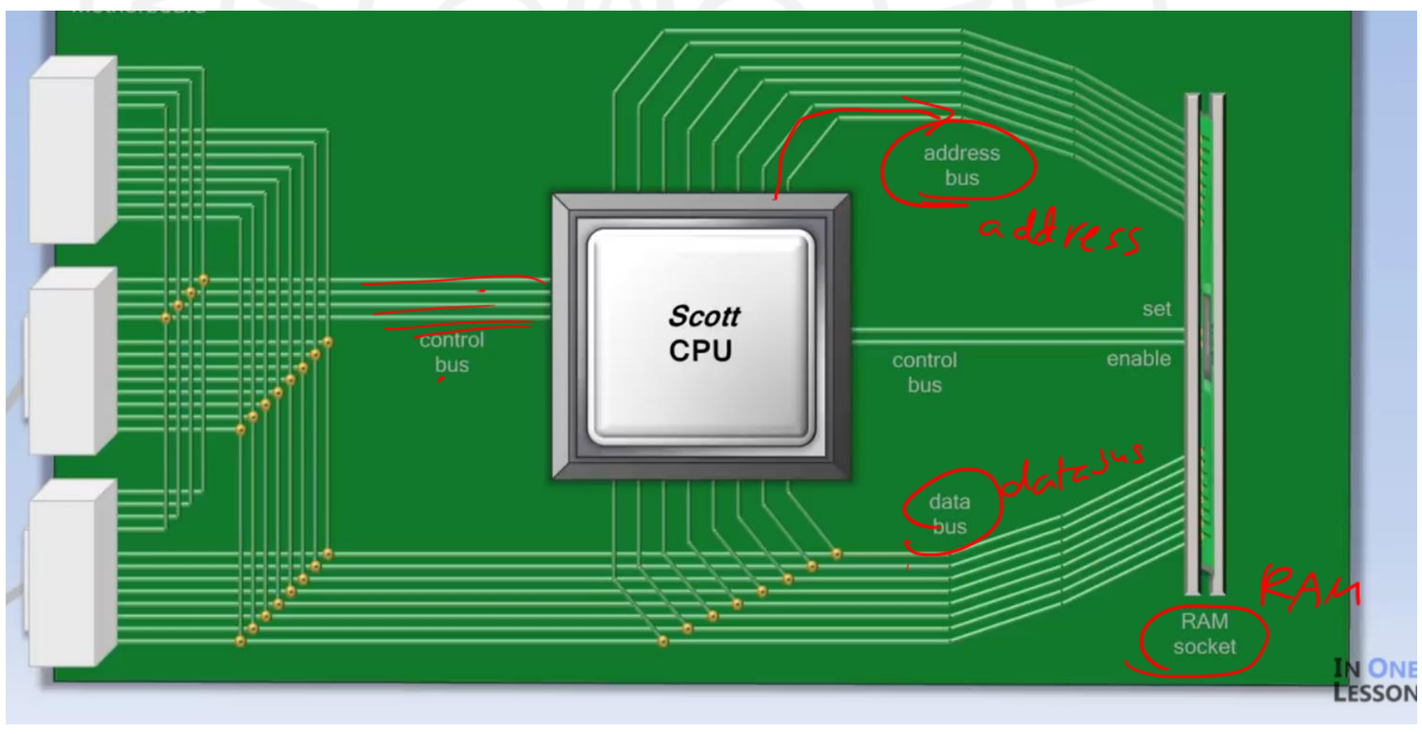
Hard disk

from Wikipedia

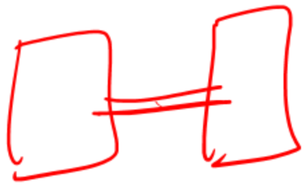
- HD: Magnetic heads
  - Data transfer rate to RAM is 100-200 MB/s.
  - 1TB-20 TB
- SSD: Solid-state drive
  - 550 MB/s
  - 128 GB - 10 TB

Connections


The Local Bus:
CPU
FSB
Northbridge
Internal Bus
Southbridge
RA


address bus
address
Scott CPU
control bus
control bus
set
enable
data bus
data bus
RAM
RAM socket
IN ONE LESSON

Peripheral component interface express



	x1	x4	x8	x16
PCIe 1.0	250MB/s	1GB/s	2GB/s	4GB/s
PCIe 2.0	500MB/s	2GB/s	4GB/s	8GB/s
PCIe 3.0	985MB/s	3.94GB/s	7.88GB/s	15.8GB/s
PCIe 4.0	1.97GB/s	7.88GB/s	15.8GB/s	31.5GB/s
PCIe 5.0*	3.94GB/s	15.8GB/s	31.5GB/s	63.0GB/s

NVlink 600 GB/sec

Bus PCIe 4.0

- Data runs on the bus.
- PCIe 4.0 x16 (128 lanes): 31.5 GB/sec.
- Compare it with the MTs of DDR4-3600, which is 25.6 GB/sec
- PCIe 5.0*: 63.0 GB/sec
- NVlink: 600 GB/sec

Back to Rome processor

Mem Hierarchy	AMD EPYC 7742 DDR4-3200 (ns @ 3.4GHz)	AMD EPYC 7601 DDR4-2400 (ns @ 3.2GHz)	Intel Xeon 8280 DDR-2666 (ns @ 2.7GHz)
L1 Cache	32KB 4 cycles *1.18ns*	32KB 4 cycles *1.25ns*	32KB 4 cycles *1.48ns*
L2 Cache	512KB 13 cycles *3.86ns*	512KB 12 cycles *3.76ns*	1024KB 14 cycles *5.18ns*
L3 Cache	16MB / CCX (4C) 256MB Total ~34 cycles (avg) *~10.27 ns*	16MB / CCX (4C) 64MB Total	38.5MB / (28C) Shared ~46 cycles (avg) *~17.5ns*
DRAM **128MB Full Random**	~122ns (NPS1) ~113ns (NPS4)	~116ns	~89ns
DRAM **512MB Full Random**	~134ns (NPS1) ~125ns (NPS4)		~109ns

Memory bottleneck

- AMD Rome 7742 has 8 channels.
- Each channel supply max of 25.6 GB/sec
- Hence, total data transfer = 25.6x8 = 204.8 GB/sec = 25.6 Giga floats/sec
- The processor can process 2 trillion float ops.
- Clearly, processor can do more than what mem can supply (80 times).

- The data in cache is accessed faster.
- In Rome, to each core, L1 cache can send 32 bytes = 4 words/cycle.
- Hence, 64cores x 4 words x 3 GHz ≈ 0.75 trillions ops can be performed in principle, if optimised to the hilt.

Your own machine

MacBook Air

M1, 2020

Name	[redacted]
Chip	Apple M1
Memory	16 GB
Serial number	[redacted]
Limited Warranty	Expires 18-Oct-2023 Details...

macOS

macOS Ventura Version 13.2.1

Displays

Built-in Retina Display 13.3-inch (2560 × 1600)

Display Settings...

Storage

Macintosh HD 99.97 GB available of 245.11 GB

- Hardware
  - ATA
  - Apple Pay
  - Audio
  - Bluetooth
  - Camera
  - Card Reader
  - Controller
  - Diagnostics
  - Disc Burning
  - Ethernet
  - Fibre Channel
  - FireWire
  - Graphics/Displays
  - Memory

Hardware Overview:

Model Name:	MacBook Air
Model Identifier:	MacBookAir10,1
Model Number:	[redacted]
Chip:	Apple M1
Total Number of Cores:	8 (4 performance and 4 efficiency)
Memory:	16 GB
System Firmware Version:	[redacted]
OS Loader Version:	
Serial Number (system):	
Hardware UUID:	
Provisioning UDID:	
Activation Lock Status:	

Memory:	16 GB
Type:	LPDDR4
Manufacturer:	Hynix

High performance computing (HPC) clusters

1. Introduction

HPC

High perf computing

- Large computing system for complex computing tasks.
  - Parallel computer for solving complex problems.
  - Large data centres for banking, cloud storage, social network, search engines (e.g., google)
  - These machines are much more powerful than desktops or servers.
  - Supercomputer: a fast computer (specially-designed or computers strung together).

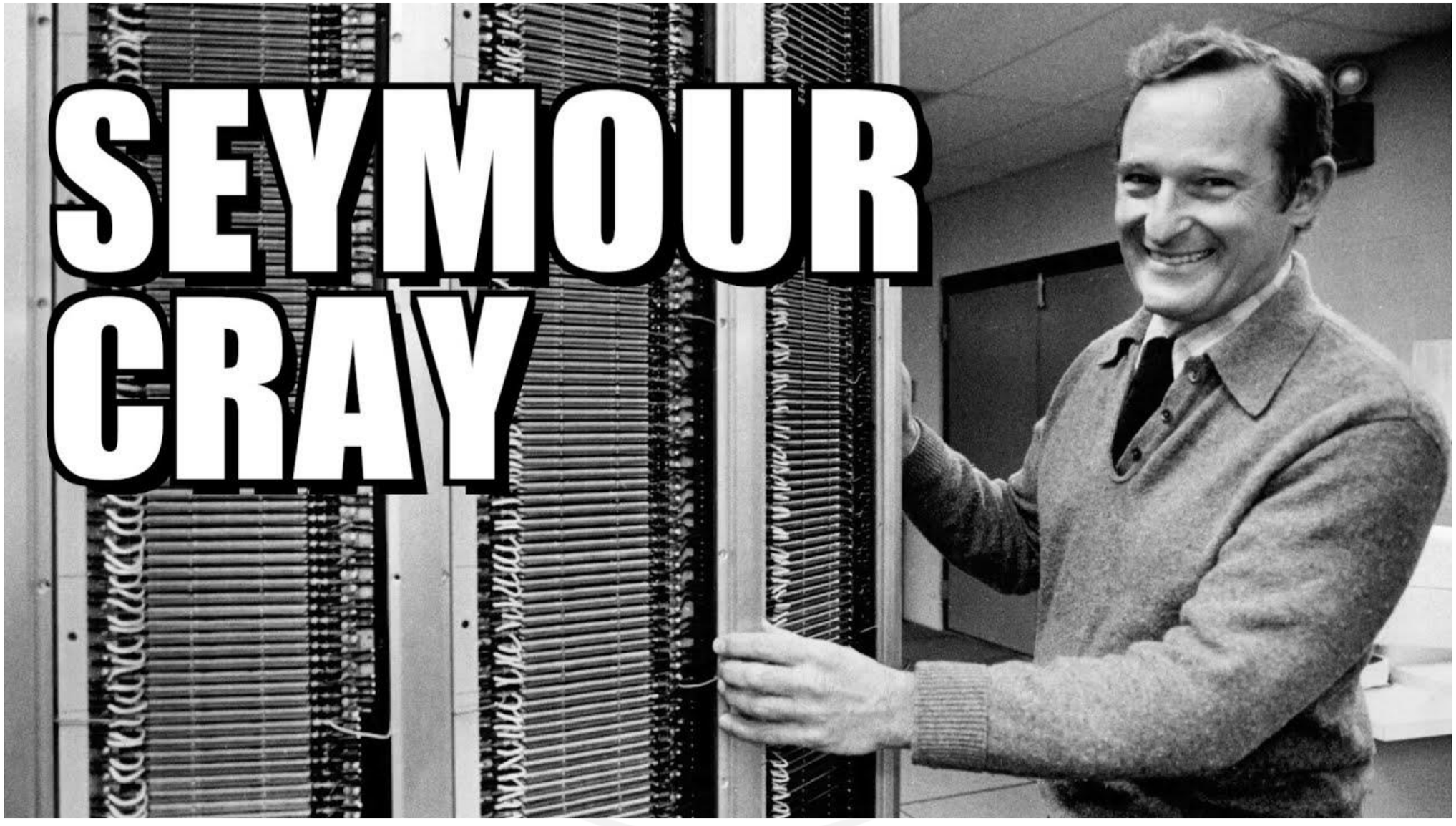
Past designs
Till 1990’s

“Anyone can build a fast CPU. The trick is to build a fast system.”

"If you were plowing a field, which would you rather use: two strong oxen or 1024 chickens?”

— Seymour Cray

Wiki



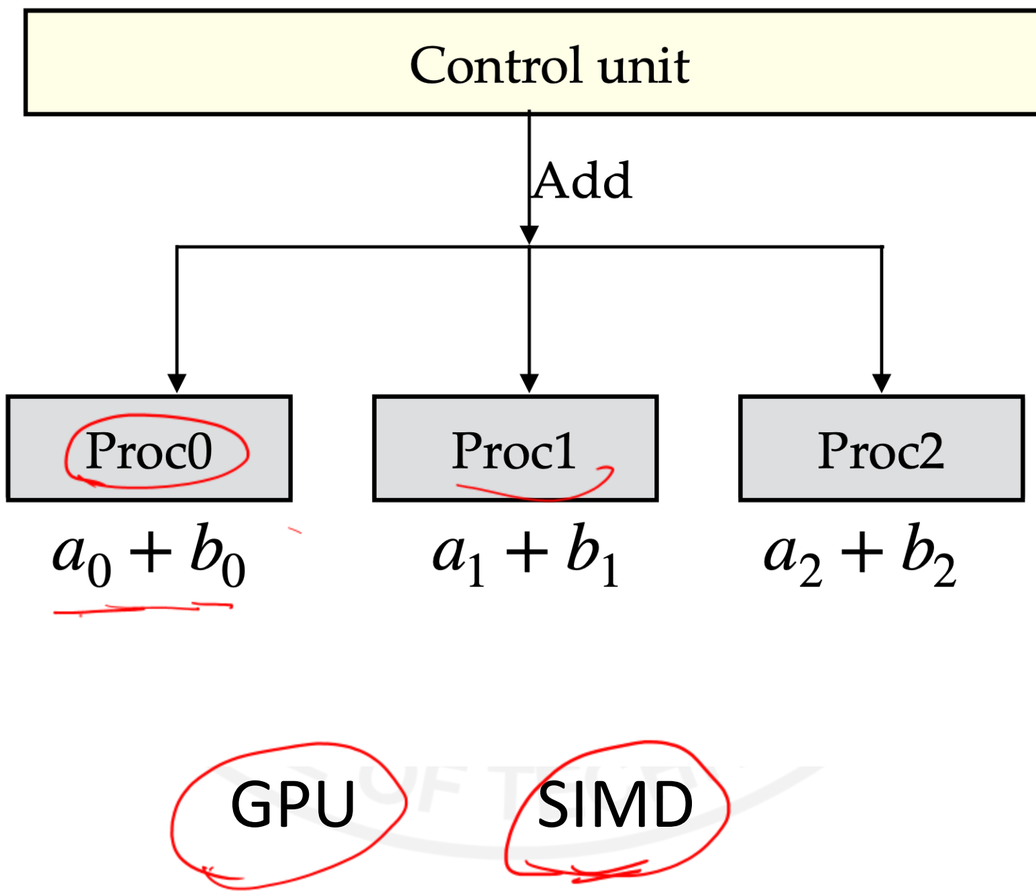
https://www.youtube.com/watch?v=htWqxABAoRE

TECH STORIES: Life & Work of Seymour Cray

Cray's Vector machines

- Early desktops/servers were quite slow.
- Cray machines: Perform many operations in a single cycle.
  - Vector machines (operate on arrays)
  - Pipelines

Array processor



GPU SIMD

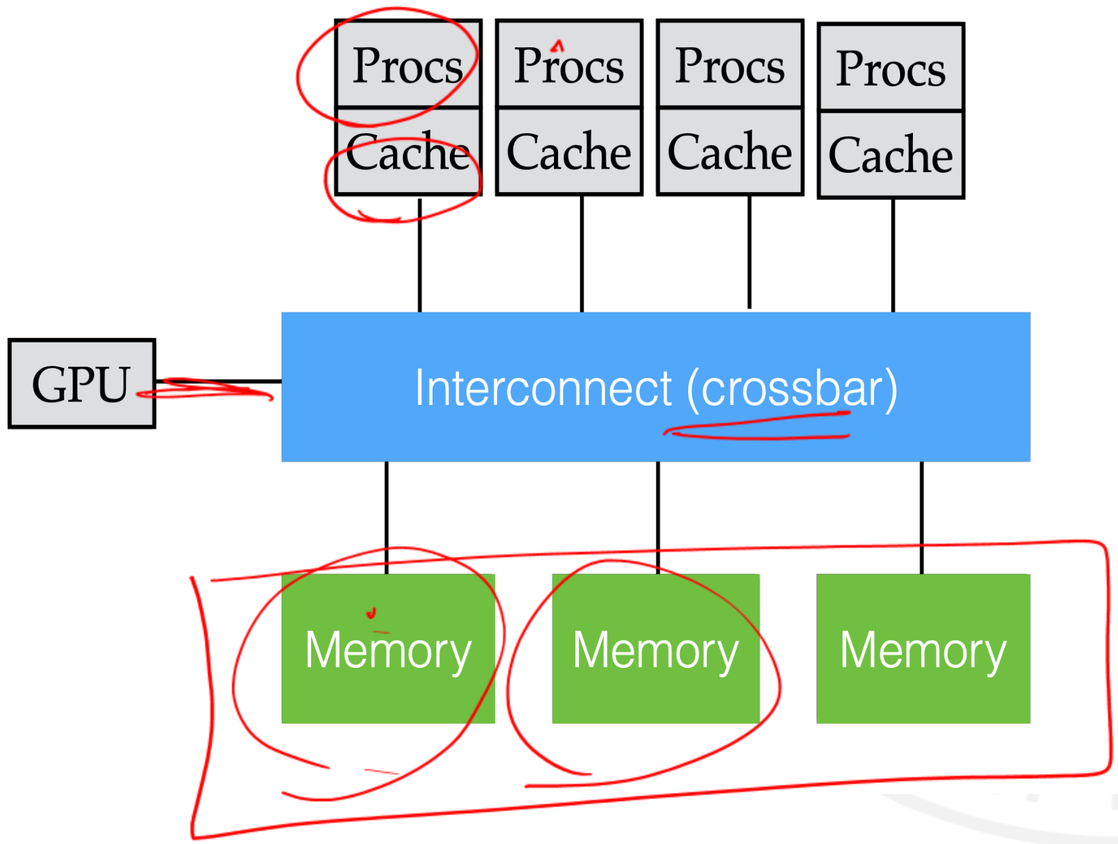
Pipelines

A*B+C using pipeline

Clock Cycle	Segment 1 R1, R2	Segment 2 R3, R4	Segment 3 R5
1	A1, B1		
2	A2, B2	A1*B1, C1	
3	A3, B3	A2*B2, C2	A1*B1+C1
4	A4, B4	A3*B3, C3	A2*B2+C2
5	A5, B5	A4*B4, C4	A3*B3+C3

2. Parallel computer Classification

Shared memory machines



Unified memory access (UMA)

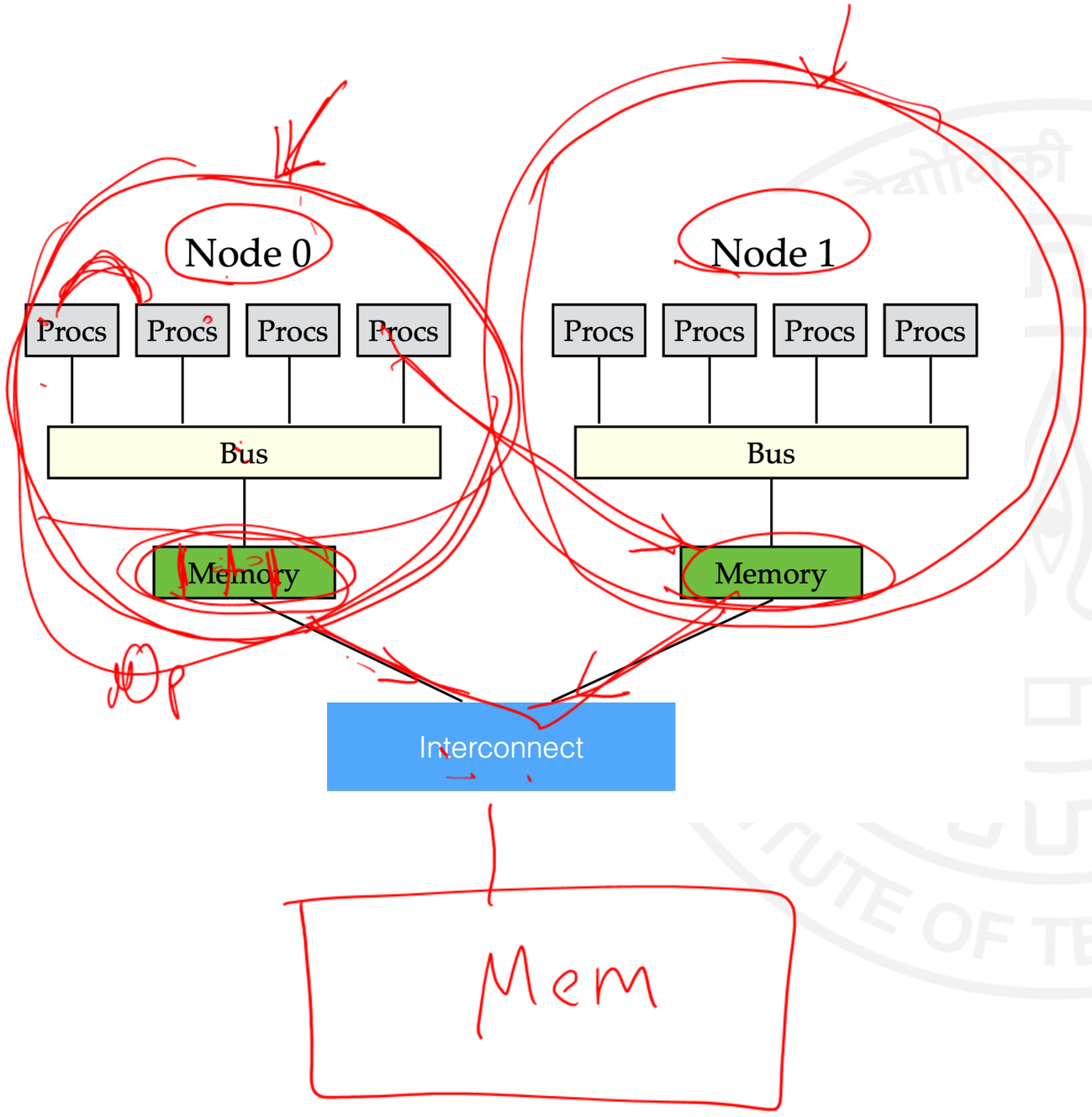
Fast memory access

Apple's M1 chip

OpenMP programming

easier than MPI

Distributed memory machines



Non-uniform memory access (NUMA)

Nodes consists of processors + mem

Nodes are connected via interconnect

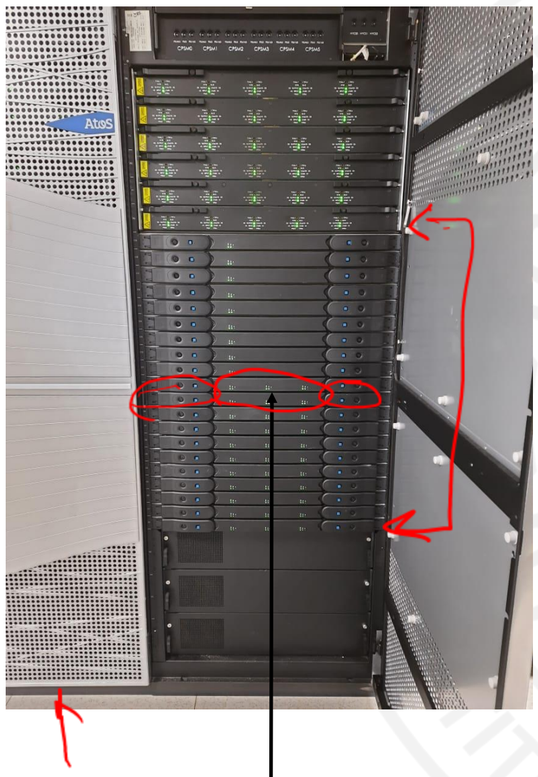
Programming paradigm like MPI

Hybrid: GPUs on many nodes

Message Passing Interface

IITK's PARAM Sanganak

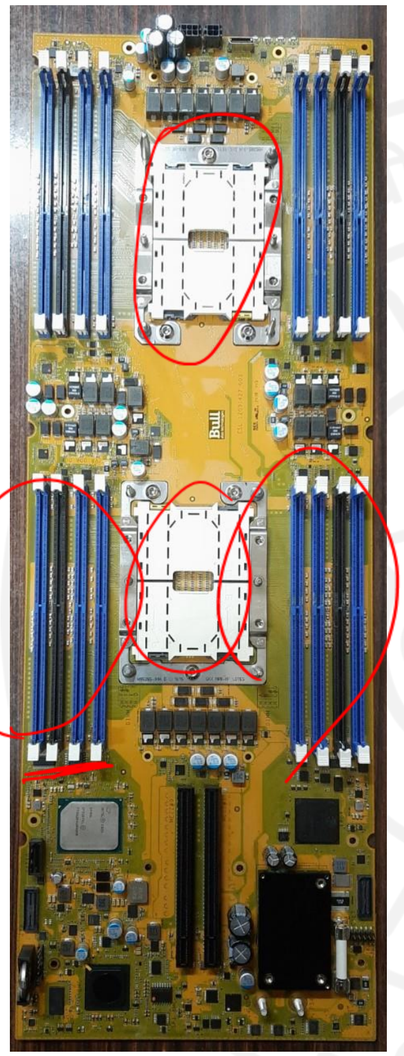
Switch



Nodes

Rack

Water cooling



Motherboard

Network card

Classification

Classes of parallelism

- Data-level parallelism: Here, many data items are operated on in parallel. Examples:
  - Many teachers are grading exam copies simultaneously.
  - Array operation C = A+B where A and B are two large arrays.
- Task-level parallelism: Here, different tasks are performed simultaneously. Examples:
  - Wars being fought at several fronts.
  - Google server searching different queries simultaneously.
  - Simultaneous simulations of Earth's atmosphere and oceans on different sets of processors of a parallel computer.

Implementation of parallelism on computers

- Instruction-level parallelism: Within the CPU itself.
  - Examples: pipelining; speculative execution
- Vector architecture and Graphic Processor Units (GPUs): Employs same operation on large data set.
  - Example: C=A+B; Rotating all the pixels of an object.

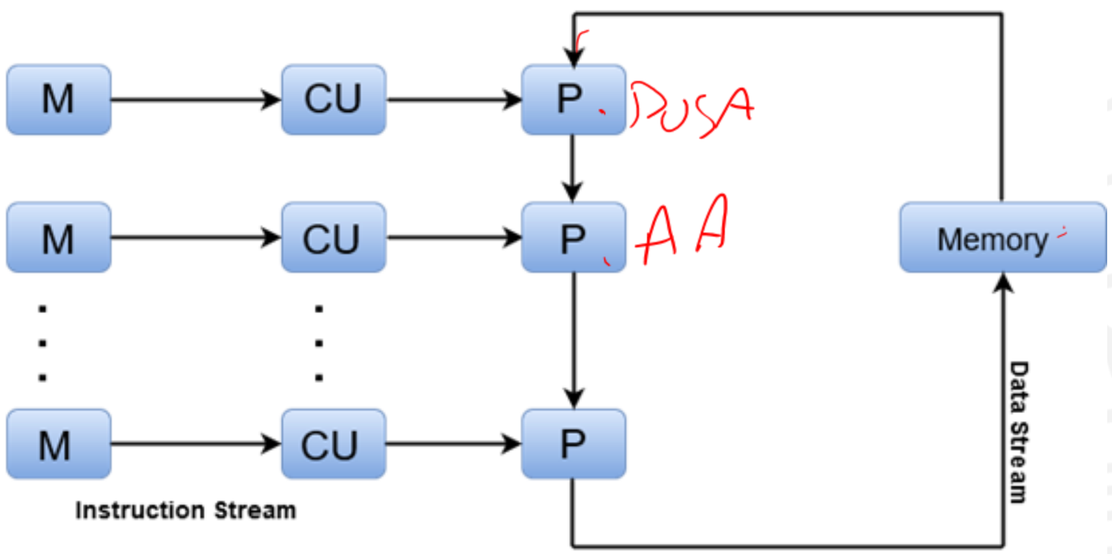
- Thread-level parallelism: In Symmetric Multiprocessors (SMP), we employ threads to divide the tasks.
  - Example: OpenMP
- Request-level parallelism: Perform decoupled tasks.
  - Example: ATMs of a bank.

Taxonomy of computers based on parallelism

- Single instruction stream, single data stream (SISD): Tasks performed in a uniprocessor. In this framework, the best option is instruction-level parallelism.

- Single instruction stream, many data stream (SIMD): Same instruction is executed by many data.
  - Threads employed to stream processors of GPU. Rotation of an image

- Multiple instruction stream, single data stream (MISD):
  A student going to different counters for registration (academic, housing countersue etc.)
- Present-day computers do not employ this strategy.



https://www.javatpoint.com/misd

- Multiple instruction stream, many data stream (MIMD): Each processor runs its own instruction on its own data. Task-level parallelism.

  - Examples: Warehouse systems (google); atmosphere-ocean simulation

Memory-based classification

- Shared memory systems:
  - memory is shared among the processes or thread.
  - OpenMP is employed for SMP systems.
  - Unified memory access (UMA)
- Distributed memory systems:
  - Each processor has its own memory.
  - Data exchanged depending on need.
  - Message Passing Interface (MPI) for such tasks.
  - Nonuniform memory access (NUMA)

3. Network

Communication between processors

Nodes

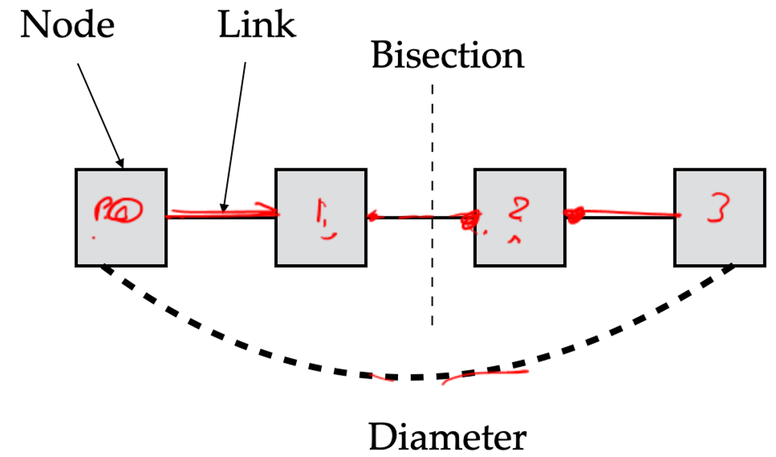
Rack

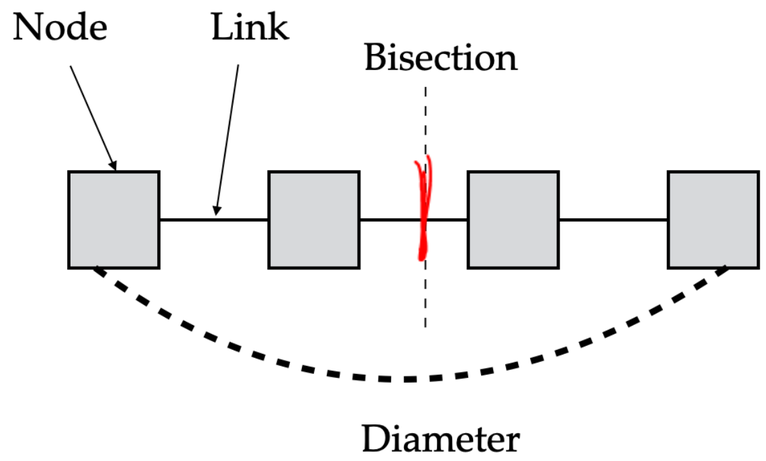
Diameter: The longest distance between any pair of nodes. Network hops.

Links: Number of links or connections in the network.

Degree: Number of in/out links at each router.

Average distance: Expected distance between two randomly selected nodes.



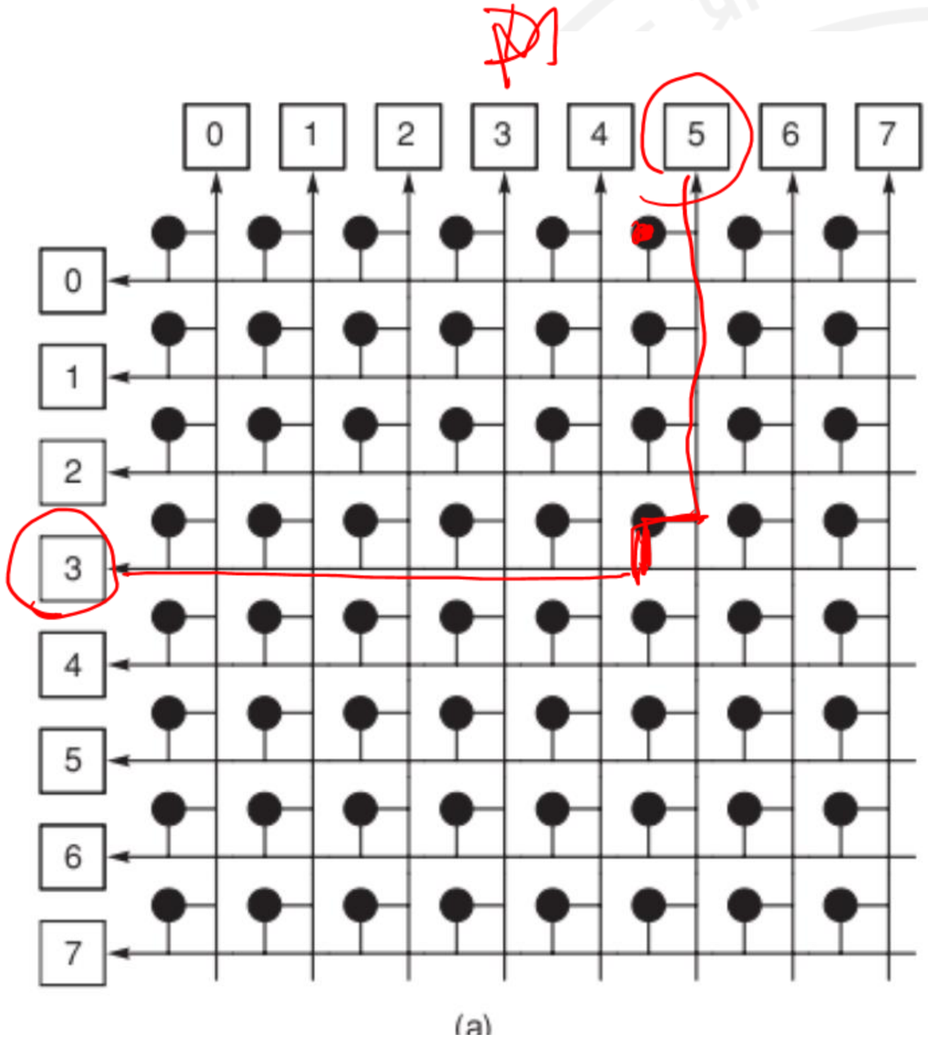


Bisection bandwidth: Minimum number of links that must be cut for partitioning the network into two equal halves.

Bandwidth: Capacity of a network to transmit the maximum amount of data in a given time.

Speed: Rate of data transfer

Crossbar switch

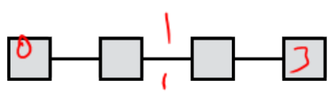


(a)

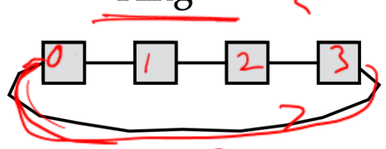
All-to-all connection

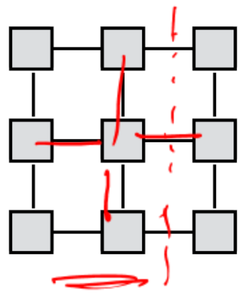
But, this arrangement is expensive.

Linear

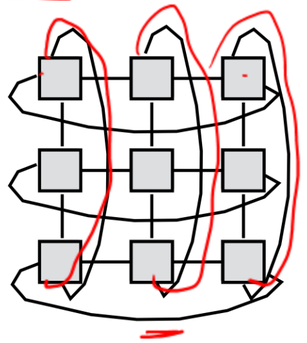


Ring





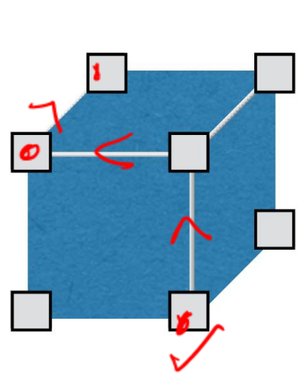
Square



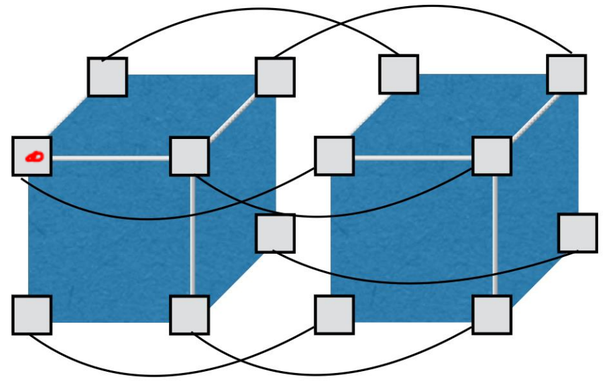
Torus-2 (T2)

Topology	Diameter	Bisection BW	# Links	Degree
Linear array	p−1	1	p−1	2
Ring	p/2	2	p	2
Square	2(√p-1)	√p	2√p(√p-1)	4
Torus-2 (T2)	√p	2√p	2p	4

Hypercube



Cube



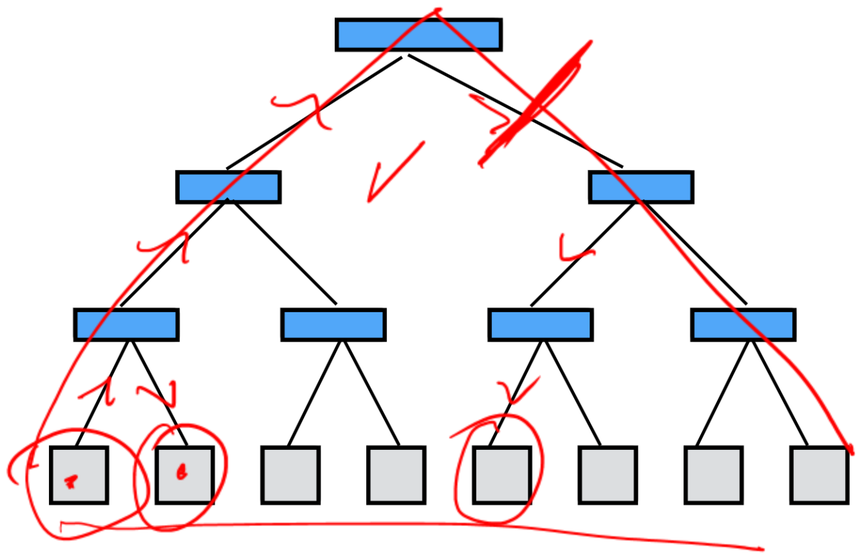
Hypercube

Torus

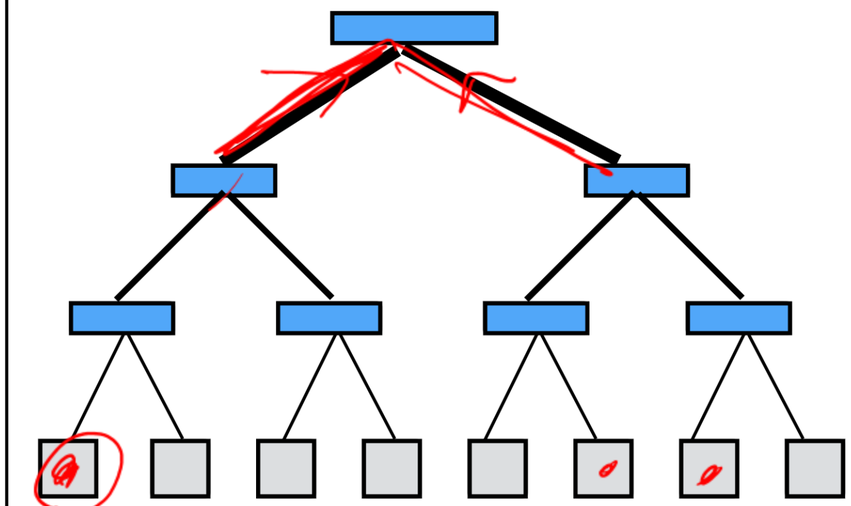
T^3

Topology	Diameter	Bisection BW	# Links	Degree
Hypercube	$\log_2(p)$	$p/2$	$(p/2)\log_2(p)$	$\log_2(p)$
Cube-3 (p=8)	3	4	12	3
Cube-4 (p=16)	4	8	32	4

Tree; Fat tree



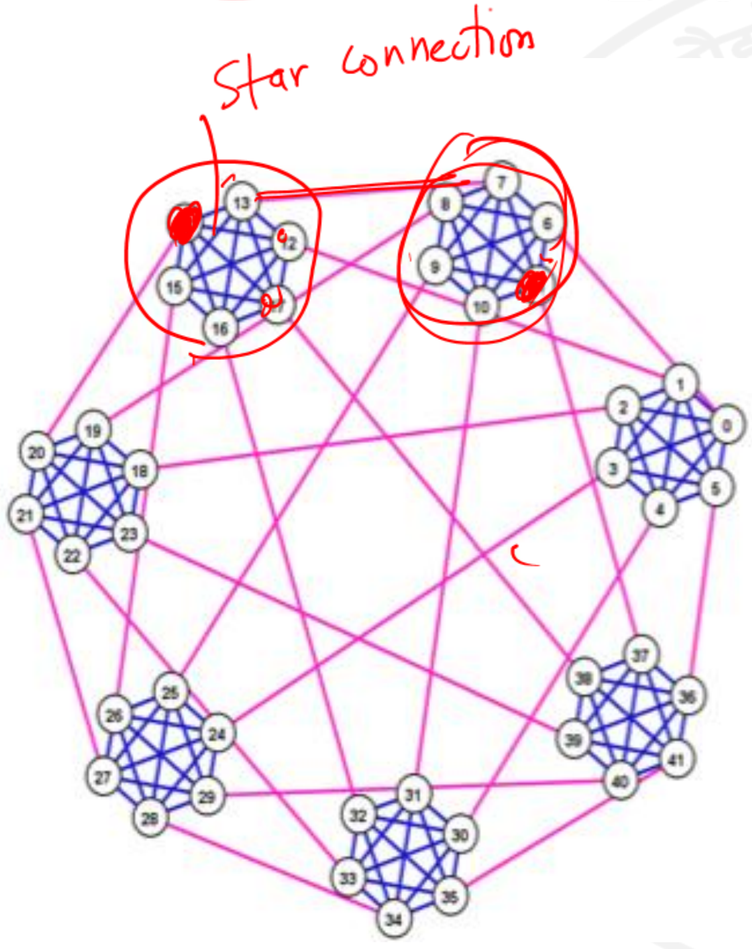
Tree



Fat tree

Topology	Diameter	Bisection BW	# Links	Degree
Binary tree	$2 \log_2(p)$	1	2(p-1)	3

Dragonfly



Fractal structure

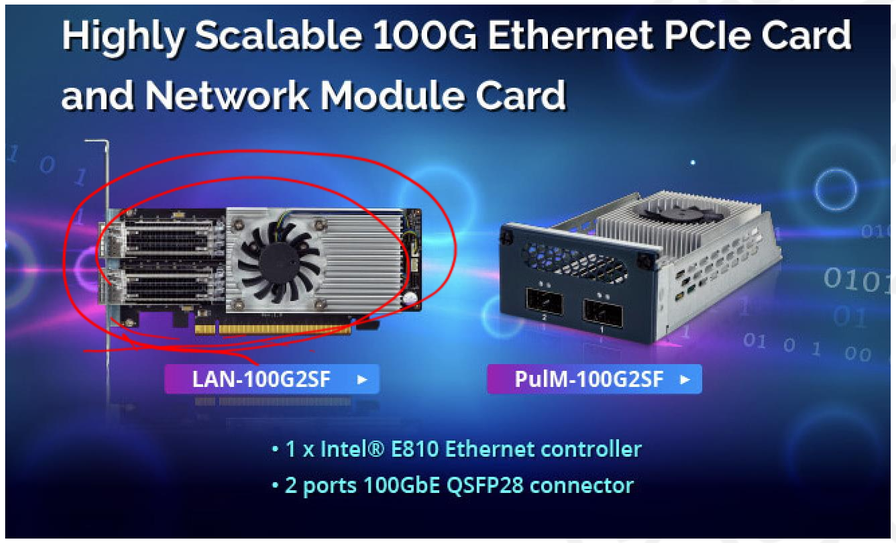
Other networks: Butterfly

Infiband switch

Wikipedia

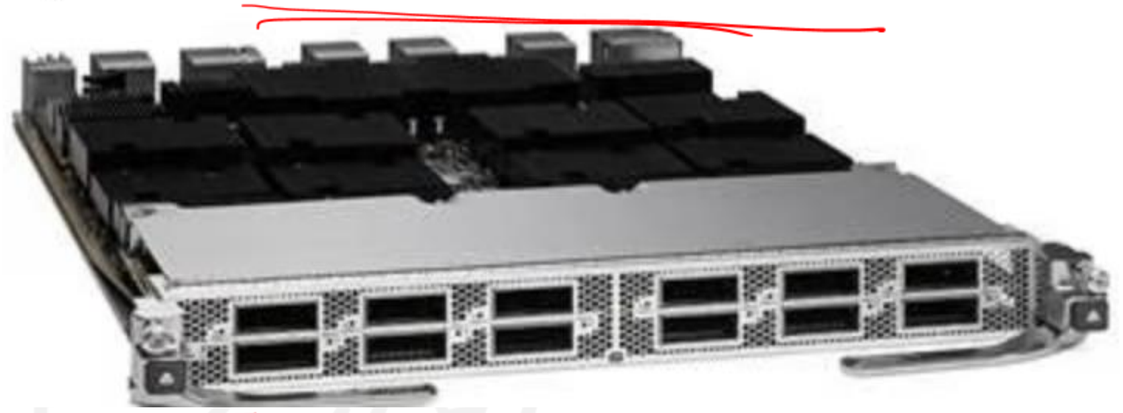
Mellanox/Nvidia

100G Ethernet Switch



https://www.ieiworld.com/en/news/con_show.php?cid=1243

Figure 1. Cisco Nexus 7700 F3-Series Module



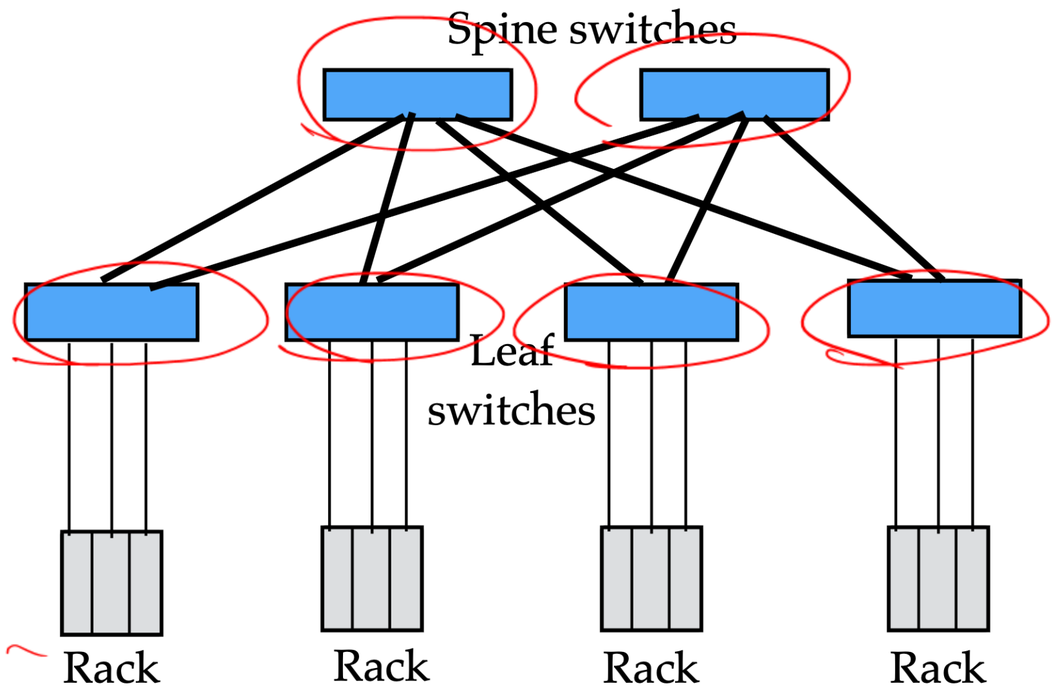
https://www.cisco.com/c/en/us/products/collateral/switches/nexus-7000-series-switches/data_sheet_c78-728423.html

Characteristics

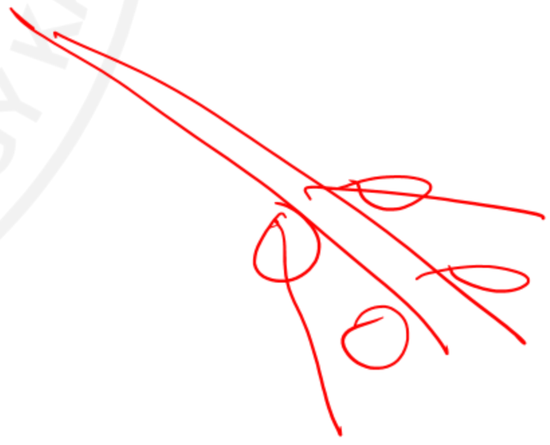
		SDR	DDR	QDR	FDR10	FDR	EDR	HDR	NDR	XDR
Signaling rate (Gbit/s)		2.5	5	10	10.3125	14.0625[6]	25.78125	50	100	250
Theoretical effective throughput (Gb/s)[7]	**for 1 link**	2	4	8	10	13.64	25	50	100	250
	for 4 links	8	16	32	40	54.54	100	200	400	1000
	for 8 links	16	32	64	80	109.08	200	400	800	2000
	for 12 links	24	48	96	120	163.64	300	600	1200	3000
Encoding (bits)		8b/10b			64b/66b				t.b.d.	t.b.d.
Adapter latency (μs)[8]		5	2.5	1.3	0.7	0.7	0.5	less?	t.b.d.	t.b.d.
Year[9]		2001, 2003	2005	2007	2011	2011	2014[7]	2017[7]	after 2020	after 2023?

Wikipedia

Connecting the nodes



Fat tree



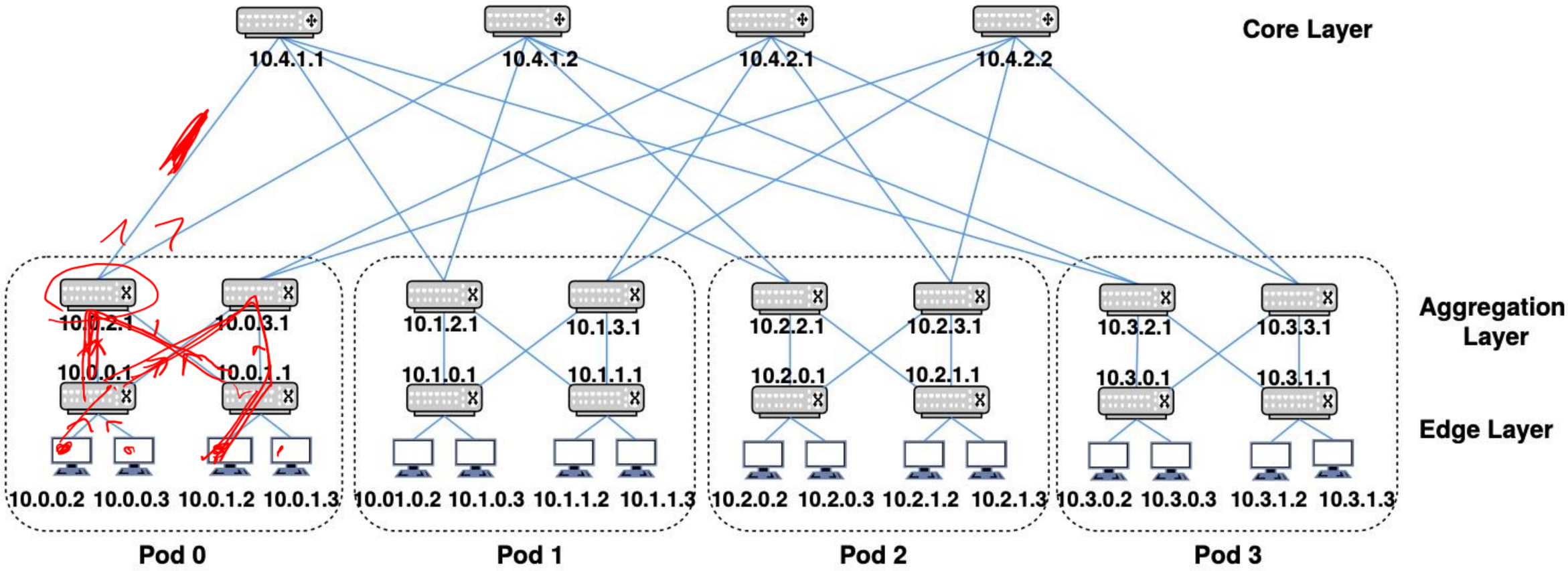


Fig. 3. Fat-Tree topology with $k = 4$.

Each switch has 4 ports.

https://web.engr.oregonstate.edu/~hamdaoui/papers/2018/DC-topology-GC-2018.pdf

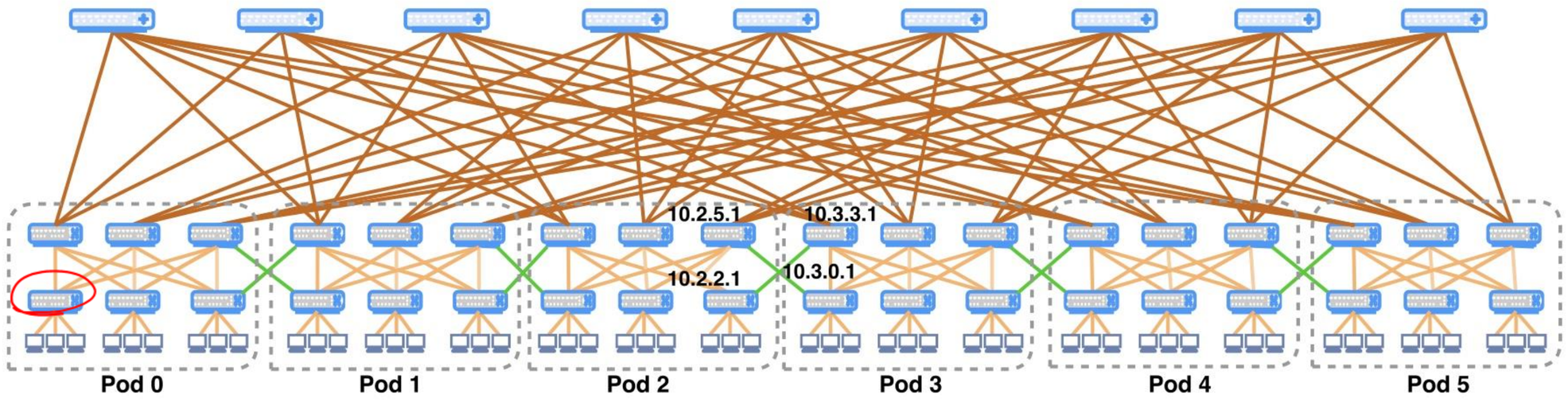
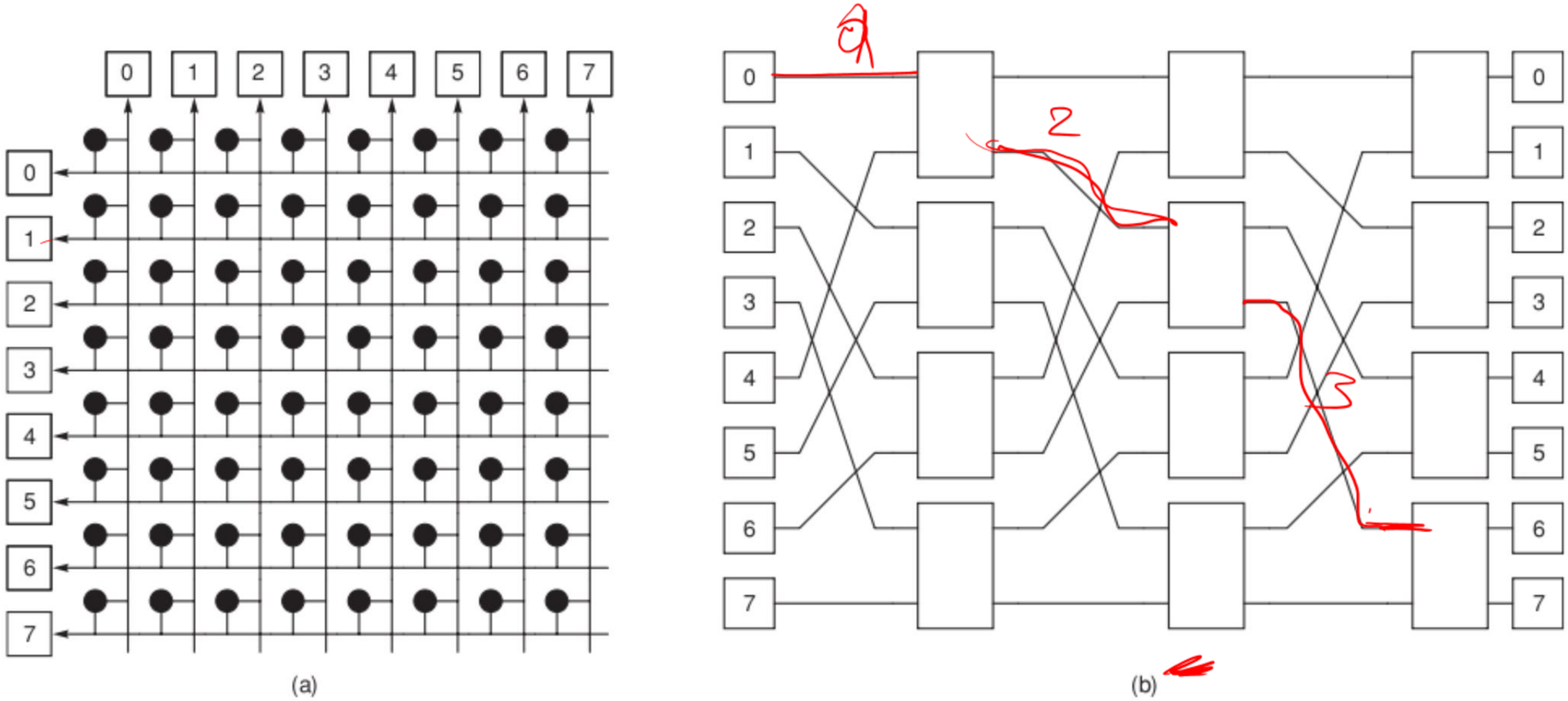


Fig. 5. The proposed `Circulant Fat-Tree` with $k = 6$.

Rethinking Fat-Tree Topology Design for Cloud Data Centers

Jarallah Alqahtani and Bechir Hamdaoui
Oregon State University, Corvallis, Oregon
{alqahtaj,hamdaoui}@eecs.oregonstate.edu



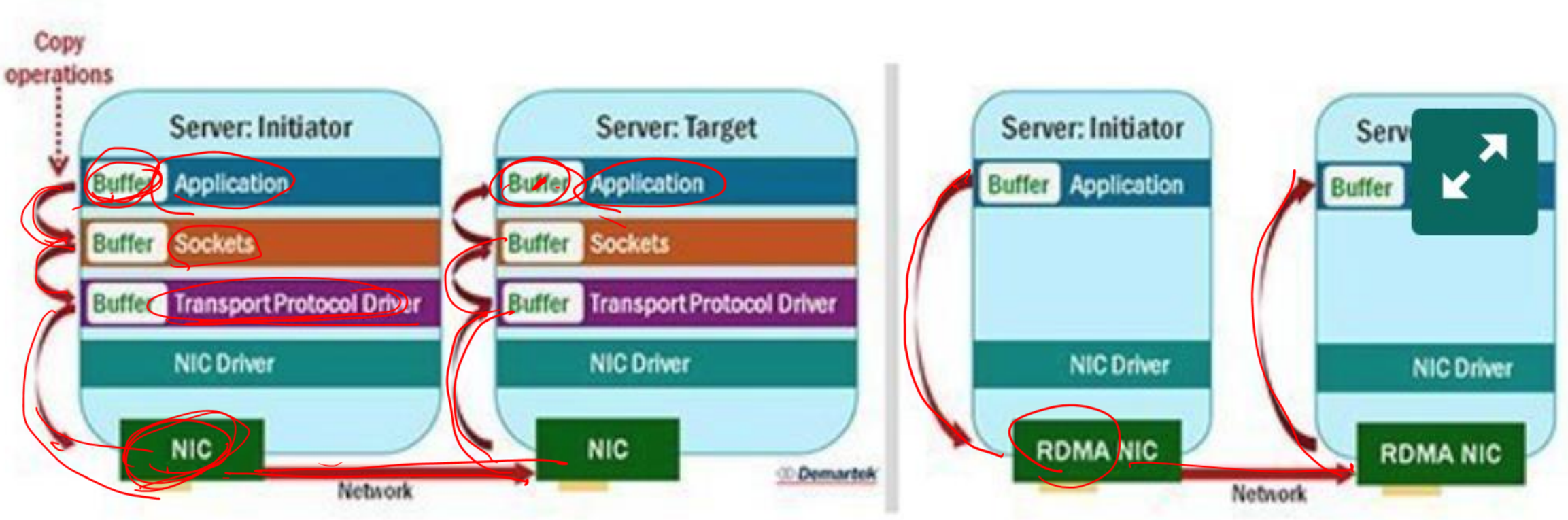
Crossbar Omega

From Hennessy & Patterson

Speeding up comm

RDMA

Remote Direct Memory Access



https://searchstorage.techtarget.com/definition/Remote-Direct-Memory-Access