Foundations of High Performance Computing

Lecture 2: HPC hardware

"Foundation of HPC" course

DATA SCIENCE & SCIENTIFIC COMPUTING

2022-2023 Stefano Cozzini

Agenda

- Why HPC is parallel ?
- Serial Computers
- Moore law/Dennard Scaling
- Parallel computers
- HPC infrastructure
- Parallel components in a cluster

Let us focus on High Performance problem

picture from http://www.f1nutter.co.uk/tech/pitstop.php

Analysis of the parallel solution

FUNCTIONAL PARTITIONING

different people are executing different tasks

DOMAIN DECOMPOSITION

different people are solving the same global task but on smaller subset

HPC

=

PARALLEL COMPUTING

HPC

=

PARALLEL

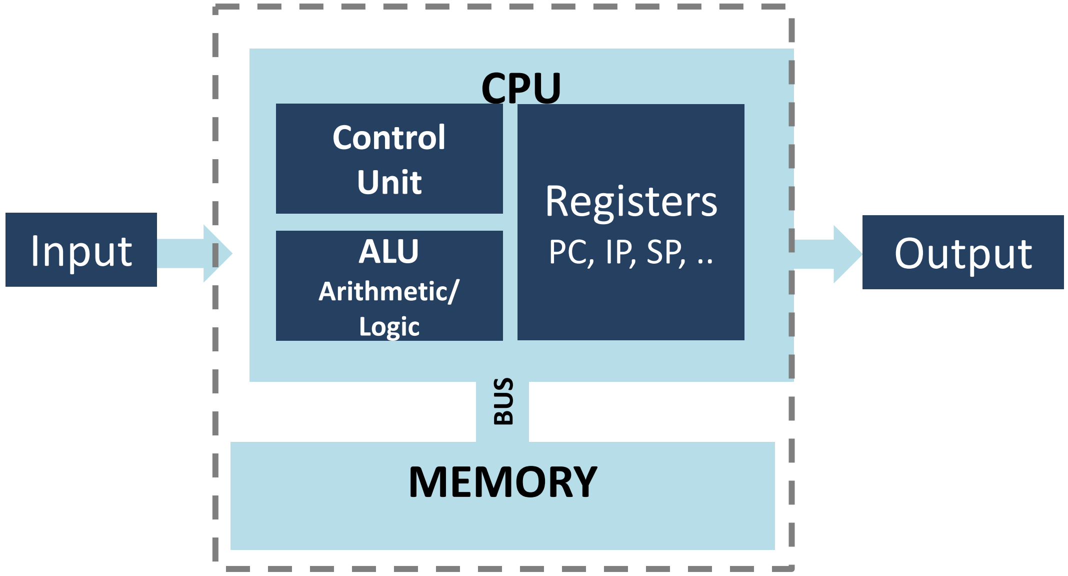
COMPUTERS

Agenda

- Why HPC is parallel ?
- Serial Computers
- Moore law/Dennard Scaling
- Parallel computers
- HPC infrastructure
- Parallel components in a cluster

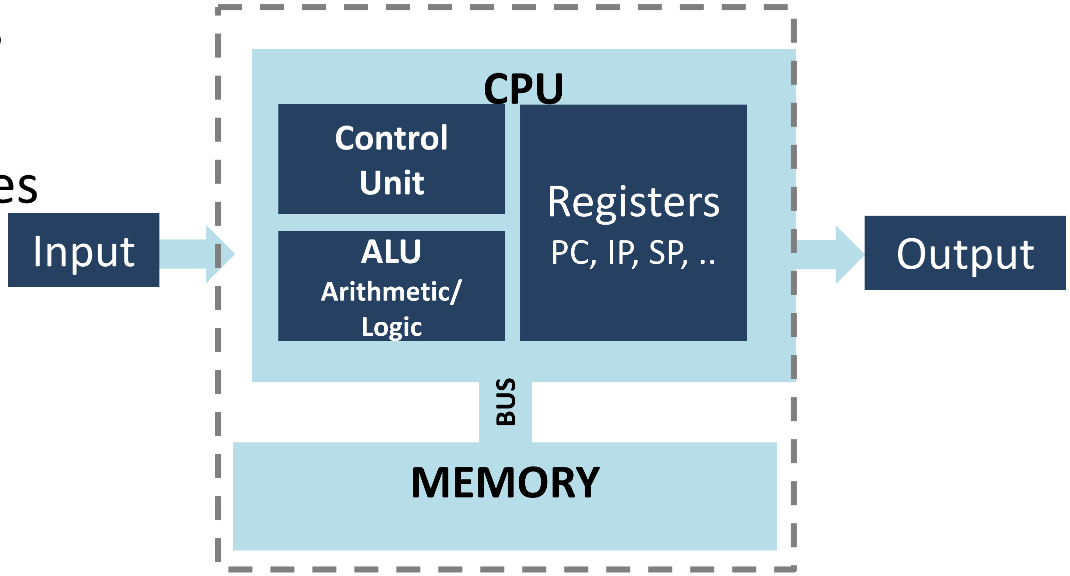
What is a serial computer ?

- Von Neumann architecture (the fundamental model)



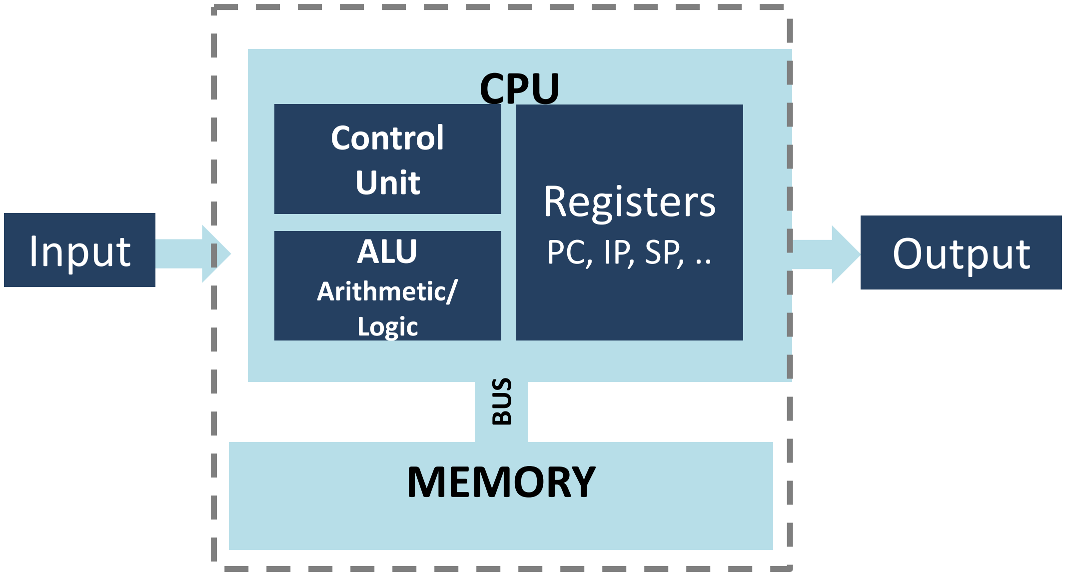
Von Neumann architecture:

- There is only one process unit (CPU)
  - Control Unit: processes instructions
  - ALU: math and logic operations
  - Register: store data



Von Neumann architecture:

- 1 instructions is executed at a time
- memory is “flat”:
  - access on any location has always the same cost
  - access to memory has the same cost than op execution

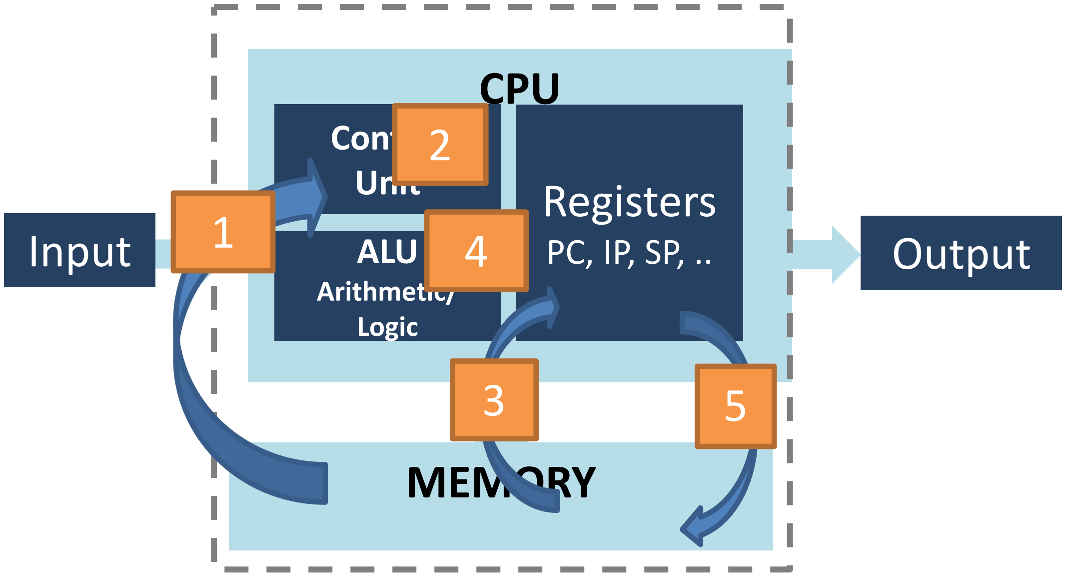


Von Neumann architecture:

5 step WORKFLOW:

1. instruction fetch
2. Instruction decode: determine operation and operands
3. Memory fetch: Get operands from memory
4. Perform operation
5. Write results back

Continue with next instruction



Instruction set architecture (ISA)

- The deeper level accessible to the programmers
- It is the boundary between SW and HW
- The interface between the programmer and the microarchitecture
- Different microarchitectures can have the same ISA (binary Compatible)
- Different generation of microarchitectures can be backward compatible
- For us: x86 instruction set

A very simple operation..

```
void store(double *a, double *b, double *c) {
 *c = *a + *b;
}
```

```
[exact@master ~]$ gcc -O2 -S -o - frammento.c
.file  "frammento.c"
.text
.p2align 4,,15
.globl store
.type       store, @function
store:
.LFB0:
.cfi_startproc
movsd       (%rdi), %xmm0   #load *a to mmx0
addsd       (%rsi), %xmm0   # load b and add to *a
movsd       %xmm0, (%rdx)  # store to C
ret
.cfi_endproc
.LFE0:
.size store, .-store
.ident      "GCC: (GNU) 4.4.7 20120313 (Red Hat 4.4.7-4)"
.section    .note.GNU-stack,"",@progbits
```

Agenda

- Why HPC is parallel ? ✓
- Serial Computers ✓
- Moore law/Dennard Scaling
- Parallel computers
- HPC infrastructure
- Parallel components in a cluster

Moore Law

- Typically stated as: “Performance doubles every X months”
- Actually, closer to: “Number of transistors per unit cost doubles every X months”

The original Moore Law

The complexity for minimum component costs has increased at a rate of roughly a factor of two per year. […]

Over the longer term, the rate of increase is a bit more uncertain, although there is no reason to believe it will not remain nearly constant for at least 10 years.

-- Gordon Moore, Electronics, 1965

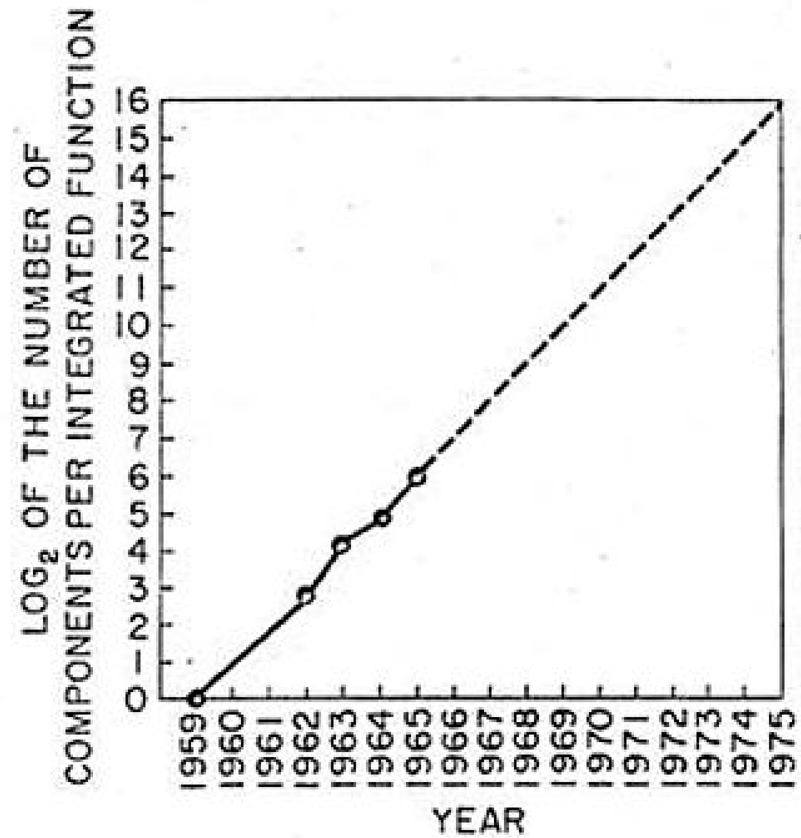


Fig. 2 Number of components per Integrated function for minimum cost per component extrapolated vs time.

Why is Moore's Law connected with processor performance?

Dennard Scaling: From Moore's Law to performance

- "Power density stays constant as transistors get smaller"

  Robert H. Dennard, 1974

- Intuitively:

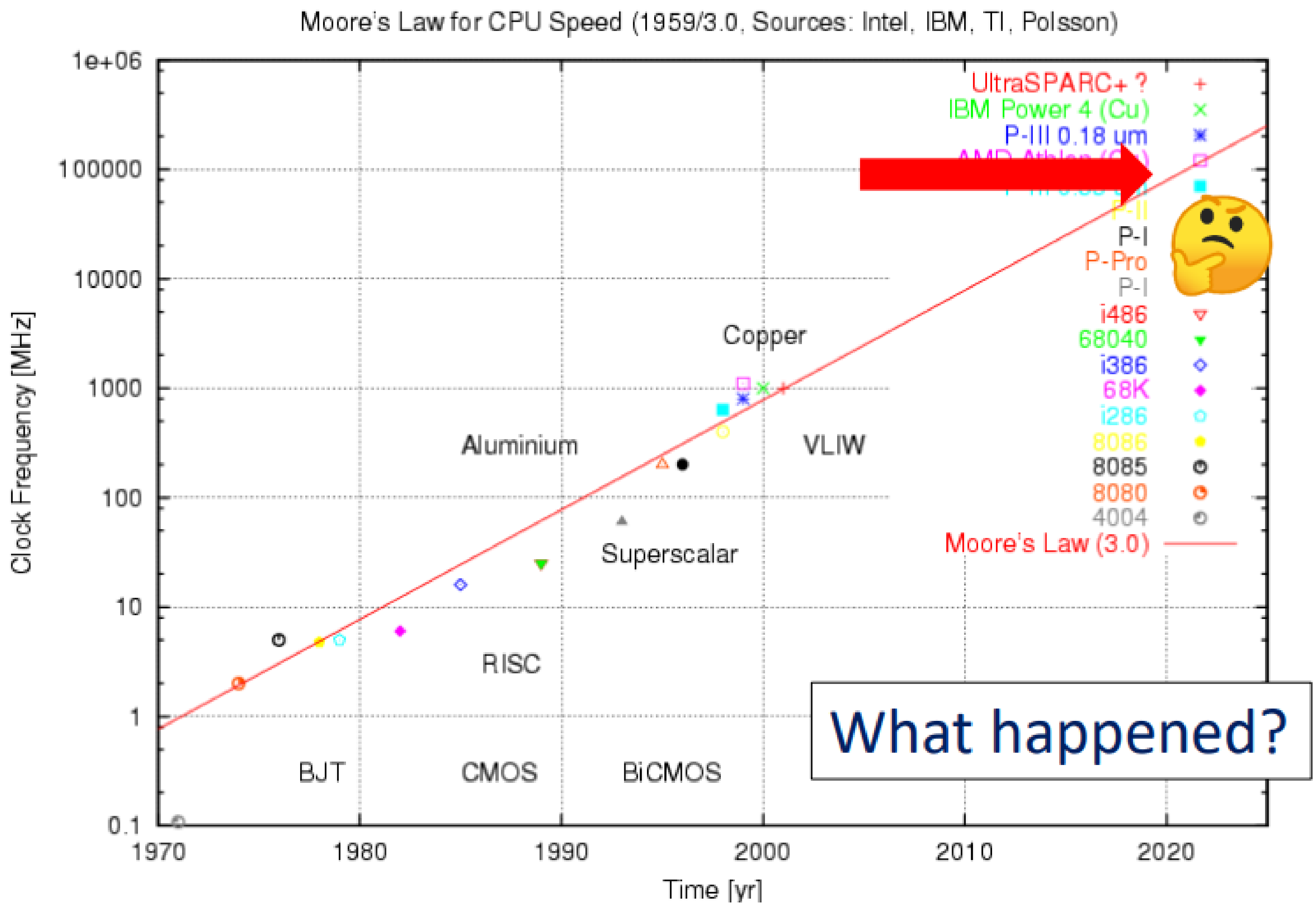
  Smaller transistors → shorter propagation delay →faster frequency

  Smaller transistors → smaller capacitance → lower voltage

  $$Power \propto Capacitance \times Voltage^2 \times Frequency$$

Moore's law → Faster performance @ Constant power!

Single-core performance scaling



What happened?

a little bit more accurate processor power consumption

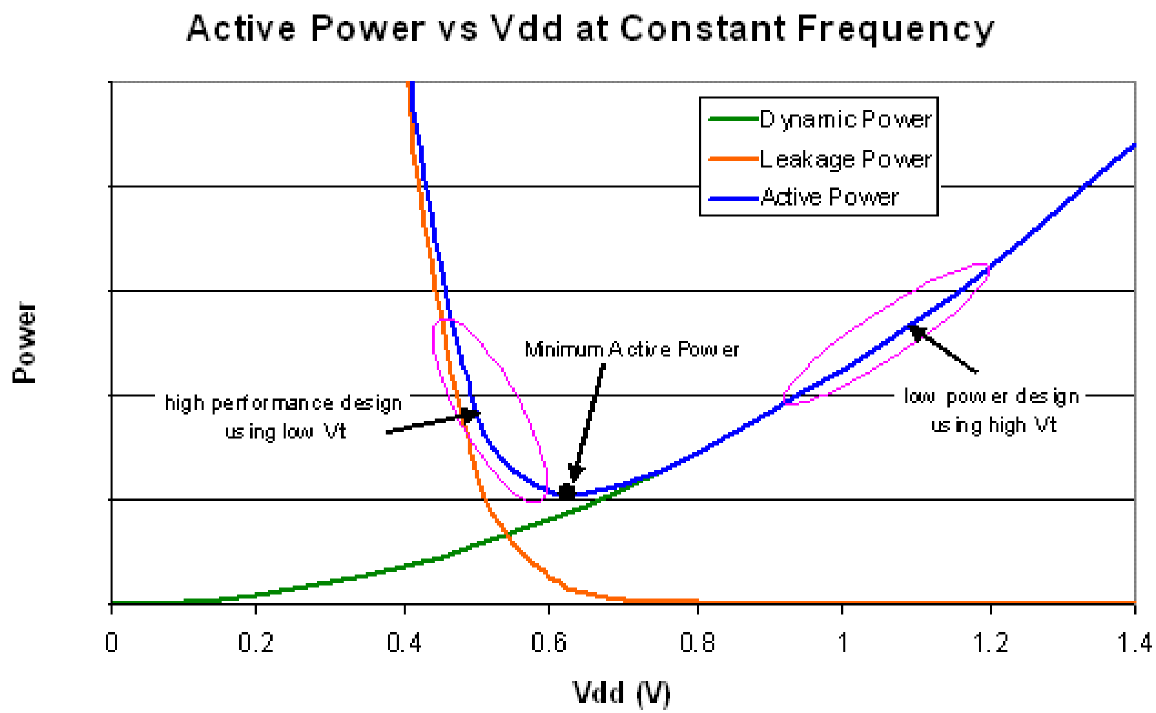
$Power =$

$ActiveTransistors \times Capacitance \times Voltage^2 \times Frequency$

D(Dynamic power)

$+ Voltage \times Leakage$

(Static power)



Both Capacitance and voltage do not scale anymore..

Capacitance: Gate-oxide technology

Voltage: leakage

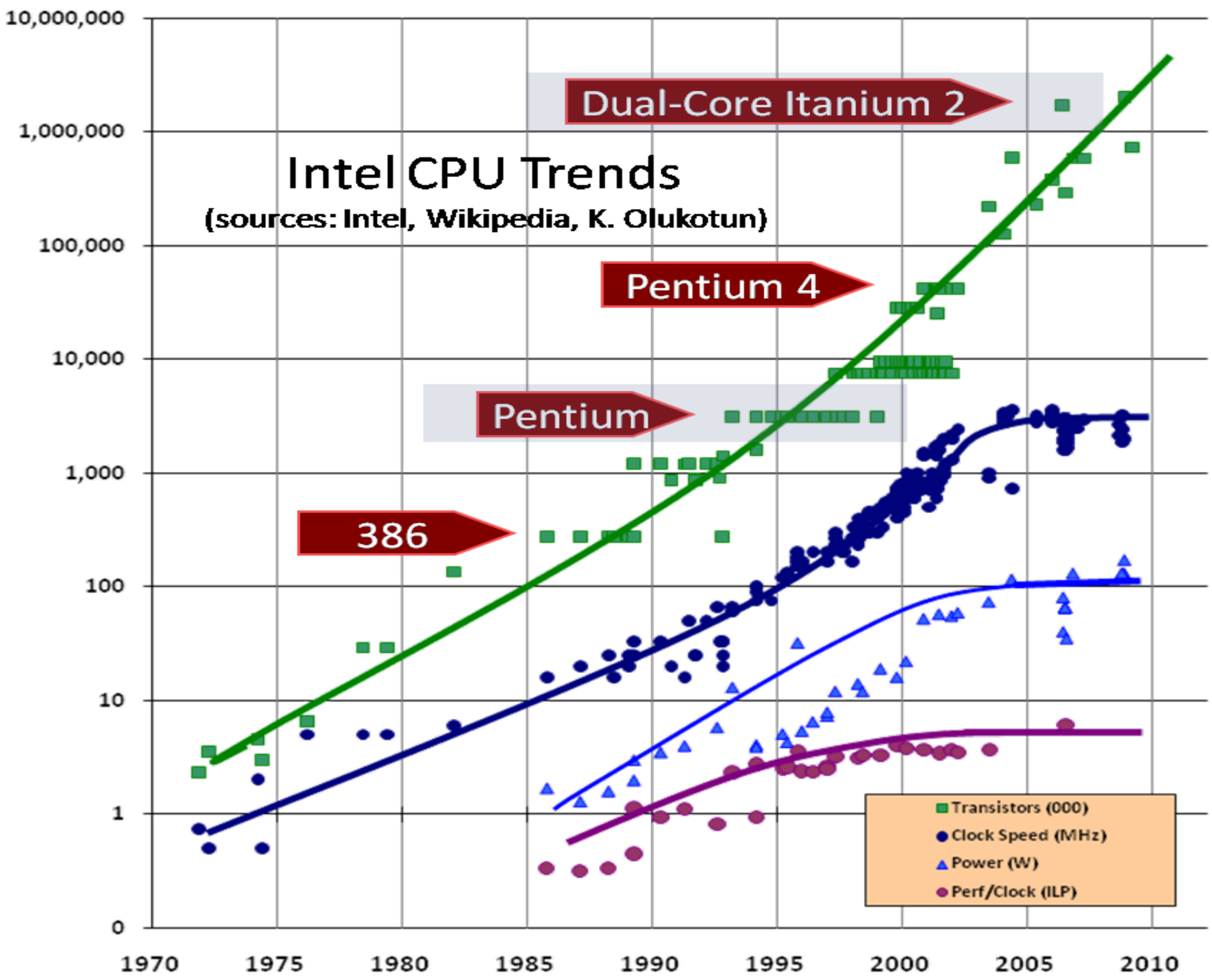
[Picture taken from: Integrated Power Management, Leakage Control and Process Compensation Technology for Advanced Processes (design-reuse.com)]

End of Dennard Scaling

- Even with smaller transistors, we cannot continue reducing power..

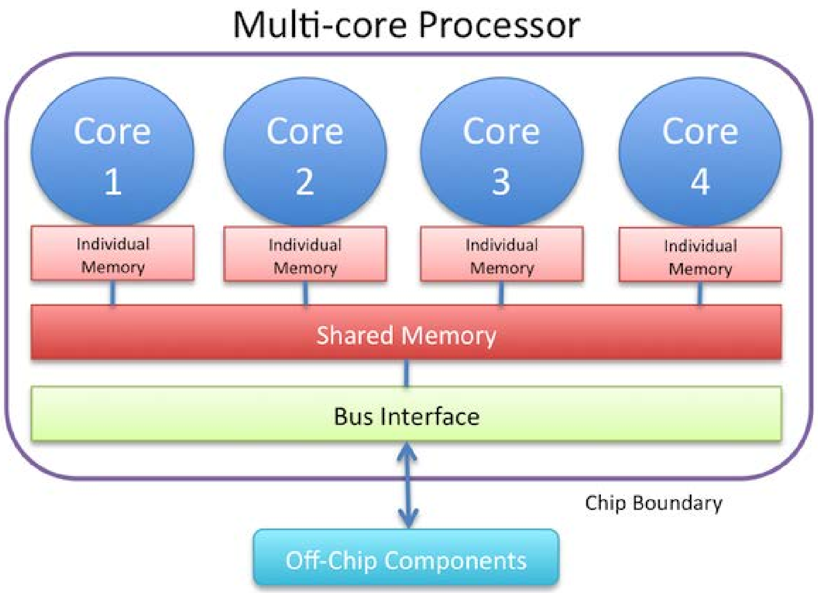
- And now ?
- 2 options:
  - Increase power (when increase frequency)
  - Stop frequency scaling…

(original) Moore law still valid…



CPU are multicore processor

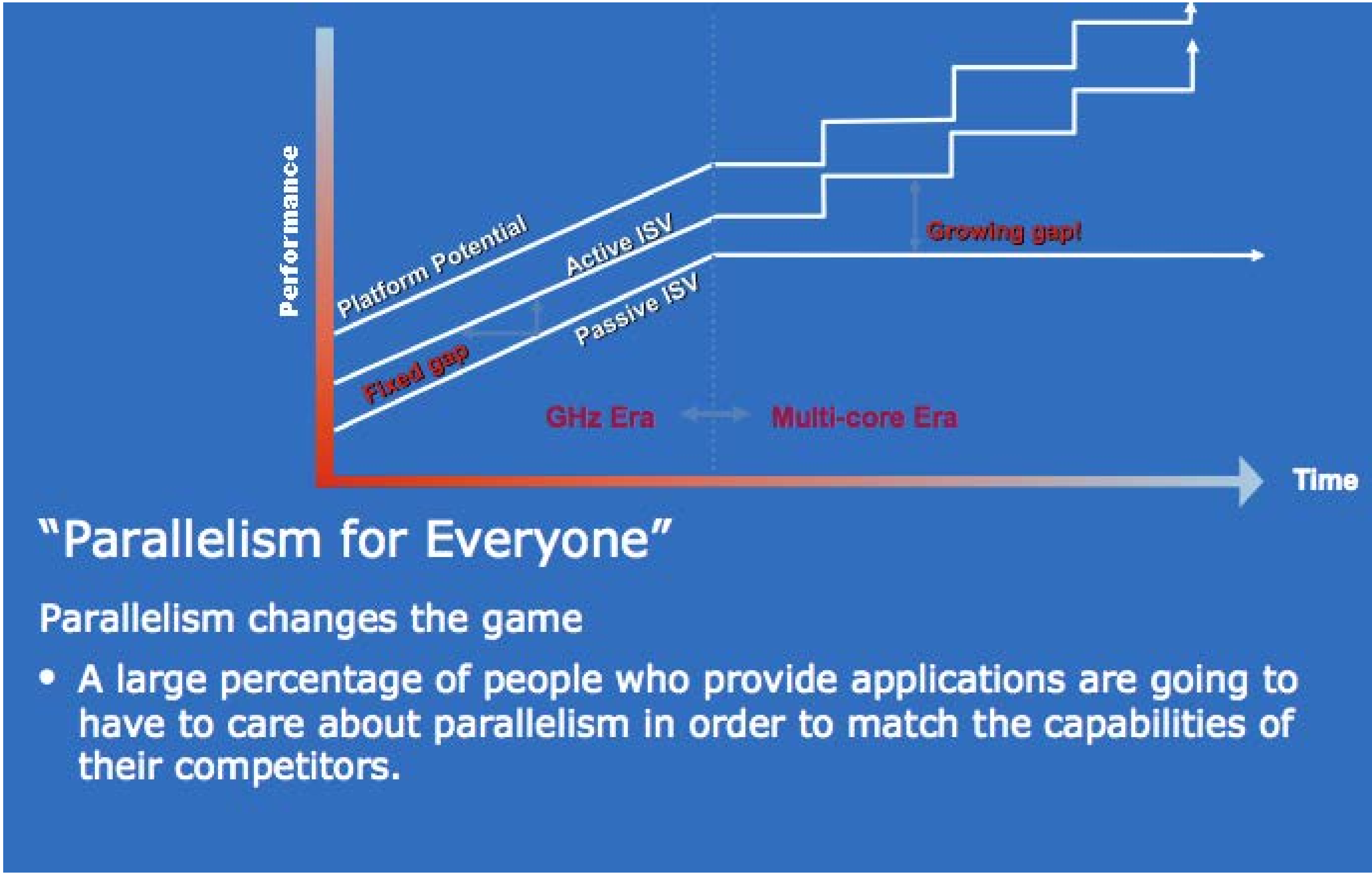
- Because of power, heat dissipation, etc increasing tendency to actually lower clock frequency but pack more computing cores onto a chip.
- These cores will share some resources, e.g. memory, network, disk, etc but are still capable of independent calculations



No more “free lunch” from 2006…

- Single core performance scaling ended.
  - Performance no longer depend on hardware scaling (i.e increase in frequency)
- Solution 1: the software solution
  - Write efficient software to make the efficient use of hardware resources
  - “Performance engineering” software, using hardware knowledge

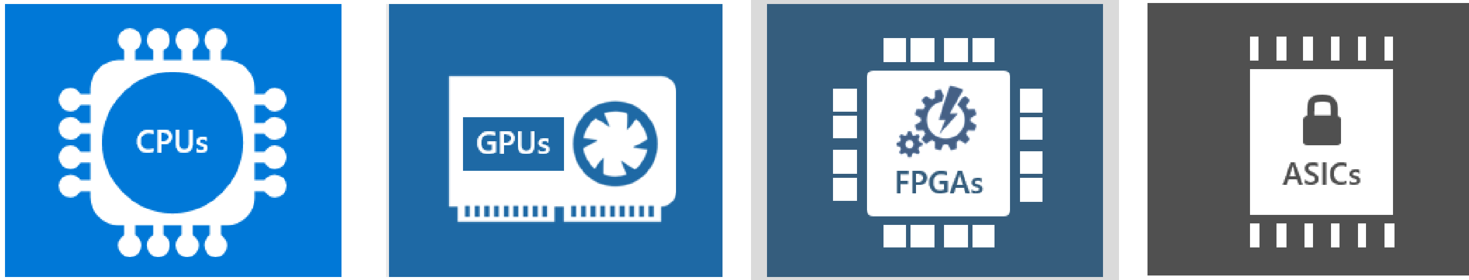
An old picture from Intel..



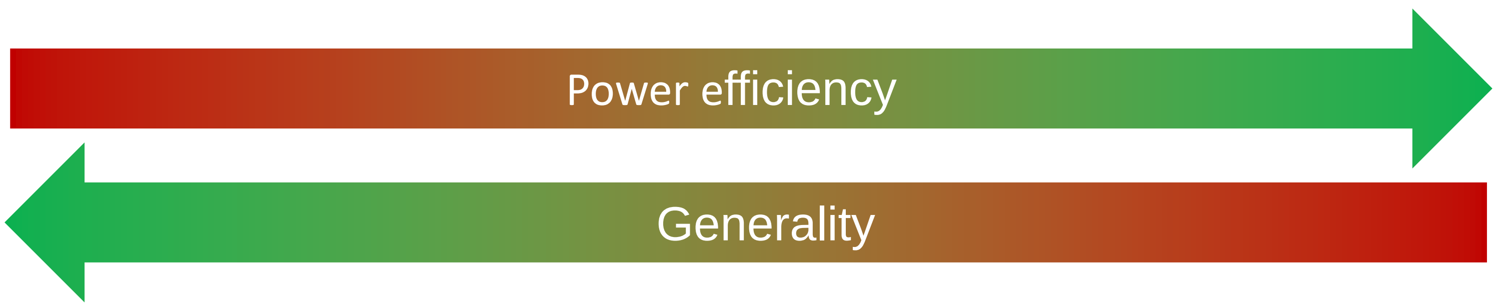
No more “free lunch” from 2006…

- Solution 2: specialized architectural solution
  - Chip space is now cheap, but power is expensive
  - Stop depending on more complex general-purpose cores
  - Use space to build heterogeneous systems, with compute engines well-suited for each application

Hardware accelerators

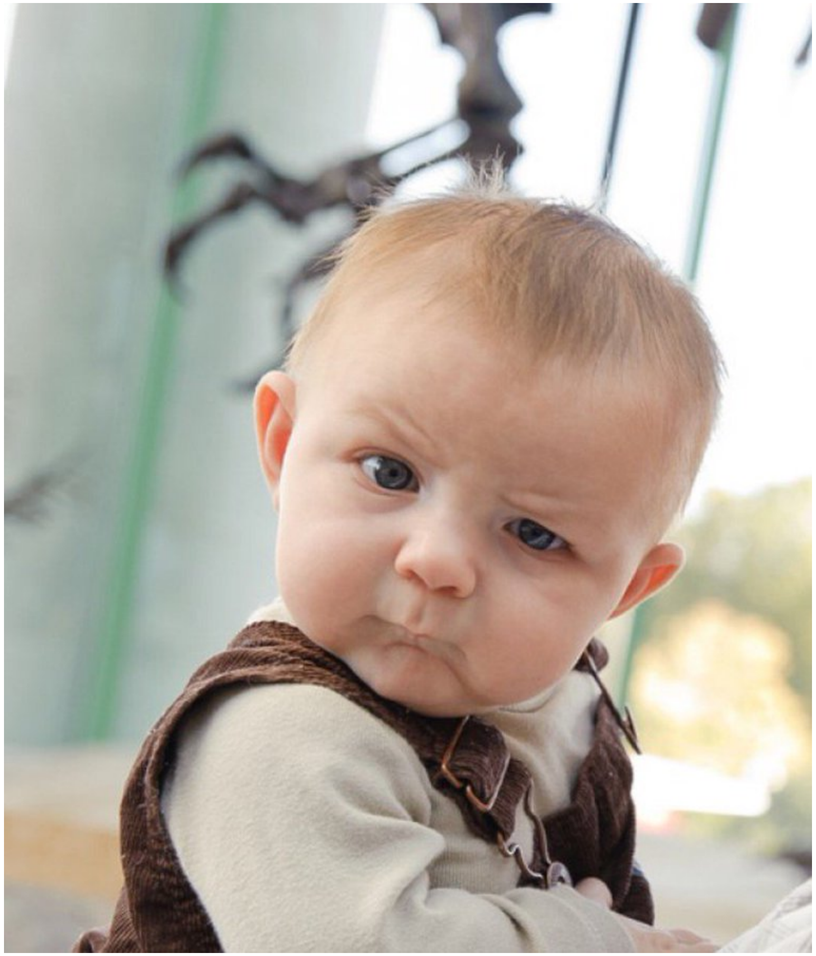


	CPU	GPU	FPGA	ASIC
Power efficiency	$5\ \frac{Gflops}{W}$	$20\ \frac{Gflops}{W}$	$70\ \frac{Gflops}{W}$	$> 70\ \frac{Gflops}{W}$



Images: https://www.microsoft.com/en-us/research/video/inside-microsoft-fpga-based-configurable-cloud/
Numbers: https://h2rc.cse.sc.edu/2015/burger_keynote.pdf

Does still exist serial computer ?



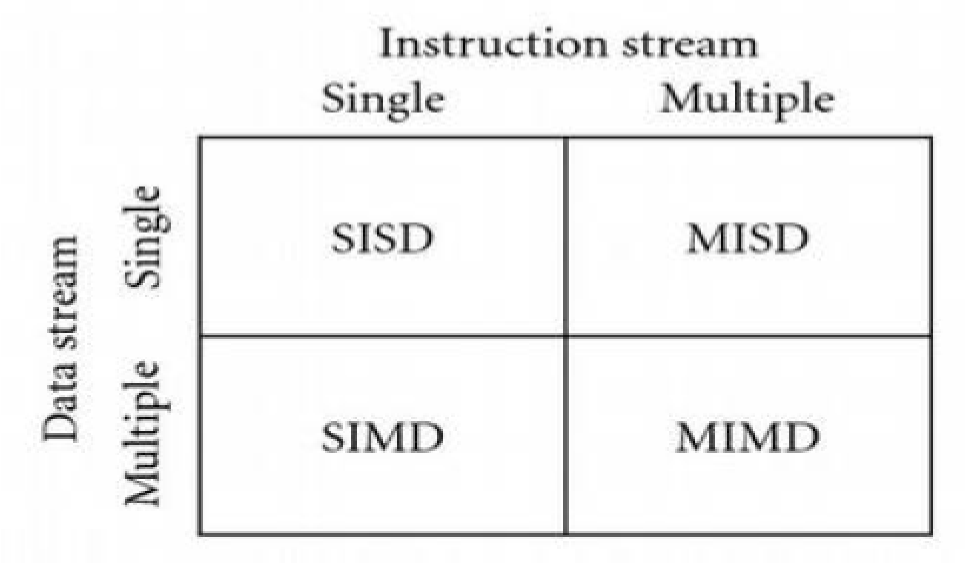
Agenda

- Why HPC is parallel ? ✓
- Serial Computers ✓
- Moore law/Dennard Scaling ✓
- Parallel computers
- HPC infrastructure
- Parallel components in a cluster

PARALLELISM IS EVERYHERE even in your laptop..

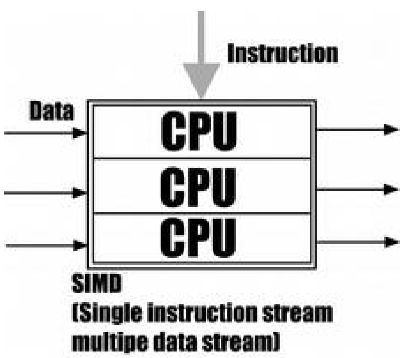
Parallel Computers

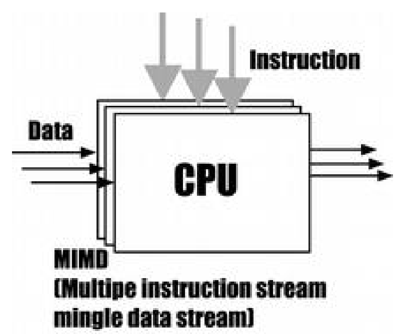
- Flynn Taxonomy (1966): may help us in classifying them:
  - Data Stream
  - Instruction Stream



Comments

- Flynn taxonomy does not help too much nowadays with modern HPC infrastructure
  - CPU and computers are changed too much in the last 50 years
- However, SIMD and MIMD concepts are still used HPC hardware

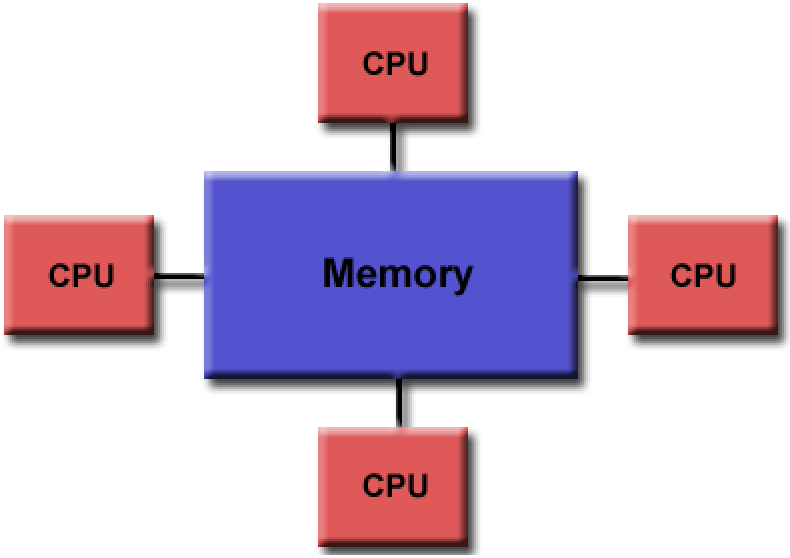




What about memory ?

- In the old time the simplest and most useful way to classify modern parallel computers was by their memory model:
  - SHARED MEMORY
  - DISTRIBUTED MEMORY

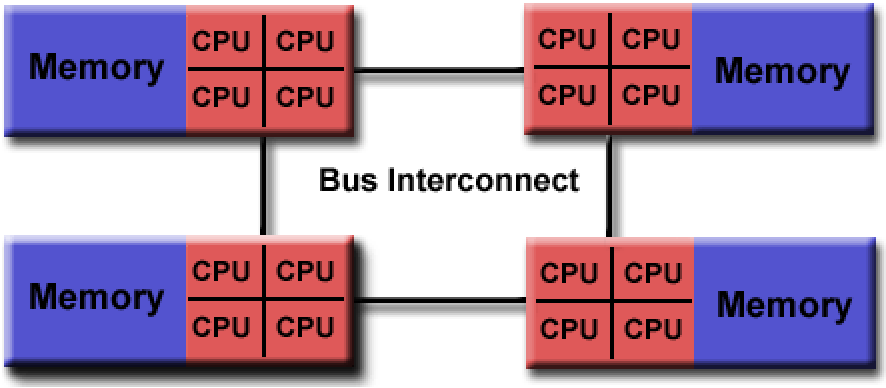
Shared memory: UMA



Uniform memory access (UMA): Each processor has uniform access to memory. Also known as symmetric multiprocessors (SMP)

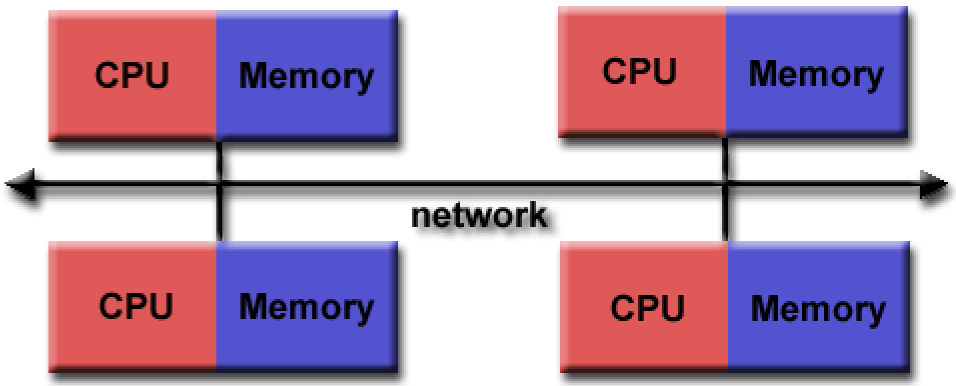
Shared memory: NUMA

Non-uniform memory access (NUMA): Time for memory access depends on location of data. Local access is faster than non-local access.



Distributed memory

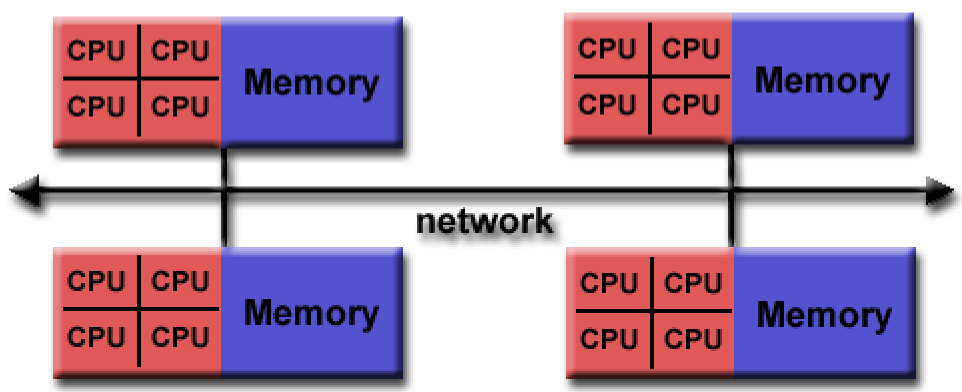
- Distributed memory
  - each processor has its own local memory. Must do message passing to exchange data between processors



ARE THESE MACHINES STILL AVAILABLE ?

Hybrid approach

The shared memory component is shared memory
The distributed memory component is the networking of multiple shared memory which know only about their own memory - not the memory on another machine.

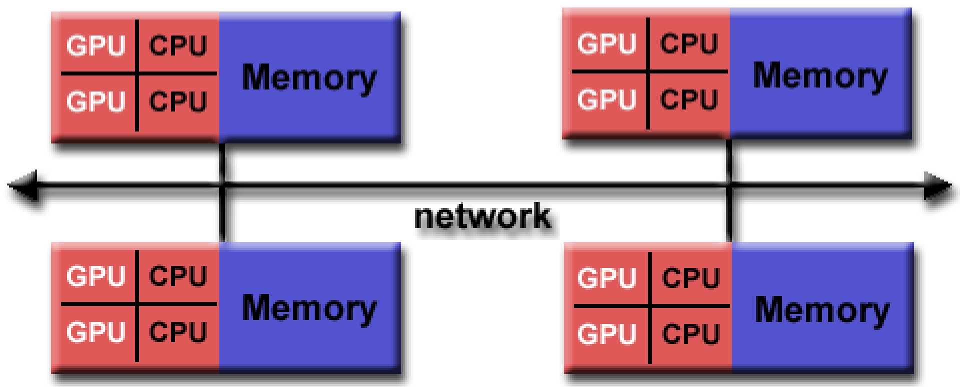


Agenda

- Why HPC is parallel ? ✓
- Serial Computers ✓
- Moore law/Dennard Scaling ✓
- Parallel computers ✓
- HPC infrastructure
- Parallel components in a cluster

Modern HPC infrastructures

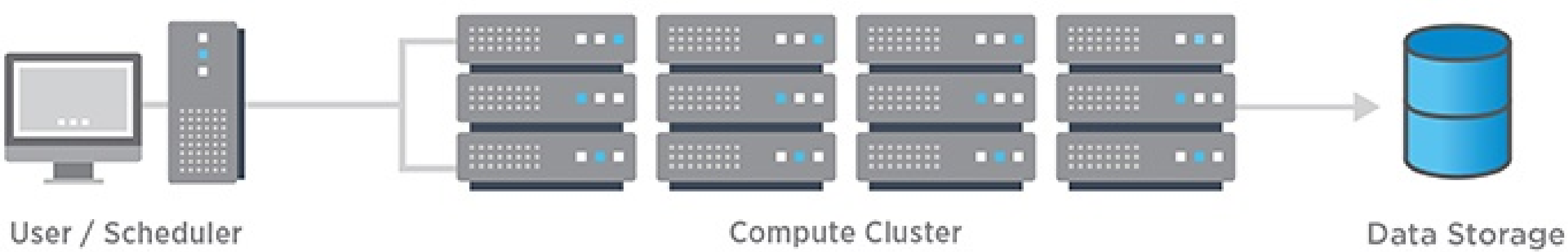
- Cluster of nodes (shared memory)



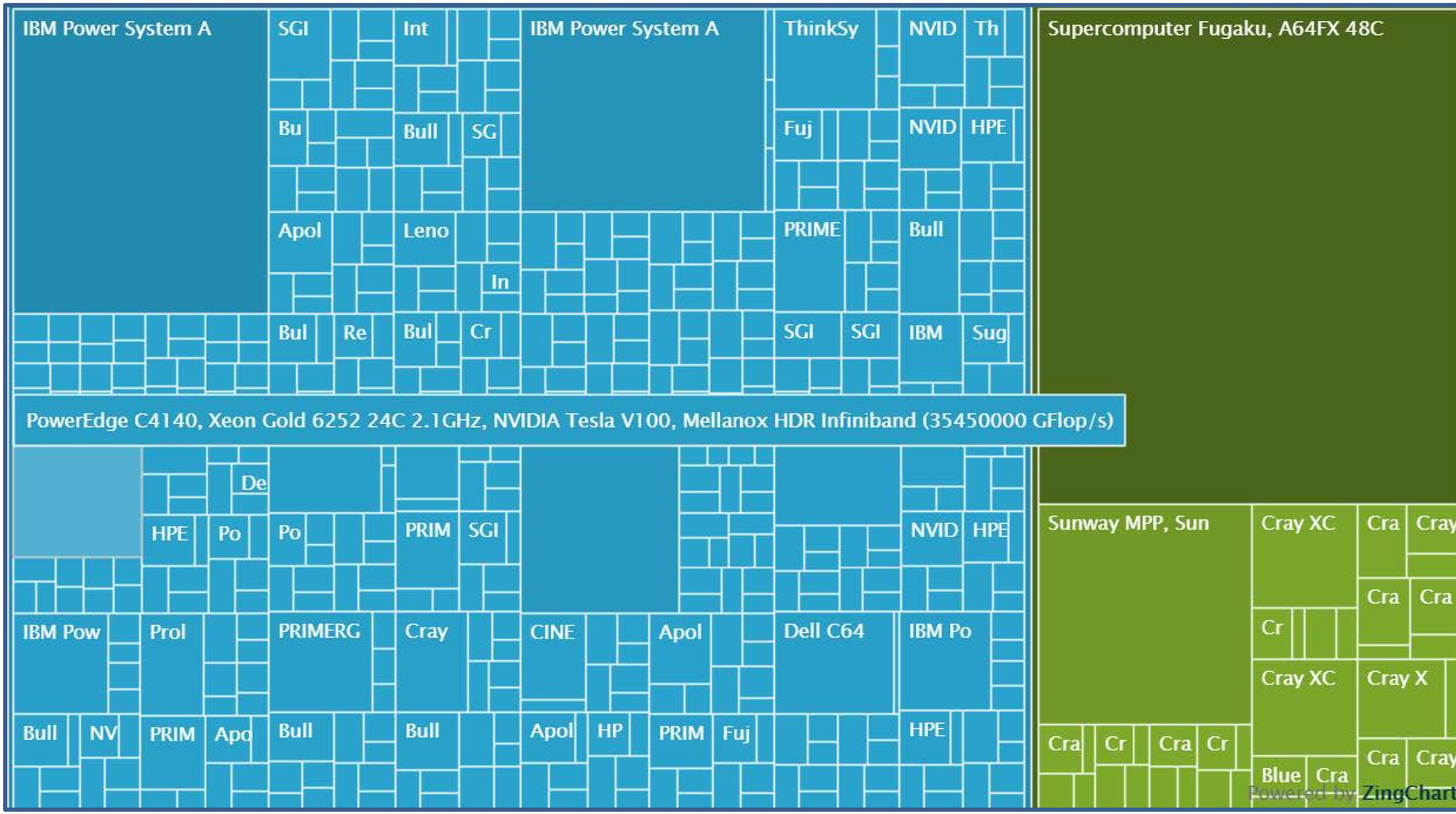
- Hybrid distributed/shared approach from memory point of view

Essential component of a cluster

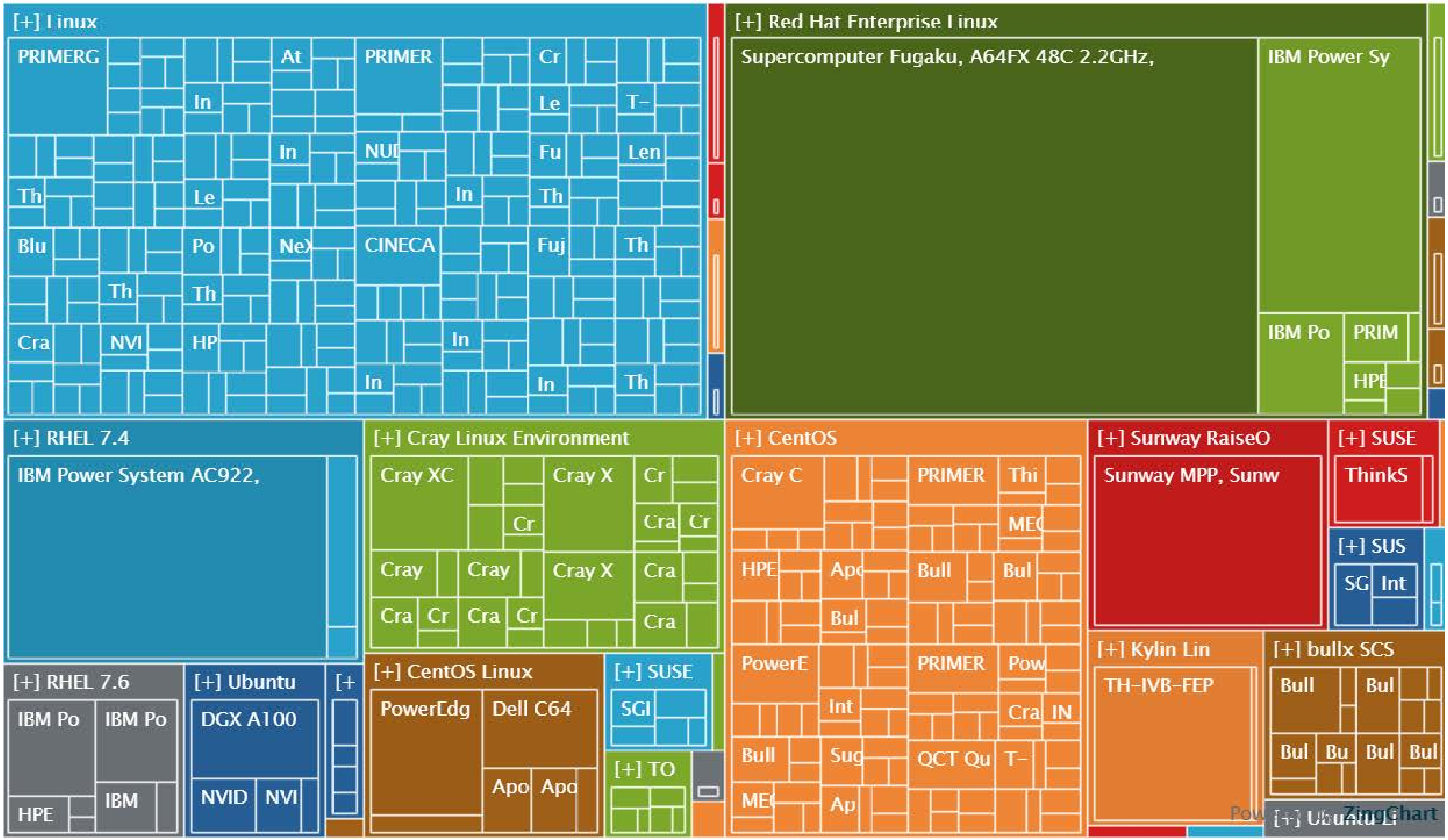
- Several computers (nodes)
  - often in special cases (1U) for easy mounting in a rack
- One or more networks (interconnects) to hook the nodes together
- Some kind of storage
- A login/access node..



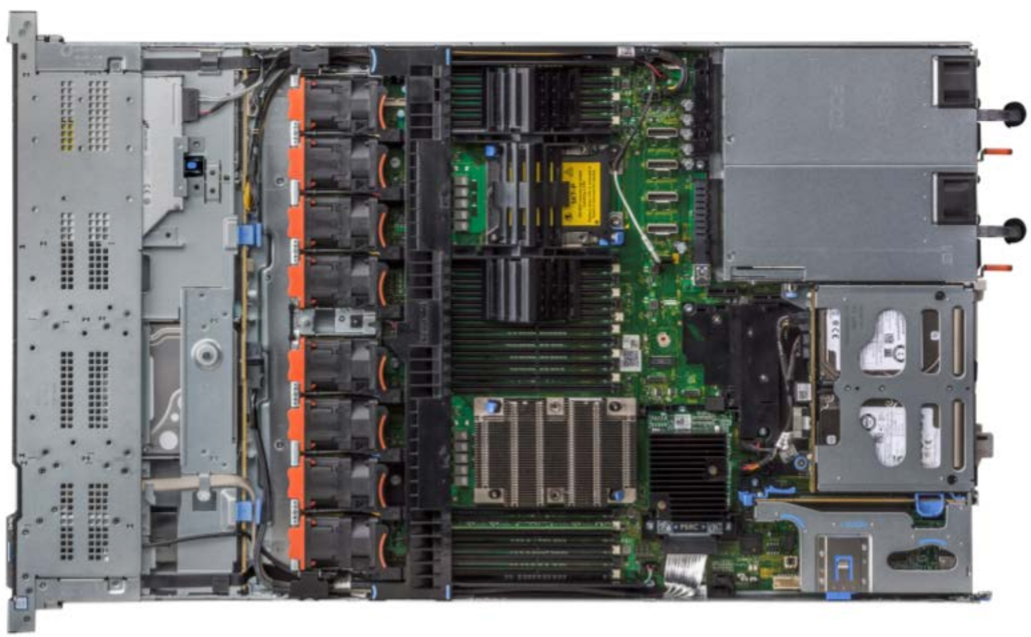
Even supercomputers are clusters !



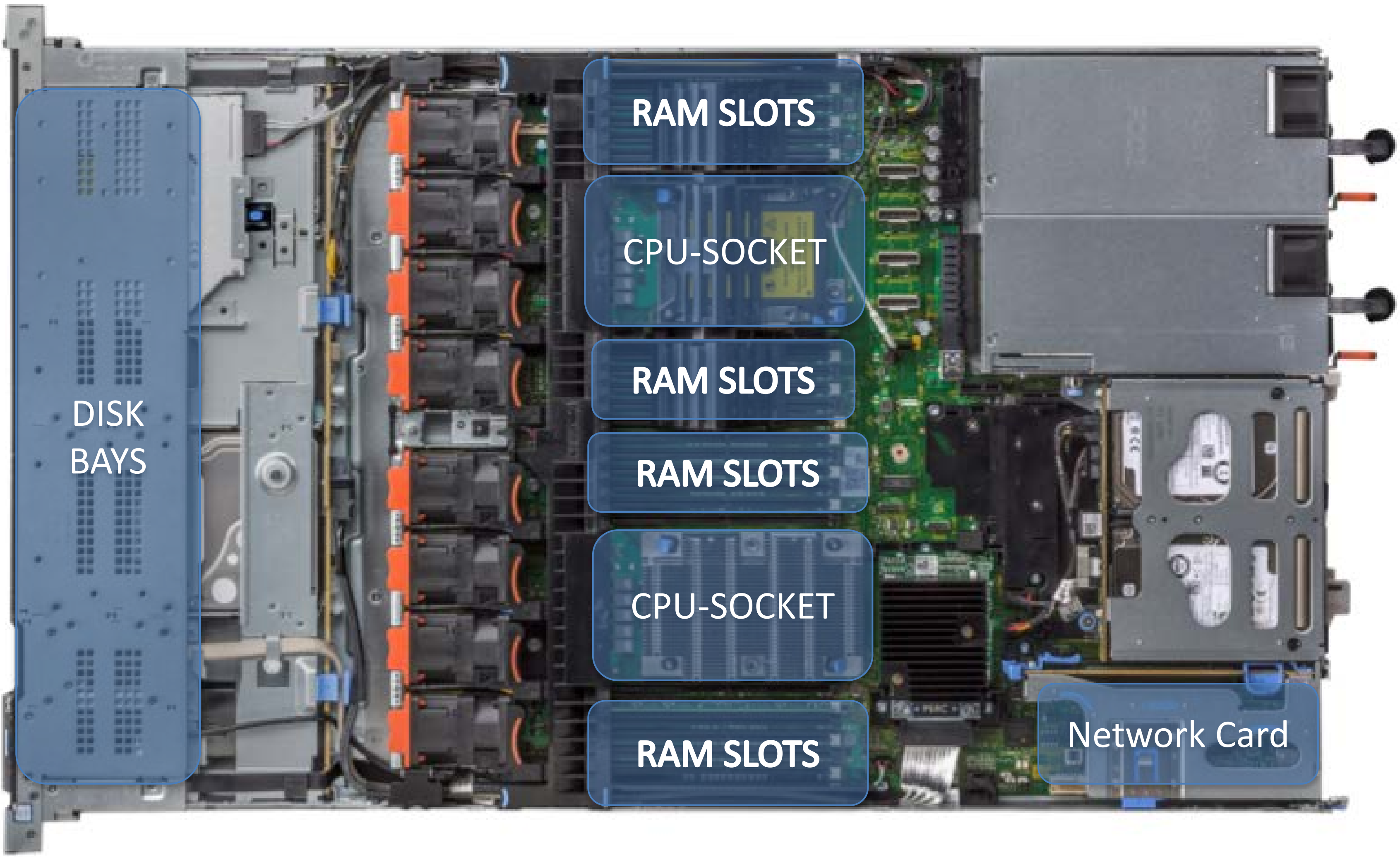
And cluster speaks the same language: LINUX



Modern 1U computing nodes

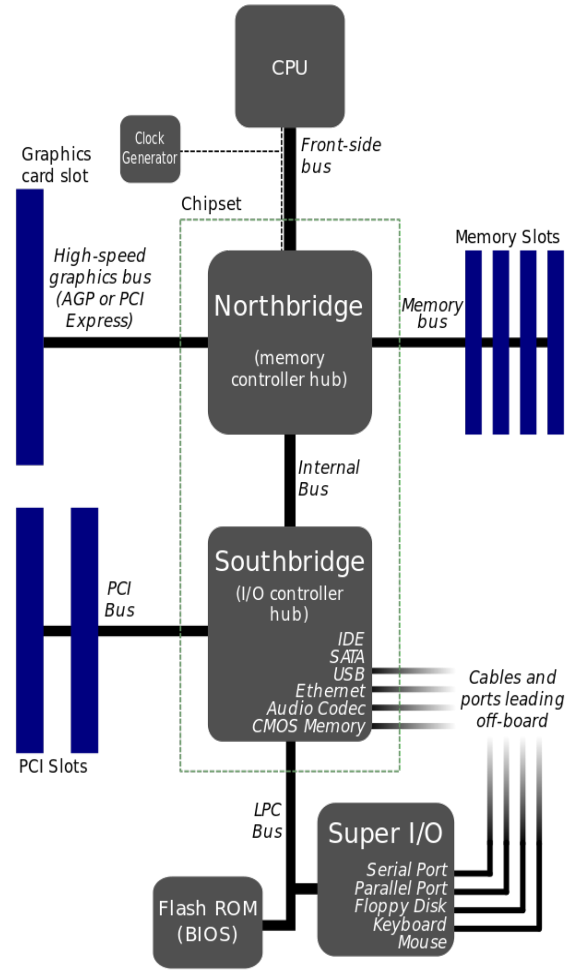


What does one node contain exactly ?



standard modern architecture

- All data communication from one CPU to another must travel over the same bus used to communicate with the Northbridge.
- All communication with RAM must pass through the Northbridge.
- Communication between a CPU and a device attached to the Southbridge is routed through the Northbridge.

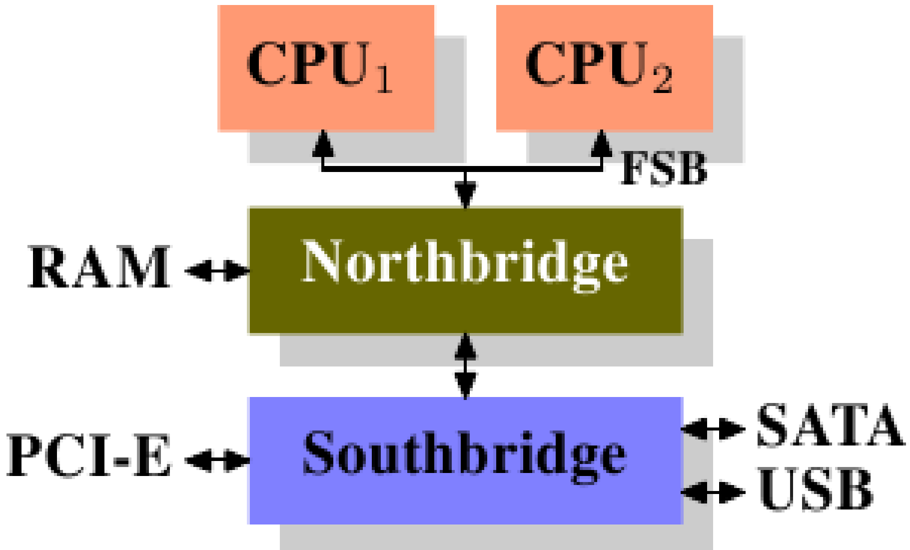


standard multisocket architecture

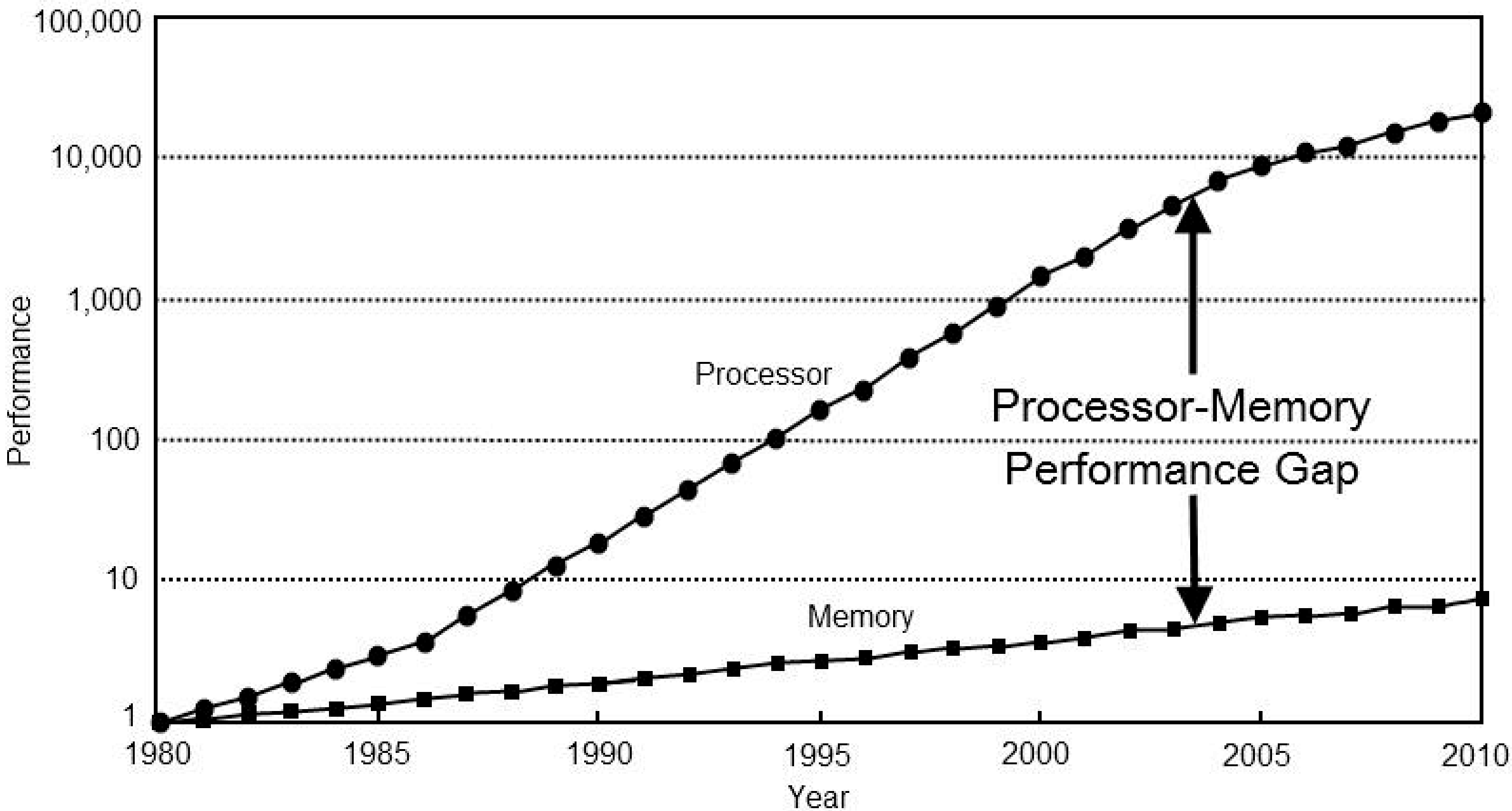
- Characteristics:

– more than one CPU !

– 64 bit address space



Memory wall problem

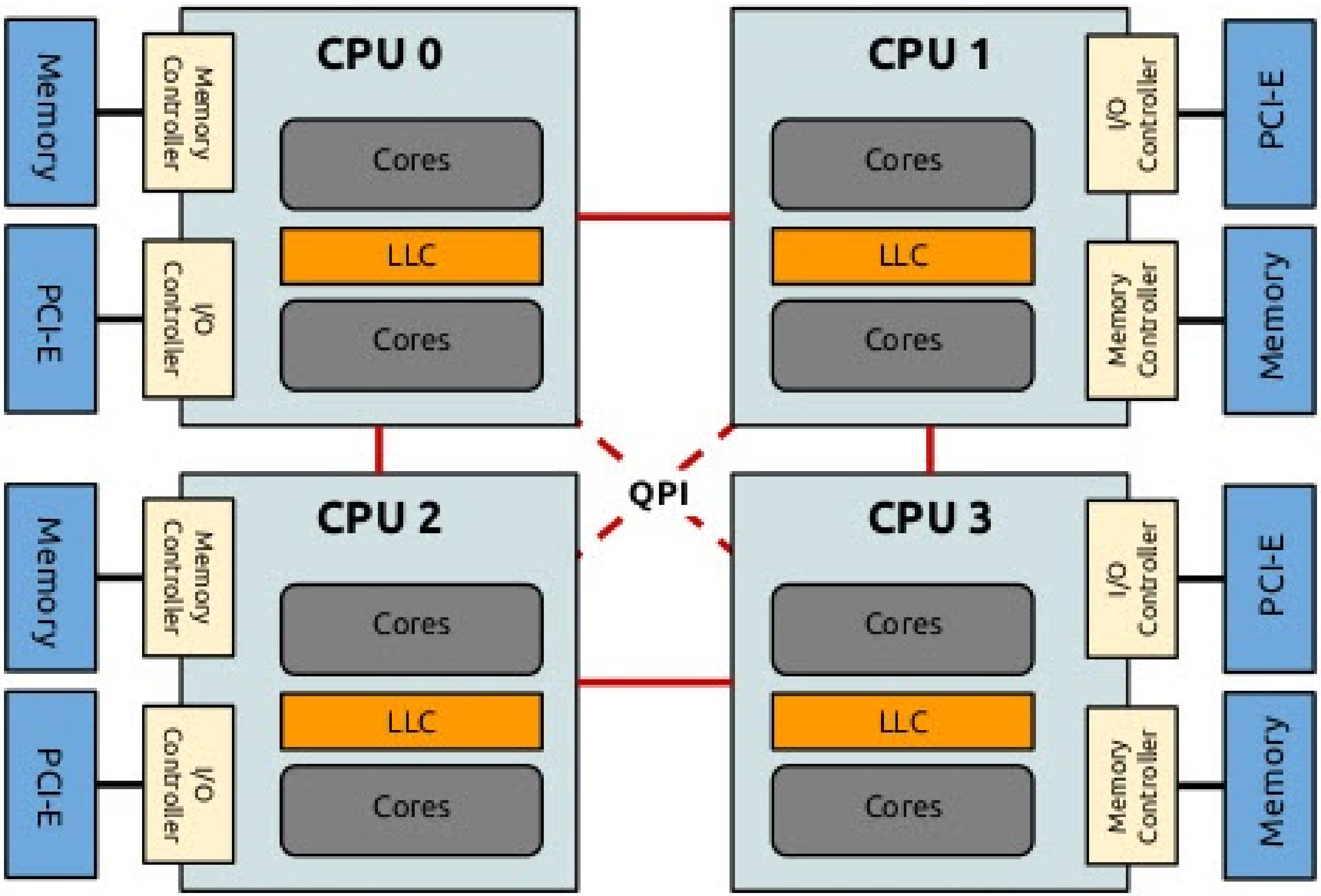


From SMP to NUMA

- FSB became rapidly a bottleneck: all the CPUs accessing memory through it
- SMP (UMA) approach no longer possible
- First NUMA architecture:
  - Hypertransport technology by AMD (2005)
- Intel came much later
  - Quick Path Interconnect (2009)
  - Fast Path Interconnect (2016)

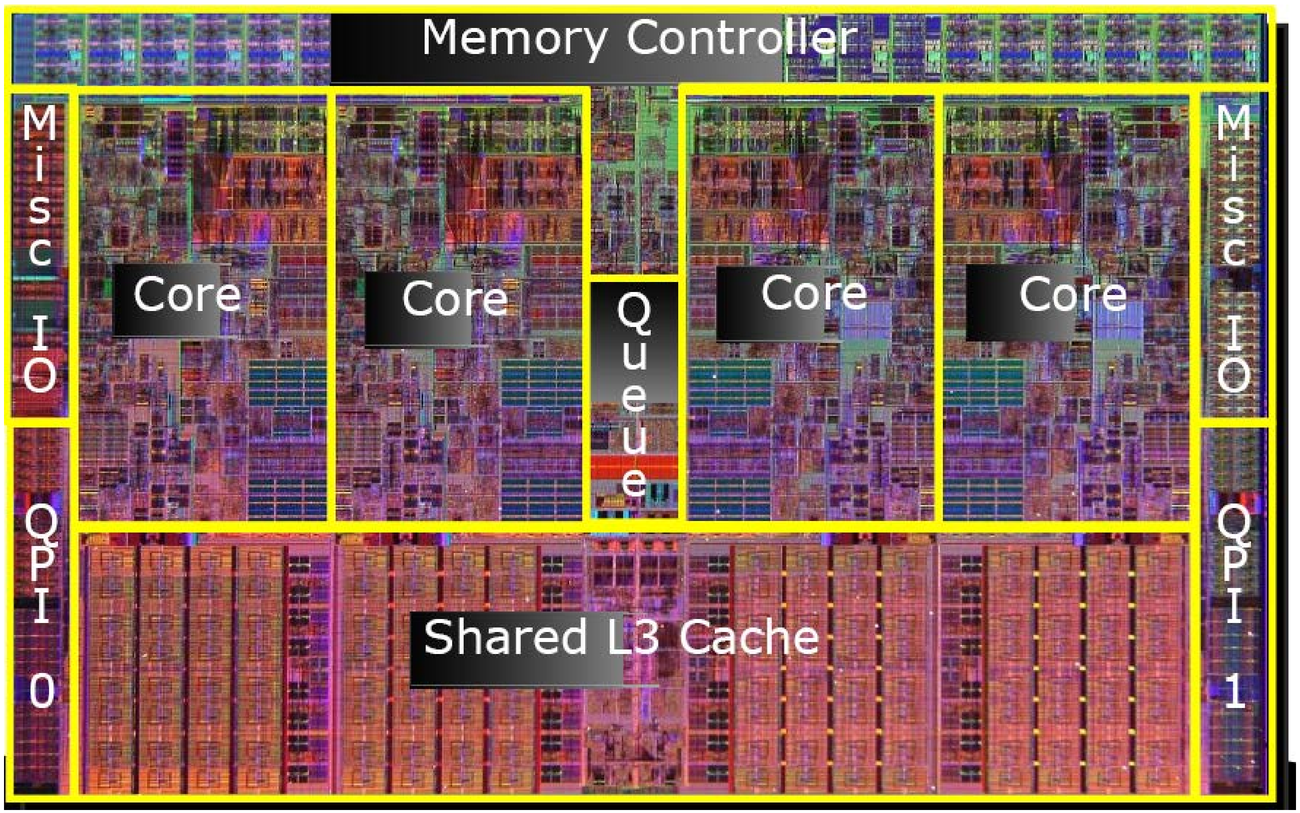
How is it logically organized ?

CPU architecture (Intel Sandy Bridge)



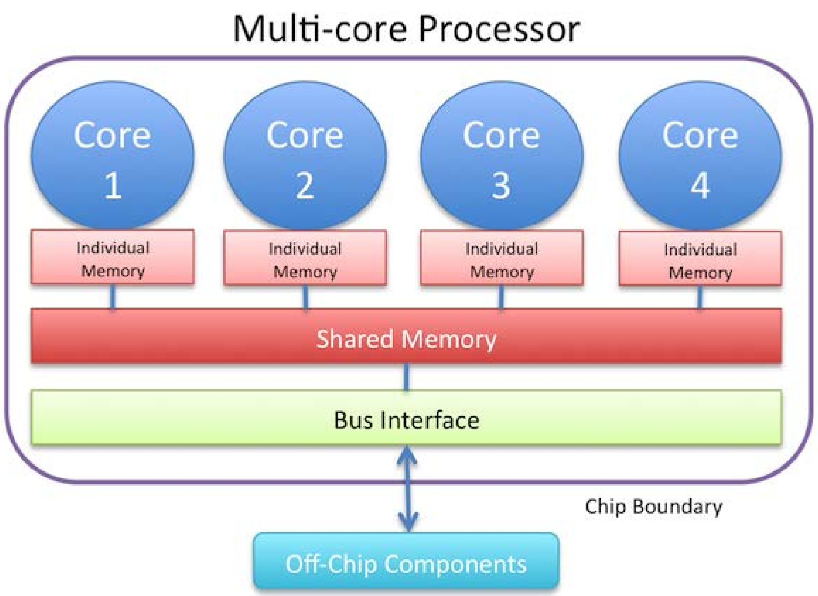
CPU level: Intel core I7

- CPU is multicore !



CPU are multicore processor

- Because of power, heat dissipation, etc increasing tendency to actually lower clock frequency but pack more computing cores onto a chip.
- These cores will share some resources, e.g. memory, network, disk, etc but are still capable of independent calculations

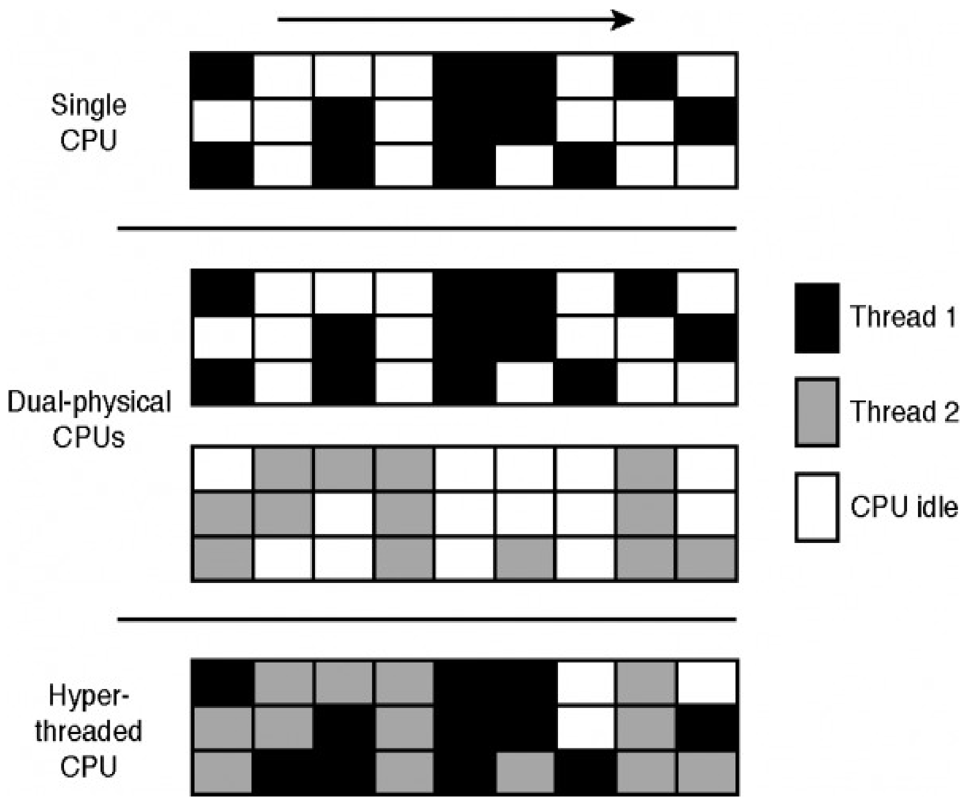


Core : definition

- A core is the smallest unit of computing, having one or more (hardware/software) threads and is responsible for executing instructions.

Hyper threading (HT)

- Intel® Hyper-Threading Technology uses processor resources more efficiently, enabling multiple threads to run on each core.
- O.S. “sees” two cores and transparently try to execute two program on two different “cores”
- Generally bad for HPC ?



Challenges for multicore

- Relies on effective exploitation of multiple-thread parallelism
  - Need for parallel computing model and parallel programming model
- Aggravates memory wall problem
  - Memory bandwidth
    - Way to get data out of memory banks
    - Way to get data into multi-core processor array
    - Memory latency
  - Cache sharing

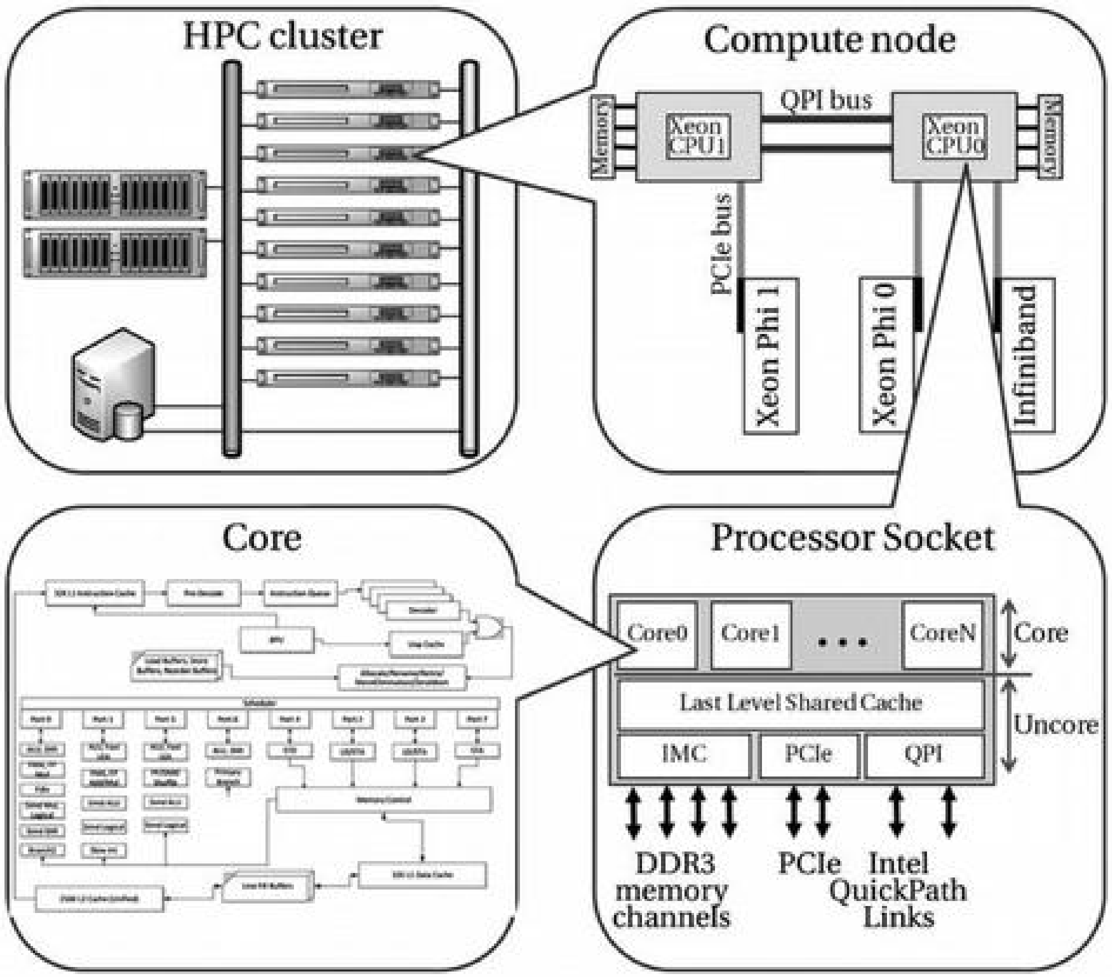
a little bit of jargon..

- Multiprocessor = server with more than 1 CPU
- Multicore = a CPU with more than 1 core
- Processor = CPU = socket

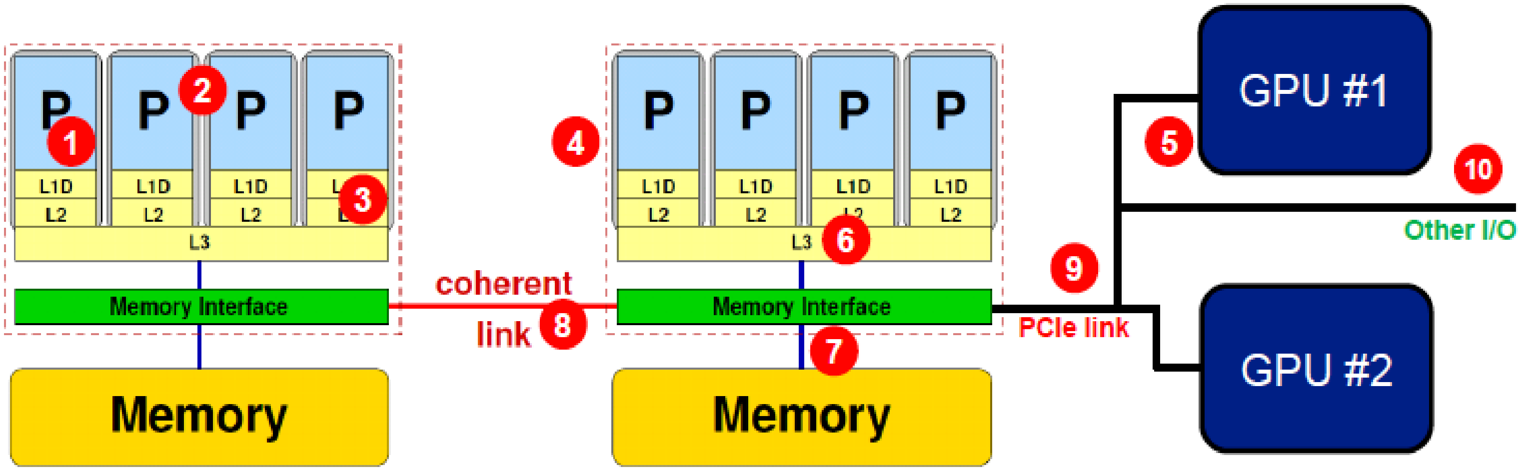
BUT SOMETIME:

- Processor = core
- a process for each processor (i.e. each core)

The building blocks of a HPC infrastructure (cluster)



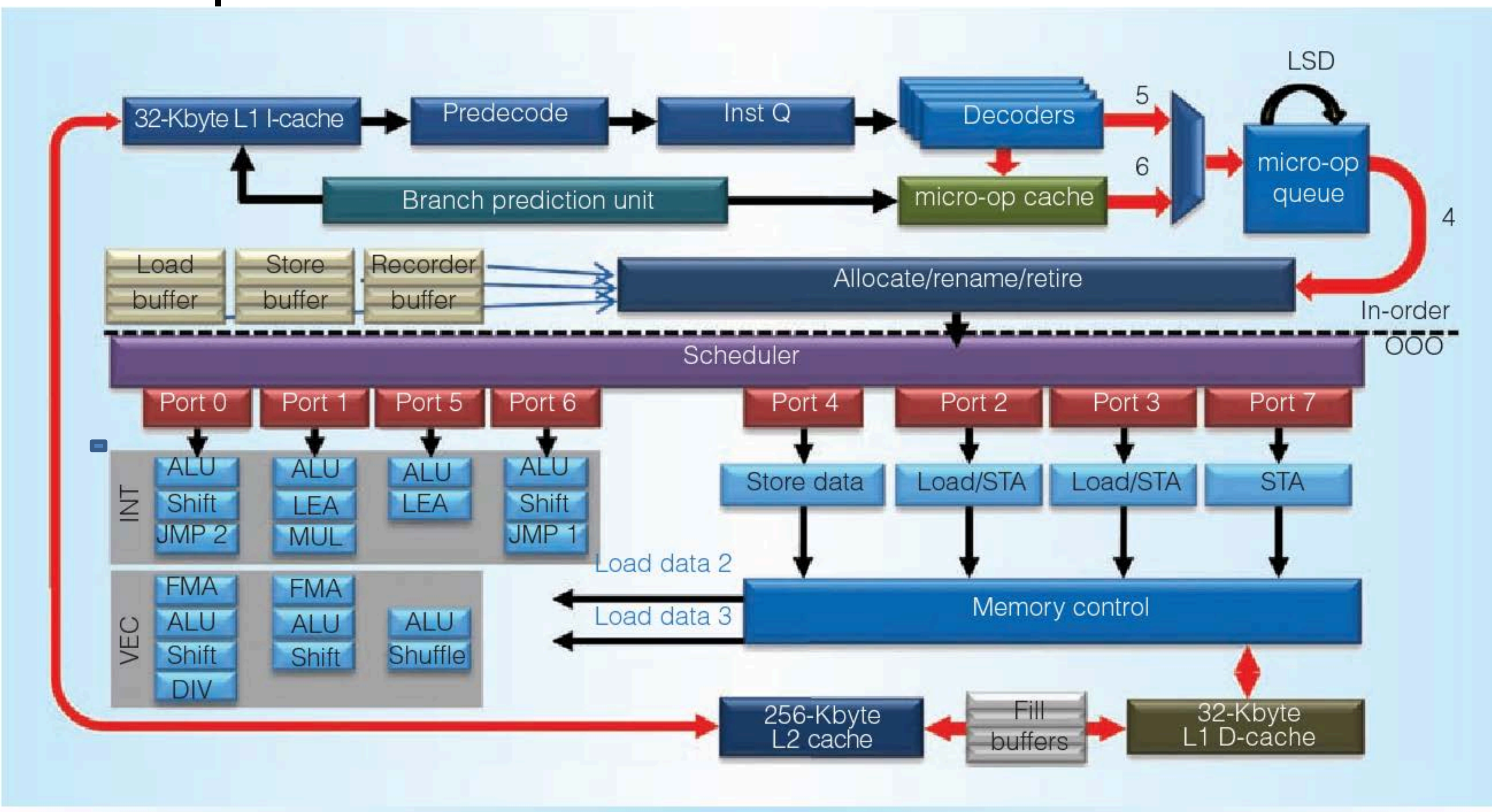
Parallellism within a HPC node



- Parallel resources
  - ILP/SIMD units (1)
  - Cores (2)
  - Inner cache levels (3)
  - Socket/ccNuma domains (4)
  - Multiple accelerator (5)

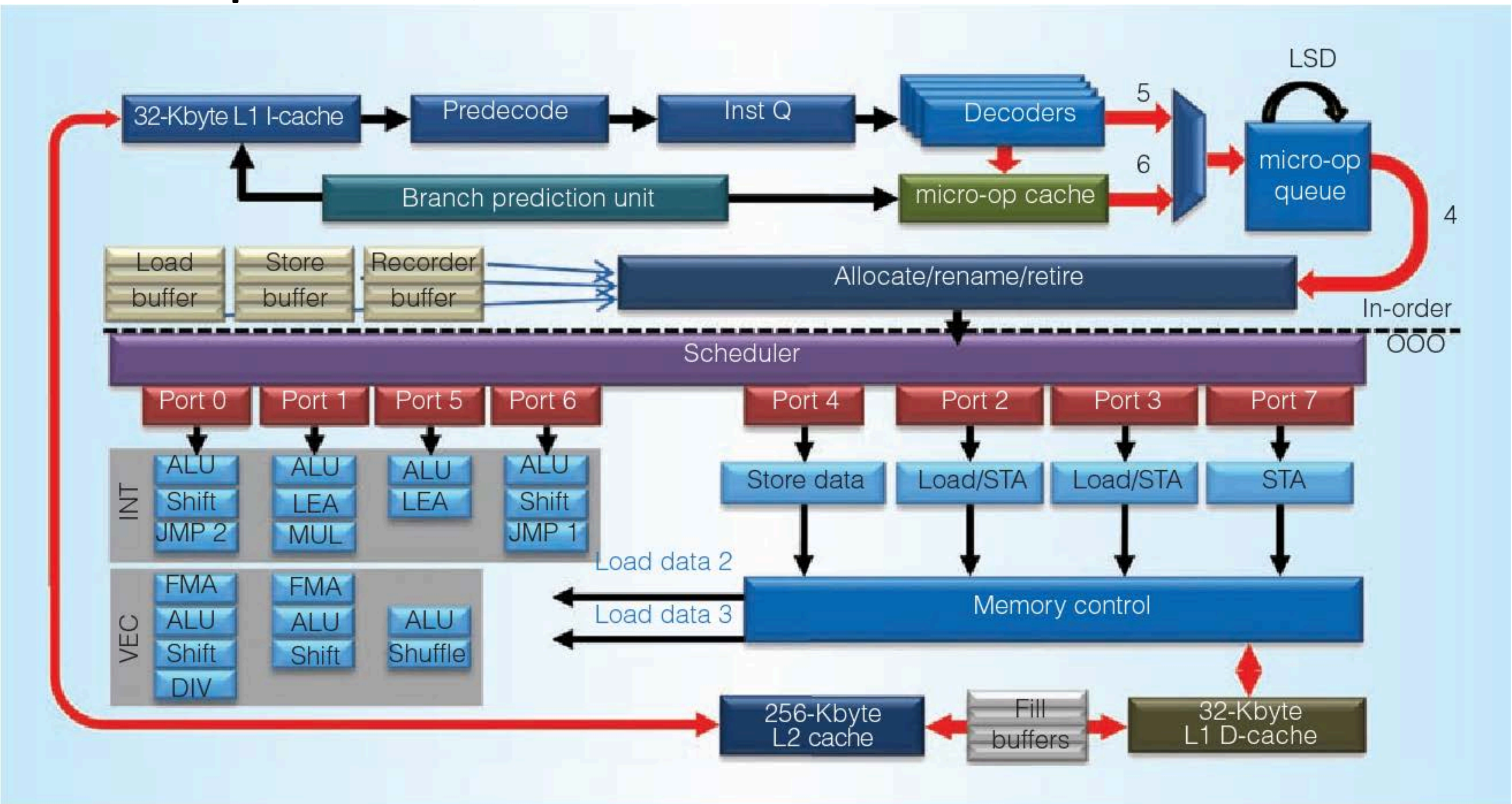
Core Level (1)

- Core can schedule instruction to more than one port at the same time

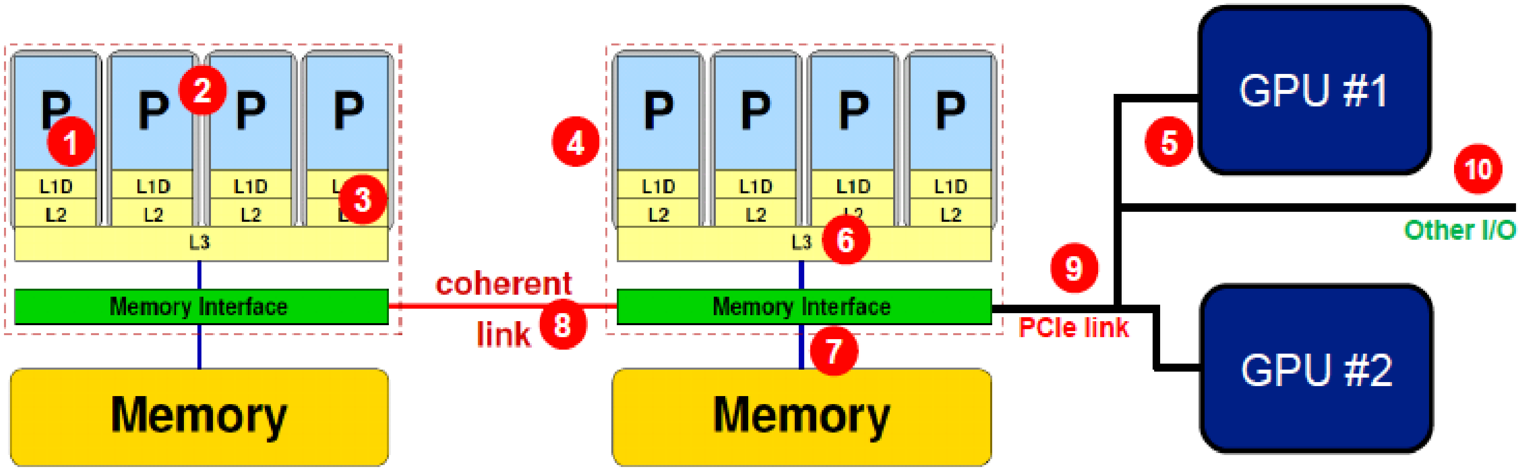


Core Level (2)

- Some port are/have SIMD devices..



Parallellism within a HPC node

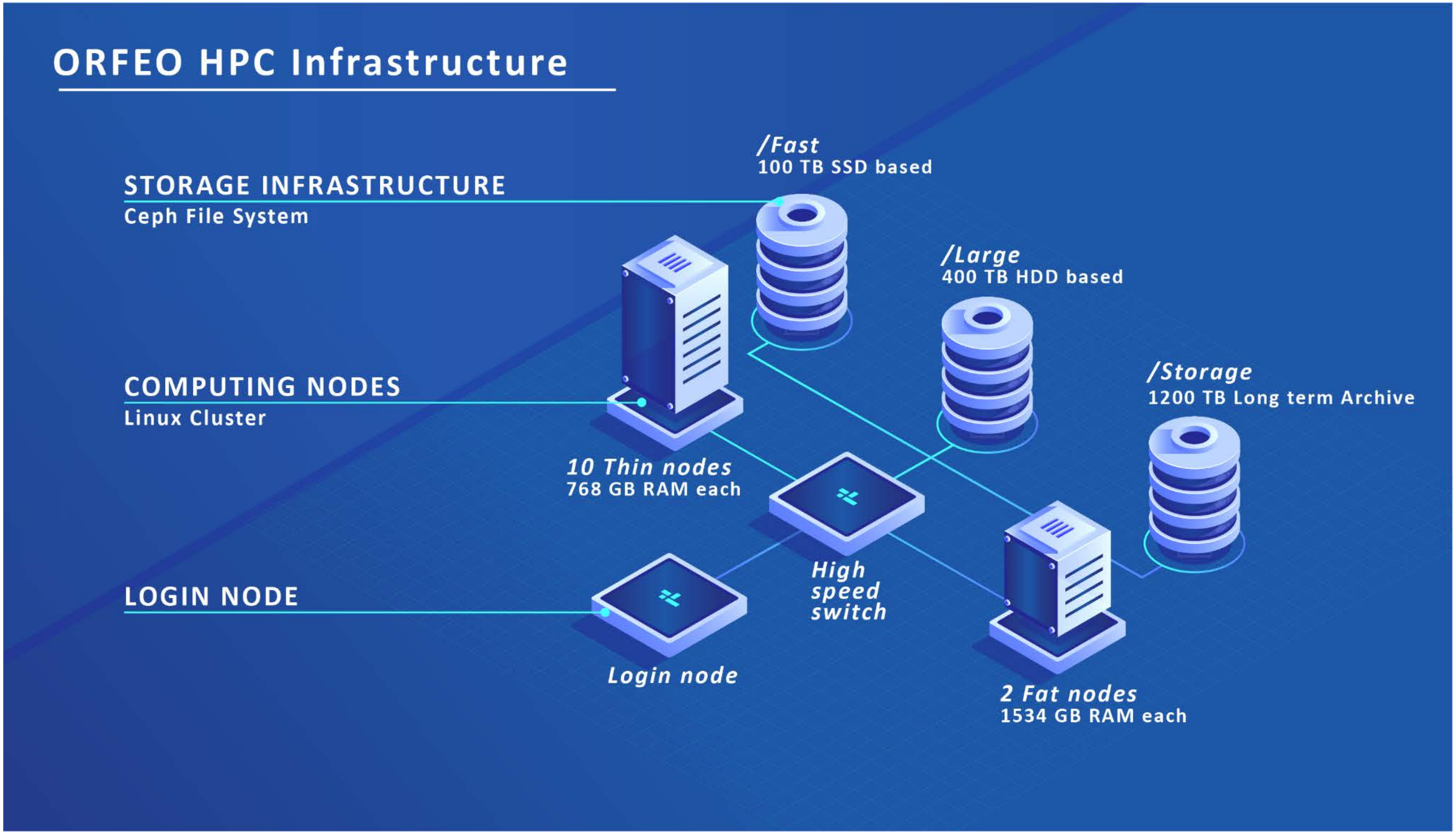


- Shared resources
  - Outer cache level per socket (6)
  - Memory bus per socket(7)
  - Intersocket link (8)
  - PCI-bus(es) (9)
  - Other I/O resources (10)

Agenda

- Why HPC is parallel ? ✓
- Serial Computers ✓
- Parallel computers ✓
- HPC infrastructure ✓
- ORFEO HPC infrastructure

A sophisticated Linux Cluster: ORFEO



ORFEO
Data Center
Open
Research
Facility for
Epigenomics and
Others

ORFEO HPC nodes..

TYPE OF NODE	RAM x nodo	CORES x nodo	GPU x nodo	Peak performance (Tflops)
10 THIN intel nodes	768 GB	24	-	1,997
2 FAT intel nodes	1536 GB	36	-	3,456
4 GPU intel nodes	256 GB	24	2 V100 (32GB)	2,073 +2x 7
8 EPYC Amd nodes (EPYC 7H12 64-Core Processor)	512 GB	128	-	?
2 DGX Nvidia Station	2048GB	128 (EPYC)	8 A100	?
TOTALE 16	**~ 15 Terabyte**	**1688**	**24**	**~ ?**

Network cluster classification

- HIGH SPEED NETWORK
  - parallel computation
  - low latency /high bandwidth
  - Usual choices: Infiniband...
- I/O NETWORK
  - I/O requests (NFS and/or parallel FS)
  - latency not fundamental/ good bandwidth
  - GIGABIT could be ok /10Gb and/or Infiniband better
- In band Management network
  - management traffic of all services (LRMS/NFS/software etc..)
- Out of band Management network:
  - Remote control of nodes and any other device

Orfeo network

- HIGH SPEED NETWORK
- I/O NETWORK

100 Gbit HDR Infiniband

- In band Management network

25Gbit Ethernet

- Out of band Management network:

1Gbit Ethernet

ORFEO storage: hardware

	FAST storage (NVMe)	FAST storage (SSD)	Standard storage (HDD)	Long term preservation
# of server	4		6	1
RAM	6 x 16GB		6 x 16GB	6 x 16GB
Disk per node	2x 1.6TB NVMe PCIe card	20 x 3.84TB	15 x 12TB	84 x 12TB + 42 x 12TB
Storage provider	CEPH parallel FS	CEPH parallel FS	CEPH parallel FS	Network FS (NFS)
RAW storage	**12TB**	**320 TB**	**1080 TB**	**1,512 TB**

I/O subsystem on ORFEO:

- Home
  - once logged in, each user will land in its home in `/u/[name_of_group]/[name_of_user]
  - e.g. the home of user area is in /u/area/[name_of_users]
  - it's physically located on ceph large FS, and exported via infiniband to all the computational nodes
  - quotas are enforced with a default limit of 2TB for each users
  - soft link are available there for the other areas

```
[cozzini@login ~]$ ls -lrt
total 548398
lrwxrwxrwx 1 cozzini area        18 Apr  7  2020 fast -> /fast/area/cozzini
lrwxrwxrwx 1 cozzini area        21 Apr  7  2020 storage -> /storage/area/cozzini
lrwxrwxrwx 1 cozzini area        21 Apr 16  2020 scratch -> /scratch/area/cozzini
```

I/O subsystem on ORFEO:

- /Scratch
  - it is large area intended to be used to store data that need to be elaborated
  - it is also physically located on ceph large FS, and exported via infiniband to all the computational nodes

```
[cozzini@login ~]$ df -h /scratch
Filesystem
Size  Used Avail Use% Mounted on
10.128.6.211:6789,10.128.6.213:6789,..:/  598T   95T  503T  16% /large
```

- /fast
  - is a fast space available for each user, on all the computing nodes
  - is intended to be a fast scratch area for data intensive application

```
[cozzini@login ~] df -h /fast
Filesystem
Size  Used Avail Use% Mounted on
10.128.6.211:6789,10.128.6.212:6789,..:/   88T  4.3T   83T   5% /fast
```

I/O subsystem on ORFEO:

- Long term storage:
  - it is NFS mounted via 50bit ethernet link
  - it is intended for long-term storage of final processed dataset

```
[cozzini@login ~]$ df -h  | grep 231
10.128.6.231:/illumina_run           128T   58T   70T  46% /illumina_run
10.128.2.231:/storage                 37T   27T  9.9T  74% /storage
10.128.6.231:/long_term_storage      128T  112T   17T  88% /long_term_storage
10.128.6.231:/analisi_da_consegnare  100T   33T   68T  33%
/analisi_da_consegnare
10.128.6.231:/onp_run_1              117T   27T   91T  23% /onp_run
10.128.2.231:/lage_archive           128T   68T   60T  54% /lage_archive
```

Completed !

- Why HPC is parallel ? ✓
- Serial Computers ✓
- Parallel computers ✓
- HPC infrastructure ✓
- ORFEO HPC infrastructure ✓