# High Performance Computing and Data Infrastructure

## Recap on HPC hardware

DATA SCIENCE &
ARTIFICIAL INTELLIGENCE

SCIENTIFIC &
DATA-INTENSIVE COMPUTING

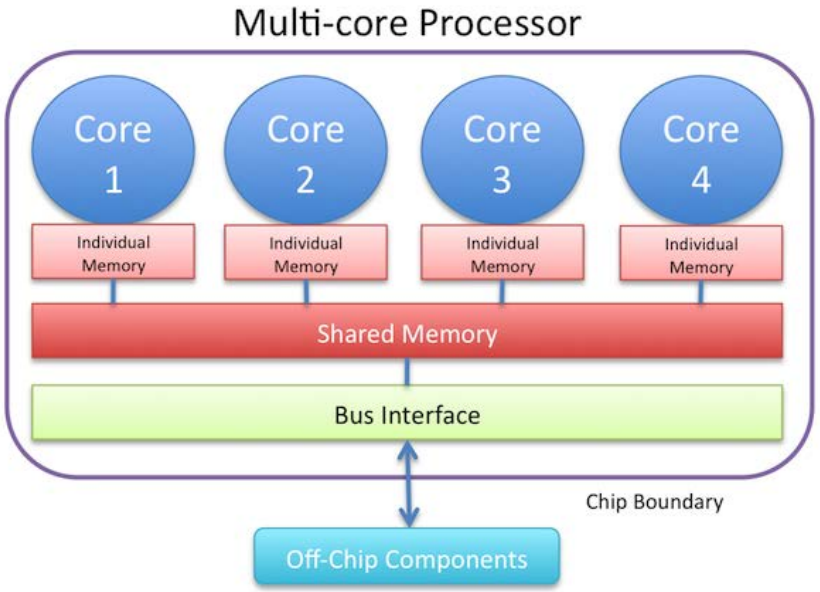
**2024-2025 @ Università di Trieste**

# Agenda

- HPC is parallel
- Serial Computers
- Moore law/Dennard Scaling
- Parallel computers
- HPC infrastructure
- ORFEO HPC infrastructure

# CPU are multicore processor

- Because of power, heat dissipation,  creasing 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

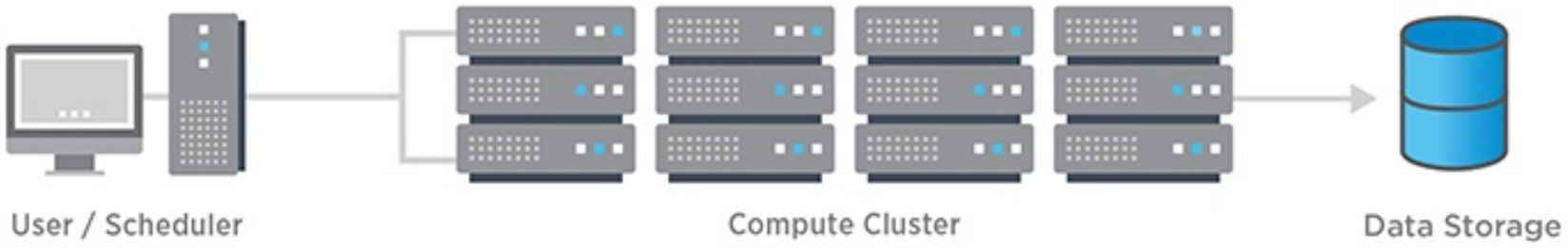


# Comments

| | HW level | SW level |
|---|---|---|
| **SISD** | A Von Neumann CPU | no parallelism at all |
| **MISD** | On a superscalar CPU, different ports executing different *read* on the same data | • ILP os same data;<br>• Multiple tasks or threads operating on the same data |
| **SIMD** | Any vector-capable hardware, the vector registers on a core, a GPU, a vector processor, an FPGA, ... | data parallelism through vector instructions and operations |
| **MIMD** | Every multi-core/processor system; on a superscalar CPUs, different ports executing different ops on different data | • ILP on different data;<br>• Multiple tasks or threads executing different code on different data. |

# Essential component of a HPC cluster

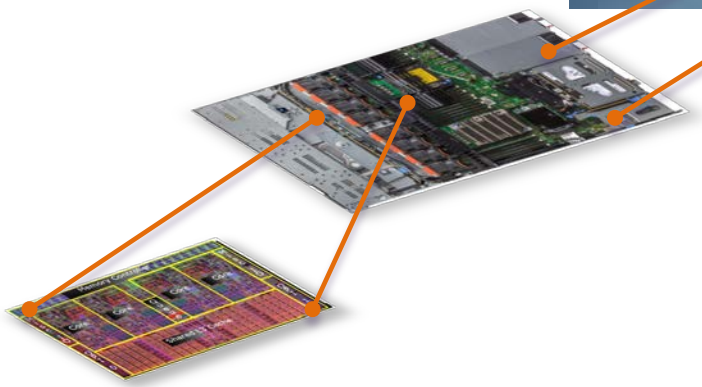
- Several computers (nodes)
  - often in special cases (1U) for easy mounting in a rack
- One or more networks (interconnects) to hook the nodes together
- Some kind of storage
- A login/access node..



# The hardware behind HPC

interconnected
racks of connected
nodes

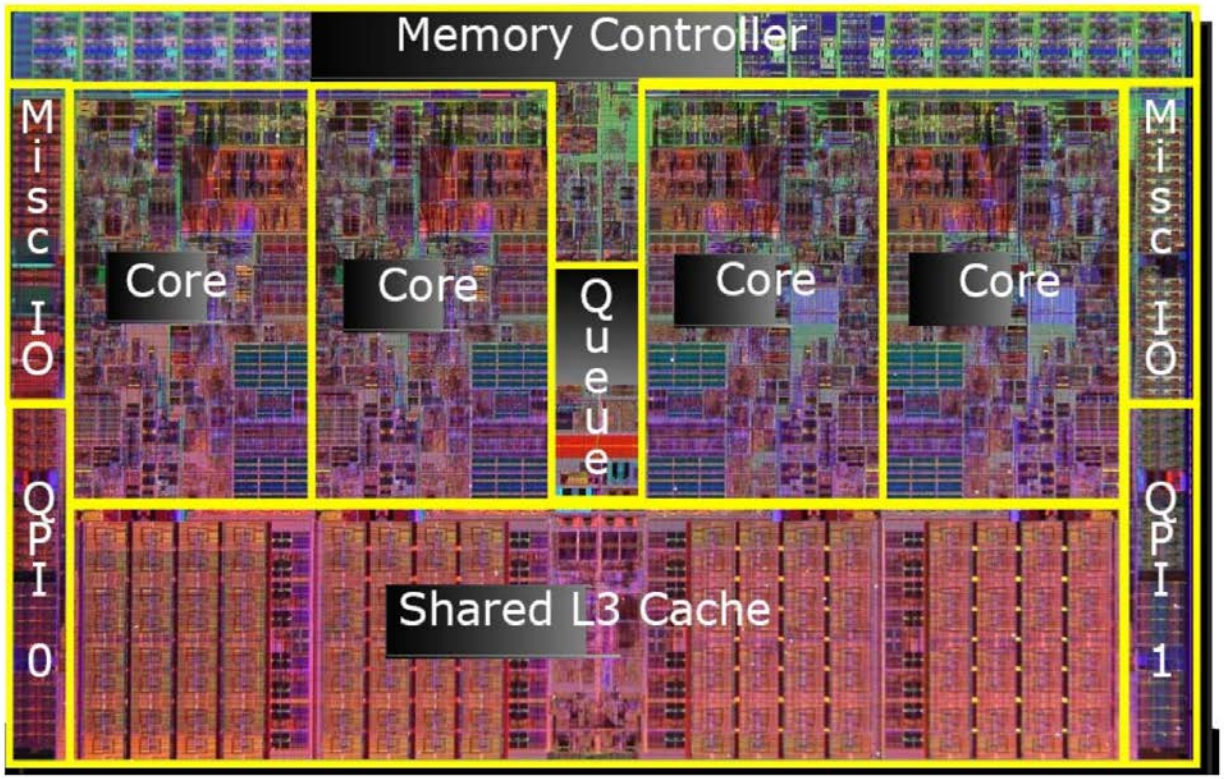
single nodes



single cpu

# Single CPU topology

**Modern CPUS**
are multi- (or many-) cores

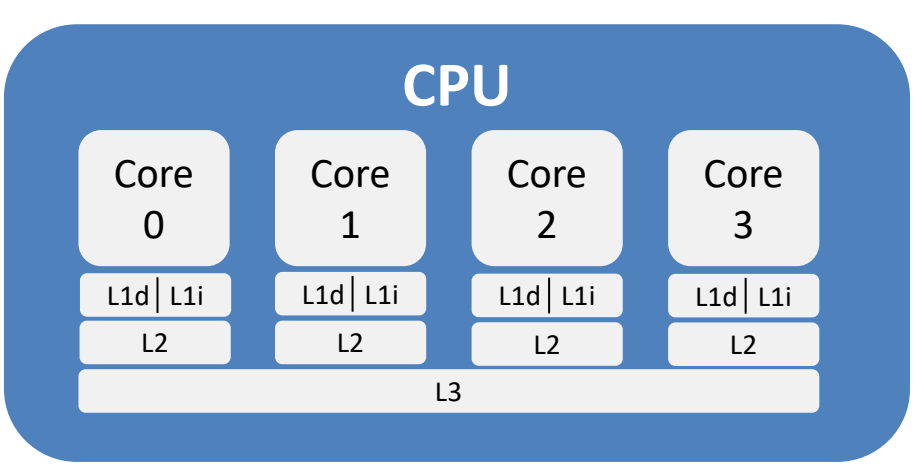


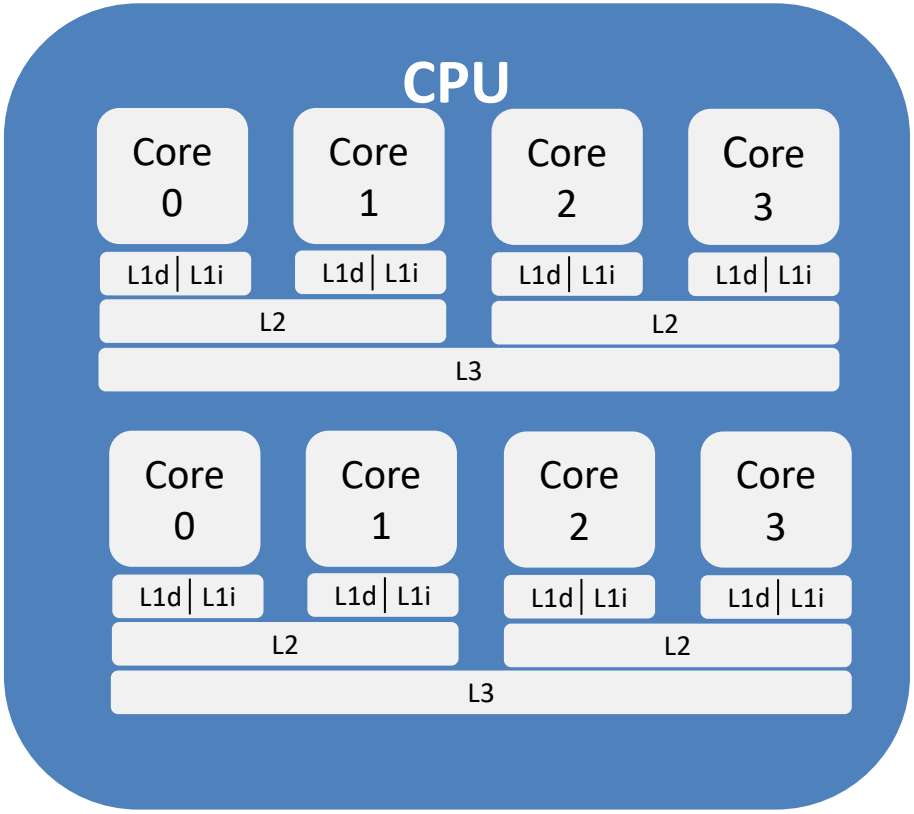
# Core : definition

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

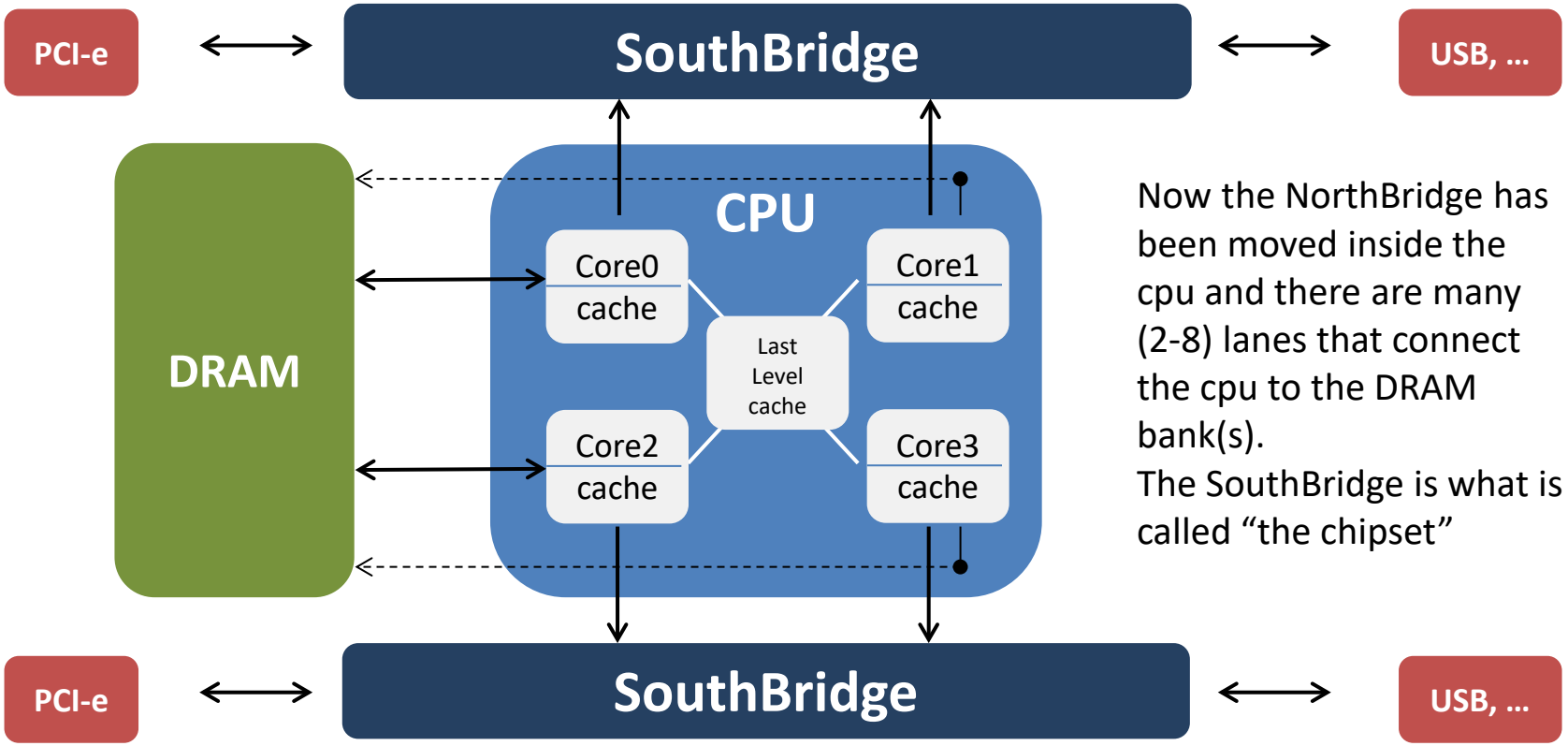
# Single CPU topology

- Cache hierarch can have different topologies



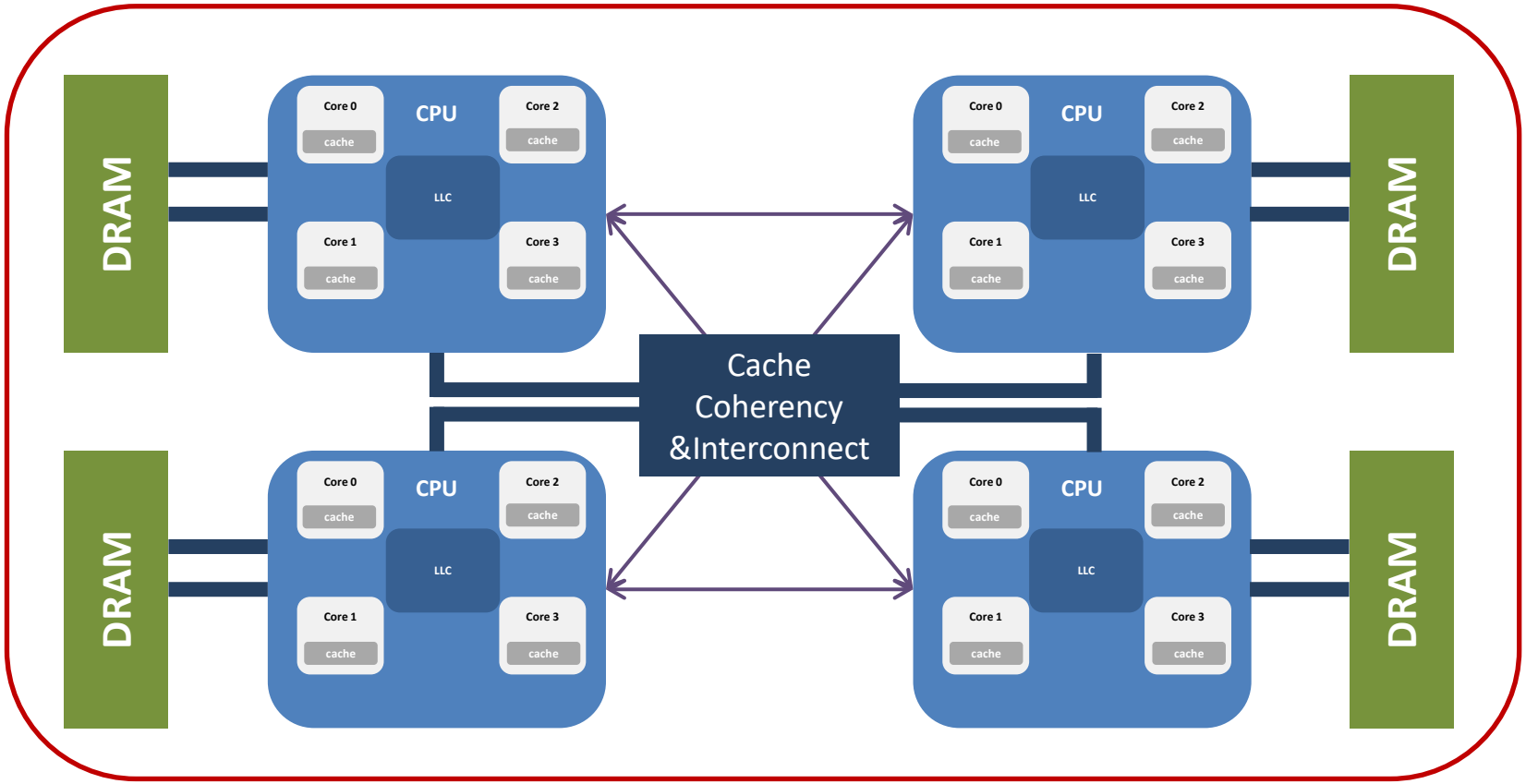



# Modern CPU layout

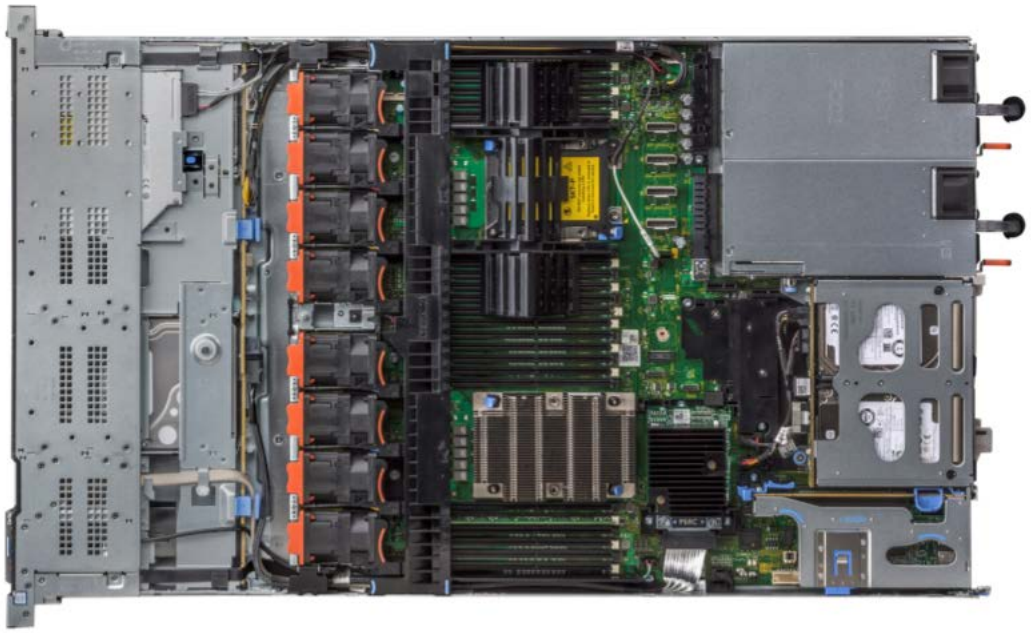



Now the NorthBridge has been moved inside the cpu and there are many (2-8) lanes that connect the cpu to the DRAM bank(s).
The SouthBridge is what is called “the chipset”

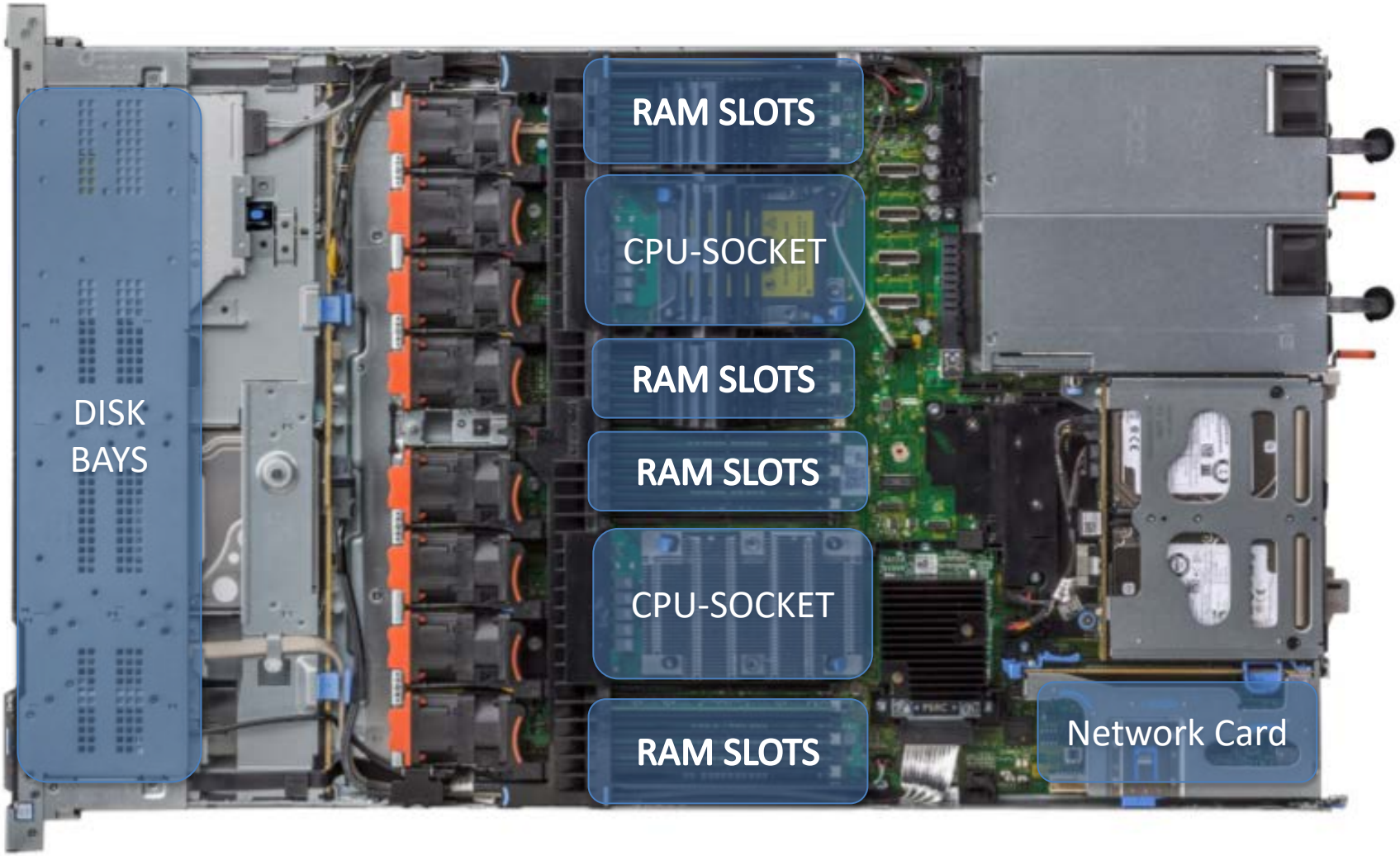
# Node topology



# Modern 1U computing nodes



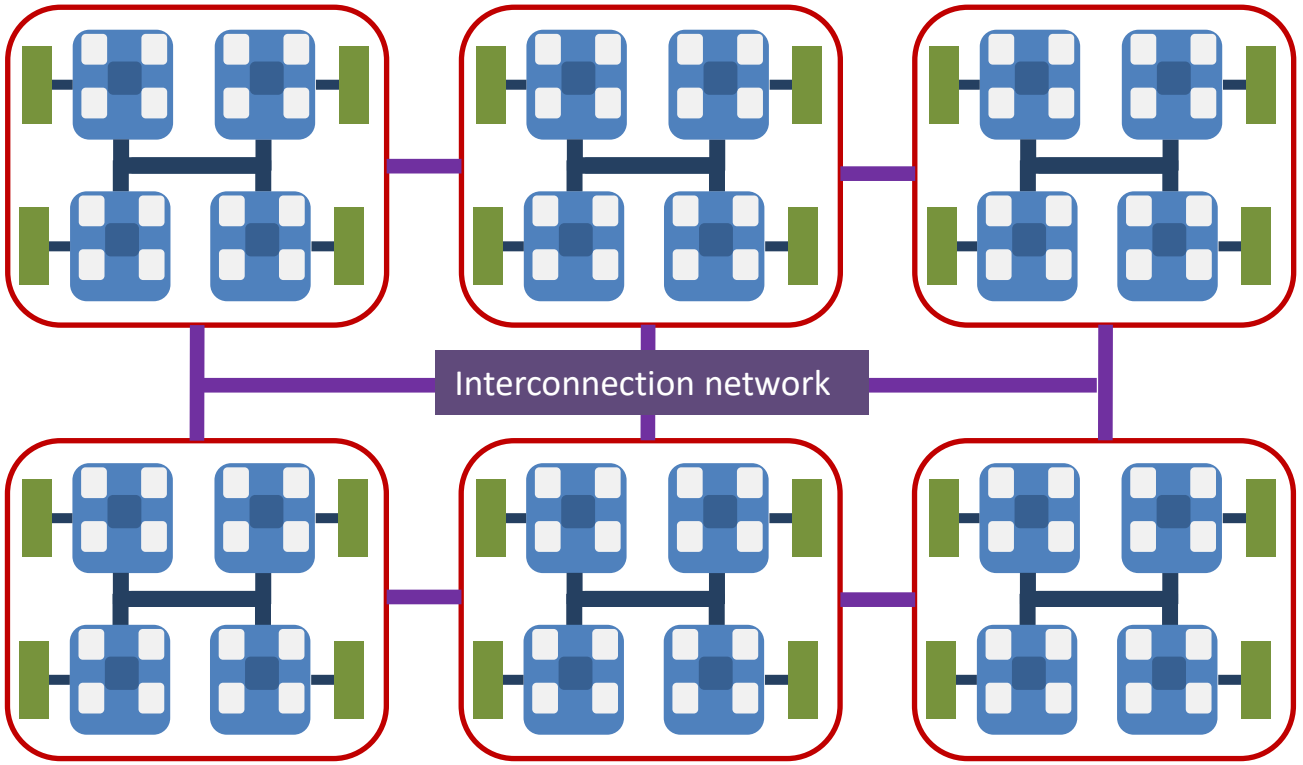
# What does one node contain exactly ?



# The overall topology

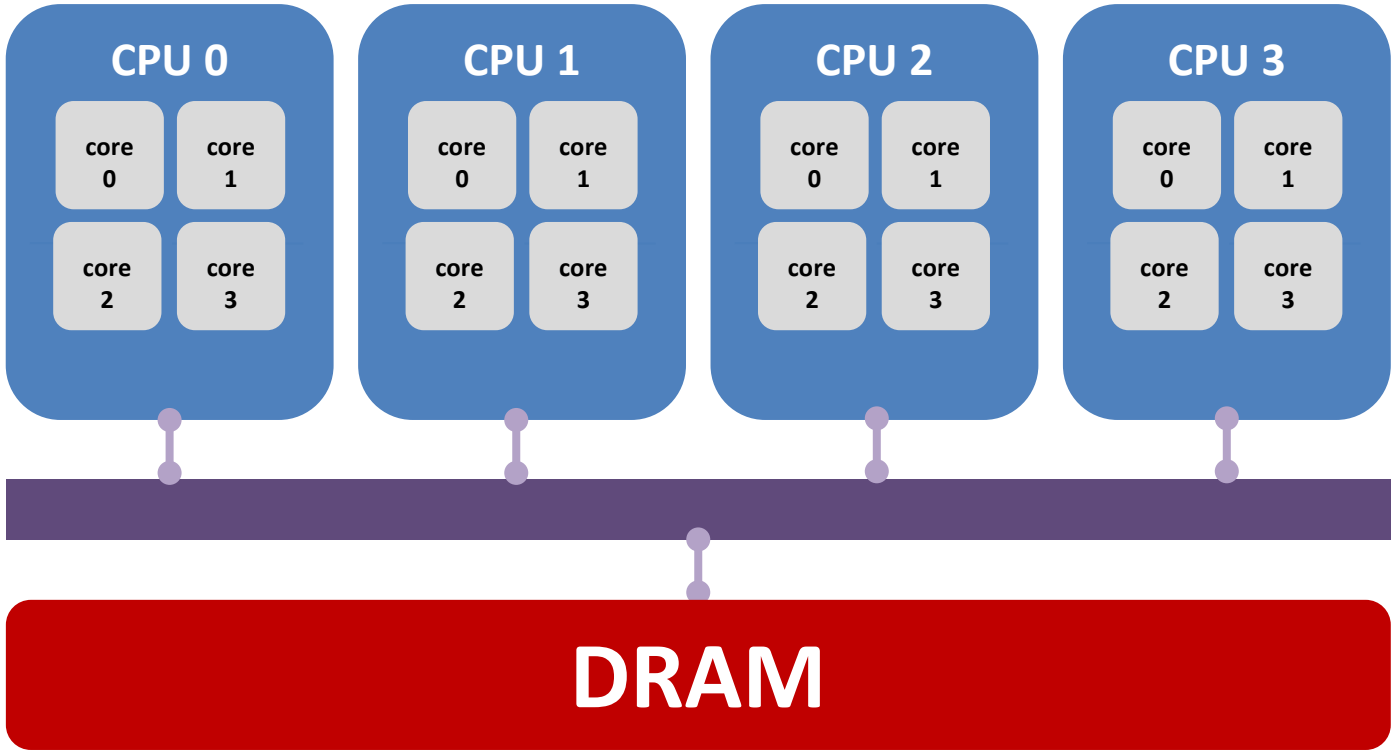
CLUSTER OF COMPUTING NODES

Note: there are many different topologies for the interconnection network.



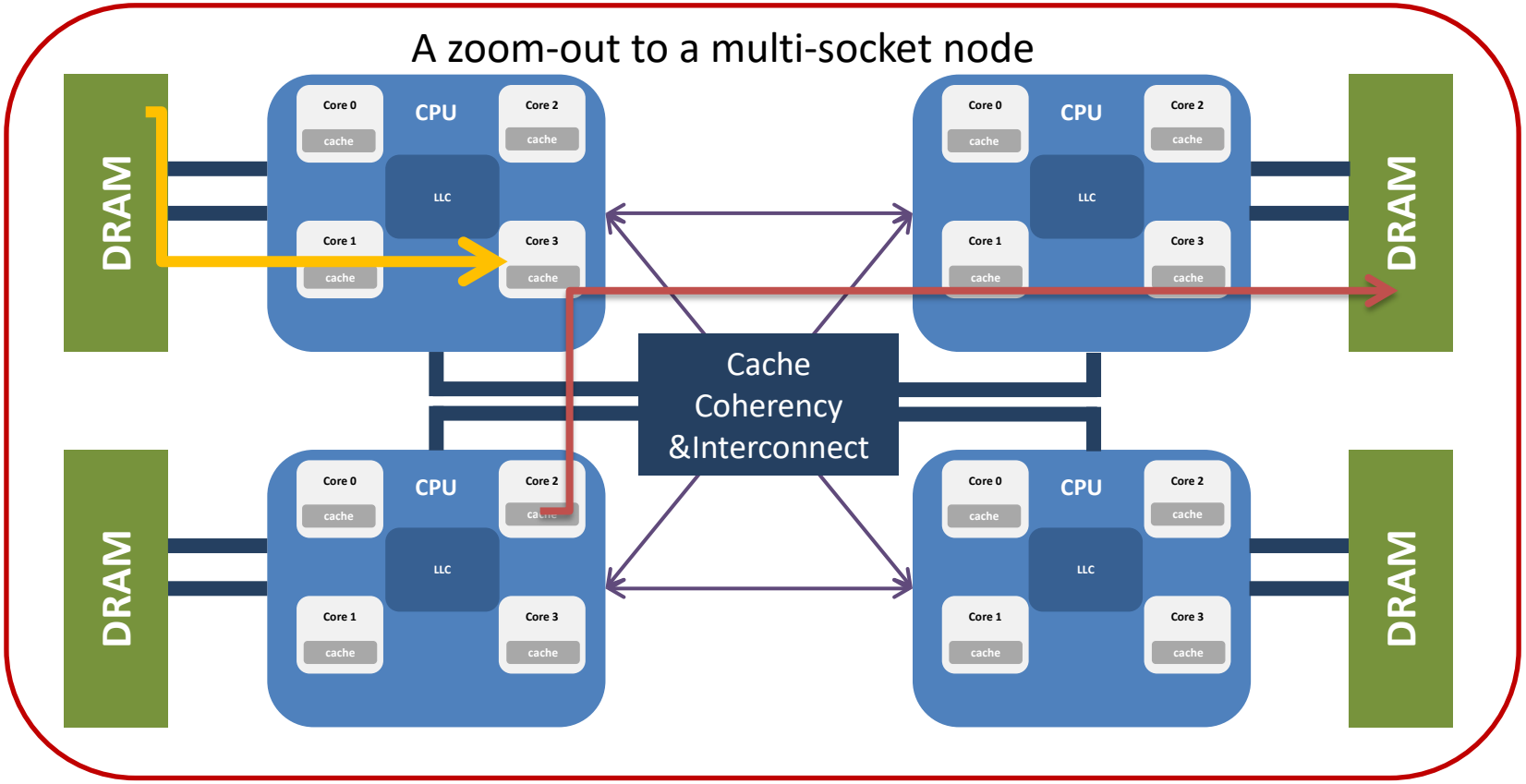
# Shared memory: UMA

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

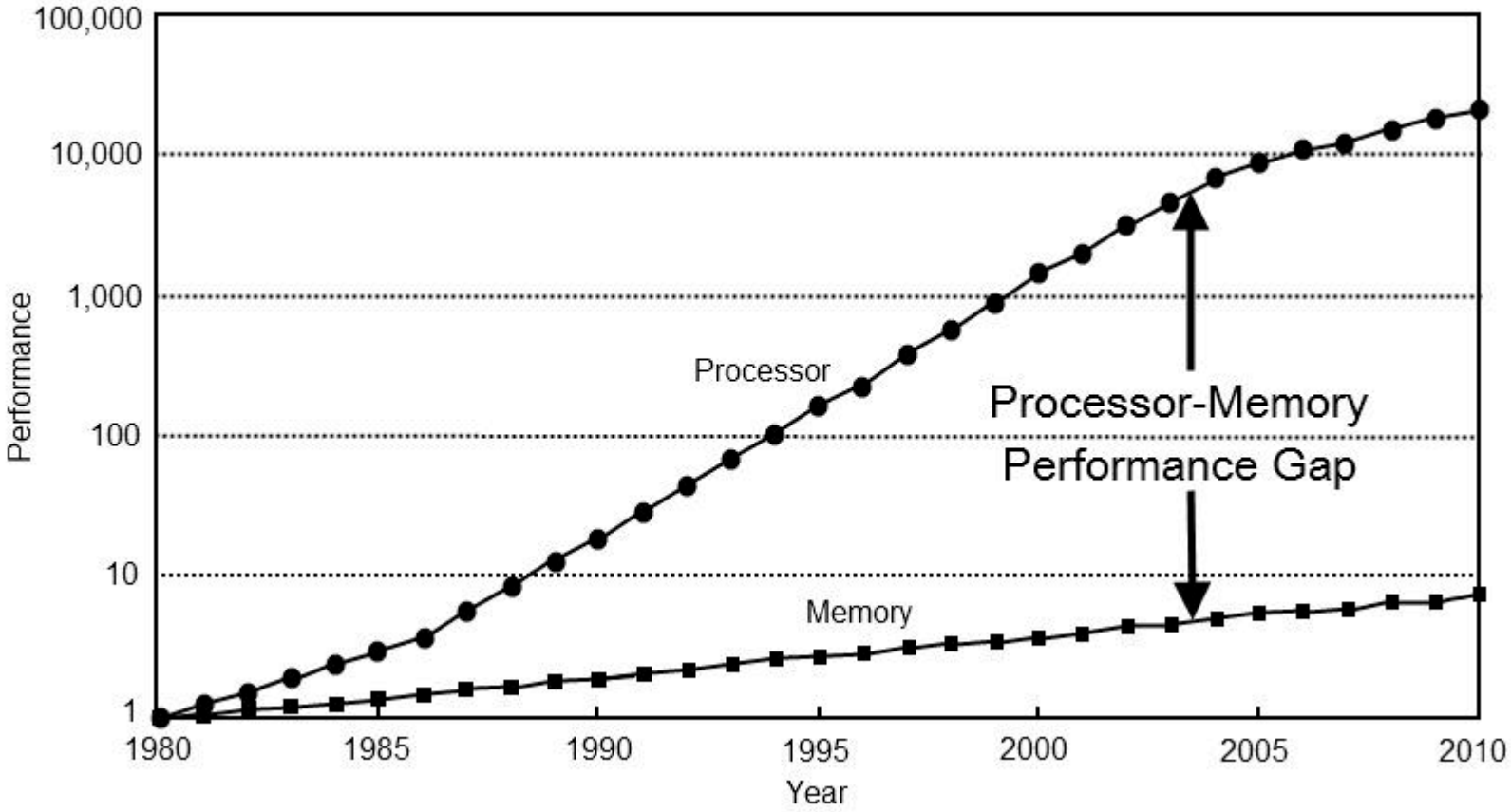


# Shared memory: NUMA

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



# Memory wall problem



# 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

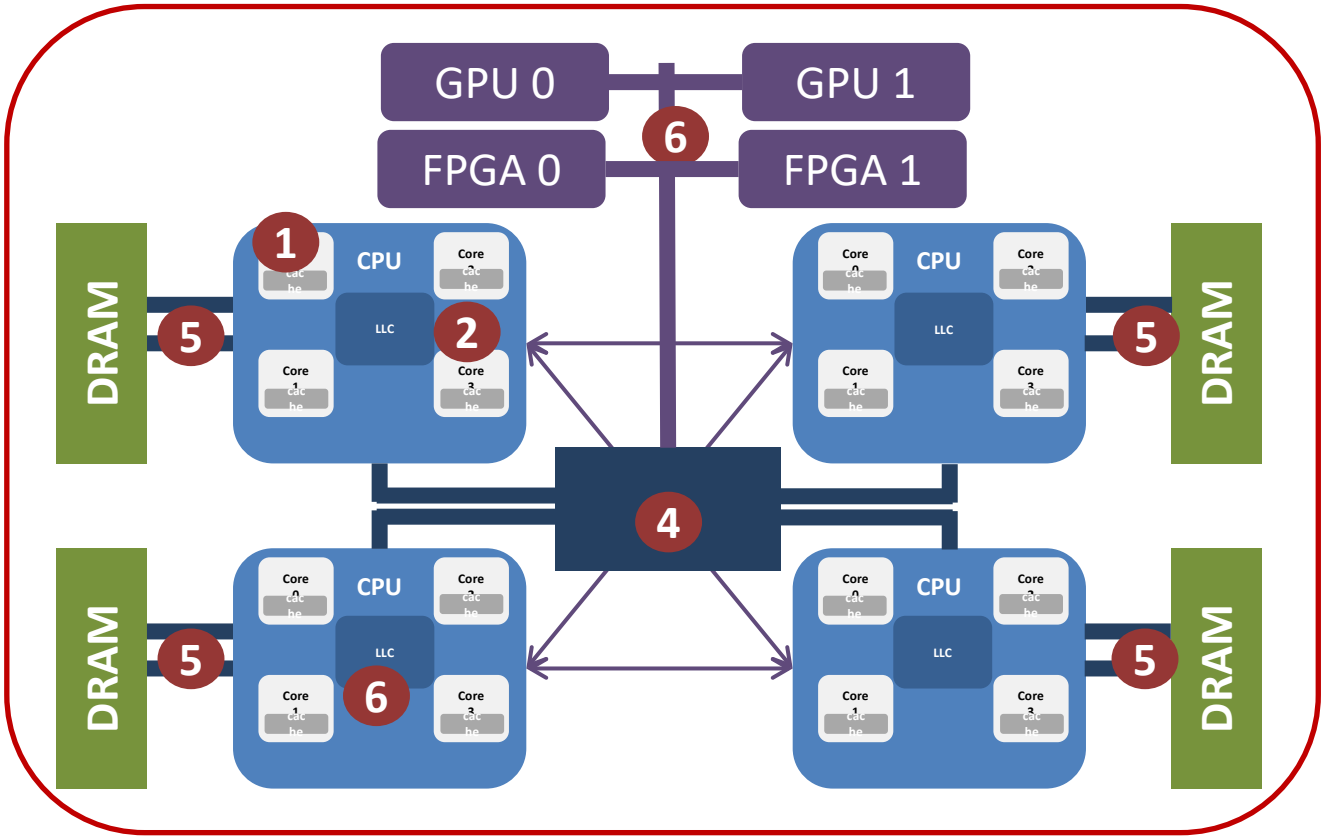
# a little bit of jargon..

- 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)

# Parallellism within a HPC node




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

# Discovering a Numa topology

- **`numactl`** tool
  it also controls the Linux NUMA policy

- **`/proc/cpuinfo`**

- **`hwloc`** (by OpenMPI)


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

# Some link to check together

- Introduction to the NVIDIA DGX A100 System — NVIDIA DGX A100 User Guide 1 documentation
- NVIDIA Hopper Architecture In-Depth | NVIDIA Technical Blog