Corso di formazione per neoassunti nelle attività di Computing

4–7 Mar 2024
LNF

Dal laptop al supercalcolo

Alessandro Costantini, Daniele Cesini, Luigi Scarponi
INFN-CNAF
alessandro.costantini <at> cnaf.infn.it

INFN and CNAF

- INFN-CNAF: the national center of INFN (Italian Institute for Nuclear Physics) dedicated to Research and Development on Information and Communication Technologies
  - https://www.cnaf.infn.it/en
- CNAF hosts the Italian Tier-1 data center for the high-energy physics experiments at the Large Hadron Collider in Geneva
- CNAF is one of the most important centers for distributed computing in Italy
  - 50000 CPU cores
  - 60 PB disk space
  - 200 PB tape library

The CNAF history in one slide

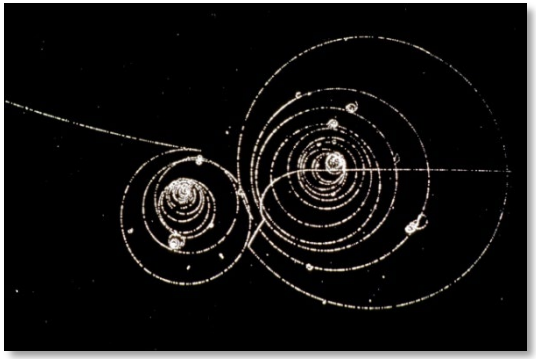
1960 | 1970 | 1980 | 1990 | 2000 | 2010 | 2020

The CNAF (National Center for Frame Analysis) was established in 1962 as a Central Facility of INFN for the analysis of the frames coming from the bubble chamber.

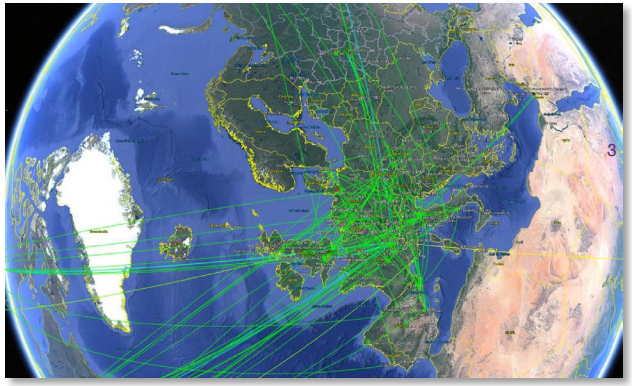
Subsequently, the CNAF developed and managed the INFN geographical network which gradually evolved into the Italian research network now managed by GARR (1980-2000)

- At the end of the 90s CNAF realizes the LHC Tier1 Data Center.
- CNAF is one of the main actors in the development of GRID World Wide Computing.

- CNAF has a solid DC Tier1 of LHC.
  Moreover it offers CPU and storage resources to more than 30 INFN physics experiments.
- It developed CLOUD services oriented to the scientific world, to the productive world, to society



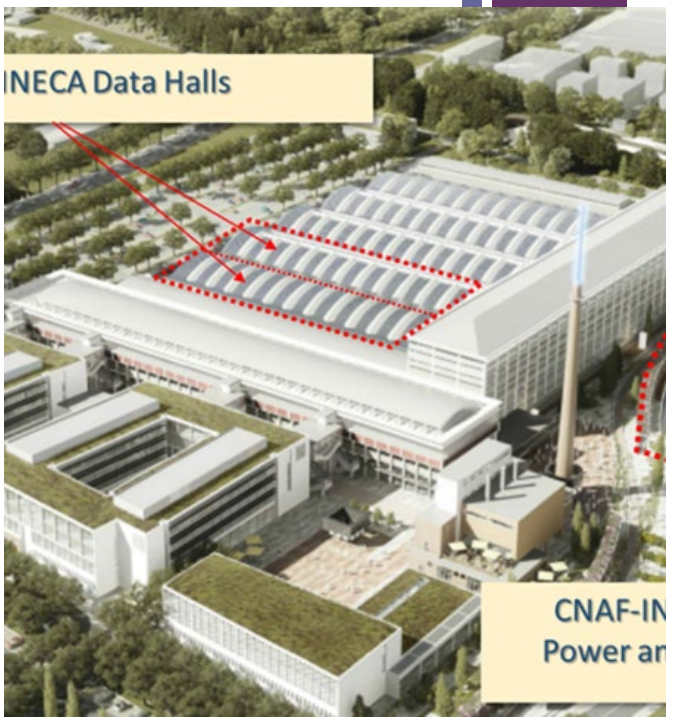
First installation of the CNAF Tier1 (2003)



Towards the new datacenter

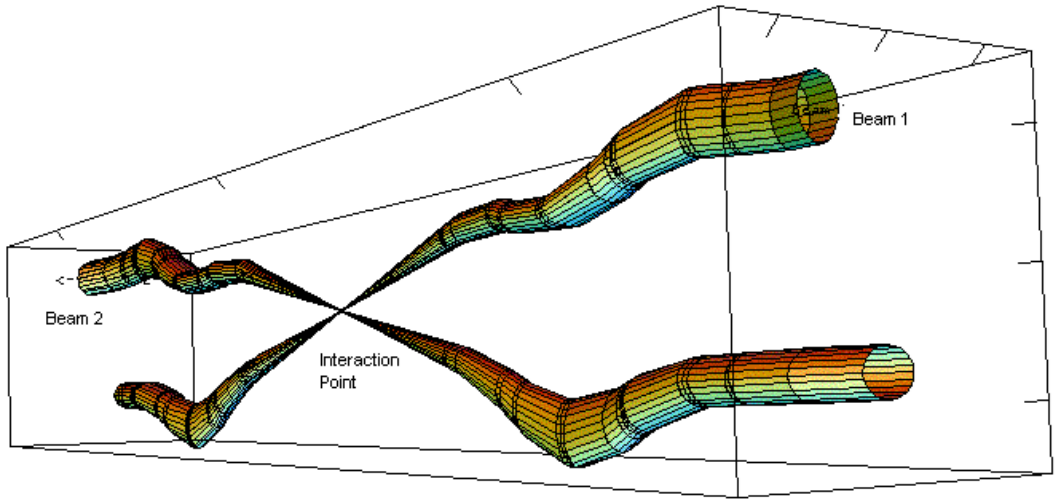
- In 2019 Emilia Romagna Region decided to promote a new district, the Tecnopolo, devoted to research, innovation and technological development
  - 3 halls assigned to ECMWF for the new data center
- A pre-exascale machine (Leonardo) funded by EuroHPC and Italian Ministry of Research and University assigned to a Consortium led by CINECA and INFN
- >~15000 m^2 (for IT and technical plants) allocated for two data center suitable to host both Leonardo and INFN Tier1

For INFN and CINECA
10MW on phase1
20MW on phase2



Main Users: High Energy Physics

Large Hadron Collider (LHC)
60 PB /year

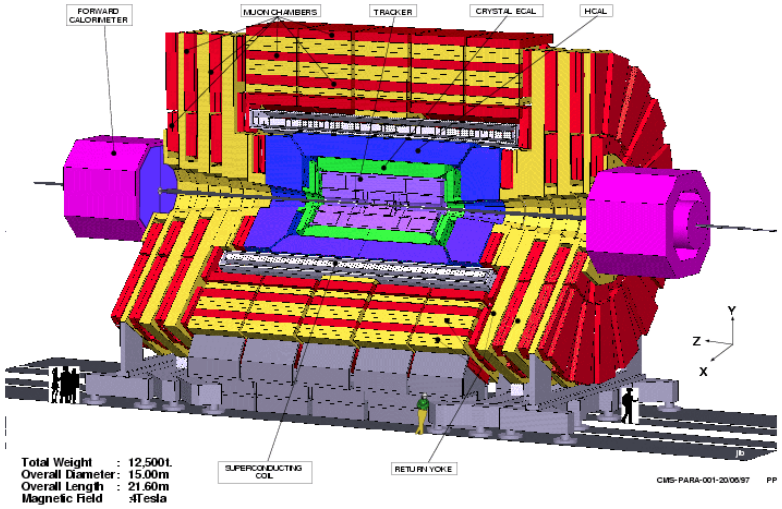


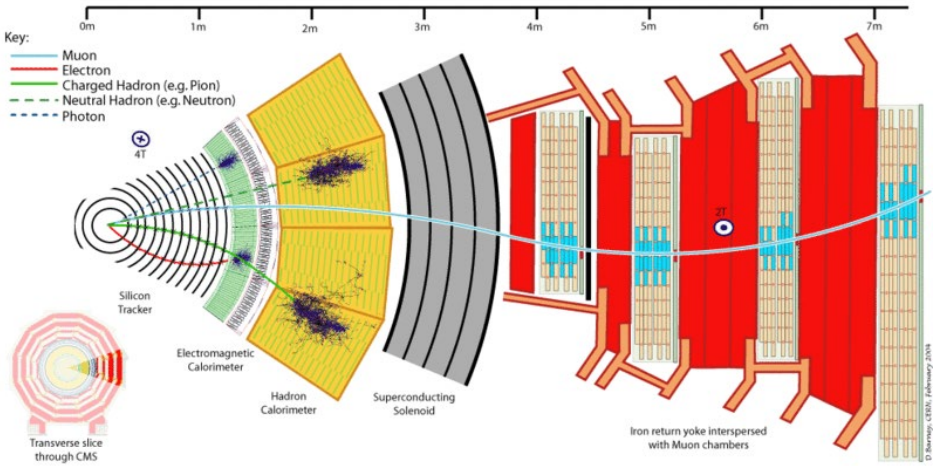
Relative beam sizes around IP1 (Atlas) in collision

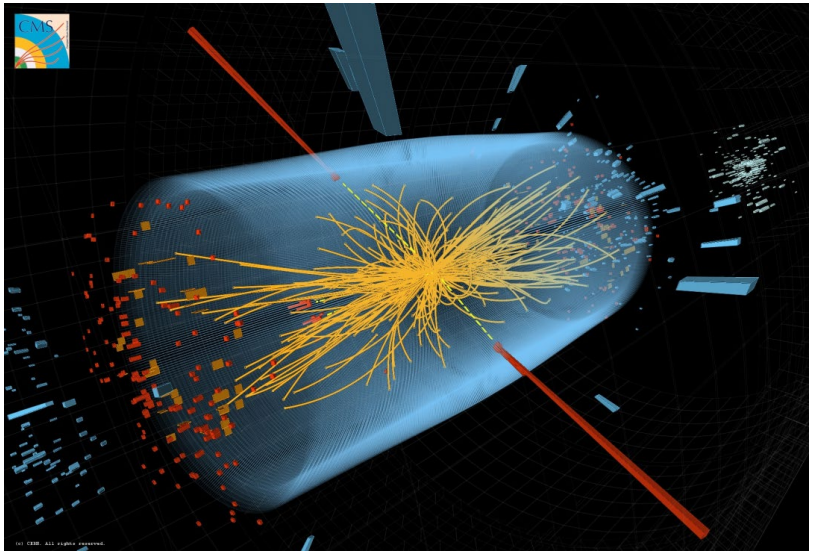
40MHz peak crossing rate (25ns)

600 million inelastic events per second.

Track reconstruction







Embarrassingly parallel applications

- Independent jobs -> no need for inter-process communication

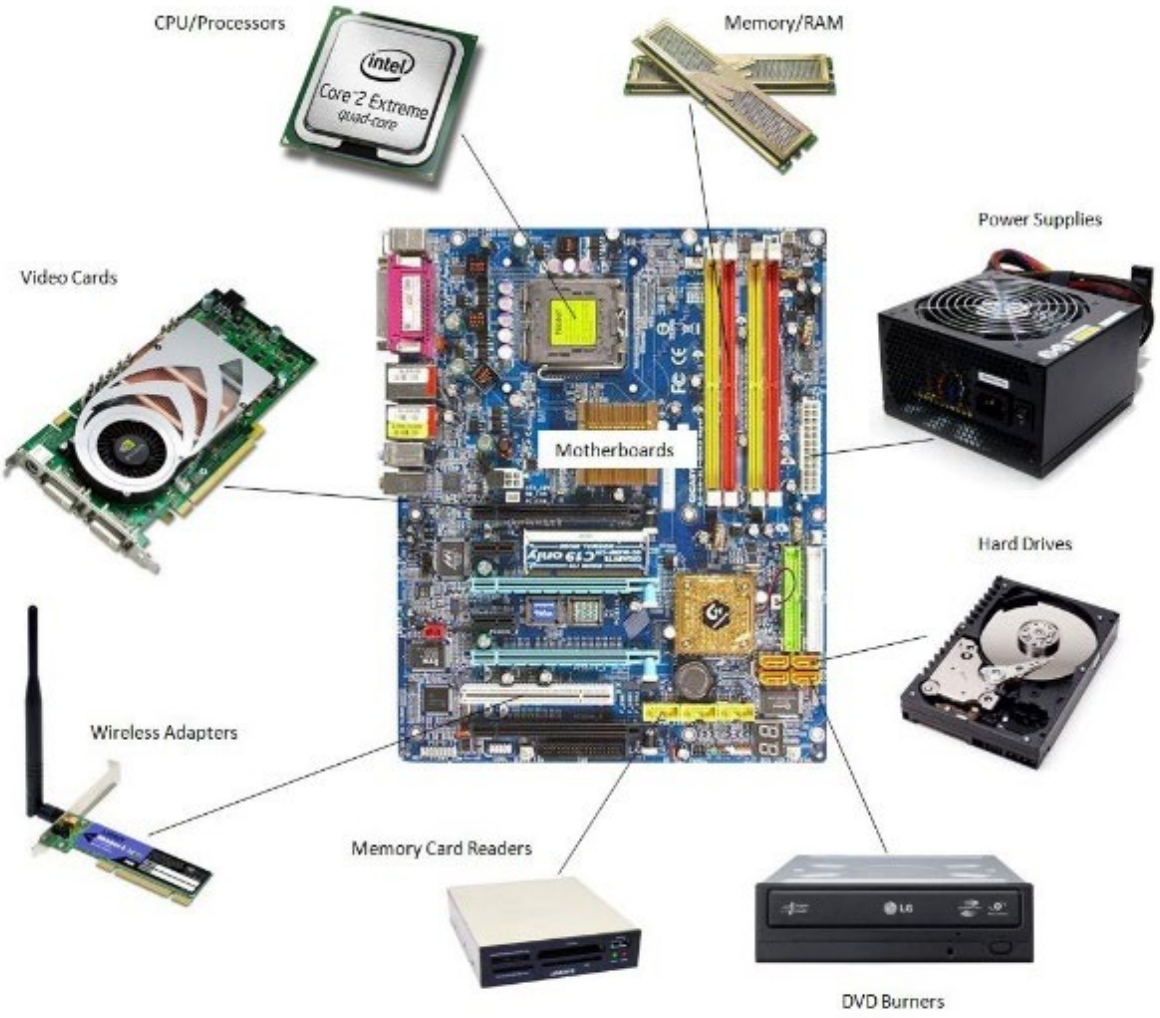
Our infrastructure, the datacenter, is shaped to address this use case

High Throughput Computing

Take away message: not all computing infrastructures fit all use cases

From the PC to the datacenter

PC components



bit and Byte

1 byte (B) = 8bit (b)

Prefixes for multiples of bits (bit) or bytes (B)						
Decimal			Binary			
Value		SI	Value		IEC	JEDEC
1000	k	kilo	1024	Ki	kibi	K kilo
1000^2	M	mega	1024^2	Mi	mebi	M mega
1000^3	G	giga	1024^3	Gi	gibi	G giga
1000^4	T	tera	1024^4	Ti	tebi	–
1000^5	P	peta	1024^5	Pi	pebi	–
1000^6	E	exa	1024^6	Ei	exbi	–
1000^7	Z	zetta	1024^7	Zi	zebi	–
1000^8	Y	yotta	1024^8	Yi	yobi	–

https://en.wikipedia.org/wiki/Template_talk:Bit_and_byte_prefixes

Binary-Byte

byte (B) is binary

- **KiB** = "kibibyte" = «kilo-binary B» = 2^10 byte = 1.024 byte
- **MiB** = "mebibyte" = «mega-binary B» = 2^10 KiB = 2^20 B = 1.048.576 byte
- **GiB** = "gibibyte" = «giga-binary B» = 2^10 MiB = 2^20 KiB = 2^30 B = 1.073.741.824 byte
- **TiB** = "tebibyte" = «tera-binary B» = 2^10 GiB = 2^20 MiB = 2^30 KiB = 2^40 B = 1.099.511.627.776 byte

The closest decimal values are

- **kB** = "kilobyte" = 10^3 byte = 1.000 byte (la "k" è minuscola per non confonderla con i gradi kelvin, "K")
- **MB** = "megabyte" = 10^3 KB = 10^6 B = 1.000.000 byte
- **GB** = "gigabyte" = 10^3 MB = 10^6 kB = 10^9 B = 1.000.000.000 byte
- **TB** = "terabyte" = 10^3 GB = 10^6 MB = 10^9 kB = 10^12 B = 1.000.000.000.000 byte

The error considering gigabyte instead gibibyte is of about 7,4%.

https://simple.wikipedia.org/wiki/Gibibyte

Processor

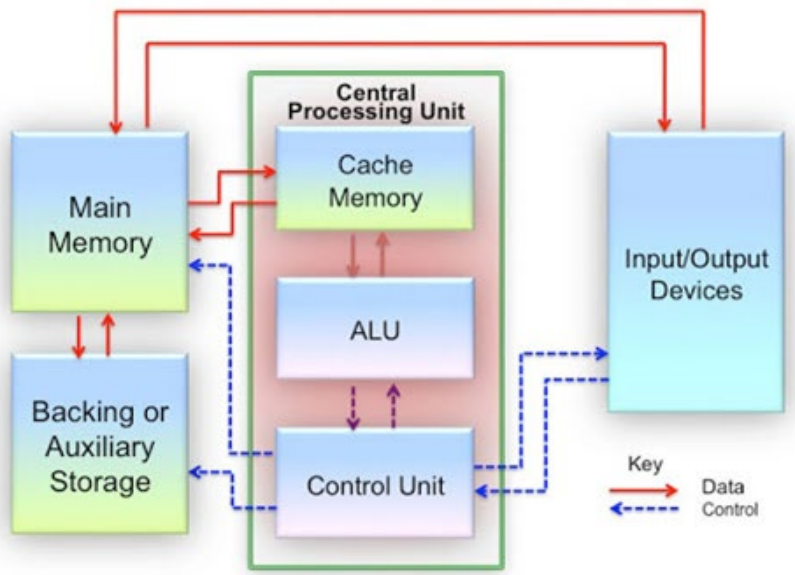
Central Processing Unit

- A central processing unit (CPU), also called a central processor or main processor, is the electronic circuitry within a computer that executes instructions that make up a computer program.

- Performs basic arithmetic, logic, controlling, and input/output (I/O) operations specified by the instructions in the program

- Principal components of a CPU include:
  - **the Arithmetic Logic Unit (ALU)**
    - performs arithmetic and logic operations
  - **processor registers**
    - supply operands to the ALU and store the results of ALU operations
  - **Control unit** that orchestrates the fetching (from memory) and execution of instructions by directing the coordinated operations of the ALU, registers and other components

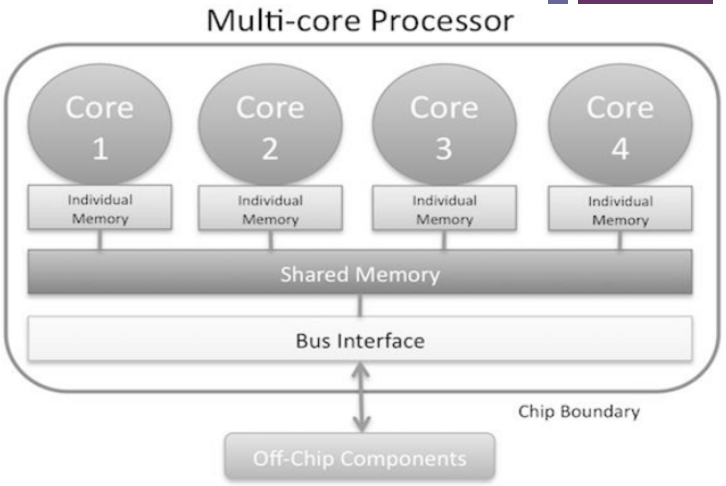
Overview of the CPU

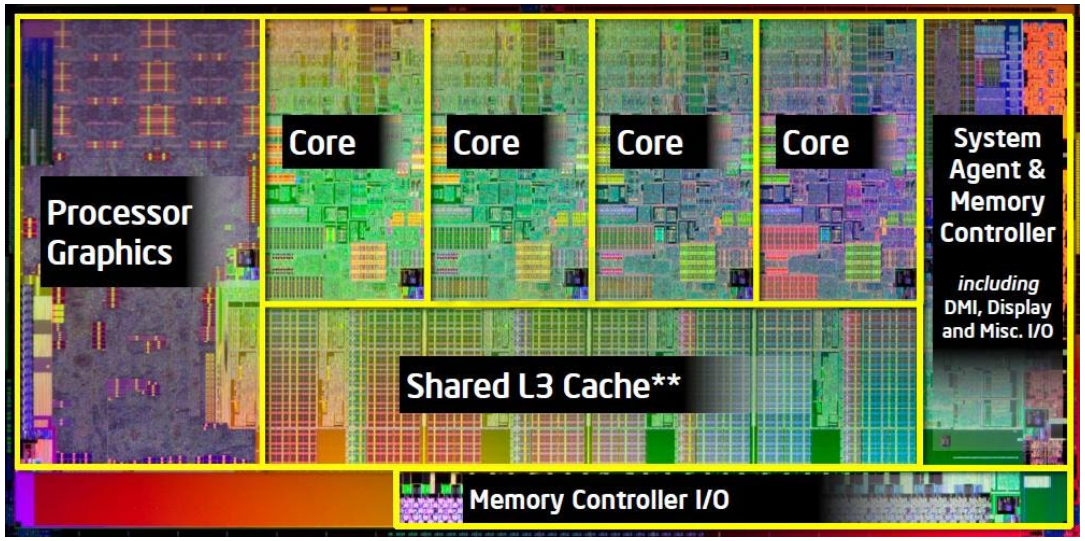


Multi-core Processors

- A **multi-core processor** is a computer processor integrated circuit with **two or more separate processing units**, called **cores**, each of which reads and executes program instructions, **as if the computer had several processors**

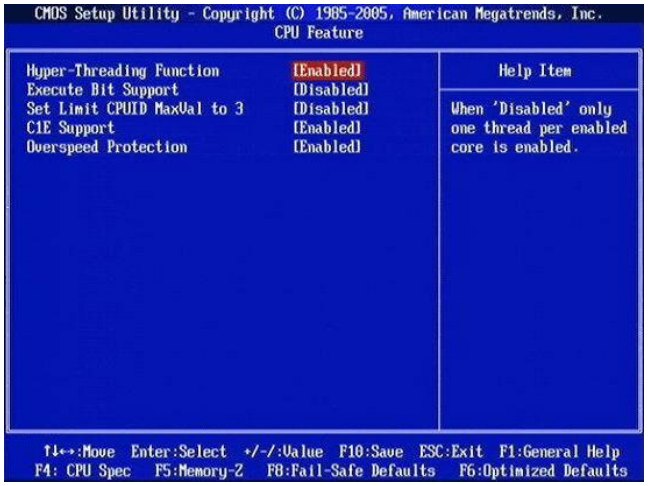
- The instructions are ordinary CPU instructions (such as add, move data, and branch) but the single processor can run instructions on separate cores at the same time
  - increasing overall speed for programs that **support multithreading or other parallel computing techniques**





Hyper-threading technology

- **Hyper-threading** (officially called Hyper-Threading Technology or HT Technology and abbreviated as HTT or HT) is **Intel's proprietary** simultaneous **multithreading** (SMT) implementation used to improve parallelization of computations on x86 microprocessors
- For each processor core that is physically present, the operating system addresses **two virtual (logical) cores** and shares the workload between them when possible
- The main function is to increase the number of independent instructions performed
- Takes advantage of superscalar architecture, in which multiple instructions operate on separate data in parallel.
- One physical core appears as two processors to the operating system, allowing concurrent scheduling of two processes per core
- According to Intel, the first hyper-threading implementation used only 5% more die area than the comparable non-hyperthreaded processor, but the performance was 15–30% better
- In real life, performance improvements are very application-dependent
  - It can also lead to a degradation of performance
    - i.e cache trash
  - Need more careful allocation of processes



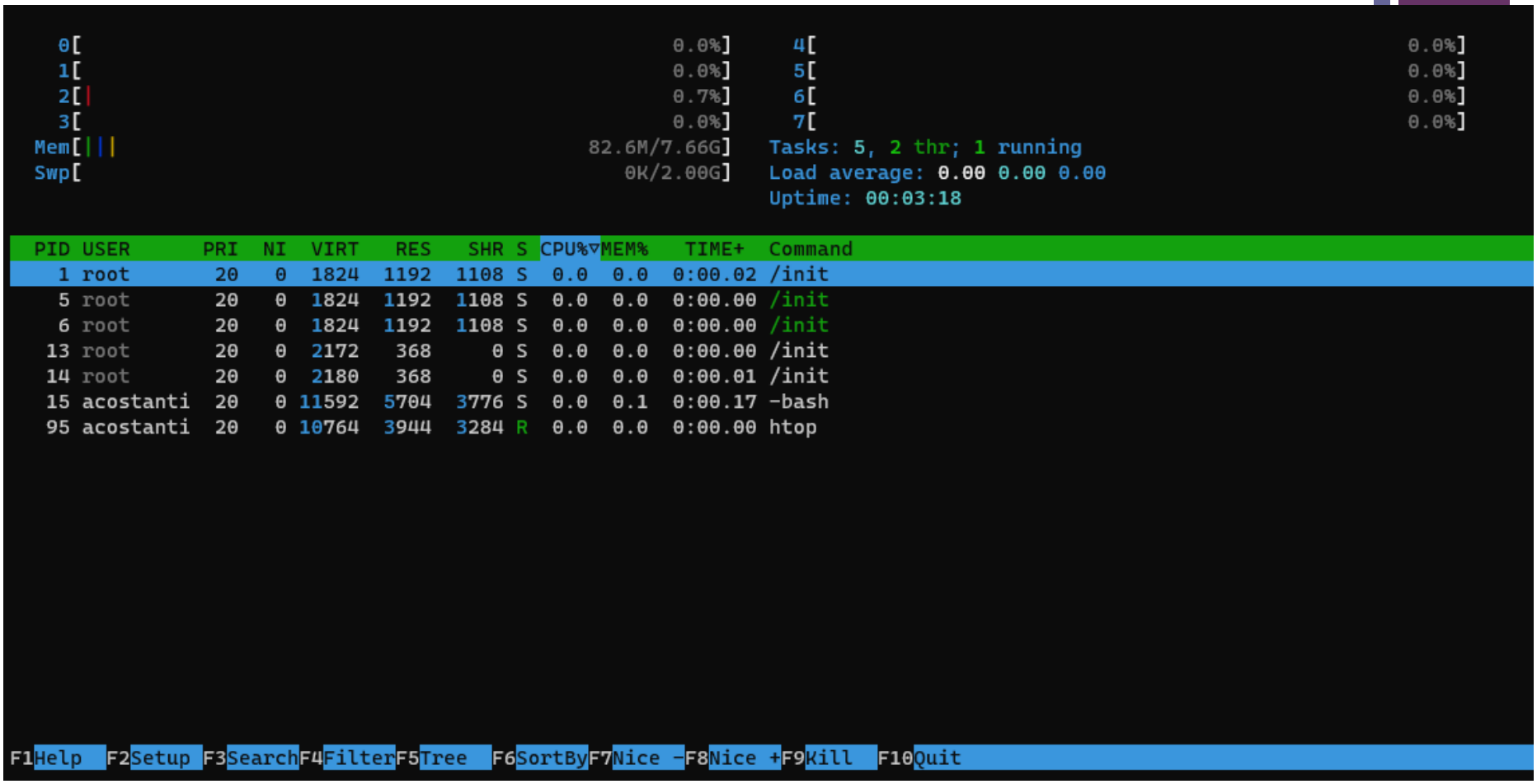
Understanding the top command


```
top - 10:33:46 up 1 min,  0 users,  load average: 0.01, 0.01, 0.00
Tasks:   5 total,   1 running,   4 sleeping,   0 stopped,   0 zombie
%Cpu(s):  0.0 us,  0.0 sy,  0.0 ni,100.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
MiB Mem :   7846.5 total,   7678.2 free,     82.8 used,     85.5 buff/cache
MiB Swap:   2048.0 total,   2048.0 free,      0.0 used.   7592.0 avail Mem

  PID USER      PR  NI    VIRT    RES    SHR S  %CPU  %MEM     TIME+ COMMAND
    1 root      20   0    1824   1192   1108 S   0.0   0.0   0:00.02 init
   13 root      20   0    2172    368      0 S   0.0   0.0   0:00.00 init
   14 root      20   0    2180    368      0 S   0.0   0.0   0:00.00 init
   15 acostan+  20   0   11460   5536   3676 S   0.0   0.1   0:00.08 bash
   90 acostan+  20   0   12944   3928   3340 R   0.0   0.0   0:00.00 top
```

top

System topology example



htop

Cat /proc/cpuinfo


```
processor       : 55
vendor_id       : GenuineIntel
cpu family      : 6
model           : 63
model name      : Intel(R) Xeon(R) CPU E5-2683 v3 @ 2.00GHz
stepping        : 2
microcode       : 57
cpu MHz         : 1200.000
cache size      : 35840 KB
physical id     : 1
siblings        : 28
core id         : 14
cpu cores       : 14
apicid          : 61
initial apicid  : 61
fpu             : yes
fpu_exception   : yes
cpuid level     : 15
wp              : yes
flags           : fpu vme de pse tsc msr pae mce cx8 apic sep mtrr pge mca cmov pat pse36 clflush dts acpi mmx fxsr sse sse2 ss ht tm pbe syscall nx pdpe1gb rdtscp lm constant_tsc arch_
perfmon pebs bts rep_good xtopology nonstop_tsc aperfmperf pni pclmulqdq dtes64 monitor ds_cpl vmx smx est tm2 ssse3 fma cx16 xtpr pdcm pcid dca sse4_1 sse4_2 x2apic movbe popcnt tsc_de
adline_timer aes xsave avx f16c rdrand lahf_lm abm ida arat epb xsaveopt pln pts dtherm tpr_shadow vnmi flexpriority ept vpid fsgsbase bmi1 avx2 smep bmi2 erms invpcid cqm cqm_llc cqm_o
ccup_llc
bogomips        : 3990.48
clflush size    : 64
cache_alignment : 64
address sizes   : 46 bits physical, 48 bits virtual
power management:
```

The file **/proc/cpuinfo** displays what type of processor your system is running including the number of CPUs present

Flags indicate the functionalities and features supported by your processor

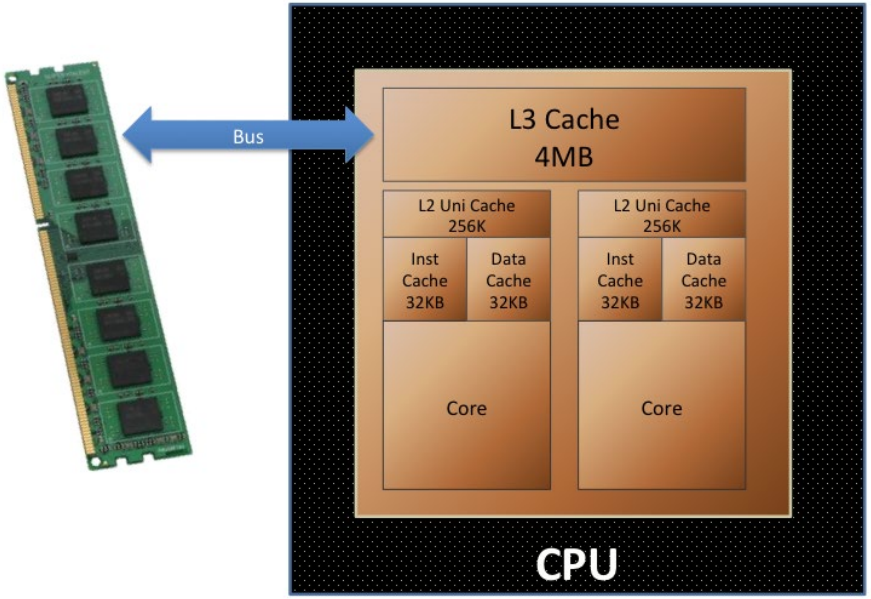
[root@hpc-200-06-29 ~]# less /usr/src/kernels/2.6.32-696.1.1.el6.x86_64/arch/x86/include/asm/cpufeature.h

Server boards

System Memory

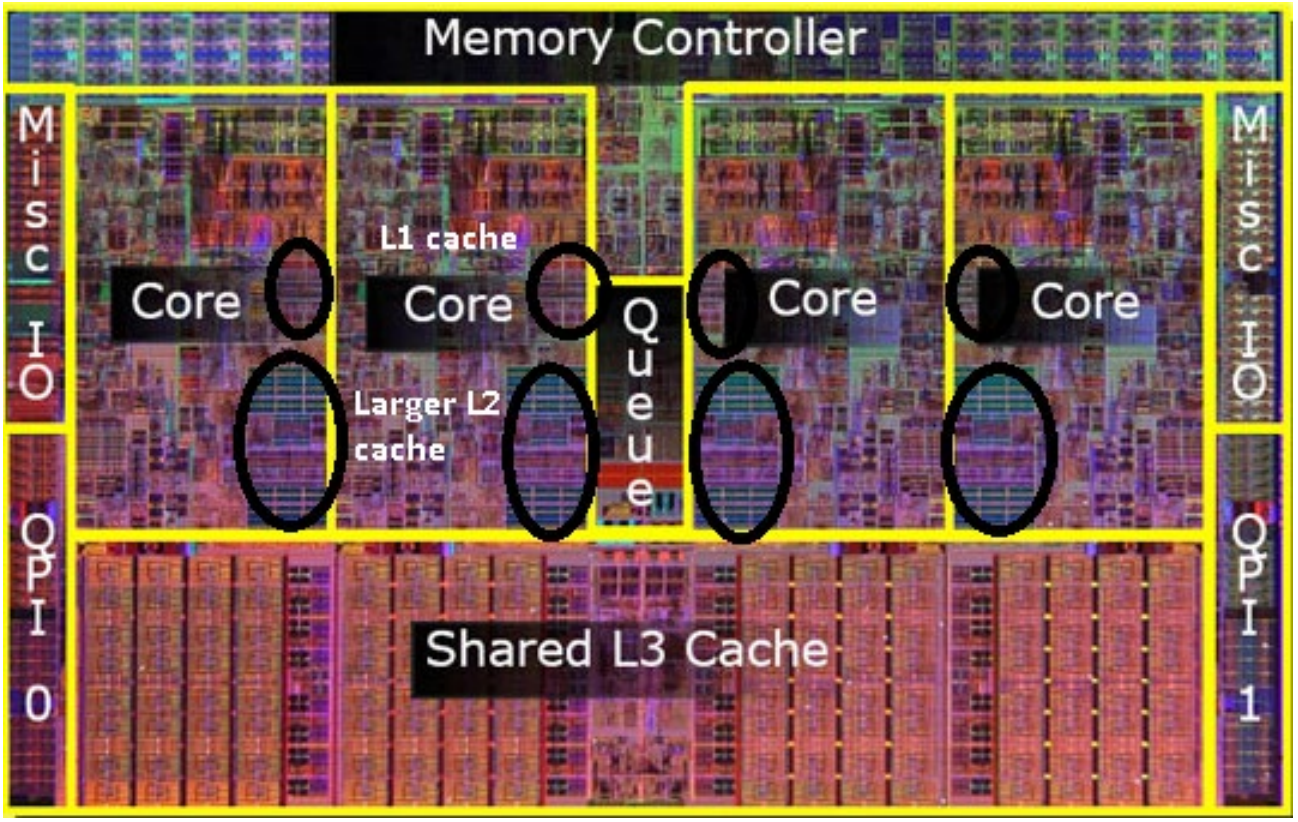
Random-Access Memory (RAM)

- **Random-access memory** is a device that is used to store information for **immediate use**
  - Can be read and changed in any order
  - Typically used to store working **data and machine code**
  - Associated with **volatile** types of memory
  - Many computer systems have a **memory hierarchy** consisting of **processor registers**, on-die **SRAM** caches, external caches, **DRAM, paging systems** and **virtual memory** or **swap space** on a hard drive



https://en.wikipedia.org/wiki/Random-access_memory

Latency Numbers every programmer should know



https://medium.com/software-design/why-software-developers-should-care-about-cpu-caches-8da04355bb8a

L1 cache access latency: 4 clock cycles
L2 cache access latency: 11 clock cycles
L3 cache access latency: 39 clock cycles
Main memory access latency: 107 clock cycles

Latency Numbers every programmer should know

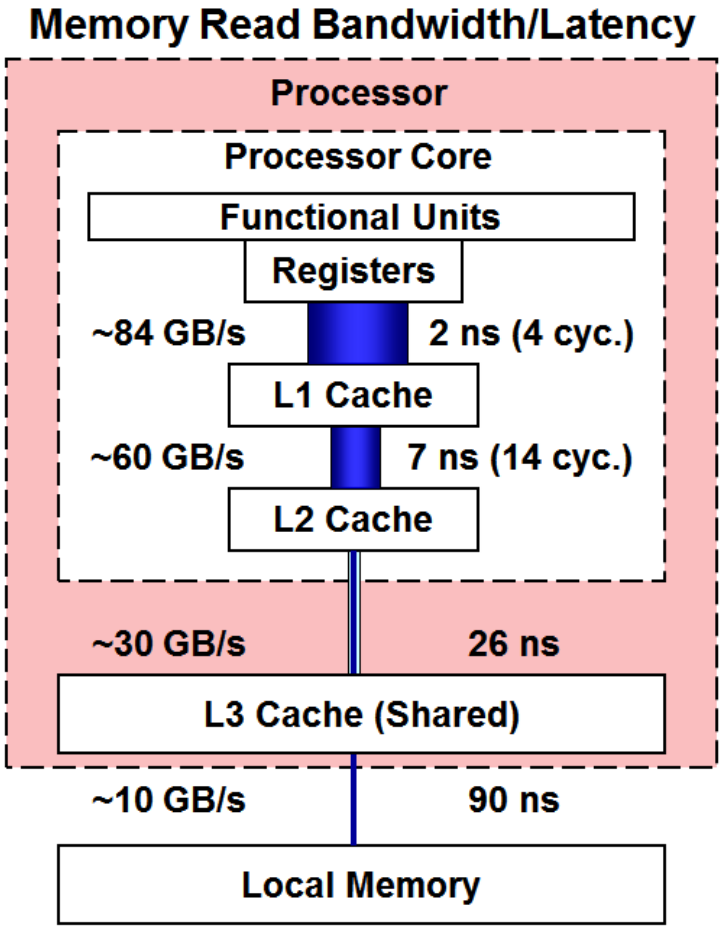

```
Latency Comparison Numbers (~2012)
----------------------------------
L1 cache reference                           0.5 ns
Branch mispredict                            5   ns
L2 cache reference                           7   ns                      14x L1 cache
Mutex lock/unlock                           25   ns
Main memory reference                      100   ns                      20x L2 cache, 200x L1 cache
Compress 1K bytes with Zippy             3,000   ns        3 us
Send 1K bytes over 1 Gbps network       10,000   ns       10 us
Read 4K randomly from SSD*             150,000   ns      150 us          ~1GB/sec SSD
Read 1 MB sequentially from memory     250,000   ns      250 us
Round trip within same datacenter      500,000   ns      500 us
Read 1 MB sequentially from SSD*     1,000,000   ns    1,000 us    1 ms  ~1GB/sec SSD, 4X memory
Disk seek                           10,000,000   ns   10,000 us   10 ms  20x datacenter roundtrip
Read 1 MB sequentially from disk    20,000,000   ns   20,000 us   20 ms  80x memory, 20X SSD
Send packet CA->Netherlands->CA    150,000,000   ns  150,000 us  150 ms

Notes
-----
1 ns = 10^-9 seconds
1 us = 10^-6 seconds = 1,000 ns
1 ms = 10^-3 seconds = 1,000 us = 1,000,000 ns
```

Credit: By Jeff Dean: http://research.google.com/people/jeff/; Originally by Peter Norvig: http://norvig.com/21-days.html#answers
Taken from https://gist.github.com/jboner/2841832

Memory Bandwidth and Latency

- **Memory bandwidth** is the rate at which data can be read from or stored into a semiconductor memory by a processor
  - usually expressed in units of **bytes/second**
- The total bandwidth is the product of:
  - **Base clock frequency**
  - **Number of data transfers per clock**: Two, in the case of "double data rate" (DDR, DDR2, DDR3, DDR4) memory
  - **Memory bus (interface) width**: Each DDR, DDR2, or DDR3 memory interface is 64 bits wide. Those 64 bits are sometimes referred to as a "line."
  - **Number of interfaces**: Modern personal computers typically use two memory interfaces (dual-channel mode) for an effective 128-bit bus width.



For example, a computer with dual-channel memory and one DDR2-800 module per channel running at 400 MHz would have a theoretical maximum memory bandwidth of:

400,000,000 clocks per second × 2 lines per clock × 64 bits per line × 2 interfaces = 102,400,000,000 (102.4 billion) bits per second (in bytes, 12,800 MB/s or 12.8 GB/s)

https://en.wikipedia.org/wiki/Memory_bandwidth

Virtual Memory and RAM Disks

- Most modern operating systems employ a method of **extending RAM capacity**, known as **"virtual memory" (Swap memory)**
  - A portion of the computer's **hard drive** is set aside for a *paging file* or a *scratch partition*
  - the combination of physical RAM and the paging file form the system's total memory
  - Excessive use of this mechanism results in **thrashing** and generally hampers overall system performance
    - hard drives are far slower than RAM

- **RAM Disk -** a portion of a computer's RAM that acts as a much faster hard drive
  - **A RAM disk loses the stored data when the computer is shut down, unless memory is arranged to have a standby battery source.**

"Memory" vs Mass Storage

- In this course do not confuse "Memory" with Mass Storage
- We will indicate with "Memory" the RAM and/or the RAM hierarchy
- With "Storage" we will indicate the mass storage which refers to the storage of large amounts of data in **persisting** media
  - **tape libraries, RAID systems**, and a variety of **computer drives** such as **hard disk drives**, **magnetic tape drives, magneto-optical disc drives, optical disc drives, memory cards, and solid-state drives**

getting the memory status


```
[root@ip-172-31-25-191 ~]# free -m
              total        used        free      shared  buff/cache   available
Mem:           3787         109        2743          16         935        3401
Swap:             0           0           0
```

man free

```
DESCRIPTION
       free displays the total amount of free and used physical and swap memory in the system, as well as the buffers and caches
       used by the kernel. The information is gathered by parsing /proc/meminfo. The displayed columns are:

       total  Total installed memory (MemTotal and SwapTotal in /proc/meminfo)

       used   Used memory (calculated as total - free - buffers - cache)

       free   Unused memory (MemFree and SwapFree in /proc/meminfo)

       shared Memory used (mostly) by tmpfs (Shmem in /proc/meminfo, available on kernels  2.6.32,  displayed  as  zero  if  not
              available)

       buffers
              Memory used by kernel buffers (Buffers in /proc/meminfo)

       cache  Memory used by the page cache and slabs (Cached and Slab in /proc/meminfo)

       buff/cache
              Sum of buffers and cache

       available
              Estimation  of  how much memory is available for starting new applications, without swapping. Unlike the data pro-
              vided by the cache or free fields, this field takes into account page cache and also that not all reclaimable mem-
              ory  slabs  will be reclaimed due to items being in use (MemAvailable in /proc/meminfo, available on kernels 3.14,
              emulated on kernels 2.6.27+, otherwise the same as free)
```

Network

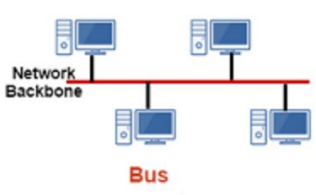
Computer Networks

- A computer network is a digital telecommunications network for sharing resources between **nodes**
- Data transmission between nodes is supported over data links consisting of **physical cable media**, such as **twisted pair** or **fiber-optic** cables, or by wireless methods, such as **Wi-Fi**, **microwave** transmission, or free-space optical communication.
- Network nodes are network computer devices that originate, route and terminate data communication
- Nodes are generally identified by network addresses
  - personal computers, phones, servers, networking hardware such as routers and switches.
  - Two such devices can be said to be networked when one device is able to exchange information with the other device

Network Topology

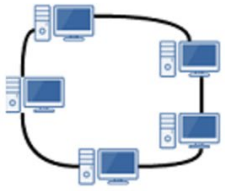
- Network topology is the layout or organizational hierarchy of interconnected nodes of a computer network.

- The topology can affect throughput and reliability
  - i.e. in a bus networks a single failure can cause the network to fail entirely.
  - In general the more interconnections there are, the more robust the network is
    - but the more expensive it is to install, configure, maintain

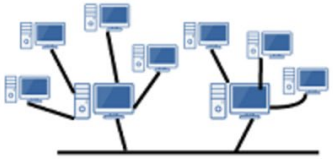


Bus

Star


Ring


Mesh


Tree


Point-to-Point

- **Bus network**: all nodes are connected to a common medium along this medium
  - used in the original Ethernet, called 10BASE5 and 10BASE2
- **Star network**: all nodes are connected to a special central node
  - Typical layout found in a Wireless LAN, where each wireless client connects to the central Wireless access point
- **Ring network:** each node is connected to its left and right neighbor node, such that all nodes are connected, and that each node can reach each other node by traversing nodes left- or rightwards
  - The Fiber Distributed Data Interface (FDDI) made use of such a topology
- **Mesh network:** each node is connected to an arbitrary number of neighbors in such a way that there is at least one traversal from any node to any other
- **Fully connected network:** each node is connected to every other node in the network
- **Tree network:** nodes are arranged hierarchically

The OSI Model

The Open Systems Interconnection model (OSI model) is a conceptual model that characterizes and standardizes the communication functions of a telecommunication or computing system without regard to its underlying internal structure and technology

OSI model				
Layer			Protocol data unit (PDU)	Function[14]
Host layers	7	Application	Data	High-level APIs, including resource sharing, remote file access
	6	Presentation		Translation of data between a networking service and an application; including character encoding, data compression and encryption/decryption
	5	Session		Managing communication sessions, i.e., continuous exchange of information in the form of multiple back-and-forth transmissions between two nodes
	4	Transport	Segment, Datagram	Reliable transmission of data segments between points on a network, including segmentation, acknowledgement and multiplexing
Media layers	3	Network	Packet	Structuring and managing a multi-node network, including addressing, routing and traffic control
	2	Data link	Frame	Reliable transmission of data frames between two nodes connected by a physical layer
	1	Physical	Symbol	Transmission and reception of raw bit streams over a physical medium

- **Layer1: Physical layer** is responsible for the transmission and reception of unstructured raw data between a device and a physical transmission medium. **It converts the digital bits into electrical, radio, or optical signals**. Layer specifications define characteristics such as voltage levels, the timing of voltage changes, physical data rates, maximum transmission distances, modulation scheme, channel access method and physical connectors

- **Layer2: The data link layer** provides node-to-node data transfer. It detects and possibly corrects errors that may occur in the physical layer. It defines the protocol to establish and terminate a connection between two physically connected devices. It also defines the protocol for flow control between them. IEEE 802 divides the data link layer into two sublayers
  - Medium access control (MAC) layer – responsible for controlling how devices in a network gain access to a medium and permission to transmit data.
  - Logical link control (LLC) layer – responsible for identifying and encapsulating network layer protocols, and controls error checking and frame synchronization.
    - The MAC and LLC layers of IEEE 802 networks such as 802.3 Ethernet, 802.11 Wi-Fi, and 802.15.4 ZigBee operate at the data link layer.
    - The Point-to-Point Protocol (PPP) is a data link layer protocol that can operate over several different physical layers, such as synchronous and asynchronous serial lines.

The OSI Model – Layer 3

- **Layer 3. The network layer** provides the functional and procedural means of transferring variable length data sequences (called **packets**) from one node to another connected in "different networks"
  - A network is a medium to which many nodes can be connected, on which every node **has an address**
    - Allow to transfer messages to other nodes connected to it by merely providing the content of **a message and the address of the destination node**
    - Find the way to deliver the message to the destination node, possibly routing it through intermediate nodes
    - If the message is too large may implement message delivery by splitting the message into several fragments at one node, sending the fragments independently, and reassembling the fragments at another node.
    - It may, but does not need to, report delivery errors
  - **Message delivery at the network layer is not necessarily guaranteed to be reliable**; a network layer protocol may provide reliable message delivery, but it need not do so

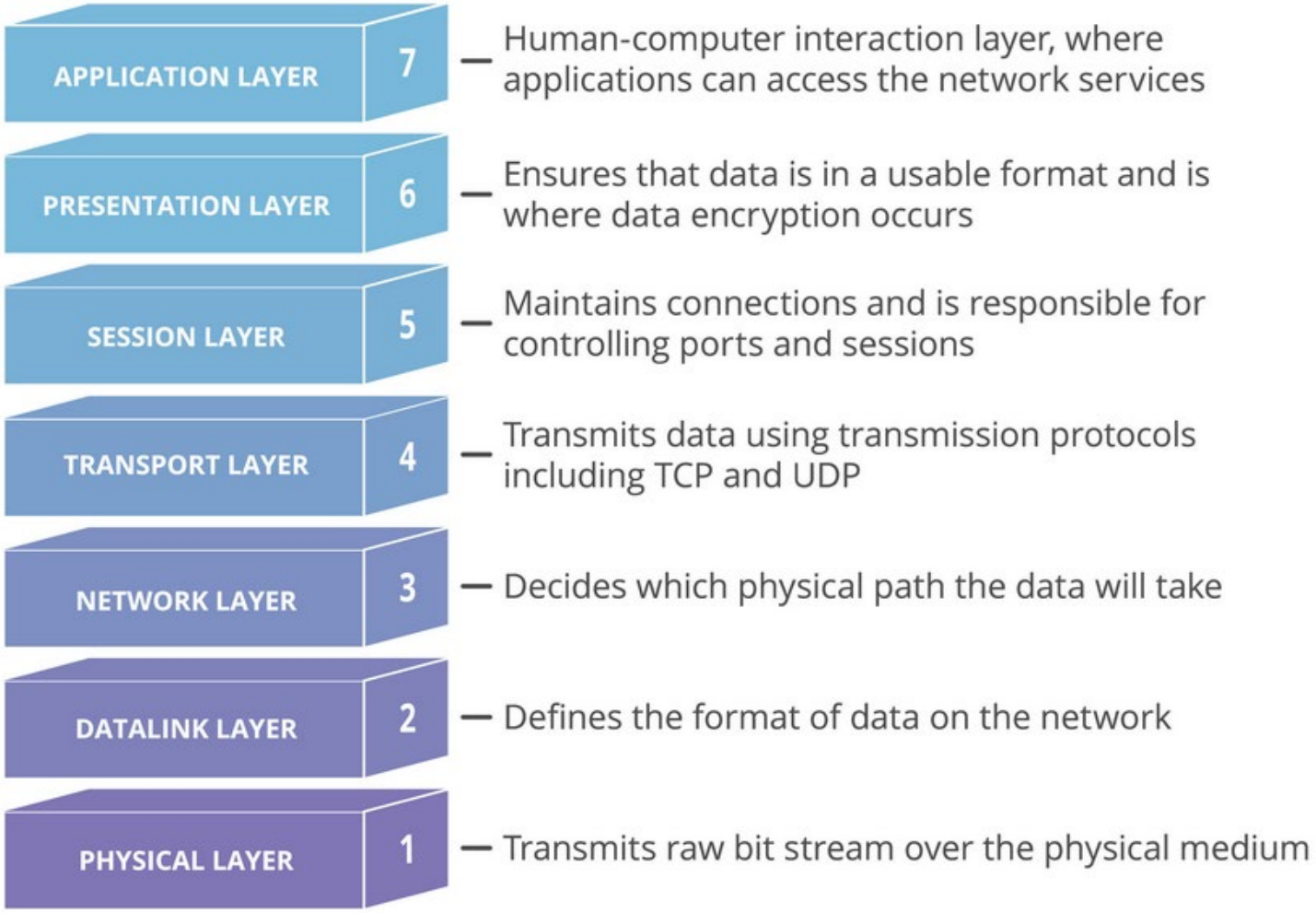
The OSI Model – Layer 4

- **Layer4: The transport layer** provides the functional and procedural means of transferring variable-length data sequences from a source to a destination host, **while maintaining the quality of service functions**.
  - controls the reliability of a given link through flow control, segmentation-desegmentation, and error control.
  - Some protocols are state- and connection-oriented. **This means that the transport layer can keep track of the segments and retransmit those that fail delivery**
  - Also provides the **acknowledgement of the successful data transmission and sends the next data if no errors occurred**.
  - The transport layer creates segments out of the message received from the application layer. Segmentation is the process of dividing a long message into smaller messages

- OSI defines five classes of connection-mode transport protocols ranging from class 0 (which is also known as TP0 and provides the fewest features) to class 4 (TP4, designed for less reliable networks, similar to the Internet)
  - Class 0 contains no error recovery and was designed for use on network layers that provide error-free connections
  - **Class 4 is closest to TCP**

The OSI Model

The following picture borrowed from Cloudflare illustrates pretty well what the OSI model is like:



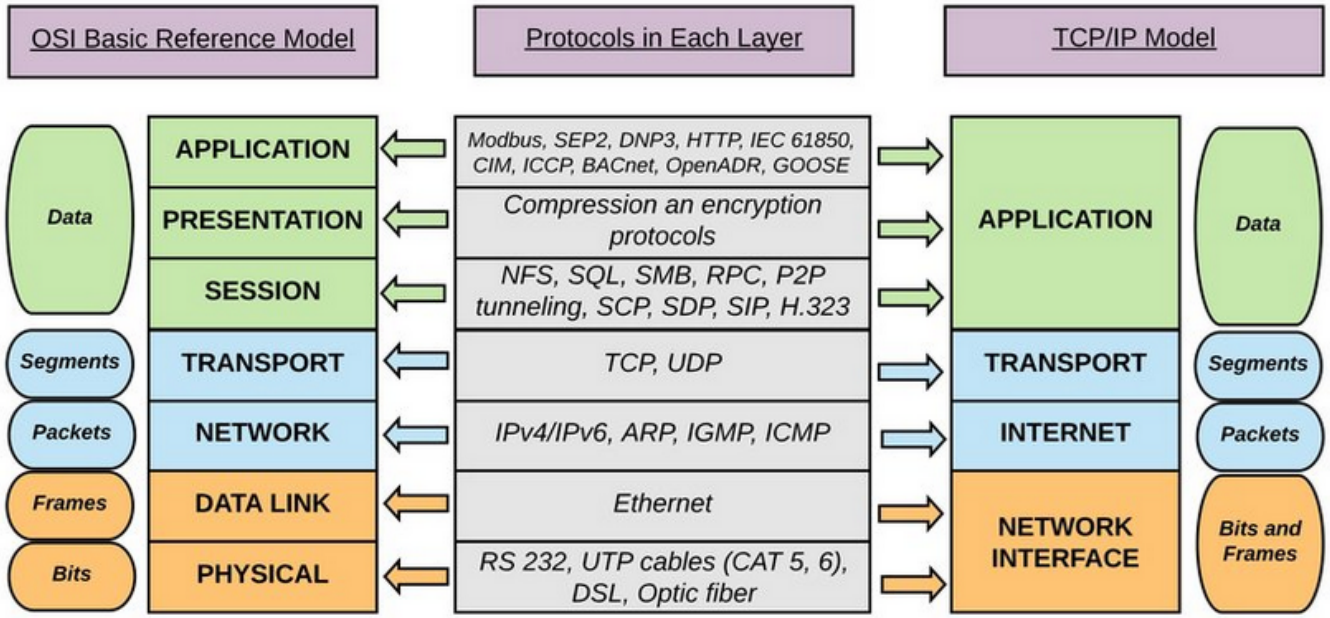
OSI and TCP/IP Model

The OSI model vs the TCP/IP model

While the OSI model is comprehensive reference framework for general networking systems, it's important to mention that the modern Internet doesn't *strictly* follow the OSI model.

The modern Internet more closely follows the simpler *Internet protocol suite*, which is commonly known as *TCP/IP* because the foundational protocols in the suite are the *TCP* (Transmission Control Protocol) and the *IP* (Internet Protocol).

The following image illustrates how the OSI and TCP/IP models relate to each other:



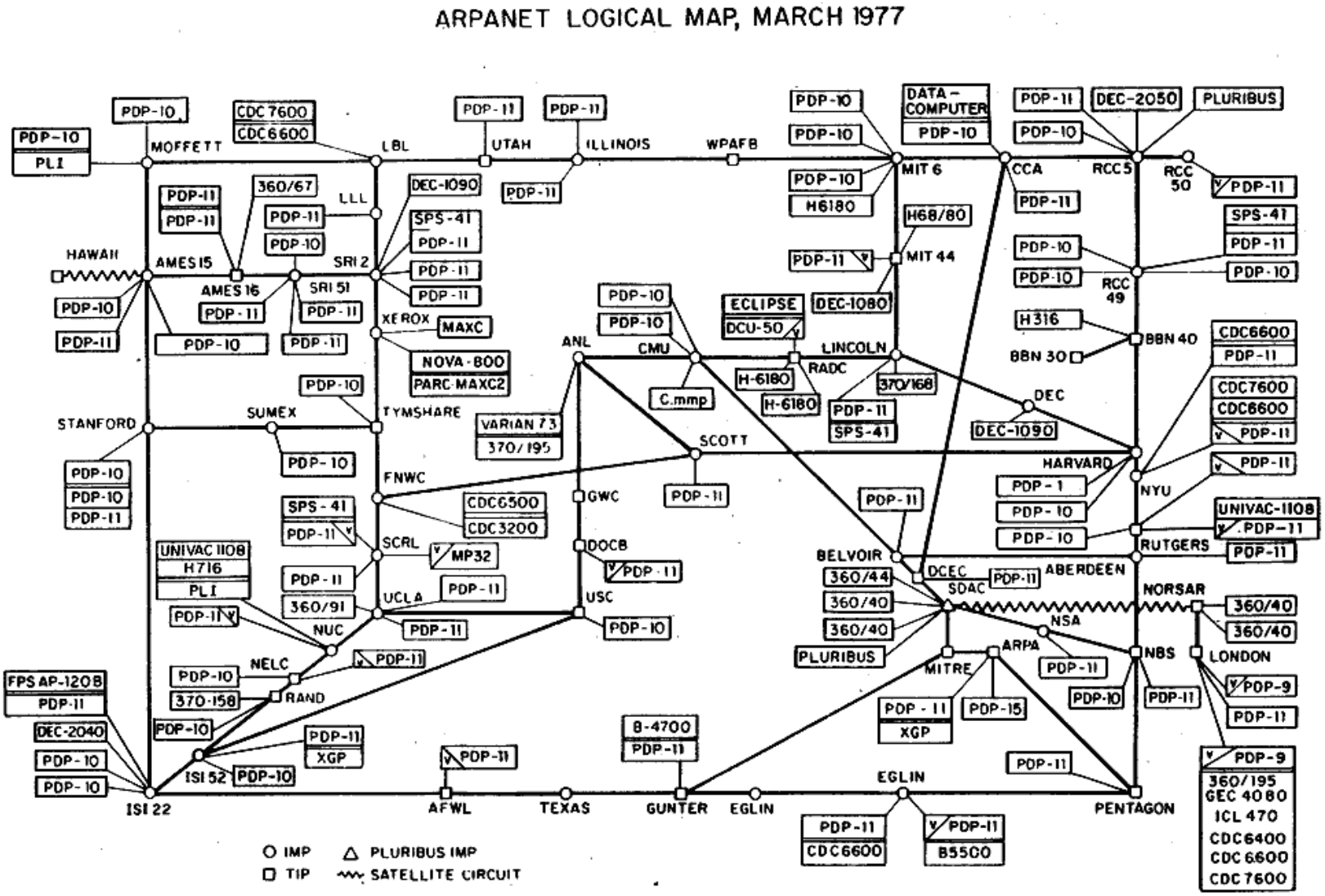
https://stackoverflow.com/questions/38596488/in-which-layer-is-http-in-the-osi-model

LAN and WAN

- **LAN**, which stands for **Local Area Network**, and **WAN**, which stands for **Wide Area Network**, are two types of networks that allow for interconnectivity between computers

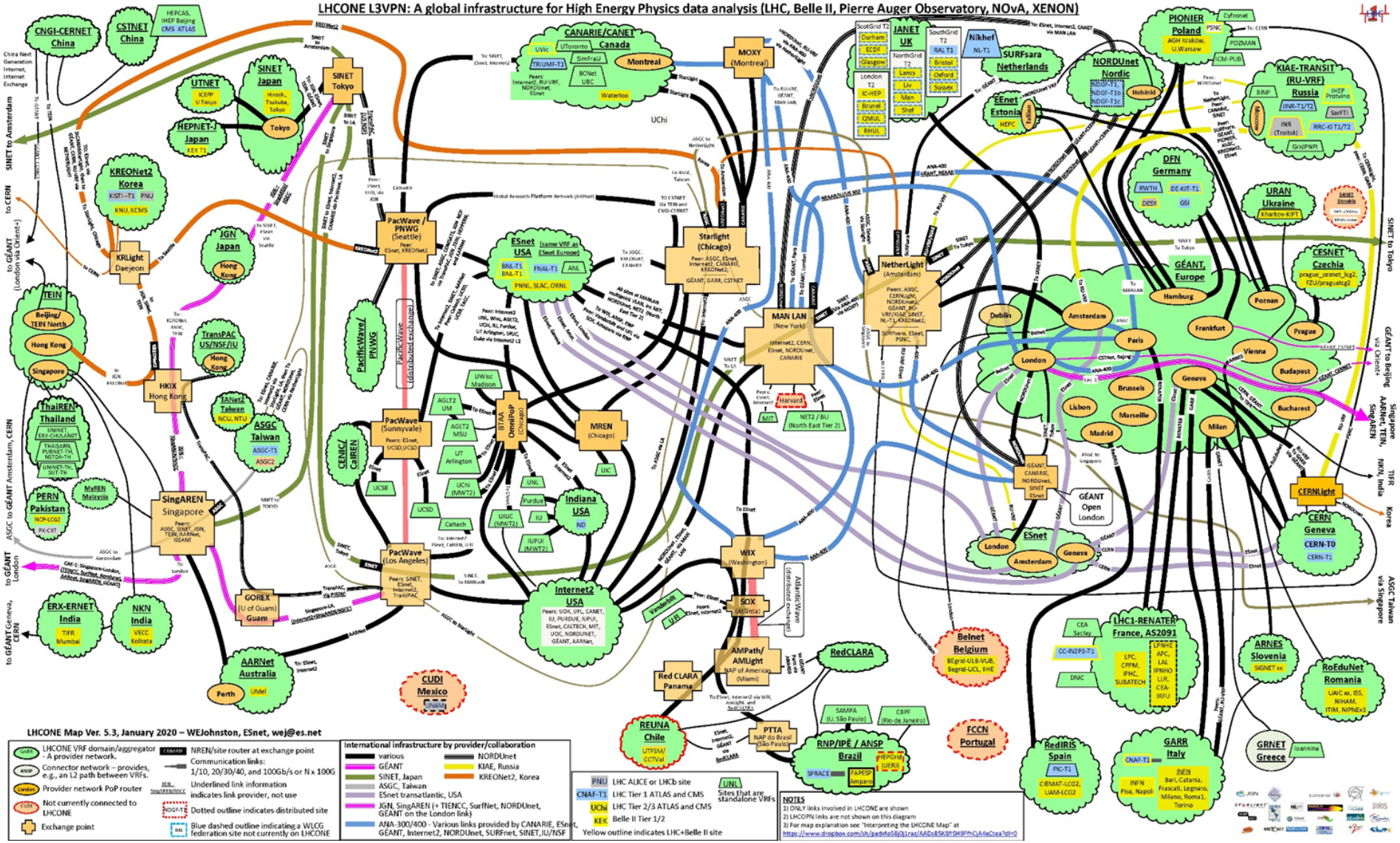
- **LANs are for smaller, more localized networking**
  - in a home, business, school
  - controlled and managed in-house by the organization where they are deployed
  - typically faster and more secure

- **WANs cover larger areas, enabling more widespread connectivity**
  - cities, nations, world
  - typically require two or more of their constituent LANs to be connected over the public Internet or via a private connection established by a third-party telecommunications provider



The **Advanced Research Projects Agency Network (ARPANET)** was the first wide-area packet-switching network with distributed control and the first network to implement the TCP/IP protocol suite. Both technologies became the technical foundation of the Internet. The ARPANET was established by the Advanced Research Projects Agency (ARPA) of the United States Department of Defense.

The LHCONE Network



The objective of **LHCONE** is to provide a collection of access locations that are effectively entry points into a network that is private to the LHC T1/2/3 sites

Packets

- Computer communication links that do not support packets, such as traditional point-to-point telecommunication links, simply **transmit data as a bit stream**
- The overwhelming majority of computer networks carry their data in **packets**
  - **A formatted unit of data**
    - A list of bits, usually a few tens of bytes to a few kilobytes long carried by a packet-switched network
    - Sent through the network to their destination
    - A longer message is packetized before it is transferred and once the packets arrive, they are then reassembled back into their original message
  - **Packets consist of two kinds of data**
    - control information
      - provides data the network needs to deliver the user data, for example, source and destination network addresses, error detection codes, and sequencing information
    - user data (payload)

MTU

- Maximum Transmission Unit **is the size of the largest protocol data unit (PDU) that can be communicated in a single network layer transaction**
  - Relates to, but is not identical to the maximum frame size that can be transported on **the data link layer**, e.g. Ethernet frame
- Larger MTU is associated with reduced overhead
- Smaller MTU values can reduce network delay
- In many cases, MTU is dependent on underlying network capabilities and must be adjusted manually or automatically so as to not exceed these capabilities

MAC Address

- **A media access control address (MAC address)** is **a unique identifier** assigned to a network interface controller (NIC) for use as a network address in communications **within a network segment**
  - Used in most IEEE 802 networking technologies, including Ethernet, Wi-Fi, and Bluetooth
  - Within the OSI network model, MAC addresses are **used in the medium access control protocol sublayer of the data link layer**
  - MAC addresses are recognizable as **six groups of two hexadecimal digits**, separated by hyphens, colons, or without a separator.
- MAC addresses are primarily assigned by device manufacturers
  - Refereed as burned-in address, Ethernet hardware address, hardware address, and physical address
  - Stored in hardware, such as the card's read-only memory, or by a firmware mechanism

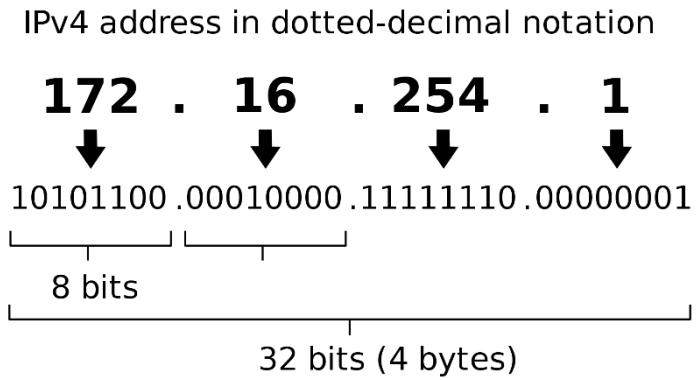
©https://en.wikipedia.org/wiki/MAC_address

Network Addresses in IPv4

- An **Internet Protocol address (IP address)** is a numerical label assigned to each device connected to a computer network that uses the Internet Protocol for communication
- Serves two main functions
  - host or network interface identification
  - location addressing
- Internet Protocol version 4 (IPv4) defines an IP address as a **32-bit number**
- IP addresses are written and displayed in human-readable notations, such as 172.16.254.1
  - The size of the routing prefix of the address is designated in **CIDR notation** by suffixing the address with the number of significant bits, e.g., 192.168.1.15/24, which is equivalent to the historically used **subnet mask** 255.255.255.0

Subnets and Subnet Mask

- IP networks may be divided into **subnetworks**
- an IP address is recognized as consisting of two parts
  - the network prefix in the high-order bits
  - the remaining bits called the rest field, host identifier
    - used for host numbering within a network
- The **subnet mask** or **CIDR notation** determines how the IP address is divided into network and host parts
  - The IP address is followed by a slash and the number (in decimal) of bits used for the network part, also called the routing prefix.
    - For example, an IPv4 address and its subnet mask may be 192.0.2.1 and 255.255.255.0, respectively. The CIDR notation for the same IP address and subnet is 192.0.2.1/24, because the first 24 bits of the IP address indicate the network and subnet



IPv4 Private Addresses

- Early network design, when **global end-to-end connectivity was envisioned for communications with all Internet hosts**, intended that IP addresses be globally unique
- However, it was found that this was not always necessary as **private networks developed and public address space needed to be conserved**
- Computers not connected to the Internet, **need not have globally unique IP addresses**
  - Today, such private networks are widely used and **typically connect to the Internet with network address translation (NAT),** when needed
- **Three non-overlapping ranges of IPv4 addresses for private networks are reserved**
  - These addresses **are not routed on the Internet** and thus their use need not be coordinated with an IP address registry.
  - Any user may use any of the reserved blocks. Typically, a network administrator will divide a block into subnets; for example, many home routers automatically use a default address range of 192.168.0.0 through 192.168.0.255 (192.168.0.0/24)

Name	CIDR block	Address range	Number of addresses
24-bit block	10.0.0.0/8	10.0.0.0 – 10.255.255.255	16 777 216
20-bit block	172.16.0.0/12	172.16.0.0 – 172.31.255.255	1 048 576
16-bit block	192.168.0.0/16	192.168.0.0 – 192.168.255.255	65 536

Network addresses in IPv6

- In **the Internet Protocol version 6** (IPv6), the address size was increased from **32 bits in IPv4 to 128 bits**
  - thus providing up to 2^{128} (approximately 3.403×1038) addresses. This is deemed sufficient for the foreseeable future
- The intent of the new design was not to provide just a sufficient quantity of addresses, but also redesign routing in the Internet by allowing more efficient aggregation of subnetwork routing prefixes.
- The large number of IPv6 addresses allows large blocks to be assigned for specific purposes and, where appropriate, to be aggregated for efficient routing
- A **unique local address (ULA)** in IPv6 address in the address range fc00::/7.
  - Its purpose is analogous to IPv4 private network addressing.
    - Unique local addresses may be used freely, without centralized registration, inside a single site or organization or spanning a limited number of sites or organizations.
    - They are routable only within the scope of such private networks, but not in the global IPv6 Internet.

Understanding ifconfig


```
[root@hpc-200-06-26 ~]# ifconfig
eth0: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 9000
        inet 131.154.184.40  netmask 255.255.255.0  broadcast 131.154.184.255
        inet6 fe80::56ab:3aff:fe13:d701  prefixlen 64  scopeid 0x20<link>
        ether 54:ab:3a:13:d7:01  txqueuelen 1000  (Ethernet)
        RX packets 636902536  bytes 2327846580678 (2.1 TiB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 466620543  bytes 2532678687597 (2.3 TiB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0
        device memory 0xc7b80000-c7bfffff

ib0: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 2044
        inet 192.168.1.26  netmask 255.255.255.0  broadcast 192.168.1.255
        inet6 fe80::211:7500:6f:2656  prefixlen 64  scopeid 0x20<link>
Infiniband hardware address can be incorrect! Please read BUGS section in ifconfig(8).
        infiniband 80:00:00:03:FE:80:00:00:00:00:00:00:00:00:00:00:00:00:00:00  txqueuelen 256  (InfiniBand)
        RX packets 370945  bytes 20984381 (20.0 MiB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 370803  bytes 22374820 (21.3 MiB)
        TX errors 0  dropped 15 overruns 0  carrier 0  collisions 0

lo: flags=73<UP,LOOPBACK,RUNNING>  mtu 65536
        inet 127.0.0.1  netmask 255.0.0.0
        inet6 ::1  prefixlen 128  scopeid 0x10<host>
        loop  txqueuelen 1  (Local Loopback)
        RX packets 5501370  bytes 855876440 (816.2 MiB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 5501370  bytes 855876440 (816.2 MiB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0
```

The loopback device

- The **loopback device** is a special, virtual network interface that **your computer uses to communicate with itself**
  - Used mainly for diagnostics and troubleshooting, and to connect to servers running on the local machine
  - does not represent any actual hardware, but exists so applications running on your computer can always connect to servers on the same machine
  - Important for troubleshooting
    - it can be compared to looking in a mirror
  - Helpful when a server offering a resource you need *is running on your own machine*

- For IPv4, the loopback interface is assigned all the IPs in the 127.0.0.0/8 address block. That is, 127.0.0.1 through 127.255.255.254 all represent your computer
  - For most purposes, though, it is only necessary to use one IP address, and that is 127.0.0.1
    - **This IP has the hostname of localhost mapped to it**

The DNS

- The **Domain Name System** (DNS) is a hierarchical and decentralized naming system for computers, services, or other resources connected to the Internet or a private network.

- It associates various information with domain names assigned to each of the participating entities.
  - It translates more **readily memorized domain names** to the **numerical IP addresses** needed for locating and identifying computer services and devices with the underlying network protocols

```
[root@ip-172-31-25-191 ~]# ping www.google.com
PING www.google.com (172.217.164.164) 56(84) bytes of data.
64 bytes from iad23s69-in-f4.1e100.net (172.217.164.164): icmp_seq=1 ttl=51 time=0.909 ms
64 bytes from iad23s69-in-f4.1e100.net (172.217.164.164): icmp_seq=2 ttl=51 time=0.947 ms
64 bytes from iad23s69-in-f4.1e100.net (172.217.164.164): icmp_seq=3 ttl=51 time=0.991 ms
64 bytes from iad23s69-in-f4.1e100.net (172.217.164.164): icmp_seq=4 ttl=51 time=0.960 ms
64 bytes from iad23s69-in-f4.1e100.net (172.217.164.164): icmp_seq=5 ttl=51 time=0.951 ms
64 bytes from iad23s69-in-f4.1e100.net (172.217.164.164): icmp_seq=6 ttl=51 time=0.968 ms
```

```
[root@ip-172-31-25-191 ~]# host  www.google.com
www.google.com has address 172.217.164.164
www.google.com has IPv6 address 2607:f8b0:4004:c09::67
[root@ip-172-31-25-191 ~]#
```

Network Protocols

- A **communication protocol** is a set of rules for exchanging information over a network.
  - In a protocol stack (also see the OSI model), each protocol leverages the services of the protocol layer below it
    - i.e. A protocol stack is **HTTP** (the World Wide Web protocol) running over **TCP** over **IP** (the Internet protocols) over **IEEE 802.11** (the Wi-Fi protocol).
      - This stack is used between the wireless **router** and the home user's personal computer when the user is surfing the web
- **IEEE 802** is a family of IEEE standards dealing with local area networks and metropolitan area networks.
  - **Ethernet**, sometimes simply called *LAN*, is a family of protocols used in wired LANs, described by a set of standards together called IEEE 802.3 published by the Institute of Electrical and Electronics Engineers
  - **Wireless LAN**, also widely known as WLAN or WiFi, is probably the most well-known member of the IEEE 802 protocol family for home users today. It is standardized by IEEE 802.11 and shares many properties with wired Ethernet

Network Bandwidth

- The term *bandwidth* defines the **net bit rate,** 'peak bit rate', 'information rate,' or physical layer 'useful bit rate', **channel capacity**, or the **maximum throughput** of a logical or physical communication path in a digital communication system
  - *Channel bandwidth* may be confused with useful data throughput (or goodput)
    - For example, a channel with *x* bps may not necessarily transmit data at *x* rate, since protocols, encryption, and other factors can add appreciable overhead

56 kbit/s	Modem / Dialup
1.5 Mbit/s	ADSL Lite
1.544 Mbit/s	T1/DS1
2.048 Mbit/s	E1 / E-carrier
4 Mbit/s	ADSL1
10 Mbit/s	Ethernet
11 Mbit/s	Wireless 802.11b
24 Mbit/s	ADSL2+
44.736 Mbit/s	T3/DS3
54 Mbit/s	Wireless 802.11g
100 Mbit/s	Fast Ethernet
155 Mbit/s	OC3
600 Mbit/s	Wireless 802.11n
622 Mbit/s	OC12
1 Gbit/s	Gigabit Ethernet
1.3 Gbit/s	Wireless 802.11ac
2.5 Gbit/s	OC48
5 Gbit/s	USB 3.0
7 Gbit/s	Wireless 802.11ad
9.6 Gbit/s	OC192
10 Gbit/s	10 Gigabit Ethernet, USB 3.1
40 Gbit/s	Thunderbolt 3
100 Gbit/s	100 Gigabit Ethernet

Network Latency

- **End-to-end delay** or **one-way delay** (**OWD**) refers to the time taken for a packet to be transmitted across a network from source to destination.
- Round-Trip Time (RTT) is the time taken to reach the destination and come back to source
- The **ping utility** measures the RTT, that is, the time to go and come back to a host.
  - Half the RTT is often used as an approximation of OWD but this assumes that the forward and back paths are the same in terms of congestion, number of hops, or quality of service (QoS).
    - This is not always a good assumption

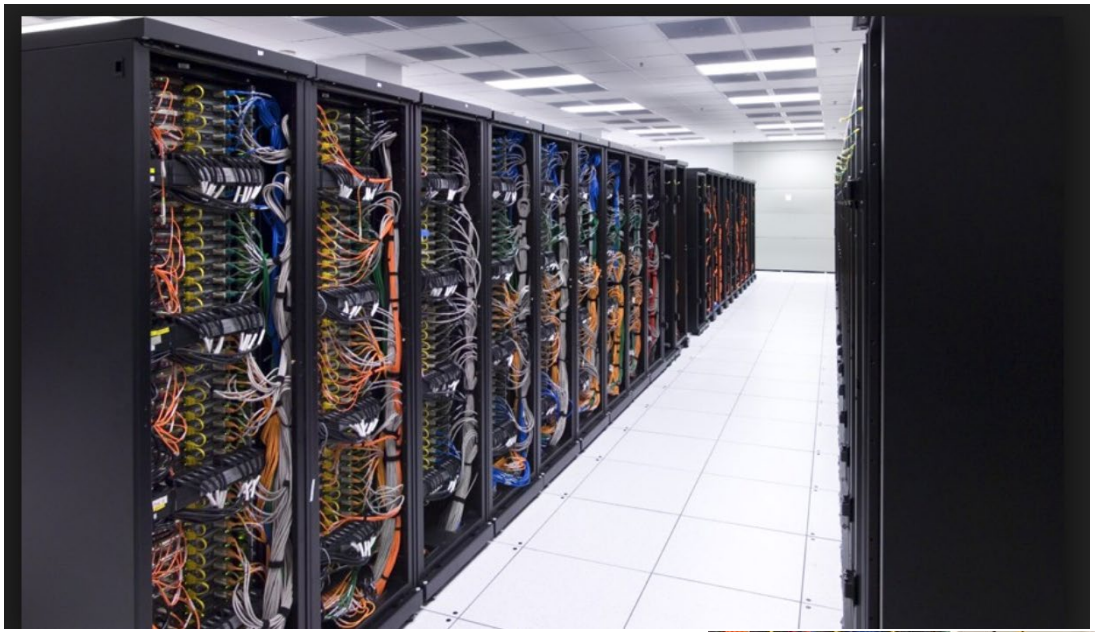
Network Adapters & PCI interface

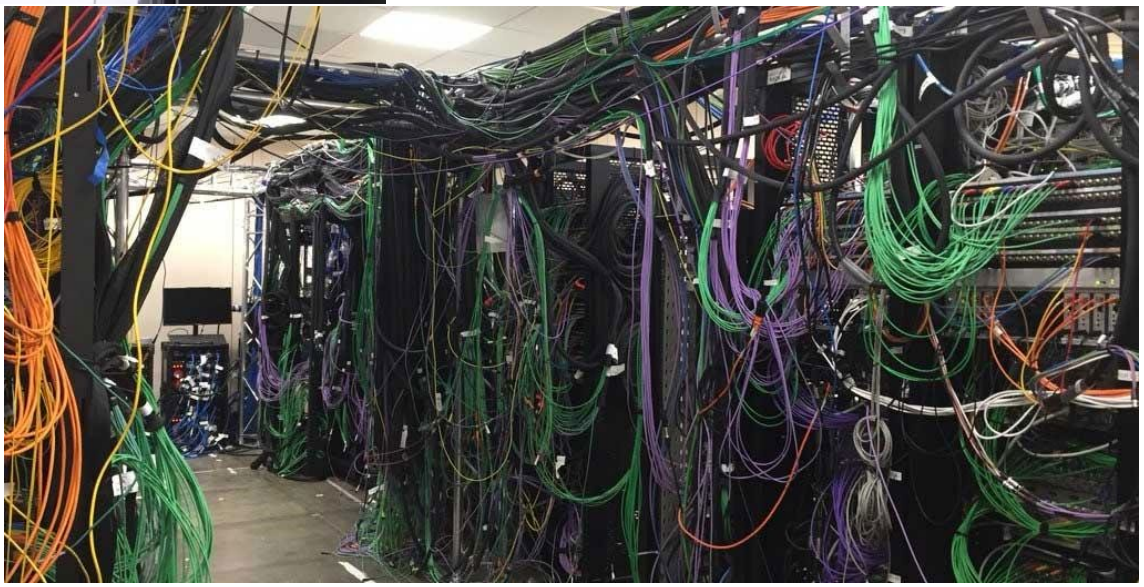
- Infiniband
- Omnipath
- Fiber Channel
- Ethernet

Transceiver and cables

- Ethernet
- Infininband/omnipath
- SAS
- Fiber

Cabling….is important





Some link performances

	Gbit/s	MB/s
USB2	0.4	50
USB3	3.2	400
SATA3	4.8	600
PCIe 3.0 x1	7.88	985
PCIe 3.0 x4	31.52	3940
PCIe 3.0 x8	63.04	7880
PCIe 3.0 x16	126.08	15760
Eth 1Gb	1	125
Eth 10Gb	10	1250
	Gbit/s	**MB/s**
IB SDR / FC-8	8	1000
IB DDR /FC-16	16	2000
IB QDR /FC-32	32	4000
IB FDR	54	6750
IB EDR	96	12000

50-125 µs latency (*)

<5 µs latency (*)

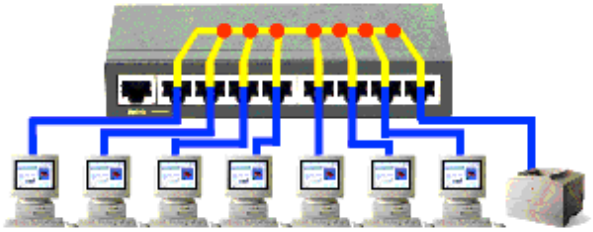
(*) https://www.cavium.com/Documents/TechnologyBriefs/Adapters/Tech_Brief_Introduction_to_Ethernet_Latency.pdf

Infiniband/Omnipath performance

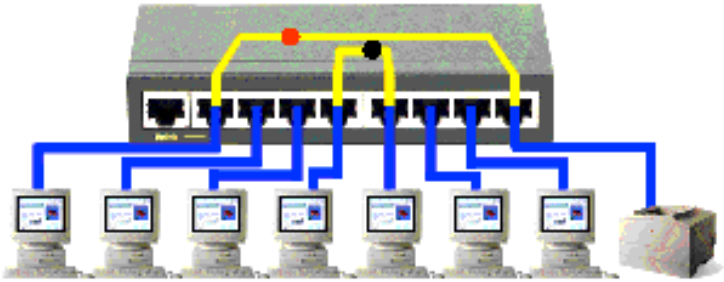
	Gbit/s (4x)	Gbit/s (12x)	latency (µs)
IB SDR	8	24	5
IB DDR	16	48	2.5
IB QDR	32	96	1.3
IB FDR-10	40	120	0.7
IB FDR	54.54	163.62	0.7
IB EDR	96.97	290.91	0.5
IB HDR	200	600	0.2
	Gbit/s		
OMP	100		0.5

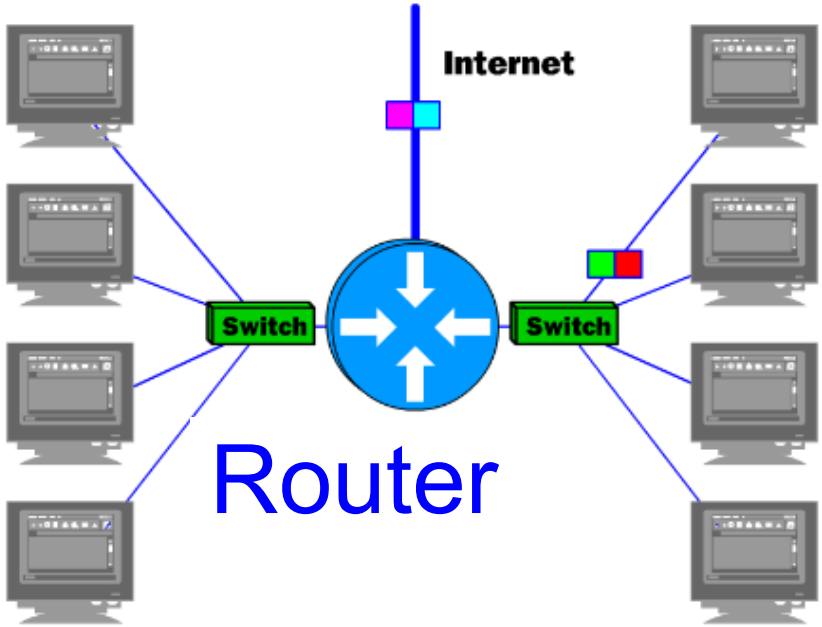
Hub, switch and router

Hub



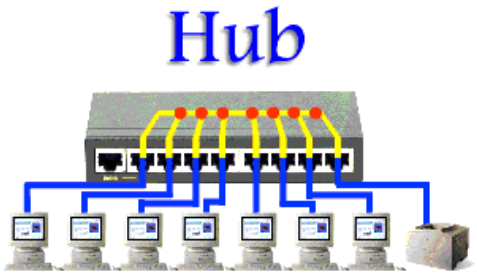
Switch





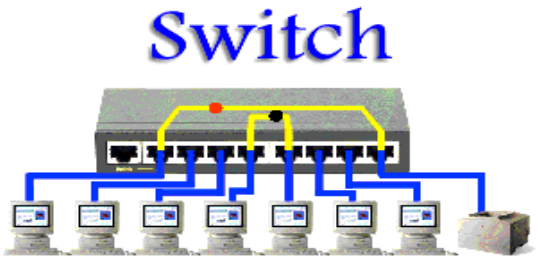
Hub, Switch and Router

- Hub
  - Used to connect segments of a LAN (Local Area Network)
  - When a packet arrives at one port, it is copied to the other ports so that all segments of the LAN can see all packets (broadcast)



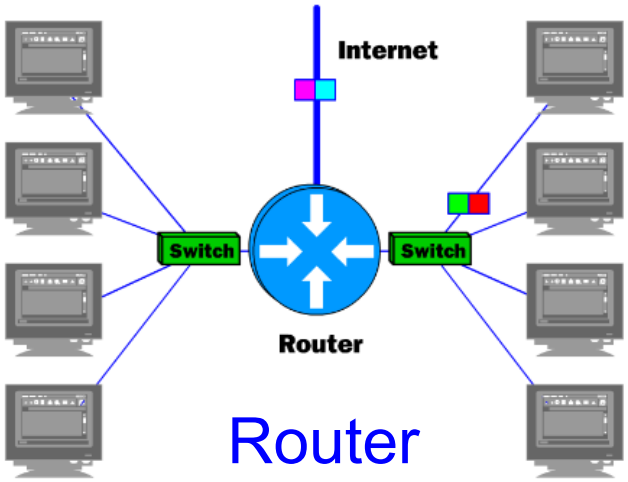
Hub, Switch and Router

- Switch
  - Operates at the data link layer (layer 2) and sometimes the network layer (layer 3) of the OSI (Open Systems Interconnection) Reference Model
  - Keeps a record of the MAC (Media Access Control) addresses of all the devices connected to it
  - Can identify which system is sitting on which port
  - When a frame is received, it knows exactly which port to send it to, without significantly increasing network response times



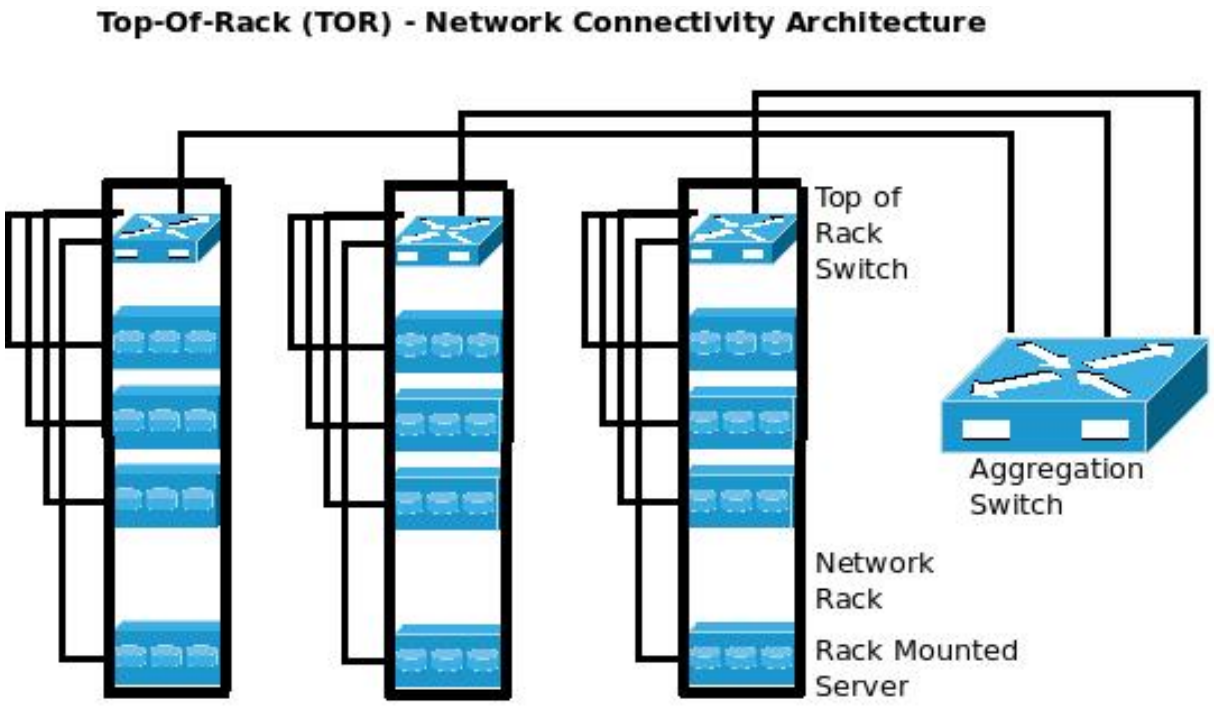
Hub, Switch and Router

- Router
  - Is connected to at least two networks, commonly two LANs or WANs (Wide Area Networks) or a LAN and its ISP (Internet Service Provider) network
  - Route packets to other networks until that packet ultimately reaches its destination

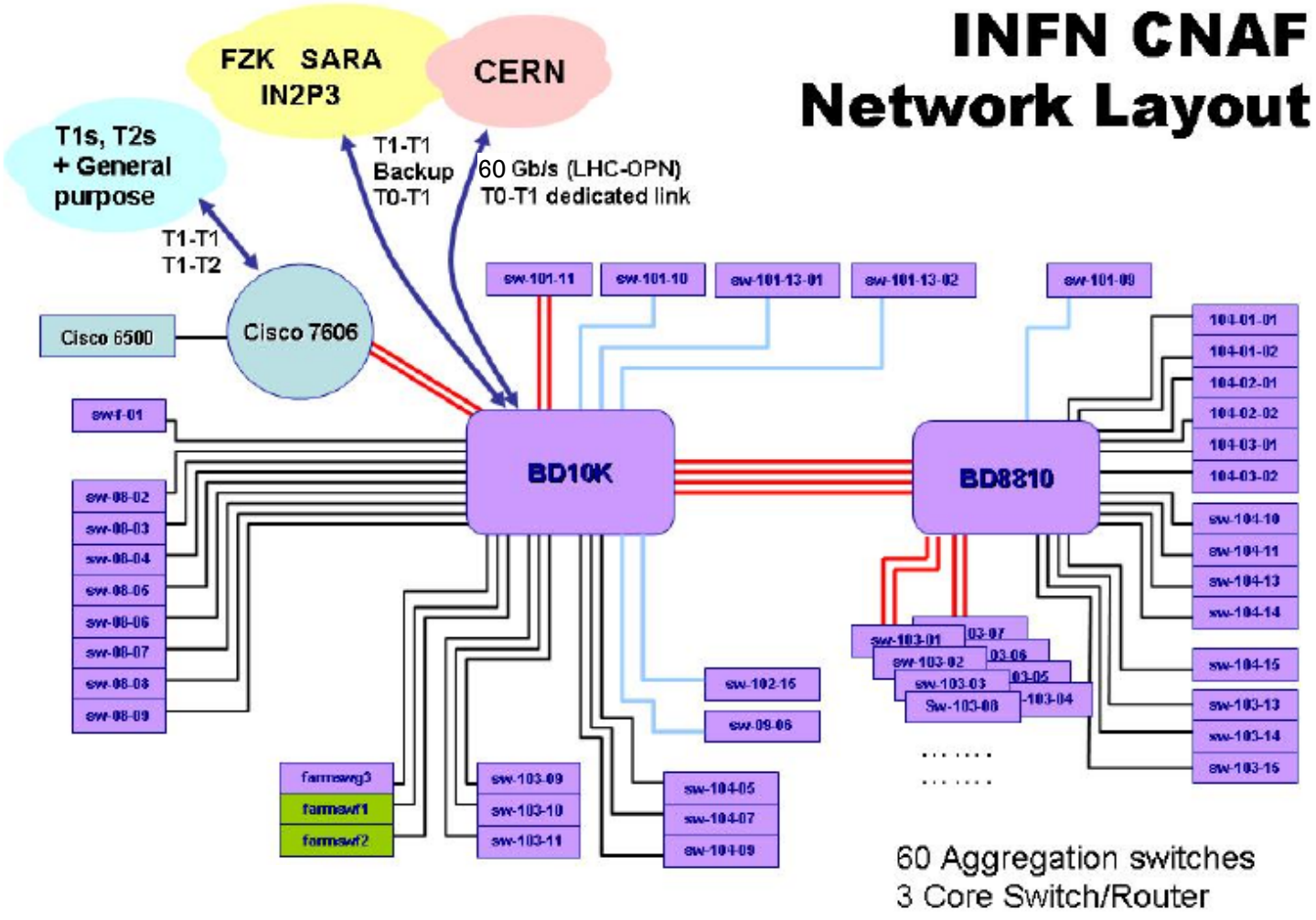


Top-of-the-rack switching

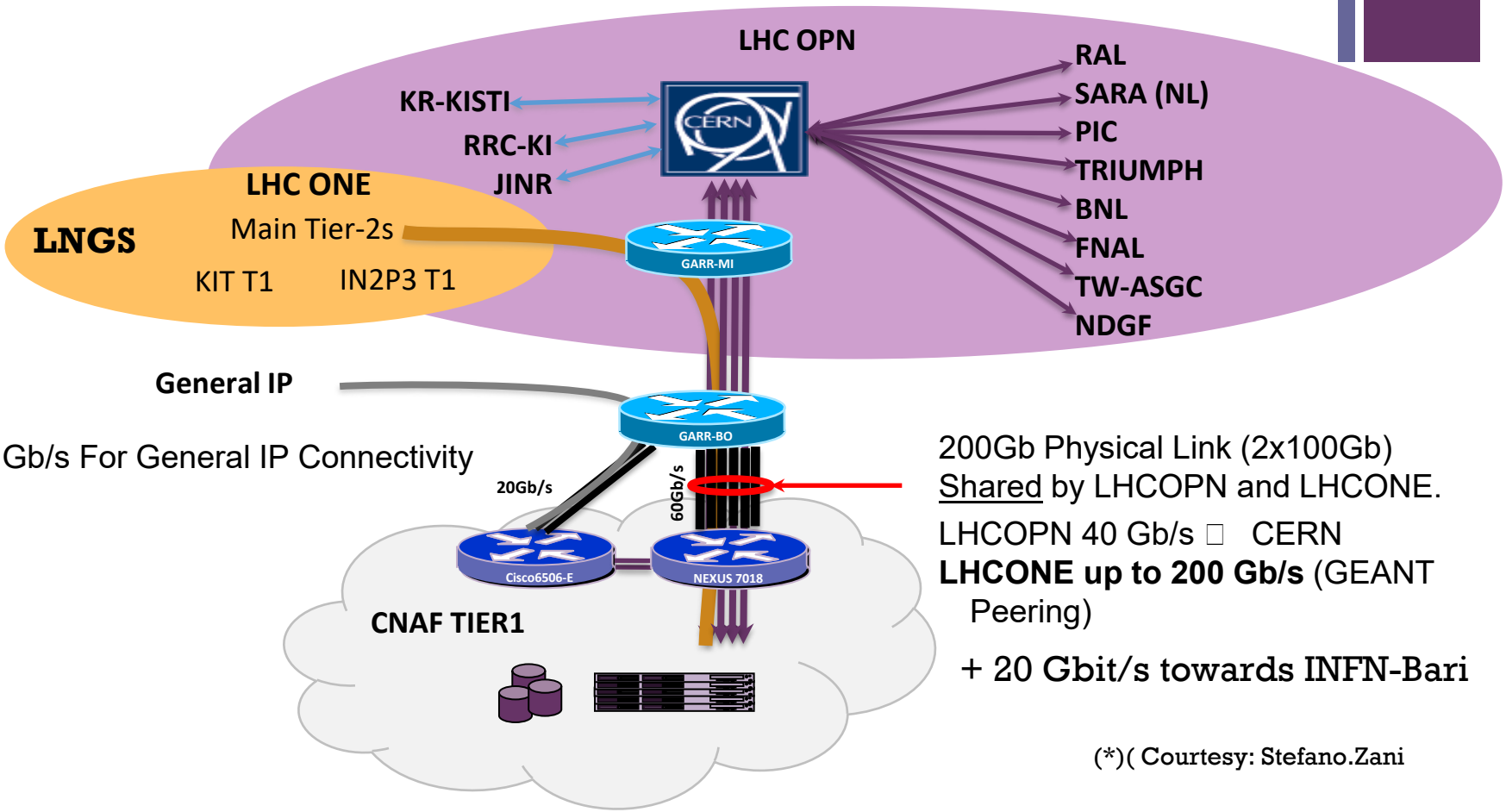
- Top-of-rack switching is a network architecture design in which computing equipment like servers, appliances and other switches located within the same or adjacent rack are connected to an in-rack network switch
- The in-rack network switch, in turn, is connected to aggregation switches via fiber optic cables



INFN-CNAF Tier1 Network topology



WAN@CNAF (Status)



60 Gb/s For General IP Connectivity

200Gb Physical Link (2x100Gb)
Shared by LHCOPN and LHCONE.

LHCOPN 40 Gb/s CERN

LHCONE up to 200 Gb/s (GEANT Peering)

+ 20 Gbit/s towards INFN-Bari

(*)(Courtesy: Stefano.Zani

Networking Useful Tools

- Measurement
  - ping
  - Traceroute/tracepath
- Status
  - ifconfig / ip addr show
  - ethtool
  - netstat
- Lookup
  - host
  - dig
  - whois

Computing

Computing Farm

- A collection of computing servers
  - It can reach millions of CPU cores
- Provides the computing power to the datacenter
- Connected with network switches and/or routers which enable communication between the different parts of the cluster and the users of the cluster

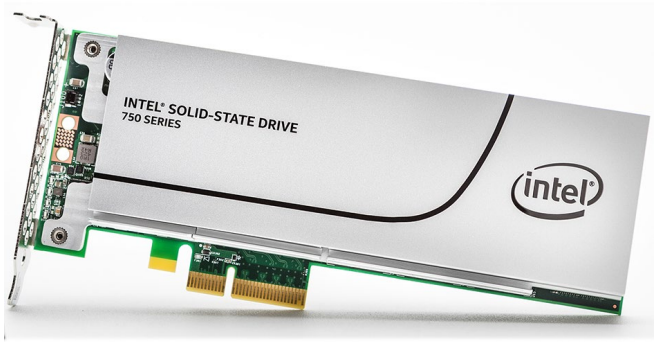
Managing the Concurrent Access

- Typically in a datacenter multiple users share the same resources
  - Optimize resource usage and avoid waste of computing power
  - Different users could have paid different shares
  - Different users could have different priorities

Storage
… a taste

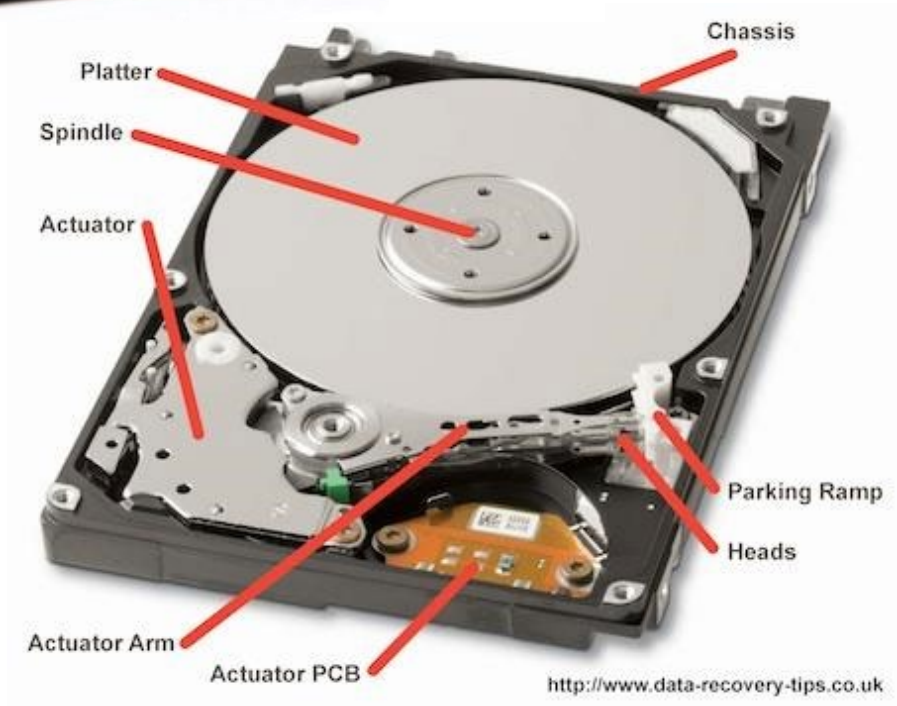
PC/Server storage

- Spinning disks
- Solid State Disks
- NVMe
- Tapes

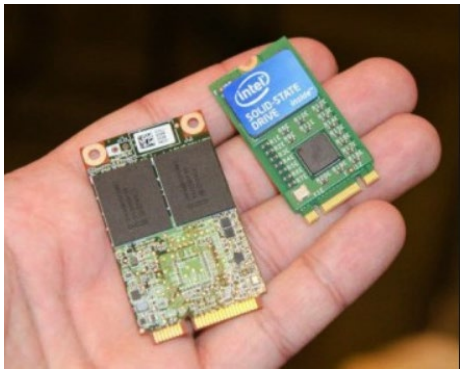


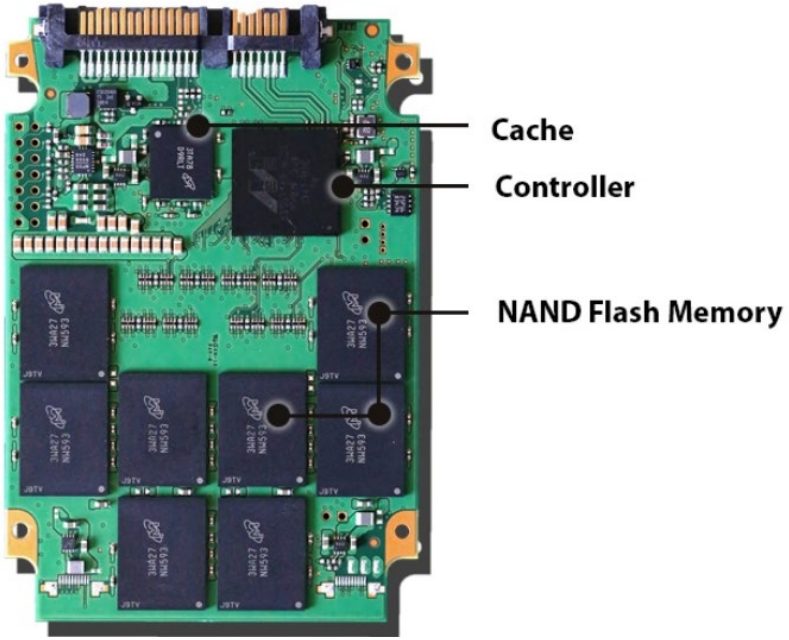
See also:
https://en.wikipedia.org/wiki/IOPS

Inside an hard disk

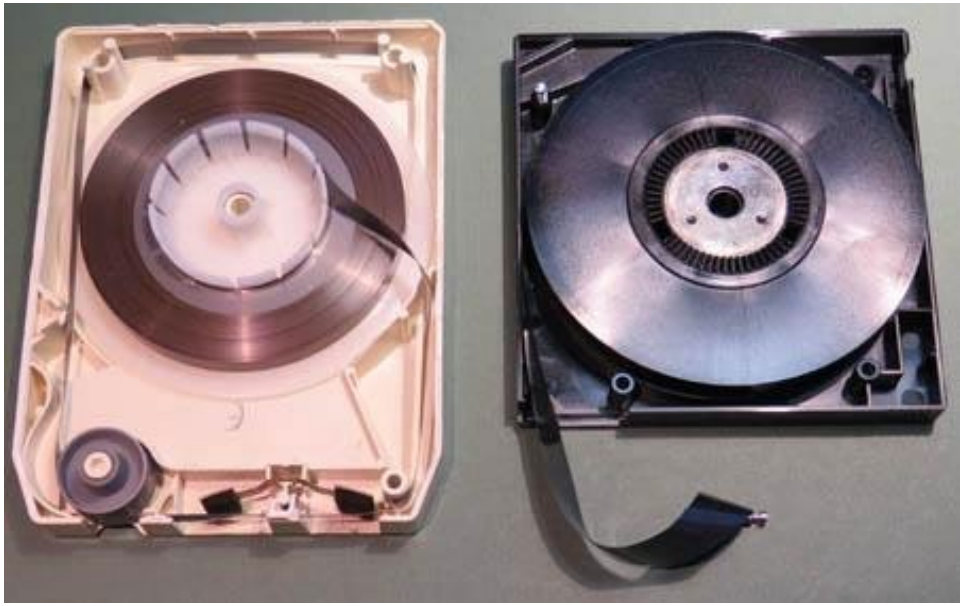


Inside an SSD





Inside a Tape Cartridge



Back to the ‘80s….

Commodore’s datassette: a 90-minutes tape (45 minutes on each side) will hold on the order of 150 kilobytes on each side if no compression or fast loader is used.

Storage systems for a PC or server

Media	MB/s	IOPS	Capacity	Cost (Purchase)
HDD – Seagate Archive	100-150	100-200	8TB	$
SSD – Samsung EVO	400-500	100k-400k	500GB	$
NVMe Intel 400	2000 (read)	450k	400GB/1TB	$$
NVMe Violin 6000	4000	1M+	10TB	$$$
Tape T10000D	250	sequential	8.5TB	($$$)* Including driver and library

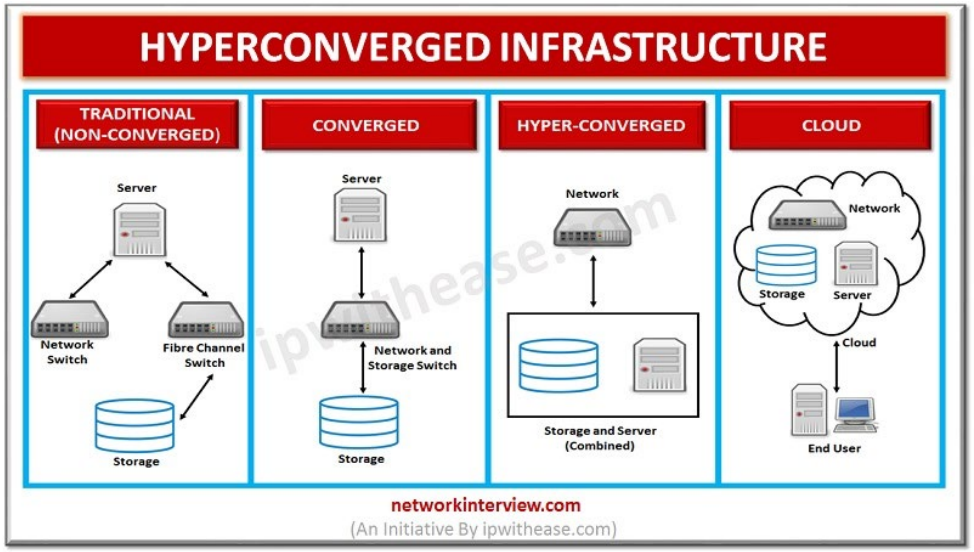
Different Quality of Services….different prices

Modern DataCenter for BigData Applications

HyperConverged architecture

Hyperconverged data center is a data center that is built on hyperconverged infrastructure (HCI), which is a software architecture that consolidates the compute, network, and storage commodity hardware.

- CPU power and storage converged towards the same building blocks of the datacenter
- Suitable for BigData analytics and MapReduce Applications
- Can be scaled easily



Provisioning and Monitoring

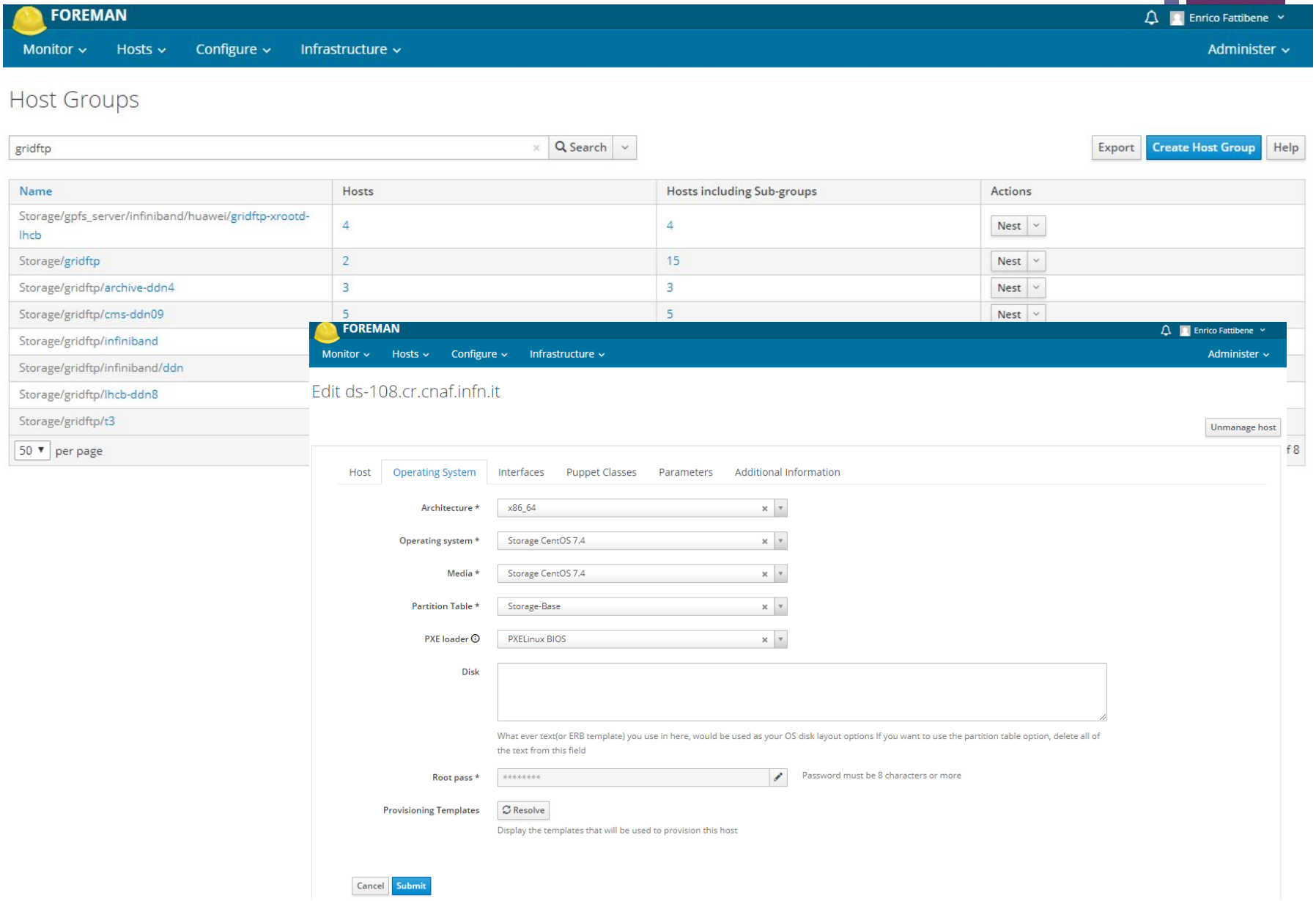
Automatic management of the datacenter resources

Automatic management of the datacenter resources

Provisioning

- Provisioning of thousands of servers in a complex datacenter can be facilitated by the usage of tools for automatic installation and configuration
- Foreman
  - Lifecycle management of bare metal or virtual servers
  - Open source
  - Web interface, API and command line
  - Definition of OS, installation media, kickstart file, etc.
    - For host groups or single hosts
  - Integrated with Puppet acts as External Node Classifier

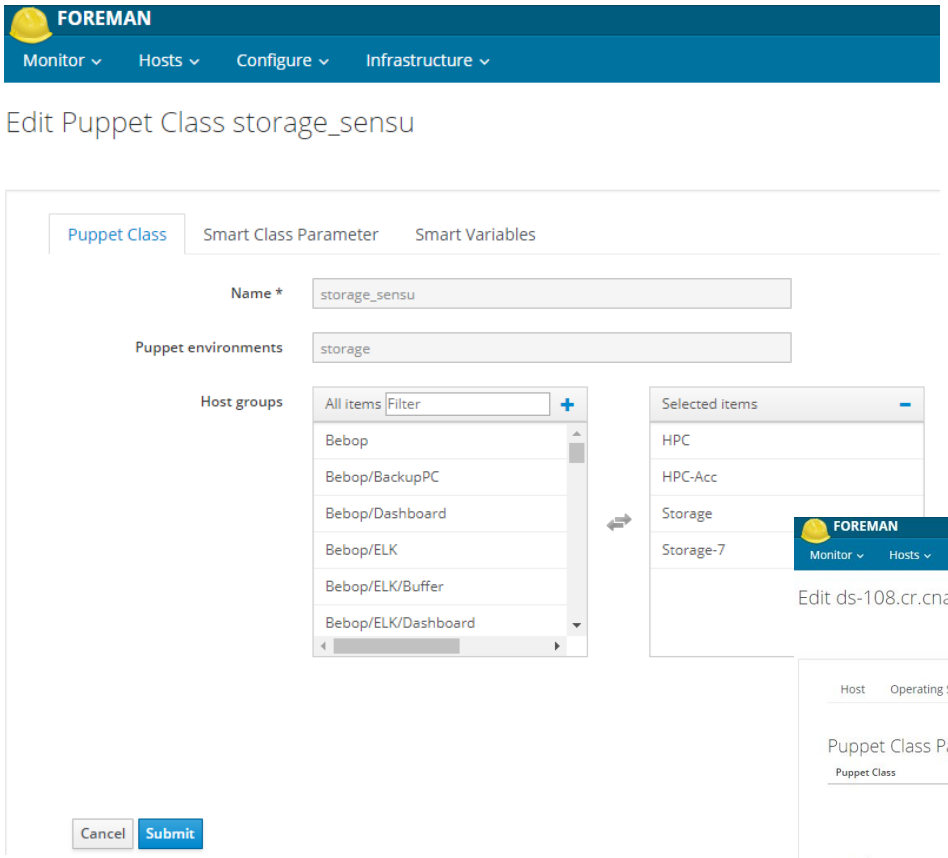
Foreman

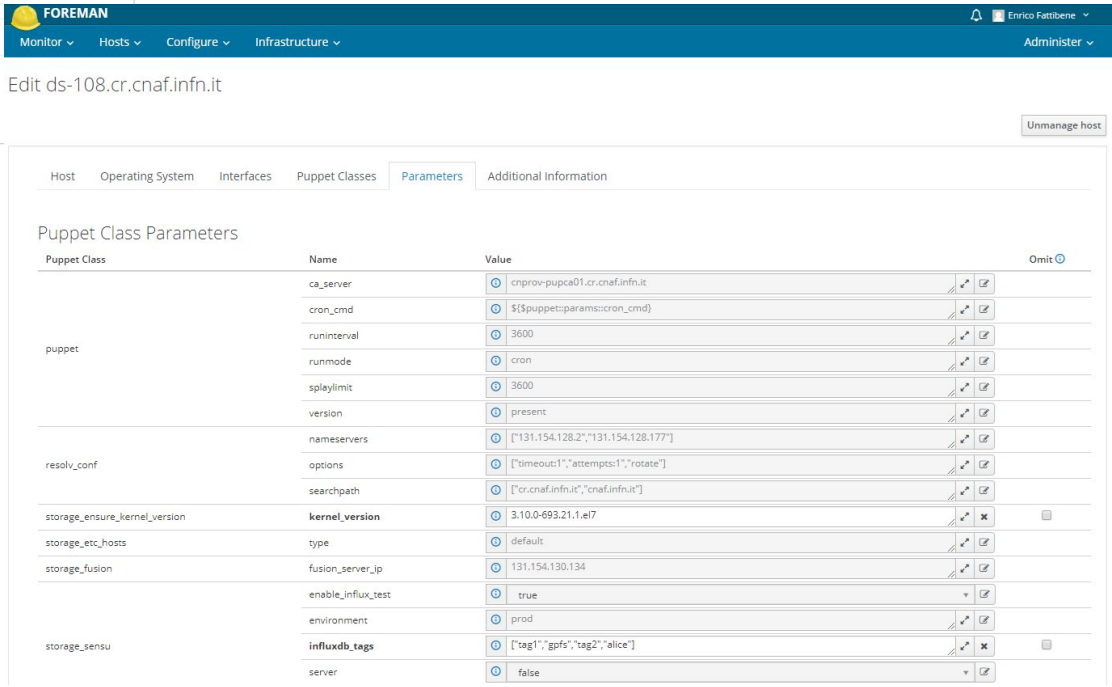

FOREMAN
Enrico Fattibene
Monitor
Hosts
Configure
Infrastructure
Administer
Host Groups
gridftp
Search
Export
Create Host Group
Help
Name
Hosts
Hosts including Sub-groups
Actions
Storage/gpfs_server/infiniband/huawei/gridftp-xrootd-lhcb
4
4
Nest
Storage/gridftp
2
15
Nest
Storage/gridftp/archive-ddn4
3
3
Nest
Storage/gridftp/cms-ddn09
5
5
Nest
Storage/gridftp/infiniband
Storage/gridftp/infiniband/ddn
Storage/gridftp/lhcb-ddn8
Storage/gridftp/t3
50
per page
FOREMAN
Enrico Fattibene
Monitor
Hosts
Configure
Infrastructure
Administer
Edit ds-108.cr.cnaf.infn.it
Unmanage host
Host
Operating System
Interfaces
Puppet Classes
Parameters
Additional Information
Architecture *
x86_64
Operating system *
Storage CentOS 7.4
Media *
Storage CentOS 7.4
Partition Table *
Storage-Base
PXE loader
PXELinux BIOS
Disk
What ever text(or ERB template) you use in here, would be used as your OS disk layout options If you want to use the partition table option, delete all of the text from this field
Root pass *
****
Password must be 8 characters or more
Provisioning Templates
Resolve
Display the templates that will be used to provision this host
Cancel
Submit

Puppet

- Automatic deployment and configuration of servers
- Open source
- Modules written in a proper declarative language
  - Many modules developed by the community are available
- Puppet master
  - Knows the desired configuration of nodes
  - compares desired with actual configuration
- Puppet agent
  - Sends to master the actul configuration
  - Executes proper actions to apply desired configuration

Puppet classes in Foreman





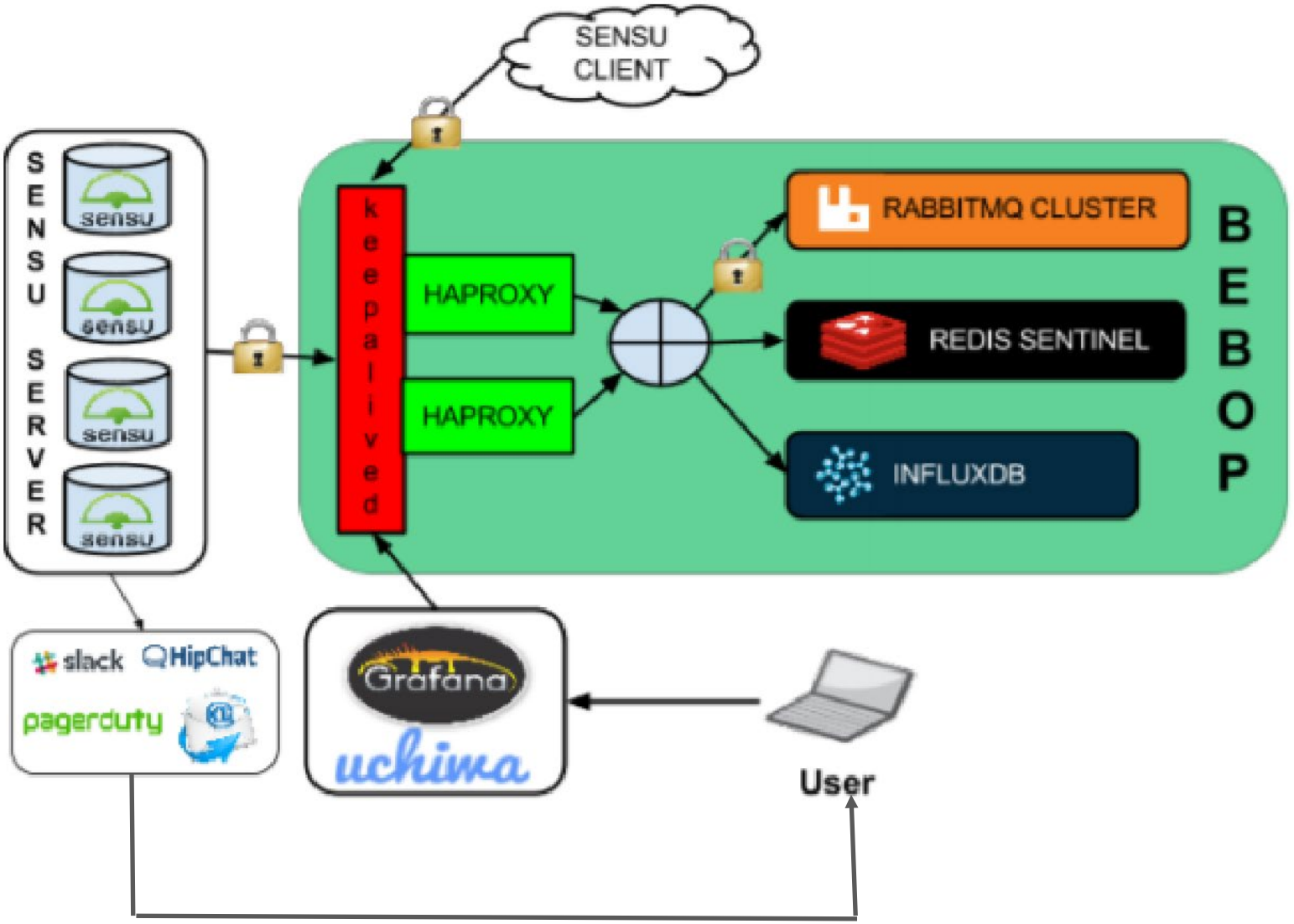
Monitoring and alarming

- Monitoring
  - Control on operation of resources and services
  - Metrics measured on hosts
  - Data analyzed via a web interface or reporting tool
- Alarming
  - System able to notify administrators in case of troubles
    - Service crash, passing of defined thresholds (i.e. for disk space usage, CPU, memory)
  - Event handling
    - Automatic Service Restart

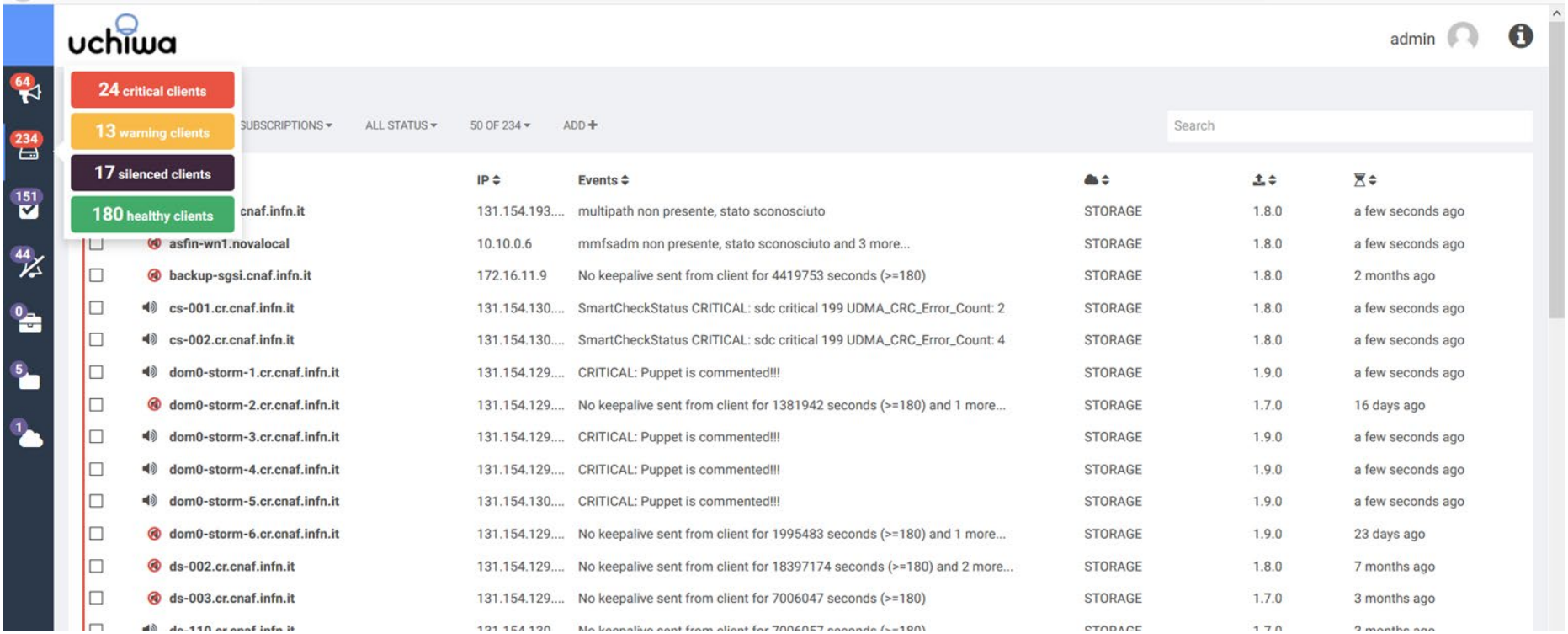
Monitoring at CNAF

- Sensu
  - For bith monitoring and alarming services
  - Retrives metrics and runs checks on hosts
- RabbitMQ
  - Manages queues of messages among Sensu server and clients
- InfluxDB
  - Time series database
  - Archives all hosts metrics
- Uchiwa
  - Web interface for Sensu
  - Used to verify metrics and checks correct execution
- Grafana
  - Web interface to create dashboard with monitoring data

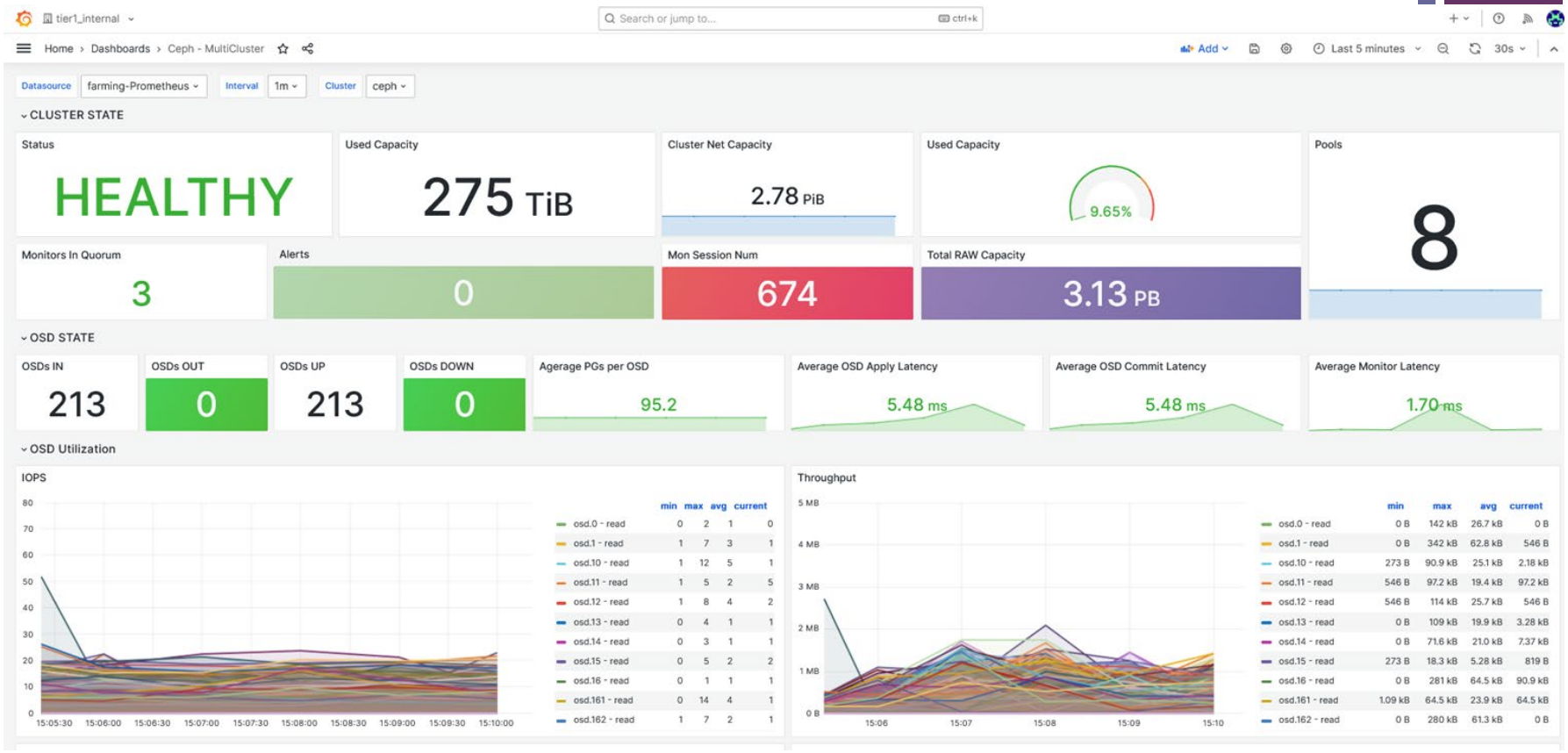
Monitoring at CNAF



Monitoring at CNAF



Monitoring at CNAF

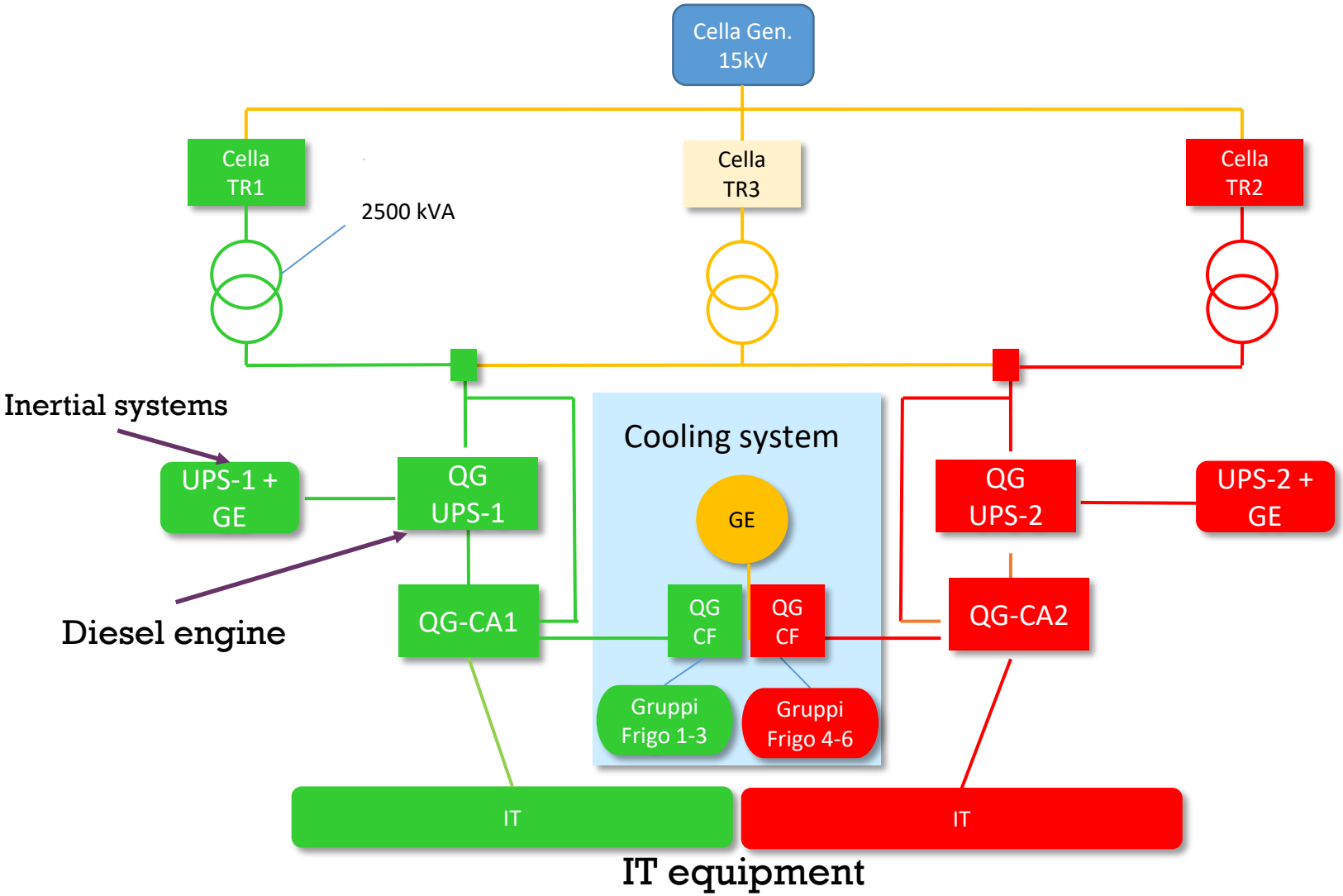


Power and Cooling

Power Infrastructure

- Provides electrical power to all the datacenter systems
- **UPS - uninterruptible power supply**
  - **Prevents failures on electrical cuts**
  - Cab be:
    - Battery based
    - Inertial system generators
    - Engine generators
    - A complex combination of the above

The CNAF power distribution system



Cooling infrastructure

- Free cooling
- Force air flow cooling
- Liquid Submersion
- Liquid Cooling
- Heat Pipes
- …and many others…..

https://en.wikipedia.org/wiki/Computer_cooling

Datacenter PUE

- **Power usage effectiveness (PUE)** is a ratio that describes how efficiently a computer data center uses energy; specifically, how much energy is used by the computing equipment (in contrast to cooling and other overhead)
- Ideally equals to 1.0
- PUE is the inverse of data center infrastructure efficiency (DCIE)

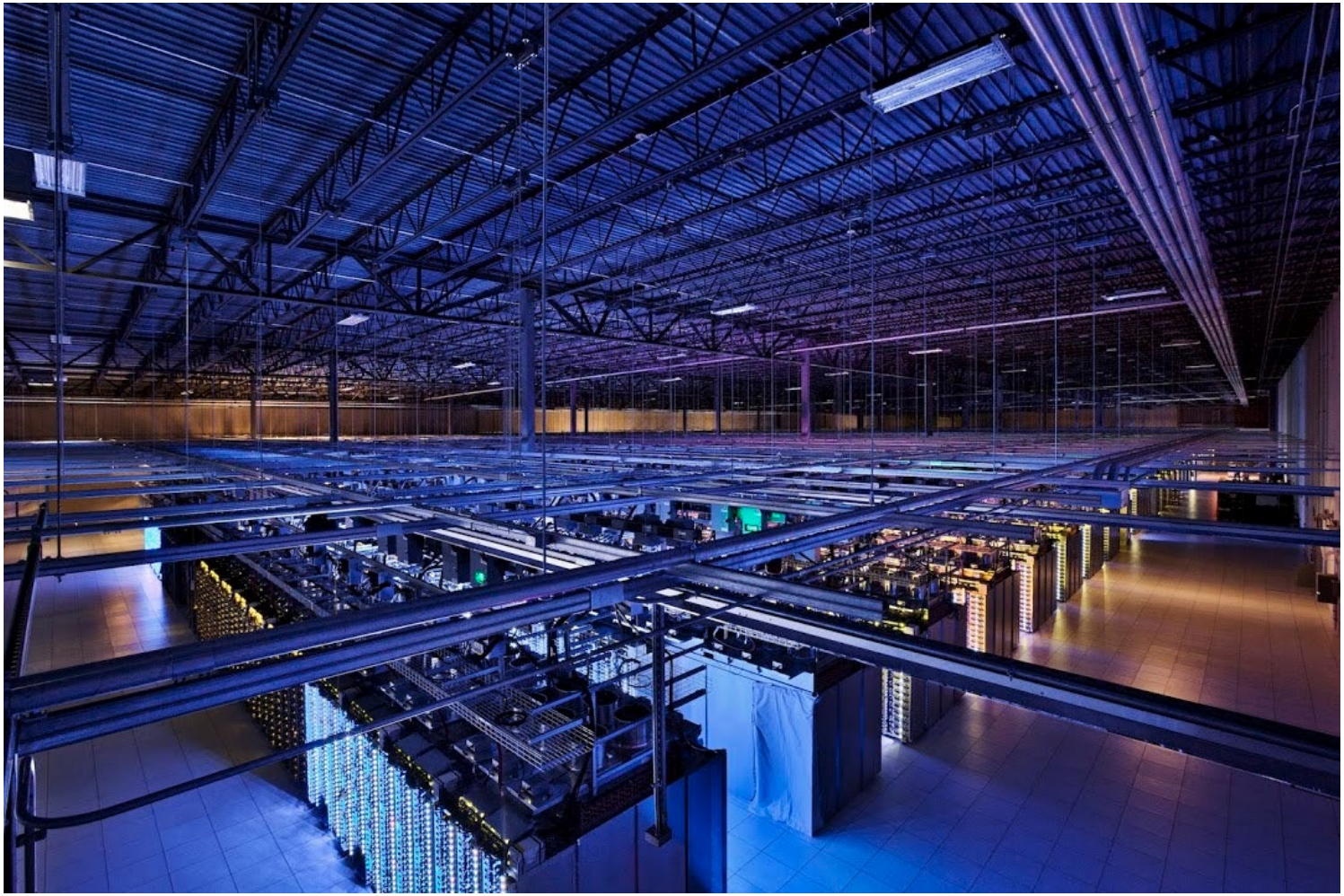
$$\mathrm{PUE} = \frac{\text{Total Facility Energy}}{\text{IT Equipment Energy}} = 1 + \frac{\text{Non IT Facility Energy}}{\text{IT Equipment Energy}}$$

PUE

- In October 2008, Google's Data center was noted to have a ratio of 1.21 PUE across all 6 of its centers
- French hosting company OVH has managed to attain a PUE ratio of 1.09 in its data centers
- In October 2015, Allied Control has a claimed PUE ratio of 1.02 through the use of 3M Novec 7100 fluid
- As of the end of Q2 2015, Facebook's Prineville data center had a PUE of 1.078 and its Forest City data center had a PUE of 1.082.
- In January 2016, the Green IT Cube in Darmstadt was dedicated with a 1.07 PUE. It uses cold water cooling through the rack doors

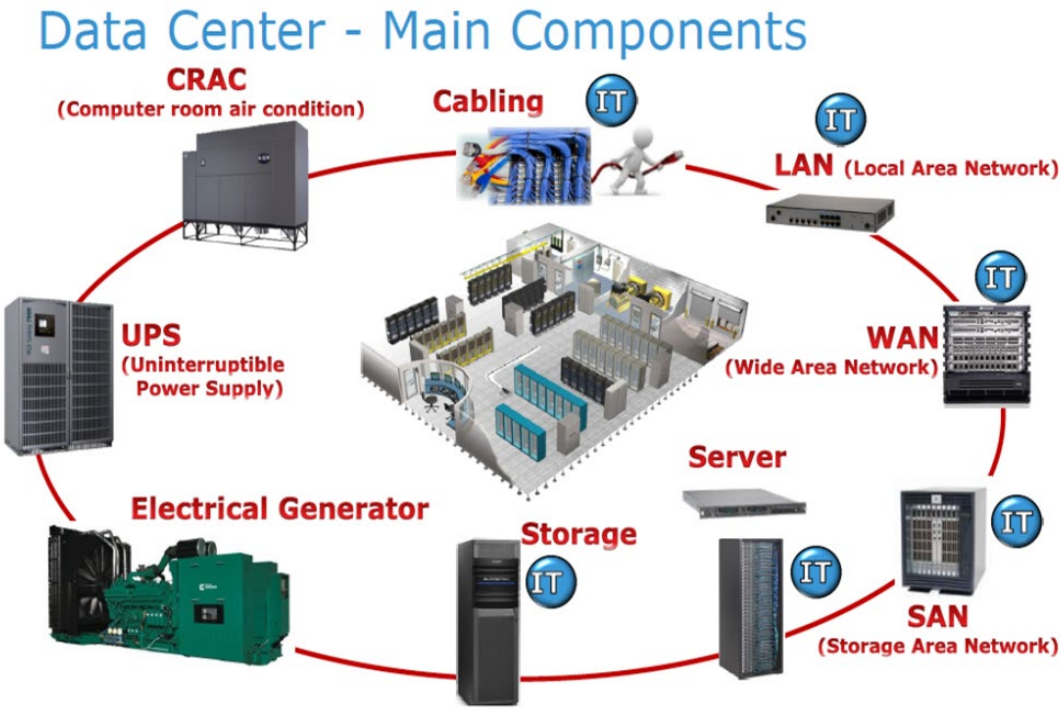
https://en.wikipedia.org/wiki/Power_usage_effectiveness

Data Center Summary



Datacenter logical components

- CPU Farm
- High speed storage - SAN
- Archive storage - TAN
- Networking facilities
- Power and Cooling infrastructures



Datacenter logical components

- CPU Farm
- High speed storage - SAN
- Archive storage - TAN
- Networking facilities
- Power and Cooling infrastructures

Evolving towards an (hyper)converged architecture in the same building blocks

Take away message on QoS – Quality of Service

- In its main components a data center is built to provide different quality of services to the users
  - Storage
    - Bandwidth
    - IOPS
    - Access latency (disks vs tape vs SSD)
    - Data redundancy and replication
  - Computing
    - CPU only and HTC resources
    - GPUs and HPC resources
    - Waiting times
  - Networking
    - Bandwidth
    - Networking
    - Redundancy

In the end, what you pay is the QoS of service connected to the resource, not the resource itself

When creating or buying a computing infrastructure (physical or virtual) it is important to analyze carefully the users/applications requirements and match them with the QoS offered by the resource providers

References

- https://en.wikipedia.org/wiki/IOPS
- https://en.wikipedia.org/wiki/Template_talk:Bit_and_byte_prefixes
- https://en.wikipedia.org/wiki/Standard_RAID_levels
- https://www.cavium.com/Documents/TechnologyBriefs/Adapters/Tech_Brief_Introduction_to_Ethernet_Latency.pdf
- https://en.wikipedia.org/wiki/Storage_area_network
- https://www.ibm.com/support/knowledgecenter/en/SSETD4_9.1.3/lsf_admin/fairshare_about_lsf.html
- https://www.ibm.com/support/knowledgecenter/en/SSETD4_9.1.3/lsf_admin/backfill.html
- https://community.fs.com/blog/do-you-know-the-differences-between-hubs-switches-and-routers.html
  https://en.wikipedia.org/wiki/Computer_cooling
- https://en.wikipedia.org/wiki/Power_usage_effectivenes
- https://puppet.com/
- https://www.theforeman.org/
- https://grafana.com/