Revision 2

C2115 Practical introduction to supercomputing

Lesson 3

Petr Kulhánek

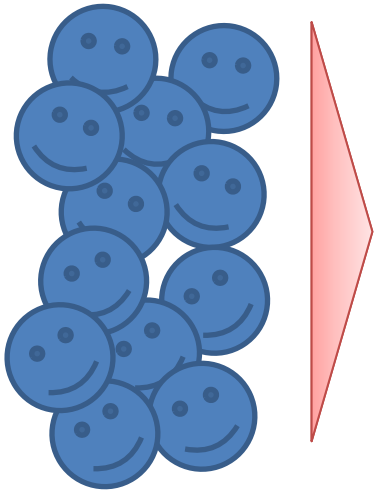
kulhanek@chemi.muni.cz

National Centre for Biomolecular Research, Faculty of Science
Masaryk University, Kamenice 5, CZ-62500 Brno

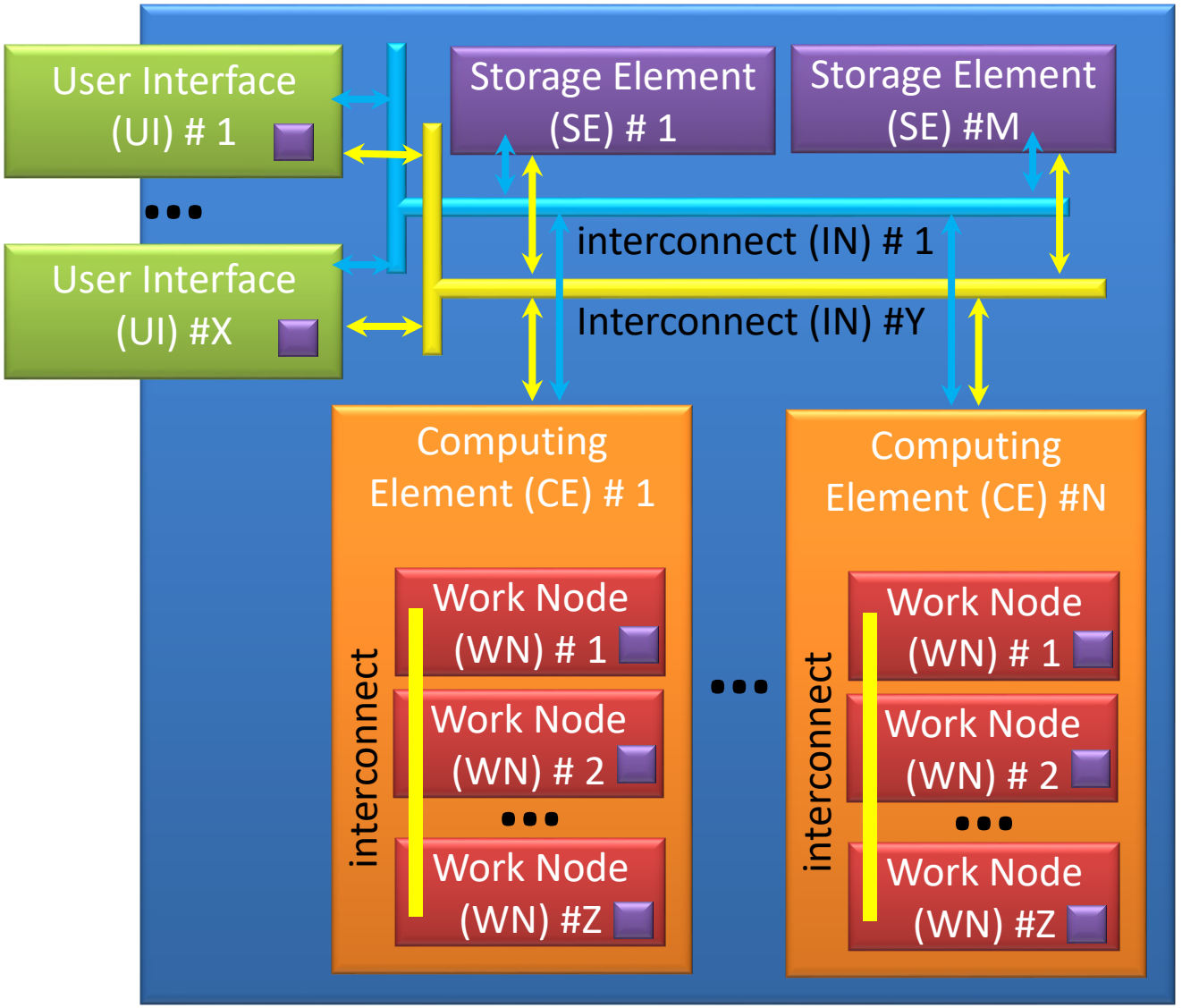
Content

- Architecture of clusters and (super)computers
  - Front nodes
  - Computational nodes and elements
  - Data storage
  - Network infrastructure

Key parts

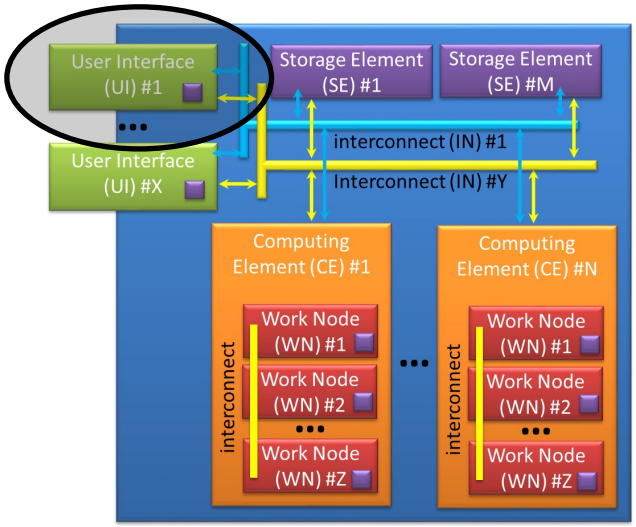


users



Frontend (UI)

Frontend (**front-end node**, **user interface**) is a computer dedicated for direct interaction with the user. The user can use it to prepare input data for tasks, **submit jobs into the batch system**, for management of jobs and manipulating job results (visualization).



The front node, unless explicitly allowed, **should not be used to run CPU and memory intensive tasks**, potential pre-processing or post-processing of job data must be entered as separate jobs into the batch system.

In small compute clusters, the front node is often also a compute node.

A cluster or supercomputer usually handles the tasks of many users, which requires only remote access to the front node. (Additionally, the front node is typically physically present in the server room, where there is relatively a lot of noise.)

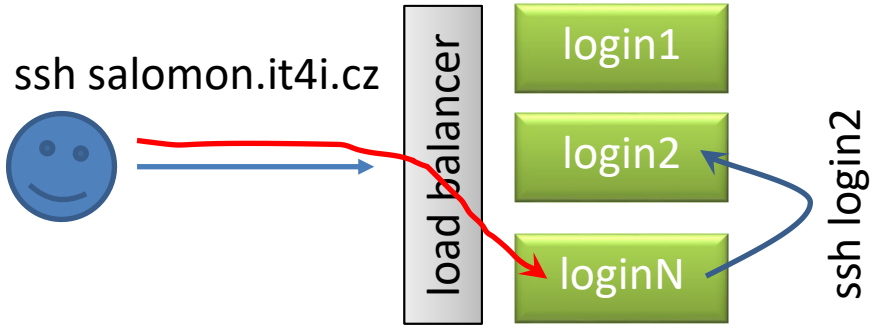
Remote access - front node (UI)

Front node (frontend) is a computer that interacts with the supercomputer.

- UI usually only offers a command line (**CLI** - command line interface) using a secure ssh connection (secure shell).
- If a graphical interface is available (**GUI** - graphical interface) it is more appropriate to use a remote desktop (VNC - virtual network connection) than export of X11 display.

direct access

indirect access

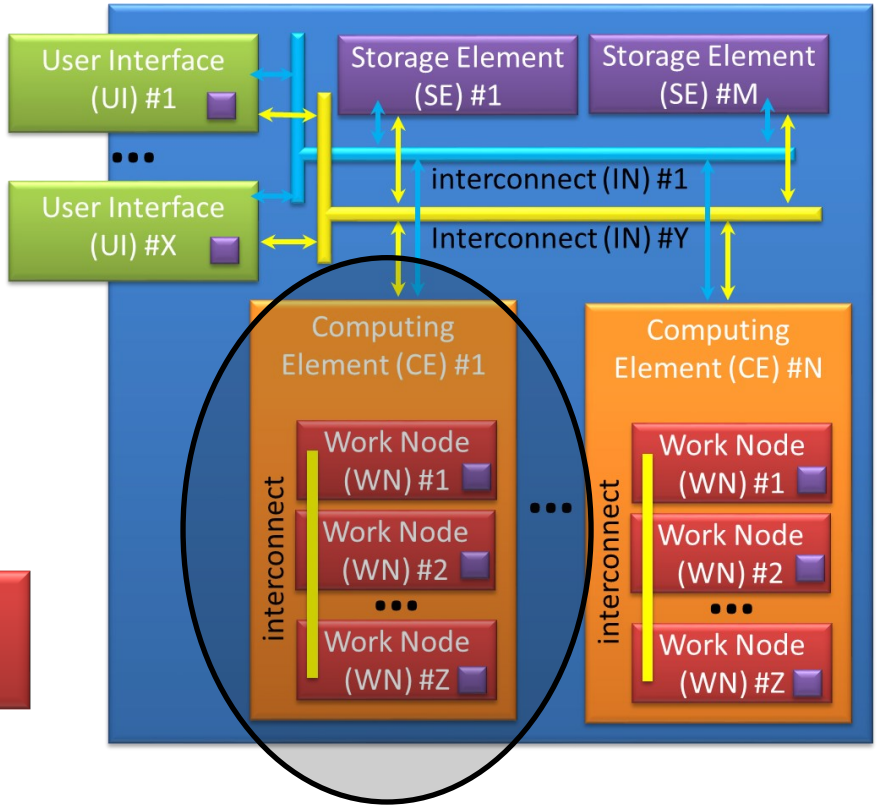


Computationally intensive tasks are NOT run on the UI!!!

Computational element (CE)

Computational element (**computational element** - **CE**), very often also referred to as **cluster**, is a grouping of computational nodes most often with the same architecture (homogeneous cluster). These nodes are usually connected by a very fast local network (Ethernet 1 Gbs, Infiniband, or a proprietary solution).

The computing node or the computer node are synonyms for the work node (**WN - worker node**).



Direct use of CE and WN is prohibited.
Jobs must be submitted to the batch system.

Computing node (WN)

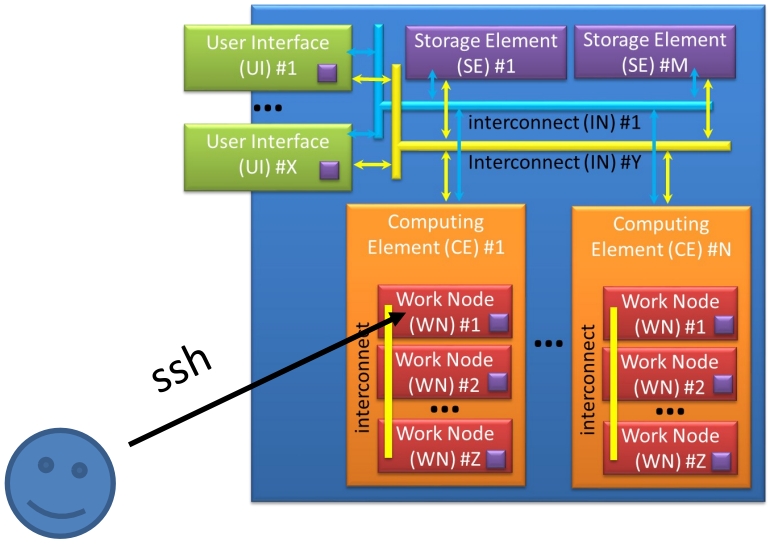
Computing node (**worker node - WN, computational node**) is a unit that acts as a stand-alone computer that is dedicated to solving user jobs. A node can handle several jobs at the same time. One job can use multiple compute nodes. However, the number of jobs should not exceed the computing resources (CPU, RAM, HDD) that this node provides. The efficient use of computing resources is taken care of by the OS (operating system) in conjunction with **batch system**.

The task can also run on nodes that are in different CEs. However, this is only suitable for a special type of jobs.

Possible organization of tasks on WN and CE:

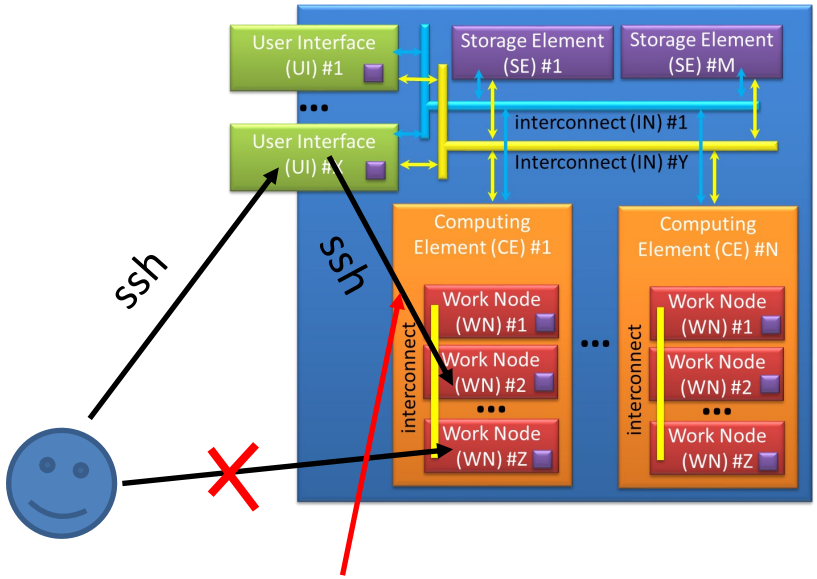
Access to WN - monitoring only

direct access



MetaCentrum,
all NCBR/CEITEC MU clusters

indirect access



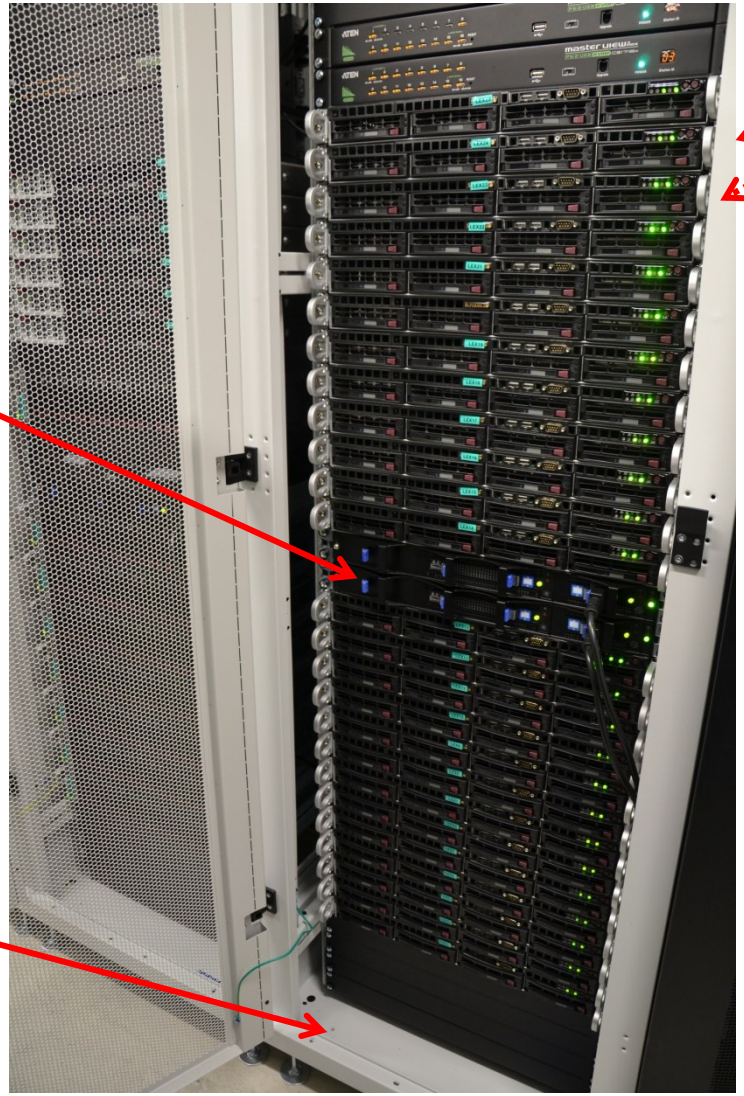
IT4I

Access is restricted to nodes where the user is running jobs.

This approach can only be used to monitor running tasks.
It MUST NOT be used to run jobs on their own

Computational element / node

LEX cluster (computational element)

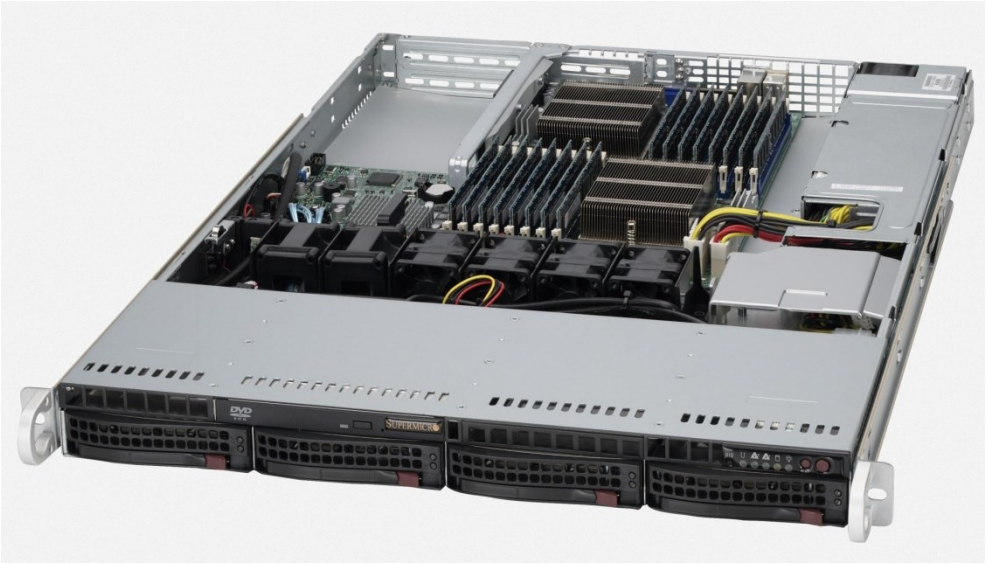


computing nodes

Infiniband Controler

rack

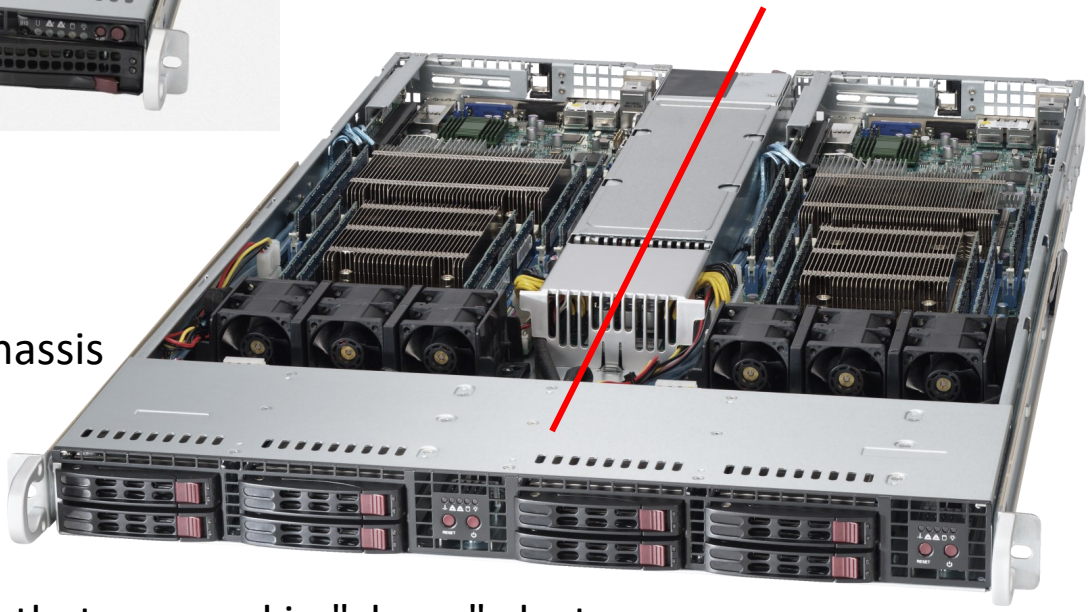
Computing node



1U height = 1.75 inches (44.45 mm)

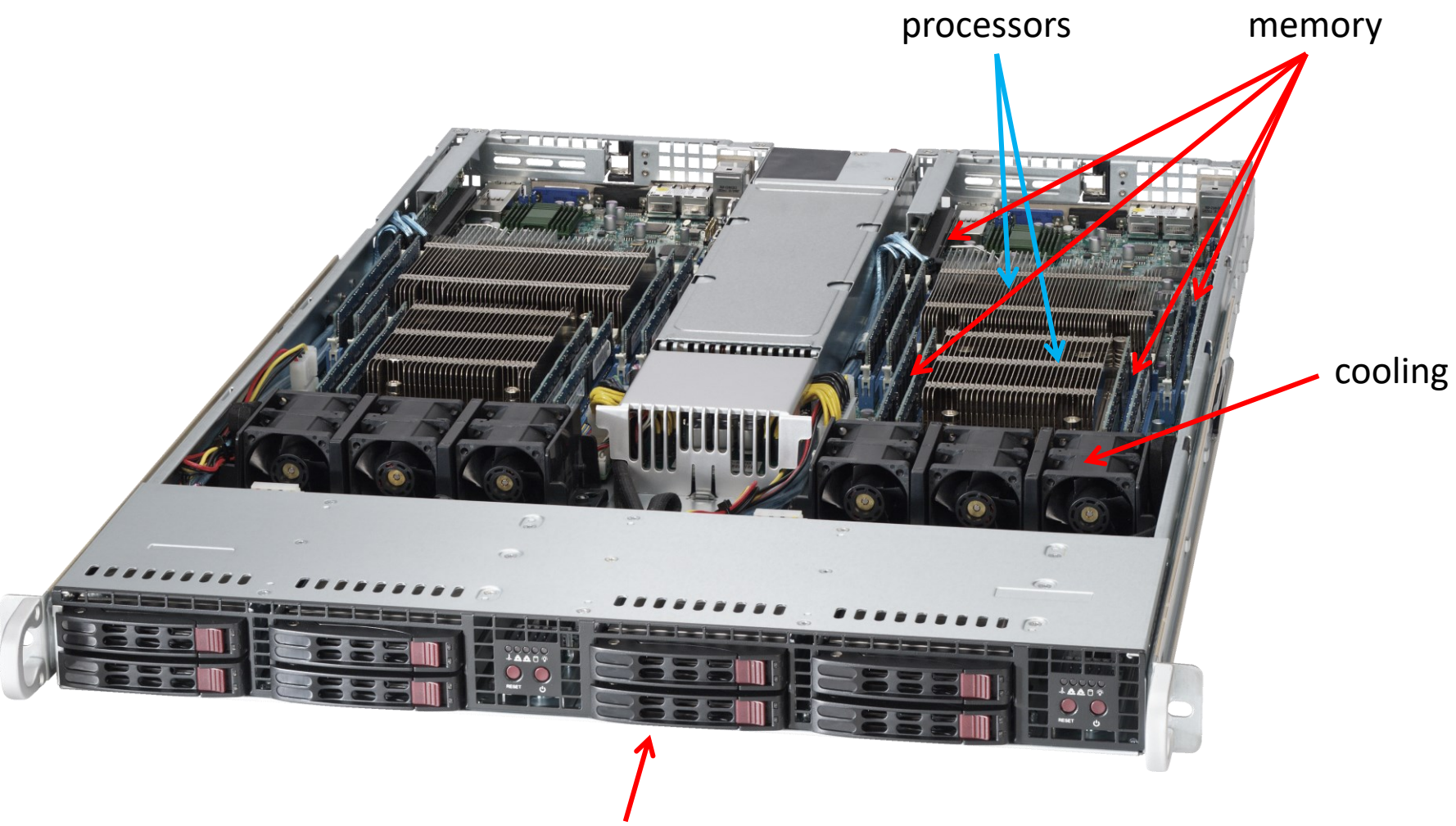
one computer in one chassis

twin - two computers in one chassis



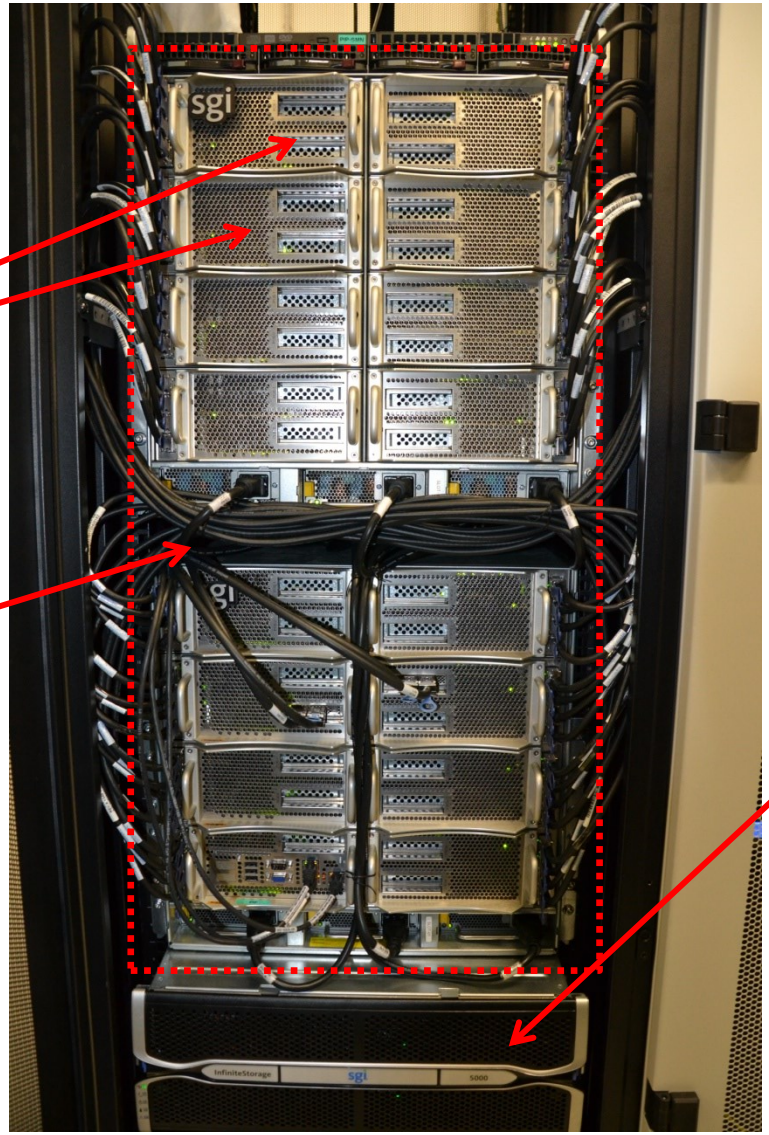
examples of typical compute nodes that are used in "cheap" clusters
- in supercomputers, a proprietary solution is mostly used

Computing node



Computing node

SGI UV2000
pip



blades

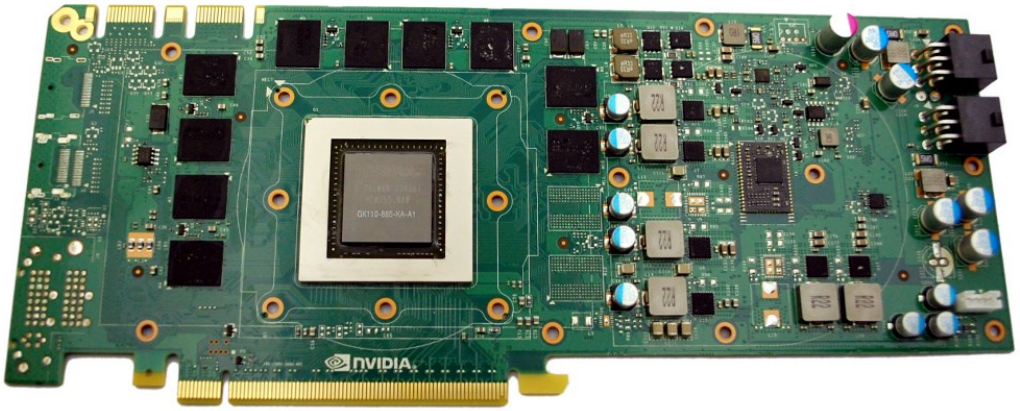
bus

one computing node
192 CPU cores, 4 TB memory

disk array - local data storage

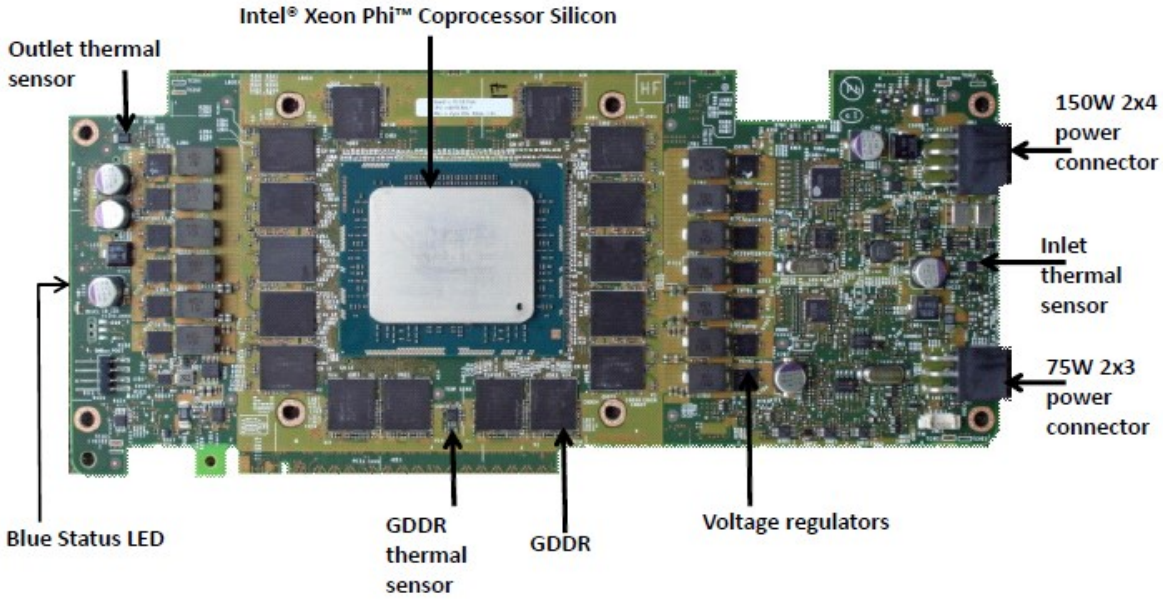
Computing node - accelerators

NVidia Tesla K20 (GPGPU)



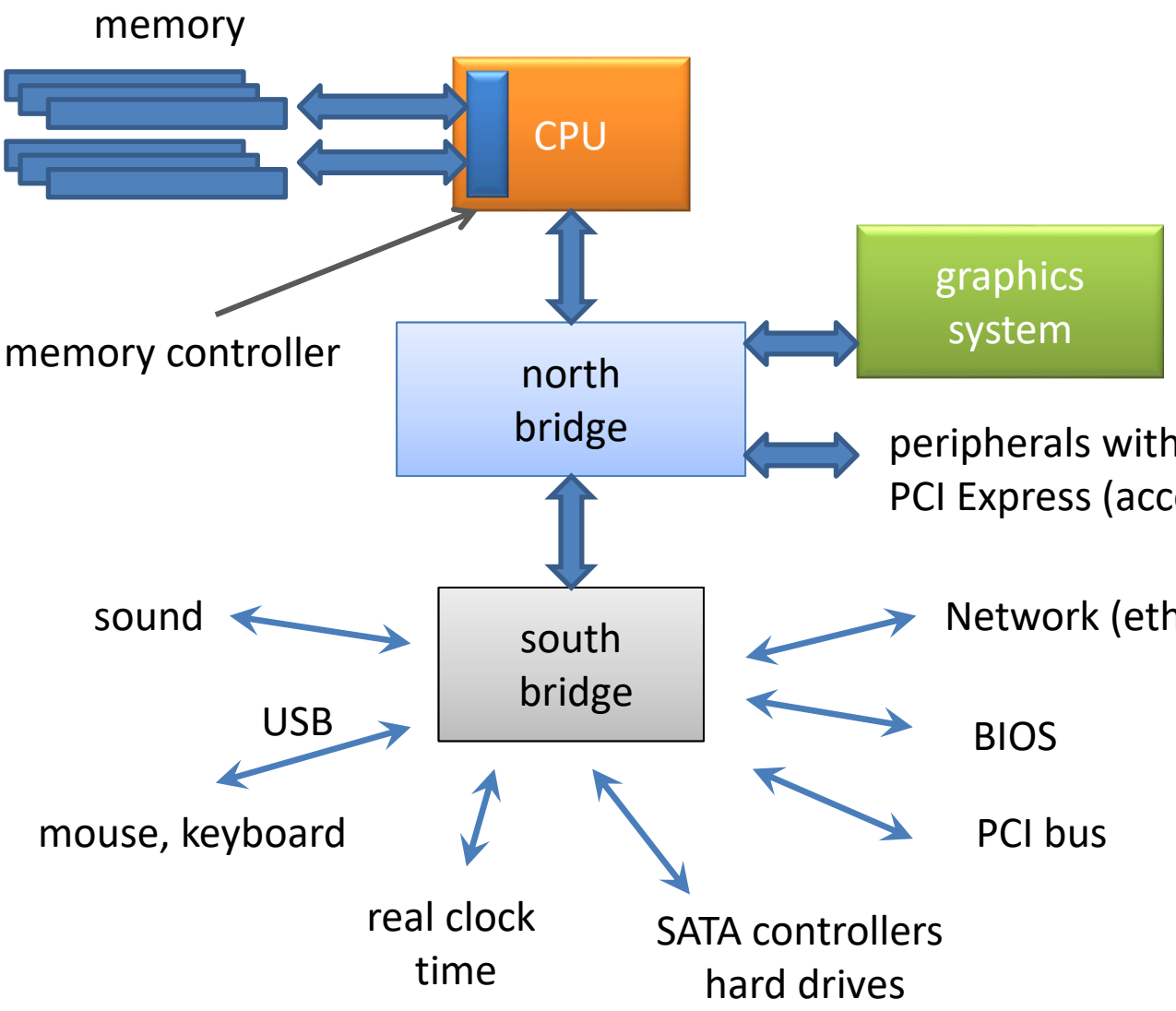
Practical:
MD simulation on GPU

Intel Xeon Phi (BALL)



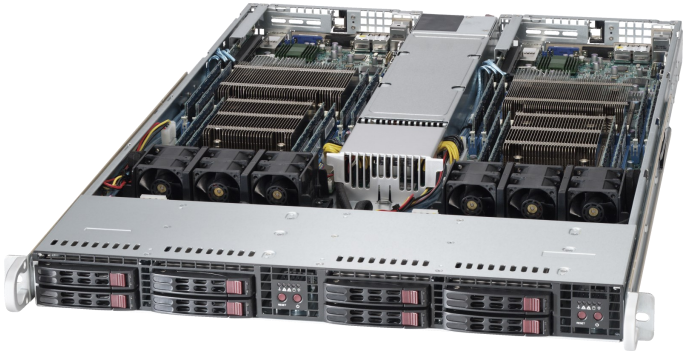
The computing power of the accelerators may exceed the performance of the installed CPUs on the computing node.

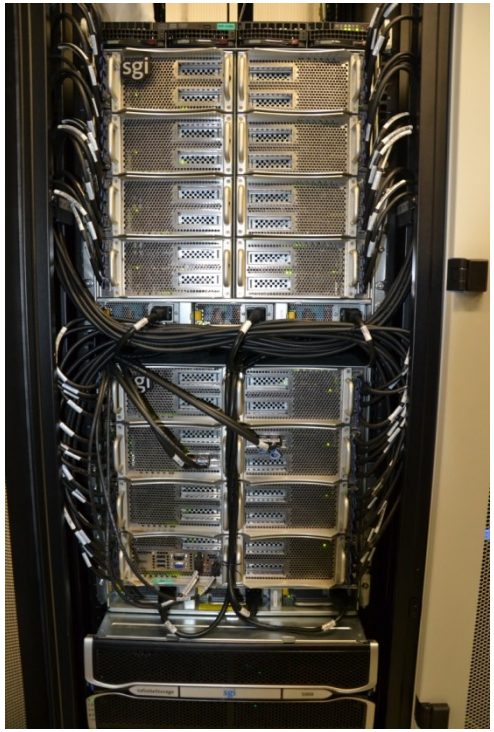
Typical computer scheme



Practical: the influence of architecture on computing power

Multiprocessor nodes





Nowadays, compute nodes contain **multiple physical processors** (minimum two), each containing **multiple computing CPU cores**. The RAM is then usually accessible at different speeds (**NUMA Non-Uniform Memory Architecture**).

The reason for this arrangement is **increasing computing power**, which, however, brings increased demands on the preparation and execution of computational tasks.

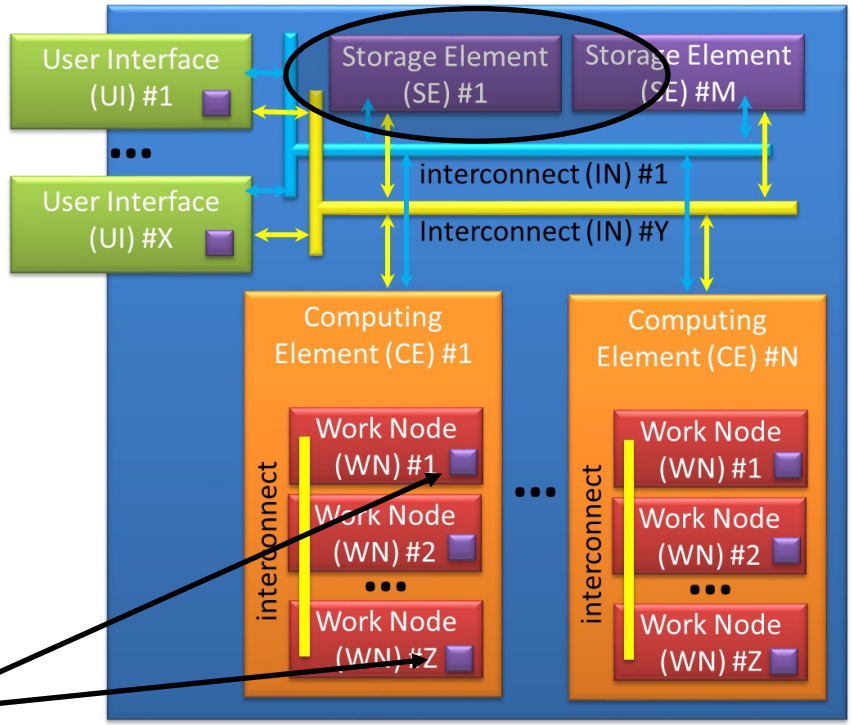
Practical:
parallelization of tasks,
running of parallel tasks,
limitations

Data storage (SE) - partitioning

Types of data storage (SE - storage element) and their use:

- **local data storage** - temporary job data
- **(remote) data storage (disk array)** - live data of jobs or solved projects
- **hierarchical data storage** - completed projects and backups

Practical:
data storage
MetaCentrum



Local data storage

- **Disk array** connected locally to the compute node.
- **HDD - hard disk** (**Hard Disk Drive**) is a device used in computers to permanently store large amounts of data by magnetic induction.
- **SSD - Solid-State Drive**, in information technology, is a type of data medium which, unlike magnetic hard disks, does not contain moving mechanical parts and has a much lower power consumption.

Local temporary storage (scratch directories) are intended for currently running tasks on the compute node.

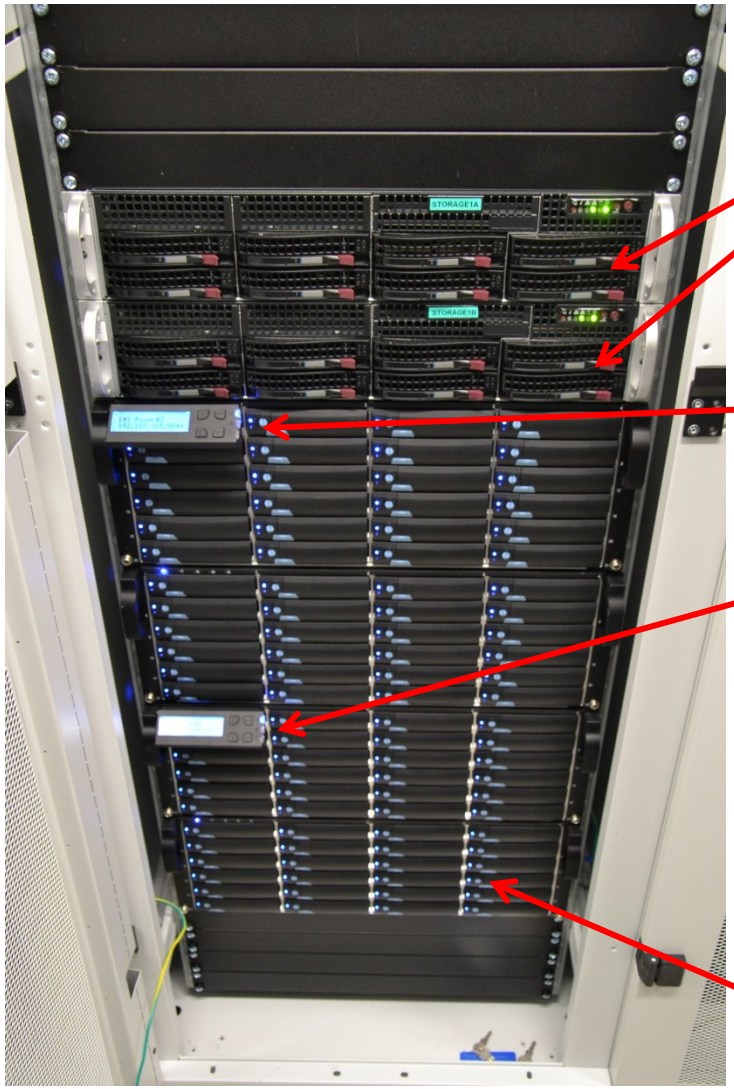
These directories MUST NOT* be used for long-term data storage.

*) Of course you can, but then don't be surprised that one day you won't find them, because the administrator, or another intelligent tool, has cleaned the storage

wikipedia.org

Disk array

brno9-ceitec 269 TiB



file servers accessing disk array data remotely via NFS (**Network File System**) protocol

RAID6 disk array

RAID6 disk array

RAID 0

Disk arrays are suitable for currently solved projects.

large amount of HDDs

Disk array - data protection

Arrays contain a large number of HDDs, which are mechanical components that are prone to failure. To reduce data corruption, data is most often protected using RAID technology.

RAID (**Redundant Array of Inexpensive/ Independent Disks**), in informatics, is a method of securing data against hard disk failure. Security is realized by specific storage of data on multiple independent disks, where the stored data is retained even if one of them fails. The level of security varies depending on the type of RAID selected, which is indicated by numbers (most often RAID 0, RAID 1, RAID 5, or more recently RAID 6).

When part of the disk array is damaged, the array runs in **degraded mode**, where another failure would be irreparable. Therefore, the so-called spare discs are used immediately as a replacement for damaged ones. Speed of ata access may be reduced during **rebuilding disk array** (new data parity calculation).

wikipedia.org

Hierarchical data storage

Hierarchical storage management (**HSM**) is a data storage technique, which automatically moves data between high-cost and low-cost storage media. While it would be ideal to have all data available on high-speed devices all the time, this is prohibitively expensive for many organizations. Instead, HSM systems store the bulk of the enterprise's data on slower devices, and then copy data to faster disk drives when needed. The HSM system monitors the way data is used and makes best guesses as to which data can safely be moved to slower devices and which data should stay on the fast devices.

brno10-ceitec-hsm
(tape robot)

HSM repositories are suitable for archiving and backing up data.

wikipedia.org

Units

Multiples of bytes V · T · E

Decimal			Binary			
Value	Metric		Value	IEC		JEDEC
1000	kB	kilobyte	1024	KiB	kibibyte	KB kilobyte
1000^2	MB	megabyte	1024^2	MiB	mebibyte	MB megabyte
1000^3	GB	gigabyte	1024^3	GiB	gibibyte	GB gigabyte
1000^4	TB	terabyte	1024^4	TiB	**tebibyte**	–
1000^5	PB	petabyte	1024^5	PiB	pebibyte	–
1000^6	EB	exabyte	1024^6	EiB	exbibyte	–
1000^7	ZB	zettabyte	1024^7	ZiB	zebibyte	–
1000^8	YB	yottabyte	1024^8	YiB	yobibyte	–

Orders of magnitude of data

original marking

wikipedia.org

Network infrastructure

Ethernet is the name of a set of technologies for local area networks (LANs) that use cables with twisted double line, optical cables for communication at transmission speeds from 10 Mbit/s to 100 Gbit/s.

InfiniBand (abbreviated IB), a computer-networking communications standard used in high-performance computing, features **very high throughput and very low latency**. It is used for data interconnect both among and within computers. InfiniBand is also utilized as either a direct, or switched interconnect between servers and storage systems, as well as an interconnect between storage systems.

Infiniband is suitable for data-intensive parallel tasks that use multiple compute nodes.

wikipedia.org

Batch system

Batch processing is the execution of a series of programs (so-called batches) on a computer without the participation of the user. Batches are prepared in advance so that they can be processed and handed over without the user's participation. All input data is prepared in advance in files (scripts) or entered using parameters on the command line. Batch processing is the opposite of interactive processing, where the user provides the required inputs only while the program is running.

Advantages of batch processing

- sharing computer resources between many users and programs
- postponing batch processing until the computer is less busy
- Eliminate delays caused by waiting for user input
- maximizing computer utilization improves investment utilization (especially for more expensive computers)

Our local clusters, MetaCentrum: PBSPro
IT4I: PBSPro

PBSPro is derived from OpenPBS.

Practical: PBSPro

wikipedia.org

Exercises 1

1. What is the name of your workstation (computer) on the WOLF cluster?
2. What is the role of this computer within the WOLF cluster?
3. Find out the names of the front nodes of the MetaCentrum virtual organization from the documentation.
4. Verify that you can log ib to one of the front nodes in MetaCentrum.
5. How many hard disks can fail in a disk group that is protected by RAID6?
6. Can RAID0 be used for data protection?
7. What is the combination of RAID6 and RAID0?
8. What type of accelerator is used in a supercomputer salomon (IT4I)?