cesnet

metacentrum

e-INFRA CZ

MetaCentrum NGI - Best Practices

Jiří Vorel

MetaCentrum User Support

April 30th, 2024
Prague

What MetaCentrum is

- **MetaCentrum is**
  - … The National Grid Infrastructure (NGI).
  - … the activity of the CESNET association.
- **MetaCentrum is available for**
  - … employees and students from Czech universities, the Czech Academy of Science, non-commercial research facilities, etc.
  - … industry and foreign partners (only for non-profit and open research).
- **MetaCentrum provides**
  - … **compute resources** (CPU, GPU), **application tools** (commercial and free/open source) and **data storage, GUI environment** (OnDemand, Matlab, Ansys, RStudio), **container solution** (Singularity/Apptainer), etc.

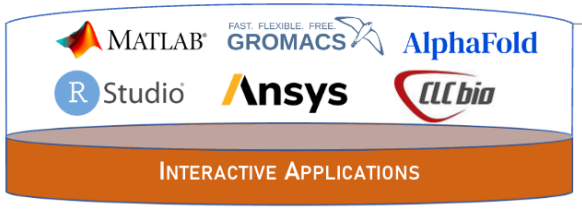
https://metacentrum.cz

https://metavo.metacentrum.cz

https://metavo.metacentrum.cz/en/application/index.html

cesnet
metacentrum

What MetaCentrum offers

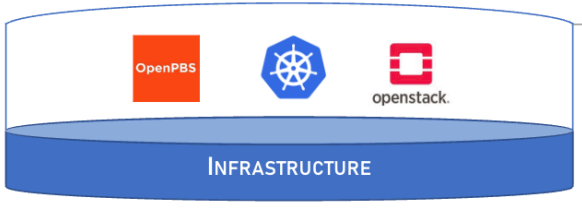


INTERACTIVE APPLICATIONS

- Matlab
- Gromacs
- AlphaFold
- CLCbio
- RStudio
- Ansys ...

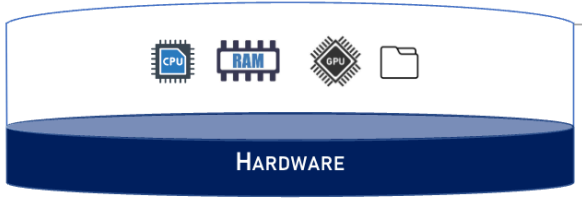
PLATFORMS

- Jupyter Notebooks (OpenPBS, K8s)
- Galaxy (OpenPBS)
- OnDemand (OpenPBS)



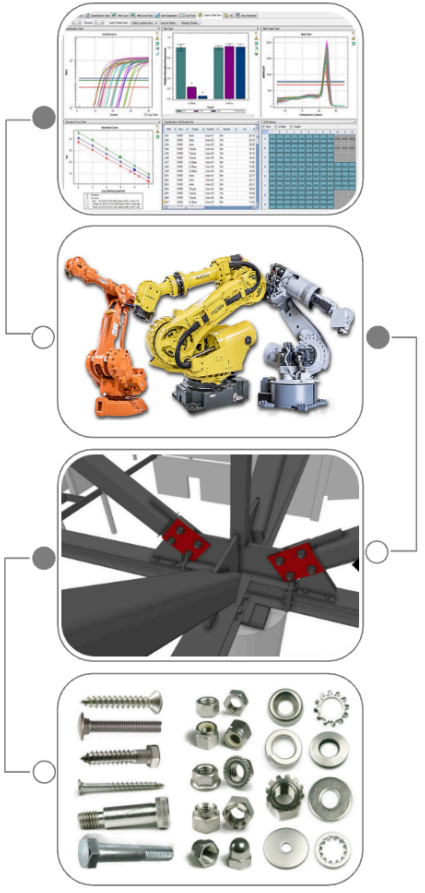
INFRASTRUCTURE

- OpenPBS for HPC
- Kubernetes (K8s) / Rancher
- OpenStack Cloud / Sensitive Cloud



HARDWARE

- 41,288 CPU cores
- 426 GPU cards with up to 80 GB RAM
- up to 10 TB RAM per node
- 33 PB of storage



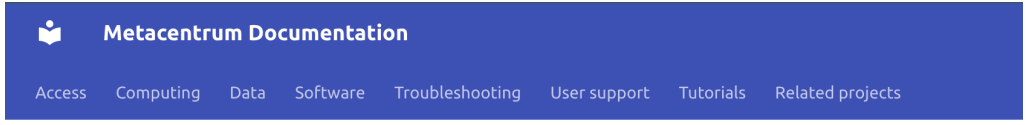
cesnet
metacentrum

Not so new "new documentation"

- The original documentation hosted on wiki.metacentrum.cz has been marked as deprecated and is not further maintained.

> **Metacentrum wiki is deprecated after March 2023**
> Dear users, due to integration of Metacentrum into https://www.e-infra.cz/en (e-INFRA CZ service), the documentation for users will change format and site.
> The current wiki pages won't be updated after end of March 2023. They will, however, be kept for a few months for backwards reference.
> The new documentation resides at https://docs.metacentrum.cz.

- We switched on the new documentation web.



Welcome

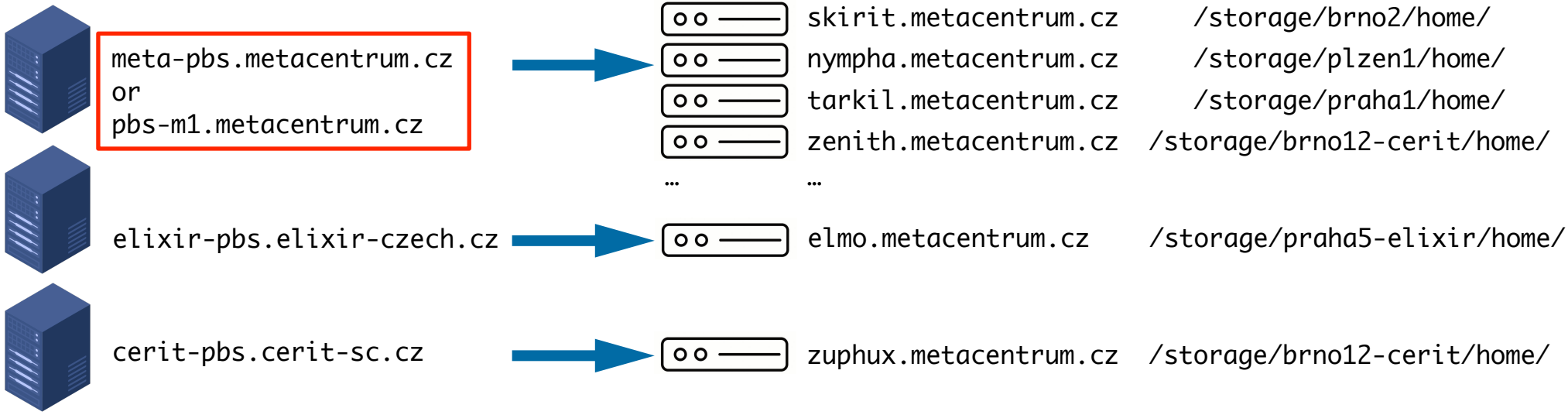
This is the documentation for users of MetaCentrum grid computing service.

https://docs.metacentrum.cz/

meta@cesnet.cz

Frontend servers

- Gateway to the entire infrastructure, **accessible via SSH protocol with a password or valid Kerberos ticket** (generated on the personal computer).
- **Do not run long and demanding calculations directly on frontends.**
- **Frontend servers can have different home directories**.



Debian 12 and new scheduler

- We introduced a new scheduler, **OpenPBS (pbs-m1.metacentrum.cz)**, which will replace the current scheduler **PBSPro (meta-pbs.metacentrum.cz)**.

https://docs.metacentrum.cz/tutorials/debian-12/

```
default@meta-pbs.metacentrum.cz    -->  default@pbs-m1.metacentrum.cz
large_mem@meta-pbs.metacentrum.cz  -->  large_mem@pbs-m1.metacentrum.cz
gpu@meta-pbs.metacentrum.cz        -->  gpu@pbs-m1.metacentrum.cz
```

- Compute nodes available in the **OpenPBS were also upgraded on the Debian 12**.
- **Debian 12 frontends** (e.g. zenith, nympha, tarkil,...) **submit jobs to OpenPBS by default**. On frontends with Debian 11, **users need to load module** openpbs before the submission.

```
(BOOKWORM)vorel@zenith:/storage$ lsb_release -a
No LSB modules are available.
Distributor ID: Debian
Description:    Debian GNU/Linux 12 (bookworm)
Release:        12
Codename:       bookworm
```

https://docs.metacentrum.cz/computing/frontends/

Debian 12 and new scheduler

- We keep OS Debian up-to-date on our nodes.
- We are **upgrading from Debian 11 (BULLSEYE) to Debian 12 (BOOKWORM)**.
- However, some libraries may be missing in the new system...

```
ImportError: libcrypto.so.1.1: cannot open shared object file: No such
file or directory
```

- Therefore, we provide universal modules with these missing libraries.

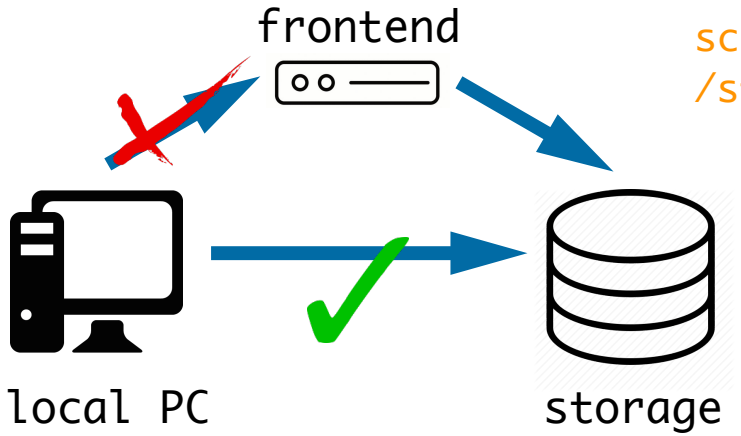
```
(BULLSEYE)vorel@skirit:~$ module ava debian*
-------------------------------- /packages/run/modules-5/debian11avx512 --
debian7/  debian8/  debian9/  debian10/  debian11/

Key:
modulepath  directory/
(BULLSEYE)vorel@skirit:~$ ls /software/debian-compat/debian11/lib
libatlas.so.3  libblas.so.3  libcrypto.so.1.1  libtiff.so.5  libwebp.so.6
```

- Users can still use other (older) modules...

https://docs.metacentrum.cz/tutorials/debian-12/

Data transfer to MetaCentrum

- Do not use frontends. Copy data (large volumes mainly) directly on the storage server. Use compressed files (.tar, .zip, .gz, etc.).
- **Very high numbers of very small files are problematic effective manipulation**.
- FTP client for Windows users (WinSCP, FileZilla, CyberDuck).




```
scp large_data.gz vorel@skirit.metacentrum.cz:\
/storage/praha5-elixir/home/vorel
```

```
scp large_data.gz \
vorel@storage-praha5-elixir.metacentrum.cz:~

scp vorel@storage-praha5-elixir.metacentrum.cz:~/
large_data.gz .
```

https://docs.metacentrum.cz/data/large-data/#large-data-handling

https://docs.metacentrum.cz/data/data-within/#moderate-data-handling

Data transfer to MetaCentrum

- Data is stored on a few independent storages.
- **All storages are accessible through all frontends and compute nodes.**
- Storages have quotas for the total volume of data and the number of files.
- MetaCentrum storage capacities are dedicated mainly to data in active usage.

Server	Directory	Backup class	Note
storage-brno11-elixir.metacentrum.cz	/storage/brno11-elixir/	2	dedicated to ELIXIR-CZ
storage-brno12-cerit.metacentrum.cz	/storage/brno12-cerit/	2	
storage-plzen1.metacentrum.cz	/storage/plzen1/	2	

https://docs.metacentrum.cz/computing/mount-storages/

Transfer of a large amount of data in job

- Into scratch (defined as `$SCRATCHDIR`).

Suitable for small data volumes (up to a few GBs).

```
cd $SCRATCHDIR
cp /storage/brno12-cerit/home/vorel/test_data.tar.gz .
scp storage-brno12-cerit.metacentrum.cz:~/test_data.tar.gz .
```

Preferred way, faster. Copy data directly from the storage server.

- From scratch.

```
# be in $SCRATCHDIR
mv result.tar.gz /storage/brno12-cerit/home/vorel
scp result.tar.gz storage-brno12-cerit.metacentrum.cz:~
clean_scratch
```

Useful utility that removes all data in `$SCRATCHDIR`.

https://docs.metacentrum.cz/computing/trap-command/ https://docs.metacentrum.cz/data/large-data/#move-data-between-storages

Local SW installations

- Users can install the software independently (in their home directories).
- No restrictions; do not violate the license terms and conditions or/and our rules.
- Users do not have sudo rights and can not right outside of the home directory.

```
(BULLSEYE)vorel@skirit:~$ apt-get install package_name
E: Could not open lock file /var/lib/dpkg/lock-frontend - open (13: Permission denied)
E: Unable to acquire the dpkg frontend lock (/var/lib/dpkg/lock-frontend), are you root?
```

- **Python** (pip with `--user` option and `$PYTHONUSERBASE`, venv).

```
(BULLSEYE)vorel@skirit:~$ module ava py-pip/
------------------------------------ /packages/run/modules-5/debian11avx512
py-pip/19.3-intel-19.0.4-hudzomi  py-pip/21.3.1-gcc-10.2.1-mjt74tn
```

- **R packages** (with `lib="user/path"`).

```
(BULLSEYE)vorel@skirit:~$ module ava r/
--------------------------- /packages/run/modules-5/debian11avx512 --------------------
r/2.14.0      r/3.1.1       r/3.5.1-gcc                r/4.0.2-intel-19.0.4-5vzfhtq
r/3.0.1       r/3.2.3-intel r/3.6.2-gcc                r/4.1.1-intel-19.0.4-ilb46fy
r/3.0.3       r/3.3.1-intel r/4.0.0-gcc                r/4.1.1-intel-19.0.4-xrup2b3
r/3.1.0       r/3.4.0-gcc   r/4.0.2-aocc-2.2.0-q43q56w r/4.1.3-gcc-10.2.1-6xt26dl
r/3.1.0shlib  r/3.4.3-gcc   r/4.0.2-aocc-2.2.0-zrf6vyw r/4.2.1-intel-19.0.4-d3gtjq7
```

https://docs.metacentrum.cz/software/install-software/

Local SW installations

- **Pre-compiled binaries** can be directly downloaded/copied into `$SCRATCHDIR`.

```
(BULLSEYE)vorel@skirit:~$ qsub -I -l select=1:ncpus=1:mem=5gb:scratch_local=1gb -l walltime=1:00:00
qsub: waiting for job 14986173.meta-pbs.metacentrum.cz to start
qsub: job 14986173.meta-pbs.metacentrum.cz ready

(BULLSEYE)vorel@elmo3-1:~$ cd $SCRATCHDIR
(BULLSEYE)vorel@elmo3-1:/scratch/vorel/job_14986173.meta-pbs.metacentrum.cz$ wget -q https://www.drive5.com/downloads/usearch11.0.667_i86linux32.gz
(BULLSEYE)vorel@elmo3-1:/scratch/vorel/job_14986173.meta-pbs.metacentrum.cz$ gunzip usearch11.0.667_i86linux32.gz
(BULLSEYE)vorel@elmo3-1:/scratch/vorel/job_14986173.meta-pbs.metacentrum.cz$ chmod u+x usearch11.0.667_i86linux32
(BULLSEYE)vorel@elmo3-1:/scratch/vorel/job_14986173.meta-pbs.metacentrum.cz$ ./usearch11.0.667_i86linux32
usearch v11.0.667_i86linux32, 4.0Gb RAM (791Gb total), 112 cores
(C) Copyright 2013-18 Robert C. Edgar, all rights reserved.
https://drive5.com/usearch
```

- **Perl** (`cpanm`) libraries.
- **Mamba/Conda/Miniconda/Micromamba** package managers.

Most preferred way. Use module `mambaforge`. https://anaconda.org/

https://docs.metacentrum.cz/software/install-software/#conda-packages

Local SW installations

- **Mamba/Conda/Miniconda/Micromamba** package managers.

Most preferred way. Use module `mambaforge`. https://anaconda.org/

```
module add mambaforge
# create new Conda environment called segemehl-0.3.4 (with python 3.8)
mamba create --prefix /storage/city/home/user_name/segemehl-0.3.4 python=3.8 -y
# activate the environment
mamba activate /storage/city/home/user_name/segemehl-0.3.4
# install the package
mamba install -c bioconda segemehl -y
# leave the environment
mamba deactivate
```

Installation

```
module add mambaforge
mamba activate /storage/city/home/user_name/segemehl-0.3.4
segemehl.x ... # run the job
mamba deactivate
```

Usage in the job

https://docs.metacentrum.cz/software/install-software/#conda-packages

Local SW installations

- **Do your compilations** (GCC, Intel oneAPI, AOCC for AMD CPUs, Open MPI, CUDA for GPU support, CMake, etc.).

```
(BOOKWORM)vorel@zenith:~$ module ava aocc/
-------------------- /packages/run/modules-5/debian12avx512 --
aocc/2.2.0-aocc-2.2.0-jzzpamo  aocc/3.2.0-gcc-10.2.1-2ttmdfs
aocc/2.2.0-gcc-8.3.0-gkqq656
```

```
(BOOKWORM)vorel@zenith:~$ module ava intel*
--- /packages/run/modules-5/debian12avx512 -
intel-mkl/                  intel-tbb/
intel-oneapi-compilers/     intelcdk/
intel-oneapi-mkl/           intelmpi/
intel-oneapi-mpi/
intel-oneapi-tbb/
intel-parallel-studio/
```

```
(BOOKWORM)vorel@zenith:~$ module ava cuda/
-------------------- /packages/run/modules-5/debian12avx512 --------
cuda/3.2-kky   cuda/10.0.130-gcc-wwf2g          cuda/11.6.2-gcc-10.2.1-nwpmxyy
cuda/4.0       cuda/10.1
cuda/4.2       cuda/10.2.89-aocc-2.2.0-eluzh4v
```

```
(BOOKWORM)vorel@zenith:~$ module ava openmpi/
------------------------------------------------- /packages/run/modules-5/debian12avx512 ---------------------------------
openmpi/0-gcc          openmpi/1.8.2-intel        openmpi/3.1.2-intel              openmpi/4.0.3-aocc
openmpi/0-intel        openmpi/1.8.2-pgi          openmpi/3.1.2-intel-cuda         openmpi/4.0.4-aocc-2.2.0-gpu-wmtsoh4
```

Writing to the root directory

- Compute nodes and frontends have **limited quotas (977 MB) for writing out of the scratch or home directory**.
- Exceeding this quota will terminate the process.
- The most common problems are caused by:
  - Write to `/tmp` (typical for local SW installation).
  - Very large `stdout` and `stderr` streams.

```
ERROR: Could not install
packages due to an
OSError: [Errno 122] Disk
quota exceeded
```

```
export TMPDIR=$SCRATCHDIR

my_app < input … 1>$SCRATCHDIR/stdout 2>$SCRATCHDIR/stderr

my_app < input … 1>/dev/null 2>/dev/null
```

- Utility `check-local-quota` can be executed on each node.

https://docs.metacentrum.cz/troubleshooting/faqs/faqs-content/disk-quota-install/

Singularity/Apptainer containers

- Apptainer (former Singularity) is an **alternative to Docker for HPC**.
- **Apptainer is compatible with all Docker images and can be used with GPUs and MPI applications**.
- MetaCentrum offers pre-built (ready-to-use) Singularity images. For example, **images available under NGC - NVIDIA GPU Cloud** (Kaldi, PyTorch, TensorFlow), **Trinity** (RNA-seq assembler), **OpenFOAM** (numerical solver), etc.).
- NGC are highly optimised for GPU-accelerated calculations.

```
(BOOKWORM)vorel@nympha:~$ singularity run /cvmfs/singularity.metacentrum.cz/NGC/PyTorch\:23.11-py3.SIF pip list | grep torch
pytorch-quantization      2.1.2
torch                     2.2.0a0+6a974be
torch-tensorrt            2.2.0a0
torchdata                 0.7.0a0
torchtext                 0.16.0a0
torchvision               0.17.0a0
```

https://docs.metacentrum.cz/software/containers/ https://docs.metacentrum.cz/computing/nvidia-gpu/

General recommendations

- **Avoid non-effective calculations**
  - Optimise your calculations for the best hardware usage (do not reserve resources which will not be used).
- **A high number of short jobs**
  - From the point of view of performance (necessary PBS hardware requirements to run every single job), an ideal job lasts at least 30 minutes. Aggregate more short jobs into bigger ones with longer total walltime.
- **Use scratch directory**
  - Temporary storage for all data necessary on the physical compute node in jobs defined by `$SCRATCHDIR`.

https://docs.metacentrum.cz/computing/scratch-storages/

Useful utilities and commands

- `go_to_scratch` **utility**
  - Monitoring of running jobs or pick-up data. Redirection to compute node.

    ```
    go_to_scratch job_ID@PBS_server_full_name
    ```

- `qextend` **utility** https://docs.metacentrum.cz/computing/extend-walltime/
  - Walltime, which could be reserved by PBS, is limited to 720 hours.
  - Users are allowed to prolong their jobs in a limited number of cases.

    ```
    qextend job_ID@PBS_server_full_name additional_walltime_hh:mm:ss
    ```

- `pbs-get-job-history` **utility** https://docs.metacentrum.cz/computing/finished-jobs/
  - Get comprehensive information about historical jobs.

    ```
    pbs-get-job-history job_ID@PBS_server_full_name
    ```

cesnet

metacentrum

e-INFRA CZ

THANK YOU FOR YOUR ATTENTION

meta@cesnet.cz