Jak sejmout MetaCentrum v 10 krocích
How to take down MetaCentrum in 10 steps

Jiří Vorel, Cesnet z. s. p. o., MetaCentrum User Support, 21. 4. 2021

- **MetaCentrum is…**

    - … The National Grid Infrastructure (NGI)

    - … a provider of computational resources, application tools (commercial and free/open source) and data storage

    - … completely free of charge

        - Users "pay" by Acknowledgement in their research publications

        https://metavo.metacentrum.cz/cs/myaccount/pubs

- **MetaCentrum is available for…**

    - … employees and students from Czech universities, the Czech Academy of Science, non-commercial research facilities, etc.

    - … industry users (only for academic and non-profit and public research)

- **Cesnet (MetaCentrum)** also integrates computing clusters of academic institutions across the Czech academic community, including **project Elixir** + cooperation with centres **CERIT-SC (MU Brno)** and **IT4Innovation (VSB Ostrava)**
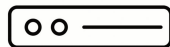
- 11 frontends submit jobs into three PBS servers
- All user home directories are available from all frontends
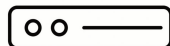
meta-pbs.metacentrum.cz →
- skirit.ics.muni.cz
- alfrid.meta.zcu.cz
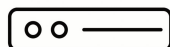- nympha.zcu.cz
- charon.nti.tul.cz
- …   …

elixir-pbs.elixir-czech.cz → elmo.elixir-czech.cz

cerit-pbs.cerit-sc.cz → zuphux.cerit-sc.cz

https://wiki.metacentrum.cz/wiki/Frontend_servers

```
my_home_pc:~$ ssh vorel@skirit.metacentrum.cz
vorel@skirit:~$
vorel@skirit:~$ pwd
/storage/brno2/home/vorel
vorel@skirit:~$ cd /storage/praha5-elixir/home/vorel
vorel@skirit:~$ pwd
/storage/praha5-elixir/home/vorel

my_home_pc:~$ ssh vorel@elmo.metacentrum.cz
vorel@elmo:~$
vorel@elmo:~$ pwd
/storage/praha5-elixir/home/vorel
```
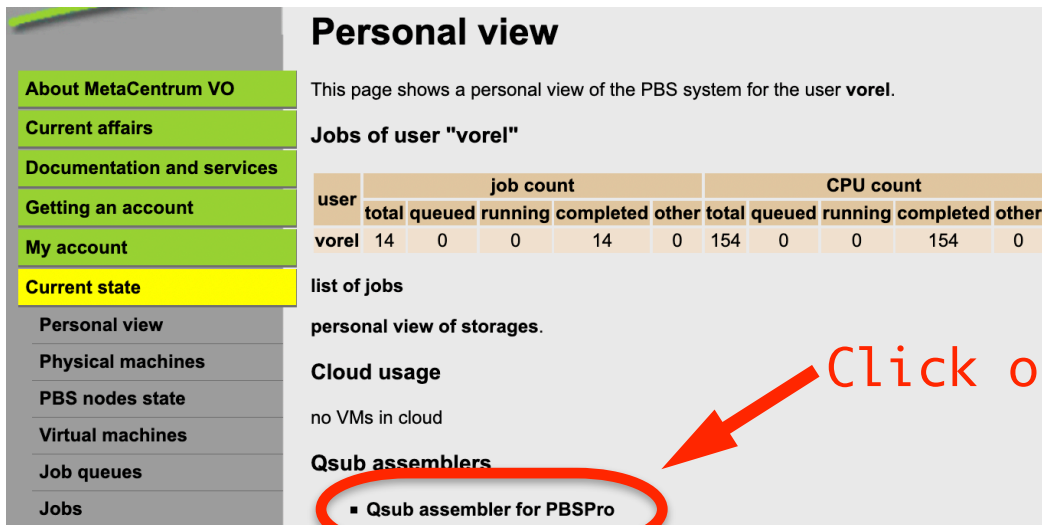
- Only a limited number of visible queues is suitable for direct use
- Which queues are most relevant for me?

Go to metavo.metacentrum.cz - Current state - Personal view - **Qsub assembler for PBSPro**
(Stav zdrojů - Osobní pohled - **Sestavovač qsub pro PBSPro**)

Personal view

This page shows a personal view of the PBS system for the user **vorel**.

Jobs of user "vorel"

user	job count					CPU count				
	total	queued	running	completed	other	total	queued	running	completed	other
vorel	14	0	0	14	0	154	0	0	154	0

list of jobs

personal view of storages.

Cloud usage

no VMs in cloud

Qsub assemblers

- Qsub assembler for PBSPro

Click on it…

About MetaCentrum VO
Current affairs
Documentation and services
Getting an account
My account
Current state
Personal view
Physical machines
PBS nodes state
Virtual machines
Job queues
Jobs

- You will be able to assemble *qsub* command and check if resources are available

Qsub assembler for PBSPro

This page assist in assembling correct parameters for the qsub command that is used for submitting jobs in PBSPro planners.

Only computing resources avaliable to the user **vorel** are offered.

```
qsub -l walltime=  24  :  0  :  0  -q  default@meta-pbs.metacentrum.cz  \
    -l select=  1  :ncpus=  4  :ngpus=  0  :mem=  50  gb  :scratch_  local  =  100  gb
    cluster ...
    city ...
    other resources ...
```

Find machines mathing the resource specification

selection from command line

qsub -l walltime=24:0:0 -q default@meta-pbs.metacentrum.cz -l select=1:ncpus=4:mem=50gb:scratch_local=100gb

selection in shell script

```
#!/bin/bash
#PBS -q default@meta-pbs.metacentrum.cz
#PBS -l walltime=24:0:0
#PBS -l select=1:ncpus=4:mem=50gb:scratch_local=100gb
#PBS -N my_awesome_job
```

✓ default@meta-pbs.metacentrum.cz
default@cerit-pbs.cerit-sc.cz
oven@meta-pbs.metacentrum.cz
gpu_titan@meta-pbs.metacentrum.cz
gpu@meta-pbs.metacentrum.cz
gpu_long@meta-pbs.metacentrum.cz
large_mem@meta-pbs.metacentrum.cz
global@meta-pbs.metacentrum.cz
backfill@meta-pbs.metacentrum.cz
elixircz@meta-pbs.metacentrum.cz
cloud@meta-pbs.metacentrum.cz
gpu_titan@cerit-pbs.cerit-sc.cz
gpu@cerit-pbs.cerit-sc.cz
uv_bio@cerit-pbs.cerit-sc.cz
phi@cerit-pbs.cerit-sc.cz
global@cerit-pbs.cerit-sc.cz
uv@cerit-pbs.cerit-sc.cz
elixir_2w_plus@elixir-pbs.elixir-czech.cz
elixir_2w@elixir-pbs.elixir-czech.cz
elixir_1d@elixir-pbs.elixir-czech.cz
large_mem@elixir-pbs.elixir-czech.cz
global@elixir-pbs.elixir-czech.cz
elixircz@elixir-pbs.elixir-czech.cz
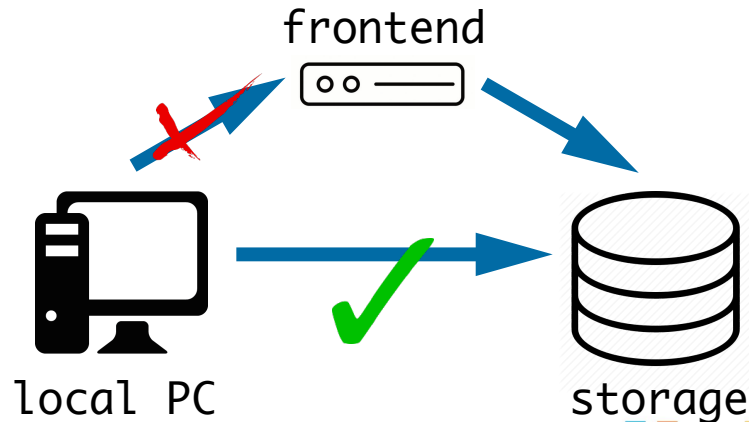
How to manage calculations and data properly

https://wiki.metacentrum.cz/wiki/Pruvodce_pro_zacatecniky

https://wiki.metacentrum.cz/wiki/Prace_s_daty

- Do not use frontend servers, copy data directly on storage, work with compressed files (.tar, .zip, .gz, etc.)

- SFTP client for Windows users (WinSCP)

frontend

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

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

local PC   storage

2. Do not run long calculations on frontends

https://wiki.metacentrum.cz/wiki/Pruvodce_pro_zacatecniky

- Is not appropriate to run long and demanding calculation directly on frontends and/or on clusters outside of PBS

- Ask for an **Interactive job…**

```
qsub -I  -l select=1:ncpus=2:mem=4gb:scratch_local=10gb -l walltime=1:00:00 \
-m abe
```

- Minimise the time lags in interactive jobs (-m flag)

  … or run a simple script for the **Batch job**

```
#!/bin/bash
#PBS -N Job_example
#PBS -l select=1:ncpus=2:mem=20gb:scratch_local=10gb
#PBS -l walltime=04:00:00
#PBS -m e

# clean scratch
trap 'clean_scratch' TERM EXIT

# define a DATADIR variable
DATADIR=/storage/city/home/user_name/dir/dir/

# copy input data to scratch directory
# varible SCRATCHDIR is set automatically
cp $DATADIR/input_data.fq $SCRATCHDIR

# move into scratch directory
cd $SCRATCHDIR

# load module for you application
module add fastQC-0.11.5

# run calculation (example)
fastqc < input_data.fq > out_results.html

# copy/more output back to DATADIR
mv out_results.html $DATADIR
```

Define resources, set job name and email alert

Scratch directory will be cleaned (more information on the next slide)

You can define as many variables as you want

- Very intensive I/O operations can cause network overload and the slowdown of central storage (`/storage/city/…`)

- Copy the input data into the scratch directory on a dedicated machine

- Variable SCRATCHDIR is set automatically

- Faster, more stable

_shared (on cluster, slower)
_ssd (faster, not everywhere)

```
qsub -I -l select=1:ncpus=1:mem=4gb:scratch_local=10gb -l walltime=1:00:00
cp my_input_data.txt $SCRATCHDIR
…
…
cp $SCRATCHDIR/my_results.txt /storage/city/home/user_name/
```

https://wiki.metacentrum.cz/wiki/Pruvodce_pro_zacatecniky#Typy_scratch_adres.C3.A1.C5.99.C5.AF

- Do not forget to clean the scratch directory when your calculation is done or have been killed by PBS

- You can do it **manually** after each finished job (but it won't be very pleasant) or **activate utility** `clean_scratch`

```
trap 'clean_scratch' TERM EXIT
cp my_input_data.txt $SCRATCHDIR
…
…
…
cp my_results.txt /storage/city/home/…  || export CLEAN_SCRATCH=false
```

- From the point of view of performance (necessary PBS hardware requirements to run every single job), an ideal job is running at least for 30 minutes

- Startup overhead may be a significant part of the whole processing time

- Try to imagine what happens when you submit 10k individual jobs at once with a real calculation time two minutes…

- Aggregate short jobs into bigger groups with longer walltime

```
#PBS -l walltime=00:30:00
```

… or more

- Computing nodes have very limited quotas (only 1 Gb) to write out of the scratch directory

- The most common problems are caused by:

  - Write to `/tmp`

  - Very large `stdout` and `stderr` streams

  ```
  export TMPDIR=$SCRATCHDIR
  ```

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

- Utility `check-local-quota` can be executed on each node (email notification)

- Optimise your calculations (hardware usage)

- Reservation of too many resources decrease your fairshare score and reduces the priority for your future jobs

  `https://wiki.metacentrum.cz/wiki/Fairshare`

- You can increase your fairshare score by acknowledgement to MetaCentrum in your publications

  `https://wiki.metacentrum.cz/wiki/Usage_rules/Acknowledgement`

- Effectivity can be checked on the computation node by standard Linux tools (`top, htop`) or on `metavo.metacentrum.cz` web portal

https://wiki.metacentrum.cz/wiki/Working_with_data#Data_archiving_and_backup

- MetaCentrum storage capacities are dedicated mainly to data for active usage

- Unnecessary data should be removed or moved to Cesnet Storage Department for long term archiving

  https://du.cesnet.cz/en/start

- MetaCentrum users can use the following archive:

  /storage/du-cesnet/home/user_name/VO_metacentrum-tape_tape-archive/

- And for backup:

  /storage/du-cesnet/home/user_name/VO_metacentrum-tape_tape/

- Parallel computing can significantly shorten the time of your job
  https://wiki.metacentrum.cz/wiki/Parallelization

- OpenMP (multiple threads) and MPI (set of nodes)

- Remember that **not all** application can utilise parallel/distributed computing

- You can request special nodes, which are interconnected by a low-latency `InfiniBand` (IB) connection to accelerate the speed of your job

```
qsub -l select=4:ncpus=4:mem=10gb:scratch_local=1gb \
     -l walltime=1:00:00 -l place=group=infiniband
```

- All users can install the software in their home folders

- Python, Perl and R libraries, Conda manager, pre-compiled binary, do your own compilations (gcc, intel, aocc), etc.

`https://wiki.metacentrum.cz/wiki/How_to_install_an_application`

- If for some reason grid infrastructure does not fulfil your expectations, maybe the MetaCentrum Cloud service would is a better choice

`https://cloud.metacentrum.cz/`
`cloud@metacentrum.cz`

- There is no reason to be afraid to use MetaCentrum

- By your activity, you are not able to "destroy" something. And if yes, it's our fault…

- Plenty of information and instructions can be found on our wiki
  https://wiki.metacentrum.cz, https://wiki.metacentrum.cz/wiki/FAQ

- When you are really lost - send an email
  meta@cesnet.cz

- Registration form for new users
  https://metavo.metacentrum.cz/en/application/index.html
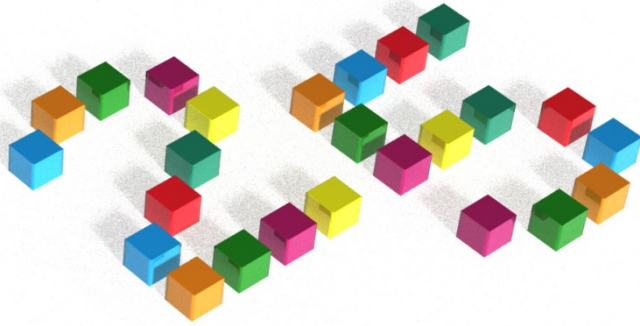
Thank you for your attention

Jiří Vorel, vorel@cesnet.cz meta@cesnet.cz