cesnet

metacentrum

How to handle MetaCentrum in a few steps

Jiří Vorel, Cesnet, MetaCentrum User Support, 15. 9. 2021

cesnet
metacentrum

- **MetaCentrum is…**
  - … a part of The National Grid Infrastructure (NGI),
  - … a provider of computational resources, application tools (commercial and free/open source) and data storage
  - … completely free of charge

- **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

https://metacentrum.cz

https://metavo.metacentrum.cz

https://wiki.metacentrum.cz

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

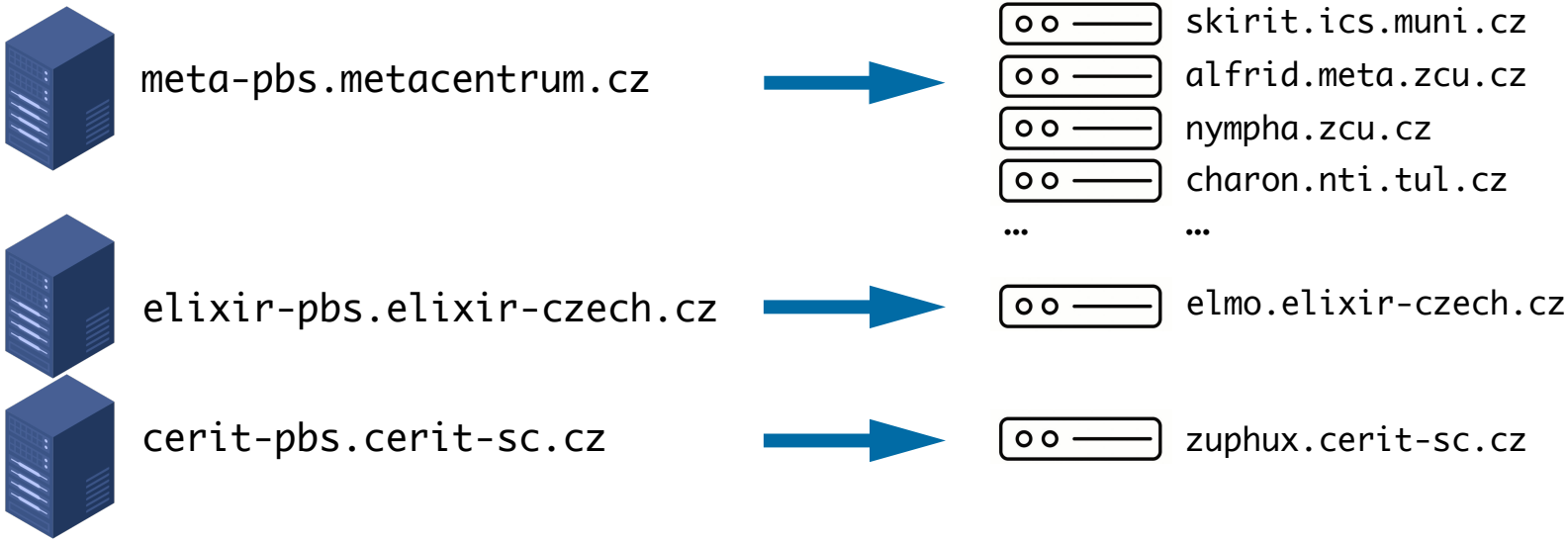
https://wiki.metacentrum.cz/wiki/FAQ/Grid_computing

https://wiki.metacentrum.cz/wiki/Reseni_problemu

Frontend servers and PBS

https://wiki.metacentrum.cz/wiki/Frontend

- 11 frontends, three PBS servers
- All user's home directories are available from all frontends



PBS and frontend servers

```
my_local_pc:~$ ssh vorel@elmo.metacentrum.cz
vorel@elmo.metacentrum.cz's password:
vorel@elmo:~$ pwd
/storage/praha5-elixir/home/vorel

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

SSH keys are not fully supported!

https://wiki.metacentrum.cz/wiki/NFS4_Servery

https://wiki.metacentrum.cz/wiki/Kerberos_authentication_system

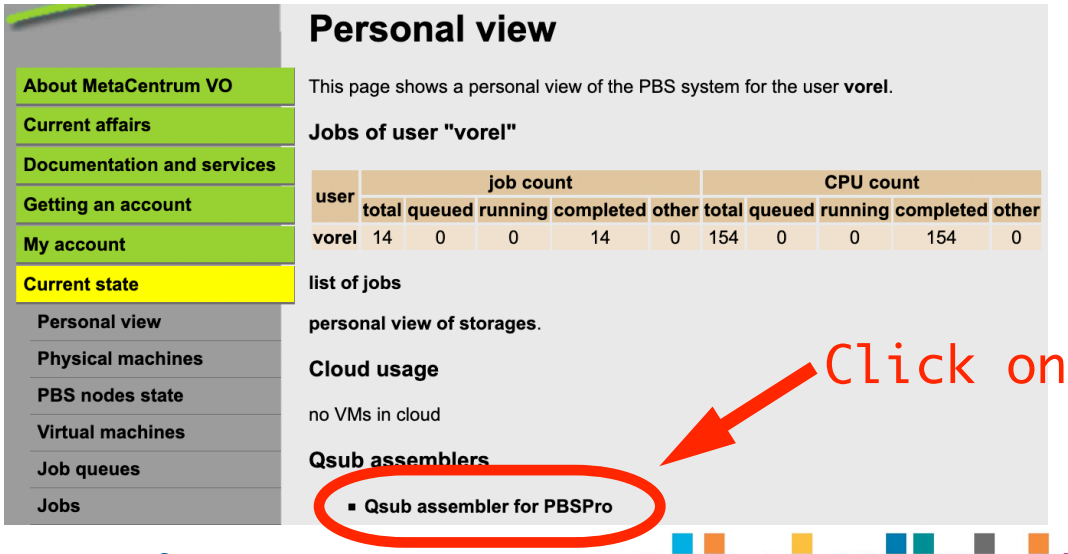
https://wiki.metacentrum.cz/wiki/Beginners_guide#Log_on_a_frontend_machine

https://wiki.metacentrum.cz/wiki/Frontend#Login_notes

Queues

- 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**)

Queues

- 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 [redacted] 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
- [redacted]
- gpu@meta-pbs.metacentrum.cz
- gpu_long@meta-pbs.metacentrum.cz
- large_mem@meta-pbs.metacentrum.cz
- global@meta-pbs.metacentrum.cz
- [redacted]
- [redacted]
- gpu@cerit-pbs.cerit-sc.cz
- phi@cerit-pbs.cerit-sc.cz
- global@cerit-pbs.cerit-sc.cz
- uv@cerit-pbs.cerit-sc.cz
- large_mem@elixir-pbs.elixir-czech.cz
- global@elixir-pbs.elixir-czech.cz

Queues

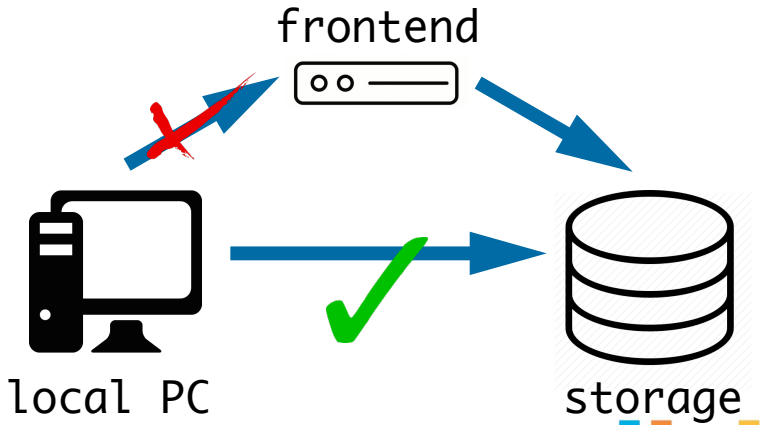
fronta		priorita	časové limity	úloh ve frontě běžících/max	hotových	celkem	max. na uživatele	max. CPU na uživatele
oven@meta-pbs.metacentrum.cz		95	0 - 720:00:00	0	10 /	0	10	20	
elixir_2w@meta-pbs.metacentrum.cz		90	24:00:01 - 336:00:00	0	0 /	0	0	1200	
elixir_2w_plus@meta-pbs.metacentrum.cz		90	336:00:01 - 720:00:00	0	0 /	0	0		2000
elixir_1d@meta-pbs.metacentrum.cz		90	00:00:00 - 24:00:00	0	0 /	0	0	1200	
gpu_titan@meta-pbs.metacentrum.cz		76	0 - 24:00:00	0	0 /	0	0		500
gpu@meta-pbs.metacentrum.cz		75	0 - 24:00:00	838	122 /	285	1466		500
gpu_long@meta-pbs.metacentrum.cz		55	0 - 336:00:00	5	24 /	6	35		200
large_mem@meta-pbs.metacentrum.cz		55	00:00:00 - 168:00:00	0	0 /	1	1	500	
q_4d@meta-pbs.metacentrum.cz		50	48:00:01 - 96:00:00	141	174 /	204	519		1000
q_2w@meta-pbs.metacentrum.cz		50	168:00:01 - 336:00:00	599	965 /	332	1896		1000
q_2h@meta-pbs.metacentrum.cz		50	0 - 02:00:00	12	13 /	41	69		
q_2d@meta-pbs.metacentrum.cz		50	24:00:01 - 48:00:00	3	33 /	14	50		1000
q_1w@meta-pbs.metacentrum.cz		50	96:00:01 - 168:00:00	2	60 /	39	101		1000
global@meta-pbs.metacentrum.cz		50	0 - 48:00:00	1	1 /	355	357	1000	1000
q_1d@meta-pbs.metacentrum.cz		50	04:00:01 - 24:00:00	46	58 /	2671	3371		1500
q_4h@meta-pbs.metacentrum.cz		50	02:00:01 - 04:00:00	1	2004 /	10227	17321		
q_2w_plus@meta-pbs.metacentrum.cz		50	336:00:01 - 720:00:00	1828	628 /	67	2523		2000
backfill@meta-pbs.metacentrum.cz		20	00:00:01 - 24:00:00	0	0 /	0	0		
elixircz@meta-pbs.metacentrum.cz		0	0 - 720:00:00	0	0 /	0	0		
cloud@meta-pbs.metacentrum.cz		0	0 - 720:00:00	0	0 /	0	0		
default@meta-pbs.metacentrum.cz		0	0 - 720:00:00	0	0 /	0	7		
gpu_titan@cerit-pbs.cerit-sc.cz		76	0 - 24:00:00	0	0 /	0	0		500
gpu@cerit-pbs.cerit-sc.cz		75	0 - 24:00:00	0	26 /	90	1254		
uv_long@cerit-pbs.cerit-sc.cz		35	96:00:01 - 168:00:00	0	0 /	0	0	8000	
uv_large@cerit-pbs.cerit-sc.cz		34	00:00:01 - 96:00:00	0	0 /	0	0	8000	
uv_bio@cerit-pbs.cerit-sc.cz		31	00:00:01 - 96:00:00	0	0 /	0	0	8000	
uv_small@cerit-pbs.cerit-sc.cz		30	00:00:01 - 96:00:00	0	21 /	30	51	8000	
phi@cerit-pbs.cerit-sc.cz		30	00:00:01 - 336:00:00	0	5 /	4	9	8000	
q_1w@cerit-pbs.cerit-sc.cz		20	96:00:01 - 168:00:00	142	146 /	108	397	2000	1000
global@cerit-pbs.cerit-sc.cz		20	0 - 48:00:00	0	0 /	0	0	1000	1000
q_4d@cerit-pbs.cerit-sc.cz		20	48:00:01 - 96:00:00	0	9 /	9	18	2000	1000
q_1d@cerit-pbs.cerit-sc.cz		20	04:00:01 - 24:00:00	1	297 /	807	1308	2000	1000
q_2w@cerit-pbs.cerit-sc.cz		20	168:00:01 - 336:00:00	7	53 /	6	67	2000	1000
q_2h@cerit-pbs.cerit-sc.cz		20	0 - 02:00:00	1	1 /	1	14	2000	
q_2d@cerit-pbs.cerit-sc.cz		20	24:00:01 - 48:00:00	0	27 /	1	28	2000	1000
q_4h@cerit-pbs.cerit-sc.cz		20	02:00:01 - 04:00:00	0	941 /	11013	33036	1000	1000
q_2w_plus@cerit-pbs.cerit-sc.cz		20	336:00:01 - 720:00:00	1	31 /	13	45	2000	1000
default@cerit-pbs.cerit-sc.cz		0	0 - 720:00:00	6006	0 /	0	6008		
uv@cerit-pbs.cerit-sc.cz		0	00:00:01 - 168:00:00	0	0 /	0	0		
elixir_2w_plus@elixir-pbs.elixir-czech.cz		90	336:00:01 - 720:00:00	0	0 /	0	0		200
elixir_2w@elixir-pbs.elixir-czech.cz		90	24:00:01 - 336:00:00	0	0 /	19	19	1200	600
elixir_1d@elixir-pbs.elixir-czech.cz		90	00:00:00 - 24:00:00	0	0 /	16	16	600	
large_mem@elixir-pbs.elixir-czech.cz		55	00:00:00 - 168:00:00	0	4 /	6	10	500	
global@elixir-pbs.elixir-czech.cz		50	0 - 48:00:00	2	71 /	591	687	1000	1000
elixircz@elixir-pbs.elixir-czech.cz		0	0 - 720:00:00	0	0 /	0	0		

https://metavo.metacentrum.cz/pbsmon2/person

1. Transfer large amount of data

https://wiki.metacentrum.cz/wiki/Pruvodce_pro_zacatecniky
https://wiki.metacentrum.cz/wiki/Prace_s_daty
https://wiki.metacentrum.cz/wiki/NFS4_Servers

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




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

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

2. Do not run long calculations on frontends

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

- It is not appropriate to run long and demanding calculations 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
```

- You can minimise the time lags in interactive jobs (`-m` flag)

2. Do not run long calculations on frontends

```
#!/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

The scratch directory will be cleaned (more information on the next slides)

You can define as many variables as you want

https://wiki.metacentrum.cz/wiki/Pruvodce_pro_zacatecniky#D.C3.A1vkov.C3.A9_.C3.BAlohy

3. Use the scratch directory

- Very intensive I/O operations can cause network overload and a slowdown of central storage (`/storage/city/…`)
- Copy the input data into the scratch directory on a dedicated machine
- $SCRATCHDIR will be 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

4. Clean the scratch directory

- Do not forget to clean the scratch directory when your calculation is finish or have been killed by PBS
- You can do it **manually** after each finished job 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
```


5. A high number of very short jobs

- 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 of two minutes
- Aggregate short jobs into bigger groups with longer walltime

```
#PBS -l walltime=00:30:00 and or more
```


6. Writing outside of the scratch directory

- Computing nodes have very limited quotas (only 1 Gb) outside of the scratch directory
- The most common problems are caused by:
  - Writing 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 notifications

7. Avoid non-effective calculation

- Optimise your calculations (hardware usage)
- Reservation of too many resources decrease your fairshare score and reduces the priority for your future calculations

  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

8. Backup and archiving

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

- MetaCentrum storage capacities are dedicated mainly to data in 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/`

9. Parallel computing and IB acceleration

- 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** applications can utilise parallel computing.
  A typical MC machine has => 32 CPUs
- 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=8:ncpus=10:mpiprocs=10:ompthreads=1:mem=100gb:scratch_local=10gb \
                -l walltime=24:00:00 -l place=group=infiniband
```


Final notes

- All users can install the software on their own
- Python, Perl and R libraries, Conda package manager, pre-compiled binary, new 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 be a better choice

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

Take-home message

- There is no reason to be afraid to use MetaCentrum
- By your activity, you are not able to "destroy" something
- You can find plenty of information and instructions on our wiki
  https://wiki.metacentrum.cz, https://wiki.metacentrum.cz/wiki/FAQ
- Are you really lost? Send an email!
  meta@cesnet.cz
- Registration form for new users
  https://metavo.metacentrum.cz/en/application/index.html

cesnet

metacentrum

Thank you for your attention

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