cesnet
metacentrum

e-INFRA cz

Gridová služba MetaCentrum - Best Practices s následnou volnou diskuzí o tématech, která Vás zajímají a nedají Vám spát

Grid service MetaCentrum - Best Practices followed by a discussion on topics that interest you and will not let you sleep

Jiří Vorel
MetaCentrum User Support
vorel@cesnet.cz meta@cesnet.cz

---

10. 5. 2022
Prague

What MetaCentrum is

- **MetaCentrum is** https://metacentrum.cz
  - … The National Grid Infrastructure (NGI) https://metavo.metacentrum.cz
  - … the activity of the CESNET association https://wiki.metacentrum.cz
  - … a provider of computational resources, application tools (commercial and free/open source) and data storage
  - … free of charge https://wiki.metacentrum.cz/wiki/Usage_rules/Acknowledgement
    - Users "pay" by Acknowledgement in their research publications
- **MetaCentrum is available for** https://metavo.metacentrum.cz/en/myaccount/pubs
  - … employees and students from Czech universities, the Czech Academy of Science, non-commercial research facilities, etc.
  - … industry users (only for non-profit and public research)

Where and how to start

- **Fill out and submit the registration form**

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

  - Select your organisation (click on the eduID logo)
  - Use your institutional username and password
  - Fill out the form and create a **strong** MetaCentrum password
  - Users must extend MetaCentrum membership from the beginning of each calendar year (typically during January)
  - MetaCentrum users obtain access to CERIT-SC resources automatically

- **Read our documentation, FAQ and tutorial for beginners**

  https://wiki.metacentrum.cz/wiki/Main_Page

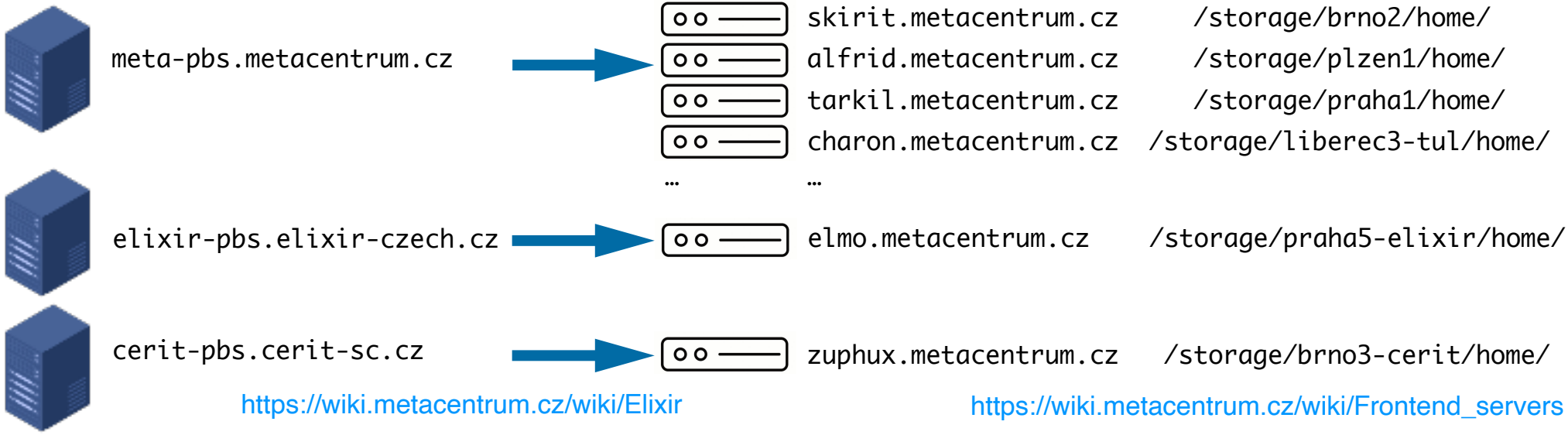
  https://wiki.metacentrum.cz/wiki/Beginners_guide

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

  https://wiki.metacentrum.cz/wiki/Troubleshooting

PBS and frontend servers

- Ten frontends (+ one alias) submit jobs to three PBS servers
- PBS (Portable Batch System) is a software that performs job scheduling
- Frontend servers can have different home directories
- All user home directories are available from all frontends



https://wiki.metacentrum.cz/wiki/Elixir

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

PBS and frontend servers

```
name@my_pc:~$ ssh vorel@perian.metacentrum.cz
vorel@perian.metacentrum.cz's password:
vorel@perian:~$ pwd
/storage/brno2/home/vorel
vorel@perian:~$ cd /storage/plzen1/home/vorel
vorel@perian:~$ pwd
/storage/plzen1/home/vorel

name@my_pc:~$ ssh vorel@minos.metacentrum.cz
vorel@minos.metacentrum.cz's password:
vorel@minos:~$ pwd
/storage/plzen1/home/vorel
```

Type a password

Where am I?

Hmm, I forgot that my data is on different storage...

Direct access to the same storage through different frontend

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

SSH keys and Kerberos authentication

- SSH keys for logging into frontends **are not fully supported**. We want to "force" you to generate a Kerberos ticket by typing the password

```
jirivorel@MacBook ~$ ssh vorel@nympha.metacentrum.cz
vorel@nympha.metacentrum.cz's password:
(BULLSEYE)vorel@nympha:~$ klist
Credentials cache: FILE:/tmp/krb5cc_1597_LTYWLt
        Principal: vorel@META

  Issued                Expires               Principal
May  6 11:22:55 2022  May  6 21:22:55 2022  krbtgt/META@META
May  6 11:22:55 2022  May  6 21:22:55 2022  afs/ics.muni.cz@META
May  6 11:22:55 2022  May  6 21:22:55 2022  krbtgt/ZCU.CZ@META
May  6 11:22:55 2022  May  6 21:22:55 2022  afs/zcu.cz@ZCU.CZ
(BULLSEYE)vorel@nympha:~$ ssh halmir1
Linux halmir1.metacentrum.cz 5.10.0-13-amd64 #1 SMP Debian 5.10.106-1+zs1 (2022-03-28) x86_64
Last login: Thu Apr 21 09:54:05 2022 from elmo2-4.hw.elixir-czech.cz
(BULLSEYE)vorel@halmir1:~$
```

Type a password

`klist` command prints the status of issued tickets

SSH keys and Kerberos authentication

```
(BULLSEYE)vorel@nympha:~$ klist
klist: No ticket file: /tmp/krb5cc_1597_rw50KaLk0H
(BULLSEYE)vorel@nympha:~$ qsub -I -l select=1:ncpus=1:mem=5gb:scratch_local=1gb -l walltime=1:00:00
No Kerberos credentials found.
(BULLSEYE)vorel@nympha:~$ ssh halmir1
vorel@halmir1's password:

(BULLSEYE)vorel@nympha:~$ kinit
vorel@META's Password:
```

`kinit` command generates new tickets

- You can have the Kerberos ticket issued on your personal computer. During the validity of the ticket, you can log in to every frontend, compute node or storage without entering a password again

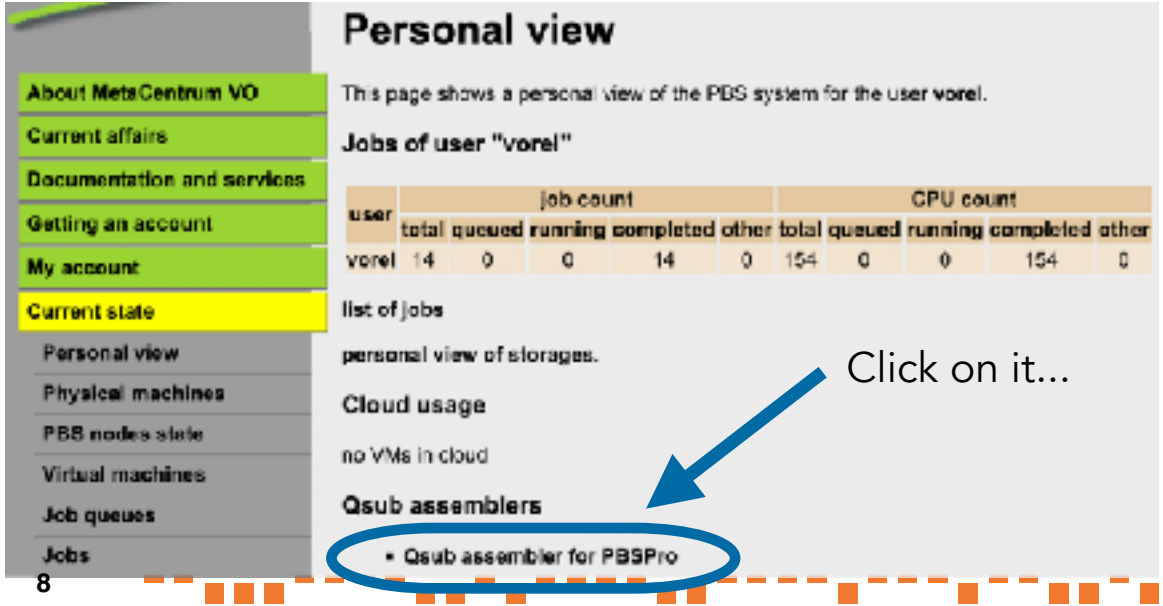
https://wiki.metacentrum.cz/wiki/Kerberos_authentication_system

https://wiki.metacentrum.cz/wiki/Kerberos_on_Windows

https://wiki.metacentrum.cz/wiki/Kerberos_on_Linux

HW resources and qsub assembler

- HW resources (CPUs, GPUs, RAM, scratch, walltime,...) are reserved by PBS
- Detailed documentation: https://wiki.metacentrum.cz/wiki/About_scheduling_system
- It requires some experience
- Helper tool for qsub command assembly



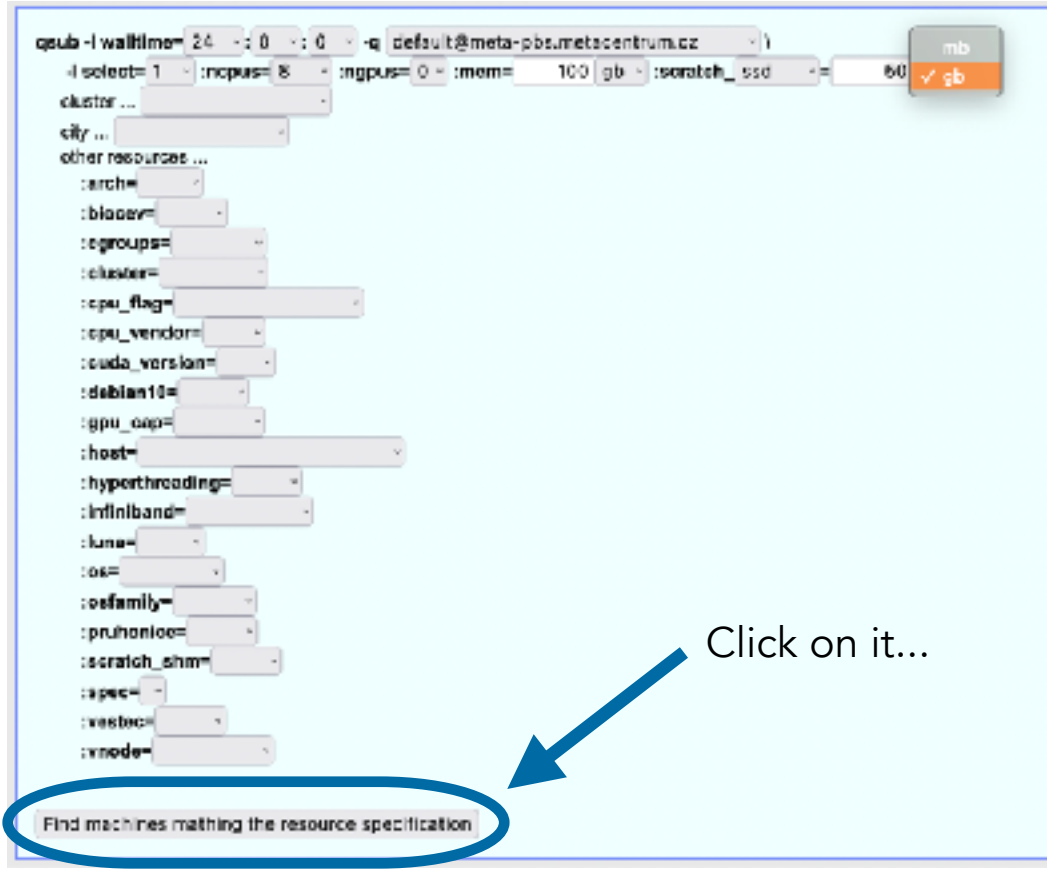
Go to `metavo.metacentrum.cz` - Current state - Personal view - **Qsub assembler for PBSPro**

(Stav zdrojů - Osobní pohled **Sestavovač qsub pro PBSPro**)

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

cesnet
metacentrum

HW resources and qsub assembler



- And you will see...

selection from command line

qsub -l walltime=24:0:0 -q default@meta-pbs.metacentrum.cz -l select=1:ncpus=8:mem=100gb:scratch_ssd=50gb

selection in shell script

```
#!/bin/bash
#PBS -q default@meta-pbs.metacentrum.cz
#PBS -l walltime=24:0:0
#PBS -l select=1:ncpus=8:mem=100gb:scratch_ssd=50gb
#PBS -N my_awesome_job
```

Result

OK

The requirement is 1 machine, and 93 such machines are free, out of 289 machines matching the requirements. The job for it.

Machines available right now

Example of a basic script for batch jobs

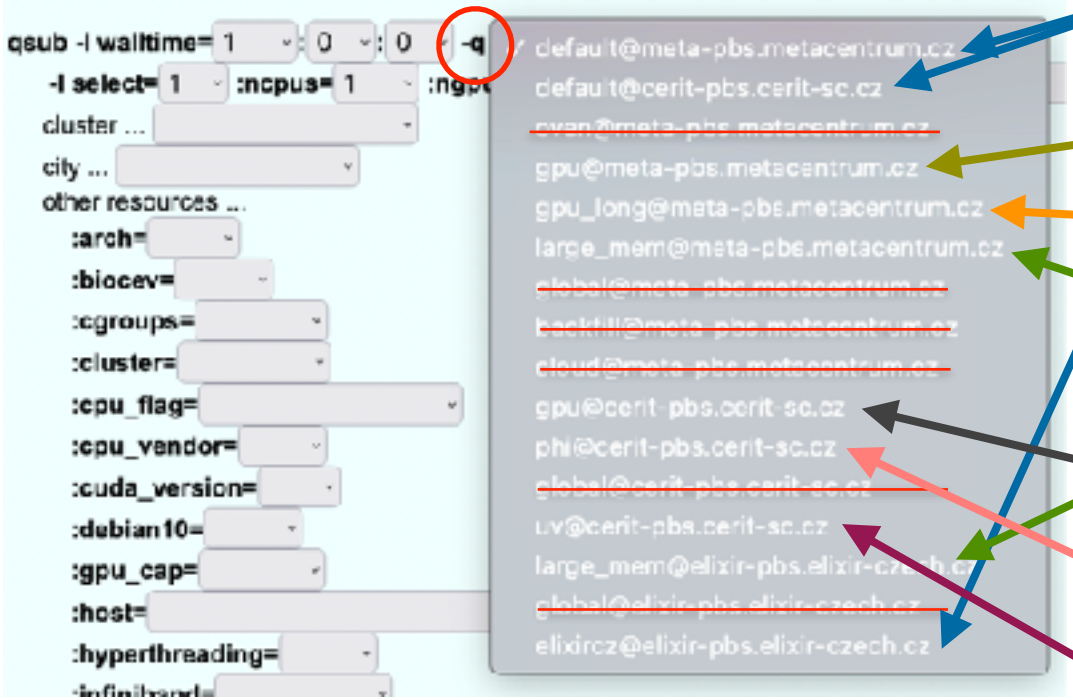
```
#!/bin/bash
#PBS -q default@meta-pbs.metacentrum.cz
#PBS -l walltime=24:0:0
#PBS -l select=1:ncpus=8:mem=100gb:scratch_ssd=50gb
#PBS -N my_awesome_job
#PBS -m e

# test if a scratch directory exists
# variable SCRATCHDIR is set automatically
test -n "$SCRATCHDIR" || { echo >&2 "Variable SCRATCHDIR is not set!"; exit 1; }

# set a DATADIR variable
DATADIR=/storage/brno12-cerit/home/vorel/data/

# copy input file "data.fa" to the scratch directory
cp $DATADIR/data.fa $SCRATCHDIR

# move into the scratch directory
cd $SCRATCHDIR

# load a module for your application
module add blast-plus/blast-plus-2.12.0-gcc-8.3.0-ohlv7b4

# run the calculation
# do not forgete to use reserved CPUs by '-num_threads' flag
# variable PBS_NCPUS is a number of CPUs requested for the entire job
blastp -query data.fa <other_parameters> -num_threads $PBS_NCPUS -out results.txt

#copy results
cp results.txt $DATADIR

# clean the scratch directory
clean_scratch
```

- Define HW resources (`-l`), queue (`-q`) and walltime (`-l`), set the job name (`-N`) and email alert (`-m`)
- You can define as many variables as you want
- Available modules can be listed by command `module ava <key_word>` on any frontend
- The scratch directory will be cleaned automatically

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

Queues

- Not all visible queues are suitable for direct use
- Explore the -q option of the qsub assembler



Queues for jobs requesting up to 720 hours

GPU jobs up to 24 hours on MetaCentrum nodes

GPU jobs up to 336 hours on MetaCentrum nodes

Queues prioritising jobs requesting more than 500 GB RAM

GPU jobs up to 24 hours on CERIT-SC nodes

Nodes with Intel Xeon Phi 7210

Individual SMP machines with OS CentOS 7

Queues

Queue default@meta-pbs.metacentrum.cz

Default queue (routing)

The queue is routing, it delivers jobs depending on their walltime to the following queues:

queue		Priority	time limits	jobs					max CPUs per user	fairshare
				queued	running /max	completed	total	max jobs per user		
q_2h@meta-pbs.metacentrum.cz	⛔	50	0 - 02:00:00	381	0 /	5678	6059		2000	
q_4h@meta-pbs.metacentrum.cz	⛔	50	02:00:01 - 04:00:00	1091	1957 /	12972	18078			
q_1d@meta-pbs.metacentrum.cz	⛔	50	04:00:01 - 24:00:00	2270	100 /	4153	6538		4000	
q_2d@meta-pbs.metacentrum.cz	⛔	50	24:00:01 - 48:00:00	126	11 /	150	287		1000	
q_4d@meta-pbs.metacentrum.cz	⛔	50	48:00:01 - 96:00:00	2036	1863 /	531	4430		1000	
q_1w@meta-pbs.metacentrum.cz	⛔	50	96:00:01 - 168:00:00	55	1507 /	1281	2944		1000	
q_2w@meta-pbs.metacentrum.cz	⛔	50	168:00:01 - 336:00:00	83	99 /	37	219		1000	
q_2w_plus@meta-pbs.metacentrum.cz	⛔	50	336:00:01 - 720:00:00	28	709 /	70	807		2000	

do not submit to the queue directly, use a routing queue instead

uv_bio@cerit-pbs.cerit-sc.cz	🔒	31	00:00:01 - 96:00:00	0	0 /
uv_small@cerit-pbs.cerit-sc.cz	⛔	30	00:00:01 - 96:00:00	20	12 /
fireprot_devel@cerit-pbs.cerit-sc.cz	🔒	50	0 - 336:00:00	0	0 /

reserved for: leontovyc_roman simekmilos vorel

Interactive job

- The opposite of batch jobs (waiting for the user's input...)
- Best choice for test calculations (which should not be run directly on frontends)
- An interactive job is requested by the qsub command with the -I (uppercase "i") option

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

```
(BUSTER)vorel@skirit:~$ qsub -I -l select=1:ncpus=4:mem=50gb:scratch_local=30gb -l walltime=1:00:00
qsub: waiting for job 11405230.meta-pbs.metacentrum.cz to start
qsub: job 11405230.meta-pbs.metacentrum.cz ready

vorel@zenon31:~$ cd $SCRATCHDIR
vorel@zenon31:/scratch.ssd/vorel/job_11405230.meta-pbs.metacentrum.cz$ module add orca/orca-5.0.1-intel-19.0.4-bnofsgq
vorel@zenon31:/scratch.ssd/vorel/job_11405230.meta-pbs.metacentrum.cz$ module list
Currently Loaded Modulefiles:
  1) metabase                                 2) openmpi/openmpi-4.0.4-intel-19.0.4-gpu-xri6uan   3) orca/orca-5.0.1-intel-19.0.4-bnofsgq
vorel@zenon31:/scratch.ssd/vorel/job_11405230.meta-pbs.metacentrum.cz$
vorel@zenon31:/scratch.ssd/vorel/job_11405230.meta-pbs.metacentrum.cz$ ...time for coffee...
-bash: ...time: command not found
vorel@zenon31:/scratch.ssd/vorel/job_11405230.meta-pbs.metacentrum.cz$ orca < input > output
```

Scratch storage

- Temporary storage on physical computing nodes
- Very intensive 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 will be set automatically
- Faster, more stable

```
qsub -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/Beginners_guide#Specify_scratch_directory

Scratch storage

- MetaCentrum offers four types of scratch
  - `scratch_local`
    - on every node, HDD, default
  - `scratch_ssd`
    - fast SSD, typically smaller in volume, not everywhere
  - `scratch_shared`
    - network volume which is shared between all clusters in a given location, not everywhere
  - `scratch_shm`
    - scratch held in RAM, very fast, on every node
    - boolean type (`True`/`False`), limited by mem parameter (`:mem=XYgb`)

https://wiki.metacentrum.cz/wiki/Scratch_storage

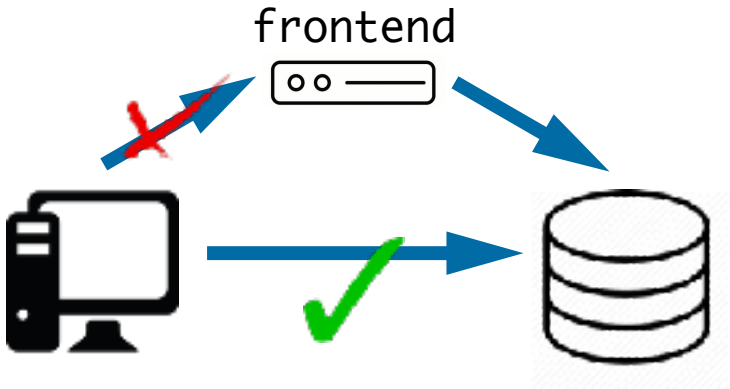
cesnet
metacentrum

e-INFRA CZ

Common issues and how to deal with them

Transfer of a large amount of data

- Do not use frontends, copy data directly on storage, use compressed files (.tar, .zip, .gz, etc.)
- SFTP client for Windows users (WinSCP, FileZilla, CyberDuck)




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

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

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

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

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 is running at least for 30 minutes
- Startup overhead may be a significant part of the whole processing time
- Aggregate short jobs into bigger groups with longer walltime

```
-l walltime=00:30:00 (and more)
```

Writing to the root directory

- Computing nodes and frontends have limited quotas (~ 1 GB) for writing out of the scratch and home directory
- Exceeding this quota will cause the termination of the process
- The most common problems are caused by:
  - Write to `/tmp`
  - Very large `stdout` and `stderr` streams

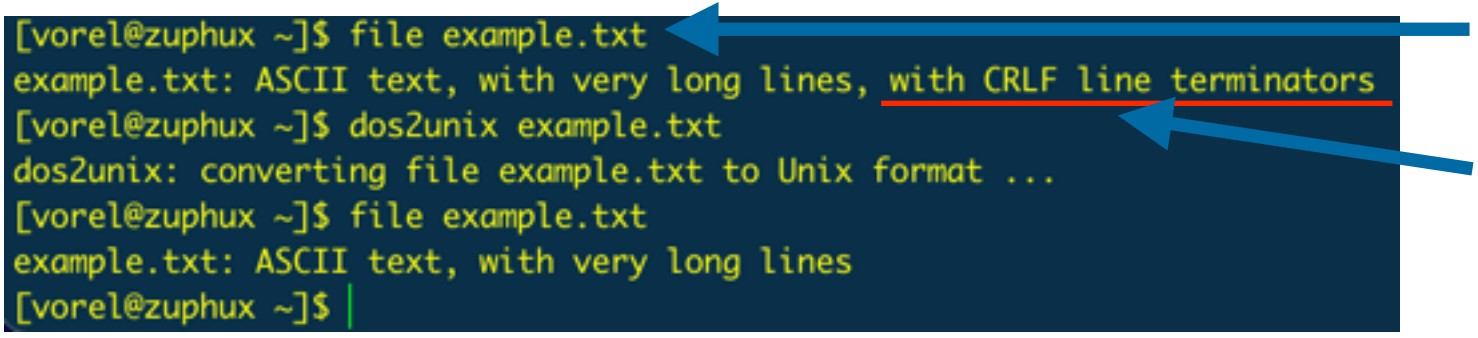
```
export TMPDIR=$SCRATCHDIR
my_app < input … 1>$SCRATCHDIR/stdout 2>$SCRATCHDIR/stderr
```

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

CRLF line terminators

- Text files created on Windows PC use more characters for termination of a line
- This format can not be read by Unix-like systems
- Individual lines are not recognised
- Utility `dos2unix` can fix the line terminators
- Typical errors:
  - `'\r': command not found  , EXIT^M: invalid signal specification`

https://owasp.org/www-community/vulnerabilities/CRLF_Injection



Toole `file` will determine the type of a file

Problem detected

Incorrectly submitted jobs

- Sometimes PBS accept a job with requirements which can never be satisfied
- Typically, this is an attempt to run the job as soon as possible.
- It's mostly counterproductive...

- Typical scenarios:
  - Incompatible Cuda versions and GPU machines
  - Wrong combinations of machines and queues
  - Combinations of parameters targeting a disparate set of machines

cesnet
metacentrum

Incorrectly submitted jobs

požadované prostředky	1:mem=16gb:scratch_local=10gb:ngpus=1:gpu_cap=cuda60:cuda_version=11.0
vytvořena	neděle 27. února 2022 19:46:54
způsobilá k běhu	neděle 27. února 2022 19:46:54
poslední změna stavu	neděle 27. února 2022 19:50:19
komentář	Can Never Run: Insufficient amount of resource: cuda_version (11.0 != ^11.2,^11.4,11.2,11.4)

https://wiki.metacentrum.cz/wiki/GPU_clusters

GPU clusters in MetaCentrum

Cluster	Nodes	GPUs per node	Compute Capability	CuDNN	gpu_cap=	cuda_version=
galcon.metacentrum.cz	galdor.metacentrum.cz - galdor26.metacentrum.cz	4x A40 48GB	8.6	YES	cuda35,cuda61,cuda75,cuda80,cuda86	11.4
ter.natur.cuni.cz	ter1.natur.cuni.cz - ter5.natur.cuni.cz	8x [illegible]	8.6	YES	[illegible]	11.2
zefron.cerit-sc.cz	zefron8.cerit-sc.cz	1x A10 24GB	8.6	YES	cuda35,cuda61,cuda75,cuda80,cuda86	11.2
zia.cerit-sc.cz	zia1.cerit-sc.cz - zia5.cerit-sc.cz	4x A100 40GB	8.0	YES	cuda35,cuda61,cuda75,cuda80	11.2
fau.natur.cuni.cz	fau1.natur.cuni.cz - fau3.natur.cuni.cz	8x Quadro RTX 5000 16GB	7.5	YES	cuda35,cuda61,cuda75	11.2
cha.natur.cuni.cz	cha.natur.cuni.cz	8x GeForce RTX 2080 Ti 11GB	7.5	YES	cuda35,cuda61,cuda75	11.2
gta.cerit-sc.cz	gta1.cerit-sc.cz - gta7.cerit-sc.cz	2x GeForce RTX 2080 Ti 11GB	7.5	YES	cuda35,cuda61,cuda75	11.2
adan.grid.cesnet.cz	adan1.grid.cesnet.cz - adan61.grid.cesnet.cz	2x Tesla T4 16GB	7.5	YES	cuda35,cuda61,cuda75	11.2
glados.cerit-sc.cz	glados2.cerit-sc.cz - glados7.cerit-sc.cz	2x GeForce RTX 2080 8GB	7.5	YES	cuda35,cuda61,cuda75	11.2
glados.cerit-sc.cz	glados1.cerit-sc.cz	TITAN V GPU 12GB	7.0	YES	cuda35,cuda61,cuda70	11.2
konos.fav.zcu.cz	konos1.fav.zcu.cz - konos8.fav.zcu.cz	4x GeForce GTX 1080 Ti 12GB	6.1	YES	cuda35,cuda61	11.2
glados.cerit-sc.cz	glados10.cerit-sc.cz - glados13.cerit-sc.cz	2x 1080 Ti GPU 12GB	6.1	YES	cuda35,cuda61	11.2
zefron.cerit-sc.cz	zefron7.cerit-sc.cz	GeForce GTX 1070 8GB	3.5	YES	cuda35, cuda61	11.2
black1.cerit-sc.cz	black1.cerit-sc.cz	Tesla P100 16GB	6.0	YES	cuda35, cuda60	11.2
grinbold.metacentrum.cz	grinbold.metacentrum.cz	2x Tesla P100	6.0	YES	cuda35, cuda60	11.2
zefron.cerit-sc.cz	zefron5.cerit-sc.cz	Tesla P4C 12GB	3.5	YES	cuda35	11.2
zubat.ncbr.muni.cz	zubat1.ncbr.muni.cz - zubat8.ncbr.muni.cz	2x Tesla K20Xm 6GB (aka Kepler)	3.5	YES	cuda35	11.2

We have only GPU machinech with cuda version 11.2 or 11.4

Incorrectly submitted jobs

požadované prostředky	1:ngpus=10:mem=300gb:scratch_local=100gb:cpu_flag=avx:mpiprocs=1:ompthreads=10
vytvořena	čtvrtek 31. března 2022 14:51:33

Probably just a typo; ngpus can be max. 4; no ncpus parameter

požadované prostředky	1:ncpus=8:cl_haldir=True:mpiprocs=8:ompthreads=1
vytvořena	středa 30. března 2022 10:31:09
způsobilá k běhu	středa 30. března 2022 10:31:09
poslední změna stavu	středa 30. března 2022 10:39:13
komentář	Can Never Run: Insufficient amount of resource: cl_haldir (True != False)

Haldir cluster has been shut down last year. GPU cluster Doom as well

požadované prostředky	1:ncpus=8:cl_doom=True:mpiprocs=8:ompthreads=1
vytvořena	středa 30. března 2022 10:24:08
způsobilá k běhu	středa 30. března 2022 10:24:08
poslední změna stavu	středa 30. března 2022 10:28:53
komentář	Can Never Run: Insufficient amount of resource: cl_doom (True != False)

Incorrectly submitted jobs

úloha	server	CPU	vyhraz. paměť	použitá paměť	jméno	uživatel	CPU čas
[redacted]	elixir	1	4gb	0b	STDIN	[redacted]	
požadované prostředky	1:ncpus=1:ngpus=2:mem=4gb:mpiprocs=1:ompthreads=1						
vytvořena	úterý 3. května 2022 16:37:25						
způsobilá k běhu	úterý 3. května 2022 16:37:25						
poslední změna stavu	úterý 3. května 2022 16:37:54						
komentář	Can Never Run: Insufficient amount of resource: ngpus (R: 2 A: 0 T: 0)						

There are no GPU machines in Elixir PBS

úloha	server	CPU	vyhraz. paměť	po
[redacted]	meta	1	4gb	
požadované prostředky	1:ncpus=1:mem=4gb:scratch_local=10gb:cluster=elmo3:mpiprocs=1:ompthreads=1			
vytvořena	pondělí 9. května 2022 21:31:34			
způsobilá k běhu	pondělí 9. května 2022 21:31:34			
poslední změna stavu	pondělí 9. května 2022 21:35:55			

Cluster elmo3 belongs under Elixir PBS

Debian11 is here

- As always, we keep Debian up-to-date on our nodes
- Now we are gradually upgrading from Deb10 (BUSTER) to Deb11 (BULLSEYE)
- However, some libraries may be missing on the new system...

```
gmx_mpi: error while loading shared libraries: libevent_core-2.1.so.6:
cannot open shared object file: No such file or directory
```

- Therefore we provide universal modules with these missing libraries

```
(BUSTER)vorel@skirit:~$ module ava debian

------------------------------------------------------------------
debian10-compat debian7-compat  debian8-compat  debian9-compat
(BUSTER)vorel@skirit:~$ module add debian10-compat
(BUSTER)vorel@skirit:~$ ls /software/debian-compat/debian*/lib
```

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

Debian11 is here

- Have you ever noticed that name in parentheses at the beginning of a line?

```
(BULLSEYE)vorel@nympha:~$    ... for Debian11


(BUSTER)vorel@nympha:~$    ... for Debian10


(STRETCH)vorel@nympha:~$    ... for Debian9
```

- Do you know what does it mean?

Debian11 is here

https://www.debian.org/releases/index.en.html

cesnet

metacentrum

e-INFRA CZ

Final notes

New services

- Kubernetes/Rancher (CERIT-SC)
  - Ready-to-use container-based applications (docker images)
  - GPU support (Nvidia A40)
  - Runs in browser with GUI
  - JupyterHub, BinderHub Nextflow, KNIME, Ansys, Rstudio, Matlab,

https://docs.cerit.io/

https://wiki.metacentrum.cz/wiki/Kubernetes_-_Rancher

https://metavo.metacentrum.cz/en/news/novinka_2022_0003.html

- `qextend` utility
  - Users are allowed to prolong their calculations on their own
  - The current limit is 2880 CPU hours, 30 days renewal period

```
qextend full_job_ID additional_walltime_hh:mm:ss
```

https://wiki.metacentrum.cz/wiki/Prolong_walltime

Take-home message

- There is no reason to be afraid to use MetaCentrum
- You can find plenty of information and instructions on our wiki

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

- If you are lost - send an email to us

  meta@cesnet.cz

- If grid infrastructure does not fulfil your expectations, maybe the MetaCentrum Cloud service would be a better choice

  https://cloud.metacentrum.cz/

cesnet
metacentrum

e-INFRA CZ

THANK YOU FOR YOUR ATTENTION

meta@cesnet.cz vorel@cesnet.cz