cesnet metacentrum

e-INFRA CZ

Grid service MetaCentrum

For scientific computations, collaborative research & its support services

Jiří Vorel, Roman Leontovyč

MetaCentrum User Support

vorel@cesnet.cz leontovyc@cesnet.cz meta@cesnet.cz

7. 2. 2023
Prague

What MetaCentrum is

- **MetaCentrum is**
  - … The National Grid Infrastructure (NGI)
  - ... the activity of the CESNET association
  - … a provider of computational **resources**, application **tools** (commercial and free/open source) and data **storage**
  - … free of charge
    - Users "pay" by Acknowledgement in their research publications
- **MetaCentrum is available for**
  - … 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)

https://metacentrum.cz

https://metavo.metacentrum.cz

https://wiki.metacentrum.cz

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

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

What MetaCentrum is

- **MetaCentrum is suitable for**
  - … individual users
  - ... projects (sharing data in a group)
  - ... institutions (we have too many resources)

https://metacentrum.cz

https://metavo.metacentrum.cz

https://wiki.metacentrum.cz

https://wiki.metacentrum.cz/wiki/Kategorie:Applications

- **MetaCentrum offers**
  - ... the principle of grid usage (privileged access for cluster owners)
  - … immediate access to HW resources
  - ... no need to submit projects
  - Mathematical/statistical software (Matlab, Mathematica, Maple, R), development tools (Intel, NVIDIA, GCC, AOCC, OpenJDK), material simulations (Ansys, OpenFOAM, Espresso)

What MetaCentrum is

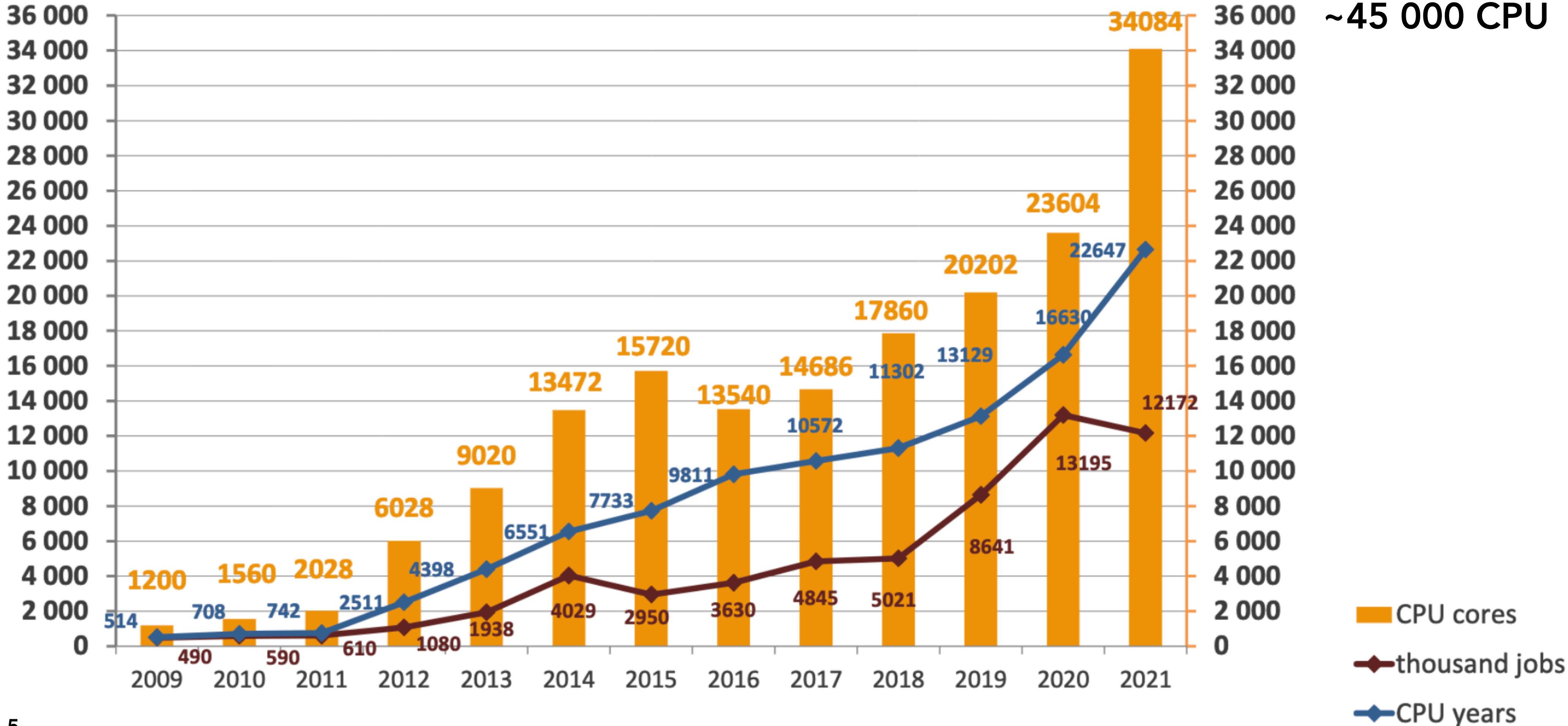
- **Examples of offered tools for bioinformatics**
  - Sequence data processing (Trimmomatic, FastQC, Bamtools, Bedtools)
  - Aligners (BWA, Bowtie, TopHat, MUMmer)
  - DNA/RNA assemblers (Velvet, MaSuRCA, SPAdes, SOAP, Flye, Trinity)
  - Annotation (Blast, InterProScan, RepeatExplorer, Maker, BUSCO)
  - Computational chemistry (Amber, DIRAC, MolPro, Gromacs, Orca)
  - Phylogenetics (BEAST, MrBayes, IQ-TREE, Kraken, RAxML, SNAPP)
  - Visualisation (Geneious, CLC-WB, PyMOL, Rstudio)

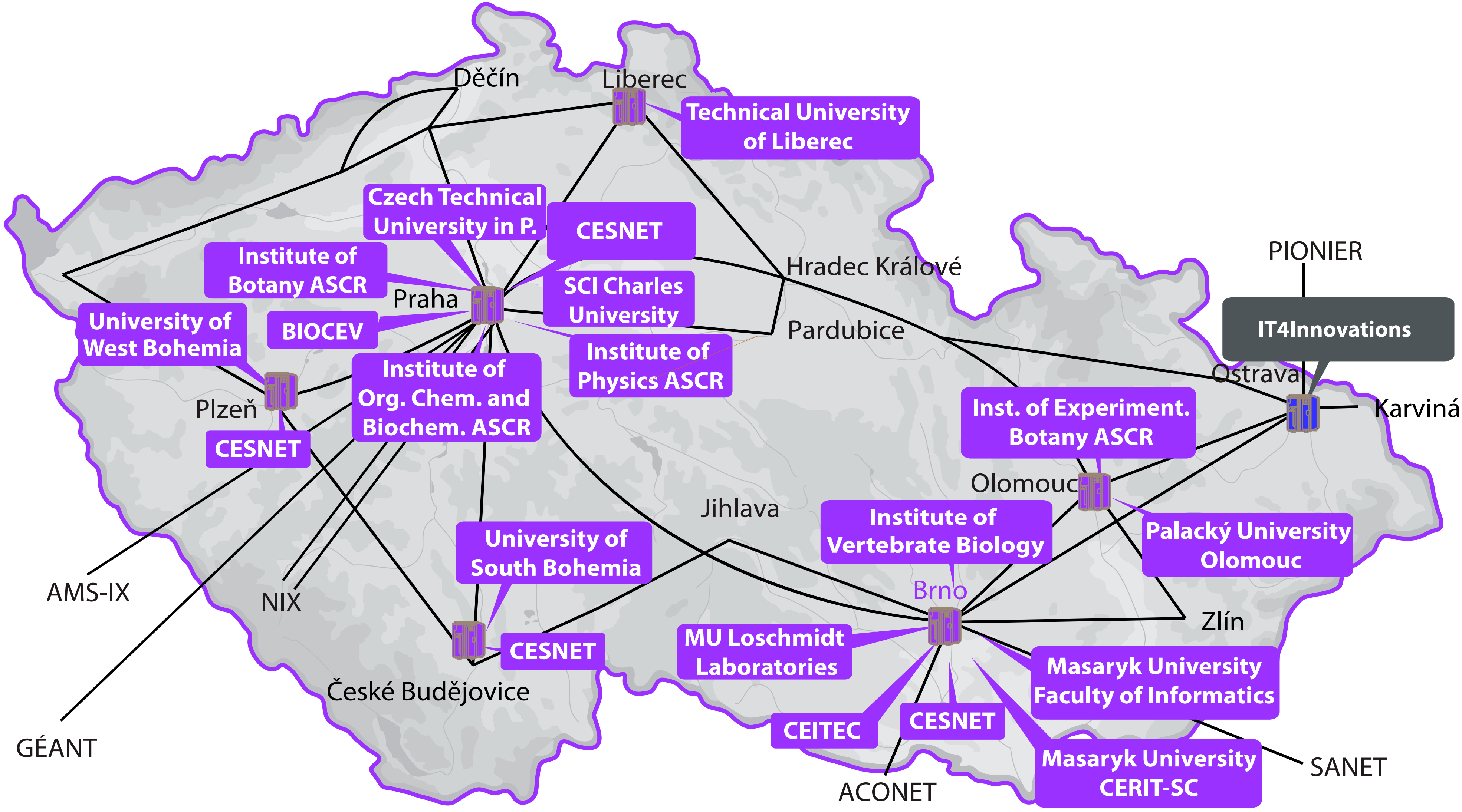
  - And much more…

Number of CPUs, executed jobs and corresponding CPU years

(Meta VO PBS)

2022:
~45 000 CPU




Děčín
Liberec
Technical University of Liberec
Czech Technical University in P.
CESNET
Institute of Botany ASCR
Praha
SCI Charles University
University of West Bohemia
BIOCEV
Institute of Physics ASCR
Hradec Králové
Pardubice
PIONIER
IT4Innovations
Ostrava
Karviná
Institute of Org. Chem. and Biochem. ASCR
Plzeň
CESNET
Inst. of Experiment. Botany ASCR
Olomouc
Jihlava
Institute of Vertebrate Biology
Palacký University Olomouc
University of South Bohemia
AMS-IX
NIX
Brno
Zlín
CESNET
MU Loschmidt Laboratories
Masaryk University Faculty of Informatics
České Budějovice
CEITEC
CESNET
GÉANT
Masaryk University CERIT-SC
SANET
ACONET

Typical HW characteristics

- **CPU**
  - ~45 000 CPU cores (x86_64) in total
  - Intel, AMD; Debian 11
  - Typically 32/64 CPU, up to 1 TB RAM (400-700 GB)
  - Special machines (up to 504 CPU, 10 TB RAM), CentOS
- **GPU**
  - 16 clusters, more than 400 GPU cards
  - NVIDIA A10, A40, A100, RTX A4000, Tesla *, GeForce *

https://wiki.metacentrum.cz/wiki/GPU_stroje

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

Frontend servers

- Gateway to the entire grid infrastructure
- Accessible via ssh with a password (ssh tickets are not fully supported)
- Frontends submit jobs to PBS servers
- Frontends are relatively small virtual machines mainly for purposes like writing scripts for batch jobs, checking applications and user data etc.

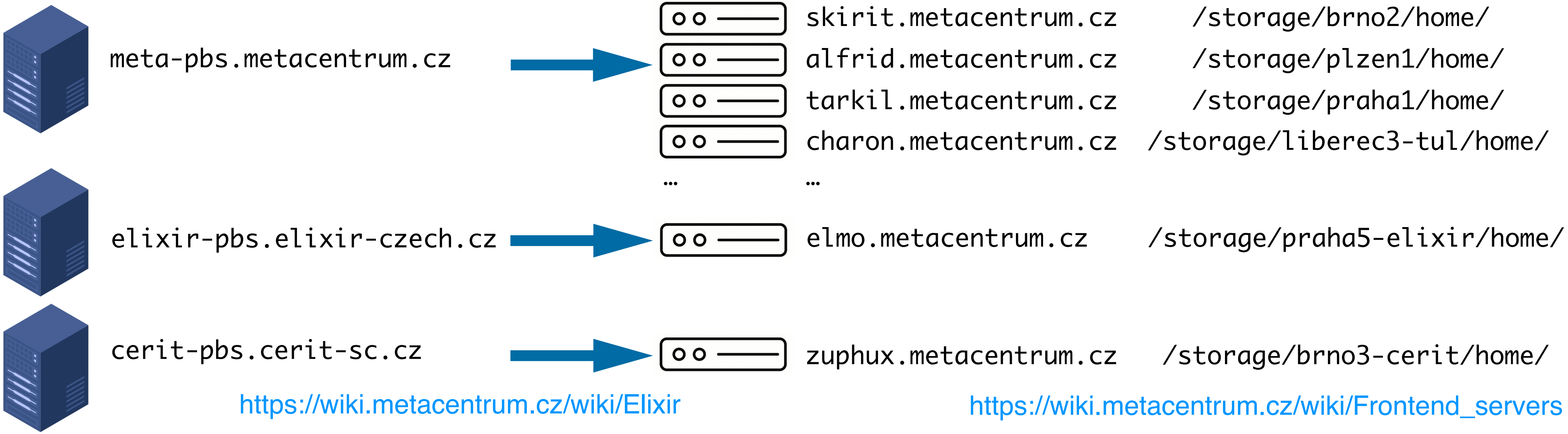
- **Do not run long and/or demanding calculations directly on frontends!**

- Frontend servers usually have different home directories
- Command line interface

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

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



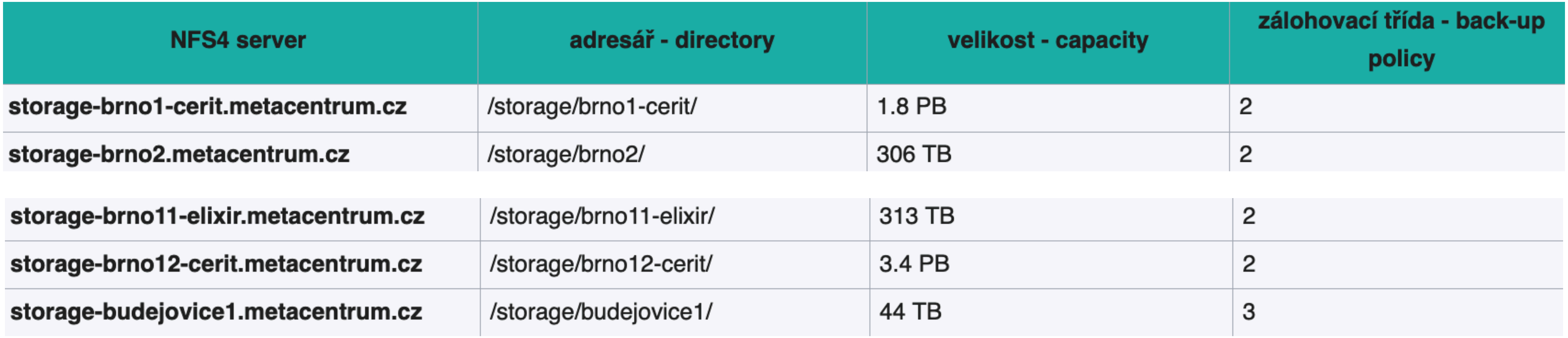
meta-pbs.metacentrum.cz	skirit.metacentrum.cz	/storage/brno2/home/
	alfrid.metacentrum.cz	/storage/plzen1/home/
	tarkil.metacentrum.cz	/storage/praha1/home/
	charon.metacentrum.cz	/storage/liberec3-tul/home/
	…	…
elixir-pbs.elixir-czech.cz	elmo.metacentrum.cz	/storage/praha5-elixir/home/
cerit-pbs.cerit-sc.cz	zuphux.metacentrum.cz	/storage/brno3-cerit/home/

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

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

NFS4 servers (storages)

- Data is stored on a few independent storages, the capacity is not infinite
- All storages are accessible through all frontends
- Storages have quotas for the total volume of data and the number of files

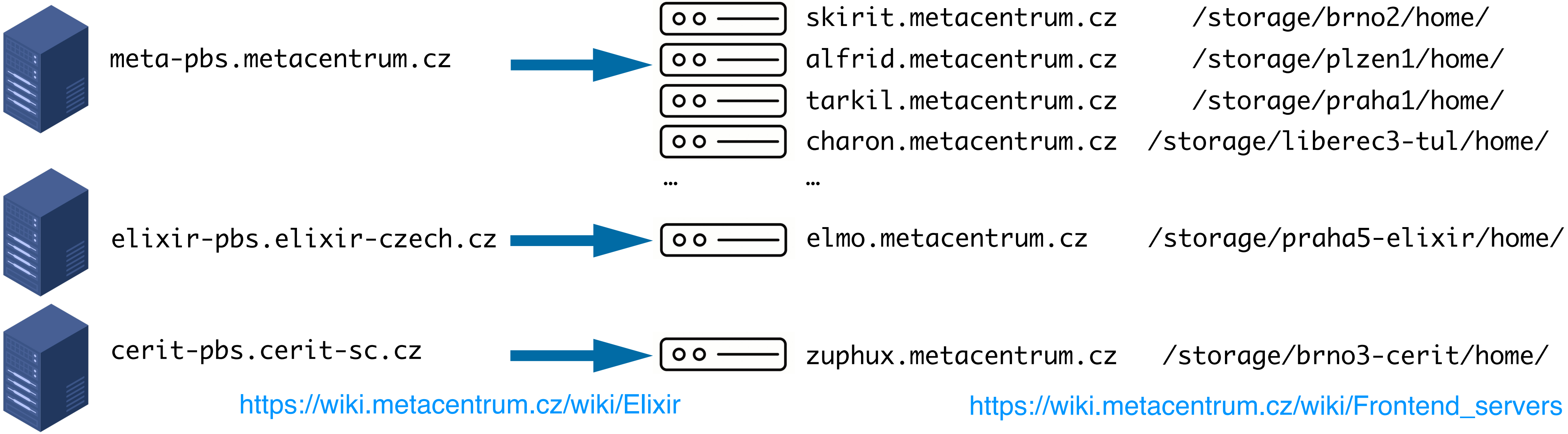


NFS4 server	adresář - directory	velikost - capacity	zálohovací třída - back-up policy
storage-brno1-cerit.metacentrum.cz	/storage/brno1-cerit/	1.8 PB	2
storage-brno2.metacentrum.cz	/storage/brno2/	306 TB	2
storage-brno11-elixir.metacentrum.cz	/storage/brno11-elixir/	313 TB	2
storage-brno12-cerit.metacentrum.cz	/storage/brno12-cerit/	3.4 PB	2
storage-budejovice1.metacentrum.cz	/storage/budejovice1/	44 TB	3

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

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



meta-pbs.metacentrum.cz

skirit.metacentrum.cz /storage/brno2/home/

alfrid.metacentrum.cz /storage/plzen1/home/

tarkil.metacentrum.cz /storage/praha1/home/

charon.metacentrum.cz /storage/liberec3-tul/home/

… …

elixir-pbs.elixir-czech.cz

elmo.metacentrum.cz /storage/praha5-elixir/home/

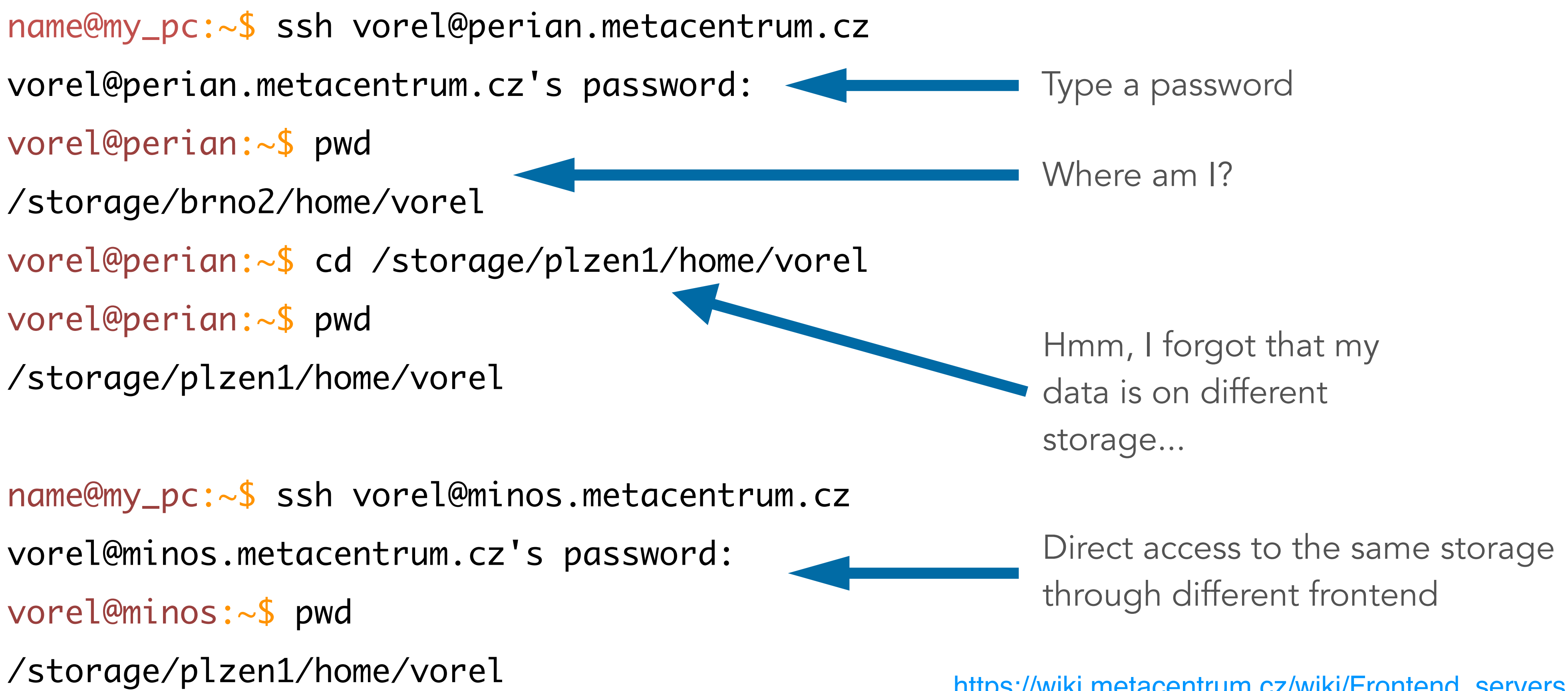
cerit-pbs.cerit-sc.cz

zuphux.metacentrum.cz /storage/brno3-cerit/home/

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?

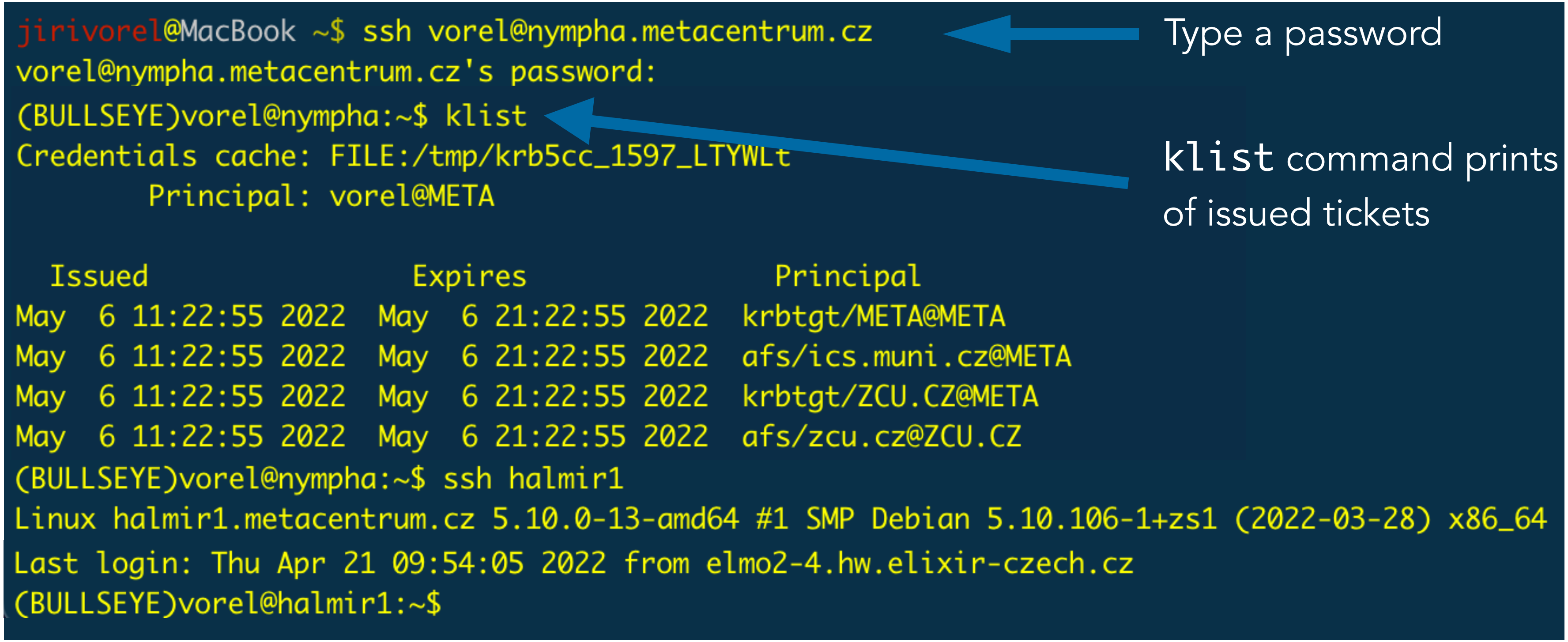
Hmm, I forgot that my data is on different storage…

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

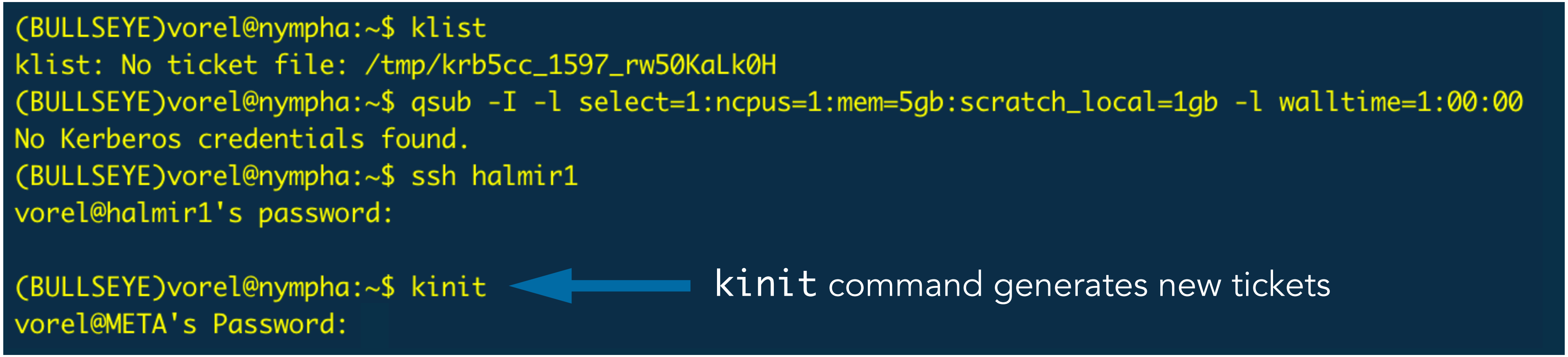


```
jirivorel@MacBook ~$ ssh vorel@nympha.metacentrum.cz
vorel@nympha.metacentrum.cz's password:
(BULLSEYE)vorel@nympha:~$ klist
Credentials cache: FILE:/tmp/krb5cc_1597_LTYWLt
        Principal: vorel@META

  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

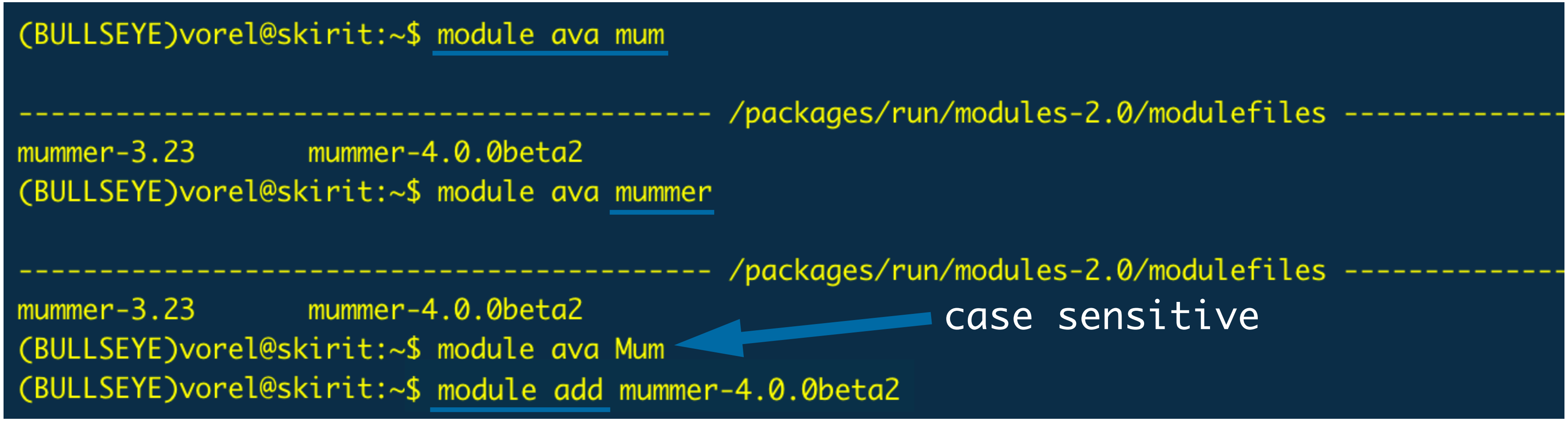
https://wiki.metacentrum.cz/wiki/Kerberos_authentication_system

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

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

Software modules

- Each software (in a specific version) is prepared as an individual module
- In theory, the module file, after activation (`module ava`), will load the main application, all dependencies and all needed libraries
- MetaCentrum contains ~3000 modules, **license agreement may be required**
- Available modules can be seen on our wiki or (much better) on the frontend

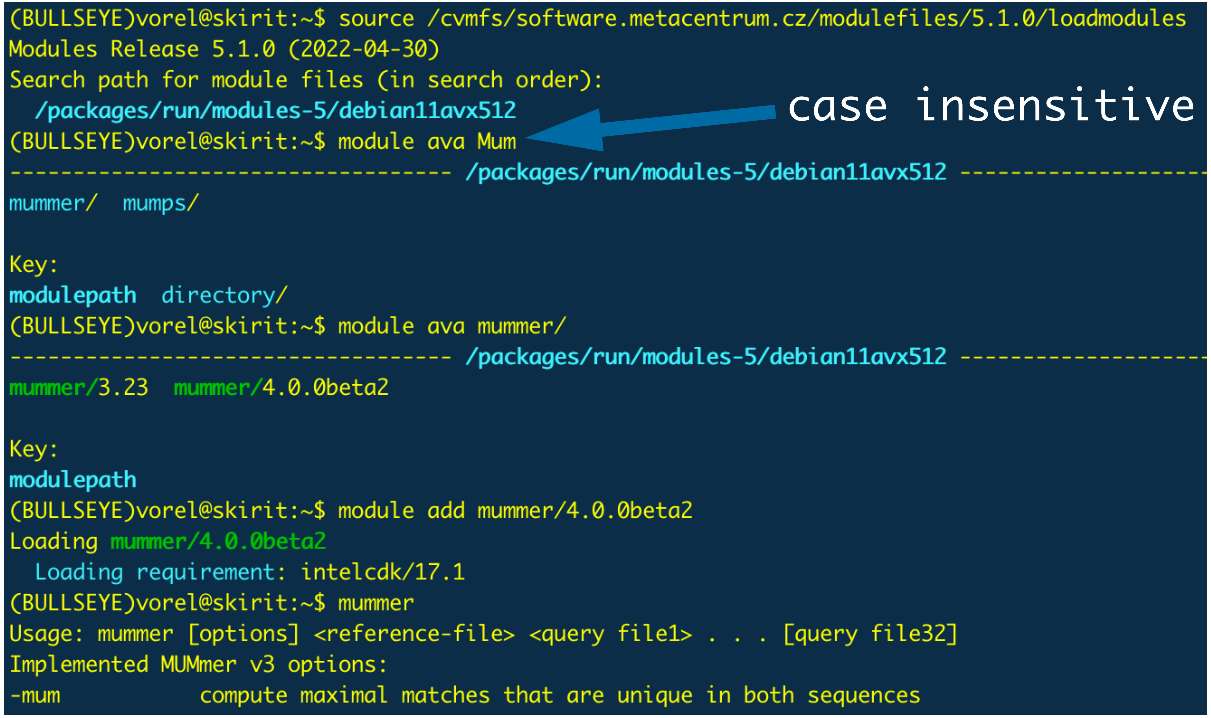


```
(BULLSEYE)vorel@skirit:~$ module ava mum

------------------------------------------- /packages/run/modules-2.0/modulefiles -------------
mummer-3.23       mummer-4.0.0beta2
(BULLSEYE)vorel@skirit:~$ module ava mummer

------------------------------------------- /packages/run/modules-2.0/modulefiles -------------
mummer-3.23       mummer-4.0.0beta2
(BULLSEYE)vorel@skirit:~$ module ava Mum
(BULLSEYE)vorel@skirit:~$ module add mummer-4.0.0beta2
```

case sensitive

Software modules

- A new version of modules will be released soon
- It can be tested now (on each frontend and machine)
- Initial activation is required (by default only on an aman cluster)



```
(BULLSEYE)vorel@skirit:~$ source /cvmfs/software.metacentrum.cz/modulefiles/5.1.0/loadmodules
Modules Release 5.1.0 (2022-04-30)
Search path for module files (in search order):
  /packages/run/modules-5/debian11avx512
(BULLSEYE)vorel@skirit:~$ module ava Mum
----------------------------------- /packages/run/modules-5/debian11avx512 --------------------
mummer/  mumps/

Key:
modulepath  directory/
(BULLSEYE)vorel@skirit:~$ module ava mummer/
----------------------------------- /packages/run/modules-5/debian11avx512 --------------------
mummer/3.23  mummer/4.0.0beta2

Key:
modulepath
(BULLSEYE)vorel@skirit:~$ module add mummer/4.0.0beta2
Loading mummer/4.0.0beta2
  Loading requirement: intelcdk/17.1
(BULLSEYE)vorel@skirit:~$ mummer
Usage: mummer [options] <reference-file> <query file1> . . . [query file32]
Implemented MUMmer v3 options:
-mum           compute maximal matches that are unique in both sequences
```

case insensitive

https://wiki.metacentrum.cz/wiki/Kategorie:Applications

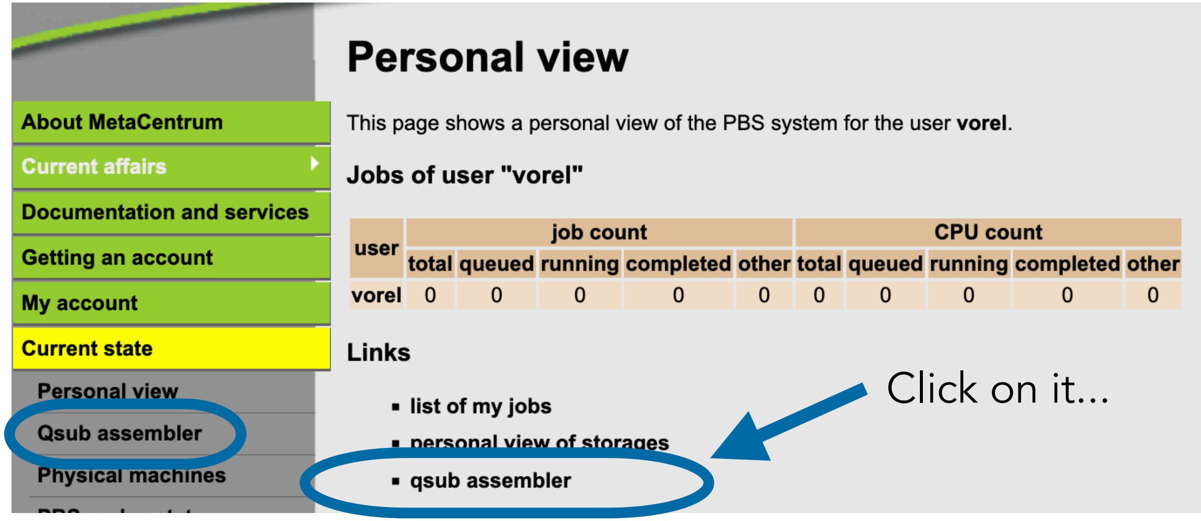
https://wiki.metacentrum.cz/wiki/Application_modules

https://wiki.metacentrum.cz/wiki/Application_modules_old

```
source /cvmfs/software.metacentrum.cz/modulefiles/5.1.0/loadmodules
```

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



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	0	0	0	0	0	0	0	0	0	0

Links

- list of my jobs
- personal view of storages
- qsub assembler

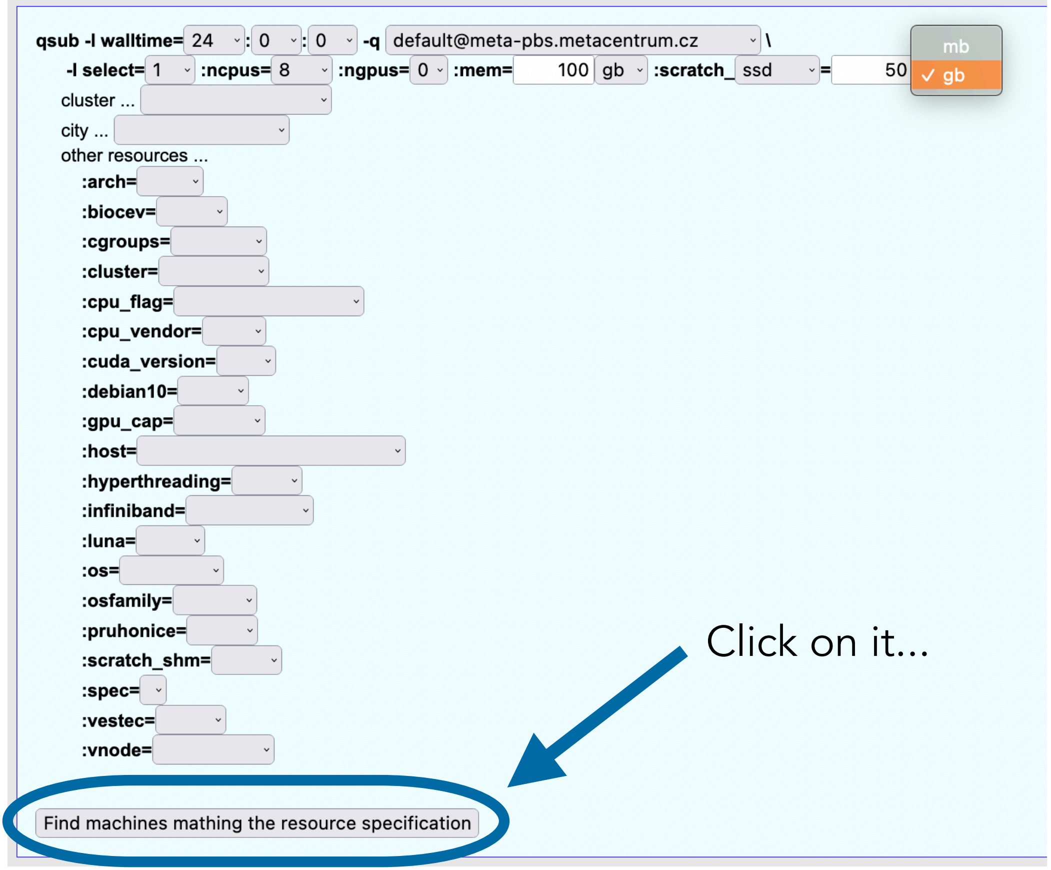
Click on it…

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

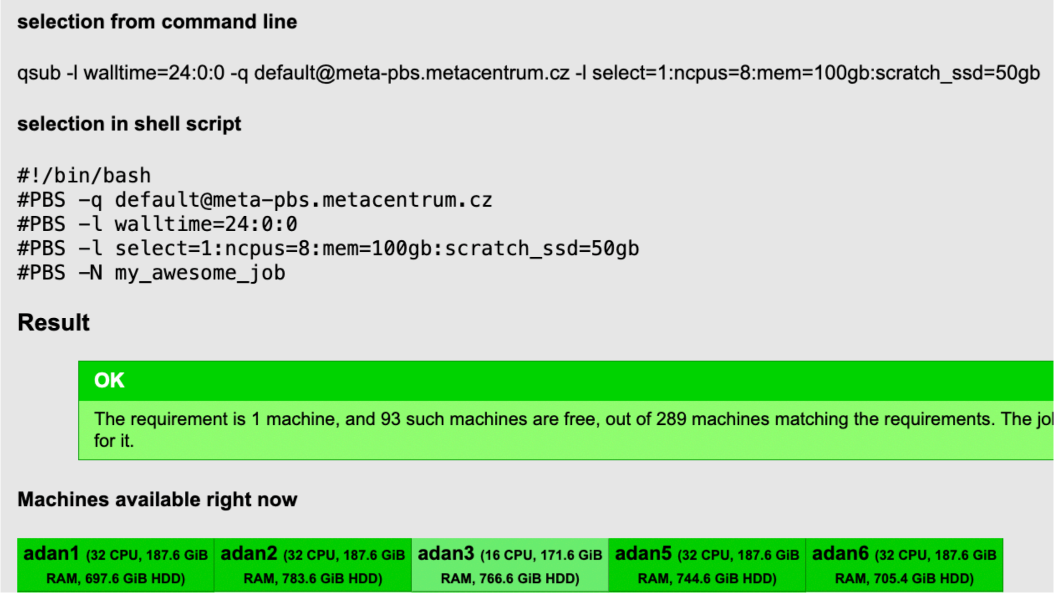
HW resources and qsub assembler



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

Click on it...

- 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

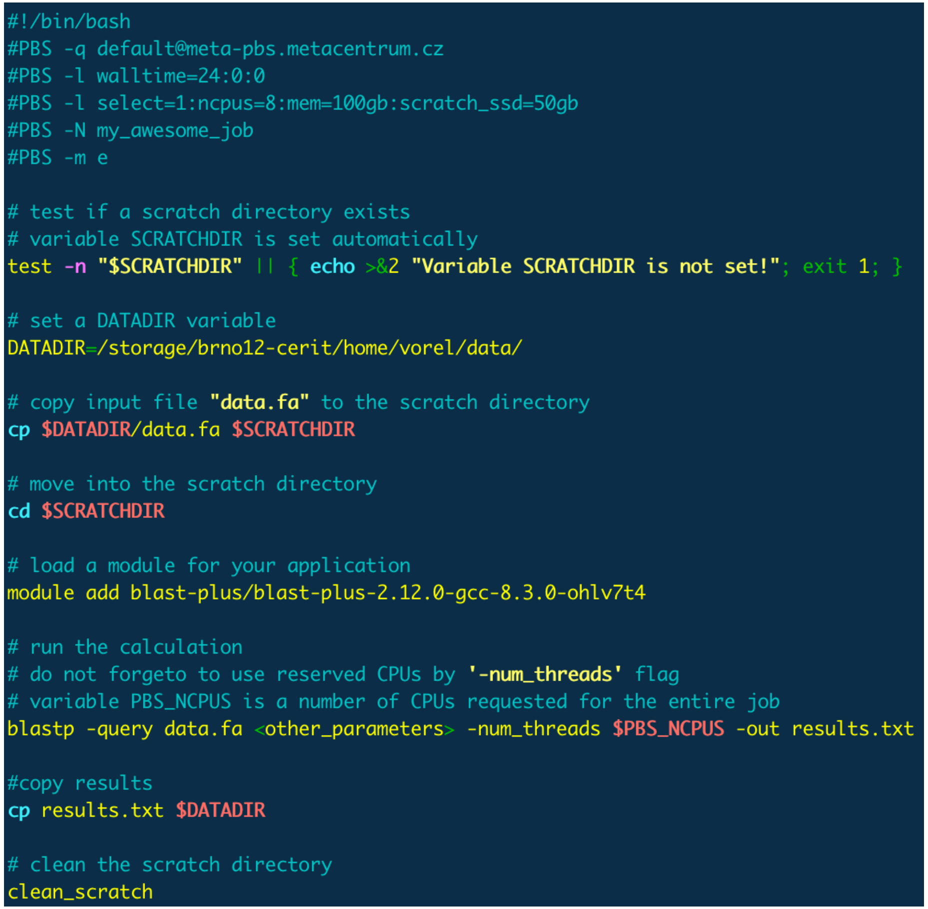
OK

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

Machines available right now

adan1 (32 CPU, 187.6 GiB RAM, 697.6 GiB HDD)	**adan2** (32 CPU, 187.6 GiB RAM, 783.6 GiB HDD)	**adan3** (16 CPU, 171.6 GiB RAM, 766.6 GiB HDD)	**adan5** (32 CPU, 187.6 GiB RAM, 744.6 GiB HDD)	**adan6** (32 CPU, 187.6 GiB RAM, 705.4 GiB HDD)

Example of a basic script for batch jobs



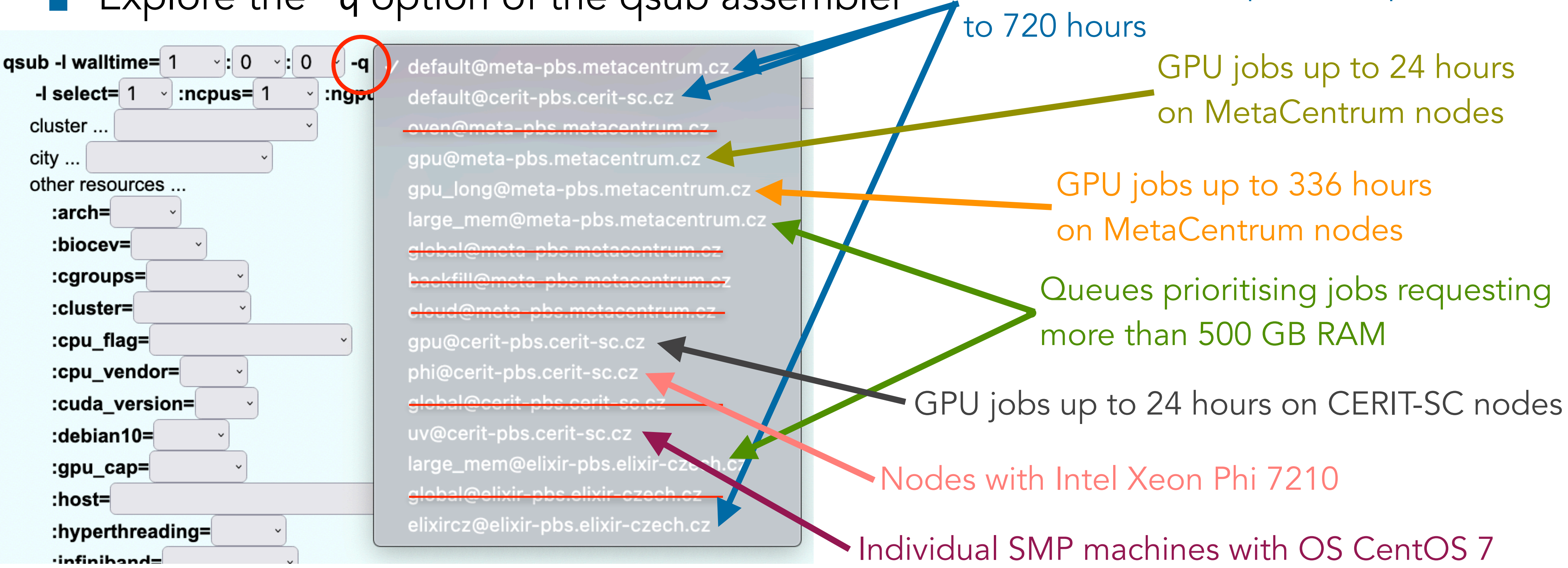
```
#!/bin/bash
#PBS -q default@meta-pbs.metacentrum.cz
#PBS -l walltime=24:0:0
#PBS -l select=1:ncpus=8:mem=100gb:scratch_ssd=50gb
#PBS -N my_awesome_job
#PBS -m e

# 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-ohlv7t4

# run the calculation
# do not forgeto 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



qsub -l walltime= 1 : 0 : 0 -q

-l select= 1 :ncpus= 1 :ngpu

cluster ...

city ...

other resources ...

:arch=

:biocev=

:cgroups=

:cluster=

:cpu_flag=

:cpu_vendor=

:cuda_version=

:debian10=

:gpu_cap=

:host=

:hyperthreading=

:infiniband=

✓ default@meta-pbs.metacentrum.cz

default@cerit-pbs.cerit-sc.cz

~~oven@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~~

~~cloud@meta-pbs.metacentrum.cz~~

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~~

elixircz@elixir-pbs.elixir-czech.cz

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

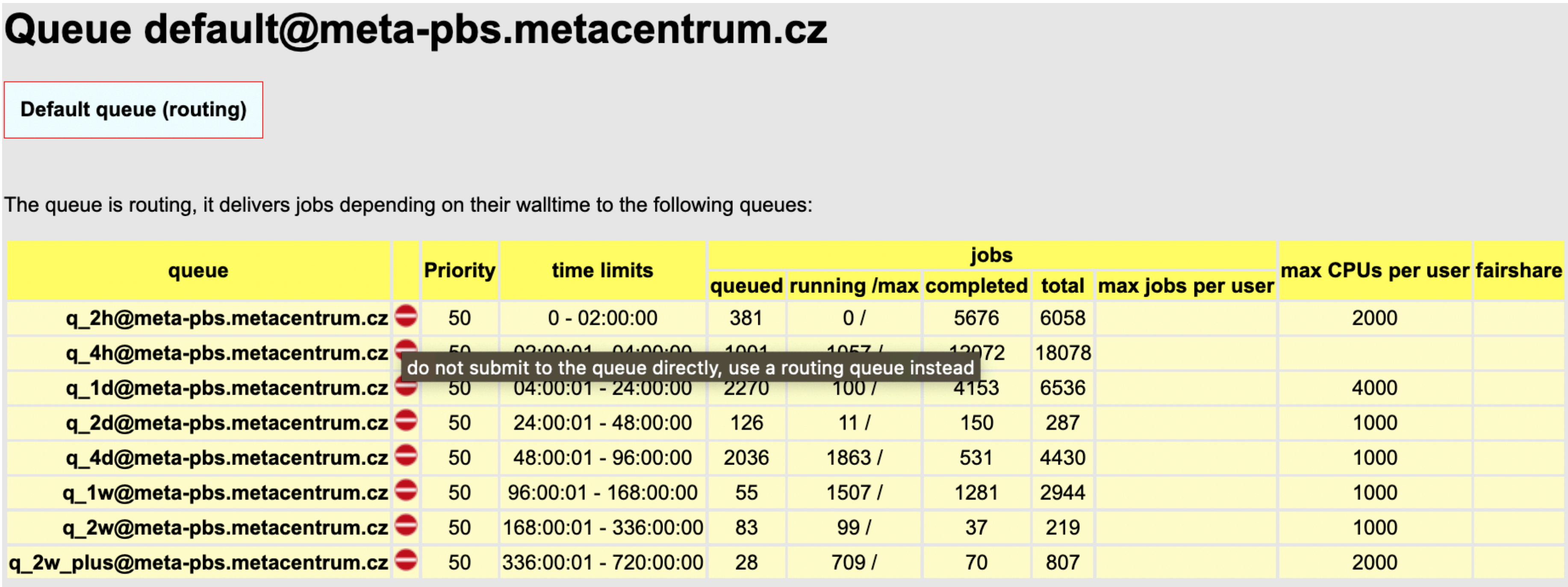
Queues prioritising jobs requesting more than 500 GB RAM

GPU jobs up to 24 hours on CERIT-SC nodes

Nodes with Intel Xeon Phi 7210

Individual SMP machines with OS CentOS 7

Queues



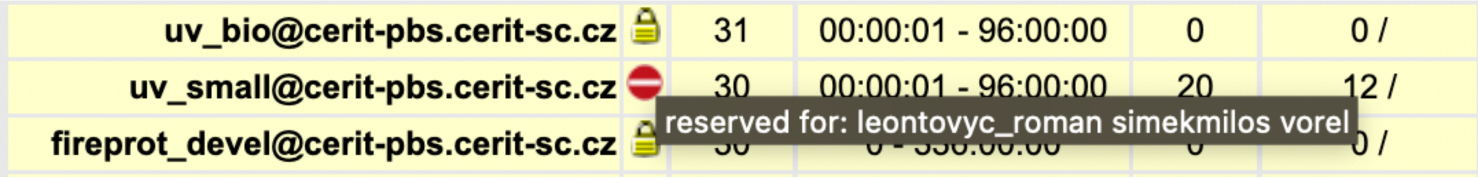
Queue default@meta-pbs.metacentrum.cz

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 /	5676	6058		2000	
q_4h@meta-pbs.metacentrum.cz	⛔	50	02:00:01 - 04:00:00	1091	1057 /	12072	18078			
q_1d@meta-pbs.metacentrum.cz	⛔	50	04:00:01 - 24:00:00	2270	100 /	4153	6536		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	

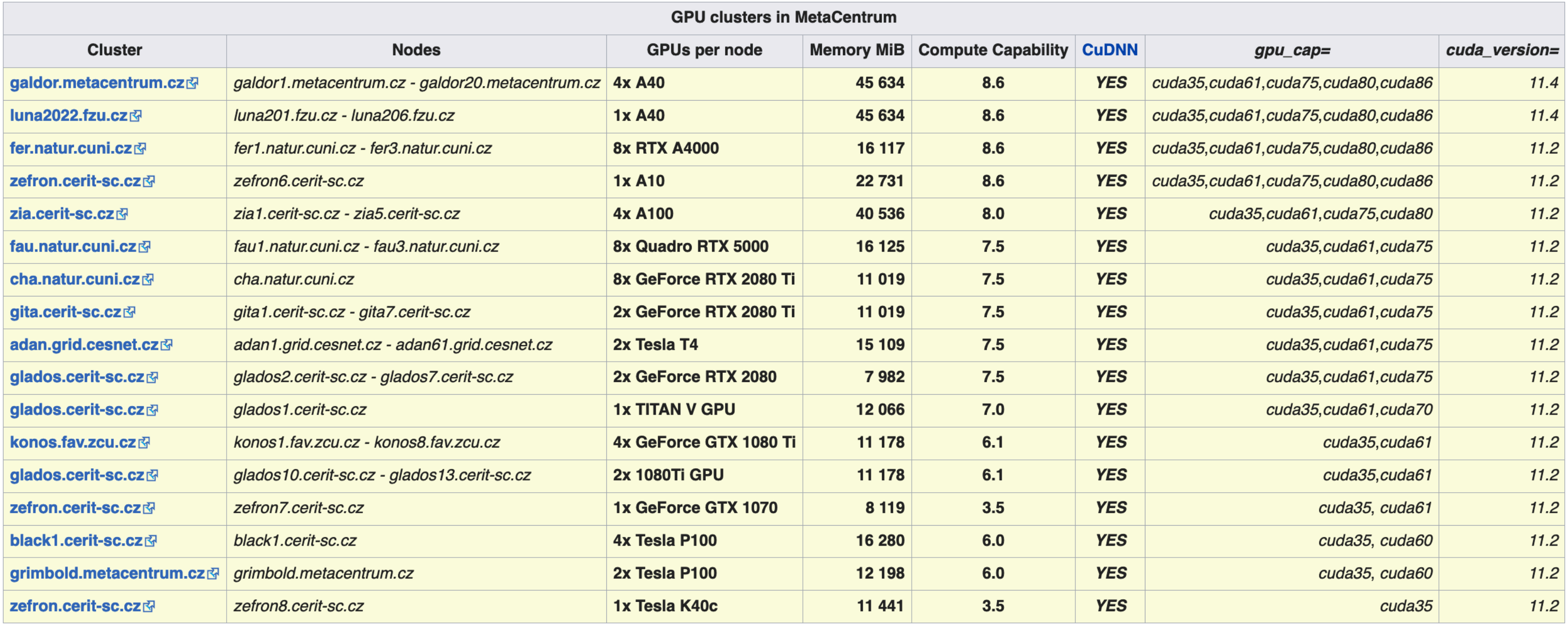
do not submit to the queue directly, use a routing queue instead



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	🔒	30	0 - 336:00:00	0	0 /

reserved for: leontovyc_roman simekmilos vorel

GPU acceleration



GPU clusters in MetaCentrum							
Cluster	**Nodes**	**GPUs per node**	**Memory MiB**	**Compute Capability**	**CuDNN**	***gpu_cap=***	***cuda_version=***
galdor.metacentrum.cz	*galdor1.metacentrum.cz - galdor20.metacentrum.cz*	4x A40	45 634	8.6	*YES*	*cuda35,cuda61,cuda75,cuda80,cuda86*	*11.4*
luna2022.fzu.cz	*luna201.fzu.cz - luna206.fzu.cz*	1x A40	45 634	8.6	*YES*	*cuda35,cuda61,cuda75,cuda80,cuda86*	*11.4*
fer.natur.cuni.cz	*fer1.natur.cuni.cz - fer3.natur.cuni.cz*	8x RTX A4000	16 117	8.6	*YES*	*cuda35,cuda61,cuda75,cuda80,cuda86*	*11.2*
zefron.cerit-sc.cz	*zefron6.cerit-sc.cz*	1x A10	22 731	8.6	*YES*	*cuda35,cuda61,cuda75,cuda80,cuda86*	*11.2*
zia.cerit-sc.cz	*zia1.cerit-sc.cz - zia5.cerit-sc.cz*	4x A100	40 536	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	16 125	7.5	*YES*	*cuda35,cuda61,cuda75*	*11.2*
cha.natur.cuni.cz	*cha.natur.cuni.cz*	8x GeForce RTX 2080 Ti	11 019	7.5	*YES*	*cuda35,cuda61,cuda75*	*11.2*
gita.cerit-sc.cz	*gita1.cerit-sc.cz - gita7.cerit-sc.cz*	2x GeForce RTX 2080 Ti	11 019	7.5	*YES*	*cuda35,cuda61,cuda75*	*11.2*
adan.grid.cesnet.cz	*adan1.grid.cesnet.cz - adan61.grid.cesnet.cz*	2x Tesla T4	15 109	7.5	*YES*	*cuda35,cuda61,cuda75*	*11.2*
glados.cerit-sc.cz	*glados2.cerit-sc.cz - glados7.cerit-sc.cz*	2x GeForce RTX 2080	7 982	7.5	*YES*	*cuda35,cuda61,cuda75*	*11.2*
glados.cerit-sc.cz	*glados1.cerit-sc.cz*	1x TITAN V GPU	12 066	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	11 178	6.1	*YES*	*cuda35,cuda61*	*11.2*
glados.cerit-sc.cz	*glados10.cerit-sc.cz - glados13.cerit-sc.cz*	2x 1080Ti GPU	11 178	6.1	*YES*	*cuda35,cuda61*	*11.2*
zefron.cerit-sc.cz	*zefron7.cerit-sc.cz*	1x GeForce GTX 1070	8 119	3.5	*YES*	*cuda35, cuda61*	*11.2*
black1.cerit-sc.cz	*black1.cerit-sc.cz*	4x Tesla P100	16 280	6.0	*YES*	*cuda35, cuda60*	*11.2*
grimbold.metacentrum.cz	*grimbold.metacentrum.cz*	2x Tesla P100	12 198	6.0	*YES*	*cuda35, cuda60*	*11.2*
zefron.cerit-sc.cz	*zefron8.cerit-sc.cz*	1x Tesla K40c	11 441	3.5	*YES*	*cuda35*	*11.2*

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

GPU acceleration

- **GPU**
  - 16 clusters, more than 400 GPU cards
  - Maximum si eight GPU cards on a single machine, typically two or four
  - Three dedicated GPU queues
    - `gpu@meta-pbs.metacentrum.cz` (up to 24 hours)
    - `gpu_long@meta-pbs.metacentrum.cz` (up to 336 hours)
    - `gpu@cerit-pbs.cerit-sc.cz` (up to 24 hours)
  - Jobs can migrate between PBS servers

```
qsub -l walltime=4:0:0 -q gpu@meta-pbs.metacentrum.cz -l \
select=1:ncpus=1:ngpus=1:mem=10gb:scratch_local=20gb
```

GPU acceleration

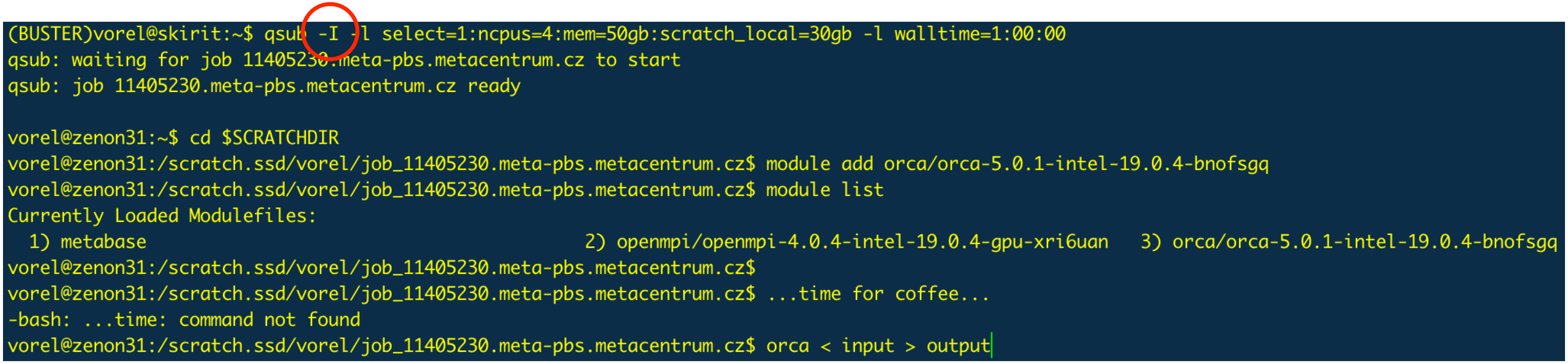
- Each GPU calculation (`ngpus=1`) needs at least one CPU (`ncpus=1`)
- Remember that the newest GPU is NOT the best for all jobs
- One GPU card per job is enough for novices
- GPU card can not be shared and is entirely dedicated to one calculation
- GPU calculations can be monitored on the same computation nodes by *nvidia-smi* or *nvtop* command
- In most cases is not wise to target one specific cluster (e.g. :cl_adan=True), select a smaller set of machines using the parameters:
  - `gpu_mem=20gb` (minimum amount of memory on card)
  - `gpu_cap=cuda80` (compute capability)
  - `cuda_version=11.4` (cuda version)

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

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

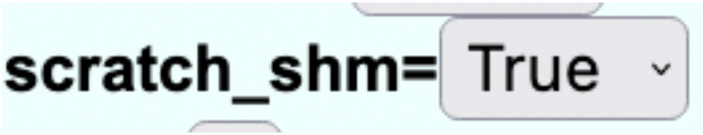
```
qsub -l select=1:ncpus=1:mem=4gb:scratch_local=10gb -l walltime=1:00:00

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


scratch_shm= True

Local installations and useful utilities

- Users can install the software in their own, do not violate the license
- 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

- `qextend` utility
  - Users are allowed to prolong their jobs in a limited number of cases

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

- `pbs-get-job-history` utility
  - Users can get complex information about their current or historical jobs

```
pbs-get-job-history job_ID
```

Backup and archiving

- 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

- 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/
```

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

How to protect home directory

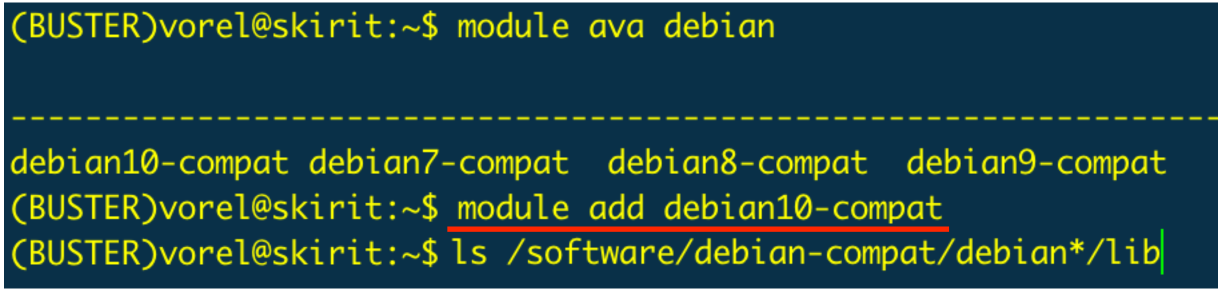
- Permission and access to the user's home directories are controlled by standard Unix rules (`chmod` command)
- **For safety reasons, only owners can write into their HOME directories** (max. is 755, an automatic script periodically corrects inappropriate settings, typically 777)
- **Default rights are not so strict:**
  - Content can be read by other MetaCentrum users
  - Subdirectories are not write protected
- If you want to hide your data, you can set very strict rules on the home directory (e.g. 700), and only you will be able to read and use the content

Debian11 is here

- As always, we keep Debian up-to-date on our nodes
- Now we are on Deb11 (BULLSEYE)
- However, some libraries may be missing in 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…

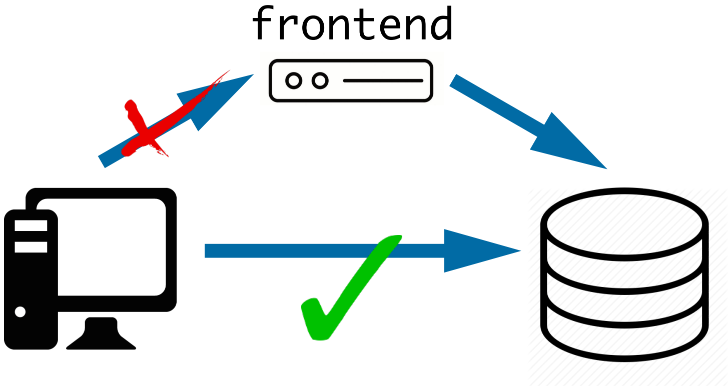
cesnet metacentrum

e-INFRA CZ

Common issues and how to prevent 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

Avoid non-effective calculations

- Optimise your calculations (hardware usage)
- Reservation of too many resources decreases 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

Use the scratch directory

- 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

Clean the scratch directory

- Do not forget to clean the scratch directory when your calculation is done or has 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
```

Do not run long calculations on frontends

- 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
```

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

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

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

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 60 minutes
- Startup overhead may be a significant part of the whole processing time
- Aggregate short jobs into bigger groups with longer walltime

```
-l walltime=01:00:00 (and more)
```

metacentrum cesnet

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

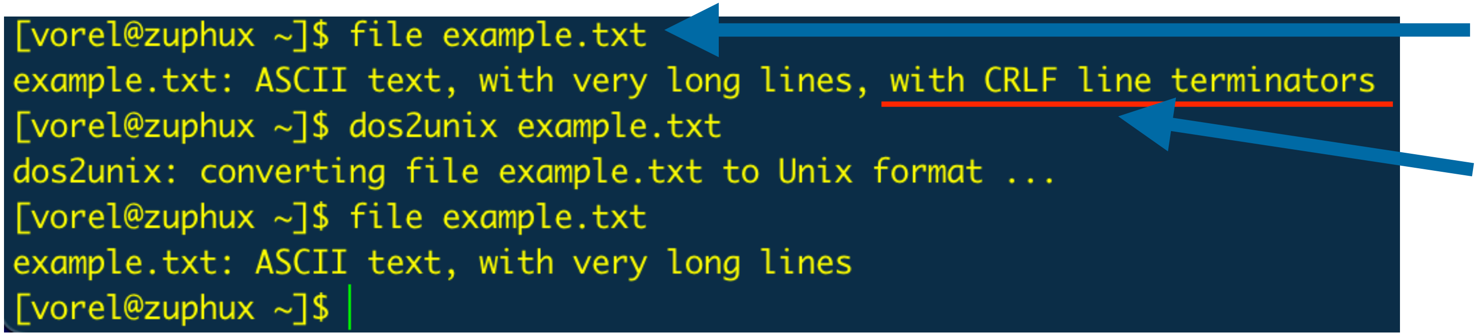
```
export TMPDIR=$SCRATCHDIR

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 MS Win. use more characters for the 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



```
[vorel@zuphux ~]$ file example.txt
example.txt: ASCII text, with very long lines, with CRLF line terminators
[vorel@zuphux ~]$ dos2unix example.txt
dos2unix: converting file example.txt to Unix format ...
[vorel@zuphux ~]$ file example.txt
example.txt: ASCII text, with very long lines
[vorel@zuphux ~]$
```

Toole `file` will determine the type of a file

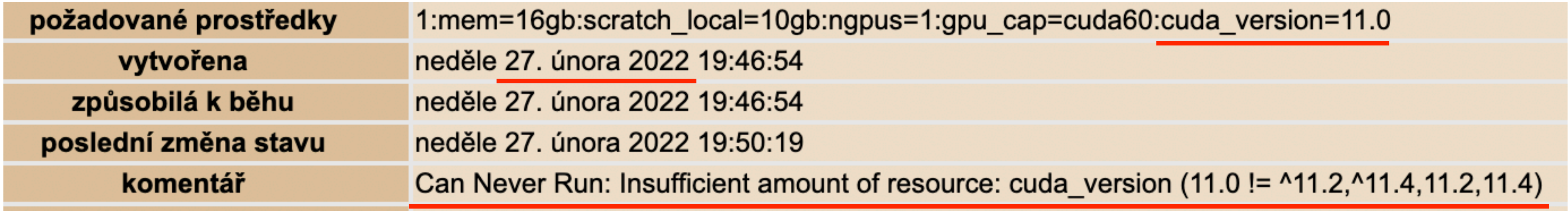
Problem detected

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...

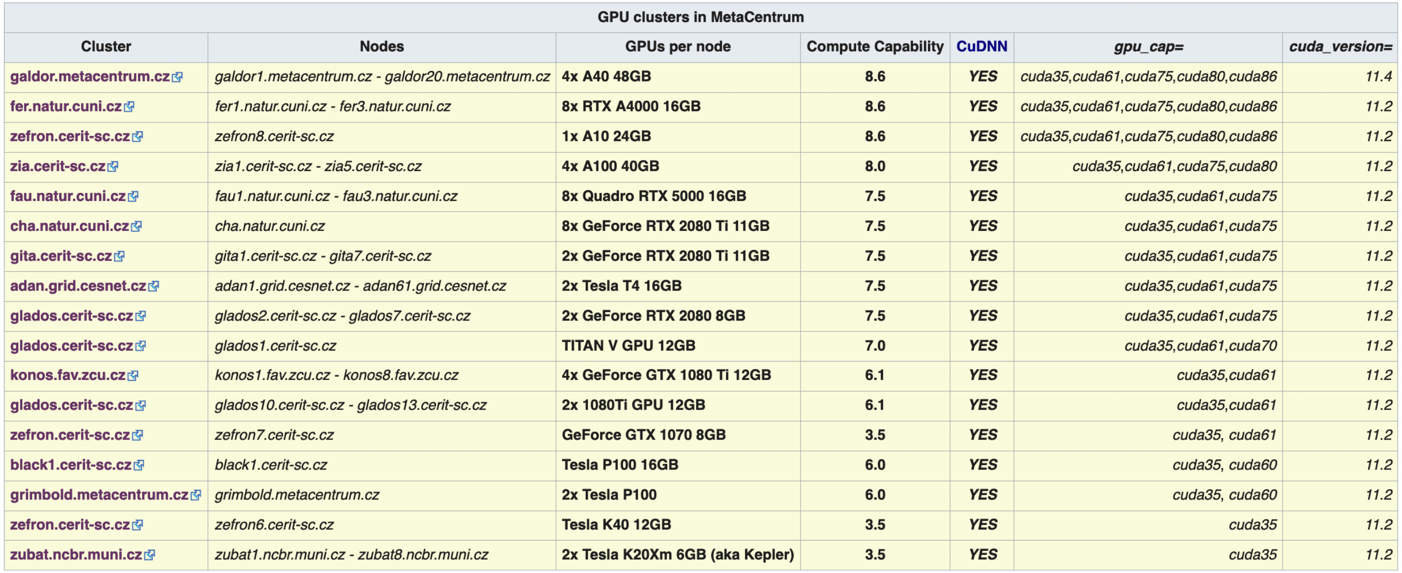
- Typical scenarios:
  - Incompatible Cuda versions and GPU machines
  - Wrong combinations of machines and queues
  - Combinations of parameters targeting a disparate set of machines

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



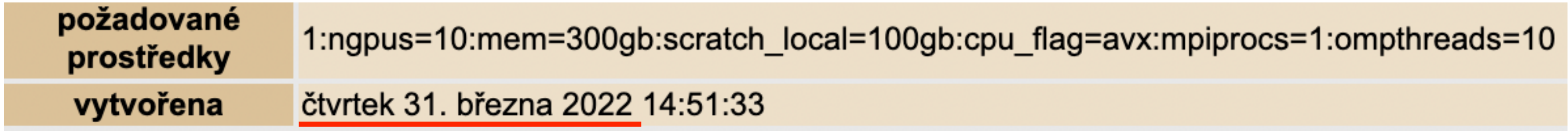
GPU clusters in MetaCentrum

Cluster	Nodes	GPUs per node	Compute Capability	CuDNN	*gpu_cap=*	*cuda_version=*
galdor.metacentrum.cz	*galdor1.metacentrum.cz - galdor20.metacentrum.cz*	4x A40 48GB	8.6	*YES*	*cuda35,cuda61,cuda75,cuda80,cuda86*	*11.4*
fer.natur.cuni.cz	*fer1.natur.cuni.cz - fer3.natur.cuni.cz*	8x RTX A4000 16GB	8.6	*YES*	*cuda35,cuda61,cuda75,cuda80,cuda86*	*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*
gita.cerit-sc.cz	*gita1.cerit-sc.cz - gita7.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 1080Ti 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*
grimbold.metacentrum.cz	*grimbold.metacentrum.cz*	2x Tesla P100	6.0	*YES*	*cuda35, cuda60*	*11.2*
zefron.cerit-sc.cz	*zefron6.cerit-sc.cz*	Tesla K40 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

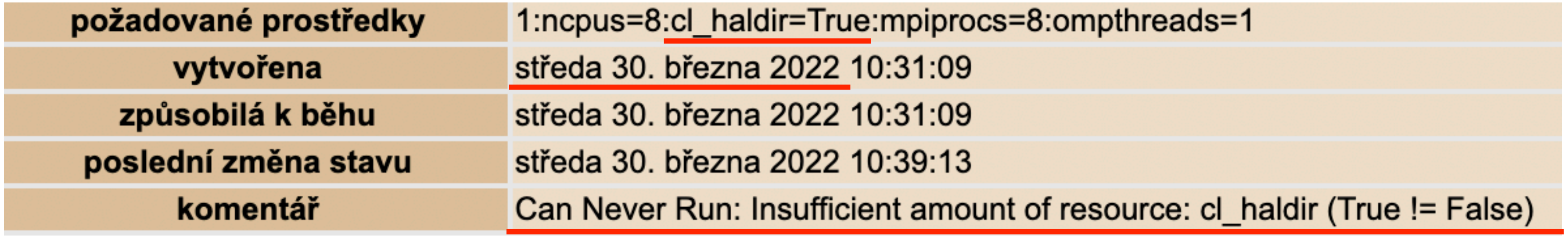
cesnet metacentrum

Incorrectly submitted jobs



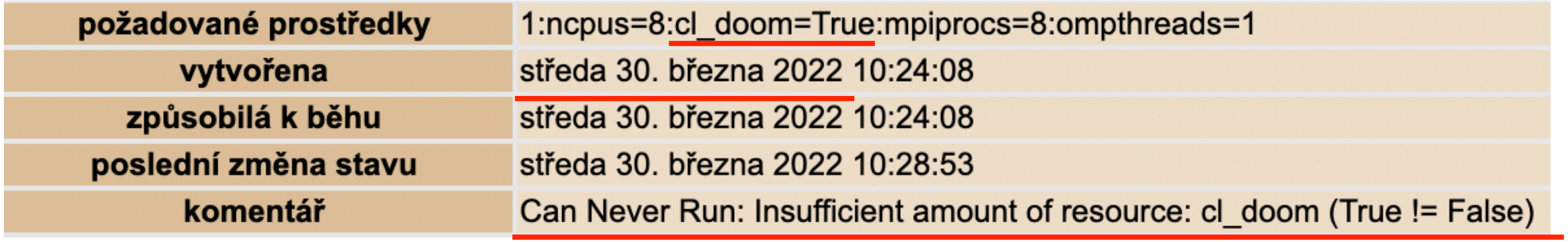
požadované prostředky	1:ngpus=10:mem=300gb:scratch_local=100gb:cpu_flag=avx:mpiprocs=1:ompthreads=10
vytvořena	čtvrtek 31. března 2022 14:51:33

Probably just a typo; ngpus can be max. 8; 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 in 2021. 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)

cesnet metacentrum

e-INFRA CZ

Other tools and services, graphical applications

Singularity containers

- Singularity (Apptainer) is an alternative to Docker
- Container system for HPC
- A container is a standard unit of software that packages up code and all its dependencies so the application runs quickly and reliably from one computing environment to another

https://wiki.metacentrum.cz/wiki/Singularity

- Saves time, prevents conflicts between applications
- Every Docker container can be converted to a Singularity image and used in MetaCentrum
- As pre-prepared Singularity images, users can use (e.g.) OpenFOAM, TE-TOOLS (RepeatMasker, RepeatModeler), Peregine (assembler for long reads)

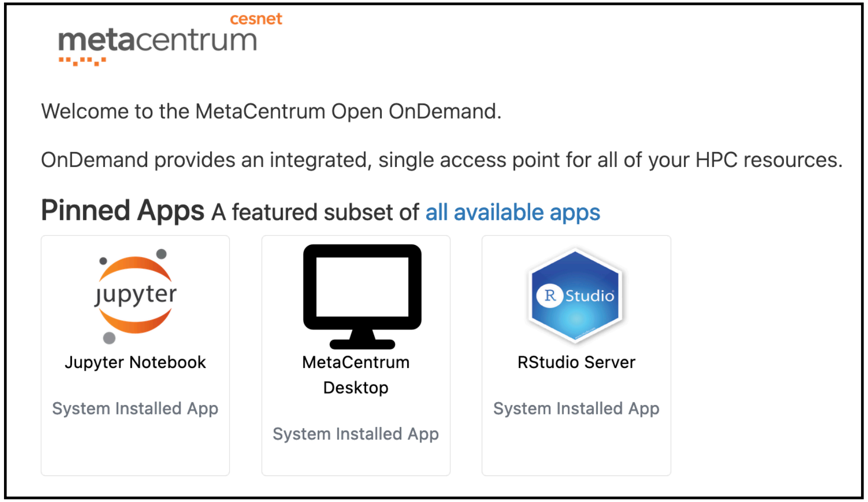
Singularity containers

- Widely used is NGC (NVIDIA GPU Cloud) package
- GPU-tuned frameworks for deep learning packed as containers, including NAMD3, Kaldi, OpenCV, PyTorch, qEspresso, **TensorFlow (22.12)**, **PyTorch (22.12)**

OnDemand

- Remote web access to supercomputers
- Currently under development, will be released soon
- GPU support will be included
- Rstudio, Jupyter notebooks, Matlab, Ansys
- MetaCentrum desktop
- Should be possible to add more tools

https://wiki.metacentrum.cz/wiki/Singularity



Welcome to the MetaCentrum Open OnDemand.

OnDemand provides an integrated, single access point for all of your HPC resources.

Pinned Apps A featured subset of all available apps

Kubernetes and Snakemake

- **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

kubernetes

RANCHER

https://docs.cerit.io/

https://wiki.metacentrum.cz/wiki/Kubernetes_-_Rancher https://metavo.metacentrum.cz/en/news/novinka_2022_0003.html

- **Snakemake**
  - Workflow management system is a tool to create reproducible and scalable data analyses (python based)

https://snakemake.readthedocs.io/en/stable/

OwnCloud and FileSender

- **OwnCloud**
  - Cloud storage with space of 100 GB per user (possible to increase)
  - User clients for Windows, Linux, OS X, iOS, Android operating systems
  - Automatic data synchronisation between several devices

https://www.cesnet.cz/sluzby/

- **FileSender**
  - Web service for sending files
  - Download link is sent to the other side, the file is stored for a maximum of one month (then is automatically deleted)
  - Connection with MetaCentre is possible

Final notes

- 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 leontovyc@cesnet.cz