Getting Started on DiRAC systems

Ed Bennett

DiRAC Federation Workshop, 2023-06-12

Slides: https://github.com/edbennett/dirac-202306

Assumptions

- You are familiar with the Unix (bash) shell
- You have used an HPC system and batch scheduler before
- You have registered with SAFE and provided an SSH key

Connecting

- Tursa:
  ssh username@tursa.dirac.ed.ac.uk
  ed25519: SHA256:QFXBZzU5vChePHu/Y/FF42Xac7w2Shb/XT4G2+vTM48
- COSMA8:
  ssh username@login8.cosma.dur.ac.uk
  ed25519: SHA256:t50+QQcNZ6QAVCfD8n4fr4gfTT22TInc2xHke8g8ZhQ
- “Two-factor” authentication: SSH key *and* login password required
  - Passphrase-protect your keys!

Login environment

- Tursa: 1TB RAM, 128 cores
- COSMA8: 2TB RAM, 64 cores
- Only for compilation/installation, not for compute-intensive/production jobs

Software environment

- Software provided via Environment Modules
- To check available modules:
  `module avail`
- To load a module:
  `module load` *`module1_name module2_name`* **…**
- Tursa: barebones stack—recommended modules:
  `cuda/11.4.1 openmpi/4.1.1-cuda11.4.1 ucx/1.12.0-cuda11.4.1`
- COSMA8 stack more full-featured

Filesystems

Tursa

- Your home directory:
  `/home/project_core/project_code/username`
- To share within your project:
  `/home/project_code/project_code/shared`
- To share with other projects:
  `/home/project_code/shared`

Filesystems
COSMA8

- Your home directory: NFS mount, slow, small quota
  `/cosma/home/`*`project`*`/`*`username`*
- Fast Lustre storage, larger quota:
  `/cosma8/data/`*`project`*`/`*`username`*
- Super-fast scratch storage for checkpointing:
  `/snap8/scratch/`*`project`*`/`*`username`*

Filesystems

Striping

- Ensure that large files (≳1TB) are striped
  - Otherwise they can take out the filesystem
  - Unless you have a one-file-per-node write pattern
- Talk to your RSEs about how to do this

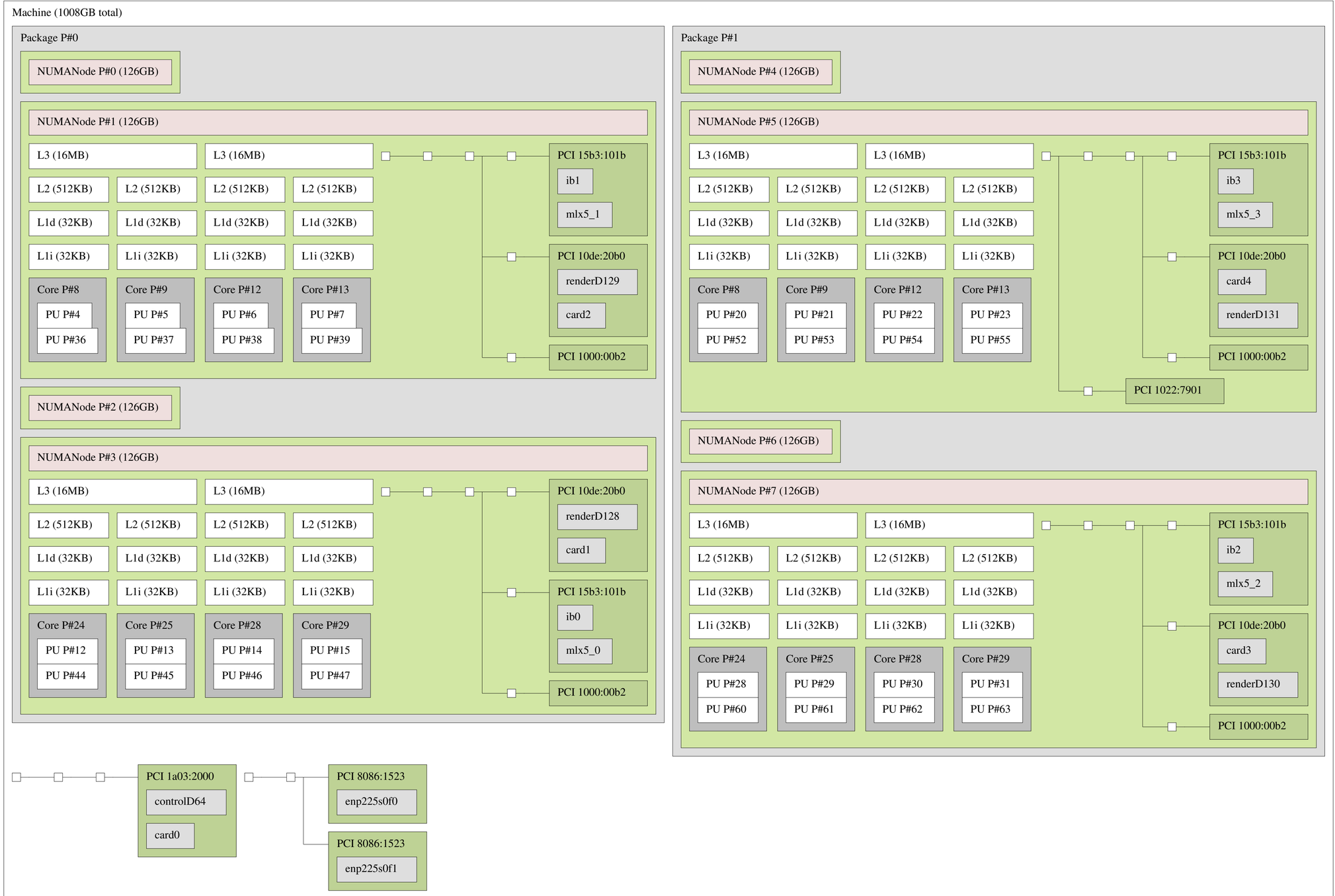
Compute nodes

Tursa

- 2 × AMD EPYC 7302
  - Each (16 cores/4 NUMA nodes/8 chiplets) × 2 hardware threads
- 4 × NVIDIA A100 [40/80GB] GPUs
- 4 × NVIDIA Networking ConnectX-6 Infiniband HDR adapters
  - Both on NUMA nodes 1, 3, 5, 7
- 1TiB RAM (128GiB per NUMA node)

Compute nodes

Tursa



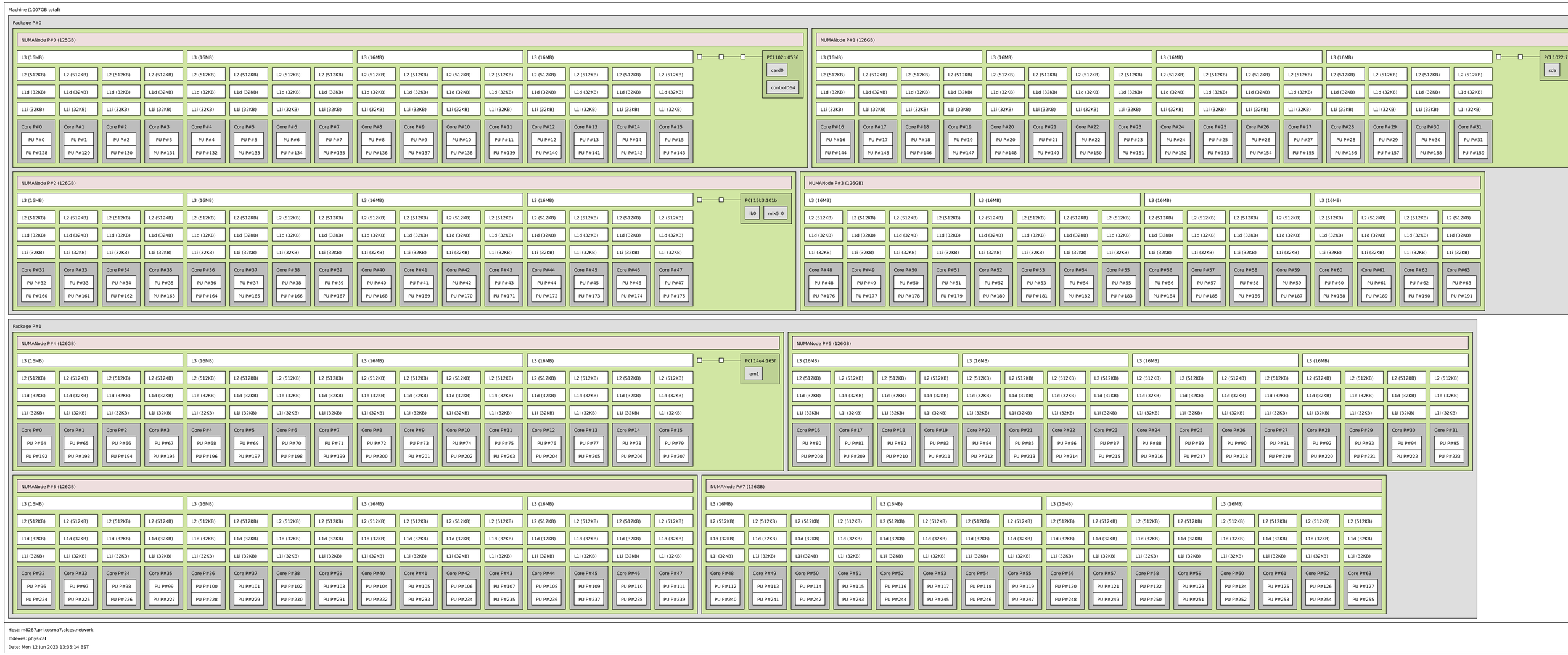
Host: tu−c0r0n45
Indexes: physical
Date: Mon 12 Jun 2023 13:19:30 BST

Compute nodes
COSMA8

- 2 × AMD EPYC 7H12
  - Each (64 cores/4 NUMA nodes/8 chiplets) × 2 threads
- 1 TiB RAM (128 GiB per NUMA node)
- 1 × NVIDIA Networking ConnectX-6 Infiniband HDR adapters

Compute nodes

COSMA8



Slurm

`sinfo`	Get overall partition status
`squeue --me`	Get status of my jobs in queue
`sbatch` *`job_script_filename`*	Submit a job to the queue
`scancel` *`job_id`*	Cancel a running or pending job

Example job script

Tursa

```
#!/bin/bash
#SBATCH --job-name NAME
#SBATCH --qos standard
#SBATCH --time 10:00
#SBATCH --account PROJECT
#SBATCH --nodes NUM_NODES
#SBATCH --ntasks NUM_TASKS
#SBATCH --ntasks-per-node 4
#SBATCH --cpus-per-task 8
#SBATCH --partition gpu
#SBATCH --gres gpu:4
#SBATCH --output %x.%j.out
#SBATCH --error %x.%j.err
#SBATCH --reservation workshop
umask 0002
```

```
module purge
module load cuda/11.4.1 \
    openmpi/4.1.1-cuda11.4.1 \
    ucx/1.12.0-cuda11.4.1

export OMP_NUM_THREADS=4
export OMPI_MCA_btl=^uct,openib
export UCX_TLS=gdr_copy,rc,rc_x,sm,cuda_copy,cuda_ipc
export UCX_RNDV_SCHEME=put_zcopy
export UCX_RNDV_THRESH=16384
export UCX_IB_GPU_DIRECT_RDMA=yes
export UCX_MEMTYPE_CACHE=n

mpirun -np ${SLURM_NTASKS} \
       --bind-to none \
       ./wrapper.sh ./my_tool
```

Wrapper script

Tursa

- Bind processes to GPUs and fabric adapters with direct connection

```
#!/bin/bash

lrank=$OMPI_COMM_WORLD_LOCAL_RANK
numa1=$(( 2 * $lrank))
numa2=$(( 2 * $lrank + 1 ))
netdev=mlx5_${lrank}:1

export CUDA_VISIBLE_DEVICES=$OMPI_COMM_WORLD_LOCAL_RANK
export UCX_NET_DEVICES=mlx5_${lrank}:1
BINDING="--interleave=$numa1,$numa2"

echo "`hostname` - $lrank device=$CUDA_VISIBLE_DEVICES binding=$BINDING"

numactl ${BINDING}  $*
```

Example job script

COSMA8

```
#!/bin/bash
#SBATCH --job-name NAME
#SBATCH --account PROJECT
#SBATCH --partition cosma8
#SBATCH --nodes NUM_NODES
#SBATCH --ntasks NUM_TASKS
#SBATCH --cpus-per-task 8
#SBATCH --time 10:00
#SBATCH --exclusive
#SBATCH --reservation onboarding

module purge
module load ucx/1.13.0rc2 oneAPI/2022.3.0

mpirun ./my_tool
```

Project codes

dp287	CompBioMed
dp288	HECBioSim
dp289	MRCGlasgow
dp290	UKAEA
dp291	Materials Science/PAX
dp292	BritLLM

Other thoughts

- We are all sharing a reservation
  - 16 nodes on COSMA8, 32 nodes on Tursa
  - Think before running long large jobs
- Tools like Miniconda, NVIDIA HPC SDK can be installed in your user space

More details

- Tursa user guide: https://epcced.github.io/dirac-docs/tursa-user-guide
- COSMA8 user guide: https://www.dur.ac.uk/icc/cosma/support/cosma8/