VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

SLURM

MAPPING AND PINNING

Ondřej Meca

IT4Innovations

SLURM

- Slurm workload manager
  - jobs scheduler for HPC clusters
  - used by all IT4I systems, LUMI, …
  - https://slurm.schedmd.com/
  - https://docs.it4i.cz/general/slurm-job-submission-and-execution/
  - https://docs.lumi-supercomputer.eu/runjobs/scheduled-jobs/slurm-quickstart/

SLURM

- `sbatch script.sh`
- `sbatch –-nodes 2 script.sh`

```
#!/usr/bin/bash
#SBATCH --job-name MyJobName
#SBATCH --account PROJECT-ID
#SBATCH --partition qcpu
#SBATCH --nodes 4
#SBATCH --ntasks-per-node 128
#SBATCH --time 12:00:00

ml purge
ml OpenMPI/4.1.4-GCC-11.3.0

srun hostname | sort | uniq -c
```

SLURM

Interactive jobs:

- `salloc -A PROJECT-ID -p qcpu -N 4 --ntasks-per-node 128 -t 2:00:00`

Start the job:

- `srun –n 128 ./app`

Get info about queued jobs:

- `squeue –-me`

Job canceling:

- `scancel JOBID`

Informations about nodes and partitions:

- `sinfo`

PARALLEL APPLICATIONS

Threaded application on a single node (OpenMP):

- salloc -N 1 –n 1 ...
- OMP_NUM_THREADS=128 srun –n 1 ./app

Pure MPI application on several nodes:

- salloc -N 4 –n 512 ...
- srun –n 512 ./app

Hybrid application on several nodes (MPI+OpenMP):

- salloc -N 4 –n 128 ...
- OMP_NUM_THREADS=4 srun –n 128 ./app

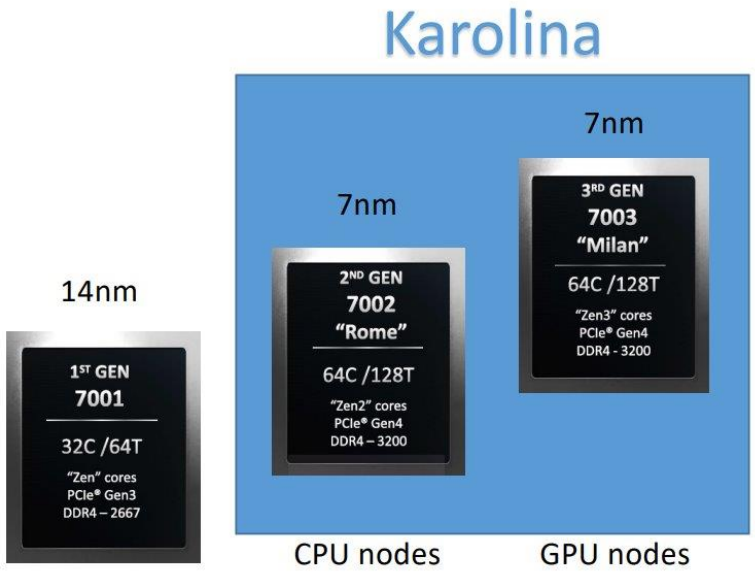
PARALLEL APPLICATIONS

Pure MPI application:

- `salloc -N 1 –n 128 ...`
- `srun –n 128 ./app`

MPI+OpenMP application:

- `salloc -N 1 –n 32 ...`
- `OMP_NUM_THREADS=4 srun –n 32 ./app`



CATEGORY	EPYC 7002 (Rome)	EPYC 7003 (Milan)
Socket	SP3	SP3
Core / Process	Zen2 / 7nm	Zen3 / 7nm
Max Core Count / Threads	64 / 128	64 / 128
L3 Cache Size	256 MB	256 MB
CCX Arch	4 Cores + 16MB	8 Cores + 32MB
Memory	8 Ch DDR4-3200, NVDIMM-N	8 Ch DDR4-3200, NVDIMM-N
PCIe Tech & Lane Count	PCIe Gen4, 128L/Socket	PCIe Gen4, 128L/Socket
Security	SME, SEV	SME, SEV, SNP
Chipset	NA	NA
Power	120W - 280W	120W - 280W

PARALLEL APPLICATIONS

Pure MPI application:

- `salloc -N 1 –n 128 ...`
- `srun –n 128 ./app`

MPI+OpenMP application:

- `salloc -N 1 –n 32 ...`
- `OMP_NUM_THREADS=4 srun –n 32 ./app`

Is it the best possible settings?

PARALLEL APPLICATIONS

I have a simple application:

- Karolina supercomputer at IT4I
- compile with OpenMPI: `mpic++ -fopenmp -O3 -march=native app.cpp -o app`
- test with different number of MPI processors up to 128

- salloc -p qcpu_exp -N 1
- export OMP_NUM_THREADS=1

- srun –n 8 ./app
- srun –n 16 ./app
- srun –n 32 ./app
- srun –n 64 ./app
- srun –n 128 ./app

PARALLEL APPLICATIONS

I have a simple application:

- Karolina supercomputer at IT4I
- compile with OpenMPI: `mpic++ -fopenmp -O3 -march=native app.cpp -o app`
- test with different number of MPI processors up to 128

- salloc -p qcpu_exp -N 1
- export OMP_NUM_THREADS=1

- srun –n 8 ./app 39.66s
- srun –n 16 ./app 17.46s
- srun –n 32 ./app 11.87s
- srun –n 64 ./app 7.31s
- srun –n 128 ./app 5.56s

Is 128 the best possible settings?

What about mapping and pinning?

MAPPING, PINNING

Mapping:

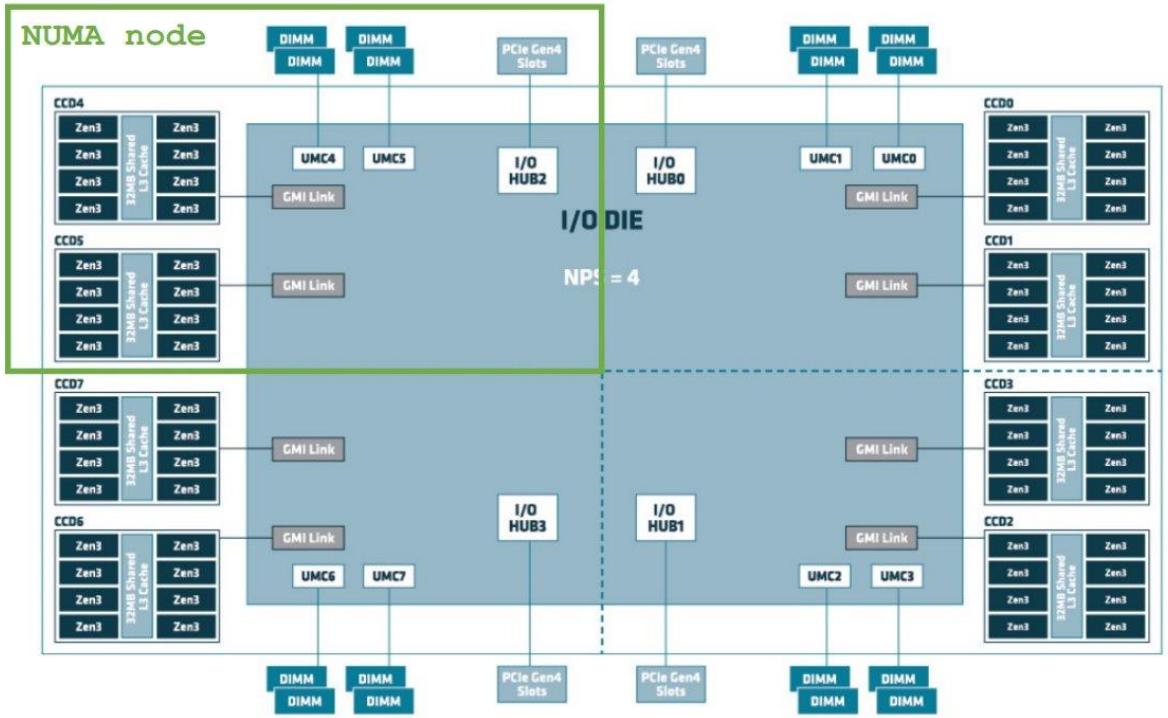
- specifies how the software components are mapped to a given hardware

Pinning, binding:

- deny migration of threads and processes to another resources

```
numactl -H
| node 0 cpus: 0   -  15
| node 1 cpus: 16  -  31
| node 2 cpus: 32  -  47
| node 3 cpus: 48  -  63
| node 4 cpus: 64  -  79
| node 5 cpus: 80  -  95
| node 6 cpus: 96  - 111
| node 7 cpus: 112 - 127
| node 0-7 size: 128 GB
```

	0	1	2	3
0	10	12	12	12
1	12	10	12	12
2	12	12	10	12
3	12	12	12	10



MAPPING, PINNING

Mapping:

- specifies how the software components are mapped to a given hardware

Pinning, binding:

- deny migration of threads and processes to another resources

OpenMPI: *mpirun -map-by {socket, numa, l3cache}*

Intel MPI*: I_MPI_PIN_DOMAIN={socket, numa, cache3}*

OpenMP: *OMP_PROC_BIND={true, close, spread}*

Slurm:

- --cpu-bind={sockets,ldoms,cores}
- -c, --cpus-per-task=<ncpus>

PARALLEL APPLICATIONS

I have a simple application:

- Karolina supercomputer at IT4I
- compile with OpenMPI: `mpic++ -fopenmp -O3 -march=native app.cpp -o app`
- test with different number of MPI processors up to 128

- salloc -p qcpu_exp -N 1
- export OMP_NUM_THREADS=1

srun –n 8 ./app 39.66s	srun –n 8 –c 16 ./app 7.11s
srun –n 16 ./app 17.46s	srun –n 16 –c 8 ./app 5.21s
srun –n 32 ./app 11.87s	srun –n 32 –c 4 ./app 5.08s
srun –n 64 ./app 7.31s	srun –n 64 –c 2 ./app 5.31s
srun –n 128 ./app 5.56s	srun –n 128 –c 1 ./app 5.56s

PARALLEL APPLICATIONS

I have a simple application:

- Karolina supercomputer at IT4I
- compile with OpenMPI: `mpic++ -fopenmp -O3 -march=native app.cpp -o app`
- test with different number of MPI processors up to 128

- salloc -p qcpu_exp -N 1
- export OMP_NUM_THREADS=1

srun –n 8 ./app	39.66s	srun –n 8 –c 16 ./app	7.11s
srun –n 16 ./app	17.46s	srun –n 16 –c 8 ./app	5.21s
srun –n 32 ./app	11.87s	srun –n 32 –c 4 ./app	**5.08s (91% of the previous best)**
srun –n 64 ./app	7.31s	srun –n 64 –c 2 ./app	5.31s
srun –n 128 ./app	**5.56s**	srun –n 128 –c 1 ./app	5.56s

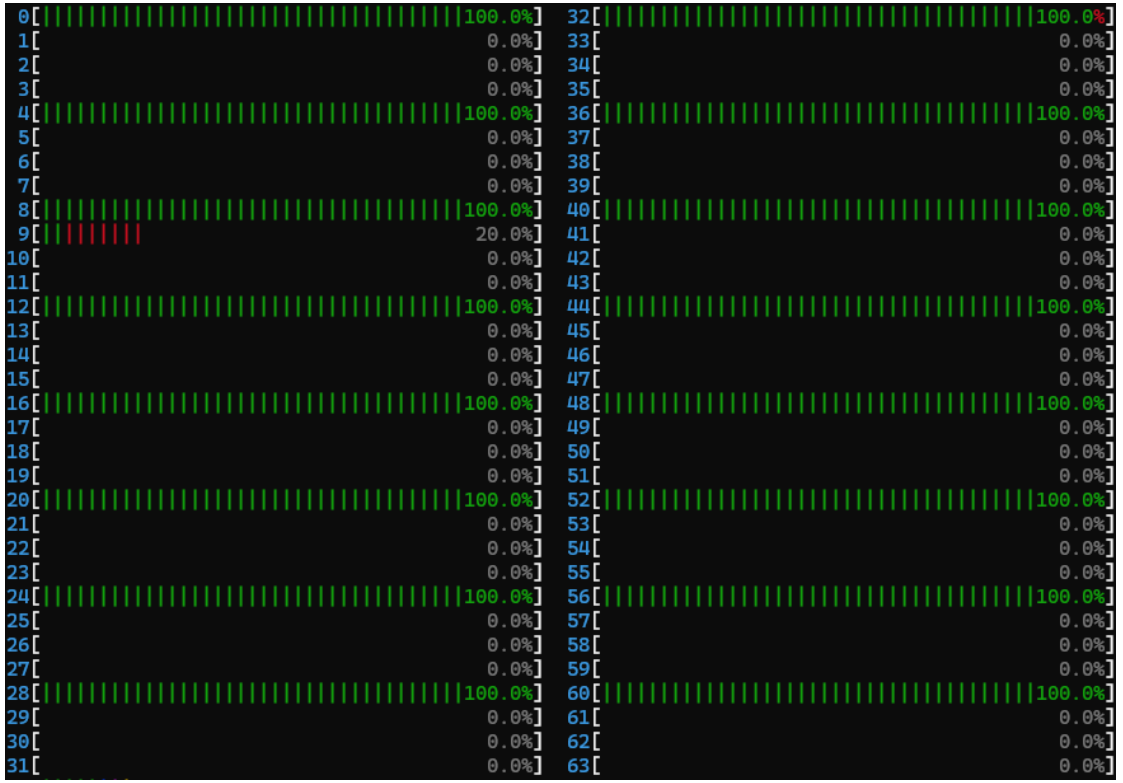
PARALLEL APPLICATIONS

I have a simple application:

- Karolina supercomputer at IT4I
- compile with OpenMPI: `mpic++ -fopenmp -O3 -march=native app.cpp -o app`
- test with different number of MPI processors up to 128

- salloc -p qcpu_exp -N 1
- export OMP_NUM_THREADS=**32**
- export OMP_PROC_BIND=**spread**

- srun -n 1 ./app 5.08s



RECOMMENDATION

Memory bound application

- Number of MPI processes / thread equal to memory channels
- Correct pinning to NUMA domains

Compute bound application

- As many MPI processes / threads as possible
- pinning to avoid migration

Your application

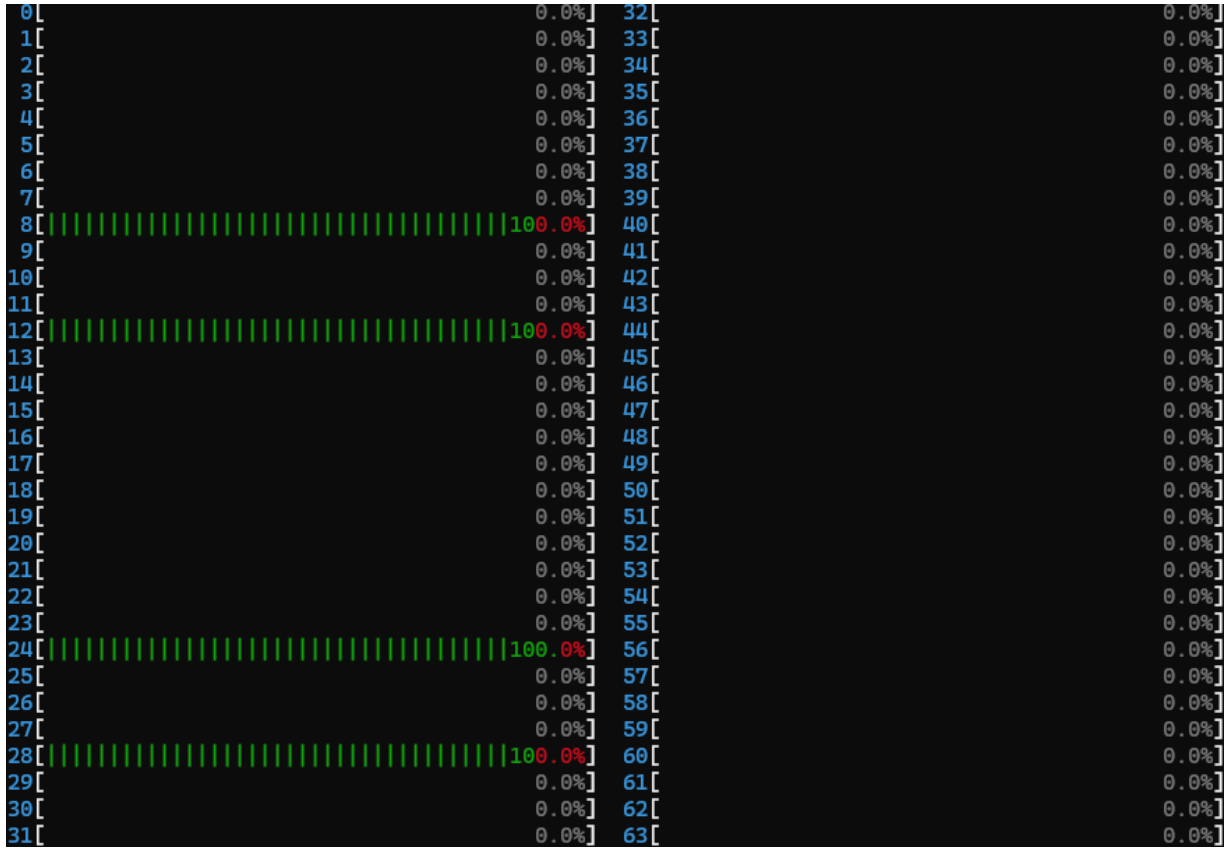
- Test performance for different number of cores per node:
  - 16, 32, 64, 128 cores per node

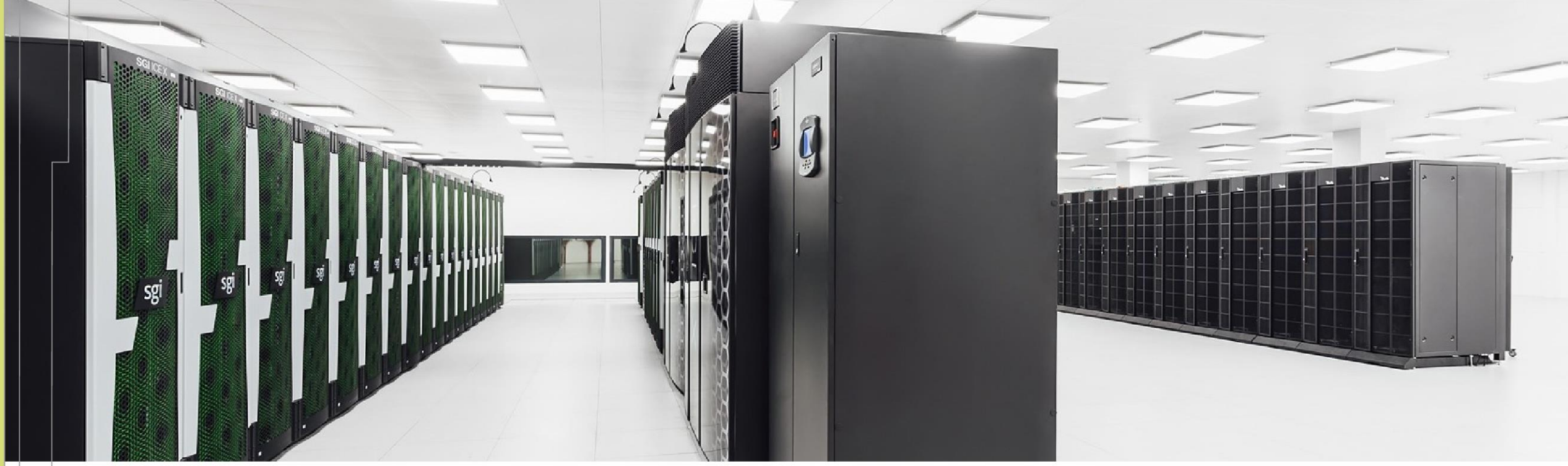
- Test different mapping / pinning options
  - close, spread

PARALLEL APPLICATIONS

More advanced binding

srun -n 4 --cpu-bind=mask_cpu: 0x100,0x1000,0x1000000,0x10000000 ./app





Ondřej Meca
ondrej.meca@vsb.cz

IT4Innovations National Supercomputing Center
VSB – Technical University of Ostrava
Studentská 6231/1B
708 00 Ostrava-Poruba, Czech Republic
www.it4i.cz

VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER