JUWELS (CLUSTER + BOOSTER)
Jülich Wizard for European Leadership Science

2021-11-22 | Damian Alvarez | HPS group @ JSC

JÜLICH
Forschungszentrum

JUWELS

Jülich Wizard for European Leadership Science

- System vendor:Bull/Atos technologies
- JUQUEEN successor
- Next-generation capability computing system @ JCS
- Modular Supercomputing Architecture
  - First (cluster) module
    - Versatile cluster architecture based on commodity multi-core CPUs
  - Second (booster) module
    - Optimized for massively parallel workloads
    - GPU-Accelerator based
- Dedicated to GCS, PRACE projects (Tier-0)

JUWELS

JUWELS cluster

- Deployed in June 2018
  - #23 in top 500 at the time, #77 now
- Bull Sequana X1000 system (10 cells)
  - Dual Intel Xeon Skylake 8168
  - 266 TByte main memory
  - 122,448 CPU cores + 71,680 FP64 CUDA cores
  - Mellanox Infiniband EDR/HDR
    - Fat tree with 2:1 prunning at L1
    - Connected with the booster via 200 IB HDR links!
- 12 Petaflop/s peak
  - 10.4 PFlop/s (CPU)
  - 1.6 PFlop/s (GPU)
- Connected to JUST for the filesystem

JUWELS

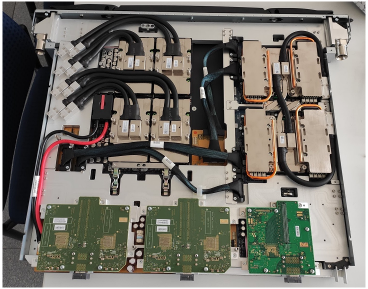
JUWELS booster

- Deployed in November 2020
  - #7 in top 500, #8 now
  - #3 in green 500 (best result for a top 10 system), #11 now
  - #5 in HPCG 500, #6 now
- Bull Sequana XH2000 system (20 cells)
  - Dual AMD EPYC 7402
  - 479 TByte main memory + 150 TByte HBM
  - 44,928 CPU cores + 12,939,264 FP64 CUDA cores
  - Mellanox Infiniband HDR
    - DragonFly+
    - Connected with the cluster via 200 IB HDR links!
- 71 Petaflop/s peak
  - 1 PFlop/s (CPU)
  - 70 PFlop/s (GPU)
- Connected to JUST for the filesystem

JUWELS BOOSTER

Node Types

- All nodes are the same type
- 936 standard compute nodes
  - Dual AMD EPYC 7402 (Zen 2 →AVX2)
  - 2x 24 cores, 2.8 GHz, 512 GB memory
  - 4x HDR-Infiniband (Connect-X6) 800 Gbps injection BW!
  - 4x NVIDIA Ampere A100, 40 GB HBM
- All compute nodes are diskless!
  - OS Image in RAM →~499GB available for users

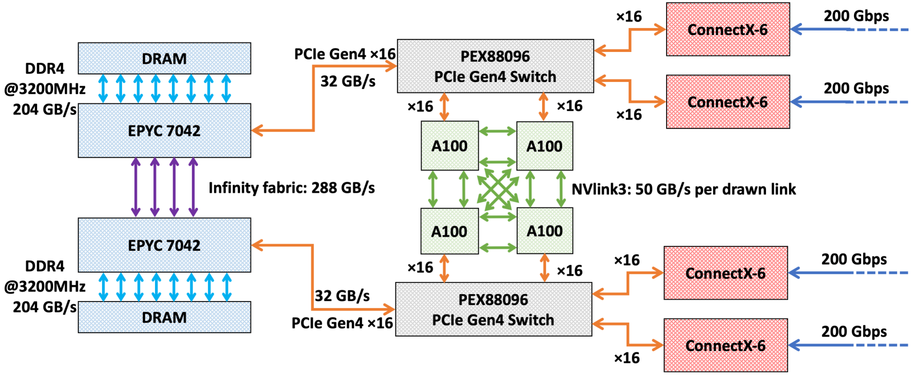


JÜLICH Forschungszentrum

JUWELS BOOSTER

Node Types

Booster node topology

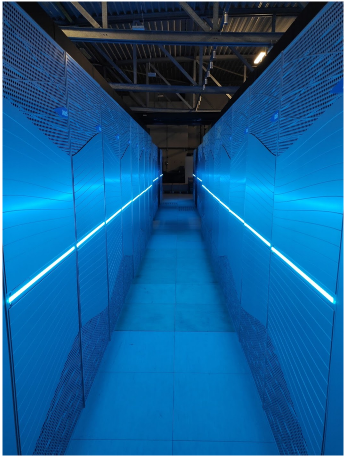


JÜLICH
Forschungszentrum

JUWELS BOOSTER LOGIN NODES

Node Types

- 4 login nodes
  - 2x 24 cores, 2.8 GHz, 512 GB memory
  - HDR-Infiniband (Connect-X6)
  - 100 GigE external network (DFN)
  - No GPUs
- Used for:
  - Compile/Submit jobs
  - Small pre- and post-processing
  - Small visualizations
  - Transfers (see also `judac.fz-juelich.de`)
- Shared nodes!
- Can submit also to the cluster →carefully



JÜLICH Forschungszentrum

JUWELS

System access

- Access with SSH keys
- Recommendation: Ed25519 or RSA (4096 bit)
- Protection of private key with non-trivial pass phrase is mandatory!

login to JUWELS

```
ssh-keygen {-t ed25519 | -t rsa -b 4096}

[Upload ssh key to judoor...]

ssh <user>@{juwels-cluster|juwelsvis|juwels-booster}.fz-juelich.de

ssh <user>@juwels[01-09].fz-juelich.de
ssh <user>@juwelsvis[00-03].fz-juelich.de
ssh <user>@juwels[21-24].fz-juelich.de
```

JUWELS

Software environment

- Operating system: CentOS 8.4 (soon Rocky Linux 8.5)
- Batch system based on SLURM/ParaStation
  - Workload management and UI => SLURM
  - Resource management => ParaStation (psid + psslurm)
- Programming environment:
  - GNU Compilers
  - Intel Professional Fortran & C/C++ Compilers
  - NVHPC Compilers
  - OpenMP (Intel, GNU, NVHPC)
  - CUDA
  - Communication libraries
    - ParaStation MPI (based on MPICH3)
    - OpenMPI
    - Intel MPI
  - Optimized mathematical libraries (Intel Math Kernel Library, etc.)
- Applications (in `$STAGES/$STAGE/software`)

JÜLICH Forschungszentrum

JUWELS

`jutil`

- Compute/data project and budget management

JUTIL example

```
# See your projects
jutil user projects

# See your compute allocation
jutil user cpuquota # (at the moment use q_cpuquota)

# See your disk quota
jutil user dataquota

# Activate environment for a given project
jutil env activate -p <project> [-A <budget>]
```

JUWELS

Accessing software (hierarchical modules)

Modules environment

```
# List modules ready to be loaded right away
ml avail
# Load the desired compiler
ml <Compiler> <MPI>
# Load the desired MPI
ml <MPI>
# Load additional applications/libraries
module load <module name>
# Search for an application/library
module spider <name>
module key <name>
```


JÜLICH Forschungszentrum

JUWELS

Filesystems on JUWELS

Filesystem	Size*	Size quota	File quota	Backup	Access
`$HOME`	2.8 TB	10 GB	40K	Tape	Everywhere
`$SCRATCH`†	9.1 PB	90 TB	4M	No	Everywhere
`$PROJECT`†	2.3 PB	16 TB	3M	Tape	Everywhere
`$FASTDATA`†	9.1 PB	Proj. dep.	Proj. dep.	Snapshot (GPFS)	Everywhere
`$DATA`†	14 PB	Proj. dep.	Proj. dep.	Snapshot (GPFS)	Login
`$ARCHIVE`†	1.9 PB	Proj. dep.	Proj. dep.	Tape (migration)	Login
HPST (soon)					Everywhere

* Total size, your usable size depends on your quota
† This variable is set after setting up the environment with `jutil`

JÜLICH Forschungszentrum

JUWELS

SLURM Partitions on JUWELS

Partition	Node type	# of nodes	Max nodes	Default time	Max time
`devel`	Normal	20	8	30m	2h
`develgpus`	GPU	10	2	1h	2h
`batch`	Normal+Large mem	2485	1024	1h	24h*
`gpus`	GPU	46	46	1h	24h*
`large`†	Normal+Large mem	2485	2485	1h	24h*
`mem192`	Large mem	240	224	1h	24h*
`esm`	Normal	6	6	30m	24h*
`maintcluster`	All	2567 (cluster)	2567 (cluster)	–	–
`develbooster`	Booster	10	4	30m	2h
`booster`	Booster	926	756	1h	24h*
`maintbooster`	Booster	936 (booster)	936 (booster)	1h	24h*
`maint`	All	3503 (all)	3503 (all)	–	–

* 6h if you have low or no contingent
† Available just on special occasions (after maintenance, scaling workshops)

JUWELS

Further information

- MOTD: Message of the Day
  - Information about preventive and emergency maintenances
  - Information about system configuration changes
- On-line documentation
  - `www.fz-juelich.de/ias/jsc/EN/Expertise/Supercomputers/JUWELS/JUWELS_node.html`
  - `apps.fz-juelich.de/jsc/hps/juwels/index.html`
  - `apps.fz-juelich.de/jsc/hps/juwels/booster-overview.html`
- User support at FZJ
  - sc@fz-juelich.de
  - Phone: 02461 61-2828