Journal of large-scale research facilities

Journal of large-scale research facilities, 7, A183 (2021) http://dx.doi.org/10.17815/jlsrf-7-183

Published: 29.10.2021

JUWELS Cluster and Booster: Exascale Pathfinder with Modular Supercomputing Architecture at Jülich Supercomputing Centre

Forschungszentrum Jülich, Jülich Supercomputing Centre *

Instrument Scientists:
- Supercomputing Support, Jülich Supercomputing Centre, Forschungszentrum Jülich, phone: +49(0)2461 61 2828, sc@fz-juelich.de

Abstract: JUWELS is a multi-petaflop modular supercomputer operated by Jülich Supercomputing Centre at Forschungszentrum Jülich as a European and national supercomputing resource for the Gauss Centre for Supercomputing. In addition, JUWELS serves the Earth system modeling community and the AI community within the Helmholtz Association as well. JUWELS currently consists of two modules. The first module deployed in 2018 is the so-called Cluster module. The Cluster is a BullSequana X1000 system with Intel Xeon Skylake-SP processors and Mellanox EDR InfiniBand. The second module deployed in 2020 is the so-called Booster module. The Booster is a BullSequana XH2000 system with 2nd generation AMD EPYC processors, NVIDIA Ampere GPUs and NVIDIA/Mellanox HDR Infiniband. This paper describes in detail the architecture of the system from a users perspective, and additionally provides further insights into the administrative infrastructure used to operate the supercomputer.

1 Introduction

The Jülich Supercomputing Centre (JSC) at the Forschungszentrum Jülich (Forschungszentrum Jülich, 2021a) has been operating some of the largest supercomputers in Europe for three decades and is part of the larger Gauss Centre for Supercomputing (Gauss Centre for Supercomputing, 2021a), offering three tier 0 supercomputers to members of the European research community. As part of this effort, JSC hosts and operates JUWELS (Jülich Wizard for European Leadership Science), a modular supercomputer whose most recent module –the JUWELS Booster, deployed in November 2020– ranks currently on its own as the most powerful supercomputer in Europe, and #7 in the world (Top500, 2021). The JUWELS Booster completes the JUWELS Cluster, the first JUWELS module deployed in 2018 (Jülich

* **Cite article as:** Jülich Supercomputing Centre. (2021). JUWELS Cluster and Booster: Exascale Pathfinder with Modular Supercomputing Architecture at Jülich Supercomputing Centre. *Journal of large-scale research facilities, 7*, A183. http://dx.doi.org/10.17815/jlsrf-7-183

Supercomputing Centre, 2019) that replaced JUQUEEN (Forschungszentrum Jülich, 2015) as the most capable system in the centre, and that is currently ranked #44 worldwide. These two modules have different node and network designs, which will be explained in detail in this article.

Figure 1: The supercomputer JUWELS in the facility of Jülich Supercomputing Centre. Left: Booster Module. Right: Cluster Module. Copyright: Forschungszentrum Jülich GmbH / Wilhelm-Peter Schneider

The investment and operational costs for JUWELS are covered by funding from the German Ministry of Education and Science (Bundesministerium für Bildung und Forschung - BMBF) and the Ministry for Culture and Science of the State North Rhine-Westphalia (Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen - MKW) via the Gauss Centre for Supercomputing (GCS). In addition, funding from the Helmholtz Association (Helmholtz Association, 2021) allows the dedicated enlargement of the system for the Earth system modelling community and the artificial intelligence (AI) community within the Helmholtz Association. Therefore, the primary users of the modules in JUWELS are the researchers integrated in the research networks of these institutions.

GCS is jointly funded by the German Ministry of Education and Science and the corresponding ministries of the three national states of Baden-Württemberg, Bavaria, and North Rhine-Westphalia. As Germany's foremost supercomputing institution, GSC integrates three national supercomputing centers: High-Performance Computing Center Stuttgart (HLRS) (High-Performance Computing Center Stuttgart, 2021), Jülich Supercomputing Centre (JSC) (Forschungszentrum Jülich, 2021b), and Leibniz Supercomputing Centre (LRZ) (Leibniz Supercomputing Centre of the Bavarian Academy of Sciences & Humanities, 2021). GCS is a hosting member of the pan-European Partnership for Advanced Computing in Europe (PRACE) (Partnership for Advanced Computing in Europe, 2021) organization consisting of 26 member countries, whose representative organizations create a pan-European supercomputing infrastructure, providing access to computing and data management resources and services for large-scale scientific and engineering applications at the highest performance level.

2 JUWELS System Details

The JUWELS system implements the modular supercomputing architecture that was pioneered through a series of EU-funded DEEP projects (Eicker et al., 2016). The JURECA system (Jülich Supercomputing Centre, 2018) was the first demonstration of this architecture in a multi-petaflop system at JSC. As the initial deployment of JUWELS in 2018, the Cluster module utilizes latest generation (at time of deployment) Intel Xeon processors and provides a familiar and versatile environment in support of

a broad class of computational workloads. The Booster module, deployed in late 2020 relies on AMD EPYC processors, but leverages the most modern GPU architecture to date –NVIDIA A100–, from which it gets most of its computational capability. The following subsections will describe the technical details of both modules.

2.1 JUWELS Cluster Module

The JUWELS Cluster is a BullSequana X1000 supercomputer. The Sequana X1000 series (Atos, 2021a) by Atos provides a high-density node integration with warm-water direct-liquid cooling capabilities. It follows a scalable hierarchical cell-based design. The JUWELS Cluster consists of ten Sequana X1000 cells with nine times 279 compute nodes (CPU-only partition) and a 10th cell with 48 GPU-accelerated compute nodes. These cells are divided in three racks. Two of these racks –named *base* and *extension*– host the compute nodes, that are inserted from the front and the back of the rack. Between these two racks is the *switch* rack. This rack hosts the cell switches –which are a custom design for the X1000 architecture– and the cell management nodes. Three major features of these switches are: 1) they are water-cooled; 2) the level-one switches have InfiniBand and Ethernet components in a single switch; and 3) they have custom connectors for the intra-rack connections.

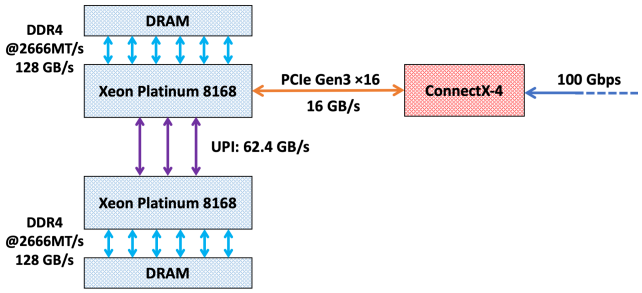
2.1.1 Node Architecture

The 2,511 computes nodes in the CPU-only partition of JUWELS are equipped with two Intel Xeon Skylake Platinum 8168 central processing units (CPUs) with 24 cores each and a base frequency of 2.7 GHz. Three nodes form together one BullSequana X1120 compute blade, as can be seen in Figure 2. 90% (2,271) of these compute nodes feature 96 GB main memory, the remaining 240 nodes offer a main memory capacity of 192 GB. The nodes are equipped with a Mellanox ConnectX-4 EDR InfiniBand host channel adapter (HCA) connected with 16 PCIe Gen 3 lanes providing a network injection bandwidth of up to 12.5 GB/s.

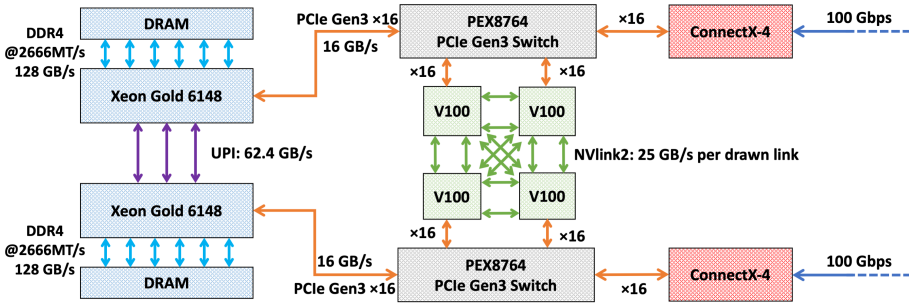
Figure 2: Example of a BullSequana compute blade encompassing three compute nodes. Please note that the shown local storage devices are not built into the JUWELS compute nodes. Copyright: Atos.

The GPU partition in the JUWELS Cluster consists of 56 compute nodes based on the BullSequana X1125 accelerator blade. The nodes are equipped with two Intel Xeon Gold 6148 processors with 20

cores each and 192 GB main memory. Each node contains four NVIDIA Volta V100 GPUs in SXM2 form factor with 5,120 CUDA cores, 16 GB high-bandwidth memory (HBM2) and a peak double precision floating point performance of 7.8 TF/s. The GPUs are connected to a PCIe Gen 3 switch via 16 lanes and to each other with two NVIDIA NVLink2 links for a total bi-directional peak bandwidth of 100 GB/s. The PCIe switch is also connected to the CPUs. The nodes are equipped with two Mellanox ConnectX-4 HCAs connecting to the PCIe switch. Due to the topology of the intra-node PCIe interconnect, GPUdirect remote direct memory access is possible for each GPU via one of the two HCAs, cf. Figure 3b.



(a) JUWELS Cluster compute node (CPU partition).



(b) JUWELS Cluster compute node (GPU partition).

Figure 3: Annotated block diagrams of JUWELS compute nodes.

In addition to the compute nodes, the JUWELS Cluster includes 12 login nodes (Sequana X430 E5 2U2S) and four visualization login nodes (Sequana X410 E5). The login nodes feature two Intel Xeon Gold 6148 CPUs and 768 GB main memory. The visualization login nodes are additionally equipped with an NVIDIA Pascal P100 GPU for visualization. All 16 login nodes are equipped with a Mellanox ConnectX-5 EDR InfiniBand HCA and a 100 Gb/s external Ethernet connectivity.

All nodes are based on the Intel Purley server platform and use fourth generation dynamic data rate (DDR4) memory with an I/O bus frequency of 1,333.33 MHz (2,667 MT/s; mega-transfers per seconds). Each processor has six memory channels for a peak transfer bandwidth of 128 GB/s per processor and 256 GB/s per node. Like any recent X86 architecture, the compute nodes are classified as a Non-Uniform Memory Architecture (NUMA). The two processors in each node are connected via two Intel Ultra Path Interconnect (UPI) links supporting up to 20.8 billion transfers per second. All compute nodes in JUWELS are disk-less. Hence, the entire operating system (including file system memory pools) is loaded into memory and occupies a share of the available main memory capacity.

The peak performance of a JUWELS Cluster compute node (excluding GPU nodes) equals 4.15 TF/s when using the base frequency for the calculation. The Skylake microarchitecture includes support for the AVX-512 Instruction Set Architecture (ISA) extension, which specifies mathematical operations on

512-bit vectors, i.e., eight double precision values at a time. Each core can perform two 512-bit wide fused multiply-add operations per cycle. In order to leverage the floating-point performance of the JUWELS compute nodes, applications need to be amendable to vectorization of the computationally intense code segments. However, the gap between nominal floating point performance and memory bandwidth has widened further with the Purley Platform. In contrast, e.g., to the now decommissioned JURECA system based on the Intel Grantley Platform with Intel Xeon E5-2680 v3 processors, the core number and vector width has doubled (i.e., the floating point peak performance has quadrupled) while the number of memory channels has increased by only 150%. This limits the sustainable performance of memory-bound codes and requires higher optimization efforts by developers.

2.1.2 Cluster InfiniBand Network Architecture

The Cluster nodes are organized in a three-level fat-tree topology. In each cell, 12 level-one switches and 12 level-two switches are located. In cells 1 to 9 (CPU-only partition), the first 9 level-one switches connects downwards to 24 compute nodes. The last 3 level-one switches connect to 21 computes nodes. All level-one switches connect upwards with 12 links to the level-two switches, i.e., a 1:2 pruning at level-one is present.

In the 10th cell (GPU partition), eight nodes connect to the first six level-one switches (16 links), and four nodes connect to switches 7 and 8. All level-one switches have twelve uplinks towards level-two, i.e., a 3:4 pruning is applied. The remaining slots in the last level-one switches in the cell were originally connecting various service nodes to the fabric. These nodes have been moved to the Booster part of the fabric during the Booster deployment.

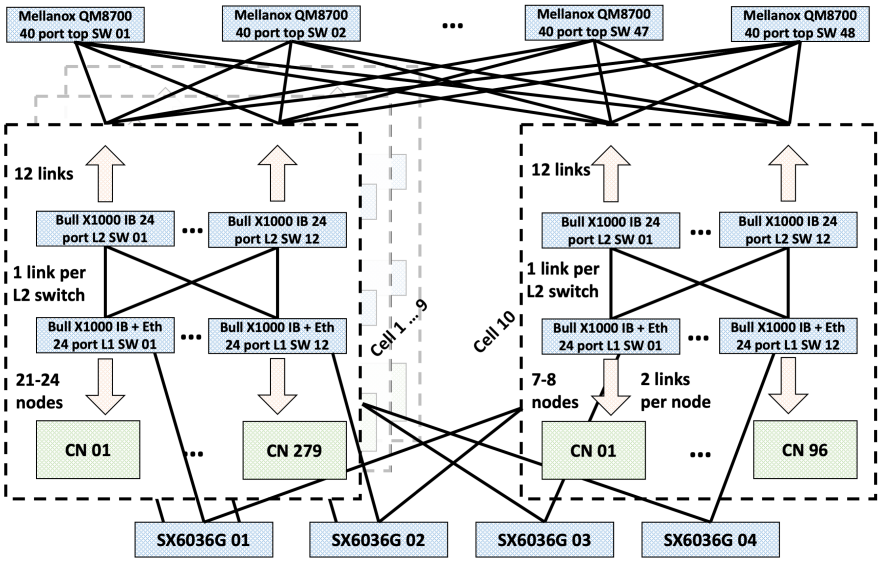


Figure 4: JUWELS Cluster InfiniBand network.

The topology is completed by 48 level-three switches that are located outside of the Sequana X1000 cells. Between level-two and level-three no pruning is present. Each cell connects with 144 uplinks to level-three (12 links per level-two switch). The intra-cell switches (level-one and level-two) are an Atos-proprietary custom switch design based on the Mellanox Switch-IB 2 technology. The level-three switches are discrete Mellanox switches. At the time of the system installation, 36-port Mellanox EDR InfiniBand switches were used. A replacement by 40-port Mellanox HDR InfiniBand (200 Gb/s) switches

was done as part of the preparation for the Booster deployment.

The JUWELS login and compute nodes access the parallel file systems exported by the central Jülich Storage cluster JUST (Forschungszentrum Jülich, 2021d) using the IBM Spectrum Scale (formerly GPFS) file system software (IBM, 2021). Users with access to several supercomputers in the high-performance computing facility at JSC can work with the same file systems on all systems so that data movement is minimized and workflows are simplified. The storage access is realized using the Internet Protocol (IP)-over-InfiniBand technology and four SX6036G InfiniBand-to-Ethernet gateways. The network performance of the I/O subsystem is 250 GB/s enabling 200+ GB/s performance on the high-bandwidth JUST file systems. The gateway switches connected to the level-one switches in the cells and the routing ensures that storage connections do not cross level-three. In consequence, each cell has a separate storage bandwidth of ca. 25 GB/s available. In addition, compute nodes connected to the same level-one switch share the same network path to each of the four gateway switches. Depending on job distribution and concurrent I/O load, different I/O bandwidth values will be observed under production conditions.

2.2 JUWELS Booster Module

The JUWELS Booster is a BullSequana XH2000 (Atos, 2021b) supercomputer. It shared with the X1000 the cooling technology. There are, however, some major differences in other aspects of the architecture. The switches are integrated in the compute rack –each one with 24 compute nodes–, instead of being separated on their own rack like in X1000. That enables a more flexible design, where the basic building block is a single XH2000 rack, instead of a cell composed of 3 racks like in the previous generation. Another difference in this area is that the switches have standard connectors, allowing to choose freely the intracell topology. These changes enabled a stratification of the hierarchy from the perspective of the different networks. In the JUWELS Booster, considering the InfiniBand network, a *cell* –or DragonFly+ group– is formed by two racks. On the other side, considering the Ethernet administrative network, an *island* is formed by ten racks. The *island managers* (ISMAs) are hosted in a separate administrative rack, instead of being integrated in the switch rack like in X1000.

2.2.1 Node Architecture

The JUWELS Booster is composed of 936 X2415 compute blades. Each one of these blades contains a single node equipped with 512 GB of memory and two AMD EPYC 7042 processors with base frequency 2.8 GHz. Each processor has 24 cores, to match the number of cores available in the Cluster partition and slightly facilitate users the usage of the system as a whole. Despite the good capabilities of these processors, the main feature of the JUWELS Booster nodes are its GPUs. In each compute node there are four NVIDIA Ampere A100 GPUs in SXM4 form factor. Each of these GPUs have 6912 CUDA cores and 40 GB of HBM2 memory. The A100 GPUs are also capable of communicating with each other bidirectionally at a rate of 200 GB/s via a NVLink3 bus. Similarly to the GPU nodes in the Cluster partition, the GPUs in the Booster nodes are connected to a PCIe Gen4 switch. These switch bridges the gap between CPUs, GPUs and HCAs. The interconnect is another major difference between the GPU nodes in the Cluster partition and the Booster nodes. Each node has four Mellanox ConnectX-6 HDR InfiniBand HCAs, so each GPU can communicate with remote nodes without interference from other GPUs on the node. The complete topology of the nodes is depicted in Figure 6.

The CPUs in the Booster nodes have four NUMA domains, two of which have direct connectivity with the PCIe switches. The memory subsystem is DDR4, clocked at 1,600 MHz (3,200 MT/s). The peak bandwidth is 204 GB/s thanks to the 8 memory channels per CPU, and the aggregated bandwidth for the four infinity fabric links connecting both CPUs is 288 GB/s. The GPUs have a much higher 1.5 TB/s of memory bandwidth.

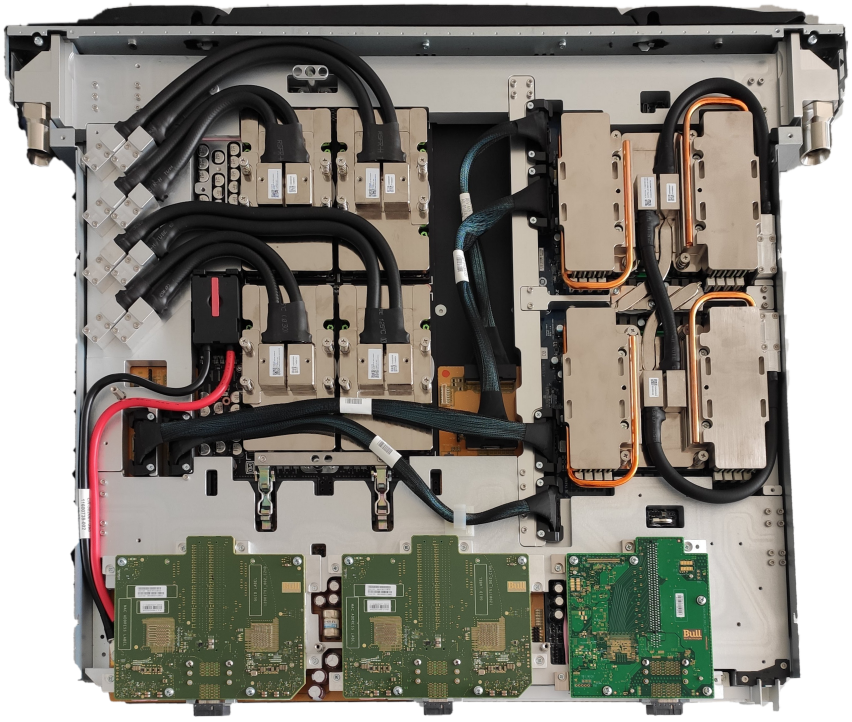


Figure 5: Example of a BullSequana compute blade with 4x A100 GPUs in a Redstone board (left), 2 AMD EPYC 7042 CPUs (right) and 4 HDR200 ConnectX-6 HCAs in 2 mezzanine cards (bottom center and left).

The AMD EPYC 7042 CPUs are AVX2 capable, each core contains two execution units, which results in a peak performance per node of 2.1 TF/s, using the base frequency. This is a lower value than what the Cluster nodes can provide. However, the strength of these nodes resides on its GPUs. The A100 peak performance in double precision is up to 19.5 TF/s when using Tensor cores (78 TF/s per node). Additionally, a major feature of these GPUs is their ability to compute in reduced precision at a very high rate, reaching up to 624 TF/s, so applications that can leverage this feature can have significant speedups.

The JUWELS Booster includes its own set of login nodes. These nodes are a simplified version of the compute nodes, having the same CPUs and memory, but lacking the GPUs for computation. These nodes are also equipped with dual port Mellanox ConnectX-6 HCAs, with one port configured as Ethernet for external access of up to 100 Gb/s, and the other port connected to the InfiniBand fabric also at a 100 Gb/s rate.

2.2.2 Booster InfiniBand Network Architecture

Unlike the Cluster InfiniBand network, the Booster topology is DragonFly+. All the switches in the compute racks are Atos BullSequana XH2000 Quantum Switches, with 40 HDR200 ports, and are based on NVIDIA QM8700 switches. In the Booster topology configuration, the level-two switches between cells are connected directly, with ten links between cell pairs (one link per level-two switch), without the need of level 3 switches. The main advantage of these topology as opposed to Fat Tree, is that the number of optical links connecting cells (or DragonFly+ groups) is reduced, and so are the number of switches, since the top-level switches are basically avoided. These characteristic makes it desirable for large networks to avoid the associated high costs of global optical links and top-level switches. The minimal number of hops between nodes in different cells is also reduced. However, the downside of this

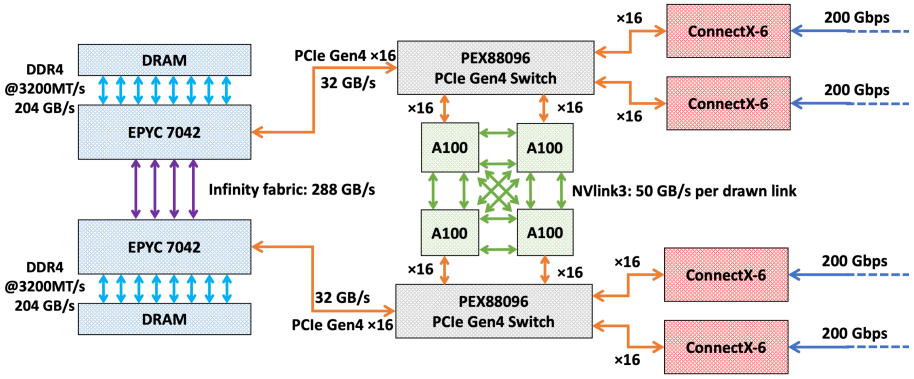


Figure 6: JUWELS Booster compute node.

topology is that the bisectional bandwidth is reduced. To counteract that fact, DragonFly+ topologies incorporate adaptive routing, where packets can be redirected via links that are not part of the shortest path between two endpoints, if the links in the shortest path are congested. That way, the reduced bisectional bandwidth becomes a limiting factor just under very particular communication patterns.

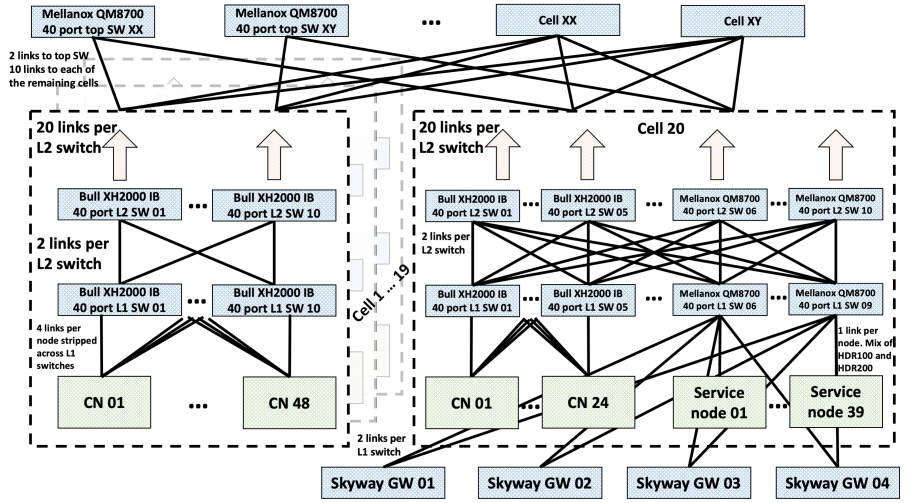


Figure 7: JUWELS Booster InfiniBand network.

Inside the DragonFly+ groups, the topology is actually a Fat Tree without pruning. Each cell is composed of two racks, and has 10 level-one switches and 10 level-two switches (5 of each per rack). The four links per node are striped across level-one switches, as opposed to be connected to the same switch. Per rack, the first 4 level-one switches have a total of 20 links towards compute nodes, whereas the last level-one switch has 16 links. In total, there are 20 uplinks at the level-one and level-two switches. Figure 7 depicts the structure of the Booster network.

One exception to this cell structure is the last cell. The reason for that is that this cell is composed of one compute rack and then a series of racks for administrative nodes, login nodes and NVIDIA Skyway InfiniBand to Ethernet gateways towards the storage cluster. All the links between level-one switches and nodes or gateways in this part of the last cell are HDR100 links, and use split cables to connect to a single HDR200 port on the switch end. The links between level-one and level-two switches are, how-

ever, HDR200. All the switches here are NVIDIA QM8700 with 40 HDR200 ports. Given the number of administrative nodes, login nodes and gateways needed, just four level-one switches are needed here, instead of the five that are used elsewhere in the network.

The gateways linking the InfiniBand and Ethernet fabrics of the supercomputer and storage, respectively, have a total of 8 InfiniBand links, and have 2 links per level-one switch. The routing on the system ensures that each island uses exclusively one of the gateways. The total bandwidth towards storage is of 400 GB/s, and each island can achieve a maximum of 100 GB/s. This bandwidth is also achievable from a single cell.

2.3 Global Network

Until now this article described the topology of the Cluster network, which was operated standalone since its deployment until the merge with the Booster, and the Booster network, which was operated as well standalone during a brief testing phase. It is noteworthy, that these two networks are currently operated as a single one, with a single active subnet manager. The following subsection describes the details of the merge of these two networks.

2.3.1 Cluster-Booster Connection

The top-level switches in the Cluster network are actually oversized, since they contain more ports than strictly necessary for a Fat Tree topology. This was a conscious design decision, to allow to merge the network of that system with a future module –the Booster module–. That way, out of the 48 top-level switches, 40 are connected to level-two switches on the DragonFly+ network of the Booster (10 HDR200 links per cell, 1 link per level-two switch). Figure 8 sketches the DragonFly+ network, and the connection to the top-level switches on the Cluster.

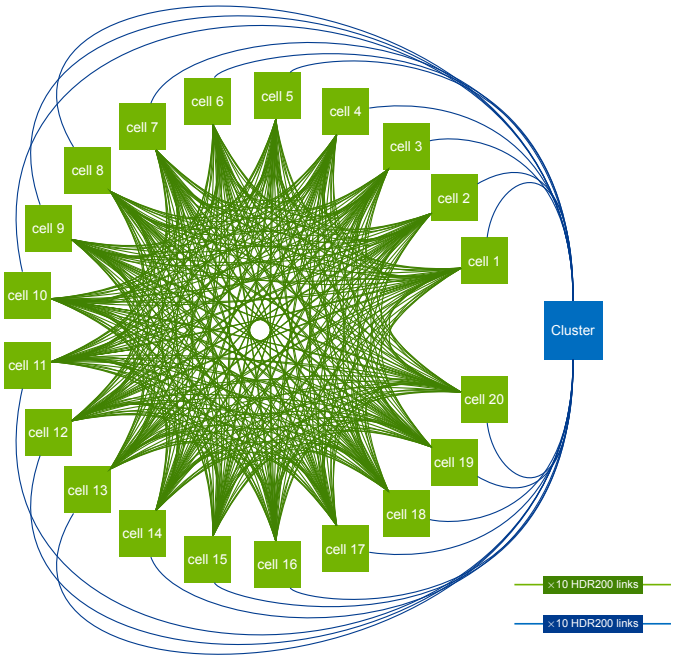


Figure 8: DragonFly+ topology on the Booster and connection to Cluster via top-level switches.

2.3.2 HPST Connection

The last element to describe in the InfiniBand fabric of the JUWELS system is the integration of the High Performance Storage Tier (HPST) I/O subsystem. HPST provides a flash-based buffering layer between the disk-based storage on JUST, and the computes nodes on JUWELS. This layer is optimized for random access and very large bandwidth. A full description will be available in a separate article describing the storage facility at JSC, and is therefore out of scope for this manuscript. However, it is noteworthy that the topology for the HPST part of the fabric is a full Fat Tree, with HDR200 links between switches, and HDR100 links between level-one switches and storage nodes (54 nodes, each one with 2 HDR100 links). The integration on the JUWELS fabric has been realized via the top-level switches in the Cluster Fat Tree, with 1 link per top-level switch (in 42 switches), or 2 links (in 6 switches).

In total, the JUWELS InfiniBand network is a mix of FDR, EDR, HDR100 and HDR200 technologies, has more than 35000 ports, 8 gateways of two different generations, 693 switches, more than 6500 end points, and three routing algorithms are used to enable optimal communication between all these components. That makes it one of the most challenging InfiniBand networks ever deployed. Figure 9 provides a snapshot view of the system, with real connections between all alive components at the time of the snapshot.

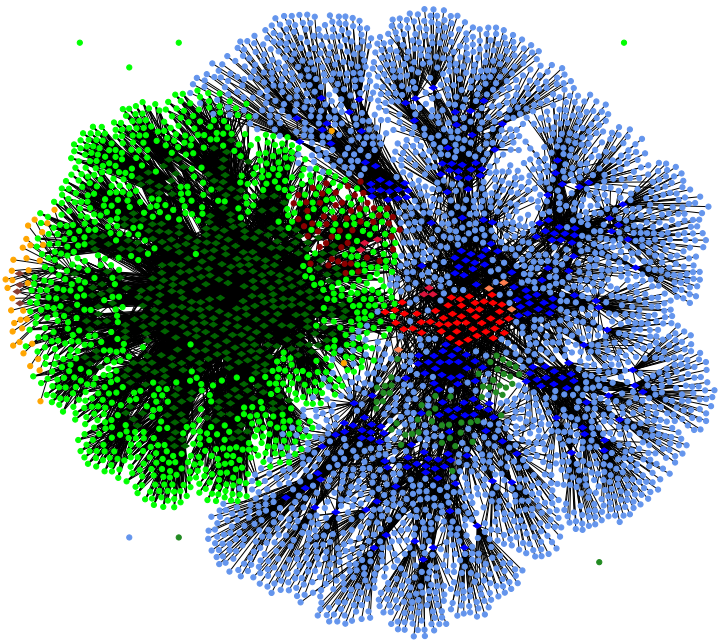


Figure 9: Visualization of the complete InfiniBand Fabric on JUWELS. Light blue dots represent Cluster nodes, dark green dots represent GPU Cluster nodes, light green dots represent Booster nodes, orange dots represent service nodes, dark red dots represent HPST nodes. Additionally: dark blue rhombi represent Cluster switches, light red rhombi represent top-level switches, dark green rhombi represent Booster switches, dark red rhombi represent HPST switches, light orange rhombi represent the Cluster Ethernet gateways, and the brown rhombi represent the Booster Ethernet gateways.

2.4 System Cooling

The JUWELS system is liquid-cooled. All major heat-producing components, such as the processors, GPUs, memory modules and network components are directly liquid-cooled. All compute racks –that is, from Cluster and from Booster– have 3 heat exchangers that couple an external cooling loop with a

rack internal cooling loop. In the Cluster, besides the compute nodes in the rack, the heat exchangers in the base rack also cool the water-cooled switches in the switch rack. In the Booster, since the switches are integrated in the compute racks, each rack is completely autonomous from the cooling point of view.

The cooling needs for Cluster and Booster are slightly different and are operated with different configuration at the facility side. At the moment, the inlet temperature on the facility side is 33°C, with a variable flow rate up to a maximum of 165 cubic meters per hour. On the Booster side, the inlet temperature is at the moment 37°C, also with a variable flow rate up to a maximum of 290 cubic meters per hour. These settings could vary during the life of the system, and will be adapted as needed.

2.5 Service Nodes

An often overlooked part of any HPC system is the set of nodes orchestrating the services needed to run a supercomputer. These nodes are outside of users' reach, but are of critical importance. Due to that, each service is redundantly configured in a minimum of 2 nodes. The nature of the service determines the redundancy strategy:

- Filesystem managers are by design redundant. To manage the filesystem a series of quorum nodes are configured. In order to perform filesystem operations, a majority of these quorum nodes need to be reachable from a given compute node. Due to this majority approach, a minimum of 5 nodes are recommended, as a node failure with a 3 node setup would tend not to reach a majority. Both Cluster and Booster have independent local GPFS clusters that import the filesystem from JUST. For this reason, each module of the system has its own set of 5 quorum nodes.
- Each island have a single island manager active at a given moment. That island manager is in charge of providing TFTP and DHCP services for node booting, DNS for name resolution, NTP for time synchronization, act as gateways for the Ethernet network within the island, and to manage the power and cooling components of the racks. These services are running in a series of containers, that are mounted at a time in a single island manager host. The container images are served from an internal storage cluster and orchestrated with ClusterLabs stack (ClusterLabs, 2021).
- SLURM is a shared resource across the whole system. As such, there is a single instance of the server and database. This instance is running inside a container, which is running in one of the 2 dedicated nodes for job scheduling purposes. The orchestration of this container is done with ClusterLabs stack.
- The InfiniBand network manager is another critical component in the system. Since there is a single InfiniBand network, there is a single subnet manager active at a given time. The design of the subnet manager software allows to have 1 or more passive subnet managers, that will take over in an event of hardware failure on the active subnet manager. In JUWELS there are 2 dedicated nodes for the InfiniBand network manager.
- The so-called master nodes is a node pair that provide access to the system for administrative purposes. On top of this pair there are the master containers, used to access the system, provide access to external networks and act as bootstrap for other administrative nodes. The high availability of these containers is orchestrated also with ClusterLabs stack.
- To monitor the state of the system there are various services deployed in a set of 3 monitoring nodes. These services are part of the ELK stack (Elasticsearch B.V, 2021), Grafana (Grafana Labs, 2021), Prometheus (Prometheus, 2021) and IBMS –an InfiniBand monitoring system from Atos–. All these services live in containers, that are moved between the 3 monitoring nodes as necessary.
- The container images served for some of the services described, are served from an internal storage cluster based on Ceph (Ceph, 2021). The redundancy of this service is built-in in the design of the software. A full description is out of scope of this article.

Table 1 summarizes the services and the redundancy strategy used for each one.

Table 1: Summary of service nodes and their redundancy strategy

	Cluster	Booster	Service Redundancy Strategy
Filesystem quorum	5 nodes	5 nodes	Built-in
Island managers	20 nodes (10 pairs)	8 nodes (4 pair)	System container
	Shared		
Job scheduler	2 nodes (1 pair)		System container
Subnet manager	2 nodes (1 pair)		Built-in
Master node	2 nodes (1 pair)		System container
Monitoring	3 nodes		System/application containers
Service storage cluster	5 nodes		Built-in

2.6 System Software

JUWELS' software stack, as well as the software stack in all the systems operated by JSC, is largely based on open-source software. On the login and compute nodes a CentOS 8 Linux operating system is used. Since the compute nodes are disk-less, only a stripped-down operating system is available on them.

All JSC-operated software is deployed and configured via Ansible (Red Hat Inc, 2021), a software provisioning and configuration management tool. The use of Ansible enables to reuse configuration across nodes and systems easily, and facilitates homogeneity, traceability and reproducibility.

JUWELS uses the open-source Slurm workload manager (SchedMD LLC, 2021) in combination with the ParaStation resource management which has a proven track record in scalability, reliability and performance on several clusters operated by JSC. The ParTec ParaStation ClusterSuite (ParTec Cluster Competence Center GmbH, 2021) is used for node imaging and health monitoring.

The management of the scientific software stack in JUWELS relies on EasyBuild (Hoste et al., 2012). Both modules, Cluster and Booster, have GCC, Intel and NVHPC compilers available. Support for AMD Optimizing C/C++ Compiler (AOCC) on the Booster is in development at the moment. The Message Passing Interface (MPI) implementations supported are mainly ParaStationMPI and OpenMPI, both being CUDA-aware for efficient internode GPU communication. On the Cluster side, IntelMPI is also offered. Different compilers, optimized mathematical libraries and pre-compiled community codes are available. We refer to the JUWELS webpage (Forschungszentrum Jülich, 2021e) for more information. Monitoring of batch jobs is possible using the latest version of the LLview (Forschungszentrum Jülich, 2021f) graphical monitoring tool.

Scientists can also use UNICORE (UNICORE, 2021) to create, submit and monitor jobs on JUWELS. In addition, the system can be accessed via the Jupyter@JSC service (Forschungszentrum Jülich, 2021c).

3 Access to JUWELS

Researchers from institutions within Germany or from an international institution with a significant German participation who are interested in conducting research with JUWELS are eligible to apply for computing time resources via GCS. An adequate proposal may be submitted in answer to corresponding computing time calls published twice a year in January/February and July/August(Gauss Centre for Supercomputing, 2021b).

All applications undergo a comparative peer-review process. The scientific quality and significance of the applications is being evaluated by national and international scientists who are experts in their respective scientific fields. Additionally, the technical feasibility of the applications is ensured by a technical assessment to enable an efficient use of the JUWELS supercomputer. The GCS and the committees of the John von Neumann-Institute (NIC) (John von Neumann Institute for Computing, 2021) are responsible for this process.

Researchers from Europe but outside of Germany who do not fulfill the criteria for the national call are advised of the possibility to apply at PRACE (Partnership for Advanced Computing in Europe, 2021).

Researchers from institutions of the Helmholtz Association working in the research field of Earth and Environment, as well as their national cooperation partners outside of the Helmholtz Association, are eligible to apply for resources in the Earth System Modeling (ESM) partition of JUWELS.

References

Atos. (2021a). *Atos Bullsequana X1000 product webpage.* Retrieved from https://atos.net/en/products/high-performance-computing-hpc/bullsequana-x-supercomputers/bullsequana-x1000

Atos. (2021b). *Atos Bullsequana XH2000 product webpage.* Retrieved from https://atos.net/en/solutions/high-performance-computing-hpc/bullsequana-x-supercomputers#bullsequana-xh2000

Ceph. (2021). *Ceph distributed storage system.* Retrieved from https://ceph.io

ClusterLabs. (2021). *ClusterLabs Stack webpage.* Retrieved from https://clusterlabs.org

Eicker, N., Lippert, T., Moschny, T., & Suarez, E. (2016). The DEEP Project An alternative approach to heterogeneous cluster-computing in the many-core era. *Concurrency and computation*, *28*(8), 2394–2411. http://dx.doi.org/10.1002/cpe.3562

Elasticsearch B.V. (2021). *ELK Stack.* Retrieved from https://www.elastic.co

Forschungszentrum Jülich. (2015). JUQUEEN: IBM Blue Gene/Q Supercomputer System at the Jülich Supercomputing Centre. *Journal of large-scale research facilities*, *1*, A1. http://dx.doi.org/10.17815/jlsrf-1-18

Forschungszentrum Jülich. (2021a). *Forschungszentrum Jülich webpage.* Retrieved from https://www.fz-juelich.de

Forschungszentrum Jülich. (2021b). *Jülich Supercomputing Centre webpage.* Retrieved from https://www.fz-juelich.de/ias/jsc

Forschungszentrum Jülich. (2021c). *Jupyter@JSC webpage.* Retrieved from https://jupyter-jsc.fz-juelich.de

Forschungszentrum Jülich. (2021d). *JUST webpage.* Retrieved from https://www.fz-juelich.de/ias/jsc/just

Forschungszentrum Jülich. (2021e). *JUWELS webpage.* Retrieved from https://www.fz-juelich.de/ias/jsc/juwels

Forschungszentrum Jülich. (2021f). *LLview webpage.* Retrieved from https://www.fz-juelich.de/jsc/llview

Gauss Centre for Supercomputing. (2021a). *Gauss Centre for Supercomputing webpage.* Retrieved from https://www.gauss-centre.eu

Gauss Centre for Supercomputing. (2021b). *HPC Access Gauss Centre for Supercomputing e.V.* Retrieved from https://www.gauss-centre.eu/for-users/hpc-access/

Grafana Labs. (2021). *Grafana: The open observability platform.* Retrieved from https://grafana.com

Helmholtz Association. (2021). *Helmholtz-Gemeinschaft Deutscher Forschungszentren e.V. (HGF) webpage.* Retrieved from https://www.helmholtz.de

High-Performance Computing Center Stuttgart. (2021). *High-Performance Computing Center Stuttgart webpage.* Retrieved from https://www.hlrs.de

Hoste, K., Timmerman, J., Georges, A., & De Weirdt, S. (2012). EasyBuild: Building Software with Ease. In *2012 SC Companion: High Performance Computing, Networking Storage and Analysis* (p. 572-582). http://dx.doi.org/10.1109/SC.Companion.2012.81

IBM. (2021). *IBM Spectrum Scale product webpage.* Retrieved from https://www.ibm.com/de-de/products/spectrum-scale

John von Neumann Institute for Computing. (2021). *John von Neumann Institute for Computing (NIC) webpage.* Retrieved from http://www.john-von-neumann-institut.de

Jülich Supercomputing Centre. (2018). JURECA: Modular supercomputer at Jülich Supercomputing Centre. *Journal of large-scale research facilities*, *4*, A132. http://dx.doi.org/10.17815/jlsrf-4-121-1

Jülich Supercomputing Centre. (2019). JUWELS: Modular Tier-0/1 Supercomputer at the Jülich Supercomputing Centre. *Journal of large-scale research facilities*, *5*, A135. http://dx.doi.org/10.17815/jlsrf-5-171

Leibniz Supercomputing Centre of the Bavarian Academy of Sciences, & Humanities. (2021). *Leibniz Supercomputing Centre webpage.* Retrieved from https://lrz.de

ParTec Cluster Competence Center GmbH. (2021). *ParTec webpage.* Retrieved from https://www.par-tec.com

Partnership for Advanced Computing in Europe. (2021). *Partnership for Advanced Computing in Europe webpage.* Retrieved from https://www.prace-ri.eu

Prometheus. (2021). *Prometheus - Monitoring system & time series database.* Retrieved from https://prometheus.io/

Red Hat Inc. (2021). *Ansible Configuration Manager webpage.* Retrieved from https://www.ansible.com

SchedMD LLC. (2021). *Slurm Workload Manager webpage.* Retrieved from https://slurm.schedmd.com

Top500. (2021). *Top500 June 2021 list.* Retrieved from https://www.top500.org/lists/2021/06

UNICORE. (2021). *Uniform Interface to Computing Resources (UNICORE) webpage.* Retrieved from https://www.unicore.eu