Journal of large-scale research facilities

Journal of large-scale research facilities, 5, A135 (2019) http://dx.doi.org/10.17815/jlsrf-5-171

Published: 06.02.2019

JUWELS: Modular Tier-0/1 Supercomputer at the Jülich Supercomputing Centre

Forschungszentrum Jülich, Jülich Supercomputing Centre *

Instrument Scientists:
- Supercomputing Support, Jülich Supercomputing Centre, Forschungszentrum Jülich, phone: +49(0)2461 61 282, 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 within the Helmholtz Association. The first module deployed in 2018, is a Cluster module based on the Bull-Sequana X1000 architecture with Intel Xeon Skylake-SP processors and Mellanox EDR InfiniBand. An extension by a second Booster module is scheduled for deployment in 2020.

1 Introduction

Since summer 2018, the supercomputer JUWELS (Jülich Wizard for European Leadership Science) is the latest leadership-class system operated by the Jülich Supercomputing Centre (JSC) at the Forschungszentrum Jülich (Forschungszentrum Jülich, 2019a) as a member of the Gauss Centre for Supercomputing (Gauss Centre for Supercomputing, 2019). JUWELS is accessible to German and European scientists as a national tier-1 and European tier-0 system, respectively. It succeeds the successful JUQUEEN IBM Blue Gene/Q system (Forschungszentrum Jülich, 2015) which was operated from 2012 to 2018. JUWELS is designed as a modular supercomputer that combines multiple architecturally diverse computing modules into a single system with software technology that enables uniform and concurrent use of the different resources. In 2018, the first module, a Cluster system with multi-core processors, was deployed. An extension with a second Booster module, utilizing processor technologies targeted specifically at massively parallel workloads, is scheduled for deployment in 2020. The traits of the Cluster and Booster hardware augment each other in terms of their versatility and performance characteristics.

* **Cite article as:** 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

Figure 1: The Cluster Module of the supercomputer JUWELS in the facility of Jülich Supercomputing Centre. Copyright: Forschungszentrum Jülich GmbH.

The JUWELS investment and operational costs 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 by the Helmholtz Association (Helmholtz Association, 2019) allows to enlarge the system by a share dedicated to the Earth system modelling community within the Helmholtz Association.

1.1 Gauss Centre for Supercomputing and PRACE

The Gauss Centre for Supercomputing (GCS) combines the three national supercomputing centers High-Performance Computing Center Stuttgart (HLRS) (High-Performance Computing Center Stuttgart, 2019), the Jülich Supercomputing Centre (JSC) (Forschungszentrum Jülich, 2019b), and the Leibniz Supercomputing Centre (LRZ) (Leibniz Supercomputing Centre of the Bavarian Academy of Sciences & Humanities, 2019) into Germany's foremost supercomputing institution. 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. GCS is a hosting member of the pan-European Partnership for Advanced Computing in Europe (PRACE) (Partnership for Advanced Computing in Europe, 2019) organization. PRACE has 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 the 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. The initial deployment of JUWELS in 2018 consists of the Cluster module that utilizes latest generation

Intel Xeon processors and provides a familiar and versatile environment in support of a broad class of computational workloads. The Booster module, will significantly amplify the performance of a more narrow class of workloads that can effectively leverage massively-parallel architectures.

2.1 Cluster Module

The JUWELS Cluster is a BullSequana X1000 supercomputer. The Sequana X1000 series (Atos, 2019) by Atos provides a high-density node integration with warm-water direct-liquid cooling capabilities. It follows a scalable hierarchical cell-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. The compute nodes are interconnected with a Mellanox InfiniBand Extended Data Rate (EDR) 100 Gb/s high-speed network for message passing and storage access.

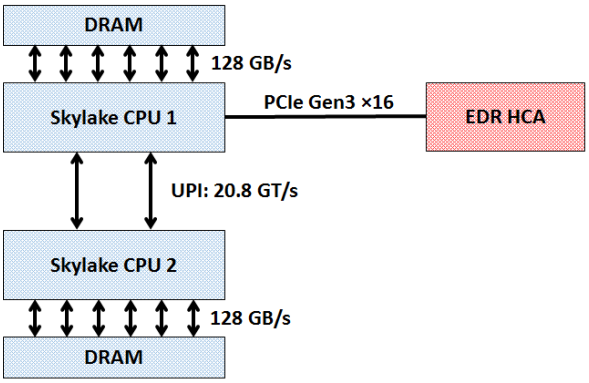
The 2,511 computes nodes in the CPU-only partition of JUWELS are equipped with two Intel Xeon Skylake Platinum 8168 central processing units (CPUs) with 24 cores each and a base frequency of 2.7 GHz. 90% (2,271) of these compute nodes feature 96 GB main memory, the remaining 240 nodes offer a main memory 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. Three nodes are collocated in a single Sequana X1120 compute blade.

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.

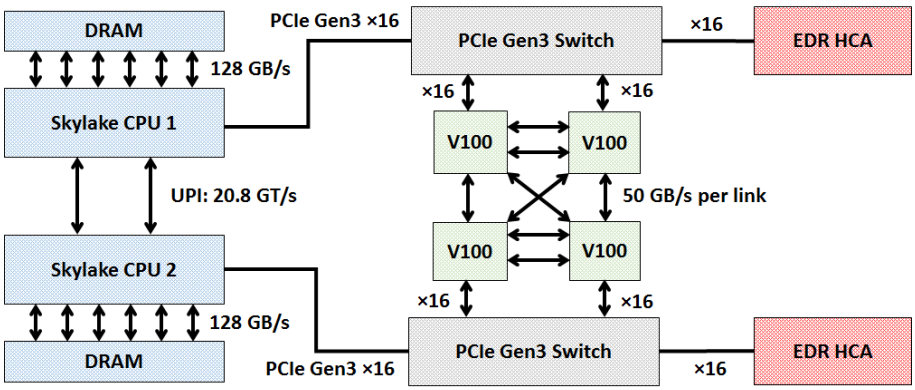
The GPU partition in the JUWELS Cluster consists of 48 compute nodes based on the BullSequana X1125 accelerator blade. The nodes are equipped with two Intel Xeon Gold 6148 processors with 20 cores each and 192 GB main memory. Each node contains four NVIDIA Volta V100 GPUs in SXM2 form factor with 5,120 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 the CPU with PCIe Gen 3 (16 lanes) and to each other with one to two NVIDIA NVLink2 links with a bi-directional peak bandwidth of 50 GB/s. The nodes are equipped with two Mellanox ConnectX-4 HCAs. 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.

In addition to the compute nodes, JUWELS 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.



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



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

Figure 3: Annotated block diagrams of JUWELS Cluster compute nodes.

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), each level one switch connects downwards to 24 compute nodes and 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 one level one switch (16 links) and twelve uplinks between level one and level two are present, i.e., a 3:4 pruning is applied. 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. 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 is scheduled for the first quarter of 2019.

The JUWELS login and compute nodes access the parallel file systems exported by the central Jülich Storage cluster JUST (Forschungszentrum Jülich, 2019d) using the IBM Spectrum Scale (formerly GPFS) file system software (IBM, 2019). 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 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 three gateway switches. Depending on job distribution and concurrent I/O load, different I/O bandwidth values will be observed under production conditions.

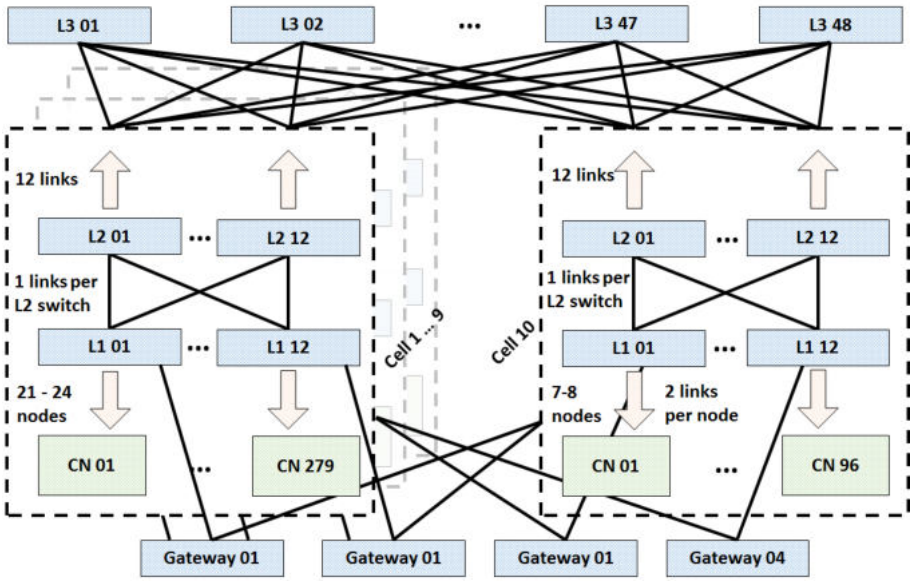


Figure 4: Schematic of the JUWELS InfiniBand network topology. Please note that the 10th cell is only half populated in the current deployment. Each cell is connected to each gateway switch with at least one link. Various service nodes that are connected to free ports in the level one switches are not shown.

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 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.

The JUWELS Cluster module is liquid cooled. All major heat producing components, such as the processors, GPUs, memory modules and network components are directly liquid cooled. The system is operated with an inlet water temperature of up to 34 °C in the primary facility water loop. The secondary closed in-rack loop operates at ca. +4 K difference. The high inlet and outlet water temperatures in the primary loop allow for power-saving free cooling against the outside air. JSC will vary water temperatures depending on environmental conditions in order to optimize operational conditions. Since Intel Xeon processors are designed to maintain a specific thermal design power, changes in the cooling parameters may slightly affect application performance for highly optimized applications.

2.2 Booster Module

An extension of the JUWELS system by a highly scalable Booster module is planned for 2020. Further details will be provided in an update of this article once the Booster module enters its production phase.

2.3 Software

JUWELS' software stack is largely based on open-source software. On the login and compute nodes a CentOS 7 Linux operating system is used. Since the compute nodes are disk-less, only a stripped down operating system is available on the compute nodes.
JUWELS uses the open-source Slurm workload manager (SchedMD LLC, 2019) 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, 2019) is used for system provisioning and health monitoring. The ParTec ParaStation software, including ParaStation MPI, are the foundation for the modular supercomputing architecture at JSC.
The Intel and ParTec ParaStation Message Passing Interface (MPI) implementations are supported. In addition the CUDA-aware MPI implementation MVAPICH2-GDR is available for mixed MPI+CUDA applications. Different compilers, optimized mathematical libraries and pre-compiled community codes are available. We refer to the JUWELS webpage (Forschungszentrum Jülich, 2019e) for more information. Monitoring of batch jobs is possible using the latest version of the LLview (Forschungszentrum Jülich, 2019f) graphical monitoring tool.
Scientists can also use UNICORE (UNICORE Forum e.V., 2019) to create, submit and monitor jobs on JUWELS. In addition, the system can be accessed via the Jupyter@JSC service (Forschungszentrum Jülich, 2019c).

3 Access to JUWELS

Scientists and engineers interested in using JUWELS for their research have to apply for JUWELS computing time resources by submitting an adequate proposal in answer to corresponding computing time calls published in January/February and July/August every year. Submitted proposals are evaluated scientifically through a competitive peer-review process. Additionally, the review process includes a technical assessment of the project regarding the efficient use of a large number of nodes on JUWELS for the proposed simulations.

Scientists and engineers whose affiliation is in Germany (or a foreign office of a German institution) and scientists working in an international institution with significant German participation can apply for computing time via the Gauss Centre for Supercomputing. Projects requesting at least 35 million core hours are classified as GCS Large Scale projects and are peer-reviewed by a committee of the GCS. Projects with up to 35 million core hours are peer-reviewed by the John von Neumann Institute for Computing (John von Neumann Institute for Computing, 2019) (NIC) on behalf of the GCS. The NIC is a joint organization of the three Helmholtz centers Forschungszentrum Jülich, Deutsches Elektronen-Synchroton (Deutsches Elektronen Synchrotron, 2019) DESY and the GSI Helmholtzzentrum für Schwerionenforschung (GSI Helmholtzzentrum für Schwerionenforschung, 2019).
European researchers that are not eligible for the national calls can apply for computing time on JUWELS via PRACE (Partnership for Advanced Computing in Europe, 2019).
Scientists from an institution of the Helmholtz Association working in the research field Earth and Environment, and their national cooperation partners outside of the Helmholtz Association as well, are eligible to apply for resources in the Earth System Modeling (ESM) partition of JUWELS.

References

Atos. (2019). *Atos Bullsequana X1000 product webpage.* Retrieved from https://atos.net/en/products/high-performance-computing-hpc/bullsequana-x-supercomputers/bullsequana-x1000

Deutsches Elektronen Synchrotron. (2019). *Deutsches Elektronen-Synchrotron (DESY) webpage.* Retrieved from http://www.desy.de

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

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. (2019a). *Forschungszentrum Jülich webpage.* Retrieved from http://www.fz-juelich.de

Forschungszentrum Jülich. (2019b). *Jülich Supercomputing Centre webpage.* Retrieved from http://www.fz-juelich.de/ias/jsc

Forschungszentrum Jülich. (2019c). *Jupyter@JSC webpage.* Retrieved from http://jupyter-jsc.fz-juelich.de

Forschungszentrum Jülich. (2019d). *JUST webpage.* Retrieved from http://www.fz-juelich.de/ias/jsc/just

Forschungszentrum Jülich. (2019e). *JUWELS webpage.* Retrieved from http://www.fz-juelich.de/ias/jsc/juwels

Forschungszentrum Jülich. (2019f). *LLview webpage.* Retrieved from http://www.fz-juelich.de/jsc/llview

Gauss Centre for Supercomputing. (2019). *Gauss Centre for Supercomputing webpage.* Retrieved from http://www.gauss-centre.eu

GSI Helmholtzzentrum für Schwerionenforschung. (2019). *GSI Helmholtzzentrum für Schwerionenforschung webpage.* Retrieved from http://www.gsi.de

Helmholtz Association. (2019). *Helmholtz-Gemeinschaft Deutscher Forschungszentren e.V. (HGF) webpage.* Retrieved from http://www.helmholtz.de

High-Performance Computing Center Stuttgart. (2019). *High-Performance Computing Center Stuttgart webpage.* Retrieved from http://www.hlrs.de

IBM. (2019). *IBM Spectrum Scale product webpage.* Retrieved from https://www.ibm.com/de-en/marketplace/scale-out-file-and-object-storage

John von Neumann Institute for Computing. (2019). *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

Leibniz Supercomputing Centre of the Bavarian Academy of Sciences, & Humanities. (2019). *Leibniz Supercomputing Centre webpage.* Retrieved from https://lrz.de

ParTec Cluster Competence Center GmbH. (2019). *ParTec webpage.* Retrieved from http://www.par-tec.com

Partnership for Advanced Computing in Europe. (2019). *Partnership for Advanced Computing in Europe webpage.* Retrieved from http://www.prace-ri.eu

SchedMD LLC. (2019). *Slurm Workload Manager webpage.* Retrieved from http://slurm.schedmd.com

UNICORE Forum e.V. (2019). *Uniform Interface to Computing Resources (UNICORE) webpage.* Retrieved from http://www.unicore.eu