Journal of large-scale research facilities

Journal of large-scale research facilities, 4, A132 (2018) http://dx.doi.org/10.17815/jlsrf-4-121-1

Published: 25.07.2018

JURECA: Modular supercomputer at Jülich Supercomputing Centre

Forschungszentrum Jülich, Jülich Supercomputing Centre *

Instrument Scientists:
- Dorian Krause, Jülich Supercomputing Centre, Forschungszentrum Jülich, phone: +49(0)2461 61 3631, email: d.krause@fz-juelich.de
- Philipp Thörnig, Jülich Supercomputing Centre, Forschungszentrum Jülich, phone: +49(0)2461 61 1472, email: p.thoernig@fz-juelich.de

Abstract: JURECA is a petaflop-scale modular supercomputer operated by Jülich Supercomputing Centre at Forschungszentrum Jülich. The system combines a flexible Cluster module, based on T-Platforms V-Class blades with a balanced selection of best of its kind components, with a scalability focused Booster module, delivered by Intel and Dell EMC based on the Xeon Phi many-core processor. With its novel architecture, it supports a wide variety of high-performance computing and data analytics workloads.

1 Introduction

Since July 2015, the Jülich Supercomputing Centre (JSC) at the Forschungszentrum Jülich (Forschungszentrum Jülich, 2018a) operates the JURECA (Jülich Research on Exascale Cluster Architectures) system as the successor of the popular JUROPA (Jülich Research on Petaflop Architectures) supercomputer. From 2015 to 2017, the JURECA Cluster (see Figure 1) served as a general-purpose supercomputing resource and, in accordance with Forschungszentrum Jülich's dual architecture strategy, augmented the leadership-class highly scalable IBM Blue Gene/Q system JUQUEEN (Forschungszentrum Jülich, 2015). In 2017, JURECA was itself augmented with a many-core processor based Booster module to enable highly scalable applications to leverage the system more efficiently. Funding for both JURECA modules was granted by the Helmholtz Association (Helmholtz Association, 2018) through the program "Supercomputing & Big Data".
The Cluster and Booster module are tightly integrated and operated as a single system following the modular supercomputing paradigm, pioneered by JSC in the context of the DEEP series of EU-funded

* **Cite article as:** 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

projects (Eicker et al., 2016). The modular supercomputing concept enables users to distribute their workloads flexibly across different, architecturally diverse, modules in order to place different phases or subroutines of their workload on the hardware best suited for the execution. The software features required to leverage this architecture is made available in the course of 2018 for the wider JURECA user community.

Figure 1: Jülich Research on Exascale Cluster Architectures (JURECA) at Jülich Supercomputing Centre. The left picture shows the Cluster module in 2015. The right picture shows the complete system after the Booster deployment in 2017. Copyright: Forschungszentrum Jülich.

The JURECA Cluster module was designed by JSC together with the hardware vendor T-Platforms (T-Platforms, 2018) to serve as a versatile scientific instrument for compute- and data-intense (simulation) science that is equally suited for capacity as for capability workloads. The JURECA Booster module was designed by JSC and Intel (Intel Corperation, 2018) in 2016 as a highly-scalable compute architecture leveraging the latest available Intel networking and processor technology. The system was delivered by Intel together with its partner Dell EMC (Dell EMC, 2018) in 2017.

2 JURECA system details

The JURECA modular supercomputer consists of two separate, but tightly integrated, compute modules. The architecture of the Cluster module is an evolution of the JUROPA architecture and follows the best-of-breed approach by combining the most advanced commodity hardware and software technologies available in the industry. The JURECA Cluster is itself a heterogeneous system offering nodes with different memory sizes (128 GiB, 256 GiB, 512 GiB as well as 1 TiB), nodes with graphics processing unit (GPU) accelerators as well as GPU-equipped nodes for visualization and other post-processing needs. The architecture of the Booster module is designed to best serve highly-scalable simulation workloads that are able to leverage the high core counts and wide vector units of the Intel Xeon Phi many-core processors.

2.1 Cluster module

The JURECA Cluster consists of 1,733 compute nodes of type T-Platforms V-Class V210S as well 75 GPU-accelerated V210F blades hosted in V5050 chassis (see Figure 2). Moreover, 64 Supermicro F618R2-RT+ twin-blade servers (512 GiB memory nodes) and 12 Supermicro 1028GR-TR visualization nodes are available.
All systems feature two Intel Xeon E5-2680 v3 12-core Haswell central processing units (Intel Corporation, 2018b) (CPUs) which support up to 24 hardware threads each. Each CPU supports the AVX 2.0 instruction set architecture extension and can perform two 256-bit (i.e., four double precision floating point numbers) wide multiply-add operations per cycle . The peak performance of a (non-accelerated) JURECA Cluster node is 0.96 TFlop/s. The maximum memory bandwidth of the node is 136 GB/s.

The two sockets are connected by a bi-directional 9.6 GT/s (Gigatransfers per second) Intel Quick Path Interconnect (QPI) link. In the JURECA Cluster module, 2133 MHz DDR4 memory technology is used. Applications that support the use of GPU accelerators can take advantage of the additional two NVIDIA K80 graphics processing units available in 75 compute nodes. The GPUs are connected with PCI Express Generation 3.0 (16 lanes) links providing a peak of 32 GB/s bidirectional host-device bandwidth. Each K80 GPU is equipped with 2×12 GB GDDR5 memory and offers 4992 CUDA cores that provide an additional 2.9 TFlop/s peak performance (5.8 TFlop/s per node) and 480 GB/s memory bandwidth per GPU. The 12 visualization nodes are equipped with two NVIDIA K40 GPUs intended for remote visualization usage.
The JURECA Cluster compute nodes are connected with Mellanox extended data rate (EDR) InfiniBand providing 100 Gb/s (12.5 GB/s) link bandwidth and MPI latencies around one microsecond. The host channel adapters (HCA) are connected via PCI Express Generation 3.0 (16 lanes). The Cluster components are interconnected in a three-level full fat tree topology which provides full bisection bandwidth and non-blocking communication for appropriate communication patterns.
A particular emphasis during the design of JURECA Cluster has been put on the storage connection in order to meet the increasing data requirements of simulation sciences as well as the needs of emerging data-intense sciences. All offered global (parallel) filesystems on JURECA are mounted from the central Jülich Storage Cluster (JUST) (Forschungszentrum Jülich, 2018c) using IBM's General Parallel Filesystem (GPFS). Users with access to several systems in the supercomputing facility at JSC work with the same filesystems on all systems so that data movement is minimized and workflows are simplified. The storage network connection of the Cluster is realized using InfiniBand-to-Ethernet gateways bridging the internal InfiniBand network with the facility's Terabit Ethernet backbone. This connection type was selected as it allows for >100 GB/s aggregate filesystem bandwidth as well as a high per-node filesystem performance that is hardware-wise only limited by the performance of the fourteen data rate (FDR) InfiniBand links (56 Gb/s) towards the gateways.

2.2 Booster module

The JURECA Booster consists of 1,6400 compute nodes of type Dell PowerEdge C6320P (see Figure 3). All systems feature one Intel Xeon Phi 7250-F CPU (Intel Corporation, 2018a) with 68 cores, a base frequency of 1.4 GHz, and 4 hardware threads per physical core. The processor package includes 16 GiB of high-bandwidth, multi-channel DRAM (MCDRAM) with a bandwidth of up to 500 GB/s. The peak performance of a Booster compute node is 3 TFlop/s. Each node is equipped with additional 96 GiB DDR4 memory clocked at 2400 MHz.
The Booster compute nodes are connected with 100 Gb/s Intel Omni-Path Architecture (OPA). The host fabric interfaces (HFI) are integrated in the CPU on the package but internally connected with PCI Express Generation 3.0 (16 lanes). The Booster components are interconnected in a three-level full fat tree topology. The Cluster and Booster modules are linked through 198 router nodes equipped with one InfiniBand HCA and one OPA HFI, enabling Cluster-Booster communication with up to 19.8 Tb/s (2.5 TB/s) bandwidth.
The Booster connects to the JUST cluster to access the same file systems as are available on the Cluster. The storage connection is realized with 26 router nodes equipped with two HFI ports and two 40 Gigabit Ethernet connections to the facility Ethernet fabric. The nominal network speed of the storage connection is 260 GB/s. In practice, due to software limitations a lower performance is observed.

2.3 Software

JURECA's software stack is largely based on open-source software. Login and compute nodes run the CentOS 7 Linux operating system with a careful setup that balances the ease of use and low entrance-barrier with the requirements, such as minimal operating system jitter, of large-scale capability clusters. JURECA uses the open-source Slurm workload manager (SchedMD LLC, 2018) in combination with

(a) Back (left) and side (right) view of a T-Platforms V-Class V210S dual-socket blade server as used in JURECA. The GPU-accelerated V210F blades host two additional PCIe devices and fit in two chassis slots.

(b) Front (left) and back (right) view of the T-Platforms V5050 chassis. Each chassis can host ten V210S or, alternatively, five V210F blades.

Figure 2: T-Platforms V-Class components used in the JURECA system. Copyright: T-Platforms.

Figure 3: Example of a Dell C6320P server system. The model used in the JURECA Booster slightly deviates from the shown version due to the utilized processor type. Copyright: Dell Technologies.

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, 2018) is used for system provisioning and health monitoring.
On JURECA, 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 JURECA webpage (Forschungszentrum Jülich, 2018b) for more information. Monitoring of batch jobs is possible using the latest version of the LLview (Forschungszentrum Jülich, 2018d) graphical monitoring tool.
Scientists can also use UNICORE (UNICORE Forum e.V., 2018) to create, submit and monitor jobs on the JURECA system.
The software functionality required for high-speed communication between Cluster and Booster via MPI is implemented in ParaStation. At the time of the Booster deployment in 2017 the software was available at proof-of-concept level. It is matured in the course of the year 2018 and is made available, along with the necessary enhancement of the workload manager, in steps to the wider JURECA community.

2.4 Hardware components

As of this writing, JURECA consists of the following hardware components. An up-to-date description of the hardware (and software) configuration of the system is maintained on the JURECA webpage (Forschungszentrum Jülich, 2018b).

2.4.1 Cluster module

- 34 racks organized in four rows
  - 1,872 compute nodes
    - 2× Intel Xeon E5-2680 v3 Haswell CPUs per node
      - 2× 24 cores, 2.5 GHz base frequency
      - Intel Hyper-Threading Technology
      - AVX 2.0 instruction set architecture extension
    - DDR4 memory technology clocked at 2133 MHz (8 channels)
      - 128 GiB memory in 1,680 nodes
      - 256 GiB memory in 128 nodes
      - 512 GiB memory in 64 nodes
    - 75 nodes equipped with 2× NVIDIA K80 GPUs each
      - 2× 4992 CUDA cores
      - 2× 24 GB GDDR5 memory
  - 12 visualization nodes
    - 2× Intel Xeon E5-2680 v3 Haswell CPUs per node
    - DDR4 memory technology clocked at 2133 MHz
      - Memory size of 1 TiB in 2 nodes
    - 2× NVIDIA K40 GPUs
      - 2× 12 GB GDDR5 memory
- Mellanox InfiniBand EDR network organized in a three-level full-fat tree topology
  - Mellanox ConnectX-4 single port host channel adapters in nodes
  - 36-port SwitchIB-based Mellanox SB7790 leaf-level switches
  - 4× SwitchIB-based Mellanox CS7500 core switches
  - 2× Mellanox SX6036G InfiniBand FDR/40 Gigabit Ethernet gateways for storage connection

2.4.2 Booster module

- 33 racks organized in three rows
  - 1,640 compute nodes
    - Intel Xeon Phi "Knights Landing" 7250-F CPU
      - 68 cores, 1.4 GHz base frequency
      - AVX-512 instruction set architecture extension
    - 16 GiB multi-channel DRAM (MCDRAM)
    - 96 GiB DDR4 memory clocked at 2400 MHz (6 channels)
  - 26 storage router nodes
    - Dual-port Intel Omni-Path host fabric interface cards
    - 2×40 Gigabit Ethernet connection to facility fabric
- Intel Omni-Path Architecture network organized in a three-level full-fat tree topology
  - On-package Omni-Path host fabric interface
  - 48-port Intel Omni-Path Edge Switch 100 switches
  - 3× Intel Omni-Path Director Class Switch 100 core switches

2.4.3 Joint infrastructure

- 12 login nodes
  - 2× Intel Xeon E5-2680 v3 Haswell CPUs per node
  - 256 GiB DDR4 memory
- 24 service nodes for system management
- 198 Cluster-Booster bridge nodes
  - 1× Mellanox ConnectX-4 single port host channel adapter directly connected to core switch line cards
  - 1× Intel Omni-Path host fabric interface card connected to edge switches

2.5 Software components

- CentOS 7 enterprise-grade Linux operating system
- ParTec ParaStation ClusterSuite
- Slurm batch system with ParaStation resource management
- Intel and ParTec ParaStation Message Passing Interface implementations
- Support for OpenMP, NVIDIA CUDA, OpenCL and OpenACC programming models

2.6 Benchmark results

Using 1,764 JURECA Cluster compute nodes without accelerators a Linpack performance of 1.42 PFlop/s was measured, placing the system on spot 50 in the November 2015 Top500 list (Top500, 2015). The Cluster module consumed on average 825 kW during the Linpack run, i.e., about 1.72 GFlop/s/W. JURECA entered the Green500 list in November 2015 on place 112 (Green500, 2015). On the High Performance Conjugate Gradients (HPCG) benchmark, JURECA Cluster achieved 68.3 TFlop/s in 2015 corresponding to place 18 in the November 2015 HPCG list (HPCG, 2015).
In 2017, following the installation of the Booster module, a combined Linpack performance of 3.78 PFlop/s was measured with 1,760 Cluster and 1,600 Booster compute nodes. The upgrade placed the system on spot 29 in the November 2017 Top500 list (Top500, 2017). With an average 2.81 GFlop/s/W, the system ranked on spot 55 in the Green500 list in November 2017 (Green500, 2017).

3 Access to JURECA

Scientists and engineers interested in using the capacities and capabilities of JURECA for their research have to apply for JURECA compute time resources by submitting an adequate proposal in answer to corresponding compute time calls published January and July every year. Submitted proposals are evaluated scientifically through a competitive peer-review process. Additionally, the review process includes a technical assessment of the applicant's ability to efficiently perform parallel computations utilizing a larger number of compute cores on JURECA.
Basically, there are two calls available twice every year: One is conducted jointly by peers in computational science and engineering at Forschungszentrum Jülich and RWTH Aachen University, accepting proposals from the two institutions only (so-called JARA-HPC/VSR Call) (Jülich-Aachen Research Alliance, 2018). The other one (NIC Call) is performed by the John von Neumann Institute for Computing (John von Neumann Institute for Computing, 2018) (NIC), a joint organization of the three Helmholtz centers Forschungszentrum Jülich, Deutsches Elektronen-Synchroton (Deutsches Elektronen Synchrotron, 2018) DESY and the GSI Helmholtzzentrum für Schwerionenforschung (GSI Helmholtzzentrum für Schwerionenforschung, 2018), accepting proposals from all other German universities and research institutions. Applicants have to demonstrate that they are qualified in their respective field and that they have an appropriate knowledge in high-performance computing.
Scientists with challenging compute- or data-intense scientific problems that require access to JURECA in order to lay the necessary software foundation for the preparation of a successful proposal can obtain a limited compute time budget on JURECA along with expert support by a JSC simulation lab (Forschungszentrum Jülich, 2018f) by answering the bi-annual call for preparatory access and support resources (Forschungszentrum Jülich, 2018e).
Between 2015 and 2018, JURECA Cluster compute time was available for all eligible scientists via NIC Calls. Starting from 2018, only the JURECA Booster module is made available via the national NIC Call for an interim period until approximately 2020. Compute time on the Cluster is only available via the JARA-HPC/VSR Call or for NIC users that can leverage the Cluster and Booster concurrently.

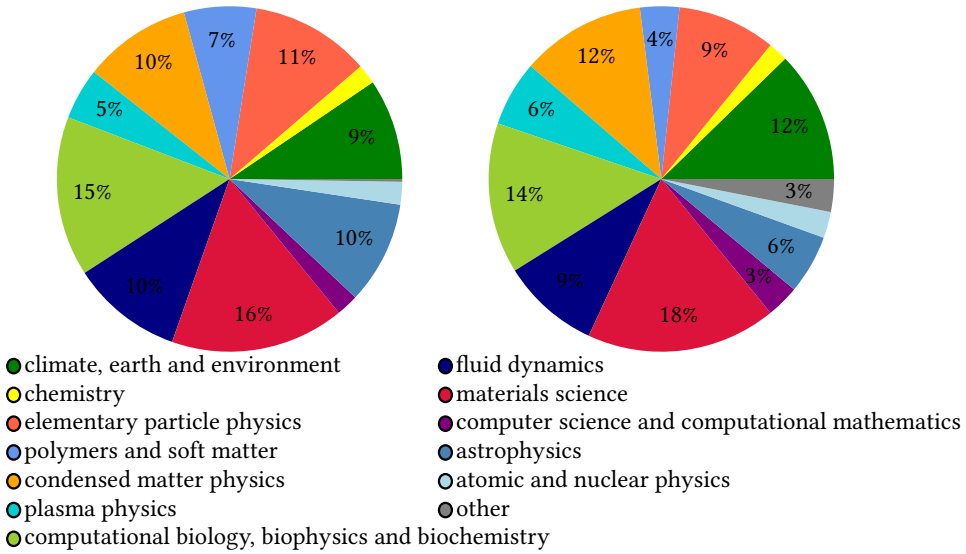

Figure 4: Allocated compute time (left) and number of projects (right) on JURECA by scientific field in the computing time period from the 1st of November 2015 to the 30th April 2016. Percentages are shown for shares above 3 %.

4 Application fields

In order to exemplify the versatility of JURECA, Figure 4 shows the allocated computing time and number of projects working on JURECA in the time frame November 2015 to April 2016, i.e., prior to the Booster addition. In this time frame, about 160 million core hours were allocated on the Cluster module to nearly 200 projects working on advanced research in different disciplines ranging from the basic sciences to engineering applications. The employed parallel applications in different projects and scientific communities range from highly-optimized applications, which are particularly tuned for JURECA's CPU architecture, to complicated multi-program simulation frameworks that are not uncommonly driven by dynamic scripting languages. The demands for hardware features (such as modern accelerators), main memory sizes, pre- and post-processing capabilities as well as input/output (I/O) performance vary similarly. JURECA's hardware design and software stack ensure that these requirements are met across the whole application spectrum.

References

Dell EMC. (2018). *Dell EMC.* Retrieved from http://www.dellemc.com

Deutsches Elektronen Synchrotron. (2018). *Deutsches Elektronen-Synchrotron (DESY).* 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. (2018a). *Forschungszentrum Jülich.* Retrieved from http://www.fz-juelich.de

Forschungszentrum Jülich. (2018b). *JURECA webpage.* Retrieved from http://www.fz-juelich.de/ias/jsc/jureca

Forschungszentrum Jülich. (2018c). *JUST webpage.* Retrieved from http://www.fz-juelich.de/ias/jsc/just

Forschungszentrum Jülich. (2018d). *LLview webpage.* Retrieved from http://www.fz-juelich.de/jsc/llview

Forschungszentrum Jülich. (2018e). *Preparatory Access to Computing and Support Resources.* Retrieved from http://www.fz-juelich.de/ias/jsc/prep-access.html

Forschungszentrum Jülich. (2018f). *Simulation Laboratories at Jülich Supercomputing Centre.* Retrieved from http://www.fz-juelich.de/ias/jsc/simlabs

Green500. (2015). *Green500 November 2015 list.* Retrieved from http://www.top500.org/green500/list/2015/11

Green500. (2017). *Green500 November 2017 list.* Retrieved from http://www.top500.org/green500/list/2017/11

GSI Helmholtzzentrum für Schwerionenforschung. (2018). *GSI Helmholtzzentrum für Schwerionenforschung.* Retrieved from http://www.gsi.de

Helmholtz Association. (2018). *Helmholtz-Gemeinschaft Deutscher Forschungszentren e.V. (HGF).* Retrieved from http://www.helmholtz.de

HPCG. (2015). *HPCG November 2015 list.* Retrieved from http://www.hpcg-benchmark.org

Intel Corperation. (2018). *Intel Corperation.* Retrieved from http://www.intel.com

Intel Corporation. (2018a). *Intel Xeon Phi Processor 7250-F.* Retrieved from https://ark.intel.com/products/94035/Intel-Xeon-Phi-Processor-7250-16GB-1_40-GHz-68-core

Intel Corporation. (2018b). *Intel Xeon Processor E5-2680 v3.* Retrieved from http://ark.intel.com/products/81908/Intel-Xeon-Processor-E5-2680-v3-30M-Cache-2_50-GHz

John von Neumann Institute for Computing. (2018). *John von Neumann Institute for Computing (NIC).* Retrieved from http://www.john-von-neumann-institut.de

Jülich-Aachen Research Alliance. (2018). *Jülich-Aachen Research Alliance – High-Performance Computing (JARA-HPC).* Retrieved from http://www.jara.org/de/research/jara-hpc

ParTec Cluster Competence Center GmbH. (2018). *ParTec webpage.* Retrieved from http://www.par-tec.com

SchedMD LLC. (2018). *Slurm Workload Manager webpage.* Retrieved from http://slurm.schedmd.com

Top500. (2015). *Top500 November 2015 list.* Retrieved from http://www.top500.org/lists/2015/11

Top500. (2017). *Top500 November 2017 list.* Retrieved from http://www.top500.org/lists/2017/11

T-Platforms. (2018). *T-Platforms.* Retrieved from http://www.t-platforms.com

UNICORE Forum e.V. (2018). *Uniform Interface to Computing Resources (UNICORE) webpage.* Retrieved from http://www.unicore.eu