JURECA: Jülich Research on Exascale Cluster Architectures

Figure 1: (Left) Front view of a V-class enclosure. (Right) Rear view of a V-class enclosure hosting ten V210S compute nodes (slim or fat nodes of type 1 in JURECA). The same chassis can host five V210F accelerated compute nodes occupying twice the space of V210S node.

According to its dual architecture strategy, Forschungszentrum Jülich is offering access to a leadership-class capability system and to a general-purpose supercomputing system. The latter meets the users' need for mixed capacity and capability computing time. Since 2009, the JUROPA (Jülich Research on Petaflop Architectures) cluster, based on Intel Nehalem CPUs and quad data rate InfiniBand networking technology, has taken up this challenge and has enabled outstanding science by researchers from around the world. Now, Jülich Supercomputing Center (JSC) has started to install the next-generation general-purpose cluster JURECA (Jülich Research on Exascale Cluster Architectures) which will supersede JUROPA.

JURECA will be based on latest-generation Intel Haswell CPUs and provide a peak floating point performance of roughly 1.8 petaflops per second – a six-fold increase over JUROPA. With its high speed connection of about 100 GiBps to the center-wide exported GPFS file systems, JURECA will not only serve the widest variety of user communities from the traditional computational science disciplines but will also be a welcoming home to data-intense science and big data projects.

JURECA will be built from V-class blade servers of the Russian supercomputing vendor T-Platforms. Once fully installed, the system will consist of more than 1,800 compute nodes with two Intel Haswell E5-2680 v3 12 core CPUs per node. About 1,680 compute nodes will be equipped with 128 GiB DDR4 main memory, i.e., more than 5 GiB per core. In support for workflows requiring more memory per core, additional 128 nodes with 256 GiB and 64 nodes with

512 GiB DDR4 RAM will be available. 75 nodes will be equipped with two NVIDIA K80 cards each, providing an additional flop rate of 430 teraflops per second for accelerator-capable applications. 12 Login nodes with 256 GiB per node will be available for workflow- and data-management as well the convenient execution of short running pre- and post-processing operations. Additionally, 12 visualization nodes with 512 GiB (10 nodes) and 1 TiB main memory (2 nodes) and two NVIDIA K40 GPUs are available for advanced visualization purposes. JURECA's visualization partition will replace the older JUVIS visualization cluster at JSC and move the analysis of data closer to its source. An overview of the different node types in JURECA is shown in Table 1. Fig. 1 and Fig. 2 show the V-class chassis and blade servers employed in JURECA.

The Haswell CPUs in JURECA support the AVX 2.0 instruction set architecture extension and can perform two 256-bit (i.e., two times 4 double precision floating point numbers) wide multiply-add operations per cycle. Due to the increased core count and improved microarchitecture of the Haswell CPUs, the peak floating point capabilities of compute nodes in JURECA are 10 times higher than that of a JUROPA node. From the user perspective, however, this performance improvement does not come for free but requires code optimizations and potentially refactoring in order to take advantage of the wider single instruction multiple data (SIMD) units of the Haswell CPUs. Since October 2014, JSC is providing JUROPA users with access to the 70 TFlops/s Haswell cluster JUROPATEST to foster such efforts.

JURECA will be interconnected with a cutting-edge Mellanox EDR (extended data rate) InfiniBand network. EDR InfiniBand with four lanes per direction achieves a unidirectional point-to-point bandwidth of 100 Gbps – a significant bandwidth improvement compared to the 4x QDR InfiniBand technology in use by JUROPA. As in JUROPA, the JURECA components will be organized in a fully non-blocking fat-tree topology. The core of the fabric is constituted by four 648-port EDR director switches connected to more than a hundred 36-port leaf switches located in the compute node racks.

Figure 2: A V210S compute node as used in the JURECA cluster. Each node hosts two CPU sockets. The V201F version, which has twice the width of a V201S module, can additionally host two PCIe-based accelerator cards.

JURECA will not be equipped with a system-private (global) file systems like JUROPA, but will be connected to JSC's storage cluster JUST4 and mount the work and home file systems from there. For users with access to multiple computing systems at JSC this consolidation of the storage landscape simplifies the data management and reduces data movement across the center. The connection to JUST4 will be based on InfiniBand and Ethernet technologies. By utilizing the high-bandwidth InfiniBand network for storage, JURECA will not only feature a high accumulated storage bandwidth but individual nodes will be able to sustain a significant portion of this total bandwidth. This design choice ensures JURECA's ability to service the widest variety of users' requirements from capacity to capability computing and traditional computational to emerging big data sciences. To bridge between the high speed InfiniBand network and JSC's Ethernet-based storage backbone network, Mellanox gateway switches are employed in an active/active mode providing high bandwidth and reliability through failover mechanisms. The employed gateway switches will each route traffic between eighteen FDR (Fourteen Data Rate) InfiniBand links on one side and eighteen 40GigE links on the other side. While JURECA is built around EDR InfiniBand technology, the storage connection is realized with lower performing FDR links due to the available market offerings.

JURECA's cutting edge hardware setup is matched by a state-of-the-art software stack. The system will be launched with a CentOS 7 Enterprise-Linux installation featuring a 3.10 Linux kernel. The main MPI (Message Passing Interface) implementation will be ParaStation MPI which, in the newest version available on JURECA, supports MPI-3.0. Additionally, Intel MPI will be supported on JURECA. JURECA will be the first large-scale system at JSC on which the open-source Slurm workload manager will be employed. In the context of the JUROPA collaboration a new plugin for the ParaStation resource management

Node Type	Number	Characteristics
Standard/Slim	1.605	2x Haswell E5-2680 v3, 128 GiB DDR4 RAM
Fat Type 1	128	2x Haswell E5-2680 v3, 256 GiB DDR4 RAM
Fat Type 2	64	2x Haswell E5-2680 v3, 512 GiB DDR4 RAM
Accelerated	75	2x Haswell E5-2680 v3, 128 GiB DDR4 RAM, 2x K80 GPUs
Login	12	2x Haswell E5-2680 v3, 256 GiB DDR4 RAM
Visualization Type 1	10	2x Haswell E5-2680 v3, 512 GiB DDR4 RAM, 2x K40 GPUs
Visualization Type 2	2	2x Haswell E5-2680 v3, 1 TiB DDR4 RAM, 2x K40 GPUs

Table 1: Overview of the different node types in the JURECA system.

daemon has been developed by ParTec together with JSC. This plugin allows for using the Slurm batch system in combination with the ParaStation resource management – the resource management of choice on JSC's clusters – without requiring additional daemons on the compute nodes that would inject spurious jitter.

Building on the JUROPA experience – where the successful collaboration between JSC and the hardware and software vendors has contributed to the exceptional life span of the system – JSC, T-Platforms and the software provider ParTec are engaging in a collaborative project to further develop and augment the JURECA system after installation and to approach urgent research questions in the scalability of large-scale cluster systems.

The installation of JURECA will proceed in two phases so as to minimize the service interruption for the users by allowing to install the first JURECA phase while JUROPA continues to operate. The first phase of JURECA consists of only six racks but delivers a performance equivalent to the current JUROPA system. It thus allows users to continue working while the JUROPA system is dismantled to free up space for the remaining 28 JURECA racks. The installation of this second phase will be done during production and only a short offline maintenance window is required to integrate phase one and two in a single fat-tree InfiniBand network and perform benchmarks as part of the acceptance procedure.

contact: Dorian Krause,
d.krause@fz-juelich.de

• Dorian Krause

Jülich
Supercomputing
Centre (JSC),
Germany