JUROPA-3 - A Prototype for the Next-Generation HPC Cluster

In preparation for a future replacement of the JUROPA HPC cluster, a prototype system called JUROPA-3 has been installed at Jülich Supercomputing Centre (JSC) in April 2013. JUROPA-3 is the outcome of a cooperation of JSC, ParTec Cluster Competence Center GmbH and T-Platforms, aiming at the development of solutions for fundamental questions in large-scale cluster computing. Topics like application check-point/restart, end-to-end data integrity, network topology, failure prediction and energy efficiency are addressed in this cooperation. The development and tuning of applications for many-core processor architectures are another key aspect of the project.

Being a prototype system, the size and overall performance of JUROPA-3 are rather limited. Nevertheless, the system architecture adapts to what is expected to be relevant for a full-sized production system. It follows the same hierachical concept already successfully applied in JUROPA. Two redundant master nodes on top serve as the administrative centre of the system.

They provide fail-over functionality and house services like the batch workload manager, master LDAP and DNS servers and cluster provisioning and management functions. On the next level, a set of administration nodes provide for distributed services used by the compute nodes on level three. Typical level-two services are replica LDAP and DNS servers and DHCP. The system is complemented by front-end nodes for user login, a Lustre storage pool with OSS/MDS servers and a GPFS gateway node.

Apart from the above mentioned research topics covered in the scope of the cooperation, there are many additional issues that will be investigated on the technical level. Different from JUROPA, where SUSE SLES 11 was used as the node operating system, Scientific Linux was chosen for JUROPA-3. One task to accomplish will be the merging of administative procedures and monitoring functions with the new operating system. An expected advantage - besides cost-effectiveness - is better support of hardware and software like Infiniband and the Lustre parallel file system. New in JUROPA-3 is also, that the majority of compute nodes run in diskless mode. This has effects on installation procedures as well as run-time behaviour. Especially the handling of GPFS (General Parallel File System) has proven to be intricate in this scenario with ParaStation being used for system image administration. Despite the diskless operation, 8 nodes are equipped with local disks, which will be used for checkpoint/restart development. A scenario to be tested is the mirroring of checkpoint data to neighboring nodes, such that the data is available for restart in the case of a node failure. 16 fat nodes possess an increased amount of main memory (256 GB and 128 GB, respectively) which allow for production-sized runs of structural mechanics applications. A small storage area (2 server nodes, 60 TB disk space) is dedicated to the development and testing of future Lustre parallel file system versions and end-to-end data integrity enhancements.

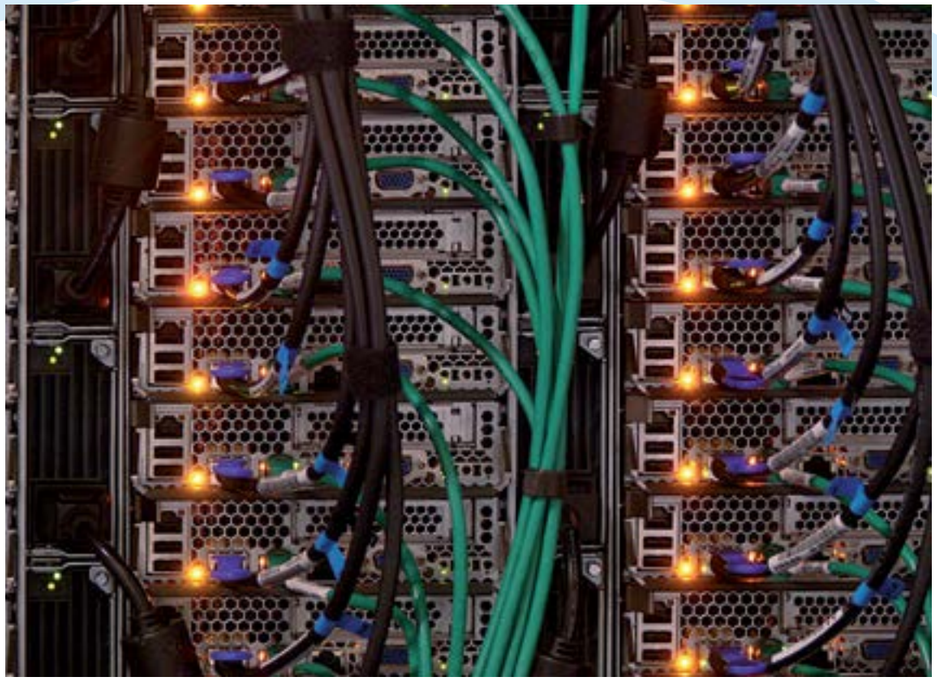

Figure 1: JUROPA-3 compute rack backview.

As mentioned above, GPFS is used as a cluster-wide file system on JUROPA-3. A GPFS gateway node with Infiniband HCA on one side and 4x10GE on the other side provides for connectivity between JUROPA-3 and FZJ's file server JUST. GPFS is mounted in addition to Lustre on all cluster nodes and serves as the main file system for home and scratch data.

Finally, the SLURM resource manager will be integrated with ParaStation and be used for job management and control. It replaces Moab and Torque used in the current JUROPA system.

System Specifications

JUROPA-3 comprises 60 compute nodes, each equipped with 2 Intel Xeon E5-2650 CPUs (Sandy Bridge-EP) providing a total of 960 processor cores with a peak performance of 15.3 teraflops. In addition, 4 compute nodes are each enhanced with 2 NVIDIA Tesla K20X GPUs and another 4 nodes with 2 Intel Xeon Phi 5110P co-processors each. The co-processors amount to an extra performance of 18.5 teraflops peak. The compute nodes, the server nodes and the storage components of the cluster are connected by a 56 Gb/s Infiniband network (FDR) with fat-tree topology providing for high communication bandwidth of parallel applications and I/O. The network topology is subject to future investigations and may be changed into hypercube or similar structure.

• Ulrich Detert

Jülich Supercomputing Centre (JSC)

Systems