GPFS Performance Aspects

To reach the maximum performance of a GPFS cluster, a number of "GPFS Golden Rules" need to be followed:

- Performance primarily depends on the number of disks within the file system (disks do not get faster anymore).
- The hardware configuration should be symmetric and all system components must be balanced.
- Settings of the GPFS parameters like the filesystem blocksize must be suitable and should match the physical layout.
- Changed characteristics from one generation of storage HW to the next (e.g. DS4000 to DS5000) should be reflected in the GPFS configuration.
- Changed usage patterns (e.g. an 5x increase in the number of files generated) will also have an impact on the filesystem performance that needs to be considered.

During the deployment phase, a number of these well-known best practices have been re-visited. Some of the design and configuration choices that were optimal with the 16-rack Blue Gene/P and generation-2007 storage hardware are no longer appropriate for a 72-rack Petaflop Blue Gene/P and generation-2009 storage hardware.

On the system level, these invaluable lessons learned (during installation, from early adopters and studies by the JSC application support team) have already been incorporated into the production setup. On the application level, efforts are ongoing to optimize the applications I/O patterns so they scale on Petaflop systems to the same degree that the MPI communication scaling has reached.

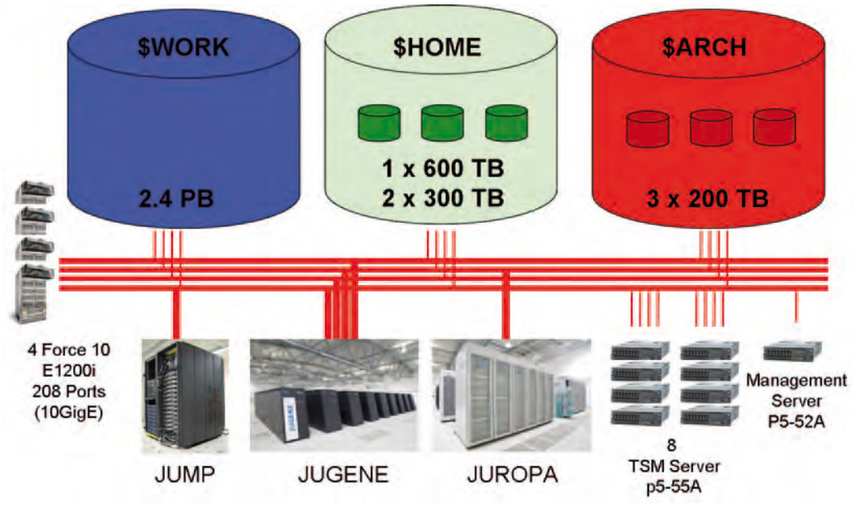

Figure 4: User view of the JUST GPFS filesystems and the connected Supercomputers at JSC

- Olaf Mextorf[1]
- Ulrike Schmidt[1]
- Lothar Wollschläger[1]
- Michael Hennecke[2]
- Karsten Kutzer[2]

[1] Jülich Supercomputing Centre, Forschungszentrum Jülich,

[2] IBM Deutschland GmbH

QPACE

QPACE (Quantum Chromodynamics Parallel Computing on the Cell) is a massively parallel and scalable computer architecture optimized for Lattice Quantum ChromoDynamics (LQCD). It has been developed in cooperation between several academic institutions (SFB TR 55 under the leadership of the University of Regensburg) and the IBM Deutschland Research and Development GmbH.

At JSC, a 4-rack QPACE system is installed. The building block is a node card comprising an IBM PowerXCell 8i processor and a custom FPGA-based network processor. 32 node cards are mounted on a single backplane and eight backplanes are arranged inside one rack, hosting a total of 256 node cards. The closed node card housing, which is connected to a liquid-cooled cold plate, acts as a heat conductor thus making QPACE a highly energy-efficient system.

To remove the generated heat a cost-efficient liquid cooling system has been developed, which enables high packaging densities. The maximum power consumption of one QPACE rack is about 32 kW. The 3D-torus network interconnects the node cards with nearest-neighbor communication links driven by a lean custom protocol optimized for low latencies. For the physical layer of the links 10 Gigabit Ethernet PHYs are used providing a bandwidth of 1 GB/s per link and direction.

In the context of PRACE work-package 8, promising technologies for future supercomputers have been evaluated and QPACE was identified as one of the advanced prototypes. In order to extend the range of applications for QPACE beyond LQCD, an implementation of the high-performance LINPACK (HPL) benchmark was a first step. Here, different from QCD applications, the transfer of large messages and collective operations must be supported efficiently. This required extensions of both network communication and the software stack.

A special FPGA bit-stream and a library of MPI functions for message passing between compute nodes has been developed at JSC in co-operation with IBM. The HPL benchmark produced excellent results (43.01 TFlop/s on 512 nodes) and the 4 rack system in Jülich corresponds to an aggregate peak performance of more than 100 TFlop/s.

Providing 723 MFlop/s/W QPACE was recognized as the most energy-efficient supercomputer worldwide and ranked top in the Green500 list at the Supercomputing Conference 2009 in Portland, Oregon.

- Willi Homberg

Jülich Supercomputing Centre, Forschungszentrum Jülich