Storage and Network Design for the JUGENE Petaflop System

In conjunction with the installation of the IBM Blue Gene/P Petaflop system JUGENE at Forschungszentrum Jülich (inSiDE Vol. 7 No. 1, p. 4), the supporting storage and networking infrastructure has been upgraded to a sustained bandwidth of 66 GByte/s, matching the enhanced computational capabilities.

Starting with the installation of the initial 16-rack JUGENE system in 2007, scalable parallel file services have been moved outside of the supercomputer systems and are now provided by a dedicated storage cluster JUST, which is based on IBM General Parallel File System (GPFS). There are three main operational considerations leading to this approach:

1. Data is more long-lived than the compute systems it is generated on.

2. There is a need to attach more than a single supercomputing resource to the parallel file system (locally and in different projects, e.g. DEISA).

3. Cutting edge supercomputers pick up many performance-related soft ware and firmware improvements that a pure storage cluster would not need.

Separating out the HPC storage greatly reduces the above interdependencies, and also makes it easier to offer site-wide parallel file services to a more heterogeneous landscape of computing and visualization resources. On the other hand, designing the networking infrastructure around it then becomes more challenging than solutions for a single supercomputer. The JUST cluster uses 10-Gigabit Ethernet as its networking technology, for several reasons:

- The I/O nodes of the Blue Gene/P (as the main client) have on-chip 10 GbE.

- Ethernet is by far the most stable, interoperable, and manageable networking solution.

- 10 GbE bandwidth is adequate, and the ultra-low latencies of other HPC interconnects are not needed for typical HPC storage traffic.

Expanding this infrastructure to support Petaflop systems requires a solution design which addresses some unique challenges which are not present (or not as relevant) at smaller scale.

- The 10 GbE network needs to provide very large port counts, and has to sustain Terabits of bandwidth between the supercomputer(s) and the storage cluster.

- The storage cluster needs to be perfectly balanced in every aspect of the hardware and software stack to be able to deliver the raw performance of thousands of disks to the end users.

- The solution needs to be resilient to faults and manageable. This becomes ever more critical with the exploding number of components both on the compute and the storage side.

- Memory is the most precious resource on petascale supercomputers, and the memory-efficient organization and scaling of application I/O patterns is an active area of our petascale research.

The Networking Layer

The networking solution needs to support a sustained GPFS bandwidth of 66 GByte/s and has to provide 10 GbE ports for 700 to 800 participants: the 600 I/O nodes of the 72-rack Blue Gene/P, at least 100 ports in the JUST storage cluster, plus connectivity to additional systems. To achieve this, JSC together with IBM started the design and planning phase well ahead of the actual deployment. Even today there is no switch solution on the market that offers 700 to 800 10 GbE ports on a single backplane, and the combination of multiple switches introduces several design and bandwidth limitations, e.g. due to port channel restrictions (IEEE802.3ad) and the spanning tree algorithm.

After discussing different network designs and even routing protocol based network solutions (e.g. equal-cost multipath), a rather flat network topology was chosen, avoiding the risk of badly converging routing protocols and additional costs for big pipes between switch fabrics. Based on the knowledge of the behaviour of GPFS as the main application, a network layout was invented that uses four switch chassis but keeps all high-bandwidth storage access traffic local within each of the switch fabrics.

To achieve this, quad-homed GPFS servers were deployed that offer direct storage access local at each switch fabric. The high number of Blue Gene/P I/O nodes are equally distributed over those 4 switch fabrics. Rather than attaching whole BG/P racks to a single fabric (which would limit the rack to that one fabric), the 8 I/O nodes of a BG/P rack are distributed over all four fabrics, so smaller BG/P jobs can also benefit from the aggregate bandwidth of all four planes.

The following candidates have been evaluated for the switch fabric:

- Cisco Nexus 7000 in its 18-slot version

- Force10 E1200i as a 14-slot switch

- Myricom Myri-10G switches

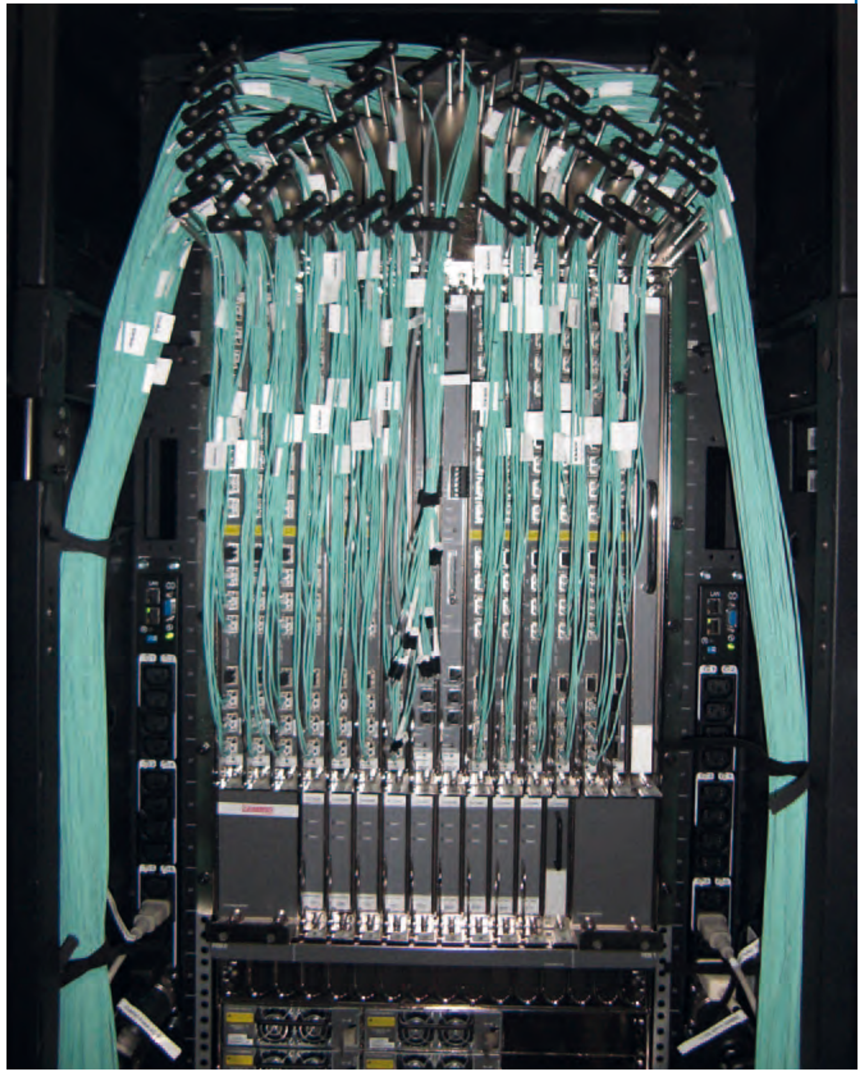

Figure 1: One of the four Force10 E1200i 10 GbE switches

While each alternative had some advantages and some disadvantages, especially in the area of new upcoming DataCenter Ethernet standards like PerPriority Flow Control (IEEE802.1Qb) and Multichassis Etherchannel, at the end the major decision criteria were the proven stable functionality of the existing single E1200 chassis for just the functions needed for JSC's network setup in combination with the assured availability of the desired hardware in sync with the overall Blue Gene/P and storage cluster deployment timeline.

Using the knowledge of Force10's module and ASIC design in combination with the known GPFS traffic patterns, JUST servers and BG/P I/O nodes are connected to the linecard ports in a layout that minimizes the congestion probability on the 4x over-subscribed linecards (which are required to achieve the large portcounts). At the end of the multi-stage deployment phase and some additional network optimizations (introduction of Jumbo frames, deployment of end-to-end flow control, tuning of network options and buffers), the chosen network design has proven to be the ideal solution for JSC's Blue Gene/P and POWER6 based storage cluster deployment. The traffic patterns observed in production mode via JSC's RMON and SFlow/Netflow based monitoring match the assumptions that were used as a base for this specific tailored network solution.

Using Ethernet as the base technology is a perfect fit for heterogeneous aspects like integrating the Nehalem-based JUROPA cluster, and interconnecting with different campus and project networks as well as visualization solutions.

GPFS Storage Design and Usage Model

The design of the hardware building blocks for the expanded JUST storage cluster is driven by three factors:

1. The target of a sustained aggregate bandwidth of 66 GByte/s.

2. The requirement to use quad-homed GPFS/NSD servers with sufficient bandwidth to drive the four 10 GbE network fabrics.

3. The requirement of a seamless migration from the initial JUST storage cluster installed in 2007 (and the goal to re-use parts of its hardware, in particular the SATA disks).

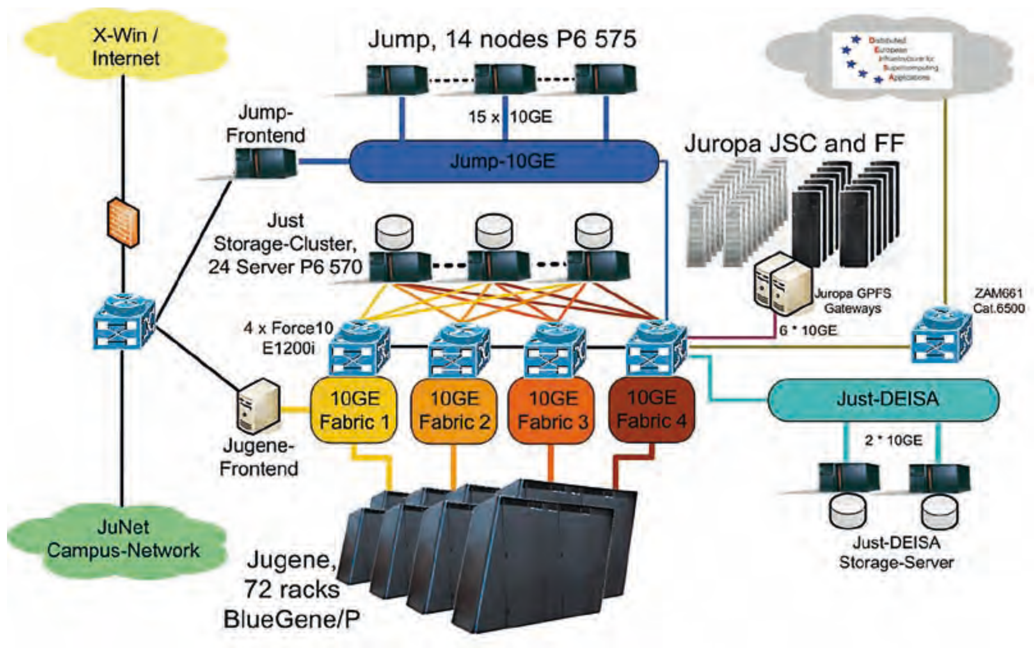

Figure 2: JSC Network Overview

As in the initial JUST cluster, the GPFS metadata is stored on dedicated metadata building blocks, with two IBM DS5300 storage subsystems and 15k FC disks. Metadata is protected both by Raid1 arrays and by GPFS replication.

Several combinations of servers and disk storage subsystems have been evaluated for the GPFS data building block. For best price/performance, nearline SATA disks are used and protected as 8+2P Raid6 arrays. The resulting GPFS data building block is shown in Figure 3:

- 3 x GPFS-DATA servers p6-570, each: 2 x CEC, 8 core, 8 x dual-port FC4 adapters, (8 ports used), 4 x 10 GigE link into BG/P network

- 2 x DS5300 storage subsystems, each: 16 x FC host ports (12 ports used), 24 x EXP5000, each @ 16 x SATA drives (5/6 are 1 TB drives, 1/6 are 500 GB drives reused from JUST-1)

To achieve the desired aggregate GPFS bandwidth, eight of these building blocks are deployed providing a total of 4.2 PByte usable capacity. This storage is partitioned into three classes of filesystems with different usage characteristics:

1. A large scratch filesystem $WORK with roughly half the capacity and bandwidth, not backed up or archived. Blue Gene/P production jobs should generally use this filesystem.

2. $HOME filesystems, with daily incremental backups.

3. $ARCH filesystems, with daily incremental backups and migration to tape storage through a combination of the IBM GPFS Information Lifecycle Management (ILM) features and IBM TSM/HSM.

Figure 4 depicts the user view of the GPFS filesystems on the expanded JUST cluster. Note that to cope with the performance impact of the increased number of files resulting from an 5x enlarged Blue Gene/P system, the number of $ARCH filesystems has been increased from one to three and users are distributed across those.

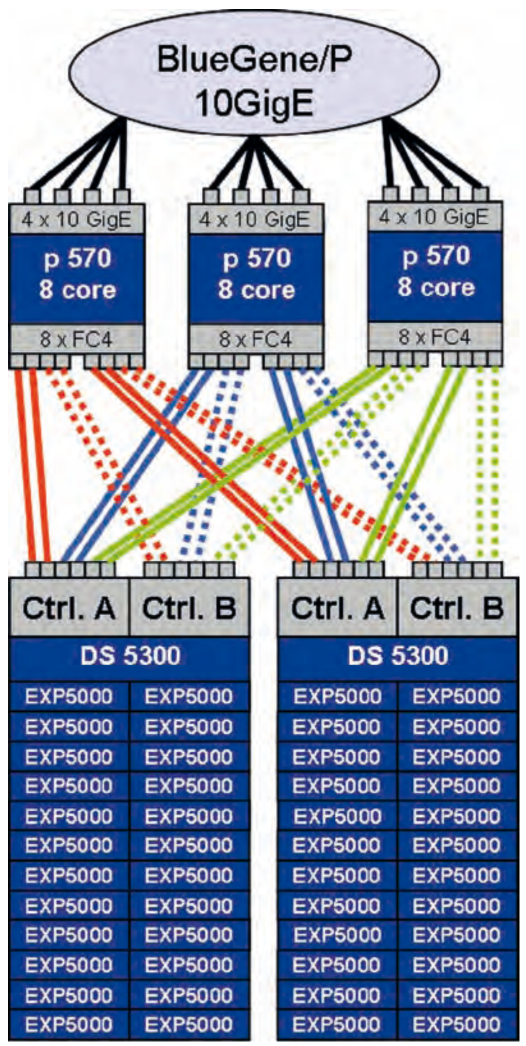

Figure 3: One JUST Building Block ($WORK uses four, $HOME and $ARCH share the other four building blocks)

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 any-more).
- The hardware configuration should be symmetric and all system compo-nents must be balanced.
- Settings of the GPFS parameters like the filesystem blocksize must be suitable and should match the physi-cal 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 num-ber 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.

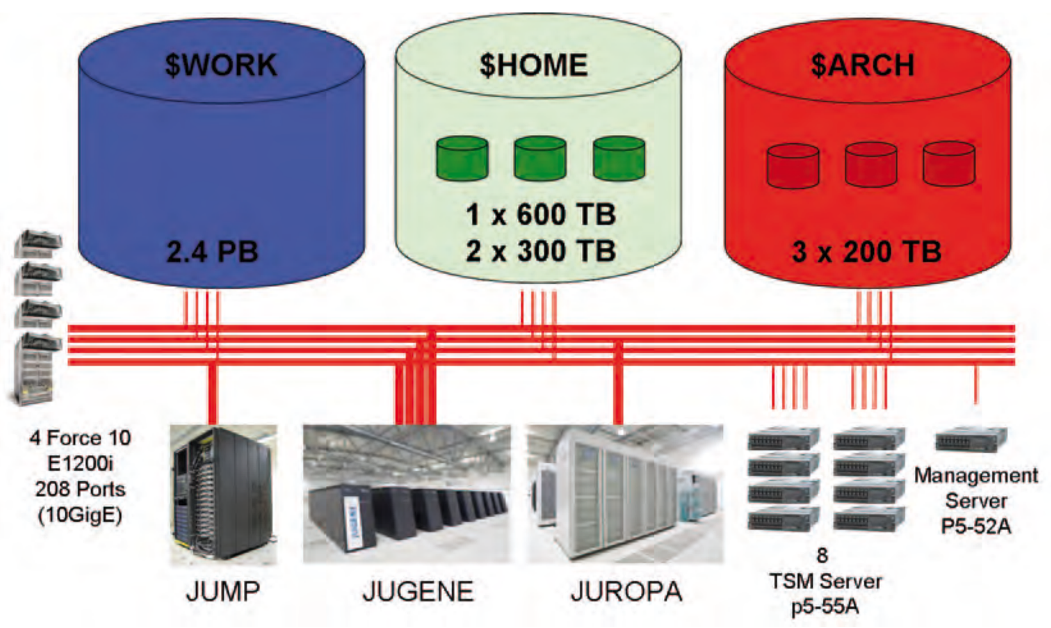

Figure 4: User view of the JUST GPFS filesystems and the connected Supercomputers at JSC

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

[1] Jülich Supercomputing Centre, Forschungs-zentrum Jülich,

[2] IBM Deutschland GmbH

Systems

QPACE

QPACE (Quantum Chromodynam-ics Parallel Computing on the Cell) is a massively parallel and scalable computer architecture optimized for Lattice Quantum ChromoDynamics (LQCD). It has been developed in co-operation between several academic institutions (SFB TR 55 under the lead-ership 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 maxi-mum power consumption of one QPACE rack is about 32 kW. The 3D-torus network interconnects the node cards with nearest-neighbor commu-nication 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 implementa-tion 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 sup-ported efficiently. This required exten-sions 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 pro-duced 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, Forschungs-zentrum Jülich