bwForCluster MLS&WISO

Sabine Richling[1], Martin Baumann[1], Stefan Friedel[2], and Heinz Kredel[3]

[1]Computing Centre, Heidelberg University
[2]Interdisciplinary Center for Scientific Computing, Heidelberg University
[3]Computing Centre, University of Mannheim

The bwForCluster MLS&WISO provides high performance compute resources for the Universities in Baden-Württemberg with focus on the research fields Molecular Life Science (MLS), Economics and Social Sciences (WISO) as well as on the development of methods for scientific computing in general. The different requirements of the communities are met by a distributed cluster concept with various node types. The cluster is connected to the scientific data storage SDS@hd with high bandwidth which facilitates the data management of complex and data-intensive workflows.

1 Introduction

The bwForCluster MLS&WISO is an entry level (Tier 3) system within the state of Baden-Württemberg's bwHPC concept for high performance computing [1]. The system is financed by the Ministry of Science, Research, and the Arts Baden-Württemberg (MKW) and the German Research Foundation (DFG) as well as the Interdisciplinary Center for Scientific Computing (IWR) of Heidelberg University. The bwForCluster MLS&WISO is a joint project of the Computing Centre (URZ) and the Interdisciplinary Center for Scientific Computing (IWR) of Heidelberg University and the Computing Centre of the University of Mannheim (RUM). The system consists of two separate clusters: The development part (Linpack performance 222.7 TFLOP/sec) and the production part (Linpack performance 291.3 TFLOP/sec). In 2015 both parts were present in the TOP500 list of the most powerful computer systems in the world [2].

2 Development Part

The development part of bwForCluster MLS&WISO (Fig. 1) is a high performance compute cluster with about 400 identical nodes and a fully non-blocking Infiniband network of speed QDR (40 GBit/sec) [3]. This cluster part is designed for compute activities related to method development in all research fields. The hardware is located in Heidelberg. The system is operated by the IWR. The development part provides a flexible service with short response times for massively parallel jobs with high scalability. The operating concept is such that it is possible to use the whole machine for a single job.

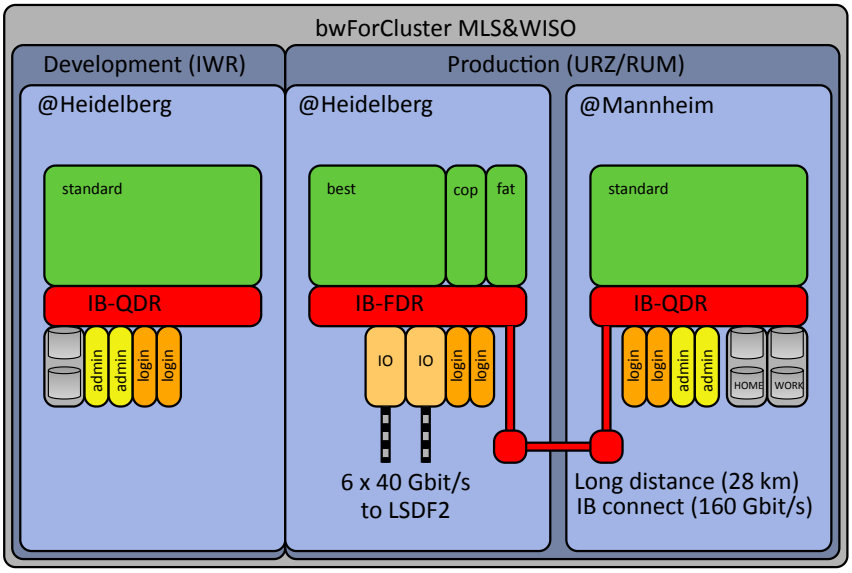

Figure 1: bwForCluster MLS&WISO: Development part and production part.

3 Production Part

The production part of bwForCluster MLS&WISO (Fig. 1) is a high performance compute cluster with about 700 nodes. The hardware is located partly in Heidelberg and partly in Mannheim. Both sites are interconnected by a long-distance Infiniband link effectively forming one single system which is jointly operated by URZ and RUM. The production part is intended for compute activities in the research fields Molecular Life Science (MLS), Economics, and Social Sciences (WISO). It provides a stable service for a large throughput of production jobs using up to 128 nodes.

All components of the cluster are connected to a high-speed Infiniband fabric for MPI communication and storage access. The Infiniband fabric in Mannheim is of speed QDR (40 GBit/sec) and fully non-blocking across all 'standard' nodes. The Infiniband fabric in Heidelberg is of speed FDR (56 Gbit/sec) and also fully non-blocking across all 'best', 'fat', and 'coprocessor' nodes. The different node types allow a high diversity of jobs: 'best' nodes have more memory and faster CPUs than 'standard' nodes, 'fat' nodes are equipped with a large amount of memory (1 TB and 1.5 TB), and 'coprocessor' nodes are equipped with GPUs (Nvidia K80) or MICs (Xeon Phi). A complete description of the node types is available in the bwHPC Wiki [4]. Two separate storage systems provide storage space for home directories (36 TB) and for workspaces (384 TB). Both storage systems are equipped with the parallel cluster file system BeeGFS [5].

4 Long-distance Infiniband Link

The long-distance Infiniband link permits the operation of the production part as one single cluster. Fig. 2 shows the interconnection in more detail. The 28 km distance between the cluster sites in Heidelberg and Mannheim is linked via a 10 GE optical fibre. Both ends are equipped with a DWDM multiplexer allowing the simultaneous transmission of different colours over one fibre. We use 16 colours to serve the two Mellanox MetroX TX6240 LongHaul appliances at each site which in turn are connected with 4 × 40 GBit/sec links to the Infiniband fabric. The total

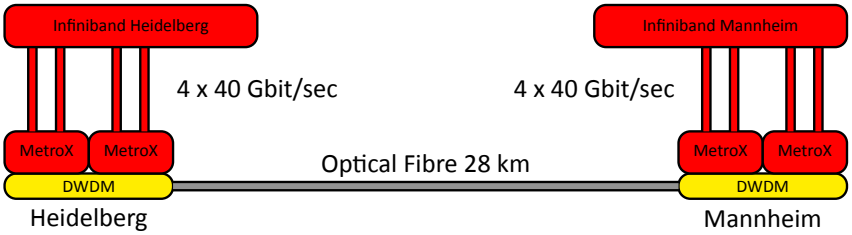

Figure 2: bwForCluster MLS&WISO: Long-distance Infiniband interconnect.

bandwidth of the long-distance interconnection is 160 GBit/sec. Latency is only slightly above the hard limit set by the speed of light. The bandwidth is broad enough to allow parallel accesses to the storage systems. For production operation, it is not allowed to use nodes from both sides for a single job, because in general only parallel jobs with low communication requirements run with sufficient performance across the long-distance link. One advantage of a two-site cluster is a higher availability. Because of the high power consumption, the compute nodes are not connected to the uninterruptible power supply. In the case of power failures or cooling problems at one site, the cluster keeps running with the compute nodes of the other site.

5 Access to SDS@hd

SDS@hd is a Scientific Data Storage for data with frequent access ('hot data') [6]. The data is stored on the second generation hardware of the Large Scale Data Facility (LSDF2) which is financed by MWK and DFG and is part of the state of Baden-Württemberg's bwDATA concept for data-intensive services [1]. The service is open to all scientist at Baden-Württemberg's Universities. Supported access protocols are NFSv4 with Kerberos, SMB 2.x/3.x as well as SSHFS.

Access to SDS@hd from all bwHPC clusters is in preparation and in case of the bwForCluster MLS&WISO already established for production use. A special feature of bwForCluster MLS&WISO is the direct and high-bandwidth connection to SDS@hd which is possible due to the physical proximity of the two systems. This allows the generation of data, e.g. with a microscope in a lab, the data analysis on the cluster, and pre- and postprocessing at the office PC of a scientist without keeping several copies of the same data and without annoying

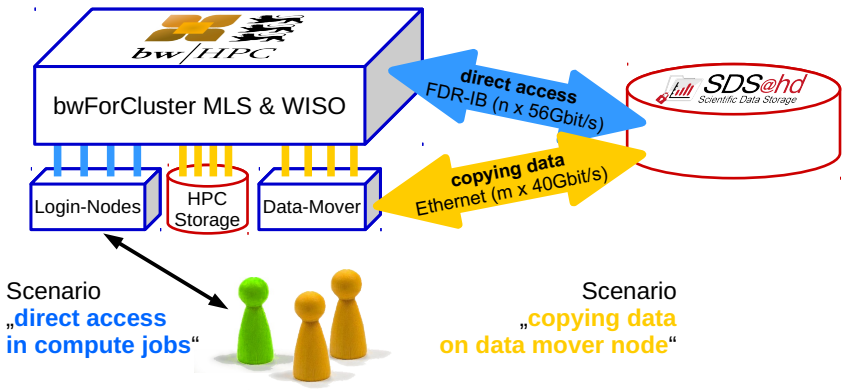

Figure 3: bwForCluster MLS&WISO: Access to SDS@hd.

manual data synchronization. For this scenario SDS@hd is mounted at the compute nodes via Infiniband (blue arrows in Fig. 3) for direct access to SDS@hd in compute jobs. If the workflow or the access pattern benefit from the parallel cluster file system, a fast data transfer via data mover nodes (IO nodes) from SDS@hd to the cluster file system is possible (yellow arrows in Fig. 3). SDS@hd is mounted at the data mover nodes via separate broad band Ethernet links to prevent copy processes from interfering with SDS@hd accesses in running jobs. Kerberos ticket management is done on the data mover nodes and is supported by e-mail notification to remind of expiring tickets.

6 bwHPC competence center MLS&WISO

User support for the bwForCluster MLS&WISO is provided by the bwHPC competence center MLS&WISO as part of the project bwHPC-C5 [7]. General support activities include assistance in access and usage of the cluster as well as in providing software for the user communities. The bwHPC competence center MLS&WISO organizes regularly the user meeting bwForTreff and an introductory HPC workshop. It also hosts workshops of external lecturers on specific software or method related topics.

In the research fields economics and social sciences many user groups apply statistical methods using software packages like R, Stata, and Matlab. The research projects deal for example with asset pricing, corporate taxation, political reforms, empirical accounting, household models, risk management, structural changes in agriculture, and energy system analysis. In social sciences data mining and language processing are fields with increasing compute demands. Here the competence center provides migration support and assists in the setup of workflows.

Scientists working in the field of molecular life science apply a broad spectrum of methods such as molecular dynamics, bioinformatics, bio-medical image processing and statistics as well as computational fluid dynamics (CFD). With the help of these methods research groups from various biological and medical institutes investigate for example structure formation processes in biological systems, folding and aggregation phenomena of proteins, membrane reorganization, evolution of vertebrate gene expression, molecular phylogenetics of birds, morphological plasticity and protein composition of neurons, biological images from fluorescent microscopy or cryo-electron microscopy, and the blood or air flow in human organs. The different node types of the cluster offer suitable hardware for the diverse methods. For example molecular dynamics and fluid dynamics jobs need many cores and benefit from the fast non-blocking Infiniband network. 'fat' nodes enable image processing to analyse very large datasets en bloc. 'coprocessor' nodes speed up the work, since more and more software packages provide GPU-enabled versions for example in molecular dynamics and image processing.

The growing number of users, high-level support teams, and community-specific software modules shows that the bwForCluster MLS&WISO has become an important tool for scientists in the fields of molecular life science, economics, and social sciences.

Acknowledgements

The bwForCluster MLS&WISO is funded by the state of Baden-Württemberg through bwHPC, by the German Research Foundation (DFG) through grant INST 35/1134-1 FUGG, and by the IWR of Heidelberg University.

References

[1] Hartenstein, H., T. Walter, and P. Castellaz. "Aktuelle Umsetzungskonzepte der Universitäten des Landes Baden-Württemberg für Hochleistungsrechnen und datenintensive Dienste." Praxis der Informationsverarbeitung und Kommunikation, Band 36, Heft 2 (2013): 99-108. http://dx.doi.org/10.1515/pik-2013-0007

[2] TOP500 Site: Heidelberg University and University of Mannheim. https://www.top500.org/site/50564

[3] bwHPC-Wiki: bwForCluster MLS&WISO Development. https://www.bwhpc-c5.de/wiki/index.php/BwForCluster_MLS&WISO_Development_Hardware

[4] bwHPC-Wiki: bwForCluster MLS&WISO Production. https://www.bwhpc-c5.de/wiki/index.php/BwForCluster_MLS&WISO_Production_Hardware

[5] BeeGFS – The Leading Parallel Cluster File System. https://www.beegfs.io/content

[6] SDS@hd – Scientific Data Storage. https://sds-hd.urz.uni-heidelberg.de

[7] bwHPC-C5: Coordinated Compute Cluster Competence Centers. http://www.bwhpc-c5.de