SuperMUC Petascale System

www.lrz.de/supermuc

System purpose

SuperMUC is the high-end supercomputer at the Leibniz-Rechenzentrum (Leibniz Supercomputing Centre, LRZ) in Garching near Munich (the MUC suffix is borrowed from the Munich airport code). With more than 241,000 cores and a combined peak performance of the two installation phases of more than 6.8 Petaflop/s (=10^{15} Floating Point Operations per second), it is one of the fastest supercomputers in the world.

SuperMUC strengthens the position of Germany's Gauss Centre for Supercomputing in Europe by delivering outstanding compute power and integrating it into the European high performance computing ecosystem. With the operation of SuperMUC, LRZ will act as a European Centre for Supercomputing and will be a Tier-0 centre for PRACE, the Partnership for Advanced Computing in Europe. SuperMUC is available to all German and European researchers to expand the frontiers of science and engineering.

SuperMUC Phase 1 and Phase 2 in the computer room

System Configuration Details

LRZ's target for the architecture is a combination of a large number of thin and medium sized compute nodes with 32 GByte (Phase 1) and 64 GByte (Phase 2) of memory, respectively, and a smaller number of fat compute nodes with 256 GByte memory. The network interconnect between the nodes allows excellent scaling of parallel applications up to the level of more than 10,000 tasks.

SuperMUC consists of 18 Thin Node Islands based on Intel Sandy Bridge-EP processor technology, 6 Thin Node Islands based on Intel Haswell-EP processor technology and one Fat Node Island based on Intel Westmere-EX processor technology. Each Island contains more than 8,192 cores. All compute nodes within an individual Island are connected via a fully non-blocking Infiniband network (Phase 1: FDR10 for the Thin nodes of Phase 1, FDR14 for the Haswell nodes of Phase 2 and QDR for the Fat Nodes). Above the Island level, the pruned interconnect enables a bi-directional bi-section bandwidth ratio of 4:1 (intra-Island / inter-Island).

In addition, SuperMIC, a cluster of 32 Intel Sandy Bridge-EP nodes each having two Intel Xeon Phi accelerator cards installed, is also part of the SuperMUC system.

Energy Efficiency by Warm Water cooling

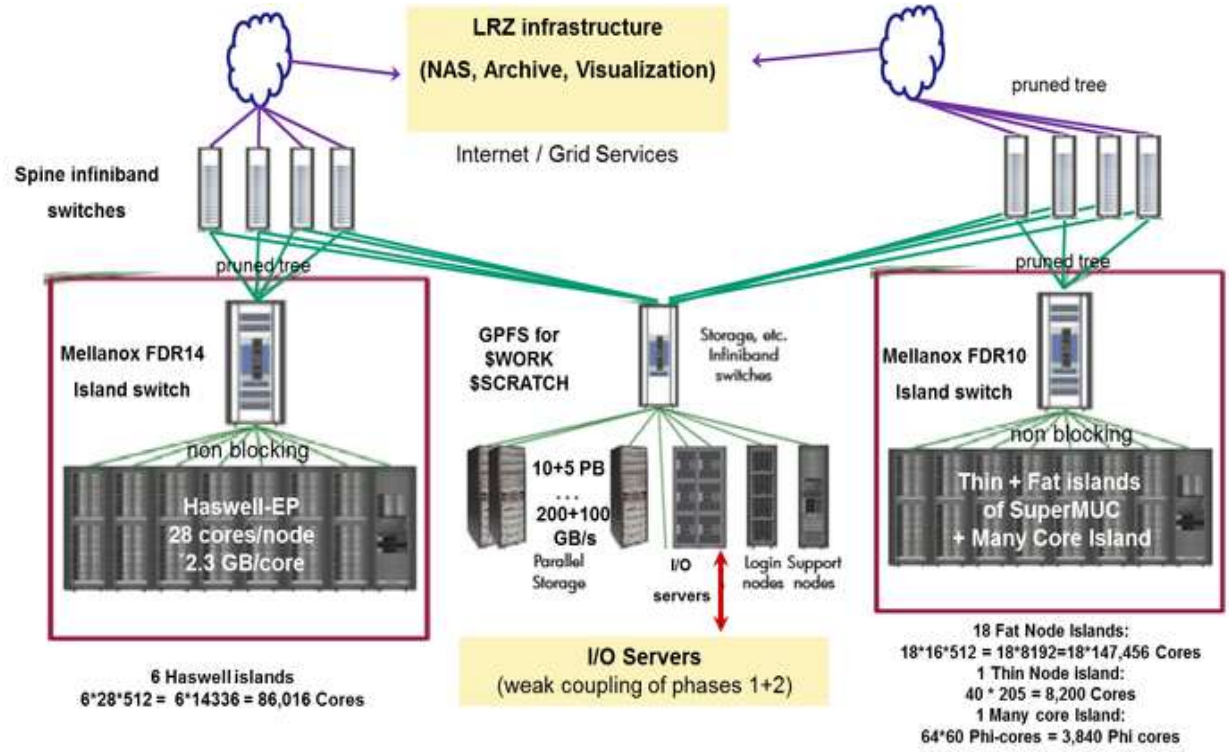
SuperMUC uses a new, revolutionary form of warm water cooling developed by IBM. Active components like processors and memory are directly cooled with water that can have an inlet temperature of up to 40 degrees Celsius. The "High Temperature Liquid Cooling" together with very innovative system software promises to cut the energy consumption of the system. In addition, all LRZ buildings will be heated re-using this energy.

Technical Data:

Installation Phase	**Phase 1**			**Phase 2**
Installation Date	2011	2012	2013	2015
Islandtype	Fat Nodes	Thin Nodes	Many Cores Nodes	Haswell Nodes
System	BladeCenter HX5	IBM System x iDataPlex dx360M4	IBM System x iDataPlex dx360M4	Lenovo NeXtScale nx360M5 WCT
Processor Type	Westmere-EX Xeon E7-4870 10C	Sandy Bridge-EP Xeon E5-2680 8C	Ivy-Bridge (IvyB) and Xeon Phi 5110P	Haswell Xeon Processor E5-2697 v3
Nominal Frequency [GHz]	2.4	2.7	1.05	2.6
Total Number of nodes	205	9216	32	3072
Total Number of cores	8,200	147,456	3,840 (Phi)	86,016
Total Peak Performance [PFlop/s]	0.078	3.2	0.064 (Phi)	3.58
Linpack Performance [PFlop/s]	0.065	2.897	n.a.	2.814
Total size of memory [TByte]	52	288	2.56	194
Total Number of Islands	1	18	1	6
Typical Power Consumption [MW]	< 2.3			~1.1
Components				
Nodes per Island	205	512	32	512
Processors per Node	4	2	2 (IvyB) 2.6 GHz + 2 Phi 5110P	2
Cores per Processor	10	8	8 (IvyB) + 60 (Phi)	14
Cores per Node	40	16	16 (host) + 120 (Phi)	28
Logical CPUs per Node (Hyperthreading)	80	32	32 (host) + 480 (Phi)	56
Memory				
Memory per Core [GByte] (typically available for applications)	6.4 (~6.0)	2 (~1.5)	4 (host) + 2 x 0.13 (Phi)	2.3 (2.1)

Installation Phase	**Phase 1**			**Phase 2**
Size of shared Memory per node [GByte]	256	32	64 (host) + 2 x 8 (Phi)	64
Bandwidth to Memory per node [Gbyte/s]	136.4	102.4	Phi: 384	137
Interconnect				
Technology	Infiniband QDR	Infiniband FDR10	Infiniband FDR10	Infiniband FDR14
Intra-Island Topology	non-blocking Tree			non-blocking Tree
Inter-Island Topology	Pruned Tree 4:1		n.a.	Pruned Tree 4:1
Bisection bandwidth of Interconnect [TByte/s]	12.5			5.1
Servers				
Login Servers for users	2	7	1	5
Storage				
Size of parallel storage (SCRATCH/WORK) [Pbyte]	15			
Size of NAS storage (HOME) [PByte]	3.5 (+ 3.5 for replication)			
Aggregated bandwidth to/from parallel storage [GByte/s]	250			
Aggregated bandwidth to/from NAS storage [GByte/s]	15			
Capacity of Archive and Backup Storage [PByte]	> 30			
System Software				
Operating System	Suse Linux Enterprise Server (SLES)			
Batchsystem	IBM Loadleveler			
Parallel Filesystem for SCRATCH and WORK	IBM GPFS			
File System for HOME	NetApp NAS			
Archive and Backup Software	IBM TSM			
System Management	xCat, Icinga, Splunk			

SuperMUC Phase1 and Phase2 are loosely coupled through the common used GPFS/GSS and NAS File systems. It is not possible to run a job across Phase1 and Phase2 and the scheduling and job classes of Phase1 and Phase2 are different. However, Phase1 and Phase2 share the same programming environment.



Schematic view of SuperMUC Phase1 + Phase2

Home file system

Permanent storage for data and programs is provided by a 16-node NAS cluster from Netapp. This primary cluster has a capacity of 3.4 Petabytes and has demonstrated an aggregated throughput of more than 12 GB/s using NFSv3. Netapp's Ontap 8 "Cluster-mode" provides a single namespace for several hundred project volumes on the system. Users can access multiple snapshots of data in their home directories. Data is regularly replicated to a separate 4-node Netapp cluster with another 2 PB of storage for recovery purposes. Replication uses Snapmirror-technology and runs with up to 2 GB/s in this setup. Storage hardware consists of >3400 SATA-Disks with 2 TB each protected by double-parity RAID and integrated checksums.

Work and Scratch areas

For highest-performance checkpoint I/O IBM's General Parallel File System (GPFS) with 12 PB of capacity and an aggregated throughput of 250 GB/s is available. Disk storage subsystems were built by DDN.

Tape backup and archive

LRZ's tape backup and archive systems based on TSM (Tivoli Storage Manager) from IBM are used for or archiving and backup. The have been extended to provide more than 30 Petabytes of capacity to the users of SuperMUC. Digital long-term archives help to preserve results of scientific work on SuperMUC. User archives are also transferred to a disaster recovery site.

Visualization and Support systems

SuperMUC is connected to powerful visualization systems: the new LRZ office building houses a large 4K stereoscopic powerwall as well as a 5-sided CAVE virtual reality environment.

For more details see: http://www.lrz.de/supermuc