Leibniz-Rechenzentrum

der Bayerischen Akademie der Wissenschaften

Description of Goods and Services

for the
European High Performance Computer
SuperMUC at LRZ

(23. September 2010)

Sep 2010 **LRZ-Bericht 2010-02**

Direktorium:
Prof. Dr. A. Bode (Vorsitzender)
Prof. Dr. H.-J. Bungartz
Prof. Dr. H.-G. Hegering
Prof. Dr. D. Kranzlmüller

Leibniz-Rechenzentrum
Boltzmannstraße 1
85748 Garching

UST-ID-Nr. DE811305931

Telefon: (089) 35831-8000
Telefax: (089) 35831-9700
E-Mail: lrzpost@lrz.de
Internet: http://www.lrz.de

Öffentliche Verkehrsmittel:
U6: Garching-Forschungszentrum

Content:

1 General information on the RFP

1.1 General objectives for the SuperMUC petascale system

The Partnership for Advanced Computing in Europe (PRACE) is preparing the formation of a persistent European HPC service, consisting of several tier-0 centres providing European researchers with access to capability computers and forming the top level of the European HPC ecosystem. The multi-Petascale supercomputer to be procured, SuperMUC, will be an important component of this European HPC service. It will also serve German universities and research institutions as a national high-end computing resource.

The expected job profile on the SuperMUC requires, simultaneously, high compute performance, a large main memory with fast access characteristics, an efficient interconnect, and mass storage with high I/O throughput. For the broad range of applications to be deployed, it is most important that the offered supercomputer provides balanced hardware and software characteristics. Since this multi-petascale computer will be composed of more than hundred thousand processing cores, a highly scalable system software stack and programming environment is of major importance for the efficient use of the system.

It is the explicit goal of the LRZ to provide these services without imposing new constraints on its current and future user base. To this end, only a general-purpose computing system will be considered for acquisition. Special-purpose computers that offer particular advantages for only a small subset of scalable programs will not be taken into consideration unless they can be used by a significant user fraction.

Following this, the primary goal for the procurement of the petascale computer is:

To establish an integrated, highly energy efficient system and a programming environment which enable the solution of the most challenging scientific problems from widely varying scientific areas.

With an expected compute power of multiple PetaFlop/s, great importance is attached to the issue of scalability. However, extensive parameter studies with a large number of medium-sized jobs are common in many scientific areas; efficient processing of these should also be possible on the new petascale computer. Supercomputing can only be achieved in conjunction with cutting-edge algorithms and program development, particularly regarding the proficiency that modern multi-core processors demand with respect to parallel program design.

Within this context, the theoretical peak performance of a computer becomes a subordinate aspect in this procurement. Instead, the expected performance for the entire spectrum of applications of our current and potential users is the dominating criterion. This will be ensured by stipulating the execution of realistic benchmark programs. While benchmarks from HLRB I users were used during the acquisition of the HLRB II, for the current petascale computer procurement appropriately scalable benchmarks from European HPC projects (DEISA, PRACE) as well as from LRZ users are included.

Given the current usage profile of the HLRB II, the acquisition goal of having a cluster containing a mixture of thin shared memory and, to a smaller extent, of fat shared memory compute nodes is chosen. These compute nodes must be connected by a scalable high-speed network that offers a well-balanced communication bandwidth between the nodes. As was already the case with the HLRB II, it is not required that the nodes and interconnect network must be entirely homogeneous. However, the complete system must be usable for a single application.

Because programs will have long run times, the offered SuperMUC has to ensure high stability under permanent load with changing usage profiles.

1.2 LRZ System Concept

In its concept for the SuperMUC, LRZ starts out from the assumption that a large proportion of applications will grow in terms of core count to the size of the current HLRB II (roughly ten thousand cores) and even beyond; a few applications will require using half or all cores of the entire system. In its concept, LRZ envisions **islands of highly interconnected compute nodes with non-blocking or nearly optimal communication links between the nodes within an island.** The size of such an island should be equal to or larger than 8192 cores.

The link bandwidth and the number of links for the **communication between these islands** may have smaller aggregate bandwidth than within an island **(pruned interconnect[1])** depending on the costs and technology which is used.

The rationale and experience for this is:

- Many communication patterns are of nearest-neighbour type and do not need a full set of links across islands due to the volume-surface ratio of data exchange.
- Large-scale applications, which need more cores than one island provides, may be scheduled round-robin over all or many islands and may consume the link bandwidth that is not used by other applications that are running entirely in one island.

Of course, if there is no substantial commercial or technical advantage to be gained by such a hierarchical pruned interconnect, this concept of islands needs not be used. **Other implementations of the internal network** like fat tree, 3-D/multi dimensional torus, hypercube etc. are also possible and welcome.

It is expected that **topology-aware placement** of jobs will be necessary for most solutions, handling of which should be flexible and easy to manage. It is also desired that **intelligent routing** algorithms prevent contention with the interconnects as well as provide failover/re-routing in case of hardware failures.

The experience of LRZ shows that parallelisation is easier and more efficient with a moderate number of high-performance nodes than with a large number of low-performance nodes. Therefore, many users will utilize **the hybrid programming model** with MPI between compute nodes and autoparallelization or OpenMP within the nodes. For several large-scale applications, it has also been demonstrated that the hybrid approach yields better performance than pure MPI. Experiences and theoretical considerations show that for achieving optimal performance with the hybrid model, a **fully thread-safe implementation of MPI** is essential.

To provide similar shared memory characteristics as those presently available on the HLRB II, the new system should have at least one island consisting of **fat shared memory nodes**, consisting of 32 or more cores and having substantially larger memory than a node in the rest of the system[2]. The bulk of the system should consist of **thin shared memory nodes** with 16 or more cores.

Accelerators (like GPGPUs) might be an appropriate way to increase the performance of some applications; it must be possible to connect them to the system. However, the accelerators themselves are not part of the procurement.

Message passing will continue to be the dominant programming paradigm and will be used for coarse-grained parallelism. Consequently, the system must be equipped with an efficient internal interconnect as well as an efficient **MPI implementation** that makes optimal use of the hardware and optionally off-loads workload to the network devices. Efficient implementations for the evolving **PGAS (Partitioned Global Address Space)** programming languages are desirable.

The complete **software** stack necessary for development and deployment of scientific and technical applications must be available (a Linux operating system suitable for supercomputing centres, highly efficient compilers, libraries, tools, ISV applications).

Together with the compute nodes, the tenderer must offer sufficiently **large and efficient mass storage subsystems** for user data. The **integrity and safety of data** has to be assured, since in the PetaByte regime dump-restore of file systems is not feasible.

The usage profile calls for two types of storage:

---

[1] The term “pruned interconnect” is used for fat tree topologies in which the connection of certain switch level hierarchies is realized with a reduced number of network links compared to a fully fledged fat tree topology (pruned fabric).

[2] SoftwareSMP may be used to build even larger virtual fat nodes with cache coherent memory.

- A **parallel file system** which should contain temporary data and the large result files. For these data, high bandwidth and moderate metadata performance are needed. Specialised support for efficient parallel I/O to a single file from multiple nodes is expected via MPI-IO.
- A highly available **home file system** for user data ($HOME), such as source and input files, libraries, and configuration data. The size of the individual files will be moderate but the number of files will be very high. Metadata performance should be very high, while bandwidth may be moderate. The requirements for reliability, availability and for data safety are very high. In addition, user data should be easily accessible from outside SuperMUC. LRZ has excellent experiences with Network Attached Storage (NAS) for this purpose. For safety and backup purposes, **snapshots** of the file system with various timestamps must be supported.

Installation of the SuperMUC shall proceed in **two phases**:

- Installation of **Phase 1** of the system should start in **2011**, as soon as the infrastructure of the new LRZ facilities is completed.
- **Phase 2**[3] might be installed approximately two years after the installation of Phase 1**.**

The **combined performance** of the Phases 1 and 2 should be at least by a factor **2.4** higher than that of Phase 1. The system should achieve one of the top positions among supercomputers worldwide.

Outside the scope of this procurement, **systems for visualization and for archiving and backup** will be acquired. Such systems should be efficiently connectable and operable together with the SuperMUC. File systems should be mutually accessible and co-scheduling with the SuperMUC should be possible.

The general concept of the system is illustrated in the following Figure. The items inside the red rectangle are integral part of the procurement, the items outside must interoperate.

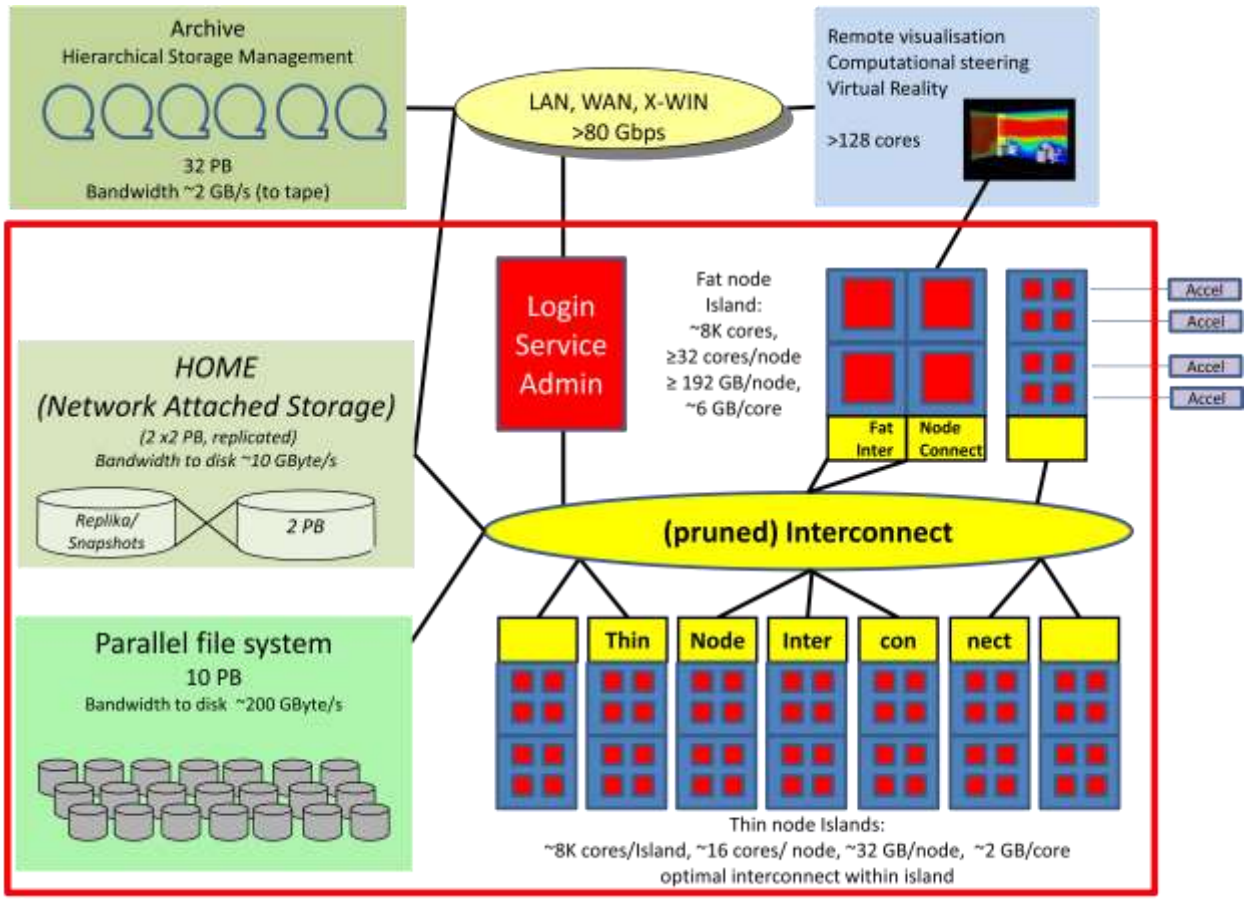


Abbildung: Concept for SuperMUC (Phase 1)

[3] For information about Phase 2 see also „Anschreiben“.

1.3 Installation Dates

The expected schedule for installation of of SuperMUC is given in the following table.

Installation of a Migration System	July 2011
Installation Phase 1	2011 (can start after building is completed)
Decommissioning of HLRB II	October 1, 2011.
System Readiness (“Betriebsbereitschaft”)	May 31, 2012.
Installation of Phase 2	approx. two years after installation of Phase 1

The installation dates given above for the Phase 1 installation make the most optimistic assumptions on the availability of the essential hardware components (e.g., processors, RAID controllers, interconnect hardware).

Since “System Readiness” of Phase 1 is shifted to 2012, a migration system must be installed at the latest in July 2011.

1.4 Categorization of the requirements for SuperMUC

The requirements of this procurement are categorized into three classes which are marked by coloured boxes:

M	Requirements within red boxes and marked by "M" are considered **mandatory**. It will constitute an exclusion criterion if they cannot be satisfied by the offered system.[4] Typically these requirements include the phrase MUST or MUST NOT.

I	Requirements within yellow boxes and marked by "I" are considered **important** for the operation and usability of the system. ´ In case that such a requirement cannot be fulfilled, alternative offers and proposals are explicitly allowed as long as they provide appropriate replacements. The proposed alternatives and their implications must be explained in detail. Typically these requests include the phrase SHOULD or SHOULD NOT.

T	Requirements within green boxes and marked by "T" are considered **targets** for optimal usage of the system. If a vendor can fulfil these requirements this will lead to a qualitatively better ranking through the evaluation procedure. Typically these requests include the word MAY or OPTIONAL or PREFERRED.

Many requirements contain checkboxes.

[X] a checkbox marked by the tenderer with "X" means that the requirement/request can be fulfilled and the feature is included in the vendor's offer.

[ ] an empty checkbox indicates that that the requirement cannot be fulfilled

1.5 Terminology

The following terms are used in this document:

- **Phase 1:** refers to goods and services delivered for the first installation phase (in 2011).
- **Phase 2:** refers to the **additional** goods and services delivered for the second phase (in 2013/2014).
- **Combined Phases 1+2:** refers to aspects of the goods and services which are accomplished by the combination of Phase 1 and Phase 2 (e.g., performance of an application which runs on the entire system).
- **Island:** refers to a group of nodes with the best interconnect network characteristics in the system (e.g., blocking factor, bandwidth, latency). Between islands the network characteristics may be degraded. If the internal network has no hierarchical structure, then the term refers to just an arbitrary grouping of processors (e.g., for use within the benchmarking procedure).
- **kByte, MByte, GByte, TByte or PByte** in the context of memory, cache, disk and filesystems means 2^{10}, 2^{20}, 2^{30}, 2^{40}, 2^{50} Byte, resp. This also applies for memory/cache/storage related bandwidths.
- **Active storage components:** refers to electrically active storage components like disks, controllers or expansion cards but not passive backplanes or disk shelves

1.6 Requirements with respect to the content of the proposal

Please explain in your proposal how you intend to fulfil the given requirements. All items mentioned in the following chapters must be covered.

If not required otherwise, only the information for Phase 1 must be provided in detail.

For Phase 2, only conceptual descriptions must be given. All information relating to Phase 2 is considered as “intended” and is not part of the procurement[5]. However, the technical possibilities and the interoperabilities with such an upgrade are assessed.

---

[4] cf. "Bewerbungs- und Vertragsbedingungen"

[5] cf. “Anschreiben”

The answers should refer only to offered goods and services. Answers and explanations must be inserted into the appropriate sections of this document and should be as concise as possible. It is preferred that the vendor use the provided fields which are labelled by:

Answers and annotations by the vendor: [Insert text here]

Additionally, the tables in chapter 3 containing the key hardware parameters of the system must be filled in for Phase 1, and where applicable and known, for Phase 2 and/or for the combined Phases 1+2[6].

The vendor is free to provide additional information at the end of any paragraph.

Questions that have not been answered or have been answered insufficiently will be considered as answered in the negative.

Any relevant difference between Phase 1 and Phase 2[7] must be explained in this description.

Wherever appropriate and possible, there should be references to common standards of information technology. Bibliographical references, also to brochures and manuals, are desirable, but can only be considered as supplementary information, i.e., they cannot replace the required answers or explanations.

The vendor may hand in the specifications stipulated below in German or English or a mixture of both.

1.7 Additional material

The following material is referenced in this document:

- Specifications of the computer rooms
- DVD containing the floor plan of the computer room, the benchmarks, this document in MS Word and PDF formats

[6] For commitments relating to Phase 2 see “Anschreiben”

[7] For commitments relating to Phase 2 see “Anschreiben”

2 Technical Requirements for the SuperMUC

The following chapter stipulates the technical requirements, which the offered system needs to fulfil.

2.1 Installation requirements

Your proposal for installation must ensure that the offered system is physically installable and operable and it must allow a reliable comprehensible prediction of air conditioning, cooling demands, and energy consumption.

2.1.1 Proposal for installation

M 1:	A technical proposal for the installation for both phases must be provided. The proposal must include a detailed space assignment plan for Phase 1 and a conceptional plan for a potential Phase 2[8].

I 1:	Information about: - the earliest installation date of Phase 1 - the earliest installation date of Phase 2[9] - the time needed to complete the installations - the length of interruption of service needed for installation of Phase 2 should be submitted.

Answers and annotations by the vendor:

[Insert text here]

2.1.2 Installation

The offered system shall be installed in the designated computer room of the new LRZ building. The **freight elevator** has clearance dimensions of

- width x depth x height = 1,95 m x 2,95 m x 2,58 m
- and a maximum payload of 3000 kg

In order to enable a transport of the system components to the HRR room in the third floor, all system components should not exceed the following dimensions:

- width x depth x height = 1,60 m x 1,60 m x 2,58 m for cubic based components, or
- width x depth x height = 1,20 m x 2,50 m x 2,58 m for very deep or wide components

The floor plan of the new computer room of LRZ is included with the tender documents. This room is reserved for Phase 1 of the offered system; it is free of pillars and has the following dimensions and false floor specifications:

- width x depth x height = 24,5 m x 21 m x 8 m (= 514,5 m^2 x 8 m)
- false floor with a maximum area load of 1500 kg per m^2 (~ 15 kN/m^2),
  a maximum "*point load*" of 500 kg and a height of 1.8 m

After shutdown and decommissioning of HLRB II the facilities actually used by HLRB II will be available for the installation of Phase 2 as well as parts of Phase 1 if the latter does not completely fit into the room reserved for its installation. A floor plan containing details of this area is included in the tender documents. The **additional** floor space available for Phase 2 and the corresponding false floor specifications are as follows:

---

[8] For commitments relating to Phase 2 see "Anschreiben"

[9] For commitments relating to Phase 2 see "Anschreiben"

- width x depth x height = 28 m x 21 m x 8 m (= 588 m^2 x 8 m)
- false floor with a maximum area load of 1500 kg per m^2 (~ 15 kN/m^2), a maximum "*point load*" of 500 kg and a height of 1.8 m

Please note that at final state one single compute room free of pillars will host Phase 1 and Phase 2. The wall seperating the compute cube extension from the old building will be dismantled after the actual HLRB II is decommissioned.

M 2:	The transportation of the offered system (this includes Phase 1 and Phase 2[10]) to the place of installation must be feasible under the constraints mentioned above. The offered system must be installable in the designated computer rooms. This includes the constraints for floorspace, power and cooling, maintenance areas, and weight of the devices. By checking the box, the vendor also declares that the complete installation process and preparation for operation of the system (for both installation phases) is incumbent on the vendor[11]. Check here if the requirement is fulfilled: **[ ]**

2.1.3 Conformity with electrical standards

LRZ provides power supply of 50 Hz alternating current, with 230 V single-phase and 400 V three-phase voltages, as commonly available in Germany. LRZ provides IEC 60309 32 Ampere five-pole, three-phase power connectors in the false floor for the electrification of the system.

M 3:	If other electrical power characteristics are needed, the required frequency and voltage transformers must be included in the tender. The installation costs for such components is part of the contract and must not be charged separately. Check here if the requirement is fulfilled: **[ ]**

The *EC Directive on Electromagnetic Compatibility of Devices* is the guideline for all manufacturers, importers, and distributors of electric and electronic devices in the European Union as far as development, production and distribution of devices, systems and equipment are concerned. In addition, the *Low-Voltage Directives 73/23/EEC* and the *Act about Product Liability* must be considered. These regulations and the corresponding directives and standards are EU stipulations that have been adopted by German Law. By attaching a "CE marking" on a device, the manufacturer gives a legally binding declaration that the device conforms to all EMC relevant directives and standards of the EU and with the regulations and laws.

M 4:	The system and its components must conform to German and European directives and laws. All necessary documentation must be submitted to LRZ before delivery of the system. Check here that this requirement is fulfilled: **[ ]**

2.1.4 Power supply, cooling, computer room conditions

The usage of highly energy efficient cooling technologies and all other aspects leading to a high Data Center Infrastructure Efficiency (DCIE) is an important goal of the SuperMUC procurement. In this spirit aspects such as free cooling and a possible re-use of waste to effectively heat the LRZ buildings are important and will be honored respectively.

[10] For commitments relating to Phase 2 see „Anschreiben"

[11] LRZ only lends the necessary technical support.

2.1.4.1 Power supply

M 5:	The total electrical power requirement of Phase 1 of the SuperMUC must not exceed 6000 kW. This does not include the additional power needed for the cooling of the system and the compute room.[12] Check here that this requirement is fulfilled: **[ ]**

M 6:	The total electrical power requirement of the combined Phases 1+2 of the SuperMUC must not exceed 7150 kW. This does not include the additional power needed for the cooling of the system and the compute room.[13] Check here that this requirement is fulfilled: **[ ]**

2.1.4.2 Cooling

To reduce cost and to improve energy efficiency one major goal of LRZ is to use free cooling to an as large as possible extent for the cooling of its HPC infrastructure in future. Hence, two distinct water cooling loops called Loop1 and Loop2 operating at different temperature levels will be available for the cooling of the system. The cooling loops will be operated at the water temperatures listed below:

- **Loop 1: T_{in}=14°C; T_{out}= 20°C; ΔT=6K;**
- Loop 2: $T_{in} \geq 30°C$; $\Delta T \geq 6K$ (The minimum T_{in} that LRZ can guarantee throughout the year is 30°C);

Here T_{in} denotes the water inlet temperature to the compute racks and T_{out} denotes the preferred water temperatures at the outlet side of the racks. The maximum cooling power of these loops in Phase 1 will be 2,0 MW for Loop 1 and 6,0 MW for Loop 2 respectively. Due to its high inlet temperature, it will be possible to operate Loop 2 without any chillers only using the free cooling capacities at the roof of the building. Loop 1 and Loop 2 are closed internal cooling loops specially dedicated for the cooling of IT equipment and separated by heat exchangers from the water cooling loops connected to the cooling towers at the roof of the building. The water within these internal loops contains no glucol and only small amounts of additives such as anticorrosives and biocides are to be expected according to VDI 2035.

T 1:	Higher water inlet temperatures T_{in} than 39 °C are desired for compute nodes. Check here that this request is fulfilled: **[ ]**

M 7:	The type of cooling system for all devices (air-cooling, water-cooling, etc.) and details about the cooling system (e.g., in- and outlet temperature, temperature variation tolerance, pressure, purity requirements, etc.) and the required environmental conditions (e.g., ambient temperature, humidity, dust free conditions, etc.) must be specified. In addition the appropriate fields in the tables in chapter 3 must be filled in.

M 8:	The heat emission of all devices of the offered system, broken down into air and water cooling, i.e., the maximum values and estimated values for permanent load, must be specified. Also the heat emission released into each of the two separate water cooling loops Loop 1 and Loop 2 must be specified. In addition the appropriate fields in the tables in chapter 3 must be filled in.

[12] cf. “Anschreiben"

[13] cf. "Anschreiben"

M 9:	At least 90 % of the waste heat of the system must be removed by means of direct or indirect[14] water cooling of the components under the following environmental circumstances: **Loop 1: T_{in}=14°C** **Loop 2: T_{in}=30°C** **Room ambient temperature of 35°C** Check here that this request is fulfilled: **[ ]**

I 2:	The total heat emission of SuperMUC Phase 1 released into the water cooling loop called Loop1 should not exceed 1.5 MW Check here that this request is fulfilled: **[ ]**

I 3:	The heat emission released into the water cooling loop called Loop2 should not exceed 4 MW in Phase 1. Check here that this requirement is fulfilled: **[ ]**

I 4:	The heat emission released into the **room air** (please note that the air handlers for room air conditioning will be connected to Loop 1) should not exceed 400 kW in Phase 1 and 700 kW in Phase 2[15]. Check here that this request is fulfilled: **[ ]**

Answers and annotations by the vendor:

[Insert text here]

[14] All water cooled system solutions which are internally still using air to remove the heat from the server components are appointed as indirect water cooled solutions in this document.

[15] For commitments relating to Phase 2 see “Anschreiben”

2.2 Hardware of the compute nodes

2.2.1 Compute node architecture

M 10:	A detailed description of the compute node and processor architecture for Phase 1 must be given In addition the appropriate fields in the table in chapter 3.2 must be filled in.

The specifications of all components must be explained in detail and the following items should be covered:

- general description of the compute node technology
- multiprocessor node architecture (e.g., NUMA, ccNUMA, etc.), if applicable
- type of processor chips
- number of offered compute nodes, number of processors, number of processor cores and size of main memory
- maximum size of main memory
- number of execution units of cores
- size and associativity of caches and cacheline lengths
- clock rates, maximum number of floating and fixed point as well as general purpose operations per clock (separately for scalar and/or vector)
- number of registers (separately for floating point, integer and general purpose operations, and for scalar and/or vector operations) and detailed description of processor registers and their functionality
- number of execution units (e.g., arithmetic and general purpose units)
- if vector units are available, describe which data types are vectorized
- memory address width
- detailed description of hardware performance counters and their usage
- intra-node processor-processor interconnect network, i.e., latencies, bandwidths, network type, interconnect topology. if applicable: intra-node hardware performance counter
- number of memory controllers, memory type and memory read/write bandwidth available per processor chip
- number and type of PCI-E and/or PCI-E Gen2 and/or other (proprietary) slots
- number and type of network and I/O links
- RAS and Security features
- hardware sensors available for measuring power consumption, temperatures (and water flow rates if applicable)
- Upgradability and expandability
- virtualization support
-

I 5:	A conceptual description for the compute node and processor architecture of Phase 2[16] should be provided. The description of Phase 2 should at least cover the following technical aspects: Description of the compute node (e.g.,processor type, clock rate, architecture and type of processing units and number of cores per processing unit, memory, intra-node processor-processor interconnect network, PCI-E or equivalent interfaces) Description of the cooling solution of compute nodes (e.g., liquid/air and direct or indirect liquid cooling)

Answers and annotations by the vendor:

[Insert text here]

2.2.2 Configuration of the compute nodes

The offered compute nodes should achieve a high compute performance (optionally obtained by auto-parallelization and/or OpenMP directives). The main memory must have an appropriate size and bandwidth in relation to the compute power and the buffers needed for message passing.

[16] For commitments relating to Phase 2 see „Anschreiben“

M 11: The vendor must disclose information about the fraction of the main memory **typically** needed for buffers for message passing (MPI) :

Amount of memory per core _____ Mbytes

Amount o f memory per node _____ Mbytes

If appropriate: Specify the scaling behaviuor of the buffer size as a function of communication partners and provide several examples for typical regular patterns (e.g. 3-D nearest neighbour, all-to-all, the MPI settings as used for the benchmark runs, etc.). Discuss the tradeoff between performance and buffer size.

The amount of MPI buffering needed for the execution of the LINPACK benchmark run as a pure MPI program on the whole Phase 1 system must be disclosed.

Amount of memory per core _____ Mbytes

Amount of memory per node _____ Mbytes

If the vendor offers a substantially larger memory than required below, this will not automatically be honored by LRZ's evaluation scheme. Therefore the vendor should explain the rationale behind this overcommitment.

Alternative memory configurations may be proposed.

I 6: All thin compute nodes should consist of shared memory nodes with 16 or more cores.

Check here that this request is fulfilled: **[ ]**

I 7: The size of memory for the thin nodes should be between 0.25 GBytes and 0.5 GBytes per floating point instruction per cycle of a thin node.[17]

0.25 GBytes per (floating point instruction per cycle of a thin node are preferred, **if the overhead of buffers for system and MPI is sufficiently small.** This means it must be ensured that at least 75% of the memory will be available for large user applications (using the complete system) .

Check here that this request is fulfilled: **[ ]**

I 8: At least one island (or approx. 8192 cores) of SuperMUC Phase 1 should consist of “fat” shared memory nodes with 32 or more cores.

Software solutions are permitted as long as they have equivalent application performance characteristics as pure hardware based solutions.

Check here that this request is fulfilled: **[ ]**

I 9: The size of memory per core of the “fat” node island should be between 4 and 8 GBytes per core. 6 GBytes are preferred by LRZ**.**

Check here that this request is fulfilled: **[ ]**

Answers and annotations by the vendor:

[Insert text here]

[17] For a node with 16 cores and 8 floating point operations per cycle per core this translates to a memory of between 2 and 4 GBytes per core. In this case, 1.5 GBytes per core for the user application are considered sufficient.

2.2.3 Main memory

M 12:	The vendor must disclose detailed information about the memory subsystem and the cache hierarchy of the compute nodes: - cache coherency protocols (overhead, impact on performance) - memory access latencies - memory and cache (read and write) bandwidths - size and associativity of caches - memory topology - access characteristics (cache line sizes, maximum number of outstanding misses, prefetches, memory interleaving, etc.) - TLB characteristics In addition the appropriate fields in the tables in chapter 3.2 must be filled in.

T 2:	Memory balance: It is desired that the theoretical memory bandwidth of a node exceeds the value of **0.3 Byte/s per 1.0 Flop/s** of theoretical peak performance. Check here that this request is fulfilled: **[ ]**

Please note that the required bandwidth for many scientific applications is well above 3 Bytes/s per Flop/s. Therefore larger bandwidth to memory than the desired one will lead to an upvaluation by the LRZ evaluation scheme. If the requirement for memory bandwidth cannot be fulfilled, large caches can be offered to compensate.

Answers and annotations by the vendor:

[Insert text here]

2.2.4 Accelerators

For some applications accelerators (particularly GPGPUs) might be an appropriate way to increase their performance on the system.

M 13:	It must be described which types of accelerators could be installed in the system, which ones are supported by the vendor, and how the accelerators can be connected to the compute nodes.

I 10:	The costs for adding in addition 100 accelerators to the thin compute nodes should be specified (Please note that accelerators are not part of the SuperMUC procurement!): Addional cost for 100 accelerators: **[Insert text here]** EUR. The additional power requirement of these 100 accelerators should be specified: Addional power requirement: **[Insert text here]** kW.

I 11:	It should be possible that the accelerators can be shut down or put into a low power standby mode when they are not needed by jobs or applications running on the nodes equipped with them. Check here that this request is fulfilled: **[ ]**

Answers and annotations by the vendor:

[Insert text here]

2.2.5 Software Solutions for Shared Memory nodes

T 3:	The possibility to aggregate several nodes into a larger virtual Shared Memory node (e.g. ScaleMP) is desired. The realisation and tradeoffs should be discussed.

Answers and annotations by the vendor:

[Insert text here]

2.2.6 Service and Login nodes

Login nodes will be used for:

- User login and all types of general interactive user activity
- Submission of batch jobs
- Compilation
- User triggered archiving of results

Service nodes will be used for:

- License serving
- Batch queuing system
- Data bases
- Backup of NAS data
- System monitoring and management

The login as well as service nodes may differ from compute nodes but should contain the same processor architecture. A standard Linux distribution should be supported for these system resources.

I 12:	In total at least **fourteen** 16-core Shared Memory nodes with a total memory of at least 1.2 TByte should be included in the offer for login and interactive work as well as other system services such as resource, backup and archiving of user data to tape as well as user management. These special nodes should include local disks for scratch and, swap and paging. The service and login nodes should have access to the user file systems of SuperMUC (home and parallel files system). The compilers, tools and libraries for program development should be available. The offer should include at least 8 login nodes containing 128 GBytes of memory each. At least two login nodes should be connected via SAN to the LRZ backup and archive system with an accumulated bandwidth of 2 Gbyte/s (see also chapter 2.5.8). The processor architectures used in the login nodes should reflect all processor architecture used in the compute nodes. The memory configuration of the service nodes will be adjusted later on to the needs of the service running on these resources. Check here that this request is fulfilled: **[ ]**

2.3 Communication network

2.3.1 Details of internal interconnect

A hierarchical interconnect may be provided consisting of islands of highly interconnected compute nodes with non-blocking or nearly optimal communication links between the nodes within an island.

T 4:	The preferred size of such an island would be equal to or somewhat larger than 8192 cores.

The link bandwidth and the number of links for the communication between these islands may have smaller aggregate bandwidth than within an island (pruned interconnect) depending on the costs and technology which is used.

M 14:	A detailed description of the internal interconnect of Phase1 (including the interconnect between the islands) must be provided: In addition the appropriate fields in the tables in chapter 3 must be filled in.

The following items must be disclosed:

- general description of network structure and node connections
- technology
- material of cables (copper, optical, etc.)
- topology
- power requirements
- bandwidth and latency (where appropriate, taking into account distance between nodes)
- scalability both in term of performance and of cost with number of compute nodes
- limitations and resource usage (e.g., number of nodes, cores, or tasks; message length; number of outstanding messages; buffer sizes etc.)
- performance of all involved single components
- limitations and resource usage (e.g., maximum number of compute nodes, message length, number of outstanding messages, buffer sizes)
- maximum time for a broadcast to reach all nodes
- whether dynamic routing is possible
- network (hardware) performance counter
- details of the interconnenct between the islands
- performance characteristics
- hardware sensors available for measuring power consumption and temperatures

I 13:	A conceptual description of the interconnect within Phase 2[18] and of the interconnect between Phase 1 and Phase 2 should be given. This description should the following technical aspects: - Network technology and topology - Performance characteristics (e.g., bandwidths and latencies) - Capabilities , e.g. dynamic routing

Answers and annotations by the vendor:

[Insert text here]

I 14:	It should be possible to off-load MPI functionality to devices of the interconnect. Check here that this request is fulfilled: **[ ]**

If available, a full description of the functionality must be provided.

[18] For commitments relating to Phase 2 see „Anschreiben“

Answers and annotations by the vendor:

[Insert text here]

T 5:	Hardware support for PGAS languages such as CAF & UPC is desired (see chapter 2.10.4).
	Check here that this request is fulfilled: **[ ]**

If available, a description of the functionality must be provided.

Answers and annotations by the vendor:

[Insert text here]

2.3.2 Interconnect within an island

I 15:	The performance of the interconnect within Phase 1 and within Phase 2[19], respectively, must be appropriate for the compute performance of the nodes The interconnect balance = aggregate theoretical link bandwidth (sum of all links, in+out) of a node divided by aggregate theoretical memory bandwidth of a node should exceed 1/30 (GByte/s / GByte/s) Check here that this request is fulfilled: **[ ]**

I 16:	The MPI latency between two arbitrary nodes (running at least one MPI thread per node) within an island should be less than 2 µs[20]. Check here that this request is fulfilled: **[ ]**

T 6:	It is desired that the MPI latency between two arbitrary nodes within an island is less than 1.5 µs. Check here that this request is fulfilled: **[ ]**

Answers and annotations by the vendor:

[Insert text here]

2.3.3 Interconnect between islands

The bisection bandwidth (measured with the same number of nodes) between islands may be less than the bisection bandwidth within one island ("pruned interconnect").

---

[19] For commitments relating to Phase 2 see “Anschreiben”

[20] The vendor may choose a method for verification, but must disclose this.

I 17:	The interconnect bandwidth between islands should not be pruned by a factor of more than 4:1 (intra:inter) Check here that this request is fulfilled: **[ ]**

I 18:	Quantify the pruning factor between the islands: insert here: _____ **: 1** (intra:inter)

I 19:	It should be possible to run a single user application on any combination of thin and fat nodes. Check here that this request is fulfilled: **[ ]**

T 7:	It is desired that the MPI latency between two arbitrary nodes of different islands should be less than 2 µs. Check here that this request is fulfilled: **[ ]**

2.3.4 Interconnect between Phase 1 and Phase 2[21]

I 20:	The interconnect between the two phases of SuperMUC may be different from that within Phase 1 or Phase 2, respectively, but it should be interoperable. Notwithstanding, running a parallel application across phase boundaries (e.g., LINPACK) should be possible in a transparent manner and with sufficient performance. Check here that this request is fulfilled: **[ ]**

I 21:	A conceptual description of the interconnect within Phase 2 and of the interconnect between Phase 1 and Phase 2 should be given. This description should at least cover the following technical aspects: - Network technology and topology - Performance characteristics (e.g., hardware and MPI link bandwidths and latencies) - Capabilies like dynamic routing etc.

I 22:	The interconnect bandwidth between Phase1 and Phase 2 should not be pruned by a factor of more than 8:1 (intra: inter) compared to that within an island.

	The appropriate fields in the table in chapter 3.4 should be filled in. Check here that this request is fulfilled: **[ ]**

T 8:	It is desired, that the MPI latency, between two arbitrary nodes of Phase 1 and of Phase 2 should not exceed 4 µs. Check here that this request is fulfilled: **[ ]**

Please describe the technical solution and potential characteristics.

[Insert text here]

[21] For commitments relating to Phase 2 see “Anschreiben”

2.3.5 Intelligent routing

I 23:	Intelligent network routing mechanisms such as error detection, congestion control and automatic re-routing of network packets in case of link failure or link congestion should be available and should be described. Check here that this request is fulfilled:	**[ ]**

If available, please describe the technical solutions and potential characteristics.

[Insert text here]

I 24:	Monitoring software for the network traffic should be available. Check here that this request is fulfilled:	**[ ]**

If available, please describe the technical solutions and potential characteristics.

[Insert text here]

2.3.6 Handling of I/O by the interconnect, special purpose interconnects and networks

If the interconnect is also used for I/O, then the vendor should discuss the implications and describe how this could affect the performance of applications and what steps can be taken to prevent such disturbances.

T 9:	It is desired that I/O will be performed over a different network than the MPI traffic and/or that means to separate and/or prioritize the application traffic are provided. Check here that this request is fulfilled:	**[ ]**

If available, please describe the technical solutions and potential characteristics.

[Insert text here]

I 25:	A dedicated network for OS services and system monitoring should be available. Check here that this request is fulfilled:	**[ ]**

If available, please describe the technical solutions in detail.

[Insert text here]

2.3.7 Connection to external networks

LRZ will provide a 40 Gbit Ethernet backbone infrastructure for the connection of SuperMUC to the outside world.

I 26:	SuperMUC Phase 1 should be connected to the outside world with an aggregated Ethernet network link bandwidth of at least 80 Gigabits per second. All network components necessary for the connection of the system to external networks should be included in the offer. Check here that this request is fulfilled: **[ ]**

I 27:	If link aggregation is used to fulfil the requested Ethernet network bandwidth, the solution should fully support the Link Aggregation Control Protocol (LACP, IEEE 802.3ad). Check here that this request is fulfilled: **[ ]**

I 28:	The connection technology to the outside world and to the visualisation systems should be described.

M 15:	The appropriate fields for external networks in the table in chapter 3.4 must be filled in.

Please describe in detail the connection of SuperMUC Phase 1 to external networks.

[Insert text here]

2.4 Configuration of Phase 2 and the combined Phases 1+2[22]

2.4.1 Extent of inhomogeneity of Phase 1 and Phase 2[23]

The SuperMUC may be inhomogeneous in the two installation phases e.g., processors of different clock frequencies may be used for Phase 1 and Phase 2 and/or shared memory nodes with different numbers of cores, processors and/or sockets may be used. An upgrade or exchange of parts of Phase 1 is also possible, not only resulting in higher performance but also in lower electrical power requirements und thus in improved TCO.

The transfer of applications between Phase 1 and 2 must be easily possible. In any case, the installation of Phase 2 must not result in an essential change of the programming paradigm i.e., programs must run with a comparable efficiency and without major alterations at the program source level.

I 29:	It should be possible to run a single application without major alterations at the source code level (e.g., LINPACK) across the entire system with sufficient performance. Check here that this request is fulfilled: **[ ]**

I 30:	It should be possible to access all user file systems from both phases. Check here that this request is fulfilled: **[ ]**

If it is not obvious in itself, please explain how these requirements are fulfilled.

Answers and annotations by the vendor:

[Insert text here]

[22] For commitments relating to Phase 2 see “Anschreiben”

[23] For commitments relating to Phase 2 see “Anschreiben”

2.4.2 Possibility of Exchange or Upgrade of Phase 1

Because energy costs consume a substantial fraction of the budget, it may be commercially advantageous to exchange or upgrade parts of Phase 1 at a later point in time with more energy-efficient technology. At the same time this may give the vendor the opportunity to offer more compute power taking TCO into account.

It must be clearly understood, that any potential improvement in performance must be based on the improvements in hardware and not on improvements in software like compiler technology, therefore the performance of the benchmarks must be re-measured shortly before and then after an optional upgrade. The interruption time for the upgrade must be appropriately compensated by higher performance.

Preferably, an optional upgrade of Phase 1 should be performed in conjunction with the installation of Phase 2.

M 16:	The performance of an optionally upgraded Phase1 (as defined by the benchmarks) must be at least that before the upgrade. Check here that this requirement is accepted **[ ]**

In case that a technology upgrade is planed for Phase 1, the system technology of the upgraded Phase 1 must be disclosed.

Answers and annotations by the vendor:

[Insert text here]

2.4.3 Performance of Phase 2[24]

At this point, it is not intended to fix the exact values of peak performance, memory bandwidth, interconnect or individual benchmarks for Phase 2 or for the combined Phases 1+2. This keeps the opportunity to negotiate the exact configuration in the forefront of installation of Phase 2.

It should be clearly understood by the vendor that the final system should be similarly balanced as Phase 1. Without specifying exact values now, all components (particularly the interconnect) should be configured in conformance with this aim.

The performance of Phase 2 in comparison to the performance of Phase 1 will be evaluated as an **improvement ratio (IR)**.

$$IR = \frac{Aggregate\ Performance\ of\ Phase\ 2}{Aggregate\ Performance\ of\ Phase\ 1}$$

The Aggregate Performance of Phase 1 will be measured during the acceptance test for Phase 1.

Details for the ascertainment of the improvement ratio are given in *Decision Criteria and Benchmark Description* in section 5.2.3.

I 31:	The improvement ratio for the added Phase 2[25] as defined in the benchmark section should be at least: **1.4** i.e., the performance of the combined Phases 1+2 is 2.4 times that of Phase 1. Check here that this requirement is fulfilled: **[ ]**

[24] For commitments relating to Phase 2 see “Anschreiben”

[25] For commitments relating to Phase 2 see “Anschreiben”

2.5 Storage

2.5.1 Storage requirements overview

Two different types of file systems must be available on SuperMUC (see 1.2):

- **Home file system** with the following characteristics: moderate bandwidth, high metadata performance, interoperability using standard NFS/CIFS protocols, replication and NDMP backup.
- **Parallel file system** with the following characteristics: high bandwidth, moderate metadata rate, support for MPI-IO and other parallel access methods. No direct possibilities for backup or replication are required.

Both file systems may be technically implemented in the same manner, but they must not share any storage or storage networking components. An additional (second) parallel file system may be implemented for Phase 2 if an extension of the filesystem from Phase 1 is not possible.

A preferred implementation for the home file system is Network Attached Storage (NAS).

M 17:	The offered storage and file system solutions must be described in detail. In addition the appropriate fields in the table in chapter 3.5 must be filled in.

Please describe the offered storage and file system solutions in detail. The description must at least contain the following information:

- Storage devices (hard disks, flash memory) used in the solution
- Type of file systems
- Relevant features, such as hardware specifications, reliability, management software and interfaces, functionalities for protection against data loss (e.g., RAID, redundancy and checksumming, replication and snapshot features)
- Hardware sensors available for measuring power consumption and temperatures
- Performance implications of rebuild operations after storage device failures as seen from file system leveltype of file system
- Upgrade / extension path and process from the capacity and performance requirements of Phase 1 to Phase 2[26]

Answers and annotations by the vendor:

[Insert text here]

2.5.2 Capacity and performance requirements

M 18:	The total usable size (as reported by the df command) of the file systems must be at least For Phase 1: 2 PB in the home file system and 2 PB for replication (4 PB total) 10 PB in the parallel file system. For the combined Phases 1 + 2[27]: 5 PB in the home file system and 5 PB for replication (10 PB total) 20 PB in the parallel file systems (max. two filesystems are allowed) Check here that this requirement is fulfilled: **[ ]**

[26] For commitments relating to Phase 2 see „Anschreiben“

[27] For commitments relating to Phase 2 see „Anschreiben“

I 32: The read and write bandwidth (as defined in *Decision Criteria and Benchmark Description* by the benchmarks in sections 6.6.1 and 6.6.2) should be at least:

For Phase 1:
10 GB/s for the primary part of the home file system using NFS
2 GB/s for the secondary part of the home file system using NFS
200 GB/s for the parallel file system

For the combined Phases 1 + 2[28]:
15 GB/s for the primary part of home file system using NFS
3 GB/s for the secondary part of the home file system using NFS
400 GB/s for the parallel file system(s)

Remark: Client-Server data integrity checking and replication can be turned off during the benchmarks. Two parallel filesystems are allowed in Phase 2 to achieve 400 GB/s.

Check here that this request is fulfilled: **[ ]**

I 33: The aggregate metadata performance (as defined in *Decision Criteria and Benchmark Description* by the benchmarks in section 6.6.3) should be at least:

For Phase 1:
100000 file creations per second for the primary part of the home file system using NFS
20000 file creations per second for the parallel file system

For for the combined Phases 1 + 2[29]:
150000 file creations per second for the primary part of the home file system using NFS
30000 file creations per second for each exiting parallel file system

Remark: Client-Server data integrity checking and replication can be turned off during the benchmarks. Metadata performance for the secondary part of the home file systeme is not specified.

Check here that this request is fulfilled: **[ ]**

2.5.3 Access to file systems

M 19: The home file system and the parallel file system must be available and accessible on all compute and login nodes.

Check here that this requirement is fulfilled: **[ ]**

I 34: The home file system should present a single namespace (one mountpoint on the client). If namespace aggregation is used, it must be possible to create at least 3000 different areas (representing different user projects) at the root (top) namespace level.
Check here that this request is fulfilled: **[ ]**

I 35: The parallel file system should present a single namespace or at least a single namespace for each of the two allowed filesystems in Phase 2[30].

Check here that this request is fulfilled: **[ ]**

[28] For commitments relating to Phase 2 see „Anschreiben“

[29] For commitments relating to Phase 2 see „Anschreiben“

[30] For commitments relating to Phase 2 see „Anschreiben“

M 20:	The home file system must be accessible using NFS v3 and NFS v4 over TCP. For NFSv4, Kerberos authentication against an Active Directory Server is required. Check here that this requirement is fulfilled: **[ ]**

I 36:	The home file system should be accessible using the CIFS protocol. Check here that this request is fulfilled: **[ ]**

M 21:	It must be possible to use MPI-IO to/from the parallel file system with appropriate performance. Check here that this requirement is fulfilled: **[ ]**

T 10:	It is desired that functionality of MPI-IO can be used with the home file system Check here that this request is fulfilled: **[ ]**

Answers and annotations by the vendor:

[Insert text here]

2.5.4 Redundancy and data protection

M 22:	All persistent storage (storage devices which are designed to store data permanently, e.g. disks or SSDs) in the file systems must be protected against at least two simultaneous failures (e.g., using RAID 6 or RAID DP or other mechanisms). Check here that this requirement is fulfilled: **[ ]**

M 23:	Automatic repair (recreation of redundancy) of failed persistent storage devices must be possible without manual intervention (e.g., using hot-spares or spare capacity). Check here that this requirement is fulfilled: **[ ]**

M 24:	If write caches are used in the solution, they must be redundant (e.g., mirrored). Check here that this requirement is fulfilled: **[ ]**

M 25:	All active components in the storage systems must be redundant without any single point of failure. Any single hardware failure in the storage systems must be handled and reported to a monitoring system without operator intervention. It must be tolerated in a way that does not cause I/O errors or data loss for running applications that access the file system. Check here that this requirement is fulfilled: **[ ]**

M 26:	The home file system must be accessible independent from the status of the compute nodes of SuperMUC or the parallel file system. Check here that this requirement is fulfilled: **[ ]**

I 37:	The primary home file system storage will be located **in another computer room** and will be operated with static UPS and diesel generator backed-up power. The home file system storage should be asynchronously replicated to a secondary storage system in **the SuperMUC** computer room which can be configured to fulfil the duties of the primary system in case of a catastrophic failure of the latter. The replication technology should be able to handle - 2 PB of data and 1 billion files in Phase 1 - 5 PB of data and 2 billion files in Phase 1+2. An aggregated replication performance of at least - 2 GB/s in Phase 1 - 3 GB/s in Phase 1+2 is required between the primary and the secondary part of the home file system. Check here that this request is fulfilled: [ ]

Please describe the replication solution in detail. The description should at least contain the following information:

- Architecture of the solution
- The configuration of the system when a 24-hour replication interval for all data in the home file system is desired
- Process of recovering from a total loss of the primary storage system
- Dependency of replication performance on the amount of data and the number of files

Answers and annotations by the vendor:

[Insert text here]

M 27:	The home file system must offer the option to keep at least 10 snapshots of all data stored on the primary part of the system. Check here that this requirement is fulfilled: **[ ]**

M 28:	All file systems must support a mechanism which allows a fast restart after system crashes or power failures (e.g., journaling of meta- and/or user data). Check here if the requirement is fulfilled: **[ ]**

The mechanisms available for fast file system recovery after crashes or power failures as well as the recovery process, recovery time and the amount of data loss in case of

- a crash of a file system client
- a crash of a file system server (if applicable)
- a temporary or permanent failure of a storage subsystem

must be described.

Please also disclose information concerning the seamless file system degradation in case of client, server (if applicable) or storage subsystem failure.

Answers and annotations by the vendor:

[Insert text here]

M 29:	The home file system must support data backup and restore using NDMP controlled by Tivoli Storage Management (TSM) servers (see also 2.5.8). Check here that this requirement is fulfilled: **[ ]**

2.5.5 Data integrity

I 38:	Means to ensure end-to-end data integrity (from client to disk) should be available for both file systems. A segmented (e.g., client-to-server + server-to-disk) implementation is allowed if it covers the whole path. Detected problems should be reported. Check here that this request is fulfilled: **[ ]**

Please describe in detail how data integrity is checked from the client to the disk.

Answers and annotations by the vendor:

[Insert text here]

M 30:	Both file systems must support storage device scrubbing (or alternative methods) which detects and corrects defective blocks. Any abnormal events must be reported. Check here that this requirement is fulfilled: **[ ]**

M 31:	The home file system must support file system checksums which can detect data corruption and misplaced blocks (e.g., “lost writes”) Note: disk sector level checksums alone do not fulfill this criterion. Check here that this requirement is fulfilled: **[ ]**

M 32:	File system consistency checking programs for all file systems are required. Check here if the requirement is fulfilled: **[ ]**

Please describe the possible procedures in detail as well as the performance of the file system checking mechanism in case a check is necessary (e.g., after media failures).

If not mentioned before, please also disclose

- facilities for online detection, correction and reporting of media errors
- any implication of the detection mechanisms on I/O performance.

Answers and annotations by the vendor:

[Insert text here]

2.5.6 File system scalability and parallelism

I 39:	The possible size for a single file in the parallel file system should be at least 250 TB. Check here that this request is fulfilled: **[ ]**

Please describe potential side effects of a configuration which allows 250 TB files.

I 40:	The scalability of all offered file system solutions should be described. The following items should be included: - maximum file system size - maximum number of files and directories - maximum numer of files and directories per directory - maximum number of sub-filesystems (if applicable) - maximum file size - maximum I/O bandwidth achievable - maximum meta data operations achievable (e.g., file creates per second) - maximum number of client nodes which can be connected to one file system -maximum number of files per directory where 90% of maximum metadata performance can be sustained

I 41:	File system parallelism: The granularity (e.g., per file, per file system) and options for tuning the parallelism of the file system should be disclosed: - multi-process access of small files - single-process access of large files - multi-process access of large files

I 42:	File system caching mechanisms: Please describe the caching model for data and metadata on clients and servers in detail, e.g., - how the file system caches interact with OS caches and - how to use or bypass caching (if applicable) as well as - any impacts on I/O performance.

Answers and annotations by the vendor:

[Insert text here]

2.5.7 Management tools for the disk subsystems and file systems

M 33:	Facilities for the efficient and scalable monitoring and management of the file systems and the underlying disks must be available, documented and accessible to LRZ administrators. This includes current file system and storage system state, events and real-time access to performance data. Check here that this requirement is fulfilled: **[ ]**

The facilities for the efficient and scalable monitoring and management of the file systems and the underlying storage systems and disks must be described in detail.

[Insert text here]

T 11:	GUI-based management and/or monitoring tools for the disk systems are desired. Check here that this request is fulfilled: **[ ]**

If available, please describe them.

[Insert text here]

I 43:	All file systems should support per-user, per-group and per-directory quotas with respect to data size and the number of files. A reporting facility should be available. Check here that this request is fulfilled: **[ ]**

Please describe in detail the user and group quota mechanism.

[Insert text here]

I 44:	All file systems should support extended Access Control Lists (ACLs) Check here that this request is fulfilled: **[ ]**

Please describe in detail the supported ACLs.

[Insert text here]

2.5.8 Tivoli Storage Management for Archive and Backup

I 45:	A fully functional TSM (Tivoli Storage Management) client should be available for the OS release used by SuperMUC to backup/archive data from both file systems. The license costs for TSM are already covered by LRZ campus licenses. Check here if the requirement is fulfilled: **[ ]**

I 46:	The TSM client supporting LAN-less backup/archive should be available for the target architecture. Check here if the request is fulfilled: **[ ]**

T 12:	A TSM client supporting serverless backup/archive is desired. Check here that this request is fulfilled: **[ ]**

If applicable please describe special backup options available, e.g.,

- snapshot support
- efficient replication
- fast search for modified files
- support for parallel backup.

Answers and annotations by the vendor:

[Insert text here]

2.5.9 GPFS-Client

I 47:	For the interoperability with other European supercomputing sites, a GPFS client should be available. Check here that this request is fulfilled: **[ ]**

I 48:	The costs for the GPFS client licences including maintenance fees for the whole system should be specified. Addional cost for GPFS client licenses: [Insert text here] EUR. Addional cost for GPFS client maintenance: [Insert text here] EUR.

2.5.10 Interoperability with a Hierarchical Storage System

A concept for hierarchical storage management (HSM) that can interoperate with SuperMUC may be described. The solution itself is not part of this procurement. If available, please describe the interoperation, implementation and functionality.

Answers and annotations by the vendor:

[Insert text here]

2.5.11 Support structures for offered storage and file system solutions

I 49:	Please describe the support structures for the offered storage and file system solutions in detail. The description must at least contain the following information: - Who is responsible for handling first, second and third level problem calls, - Who is responsible for hardware and software enhancements of the proposed storage and file system Solutions.

Answers and annotations by the vendor:

[Insert text here]

2.6 Reliability and fault tolerance

2.6.1 Detection of hardware faults

M 34:	Facilities for the online detection and monitoring of hardware errors (e.g., faulty memory modules, processors, fans, network links, and switches) must be provided. Check here that this requirement is fulfilled: **[ ]**

Particularly describe:

- the facilities and processes for diagnosis and error correction
- whether there are any spare parts and how a high availability of the total system will be realized

Answers and annotations by the vendor:

[Insert text here]

2.6.2 Mean time to Interrupt

M 35:	An estimate and rationale for the expected mean time to interrupt leading to aborts of user jobs for the overall system and for a single 8k cores island must be given.

Answers and annotations by the vendor:

[Insert text here]

2.6.3 Fault Isolation

I 50:	A seamless degradation of the system should be feasible in case of a failure of a single hardware component, like a network switch, network link, compute node or I/O node. Check here that this request is fulfilled: **[ ]**

Please describe in detail how the operation of the system with lower performance in case of failure of a component (e.g., a CPU, a compute node, an I/O node, part of the main memory, part of the I/O subsystem) will continue. Please describe how the repaired components are returned to operation. In which cases is this possible without interruption of normal operations, and in which cases can an interruption not be avoided?

Describe which components may cause an interruption of the complete system or larger proportions of it e.g., central switches. Describe the strategies for problem resolution and the provisioning of spare parts.

Answers and annotations by the vendor:

[Insert text here]

2.7 Operating system

System components for resource administration and batch administration as well as other components that are essential for system operation are considered part of the operating system.

2.7.1 Standards

M 36:	The operating system must be complete, stable and appropriate for production usage in a supercomputing center. It must allow flexible administration and control of jobs in interactive and batch mode. The system must be delivered with a 64-bit operating system using a 64-bit kernel. Check here that this requirement is fulfilled: **[ ]**

I 51:	If a light-weight OS or micro-kernels are used on the compute nodes, the vendor should disclose the capabilities and the differences to a full featured Linux OS.

Answers and annotations by the vendor:

[Insert text here]

I 52:	The operating system should be based on Linux and should be compatible with the X/Open Standard POSIX 1003 (ISO/IEC 9945). The OS licenses and maintenance fees should be part of the proposal. Check here that this request is fulfilled: **[ ]**

I 53:	It should be possible to use IPv6 for all SuperMUC services which have to be reached from external networks or which must connect to external services (e.g., login, ntp, DNS, …). Check here that this request is fulfilled: **[ ]**

T 13:	The OS preferred by LRZ is "SuSE Linux Enterprise Server (SLES)". Check here that this request is fulfilled: **[ ]**

Answers and annotations by the vendor:

[Insert text here]

2.7.2 Diskless nodes

The thin and fat compute nodes may be operated without disks, since there is no intent to use swap space on the compute nodes.

T 14:	The diskless provisioning and operation of the compute nodes should be discussed.

Answers and annotations by the vendor:

[Insert text here]

2.7.3 Checkpoint/Restart & Suspend/Resume

As a provision against unplanned interrupts, the availability of checkpoint/restart facilities may support efficient job processing.

If available, describe the possibilities for checkpointing including information about.

- possibilities of system wide, preventive checkpointing
- multi-threaded or MPI applications
- integration with the batch-system
- restart on different nodes than those the checkpoint was written from
- limitations

Suspending a job means interrupting execution of a job and keeping the job and its environment in virtual memory (e.g., the the paging space if available), so that all processor resources used by this job are released and can be used by other jobs. It is especially useful for short term scheduling of jobs on a filled machine or for scheduling large job by suspending a few smaller ones.

If available, describe the capabilities of job suspend/resume.

Answers and annotations by the vendor:

[Insert text here]

2.7.4 Operating System induced scheduling noise (OS jitter)

On many computers, system activities can run without interfering with the application as long as there is a spare processor available in each node to absorb them. If there is no spare processor, an application-assigned processor may be temporarily used by the OS to handle the system activity. Doing so in an unsynchronised manner will introduce a performance reduction, particularly harmful for large jobs.

I 54:	Global thread-level synchronisation mechanisms for system services such as clock, OS deamons, semaphores/barriers should be available. Check here that this request is fulfilled: **[ ]**

I 55:	Other provisions to minimize system activities that may disturb user applications should be available e.g., stripped OS. Check here that this request is fulfilled: **[ ]**

If available, please describe these provisions.

Answers and annotations by the vendor:

[Insert text here]

2.8 Job Management

A highly scalable resource manager and scheduling system is crucial for the smooth operation of a petascale system. It must be tolerant against system failures, including failure of the node executing its control functions.

The term *Batch System*, as used here, denotes the software for optimizing the computational workloads by combining resource managers and scheduling systems.

M 37:	The batch system, its components, capabilities and restrictions must be described.

I 56:	The batch system should be able to handle hundred thousands of processing cores and ten thousands of compute nodes. Check here that this request is fulfilled: **[ ]**

I 57:	The batch system should be able to classify batch jobs into different job classes. Depending on the chosen job policies, it should be possible to charge for used resources on a per-user, per-project and on a per-job basis. Check here that this request is fulfilled: **[ ]**

I 58:	The batch system should be capable of starting very large jobs in a reasonably short time without too much loss from empty nodes (e.g., through backfill scheduling). Check here that this request is fulfilled: **[ ]**

I 59:	The batch system should be capable of supporting Grid software, particularly Globus and UNICORE. Check here that this request is fulfilled: **[ ]**

I 60:	The batch system should be **aware of the topology of the interconnect** i.e., be able to perform an optimal placement of jobs to nodes and islands and be able to interoperate with the MPI-implementation's and the operating system's facilities to perform optimal placement of tasks and threads on each compute node. Check here that this request is fulfilled: **[ ]**

I 61:	The batch scheduling system should be **energy-aware** i.e., it should support switching unused components to energy-saving mode (deep sleep mode, etc.) as well as the application dependant optimization of processor frequencies. Check here that this request is fulfilled: **[ ]**

I 62:	The batch system should support the control of resource limits at the node level, e.g. memory usage. Check here that this request is fulfilled: **[ ]**

I 63:	The batch system should support the reservation and allocation of nodes for interactive usage. Check here that this request is fulfilled: **[ ]**

I 64:	Means to prevent normal user login to batch nodes should be available. The MPI implementation should be able to cope with this restriction. Check here that this request is fulfilled: **[ ]**

I 65:	Means to gather OS information of nodes running a batch job should be available for the users. Check here that this request is fulfilled: **[ ]**

Please describe in details the batch system, its components, capabilities and restrictions.

Answers and annotations by the vendor:

[Insert text here]

2.9 System Administration and Monitoring

2.9.1 Configuration Management

All offered administrative programs must provide the capability to efficiently, securely, reliably, and scalably perform administrative measures across the complete system.

M 38:	The capabilities for configuration management (particularly with respect ot the above properties) must be described.

I 66:	Mechanisms for the system administrator should be available that allow the detection of divergences from a well-defined configuration (e.g., alterations of access rights or exchanged files or commands). Check here that this request is fulfilled: **[ ]**

Please describe the facilities for configuration management.

Answers and annotations by the vendor:

[Insert text here]

2.9.2 System monitoring

I 67:	An error tracking and reporting mechanism in cases of operating system errors (e.g., system dumps) should be available and supported by the vendor. Check here that this request is fulfilled: **[ ]**

I 68:	Detailed and complete log files that store all security relevant incidents (e.g., every command needing extended system rights and its context, such as parameters and options) should be available. The format of log files should be uniform, informative, readable and secure. Facilities should be available to process the log files efficiently and scalably across the whole system. Suitable interfaces for the processing of log files should be provided. Check here that this request is fulfilled: **[ ]**

I 69:	Scalable analysis tools interpreting the internal state of the system (e.g., tables of the OS) as well as monitors for the measurement of important performance parameters (I/O behavior, disk access behavior, CPU load, memory load, paging rate, etc.) should be provided with an easy-to-interpret output. Check here that this request is fulfilled: **[ ]**

T 15:	Graphical high level system overwiew with the ability to zoom into a detailed display of interesting components (I/O behavior, disk access behavior, CPU load, memory load, paging rate, etc.) are desired. Check here that this request is fulfilled: **[ ]**

I 70:	Scalable tools to extract system-wide hardware performance information like number of floating point operations, number of integer operations, memory references, etc. (like pfmon) should be available and supported by the vendor. This information should be available to the system administrators on a per CPU basis without any major impact on user codes and without the necessity of any specific changes to those codes. Check here that this request is fulfilled: **[ ]**

I 71:	It should be possible to monitor the interconnect of the system e.g., determine the number and size of packets and the amount of data sent or received. Check here that this request is fulfilled: **[ ]**

I 72:	Detection of hardware errors should be carried out automatically and be automatically reported to the vendor. Check here that this request is fulfilled: **[ ]**

I 73:	Hardware sensors or other means to separately measure the power consumption and temperatures of all compute nodes, network switches and storage should be delivered with the system. Check here that this request is fulfilled: **[ ]**

I 74:	Scalable tools to extract, accumulate and display information such as the power consumption and temperatures of compute nodes, network switches and storage should be available and supported by the vendor. This information should be available to the system administrators on a per node, per network switch and per disk shelf and disk controller basis. Check here that this request is fulfilled: **[ ]**

T 16:	Integration and/or interoperability of the offered monitoring tools with Nagios is desired. Check here that this request is fulfilled: **[ ]**

T 17:	For integration with other monitoring tools used at LRZ, an SNMP agent with Management Information Bases (standard MIB and host resources MIB) for automatic system monitoring by means of network management systems is desired. Check here that this request is fulfilled: **[ ]**

Please describe the facilities for system performance and energy monitoring, error tracking as well as for system performance analysis and tuning in detail.

Answers and annotations by the vendor:

[Insert text here]

2.9.3 System restarts and upgrades

I 75:	The time required for - a complete system reboot - for a reboot and re-integration into to whole system of an island - for the reboot and re-integration into to whole system of a node should be specified.

I 76:	The time required for an operating system upgrade or complete installation of a new operating system on all nodes should be well below 24 hours. Check here that this request is fulfilled: **[ ]**

I 77:	For testing new versions of the OS and the interoperability with customer programs, it must be possible to run a different version of the OS in a well definded and secluded part of the system, e.g. an island. Check here that this request is fulfilled: **[ ]**

Answers and annotations by the vendor:

[Insert text here] Operating system upgrades

2.9.4 Security

Bugs and security issues discovered during operation of the system must be eliminated reliably and within a reasonable response time.

M 39:	The concept for reaction to security incidents (including reaction times, assignment of responsibilities and potentially cooperation with CERT) must be described. Check here that this requirement is fulfilled: **[ ]**

Answers and annotations by the vendor:

[Insert text here]

2.9.5 Authentification

I 78:	Pluggable authentication modules should be supported; specifically, modules for LDAP and Kerberos5 should be provided. Check here that this request is fulfilled: **[ ]**

T 18:	Other means for secure user authentification methods such as SmartCards may be of interest.

If available, please shortly describe such mechanisms and their costs.

Answers and annotations by the vendor:

[Insert text here]

2.10 Software

The offered software stack must be reliable and scalable. Even for jobs which use the whole system, severe bottlenecks and/or resource limitations must not occur.

For every software package mentioned in this section, please give an indication if only third-party manufacturers can provide it.

2.10.1 Message Passing Interface

The MPI implementation must be scalable i.e., even for very large applications, the resource usage (e.g., buffer memory) must not be excessive, start-up times and connection times must be moderate. Process management for start-up and for termination must be reliable and predictable.

M 40:	A scalable MPI implementation must be provided which is capable of efficiently running jobs up to the size of the whole system without excessive resource usage. Check here that this requirement is fulfilled: **[ ]**.

Describe your implementation of MPI in detail, particularly

- scaling behaviour
- resource usage and its scaling behaviour
- implementation and restrictions for Alltoall type of collective communication
- offloading functionality

I 79:	The maximum startup time for a pure MPI application spanning the whole system should be specified. (The startup time is defined as the time between the start of mpiexec to the time after MPI_INIT i.e., it includes the time for the spawning of the MPI processes and the time to establish the connections between the processes.)

Maximum Startup time:	[Insert text here] Minutes

M 41:	The MPI implementation must conform to the MPI 1.3 standard. Check here that this requirement is fulfilled: **[ ]**.

M 42:	At least the following features from the MPI 2.2 standard must be available: - one-sided communication - MPI-IO (optimized for use with the Parallel Storage) - extended collective operations - language bindings for Fortran and C (*C++ is not mandatory*). Check here that this requirement is fulfilled: **[ ]**

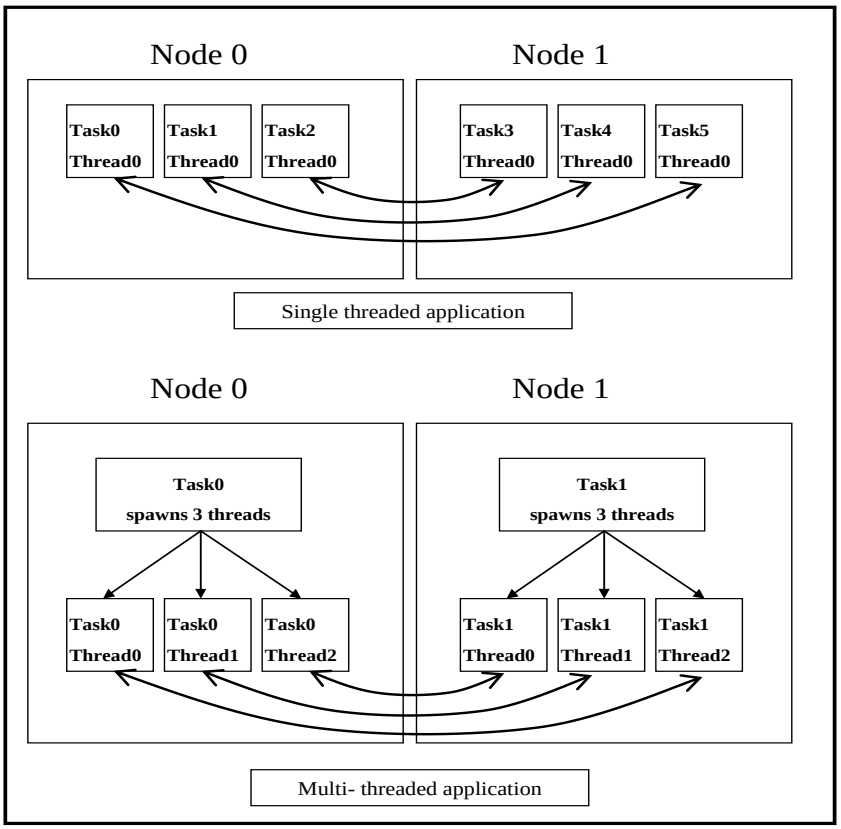
I 80:	The complete MPI 2.2 standard should be supported Check here that this request is fulfilled: **[ ]**

M 43:	It must be possible to use the parallel file system with high efficiency and performance for MPI-IO. Check here that this requirement is fulfilled: **[ ]**

I 81:	It should be possible to use the home file system (see 2.5.1) for MPI-IO.	
	Check here that this request is fulfilled:	**[ ]**

T 19:	It is desired that the scalability features from the upcoming MPI 3.0 standard (e.g., non-blocking collectives) become available at latest one year after the release of this standard.	
	Check here that this request is fulfilled:	**[ ]**

For large-scale parallel programs running on 10,000 or more cores, it is expected that hybrid parallelism will in many cases provide better scalability than single-threaded MPI. To enlarge the set of programs that can profit from such a hybrid programming style, it is essential that the overhead incurred by technical requirements on the implementation of the multi-threaded MPI library is not excessive. The following figure serves to illustrate the issue:



In both situations the same hardware resources are available to the program; the top panel shows how the resources might be used by a single threaded MPI program with 6 tasks. The bottom panel shows a functionally equivalent panel performing multi-threaded execution using 2 MPI tasks with 3 OpenMP threads per task, where MPI calls are performed from within the threaded regions.

M 44:	A thread-safe MPI implementation (MPI_THREAD_MULTIPLE) must be provided.	
	Check here that this requirement is fulfilled:	**[ ]**

I 82:	The MPI performance of multi-threaded programs should not be substantially lower than that of equivalent multi-tasked programs.	
	Check here that this request is fulfilled:	**[ ]**

I 83:	For the MPI implementation, deadlock detection mechanisms should be provided.	
	Check here that this request is fulfilled:	**[ ]**

I 84:	Front ends (e.g., mpif90, mpicc, mpicxx, mpiexec) to the MPI subsystem should be provided for alternative compilers (e.g., PGI, gcc). Check here that this request is fulfilled: **[ ]**

Answers and annotations by the vendor:

[Insert text here]

2.10.2 Other message passing libraries

What other libraries or dialects for message passing are available (e.g., MPICH, OpenMPI, TCGMSG, ARMCI, GASnet) and can be used with the interconnect (not considering TCP/IP)?

Answers and annotations by the vendor:

[Insert text here]

2.10.3 Compilers

M 45:	A least one set of supported, optimizing compilers including a Fortran, a C and a C++ compiler must be provided. Check here that this requirement is fulfilled: **[ ]**

The following requirements refer to this “mainline” set of compilers.

I 85:	The Fortran compiler should support the following language features: 1. A full implementation of the language specifications of Fortran 95 (ISO/IEC 1539-1 : 1997). 2. Extensions to allocatable objects as specified in ISO/IEC 15581 TR : 1998(E) 3. The C interoperability features from the Fortran 2003 standard (ISO/IEC 1539-1 : 2004), as described in section 15 of the standard document (the C compiler must be able to co-operate with the Fortran implementation as a companion processor); I/O processing with STREAM access as described in section 9. 4. Procedure pointers (also as type components) and the IMPORT statement, as described in section 12 of the Fortran 2003 standard; the VALUE attribute as described in section 5 of the Fortran 2003 standard. Check here that this request is fulfilled: **[ ]**

I 86:	The C Compiler should at least support a full implementation of the standard ISO-Standard ISO/IEC 9899:1999 („C99“): Check here that this request is fulfilled: **[ ]**

I 87:	The C++ Compiler should at least support an implementation of the standard ISO/IEC 14882-2003 („C++03“), including the C++ standard library. Check here that this request is fulfilled: **[ ]**

I 88:	The Fortran, C, and C++ Compilers should support the OpenMP standard version 3.0 Check here that this request is fulfilled: **[ ]**

T 20:	Full support for the Fortran 2003 standard (ISO/IEC 1539-1 :2004) is desired. Check here that this request is fulfilled: **[ ]**

T 21:	Support for the following features from the Fortran 2008 (draft) standard is desired: - Submodules and - Coarrays Check here that this request is fulfilled: **[ ]**

2.10.4 Other compilers, libraries and development tools

I 89:	A suitable licensing including the license fees for compilers, libraries, and development tools should be part of the offer and be included in the maintenance costs.[31] Please describe the licensing schemes.

I 90:	If compilers, libraries, and development tools are delivered by a third party, specify what level of support is available for each delivered package and whether LRZ can be provided with direct access to the support structure of the software provider.

I 91:	Specify the vendor's scope of responsibility and warranty for the programming environment, particularly compilers, libraries, and performance analysis tools.

I 92:	Emerging compilers based on PGAS concepts should be available, e.g. Check individually for available implementations: - CAF (Coarray Fortran) **[ ]** - UPC (Unified Parallel C) **[ ]**

Answers and annotations by the vendor:

[Insert text here]

2.10.5 Programming environment

M 46:	The programming environment and its tools for performance measurement and analysis must be described.

T 22:	An integrated programming environment that allows to develop, start, debug and optimize parallel programs and to visualize the performance data is desired (e.g., via integration into Eclipse) Check here that this request is fulfilled: **[ ]**

[31] LRZ already has campus licences for some compilers, libraries and tools which may be used or extended

T 23:	It is desired, that the MPI implementation as well as the tools for performance analysis support the STF and /or OTF trace file formats. Check here that this request is fulfilled: **[ ]**

I 93:	For monitoring message passing programs, Vampir NG/Vampirtrace and/or the Intel Cluster Tools should be available. Check here that this request is fulfilled: **[ ]**

M 47:	Access to the hardware performance counters from user space must be available and be supported. Check here that this requirement is fulfilled: **[ ]**

I 94:	The operating system should provide support for the performance application programming interface (PAPI, see http://icl.cs.utk.edu/papi). Check here that this request is fulfilled: **[ ]**

T 24:	Means for automatic calling of user-specified functions at subroutine entry and exit (user hook functions for inserting of user specified performance libraries) are desired. Check here that this request is fulfilled: **[ ]**

Answers and annotations by the vendor:

[Insert text here]

2.10.6 Debuggers

M 48:	A scalable debugger with an GUI must be available. Check here that this requirement is fulfilled: **[ ]**

I 95:	The Totalview or Allinea DDT debugger should be provided. The number of MPI tasks supported by the debugger and the license should be at least 1024. The license and maintenance costs should be included in the offer. Check here that this request is fulfilled: **[ ]**

Describe the test aids, debuggers, etc. which are available and their features.

Answers and annotations by the vendor:

[Insert text here]

2.10.7 Libraries

M 49:	At least the following highly efficient and optimized libraries must be available: - a scientific library including FFT routines, random number generators, linear algebra etc. - optimized BLAS and LAPACK (serial and shared memory parallel version) - ScaLAPACK, BLACS - PETsc - FFTW - Global Arrays Check here if these requirements is fulfilled: **[ ]**

I 96:	NAG Fortran Libraries (serial, parallel, and shared memory parallel versions) should be available.
	Check here if these request is fulfilled: **[ ]**

Different applications might use the I/O system in different ways. For example, if I/O requests are large, library buffering is unnecessary. On the other hand, if all I/O requests are very small, an increased library buffer might improve performance. Specific hints to the MPI-IO interface are also useful to improve performance.

T 25:	Special libraries for tuning and performance enhancement of I/O are desired.

If such solutions are available, please describe their capabilities:

Answers and annotations by the vendor:

[Insert text here]

2.10.8 Grid Software

I 97:	UNICORE and Globus should be available. Check here that this requirement is fulfilled: **[ ]**

2.11 Training, user support and technical support

2.11.1 Documentation

M 50:	The documentation for the system and the software products must be provided. Check here that this requirement is fulfilled: **[ ]**

I 98:	Documentation for the end user should be electronically accessible[32]. Check here that this request is fulfilled: **[ ]**

2.11.2 Introduction to the system

I 99:	Training for the LRZ system administrators as well as an introduction for selected users and LRZ staff to usage of the system should be provided. Check here that this request is fulfilled: **[ ]**

2.11.3 System-related support

Typical activities of the system support personnel include, but are not limited to:

- Installation and maintenance of all system software components provided by the vendor
- Troubleshooting of system incidents
- Assistance in all system administration related topics
  - Assistance in configuration and policy definition and enforcement of the batch queuing system
  - OS and system performance tuning
  - Installation and performance tuning of file systems
  - Configuration and tuning of system management software
- Training of LRZ system administrators

M 51:	For the duration of operation of the system, the vendor must provide **on-site** personnel responsible for system-related support, critical failure analysis and maintenance management of the system.. **On-site** availability of at least one skilled person from 8 am till 6 pm during normal working days[33]. **On-call** availability of at least one skilled person from 8 am till 6 pm during weekends and official Bavarian legal holidays[33]. Check here if the requirements are fulfilled: **[ ]**

M 52:	During the first year of operation[34], the vendor must provide **additional on-site** personnel which, together with LRZ staff, are in charge of configuring and putting the system into user operation. These personnel must have excellent knowledge of the system and must maintain the necessary contacts to the vendor. Check here if the requirements are fulfilled: **[ ]**

T 26:	Please describe additional possibilities you could offer to LRZ with respect to system administration and tuning.

Office space for vendor personnel will be provided by LRZ.

Answers and annotations by the vendor:

[Insert text here]

[32] Access restrictions may be applied.

[33] An intermission interval for the vendor personnel of up to 2 hours is permitted within the 8 am to 6 pm time window.

[34] Starting after acceptance of Phase1

2.11.4 User support

A supercomputer requires staff-intensive user support; LRZ here depends on the manufacturer's assistance.

Typical activities of the support personnel include, but are not limited to:

- Handling of user related incidents, which cannot be dealt with by LRZ support staff
- Optimization of user codes in case of severe performance problems
- Assistance in optimization and porting of third party software, libraries and tools
- Assistance on definition of complex workflows, including data handling, grid usage and and visualisation, pre- and postprocessing of data
- Training of users on the vendor specific topics like batchsystem or tools

M 53:	For the duration of operation of the system, the vendor must provide software engineering personnel (at least one full time equivalent) with deep knowledge about the system architecture and programming environment that assists with optimization-related problems of users. This vendor support may be realized partly off-site. Check here if the requirements is fulfilled: **[ ]**

T 27:	It is desired that the vendor provide additional support for obtaining optimal performance and scalability as well as porting and optimisation of specific software packages. Please describe the extent of such support.

This personnel need not be on-site at all time, but must be readily schedulable.

Answers and annotations by the vendor:

[Insert text here]

2.12 Maintenance

M 54:	An 365 x 10 hours on-site **hardware** maintenance contract fulfilling the following conditions must be offered: During peak operation times (from 8 am till 6 pm), the availability of technicians and parts on-site within 4 hours after problem report must be guaranteed for critical hardware errors (errors which have a high impact on the SuperMUC service quality). Check here if the requirements are fulfilled: **[ ]**

M 55:	A 365 x 10 hours **software maintenance contract** fulfilling to the following conditions must be offered: Response to a critical software error must happen within four hours. Check here if the requirements are fulfilled: **[ ]**

M 56:	The total cost for hardware and software maintenance for a 5 year term of operation must be specified in the following table:

Total cost for HW maintenance:	_____ €
Total cost for SW maintenance:	_____ €

Further details about guaranteed repair times and possible delivery times for spare parts should be specified. These details will be regulated in the maintenance contract.

[Insert text here]

2.13 References

The tenderer should provide a list of installed and projected systems of similar type and/or size.

Answers and annotations by the vendor:

[Insert text here]

2.14 Collaboration with the vendor

T 28:	A close collaboration with the vendor is desired.

Please describe the scope, extent and organizational structure of such a collaboration.

Answers and annotations by the vendor:

[Insert text here]

2.15 Migration System

M 57: A migration system must be installed until end of July 2011, if the installation and acceptance of Phase 1 cannot be completed until the end of 2011.

Check here if the requirement is fulfilled: **[ ]**

The migration system might consist of an early delivery of the fat node island, or parts of the migration system may be reused for the fat node island of SuperMUC.

The filesystem for the migration system might be an early delivery of the proposed NAS, or parts of filesystem may be reused for SuperMUC (e.g., for the replication system).

The size and characteristics of the migration system must be such that it can handle the workload of the current HLRB II by the end of the third quarter of 2011.

M 58: The migration system must have the following characteristics:

- at least 60 Tflop/s peak performance
- at least 8000 cores
- at least 4 Gbyte of main memory per core
- at least 32 cores per shared memory node
- at least QDR Infiniband connection between the nodes
- at least 300 (usable) TByte of disk space
- at least 2 GByte/s aggregate bandwidth for writing to the disks
- at least 2 GByte/s aggregate bandwidth for reading from the disks
- versions of the operating system, batch system, programming environment and tools which are proposed for SuperMUC
- 2x10 GB/s Ethernet connection
- appropriate maintenance and support

Check here if the requirements are fulfilled: **[ ]**

M 59: A description of the compute node, of the processor architecture, of the required power and cooling infrastructure and of the internal network must be provided.

The appropriate fields in Table 3.6 must be filled in.

Answers and annotations by the vendor:

[Insert text here]

3 Tables of Key System Parameters

3.1 Environment

Item and Unit	Phase 1
Footprint of the system incl. maintenance areas: length x width (m x m)	
Number of racks	
Dimension of system racks: width x depth x height (m x m x m)	
Total Weight (kg)	
Specific weight per square meter (kg/m^2)	
Point Load (N)	
Ambient temperature (range) (°C)	
Relative humidity (range) (%)	
Air purity of computer room (µg/m^3)	
Cooling type of compute nodes (liquid or air); if applicable, the type of liquid must be specified	
Cooling type of disk storage (liguid or air); if applicable, the type of liquid must be specified	
Maximum tolerable supply temperature of cooling Loop 2 (°C)	
Inlet temperature of compute node coolant (range) (°C)	
Outlet temperature of compute node coolant (range) (°C)	
Inlet temperature of disk storage coolant (range) (°C)	
Outlet temperature of disk storage coolant (range) (°C)	
Expected electrical power consumption of total system (kW)	
Peak[35] electrical power consumption of total system (kW)	
Expected electrical power consumption of thin nodes (kW)	
Peak[36] electrical power consumption of thin nodes (kW)	
Expected electrical power consumption of fat nodes (kW)	
Peak[37] electrical power consumption of fat nodes (kW)	
Expected electrical power consumption of disk storage (kW)	

[35] The peak electrical power consumption of the whole system is defined as the sum of the following power consumption values:

- Power consumption of the system when running the Linpack benchmark at maximum processor frequency on all nodes
- maximum power which can be consumed in theory by all storage and network components of the system.

[36] The peak electrical power consumption of thin nodes is defined as the sum of the power consumed by all thin nodes when running the Linpack benchmark at maximum processor frequency on all corresponding nodes.

[37] The peak electrical power consumption of fat nodes is defined as the sum of the power consumed by all fat nodes when running the Linpack benchmark at maximum processor frequency on all corresponding nodes.

Peak electrical power consumption of disk storage (kW)	
Expected electrical power consumption of communication network (kW)	
Peak electrical power consumption of communication network (kW)	
Expected heat emission of total system into air (kW)	
Expected heat emission of total system into water Loop1[38] (kW)	
Expected heat emission of total system into water Loop2 (kW)	
Peak heat emission of total system into air (kW)	
Peak heat emission of total system into water Loop1[38] (kW)	
Peak heat emission of total system into water Loop2 (kW)	
Expected heat emission of disk storage into air (kW)	
Expected heat emission of disk storage into water Loop1[38] (kW)	
Peak heat emission of disk storage into air (kW)	
Peak heat emission of disk storage into water Loop1[38] (kW)	
Expected heat emission of network components into air (kW)	
Expected heat emission of network components into water Loop1[38] (kW)	
Peak heat emission of network components into air (kW)	
Peak heat emission of network components into water Loop1[38] (kW)	
Voltage (V)	
Number of electrical phases	
Frequency of electrical current (Hz)	

3.2 Compute nodes

Item and Unit	Phase 1 Thin Nodes	Phase 1 Fat Nodes	Phase 1 Thin+Fat Nodes
Quantity and Performance			
If applicable: Number of islands			
Number of nodes			
Number of cores			
Type of processor chips			
Total peak floating point performance (sum of all cores, only multiply/add, no div or sqrt) (PFlop/s)			
Total peak number of instructions per second (PInst/s)			
Number of processors per node			
Number of cores per node			
Frequency of a core (GHz)			

[38] Please only specify the heat which will be emitted via indirect water cooling of in this field (e.g., do not add the heat which will be emitted into air)

Peak floating point performance of one core (only multiply/add, no div or sqrt) (GFlop/s)			
Peak number of instructions per second of one core (GInst/s)			
Memory and cache			
Total size of memory (TByte)			
Size of memory for one core (GByte/core)			
Size of memory of one node (GByte/node)			
Maximum possible size of memory of one node(GByte/node)			
Total Bandwidth to local memory of one node (GByte/s)			
(Share of theoretical) Bandwidth to local memory of one core (GByte/s)			
Interleaving of memory access			
If applicable: Size of L1 data cache (kByte)			
If applicable: Size of L1 instruction cache (kByte)			
If applicable: Size of L2 data cache (MByte)			
If applicable: Size of L3 data cache (MByte)			
If applicable: Bandwith of L1 data cache (GByte/s/core)			
If applicable: Bandwith of L2 data cache (GByte/s/core)			
If applicable: Bandwith of L3 data cache (GByte/s/core)			
If applicable: Cache line size of L1 data cache (Byte)			
If applicable: Cache line size of L2 data cache (Byte)			
If applicable: Cache line size of L3 data cache (Byte)			
If shared by more than one core: Size of L2 data cache per core (MByte)			
If shared by more than one core: Size of L3 data cache per core (MByte)			
Latency to local memory (Cycles)			
If applicable: Latency to memory on remote node(s) (Cycles)			
If applicable: Number of SSE/AVX registers (per processor core)			
If applicable: Length of SSE/AVX registers (per processor core)			
Number of floating point registers (per processor core)			

3.3 System interconnect

Item and Unit	Phase 1
Compute node interconnect network (Type)	
Network topology	
Number of (bidirectional) links per node*	
Theoretical bandwidth of one link (bidirectional) (GByte/s)*	

Theoretical bandwidth of one node (bidirectional) (GByte/s)*	
Theoretical bisection bandwith of the entire system (GByte/s)	
Maximum time for a broadcast (1 Byte length) to reach all nodes (µs)	
Maximum MPI send-receive (1 Byte length) latency between two arbitrary communication partners (blades or nodes) (µs)	
Minimum MPI send-receive (1 Byte length) latency between two arbitrary communication partners (blades or nodes) (µs)	

*If thin and fat nodes differ, provide both numbers.

3.4 Connection of the system to external networks

Item and Unit	Phase 1
Number and type of interfaces to the external network	
Aggregate bandwidth to the external network (GByte/s)	

3.5 Disk storage

Item and Unit	Phase 1
Number of I/O nodes	
Size of SAN/DAS user storage (TByte)	
Size of NAS user storage (TByte)	
Aggregate theoretical bandwidth to/from SAN/DAS storage (GByte/s)	
Aggregate theoretical bandwidth to/from NAS storage (GByte/s)	

Migration System

Item and Unit	Migration System
Footprint of the system incl. maintenance areas: length x width (m x m)	
Dimension of system racks: width x depth x height (m x m x m)	
Total Weight (kg)	
Specific weight per square meter (kg/m^2)	
Point Load (N)	
Ambient temperature (range) (°C)	
Relative humidity (range) (%)	
Air purity of computer room (µg/m^3)	
Peak[39] electrical power consumption of the of the migration system (kW)	
Expected electrical power consumption of the of the migration system (kW)	
Voltage (V)	

[39] The peak electrical power consumption of thin nodes is defined as the sum of the power consumed by all thin nodes when running the Linpack benchmark at maximum processor frequency on all corresponding nodes.

Number of electrical phases	
Frequency of electrical current (Hz)	
Number of nodes	
Number of cores	
Type of processor chips	
Number of cores per node	
Frequency of a core (GHz)	
Peak floating point performance of migration system (Tflop/s)	
Peak floating point performance of one core (only multiply/add, no div or sqrt) (GFlop/s)	
Total size of memory (TByte)	
Size of memory for one core (GByte/core)	
Size of memory of one node (GByte/node)	
Maximum possible size of memory of one node(GByte/node)	
Total Bandwidth to local memory of one node (GByte/s)	
(Share of theoretical) Bandwidth to local memory of one core (GByte/s)	
Compute node interconnect network (Type)	
Network topology	
Number of (bidirectional) links per node*	
Theoretical bandwidth of one link (bidirectional) (GByte/s)*	

4 Risks and Mitigations

4.1 Assessment of the risks

Please comment on the following risks, assess their probability and severity and describe mitigation measures. Such measures may include proactive steps, taken before the installation starts. Although some of the risks are within the responsibility of LRZ, vendors are asked for their assessment and suggestions for possible countermeasures or mitigations.

M 60: For each identified risk, two parameters have to be assessed:
severity (severity of consequences)
probability (likelihood of occurrence)

I 100: Potential mitigation measures should be described.

T 29: The vendor is encouraged to describe additional risks.

For severity and probability, we have chosen a scale with 4 levels shown in the Table below. This is a good practice since there is no "medium" value which prevents choosing a too neutral or average value too often.

Scale	1	2	3	4
Severity	very low (almost no impact)	noticeable impact	strong impact	major impact (may lead to project failure)
Probability	unlikely	possible	high	almost certain

4.2 Procurement related risks

Protests or objections by competitors may delay the procurement process. In the worst case it must be completely restarted.

Severity: [Insert text here]
Probability: [Insert text here]
Mitigation: [Insert text here]

4.3 Infrastructure risks

The infrastructure of the computing centre may not be ready. Technical installations for power supply and cooling may be insufficient or may not obey the required specifications. The electrical power supply may turn out to be too unstable. Specifications provided by LRZ prior to the installation may be incorrect

Severity: [Insert text here]
Probability: [Insert text here]
Mitigation: [Insert text here]

4.4 Risks that May Prevent Systems from Becoming Operational at All

The system may not pass the acceptance tests because it fails to satisfy the contracted commitments for performance, functionality, reliability, availability, usability etc.

Severity: [Insert text here]
Probability: [Insert text here]
Mitigation: [Insert text here]

The vendor or one of its main suppliers may cease operation e.g., due to bankruptcy. In this case it may not be possible to build, deliver or operate a projected system or parts of it.

Severity: [Insert text here]
Probability: [Insert text here]
Mitigation: [Insert text here]

4.5 Risks that May Delay System Operation

Technology road maps may be considerably delayed, vendors or one of their suppliers cannot adhere to contracted time schedules e.g., because technical problems delay the availability of components or their interoperability. It may take considerable time to fix these technical problems.

Severity: [Insert text here]
Probability: [Insert text here]
Mitigation: [Insert text here]

The vendor or one of its main suppliers may get into a **financial instability**. This may cause delays in the production process. It may also reduce the availability of human staff on the vendor's or its supplier's side..

Severity: [Insert text here]
Probability: [Insert text here]
Mitigation: [Insert text here]

4.6 Risks that May Limit the Usability of Systems

Contracted technical properties not available. The vendor or one of its suppliers cannot adhere to contracted technical properties. As can be seen from known issues with global parallel file systems this risk includes (system) software. It may take considerable time to fix these technical problems.

Severity: [Insert text here]
Probability: [Insert text here]
Mitigation: [Insert text here]

Crucial components failures. This includes failures of cooling components, water leakages etc.

Severity: [Insert text here]
Probability: [Insert text here]
Mitigation: [Insert text here]

Operational limits (power, cooling, temperature) may be exceeded, e.g., on very hot summer days.

Severity: [Insert text here]
Probability: [Insert text here]

Mitigation: [Insert text here]

Software scalability limits: Software (e.g. OS, batch system, MPI, parallel file sytem etc.) may be not mature enough to scale with the size of the system, the number of jobs, etc.

Severity: [Insert text here]
Probability: [Insert text here]
Mitigation: [Insert text here]

Although a system passed the acceptance test it may **not be reliable** enough for long program runs. Possible causes may be insufficient node reliability, insufficient reliability of the communication network or insufficient reliability of the I/O (most likely for shared global file systems).

Severity: [Insert text here]
Probability: [Insert text here]
Mitigation: [Insert text here]

4.7 Problems with Usage of Systems

Applications do not run Examples are programs that do not compile/link and programs that produce wrong results. The main causes are compiler errors, library errors, CPU errors and errors in the communication hardware (including firmware) and communication software.

Applications suffer from unexpectedly bad performance Bad performance may be related to the code generation of a compiler, to a suboptimal implementation of library routines or missing optimized library routines for special purposes

Tools for performance analysis and debugging turn out to be too unstable or immature for working on large-scale applications

Severity: [Insert text here]
Probability: [Insert text here]
Mitigation: [Insert text here]

4.8 Risks with procurement and installation of Phase 2[40]

Funding for Phase 2 may be delayed or totally cancelled.

Severity: [Insert text here]
Probability: [Insert text here]
Mitigation: [Insert text here]

Technology road maps may be considerably delayed, vendors or one of their suppliers cannot adhere to contracted time schedules.

Severity: [Insert text here]
Probability: [Insert text here]
Mitigation: [Insert text here]

[40] For commitments relating to Phase 2 see „Anschreiben“

4.9 Financial and fiscal risks

Tax iincrease**s** (VAT). higher **energy consumption** than expected **or exchange rate drops** may make the veondor's calculations invalid.

Severity: [Insert text here]
Probability: [Insert text here]
Mitigation: [Insert text here]

4.10 Other Risks

Please describe additional risks which appear significant, assess these risks, where applicable and describe potential countermeasures.

Risk Description: [Insert text here]
Severity: [Insert text here]
Probability: [Insert text here]
Mitigation: [Insert text here]

5 Summary of Mandatory Requirements

Please check that all mandatory information is included in your proposal: