Leibniz-Rechenzentrum
der Bayerischen Akademie der Wissenschaften

LRZ's position in European supercomputing

The Leibniz Supercomputing Centre (Leibniz-Rechenzentrum, LRZ) is part of the Bavarian Academy of Sciences and Humanities (Bayerische Akademie der Wissenschaften, BADW). LRZ is member of the Gauss Centre for Supercomputing (GCS), the alliance of the three national supercomputing centres in Germany (JSC-Jülich, HLRS Stuttgart, LRZ-Garching). It is a central site for large scale data archiving and backup, networking and general IT services for universities and research institutions. LRZ has been an active player in the area of high performance computing for over 25 years and provides computing power on several different levels.

Since 2012, LRZ operates the supercomputer "SuperMUC" as one of the GCS Tier-0 systems open to PRACE users with a peak performance of 3.2 Petaflop/s and more than 320 Terabytes of main memory. SuperMUC will be upgraded to a peak performance of 6.4 Petaflops, in the first half of 2015.

LRZ demands for future system architectures and system software

Highly scalable general purpose system architecture

The job profile on LRZ's current supercomputer SuperMUC has the following characteristics:

- Execution of a large number of different projects respectively different applications (more than 170).
- So far, about 10 scientific applications are currently able to use all 150k compute cores of the system in one program instance efficiently.
- Most applications need more than 1.0 GBytes of main memory per compute core.
- Data centric applications with new demands particularly with regard to I/O and long term online storage of data as well as main memory (e.g, very huge non-volatile memory) are steadily growing.

For the broad range of applications which need to be executed on LRZ's supercomputers, it is most important that the future HPC system architectures provide balanced hardware and software characteristics. Since future supercomputer will be composed of several hundred thousand processing cores, a highly scalable system software stack and programming environment will be of major importance for the efficient use of the system.

It is the explicit goal of the LRZ to provide supercomputing services without imposing additional constraints on its current and future user base. Hence, special-purpose computers that offer particular advantages for only a small subset of scalable programs are no suited HPC architecture for our broad HPC user and application base.

Following this, our primary goal for the procurement of future computers in the 100 to 1000 Petascale performance area 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.

In this context, the theoretical peak performance of a computer becomes more and more a subordinate aspect. Instead, the expected performance and energy to solution for the entire spectrum of applications of our current and future scientific projects will also be in future our dominating selection criterion.

Energy efficiency

Recent trends in HPC reveal that energy efficiency of supercomputers is a challenge more than ever. The power consumption of supercomputers has reached a level of more than 10 Megawatts, yet it continues to grow. Hence, increasing the energy efficiency of supercomputers is today one of the main goals of the HPC community. This is also reflected by the popular Green500 list, which lists the 500 most energy efficient supercomputers worldwide in relation to executing the Linpack benchmark on the system.

It is widely known that the Linpack is rather unlikely to be a good measure for true application performance since it mainly stresses processor performance, while obtaining high real world application performances generally also require high memory, network and I/O performance – in other words, a good system balance. Hence, the Green500 list is definitely no good measure for HPC system energy efficiency for real world applications. LRZ has therefore developed the 4-pillar framework for energy-efficient HPC. It consists of the pillars "Energy Efficient Compute Centre Infrastructures", "Energy Efficient HPC System Hardware", "Energy Efficient HPC System Software" and "Energy Efficent HPC Applications".

LRZ uses an integrated approach to maximize the energy efficiency of supercomputers. We therefore perform R&D work on all 4 pillars with a focus on pillars one, three and four. Using this holistic approach, important requirements for future generations of supercomputers are as follows:

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 b. München
Telefon: +49 89 35831 - 8000
Telefax: +49 89 35831 - 9700
USt-IdNr.: DE811335517
E-Mail: lrzpost@lrz.de
Internet: www.lrz.de

- Energy-aware system software and performance tools supporting the
  - optimal placement of application threads on the network as well as the
  - automatic minimization of energy to solution for large real world applications within a constant energy delay product through the use of energy-aware system software.
- Highly optimized numerical libraries
- Enabling chiller-free system cooling in all European climate zones and the re-use of system waste heat
  - Use of direct liquid technologies for all HPC system components
  - Use of CMOS components able to support the ASHRAE W5 inlet temperature spectrum (40°C – 55°C)
- Implementation of a central data base containing important environmental sensors such as outside whet bulb temperature and room temperatures together with site infrastructure and system energy counters as well as important system performance counters
- Standardization of system energy sensors and all corresponding APIs
  - Use of calibrated sensors
  - Clear declaration of sensor accuracy

Planned LRZ R&D activities in areas mentioned above will be the enhancement of existing LRZ tools such as the system performance monitoring framework PerSyst and the Power Data Aggregation Monitor (PowerDAM) as well as the development of an energy-aware scheduling plugin for the slurm batch scheduler.

Supercomputing issues to be solved

Many issues mentioned by the DARPA Exascale computing study can already be observed in today's supercomputers. Above all, deficiencies in the fault tolerance and resiliency of important system software such as network routing algorithms, parallel file systems and run time systems are major reasons for system outages or application aborts. Hence, further R&D efforts in these areas will be needed in order to facilitate the efficient use of future generations of supercomputers.
Future supercomputers will house hundreds of thousands of processor cores, at least partly of heterogeneous nature. The present programming paradigms and the majority of the widely used algorithms presently employed will not be sufficient to efficiently use the computing resources available in future systems. Also numerical libraries, parallel run time systems, performance tools, system management and monitoring tools, and operating systems as presently available are not up to the demands that the future systems will pose. Hence research in the following areas will be needed to overcome this situation:

- New numerical algorithms that can take sufficient advantage of the available computing and communication resources.
- Efficient parallel programming paradigms, novel programming languages, compilers, numerical libraries and parallel debuggers
- Special languages and environments for hybrid processor architectures
- Parallel file systems
  - Metadata performance
  - Scalability and reliability
  - Data protection and recovery mechanisms
  - Mechanisms for end-to-end data integrity
  - Distributed RAID mechanisms
  - Analysis and diagnostics database
  - Data mining and visualization tools
- Highly scalable performance tools
- Highly scalable system management and monitoring tools
  - Energy- and network topology-aware batch schedulers
  - Monitoring and correlation of system events
  - Databases for root cause analysis and diagnostics

Current applications run at efficiencies of 5 to 10% in many cases. Hence, 90% or more of the energy that is needed to operate current and future systems will be wasted if we do not address this challenge. This is not new insight. Already

with the current HPC systems, legacy applications struggle with achieving decent efficiencies. In most cases, complete application re-writes or new algorithmic developments will be necessary to fully leverage the compute power of future supercomputers. Hence Many-Peta- to Exascaling of applications requires a close cooperation to be established with the numerical mathematics community as well as representatives from main scientific domains and code developers at research sites as well as ISV companies.

For this reason, LRZ has started the partnership initiative computation sciences (πCS) in which experts of different science domains such as astrophysics, geophysics, life and environmental sciences cooperate closely with LRZ computational sciences experts to improve the efficiency of selected application codes.