Kriterien für ein PetaFlop System

Rainer Keller, HLRS
Context: Organizational

- HLRS is one of the three national supercomputing centers in Germany.
- The national supercomputing centers are working together in the Gauss Centre for Supercomputing GCS.
- GCS is the means to contribute to the Partnership for Advanced Computing in Europe (PRACE).
- All centers work within PRACE towards a European HPC Infrastructure and perform research with all PRACE partners towards Exascale computing.
- Additionally HLRS is responsible within PRACE and GCS for the support of the engineering community and the definition of the industrial offer.
Context: Systems in Stuttgart

As a tradition, large variety of HPC systems offered:

Cray XE6:
Phase 1, Step0

Clusters, e.g.:
• Laki 62 TF
• BW-Grid 14 TF

New architectures, e.g.:
• Cluster of Cell (2008)
• MD Grape (2007)
The petaGCS project is a BMBF funded project covering the national share for investment and operation of national supercomputing in Germany:

- Covers currently Phase1 of all GCS centers
- Next phases will be covered in a similar way
- 50% co-funding is provided by the regional governments

For HLRS: the Ministry of Science, Research and the Arts Baden-Württemberg
Context: Main User’s Research Projects

- Aeroacoustics
- Aerodynamics
- Astrophysics
- Bioinformatics
- Combustion
- Fluid-Structure Interaction
- Helicopter Aerodynamics
- Meterology
- Medical Imaging
- Nanotechnology
- Solid State Physics
- Turbo Machinery
- Turbulence Phenomena
Constraints: Diversity User Community

- Test users/Beginners
- Early Developers
- Hundreds of TeraFlop/s
- Requires wide spread training & teaching expertise
- Federal project users
- Hundreds of TeraFlop/s
- PetaFlop/s high-end capability
- Advanced users/ Heroes
- Demands for strong collaboration in particular with the vendor
- Courses: MPI & OpenMP
  Fortran for Scientific Computing
  Coarray Fortran & UPC
  GPU Programming
  Single CPU Optimization
  CFD
- Scalability Workshops
  Tools Workshops
  Co-Development
  …
- Constraints: Diversity User Community
- Early Developers
- Test users/Beginners
- Federal project users
- Advanced users/ Heroes

:: 20.5.2011 :: Kriterien für ein Petaflop System :: 6
Sustained Application Performance

- How many applications?
- How to balance between them?
- How to avoid an unfair testcase?
- Real applications or kernels?
- Memory Bound
- CPU Bound
- Communication intensive
- Disk IO intensive
- Grand Challenge
- Open Source
- Petascale Potential
- Important Application Domain
- Real Testcases

:: 20.5.2011 ::

Kriterien für ein Petaflop System :: 7
Tocal cost of ownership

- Less Power = Less Compute Nodes?
- How many people?
- Max power is theoretical max or HPL load or?
- MTBF versus Time to Repair
- Average Power consumption
- Early delivery \(\rightarrow\) Higher Power costs

TCO

```
20.5.2011
Kriterien für ein Petaflop System
```
Procurement Constraints: Summary

• Application performance in engineering domain is more important than peak performance.
• Programmability of the system (Software Stack/libraries, Support of diverse programming models, Available Tools, Adequate memory size, ratio and speed, …) is key!
• Performance of I/O to pre-/postprocess and visualisation is essential
• System architecture should be designed for highly scalable codes with low latency, low communication memory footprint and with high bandwidth interconnect
• Heterogeneous Customer Groups (Industry, Local, Regional, EU Wide)
• Future system must fit into the existing machine room within the planned update to 5MW Power & Cooling capacity
• Total Cost of Ownership is relevant and part of the selection criteria
• New System must support migration path for users of existing systems
  – NEC SX-9: How to attract users of Vector Systems?
  – NEC Nehalem: How to convince people to move to the new system?
  – Vendors: Provider Collaboration needs special attention
Phase 1 Step 1: Hermit 1
3rd PRACE TIER-0 System
Hermit: Infrastructure bring-up

- August 2010
- November 2010
- January 2011
- February 2011

- Finished: April 2011
Phase 1 Step 0: Configuration

- Configuration of **Test** System:
  - Cray XE6 (1 cabinet)
    - 84 compute nodes
    - AMD Magny Cours
      - 8 core/socket, 2.0 GHz
    - 32 GB memory
    - (1344 total cores)
  - 12 service nodes
  - AMD Istanbul
    - 6 core/socket, 2.2 GHz
  - 16 GB memory
  - (72 total cores)
Phase 1 Step 1: Overview

- **Configuration:**
  - Peak Performance ~1PF
  - Overall >3500 nodes
  - Each node has 2 sockets
    - AMD Interlagos @ 2.3GHz,
      - 16 Cores each → ~112k cores
  - Nodes with 32GB and 64GB memory
  - 2.7PB storage capacity @ ~150GB/s IO bandwidth
  - External Access, Pre- & Postprocessing and Remote Visualization Nodes
  - ~2MW maximal power consumption

- **Essential:** intensive support by on-site staff and a collaboration agreement between HLRS and CRAY
- **Support for ISV Codes under CLE („native“) or Cluster Compatibility Mode (CCM).**
- **Part of the acceptance tests are validation of sustained application performance spanning across the full system**
Kriterien für ein Petaflop System

:: 20.5.2011 ::

:: 14 ::
Visualization Servers allow remote graphical access
Full Hardware accelerated graphics through dedicated remote graphics hardware as well as VirtualGL

Powerful Graphic Cards with GP-GPU post processing support beyond the capabilities of a typical end-user system

Visualization of very large datasets without the need to move data outside the data center

Direct link to parallel Visualization on the compute nodes

Access to applications without the need for a local deployment
Phase 1 Step1: Storage Solution

• Lustre based solution for fast disk space
  – 2.7 PB Lustre Workspace capacity
  – 16 DDN SFA10k controllers
  – Integrated into the overall HLRS environments

• HLRS wide Home Space (60TB)

• Local and fast storage capacity with 20TB for Pre- and Postprocessing servers

• Integrated with the existing HPSS system using a data mover concept
A glimpse on Phase 1 Step 2 – Hermit 2

- Step will run in parallel to Step 1 realising an integrated system
- Goal: maintain similar software stack for both installation steps
- Expected architectural changes:
  - Aries interconnect
  - Newest generation of CPUs
  - Partially relying on accelerators
  - Updated storage infrastructure
  - Additional external servers
- Significantly increased sustained application performance
- Scheduled for Autumn 2013
- Overall peak performance of complete Phase 1 will be >5PF
- Specification is driven by agreed sustained application performance and **not** peak performance tough
Thank You very much!

Any Questions?