

HPC Infrastructure Evolution@HLRS

Stefan Wesner, Managing Director

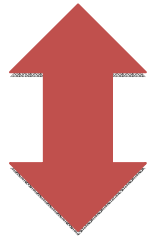


.....

OVERALL CONTEXT

Context: Organizational

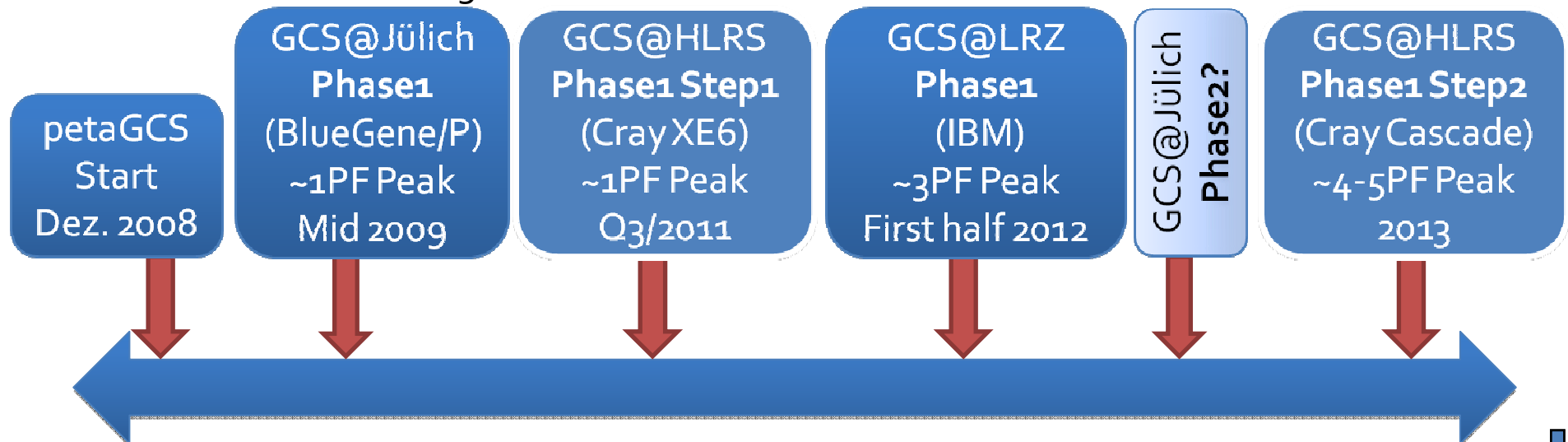
- HLRS is one of the three national supercomputing centers in Germany
- The national supercomputing centers are working together in the Gauss Centre for Supercomputing **GCS**
- GCS is the mean to contribute to the Partnership for Advanced Computing in Europe (PRACE)
- All centers work within PRACE towards 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



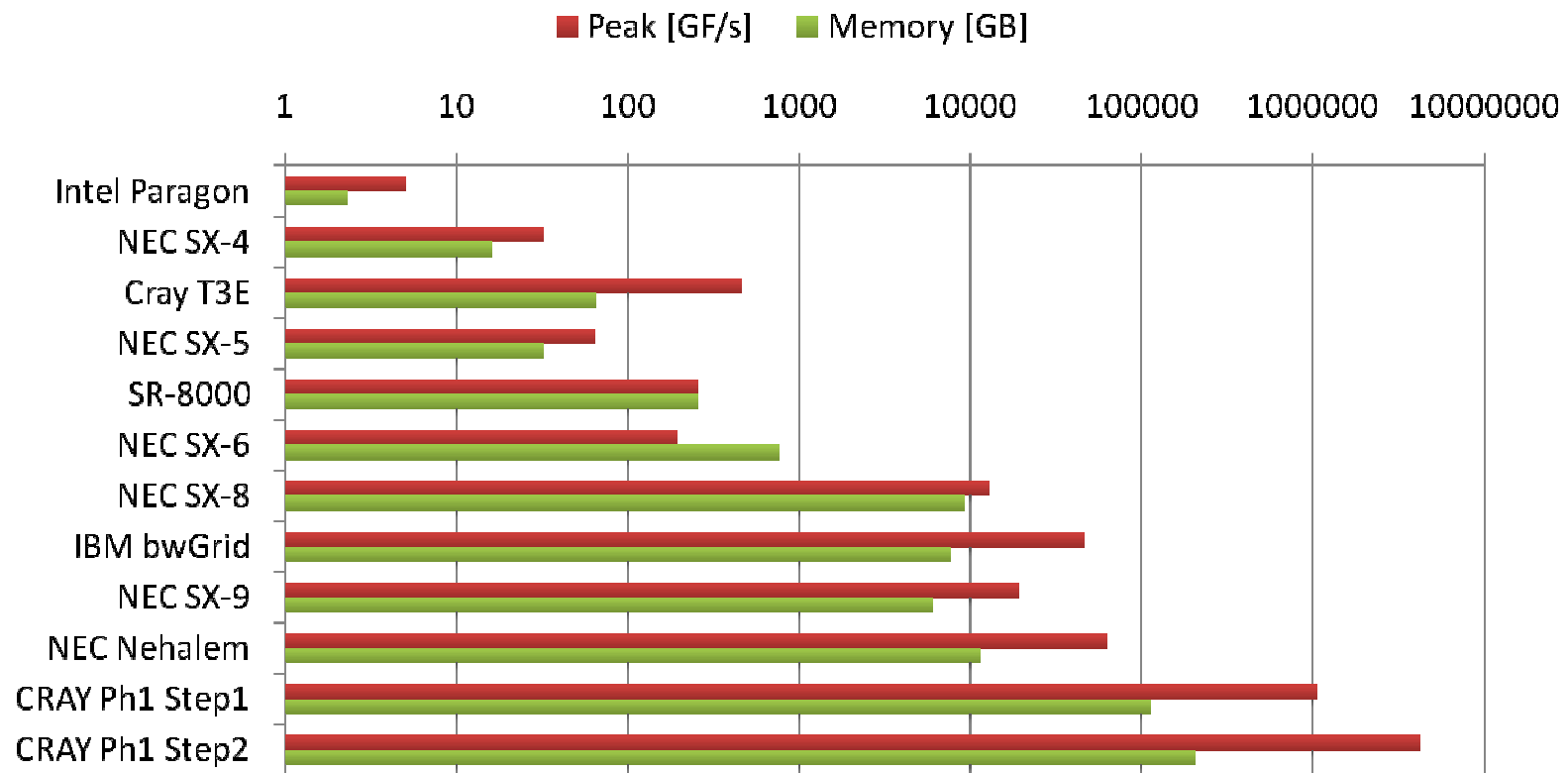
PRACE

Context: The petaGCS Project (Phase1)

- 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 this is the Ministry of Science, Research and the Arts Baden-Württemberg



Context: History




Phase2

.....

PROCUREMENT PROCESS

Procurement: Constraints

- Performance of typical applications in the engineering domain is more important than peak performance
- Programmability of the system (Software Stack, Support of diverse Programming Models, Available Tools, Adequate Memory Size and Speed, ...) is key
- Reliability of the file system is essential
- System architecture should be designed for highly scalable codes with low latency, low memory footprint for communication 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-g -> How to attract users of Vector Systems?
 - NEC Nehalem -> How to convince people to move from Standard IB Based environments (ISV Software!) to the new system?
 - Vendor<->Provider Collaboration needs special attention

Procurement: Major time consumers (preparing the RFP)

- How much GPGPUs/Accelerators do we need and when?
- Is PGAS ready for a major uptake with the new system features?
- How do we allow our users to perform pre- and postprocessing of data without moving (too much) data around?
- How hard will the change for a user familiar with the Vector Architecture be? How can they be supported appropriately?
- Memory Bandwidth per Flop is decreasing. What is the concrete impact? How do we measure the impact?
- How can we appreciate power efficiency of systems?

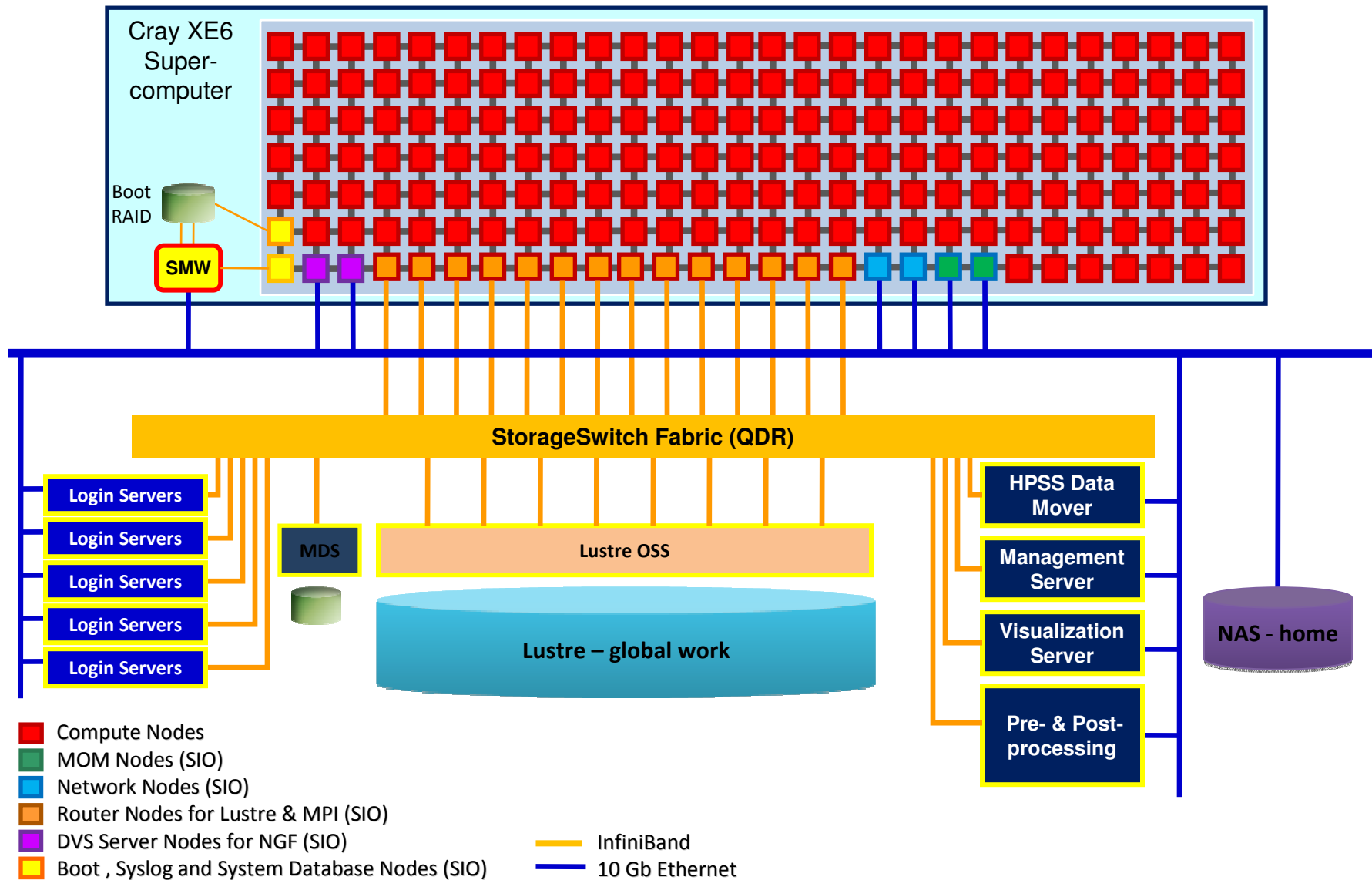
.....



KEY SYSTEM PROPERTIES

PHASE 1 STEP1

Conceptual Architecture



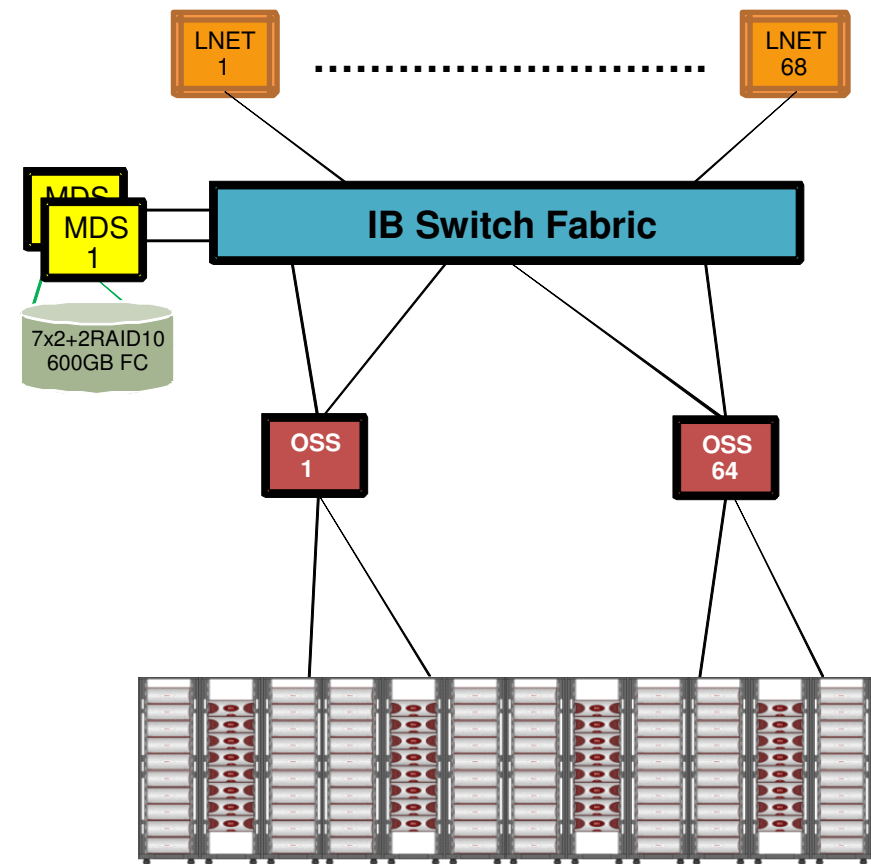
Phase 1 Step 1: Overview

- Configuration:
 - Peak Performance ~ 1PF
 - More than 3500 nodes
 - Each Node will have 2 sockets
 - AMD Interlagos @ 2.3GHz 16 Cores each leading to >100.000 cores
 - Nodes with 32GB and 64GB memory reflecting different user needs
 - 2.7PB storage capacity @ ~ 150GB/s IO bandwidth
 - External Access Nodes, Pre- & Postprocessing Nodes, Remote Visualization Nodes
 - ~2MW maximal power consumption
- Essential part of the contract is an intensive support by on-site staff and a collaboration agreement between HLRS and CRAY
- Support for ISV Codes depending on the app under CLE („native“) or CCM
- Part of the acceptance tests are validation of sustained application performance spanning across the full system



Phase 1 Step1: Storage Solution

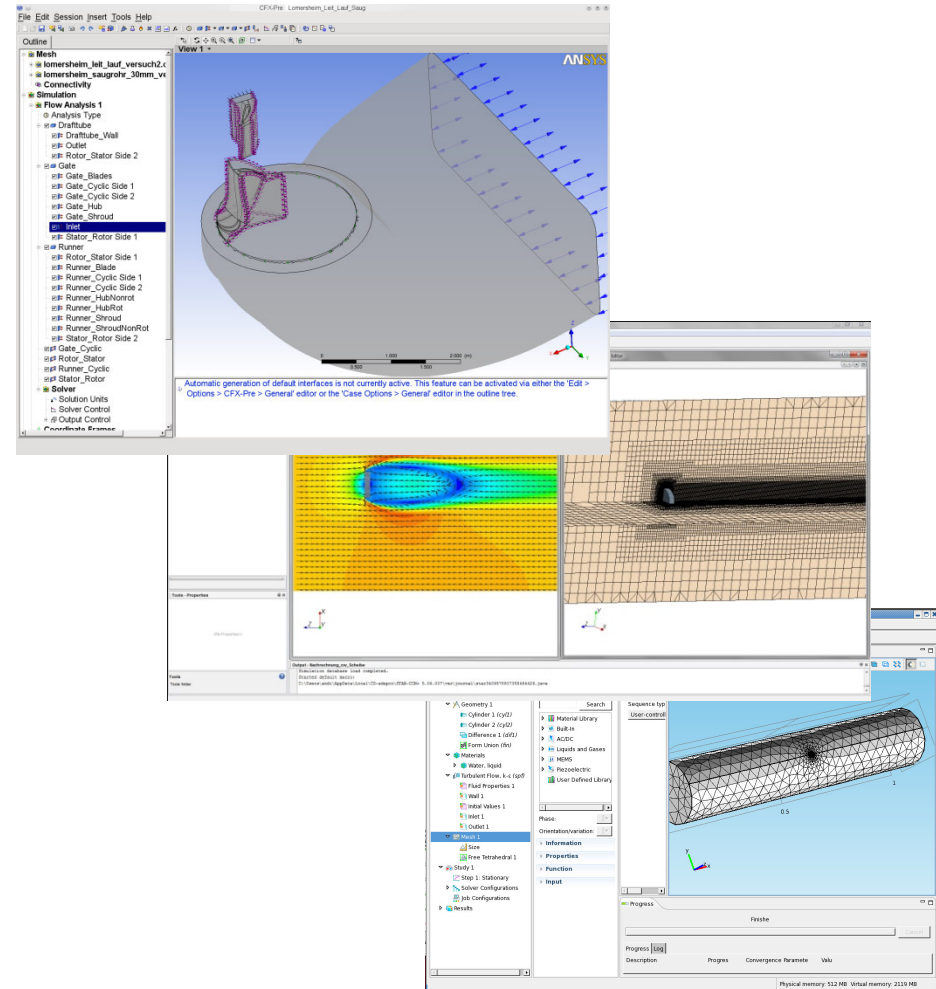
- Lustre based solution for the fast disk space
 - 2.7 PB Lustre Workspace capacity
 - Realised with 16 DDN SFA10k controllers
 - Integrated into the overall HLRS environments
- HLRS wide Home Space with 60TB capacity
- Local storage capacity with 20TB for Pre- and Postprocessing servers
- Integrated with the existing HPSS system using a data mover concept



3840TB Total Raw Capacity

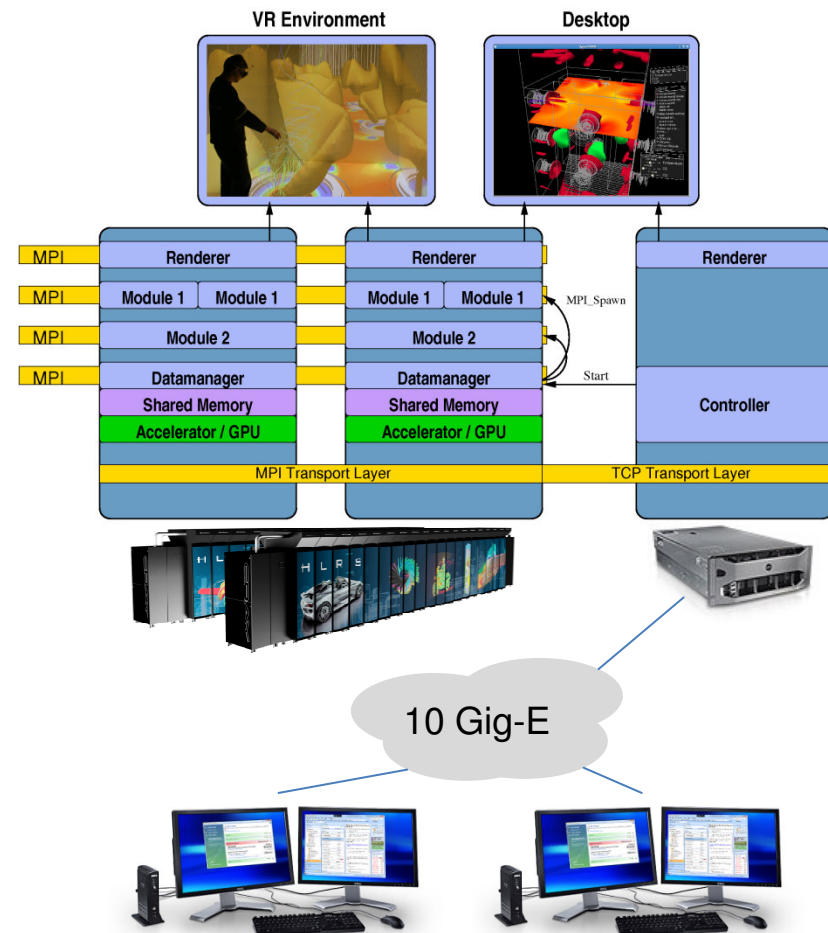
Phase 1 Step 1: Pre- and Postprocessing

- Pre- and Postprocessing servers support users in their workflow with large memory and full Linux environment
- Phase 1 Step 1 will have several Pre-/Postprocessing servers with up to 1TB of main memory
- The following step will add more servers with up to 2TB main memory
- Resources will be also controlled by the scheduling system



Phase 1 Step 1: Remote Visualization Servers

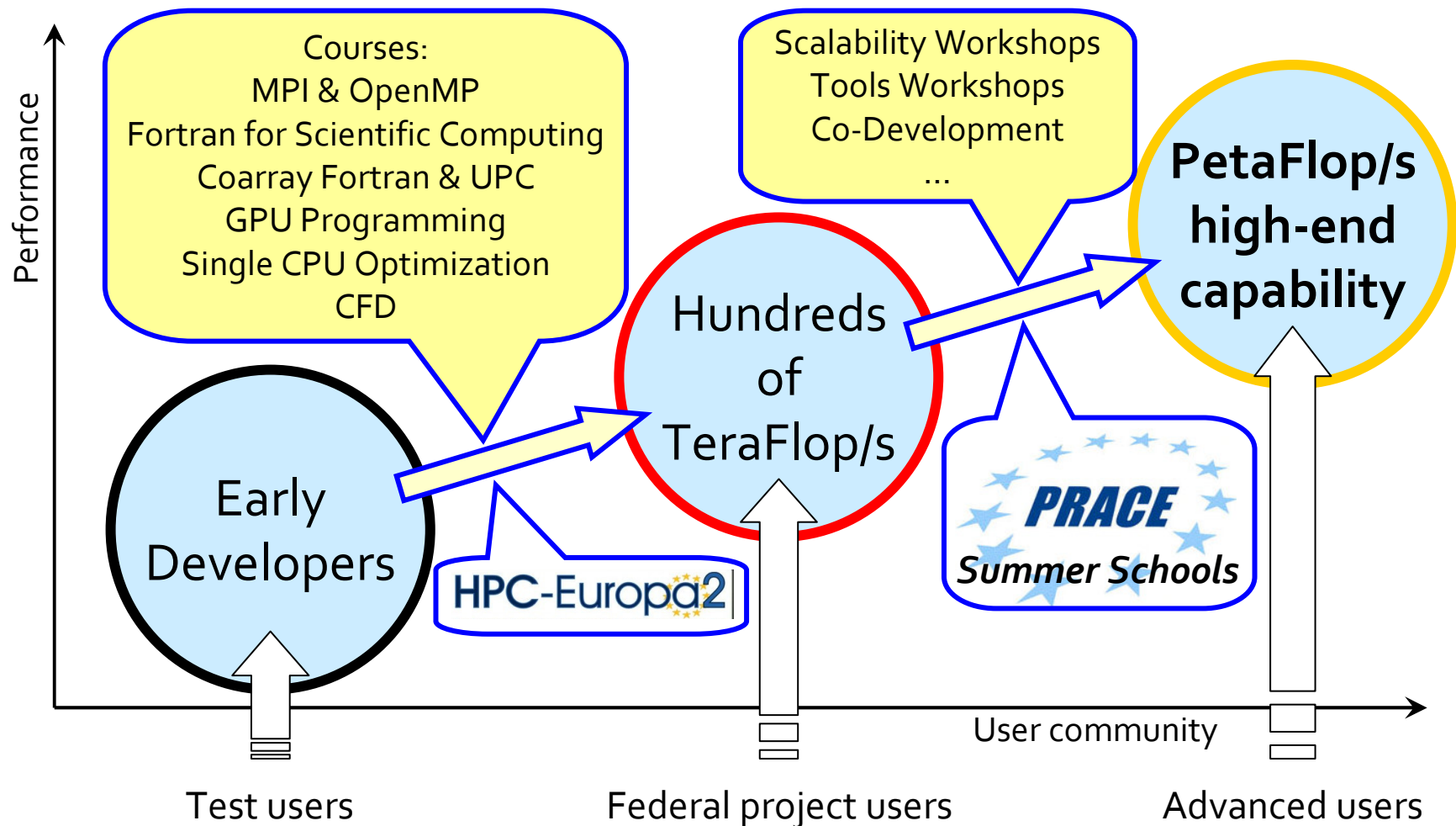
- Visualization Servers allow remote graphical access
 - Full Hardware accelerated graphics through dedicated remote graphics hardware as well as VirtualGL
 - Powerful Graphic Cards with GP-GPU post processing support beyond the capabilities of a typical end-user system
 - Visualization of very large 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



A glimpse on Phase 1 Step 2

- Step 2 will run in parallel to Step 1 as a single system from the user perspective
- Goal is to maintain similar software stack for both installation steps
- Expected architectural changes
 - Aries interconnect
 - Newest generation of CPUs
 - Partially relying on accelerators
 - Updated storage infrastructure e.g. 5.8 PB Lustre workspace
 - Additional external servers
- Significantly increased sustained application performance
- Scheduled for Autumn 2013
- Overall peak performance of complete Phase 1 will be >5PF





Evolution of the User Community



Courses — 2011

This year



Jan.	17–21	HLRS	Stuttgart	Fortran for Scientific Computing
Jan.	24–25	HLRS	Stuttgart	GPU Programming using CUDA (optional course)
Feb.	02–04	HLRS	Stuttgart	Cray XE6 Optimization Workshop 
Feb.	14–17	ZIH	Dresden	Parallel Programming
Feb.	28–Mar.3.	HLRS	Stuttgart	Iterative Gleichungssystemlöser und Parallelisierung
Mar.	28–30	HLRS	Stuttgart	VI-HPS Tuning Workshop 
Apr.	04–08	HLRS	Stuttgart	Computational Fluid Dynamics
Apr.	18–19	HLRS	Stuttgart	Plattformen am HLRS – 2nd day: Cray XE6 Optim. 
July	04–06	HLRS	Stuttgart	GPU Programming using CUDA
July	07–08	HLRS	Stuttgart	Introduction to UPC and Co-Array Fortran
July	11–15	HLRS	Stuttgart	Fortran for Scientific Computing
Aug.	01–03	TUHH	Harburg	Parallel Programming
Sep.	06–07	HLRS	Stuttgart	Cray XE6 Optimization Workshop
Sep.	12–16	LRZ	Garching	Iterative Gleichungssystemlöser und Parallelisierung
Sep.	26–30	RWTH	Aachen	Computational Fluid Dynamics
Oct.	10–12	HLRS	Stuttgart	Parallel Programming (in English)
Oct.	13–14	HLRS	Stuttgart	Advanced Topics in Parallel Programming (in English)
Nov.	02–04	HLRS	Stuttgart	Cray XE6 Optimization Workshop 
Nov.	28–30	JSC	Jülich	Parallel Programming
Dec.	5-7 / 8–9	HLRS	Stuttgart	CUDA / UPC and Co-Array Fortran

.....

CURRENT STATUS

Status: Infrastructure update



HLRS Phase 1 Step 0 configuration

- XE6 s/n 4157 (1 cabinet)
- 84 compute nodes, 32 GB memory, AMD Opteron 8-Core, 2.0 GHz Magny Cours (1344 total cores)
- 12 service nodes, 16 GB memory, AMD Opteron 6-Core, 2.2 GHz Istanbul (72 total cores)

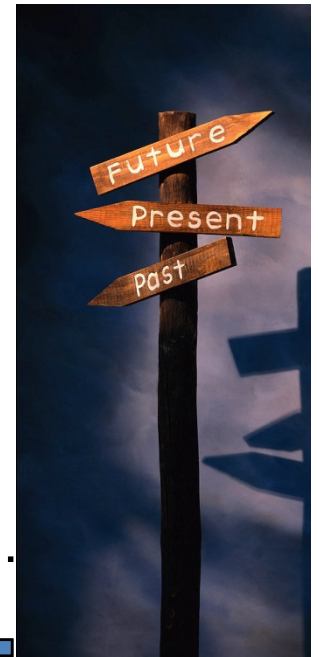


Targets for Phase 1 Step 0: HLRS View

- Experiment with different configuration options meeting the site specific requirements
- Work on tools/scripts making migration from current systems to the new systems as painless as possible
- Get Local Staff trained and prepared for the big system
- Evaluate new software tools (in particular software currently in beta and final in autumn)
- Understand new hardware features (e.g. Gemini), software tools und experiment with in-house applications to prepare for consultancy of our users

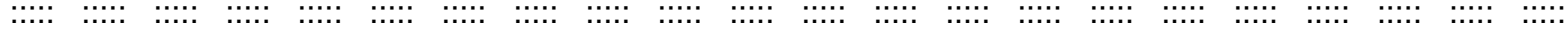
Targets for Phase 1 Step 0: User View

- More than 90 projects have already access to the test system
- Several Training events with partially more than 70 participants have been organized
- Major focus of collaboration between HLRS Users, HLRS Staff and Cray is
 - Support users in porting their applications on the new system
 - Mimic the HLRS specific environment as close as possible for painless migration
 - Evaluate performance of ISV applications
 - Finalize the configuration and validate the external server setting
 - Realize an integration into the HLRS environment
- Interest and Expectations are high and lot's of work is ahead..



Phase 1 Step 0: Access to the Test System

- All confirmed federal user projects (from NEC SX-9 and NEC Nehalem) have free access to the Cray XE6 test system
- New users should apply via
 - www.hlrs.de → Systems → Access and Usage Models
 - How to get an account
 - Research Access to the National Supercomputers through Review Procedure
 - Test account
 - They should apply for a NEC Nehalem test account
 - The test account for the Cray XE6 is automatically added
- The system is for **testing** and **porting**
- **No production** intended



FUTURE OF CURRENT SYSTEMS

Future of current systems (based on current planning)

- Cray XE6 Test system:
 - will be upgraded to be part of Phase 1, Step 1
- NEC SX-9:
 - End 2011, shut-down of whole system or parts of it
- NEC Nehalem Cluster:
 - Expected to be still available in 2012
- bwGrid:
 - Expected to be still available in 2012
- Cray XT5m:
 - Expected to be still available in 2012



THANK YOU! ANY QUESTIONS?