HiPerCH Workshop
8th – 11th April 2013

High Performance Computing Hessen

TECHNISCHE UNIVERSITÄT DARMSTADT

Introduction: System Software – TU Environment

Dr. Andreas Wolf, (HRZ) TU Darmstadt

HRZ Hochschulrechenzentrum

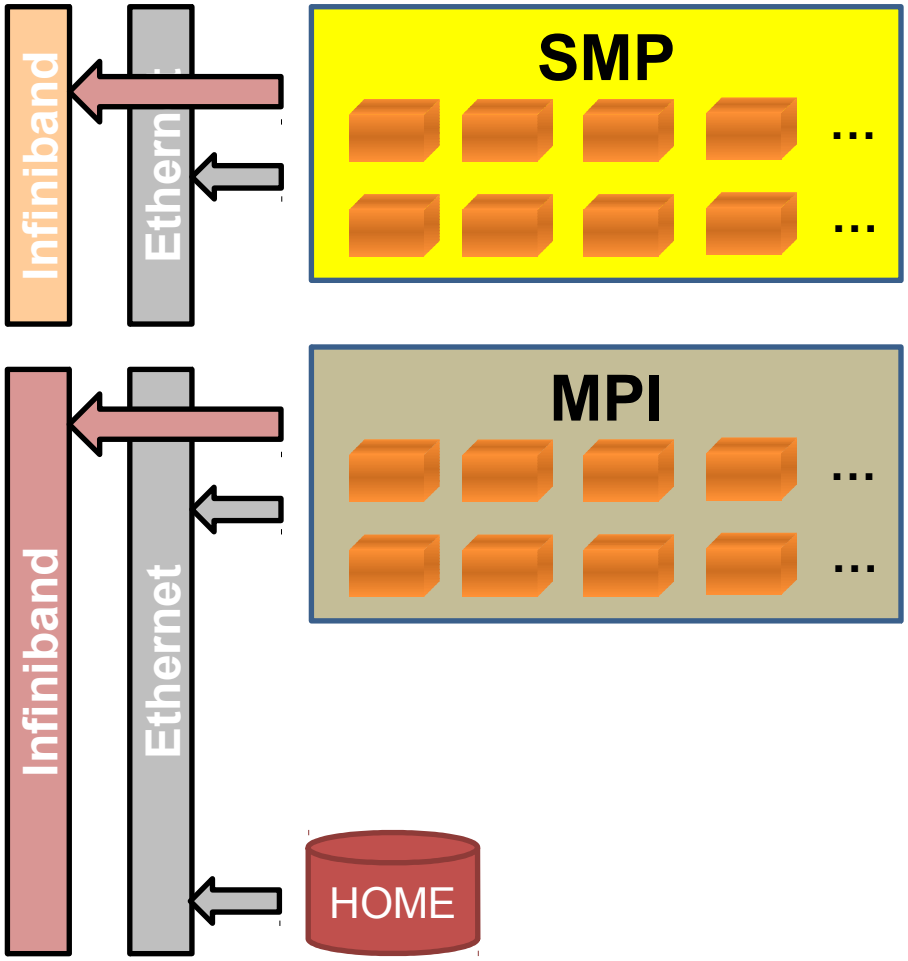
Overview

TECHNISCHE UNIVERSITÄT DARMSTADT

- Current and next system (Hardware)
- Software packages
  - Module system
- Batch system
  - Queueing rules
  - Commands
  - Batch script examples for MPI and OpenMP
  - Hints and tricks
- Access requirements

- Current and next system (Hardware)
- Software packages
  - Module system
- Batch system
  - Queueing rules
  - Commands
  - Batch script examples for MPI and OpenMP
  - Hints and tricks
- Access requirements

A part of the new System – UCluster



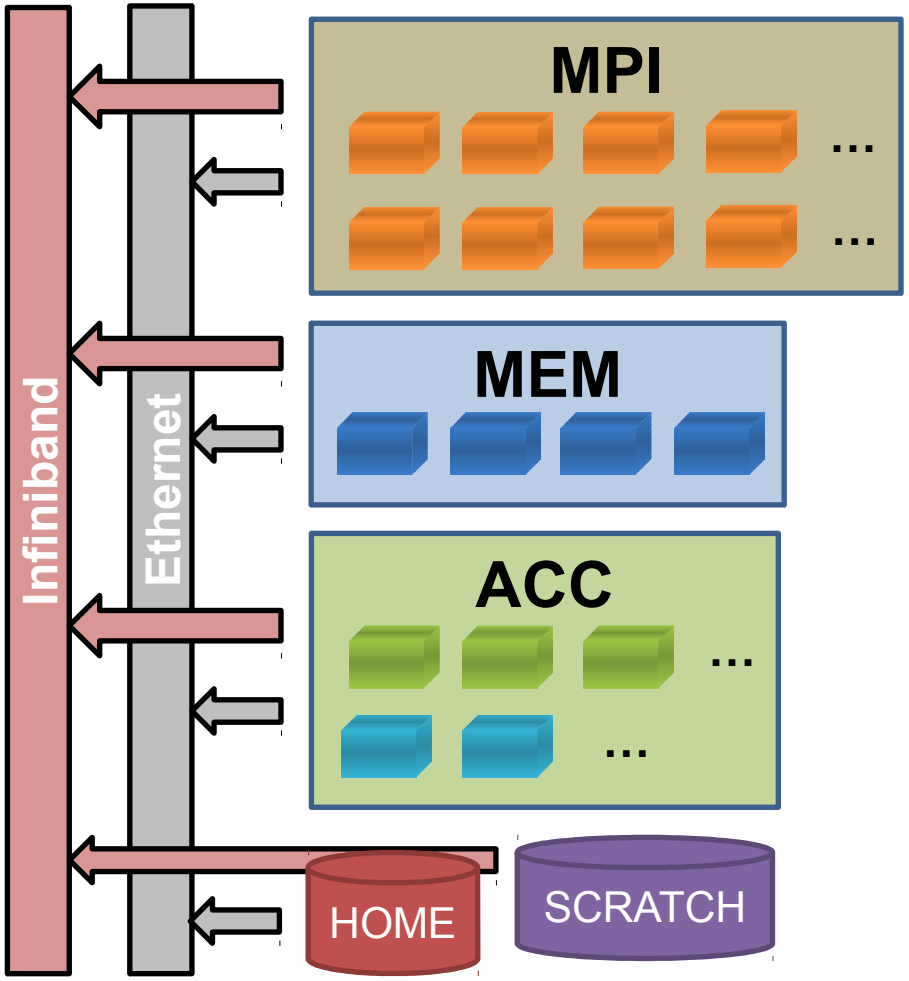
32x **SMP** (**ICluster** – older System)
- 4 Processors, AMD Opteron
  - each 12 Cores, 2.6 GHz
- 64-128 GByte

5x32 **MPI** (**UCluster**)
- 2 Processors, Intel Sandybridge
  - each 8 Cores, 2.6 GHz
- 32 GByte

File System
- **Home**: 500 TByte

Infiniband (only island internally)
- **FDR-10**

Hardware 2013 – Phase I



704 x **MPI** (inclusive UCluster)
- 2 Processors, Intel Sandybridge
  - each 8 Cores, 2.6 GHz
- 32 GByte (10% 64 GByte)

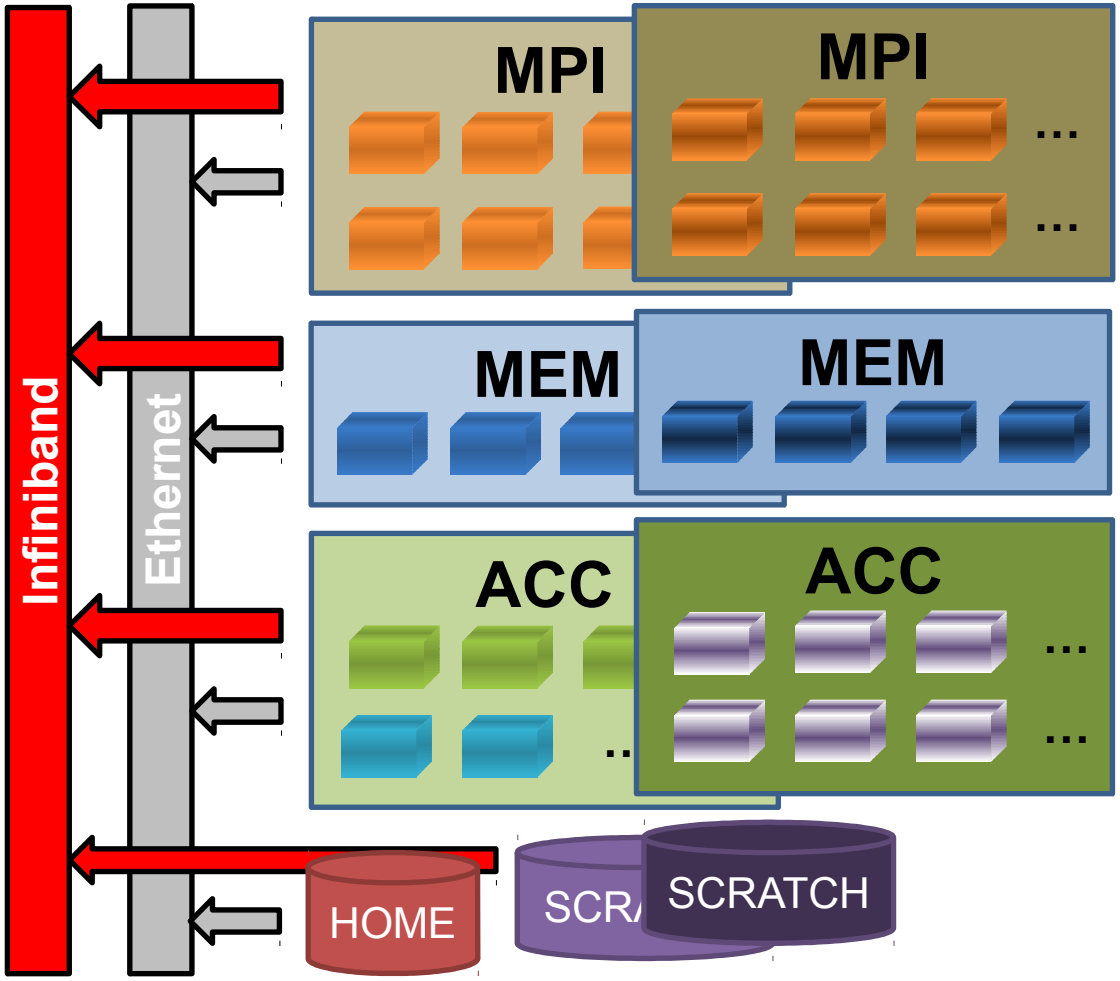
4 x **MEM**
- 8 Processors, each 8 Cores
- 1024 GByte

64 x **ACC**
- 2 Processors + 2 Accelerators
  - Nvidia Kepler
  - Intel Xeon Phi (former MIC)
- 32 GByte

File Systems
- **Scratch**: 768 TByte, 20 GB/s
- **Home**: 500 TByte, 5 GB/s

Infiniband (islands interconnected)
- **FDR-10**

Hardware 2014 – Phase II



Additional **MPI**
- 2 Processors
  - Successor architecture

4x Additional **MEM**
- 4 Processors
- Successor architecture
- 1024 GByte

Additional **ACC**
- 2 Processors
- 2 Accelerators
  - Successor architecture

File Systems
- **Scratch**: +768 TByte
  - Overall 1.5 PByte

Infiniband
- **FDR**, 54 Gbit/s, ~1 µs Latency

- Current and next system (Hardware)
- Software packages
  - Module system
- Batch system
  - Queueing rules
  - Commands
  - Batch script examples for MPI and OpenMP
  - Hints and tricks
- Access requirements

Software available / installable

TECHNISCHE UNIVERSITÄT DARMSTADT

Operation System

- SLES (SP2), x86-64, 64Bit

System Tools

- GCC 4.4.6, 4.7.2 …
- Intel 12.1.0, 13.0.1 (incl. Intel Cluster Studio XE → Wednesday)
- PGI 13.1
- ACML, Intel-MKL, SCALAPACK etc.
- OpenMPI, Intel-MPI, ...
- Totalview (→ Tuesday), Vampir (→ Tuesday)

Applications

- Ansys 140, 145 ...
- Abaqus 6.12-1
- Matlab 2012a
- COMSOL 4.3

Modular Load and Unload – 1

TECHNISCHE UNIVERSITÄT DARMSTADT

> **module list**

- Shows all currently load software environments (load packages of the user)

> **module load** <module name>

- Loads a specific software environment module
  - Only when the module is successfully loaded – the software is really useable!

```
/home/user/@hpa0392:~> module load ansys/145
Loaded module ansys/145
/home/user/@hpa0392:~>
```

> **module unload** <module name>

- Unloads a software module

Modular Load and Unload – 2

TECHNISCHE UNIVERSITÄT DARMSTADT

> **module avail**

- Shows all available software packages currently installed

```
/home/user/:~> module avail

------------------------------------------------------------------------------
cluster-tools/5.2 dot                freeipmi/1.0.2    ipmitool/1.8.11   module-info

------------------------------------------------------------------------------
abaqus/6.12-1(default)                        comsol/v-4.3a
acml/5.1.0_gcc-4.7.0(default)                 fftw2/2.1.5_gcc-double(default)
acml/5.1.0_gcc-4.7.0mp                        fftw2/2.1.5_gcc-float
acml/5.1.0_gcc64-4.7.0                        fftw3/3.2.2_gcc(default)
acml/5.1.0_gcc64-4.7.0mp                      gcc/4.3.6
acml/5.1.0_intel-13.0.1                       gcc/4.4.6
acml/5.1.0_intel-13.0.1mp                     gcc/4.6.2
acml/5.1.0_intel64-13.0.1                     gcc/4.7.0
acml/5.1.0_intel64-13.0.1mp                   gcc/4.7.2(default)
ansys/130                                     globalarrays/5.0.2_gcc(default)
ansys/140                                     hdf5/1.6.10(default)
ansys/145(default)                            hwloc/1.1.1(default)
comsol/v-4.3(default)                         icem/140(default)
/home/user/:~>
```


HRZ Hochschulrechenzentrum

- Current and next system (Hardware)
- Software packages
  - Module system
- Batch system
  - Queueing rules
  - Commands
  - Batch script examples for MPI and OpenMP
  - Hints and tricks
- Access requirements

Queueing – Scheduling

TECHNISCHE UNIVERSITÄT DARMSTADT

- Different queues – for different purposes
  - **deflt** – Limited to maximal 24 hours
    - Main part of all computing nodes
  - **long** – Limited to maximal 7 days
    - Very small part (~ 10%) of all computing nodes
  - **short** – Limited to maximal 30 minutes
    - Depending on the demand – some nodes, but with highest scheduling priority
- Advantages
  - Maintainability of the most computing nodes within 24 hours
    - Because of the main focus to 24 hours jobs
  - Small test jobs (30 minutes) will scheduled promptly

Batch System - LSF

TECHNISCHE UNIVERSITÄT DARMSTADT

Why using LSF?

- Scalability for a large number of nodes
- Professional support (for fine tuning)
- WebGUI – graphical front-end for
  - Creating
  - Submitting
  - Monitoring

  of batch jobs

  (At present unfortunately not ready for use – later)

➔ Usability also from a Windows client

Batch System Commands - LSF

TECHNISCHE UNIVERSITÄT DARMSTADT

> **bsub <** <batch script>

- Submit a new batch job to the queueing system

> **bqueue**

- Shows all presently submitted or active batch jobs and their batch-ID numbers

> **bkill** <batch-ID>

- Deletes own batch jobs (with ID ...)

> **bjobs** <batch-ID>

- Shows specific configuration or runtime information for a batch job

LSF – bsub & bkill

TECHNISCHE UNIVERSITÄT DARMSTADT

```
/home/user/@hpa0392:~/memory_bandwidth> bsub < run_mb64.sh
Job <1796> is submitted to default queue <deflt>.
/home/user/@hpa0392:~/memory_bandwidth> bkill 1796
Job <1796> is being terminated
/home/user/@hpa0392:~/memory_bandwidth>
```

LSF - bqueue

TECHNISCHE UNIVERSITÄT DARMSTADT

```
/home/user/@hpa0392:~/memory_bandwidth> bqueues
QUEUE_NAME      PRIO STATUS          MAX JL/U JL/P JL/H NJOBS  PEND   RUN  SUSP
long              43  Open:Active      -    -    -    -     0     0     0     0
short             43  Open:Active      -    -    -    -     0     0     0     0
deflt             43  Open:Active      -    -    -    -    64     0    64     0
/home/user/@hpa0392:~/memory_bandwidth>
```


HRZ Hochschulrechenzentrum

LSF - bjobs

```
/home/user/@hpa0392:~/memory_bandwidth> bjobs
JOBID   USER    STAT  QUEUE      FROM_HOST   EXEC_HOST   JOB_NAME   SUBMIT_TIME
1797 <UserName>RUN   deflt      hpa0392.ida 64*i11     <JobName>  Feb 26 21:40
/home/user/@hpa0392:~/memory_bandwidth>
```

- Current and next system (Hardware)
- Software packages
  - Module system
- Batch system
  - Queueing rules
  - Commands
  - Batch script examples for MPI and OpenMP
  - Hints and tricks
- Access requirements

Batch Script for running a MPI Program - 1

TECHNISCHE UNIVERSITÄT DARMSTADT

```
#Job name
#BSUB -J MPItest

#File / path where STDOUT will be written, the %J is the job id
#BSUB -o /home/<TU-ID>/MPItest.out%J

#File / path where STDERR will be written, the %J is the job id
#BSUB -e /home/<TU-ID>/MPItest.err%J

#Request the time you need for execution in minutes
#The format for the parameter is: [hour:]minute,
#that means for 80 minutes you could also use this: 1:20
#BSUB -W 10

#Request virtual memory you need for your job in MB
#BSUB -M 18000
```

suppress full path

Batch Script for running a MPI Program - 2

TECHNISCHE UNIVERSITÄT DARMSTADT

```
...
#Request the number of compute slots you want to use
#BSUB -n 64

#Specify the MPI support
#BSUB -a openmpi

#Define the host type at the job level
#BSUB -R "select[type=any]"

#Specify your mail address
#BSUB -u <email address>

#Send a mail when job is done
#BSUB -N
```

Batch Script for running a MPI Program - 3

TECHNISCHE UNIVERSITÄT DARMSTADT

```
...

module load openmpi/1.6.4

module list

cd ~/<working path>

mpirun  Program

echo "$LSB_HOSTS" | sed -e "s/ /\n/g" > hostfile.
$LSB_JOBID
mpirun  -n 64  -hostfile host
```

OpenMPI 1.6.4 and newer

MPI (without LSF Support)
→ currently Intel MPI

HRZ Hochschulrechenzentrum

Batch Script for running a OpenMP Program

TECHNISCHE UNIVERSITÄT DARMSTADT

```
...

#Specify the OpenMP support
#BSUB -a openmp

#export OMP_NUM_THREADS=16

<OpenMP program call>
```

- Current and next system (Hardware)
- Software packages
  - Module system
- Batch system
  - Queueing rules
  - Commands
  - Batch script examples for MPI and OpenMP
  - Hints and tricks
- Access requirements

„What can be go wrong“ - Hints for Performance

TECHNISCHE UNIVERSITÄT DARMSTADT

- Batch job didn't start (also after days/weeks)
  - Resource requirements
    - Batch system gives a concrete error message
    - Resources currently not installed (not available)
      - Often the batch system can't decide, that a requirement will never be fulfilled (e.g. nodes and memory can later be increased)
- Some computing cores of a node will not be used
  - During requirement defining, please take care of the hardware
- Every time a different runtime for the same job (workload)
  - (Runtime can differ up to 50%)
  - → The use of processor binding is often the solution
  - → Take care of the ccNUMA characteristics of the used system

16, 48, 64 Cores

Processor Binding

TECHNISCHE UNIVERSITÄT DARMSTADT

MPI

- Additional parameters (MPI specific): **-bind-to-core -report-bindings**

(OpenMPI)

mpirun **-bind-to-core -report-bindings** **Program**

OpenMP

- Compiler specific environment variables

(GNU)

export GOMP_CPU_AFFINITY="0-15"

MPI vs. Compiler – Support matrix (in development)

TECHNISCHE UNIVERSITÄT DARMSTADT

Compiler / MPI	OpenMPI	Intel-MPI	(MPICH)	(MVAPICH)
GNU (4.7.x)	x	(x)	(later)	(later)
Intel (13.x)	(later)	(x)	(?)	(?)
PGI (13.x)	(later)	(?)	(?)	(?)

Using Intel Compiler

- **module load intel** (loads 13.0.1 without tools = default)
- But (!) with Intel-MPI
  - → **module load intel/13.0.1_icsxe** (all Intel tools will be loaded)
- Important: it is necessary to load the appropriate GCC version
  - → **module load gcc** (4.5 or higher)

Using Intel-MPI

- with Intel Compiler
  - **mpiicc** (C), **mpiicpc** (C++), **mpiifort** (Fortran 77/90)
- with GNU Compiler
  - **mpicc** (use of gcc), **mpicxx** (use of g++), **mpifc** (use of gfortran)

- Current and next system (Hardware)
- Software packages
  - Module system
- Batch system
  - Queueing rules
  - Commands
  - Batch script examples for MPI and OpenMP
  - Hints and tricks
- Access requirements

Access Requirements for the new Cluster

TECHNISCHE UNIVERSITÄT DARMSTADT

- Because of the size of the system
  - Each potential user needs to be proved first against the export limitations of the "Bundesamt für Wirtschaft und Ausfuhrkontrolle (BAFA)" - Export control/Embargo
- E-Mail at HHLR@hrz.tu-darmstadt.de
  - The new "user rules" document ("Nutzungsordnung")
    - Names, TU-ID, Email address, Institute and Institution affinity, **Citizenship**
    - **Project title**
    - Reports
    - No private data (email, pictures etc.)
    - No commercial use
    - Limited data storage life

Access to the UCluster - for the Lab

TECHNISCHE UNIVERSITÄT DARMSTADT

- per SSH
- Login-Knoten: ucluster1.hrz.tu-darmstadt.de
  ucluster2.hrz.tu-darmstadt.de

```
/home/user/@client:~> ssh username@ucluster1.hrz.tu-darmstadt.de
username@ucluster1.hrz.tu-darmstadt.de's password:
```


HRZ Hochschulrechenzentrum

HPC Newsletter

TECHNISCHE UNIVERSITÄT DARMSTADT

Email Newsletter: HPC@lists.tu-darmstadt.de

Newsletter subscription

- https://lists.tu-darmstadt.de/mailman/listinfo/hpc
- Information about
  - Planned Events, User meetings
  - Planned Lectures, Workshops etc.
  - Common information of the system / news

Thank you for your attention

Questions ???