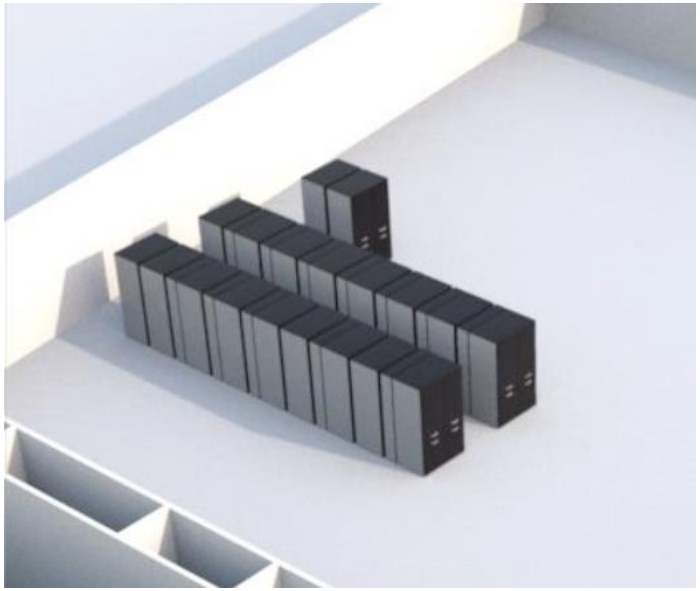
Computer systems and software environment

Dirk Schmidl

RWTH AACHEN UNIVERSITY

Computer System

- **CLAIX – Phase I**
  - → Two Rack-Rows + 2 Racks
- **Test operation since November 2016**
  - → Tier-2 HPC-System
- **Production mode since January 2017**
- **TOP500 in November 2016:**
  - → 506.87 TFLOPS (528 TFLOPS Peak)
  - → Rank 309 in TOP500, 96% Efficiency
- **15000+ Intel Xeon Cores**
- **100 Gbit/s Intel Omni-Path HPC Network, 3.3 PB Lustre Storage**



Computer System

Orchestrating a brighter world

NEC

water cooled
(closed)

air cooled
(open)

Technical Specification: MPI Node

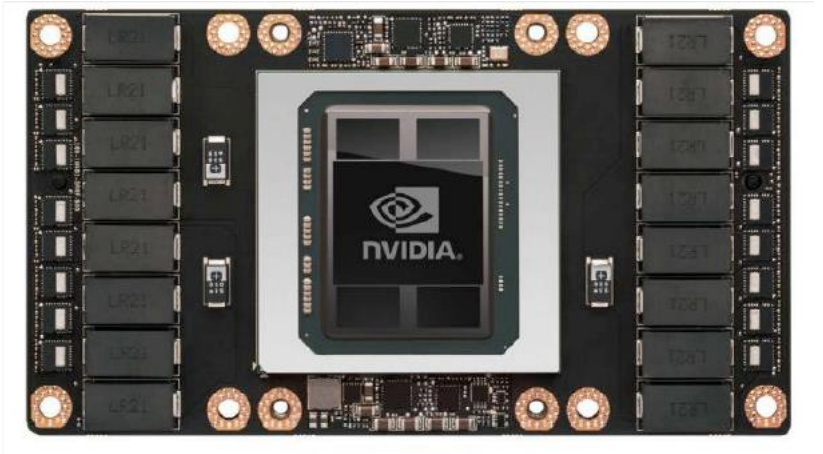
- **609 Standard MPI Nodes**
  - → 2x Intel Xeon Broadwell-EP (E5-2650v4): 12 Cores, 2.2 GHz
    - → Turbo: 2.9 GHz (2.5 GHz for AVX-Instructions)
  - → 128 GB DDR4-2400 Main Memory
  - → Intel Omni-Path x16 HCA
- **4 Servers in 2 HU**
- **air cooled nodes in water cooled closed racks**

Intel Xeon Broadwell

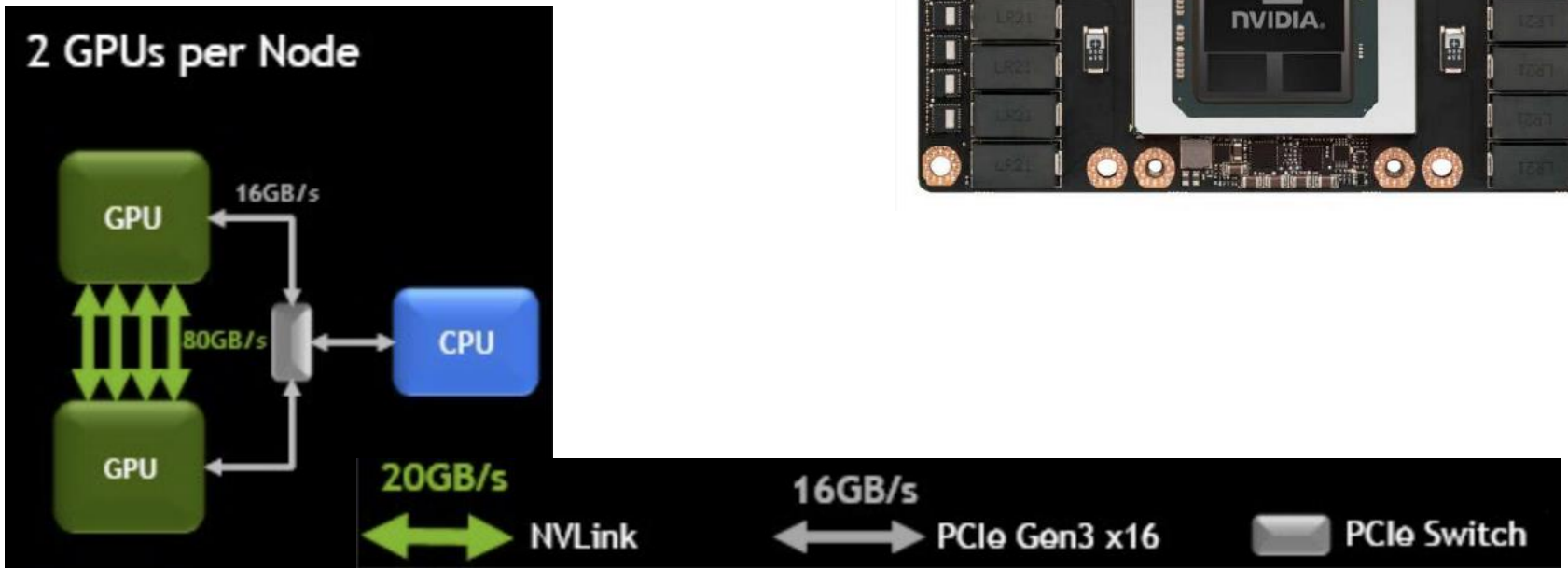
- **Product name: E5-v4**
  - → Intel: the latest and greatest (for now...) ;-)
  - → 14 nm technology
- **Our choice:**
  - → “Advanced” Segment: full memory bandwidth and large caches
  - → DDR4-2400 Memory, 9.6 GT/s QPI
  - → Modell with 12 Cores: Optimized for prize performance of our Jobmix

Technical Specification : GPU Node

- **10 GPU nodes**
  - → Host as MPI node, plus:
  - → 2x NVIDIA P100 (Pascal Architecture)



- **Connected via NVLink**



Technical Specification : SMP Node

- **8 Standard SMP Nodes**
  - → 8x Intel Xeon Broadwell-EX (E7-8860v4): je 18 Cores, 2.2 GHz
  - → 1 TB Main Memory
  - → 2x 2 TB NVMe SSD
  - → Intel Omni-Path x16 HCA

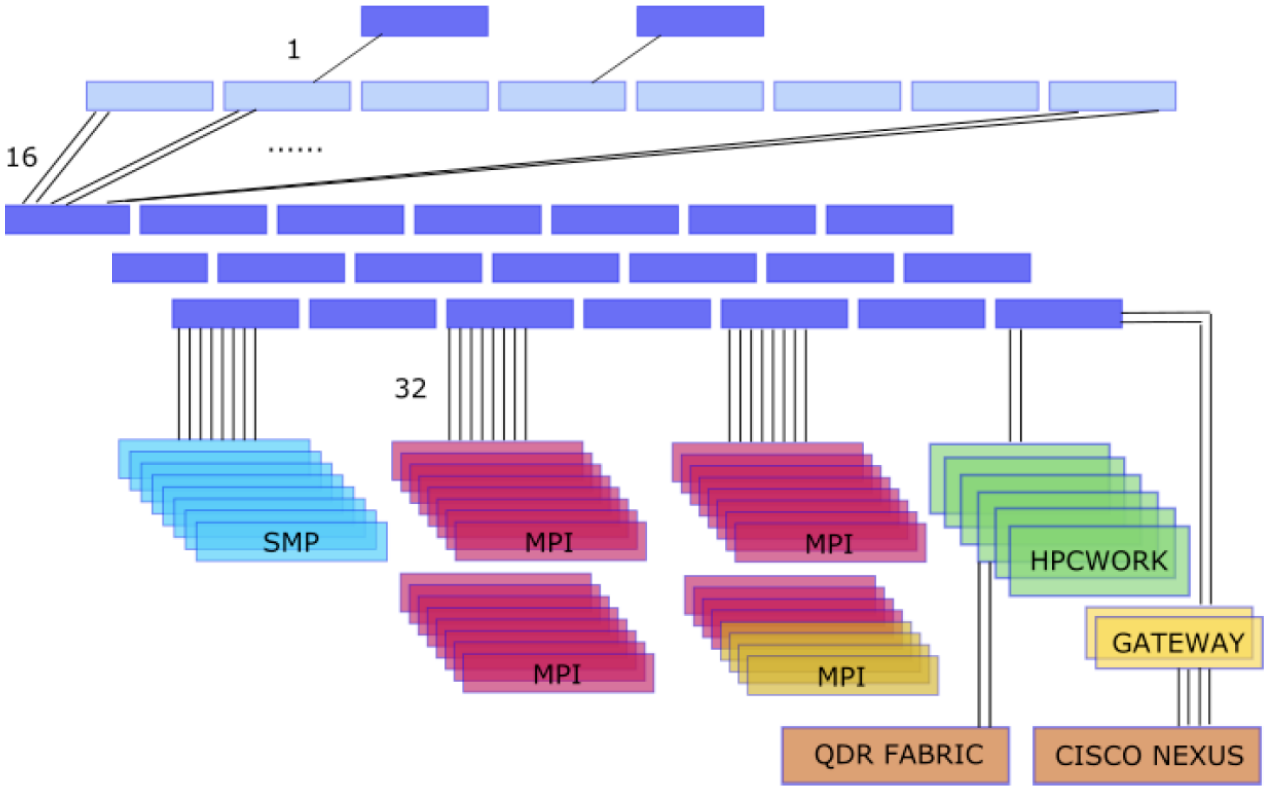
- **2 Systems with 1x NVIDIA P100**
  - → PCIe Card
- **Air cooled systems**

Technical Specification : NVMe node

- **8 NVMe nodes**
  - → Host as standard MPI node, plus:
  - → 2 TB NVMe SSD
- **Currently: Evaluation of usage modes**
  - → File system
  - → Mapped Memory
  - → Allocated Memory
- **Evaluation of usage scenarios**
  - → In-situ analysis
  - → I/O-Nodes

HPC network

- **Intel Omni-Path network**
  - → 1:2 Blocking
  - → x16 PCIe nodes (100 Gbit/s)



- **2 Omni-Path to Eth Gateways**
  - → 2x 40 Gbit/s per Gateway

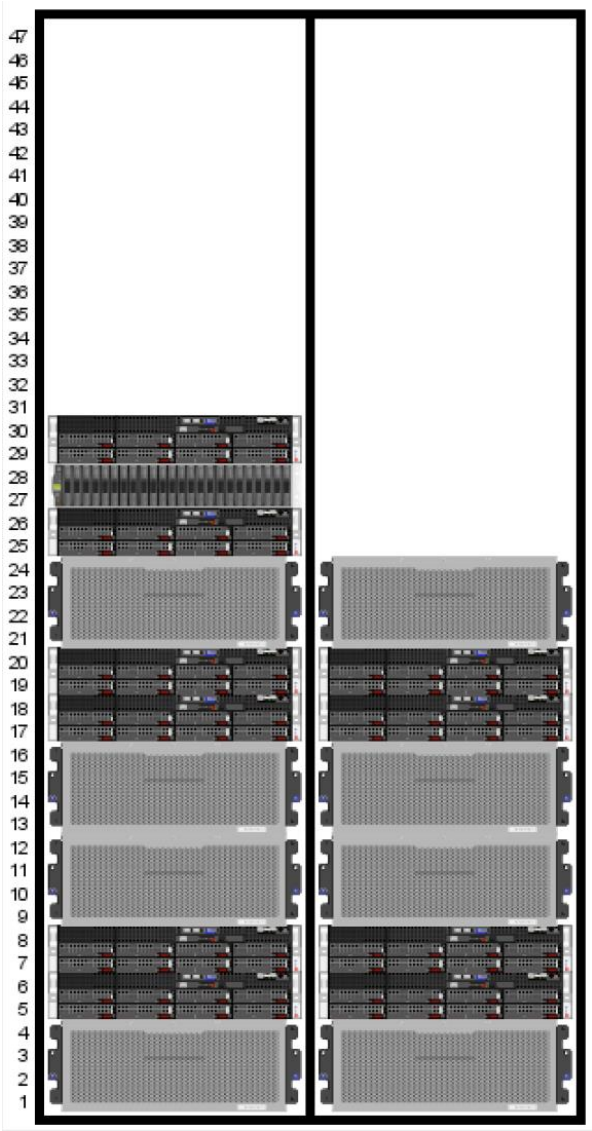
- **MPI latency (node to node): 0.93 usec**
- **MPI bandwidth (node to node): 93 Gbit/s**

Lustre Storage

- **Mount: $HPCWORK**
- **Capacity: 3300 TB**

- **Performance(bandwidth)**
  - → 55+ GB/s write
  - → 50+ GB/s read

- **First Installation of LXFS with ZFS in europe**
  - → Goal: End-to-end Integrity



Outlook

- **Evaluation of further architectures for 2018**
  - → Evaluation Cluster of Intel Xeon Phi Processor nodes
    - →Many-Core Technology: 240+ Cores per processor
  - → Cooperation with NEC: Evaluation of AURORA (Vectorarchitecture)

- **2018: Switch to warm water cooling**

How to login

Frontends

cluster.rz.RWTH-Aachen.DE	cluster2.rz.RWTH-Aachen.DE
cluster-x.rz.RWTH-Aachen.DE	cluster-x2.rz.RWTH-Aachen.DE
cluster-linux.rz.RWTH-Aachen.DE	cluster-linux-nehalem.rz.RWTH-Aachen.DE
cluster-linux-xeon.rz.RWTH-Aachen.DE	cluster-linux-tuning.rz.RWTH-Aachen.DE
cluster-copy.rz.RWTH-Aachen.DE	cluster-copy2.rz.RWTH-Aachen.DE
login.hpc.itc.rwth-aachen.de	**login2.hpc.itc.rwth-aachen.de**

- Use frontends to develop program, compile applications, prepare job scripts or debug programs
- login / SCP File transfer:

```
$ ssh [-Y] user@cluster.rz.rwth-aachen.de
$ scp [[user@]host1:]file1 [...] [[user@]host2:]file2
```

LSF: Job Submission (1/3)

- **How to submit a job**

  → `$ bsub [options] command [arguments]`

- **General parameters**

Parameter	Description
`-J <name>`	Job name
`-o <path>`	Standard out
`-e <path>`	Standard error
`-B`	Send mail when job starts running
`-N`	Send mail when job is done
`-u <mailaddress>`	Recepient of mails
`-P <projectname>`	Assign the job to the specified project (e.g. jara, integrative hosting costumers)
`-U <reservation>`	Use this for advanced reservations

LSF: Job Submission (2/3)

- **How to submit a job**

  → `$ bsub [options] command [arguments]`

- **Parameters for job limits / resources**

Parameter	Description
`-W <runlimit>`	Sets the hard runtime limit in the format [hour:]minute [default: 15]
`-M <memlimit>`	Sets a **per-process** memory limit in MB [default: 512]
`-S <stacklimit>`	Set a **per-process** stack size limit in MB [default: 10]
`-x`	Request node(s) exclusive
`-R "select[hpcwork]"`	ALWAYS set if you using the HPCWORK (Lustre file system)

LSF: Job Submission (3/3)

- **How to submit a job**
  - → `$ bsub [options] command [arguments]`
- **Parameters parallel jobs**

Parameter	Description
`-n <min_proc>[,max_proc]`	Submits a parallel job and specifies the number of processors required [default: 1]
`-a openmp`	Use this to submit a shared memory job (e.g. OpenMP)
`-a {open|intel}mpi`	Specify the MPI (remember to switch the module for Intel MPI)
`-R „span[hosts=1]“`	Request the compute slots on the same node
`-R „span[ptile=n]“`	Will span *n* proceses per node (hybrid)

→ `$MPIEXEC $FLAGS_MPI_BATCH a.out`

LSF: Using the Magic Cookie

- **You can use the magic cookie `#BSUB` for a batch script `job.sh`**

```
#!/bin/zsh
#BSUB -J TESTJOB                    #Job name
#BSUB -o TESTJOB.o%J                #STDOUT, the %J is the job id
#BSUB -e TESTJOB.e%J                #STDERR, the %J is the job id
#BSUB -W 100                        #Will run max 100 minutes
#BSUB -M 1024                       #Request 1024 MB memory
#BSUB -u user@rwth-aachen.de        #Specify your mail
#BSUB -N                            #Send a mail when job is done
cd /home/user/workdirectory         #Change to the work directory
a.out                               #Execute your application
```

- **Submit this job**

→ `$ bsub < job.sh`

LSF: Job Status (1/2)

- **Use `bjobs` to display information about LSF jobs**

→ `$ bjobs [options][jobid]`

```
JOBID   USER    STAT  QUEUE      FROM_HOST   EXEC_HOST   JOB_NAME   SUBMIT_TIME
3324    tc53084 RUN   serial     linuxtc02   ib_bull     BURN_CPU_1 Jun 17 18:14
3325    tc53084 PEND  serial     linuxtc02   ib_bull     BURN_CPU_1 Jun 17 18:14
3326    tc53084 RUN   parallel   linuxtc02   12*ib_bull  *RN_CPU_12 Jun 17 18:14
3327    tc53084 PEND  parallel   linuxtc02   12*ib_bull  *RN_CPU_12 Jun 17 18:14
```

Option	Description
-l	Long format – displays detailed information for each job
-w	Wide format - displays job information without truncating fields
-r	Displays running jobs
-p	Displays pending job and the **pending reasons**
-s	Displays suspended jobs and the suspending reason

→ LSF can display the reasons for a pending job

LSF: Job Status (2/2)

- **Use `bpeek` to display stdout and stderr of an running LSF job**

  → `$ bpeek [options][jobid]`

```
<< output from stdout >>
Allocating 512 MB of RAM per process
Writing to 512 MB of RAM per process
PROCESS 0: Hello World!
PROCESS 1: Hello World!
[ application output ]
<< output from stderr >>
```

- **Remove a job from the queue**

  → `$ bkill [jobid]`

- **Remove all jobs from the queue**

  → `$ bkill 0`

during this week

#BSUB -P hpclab
#BSUB -U PPCES17

Module System (1/2)

- **Many compilers, MPIs and ISV software**
- **The module system helps to manage all the packages**
  - → List loaded modules
    - → `$ module list`
  - → List available modules
    - → `$ module avail`
  - → Load / unload a software
    - → `$ module load <modulename>`
    - → `$ module unload <modulename>`
  - → Exchange a module (Some modules depend on each other)
    - → `$ module switch <oldmodule> <newmodule>`
  - → Reload all modules (May fix your environment, especially with a NX session)
    - → `$ module reload`
  - → Find out in which category a module is:
    - → `$ module apropos <modulename>`

Module System (2/2)

```
$ module avail
------------ /usr/local_rwth/modules/modulefiles/linux/x86-64/DEVELOP ------------
cmake/2.8.5(default)   inteltbb/4.1(default)
cuda/40                intelvtune/XE2013U02(default)
cuda/41                likwid/system-default(default)
cuda/50(default)       nagfor/5.2
ddt/2.6                nagfor/5.3.1(default)
ddt/3.0(default)       openmpi/1.5.3
gcc/4.3                openmpi/1.6.1(default)
gcc/4.5                openmpi/1.6.1mt
gcc/4.6                openmpi/1.6.4
gcc/4.7                openmpi/1.6.4mt
...                                    ...
------------ /usr/local_rwth/modules/modulefiles/C      AL ------------
BETA      DEPRECATED GRAPHICS   MATH      TECHNICS
CHEMISTRY DEVELOP    LIBRARIES  MISC      UNITE
```

modules

Tools used during this workshop are placed in the **UNITE** category.

categories

Questions?