PC² User Meeting

- Technical Overview of the Noctua HPC System -

Jens Simon

PC² - Paderborn Center for Parallel Computing

December, 10th 2018

https://pc2.uni-paderborn.de

Noctua

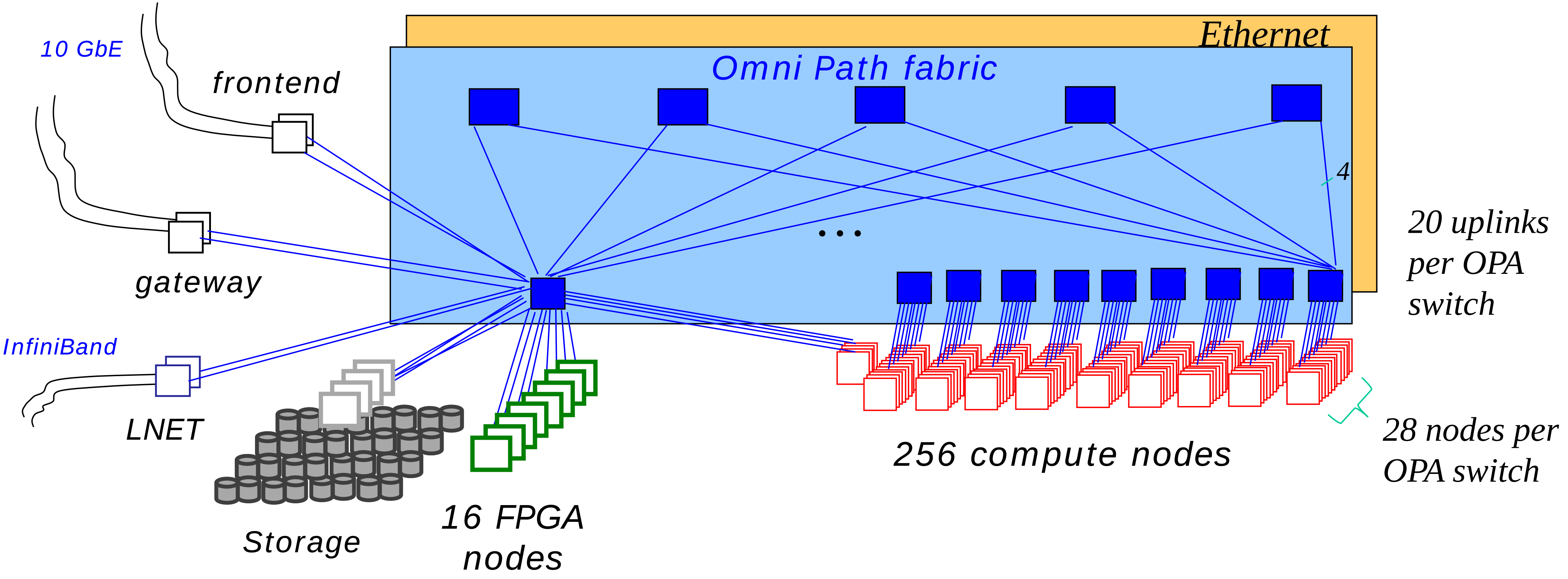
Frontside: Cold air intake

Backside: Cooled backdoors

Noctua: Cray CS 500 Storm

- 256 compute nodes
  - 2x Intel Xeon Gold 6148, 40 cores, 2.4 GHz
  - 192 GiB main memory
- 16 FPGA nodes
  - Intel Xeon 6148+6148F, 192 GiB
  - each with 2 Bittware Stratix 10
- Parallel file system
  - Lustre
  - 720 TB disk capacity
- Interconnect Intel Omni-Path
  - 100 Gbit/s network
  - Blocking factor 1:1,4

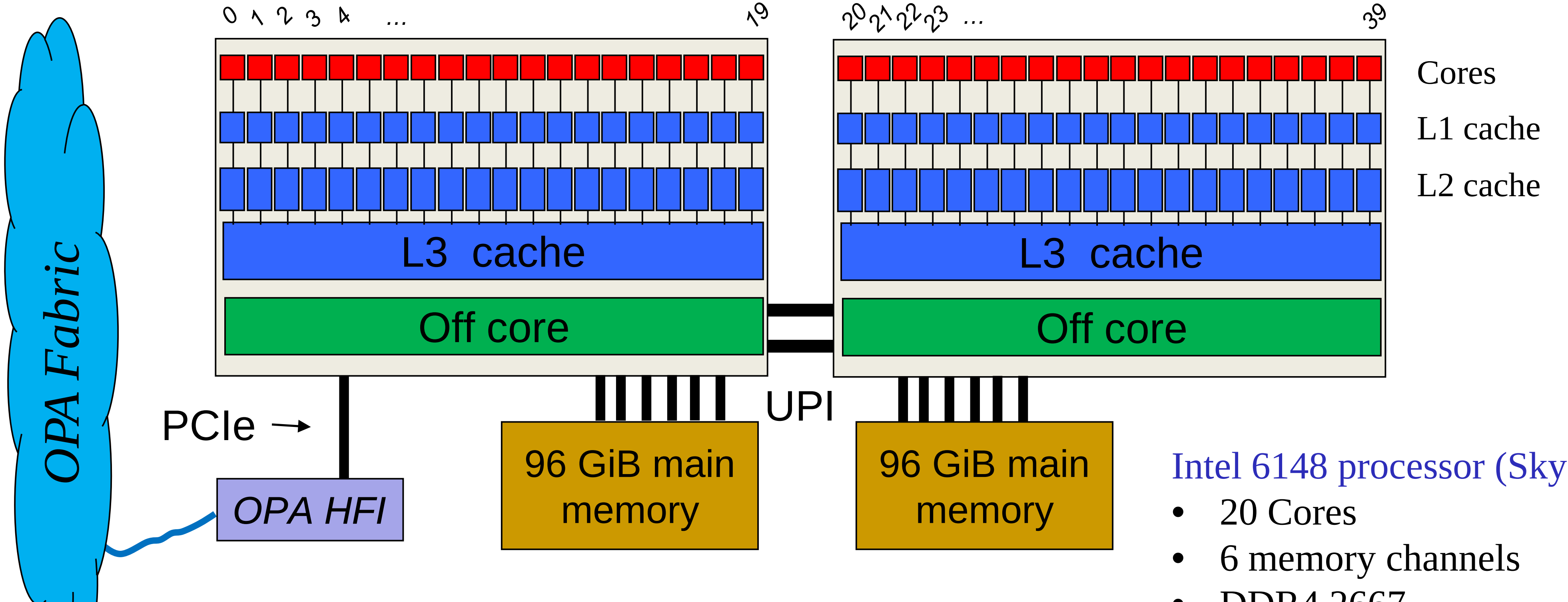
Noctua: System Architecture



Paderborn Center for Parallel Computing

Noctua: Compute Node

Intel Xeon Gold 6148 Node Architecture



272 nodes

Intel 6148 processor (Skylake)

- 20 Cores
- 6 memory channels
- DDR4 2667
- 10.4 GT/s UPI
- AVX-512 instructions
- …

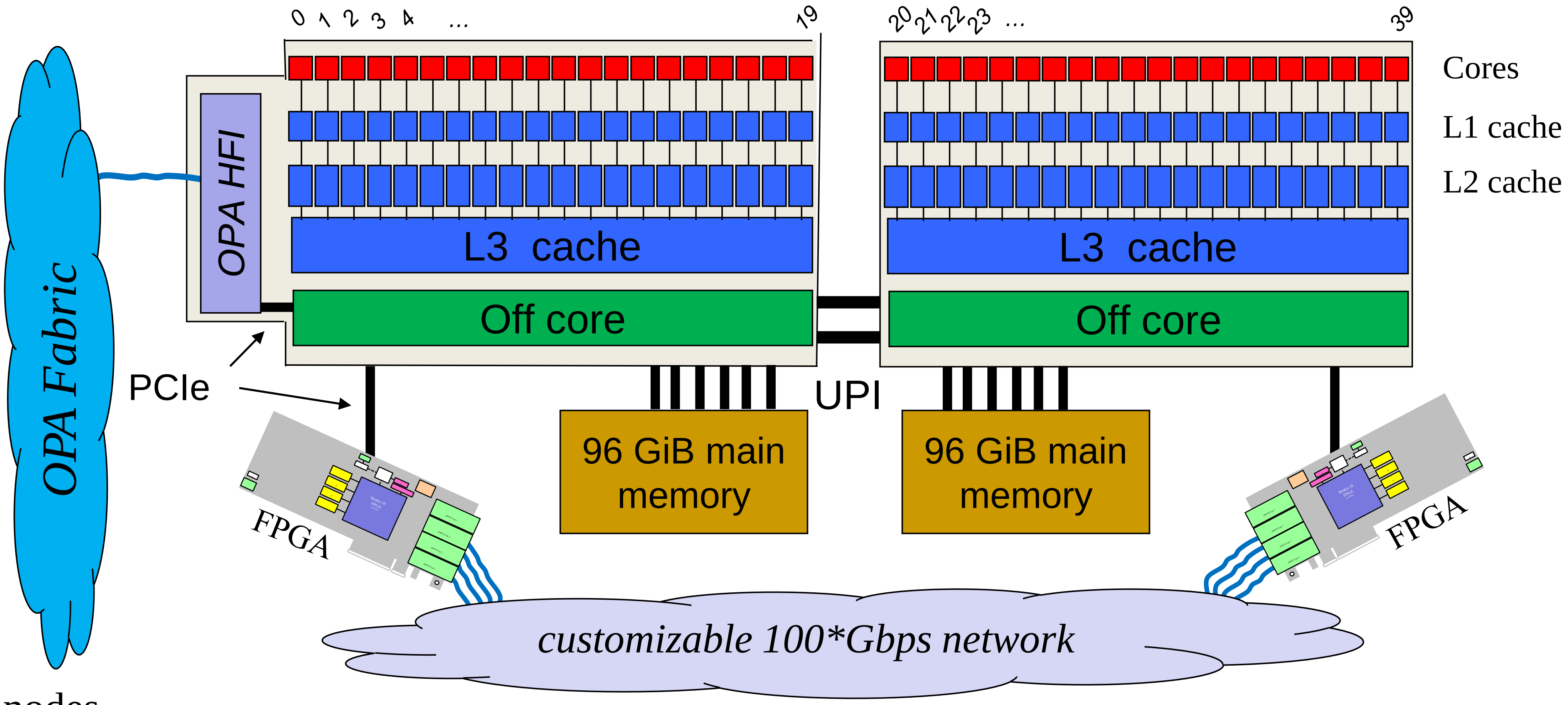
Noctua: FPGA Node

CPU with integrated Fabric

Fabric passive cable

Fabric through carrier card

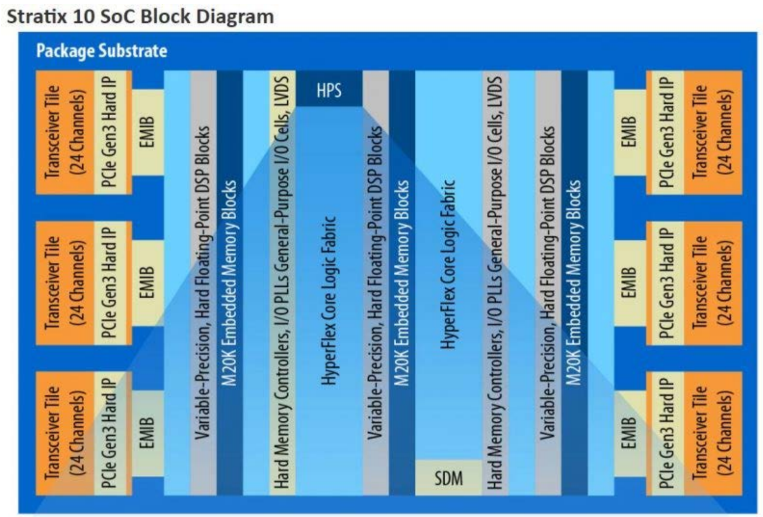
Intel Xeon Gold 6148 (F) Node Architecture



272 nodes

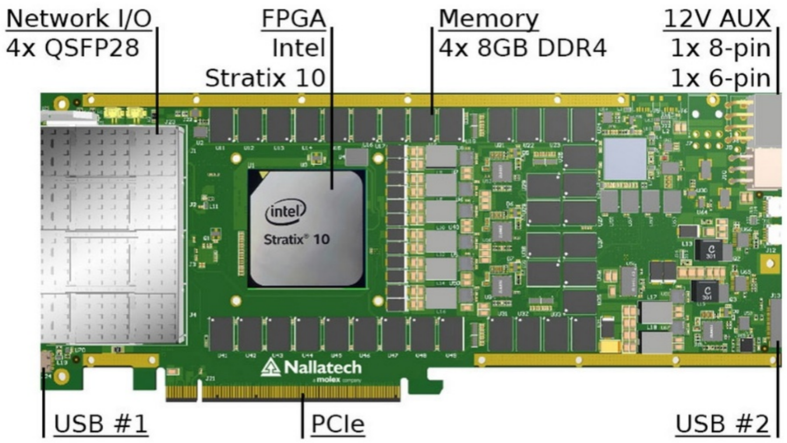
32 FPGAs

Bittware 520 N with Intel Stratix 10 FPGA



Stratix 10 SoC Block Diagram
Package Substrate
Transceiver Tile (24 Channels)
PCIe Gen3 Hard IP
EMIB
Variable-Precision, Hard Floating-Point DSP Blocks
M20K Embedded Memory Blocks
Hard Memory Controllers, I/O PLLs General-Purpose I/O Cells, LVDS
HPS
HyperFlex Core Logic Fabric
SDM

Source: Intel Corp.



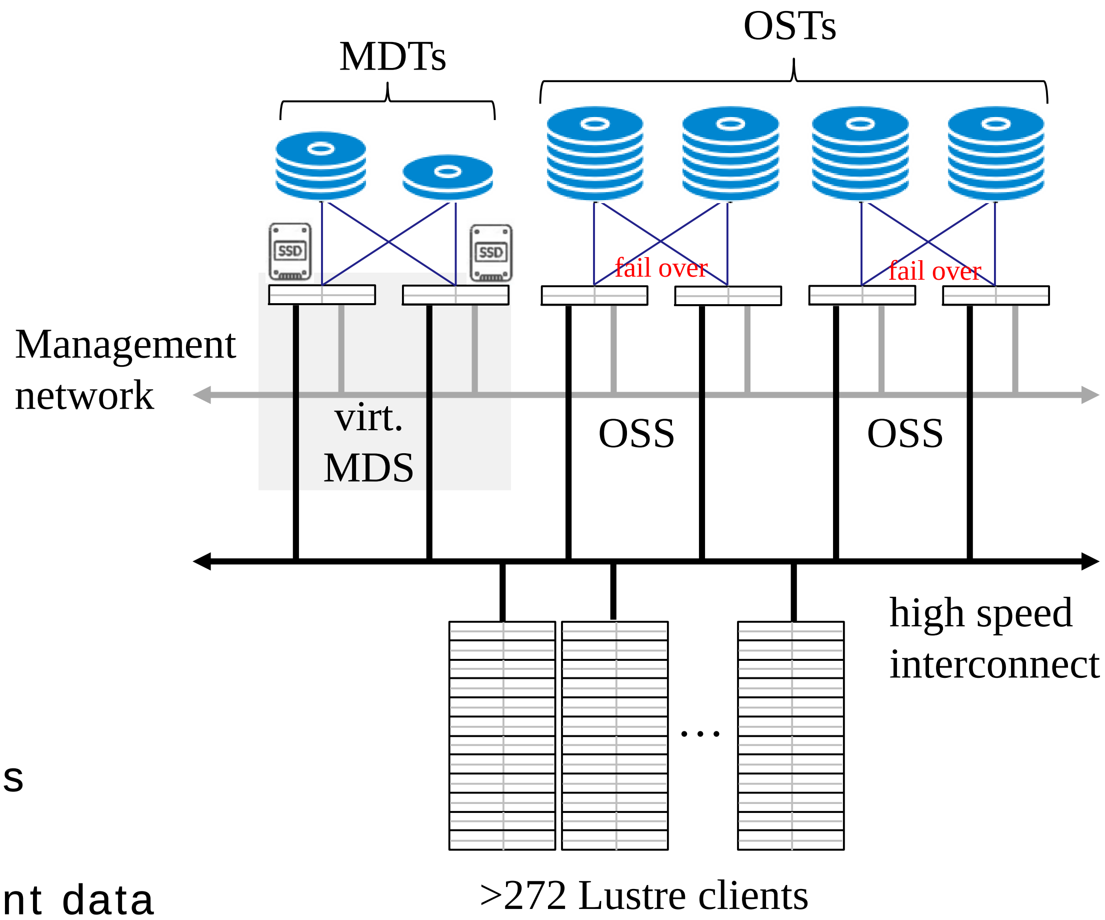
Source: Bittware, Inc.

Feature	Stratix 10 GX2800
Logical elements (LEs)	2,753,000
Adaptive logic modules (ALMs)	933,120
ALM registers	3,732,480
M20k memory blocks	11,721
M20k memory size (Mb)	229
variable precision DSP blocks	5,760
18 x 19 multipliers	11,520
Peak fixed point perf. (TMACS)	23.0
Peak FP SP perf. (TFLOPS)	9.2

FPGA network	
FPGA to FPGA latency	590 ns (w/o switch)
bidir. bandwidth	9.2 GiB/s (1 port)
bidir. bandwidth	36.8 GiB/s (4 ports)

ClusterStore with Lustre FS

- 720 TByte storage capacity
- Lustre file system with
  - one virtualized Management and Metadata Server (MDS) with 2 Metadata Targets (MDTs)
  - 4 Object Storage Targets (OSTs)
  - two SSDs for small files



Key points

- high performance through parallelism
  - best performance from multiple Clients writing to multiple OSTs
- achieve high bandwidth to/from a small number of files
  - used as a scratch file system
  - good match for scientific datasets and/or checkpoint data
- not designed to handle large numbers of small files
  - potential bottle necks at the MDS when files are opened
  - data will not be spread over multiple OSTs
  - not a good choice for program compilation

Noctua1 vs OCuLUS (node basis)

Metric	Noctua1 node	OCuLUS node	Difference
Processor	2 x Intel Gold 6148	2 x Intel E5-2670	AVX-512
# Cores	40	16	x 2.5
Frequency	2.4 GHz	2.6 GHz	
Main memory cap. [GiB]	192	64	x 3
Main memory bw [GiB/s]	184	71	x 2.6
Linpack perf. [TFlop/s]	2.1	0.31	x 6.8
SpecFP CPU 2006 rate	1,400	480	x 2.9
SpecINT CPU 2006 rate	1,950	624	x 3

Noctua1 vs. OCuLUS (total)

Metric	Noctua1	OCuLUS	Diff.
# nodes	272	616	- 56%
# cores	10.880	9.856	+ 10%
Total memory [TiB]	52.2	41.2	+ 27%
HP-Linpack [TFlop/s]	537	188.7	x 2.8
Accu. STREAM [GiB/s]	50,150	43,700	+ 15%
Accu. SpecFP2006 rate	381,000	295,700	+ 29%
Accu. SpecINT2006 rate	516,800	381.900	+ 35%
MPI network [Gbps]	100 (Omni Path)	40 (InfiniBand QDR)	x 2.5
blocking factor	1:1.4	1:2	
latency [µs]	1.24	1.9	- 35%
bandwidth [GByte/s]	24,5	7	x 3.5
Storage Capacity [TB]	720	500	+ 44%
max. bandwidth [GiB/s]	20	25	- 20%
Power consumption [kW]	164	230	- 29%

Project Membership

- Users must apply for HPC system usage (project application)
  - More than one user can be part of a project
  - One user can be part of different projects
  - Each PC² project is associated with a “UNIX group”

- LDAP groups are used for
  - System access, WLM–limits, accounting, self-service, Software ACLs

- Tight integration with IMT services to allow
  - Light-users (external, non UPB members)
  - Self service (Sniff accounts, Group membership, ...)
  - Filesystem access (/upb/departments/pc2)
  - Service-status information

Group Based Directories

- Scratch and permanent data has to be separated.
- Capacities are requested in project application and set quotas are enforced.
- Campus Storage (CST) is available on all HPC systems (Noctua, OCuLUS, etc.)
- HPC systems with a fast local parallel file system (PFS) (Noctua, OCuLUS)

Environment Variable	Points to	Purpose	Initial Quota	Backup
HOME	Absolute path to PC^2 wide home directory (CST)	permanent, small data, per user	5GB	Yes
PC2DATA	Absolute path to the PC2^2 wide group data directories (CST)	permanent, data, per group	requested on application	Yes
PC2PFS	Absolute path to group scratch directories on PFS of the HPC system	temporary, fast, scratch data, per group	requested on application	No
PC2SCRATCH	Absolute path to the PC2^2 wide group scratch directories	temporary, scratch data, per group	requested on application	No
PC2SW	Absolute path to the HPC software	pre-installed SW, read only	-	-
TMPDIR CCS_TMPDIR	Absolute path to the node local disks	temporary, created by WLM	Node specific	No

Noctua: System Access

- IMT user account and at least one approved PC² project is needed

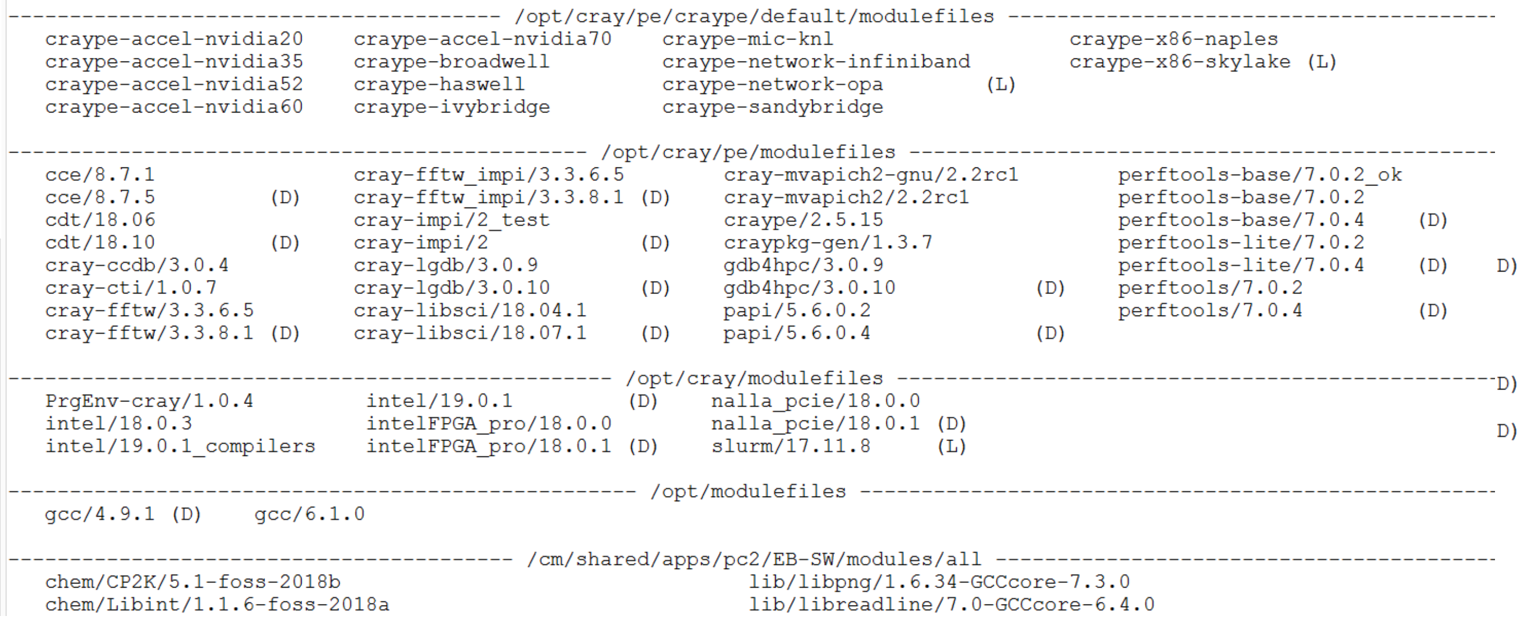
- Login (Jump) server
  - *fe.noctua.pc2.uni-paderborn.de*
- front ends accessible from login server
  - *ssh noctua*, or
  - *ssh noctua-last*

- On all Noctua nodes a standard environment is initially set (modulefiles)
  - *pc2fs*
    - File paths to the PC² file systems
  - *slurm*
    - Workload manger
  - *craype-x86-skylake* and *craype-network-opa*
    - Sets proper processor (Skylake) and MPI network (Omni Path) for Cray PE

Noctua: Modules

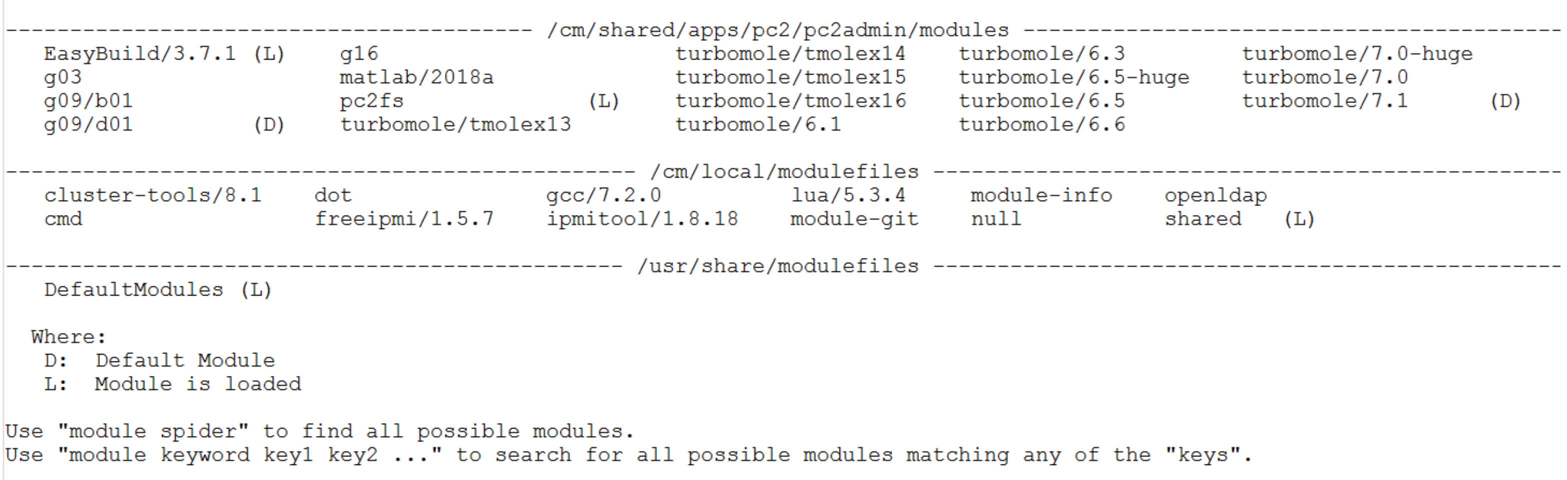
- Module system (Lua based)
  - Provides a flexible user environment
  - Modulefile contains information needed to configure the shell for an application
- EasyBuild is used
  - Manage scientific software on HPC systems
  - Fully autonomous builds with build logging, automatic dependency resolution, …
  - Lots of software packages are supported
    - https://easybuild.readthedocs.io/en/latest/version-specific/Supported_software.html
  - A hierarchical module naming scheme is used
    - chem, compiler, devel, lang, lib, math, mpi, numlib, system, toolchain, tools, vis, …
  - Users can create their own modules / applications in their directories

Noctua: Modulefiles



```
---------------------------------------- /opt/cray/pe/craype/default/modulefiles ----------------------------------------
   craype-accel-nvidia20    craype-accel-nvidia70    craype-mic-knl                    craype-x86-naples
   craype-accel-nvidia35    craype-broadwell         craype-network-infiniband         craype-x86-skylake  (L)
   craype-accel-nvidia52    craype-haswell           craype-network-opa         (L)
   craype-accel-nvidia60    craype-ivybridge         craype-sandybridge

----------------------------------------------- /opt/cray/pe/modulefiles -----------------------------------------------
   cce/8.7.1                cray-fftw_impi/3.3.6.5        cray-mvapich2-gnu/2.2rc1       perftools-base/7.0.2_ok
   cce/8.7.5         (D)    cray-fftw_impi/3.3.8.1  (D)   cray-mvapich2/2.2rc1           perftools-base/7.0.2
   cdt/18.06                cray-impi/2_test              craype/2.5.15                  perftools-base/7.0.4    (D)
   cdt/18.10         (D)    cray-impi/2             (D)   craypkg-gen/1.3.7              perftools-lite/7.0.2
   cray-ccdb/3.0.4          cray-lgdb/3.0.9               gdb4hpc/3.0.9                  perftools-lite/7.0.4    (D)    D)
   cray-cti/1.0.7           cray-lgdb/3.0.10        (D)   gdb4hpc/3.0.10          (D)    perftools/7.0.2
   cray-fftw/3.3.6.5        cray-libsci/18.04.1           papi/5.6.0.2                   perftools/7.0.4         (D)
   cray-fftw/3.3.8.1 (D)    cray-libsci/18.07.1     (D)   papi/5.6.0.4            (D)

------------------------------------------------- /opt/cray/modulefiles -------------------------------------------------D)
   PrgEnv-cray/1.0.4         intel/19.0.1          (D)    nalla_pcie/18.0.0
   intel/18.0.3              intelFPGA_pro/18.0.0         nalla_pcie/18.0.1  (D)                                        D)
   intel/19.0.1_compilers    intelFPGA_pro/18.0.1  (D)    slurm/17.11.8      (L)

--------------------------------------------------- /opt/modulefiles ---------------------------------------------------
   gcc/4.9.1  (D)    gcc/6.1.0

----------------------------------------- /cm/shared/apps/pc2/EB-SW/modules/all -----------------------------------------
   chem/CP2K/5.1-foss-2018b                                 lib/libpng/1.6.34-GCCcore-7.3.0
   chem/Libint/1.1.6-foss-2018a                             lib/libreadline/7.0-GCCcore-6.4.0
```

...



```
------------------------------------------ /cm/shared/apps/pc2/pc2admin/modules ------------------------------------------
   EasyBuild/3.7.1  (L)    g16                        turbomole/tmolex14     turbomole/6.3          turbomole/7.0-huge
   g03                     matlab/2018a               turbomole/tmolex15     turbomole/6.5-huge     turbomole/7.0
   g09/b01                 pc2fs               (L)    turbomole/tmolex16     turbomole/6.5          turbomole/7.1       (D)
   g09/d01          (D)    turbomole/tmolex13         turbomole/6.1          turbomole/6.6

------------------------------------------------------ /cm/local/modulefiles ------------------------------------------------------
   cluster-tools/8.1     dot                  gcc/7.2.0           lua/5.3.4     module-info     openldap
   cmd                   freeipmi/1.5.7       ipmitool/1.8.18     module-git    null            shared    (L)

-------------------------------------------------- /usr/share/modulefiles --------------------------------------------------
   DefaultModules  (L)

  Where:
   D:  Default Module
   L:  Module is loaded

Use "module spider" to find all possible modules.
Use "module keyword key1 key2 ..." to search for all possible modules matching any of the "keys".
```

Noctua: Program Development Environments

- Intel Parallel Studio Cluster Edition (version 19.0.1)
  - C/C++ and Fortran
  - Python
  - Intel Math Kernel Libraries (MKL)
  - Intel MPI
  - Intel Data Analytics Acceleration Library, Integrated Performance Primitives, Threading Building Blocks
  - Intel VTune, Advisor, Inspector, Trace Analyzer and Collector
- Cray Programming Environment (version 2.5.15)
  - Cray Compiler Environment “CCE”
    - C/C++ and Fortran 2008
    - OpenMP 4.1, MPI 2.2, UPC 1.2, OpenACC 2.0, LibSci, LibSci_ACC
  - Cray Performance Measurement, Analysis, and Porting Tools
    - Performance and Analysis Tool CrayPAT
    - Visualization Tool Cray Apprentice2
    - Porting Tool Cray Reveal
- Intel FPGA SDK for OpenCL (version 18.0.1)
- and lots of GNU tools

Cray Performance Tools

- CrayPAT profiles executables
  - Timing and hardware performance counter measurements
  - Collect and show program top time consumers and bottlenecks
  - Automatic generation of observations and suggestions
  - Data collection and presentation of computation, communication, I/O, and memory statistics
  - CrayPAT lite is a simplified, easy-to-use version of CrayPAT
- Visualization of performance data with Cray Apprentice2
  - Reports and graphical formats
  - GUI
  - Runs on Windows, MacOS, and Linux using the platform-independent data files
- Code-restructuring assistant Reveal
  - Helps developers to add additional levels of parallelism
  - Assists with parallelizing more complicated loops
  - Combining performance statistics and program source code

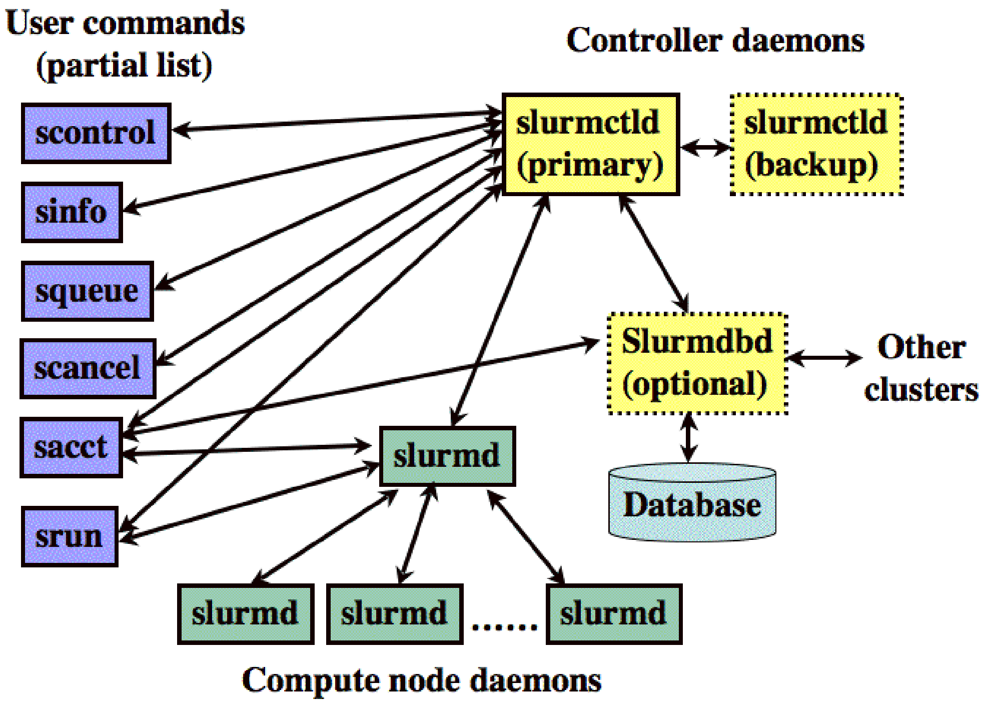
FPGA Development Tools

- SDK, development and emulation on all nodes
  - Module *intelFPGA_pro*, current version 18.0.1
  - Wrap performance critical code into OpenCL kernel
    - *aoc* compiles code into hardware structure
  - Perform software emulation to ensure correctness
  - Generated reports provide insights on performance and resources
  - Time-consuming hardware generation only as last step

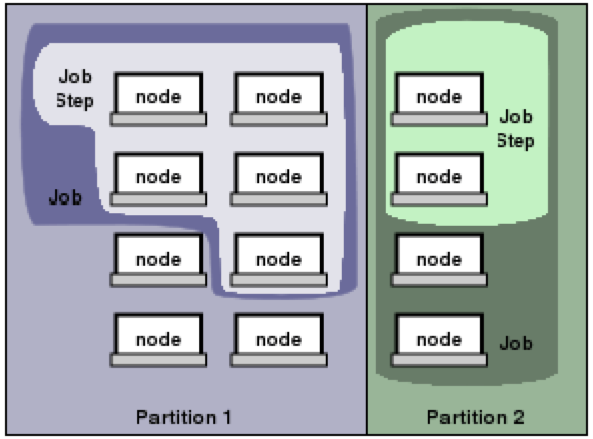
- Hardware execution on FPGA-nodes
  - `srun --partition=fpga --constraint=18.0.1`
  - Drivers and infrastructure provided by *constraint* for all installed versions

SLURM

- User commands
  - sacct, salloc, sattach, sbatch, sbcast, scancel, scontrol, sinfo, smap, squeue, srun, strigger, sview
- Managed entities
  - Jobs (allocation of resource assigned to a user for a specified amount of time)
  - Job steps (sets of parallel tasks within a job)
  - Resources are nodes, processors, memory, …
  - Nodes are logically organized into possibly overlapping partitions (aka queues)
- lots of documention available https://slurm.schedmd.com/



Source: SchedMD LLC



Queue Configuration

Name	Max. Nodes*	Max. Runtime*
short	2	30m
test	50	30m
batch	100	24h
long	50	21d
fpga	16	2h
all	272	12h

* Initial settings

Projects are restricted in maximum number of nodes and to certain queues.

The queue configuration is subject of change.

Applications / Libraries / Tools

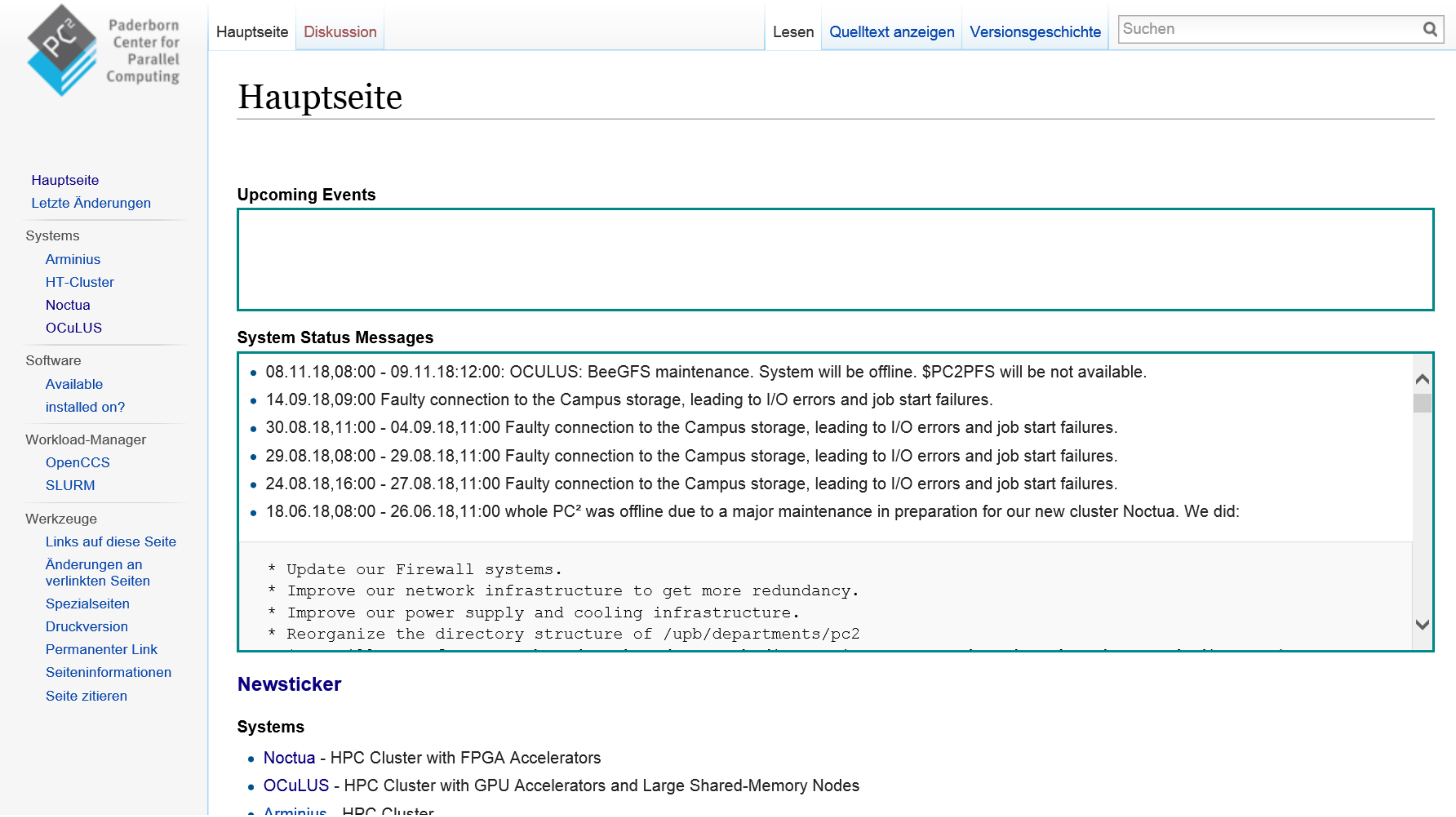
CP2K
Gaussian
Turbomole

IntelMPI
OpenMPI
FFTW
MKL
OpenBLAS
ScaLapack

Easybuild
Python3
Valgrind
gdb4hpc

… and further applications on request. Please contact us.

https://wikis.uni-paderborn.de/pc2doc/Hauptseite



Hauptseite

Upcoming Events

System Status Messages

- 08.11.18,08:00 - 09.11.18:12:00: OCULUS: BeeGFS maintenance. System will be offline. $PC2PFS will be not available.
- 14.09.18,09:00 Faulty connection to the Campus storage, leading to I/O errors and job start failures.
- 30.08.18,11:00 - 04.09.18,11:00 Faulty connection to the Campus storage, leading to I/O errors and job start failures.
- 29.08.18,08:00 - 29.08.18,11:00 Faulty connection to the Campus storage, leading to I/O errors and job start failures.
- 24.08.18,16:00 - 27.08.18,11:00 Faulty connection to the Campus storage, leading to I/O errors and job start failures.
- 18.06.18,08:00 - 26.06.18,11:00 whole PC² was offline due to a major maintenance in preparation for our new cluster Noctua. We did:

```
 * Update our Firewall systems.
 * Improve our network infrastructure to get more redundancy.
 * Improve our power supply and cooling infrastructure.
 * Reorganize the directory structure of /upb/departments/pc2
```

Newsticker

Systems

- Noctua - HPC Cluster with FPGA Accelerators
- OCuLUS - HPC Cluster with GPU Accelerators and Large Shared-Memory Nodes
- Arminius - HPC Cluster