JÜLICH
FORSCHUNGSZENTRUM

JuRoPA – Juropa-JSC / HPC-FF

An Overview

21 May 2012 | Ulrich Detert

JuRoPA

- ***JuRoPA***
  - *Jülich Research on Petaflop Architectures*
  - *Bull, Sun, ParTec, Intel, Mellanox, Novell, FZJ*
  - ***JUROPA (Juropa-JSC)***
    - *FZJ production system*
      - *NIC and VSR projects*
      - *Commercial customers*
    - *PRACE Tier 1 system*

  - ***HPC-FF***
    - *HPC-FF: High Performance Computing For Fusion Dedicated to European Fusion Research Community*

Juropa Components (1)

- ***JUROPA***
  - *Sun Constellation System*
  - *Infiniband QDR*
  - *2208 compute nodes:*
    - *2 Intel Nehalem-EP quad-core processors (Xeon X5570)*
      - *2.93 GHz*
    - *24 GB memory (DDR3, 1066 MHz)*
    - *IB QDR HCA (QNEM - Network Express Module)*
  - *17664 cores*
  - *207 TF peak*

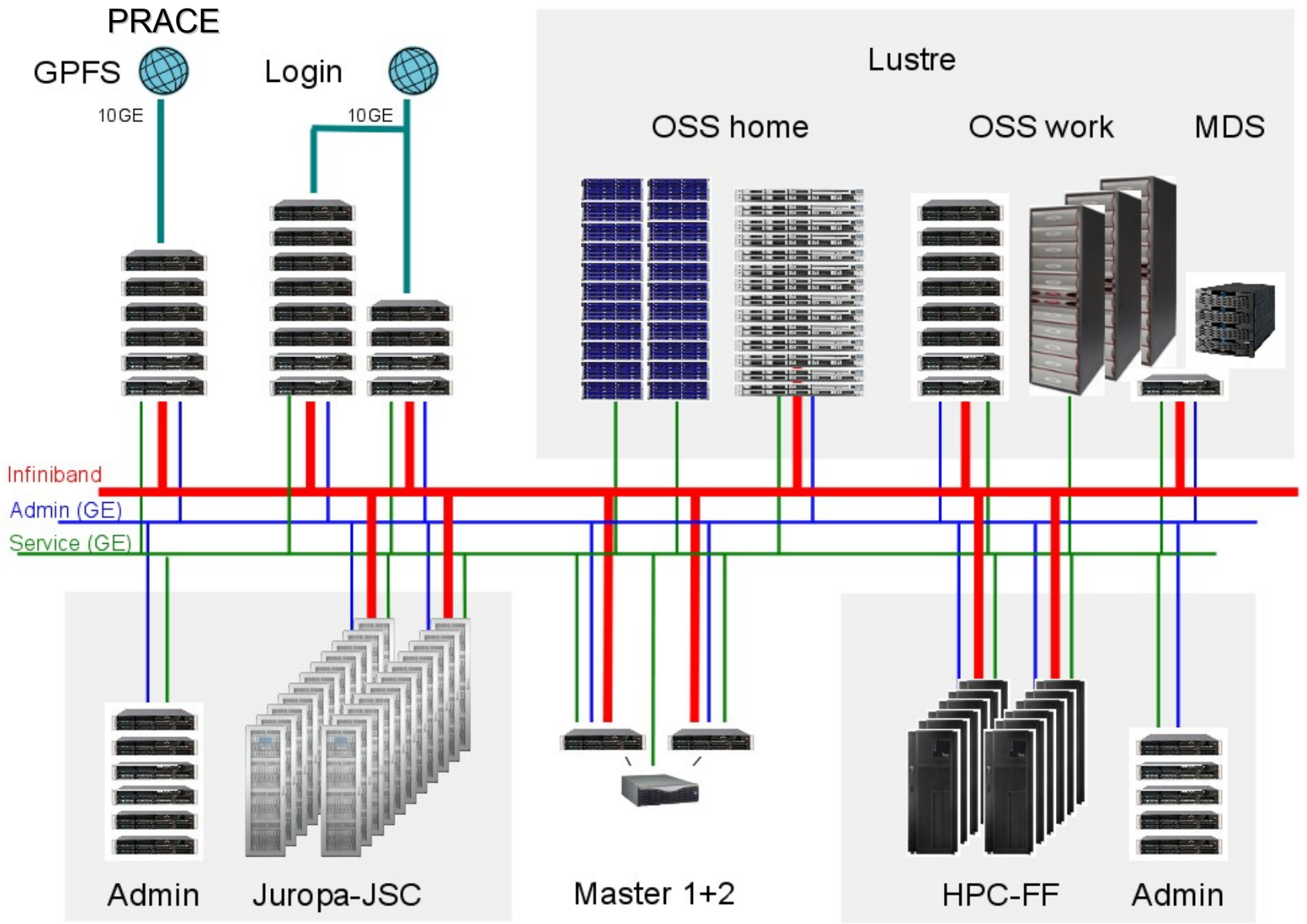
Juropa Components (2)

- ***HPC-FF***
  - *Bull NovaScale R422-E2*
  - *Infiniband QDR*
  - *1080 compute nodes*
    - *2 Intel Nehalem-EP quad-core processors (Xeon X5570)*
      - *2.93 GHz*
    - *24 GB memory (DDR3, 1066 MHz)*
    - *Infiniband Mellanox ConnectX QDR HCA*
  - *8640 cores*
  - *101 TF peak*

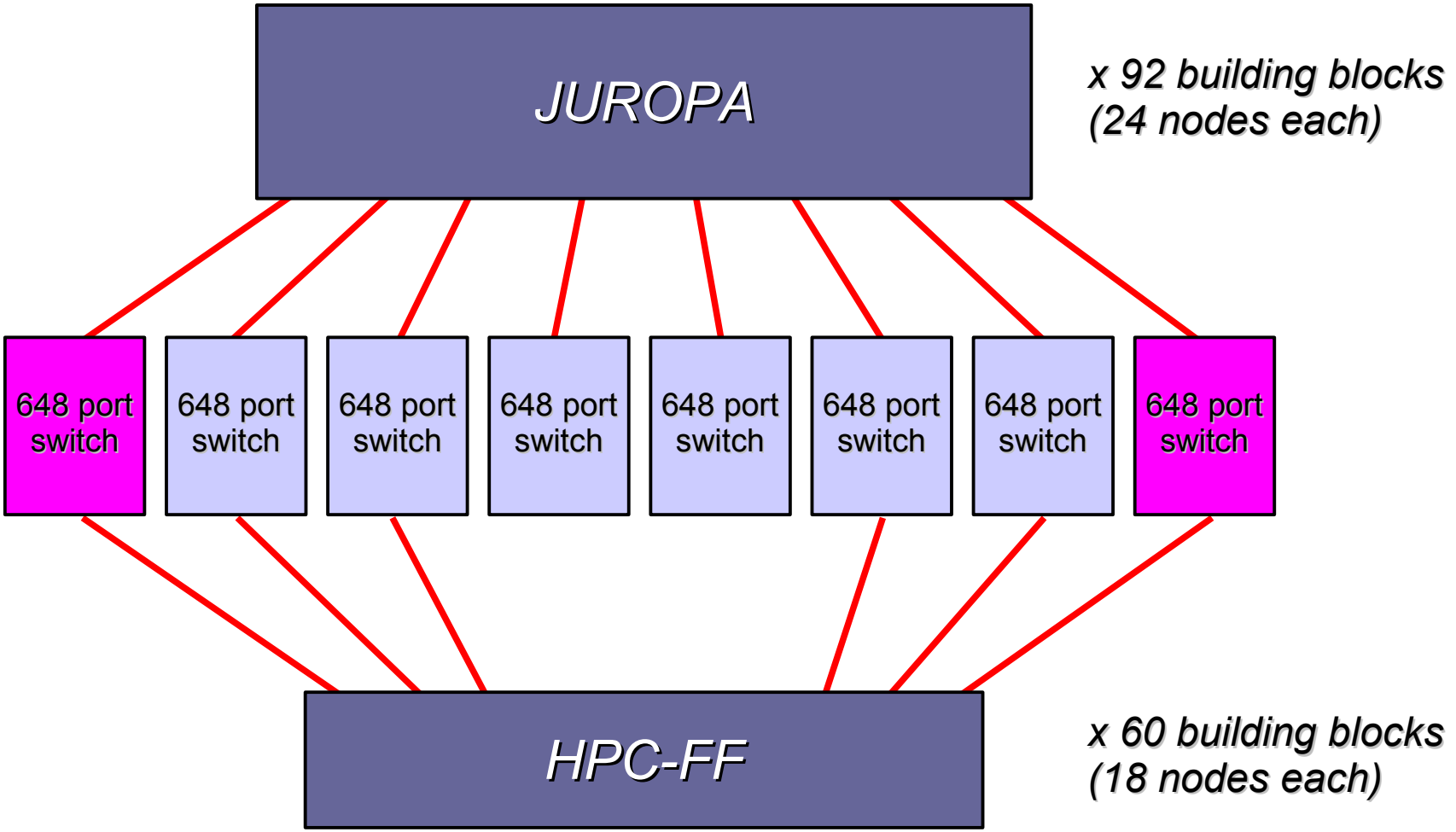
Juropa Components (3)

- ***Lustre Storage Pool***
  - *4 Meta Data Servers (MDS)*
    - *Bull NovaScale R423-E2 (Nehalem-EP 4-core/Westmere 6-core)*
    - *100 TB for meta data home and work (EMC² CX4-240)*
  - *14 Object Storage Servers (OSS) (**home**)*
    - *Sun Fire X4170 Server*
    - *500 TB user data*
  - *8 Object Storage Servers (OSS) (**home**)*
    - *Bull NovaScale R423-E2*
    - *500 TB user data*
  - *8 Object Storage Servers (OSS) (**work**)*
    - *Bull NovaScale R423-E2*
    - *800 TB user data*
  - *Aggregated data rate ~40 GB/s*

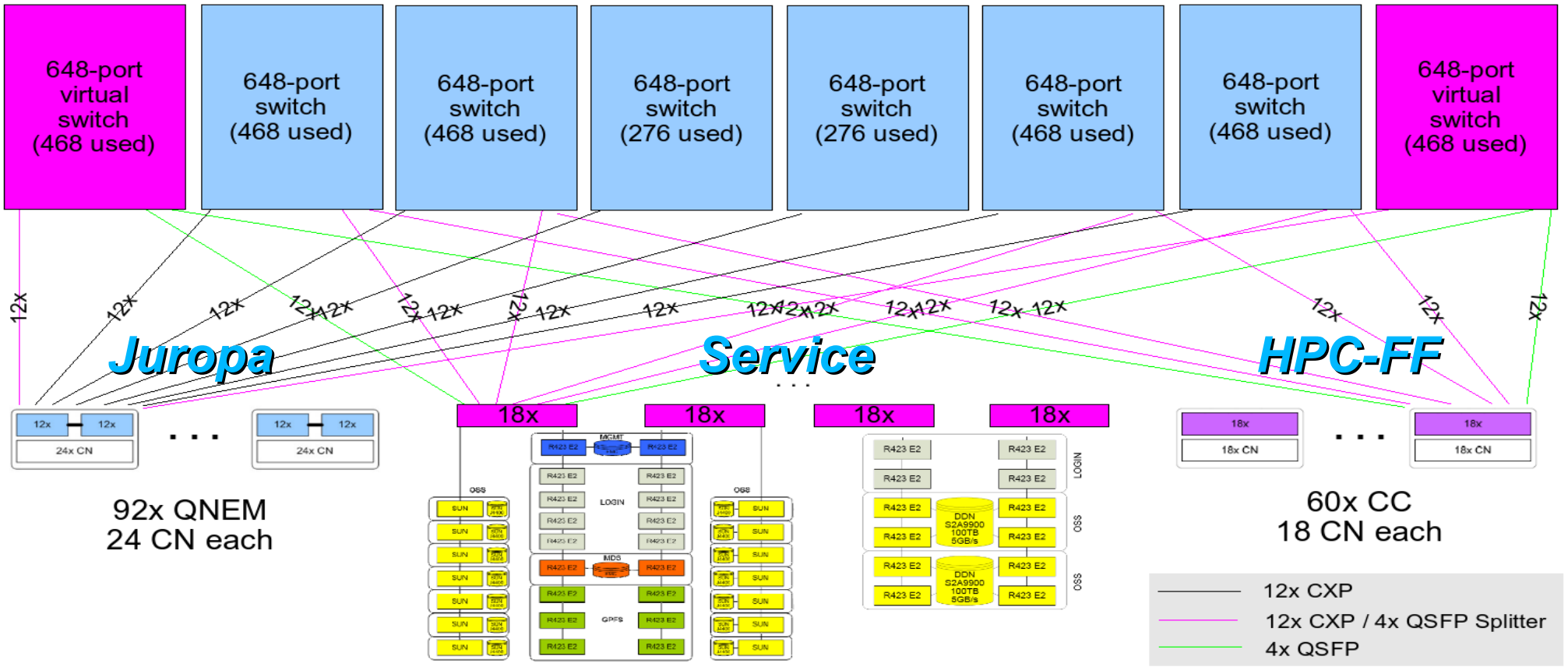
Juropa Architecture



Infiniband Topology (Fat Tree)



Infiniband Topology (Fat Tree)



- *23 x 4 QNEM modules, 24 ports each*
- *6 x M9 switches, 648 ports max. each, 468/276 links used*
- *Mellanox MTS3600 switches (Shark), 36 ports, for service nodes*
- *4 Compute Sets (CS) with 15 Compute Cells (CC) each*
- *CC with 18 Compute Nodes (CN) and 1 Mellanox MTS3600 (Shark) switch each*
- *Virtual 648-port switches constructed from 54x/44x Mellanox MTS3600*

Software (1)

- ***Operating System***
  - *SUSE SLES 11 SP1*
- ***Cluster Management***
  - *ParaStation*
  - *GridMonitor, Jumpmon*
- ***Batch System***
  - ***Torque** Resource Manager (start jobs, return output etc.)*
  - ***Moab** Workload Manager (priorities, accounting, job chains)*
    - *User command line interface (job start, status, cancel etc.)*
- ***Compiler***
  - *Intel Professional Fortran, C/C++*

Software (2)

- ***Libraries***
  - *Intel Math Kernel Library (**MKL**)*
    - *BLAS, LAPACK, ScaLAPACK1, Sparse Solvers, Fast Fourier Transforms, Vector Math etc.*
    - *Highly optimized for Intel CPU architecture*
  - */usr/local*
    - *~ 90 packages from adf, amber … to wsmp, zlib*
- ***MPI - Message Passing Interface***
  - *ParTec MPI (based on MPICH2)*
- ***OpenMP - Memory-Parallel Multi-Threading***
  - *Intel*
- ***Unicore, PRACE (gsissh, gridftp)***

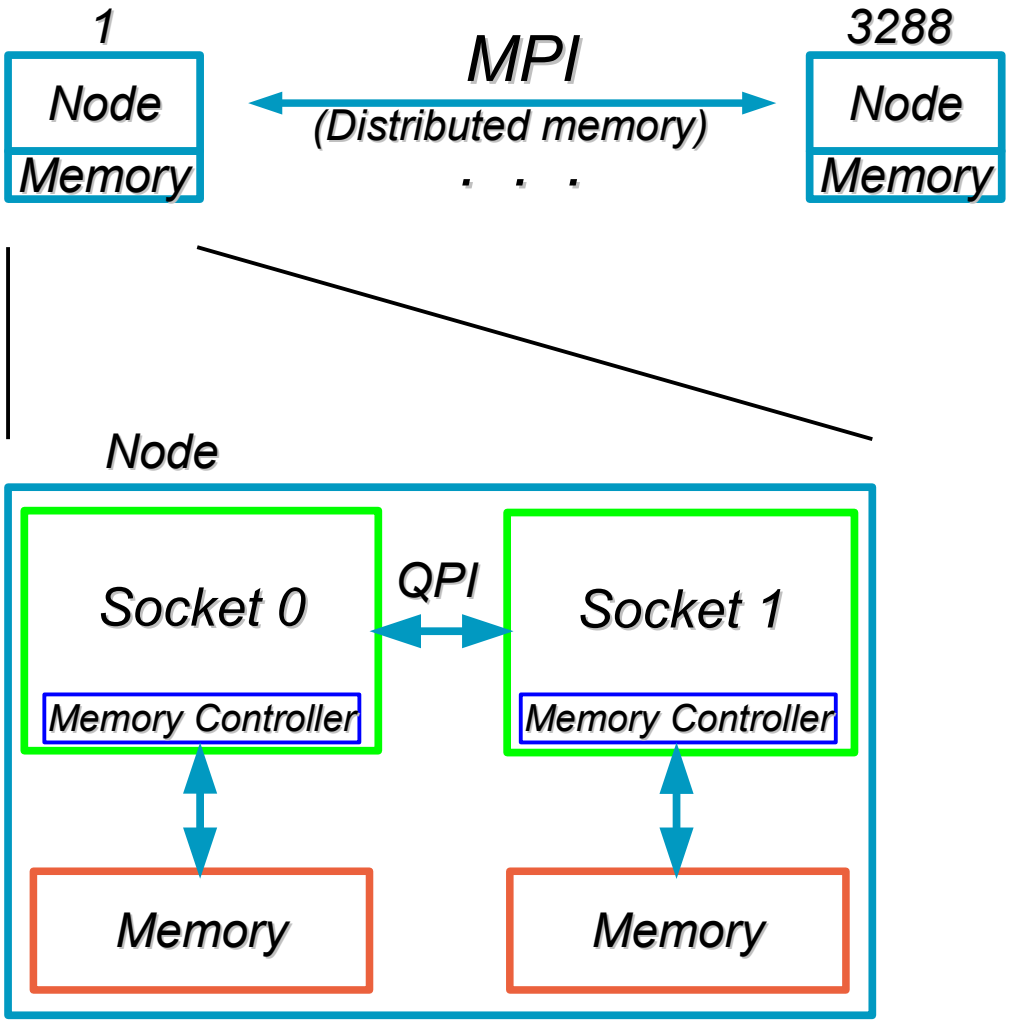
Modules

- ***Modules allow to switch between versions of a specific software or library***
  - `module avail`
    - *shows available modules*
  - `module list`
    - *lists loaded modules*

```
Currently loaded modulefiles:
1) parastation/mpi2-intel-5.0.26-1
2) mkl/10.2.5.035
3) intel/11.1.072
```

  - `module load | unload`
    - *load / unload a module*
  - `module help`
    - *list usage information*

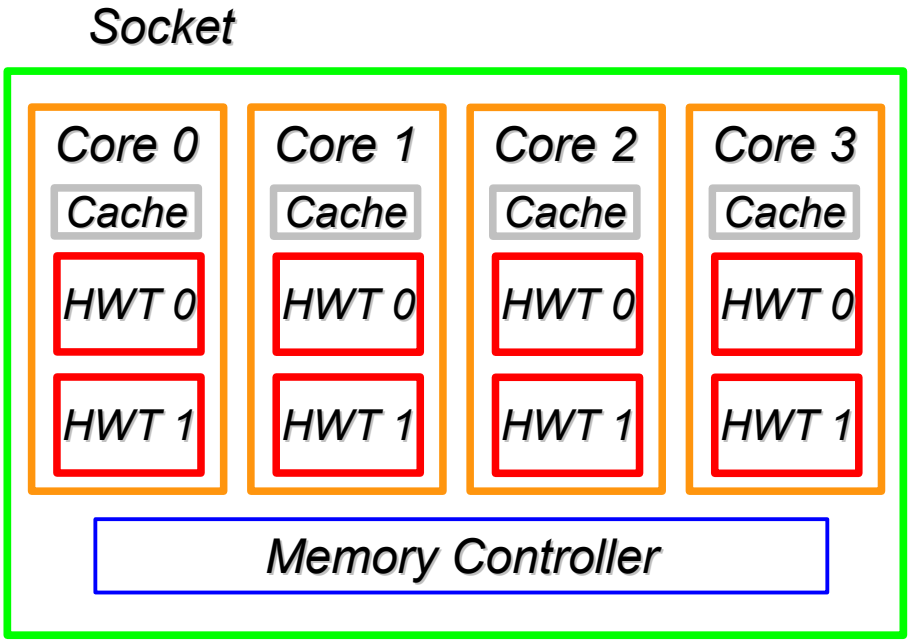
What is Simultaneous Multi-Threading (SMT)?



***Compute nodes**: mostly independent of each other, local memory, shared network access*

***Socket**: full-featured quad-core Nehalem-EP CPU with its own memory controller and Quickpath Interconnect (QPI)*

What is SMT?



***Core**: full-featured processor with its own register set, functional units and caches. Memory access is shared among cores.*

***HWT - Hardware Thread**: Collection of registers and functional units ("virtual core") that are replicated in each core. Non-replicated units like **caches** are shared among HWTs.*

- ***Process (Task):***
  - *Computational entity that possesses its own copy of program code and data.*
- ***Thread:***
  - *Computational entity that shares program code and data with other threads. Threads may have (a limited amount of) non-shared, local data.*

Node Types (1)

- ***Compute nodes***
  - *Intended for batch jobs*
    - *includes „interactive“ batch (JSC: 32 / HPC-FF: 16 nodes max.)*
      - *charged for connect time*
  - *No direct login on compute nodes (except interactive batch)*
  - *Exclusive usage by one user/job (no node sharing)*
    - *smallest reservation unit is one node (= 8 cores / 16 with SMT)*
    - *charged for wall-clock time*
    - *„unlimited“ access to existing resources*
      - *~ 22 GB memory, 8 processors (16 with SMT), wallclock time limit 24 h*

Compute Nodes - Available Memory

- ***Communication Scalability***
  - *Memory consumption for static IB communication buffers depends on the number of communicating tasks*

    Example:

    ```
       Default requirement per connection = 0.5 MB
    => An 8-core node that connects to 8 x 512 tasks
       consumes 16 GB just for buffers
    ```

  - *Solutions*
    - *Smaller / less buffers:*

      ```
      PSP_OPENIB_SENDQ_SIZE=3..16 (default: 16)
      PSP_OPENIB_RECVQ_SIZE=3..16 (default: 16)
      ```

    - *Buffer allocation „on-demand“:*

      ```
      export PSP_ONDEMAND=1
      ```

      or

      ```
      mpiexec --ondemand ...
      ```

Node Types (2)

- ***JSC Login nodes (juropa, juropa01 .. 07)***
  - *Access for non-HPC-FF users only*
  - *Intended for interactive work*
    - *program development (edit, compile, test)*
    - *pre and postprocessing*
    - *access to home filesystem and work (Lustre)*
    - *no production jobs here! (cpu time limit 30 min.)*
      - *use command `ulimit -Sa` or `ulimit -Ha` to display limits*
- ***HPC-FF Login nodes (hpcff, hpcff01 .. 03)***
  - *Access for HPC-FF users only*
  - *Same functionality as JSC Login nodes*

Node Types (3)

- ***GPFS nodes (juropagpfs, juropagpfs01 .. 05)***
  - *Access for JSC **and** HPC-FF users*
  - *Access to GPFS file systems (mounted also on JUGENE)*
  - *Intended for data manipulation*
    - *copy data to and from GPFS/Lustre*
    - *restore data from TSM backup*
    - *import/export data to/from external sources*
    - *same limits as on Login nodes (except **CPU limit: 360 min**)*
- ***GPFS nodes 04 and 05 for „big-memory“ requests:***
  - *192 GB memory*

Interactive performance might be degraded on GPFS nodes due to heavy data traffic.
***Recommendation:** Use Login nodes, if GPFS file system, large memory, higher CPU time limit or connection to PRACE network is not needed.*

Accessing the System

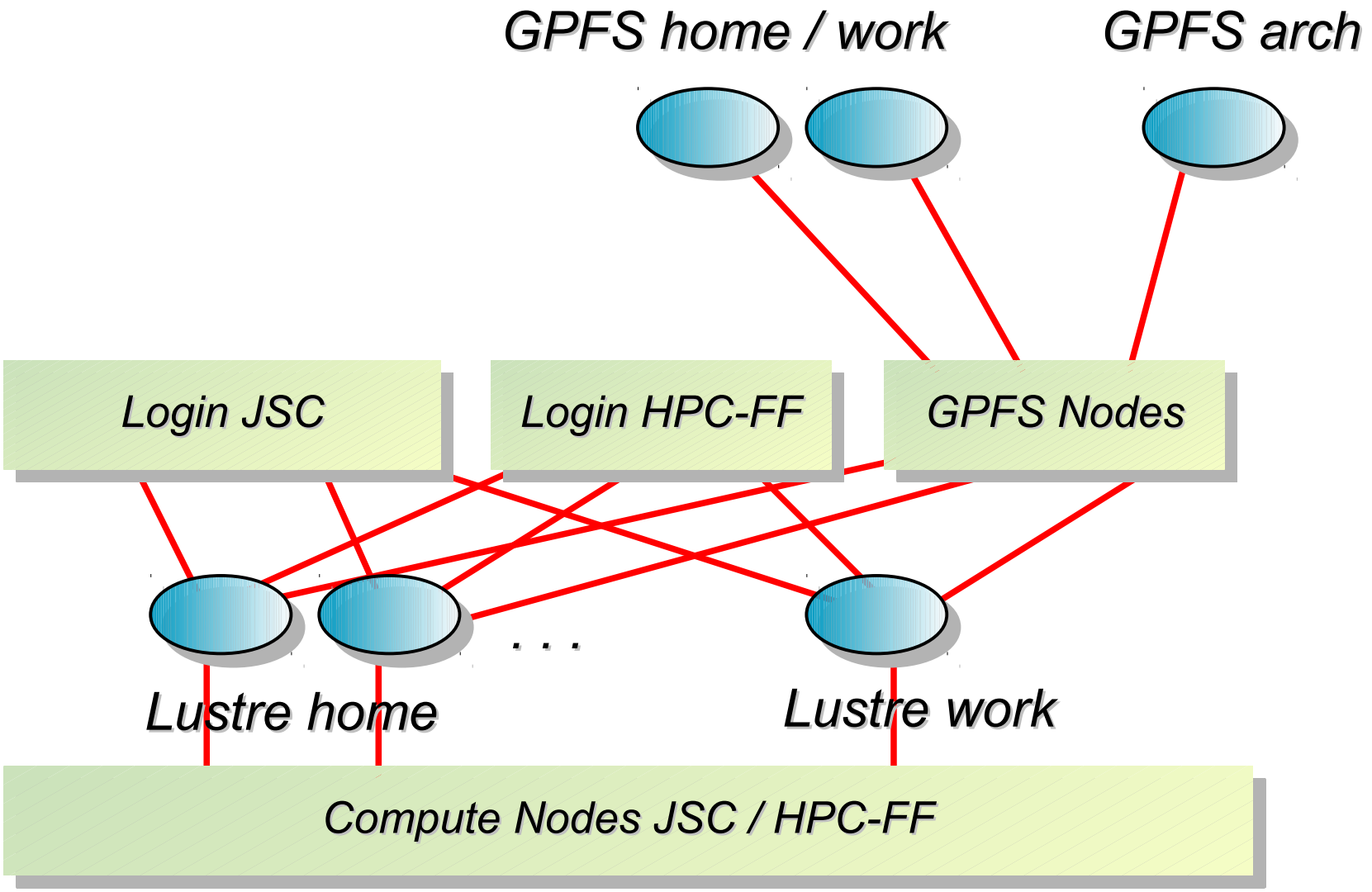
- ***Login nodes JSC***
  - `ssh [-X] <user>@juropa.fz-juelich.de`
    - *Login nodes 01..07 are selected „round-robin“*
    - *Hostnames:* `jj28l01 .. 07`
  - `ssh [-X] <user>@juropa01.fz-juelich.de`
    - *Access specific node juropa01*
- ***Login nodes HPC-FF***
  - `ssh [-X] <user>@hpcff.fz-juelich.de`
    - *Login nodes 01..03 are selected „round-robin“*
    - *Hostnames:* `jf29l01 .. 03`
  - `ssh [-X] <user>@hpcff01.fz-juelich.de`
    - *Access specific node hpcff01*

Accessing the System

- ***GPFS nodes***
  - `ssh [-X] <user>@juropagpfs.fz-juelich.de`
    - *GPFS nodes 01..05 are selected „round-robin“*
    - *Hostnames:* `jj28g01 .. 05`
  - `ssh [-X] <user>@juropagpfs01.fz-juelich.de`

- ***Access to all nodes requires SSH key***
  - *To be provided when applying for an account (JSC Dispatch)*
  - *Do provide a passphrase for your ssh key!*
  - *Password access is **not** possible!*

File Systems Overview



File Systems (1)

- ***Lustre***
  - *Mounted on Login, GPFS and Compute nodes*
    - *$**WORK** = $LUSTREWORK = `/lustre/jwork`*
      - *800 TB, shared between JSC and HPC-FF*
      - *group quota: 3 TB, 2 million files*
      - *no backup*
      - *files older than 28 days will be deleted*
      - *recommended for large **temporary** files and **high performance requirements***
    - *$**HOME** = $LUSTREHOME = `/lustre/jhome1 .. 24`*
      - *from 29 to 62 TB per file system, distributed among user groups*
      - *group quota: 3 TB, 2 million files*
      - *daily backup*
      - *recommended for **permanent** program data with **low performance requirements** (e.g. program sources, input files, configuration data)*

File Systems (2)

- ***GPFS***
  - *Mounted on GPFS nodes only*
  - *Shared with JUGENE, mounted from JUST file server*
  - *Only available with valid JUGENE user ID*
    - *$**GPFSWORK** = `/gpfs/work`*
      - *2.1 PB, mounted from JUST file server*
      - *modest performance (~ 200 - 300 MB/s)*
      - *details on sizes, quota, backup etc. => tomorrow*
    - *$**GPFSHOME** = `/gpfs/homea, .. homec`*
      - *mounted JUGENE home file systems*
      - *details on sizes, quota, backup etc. => tomorrow*
    - *$**GPFSARCH** = `/gpfs/arch, ..1, ..2`*
      - *automatic data migration to/from tape library*

Backup

- ***Backup is done for the Lustre home file systems, GPFS/home and GPFS/arch***
  - *Daily backup with TSM*
  - *Restore of user data can only be done on the GPFS nodes:*
    `ssh -X <user>@juropagpfs.fz-juelich.de`
    `adsmback`
    Select ***`home, arch`*** *or* ***`gpfshome`***
  - *This opens a panel for interactive restore*
    Select => Restore
    => File Level
    Choose your files/directories to restore
    => Restore

Further Information

- ***Regular preventive maintenance on Thursdays***
  - *See „Message of today“ at login*
- ***Juropa and HPC-FF on-line documentation***
  - *http://www.fz-juelich.de/ias/jsc/juropa/*
- ***User support at FZJ***
  - *sc@fz-juelich.de*
  - *Phone: 02461 61-2828*
- ***HPC-FF person of contact***
  - *Alexander Schnurpfeil: sc@fz-juelich.de*