JUWELS BOOSTER

SYSTEM DESIGN INFO FOR ESM COMMUNITY

4. FEBRUARY I JUWELS BOOSTER PROJECT TEAM

Slides:
http://bit.ly/booster-esm-uf

ESM Advanced Earth System Modelling Capacity

JÜLICH Forschungszentrum

PROJECT TEAM

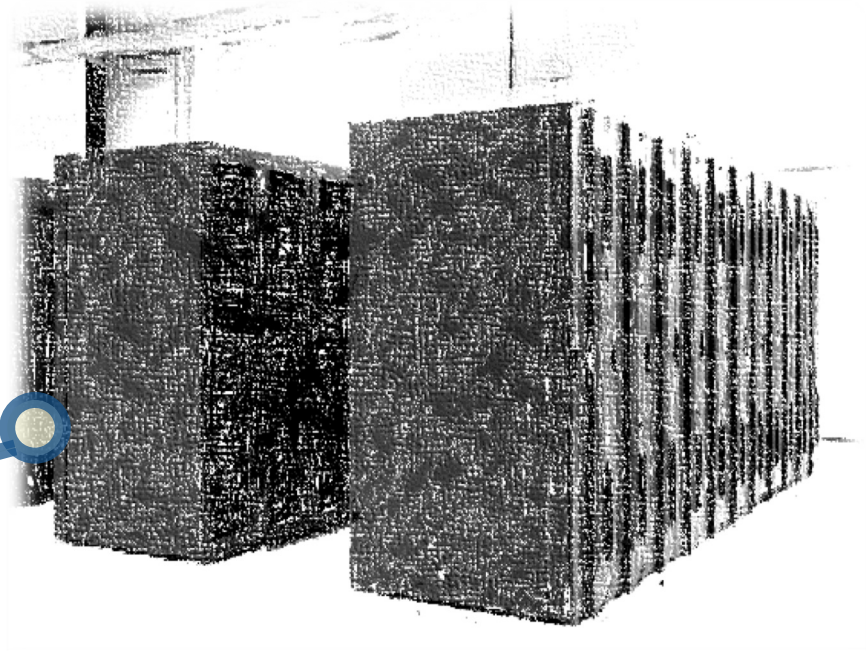
- Andreas Herten, Nvidia Application Lab at Jülich, Jülich Supercomputing Centre
- Dorian Krause, High-Performance Computing Systems, Jülich Supercomputing Centre

Slides:
http://bit.ly/booster-esm-uf

EXPANDING JUWELS

JUWELS Cluster

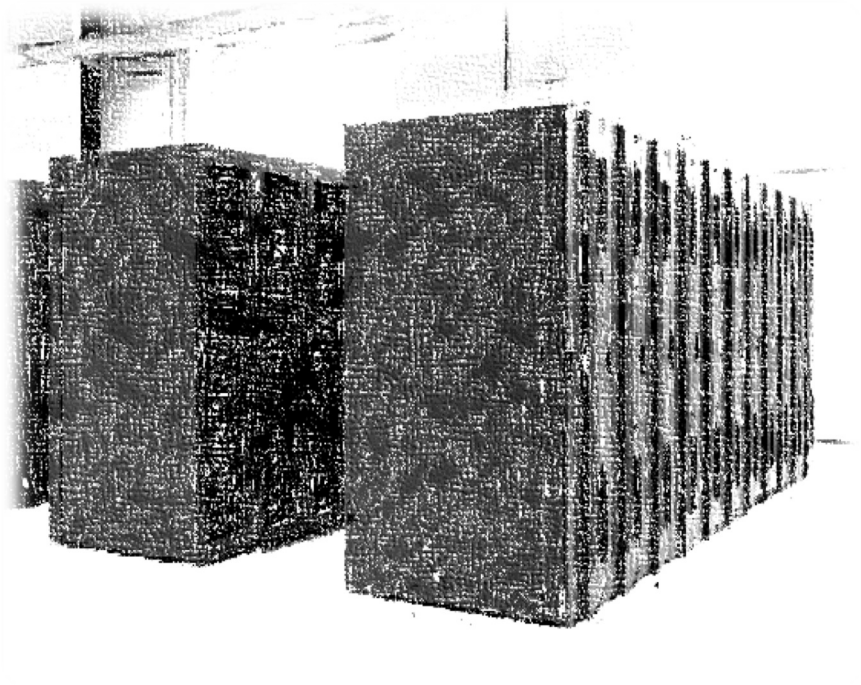
- 2511 compute nodes based on dual-socket Intel Xeon Skylake
- 48 GPU nodes (4 × V100 w/ NVLink2)
- Mellanox InfiniBand EDR100 (100 Gbit/s) network
- Fat-tree topology (1:2@L1)
- 12 PF/s



JUWELS Booster

- Installation in 2020
- Focus on massively-parallel and machine learning applications
  - GPUs
  - Balanced network
- JUWELS performance: ~70 PF/s

JUWELS BOOSTER KEY FIGURES

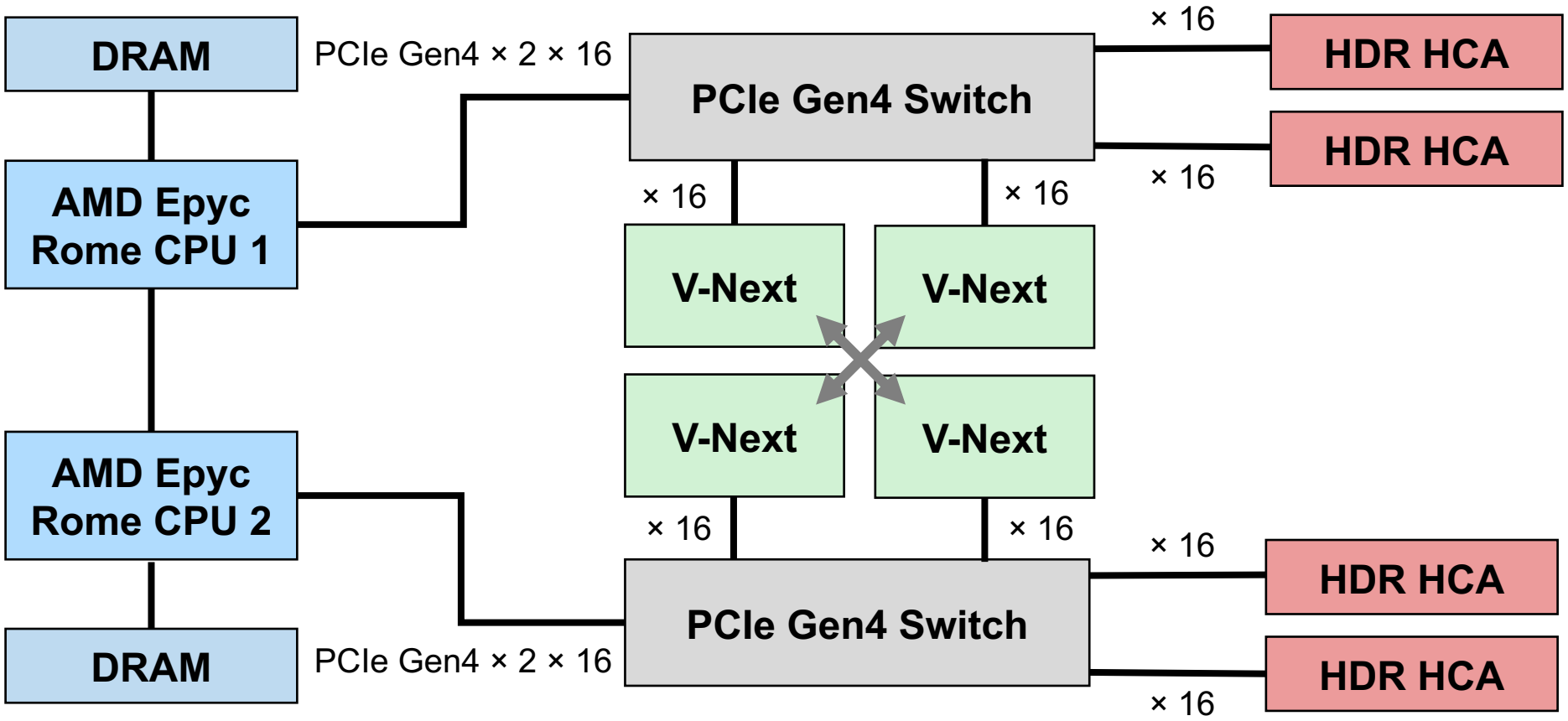


- Atos BullSequana XH2000 system; warm-water cooled; structured into cells (2 racks)
- Node Design
  - CPUs: 2 × AMD Epyc Rome (24 cores)
  - GPUs: 4 × Nvidia Volta-Next; NVLink
  - Network: 4 × Mellanox HDR200 (4 × 200 Gbit/s)
  - Memory: 512 GB DDR4
- DragonFly+ network topology
- 400+ GB/s I/O performance to JUST, up to 1 TB/s to HPST access
- **Partners**: Atos, ParTec, Nvidia, Mellanox

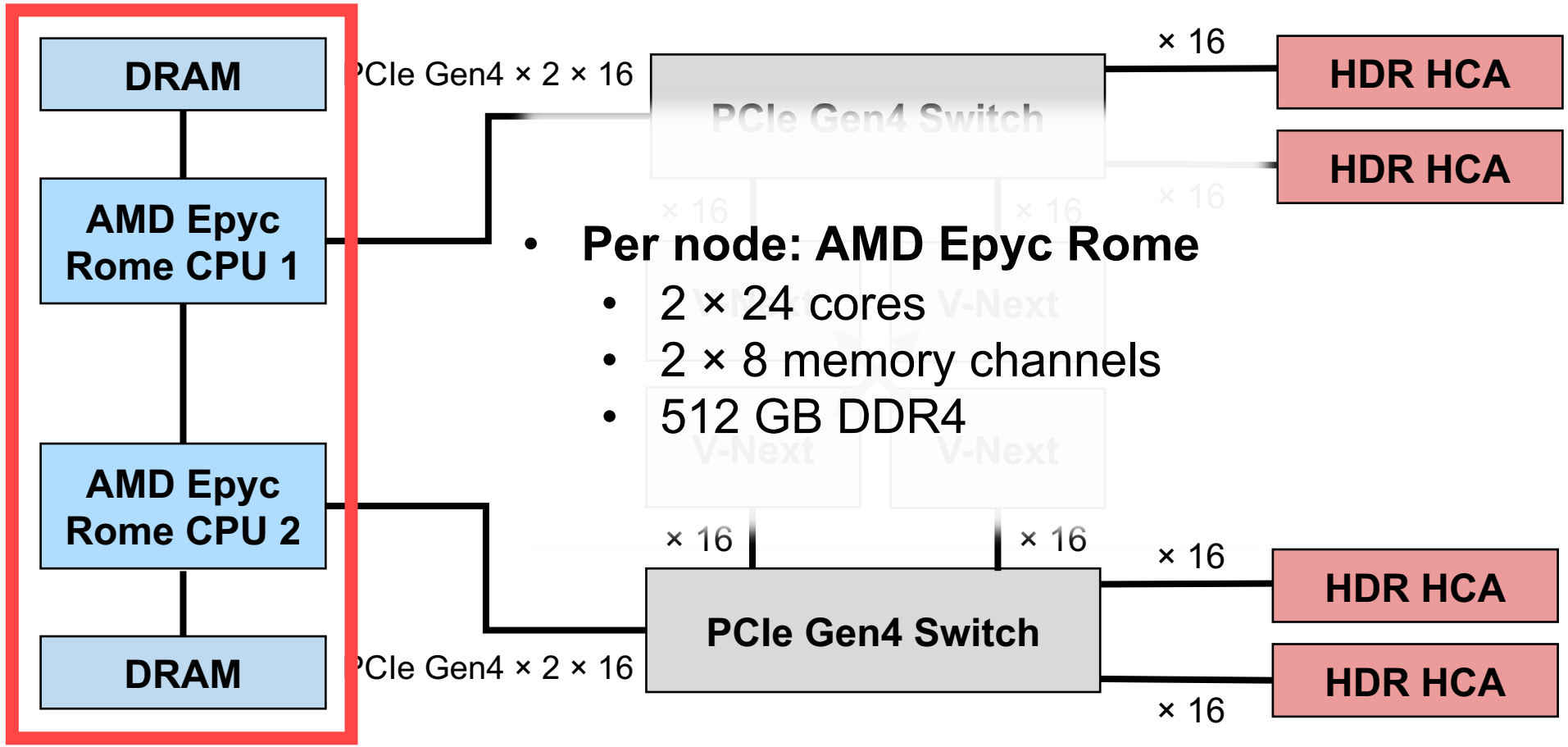
JUWELS BOOSTER NODE DESIGN

© Atos

JUWELS BOOSTER NODE DESIGN

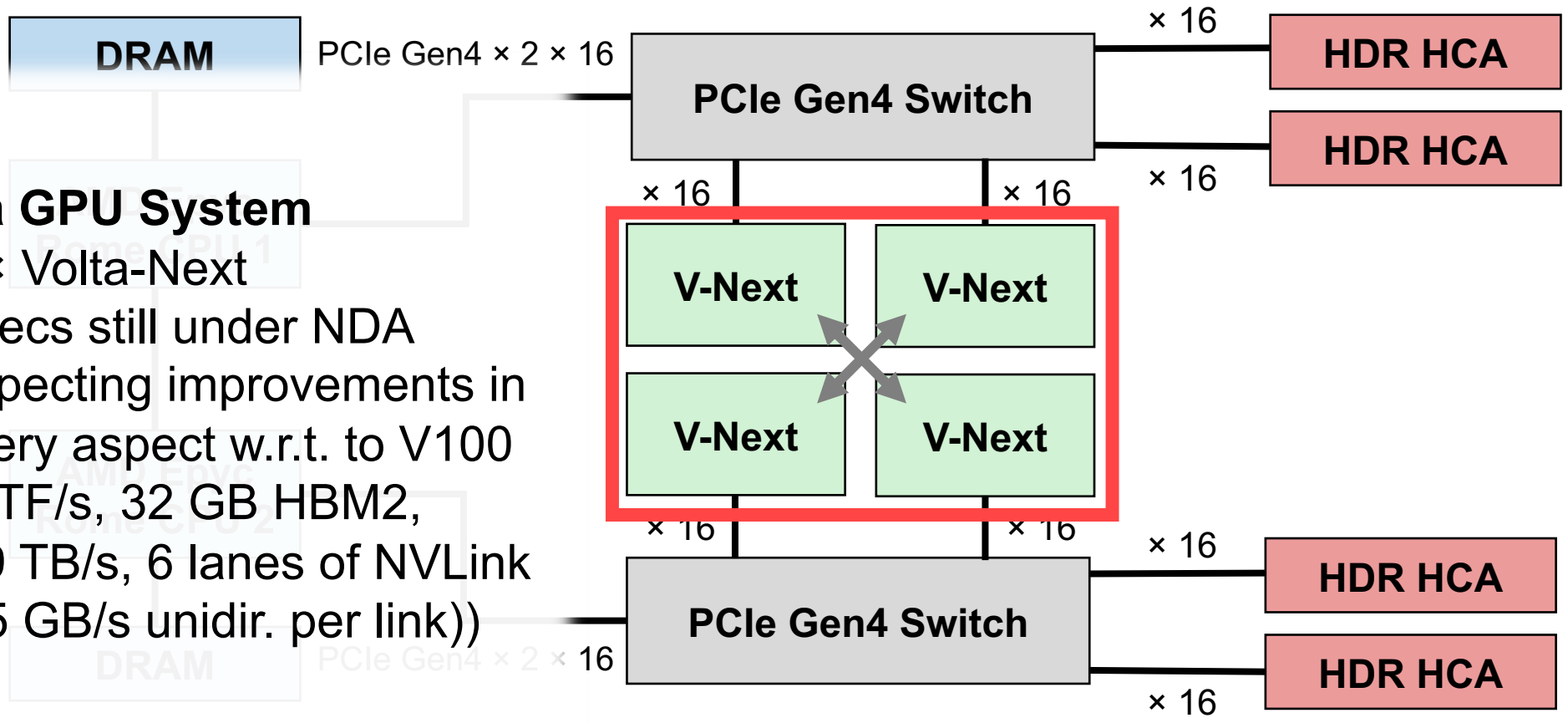


JUWELS BOOSTER NODE DESIGN

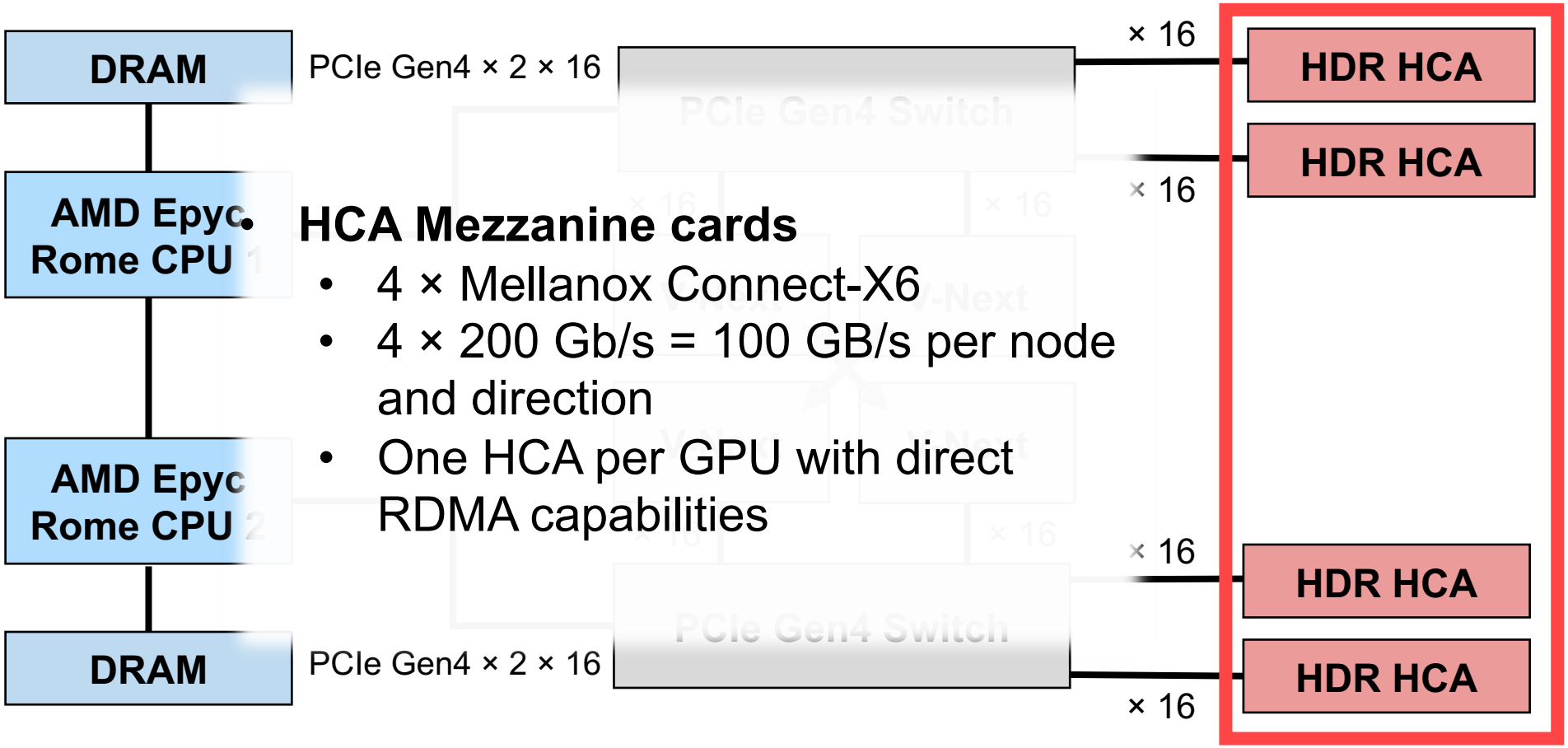


JUWELS BOOSTER NODE DESIGN

- **Nvidia GPU System**
  - 4 × Volta-Next
  - Specs still under NDA
  - Expecting improvements in every aspect w.r.t. to V100 (7 TF/s, 32 GB HBM2, 0.9 TB/s, 6 lanes of NVLink (25 GB/s unidir. per link))



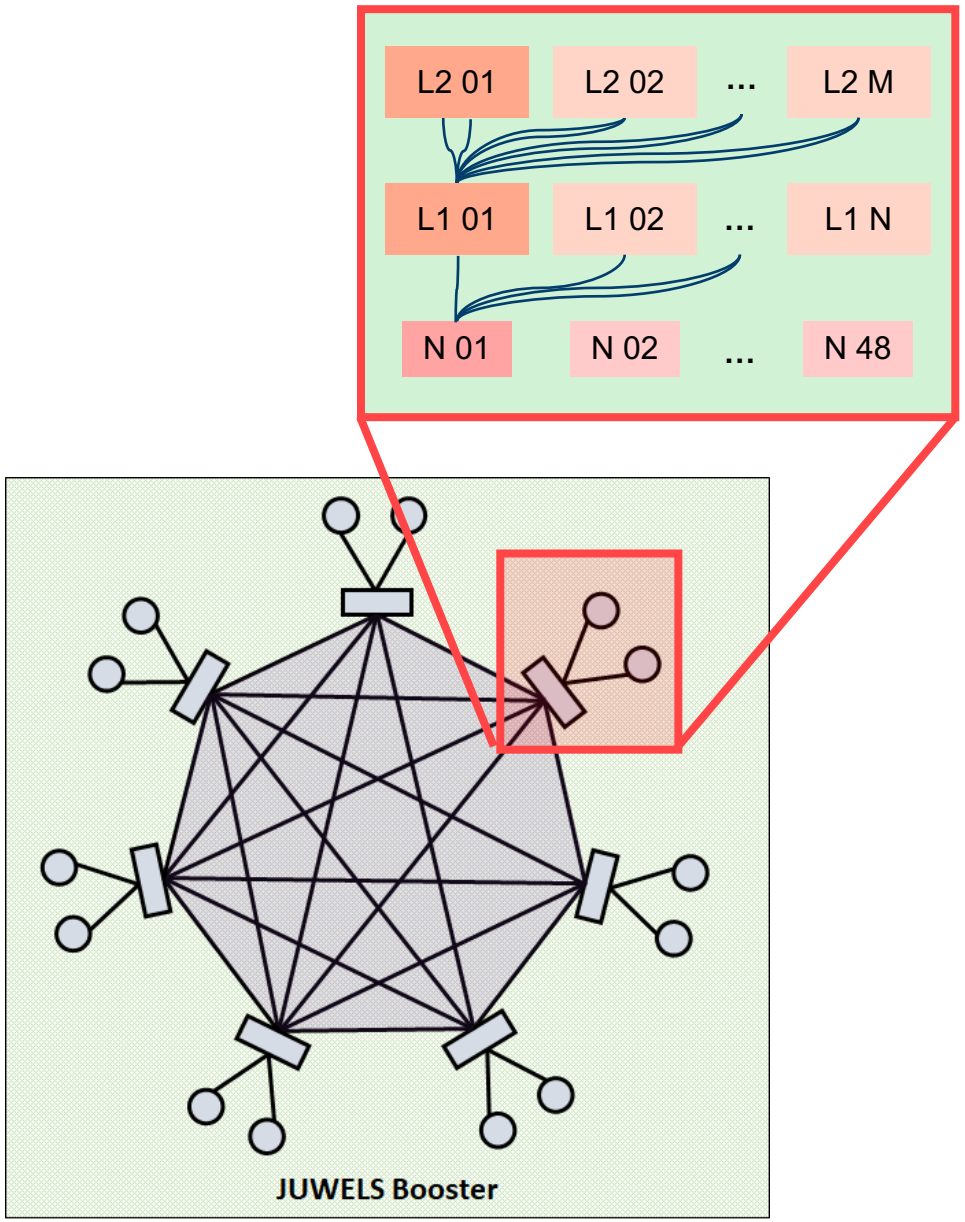
JUWELS BOOSTER NODE DESIGN



HCA Mezzanine cards

- 4 × Mellanox Connect-X6
- 4 × 200 Gb/s = 100 GB/s per node and direction
- One HCA per GPU with direct RDMA capabilities

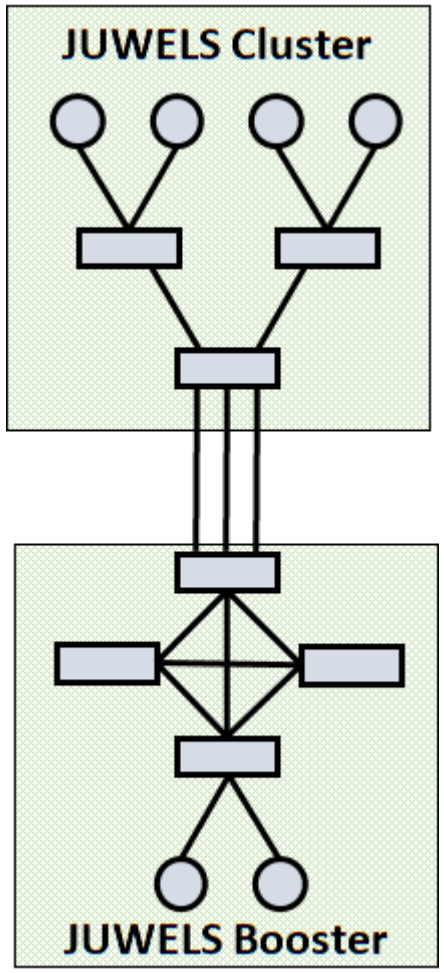
JUWELS DRAGONFLY+ BOOSTER TOPOLOGY



- Non-minimal adaptively routed two-level topology
- Cells with 48 nodes interconnected in two-level non-blocking fat-tree topology
  - 1 link per GPU
  - → High bandwidth in cell
- All-to-all connectivity between cells
  - ~250 GByte/s cell-to-cell bandwidth (static)
    → Effective bandwidth can be several time higher via adaptive routing
- Up to 5 TB/s between Cluster and Booster

JÜLICH Forschungszentrum

JUWELS AS A MODULAR SYSTEM: CLUSTER + BOOSTER

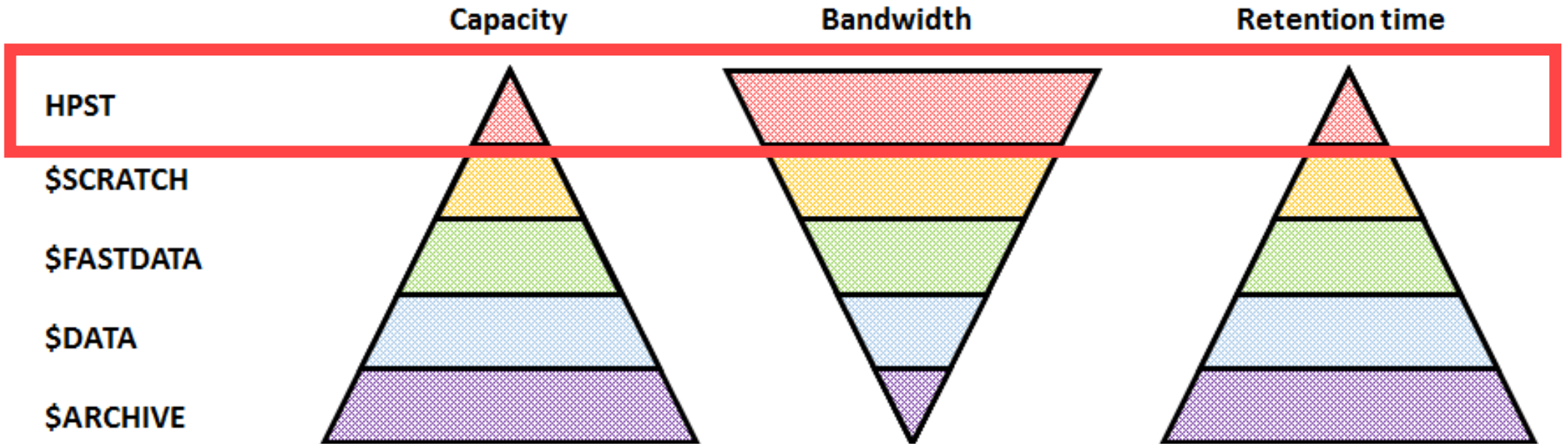


- **Cluster:** Fat-tree with 1:2 pruning @ L1 with ca. 2600 endpoints
  - Technology: EDR
- **Booster:** DragonFly+
  - Adaptive routing
  - Technology: HDR
- **Goal:** High-performance for jobs on Booster and jobs spanning Cluster + Booster
- Booster has dedicated path to **`$SCRATCH`**, **`$PROJECT`**, etc.
- Booster accesses the HPST via Cluster links

JUST HPST ACCESS

IME® © DDN

- **JUST5 High Performance Storage HPST**
  - NVM storage/caching layer for I/O acceleration with direct InfiniBand integration
    → Accessible via POSIX and MPI-I/O on JUWELS and JURECA with coherency
- High Performance: ~1 TB/s



JUWELS BOOSTER SYSTEM SOFTWARE

- Setup builds on JUWELS Cluster environment
- Communication (using GPUDirect GPU↔GPU features)
  - MPI: ParaStationMPI (incl. new CUDA-awareness), Open MPI, MVAPICH?
  - NCCL
  - NVSHMEM (if used?)
- CUDA: Latest CUDA versions, incl. new Nsight Systems, Nsight Compute
- ScoreP, Vampir, … other profiling tools available
  (see https://fz-juelich.de/ias/jsc/msa-seminar talk #2)

JUWELS BOOSTER: HOW/WHERE TO PREPARE?

- *Timeline: User access end of 2020*
- Single-**GPU Nodes**: JUWELS Cluster very similar
  - 2 × Intel Skylake, 4 × Nvidia V100 16 GB, 2 × InfiniBand EDR
  - ⇨ `develgpus` partition
  - `cesmtst` project
  - Test accounts
- **Scaling** on GPU System: CSCS Piz Daint
  - Aries interconnect, DragonFly topology
  - 1 × Intel Haswell, 1 × Nvidia P100 16 GB
  - Access for *Development Projects* (<36k NH) → https://www.cscs.ch/user-lab/allocation-schemes/

- **GPU Seminar**: Tuesdays, 15:00, JSC Lecture Room → https://fz-juelich.de/ias/jsc/msa-seminar
- Helmholtz **GPU Hackathon**, Berlin, 9 – 13.3.2020 → https://www.gpuhackathons.org/
- HPC **Tunathon**, Jülich, ~25 – 29.5.2020 → http://fz-juelich.de/ias/jsc/2020/HPC-tuna

- JSC **Nvidia Application Lab** at Jülich → https://fz-juelich.de/ias/jsc/nvlab
- JSC **SimLab Climate Science** → https://fz-juelich.de/ias/jsc/slcs

JÜLICH APPLICATION LAB NVIDIA

Questions?