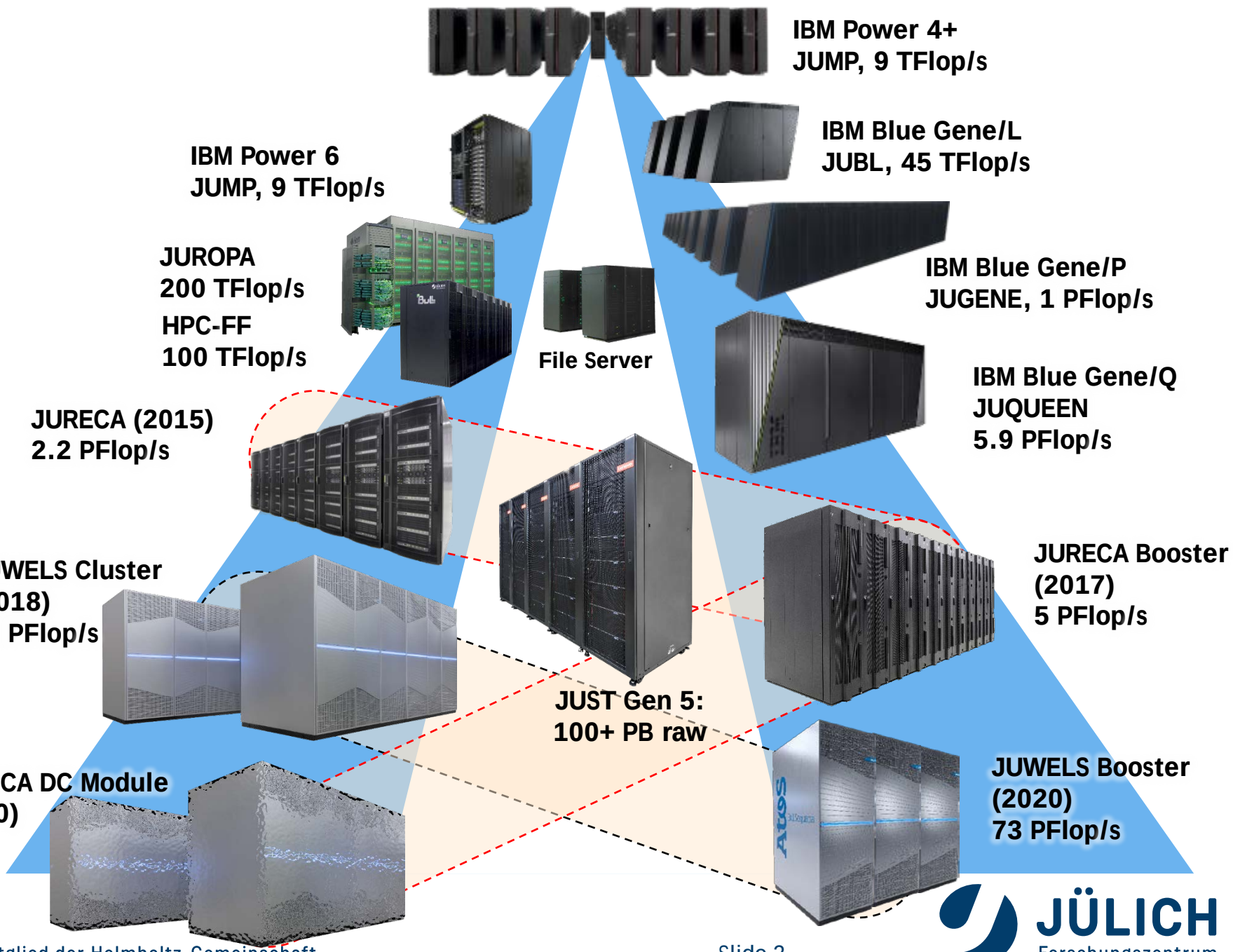
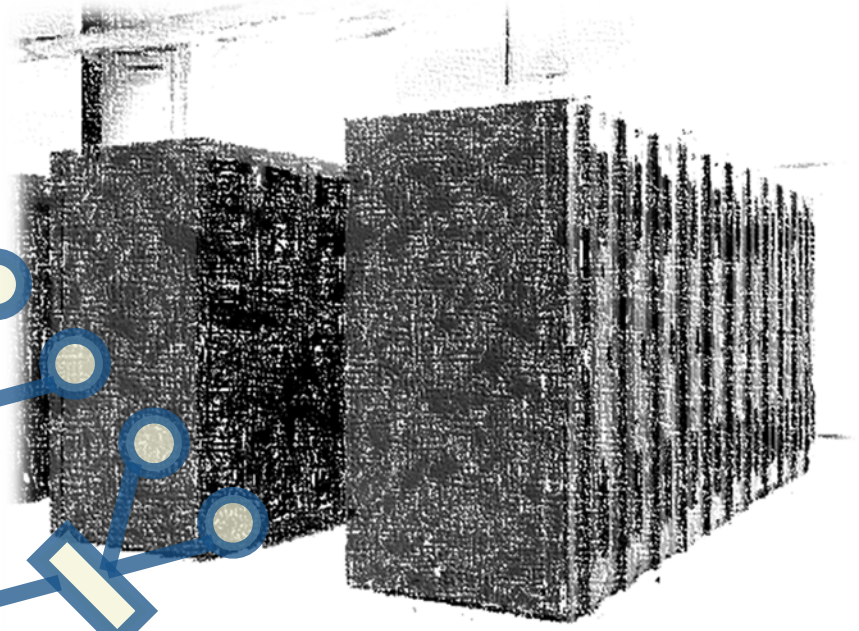
JUWELS BOOSTER

DESIGN OF JSC'S GPU-BASED FLAGSHIP SYSTEM

07.07.2020 I D. ALVAREZ, D. KRAUSE (ON BEHALF OF THE PROJECT MEMBERS)


IBM Power 4+
JUMP, 9 TFlop/s
IBM Power 6
JUMP, 9 TFlop/s
IBM Blue Gene/L
JUBL, 45 TFlop/s
JUROPA
200 TFlop/s
HPC-FF
100 TFlop/s
File Server
IBM Blue Gene/P
JUGENE, 1 PFlop/s
IBM Blue Gene/Q
JUQUEEN
5.9 PFlop/s
JURECA (2015)
2.2 PFlop/s
JUWELS Cluster
(2018)
12 PFlop/s
JUST Gen 5:
100+ PB raw
JURECA Booster
(2017)
5 PFlop/s
JURECA DC Module
(2020)
JUWELS Booster
(2020)
73 PFlop/s

JUWELS



JUWELS Cluster (2018)

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

JUWELS Booster

- 936 nodes (4× A100 w/ NVLink3)
- Mellanox InfiniBand HDR200: DragonFly+ network
- Focus on massively-parallel and learning applications
  - GPUs
  - Balanced network
- 73 PF/s peak

JUWELS BOOSTER IN A NUTSHELL (1/2)

- 936 compute nodes
  - 2× 24-core AMD Epyc Rome CPUs
  - 512 GB DDR memory
  - 4× Nvidia A100 GPUs
    - 9.7 / 19.5 TF/s peak
    - 40 GB HBM2 memory
    - 1.5 TB/s memory performance
    - NVLink3
  - One HDR200 InfiniBand adapter per GPU
- DragonFly+ network topology with 20 cells
  - End-to-end 200 Gb/s HDR200
  - 40 Tb/s connection to Cluster

© Atos

JUWELS BOOSTER IN A NUTSHELL (2/2)

- High-performance storage sub-system
  - 400+ GB/s I/O performance to JUST-DSS
  - Up to 1 TB/s I/O performance to JUST-IME
- Bull Sequana XH2000 system with warm-water cooling
  - 37 °C inlet temperature
- Co-design with Atos, ParTec, NVIDIA, Mellanox since 2018
  - Hardware, Integration & Software
- **User groups:** GCS users, Helmholtz ESM community, Helmholtz AI community

© Atos

JUWELS CHARACTERISTICS (1/2)

JUWELS Cluster (w/o GPU nodes)

- 2511 nodes
- 48 cores / node
- **384** FP64 units (CPU)
- 4.15 TF per node
- 10.6 PF peak performance
- 240 K concurrency
- ≥96 GB main memory
- **0** B high-bw memory
- 264 TB total memory
- 256 GB/s memory bw per node
- 0.6 PB/s memory bw

JUWELS Booster

- 936 nodes
- 48 cores / node
- **3456** FP64 units (GPU)
- 78 TF per node
- 73 PF peak performance
- **»52 M** concurrency
- 512 GB main memory
- **160** GB high-bw memory
- 629 TB total memory
- 6 TB/s memory bw per node
- 5.6 PB/s memory bw

JUWELS CHARACTERISTICS (2/2)

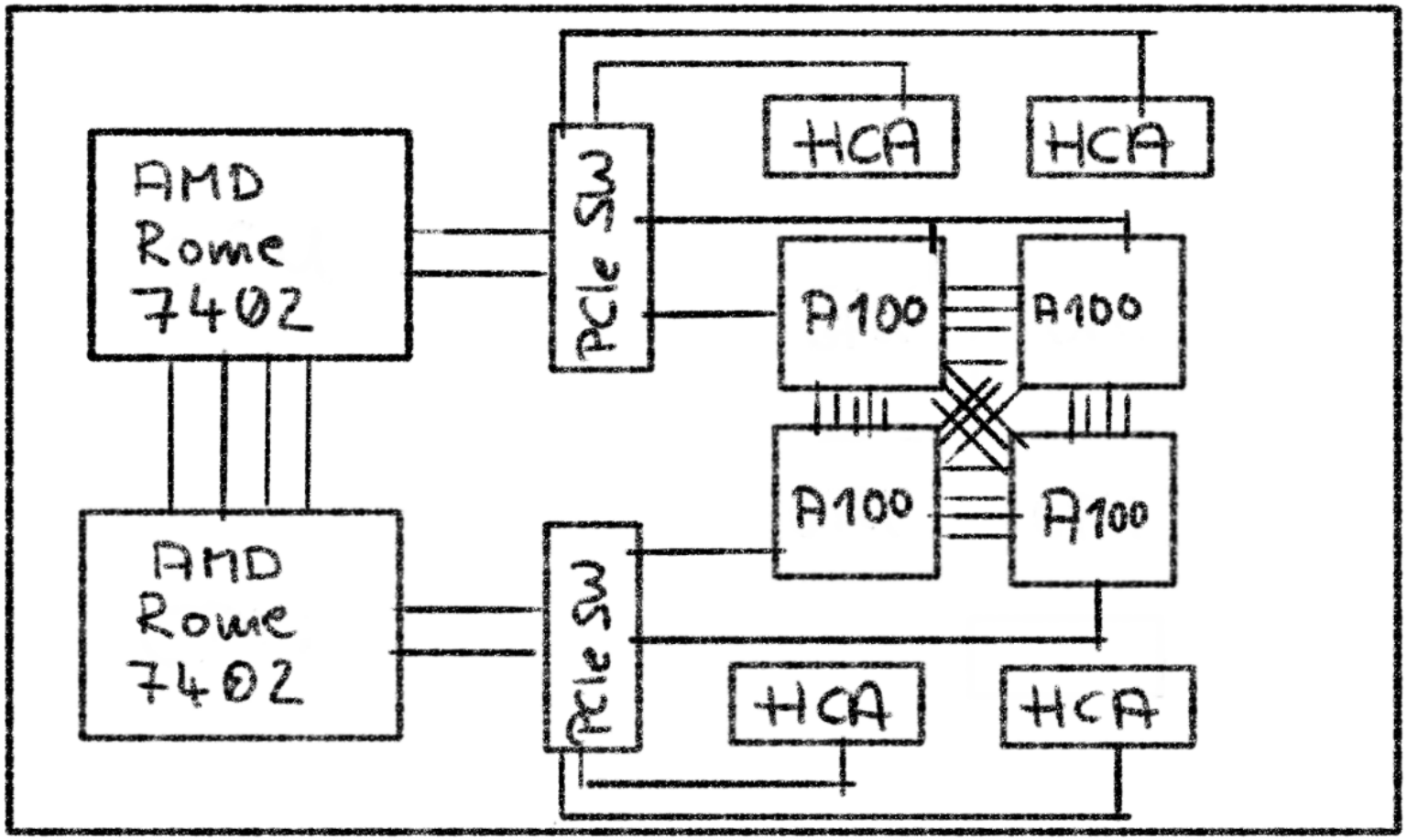
JUWELS Cluster (w/o GPU nodes)

- 100 Gb/s link speed (**EDR**)
- 1 NIC per node
- 100 Gb/s node injection bw
- 24.1 Gb per TF
- 251 Tb/s injection bw
- FT topology
- 63 Tb/s global bandwidth
- detem. routing
- 4+ GB/s I/O bw per node
- 10 GB/s I/O bw per node (IME)

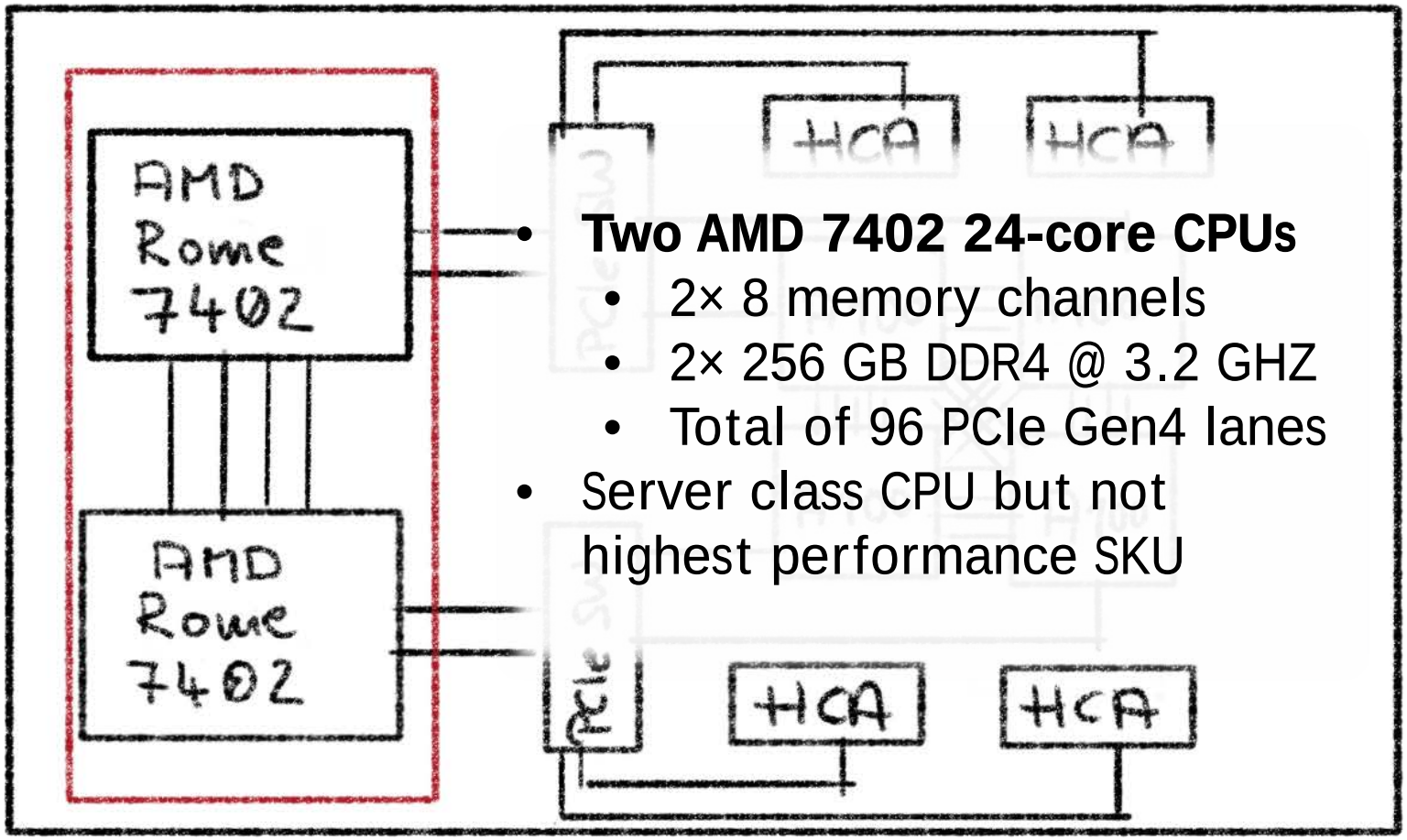
JUWELS Booster

- 200 Gb/s link speed (**HDR**)
- 4 NICs per node
- 800 Gb/s node injection bw
- 10.3 Gb per TF
- 749 Tb/s injection bw
- DF+ topology
- 200 Tb/s global bandwidth
- adaptive routing
- 10+ GB/s I/O bw per node
- 90+ GB/s I/O bw per node (IME)

JUWELS BOOSTER NODE DESIGN



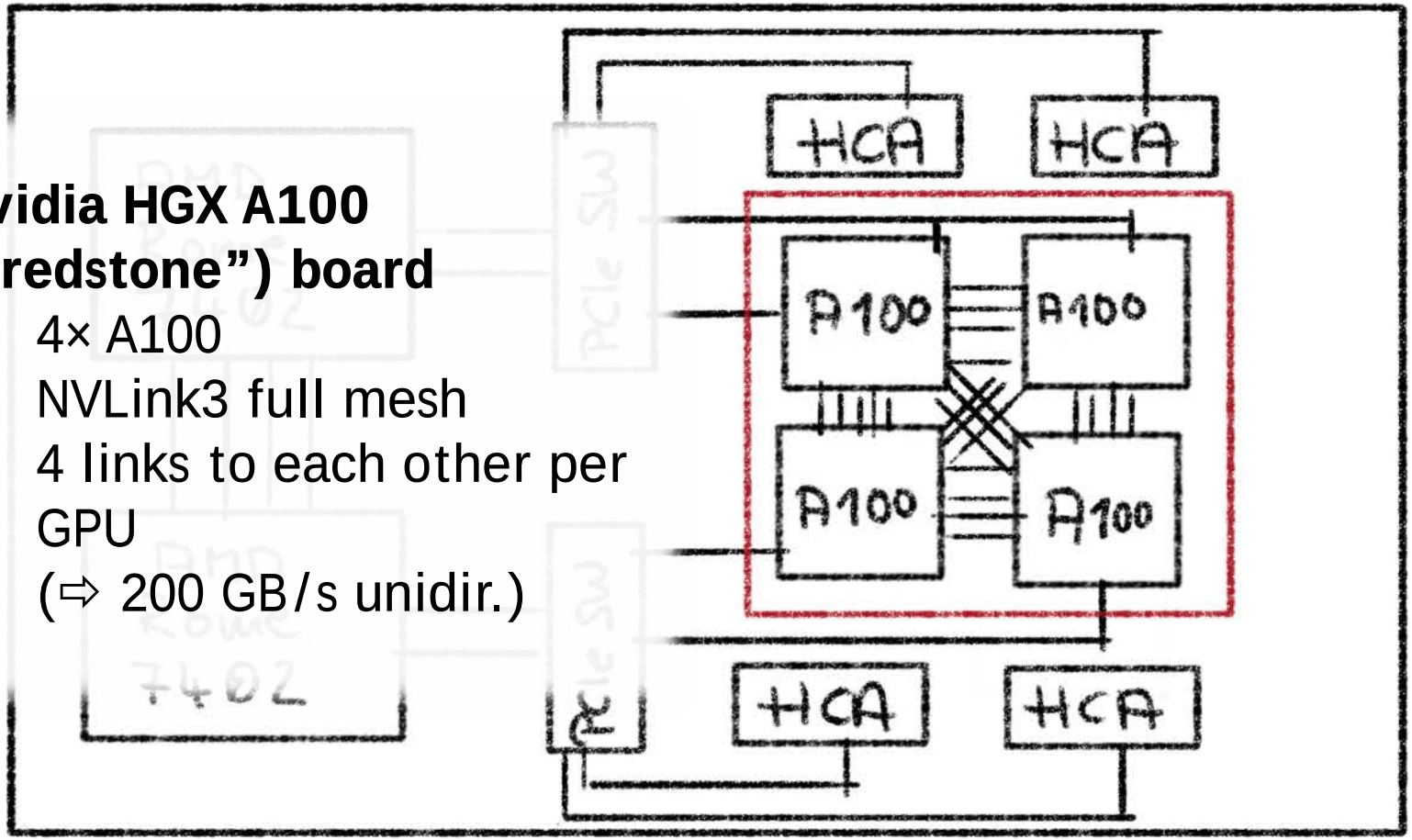
JUWELS BOOSTER NODE DESIGN



- **Two AMD 7402 24-core CPUs**
  - 2× 8 memory channels
  - 2× 256 GB DDR4 @ 3.2 GHZ
  - Total of 96 PCIe Gen4 lanes
- Server class CPU but not highest performance SKU

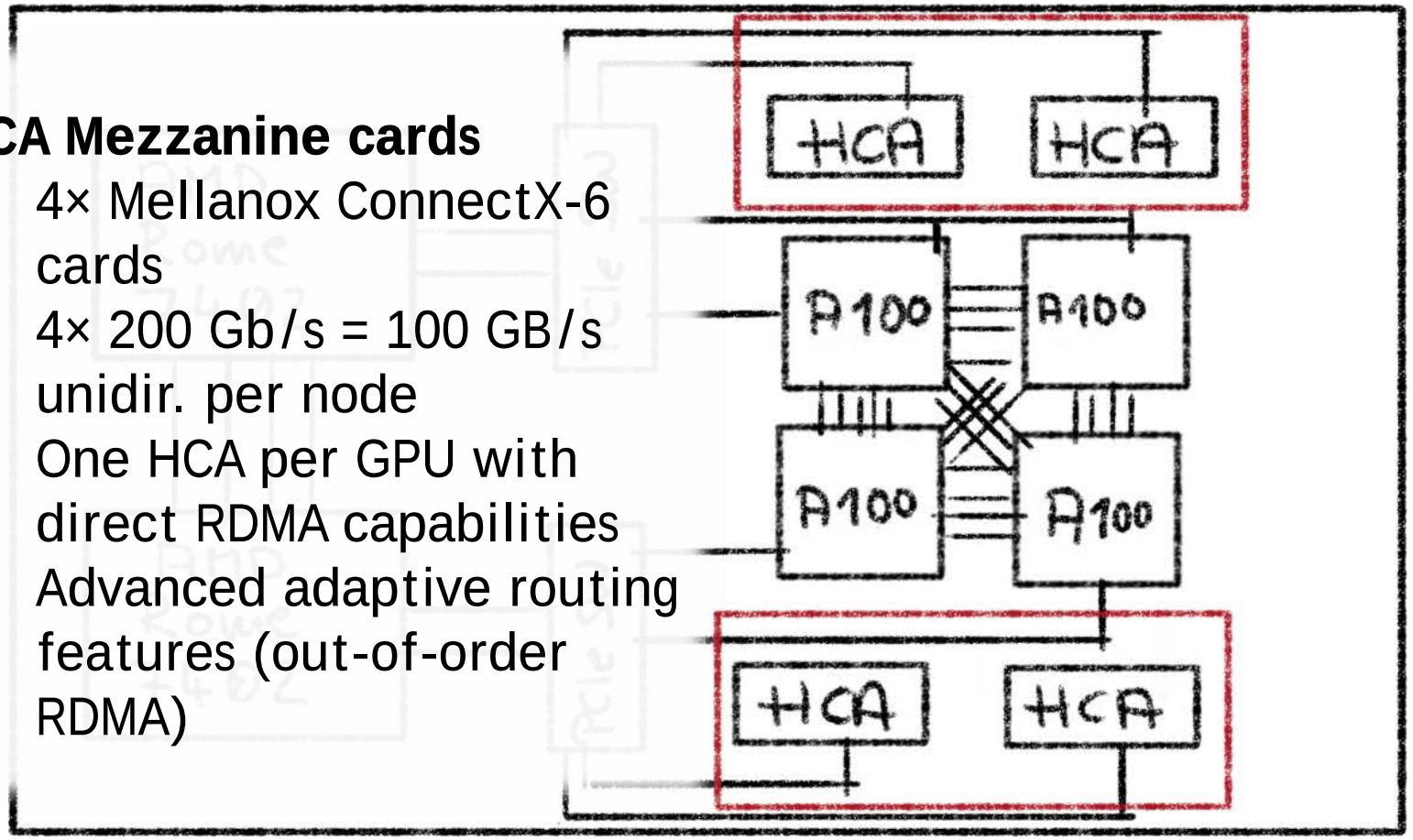
JUWELS BOOSTER NODE DESIGN

- **Nvidia HGX A100 (“redstone”) board**
  - 4× A100
  - NVLink3 full mesh
  - 4 links to each other per GPU
    (⇨ 200 GB/s unidir.)

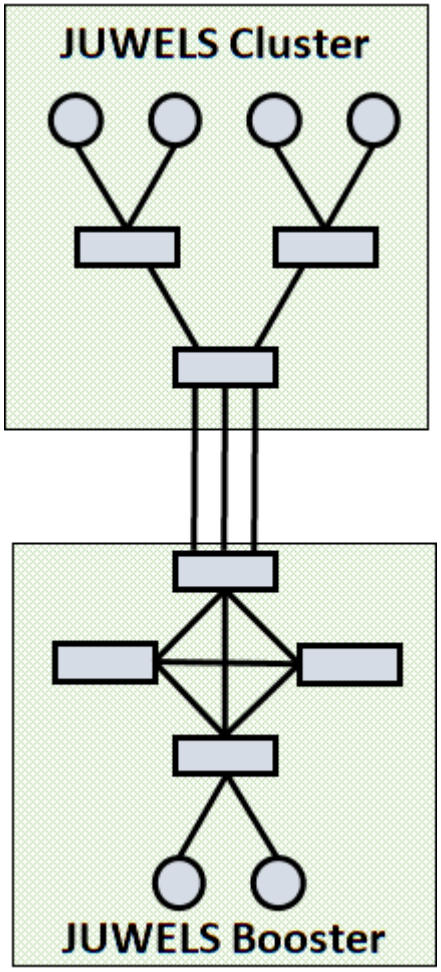


JUWELS BOOSTER NODE DESIGN

- **HCA Mezzanine cards**
  - 4× Mellanox ConnectX-6 cards
  - 4× 200 Gb/s = 100 GB/s unidir. per node
  - One HCA per GPU with direct RDMA capabilities
  - Advanced adaptive routing features (out-of-order RDMA)

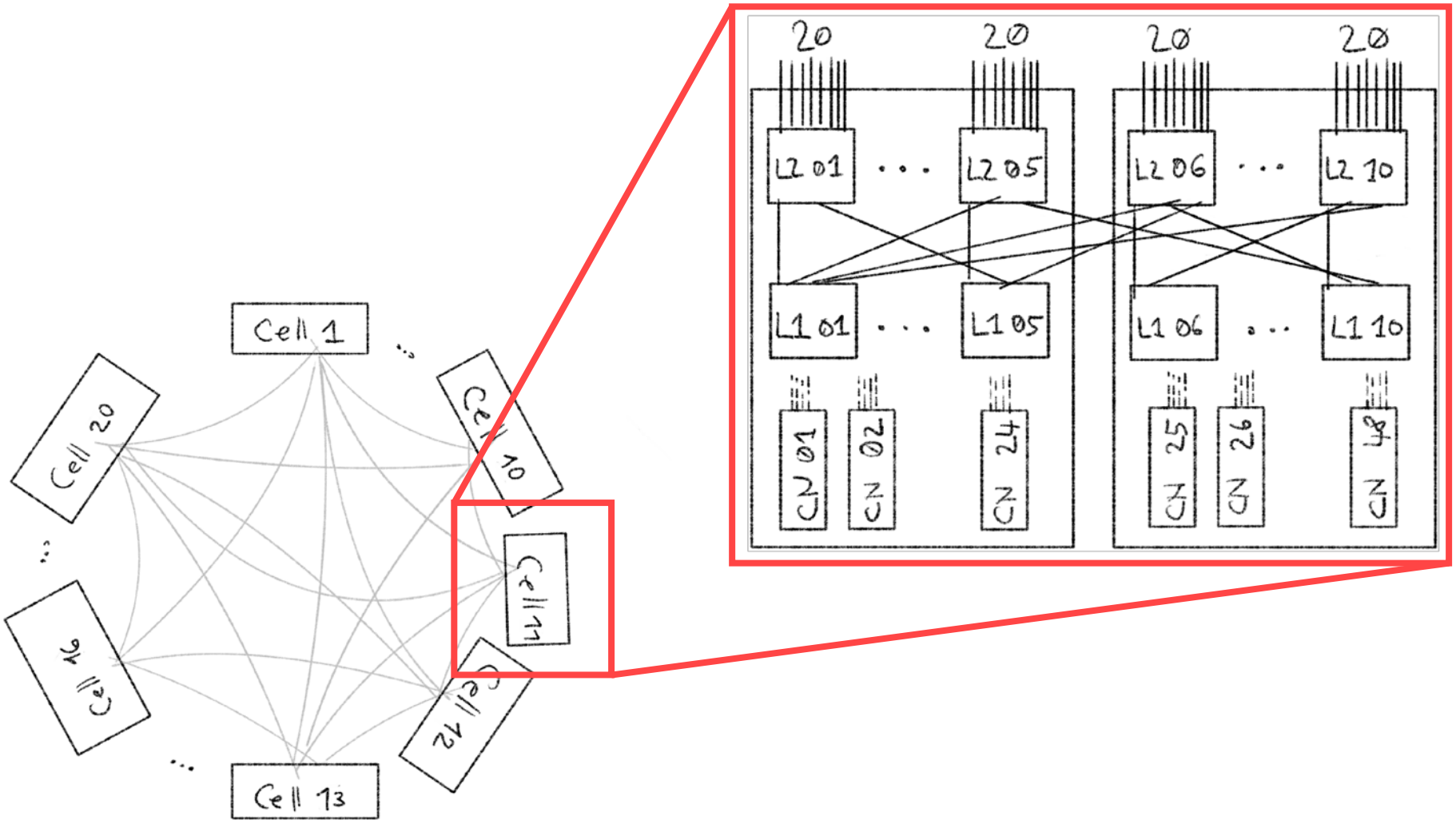


NETWORK TOPOLOGY: OVERVIEW



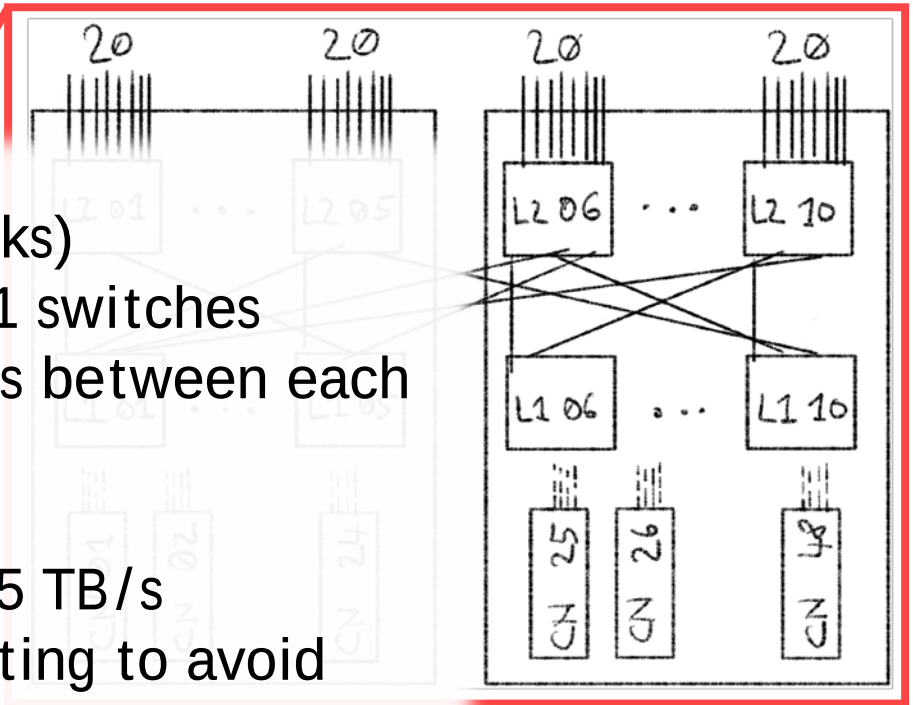
- **Cluster:** Fat-tree with 1:2 pruning @ L1
  - 2623 endpoints
  - Technology: EDR + FDR
- **Booster:** 20 cells DragonFly+
  - 3744 endpoints
  - Adaptive routing
  - Technology: HDR200
- **Goal:** High-performance for jobs on Booster and jobs spanning Cluster + Booster

JUWELS DRAGONFLY+ BOOSTER TOPOLOGY

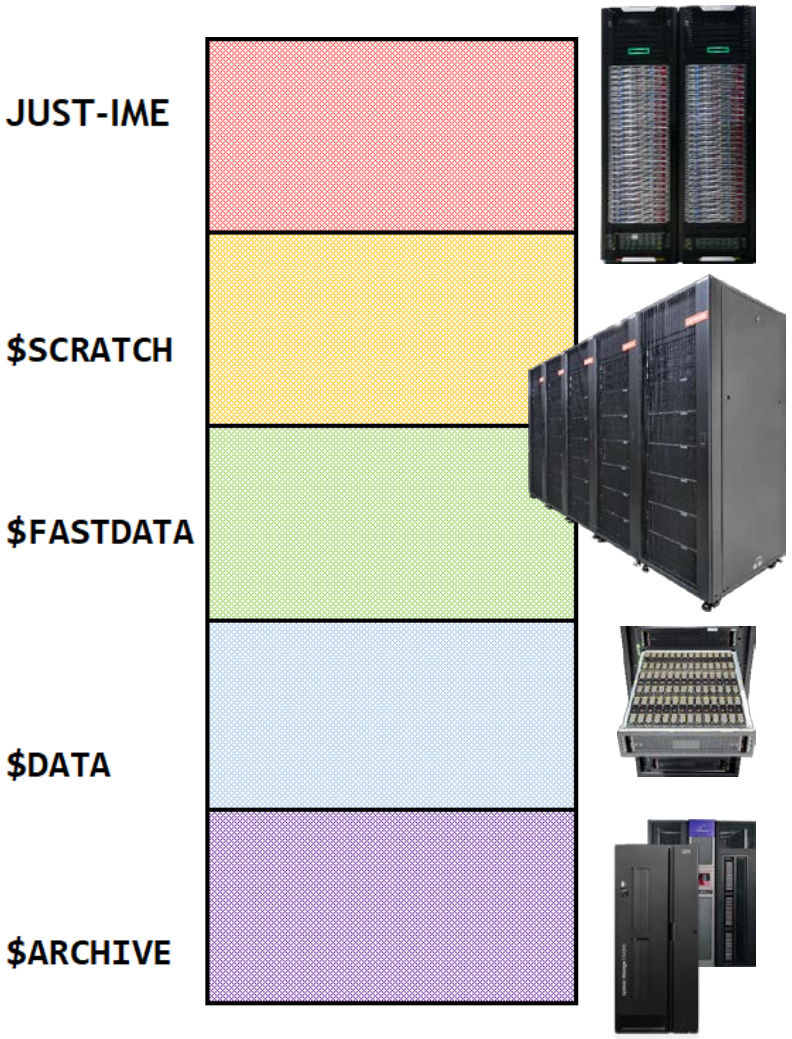


JUWELS DRAGONFLY+ BOOSTER TOPOLOGY

- 48 nodes per cell (two racks)
  - 4× links striped over L1 switches
- Bundle width 10 ⇨ 10 links between each cell
  - 250 GB/s
  - Global bandwidth: 1.25 TB/s
- Non-minimal adaptive routing to avoid congestion
- 200 links (10 per cell) to Cluster
  - 5 TB/s



JUST: MULTI-TIER STORAGE SYSTEM



Bandwidth optimized, capacity limited storage (NVM based)

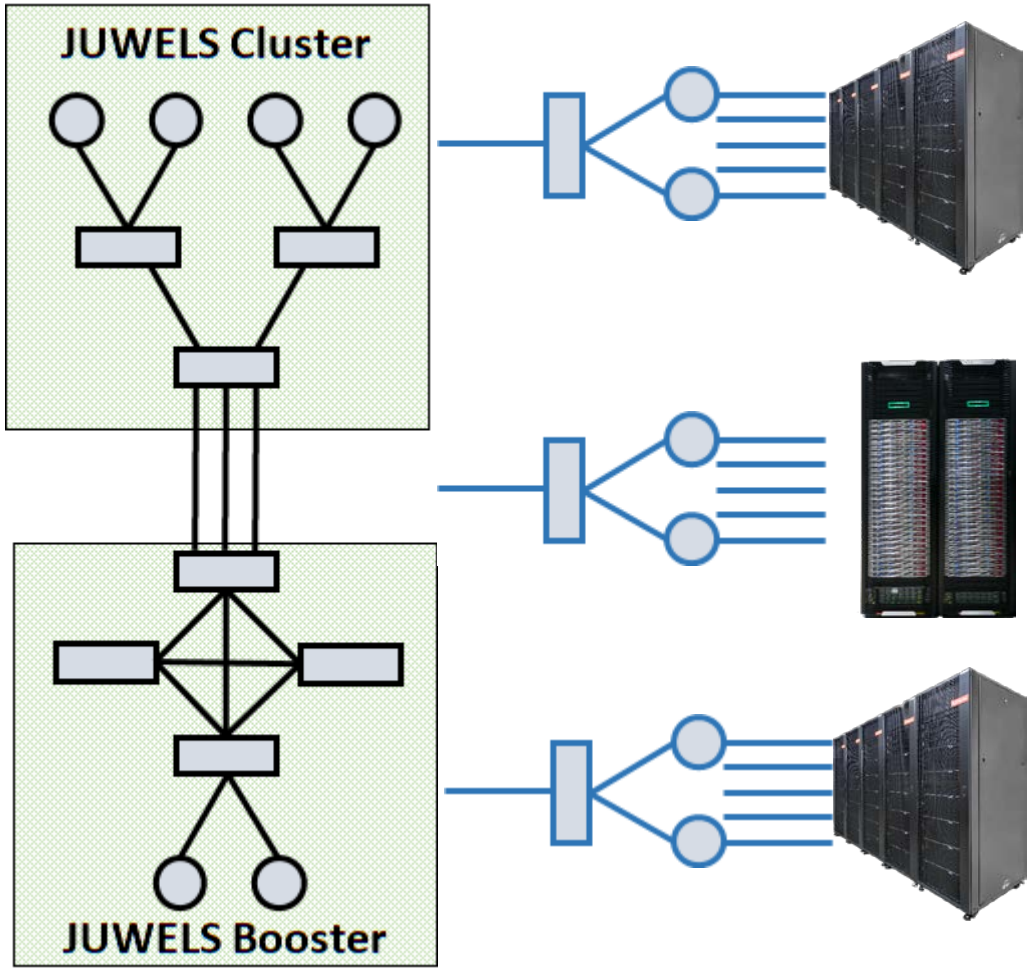
Capacity and bandwidth balancing HPC storage (temp.)

Capacity and bandwidth balancing HPC storage (pers.)

High capacity, low bandwidth storage (campaign use cases)

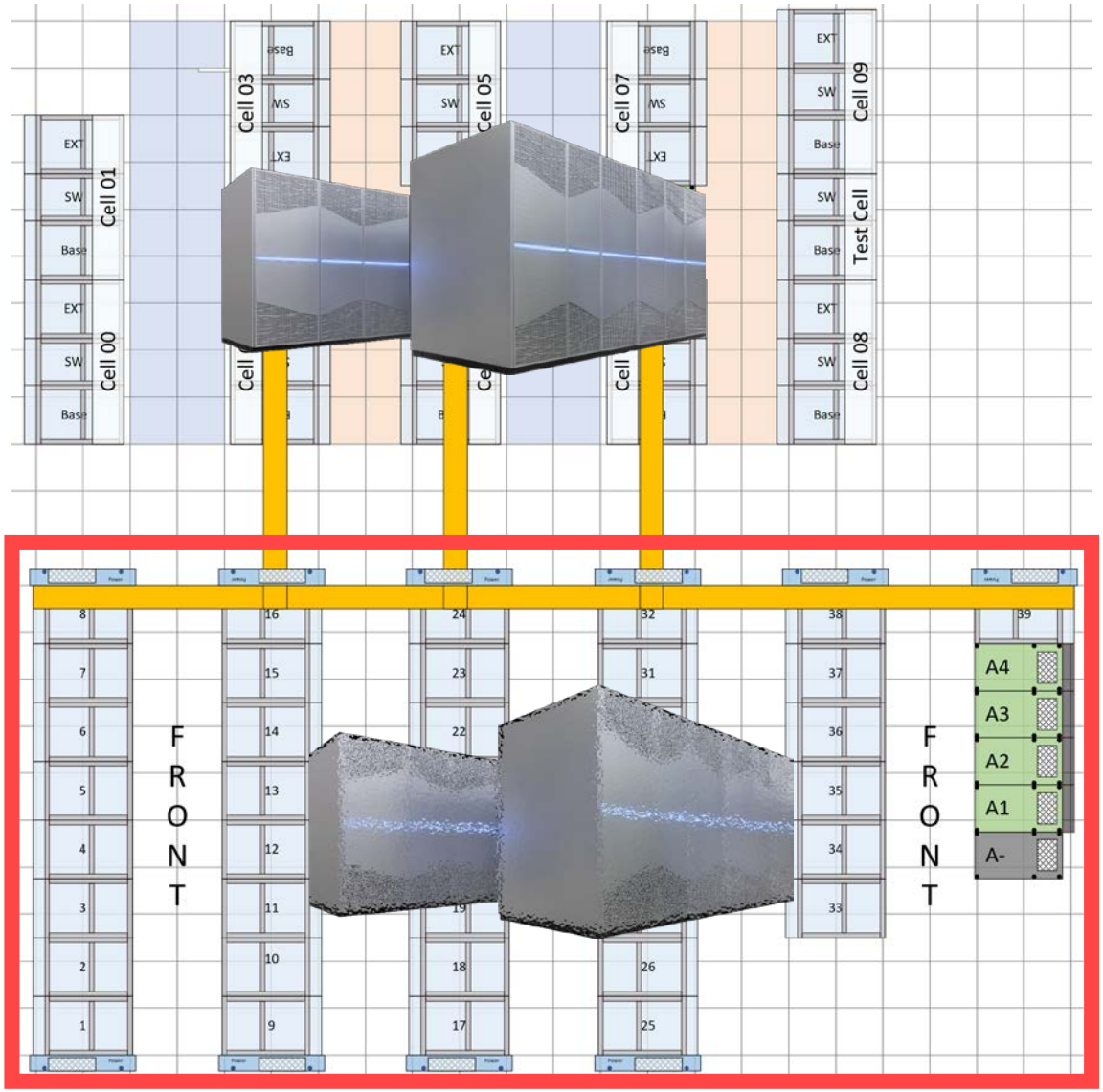
Archival storage based on high latency media

NETWORK TOPOLOGY: EXTERNAL SYSTEMS



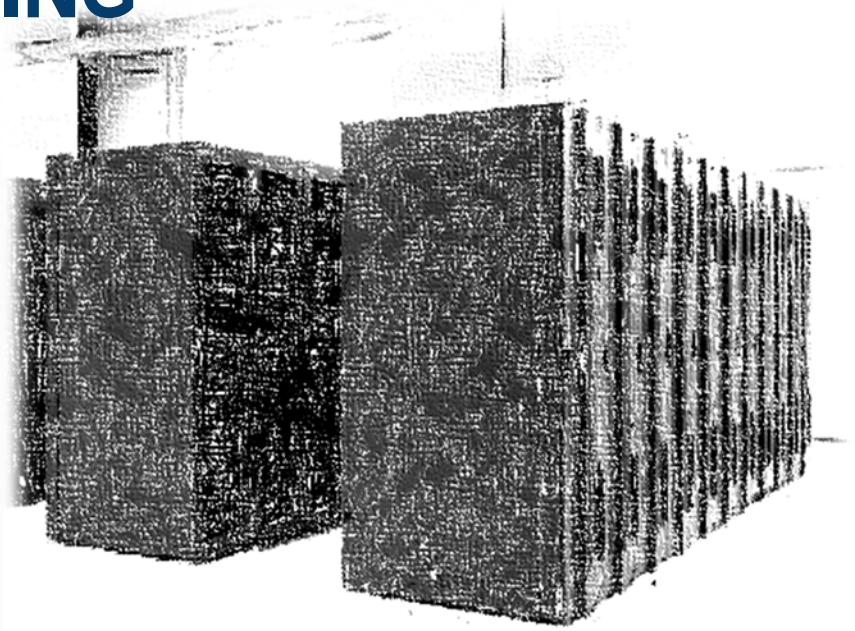
- **JUST5-DSS for Cluster**
  - 4× Mellanox SX6036G @ L0 in tree
- **JUST5-IME for C+B**
  - HDR200-based integration in **Cluster FT** with 56 ports (1.4 TB/s)
- **JUST5-DSS for Booster**
  - 8× Mellanox Skyway GW in cell 20
  - 800 GB/s network bw.

JUWELS INFRASTRUCTURE



- 39 Sequana XH200 racks
  - 6 rows
- 2.3 MW expected HPL power consumption
  - Transformers 1 & 2 for machine hall fully utilized

JUWELS BOOSTER: COOLING



- Direct warm-water cooling
  - 34 - 37 °C facility-side inlet temperature
  - Will enable operation with dry coolers
- Free cooling equipment planned for Mid 2021
  - Extension to machine hall building

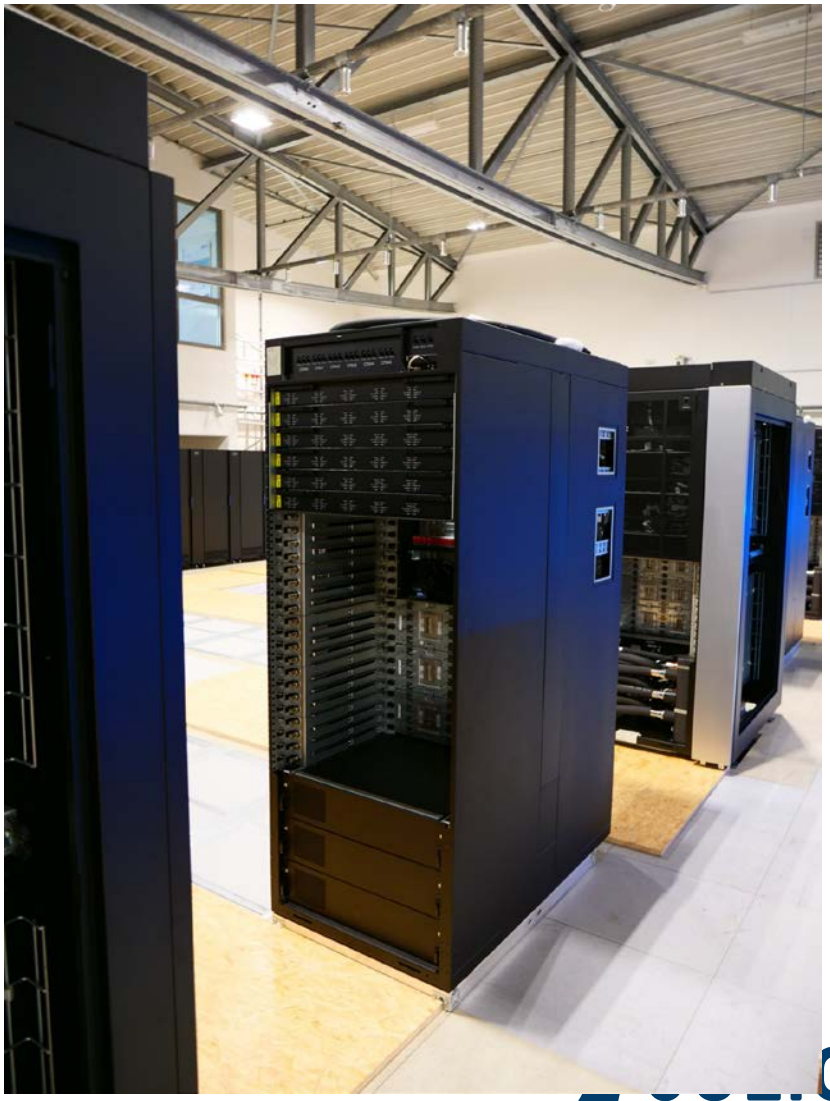
JUWELS BOOSTER SYSTEM SOFTWARE

- **ParaStation** as core enabler of the Modular Supercomputing Architecture
  - Resource Management
  - MPI Implementation (MPICH-based)
  - Extensions to support multi-GPU nodes
    - CUDA awareness, affinity management, RDMA and multi-rail capabilities
- Open MPI supported
- **Slurm** as Workload Manager for JUWELS
- Red Hat Enterprise Linux and CentOS 8
  - Upgrade of Cluster together with merge planned

JUWELS BOOSTER: CURRENT STATUS

JUWELS BOOSTER: CURRENT STATUS

JUWELS BOOSTER: CURRENT STATUS



JUWELS BOOSTER: CURRENT STATUS

- Deployment project has been affected by Covid-19 counter measures in different countries
- A100 GPUs are GA since May
- Service partition delivered
  - Software installation of service nodes in progress
  - Complete with exception of Mellanox Skyway InfiniBand-to-Ethernet gateways
- Delivery of XH2000 racks on-going
- Connection of Ethernet networks scheduled for Early July

JUWELS BOOSTER: TENTATIVE TIME LINE

- 2nd half of August: Share of system available for internal tests
  - Planned start of early access tests with selected applications/use-cases
- September: Integration of Cluster and Booster InfiniBand networks
- October: Stabilization, benchmarking & acceptance testing
- Beginning of November: Start of production

THANK YOU