EVIDEN | JÜLICH Forschungszentrum

Evolution of the Sequana System Architecture

The Past, the Present and the Future

Dominik Friedrich – HPC Solution Architect

Introduction to Supercomputing at JSC – Theory & Practice
Jülich, May 24 2024

EVIDEN | JÜLICH Forschungszentrum

Content overview

01
Sequana 1: X1000
JUWELS Cluster

02
Sequana 2: XH2000
JUWELS Booster, JURECA-DC

03
Sequana 3: XH3000
JUPITER

04
Q&A

Pre Sequana Era

Bull B700 DLC Solution

- Chassis based approach
- Direct Liquid Cooling on blades and switches
- Design with 9 blades (18 nodes) per Chasssis was directly linked to 36-Port IB Switches

- Many Sequana features already present:
  - All in one approach
  - Central Power (54V DC)
  - Free Cooling & Heat reuse

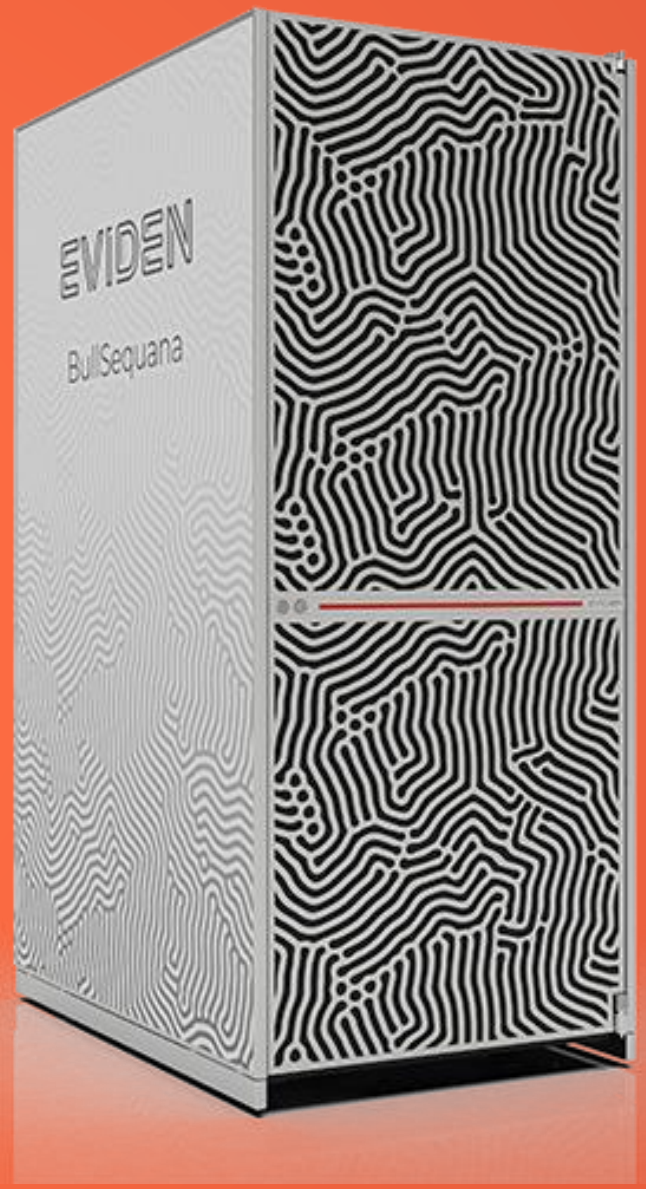
- Installations in Germany
  - TU-Dresden: Taurus
  - DKRZ Hamburg: Mistral

EVIDEN | JÜLICH Forschungszentrum

EVIDEN | JÜLICH Forschungszentrum

The Past:

Sequana 1 - X1000



Bull Sequana X1000 cell technology

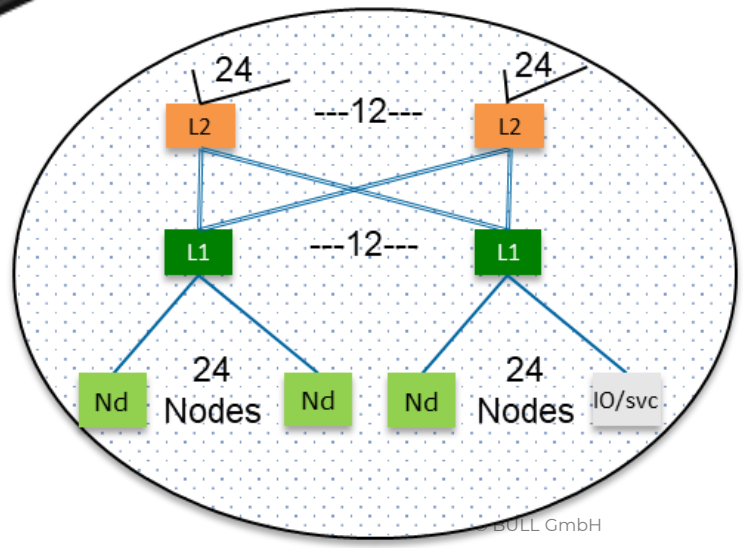
Bull Sequana X1000 cell

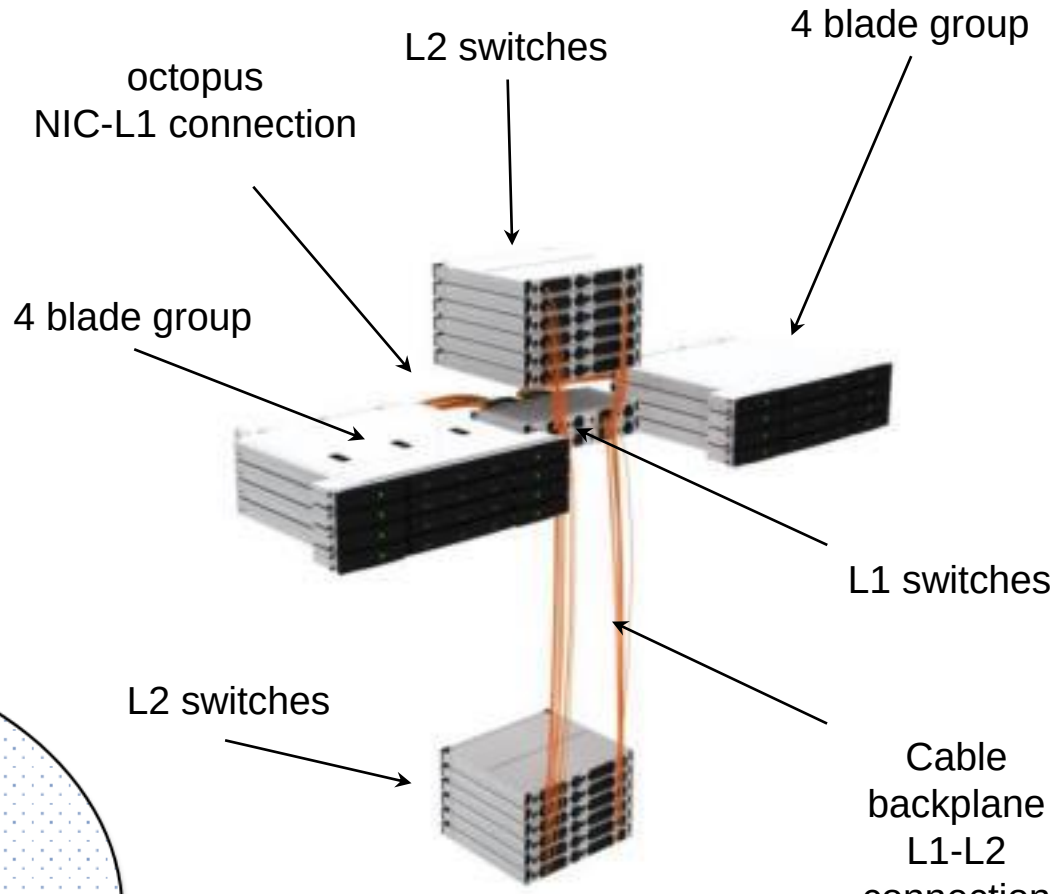
support for several types of CPU and GPU compute nodes

- **3 cabinets**
  - 2 compute cabinets
  - 1 x L1 & L2 interconnect switches and management server cabinet
- **Up to 288 compute nodes (96 blades)**
  - Supports Intel Xeon Broadwell-EP processor
  - Supports Intel XeonPhi Knights Landing (KNL) processor
  - Supports Intel Xeon Skylake-EP processor
  - Supports Nvidia GPU Pascal accelerator
- **2 interconnect technologies supported**
  - InfiniBand EDR
  - Bull eXascale Interconnect (BXI)
- **Full Direct Liquid Cooling**
  - compute blades
  - L1 & L2 interconnect switches
  - Power supplies (end 2016)
- **Island Management and Administration**
  - Redundant server with
  - Shared storage

EVIDEN JÜLICH Forschungszentrum

Bull Sequana X1000 – embedded interconnect





Fast Interconnect layout

Bull Sequana X1000 (JUWELS Cluster)

Lessons Learned

PRO:

- **Modular system platform**
- **Blade system**
- **Multiple blade types**
- **Cell Concept as building block**
- **Direct Liquid Cooling**
- **Up to 40°C warm water as inlet temperature (free cooling)**
- **All in one approach (Compute, interconnect, power, cooling)**

CON:

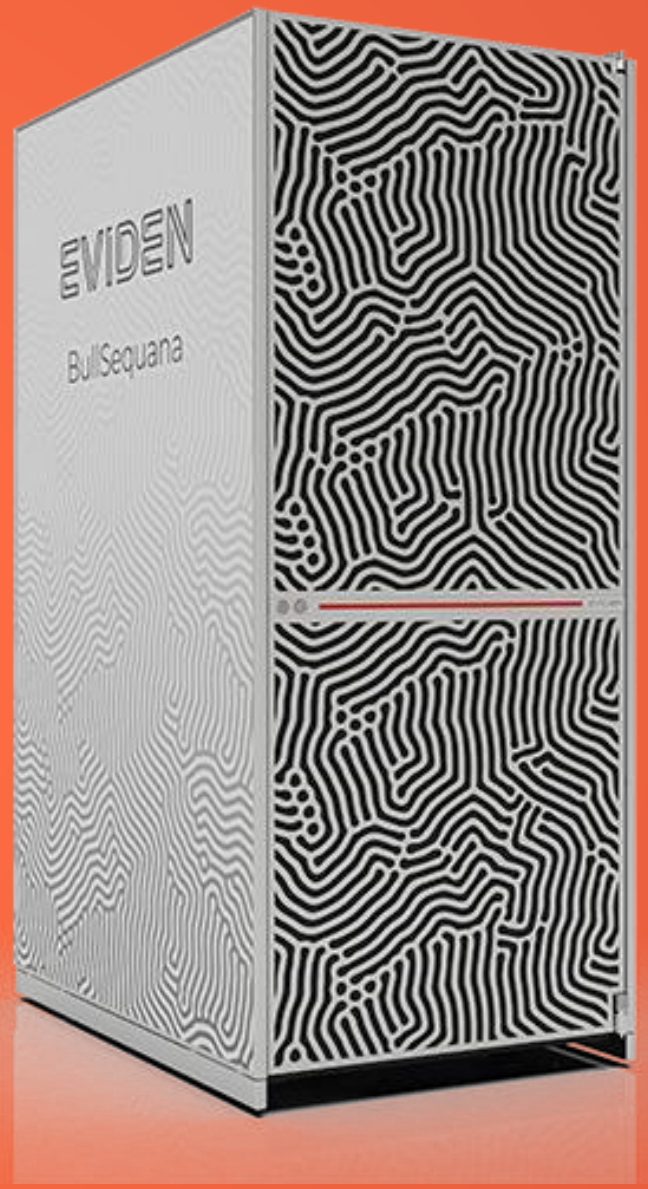
- **Fixed Cell Size (288 nodes) as building block**
- **Fixed interconnect topology (L1 and L2)**
- **Proprietary switch design**
- **Missing flexibility with EDR (only 2:1 Fat-Tree)**
- **Air-Cooled components (ISMAs, PSUs)**

EVIDEN JÜLICH Forschungszentrum

EVIDEN | JÜLICH Forschungszentrum

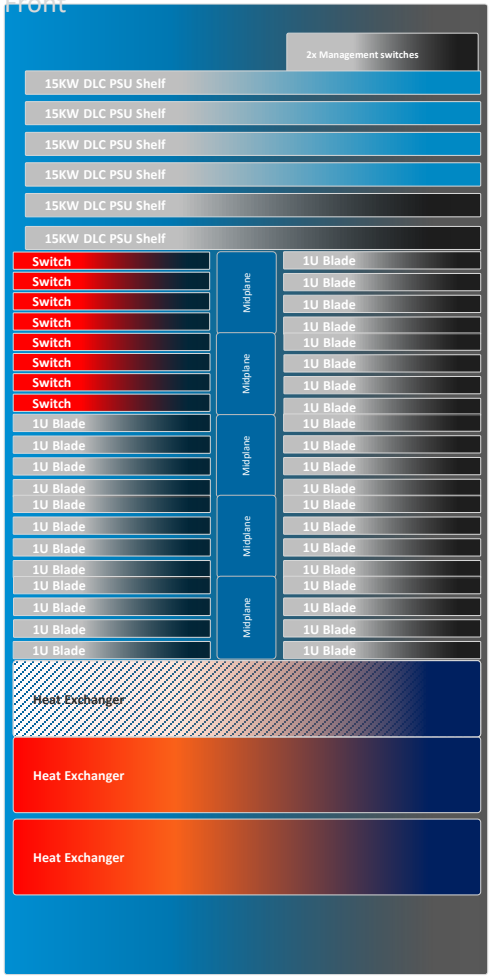
The Present:

Sequana 2 – XH 2000



BullSequana XH2000 Details

Side View



Front View

PDU + Power controller

up to 6 x 15KW DLC shelves
(Optional redundancy)

2 x Leaf Eth switches

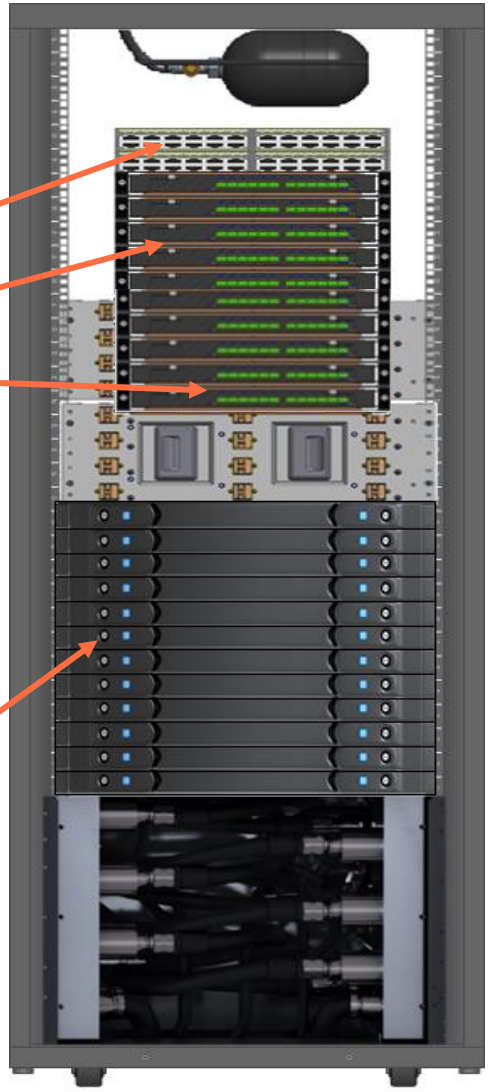
up to 10 switches

4 to 20 compute blades

up to 12 compute blades

up to 3 Hydraulic chassis
(2+1 optional redundancy)

Rear View

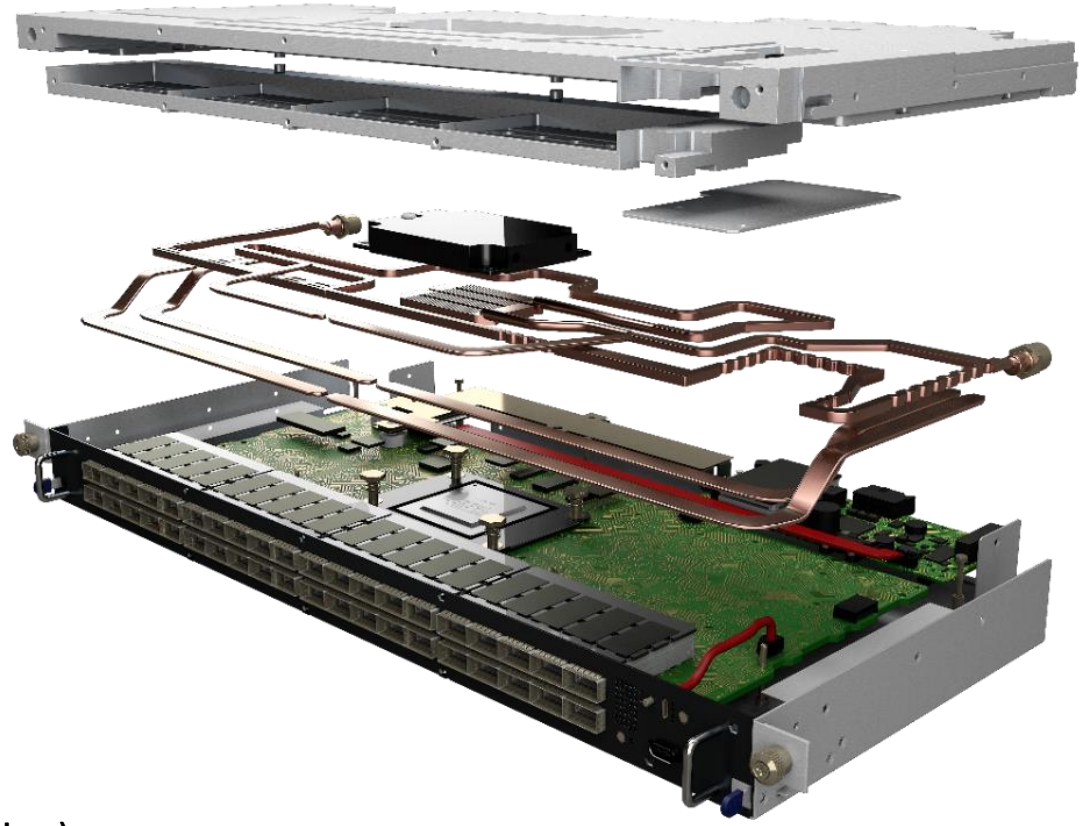


BullSequana XH2000 Networking

New HDR interconnect

DLC cooled Mellanox HDR switch

- 40 X HDR 200Gb/s ports in a 1U switch
- 80 X HDR100 100Gb/s ports in a 1U switch
- 16Tb/s aggregate switch throughput
- Up to 15.8 billion messages-per-second
- 90ns switch latency
- Atos Cold Plate – DLC

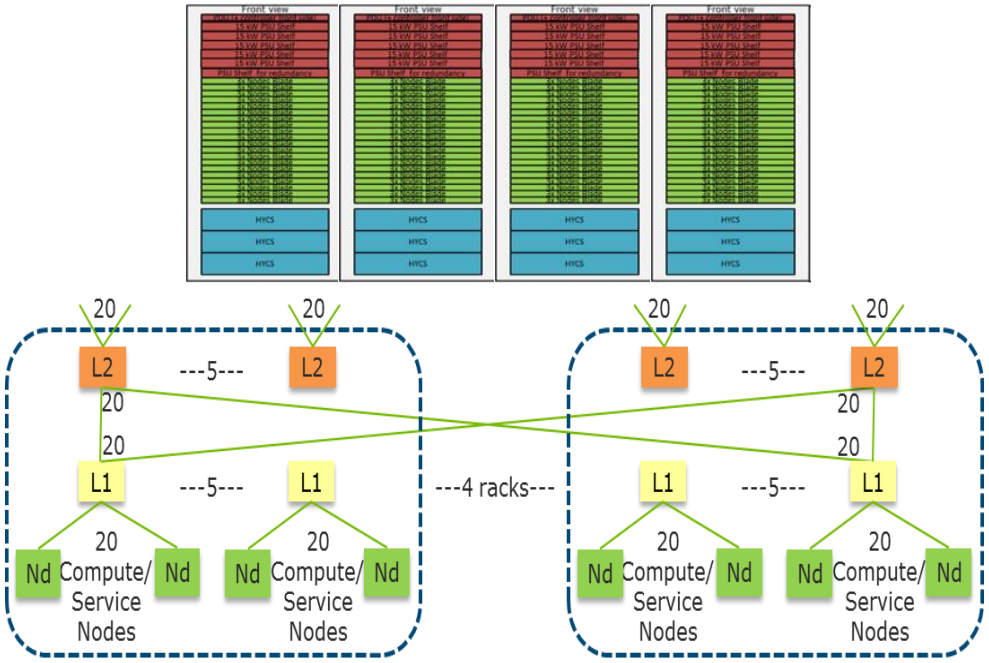


HDR Flexible Sideplane

- 4 blades / up to 12 nodes HDR Sideplane
- QSFP connectors, HDR and HDR100 option (Y cables in SOH)
- Single connector for Sideband and management (up to 12 nodes)

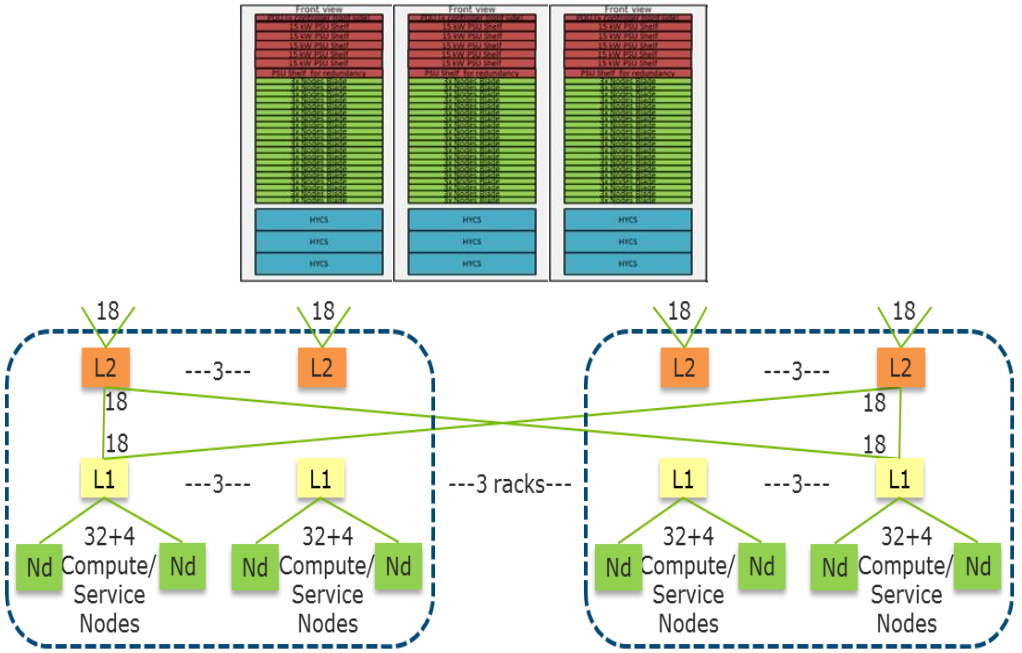
BullSequana XH2000 Topology options

Cell Design: NON BLOCKING Fat Tree (HDR & HDR100)



Single Sequana Cell: **HDR FT 1:1**

- 4 racks
- 384 Compute Nodes
- 40 HDR switches
- **9,6** CN/SW



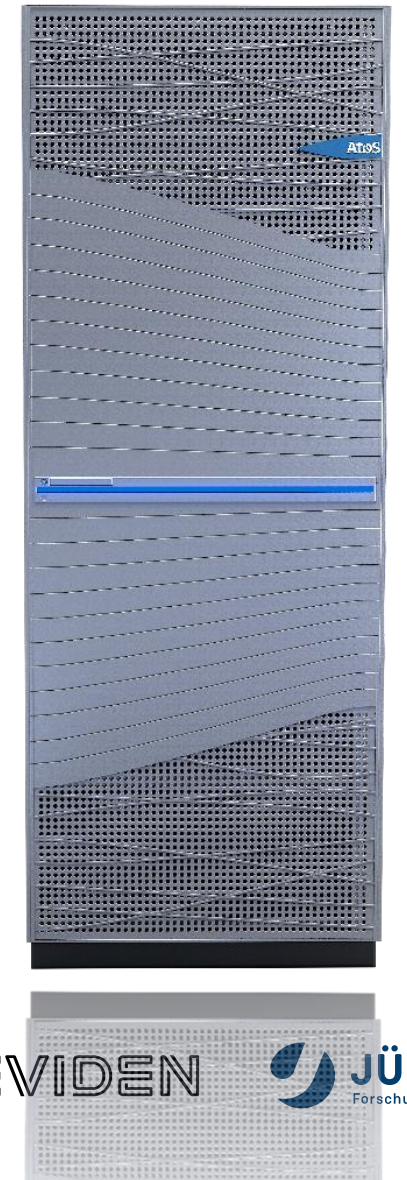
Single Sequana Cell: **HDR100 FT 1:1**

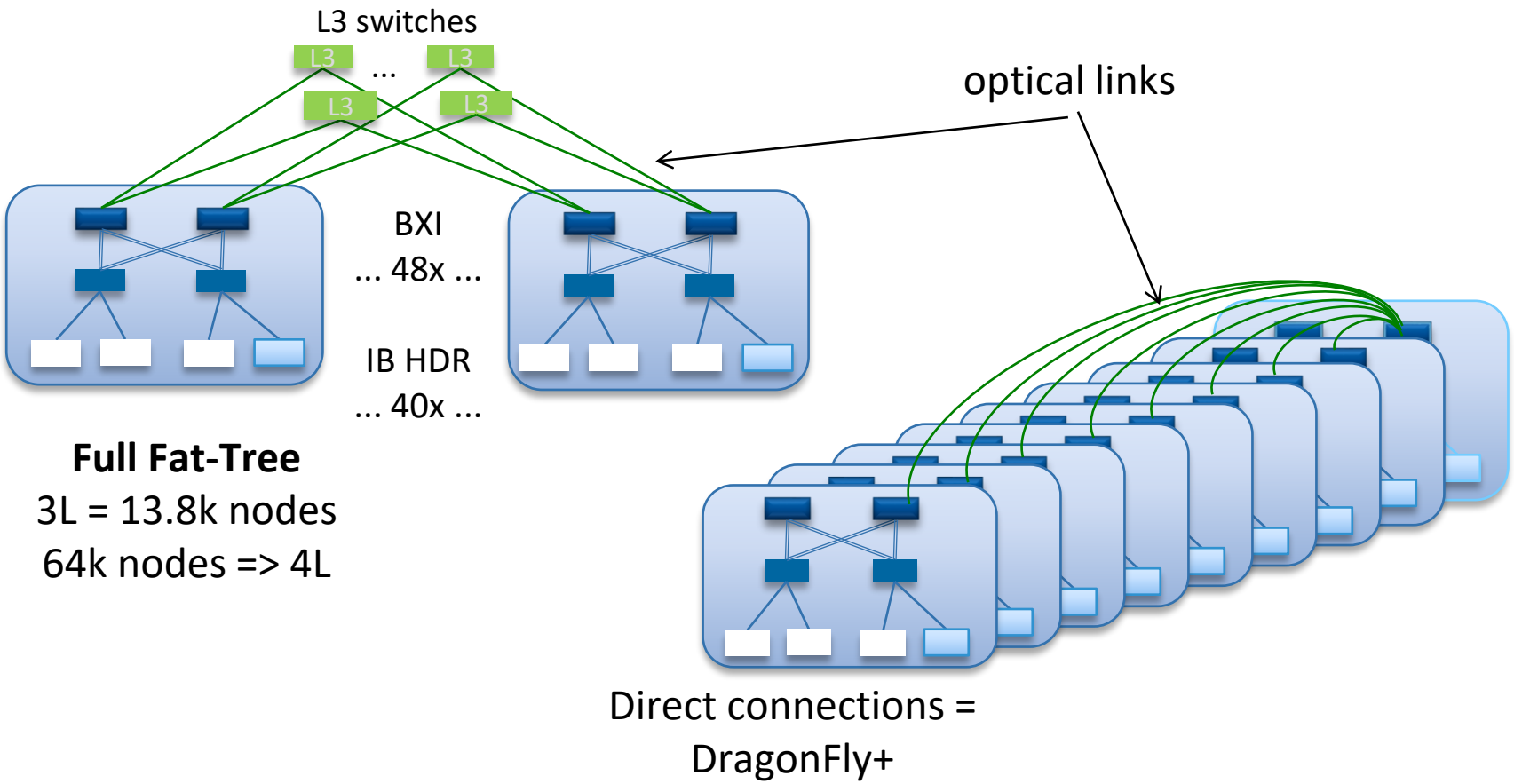
- 3 racks
- 288 Compute Nodes +18-36 IO
- 18 HDR switches
- **16** CN/SW

EVIDEN JÜLICH Forschungszentrum

BullSequana XH2000 Networking

Best in class Interconnect flexibility

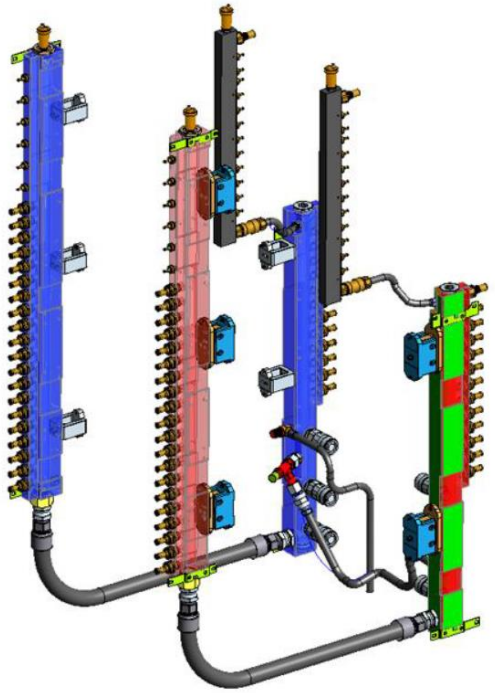


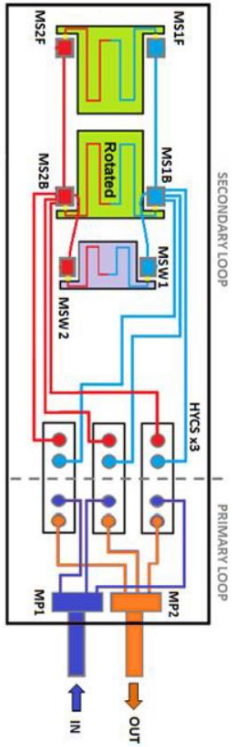


BullSequana XH2000 Cooling

Optimized Power Usage Effectiveness

BullSequana XH2000: > 95% cooling efficiency

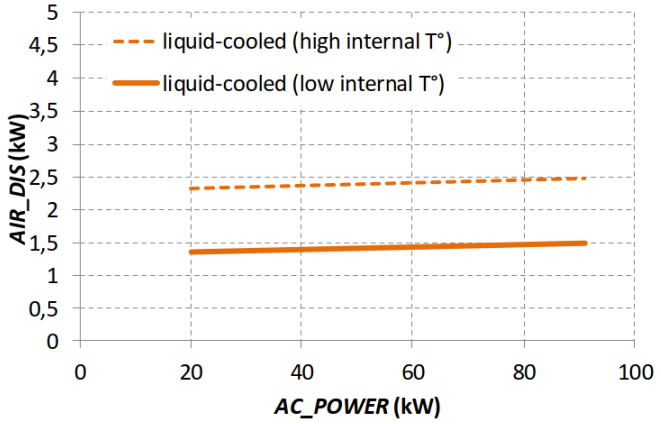


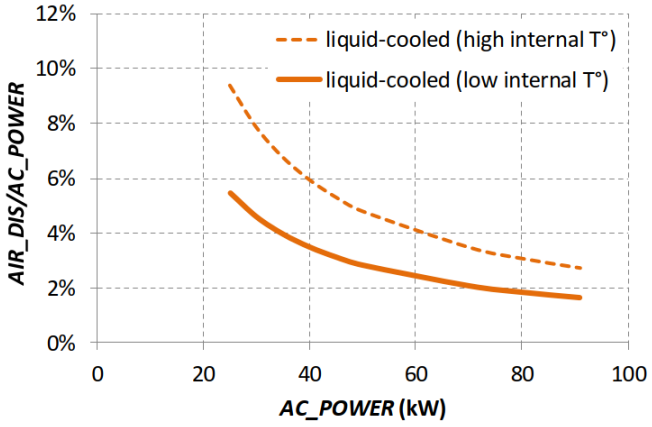


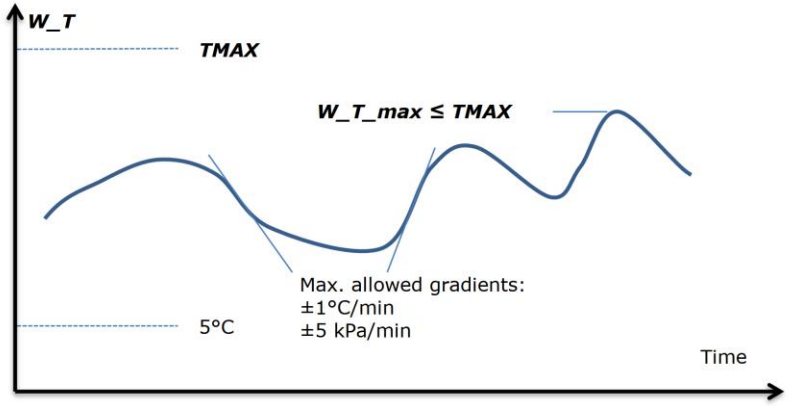
Fan less architecture :

- Warm water up to 40°C (104°F) inlet
- Heat rejected in air is almost constant
  - Pumps, radiation and normal convection ~1,5 kW / rack to 2,5kW / rack
  - DC power heat rejection : 0,5% of the power consumption

- 2 modes of operation: low & high internal temperature







EVIDEN JÜLICH Forschungszentrum

Bull Sequana XH2000 (JUWELS Booster, JURECA-DC)

Lessons Learned

PRO:

- **Modular system platform**
- **Blade system**
- **Multiple blade types**
- **Switch blade based on standard technology**
- **Rack (96 nodes) as technological building block**
- **Cell Concept as logical building block**
- **Direct Liquid Cooling (fanless rack)**
- **Up to 40°C warm water as inlet temperature (free cooling)**
- **All in one approach (Compute, interconnect, power, cooling)**

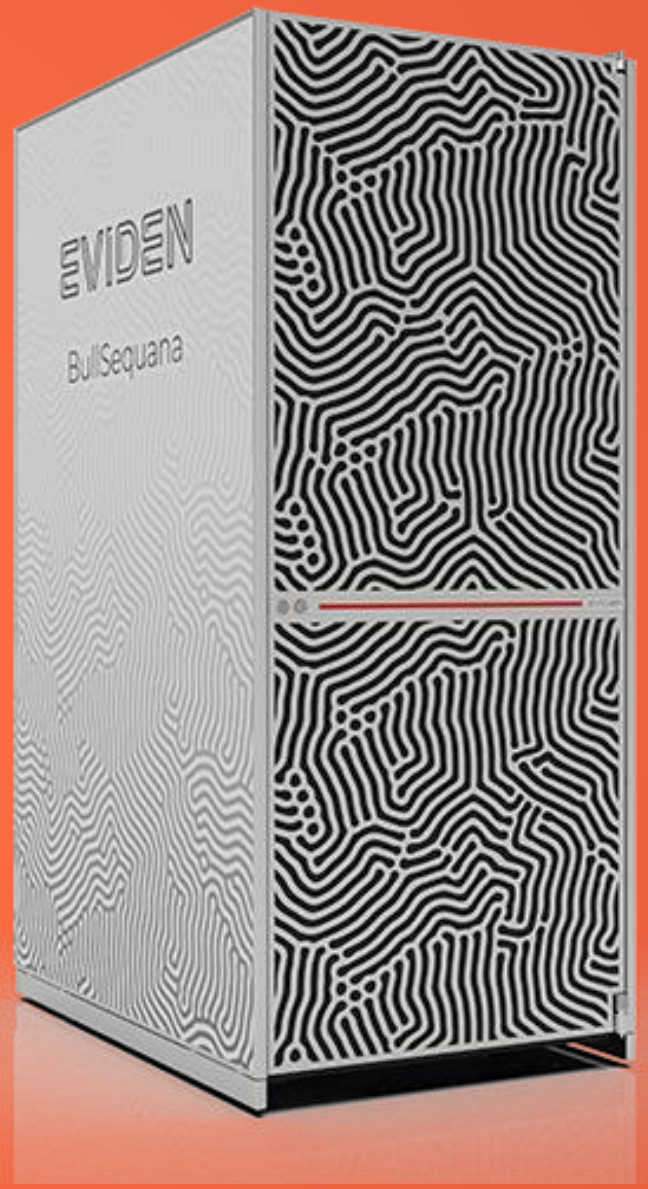
CON:

- **90kW power not sufficient for future technology**
- **Using Midplane for the high speed interconnect can be a limitation**
- **Different form factor for compute and switch blades can be a limitation – or leads to ineffective use of rack space**

EVIDEN | JÜLICH Forschungszentrum

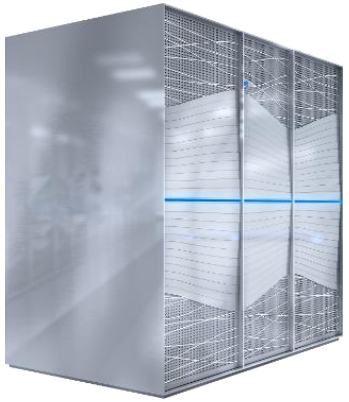
The Future (Now):

Sequana 3 – XH 3000



BullSequana X high-end platform evolution

Relentless pursuit to bring more performance and flexibility to our customers



BullSequana X1000

- 2nd generation DLC
- 40°C inlet water temperature support
- Increased density
- Designed to better scale to large Petascale systems
- Support of InfiniBand HDR high-speed interconnect technology

BullSequana XH2000

- 3rd generation DLC with introduction of DLC PSUs
- “All-In-One” Rack form factor to support smaller systems all the way up to Exascale systems
- Increased flexibility of compute and interconnect technologies supported
- Support of InfiniBand HDR high-speed interconnect technology

BullSequana XH3000

- 4th generation DLC
- Substantial increase of power and cooling envelope
- Increased flexibility of compute and interconnect technologies supported
- Support of InfiniBand NDR high-speed interconnect technology
- Standardized design to support OpenSequana program

BullSequana XH3000 Infrastructure overview

Architecture Overview

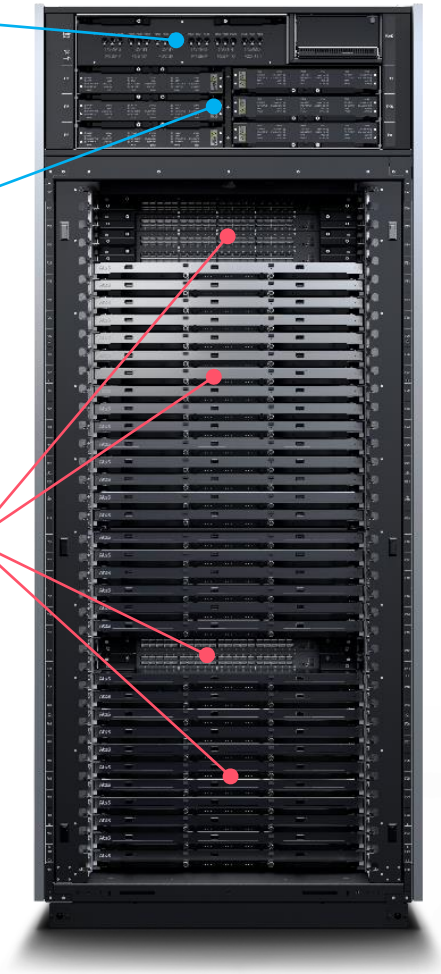
1x Front Compute PDU

Power Group

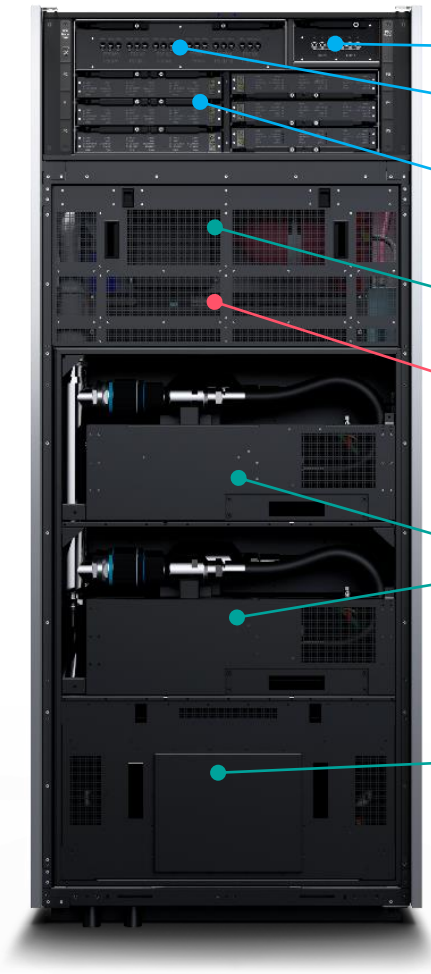
6x Front DLC PSU Shelves

DLC ready Compute Rack

38x standardized slots for Compute & Switch blades



Front view



Rear view

1x Hydraulic PDU

1x Rear Compute PDU

6x Rear DLC PSU Shelves

1x Expansion Tank

2x Management switches

2x Hydraulic Pumps

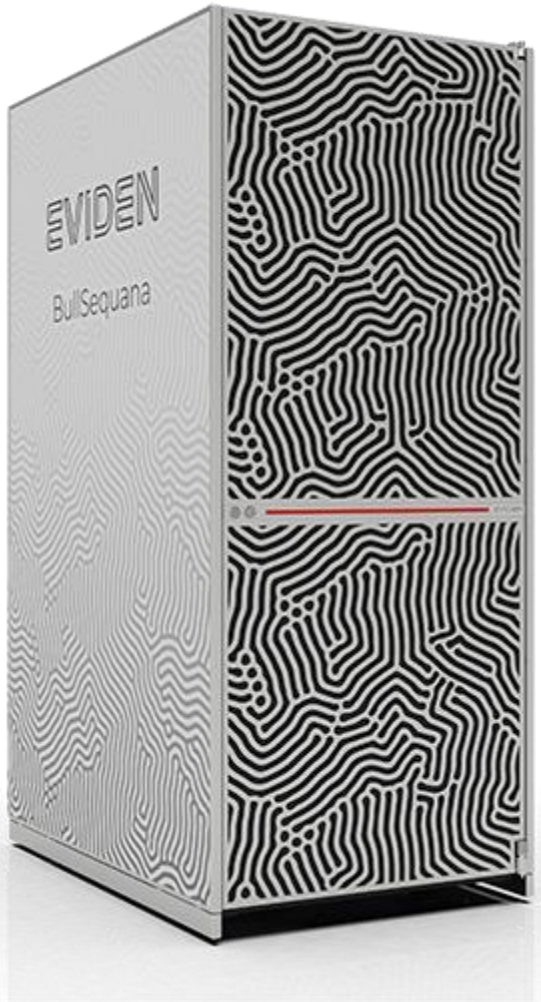
Hydraulic Group

1x Heat Exchanger

EVIDEN | JÜLICH Forschungszentrum

BullSequana XH3000 Cooling

Optimized Power Usage Effectiveness



BullSequana XH3000: >97% Warm Water Cooled

Fan less architecture :

- Warm water up to 40°C inlet
- Heat rejected in air is almost constant
  - Pumps, radiation and normal convection ~1,5 kW / rack to 2,5kW / rack
  - DC power heat rejection : 0,3%-0,5% of the power consumption

Full rack running linpack : 120 kW

- 97% efficiency at Low Internal temperature: 3,6 kW Air dissipation
- 95% efficiency at High Internal temperature: 6 kW Air dissipation

Full rack 1/3 of load : 40 kW

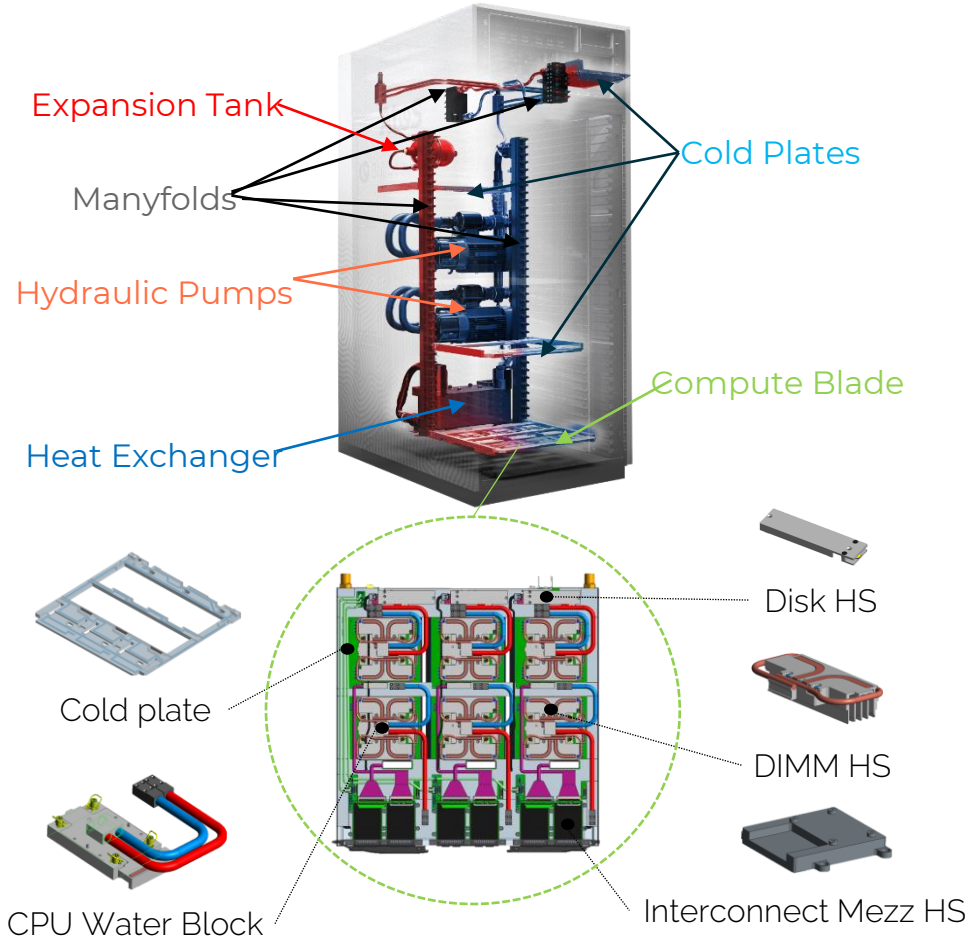
- 92% efficiency at Low Internal temperature: 3,2 kW Air dissipation
- 87% efficiency at High Internal temperature: 5,2 kW Air dissipation

BullSequana XH3000 Infrastructure

Hydraulic architecture overview

Hydraulic architecture is composed of several elements:

- That are part of the rack:
  - 2 hydraulic pump modules managed by 2 HMCs
  - 1 common heat exchanger with 2 primary valves
  - 2 sets of manifolds:
  - One for compute, switch and administration blades
  - One for power shelves
  - 1 expansion tank
- That are part of the blades:
  - Water blocks for CPU/GPU cooling in compute blades
  - Heat spreaders for DIMM, Interconnect mezzanine and disk in compute blades
  - Cold plates for other motherboard components in compute, switch and administration blades



BullSequana XH3000 Power group

Power architecture overview

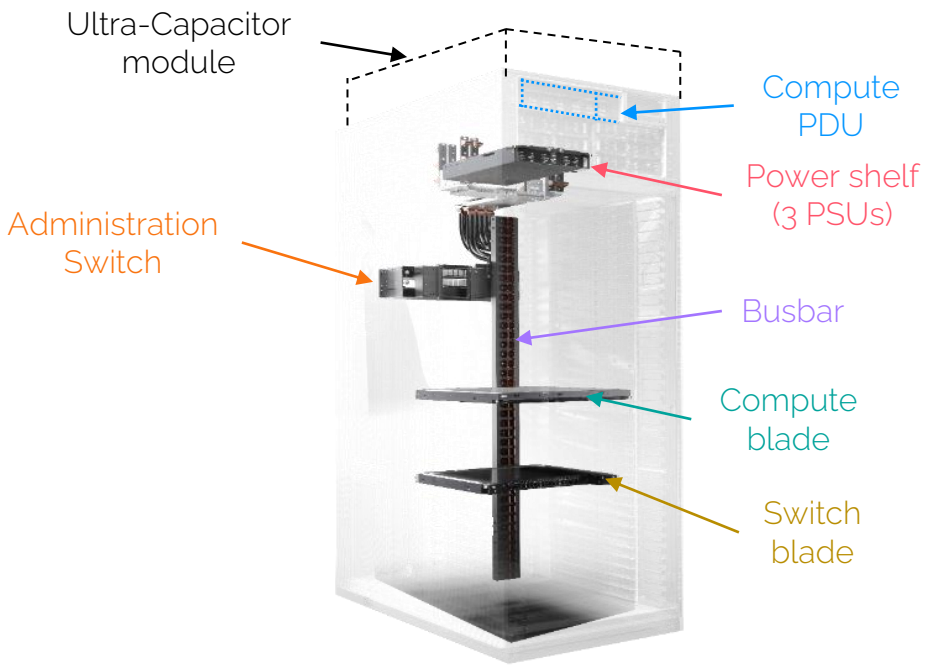
Power architecture is composed of several elements:

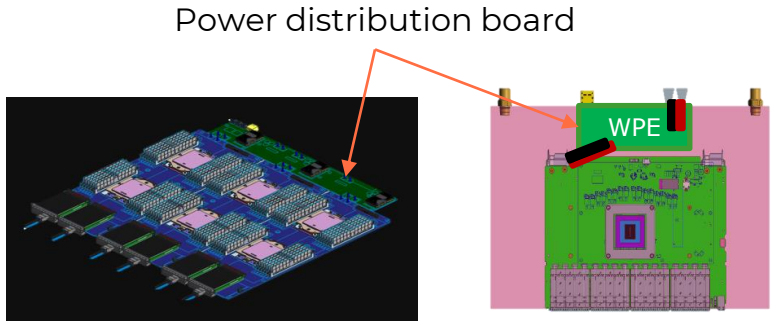
- Power Group P3G mounted on top of the rack:
  - Power Distribution Units (PDU): 2 compute PDUs (one at the front, one at the rear of the rack), 1 hydraulic PDU (at the front of the rack)
  - Power Shelves. One shelf contains 3 Titanium Power Supply Units of 4,2kW each. Max of 12 Power Shelves per rack (147kW + 4,2kW redundancy)
- Power distribution busbar inside the rack
- Power distribution board inside each blade
- Ultra-capacitor module mounted on top of P3G

Power shelves and blades are Direct Liquid Cooled and "hot-plug"

Possibility to assemble Power group at customer site

- Standard is assembled at factory
- When height constraint in data center below 2,25m



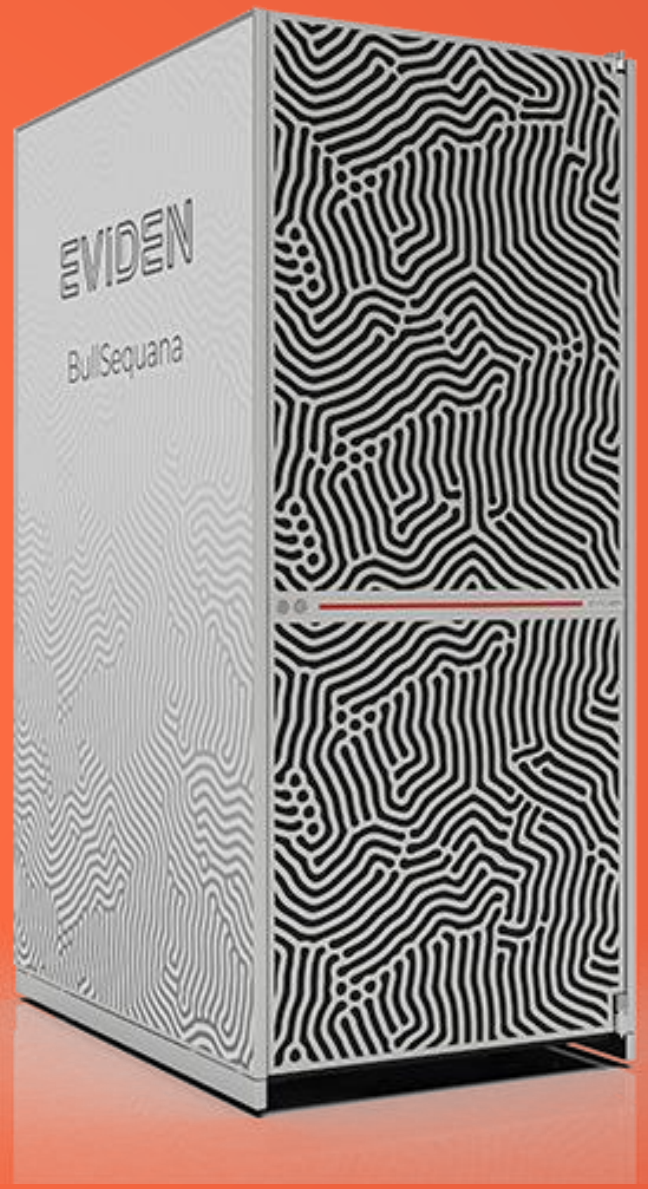


EVIDEN | JÜLICH Forschungszentrum

First to the future

JUPITER (ExaScale)

System Architecture



System Solution

JUPITER: 1st European ExaScale System

> 6000 Compute Nodes

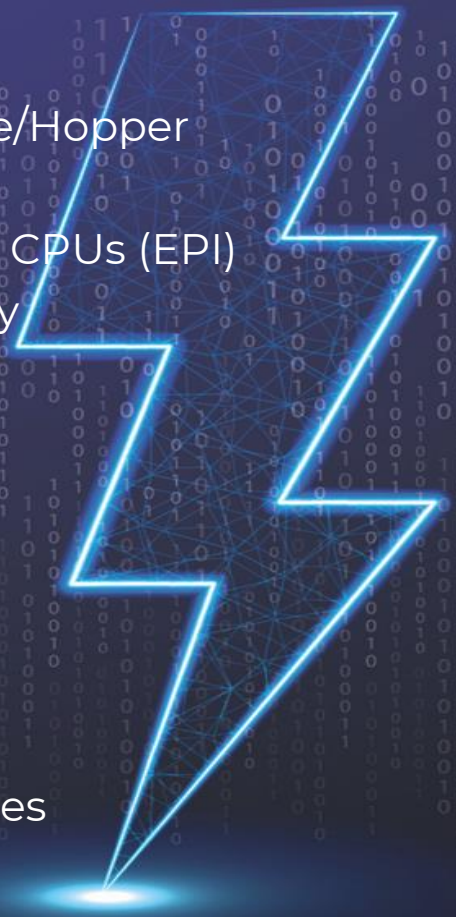
- > 5.000 GPU nodes
- > 20.000 Nvidia Grace/Hopper
- > 1.000 CPU Nodes
- > 2.000 Sipearl Rhea1 CPUs (EPI)
- > 14 PB main memory

Flash Storage

- > 20 PB
- > 2 TB/s Bandwidth

Service Nodes

- Login Nodes
- Admin & Service Nodes



High Speed Interconnect

- NDR Infiniband – Fully non blocking

Footprint

- 25 Sequana Cells (5x XH3000 cabinet)
- 5 Standard Racks (Service & Flash Storage)

Energy Efficiency

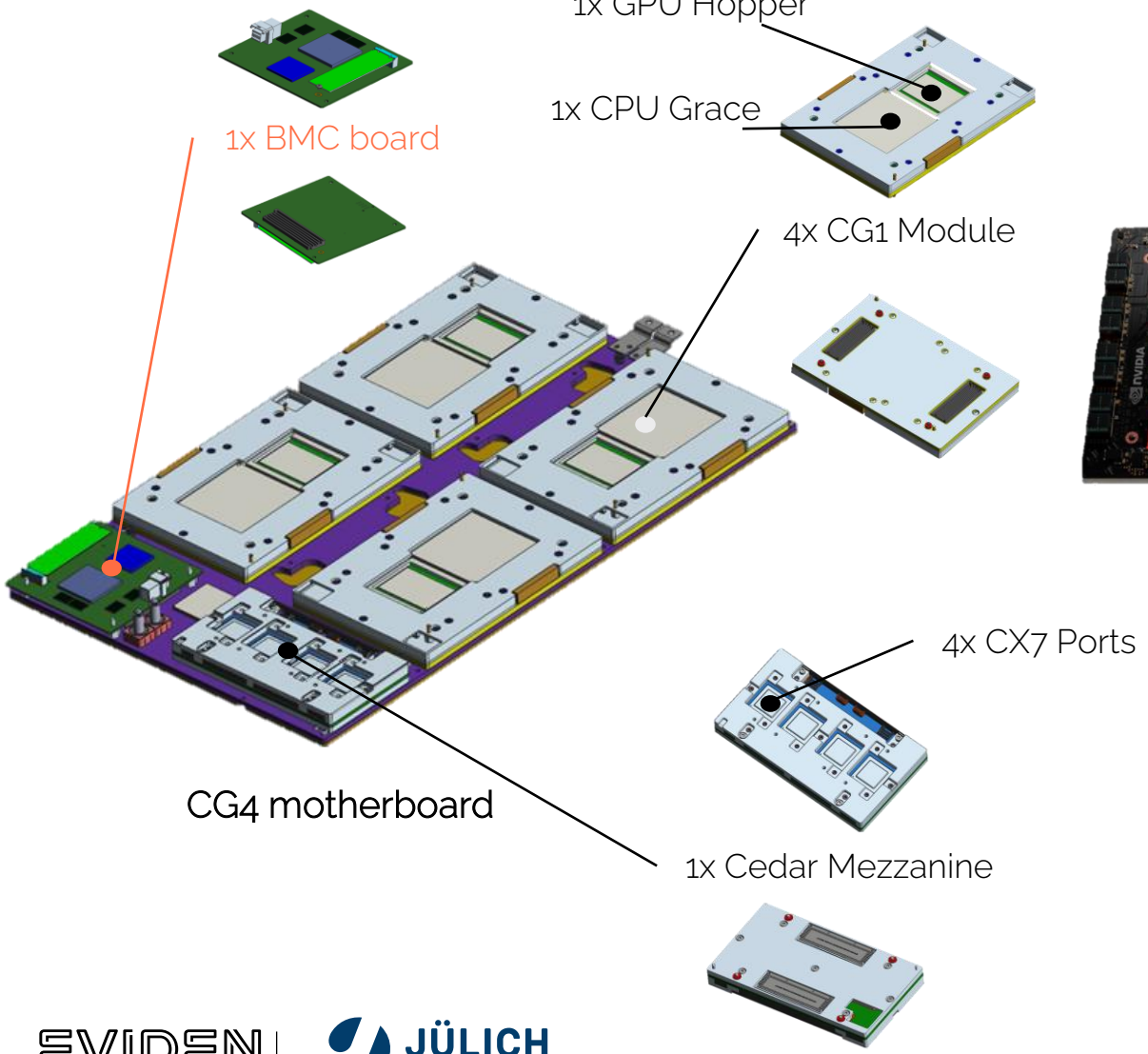
- PUE factor of 1,03
- Enhanced DLC – up to 40°C for inlet water and 97% DLC efficiency

Performance: 1 EFlop/s HPL

BullSequana X3515-HMQ Grace-Hopper Blade

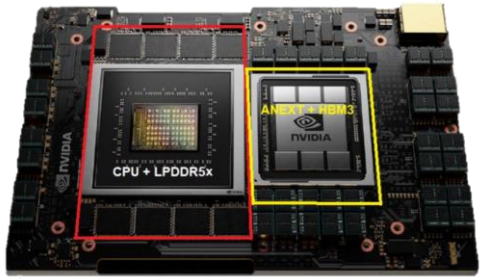
First to the future

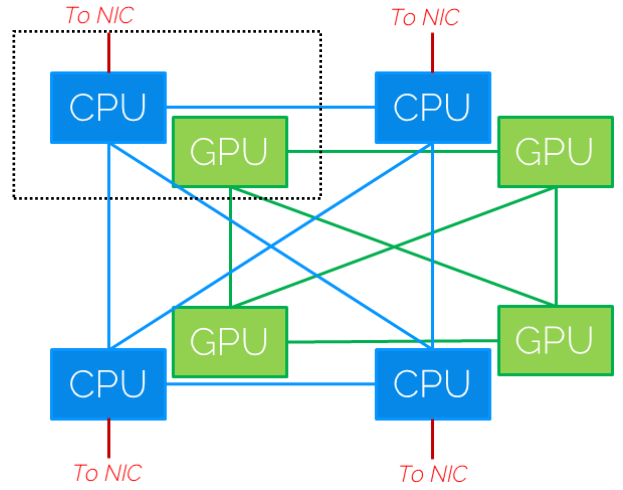
GPU Node



IN DEVELOPMENT

CG1 module



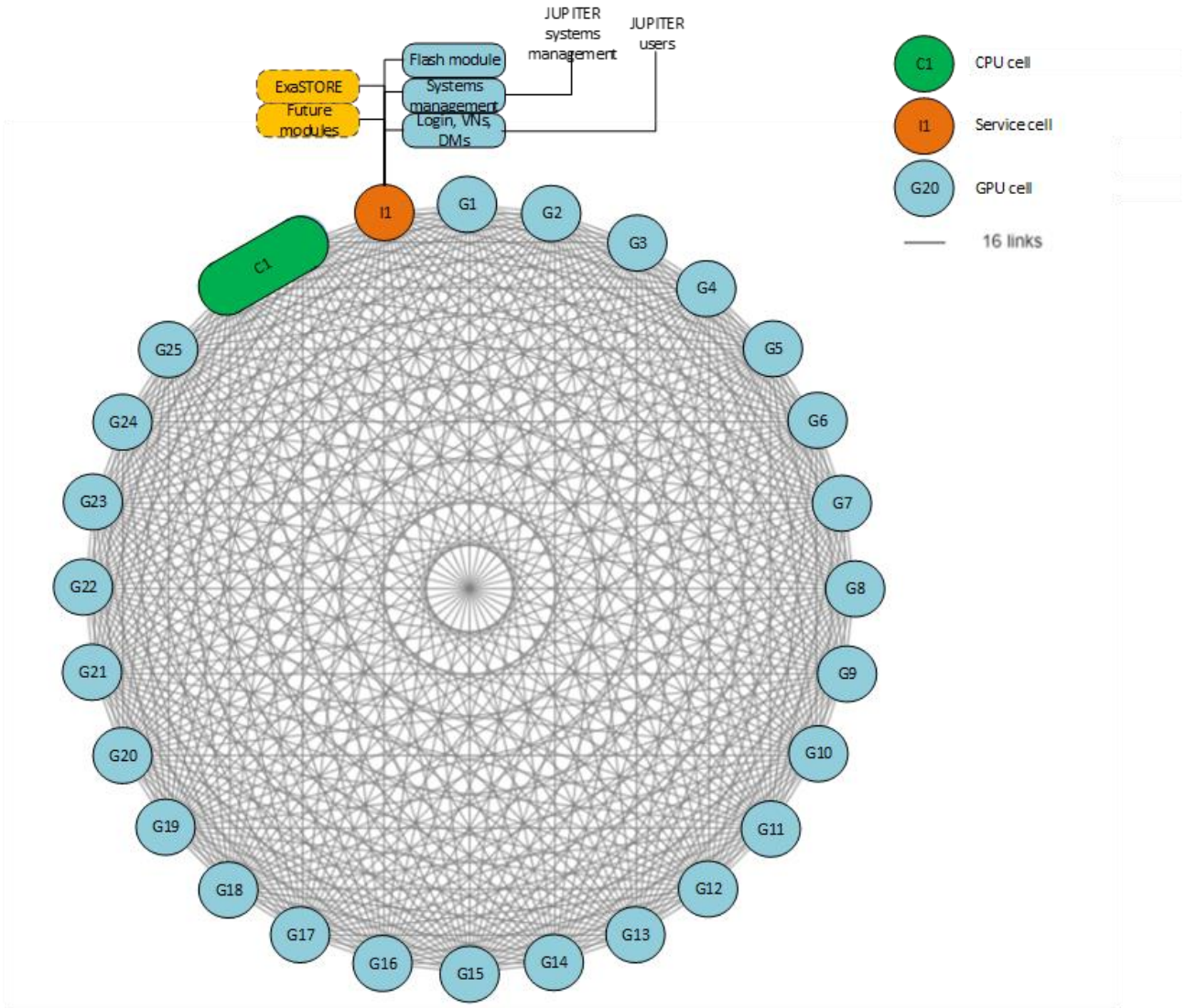


- All-to-all NVLink between CPU (C-link)
- All-to-all NVLink between GPU (G-link)
- Coherent memory space
- GPU direct access to NIC

EVIDEN | JÜLICH Forschungszentrum

DragonFly+ with Cluster-Cell and I/O-Cell

First to the future



IO Cell contains:

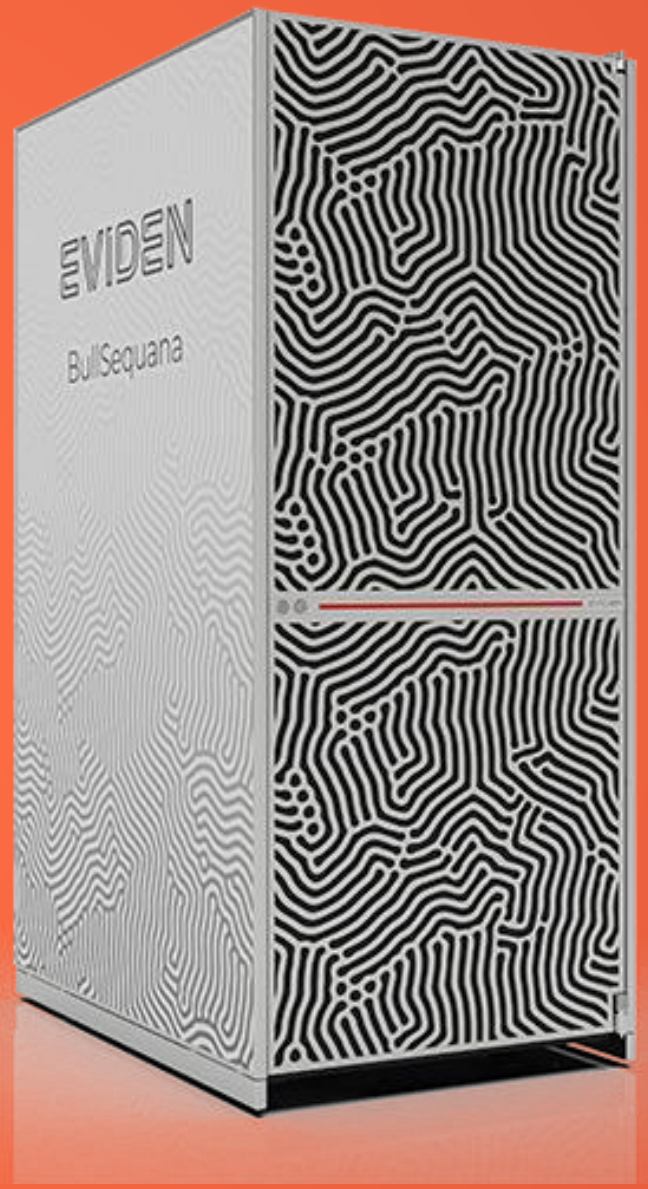
- Flash storage
- Management nodes
- Login nodes
- Any other peripheral nodes which may be needed such as pre- post- processing or visualization

- (ExaStore Storage)
- (Future Modules)

EVIDEN JÜLICH Forschungszentrum

EVIDEN | JÜLICH Forschungszentrum

Questions?



EVIDEN

Thank you!

For more information please contact:
Dominik Friedrich
HPC Solution Architect
Email: dominik.friedrich@eviden.com