Evolution of the Sequana System Architecture

The Past, the Present and the Future

Dr. Thomas Warschko – Technical Director Extreme Computing Germany

Introduction to Supercomputing at JSC – Theory & Practice
Jülich, November 23rd 2023

Nov 2023

an atos business

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

05

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 (54C DC)
  - Free Cooling & Heat reuse

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

EVIDEN | JÜLICH Forschungszentrum

The Past:

Sequana 1 - X1000

Bull Sequana X1000

the Bull exascale generation of supercomputer

- **Open and modular platform designed for the long-term**
    - To preserve customer investments
    - To integrate current and future technologies
    - Multiple compute nodes: Xeon-EP, Xeon Phi, Nvidia GPUs, other architectures…
- **Scales up to tens of thousands of nodes**
    - Large building blocks to facilitate scaling
    - Large systems with DLC: 250-64k nodes
- **Embedding the fastest interconnects**
    - Multiple Interconnects: BXI, InfiniBand EDR/HDR
    - Optimized interconnect topology for large basic cell / DLC (288 nodes)
    - Fully non-blocking within Cell
- **Ultra-energy efficient**
    - Enhanced DLC – up to 40°C for inlet water and ~100% DLC

2015-2019

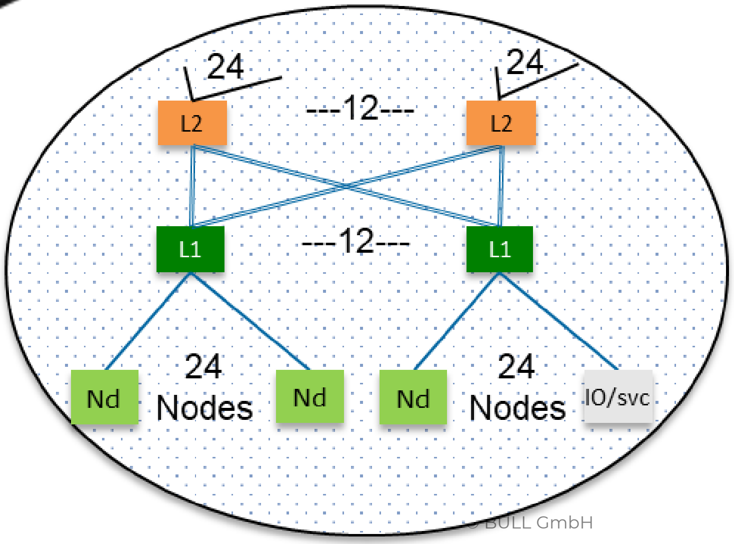
Bull Sequana X1000 cell technology

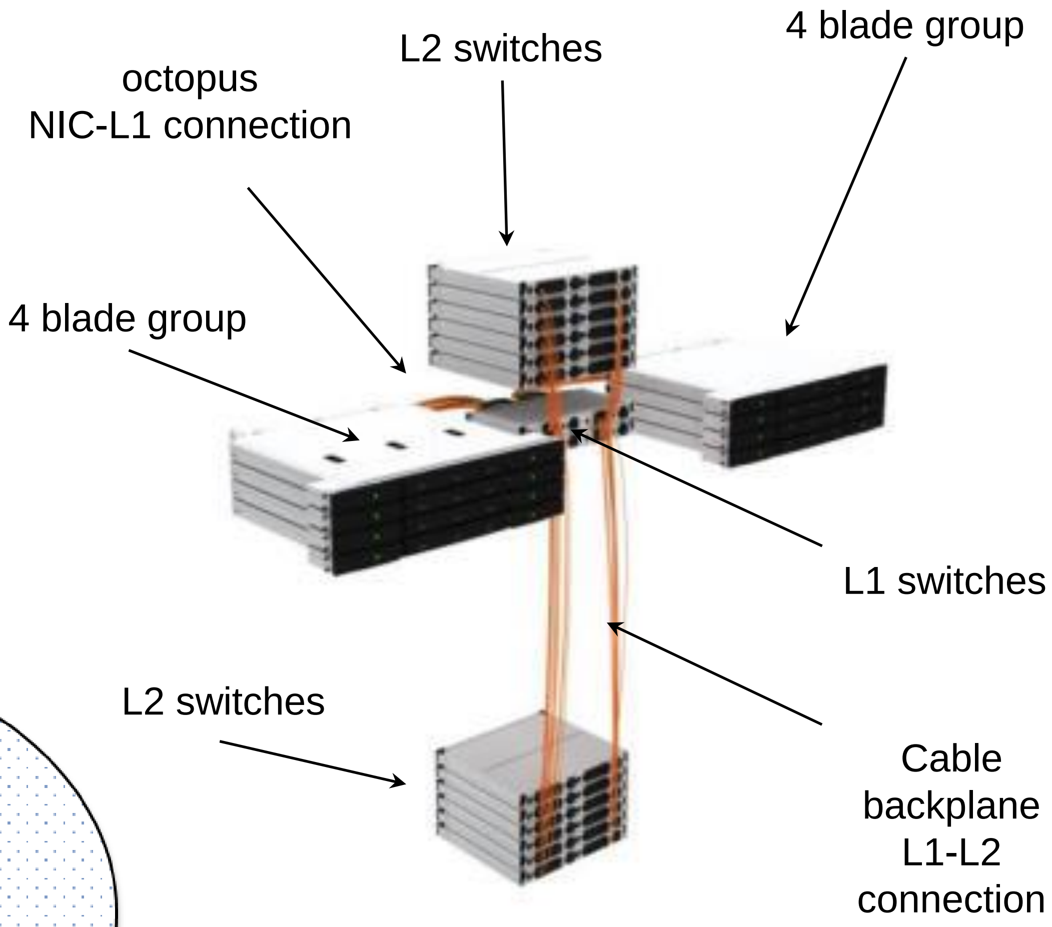
Bull Sequana X1000 cell

support several types and generations of compute nodes either with conventional processors or with accelerators

- **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

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)**

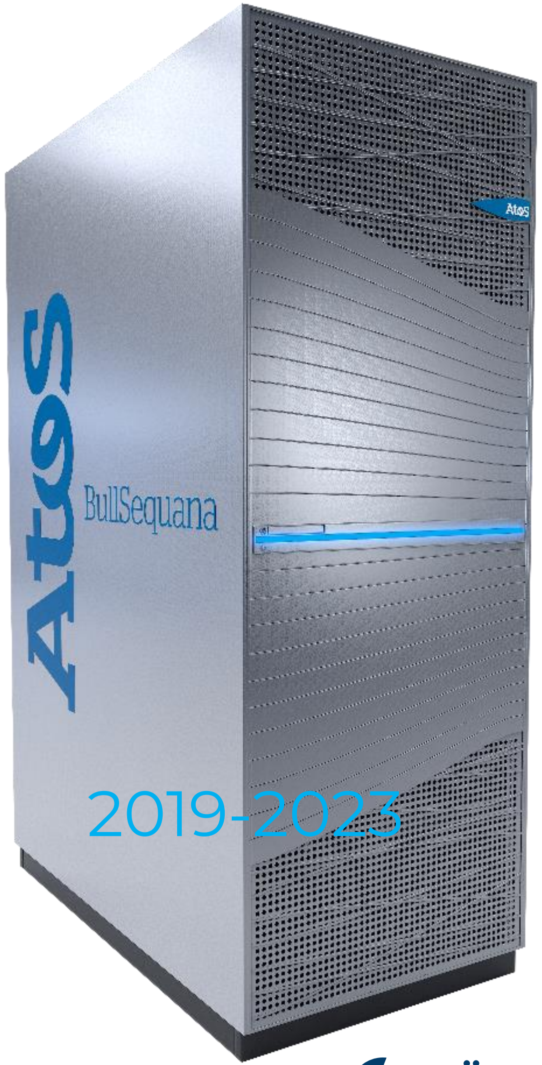
2015-2019

The Present:

Sequana 2 – XH 2000

What is BullSequana XH2000?

A new generation of Sequana X1000



2019-2023

- **XH2000 is not a new machine, it is a natural evolution of X1000:**
  - XH2000 is compatible with existing and future blades
  - XH2000 reuse as much as possible X1000 components in order to protect Atos investments
  - XH2000 will be able to scale to Exascale
- **XH2000 leads to cost optimization**
- **XH2000 embeds new features:**
  - XH2000 introduces support for new technologies such as Mellanox HDR, new fabric topologies, new pruning ratios, Fast Ethernet
  - XH2000 improves infrastructure costs by at least 10% compared to X1000
  - XH2000 provides access to new markets:
    - Entry level configurations
    - Configurations up to 800 nodes should be installed (SW) in less than 3 days
  - XH2000 provides optional redundancy features (compared to X1000 where they are embedded)

BullSequana XH2000 Overview

One 42U cabinet with:

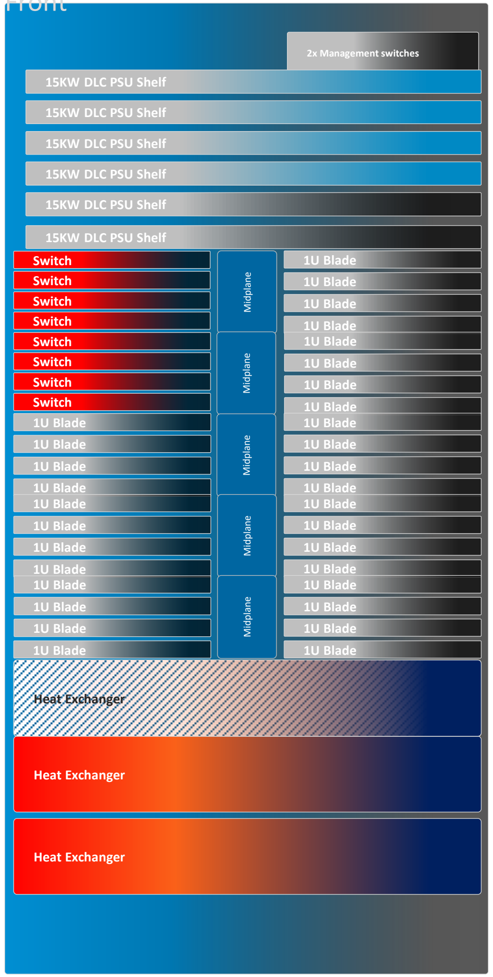
- up to 32 compute DLC blades / 96 compute nodes
  - 20 on front side, 12 on rear side
- up to 6 liquid-cooled PSU shelves (up to 30 liquid cooled PSUs)
- fanless design
- 2 HYC, optional 3rd HYC for 2+1 redundancy
- 2 Leaf Ethernet modules
- Up to 10 Interconnect DLC Switches
  - HDR100 & HDR200 in Phase 1
  - BXI and Fast Ethernet in Phase 2
- 1 Power distribution unit with 3x 63A tri-phase cables
- Power and signal connections at the top of rack

Power and cooling capacity: 15 to 90kW

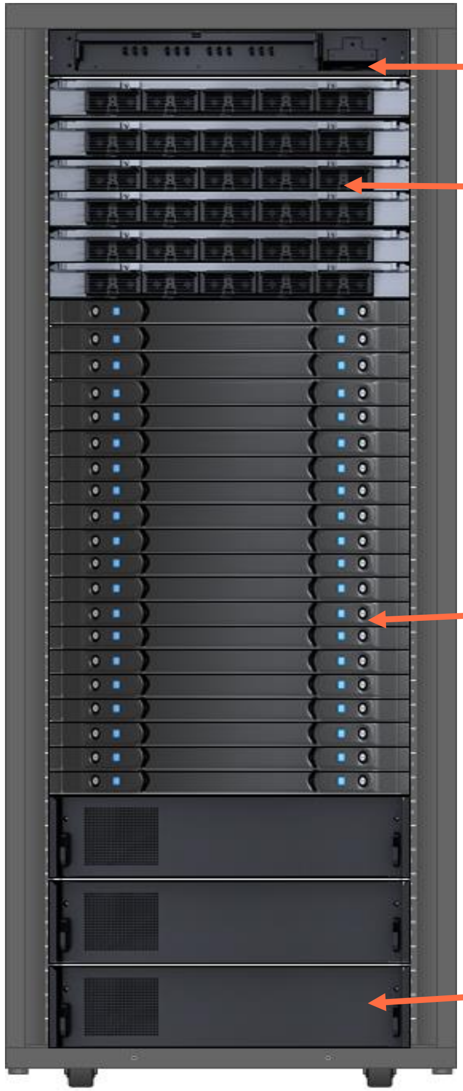
BullSequana XH2000 Details

Side View

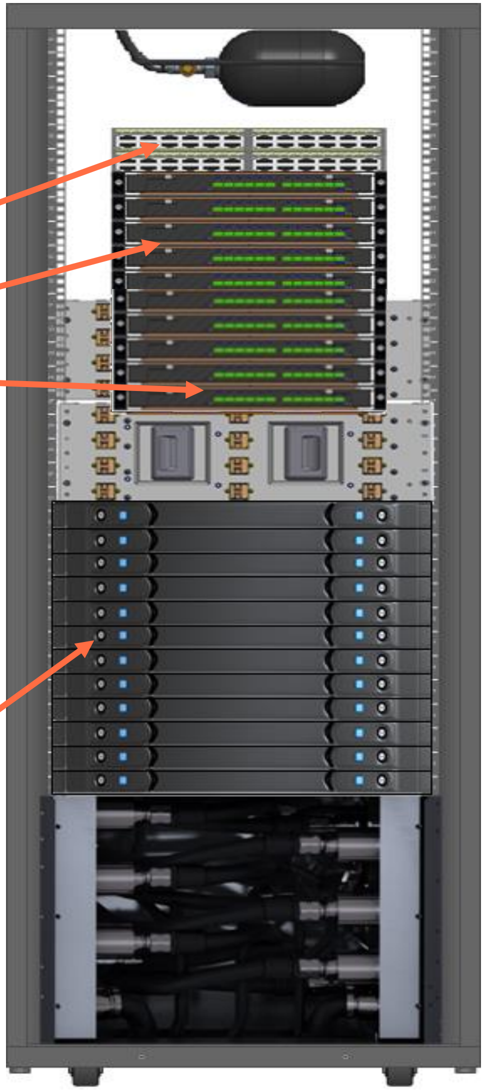
42U Cabinet
Back
Front



Front View



Rear 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

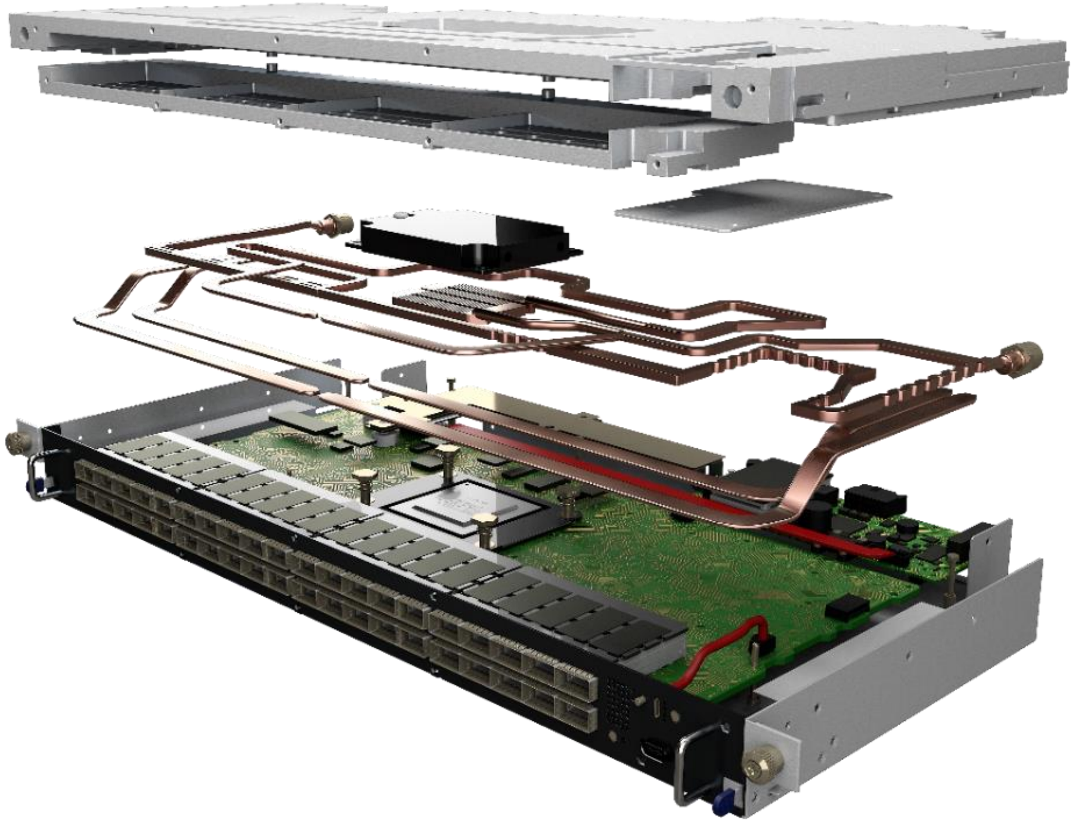
up to 3 Hydraulic chassis
(2+1 optional redundancy)

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

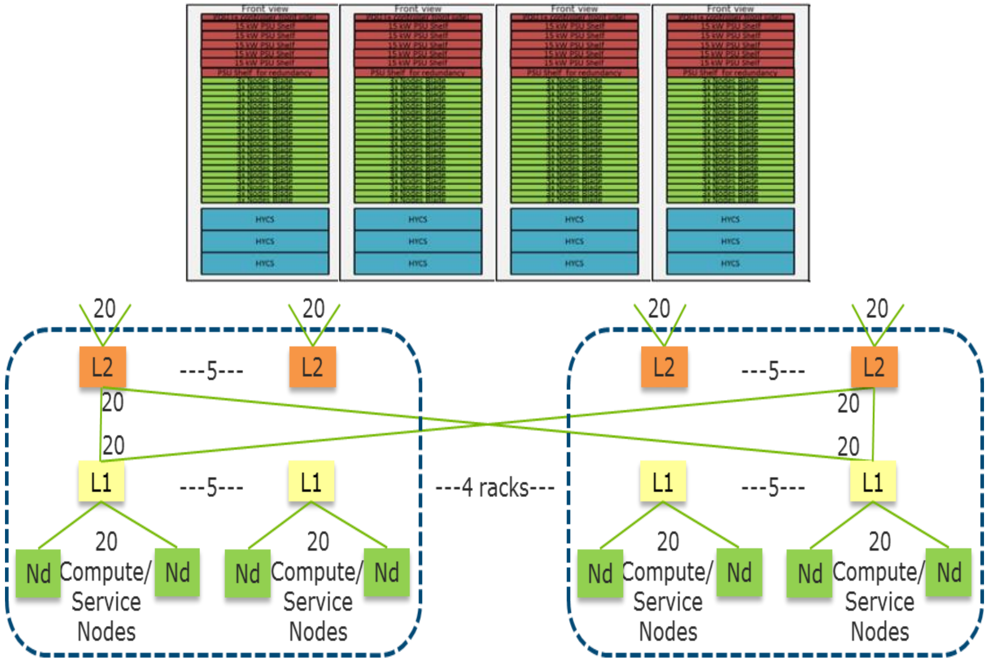


HDR Flexible Sideplane

- 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)

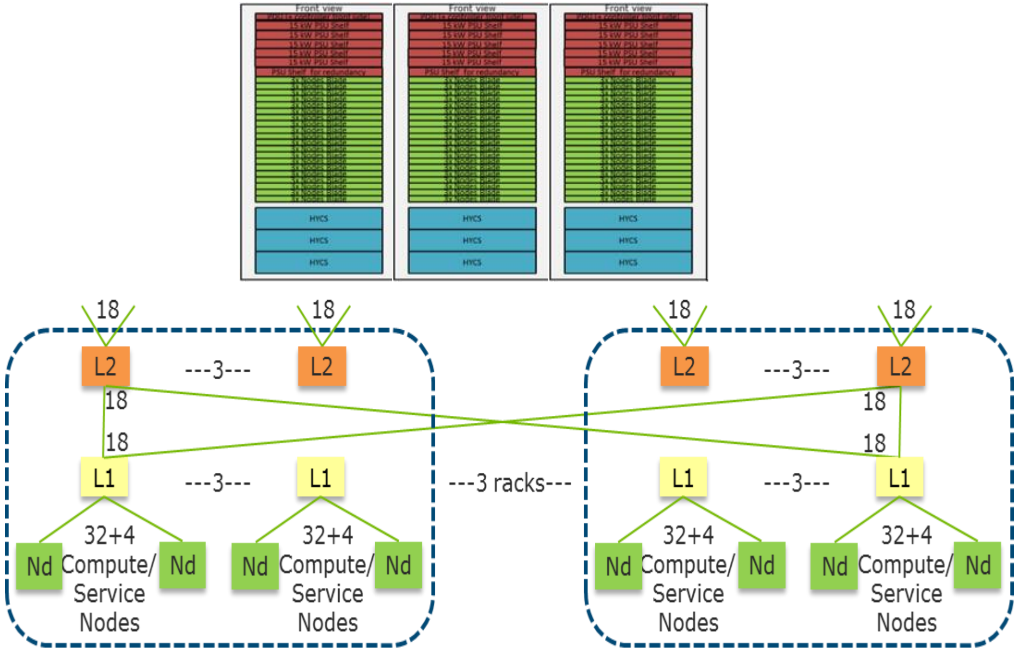
BullSequana XH2000 Topology options

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



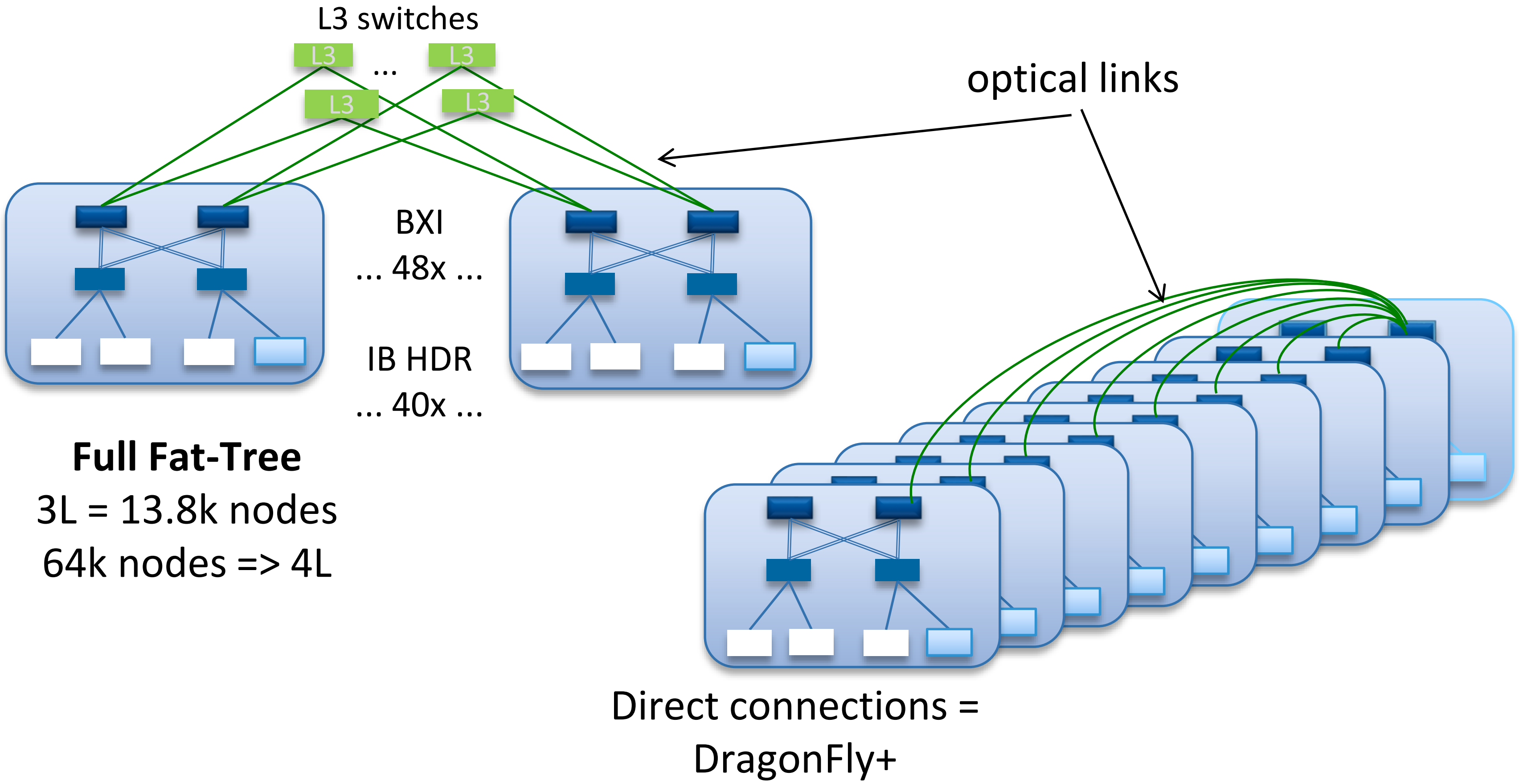
Single Sequana Cell: **HDR100 FT 1:1**

- 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



L3 switches

optical links

BXI
... 48x ...

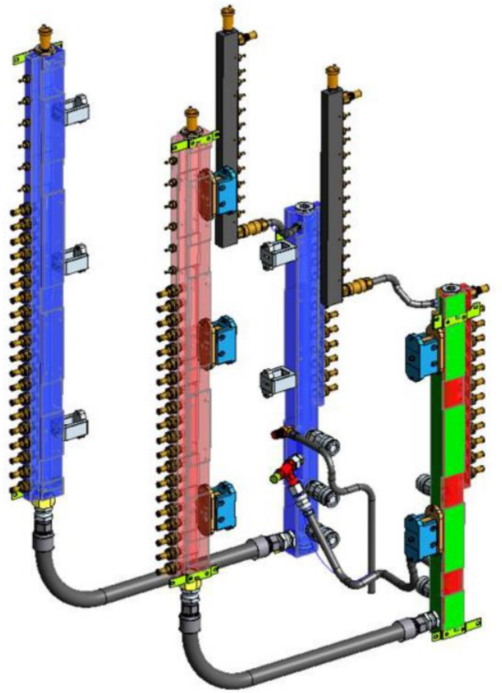
IB HDR
... 40x ...

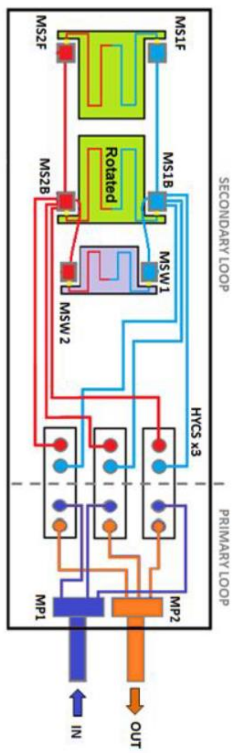
Full Fat-Tree
3L = 13.8k nodes
64k nodes => 4L

Direct connections =
DragonFly+

BullSequana XH2000 Cooling

Optimized Power Usage Effectiveness



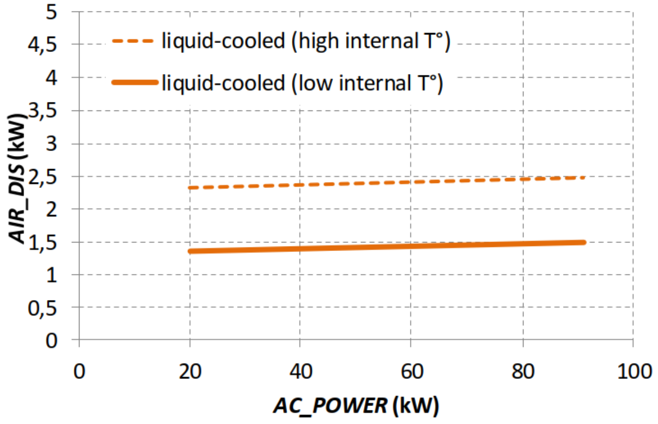


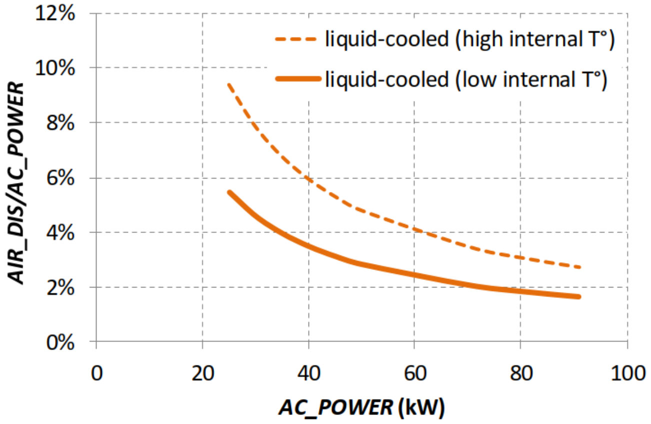
BullSequana XH2000: > 95% cooling efficiency

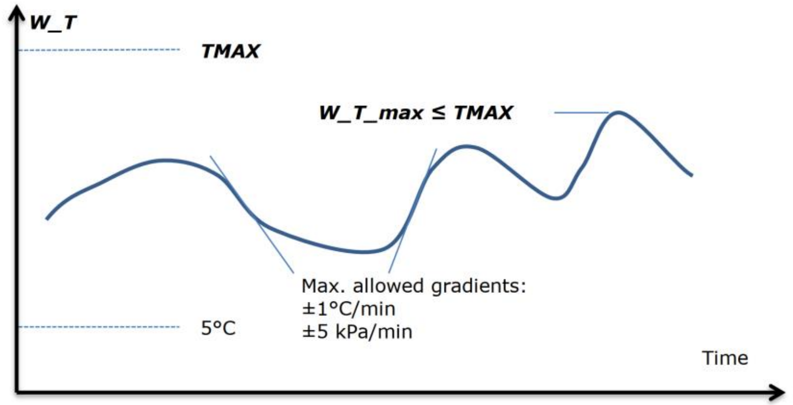
Fan less architecture :

- 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







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

EVIDEN | JÜLICH Forschungszentrum

BullSequana XH3000

A fully integrated Direct Liquid Cooled (DLC) custom platform

- BullSequana XH3000 is an Atos custom designed platform that integrates:
  - DLC ready infrastructure with power and cooling distribution,
  - DLC compute nodes (or servers),
  - DLC high-speed interconnect switches with high-speed cabling, and
  - DLC administration switches

BullSequana XH3000
Full DLC platform

=

Custom DLC Rack infrastructure to distribute power & cooling to all the elements hosted within the rack

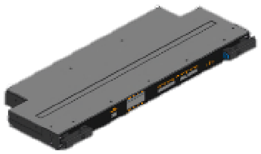
+

Custom DLC 1U Compute blades to provide compute processing power.

+

Custom DLC 1U High-Speed Interconnect switches & cabling to provide a high-speed network to exchange data between compute blades

+

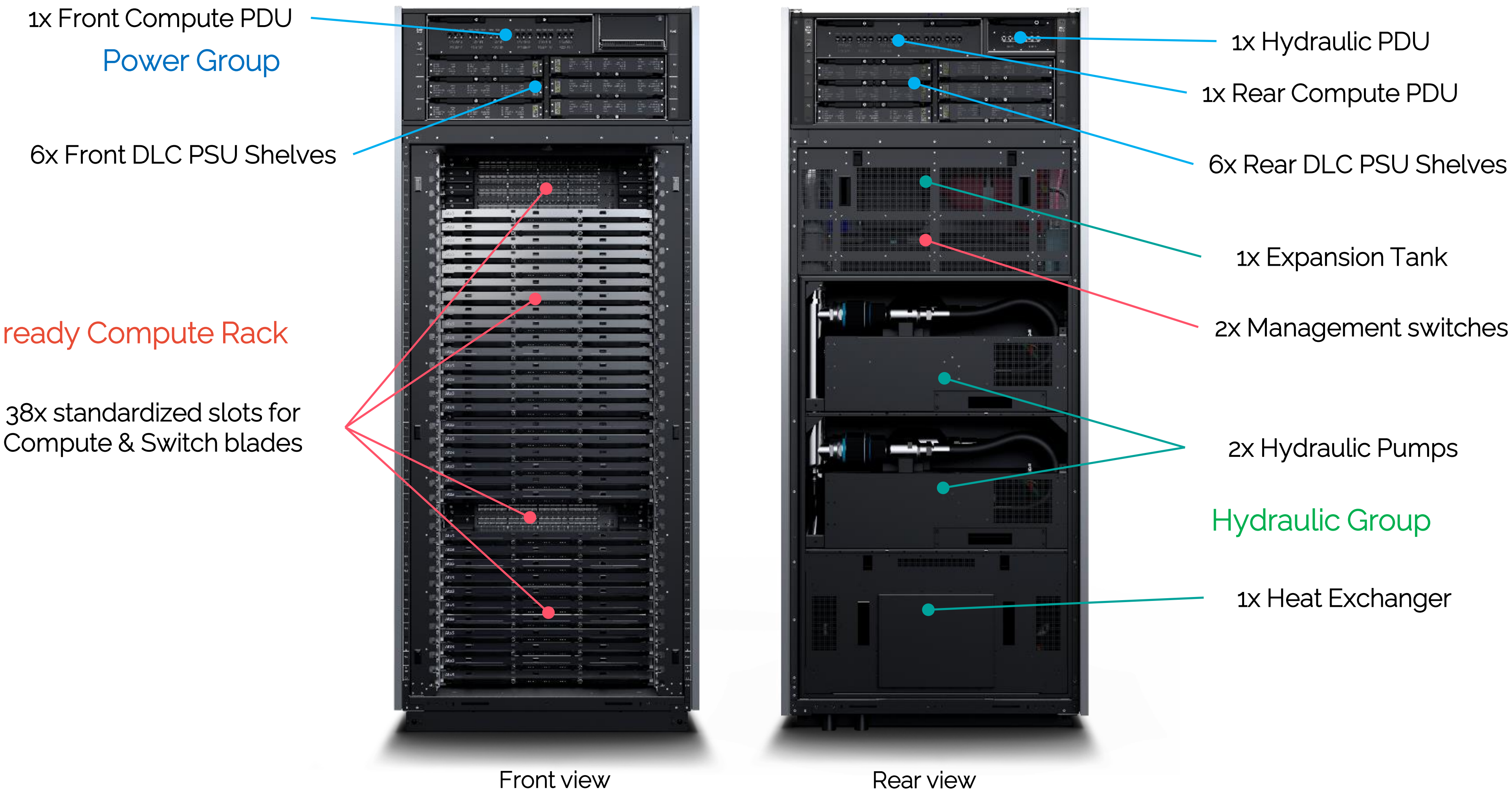


Custom DLC 1U Administration switches & cables to manage the different elements hosted within the rack

- All components within the rack are DLC with warm water up to 40°C to provide maximum performance, density and the lowest Total Cost of Ownership possible

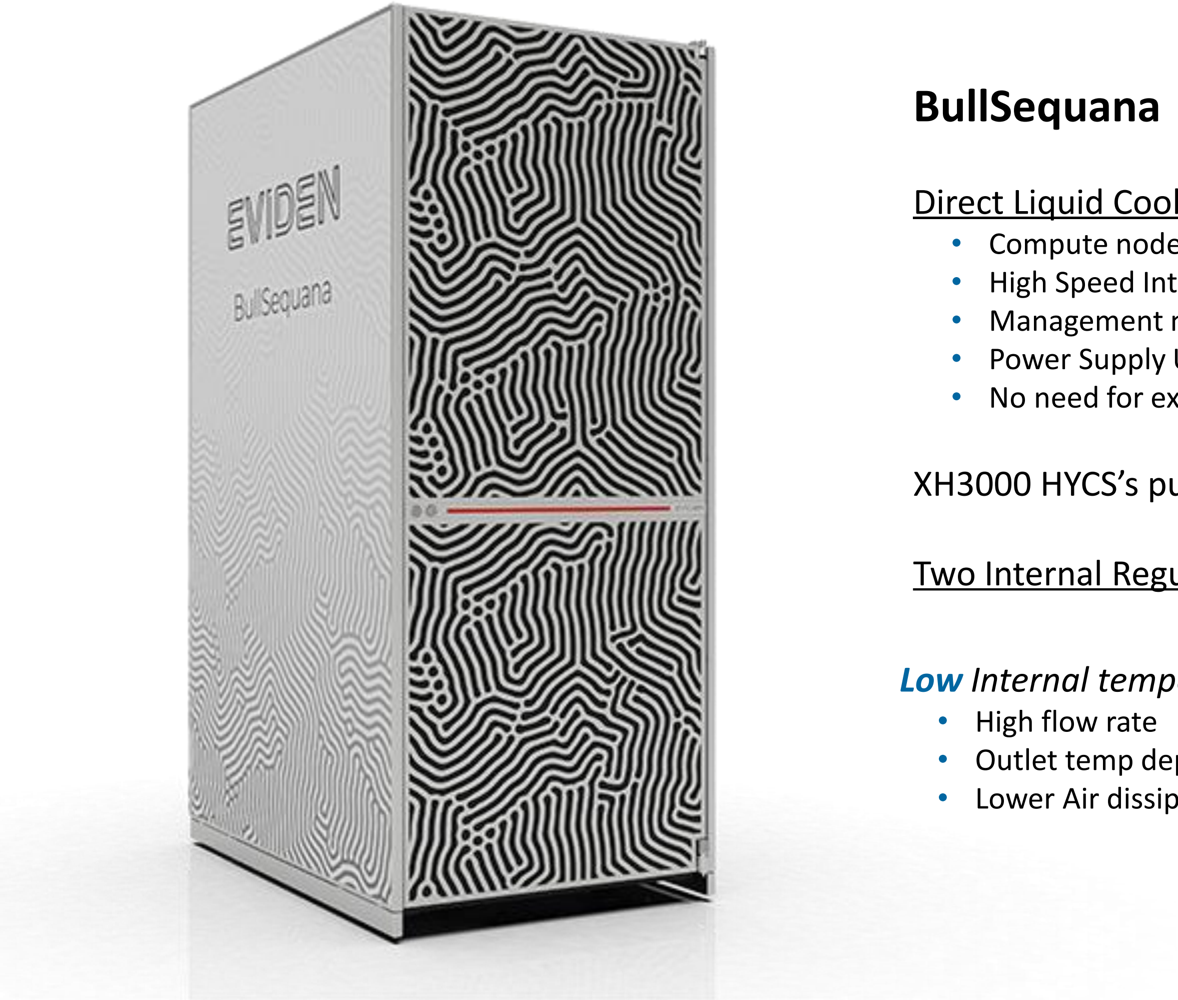
BullSequana XH3000 Infrastructure overview

Architecture Overview



BullSequana XH3000 Cooling

Optimized Power Usage Effectiveness



BullSequana XH3000: a fanless innovative cooling solution

Direct Liquid Cooling :

- Compute nodes (CPU, Memory, Drives, GPU)
- High Speed Interconnect: HDR, BXI & High-Speed Ethernet switches
- Management network: Intra Rack management switches
- Power Supply Unit: DLC PSU shelves
- No need for external CDU, they are integrated and redundant

XH3000 HYCS's pumps are not liquid cooled but are still fanless

Two Internal Regulation Temperature modes:

***Low** Internal temperature*

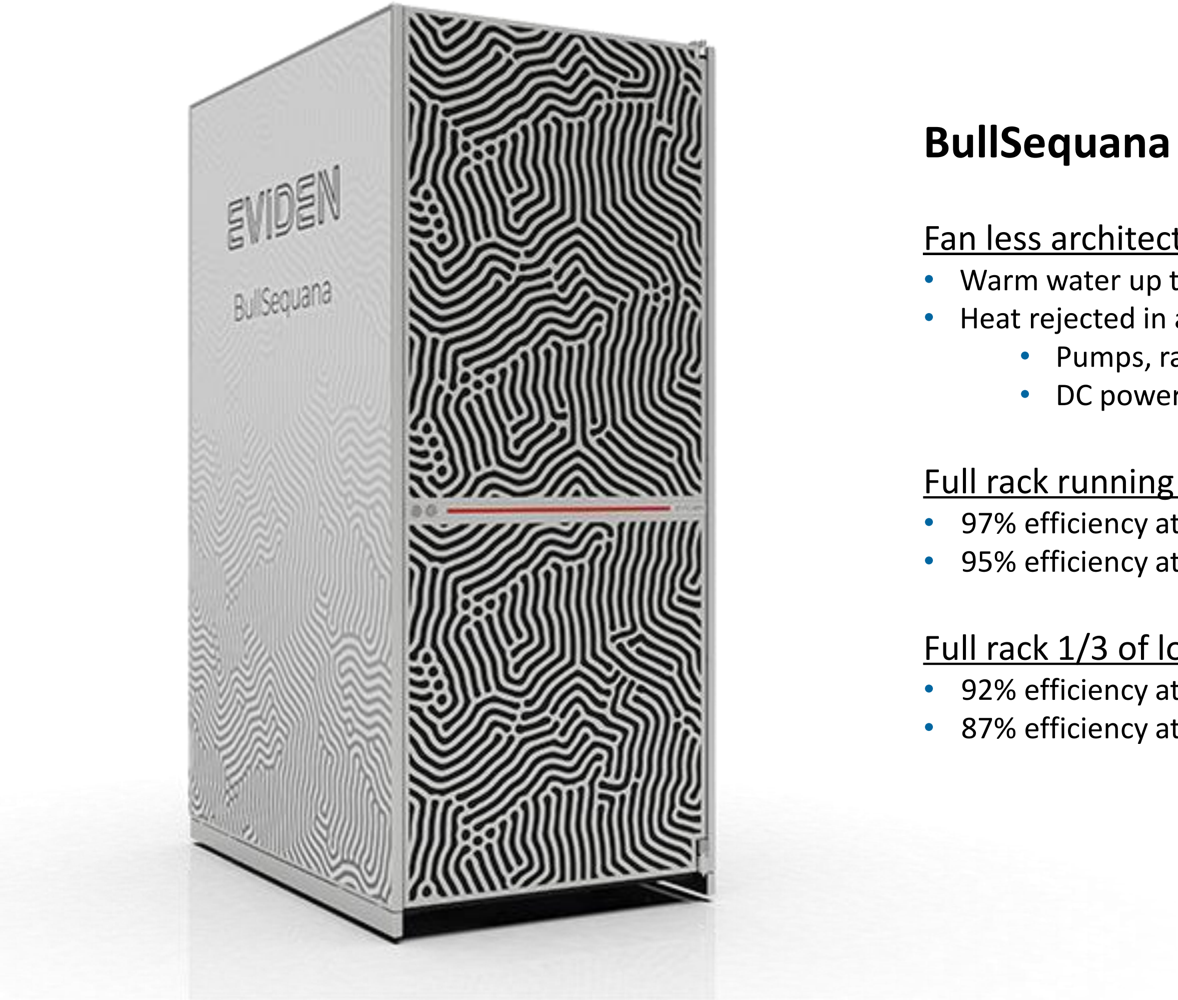
- High flow rate
- Outlet temp depend on Inlet
- Lower Air dissipation

***High** Internal temperature*

- Low flow rate
- Outlet temp always at max.
- Heat reuse

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

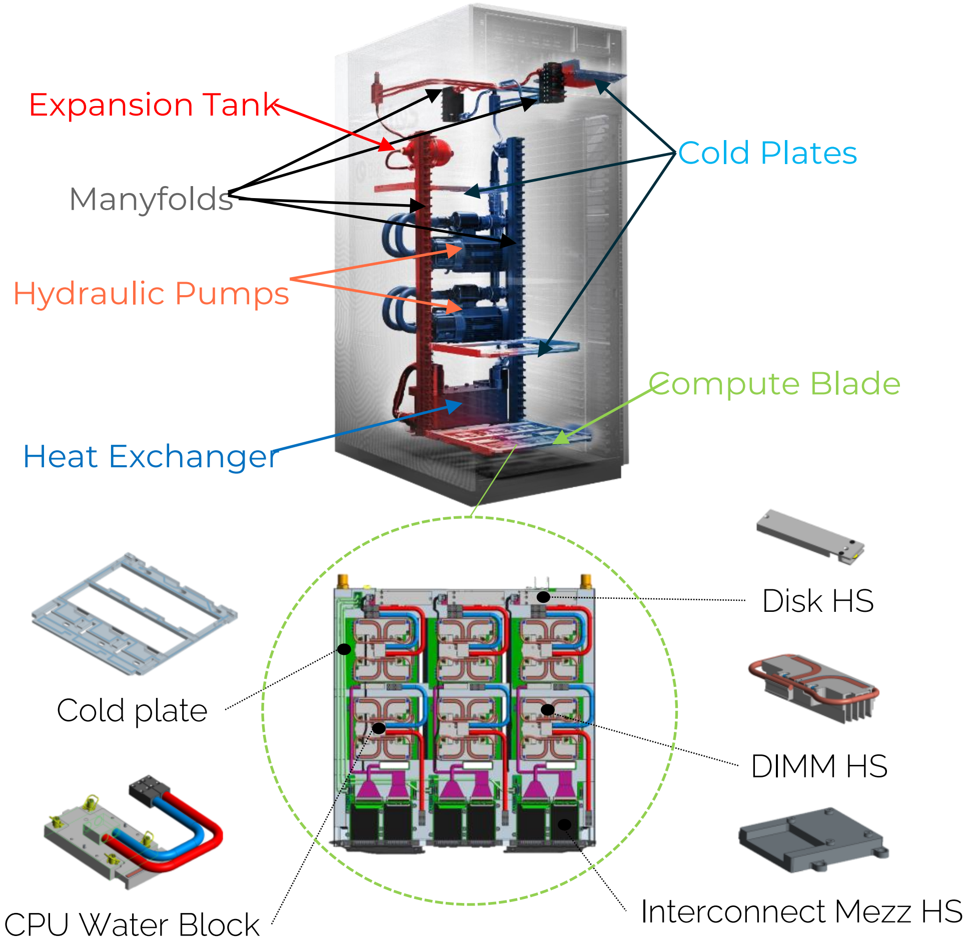
- 92% efficiency at Low Internal temperature: 3,2 kW Air dissipation
- 87% efficiency at High Internal temperature: 5,2 kW Air dissipation

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

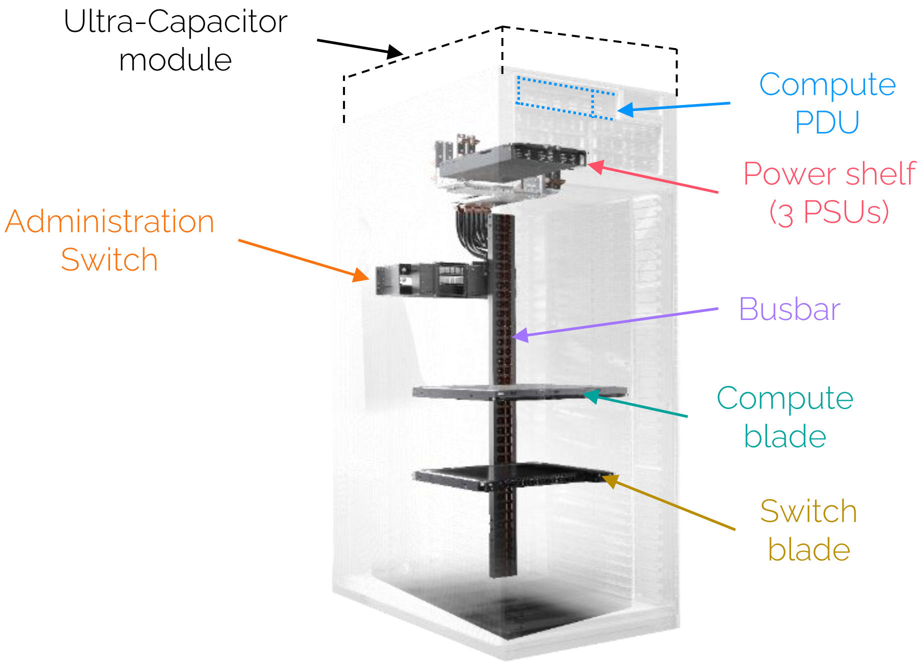
Power architecture is composed of several elements:

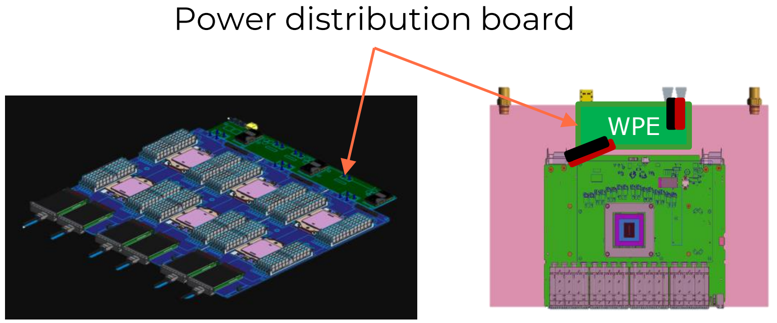
- Power Group P3G mounted on top of the rack:
  - Power Distribution Units (PDU): 2 compute PDUs (one at the front, one at the rear of the rack), 1 hydraulic PDU (at the front of the rack)
  - Power Shelves. One shelf contains 3 Titanium Power Supply Units of 4,2kW each. Max of 12 Power Shelves per rack (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





Compute blade boards

Switch blade board

First
to the
future

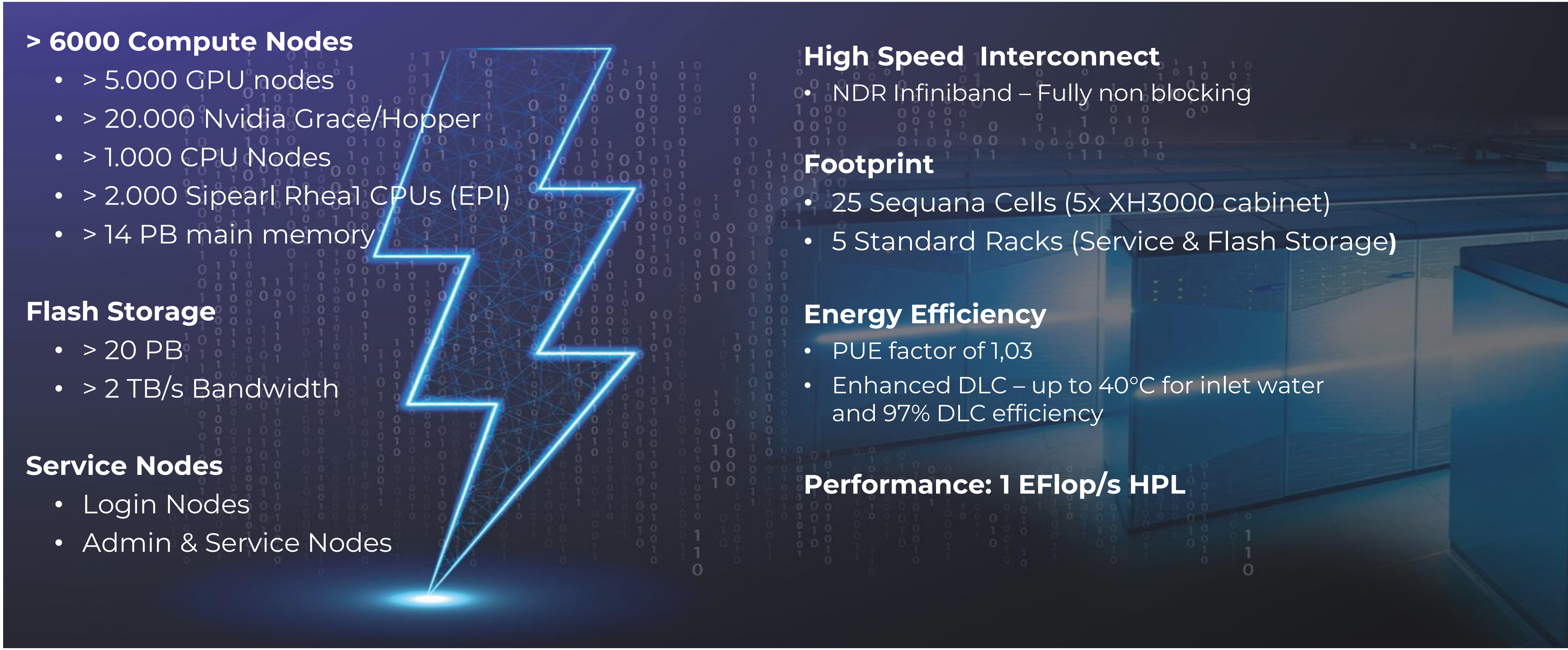
JUPITER (ExaScale)

System Architecture

System Solution

JUPITER: 1st European ExaScale System

First to the future



> 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

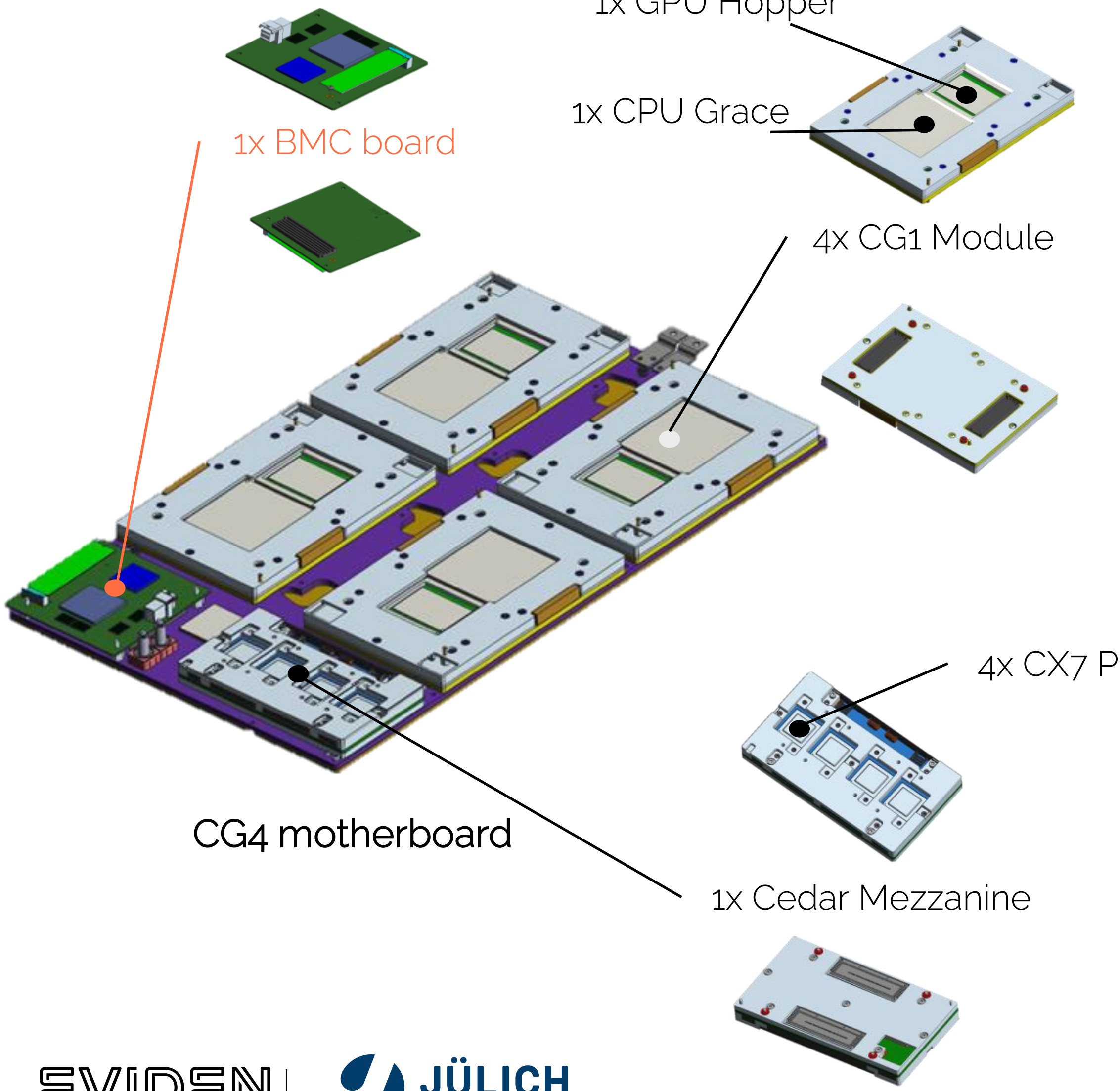
Performance: 1 EFlop/s HPL

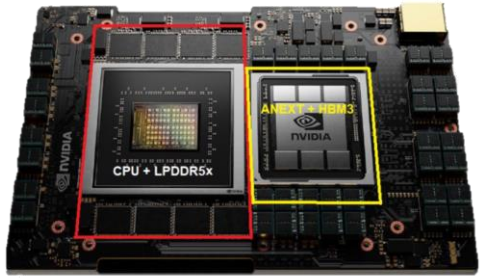
BullSequana X3515-HMQ Grace-Hopper Blade

First to the future

NVIDIA

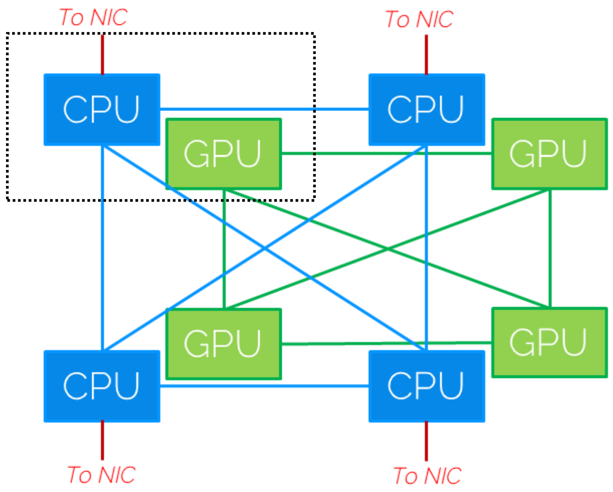
GPU Node





IN DEVELOPMENT

CG1 module

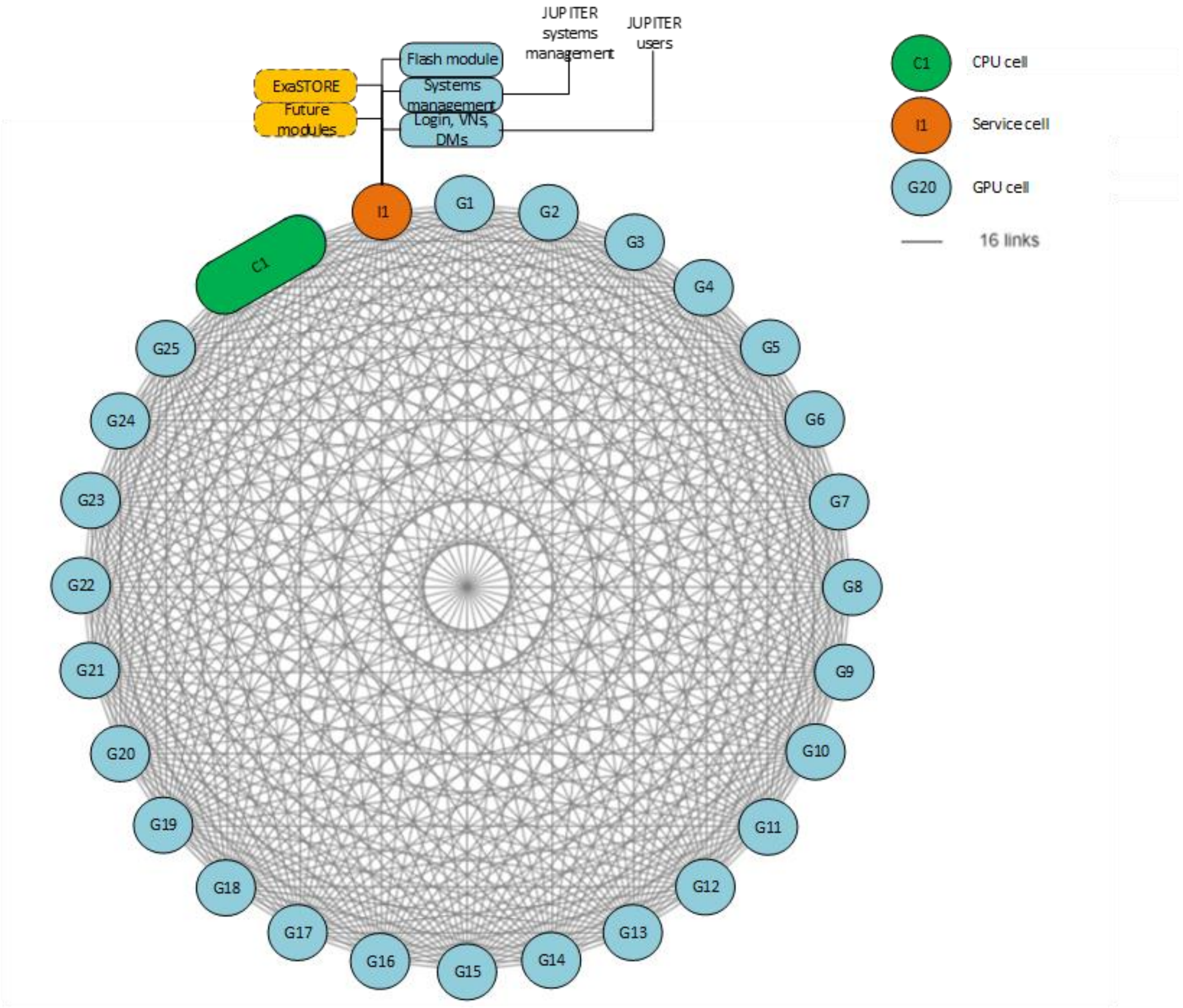


- All-to-all NVLink between CPU (C-link)
- All-to-all NVLink between GPU (G-link)
- Coherent memory space
- GPU direct access to NIC

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

© BULL GmbH

Questions?

EVIDEN

Thank you!

For more information please contact:

Dr. Thomas Warschko
Technical Director Extreme Computing Germany
Email: thomas.warschko@eviden.com