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 24th 2022

Atos

Evolution of the Sequana System Architecture
The Past, the Present and the Future

01. Sequana 1: X1000 – JUWELS Cluster
02. Sequana 2: XH2000 – JUWELS Booster, JURECA-DC
03. Sequana 3: XH3000
04. Q&A

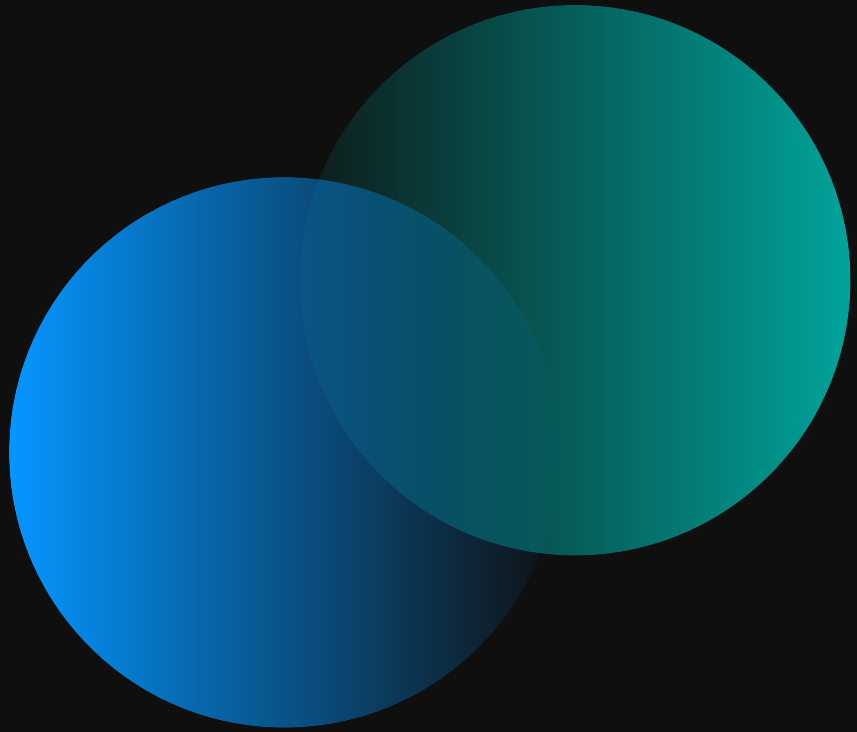
Pre Sequana Era
Bull B700 DLC Solution

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

- Many Sequana features already present:
  - All in one approach
  - Central Power (54C DC)

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

01. The Past:
Sequana 1 – X1000

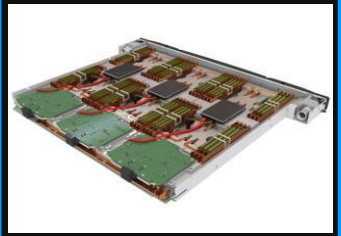


Bull Sequana X1000
the Bull exascale generation of supercomputer

- **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

Bull Sequana X1000 cell technology

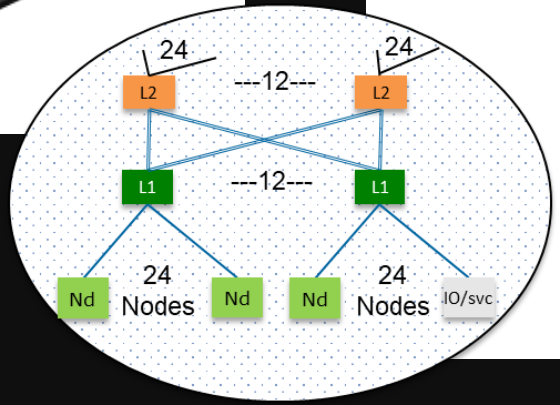
Bull Sequana X1000 cell

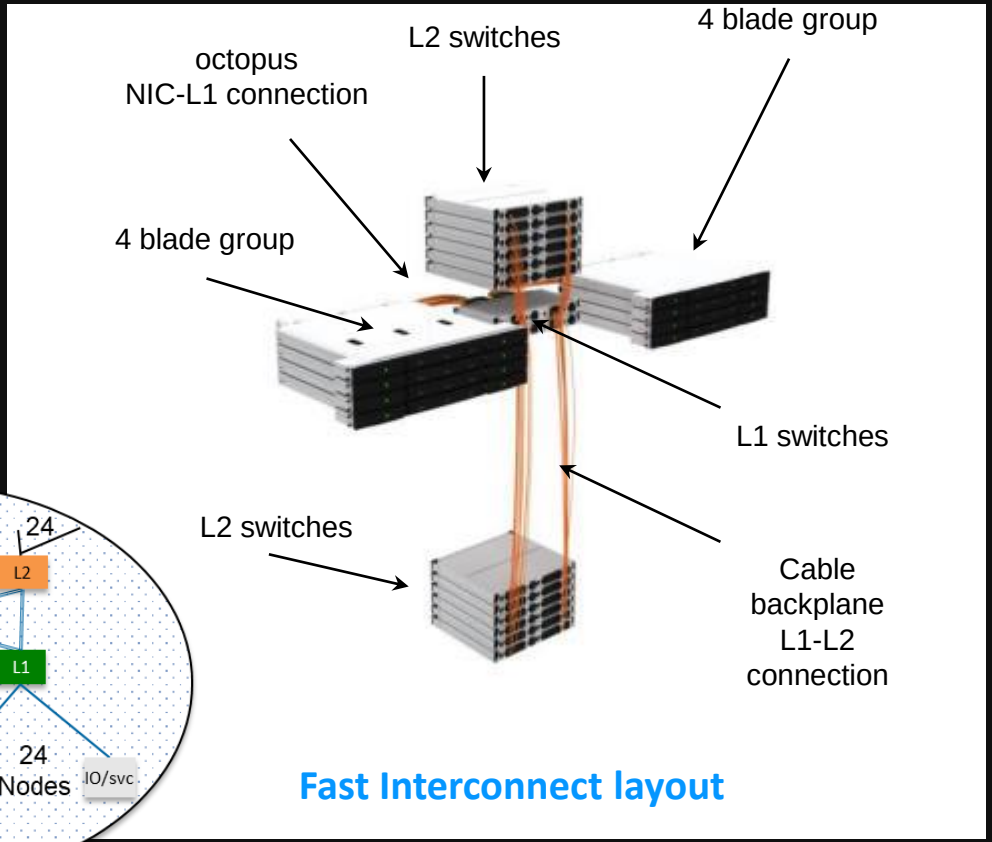


support several types and generations of compute nodes either with conventional processors or with accelerators

- 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



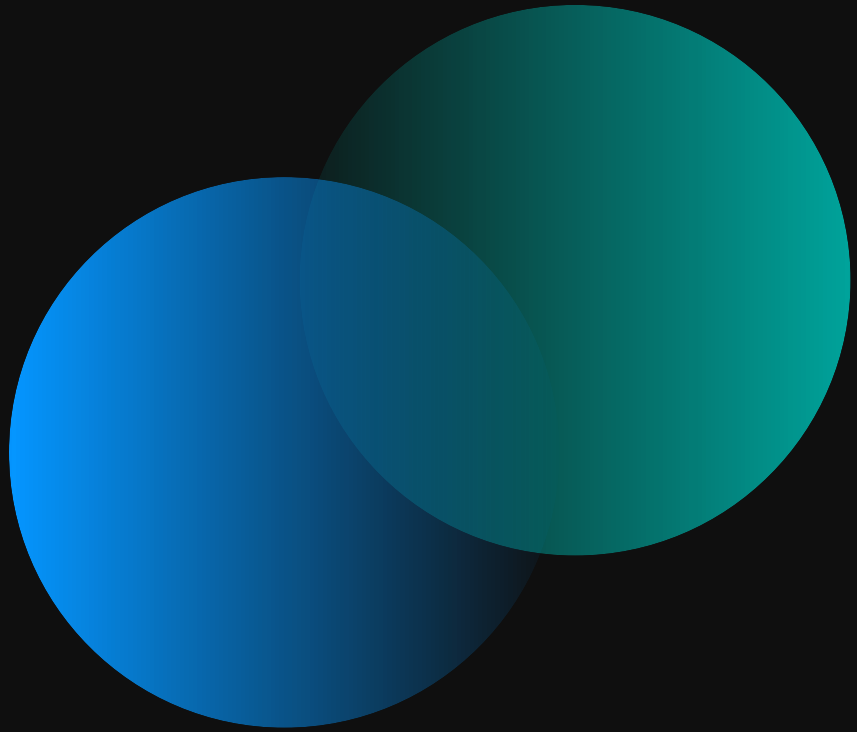


Bull Sequana X1000 (JUWELS Cluster)
Lessons Learned

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)**

02. The Present:
Sequana 2 - XH2000



What is BullSequana XH2000?

A new generation of Sequana X1000

- **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:

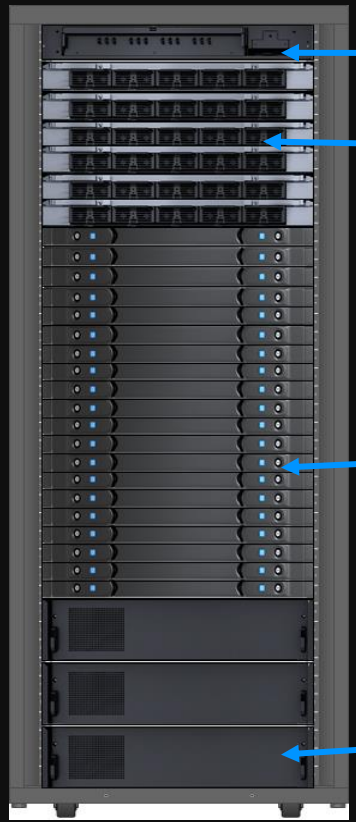
- up to 32 compute DLC blades/96 compute nodes
  - 20 on front side, 12 on rear side
- up to 6 liquid-cooled PSU shelves (up to 30 liquid cooled PSUs)
- fanless design
- 2 HYC, optional 3rd HYC for 2+1 redundancy
- 2 Leaf Ethernet modules
- Up to 10 Interconnect DLC Switches
  - HDR100 & HDR200 in Phase 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

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)

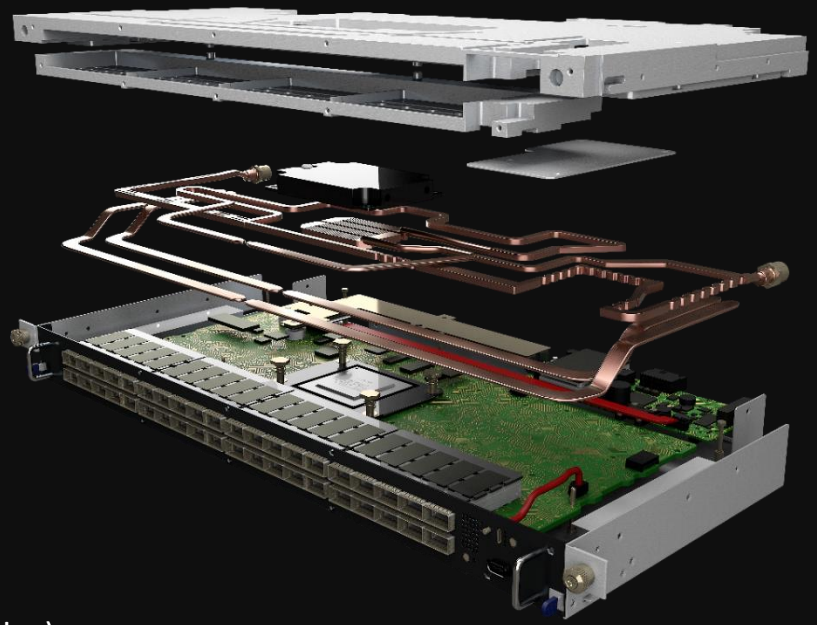
Rear View

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

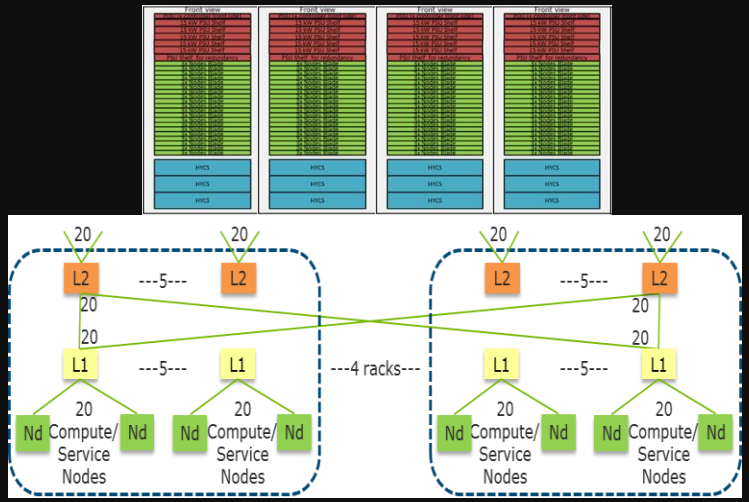


HDR Flexible Sideplane

- 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)

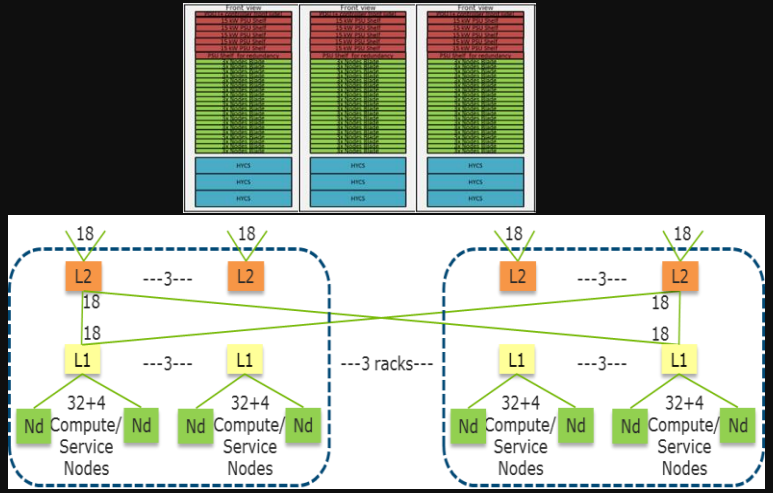
BullSequana XH2000 Topology options

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

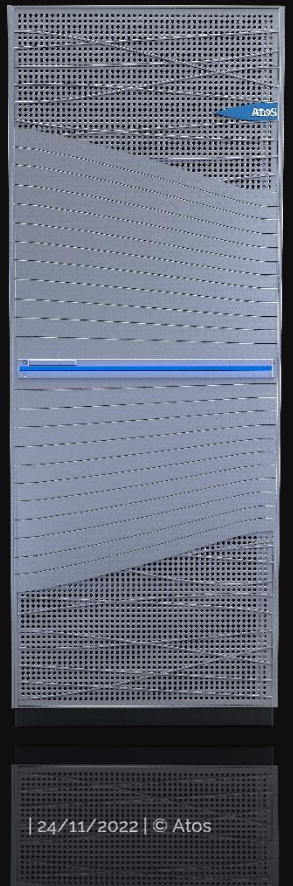


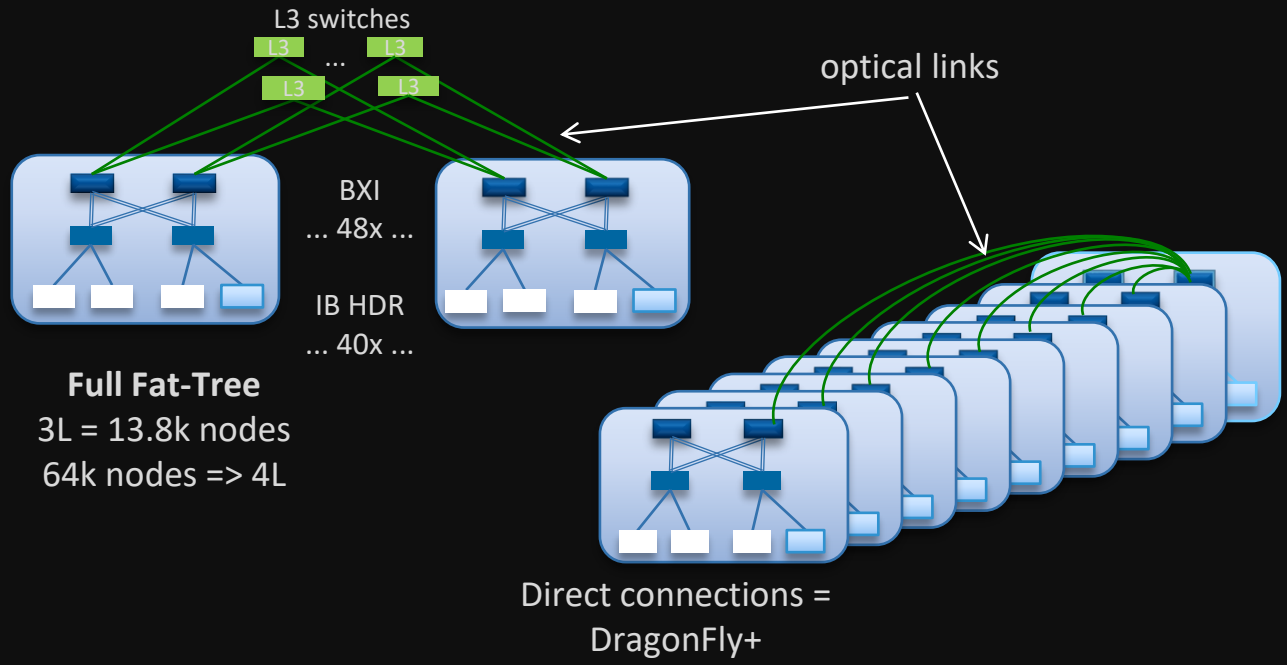
Single Sequana Cell: HDR100 FT 1:1

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

BullSequana XH2000 Networking

Best in class Interconnect flexibility





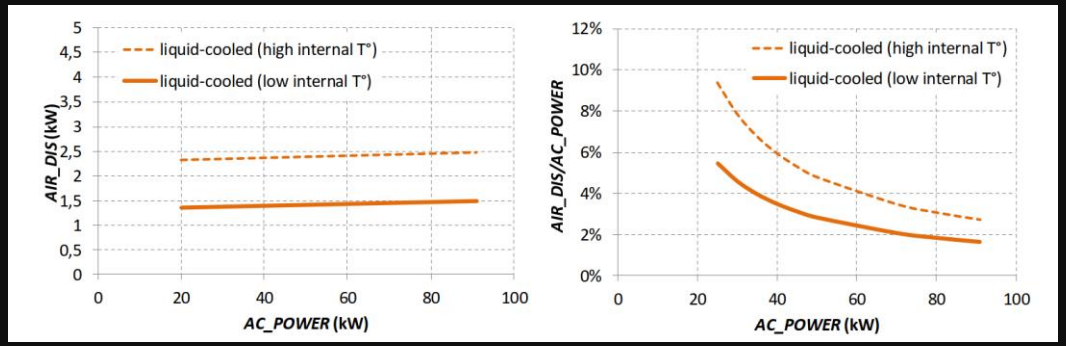
BullSequana XH2000 Cooling
Optimized Power Usage Effectiveness

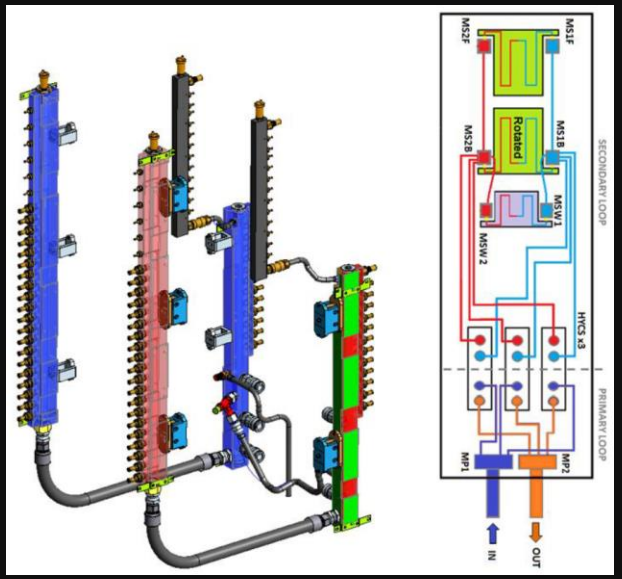
BullSequana XH2000: > 95% cooling efficiency

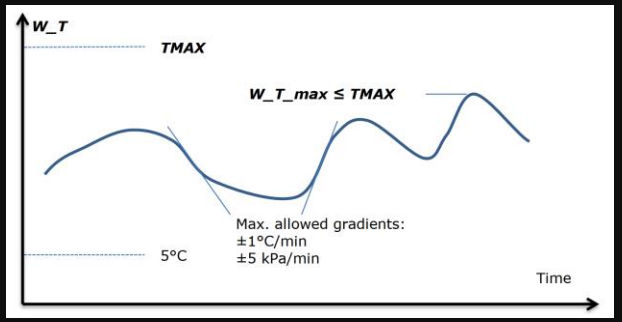
<u>Fan less architecture :</u>

- 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

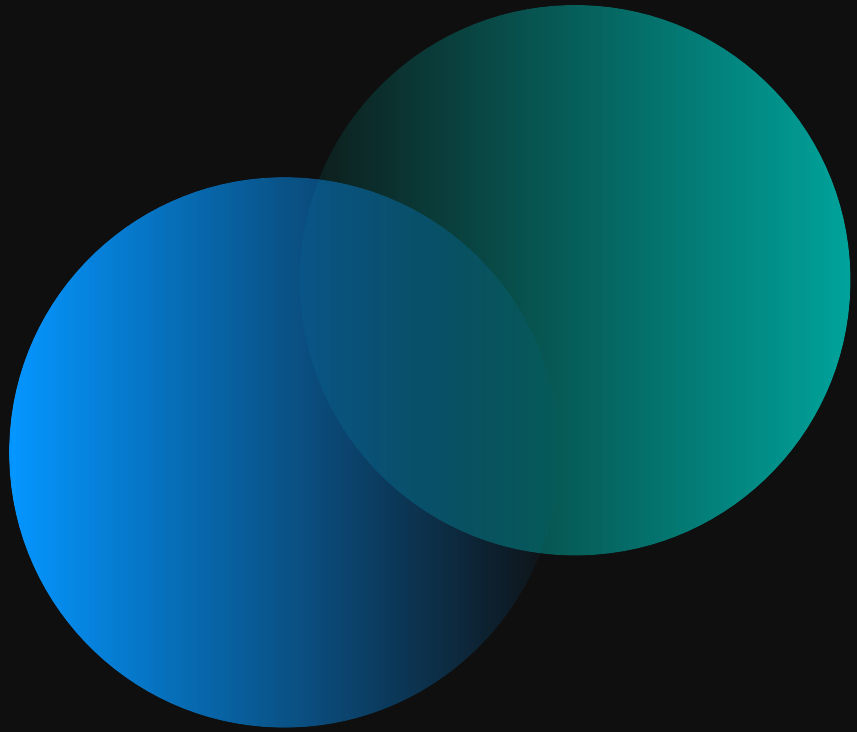
PRO:

- **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**

03. The Future:
Sequana 3 – XH3000



Bull Sequana XH3000

Keeping the pace

Stick to the successful features:

- **Modular system platform**
- **Blade system**
- **Multiple blade types**
- **Switch blade based on standard technology**
- **Rack 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)**

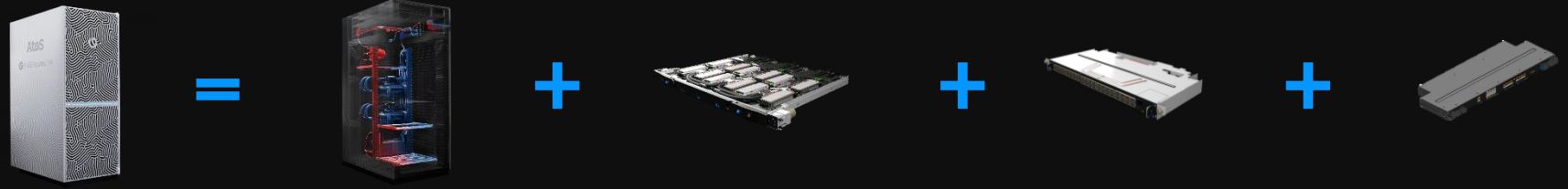
Adding new features and capabilities:

- **Increased power and cooling (to keep up with future technology)**
- **Support latest/future CPUs, GPUs, Interconnects and more**
- **Unified form factor for Blades**

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

For detailed information please consult the latest factsheets on our web-site:

https://atos.net/en/solutions/high-performance-computing-hpc/bullsequana-x-supercomputers

Questions ?

Thank you!

For more information please contact:

Dr. Thomas Warschko
Email: thomas.warschko@atos.net

Atos