AI Factories

“Technical and Sales”

Vladimir Vukovic

Staff Solution Manager Architecture

Supermicro Netherlands

- AI Factory
- Direct Liquid Cooling
- Air Cooled racks

AI Factory

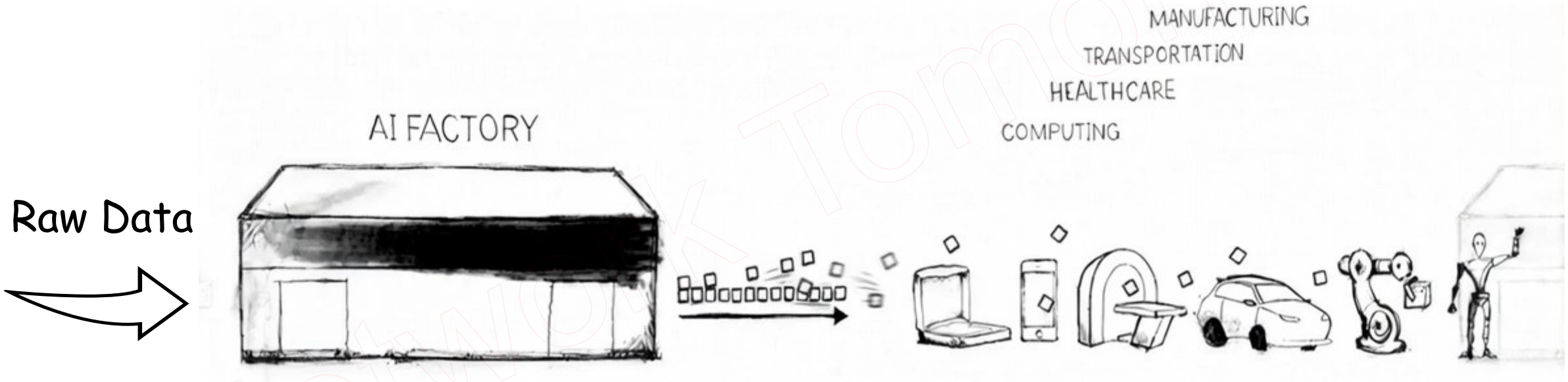
Academia, SMB or Enterprise

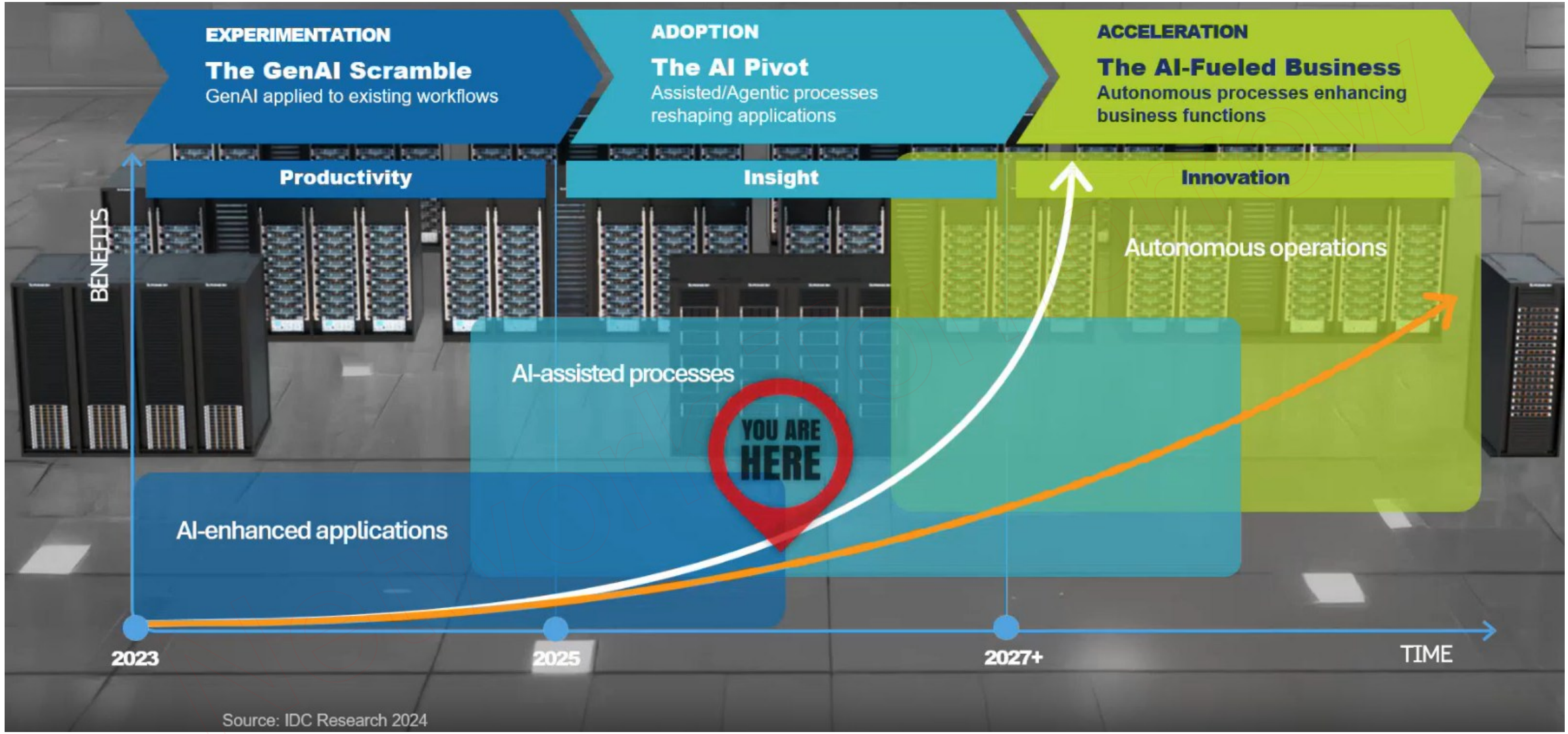
Industrial Factory

Raw material

Product

AI Factory

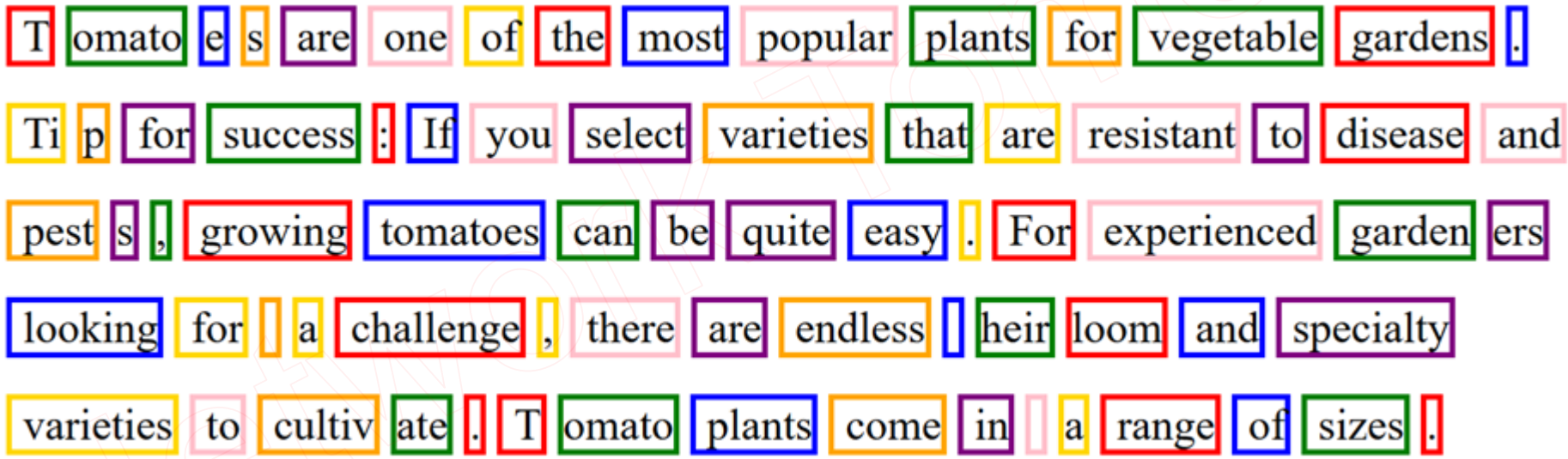



EXPERIMENTATION
The GenAI Scramble
GenAI applied to existing workflows
ADOPTION
The AI Pivot
Assisted/Agentic processes reshaping applications
ACCELERATION
The AI-Fueled Business
Autonomous processes enhancing business functions
Productivity
Insight
Innovation
BENEFITS
Autonomous operations
AI-assisted processes
YOU ARE HERE
AI-enhanced applications
2023
2025
2027+
TIME
Source: IDC Research 2024

TOKEN

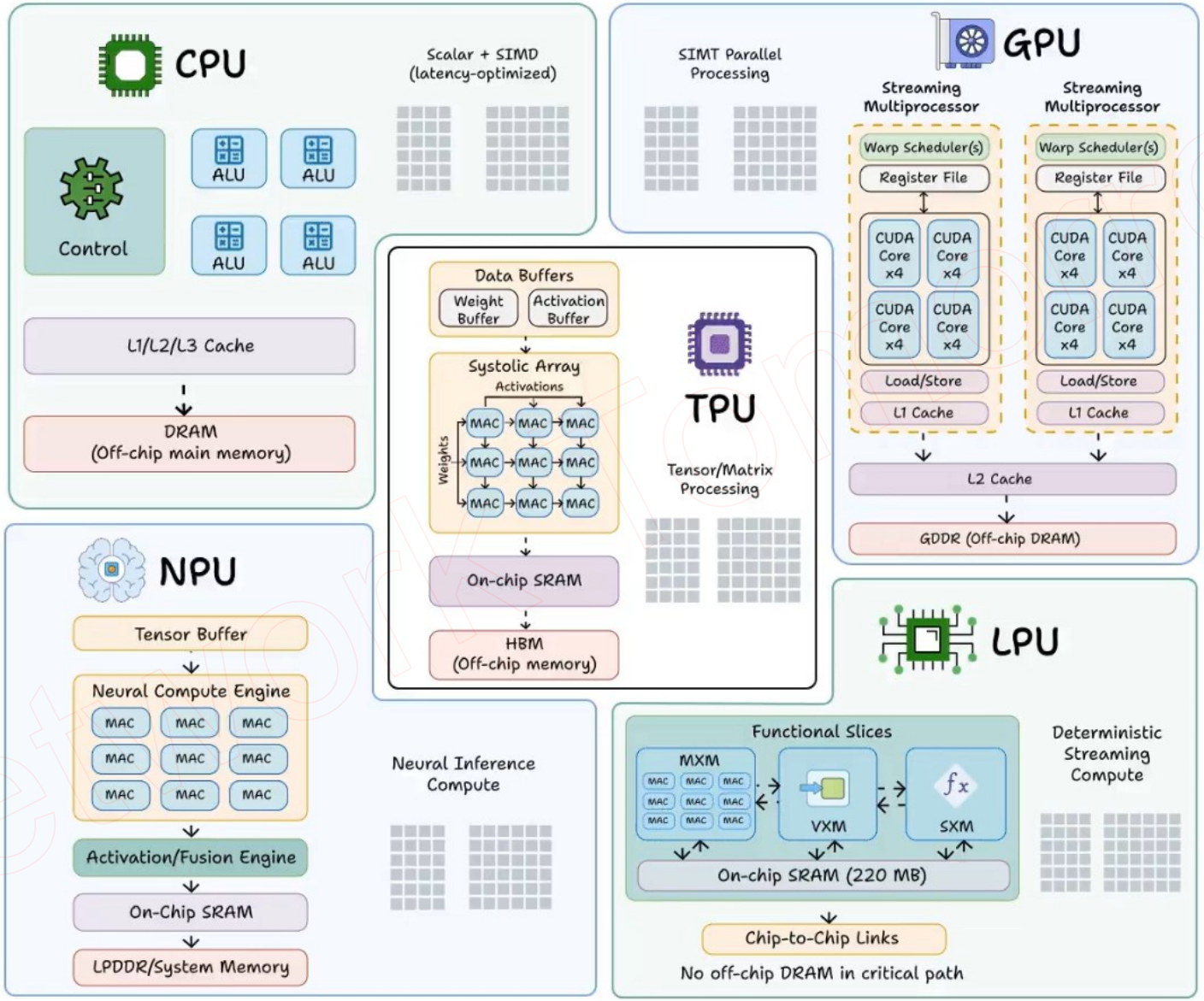
Tokenization is a preprocessing technique used in Natural Language Processing (NLP)

NLP algorithms segment large texts into smaller tokens that NLP tools can process.

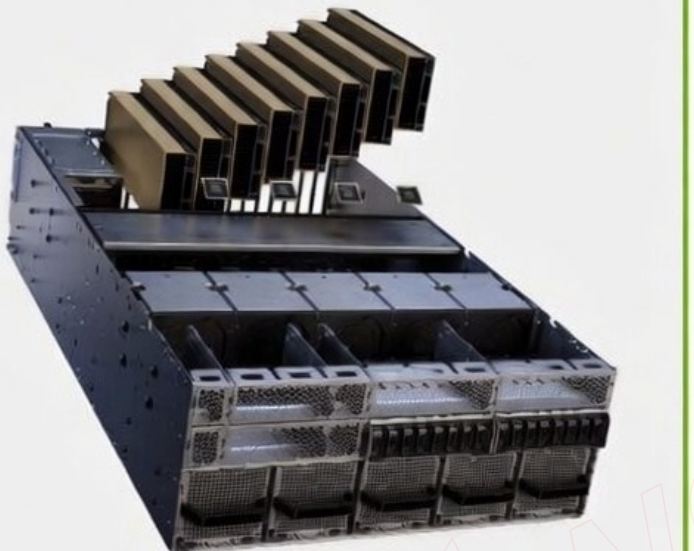


This illustration shows how an NLP algorithm might break text into a collection of tokens.

5 AI Compute Architectures



AI Machines for Manufacturing



RTX PRO Server

Multi-Workload Acceleration,
Enterprise DC Compatibility

HGX B300

AI-Optimized Performance for Training and
Large Model Inference

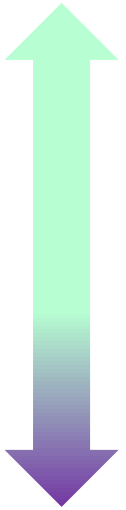
GB300 NVL72

Exascale Computer in a Rack for Massive
LLM Training and Inference

	A100 (EoL)	H200	H200 NVL	RTX Pro 6k BSE	B200	B300
Form Factor	SXM	SXM	PCIe	PCIe	SXM	SXM
FP64	9.7TF	34TF	30TF	2TF	37TF	1.2TF
FP64 Tensor Core	19.5TF	67TF	60TF	-	37TF	1.2TF
FP32	19.5TF	67TF	60TF	120TF	75TF	75TF
TF32 Tensor Core (Sparce)	312TF	989TF	835TF	468TF	2,200TF	2,200TF
FP16/BF16 Tensor Core (Sparce)	624TF	1,979 TF	1,671TF	935TF	4,500TF	4,500TF
FP8/FP6 Tensor Core	-	3,958TF	3,341TF	1,870TF	9,000TF	9,000TF
INT8 Tensor Core (Sparce)	1,248TF	3,958TF	3,341TF	1,870TF	9,000TF	9,000TF
Peak FP4 AI (Sparce)	-	-	-	4TF	18,000TF	18,000TF
Peak FP4 AI (Dense)	-	-	-	2TF	9,000TF	14,000TF
Ray Tracing Core perf	-	-	-	355TF	-	-
GPU Mem	80GB HBM2e	141GB HBM3e	141GB HBM3e	96GB GDDR7	180GB HBM3e	288GB HBM3e
TDP	400W	700W	600W	600W	1000W	1100W

> Power

Precision



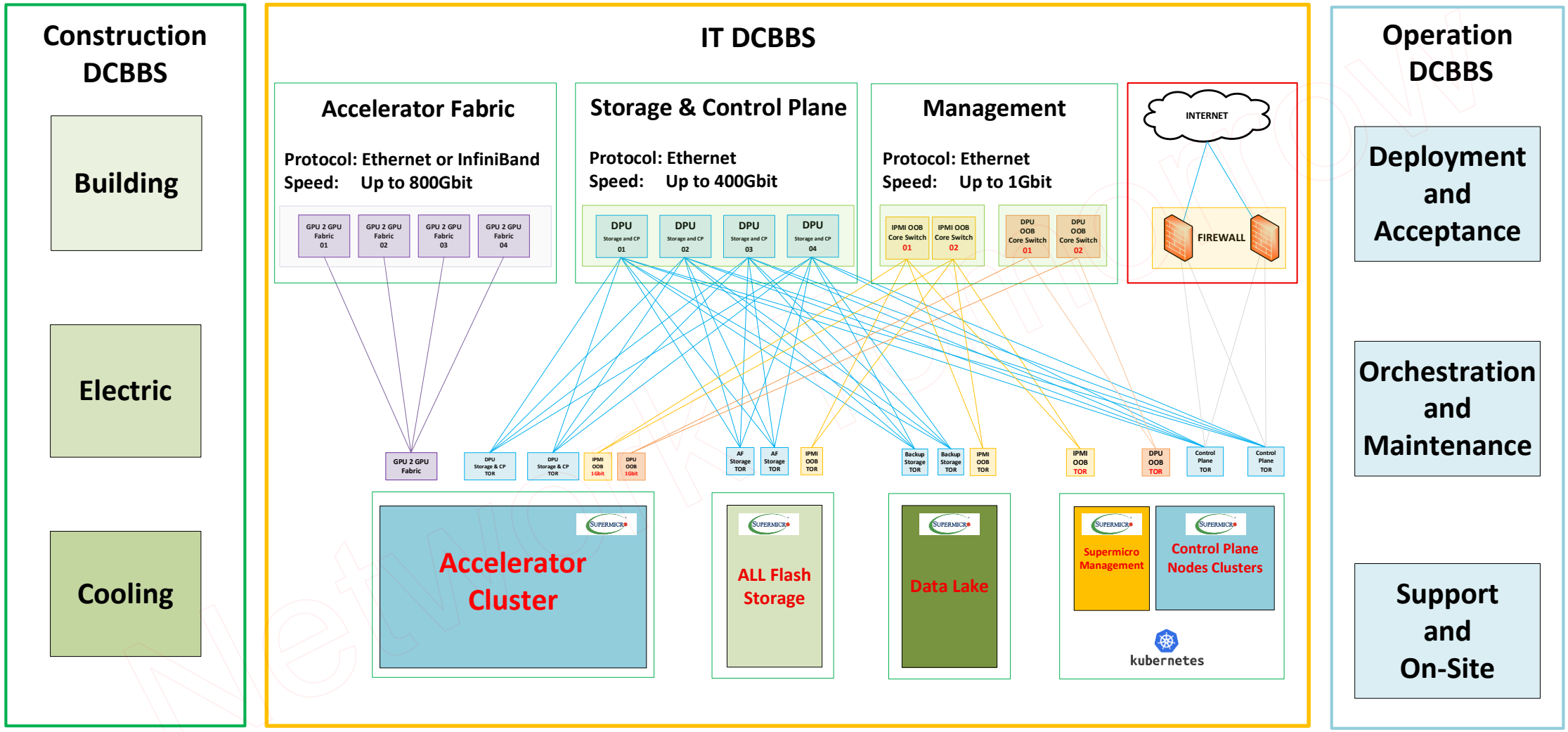
Speed

< Power

Select the Best GPU for the Job

	Use Case	
Large LLM Model Training and Inference	Foundational Model Training	B300 / GB300
	Fine Tuning / Inference (B300 / GB300)	B300 / GB300
	Double Precision (FP64) Workloads?	GB300
	Memory-Dependent Workloads (HBM3e req.) ?	B300 / GB300
Compute-heavy Workloads No Graphics Power Constraints	Mid to Large LLM Inference (>70B) / FP64); <20kW Rack	H200 NVL Enterprise Server
Mix of Accelerated Workloads Flexible + Lower Power	Fine-Tuning/Inference/RAG (< 70B Parameters)	RTX PRO Server
	Early in AI Journey? Data Center Refresh?	RTX PRO Server
	<20kW Data Center Rack? PCIe Flexibility?	RTX PRO Server
	Visual Computing Workloads (Graphics, Design, Digital Twins)?	RTX PRO Server
	Additional Workloads (Data Analytics, Simulation, FP32 HPC)?	RTX PRO Server

AI Factory Component Logics



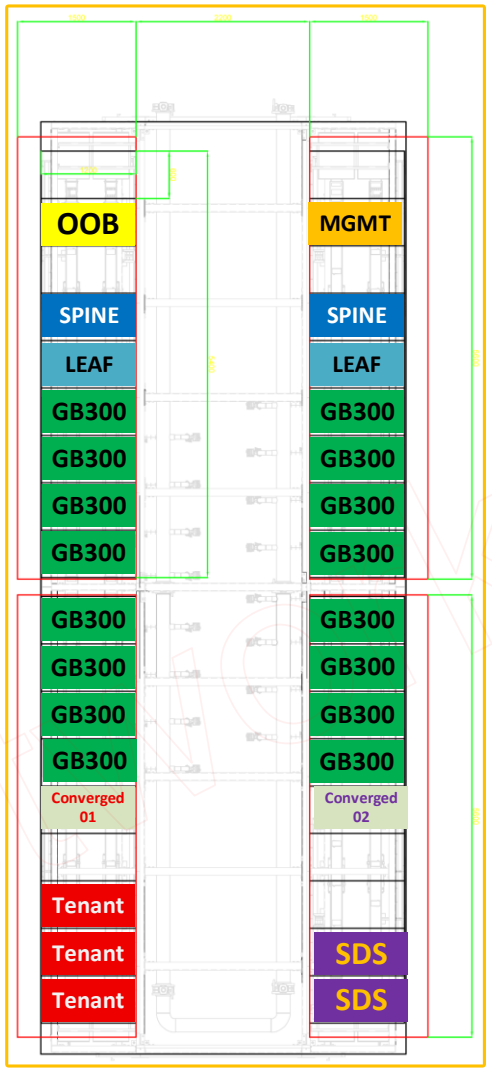
DCBBS – Data Center Building Block Solution

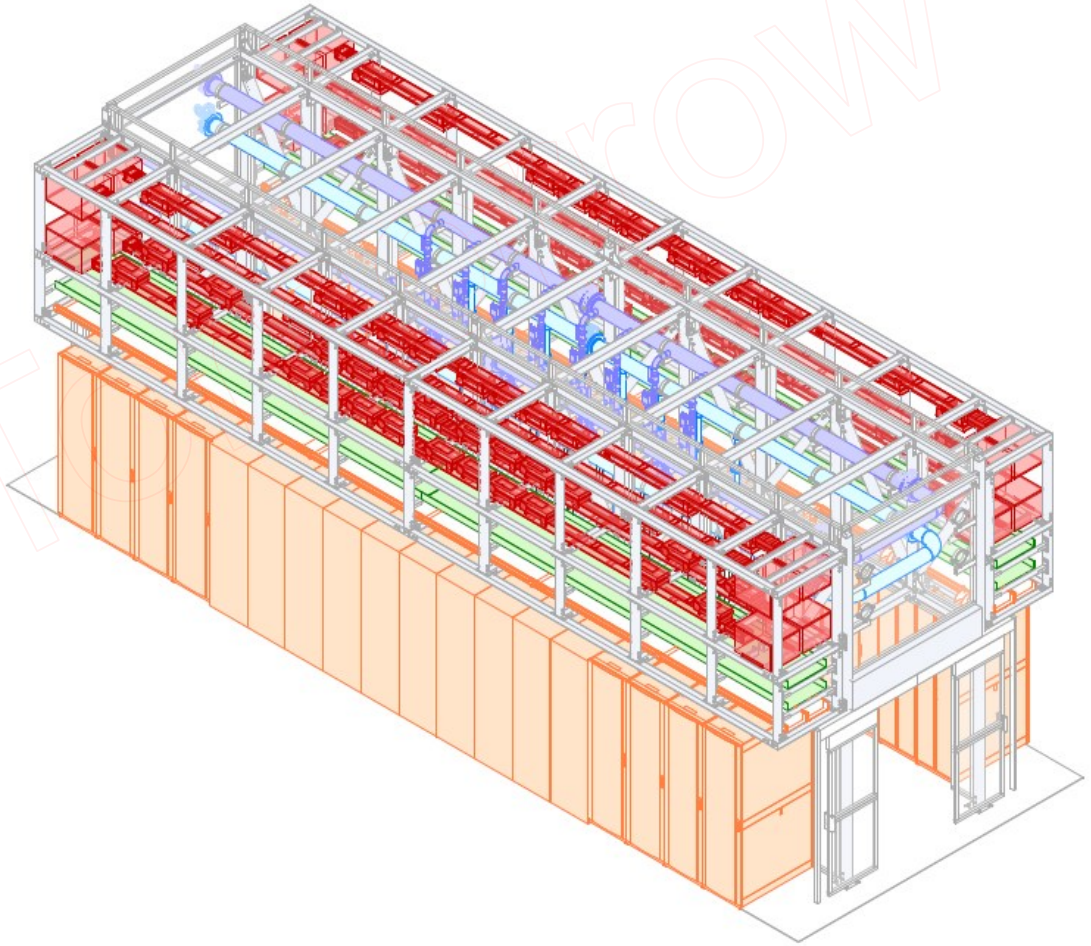
https://www.supermicro.com/en/solutions/dcbbs

POD Example

GB300	GB300 NVL72 Racks
LEAF	InfiniBand Leaf Racks
SPINE	InfiniBand Spine Racks
Converged 01	Converged 01 Racks
Converged 02	Converged 02 Racks
OOB	OOB Network Rack
SDS	Software Defined Storage
MGMT	Management Rack
Tenant	Tenant Racks with Servers

MDC POD 1

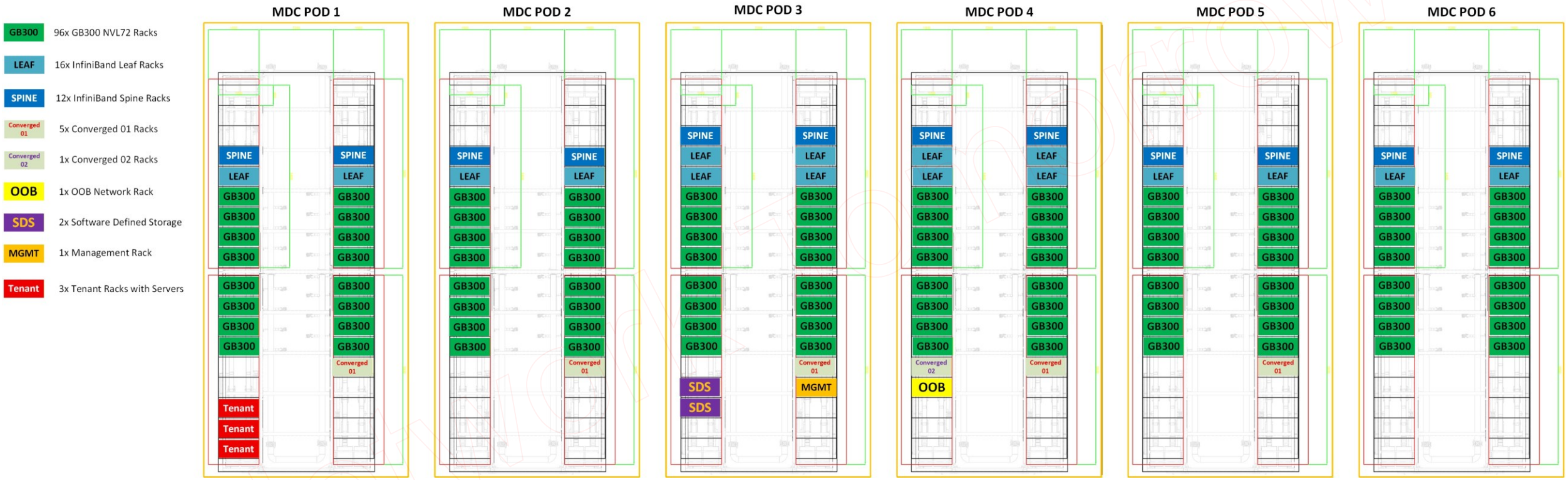




Schneider Electric
PREFABRICATED MODULAR DATA CENTERS SECURE POWER
www.se.com

SUPERMICRO

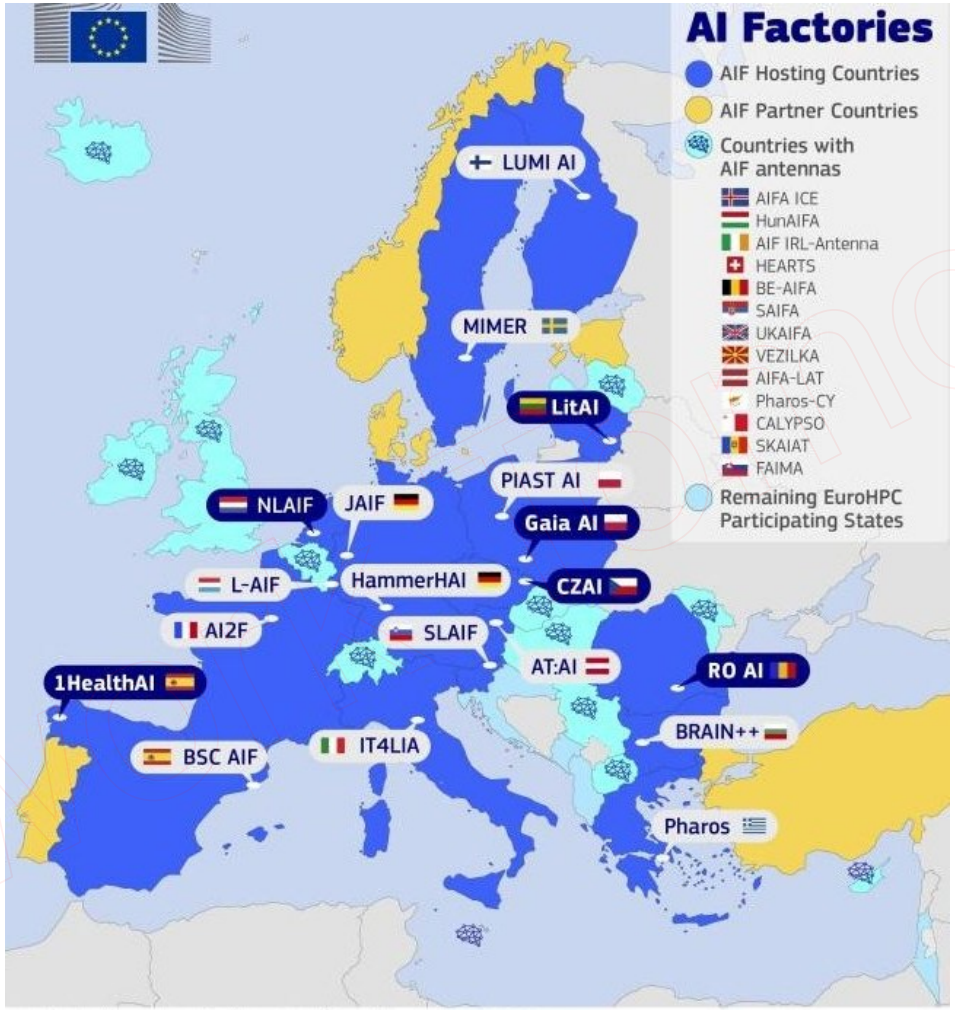
Project example



Quote break-out example for AI Clusters

	Cluster Part	Description	Quote Number	Quantity
1	**Accelerator Cluster**	GPU rack	SRS-GB300-NVL4-M0-customer(QN)	**96**
2	**E/W Infiniband**	Rack for East / West network	SRS-48UNTW-EWSP-01-customer(QN)	**12**
3	**N/S Converged / Ethernet**	Rack for North / South network	SRS-48UNTW-CONV-01-customer(QN)	**5**
			SRS-48UNTW-CONV-02-customer(QN)	**1**
4	**OOB Network / Ethernet**	Rack for Management switch	SRS-48NTWK-OOBN-01-customer(QN)	**2**
5	**All Flash Storage**	Rack Software Defined Storage	SRS-48UNTW-SDS-customer(QN)	**2**
		System for SDS	86015000-6688	**1**
6	**MGMT Rack**	Rack for Management Nodes	SRS-48CLDC-SCCA-01-customer(QN)	**1**
		SMC SYS for SCC	8601500-8688	**2**
		SYS for Admin K8S nodes	8601500-8688	**3**
7	**TENANT rack**	Rack for Tenant Servers	SRS-48CLDC-TENT-01-customer(QN)	**3**
		System for Tenant	8601500-8688	**21**
8	**ON-Site Services**	Deployment, acceptance tests and benchmarks	SOQ for All Supermicro rack	**1**
9	**On-Site Deployment**	L11 + L12 Testing	SOQ for rack level and cluster level test	**1**
10	**Enteprise Software**	Can be Purchased via: Supermcro or from local Distributor	860150-8688, 96x Lic	**1**
	Enteprise Services		860150-8688	**1**
11	**Schneider Electric**	Modular DC, In-Row CDU and Airconditioning	PODs & CDUs & CRAC	**1**

EuroHPC AI Factories



Note:

Other verticals are coping request framework

https://www.eurohpc-ju.europa.eu/eurohpc-ju-selects-six-additional-ai-factories-expand-europes-ai-capabilities-2025-10-10_en

Example of Award Criteria

Criterion	Score calculation	Score	Min	Total Score
1. Technical Value of the system design Section 3 of Technical Specification	For Sub-criterions 1.1 and 1.2, form a score from the individual scores s_i Target Capabilities: $TV_x = \Sigma_i s_i^X$ 0-50pt. For Sub-criterion 1.3: Qualitative evaluation: 0-6pt.	0-56 pt [42.43%]	29 pt	
2. Cost Performance Analysis Section 4 of Technical Specification	Score according to the described methodology in section 12.1.8 of the Technical Specification. The final Cost Performance Analysis score formula is the following: $\boldsymbol{Score(v) = \left[\frac{\rho(v)}{\max_v\{\rho(v)\}} \cdot \frac{Q(v)}{1.1}\right] \cdot 40}$	0-40 pt [30.30%]	22 pt	0-132 pt
3. Quality of services Sections 5, 6 and 7 of Technical Specification	Qualitative evaluation of the services provided, C3	0-16 pt [12.12%]	9 pt	
4. EU added value Section 8 of Technical Specification	Qualitative Evaluation based on Section 7 of Technical Specification. Qualitative evaluation, C4	0-20 pt [15.15%]	11 pt	

Couple of Key Points

- Early Engagement as Much as Possible (RFI Stage is preferred)
- Customer Used Case (HPC or NCP)
- Reference Architecture is starting point
- Collaboration – Internally and Externally
- Multi months project
- Components Availability and Forecast
- TTO – Time To Operational

Direct Liquid Cooling

DLC

Largest AI Supercomputer Powered by Supermicro

World's Largest Liquid-Cooled AI Cluster

- 100,000 NVIDIA H100 Tensor Core GPUs
- Multi-billion-dollar cluster, deployed in 122 days
- The basic building block for this Supercomputer is the Supermicro liquid-cooled rack
- 8x 4U servers each with 8 GPUs, for a total of 64 GPUs per rack, plus a CDU
- 1,500+ Racks, 1,500+ CDUs, 300K+ cables
- Spectrum-X (400GB) Ethernet
- Supermicro's design is from the ground up to be liquid-cooled, and all from one vendor

Supermicro NVIDIA HGX Systems

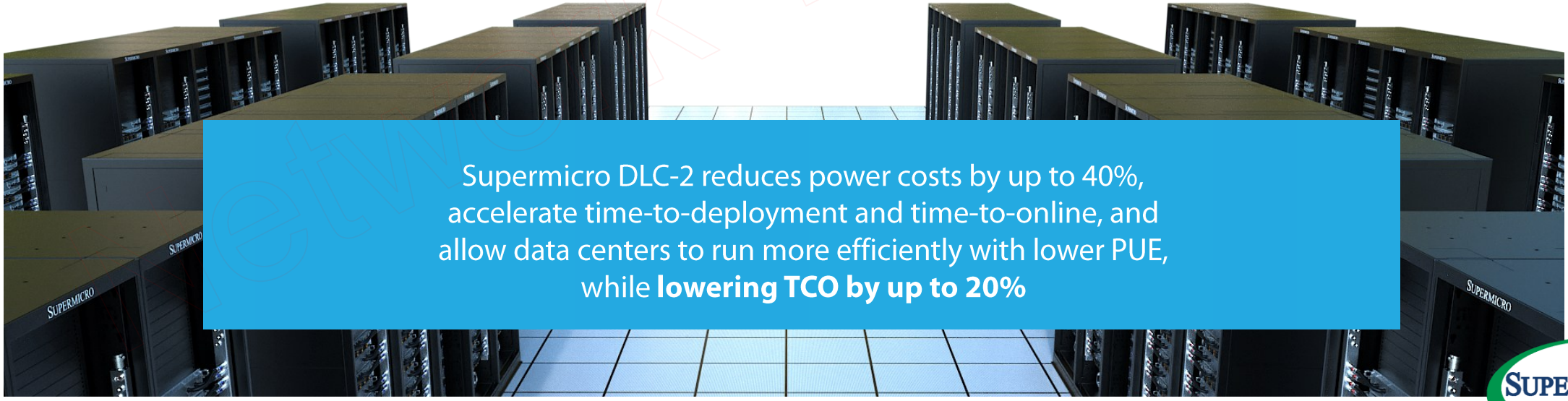
Supermicro Liquid Cooling Total Solution

NVIDIA Spectrum-X Ethernet networking

Data Center End-to-End Liquid-Cooling Solution

Up to 40% More AI per Watt with The Next Generation Direct-to-Chip Liquid Cooling Solutions

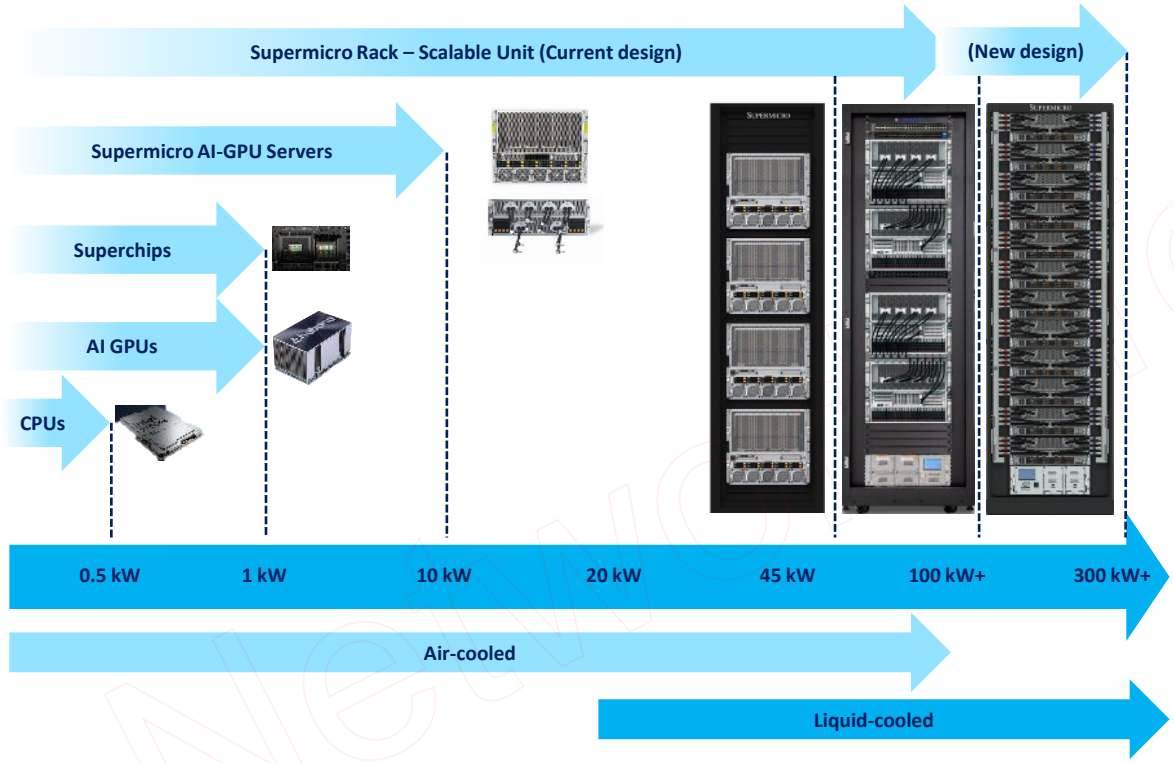
Power Savings	Water Savings	System Heat Capture	Quiet Data Center	Space Savings
up to **40%**	up to **40%**	up to **98%**	as low as **50dB**	up to **60%**
Savings in entire data center (vs. Air-cooling) by using Supermicro DLC-2	Savings with 45 °C warm water operation and eliminating chilled water and compressor	Heat capture in DLC-2 Liquid-cooling with CPU, GPU, PCIe Switch, DIMM, VRM, PSU, and more	Significantly reduces noise with less fans and fan speed. As quiet as a library	Savings with more than 2.5x compute density compared to air-cooled systems

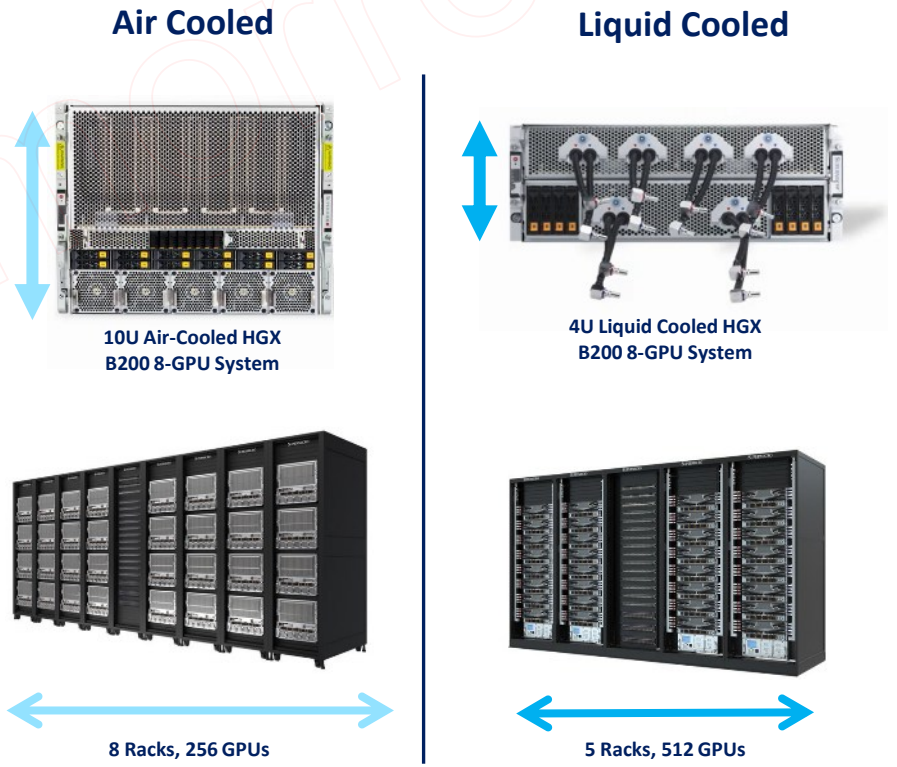


Supermicro DLC-2 reduces power costs by up to 40%, accelerate time-to-deployment and time-to-online, and allow data centers to run more efficiently with lower PUE, while **lowering TCO by up to 20%**

Why plan for liquid cooling

Power hungry AI Servers driving higher energy consumption



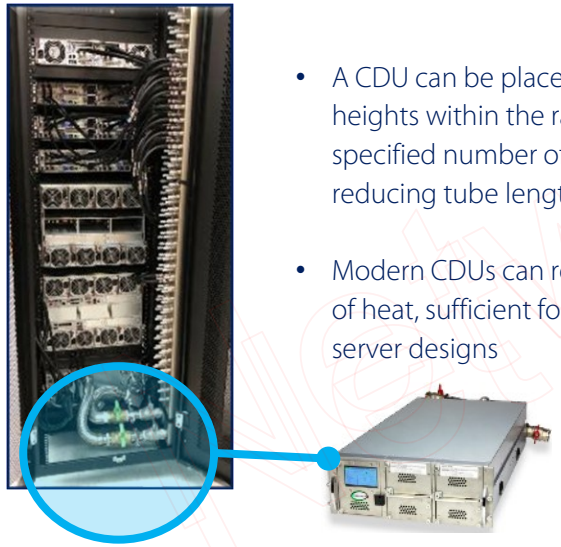


Higher density with lower energy cost, rack space and server noise

Supermicro Liquid Cooling Solutions

In-Rack CDU

- Installed within the server rack
- Directly cools IT equipment at the source
- Provides localized cooling with shorter coolant lines



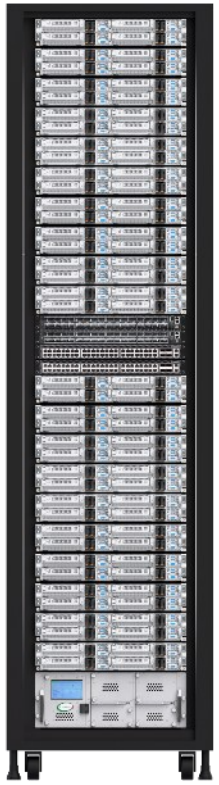
- A CDU can be placed at different heights within the rack to cool a specified number of servers, reducing tube lengths
- Modern CDUs can remove 100+ kW of heat, sufficient for most of today's server designs

In-Row CDU

- Located outside the server rack or data center
- Cools coolant circulated throughout the facility
- Supports larger cooling capacities but requires longer coolant lines
- Allows more flexibility in rack design and layout

Full Portfolio of Liquid Cooling Solutions

Liquid Cooled Rack Configurations



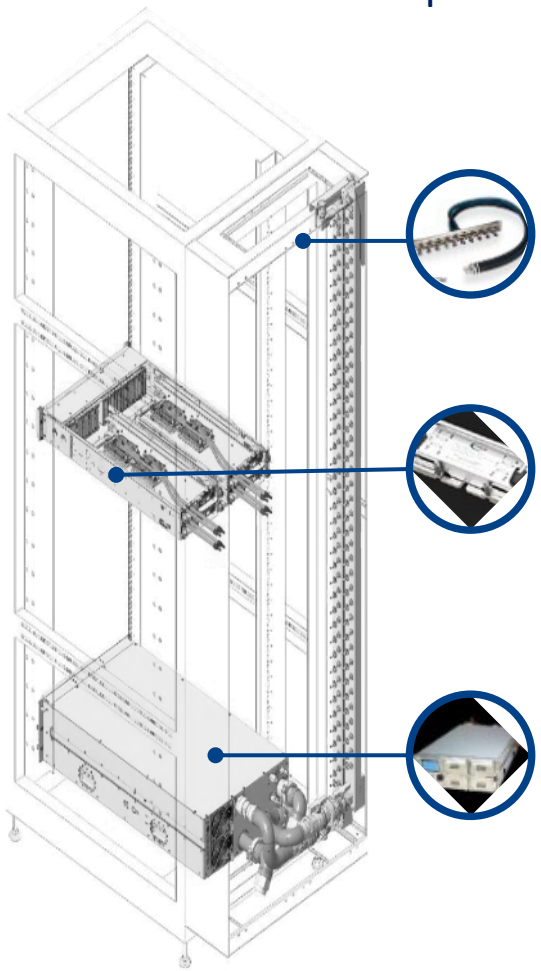
High Density Compute Rack

Multi-Node

GPU System

High Density GPU System

Components



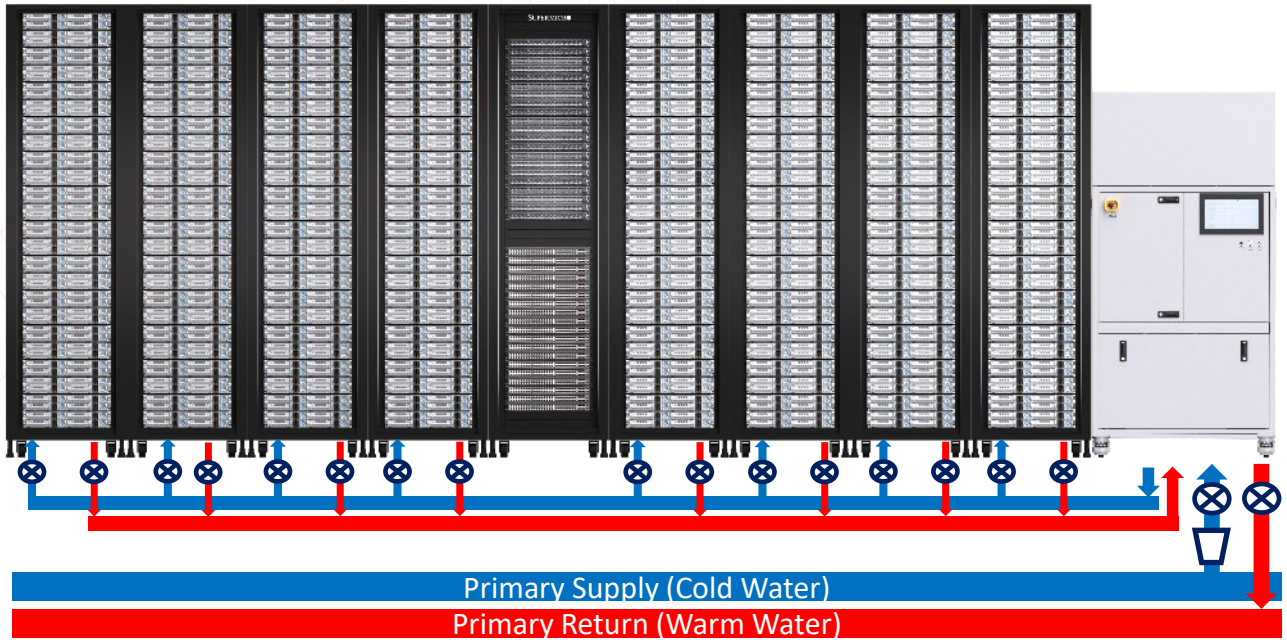
Cooling Tower

FlexTwin™ Rack Optimized Liquid Architectures

In-Rack CDU
Reference Configuration

High-Daensity
Compute Rack

In-Row CDU
Reference Configuration



Air vs DLC

Why Use A Liquid To Air Sidecar

Some solutions are only available liquid cooled, sidecar enables use inside air cooled DC without the need for liquid cooling infrastructure.

In cases where higher density is needed, the sidecar can offer an advantage over air cooled using liquid systems without the need for liquid cooling infrastructure.

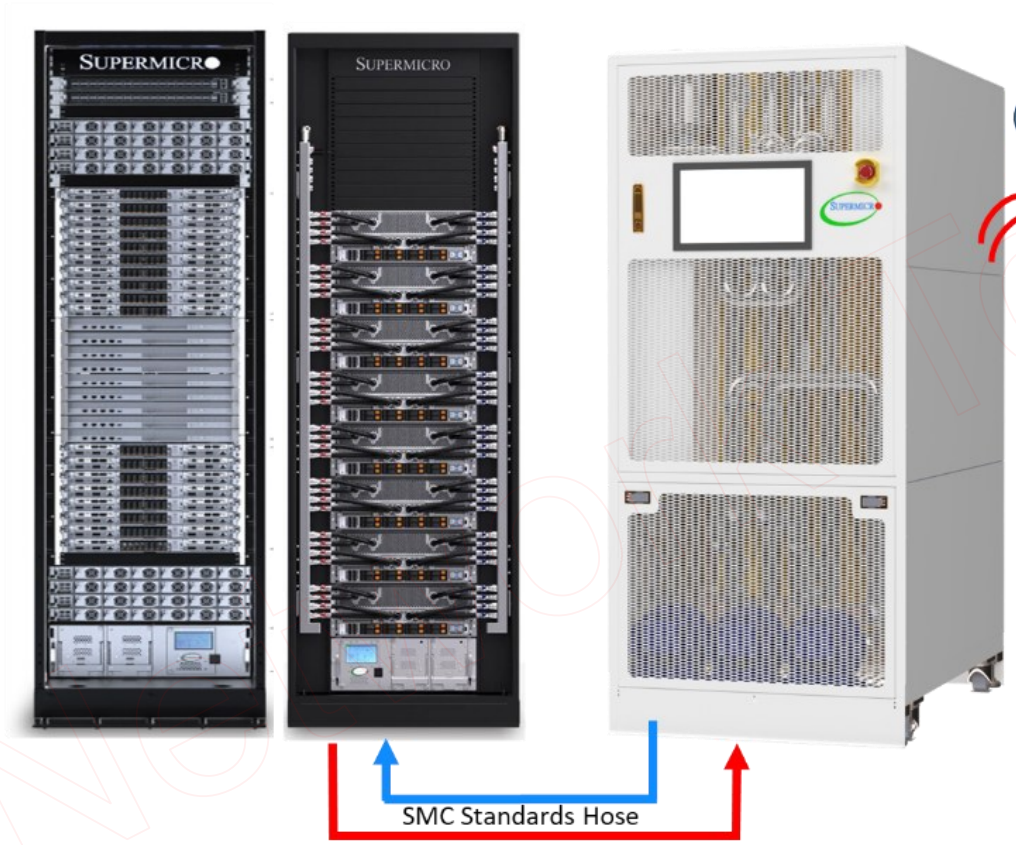
Air vs DLC solution difference overview

Project need	Air Cooling 1x system per rack	Liquid Cooling 8x System per rack	Difference
Number of racks	**144**	**18 + 18x Side car**	DC footprint + CDU vs Side Car price is similar
Rack only price difference: **3000$ per rack**	144x 3000$ = **432.000$**	18x 3000$= **54.000$**	**~378.000$** price difference for rack only
PDU only price difference: **1000$ per PDU** **(2x PDU per rack for air cooling)** **(6x PDU per rack for DLC cooling)**	144x (2x 1000$) = **288.000$**	18x (6x 1000$)= **108.000$**	**~180.000$** price difference for PDU only
Cooling savings	**PUE 1.8**	**PUE 1.4 - 1.5**	Savings on electricity for cooling
Power consumption per system	**14.5kW**	**13.2kW**	Power difference Air vs DLC on system level is **1.3 KW**
Power consumption Cluster	**14.5kWx 144 = 2.08 MW**	**13.2kW x 144 = 1.90 MW**	Power difference Air vs DLC on cluster level is **189 KW**

General idea:

Invest in Data Center to support Cooling efficient technologies

Supermicro's Sidecar Liquid to Air Solution



Cooling Solution Scenario (B200)	
Server (HGX) QTY per Rack	8
B200 QTY per Rack	64x GPUs
Server Power Budget	12 kW
Rack Power Budget	96 kW
Liquid Cooled Heat	88 kW
Rack No per Sidecar	2

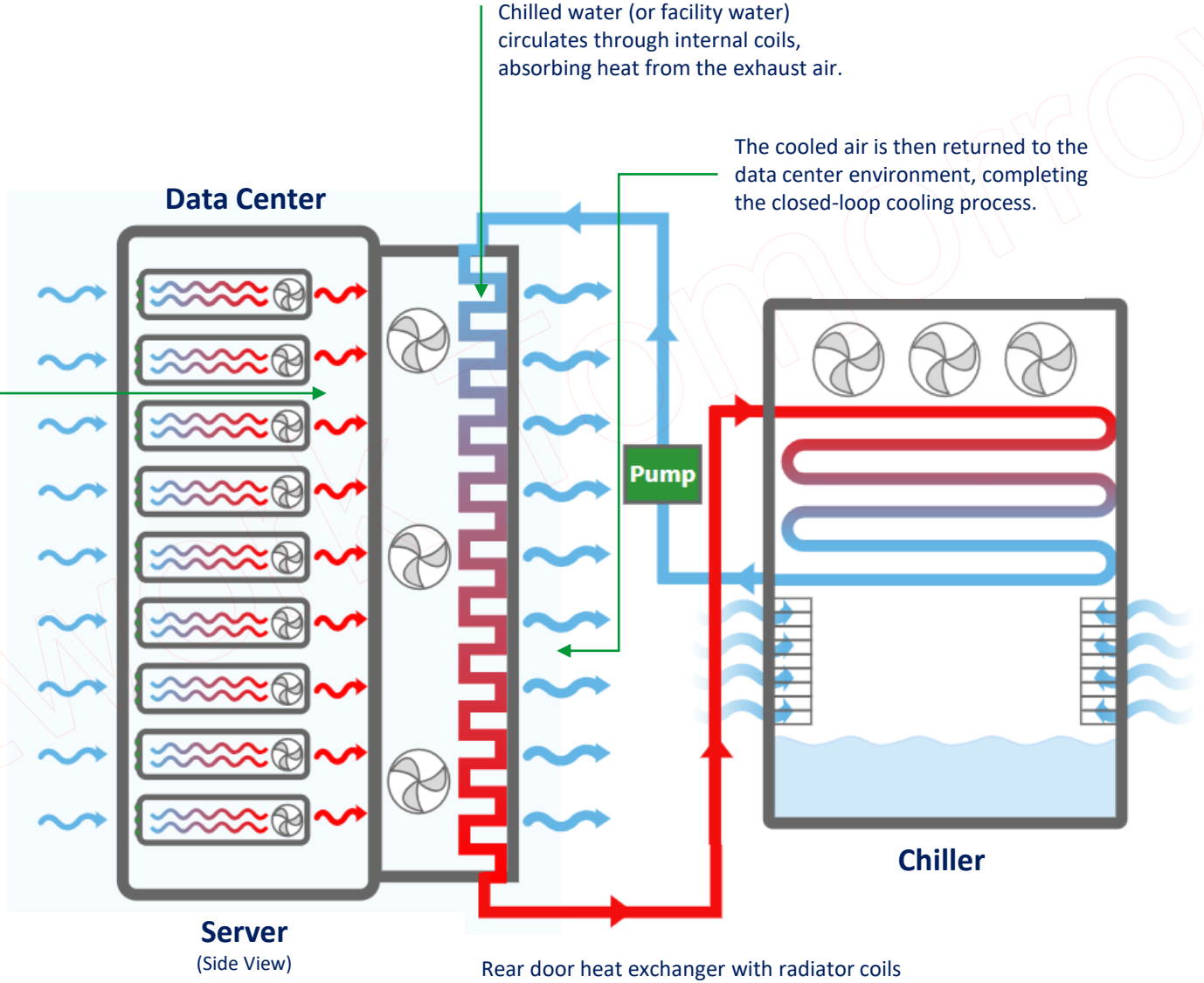
Sidecar Liquid to Air CDU Solutions

Item	Specification
Nominal Cooling Capacity	200kW
Pump Redundancy	N+1
Integrated Variable Speed Drives (VFD's)	Yes
Power Consumption (kW)	19
Dimensions (H x W x D) (mm)	2,000 x 1,100 x 1,415
Weight	<1,20 kg, including coolant <1,000 kg, excluding coolant
Dual Power Feeds (ATS)	Standard
Secondary Connection	2.5" sanitary clamp fittings top and bottom
Ambient Temp	5° to 40° C, 8% to 85% (RH)
Coolant	Propylene Glycol 25wt% 200L
Monitoring	Temperature, Pressure, Flowrate, Water quality, Leak, Power Consumption (redundant)
Communications	Web UI, Modbus, SNMP, Redfish/Restful API

Side Car On-Site Deployment

Air Cooled Racks

What is RDHx?



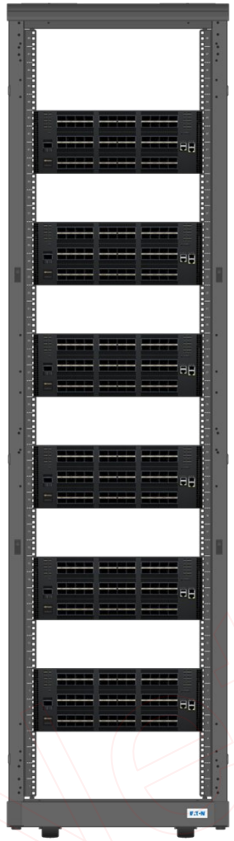
Why RDHx?

- Scalable cooling without infrastructure overhaul
  - Installation in standard racks
  - Reduces strain on room-level systems as rack power climbs
- Flexible deployment and minimal disruption
  - Facilities can deploy them incrementally, starting with one rack and scaling as needed
  - Most suitable liquid-cooling entry solution for legacy data center retrofits.
- More computing capacity within the same footprint
- Building readiness for long-term cooling
- Reduce Data Center build cost!
- Energy Efficiency!

Network Rack elevation

InfiniBand

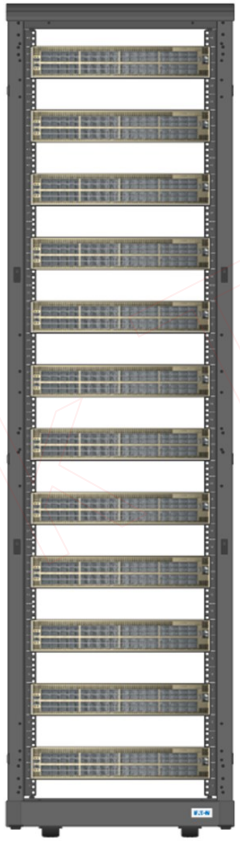
East - West



Power Consumption per Switch	**Max power: 7,000W**
Power Consumption per Rack	**Max power: 42,000W**

Ethernet

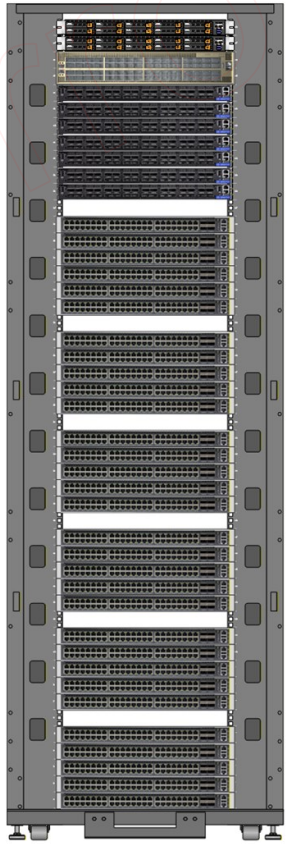
North - South



Power Consumption per Switch	**Max power: 2,080W**
Power Consumption per Rack	**Max power: 25,000W**

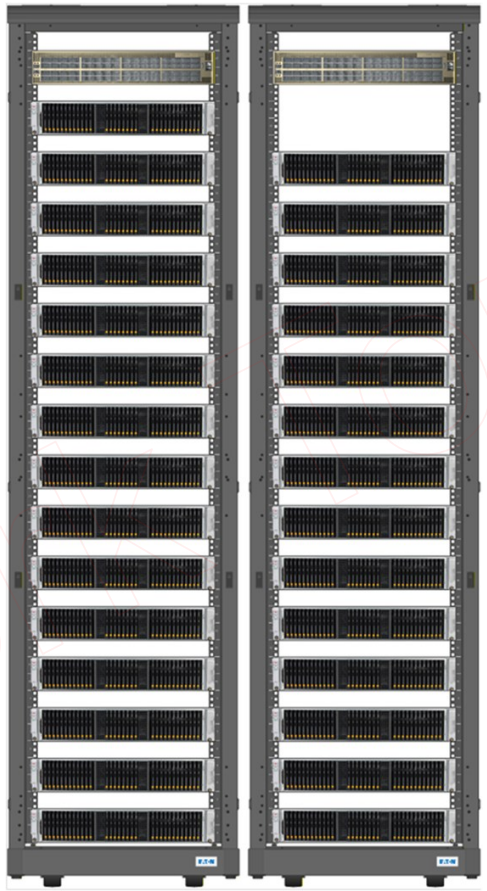
Ethernet

OOB / IPMI



Power Consumption per Switch	1x SN5610: 2.08KW 31x MSN2201: 98W 7x SN4700: 630W 2x UFM: 826.6W
Power Consumption per Rack	**Max power: 9,104W**

Storage Rack elevation



All Flash Storage Solution
5PB Usable

Power Consumption per Server	**Max power per server: 1,617W** **SN5610 Typical power with 64 optical modules: 2.08KW**
Power Consumption per 1x Rack	**Max power: 24.257W**

SMCI RDHx Features Overview

Form Factor Design
- Fits 42 to 52U height.
- 600 – 800mm (23.62" - 31.50") width

Control System
- 7” touch-screen HMI display.
- Cooling modes programmable.
- Temperature, pressure , flowrate monitoring.
- Condensation and leak detection features.
- SNMP, Modbus, Redfish API protocols
- SMC SuperCloud Composer, and WebUI monitoring interfaces.

LED Status Indicator
- Visual monitoring of system health

*42U height ready in late Q4 25’, for early deployment project , please contact us

Primary Loop Design
- Quick-disconnect design to hose for easy commissioning and deployment.
- Top or bottom feed liquid design available.
- Electronic valve and flowrate sensor to control and monitor primary circuit.

4-5 Hot-Swappable EC centrifugal Fan
- Fast and easy replacement.
- Redundant fans features with zero down time during service
- Optimized performance and power efficiency.
- Temperature and differential pressure control modes.

Coil(Heat exchanger)
- 35kW - Max 80KW cooling capability per rack

Super Micro Computer, Inc. may make changes to specifications and product descriptions at any time, without notice. The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors. Any performance tests and ratings are measured using systems that reflect the approximate performance of Super Micro Computer, Inc. products as measured by those tests. Any differences in software or hardware configuration may affect actual performance, and Super Micro Computer, Inc. does not control the design or implementation of third party benchmarks or websites referenced in this document. The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to any changes in product and/or roadmap, component and hardware revision changes, new model and/or product releases, software changes, firmware changes, or the like. Super Micro Computer, Inc. assumes no obligation to update or otherwise correct or revise this information.

SUPER MICRO COMPUTER, INC. MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

SUPER MICRO COMPUTER, INC. SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL SUPER MICRO COMPUTER, INC. BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF SUPER MICRO COMPUTER, Inc. IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

ATTRIBUTION

www.supermicro.com