Thermal Techniques for Data Center Compute Density

Tom Garvens

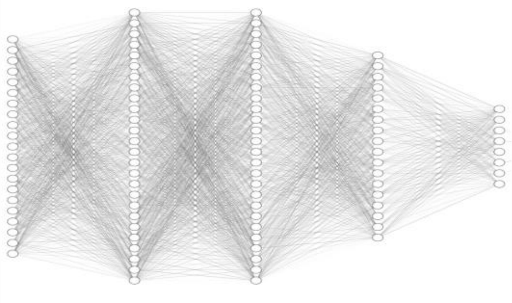
Supermicro

VP Hardware Solutions

Agenda

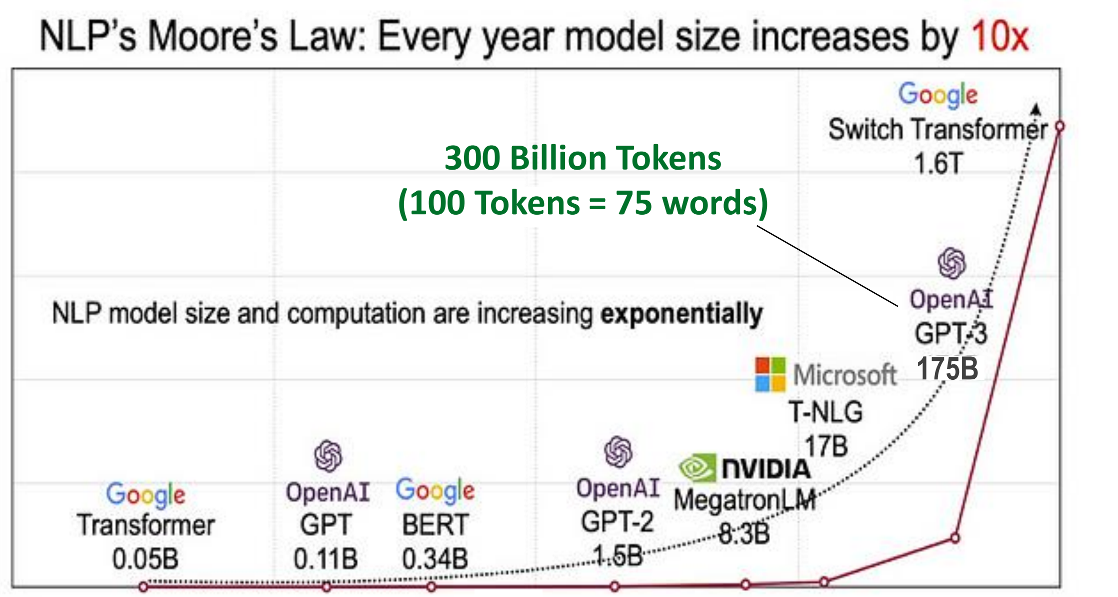
- GenAI LLM Era
- Data Center Power and Cooling Challenges
- Solutions and TCO
- Future Trends

GenAI LLM Era



LLM Parameters (GPT-3 175B) are like adjustable dials in a complex machine.

More adjustments = more optimization (for LLM's this is more nuanced text)


NLP's Moore's Law: Every year model size increases by 10x

ChatGPT

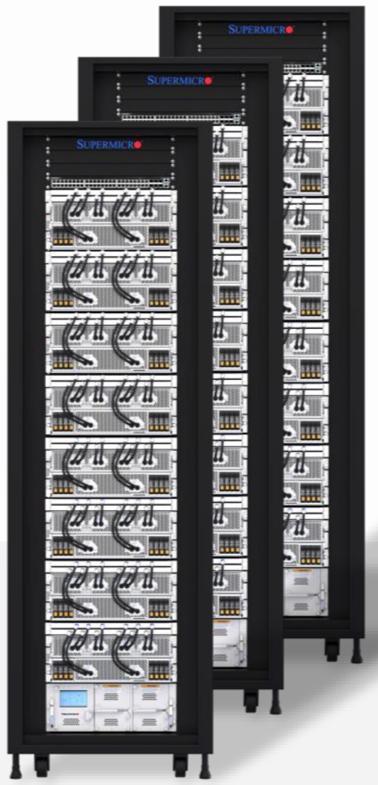
Gemini

LLMs

+

Petascale Data Sets

=

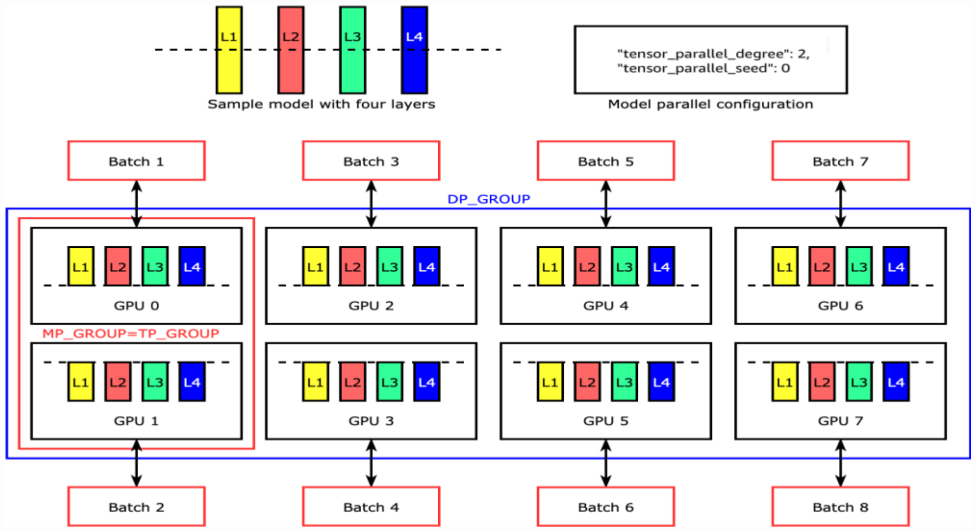


Massive GPU Compute

LLM Training: Tensor Parallelism & Model Pipelines

Models and Data Sets must be sub-divided to fit into GPU memory for performance (time) optimization

GPT-3: 175 billion parameters and 300 billion tokens. On 1024 A100 GPUs (80GB HBM each) -> 140 TFLOPs per GPU and the time required to train is 34 days.

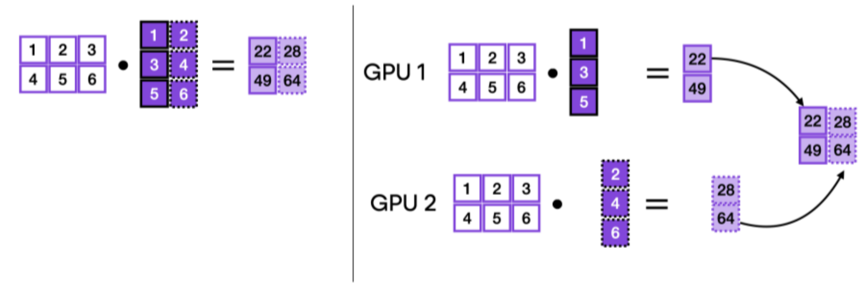


```
"tensor_parallel_degree": 2,
"tensor_parallel_seed": 0
```

Tensor Parallelism

- Reduces the required pipeline depth
- Enables matrix operations across GPU's

For example, split the matrix multiplication by column



$\begin{bmatrix}1&2&3\\4&5&6\end{bmatrix} \cdot \begin{bmatrix}1&2\\3&4\\5&6\end{bmatrix} = \begin{bmatrix}22&28\\49&64\end{bmatrix}$

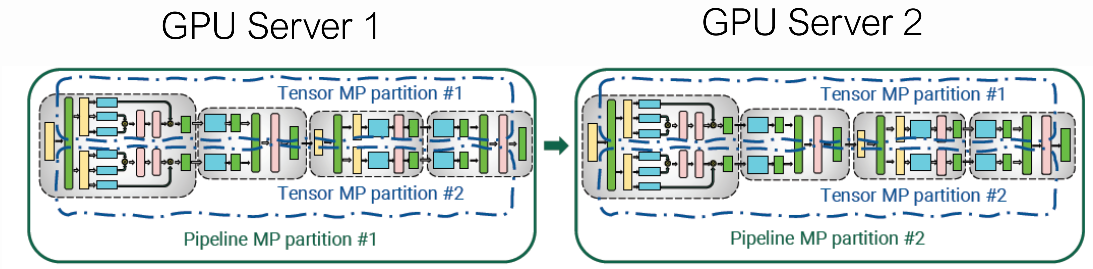
GPU 1 $\begin{bmatrix}1&2&3\\4&5&6\end{bmatrix} \cdot \begin{bmatrix}1\\3\\5\end{bmatrix} = \begin{bmatrix}22\\49\end{bmatrix}$

GPU 2 $\begin{bmatrix}1&2&3\\4&5&6\end{bmatrix} \cdot \begin{bmatrix}2\\4\\6\end{bmatrix} = \begin{bmatrix}28\\64\end{bmatrix}$

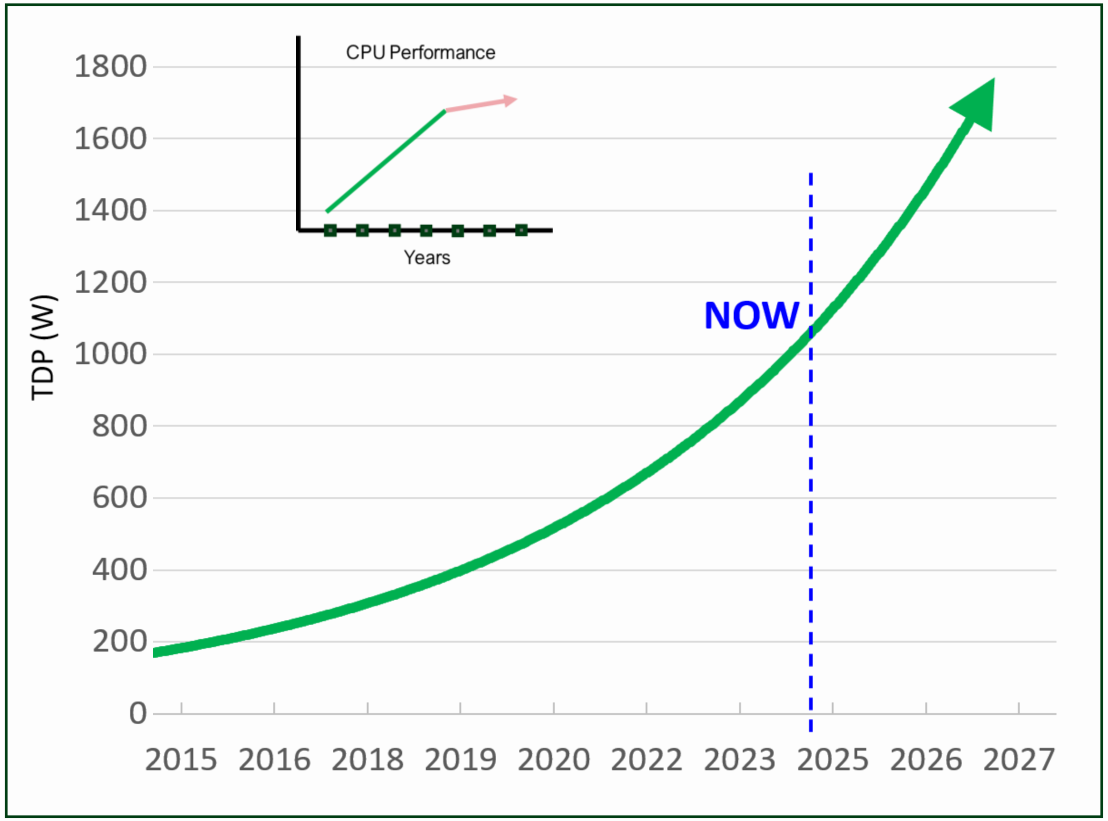
$\begin{bmatrix}22&28\\49&64\end{bmatrix}$

Tensor Processing

- Heavy linear algebra matrix multiplication with bulk data transfers between GPU's
- Partition size defined by GPU memory in coherent domain

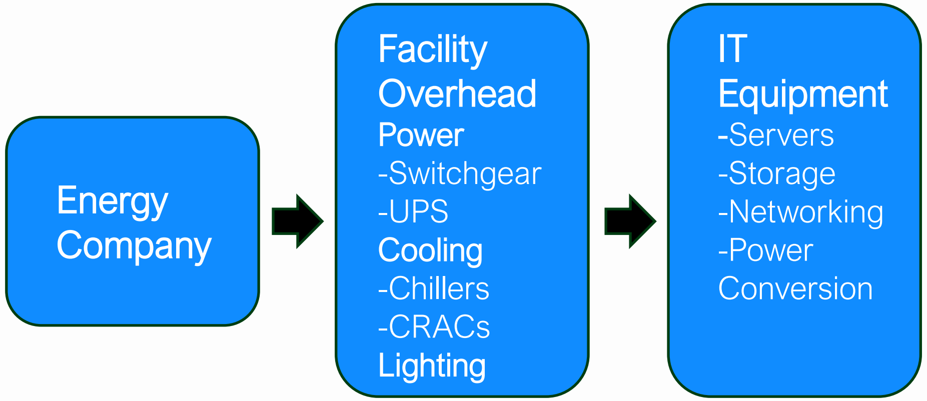


Data Center Challenges – Rise of the GPU



- xPU Thermal Design Points increasing
  - CPUs getting hotter (500W+ H2, 2024)
  - GPUs and AI Accelerators significantly hotter and more power hungry (1000W+ H2, 2024)
  - AI GPU training servers consume 10kW+ per server
- Silicon max temp specs decreasing
- Thermal density continues to compress at the silicon and system level
- Most data centers were designed for general compute & storage resulting in severe power & cooling constraints for today’s AI Era
  - <15kW/rack is common. 120kW/rack is here.
  - Long lead times to build new data centers

Data Center Power Efficiency (PUE)



Energy Company

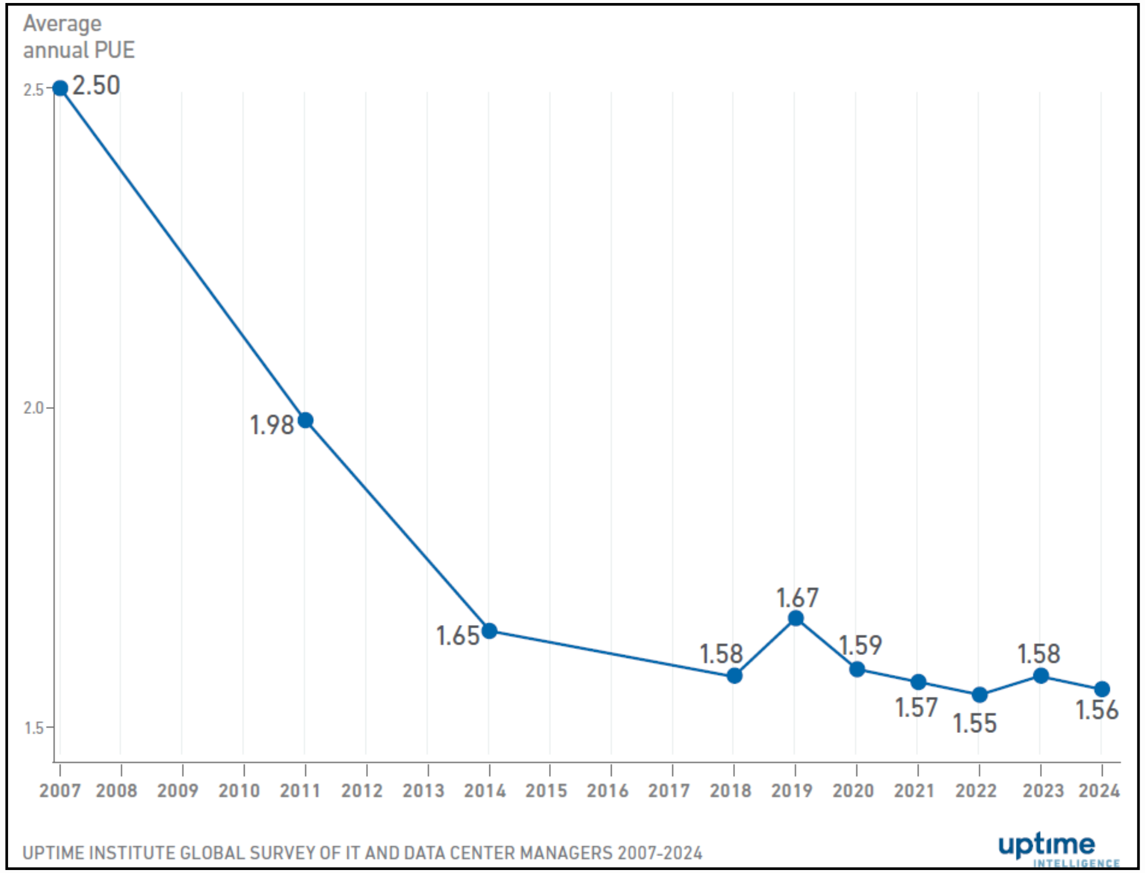
Facility Overhead Power
-Switchgear
-UPS
Cooling
-Chillers
-CRACs
Lighting

IT Equipment
-Servers
-Storage
-Networking
-Power Conversion

$$\text{PUE} = \frac{\text{Total Facility Power}}{\text{IT Equipment Power}}$$

PUE	Level of Efficiency
3.0	Very Inefficient
2.5	Inefficient
2.0	Average
1.5	Efficient
1.2	Very Efficient
1.05	**Extremely Efficient**

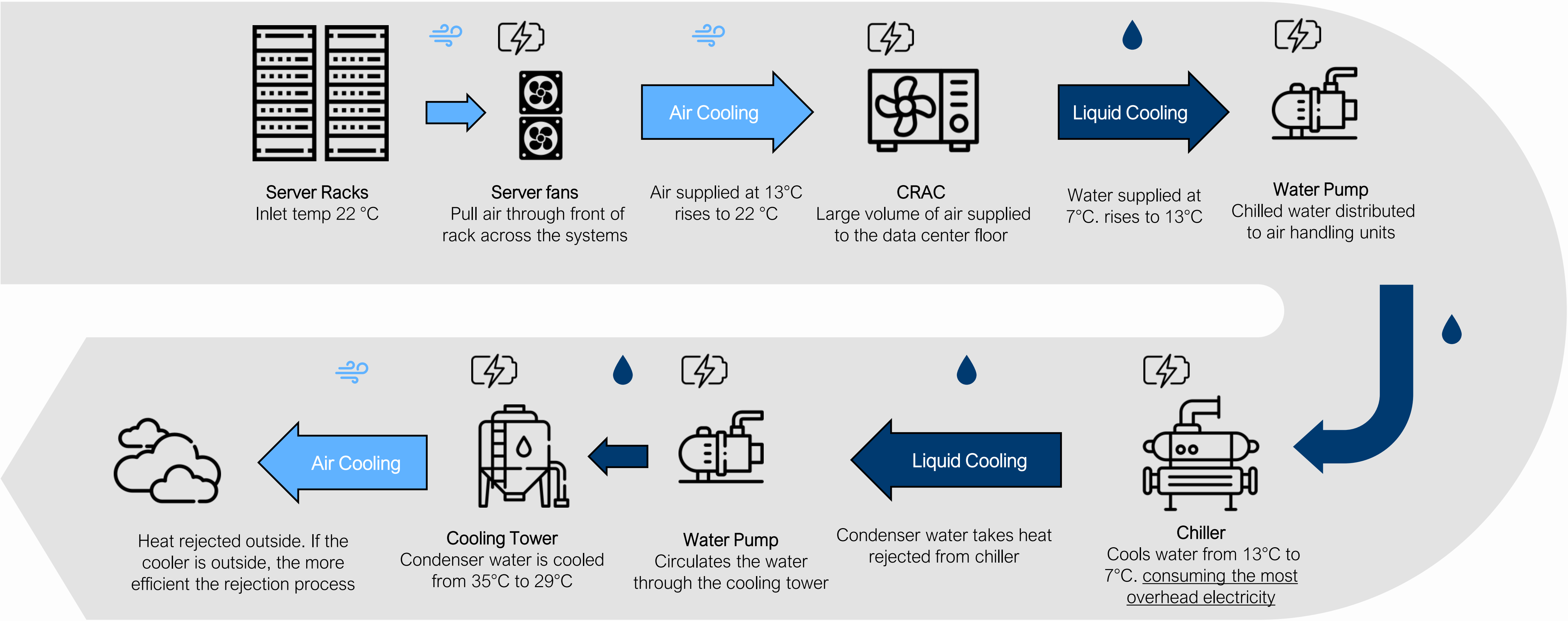
PUE of 1.0 means that the data center is perfectly efficient



UPTIME INSTITUTE GLOBAL SURVEY OF IT AND DATA CENTER MANAGERS 2007-2024

Air Cooling

The largest energy consumers in the traditional cooling loop are the fans & CRAC blowing air into data center floors and the chiller that provides chilled water. Miles of vents in larger facilities.



Server Racks
Inlet temp 22 °C

Server fans
Pull air through front of rack across the systems

Air supplied at 13°C rises to 22 °C

CRAC
Large volume of air supplied to the data center floor

Water supplied at 7°C. rises to 13°C

Water Pump
Chilled water distributed to air handling units

Chiller
Cools water from 13°C to 7°C. consuming the most overhead electricity

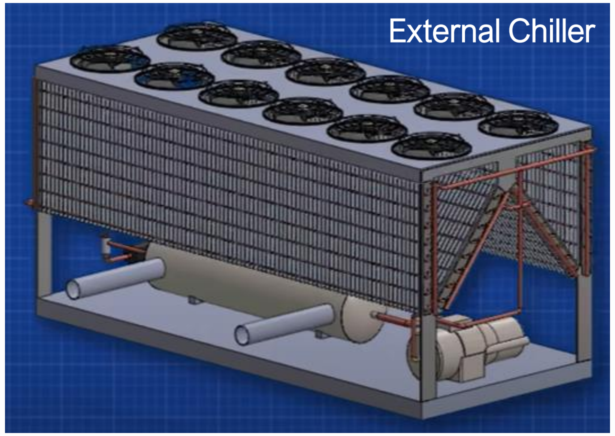
Condenser water takes heat rejected from chiller

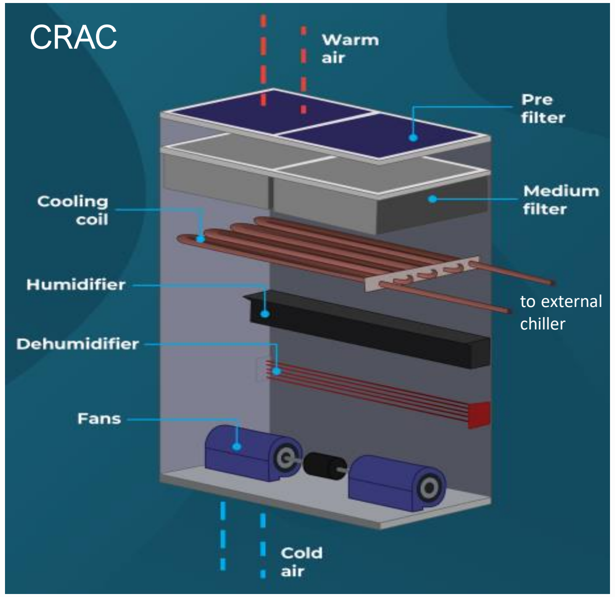
Water Pump
Circulates the water through the cooling tower

Cooling Tower
Condenser water is cooled from 35°C to 29°C

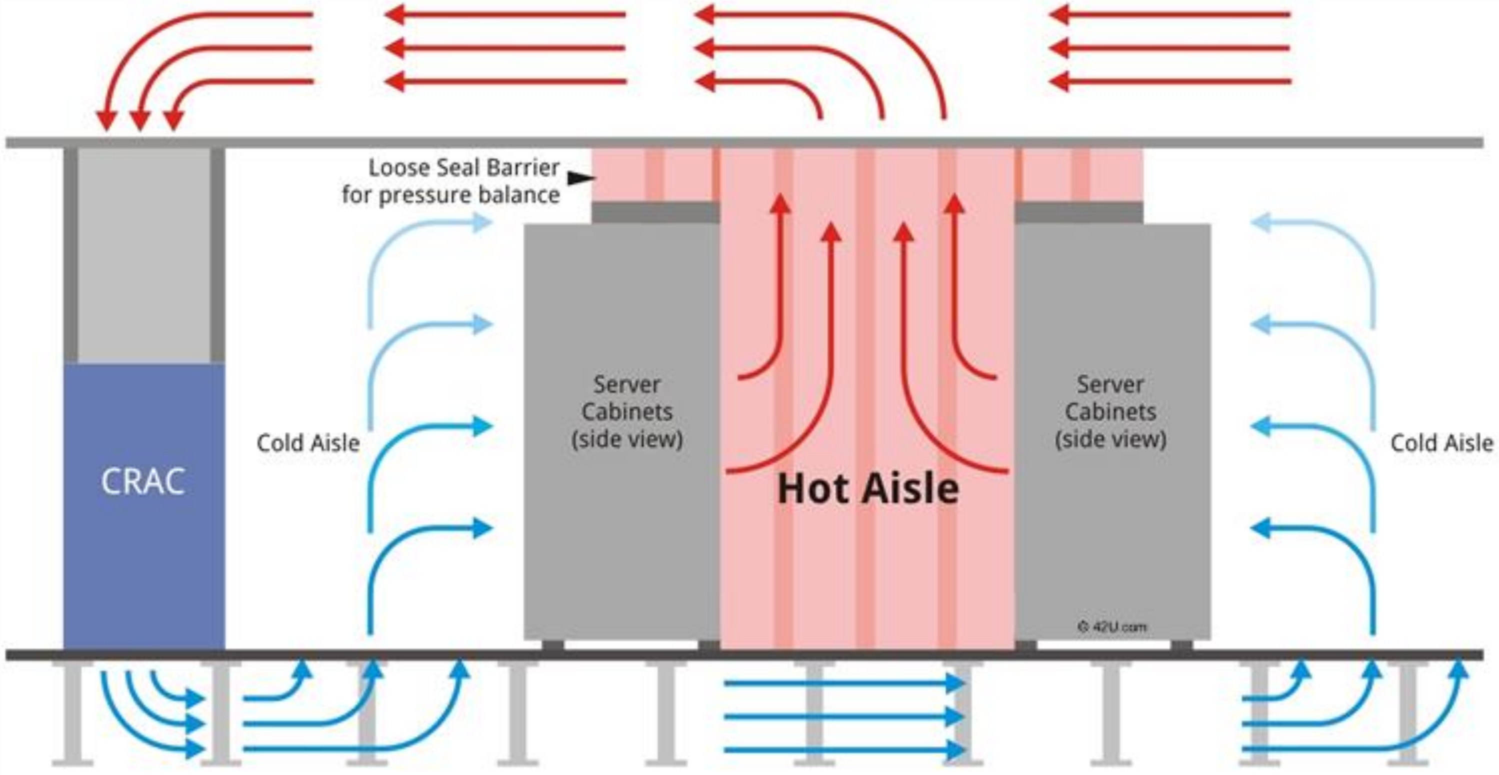
Heat rejected outside. If the cooler is outside, the more efficient the rejection process

Optimized Air-Cooled Data Center



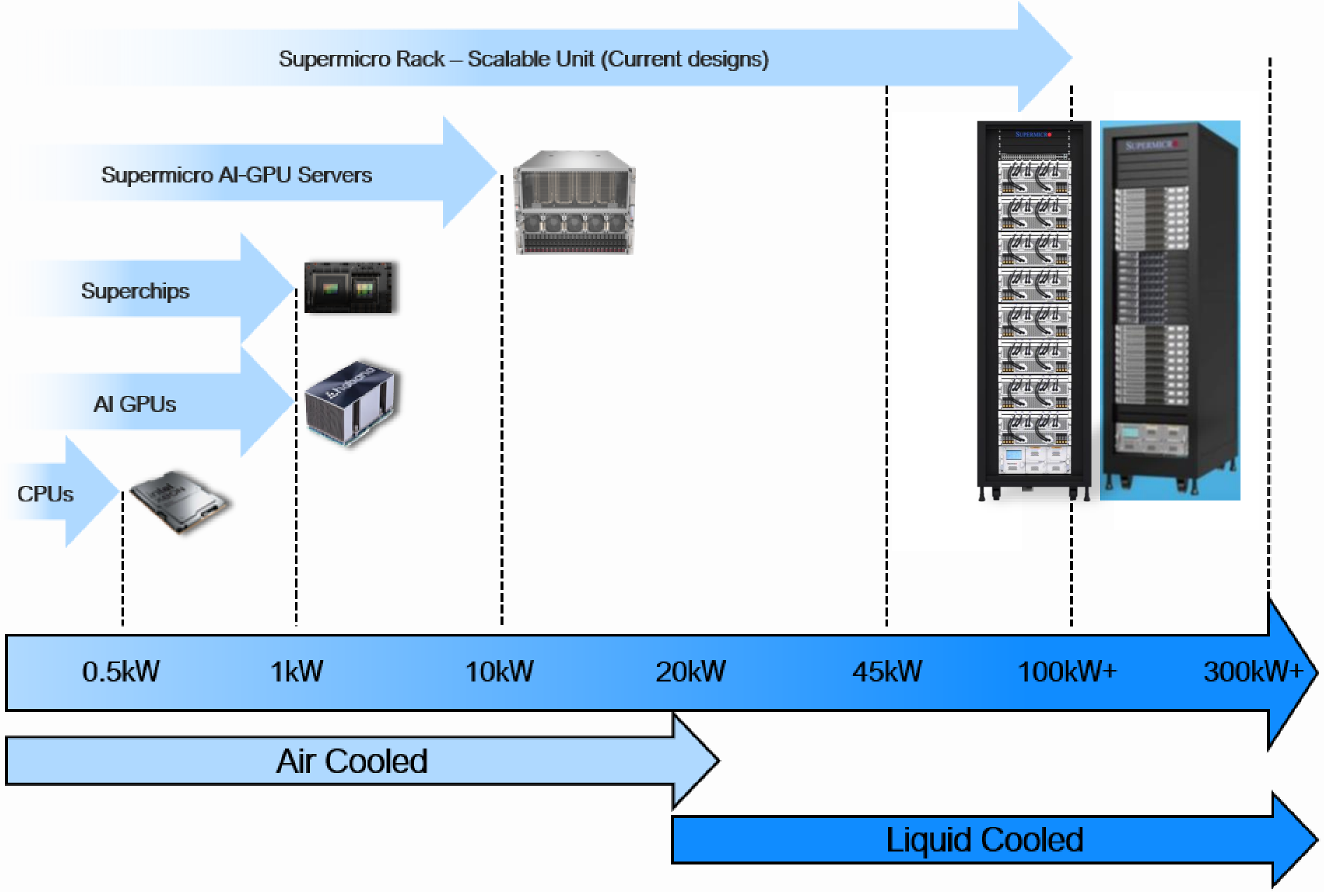


Hot Aisle Contained Data Center Thermal System


Loose Seal Barrier for pressure balance
CRAC
Cold Aisle
Server Cabinets (side view)
Hot Aisle
Server Cabinets (side view)
Cold Aisle
© 42U.com

The Opportunity: Liquid-Cooled Data Centers

Liquid-cooling vastly reduces power costs compared to air-cooling, reducing customer TCO while minimizing environmental impacts.



Water has significantly higher thermal conductivity than air (molecules are closer together and have stronger bonds)

Up to

92%

Reduction of server cooling power

Up to

40%

reduction in electricity costs for entire data center

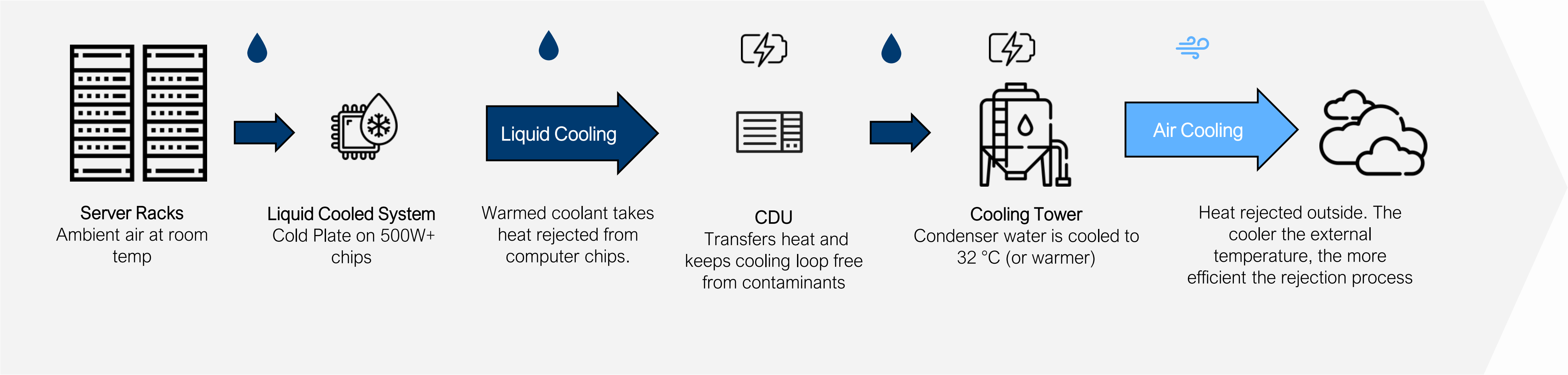
Up to

55%

reduction in data center server noise

Liquid Cooling

With a liquid cooling system, large energy-consuming components are eliminated.



Server Racks
Ambient air at room temp

Liquid Cooled System
Cold Plate on 500W+ chips

Warmed coolant takes heat rejected from computer chips.

CDU
Transfers heat and keeps cooling loop free from contaminants

Cooling Tower
Condenser water is cooled to 32 °C (or warmer)

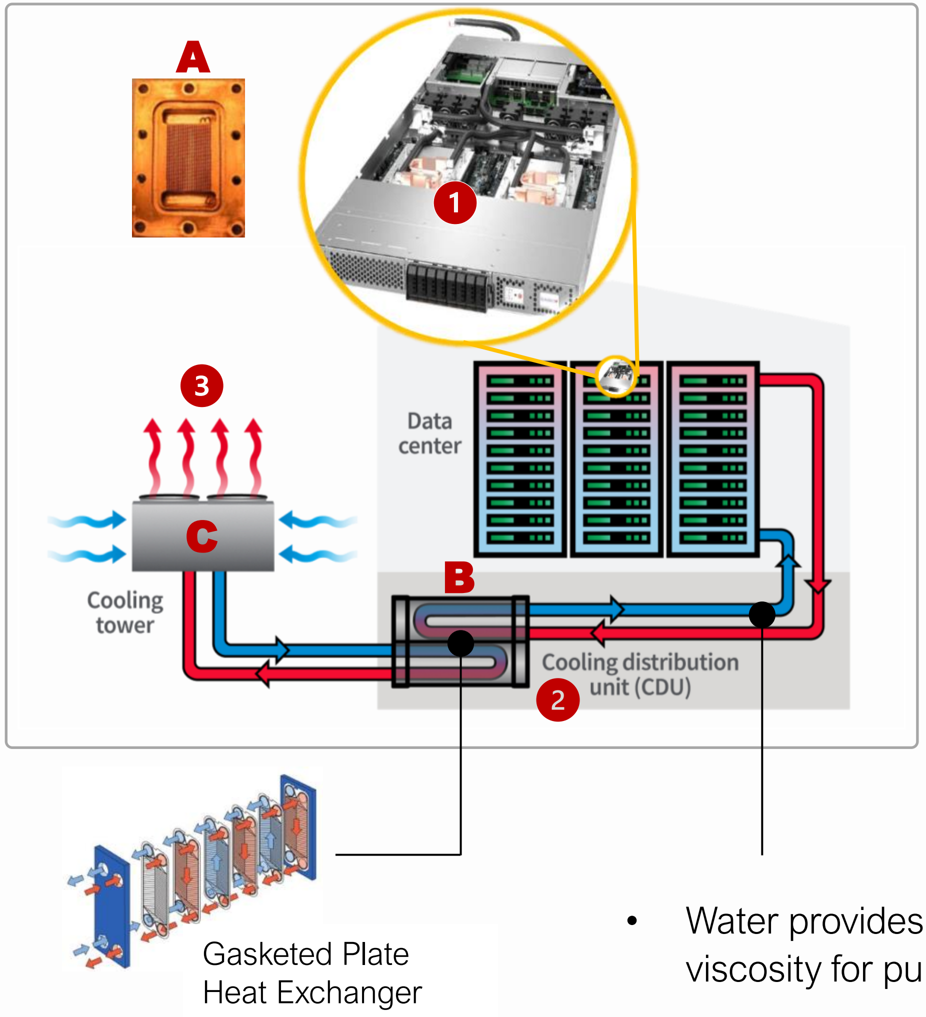
Heat rejected outside. The cooler the external temperature, the more efficient the rejection process

Fans

CRAC

Chiller

Direct Liquid Cooing (DLC) Overview


A
1
3
C
Cooling tower
Data center
B
Cooling distribution unit (CDU)
2
Gasketed Plate Heat Exchanger

How it Works

1 Liquid coolant flows through cold plates directly mounted on the heat-generating server components (such as CPUs or GPUs)

2 The heated liquid is then cooled through a liquid-to-liquid CDU, either contained within the rack or externally. Each cooling loop is isolated.

3 An external system cools the liquid through a liquid-to-liquid process using a Cooling Tower

Key Components

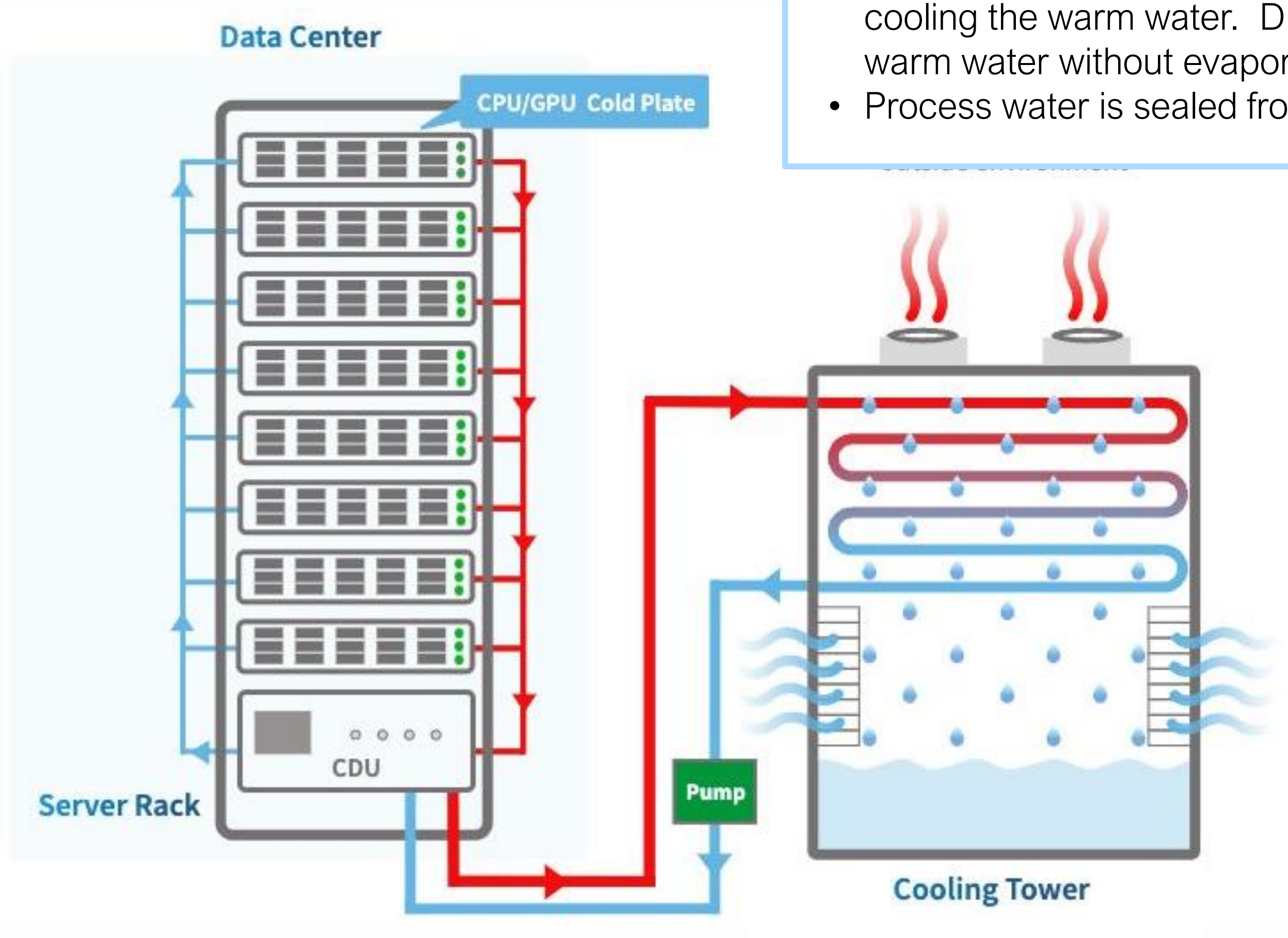
A	Cold Plates (Cooling units)
B	Cooling Distribution Unit (CDU)
C	Cooling Tower

NOTE: Server system fans still needed to operate at low levels to cool peripheral components (memory, storage, voltage regulation)

- Water provides high heat capture, but is often mixed with glycol - reduces heat capture but increases viscosity for pumping efficiency, lowers freezing point, and adds corrosion/bio-growth benefits
- Nano-fluid water additives can improve heat transfer further by evenly suspending nano-particulates in glycol-based solutions

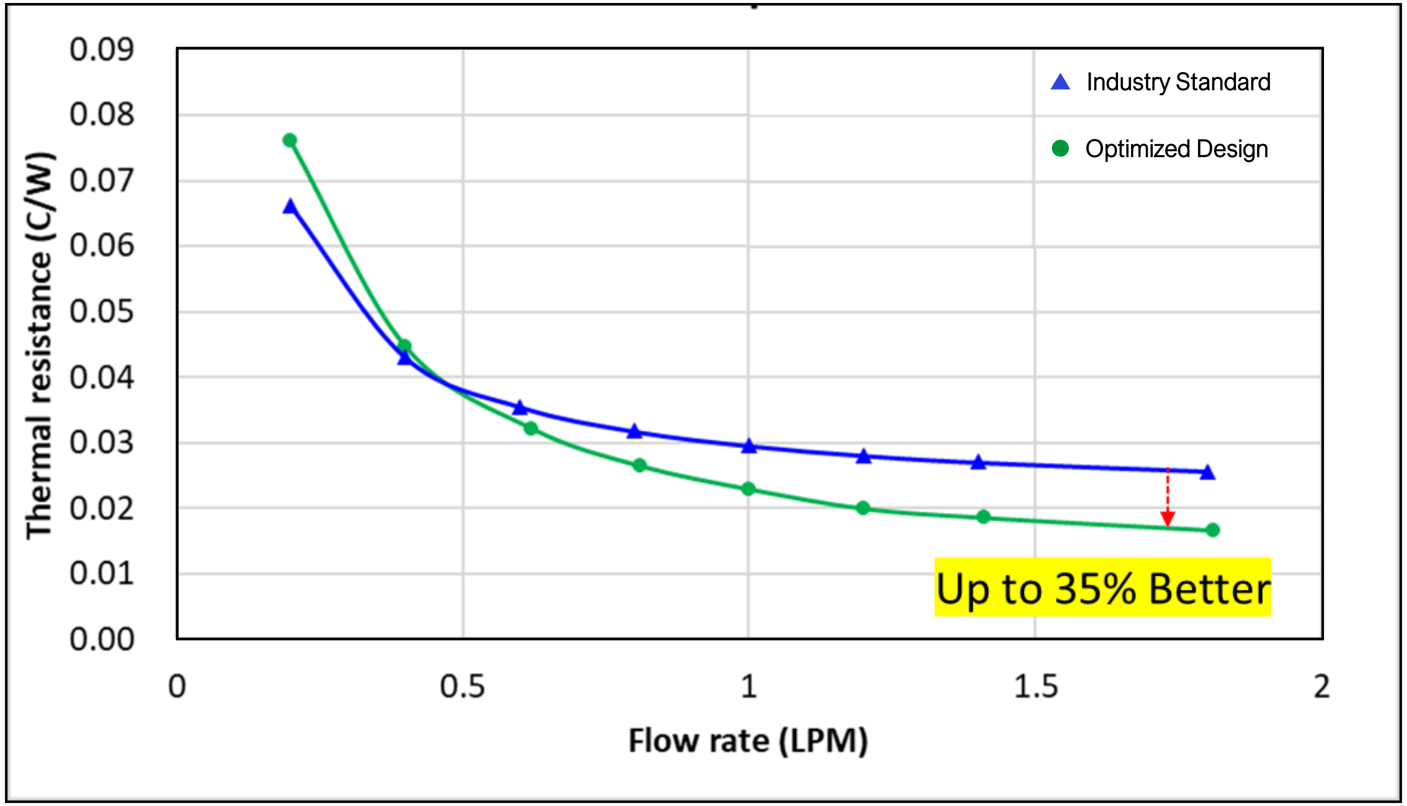
DLC Cooling Tower

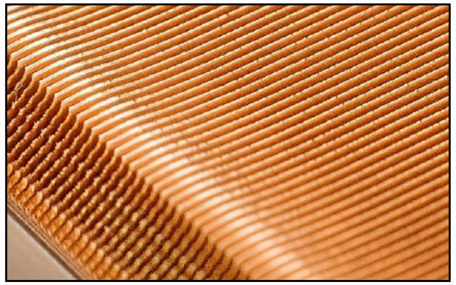
- Warm process water flows through Cooling Tower coils
- Wet Cooler: An isolated water source is sprayed over the coils, cooling the warm water. Dry coolers use ambient air to cool the warm water without evaporation
- Process water is sealed from pollutants

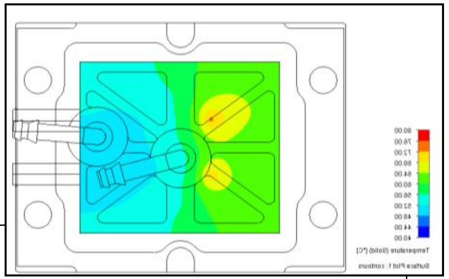


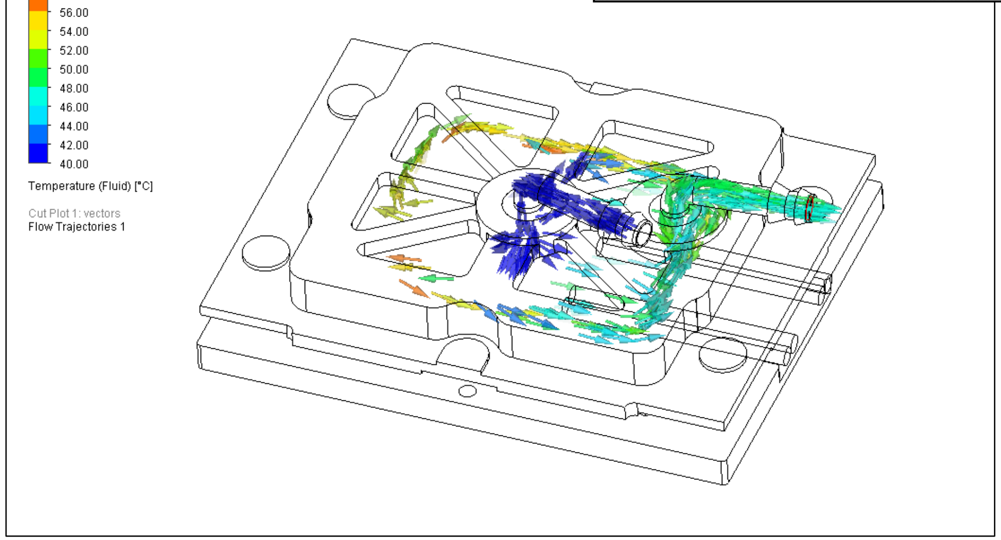
Cold Plate Design Optimization

Micro-Channel Heat Transfer Optimization: analytical design methods incorporating silicon heatmap, CFD, and additive manufacturing capabilities used to improve performance compared with market standard designs









CDU Design Tenets

In-Rack CDU

Front

Back

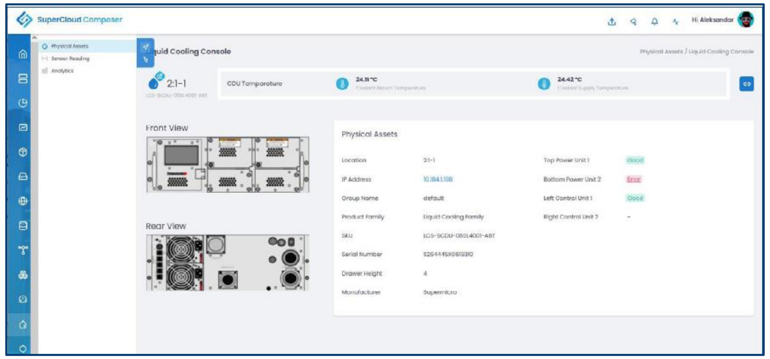
Capacity (example)

- 100kW load @ 32C facility water
- 60kW load @ 40C facility water

Features, Availability and Manageability

- Automatic anti-condensation control
- Supports up to 45C facility water
- N+1 hot-swappable pumps and PSU’s
  - Replace pumps in 2 minutes
  - Replace PSU in 1 minute
- Intelligent monitoring and control
  - Coolant pressure, flow rate and leakage
  - Real time sensor reading with historical data availability
- Touch panel and remote access
- Fully integrated with system management SW

In-Row CDU



Liquid Cooling Console

Physical Assets / Liquid Cooling Console

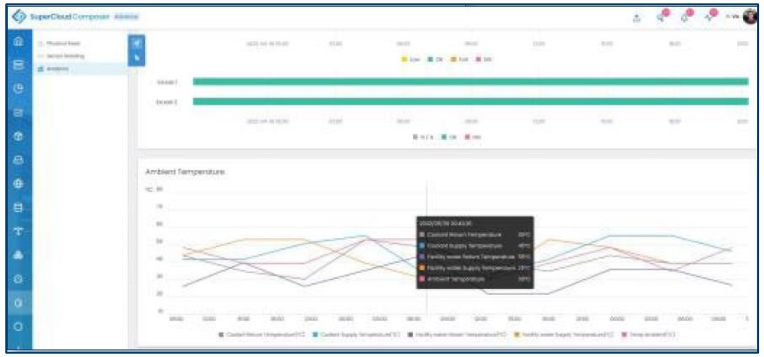
2:1-1

LCS-SCDU-080L4001-ABT

CDU Temperature: 24.11 °C Coolant Return Temperature; 24.42 °C Coolant Supply Temperature

Physical Assets

Location	2:1-1	Top Power Unit 1	Good
IP Address	10.184.1.198	Bottom Power Unit 2	Error
Group Name	default	Left Control Unit 1	Good
Product Family	Liquid Cooling Family	Right Control Unit 2	-
SKU	LCS-SCDU-080L4001-ABT		
Serial Number	S2S4445X0619310		
Drawer Height	4		
Manufacturer	Supermicro		

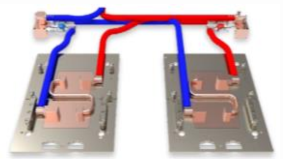


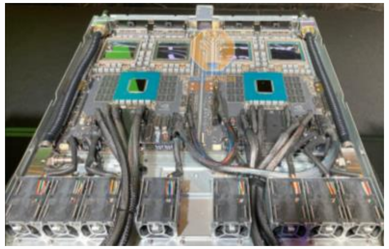
Direct Liquid Cooling Data Center Bonus Benefits

- Liquid can be directed to the heat source (DLC), improving performance
  - Airflow shadowed CPU’s or GPU’s result in inefficient cooling
- Carbon emission reduction directly correlated to power reduction %
- Liquid cooling increases rack and floor density as tall, air cooled heatsinks are not needed.
- Liquid cooled data centers can provide heating for on-site or nearby non-IT environments
- Liquid cooling is the only technology that enables “future” GPU solutions

120kW+

10kW+





Direct Liquid Cooling vs. Air Cooling $ Advantage

“It’s all about power reduction”

CAPEX

1. IT power load determines datacenter build cost
2. 15%+ power load reduction in DLC GPU servers vs. air cooled
3. CAPEX savings from a DLC optimized datacenter pays for DLC cooling infrastructure

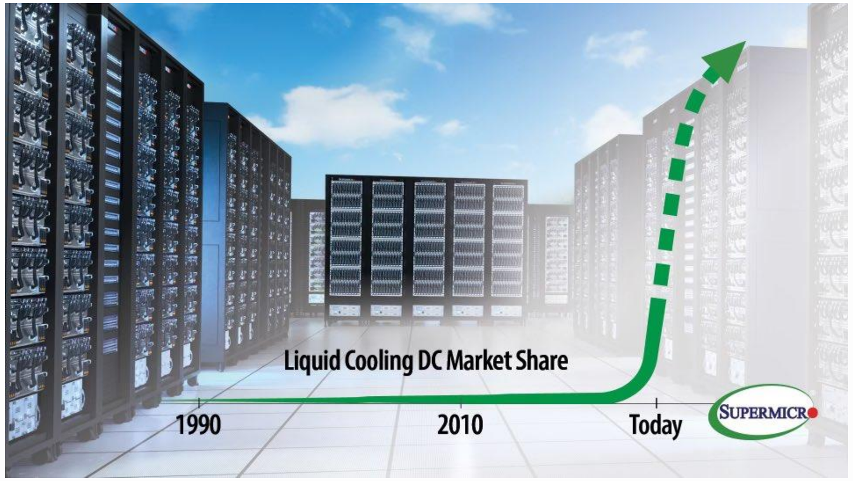
OPEX

1. 15%+ power load reduction in DLC GPU servers vs. air cooled (already stated above)
2. 80%+ power load reduction in Datacenter cooling for a DLC optimized datacenter (reduction of HVAC, compressors, fans, pumps)
3. 40% OPEX savings from lower utility bills (DLC GPU Servers + Datacenter cooling) – Applies to greenfield or retrofit datacenter
4. $60M OPEX savings over 5yrs for 1K DLC GPU servers (8K GPU’s) @ $0.18/kWh

Direct Liquid Cooling Today

DLC Past Concerns	Supermicro DLC Today
Long Lead Time	Delivering DLC in 2-4 weeks with a given forecast
More expensive	Liquid cooling infra can be free in an optimized datacenter
Reliable	Demonstrated performance and uptime at scale

0-3% Net Savings on initial CAPEX – Mid-Size LLM HW Requirements



4U SXM/OAM GPU

8U SXM/OAM GPU

2U2N / 2U4N BigTwin®

SuperBlade®

1U / 2U Hyper

4U4N / 4U8N FatTwin®

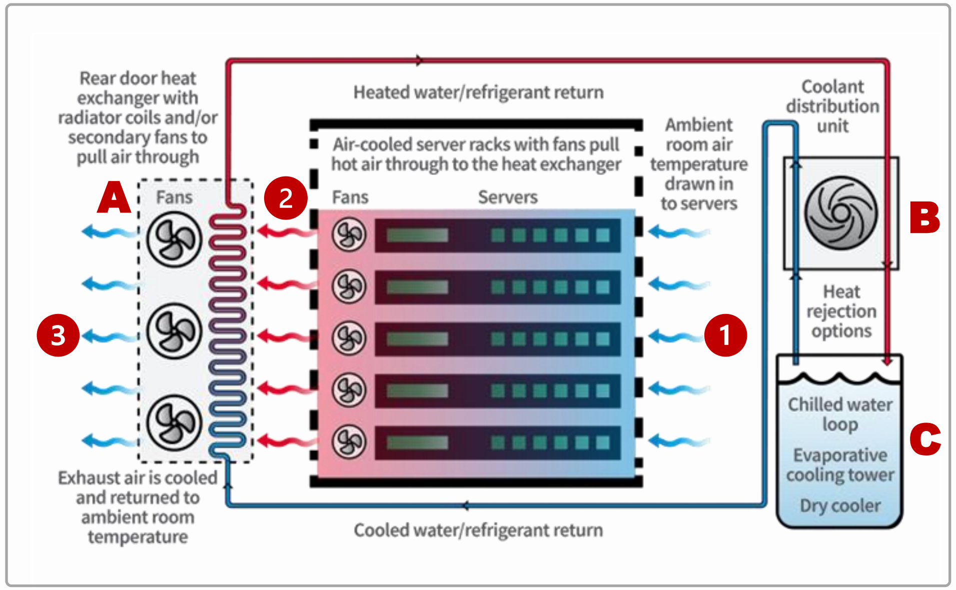
Optimized Cold Plates

https://www.supermicro.com/manuals/brochure/Brochure-Liquid-Cooling-Solutions.pdf

Rear Door Heat Exchanger (RDHx)

This option includes installing a panel on the back of the server rack with a chilled water heat exchanger

RDHx Diagram


Rear door heat exchanger with radiator coils and/or secondary fans to pull air through
Heated water/refrigerant return
Coolant distribution unit
Air-cooled server racks with fans pull hot air through to the heat exchanger
Ambient room air temperature drawn in to servers
A
Fans
2
Fans
Servers
B
3
1
Heat rejection options
Chilled water loop
C
Evaporative cooling tower
Dry cooler
Exhaust air is cooled and returned to ambient room temperature
Cooled water/refrigerant return

How it works

1 Ambient air temperature drawn into the servers

2 Air-cooled server racks push hot air into the RDHx

3 The RDHx and the coolant absorbs the heat, returning cooler air to the data center around ambient temperature

Key Components

A RDHx. With radiator coils and/or secondary fans to pull air through.

B Cooling Distribution Unit (CDU)

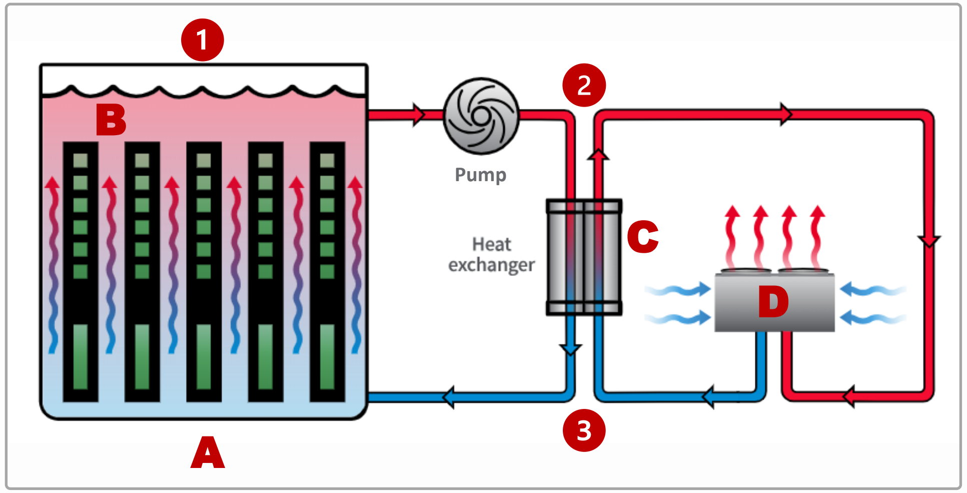
C Water Chiller

This option features efficient heat removal and compatibility with existing air-cooled IT infrastructure. Many customers will use this as a bridge solution to improve efficiency in existing environments.

Immersion Liquid Cooling (ILC)

In this option, the entire server is immersed in a dielectric coolant which cools all components.

Immersion Diagram



How it works

1 Servers or other IT components are submerged in a thermally conductive dielectric liquid

2 The liquid gets pumped to a heat exchanger where heat is transferred to a cooler water circuit

3 The warm liquid is then chilled and brought back into the tank

Key Components

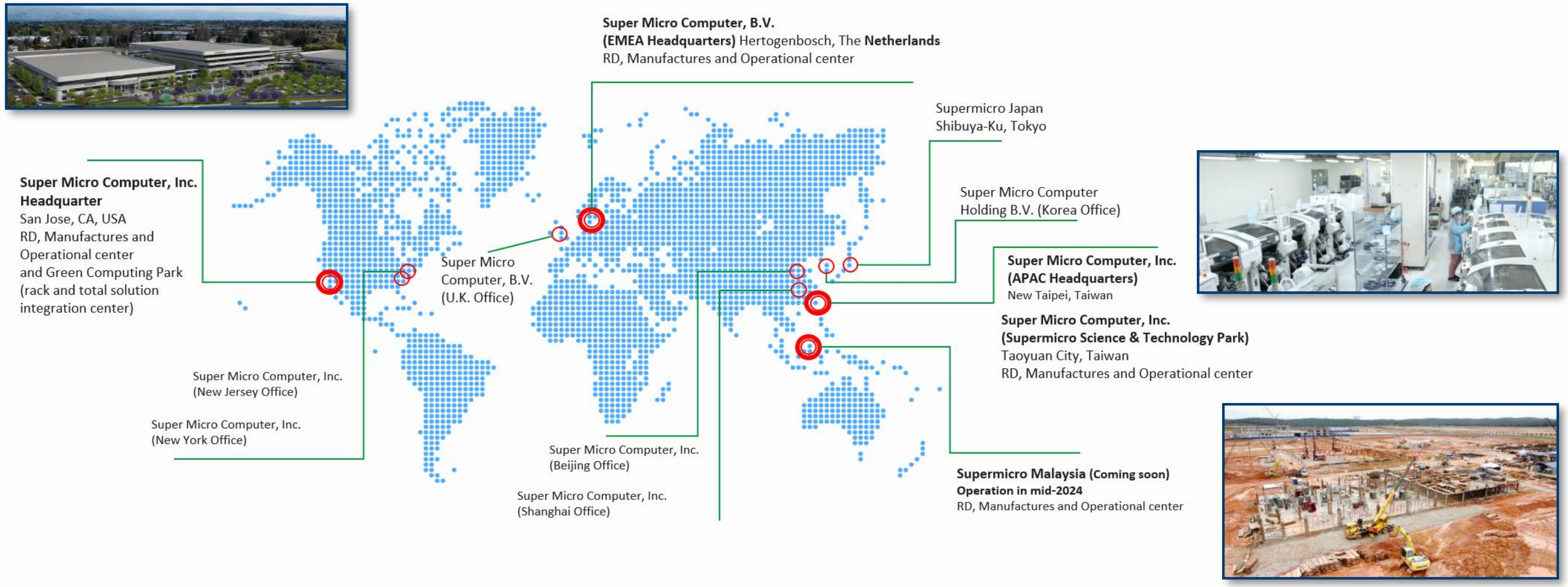
A	Tank
B	Dielectric liquid
C	Liquid-to-liquid heat exchanger
D	Cooling Tower

This option enables excellent efficiency capturing 100% of the thermal energy. Challenges: component compatibility with cable insulation, connectors, transformers, and PCBA solder flux as they were designed for air-cooling. Also, a significant change in serviceability workflow.

Supermicro Overview

Total Revenue	$7.1B (FY2023) $14.9B (FY2024)
Production	Major locations: US, NL, TW and MY (more locations under planning)
Human Resources	6000+ worldwide, >50% engineering

- Design, Manufacturing and Services under the same roof. Optimizes TTM
- Datacenter to Edge; compute, storage, and networking with systems management from chip to cooling tower
- Customer on-site Deployment Services
- Rack-scale design, validation, PCBA through L12 manufacturing and delivery



Super Micro Computer, Inc. Headquarter
San Jose, CA, USA
RD, Manufactures and Operational center
and Green Computing Park
(rack and total solution integration center)

Super Micro Computer, Inc.
(New Jersey Office)

Super Micro Computer, Inc.
(New York Office)

Super Micro Computer, B.V.
(U.K. Office)

Super Micro Computer, B.V.
(EMEA Headquarters) Hertogenbosch, The **Netherlands**
RD, Manufactures and Operational center

Supermicro Japan
Shibuya-Ku, Tokyo

Super Micro Computer
Holding B.V. (Korea Office)

Super Micro Computer, Inc.
(APAC Headquarters)
New Taipei, Taiwan

Super Micro Computer, Inc.
(Supermicro Science & Technology Park)
Taoyuan City, Taiwan
RD, Manufactures and Operational center

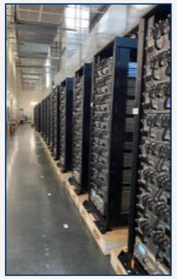
Super Micro Computer, Inc.
(Beijing Office)

Super Micro Computer, Inc.
(Shanghai Office)

Supermicro Malaysia (Coming soon)
Operation in mid-2024
RD, Manufactures and Operational center

5000+ Racks per month global capacity

2000+ DLC Racks per month

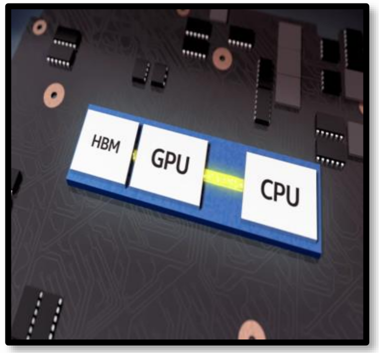


AI Crystal Ball

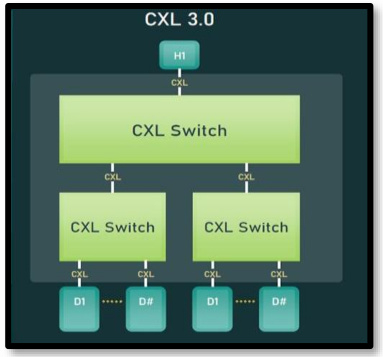
- Domain-specific and country-specific LLMs. Multi-modality
- LLM’s and SLM’s co-exist and unite.
- AI Inference everywhere driving edge and environmental expansion cooling needs
- Further Compute Architecture Optimizations
  - Data Center vs. Edge, Training vs. Inference, Transformer vs. Flexible Accelerator
  - Memory Tiering
- Direct Liquid Cooling is rapidly expanding today. Catalyst for immersion to take off?
- How far is Artificial General Intelligence (AGI)?

Domain-specific LLM

AI Inference everywhere

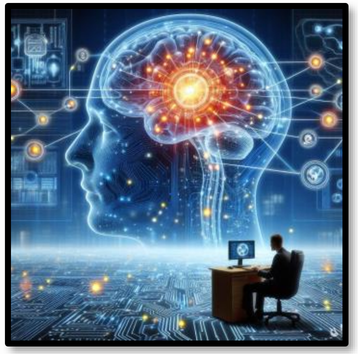


CPU+GPU+HBM



CXL Technology

Liquid Cooling



AGI

Thank You

SUPERMICRO®

www.supermicro.com