LINE Yahoo! US Data Center Technology at the Forefront
LLM (Large Scale Language Models) and the Challenge of Water Cooling Technology

Norifumi Matsuya

Self-introduction

Self-introduction in JAPAN

- **Norifumi Matsuya**

- From 2000 to 2014,
  I was in charge of infrastructure design, construction and operation at Yahoo!

- Main Achievements
  - BGP in production network
  - Obtained AS from JPNIC, APNIC, and ARIN
    (AS23816, AS24572, AS18140)
  - Backbone network virtualization and VPLS
  - Creating Japan's largest Openstack cluster
  - Construction of domestic data centers
    (Tokyo, Osaka, Chubu region, Kyushu region, Tohoku region)

US Data Center Tour '09

After visiting a data center in the US, the decision was made to expand into the US

Self-introduction in the US

Actapio Data Center

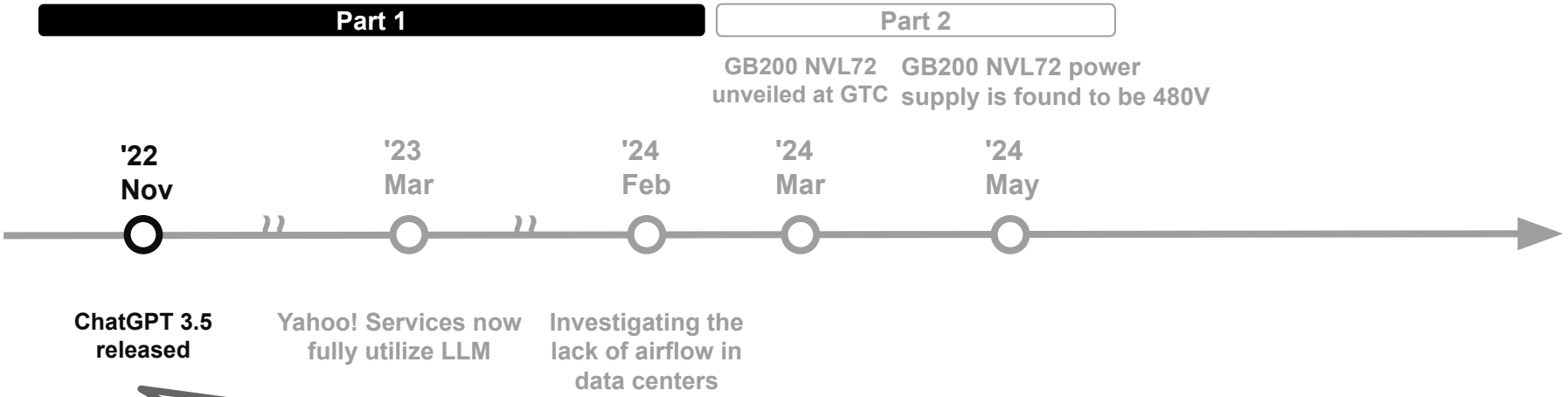
- Moved to the US in 2014 and seconded to Actapio
- Actapio is a US subsidiary of LINE Yahoo Japan Corporation, and
  a company that builds data centers and operates servers for AI and big data.
- Currently, while running a company, I am in charge of OCP (Open Compute Project) server operation and AI business development.

AGENDA

- **Today's announcements' overview**
- **Part 1: Water-cooling technology can solve the problem of insufficient airflow from air conditioners**
  - Emergence of LLM and changes in server resources
  - Current state of data centers and predicted air conditioner air volume problems
  - Considering solutions to insufficient airflow from air conditioners
- **Part 2: Server proximity constraints that can be solved with liquid cooling technology**
  - The emergence of GB200 NVL72 (GPU server) and its background
  - Change power supply to 480V (under investigation)
- **summary**

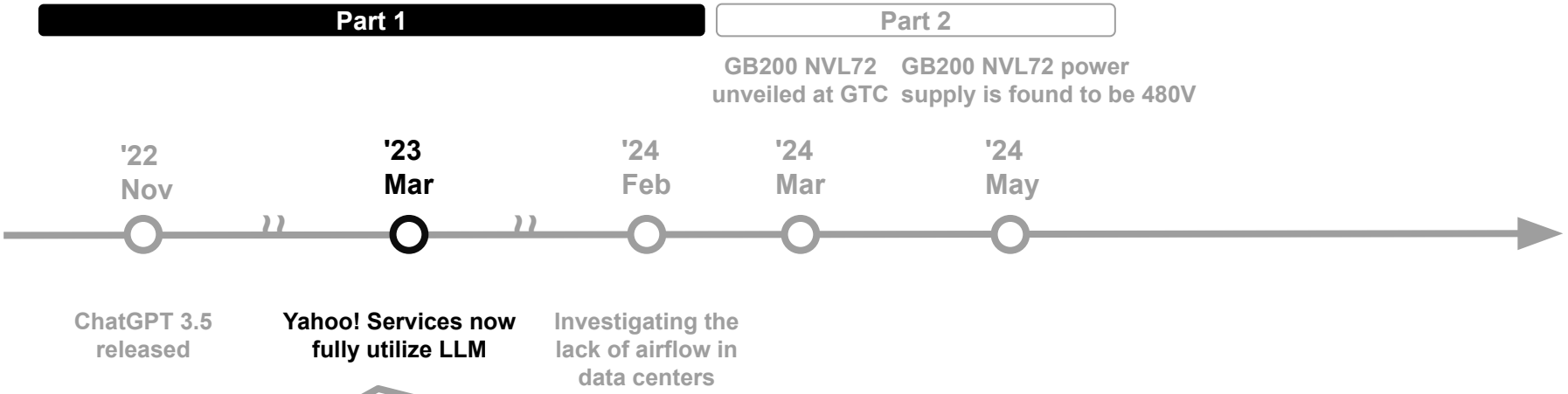
Today's announcements’ overview

Today's announcement



- ChatGPT is a conversational **LLM** (large-scale language model)
- LLM uses **a transformer model** and is trained on **large amounts of text data**
- Transformer is a neural network architecture that demonstrates **high performance in natural language processing**

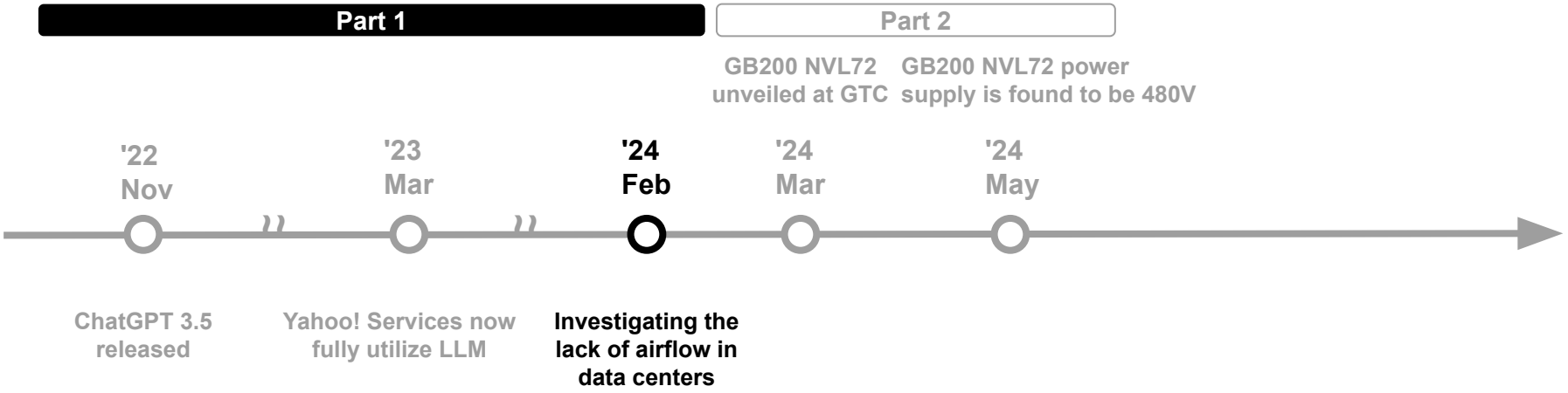
Today's announcement



- LLMs require **large amounts of data** and **enormous computing resources**
- This results in a significant increase in server **workload.**

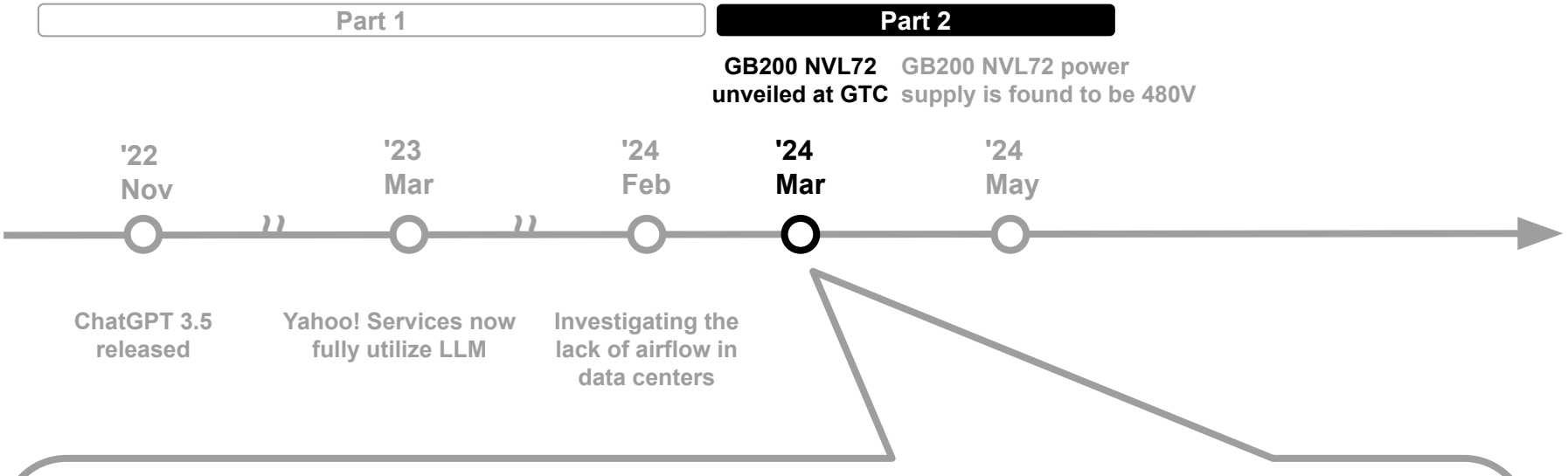
- Check each resource in the data center and confirm that there is no problem
- Obtained information from a U.S. company about insufficient airflow in data center air conditioners

Today's announcement



- **water-cooling technology** can allow
  - **Improved cooling efficiency**
  - **Solves the problem of insufficient airflow**
  - **More servers can be installed**

Today's announcement



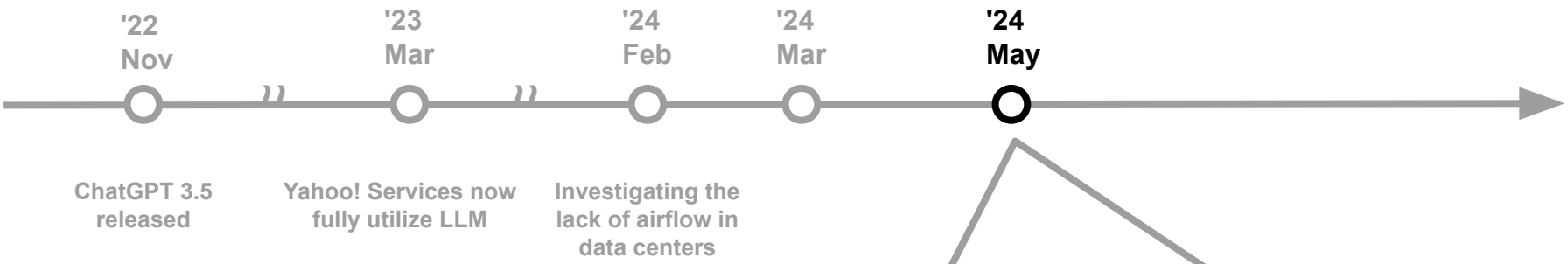
- In March 2024, NVIDIA will announce a new GPU server, the GB200 NVL72, at the GTC (GPU Technology Conference).
  - **Water-cooling technology** is essential for the GB200 NVL72
- Interview with a US company on why water cooling technology is necessary
  - The reason is that **high-bandwidth communication** between GPUs is required to shorten the model training time.

Today's announcement

Part 1 | **Part 2**

GB200 NVL72 unveiled at GTC

GB200 NVL72 power supply is found to be 480V



- GB200 NVL72 power supply is **480V input**
- This presentation explains the current situation.

Emergence of LLM and changes in server resources

We are here!

Part 1

Part 2

GB200 NVL72 unveiled at GTC

GB200 NVL72 power supply is found to be 480V

'22 Nov

'23 Mar

'24 Feb

'24 Mar

'24 May

ChatGPT 3.5 released

Yahoo! Services now fully utilize LLM

Investigating the lack of airflow in data centers

November 2022 ChatGPT 3.5 release

November 2022 ChatGPT 3.5 release

How is it different from previous AI?

Conventional "tabular data learning" and Differences in "Transformer Learning for LLM"

Features	Tabular format data	LLM (Large Scale Language Models)
Data Type	Tabular data (e.g. Excel, CSV)	Text data
Special skill	Numerical prediction, classification, data analysis	Text generation, translation, chatbots
Training Data	Structured data (numerical, categorical)	Large amount of document data
Model Scale	Relatively small scale (depends on features and number of data)	Large scale (billions of parameters)
Flexibility	Specialized for specific datasets	Capable of handling a wide variety of tasks

Conventional "tabular data learning" and Differences in "Transformer Learning for LLM"

Features	Tabular format data	LLM (Large Scale Language Models)
	CSV)	Text data
	sification, data	Text generation, translation, chatbots
	al)	**Large amount of document data**
Model Scale	Relatively small scale (depends on features and number of data)	**Large scale (billions of parameters)**
Flexibility	Specialized for specific datasets	Capable of handling a wide variety of tasks

It requires more data compared to the previous AI.

LLM also available for Yahoo! JAPAN services



- Yahoo! JAPAN uses LLM to **"understand language"** in user posts
- Significant **improvement in accuracy** achieved by using LLM

However, there are also major challenges

LLM also available for Yahoo! JAPAN services



- Yahoo! JAPAN uses LLM to **"understand language"** in user posts
- Significant **improvement in accuracy** achieved by using LLM

However, there are also major challenges

Changes in Server Workloads (with A100 GPU)

Tabular format data Workload 30%-80% for about a week

LLM Workload **80%-100% continue for 1-2 months**

Current status of data centers and predicted air conditioner air volume problems

We are here!

Part 1

Part 2

GB200 NVL72 unveiled at GTC

GB200 NVL72 power supply is found to be 480V

'22 Nov — ChatGPT 3.5 released

'23 Mar — **Yahoo! Services now fully utilize LLM**

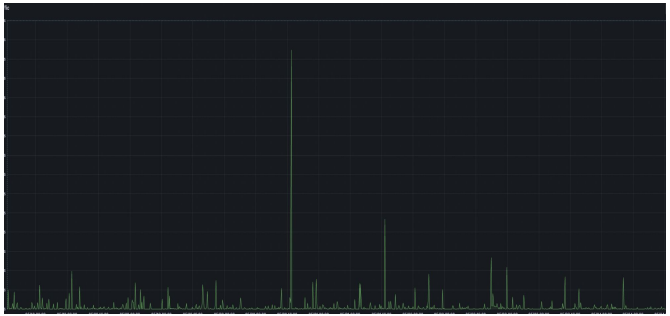
'24 Feb — Investigating the lack of airflow in data centers

'24 Mar

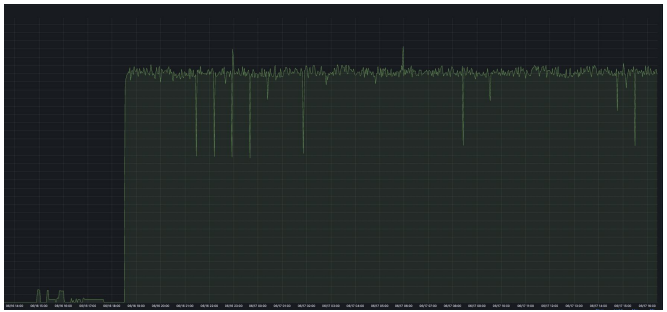
'24 May

Network: Comparison of server bandwidth between LLM and another system (Hadoop)

Hadoop server interface graph



LLM server interface graph

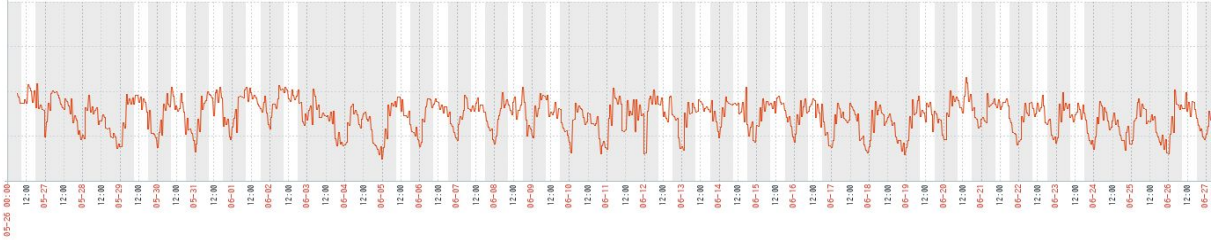


	Server port bandwidth capacity	Usage situation
Hadoop server interface	25Gbps	13.50 Gbps (**54%**)
LLM Server Interface	100 Gbps	23.5Gbps (**24%**)

- There is enough bandwidth and no errors or drops.
- **No network issues have occurred**
- * Some ingenuity when creating the model also played a role.

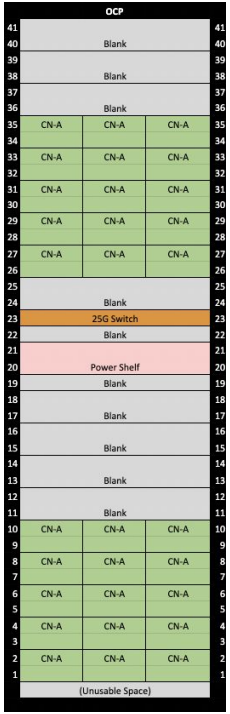
Power: Comparison of server power consumption between LLM and another system (Hadoop)

Hadoop Server (Maximum operating rate: 76.12%)

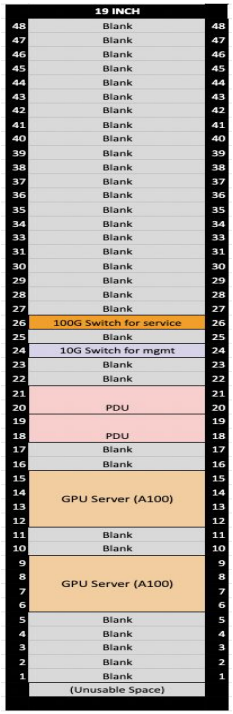


LLM Server (Maximum operating rate: **95.44%**)

Hadoop Rack
(30 nodes / rack)

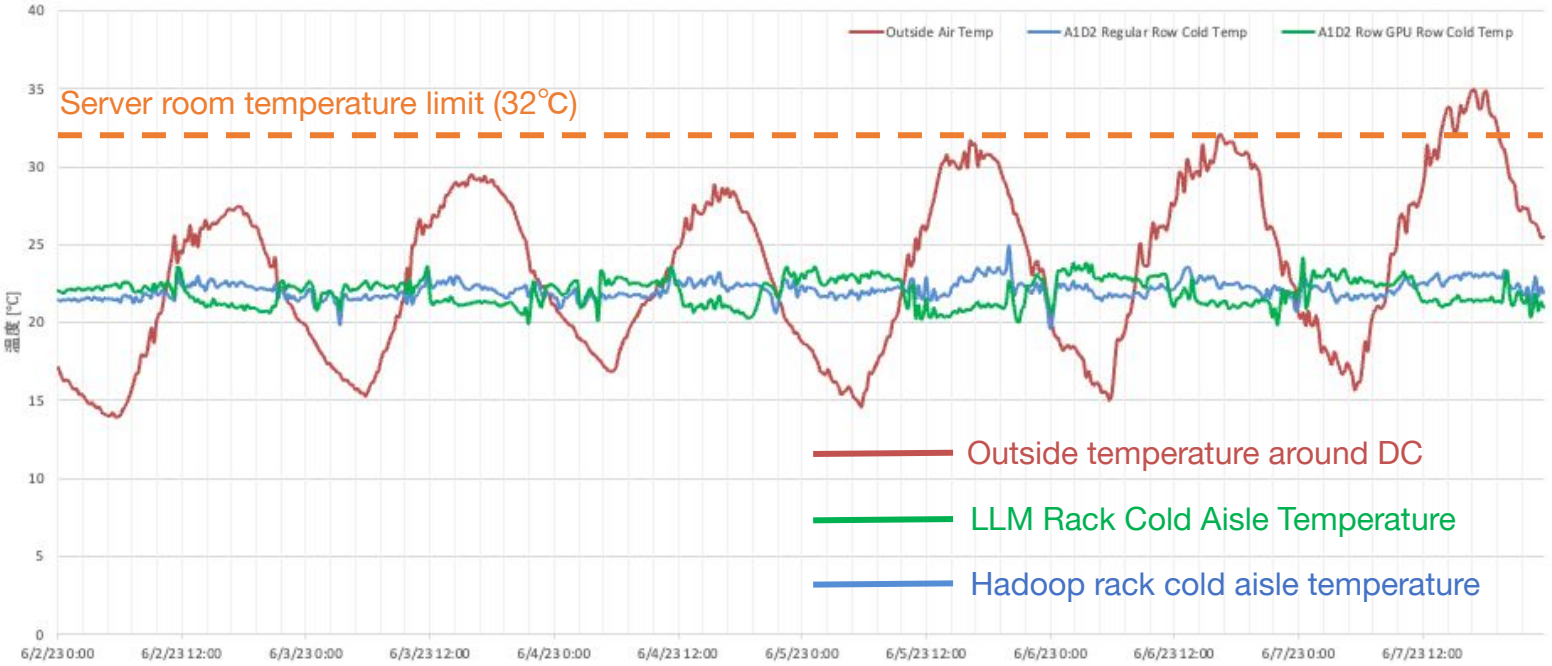


LLM GPU Rack
(2 nodes / rack)



Server workloads remain high for long periods, causing **increased power consumption**

Cooling: Temperature comparison between LLM and another system (Hadoop)



In terms of temperature, there is no significant difference between the LLM and the Hadoop rack rows.

There are no temperature issues

Currently **no issues with network, power and cooling**

Discussion with US company about water cooling technology

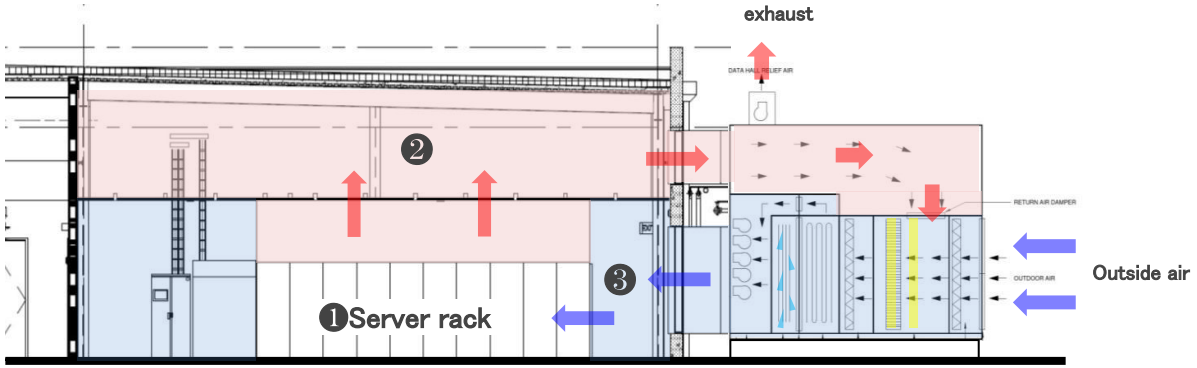
- GPU TDP (Thermal Design Power) and other factors are on the rise
- Increased TDP maximizes server fan speed
- There is a concern that the increased airflow from servers will lead to an increase in the airflow from air conditioners.

Currently no issues with network, power or cooling

Discussion with US company about water cooling technology

- GPU TDP (Thermal Design Power) and other factors are on the rise
- Increased TDP maximizes server **fan speed**
- There is a concern that the **increased airflow** from servers will lead to an increase in the airflow from air conditioners.

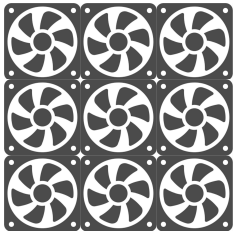
Air conditioning system and cross-section of Actapio data center



- Increased airflow in server rack in ❶
- An increase in pressure in the hot aisle in ❷ and a decrease in pressure in the cold aisle in ❸ occured
- Sufficient airflow from the air conditioner is required to exhaust and discharge the air to areas ❷ and ❸.
- If the air conditioner does not have enough airflow, the pressure balance inside the data center cannot be maintained and temperature control becomes impossible.

For data center air conditioners to operate without problems, airflow **of the air conditioner** must be more than the server airflow

Server airflow

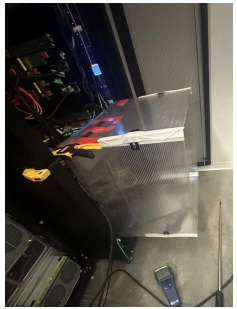


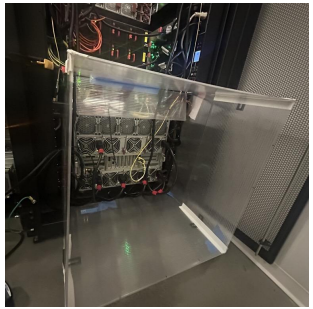
Air conditioner air volume

Measuring the airflow of an air-cooled GPU server

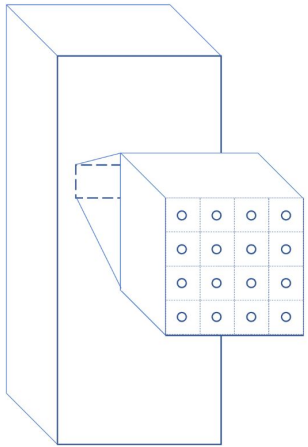
- Servers are measured at full load
- Surround the rear of the server with a plastic wind shield
- The cross section was divided into 4x4 sections with string, and each section was measured with an anemometer

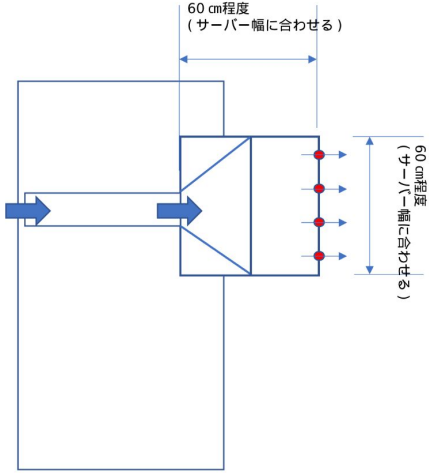
Actual situation





Measuring method

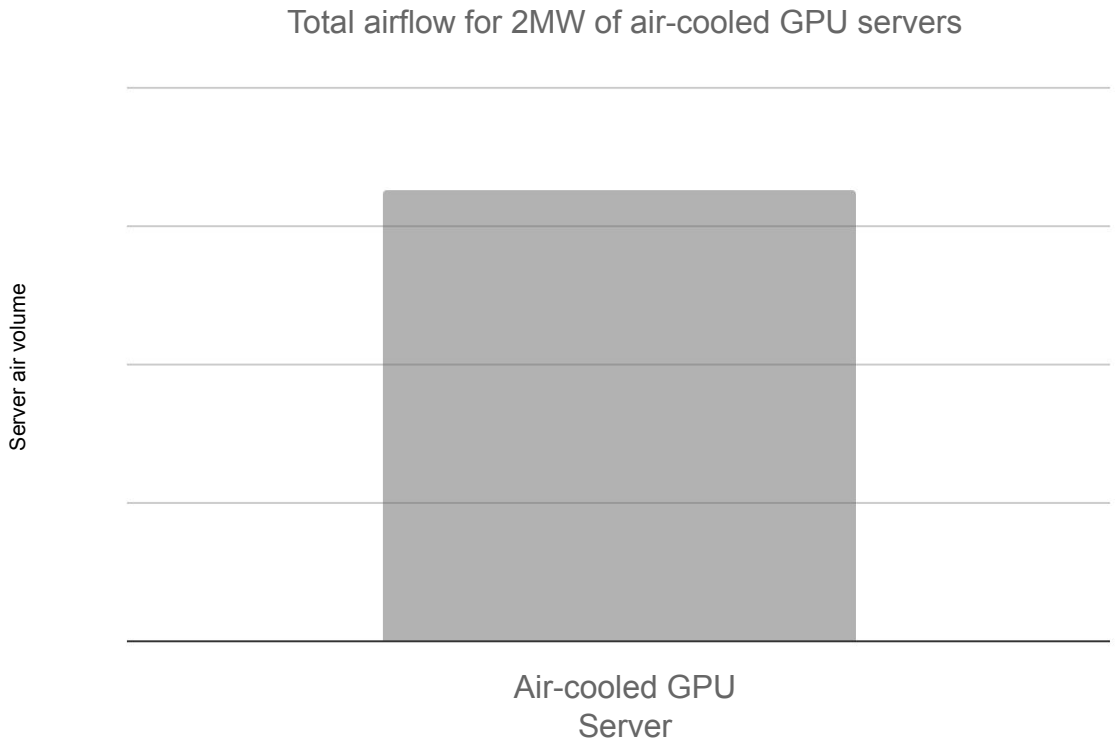




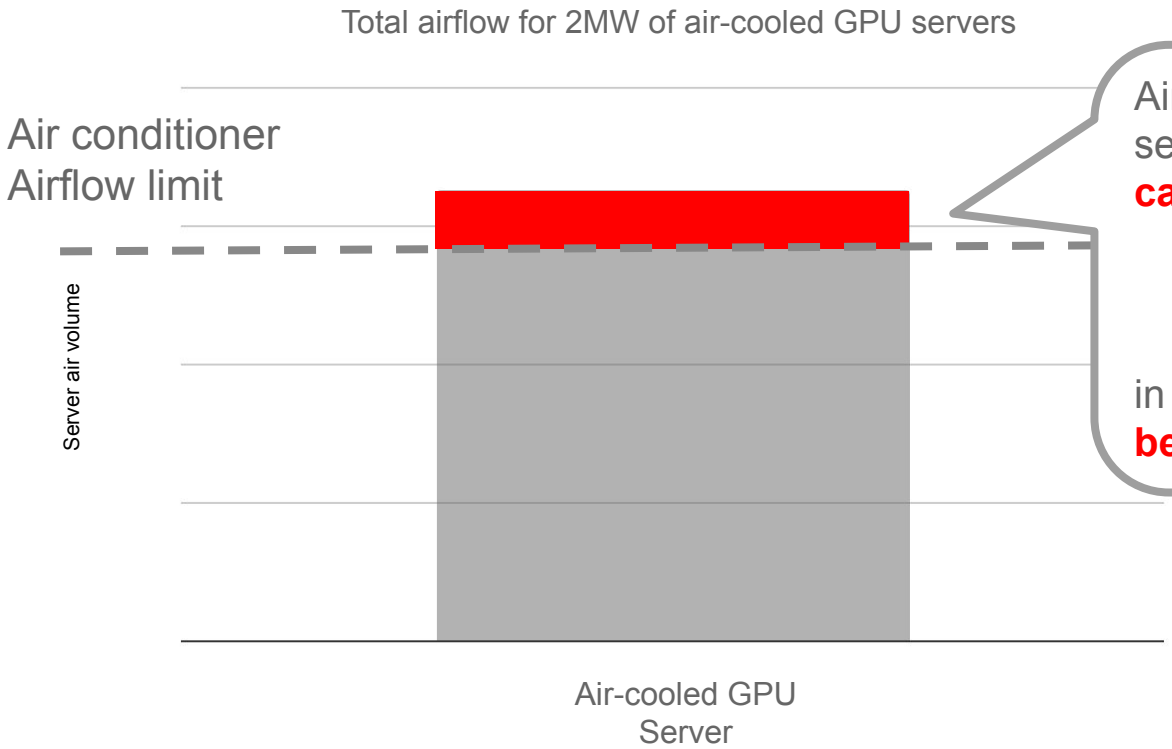
Anemometer

* Measurement method: JIS A 1431 (https://kikakurui.com/a1/A1431-1994-01.html)

Actapio's 2MW server room
Scenario for fully installed air-cooled GPU servers for LLM



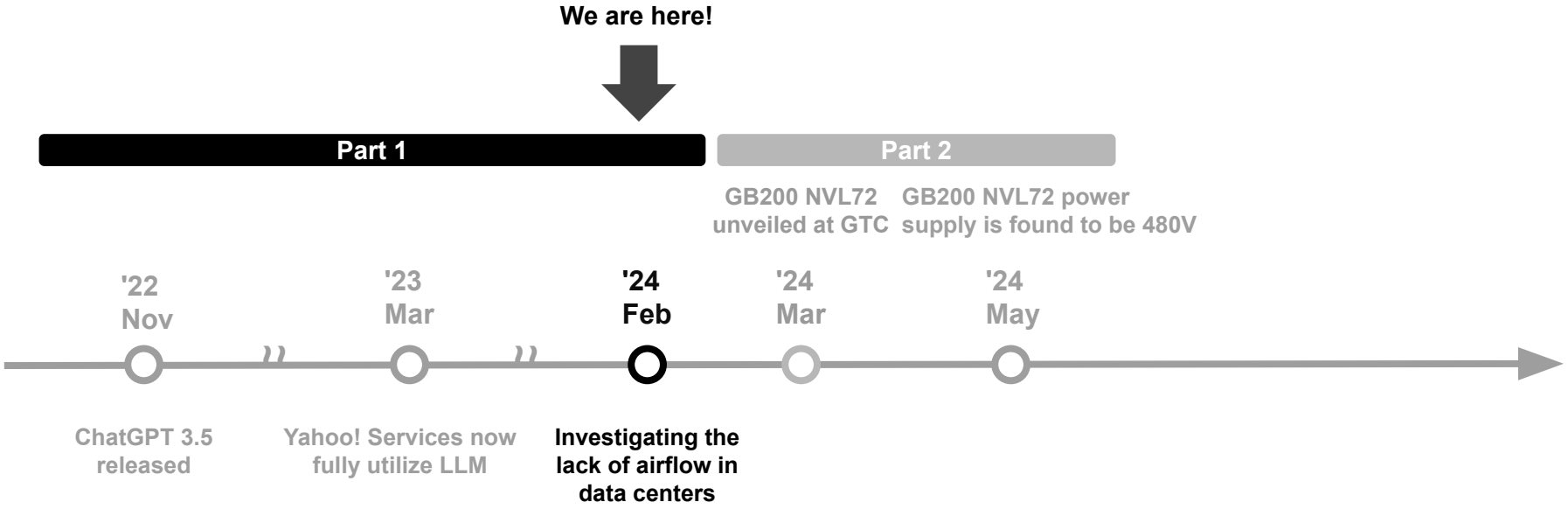
Actapio's 2MW server room
Scenario for fully installed air-cooled GPU servers for LLM



Airflow for air-cooled GPU servers **exceeds air conditioner capacity**

in the server room **can no longer be controlled.**

Considering solutions to insufficient airflow from air conditioners



Solutions to insufficient airflow from air conditioners

Option 1: Limiting the number of installed servers

Option 2: Increase air conditioning capacity

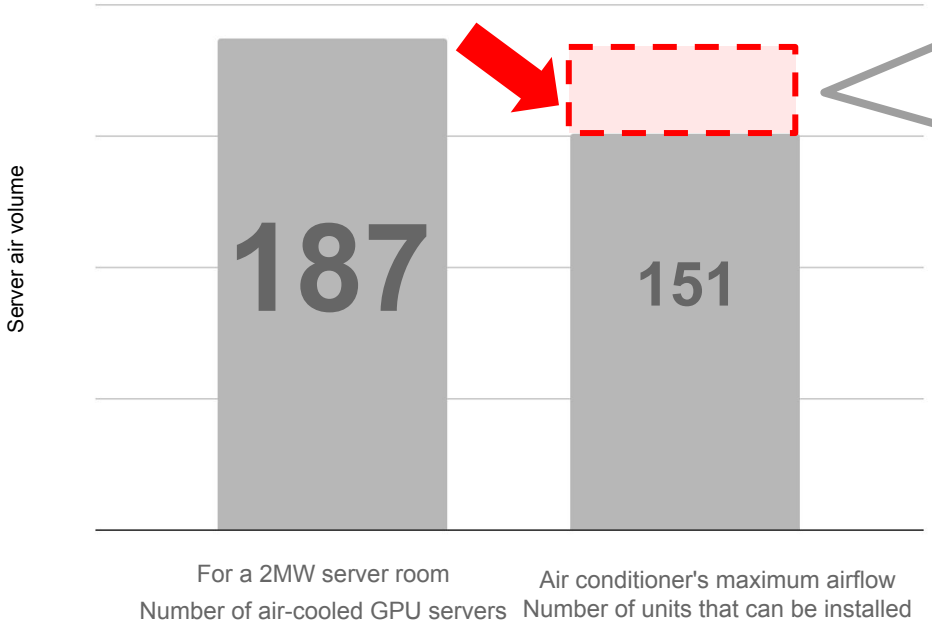
a. Addition of air conditioners
b. Increasing air conditioner specifications

Option 3: Consideration of water cooling method: DLC (Direct Liquid Cooling)

a. L to L: Liquid to Liquid
b. AALC: Air Assisted Liquid Cooling

Option 1: Limiting the number of installed servers

-36 units



Reducing the number of installed servers reduces the amount of airflow

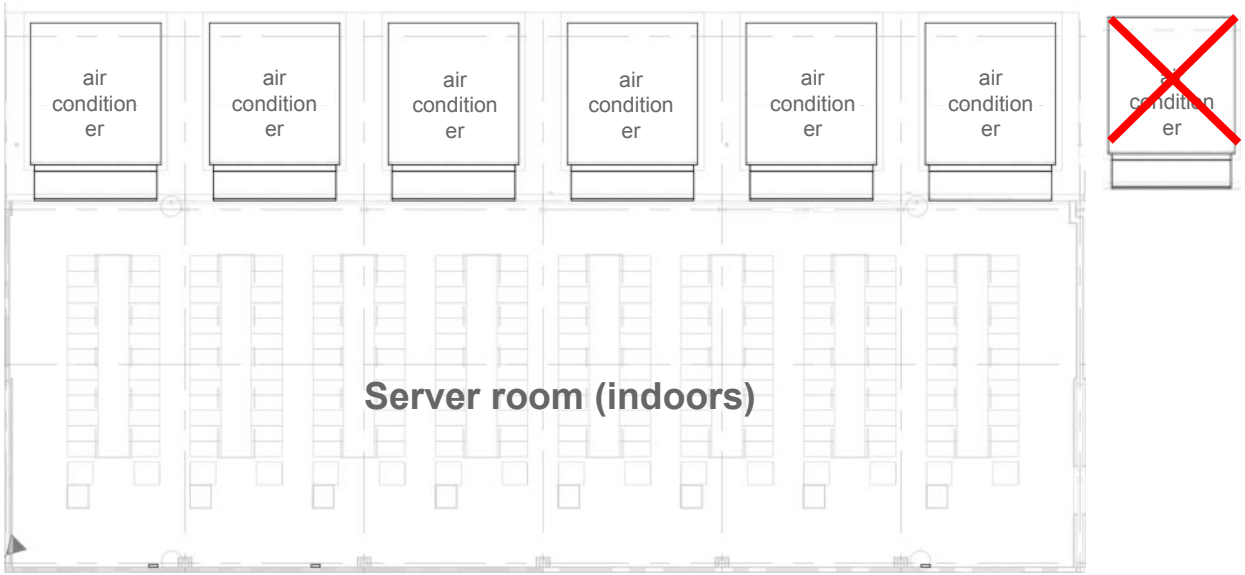
Need to **reduce 36 GPU servers**

Option 2: Increase air conditioning capacity

a.Additional **air conditioners**

The layout of the data hall (see below) requires additional physical **capacity.**

Server room top view

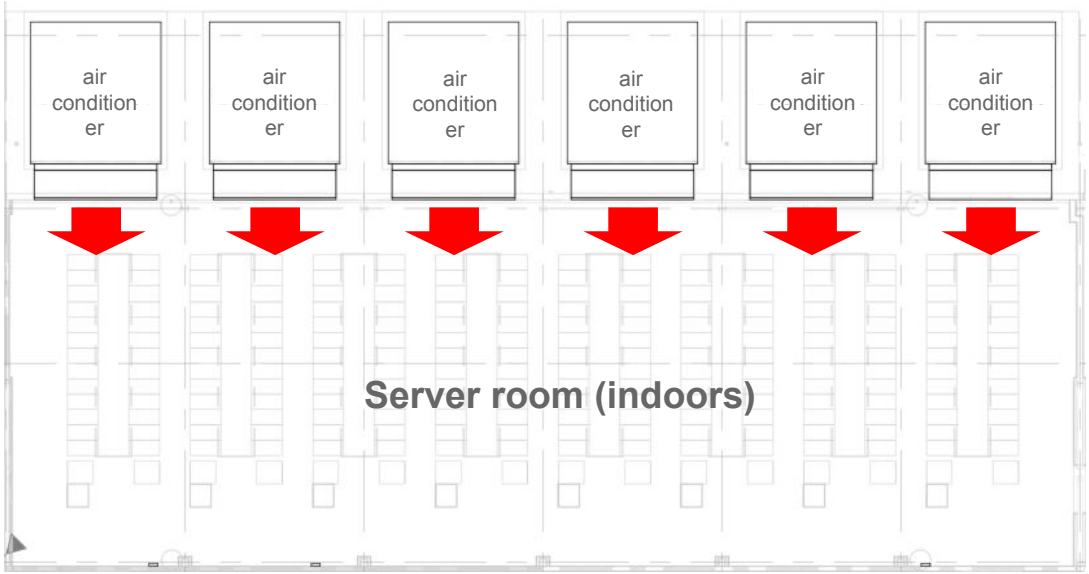


Option 2: Increase air conditioning capacity

b. **Increasing the specifications of air conditioners**

Upgrading the specifications of existing air conditioners is costly and **difficult , including the construction required**

Server room top view



Option 3: Consider water cooling method

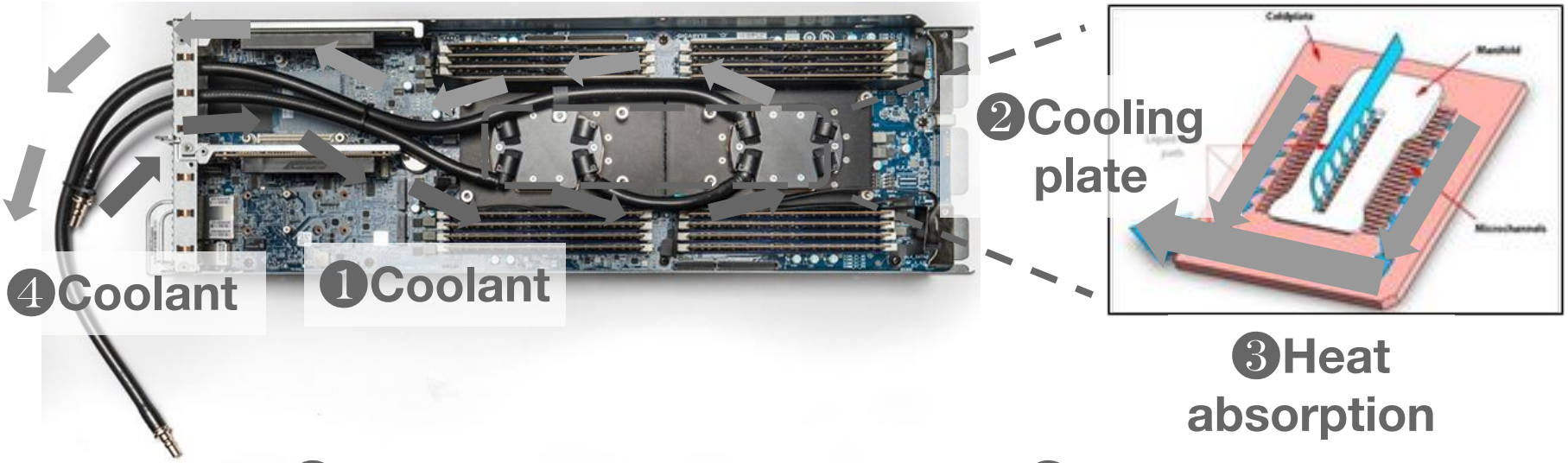
Water cooling methods (DLC) can be classified into two types

DLC (Direct Liquid Cooling)

1. L to L: Liquid to Liquid
2. **AALC: Air Assisted Liquid Cooling**

Among the water cooling methods, we are considering AALC and we will explain why

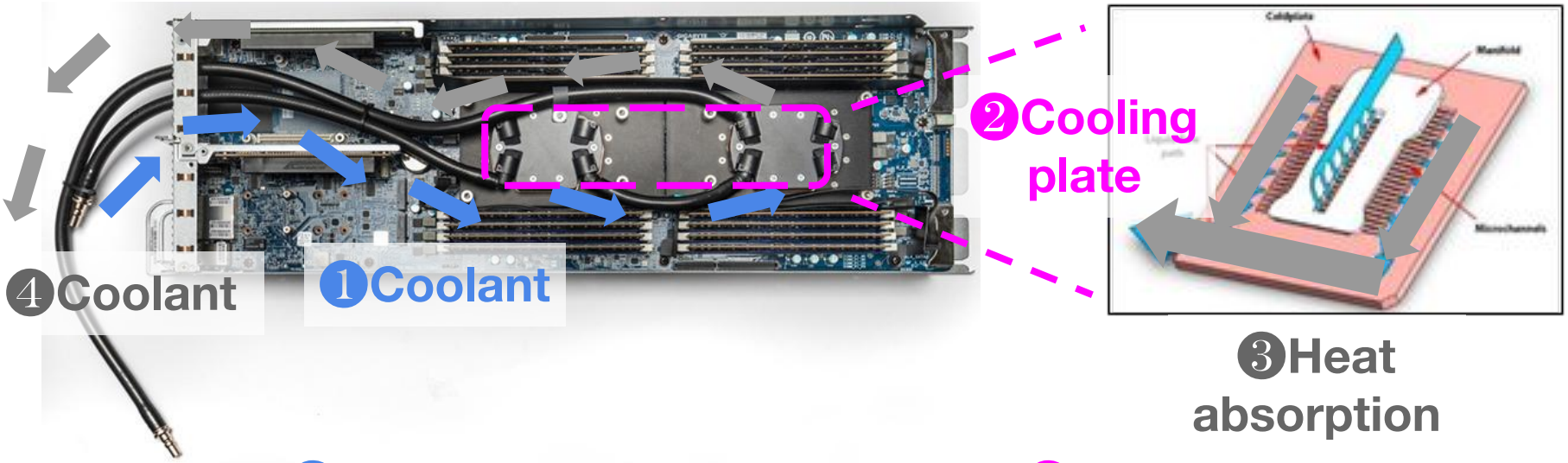
About DLC (Direct Liquid Cooling)



- ❶**The cold coolant** is sent to the ❷**Cooling plate**
- ❷**Cooling plate** ❸**absorbs heat** from the processor
- ❹**The warm coolant** is sent to the heat exchanger

Source: https://www.gigabyte.com/Solutions/coolit-systems-dlc
Dell Technologies Direct Liquid Cooling Support for New PowerEdge Servers | Dell Technologies Info Hub

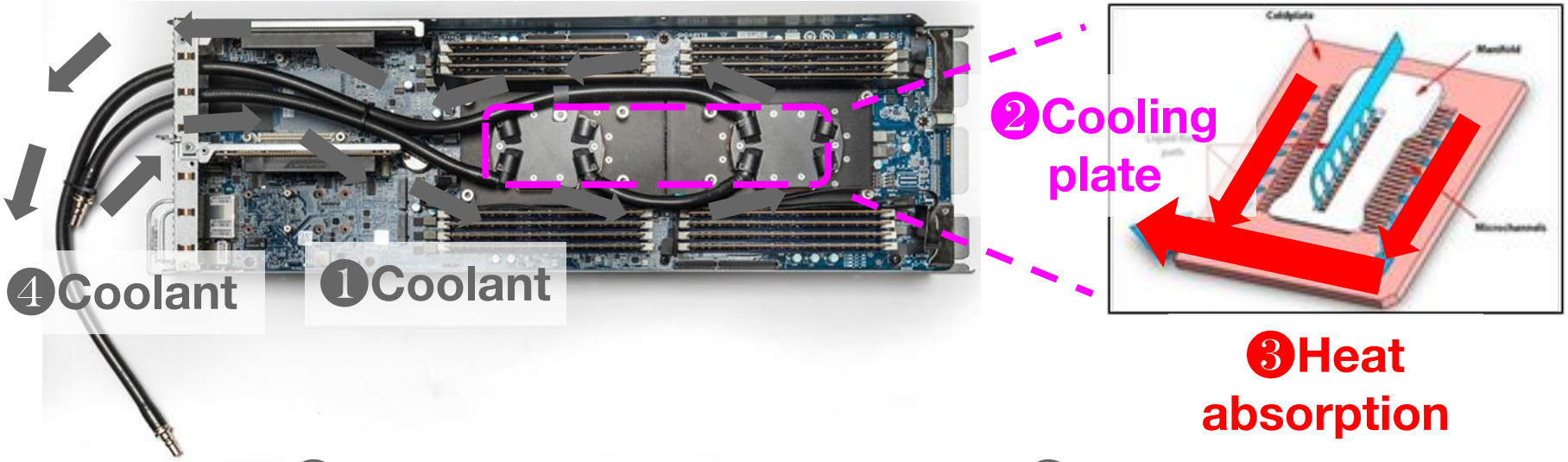
About DLC (Direct Liquid Cooling)



- ❶**The cold coolant** is sent to the ❷**Cooling plate**
- ❷**Cooling plate** ❸**absorbs heat** from the processor
- ❹**The warm coolant** is sent to the heat exchanger

Source: https://www.gigabyte.com/Solutions/coolit-systems-dlc
Dell Technologies Direct Liquid Cooling Support for New PowerEdge Servers | Dell Technologies Info Hub

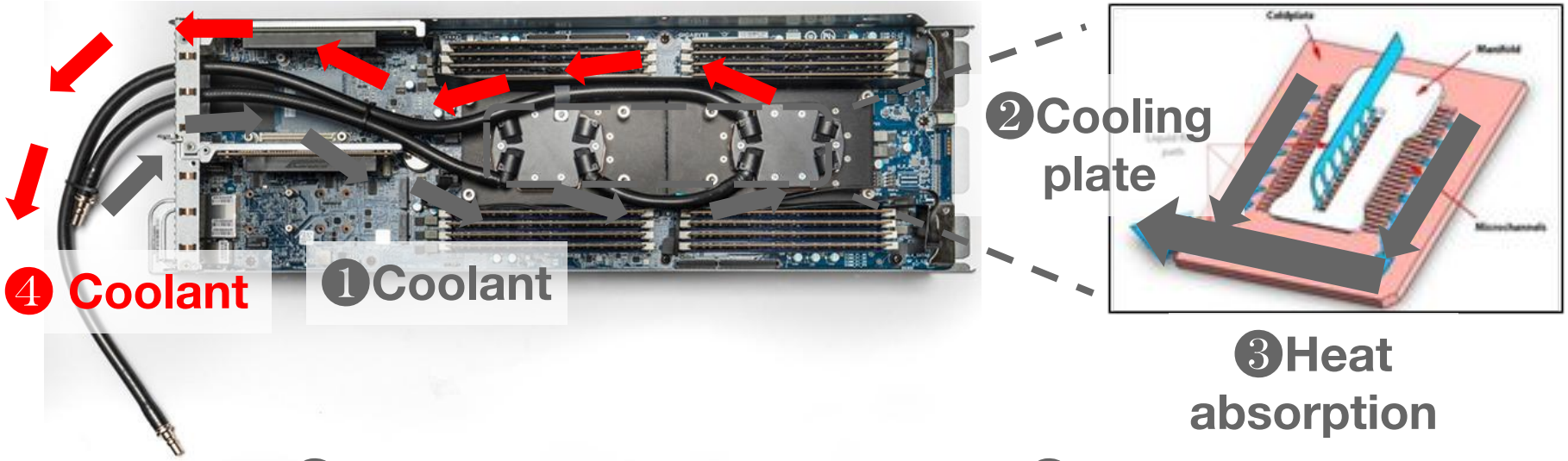
About DLC (Direct Liquid Cooling)



- ❶**The cold coolant** is sent to the ❷**Cooling plate**
- ❷**Cooling plate** ❸**absorbs heat** from the processor
- ❹**The warm coolant** is sent to the heat exchanger

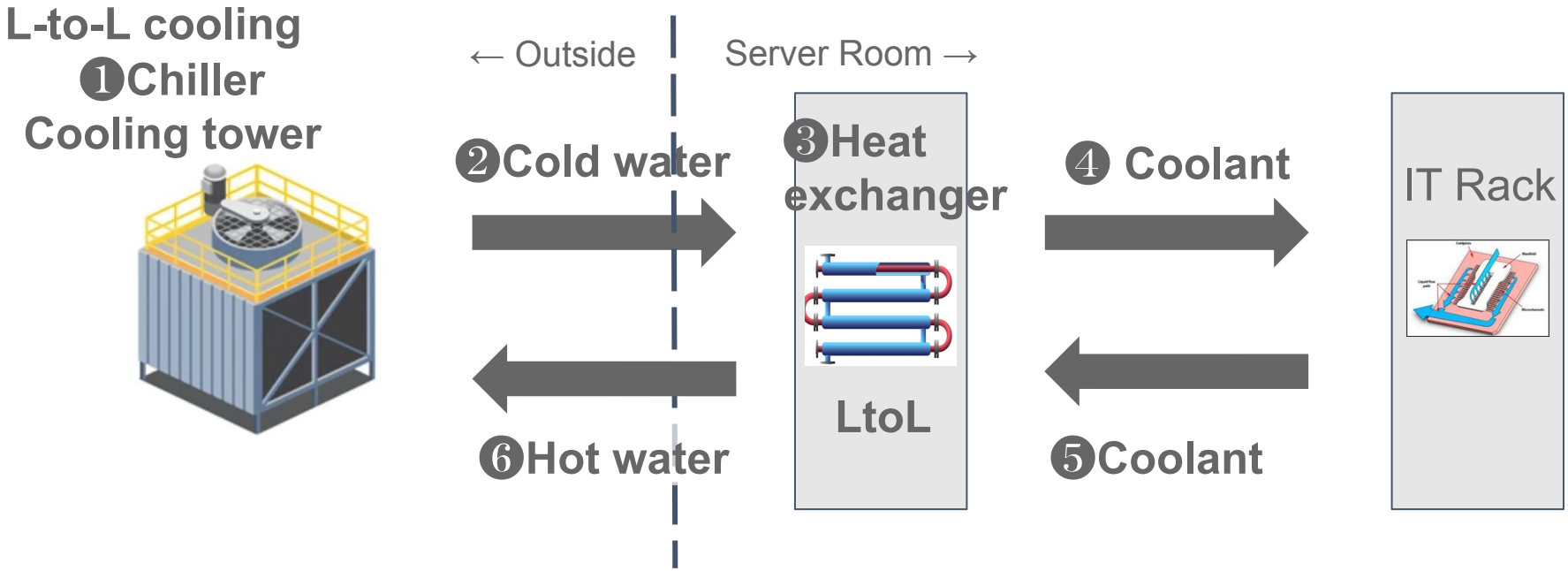
Source: https://www.gigabyte.com/Solutions/coolit-systems-dlc
Dell Technologies Direct Liquid Cooling Support for New PowerEdge Servers | Dell Technologies Info Hub

About DLC (Direct Liquid Cooling)



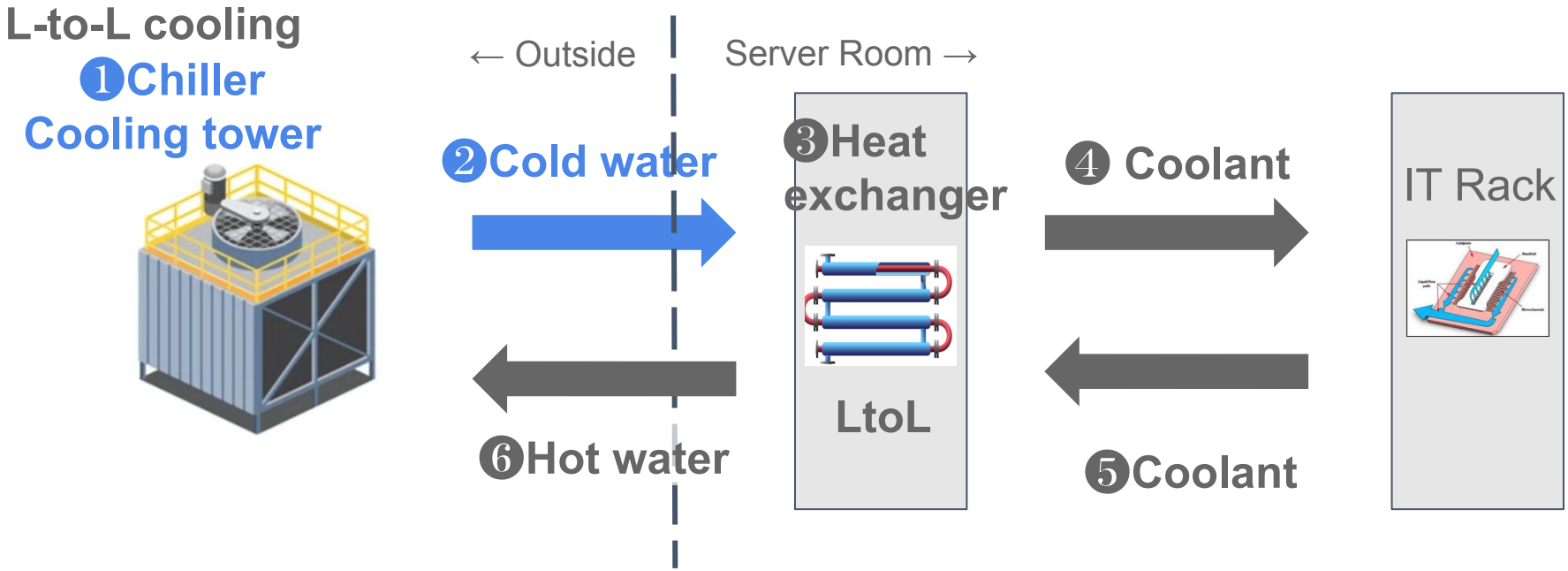
- ❶**The cold coolant** is sent to the ❷**Cooling plate**
- ❷**Cooling plate** ❸**absorbs heat** from the processor
- ❹**The warm coolant** is sent to the heat exchanger

Source: https://www.gigabyte.com/Solutions/coolit-systems-dlc
Dell Technologies Direct Liquid Cooling Support for New PowerEdge Servers | Dell Technologies Info Hub



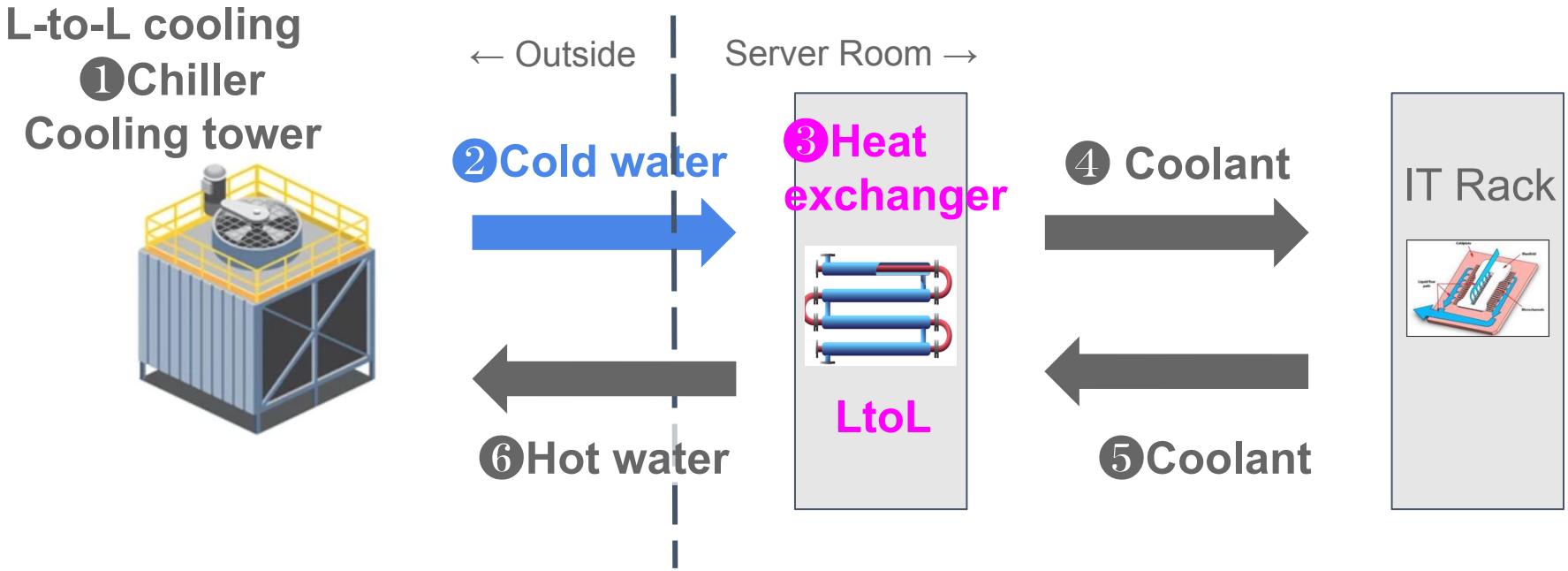
- ❶Chillers and cooling towers are used to make ❷chilled water
- ❷The cold water exchanges heat with the cooling liquid in the server room at ❸LtoL
- ❹ The cold liquid is sent to the server's cooling plate
- ❺The warm coolant is returned to the ❹CDU, where it exchanges heat with the cold water.
- ❻The hot water is returned to ❶the chiller or cooling tower to produce ❷cold water again.

Source: Dell Technologies Direct Liquid Cooling Support for New PowerEdge Servers | Dell Technologies Info



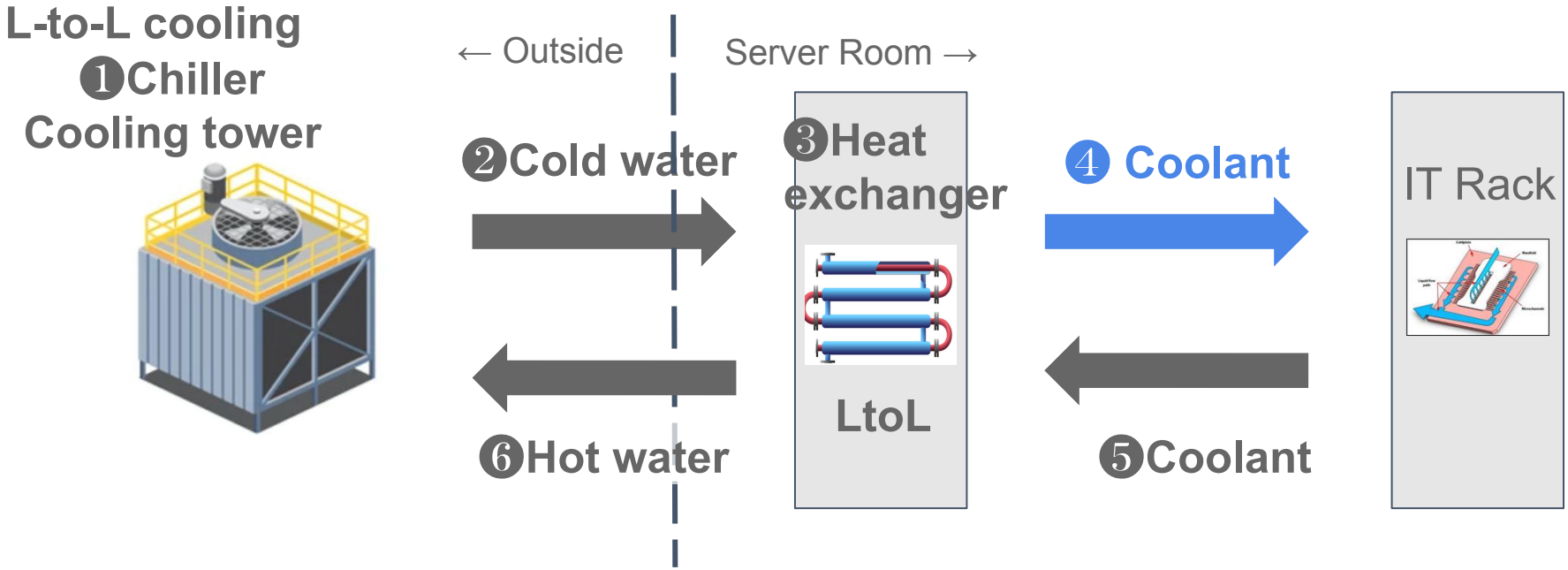
- ❶Chillers and cooling towers are used to make
- ❷The cold water exchanges heat with the cooling liquid in the server room at ❸LtoL
- ❹ The cold liquid is sent to the server's cooling plate
- ❺The warm coolant is returned to the ❹CDU, where it exchanges heat with the cold water.
- ❻The hot water is returned to ❶the chiller or cooling tower to produce ❷cold water again.

Source: Dell Technologies Direct Liquid Cooling Support for New PowerEdge Servers | Dell Technologies Info



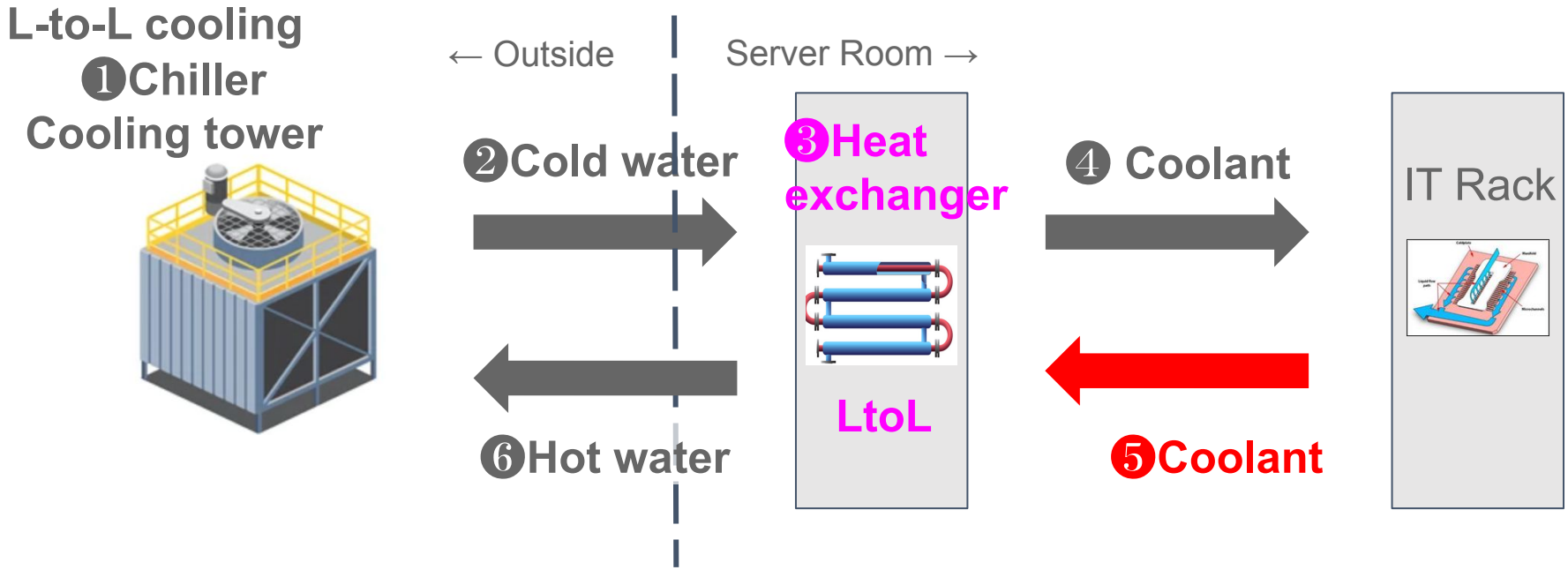
- ❶**Chillers and cooling towers are used** to make
- ❷**The cold water** exchanges heat with the cooling liquid
- ❹ **The cold liquid** is sent to the server's cooling plate
- ❺**The warm coolant** is returned to **the** ❹**CDU** , where it exchanges heat with the cold water.
- ❻**The hot water** is returned to ❶**the chiller or cooling tower** to produce

Source: Dell Technologies Direct Liquid Cooling Support for New PowerEdge Servers | Dell Technologies Info



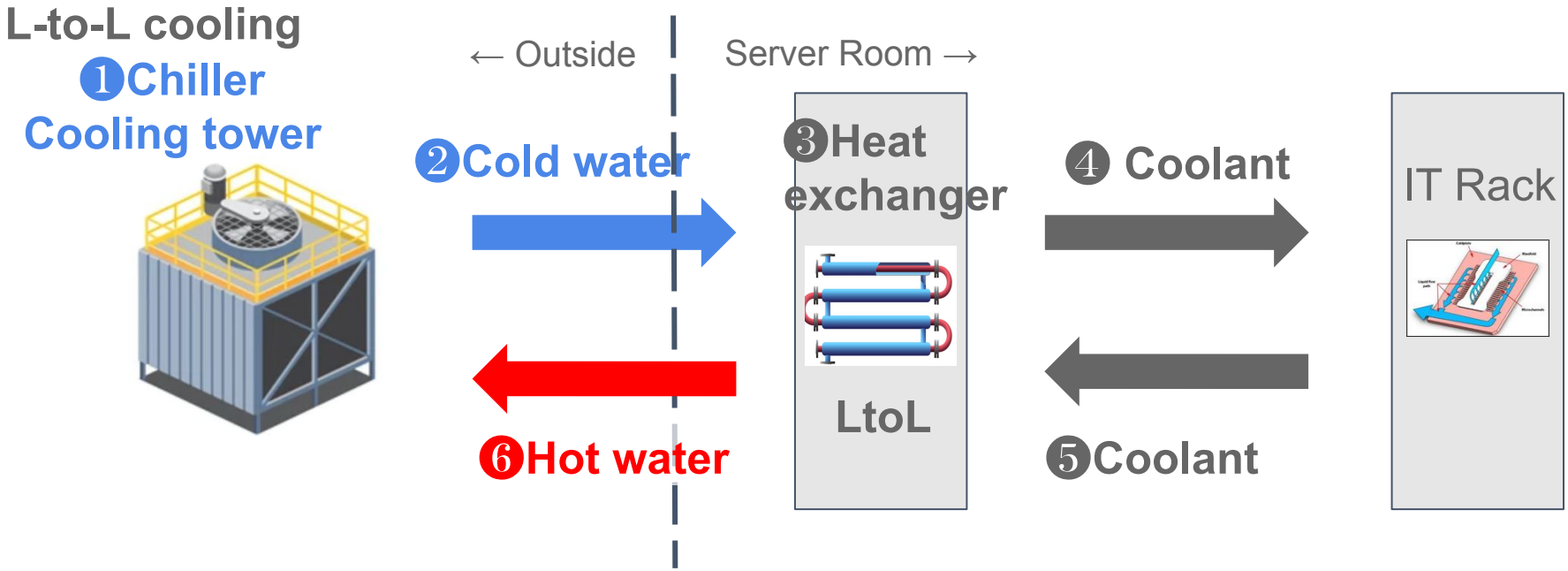
- ❶**Chillers and cooling towers are used** to make
- ❷**The cold water** exchanges heat with the cooling liquid
- ❹ **The cold liquid** is sent to the server's cooling plate
- ❺**The warm coolant** is returned to ❸**LtoL** and exchanges heat with the cold water.
- ❻**The hot water** is returned to ❶**the chiller or cooling tower** to produce

Source: Dell Technologies Direct Liquid Cooling Support for New PowerEdge Servers | Dell Technologies Info



- ❶**Chillers and cooling towers are used** to make
- ❷**The cold water** exchanges heat with the cooling liquid
- ❹ **The cold liquid** is sent to the server's cooling plate
- ❺**The warm cooling liquid** is returned to ❸**LtoL** and exchanges heat with the cold water.
- ❻**The hot water** is returned to ❶**the chiller or cooling tower** to produce

Source: Dell Technologies Direct Liquid Cooling Support for New PowerEdge Servers | Dell Technologies Info



- ❶**Chillers and cooling towers are used** to make
- ❷**The cold water** exchanges heat with the cooling liquid
- ❹ **The cold liquid** is sent to the server's cooling plate
- ❺**The warm cooling liquid** is returned to ❸**LtoL** and exchanges heat with the cold water.
- ❻**The hot water** is returned to ❶**the chiller or cooling tower** to produce

Source: Dell Technologies Direct Liquid Cooling Support for New PowerEdge Servers | Dell Technologies Info

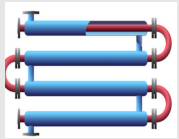
L-to-L cooling

← Outside | Server Room →

❶Chiller
Cooling tower

❷Cold water

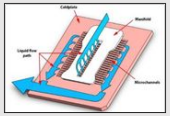
❸Heat exchanger



LtoL

❹ Coolant

IT Rack



❺Coolant

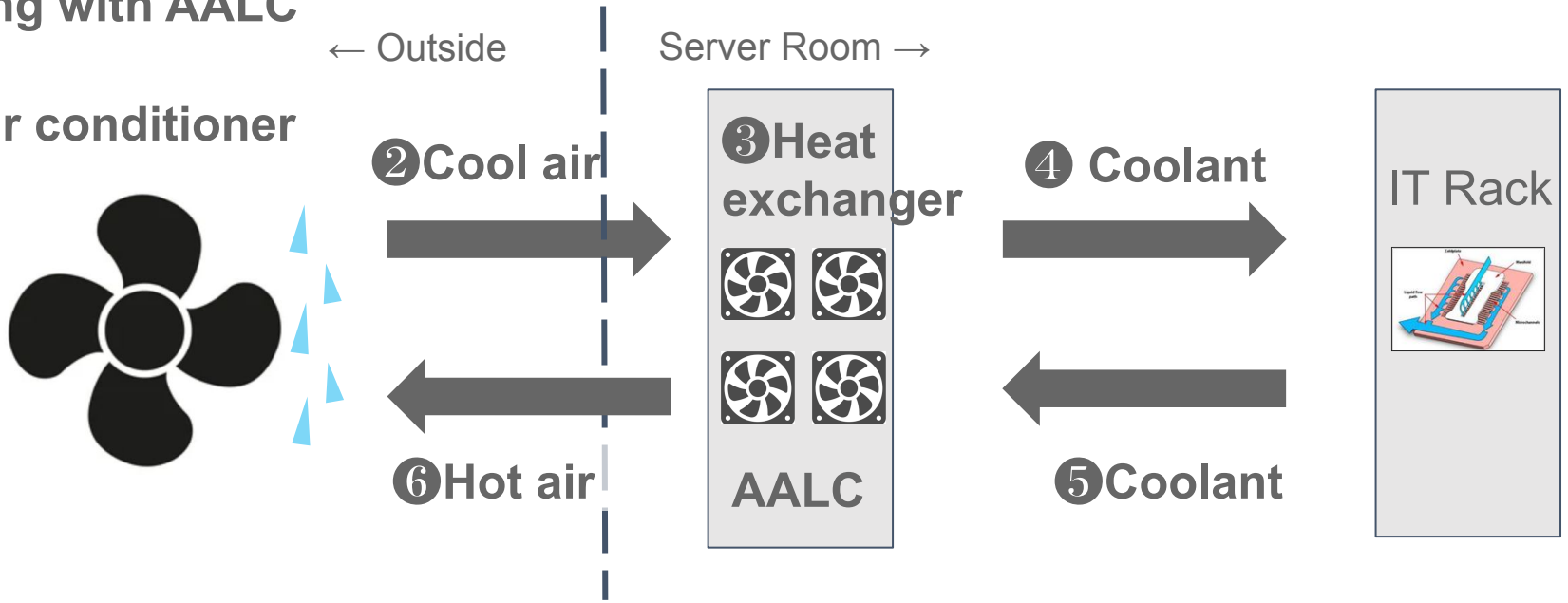
❻Hot water

air conditioner

In addition to the existing air conditioners, chilled water equipment and water pipes to the server room are required, which **increases costs and delivery times.**

Source: Dell Technologies Direct Liquid Cooling Support for New PowerEdge Servers | Dell Technologies Info

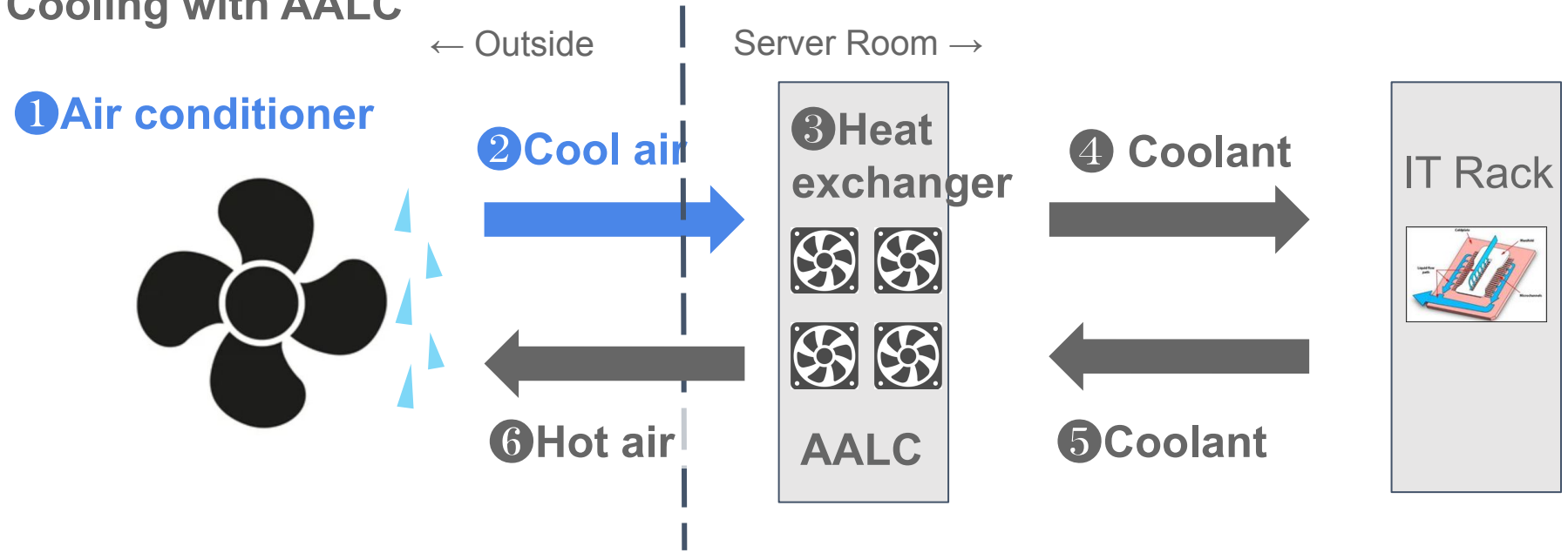
Cooling with AALC



- Air **conditioner** ❷ creates cool air
- ❷**The cold air** exchanges heat with the cooling liquid in
- ❹ **The cold liquid** is sent to the server's cooling plate
- ❺**The warm cooling liquid** is returned to ❸**AALC** and exchanges heat with the cold air.
- ❻**The warm air** is returned to ❶**the air conditioner , where it again** creates

Source: Dell Technologies Direct Liquid Cooling Support for New PowerEdge Servers | Dell Technologies Info

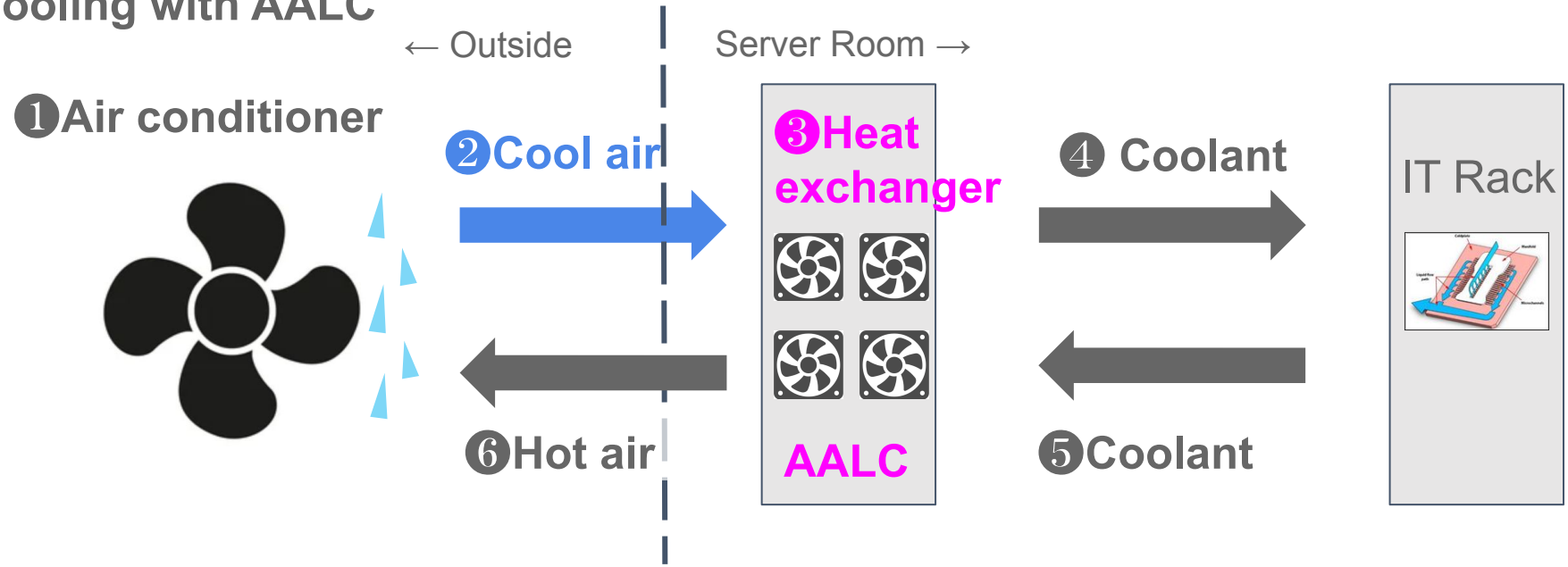
Cooling with AALC



- Air **conditioner** ❷ creates cool air
- ❷**The cold air** exchanges heat with the cooling liquid in
- ❹ **The cold liquid** is sent to the server's cooling plate
- ❺**The warm cooling liquid** is returned to ❸**AALC** and exchanges heat with the cold air.
- ❻**The warm air** is returned to ❶**the air conditioner , where it again** creates

Source: Dell Technologies Direct Liquid Cooling Support for New PowerEdge Servers | Dell Technologies Info

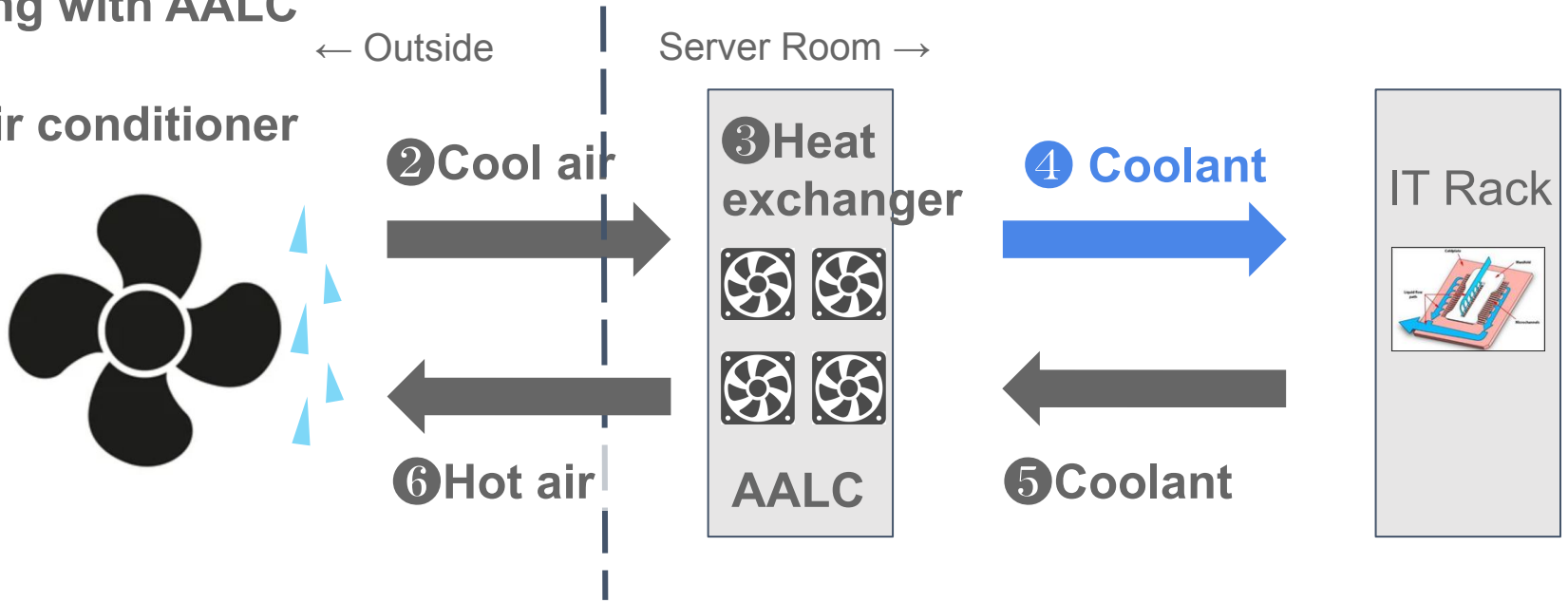
Cooling with AALC



- Air **conditioner** ❷ creates cool air
- ❷**The cold air** exchanges heat with the cooling liquid in
- ❹ **The cold liquid** is sent to the server's cooling plate
- ❺**The warm cooling liquid** is returned to ❹**AALC** and exchanges heat with the cold air.
- ❻**The warm air** is returned to ❶**the air conditioner , where it again** creates

Source: Dell Technologies Direct Liquid Cooling Support for New PowerEdge Servers | Dell Technologies Info

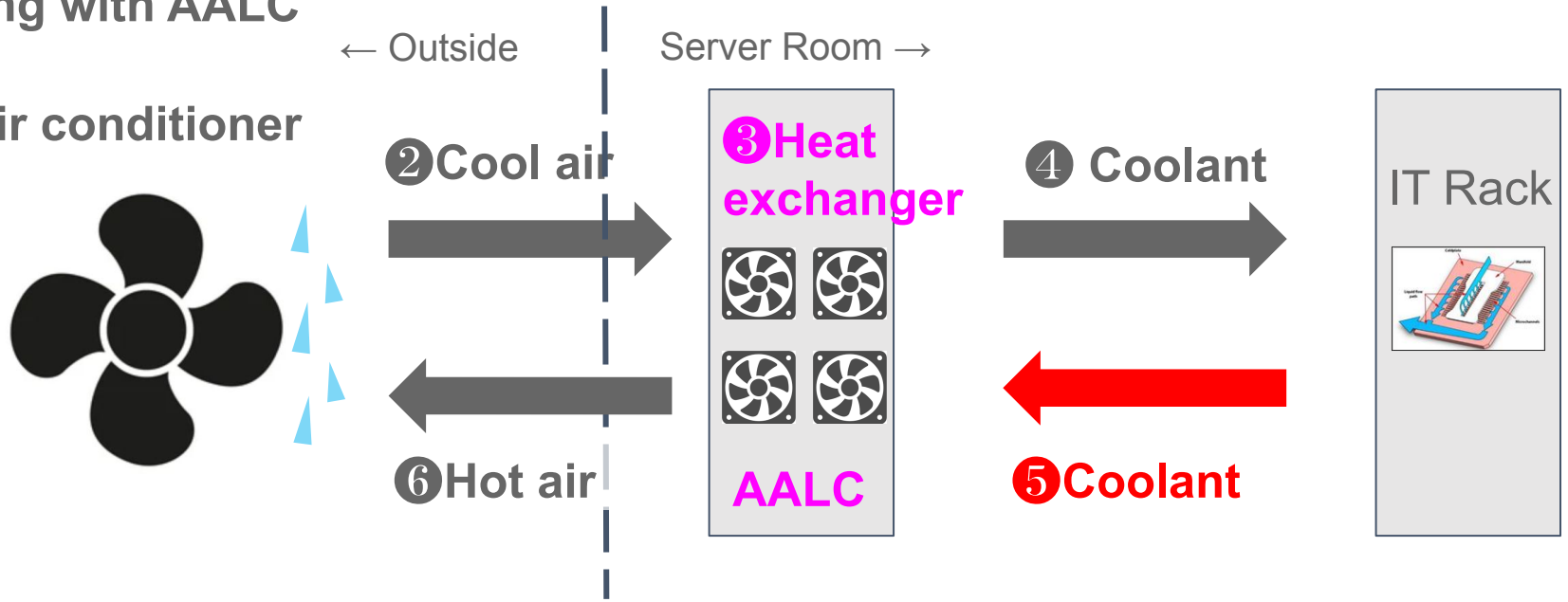
Cooling with AALC



- Air **conditioner** ❷ creates cool air
- ❷**The cold air** exchanges heat with the cooling liquid in
- ❹ **The cold liquid** is sent to the server's cooling plate
- ❺**The warm cooling liquid** is returned to ❹**AALC** and exchanges heat with the cold air.
- ❻**The warm air** is returned to ❶**the air conditioner , where it again** creates

Source: Dell Technologies Direct Liquid Cooling Support for New PowerEdge Servers | Dell Technologies Info

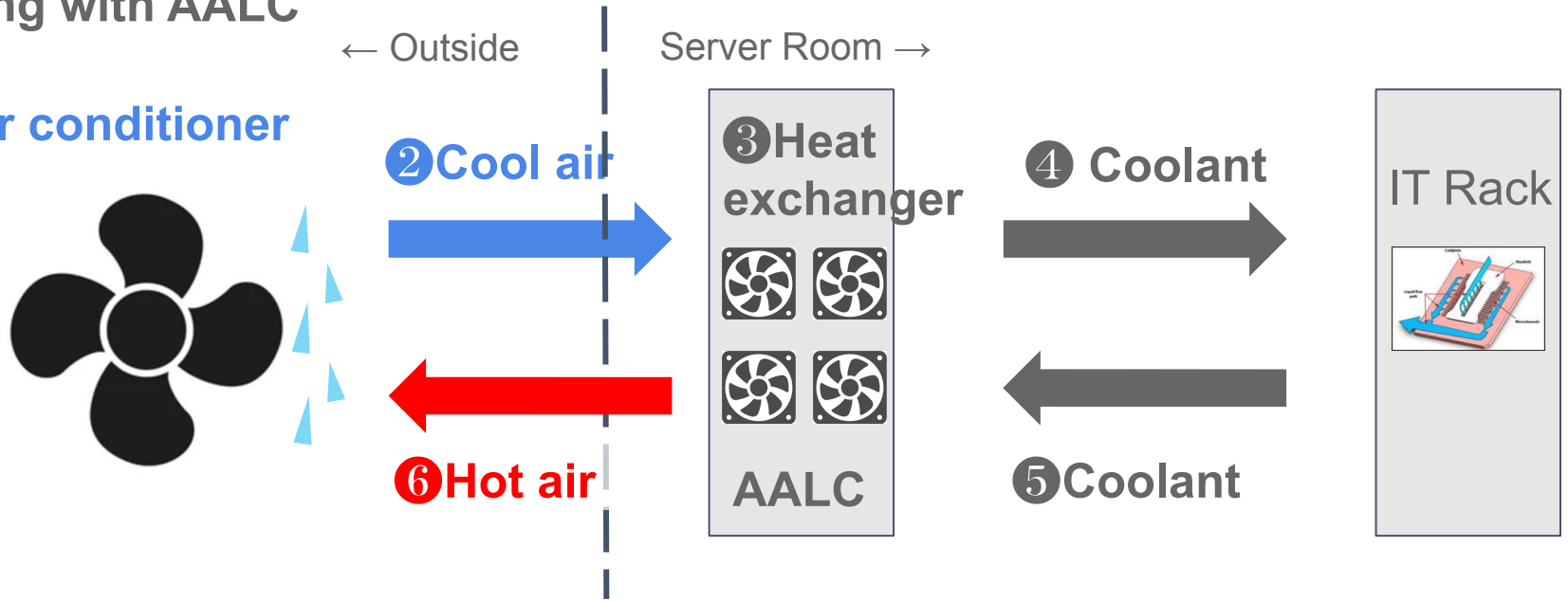
Cooling with AALC



- Air **conditioner** ❷ creates cool air
- ❷**The cold air** exchanges heat with the cooling liquid in
- ❹ **The cold liquid** is sent to the server's cooling plate
- ❺**The warm cooling liquid** is returned to ❹**AALC** and exchanges heat with the cold air.
- ❻**The warm air** is returned to ❶**the air conditioner , where it again** creates

Source: Dell Technologies Direct Liquid Cooling Support for New PowerEdge Servers | Dell Technologies Info

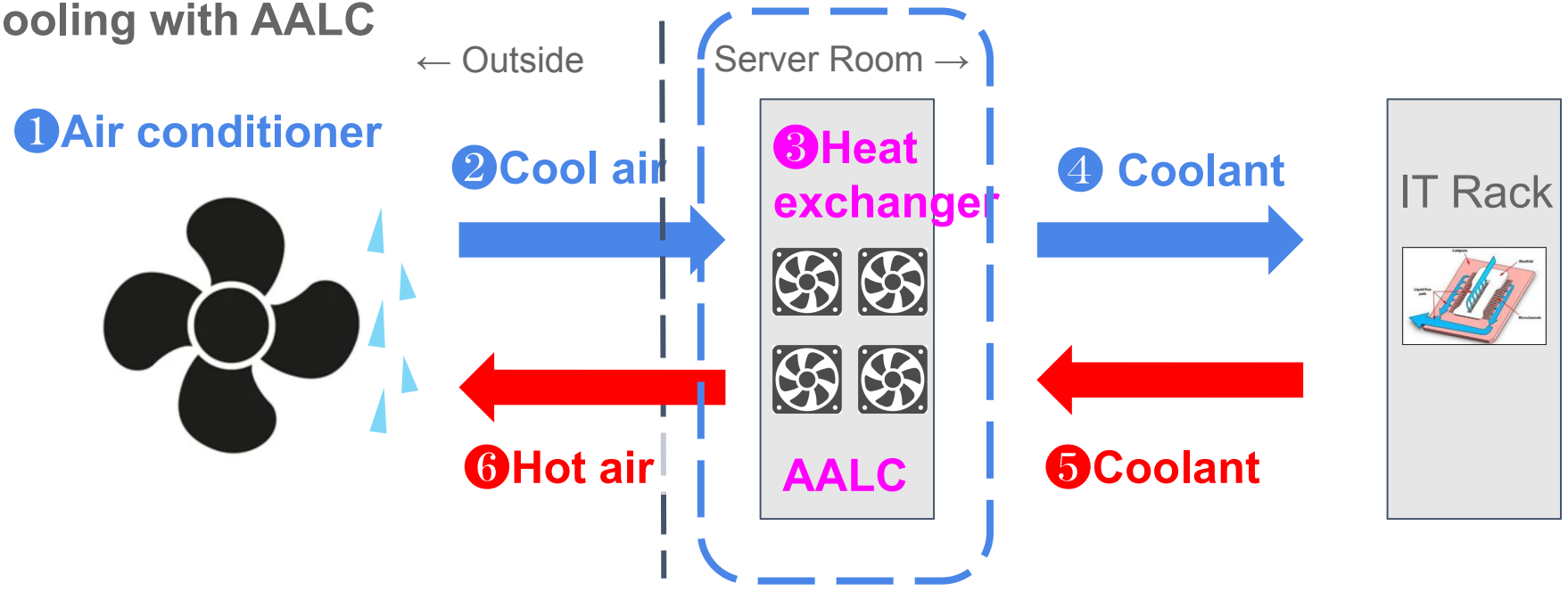
Cooling with AALC



- Air **conditioner** ❷ creates cool air
- ❷**The cold air** exchanges heat with the cooling liquid in
- ❹ **The cold liquid** is sent to the server's cooling plate
- ❺**The warm cooling liquid** is returned to ❹**AALC** and exchanges heat with the cold air.
- ❻**The warm air** is returned to ❶**the air conditioner , where it again** creates

Source: Dell Technologies Direct Liquid Cooling Support for New PowerEdge Servers | Dell Technologies Info

Cooling with AALC



Since it is only installed in the server room, it has a lower cost and shorter delivery time

Source: Dell Technologies Direct Liquid Cooling Support for New PowerEdge Servers | Dell Technologies Info Hub

We tested the reduction in airflow in server rooms with AALC.

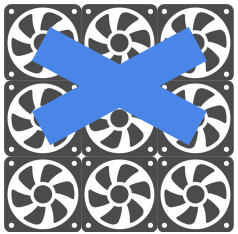
By using AALC, the amount of air flow in the server room was reduced.

We tested the reduction in airflow in server rooms with AALC.

By using AALC, the amount of air flow in the server room was reduced.

the server airflow and AALC airflow **are lower than the air conditioner airflow.**

Server air volume



+

AALC Air Volume

Rack type AALC

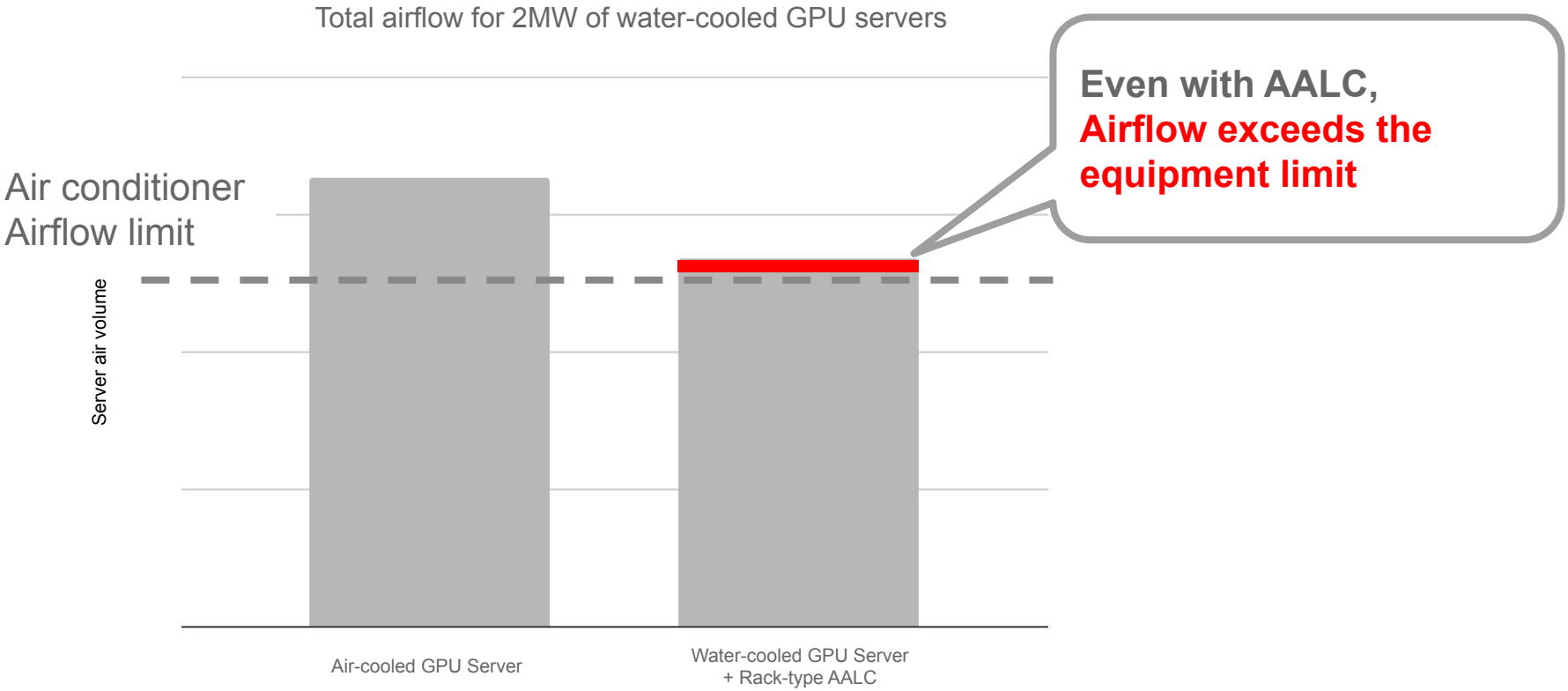
<

Air conditioner air volume

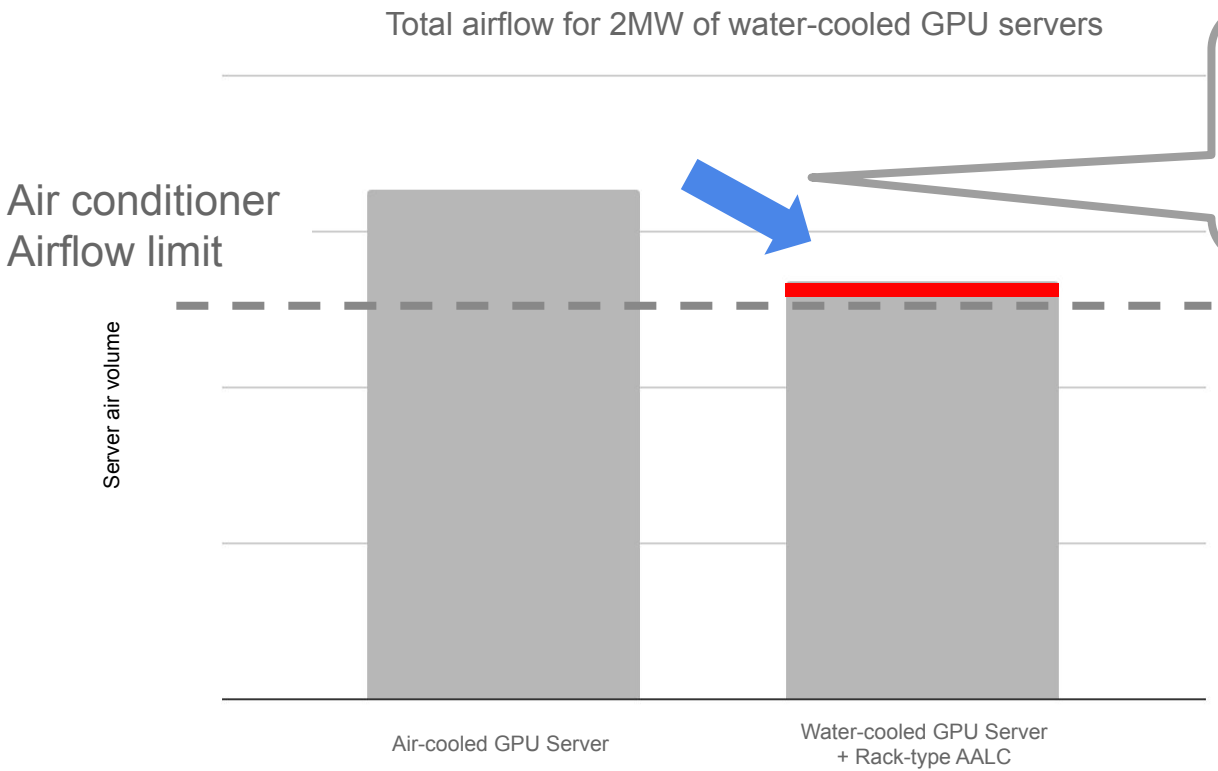
Key point: The amount of airflow for the GPU and CPU **has decreased** , while the amount of airflow for the AALC **has increased.**

Source: https://www.coolitsystems.com/products/#cdu

Actapio's 2MW server room with water-cooled GPU servers for LLM and a scenario with a full AALC installation

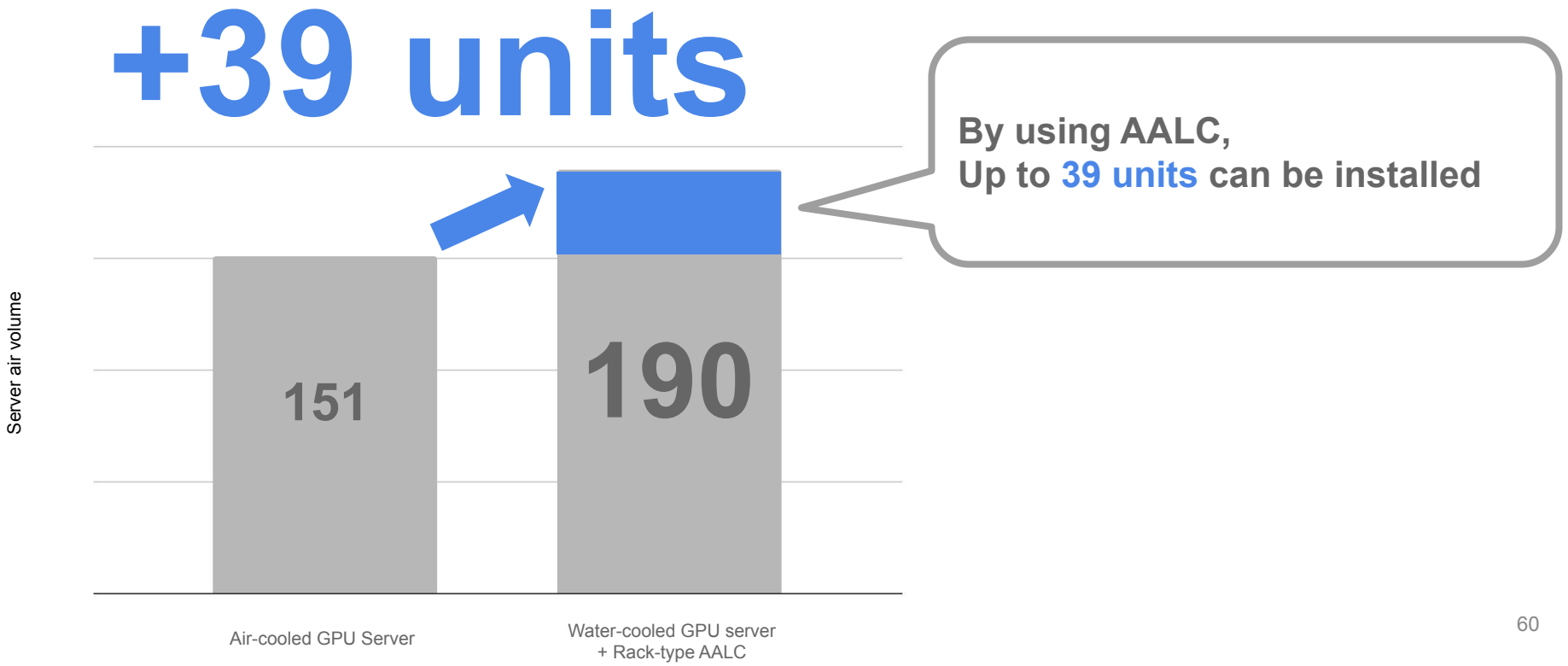


Actapio's 2MW server room with water-cooled GPU servers for LLM and a scenario with a full AALC installation



However, it was confirmed that the use of AALC can significantly reduce the air volume

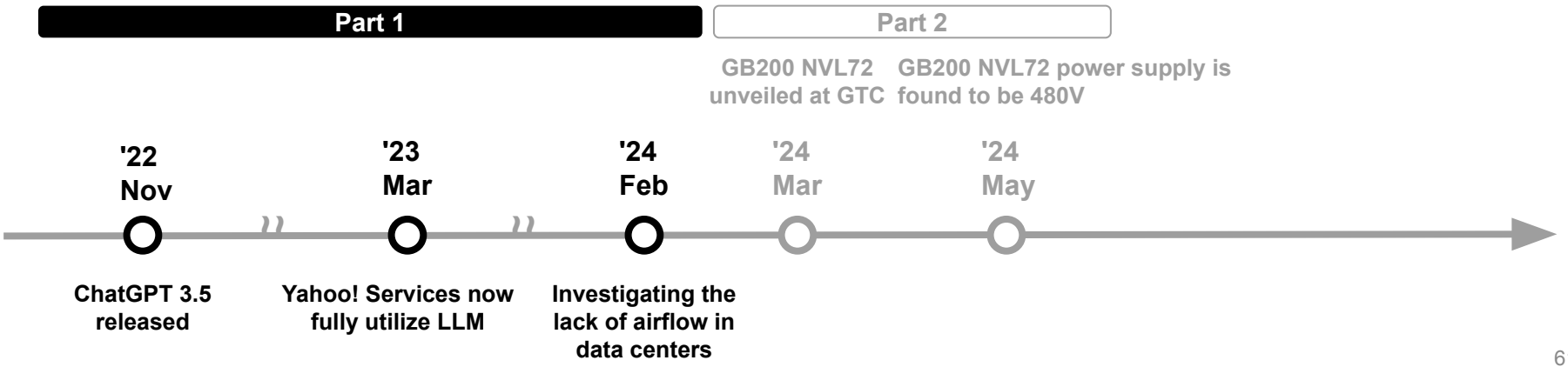
Actapio's 2MW server room with water-cooled GPU servers for LLM and a scenario with a full AALC installation



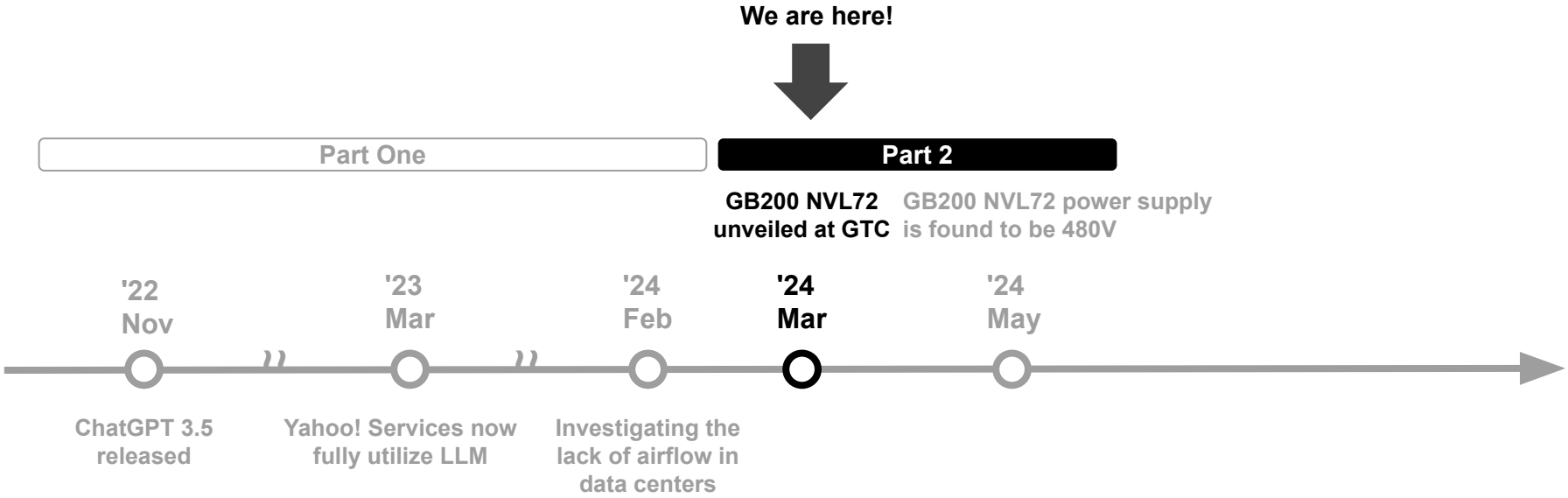
Part 1 Summary

In part 1 we covered

- LLMs require large amounts of data and enormous computing resources
- As a result, server workloads increase, straining data center resources
- I was told by a U.S. company that there may be a shortage of airflow in air conditioners
- air conditioning system airflow shortages **with AALC**



The emergence of GB200 NVL72 and its background



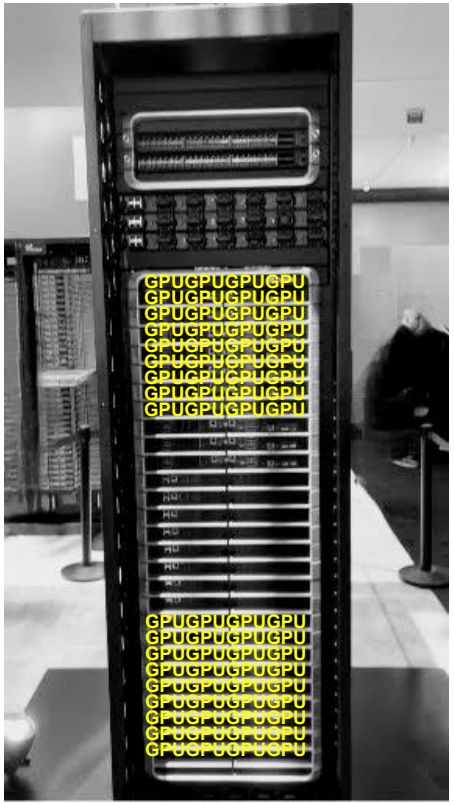
What is GB200 NVL72?

- A GPU system for LLM with 72 of the latest GPUs installed in one rack
- Provided as a complete rack, not as individual servers
- All GPUs in the rack are interconnected via NVLink, enabling high-speed inter-GPU communication.
- Compared to the previous model chip, H100, the training speed is 4 times faster and the power efficiency is 1/25
  (for GPT-MoE-1.8T model)

*Source: https://www.nvidia.com/en-us/data-center/gb200-nvl72/

What is GB200 NVL72?

- A GPU system for LLM with 72 of the latest GPUs installed in one rack
- Provided as a complete rack, not as individual servers
- All GPUs in the rack are interconnected via NVLink, enabling high-speed inter-GPU communication.
- Compared to the previous model chip, H100, the training speed is 4 times faster and the power efficiency is 1/25
  (for GPT-MoE-1.8T model)



*Source: https://www.nvidia.com/en-us/data-center/gb200-nvl72/

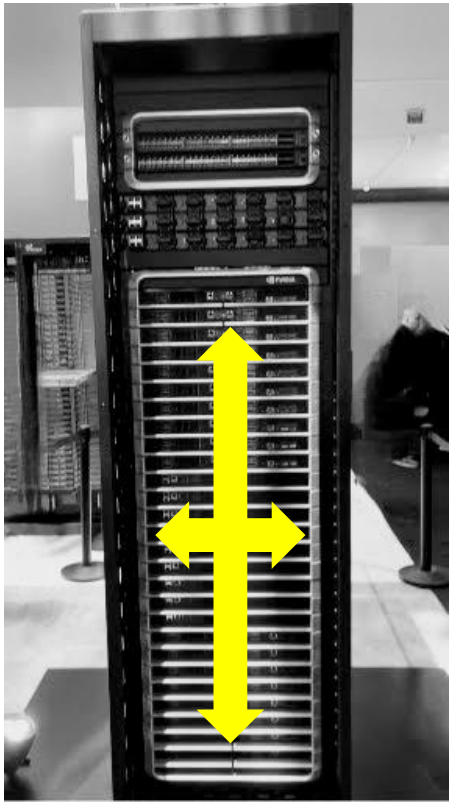
What is GB200 NVL72?

- A GPU system for LLM with 72 of the latest GPUs installed in one rack
- Provided as a complete rack, not as individual servers
- All GPUs in the rack are interconnected via NVLink, enabling high-speed inter-GPU communication.
- Compared to the previous model chip, H100, the training speed is 4 times faster and the power efficiency is 1/25
  (for GPT-MoE-1.8T model)

*Source: https://www.nvidia.com/en-us/data-center/gb200-nvl72/

What is GB200 NVL72?

- A GPU system for LLM with 72 of the latest GPUs installed in one rack
- Provided as a complete rack, not as individual servers
- All GPUs in the rack are interconnected via NVLink, enabling high-speed inter-GPU communication.
- Compared to the previous model chip, H100, the training speed is 4 times faster and the power efficiency is 1/25
  (for GPT-MoE-1.8T model)



*Source: https://www.nvidia.com/en-us/data-center/gb200-nvl72/

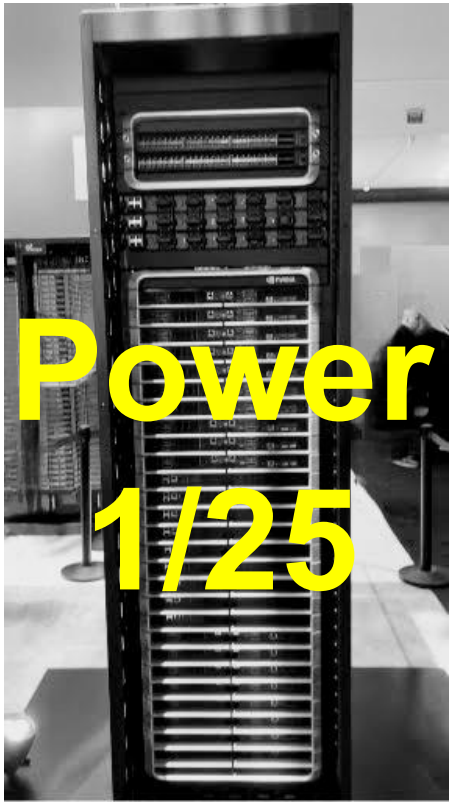
What is GB200 NVL72?

- A GPU system for LLM with 72 of the latest GPUs installed in one rack
- Provided as a complete rack, not as individual servers
- All GPUs in the rack are interconnected via NVLink, enabling high-speed inter-GPU communication.
- Compared to the previous model chip, H100, the training speed is 4 times faster and the power efficiency is 1/25
  (for GPT-MoE-1.8T model)

*Source: https://www.nvidia.com/en-us/data-center/gb200-nvl72/

What is GB200 NVL72?

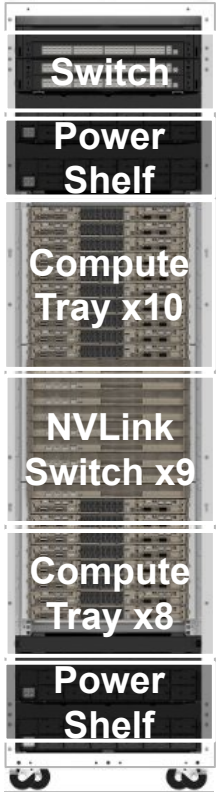
- A GPU system for LLM with 72 of the latest GPUs installed in one rack
- Provided as a complete rack, not as individual servers
- All GPUs in the rack are interconnected via NVLink, enabling high-speed inter-GPU communication.
- Compared to the previous model chip, H100, the training speed is 4 times faster and the power efficiency is 1/25
  (for GPT-MoE-1.8T model)



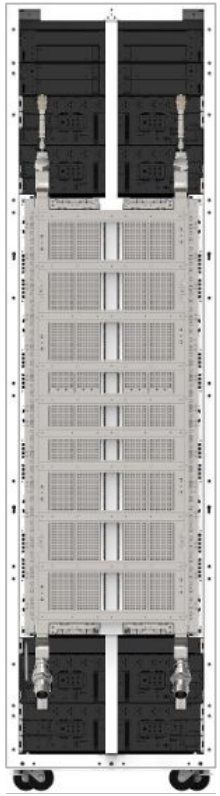
*Source: https://www.nvidia.com/en-us/data-center/gb200-nvl72/

Features of GB200 NVL72

Front View



Rear view

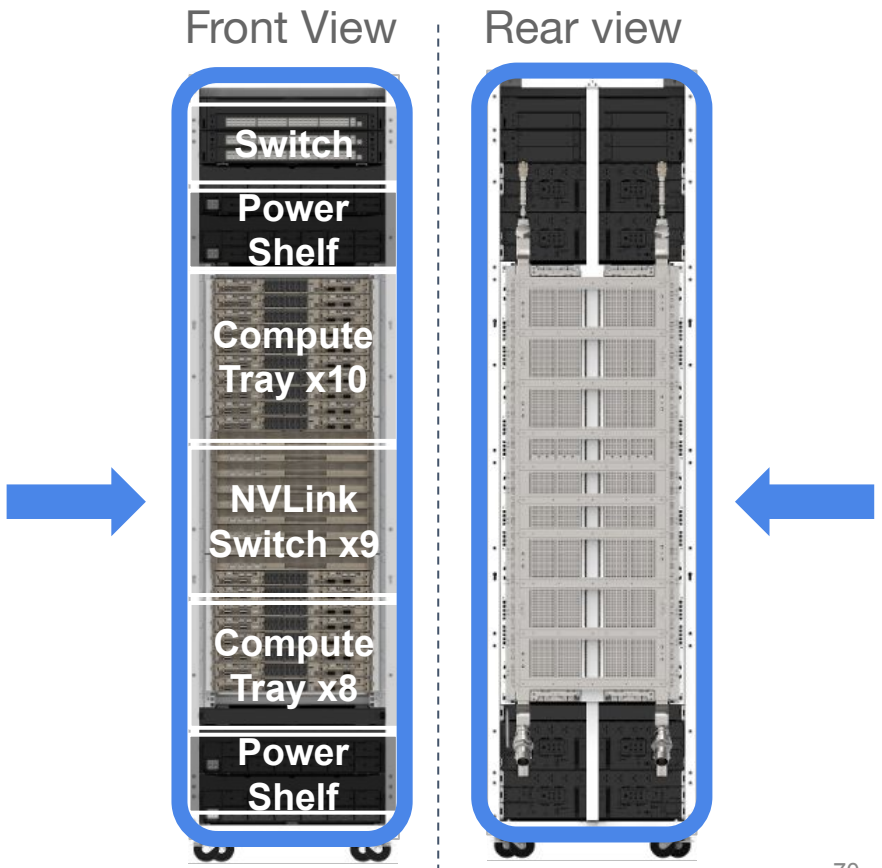


Sources: - Ingrasys (Diagram)
- https://www.youtube.com/watch?v=u9yB67nau7Q&ab_channel=InventecDataCenterSolutions (Specifications)

Features of GB200 NVL72

Rack: **OCP** ORv3 **rack**

- The ORv3 rack has a rack width of **21 inches.**
- **Both 19-inch and 21-inch** devices can be installed
- GB200 NVL72 is 19 inches except for the Power Shelf

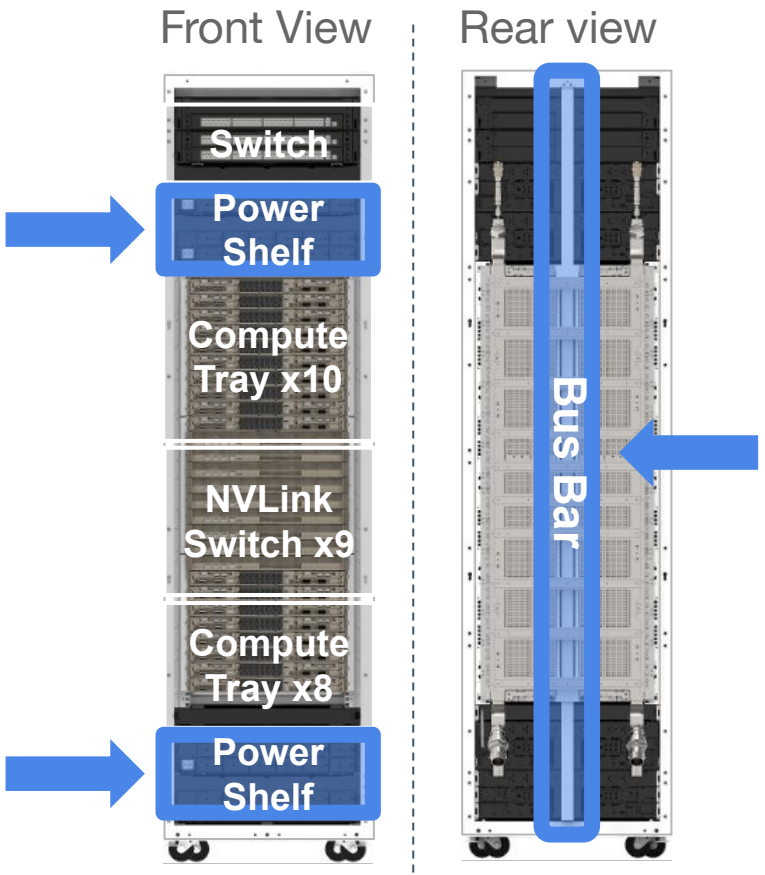


Sources: - Ingrasys (Diagram)
- https://www.youtube.com/watch?v=u9yB67nau7Q&ab_channel=InventecDataCenterSolutions (Specifications)

Features of GB200 NVL72

Power: **OCP bus bar** and **power shelf**

- Bus Bar is **48V DC voltage**
- allows you to add additional backup units and electrical power units if needed

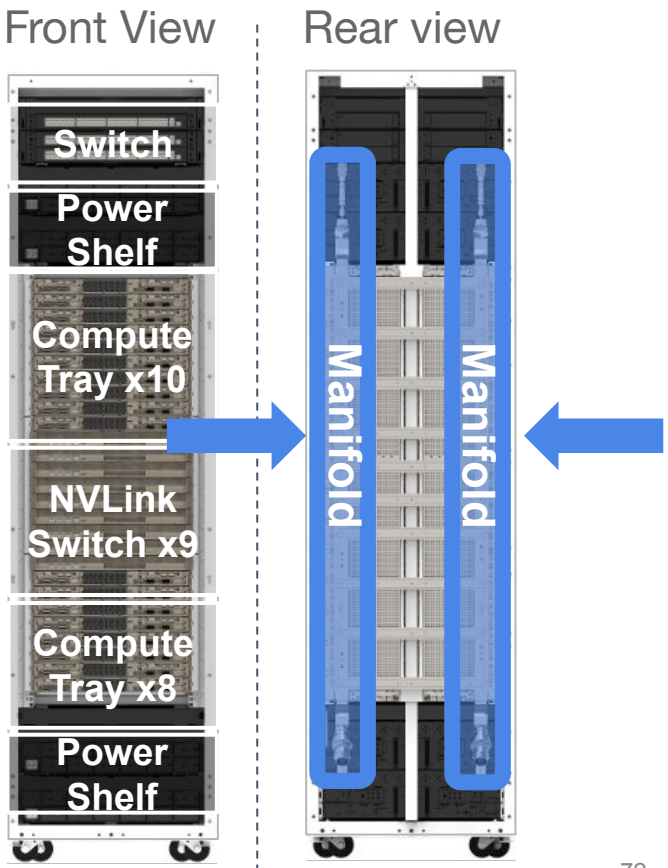


Sources: - Ingrasys (Diagram)
- https://www.youtube.com/watch?v=u9yB67nau7Q&ab_channel=InventecDataCenterSolutions (Specifications)

Features of GB200 NVL72

Manifold: **Collective piping for water cooling**

- Distributes coolant to each IT equipment in the rack
- Universal quick disconnect for quick and secure connection and disconnection
- Standardization with OCP ensures compatibility

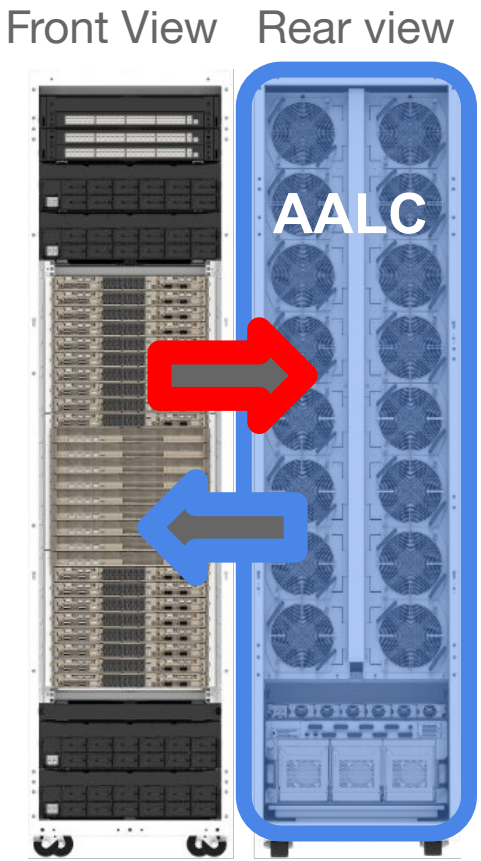


Sources: - Ingrasys (Diagram)
- https://www.youtube.com/watch?v=u9yB67nau7Q&ab_channel=InventecDataCenterSolutions (Specifications)

Features of GB200 NVL72

Cooling system: **Liquid cooling (DLC)** is required

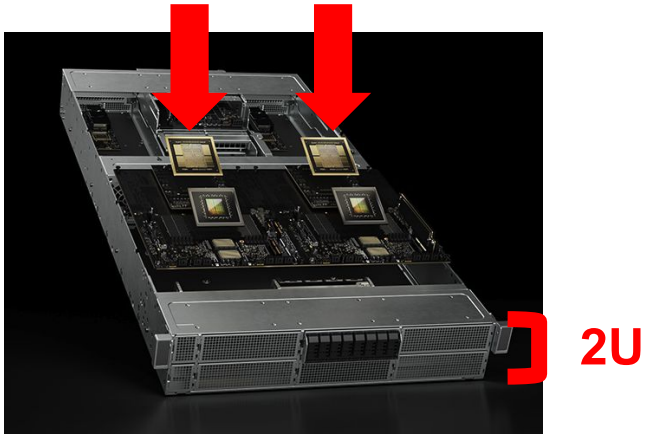
- It seems that each company is thinking about a model that connects with a rack-type AALC



Source: Ingrasys (Figure)

Why? Water cooling is essential

- The GB200 NVL2 is equipped **with two** NVIDIA Blackwell GPU chips and **has an air-cooled heat sink in a 2U** size.



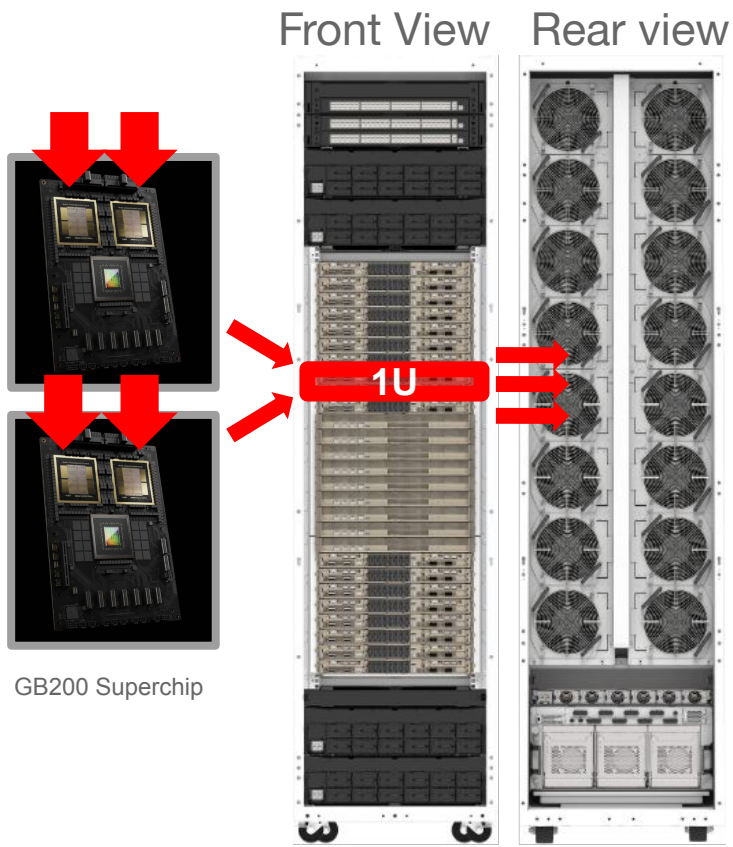
GB200 NVL2

Front View Rear view

Sources: - Ingrasys (rack diagram)
- https://www.nvidia.com/en-us/data-center/gb200-nvl2/ (GB200NVL2 diagram)

Why? Water cooling is essential

- GB200 NVL72 **has four NVIDIA Blackwell GPU chips in 1U**
- It is difficult to place an air-cooled heat sink in 1U **due to space limitations.**
- Therefore, it was necessary to use water cooling technology to carry heat

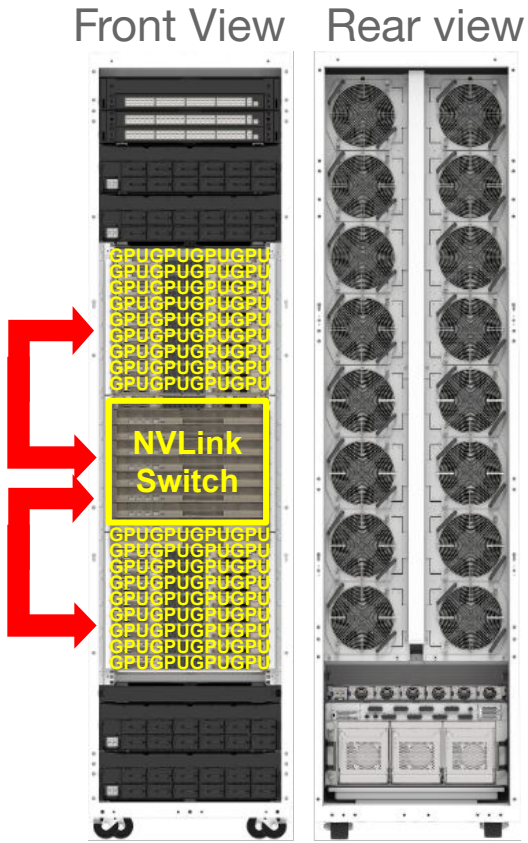


Sources: - Ingrasys (rack diagram
)

Why? Water cooling is essential

Why 1U?

- To connect 72 GPUs with **NVLink , you need to connect them with copper wires.**
- **The length** of the copper is limited.
- Therefore, it is necessary to place GPUs (servers) close to each other



Source: Ingrasys (Figure)

Why? Water cooling is essential

Why 1U?

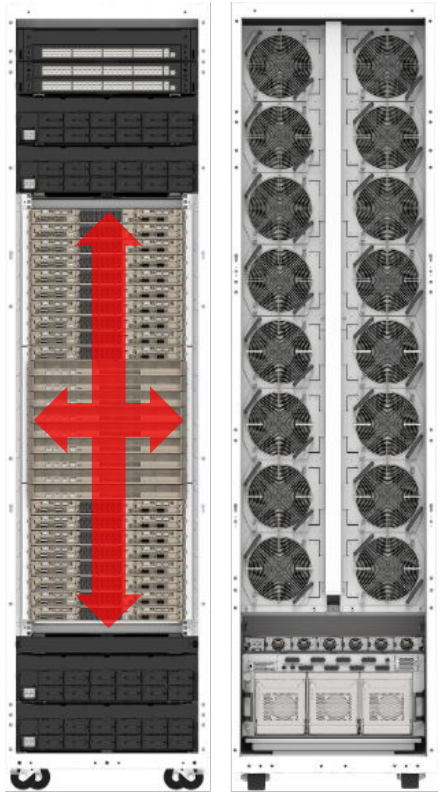
Why? Do I have to connect via NVLink?

- The challenge of LLM is **to shorten the model learning time**
- To shorten the model learning time, **high-speed communication between GPUs** is essential.
- High-speed communication is possible by connecting GPUs with NVLink.

Speed comparison (bi-directionally)

	GB/s	Gbps
NVLink 5th	**1,800**	**14,400**
PCIe-7 x16	512	4,096
800G Ether	200	1,600
400G Ether	100	800

Front View Rear view



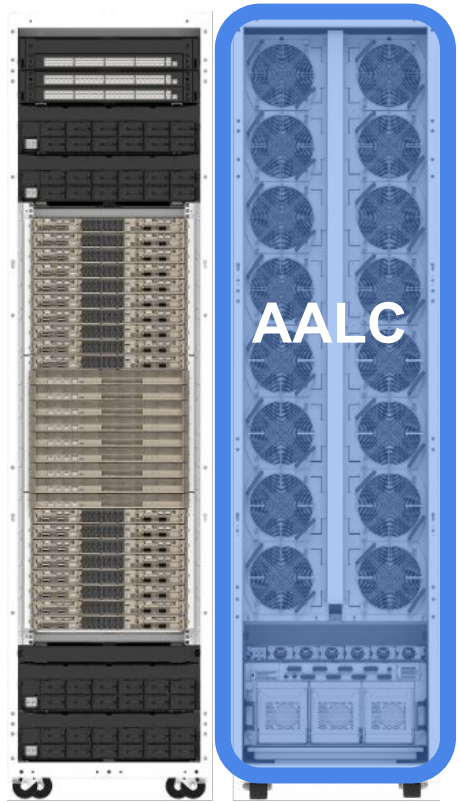
Source: - Ingrasys (figure)
- https://www.nvidia.com/en-us/data-center/nvlink/ (NVLink speeds)

Why? **" Water cooling is essential"**

Why 1U?

Why? Do I have to connect via NVLink?

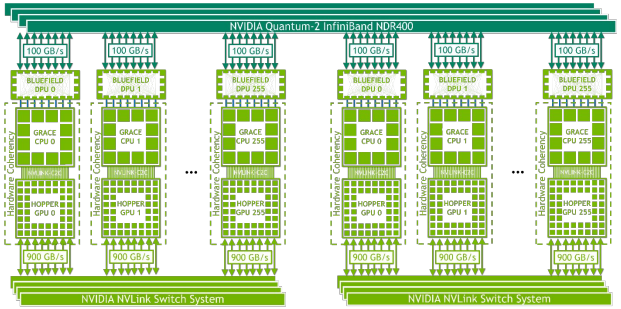
" To shorten the model learning time, which is an issue for LLM."



Source: Ingrasys (Figure)

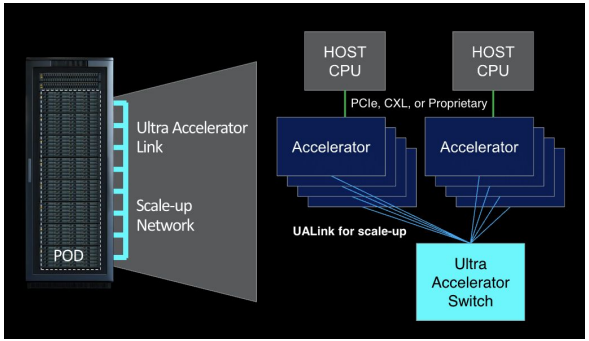
Extra point: Processor to Processor Networks : The Future of Connectivity?

NVlink



VS

UALink (Ultra Accelerator Link)



Developer: NVIDIA
Main use: Data transfer between processors
Features: High bandwidth, low latency, connects 576 GPUs
Scalability: Connect up to 576 GPUs (GH200 shown)

Developers: Intel, AMD, Broadcom, Cisco, HP, Google, Microsoft, Meta
Main use: Data transfer between processors
Features: High bandwidth, low latency, open standard
Scalable: Connect up to 1,024 AI accelerators

Source: https://techcrunch.com/2024/05/30/tech-giants-form-new-group-in-effort-to-wean-off-of-nvidia-hardware
https://developer.nvidia.com/blog/nvidia-grace-hopper-superchip-architecture-in-depth//

480V power supply (new challenge)

We are here!

Part One

Part 2

GB200 NVL72 unveiled at GTC

GB200 NVL72 power supply is found to be 480V

'22 Nov — ChatGPT 3.5 released

'23 Mar — Full-scale use of LLM begins on Yahoo! services

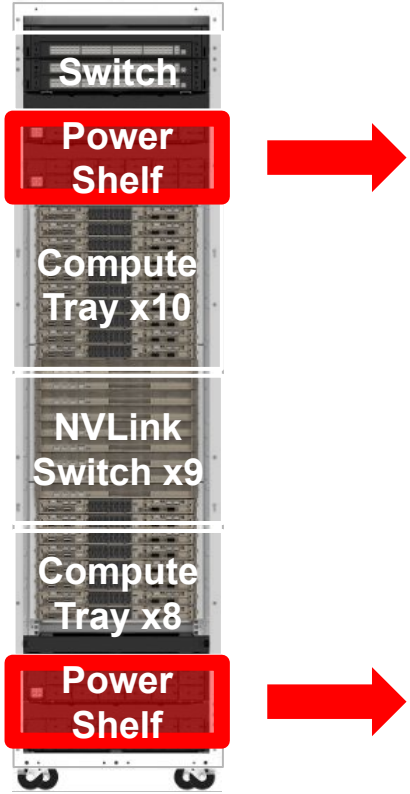
'24 Feb — Investigating the lack of airflow in data centers

'24 Mar

'24 May

GB200 NVL72 is a 120kw monster machine with 1 rack

Front View

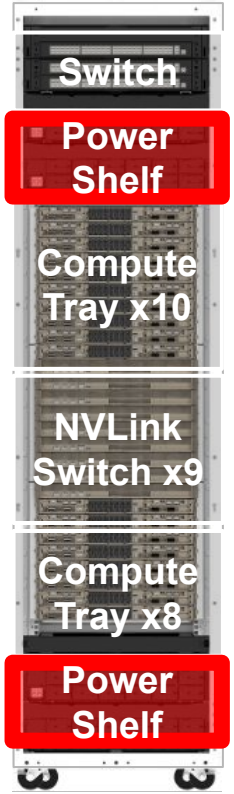


High power of 120 kW is required because it is packed into 1 Rack.

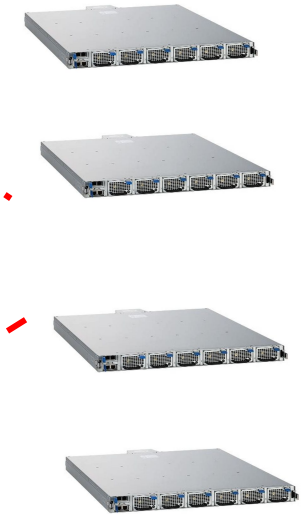
Source: Ingrasys (Figure)

GB200 NVL72 is a 120kw monster machine with 1 rack

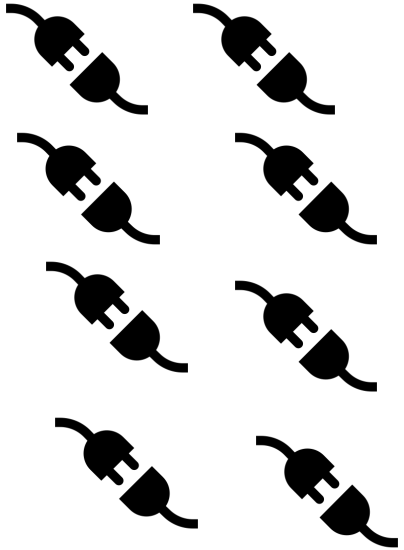
Front View



30kw or more x 4 units



480V x 8 19.0A

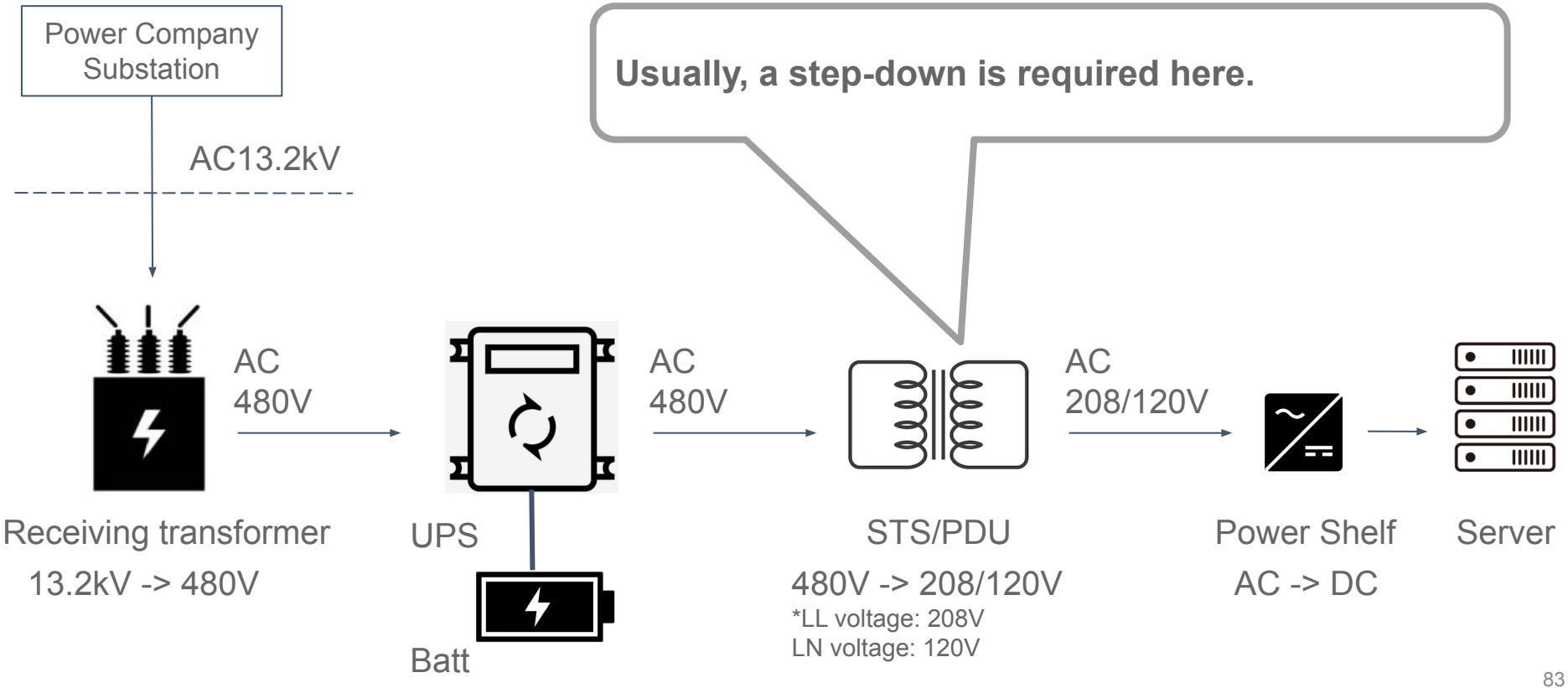


Because high voltage is used to connect to the server rack, Safety concerns

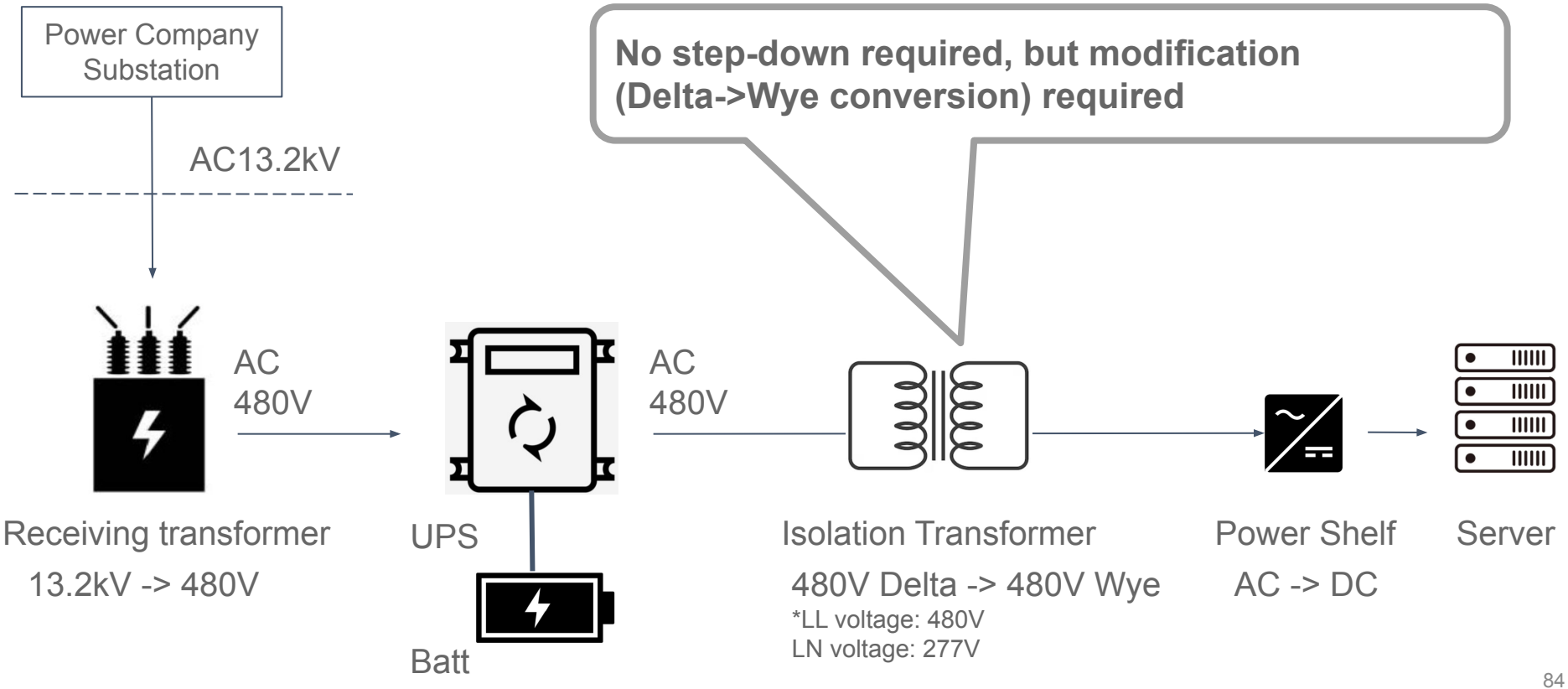
Source: Ingrasys (Figure)
https://www.murata.com/en-us/products/power/open-compute/overview/lineup/power-shelf

Power Supply Scheme

In normal cases

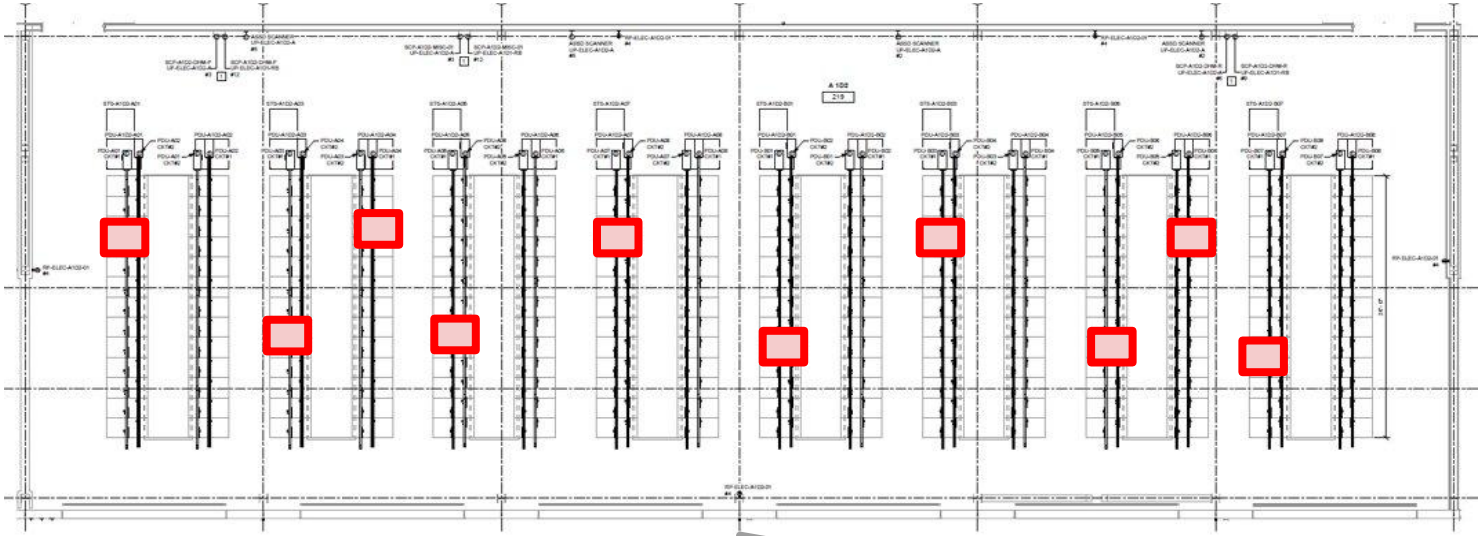


Power Supply Scheme **For 480V**



To be continued 😢

Only about 10 sets can fit in a 2MW server room.

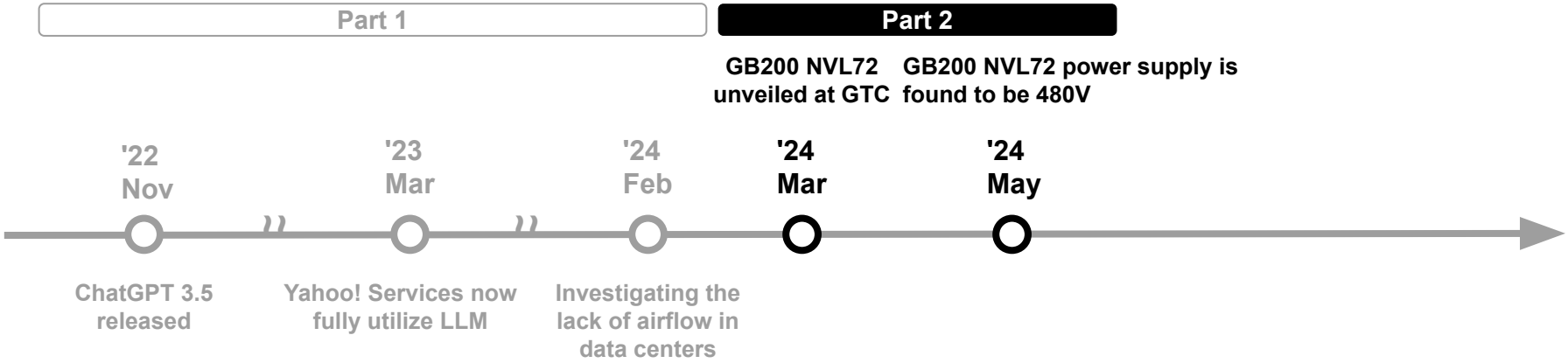


It's not space efficient, but the data center **land and building opex are so small** that it doesn't matter 🤷

Part 2 Summary

In part 2 we covered

- **Reducing model learning time** is a major issue
- The GB200 NVL72 technology announced at GTC **is a solution to this problem.**
- **Proximity makes it difficult to cope with heat sinks, and water cooling technology is required for efficient cooling**
- The servers packed into a single rack **created a new challenge: the power supply.**

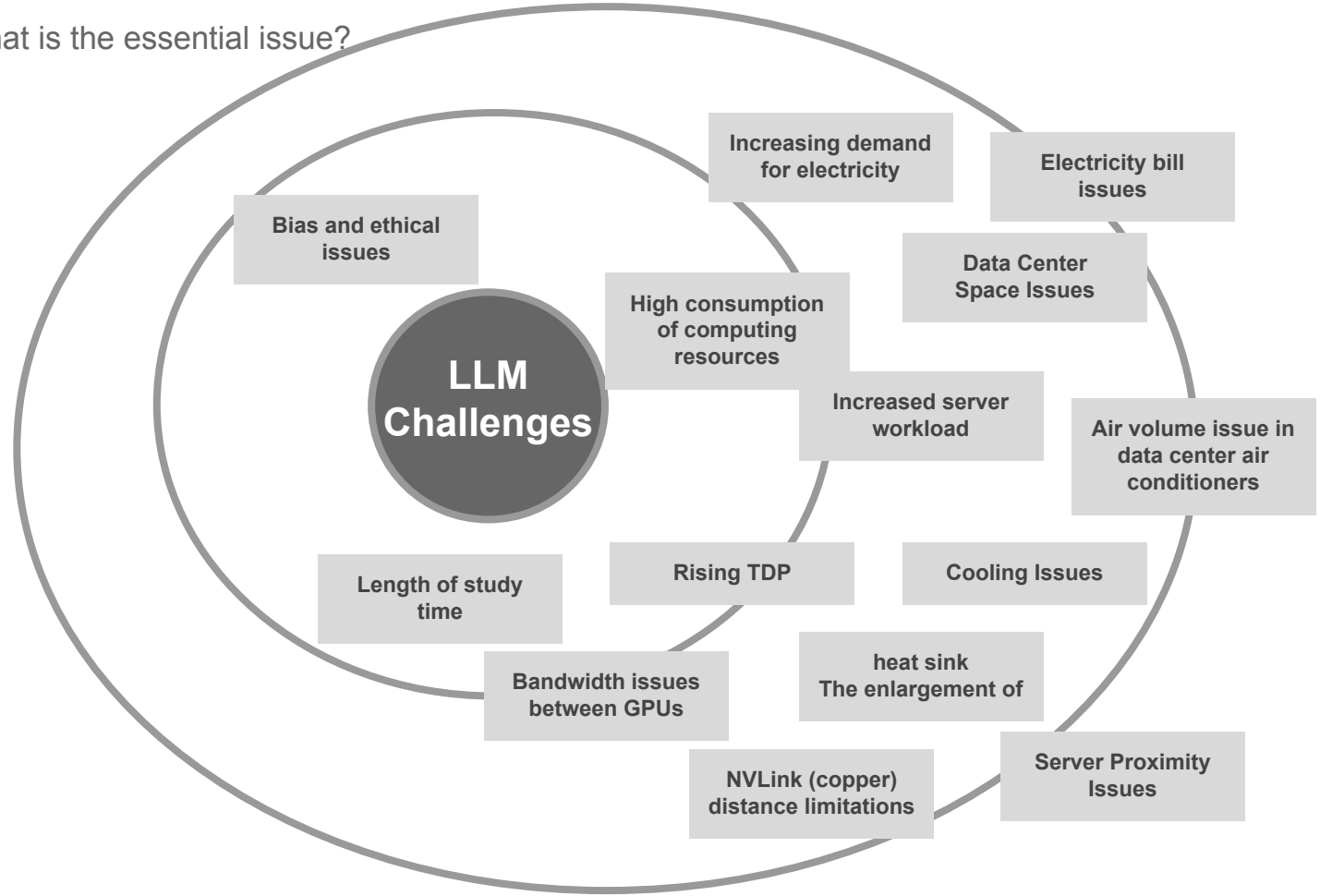


Summary

- **The popularity of LLM** has brought new challenges to our data center operations.
- LLM creates increased server workloads and associated cooling challenges
- Specifically, it surfaced as a problem of insufficient airflow in the air conditioning system and server proximity constraints

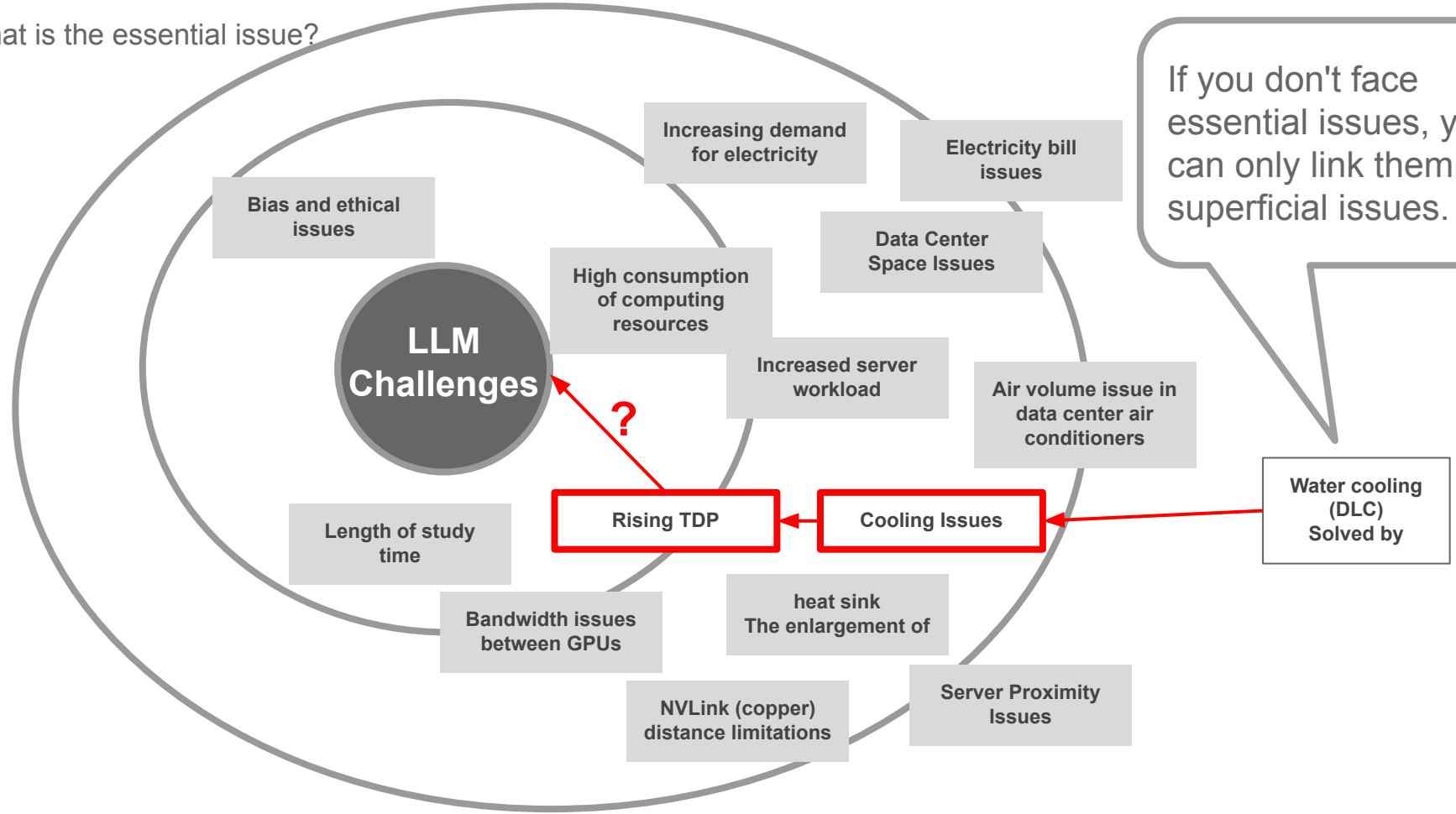
- We have spoken with U.S. companies and hyperscalers at GTC, OCP Summits, and elsewhere, and carefully considered that water cooling technology is the key to solving these "essential challenges"

What is the essential issue?



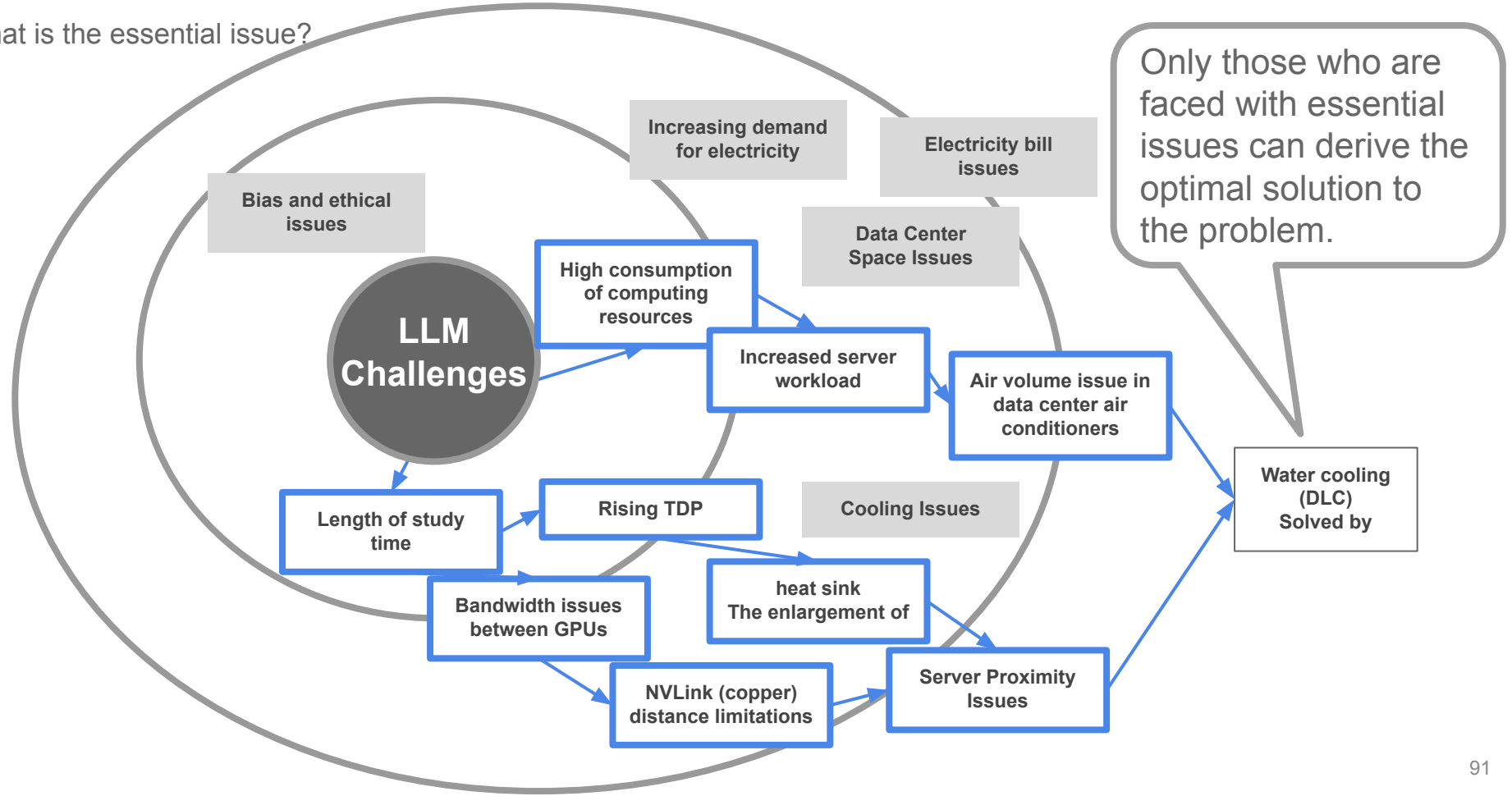
Water cooling (DLC) Solved by

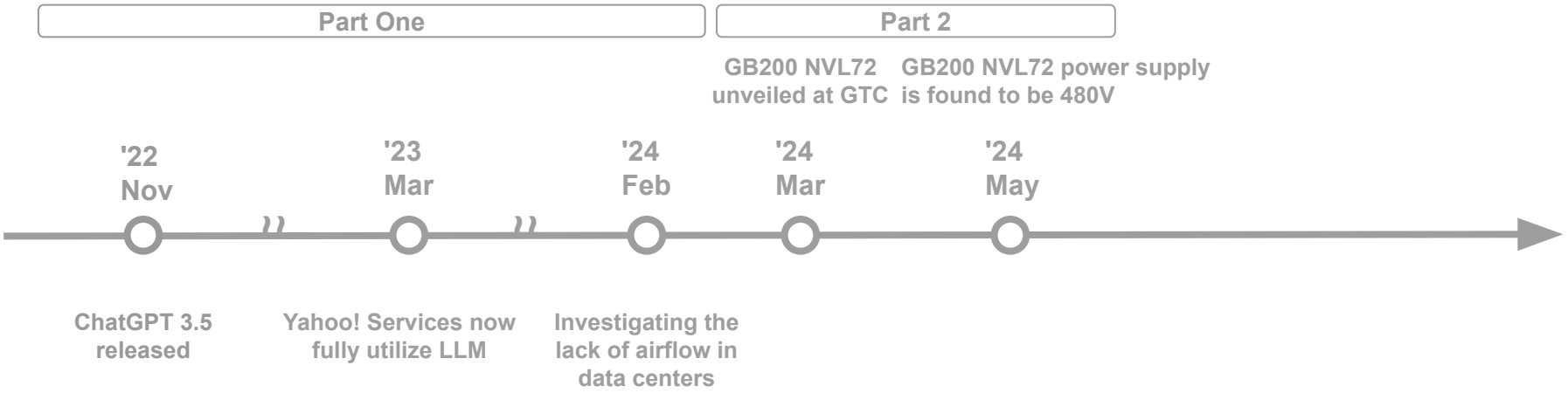
What is the essential issue?



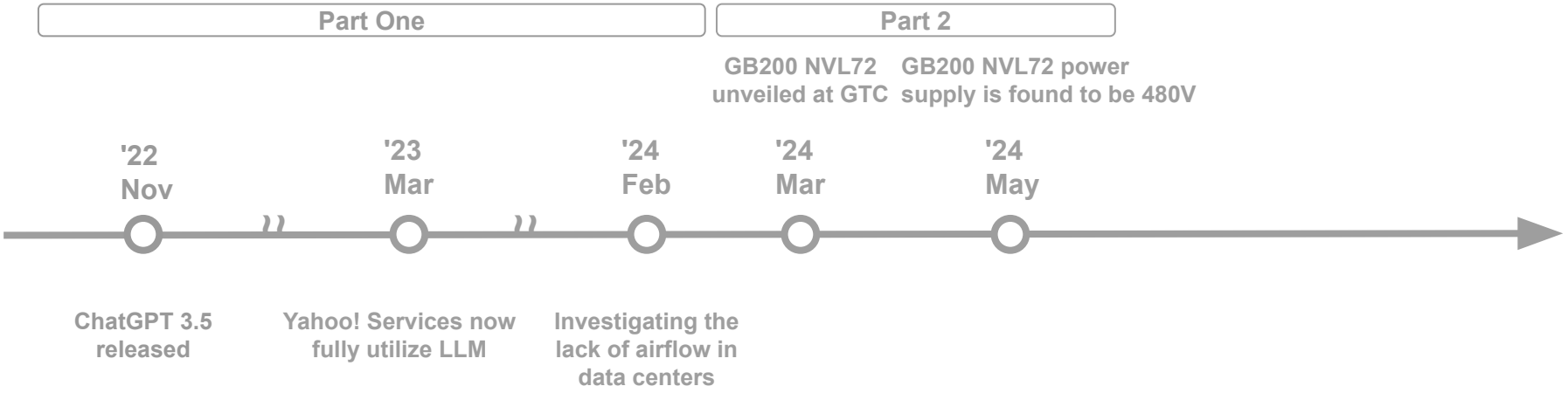
If you don't face essential issues, you can only link them to superficial issues.

What is the essential issue?





U.S. companies (hyperscalers) and others are working to solve the problems they faced at this point



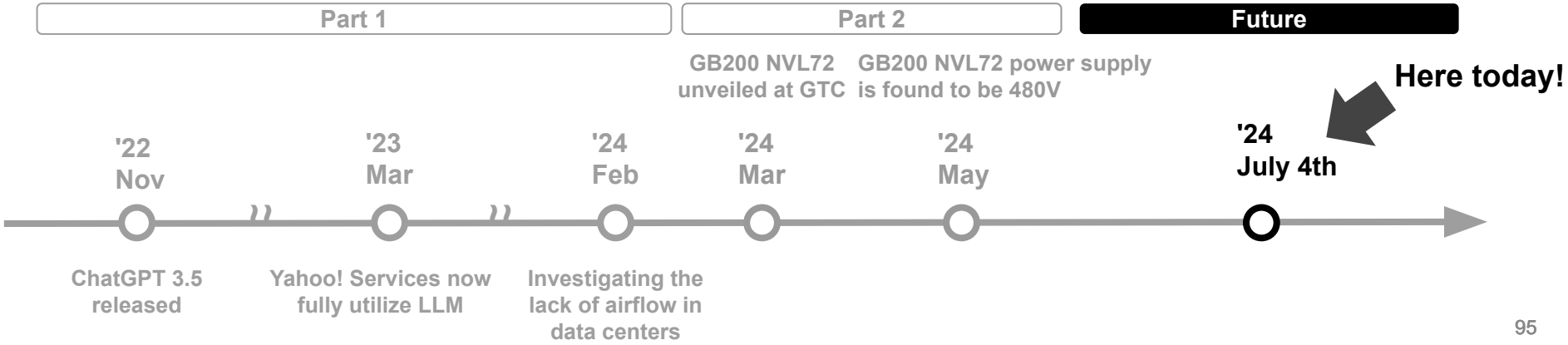
- It is at this point that we **begin to encounter challenges**
- We encounter problems **later than American companies**

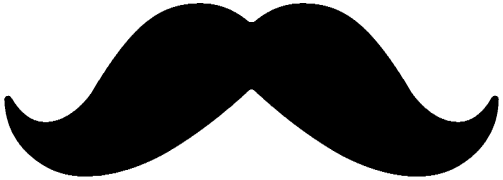
How to avoid mistaking the solutions to future problems
for those who have not faced any challenges yet

- Gather information on what the essential challenges are for those who face challenges
- In terms of LLM, **hyperscale companies such as US companies**
- It is important to immediately put the information gathered into practice in the near future (in the United States)

- Actapio's problem-solving approach
  - We collect, verify, and implement information from people who have faced challenges (U.S. companies), so we are able to select the best technology

- Vision for the future
  - **The "users"** are left behind
  - In order to lead the industry, we must become a "pioneer in problem solving"
  - As we look towards the future, will we aim to be on the "user side" that quickly adopts technology, or will we aim to be "pioneers in solving problems" and become industry leaders?
  - Whichever path you choose, let's start thinking together **today** about what we need to do to pave the way for the future.





THANKS

Our Team

Aaron Kirkpatrick
Andrew Arnold
Anthony Skwiat
Atsuko Ishigaki
Eiji Kawauchi
Jason Van Winkle
J.D. Salling
Kai Fukazawa
Ken Tairabune
Koyo Uemizu
Kyoya Torikai
Hideki Mikami
Hisatomo Tanaka
Marie Rudolph
Masahiko Matsui
Masayuki Ashida
Osamu Kurino
Shinichiro Okamoto
Susan McKee
Takashi Watanabe
Takuya Kitano
Takumi Uematsu
Tatsumi Yuusuke
Travis Mather
Yoshimura Tsubasa
Yuji Kohata
Yukihito Imai

Our Datacenter