AI-Factories Infrastructure: Design, Power & Cooling Dynamics

Summary

Direct liquid cooling (DLC) remains a hot topic among data center owner/operators – and results from Uptime Institute's 2024 Cooling Systems Survey (n=964) provide evidence why.

A total of 22% of respondents report their organization is making some use of DLC in their facilities. Three in five (61%)say they are not currently using DLC but would consider doing so in the future.

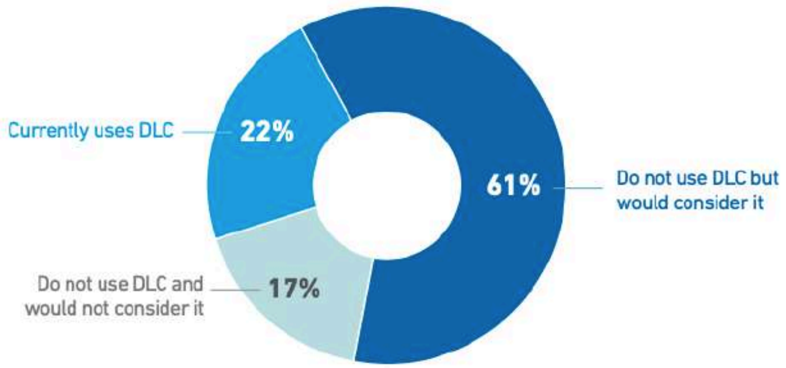
Nearly half of DLC users report that less than 10% of their organization is using it in "only a few servers," while 47% say DLC is being used in "some IT racks."

About this survey
Uptime's 2024 Data Center Cooling Systems Survey was conducted between February 8, 2024, and March 13, 2024, and had a total of 964 data center industry respondents.

The questions in this report were answered by data center owner/operators, with the exception of question VII which was answered by both owner/operators and vendors/product providers.

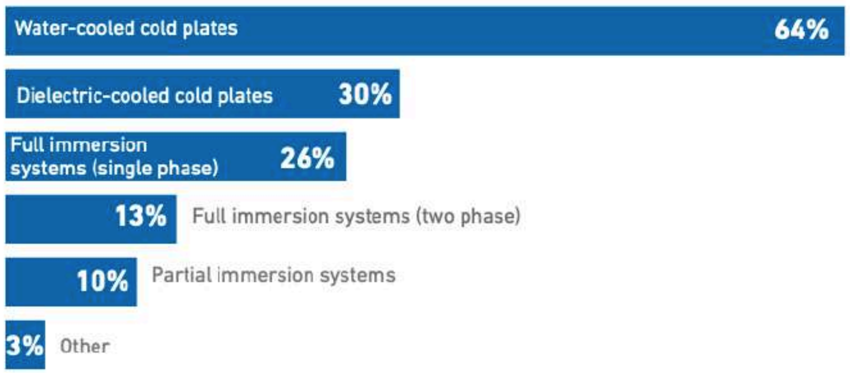
2024 use of direct liquid cooling

I. Do you currently use, or are you considering using, any direct liquid cooling technology in your owned data centers? (n=453)



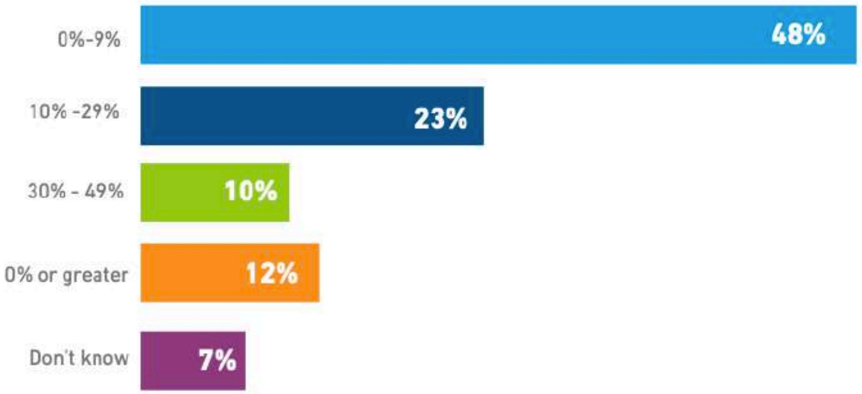
Types of direct liquid cooling being used

II. Which types of direct liquid cooling are currently used in your data center / colocation facilities? Choose all that apply. (n=94)



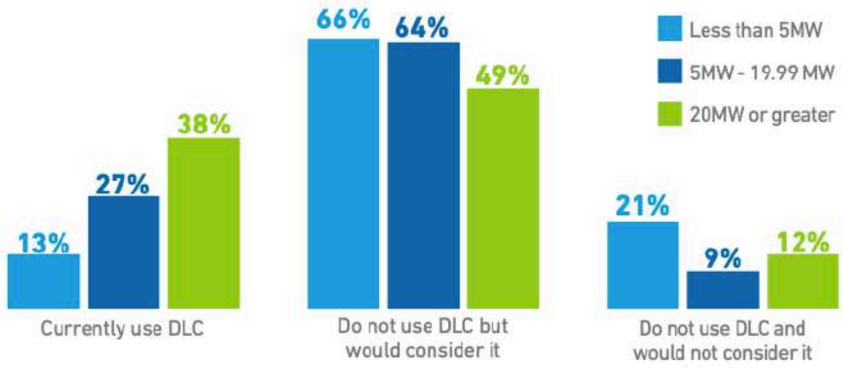
Percentage of IT racks incorporating direct liquid cooling

III. What percentage of your organization's IT racks would you say incorporate direct liquid cooling? (n=94)



2024 use of direct liquid cooling (by size of largest data center)

IV. Do you currently use, or are you considering using, any direct liquid cooling technology in your owned data centers? (n=397)

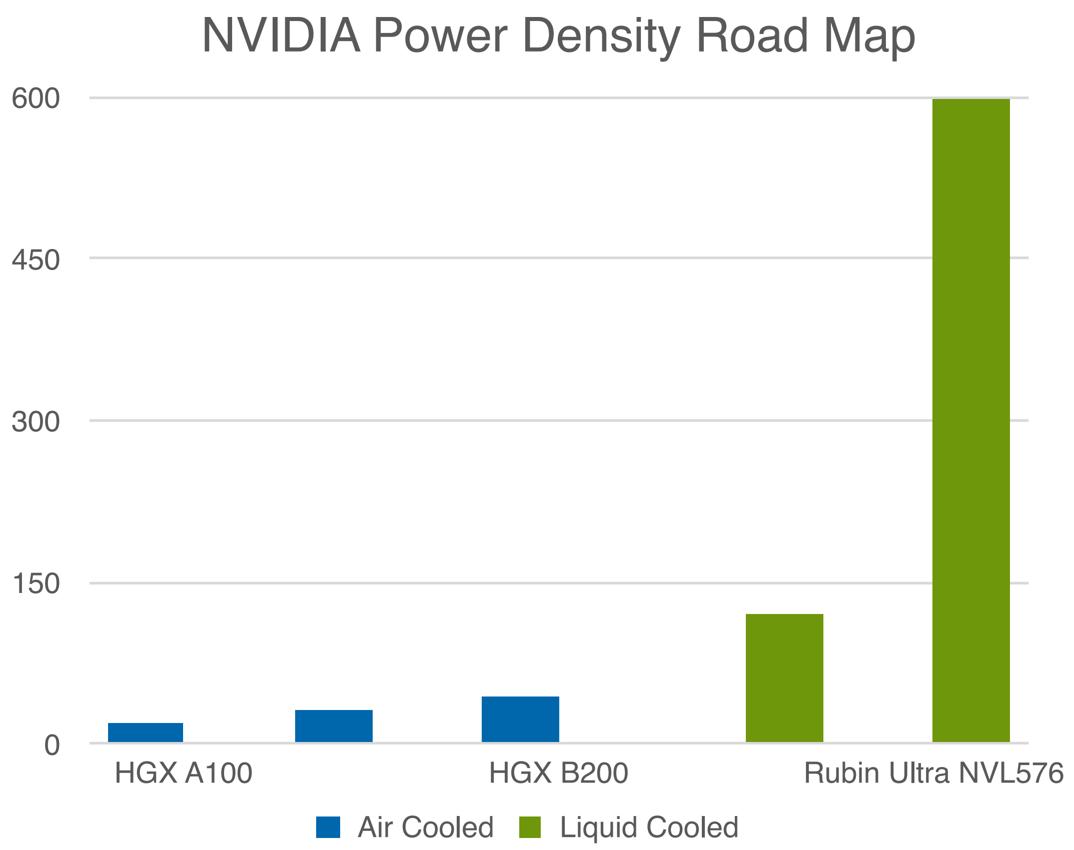


Source: Uptime Intelligence

uptime INSTITUTE

Rack Power Densities

Platform	GPU Architecture	GPU Count	GPU Memory	Memory Bandwidth	Target Use Case	Power & Cooling
HGX A100	Ampere	Up to 8	Up to 80GB	~1.6 TB/s	AI training/inference (2021-2022)	Moderate
HGX H100	Hopper	Up to 8	80-96GB	3.35 TB/s	Large-scale AI workloads (2023+)	Higher power/ cooling
HGX B200	Blackwell	8	192GB	8 TB/s	Next-gen extreme AI training/inference	High, liquid cooled
GB200 NVL72	Blackwell + Grace	72 GPUs +	1.44TB+	Rack-scale fabric	Multi-trillion parameter model training	Very high, liquid
Rubin Ultra NVL576	Blackwell	576 GPUs	Ultra-high	Ultra-high	Hyperscale AI factory deployments	Ultra-high, liquid



uptime INSTITUTE

DESIGN CONSIDERATIONS

AI Work Loads

- Space Requirements: Liquid cooling systems may require additional space for coolant distribution units (CDUs), heat exchangers, piping, pumps, and tanks, depending on the type of liquid cooling (cold plate or immersion).
- Floor Load Capacity: Ensure the data center floor can support the weight of liquid-cooled racks and equipment, especially for immersion cooling tanks.
- Floor Plan: For immersion cooling, modifications may be needed to accommodate tanks and provide appropriate access for maintenance and installation.
- Dielectric Fluid Selection: Choose a suitable dielectric fluid (non-conductive liquid) for immersion cooling. Consider factors like thermal conductivity, viscosity, chemical stability, cost, and environmental impact.
- Coolant Quality and Maintenance: For both cold plate and immersion cooling, regular monitoring of coolant quality is essential to prevent contamination, corrosion, or degradation.
- Leak Detection and Containment: Implement leak detection systems and secondary containment measures to quickly identify and contain leaks, minimizing potential damage and ensuring safety.
- Heat Exchangers and CDUs: Design and place heat exchangers and CDUs to effectively transfer heat from the coolant to the external environment.

Design

Example

A 60 MW Data Center direct-to-chip cold plate reference design is used for illustration.

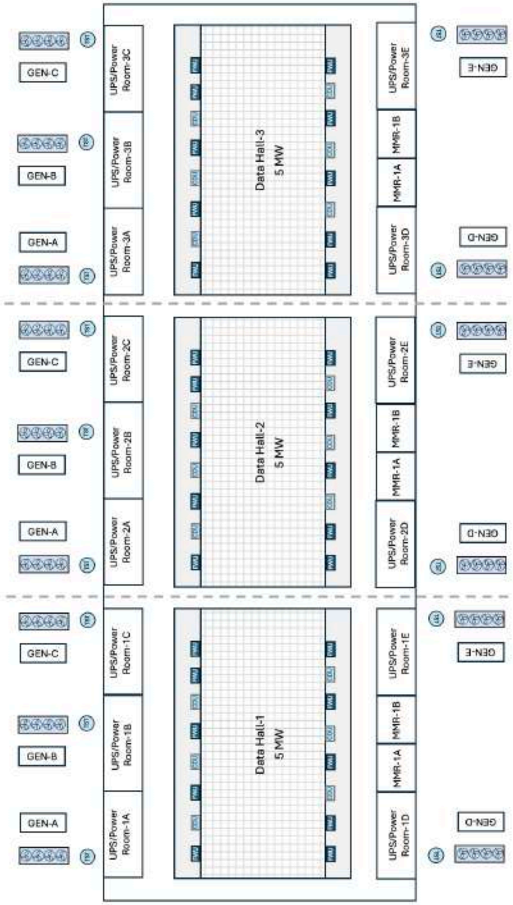
Tier Standard - Topology

Design

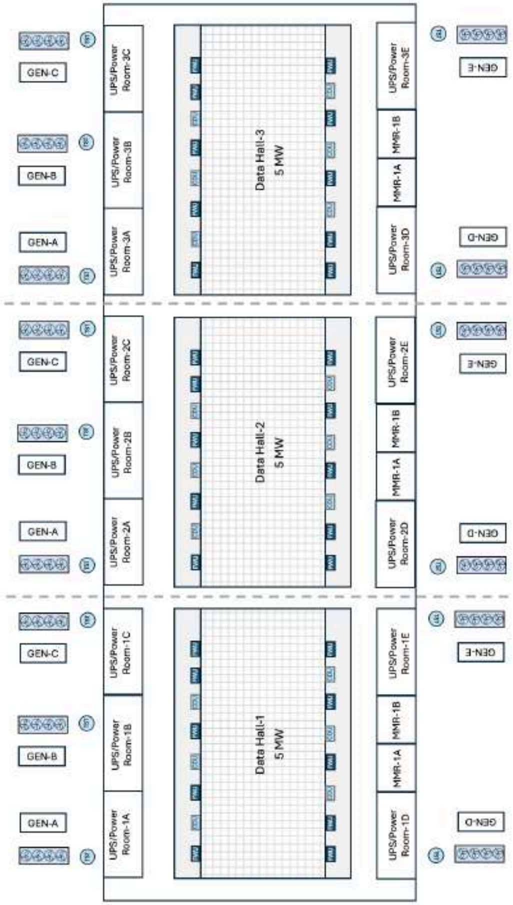
	Tier I	Tier II	Tier III	Tier IV
Minimum Capacity Components to Support the IT Load	N	N+1	N+1	N After any Failure
Distribution Paths - Electrical Power Backbone	1	1	1 Active and 1 Alternate	2 Simultaneously Active
Critical Power Distribution	1	1	2 Simultaneously Active	2 Simultaneously Active
Concurrently Maintainable	No	No	Yes	Yes
Fault Tolerance	No	No	No	Yes
Compartmentalization	No	No	No	Yes
Continuous Cooling	No	No	No	Yes

Table 1: Tier Requirements Summary

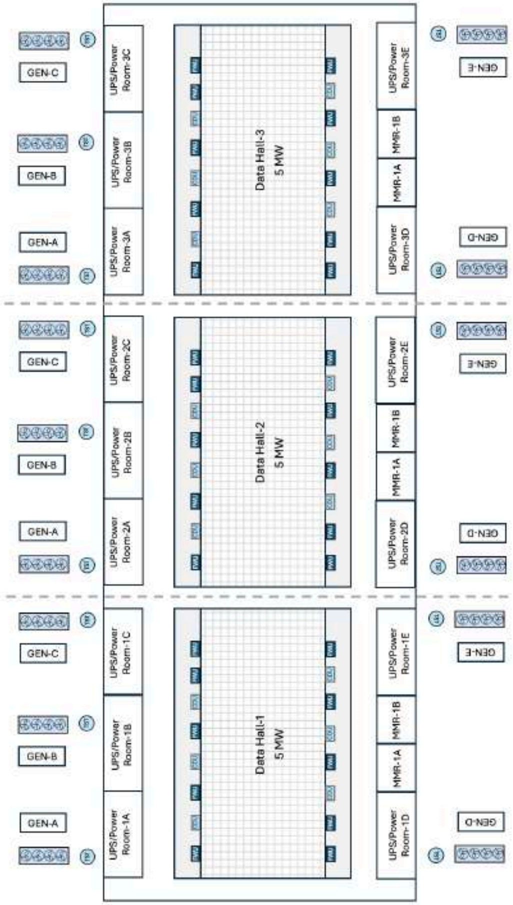
Reference Design 60 MW (Tier III, CW, 130 kW/ Rack)



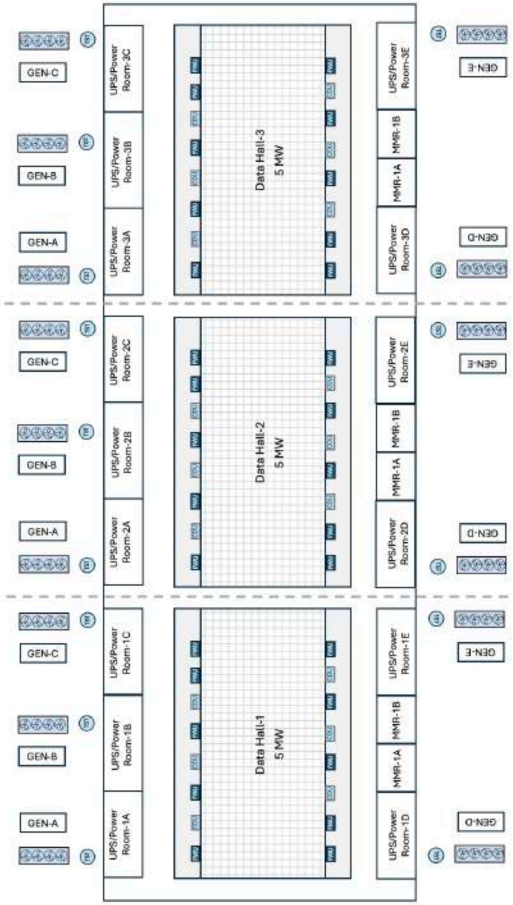
15 MW



15 MW



15 MW

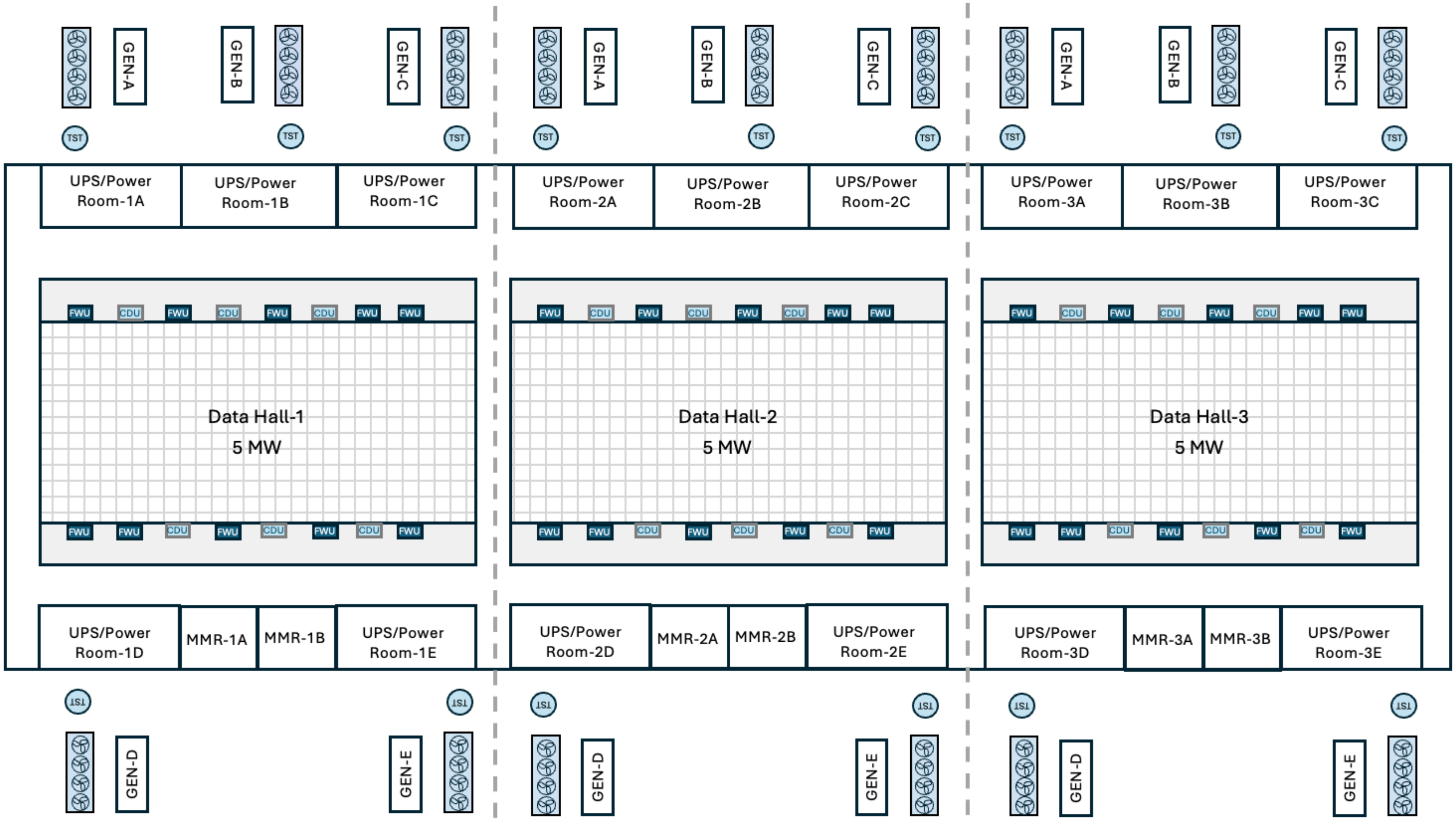


15 MW

uptime INSTITUTE

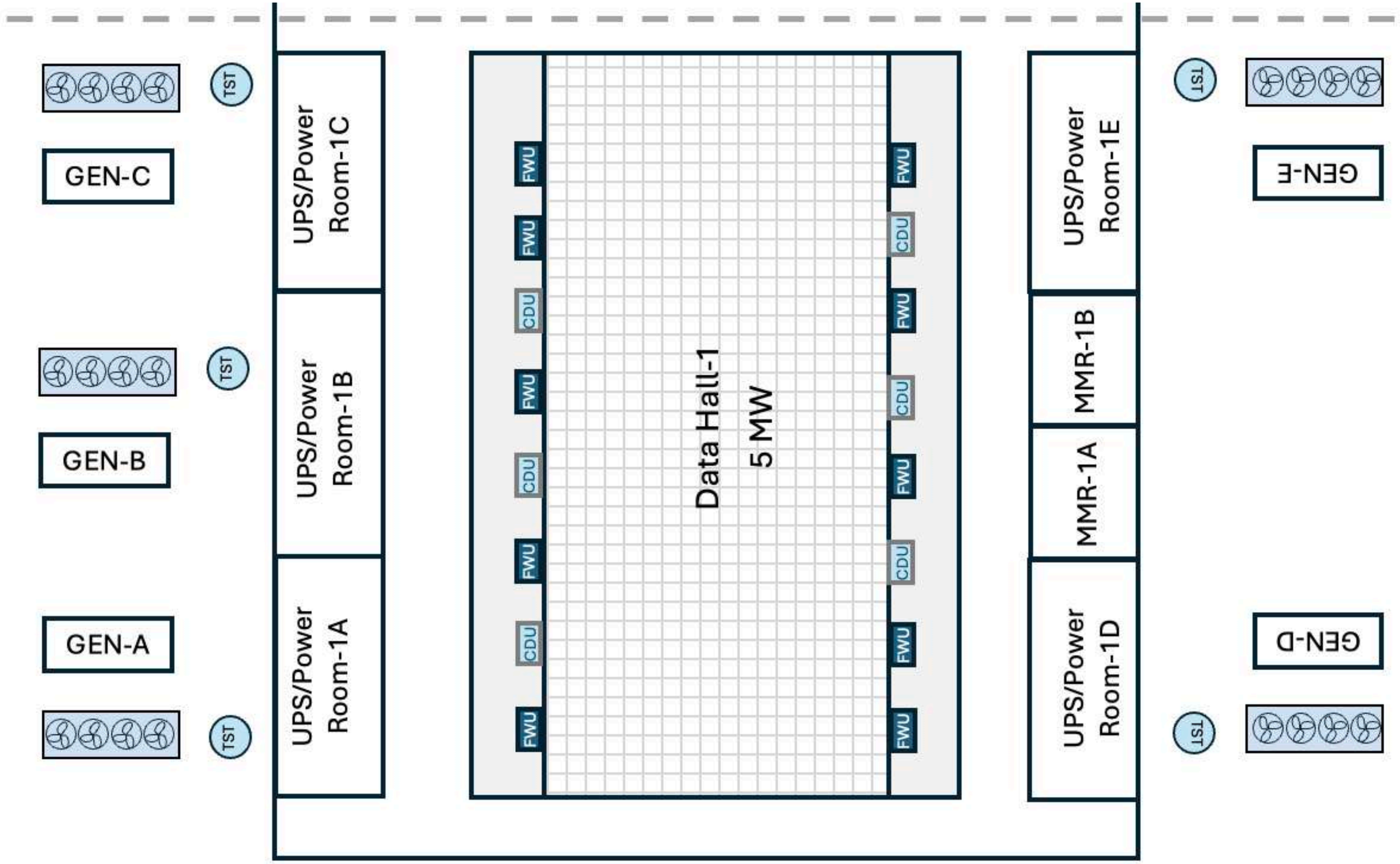
Reference Design 60 MW (Phase-1, 15 MW, Tier III, CW, 130 kW/Rack)

15
MW



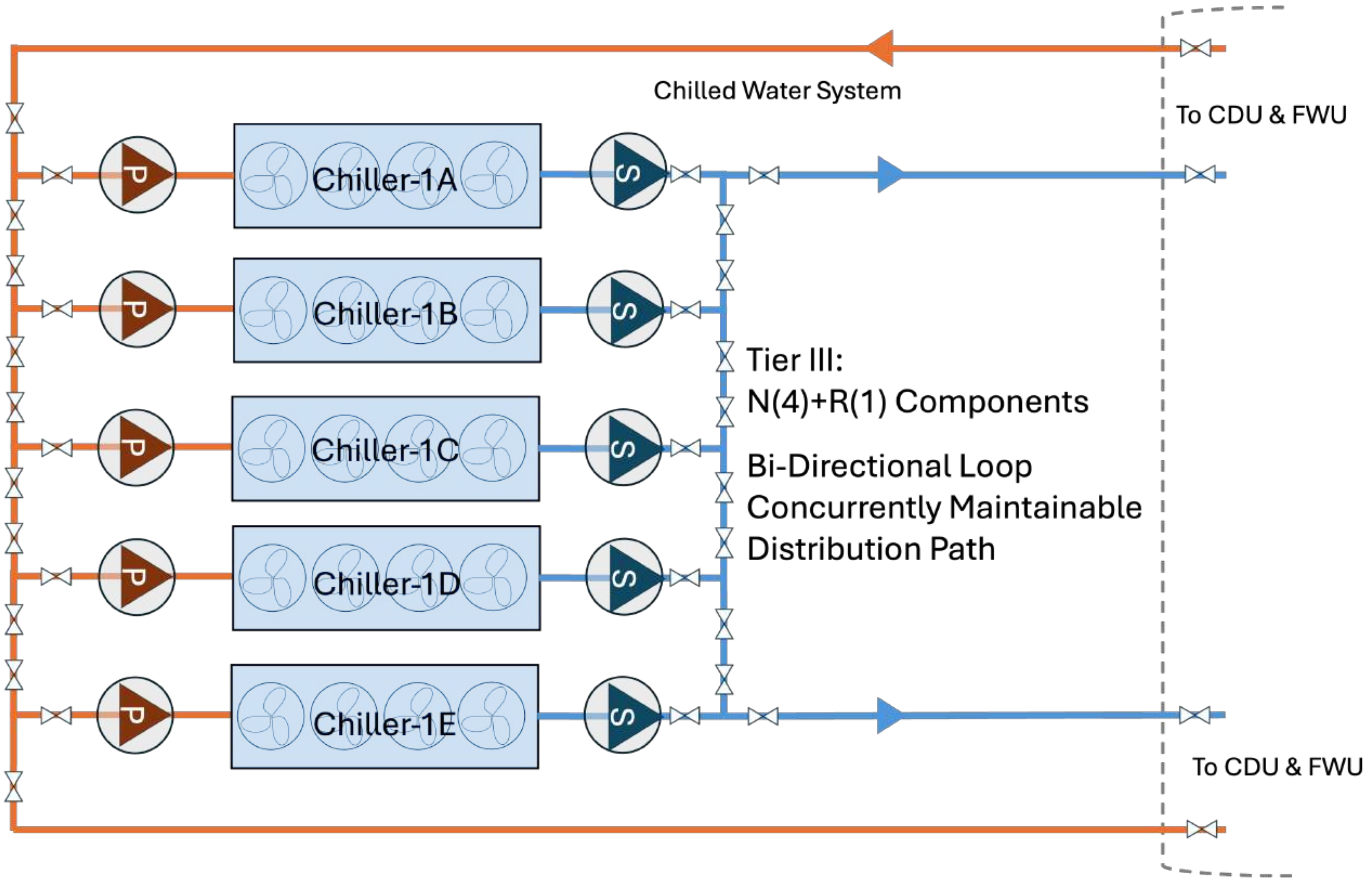
uptime INSTITUTE

Reference Design 60 MW (Base Module, 5 MW, Tier III, CW, 130 kW/ Rack)



uptime INSTITUTE

Chiller Plant Design (N+1, 5 MW, Tier III, CW, 130 kW/Rack)



Technology Cooling Design (N+1, 5 MW, Tier III, CW, 130 kW/ Rack)

CDU: 1,250 kW

Each Loop: N+1,

N=2 (2,500 kW)

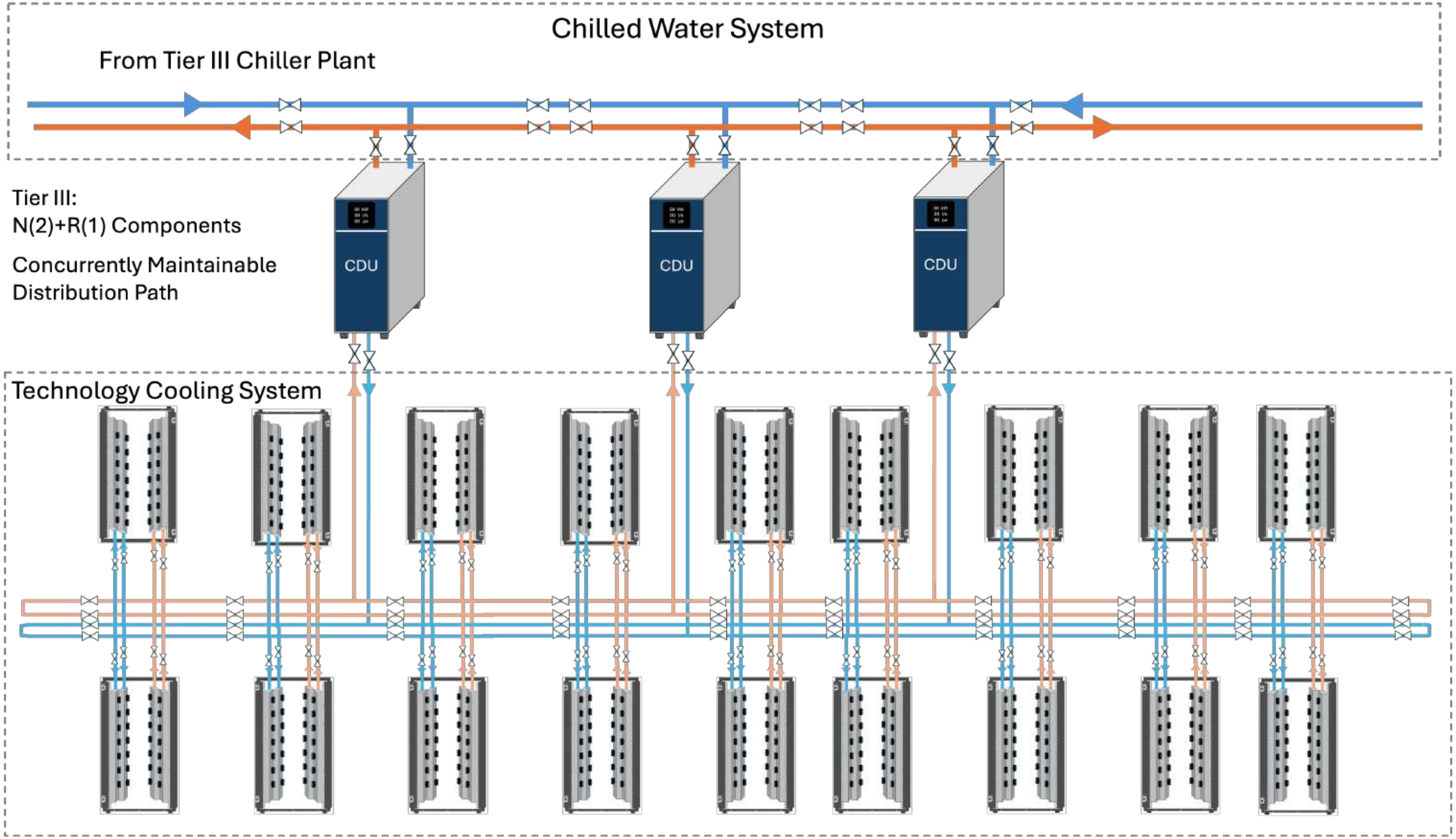
IT Racks: 18 No.

DLC @ 95% = 2,223

kW

DLC @ 55% = 1287

kW



uptime INSTITUTE

Technology Cooling Design (N+1, 5 MW, Tier III, CW, 130 kW/ Rack)

CDU: 1,250 kW

Data Hall: 2 Loops

N+2, N=4 (5,000 kW)

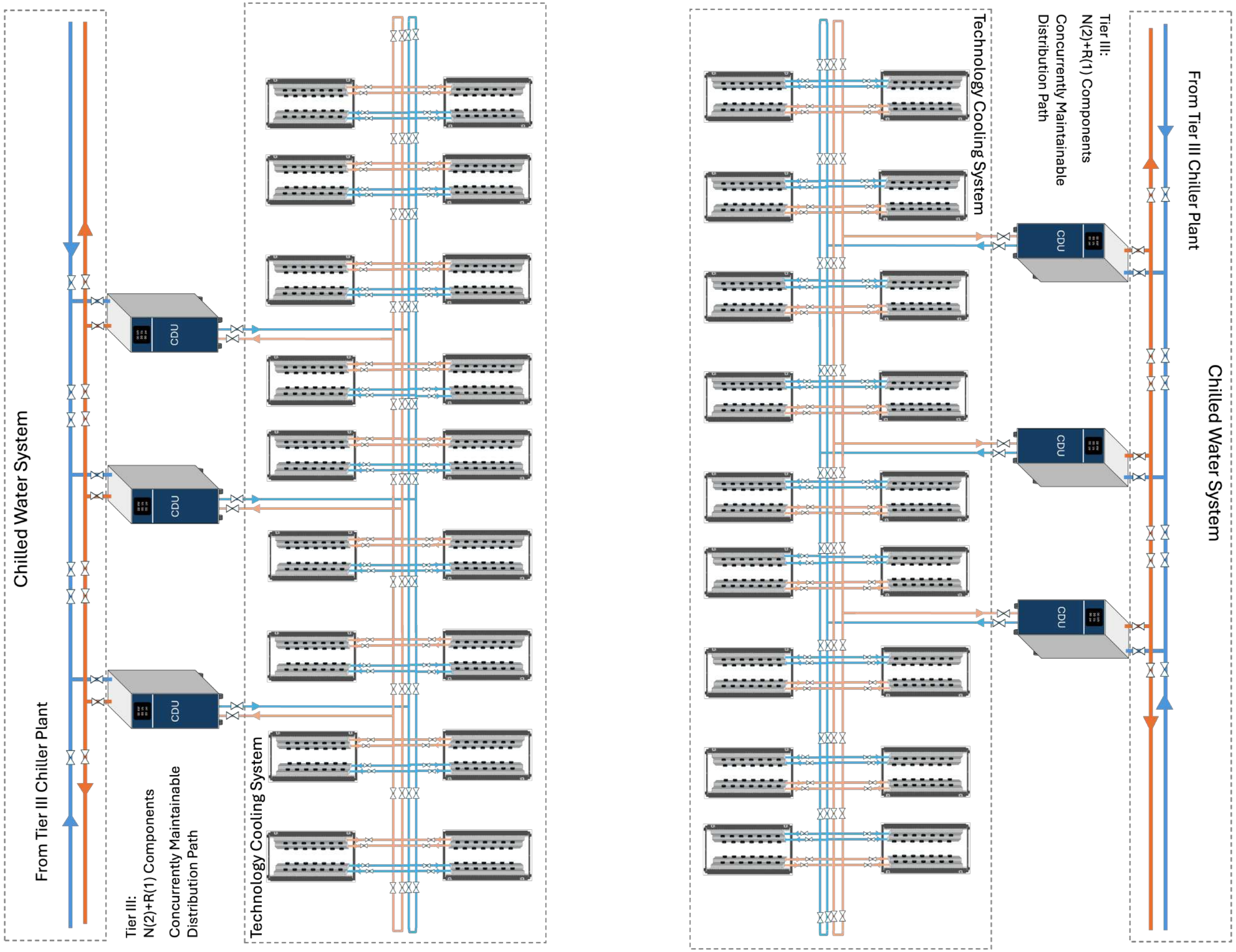
IT Racks: 36 No.

DLC @ 95% = 4,446

kW

DLC @ 55% = 2,574

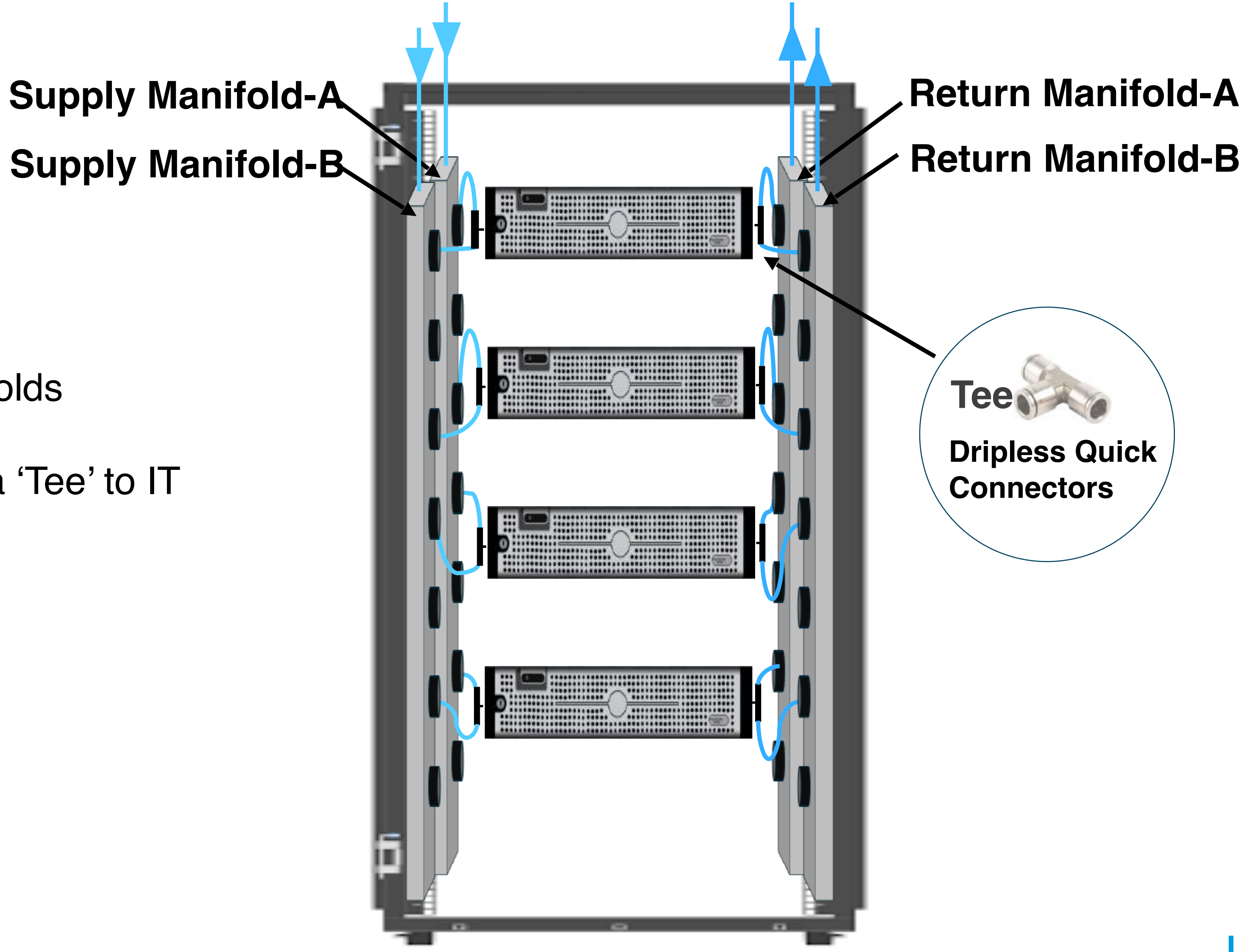
kW



uptime INSTITUTE

Rack Cooling Design (N+1, 5 MW, Tier III, CW, 130 kW/Rack)

- ✓ Rack Power Density: 130 KW
- ✓ Redundant Set of Rack Manifolds
- ✓ Dripless Quick Connectors via ‘Tee’ to IT Assets



uptime INSTITUTE

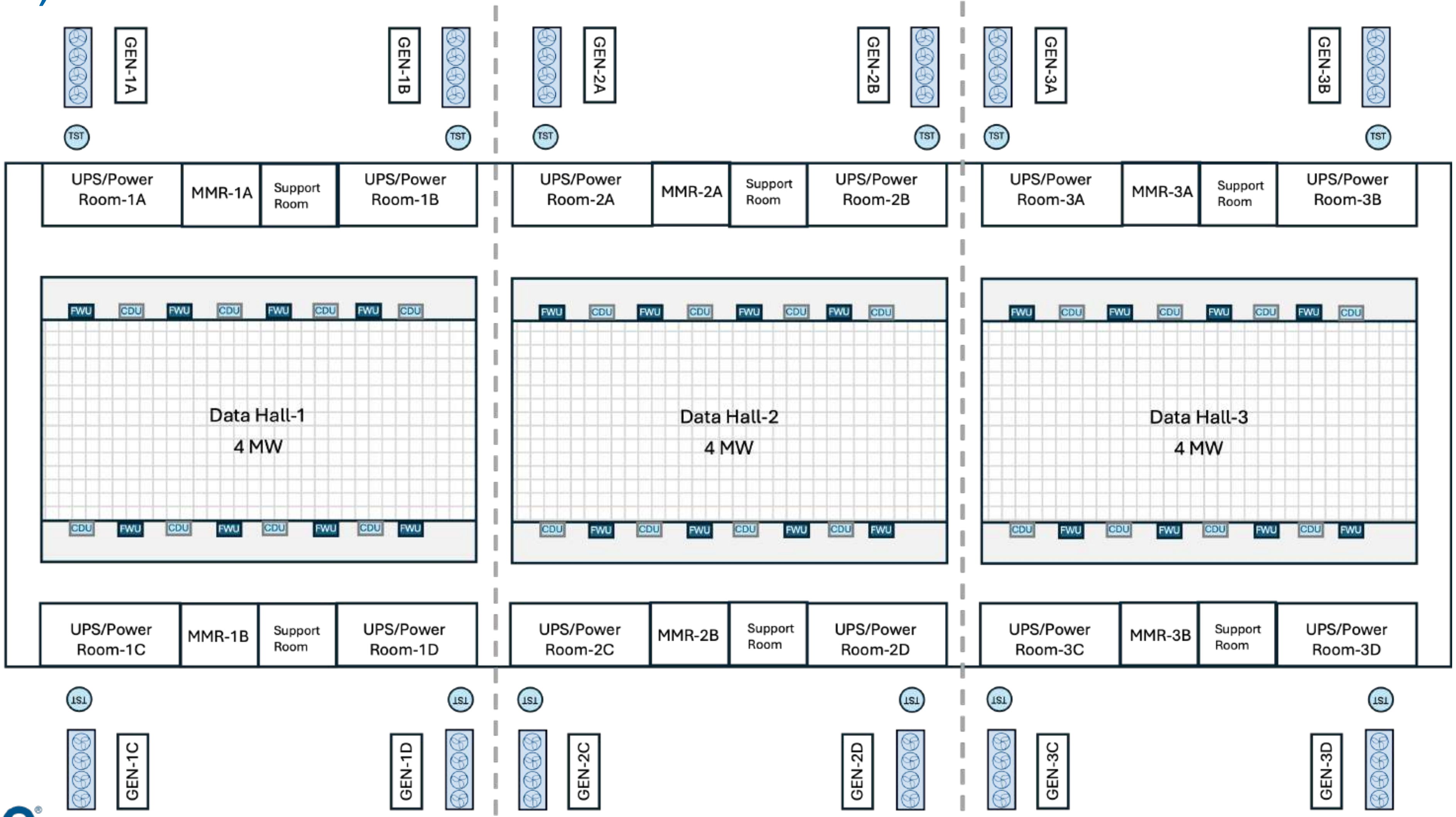
Reference Design 48 MW (Tier IV, CW, 130 kW/ Rack)

12 MW 12 MW 12 MW 12 MW

uptime INSTITUTE

Reference Design 48 MW (Phase-1, 12 MW, Tier IV, CW, 130 kW/Rack)

12 MW

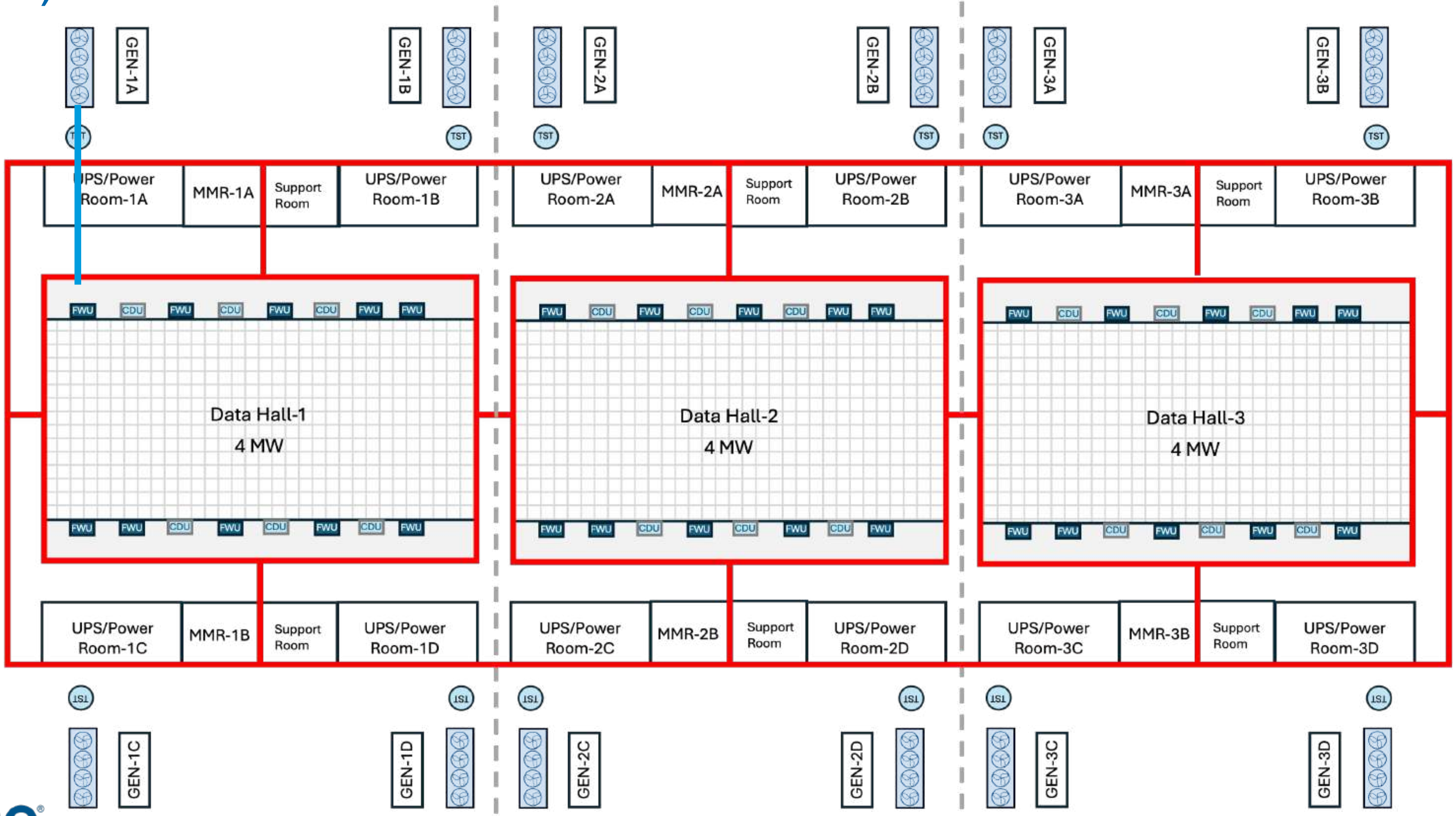


uptime INSTITUTE

Reference Design 48 MW (Phase-1, 12 MW, Tier IV, CW, 130 kW/Rack)

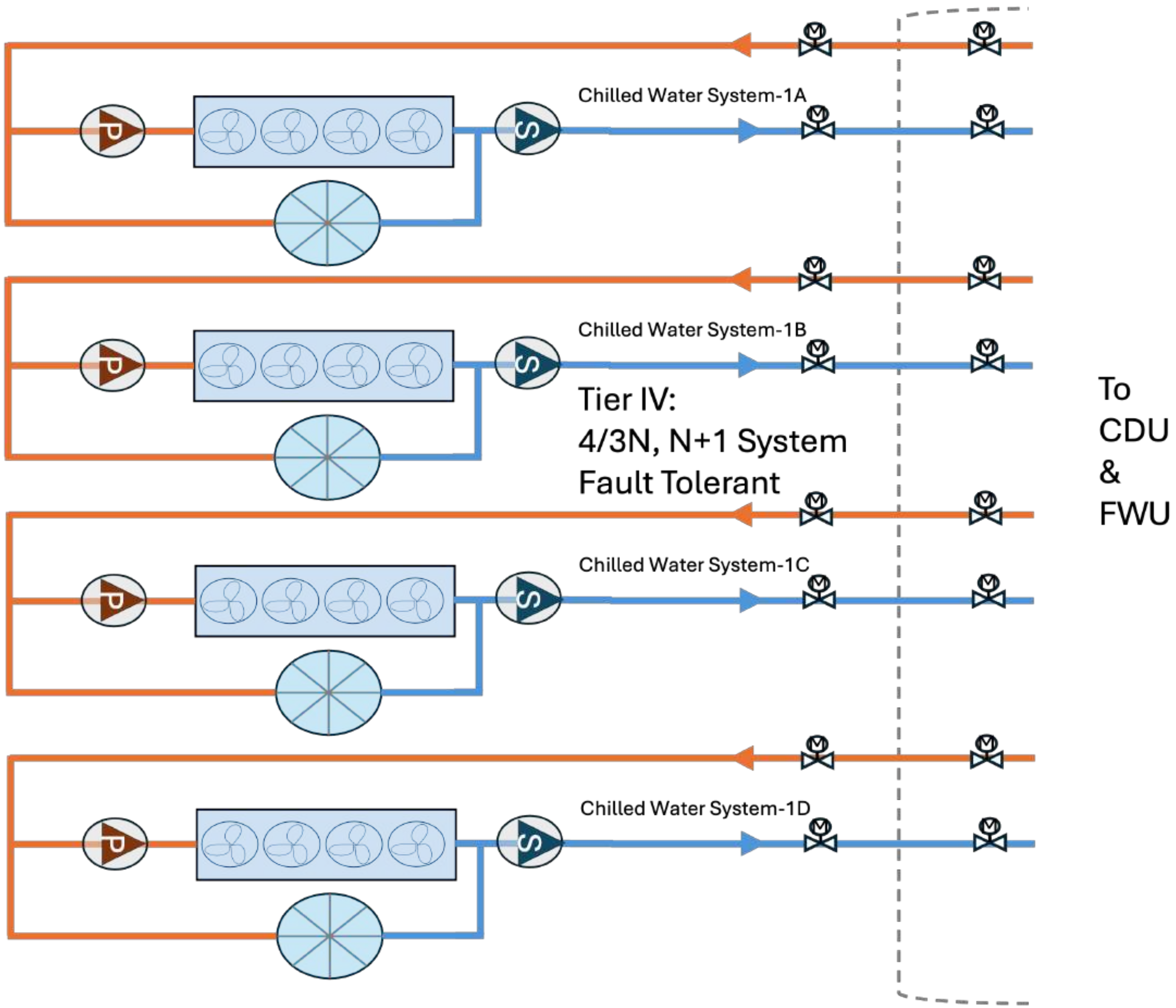
Compartmentalization

12 MW



uptime INSTITUTE

Chiller Plant Design (N+1, 4 MW, Tier IV, CW, 130 kW/Rack)



Technology Cooling Design (N+R, 4 MW, Tier IV, CW, 130 kW/ Rack)

CDU: 750 kW

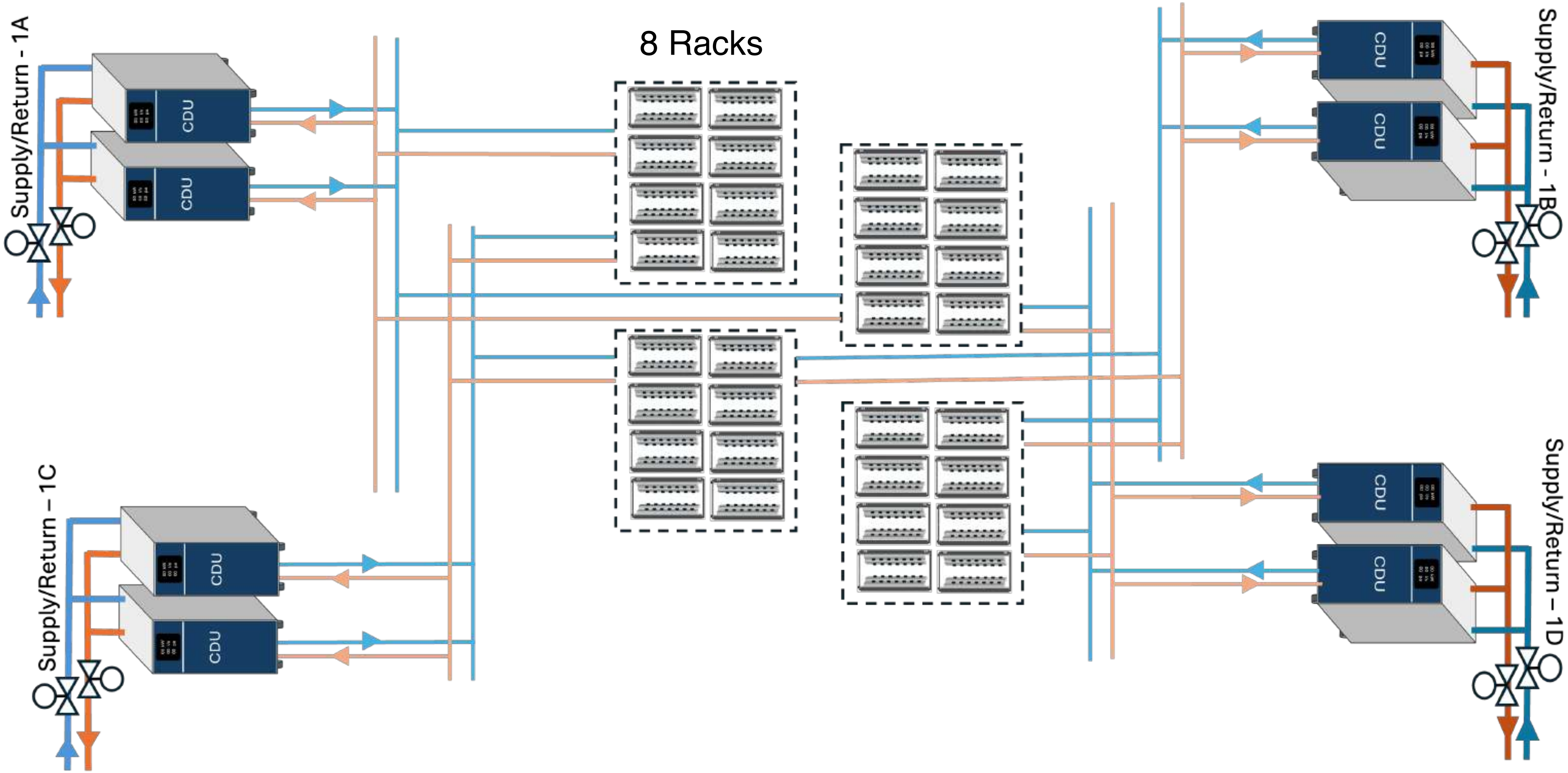
N+R, N=6 (4,500 kW)

R=2 (1500 KW)

IT Racks: 32 No.

DLC @ 95% = 3,800 kW

DLC @ 55% = 2,200 kW



uptime INSTITUTE

Technology Cooling Design (N+R, 4 MW, Tier IV, CW, 130 kW/ Rack)

CDU: 750 kW

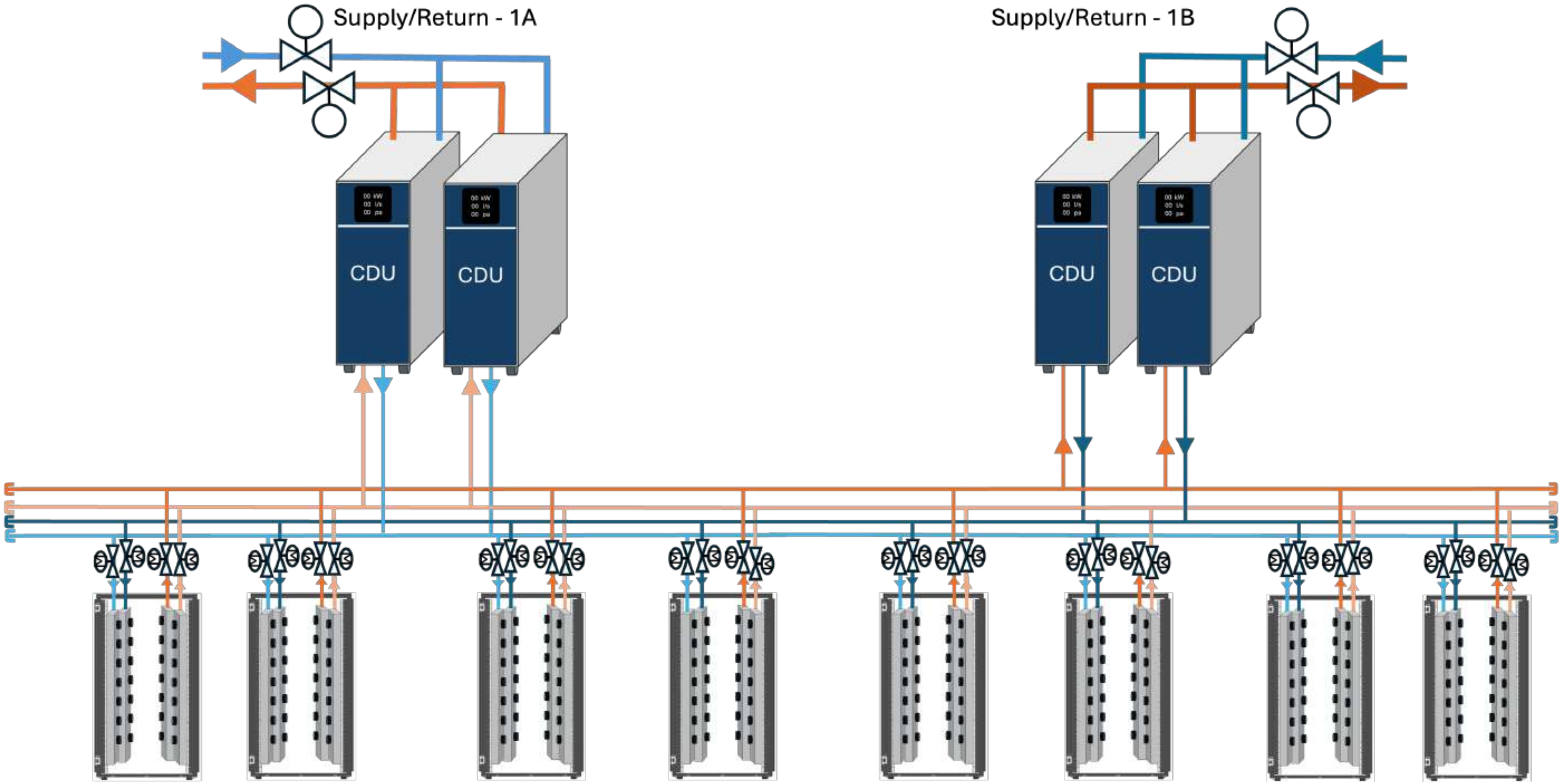
N+R, N=6 (4,500 kW)

R=2 (1500 KW)

IT Racks: 32 No.

DLC @ 95% = 3,800 kW

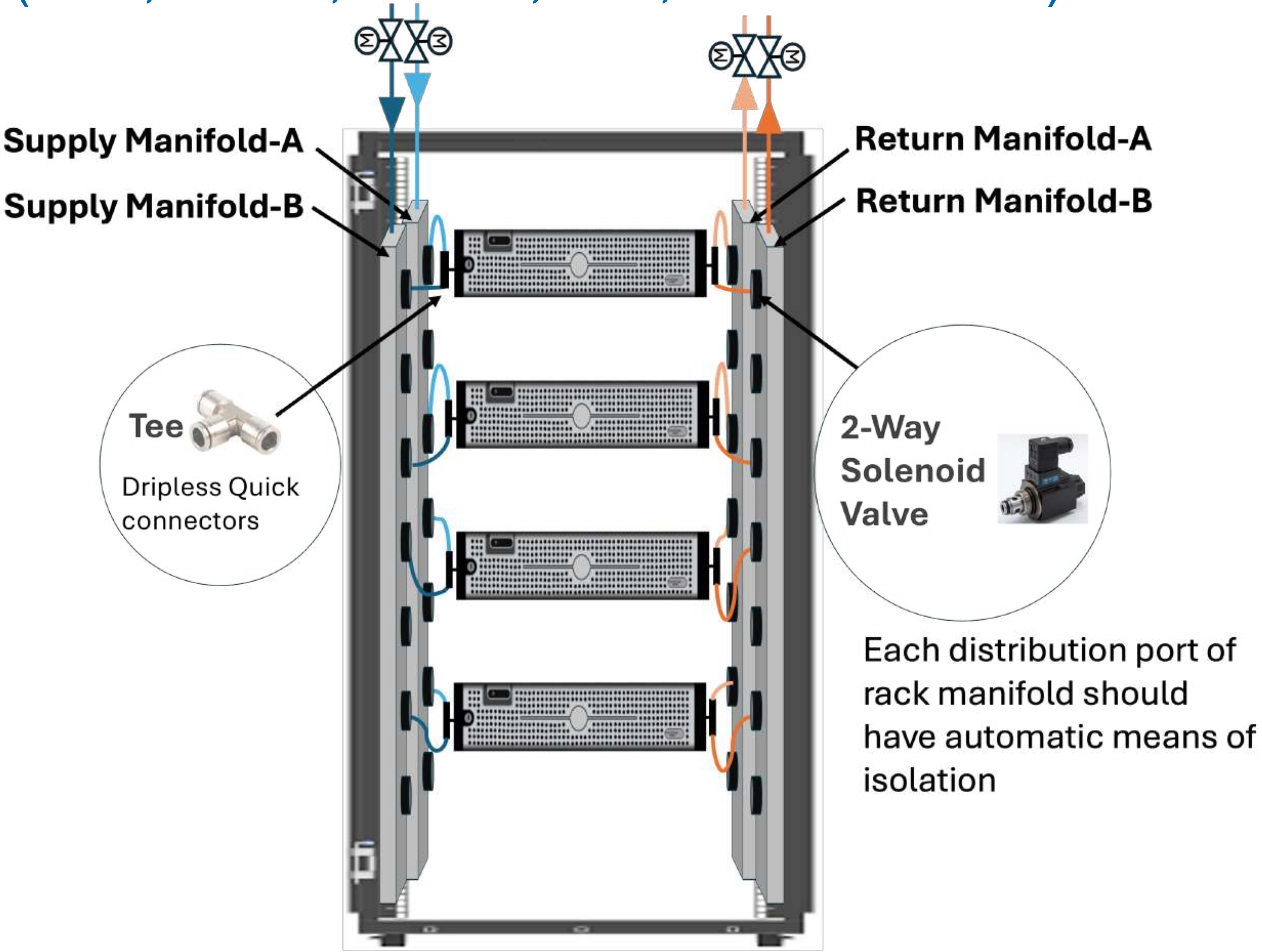
DLC @ 55% = 2,200 kW



uptime INSTITUTE

Rack Cooling Design (N+1, 4 MW, Tier IV, CW, 130 kW/Rack)

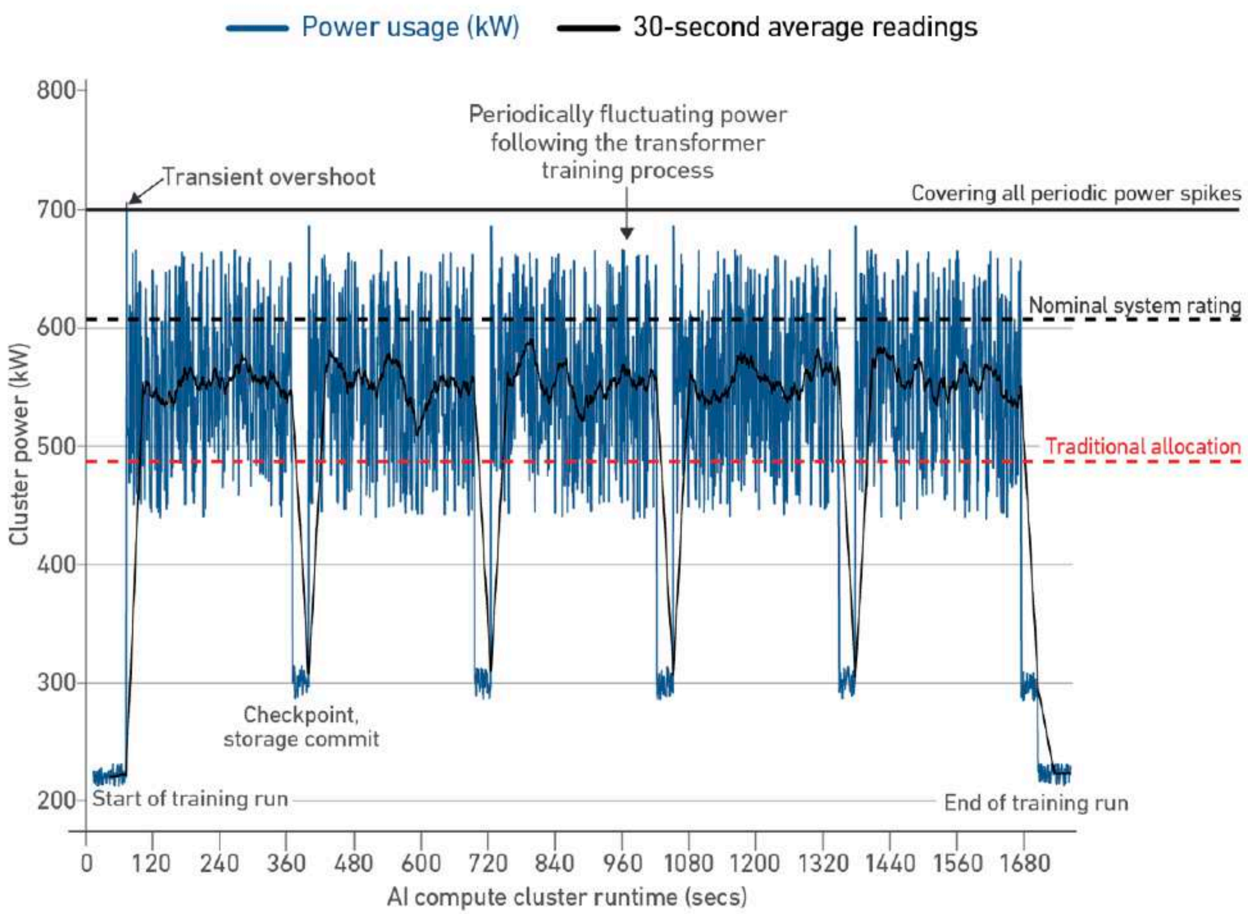
- ✓ Rack Power Density: 130 KW
- ✓ Redundant Set of Rack Manifolds
- ✓ Dripless Quick Connectors via 'Tee' to IT Assets
- ✓ Solenoid valves for autonomous isolation
- ✓ Leak detection
- ✓ Continuous Cooling



uptime INSTITUTE

AI Loads Power Profile

Figure 1 Power profile of a GPU-based training cluster (algorithmic not real-world data)



UPTIME INSTITUTE 2025

https://intelligence.uptimeinstitute.com/resource/electrical-considerations-large-ai-compute

Commissioning

Level 1: Factory Witness Testing

Performance of equipment at the factory before it is delivered to the site. This includes testing critical components such as generators, uninterruptible power supplies (UPS), chillers, and other major systems to ensure they meet the required specifications.

Level 2: Site Acceptance Testing (SAT)

Site Acceptance Testing to ensure all equipment has been shipped and installed correctly.

Level 3: Pre-Functional Testing (PFT)

Pre-functional testing to verifying that all individual systems (e.g., electrical, mechanical, and plumbing) are installed and functioning correctly on their own.

Level 4: Functional Performance Testing (FPT)

Integrated operation of all systems to ensure they work together as designed.

Level 5: Integrated System Testing (IST)

Integrated System Testing is the final phase of commissioning, where we validate the complete data center infrastructure under real-world conditions.

AI Simulators:

AI factories require specialized load banks to mimic actual power and cooling profile of the AI work loads to perform integrated system testing before handover.

Operations

Cooling System Management

- Coolant Quality and Maintenance
- Leak Detection and Containment
- Pumping and Circulation Monitoring

Environmental Control and Safety

- Temperature and Humidity Monitoring
- Safety Protocols
- Emergency Response

Operational Efficiency and Optimization

- Energy Management
- Heat Rejection Management
- Capacity Planning and Scalability

Maintenance and Lifecycle Management

- Routine Maintenance
- Component Longevity

Data Center Capacity Management

- Flexible Layouts
- Weight Management

Compliance and Documentation

- Regulatory Compliance
- Operational Documentation

Training and Staffing

- Specialized Training
- Technical Expertise

Training:

Understanding of specific design and equipment requirements in conjunction with system operations.

Visit uptimeinstitute.com for more information.

Uptime Institute is the Global Digital Infrastructure Authority. Uptime Institute is headquartered in New York, NY, with offices in London, Sao Paulo, Dubai, Riyadh, Singapore, and Taipei.

Uptime Institute
405 Lexington Avenue
New York, NY 10174

uptime INSTITUTE