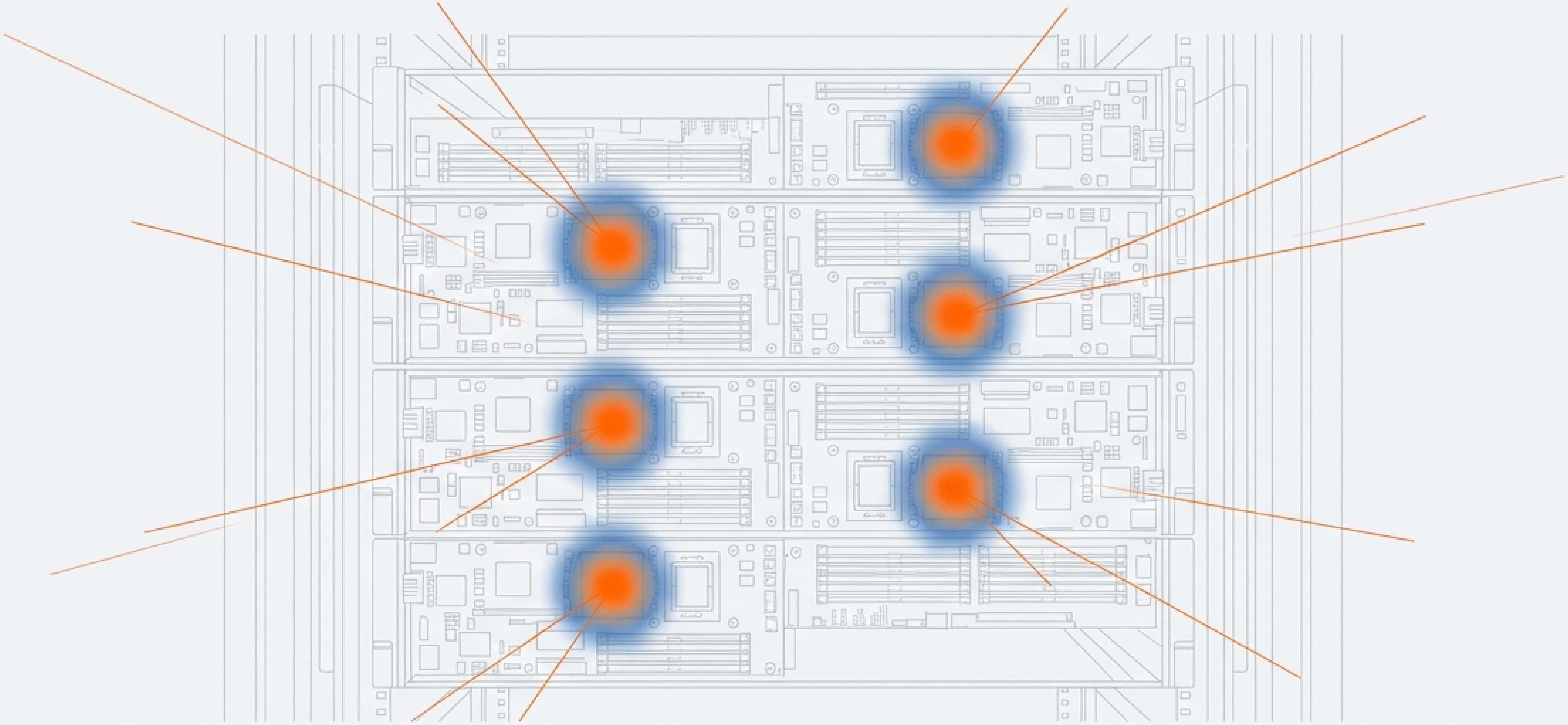
The Unrelenting Rise of Heat: Navigating the New Thermal Frontier in Data Centers

An executive briefing on the technologies, trade-offs, and future of data center cooling in the AI era.



Synthesized from analysis by JLL, Uptime Institute, ASHRAE, Data Center Frontier, and Equinix.

AI is Pushing Rack Densities Beyond the Limits of Traditional Air Cooling

50-100+ kW

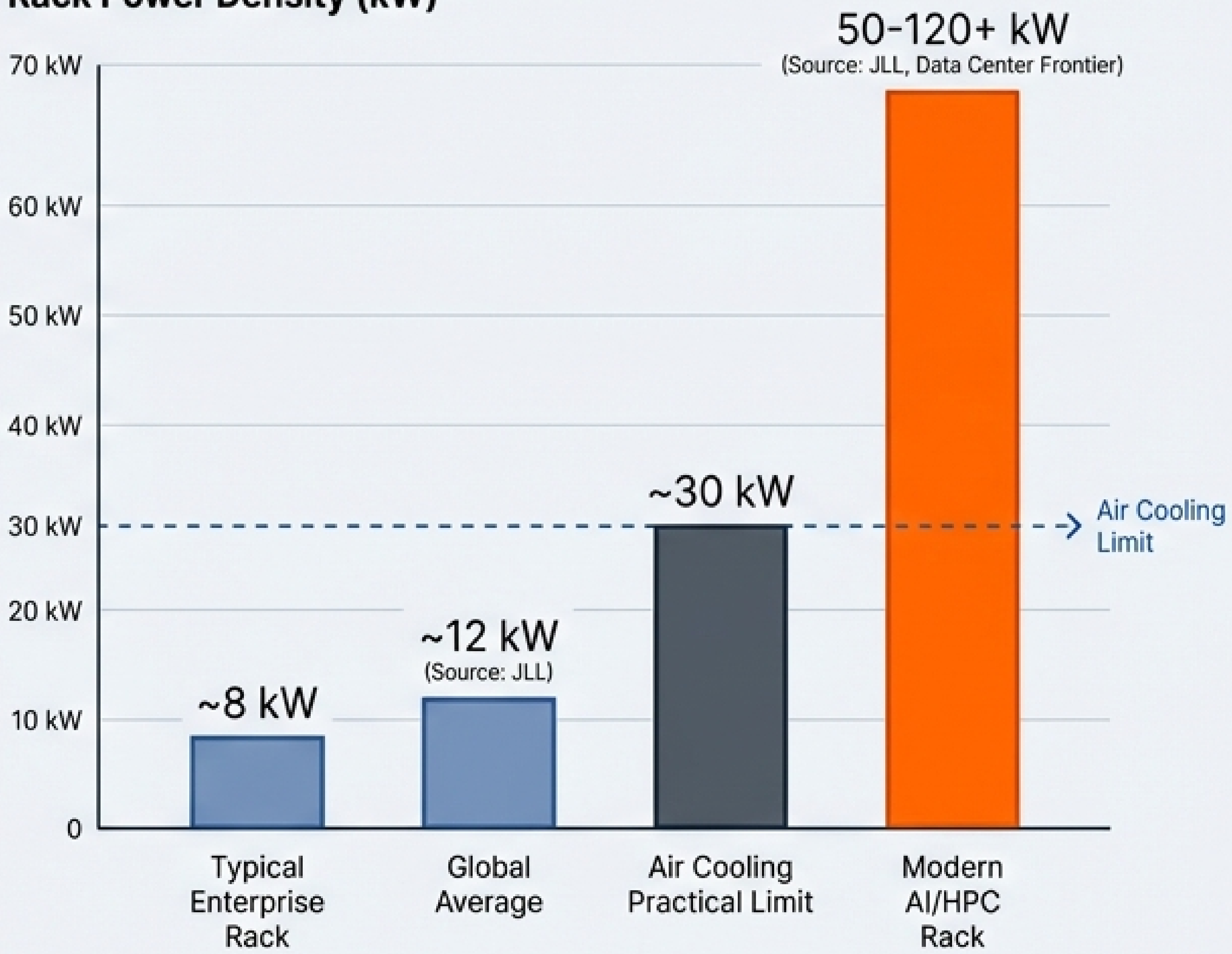
Rack power density for cutting-edge AI/HPC deployments, shattering the global average of ~12 kW. (Source: JLL)

The AI Accelerator: AI and HPC workloads are driving unprecedented computational density, concentrating heat generation in smaller footprints.

The Physical Limit: Most sources agree that conventional air cooling is "tapped out" and becomes impractical or inefficient beyond ~30 kW per rack. (Source: Data Center Frontier, Vertiv)

The Consequence: Without a paradigm shift in thermal management, performance throttling, component damage, and operational failure are inevitable.

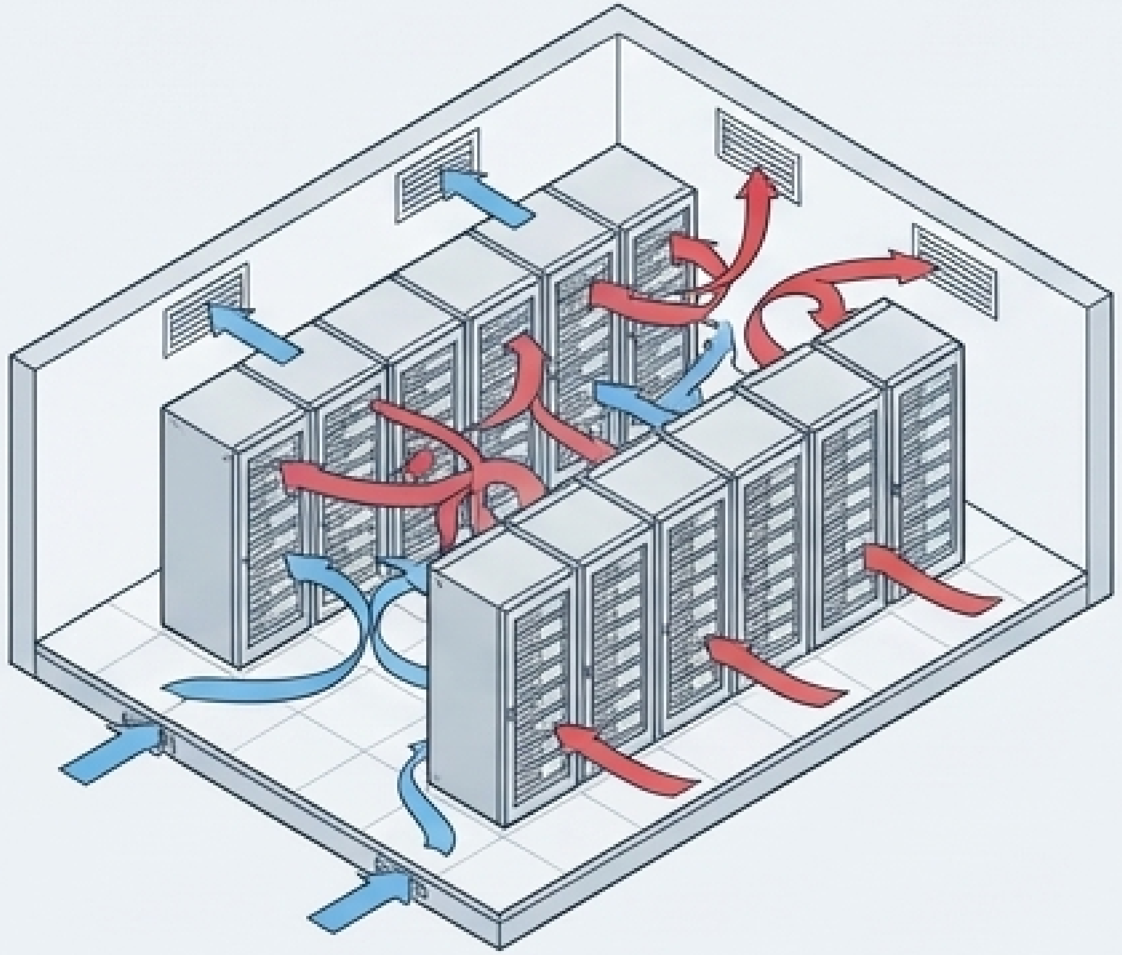
Rack Power Density (kW)



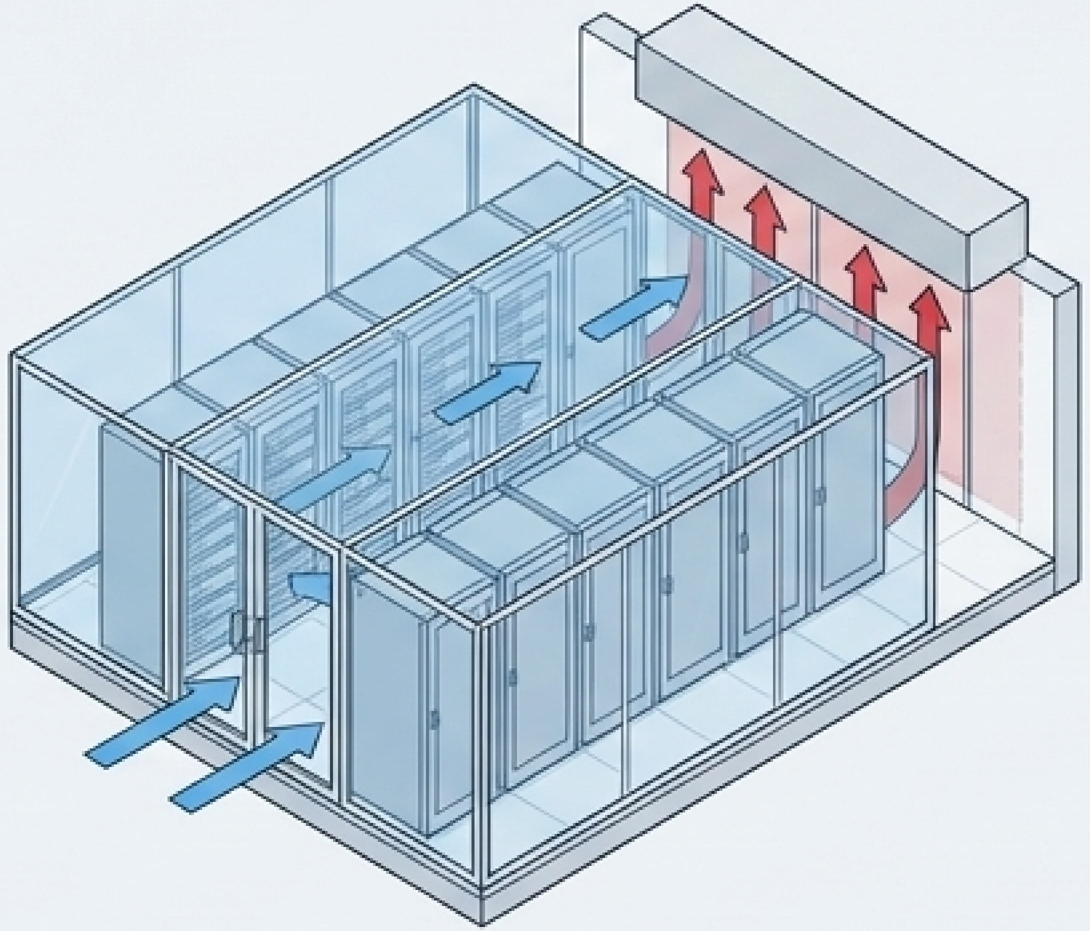
Mastering Airflow is the Non-Negotiable Foundation of Thermal Efficiency

The primary goal is to prevent the recirculation of hot exhaust air into the cold server intakes. Hot aisle/cold aisle containment is the industry-standard best practice.

Before Containment
Chaotic Air Mixing



After Containment
Clean Separation

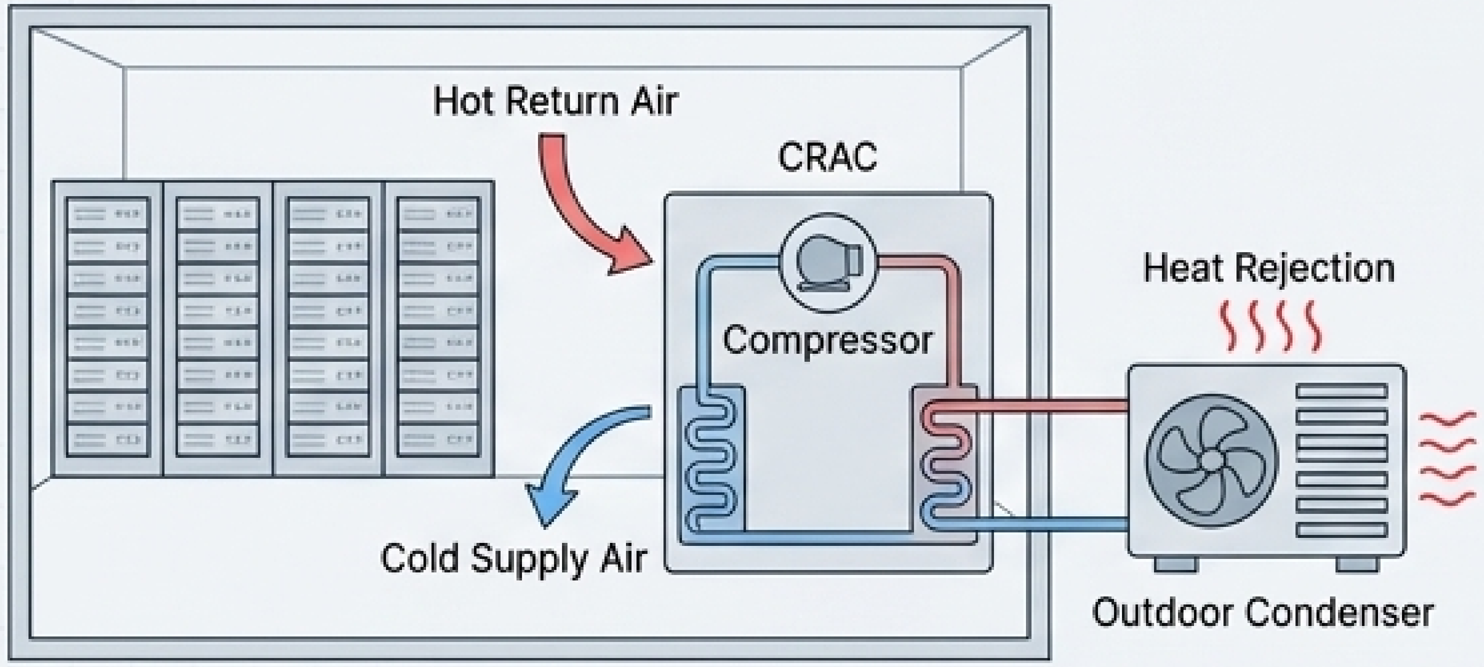


- **10-35%** reduction in cooling energy consumption by eliminating hot/cold air mixing. (Source: DOE/Energy Star)
- **20-25%** savings in fan energy by using variable speed fans that no longer have to fight recirculation. (Source: DOE)
- **Enables** higher, more efficient setpoint temperatures for chilled water or AC units.
- Underscoring its importance, roughly **two-thirds of large data centers** have already implemented hot/cold aisle arrangements. (Source: Energy Star)

NotebookLM

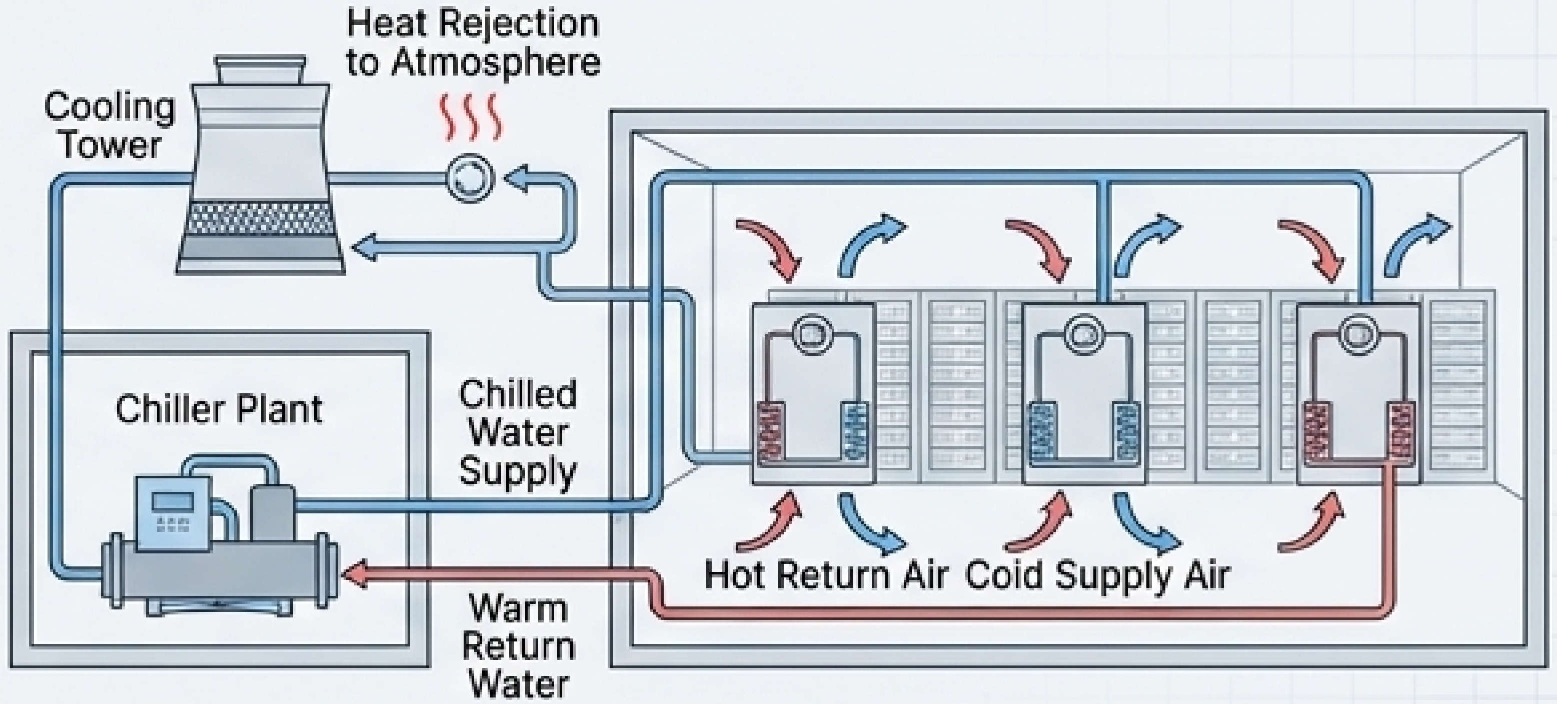
The Air Cooling Toolkit Ranges from Self-Contained Units to Centralized Chiller Plants

CRAC (Computer Room Air Conditioner)



- Direct-Expansion (DX) refrigerant cycle.
- Self-contained and simple to deploy.
- **Best for:** Small to mid-sized facilities, edge deployments, or locations with water scarcity. (Source: ASHRAE, Park Place)

CRAH (Computer Room Air Handler) with Chiller Plant

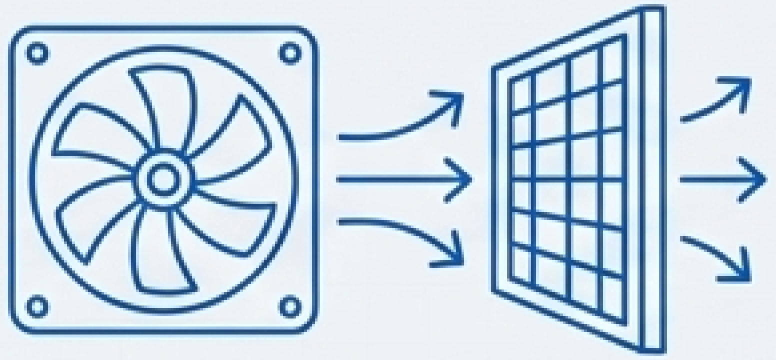


- Uses chilled water/glycol as the cooling medium.
- Higher upfront cost but superior efficiency and capacity at scale.
- **Best for:** Large enterprise and hyperscale data centers. (Source: DOE, Park Place)

NotebookLM

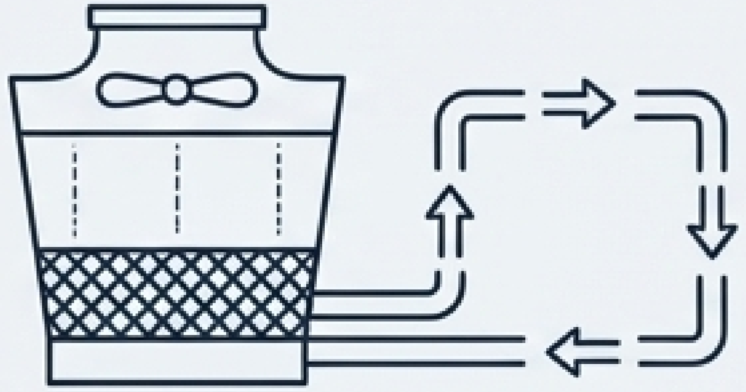
Economization Leverages Local Climate to Drastically Reduce Mechanical Cooling Energy

“**Free cooling**” uses favorable outside air or water temperatures to cool the facility, allowing energy-intensive chiller compressors to be turned off for thousands of hours per year.



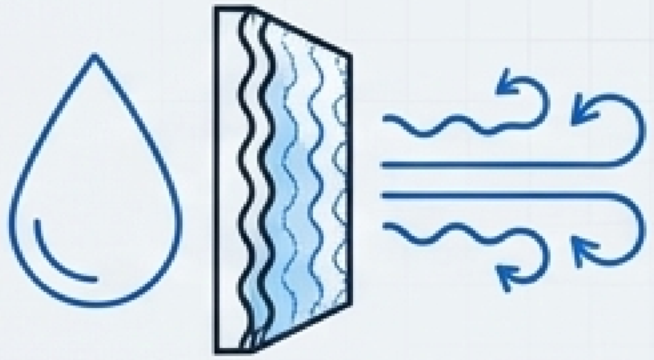
1. Air-Side Economization

Directly brings filtered, cool outside air into the data hall. A common strategy for hyperscalers like Meta in temperate climates.



2. Water-Side Economization

Uses a cooling tower or dry cooler to produce chilled water without running a chiller compressor when outside wet-bulb temperatures are low.



3. Adiabatic/Evaporative Assist

Uses water evaporation (like a “swamp cooler") to pre-cool the incoming air, extending the hours of economization. Can be direct (air passes through wetted media) or indirect (uses a heat exchanger to avoid contamination). (Source: Data Center Frontier)

ASHRAE’s expanded thermal envelopes (Classes A1-A4) were created to facilitate more economizer hours by allowing IT equipment to safely operate at **higher inlet temperatures** (e.g., up to **45°C / 113°F** for A4). (Source: ASHRAE)

Optimizing Cooling is a Balancing Act Between Energy and Water Consumption

The Plateauing Metric (PUE)

Power Usage Effectiveness (PUE) = Total Facility Energy / IT Equipment Energy. A perfect score is 1.0.

After years of improvement, the industry average PUE has plateaued at **~1.58**, meaning cooling and overhead still consume ~58% of the power used by the IT load.
(Source: Uptime Institute, 2023)

The Ascendant Metric (WUE)

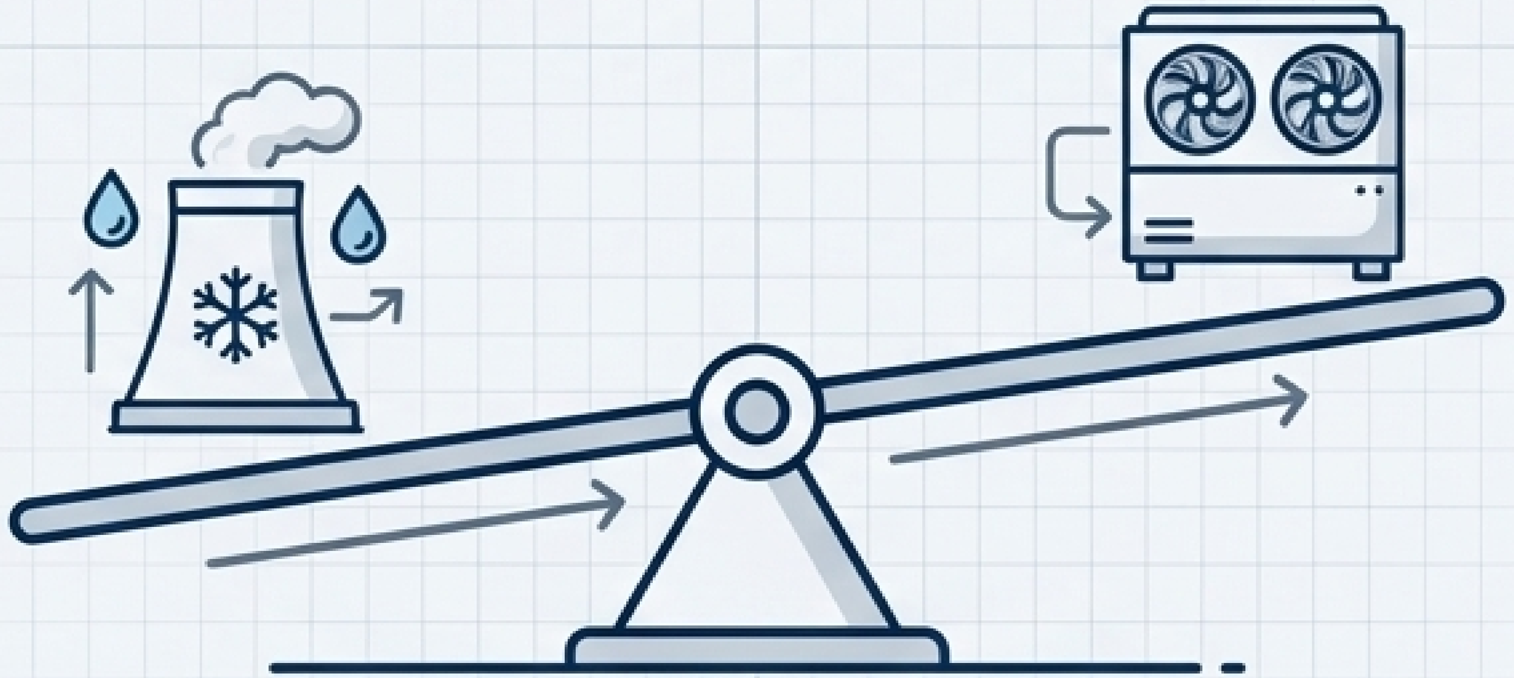
Water Usage Effectiveness (WUE) = Annual Water Usage / IT Equipment Energy (Liters/kWh). A score of 0 is best.

Introduced to quantify the water footprint of cooling, especially with the rise of evaporative methods.

Low PUE
(Good Energy Efficiency)

Aggressive Evaporative Cooling

Low energy use (good PUE), but high water consumption. **WUE can be up to 2.5 L/kWh**.
(Source: Equinix)



Low WUE
(Good Water Efficiency)

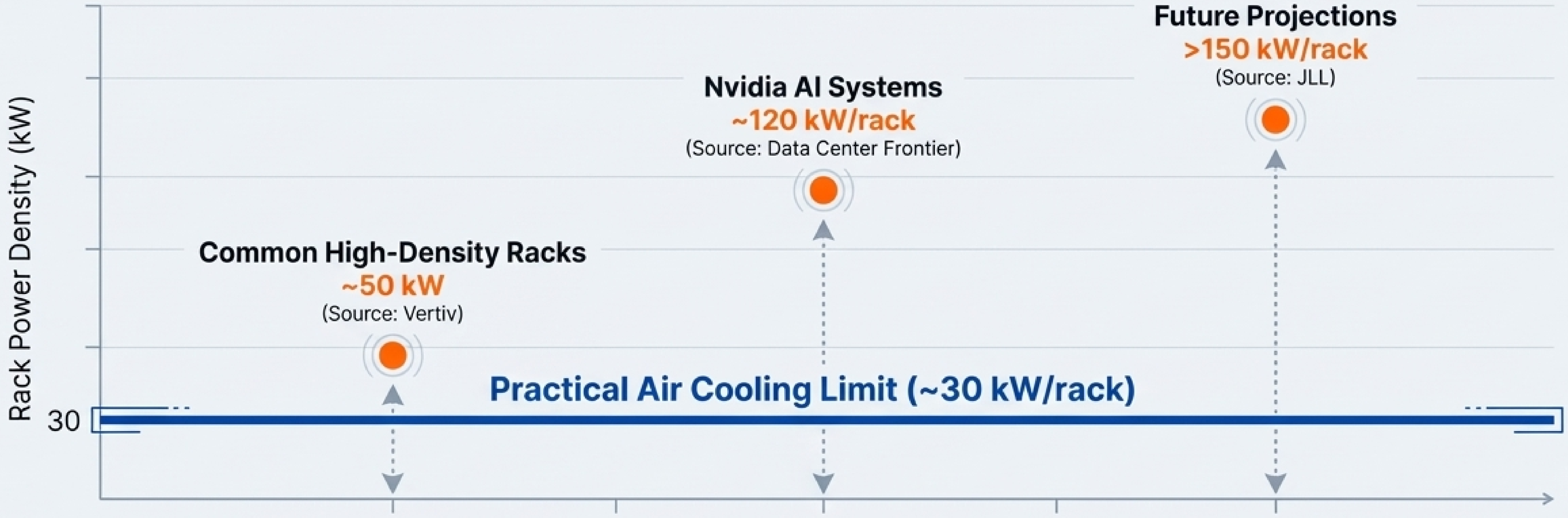
Air-Cooled Chillers (no evaporation)

Zero water use (**WUE = 0**), but higher energy consumption (worse PUE).
(Source: Equinix)

A low PUE can be misleading if it comes at the cost of unsustainable water use in a water-stressed region. Operators must now monitor and report both metrics.

The Inflection Point: Liquid Cooling is Now an Imperative for AI Infrastructure

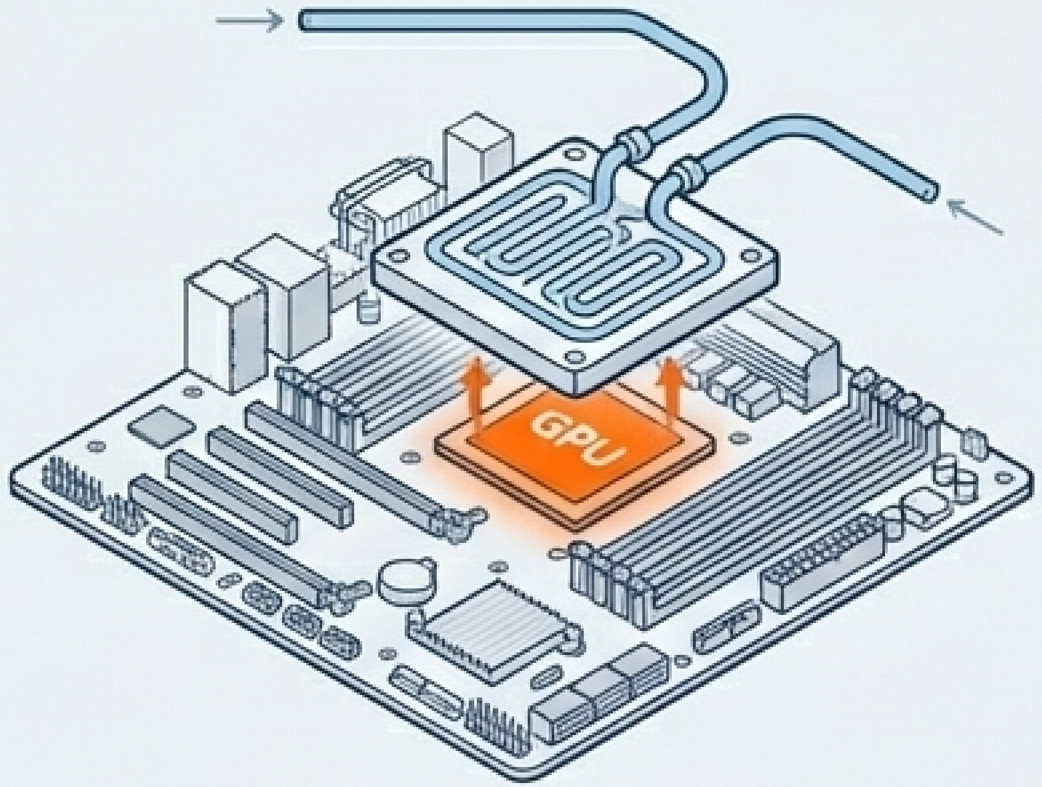
Liquids like water or dielectric fluid have thousands of times the heat capacity of air, enabling far more efficient heat removal directly at the source. (Source: Vertiv)



> “Liquid cooling is now seen as the **‘only viable path’** to support future AI/HPC loads at scale.” (Paraphrased consensus from sources)

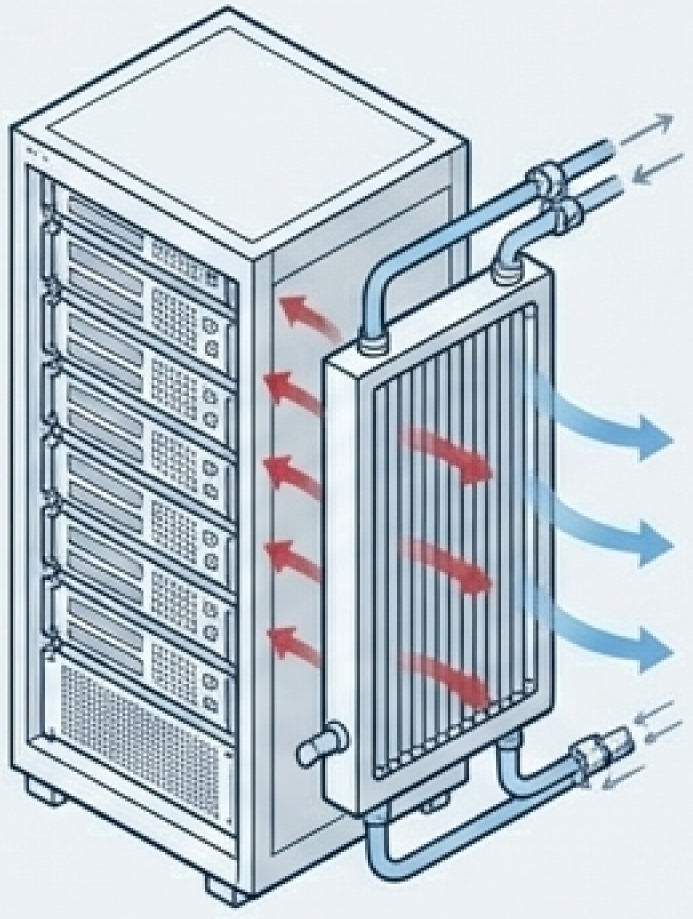
The Liquid Cooling Portfolio Offers a Spectrum of Solutions from Hybrid to Full Immersion

Direct-to-Chip (DLC) / Cold Plates



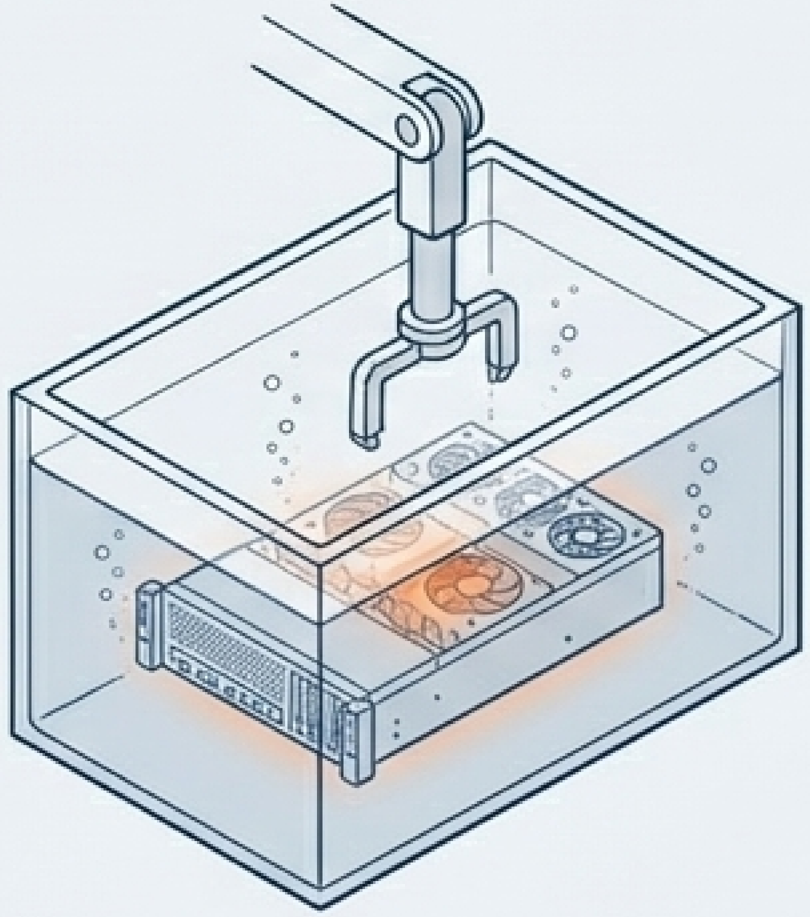
A pumped liquid flows through cold plates attached to key components. Highly efficient, capturing **~70-80% of the server heat** at the source. The remaining ~20-30% is handled by slower, quieter fans. Compatible with traditional rack form factors.

Rear-Door Heat Exchanger (RDHx)



A passive or active ‘radiator door’ that intercepts hot exhaust air before it enters the room. A common and effective retrofit solution for boosting the density of existing air-cooled racks.

Immersion Cooling



Servers are fully submerged in fluid, which directly absorbs 100% of the heat. Allows for the removal of all server fans, maximizing density and efficiency. Projected to become common for racks exceeding **150 kW**. (Source: JLL)

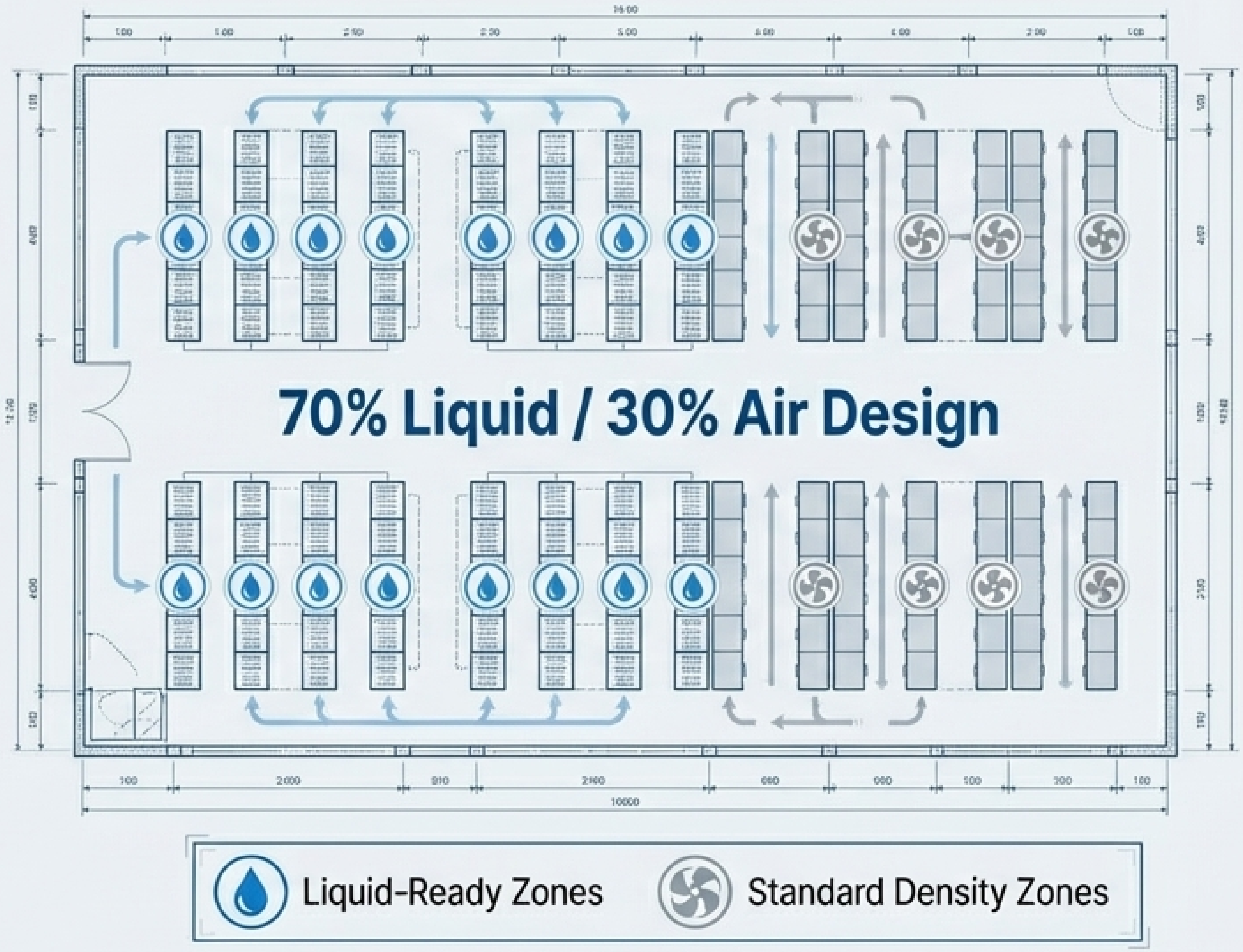
Liquid Is the New Default for High-Density Deployments, Not a Future Option

Market Evidence

- New AI-oriented data center builds are **now defaulting to liquid cooling infrastructure**. (Source: JLL, 2025 Global Outlook)
- A common strategy in new deployments is a **70% liquid / 30% air** cooling mix, designed for flexibility and future growth. (Source: JLL)

Standards Evolution

- ASHRAE has introduced a new environmental class, **Class H1**, specifically for high-density, liquid-cooled IT equipment.
- Unlike other classes that allow wide temperature ranges, H1 recommends a narrow air range of **18–22°C** for the minimal remaining air cooling, signaling an environment optimized for liquid's primary role. (Source: ASHRAE)



Cooling Strategy is Fundamentally Dictated by Local Climate and Resources

Hot & Dry (e.g., Phoenix, US)

Strategy: Heavy reliance on direct evaporative cooling.

Outcome: Excellent PUE, but **extremely high water usage** (WUE). Some data centers consume **millions of gallons of water per day**, facing scrutiny from local authorities. (Source: ASCE)

Hot & Humid (e.g., Southeast Asia)

Strategy: Primarily high-efficiency, closed-loop chiller plants. Air-side economization is impractical due to high ambient enthalpy.

Outcome: Higher PUE due to year-round mechanical cooling load. Water is used in cooling towers but not for direct evaporation.

Temperate (e.g., Northern Europe)

Strategy: The "sweet spot" for free cooling. Extensive use of air-side and water-side economization for 8-10 months per year.

Outcome: Achieves the **lowest PUEs** in the industry, often in the **1.1—1.3 range**.

Cold (e.g., The Nordics)

Strategy: Year-round free cooling is possible. The main challenge shifts from heat rejection to heat reuse.

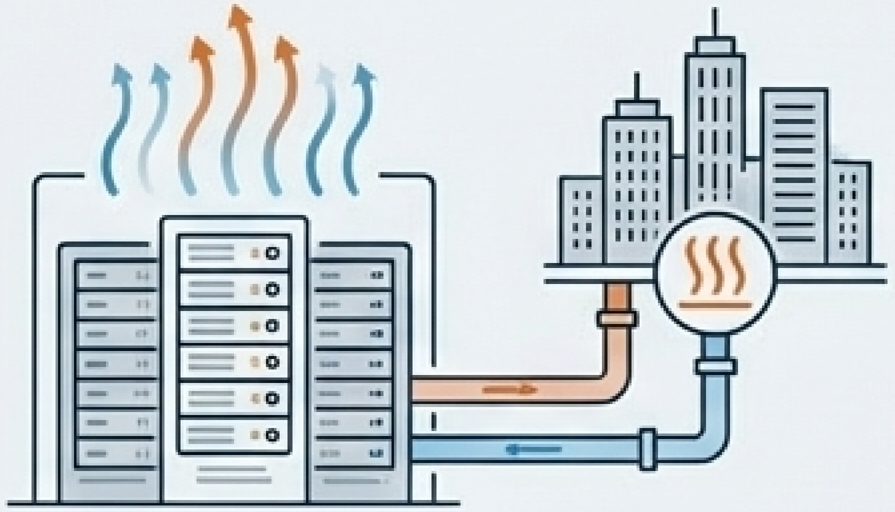
Outcome: Ultra-low PUE and a unique opportunity to integrate with district heating systems.

Special Consideration

High-altitude locations require **derating inlet temperatures (~1°C per 300m above 900m)** to compensate for thinner air's reduced cooling capacity. (Source: ASHRAE)

True Sustainability Involves a Holistic Approach Beyond PUE and WUE

1. Waste Heat Reuse

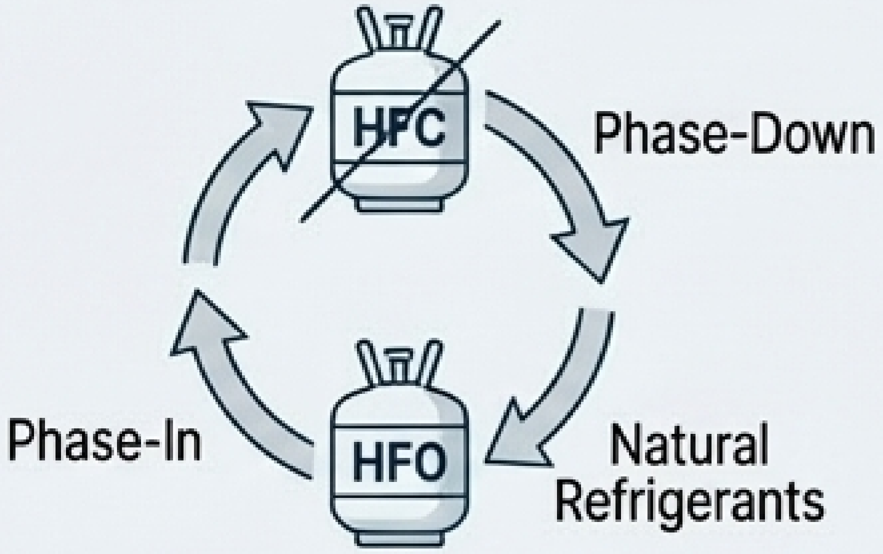


Concept
Transforming waste server heat from a liability into a valuable community resource.

Northern Europe leads with ~60 data centers feeding district heating grids. Microsoft's projects in Finland will **heat ~250,000 residents' homes**.
(Source: Uptime Institute, Bloomberg)

Often requires heat pumps to boost low-grade (~30-45°C) water to usable temperatures.

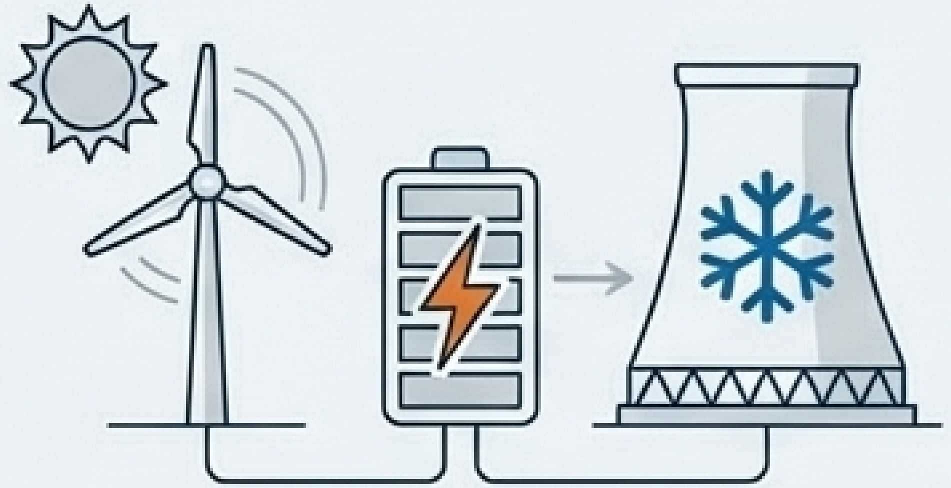
2. Refrigerant Transition



Mandate
Global regulations (e.g., EU F-gas, US AIM Act) are phasing down high-GWP (Global Warming Potential) HFC refrigerants like R-410A.

New systems are moving to lower-GWP alternatives like HFO blends (e.g., R-513A) and natural refrigerants (ammonia, CO_2).

3. Renewable Energy Integration



Concept
Using thermal storage (ice or chilled water tanks) to shift cooling's electrical load.

Application
Chillers run to create 'coolth' when renewable energy (solar/wind) is abundant and cheap, and the stored cooling is used during peak grid demand or when renewables are offline.

(Source: Trane, NREL)

Reliability Remains the Paramount Mandate: Cooling Must Be as Resiltient as Power

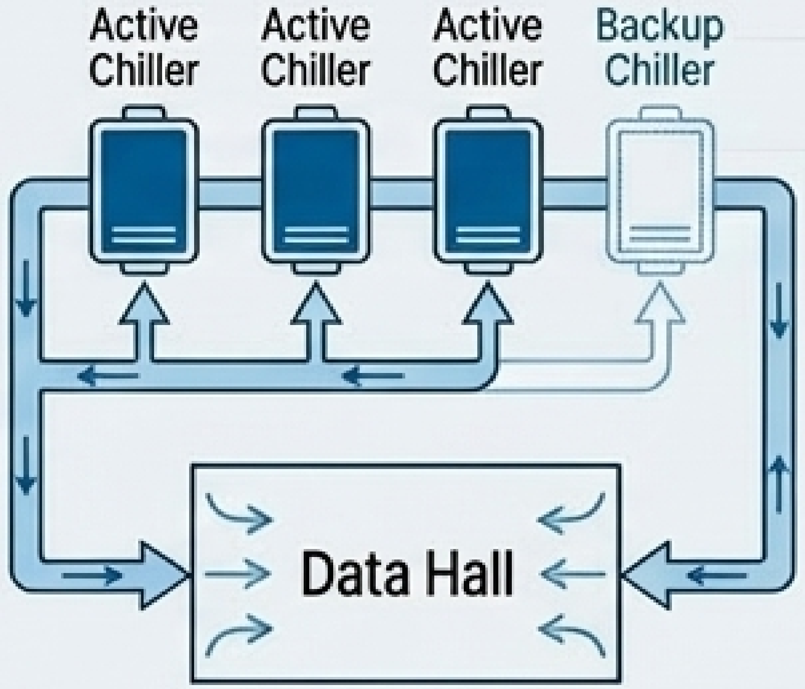
A single cooling failure can cause a catastrophic thermal runaway in minutes. Redundancy is not optional.

Power backup (UPS and generators) must be provisioned for all critical cooling equipment with the same rigor as the IT load.

N+1 Redundancy (Tier III - Concurrently Maintainable)

Having at least one independent backup component. If a facility needs 10 cooling units (N=10), it will have 11 installed.

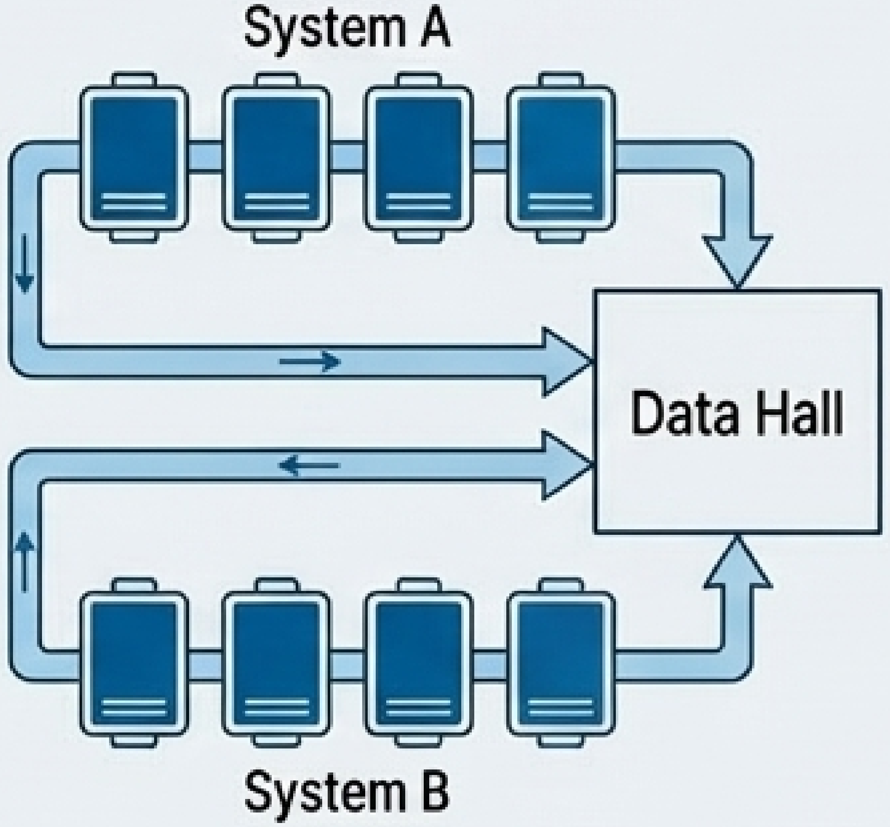
Allows for any single component to be taken offline for maintenance without impacting the IT load. (Source: CoreSite, PhoenixNAP)



2N Redundancy (Tier IV - Fault-Tolerant)

Two completely independent, duplicated cooling systems (“A” side and ”B” side), each capable of handling the full IT load.

Can withstand a major failure of an entire cooling loop without any downtime. (Source: Digital Realty)



AI-Driven Controls are Unlocking a New Level of Dynamic Cooling Efficiency

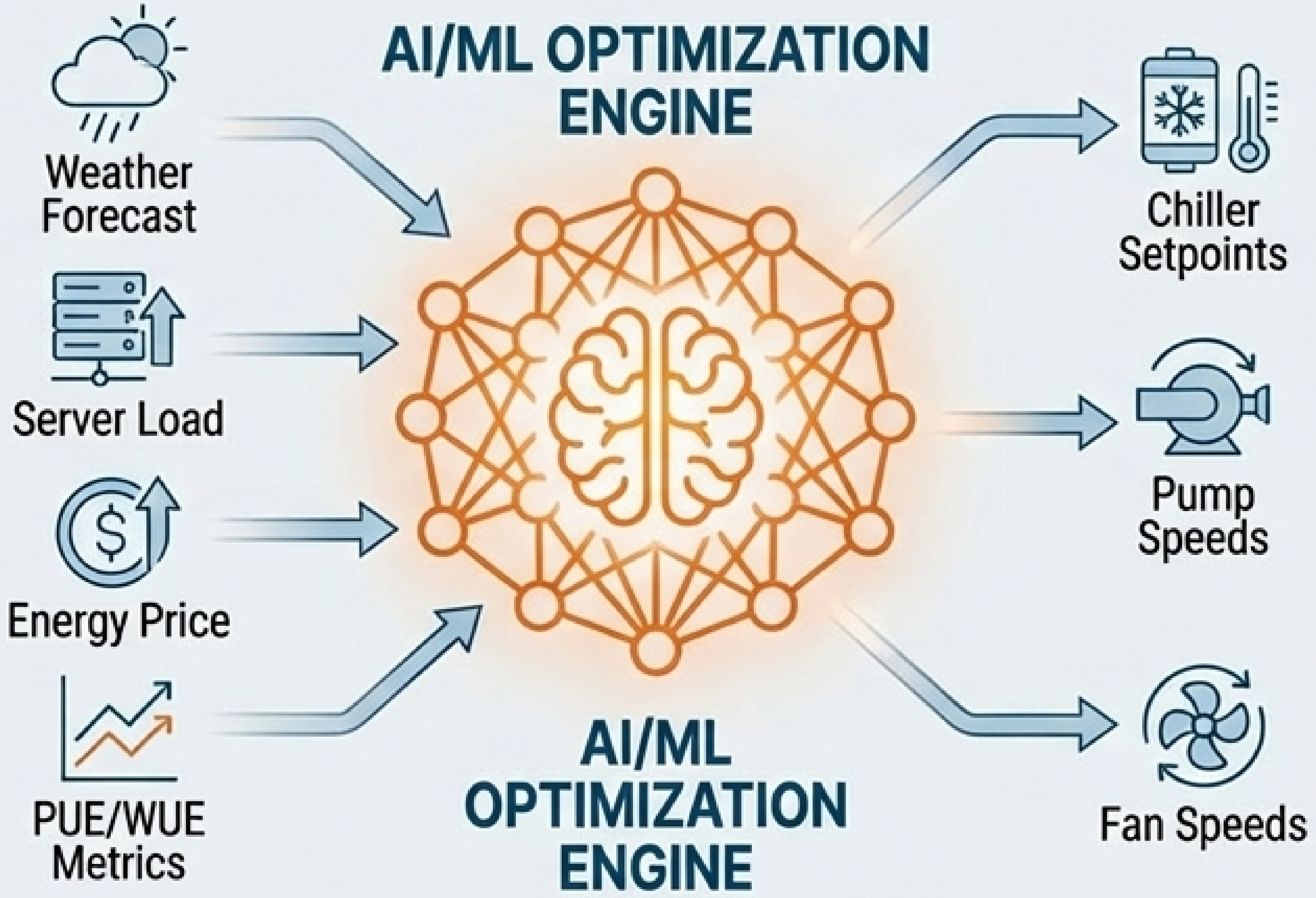
Moving from static, human-set temperature points to **predictive, real-time optimization** that constantly adapts to IT load, weather, and energy prices.

Google DeepMind (2016)

The seminal project that proved the concept. Used machine learning to predict thermal loads and optimize chiller plant operations.

40% reduction in cooling energy usage

15% improvement in overall PUE.
(Source: DeepMind)



Weather Forecast

Server Load

Energy Price

PUE/WUE Metrics

AI/ML OPTIMIZATION ENGINE

Chiller Setpoints

Pump Speeds

Fan Speeds

AI/ML OPTIMIZATION ENGINE

Meta (Facebook) (2024)

The next generation of AI control. Uses simulator-based reinforcement learning (RL) to manage airflow systems.

20% reduction in fan energy

4% savings in water on average across their fleet.

(Source: Meta Engineering Blog)

Modular and Prefabricated Systems Provide the Agility to Scale Cooling On-Demand

The Challenge: AI demand growth is rapid and unpredictable, making traditional, multi-year construction cycles for cooling plants a business risk.

The Solution: Modular Cooling: Self-contained, factory-built cooling units (e.g., skid-mounted chiller plants, containerized liquid cooling loops, evaporative modules) that are delivered and installed on-site.

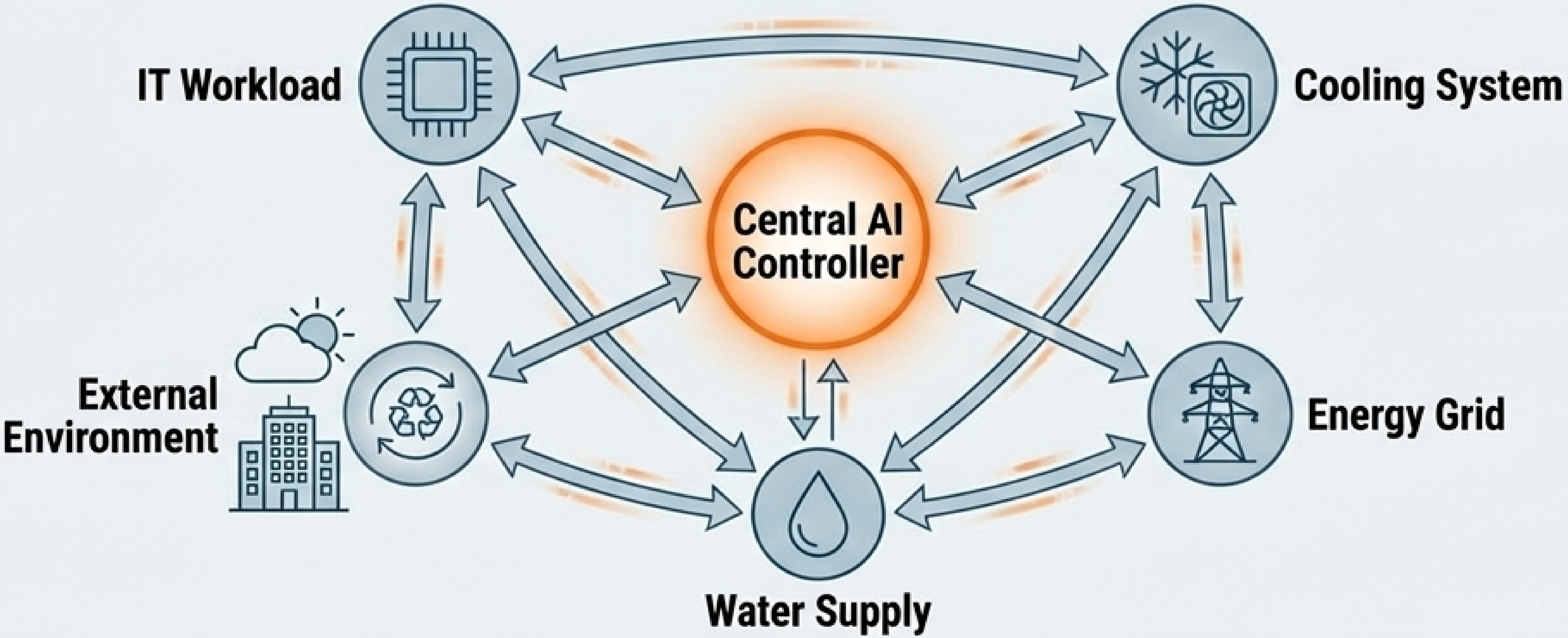
Key Advantages:

- **Speed to Market**: Drastically reduces deployment time from years to months.
- **Scalability**: Allows operators to adopt a 'pay-as-you-grow' model, adding cooling capacity in lockstep with demand.
- **Standardization**: Pre-tested, factory-built quality reduces commissioning risks and ensures predictable performance.

> This approach is highly attractive as **AI demand growth is both rapid and uncertain**, allowing infrastructure to scale dynamically. (Paraphrased consensus from sources like Schneider, Vertiv)

The Future is Integrated: From Thermal Management to Resource Orchestration

- **From**: A siloed function focused solely on rejecting waste heat.
- **To**: An intelligent, integrated system that holistically manages and orchestrates multiple resources.



The ultimate goal is **thermal-aware workload scheduling**, where the data center's control plane can actively shift computational jobs to different servers, racks, or even different facilities based on available cooling capacity, energy cost, and thermal efficiency. Cooling and computing become a single, co-optimized system.