The Power Conundrum: Cooling to the Rescue?

Carrie Goetz

ACCELSIUS™
www.accelsius.com
info@accelsius.com

INTRODUCTION

Power, cooling, and compute are the fundamental trinity of the data center ecosystem. Adding compute resources creates the need for additional power and cooling. Not having enough of either negates the ability to add compute without drastically changing the dynamic. Conversely, it's important to understand that additional cooling will also require additional power; all ships rise with the tide, so to speak.

Power sources are not infinite. Very real hard limits govern power. In fact, many utilities struggle to bring capacity online as fast as the demand requires. While renewables show some promise, a vast amount are not yet grid-connected and are running into obstacles and transmission line deficiencies. So where does that leave power? In the absence of new power, the remaining solution is to become far more energy efficient with the current power demand. Aiming for energy efficiency is also subject to cooling decisions. With roughly 50% of overall power consumed by cooling, cooling load is a primary factor for efficiency and overall power costs. All of this creates a difficult problem to unravel. How do you become more efficient while meeting demands at record speeds?

CONTENTS

"How do you become more efficient while meeting demands at record speeds?"

Site selection is a potential answer. Building where grid stress is lower and available power is higher will provide one level of assistance to the power conundrum. Organizations are looking to power companies to help them become "green" while delivering the needed MW for operations. According to Statista, power per kWh in the US ranges from the most expensive cost in Hawaii of $40.59 to the lowest of $8.29 in Iowa. However, transmission fees, demand charges, and peak demand fees are frequently added onto these base charges. Areas of the country near stock exchanges, subsea landing stations, and other services that make a site attractive may also lead to higher fees on top of power charges.

Beyond the location a company chooses for its data center, other factors will alter the demand for power after a site goes live. Data Center Infrastructure Management (DCIM) solutions help data center operators analyze, model, and plan for power and cooling fluctuations across the data center whitespace. The practice involves looking at loads across the footprint to guarantee enough power to operate. Understanding fluctuations, peak demand, and operational shifts is part of this demand response. Further planning for new hardware and technologies is imperative to guarantee smooth operations even during periods of growth. Efficiency gains happen with the best information and intentions.

Driving Growth

Growth is inevitable. Nowadays, AI is leading the growth of today's data centers. In fact, AI and data volumes are driving the repatriation of many data stores from the cloud.

A single AI search query can use up to five times the power of the same query on a search engine. With generative AI, where the outcome is generated based on variables, the power consumed can be 20 to 30 times greater than a single text search. The higher power demand in turn increases the need for cooling. In other words, if power is 20 to 30 times greater, a facility must plan cooling capacity to be 20 to 30 times greater in response. This level of cooling is not possible with much of the current cooling infrastructure.

Repatriating data from the cloud for machine learning and AI, or simply controlling cloud spending, drives demand back to the edge (or non-core data centers). About 80% of companies in an IDC study said that they are considering some form of cloud repatriation. This mindset shift will lead to new data center builds as well as revamping and retrofitting existing data centers. Understanding that many of these data centers were abandoned due originally to capacity limits, the facilities teams are now left to reinvent the room for repatriated demands. Workloads may consist of their entire stack, part of their IT load, or only their AI system. Higher density loads, such as AI, may lead to fewer servers but at a higher power and cooling density. In fact, the cooling for new AI loads may be completely different with greater efficiency. An upgrade or replacement of cooling technologies to support higher power and thermal density for a section of the room will render a previously unusable room usable.

Industry 4.0, smart communities, precision agriculture, autonomy, and various applications similarly drive power and site demands; again, much of the growth is at the edge. New infrastructure projects aim to increase communications speeds to and from multiple locations. The increase in demand for new users, bandwidth, storage, and applications creates the perfect storm for growth, regardless of capacity constraints. We simply must tackle capacity limits from all fronts. We must function more efficiently within our facilities while building additional capacity to accommodate growth.

Most equations for capacity planning revolve around the trinity of power, cooling, and compute. You must have enough power to both cool the room and power the compute. You must have enough cooling to maintain operations, as computing equipment shuts down when it gets too hot. You must have computing to make the power and cooling necessary. One such measurement used to express the ratio is PUE (power usage effectiveness):

PUE = total power / IT equipment power consumption

The closer to “1” a facility is, the better its overall efficiency. Reaching “1” is a statistical impossibility, although working towards this mythical “1" increases efficiency. But how is efficiency improved in an operational site? There are only a few ways to “find” capacity. A facility can free up existing capacity by moving to more efficient equipment, layouts, and other improvement projects. It can increase capacity through expansion projects. Or, a facility can free up capacity spoken for but unused in the form of stranded power.

“Decreased power for cooling increases power available for other needs within the data center ecosystem.”

Efficiency Gains

If you examine in isolation the systems that consume power, lose power in DC to AC conversions, and deliver power with transmission losses, no singular approach will lead to the best efficiency gains. Instead, combining efforts across both power and cooling platforms is ideal. Decreased power for cooling increases power available for other needs within the data center ecosystem. Adding direct to chip further lowers power consumption by reducing power normally attributed to server fan and cooling fan activity. AC to DC conversions will reduce loss. Upgrading older distribution and cooling equipment to more efficient options will lessen power capacity drains.

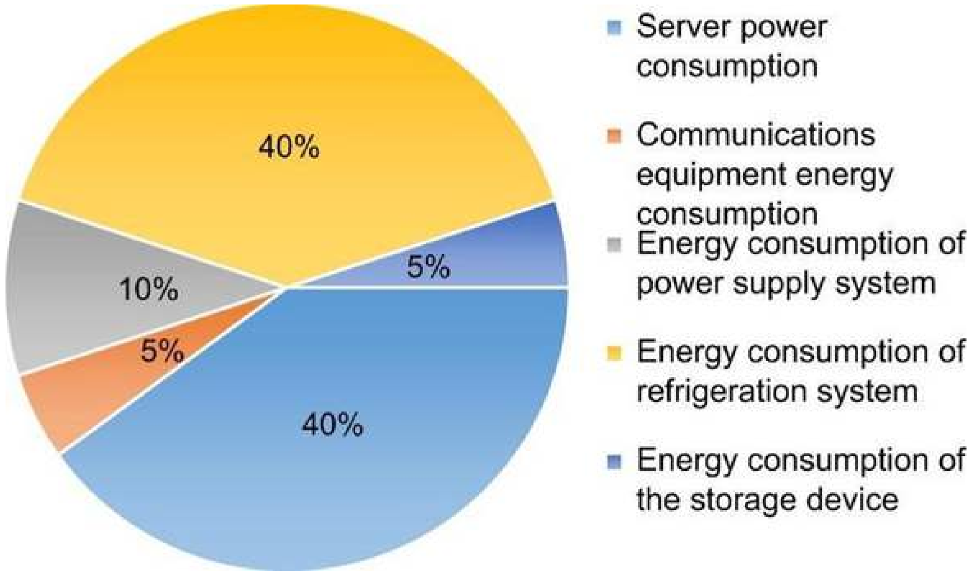


Fig. 1: Data Center Energy Breakdown

The above shows a typical breakdown of data center power. According to the American Council for an Energy Efficient Economy, of the 40% attributed to server power consumption, about 11% of the overall server power consumption is attributed to the fan power.

Computer hardware selection is one such power draw consideration. Hardware for servers, storage, networking, wide area access, and intrusion prevention consume power. CPUs are typically limited to a few cores to process information. Conversely, GPUs have multiple cores and can process multiple instructions simultaneously in concert with CPUs. The gains in processing speed and cores increase power demand further, resulting in increased demands for cooling and heat rejection.

There are programs that will pore over your stack and return the most efficient hardware. Alternatively, test and development departments are often tasked with finding the most efficient hardware for computing, continuously evaluating the next generation of hardware. Frequently, enterprises can consolidate multiple disparate systems into one hardier system, but at the cost of higher rack power density. For example, when examining a full rack of the latest iteration of Nvidia's Blackwell® system, which consists of GPUs, CPUs, networking, diagnostics, and graphics processing unit (GPU) microarchitecture, one cabinet clocks in at a whopping 120kW, compared to the average data center with 5-15kW per rack of consumption. Hyperscalers and larger enterprise data centers tend to have a slightly higher power density than the average and are creeping up rapidly due to AI and other digital demands.

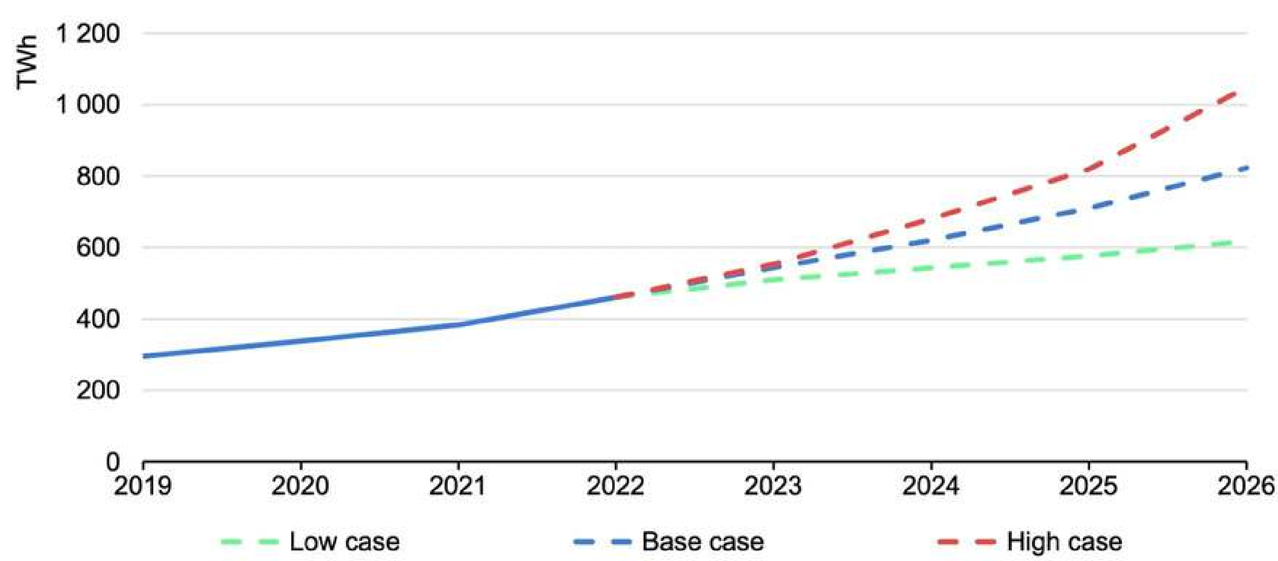
Fig. 1 Source: Ramdas, Shrinath & Rajmane, Pavan & Chauhan, Tushar & Misrak, Abel & Agonafer, Dereje. (2019). Impact of Immersion Cooling on Thermo-Mechanical Properties of PCB's and Reliability of Electronic Packages. 10.1115/IPACK2019-6568.

It doesn't take a lot of math to realize that new facilities and those implementing newer technologies, like AI, will demand significantly more power than a data center with none.

Highly redundant data centers suffer from higher percentages of stranded power. Power is stranded due to equipment mismatches, changes over time, predicted or modeled capacity allocations, or lack of use caused by redundant or "just in case" allocations, among other reasons. Reclaiming stranded power from cooling and compute loads will help add capacity from a power perspective. Likewise, moving to more effective and efficient cooling methodologies, in whole or in part, will free up power capacity for computing, or provide relief for what may be demanded from new power-hungry applications.

According to the IEA (International Energy Agency), power consumption of data centers could reach more than 1,000 terawatt hours by 2026. The increase is, in large part, due to AI and blockchain technologies. An increase in power demands an increase in cooling to support our compute loads.

Global electricity demand from data centres, AI, and cryptocurrencies, 2019-2026



Notes: Includes traditional data centres, dedicated AI data centres, and cryptocurrency consumption; excludes demand from data transmission networks. The base case scenario has been used in the overall forecast in this report. Low and high case scenarios reflect the uncertainties in the pace of deployment and efficiency gains amid future technological developments.

Sources: Joule (2023), de Vries, The growing energy footprint of AI; CCRI Indices (carbon-ratings.com); The Guardian, Use of AI to reduce data centre energy use; Motors in data centres; The Royal Society, The future of computing beyond Moore's Law; Ireland Central Statistics Office, Data Centres electricity consumption 2022; and Danish Energy Agency, Denmark's energy and climate outlook 2018.

The Cooling Conundrum

More power consumed means more BTUs of heat rejection are required. In general, to cool for the heat load of 1kW, 3412 BTUs of heat rejection are required. The efficiency of the heat rejection equipment will help dictate the amount and speed of heat rejection. In general, there are a few methods to cool or remove heat. They are generally divided into air, water, and other liquid technologies. Water is 3500 times more efficient at attracting heat than air. Other liquid cooling methods use refrigerants and non-dielectric fluids. Liquid cooling methods are more efficient than air removal. Closely coupled cooling, where the heat rejection is as close as possible to the heat source, further aids heat removal efficiency. In the example below, if we input our parameters using The Green Grid calculator, we know that using direct-to-chip cooling is 49% more efficient than with air.

	Direct To Chip Liquid Cooling	Air Cooling
PUE	1.172	1.338
WUE	0	2.054
Total KWh	15,033,766	29,641,046
KWh Savings vs Air	**49%**	-
White Space	18,938	78,750
White Space Savings vs Air	76%	-
First Cost	$6,719,749	$22,224,411
10-Year TCO	$29,009,760	$71,535,317
TCO Savings	59%	-
Breakeven	Year 3	-
10-Year OpEx Savings vs Air	55%	

Water is a scarce resource in some parts of the country. Until recently, fewer liquid cooling options were available to answer water shortages, leaving less efficient air for heat rejection needs. Newer cooling methods show great promise in answering the cooling conundrum if a facility can use less power for cooling while providing the same heat rejection, which frees up power elsewhere for computing, networking, and other resources within the facility. This in turn provides a more efficient cooling balance across the whitespace.

Close-coupled solutions prove more efficient than whole-room solutions. The principle is to align our heat rejection equipment closer to the heat load needing removal. With more closely coupled cooling, the heat exchange is quicker. It creates a shorter cycle, or closer union, between the heat source and the removal method.

Liquify Your Chill — Closely

Due to the superior performance of liquid cooling, we know that liquid cooling can perform the same heat rejection with less effort. The type of liquid will have a bearing on the speed of the heat transfer. Further close coupling, as outlined above, provides additional efficiencies. Therefore, liquid cooling is available to help increase efficiency and step up cooling capacity for higher power applications.

While some people push back on liquid cooling in the data center, the first computer was liquid-cooled. There simply wasn’t enough air to cool the early processors. Fast forward past other cooling methods, and we are returning to liquid cooling. It’s not just that a heat rejection model can remove heat; we must also consider thermal flux. Thermal flux is how much heat is transferred in a fixed unit in a fixed amount of time. This term is used interchangeably with thermal density. It makes sense that the hotter the air needing removal, the more energy it will take. If the cycle happens once, we remove heat once. We know in data centers, we must remove that heat over and over as the source of that heat is active and continues to produce heat.

Immersion cooling is one such liquid cooling methodology. Servers are immersed in liquid-filled “cabinets” that are horizontally situated. Onsite whitespace crane-like apparatuses insert and remove the servers, allowing the non-dielectric fluid to drain before the server is changed or repaired. As the chips are mounted on the server boards within the liquid, the efficiencies gained over traditional methods are measurable. The Open Compute Project is working on incorporating immersion into one of their designs.

Meanwhile, direct-to-chip (or D2C) cooling technologies take closely coupled cooling to the closest coupling possible. With D2C cooling, a plate sits on top of the processing unit where the heat is generated. A series of manifolds directs the heated liquid away from the chip and into a coolant distribution unit. The entire assembly operates similarly to its larger counterpart, the rear door heat exchanger, but rather than cooling an entire rack at a distance of a couple of feet, direct-to-chip takes the cooling directly to the heat source, the processor itself.

As noted, liquid cooling, and closely coupling that cooling to the heat source, are key to cooling future AI chips. There are two types of direct-to-chip cooling. The first, single-phase, uses water-based liquids that stay liquid throughout the thermal heat transfer. We can improve upon that number yet again with two-phase liquid cooling. Two-phase uses a dielectric refrigerant that nucleates, allowing for even greater heat removal efficiency. This method is sometimes referred to as refrigerant-based cooling. According to Nvidia, liquid cooling offers a significant cost reduction over air-cooled counterparts and with a smaller footprint. In fact, in a recent Data Center World presentation on liquid cooling, Mohammad Tradat, PhD, stated that Nvidia's new AI chips will require two-phase liquid cooling.

"Air is lazy."

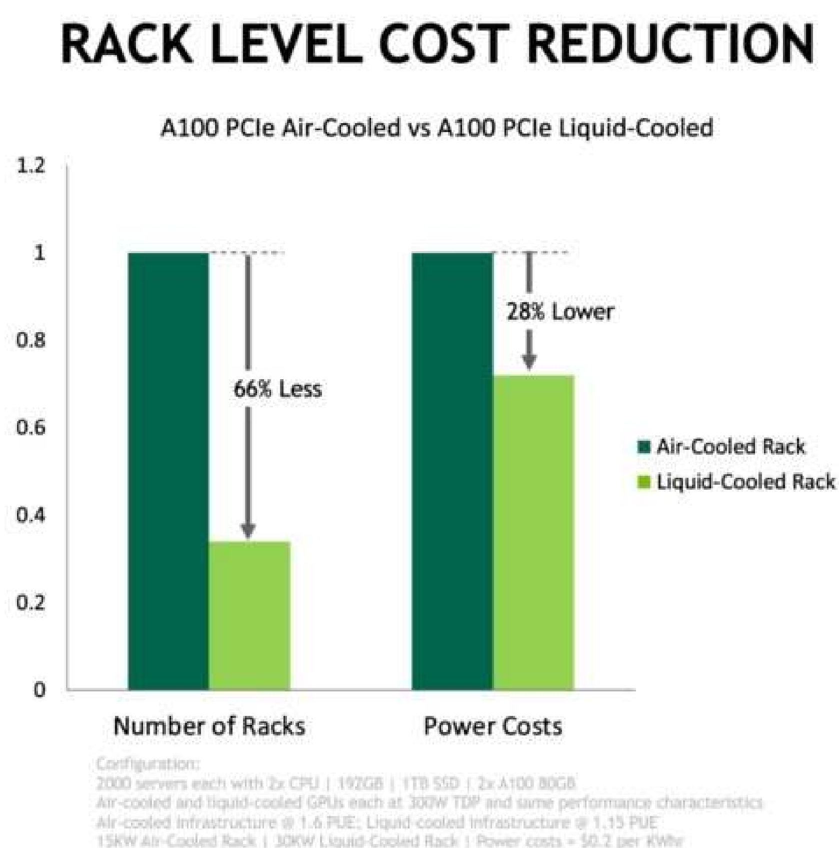
Nvidia's Ali Heydari has affirmed that not only is direct-to-chip (D2C) cooling more economical, efficient, and lends itself well to hybrid data centers with varied cooling capacity, but, according to their studies, D2C allows silicone to have a longer lifecycle when directly coupled with the processors. Further, removing the thermal volatility extends the life and reliability of the silicone due to thermal consistency. D2C rejects the heat in a surgical manner. Extending the life of electronics, especially expensive ones, provides a better return on investment.

Ali continued: "We have been living with air for 70 years. Air is lazy. The laws of physics dictate we move from air to liquid cooling." Two-phase offers the benefit of liquid, with the added benefit of refrigerants' enhanced heat rejection properties. The use of these two-phase refrigerants will help alleviate corrosion and erosion problems created when thermal flux creates the need for more frequent coolant turnover.

Liquid Cooling in Action

Due to this directly coupled arrangement, the heat from the processor never hits the room air. NeuCool, from Accelsius, is a complete liquid cooling solution that uses a highly efficient two-phase process and a dielectric refrigerant that is entirely safe for electronics. The NeuCool system supports 1500W+ per socket and up to 100kW per rack (80kW direct-to-chip cooling), efficiently cooling current and future generations of high-performance CPUs & GPUs. NeuCool is a tailorable solution adaptable to specific computing needs and infrastructure systems, whether a company is adding spot cooling capacity or wishes to supply the solution across the entire floor for high-performance centers.

As companies work go create efficiency gains across their data center floors, every part of the ecosystem plays a part in the balance. Liquid cooling, particularly direct-to-chip two-phase cooling, is an excellent tool to conserve power overall. Energy not spent on cooling is energy that can be used for computing. The need for computing power will never wane. The trick is to ensure that a facility is using the correct facilities to support that compute in the most efficient manner. D2C cooling is one such solution to change capacity ratios, freeing up more power for computing.



ACCELSIUS™

info@accelsius.com

www.accelsius.com