October 23, 2024

Liquid Cooling: The Solution To AI's Heat And Water Problem

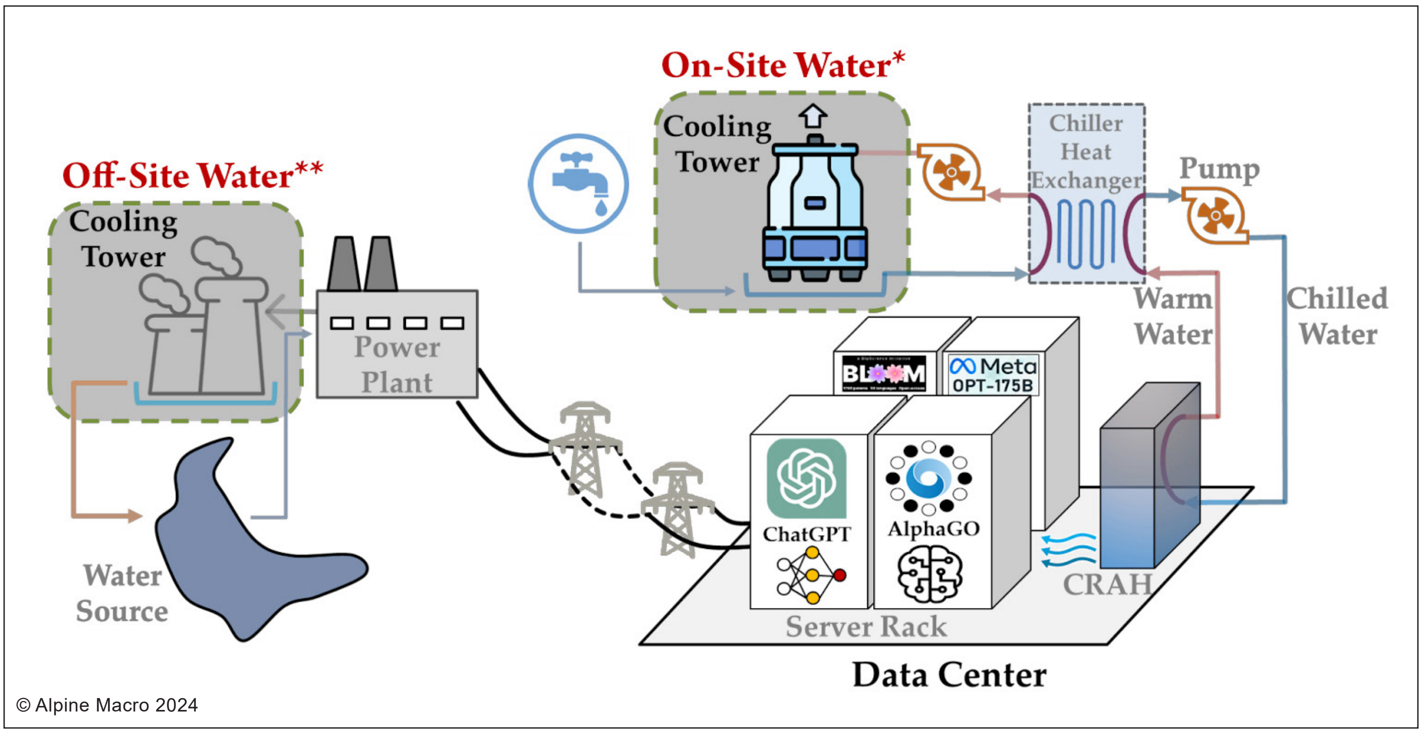
In This Report

AI's Resource Use Dilemma 2
More Resource Usage Even With More Efficient Chips 3
Next-Generation Liquid Cooling 4
Investment Considerations 5

The computational power required for AI development is doubling roughly every 100 days. As a result, the number of data centers, which stands at roughly 8,000+ globally, is exponentially scaling. The cooling and electricity generation requirements of the data center buildout are unsustainably water intensive (Chart 1).

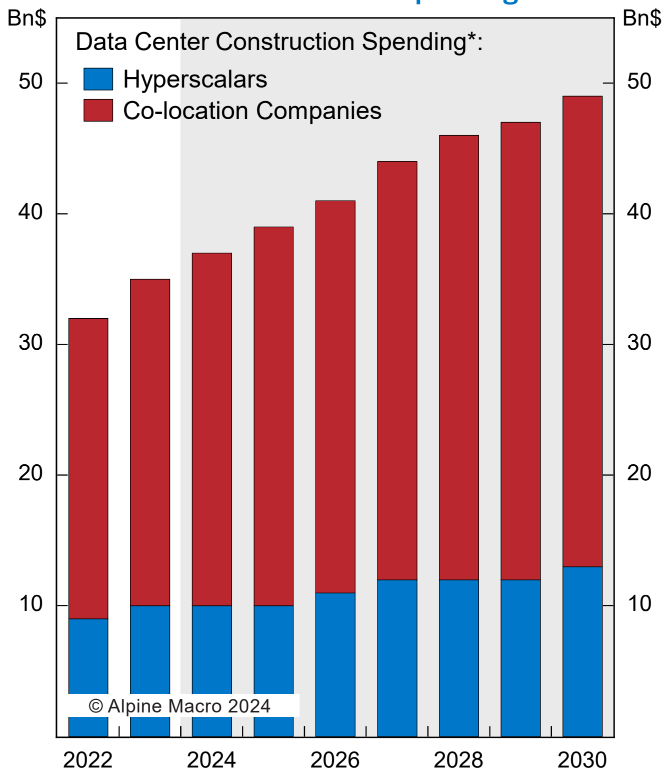
In North America alone, data center construction and investment (Chart 2) is soaring with supply under construction hitting a record 3.9 gigawatts, a ≈ 70% increase from a year ago, notes CBRE Group. In the first half of 2024, over 500 MW of new data centers, roughly equivalent to all the existing capacity in Silicon Valley, were rolled out in North America. These high-density data centers generate copious amounts of heat and must be cooled down

Chart 1 An Example Of Data Center's Operational Usage



*On-site water for server cooling using cooling towers
**Off-Site water usage for electricity generation
Source: OECD

Chart 2 Ever-Growing Data Infrastructure Spending



*Includes construction spending by providers. Excludes enterprise spending and any other capital expenditure outside of construction (such as equipment);
Note: Shaded area denotes forecast period; source: Synergy Research Group

as temperature directly affects semiconductor performance. Currently, ≈80% of data centers use sophisticated air-conditioning systems to keep temperatures below 82°F (just under 28°C). Some high-performance computers require server-room temperatures below 50°F (10°C). Temperature is one of the largest external factors that impacts chip functionality. This has resulted in cooling becoming the fastest growing physical infrastructure cost for data centers, growing at a 16% CAGR, notes Bloomberg.

The problem with current cooling systems is their heavy consumption of both water and energy, with cooling alone making up about 40% of a data center's total power usage. This significantly contributes to AI's unsustainable carbon footprint. For example, Google's emissions have surged by 48% since 2019 due to the AI boom. In terms of water use, Google's data centers used over 6.1 billion gallons of water last year, a 17% increase from 2022. It is worth noting that both Google and Microsoft have pledged to be both carbon neutral and water positive by 2030. Without improvements in data center cooling, tech companies will impede AI development and will fail to meet sustainability goals.

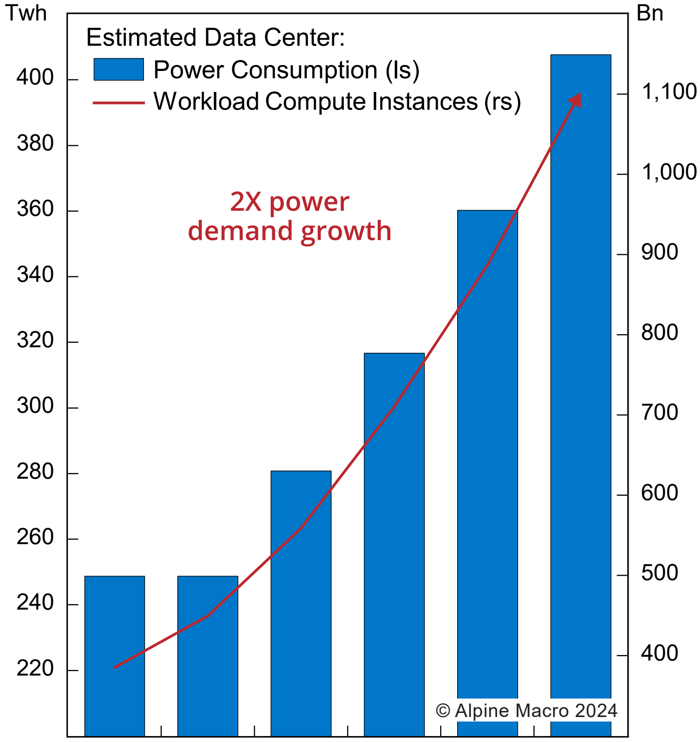
A fundamental overhaul of data center cooling is underway and represents a unique AI investment avenue. This report will argue that next-generation liquid-cooling technologies that use advanced liquids are on track for swift adoption and will become the default data center cooling technology. Next generation chips are not compatible with traditional cooling methods.

AI's Resource Use Dilemma

While data centers have historically been guzzlers of both water and power, AI has sent this into overdrive (Chart 3). By 2030, U.S. data center electricity consumption could more than double, reaching 9% of total electricity demand in the U.S. according to The Electric Power Research Institute.

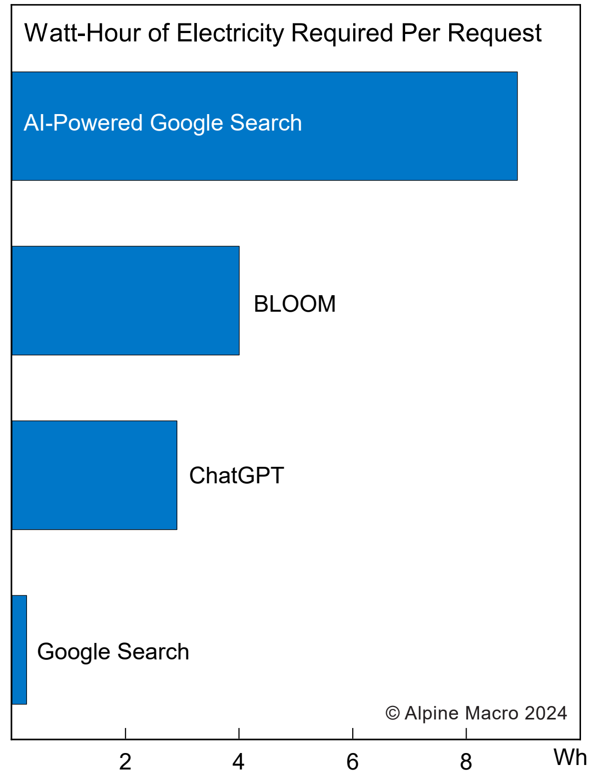
Unsustainable water use across both model training and inferencing must be addressed as AI scales. Water is the globe's scarcest natural resource, with 4 billion people experiencing severe water scarcity for at least one month annually. Last year, U.S. data centers consumed over 75 billion gallons of water. A recent study by the University of California, Riverside, predicts that data centers

Chart 3 More Computing, More Energy



Source: Goldman Sachs, IEA, Cisco

Chart 4 Resource Intensity Of AI Inference



Source: LA times, Alex de Vries

could raise their combined water consumption to 6.6 billion cubic meters annually by 2027 – about half of the U.K.'s yearly water withdrawal. A single 100-word email generated by ChatGPT-4 requires 519 milliliters of water, which is just over 1 bottle of water. This is significantly higher than a simple google search (**Chart 4**).

More Resource Usage Even With More Efficient Chips

Although individual chips are becoming more efficient on a per watt basis, overall power requirements are still increasing due to the growing complexity and volume of calculations. The more transistors on a chip, the higher the power demand. For example, Nvidia's A100 chip requires 400 watts, while the H100 chip which has more transistors requires 700 watts. Nvidia's Blackwell chip will soon consume 1,000 watts of electricity per chip – over 4x the estimated power needed to run a MacBook Air for eight hours. Power requirements and temperature have a strong positive correlation, so any increase in rack density must be paired with enhanced cooling capabilities.

In a data center, chips are housed inside of server racks. Think of chips as books and the server rack as the library shelf. While increasing the "density" of a rack via incorporating more chips increases computing power, it also requires more electricity. Previously, server racks were considered high density at 30-50 KW. Now, some racks comprised of new chips are pushing 100 KW. Yet, according to industry estimates, less than 5% of current data centers globally can support even 50 KW per rack-highlighting a significant adoption readiness gap. Computing demand is so high that any data center is currently indispensable (**Chart 5**).

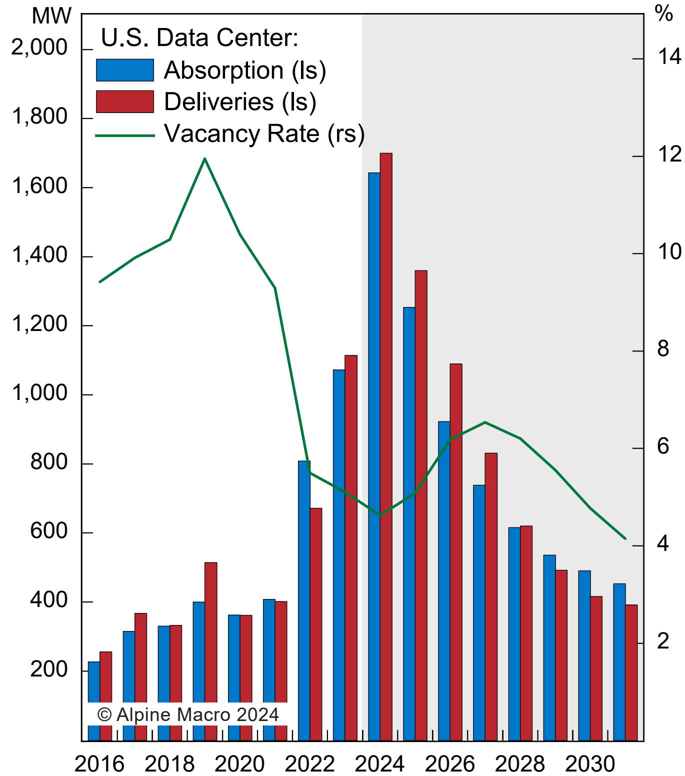
Traditional air cooling and even chilled-water cooling, which absorbs four times more heat, are unable to maintain proper temperatures in data centers as next-generation chips produce more heat than ever before. Server rack density will only continue to increase.

Next-Generation Liquid Cooling

Liquid cooling offers distinct benefits compared to traditional data center cooling systems. Importantly, liquid cooling boasts streamlined infrastructure that improves scalability compared to chilled-water setups. In addition, liquid cooling leverages specialized fluids that remove the need for advanced refrigeration systems by directly capturing and dissipating heat from server components. Use cases indicate that liquid cooling can significantly lower facility power usage. For example, Supermicro's liquid cooling systems enable data centers to reduce power consumption by 30-40%. As an added benefit, liquid-cooled data centers promote higher server density as they can accommodate twice the computing power within the same space, notes Nvidia. Hewlett-Packard reports that their liquid cooling technology can reduce data center carbon footprints by 87% and reduce annual operating cost by 86%.

Overall, liquid cooling is the most efficient heat-removing method outperforming air cooling by

Chart 5 Limited Data Center Vacancy Rates Regardless Of Supporting Infrastructure Capabilities



Note: Shaded area denotes forecast period; source: CBRE

2-10x, according to Intel. The technology also facilitates uniform cooling that eliminates hidden infrastructure "hot spots". As a result, liquid cooling's benefits have positioned it for substantial market growth. A recent analysis by the Dell'Oro Group found liquid cooling adoption has reached an inflection point this year and will become a $15 billion opportunity over the next five years.

Currently, 95% of current data centers use air cooling with less than 5% using advanced liquid cooling, notes Morgan Stanley. The technology's growth potential is massive and demand for liquid cooling could become one of the most pressing bottlenecks to boosting data center capabilities. The nation that realizes the liquid cooling competitive advantage is positioned to win the next leg of the AI race, where models become more complex and accurate.

There are three main types of advanced liquid cooling systems that require varying degrees of modification to existing data center infrastructure.

1. **Immersion cooling:** submerges critical data center components in a non-conductive dielectric liquid. While immersion cooling boasts the highest energy efficiency, it requires extensive CAPEX and is best integrated into new data center buildouts.

2. **Direct-to-chip cooling:** involves circulating liquid coolant across key heat-producing components like GPUs. While this approach offers the second best heat absorption, it does require adjustments to server design. Nvidia has announced that its next-generation DGX servers will incorporate this liquid cooling approach.

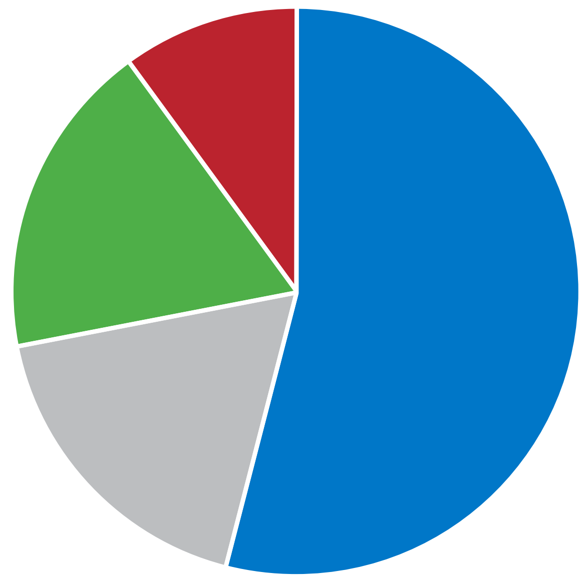
3. **Rear-door heat exchangers:** offers a blended cooling approach that can be incorporated without significant infrastructure modifications. It combines traditional air cooling with liquid-filled radiators attached to the back of server racks. This cooling approach is mostly a "stopgap" but is an appealing upgrade to outdated cooling infrastructure.

Investment Considerations

As the computing demands of AI continue to increase, we are in the camp that the liquid cooling approaches highlighted above will realize widespread adoption (Chart 6). As server racks rapidly scale, data centers that do not modernize their cooling infrastructure will simply be unable to support the next-generation chips that are critical to cutting-edge AI development. Such data centers risk becoming stranded assets.

Chart 6 Liquid Cooling Breakdown

Estimated Share of Data Centers
By Liquid Cooling Technology (2028)



Note: A single-phase fluid remains in its liquid form during the entire cooling process, while a two-phase fluid undergoes a phase change and becomes a gas.
Source: Dell'Oro Group

Liquid cooling is not a new technology. However, until recently, it has realized minimal adoption as air and water-based cooling approaches have sufficed. This is no longer the case, and liquid cooling is evolving into the *"de-facto"* cooling method for data centers seeking compatibility with next-generation chips. AI leaders including Nvidia, Microsoft, and Google have announced their aim to prioritize using liquid cooling across future data centers builds. The sector is poised to experience a nascent demand shock.

As of now, there are no pure play liquid cooling companies. However, there are several compelling candidates with exposure to liquid cooling solutions. Asian solution providers include Auras Technology Co., Ltd. (3324.TW) and Lite-On Technology Corporation (2301.TW), while Hewlett Packard Enterprise (HPE) and Vertiv (VRT) provide Western exposure.

Noah Ramos
Global Strategist

A Unique Mind On The Markets