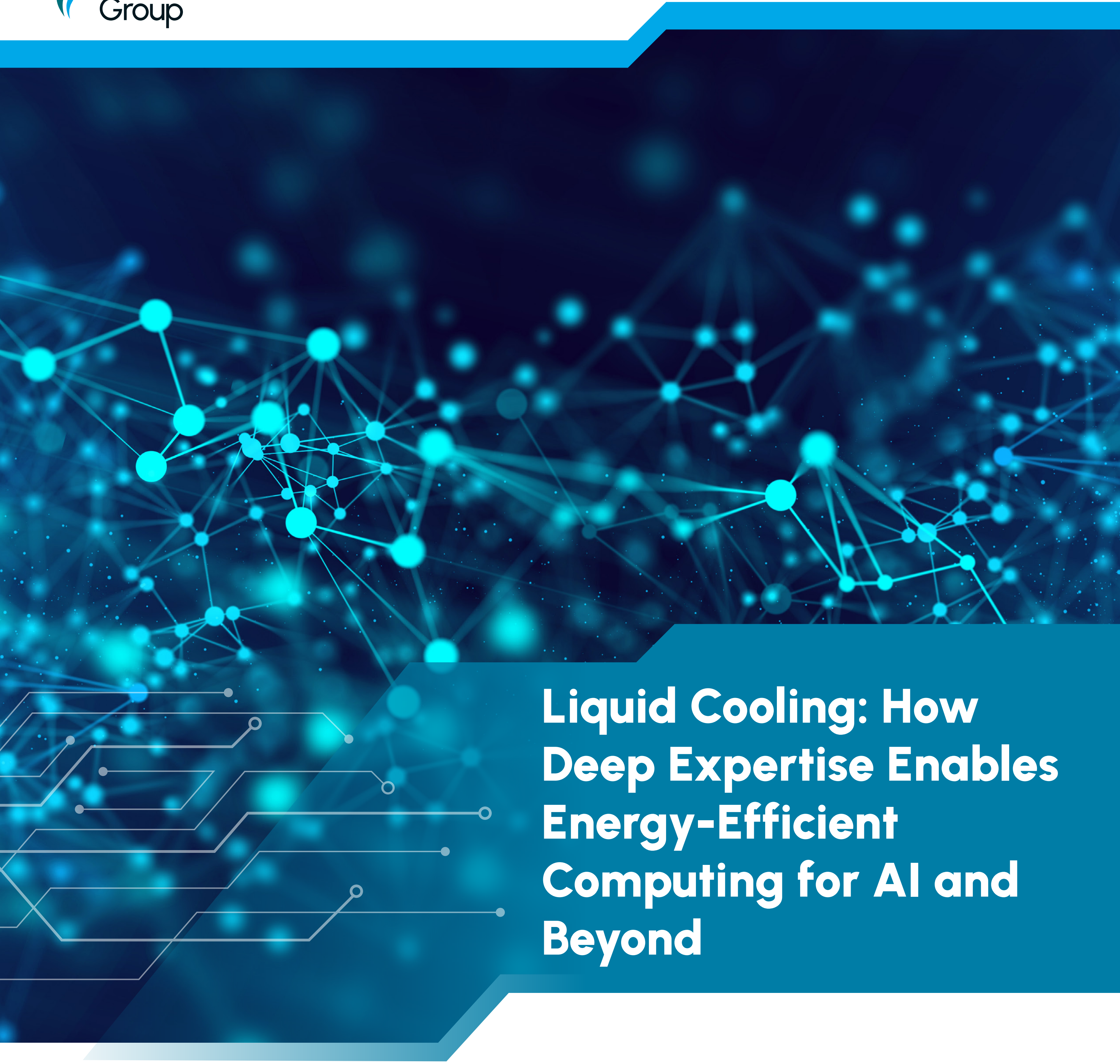
The Futurum Group



Liquid Cooling: How Deep Expertise Enables Energy-Efficient Computing for AI and Beyond

AUTHOR

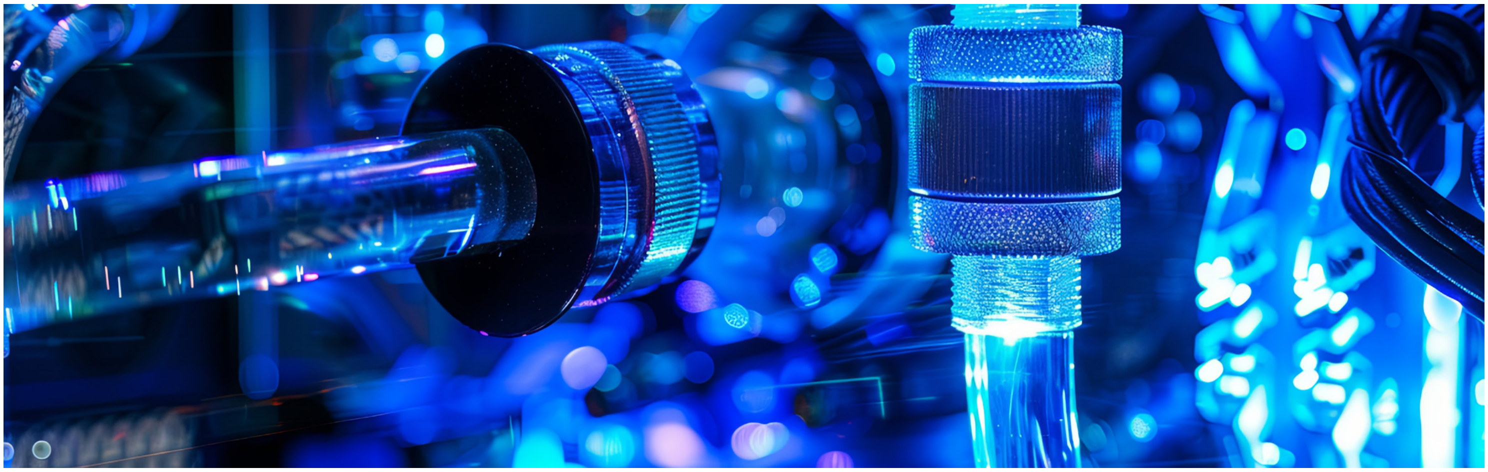
Steven Dickens
Chief Technology Advisor | The Futurum Group

Ron Westfall
Research Director | The Futurum Group

AUGUST 2024

IN PARTNERSHIP WITH

Lenovo | NVIDIA.



Executive Summary

Artificial Intelligence (AI) and, in particular, Generative AI (GenAI) applications, are taking the world by storm. Across a wide range of industries, AI is enabling breakthroughs in everything from retail to manufacturing, from financial services to healthcare, from basic research to cybersecurity.

These breakthroughs are enabled by computing technology beyond our imagination just a few years ago. Yet, as this technology progresses, the power consumption for virtually every component of a server—CPUs, memory, accelerators, networking, etc.— has increased 200%, on average, over the past decade. These power-hungry systems needed for AI workloads are filling up data centers across the world and are a major contributor to the overall power increase seen in the data center space. In the US alone, energy demand due primarily to AI data centers will increase by 5.1% to 9.1 % of the US's overall energy demand by 2030. Today the US's energy consumption is only 4% according to the Electric Power Research Institute.

The Joule effect, in essence, states that when an electric current flows through a conductor with resistance, that electric energy is converted into heat. This means that as more power is consumed, more heat is generated. Dealing with the heat generated by the increased power demands of AI and High-Performance Computing (HPC) has become a challenge for many organizations. Large increases in power have downstream effects. Higher powered components like CPUs, GPUs, memory, etc. generate more localized heat, and heat is the enemy of silicon. Heat that is not dissipated limits the performance of CPUs and GPUs and can damage the silicon itself and/or cause severe issues like overheating and fire.

Against this backdrop, not only are silicon vendors working to improve the efficiency of individual components, but liquid cooling approaches are increasingly becoming a focus for both silicon and system designers because of the ability of liquid to absorb heat faster and at higher capacities than traditional air-cooling methods. Liquid cooling can also help meet the escalating demand to enhance data center energy efficiency for expanding AI workload environments, due to downstream effects such as the elimination of air conditioning power consumption, power to chill water and the power needed for fans to move air. This includes managing and optimizing energy consumption by using liquid cooling technologies with higher inlet temperatures (no need to chill water) all the while ensuring that system performance is not compromised and core density per data center is not affected.

Moving to liquid cooling can also assist with a company's ability to work toward Environmental, Social, and Governance (ESG) goals that relate to carbon emissions and sustainable practices. If a company or data center co-location provider does not modernize but maintains the status quo of air cooling, then the energy explosion of AI silicon power will be exacerbated by the need for colder air that will need to move at faster rates. More power will be required for air handling vs. IT consumption and that will indirectly handcuff a company from being able to reduce their carbon footprint. All told, the environmental impact of AI infrastructure, particularly due to high energy demands, poses challenges for companies striving to balance technological advancements with their ESG commitments. Liquid cooling can help address these issues.

We assess why Lenovo Neptune® 6th generation liquid cooling solutions are superior at enabling AI workload optimization without compromising organization-wide sustainability goals. This report focuses on the deep and long-standing investments Lenovo has made in engineering and innovation when it comes to liquid cooling. We examine how the system and solution level engineering for Lenovo Neptune® liquid cooling show higher quality and metrics that compare favorably against industry standards.

Lenovo has a long history of liquid cooling innovation for x86-based systems that started in its IBM days before the acquisition of IBM's System X business in 2014. Lenovo's 12 years of delivering production-level liquid cooling on x86-based systems (both with and without GPUs) show in the company's approach from its details in design, manufacturing and delivery process to overall quality and support processes.

The company's engineering differences ultimately provide a competitive advantage with higher quality and heat capture breakthroughs. Lenovo's deep experience and understanding of conductive materials and fluid behaviors across the field and their unique approach in designing liquid-cooled solutions deliver breakthrough system quality plus a competitive advantage in providing energy-efficient computing.

Section 1: The Evolving Liquid Cooling Landscape

It's no stretch to say that AI is revolutionizing a wide variety of industries. AI is helping retailers improve the customer experience, reduce inventory losses, and increase profits. It's helping researchers develop new pharmaceuticals, explore the origins of the universe, and work to mitigate climate change. AI is helping manufacturers improve quality, enhance process monitoring and provide predictive maintenance; enabling healthcare professionals to reduce overhead, improve diagnoses, and spend more time with their patients. And in the realm of financial services, AI is improving fraud detection and enabling faster, more accurate investment decisions.

These breakthrough applications are enabled by computing technology beyond our imagination just a few years ago. As this technology progresses, the power consumption for virtually every component of a server—CPUs, memory, accelerators, networking, etc.— has grown roughly 200% over the past decade. Processing power has increased to up to 500W for CPUs and 1000W+ for GPUs, with corresponding power increases for other server components. The overall power requirements for a single rack have increased from about 15kW to up to 100kW+.

Dealing with the heat produced by all this electrical power has become a challenge for many customers. Clearly, alternative methods of cooling are needed as customers try to balance their desire for new AI applications, greater IT performance, and the economic and environmental realities of server, rack and data center cooling. Liquid cooling has emerged as a key enabler for the new generation of AI and GenAI applications.

Let's look more deeply at the drivers for liquid cooling. Not only are the individual server components we've mentioned consuming increasing amounts of power, but AI and HPC workloads are altering the power demands at a data center level. AI training is synchronous, with all the elements in the cluster working in concert at immense intensity on a single service. The size of an AI training cluster can be massive; it can require 5,000 servers at minimum with each server using several CPUs and GPUs. And with a single server requiring ~10kW of power, it's easy to see how the problem is compounded at scale. AI workloads can consume up to three times the power of typical cloud workloads, producing a corresponding increase in heat.

We have discussed how an increase in device power influences the overall data center. Let's look at other relationships that exist for these next-generation systems that are designed to support these advanced computing workloads.

Next-generation advanced computing presents high-level system design challenges across three key areas: thermal, mechanical, and electrical:

- **Thermal:** Hotter components overall (e.g., Voltage Regulators, CPUs, GPUs, Memory, Networking, etc.), lower silicon threshold temps required generation to generation, geographical challenges (hot arid data center locations)

- **Mechanical:** Dramatic increase in socket sizes to accommodate higher transistor counts, dual in-line memory module (DIMM) quantities per CPU or GPU are increased, and Voltage Regulator (VR) quantities, network card sizes along with an increase in board layers and increased silicon height

- **Electrical:** High demand for ≥600W CPUs, ≥1200W GPUs, and move toward 48V input. One AI system can go beyond 10kW.

Next, let's look a little more deeply into how the two methods, liquid cooling and air cooling, differ and why it matters to AI. Liquid has a higher thermal capacity than air. Thermal capacity can be understood as how much heat a medium can absorb before its temperature changes. The thermal capacity of liquid is much greater than that of air, which simply means liquid can absorb a significantly higher amount of heat before the liquid raises its temperature, assuring that the heat rejection process is significantly more efficient.

There's a data center footprint issue to cooling as well. With air cooling, heat sink performance is tied to the overall surface area. The larger the surface area, the better the heat sink performance. Increasing the surface area of the heat sink can be accomplished by expanding it horizontally or vertically or both. Ultimately, this increase in surface area results in larger systems, height-wise, thus diminishing the amount of computing power that can go into a data center. Translated, that means less computing for an increase in power!

Liquid cooling, however, can maintain more compute power in traditional server form factors because it's so efficient at absorbing heat, and the effect of surface area as a relation to performance is dramatically decreased, thus allowing for denser computing in a traditional computing footprint. Liquid-cooled servers require less space, generate less noise (lower fan power due to reduced need for high air flow), and enable an overall cooler data center environment.

All these factors have driven an intense amount of recent investment in liquid cooling in the data center industry. We find that there is a general perception that all liquid cooling solutions are the same with scant difference among vendors. However, in our assessment, we have found that is not the case at all. Lenovo's 12 years of experience in liquid cooling technology development has enabled the company to build solutions that not only meet but greatly exceed general standards and in our assessment are significantly better than the competition in design, manufacturing, delivery, and services.

Section 2. Why Lenovo Neptune® Is the Proven Solution for AI Workloads: Superior Liquid Cooling Technology

We examine why Lenovo's 6th Generation Neptune® Liquid Cooling solution offers a robust solution available today to meet distinct AI workload demands. The Lenovo solution provides significantly higher quality and delivers more efficiency than air cooling in key areas such as heat dissipation, all the while ensuring peak performance and greatly enhancing energy efficiency. Fundamentally, the Lenovo Neptune® solution provides a holistic cooling approach that we identify as market leading across the data center realm.

First, the Lenovo Neptune® technology portfolio stands out. Lenovo Neptune® is built on more than a decade of liquid cooling expertise with more than 40 patents, playing an integral role in large-scale supercomputing and AI cluster implementations, enabling organizations to deploy high-performance AI at any scale. Lenovo was the first to deliver a high inlet temperature liquid-cooled petascale supercomputer at the Leibniz Supercomputing Centre (LRZ) in June of 2012 that debuted at #4 of the list of Top500 Supercomputers (www.top500.org), which measures the fastest 500 systems in the world.

This long experience with liquid cooling has enabled Lenovo to fine tune its technology and manufacturing prowess. Some of the superior Lenovo technology distinctives we have observed include:

- **Hosing (Internal and External):** Lenovo uses high quality Ethylene Propylene Diene Monomer (EPDM) hoses. EPDM is known for its high PSI strength and low durometer range which indicates greater flexibility than the plastic tubing used by some others in the market. These hoses are also treated internally with peroxide which prevents hose corrosion and degradation.

- **Main Cooling Loop:** Lenovo's main internal loops and cold plates use copper instead of fluorinated ethylene propylene (FEP) because of copper's historic record as an excellent electrical conductor, strength plus durability, and more efficient dissipation of heat.

- **Optimized, patented cold plates:** Lenovo's low pressure-drop cold plate design maximizes heat extraction for accelerators consuming ~700W now and 1000W+ in the future.

- **Water First:** Lenovo Neptune® uses water rather than other fluids such as polyethylene glycol, since water has one of the highest thermal capacities, is easily sourced, is safe for plants and animals, and requires no messy hazmat cleanup.

- **Warm water vs chillers:** Lenovo Neptune® direct water-cooling technology can use inlet temperatures up to 45°C rather than requiring pre-chilled water, reducing the energy requirements for the overall data center and eliminating the need for specialized data center air conditioning equipment. Additionally, the hot output water can be repurposed within the facility, turning waste heat into value due to the high energy content stored in the form of heat which can range in the 65°C range.

- **Brazed Joints:** Lenovo uses brazed joints rather than O-ring face seal (ORFS) fittings since they are known for their leak-free performance, support and stability (eliminating twisting of the tube).

Industry credentials include Lenovo's established supercomputing pedigree according to the TOP500 supercomputing list. Lenovo has more TOP500 systems than any other vendor (as of the latest list released in June 2024). Nine out of ten of Lenovo's fastest TOP500 systems use Neptune®-based liquid cooling.

Secondly, Lenovo's manufacturing process and delivery approach align with the emerging requirements of sustainable data centers. Lenovo developed Supply Chain Intelligence (SCI), an AI-powered solution that continuously analyzes supply chain data to identify potential issues and resolve them rapidly.

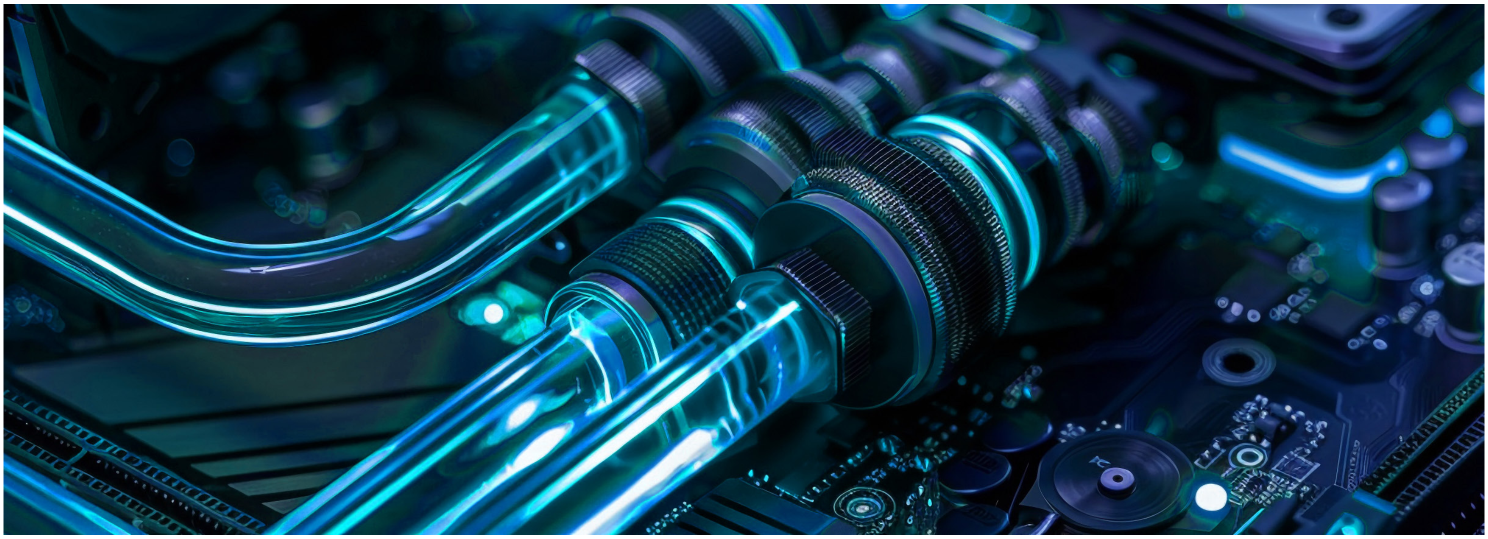
Of key importance, Lenovo Neptune® servers are tested together as a solution, rather than just as single systems. Systems are filled with deionized water, brought online and stress tested to ensure that overall system operation and water flow meet high standards before they are packaged and readied for shipment to a customer by draining the fluid and pressurizing the loop with Nitrogen. Shipping without liquid in the loop produces a competitive advantage over other vendors who ship with liquid and have experienced component damage and liquid leaks during transport.

Thirdly, we find that Lenovo's long experience with liquid cooling shines through in the solution's support and maintainability as well. Lenovo provides one-stop, first-line support for all Neptune® solution equipment and components, from servers to cooling distribution units (CDUs), unlike other providers who generally require each vendor to support its own solution components. And Neptune® chassis-enclosed systems automatically connect to water when installed, so that individual Neptune® servers can be serviced while others remain in production.

The Lenovo Neptune® Liquid Cooling solution is strengthened through the company's ability to deliver pre-optimized, tested, and certified IT solutions. This helps to ensure that Lenovo Neptune® offers high-performance, scalable, and energy-efficient solutions through integration of factory-tested components to lock in reliability and cost-effectiveness.

Finally, Lenovo Neptune® Liquid Cooling delivers whisper-quiet high performance with lower power consumption in a compact footprint, so customers can achieve higher density and more output from their data center. During operation, the direct water-cooling solution recycles loops of warm water to cool data center systems and keeps all server components cool, decreasing the need for noisy and power-hungry system fans in data center operation.

It is noteworthy that Digital Realty recently selected Lenovo Neptune® cooling technology in the building of its high-density colocation offering throughout more than half its data centers worldwide. From our view, this represents a major endorsement in validating the production readiness of the Lenovo 6th generation Neptune® portfolio in meeting the distinct challenges of high-density workloads presented by AI.



Section 3: Lenovo Neptune® Liquid Cooling Technology Benefits and Competitive Advantages - A Proven Track Record of Solution, Leadership, and Tech Innovation

Lenovo Neptune® is not a singular technology that is applied uniformly across its entire ThinkSystem portfolio. Neptune® provides a holistic cooling approach that doesn't just cool CPUs; it includes options for direct warm water cooling of entire systems, individual components, and rack-level rear door water cooling. Such a comprehensive approach underpins the ability to optimize performance and efficiency and to tailor solutions appropriate for individual customers and installations:

- **Direct Water Cooling (DWC):** The solution offers DWC of individual system components, available across Lenovo's ThinkSystem portfolio, as well as full-system cooled servers in which DWC uses parallel liquid flow over all heat-producing components (CPUs, Accelerators, DIMMs, Networking, etc.) with a water loop that can capture up to 98% of the system heat and reduce system level and data center power consumption by up to 40 percent.

- **Liquid Assisted Cooling:** Lenovo Neptune® liquid-assisted cooling provides the benefits of liquid in an air-cooled system. With either a thermal transfer module (TTM) or liquid-to-air heat exchanger (L2A), traditional air-cooled systems benefit from liquid cooling with specially designed heat handling, all without additional plumbing. This supports customers who choose to keep the status quo for air-cooled data centers without compromise in introducing liquid cooling.

- **Rack Water Cooling:** The Rear-Door Heat Exchanger (RDHX) delivers 100% heat removal efficiency without requiring moving parts or power. RDHX works with standard air-cooled servers, storage, and networking without modification, easing assimilation into existing data center infrastructure and taking advantage of in-rack CDUs that provide efficiency gains over traditional computer room air conditioning (CRAC) units.

We find that Lenovo Neptune® technology has proven instrumental in supporting top priority use cases, including engineering, modeling, simulation workloads, and animation movie production that all require compute power at peak performance. For example, Lenovo Neptune® is decisive in key sectors such as fintech, computer-aided engineering and computational fluid dynamics (CAE/CFD), electronic design automation (EDA), weather and climate modeling and forecasting, bioinformatics, geospatial/energy, earth sciences research, and animation render farms. Lenovo's long history in HPC has positioned them well for tackling the future of AI and all the downstream power, cooling and sustainability challenges that are related to data center expansions to support AI.

Moreover, the Lenovo Neptune® liquid cooling solution has won the 2024 Business Intelligence Group Sustainability Product of the Year award and is CRN's 2024 Best Green Energy Product of the Year, winner of the 2024 SEAL Sustainable Product Award and HPCwire's Best HPC Server Product or Technology for 2023, further boosting its ecosystem-wide ESG credentials.

Neptune® liquid cooling solutions benefit immensely from Lenovo's partnership acumen as exemplified by the NVIDIA partnership. This includes new comprehensive services powered by NVIDIA through the Lenovo AI Center of Excellence targeted at GenAI, which has been identified as the worldwide top tech investment priority of business and IT decision makers. For instance, Lenovo ThinkSystem SR780a server uses Neptune® liquid cooling to attain an ultra-efficient Power Usage Effectiveness (PUE, which is the ratio of how much power a data center consumes overall compared to the power used by just the IT equipment) of 1.1. By implementing direct water-cooling of CPUs, GPUs, and NVIDIA NVSwitch technology, this system can sustain maximum performance without reaching thermal limits.

The Lenovo AI Fast Start service delivers live solutions to showcase GenAI business, operational, and technology results. As such, businesses can swiftly scale and turbocharge AI using full-stack NVIDIA-based technologies through Lenovo AI Fast Start for NVIDIA AI Enterprise, while new Lenovo AI Fast Start for NVIDIA NIM inference microservices provide developers with easy-to-manage containerized and optimized inference engines for NVIDIA AI Foundation models available from NVIDIA.

Section 5: Conclusions and Recommendations

AI applications, based on purpose-built accelerator technologies, are dramatically improving results across many industries. Liquid cooling technology is emerging as an essential enabler for the highest performance AI, and based on Futurum's analysis, Lenovo Neptune® liquid cooling is the industry leader in the liquid cooling realm. Lenovo's decade-plus experience with liquid cooling has led to best-in-class design, manufacturing, delivery and support.

From our perspective, the Lenovo Neptune® liquid cooling value proposition is thoroughly validated by customer endorsement in production networks. Geely Auto R&D deployed an HPC platform based on Lenovo ThinkSystem SD650 V3 servers with Neptune® water cooling, bolstering performance by 35% and creating a runway for more R&D innovation. Geely selected Lenovo to deploy an on-premises HPC cluster to avoid cloud cost and complexity. The move delivered immediate benefits as the new cluster increased HPC performance by 35% while cutting power consumption by 1 million kWh/year.

With these considerations, we make the following observations to organizations in evaluating Lenovo's Neptune® solution to fulfill their liquid cooling requirements.

Holistic Approach. Lenovo Neptune® includes direct warm water cooling of entire systems, water cooling of individual components, and liquid-assisted cooling for air-cooled systems resulting in a comprehensive approach that optimizes performance and power.

Competitive Advantages. Lenovo's 12 years' experience of delivering production-level liquid cooling solutions, including EPDM hosing, water-first, copper-based cooling loops, and use of brazed joints, delivers competitive advantages in design, manufacturing and delivery process, overall quality and support process.

Sustainability/ESG Fulfillment. Lenovo Neptune® is built to align with key sustainability objectives such as carbon reduction without compromising computing power that organizations need to prioritize in evaluating liquid cooling systems.

Important Information About this Report

CONTRIBUTORS

Steven Dickens
Chief Technology Advisor | The Futurum Group

Ron Westfall
Research Director | The Futurum Group

PUBLISHER

Daniel Newman
CEO | The Futurum Group

INQUIRIES

Contact us if you would like to discuss this report and The Futurum Group will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "The Futurum Group." Non-press and non-analysts must receive prior written permission by The Futurum Group for any citations

LICENSING

DISCLOSURES

The Futurum Group provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

ABOUT LENOVO AND NVIDIA

Lenovo brings the new era of AI-powered innovation to everyone. Our full-stack portfolio delivers powerful, flexible, and responsible AI solutions to transform industries and empower individuals. We create a future of Smarter AI for all. At Lenovo, we believe the future of AI involves the co-existence of public and enterprise AI. Lenovo brings AI to you and your data.

In partnership with NVIDIA, hybrid AI solutions are purpose built through engineering collaboration to efficiently bring AI to customer data, where and when users need it the most, advancing Lenovo's vision to enable AI for all and delivering time to market support of breakthrough technologies and architecture for the next generation of generative AI. Lenovo hybrid solutions, already optimized to run NVIDIA AI Enterprise software for secure, supported and stable production AI, also provide developers access to NVIDIA microservices, including NVIDIA NIMs and NeMo Retriever.

ABOUT THE FUTURUM GROUP

The Futurum Group is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

CONTACT INFORMATION

The Futurum Group LLC | futurumgroup.com | (833) 722-5337 |