VERTIV.

AI workloads and the future of IT infrastructure

Finding the right roadmap
to navigate uncertain times

White paper

Table of contents

Data as the new economic engine

People often say "data is the new oil." But just like crude, data needs to be refined.

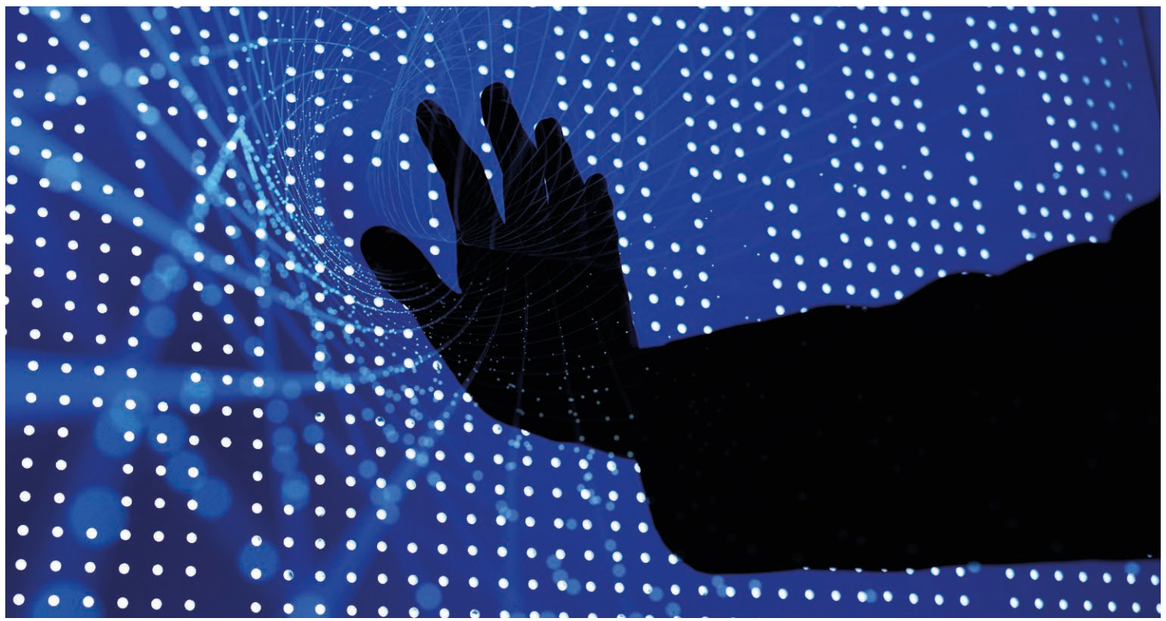
Like crude, data has the potential to transform the global economy. However, just as it took advances in refining technology to turn crude into the 20th century's engine of economic growth, artificial intelligence (AI) is likely the key to making data the basis of the 21st-century economy.

According to McKinsey & Company, generative AI alone could add the equivalent of up to $4.4 trillion annually to the global economy, including up to $100 billion in the telecommunications sector, $130 billion in media, and $460 billion in high tech.[1]

And the power of AI will not only be measured in dollars, euros, yuan, or yen. It will also be measured by breakthroughs that will impact the quality of human life. We are already on the precipice of major breakthroughs, like turning large amounts of medical data into cures for diseases like cancer, and even helping develop novel solutions to combat global climate change.

> **>$25 trillion**
>
> **The potential increase in total global GDP as the result of AI by 2030[2]**

However, unlocking the promise of AI will take an entire ecosystem to support the technology. The data center architecture and the critical digital infrastructure supporting it will have to undergo a major transformation to support the demands of AI workloads. And that ecosystem will have to be designed to make a lasting difference, not for obsolescence.



1 McKinsey Digital; "The economic potential of generative AI: The next productivity frontier," published June 14, 2023; https://www.mckinsey.com/ capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier; accessed May 30, 2024

2 ibid.

The critical digital infrastructure that enables AI

Current trends suggest that 19% of power usage in data centers will be associated with AI by 2028.[3] So, what does this mean for data centers, especially when it comes to architecture, energy usage, and cooling?

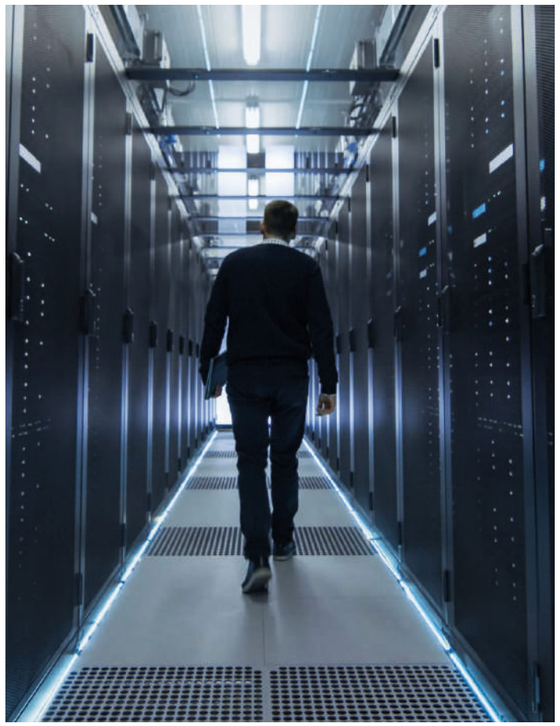
While it may not be front-page news, the meteoric rise of AI is already disrupting IT architecture and data center critical digital infrastructure. But the effect up to this point is only the tip of the iceberg. The inputs associated with AI are creating a potential bottleneck in chip size, rack weight, and data center power and cooling. Spending on critical digital infrastructure for GenAI will top $18 billion in 2024, increasing to more than $48 billion by 2027.[4]

The level of investment indicates just how profoundly the rise of generative AI will affect critical digital infrastructure. The revolution in AI will necessarily lead to a revolution in how data centers operate. So, while most people are focused on the economic benefits of AI, obtaining these benefits will require predicting, understanding, and solving for the emerging infrastructure challenges.

To fully take advantage of the biggest development in the global economy since the rise of the oil age, anyone who is relying on AI to power their growth will need to work with the right partners and formulate a plan for how to address this emerging infrastructure challenge.

> "When the metaphor of data being the new oil was first floated, it wasn't necessarily accurate. Now we have an incredible tool in AI that can mine data and unleash its hidden value. But this also means you have to be prepared for the profound changes in infrastructure necessary to support AI."

Stephen Liang
Chief Technology Officer and
Executive Vice President, Vertiv



Efficiency is crucial. But it won't be a silver bullet

AI is built on many things, but one of the key elements is microprocessors. The chips used to train AI models require a significant amount of power and generate a corresponding amount of heat.

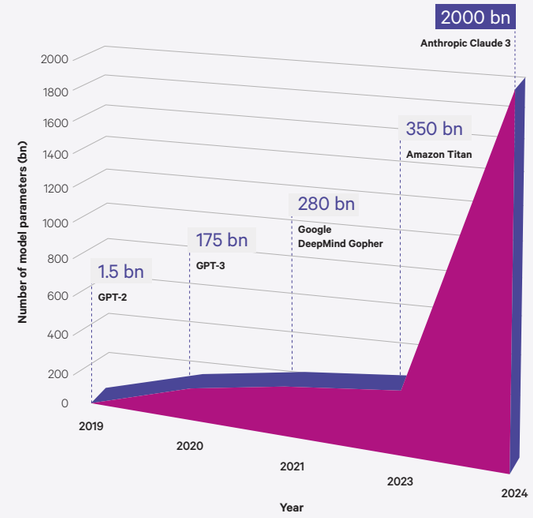
GPUs are the current chip of choice for running AI workloads and using GPUs in place of CPUs to run parallel compute workloads is 100x more efficient from a power consumption standpoint. However, while GPUs are more efficient, compute requirements will keep increasing exponentially, far outpacing any increase in chip efficiency.[5] So, while GPUs will undoubtedly produce much more compute for the same amount of power used, the demands of AI mean that the amount of total power used will still increase.[6]

Significant funds will likely be dedicated to the development of more efficient GPUs. With that amount of investment, it would be foolhardy to not expect progress on the efficiency front.

However, if the current trend continues, the models that make generative AI better will also continue to get bigger. The competition among major industry players like OpenAI, Anthropic, Google, and Meta are being driven by the use of ever-larger models, with the result that models have been doubling their computing needs every six months.[7]

Beyond exponential growth[8]

Will the past predict the future in the growth of LLMs?



The incredible output associated with AI certainly justifies the increasing power inputs, but it remains a fact that advances in AI capabilities seem to also require ever-larger models.

The demand for accuracy is partly behind the continued growth of these models, but the growth goes beyond simple precision. New functionality will likely demand even bigger models to accomplish tasks like performing multi-step work rather than just responding to prompts or determining which algorithmically available answer best meets the user's needs.

While it is true smaller models are showing promising results, the demand among users for both greater accuracy and functionality means these smaller models will likely remain niche.

At the same time the user base for AI is rapidly expanding. Individual consumers, business of all types, and even governmental organizations are driving a burgeoning demand for AI.

3 Goldman Sachs; "AI is poised to drive 160% increase in data center power demand;" May 14, 2024. https://www.goldmansachs.com/insights/articles/AI-poised-to-drive-160-increase-in-power-demand. accessed August 5, 2024

4 Villars R. et al. "Worldwide Core IT Spending for GenAI Forecast, 2023–2027: GenAI Is Triggering Hyper-Expansion of AI Spending;" IDC #US51539723, December 2023

5 Newmark; "2023 U.S. Data Center Market Overview & Market Clusters;" January 2024; https://www.nmrk.com/insights/market-report/2023-u-s-data-center-market-overview- market-clusters. accessed July 10, 2024

6 ibid.

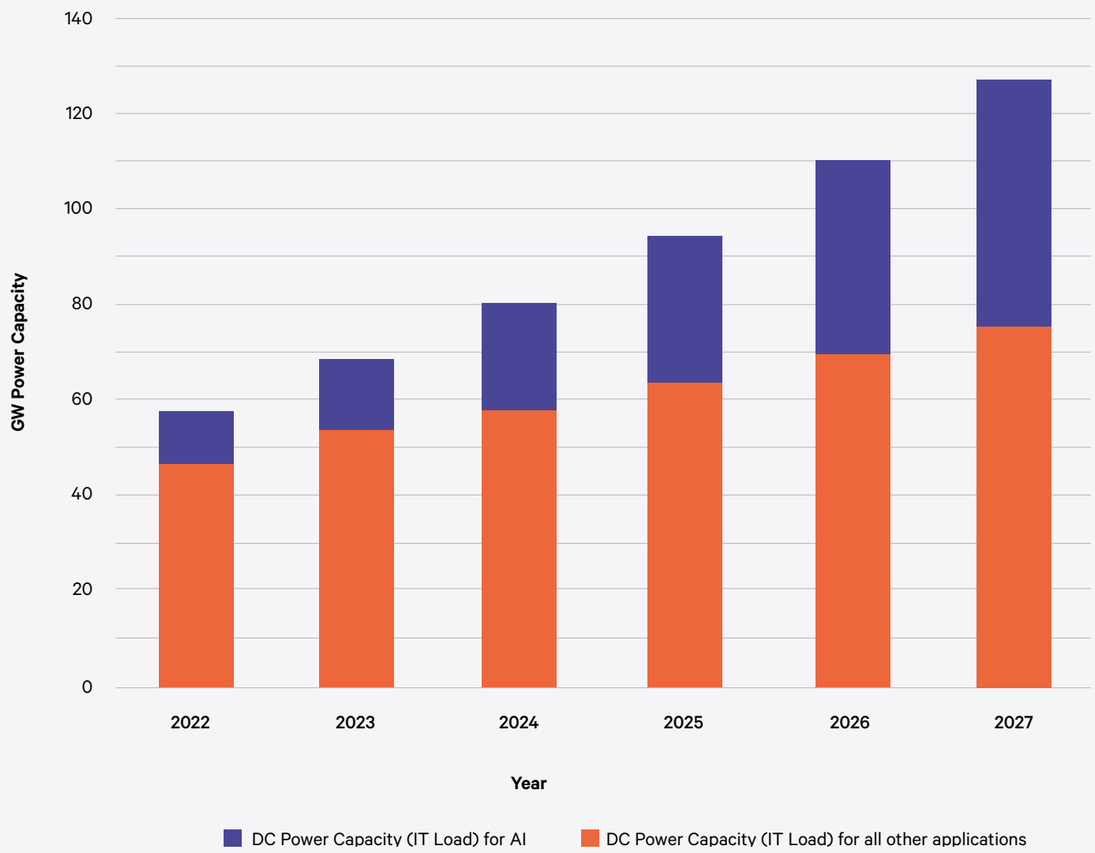
7 Epoch AI; "Notable AI Models;" June 19, 2024; https://epochai.org/data/notable-ai-models. accessed July 10, 2024

8 Adapted from The Rise and Rise of A.I. Large Language Models (LLMs) & their associated bots like ChatGPT, by McCandless, D. Evans T. and Barton P, 2024 (https://informationisbeautiful.net/visualizations/the-rise-of-generative-ai-large-language-models-llms-like-chatgpt). In the public domain

Making sense of the future

According to Jevons paradox, an increase in a resource's efficiency will generate an increase in consumption rather than a decrease. Given the many factors that seem to be driving both demand for AI and the use of larger models, it is difficult to see how chip or model efficiency gains will completely make up for the many compute-intensive developments in the AI space. In other words, the future may be more efficient, but IT load capacity will continue to increase throughout the decade despite gains in efficiency.

AI is a driving force in the growth of IT load capacity[9]

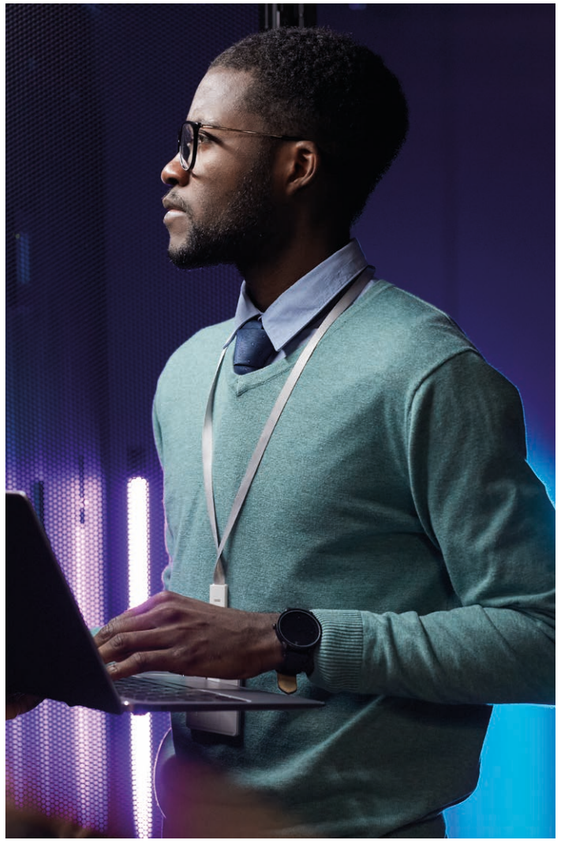


[9] Galabov, V and Sukumaran M; "Cloud and Data Center Market Snapshot – November 2023;" Omdia; November 2023

The American baseball player and amateur philosopher Yogi Berra once said, "It's tough to make predictions, especially about the future." While this may seem particularly apt with so many known unknowns in the development of AI infrastructure, based on current trends and likely technological developments it is possible to anticipate the greatest challenges to AI deployment and usage, especially within the data center context. So, what are the most important elements to follow if you are interested in making your operations future-ready?

Specifically, the following trends should be on everyone's radar.

- **Changes in output capacity:** We need to be prepared for a world where output capacity grows. Rack loads, PDUs, UPSs, switchgear, and CDUs will all increase in capacity and have a corresponding knock-on effect throughout the overall infrastructure.
- **Changes in overall scale:** Arguably, the most obvious knock-on effect will involve facilities also getting bigger. Today's 3 MW blocks will be tomorrow's 20MW blocks.
- **The quest for always-on power:** Bigger also means more power and heat. And more heat means hybrid and liquid cooling are the future. Since the GPUs of the future must be continuously cooled, always-on power is an imperative.
- **AI means unconventional load profiles:** AI is associated with power loads that can pulse from a 10% idle to a 150% overload in a flash.
- **Pressure to refresh and upgrade:** Data center operators will need to determine the best way to retrofit their existing operations to meet the demands listed above.



"I think the architecture is going to change in important ways, and a crucial part of being successful is anticipating these changes. For example, it's important to ask, 'How do we plan for 250 kilowatt racks or greater?' The way people do it currently is unworkable, so the time to think about change is now."

Gregory Ratcliff
Chief Innovation Officer, Vertiv

Major changes are coming in architecture, power, and cooling

A summary of the biggest development this decade

Current trends point to major changes coming in the overall IT infrastructure. But what specific changes will we see and when will these happen?

Determining the answers to these questions begins at the most basic level. As GPU architecture, rack architecture, and row architecture evolve, we will see a corresponding evolution in both power and cooling.

While no one can be truly certain about the exact future of IT architecture, looking at the basic drivers of change is the best way to determine what that future will look like.

	2024	2025	2027**	2029+**
	Currently shipping AI GPU designs primarily based on air-cooling or hybrid liquid and air-cooling for greater operating efficiency. [10]	Next-release AI GPU designs primarily based on one-phase liquid-cooling to the chip and air-cooling for residual heat load.[11]	Next-generation (not announced but projected) AI GPU designs primarily based on increased flow and lower temperature for increased performance.[12]	Future-generation (projected trajectory) AI GPU designs primarily based on liquid-cooling to the chip or some form of contained immersion.[13]
GPU Architecture	Up to 32x NVIDIA Hopper H100 GPU cores (or similar) per rack, configured in row clusters /pods up to 240x GPU cores.[14]	Up to 72x NVIDIA Blackwell B200 GPU cores (or similar) per rack, configured in row clusters/pods up to 720x GPU cores.[15]	Next-generation GPU cores (12-month release-cycles), in similar row cluster/pod configurations.	
Rack Architecture[16]	Up to 50kW per rack with 400/230VAC N+1 power distribution. Cooled using liquid-to-air rear-door heat exchangers.	Up to 140kW per rack with 400/230VAC or 400VDC N+1 power distribution. Cooled using one-phase liquid-to-chip cold-plates (80% of heat load), augmented by liquid-to-air rear-door heat exchangers (20% of heat load).	Up to 300kW per rack with 400/230VAC or 400VDC N+1 power distribution. Cooled using liquid-to-chip cold plates, augmented by liquid-to-air rear door heat exchanger.*	Densities up to 1MW per rack. 480/277V AC, 600/347V AC or 800V DC by 2031 should be considered in the design. 400/230VAC or 400VDC N+1 power distribution.* Cold-plate cooling (100% of heat load) should be considered in the design.

*Rack form factor will likely change in size and configuration from 2027 onwards and will likely not correspond to the EIA and OCP versions used today.

	2024	2025	2027**	2029+**
Row Architecture[17]	Overhead rigid busway @ 400A with tap-boxes, providing maximum flexibility for drop-in replacement and GPU refresh, is recommended. Field-wired dedicated circuits from PDU can be considered for cost. Row-based 600kW cooling distribution units and perimeter computer-room air-handlers.	Overhead rigid busway @ 2000A with tap-boxes, providing maximum flexibility for drop-in replacement and GPU refresh, is recommended. Field-wired dedicated circuits from PDU are not recommended due to cost of installation and inflexible wire gauge. Row-based 1350kW cooling distribution units and perimeter computer-room air-handlers.	Overhead rigid busway @ 4000A with tap-boxes, providing maximum flexibility for drop-in replacement and GPU refresh, is recommended. Field-wired dedicated circuits from PDU are not recommended. Row-based 3000kW cooling distribution units and perimeter computer-room air-handlers.	Overhead rigid busway @ 6000A and up with tap-boxes, providing maximum flexibility for drop-in replacement and GPU refresh, is recommended. Field-wired dedicated circuits from PDU are not recommended. Multi-MW cooling distribution units.
Power Management	Minimum Tier II or Tier III utility feed supported by N+1 standby power generation with onsite Battery Energy Storage Systems (BESS) or similar should be considered for grid interaction, arbitrage and resiliency.		Minimum Tier II or Tier III utility feed supported by N+1 standby power generation with Distributed Energy Resources (DER), onsite Battery Energy Storage Systems (BESS) or fuel-cells should be considered for fully dynamic grid interaction, arbitrage and resiliency. Integrated high-speed power monitoring at the tap-box (busway) with intelligent thermal control is required (Vertiv proprietary).	
GPU Workload	Up to 160%	Up to 160%	Up to 160%	Up to 160%
	We are assuming initial GPU deployments to be >20% of total workloads for enterprise up to 2025 and growing to 100% of total load by 2027.			
Thermal Cycle[18]	Direct expansion (DX) and chilled water (CW).		High-temperature chilled water loop.	Chilled water loop.
Heat Reuse	Use organic Rankine cycle (ORC), direct air capture (DAC), and absorption chillers.			
	More than 25% heat reuse.	More than 25% heat reuse.	More than 25% heat reuse.	More than 80% heat reuse.
Controls	Capture operating telemetry in datalake and implement condition-based maintenance (CBM).		Implement data center-wide digital twins and leverage continuous telemetry inputs for optimization of operation using AI.	Self-optimization and autonomous robotic operation.

**The technical and management details, estimations, and projections cited here are based on Vertiv research, internal expertise, and currently available information from Vertiv partners in the industry. Analyzed forecasts are based on the available design and infrastructure technologies as of this white paper's date of publication. Projections for illustrative purposes only.

The upward trajectory in rack density triggered by the current wave of AI will not be a historical blip. To meet the growing demand for AI, rack architecture will increase from the upper 30kW to 300-600kW densities in the near-term and possibly 1MW and above by 2030.

Of course, there will remain a level of uncertainty around this number based not only on technological developments, but also on the demand for AI. However, it is reasonable to expect that the future will be based on the growing rack density required by AI and the technological improvements in critical infrastructure necessary to meet these demands.

> "Things are changing faster now than at any time in the last 30 years. To manage this change we need to work together as an industry to establish best practices and reference designs. It's key to being prepared for the changes that are already happening."
>
> **Peter Panfil**
> Vice President, Global Power, Vertiv

[10] NVIDIA; "NVIDIA H100 Tensor Core GPU Architecture;" https://resources.nvidia.com/en-us-tensor-core/gtc22-whitepaper-hopper?ncid=no-ncid; accessed August 12, 2024

[11] NVIDIA; "NVIDIA Blackwell Architecture Technical Brief;" https://resources.nvidia.com/en-us-blackwell-architecture?ncid=no-ncid; accessed August 12, 2024

[12] Trueman, C. "Stack Infrastructure to support AI workloads requiring up to 300kW-per-rack will be achieved through closed-loop water cooling systems;" Data Center Dynamics; January 8, 2024; https://www.datacenterdynamics.com/en/news/stack-infrastructure-to-support-ai-workloads-requiring-up-to-300kw-per-rack/; accessed August 14, 2024

[13] ibid.

[14] NVIDIA; "NVIDIA H100 Tensor Core GPU Architecture;" https://resources.nvidia.com/en-us-tensor-core/gtc22-whitepaper-hopper?ncid=no-ncid; accessed August 12, 2024

[15] NVIDIA; "NVIDIA Blackwell Architecture Technical Brief;" https://resources.nvidia.com/en-us-blackwell-architecture?ncid=no-ncid; accessed August 12, 2024

[16] Wang, S. "2024 Trends to Watch: Data Center Physical Infrastructure;" Omdia; December 22, 2023; https://omdia.tech.informa.com/om033896/2024-trends-to-watch-data-center-physical-infrastructure; accessed August 12, 2024

[17] ibid.

[18] Wang, S. "AI-driven Data Center Cooling Systems and Technologies;" Omdia; August 2, 2024; https://omdia.tech.informa.com/blogs/2024/aug/ai-driven-data-center-cooling-systems-and-technologies; accessed August 15, 2024

While there may be uncertainty around how quickly rack architecture will evolve, especially in the longer term, there is little doubt that changes in architecture will require new approaches to cooling. Current AI GPU designs are mostly based on air-cooling.

However, the tremendous growth in power usage will mean various forms of hybrid and liquid cooling are the future. Liquid cooling is significantly more efficient than air cooling, and that efficiency is extremely important to developing the infrastructure necessary for AI. By the end of the decade, we will see data centers primarily rely on liquid-cooling to the chip, self-contained immersion and air-cooling for residual heat loads.

> “The rise of liquid cooling should raise quite a few questions for those who run data centers. For example, ‘How do I make sure the liquid loop is stable? How do I make sure it is redundant? How do I put a rack into the system? How do I take a rack out of the system?’ These are all things people need to think about with the changes that are coming.”

Steve Madara
Vice President, Global Cooling, Vertiv

Cooling is not the only concern. The data centers of the future will also have to be flexible enough to deal with “pulse loads,” or periodic loads with high power inrush in a very short time. AI training clusters are known to exhibit short-duration spikes in electrical current during certain compute cycles, so overload provisions will become de rigueur.

But data centers likely have a bigger headache than just spikes in current when it comes to power. In the United States alone total electricity consumption is expected to grow based almost entirely on demand from data centers. According to data from Goldman Sachs, there will be a 15% compound annual growth rate in data center power demand through 2030, with data centers making up 8% of total US power demand by the end of the decade, up from about 3% currently.

> “One of our roles today is making what power is there more available and ensuring it is used more efficiently at the data center. You need to make sure you use the power you have effectively.”

Stephen Liang
Chief Technology Officer and Executive Vice President, Vertiv

Although it is difficult to pinpoint all the exact technological milestones that will occur this decade and beyond, the broad brushstrokes of the future are becoming more clear. We know rack densities will increase. We know cooling methods and management will need to evolve. We know power and load management are quickly becoming major challenges.

Now we need to apply what we know to inform the development of products and solutions that will resolve the needs of data centers both today and tomorrow. And we need to do it quickly if we want to realize the full potential of AI.

For example, we should plan for increased rack densities by developing high-density rack power distribution units for managing power distribution more effectively in quickly evolving data centers. We should also revisit the existing technologies for liquid cooling and ask ourselves how we can be ready for a world where two-phase liquid-to-chip cold-plate cooling is a necessity. And we should look at actual AI load profiles and develop ways to optimize performance based on the profiles we are seeing now and what we anticipate seeing as the AI revolution marches forward.

The key to adapting? Leveraging tech and building good partnerships

But accomplishing all these things is not necessarily easy. While preparing yourself for the changes being brought about by AI is crucial, it is equally important to recognize that any attempt to make your operations future-ready cannot be done in a vacuum.

NVIDIA CEO Jensen Huang effectively summed up the situation when he said, “AI (is) not a chip problem. It’s a reinvention-of-computing problem. You can’t solve this new way of computing by just designing a chip. Every aspect of the computer has fundamentally changed.”

How we support and leverage AI is also changing. Already we are seeing the rise of the AI factory, a centralized hub that allows experts, data scientists, and engineers to collaborate on developing, deploying, and scaling AI-based solutions.

As AI grows so will the factory model. Functioning similarly to traditional factories but creating actionable intelligence and new insights instead of physical products, the model needs complex infrastructure and a holistic design to work.

That effectively means no single player in the space has all the answers. Therefore, thriving in an AI-dominated world will mean going beyond just solving a single problem with power, cooling, or chips. Instead, the smart operators will see how the problems they face fits into the overall ecosystem.

That is why being future-ready will require developing strong relationships with knowledgeable partners who are not just watching the changes brought about by AI but are driving those changes.

Even for the leading firms in the AI space, bringing together expertise from the most experienced players is an important part of maximizing AI’s benefits. Integrating the perspectives of everyone from chip developers to power managers is key to making smart predictions about the demands of AI and having the capacity to meet those demands.

What does this look like in practice? Because elements like innovative chips will have an enormous effect on the future architecture and operation of data centers, Vertiv has been working closely with leading AI chipmakers, serving as consultant on rapidly developing chip technology and architect for the infrastructure necessary to house these chips. As part of various global programs involving tech innovators who offer solutions built on or powered by technologies, we can help ensure that the critical digital infrastructure and GPUs work together to deliver effective performance.

And by offering our expertise to networks that also include leading software vendors, cloud service providers, solution providers, and system integrators, we are strengthening the work we do today and preparing for changes in the future by integrating Vertiv’s full portfolio of power and cooling solutions into the latest technologies from across the sector.

Because in today’s market a deep bench of unmatched technical expertise is very important, but that expertise needs to work in conjunction with other parts of the developing AI ecosystem if you truly want to future proof your operations.

> “Given the unprecedented changes data centers are beginning to face with the rise of AI, I said to myself, ‘I think we need to be calling on the chip guys and specifically the disruptive chip guys. We need to learn what are they going to make and build that's going to disrupt our ecosystem.”

Greg Stover
Global Director, Hi-Tech Development, Vertiv

What will the future of IT infrastructure look like? Here's what leading experts are saying

The need to support AI is already ushering in major changes in data center infrastructure. However, the biggest developments are yet to come. Over the remainder of the decade, AI will be the driving force behind a cascade of changes, with new GPU architecture, rack architecture, and row architecture all contributing to significant changes in IT infrastructure and critical digital infrastructure. What will it take to leverage the changes and seize the opportunities presented by the rise of AI?

Powering the future means looking beyond the obvious

Peter Panfil
Vice President, Global Power, Vertiv

Peter leads strategic customer development for the Vertiv power business. He is skilled at solving customer challenges with the latest power and control technologies delivering availability, scalability, and efficiency levels to meet diverse customer and sustainability needs. Peter has deep knowledge in power solutions to support the unique needs of AI applications, and he is an advocate of the "Bring Your Own Power" approach to solving utility dependence issues. With more than 30 years in the critical infrastructure space, he has held executive positions including VP Engineering and VP/GM AC Power prior to his current responsibilities. He is a frequent presenter and spokesperson for industry trade shows, conferences and media outlets serving the IT, facilities and engineering industry, and a published author via his contribution to the 2024 book "Greener Data Vol. 2".

"

For a long time we've lived happily in the data center world with three-MW blocks directly coupled to a three-MW generator at 4,000-amp bus. Those days are ending. With the rise of AI, it is common now to start talking about 10-MW blocks and even 20-MW blocks.

So, when it comes to power use in data centers, it's no secret that we are looking at a massive increase in size and scale. The bottom line is: Operators will need to figure out where that power is going to come from.

But just as importantly, AI will also result in spiky loads associated with training generative AI and a demand for continuous power associated with cooling needs. These last two factors are relatively recent needs and will be a growing challenge across the industry in the years to come.

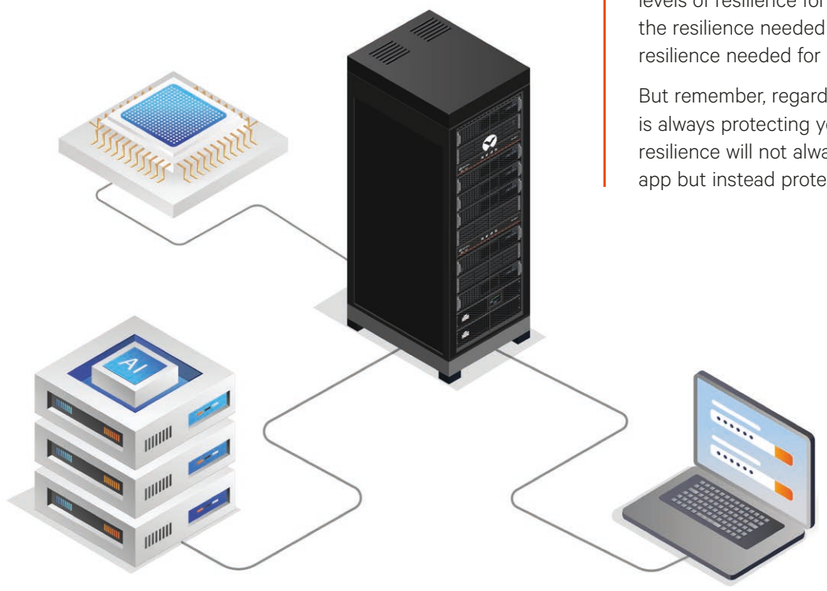
So, what is the secret to meeting these power challenges? I think a big part will be anticipating the more specific changes in power. You need to know where the chip makers are headed, and you need to know what the IT industry insiders think.

When it comes to power solutions, the innovators are already asking questions like 'Can I simplify this and deploy it easily and at low-cost?' and 'Can I be sure this solution is reliable?' The answers in my mind are definitely 'Yes.' However, it's important to understand that getting to 'yes' will involve working with the right people; people who understand both the latest chips and the latest power solutions."

Use resilience to protect your assets

Stephen Liang
Chief Technology Officer and Executive Vice President, Vertiv

Stephen Liang aligns Vertiv's technology strategies and resources to prioritize the voice of the customer in product development and research. He has three decades of experience in power supply solutions, manufacturing operations, and R&D. With a bachelor's and a master's degree in mechanical engineering from the Massachusetts Institute of Technology, he combines both technical knowledge and long-standing experience in his vision for how technology is changing IT infrastructure.



"

I have a long experience with IT infrastructure and can say without a doubt that technological change is happening faster than ever in the sector. One reason is because we expect the technology to do more than ever before and are asking the overall infrastructure to do the same.

We have seen an immense investment in valuable next-generation computing chips and server systems, much of it related to AI. This investment means we have more riding on keeping these systems up and running than ever before. To get the most out of these assets, our customers will need them to run 24/7. Therefore, I think one of the biggest watchwords for the future is resilience. Obviously, having a resilient IT infrastructure has always been a priority. But in the future resilience will be more important, and more complex, than ever.

What do I mean by that? At Vertiv, we have a long track record of forming partnerships with chip developers and manufacturers to gain insights, all while providing continuous power and continuous cooling; it's in our DNA. We can see how AI demands often require different levels of resilience for different functions. For example, the resilience needed for training is different from the resilience needed for inference.

But remember, regardless of your end goal, the key is always protecting your investment. In the future, resilience will not always be about protecting the app but instead protecting the asset value."

Learn to leverage what you have

Martin T. Olsen
Senior Vice President, Global Product Strategy, Vertiv

Martin Olsen is responsible for global product strategy and planning. His work drives roadmap interoperability of products to deliver complete solutions, investment scope, evaluation, and prioritization across Vertiv's global business units. With experience in various leadership positions in engineering, strategy, and planning, he is an expert on global critical infrastructure in the data center industry. Martin holds several thermal US patents, and earned his Bachelor of Science in Marketing and Business Analytics from University of Syddanmark, Kolding, Denmark and completed the Harvard Business School alumni-status General Management Program.

“

We are living in a world where the rise of AI means rack densities could be 1MW per rack by the end of the decade. For those who already have significant infrastructure investments in place, they need to be thinking about how they can
make the most out of these investments while still being prepared for a new scale of demands in the future.

In other words, in an environment that is rapidly changing, how do you leverage what you already have in place without risking becoming obsolete?

Many data center operators will need to consider retrofitting their assets. But as rack densities continue to grow at an accelerated rate, simply deciding upon a retrofit will not be enough. Instead, you will need to consider what densities you are retrofitting for and what cooling technologies will be necessary as part of your retrofit.

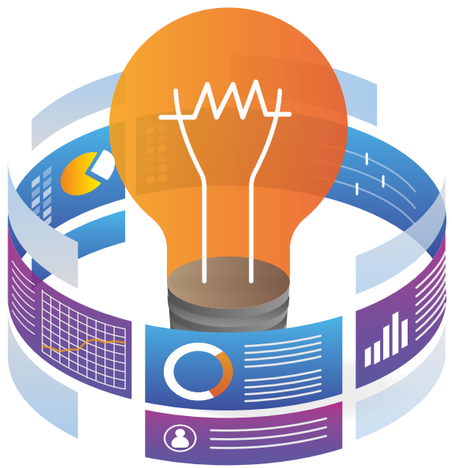
Deploying and using AI in business is very transformative, and customers are acutely aware of the phrase "Disrupt or be Disrupted" by John Chambers, former CEO and Chairman Emeritus of Cisco Systems. They want to do things in a way that differentiates them from their competitors. When it comes to thinking about a retrofit, that means making a decision only after seeing all the options in front of you and considering how these options can work together intelligently.

That can be a tall order. After all, how do you know what all your options are? What's the best way to see both the details and the big picture at the same time? In my mind, the element that will make a data center operator truly ready for the future will be finding partners that not only have broad knowledge of emerging technologies, but also how these technologies can work together to meet future demands. Without that partner, I think data center operators risk watching the future pass them by.”

Where to look for the future's most innovative products

Greg Ratcliff
Chief Innovation Officer, Vertiv

Greg Ratcliff is a 30+ year industry veteran with an emphasis on data-driven innovation. He led the global advanced analytics group for the services team, specializing in real-time data and connecting Vertiv's nearly 1 billion operating products to the Vertiv cloud. Greg builds relationships with technology partners, universities, and Vertiv engineering teams to ensure the company is future-focused and ready to support change in the industry. His educational background includes undergraduate degrees in Applied Mathematics and Computer Engineering with minor studies in Electrical Engineering from The Ohio State University, an MBA degree from Phoenix University, and specialized training focused on Agile project management of IoT and Big Data projects from Liberty University.



“

I like to think about R&D in three contexts: Now, near, and far. The 'now' is where you find products you can use today. These are essential to maintaining contemporary IT infrastructure.

But any company wanting to keep pace with the rapid changes already underway can't afford to ignore the 'near' and 'far.' The technologies essential to IT infrastructure will be significantly different from what we employ today, which means keeping one eye on the developmental pipeline is an essential part of being prepared for the future.

In my world, one way we prepare for the future is by having active relationships with dozens of universities. Some of these relationships are formal, like The Center for Energy-Smart Electronic Systems (ES2). ES2 is a National Science Foundation/Industry/University Cooperative Research Center (I/UCRC) established to develop tools to maximize energy efficiency for the operation of electronic systems, including data centers, 'from the chip to the cooling tower.' Others are less formalized, and very valuable, relationships with academia.

But these relationships with universities don't mean we live in an ivory tower. In fact, Vertiv's end-use customers are often an important part of these relationships. I like to think of it as a three-legged stool that supports practical innovation. Innovations may come from a university lab, but Vertiv and our customers play an essential role in turning the innovation into a product ready to meet the demands of the wider market.

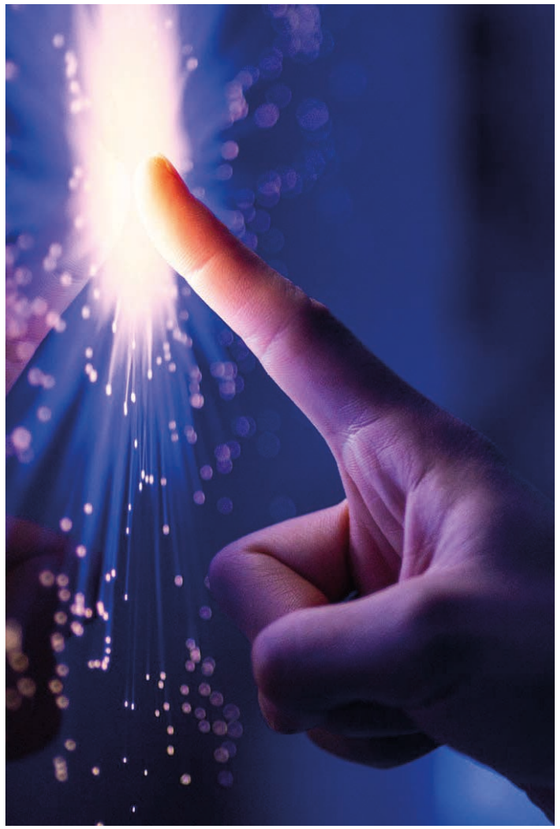
And innovations that turn into real products will be an essential part of the evolution of critical digital infrastructure. If you want to be ready for the future, partner with a company that has deep ties with tech leaders and innovators. Because in today's environment you may be focused on the 'now,' but you also need to keep an eye on the 'near' and 'far.'”

Where to go from here?

The rise of AI has made this truly an exciting time in the world of computing. But with excitement and possibilities come questions and uncertainty. Many entities with data centers are scrambling to add AI capacity to existing operations. At the same time, those planning new data centers are wondering exactly how much AI capacity they should include in their blueprints. And what will the rack architecture look like? How about power and cooling needs?

Answering these types of questions will be difficult, but one way to make it easier is to ensure you are designing your critical digital infrastructure with leaders in AI development, not followers. Because in a world where change is both rapid and certain, the incorrect approach could be an investment in a solution that becomes out of date shortly after you deploy it. By being smart and deliberate, you can avoid designing into obsolescence and instead create something more sustainable, reliable, and long-lasting.

Thriving in the AI revolution means choosing your solutions wisely, but it means being even wiser about who provides your solutions. You not only need the right technology, but also a partner with a broad understanding of the end-to-end power train and thermal chain, and with experts who can advise on powering your AI needs both today and tomorrow.



> “Your ideal partner is one with a broad portfolio of power and cooling technology, with a large service network and global footprint. Why? Because you need a partner that can enable you to be transformative, and to be transformative at this stage in the game with AI you need both speed and scale.”

Martin Olsen
Senior Vice President, Global Product Strategy, Vertiv

References

Adapted from The Rise and Rise of A.I. Large Language Models (LLMs) & their associated bots like ChatGPT, by McCandless, D. Evans T. and Barton P, 2024 (https://informationisbeautiful.net/visualizations/the-rise-of-generative-ai-large-language-models-llms-like-chatgpt). In the public domain

Epoch AI; “Notable AI Models;” June 19, 2024; https://epochai.org/data/notable-ai-models, accessed July 10, 2024

Galabov, V and Sukumaran M; “Cloud and Data Center Market Snapshot – November 2023;” Omdia; November 2023

Goldman Sachs; “Generational growth: AI, data centers and the coming US power demand surge;” published April 28, 2024; https://www.goldmansachs. com/intelligence/pages/gs-research/generational-growth-ai-data-centers-and-the-coming-us-power-surge/report.pdf; accessed May 30, 20024

Goldman Sachs; “AI is poised to drive 160% increase in data center power demand;” May 14, 2024. https://www.goldmansachs.com/insights/articles/AI-poised-to-drive-160-increase-in-power-demand. accessed August 5, 2024

McKinsey Digital; “The economic potential of generative AI: The next productivity frontier;” published June 14, 2023; https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier; accessed May 30, 2024

Newmark; “2023 U.S. Data Center Market Overview & Market Clusters;” January 2024; https://www.nmrk.com/insights/market-report/2023-u-s-data-center-market-overview- market-clusters, accessed July 10, 2024

NVIDIA; “NVIDIA H100 Tensor Core GPU Architecture;” https://resources.nvidia.com/en-us-tensor-core/gtc22-whitepaper-hopper?ncid=no-ncid; accessed August 12, 2024

NVIDIA; “NVIDIA Blackwell Architecture Technical Brief;” https://resources.nvidia.com/en-us-blackwell-architecture?ncid=no-ncid; accessed August 12, 2024

Trueman, C. “Stack Infrastructure to support AI workloads requiring up to 300kW-per-rack will be achieved through closed-loop water cooling systems;” Data Center Dynamics; January 8, 2024; https://www.datacenterdynamics.com/en/news/stack-infrastructure-to-support-ai-workloads-requiring-up-to-300kw-per-rack/; accessed August 14, 2024

Villars R. et al. “Worldwide Core IT Spending for GenAI Forecast, 2023–2027: GenAI Is Triggering Hyper-Expansion of AI Spending;” IDC #US51539723, December 2023

Wang, S. “2024 Trends to Watch: Data Center Physical Infrastructure;” Omdia; December 22, 2023; https://omdia.tech.informa.com/om033896/2024-trends-to-watch-data-center-physical-infrastructure; accessed August 12, 2024

Wang, S. “AI-driven Data Center Cooling Systems and Technologies;” Omdia; August 2, 2024; https://omdia.tech.informa.com/blogs/2024/aug/ai-driven-data-center-cooling-systems-and-technologies; accessed August 15, 2024

Vertiv.com | **Vertiv Headquarters,** 505 N Cleveland Ave, Westerville, OH 43082, USA

SL-71324 (09/24)