The AI Factory:
Data Centers at the Heart of the Action

Steven Carlini
Chief AI Advocate,
AI and Data Centers

Schneider Electric

Foreword

Since the introduction of ChatGPT less than four years ago, it would have been hard to imagine the ways AI in all its forms – industrial, physical, generative, and agentic – would permeate our daily lives. Although we are at the start of AI's global journey, we have witnessed it improve both the effectiveness and efficiency of systems and processes in business and government.

Like the internet, which was the last great technology wave, AI depends on IT hardware and application software running in data centers. Unlike the internet buildout when data centers were in the background, the AI buildout places data centers front and center – at the heart of the action. Data centers now make headlines and are a hot topic from cocktail parties to congressional hearings.

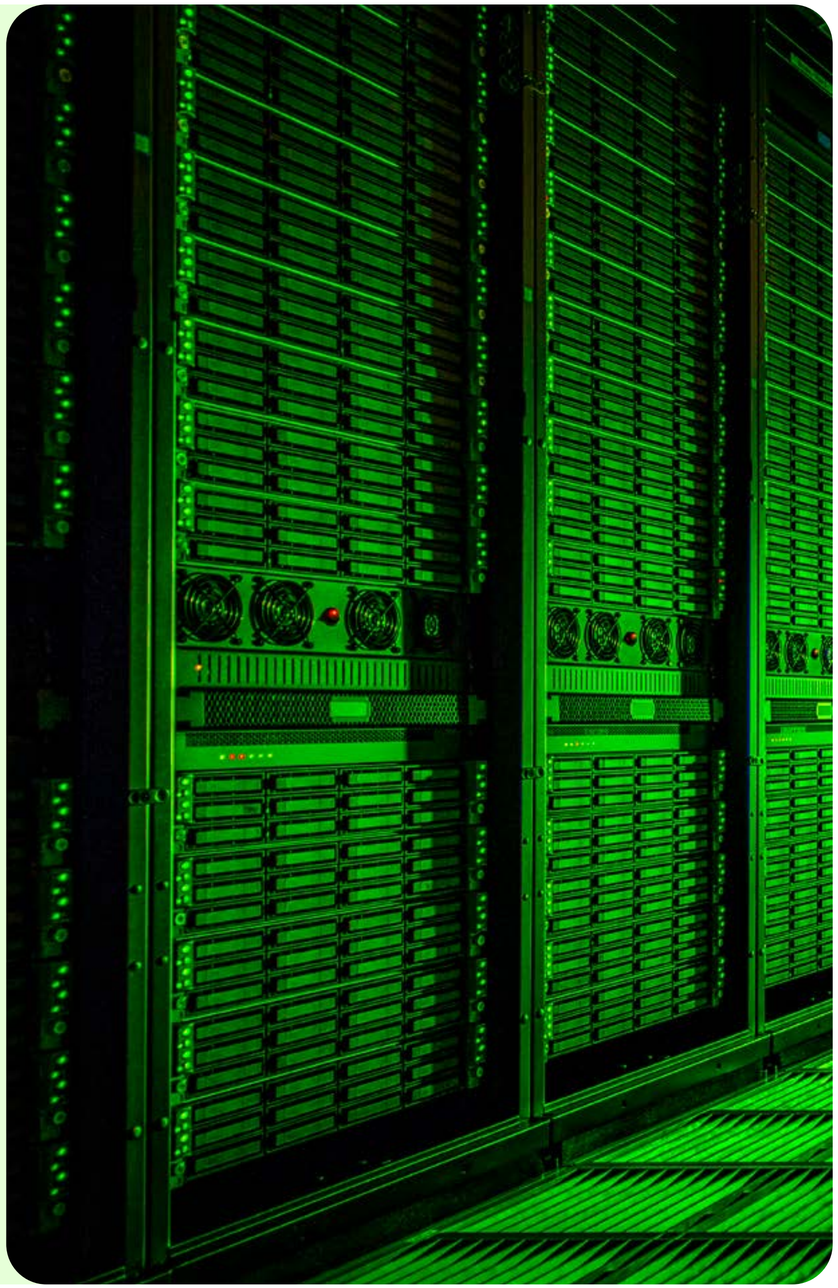
Some people are concerned about AI moving so quickly and AI itself faces many challenges. I find AI a fascinating technology. In this ebook, I share some intriguing subjects concerning AI and data centers and provide analysis and commentary.

Topics include AI as a five-layer cake and a metric called tokens per watt that quantifies the work of AI factories. I explain why liquid cooling these data centers is harder than it looks, examine how data centers are addressing utility power challenges, and preview innovation that will take AI factory racks to 1MW using 800 volts DC while digging into how the industry is addressing labor shortages.

Let's look at the engine that is empowering AI growth: **data centers**.

Steven Carlini
Chief AI Advocate,
AI and Data Centers
Schneider Electric

Table of contents



Ways To Reduce The Environmental Impact Of AI Factory Data Centers At Every Layer

AI factories are built on a five-layer architecture — energy, chips (accelerated compute), data centers, models, and applications — and each layer influences performance, efficiency, and sustainability.

Schneider Electric Chief AI Advocate Steven Carlini echoes NVIDIA Founder and CEO Jensen Huang's comparison of AI to a 5-layer cake and takes the analogy one step further.

After breaking down the layers to reveal how they are interdependent from the base through the fillings to the icing atop the AI cake, Carlini dives into each layer to reveal the path to more sustainable AI factory operations.

Businesses, governments, and people across the globe are driving the demand to increasingly automate our world. As a result, AI is being evolved and scaled but many don't understand what is involved. To help illuminate how AI works, NVIDIA Founder and CEO Jensen Huang has described the technology stack needed for AI as a 5-layer cake. His analogy encapsulates how each tech layer is interdependent on the layer below it. Here is a quick breakdown of the five layers in the AI cake:

1 **Energy (The Base):**
Electricity is the fundamental AI fuel

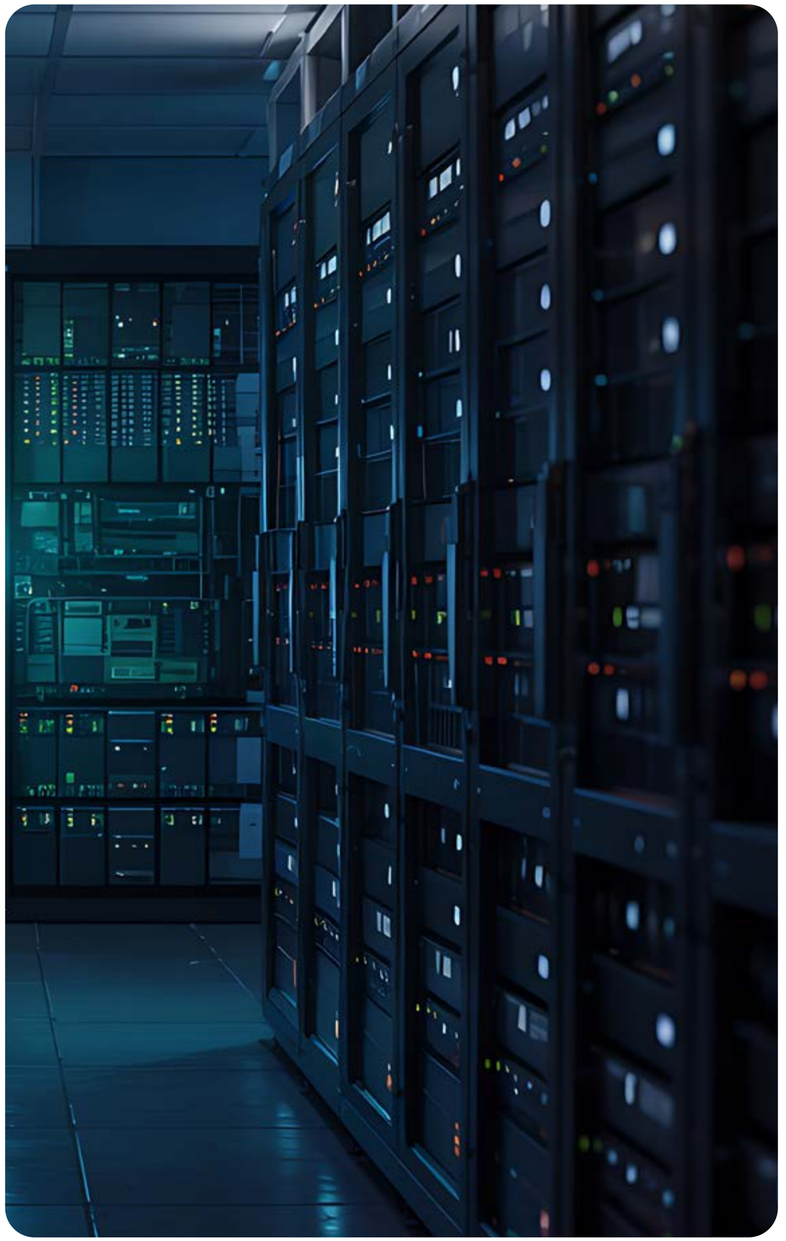
2 **Chips (Accelerated Compute):**
GPUs and specialized accelerators that perform massive calculations

3 **Data Centers (Facilities):**
Physical buildings, racks, cooling, power back-up, storage and distribution with orchestration

4 **Models (AI Core):**
Foundation models (LLMs, etc.,) frontier models (most advanced) as well as open source (publicly accessible)

5 **Applications (The Top):**
The top layer where content generation and problem solving is delivered through agents

I believe every layer of the AI cake plays a significant role not only when it comes to delivering AI but also in the impact to the environment. I want to build on Huang's creative analogy by revealing how to boost sustainability at each level. The time is right because: energy generation from fossil fuel plants has high CO2 emissions; today's GPUs use more electricity in every generation; data centers can have inefficient and energy-wasting power distribution and cooling and high carbon emission generators; and AI models are not optimized to answer complex queries and applications are immature.



Deep diving into each layer to find out how to mitigate the environmental impact

I want to dive into each of the layers and ask: “What can be done to help reduce the environmental impact?”

Energy (The Base)

First of all, data center operators are relying on voluntary net-zero commitments and renewable electricity adoption targets to decarbonize their operations while direct regulatory push remains limited. As an industry, data centers are the largest investors that are funding low carbon power generation through power purchase agreements (PPAs) and renewable energy credits (RECs). These deals fund clean electricity for new wind and solar projects.

Secondly, to lower emissions, I recommend locating the data center adjacent to a zero Scope 2 power plant such as a hydro plant or a nuclear plant. In addition to eliminating Scope 2 carbon emissions, this step also avoids about 6% transmission and distribution (T&D) losses and the associated carbon emissions. In the future, the industry will leverage small modular reactors (SMRs) either on premise or adjacent. Many people forget that data centers require redundant power sources. I view this requirement as an opportunity for grid operators to leverage ADMS (Advanced Distribution Management System) software to create smarter, more resilient grids featuring automated grid analysis for optimized workflows and improved efficiency. This AI enhances the integration of renewables by analyzing real-time data for better decision-making, optimizing the grid operations to minimize carbon emissions.

Chips (Accelerated Compute)

In the AI era, leading chip companies are launching new generations every year, which is almost a double power use profile. However, these new generations of GPUs deliver 10-100 times more tokens per watt for specific AI inference workloads. It is important to keep in mind that more generations of these GPUs are planned so the inflection point where the performance satisfies the need appears to be far in the future.

Data Centers (Facilities)

The main job of the data center is to house the accelerated compute and enable it to run at its potential with as little electricity use as possible. The data centers can minimize the environmental impact of the facility by starting with green cement and steel. Accelerated compute will also migrate 800 VDC power distribution to the servers starting in 2027. It will use a single-step AC/DC conversion resulting in fewer transformer losses and a more direct power flow as well. It will cut copper usage by up to 45% compared to traditional architectures.

For cooling, data center operators spend an estimated $2.5M per megawatt (MW) per year, which amounts to around $1M spent annually on cooling-related energy in a legacy data center. Switching from air-cooled servers to direct-to-chip (DTC) liquid cooled servers will result in overall data center energy decreases of 6.5% and a cooling system energy reduction up to 67%.

Running liquid cooling at higher water temperatures can mean less energy used for cooling but it risks running chips closer to their thermal limit. Water temperatures in liquid-cooled systems today for AI factories seem to be converging around 80-86°F (27-30°C) range, not the 104-122°F (40-50°C) that may be possible if the data center is located in a colder climate.

Continues onto the next page

The colder the climate, the more efficiently a data center's cooling system operates by allowing for more chiller economizer hours or use of a dry cooler. With higher water temperatures, heat reuse can be another benefit of liquid cooling. Instead of rejecting the data center's waste heat to the outdoors, the heat energy is redirected to applications such as district heating systems, industrial processes, and greenhouses.

A two-four hour lithium-ion battery energy storage system (BESS) can transform a data center into a distributed energy resource (DER) living on the grid. Along with acting as a backup power for the data center during extended outages, they can also be used to lower carbon emissions. When this BESS system is charged with renewable energy (wind, solar, hydro, nuclear), it emits zero carbon during operation. And when there is a surplus in renewable supply, instead of curtailing renewable production, this surplus energy is used to charge the BESS.

Models (AI Core)

AI models (large and small) are ongoing in the need to create new algorithms that enhance efficiency, accuracy, and scale. If the models require less "brain power" or fewer processing cycles, they will use less energy and emit less carbon. AI will evolve and have better reasoning and problem-solving capabilities including agentic. This is where models can start acting unchecked, leveraging breakthroughs like unsupervised learning and reinforcement learning. While "thinking" takes longer up front, it will lead to faster, higher-quality final outputs.

Applications (The Top)

As we are still in the initial rollout phase of AI, many of the applications are limited in scope and value delivered. Capabilities will increase as GPUs become more powerful and models advance and mature, resulting in a lower carbon profile.

AI from the base to the fillings to the icing atop the cake

AI is comprised of a 5-layer technology stack with interdependencies. The footprint of this technology stack is expanding through the need for more grid power, higher performance compute, smarter data centers, and better performing models and applications.

Fortunately, as AI scales, opportunities will continue to emerge that will make a major impact on reducing the carbon footprint - from funding and operating on low carbon power sources to leveraging AI software to optimize the use of low carbon power sources. And there are innovative technologies like 800 VDC power distribution, BESS and liquid cooling, optimizing the performance of the liquid cooling via advanced controls, and innovating and evolving at the AI model and application layer.

From the base of the AI cake, through the fillings, to the icing atop the cake, AI will continue to evolve. It will be smarter and more efficient, resulting in each layer providing a path for AI factory operations to emit far fewer carbon emissions.

Why ‘Tokens Per Watt’ Is Crucial For Measuring AI Efficiency

Tokens per watt is an extremely useful metric showing how much “work” an IT system can produce for every watt of power consumed.

Many companies want to understand power utilization to feel good about the AI they will come to rely on. To help organizations, the time has come for the data center industry to agree upon using a new metric, one that is both intuitive and informative. This metric should bridge the worlds of power use and AI compute work output.

This is why I think tokens per watt should come to the forefront.

Usefulness Of Tokens Per Watt

You may ask, "What are tokens?" Simply put, tokens are the language that AI models speak. Text, images, audio clips and videos are broken into logical and descriptive pieces that all AI models can process.

Tokens are essential to understand for one important reason: They are how people pay for AI working models, also known as reasoning or inference models. In working AI models, tokens are used for input queries as well as output intelligence of prediction, content generation and reasoning. Users can pay based on token use. For ChatGPT's GPT-5, text input queries are $1.25 per 1 million tokens. Output responses are $10 per 1 million tokens. Image and audio token prices are higher.

Tokens per watt is, therefore, an extremely useful metric showing how much "work" an IT system can produce for every watt of power consumed.

At the micro level, companies can use the tokens per watt metric to grade IT performance and as an ROI to justify purchases. As GPUs evolve, there is usually an order of magnitude increase in performance, but also an increase in power usage. This juncture is where tokens per watt come into play, as the computing performance work output increase is typically much higher than the electric power use increase.

From a macro level, the tendency is to look at the gross power increase from data centers that will power AI. Tokens per watt can put a different lens on the topic by showing that, while power use is increasing, the work output is advancing at a much greater speed.

For years, IT performance advanced at a relatively slow pace compared to today's advances. Now, it's "accelerating compute" versus power use, and tokens per watt can quantify compute efficiency going forward.

The Need For Metrics Measuring Computing Output

While society's demand for increased automation through accelerated compute and AI is driving electricity demand, what's missing in the discussion is the "computing bang for the power used." You may argue that we already have metrics. That's true, but they're lacking in measuring computing output.

For decades, Moore's Law was touted as a metric to gauge computing progress. Published in 1965 by Intel cofounder Gordon Moore, the law essentially predicted that the number of transistors in a dense integrated circuit would double every two years, increasing processing power by 1.5 times or two times for the same power used. However, the law focused on central processing units (CPUs), not the GPUs used for AI today. It is also becoming less relevant due to, among other reasons, the speed of light imposing a natural limitation on the number of computations a single transistor can process.

Another commonly used metric, power utilization effectiveness (PUE), is a ratio that divides the power coming into your data center by the power used directly by the IT, with a perfect rating of 1.0. But people are used to efficiency as a percentage, so the ratio is often confusing. PUE also doesn't require real-time, continuous data monitoring, usually only one data point per week or month, allowing the operator to choose the most beneficial reporting time or turn up the IT power use at the measurement time.

Floating-point operations per second (FLOPS) per watt is a metric used in high-performance computing (HPC), but it can be misleading. Peak FLOPS are the standard metric. Tasks relying on FLOPS, like certain machine learning models, may have high FLOPS per watt. However, applications with less intensive floating point demands or tasks limited by other factors (low memory bandwidth or high latency) diminish their significance as a measure.

Targeting Improvements In Facilities

Data center operators have been measuring the power used—kilowatt-hour (kWh)—for a couple of decades for PUE reporting, carbon emission calculations and to target efficiency improvements. For kWh and other metrics like tokens per watt, operators need accurate power measurements. How and where power is measured depends on the type of facility and the IT workload involved:

- **For stand-alone facilities:**
  Get the power usage from the main central breakers data or the utility meter.

- **In mixed-use facilities:**
  Where the data center occupies only part of the facility - power usage can typically be measured at the uninterruptible power supply (UPS) connected to the main power feed or breaker serving that room. However, because cooling systems like chillers are often shared across the entire building, you'll need to allocate a portion of the cooling energy specifically to the data center's IT equipment.

- **For mixed, accelerated compute AI:**
  As well as for general-purpose IT or non-accelerated cloud IT, power use data needs to be taken at the individual server outlets from the power distribution units (PDUs).

- **For a rack of accelerated compute AI:**
  Branch feed meters in the busway provide the power data.

- **For an AI cluster:**
  Power is often delivered through large feeder breakers connected to a central PDU, switchgear or a dedicated UPS. These components can monitor and report the power consumption.

In all cases, total power usage should include both the electricity consumed by the hardware and the additional power required for cooling. To calculate tokens per watt, divide the number of tokens generated in one hour by the total power used during that same hour.

It's Time For Tokens Per Watt

As AI becomes increasingly central to business operations, leaders should begin evaluating infrastructure not just by power cost or capacity, but by how much useful AI output—tokens—is being produced per watt. That's the future of responsible, performance-driven AI deployment.

By providing this insight, the data center industry can demonstrate that AI is implemented responsibly.

This article was originally published by Forbes Technology Council.
Click here to access

Why Liquid Cooling For AI Data Centers Is Harder Than It Looks

Cooling is a complex architecture, and extreme densities require specialized expertise in designing, procuring, deploying, operating, and maintaining systems.

While liquid cooling is rightly considered an emerging technology, it's not new. Early IBM mainframes from the 1960s and Cray supercomputers featured liquid cooling.

Notably, a full-time technician was included in the Cray system purchase for installation, operation and maintenance.

Why AI Is Accelerating Liquid Cooling

Today, generative AI is reshaping the way compute and data centers are designed. Accelerated compute servers now incorporate two to 16 graphic processing units (GPUs) per server, alongside central processing units (CPUs) and even data processing units (DPUs). These servers are powerful number crunchers optimized for AI model training, but they consume over 20 times the power of standard Intel-based CPU cloud servers—and output 20 times more heat per server.

This heat output means these servers can only be liquid-cooled. Most now come standard with input and output piping for circulating liquid coolant.

Managing Heat: Power, Density And Design Challenges

Rack thermal demands have surged alongside each new generation of GPU-accelerated servers. When fully loaded into a rack, the latest NVIDIA-based GPU servers require 132 kW of power—and densities continue to increase. The next generation, expected in under a year, will require 240 kW per rack.

The dominant cooling method is direct-to-chip, or cold plate cooling. However, as the name suggests, it cools only the chips—not the rest of the components in the chassis or rack. Because liquid systems cool only the chips, supplemental air cooling still covers 20% to 30% of the total thermal load.

Cooling Is A Complex Architecture

Whether you're a large enterprise or a seasoned data center operator, it's unlikely you have the in-house expertise to design and deploy hybrid (liquid and air) cooling systems at these extreme densities. Specialized expertise is essential in designing, procuring, deploying, operating and maintaining such systems.

Direct-to-chip systems require two separate cooling loops—one for the IT room, another for heat rejection. Cooling distribution units (CDUs) interface between the two. When designing these systems, select a partner experienced with the full cooling architecture: manifolds, piping, CDUs, chillers, pumps and cabinets.

These components must function together, requiring compatibility, integrated controls and performance tuning. Choose vendors familiar with piping, fluid dynamics, pressure and flow rates—and ideally, ones that offer warranties and have certifications from GPU manufacturers.

The Role Of Simulation And Software

Given the extreme heat densities, trial-and-error approaches will extend the "time to cooling" and reduce the odds of success. Choose a partner that uses digital twin modeling and simulation to validate the system design virtually before deployment.

Vendors that work directly with GPU manufacturers have conducted lab testing or have proven deployments should be prioritized. Some vendors also offer pre-engineered and prefabricated cooling systems, which accelerate deployment and reduce risk.

Downtime Is Not An Option

At these densities, even a brief interruption in liquid flow can lead to thermal throttling or overheating in seconds. CDUs must include redundancy—dual pumps and power supplies should be standard.

Uninterruptible power supplies must support CDUs to ensure continuity during transitions to backup systems or generators. Leak detection software is also critical in the data center's white space; even a small leak can crash a server or cluster.

Optimization Requires AI, Too

Once operational, your liquid cooling system needs continuous tuning. Precision matters: even minor temperature increases can degrade GPU performance and slow down AI model training.

AI software can dynamically adjust cooling system parameters—like water temperatures, flow rates and airflow—in real time. These systems can even learn from operational data to optimize performance over time.

Choose Vendors With An Eye On The Future

The pace of GPU evolution is placing intense demands on cooling vendors. When selecting a partner, ask about their technology roadmaps—can they support future generations of GPUs with even higher thermal densities?

Liquid cooling may still be categorized as "emerging," but it is quickly becoming essential infrastructure. Companies aiming to scale AI must partner with vendors capable of supporting today's and tomorrow's requirements.

This article was originally published by Forbes Technology Council.

Click here to access

Innovative Ways Data Centers Can Supplement Electricity From The Grid

Data centers are mainly business critical, but they are quickly becoming life critical as more industries digitize and automate core processes and systems.

An electric utility has traditionally been defined as “a company that generates, transmits, and distributes electricity for public use.”

Data center customers are public users, and their facilities are filled with IT equipment, power systems, and cooling systems that all run on electricity as the prime power source. Prime power is defined as the main source of power for the data center, whereas back-up power (emergency power) is to be used when the prime power source fails.

For Forbes, I have written about the growing number of and complexity of back-up power systems for data centers. Back-up power is a hot topic as the electricity coming out of the grid is becoming more distributed, unreliable, and harder to find. Emerging as an even hotter topic is the mismatch of the exponential growth for data centers driven by AI computing vs the growth of electric utilities.

Electric utilities as an industry have benefited from years of little or no growth in demand while rolling out low carbon power sources (wind, solar, hydro) and decommissioning coal plants and nuclear plants. Today, they are shifting gears into fast growth mode, adding or extending electric generation capacity. However, the capital expenses, regulations, access rights for distribution and substations, and planning cannot keep up with emerging electricity demand.

Data centers are mainly business critical, but they are quickly becoming more and more life critical as more industries digitize and automate core processes and systems. Remember the recent global IT outage that brought many industries to a halt? Data center developers are now going to extraordinary lengths to source reliable power sources at scale. With increasing grid constraints, exploring on-site power or adjacent power is often the only way a data center development will be able to get approval. In my opinion, there are three paths developers can take to provide prime power in the short, medium, and long term.

Short term – Deploy existing solutions on-site

Natural gas power turbines:

While the turbines exist and can supply a high level of electricity, the natural gas power grids across the globe are certainly not as ubiquitous as the electric grid. Additionally, natural gas providers today do not have to meet the same reliability standards as the electric grid, but this is evolving. For example, U.S. Energy Transfer and Williams Cos., are in discussions with data center operators about the possibility of building pipelines directly to their facilities. In Ireland, several data center projects are proposing connections to the gas network instead of the electrical grid.

On-site solar and wind with battery stabilization:

These solutions are site dependent as real estate for solar and wind are needed for operation. While low carbon, these systems operate intermittently but can be integrated into a prime power ecosystem cooperating with electric utilities.

Fuel cells:

Fuel cells have been around now for a few decades and there are many plans and proposals in the works for them as prime power sources for data centers. Unfortunately, hydrogen, the power source for fuel cells, needs to be derived from water or natural gas and is dependent on large amounts of electricity, which kind of defeats the purpose.

Medium term – Purchase and rejuvenate existing power plants

Microsoft is partnering with Constellation Energy on a power deal to help resurrect a unit of the Three Mile Island nuclear plant in Pennsylvania that was closed in 2019. While no permits have been issued, the expectation is the plant will be operational in 2028 with a cost of slightly less than $2 billion. Amazon Web Services (AWS) recently bought a 2.5 GW nuclear-power plant from distressed Talen Energy for $650 million and Google is in talks with utilities in the United States and other countries to assess nuclear power as a possible energy source.

Long term – SMRs (Small Modular Reactors) on-site

There has been incredible enthusiasm from the data center industry for the promise of SMRs – safe, reliable, efficient, small footprint, air cooled, and runs on used, discarded uranium that has been refurbished – but SMRs still need to go through testing and regulatory approval. Additionally, while not the same nuclear technology, they come with a negative perception that must be overcome. Google announced recently that it has signed a deal with nuclear startup Kairos Power to build seven small reactors to supply electricity to its data centers by 2030, which seems quite aggressive and speaks to the strong desire for SMRs.

Emerging ecosystems of prime power sources

While there is no silver bullet to supply prime power to the data center market, data center power demands will continue to grow very quickly, and many data center developers are looking at on-site or adjacent power generation as a solution. I predict we will see ecosystems emerge of prime power sources where the simple electric grid as the source morphs into a complex ecosystem of power systems including: electric utilities, natural gas utilities, on-site natural gas turbines, on-site fuel cells, on-site or adjacent wind and solar, adjacent nuclear and on-site or adjacent SMRs. Of course, this much power complexity will need software controls and automation to match the power delivery to the desired data center operators' criteria whether it's lowest cost, lowest possible carbon emissions, or the highest reliability. Rest assured, these software automation solutions are being developed by Schneider Electric.

This article was originally published on the Schneider Electric Blog Site.

Click here to access

The 1 MW AI IT Rack Is Coming, And It Needs 800 VDC Power

Due to the laws of physics, 800 VDC is necessary for single IT racks 400kW and up to 1 MW. The architecture solves numerous problems.

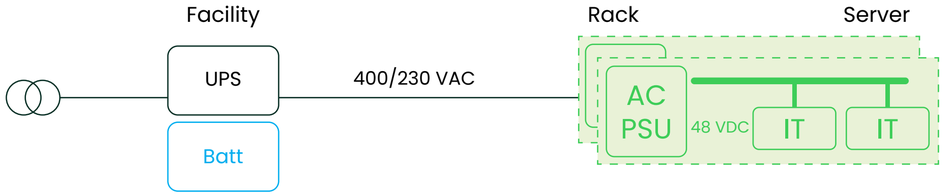
It seems like the 1890's again!

Why?

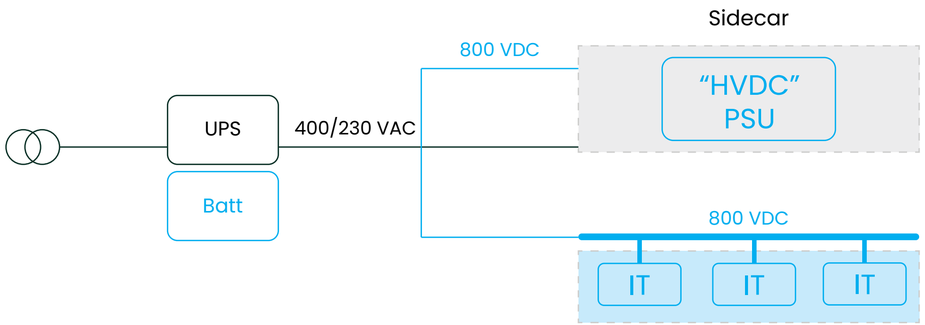
Back then a debate was raging between alternating current (AC) and direct current (DC) systems as the choice for America's electric grid. This was a pivotal moment in the history of electricity as Nikola Tesla was championing AC power that oscillates and Thomas Edison was championing DC, which flows steadily in one direction.

AC power won because of the ability to transform AC to high voltages to transmit long distances, which was not available for DC voltage. Higher voltage means lower current, lower voltage means higher current, and lower current means much smaller wires are needed. For example, running high voltage (35 kV) AC power lines long distance may be only 1 inch (2.54 cm) thick, but running them at low voltage means they would be about 6 feet (183 cm) thick, which is not only impractical, it's impossible!

This brings us to the modern day issue, which is the fast-moving rack power densities for accelerated compute platforms like the NVIDIA GB300 NVL72 that runs 72 GPUs in parallel at 142 kW per rack. Power must be transformed from the utility, most likely around 35kV down to 12V into the server chassis. The two main power distribution approaches feeding into the servers today are 400V 3 Phase AC and 48 VDC to the rack. Both of these approaches become difficult at 200 kW per rack and impossible at 400 kW per rack, which correlate with the NVIDIA Kyber and NVIDIA Rubin Ultra platforms.



At GTC 2025, NVIDIA exhibited an 800 VDC sidecar PSU (power supply unit) to power 576 of the Rubin Ultra GPUs in a single Kyber rack.



The benefits of 800 VDC architecture

Due to the laws of physics, 800 VDC is necessary for single IT racks 400kW and up to 1 MW! The Rubin Ultra GPUs in a single Kyber rack will begin shipping in 2027. Schneider Electric will have its sidecar in the market well before the release of the Rubin Ultra's. The Schneider sidecar technical specifications and reference design will also be available to engineers and data center operators well in advance to plan for deployment.

This 800 VDC architecture solves many problems:

Space limitations
Smaller cables and busbars provide connection flexibility.

Reduced copper usage
Weight and cost savings.

Higher efficiency
Reduced thermal losses.

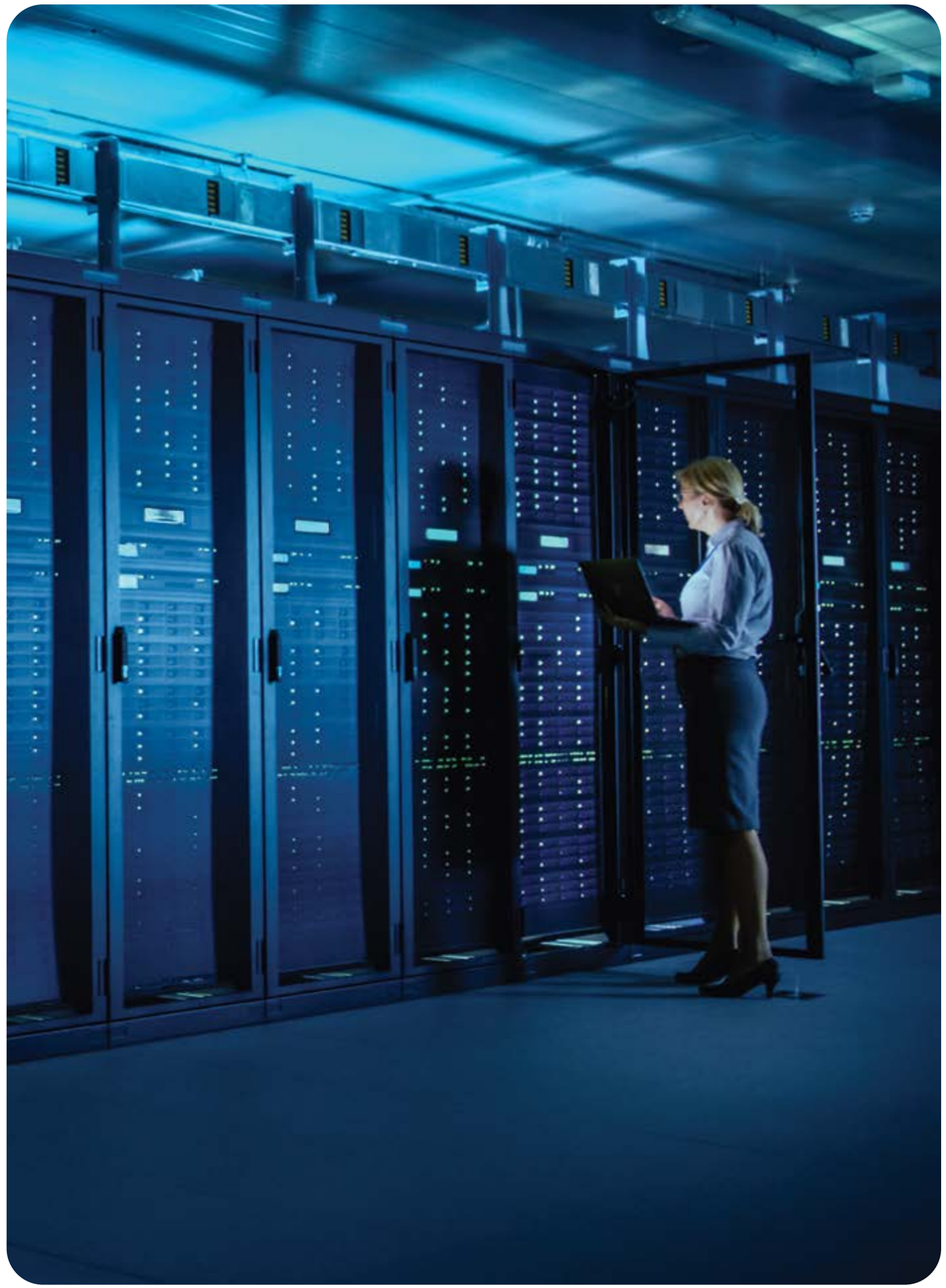
With a single-step AC/DC conversion, there are fewer transformer losses and a more direct power flow. There is also reduced electrical complexity and maintenance and management needs. DC power also brings in the use of diodes and overcurrent circuit protection, which are extremely efficient and reliable.

Schneider Electric's 800VDC sidecar not only aligns with emerging standards from NVIDIA, it supports Google, Meta, and others. We plan to offer advanced functionality "live swap" capabilities to dramatically simplify maintenance and reduce repair time.

Our commitment to supporting 800 VDC power

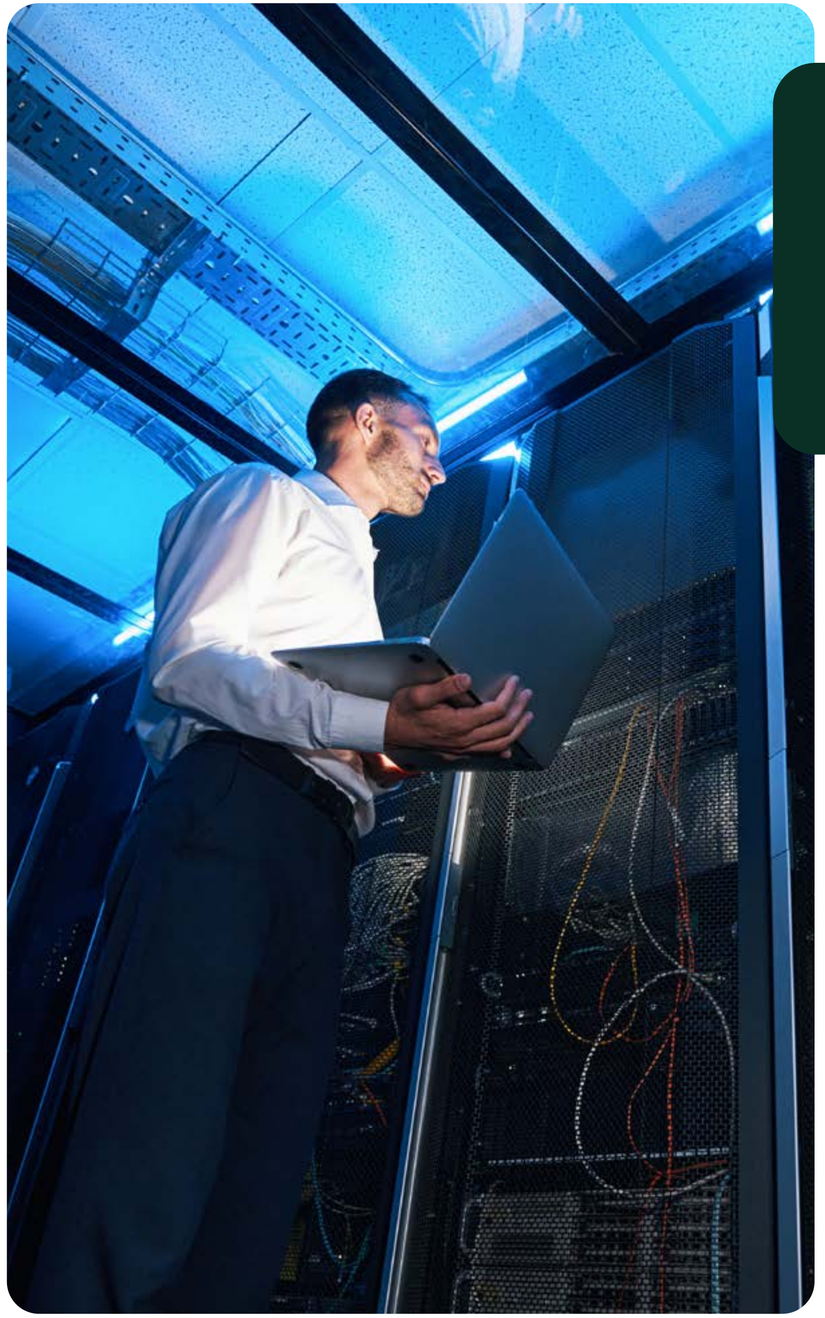
At Schneider Electric, we actively collaborate with NVIDIA and are committed to releasing power and cooling in advance of every future NVIDIA platform evolution. The 800 VDC sidecar is the first solution on the way to 1 MW IT racks but it won't be the only solution. We plan to continuously innovate power distribution and back-up solutions to drive increased resiliency, availability, and efficiency while simplifying deployment, operation, and maintenance. To learn more about our commitment, read our press release.

This article was originally published on the Schneider Electric Blog Site.
Click here to access



Mind The Gap: Bridging AI Talent Shortages In Data Centers

As AI increases demand and strains data centers to the limit, data center operators can address the AI talent gap and take steps to alleviate the situation.



As data center operators race to build and expand facilities, they are encountering shortages of the specialized skills needed to design, build, and operate AI-factory environments. From construction crews to cybersecurity experts to AI engineers to HVAC techs, to diesel mechanics and electricians, the workforce simply isn't keeping pace with the speed of construction and innovation. The result is a growing concern that talent, rather than shortage of power or semiconductors, may become the primary barrier to scaling artificial intelligence.

With power density of AI factories roughly doubling every year, many companies do not have the extremely talented engineers to keep up. Advanced design and simulation software can help make digital twins of the power system, cooling systems, and the entire data center. Also, prefabricated modules are available for the IT room, cooling, and power systems. Once the necessary site preparation is complete, they function as plug-and-play solutions. Built and tested in factories, they can reduce design time and costs while accelerating data center deployment.

Construction labor pains

Current skill shortages affect almost every aspect of data center operations, starting with power distribution, cooling, and building construction. As data center operators and hyperscalers rush to add capacity, they are finding that construction companies also face staffing challenges, in addition to slow permitting processes and access to power. Call it a perfect storm.

According to McKinsey, there aren't enough trained professionals for the electrical and mechanical installation work required in data center construction projects.

The U.S. construction industry is currently short roughly 439,000 skilled workers, and more than half of data center construction sites report disruptions due to staffing shortages, contributing to extended project backlogs—a dynamic that especially affects rural builds and drives large contractors to partner with smaller regional companies to secure needed labor.

Operational staffing challenges

Once a data center is built or expanded, skill shortages don't go away. All types of talent are needed to update, maintain, and operate equipment. But candidates for just about every role, from heating and cooling specialists to electrical engineers to project managers, are in short supply.

Many IT positions are difficult to fill. Cybersecurity talent shortages, for instance, are a chronic issue. And AI is so new that not enough skills have been developed to fill positions. According to Uptime Institute, 51% of data center operators struggled to find qualified candidates in 2024. The biggest gaps were in junior and mid-level operations roles, followed by shortages in operations management, mechanical, and electrical positions.

In addition, retention will remain a challenge, with pending retirements and data center talent leaving for other opportunities. New research from DataX Connect forecasts that 40% of data center professionals plan to leave their roles despite rising salaries, potentially widening the talent gap. Why? Working in the data center industry is extremely demanding, with long hours and high-pressure projects, so professionals are looking for more than just a salary; they want benefits, flexibility with work-life balance, and visible investment in their career development.

Filling the data center talent gaps

It's true that no single solution exists to address data center staffing gaps. But operators can take steps to alleviate the situation:

Train existing staff

Organizations often overlook existing talent, focusing on outside recruitment to fill positions. Assess your departments to identify promising talent who can be trained in AI and other areas with shortages.

Focus on retention

Offer continuous training, clear career paths, supportive leadership, and a strong workplace culture that reduces burnout and keeps employees engaged.

Outsource tasks

In some cases, data center operators can outsource simpler, routine tasks, allowing in-house staff to focus on more critical functions.

Automation

Leverage AI and robotics to automate tasks, which will help fill staffing gaps. Still, it raises a chicken-and-egg issue: You need AI talent to fill jobs created by AI before that talent is available.

Work with academia

Forge relationships with universities, colleges, technical institutes, and high schools to groom the next batch of data center workers. Offer internships and apprenticeships to give promising young professionals a career start.

Partner with a vendor

Lastly, operators don't have to tackle the AI talent gap and its related challenges alone. Schneider Electric offers a broad portfolio of data center solutions designed to support AI deployments from design through operations. These solutions help bridge skills gaps, streamline implementation, and optimize AI infrastructure.

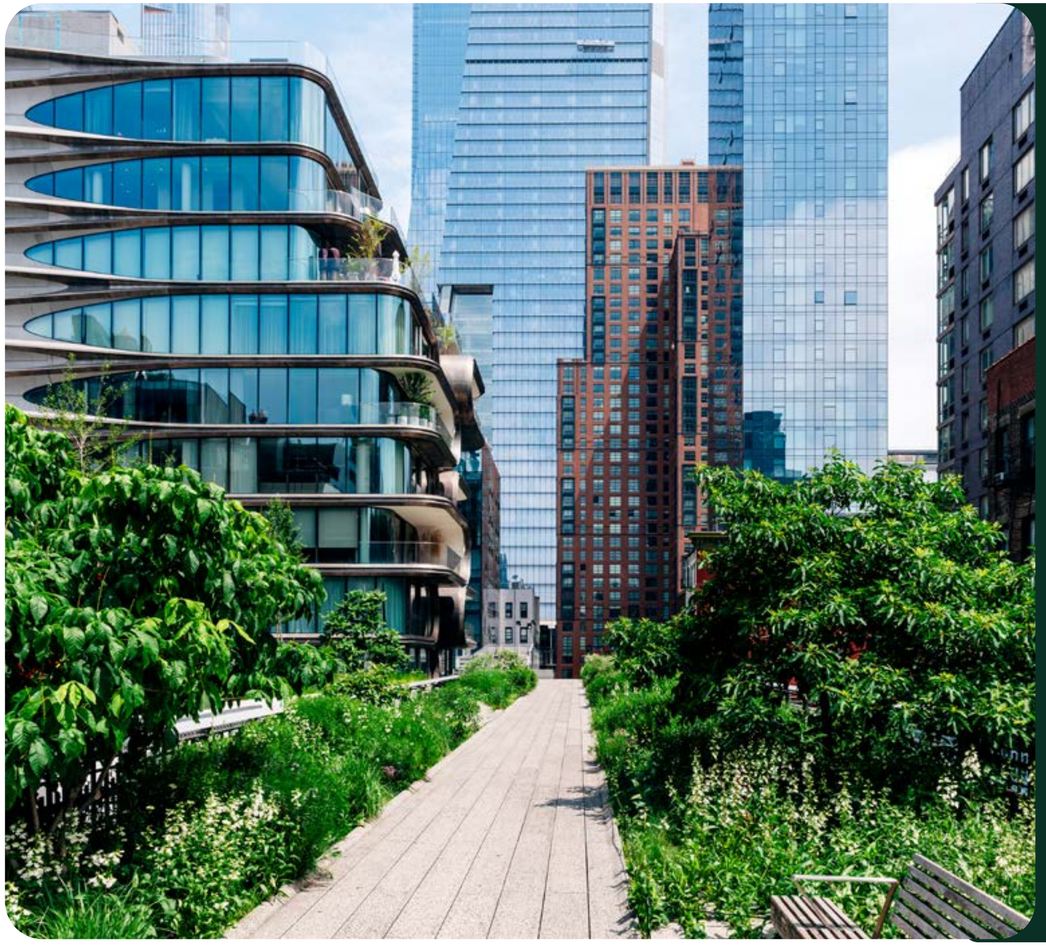
Addressing the AI talent gap

As AI increases demand and strains data centers to the limit, the talent shortage is proving to be a significant challenge. We may eventually see highly autonomous facilities, but we're not there yet—and human expertise is still critical to operations.

In the meantime, taking practical steps—such as upskilling and retaining existing staff, outsourcing routine tasks, leveraging automation, and building partnerships with schools and technology vendors—can make a measurable impact. By addressing the talent gap today, operators can stay ahead of accelerating AI-driven growth rather than constantly reacting to it.

Lastly, discover how Schneider Electric's EcoStruxure™ Pod Data Center solutions can adapt to your facility's specific requirements and simplify deployment, helping your team scale faster with less strain on internal resources.

This article was originally published on the Schneider Electric Blog Site.
Click here to access



Author

Steven Carlini

Chief AI Advocate,
AI and Data Centers
Schneider Electric
LinkedIn
Forbes

With his expertise in data centers and AI coupled with a passion for innovation and the ability to forecast technology trends, Steven is an instrumental member of Schneider Electric's global leadership team.

A popular speaker at business and technology summits, his focus areas include AI, accelerated computing, data centers, power systems, liquid cooling, sustainability, and cloud and edge computing.

Sought-after by the media for expert commentary, he's been featured in **CNBC, The Financial Times, Bloomberg, Barron's, The Economist, BBC**, and more.

A member of: **Capacity Media's** Capacity POWER 100, a list of "trailblazers, innovators and leaders driving the global digital infrastructure space"; Data Economy's Future 100, the top 100 people to watch; Forbes Technology Council; and the World Economic Forum's 5G-Next Generation Networks Programme. Please follow Steven on LinkedIn.

Schneider Electric