AI FACTORY: A CASE STUDY FOR TOTAL COST OF OWNERSHIP

WITH THE EXPECTED EXPLOSION OF GENERATIVE ARTIFICIAL INTELLIGENCE, RESPONSIBLE DESIGN AND LOCATION CHOICES MUST BE MADE FOR A SUSTAINABLE FUTURE.

By: Laura A. Laltrello, Applied Digital

Contributors: Todd Gale, Etienne Snyman, Nick Phillips, Lewis Maggio and Koen Lock

June 15, 2025

AI FACTORY: A CASE STUDY FOR TOTAL COST OF OWNERSHIP

EXECUTIVE SUMMARY

The explosive growth in AI computing demand (projected to require 171-219GW globally by 2030)[1] necessitates rethinking data center design and location strategies. By leveraging stranded power sources and climates optimized for free cooling, AI factories can achieve significantly lower total cost of ownership while supporting the infrastructure demands of generative AI applications.

The analysis suggests shifting away from traditional hubs like Northern Virginia toward more optimal regions with abundant, low-cost power, and favorable cooling conditions.

KEY FINDINGS

- An **AI factory** is a **new category** of data center with **unique requirements for design** and **location** to **optimize Total Cost of Ownership (TCO)**.
- An AI factory energizes and cools IT equipment with **fifteen to thirty times the power density** of a traditional data center, which:
    - **increases the power capacity** and **density** requirements.
    - shifts **latency prioritization to inside** the data center.
    - and shifts most **cooling** from air **to liquid.**
- Higher power density forces a **novel approach** to **source enough sustainable power** on an **already constrained resource**.
- **Liquid cooling** provides an opportunity to **leverage free cooling** where the **ambient air does the work** instead of mechanical refrigeration.
- These choices **optimize TCO by optimizing both the supply and the demand for power**, the most constrained resource in the ecosystem.
- This results in **lower cost for power from more renewable sources** and **reduced power consumption, reducing electricity costs**, the largest operating expense in a data center.
- Site selection and design can **reduce the total cost of ownership by $30-$50M** per year for a 100MW data center over the current industry standard, or **$0.9 - $1.5 billion for a thirty-year lifespan**.
- Furthermore, a greenfield 100MW AI Factory can have a **lower total cost of ownership by $60 - $90M** per year **or $1.8 - $2.7 billion** for a thirty-year lifespan, compared to traditional data center facilities.

WITH THE EXPECTED EXPLOSION OF GENERATIVE ARTIFICIAL INTELLIGENCE, RESPONSIBLE DESIGN AND LOCATION CHOICES MUST BE MADE FOR A SUSTAINABLE FUTURE.

[1] Srivathsan, Sorel, Sachdeva, & Arjita Bhan, 2024

UNDERSTANDING AI FACTORY REQUIREMENTS

THE NEED FOR AN AI FACTORY

Generative AI, or GenAI, is driving the rapid expansion of the data center market. ChatGPT alone, the most popular GenAI engine, answers over one billion queries per day[2] and the platform is only two and a half years old. To put this into perspective, it took Google nearly 11 years after launch to reach a billion queries in one day.

GenAI is a fundamentally new paradigm for processing information and solving problems. Traditional computation that has driven the data center market to date follows a deterministic, rules-based processing approach where precise instructions produce predictable outputs. GenAI uses probabilistic models trained on vast datasets to generate responses based on learned patterns rather than explicit programming.

Applications running in a traditional data center typically scale linearly with computational complexity and can run efficiently on modest hardware. GenAI scales exponentially requiring enormous computational resources for training and inference, requiring specialized hardware like GPUs, and consuming significantly more energy per operation.

GenAI demands new hardware: GPUs and CPUs working in tandem, spaced closely together. This new hardware in turn demands a new data center: an AI Factory that can manage power, cooling, and latency requirements created by the tens of thousands of tightly spaced, parallel processing chips working to solve complex problems.

WHAT IS AN AI FACTORY?

An AI Factory must solve the new challenges of GenAI requirements: the power density required from tightly spaced, power-hungry GPUs and CPUs driving learning and inference through parallel processing, which drives heat density. The power density and increased power requirements exceed the tipping point of air cooling and expand the scale of the data center itself.

Two-thirds of today's data centers contain racks requiring between four and nine kW of power, and less than two percent of data centers have racks with greater than 50kW[3]. By stark contrast, the newest generation in GenAI computing, the NVL72, demands more than 100kW, like SuperMicro's GB200 NVL72 SuperCluster, requiring 132kW[4]. At fifteen to thirty times the power density, moving to over 1000 watts per square foot, the cooling and power strategies must be redesigned.

AI FACTORY VS. TRADITIONAL DATACENTER: DIFFERENT OPTIMIZATIONS FOR DIFFERENT GOALS

Traditional data center:

- Located near users to minimize latency.
- Designed for reliability and continuous operation.
- Optimized for diverse workloads and client-server transactions.
- Success is measured by uptime and response time to external requests.
- Power and heat density acceptable for air cooling.
- COLOs often multi-tenant.

AI Factory:

- Optimized for internal communication between processors.
- Designed for massive parallel computing power.
- Complex learning and inference with large datasets.
- Success measured by computational throughput and efficiency.
- Power and heat density force liquid cooling.
- Single tenants allow custom optimization for cooling & network.

[2] https://www.demandsage.com/chatgpt-statistics/#

[3] Donnellan, et al., 2024

[4] https://www.supermicro.com/manuals/brochure/Brochure-AI-SuperCluster-NVIDIA-GB200-NVL72.pdf

WHY THIS MATTERS

The exploding demand from GenAI constrains the most critical resource – **power** - with estimates of global demand ranging from 55 GW in 2023 to between **171 to 219GW** of power required by data centers **by 2030**[5]**.** This strains our infrastructure and the resources that supply it. In the United States alone, data center domestic energy usage is expected to double or triple by 2028, reaching between 74 and 132 gigawatts, or 6.7% to **12% of total U.S. electricity consumption,** according to the Lawrence Berkeley National Laboratory[6].

By revisiting the basis of design and the location selections for AI Factories separately from traditional data centers, it is possible to find efficient, effective, and more sustainable ways to support this growing demand.

This whitepaper explores the design and location choices to optimize the total cost of ownership for an AI Factory. While the choices in design and location must consider the same factors as a traditional data center, the importance shifts to power and cooling.

POWER DENSITY CAN INFLUENCE DESIGN AND LOCATION

AI FACTORIES USE OF STRANDED POWER

The power density required for the rack extends to the AI Factory itself, as large clusters are required to train or infer from large datasets, with a cluster of GPUs working together to solve a problem - much like neurons working together in a brain to produce an output. However, unlike the human brain, which processes roughly one exaFLOP (10^{18} FLOPS) per second[7], modern AI clusters can achieve over 260 exaFLOPS[8]. This exacerbates the need for at-scale power access with locations that can supply a gigawatt or more.

Over the past two decades, the development and construction of renewable generation resources have outpaced the ability of utilities to construct sufficient transmission capacity to move that electricity to large load centers, such as cities. This leads to congestion on the transmission systems. Grid operators manage congestion by providing economic signals, in the form of negative power prices, to generators located in areas with insufficient transmission capacity. Owners of renewable generation can support some negative electricity prices as they have other, offsetting revenue streams, such as production tax credits and renewable energy credits, to offset the loss associated with the electricity. However, supply and demand dictate that the price of electricity decreases until such time as the transmission system is at equilibrium. The resulting curtailed, or negative-priced power, is referred to as "stranded power."

Developing data centers at areas with stranded power can benefit all parties involved:

- Data center – gains access to available capacity at competitive electricity prices
- Renewable energy generators – Incur fewer curtailments and improve interconnection applications approvals.
- Grid Operators – improve grid reliability with a more balanced grid, requiring less intervention.
- Other grid customers – share benefits of lower cost by reducing transmission system expansions.

[5] Srivathsan, Sorel, Sachdeva, & Arjita Bhan, 2024

[6] DOE Releases New Report Evaluating Increase in Electricity Demand from Data Centers., 2024

[7] Madhavan, 2023

[8] Thiagarajan, 2024

SHIFT TO LIQUID COOLING INCREASES FREE COOLING POTENTIAL

As a rule of thumb, power density exceeding 50kW per rack requires liquid cooling. Liquid cooling is a more efficient medium for heat rejection, and direct-to-chip liquid cooling cold plates reduce heat at the source. Air cooling is less targeted, making liquid cooling in a data center far more efficient than air. Liquid cooling systems can transfer heat when the outdoor temperature is only 10 – 18^{O} F below the coolant temperature, increasing the locations that can benefit from free cooling. Air cooling requires a 25 – 45^{O} delta to achieve the same result.

Free cooling uses ambient air to cool the liquid directly with no need for mechanical refrigeration. This liquid is routed back through the data center onto the cold plate to cool the chip, then cycled outside into a dry cooler where the ambient air dissipates the heat from the liquid.

Cooling towers and dry coolers used in liquid systems are more efficient at rejecting heat to ambient air than direct air cooling of servers. The large surface areas and optimized airflow in these heat rejection systems create better heat transfer conditions than cooling servers directly with ambient air.

Dry coolers and cooling towers use ambient air to cool the liquid; therefore, the power consumption for those technologies is less compared to mechanical chillers, which is referred to as "free cooling." Using free cooling with a liquid cooling system is more practical as it mixes with mechanical cooling more seamlessly than air cooling. As outdoor temperatures rise, chillers can supplement the dry coolers as needed. Air systems often have a more abrupt transition between free cooling and mechanical cooling modes, further reducing free cooling's applicability with air-based cooling systems.

In summary, the shift to liquid cooling creates an opportunity to effectively use free cooling to cool an AI Factory.

WHICH MARKETS ARE BEST OPTIMIZED FOR AI FACTORIES

An AI Factory can optimize TCO in locations with stranded power and that can utilize free cooling. Selection criteria also include locating where fiber routes are available and redundant, the land is both legally and environmentally buildable, and there is a workforce to support operations. Regulatory environment and tax incentives are also considerations.

Northern Virginia, currently considered the data center capital of the world, has over 51 million square feet of data center space consuming 4 GW of power. This is equivalent to the electricity consumption of 800,000 homes.[9]

A single future AI Factory needing 4GW of power is imaginable and probable. Such a facility would likely be a small fraction of the 51 million square feet of data center space in Northern Virginia, yet equal to the largest data center city in power required globally today. To build such facilities responsibly, locations addressing all the above-mentioned TCO factors should be prioritized.

WHERE ARE AI FACTORIES BEING BUILT?

With Northern Virginia's strengths in data center infrastructure, it remains a sought-after location for building an AI factory. However, Virginia has become constrained for power, **using more than it generates.** Additionally**,** the state is focused on shifting to intermittent generation, which may not align well with the reliability requirements of AI Factories. Virginia does not have excess generation capacity to support significant data center growth. Virginia has 174 equivalent days per year below 56^{O}F.

Texas has seen the most growth, and according to Collier's, the Dallas area could grow to the second-largest U.S. market if all its planned capacity comes online[10]. Texas is the largest electricity-generating state in the US and has adopted

[9] Clabaugh, 2024

[10] Saavedra & Seaward, 2025

favorable data center builder incentives and regulations, making it an attractive market. Texas also has a diverse workforce with approximately 1,300 people moving to Texas every day[11].

While Texas generates the most power, it also consumes the most. Texas currently has a generation capacity of 155GW annually[12]. The Electric Reliability Council of Texas ("ERCOT"), the grid operator, released a Long-Term Load Forecast Update in April 2025 showing a forecasted load of up to 218 GW by 2031. Based on experience from '22 to '24, ERCOT is delaying in-service dates for all new large loads by 180 days, putting data centers six months behind their requested dates[13]. In terms of cooling, Texas has a warmer climate than Virginia, with 2,273 hours below the 56° threshold needed to maximize free cooling. **The delayed supply line for power and only 95 days of free cooling suboptimize Texas for an AI Factory.**

Like Texas, North Dakota is blessed with natural energy resources and generates 42.1TWh (4.8GW) of electricity with a net summer electricity capacity of 9.4GW. North Dakota has more capacity than electricity it needs to generate due to demand. Unlike Texas, North Dakota uses significantly less than it produces and exports 33% out of state. With 34% sourced by wind energy, and most of the remainder sourced from reliable hydrocarbon-based generation. North Dakota's wind energy is often in regions with more supply than demand and limited transmission to move the generated capacity, creating stranded power. Utilizing this stranded power along with the lowest commercial and residential electricity rates in the United States[14] makes **North Dakota a great candidate for AI Factories**.

With abundant cost-effective supply, North Dakota's colder climate - ~5,280 hours (about 7 months) or the equivalent of 220 days annually below 56ºF - lowers power demand. This provides almost **four more months annually of free cooling than Texas, and a month and a half more than Virginia. This means an AI Factory can cool using no mechanical cooling for a significant portion of the year. This saves energy and improves** Power Usage Effectiveness or **PUE.**

STRANDED POWER, LOW ELECTRICITY RATES, AND ABILITY TO LEVERAGE FREE COOLING MAKE NORTH DAKOTA AN IDEAL STATE FOR SUSTAINABLE AI FACTORIES WITH BEST IN CLASS TCO.

[11] Texas Infrastructure Report Card 2025, 2025

[12] ASCE, 2025

[13] 8.1 Long Term Load Forecast Update, 2025

[14] U.S. Energy Information Administration, 2025

HOW THESE DECISIONS IMPACT TOTAL COST OF OWNERSHIP

For AI factory tenants, lease rates and electricity costs comprise over 80% of total operating expenses, **making design and location** the most **critical factors** in **optimizing Total Cost of Ownership** (TCO). Strategic decisions in these areas can **reduce annual operating costs by $30-50 million** for a typical **100MW facility** compared to a similar liquid cooled facility in a warmer climate and less efficient design through improved Power Usage Effectiveness (PUE), electricity rates, and Water Usage Effectiveness (WUE) and **$60 - $90M** compared to the industry's current data center fleet.

LOCATION: THE 0.15 PUE ADVANTAGE

In their research on climate- and technology-specific PUE and WUE, Lei and Masanet simulate ten different data center designs in fifteen different climates. "As expected, the simulated PUE ranges are lower in cooler climates."[15] Based on their research, the climate difference between North Dakota to Texas impacts PUE between 0.07 and 0.17 points. This small difference has massive cost implications:

North Dakota vs. Texas Comparison:

- **North Dakota**: ~5,280 hours annually below 56°F (**220 days of free cooling**)
- **Texas**: ~2,273 hours annually below 56°F (**95 days of free cooling**)

Cost Impact: The additional 125 days of free cooling in North Dakota reduces annual electricity costs by **$5 - $15 million** for a 100MW facility**.**

STRANDED POWER: ACCESSING BELOW-MARKET ELECTRICITY RATES

AI factories can capitalize on stranded renewable power - electricity that would otherwise be curtailed (reduced generation) due to transmission constraints. This creates a triple benefit:

- **Cost Savings:** Negative-priced or significantly discounted electricity during peak renewable generation periods.
- **Grid Stabilization Value:** Utilities offer preferential rates to loads that can absorb excess renewable generation.
- **Transmission Cost Avoidance:** Avoid high-demand urban markets eliminates transmission congestion charges.

Example: North Dakota's abundant wind generation creates regular stranded power opportunities, with electricity rates among the lowest in the US.
Cost Impact: An improvement of four cents per kilowatt-hour, the anticipated rate difference between North Dakota's rates and those in Texas or Virginia, reduces annual electricity costs by **$40 - $50M** for a 100MW facility.

WATER USAGE: DRIVING SUSTAINABLE AI FACTORIES AND MANAGING INCREASED REGULATION

While water costs represent a smaller portion of direct operating expenses, WUE optimization drives significant TCO benefits:

Direct Water Cost Savings: A 100MW facility with industry-average 1.8 L/kWh WUE consumes **416 million gallons** annually, or the **equivalent of 3,500 four person households**. **A closed-loop liquid cooling solution can reduce** this to 0.0052 L/kWh WUE, the equivalent of **one four-person household**.
Regulatory Risk Mitigation: Multiple states are banning high water-consumption data centers. Facilities with poor WUE face:

- Permitting delays and increased compliance costs
- Potential operational restrictions during drought conditions
- Reputational costs affecting corporate sustainability commitments.

Design Impact on WUE: Air cooling achieves zero water consumption, but at the cost of 0.2 points higher PUE. **Advanced liquid cooling systems with closed-loop designs optimize both metrics simultaneously**.

---

[15] Nuoa Lei E. M., 2022

Cost Impact: The water savings are $1.5 - $2M per year for 416 million gallons of water.

QUANTIFYING THE COMBINED TCO IMPACT

100MW AI FACTORY COST COMPARISON

	Baseline	Improved	Optimized
	Industry Average	Liquid Cooled Site / Reduced Rates	Closed Loop Liquid Cooled, ND Climate
PUE	1.58	1.33	1.18
Electricity Rate ($/kWh)	0.09	0.07	0.04
Annual Electricity Cost ($/million)	$125	$82	$41
Water Consumption (in million gallons)	416	416	0
Water cost	$1.7	$1.7	$0.0

Moving to a high-density liquid hybrid cooled site, as most AI Factories will do, the PUE will improve ~0.2 – 0.25. When combined with the location of North Dakota and the advantage of free cooling to optimize design and location, this generates an annual savings of $85M per year vs. the average data center today, and $43M vs the likely AI Factory placed in a warmer climate.

OTHER CONSIDERATIONS

North Dakota offers available, sustainable power at the lowest rate in the nation and a climate suited for free cooling. What about the other considerations for an AI Factory, such as fiber connectivity and latency, regulatory environment, and workforce availability?

GRID RELIABILITY

In addition to the climate benefits of North Dakota, it has the most stable power grid. In contrast, Texas has the least stable grid, accounting for 13% of the US outages in 2024[16]. During outages, AI factories must generate their own power at approximately 10× grid rates.

FIBER CONNECTIVITY

Fiber connectivity, the connection from data centers to other data centers and to end users, is measured by latency. To reduce latency, most data center providers seek two or more distinct fiber paths to a potential site. AI Factories need low latency (<100ms) to moderate latency (<1s) for applications like chatbots, conversational AI, real-time recommendations, live video analysis, web search, or image classification. Less than 100ms feels instantaneous to most users.

Average Latency (ms)	New York	Chicago	Seattle	Atlanta	Los Angeles	San Francisco
Ellendale, ND	35	16	26	39	53	45
Abilene, Texas	44	41	68	19	21	31
Northern Virginia	9	18	58	16	76	68

With the number of fiber routes placed in the United States, the current network latency is more tied to the distance between the two cities than to the remoteness of any given location. For all three sites, the latency is less than 100ms

[16] Texas Infrastructure Report Card 2025, 2025

across the country. While it is not ideal for all locations for some specialized very low latency applications, it is more than acceptable for most GenAI training and inference applications.

REGULATION AND TAXES

Each of Texas, Virginia, and North Dakota has tax incentives and a regulatory environment advantageous to data centers.

Texas and Virginia have well-established regulations with tax incentives. Texas offers tax incentives conditioned upon a minimum investment of $200M of capital investment over five years, 100,000 square feet in size, and a minimum of twenty jobs created. Virginia's standards are similar, requiring a minimum of $150 million in capital investment and at least fifty new jobs. Virginia is now facing growing regulatory headwinds with proposed restrictions on siting, reporting energy requirements, and environmental controls, which could impact future development.

North Dakota has the lowest required investment for a new data center. A new data center must be at a minimum of 15,000 sq. ft. with 50% or more of its area used for data processing.

WORKFORCE AVAILABILITY

North Dakota has the most constrained labor market of the three states. It offers more jobs than workers and is new to the data center industry. However, history has repeatedly proven that **jobs create towns** not towns create jobs. This has been true from the gold rush to the construction of the transcontinental railroad and the Industrial Revolution. Thus, GenAI, will be its own combination of all three. North Dakota has a history of successfully importing workforce to support a new industry. For example, the Bakken oil boom in North Dakota, which began in earnest around 2006-2008, transformed Watford City from a small agricultural town into a major oil hub. The city's population exploded from about 1,744 residents to nearly 7,500 in one year[17].

POLARIS FORGE: A CASE FOR DIFFERENT BY DESIGN

The design elements discussed in this paper are precisely what led Applied Digital to Ellendale, ND, and to other sites in our pipeline. Ellendale sits adjacent to **stranded power** from wind farms taking advantage of a renewable energy source and best-in-class electricity rates. We worked with the utility provider to help transmit that wind energy, when available, to our facility resulting in reduced power costs for the site and for the community.

Our North Dakota location is optimal for our closed-loop direct-to-chip liquid-cooled design. This design uses dry coolers supplemented by mechanical refrigeration. **Closed loop** means the liquid flows through the system continuously without need for replacement, reducing the **water needs** in the data center to **less than a single four-family household**. **Direct-to-chip** efficiently **cools the source** of heat through a cold plate on the processor. Our dry coolers can take advantage of North Dakota's cooler ambient air, which offers **free cooling 220 plus days per year.** The PUE design specification for Ellendale is currently 1.18, equivalent to the PUE cited in the paper. With the amount of free cooling in North Dakota, we expect the operational PUE to be closer to 1.10, driving additional savings in electricity costs and a more sustainable site.

Like Watford City, Ellendale is a small agricultural town. To prepare for the growth, we have partnered with the local government, citizens, and business owners to build responsibly and sustainably. For our workforce, we partnered with the business community to build twenty homes + thirty-eight apartments to start fulfilling their housing needs.

[17] The Dickinson Press, n.d.

At Applied Digital, we understand the strategic importance of designing solutions for power and cooling that consider the long-term costs and needs of our customers. From site selection to community engagement to build and operate, we strive for best-in-class total cost of ownership with sustainable solutions that integrate into the local communities.

AI Factories can drive the betterment of communities with **responsible**, **thoughtful,** and **sustainable** choices in location and design.