Growing Data Center Capacity and Reducing Grid Impact: Strategies for Efficient, Safe, and Reliable Facilities

Kieren McCord, Bing Liu, Pacific Northwest National Laboratory
Kaustubh Phalak, Jacobs
Jamison Caldwell, Ionel Petrus, Jason Smith, SmithGroup
Todd Sims, Casey Granata, National Electrical Manufacturers Association
Leigh Lain Walker, Thomas Loxley, ASHRAE

ABSTRACT

Data centers are essential infrastructure, enabling American digital innovation and economic growth. Demand for data center capacity is substantial and projected to continue to rise sharply into the future, due to expansion in cloud computing, data traffic, and artificial intelligence (AI). Accelerated servers are driving long-term trends toward higher rack power densities, and higher total facility power requirements. Combined with emerging power constraints, these trends highlight that a cohesive strategy to design and build efficient data centers with reliable performance is imperative. Furthermore, unique characteristics of AI data centers, such as rapid power ramp rates, underscore the need for inventive solutions that meet the need for data center capacity growth while keeping the grid reliable and energy affordable. Pacific Northwest National Laboratory, ASHRAE, and the National Electrical Manufacturers Association (NEMA) developed an AI Data Center Energy Performance Framework (Framework) to address these needs. Our paper presents an overview of the Framework and details some specific solutions, including energy storage, two-way uninterrupted power supply (UPS), microgrids, waste heat export, cooling efficiency, and demand flexibility. All are opportunities to increase data center capacity with reduced impact on the power grid, while emphasizing reliability, safety, and performance in all aspects of the data center facility. This paper addresses cooling efficiency technologies including adiabatic, free air, liquid cooling, and others at a high level. Real-world case studies for system optimization and impact on power usage effectiveness (PUE) and water usage effectiveness (WUE) are presented. The strategies are practical and applicable to both new and existing facilities in varied environments. The newly developed Framework will enable operators to increase return on investment (ROI) while maintaining operational effectiveness and dependability of data centers.

Introduction

With substantial data center growth in the last decade and widely varying projections for future growth, one underlying factor scholars and industry agree on is that data center computation and cooling needs are growing rapidly, which results in substantial increases in energy demand. Increases in hardware efficiency still do not outpace increases in training computation and duration of AI models (EPRI 2025). Rapid growth in any industry can result in increased demand on the power grid, higher demand for sector-specific resources, and needs for innovative strategies to sustainably develop the built infrastructure needed to house the developing industry resources. What makes AI data center growth so unique and challenging is both the absolute magnitude of the energy requirements as well as the magnitude of the temporal fluctuations. The most recent U.S. Data Center Energy Usage Report notes that as server power scaling ability has increased, the actual electricity demand for a particular server can fluctuate widely over the course of a day (Shehabi et al. 2024). In a data center full of these servers, this range of demand can multiply to create substantial and rapid swings in energy demand (Chen et al. 2025). Uncertainties in true demand from a large facility can pose challenges for utilities, such as when requested load growth is not always realized (Bessasparis et al. 2025). Current data center development has seen some geographical clustering (Bessasparis et al. 2025), and experts suggest that future development will include "concentrated and distributed data center loads" and that demand flexibility will be a key consideration wherever development happens (EPRI 2025).

To understand the factors involved in predicting and optimizing power usage in data centers, it is important to understand the two main components of data center power loads; First, the IT infrastructure, which provides the computational capabilities which constitute the core purpose of a data center, and second, the space conditioning, which can include a variety of cooling strategies to maintain IT equipment temperatures conducive to efficient operations. The ratio of the total facility energy used to the energy divided by the IT equipment is known as power usage effectiveness (PUE). The closer PUE comes to 1, the more energy efficient the facility is. While historical increases in efficiency have helped lower PUE a full point in the last couple decades, since around 2020, PUE values have stagnated around 1.55–1.59, with larger (and newer) facilities seeing the highest efficiencies, in part due to using newer and more efficient air flow management and controls in the cooling equipment (Davis 2024; Bizo 2023).

This paper makes two important contributions. First, this work presents the AI Data Center Energy Performance Framework (Framework) that was collaboratively authored by PNNL, ASHRAE, NEMA, and a variety of guest industry authors. This paper serves to introduce this work by presenting the motivation, methods, and high-level content contained therein. The Framework fills an established industry need for consolidated and updated guidance as the industry rapidly advances. Second, this paper provides two tangible case studies on Framework topics that illustrate implementable best practices for efficient data center design and operation.

Overview of the AI Data Center Framework for Efficiency and Resilience

In early 2025, PNNL, ASHRAE, and NEMA convened a group of 50+ industry professionals from sectors including electrical components and power systems; mechanical, electrical and cooling equipment manufacturers; data center design, engineering, and consulting firms; IT and networking companies; data center operators; real estate and financial investment; and professional and academic organizations. At this convening, participants were polled with the goal of understanding the needs of industry regarding AI data centers. Consensus around a variety of challenges emerged. Respondents cited design and modeling gaps, noting a lack of universal tools and accurate models to support architecture, product selection, and real-world cooling system performance, especially for AI-driven workloads. They noted a need for better data on AI versus non-AI energy use, as well as clearer visibility into emerging data center design trends and IT equipment shifts. They called out infrastructure and resource constraints, noting concerns around grid limitations, renewable development delays, water scarcity, and availability of land and skilled labor. Respondents described challenges with standards and best practices, noting missing or fragmented standards, particularly for liquid cooling, cold plate systems, and AI data center operation, along with a lack of integration across initiatives like

ASHRAE and Open Compute. Finally, they called for broader community planning, future-proofing for rapid load growth, and better education on emerging standards.

In response to the calls from industry to foster collaboration between disparate stakeholders, provide clear direction and best practices, and provide a mechanism for knowledge sharing and engagement for industry leaders, the AI Data Center Energy Performance Framework was created and made publicly available (McCord et al. 2026). The technical systems scope of the Framework covers mechanical and electrical systems in AI data centers. The practical scope of the Framework considers multiple phases of the data center development process, from early project initiation and design and through to operations and maintenance phases. The following section provides key takeaways from each Framework chapter.

Chapter-by-Chapter Highlights

Planning & Siting: Speed to market is front of mind in data center development, and decisions made in the pre-design phase have outsized impacts on project timelines. Selecting a project site requires balancing a variety of often competing priorities such as cost profiling, grid proximity and capacity, water availability, environmental assessment, and permitting. It is at this stage that determination of resiliency requirements, such as the Uptime Institute Tier classification for resilience, determines the number and types of systems involved, which drives costs, timelines, and site needs (Uptime Institute 2018).

Integrated Design Principles: Data center development requires coordination between a variety of sectors, including key players such as architectural, mechanical, electrical and IT engineers. Data centers dominated by AI loads require a higher density of cooling infrastructure (gray space) to serve the IT infrastructure (white space), necessitating careful coordination between these key performers.

Energy & Thermal Efficiency: A key motivation for needing better thermal efficiency in AI data centers is the higher rack-level power densities of GPU clusters compared to traditional CPU computing. As energy demand continues to increase, new sources of grid energy may not be available fast enough to meet demand. Parts of the solution will likely have to come from on-site energy generation and increases in cooling efficiency. High level recommendations for cooling efficiency include hot/cold aisle containment, right-sizing airflow, raising supply air temperature, and continuous monitoring. Additionally, implementing economization and free cooling, such as airside economizers, waterside economizers/dry coolers, and refrigerant free-cooling direct expansion (DX) systems can help optimize cooling efficiency. Emerging cooling architectures, such as liquid-assisted cooling in rear-door heat exchangers, direct-to-chip, or immersion cooling can help meet the demands of higher AI-driven rack densities. Another important consideration in thermal efficiency that can expand beyond the data center boundaries is the opportunity to utilize waste heat from the data center by utilizing heat recovery where there is a viable heat sink, such as nearby buildings or any on-site processes that require heat. Liquid cooling can provide higher-grade heat that is better suited for reuse. Geothermal solutions such as integrated thermal energy networks provide a high potential opportunity to leverage the heat resources produced by data center operation. Finally, controls and monitoring can provide opportunities to create feedback loops that continuously optimize the energy use of the facility.

Grid-Interactive Design: While available grid power can be a limiting factor for data center development, two-way communication and interactivity with the grid can be an asset. Pilot studies have shown that with demand flexibility, AI data centers' demand can maintain quality of service and still see significant (e.g., 25%) reductions during peak periods, a strategy that has the potential to help stabilize grid loads (Colangelo et al. 2025). In fact, a recent analysis by ACEEE determined that energy efficiency and demand flexibility are the most cost effective ways we currently have to reduce peak demand and overall electricity consumption (Specian and Aquino 2026). Bidirectional energy flow, integration with renewable energy, and demand responsive system design can increase the resiliency and efficiency of AI data centers (Hirschboeck et al. 2026).

Resilient Design: Critical data center operations must be resilient against outages. To design for current operation and future scalability, it is helpful to ensure voltage tolerances and equipment reusability to accommodate future transitions from 800 V DC to higher voltage levels such as 1500 V DC. Microgrid systems can be incorporated into data center design and can be optimized to ensure resilience and reliability of data center operations. Two-way uninterrupted power supply (UPS) and energy storage system compatibility can ensure resilience against outages and increase uptime percentage. Even the type of insulation used in the wiring of the electrical systems can have a substantial impact on facility reliability and resilience.

Commissioning and Performance Validation: As data center development continues to accelerate, commissioning and performance validation are crucial elements for ensuring that new projects perform as designed and for acquiring and sharing accurate data and lessons learned to apply to the next project. Commissioning ensures equipment cleanliness and preparation for performing its intended task. For example, for systems that use liquid cooling, proper cleaning, flushing, and passivating of the hydronic systems help avoid equipment fouling during deployment, which can eliminate massive project delays. It is important that discipline-specific subject matter experts work together on commissioning activities to ensure proper function and interoperability of all equipment.

Operations and Maintenance: While there is much public attention on new construction, it is essential that existing data center infrastructure is properly operated and maintained to ensure that the reliability and efficiency continue to meet targets over time. AI and machine learning (ML) techniques can be utilized for predictive maintenance in avoiding outages. Proper maintenance of systems can protect against cyber threats, prepare for disaster resilience, and optimize energy usage. Upskilling the workforce will be critical in ensuring maintenance of emerging technology configurations.

Retrofit & Modernization Strategies: As data center demand increases, there is a large capability gap between traditional data center infrastructure and systems that can support AI workloads. Some key recommendations for data centers aiming to expand operation to support AI workloads include adopting a hybrid cooling architecture and retrofitting power infrastructure to handle load volatility by increasing "headroom" in switchgear and UPS systems. Digital twin operation can allow simulation of increased loads and various scenarios to understand if a system can handle the loads before actual physical deployment.

Tools, Standards, & Resources: The Framework concludes with a centralized pool of resources that includes tools, standards, and other resources that are relevant to efficient data center development. These include NEMA design considerations, ASHRAE resources, and a variety of bridges and links to other resources such as content from EPRI, Uptime Institute, and other industry leaders.

The remainder of the technical content in this paper outlines two case studies that provide additional depth about select topics in the Framework. First, a case study on integrated design demonstrates impacts of cross-sector coordination on data center development efficiency. Next, a case study comparing advanced cooling configurations shows impacts on PUE and WUE, demonstrating key concepts from the Integrated Design and Energy & Thermal Efficiency chapters of the Framework.

Case Study 1: Impacts of Integrated Design on Data Center Development and Operational Efficiency

Case study 1 describes integrated design principles that intentionally align architectural, site, structural, mechanical, electrical, plumbing, and landscape systems to improve performance, resiliency, community compatibility, and long-term adaptability of AI-focused data centers. The principles emphasize early cross disciplinary coordination to avoid optimization of a single building system at the expense of overall design quality.

The case study is divided into four main topics, each of which details several associated design principles. A brief synopsis of how the design principles were specifically applied to a recent project case study is included. Although the Client and location of the project are confidential, the exercise examined the development of a master plan for a 450 MW greenfield data center campus. This project effectively illustrates the application of each of the proposed integrated design principles to a real-world scenario. For the purposes of this framework, the project will be referred to as "Greenfield 450."

Integrated Design Topic 1: Building Architecture

The building form and material selection was carefully considered in the Greenfield 450 master plan to balance technical performance, constructability, adaptability, and environmental response. Unlike conventional buildings, the form was primarily driven by the infrastructure requirements of the IT systems and the associated power and cooling systems. Still, a responsible architectural design process must also respond to site, climate, and community context. Thoughtful design at Greenfield 450 allowed the building facade to become a visible expression of the building's performance, communicating innovation, transparency and environmental responsibility. Architecture, materiality, and site design sought to create a campus of buildings that felt approachable, visually compelling, and scaled appropriately for those who work at the facilities and the surrounding community.

Building Form: Long-span, modular building forms were studied to enable clear separation and adjacency of data halls, mechanical galleries, and electrical rooms. This practice was aimed at minimizing distribution losses and improving system maintainability. The use of linear or bar-type forms supports a phased deployment of IT load, while allowing modular scaling of mechanical and electrical systems without significant downtime or rework. At Greenfield 450, structural grids were planned for flexibility of rack layouts, overhead power and cooling distribution, and future increases in rack weight to preserve flexibility as AI compute densities evolve. The master plan explored multiple shell-based building configurations (three, four, or five shells were studied), with each shell organized around 6 MW modular infrastructure increments. Building length and structural bay spacing were directly derived from mechanical and electrical module dimensions, ensuring that future expansion or densification can occur without altering the primary structural or architectural form. A centralized campus plant was compared to a series of distributed utility plants (local to each shell building) were tested against building geometry to preserve long-term flexibility without locking the site into a single cooling or power topology.

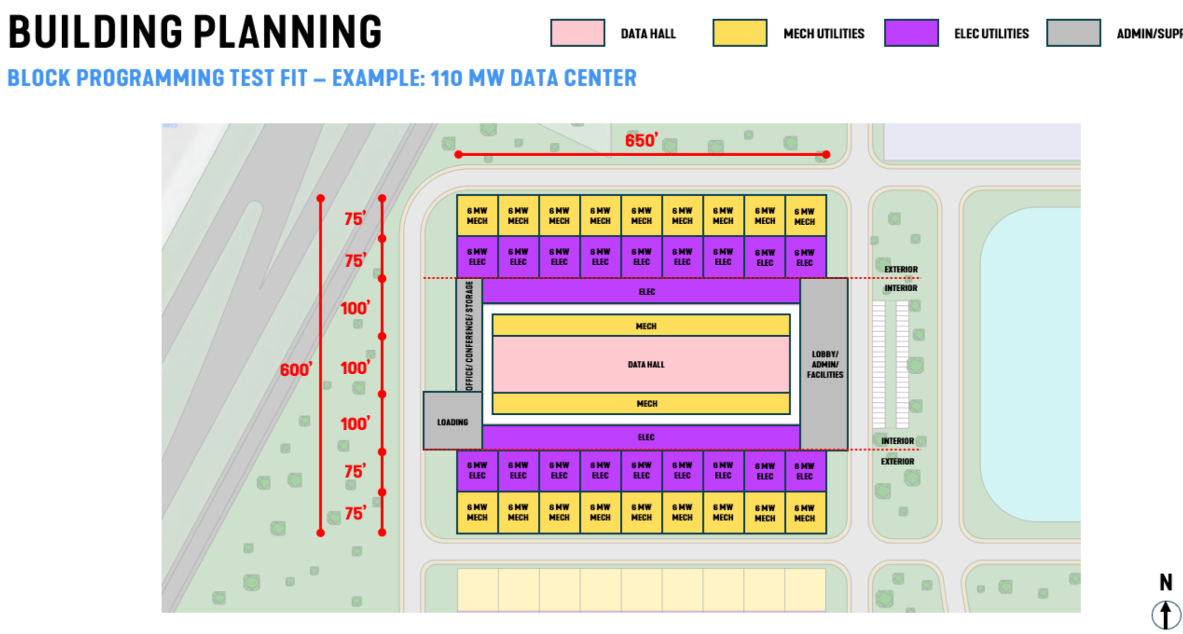

Figure 1. The Greenfield 450 master plan proposed options for 3, 4 and 5 data center building concepts. Shown above is the planning for the 4-building concept, demonstrating the long-span, modular intent. SmithGroup Image, 2026.

Façade: Exterior envelope systems prioritized durability, thermal performance, and consider resilience over transparency. High-performance opaque assemblies, such as insulated precast concrete, metal panel rainscreens, or composite wall systems, are effective in reducing heat gain, limiting air infiltration, and support acoustic mitigation. Glazing is limited to administrative or occupied spaces and is coordinated with building massing and siting such that daylight and views are applied where most beneficial to occupants, while maintaining security and thermal efficiency in data hall zones. Due to the sprawling nature of this building type, the roof 'façades' were given unique consideration for the Greenfield 450 master plan. Considerations were made for longevity and durability as well as visual appeal and potential for participation in renewable energy generation.

Materials: Integrated design excellences prioritize durable, high-quality materials that create warmth, human scale and visual richness while advancing environmental responsibility.

Buildings should emphasize materials with low embodied carbon and carbon-sequestering potential, avoiding red-list materials and prioritizing transparency in material health.
For Greenfield 450, structural systems favor steel or hybrid systems sized for heavy equipment loads and future flexibility, while envelope materials were proposed based on long service life, resistance to moisture, and compatibility with modular penetrations. Roof assemblies are designed for high reflectivity, future photovoltaic readiness, and minimized penetrations to protect mission critical spaces below.

Integrated Design Topic 2: Management of Community Impact

As AI data centers become more prevalent in local communities, it is essential that the early planning and design stages of any project engage with community members and reinforce design principles that focus on positioning the data center as a meaningful and responsible ("good") neighbor. Public realm improvements or benefits from data center construction help these highly technological projects become an integral component of civic infrastructure. Community impact mitigation is most effective when addressed holistically across disciplines rather than through isolated add-on measures.

Acoustic and Visual Impact Mitigation: On the Greenfield 450 project, the noise from generators, cooling equipment, and utility yards is managed through site planning first, followed by architectural and mechanical strategies. Locating major noise sources away from sensitive edges, using building mass as a sound buffer, and integrating acoustic barriers into landscape and enclosure systems reduces reliance on mechanical attenuation alone. Equipment selection, operating strategies, and enclosure design are coordinated to meet acoustic criteria without compromising operational reliability or efficiency. Visual impacts of such a large campus are addressed through building scale modulation, articulated façades, and layered landscape buffers. Rather than concealing infrastructure entirely, the design expressed the buildings as a deliberate, well integrated element of the site, using material texture, rhythm, and setbacks to reduce perceived bulk.

Traffic, Access, and Operational Interfaces: Early coordination between architectural, site, civil, Early coordination between architectural, site, civil, and operations planning are necessary to separate service traffic from public and employee access. The design for Greenfield 450 considered loading docks, fuel delivery, and maintenance routes to minimize conflicts with surrounding roadways and neighborhoods. Additionally, the access patterns considered future phased construction and operational growth strategies to prevent repeated disruption to traffic patterns or building uptime over the campus lifecycle.

Figure 2. Greenfield 450 campus proposed a combination of existing and new forested areas, and took advantage of existing roadways, to buffer commercial development to the west and residential development to the east from operations on the project site. SmithGroup Image, 2026.

Integrated Design Topic 3: Heat Reuse

AI data centers generate substantial quantities of recoverable thermal energy. Integrated, early, design decisions can significantly improve the feasibility and efficiency of heat capture and reuse at building, campus, or district scales. Facilities that transform excess heat into community benefit redefine the relationship between data infrastructure and the community (see Topic 2).

Scales of Reuse: Each building at Greenfield 450 was designed to consider direct reuse of recovered heat to offset space heating, domestic hot water, and any remaining process loads directly within the facility, reducing energy use. Architectural zoning placed heat accepting spaces, including offices, support areas, and ancillary functions, together to limit the scale of the heating distribution system to minimize distribution loses. Campus scale considerations at Greenfield 450 addressed site and utility planning needed to reserve corridors and connection points for future heat export to adjacent buildings, campuses, or even a district-level energy networks. Heating at the district scale has the potential to benefit residential, light commercial, research / industrial commercial and agricultural operations alike on adjacent properties. Even though no immediate district-scale demand for heat user existed at the time of Greenfield 450 plan development, providing physical and hydraulic "tie in readiness" preserves long-term optionality at minimal incremental cost. Early coordination is essential to avoid future costly add-on projects, such as considering tunneling under nearby roadways. Every new data center has the potential to prescribe a set of district level "heat reuse" design objectives that all new construction within the district must consider in the design of their building mechanical systems. For comparisons sake, it was estimated that the waste heat reuse potential of the Greenfield 450 project would be enough to provide heating capacity to approximately 60,000 residential homes and 45M gsf of commercial, mixed-use, construction.

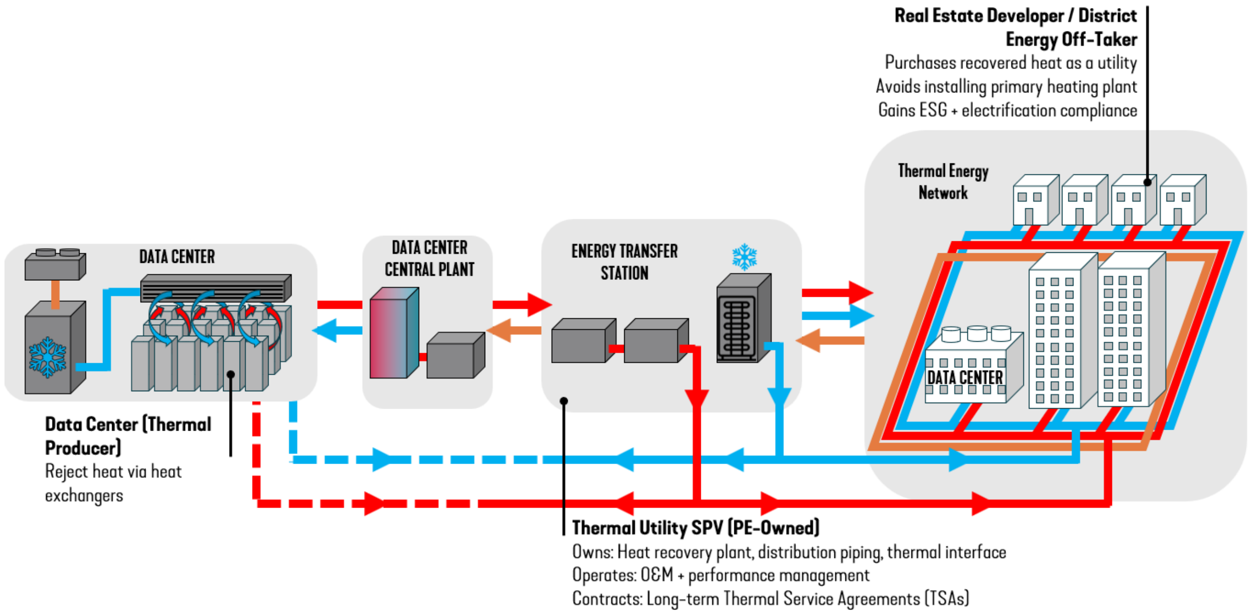

Figure 3. This potential heat reuse schematic demonstrates the components in the system and the various levels of ownership. SmithGroup Image, 2026.

Cooling System Architecture: Mechanical systems are planned around elevated temperature cooling loops that support both air-cooled and direct-to-chip liquid-cooled IT equipment. Higher return water temperatures increase the quality of waste heat, expanding opportunities for reuse while improving cooling system efficiency. Structural and spatial allowances are made for heat exchangers, piping, and future connections required to enable heat reuse beyond the building scale without disrupting ongoing data center operations. The architectural and site planning must consider the need for future energy transfer stations as a component of the overall site design.

Integrated Design Topic 4: Holistic Building Management Systems

The complexity and criticality of AI data centers demand building management strategies that transcend traditional system silos. Although Greenfield 450 was a master planning study and not a design project, it was important that the plan set early expectations around the intended scope of control systems to ensure conceptual cost modeling incorporated the prescribed holistic approach.

Integrated Controls Ecosystem: Rather than operating BAS, EPMS, DCIM, and IT management systems independently, the integrated design approach establishes a coordinated data architecture from the outset. Each platform retains its domain authority while sharing data through secure, standardized interfaces. This enables real-time visibility into power, cooling, water, and IT performance at a system-of-systems level. The Greenfield 450 master plan was developed to support either a single use or a co-location ownership model. Planning for monitoring and controls systems to include a high degree of metering, monitoring and control across multiple system-level platforms supports the design flexibility required to accommodate variation in operational ownership across the campus.

Operational Optimization and Resilience: The cross-platform integration required at Greenfield 450 plan also supports advanced analytics, fault detection, and predictive maintenance. By correlating electrical load, thermal performance, and IT utilization, operators can identify inefficiencies, anticipate failures, and optimize system response dynamically. Redundancy strategies are reinforced through coordinated alarms, sequencing, and automated responses across platforms.

Future-Ready Architecture: The controls framework is designed to be extensible, supporting future AI-assisted optimization, new sensing technologies, and evolving operational requirements without wholesale replacement. Governance, cybersecurity, and data ownership considerations are embedded in master plan to facilitate long term reliability and trust.
The Greenfield 450 project assumed a long-term, phased development horizon. The plan explicitly avoids prescriptive controls solutions, instead preserving architectural, electrical, and network pathways that allow future platforms, including those based driven by AI, to be layered onto the infrastructure over time.

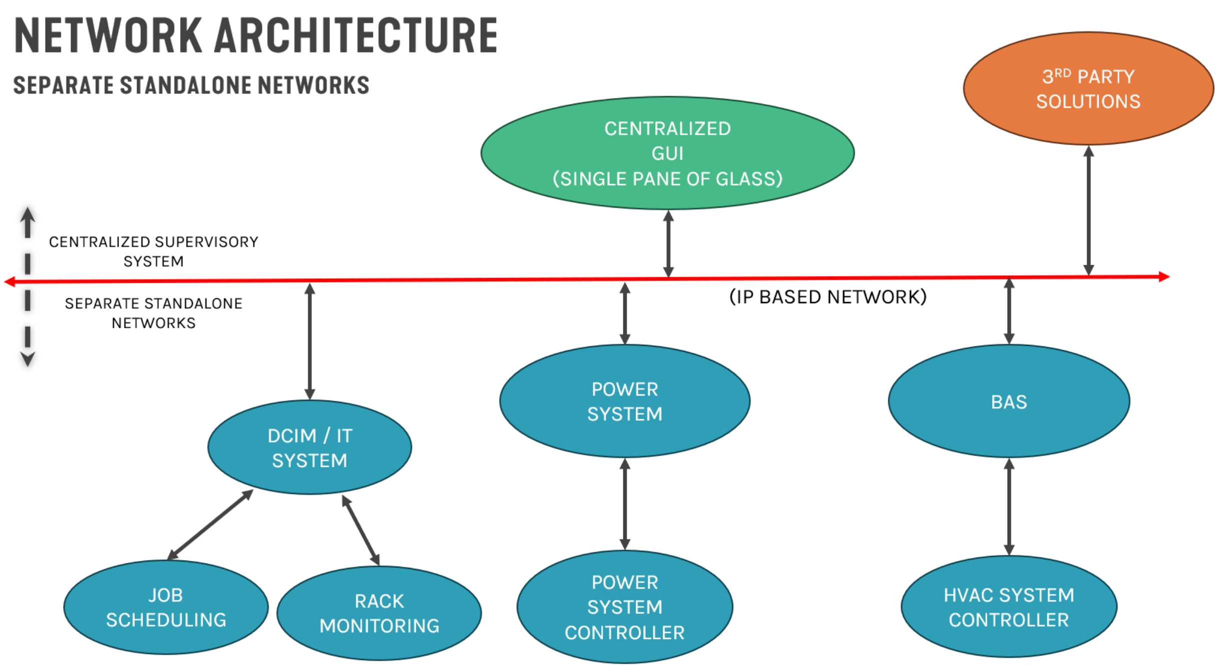

Figure 4. Holistic data center control system schematic. SmithGroup Image, 2026.

Integrated design is fundamental to the master plan that was create for Greenfield 450. By aligning building form and materials, understanding community integration, preparing for a heat reuse strategy, and adopting a holistic controls architecture from the earliest planning stages, the master plan created a roadmap to achieve superior performance, resilience, and adaptability while limiting the perceived burden to the utility infrastructure and surrounding neighborhood environment.

Case Study 2: Comparative Performance Analysis of Advanced Cooling Configurations for High-Density AI Data Centers

Introduction: Cooling system selection is a defining factor in data center efficiency, water consumption, and grid impact. As AI workloads drive facilities toward 100+ MW scale, the choice of cooling architecture (e.g., air-cooled versus water-cooled, dry versus evaporative heat rejection, and mechanical versus free-cooling operation) together determine the partial power usage effectiveness (pPUE) and water usage effectiveness (WUE) that a facility will achieve. This case study used an 8,760-hour simulation to evaluate five cooling system configurations across three U.S. climates and four chilled water supply temperature (CHWS) setpoints, quantifying the pPUE–WUE tradeoff frontier and revealing how climate and supply temperature interact to determine optimal architecture selection.

All efficiency values are partial PUE (pPUE), where

$$pPUE = \frac{(P_{IT} + P_{cooling})}{P_{IT}}$$

and $P_{cooling}$ includes only the modeled cooling subsystem chillers, pumps, fans, dry fluid coolers, and adiabatic pads. Total site PUE is typically 0.05 to 0.15 higher.

Methods

A physics-based hourly simulation engine modelled each architecture as interconnected components sized from manufacturer performance maps. For each of the 8,760 TMY hours modeled, the engine determined the operating mode (mechanical, free cooling, hybrid, or adiabatic-assisted) based on ambient conditions and computed component-level power and water consumption. Five configurations were evaluated (Table 1), simulated at 100 MW IT load with 28,500 TR total cooling capacity and N+X redundancy [where X = roundup(0.1N)]. Three representative climate zones, Albuquerque, NM (ASHRAE 4B, hot-dry), Aiken, SC (3A, hot-humid), and Idaho Falls, ID (6B, cold) were selected to span U.S. thermodynamic conditions. Each architecture was run at four CHWS setpoints (65°F, 72°F, 78°F, 86°F), yielding 60 annual simulations.

Table 1. Cooling System Configurations

#	Architecture	Description	Water
1	ACC	Air-cooled chiller w/ integrated dry cooler (mech. + free cooling)	None
2	WCC+DFC	Water-cooled centrifugal chiller + dry fluid cooler (strict modes)	None
3	WCC+DFC-H	WCC + DFC with hybrid free-cooling trim mode	None
4	WCC+DFC-A	WCC + DFC + adiabatic pre-cooling pad	Minimal
5	ACC+FC-A	Air-cooled chiller + free-cooling coil + adiabatic pad	Minimal

Architectures 2–4 represent variants of a 1 GW AI factory reference design (Trane Technologies 2026) using a water-cooled centrifugal chiller paired with a dry fluid cooler (DFC) for waterless free cooling. The hybrid variant (3) adds a trim mode where the chiller supplements the DFC during shoulder conditions. The adiabatic variant (4) adds pre-cooling pads upstream of the DFC to further extend free cooling at the cost of modest water consumption.

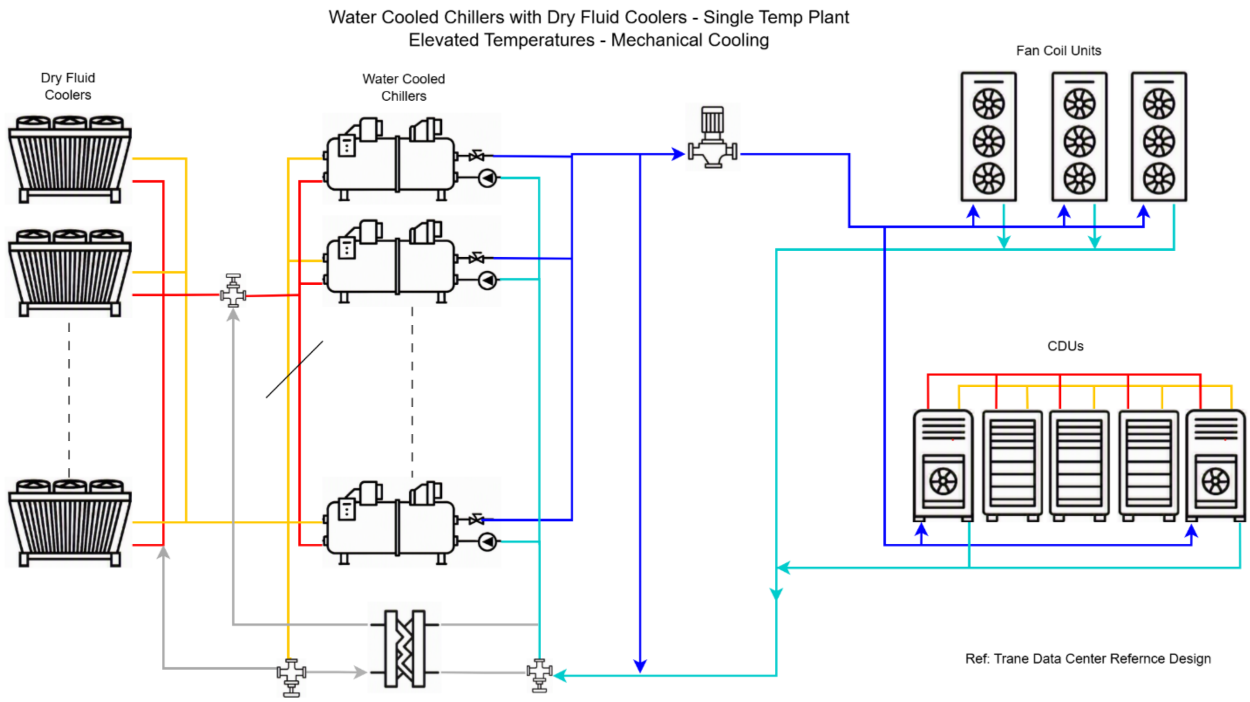

Figure 1. WCC + Dry Fluid Cooler architecture schematic (Architectures 2-4)

Results

A central finding is that advanced chilled-water architectures with dry free cooling achieve lower pPUE than the ACC baseline while consuming zero water. The WCC+DFC and WCC+DFC-H architectures deliver pPUE values of 1.05–1.08 with WUE = 0 across all climates. The only architecture with measurable water use is WCC+DFC-A, at 0.094–0.142 L/kWh (~22–33 million gal/yr for 100 MW). This targeted adiabatic strategy uses water selectively during the hottest hours to suppress peak pPUE excursions while maintaining near-zero water use for the majority of the year.

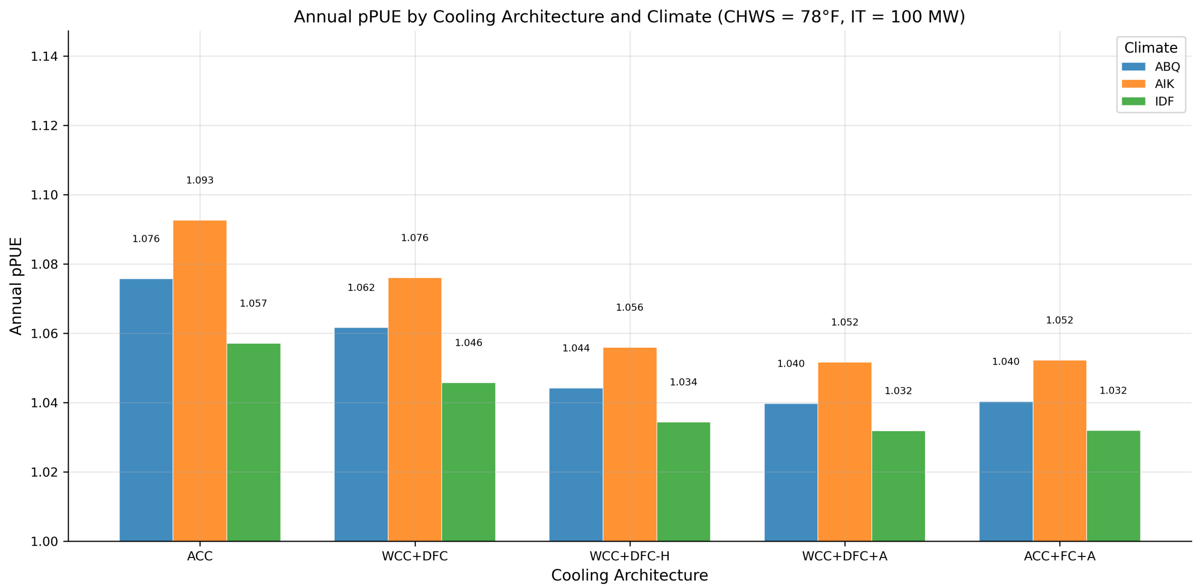

Figure 2. Annual pPUE by architecture at CHWS = 78°F across three climates.

CHWS Temperature Sensitivity: Raising supply temperature is a powerful efficiency lever for free-cooling architectures (Figure 3). The WCC+DFC-H architecture shows the strongest response: raising CHWS from 65°F to 86°F reduces annual pPUE by 0.039 in Albuquerque— equivalent to approximately 34 GWh of annual cooling energy savings for a 100 MW facility. At CHWS = 86°F, WCC+DFC-H operates in free-cooling or hybrid mode for 94.3% of the year. By contrast, the ACC pPUE plateaus between 78°F and 86°F, indicating diminishing returns for air-cooled systems at elevated supply temperatures.

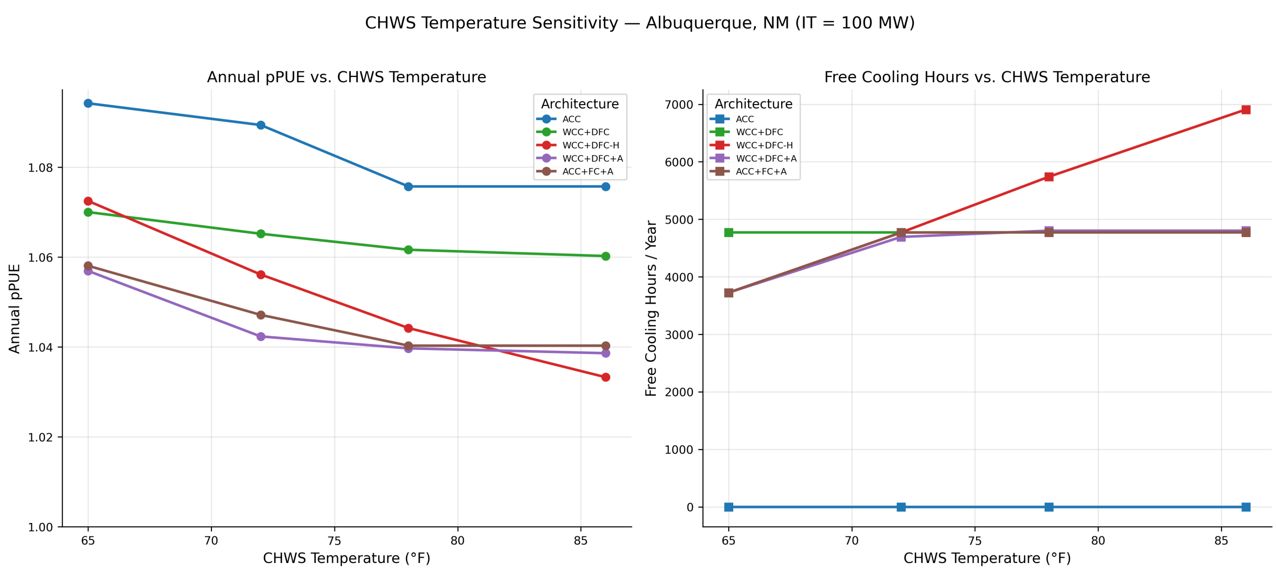

Figure 3. Annual pPUE sensitivity to CHWS temperature setpoint

Peak pPUE and Infrastructure Sizing: Annual pPUE masks significant hourly variation. Architectures with the lowest annual pPUE often exhibit the largest peak excursions: ACC+FC-

A achieves an annual pPUE of 1.040 in Albuquerque but spikes to 1.257 during the hottest hours, a gap of 0.217. The WCC+DFC architecture shows a more moderate gap of 0.097. For a 100 MW facility, a peak pPUE of 1.26 versus 1.16 translates to about 10 MW of additional peak cooling power demand. Because data centers are power-constrained, peak pPUE not annual pPUE governs electrical infrastructure sizing, capital cost, and grid interconnection capacity.

Equipment Sizing as a Water-Elimination Lever: Simulations shows that deploying 33% to 50% more dry cooler capacity raises the dry-to-adiabatic switchover temperature from between 78°F and 88°F to between 92°F and 98°F, reducing annual water consumption by more than 90%. This capital-for-water tradeoff enables near-zero-water operation even in hot-dry climates and should be evaluated alongside architecture selection during facility design.

Key Findings:

- Dry free-cooling architecture achieves lower pPUE than air-cooled baselines while consuming zero water. The WCC+DFC-H (hybrid) achieves annual pPUE as low as 1.05 in favorable climates, demonstrating that waterless operation does not require sacrificing cooling efficiency.
- CHWS temperature is one of the important efficiency levers. Raising the supply setpoint from 65°F to 86°F reduces annual pPUE by up to 0.04 for hybrid free-cooling architectures (about 34 GWh/yr savings for 100 MW).
- Peak pPUE governs facility power provisioning: Low-average architectures can have peak excursions to 1.26, requiring infrastructure sized for 26% cooling overhead versus the 6% suggested by the annual average.
- Targeted adiabatic cooling and equipment oversizing provide complementary water-minimization strategies, enabling near-zero WUE while maintaining best-in-class pPUE performance.

Overall, this case study presented a comparison between a variety of cooling architectures for AI data centers, documenting efficiency tradeoffs and energy and water usage comparisons. These comparisons enable informed decision-making when considering system types for large-scale data center design.

Concluding Discussion

This paper presents the AI Data Center Energy Performance Framework, an open access resource providing high-level guidance for efficient data center development across the project lifespan. Two case studies are discussed. The first shows a system type comparison between advanced cooling techniques and subsequent impacts on PUE and WUE. The second details measurable impacts of implementing an integrated design approach to data center development.

The Framework described in this paper is a living document. The principles, case studies, and guidance are housed in a digital format that can be updated as new information becomes available. This format was chosen because this field is rapidly advancing.

Future work will include continuous monitoring of industry developments to see where updates are needed in the Framework. Additionally, we will continue to seek additional case studies to better understand the geographical, capacity, and other considerations and variations in how these principles are applied.

References

Bessasparis, Stephen, Shana Ramirez, Emily Zhao, Ye Zhen, Kush Patel, and Isabelle Riu. 2025. *Forecasting Large Loads in the Age of AI and Data Centers*. Energy+Environmental Economics (E3). https://www.ethree.com/wp-content/uploads/2025/12/E3Whitepaper_DataCenterForecasting.pdf.

Bizo, Daniel. 2023. "Global PUEs — Are They Going Anywhere?" *Uptime Institute Blog*, December 4. https://journal.uptimeinstitute.com/global-pues-are-they-going-anywhere/.

Chen, Xin, Xiaoyang Wang, Ana Colacelli, Matt Lee, and Le Xie. 2025. "Electricity Demand and Grid Impacts of AI Data Centers: Challenges and Prospects." arXiv:2509.07218. Preprint, arXiv, November 26. https://doi.org/10.48550/arXiv.2509.07218.

Colangelo, Philip, Ayse K. Coskun, Jack Megrue, et al. 2025. "AI Data Centres as Grid-Interactive Assets." *Nature Energy*, December 5, 1–8. https://doi.org/10.1038/s41560-025-01927-1.

Davis, Jacqueline. 2024. "Large Data Centers Are Mostly More Efficient, Analysis Confirms." *Uptime Institute Blog*, February 7. https://journal.uptimeinstitute.com/large-data-centers-are-mostly-more-efficient-analysis-confirms/.

EPRI. 2025. *Scaling Intelligence: The Exponential Growth of AI's Power Needs*. White Paper. EPRI. https://www.epri.com/research/products/000000003002033669.

Hirschboeck, Peter Hans, Lisa Nguyen, Alex Hall, and Mark J. Cira. 2026. *Data Center Flexibility: Coordinating America's Digital and Energy Infrastructure*. impactECI. https://drive.google.com/file/d/1Mhi1HHY_f--Bgd8RwtIywm97_FH000Rp/view?usp=embed_facebook.

McCord, Kieren, Bing Liu, Thomas Loxley, Leigh Lain Walker, Casey Granata, and Todd Sims. 2026. "AI Data Center Energy Performance Framework." June 10. https://www.ashrae.org/test-pages/ai-data-center/ai-data-center-framework.

Shehabi, Arman, Alex Newkirk, Sarah J. Smith, et al. 2024. *2024 United States Data Center Energy Usage Report*. December 20. https://doi.org/10.71468/P1WC7Q.

Specian, Mike, and Alex Aquino. 2026. *Faster and Cheaper: Demand-Side Solutions for Rapid Load Growth*. ACEEE. https://www.aceee.org/sites/default/files/pdfs/faster_and_cheaper_-_demand-side_solutions_for_rapid_load_growth.pdf.

Trane Technologies. 2026. *High-Efficiency Cooling for the 1 GW AI Factory – CenTraVac® Simplex Reference Design*.

https://www.trane.com/content/dam/Trane/Commercial/global/markets/data-centers/Trane-ReferenceDesign-1GW-v8.pdf.

Uptime Institute. 2018. *Data Center Site Infrastructure Tier Standard: Topology*. https://uptimeinstitute.com/uptime_assets/889d381be3ad2900e838d497ba1ed0a2dbda4bbd31454b2d475eb986f2af3a55-00001E.pdf.

AI-assisted Tools: Disclosures and Responsibility

For the introductory material and AI Data Center Energy Performance Framework overview, PNNL's AI Incubator and Microsoft Copilot were used for light editing and copyediting of text. No AI-produced writing was copied or directly leveraged. The authors reviewed and verified all content, calculations, and citations. For the Greenfield 450 case study on integrated design Microsoft Copilot was used to provide review and summary of the SmithGroup case study report based on the topics and sub-topics selected by the SmithGroup authors. All concepts and language used in the case study originate in the SmithGroup report. The authors reviewed and verified all content and calculations.