WHITEPAPER

Bridging the White and Gray Space Data Center Silos

Integrating Workload Clustering and Zonal Cooling to Maximize Compute Per Megawatt™

March 2026

PADO AI ORCHESTRATION

Table of contents

Abstract

Most data centers operate at only 40–60% of their true potential. The remaining capacity isn't missing—it's stranded by misalignment in how workloads are organized and inefficiencies in how cooling resources are deployed. Unlocking this capacity means unlocking revenue.

This white paper details PADO's integrated optimization framework, which uses Workload Clustering and Zone-Targeted Cooling to minimize stranded capacity by decoupling power constraints with compute constraints. By bridging the traditional silos between white space (predictive job packing) and gray space (zonal cooling), PADO maximizes Compute Per Megawatt™. Historically, the operation of the white and gray space has occurred in silos with separate entities optimizing to different goals and in the process stranding capacity. The PADO approach, as described in this paper, connects the white and the gray, creating transparency in the white space to eliminate operational silos, thus increasing throughput via workload clustering with HVAC zone targeting.

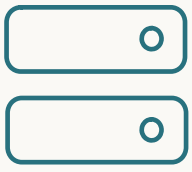
Workload Clustering

Understanding specific job demands is essential for optimization. Different workloads—from steady AI training to bursty inference—impose unique demands on power and infrastructure. Clustering these jobs by a broader set of attributes (thermal profile, sovereignty, priority, etc.) allows for more precise capacity planning and cooling.

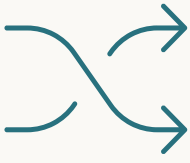
Zone-Targeted Cooling

This strategy complements white space optimization by matching cooling delivery to a zone's specific thermal needs. This precision allows for increased temperature setpoints in cooler zones while maintaining safety in high-density areas through variable control.

15-25%
Improvement in effective utilization



8-15%
Improvement in PUE



20-40%
Improvement in throughput density

Introduction

While legacy enterprise racks typically operated at 5–10 kW, modern AI clusters have quickly pushed past the 40–60 kW threshold as demand for higher compute density continues to grow.

NVIDIA's Blackwell GB200 NVL72 architecture now benchmarks this shift, with a single rack engineered for a power draw of approximately 120 kW—a massive leap from the 10.2 kW per node required by previous DGX H100 systems. This 12x density increase makes reclaiming every watt of underutilized "stranded" power an operational necessity.

This power density increase creates cascading challenges:

Cooling capacity constraints
Traditional air cooling systems designed for 5–10 kW/rack cannot adequately cool 40+ kW/rack deployments [Patterson2008]

Power distribution limitations
Existing electrical infrastructure may lack capacity for high-density zones [Rasmussen2011]

Stranded capacity
Thermal or power constraints in one area prevent utilization of available capacity elsewhere, creating "stranded" white space [Greenberg2009]

Thermal hotspots
Concentrated heat sources lead to localized overheating, risking equipment failure [Moore2006]

Despite the urgent demand for new infrastructure, a persistent paradox remains: average server utilization in traditional enterprise data centers typically lingers between 6% and 12%, leaving a vast majority of expensive compute capacity idle and unproductive [McKinsey 2023]. This idle silicon creates a massive economic and environmental burden as organizations struggle to secure scarce power.

As chip architectures move rapidly from general-purpose CPUs to specialized GPUs, TPUs and task-specific ASICs, the useful life of a hardware asset is being decoupled from its accounting depreciation schedule. Data centers have the immediate ability to maximize utilization through portable workloads by moving jobs dynamically between chip types to ensure every watt of power is utilized profitably. Operators must move beyond static infrastructure and SLA models and embrace a strategy that balances extreme power density with the technical flexibility to shift workloads as hardware and demand evolve.

Problem Statement

Stranded Capacity

Stranded capacity is where available white space cannot be utilized due to localized constraints such as [Greenberg2009]:

Power stranding:
Cooling capacity exists but power circuits are exhausted

Cooling stranding:
Power capacity exists but cooling is inadequate

Network stranding:
Compute capacity exists, but network limits deployment.

Optimization problem

In data centers with heterogeneous workloads and fixed infrastructure capacity, the challenge is optimizing workload placement and cooling to maximize compute throughput per unit area while minimizing energy use and maintaining ASHRAE-compliant thermal conditions.

This is a multi-objective optimization problem with the following objectives

- Maximize utilization of compute resources
- Minimize PUE through cooling optimization
- Maximize throughput density
  (e.g. tokens/sqft for AI workloads)

Subject to constraints

- Thermal limits per ASHRAE guidelines
- Power capacity at rack, row, and facility levels
- Cooling capacity by zone
- Workload placement constraints
  (data locality, latency requirements)

Consider a typical scenario

A data center row has available rack space, but the cooling capacity for that zone is already maxed out by neighboring high-intensity workloads. Adding more equipment would cause overheating. The space is available in theory but unusable in practice.

Traditional Data Center Layout

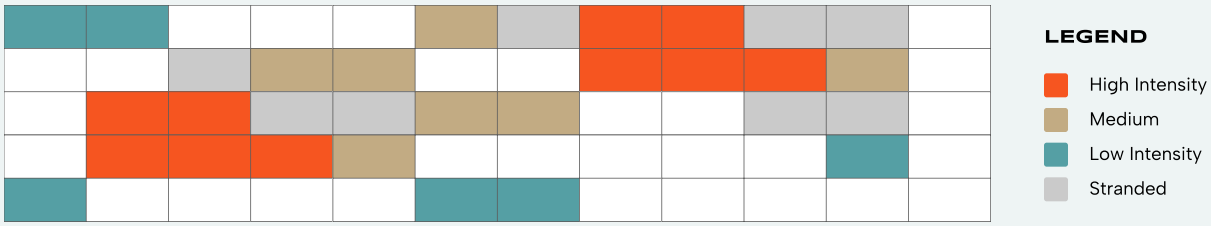


Figure 3: Stranded capacity scattered throughout a traditional data center floor

Why Traditional Approaches Fall Short

Traditional data center management treats the facility as a uniform resource pool. Workloads are placed based on available slots, and cooling is provisioned uniformly across the floor. This approach worked adequately when workloads were relatively homogeneous, but it fails in today's environment for three key reasons:

Workload diversity creates thermal noise

When AI training jobs sit next to web servers, and batch analytics runs alongside real-time inference, the resulting heat map becomes unpredictable. Cooling systems designed for average loads either over-cool some areas (wasting energy) or under-cool others (creating thermal risks).

Static provisioning cannot adapt

Traditional cooling infrastructure is sized for worst-case scenarios across the entire facility. This conservative approach ensures safety but guarantees inefficiency.

Utilization variability compounds the problem

Workloads with highly variable resource demands require larger safety margins, further reducing effective capacity.

Overview: PADO Solution

Our Approach: Increase Throughput via Workload Clustering with HVAC Zone Targeting

The White Space Optimization system addresses this problem through two complementary strategies:

Workload Clustering

Grouping similar workloads together to reduce utilization variability, improve packing efficiency, and create predictable thermal zones. This approach is inspired by bin-packing algorithms [Coffman1996] and VM placement research [Verma2008].

HVAC Zone Targeting

Dynamically allocating cooling resources based on actual heat load rather than static provisioning. This enables temperature setpoint increases in cool zones while maintaining thermal safety in hot zones, following the variable air volume (VAV) paradigm [ASHRAE2019].

The aggregate value of these strategies can be measured Compute Per Megawatt™

Maximizing Compute Per Megawatt™ serves as a comprehensive KPI for the modern AI data center, shifting the industry's focus from mere energy savings (PUE) to revenue-generating output. It is a vital metric because it quantifies the "stranded capacity" typically lost to thermal imbalances and inefficient workload distribution, allowing operators to visualize exactly how much compute they are leaving on the table.

What makes this metric unique to PADO is its integrated nature: while traditional tools silo "Gray Space" (facility power/cooling) and "White Space" (IT workloads), PADO's formula treats them as a single, interdependent physics problem. By mathematically linking compute rates, cluster locality, and HVAC zone targeting, **PADO provides the only framework that can predictably convert a saved watt of cooling into an active watt of compute, effectively allowing data centers to increase their productive capacity without requesting a single additional megawatt from the utility grid**.

White Space

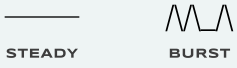
Understanding what runs inside a data center is essential for optimizing its operation. As the fundamental resource-consuming units of work in data centers, different job types impose unique demands on power, cooling, and computational infrastructure.

What is a Job?

In data center terminology, a job represents a discrete unit of computational work that:

- Arrives at the data center at some time (either scheduled or on-demand)
- Consumes resources including CPU cycles, GPU compute, memory, storage, and network bandwidth
- Generates heat as a byproduct of computation
- Completes after some duration, releasing its resources

Why Job Characteristics Matter



Different jobs create fundamentally different demands on data center infrastructure:

- **Power consumption:** Jobs vary from 0.5 kW (simple web services) to 40+ kW (GPU training clusters)
- **Duration:** Some jobs run for milliseconds, others for weeks
- **Predictability:** Training jobs have steady, predictable power draw; inference jobs fluctuate with user demand
- **Resource mix:** Some jobs need CPUs only; others require specialized hardware like GPUs or TPUs

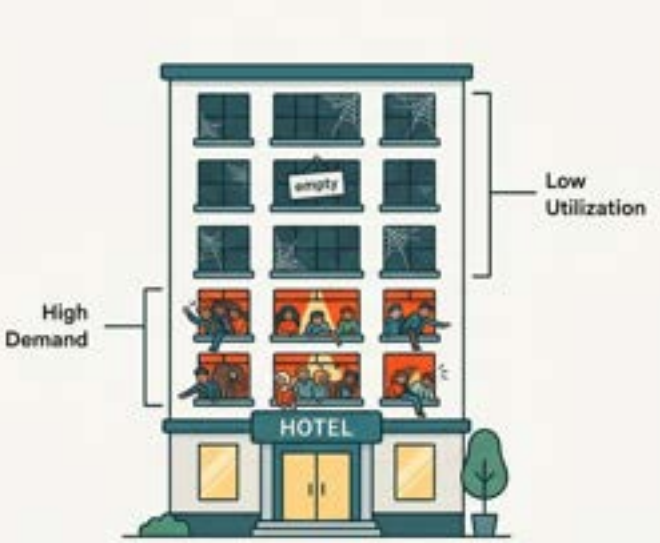
These differences impact cooling needs, capacity planning, and operating costs.

Racks running steady AI training need consistent cooling, while bursty inference racks experience rapid thermal cycling.

The Core Challenges: Variability and Packing

Imagine you manage a hotel chain with 100 rooms across 10 hotels. On average, 60 rooms are occupied each night—a 60% occupancy rate. But this average hides important details:

Utilization Variability: The Hotel Analogy

Scenario A: High Variability	Scenario B: Low Variability
	
Inefficient Waste Some hotels are 95% full while others are 25% full. The busy hotels turn away customers (lost revenue), while empty hotels waste resources (heating, staffing).	**Optimized Efficiency** All hotels hover around 60% occupancy. No customers are turned away, and resources are used proportionally.

Both scenarios have the same average occupancy, but Scenario B is far more efficient and profitable

Data centers face the same challenge:

When some racks run at 90% while others sit at 30%, the data center experiences

Thermal hotspots
High-utilization racks generate concentrated heat that strains cooling systems

Stranded capacity
Low-utilization racks represent expensive equipment sitting idle

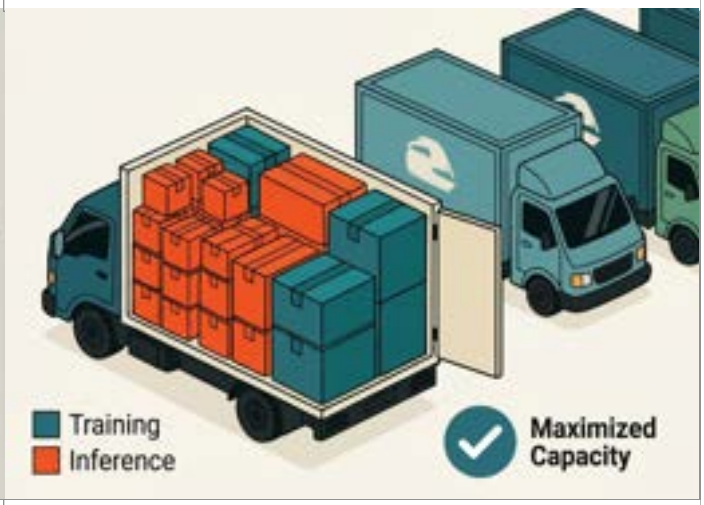
Operational complexity
Racks require varying levels of management

Reduced reliability
Overloaded racks are more prone to failures

Bin Packing

Imagine you're moving and need to pack boxes into a moving truck:

Job Packing: The Moving Day Analogy

Inefficient: Haphazard Placement	Efficient: Workload Clustering
	
Scattering diverse jobs randomly leads to large gaps and wasted “stranded” capacity.	Grouping similar items by size and demand allows for tighter, high-density packing.

Result: Run more workloads on the same infrastructure

- You have items of various sizes (small books, medium appliances, large furniture)
- Each truck has a weight/volume limit it cannot exceed
- Your goal: use as few trucks as possible while not overloading any truck

An inexperienced packer loads trucks haphazardly, leaving many half-empty.
An expert groups similar items to fill each truck efficiently.

In data center terms:

Items
Jobs with power demands (the "weight" of each workload)

Bins
Racks with power capacity (the "truck capacity")

Goal
Maximize utilization without exceeding any rack's power limit

Intelligent Workload Clustering

This strategy is deceptively simple in concept, but objectively hard in practice given the speed of jobs, unpredictability of jobs and segmented infrastructure. Intelligent workload clustering groups similar workloads together. Rather than scattering AI training jobs, web services, and batch analytics randomly across the floor, organize them into dedicated zones based on their operational characteristics.

Figure 1: Workload clustering transforms a chaotic mix of job types (left) into organized zones where similar workloads are co-located (right). This enables zone-specific cooling and reduces resource contention.

Implementing this requires a departure from static rack-level assignments in favor of intelligent workload clustering. By categorizing workloads into specific profiles, the system can dynamically map jobs to the optimal zone.

Workload Clustering Variables include, but are not limited to:

Resource Profile
Defines hardware needs including CPU, GPU, memory, storage, and network ratios.

Power Demand Characteristics
Tracks average power draw and consumption variability across workloads over time.

Temporal Patterns
Analyzes job duration, arrival rates, and burst or diurnal workload patterns.

Thermal Compatibility
Identifies heat profiles to group workloads for efficient zone-based cooling.

Data Sovereignty & Compliance
Defines data residency and regulatory constraints.

Workload Priority & Preemption
Defines job priority and SLA tiers during power constraints.

Forecast

By integrating physics-based heat transfer modeling with queuing theory, the platform simulates millions of scenarios to predict thermal and power dynamics on a rack-by-rack basis. This forecast provides transparency from the white space to the gray space and allows operators to transition from conservative "worst-case" static planning to dynamic operation. The primary benefit is the creation of dynamic power headroom.

PADO and its partners leverage these forecasts to drive multi-layer optimization across the data center's physical and digital infrastructure. By feeding predictive workload data directly into the Building Management System (BMS), PADO proactively modulates cooling loops and airflow to meet impending thermal loads. Simultaneously, the platform synchronizes these forecasts with onsite energy assets—including Uninterruptible Power Supplies (UPS) and Battery Energy Storage Systems (BESS)—to buffer peak demand, manage state-of-charge for grid-response readiness, and ensure seamless power continuity during high-density workload shifts.

The benefits of clustering become clear:

Predictable thermal profiles
Each zone has a characteristic heat load, enabling targeted cooling.

Reduced contention
Jobs with similar resource profiles interfere less with each other.

Improved packing
Homogeneous demand within zones allows tighter bin-packing.

Simplified operations
Capacity planning becomes zone-based rather than rack-by-rack.

Grey Space

Data centers are essentially large-scale heat management systems. Every watt of electricity consumed by IT equipment is ultimately converted to heat, which must be removed to prevent equipment damage. Understanding this thermal challenge is essential for optimization.

The Cooling Challenge

Consider a 10 MW data center

IT equipment generates 10 MW of heat continuously	This equals approximately 34 million BTU/hour	Equivalent to the heat output of 3,400 homes	Concentrated in approx 50,000 sq ft

Without active cooling, server inlet temperatures would rise above safe limits within minutes, causing thermal throttling and eventual hardware failure [Moore2006].

The Problem with Uniform Cooling

Traditional data centers supply uniform airflow across all zones.

This creates two problems:

Overcooling of low-load zones
Zones with light workloads receive more cooling than needed, wasting fan and chiller energy.

Undercooling of high-load zones
Hot zones may not receive enough airflow, causing hot spots and thermal throttling.

Zone-Targeted Cooling

The second strategy complements the white space optimization strategy:

Once workloads are clustered into zones with predictable thermal profiles, cooling can be precisely targeted to each zone's actual needs.

Precision Over Uniformity: Zone-Target Cooling

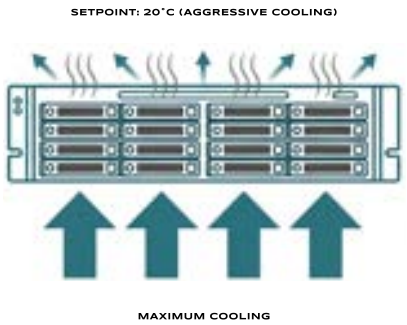
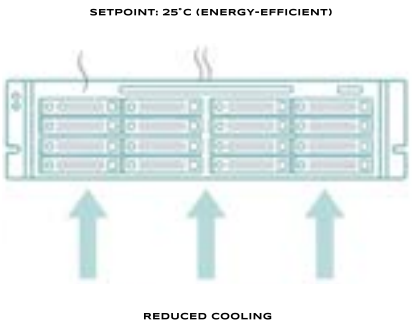
High-Intensity Zone	Low-Intensity Zone
	
High-intensity zones receive aggressive cooling matched to their heat output.	**Low-intensity** Low-intensity zones run at higher ASHRAE-compliant temperature setpoints, reducing cooling energy.

Figure 5: Zone-targeted cooling matches cooling intensity to actual heat load

Traditional cooling treats every rack like the hottest; zone targeting cools by actual need.

Many data centers maintain uniform, conservative cooling across the floor to protect hot spots, often wasting energy by over-cooling racks that don't require it.

Liquid Cooling

Enabling precision thermal control for high-density AI environments

- Integrates with Direct-to-Chip (D2C) plates and Coolant Distribution Units (CDUs) to manage high-density AI thermal loads
- Synchronizes coolant flow with real-time GPU utilization for zone-specific thermal relief
- Eliminates over-cooling of lower-density racks, improving overall energy efficiency
- Feeds predictive workload signals into the BMS for proactive pump and loop adjustments
- Replaces static safety margins with dynamic, zone-level setpoints to support 100kW+ per rack scaling

Battery Energy Storage Systems (BESS)

Turning energy storage into intelligent, revenue-generating infrastructure

- Transforms batteries into compute-aware, revenue-generating infrastructure
- Reduces peak demand charges and improves site-level energy economics
- Enables energy arbitrage and renewable load shifting
- Strengthens backup readiness through outage simulation and SLA prioritization
- Qualifies assets for grid services while ensuring SLA-safe dispatch participation

Control Systems Integration

Creating truly AI driven Building Management Systems (BMS)

- Control algorithm computes optimal airflow distribution – cooling can become demand-shaped, not rule-based.
- Variable frequency drives (VFDs) adjust fan speeds – thermal optimization can dynamically happen in rate-limited, safety-checked, logged & reversible approaches.
- Feedback loop maintains target temperatures – even under volatile workloads.
- Integration with Existing BMS/EMS Platforms – providing an intelligence layer, not a replacement.

Preserve Infrastructure Resilience

Turning white and gray space data into predictive, applied operating envelopes

- Operate closer to peak utilization without crossing safe boundaries, protecting GPUs, CPUs, and power electronics.
- Thermal spikes are can be managed proactively, not reactively—preventing hotspot formation and emergency throttling.
- Implement thermal alerts before reaching critical thresholds – shift from firefighting to preventive control, reducing incident frequency and downtime.
- Extended equipment life, stronger warranty positions, and lower lifecycle CapEx.

Optimization: How the Two Strategies Work Together

The magic happens when workload clustering and zone-targeted cooling work in concert:

1 Assign Zones

Physical zone assignment maps these logical clusters to physical locations on the data center floor.

3 Analyze Workloads

Identify each job type's operational profile: resource demand, utilization patterns, and thermal output.

2 Cluster Groups

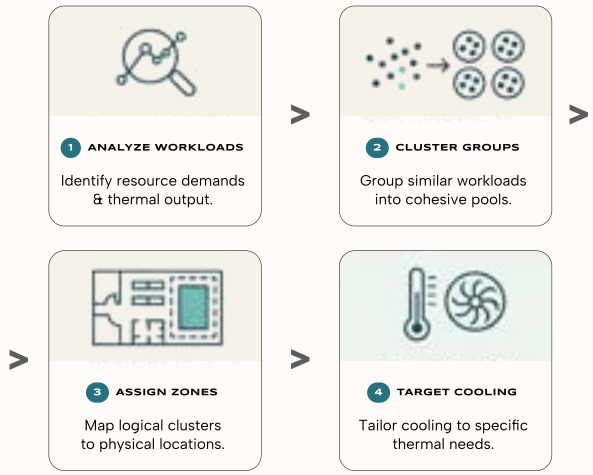
Group similar workloads into cohesive workload pools.

4 Cooling Target

Cooling optimization tailors each zone's cooling parameters to its specific thermal requirements.

The result: a smarter data center

Resources flow where needed, cooling targets demand, and stranded capacity becomes usable.



Result: Higher Utilization + Lower Consumption

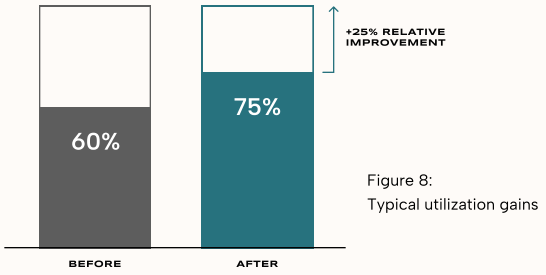


Figure 8:
Typical utilization gains

The Logic: Reduced Variability

Standard Deviation (σ) drops from 18% to 11%. Grouping similar workloads creates a higher safe utilization floor without risking hotspots.

The Result

Gaining capacity without adding hardware.

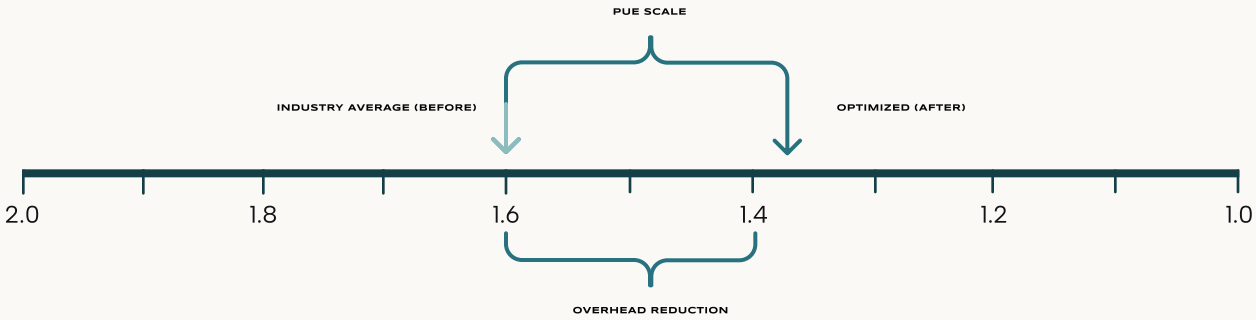
Real World Impact

Organizations implementing intelligent workload placement and zone-targeted cooling consistently achieve significant improvements across multiple dimensions.

Utilization Improvements

When similar workloads are grouped together, variability within each zone decreases. With more predictable demand, operators can safely run infrastructure at higher utilization without increasing the risk of contention or thermal issues.

Organizations typically see 15–25% relative utilization gains. A facility running at 60% utilization can often reach 72–75% after optimization, effectively unlocking additional capacity without adding hardware.



Impact on a 10 MW Facility:

- ~1 MW of Saved Overhead Power
- Hundreds of thousands in annual electricity savings
- Power freed up for actual compute

PUE Improvements

Zone-targeted cooling delivers substantial energy savings. By matching cooling intensity to actual thermal load in each zone, facilities avoid the waste inherent in uniform over-cooling.

Zone-targeted cooling delivers substantial energy savings. By matching cooling intensity to actual thermal load in each zone, facilities avoid the waste inherent in uniform over-cooling.

For a facility consuming 10 MW of power, a PUE improvement from 1.60 to 1.44 represents approximately 1 MW of saved overhead power—translating to hundreds of thousands of dollars in annual electricity savings.

- Predictable thermal profiles: Each zone has a characteristic heat load, enabling targeted cooling.
- Reduced contention: Jobs with similar resource profiles are less likely to interfere with each other.
- Improved packing: Homogeneous demand within zones allows tighter bin-packing.
- Simplified operations: Capacity planning becomes zone-based rather than rack-by-rack.

Key Insight

Power Usage Effectiveness (PUE) measures how efficiently a data center uses power. A PUE of 1.60 means 60% of power goes to cooling and infrastructure overhead. Reducing PUE to 1.44 means more power reaches computing equipment—and less is wasted.

Limitations of Power Usage Effectiveness (PUE)

Critics have noted that PUE alone does not capture overall efficiency [Brady2013]:

- **IT equipment utilization is ignored:** A facility with PUE 1.2 but 20% server utilization is less efficient than one with PUE 1.5 and 80% utilization
- **Measurement inconsistencies:** Different measurement methodologies yield incomparable results [Uptime2014]
- **Perverse incentives:** Operators may over-cool to protect PUE targets rather than optimizing holistically

This motivates our multi-metric approach, considering utilization, PUE, and throughput density simultaneously.

Throughput Improvements

The combined effect of higher utilization and lower cooling overhead is dramatically improved throughput—more computing work accomplished per unit of floor space.

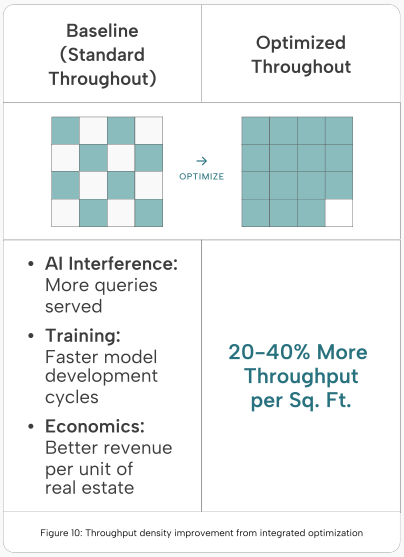


Figure 10: Throughput density improvement from integrated optimization

For AI inference workloads, this means more model queries served per facility. For training workloads, faster model development cycles. For any compute-intensive application, better economics per unit of real estate.

Sustainability Benefits

Efficiency improvements deliver environmental benefits:

Environmental Impact

Reduced carbon footprint
Less energy consumed means fewer emissions

Lower water consumption
Efficient cooling reduces evaporative water use

Extended facility lifespan
Better utilization delays new construction

Measurable sustainability
Clear data for ESG reporting

Revenue Generation Drivers

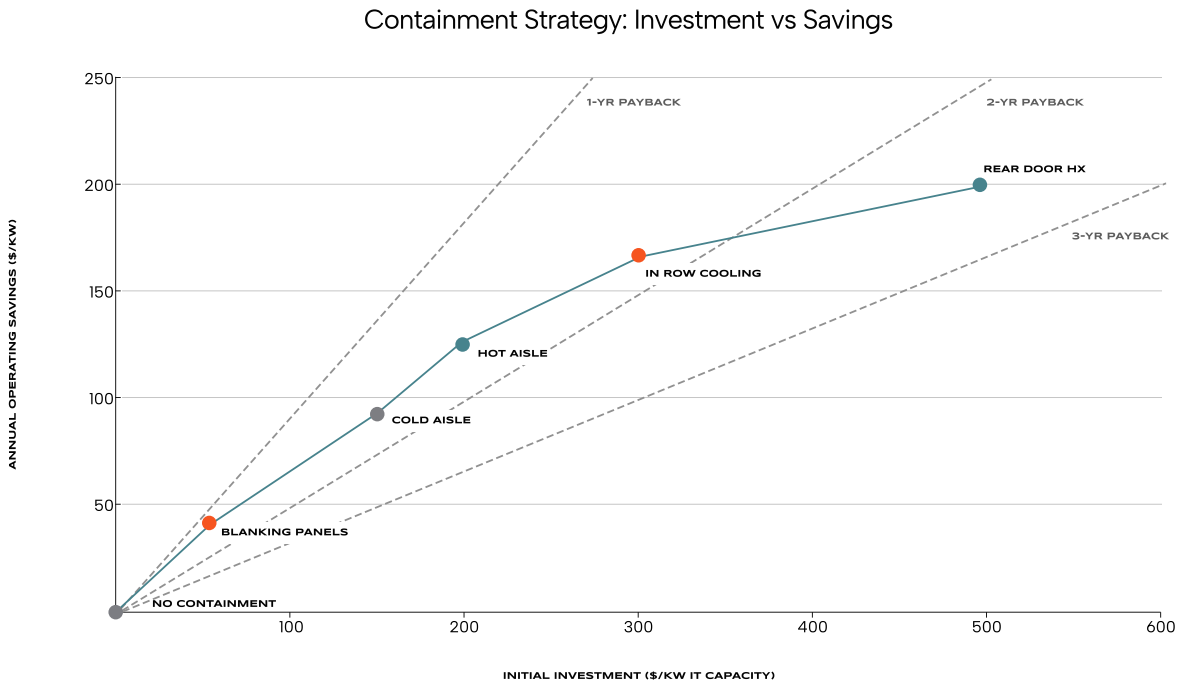
Increased billable capacity
More compute available for revenue workloads

Faster time to market
Deploy services using unlocked capacity

Competitive pricing
Lower compute costs enable stronger pricing

Avoided capital expenditure
Delayed construction preserves capital

Cost Savings (Secondary Benefits):



Energy cost reduction
PUE improvements of 8–15% translate to meaningful electricity savings

Deferred infrastructure investment
Maximizing existing capacity delays costly expansion projects

Reduced maintenance complexity
Organized workload zones simplify operations to conservative cooling and unmanaged workload variability. This study demonstrates how PADO's White Space Optimization—integrating Workload Clustering with HVAC Zone Targeting—reclaims this capacity in a 1,000 NVIDIA H100 GPU reference architecture.

The Optimization Levers

Workload Clustering

Segmenting workloads into four thermal zones reduced utilization variability by 39%, allowing for higher safe packing density.

HVAC Zone Targeting

Transitioning from monolithic cooling to zonal VAV control reduced total cooling power by 40%.

Optimization Impact: Baseline vs. PADO Orchestrated

The following table integrates facility specifications with real-world performance gains.

Metric	Baseline (Pre-Opt)	PADO Optimized	Delta / Improvement
Active Compute	1,000 H100 GPUs	1,000 H100 GPUs	Reference Constant
Mean GPU Utilization	60%	72%	+20% Yield
IT Power Capacity	8.33 MW	8.33 MW	Fixed Asset
Cooling Power Req.	3.33 MW	2.00 MW	-1.33 MW Saved
Total Facility Power	13.33 MW	12.00 MW	-10% OpEx
Facility PUE	1.60	1.44	-10% Efficiency
Monthly Throughput	933.12 B Tokens	1,119.74 B Tokens	+186.6 B Tokens
Throughput Density	888 Tokens/sqft/s	1,066 Tokens/sqft/s	+20% ROI/sqft

Financial & Strategic Outcome

By converting 1.33 MW of "thermal overhead" into active compute capacity, the facility achieves a significant shift in unit economics:

- Annual Revenue Uplift: +$9.86M (based on $4.40/M tokens)
- Annual Energy Savings: +$932k (at $0.08/kWh)
- Total Annual Benefit: +$10.79 Million
- 5-Year Project NPV: $32.26 Million (20% discount rate)

PADO's integrated optimization workflow, transforms the data center from a static utility into a dynamic compute engine. In this 1,000 GPU scenario, the platform successfully recovered enough stranded power to support 20% higher throughput while simultaneously reducing the total power draw by 10%.

Conclusion

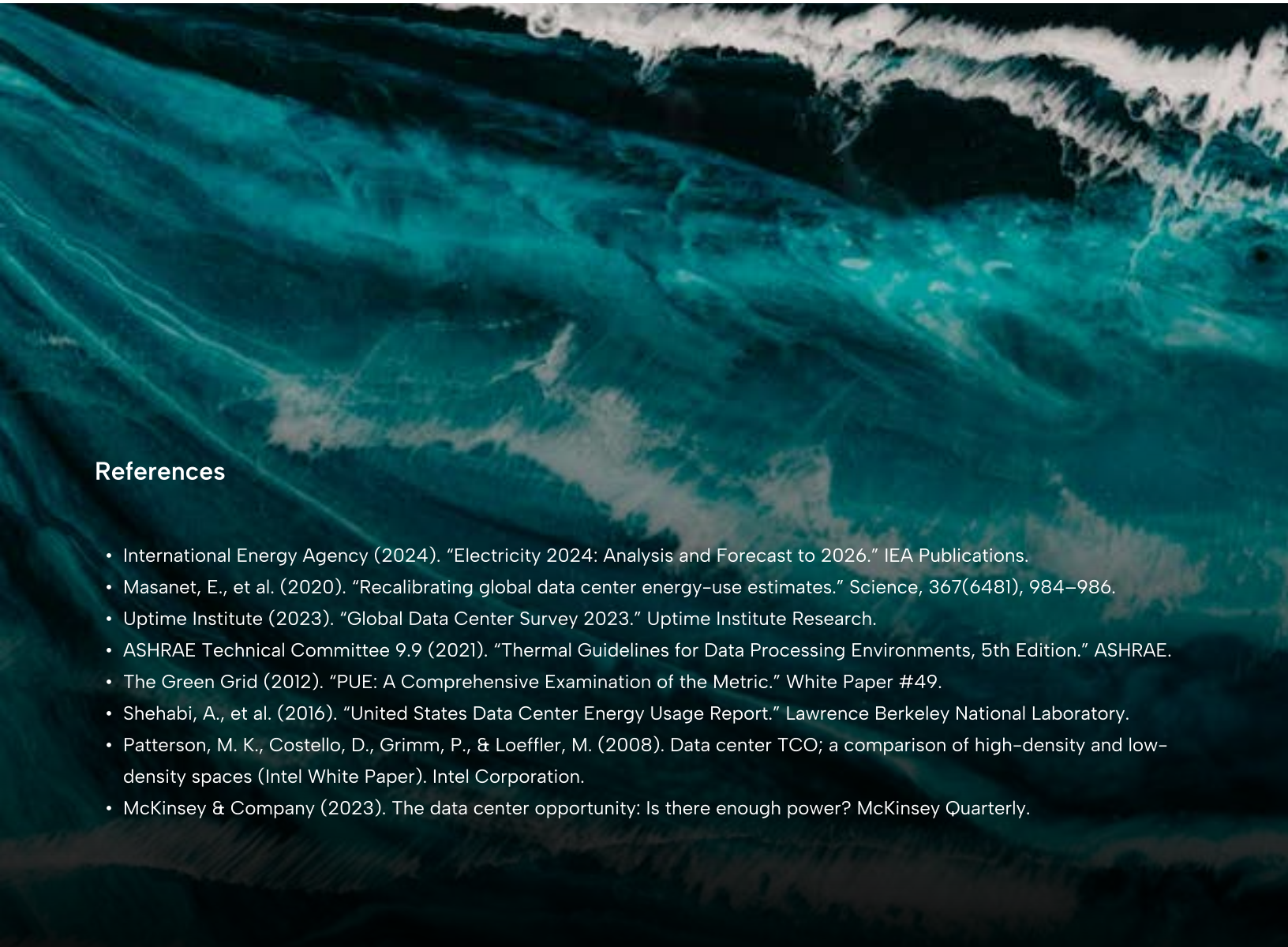
The transition to high-density AI infrastructure has made legacy, static data center management obsolete. As rack densities reach 120kW and beyond, the traditional silos between white space (IT) and gray space (facilities) create significant stranded capacity that operators can no longer afford to ignore.

PADO's integrated optimization framework bridges this gap, transforming the data center into a dynamic, physics-aware compute engine. By synchronizing Workload Clustering with Zone-Targeted Cooling, the platform eliminates the need for wasteful, facility-wide safety margins and replaces them with precise, predictive control.

The results are clear: by reclaiming stranded capacity, operators can realize a 20% increase in throughput and a 10% improvement in PUE within their existing footprint.

Ultimately, PADO's approach ensures that as silicon continues to specialize and power demands grow, the data center remains a flexible asset capable of Maximizing Compute Per Megawatt™ and driving superior unit economics.

Ready to learn more? | pado.co | info@pado.co



References

- International Energy Agency (2024). "Electricity 2024: Analysis and Forecast to 2026." IEA Publications.
- Masanet, E., et al. (2020). "Recalibrating global data center energy-use estimates." Science, 367(6481), 984–986.
- Uptime Institute (2023). "Global Data Center Survey 2023." Uptime Institute Research.
- ASHRAE Technical Committee 9.9 (2021). "Thermal Guidelines for Data Processing Environments, 5th Edition." ASHRAE.
- The Green Grid (2012). "PUE: A Comprehensive Examination of the Metric." White Paper #49.
- Shehabi, A., et al. (2016). "United States Data Center Energy Usage Report." Lawrence Berkeley National Laboratory.
- Patterson, M. K., Costello, D., Grimm, P., & Loeffler, M. (2008). Data center TCO; a comparison of high-density and low-density spaces (Intel White Paper). Intel Corporation.
- McKinsey & Company (2023). The data center opportunity: Is there enough power? McKinsey Quarterly.