2025 33rd International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS) | 979-8-3315-5760-7/25/$31.00 ©2025 IEEE | DOI: 10.1109/MASCOTS67699.2025.11283387

CAPLAI: AI-assisted Lifecycle Provisioning for GPU data centers

Chengyi Nie*, Anna Xing†, Imran Latif‡, Zhenhua Liu*

*Stony Brook University, Stony Brook, NY, USA
{nie.chengyi, zhenhua.liu}@stonybrook.edu
†Ward Melville High School, East Setauket, NY, USA
annaxing2020@gmail.com
‡Brookhaven National Laboratory, Upton, NY, USA
ilatif@bnl.gov

***Abstract*—The accelerating deployment of artificial intelligence (AI) workloads, driven by recent advancements in AI technology, has significantly increased the demand for computing resources and supporting infrastructure. However, the physical and electrical capacity of data centers cannot scale at the same pace, which introduces a new challenge: accommodating rising compute demand under stringent power and space constraints. New high performance GPUs offer better power efficiency compared to previous generations, and liquid cooling systems are significantly more efficient than traditional air cooling. These advancements create an opportunity to upgrade data centers that accommodate more AI workloads with limitations in physical infrastructure and power availability.**

We propose CAPLAI, *Capacity-Aware PLanning for AI infrastructure*, an AI-assisted stochastic optimization framework for lifecycle planning in GPU data centers. Our method adopts large language models (LLMs) to generate diverse and plausible future scenarios that capture demand growth, hardware efficiency decay, electricity prices, and resale market trends. These scenarios feed into a stochastic optimization model that determines GPU purchase, retirement, and cooling infrastructure upgrades. We evaluate our framework using real-world traces and constraints derived from a large-scale AI data center at Brookhaven National Lab. Compared to conventional threshold-based heuristics, our approach increases the effective GPU computing capacity within the same power limit by up to 36% and reduces lifecycle operating cost by up to 32%. Results demonstrate that capacity-aware, AI-guided planning significantly enhances efficiency and robustness with the escalating demand and infrastructural limits.

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

The advancement of artificial intelligence (AI) has led to the wide adoption of large language models (LLMs) and generative AI services, resulting in significant growth in GPU-based data centers. Data centers are critical infrastructures, providing the computational backbone necessary for both the AI models' training and inference. According to recent industry analyses, the computing capacity of AI data centers needs to grow by 20–30% annually to meet demand, resulting in significantly higher GPU deployment and corresponding energy consumption [1].

A major problem of AI data centers is the high power density of modern AI hardware. For example, a Nvidia H100 GPU server can draw up to 6 kilowatts, compared to a typical CPU server—with a similar physical size in the rack—which consumes around 500 watts [2]. Standard air cooling is becoming insufficient as rack power exceeds 30 to 50 kW [3]. While a standard 42U rack can host 42 1U servers [4]–[6], it can only hold 5 to 8 GPU servers, which leads to significant space underutilization and low computing power. Fan power increases cubically with speed, but airflow delivers diminishing cooling returns due to air's low heat capacity. The consequence is not only energy waste but also potential performance bottleneck. In many facilities today, nearly half of the power goes to cooling overhead rather than computing work [7]. Since the power supply to a data center is often fixed, increased cooling power inevitably reduces computing capability, especially as computing demand continues to rise. Liquid cooling transfers heat more efficiently, reducing both energy waste and thermal risk. With high-thermal-capacity fluids in direct contact with hot components, liquid cooling can remove heat far more efficiently than air [8], [9]. Hence, liquid cooling enables much higher compute density per rack and extends the hardware lifespan [10]. This means more GPUs and more computing power can be packed into the same rack, supporting the growing AI workloads within existing data center infrastructure and power supply.

Traditional methods for GPU lifecycle planning often rely on static replacement cycles or heuristic-based strategies [11], [12]. It's challenging for heuristic methods to address the complex uncertainties inherent in GPU data center operations and accommodate various scenarios. For instance, workload demands for AI services can be extremely volatile due to variable user adoption, service expansion, and technological advancements [13]. Moreover, hardware performance typically decayed non-linearly over time due to aging, frequent failures, and workload variability [14]. The rapid release of new GPU models also contributes to fluctuations in resale values, making financial planning for hardware upgrades and retirements more complicated. Therefore, heuristic methods often lead to scenarios of either over-provisioning, resulting in unnecessary capital expenditure and wasted resources, or under-provisioning,

Authorized licensed use limited to: SUNY AT STONY BROOK. Downloaded on March 04,2026 at 17:50:40 UTC from IEEE Xplore. Restrictions apply.

which causes degradation in service performance and a loss of competitive advantage [15], [16].

We propose an AI-assisted stochastic optimization framework that leverages large language models (LLMs) for various scenario generation. The proposed methodology uses LLMs to formulate comprehensive future scenarios that simulate realistic variations in workload demand, GPU performance decay, new hardwares, and market dynamics. The LLM generated scenarios provide large number of different inputs to our stochastic optimization model, which makes optimal decisions regarding GPU procurement, retirement timing, and necessary cooling infrastructure upgrades under localized power and space capacity constraints, while considering carbon emissions throughout the lifecycle.

The main contributions of this paper can be summarized as:

- We propose an LLM-based data center lifecycle planning framwork that captures complex uncertainties in AI workload demands, hardware depreciation, and evolving power/cooling limitations.
- We formulate a stochastic optimization problem that simultaneously considers the scheduling of upgrades and retirements for the data center facilities, aiming to minimize total lifecycle costs and maximize the space utilization in data centers.
- Through an empirical evaluation with the data collected from the data center of Brookhaven National Lab, we demonstrate that Caplai significantly outperforms conventional heuristic methods across various demand growth scenarios, increases the effective computing capacity within the same power limit by up to 36%, and reduces the data center upgrade cost by up to 32%.

II. Background

A. Data Center Infrastructure Constraints

Limited Power and Space Capacity: The data center facility can be equipped with a fixed amount of power to GPU racks and a limited number of racks, thereby capping the number of high-wattage devices that can operate. Each V100 or A100 GPU uses hundreds of watts [3]–[6]; adding more GPUs or higher-TDP models risks overloading the power budget. Therefore, any increase in computing must come with better FLOPs per Watt efficiency rather than simply more power draw. Newer GPUs usually have greater performance per watt, which gives the potential for GPU upgrade.

Cooling Limitations: The existing racks use air cooling, sufficient for 300–400W GPUs but challenged by the heat generation of newer 700W devices. High-power GPUs like the H100 in SXM form factor (700W TDP) [8], [9], cannot be densely deployed with standard air cooling, as they would overheat or cause performance drop. Upgrading to liquid cooling in the data center could accommodate H100s at full performance. However, this upgrade has a one-time capital cost and operational changes. The trade-off is that liquid cooling would allow higher rack densities and improved energy efficiency to support the next generation of accelerators.

Hardware Depreciation & Residual Value: GPUs are expensive equipment, but they don't retain value for long in the fast-moving AI market. The resale values of older GPUs decay over time, especially as newer generations become available quickly. The current V100s, now two generations old, still have value on secondary markets. However, A100s will likewise lose value after H100s and BL100s are announced [17]. This caused the intuition to retire older GPUs earlier to get funds that can make new purchases. On the other hand, running hardware to its full useful life can maximize the return on investment. Hence, retirement timing is a balancing act between performance needs and residual value recovery.

B. GPU computing efficiency study

To have a better understanding of the computing to power performance of different generations of GPUs, we compare NVIDIA's V100, A100, and H100 GPUs in terms of performance and suitability for large AI workloads [18]–[20]. Table I summarizes key specifications and metrics. The H100's Hopper architecture introduces faster memory (HBM3) and more compute units, delivering significantly higher throughput. And H100 supports 8-bit floating point (FP8) calculations and a Transformer Engine, enabling up to four times higher performance on AI workloads at lower precision than A100. Compared to the V100, the H100 has 8.3 times more computing power and 2.3 times higher rated power, making it approximately 3.6 times more power-efficient.

TABLE I
EVOLUTION OF NVIDIA DATA CENTER GPUS

Model	Year	TDP	Memory Bandwidth	Memory (GB)	FP16 (TFLOPS)
V100	2017	300W	900 GB/s	32	120
A100	2020	400W	2039 GB/s	80	312
H100	2022	700W	3000 GB/s	80	1000

III. Related Work

A. GPU Lifecycle Planning in Data Centers

Current data center lifecycle planning usually uses static provisioning or heuristic-based strategies. It is challenging for those methods to address the dynamic AI workload demand and the uncertainties in hardware performance degradation. These methods could cause low computing resource utilization and higher operational costs [12], [21].

Recent studies have explored more dynamic strategies for managing the GPU lifecycle. For example, some approaches use predictive models to forecast the workload demands in data center [22], [23]. However, these methods often lack integration with well-defined optimization frameworks. A method that combines comprehensive predictions with a robust optimization framework is needed to address the problem.

Authorized licensed use limited to: SUNY AT STONY BROOK. Downloaded on March 04,2026 at 17:50:40 UTC from IEEE Xplore. Restrictions apply.

B. AI-Driven Scenario Generation

LLMs have been widely deployed to enhance scenario generation for case studies and what-if studies [24]. The AI-driven approaches can generate diverse and reasonable future scenarios [25]–[29], to capture potential variations in workload demands, hardware performance, and market dynamics. With LLMs, it's easier for data center operators to anticipate and prepare for different possible futures, and improving the robustness of their planning processes. Besides these advancements, integrating AI-generated scenarios into formal optimization models remains a challenge. Existing approaches only adopt scenario generation but haven't incorporate these scenarios into stochastic optimization frameworks that can guide decision-making under uncertainty [30]. Therefore, we need methodologies that blend AI-driven scenario generation with optimization techniques to enhance lifecycle planning in GPU data centers.

C. Stochastic Optimization for Data Center Management

Stochastic optimization has been widely used for managing uncertainties in data center resource management. By modeling uncertainties in workload demands, hardware performance, and energy costs, we can reduce the operation cost and improve the resource utilization and power efficiency [23], [31]–[34]. Previous work applied stochastic optimization to specific aspects of data center operations, such as energy management and virtual machine placement [35], [36]. However, the work often focuses on single-component optimization rather than a holistic approach to lifecycle planning.

IV. Problem formulation

A. Model Overview

We formulate a stochastic optimization model for data center lifecycle planning over a multi-period horizon $t = 1, \ldots, T$. In our optimization model, we consider multiple server types $i \in I$ for different servers, including hybrid racks that integrate CPU-only nodes and GPU-accelerated nodes. The heterogeneous setting enables the model to capture the power–performance coupling inherent between CPUs and GPUs. Each server type is characterized by distinct compute capacities and power consumption. Our formulation takes carbon emissions into consideration, explicitly capturing the embedded emissions from hardware manufacturing and disposal and the operational emissions generated by energy usage. These emissions are modeled as emission costs [37].

For more realistic modeling, we set a hard limit $p^{\max}$ on total electricity power draw from the grid for both IT equipment and cooling infrastructure. In this framework, we use power oversubscription through statistical multiplexing to ensure that the expected total power consumption remains compliant with the established grid capacity limits. Lastly, the model considers the different in price and efficiency of multiple cooling technologies $c \in C$, such as air cooling, liquid cooling, and immersion cooling. Each cooling strategy has corresponding efficiency factor η_c, one time purchase cost, and operational expenditures (CAPEX/OPEX).

Operating costs in the model are formulated as operational and maintenance expenses related to labor, hardware maintenance, and facility management. the operational cost function considers fluctuations in electricity pricing c_t^{elec} and time-varying emission factors ϕ_t. To address uncertainties in data center operations, our optimization approach considers the uncertainty of workload demands, GPU performance degradation, and hardware resale values.

The target of the optimization model is to minimize total lifecycle cost with the consideration of environmental impact by optimally deciding purchase, operation, and retirement of heterogeneous GPU servers and cooling systems under realistic constraints.

TABLE II
Notation Table

Symbol	Description
Indices and Sets	
$t \in \{1, \ldots, T\}$	Time periods.
$i \in I$	Server types (CPU–GPU combinations).
$c \in C$	Cooling technologies.
$\omega \in \Omega$	Stochastic scenarios.
Parameters	
k_i	Compute capacity per server type i.
p_i^{IT}	IT power per server type i at full utilization.
η_c	Cooling overhead factor for technology c.
$p^{\max}$	Grid power limit.
$d_{t,\omega}$	Workload demand in period t, scenario ω.
$\alpha_{i,\tau}$	Performance degradation factor of server i with age τ.
$c_{i,c}^{\text{cap}}$	Capital cost per server i with cooling technology c.
c_i^{maint}	Maintenance cost per server i.
c_t^{elec}	Electricity price in period t.
ϕ_t	Carbon intensity of electricity in period t.
ζ_i^{emb}	Embedded carbon for manufacturing server i.
ζ_i^{disp}	Disposal emissions for server i.
$v_{i,t,\omega}$	Resale value per server i in period t, scenario ω.
Decision Variables	
$x_{i,c,t}^{\text{new}}$	Number of server i with cooling c purchased at time t.
$x_{i,t}^{\text{retire}}$	Number of server i retired at time t.
$n_{i,c,t}$	Number of active server i with cooling c at time t.
$y_{t,\omega}$	Workload served in period t, scenario ω.

B. Objective Function

Table II describes the notations of the problem formulation. The object is to minimize expected total lifecycle cost:

$$\min_{x,N} \mathbb{E}_\omega \Bigg[\sum_{t,i,c} c_{i,c}^{\text{cap}}\, x_{i,c,t}^{\text{new}} + \sum_t \Big(c_t^{\text{elec}}\, e_{t,\omega} + \sum_{i,c} c_i^{\text{maint}}\, n_{i,c,t} \Big) + \kappa \sum_t \Big(\phi_t\, E_{t,\omega} + \sum_{i,c} \zeta_i^{\text{emb}}\, x_{i,c,t}^{\text{new}} + \sum_i \zeta_i^{\text{disp}}\, x_{i,t}^{\text{retire}} \Big) - \sum_{t,i} v_{i,t,\omega}\, x_{i,t}^{\text{retire}} \Bigg].$$

Authorized licensed use limited to: SUNY AT STONY BROOK. Downloaded on March 04,2026 at 17:50:40 UTC from IEEE Xplore. Restrictions apply.

This objective function captures the cost of data center lifecycle management under uncertainties. The first term represents the CAPEX for acquiring and deploying new servers with specific cooling technologies. The second term accounts for OPEX, combining electricity costs and maintenance costs of active servers. The third component internalizes the carbon cost by monetizing both operational emissions (due to electricity use) and embedded emissions (from manufacturing and disposal of hardware). Finally, the salvage value of retired servers is deducted, reflecting the financial benefit from hardware resale. In our model, we keep both the salvage value and the disposal carbon emissions of retired servers. In practice, some hardware may be resold or recycled, we conservatively assume that a portion of embodied carbon remains attributable to the data center that retires the equipment. This reflects the fact that resale does not eliminate eventual disposal and associated emissions. In this model, the salvage value captures financial recovery, while the disposal carbon term internalizes the environmental cost associated with lifecycle completion. By optimizing this objective, the model balances the trade-offs among investments, operational efficiency, and long-term value recovery across multiple planning scenarios.

C. Constraints

The optimization is subject to constraints ensuring feasibility and service requirements:

Demand Satisfaction. Each period, available computing capacity must meet workload demand (scenario ω):

$$\sum_{i\in I}\sum_{c\in C} \alpha_{i,\tau(i,t)}\; k_i\; n_{i,c,t} \;\geq\; d_t, \quad \forall t,$$

where $\tau(i,t)$ is the age of servers type i at period t. Performance degradation and oversubscription are captured by α. A stricter probabilistic or chance constraint can be applied, but here we ensure deterministic peak-demand satisfaction.

Power Capacity Constraint. The total power draw from IT equipment plus cooling cannot exceed grid capacity $p^{\max}$:

$$\sum_{i,c} \eta_c\, p_i^{\text{IT}}\, u_{i,t}\, n_{i,c,t} \leq p^{\max}, \quad \forall t,$$

where the server utilization fraction $u_{i,t}$ is set at typical (e.g., average or high percentile) values to allow power oversubscription. A high-probability constraint or dynamic power-capping could ensure reliability.

Server Inventory Dynamics. Server inventory updates with procurement and retirement:

$$n_{i,c,t} = n_{i,c,t-1} + x^{\text{new}}_{i,c,t} - x^{\text{retire}}_{i,c,t}, \quad \forall i,c,t.$$

with initial condition $n_{i,c,0}$ provided, and constraints $n_{i,c,t}, x^{\text{new}}_{i,c,t}, x^{\text{retire}}_{i,c,t} \geq 0$.

Cooling Capacity and Selection. Limits on deployment of cooling technologies:

$$\sum_i n_{i,\text{immersion},t} \leq l \cdot r_{\text{rack}},$$

where l is rack limit and r_{rack} servers per rack. Cooling efficiency (η_c) captures the CAPEX/OPEX trade-offs implicitly.

Operational Cost Calculation. Energy cost is calculated as

$$e_{t,\omega} = \left(\sum_{i,c} \eta_c p_i^{\text{IT}} u_{i,t,\omega} n_{i,c,t}\right) \Delta t,$$

with Δt period duration in hours.

Non-negativity and Integrality. Non-negativity and integrality constraints:

$$x^{\text{new}}_{i,c,t}, x^{\text{retire}}_{i,c,t}, n_{i,c,t} \geq 0.$$

Integrality may be relaxed for strategic planning. Logical constraints (no retirements exceeding existing stock):

$$x^{\text{retire}}_{i,t} \leq \sum_{\tau=1}^{t} x^{\text{new}}_{i,c,\tau} - \sum_{\tau=1}^{t-1} x^{\text{retire}}_{i,\tau}.$$

D. Assumptions of Modeling

We make several reasonable modeling assumptions. First, we assume the workload demand d_t is primarily GPU-intensive, with CPU requirements implicitly accounted for within the GPU-based server configurations. Each server type i is characterized by fixed compute capacity k_i and IT power consumption p_i^{IT}. Embedded carbon emissions are modeled with embodied emissions ζ_i^{emb} and disposal emissions ζ_i^{disp}, they account for emissions from manufacturing, deployment, and representing end-of-life impacts. To model realistic operational behavior, we incorporate typical utilization rates u_i, allowing for power oversubscription under the assumption that the risk of simultaneous peak usage can be managed. Cooling overheads are captured using constant efficiency factors η_c, with each server associated with a single cooling technology. Operational costs are simplified through linear per-server maintenance costs c_i^{maint}, which approximate a range of variable expenses such as labor and upkeep. We allow for time-varying electricity prices c_t^{elec} and carbon emission prices ϕ_t. We use a scenario-based stochastic modeling approach to account for uncertainty. Each scenario ω has an corresponding probability p_ω. The aging factor $\alpha_{i,\tau}$ measures the performance degradation over time. Finally, resale prices $v_{i,t}$ are treated as uncertain and incorporated into the decision-making process via their expected values.

V. LLM–Assisted Scenario Generation

Data-center capacity planning is a problem with high uncertainty in future workloads, technology roadmaps, and energy constraints. Classic *scenario planning* (*intuitive logics*, trend extrapolation, Monte-Carlo simulation, etc.) relies on experts to distill a handful of narrative futures [22], [23]. Furthermore, such human-only workflows are **slow** (weeks of facilitation), **costly** (domain specialists, external consultants), and prone to cognitive bias [26]. Recent work demonstrates that large language models (LLMs) can draft coherent, diverse scenario narratives in minutes and at negligible marginal cost [27]–[29]. We extend this insight with a workflow that couples human interpretability with formal stochastic modelling via varying parameter distributions.

Authorized licensed use limited to: SUNY AT STONY BROOK. Downloaded on March 04,2026 at 17:50:40 UTC from IEEE Xplore. Restrictions apply.

A. Human vs. LLM Scenario generation

Traditional scenario workshops rarely produce a large number of scenarios bacause of the limitation from meeting time, experts' bandwidth, and cognitive bias typically cap coverage at three to five cases. A single LLM prompt can deliver dozens of coherent storylines in minutes. Human discussions can add valuable insight, but may also introduce distractions. When human experts do brainstorming, using high-temperature sampling in an LLM can achieve a similar effect with much less effort. What used to take days of discussion can now be accomplished in a two-hour LLM sampling, making scenario planning accessible even with limited resources.

B. AI-Augmented Workflow

Firstly, we select a small set of macro cases, for example, LLM demand is increasing exponentially. For each case, we use a chain-of-thought prompt that first analyzes key drivers, and then we create a timeline for this macro case. Finally, generate a detailed description with key timeline and events information. The domain expert agent reviews the output, corrects factual errors, and extracts the scenarios most relevant to the optimization engine. Next, we translate quantitative features (e.g., "demand grows 20% per quarter") into baseline $\{\mu_t^{(m)}\}_{t=1}^{T}$ wiht parsing templates. We design an explicit noise model for each baseline, such as $d_{t,\omega}^{(m)} \sim \text{LogNormal}(\mu_t^{(m)}, \sigma^2)$.

C. Illustrative Example

Suppose we solicit the following scenario from the LLM:

> *"Over 2025–2027, LLM services expand aggressively, doubling demand yearly. NVIDIA's next-gen GPU arrives in mid-2026, older GPU resale values down by 40%. Electricity costs rise slowly, by 5% per year."*

From this description, we extract baseline: (i) demand grows 20–25% each quarter, (ii) new GPU releases occur in Q2 2026, and (iii) older GPU resale values drop abruptly around Q2–Q3 2026. Then, for each quarter t, we define random variables—e.g., $d_{t,\omega} \sim \mathcal{N}(\mu_t, \sigma^2)$—where μ_t follows the LLM's stated 20–25% growth pattern, and σ^2 is set by our domain knowledge or stress-testing strategy (e.g., a 10% standard deviation around the mean). We repeat this for GPU performance and resale value if needed. With AI-assisted scenario generation, we can easily explore a massive number of what-if studies that cover different future demands and market trends.

VI. Evaluation

We assess the proposed *AI-assisted stochastic planner* from four complementary perspectives:

1) **Data center capacity expansion** with space and power limits (§VI-B).
2) **End-to-end life-cycle cost saving** against a threshold-based heuristic (§VI-C).
3) **Quality of LLM-generated scenarios** and their impact on planning robustness (§VI-D).
4) **Sensitivity to carbon price** to highlight the importance of explicitly pricing emissions (§VI-E).

We evaluate our stochastic optimization approach by simulating a mid-sized data center's GPU lifecycle planning.

AI-Assisted Stochastic Planning: Our method uses the optimization framework described in Section IV with scenarios derived from LLM-generated scenarios and noise.

Heuristic Upgrade Strategy: Upgrades are triggered by simple thresholds (e.g., "Buy new GPUs if current utilization exceeds 90%"), and cooling upgrades occur only when power limits are exceeded for two consecutive quarters.

A. Simulation Setup

Table III summarizes the key input parameters for the simulation. We generate three different demand growth trends (*High*, *Moderate*, *Low*). Within each scenario, we sample Ω_s realizations from Gaussian distributions around that scenario's baseline for workload demand, GPU resale values, and operational cost. The Gaussian distribution's mean μ is to the baseline value, and the variance σ^2 is estimated using each price's historical variance within a short period of time.

TABLE III
Key Simulation Parameters

Parameter	Value
Time Horizon	12 quarters (3 years)
GPU Types	{V100, A100, H100}
Purchase Costs	A100 = \$10k, H100 = \$20k
Power Caps (kW)	Air: 3; Liquid: 12; Immersive: 18
Cooling Upgrade Cost	Liquid = \$5k, Immersive = \$8k
Scenarios ($\mathcal{S}$)	High, Medium, Low Demand Growth
Micro Draws (Ω_s)	20 draws/scenario for random demand
Electricity Rate	\$0.10/kWh
Total power	2MW

B. Capacity Expansion Under Fixed Space & Power Caps

We conduct a stress-test experiment to validate the computing capacity improvement: Firstly, fix the rack count and 2 MW facility power limit from Table III. Each scenario's baseline demand trace d_t^{base} is multiplied by a scalar load factor $\alpha \in [\,1, 2.0\,]$.[1] Then, for a given α, we solve the optimization model (§IV) and check feasibility of the resulting deployment in every quarter: rack-count $\leq X$ *and* total power draw ≤ 2 MW. Lastly, using binary search on α, we record the *maximum feasible load factor* $\alpha^\star$ for both the heuristic baseline and our AI-assisted planner.

a) Results.: Table IV reports the peak quarterly kilo GPU-hours that can be satisfied (Demand $= \alpha^\star \sum_t d_t^{\text{base}}$) alongside the percentage headroom gained.

b) Takeaways.: The optimized plan unlocks **32–36%** more compute capacity within the same physical envelope by: (i) *early migration* to higher–FLOP/W GPUs, upgrade V100 and A100 to H100, (ii) *rack-level liquid cooling* that raises the per-rack power budget from 3 kW to 12 kW, and (iii) *retiring* legacy V100 nodes whose poor perf/W otherwise spend a lot on the power cap. This highlights that the planner improves the data center capacity based on the power and space limits.

[1] Values $\alpha > 1$ represent additional clusters or external tenant jobs that would otherwise be off-loaded to another site or cloud.

Authorized licensed use limited to: SUNY AT STONY BROOK. Downloaded on March 04,2026 at 17:50:40 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
PEAK DEMAND SERVED UNDER THE EXISTING 2 MW AND RACK-COUNT CAPS.

Scenario	Heuristic	Optimized	Improve
Fast-Growth	420	572	+36%
Linear-Growth	310	419	+35%
Slow-Growth	150	198	+32%

C. End-to-End Cost Optimization

We measure *total life-cycle cost*: purchases + cooling upgrade + electricity expense − salvage. Figure 1 compares the proposed method and the baseline under three distinct demand growth scenarios. From our simulations, the cost saving comes from retiring the old GPUs with high residual value and the accurate prediction of the demand.

In the *Fast-Growth* scenario, our method achieves a total cost of $283M compared to $397M for the manual approach, yielding a savings of about 28.7%. Under *Linear-Growth* demand, the cost difference is $221M vs. $316M with 30% savings. For *Slow-Growth* demand, the costs are $132k vs. $196k, saving up to 32%.

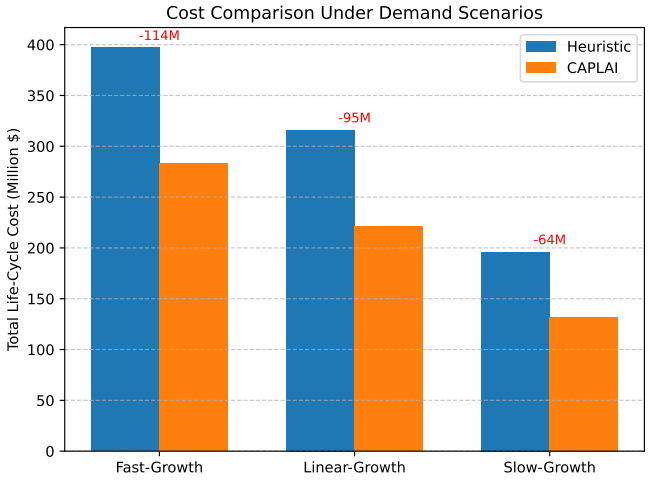

Fig. 1. Illustrative total life-cycle cost under CAPLAI and the manual heuristic.

D. Evaluating LLM-Generated Scenarios

1) End to end Robustness: We benchmark the LLM pipeline against two baselines: (1) **Historical-Fit (HF):** scenarios obtained by ARIMA forecasting of the past 12 quarters. (2) **Expert-Workshop (EW):** three hand-crafted narratives produced by a domain expert panel.

a) Diversity & Coverage.: We compute pairwise Jensen–Shannon divergence of normalized demand trajectories. As shown in figure 2, the LLM set achieves a mean JSD of 0.34, versus 0.21 (HF) and 0.26 (EW), indicating broader coverage of potential futures.

b) Realism.: To test internal coherence, we sample 1 000 synthetic quarterly time-series from each scenario family and compare them against an held-out 2024–2025 trace from a public LLM inference service [38]. The Kolmogorov–Smirnov distance between empirical CDFs of quarterly growth factors is smallest for LLM (0.12) relative to HF (0.19) and EW (0.16).

c) Planning Robustness.: Running our optimizer with HF and EW scenarios increases mean life-cycle cost by 6.3 % and 3.8 %, respectively, compared with LLM scenarios, because unanticipated demand spikes in 3Q-4Q 2026 trigger emergency GPU purchases at higher spot prices.

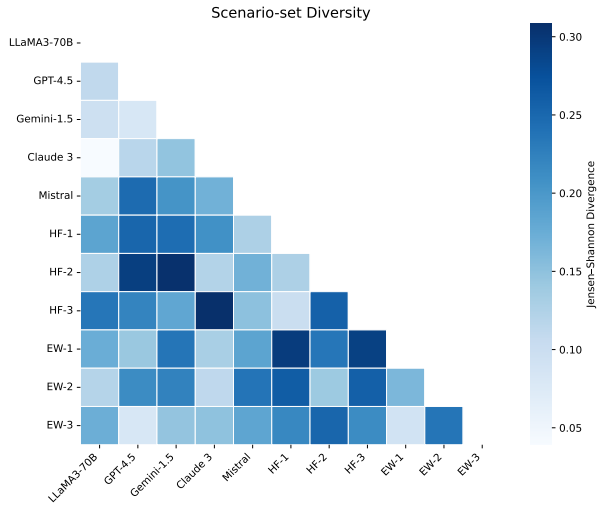

Fig. 2. Scenario-set diversity measured by pairwise Jensen–Shannon divergence (lower-triangle heatmap); darker means more dissimilar.

2) Historical Replay Back-Test (2016–2024): To validate that our Scenario-Generation Agent (SGA) improves *forecast uncertainty*—and not merely mean-squared error—we replay eight years of production traces (2016–2024) under a strict walk-forward protocol.

a) Metric.: Let $h \in \{1, \dots, H\}$ index the H monthly evaluation cut-offs. For each h we generate a scenario deck $\{\tilde{d}_t^{(k)}\}_{k=1}^{K}$ covering the 12-month forecast window $[h, h+T)$, and compute its variance $\sigma^2_{\text{forecast},h}$; the realised trace over the same window has variance $\sigma^2_{\text{real},h}$. We have *variance-error*

$$\text{VarErr} \;=\; \frac{1}{H}\sum_{h=1}^{H}\left(\sigma^2_{\text{forecast},h} - \sigma^2_{\text{real},h}\right)^2. \qquad (1)$$

b) Baselines.: Table V lists the naïve models we benchmark against. All share the same input window—*the last 12 months visible at* h—and require **no** hyper-parameter tuning beyond what is shown.

TABLE V
DEMAND-FORECAST BASELINES USED IN THE REPLAY EXPERIMENT.

Baseline	Forecast rule	Hyper-parameters
Linear trend	OLS on last 12mo	none
Exponential	$d_{t+1} = d_t \exp(\beta)$	$\beta =$ CAGR(12mo)
Seasonal naïve	Repeat from $t-12$	n/a

Authorized licensed use limited to: SUNY AT STONY BROOK. Downloaded on March 04,2026 at 17:50:40 UTC from IEEE Xplore. Restrictions apply.

c) Walk-forward protocol.: For each monthly horizon h from January 2016 through December 2024, we first freeze all information available up to month h. The Scenario-Generation Agent (SGA) then draws $K = 200$ stochastic trajectories with the LLM, after which every baseline in Table V is refit on the same historical window. Finally, we assess forecast quality by computing the variance-error metric in Eq. (1) over the next $T = 12$ months.

d) Results.: As shown in Figure 3, across $H = 108$ rolling horizons, SGA reduces the variance-error metric by $42\% \pm 7$ pp on average relative to the strongest naïve baseline. The single largest improvement 64% occurs in March 2020, when pandemic-induced demand spikes invalidate all simple trend assumptions.

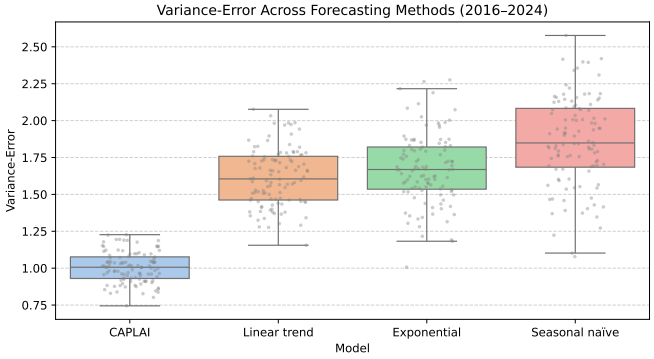

Fig. 3. Percentage reduction in variance-error per horizon (box = IQR, whiskers = 5–95th pct.).

E. Carbon-Price Sensitivity

Current Carbon price is around \$30\$/tCO$_2$e [37]. To evaluate the decision making with different importance level to carbon emission, we vary the social cost of carbon ($c^{\mathrm{CO_2}}$) from \$30 to \$2000 \$/tCO$_2$e, applying it to both electricity and the embodied emissions of new GPUs. Figure 4 illustrates the optimal cost improvement for different carbon prices. At \$50 \$/t, the planner already shifts 18% of quarterly demand from V100s to A100s thanks to their higher FLOPs/W. When the price exceeds \$100 \$/t, liquid cooling becomes cost-effective a full year earlier—from Q2 2026 to Q2 2025—because the improved power-usage effectiveness (PUE) reduces cumulative emissions. Despite the increase from \$0 to \$200 \$/t, the total discounted cost rises by only 6%, indicating that proactive hardware choices absorb most carbon penalties.

VII. Discussion

CAPLAI is a framework that guides the data center lifecycle planning. However, there are some limitations of the current system. First of all, the data center upgrade and retirement plans could not be directly evaluated before we select and implement the lifecycle decision. Therefore, we chunk the long-term (three-year) planning into shorter time periods (a quarter) plans. This strategy gives us the space to fine-tune the data center lifecycle planning with new incoming

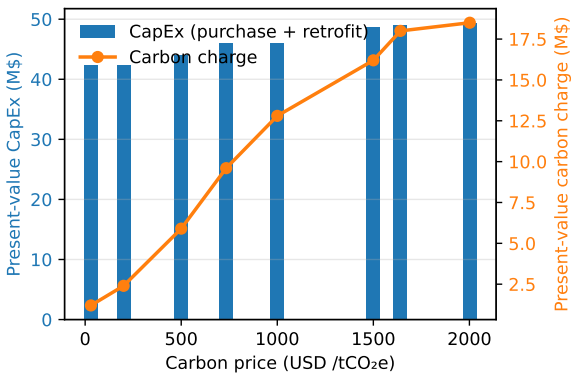

Fig. 4. Impact of carbon price on optimal upgrade timing (bars: purchase + retrofit capex, lines: carbon charge).

information. The second limitation is the challenge to reduce the hallucinations of LLMs in CAPLAI. In our evaluation, we observed that the macro cases' quality is more easily affected by hallucinations compared with the micro cases. This is mainly because the micro cases are generated under the guidance of mathematical distributions, but the macro cases are mainly generated from the few-shot learning of LLMs. Although we have the fact validation agent, the macro case generator is still the bottleneck of CAPLAI's performance. One potential method to improve the macro case generation is to keep optimizing the macro case generation's guidance [24]. The other method is to improve the capability of the fact validation agent.

A promising future direction of CAPLAI is extending the framework into a closed-loop, multi-agent system for end-to-end infrastructure management. Scenario generation agent uses LLMs to translate high-level business factors—such as budget limits, risk level, and SLOs into structured planning constraints. We can use real-time monitoring data from job traces and the cooling system to continuously fine-tune scenario realism and enable dynamic updates to workload models. Finally, the real-time manager agent uses a digital twin of the facility, adjusts workloads and infrastructure operations, and feeds telemetry back into the planning loop.

VIII. Conclusion

We propose an AI-assisted stochastic optimization framework, CAPLAI, that balances GPU data center upgrades and retirements under uncertainty. Using LLM-generated macro and micro scenarios, CAPLAI captures demand and hardware trends and proposes the optimal data center lifecycle strategies. Evaluation shows CAPLAI outperforms threshold-based strategies, increasing compute capacity by up to 36% and reducing lifecycle costs by up to 32%. Results highlight the promise of combining LLMs with optimization for robust data center infrastructure planning.

Acknowledgments

We thank the anonymous reviewers for their insightful comments. This work is supported by the US National Sci-

Authorized licensed use limited to: SUNY AT STONY BROOK. Downloaded on March 04,2026 at 17:50:40 UTC from IEEE Xplore. Restrictions apply.

ence Foundation under grant numbers CNS-2106027, CNS-2214980, CNS-2146909, and CCF-2046444.

REFERENCES

[1] "The state of ai: How organizations are rewiring to capture value," *McKinsey Company*, Mar. 2025, accessed: 2025-06-01. [Online]. Available: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

[2] V.-G. Anghel, "Liquid cooling in the generative ai era," *Data Center Dynamics*, January 2025. [Online]. Available: https://www.datacenterdynamics.com/en/analysis/liquid-cooling-in-the-generative-ai-era/

[3] C. Tuttle, "When to move from air cooling to liquid cooling for your data center," *Dell Technologies Blog*, November 2024. [Online]. Available: https://www.dell.com/en-hk/blog/when-to-move-from-air-cooling-to-liquid-cooling-for-your-data-center

[4] NVIDIA Corporation, "Data center best practices with dgx b200," NVIDIA, Tech. Rep., 2024. [Online]. Available: https://docs.nvidia.com/nvidia-dgx-superpod-data-center-best-practices-with-dgx-b200.pdf

[5] C.-I. Hsu and L. Zhou, "Rack pdu for green data centers," in *Data Center Handbook*. Wiley, 2021, ch. 16, pp. 263–290.

[6] Dgtl Infra. (2023) Data center racks, cabinets, and cages: An in-depth guide. [Online]. Available: https://dgtlinfra.com/data-center-racks-cabinets-cages/

[7] Q. Zhang, Z. Meng, X. Hong, Y. Zhan, J. Liu, J. Dong, T. Bai, J. Niu, and M. J. Deen, "A survey on data center cooling systems: Technology, power consumption modeling and control strategy optimization," *Journal of Systems Architecture*, vol. 119, p. 102253, 2021. [Online]. Available: https://doi.org/10.1016/j.sysarc.2021.102253

[8] A. H. Khalaj and S. K. Halgamuge, "A review on efficient thermal management of air- and liquid-cooled data centers: From chip to the cooling system," *Applied Energy*, vol. 205, pp. 1165–1188, 2017. [Online]. Available: https://doi.org/10.1016/j.apenergy.2017.08.037

[9] S. Xu, H. Zhang, and Z. Wang, "Thermal management and energy consumption in air, liquid, and free cooling systems for data centers: A review," *Energies*, vol. 16, no. 3, p. 1279, 2023. [Online]. Available: https://www.mdpi.com/1996-1073/16/3/1279

[10] Wikipedia contributors, "Computer cooling," https://en.wikipedia.org/wiki/Computer_cooling, 2025, accessed: 2025-06-01.

[11] "Ask an engineer: What is node lifecycle management and why does it matter for ml training and inference?" https://www.coreweave.com/blog/what-is-node-lifecycle-management-ml-training-and-inference, August 2024, accessed: 2025-06-01.

[12] X. Chen, H. Xu, and C. Xu, "Optimizing dynamic data center provisioning through speed scaling: A primal-dual perspective," in *Proceedings of the 36th ACM Symposium on Parallelism in Algorithms and Architectures*, ser. SPAA '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 89–100. [Online]. Available: https://doi.org/10.1145/3626183.3659956

[13] D. Crawford, J. Wang, and R. Singh, "Ai's trillion-dollar opportunity," *Bain Company*, September 2024. [Online]. Available: https://www.bain.com/insights/ais-trillion-dollar-opportunity-tech-report-2024/

[14] T. Gonçalves, A. Beck, and A. Lorenzon, "Balancing performance and aging in cloud environments," in *Proceedings of the 14th International Conference on Cloud Computing and Services Science - Volume 1: CLOSER*, INSTICC. SciTePress, 2024, pp. 216–223.

[15] X. Fan, W.-D. Weber, and L. A. Barroso, "Power provisioning for a warehouse-sized computer," *SIGARCH Comput. Archit. News*, vol. 35, no. 2, p. 13–23, Jun. 2007. [Online]. Available: https://doi.org/10.1145/1273440.1250665

[16] Y. Cheng, Z. Chai, and A. Anwar, "Characterizing co-located datacenter workloads: An alibaba case study," 2018. [Online]. Available: https://arxiv.org/abs/1808.02919

[17] Lepton AI, "The missing guide to the h100 gpu market," September 2024, accessed: 2025-06-01. [Online]. Available: https://blog.lepton.ai/the-missing-guide-to-the-h100-gpu-market-91ebfed34516

[18] "Nvidia tesla v100 gpu accelerator," NVIDIA Corporation, Tech. Rep., 2017, accessed: 2025-06-01. [Online]. Available: https://images.nvidia.com/content/technologies/volta/pdf/tesla-volta-v100-datasheet-letter-fnl-web.pdf

[19] "Nvidia a100 tensor core gpu," NVIDIA Corporation, Tech. Rep., 2020, accessed: 2025-06-01. [Online]. Available: https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-us-nvidia-1758950-r4-web.pdf

[20] "Nvidia h100 tensor core gpu," NVIDIA Corporation, Tech. Rep., 2022, accessed: 2025-06-01. [Online]. Available: https://www.nvidia.com/content/dam/en-zz/Solutions/gtcs22/data-center/h100/PB-11133-001_v01.pdf

[21] J. Wang, B. Palanisamy, and J. Xu, "Sustainability-aware resource provisioning in data centers," in *2020 IEEE 6th International Conference on Collaboration and Internet Computing (CIC)*, 2020, pp. 60–69.

[22] S. Rallabandi, "Ai-driven capacity planning in large-scale infrastructure: A comparative analysis of lstm networks and traditional forecasting methods," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 10, no. 6, pp. 950–962, November 2024, accessed: 2025-06-01. [Online]. Available: https://ijsrcseit.com/index.php/home/article/view/CSEIT241061139

[23] J. Comden, S. Yao, N. Chen, H. Xing, and Z. Liu, "Online optimization in cloud resource provisioning: Predictions, regrets, and algorithms," *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 3, no. 1, Mar. 2019. [Online]. Available: https://doi.org/10.1145/3322205.3311087

[24] E. Xing, M. Deng, J. Hou, and Z. Hu, "Critiques of world models," 2025. [Online]. Available: https://arxiv.org/abs/2507.05169

[25] Y. Liu, H. Du, D. Niyato, J. Kang, Z. Xiong, Y. Wen, and D. I. Kim, "Generative ai in data center networking: Fundamentals, perspectives, and case study," 2024. [Online]. Available: https://arxiv.org/abs/2409.09343

[26] Y. Mei, T. Nie, J. Sun, and Y. Tian, "Llm-attacker: Enhancing closed-loop adversarial scenario generation for autonomous driving with large language models," 2025. [Online]. Available: https://arxiv.org/abs/2501.15850

[27] H. Jung, W. Seo, S. Song, and S. Na, "Toward value scenario generation through large language models," in *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, ser. CSCW '23 Companion. New York, NY, USA: Association for Computing Machinery, 2023, p. 212–220. [Online]. Available: https://doi.org/10.1145/3584931.3606960

[28] C. Chang, S. Wang, J. Zhang, J. Ge, and L. Li, "Llmscenario: Large language model driven scenario generation," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2024.

[29] X. Li, E. Liu, T. Shen, J. Huang, and F.-Y. Wang, "Chatgpt-based scenario engineer: A new framework on scenario generation for trajectory prediction," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 3, pp. 4422–4431, 2024.

[30] D. Islip, R. H. Kwon, S. Bae, and W. C. Kim, "Contextual scenario generation for two-stage stochastic programming," 2025. [Online]. Available: https://arxiv.org/abs/2502.05349

[31] Z. Liu, Y. Chen, C. Bash, A. Wierman, D. Gmach, Z. Wang, M. Marwah, and C. Hyser, "Renewable and cooling aware workload management for sustainable data centers," *SIGMETRICS Perform. Eval. Rev.*, vol. 40, no. 1, p. 175–186, Jun. 2012. [Online]. Available: https://doi.org/10.1145/2318857.2254779

[32] T. N. Le, Z. Liu, Y. Chen, and C. Bash, "Joint capacity planning and operational management for sustainable data centers and demand response," in *Proceedings of the Seventh International Conference on Future Energy Systems*, ser. e-Energy '16. New York, NY, USA: Association for Computing Machinery, 2016. [Online]. Available: https://doi.org/10.1145/2934328.2934344

[33] T. Chen, X. Wang, and G. B. Giannakis, "Cooling-aware energy and workload management in data centers via stochastic optimization," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 2, pp. 402–415, 2016.

[34] S. Baxi, K. Cummings, A. Jacquillat, S. Lo, R. McDonald, K. Mellou, I. Menache, and M. Molinaro, "Online rack placement in large-scale data centers," 2025. [Online]. Available: https://arxiv.org/abs/2501.12725

[35] L. Yu, T. Jiang, and Y. Zou, "Distributed real-time energy management in data center microgrids," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, pp. 3748–3762, 2016.

[36] H. Jin, D. Pan, J. Xu, and N. Pissinou, "Efficient vm placement with multiple deterministic and stochastic resources in data centers," in *2012 IEEE Global Communications Conference (GLOBECOM)*, 2012, pp. 2505–2510.

[37] KraneShares, "Kraneshares global carbon strategy etf (krbn)," 2025, accessed: 2025-06-01. [Online]. Available: https://kraneshares.com/krbn/

[38] S. Singh, "Chatgpt statistics (2025) – daily & monthly active users," https://www.demandsage.com/chatgpt-statistics/, August 14 2025, demandSage; accessed [date you accessed it].

Authorized licensed use limited to: SUNY AT STONY BROOK. Downloaded on March 04,2026 at 17:50:40 UTC from IEEE Xplore. Restrictions apply.