

Sustainable Grid through Distributed Data Centers

Spinning AI Demand for Grid Stabilization and Optimization

Scott C Evans¹, Nathan Dahlin³, Ibrahima Ndiaye², Sachini Piyoni Ekanayake¹,
Alexander Duncan¹, Blake Rose¹, Hao Huang¹

GE Vernova Advanced Research Center: ¹AI-Machine Learning, Robotics Lab; ²Electrification Mission

³ECE Department, University at Albany SUNY

¹evans@ge.com, ³ndahlin@albany.edu, ²Ndiaye@ge.com, ¹Sachini.Ekanayake@ge.com,

¹Alexander.K.Duncan@ge.com, ¹Blake.Rose@ge.com, ¹Hao.Huang@ge.com

Abstract—We propose a disruptive paradigm to actively place and schedule TWhrs of parallel AI jobs strategically on the grid, at distributed, grid-aware high performance compute data centers (HPC) capable of using their massive power and energy load to stabilize the grid while reducing grid build-out requirements, maximizing use of renewable energy, and reducing Green House Gas (GHG) emissions. Our approach will enable the creation of new, value adding markets for spinning compute demand, providing market based incentives that will drive the joint optimization of energy and learning.

Index Terms—grid stability, entropy economy, renewable energy, distributed artificial intelligence, high-performance computing, monte-carlo simulations.

I. INTRODUCTION

Data center HPC energy demand totaled 200TWh (4% of US electricity) in 2022 and is expected to grow to 260TWh (6%) by 2026 [1] and to 9.1% of US electricity consumption by 2030 [2]. This growth trends in concentrated areas, creating uneven geographic distribution and forcing significant expansion where green generation resources may not be abundant, affordable, or optimally used. Buildout to energize this +2x demand growth under current paradigms is projected to require US grid expansion by +1x at insurmountable costs (at least +4x) [3]. In this paper, we describe a paradigm of distributed HPC that leverages the near zero cost of moving information over the internet, enabling the seamless routing of energy intensive, massively parallelizable HPC Monte-Carlo (HPCMC) and other high value AI jobs to available green energy capacity. This process will stabilize the grid, cut grid build-out requirements by half, and maximize the use of renewable, GHG-free energy.

Our approach provides grid-centric, disruptive methods to place and schedule critical portions of HPC loads, allowing lower peak demand requirements for grid capacity and reliability while minimizing GHG emissions. Our vision of a sustainable grid through “spinning demand” served to distributed data centers is shown in Fig. 1, where a pooling agent distributes a backlog of HPCMC jobs throughout the grid to level load, maximize renewables, and stabilize the grid.

© 2025 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

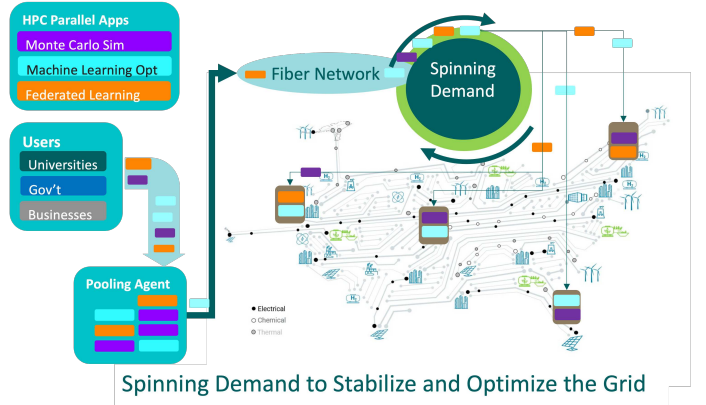


Fig. 1. Vision for Sustainable Grid through distributed data centers: spinning demand of batch-able compute jobs distributed to stabilize the grid.

Our paper is structured as follows: After reviewing previous work, we describe our approach to providing grid-centric disruptive methods to prescribe and schedule a critical portion of HPC loads. We then present simulation results that illustrate the efficacy of our paradigm, followed by analysis of market operation and pricing, and cost-benefit analysis.

The key contributions of this work include: (i) A novel grid-centric paradigm focused on the distribution of batchable compute loads. (ii) An algorithm and accompanying simulations demonstrating benefits including reductions in renewable curtailment and required spinning reserve, with particular attention to the impact of HPC and renewable co-location. (iii) Assessment of the sensitivity of economic benefits of our approach to factors including end user flexibility, degree of co-location of HPC and renewable resources, and renewable supplier market participation rules.

II. PRIOR WORK

Our vision builds upon the state of art, such as “Zero Carbon Grid” [4] paradigms and related technologies, by jointly optimizing the electric and HPC loads, making possible solutions previously not considered. Today, grid stabilization is achieved via costly high-capacity capacitors, synchronous power condensers, and the overprovisioning of spinning reserve and the grid capacity. We leverage theoretical concepts put

forward in GE Vernova’s “Green AI” and “Entropy Economy” initiatives [5] that show promise for the joint optimization of computational learning and energy, and harvest these concepts with a grid centric view. Specifically, [6] describes how to optimally schedule energy intensive ML tasks in a distributed way by allocating load to where the free energy exists, such that the priority of the task completion is also achieved.

In recent years the adoption of deep learning models has grown rapidly due to their superior performance in areas like natural language processing (e.g., large language models (LLMs)). At the same time, the complexity of these models has grown at a comparable pace, such that both training and inference require higher energy consumptions [7]–[9]. In this setting, [10] proposed a streaming neural network that uses incremental weighted Principal Component Analysis (PCA), facilitating swift and precise classification tasks given restricted memory and energy resources while ensuring high processing speed. Further, to amplify learning while jointly minimizing energy and carbon costs, [11] introduces a concept of “Additive AI”, designed to harmonize with the Kolmogorov Complexity and Kolmogorov Structure Function.

At the same time, tracking CO2 emission from various machine learning models has emerged [12]. For instance, [13] presents an accurate system-wide power analysis and carbon emission tracking method called “PACT” which measures the total power consumed by hardware resources when running benchmark machine learning tasks. To this end, [14] discusses the advantage of finite-time parallel computing over serial computing using insights from non-equilibrium thermodynamics. In parallel to the advancement of LLMs, [15], [16] present techniques to measure and reduce energy consumption for training and inference for language models including avenues like reducing carbon impact. Further, [17] explores the impact of energy efficiency of inference in deep learning runtime infrastructure by comparing different execution providers using a single model i.e., ResNet. Since many LLMs are trained once and used many times later, [18] analyses the inference costs in the applications of computer vision and natural language processing in the aspect of AI and energy consumption.

On the other hand, it is important to focus on how this energy increment affects the power grid. To this end, [19] emphasizes the impact on grid reliability due to the addition of load from tied to the exponential growth in machine learning models, while [20] emphasizes the impact of AI on power demand and need for power distribution and management (e.g., energy storage systems and smart grid technologies) to support the growing AI industry. Regarding state-of-the art methods deployed today, [21] describes development of battery energy storage systems to support variability introduce in the grid due to renewables and [22] describes the synchronous condenser usage for weak grids. Still, these existing grid stability methods do not consider explicitly leveraging the emerging data center energy demand [1].

III. PROPOSED PARADIGM

A. Parallelization of Computational jobs

Traditionally, HPC vendors have required uninterruptible power and energy supply to support their ever-growing energy and computational needs. While some HPC functions cannot be easily interrupted, and thus require consistent, even backup power at the ready, not all computational jobs require this highest level of Quality of Service (QoS) for time and energy, and many can be paused, delayed, or even discarded without undermining the overall computational goals. To this end, we exploit massively parallelizable Monte-Carlo (MC) and other AI jobs comprising up to 10% of HPC load. These jobs will provide sufficient initial “Spinning Demand” throughout the grid, creating breakthrough opportunities for grid planning, stabilization, efficiency, GHG reduction, and renewable adoption. Inclusion of other types of HPC, and even integration with other computational loads are discussed below.

1) *Progression of Compute Loads:* We envision the inclusion of HPC computation jobs will move from simple to more complex, starting with HPCMC jobs as discussed above. These HPCMC jobs, pervasive in epidemiology, weather, scientific computing, and machine learning model optimization and validation, share properties of: a) being parallelizable into independent, simultaneously executed jobs, b) lacking strong QoS requirements (on scale of minutes, hours, or even days), and c) lacking strong privacy or security requirements. As the Sustainable Grid Marketplace grows, inclusion of more complicated, yet parallelization computational loads are envisioned. Federated Learning, and even distributed large language model learning, are examples of AI computational loads that benefit from parallelization, but have higher QoS and inter-process communication requirements that we will consider as the sustainable grid progresses.

2) *Integration of Compute Loads: Learn While Mining:* A computational load that is even less complex than HPCMC is proof of work crypto mining. These loads are ubiquitous, and already being used for load leveling activities, but have been resisted due to their high energy cost. One way the carbon footprint of proof of work crypto mining can be improved is by leveraging it for scientific machine learning as discussed in [23]. Monte Carlo and Proof of Work computational loads share in common the need for massive amounts of pseudo-random numbers in order to function. One can think of proof of work crypto mining as an extremely aggressive form of rejection sampling. The inefficiency of proof of work crypto lies in the fact that the only value add done from the terahashes of random numbers that are generated is the rare authentication of a transaction. In Fig. 2 we show the energy benefit of harvesting the random numbers generated from proof of work mining to simultaneous MC simulations. Here we see a 10% energy gain from simultaneous “Learn While Mining” over separate mining and MC simulations. Thus jointly conducting these and other types of computational jobs can lead to energy benefits.

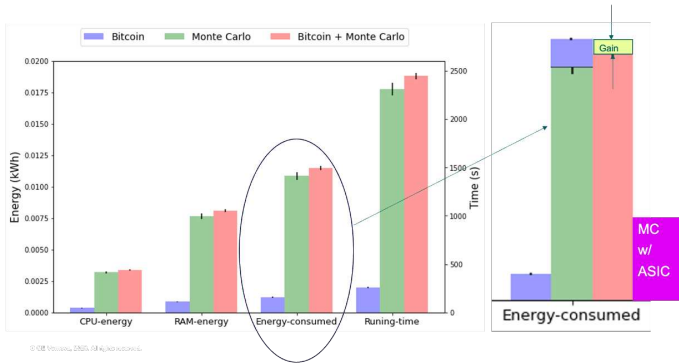


Fig. 2. Energy benefit of joint Monte Carlo and proof of work mining

B. Coordinated Load Placement

Rather than “Demand Response,” which reacts to loads and responds to some objective, our approach will “Actively Place,” and serve at its discretion HPCMC load, leveraging the near zero cost of moving data and compute jobs over the internet to reduce or remove infrastructure requirements. This paradigm leverages the fact that moving energy is very expensive, at \$41.5/MWhr [24] for 1000 miles, whereas moving information is essentially free [25] depending on the platform. Thus, the most affordable way to optimize the intermodal energy superhighway is to transport demand using the information grid (the internet), and leverage a portion of the growing demand from HPC to optimize and stabilize the grid itself. Specific metrics measuring progress in economic and security priorities for the US including the provision of energy technologies that reduce GHG emissions, improving efficiency of virtually all economic sectors as the AI revolution continues, and increasing the resilience and reliability of energy infrastructure are as described in Table I.

TABLE I
EXPECTED BENEFITS

Metric	State-of-the-art	Target
Stability	Freq Regulation	30%+ Cost Reduction
Grid Expansion	+1x	+0.5x
Renewables curtailed	2-6%	< 0.5% (10x reduction)
HPC GHG Emissions	.86lb/kWh	.75lb/kWh

C. HPC and Grid Stability Support

The most fundamental operation in maintaining grid reliability is balancing supply and demand. Indeed, at any given time the power demand from the load is required to be satisfied by the generation otherwise the grid frequency deviates from its synchronous value (e.g., 60Hz in North America, 50Hz in Europe) which can ultimately lead to grid collapse if not corrected rapidly. In the era of high penetration of HPC and variable renewable energy (VRE) resources such as wind and solar, the grid stability can be significantly challenged by the high volatility of both load and generation. The high dynamics of HPC can be turned into an opportunity to support the

grid by providing ancillary services that will increase grid stability while reducing the size and cost of ancillary equipment otherwise required. The grid regulates frequency using two key mechanisms: regulating reserve, including autonomous dispatch or the so-called automatic generation control (AGC), and contingency reserve, including both spinning and non-spinning reserve [26]. Regulation and contingency reserves differ essentially by their time scale requirements and the opportunity and magnitude of their intervention. In the US, regulation reserve (also known as primary frequency control) is a continuous operation (non-event) that deploys a dispatch command to generators every 4-6 secs, while contingency reserve is an event-based operation. Spinning reserve (already connected resources) and non-spinning reserve (ready to connect upon request) are required to supply power usually within 10 minutes to help the frequency return to nominal value upon a significant event (e.g. generation trip). Demand response is another form of contingency reserve. Another key difference between regulation and contingency reserve is the directionality of the power contribution. While regulation reserve can increase or decrease the power based on the received control signal (up and down regulation, respectively), contingency reserve is expected to only inject power to restore a collapsing frequency. This sets the boundary of where HPC can contribute to support the grid. HPC provides the ability to stack compute load and build an instantaneous power demand to help stabilize grid frequency. If this computational power can be modified in real-time to go up and down and follow an automatic dispatch signal such as AGC, they can become a very powerful asset to provide frequency regulation service to the grid. However, regulation reserve service has strict rules to be complied to for eligibility. The most critical include the capability range for ramp up and ramp down which may or may not be symmetrical, ability to follow AGC setpoints adjusted every four to six seconds, minimum duration capability (i.e., minimum duration time for same response direction) which is typically 15 minutes. With future hyperscale data center size ranging from tens of MWs to gigawatts [27], HPC have the advantage of size (up to several hundreds MW), modularity (adjustable up and down range from same HPC), and high flexibility. Table II summarizes key differences between the regulation reserve and secondary contingency reserve and the eligibility of HPC for each.

The ability of HPC as an asset for regulation reserve is clear. However, its eligibility to participate in contingency reserve such as the secondary frequency regulation is less straightforward in current electricity market practices. Indeed, the contingency reserve requires eligible resources to take rapid action and inject power to the grid upon request to restore the grid frequency to nominal value. Unlike battery energy storage, HPC is a purely “load” resource i.e. it can only absorb power from the grid; therefore, power injection is not an option. However, if the consequence of injecting power under contingency is to reduce the imbalance within an area control and avoid underfrequency load shedding (UFLS), it is conceivable that a rapid ramp down of large flexible

TABLE II
KEY GRID RELIABILITY SERVICES AND ELIGIBILITY OF HPC

Service	Regulating Reserve	Primary Contingency Reserve	Secondary Contingency Reserve
Main function	Continuous Frequency regulation	Demand Response	Spinning and Non-spinning reserve
Trigger	Autonomous dispatch (AGC)	Real-time dispatch or Major Event	Major event (e.g loss of generation)
Response type	Upward and Downward	Downward	Upward
Response time	Follow setpoint every 4 to 6 secs	Within 10min	Within 10min
Capability basis	5min Ramp Up/Down	Undefined	10-min Ramp
Minimum capacity duration	15min (typical)	Minutes to several hours	30-60min
HPC qualification	Very Likely	Very Likely	Possible

load such as HPC can presumably provide the same result. If not readily eligible for “spinning” reserve, HPC can at least be qualified for demand response. Demand response is often considered as a form of secondary contingency but due to eligibility requirements such as its autonomous characteristic, required response time, or directionality, it is effectively a primary frequency response mechanism that can suit very well the capabilities of HPC. Although the participation of HPC in demand response will not impact the system inertia as with traditional spinning and non-spinning resources (e.g. gas or hydro plants), its potential fast response can improve the ROCOF (rate-of-change-of-frequency) and ultimately the frequency nadir. The biggest challenge for HPC in providing ancillary services and offsetting grid investment to lower the cost of regulation and contingency reserves is understanding the market rules. Indeed, in many parts of the world, the grid is segmented into regions or balancing authorities administrating the energy markets and the frequency regulation mechanisms. In the US for example, those will be the ISOs (Independent System Operators). Each ISO, due to their generation mix, grid inertia, marginal energy price, etc. will have specific rules affecting the eligibility of HPC to provide ancillary services for grid stability. If forecasting computational load across a region (e.g across US or Europe) allows HPC participants to share the load, the load stacking towards building a profile modifiable in real-time to respond to an AGC signal or a contingency event will be done at the transmission system operator (TSO) level. Fig. 3 shows a notional segmentation of distributed HPC load that can support participation in frequency markets in different TSO. Each TSO as highlighted in Fig. 3 will have different market rules and eligibility [28]. Eligibility requirements and rewards will not be the same per region. In Fig. 3, HPC n is an aggregation of regional HPC participants. An additional HPC optimal dispatch within the ISO, constrained or incentivized by ramping capability requirements and other economic and security criteria will need to be developed to appropriately schedule the computational demand for real-time operation.

Despite the complexity of grid ancillary service mechanisms, the trend for modernizing grids is a larger utilization of ancillary services due to greater volatility of both the generation and load. In this context, distributed HPC load is positioned as an important resource, as it is potentially more powerful, cost-effective, and carbon neutral than battery energy storage, aggregated electric vehicles, or hydrogen electrolyzer plants.

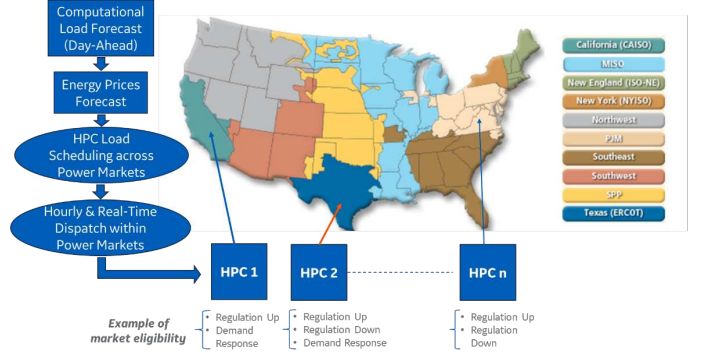


Fig. 3. Potential utilization of HPC in US electric power markets

IV. HPC LOAD PLACEMENT & SCHEDULING ALGORITHM AND SIMULATION RESULTS

We demonstrate the intuition of these concepts via simulation and analysis of a 3-node system. As pictured in the upper portion of Fig. 4, the system includes a solar farm, a wind farm, and a gas turbine plant located at node 1, node 2, and node 3, respectively. We consider two HPC configuration cases: (a) 100 MW HPC concentrated at node 3 alongside the gas turbine (b) 100 MW HPC at node 3 augmented with 33 MW HPC at both node 1 and node 2. Local electricity demand at each node is modeled as a Gaussian random walk with a standard deviation $\sigma = 5$, as are wind and solar generation. Compute load arrives at a centralized demand server, to be served across the case-dependent available HPC capacity. Note that we impose generation capacity limits across all nodes as well. Costs for generation and movement of energy, including carbon credit offsets, are described in Table III.

TABLE III
COST ASSUMPTIONS IN SIMULATION

Source	Cost (\$/MWhr)
Solar Farm (node 1)	\$10/MWhr
Wind Farm (node 2)	\$20/MWhr
Gas Power Plant (node 3)	\$50/MWhr
Energy Transport > 1000 miles	\$40/MWhr

Taking on the perspective of a centralized scheduler, one approach to matching supply to demand is summarized in Algorithm 1. At all nodes, local non-compute demand is met

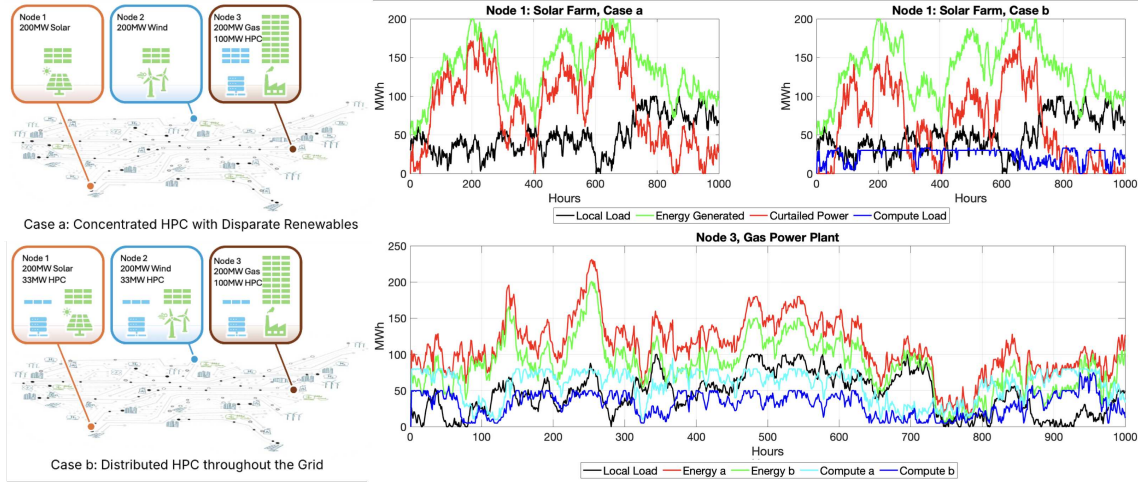


Fig. 4. (Left) Example simulation HPC configurations. (Right) Energy generation, curtailment, and usage time series in simulation cases a and b.

to the extent possible by local generation, with the gas turbine plant acting as a slack bus in scenarios involving renewable shortfalls. Any excess renewable generation is used to serve the centralized compute load, with the gas generator again acting as a slack bus to the extent possible.

Algorithm 1 Load Optimization Algorithm

- 1: **Input:** Set of Nodes $N = \{n_i, i = 1 : M\}$ with properties: $E_{Capacity}$, $C_{Capacity}$, and $CarbonIntensity$
- 2: **for** each time step **do**
- 3: $\forall n \in N$: Calculate excess energy and excess compute capacity for each node and sort by cost
- 4: **while** There exist excess energy at low cost nodes **do**
- 5: Calculate Cost to serve non-compute loads
- 6: Move Energy as to meet Non Compute Demand, minimizing cost
- 7: Update excess energy and compute capacity for each node and sort by cost
- 8: Calculate maximum compute load that can be moved
- 9: Move compute load to lowest cost nodes until compute load met
- 10: **end while**
- 11: **end for**
- 12: **Output:** Compute load transfer profile

As shown in Fig. 4, curtailed generation levels at the solar farm are higher in case (a) than in case (b) due to the ability to flexibly use excess energy to process compute load, e.g., HCPMC. The lower graph shows the overall impact of distributed HPC on the gas power plant. In case (a) the plant must make up for power deficits in the system, while in case (b) it consistently operates at lower power level, thanks to the distributed compute load serviced by renewable sources. An expanded simulation shows the results for varying levels of HPC distribution, shown in Fig. 5 where the simple experiment described above is repeated for 100 bootstraps of Gaussian random walks of power generation and load for levels of

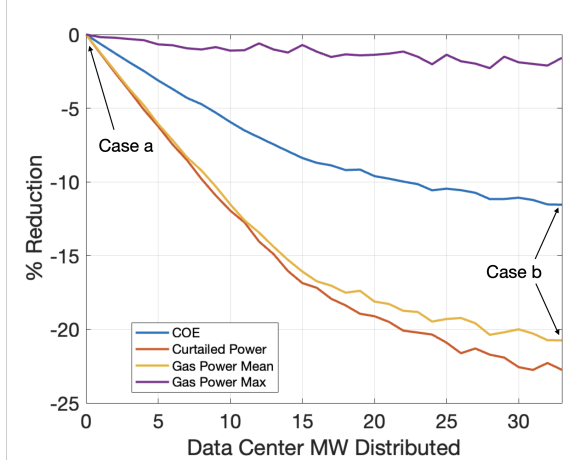


Fig. 5. Cost of energy (COE), renewable curtailment, and gas power utilization statistics with increasingly distributed HPC.

distributed HPC between zero (case (a) above) and distribution of additional HPC capacity at the other nodes (33 MW of compute each, case (b)). The cost of electricity, including generation and movement, drops by approximately 12% as HPC capacity becomes more distributed. Similarly, both curtailed renewable production and mean gas power generation levels decline with more uniform HPC distribution. An additional benefit is the peak gas generator power also drops by 2-3% as the distribution of compute load increases, implying that grid spinning reserve and capacity levels can be reduced.

Fig. 6 further explores nodal operations when the compute load is distributed in a scenario marked by wind farm underproduction. here we observe that with the distributed HPC load among renewable energy sources, compute load served by gas production at node 3 decreases in case (b). This is because we take advantage of curtailed power generated by renewable sources by transferring compute load appropriately.

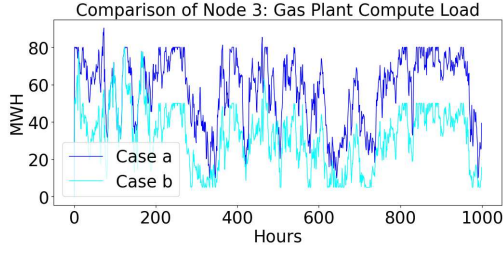


Fig. 6. Comparison of gas power plant node compute load for case a and b.

V. MARKET OPERATION AND PRICING

As mentioned in Section III-A, a significant portion of compute load is flexible in time. Thus, while HPC capacity may be available at a given moment, some consumers may be amenable to adjusting their usage of these resources depending upon their willingness to pay a given price. On the other hand, renewable producers, particularly those with excess generation slated to be curtailed, may be willing to serve compute load to a degree depending upon prevailing market prices. In this section, we examine a stylized spot HPC capacity market, assessing the impact of HPC distribution across the system described in Section III-B. We assume for simplicity that all HPC capacity in the system is sold in this spot market, after electrical loads are served. Therefore, any renewable adjacent HPC is powered by production that would otherwise be curtailed.

We focus here on the three-node setting of Section IV and a market in which the three suppliers each offer supply schedules at each time slot mapping prices to amounts of HPC capacity provided to a single customer, e.g., a compute load aggregator. We assume that compute load arriving at a given time may be only partially served, incurring per unit monetary shortfall penalty θ , and that the consumer solves the following price selection problem at each time step, which balances the cost of compute with a penalty for supply shortfall

$$\min_{p \geq 0} p \sum_{i=1}^3 s_i(p) + \theta(d - s_i(p))_+, \quad (1)$$

where $x_+ = \max\{x, 0\}$, p is the price selected. We suppress the dependency of compute demand d and the supply functions s_i on time. Regarding unserved compute load at a given time step, we later compare two options: unserved load is either discarded entirely or added to a queue, so that it rolls over and contributes to the amount d in (1) at the following time step.

Turning to pricing, adhering to the well-known merit order supply paradigm for electricity markets [29], one way to price compute jobs as they are distributed geographically is as follows: make use of computing capacity in order of increasing unit cost bids, and set the per unit price equal to the rate charged by the final or marginal supplier. In this setting, bids take the form $s_i(p) = 0$ for $p < c_i$, where c_i is a constant marginal cost, and $s_i \in [0, \bar{s}_i]$ for $p \geq c_i$, where \bar{s}_i is the maximum capacity supplier i can provide. The lefthand plot in Figure 7 shows representative individual and market-wide supply curves

under this approach. Note that the flat portions of the renewable supply curves may correspond to a value below the maximum HPC capacity, whenever curtailable energy levels fall below these maximum values. In the context of such bids, θ essentially sets the customer's highest acceptable price in the sense that no supply will be requested from supply i with $c_i > \theta$.

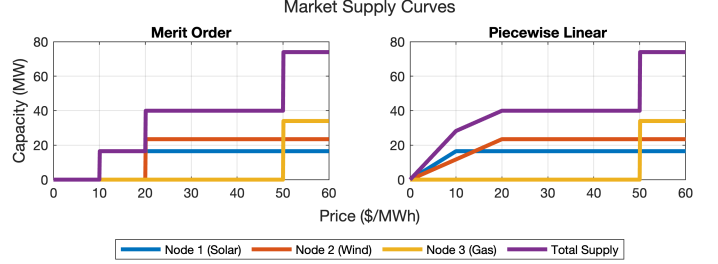


Fig. 7. Alternative market-wide supply curves arising from merit order or linear supply function bidding from of renewable-backed HPC providers.

As we assume that renewable generation is curtailed if not committed to HPC, solar and wind farm operators may be willing to offer HPC capacity for rates below those reported in Table III. One relatively simple alternative bid format to the merit order approach capturing this willingness consists of piecewise linear supply functions (PLSF) specified as linear ramps from zero to the full market rate at each renewable supplier's (time dependent) maximum capacity. The righthand plot in Fig. 7 shows adjusted supply curves following this format. Note that we assume the gas turbine supply curve is unchanged as production is fully controllable. Similar curves on the demand side in related settings have been shown to be helpful in improving productivity and lowering costs [30].

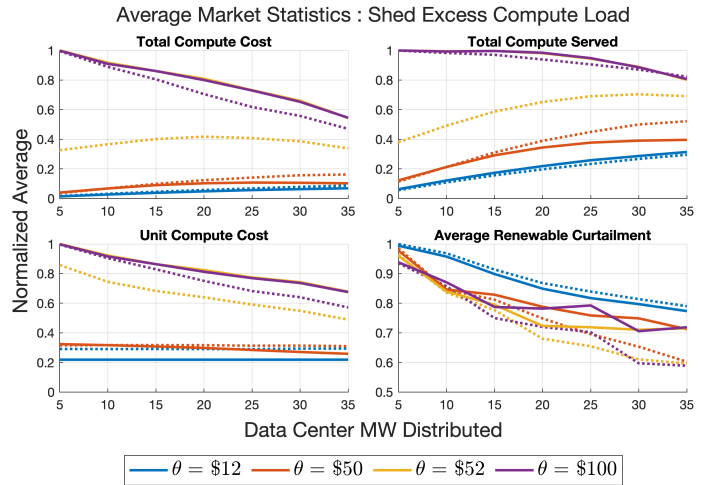


Fig. 8. Mean metrics over 1000 traces (solid : merit order, dotted : piecewise linear supply functions).

Figs. 8 and 9 illustrate outcomes associated with these bid formats assuming unmet compute load is shed or rolled forward in time, respectively, for select values of θ , in terms of average compute total compute cost, service and unit cost (overall

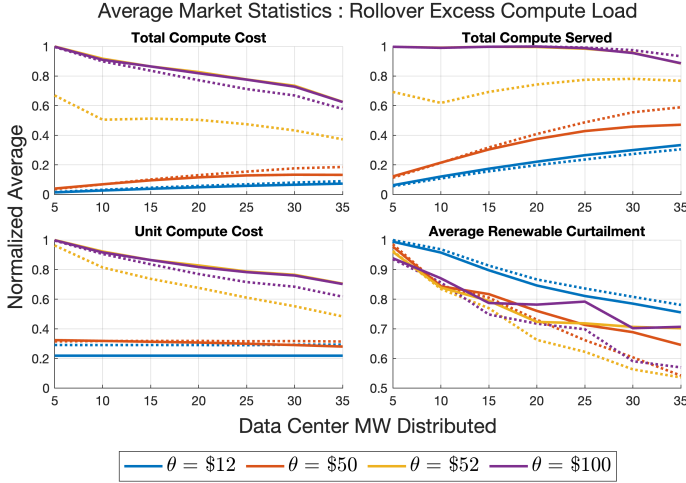


Fig. 9. Mean metrics over 1000 traces (solid : merit order, dotted : piecewise linear supply functions).

service cost divided by load served), along with the average amount of renewable production curtailed after commitments are made to HPC. Averages are taken over 1000 traces each, with each trace consisting of 5 minute interval time steps over 24 hours. In these simulations, we held the total network HPC level constant at 100 MW, and horizontal axis values in Figs. 8 and 9 refer to the capacity located at each node 1 and 2. For instance, at the ‘30’ tick marks, 60 MW of HPC capacity is evenly split between nodes 1 and 2, with 40 MW left to node 3. Each subplot in either figure is individually normalized by the maximum of the raw data underlying the curves displayed. Solid lines correspond to merit order supply, while dotted lines are used for PLSF bids. Fig. 8 plots results assuming that any excess compute load remaining after the consumer selects their price point via 1 is discarded, while Fig. 9 displays results when the consumer simply rolls remaining compute load over to the next slot, until the end of a given 24 hour trace.

With respect to the parameter θ , in view of Table III, under the merit order approach, the \$10/MWh rate offered by the solar farm represents the lowest θ value for which *any* HPC capacity can be purchased, while the \$50/MWh rate offered by the gas turbine represents the minimum θ value required to make use of the entire network’s HPC capacity, regardless of location. Note that under merit order market operation, mean cost and service totals are essentially the same even as θ nearly doubles from \$52 to \$100, while under the PLSF format, a tradeoff is evident: lower willingness to pay translates to lower total compute served, at lower total and per unit costs. While not plotted, a continuum of such curves exists for the PLSF setting, both between the \$12 and \$50 curves, as well as the \$52 and \$100 curves, reflecting greater flexibility in terms of how the market responds to varying customer preferences.

At a high level, both Figs. 8 and 9 demonstrate that PLSF bids yield reduced average renewable curtailment, particularly at high levels of renewable HPC colocation. Assuming a high

willingness to pay (or equivalently, eagerness to meet demand quickly) on the part of the HPC customer, PLSF bids offer comparable total compute service levels at reduced total and per unit costs. For lower values of θ , overall and per unit costs increase slightly under PLSF bids, while compute load served increases as θ approaches \$50.

Concentrating on the comparison between shedding or rolling over excess compute load at the end of a given time slot, the latter yields higher levels of average total compute served, and lower levels of renewable curtailment, particularly for high values of θ . In terms of compute flexibility in time, these two approaches represent outcomes under two extremes - either compute load is to be served immediately, or it can essentially be delayed indefinitely, or at least to the end of a particular trace (at cost). We leave deadline-aware scheduling scenarios falling between these two extremes for future work.

VI. COST BENEFIT ANALYSIS

The California Independent System Operator’s (CAISO) weighted real-time average prices in 2022 were \$22.98 per/MWh for regulation down, \$12.53/MWh for regulation up, \$8.51/MWh for spinning reserve, and \$2.08/MWh for non-spinning reserve [31]. By utilizing renewable assets at max capacity, and providing schedulable real-time variable loading, utilities can release a controllable percentage of their generational capacity to the grid downstream of the schedulable load.

This paradigm allows these assets to operate on very short time scales as spinning reserve as well as regulation up or down within operational bounds that do not involve the dynamics or time scales of the turbines. Adding this utility-controlled demand response allows for operating assets at close to steady state and shaping the output to the grid based on the relative energy cost of selling output, or consuming it to sell as compute capability. Co-located HPC allows for the introduction of a salvage value for otherwise curtailed power, capitalizing on the opportunity cost of previously unutilized generation capacity to provide compute as a service. CAISO curtailed 2.4 TWh in 2022 that otherwise could have been generated and re-directed to revenue generating compute services rather than being curtailed [32].

Within the current grid control paradigm, marginal increases in renewable power installations create commensurate increases in the curtailment rate. By introducing controllable loading, this paradigm can be flipped, reducing the curtailment rate with the introduction of net demand that is coordinated by the utility and optimized based on marginal energy costs. Based on a study from UC Davis, an estimated marginal 1 MWh increase in hourly net demand added in California results in a reduction in the average daily curtailment rate for utility-scale solar and wind, with a marginal reduction in the curtailment rate of 1.5% for this newest MWh of load. With additional granularity for time of day, scheduling this marginal demand for times of peak effect (10am in this case), one could expect as much as a 5.3% reduction in marginal curtailment rate [33]. The effect of this co-located compute would be beneficial to

the gas industry as well. Allowing turbines to operate above capacity with downstream load to filter its output. This method can have a smoothing effect on the ramp rates necessary to meet variable demand helping to extend the lifetime of a turbine.

VII. CONCLUSIONS AND NEXT STEPS

We plan to extend this work to the deployment of HPC loads towards grid stability related services including voltage and frequency regulation using test systems, e.g., IEEE benchmarks, and expand generation modeling to include coal and other carbon intensive sources. Accurate grid modeling and simulation will be essential to analyze the grid's behavior under various energy portfolios and load conditions (e.g., local and compute load) when coupled with our HPC scheduling approach. Regarding market design, we plan to incorporate demand-side bidding and examine two-sided market clearing, including equilibrium analysis and strategic behavior, and consider issues around compensating flexibility, as well as the role of HPC workload aggregators. In total, we aim to build a comprehensive approach to efficiently maintaining grid stability and reliable power delivery that leverages spinning HPC demand.

REFERENCES

- [1] P. INTERNATIONAL ENERGY AGENCY, "Electricity 2024," 2024, p. Licence: CC BY 4.0. [Online]. Available: <https://www.iea.org/reports/electricity-2024>
- [2] I. Electric Power Research Institute, "Analyzing artificial intelligence and data center energy consumption," 2024, p. Whitepaper. [Online]. Available: <https://www.epri.com/research/products/000000003002028905>
- [3] DOE, "National transmission needs study," October 2023.
- [4] U. of Chicago, "Zero-carbon cloud and sustainable cloud computing." [Online]. Available: <https://people.cs.uchicago.edu/~aachien/lssg/research/zcccloud/>
- [5] J. Xing, B. Acun, A. Sundarajan, D. Brooks, M. Chakkaravarthy, N. Avila, C.-J. Wu, and B. C. Lee, "Carbon responder: Coordinating demand response for the datacenter fleet," *arXiv preprint arXiv:2311.08589*, 2023.
- [6] S. P. Ekanayake, T. Shah, and S. Evans, "Optimizing emissions for machine learning training," in *2023 IEEE Conference on Technologies for Sustainability (SusTech)*, 2023, pp. 237–238.
- [7] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, and J. Gao, "Large language models: A survey," *arXiv preprint arXiv:2402.06196*, 2024.
- [8] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang *et al.*, "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1–45, 2024.
- [9] T. Yarally, L. Cruz, D. Feitosa, J. Sallou, and A. Van Deursen, "Uncovering energy-efficient practices in deep learning training: Preliminary steps towards green ai," in *2023 IEEE/ACM 2nd International Conference on AI Engineering–Software Engineering for AI (CAIN)*. IEEE, 2023, pp. 25–36.
- [10] H. Huang, T. Shah, S. Evans, and S. Yoo, "Energy efficient streaming time series classification with attentive power iteration," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 11, 2024, pp. 12 574–12 582.
- [11] S. C. Evans, T. Shah, H. Huang, and S. P. Ekanayake, "The entropy economy and the kolmogorov learning cycle: Leveraging the intersection of machine learning and algorithmic information theory to jointly optimize energy and learning," *Physica D: Nonlinear Phenomena*, vol. 461, p. 134051, 2024.
- [12] S. Budennyy, V. Lazarev, N. Zakharenko, A. Korovin, O. Plosskaya, D. Dimitrov, V. Arkhipkin, I. Oseledets, I. Barsola, I. Egorov, A. Kosterina, and L. Zhukov, "Eco2ai: carbon emissions tracking of machine learning models as the first step towards sustainable ai," 2022. [Online]. Available: <https://arxiv.org/abs/2208.00406>
- [13] A. Ukarande, T. Basaklar, M. Cao, and U. Ogras, "Pact: Accurate power analysis and carbon emission tracking for sustainability," in *Proceedings of the 29th ACM/IEEE International Symposium on Low Power Electronics and Design*, 2024, pp. 1–6.
- [14] M. Konopik, T. Korten, E. Lutz, and H. Linke, "Fundamental energy cost of finite-time parallelizable computing," *Nature Communications*, vol. 14, no. 1, p. 447, 2023.
- [15] J. McDonald, B. Li, N. C. Frey, D. Tiwari, V. Gadepally, and S. Samsi, "Great power, great responsibility: Recommendations for reducing energy for training language models," in *NAACL-HLT*, 2022. [Online]. Available: <https://api.semanticscholar.org/CorpusID:248887431>
- [16] M. F. Argerich and M. Patiño-Martínez, "Measuring and improving the energy efficiency of large language models inference," *IEEE Access*, 2024.
- [17] N. Alizadeh and F. Castor, "Green ai: A preliminary empirical study on energy consumption in dl models across different runtime infrastructures," in *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI*, 2024, pp. 134–139.
- [18] R. Desislavov, F. Martínez-Plumed, and J. Hernández-Orallo, "Compute and energy consumption trends in deep learning inference," *arXiv preprint arXiv:2109.05472*, 2021.
- [19] Y. Li, M. Mughees, Y. Chen, and Y. R. Li, "The unseen ai disruptions for power grids: Llm-induced transients," *arXiv preprint arXiv:2409.11416*, 2024.
- [20] Sourceability, "The need for a strong power grid infrastructure in the age of ai," 2024. [Online]. Available: <https://sourceability.com/post/the-need-for-strong-power-grid-infrastructure-in-the-age-of-ai>
- [21] D. Zhang, G. Shafiullah, C. K. Das, and K. W. Wong, "Optimal allocation of battery energy storage systems to improve system reliability and voltage and frequency stability in weak grids," *Applied energy*, vol. 377, p. 124541, 2025.
- [22] M. O. Qays, I. Ahmad, D. Habibi, and M. A. Masoum, "H ∞ control system for synchronous condensers in renewable energy-integrated weak grids," *CSEE Journal of Power and Energy Systems*, 2025.
- [23] S. Evans, H. Huang, T. Shah, B. Rose, B. Wilson, I. Claydon, D. Vernoooy, and W. Ziminsy, "Reduced-energy random number generation for dynamic grid stabilization of a power system," U.S. Patent Application 63/604,517, Filed 11/30/2023.
- [24] D. DeSantis, B. D. James, C. Houchins, G. Saur, and M. Lyubovskiy, "Cost of long-distance energy transmission by different carriers," *iScience*, vol. 24, no. 12, p. 103495, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2589004221014668>
- [25] U. E. I. Administration, "Solar and wind power curtailments are rising in california," 2023. [Online]. Available: <https://www.eia.gov/todayinenergy/detail.php?id=60822>
- [26] C. . . EPRI, Palo Alto, "Ancillary services in the united states: Technical requirements, market designs and price trends." [Online]. Available: <https://www.epri.com/research/products/000000003002015670>, LastOnlineAccessNovember11,2024
- [27] IBM, "What is a hyperscale data center?" [Online]. Available: <https://www.ibm.com/topics/hyperscale-data-center>
- [28] FERC, "Electric power markets – national overview." [Online]. Available: <https://www.ferc.gov/electric-power-markets>, LastOnlineAccessonNovember11,2024
- [29] F. Sensfuß, M. Ragwitz, and M. Genoese, "The merit-order effect: A detailed analysis of the price effect of renewable electricity generation on spot market prices in germany," *Energy policy*, vol. 36, no. 8, pp. 3086–3094, 2008.
- [30] M. A. Islam, X. Ren, S. Ren, and A. Wierman, "A spot capacity market to increase power infrastructure utilization in multi-tenant data centers," in *Proceedings of the 2017 ACM SIGMETRICS/International Conference on Measurement and Modeling of Computer Systems*, 2017, pp. 19–20.
- [31] "2022 annual report on market issues and performance." [Online]. Available: <https://www.caiso.com/documents/2022-annual-report-on-market-issues-and-performance-jul-11-2023.pdf>
- [32] "Solar and wind power curtailments in california, 2023." [Online]. Available: <https://www.eia.gov/todayinenergy/detail.php?id=60822>
- [33] K. Novan and Y. Wang, "Estimates of the marginal curtailment rates for solar and wind generation." [Online]. Available: https://arefiles.ucdavis.edu/uploads/pub/2023/04/11/curtailment_paper.pdf