THE GRID AS OBSERVABLE STATE: INFRASTRUCTURE-AWARE RESILIENCE FOR LARGE-SCALE TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale distributed training relies on checkpointing to recover from failures. Its resilience also depends on the infrastructure shared by the sites that host it. We argue that resilience evaluations should make this shared exposure visible alongside hardware failure statistics.

Using five years of hourly ERCOT electricity-price data, we characterize the spatial and temporal structure of negative-price episodes. In 2025, negative-price indicators at the studied North Texas substations have pairwise correlations of 0.875–0.933. In the corridor-level series, episodes last 3.8 hours on average, and 46 of 56 episode starts (82.1%) fall between 03:00 and 12:00 CST, a window covering 37.5% of the day. These observations reveal shared market exposure and concentrated timing, two dimensions of infrastructure context relevant to resilience evaluation.

Whether market conditions affect training availability depends on the facility's operational response. Under an explicit episode-to-interruption scenario, a checkpointing illustration connects the observed timing to a break-even criterion involving interruption exposure, checkpoint effort, and recovery costs. Comparisons at matched nominal checkpoint frequency show why resource accounting matters when assessing adaptation. Together, the empirical findings and conditional analysis motivate evaluating observable infrastructure signals alongside facility telemetry to identify when they can improve training-resilience decisions.

1 INTRODUCTION

For a training run spanning weeks and thousands of accelerators, recovery is part of the job. Modern training systems periodically checkpoint model state and restart from the most recent checkpoint after a failure (Shoeybi et al., 2019; Jiang et al., 2024; Grattafiori et al., 2024). Each checkpoint leaves a shorter path back after a failure, at the price of additional I/O. The interval determines the balance: save state more often and lose less work, or save less often and risk more recomputation.

Classical checkpointing analysis considers a stationary failure process with a constant hazard. Under the standard exponential/Poisson model, the failure rate, or equivalently the mean time between failures (MTBF), determines a static checkpoint interval, with the Young–Daly approximation

$$\Delta^* = \sqrt{\frac{2r}{\lambda}},$$

where r is the checkpoint-write cost and λ is the failure rate (Young, 1974; Daly, 2006). This is a useful baseline, not a universal description of system failures: large-system failure measurements already show non-exponential interarrival times (Schroeder & Gibson, 2006). Our question is what additional information shared external infrastructure can provide for resilience decisions.

A training job's exposure does not end at the rack. We study the power grid as a shared infrastructure layer and ask: *does the grid expose structure that could be useful for training resilience?* We study this question using five years of hourly ERCOT real-time locational marginal price (LMP) data. Negative LMPs are not training failures, nor do they imply that a datacenter was curtailed. They are an observable market signal that can reflect local oversupply and transmission constraints (Schweppe et al., 1988; Hogan, 1992). We use negative-LMP episodes as a proxy for local grid state and study their spatial and temporal structure.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Our position. Training-resilience evaluations should report shared infrastructure exposure alongside hardware failure statistics. Observable infrastructure signals should be assessed for their incremental predictive value and their effect on total recovery cost. Our ERCOT case study motivates this evaluation by documenting spatial dependence, persistence, and time-varying entry risk in a public grid-price signal.

We separate two layers: observable market-state dynamics and the facility response that determines whether training is interrupted. Let G_t denote the observable grid-price state and I_t an actual training interruption. Our data observe G_t, but not I_t. Whether a grid episode affects a training job depends on facility-specific factors such as the interconnection agreement, onsite backup, notice period, and scheduler behavior. The relevant systems quantity is therefore

$$P(I_{t+1} = 1 \mid G_t, h_t, X_t),$$

where h_t is time of day and X_t represents facility-specific conditions. We characterize the structure of G_t and then study the checkpointing implications under an explicit mapping from grid episodes to training interruptions.

Grid-price states are structured. We focus on a North Texas corridor containing substations close to large compute facilities, with comparison sites in Austin and Houston. The resulting grid-price signal shows strong spatial and temporal structure. Within the North Texas corridor, negative-LMP indicators are highly correlated across substations, with pairwise correlations of $\rho = 0.875$–0.933. The state is also persistent: in 2025,

$$P(C_{t+1} = C \mid C_t = C) = 0.737,$$

and observed episodes have a mean duration of approximately 3.8 h. This transition probability measures serial persistence; a geometric duration model provides a descriptive reference for the observed run lengths.

The entry process also varies strongly with time. In 2025, 46 of the 56 episode starts occur between 03:00 and 12:00 CST, a window covering 37.5% of the day. The North Texas $P(N \to C)$ rate increased from 0.00466 per hour in 2021 to 0.00655 in 2025, a 41% increase. In 2025, the North Texas sample contains 56 corridor-level negative-LMP episodes, with a maximum observed duration of 22 h.

Implications for training resilience. The main systems implication is that external interruption risk need not be represented only by a single failure rate. If the probability of interruption depends on observable grid state, then a scheduler can in principle condition checkpointing, placement, or other resilience decisions on that state:

$$P(I_{t+1} = 1) \quad \longrightarrow \quad P(I_{t+1} = 1 \mid G_t, h_t, X_t).$$

This is different from simply estimating a better MTBF. It allows the system to respond to changes in risk over time and to shared conditions outside the compute cluster.

We illustrate the decision implications using checkpoint scheduling. A time-of-day-aware policy allocates more checkpoints to hours with high observed episode-entry concentration and fewer to other hours. Under a specified episode-to-interruption scenario, the analysis connects this temporal allocation to rollback and total recovery overhead. Comparisons at matched nominal checkpoint frequency separate timing from checkpoint effort, while boundary-write accounting tests the sensitivity of the comparison. The resulting break-even criterion links the observed timing of episodes to the conditions under which adaptation can reduce total overhead.

Contributions. This paper makes three contributions.

- **We characterize the spatial and temporal structure of an external infrastructure signal relevant to large-scale training.** Using five years of ERCOT bus-level LMP data, we find strong within-corridor dependence, multi-hour persistence, and large intraday and year-to-year variation in grid-state transitions.
- **We formulate the training-resilience problem in terms of observable infrastructure state.** We explicitly separate grid-price episodes from facility curtailments and training interruptions, and study when conditioning on external state can provide information beyond a stationary failure rate.

- **We illustrate the conditional checkpointing implications and derive a break-even criterion.** A time-of-day-aware policy connects the observed episode timing to rollback under an explicit interruption model. Matched nominal checkpoint budgets distinguish timing from checkpoint effort. The break-even condition links episode concentration, interruption exposure, and additional checkpoint and restart costs.

2 Checkpointing Under Static and State-Dependent Risk

Consider a distributed training job that checkpoints every Δ hours at cost r. If interruptions follow a Poisson process with constant rate λ, the leading-order expected overhead is

$$J_{\text{static}}(\Delta) = \frac{r}{\Delta} + \frac{\lambda\Delta}{2}, \tag{1}$$

where the two terms represent checkpoint I/O and expected rollback work. Minimizing (1) gives the Young checkpoint interval

$$\Delta^* = \sqrt{\frac{2r}{\lambda}} \tag{2}$$

(Young, 1974; Daly, 2006). This model summarizes interruption risk by a single average rate λ and is appropriate when that rate is approximately constant and no additional observable state is predictive of future interruptions.

External infrastructure motivates a state-dependent formulation. Let G_t denote an observable infrastructure state and I_t an actual training interruption. The relevant risk may instead be

$$P(I_{t+1} = 1 \mid G_t, h_t, X_t), \tag{3}$$

where h_t is time of day and X_t represents facility-specific conditions such as backup generation, interconnection arrangements, and scheduler behavior. The empirical analysis characterizes G_t; the decision analysis makes the facility-response assumptions explicit. Together, they identify the observed structure and the conditions under which it could guide checkpoint allocation.

We study three forms of structure. First, correlated grid states across substations imply shared spatial exposure. Second, dependence between G_t and G_{t+1} makes the current state informative about near-term grid conditions; persistence is analyzed separately from whether episode durations are geometric. Third, the entry probability

$$p_{NC}(h) = P(G_t = C \mid G_{t-1} = N, h_t = h) \tag{4}$$

may vary across hours or years, so a single stationary rate can average over useful predictive information.

Section 3 measures these properties in ERCOT data. Section 4 then evaluates their checkpointing implications under an explicit mapping from grid-price episodes to training interruptions.

3 Empirical Evidence: Spatial and Temporal Structure of Grid-Price States

3.1 Data and State Construction

We use hourly ERCOT real-time settlement point prices from January 2021 through December 2025 at the individual bus-node level. Our primary sample is the North Texas corridor, where we aggregate nine bus nodes across Navarro Mills (four nodes) and Venus Switch (five nodes).

Let $\bar{\lambda}_t$ denote the corridor-average real-time LMP at hour t. We define the observed grid-price state as

$$G_t = \begin{cases} C, & \bar{\lambda}_t < 0, \\ N, & \bar{\lambda}_t \geq 0, \end{cases} \tag{5}$$

where C denotes a negative-price state and N a non-negative-price state. A negative-LMP episode is a contiguous sequence of hours for which $G_t = C$.

Negative LMPs can arise from oversupply, congestion, and bidding incentives (Schweppe et al., 1988; Hogan, 1992); we use them as a market-state indicator. Empirical work links negative prices to limits on power-system flexibility (Nicolosi, 2010). All clock times use fixed CST (UTC-6), without daylight-saving adjustment. The nine nodes are NAVARRO1_8, NAVARRO1_8Z, NAVARRO1_9, NAVARRO1_9W, VENSW_0001, VENSW_0701, VENSW_0801, VENUS_N5, and VENUS_S5. The hourly corridor value is their arithmetic mean. The 2025 series has 8,759 observed timestamps; the missing hour, March 8 at 20:00 CST, lies between two non-negative observations and is not imputed.

Transition estimates use adjacent available observations, including the single missing-hour gap as an $N \to N$ pair; it does not alter the observed episode count. We divide the number of transitions from state i to state j by the number of available preceding-state observations in i, excluding the terminal observation. Episode starts are transitions from N to C; observed negative-state run lengths give episode durations. For the spatial analysis, we first average prices within each substation, threshold at zero, and compute Pearson correlations between the resulting binary series at matched timestamps. Figure 1(b) shows the Navarro Mills, Venus Switch, Ennis, and Limmer subset.

A negative price means that the market price of electricity at a location is below zero; it does not by itself mean that electricity is unavailable. The purpose of this threshold is to define a transparent, reproducible market state. Geographical proximity identifies candidate shared exposure, but does not establish the electrical supply path to a particular campus. Accordingly, the spatial results describe dependence between observed substation price states. Whether a facility can exploit or must respond to those states is a separate question about its contracts and operations.

Table 1 summarizes the six substations used in the analysis. Navarro Mills and Venus Switch define the North Texas compute corridor. Limmer and Teal Switch provide a same-sector comparison in the Austin area, while Apache provides an industrial comparison outside the North Texas corridor. Ennis is included as a within-corridor comparison site: it shares the North Texas transmission region with Navarro Mills and Venus Switch but has no hyperscale-campus match in our geospatial merge. This comparison helps distinguish corridor-level structure in the price signal from patterns specific to datacenter-proximate substations.

Table 1: Substations used in the empirical analysis. DC$+1$ denotes a substation within 1 mi of a matched hyperscale campus; DC$=0$ denotes no campus match; DC-1 denotes an industrial comparison outside the North Texas corridor.

Substation	**Role**	**Corridor**	**Nodes**	**Selection criterion**
Navarro Mills	DC$+1$	N. Texas	4	$<$0.4 mi to $\geq$200 MW campus
Venus Switch	DC$+1$	N. Texas	5	$<$0.7 mi to $\geq$58 MW campus
Ennis	DC$=0$	N. Texas	4	Same region; no campus match
Limmer	DC$+1$	Austin	6	$<$0.3 mi to 48 MW campus
Teal Switch	DC$+1$	Austin	6	$<$0.4 mi to same campus
Apache	DC-1	Houston	4	Industrial comparison; out of corridor

3.2 Shared Spatial Exposure

Figure 1(b) shows pairwise correlations of the negative-LMP indicator across substations in 2025. Within North Texas, correlations range from $\rho = 0.875$ to 0.933. The Navarro–Ennis pair has correlation $\rho = 0.933$, exceeding the Navarro–Venus pair ($\rho = 0.875$), even though Ennis has no hyperscale-campus match in our geospatial merge. This suggests that a substantial part of the observed dependence is associated with conditions shared across the transmission corridor, rather than with a single datacenter-proximate site.

The North Texas pairs shown in Figure 1(b) have correlations of 0.875 (Navarro–Venus), 0.933 (Navarro–Ennis), and 0.908 (Venus–Ennis). Their correlations with Limmer range from 0.734 to 0.753. The observed dependence therefore includes both a strong within-corridor component and substantial cross-corridor dependence.

For training systems, the implication is shared external exposure. Replication can protect against hardware failures when replicas do not share the same failure source. By contrast, replicas located

within the same transmission corridor may remain exposed to the same grid conditions. The measured dependence identifies shared exposure in G_t. Its consequences for training interruptions I_t depend on facility interconnections, backup supply, and operating policy.

3.3 Temporal Persistence and Episode Durations

We next examine the temporal structure of the North Texas grid-price state. Table 2 reports the empirical transition matrix for 2025. The probability of remaining in the negative-price state is

$$P(G_{t+1} = C \mid G_t = C) = 0.737,$$

compared with

$$P(G_{t+1} = C \mid G_t = N) = 0.0066.$$

Thus, once the corridor enters state C, the current grid state is highly informative about the state in the following hour. There were 56 negative-LMP episodes in 2025, with a maximum observed duration of 22 h.

Table 2: Empirical transition matrix for the North Texas grid-price state in 2025. States are NON-NEGATIVE ($\bar{\lambda}_t \geq 0$) and NEGATIVE ($\bar{\lambda}_t < 0$).

	NON-NEGATIVE	NEGATIVE
NON-NEGATIVE	0.9934	0.0066
NEGATIVE	0.2629	0.7371

56 episodes in 2025; maximum duration = 22 h.

The transition matrix shows strong persistence, but persistence alone does not imply that episode durations are non-memoryless. A homogeneous two-state Markov chain with constant $P(C \to C) = p$ produces geometric run lengths,

$$P(D = d) = (1 - p)p^{d-1}, \tag{6}$$

with expected duration

$$E[D] = \frac{1}{1 - p}. \tag{7}$$

For $p = 0.737$, this gives an expected episode duration of approximately 3.8 h.

Figure 1(a) compares the observed run lengths with a geometric reference fitted to the same mean. This comparison is descriptive; the transition matrix establishes serial dependence in the state sequence without requiring rejection of a geometric duration model.

The two transition probabilities answer different questions. The entry probability describes how often a new episode begins when the preceding state is non-negative; the persistence probability describes continuation after an episode has already begun. In our one-interruption-per-entry scenario, a 22 h episode generates one interruption, not 22 independent failures. Duration and onset therefore should not be used interchangeably as the failure rate in a checkpoint calculation. Observing a negative price after entry also does not provide advance warning of that entry. Our checkpoint example uses the historical time-of-day profile to allocate effort before a possible interruption; it does not treat the contemporaneous price as a forecast.

Persistence therefore characterizes the continuation of an observed market state, while entry timing determines the rollback calculation in the scenario below.

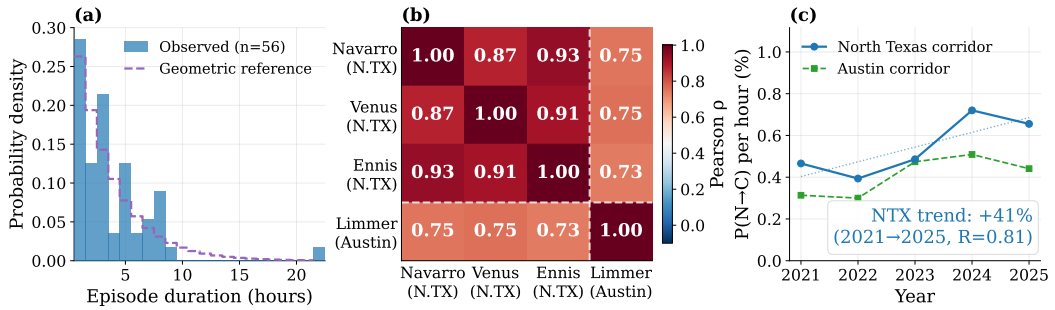


Figure 1: **Spatial and temporal structure of ERCOT negative-LMP states** (North Texas, 2025 unless noted). (a) Observed episode durations ($n = 56$, including the 22 h maximum) and a geometric reference fitted to the same mean. (b) Pairwise correlations of negative-LMP indicators across substations; within-corridor correlations are $\rho = 0.875$–0.933. (c) Annual entry rate $P(G_{t+1} = C \mid G_t = N)$ from 2021–2025; the North Texas rate increases by 41%.

3.4 Time-Varying Entry Risk

We next examine whether the probability of entering the negative-price state is stable over time. Figure 1(c) shows the annual transition rate

$$P(G_{t+1} = C \mid G_t = N)$$

for the North Texas and Austin corridors from 2021 to 2025. In North Texas, the entry rate increased from 0.00466 per hour in 2021 to 0.00655 in 2025, a 41% increase ($R = 0.81$). Austin also shows an upward trend over the same period ($R = 0.77$), although its entry rate remains below North Texas throughout the sample.

The North Texas entry rate is approximately $1.49\times$ the Austin rate in 2021 and $1.51\times$ in 2025. These are descriptive comparisons across five annual observations, rather than an attribution of changes to a single physical cause. Renewable generation and transmission constraints provide relevant context (Denholm et al., 2015; Bird et al., 2016), but the price panel alone does not identify their separate effects.

A single average compresses this timing into one number. If grid state is informative about interruption risk, a stationary rate can miss useful intraday and longer-run variation. Section 4 examines the checkpointing implications under an explicit interruption mapping.

4 Checkpointing Under Time-Varying Grid Risk

Section 3 shows that the probability of entering the negative-price state varies substantially over the day. In 2025, 46 of 56 episode entries occur in $[03{:}00, 12{:}00)$ CST: 82.1% of entries are concentrated in 37.5% of the day. We use this observed concentration to illustrate how checkpoint effort can be allocated across periods with different entry risk.

The fraction $46/56$ describes where episode starts fall, not an 82.1% probability of interruption during any given morning. In contrast, $p_{NC}(h)$ divides entries at hour h by preceding non-negative observations eligible to enter an episode. Our rollback calculation allocates loss over observed entries; forecasting would require the latter denominator and evidence linking price states to interruptions.

We use this structure to study a simple conditional checkpointing rule. Because the ERCOT data do not contain facility-level training interruptions, the calculation requires an explicit mapping from grid-price episodes to interruptions. We therefore consider a counterfactual scenario in which each negative-LMP episode produces one training interruption at episode entry. Under this mapping, the calculation illustrates how observed temporal concentration enters a checkpoint-cost comparison and identifies the conditions under which adaptation can pay off.

4.1 Policy Structure

We compare a static checkpoint policy with a time-of-day-aware policy. The static baseline uses a fixed checkpoint interval Δ_{static} throughout the day.

For the time-of-day-aware policy, let $\mathcal{H}_{\text{high}}$ denote a specified morning window with elevated episode-entry concentration in the observed sample. The checkpoint interval is

$$\Delta(h) = \begin{cases} \Delta_{\text{high}}, & h \in \mathcal{H}_{\text{high}}, \\ \Delta_{\text{low}}, & h \notin \mathcal{H}_{\text{high}}. \end{cases} \tag{8}$$

For the 2025 North Texas calibration, we use $\mathcal{H}_{\text{high}} = [03{:}00, 12{:}00)$ CST, $\Delta_{\text{high}} = 1\,\text{h}$, and $\Delta_{\text{low}} = 6\,\text{h}$. We also place a checkpoint at the start of the high-risk window so that work accumulated during the preceding low-risk period is not carried into the hours with the highest observed entry probability.

At execution time, this rule requires only the clock and a previously estimated risk profile. Here, the 2025 sample supplies the calibration; the calculation is a retrospective illustration rather than an out-of-sample forecasting evaluation.

4.2 Conditional Rollback Savings and Total Overhead

We first isolate rollback-related forward compute. Let $T = 8760\,\text{h}$, let $\mu = 56$ be the number of North Texas negative-LMP episodes observed in 2025, and let $f_{\text{high}} = 46/56$ denote the observed fraction of episode entries in $[03{:}00, 12{:}00)$ CST. We use calendar-year normalization T. The following first-order approximation assumes interruption phase is uniform within a checkpoint interval, giving mean rollback $\Delta/2$; it does not replay the exact schedule at window boundaries.

Under the assumed one-interruption-per-episode mapping, expected rollback-related forward compute for a static checkpoint interval is

$$W_{\text{static}} = \frac{\mu}{T}\frac{\Delta_{\text{static}}}{2}. \tag{9}$$

For the time-of-day-aware policy,

$$W_{\text{tod}} = \frac{\mu}{T}\left[f_{\text{high}}\frac{\Delta_{\text{high}}}{2} + (1 - f_{\text{high}})\frac{\Delta_{\text{low}}}{2}\right]. \tag{10}$$

Using $\Delta_{\text{static}} = 4\,\text{h}$, $\Delta_{\text{high}} = 1\,\text{h}$, and $\Delta_{\text{low}} = 6\,\text{h}$ gives

$$W_{\text{static}} = 1.28\%, \qquad W_{\text{tod}} \approx 0.61\%.$$

The 52.7% rollback reduction compares policies with different checkpoint effort. To separate timing from effort within the approximation, let $a = 9/24$ be the high-risk fraction of the day. The time-of-day policy's nominal checkpoint frequency is

$$\nu_{\text{tod}} = \frac{a}{\Delta_{\text{high}}} + \frac{1-a}{\Delta_{\text{low}}} = \frac{23}{48}\,\text{h}^{-1}. \tag{11}$$

A static interval $1/\nu_{\text{tod}} = 48/23 \approx 2.09\,\text{h}$ matches this nominal frequency. Its annual rollback is $56(24/23) = 58.43\,\text{h}$, versus $46(1/2) + 10(6/2) = 53\,\text{h}$ for the time-of-day policy, a 9.3% reduction. Nominal frequency is a continuous-rate approximation, not an executed checkpoint count. As a sensitivity case, if the boundary checkpoint adds one distinct write per day, matching $\nu_{\text{tod}} + 1/24$ gives a static interval of $1.92\,\text{h}$ and rollback of $53.76\,\text{h}$. Holding the same uniform-phase rollback estimate at $53\,\text{h}$ yields a 1.4% reduction. This case varies checkpoint accounting; it does not replay boundary timing.

Rollback loss is only one component of training overhead. For a checkpoint policy π, write total overhead as

Table 3: Conditional annual rollback under the episode-to-interruption mapping and uniform-phase approximation. The 1 h/6 h time-of-day (TOD) policy loses 53 h in each comparison. Matching refers to nominal write frequency, assuming equal cost per write; I/O and restart costs are not included. The extra daily write is an accounting sensitivity assumption.

Static reference	Interval (h)	Rollback (h)	TOD reduction
Specified 4 h baseline	4.00	112.00	52.7%
Matched nominal frequency	2.09	58.43	9.3%
Matched with one extra write/day	1.92	53.76	1.4%

$$J(\pi) = W_{\text{rollback}}(\pi) + C_{\text{checkpoint}}(\pi) + C_{\text{restart}}(\pi). \tag{12}$$

For the 1 h/6 h rule versus static 4 h, the equations give

$$W_{\text{static}} - W_{\text{tod}} = \frac{\mu}{T}(2.5 f_{\text{high}} - 1). \tag{13}$$

Rollback decreases whenever $f_{\text{high}} > 0.4$; the observed $46/56$ concentration exceeds this threshold. If a fraction q of episodes interrupts training independently of entry hour, this saving scales by q, while the nominal cost of scheduled checkpoints does not. This identifies the trade-off without assuming that every price episode reaches a training job.

More generally, a policy reduces total overhead only when

$$q(W_{\text{static}} - W_{\text{tod}}) > \Delta C_{\text{checkpoint}} + \Delta C_{\text{restart}}, \tag{14}$$

where the W terms use one interruption per episode, and each ΔC is the time-of-day policy's additional cost relative to static checkpointing, evaluated under the same interruption exposure q. Against static 4 h, the rollback saving is 59 h per year, or 0.674 percentage points of the calendar-year normalization. At exposure q, additional checkpoint and restart overhead must therefore remain below $59q$ job-level equivalent hours. Table 3 shows why the reference policy matters: matching nominal write frequency leaves a much smaller rollback margin. These comparisons identify a conditional break-even criterion, rather than a full-cost optimum.

5 Relation to Prior Work and Scope

Failure models and recovery mechanisms. Predictive checkpointing studies how forecast accuracy and advance warning affect checkpoint decisions (Aupy et al., 2014). Complementary work changes the cost of recovery: SCR uses multilevel checkpoint storage (Moody et al., 2010), CheckFreq adapts iteration-level checkpoint frequency and pipelines I/O (Mohan et al., 2021), and Gemini combines replicated in-memory checkpoints with placement and traffic scheduling (Wang et al., 2023). Bamboo instead uses redundant computation to tolerate preemptions in pipeline-parallel training (Thorpe et al., 2023). Infrastructure-conditioned risk could guide when or where to use these mechanisms; it does not replace them. Their different costs also explain why a rollback percentage alone cannot rank resilience policies.

Energy-aware operation and grid coordination. Carbon-aware computing shifts flexible work using energy-system signals (Radovanović et al., 2023). Carbon Explorer jointly studies renewable supply, storage, and workload scheduling for datacenter design (Acun et al., 2023); Carbon Responder allocates demand-response effort across heterogeneous workloads using performance models (Xing et al., 2023). More recently, OpenG2G couples measured AI-service characteristics with grid simulation to study runtime coordination (Chung et al., 2026). These works establish infrastructure-aware operation as an active systems direction. Our focus is the structure of an observed grid-price signal and the conditional recovery-cost implications of its timing. Planned load modulation, carbon reduction, and recovery from interruptions involve different objectives and operational constraints.

Scope. The transition matrix describes the selected ERCOT price panel. Geographical proximity does not verify electrical supply connections, and transfer to other markets or facilities requires independent validation.

6 From Infrastructure Context to Resilience Decisions

We propose that resilience evaluations report infrastructure exposure and test the incremental value of information available before a decision. The practical question is which additional observation changes the best action. Shared market conditions can lead to different outcomes at facilities with different backup supply, contractual obligations, and scheduling policies.

Record the operational context. A useful reporting unit is a training interruption linked to its operational context. Relevant fields include the interruption timestamp, affected capacity, available advance notice, recovery procedure, checkpoint-write cost, and whether the interruption was associated with a facility or infrastructure event. Recording a shared infrastructure domain can help distinguish independent equipment failures from common exposure. Market prices can accompany these records as an observable signal; they should not substitute for the event label.

Compare information available before the decision. A stationary-rate baseline asks whether timing information matters; a time-of-day baseline asks whether a live signal adds to recurring patterns. Facility telemetry and dispatch notices are further alternatives. Information must arrive early enough for the action: completing a checkpoint before an interruption requires different lead time from planning recovery afterward. Calibration and evaluation periods should preserve this ordering.

Comparisons should match measured checkpoint effort or tune policies to the same total-cost objective. Our nominal-frequency comparison is an analytical step toward this requirement; an implementation must count boundary writes, checkpoint duration, and recovery behavior explicitly. The four-hour reference is a specified baseline, not an optimized policy.

Match the objective to the action. Rollback savings matter only in relation to the extra costs of the action. Checkpoint writes can extend elapsed time; placement and replication add communication and capacity costs. Evaluations should report these costs alongside recomputation, rather than ranking policies by rollback alone.

Specify when adaptation is warranted. If backup supply isolates training from grid events and no external signal improves interruption prediction, an equipment-based risk model may suffice. If facility telemetry or dispatch notices provide the actionable information, public prices must add value beyond them. Infrastructure signals become useful when they improve decisions with sufficient lead time and a net cost benefit. Our position calls for these comparisons, not automatic deployment of a price-triggered checkpoint rule.

7 Conclusion

Shared infrastructure exposure has measurable spatial and temporal structure. In the 2025 North Texas sample, substation-level negative-price indicators have pairwise correlations of 0.875–0.933, and 82.1% of corridor-level episode entries fall within nine morning hours. These observations make a concrete case for including infrastructure context in training-resilience evaluations. The conditional checkpoint illustration connects the observed timing to recovery costs: its matched nominal-budget comparisons and break-even criterion identify the assumptions and costs that determine whether adaptation pays off.

We advocate evaluating infrastructure signals alongside facility telemetry and reporting the infrastructure context of training interruptions. The aim is concrete: identify which signals help a scheduler make a better decision, for which facilities, and at what cost.

Reproducibility statement

The sample selection and negative-price state definition are given in Section 3.1; Sections 3.2–3.4 describe the empirical summaries. Section 4 specifies the counterfactual interruption mapping, checkpoint intervals, and analytical rollback calculation, including the nominal checkpoint-frequency comparison.

AI USAGE DISCLOSURE

Generative AI tools assisted manuscript editing and polishing, literature discovery, and reference formatting. The authors remain responsible for the final manuscript.

REFERENCES

Bilge Acun, Benjamin Lee, Fiodar Kazhamiaka, Kiwan Maeng, Udit Gupta, Manoj Chakkaravarthy, David Brooks, and Carole-Jean Wu. Carbon Explorer: A holistic framework for designing carbon aware datacenters. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, pp. 118–132, 2023. doi: 10.1145/3575693.3575754. URL https://doi.org/10.1145/3575693.3575754.

Guillaume Aupy, Yves Robert, Frédéric Vivien, and Dounia Zaidouni. Checkpointing algorithms and fault prediction. *Journal of Parallel and Distributed Computing*, 74(2):2048–2064, 2014. doi: 10.1016/j.jpdc.2013.10.010.

Lori Bird, Debra Lew, Michael Milligan, E. Maria Carlini, Ana Estanqueiro, Damian Flynn, Emilio Gomez-Lazaro, Hannele Holttinen, Nickie Menemenlis, Antje Orths, Peter Børre Eriksen, J. Charles Smith, Lennart Soder, Poul Sorensen, Argyrios Altiparmakis, Yoh Yasuda, and John Miller. Wind and solar energy curtailment: A review of international experience. *Renewable and Sustainable Energy Reviews*, 65:577–586, 2016. doi: 10.1016/j.rser.2016.06.082.

Jae-Won Chung, Zhirui Liang, Yanyong Mao, Jiasi Chen, Mosharaf Chowdhury, and Vladimir Dvorkin. OpenG2G: A simulation platform for AI datacenter-grid runtime coordination. *arXiv preprint arXiv:2605.05519*, 2026. doi: 10.48550/arXiv.2605.05519. URL https://arxiv.org/abs/2605.05519.

John T. Daly. A higher order estimate of the optimum checkpoint interval for restart dumps. *Future Generation Computer Systems*, 22(3):303–312, 2006. doi: 10.1016/j.future.2004.11.016.

Paul Denholm, Matthew O'Connell, Gregory Brinkman, and Jennie Jorgenson. Overgeneration from solar energy in California: A field guide to the duck chart. Technical Report NREL/TP-6A20-65023, National Renewable Energy Laboratory, 2015.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. doi: 10.48550/arXiv.2407.21783. URL https://arxiv.org/abs/2407.21783v3.

William W. Hogan. Contract networks for electric power transmission. *Journal of Regulatory Economics*, 4(3):211–242, 1992. doi: 10.1007/BF00133621.

Ziheng Jiang, Haibin Lin, Yinmin Zhong, Qi Huang, Yangrui Chen, Zhi Zhang, Yanghua Peng, Xiang Li, Cong Xie, Shibiao Nong, et al. MegaScale: Scaling large language model training to more than 10,000 GPUs. In *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, pp. 745–760. USENIX Association, 2024. URL https://www.usenix.org/conference/nsdi24/presentation/jiang-ziheng.

Jayashree Mohan, Amar Phanishayee, and Vijay Chidambaram. CheckFreq: Frequent, fine-grained DNN checkpointing. In *19th USENIX Conference on File and Storage Technologies (FAST)*, pp. 203–216. USENIX Association, 2021. URL https://www.usenix.org/conference/fast21/presentation/mohan.

Adam Moody, Greg Bronevetsky, Kathryn Mohror, and Bronis R. de Supinski. Design, modeling, and evaluation of a scalable multi-level checkpointing system. In *2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, pp. 1–11, 2010. doi: 10.1109/SC.2010.18.

Marco Nicolosi. Wind power integration and power system flexibility: An empirical analysis of extreme events in Germany under the new negative price regime. *Energy Policy*, 38(11):7257–7268, 2010. doi: 10.1016/j.enpol.2010.08.002.

Ana Radovanović, Ross Koningstein, Ian Schneider, Bokan Chen, Alexandre Duarte, Binz Roy, Diyue Xiao, Maya Haridasan, Patrick Hung, Nick Care, Saurav Talukdar, Eric Mullen, Kendal Smith, MariEllen Cottman, and Walfredo Cirne. Carbon-aware computing for datacenters. *IEEE Transactions on Power Systems*, 38(2):1270–1280, 2023. doi: 10.1109/TPWRS.2022.3173250.

Bianca Schroeder and Garth A. Gibson. A large-scale study of failures in high-performance computing systems. In *International Conference on Dependable Systems and Networks (DSN)*, pp. 249–258, 2006. doi: 10.1109/DSN.2006.5. URL https://www.cs.toronto.edu/~bianca/papers/dsn06.pdf.

Fred C. Schweppe, Michael C. Caramanis, Richard D. Tabors, and Roger E. Bohn. *Spot Pricing of Electricity*. Kluwer Academic Publishers, 1988.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-LM: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

John Thorpe, Pengzhan Zhao, Jonathan Eyolfson, Yifan Qiao, Zhihao Jia, Minjia Zhang, Ravi Netravali, and Guoqing Harry Xu. Bamboo: Making preemptible instances resilient for affordable training of large DNNs. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, pp. 497–513. USENIX Association, 2023. URL https://www.usenix.org/conference/nsdi23/presentation/thorpe.

Zhuang Wang, Zhen Jia, Shuai Zheng, Zhen Zhang, Xinwei Fu, T. S. Eugene Ng, and Yida Wang. Gemini: Fast failure recovery in distributed training with in-memory checkpoints. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, pp. 364–381. ACM, 2023. doi: 10.1145/3600006.3613145. URL https://www.cs.rice.edu/~eugeneng/papers/SOSP23.pdf.

Jiali Xing, Bilge Acun, Aditya Sundarrajan, David Brooks, Manoj Chakkaravarthy, Nikky Avila, Carole-Jean Wu, and Benjamin C. Lee. Carbon Responder: Coordinating demand response for the datacenter fleet. *arXiv preprint arXiv:2311.08589*, 2023. doi: 10.48550/arXiv.2311.08589. URL https://arxiv.org/abs/2311.08589.

John W. Young. A first order approximation to the optimum checkpoint interval. *Communications of the ACM*, 17(9):530–531, 1974. doi: 10.1145/361147.361115.