# AI HPC Data Centers and Demand Response: Technical Reference Guide

AI HPC data centers can provide meaningful grid flexibility, but with significantly different characteristics than Bitcoin mining. **Key finding: 10-25% of total facility load is typically flexible for demand response**, with response times ranging from milliseconds (GPU DVFS) to minutes (workload migration). Pre-training workloads offer the greatest flexibility potential at 20-40% interruptibility, while real-time inference remains nearly inflexible at 0-10%. The economic calculus differs fundamentally from Bitcoin mining—compute value typically exceeds DR revenue by 10:1 or more, making AI operators price-insensitive to electricity costs.

---

## Load shedding capability: MW/min specifications and ramp rates

AI HPC data centers can shed load across multiple time scales, from hardware-level power capping in milliseconds to facility-wide workload reduction over minutes.

### GPU-level response times

NVIDIA GPU Dynamic Voltage and Frequency Scaling (DVFS) operates at remarkably fast timescales. **Hardware DVFS can adjust frequencies within nanoseconds**, (washington) though software-controlled frequency changes through nvidia-smi operate on longer cycles. The GreenLLM DVFS controller demonstrates runtime frequency adjustment every **20 milliseconds** on A100 GPUs, achieving up to **34% energy reduction** with less than 3.5% SLO violations. (arXiv)

Power measurement granularity creates a practical constraint: nvidia-smi samples power values only **25ms out of every 100ms** on H100/A100 GPUs, meaning operators see only 25% of actual power behavior. AMD GPUs can switch clock frequencies within a few milliseconds, while NVIDIA requires approximately **100ms for software-initiated frequency transitions**. (Pado Pado)

### Achievable power reduction percentages

| Control Method | Power Reduction | Response Time | Notes |
|---|---|---|---|
| GPU DVFS (frequency scaling) | 19-34% | 20-100ms | Less than 4% performance loss |
| GPU power capping | 20-30% | Sub-second | Enforced by driver |
| SLURM power balancing | 10-30% | 30-60 seconds | Cluster-wide redistribution |
| Workload migration | 25-40% | 5-30 minutes | Requires checkpoint |
| Cooling setpoint increase | 10-15% | Minutes to hours | Per 2-3°C increase |
| Full job preemption | Up to 95% | Minutes | Training jobs only |

Lawrence Berkeley National Laboratory field studies demonstrated **25% demand savings at the data center level** and **10-12% savings at whole building level** through coordinated workload shifting and cooling adjustment. (Lbl) Texas ERCOT projections indicate that by 2030, up to **50% of the expected 35 GW** data center capacity could provide emergency reliability support under the right conditions.

### Maximum instantaneous load shed

For emergency response, AI data centers with proper automation can achieve near-complete load reduction similar to Bitcoin miners, but with significant caveats. The practical maximum depends on workload composition—facilities running primarily batch training can shed **80-95%** of IT load with checkpoint overhead, while inference-heavy facilities may be limited to **10-25%** reduction without SLA violations.

---

## Advance notice requirements for demand response participation

The advance notice required varies dramatically by DR product type, creating a natural segmentation of which products AI data centers can realistically participate in.

### Grid operator DR product requirements

| DR Product | Response Time | AI Data Center Capability | Gap Assessment |
|---|---|---|---|
| **Frequency Regulation** | 2-4 seconds | Limited (UPS only) | **Significant gap**—incompatible with IT load adjustment |
| **Spinning Reserve** | 10 minutes | Moderate | **Moderate gap**—requires automation, telemetry |
| **Non-Spinning Reserve** | 30 minutes | Good | **Small gap**—sufficient for workload migration |
| **Economic DR** | Day-ahead/Hour-ahead | Excellent | **No gap**—aligns with scheduling capabilities |
| **Emergency DR** | Hours ahead | Good | **Small gap**—island mode feasible |

### ISO/RTO specific requirements

**ERCOT Emergency Response Service (ERS)** requires 10-minute (ERS-10) or 30-minute (ERS-30) response (ERCOT) (PCI Energy Solutions) with **2-second telemetry** and under-frequency relay at 59.7 Hz. (Cpowerenergy)

Minimum curtailable load is 100 kW (Bridgevue Energy) with 95% availability threshold. (Enel North America) Maximum deployment is **24 cumulative hours per contract period**. (Legal Information Institute)

**PJM** operates synchronized reserve with 10-minute response, day-ahead scheduling reserve with 30-minute response, and regulation following 4-second AGC signals. Economic Load Response Participants must register with a Curtailment Service Provider.

**CAISO Proxy Demand Resource (PDR)** has Must Offer Obligation requiring bids whenever capacity is available, with minimum **4 hours per dispatch, 24 hours per month**. Telemetry is required for ancillary services participation.

### What AI operators actually need

Google's demand response approach reveals practical requirements: their algorithm generates **hour-by-hour instructions** based on forecasted grid events, shifting non-urgent compute tasks during peak periods. Critical services (Search, Maps, healthcare cloud) remain protected. This day-ahead to hour-ahead notice requirement aligns well with economic DR products but excludes fast-response ancillary services.

The DCFlex Initiative pilot in Phoenix achieved **10-40% flexibility** with hour-ahead coordination, (ieee) (IEEE Spectrum) demonstrating that meaningful load reduction is achievable with proper advance notice. (IEEE Spectrum) However, even Google acknowledges that "high reliability requirements for critical services limit how much flexibility any data center can offer." (Google)

---

## Demand response capability differences by AI workload type

### Pre-training workloads (large-scale model training)

Pre-training represents the **most flexible AI workload category** for demand response, though significant constraints remain.

**Checkpoint characteristics:**

- **Frequency:** Every 10-100 iterations (infrequent to reduce overhead)
- **Size:** 100 GB to 2+ TB for GPT-4 class models
- **Overhead:** 5-43% of training time depending on storage performance
- **Modern solutions:** Asynchronous checkpointing reduces overhead to under 5% (arXiv)

**Interruption costs are substantial.** For a 256-instance P5 cluster at $55/hour, each disruption costs approximately **$4,693** (10 minutes lost work plus 10 minutes recovery). A month-long training run with daily disruptions adds **$141,000** in costs and 10+ hours delay.

**Gang scheduling creates all-or-nothing constraints.** Tensor parallelism accounts for over 75% of inter-GPU bytes transferred, (Frontier Models) requiring all GPUs to run simultaneously. Partial curtailment is impractical—either checkpoint and stop completely, or continue running.

**Flexibility rating: 20-40% interruptible** with proper orchestration and checkpointing infrastructure.

### Fine-tuning / post-training workloads

Fine-tuning offers **higher flexibility than pre-training** due to smaller scale and shorter duration.

- **GPU requirements:** 1-4 GPUs for 7B models versus thousands for pre-training
- **Duration:** Hours to days versus weeks to months
- **Checkpoint size:** Much smaller with techniques like LoRA/QLoRA (10x memory reduction)
- **Already runs on spot instances:** Many organizations fine-tune on preemptible capacity

**Flexibility rating: 40-60% interruptible** with minimal economic penalty.

### Inference workloads (serving predictions)

Inference flexibility depends entirely on latency requirements and presents a bimodal distribution.

**Latency-sensitive inference (0-10% flexible):**

- Real-time chat/API: P99 under 300ms required
- Banking/transactions: P99 under 100ms
- Voice assistants: Sub-second response mandatory
- 53% of users abandon applications if load time exceeds 3 seconds (ControlPlane)

**Batch inference (80-100% flexible):**

- Offline analytics
- Video processing (YouTube encoding)
- Document classification
- Non-real-time recommendations

**Microsoft Azure OpenAI offers 99.9% latency SLA** for token generation, (Neowin) indicating the industry prioritizes reliability over flexibility. Cold start penalties (minutes to load large models) exceed typical preemption warning windows (30 seconds to 2 minutes), making dynamic scaling impractical for latency-sensitive workloads. (arXiv)

---

## Comparison: AI HPC versus Bitcoin mining demand response

Bitcoin mining represents the gold standard for demand response loads, and AI HPC falls significantly short on nearly every dimension.

### Head-to-head comparison matrix

| Dimension | Bitcoin Mining | AI Training | AI Inference |
|---|---|---|---|
| **Shutdown time** | <15 seconds | 5-30+ minutes | Seconds (with SLA impact) |
| **Startup time** | Minutes | 15-30+ minutes | Seconds |
| **State preservation** | Not required | Required (checkpoint) | Not required |
| **Checkpoint size** | N/A | 100 GB - 2 TB | N/A |
| **Checkpoint overhead** | 0% | 5-43% of training time | N/A |
| **Economic sensitivity** | High (curtail >$70-85/MWh) | Low | Low |
| **SLA constraints** | None | Internal deadlines | Strict (ms-level) |
| **Interruptibility rating** | ★★★★★ | ★★☆☆☆ | ★★☆☆☆ |

### Why Bitcoin mining is "ideal" for demand response

A November 2023 paper co-authored by former ERCOT CEO Brad Jones identifies the optimal DR load characteristics: **instantaneous reactivity, price signal responsiveness, continuous curtailment capacity, minimal opportunity cost, high granularity, and significant MWh scale.** Bitcoin mining achieves all six; AI achieves only scale.

**Riot Platforms earned $31.7 million in demand response credits in August 2023**—nearly 4x their $8.6 million from selling mined Bitcoin that month. This economic structure creates strong incentives to curtail. By contrast, AI operators are **relatively insensitive to electricity prices** because highly lucrative workloads justify premium power costs.

### Economic incentive differences

| Factor | Bitcoin Mining | AI Training |
|---|---|---|
| **Value per hour of operation** | $50-100/MWh equivalent | Training: $100M+ total cost |
| **Price threshold to curtail** | $70-85/MWh | Rarely economically justified |

| Factor | Bitcoin Mining | AI Training |
|---|---|---|
| **DR payment vs. operational revenue** | Often DR > mining during peaks | DR << training value |
| **Decision complexity** | Simple ROI calculation | Multi-factor optimization |

### Control system architecture comparison

**Bitcoin mining** uses dedicated DR software (Lancium Smart Response, Foreman, Braiins) that directly integrates with grid operators and provides **automated curtailment within 15 seconds** based on price signals. Approximately 750 MW of Texas miners hold CLR (Controllable Load Resource) designation with 2,600 MW awaiting approval.

**AI HPC** relies on cluster schedulers (SLURM, Kubernetes, Run:ai) designed for job management rather than grid response. No standardized protocol exists for data center DR participation, and most facilities require manual coordination with utilities.

---

## Algorithmic and electrical control mechanisms

This section focuses exclusively on software, algorithmic, and electrical controls—not behind-the-meter generation or batteries.

### GPU power management APIs

**NVIDIA NVML (Management Library)** provides programmatic power control:

| Function | Purpose |
|---|---|
| nvmlDeviceSetPowerManagementLimit() | Set GPU power cap in watts |
| nvmlDeviceGetPowerUsage() | Read current power draw |
| nvmlDeviceSetApplicationsClocks() | Pin GPU clocks to specific frequency |
| nvmlDeviceGetEnforcedPowerLimit() | Query actual enforced limit |

**Power control hierarchy:** VBIOS defines maximum possible TGP → nvidia-smi sets user limit via host → SMBPBI provides out-of-band control via BMC.

**NVIDIA Data Center GPU Manager (DCGM)** enables cluster-wide GPU orchestration, health monitoring, and power management with polling precision suitable for **millisecond-level kernel execution**. Nebius

### GPU power limit specifications

| GPU Model | TDP | Configurable Range | Notes |
|---|---|---|---|
| H100 SXM5 | 700W | Configurable | DGX uses 6 PSUs for 4+2 redundancy |
| H100 PCIe | 350W | 200-350W | Standard configuration |
| H100 NVL | 400W | Per documentation | Liquid-cooled variant |
| A100 SXM4 | 400W | Up to 500W with CTS | Custom thermal solution |
| A100 PCIe | 250-300W | SKU dependent | |

### SLURM power management configuration

SLURM's power management plugin enables facility-wide power capping:

```
PowerParameters:
  balance_interval=60      # Seconds between cap rebalancing
  cap_watts=1800000        # Total facility power limit
  decrease_rate=30         # Reduce cap by 30% when under threshold
  increase_rate=10         # Increase cap by 10% when approaching limit
  lower_threshold=90       # Begin reducing at 90% utilization
  upper_threshold=98       # Hard cap at 98%
```

Jobs can request specific CPU frequencies: `--cpu-freq=2400000-3000000:ondemand`. Energy accounting adds less than 0.6% overhead in consumption and 0.2% in execution time.

### Kubernetes GPU power management

**NVIDIA GPU Operator** enables:

- Time-slicing for container GPU sharing
- Multi-Instance GPU (MIG) partitioning on A100/H100 (up to 7 isolated instances)
- Health monitoring to prevent scheduling on faulty devices

**Power-aware scheduling plugins** minimize power consumption increase per task assignment, achieving **20%+ power savings** versus baseline schedulers in testing with Alibaba GPU traces.

### Dynamic voltage and frequency scaling (DVFS) capabilities

DVFS achieves **19-34% energy reduction** with less than 4% performance impact across multiple studies:

| Study | Workload | Energy Savings | Performance Impact |
|---|---|---|---|
| ACM HotPower 2013 | General GPU | 19.28% | ≤4% |
| ACM e-Energy 2019 | DNN training | 8.7-23.1% | Minimal |
| ACM e-Energy 2019 | DNN inference | 19.6-26.4% | Varies |
| Springer ISC 2024 | Mixed precision | 6-45% | Application dependent |
| UW IEEE CAL 2023 | Data center | 20% peak power | <1% performance |

**Key limitation:** A100 memory clock cannot be independently adjusted due to HBM2 architecture. MIG partitions share clock domains, preventing individual partition power optimization.

---

## Flexible load portion by category

### IT load breakdown (50-60% of total facility power)

| Component | % of IT Load | Flexibility | Notes |
|---|---|---|---|
| Training workloads | 20-35% | 20-40% | With checkpointing |
| Batch inference | 10-20% | 80-100% | Highly deferrable |
| Real-time inference | 30-50% | 0-10% | SLA-bound |
| Fine-tuning | 5-15% | 40-60% | Medium flexibility |
| Networking | 5-8% | 0-5% | Generally fixed |
| Storage | 5-10% | 10-20% | HDD spin-down possible |

**Overall IT flexibility: 15-30%** of IT load can be interrupted or deferred with proper orchestration.

### Cooling load breakdown (20-40% of total facility power)

Cooling flexibility depends on thermal mass and time constants:

| Component | Thermal Time Constant | Flexibility Strategy |
|---|---|---|
| Building thermal mass | 15-60 minutes | Raise setpoint 2-3°C for 10-15% savings |
| Chilled water systems | 5-15 minutes | Chiller cycling for 20-30% reduction |
| CRAH/CRAC fans | Minutes to hours | Speed reduction for 5-15% savings |
| Server thermal inertia | Seconds to minutes | Limited buffer |

**Cooling flexibility: 15-30%** sustainable for 15-60 minute windows.

Facility PUE determines cooling as percentage of total power:

- **Hyperscale (best):** PUE 1.06-1.12, cooling 6-12% of total
- **Modern efficient:** PUE 1.15-1.25, cooling 15-25%
- **Average colocation:** PUE 1.3-1.6, cooling 30-40%

### Auxiliary loads (5-15% of total)

| Component | % of Total | Flexibility |
|---|---|---|
| UPS losses | 3-8% | 0-5% |
| PDU losses | 2-4% | 0% (fixed) |
| Lighting | 1-3% | 70-90% |
| Security | 0.5-1% | 0% (critical) |
| Fire suppression | <0.5% | 0% (critical) |

**Auxiliary flexibility: ~1-3%** of total facility power.

### Aggregate flexibility by time horizon

| Time Horizon | Flexible % of Total Load | Primary Sources |
|---|---|---|
| **5 minutes** | 5-15% | Training pause, cooling thermal mass |
| **15 minutes** | 10-25% | Above + chiller cycling, batch defer |

| Time Horizon | Flexible % of Total Load | Primary Sources |
|---|---|---|
| **1 hour** | 15-35% | Above + workload migration |
| **4 hours** | 20-40% | Full training deferrals |
| **24+ hours** | 30-50% | Geographic shift, batch rescheduling |

## Technical barriers and considerations

### Checkpoint overhead: The dominant constraint

Checkpointing large AI models imposes significant time and cost burdens. For GPT-4 scale models, checkpoint sizes reach **100 GB to 2+ TB**, requiring 15-30 minutes to save with traditional storage systems. Microsoft Learn Modern solutions reduce this substantially:

- **Google multi-tier checkpointing:** Saves scale to under 5 minutes, restore in under 1 minute Google Cloud
- **AWS SageMaker checkpointless training:** 80-93% reduction in recovery time AWS
- **Microsoft Nebula:** 95-99% reduction (hours to seconds) Microsoft Learn Microsoft Learn

Without these optimizations, checkpointing overhead consumes **12% average** and up to **43%** of total training time. Microsoft Learn

### Gang scheduling creates all-or-nothing constraints

Distributed training requires synchronized execution across all GPUs. NCCL collective operations (AllReduce) have a **10-minute default timeout**—any rank that fails brings the entire job down. PyTorch The new NCCL 2.27 ncclCommShrink feature allows dynamic GPU exclusion, but recovery still requires collective reinitialization across surviving ranks. NVIDIA Developer

### Economic barriers outweigh DR revenue

| Cost Category | Impact |
|---|---|
| Lost compute per interruption | $4,693 (256 P5 instances, 20 min total) |
| Month-long training with daily disruptions | +$141,000, +10 hours |
| 1 hour curtailment, 100 MW H100 cluster | ~$500K-1M lost compute |
| DR revenue potential | $5-50/MWh |

The **10:1 or greater ratio** of compute value to DR revenue makes participation economically challenging without policy mandates.

### Integration gaps

Most data centers lack:

- Automated systems to receive grid operator signals
- Compatible telemetry/metering infrastructure (2-6 second intervals for ancillary services)
- Standardized protocols for DR participation
- Software integration between cluster schedulers and utility systems

---

## Industry examples and case studies

### Google: Most advanced DR participant

Google leads hyperscaler DR participation with documented deployments:

- **Europe (Winter 2022-23):** Daily power reductions during peak periods (5pm-9pm)
- **Taiwan (2022-23 summers):** Daily peak hour reductions with Taiwan Power Company
- **Oregon, Nebraska, Southeast US:** Response to extreme weather events
- **Fort Wayne, Indiana:** New agreement with Indiana Michigan Power
- **DCFlex pilot (Lenoir, NC):** Testing workload choreography with Duke Energy

Google's Carbon-Intelligent Computing Platform generates **hour-by-hour instructions** automatically, limiting non-urgent tasks while protecting Search, Maps, and Cloud services.

### DCFlex Initiative: Industry collaboration

The EPRI-led DCFlex Initiative represents the most promising path forward with **45 collaborators** including Google, Meta, Microsoft, Nvidia, and Oracle. Three pilot sites announced in June 2025:

1. **Lenoir, NC** (Google/Duke Energy): Workload choreography testing
2. **Phoenix, AZ** (Nvidia/Oracle/Emerald AI/Salt River Project): **10-40% flexibility** achieved in simulated peak events
3. **Paris, France** (Data4/Schneider Electric/RTE): UPS systems for grid stability (ieee)

### Scale of the opportunity

Duke University research shows that even minimal flexibility delivers substantial grid benefits:

- **0.25% annual curtailment** (~87 hours/year partial load) enables 76 GW of new load
- **0.5% annual curtailment** (~100 hours partial) enables 98 GW new load (RMI)
- **1% annual curtailment** enables 126 GW new load (Deloitte Insights)

Average curtailment events last approximately **2 hours**, and 90% of events retain at least 50% of load—partial curtailment is the norm, not complete shutdown.

---

## Reference parameters for antigravity energy optimizer

### Workload flexibility parameters

```yaml
```

```
PRE_TRAINING:
  interruption_cost_hours: 0.5-4
  checkpoint_overhead_pct: 2-10%
  restart_time_minutes: 5-30
  min_run_duration_hours: 2-4
  flexibility_pct: 20-40%
  gang_scheduling: true
  economic_threshold_usd_per_mwh: rarely_economically_justified

FINE_TUNING:
  interruption_cost_hours: 0.1-1
  checkpoint_overhead_pct: 1-5%
  restart_time_minutes: 1-10
  min_run_duration_hours: 0.5-2
  flexibility_pct: 40-60%
  gang_scheduling: partial

BATCH_INFERENCE:
  interruption_cost_hours: 0
  restart_time_minutes: <1
  flexibility_pct: 80-100%
  deferrable_hours: 4-24+

REALTIME_INFERENCE:
  interruption_cost: SLA_VIOLATION
  flexibility_pct: 0-10%
  latency_sla_ms: 100-300 (P99)
```

### Facility flexibility parameters

yaml

```yaml
COOLING:
  thermal_time_constant_minutes: 15-60
  setpoint_increase_potential_C: 2-5
  power_reduction_per_degree_pct: 3-5%
  chiller_cycling_duration_min: 15-30
  max_reduction_pct: 20-30%

IT_LOAD:
  flexible_portion_pct: 15-30%
  response_time_minutes: 1-15
  gpu_dvfs_response_ms: 20-100
  slurm_balance_interval_sec: 30-60

TOTAL_FACILITY:
  typical_flexible_pct: 10-25%
  peak_flexible_pct: 20-40%
  annual_curtailment_hours: 50-200
  avg_event_duration_hours: 2-3
```

### Power breakdown (PUE 1.2 facility)

```yaml
TOTAL_FACILITY_POWER: 100%
IT_LOAD: 83%  # (1/PUE)
  GPU_TPU: 60-70%
  CPU: 10-15%
  NETWORKING: 5-8%
  STORAGE: 5-10%
COOLING: 12-15%
  CHILLERS: 6-10%
  CRAH_AHU: 4-6%
  PUMPS: 1-2%
OTHER: 3-5%
  UPS_LOSSES: 2-3%
  PDU_LOSSES: 0.5-1%
  LIGHTING_MISC: 0.5-1%
```

### DR product compatibility matrix

```yaml
```

```
FREQUENCY_REGULATION:
  response_time: 2-4_seconds
  ai_datacenter_compatible: false
  pathway: UPS_only

SPINNING_RESERVE:
  response_time: 10_minutes
  ai_datacenter_compatible: moderate
  requires: automation_telemetry

NON_SPINNING_RESERVE:
  response_time: 30_minutes
  ai_datacenter_compatible: good
  enables: workload_migration

ECONOMIC_DR:
  response_time: day_ahead_or_hour_ahead
  ai_datacenter_compatible: excellent
  best_fit: training_workloads

EMERGENCY_DR:
  response_time: hours_ahead
  ai_datacenter_compatible: good
  mechanism: island_mode_or_full_curtailment
```

---

## Conclusions: Key insights for implementation

AI HPC data centers represent a fundamentally different demand response resource than Bitcoin mining. While the scale potential is comparable—hundreds of megawatts to gigawatts—the operational constraints create a more complex optimization problem.

**The 10-25% flexibility range is achievable** for most facilities with proper workload segmentation. Batch training and non-real-time inference provide the primary flexible capacity, while production inference systems remain protected. This aligns naturally with day-ahead and hour-ahead economic DR products rather than fast-response ancillary services.

**Checkpoint technology is advancing rapidly.** Google, AWS, and Microsoft investments in multi-tier and asynchronous checkpointing are reducing the dominant technical barrier. Facilities implementing these solutions can achieve meaningful flexibility with substantially lower overhead than legacy approaches.

**Economic incentives require policy intervention.** Unlike Bitcoin mining where DR payments often exceed operational revenue during peaks, AI compute value typically exceeds DR revenue by 10:1 or more. Texas SB6 mandating curtailment for loads exceeding 75 MW represents the emerging regulatory approach to ensure data centers contribute to grid reliability.

**The DCFlex Initiative provides a roadmap.** With 45 collaborators and 10 pilot sites planned, industry-standard practices for AI data center flexibility are emerging. The Phoenix pilot's demonstration of 10-40% flexibility with coordinated workload choreography establishes a practical benchmark for what optimized facilities can achieve. IEEE Spectrum