

Data Center Briefing

February 03, 2026

Global

Key themes:

Enterprise AI supply chain security and offline deployment; Large swings in inference energy from serving choices; Data centre load flexibility for grid frequency regulation; GPU serving throughput gains via concurrency control; Autoscaling ML pipelines with LLM-based controllers; CXL memory compression/precision scaling for long-context inference; Cross-vendor GPU clustering via heterogeneous collectives; Custom congestion control using LLMs and BPF

Global data centres & AI infrastructure briefing (UTC 2026-02-03)

Audience: Institutional asset managers and infrastructure fund managers focused on data centres, power, and grid infrastructure.

Top news (3)

1. **Hardening the enterprise AI supply chain + enabling offline deployment:** [Dell Enterprise Hub strengthens AI supply chain security and deployment](#) highlights multi-layer image integrity and scanning (daily Amazon Inspector scans, Cosign-signed images, SHA-384 manifest hashes, and model malware/pickle scanning) alongside support for fully offline, on-prem deployments.
2. **Inference energy and cost can swing by orders of magnitude on the same GPUs:** [LLM Inference Efficiency: Quantization, Batching, and Serving Strategies](#) reports that system-level serving choices (quantization, batching, configuration) materially change energy use on NVIDIA H100s; it also reports arrival shaping can reduce per-request energy by up to **100x**.

3. **Data centres as grid assets (frequency regulation + emissions lens):** [Data center load flexibility for grid frequency regulation and emissions](#) introduces an “Exogenous Carbon” metric and the “EcoCenter” framework to coordinate GPU data centre load flexibility for frequency regulation, with the authors reporting potential savings that can outweigh operational emissions.
-

Key deals and projects

No M&A, financing, lease, land, or build project announcements were included in today's story set.

Technology and operations (what could matter for capex, utilisation, and performance)

AI platform security, verification, and enterprise deployment

- [Dell Enterprise Hub strengthens AI supply chain security and deployment](#)
 - **Security controls described:** daily Amazon Inspector scans; Cosign-signed images; SHA-384 manifest hashes; model malware/pickle scanning.
 - **Enterprise verification:** integrity and scan results are published with images (noted as being available to verify).
 - **Deployment angle:** enables **fully offline on-prem deployments**, which can be relevant where air-gapped or regulated environments constrain cloud usage.

Serving efficiency, autoscaling, and concurrency control (GPU utilisation and tail latency)

- [CONCUR: Congestion-Based Concurrency Control for Agentic LLM Inference](#)
 - Introduces an admission/concurrency control layer to avoid GPU KV-cache “middle-phase thrashing” in agentic batch inference.
 - Reported throughput improvements: **up to 4.09x** (Qwen3-32B) and **1.9x** (DeepSeek-V3), while remaining compatible with existing serving systems.
- [SAIR: Cost-Efficient Multi-Stage ML Pipeline Autoscaling via LLM](#)
 - Uses an LLM as an in-context RL controller for autoscaling multi-stage inference pipelines.

- Reported outcomes (under stated assumptions): up to **50%** P99 latency improvement and up to **97%** effective cost reduction; **86%** bottleneck detection accuracy.
- [Serverless GPU Architecture for Enterprise HR Analytics BDaaS Implementation](#)
 - Presents a Helm-packaged BDaaS blueprint integrating a single-node serverless GPU runtime with TabNet for compliance-aware HR analytics.
 - Benchmarks reported vs Spark baselines: up to **4.5×** higher throughput, **98×** lower latency, and **~90%** lower cost per 1K inferences; compliance overhead **~5.7 ms** (p99 < **22 ms**).

Memory and interconnect efficiency (CXL and model serving at long context)

- [TRACE: Lossless Compression and Precision Scaling for CXL Bandwidth](#)
 - Proposes device-internal transforms enabling lossless compression and precision-proportional fetch for CXL memory.
 - Reported footprint reductions: BF16 weights **25.2%**, BF16 KV **46.9%**.
 - System modeling claim: throughput for GPT-OSS-120B-MXFP4 at 128k tokens increases from **16.28** to **68.99 tok/s** (**4.24×**).
 - Overheads reported: area **7.2%**, power **4.7%**, latency **6.0%**.

Heterogeneous GPU clusters (multi-vendor flexibility)

- [HetCCL Accelerates LLM Training with Heterogeneous GPU Clusters](#)
 - Presents a collective communication library unifying vendor-specific backends to enable RDMA-based cross-vendor GPU communication without driver changes.
 - Evaluation summary: matches NVIDIA NCCL and AMD RCCL performance in homogeneous setups and scales in heterogeneous environments.

Networking controls and custom congestion control

- [NECC: Enabling Non-Expert Customized Congestion Control with LLMs](#)
 - Uses LLMs plus the Berkeley Packet Filter (BPF) interface to model, implement, and deploy custom congestion control algorithms.
 - Paper noted as an accepted manuscript for IEEE ICC 2025 with “promising” results using real-world CCAs.

Power, grid, and interconnection highlights

- [Data center load flexibility for grid frequency regulation and emissions](#)
 - Positions GPU data centres as flexible loads to help provide **grid frequency regulation**.

- Introduces the **Exogenous Carbon** metric and **EcoCenter** coordination framework.
- Reported finding: exogenous carbon savings can often outweigh operational carbon emissions (per the preprint's results summary).

Policy and regulation

No policy, permitting, tariff, interconnection-queue, or regulatory changes were included in today's story set.

2-line wrap

Today's flow was dominated by AI infrastructure enablement: securing model supply chains, raising GPU utilisation, and reducing memory/interconnect bottlenecks.

Separately, load-flexibility research continues to frame GPU data centres as controllable grid resources for frequency regulation and emissions outcomes.

TELBO ● RG