Resiliency Requirements for AI training Deployments

January 7, 2025

North America Roundtable Discussion

Readout produced by Michael O'Neil, Contributing Analyst

uptime INSTITUTE

Smarter Together

The Uptime Network is a community of data center owners and operators under mutual NDA. No member organizations or individuals are named.

This readout captures discussions between Uptime Network members, Uptime Intelligence Senior Research Director Owen Rogers and Uptime Vice President, Global Service Management Naveed Saeed in January 2025.

Based on an iterative, collaborative process between Network members and Uptime technical SMEs, the intent is to develop practical guidance, which will be shared with the Network when ready.

Email suggestions for future topics to Matt Stansberry: mstansberry@uptimeinstitute.com.

uptime® INSTITUTE

January 8

Meeting Notes

Uptime Network convened the North America roundtable, including senior operations managers from ten member organizations, representing finance, manufacturing, colocation and healthcare industries.

10 Member Organizations

Topics discussed:

Defining terms, processes and IT operational practices for AI training deployments.

Evaluating facility resiliency requirements for these workloads.

Unresolved questions for mechanical and electrical infrastructure suppliers.

uptime INSTITUTE

Business Drivers for Training AI Models

For the purposes of the discussion, AI training was defined as the process of teaching an artificial intelligence (AI) system, typically a machine learning model, to perform specific tasks by exposing it to a dataset and using algorithms to adjust its internal parameters to optimize performance.

Uptime shared the belief that many companies will use pre-trained systems, while others will opt to train their own models for customization, first-mover advantage, data security concerns, and/or economies of scale.

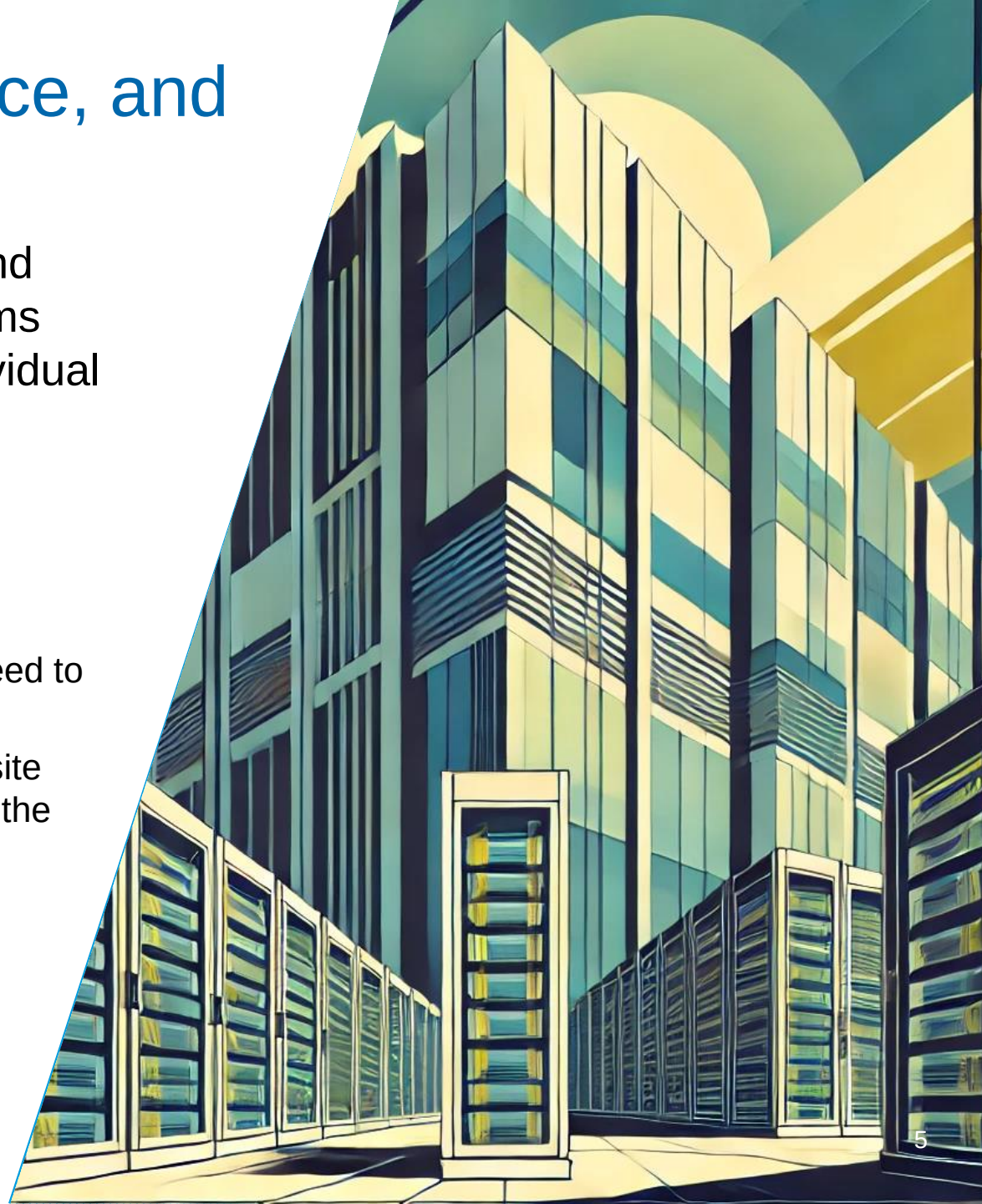
Uptime noted training a model can take weeks or months and is a one-off batch process, and is performed in advance of value generation, before AI models are embedded in enterprise business applications.

Deploying large-scale GPU clusters for generative AI training has high capital costs and will require specialized infrastructure and expertise, driving many enterprises to the cloud (Most AI models will be trained in the cloud, Dec 2024).

Previous discussions with Uptime Network members and the attendees on this roundtable suggested some organizations are planning to train proprietary AI instances in their own facilities for strategic reasons, and others' plans are still undetermined.

Balancing Resiliency, Performance, and Cost

- Uptime and Network members expect high cost and performance requirements for GPU training systems will not allow for redundancy or failover at the individual machine level.
- Resiliency at the data center facility layer may be critical for businesses where corporate strategy requires faster training results:
  - If an entire site goes down, the model can be lost and will need to start over.
  - If a site is not at least concurrently maintainable (Tier III), a site could experience weeks of downtime or reduced capacity in the event of a required maintenance shutdown.



uptime INSTITUTE

Trade-offs: Checkpoints and Model Stability

Uptime defined the following terms:

Epoch

A complete pass of the training dataset through a training cycle. Larger epochs are more efficient because they can ingest more data and identify more connections; smaller epochs can be restarted with minimal loss in the event of a crash/outage.

Checkpoint

A snapshot of results by writing in-progress models to disk. More frequent checkpoints avoid losses in the event of a crash/outage, but higher frequencies can drive erratic CPU and disk performance.

Network members and Uptime observed there is no framework for optimizing epoch size and checkpoint frequency. Firms looking to build or enhance resiliency through software controls will need to establish their own approaches to key parameters.

uptime® INSTITUTE

Challenges for Hosting AI Training in Existing Sites

Most facilities are designed for racks of up to ~15 kw; air cooled facilities with hot/cold aisle containment may accommodate up to ~20kw/rack.

Facilities with rear door heat exchangers can accommodate racks up to about 40kw. However, other constraints (mostly, power) will limit these racks to about 25% of floor space capacity: e.g., a data hall designed to hold 400 10kw racks will accommodate 100 40kw racks.

Other cooling strategies may increase rack density beyond 40kw, but the floor space utilization constraints remain.

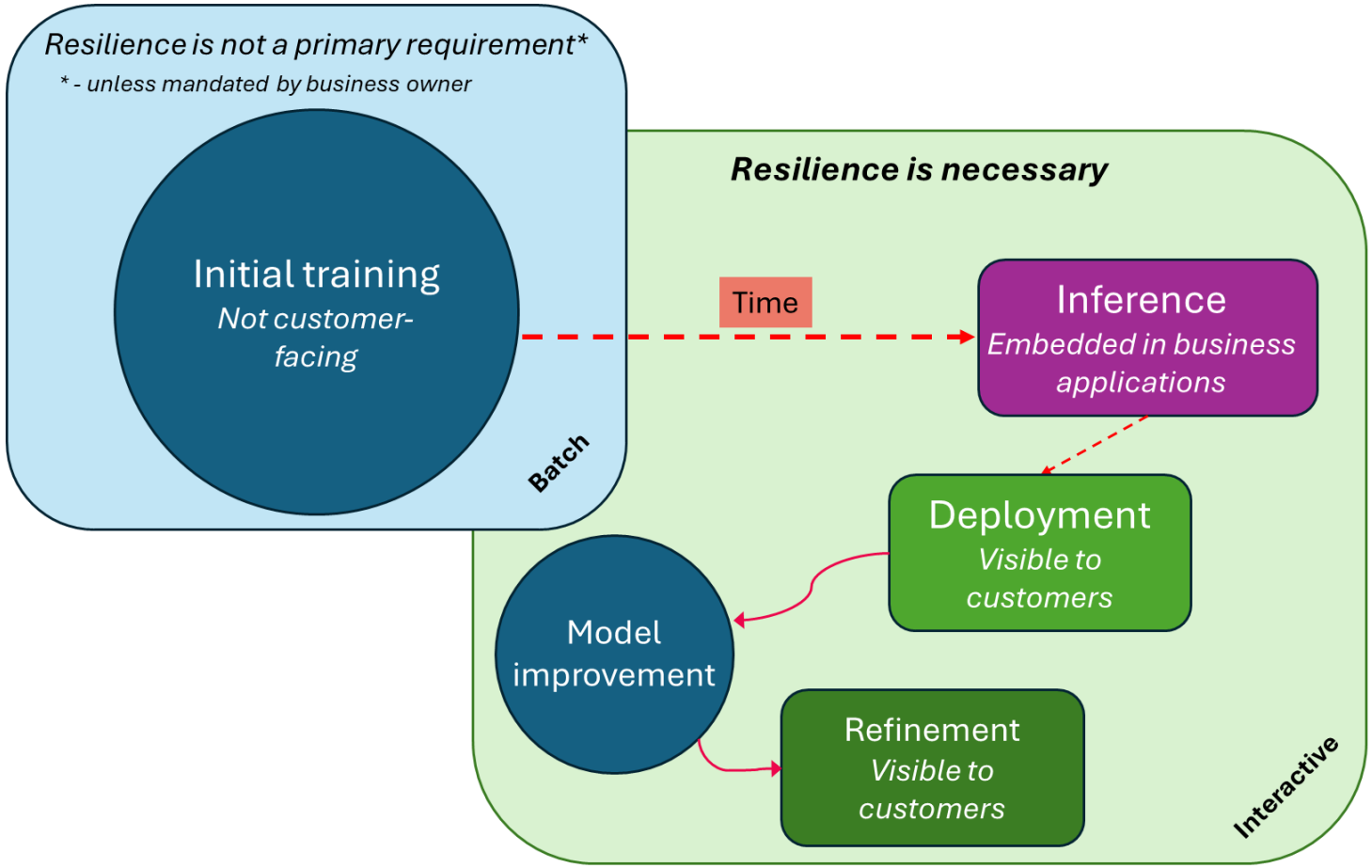
A member from a financial organization raised concerns about UPS suppliers' response to erratic power profiles in AI training clusters (see UI Note 387)

An operator from a colocation company observed that colocation providers are going to struggle to explain the massive change in white-space/gray-space ratios. The densification push is throwing data center designs off balance, with the same IT space requiring increasingly larger support infrastructure.

Future Technical and Business Imperatives

Uptime and Network members reached consensus that once the trained model is embedded in an inference engine and deployed to (internal and/or external) customers, resilience becomes more critical.

Inference engines may be deployed at the edge to support latency-sensitive applications, while training still occurs in central facilities. As training cycles compress to support model refinement, training may require more resiliency (see diagram).



uptime INSTITUTE

Key Challenges and Strategies

Deployment of AI-capable infrastructure is a real challenge.

- Members noted a need for significant power and cooling capacity, requirement for expensive server clusters that require high utilization and may not hold their value through the entire depreciation cycle.
  - **Possible strategies**: Use modular approaches for power, deploy incrementally, align cooling with specific zone requirements; mixed-tier facilities.

Correctly scoping the AI training environment is a significant challenge.

- Over-scoping results in unused – and very expensive – capacity.
- Under-scoping would require adding resources – which could result in a need to rearchitect power and cooling systems, increasing cost and extending payback timeframes.
  - **Possible strategy**: Investigate whether some (or all) of AI training requirements can be run in a specialized AI-as-a-Service environment such as CoreWeave or Lambda Labs, etc.

Lessons Learned / Recommendations

- Initial step: Determine whether your organization's AI training objectives are best met by on prem deployment vs. third party training (stay tuned for Network member opportunities to model sample deployments against Uptime's FORCSS framework).
- Anticipate perspective differences between IT operations and business units on resiliency requirements. Understand that epochs and checkpoints incur performance and delivery risk and establish a strategy that is acceptable to all essential stakeholders.
- High utilization rates will be necessary to justify AI infrastructure investments, so organizations will want to identify a 'long list' of beneficial AI training projects. Ensure your organization is tracking the depreciation challenges.
- Build a horizon that extends beyond initial training to inference engines embedded in applications, deployed into production, and refined through ongoing training. Resiliency requirements change as the AI workload becomes interactive and is exposed to customers; infrastructure plans need to align with these evolving needs.
- With respect to facility planning, recognize that in the near term, white space utilization ratios will be negatively impacted by density (fewer racks than planned), while in the longer term, the ratio of gray space to white space will increase "massively".
- Understand that inference workloads may run in many different environments, including at the edge, in IoT/OT devices, etc. Plan for secure network connections that are reliable and fast enough to meet availability and latency targets.

Take the Next Steps

- Tuesday, January 28 Owen Rogers and Naveed Saeed will host the next roundtable on this topic for Middle East/Europe (7am ET - 12pm GMT). The session is open to any member and registration is here.
- Discuss AI resiliency requirements with your peers in-person at the Uptime Network Americas Spring Conference: March 18-19, Wyndham Atlanta Buckhead Hotel & Conference Center. Registration is open now as tour spaces are limited.
- Further Reading:
  - AI to trigger radical overhaul of data center electrification
  - Why bigger is not better: gen AI models are shrinking
  - How generative AI learns and creates using GPUs
- Email Matt Stansberry to request a private briefing on this topic with Uptime Intelligence analysts and/or Uptime Technical Consultants: mstansberry@uptimeinstitute.com.

uptime INSTITUTE

uptime® INSTITUTE

Visit www.uptimeinstitute.com/ui-network for more information.

Uptime Institute
405 Lexington Avenue
New York, NY 10174