Assessment Criteria

Performance & Scalability

- Ability to handle large-scale AI workloads, including LLM training, fine-tuning, and high-throughput inference.
- GPU-to-GPU communication bandwidth and latency within the server.
- Memory capacity and bandwidth to support very large models.
- Overall compute throughput and suitability for scaling future workloads.

Infrastructure Fit & Reliability

- Compatibility with enterprise data center power, cooling, and rack standards.
- Efficiency of the proposed system in terms of power draw and cooling requirements.
- Server reliability and ability to minimize downtime during operation.
- Quality and integration of pre-installed networking, storage, and memory subsystems.
- The site currently operates under standard enterprise-class data center requirements (rack-mounted equipment, conditioned cooling, and high-speed network switching). The remote site does not support network switching above 25 Gbit; therefore, providers should avoid use OCP 3.0 or higher-speed adapters. Each rack provides a maximum of 30A power capacity, and solutions must operate within this limit. Vendors should design their proposals to meet typical Tier II/Tier III data center standards, and any special requirements (e.g., liquid cooling, high power draw) must be explicitly identified in their proposal.

Cost & Lifecycle Considerations

- Total cost of ownership, including hardware, power, and maintenance.
- Availability of the GPUs used (whether newer-generation GPUs still in production or older-generation models).
- Long-term warranty and availability of replacement parts.
- Availability of vendor support for operation, updates, and maintenance.

Multi-Network Deployment & Support

- The proposed solution demonstrates physical and logical isolation of server(s) and GPUs for each network.
- Vendor documentation clearly describes how the solution maintains separation, mitigates cross-network risk, and supports compliance with security policies.
- Vendor’s ability to transport, deliver, and support installation at the remote site.
- Delivery and deployment timeline.
- Level of pre-configuration for software, drivers, and libraries to enable rapid deployment.