

Multicast Use Cases for Large Language Model Synchronization

draft-liu-rtgwg-llmsync-multicast-00

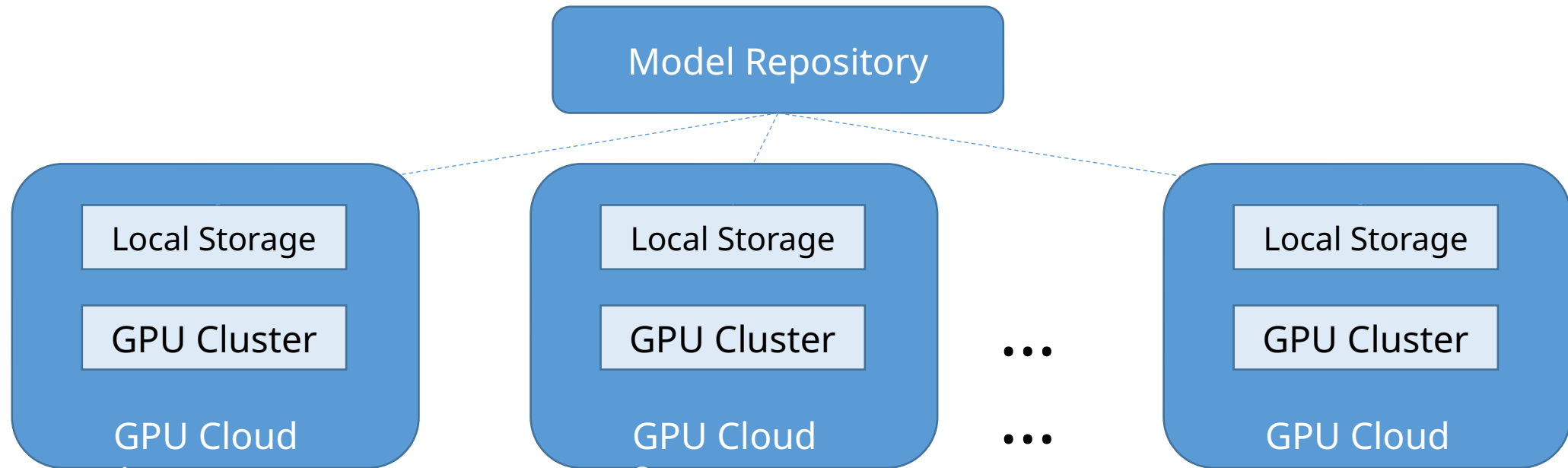
IETF 125 Shenzhen

Yisong Liu(China Mobile)

Sandy Zhang(ZTE)

Junye Zhang(China Mobile)

LLM Synchronization in Inference Clouds Scenario



- **Emerging Inference Cloud Services** : Deliver large-scale real-time inference, fine-tuning and model optimization services on GPU cloud platforms
- **Multi-Cloud LLM Synchronization**: Centralized model repositories automatically replicate and sync LLMs to geographically distributed GPU clouds in different regions or different carrier networks

LLM Synchronization in Inference Clouds Challenge

- High Concurrency

- ✓ A popular large model with the size of 70GB to 1TB may be downloaded simultaneously across dozens of GPU clouds
- ✓ Leading to I/O bottlenecks at the storage repository, delaying model distribution at scale

- Cold Start Latency

- ✓ Inference services cannot start until the model is fully downloaded to the GPU cloud
- ✓ Low download efficiency leads to significant cold start latency, delaying user access to inference

Though separate from training and inference, this synchronization process directly affects the efficiency and reliability of inference service delivery

Why Multicast is needed

- Synchronizing large models to multiple GPU clouds is a typical multicast use case
- Reduces I/O bottlenecks from simultaneous downloads, improves transmission efficiency, and minimizes cold start latency
- GPU clouds span multiple regions and operators, multicast technology capable of operating across core and metro networks is required

Candidate Multicast Technologies Analysis

- PIM-SM
 - ✓ Requires a multicast tree to be established in advance
 - ✓ All nodes along the path must maintain state information
 - ✓ Slow to respond to network topology changes
 - ✓ Suitable for scenarios where the set of destination GPU clouds is relatively fixed
- SR-P2MP
 - ✓ Relies on a controller to implement multicast traffic engineering
 - ✓ Replicating nodes require state; a multicast tunnel must be established beforehand
 - ✓ Slow to respond to network topology changes
 - ✓ Suitable for scenarios where the set of destination GPU clouds is relatively fixed
- BIER
 - ✓ A stateless multicast technology and no need to establish a multicast tree in advance
 - ✓ Responds quickly to network topology changes
 - ✓ No requirement that the destination GPU clouds set be fixed

Next Steps

- Solicit WG feedback
- Discuss more detailed requirements and potential gaps