LLM Traffic Implications for Peering Strategy

Tomás Lynch - AS20473 - LACNIC 45 - Panamá

VULTR

Disclaimer

- I am presenting what is happening today in my Network

LLM Traffic Profiles

Carga de datos

Dataset Ingestion

- Large volumes of data moved into the training environment
- Example: “Here is a massive dataset about cats.”

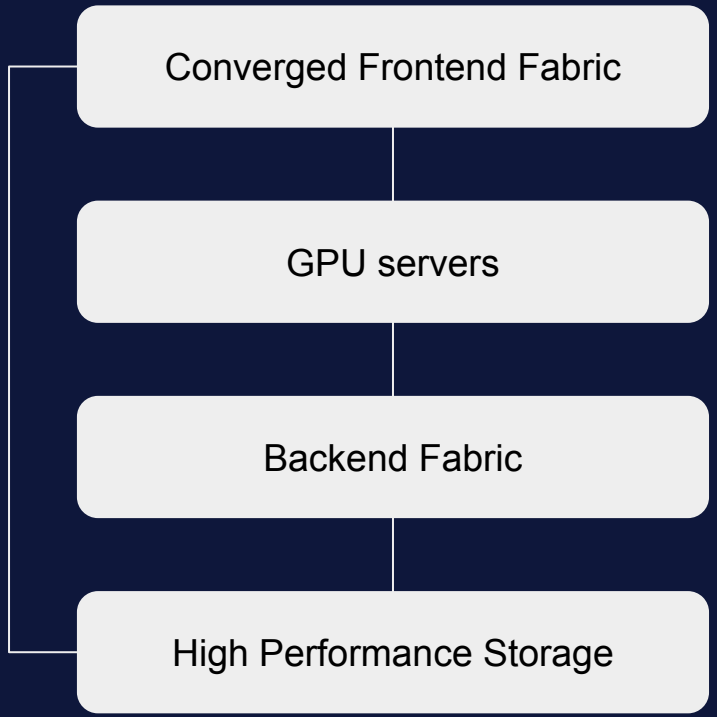
Training

- Model processing and learning from that dataset
- Example: “The model learns patterns about cats.”

Inference

- User interaction with the trained model
- Example: “How do I unclog a toilet?”

GPU Cluster Network Architecture



Converged Frontend Fabric
- Leaf-Spine fabric
- Predominantly North-South traffic
- Inference and dataset ingestion traffic

GPU servers
- Accelerated compute
- Termination point for inference sessions
- Source and sink for backend training traffic

Backend Fabric
- Non-blocking Leaf-Spine fabric
- East-West traffic only
- Training traffic

High-Performance Storage
- Training datasets and checkpoints
- East–West training traffic
- North-South dataset ingestion traffic

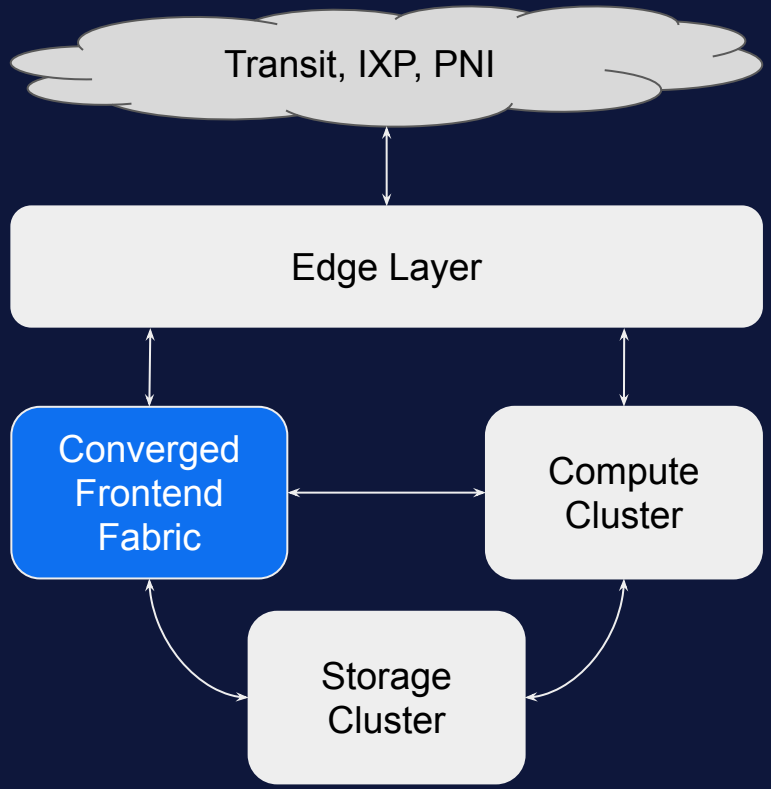
Converged Frontend Fabric Interaction

Edge Layer - North-South Traffic

- Transit, IXPs, and peers
- Shared external connectivity for multiple workload types
- Usually oversubscribed
- 100GE, 400GE interfaces

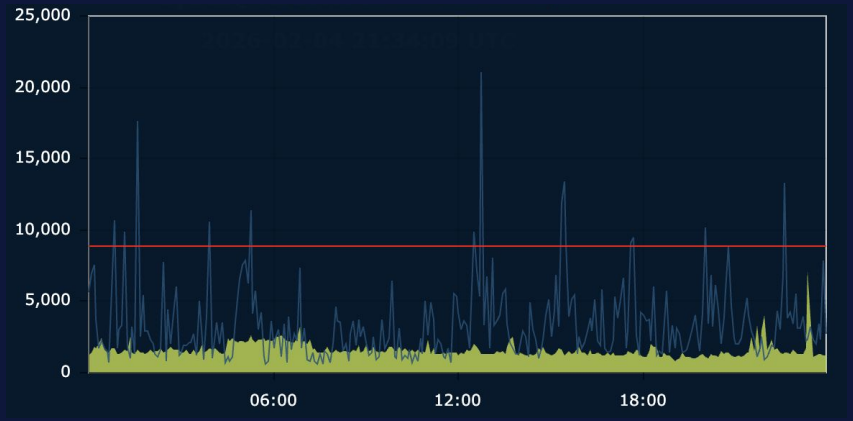
Other Clusters - East-West Traffic

- VM and bare metal compute clusters
- General-purpose storage clusters
- Almost non-blocking
- 400GE, 800GE interfaces



GPU and Compute Traffic Pattern Comparison

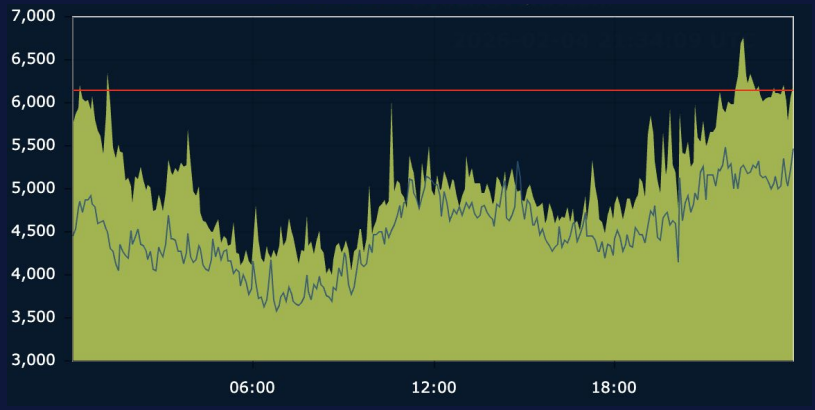
GPU Frontend traffic - 24 Hours



Outbound to GPU (blue) has many spikes

Inbound from GPU (green) is almost constant

Compute Frontend traffic - 24 Hours



Outbound and inbound follow a similar pattern

Traditional sinusoid traffic

Inference and Dataset Ingestion Traffic

Inference Traffic Characteristics

- Low volume
- Constant pattern
- High frequency
- Bidirectional
- Geographically dispersed

Dataset Ingestion Traffic Characteristics

- Massive volume
- Bursty, intermittent pattern
- Low frequency
- Mostly inbound
- Geographically concentrated

Different Traffics, Different Peering Strategies

Inference Traffic Peering Strategy

User-driven, demand-correlated traffic growth

Session-based, asymmetric north–south traffic

Traffic patterns similar to traditional networks

Dataset Ingestion Traffic Peering Strategy

Predominantly inbound over transit

Batch-oriented, throughput-driven traffic

Direct impact on 95th percentile and transit cost

Latency-Driven peering strategy

Cost-Driven peering strategy

Inference Traffic: Cost and Peering Considerations

Inference traffic is growing, more people are using LLM tools.

Cannot be coordinated with end customers

- IXPs, IXPs, IXPs
- PNIs, PNIs, PNIs

Dataset Ingestion Traffic: Cost and Peering Considerations

High-impact on transit cost (95th percentile driven)

Customer coordination and workload shaping

- Can ingestion be rate-limited?
- Can transfers be scheduled or spread over time?

Economic mitigation options

- Increase minimum commitments
- Add or rebalance transit capacity
- Establish direct connectivity to GPU customers

Not related to peering but worth mentioning

Ingestion vs attack ambiguity

- Validate DDoS detection and mitigation thresholds

Mixed flow sizes

- Congestion management becomes critical

Thank you!
Questions?