

mrLLM: Fast Multi-Region LLM Inference using Learned Adapters

Marilyn Rego¹, Maxwell Kumbong², Hermann Kumbong³, Ertza Warraich⁴, Muhammad Shahbaz¹

¹University of Michigan ²Swarthmore College ³Stanford University ⁴Purdue University

Abstract and Motivation

Challenges in Distributed Inference

Challenge 1: Resource Scarcity

High-VRAM GPU instances are geographically scarce and unevenly distributed across regions and providers.

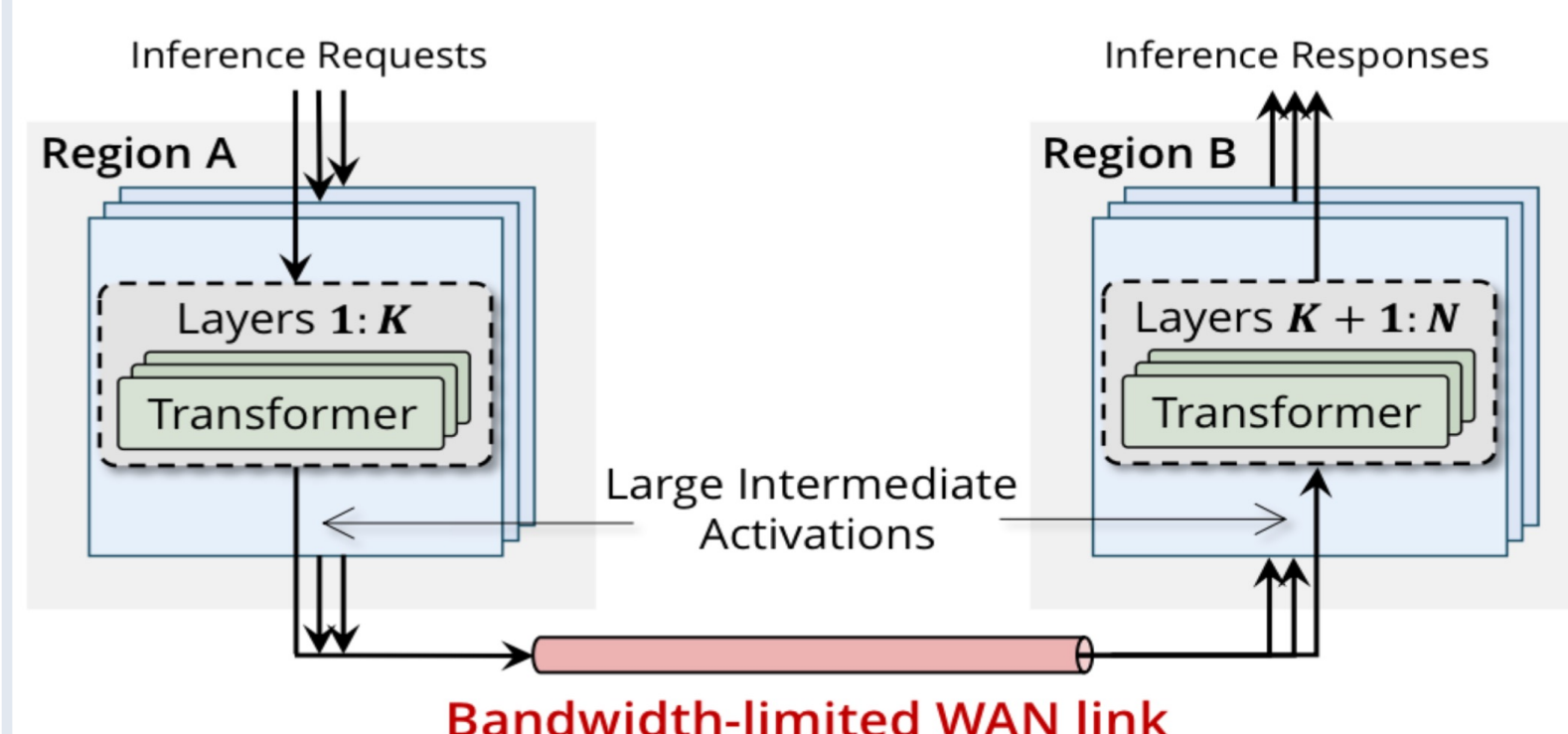
Challenge 2: Cost Inefficiency

GPU pricing varies significantly across regions and providers, while spot instances offer large but underutilized cost savings opportunities.

Challenge 3: Fragmented Infrastructure

Available compute is distributed across regions/providers but connecting them efficiently is constrained by limited and variable network bandwidth (4–10 Gbps WAN).

Communication, not compute, is the bottleneck in WAN



Large activation tensors must traverse bandwidth-limited WAN links at every pipeline boundary, making data transfer the dominant cost on the critical path.

Our Design Addresses This By:

1. Enabling Cross-Provider Pipeline Parallelism

Split model stages across the cheapest available regions/providers to exploit resource availability and price differences.

2. Compressing Activation Communication

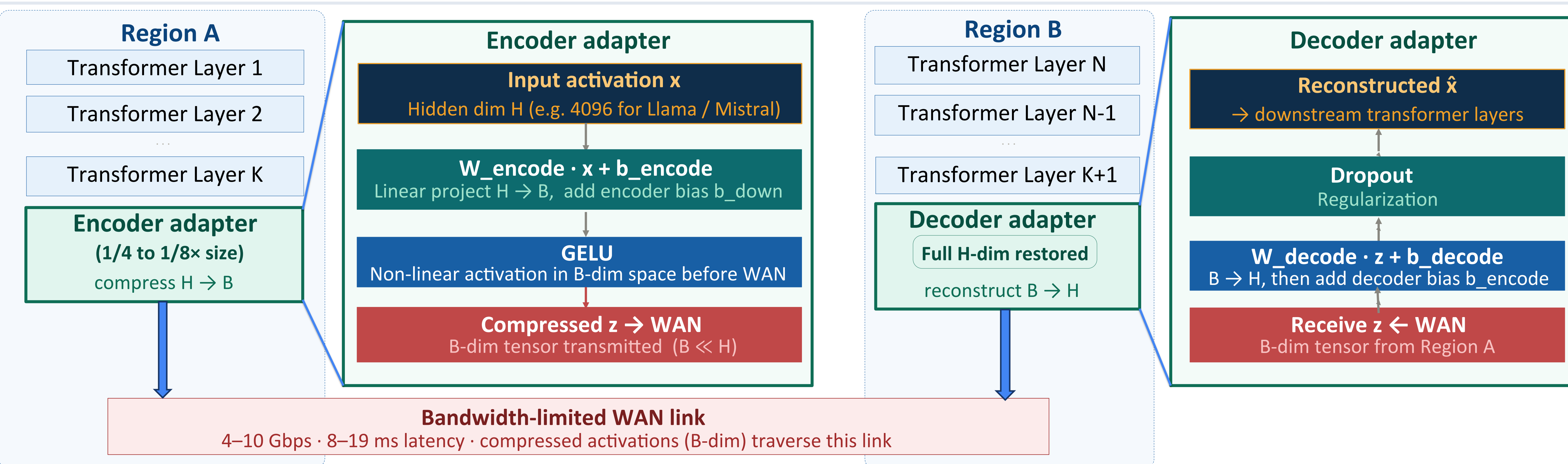
Reduce the size of transmitted activations to fit within realistic WAN bandwidth budgets.

3. Preserving Model Fidelity Under Compression

Use learned, model-aware adapters to maintain downstream accuracy despite aggressive communication reduction.

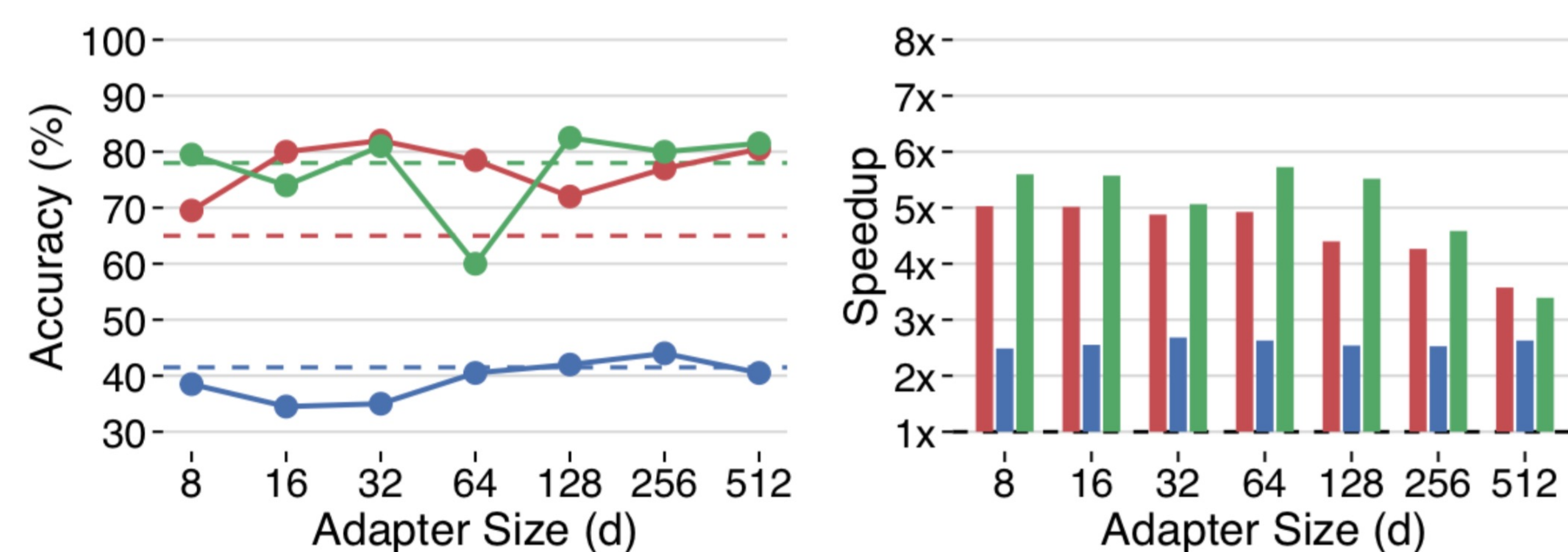
Design of mrLLM

mrLLM: adapter-based activation compression for distributed inference



Evaluation and Results

● Llama-3.1-8B ● GPT-2 XL ● Mistral-7B



Key Results

Activation compression → 4.08× ± 1.23 speedup (avg)

Llama-3.1-8B: 4.58× speedup • +12.1 pts accuracy (up to +17)

GPT-2 XL: 2.58× speedup • ~no loss (recovers to +2.5)

Mistral-7B: 5.07× speedup • best throughput, minor variance

Takeaway: 4–5× faster inference with minimal or improved accuracy