Insights into DeepSeek-V3: Scaling Challenges and Reflections on Hardware for AI Architectures

Submitted May 14th 2025

Presented By: Nathan Kotni

UCDAVIS

Section 1: Introduction

Motivation

- LLM scaling faces new bottlenecks in the current predominant hardware architectures
    - Memory capacity, computational efficiency, and interconnection bandwidth
- DeepSeek-V3: a real system trained at scale on 2,048 NVIDIA H800 GPUs
- Breaking down how a hardware-aware model tackled these challenges

DeepSeek-V3 Overview

- 671B MoE model
- Only ~37B active params/token
- Unique features
    - Multi-head Latent Attention
    - FP8 8-bit training pipeline

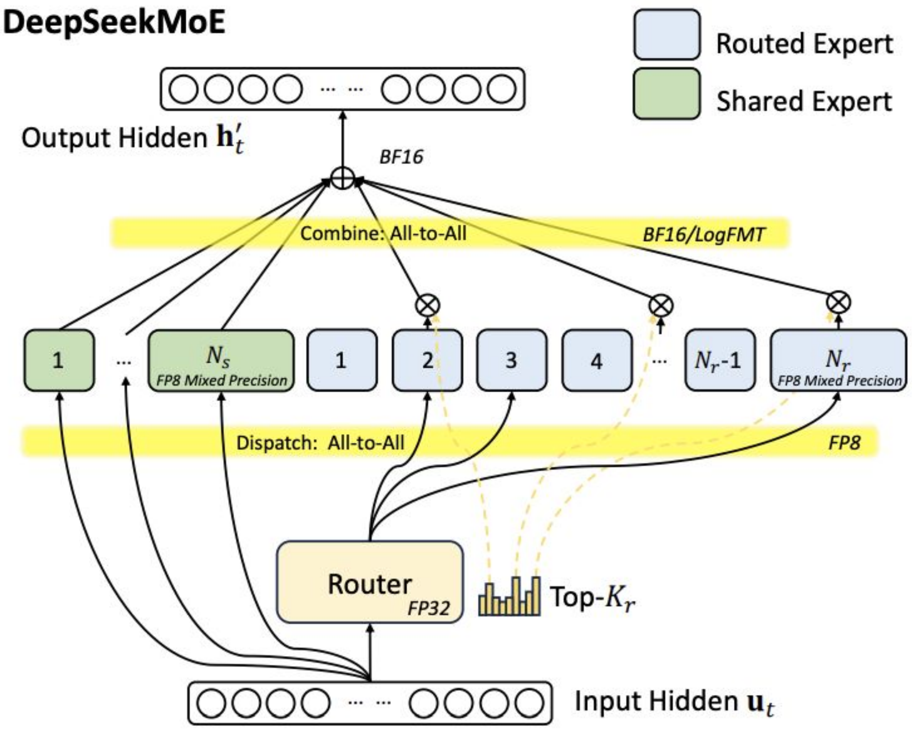
Design Philosophy

- Hardware-first model design
- Large companies often opt for more compute
- For DeepSeek’s team: Efficiency > raw parameter size
- Co-design approach

Section 2: Model Architecture Elements

Mixture of Experts

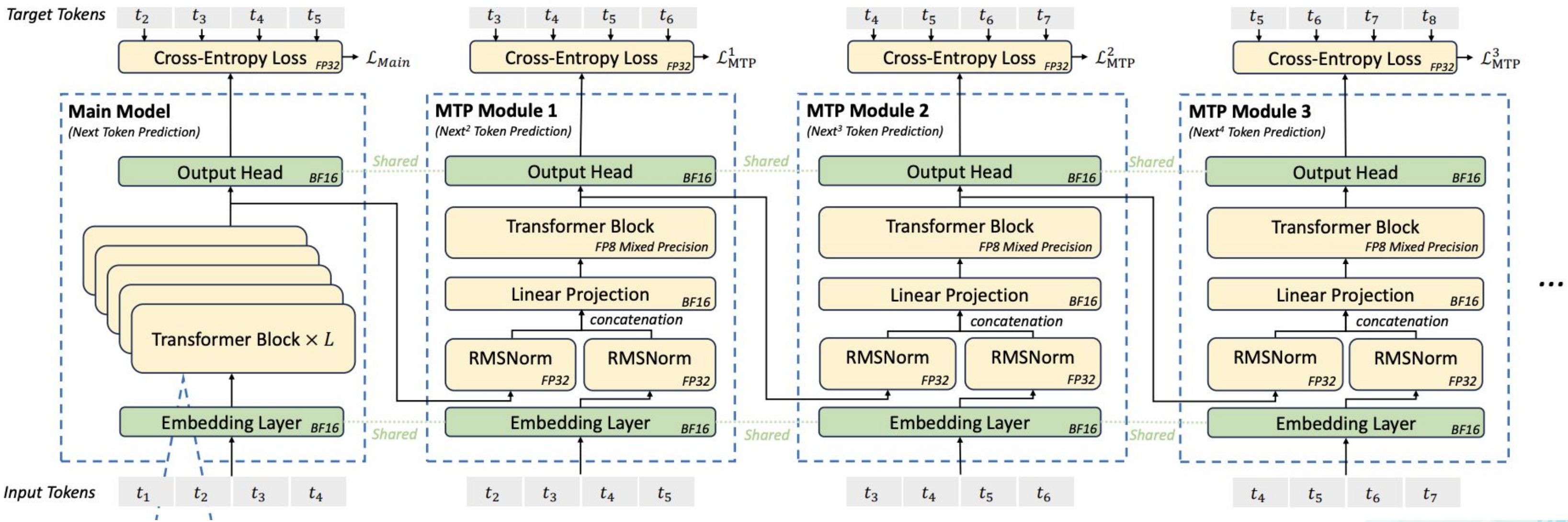
- Many experts in a network, each token only activates two
- Routed experts and shared experts
- Expert Parallelism Routing routing requires All-to-all network communication



From Figure 1 of Paper

Multi-Token Prediction

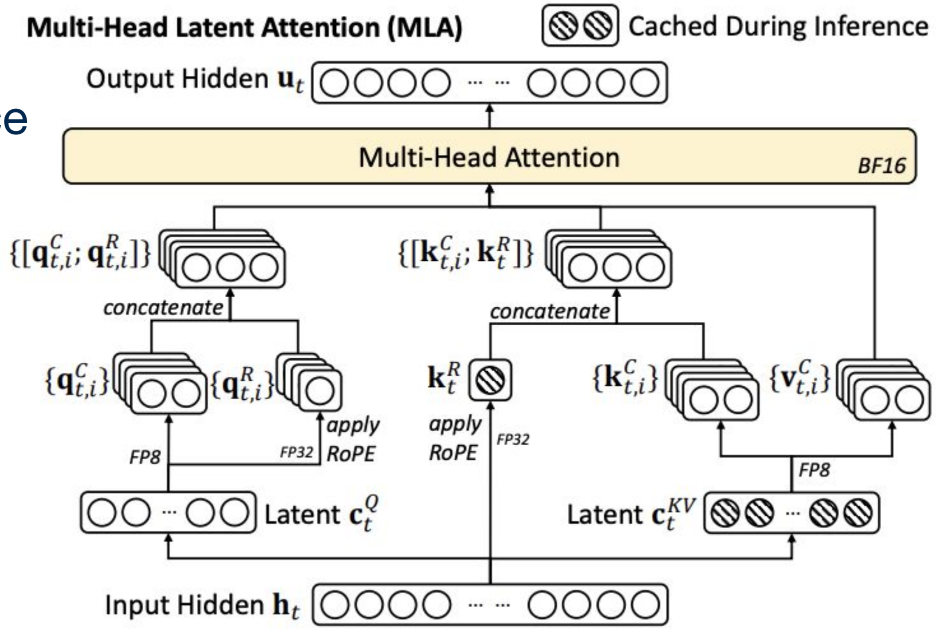
- Adding lightweight modules to predict several subsequent tokens
- Speeds up inference, real world practice data: 80-90% secondary token acceptance with TPS increased by 1.8



From Figure 1 of Paper

Multi-Head Latent Attention

- Compress KV-cache into latent space
- Reduces cache per token by 7.28 times compared to LLaMA-3.1 405B
- Maintains high accuracy
- Allows longer context windows
- Lowers serving costs and feasible on smaller GPUs



From Figure 1 of Paper

FP8 Mixed Precision Training

- Core application of low-precision design
- Using FP8 to store activations and weights reduces memory usage and increases training speed
- Minimal accuracy degradation (<0.25%)
- Limitations:
    - FP8 Accumulation Precision, Fine-Grained Quantization Challenges

Section 3: System Innovations

Dual Micro-Batch Overlap

- Split MLA and MoE operations into 2 distinct stages, form 2 micro-batches
- Handle the computation of one batch while the communication steps like dispatch of the other batch run in parallel on a separate worker, then swap
- Full GPU utilization throughout

Node-Limited Routing

- Locality-aware MoE routing: tokens prefer experts residing on the same GPU node
- Reduces cross-node all-to-all communication, main MoE bottleneck
- Prevents network congestion
- Improves training stability and inference throughput by smoothing communication load
- Spills over to external nodes only when necessary, balancing efficiency and model quality

Multi-Plane Fat-Tree

- High-bandwidth switch network made to avoid bottlenecks at upper layers
- Multiple parallel paths between nodes
- Ideal for MoE all-to-all communication,
- Enables stable, scalable training with consistent performance across thousands of GPUs

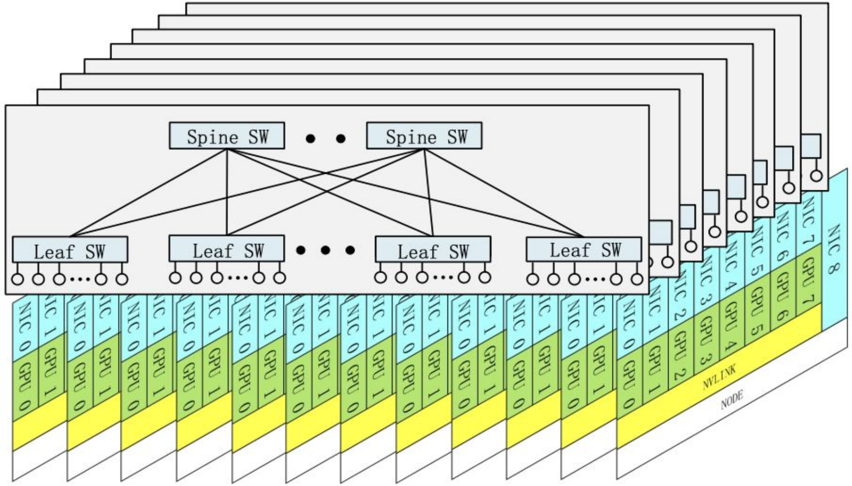


Figure 3: Eight-plane two-layer fat-tree scalue-out network: Each GPU and IB NIC pair belongs to one network plane. Cross-plane traffic must use another NIC and PCIe or NVLink for intra-node forwarding.

Section 4: Conclusion

Future Hardware Needs

- More robust interconnects (beyond NVLink)
- Optimizations for all-to-all Dispatch and Combine communication
- Built-in ordering guarantees for memory-semantic communication
- Memory-centric architectures

Overview

- Scaling can be accessible without excessive compute
- Efficient LLMs reduce cost barriers
- New ways forward
- Inspires smaller and open source teams to continue innovating

Thank You!

Questions?