Advanced Computer Networks

Data Center Network for GPUs (II)

CS740

https://pages.cs.wisc.edu/~mgliu/CS740/F25/index.html

Ming Liu
mgliu@cs.wisc.edu

Outline

- Last lecture
  - Data Center Network For GPUs (I)

- Today
  - Data Center Network For GPUs (II)

- Announcements
  - Project Presentation on 12/04/2025 and 12/09/2025

Insights into DeepSeek-V3

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

Chenggang Zhao, DeepSeek-AI, Beijing, China, chenggangz@deepseek.com
Chengqi Deng, DeepSeek-AI, Beijing, China, cq.deng@deepseek.com
Chong Ruan, DeepSeek-AI, Beijing, China, chong.ruan@deepseek.com
Damai Dai, DeepSeek-AI, Beijing, China, damai.dai@deepseek.com
Huazuo Gao, DeepSeek-AI, Beijing, China, gaohuazuo@deepseek.com
Jiashi Li, DeepSeek-AI, Beijing, China, js.li@deepseek.com
Liyue Zhang*, DeepSeek-AI, Beijing, China, ly.zhang@deepseek.com
Panpan Huang, DeepSeek-AI, Beijing, China, pp.huang@deepseek.com
Shangyan Zhou, DeepSeek-AI, Beijing, China, sy.zhou@deepseek.com
Shirong Ma, DeepSeek-AI, Beijing, China, mashirong.2000@deepseek.com
Wenfeng Liang, DeepSeek-AI, Beijing, China, wenfeng.liang@deepseek.com
Ying He, DeepSeek-AI, Beijing, China, ying.he@deepseek.com
Yuqing Wang*, DeepSeek-AI, Beijing, China, wangyq@deepseek.com
Yuxuan Liu, DeepSeek-AI, Beijing, China, liuyuxuan@deepseek.com
Y.X. Wei, DeepSeek-AI, Beijing, China, weiyx@deepseek.com

Abstract

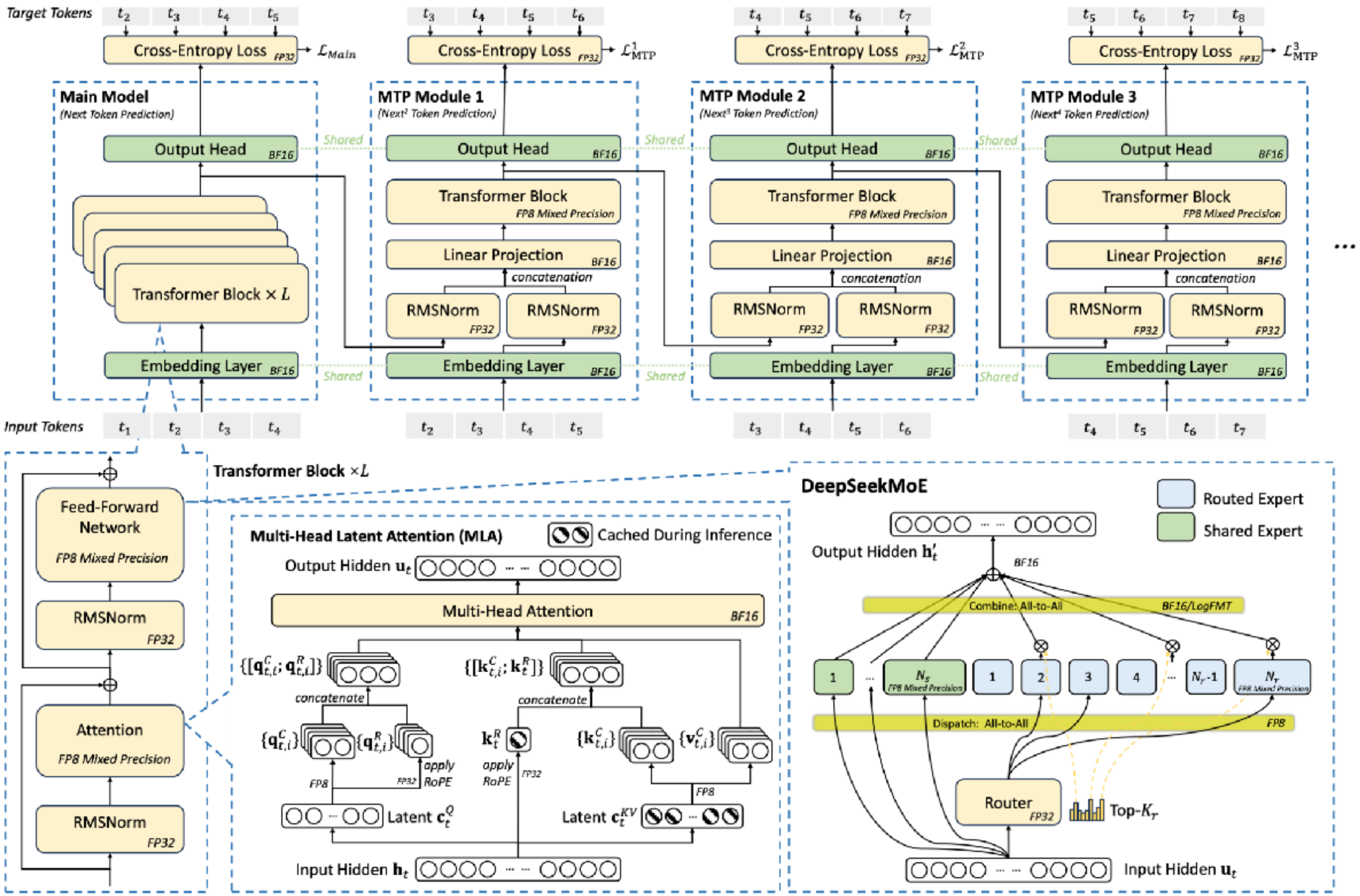
The rapid scaling of large language models (LLMs) has unveiled critical limitations in current hardware architectures, including constraints in memory capacity, computational efficiency, and interconnection bandwidth. DeepSeek-V3, trained on 2,048 NVIDIA H800 GPUs, demonstrates how hardware-aware model co-design can effectively address these challenges, enabling cost-efficient training and inference at scale. This paper presents an in-depth analysis of the DeepSeek-V3/R1 model architecture and its AI infrastructure, highlighting key innovations such as Multi-head Latent Attention (MLA) for enhanced memory efficiency, Mixture of Experts (MoE) architectures for optimized computation-communication trade-offs, FP8 mixed-precision training to unlock the full potential of hardware capabilities, and a Multi-Plane Network Topology to minimize cluster-level network overhead. Building on the hardware bottlenecks encountered during DeepSeek-V3's development, we engage in a broader discussion with academic and industry peers on potential future hardware directions, including precise low-precision computation units, scale-up and scale-out convergence, and innovations in low-latency communication fabrics. These insights underscore the critical role of hardware and model co-design in meeting the escalating demands of AI workloads, offering a practical blueprint for innovation in next-generation AI systems.

CCS Concepts

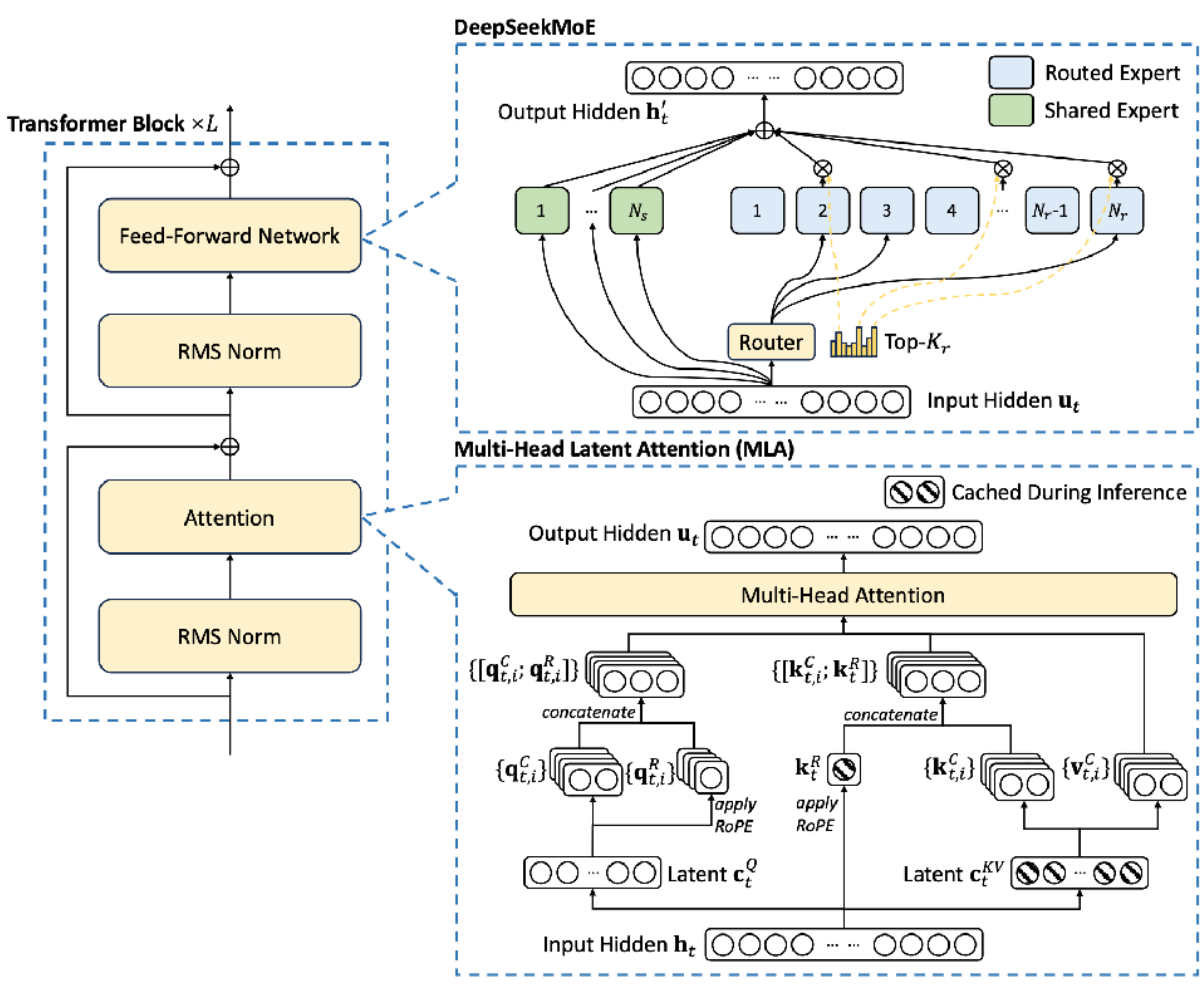
• **Computer systems organization → Architectures.**

- DeepSeek-V3 Overview
- Low-Precision
- Interconnect
- Cluster Network
- Looking Forward: Challenges

Basic Architecture of DeepSeek-V3

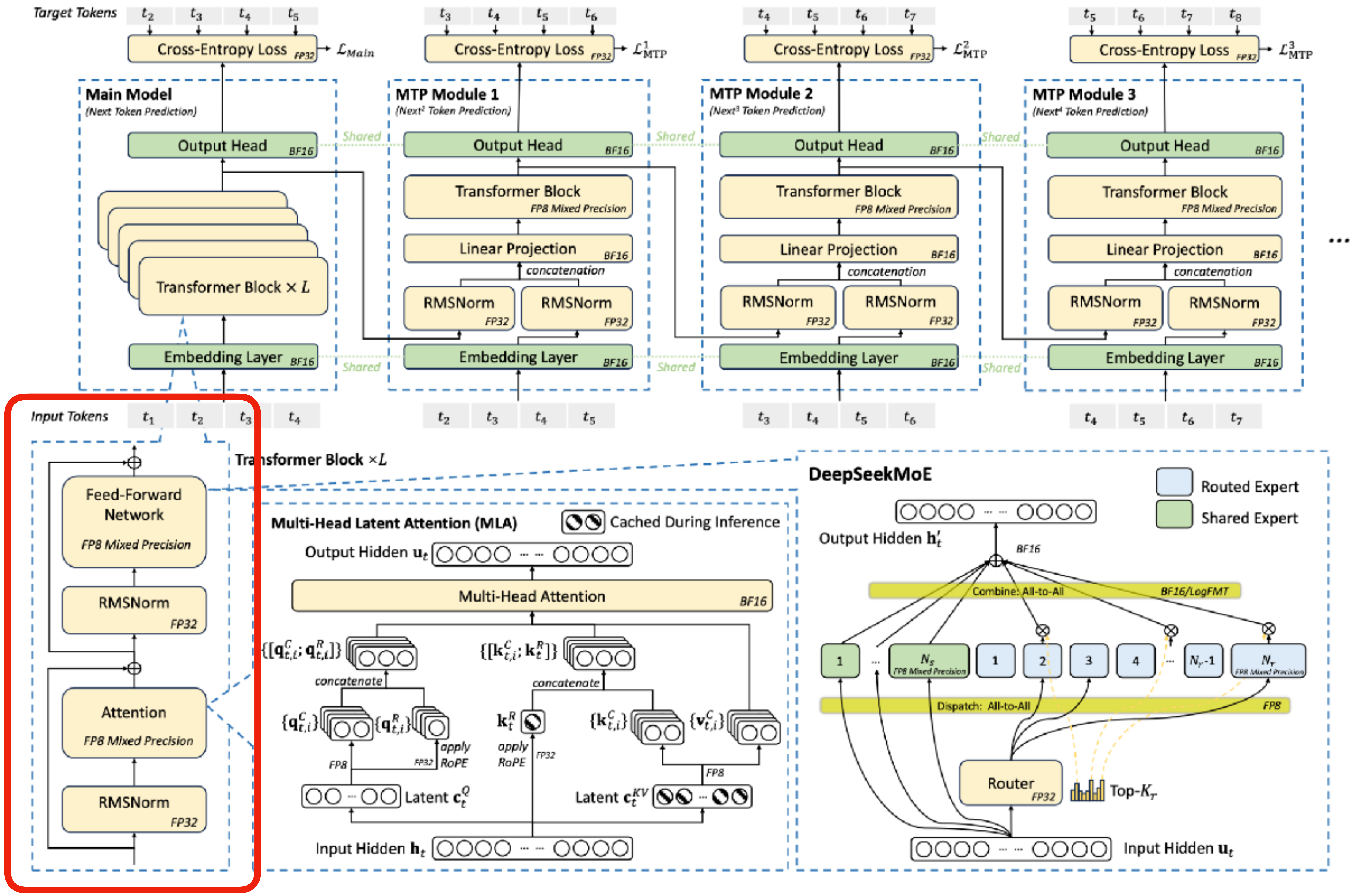


DeepSeek-V2

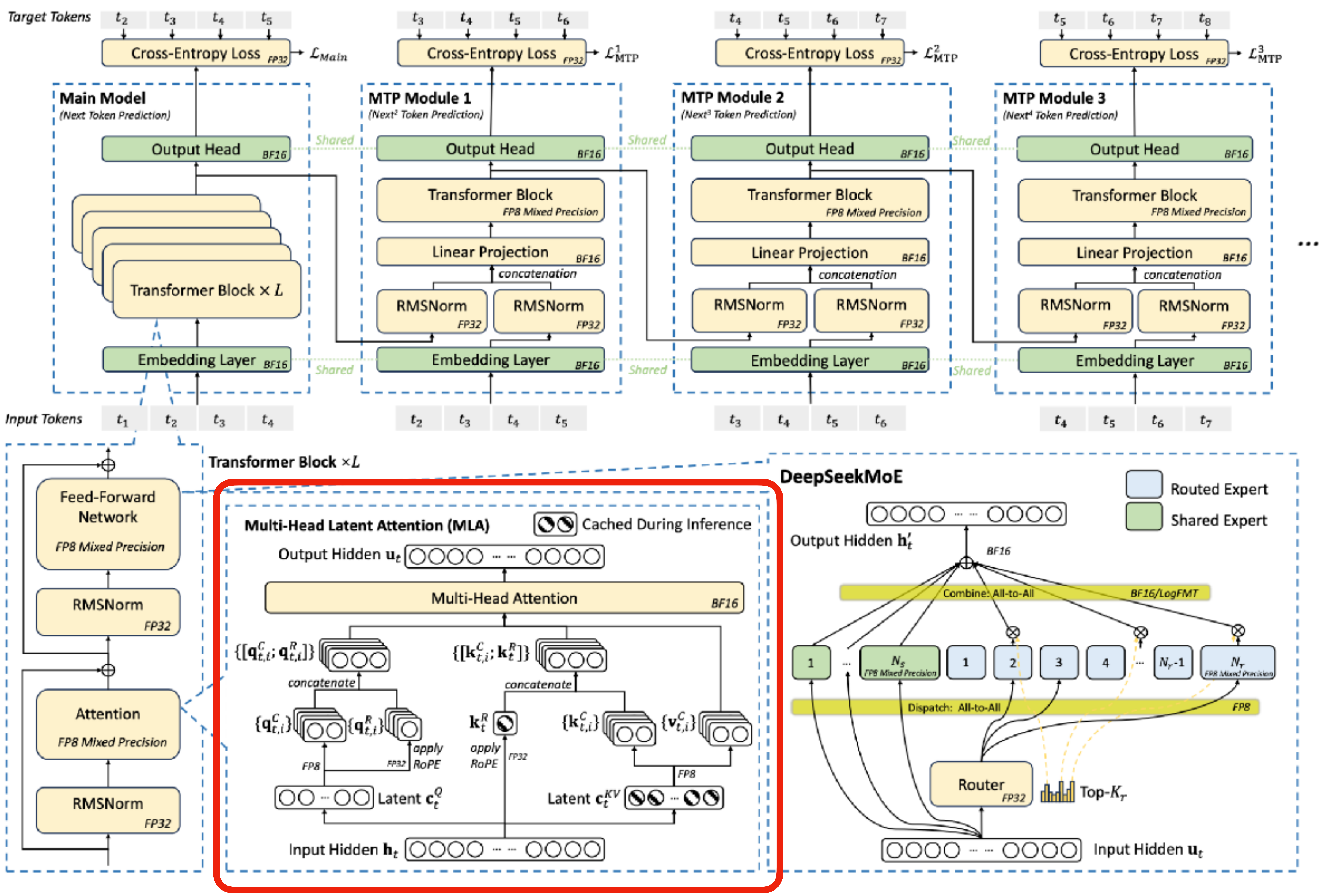


Transformer Block ×L
Feed-Forward Network
RMS Norm
Attention
RMS Norm
DeepSeekMoE
Output Hidden $\mathbf{h}_t'$
Routed Expert
Shared Expert
1
N_s
1
2
3
4
N_r-1
N_r
Router
Top-K_r
Input Hidden $\mathbf{u}_t$
Multi-Head Latent Attention (MLA)
Cached During Inference
Output Hidden $\mathbf{u}_t$
Multi-Head Attention
$\{[\mathbf{q}_{t,i}^C; \mathbf{q}_{t,i}^R]\}$
$\{[\mathbf{k}_{t,i}^C; \mathbf{k}_t^R]\}$
concatenate
concatenate
$\{\mathbf{q}_{t,i}^C\}$
$\{\mathbf{q}_{t,i}^R\}$
apply RoPE
$\mathbf{k}_t^R$
apply RoPE
$\{\mathbf{k}_{t,i}^C\}$
$\{\mathbf{v}_{t,i}^C\}$
Latent $\mathbf{c}_t^Q$
Latent $\mathbf{c}_t^{KV}$
Input Hidden $\mathbf{h}_t$

Transformer-based



Multi-head Latent Attention: Memory Bottleneck

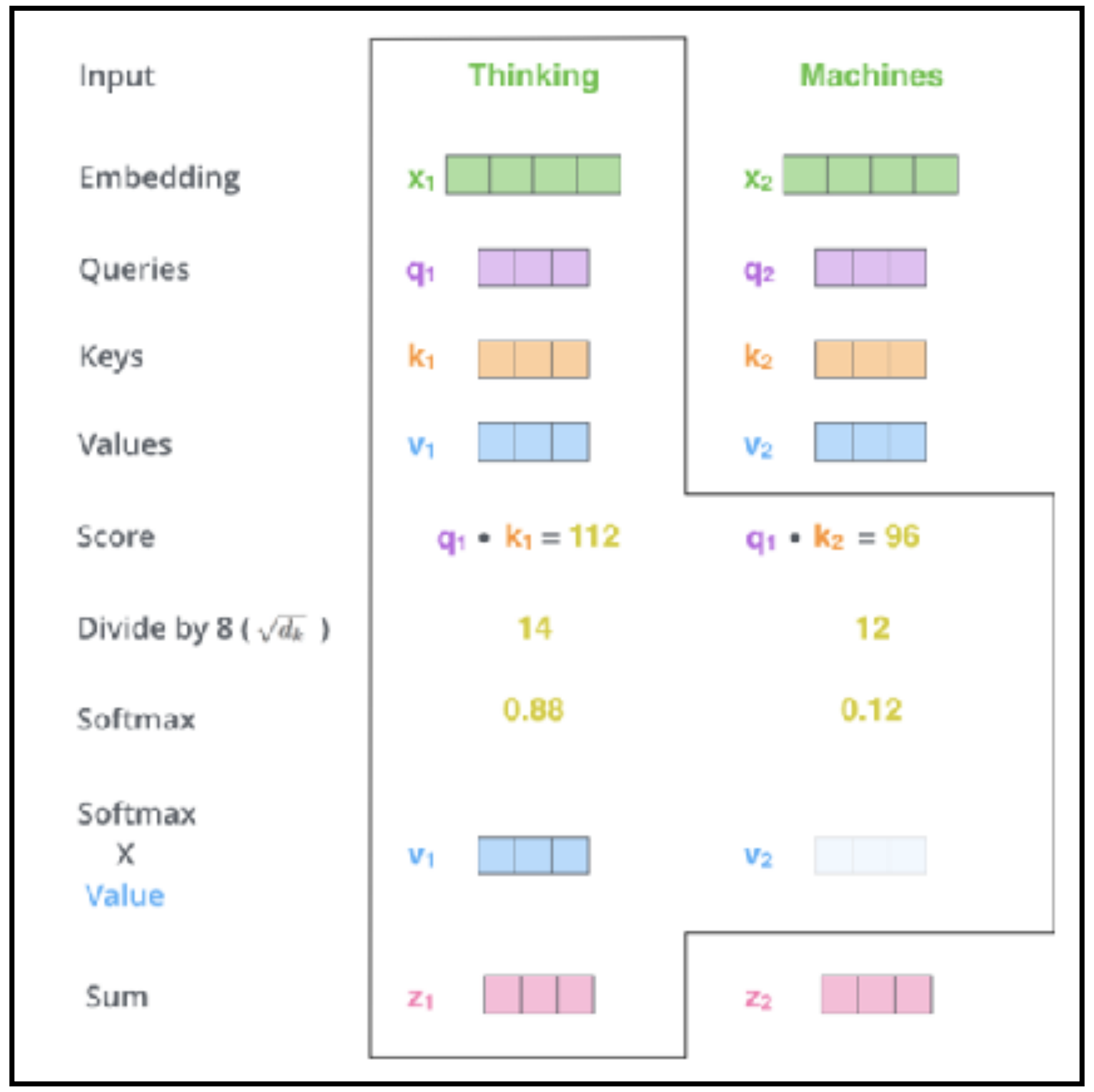


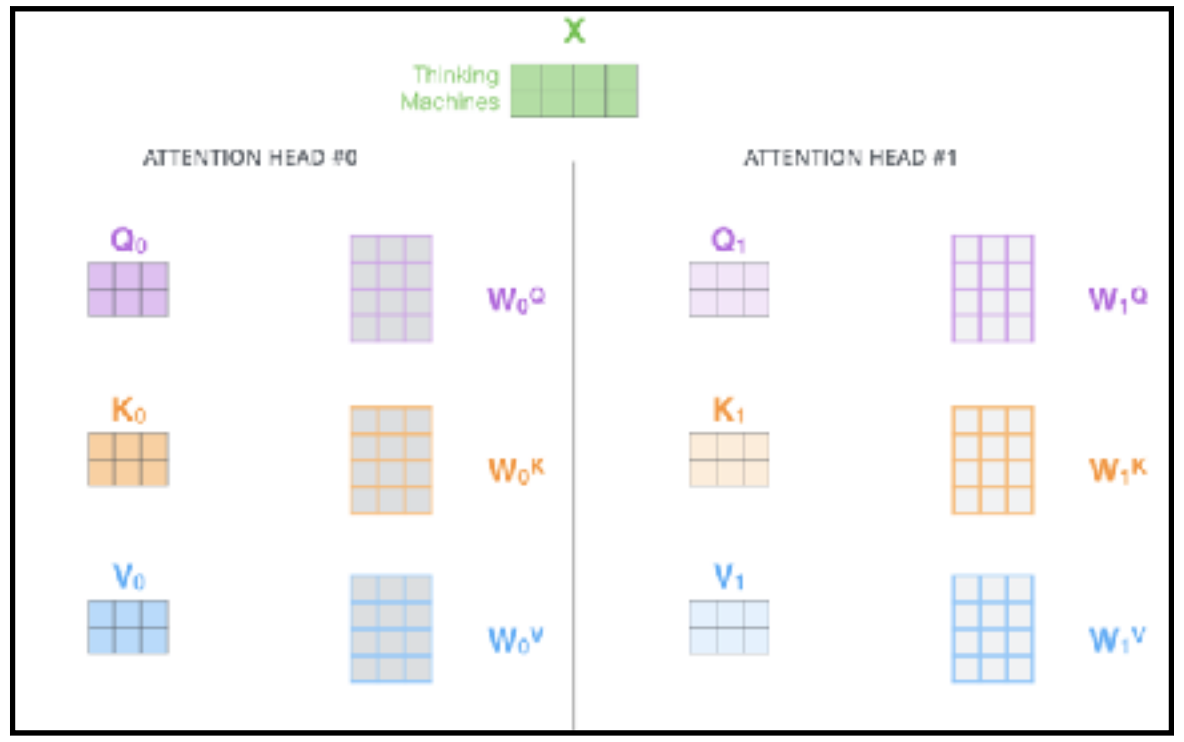
Multi-Head Attention

- Reduce KV Cache with MLA
  - Multi-Head Attention: try to find *correlation* between inputs

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

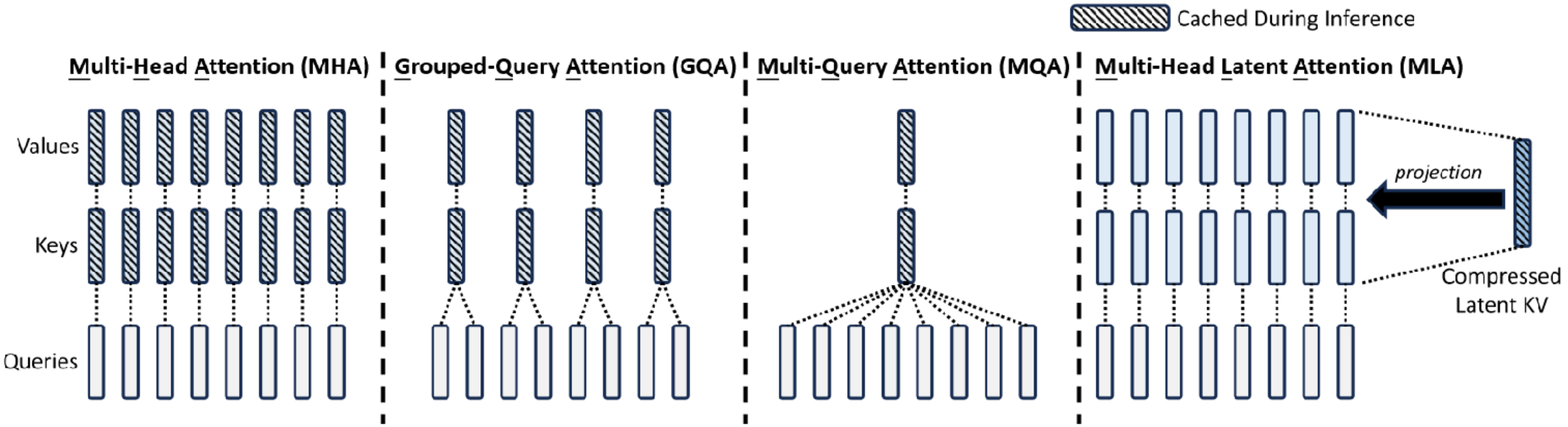
$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head_1}, ..., \mathrm{head_h})W^O$$
$$\text{where } \mathrm{head_i} = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$





Memory Efficiency

- Multi-Head Attention with Causal Mask:
- The i-th row represents the i-th token, depending on the KVs of the token 1, 2, …, i-1
- Some existing KV cache optimizatins



Memory Efficiency (cont'd)

- Reduce KV Cache with MLA
- Idea: cache the compressed latent KV, then convert it back to KVs with project matrix

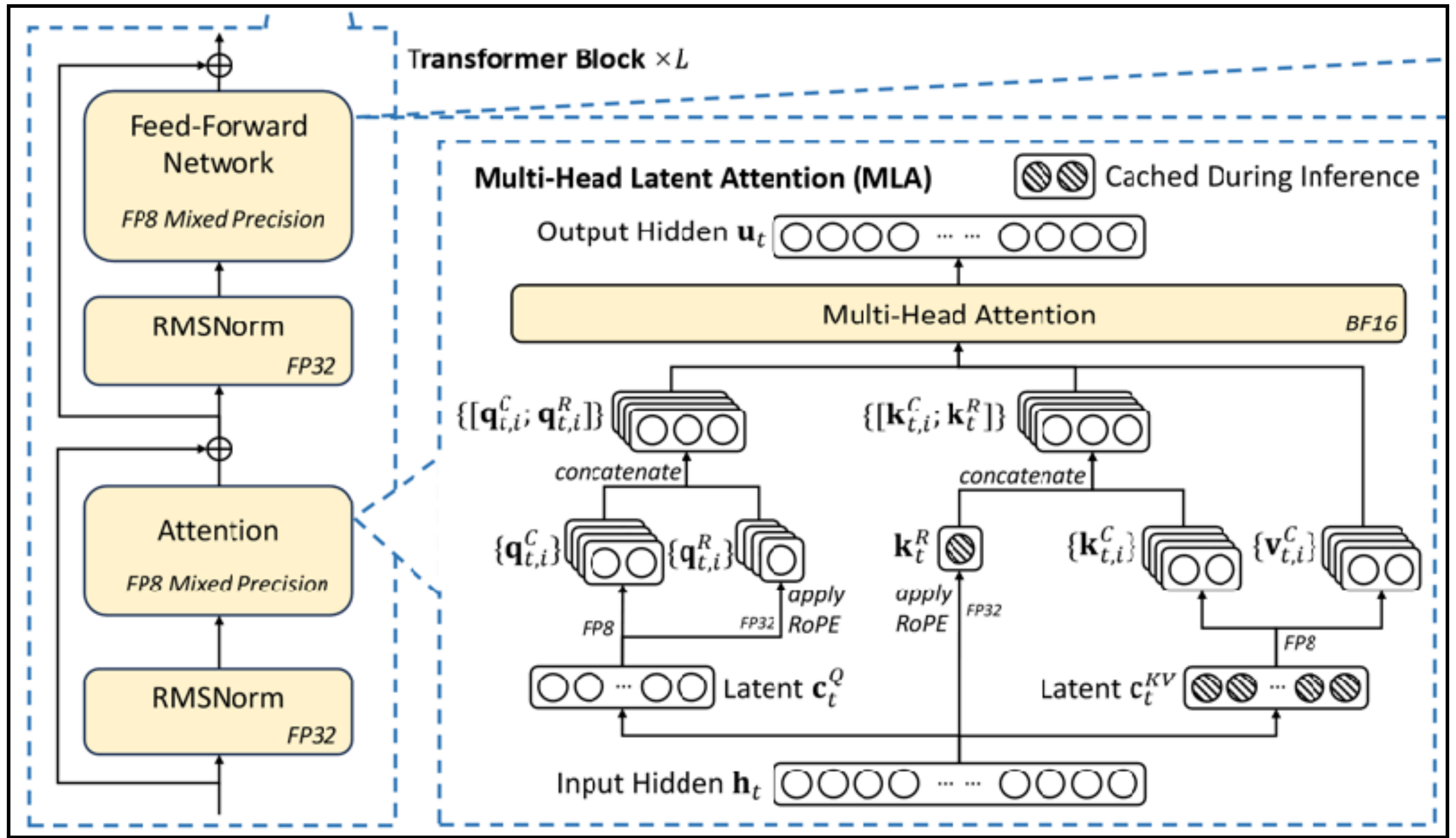
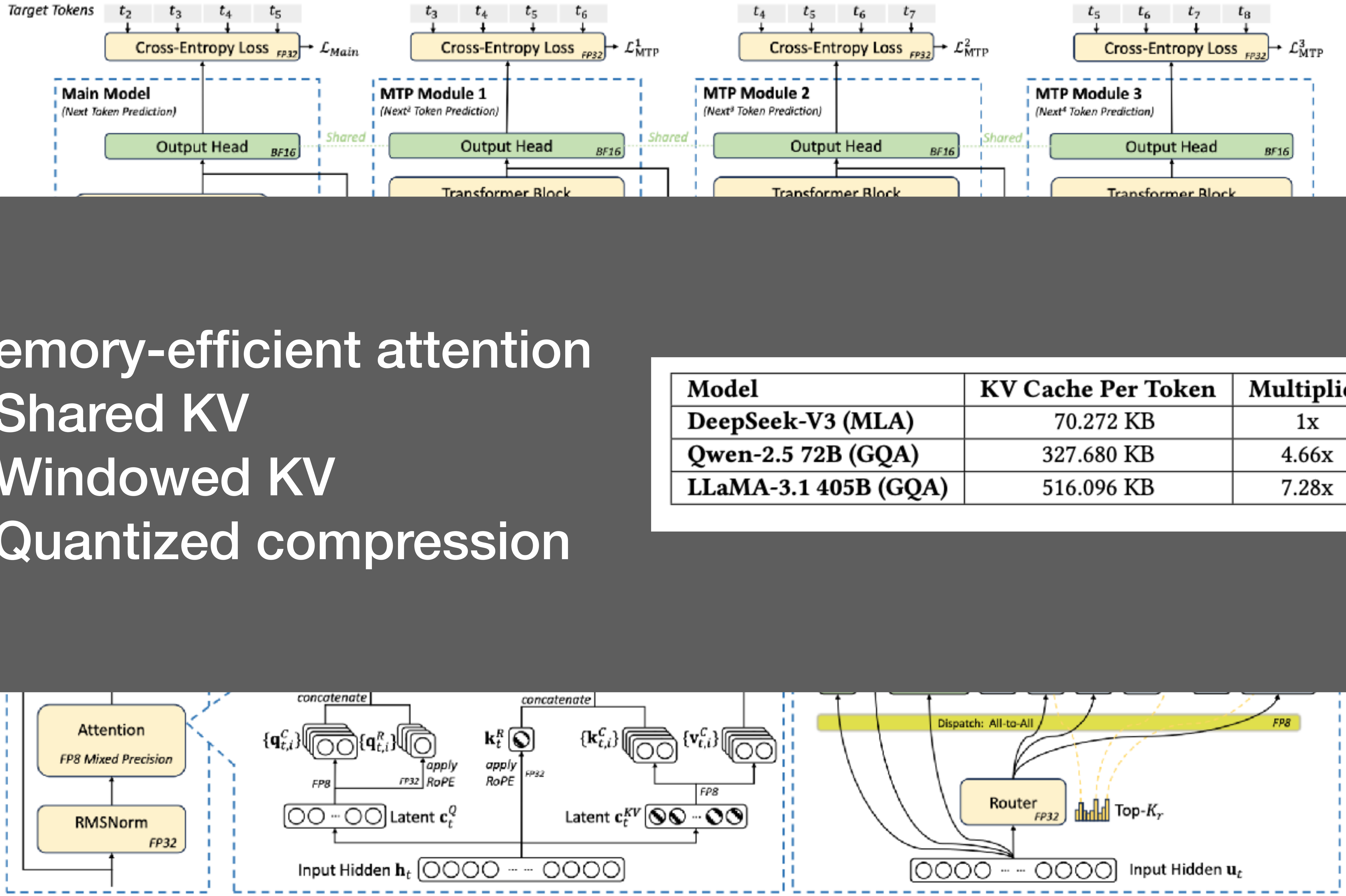


Table 1: KV cache size comparison (BF16 precision): DeepSeek-V3 (MLA) largely reduces KV cache size compared to other models using GQA.

Model	KV Cache Per Token	Multiplier
DeepSeek-V3 (MLA)	70.272 KB	1x
Qwen-2.5 72B (GQA)	327.680 KB	4.66x
LLaMA-3.1 405B (GQA)	516.096 KB	7.28x

Other Memory Optimizations

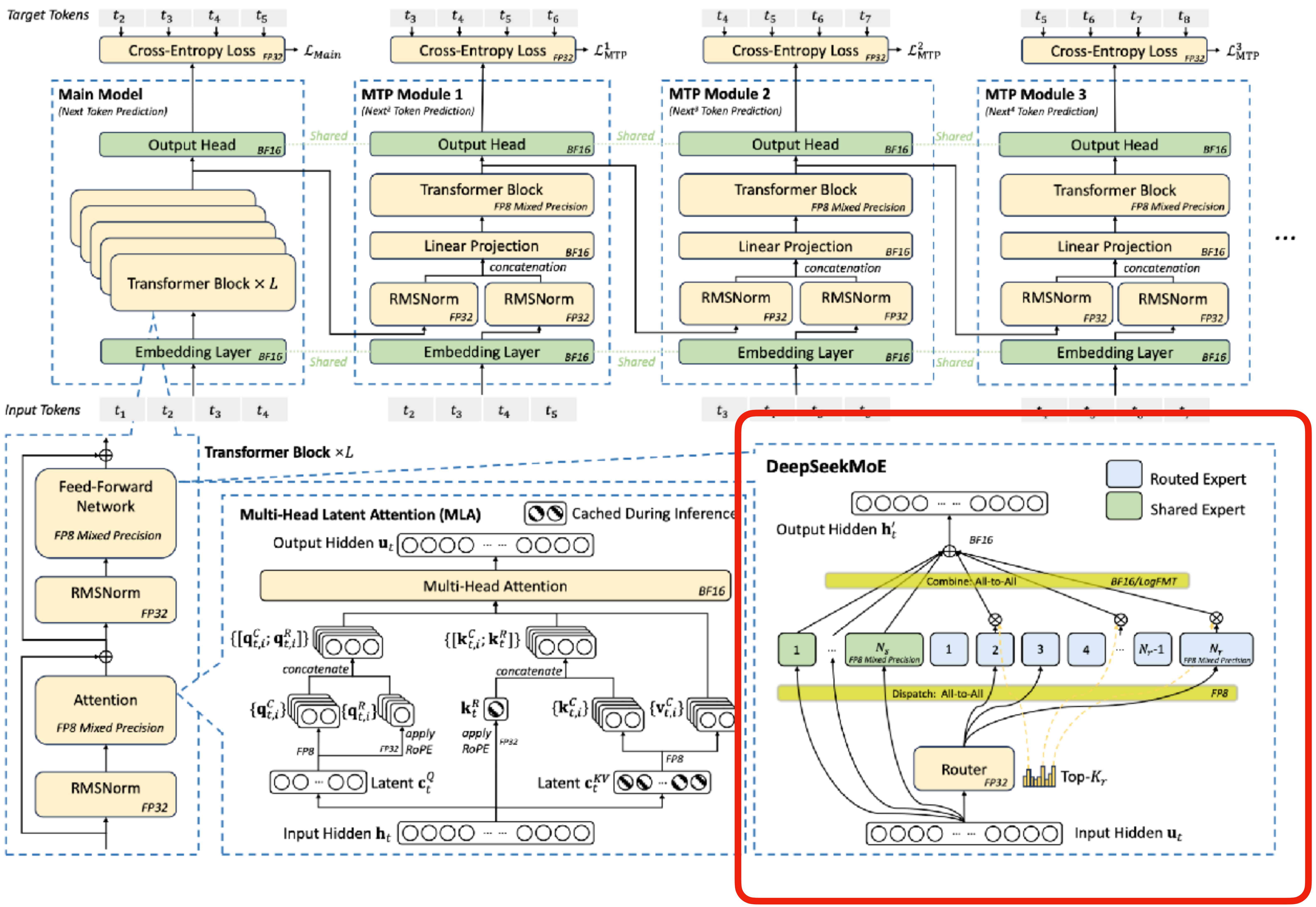


Memory-efficient attention

- Shared KV
- Windowed KV
- Quantized compression

Model	KV Cache Per Token	Multiplier
DeepSeek-V3 (MLA)	70.272 KB	1x
Qwen-2.5 72B (GQA)	327.680 KB	4.66x
LLaMA-3.1 405B (GQA)	516.096 KB	7.28x

DeepSeekMoE: Efficient Sparse Computing



MoE Models

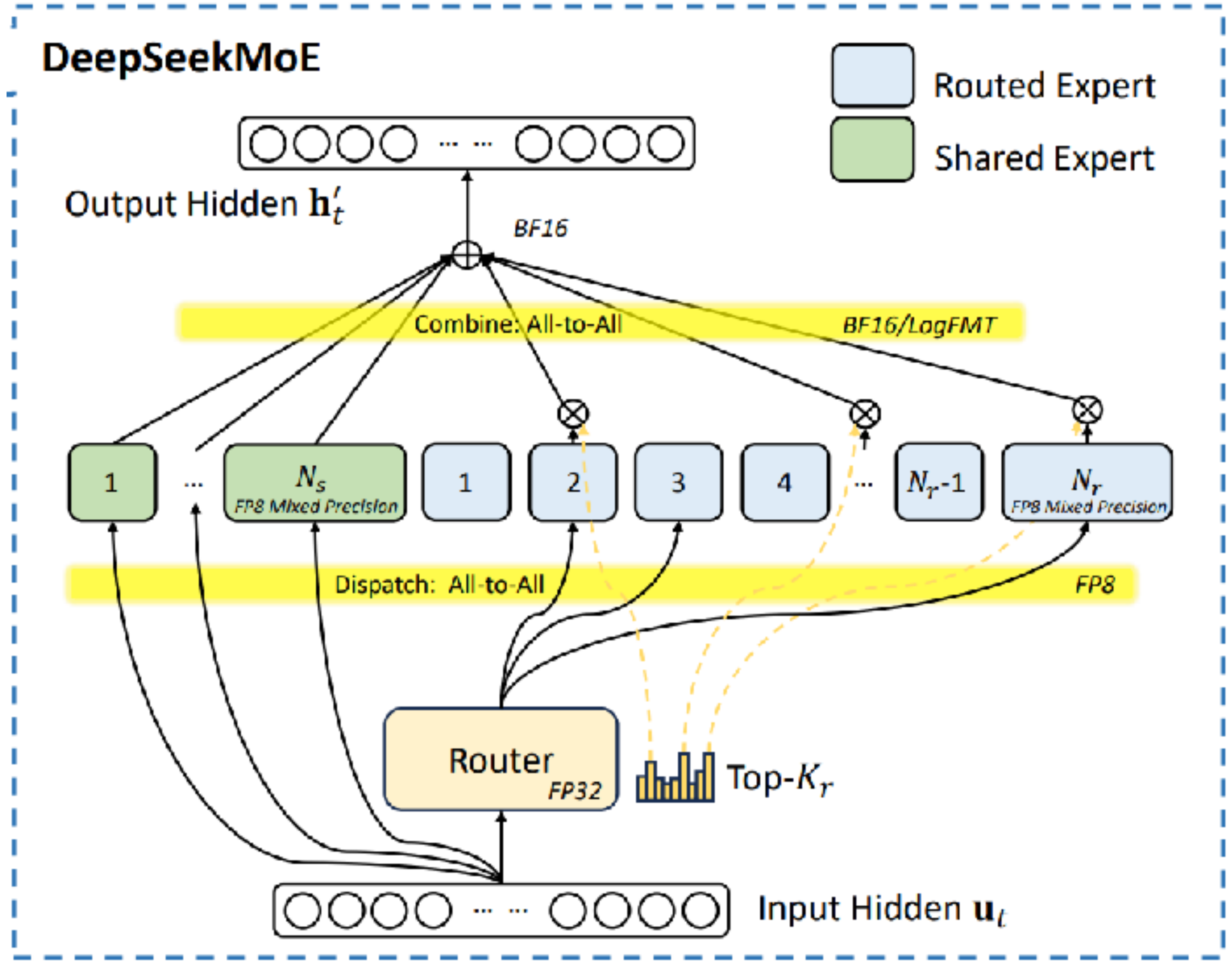
- Experts feature fine-grained.
- Shared Experts are vital
- Activated parameters are largely reduced.
- Fine-grained experts are friendly to deploy.

$$\mathbf{h}'_t = \mathbf{u}_t + \sum_{i=1}^{N_s} \mathrm{FFN}_i^{(s)}(\mathbf{u}_t) + \sum_{i=1}^{N_r} g_{i,t}\, \mathrm{FFN}_i^{(r)}(\mathbf{u}_t),$$

$$g_{i,t} = \frac{g'_{i,t}}{\sum_{j=1}^{N_r} g'_{j,t}},$$

$$g'_{i,t} = \begin{cases} s_{i,t}, & s_{i,t} \in \mathrm{Topk}(\{s_{j,t} | 1 \leqslant j \leqslant N_r\}, K_r), \\ 0, & \text{otherwise}, \end{cases}$$

$$s_{i,t} = \mathrm{Sigmoid}\left(\mathbf{u}_t{}^T \mathbf{e}_i\right),$$



MoE Models

- Experts feature fine-grained.
- Shared Experts are vital
- Activated parameters are largely reduced

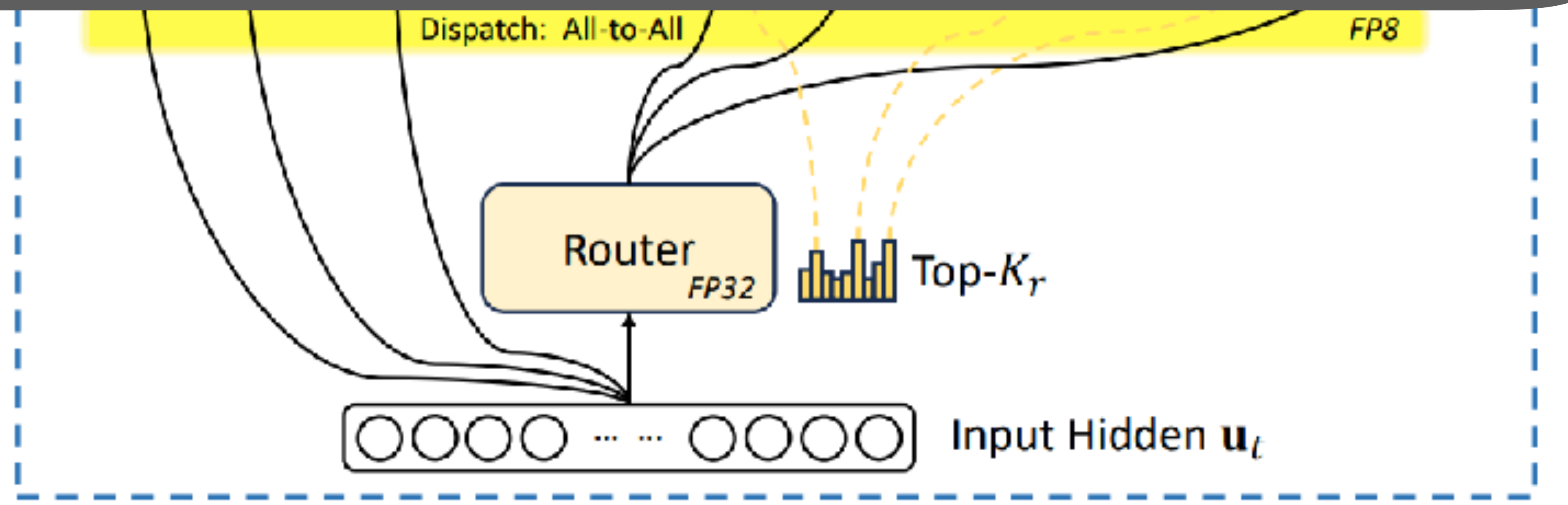
DeepSeek-V2
- 236B parameters
- 21B parameters activated per token

DeepSeek-V3
- 671B parameters
- 21B parameters activated per token

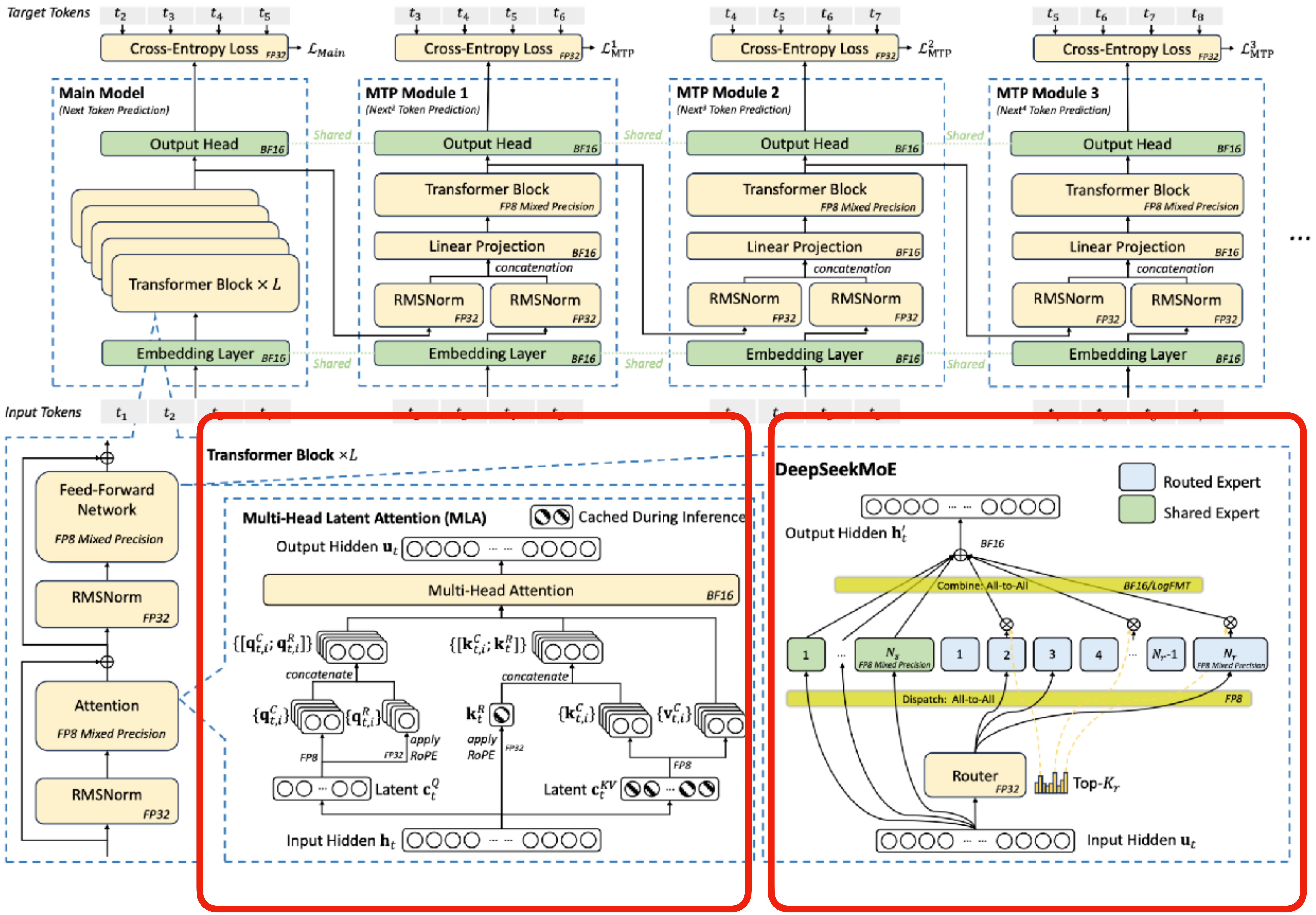
Model	Size	Training Cost
DeepSeek-V2 MoE	236B	155 GFLOPS/Token
DeepSeek-V3 MoE	671B	250 GFLOPS/Token
Qwen-72B Dense	72B	394 GFLOPS/Token
LLaMa-405B Dense	405B	2448 GFLOPS/Token

$$g'_{i,t} = \begin{cases} s_{i,t}, & s_{i,t} \in \operatorname{Topk}(\{s_{j,t} | 1 \leqslant j \leqslant N_r\}, K_r), \\ 0, & \text{otherwise}, \end{cases}$$

$$s_{i,t} = \operatorname{Sigmoid}\left(\mathbf{u}_t{}^T \mathbf{e}_i\right),$$



Decouple MLA and MoE



DualPipe and Computing-Communication Overlap

- 4 components
  - Attention (BWF)
  - All-to-all dispatch
  - MLP (BWF)
  - All-to-all combine (BF)

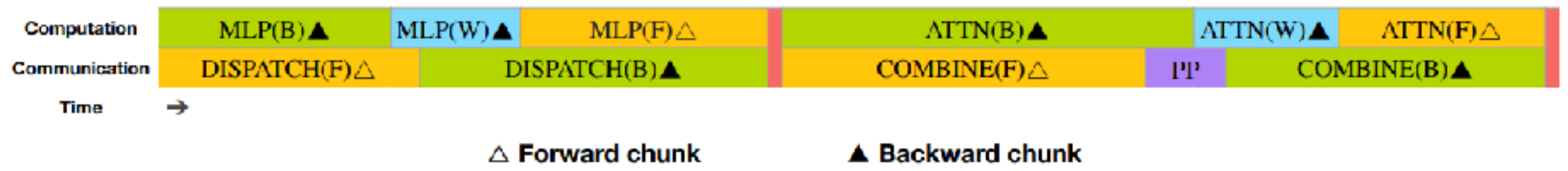


Figure 4 | Overlapping strategy for a pair of individual forward and backward chunks (the boundaries of the transformer blocks are not aligned). Orange denotes forward, green denotes "backward for input", blue denotes "backward for weights", purple denotes PP communication, and red denotes barriers. Both all-to-all and PP communication can be fully hidden.

Figure 5 | Example DualPipe scheduling for 8 PP ranks and 20 micro-batches in two directions. The micro-batches in the reverse direction are symmetric to those in the forward direction, so we omit their batch ID for illustration simplicity. Two cells enclosed by a shared black border have mutually overlapped computation and communication.

Back-of-the-envelope Calculation

- Read Sec 2.3.2 carefully

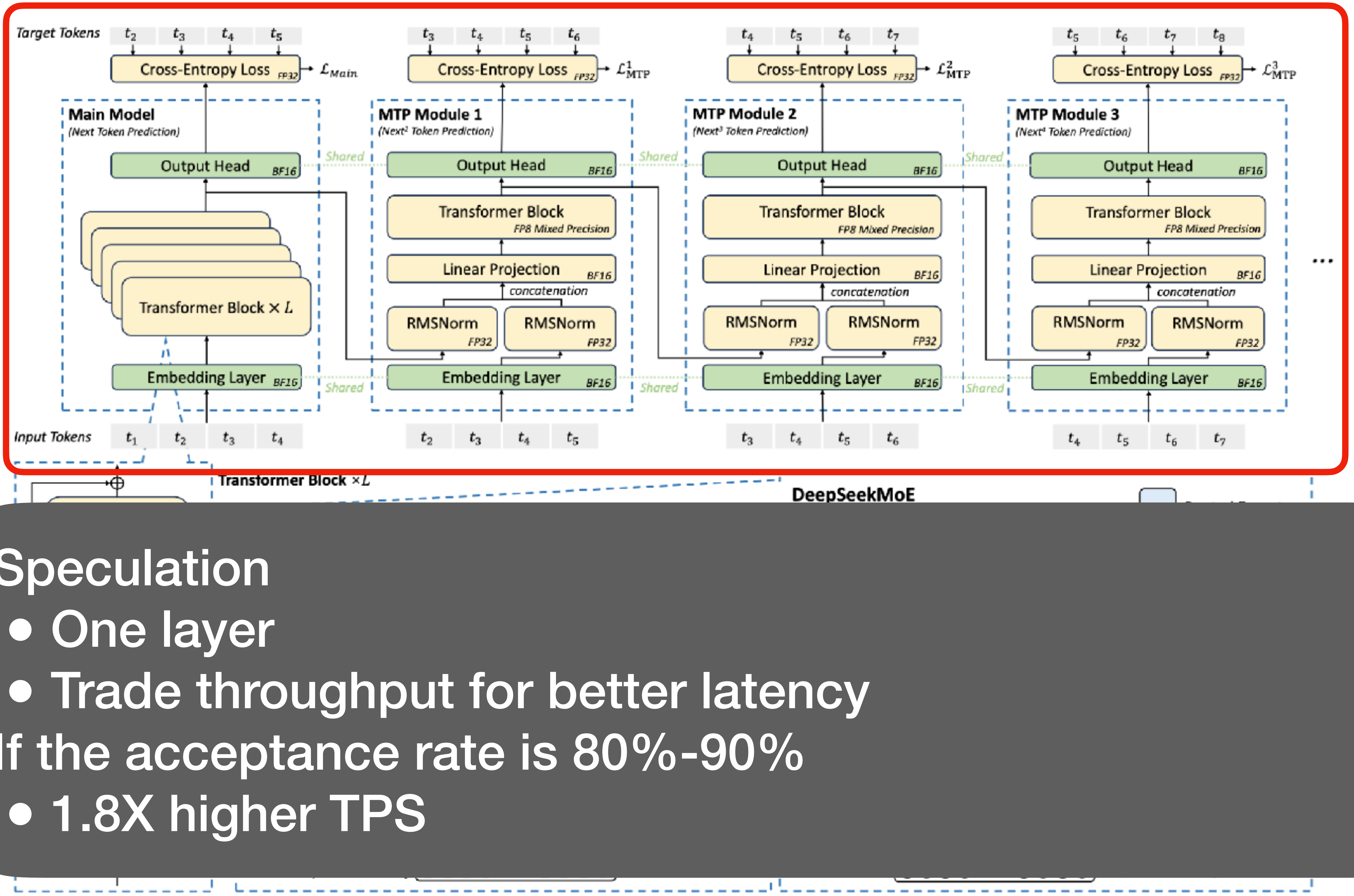
$$\text{Comm. Time} = (1\text{Byte} + 2\text{Bytes}) \times 32 \times 9 \times 7\text{K}/50\text{GB/s} = 120.96\mu s$$

$$\text{Total Time Per Layer} = 2 \times 120.96\mu s = 241.92\mu s$$

$$\text{Total Inference Time} = 61 \times 241.92\mu s = 14.76\text{ms}$$

$$\text{Comm. Time} = (1\text{Byte} + 2\text{Bytes}) \times 32 \times 9 \times 7\text{K}/900\text{GB/s} = 6.72\mu s$$

Multi-Token Prediction (MTP) Framework



Speculation

- One layer
- Trade throughput for better latency

If the acceptance rate is 80%-90%

- 1.8X higher TPS

Multi-Token Prediction (MTP) Framework (cont'd)

$$\mathbf{h}_i'^k = M_k[\mathrm{RMSNorm}(\mathbf{h}_i^{k-1}); \mathrm{RMSNorm}(\mathrm{Emb}(t_{i+k}))],$$

$$\mathcal{L}_{\mathrm{MTP}}^k = \mathrm{CrossEntropy}(P_{2+k:T+1}^k, t_{2+k:T+1}) = -\frac{1}{T}\sum_{i=2+k}^{T+1} \log P_i^k[t_i],$$

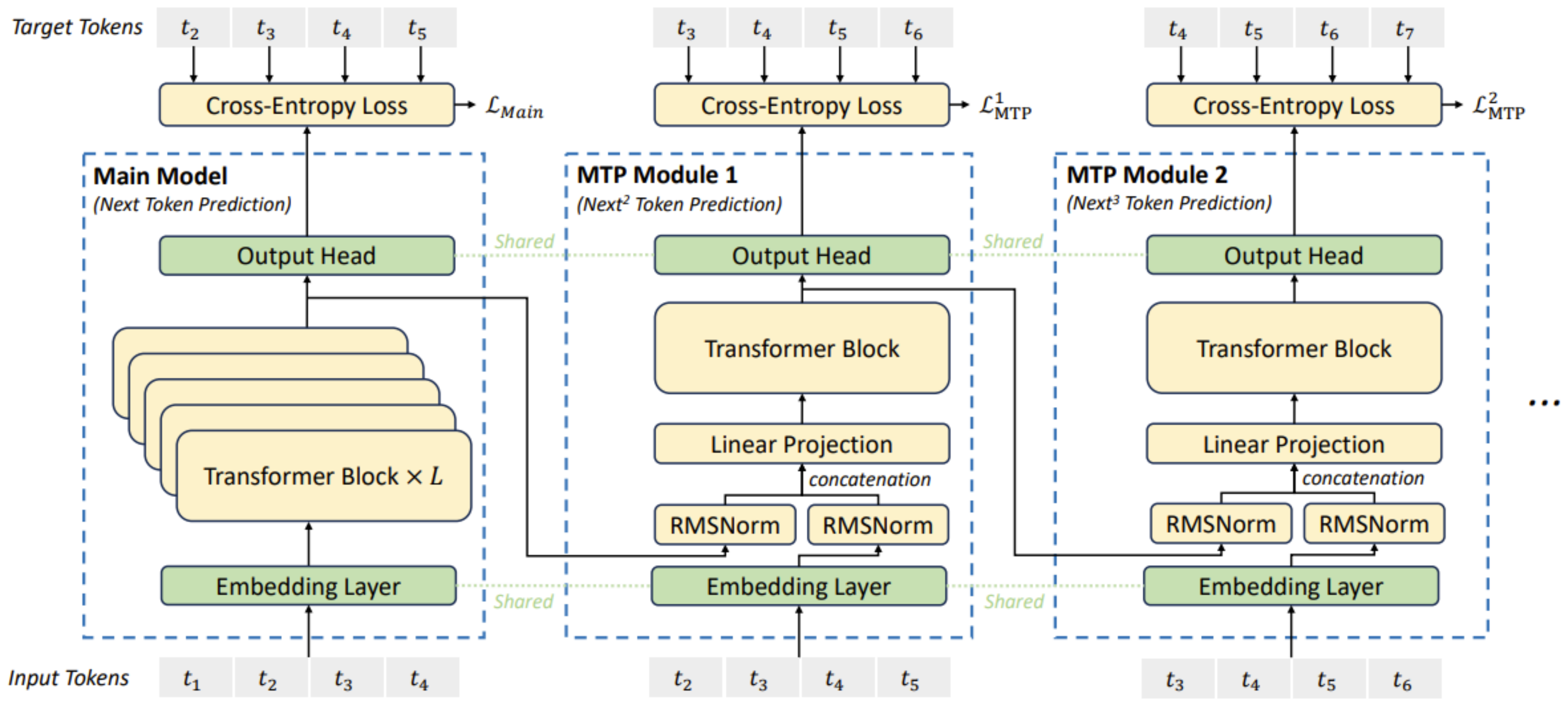


Figure 3 | Illustration of our Multi-Token Prediction (MTP) implementation. We keep the complete causal chain for the prediction of each token at each depth.

Insights into DeepSeek-V3

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

Chenggang Zhao
DeepSeek-AI
Beijing, China
chenggangz@deepseek.com

Chengqi Deng
DeepSeek-AI
Beijing, China
cq.deng@deepseek.com

Chong Ruan
DeepSeek-AI
Beijing, China
chong.ruan@deepseek.com

Damai Dai
DeepSeek-AI
Beijing, China
damai.dai@deepseek.com

Huazuo Gao
DeepSeek-AI
Beijing, China
gaohuazuo@deepseek.com

Jiashi Li
DeepSeek-AI
Beijing, China
js.li@deepseek.com

Liyue Zhang*
DeepSeek-AI
Beijing, China
ly.zhang@deepseek.com

Panpan Huang
DeepSeek-AI
Beijing, China
pp.huang@deepseek.com

Shangyan Zhou
DeepSeek-AI
Beijing, China
sy.zhou@deepseek.com

Shirong Ma
DeepSeek-AI
Beijing, China
mashirong.2000@deepseek.com

Wenfeng Liang
DeepSeek-AI
Beijing, China
wenfeng.liang@deepseek.com

Ying He
DeepSeek-AI
Beijing, China
ying.he@deepseek.com

Yuqing Wang*
DeepSeek-AI
Beijing, China
wangyq@deepseek.com

Yuxuan Liu
DeepSeek-AI
Beijing, China
liuyuxuan@deepseek.com

Y.X. Wei
DeepSeek-AI
Beijing, China
weiyx@deepseek.com

Abstract

The rapid scaling of large language models (LLMs) has unveiled critical limitations in current hardware architectures, including constraints in memory capacity, computational efficiency, and interconnection bandwidth. DeepSeek-V3, trained on 2,048 NVIDIA H800 GPUs, demonstrates how hardware-aware model co-design can effectively address these challenges, enabling cost-efficient training and inference at scale. This paper presents an in-depth analysis of the DeepSeek-V3/R1 model architecture and its AI infrastructure, highlighting key innovations such as Multi-head Latent Attention (MLA) for enhanced memory efficiency, Mixture of Experts (MoE) architectures for optimized computation-communication trade-offs, FP8 mixed-precision training to unlock the full potential of hardware capabilities, and a Multi-Plane Network Topology to minimize cluster-level network overhead. Building on the hardware bottlenecks encountered during DeepSeek-V3's development, we engage in a broader discussion with academic and industry peers on potential future hardware directions, including precise low-precision computation units, scale-up and scale-out convergence, and innovations in low-latency communication fabrics. These insights underscore the critical role of hardware and model co-design in meeting the escalating demands of AI workloads, offering a practical blueprint for innovation in next-generation AI systems.

CCS Concepts

• **Computer systems organization → Architectures.**

- *DeepSeek-V3 Overview*
- Low-Precision
- Interconnect
- Cluster Network
- Looking Forward: Challenges

Quantization

- Primarily used in inference, not training
- NVIDIA’s Transformer Engine supports FP8 mixed-precision
- FP8-compatible training framework
- Co-design between algorithm and infrastructure

Low-Precision Driven Design

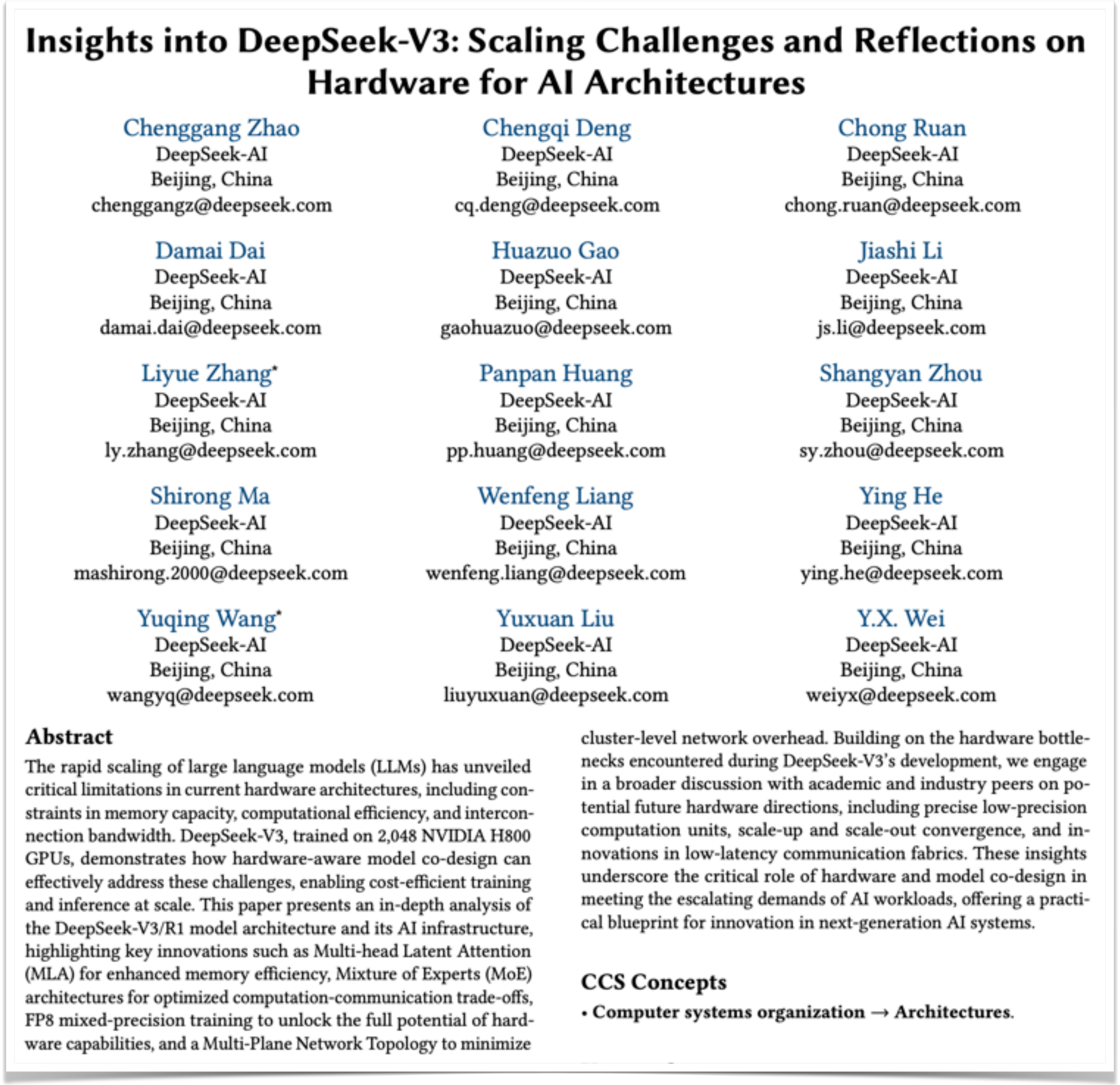
- FP8 mix-precision training: more data & high bandwidth
  - E4M3: higher precision; weights/activations
  - E5M2: low precision: gradients

- Challenges
  - Limited accumulation precision: FP22 (E8M13) addition registers
  - Fine-grained quantization: sub-tensor quantization parameters
  - Different parts have different parameters
  - Fine-grained de-quantization done by CUDA cores = data movement

Low-Precision Driven Design (cont'd)

- LogFMT: Communication Compression
  - Quantize to a dynamically-ranged log space
  - Step = (max-min) / [2^(n-1) - 2]
  - Using n=8 bits: better accuracy than E4M3/E5M2
  - Tokens dispatched using fine-grained FP8 quantization in EP parallelism
  - LogFMT-nBit: n is the number of bits with the leading 1 bit as the sign bit
  - n=8, better than E4M3 and E5M2
  - n=10, similar to BF16 combine

Insights into DeepSeek-V3



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

Chenggang Zhao
DeepSeek-AI
Beijing, China
chenggangz@deepseek.com

Chengqi Deng
DeepSeek-AI
Beijing, China
cq.deng@deepseek.com

Chong Ruan
DeepSeek-AI
Beijing, China
chong.ruan@deepseek.com

Damai Dai
DeepSeek-AI
Beijing, China
damai.dai@deepseek.com

Huazuo Gao
DeepSeek-AI
Beijing, China
gaohuazuo@deepseek.com

Jiashi Li
DeepSeek-AI
Beijing, China
js.li@deepseek.com

Liyue Zhang*
DeepSeek-AI
Beijing, China
ly.zhang@deepseek.com

Panpan Huang
DeepSeek-AI
Beijing, China
pp.huang@deepseek.com

Shangyan Zhou
DeepSeek-AI
Beijing, China
sy.zhou@deepseek.com

Shirong Ma
DeepSeek-AI
Beijing, China
mashirong.2000@deepseek.com

Wenfeng Liang
DeepSeek-AI
Beijing, China
wenfeng.liang@deepseek.com

Ying He
DeepSeek-AI
Beijing, China
ying.he@deepseek.com

Yuqing Wang*
DeepSeek-AI
Beijing, China
wangyq@deepseek.com

Yuxuan Liu
DeepSeek-AI
Beijing, China
liuyuxuan@deepseek.com

Y.X. Wei
DeepSeek-AI
Beijing, China
weiyx@deepseek.com

Abstract

The rapid scaling of large language models (LLMs) has unveiled critical limitations in current hardware architectures, including constraints in memory capacity, computational efficiency, and interconnection bandwidth. DeepSeek-V3, trained on 2,048 NVIDIA H800 GPUs, demonstrates how hardware-aware model co-design can effectively address these challenges, enabling cost-efficient training and inference at scale. This paper presents an in-depth analysis of the DeepSeek-V3/R1 model architecture and its AI infrastructure, highlighting key innovations such as Multi-head Latent Attention (MLA) for enhanced memory efficiency, Mixture of Experts (MoE) architectures for optimized computation-communication trade-offs, FP8 mixed-precision training to unlock the full potential of hardware capabilities, and a Multi-Plane Network Topology to minimize cluster-level network overhead. Building on the hardware bottlenecks encountered during DeepSeek-V3's development, we engage in a broader discussion with academic and industry peers on potential future hardware directions, including precise low-precision computation units, scale-up and scale-out convergence, and innovations in low-latency communication fabrics. These insights underscore the critical role of hardware and model co-design in meeting the escalating demands of AI workloads, offering a practical blueprint for innovation in next-generation AI systems.

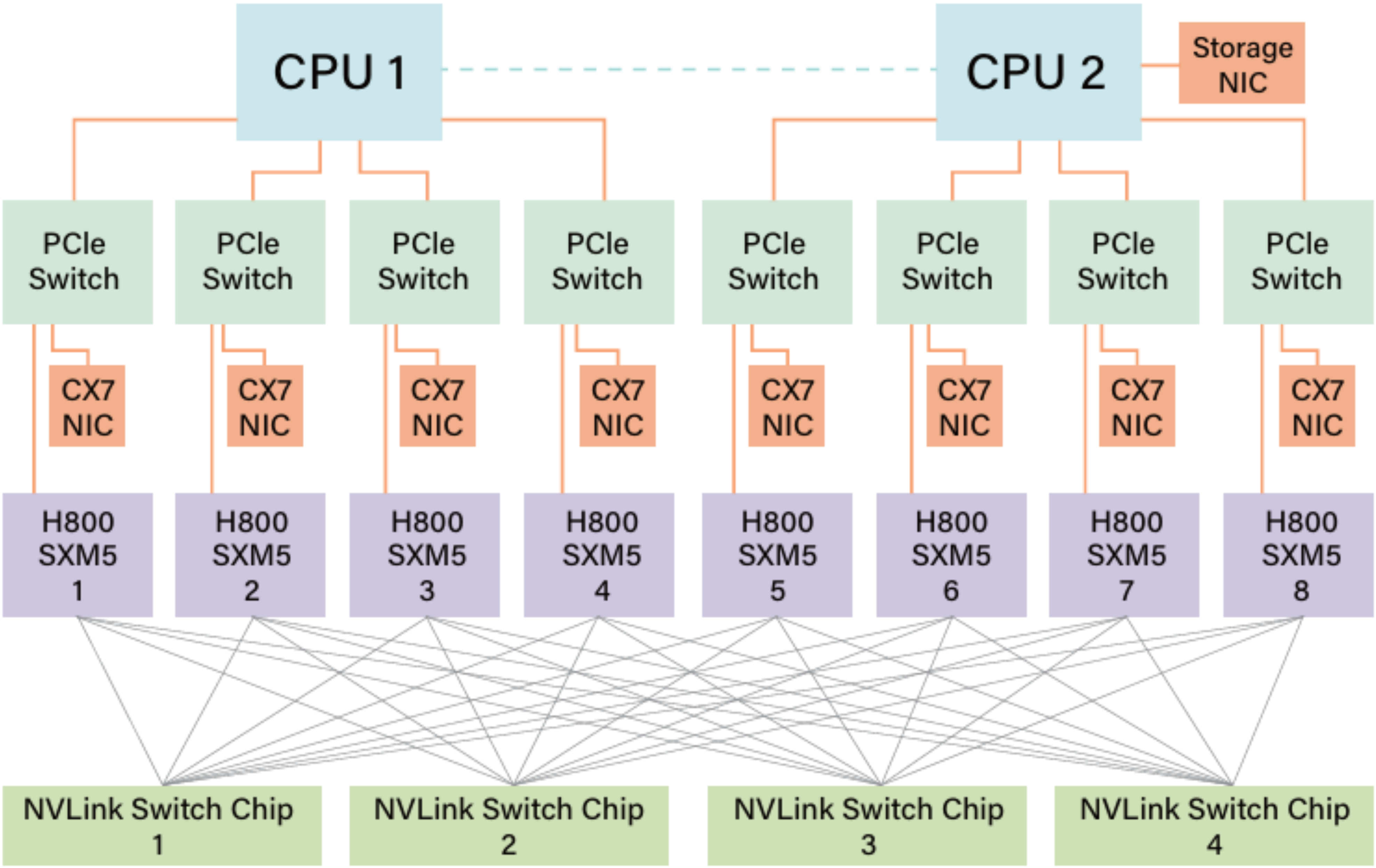
CCS Concepts

• **Computer systems organization → Architectures.**

- *DeepSeek-V3 Overview*
- *Low-Precision*
- Interconnect
- Cluster Network
- Looking Forward: Challenges

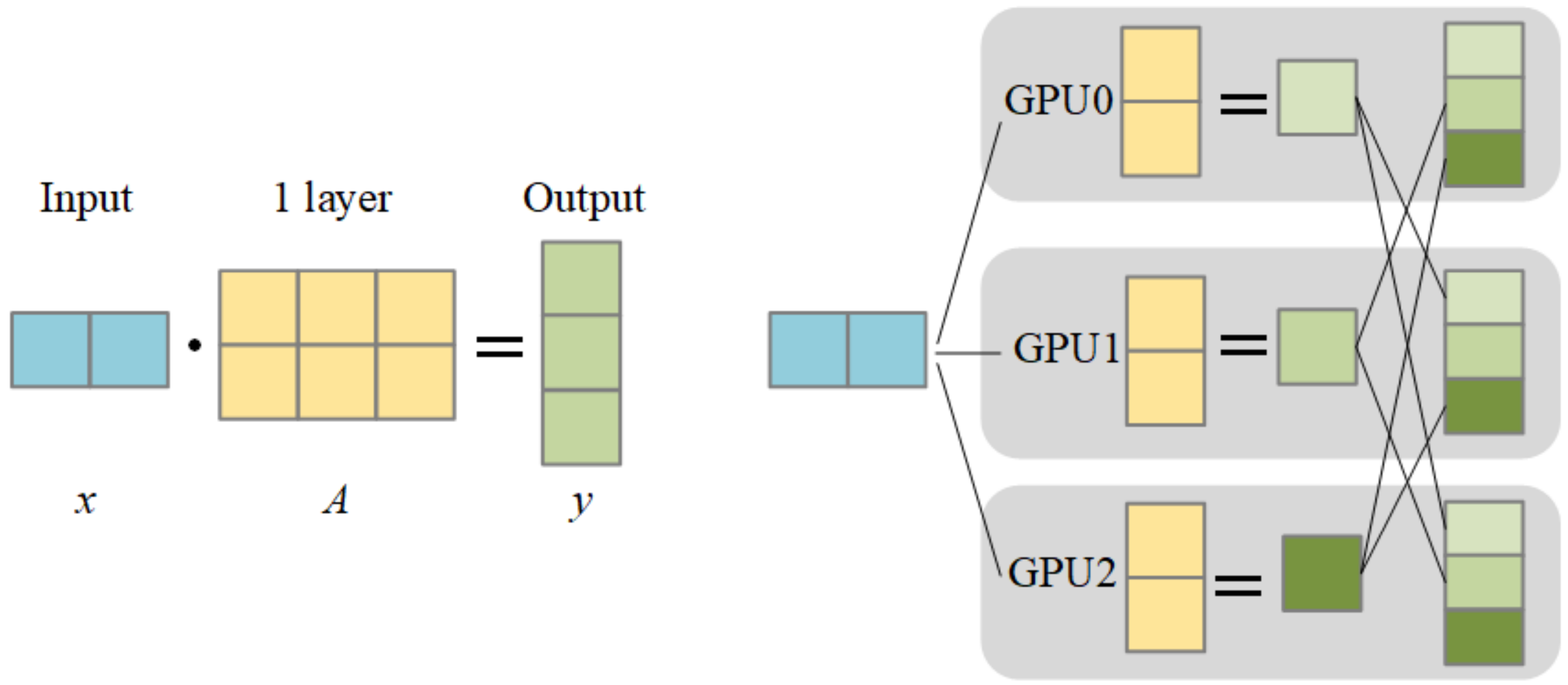
Hardware Architecture

- 2x CPU
- 8x NVIDIA H800
- NVLink BW: 400GB/s
- 8x IB CX7 NIC



NVLink Bandwidth is limited

- Avoid Tensor Parallelism (TP)
  - Disabled during training
  - Enabled during inference to reduce latency



NVLink Bandwidth is limited (cont'd)

- Enhanced Pipeline Parallelism (PP)
  - Dual Pipe
  - Overlap attention and MoE computation with MoE communication

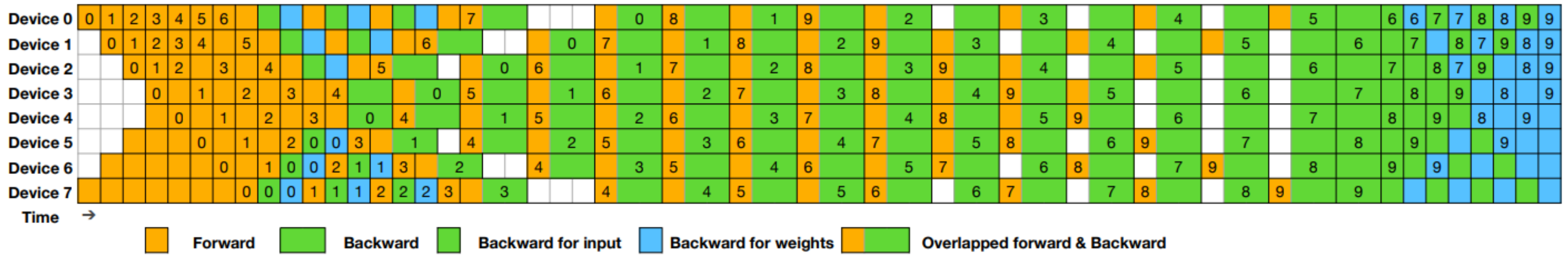
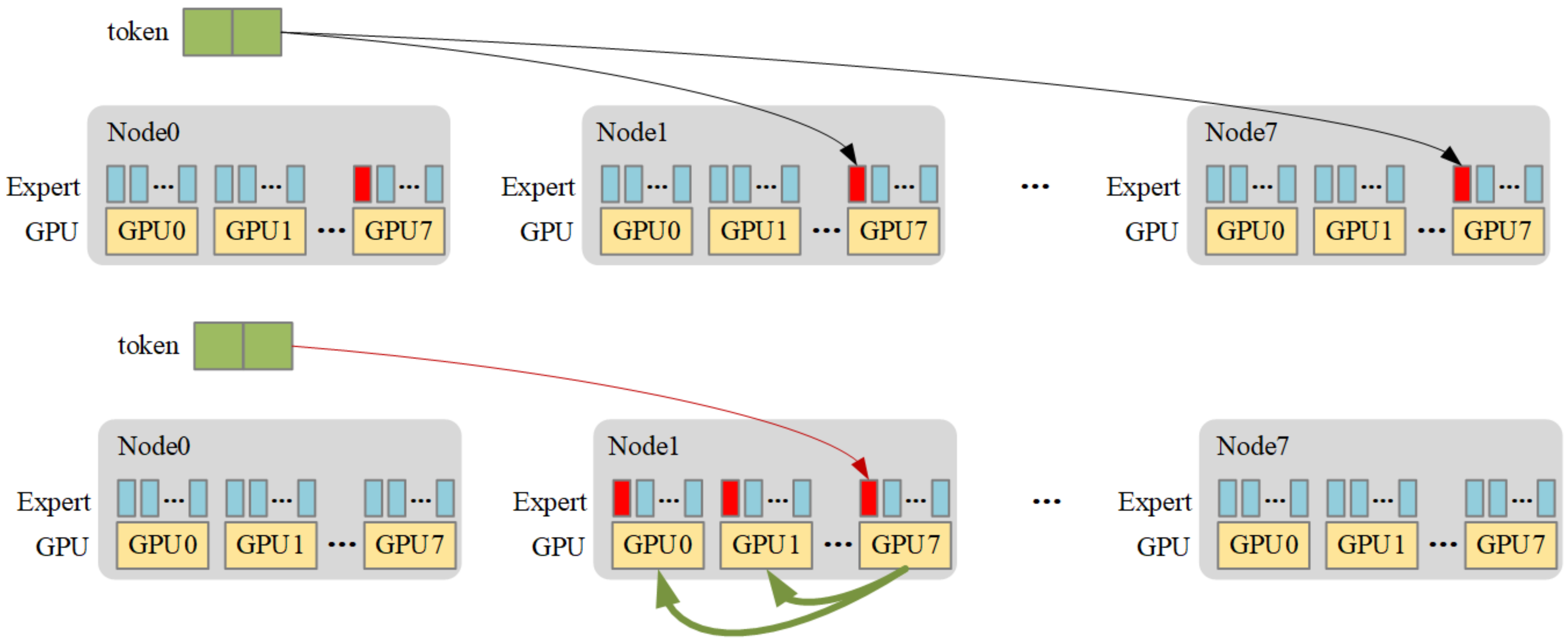


Figure 5 | Example DualPipe scheduling for 8 PP ranks and 20 micro-batches in two directions. The micro-batches in the reverse direction are symmetric to those in the forward direction, so we omit their batch ID for illustration simplicity. Two cells enclosed by a shared black border have mutually overlapped computation and communication.

- Accelerated expert parallelism (EP)
  - NIC x8, all-to-all 40GB/s
  - All-to-all EP implementation, DeepEP

Intra-node has higher bandwidth

- Bandwidth, intra-node v.s. inter-node = 160GB/s: 40GB/s
- Node-limited routing: up to 4 nodes



Fewer GPU SMs for computing

- Limitation: fewer GPU SM for computing
  - Network message handling (e.g., filling QPs and WQEs)
  - Data forwarding over NVLink
  - Up to 20 SMs

- Tasks can be offloaded to the NIC
  - Forwarding data, transport, reduce operations
  - Manage memory layout, data type cast

Fewer GPU SMs for computing

- Limitation: fewer GPU SM for computing
  - Network message handling (e.g., filling QPs and WQEs)
  - Data forwarding over NVLink

Ideally,

- Unified network adapter: Design NICs or I/O dies that are connected to unified scale-up and scale-out networks
- Dedicated communication co-processor: packet processing offloading
- Support flexible forward, broadcast, and reduce mechanisms
- Support hardware synchronization primitives

Cannot allocate BW of different traffics on NVLink and PCIe

- Inference
  - PCIe contention, loading KV cache + EP communication
- Ideally: Dynamic NVLink/PCIe traffic prioritization
  - Different priorities of traffic related to EP, TP, and KV cache transfers
- Ideally: I/O Die chiplet integration
  - Integrate NICs directly to the I/O die and connect them to the compute die
- Ideally: CPU-GPU interconnects within the scale-up domain
  - Connect CPUs and GPUs with NVLink or dedicated high-bw fabrics

Insights into DeepSeek-V3

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

Chenggang Zhao
DeepSeek-AI
Beijing, China
chenggangz@deepseek.com

Chengqi Deng
DeepSeek-AI
Beijing, China
cq.deng@deepseek.com

Chong Ruan
DeepSeek-AI
Beijing, China
chong.ruan@deepseek.com

Damai Dai
DeepSeek-AI
Beijing, China
damai.dai@deepseek.com

Huazuo Gao
DeepSeek-AI
Beijing, China
gaohuazuo@deepseek.com

Jiashi Li
DeepSeek-AI
Beijing, China
js.li@deepseek.com

Liyue Zhang*
DeepSeek-AI
Beijing, China
ly.zhang@deepseek.com

Panpan Huang
DeepSeek-AI
Beijing, China
pp.huang@deepseek.com

Shangyan Zhou
DeepSeek-AI
Beijing, China
sy.zhou@deepseek.com

Shirong Ma
DeepSeek-AI
Beijing, China
mashirong.2000@deepseek.com

Wenfeng Liang
DeepSeek-AI
Beijing, China
wenfeng.liang@deepseek.com

Ying He
DeepSeek-AI
Beijing, China
ying.he@deepseek.com

Yuqing Wang*
DeepSeek-AI
Beijing, China
wangyq@deepseek.com

Yuxuan Liu
DeepSeek-AI
Beijing, China
liuyuxuan@deepseek.com

Y.X. Wei
DeepSeek-AI
Beijing, China
weiyx@deepseek.com

Abstract

The rapid scaling of large language models (LLMs) has unveiled critical limitations in current hardware architectures, including constraints in memory capacity, computational efficiency, and interconnection bandwidth. DeepSeek-V3, trained on 2,048 NVIDIA H800 GPUs, demonstrates how hardware-aware model co-design can effectively address these challenges, enabling cost-efficient training and inference at scale. This paper presents an in-depth analysis of the DeepSeek-V3/R1 model architecture and its AI infrastructure, highlighting key innovations such as Multi-head Latent Attention (MLA) for enhanced memory efficiency, Mixture of Experts (MoE) architectures for optimized computation-communication trade-offs, FP8 mixed-precision training to unlock the full potential of hardware capabilities, and a Multi-Plane Network Topology to minimize cluster-level network overhead. Building on the hardware bottlenecks encountered during DeepSeek-V3's development, we engage in a broader discussion with academic and industry peers on potential future hardware directions, including precise low-precision computation units, scale-up and scale-out convergence, and innovations in low-latency communication fabrics. These insights underscore the critical role of hardware and model co-design in meeting the escalating demands of AI workloads, offering a practical blueprint for innovation in next-generation AI systems.

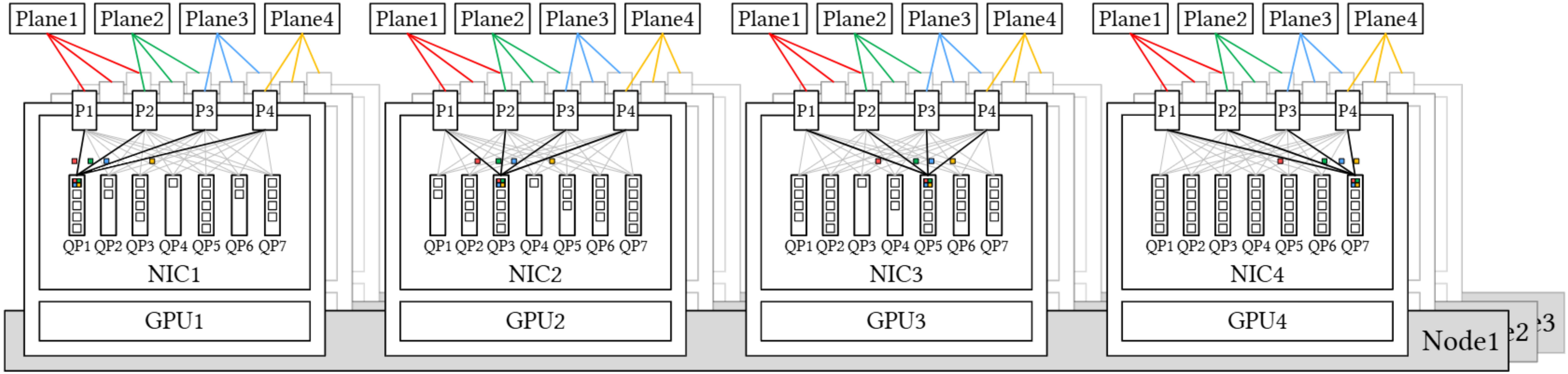
CCS Concepts

• **Computer systems organization → Architectures.**

- *DeepSeek-V3 Overview*
- *Low-Precision*
- *Interconnect*
- Cluster Network
- Looking Forward: Challenges

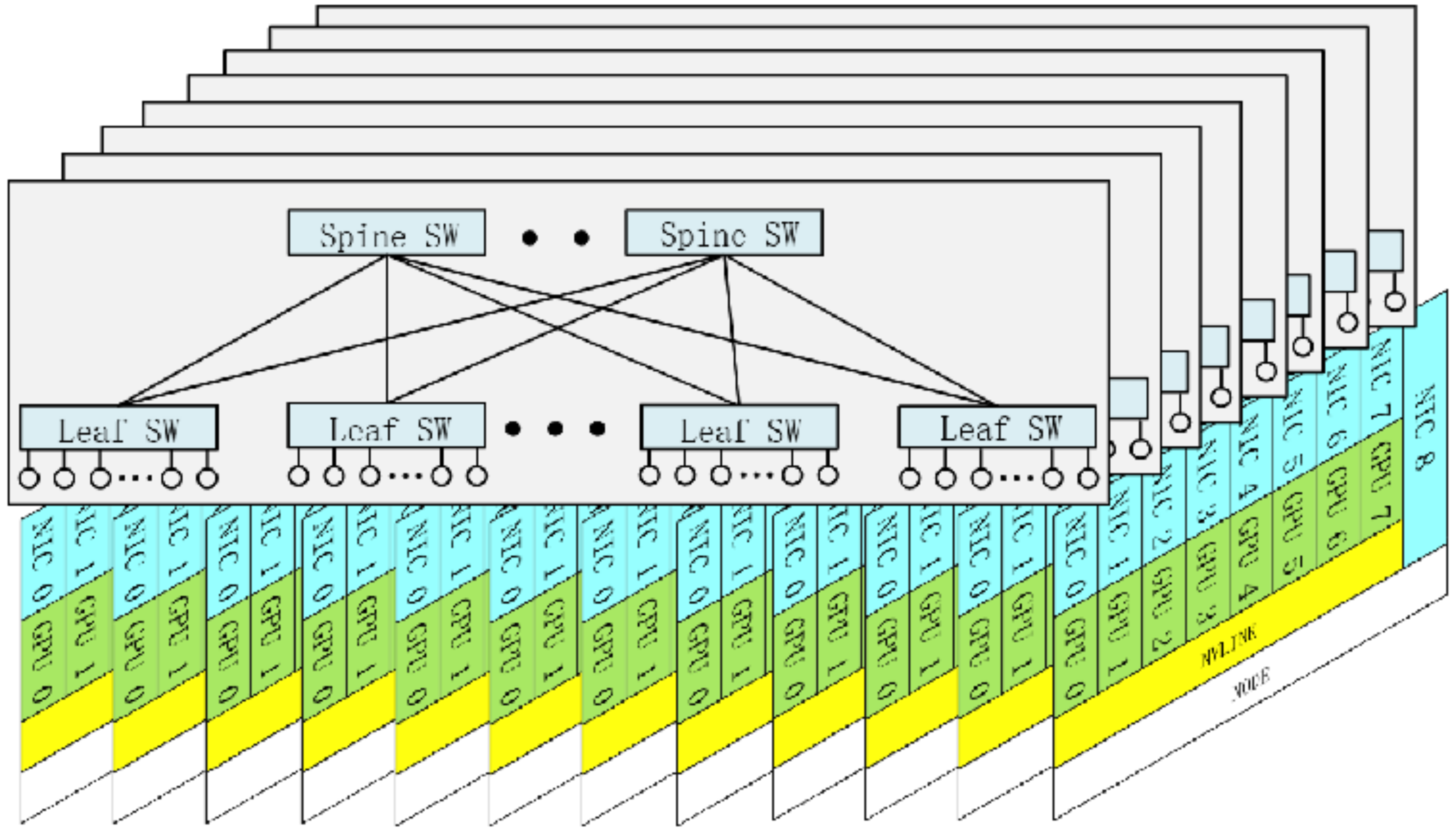
Network Topology

- Multi-plane



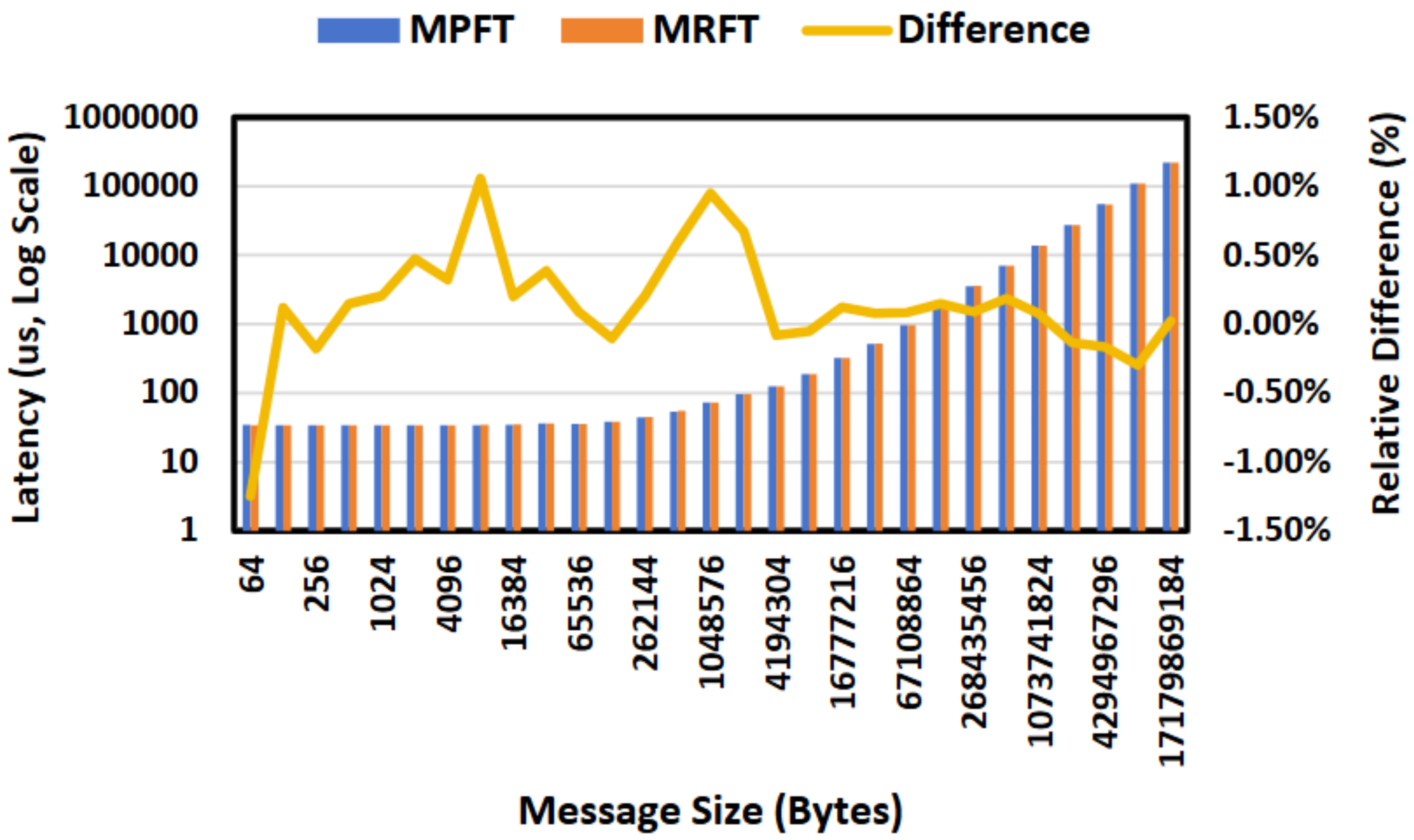
MPFT

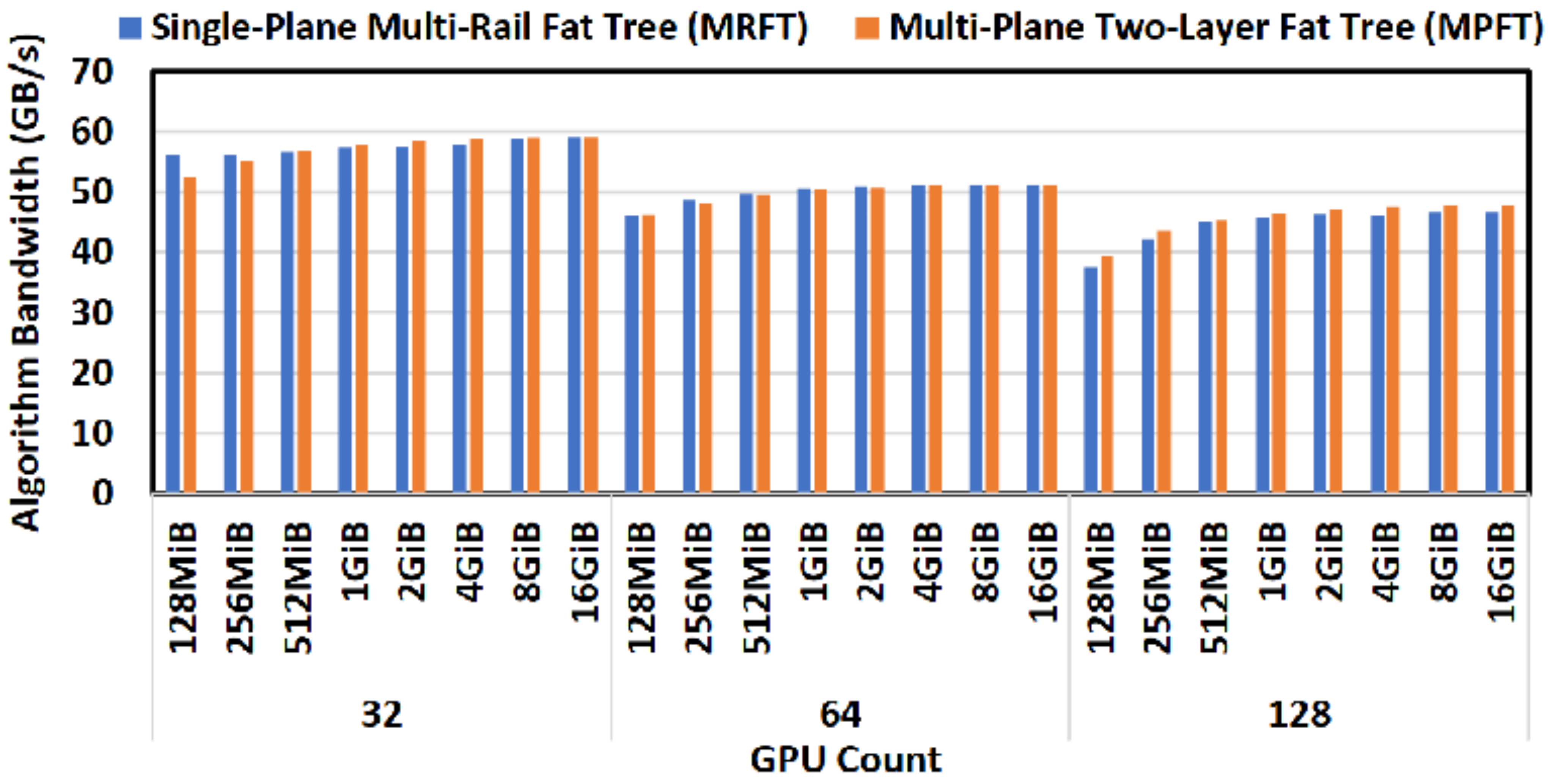
- Benefits
  - Subset of MRFT (NCCL works)
  - Cost
  - Traffic isolation
  - Latency reduction (low diameter)
  - Robustness (better with MP NIC)



Metric	FT2	MPFT	FT3	SF	DF
Endpoints	2,048	16,384	65,536	32,928	261,632
Switches	96	768	5,120	1,568	16,352
Links	2,048	16,384	131,072	32,928	384,272
Cost [M$]	9	72	491	146	1,522
Cost/Endpoint [k$]	4.39	4.39	7.5	4.4	5.8

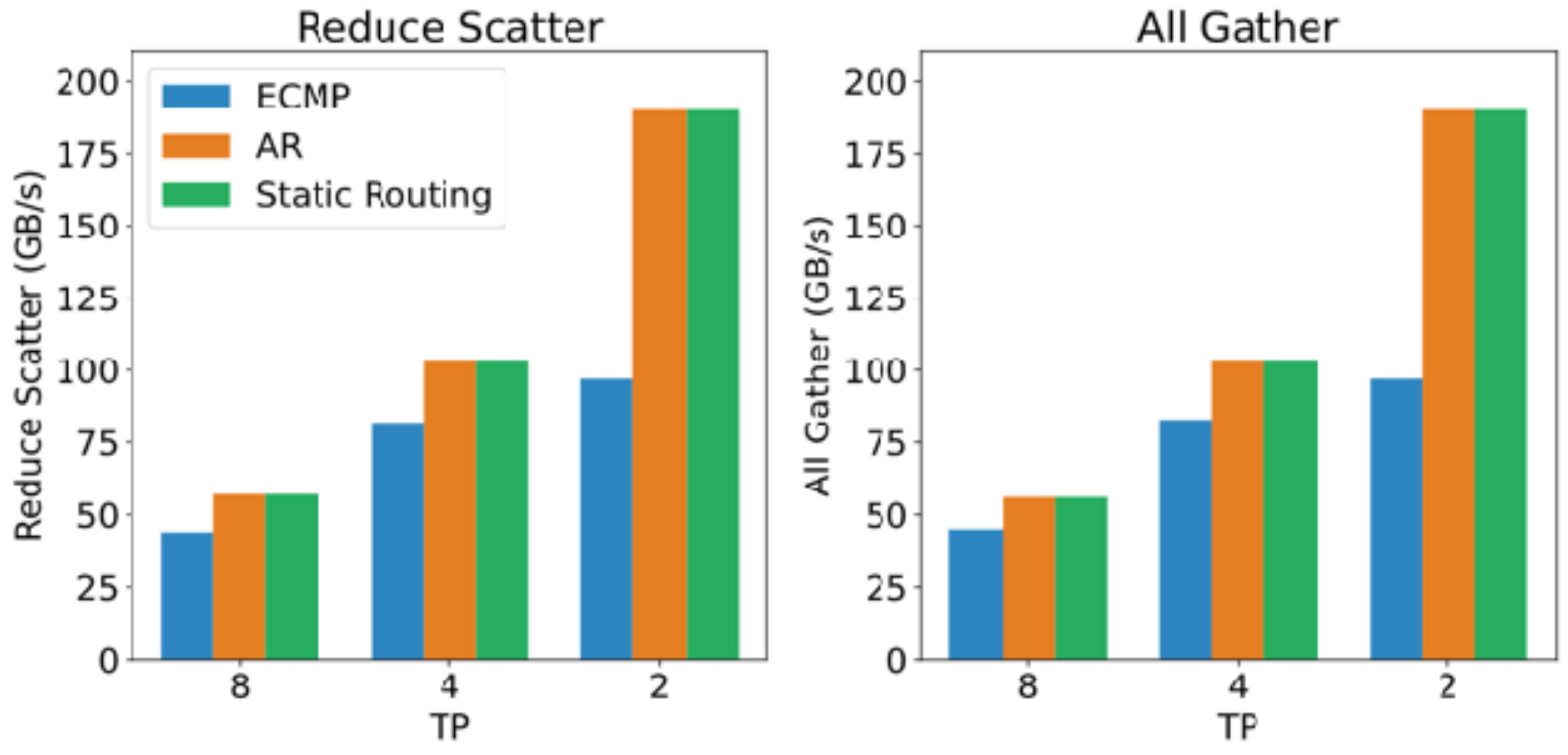
MPFT Performance





Low latency Networks

- IB over RoCE
  - Low latency
  - High cost
  - Low-radix switches
- RoCE potential improvements
  - Specialized RoCE switches for low latency
  - Adaptive routing
  - Improved CC
- CPU-GPU communication
  - GPUDirect (IBGDA)



Link Layer	Same Leaf	Cross Leaf
RoCE	3.6us	5.6us
InfiniBand	2.8us	3.7us
NVLink	3.33us	-

Insights into DeepSeek-V3

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

Chenggang Zhao
DeepSeek-AI
Beijing, China
chenggangz@deepseek.com

Chengqi Deng
DeepSeek-AI
Beijing, China
cq.deng@deepseek.com

Chong Ruan
DeepSeek-AI
Beijing, China
chong.ruan@deepseek.com

Damai Dai
DeepSeek-AI
Beijing, China
damai.dai@deepseek.com

Huazuo Gao
DeepSeek-AI
Beijing, China
gaohuazuo@deepseek.com

Jiashi Li
DeepSeek-AI
Beijing, China
js.li@deepseek.com

Liyue Zhang*
DeepSeek-AI
Beijing, China
ly.zhang@deepseek.com

Panpan Huang
DeepSeek-AI
Beijing, China
pp.huang@deepseek.com

Shangyan Zhou
DeepSeek-AI
Beijing, China
sy.zhou@deepseek.com

Shirong Ma
DeepSeek-AI
Beijing, China
mashirong.2000@deepseek.com

Wenfeng Liang
DeepSeek-AI
Beijing, China
wenfeng.liang@deepseek.com

Ying He
DeepSeek-AI
Beijing, China
ying.he@deepseek.com

Yuqing Wang*
DeepSeek-AI
Beijing, China
wangyq@deepseek.com

Yuxuan Liu
DeepSeek-AI
Beijing, China
liuyuxuan@deepseek.com

Y.X. Wei
DeepSeek-AI
Beijing, China
weiyx@deepseek.com

Abstract

The rapid scaling of large language models (LLMs) has unveiled critical limitations in current hardware architectures, including constraints in memory capacity, computational efficiency, and interconnection bandwidth. DeepSeek-V3, trained on 2,048 NVIDIA H800 GPUs, demonstrates how hardware-aware model co-design can effectively address these challenges, enabling cost-efficient training and inference at scale. This paper presents an in-depth analysis of the DeepSeek-V3/R1 model architecture and its AI infrastructure, highlighting key innovations such as Multi-head Latent Attention (MLA) for enhanced memory efficiency, Mixture of Experts (MoE) architectures for optimized computation-communication trade-offs, FP8 mixed-precision training to unlock the full potential of hardware capabilities, and a Multi-Plane Network Topology to minimize cluster-level network overhead. Building on the hardware bottlenecks encountered during DeepSeek-V3's development, we engage in a broader discussion with academic and industry peers on potential future hardware directions, including precise low-precision computation units, scale-up and scale-out convergence, and innovations in low-latency communication fabrics. These insights underscore the critical role of hardware and model co-design in meeting the escalating demands of AI workloads, offering a practical blueprint for innovation in next-generation AI systems.

CCS Concepts

• **Computer systems organization → Architectures.**

- *DeepSeek-V3 Overview*
- *Low-Precision*
- *Interconnect*
- *Cluster Network*
- Looking Forward: Challenges

Challenge #1: Reliability

- Limitations:
  - Interconnect failures
  - Single hardware failures
  - Silent data corruption

- Suggestions
  - Built-in reliability support
  - Hardware vendors provide diagnostic toolkits

Challenge #2: Fast CPU and Memory

- Limitations:
  - CPU: handle the PCIe transactions
  - High memory bandwidth given PCIe Gen5/6
  - Kernel launch and network processing still rely on the CPU

- Suggestions
  - General-purpose computation should also be fast in a balanced system

Challenge #3: Networking in the AI Infrastructure

- Call for actions:
    - Co-packaged optics —> The opportunity for silicon photonics comes
    - Lossless network based on credit
    - Adaptive routing with a controlled and fast feedback
    - Fault-tolerance, including detection, localization, and fix
    - Inferences need better networking resource management

Challenge #4: Memory-Semantic Communication

- Call for actions:
  - Load-store is strongly needed
  - Transparent ordering from networking gears

Challenge #5: In-Network Computation and Compression

- Call for actions:
  - Compute-network co-design

Challenge #6: Memory-Centric Innovations

- Limitations:
  - Memory bandwidth is limited

- Suggestions:
  - High-bandwidth
  - System-on-Wafer

Summary

- Please teach us!
- We are happy to learn from you!
- See you on Thursday and Tuesday