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

Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Huazuo Gao, Jiashi Li, Liyue Zhang, Panpan Huang, Shangyan Zhou, Shirong Ma, Wenfeng Liang, Ying He, Yuqing Wang, Yuxuan Liu, Y.X. Wei
DeepSeek-AI
Beijing, China

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**.

Keywords

Large Language Model, Mixture-of-Experts, Deep Learning, FP8 Mixed-Precision Training, Multi-Plane Network, Co-Design

ACM Reference Format:
Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Huazuo Gao, Jiashi Li, Liyue Zhang, Panpan Huang, Shangyan Zhou, Shirong Ma, Wenfeng Liang, Ying He, Yuqing Wang, Yuxuan Liu, Y.X. Wei . 2025. Insights into DeepSeek-V3: Scaling Challenges and Reflections on Hardware for AI Architectures. In *Proceedings of the 52nd Annual International Symposium on Computer Architecture (ISCA '25), June 21–25, 2025, Tokyo, Japan.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3695053.3731412

ISCA '25, June 21–25, 2025, Tokyo, Japan

ACM ISBN 979-8-4007-1261-6/2025/06
https://doi.org/10.1145/3695053.3731412

1 Introduction

1.1 Background

Large Language Models (LLMs) have undergone rapid evolution in recent years, driven by iterative advancements in model design computational power, and data availability. In 2024, groundbreaking models such as GPT4o [59], LLaMa-3 [3], Claude 3.5 Sonnet [8] Grok-2 [73], Qwen2.5 [75], Gemini-2 [37] and our DeepSeek-V3 [26] have showcased remarkable progress, further narrowing the gap towards Artificial General Intelligence (AGI). As the Scaling Laws [45] shows, increasing model size, training data, and computational resources leads to substantial improvements in model performance underscoring the pivotal role of scaling in advancing AI capabilities Collectively, these developments have ushered in an era where scaling model size and computational power is seen as the key to unlocking higher levels of intelligence.

Recent developments, reasoning models such as OpenAI's o1/o3 series models [60, 61], DeepSeek-R1 [28], Claude-3.7 Sonnet [9] Gemini 2.5 Pro [38], Seed1.5-Thinking [68] and Qwen3 [71] have demonstrated not only the benefits conferred by large-scale architectures, but also the necessity of improving inference efficiency particularly in handling longer contexts and achieving greater reasoning depth. These advancements underscore the need for faster and more efficient inference, consequently placing ever-increasing demands on computational resources.

To meet these challenges, industry leaders such as Alibaba ByteDance, Google, xAI and Meta have deployed colossal training clusters [33, 42, 43, 56, 62, 74], featuring tens or even hundreds of thousands of GPUs or TPUs. While such massive infrastructures have enabled the development of state-of-the-art models, their exorbitant costs present significant barriers for smaller research teams and organizations. Despite these barriers, open-source startups such as DeepSeek [23–26, 28] and Mistral [41, 55] are also striving to develop state-of-the-art models. Among them, DeepSeek has especially demonstrated that effective software-hardware co-design can enable cost-efficient training of large models, leveling the playing field for smaller teams.

Building on this tradition, DeepSeek-V3 [26] represents a new milestone in cost-effective training. By leveraging just 2,048 NVIDIA H800 GPUs, DeepSeek-V3 achieves state-of-the-art performance This achievement aligns with the commitment to advance AI through practical and scalable solutions, as previously demonstrated in the cost-effective architecture of Fire-Flyer AI-HPC [7]. The practices and insights derived from DeepSeek-V3 demonstrate how existing hardware resources can be harnessed to their fullest potential offering valuable lessons for the broader AI and HPC communities

Authors are listed in alphabetical order of their first names. Yuqing Wang and Liyue Zhang are the corresponding authors of this p om)