# **关于DeepSeek爆火现象的思考**

原创 Eagle 当下瞬间进入历史 *2025年02月11日 15:32* *北京*

2024年圣诞节后一天DeepSeek V3发布，三周以后推理版本R1发布，先是引起业内热议，然后似乎突然之间成为线上线下，政府会议，街头巷尾，甚至大洋两岸的热门话题，完全出乎意料，火爆程度堪比两年前ChatGPT问世时的景象。全球股票市场也做出不同的应激反应，老美那边的智算巨头Nvidia股价应声大跌17%，而国内各路科技股乘风狂欢。

在这次爆火之前，DeepSeek团队相对低调。他们第一次为大众所知可能是在去年中发布V2版本时，因为设定的服务价格只有不同竞争对手的十分之一到百分之一，被称为AI界的的“拼多多”。

在V3版本发布后几天，传言雷军千万年薪挖角DeepSeek团队的95后工程师罗福莉，也为DeepSeek快速进入大众视野点了一把火。R1版本的接续发布，性能逼近OpenAI最先进推理模型o1，叠加上DeepSeek训练集群只有2000块阉割版的H800（对比Llama3的16000块满血的H100训练集群），DeepSeek领导梁文锋被邀请参加总理座谈会，受到中美媒体关注，这个年轻的团队最终引发海啸般的赞誉和关注。

**DeepSeek的成功要素**

DeepSeek的成功主要是在AI行业军备竞赛趋于白热化，地缘政治日趋紧张的背景下，通过扎实的技术创新和充满勇气的开放，赢得了行业内外的一致尊敬。

**DeepSeek的技术创新**

*基础设施方向*

最早关注到DeepSeek是他们一篇AI基础设施优化的文章（Fire-Flyer AI-HPC），当时一直在思考AI Infra领域的问题和机会。

他们在10000块PCIe连接的A100集群上通过软硬件联合优化设计，取得了接近Nvidia DGX集群的性能，而成本降低一半。通过自研底层通信和调度库，在PCIe连接的A100系统上性能超过了Nvidia NCCL库，减少了对超高速NVlink连接的依赖。这个工作也成为他们DeepSeek V3的成功的基础之一。

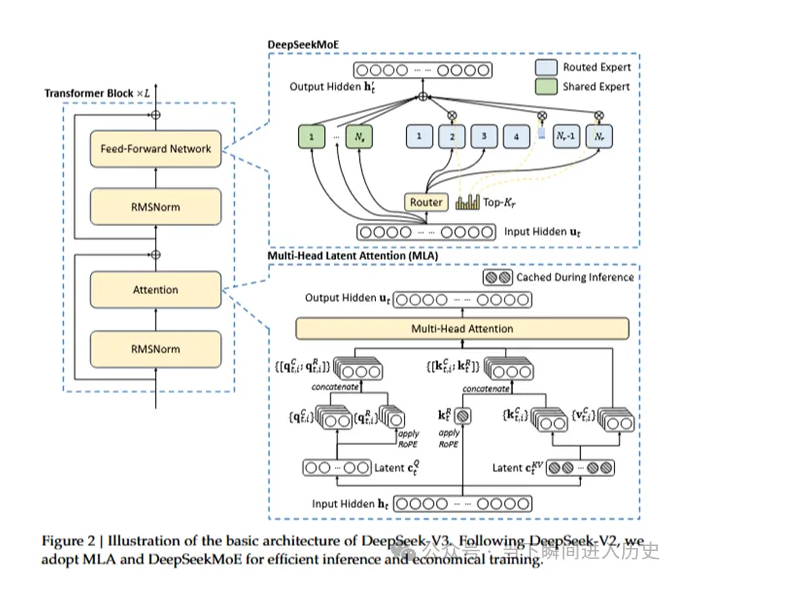
V3训练使用的GPU H800互联带宽被大幅阉割，他们通过复杂精细的系统联合优化设计，在只有八分之一算力规模的集群（2000 H800 vs Llama 16000 H100）上取得领先的性能。正是这一成功背后的成本曲线改变，直接激起了后来资本市场的巨震。

V3集群优化的流水线并行（Pipeline Parallelism）和专家并行（Expert Parallelism）算法都是大型复杂分布式系统建设的难题，需要对底层硬件拓扑架构和上层负载特性有深刻的理解并能设计有效的方案。

DeepSeek团队在这些系统领域的成功为中国AI行业突破芯片卡脖子困难带来了新的希望。一方面可以利用对上层大模型复杂特性的深刻理解改善芯片系统的设计，另一个方面是用算力性能低的芯片构建大规模集群系统。当然这是一个超级系统工程，需要整个技术栈生态联合协作，目前DeepSeek还是基于Nvidia平台实现的系统，移植国产平台需要大量基础工作，如何有效的推进并不简单是一个技术问题。

*模型方向*

DeepSeek模型方向的创新主要是对注意力机制的优化（MLA）和专家模型的优化（DeepSeekMoE）。这两个优化更多还是偏向系统层面的优化，目的在于提高效率，对于模型的智能水平提升并不直接产生贡献，间接的贡献在于成本降低可以支持将更多的算力用于模型扩展（Scalling）。

DeepSeek最重要的创新是MLA，它利用矩阵降秩计算大幅降低算力资源使用，考虑到降秩可能会影响训练性能，在预训练阶段很少有人使用，是微调阶段常用的办法，Stable Diffusion改善Diffusion模型的图像生成性能是类似的思路。

DeepSeek在训练阶段还采用了复杂精细的混合量化方法和大量优化计算内存通信平衡的算法，整个系统的工程优化程度让人惊叹。这些优化可能和他们面临的资源限制直接相关，也显示了团队深厚的工程实力。

美国主要头部玩家可能受Rich Sutton“The bitter lesson”的影响倾向于用简单通用的方向，通过增加系统规模实现模型的提升。

DeepSeek的优化算法紧密耦合于他们的模型架构和硬件系统设计，是定制的特定算法。目前他们在论文中多数给出了详细的说明，其他团队可以借鉴，这也收到了业内专家的称赞，这些算法能否抽象成更加通用的方法给更多人使用是更高的标准，这个方面Meta贡献的PyTorch是个典范。

DeepSeek R1推理模型探索了利用强化学习RL增强模型推理能力，性能接近OpenAI o1模型能力，发布时间也只落后4个月，这个时间比大家普遍认为的AI中美差距2年大为缩短，可能对AI方向的地缘政治影响深远。Anthropic CEO公开发文呼吁进一步加强GPU出口管控，限制推理型芯片H20的出口，而国内也有人喊出DeepSeek是国运级的科技突破。

RL用于大模型能力提升尤其是推理能力提升是这两年的热门话题，从OpenAI神秘的传说项目Q\*开始就牵动人心。R1试验了从基础模型直接应用强化学习的可行性，为减少资源消耗没有采用同等规模的批评模型（Critic Model）而代以直接评估一组采样来优化（GRPO），采用基于规则的奖励模型，总体取得了很好的性能，尤其在编码和数学等规则明确的领域非常成功。

R1另一个可贵的实验是将模型优秀的推理能力成功的蒸馏（Distill）到不同的模型中，为未来模型能力的组合开辟了路径。

**DeepSeek成功的影响**

DeepSeek开放了所有模型的权重，对模型使用采用了最友好的MIT许可，也用两篇论文公开了主要的创新方法，透明度和Llama3的文章同为业内专家赞赏，模型许可开放性超过了Llama3的水平。开源模型的每一次进步可能都会重塑整个行业的模型生态，以至于OpenAI CEO Sam Altman有点怀疑选择闭源的决策是否正确。

DeepSeek模型的大量优化也将让部署模型所需资源大幅降低，会直接推动下游应用的成本降低和场景扩大，这些变化会推动推理场景的资源需求增长。训练成本曲线的变化既会改变训练算力供应的需求，也可能加速下一代模型的开发进程。

**未来的发展**

老美领先的玩家都在致力于下一代模型乃至AGI模型的开发，几大巨头今年的资本开支都接近或达到千亿美元级别，OpenAI牵头的StarGate项目计划4年投入5000亿美元就是典型代表。模型训练集群正在从10万卡向百万卡爬升，GW级的数据中心建设正从蓝图走向现实中。

DeepSeek代表的极致优化系统的方法能走多远，算力卡脖子问题是否已经消失都不是确定答案。DeepSeek团队年轻人身上展示出来的青春勇气确实让人看到中国年轻一代的希望，全球AI领域也到处是中国年轻工程师的身影，希望国内有更多的年轻人和团队不断涌现出来，行业少一点浮躁和投机炒作，更多扎实的创新，我们就可以对未来更乐观。

​