



deepseek

DS 开源周 回顾与思考



ZOMI

Context

1. DeepSeek 一周开源情况汇总
2. DeepSeek 开源周对业界思考



01

一周开源情况汇总



一周开源情况汇总

项目名称	GitHub地址	简介	应用领域	优化分类	技术意义	引用技术
FlashMLA	https://github.com/deepsseek-ai/FlashMLA	专为Hopper架构优化的MLA Kernel, 支持变长序列	MoE 推理加速 (长文本生成、对话系统)	推理	降低高性能推理门槛, 推动AI普惠化, 尤其适合中小开发者低成本部署	融合FA 2/3 注意力优化与cutlass硬件适配, 针对Hopper Tensor Core定制
DeepEP	https://github.com/deepsseek-ai/DeepEP	MoE 专用通信库, 支持EP 与FP8 低精度通信, 提供高吞吐、低延迟的GPU Kernel 实现	MoE 训练与分布式推理	训练与推理	突破MoE模型的通信瓶颈, 推动千亿级模型实用化部署	基于 NVSHMEM 优化通信协议, 结合自研低精度专家分发算法
DeepGEMM	https://github.com/deepsseek-ai/DeepGEMM	高性能 FP8 矩阵运算库, Hopper 峰值性能1350+ TFLOPS, 显存占用较FP16减少50%	MoE 架构计算优化、边缘设备轻量化部署	训练与推理	推动行业向低精度计算迁移, 解决千亿模型内存墙问题	集成Hopper 架构 FP8 Tensor Core 指令集, 动态量化策略平衡精度与效率
DualPipe	https://github.com/deepsseek-ai/DualPipe	双向流水线并行框架, 通过计算与通信重叠减少流水线气泡, 提升训练效率	超大规模 MOE 训练 (DeepSeek-V3 128K上下文)	训练	解决流水线并行中的资源闲置问题, 显著降低训练成本	创新流水线调度算法, 结合跨节点全对全通信优化
3FS	https://github.com/deepsseek-ai/3FS	分布式训练存储优化方案, 支持高效数据分片与缓存管理	超大规模训练数据加载加速、分布式检查点存储	训练	缓解训练I/O瓶颈, 提升GPU集群利用率	推测采用分阶段数据预加载与内存映射技术



Day1: FlashMLA



DeepSeek  @deepseek_ai · Feb 24



 Day 1 of [#OpenSourceWeek](#): FlashMLA

Honored to share FlashMLA - our efficient MLA decoding kernel for Hopper GPUs, optimized for variable-length sequences and now in production.

- ✓ BF16 support
- ✓ Paged KV cache (block size 64)
- ⚡ 3000 GB/s memory-bound & 580 TFLOPS

[Show more](#)

 What is FlashMLA's optimization method?

 515

 1.8K

 10K

 1.5M



Day2: DeepEP



DeepSeek  @deepseek_ai · Feb 25



 Day 2 of [#OpenSourceWeek](#): DeepEP

Excited to introduce DeepEP - the first open-source EP communication library for MoE model training and inference.

- ✓ Efficient and optimized all-to-all communication
- ✓ Both intranode and internode support with NVLink and RDMA
- ✓

[Show more](#)

 471

 1.4K

 8.4K

 1.2M



Day3: DeepGEMM






DeepSeek  @deepseek_ai · Feb 26



 Day 3 of [#OpenSourceWeek](#): DeepGEMM

Introducing DeepGEMM - an FP8 GEMM library that supports both dense and MoE GEMMs, powering V3/R1 training and inference.

-  Up to 1350+ FP8 TFLOPS on Hopper GPUs
-  No heavy dependency, as clean as a tutorial
-  Fully Just-In-Time compiled

[Show more](#)

 415

 1.5K

 6.5K

 850K



Day4: DualPipe



DeepSeek  @deepseek_ai · Feb 27



 Day 4 of [#OpenSourceWeek](#): Optimized Parallelism Strategies

✅ DualPipe - a bidirectional pipeline parallelism algorithm for computation-communication overlap in V3/R1 training.

 [github.com/deepseek-ai/Du...](#)

✅ EPLB - an expert-parallel load balancer for V3/R1.



[Show more](#)

deepseek-ai/




Day5: 3FS




DeepSeek  @deepseek_ai · Feb 28



 Day 5 of [#OpenSourceWeek](#): 3FS, Thruster for All DeepSeek Data Access

Fire-Flyer File System (3FS) - a parallel file system that utilizes the full bandwidth of modern SSDs and RDMA networks.

 6.6 TiB/s aggregate read throughput in a 180-node cluster

 3.66 TiB/min

[Show more](#)

 445

 1.7K

 10K

 2.5M



Day6: Inference system Overview

**DeepSeek**  @deepseek_ai · Mar 1  

 Day 6 of #OpenSourceWeek: One More Thing – DeepSeek-V3/R1 Inference System Overview

Optimized throughput and latency via:

-  Cross-node EP-powered batch scaling
-  Computation-communication overlap
-  Load balancing

Statistics of DeepSeek's Online Service:

 73.7k/14.8k

[Show more](#)

 How is profit margin calculated?


How will DeepSeek expand margin

 569

 1.5K

 7.9K

 2.5M



一周开源情况汇总

项目名称	GitHub地址	简介	应用领域	优化分类	技术意义	引用技术
FlashMLA	https://github.com/deepsseek-ai/FlashMLA	专为Hopper架构优化的MLA Kernel, 支持变长序列	MoE 推理加速 (长文本生成、对话系统)	推理	降低高性能推理门槛, 推动AI普惠化, 尤其适合中小开发者低成本部署	融合FA 2/3 注意力优化与cutlass硬件适配, 针对Hopper Tensor Core定制
DeepEP	https://github.com/deepsseek-ai/DeepEP	MoE 专用通信库, 支持EP 与FP8 低精度通信, 提供高吞吐、低延迟的GPU Kernel 实现	MoE 训练与分布式推理	训练与推理	突破MoE模型的通信瓶颈, 推动千亿级模型实用化部署	基于 NVSHMEM 优化通信协议, 结合自研低精度专家分发算法
DeepGEMM	https://github.com/deepsseek-ai/DeepGEMM	高性能 FP8 矩阵运算库, Hopper 峰值性能1350+ TFLOPS, 显存占用较FP16减少50%	MoE 架构计算优化、边缘设备轻量化部署	训练与推理	推动行业向低精度计算迁移, 解决千亿模型内存墙问题	集成Hopper 架构 FP8 Tensor Core 指令集, 动态量化策略平衡精度与效率
DualPipe	https://github.com/deepsseek-ai/DualPipe	双向流水线并行框架, 通过计算与通信重叠减少流水线气泡, 提升训练效率	超大规模 MOE 训练 (DeepSeek-V3 128K上下文)	训练	解决流水线并行中的资源闲置问题, 显著降低训练成本	创新流水线调度算法, 结合跨节点全对全通信优化
3FS	https://github.com/deepsseek-ai/3FS	分布式训练存储优化方案, 支持高效数据分片与缓存管理	超大规模训练数据加载加速、分布式检查点存储	训练	缓解训练I/O瓶颈, 提升GPU集群利用率	推测采用分阶段数据预加载与内存映射技术



02

开源周对业界思考



Context

1. 大模型算法演进的改变
2. 国产芯片厂商春天与寒冬
3. 大模型厂商竞争加剧
4. MaaS 平台服务 & 转型



1、大模型算法演进的变化

1. **打破 AI 星际之门**：让 OpenAI 不再神话，走出国产独立自主的开源路线，让 QWEN 等大厂不再走闭源路线。让 AI 走进千行百业，推动大模型在更多领域应用。
2. **FP8 计算生态崛起**：DeepGEMM 支持 FP8，Hopper 架构实现 1350+ TFLOPS 性能，推动大模型算法向低精度计算迁移，显存占用减少的同时提升吞吐量。
3. **MoE 模型落地加速**：DualPipe、DeepEP 等加速训练，解决了 MOE 计算效率问题，未来 Scaling Law 可能全面转向 MoE 架构。
4. **从堆料回归到算法创新**：为大模型算法的创新和优化提供了新的思路和方法，聚焦 Infra 层的加速提升算力利用率。Scaling Law 的新趋势。



1、大模型算法演进的改变

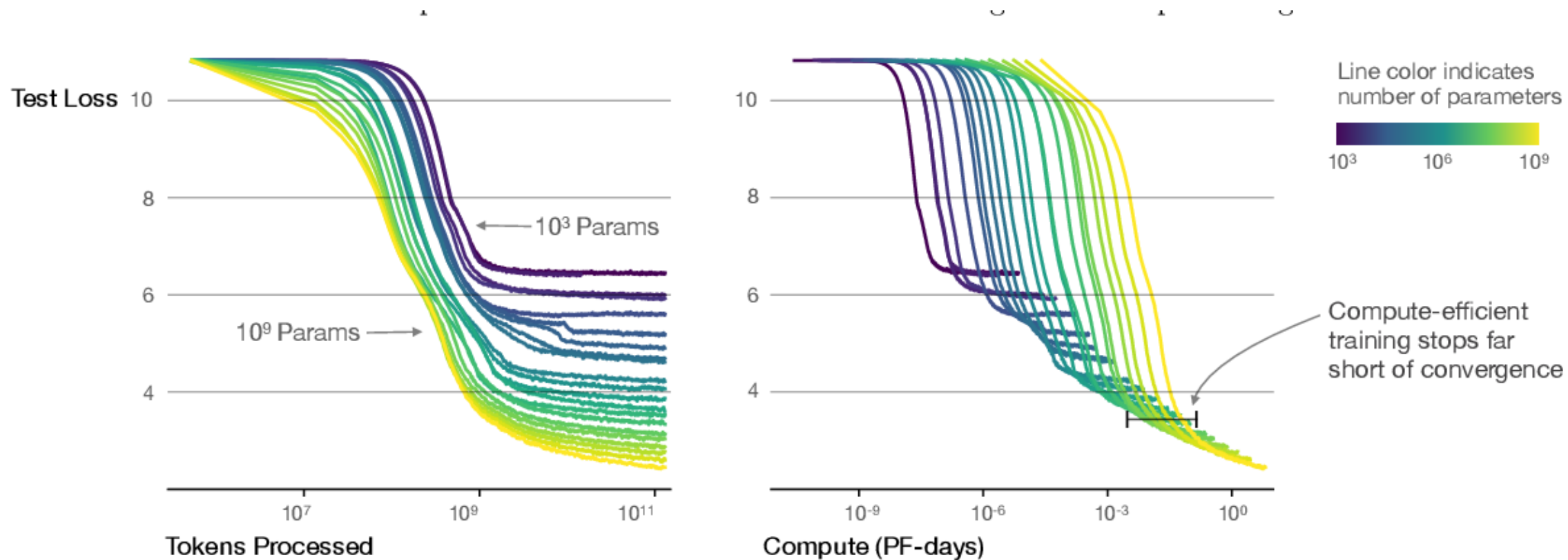


Figure 2: We show a series of training runs with model sizes ranging from 10^3 to 10^9 .

2、国产芯片厂商春天与寒冬

1. **短期算力投资逻辑转变：** DS 低成本降低对万卡/十万卡 AI 集群建设需求。长期来看，分布式推理需求推动边缘计算和专用芯片（ASIC to MOE、RISC-V）发展，为国产芯片厂商提供了新机遇。
2. **国产芯片加速布局：** 迫使国产玩家积极优化推理侧解决方案（PD 分离、EP 并行、MTP 优化）；需要同更低精度格式 FP8；如何做好开源开放（如 PTX、NVSHMEM、CUDA），让生态伙伴（如 DS 等）参与。
3. **市场需求与份额增长：** 打开星际之门 🚪，使得更多企业和开发者能够参与到 AI 应用的开发中，从而带动了对国产芯片的需求增长。
4. **技术创新与升级：** 暴露了现有国产 AI 芯片设计缺陷（FP8、GPU Direct），促使国产芯片厂商借鉴，重新设计内部计算单元和通信总线，提升芯片性能和效率，激发国产芯片厂商在 AI 芯片领域创新。

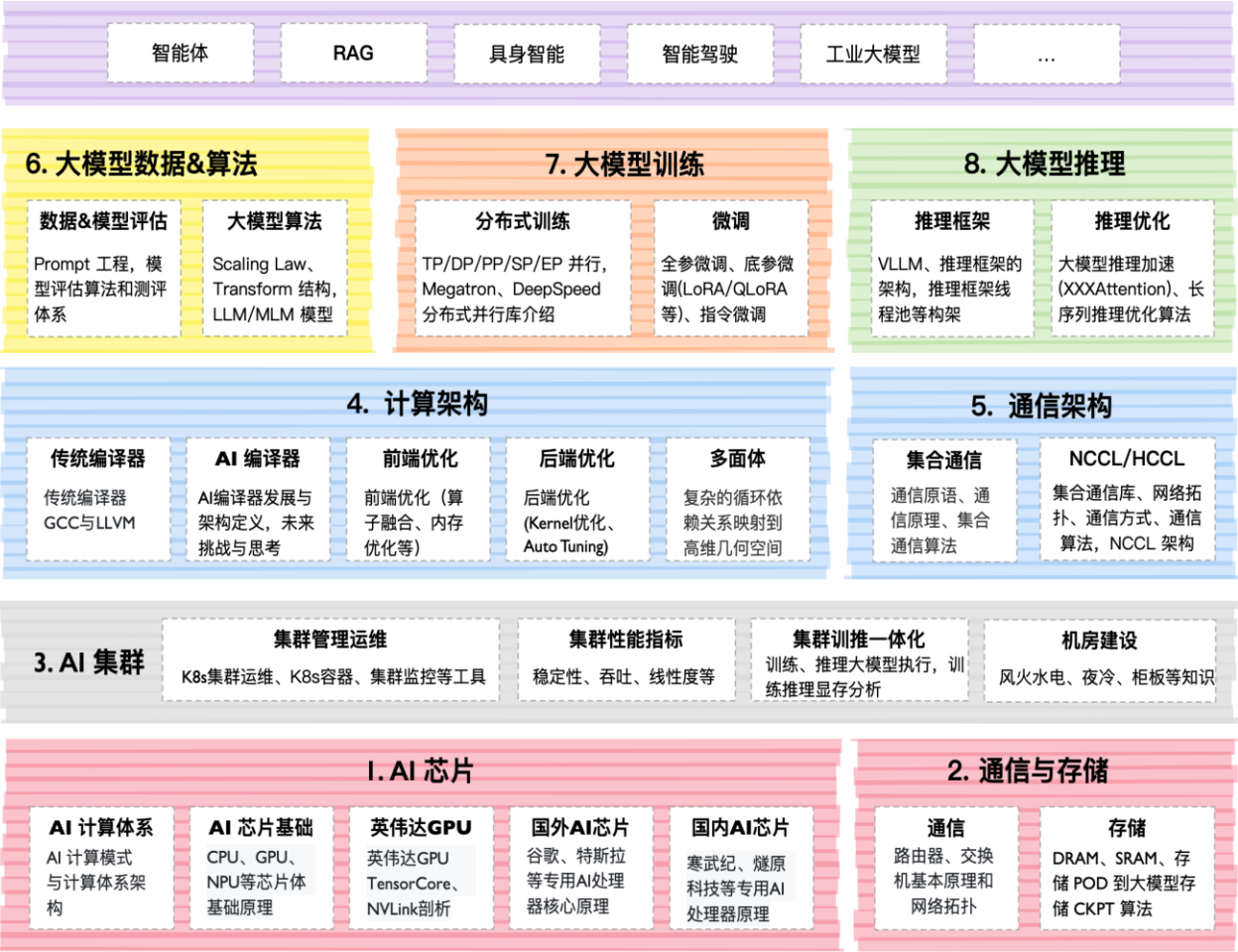


2、国产芯片厂商春天与寒冬

AI 系统 + 大模型全栈架构图



<https://github.com/chenzomi12/Allnra>

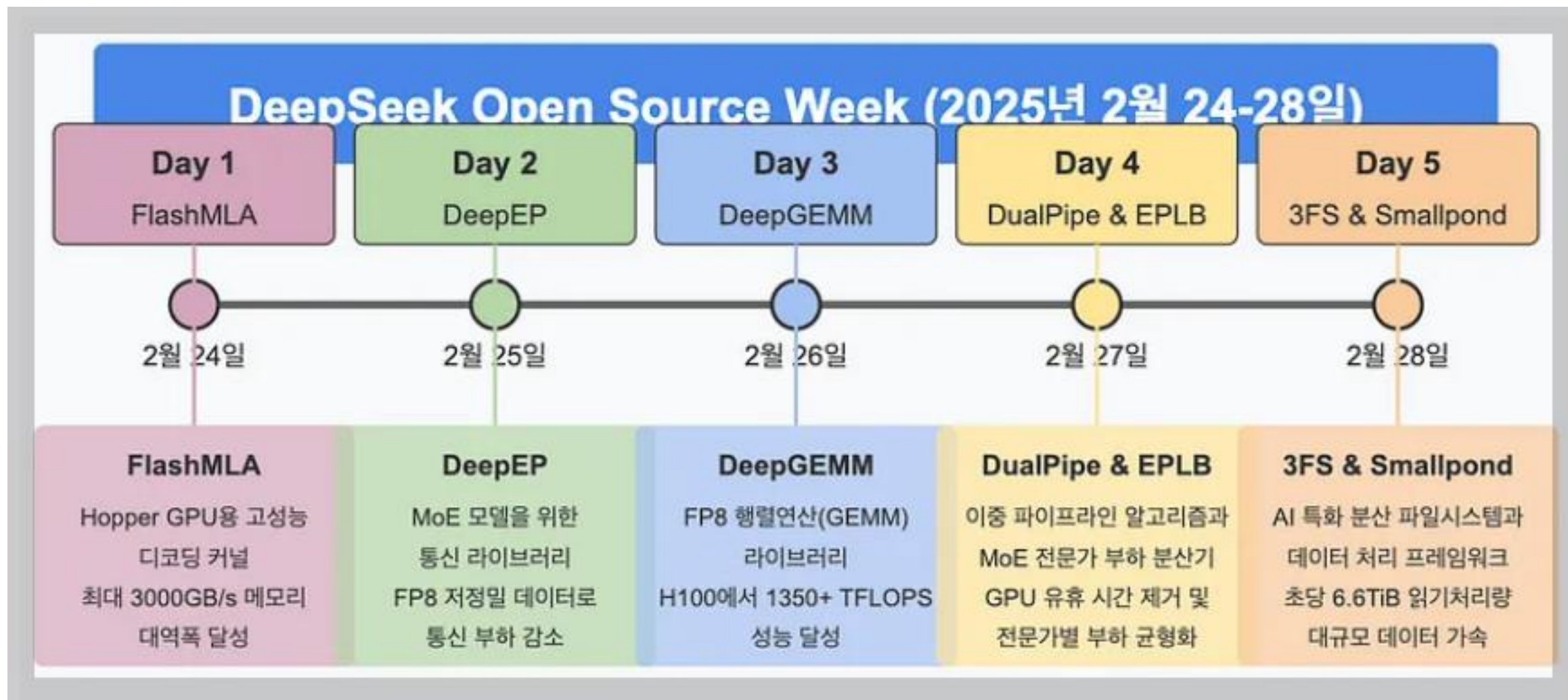


3、大模型厂商竞争加剧

1. **开源倒逼技术透明化：** FlashMLA、DeepGEMM等核心算法库，公开模型架构与调参方法，推动行业从“黑盒模型”向透明化发展。百度文心一言要开源，QWen 也要开源了！❤
2. **竞争格局新变化：** 迎来全球开发者复现热潮（UC 伯克利复现 R1），开源模型加速技术民主化，中台开源人员压力陡增；中小厂商通过生态协作快速追赶头部企业；行业竞争从“模型能力”转向“数据与场景适配性”；
3. **技术转型到 MoE 架构：** DS 以成本优势吊打 LLAMA & Grok，迫使大厂商优化 AI Infra & 算法架构，转向 MoE 和低精度计算（FP8）以降低训推成本；甚至部分厂商会从国产路线转回 NV 技术路线。
4. **放弃自研大模型华丽转身：** 算法和 AI Infra 人才储备不足，手头卡数资源不足，资金不足等原因，迫使从预训练大模型转向提供大模型服务（垂直领域后训练、微调、蒸馏、MaaS 平台等）。



3、大模型厂商竞争加剧

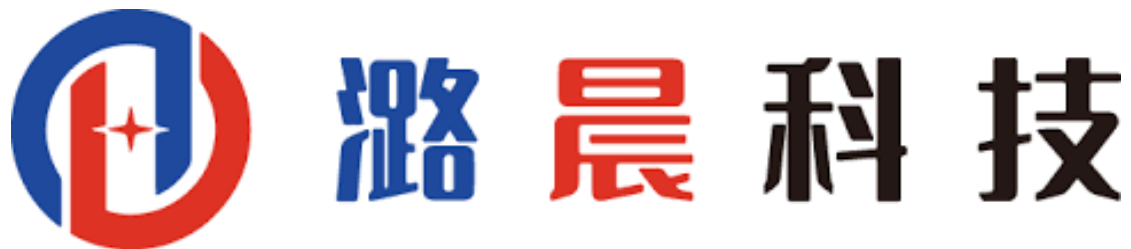


4、MaaS 平台服务 & 转型

- **算力服务调整：** 推理算力需求呈指数级增长。MaaS 和云平台厂商积极调整算力服务策略，加大对推理算力投入，满足低成本、高效能推理算力的需求。maybe 🐘 云厂商可能会推出针对推理场景的算力套餐和优化方案。
- **迫使加大优化投入：** DS 一天的总收入为 \$562,027，成本利润率 545%，刺激 MaaS 和云平台厂商加快对 DS 适配和优化，提升平台性能、稳定性，提升 QPS。如优化云平台计算、存储和网络资源分配，提高运行效率和用户体验。
- **商业模式创新：** 提出将 DeepSeek 模型与端侧硬件结合一体机，推出高算力模组，探索“端云协同”的新盈利模式。



4、MaaS 平台服务 & 转型



坑人的硅基流动



尤洋

2022 年度新知答主

+ 关注他

2 人赞同了该文章

我评论一下潞晨科技的竞争对手硅基流动。

我本来不想发这些东西，但是硅基流动的袁进辉老师频繁在朋友圈里阴阳我。这家公司疑似组织水军在网长期黑我。今天DeepSeek有一篇文章指向我，他也在那里煽风点火。

忍无可忍，我写一篇文章回应一下。

硅基流动三周前网站访问量大增，原因是什么呢？（不过我看最近已经跌到巅峰的30%，很可能是昙花一现。）

原因：

1. 牺牲员工的春节假期，绑上华为春节假期期间最早发出公众号和可用的DeepSeek API，由于华为在中国的地位，让人联想到AI全栈国产化。激起了国人的兴趣。宣传效果很好。
2. 邀请码直接送代金券，拉人头在小红书上快速形成病毒式扩散。邀请人和被邀请人都能获得14元。有很多小红书用户刷到了上千元。



袁进辉

无语。我们团队愿意拼搏抓一个机会会有什么错？邀请用户送点免费券有什么错？很多应用都这么做，海外也有；春节那几天，全民都想访问DeepSeek而不得时，我们提供了仅有的一个稳定的服务，用户愿意过来有什么错？来的人太多了，网站被挤爆了，付费用户也用不了了，只好辟出一块资源作Pro版给付费用户，保障付费用户体验有什么错？我们免费版现在体验也改善了，Pro版一直是业内最稳定的服务之一。硅基流动工程师在几年前OneFlow时就开源过一批比英伟达官方实现还要快的算子，嗨被潞晨科技抄袭，只是为了给对方留个面子没有公开，现在竟然这样诋毁我们。







Thank you

把AI系统带入每个开发者、每个家庭、
每个组织，构建万物互联的智能世界

Bring AI System to every person, home and
organization for a fully connected,
intelligent world.

Copyright © 2024 XXX Technologies Co., Ltd.
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. XXX may change the information at any time without notice.



ZOMI

GitHub github.com/chenzomi12/AllInfra



ZOMI