LLM 训练推理加速在阿里巴巴的实践

杨斯然 / 刘侃

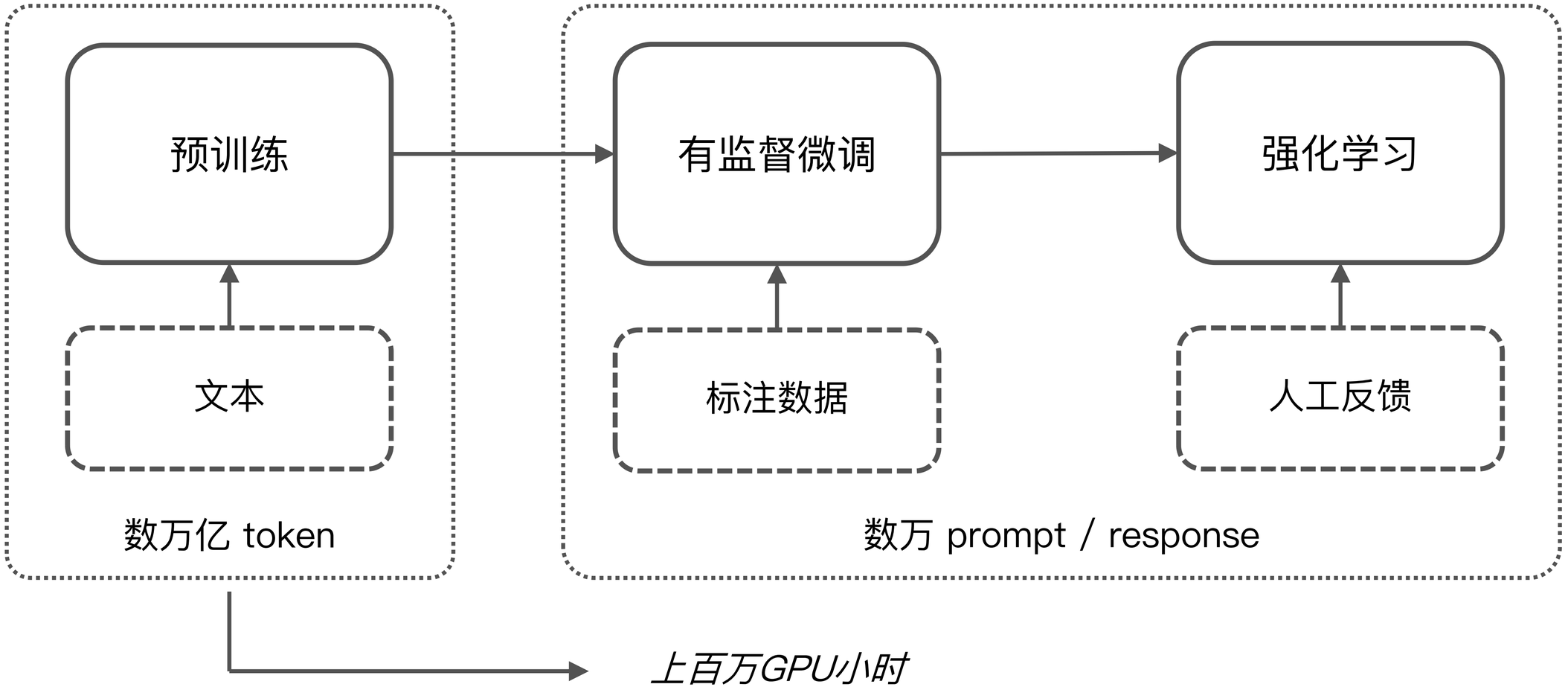
大语言模型训练和Megatron-LLaMA框架

- 大语言模型训练的过程
- 大语言模型训练的算法和问题
- 大语言模型中的模型并行
- Megatron-LLaMA框架的计算和通信并行
- Megatron-LLaMA框架的3D并行调优
- Megatron-LLaMA框架应用到LLaMA模型中
- 小结

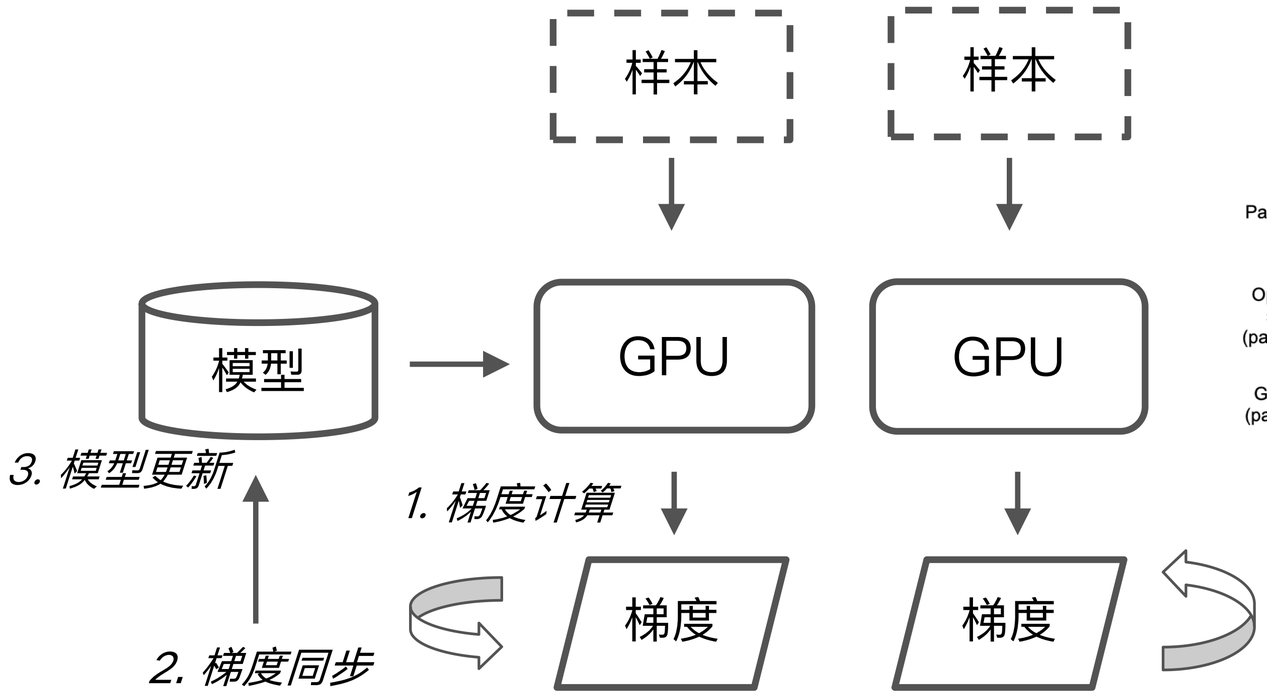
大语言模型的应用

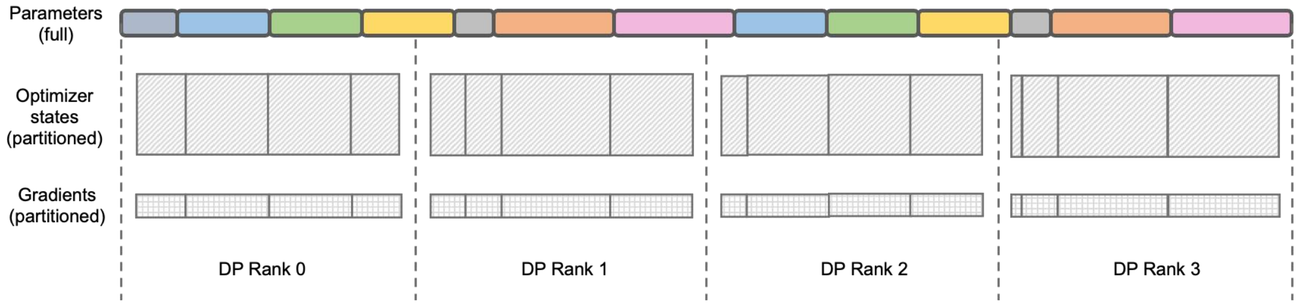

LLM应用
Agent
米哈游-智能NPC
有鹿机器人-具身智能
Character AI-角色扮演
代码智能
Github Copilot
aiXcoder (硅心科技)
AI问答引擎
Perplexity AI
Bing Chat
天工搜索
教育
好未来
网易有道
科学研究
Cradle-蛋白质设计
MedWise-临床判断
DeepMind
ChatBot
文心一言
ChatGPT
Bard
文案生成
讯飞听见写作
小红书生成器
阅文妙笔
企业服务
智能运维
法律助手
金融助手

大语言模型：训练过程


预训练
有监督微调
强化学习
文本
标注数据
人工反馈
数万亿 token
数万 prompt / response
上百万GPU小时

大语言模型训练：算法和问题



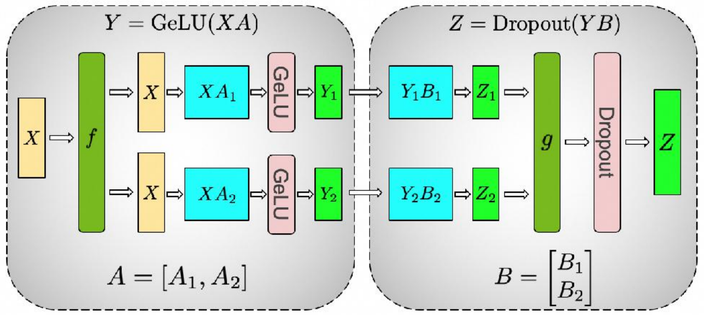


	需求	限制
模型存储	13B模型：156GB 65B模型：780GB	显存容量：80GB
梯度同步	13B模型：52GB 65B模型：260GB	网络带宽：100GB/s

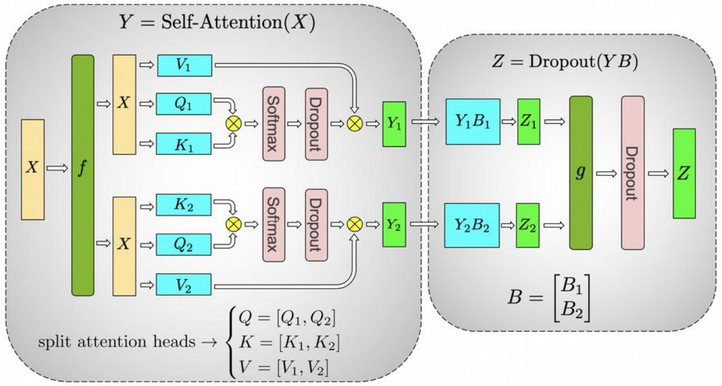
DeepSpeed：ZeRO优化器

1. 优化状态切分到数据并行的各个rank上
2. 部分通信和梯度计算并行
3. 激活重算，节省梯度计算过程显存

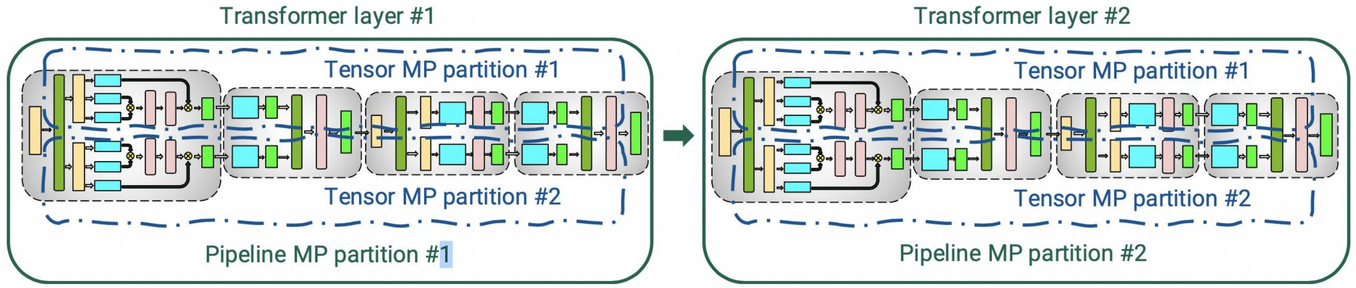
大语言模型训练：模型并行

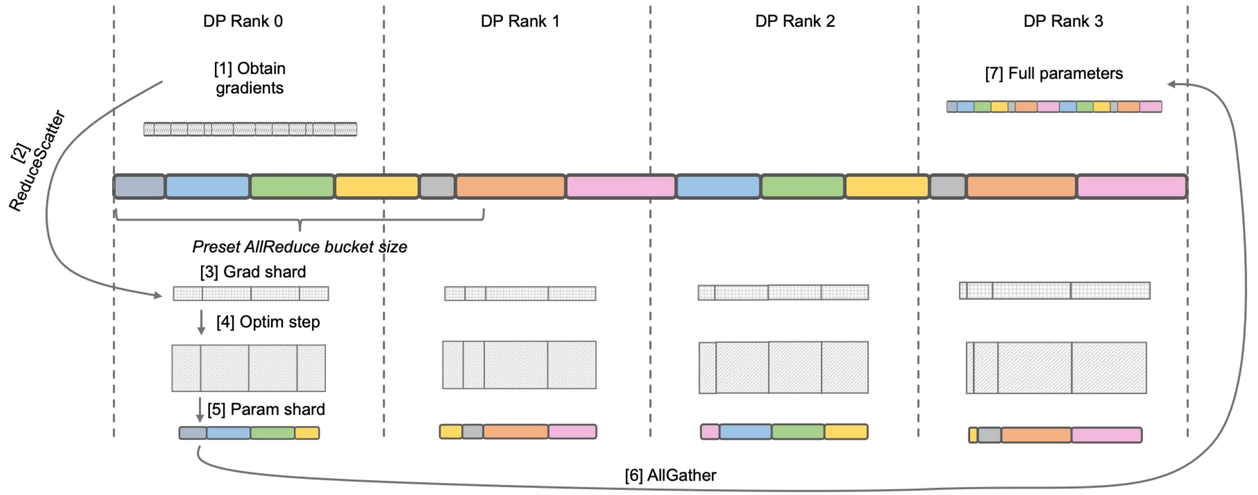


(a) MLP



(b) Self-Attention





Megatron-LM：3D模型并行

1. MLP层和Attention层切分到多张卡上，通常用NVLink互联
2. 模型的分为多个Stage，切分到多台机器上 3. 分布式优化器

参考文献：

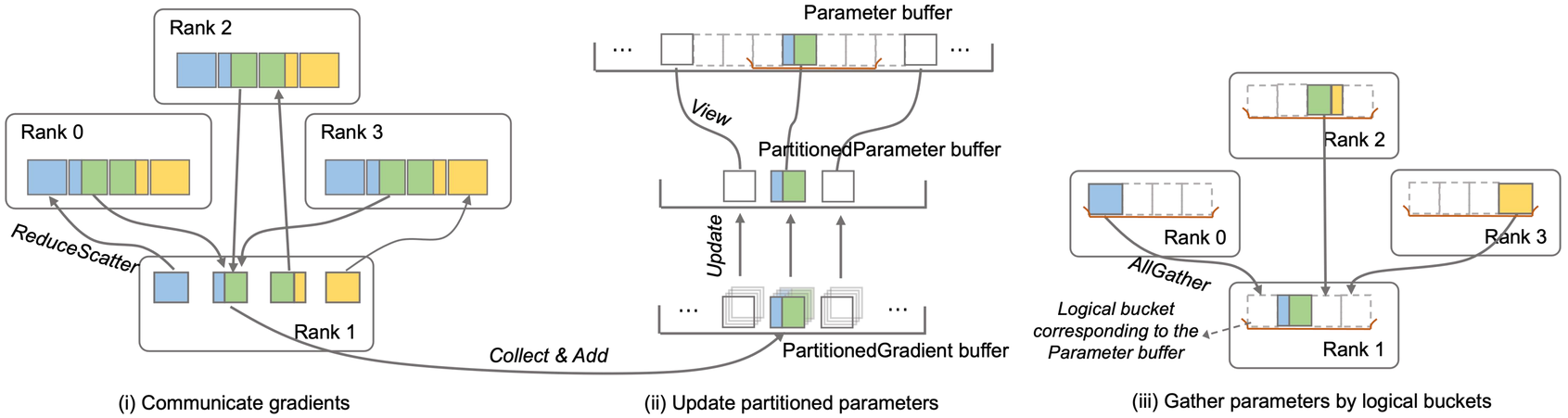
1. *Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism*
2. *Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM*

Megatron-LLaMA框架：计算通信并行

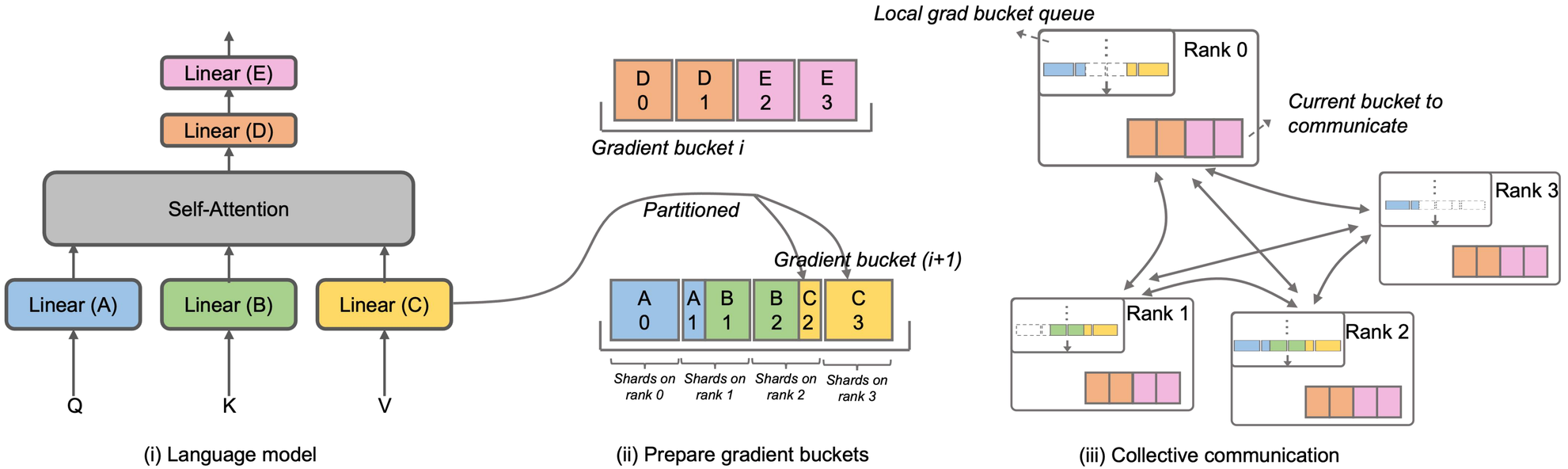
Megatron-LM 分布式优化器时间线

Megatron-LLaMA
通信遮盖分布式优化器时间线

Megatron-LLaMA
参数分片和梯度更新



Megatron-LLaMA框架：计算通信并行

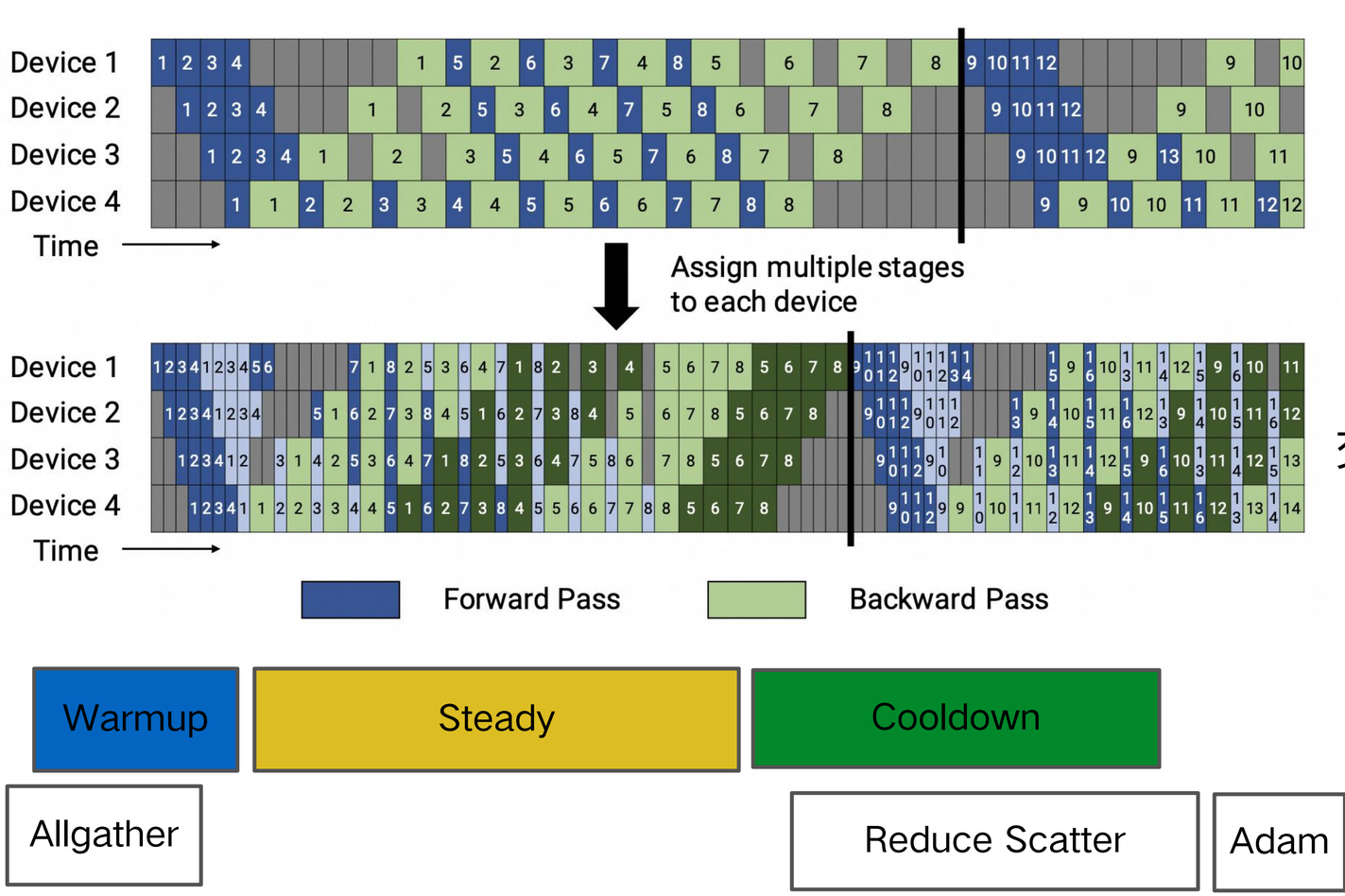


(i) Language model　(ii) Prepare gradient buckets　(iii) Collective communication

LLaMA-13B 训练时间

	256 x A100 80GB	512 x A100 80GB
Megatron-LLaMA	1890 (23.9 天)	1845 (12.2 天)
Megatron-LM	1630 (27.8 天)	1430 (15.8 天)

Megatron-LLaMA框架：3D并行调优



Tensor并行开销：正比于TP Size
Pipeline并行开销：正比于 (PP Size – 1)

交错Pipeline并行开销：正比于 (PP Size – 1) / V

Megatron-LLaMA
通信遮盖应用到Pipeline并行

参考文献：
Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM

Megatron-LLaMA框架：LLaMA模型应用

```
sh tools/checkpoint_conversion/hf_to_megatron.sh

torchrun --nproc_per_node=8 pretrain_llama.py \
    --tensor-model-parallel-size 2 \
    --pipeline-model-parallel-size 1 \
    --overlapped-distributed-optimizer \
    --reduce-bucket-size 4e8 \
    --tokenizer-type=PretrainedFromHF

sh tools/checkpoint_conversion/megatron_to_hf.sh
```

优化器设置

通信分片设置

Tokenizer设置

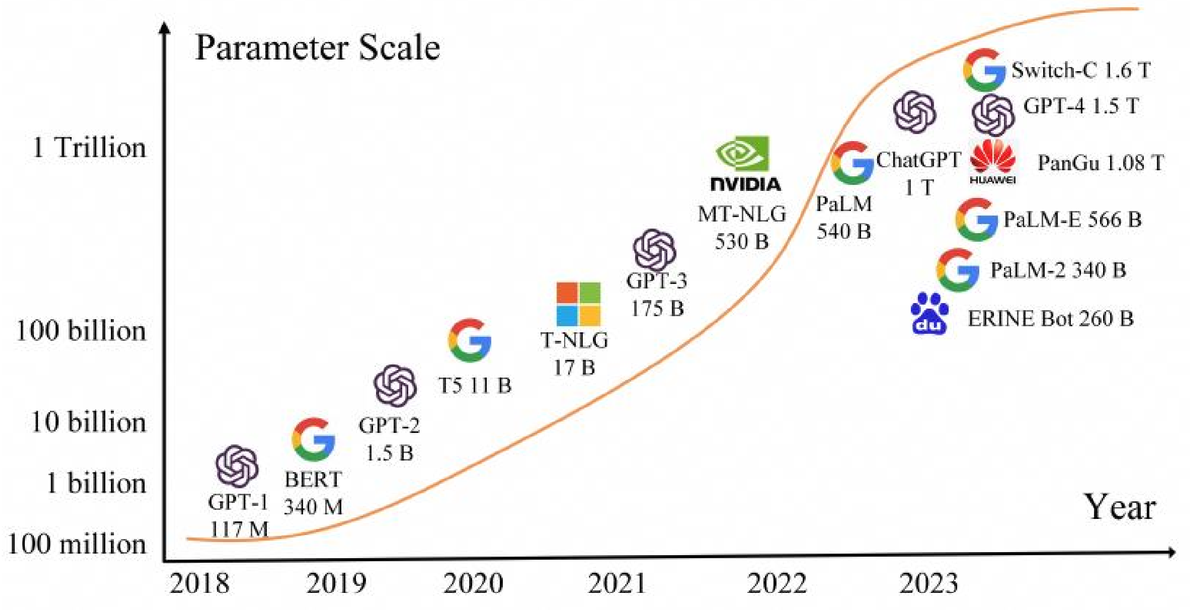
大语言模型训练：小结

- 大语言模型的训练需要消耗大量的GPU资源
- 算力、显存、通信是三大核心瓶颈
- 框架可以提供在这三者之间平衡和优化的手段
- Megatron-LLaMA通过在Megatron-LM的基础上进行通信/计算模式优化
- 同时提供LLaMA相关结构以及HuggingFace生态兼容工具
- 开源地址：https://github.com/alibaba/Megatron-LLaMA

LLM 推理和 rtp-llm 框架

- LLM 推理的趋势
- LLM 推理的应用场景和挑战
- LLM 推理的核心问题和优化方法
- rtp-llm 框架简介
- rtp-llm 框架实践 — 淘宝问问 KVCache 复用
- rtp-llm 框架实践 — Query 改写极致延迟优化
- rtp-llm 框架实践 — Speculative 近似方法
- 小结

LLM 推理的趋势



Source: https://arxiv.org/pdf/2312.06261.pdf

LLM — 新的信息基础设施和应用底座

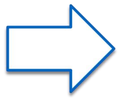
- 级大的推理算力需求
- 极高的延迟响应要求
- 极致的成本考验

LLM 的应用场景和挑战

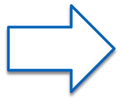
聊天应用	流式、token 延迟敏感, 30ms/token
代码补全	单轮、端到端延迟敏感, ~1s
Query 改写	单轮、端到端延迟敏感, ~50ms
离线评测	利用率高，最大化吞吐
模型多租	利用率低，最小化成本

LLM 推理的核心问题和优化方法

- 算力需求
- 延迟要求
- 成本考验



- 大模型做小
- 算力做强
- 请求/部署 复用



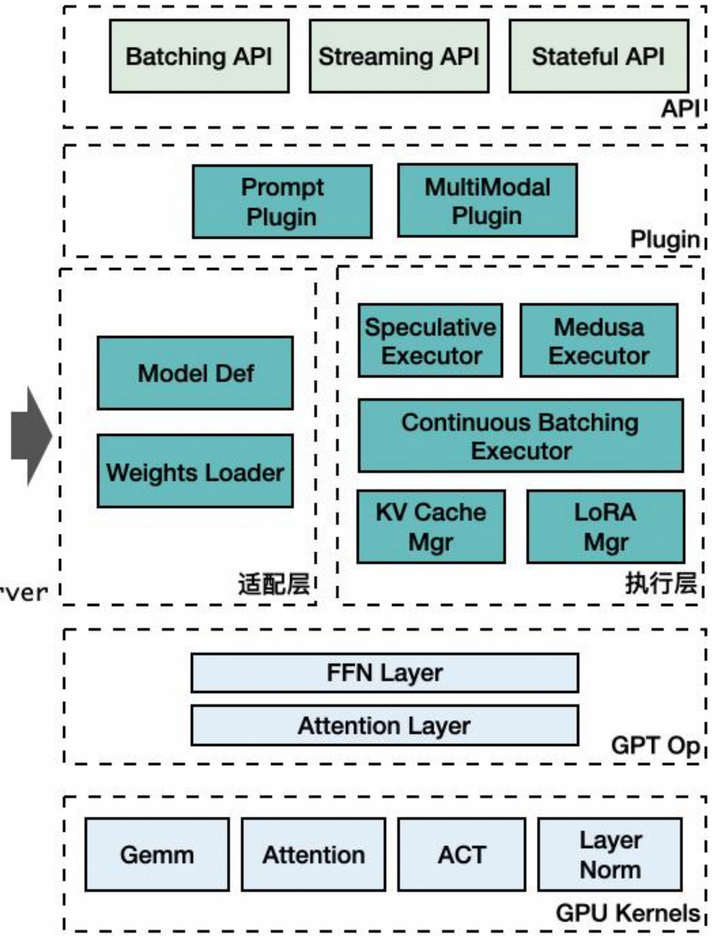
- 量化/剪枝/蒸馏
- 小模型辅助加速

- 高性能异构计算硬件
- 多机/多卡
- 软硬件结合优化

- Continuous Batching
- Multi Tenant

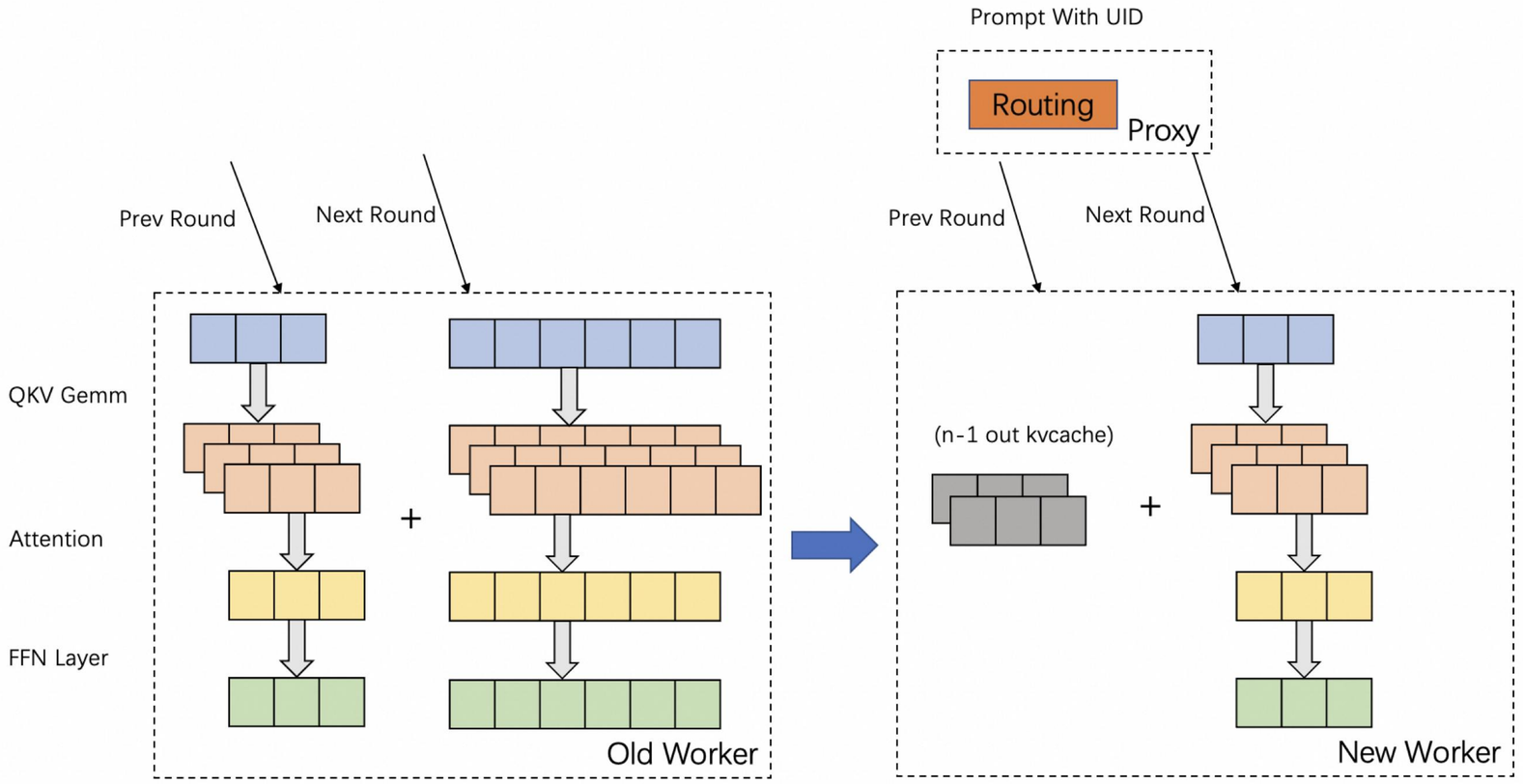
rtp-llm 框架

```
CHECKPOINT_PATH=/path/
TOKENIZER_PATH=/path/
MODEL_TYPE=llama
WEIGHT_TYPE=int8
python3 -m maga_transformer.start_server
```



- 全面适配主流开源模型/内部模型
- 配置化定义，统一模型执行流程
- 动态/静态添加 LoRA
- 动态/静态量化
- 支持多种辅助推理加速方法
- 支持多模态模型

rtp-llm 实践–淘宝问问 KVCache 复用



System Prompt?
无解，重来

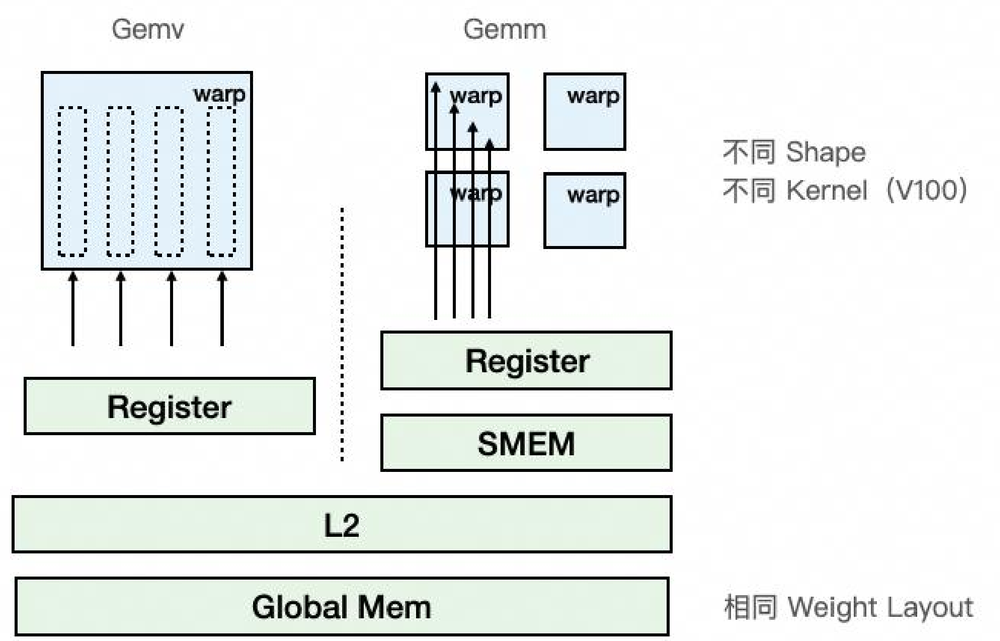
Cache 迁移?
显存不够，无解
需要复杂的 Proxy 负载均衡策略

rtp-llm 实践-Query 改写极致延迟优化

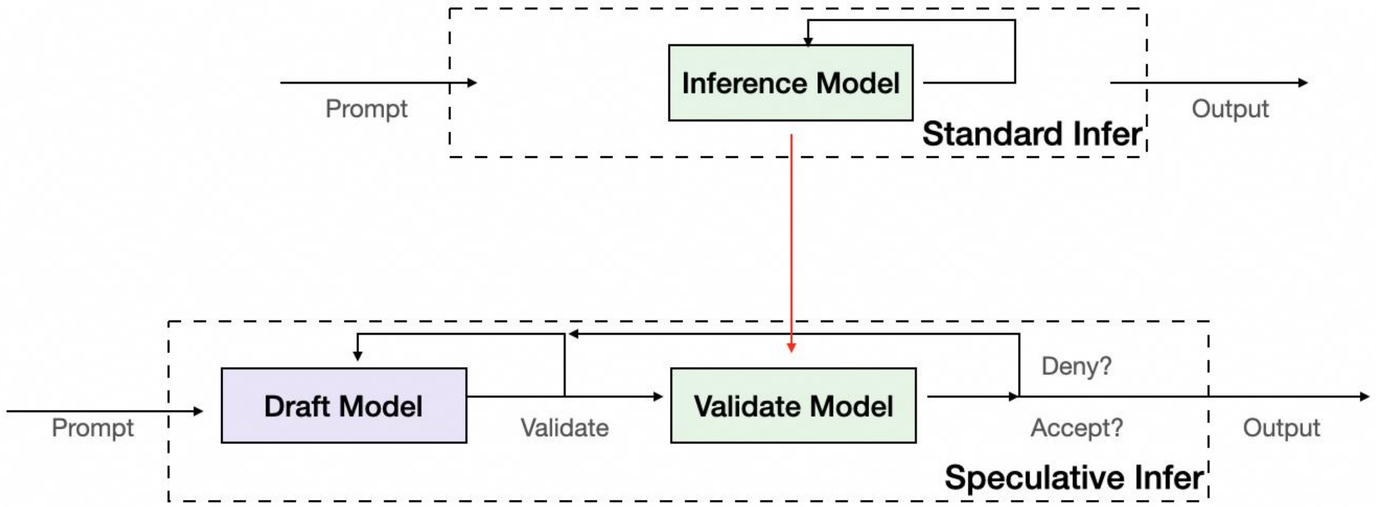
大模型变小

	缺点	优点
量化	Scaling 有限	应用部署简单
蒸馏	训练复杂， 需要较多实验，精细调优效果	参数量减少明显 加速适配简单
剪枝	剪裁比较多的是 HEAD 而不是 FFN，加速较弱	训练相对简单 效果损失可控

CUDA Kernel 优化



rtp-llm 实践-Speculative 近似方法



- Draft Model 复杂度要足够低
- Draft Model 选择（Speculative Decoding，Medusa，REST）

延迟对比		
A 场景（S-D）	B 场景（S-D）	C 场景（Medusa）
-37.7%	+0.19%	-50%

LLM 推理：小结

- LLM 推理需要解决 算力/延迟/成本 三方约束问题
- rtp–llm 框架提供了多种平衡和优化手段
- LLM 推理是一个复杂的体系工程，基于框架我们还需要成熟的推理产品

THANKS

软件正在重新定义世界

Software Is Redefining The World