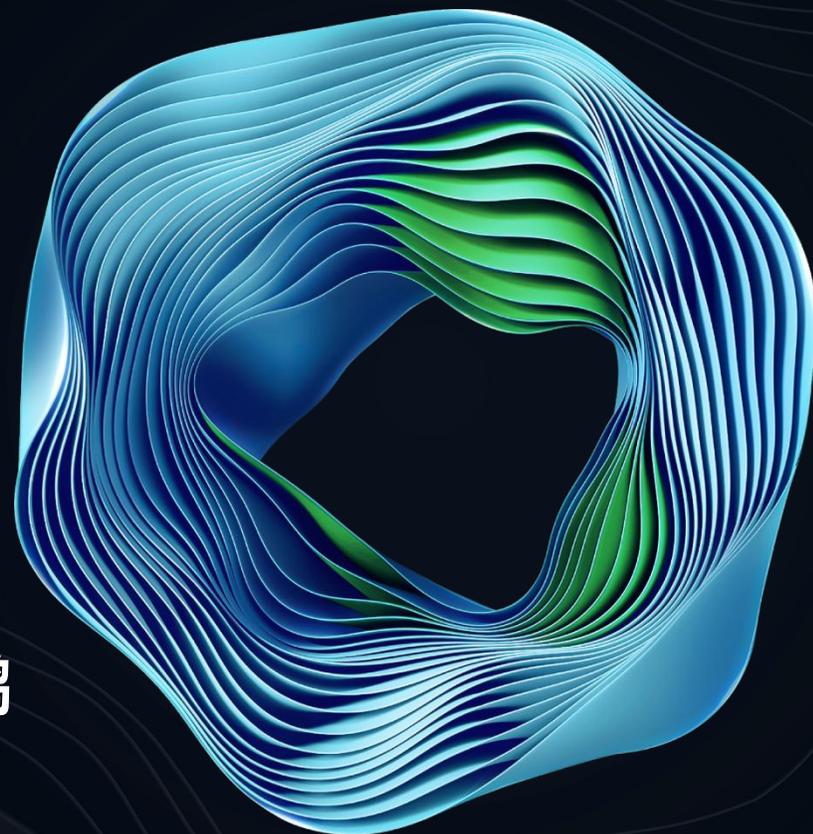


新华三智慧IT 成就AIGC时代 智造算力引擎

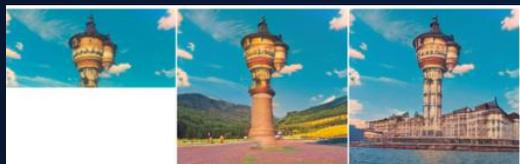
新华三云与计算存储产品线资深架构师 张鹏



ChatGPT与AIGC



AIGC可以生成图像、文本、视频、音频等多种内容模态



图像

草图到图像
图像补全
图像编辑
3D图像生成
.....



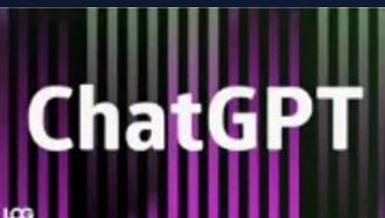
视频

视频草图到视频
视频预测、视频编辑
视频画质增强、风格迁移
.....



音频

文本合成语音
语音克隆
音乐生成
...



文本

文本创作
代码生成
对话问答
...



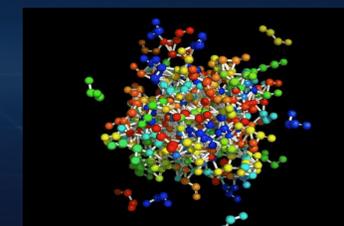
- 输入文本**
1. 一只逼真的泰迪熊正在旧金山的海里游泳
 2. 泰迪熊潜入水中
 3. 泰迪熊和五颜六色的鱼一起在水里游来游去
 4. 一只熊猫正在水下游泳

多模态

文本到图像
文本到视频
文本到3D
.....

其他

科研如蛋白质生成
TRDiffusion模型



Gartner预测, 2023年将有20%的内容被生成式AI所创建;
2025年, Gartner预计生成式AI产生的数据将占有所有数据的10%, 而今天这个比例不到1%

大模型应用覆盖企业全业务流程

生产

- **产品/工业设计**，产品工业设计、3D模型快速生成
- **软件开发**，代码补全与生成、BUG检查修复、代码注释生成
- **智能流水线**，智能化响应和调整生产计划

交易

- **助力营销沟通**，会议实时提词，沟通话术辅助改善
- **营销内容生成投放**，语料、海报等快速生成，差异化投放
- **采购&营销策略优化**，智能响应调整优化
- **同报表生成**，AI风险评估及合规审查

服务

- **智能售前咨询**，基于产品语料，专业话术，高效答疑
- **数字人服务**，数字人提供导览、播报、讲解等服务
- **智能售后客服**，问题响应范围提高，质量全天候保障

运营

- **办公模式革新**，文本写作及信息处理，工作指令自动执行
- **人力资源管理进化**，职位自动发布、筛选和面试，AI培训与问答
- **财务&法务提速**，合同报表生成，AI风险评估及合规审查

全域AI应用，赋能企业全流程生产力进化

训练大模型需要大规模算力做支撑

典型大模型的训练和部署，对AI算力消耗巨大

GPT-3模型

- 1750亿个参数，45TB训练语料
- 训练175B的PPO-ptx模型需要60pflops/s-days
- 训练GPT-3算力消耗约3,640pflops/s-days
- ChatGPT按1300万/天访问量，估算需要3万多张A100 GPU

谷歌PaLM模型

- 5400亿个参数
- 2.5亿个图像文本对数据集
- 2.56*E24FLOPs
- 消耗算力29600pflops/s-days

DALL.E模型

- 120亿个参数
- 2.5亿个图像文本对的数据集上训练
- 1*E22FLOPs

Stable Diffusion模型

- 8-9亿个参数
- 25亿个图像文本对数据集 (LAION-5B)
- 4000块 A100 GPU上训练

针对超大模型训练，Nvidia GPU是主力，国产卡在全力追赶



算力相对较强的国产卡如寒武纪和壁仞，均受到限制无法继续生产

另外，大规模的AI计算，对绿色能效、液冷方面有很强需求

训练大模型需要大规模算力做支撑

GPT-3模型所需的GPU显存

- 参数量 (FP16精度) : 350GB (175B*2bytes)
- 梯度 (FP16精度) : 350GB (175B*2bytes)
- 优化器状态 (FP32精度, 包括权重、均值、平方) : 2100GB (175B*12bytes)
- 总计: **2800GB** (350GB+350GB+2100GB)
- 满足上述显存的GPU数量: **35张A800 80G**

GPT-3模型所需的GPU算力

- 每次迭代所需的算力 (参考 <https://arxiv.org/abs/2104.04473>) :

$$96BSlh^2 [1 + S/(6h) + V/(16lh)] = 4.5 \text{ ExaFLOPs}$$

B: batch size, S: sequence length, l: transformer layer number, h: hidden size,
V: vocabulary size

- 迭代~95000次所需的算力: **430 ZettaFLOPs** (1 ZettaFLOP=1024 ExaFLOPs)
- 在**128个HGX A800 8GPU模块** (1024张A800 80G) , 所需训练时间 :

$$430 * 1024^3 / (2496 * 50\%) / 3600 / 24 / 128 = 34 \text{ 天}$$

-已知1个HGX A800 8GPU模块可提供2496 teraFLOPS算力, 假设在集群训练中能发挥50%性能

- ChatGPT模型的训练时间可按如下公式估算:

$$\text{Training time(s)} \approx 8TP/nX$$

P: parameters, 模型的参数量

T: Tokens, 数据集的数据量大小

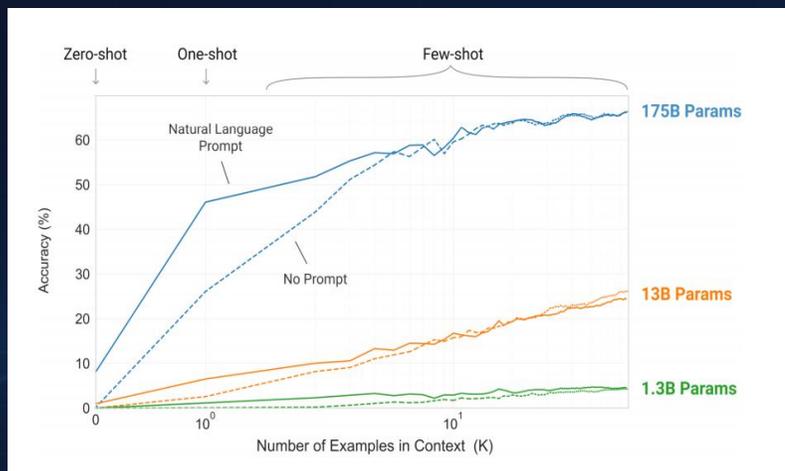
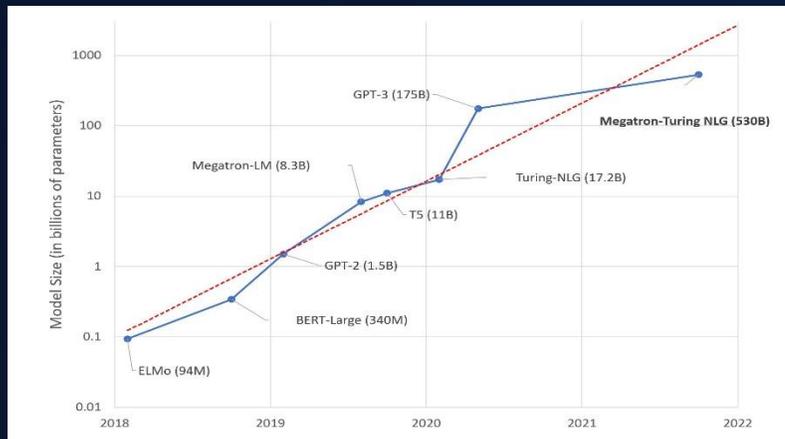
n: GPUs, GPU的数量

X: Tflops, GPU实际能达到的算力

- 此外, 针对GPT模型参数量的大小, Nvidia也给出了一个大概的GPU集群规模

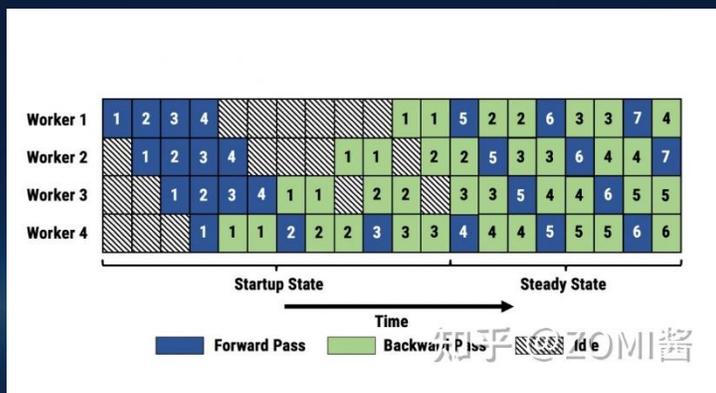
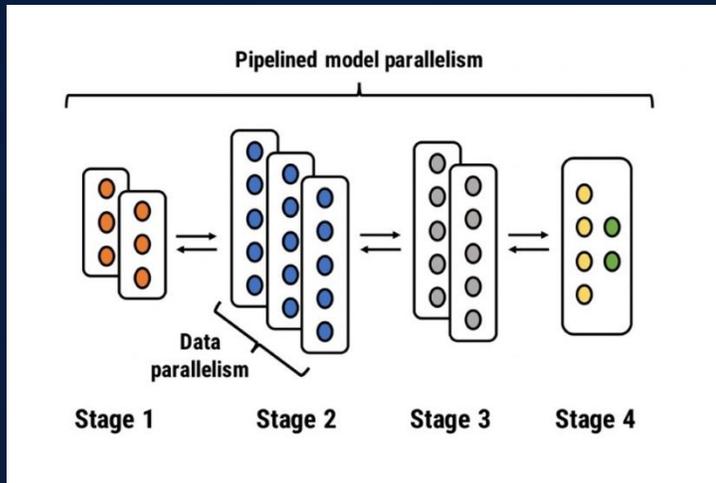
Model Size	Number of A800 GPUs	Number of DGX Servers
1.7B	32	4
3.6B	64	8
7.5B	128	16
18B	256	32
39B	512	64
76B	1024	128
145B	1536	192
310B	1920	115
530B	2520	315
1T	3072	384

AI大模型对计算资源的要求



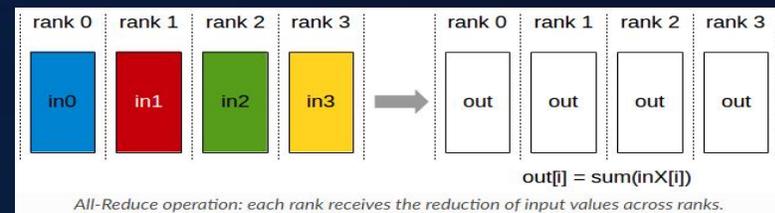
模型参数逐年增长，训练所需的算力水涨船高

- NLP属于预训练模型，为保证算法准确，参数逐年指数增长
- Zero Shot Prompting需要通过大模型保证输出效果
- 模型训练时实际内存占用是参数的**5倍以上**

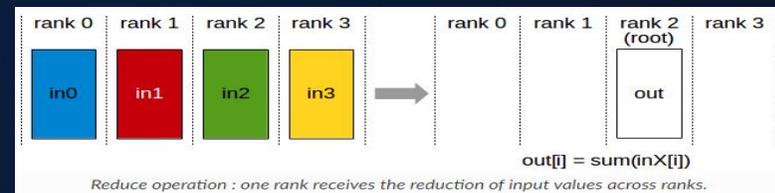


需要通过多机多卡GPU集群保证算力需求

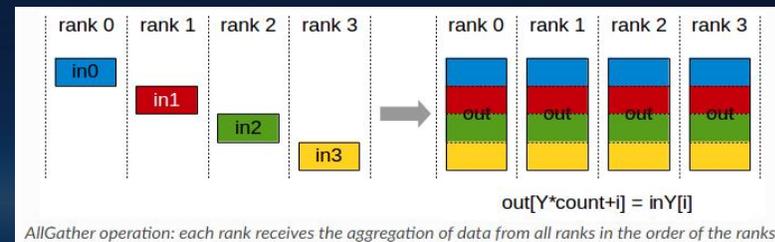
- 流水线并行需要对模型和数据进行分片，涉及GPU间fullmush和串行通信
- 提升了GPU利用率，但对数据集**加载速度**有较高要求



AllReduce, 计算所有GPU数据后保存到所有GPU



Reduce, 计算所有GPU数据后保存在单个GPU



AllGather, 汇聚所有GPU中数据后保存到所有GPU

需要通过高速通信保证GPU集群训练效率

- 训练期间GPU通信存在P2P (1对1) 和Collective通信 (1对多或多对多)
- 其中Collective需要高速通信以匹配GPU的算力

GPU服务器基础设施 — 从推理到训练全场景覆盖

持续创新的G6新品

多样算力先锋

H3C UniServer R4950 G6

192核 10个 512TB
处理器 PCIe 5.0槽位 存储空间

专业算力旗舰

H3C UniServer R4900 G6 Ultra

120核 4张 400Gb/s
处理器 PCIe5.0 双宽GPU IB网络

混合算力引擎

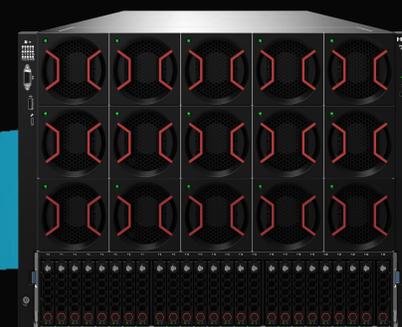
H3C UniServer R5350 G6

10块 400Gb/s 4种
双宽GPU 极速网络 智能拓扑

智能算力旗舰

H3C UniServer R5500 G6

7倍 400Gb/s 12块
AI算力提升 极速网络 NVMe SSD



推理/小规模训练

1~4 A800-AI训练/HPC
1~4 A30-AI训练/推理
1~4 A10-AI推理
1~8 T4-AI推理/智能监控

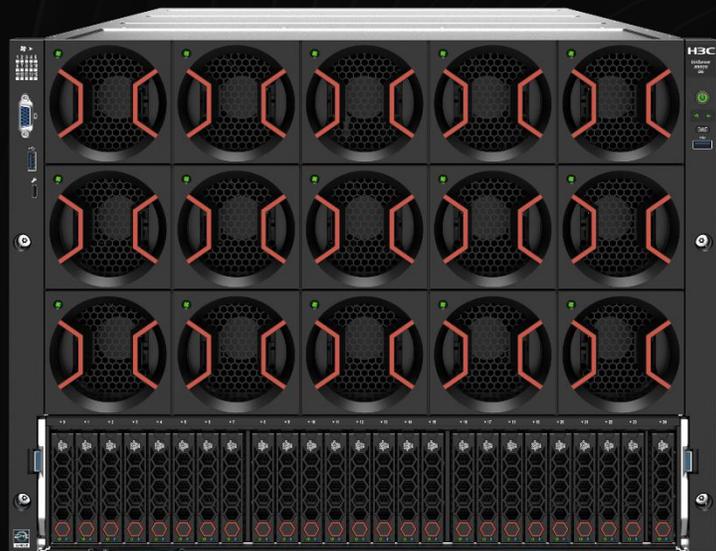
大规模推理/训练

HGX A800 4-GPU-AI训练/HPC
4~8 A800-AI训练
4~8 A30-AI训练/推理
4~8 A10-AI推理
8~16 T4-智能监控

大规模训练

HGX A800 8-GPU-AI训练
*液冷解决方案已在研，请各位关注
GPU液冷需求，现有可支持方案为
R5500 G5 INTEL平台

H3C UniServer R5500 G6 — 智能算力旗舰



强劲算力，助力AI业务高效运转

适配最新一代NVIDIA企业级GPU模组，性能对比上一代产品提升3.4倍，为AI业务提供强劲算力



灵活拓扑，适配不同AI场景需求

2种GPU拓扑设计，双平台CPU设计，灵活适配客户不同AI场景需求



模块化设计，轻松简便运维

系统解耦，模块化设计，无需下架可进行运维；GPU与计算节点分开独立供电，保障业务稳定运转

双平台CPU

2颗第四代英特尔®至强®可扩展处理器/2颗AMD EYPC™处理器

强劲算力

32 PFLOPS

灵活拓扑

2种GPU拓扑

模块化设计

灵活部署CPU/AI平台

[注] 1. 32 PFLOPS=3958 TFLOPS*8 (H800 FP8) 2. 3.44=67/19.5 (FP32)