**重点内容关注标红部分。**

**问答环节：**

**Q：英伟达GH200对光模块和PCB厂商有什么影响?**

1）光模块：光模块更多在系统之间的高速互联，除了100G和200G国内兼容性好，英伟达CX7这种400G高端场景，国产支持不太好。另外这些系统实际对光模块需求量不是很大，除非是大规模集群设计。

综合认为在大规模AI场景中，国内光模块厂商不会扮演重要角色；他们更多在信创、传统数据中心（25G、100G）这些中低端市场有一些份额。

2）PCB：对PCB要求越来越高，封装对基板要求也更高；高端PCB需求可能会高速增长。

**Q：PCB的ASP会提升，但PCB用量是下降的?**

整体PCB量没有明显下降，而且ASP上升，对PCB厂商是好事。

**Q：GH200、MI300封装技术的优点是什么？渗透速度会如何?**

对于需要完整一体化方案的场景有优势，高度集成、性能提升30%-40%（保证CPU和GPU之间内存一致性），对于英伟达推广自己云服务也是好的解决方案。

但对于很多互联网厂家不是很愿意，他们希望更开放的设计，否则只能绑定在上述产品架构上，而且价格溢价能力变弱；大的市场可能不会有很大市场份额，小于10%。

英伟达推出GH200主要用于弥补CPU体系不足，构建完整的生态，不要过分依赖于CPU x86环境；AMD则是主要用于跟英伟达竞争。

**Q：寒武纪跟百度合作的情况怎么样，百度有没有使用思元590?**

百度文心一言没有使用思元590，只是早期做了适配，小规模部署500多片，实际上线并没有使用；目前主要用A100和昆仑芯2代。

之前是建立开发团队配合开发，但实际部署结果来看，590性能指标不如A100，而且架构不太兼容、难度大，所以百度没有用.

寒武纪目前的产品不太适合大模型迭代，软件生态问题比较大；比如百度模型在不断迭代中，而每次迭代都需要思元590进行适配和优化，工作量太大，不适合百度开发；另外架构和指令集都比较特殊，不可控因素太多

未来发展上需要进行主流几个框架的支持，但目前支持都不太好，大模型场景使用有难度。

**Q：快手传言使用了寒武纪思元590?**

快手没有使用大模型，只是使用了一些传统AI技术、小模型，不涉及大规模系统并行，思元590可以支持。

**Q：国内芯片厂商里沐曦相对好一些?**

沐曦综合情况好一些（软件跟CUDA兼容，团队是AMD原来开发MI200的核心团队），但产品还没出来，只能做初步评价。

**Q：A100和H100在国内受限，国内下游厂商是不是对国产芯片持开放态度?**

态度开放，都需要找一些替代产品进行平衡，但性价比是关键因素。

**Q：模型迭代是否使得推理算力需求降低?**

未来头部大厂会出现预训练大模型，而更多模型会是垂直领域小模型，对算力需求碎片化；大规模算力只有头部厂商有需求，其他厂家只需要小模型、小算力。

不过整体需求还是快速增长，特别是推理需求，训练需求可能慢慢放缓。

**Q：国内厂商算力储备大概什么量级？向英伟达采购量增长多快?**

目前大厂各自手中估计有2k-3k片A100存货量，此前购买的大部分被常规业务占有，比较难拿出富裕算力（除非要把现有业务停掉，能凑出万片左右进行训练）。

国内互联网厂家3月份向英伟达进行38亿美元采购，年底才能陆续交货，持续交货到2024年；所以后面短期可能增速变慢。

**Q：AMD MI300的性能怎么样？进展如何？价格水平？未来空间?**

1）性能：MI300等比性能接近MI250两倍，整体性能应该是H100的1.5-2倍；核心架构类似英伟达GH200；软件支持对CUDA兼容。

2）进展：国内对应叫MI388，8月份可能提供测试样品；认为是非常强的产品。

3）价格：MI388 国内大概2.2-2.4万美金，非常接近H800。

4）用途场景：跟GH200完全一致，可以用于HPC、AI计算（性能很高，因为集成了CPU，不再需要单独CPU，板子上只需要MI300）。

5）竞争空间：对英伟达GH200、H100形成竞争压力，性价比具备优势；但在AI领域，MI300还有一定距离，主要是软件兼容性有一定差距（虽然支持CUDA兼容，但维护团队不够，并且跟目前典型大模型还没有适配案例，客户不了解）。

**Q：MI300下游客户有谁?**

国际主要是HPC场景，比如政府超算；国内字节、腾讯在做测试，但因为软件还不是很成熟，所以只能做算力评估，还不能做综合性能、大模型使用场景的评估。

**Q：哪些厂家扶持AMD竞争NVDA?**

国内外厂家都积极在跟AMD接触，比如微软、AMD、字节、腾讯，他们都对英伟达的溢价和垄断体系有一定的诟病。

**PPT环节：**

## 大模型生产流程更简单，但对基础设施要求更高

### 当前大模型生产范式的转变


传统AI实现


大模型实现

## 国内外算力市场发展区别：

1）国内厂商主流集群规模小（比如A100），很难有专门用于大模型训练的集群；国外有大量主流集群

2）国内开发框架不开源，模型市场分化；国外集中单一，开源，生态好

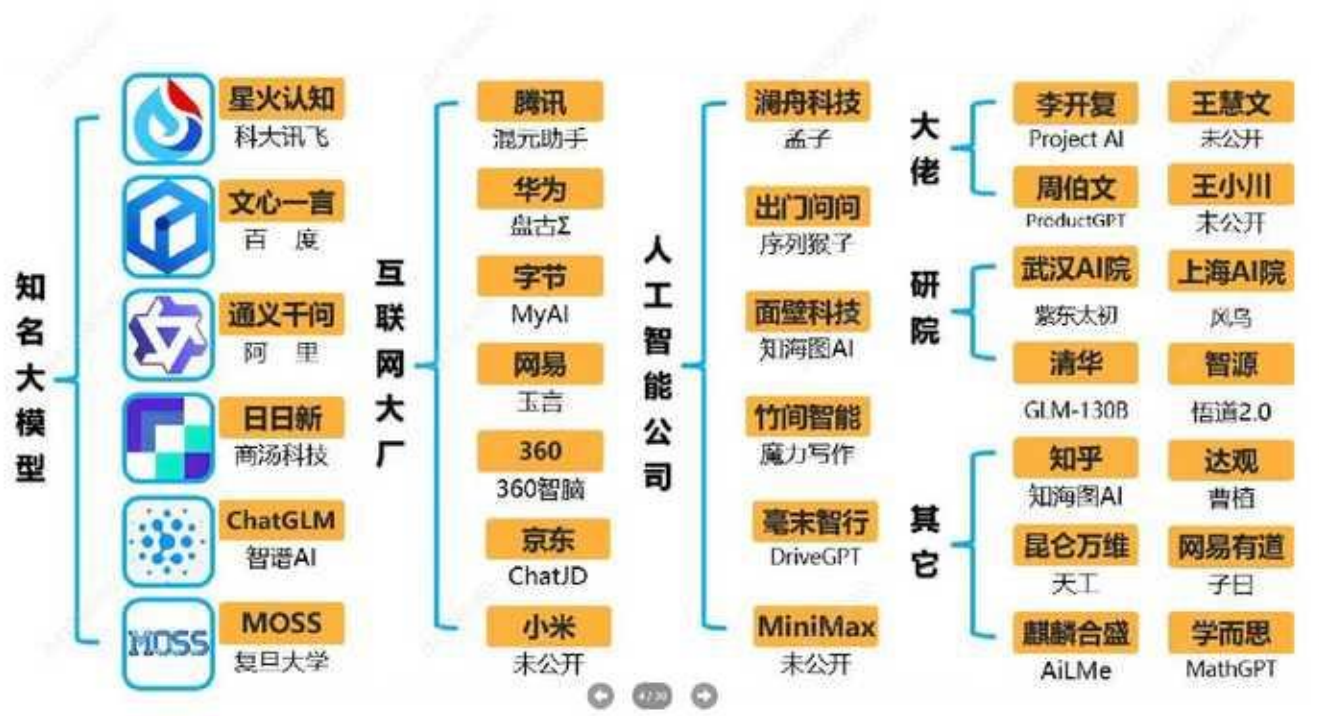
3）具体模型上，国内大部分是基于国外开源进行微调，多数没有掌握核心；国外开始向垂直领域渗透

4）应用方面，国内以内场应用为主，节奏较慢

### AIGC市场发展分析

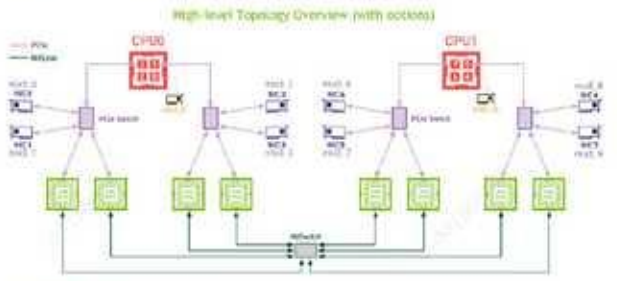
| | 国内 | 国外 |
|---|---|---|
| 算力基础设施 | 云服务商和国家实验室 | 云服务商 |
| | 阿里、腾讯、百度、国家超算中心和实验室<br>千片规模上流GPU集群资源 | AWS、微软、谷歌<br>万片规模上流GPU集群资源 |
| 大模型开发框架 | 腾讯PatricStar、达摩院Whale、华为MindSpore、百度PaddlePaddle | 微软DeepSpeed、谷歌Tenseflow、Pytorch |
| 大模型 | 智源研究院“悟道”GLM（1300亿参数）<br>华为盘古CV（2000亿参数）<br>阿里达摩院M6、通义千问（10T参数）<br>Baidu文心一言（ERNIE Bot 2600亿参数）<br>浪潮“源”（2450亿参数）<br>腾讯混元（1T参数） | OpenAI、Huggingface、Stability AI、Meta OPT&LLaMA、<br>Google LaMDA/MUM, PaLM, Microsoft Megatron-T、Databrick Dolly<br>趋向垂直领域的小规模参数模型如金融、医疗等 |
| 应用 | 内场应用为主<br>SaaS发展滞后，集中在文本和图片生成相关的文化与艺术类场景<br>科大讯飞、拓尔思等 | 内场应用，SaaS（Jasper、Stability、Play.ht、Midjourney） |
| 模型 | 主要以开源模型为基础，如GPT-1/2，Megatron-T5 | 模型以GPT-3, GPT-4规模和技术为核心 |
| 自研算力加速卡 | Baidu 昆仑芯3，Tencent，阿里 | Meta MTIA v1(7nm 128G LPDDR5 FP16 51.2TFlops, INT8 102TOPs, TDP 25W)<br>Microsoft Athena<br>Google TPU v4 |

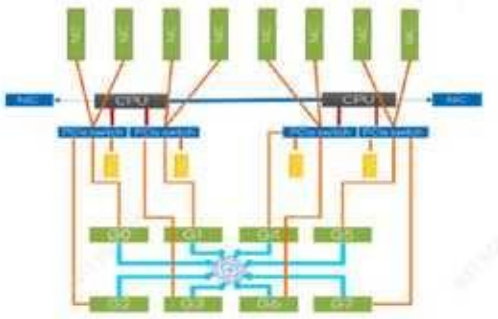
## 国内大模型偏好：看好智谱、复旦，在部分研究场景有优势。



# 大模型对算力架构的影响

目前可以参考的有英伟达的GPU （A100和H100），Google的TPU+Tensorflow，华为昇腾Atlas800-MindSpore，英特尔SPR+Habana/XeGPU



- 硬件挑战
  - 单卡：更大的算力，更多高速内存
  - 多卡：高带宽、低延时卡间互联
  - 集群：分布式，长时间稳定运行，高效利用所有硬件资源
- 软件挑战
  - 系统和集群管理
  - 大规模分布式训练和推理的算法和框架建设及优化
    - Data Parallel
    - Model Parallel
    - Pipeline Parallel
    - Activation Checkpointing
    - Offloading
    - …
  - 巨量数据预处理及优化

# 大模型部署的关键影响因素

## 部署技术：

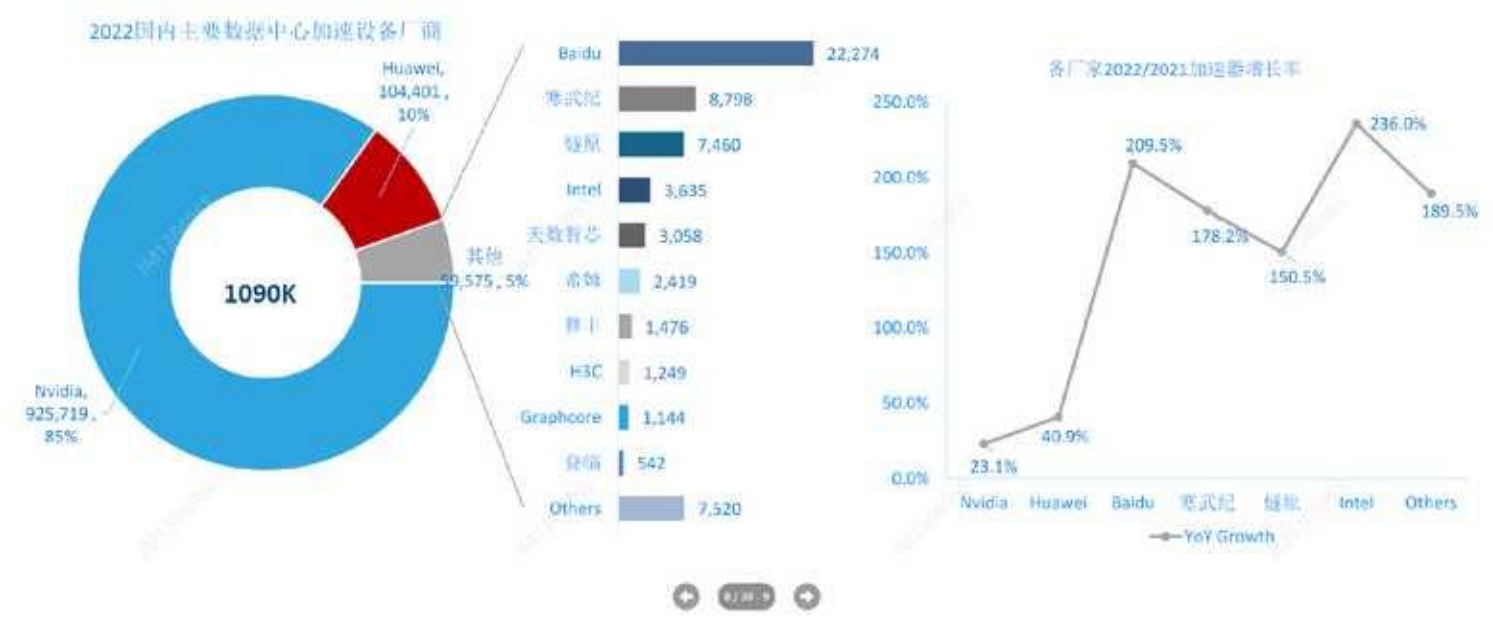
- 硬件成本和能效将影响用户的选择，相对V100、A100的性价比提升
- 单芯片硬件利用率，片上内存（on-chip memory）和容量将能够减少延迟
- 芯片到芯片的扩展影响模型和数据并行。芯片间互连的带宽（PCIe，专有，CCIX，以太网）
- 系统到系统的扩展，IB，以太网或其他

## 使用成本：

- 算力（硬件或云）和数据（成本和质量）是大模型的核心
- 日常大模型Fine Tune的成本约为150万美金/次（训练$0.0300 / 1K tokens；推理$0.1200 / 1K tokens ）
- 应用价格随用户访问量线性增长，每个word $0.0003，每次提问成本约2-5美分。

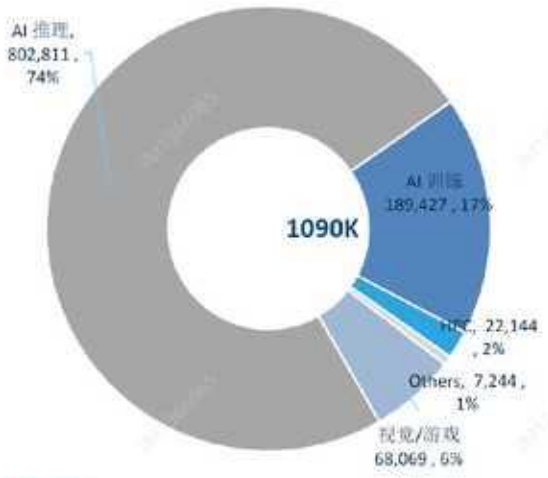
硬件出货量：23年市场增速快，英伟达季度增速更快（原因考虑是单季客户爆发性增长，全年可能不如Q2）；24年之后算力普及，增速进入平稳。


| 全部数据中心加速卡 | 1,089,695 | 1,387,334 | 1,687,744 | 1,998,417 | 2,216,550 | 2,425,292 |
|---|---|---|---|---|---|---|
| 增长 | 28.7% | 27.3% | 21.7% | 18.4% | 10.9% | 9.4% |




训练和推理比例：国内训练和推理比例一般是1:4，而国外1:7

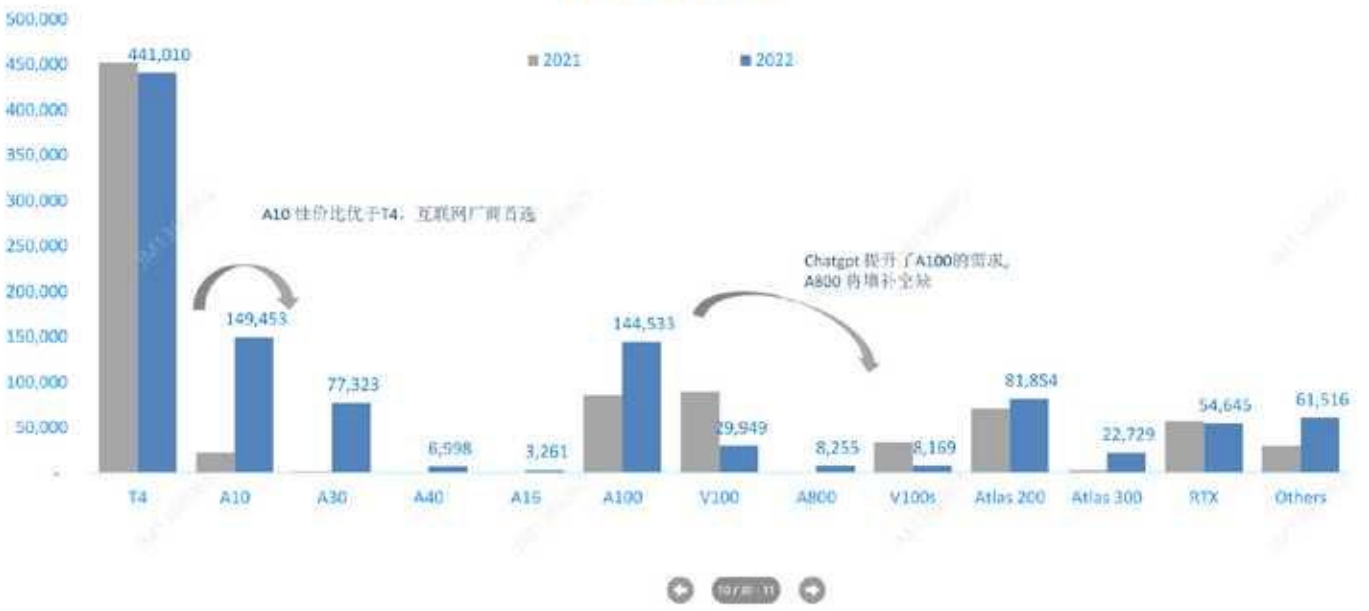
2022国内主要数据中心加速设备




按模型统计

| AI 训练 | | AI 推理 | | 视觉/游戏 | | HPC | |
|---|---|---|---|---|---|---|---|
| A100 | 127K | T4 | 427K | RTX3090 | 20K | A100 | 14K |
| V100 | 26K | A10 | 164K | RTX3080 | 13K | V100 | 2K |
| A800 | 8K | Atlas 200 | 81K | RTX3060 | 7K | A40 | 0.9K |

市场增长率




目前认为只有GPGPU/ASIC能满足大模型算力场景；存算一体也能用，其他像CPU、DPU、FPGA已经不太能使用大模型场景。

# 算力芯片大模型性能比较

| | GPGPU/ASIC | CPU | DPU | FPGA | 存算一体芯片<br>如 Cerebras |
|---|---|---|---|---|---|
| 训练FP32精度 | 支持，多数在几十TF以上 | 支持，多数十几个TF | 支持，几个TF | 不支持 | 支持 |
| 推理FP16精度 | 支持，多数在上百TF | 支持，多数几十TF | 支持，十几TF | 不支持 | 支持，上百TF |
| 推理混合精度 | 支持 | 支持，性能一般 | 不支持 | 不支持 | 支持 |
| 内存 | 支持片上内存 | 不支持片上内存 | 片上内存空间有限 | 片上内存空间有限 | 片上内存空间有限 |
| 片间互联 | 支持 | 支持，仅有双路或4路16G/s | 不支持 | 有限支持 | 有限支持 |
| 主要特点 | • 数量众多的运算单元，采用极简的流水线进行设计，适合计算密集、易于并行的程序<br>• 具备片间并行计算能力 | • 管理、调度能力强<br>• 运算单元数目相对较少，适合相对复杂的串行运算 | • 计算能力有限，只能针对特定任务提供卸载<br>• 不支持并行化<br>• 开发环境不成熟，开发工作量大 | • 灵活性高，可定制开<br>• 延迟低<br>• 可以支持并行计算<br>• 成本和技术壁垒高 | • 计算单元多，支持多核并行计算<br>• 延迟低<br>• 可以支持训练和推理 |
| 针对大模型的支持 | • 满足大模型在计算精度，算力，并行化和内存优化的要求<br>• 软件生态成熟，针对大模型已经有较长时间的实用案例。 | • 支持大模型训练但成本高，难度大<br>• 大模型推理性价比低于GPU和ASIC | • 可以配合CPU+GPU解决部分数据预处理需求<br>• 不能独立支持大模型推理和训练任务 | • 可用于小规模推理部署<br>• 价格较高，不适于大规模部署<br>• 计算能力和峰值性能不如GPU<br>• 效率和功耗上劣于专用芯片ASIC | • 支持大模型推理和训练任务<br>• 开发和部署成本高 |

# 当前人工智能领域的主流芯片厂商及型号

| | NVIDIA V100 | NVIDIA A100 80G SXM | NVIDIA H100 SXM | NVIDIA H100 PCIe | AMD MI250X | Habana Gaudi2 | Intel Xeon HBM |
|---|---|---|---|---|---|---|---|
| TDP | 300W | 400W | 700W | 350W | 560W | 600W | 350W |
| FP64 | 7.8 | 9.7 | 30 | 24 | 47.9 | N/A | 7.8 |
| FP32 | 15.7 | 19.5 | 60 | 48 | 47.9 | 116 | 15.7 |
| TF32 TF | N/A | 156* | 500* | 400* | N/A | 211 | 72.4 |
| FP16/F16 TF | 31.4/125 | 312* | 1000* | 800* | 383 | 433 | 209 |
| Int8 TOPs | N/A | 624* | 2000* | 1600* | 383 | 1792 | 430 |
| FP8 TF | N/A | N/A | 2000* | 1600* | N/A | 1792 | |
| GPU Memory | 32GB HBM2 | 80G HBM2e | 80GB HBM3 | 80GB HBM3 | 128GB HBM2e | 96GB HBM2e | 64GB HBM2e |
| Memory Bandwidth | 900 GB/s | 2 TB/s | 3 TB/s | 2 TB/s | 3.2 TB/s | 3.2 TB/s | 3.2 TB/s |
| LLC | 16 MB | 40 MB | 50 MB | 50 MB | 16 MB | 48 MB | 112 MB |
| GPU-to-GPU Bandwidth | 300GB/s | 600 GB/s | 900 GB/s | 600 GB/s | 800 GB/s | 600 GB/s | 48 GB/s |
| 价格区间（$） | 4-5K | 10K | 25K | 22K | 8K | 10K | 17K |

| | A30 | A2 |
|---|---|---|
| Process Node | TSMC 7nm | Samsung 8nm |
| TDP | 165W | 60W |
| FP64 | 5.16 | 0.14 |
| FP32 | 10.32 | 4.5 |
| TF32 | 82 | 9 |
| FP16 | 10.32 | 4.5 |
| INT8 | 330 | 36 |
| PCIe Interconnects | PCIe 4.0 x16 +NVLink | PCIe 4.0 x8 |
| Memory Capacity | 24 GB HBM2e | 16 GB GDDR6 |
| Bandwidth | 933.1 GB/s | 200 GB/s |
| 价格区间（$） | 5K | 1.5K |

- NVIDIA 宣传的2x TF32, FP16, BF16, Int8 和 FP8 性能源于Tensor Cores sparsity
- 推理市场X86 CPU占有>50%的市场份额

华为：昇腾910由于不太支持FP32，必须依赖华为自身软件生态、需要华为深度优化及代码移植，开源大模型很难在910上使用；920能够达到A100性能1.7倍水平，不过供货量可能不足，价格可能也居高。

阿里：产品低调，担心美国调查。

昆仑芯：年底计划做3代，目标是训练，但实际看可能更适合推理。

沐曦：N100出货几百片、几十万元，下一代产品C100，目标训练场景，性能对标H100，并且兼容CUDA，比较期待。

寒武纪：思元590整体算力综合性能大约是A100的70%，指令兼容性差，影响部署；思元590B下一代产品，性能指标大约是A100的1.5倍，但同样面临软件生态影响，以及供应链问题。

# 国内算力芯片的发展情况-1

| Vendor | Card | Processor | Type | Peek Performance | Power | Interface | Features | Fabrication | Comment |
|---|---|---|---|---|---|---|---|---|---|
| Graphcore | IPU M2000 | GC200 IPUs | Training | 62.5TFlops@FP32<br>250TPFLOPS@FP16 | 300w | PCIe Gen4.0x16 | 900MB SRAM(65TB/s)<br>IPU-LinkTM - 512Gbps | 7nm | |
| Huawei | Atlas300-9000 | 昇腾910 | Training | 256 TFLOPS@FP16<br>512 TOPS@INT8 | max 310w | PCIe Gen4.0x16 | 32GB HBM+16GB高级大容量内存<br>支持HCCS和100G RoCE v2 | 7nm+ | 受供应链影响较大<br>新产品 SMIC 920（910 chiplet）8月提供样品<br>主要适用于特定应用，<br>仅面向政府资助项目，<br>需要支持自有框架，供应链问题很大 |
| | Atlas300-3000 | 昇腾310 | Inference | card:<br>64 TOPS INT8<br>chip:<br>16 TOPS@INT8<br>8 TFLOPS@FP16 | 67w / 8w | PCIe Gen3.0x16 | H264/H264 64 decoders / 4 encoders<br>LPDDR4x32GB, | 12nm FFC | |
| Alibaba | | 含光800 | Inference | 800TOPS ?<br>resnet50 IPS@Watt:<br>78563@276<br>53983@108<br>37500@75<br>25000@50<br>12500@25 | TDP250W | PCIe Gen4.0x16 | | 12nm | 含光因不支持FP适用场景有限，已停产。<br>GPGPU对标H100预计2023年上半年工程测试，无对外出货计划，目前项目暂停 |
| 昆仑芯 | 昆仑芯2代 | R200 | Inference | 128Tflops@FP16<br>256 TOPS@INT8 | 120W | PCIe Gen4.0x16 | 集成了ARM处理器<br>GDDR6 | 7nm | 合计出货约2万片，主要用于Paddle场景 |
| 壁仞 | Antoum-S30 | Antoum | Inference | | | PCIe Gen4.0x16 | 最高20倍稀疏 | 12nm | ResNet-50 95784fps |
| 沐曦 | N100/C100 | | Training/<br>Inference | 50TFlops@FP64 | 600W | PCIe Gen4.0x16 | | 7nm | 目标H100，宣传兼容CUDA，2023年初full mask（推理+训练） |

## 国内算力芯片的发展情况-2

| Vendor | Card | Processor | Type | Peek Performance | Power | Interface | Features | Fabrication | Comment |
|---|---|---|---|---|---|---|---|---|---|
| Cambricon 寒武紀 | 思元590 | 思元590 | Training | 24TFlops@FP32<br>200TFlops@FP16/BF16 | 400w | PCIe Gen4.0x16 | 80GB HBM2e<br>内置 MLU-Link 512GB/s | 7nm | 算力指标1.3倍，综合性能>70%英伟达A100<br>590B（590 chiplet）性能对标英伟达A100。<br>指令集兼容性差，主要依赖政府资助，客户体验很差，以项目投资入股为主 |
| | 思元370-X8 | 思元370 | Training | 24TFlops@FP32<br>96TFlops@FP16/BF16<br>128TOPS@INT16<br>256TOPS@INT8 | 250w | PCIe Gen4.0x16 | 48GB LPDDR5<br>内置 4个50Gbps MLU-Link<br>200GB/s聚合带宽<br>内置视频编解码 | 7nm | |
| | 思元270-F4/S4 | 思元270 | Inference | 128TOPS@INT8<br>256TOPS@INT4<br>64TOPS@INT16 | 150w/75W | PCIe Gen3.0x16 | 16GB DDR4 | 16nm | |
| | 思元100-C | 思元100 | Inference | 16 TFLOPS@FP16<br>32 TOPS@INT8 | 110w | PCIe Gen3.0x16 | 8GB/16GB<br>decoder | 16nm | |
| Enflame 燧原 | 云燧I20 | 邃思2.5 | Inference | 20TFLOPS@FP32<br>80TFLOPS@FP16<br>80TFLOPS@BF16 | 225w | PCIe Gen4.0x16 | 16GB HBM<br>200GB ESL互联技术 | 12nm | 1代表现一般，性能只有V100的50%；2代已经量产，性能达到A100的70%左右；TSMC投片成功<br>内出货2K，主要依赖自经政府项目 |
| | 云燧T21 | 邃思2.0 | Training | 32TFLOPS@FP32<br>128TFLOPS@TF32/FP16/BF16<br>256TOPS@INT8 | 300w | PCIe Gen4.0x16 | 16GB HBM2E<br>200GB ESL互联技术<br>OCP OAM | 12nm | |
| 壁仞科技 | BR104-300W | BR104, BR100 | Training<br>Inference | 128/240TFlops@FP32<br>256/480TFlops@TF32<br>512/960TFlops@BF16<br>1024/1920TOPS@INT8 | 300/550W | PCIe Gen5.0x16 | 150MB SRAM<br>32/64GB HBM2E<br>0.819/1.64GB/s 带宽<br>OAM架构，支持CXL，内置视频编解码<br>180GB/s互联 | 7nm | 自研BIRENSUPA平台，BR100受美国禁令重新设计，降低性能，预计6月提交美国商务部申请许可 |
| 登临科技 | Goldwasser | Goldwasser XL | Inference | 128TFlops@FP16<br>512TOPS@INT8 | 200W | PCIe Gen3.0x16 | 64GB 内置视频编解码 | 16nm | 软件定义的片内异构 |
| 天数智芯 | | 智铠100 | training | 37TFlops@FP32<br>147TFlops@FP16/BF16<br>147TOPS@INT16<br>295TOPS@INT8 | 300W | PCIe Gen4.0x16 | 32GB HBM2 | 7nm | 主要是落地政府AI数据中心，10K的订单，软件使用效果不 |

景嘉微、芯动、摩尔线程、兆芯：都是GPU，整体性能都比较低，除了摩尔线程能满足AI小模型训练和部分推理场景，但软件不太行，只有小量测试使用；其他3家都不太能用于AI。

海光：类似AMD第一代产品MI50，性能类似英伟达P100，软件生态不错，可以用于大模型；但海光整体策略放在HPC领域中，在AI领域没有多少投入，使用案例少。

## 国内算力芯片的发展情况-3

| | 景嘉微 | 芯动 | 摩尔线程 | 兆芯 | 海光 | AMD | 英伟达 |
|---|---|---|---|---|---|---|---|
| 型号 | JM9系列 | 风华一号 | MTT S3000 | Arise-GT10C0 | 深算一号 | MI100 | A100 |
| 制程 | N/A | 12nm | 12nm | 28nm | 7nm | 7nm | 7nm |
| 核心数量 | NA | NA | 4096 MUSA | NA | 4096（64CU） | 7680（120CU） | 2560 CUDA Core<br>640 Tensor Core |
| 时钟频率 | 1.5GHz | NA | 1.9GHz | NA | 1.5GHz（FP64）<br>1.7GHz（FP32） | 1.5GHz（FP64）<br>1.7GHz（FP32） | 1.53GHz |
| 显存容量 | 8GB | 8GB/16GB | 32GB | 4GB | 32GB | 32GB | 80GB |
| 显存类型 | DDR4 | GDDR6/6X | GDDR6 | DDR4 | HBM2 | HBM2 | HBM2e |
| 显存带宽 | 25.6GB/s | 304GB/s | 448GB/s | 25.6GB/s | 1024GB/s | 1228GB/s | 2039GB/s |
| FP32算力 | 1.5TFlops | 5/10TFlops | 15.2TFlops | 1.5TFlops | 11.5TFlops | 46.1TFlops | 19.5TFlops/156TF |
| 总线接口 | PCIe4.0x8 | PCIe4.0x16 | PCIe5.0x16 | PCIe4.0x16 | PCIe4.0x16 | PCIe4.0x16 | PCIe4.0x16 |
| GPU互联 | None | None | None | None | 184GB/s（xGMIx2） | 276GB/s（Infinity Fabric） | 600GB/s（NVLink） |
| TDP | 50W | 50W | 35W | 50W | 350W | 300W | 400W |
| 等效性能 | GTX1080 | RTX 2060 | RTX3060 | GTX1050 | P100 | V100 | |
| 软件 | OpenCL | OpenCL | CUDA on MUSA | OpenCL | ROCm（Radeon Open Compute platform） | | |

## 国内算力芯片在大模型应用中的问题

**当前国产算力芯片在运行大模型时的设计不足：**

- 从技术和经济角度衡量，训练产品与A100仍有1-2代硬件差距，推理产品进入主流企业仍有较高门槛：1.3-2x TCO，同时面临美国半导体法案的限制；
- 缺乏片间和系统间互联的解决方案（以免费CCIX为主，生态不完整，缺少实用案例，无NV-Link类似的协议），主要以单机或单卡为主；并行化和线性扩展能力还无法验证，尤其是大规模部署时的线性性能；
- 大模型多数需要在专有框架下才能发挥性能，软件生态差距明显，移植灵活性，产品易用性与客户预期差距较大；
- 产品研发能力（设计与制程），核心IP（HBM，接口等）等不足，阻碍了硬件的性能提升。

**建议：**

- 大模型推理和训练都对加速芯片计算能力和片上内存容量提出直接要求，成熟和高效的片间互联解决方案，与网络和存储厂商合作实现内存一致性，减少数据拷贝，降低延迟，解决系统间互联，确保集群规模化的线性性能；
- 提供完善的软件栈，支持模型与数据并行；
- 未来关注Chiplet、互联在中国大算力芯片中的机会，系统厂商在异构计算（一体机）和AI服务器架构设计（散热）的作用。

# 国内大模型发展问题

- 国内大模型都是基于GPT-1,2开源模型开发。缺乏对模型技术和架构的深入研究，主要是增加参数和中文数据集，与GPT-3/4技术差距比较明显。
- 开源模型普遍存在准确度的问题，会造成使用中额外的成本和影响用户体验。支持大模型的框架较少，并且不开放使用，不利于生态和合作推广，不利于未来的发展。
- 国内数据开放程度不高，数据孤岛现象严重，不支持爬虫或者开放数据搜索，影响大模型泛华。
- 国内算力芯片除了升腾可以用于特殊订制化大模型训练外，没有可实用的解决方案，推理需要解决性价比的问题。
- 大模型开发者要在数据安全、合规、伦理等方面解决使用者的顾虑。

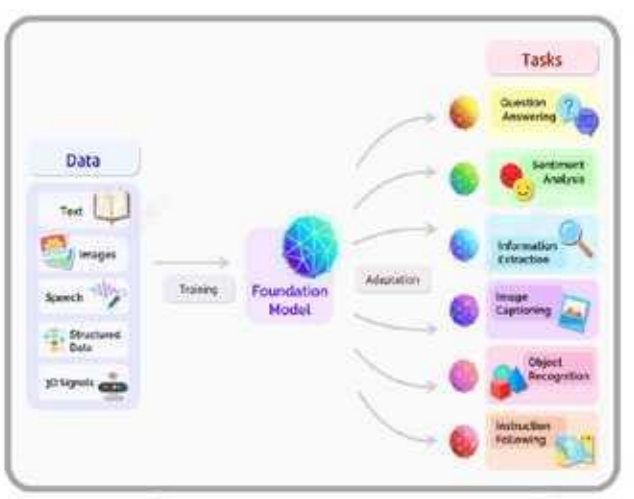
# 总结

- 大模型的发展带来了AI上下游产业的一轮风潮，如：CPU、GPU、FPGA、 DPU、IP、PCB、Chiplet、CPO（共封装光学），并延伸至硅光（Silicon photonics）等领域。这些领域是否真正会伴随大模型爆发。
- AIGC的发展对软件和应用的影响还没有全面爆发，未来与AIGC相关领域的机会（传媒、电商、影视娱乐、教育、金融、医疗与工业制造）。
- 美国对中国的高科技限制政策不仅对半导体硬件，还对软件的发展产生了巨大影响。
- 国内企业的能力和策略将决定大模型技术演进、发展和应用。
- 目前与国外企业尤其是大模型技术相关领域，中美企业差距较大。
- 关注具备 AI 模型、算法技术优势的科技公司；具备对 AI 商业化实际应用场景理解和业务优势的厂商；具备 AI 硬件优势的厂商


ChatGPT相对GPT-3的进化更多是巧妙利用了人工标注，而非深刻模型本身的架构突破，基本原理仍然是根据已有的文本，生成一个符合人类书写习惯的下一个合理内容

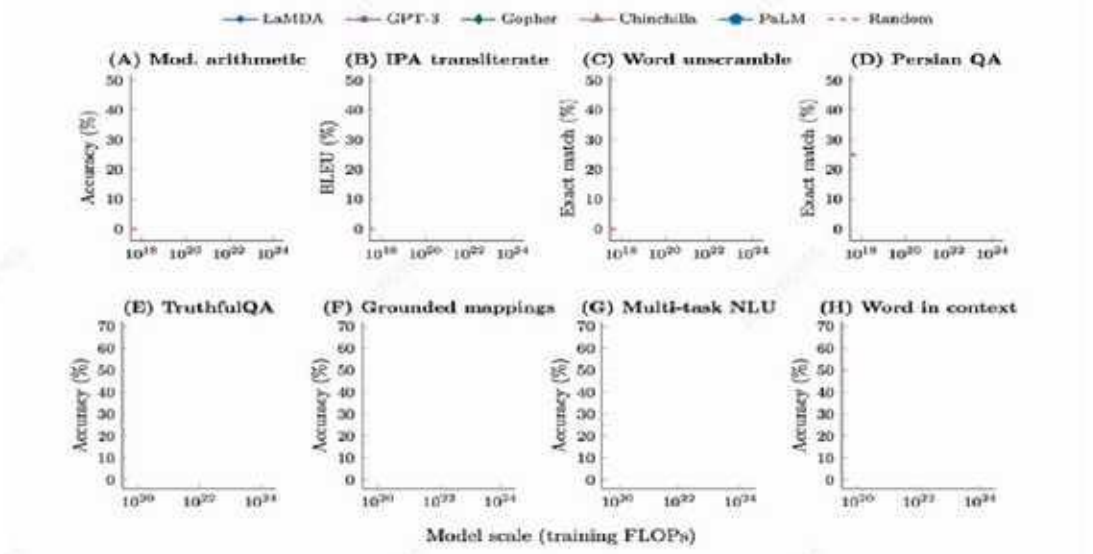
# 当前大模型的市场

- 大模型可以集中来自各种模态的所有数据信息，然后可以适用于广泛的下游任务（多模态）。
- 预训练大模型 + 下游任务微调。大规模预训练可以有效地从大量标记和未标记的数据中捕获知识，通过将知识存储到大量的参数中并对特定任务进行微调，极大地扩展了模型的泛化能力。
- 随着新的大模型的提出，数据规模和模型规模的不断增大，模型精度也得到了进一步提升准确率不断取得突破的趋势。例如：机器对自然语言理解能力的不断提升。
- 大模型的自监督学习方法，可以减少数据标注，在一定程度上解决了人工标注成本高、周期长、准确度不高的问题。



| | 模型 | 总计算量（PFlop/s-day） | 总计算量（Flops） | 参数量（百万个） | 令牌数量（十亿） |
|---|---|---|---|---|---|
| T5 模型 | T5-Small | 2.08E+00 | 1.80E+20 | 60 | 1000 |
| | T5-Base | 7.64E+00 | 6.60E+20 | 220 | 1000 |
| | T5-Large | 2.67E+01 | 2.31E+21 | 770 | 1000 |
| | T5-3B | 1.04E+02 | 9.00E+21 | 3000 | 1000 |
| | T5-11B | 3.82E+02 | 3.30E+22 | 11000 | 1000 |
| BERT 模型 | BERT-Base | 1.89E+00 | 1.64E+20 | 109 | 250 |
| | BERT-Large | 6.16E+00 | 5.33E+20 | 355 | 250 |
| | ROBERTa-Base | 1.74E+00 | 1.50E+21 | 125 | 2000 |
| | ROBERTa-Large | 4.93E+01 | 4.26E+21 | 355 | 2000 |
| GPT 模型 | GPT-3 Small | 2.60E+00 | 2.25E+20 | 125 | 300 |
| | GPT-3 Medium | 7.42E+00 | 6.41E+20 | 356 | 300 |
| | GPT-3 Large | 1.58E+01 | 1.37E+21 | 760 | 300 |
| | GPT-3 XL | 2.75E+01 | 2.38E+21 | 1320 | 300 |
| | GPT-3 2.7B | 5.52E+01 | 4.77E+21 | 2650 | 300 |
| | GPT-3 6.7B | 1.39E+02 | 1.20E+22 | 6660 | 300 |
| | GPT-3 13B | 2.68E+02 | 2.31E+22 | 12850 | 300 |
| | GPT-3 175B | 3.64E+03 | 3.14E+23 | 174600 | 300 |

## 大模型部署的问题与挑战-1




- 各类大模型应用都需要强算力
- 算力随准确度呈指数增长
- 准确度源于模型参数与数据

## 大模型近期发展方向

### 大模型生态

- 目前部分深度学习框架，如Pytorch和Tensorflow，没有办法满足超大规模模型训练的需求，微软基于Pytroch开发了DeepSpeed，腾讯基于Pytroch开发了派大星PatrieStar，达摩院基于Tensoflow开发的分布式框架Whale，华为昇腾的MindSpore、百度的PaddlePaddle，基于原生的AI框架支持超大模型训练。
- 大模型开源生态，Huggingface、Stability AI、Meta开源的OPT（175B）、阿里巴巴达摩院的中文模型开源社区“魔搭”（ModelScope）等。

### 大模型应用场景

- 智源研究院“悟道”GLM（1300亿参数）大模型用于冬奥手语播报数字人
- 华为盘古CV（2000亿参数）大模型针对无人机电力智能巡检
- 阿里达摩院开发的M6大模型智能生成内容文案，多模态特征提取，用于客服应答和商品标记
- Baidu文心一言（ERNIE Bot 2600亿），浪潮“源”（2450亿参数）

# 当前大模型的市场 GPT-4

| | GPT-4 | 相比GPT-3.5的改进 |
|---|---|---|
| 模型发布时间 | 2021年开始研发<br>2022年8月完成训练<br>2023年3月14日上线 | 2022年12月上线 |
| 参数规模 | 未知，内容窗口支持32000个token（约25000单词或50页文本） | 175B （3000单词或6页文本） |
| 训练 | SFT（Supervised FineTuning）预训练、RLHF奖励模型训练、PPO（近端策略优化）算法微调<br>数据集截止到2021年9月<br>经过对抗性测试，训练高速稳定 | 在PPO微调新采用了基于规则的奖励模型（RBRM） |
| 模型输入特点 | 多模态语言模型<br>支持多种语言文字、图形和视频的输入和文本输出 | 文本输入和输出，以英文为主，其他语言准确率不高 |
| 功能特点 | 增强的推理、学习，“涌现”上下文学习和理解能力 | 复杂和细微任务处理<br>安全性（50余人安全专家）和一致性 |
| 潜在应用场景 | 生成、编辑和迭代 | 文字生成、编辑与应答 |
| | 模型、硬件、训练、数据集未开放 | 除模型外其他信息开放 |

# 大模型未来发展方向—应用层

| 分类 | 具体分支 | 应用领域 | 代表公司案例 |
|---|---|---|---|
| 文本生成 | 文本理解 | 话题解析、文本情感分析 | 科大讯飞，阿里巴巴和微软亚洲研究院在文本理解挑战赛的完全匹配得分均超过人类得分 |
| | 结构化写作 | 新闻撰写 | Automated Insights开发的Wordsmith可以生成评论文章 |
| | 非结构化写作 | 营销文案，剧情续写 | Jasper平台为社交媒体、广告营销、博客等产出标题，文案，脚本，文章等 |
| | 交互性文本 | 客服，游戏 | OpenAI与Latitude推出的游戏AI Dungeon，可根据输入的动作或对话生成个性化内容 |
| 音频生成 | 语音克隆 | 地图导航 | 百度地图可根据输入音频，生成专属导航语音 |
| | 语音机器人 | 客服，销售，培训 | 思必驰拥有外呼机器人、呼入机器人、陪练机器人等产品 |
| | 音乐生成 | 播客，电影，游戏 | OpenAI的MuseNet可利用10种乐器共同生成4分钟音乐作品 |
| 图像生成 | 图像编辑与融合 | 设计，电影 | 谷歌的Deep Dream Generator可上传图像并选择风格，生成新图像 |
| | 2D图像生成3D模型 | 游戏，教育，产品测试 | 英伟达的GANverse3D可利用汽车照片生成3D模型，并在NVIDIA Omniverse中行驶 |
| 视频生成 | 画质增强修复 | 视频插帧、视频细节增强、老旧影像的修复与上色 | 当虹科技的画质增强修复技术帮助视频画质提升 |
| | 切换视频风格 | 电影风格转换，医学影像成像效果增强 | 腾讯天衍工作室在结直肠内镜项目中切换视频风格，优化医学影像视觉效果 |
| | 动态面部编辑 | AI换脸 | Akool的faceswap平台拍摄样本视频便可编辑、替换模特面部 |
| | 视频内容创作 | 制作电影预告片、赛事精彩回顾 | IBM的Watson制作了20世纪福克斯的科幻电影《Morgan》的预告片 |
| 跨模态生成 | 文本生成图像 | 传媒，娱乐 | OpenAI的DALL E 2可通过输入文字生成高仿真图像 |
| | 文本生成视频 | 电影，短视频创作 | Meta的Make-A-Video输入文本可生成数秒的视频 |
| | 图像/视频生成文本 | 搜索引擎，问答系统 | 谷歌的MUM模型支持多模态复杂信息搜索 |
| | 文本生成代码 | Copilot | OpenAI的Codex模型可将自然语言翻译成代码 |

GPT-3用预测下一个字的方式来表达智能，根据上段文本生成下段文本

由人工标记出一些问题，并写出参考答案给AI，用这些标注好的问答数据集去微调GPT-3模型，“收集示例数据，训练一个有监督的模型”

“收集比较数据，训练一个奖励模型”

“根据奖励模型，对有监督模型进行持续的强化学习”

RLHF（从人类反馈中进行强化学习）