Global Computing Consortium 全球计算联盟

Open AI Infra A Chapter of GCC

2026 Open AI Infra Summit

兆瓦级算力系统技术探索与规划

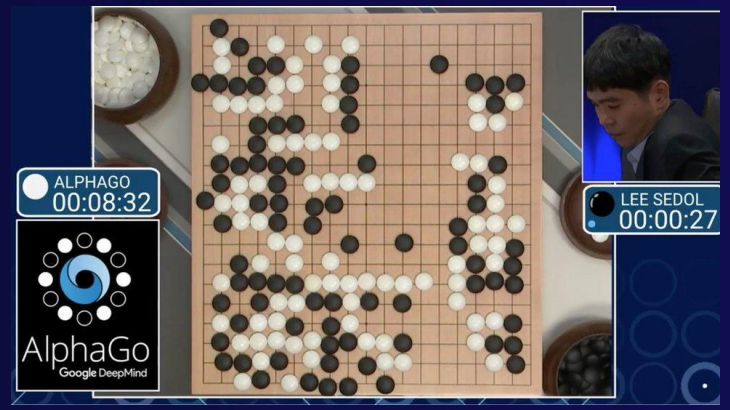
龙盘

OAII社区管理委员会联席主席

华为公司计算产品线研发副总裁

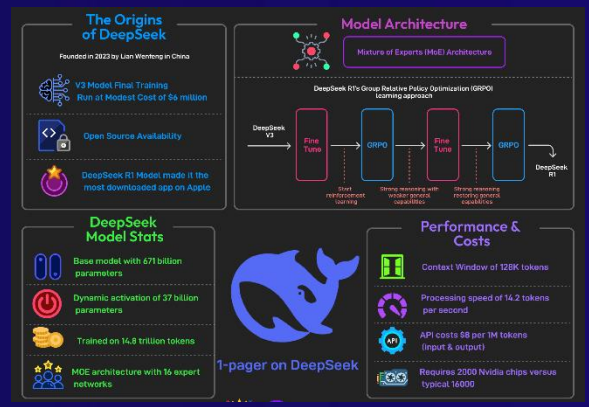
Open创新理念，AI日新月异，Infra持续探索

Open AI Infra
A Chapter of GCC

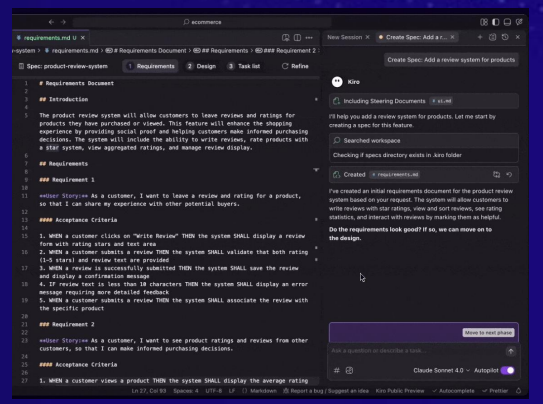


AlphaGo战胜世界冠军

ChatGPT开创聊天机器人

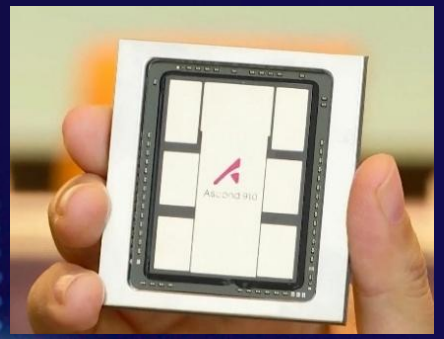


Deepseek极大降低成本



AI生产力时代即将到来

2016 2019 2023 2024 2025 2026



昇腾处理器

灵衢互联协议

液冷数据中心

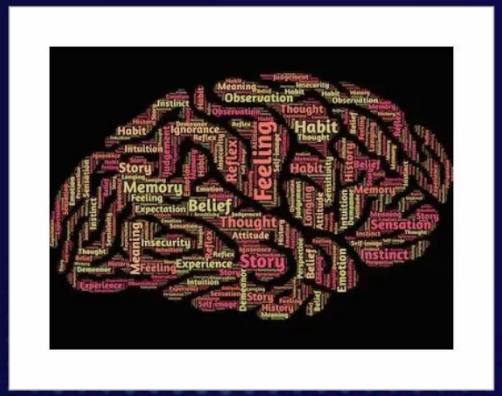
支撑智能体具备生产力的要素

学富五车

- ✓ 万亿级参数，万亿级语料；
- ✓ 10万卡集群，海量带宽；
- ✓ 全机运行98%以上可用度；

思维敏捷

- ✓ 首Token时延，每秒Token数；
- ✓ MoE模型，大EP并行架构；
- ✓ 低延迟，大带宽，内存语义；



过目不忘

- ✓ M级长序列，个性化长记忆；
- ✓ 100TB级共享内存池；
- ✓ 10PB级海量KV缓存池；

品行端正

- ✓ 运行中对齐，敏感词过滤；
- ✓ 测试中思维，多链对比；
- ✓ 海量向量数据库高效查找；

AIDC的特征：兆瓦级算力系统，吉瓦级数据中心，重构供电/散热/光互连

兆瓦级算力系统给AIDC带来的挑战与机会

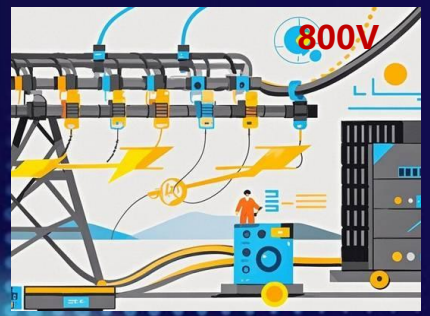
Open AI Infra
A Chapter of GCC

系统技术

1）算力芯片微指令；2）分级内存子系统；3）芯片组协同架构；4）超节点互联协议；5）并行计算软件框架；6）DFx技术；

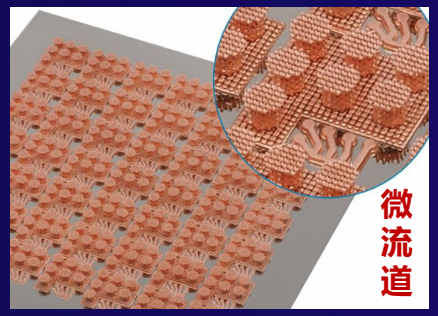
高效供电

- ✓ 功率跃增后需要提高母线电压，降低铜缆损耗、提升转换效率；
- ✓ 800V机房 and Sidecar供电两种模式，在较长时间内共存；
- ✓ 建设投资：5~8RMB/瓦；



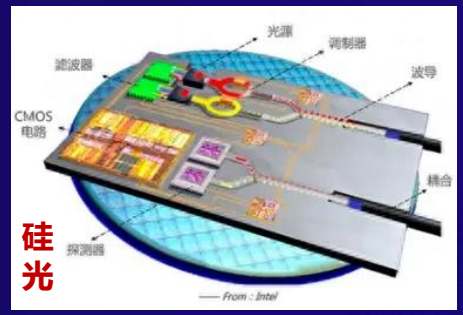
闭环液冷

- ✓ 微流道等先进换热技术对杂质颗粒、离子浓度更敏感；
- ✓ 工质环路内置检测/告警/过滤功能，成为可信赖的闭环系统；
- ✓ 建设投资：5~8RMB/瓦；



高密光互连

- ✓ 智算系统互连带宽需求以HBM带宽为锚点，等比例增长；
- ✓ Scale-Up：HBM的10~20%；
- ✓ Scale-Out：HBM的1~2%；
- ✓ 光互连占算力投资的5~8%；



谢谢您的聆听！