燧原科技宣布升级企业战略——全面打造AIGC时代的基础设施

上海2023年3月14日 /美通社/ -- 燧原科技宣布升级企业战略：以全栈软硬件和集群产品为数字底座，结合MaaS (Model as a Service) 的业务模式，全面打造人工智能技术生成内容 (AIGC) 时代的基础设施。

一、燧原大模型技术与产品深耕已久

燧原科技自创业之初就瞄准云端训练产品市场，以突破高难度训练芯片为目标，2019年推出的第一代产品云燧T10/T11是国内第一款具备FP32高精度算力的训练产品。其创新的片间高速互联，以及软件栈聚焦优化数据并行与模型并行等大规模集群的训练能力，奠定了今天燧原的训练产品在国内大规模集群训练场景的领先优势。

基于云燧T11的冷板式液冷方案，燧原科技为之江实验室构建了超千卡规模AI液冷集群，以赋能文本到视频生成的场景为目标，成功支持了GPT-2、源1.0及实验室自研蛋白质结构预测等多个超大规模巨量模型的高效训练。液冷智算集群也顺应国家的绿色低碳环保的要求，PUE经实测最低可降至1.08。

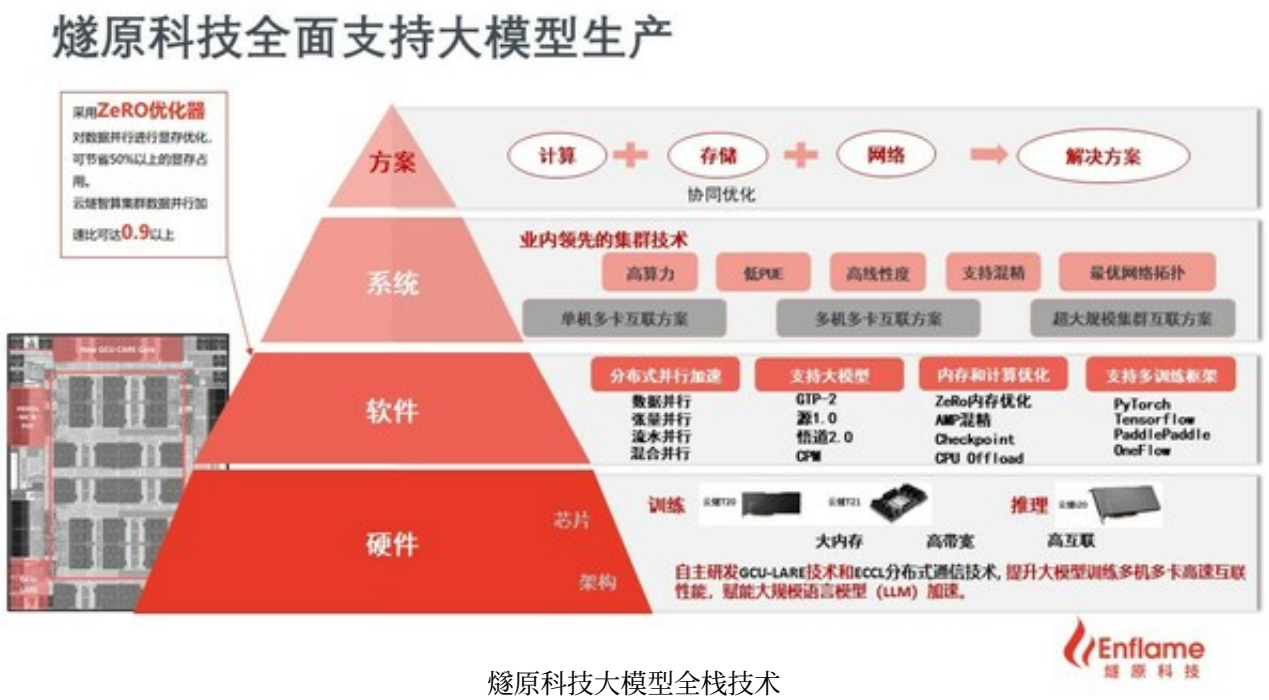
历经五年的产品迭代和优化，燧原科技现已拥有从硬件、软件到系统的全栈解决方案，结合云燧训练和推理产品在行业落地打磨的实践经验，可为客户提供丰富多样的人工智能系统软硬件产品，全方位降低AI算力中心部署和应用成本。

二、厚积薄发，燧原科技发力AIGC

在大模型标杆项目落地经验指引下，以大幅缩短大模型开发与应用周期为目标，针对大模型场景下的算力需求特性，燧原科技现已针对大模型场景形成从硬件、软件、系统方案的全栈技术，全面支持大模型生产，包括但不仅限于：

- 大模型现有生态接入：支持PyTorch、TensorFlow、PaddlePaddle、OneFlow、Megatron-LM、FairScale等主流AI框架和分布式加速库，支持GPT-2、源1.0、悟道2.0、CPM等主流AIGC大模型的Pretrain和Finetune。
- 大模型极致性能提供：采用自主研发的GCU-LARE技术和ECCL分布式通信技术，提升大模型训练多机多卡高速互联的性能，支持数据并行、模型并行、流水线并行和混合并行等并行加速功能，支持Activation Checkpointing、ZeRO优化器、CPU offload、AMP（自动混合精度）等算力和显存优化方法，可快速高效地进行大模型训练。
- 大模型训练TCO优化：从大模型应用端到端、技术全栈角度，燧原科技提供一体化大规模AI算力集群方案 -- 云燧智算机（CloudBlazer POD），方案采用一体化设计，是专为人工智能场景下计算、存储、网络、软硬协同设计的标准化产品，以"全局优异"为目标，大幅降低建设满足大模型场景需求的AI算力基础设施的总拥有成本 (TCO)。
- 大模型业务高性价比推理加速：凭借云端推理产品云燧i20通过互联网社交应用服务上亿规模用户的成功经验，同时与广泛的落地场景进行打磨，燧原的推理产品在支持Stable

Diffusion、GPT-2、T5等AIGC大模型推理上具备高性价比，加速AIGC相关下游场景的商业落地。



燧原科技大模型全栈技术

燧原科技创始人、COO张亚林表示："人工智能技术的发展正在步入一个全新的阶段，AIGC内容生成类模型以及所生产的内容生动反映了人工智能从感知、认知进阶到生产，也正在重构互联网商业模式，催生数字经济新突破。在以ChatGPT和Stable Diffusion为代表的AIGC技术浪潮下，燧原科技凭借在大模型训练及推理的产品技术优势，结合MaaS的业务模式，构建AIGC时代的基础设施底座。"

原文地址：http://www.china-nengyuan.com/news/192863.html