曙光 GPU 计算平台推动咪咕公司在人工智能领域产品研究

案例简介

咪咕公司利用 NVIDIA GPU 建立人工智能高性能计算集群，加速在图像、语音识别等人工智能领域的技术和产品研究。

Case Introduction

MIGU company used NVIDIA GPU to build the high performance computing cluster for artificial intelligence in order to accelerate the technology and product research in the field of artificial intelligence, such as image recognition, speech recognition, etc.

背景

咪咕文化科技有限公司是中国移动面向移动互联网领域设立的，负责数字内容领域产品提供、运营、服务一体化的专业子公司，下设咪咕音乐、咪咕视讯等 5 家子公司。公司中大量的产品在往智能方向转化，需要在图像、视频和语音等领域进行大量的技术探索和研究，为满足计算需求，计划建设人工智能 GPU 计算平台，具体要求如下：

1. 整套 AI 平台需分为训练模块和推理模块，以满足咪咕的算法研究和线上业务；

2. AI 平台训练模块需满足整个咪咕公司的算法研究工作，用于支撑公司产品中的视频、图像和语音等领域的算法调优和开发，训练模块 GPU 整体的单精度计算能力不低于 1.68 Pflops；

3. AI 平台推理模块需满足咪咕公司的线上业务，用于支撑产品上线前的大量测试和产品上线后的高吞吐量计算，推理部分 GPU 的整体单精度计算能力不低于 110 Tflops，同时需支持 INT8 类型，减少计算精度以加速推理速度；

4. 需提供专业的 AI 平台管理工具，支撑平台管理人员对平台硬件的实时监控和管理；

5. 提供专业远程运维工具，能够远程对集群进行监控，可以协助平台运维人员对集群进行故障定位和故障处理等；

挑战

咪咕公司在音乐、视频、阅读、游戏、动漫数字内容等多个业务板块有众多的智能产品，需要大量的算法研究和开发工作，为了满足产品的快速迭代，算法的调优时间需要尽可能的缩短，所以对 AI 平台训练模块的计算能力提出了更高的要求；

咪咕公司每天都有大量的新数据产生，为了能从这些数据中挖掘更多有价值的信息，会有大量的数据参与网络的训练，为了能够更好的拟合这些数据，单任务会使用多个计算节点，对节点之间的网络提出了更高的要求；

咪咕公司有着大量的线上业务，为了能得到更好的用户体验，需要对大量的数据进行线上推理，为满足实时性，特别是对于计算量大的推理任务，需要 GPU 作为支撑，同时在不影响准确性的前提下需充分利用 GPU 的计算性能，在某些应用下需使用 GPU 的 INT8 计算特性；

为满足大量的训练需求和线上业务的稳定运行，除了从软件层面制定容灾策略之外，对 GPU 服务器和 GPU 卡本身的稳定性也提出了更高的要求，能够满足 7*24 小时的高压稳定运行是必不可少的；

AI 平台有大量的硬件设备，对整个平台的运维和管理都是有挑战的，不仅对管理和运维人员本身要求高，同时需要专业的平台管理软件和运维工具。

方案

针对咪咕公司的具体需求，结合公司的实际业务场景，曙光公

司提供了一套基于 NVIDIA GPU 的人工智能平台解决方案。整个 AI 平台分为训练和推理两个模块，对应咪咕公司的算法研究和线上业务。

训练模块采用曙光 W580-G20 GPU 服务器，内置 4 颗 NVIDIA P40 GPU，单节点可以达到 48Tflops 的单精度计算能力，采用双万兆互连，满足大量研究人员的接入使用和单任务的扩展性。

推理模块采用曙光 W580-G20 GPU 服务器，内置 4 颗 NVIDIA P4 GPU，单节点可以达到 22 Tflops 的单精度计算能力，同时针对推理应用的特殊需求，可使用 GPU 的 INT8 计算特性，单节点可以达到 88TOPS，同时为了满足高吞吐量，配置双万兆网络连接。

为了简化设备的管理和运维，配备曙光专业的平台管理软件 GridView，满足节点视图、节点监控和异常报警等多种功能，同时专业远程运维软件 EasyOP 可以协助咪咕运维人员完成集群运维和远程技术支持。

影响

为了满足本次咪咕公司的 AI 平台建设，曙光与咪咕进行了大量的前期测试与实验，先后独家中标咪咕文化和咪咕音乐先期实验性 GPU 服务器采购项目，对 GPU 在咪咕公司的后续采购做了前期准备。

本次曙光在咪咕 GPU 服务器集采项目的成功入围，证明了曙光在 AI 平台建设上的专业技术能力，同时对 NVIDIA GPU 人工智能计算平台在运营商中的落地产生重要影响。