解决方案简介

宝德联手英特尔打造支持 Analytics Zoo Cluster Serving 的自动分布式可扩展推理平台

概述

凭借在挖掘数据丰富的内在信息、拟合能力、数据扩展性等方面的优势，深度学习正在成为大量企业用户部署人工智能（AI）应用的优先选择。但与此同时，深度学习相较普通的机器学习应用，在算法部署、模型设计、算力基础设施构建等方面有着更高的门槛。

为帮助企业用户降低深度学习部署与应用门槛，构建软硬一体的分布式全栈推理平台，宝德推出了基于 AI 推理服务器 PR2715P2，可支持 Analytics Zoo Cluster Serving 的自动分布式可扩展推理平台。宝德 PR2715P2 采用了全新设计，融合了第二代英特尔® 至强® 可扩展处理器和英特尔® 深度学习加速（英特尔® DL Boost）技术，在性能与功耗之间实现了有效平衡，为深度学习应用提供了一个高效能的基础平台。

挑战：如何加速深度学习部署与优化

随着深度学习算法的不断创新，越来越多的应用需要对深度学习模型进行大规模和实时的分布式推理服务。虽然已经有一些工具可用于模型优化、服务、集群调度、工作流管理等相关任务，但是对于许多深度学习的工程师和科学家而言，开发和部署能够透明地扩展到大型集群的分布式推理工作流仍然是一个严峻的挑战。

为了便于构建和生成面向大数据的深度学习应用程序，英特尔推出了 Analytics Zoo 平台。该平台提供了统一的数据分析 + AI 平台，可将 TensorFlow、Keras、Pytorch、BigDl Spark、Flink 和 Ray 程序无缝集成到一个统一的数据分析流水线中，用于分布式训练或预测，方便用户构建深度学习应用。整个流水线可以透明地扩展到运行在由搭载英特尔® 至强® 处理器的服务器组成的 Hadoop/Spark 集群上，以进行分布式训练或推理。

Analytics Zoo 在较新的版本中还提供了对于 Cluster Serving 的支持，构建了轻量级、分布式、实时的模型服务解决方案。Analytics Zoo Cluster Serving 支持多种深度学习模型，提供了一个简单的发布/订阅 API，可支持用户可轻松地将他们的推理请求发送到输入队列。然后，Cluster Serving 将使用分布式流框架在大型集群中进行实时模型推理和自动扩展规模。

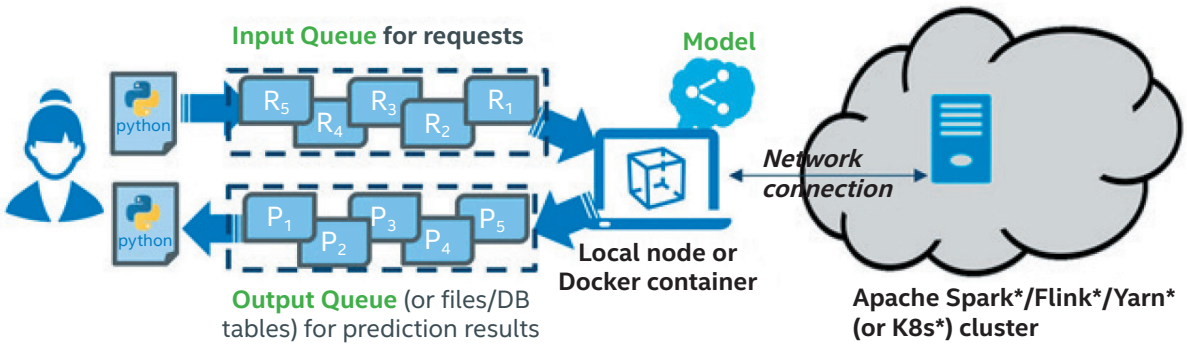

图 1. Analytics Zoo Cluster Serving 解决方案总体框架

要部署基于 Analytics Zoo Cluster Serving 的深度学习算法与应用，企业需要进行硬件选型、优化与验证，以提供高性能的算力支撑，但在此过程中，企业也面临着艰巨的挑战。首先，传统的解决方案并非是全栈设计，需要在硬件选型、软硬件适配与优化等方面耗费大量的时间与精力，也容易带来总体拥有成本（TCO）的上升。

其次，深度学习算法与应用对于 AI 性能有着很高的要求，未针对 AI 进行性能优化的 CPU 在运行效率上存在明显瓶颈。GPU 服务器虽然能够提供充足的算力支持，但是成本相对较高，应用范围受到较多的局限，而且其需要专门的部署与调优，不利于深度学习应用的快速上市。

解决方案：基于宝德 PR2715P2 AI 推理服务器的自动分布式可扩展推理平台

搭载第二代英特尔至强可扩展处理器的宝德 PR2715P2 AI 推理服务器全面支持 Analytics Zoo Cluster Serving 分布式推理技术方案。该推理服务器是一款软硬一体的分布式全栈推理解决方案，专为高性能计算、高级人工智能分析任务而设计，具有出色的性能功耗比。

图 2. 宝德 PR2715P2 AI 推理服务器

第二代英特尔至强可扩展处理器专为数据中心现代化革新而设计，能够提高各种基础设施、企业应用及技术计算应用的运行效率，进而改善总体拥有成本（TCO），提升用户生产力。它拥有更高的单核性能，能够在计算、存储和网络应用中，为计算密集型工作负载提供高性能和可扩展性。得益于英特尔® 超级通道互联（英特尔® UPI）、英特尔® Infrastructure Management 技术（英特尔® IMT）、英特尔® 高级矢量扩展指令集 512（英特尔® AVX-512）等领先功能，它可满足严苛的 I/O 密集型工作负载的需求。

此外，第二代英特尔至强可扩展处理器内置人工智能加速，并已针对工作负载进行优化，能够为各种高性能计算工作负载、AI 应用以及高密度基础设施带来一流的性能和内存带宽。同时，采用矢量神经网络指令（VNNI）的英特尔® 深度学习加速（英特尔® DL Boost）显著提高了人工智能推理的表现，与上一代产品相比，性能提升高达 14 倍。这使其成为拓展 AI 应用的卓越基础设施。

在宝德 PR2715P2 AI 推理服务器搭建的平台上，用户只需要准备 Analytics Zoo Cluster Serving 的 Docker Image、配置文件、训练好的模型（当前支持的模型包括 TensorFlow、PyTorch、Caffe、BigDL 和 OpenVINO™ 的模型）及推理数据，即可在几分钟内启动并运行推理应用。

通过加入对于 Analytics Zoo Cluster Serving 的支持，宝德一体化分布式可扩展人工智能推理方案显著降低了配置和流程的复杂程度，在提供高级定制化服务的同时，有效降低了 TCO。这种全新的集群模型服务支持有助于简化用户的分布式推理工作流，提高工作效率，并为深度学习场景带来领先的性能。

效果：自动分布式可扩展推理平台加速深度学习部署与应用

通过部署基于宝德 PR2715P2 AI 推理服务器的自动分布式可扩展推理平台，用户能够实现如下价值：

- **提升深度学习的部署速度：**得益于 Analytics Zoo Cluster Serving + 宝德 PR2715P2 AI 推理服务器的全栈推理解决方案，用户可以节省在平台搭建、软硬件优化与验证等方面的大量时间，快速部署深度学习应用。
- **提高系统整体算力：**第二代英特尔至强可扩展处理器具备强大 AI 性能，为该解决方案的算力奠定了坚实基础，在组建分布式集群之后，可以应对大型深度学习负载。
- **确保面向未来的扩展性：**该解决方案不仅在软件层面实现了自动扩展规模，还通过服务器的分布式部署提供了敏捷扩展能力，能够敏捷应对未来的深度学习对于基础设施的要求。

展望：宝德与英特尔协力推动 AI 技术发展

人工智能技术与应用是数字化转型的关键技术方向。多年来，宝德与英特尔等合作伙伴构建了繁荣的 AI 生态，不断推动软硬一体化的 AI 方案的创新，为机器学习、深度学习等技术提供基础设施支撑。同时，双方还不断针对行业需求拓展 AI 解决方案落地场景，深化产品与产业的应用融合。

基于 Analytics Zoo Cluster Servining 和宝德 PR2715P2 AI 推理服务器的自动分布式可扩展推理平台是宝德与英特尔合作的重要成果，不断证明着其在降低深度学习部署门槛、加速深度学习推理等方面的价值。双方还将进一步把高性能计算数据分析和人工智能加速整合到单一的计算环境中，并提供新的内存和存储模式，为计算引擎提供支持，进而解决高性能计算系统面临的独特挑战。

附：宝德 PR2715P2 AI 推理服务器配置

特性	PR2715P2 技术规格
形态	2U 机架服务器
处理器数量	1/2 个
处理器型号	第二代英特尔® 至强® 可扩展处理器
内存	24 DIMM 插槽，支持 2933MHz DDR4 的 RDIMM 内存，最大支持 3TB，可选支持 AEP 内存
硬盘	**前置：** 可支持 8 个（默认）/12 个 2.5/3.5 英寸 SAS/SATA/SSD 硬盘，或 24 个 2.5 英寸 SAS/SATA/SSD 硬盘 **内置：** 可支持 2 个 2.5 英寸 SAS/SATA/SSD 硬盘，或 2 个 2.5 英寸 U.2 SSD 硬盘，和 1 个 PCI-E M.2 SSD **后置：** 可支持 2 个 2.5 英寸 SAS/SATA/SSD 硬盘，或 2 个 2.5 英寸 U.2 SSD 硬盘
Raid 支持	支持 SATA RAID0、1、10， 可选配支持 SAS RAID0、1、10、5、50、6、60 等，RAID 无缓存/1 GB/2 GB 缓存， 可选缓存掉电保护
板载网络	集成 2 个 Intel x722 千兆 RJ45 网口；可选配千兆及万兆 OCP 模块/网卡
PCIE 扩展	默认提供 6 个 PCI-E3.0 标准插槽，1 个专用插槽（PCI-E3.0），最多提供 8 个 PCI-E3.0 标准插槽，1 个 OCP 专用插槽: **PCI-E 插槽 1（CPU0 引出）：** 默认转接卡支持 2* 全高 PCI-E3.0x8（in*16）插槽，1* 全高 PCI-E3.0x8 插槽，或可选支持 1* 全高双宽 PCI-E3.0x16 插槽，1* 全高 PCI-E3.0x8 插槽； **PCI-E 插槽 2（CPU1 引出）：** 默认转接卡支持 2* 全高 PCI-E3.0x8（in*16）插槽，1* 全高 PCI-E3.0x8 插槽，或可选支持 1* 全高双宽 PCI-E3.0x16 插槽，1* 全高 PCI-E3.0x8 插槽； **PCI-E 插槽 3（CPU1 引出）：** 可选转接卡支持 1* 半高 PCI-E3.0x8 插槽，或 2* 半高 PCI-E3.0x8 插槽； **专用插槽（CPU0 引出）：** 1* OCP 插槽（PCI-E3.0x8）
其他端口	USB3.0 接口：5 个（前部 2 个，后部 3 个） VGA 接口：2 个（前部 1 个，后部 1 个） 串行接口：1 个（后部 1 个） 管理网口：1 个（后部 1 个）
风扇	6 个热插拔冗余风扇
电源	标配 1 个 800W 白金交流电源模块；支持 1+0 单电源模式，或 1+1 冗余电源模式；电源模块可选 800W/1200W 白金交流电源模块，和 800W 直流 -48V 电源模块
管理	支持 IPMI2.0，对外提供 1 个 100/1000 Mbps RJ45 管理网口，支持 iKVM 远程管理
显示控制器	集成 ASPEED AST2500
CD-ROM/驱动器	选配 SATA/USB 接口光驱
支持的操作系统	Windows Server 2012 R2 Red Hat* Enterprise Advanced Server 7.5
供电	220V AC /240V DC
物理尺寸	高 87mm * 宽 438mm * 深 735mm，支持 19 英寸机柜，最低配置毛重约 19KG
环境及规范	
环境温度	运行时 10℃ 至 35℃ 非运行时 -40℃ 至 +55℃ 周围环境
相对湿度	非运行时 95%，于 25°C 至 30°C 温度下不凝结
噪声	运行模式中，于侧位测量声压 <50dBA；环境温 <28°C 时测得声强为 6.2BA
静电释放	每项英特尔环境温度测试规范 15KV
安全标准（中国）	CCC

关于宝德

深圳市宝德计算机系统有限公司成立于 2003 年，以服务器和 PC 整机研发、生产、销售和为客户提供云计算综合解决方案为主营业务，致力于成为中国领先的 IT 产品和解决方案提供商，为互联网、教育、广电、安全、金融、电信、税务、交通、电力、医疗等行业提供尖端的 IT 产品和服务。多年来，在强者林立的中国服务器市场，凭借先进的技术和独特的软硬件综合实力，宝德服务器市场占有率连续多年稳居国内前茅。

关于英特尔

英特尔（NASDAQ：INTC）作为行业引领者，创造改变世界的技术，推动全球进步并让生活丰富多彩。在摩尔定律的启迪下，我们不断致力于推进半导体设计与制造，帮助我们的客户应对最重大的挑战。通过将智能融入云、网络、边缘和各种计算设备，我们释放数据潜能，助力商业和社会变得更美好。如需了解英特尔创新的更多信息，请访问英特尔中国新闻中心 newsroom.intel.cn 以及官方网站 intel.cn。

本文并未（明示或默示、或通过禁止反言或以其他方式）授予任何知识产权许可。英特尔未做出任何明示和默示的保证，包括但不限于，关于适销性、适合特定目的及不侵权的默示保证，以及在履约过程、交易过程或贸易惯例中引起的任何保证。

英特尔运营所需的任何商品和服务预测仅供讨论。就与本文中公布的预测，英特尔不负有任何购买责任。本文中提供的所有信息可在不通知的情况下随时发生变更。关于英特尔最新的产品规格和路线图，请联系您的英特尔代表。

英特尔技术特性和优势取决于系统配置，并可能需要支持的硬件、软件或服务得以激活。产品性能会基于系统配置有所变化。没有任何产品或组件是绝对安全的。更多信息请从原始设备制造商或零售商处获得，或请见 intel.com

在特定系统的特殊测试中测试组件性能。硬件、软件或配置的差异将影响实际性能。当您考虑采购时，请查阅其他信息来源评估性能。关于性能和基准测试程序结果的更多信息，请访问：www.intel.com/benchmarks

英特尔并不控制或审计第三方数据。请您自行审核该等内容、咨询其他来源，并确认提及数据是否准确。

英特尔、英特尔标识以及其他英特尔商标是英特尔公司或其子公司在美国和/或其他国家的商标。文中涉及的其它名称及品牌属于各自所有者资产。

© 英特尔公司版权所有