观察文章

人工智能如何重新定义计算系统与架构

作者：Infra 观察家

人工智能时代的到来，使得数据及计算能力的重要性再次升级，以往 x86 架构下以 CPU 为核心的计算性能提升到达瓶颈，呼唤计算力的重构。

IBM 推出“认知系统”：面向 AI 提供“一步认知”的 Power 架构

去年 9 月，IBM 推出了面向高性能计算的 IBM Power Systems S822LC 服务器（Minsky 服务器），其使用的 NVidia 开发的通信协议 NVLink，能够迅速在 CPU 及 GPU 之间建立连接，IBM 成为了目前首家同时是唯一一家采用这一技术的公司。作为唯一拥有 CPU:GPU NVLink 的架构，Minsky 服务器使用 NVIDIA Tesla P100 解决高性能计算及人工智能对于计算能力提出的新挑战，在加速计算性能的同时，增强系统的可编程性和可访问性，并消除 PCI-E 瓶颈。目前，这一架构已经实现两倍于 x86 系统的性能优势。

基于这一面向高性能计算而优化的计算系统，IBM 进一步推出了“认知系统”(Cognitive Systems)，面向机器学习、深度学习、自然语言处理、实时高级分析等人工智能相关工作负载，通过 BlueMind 深度学习云平台、PowerAI 深度学习框架，以及使用 NVLink 技术的 Minsky 服务器提供一个硬件+软件整合的解决方案，为主流人工智能框架更顺畅地运行在 Power Systems 上提供可能。

面向数据而设计的 IBM Cognitive Systems 包括以下几大亮点：

- **PowerAI 深度学习框架：**PowerAI 包含了主流深度学习软件框架，例如用于模型训练的 TensorFlow、Caffe、Torch、Theano 以及关联库的 cuDNN 等，通过不断基于 Minsky 服务器优化性能，PowerAI 为主流深度学习框架工具包交付企业级的支持。并帮助开发人员提升易用性、面向数据科学家缩短模型训练时间。

Powered By IBM

- **BlueMind 深度学习平台：**基于 Spark 大数据平台框架，能够进行深度学习平台资源管理、调度，拥有优异的并行效率和扩展性能，并且具有丰富的深度学习功能，可帮助用户在集群或云环境中快捷高效地开发和部署深度学习应用。
- **Spectrum conductor 软件定义架构：**软件定义基础架构解决方案，最大程度发挥系统潜力，并降低与网络和集群计算相关的直接成本。
- **OpenCAPI 标准：**通过 OpenCAPI 联盟对于这一开放式标准的探索，OpenCAPI 标准总线将加快数据在数据中心各个层面的迁移速度，每个通道的数据都可达 25Gbps，从而进一步提升面向数据密集型工作负载的 Power 系统的性能优势。

计算系统与架构持续演进——计算系统的重塑迫在眉睫

IBM Cognitive Systems 的推出，与近来计算系统的演变趋势紧密相关。以往，数据是集中式的，传统的计算逻辑往往采取程序化的计算方式，以 CPU 为中心提升计算性能，探索如何有效加速计算流程从而达到更好的数据处理效果。然而随着分布式数据、非结构化数据逐渐成为主流，计算正在从“程序化的计算”向“以数据处理为中心的计算”演进。对于数据价值的挖掘也更需要内外部结合，将数据整合在一起提供更以数据为中心、更具效率与智慧的计算，而非仅仅是对数据的流程化的处理。

人工智能相关负载需要高性能的数据传输，并需要具有最佳准确度的训练模型，以高效、快速地找出“超级参数”，从而大幅节省模型的训练时间。因此，AI 堆栈的基础首先是正确的硬件：带有加速器的服务器，以及正确的存储设备。

GPU 加速计算良好地适应了深度学习训练“计算密集”这一特性，具有最高 CPU-GPU 带宽的服务器能够实现高性能的数据传输，而这一点正是规模更大、更为复杂的深度学习模型所需的优势。

正如 **IBM 认知系统高级副总裁 Bob Picciano** 而言：“IBM 认为，基础架构世界不再以 CPU 为核心，而是从以 CPU 为中心追求计算性能，转向追求整体计算系统的效能。未来，

计算能量不仅是存在于CPU，而是广泛涉及到包含GPU、FPGA甚至内存计算等其他设备。IBM在探索的是，如何利用新的技术，让这些计算能力从以往的程序式计算向未来的有关认知的计算模式转型，这也是IBM在未来有关计算的发展方向。IBM非常看重在认知时代下对计算系统的重塑，希望通过IBM Cognitive Systems，确保CPU、GPU、I/O、内存等结合在一起，一站式地提供客户应用人工智能所需的计算能量。”

合力创新为人工智能架构开辟用武之地

除不断革新计算架构以外，IBM还不断通过OpenPOWER基金会，联合合作伙伴共同让面向AI的创新更具用武之地。通过与 OpenPOWER 基金会成员合作，IBM 推出了OpenPOWER LC系列的三种新服务器。IBM通过与NVidia及赛灵思合作，加速计算性能的提升及连接效率，已经实现了比x86高3-5倍的CPU和GPU间I/O带宽，使机器的训练时间从几天缩短到几小时甚至几分钟。

此外，由IBM 、AMD、Google、Mellanox以及Micron创建的OpenCAPI联盟将进一步探索如何打开CPU之间的链接，通过全新的“OpenCAPI”标准满足高性能异构计算的需求，并将在今年下半年发布的Power9服务器中率先应用OpenCAPI，促进OpenPOWER基金会成员开展基于OpenCAPI的创新。

在前不久举办的NVidia GTC开发者大会中，AI及深度学习技术不仅成为了会议演讲的主题焦点，也在其发布的多项革新性产品中呈现基础性地位。而近日，赛灵思宣布，和IBM 联手利用PCI Express Gen4，超越目前广泛采用的PCI Express Gen3标准，率先将加速器和CPU之间的互联性能提升一倍。可见在AI应用场景成为大势所趋的今天，联合开展新技术探索将能收获事半功倍的效果。正如**IBM大中华区硬件系统部服务器解决方案副总裁施东峰**所言：“IBM希望通过OpenPOWER基金会，将我们合作伙伴的创新成果得到最大程度的展现，并将其转化为AI机会。我们要确保集众人的智慧，探索出最佳的整合架构，为我们的客户提供在认知计算时代下的有效的应对方案。”

了解更多有关IBM Cognitive Systems（IBM认知系统）的信息，请访问以下链接
http://www-03.ibm.com/systems/cn/power/hardware/hpc/index.shtml

-完-