案例研究

英特尔® 至强® 处理器E5产品家族
英特尔® 高级矢量扩展（英特尔® AVX）指令集
公有云服务
人工智能在线服务

云上智多星

UCloud借力英特尔技术，打造易部署、低TCO的人工智能在线服务

“企业要构建自己的AI在线服务系统并非易事，无论是IT基础设施的建设还是AI框架的部署，都需要耗费大量人力、物力。如果在IT系统、AI框架上选择失误，则会前功尽弃，这给AI项目的发展和普及制造了很高的门槛。我们的目标是帮助用户像使用云主机、云存储这些成熟的云产品一样使用AI在线服务。通过充分利用英特尔® 至强® 处理器E5产品家族的高可扩展性及英特尔® AVX，我们的UAI-Service正逐渐走近这一目标。”

叶理灯
创新产品线总监
UCloud

上海优刻得信息科技有限公司（以下简称UCloud*）是目前国内领先的公有云服务提供商之一。它深耕云计算行业多年，在全球拥有19个数据中心，对互联网、移动互联网、传统企业在不同场景下的业务需求有着深入了解，可以提供计算、网络、存储、CDN、数据分析等多种IT全局解决方案。截至今天，UCloud已为5万余家企业级客户提供了服务，业务范围覆盖游戏、电商/O2O、互联网金融、互联网医疗、音视频等19个细分领域，间接服务的用户数量超过10亿，部署在UCloud平台上的客户业务总产值逾千亿元人民币。

UCloud不仅是一家云计算企业，也是一个资源整合的创新云服务平台。它通过与产业链上下游优质企业的通力合作，打造出健康、稳定的一站式服务生态。在人工智能（Artificial Intelligence，AI）如火如荼的今天，很多初创企业和传统企业都选择以AI为契机开拓市场，但同时也面临着缺乏高效部署AI能力的难题。为此，UCloud基于英特尔® 至强® 服务器平台，充分发掘和利用英特尔® 高级矢量扩展（英特尔® AVX）指令集相关处理单元的潜能，推出了UCloud AI 在线服务（UCloud AI online Service，UAI-Service*），其具备的大规模分布式计算平台可以满足企业在图像识别、自然语言处理等多个AI领域的在线服务应用需求。

面临挑战

企业的AI之路并非坦途大道：无论是初创企业踏上AI创新之旅，还是传统企业希冀借助AI之力调转航向，实现转型或升级，AI系统的设计、部署和运维都需要巨大、多维度的投入且困难重重，在决策选型过程中稍有不慎，都会带来巨大的沉没成本，令许多企业望而生畏。

AI的高成本正侵蚀企业的总拥有成本（Total Cost of Ownership，TCO）：AI能力提升的背后，可能会给企业带来巨大的成本开支，如何在性能和成本之间达到平衡？这一问题让许多企业决策者感到苦恼。

解决方案

UCloud UAI-Service：面向初创企业、传统企业AI转型而生的UCloud UAI-Service，旨在提供易部署、易运维、更安全以及多AI框架支持的AI在线服务节点，可助力企业完成AI模型部署这一关键环节，并在图像识别、机器学习等多个AI领域满足企业用户的需求。

英特尔® 至强® 处理器E5产品家族及英特尔® AVX: 通过与英特尔的紧密技术合作, UAI-Service一方面巧妙地利用云主机中英特尔® 至强® 处理器E5产品家族的空闲处理能力, 将其英特尔® AVX能力用于支持和加速AI在线服务; 另一方面, 利用该处理器产品家族强大的可扩展性进行弹性部署, 用低成本获得高性能, 降低用户的TCO。

成果

真正推动AI技术的普及, 助其持续发展: UCloud推出的UAI-Service将身处技术"深闺"中的AI技术和应用进一步平民化、实体化。通过PaaS的方式, 让更多有志于在AI领域开拓进取的企业能够获取出色的AI部署能力, 进而让整个AI产业实现"小步快跑"的前进节奏。

更有效利用空闲计算资源、节约用户成本支出: UAI-Service创新地利用英特尔® 至强® 处理器E5产品家族的空闲处理能力, 是对空闲计算资源再利用的有效尝试, 其成功实践令成千上万的数据中心处理器空闲能力得以充分利用。这既降低了企业用户的TCO, 也达到了环保节能的效果。

横空出世的AlphaGo, 让AI成为近两年来人们持续关注的热点。而AI也正在走出象牙塔, 走近普通企业和大众, 并开始在经济和民生层面扮演起越来越重要的角色。从机器学习、模式识别到自动驾驶、机器视觉, 不但众多初创企业将AI研发作为扬帆起航的契机, 许多传统企业也将其作为自身转型升级所必备的利器。

但AI系统的建设并非易事, 企业AI系统的建设可分为"数据收集"、"模型训练"及"模型部署"三个步骤, 每个步骤都会带来复杂的IT系统建设及运维工作。随着各类大数据、云计算技术方案的日趋成熟, "数据收集"和"模型训练"的工作正逐渐转移到云上, 形成了成熟的云化方案, 而AI模型部署的云化还存在许多问题: 一方面, 多种多样的AI框架需要企业制订和执行不同的部署策略, 难免因此产生高昂的运营成本; 另一方面, 主要用于模型训练的GPU平台在模型部署中不仅部署成本较高, 而且在扩展性上的表现也不够理想。

UCloud推动的UAI-Service, 就是针对上述AI模型部署难题而生的创新方案。UCloud的工程师们创造性地利用了虚拟云主机上英特尔® 至强® 处理器E5产品家族的空闲计算资源, 借助英特尔® AVX的能力, 来提供专注于AI模型部署的AI在线服务。英特尔® 至强® 处理器强大的可扩展性也帮助UAI-Service获得了快速便捷部署的能力, 并显著降低了企业运行AI在线服务的成本支出。

让使用AI服务像使用云主机一样便捷

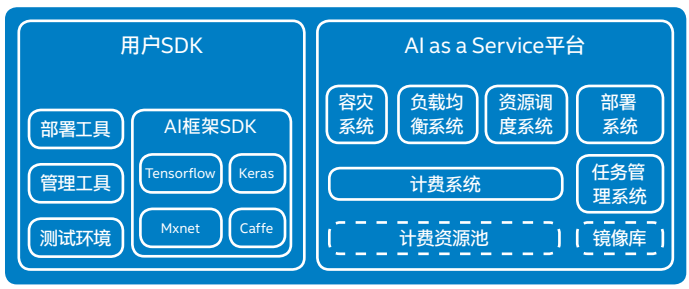
"简单来讲, AI的三部曲可以分为大数据收集, AI模型训练和AI在线服务。"UCloud创新产品线总监叶理灯这样描述企业AI系统建设, "此前, 针对前两步, UCloud都已经为用户提供了成熟的云主机、云存储、云网络等解决方案。"

图一 UCloud企业AI系统建设框架

但三部曲的最后乐章, 却还面临诸多问题。一方面, 企业用户在基于AI进行业务创新时, 常常面临众多的业务流程, 如何将不同的业务流程与AI在线服务一一映射, 这对AI在线服务的部署、可管理性及可扩展性提出了巨大的挑战; 另一方面, 面对众多的AI框架, 企业运维人员总有无从着手的烦恼, 因为他们需要为各个框架开发和配置不同的接口, 工作量巨大。为解决AI系统建设这"最后一公里"的问题, UCloud提供了UAI-Service, 它能基于大规模分布式计算平台为用户提供AI在线服务。

在实际任务部署中, UAI-Service为用户提供了"两步走"的部署模式。首先, 向用户提供SDK工具包, 内含接口代码框架、代码和数据打包模板以及第三方依赖库描述模板。用户只需根据SDK工具包内的代码框架编写接口代码, 准备好相关代码和AI模型以及第三方库列表, 就可以通过打包工具一键完成任务的在线部署。

任务打包完毕后, 用户可以通过UAI-Service分布式的AI在线服务PaaS平台进行后续管理和维护。该平台可以同时管理上千个计算节点, 每个计算节点都是同构节点, 具有相等的计算能力, 并拥有自动请求负载均衡、自动资源管理的功能。用户只需要将业务部署在平台上, 就无须操心其后续的运维。



图二 UAI-Service任务部署流程

"UAI-Service给用户带来的最大优势, 就是省去了部署AI在线服务时的大量繁琐工作, 让用户可以将宝贵的资源聚焦在自身的业务上。"在UCloud叶理灯看来, 如果每一个企业用户在部署自己的AI服务时, 都需要通盘考虑容灾、安全性、资源调度或者负载均衡, 那么企业在人力资源和成本上的支出将是沉重不堪的。UAI-Service将这些工作都内化为SDK包和PaaS平台服务, 用户只需要像使用云主机或者云存储服务那样, 轻松将所需的功能或服务

配置在一起就可以使用，而且UAI-Service还可以自动将分布式部署的四大要素——负载均衡、自动扩容、分布式容灾以及海量计算资源进行有效配置。

UAI-Service另一个重要优势是平台内数据的安全性。首先，由于UAI-Service在每个虚拟机上只部署一个AI节点，因此做到了各个AI任务之间的隔离；其次，由于平台本身并不涉及AI训练数据以及训练方法，在运行时仅需模型文件及接口代码，杜绝了数据外泄的可能；此外，UCloud还基于SDN技术实现了网络链路层的隔离，使每个客户的AI Service项目子网之间相互隔离，提升了网络安全性；最后，UCloud基于SDN技术实现了网络链路层的隔离，使每个客户的AI Service项目子网之间相互隔离，提升了网络安全性。在UAI-Service中，用户之间的AI模型和接口代码是安全隔离的，全自动化的部署过程使运维人员无权访问敏感数据，进一步提升了数据的安全性。

UAI-Service的通用性优势，解决了企业面对不同AI框架时的“选择障碍症”。UAI-Service对主流的AI框架，包括TensorFlow*、Keras*、Caffe* 和MXNet* 等都提供良好的支持，企业可以根据自己的业务需求来选择不同的AI框架进行接入。

在传统的AI框架以外，UAI-Service还与英特尔一起，引入了性能更佳的AI框架：面向英特尔® 架构优化的Caffe框架。这一版本的Caffe框架与传统AI框架相比，能更好地支持英特尔® 至强® 处理器产品家族和英特尔® 至强融核™ 处理器产品家族，并集成了最新版本的英特尔® 数学核心函数库2017，能更高效地利用英特尔® AVX的处理能力。

源自UCloud的一系列测试结果表明，借助面向英特尔® 架构优化的Caffe框架，测试系统同时运行的线程数量可以得到显著增加。基于该框架，测试系统的执行时间也能从最初未修改前的37秒缩短至优化后的3.6秒，整体执行性能提高了10倍以上。事实证明，通过采用这一框架，UAI-Service的AI在线服务效率得到了极大的跃升[1]。

以低TCO获取高效AI服务

现阶段，致力于AI开发和创新的企业，多为初创型企业，或者是正在谋求业务转型和升级的传统企业，因此对TCO的控制尤为敏感。如何获取高性价比的在线AI能力是企业用户们普遍关心的热点话题。

由于在图像识别、自然语言处理等AI正在发挥重要作用的领域中，往往需要用到大量的浮点运算，因此在人们的传统观念里，只针对浮点计算提供加速的GPU平台，似乎更适于AI系统的构建。但在AI模型的部署实践中，GPU动辄高达数万元人民币的售价极其昂贵，而且由于其扩展性不足，一旦部署，通常就只能固定执行单一的计算工作，难以随时根据工作任务的调整和变化实现及时的适配。

这样一来，UCloud就盯上了数据中心内大量部署的、每台服务器都会配备的通用处理器。“通过技术分析，我们发现虚拟云主机中的处理器，处于工作状态的主要都是简单指令集，而英特尔® 至强® 处理器集成的英特尔® AVX则并没有被充分利用。”UCloud叶理灯表示：“这意味着以浮点计算性能著称的英特尔® AVX的能力，或许可以为我们提供更适用的解决方案。”

英特尔® AVX是一套完整的单指令多数据（Single Instruction Multiple Data，SIMD）指令集规范，其最大的优势在于支持256位矢量计算，大大提升了处理器的浮点计算性能。其具备的增强数据重排能力，也能更有效地存储、读取数据。在充分认识到了英特尔® AVX及其处理单元的特性和优势之后，UCloud的工程师们开始了一项大胆的创新：利用各个虚拟机中此前未能“物尽其用”的英特尔® AVX能力，来满足AI在线服务的计算需求。

为了实现这一创举，UCloud与英特尔的工程师们携手优化了英特尔® AVX在AI在线服务中的应用表现，经过反复的优化与验证，AI在线服务的重要技术指标——时延被成功降低到了数百毫秒，完全能够满足UCloud用户的实际应用需求。

在时延这一性能指标达标的同时，英特尔® 至强® E5处理器产品家族出色的可扩展性也开始释放其强大的应用潜力。在数据中心内、服务器中配备的无数英特尔处理器都可以被扩展到系统中，来进一步强化AI在线服务所需的浮点计算能力，这是一种远比GPU方案经济高效得多的解决方案，毕竟，这些处理器节点已经是UCloud的既有投资，无需再为此多支出一分钱。

“这就是英特尔处理器强大的可扩展性带来的力量。在云计算平台上，处理器资源能够迅速地进行海量扩容，按我们目前的解决方案，即在每一个虚拟机上都部署一个AI在线服务计算节点，这意味着我们的AI在线服务未来可以根据用户需求得到迅速且海量的扩容能力，同时还不需要额外支付太多成本。”UCloud叶理灯满意地说。

[1]测试数据来自于英特尔网站：【面向英特尔® 架构优化的Caffe*：使用现代代码技巧】https://software.intel.com/es-es/node/702373?language=de）

为了对这一创新成果进行验证，UAI-Service已在200多个基于英特尔® 至强® E5处理器产品家族的虚拟机节点上部署了AI在线服务计算节点，验证测试的结果表明：基于英特尔至强处理器的AI在线服务计算节点完全能满足用户对性能的要求，在帮助用户有效降低TCO的同时，也顺应了数据中心环保节能的发展方向。

最新进展：让机器更快识别"喜怒哀乐"

在前期成功开发和测试的基础上，UAI-Service最近又在人脸识别技术的应用上实现最新进展。人脸识别一直是人工智能的一个重要研究方向，而基于英特尔至强服务器平台，利用英特尔® AVX来支持的UAI-Service，已在人脸表情识别的一系列测试中达成了优异的表现，验证了其能帮助零基础用户获得强大AI能力的潜能。

在测试中，UCloud选用了Tensorflow提供的TF-Slim实验库以及人脸表情识别公开数据库fer2013，其中共包含35887张人脸图片，各测试样本在年龄、面部等方面有较大差异性，这使该项技术测试具备了巨大的挑战性。

而测试结果表明：在有并发的前提下，UAI-Service AI在线服务的性能普遍高于8核8G云主机的性能，刚刚得到的测评数据表明，目前并发数为8-16个节点时，AI在线服务在性能上基本与基于GPU的方案相仿，这不仅说明在UAI-Service AI在线服务上部署人脸表情识别应用可以带来出色的成效，还证明基于英特尔® AVX支持的UAI-Service在人工智能应用中完全具备了与传统方案相媲美的能力。

结论：

以此前在UAI-Service上的成功协作为基础，UCloud未来还计划进一步深化与英特尔的创新协作，包括将最新的、面向英特尔处理器优化的AI框架引入UAI-Service在线服务平台，并将充分发掘新一代英特尔® 至强® 可扩展处理器的能力，特别是其集成的全新英特尔® AVX-512带来的更为强悍的浮点运算能力，来进一步优化AI在线服务，让专注于AI创新和应用的企业用户，能继续在合理的成本条件下，获取更强的AI计算能力支持。

经验：

AI在线服务的普及，不仅需要在技术上予以突破，其部署的便捷性、与现有云计算能力的结合程度以及在分布式集群上部署的可行性，也在深刻影响着企业用户的AI研发和应用进程。正是因为准确捕捉到了用户的痛点和具体需求，UCloud的UAI-Service才赢得了用户的青睐。

受数据中心内普遍存在的处理器计算资源闲置现象的启发，创造性地将其空闲的浮点计算能力投入到AI在线服务中，这不仅是技术上的创新，也是AI处理工作模式上的全新探索和尝试，它既能有效帮助企业用户降低TCO，也顺应了数据中心环保节能的发展趋势。

0917/SHA/PMG/XX/PDF

336454-001CN