解决方案白皮书

基于第四代英特尔® 至强® 可扩展处理器的用友智能 OCR 服务

挑战

为了满足特定服务级别协议（SLA）的要求，智能 OCR 的端到端推理流程需要保证在特定的时间内完成（例如小于 3 秒），需要强大的算力予以支撑。虽然基于独立 GPU 的推理方案能够满足性能需求，但也给企业带来了一些挑战：

独立 GPU 不仅采购或租用成本相对较高，而且部署和运维都会带来一定的成本。考虑到大量 OCR 场景不需要太高性能，采用 CPU 的方案将更具成本效益；

GPU 服务器通常是专用服务器，对于企业用户而言，如果升级为基于 GPU 的深度学习方案，就需要客户同时升级硬件平台，无法充分利用现有的 CPU 服务器资源；

要想充分释放 CPU 服务器的 AI 推理潜力，需要充分利用 CPU 的高级硬件能力与软件优化，更好地推动软硬件融合，提升性能表现。

解决方案概述

加速人工智能（AI）、大数据、云原生等数智化技术的落地，推动业务流程的数智化重塑，当前已经成为企业提升工作效率、挖掘数据价值、助力商业创新的重要方式。在此过程中，由 AI 赋能的智能光学字符识别（OCR）应用发挥着日益重要的价值。智能 OCR 能够将相当一部分的文字输入、单据识别等工作转为自动化流程，同时有效应对传统 OCR 技术在识别准确率、复杂环境应对能力等方面的缺陷。目前，智能 OCR 已经在金融、政府、制造、互联网、医疗等行业的单据识别、信息流入、图片翻译、车牌识别等场景得到广泛应用，帮助企业有效地释放人力资源，提升工作效率，为广泛的数智化应用提供基础能力支撑。

为了解决智能 OCR 应用在算力资源开销等方面的挑战，用友与英特尔进行合作，在用友商业创新平台（BIP）中，采用了基于第四代英特尔® 至强® 可扩展处理器的服务器作为智能 OCR 应用的基础算力设备。通过该方案的应用，用友能够通过 CPU 实现高速 AI 推理，不仅避免了独立 GPU 所带来的高昂支出，而且能够充分利用现有的 CPU 资源，实现了更高的灵活性、敏捷性。

集成于用友 iuap 智能中台的智能 OCR 应用

用友商业创新平台 YonBIP 是用友采用新一代信息技术，按照云原生（含微服务）、元数据驱动、中台化和数用分离的架构设计，涵盖平台服务、应用服务、业务服务与数据服务等形态，集工具、能力和资源服务为一体，服务企业与产业商业创新的平台型、生态化的云服务群。YonBIP 具有数字化、智能化、高弹性、安全可信、平台化、生态化、全球化和社会化八大特性，是企业通过数智化实现商业创新发展的使能平台。

“第四代英特尔® 至强® 可扩展处理器集成了英特尔® AMX 这一高级硬件特性，使得其人工智能加速能力远高于上一代产品。通过应用第四代英特尔® 至强® 可扩展处理器，并结合我们在深度学习模型方面的持续优化，我们为用户提供了以智能 OCR 为代表的高效、卓越的智能化服务，助力支撑企业智慧大脑，赋能管理者商业创新和智慧管理，提升员工工作效率与用户体验。”

方高林
用友智能中台总经理

用友 yonyou

iuap 作为全新一代商业创新平台的支撑底座，服务于成长型、大型及巨型企业数智化转型，助力企业提升数字化技术驾驭能力。其基于技术平台、数据中台、智能中台及业务中台，为企业提供了中台化构建能力、多云环境下的混合云开放集成互联互通能力、技术普惠化下的低代码开发和数智能力自助等应用快速构建能力。iuap 提供开放共享的生态连接，赋能客户、生态伙伴、社会，共享共创，成就数智企业，推动商业创新。

iuap 的 AIPaaS 智能中台提供了低门槛、向导式、高效率的 AI 服务，支撑企业全价值链、全场景的泛在智能和群体智能应用。智能 OCR 作为 AIPaaS 智能中台的一个典型 AI 应用，支撑了 VPA/RPA 等机器人智能服务，助力以数据驱动为核心的智慧管理。

图 1： 用友 YonBIP PaaS 云平台 iuap AIPaaS 智能中台

采用第四代英特尔® 至强® 可扩展处理器提升 OCR 推理性能

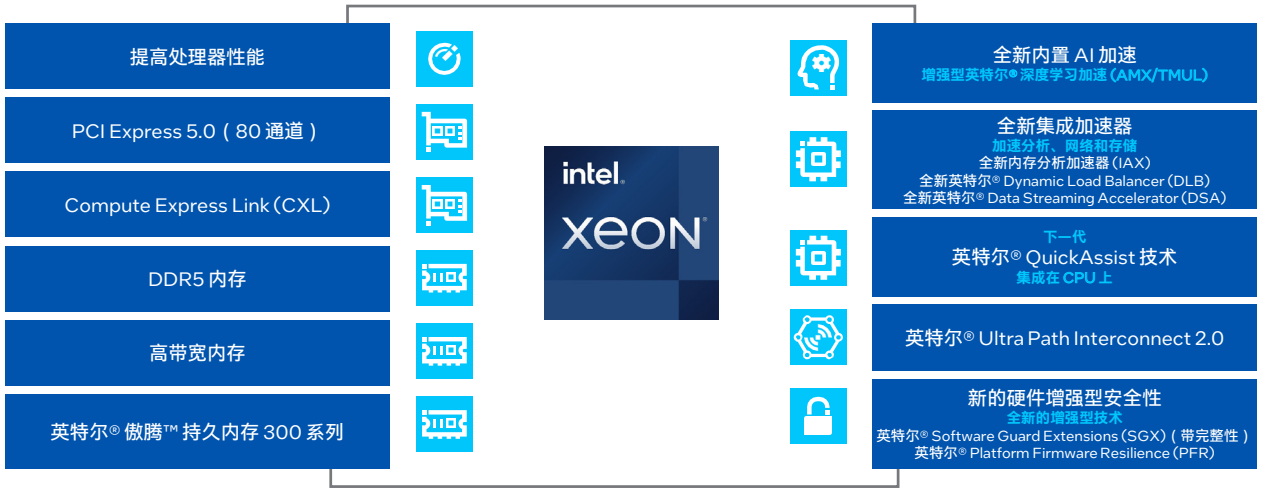
为了进一步释放 CPU 在 AI 推理流程的潜力，用友将基于第三代英特尔® 至强® 可扩展处理器的云服务器作为智能 OCR 模型推理的基础平台，并进行了性能验证，证明能够满足实际场景对于 OCR 推理的性能需求。

第四代英特尔® 至强® 可扩展处理器通过创新架构增加了每个时钟周期的指令，每个插槽多达 60 个核心，支持 8 通道 DDR5 内存，实现了 50%[1] 的内存带宽提升，并通过每 PCIe 5.0（80 个通道）实现了 2 倍[2] 的 PCIe 带宽提升，整体可实现 60%[3] 的代际性能提升。第四代英特尔® 至强® 可扩展处理器提供了现代性能和安全性，可根据用户的业务需求进行扩展。借助内置的加速器，用户可以在 AI、分析、云和微服务、网络、数据库、存储等类型的工作负载中获得优化的性能。通过与强大的生态系统相结合，第四代英特尔® 至强® 可扩展处理器能够帮助用户构建更加高效、安全的基础设施。

[1,2,3] 实际性能受使用情况、配置和其他因素的差异影响。更多信息请见 www.Intel.com/PerformanceIndex

第四代英特尔® 至强® 可扩展处理器在 AI 性能上更进一步。该处理器内置了创新的英特尔® AMX 加速引擎。英特尔® AMX 针对广泛的硬件和软件优化，通过提供矩阵类型的运算，显着增加了人工智能应用程序的每时钟指令数 (IPC)，可为 AI 工作负载中的训练和推理提供显著的性能提升。

用友充分利用了第四代英特尔® 至强® 可扩展处理器上集成的英特尔® AMX 特性，并将 OCR 模型从 FP32 量化到 INT8/BF16，以提升推理性能。



50%[4] 核数增加　50%[5] 内存带宽增加　2x[6] PCIe 带宽提升　60%[7] 代际性能提升

在测试中，用友首先对比了英特尔® 至强® 铂金 8380 处理器（第三代英特尔® 至强® 可扩展处理器，仅采用矢量神经网络指令 VNNI）与英特尔® 至强® 铂金 8480+ 处理器（第四代英特尔® 至强® 可扩展处理器，采用英特尔® AMX）的性能。测试数据如图 2 所示，得益于处理器基础算力的提升以及英特尔® AMX 的应用，第四代英特尔® 至强® 可扩展处理器将 OCR 模型的推理性能提升了 3.42 倍[8]。

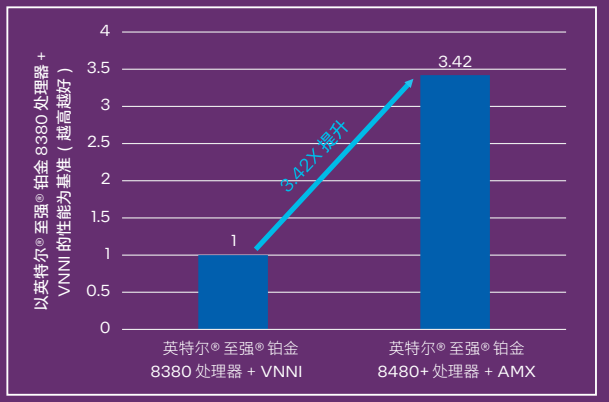


图 2. 第三代/第四代英特尔® 至强® 可扩展处理器 OCR 算法推理性能对比（时延 <30ms）

[4,5,6,7] 实际性能受使用情况、配置和其他因素的差异影响。更多信息请见 www.Intel.com/PerformanceIndex

[8] 截止 2022 年 8 月由英特尔开展的测试。测试配置 1：单节点，双路英特尔® 至强® 铂金 8380 处理器，40 核，启用超线程，启用睿频加速技术，256 GB 总内存（16 插槽/32 GB/3200 MHz），SE5C620.86B.01.01.0005.2202160810，0xd000375，Ubuntu 22.04.1 LTS，5.19.0-051900-generic，gcc 11.2，Yonyou OCR v1，OpenVINO 2022.2.0-custom_onednn2.6_9a3a3181e7056dcf7ccd3a16e599e6882a4edc23，Yonyou OCR v1，OneDNN 2.6。测试配置 2：单节点，双路英特尔® 至强® 铂金 8480+ 处理器，56 核，启用超线程，启用睿频加速技术，256 GB 总内存（16 插槽/32 GB/4800 MHz），EGSDCRB1.SYS.0085.D15.2207241333，0x2b000070，Ubuntu 22.04.1 LTS，5.19.0-051900-generic，gcc 11.2，Yonyou OCR v1，OpenVINO 2022.2.0-custom_onednn2.6_9a3a3181e7056dcf7ccd3a16e599e6882a4edc23，Yonyou OCR v1，OneDNN 2.6。实际性能受使用情况、配置和其他因素的差异影响。更多信息请见 www.Intel.com/PerformanceIndex

随后，用友还在基于英特尔® 至强® 铂金 8480+ 处理器的平台上，测试了将模型从 FP32 量化到 INT8/BF16 后的性能表现。测试数据如图 3 所示，量化至 INT8/BF16 使得模型性能提升了 7.3 倍[9]。

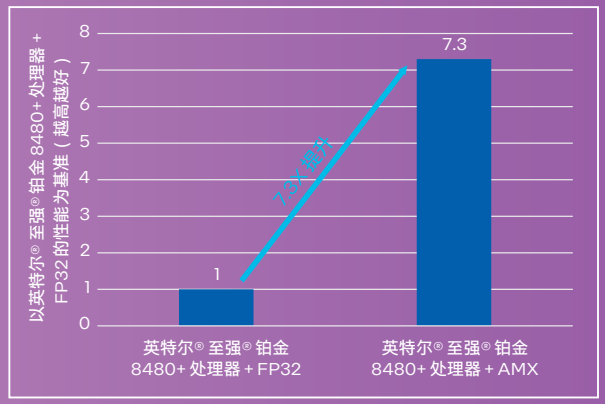


图 3. 第四代英特尔® 至强® 可扩展处理器使用 FP32 与 AMX 的性能对比

收益

- 通过性能优化，显著提升了基于 CPU 的 AI 推理性能，无需使用专门的基于 GPU 的硬件来进行推理，不仅能够降低硬件的采购成本，相应的空间、功耗、软硬件调优等成本也得到显著降低，有助于提升 OCR 应用的投资回报率（ROI）。
- 方案能够有效利用现有的 CPU 服务器资源，用户无需额外采购/租用 GPU 服务器，提升了基础设施的灵活性，也避免了 GPU 服务器部署、维护等带来的资源损耗。
- 该应用实践为用友 YonBIP 用户的云实例/硬件选型提供参考，用户可以根据实际的性能需求，选择更适用的云实例。

展望

在双方已有合作成果的基础上，用友与英特尔探索了第四代英特尔® 至强® 可扩展处理器在智能 OCR 推理中的巨大潜力，证明了第四代英特尔® 至强® 可扩展处理器在架构、高级硬件特性等方面的提升有助于加速 AI 负载。在切换到基于第四代英特尔® 至强® 可扩展处理器的基础设施平台之后，用友将有望提升智能中台在敏捷性、经济性等方面的优势，为企业带来更加高效的数智化服务。

面向新一轮商业创新的大潮，用友与英特尔将围绕 IT 基础设施重构与优化、数智化应用创新等方面进行广泛合作，聚合企业服务生态圈，为客户提供基于新一代数智技术、真正云原生技术架构、创新应用架构的平台和应用服务，提供随需而用的企业云服务和无处不在的数智价值，助力企业实现业务运营和管理的数智化升级与转型。

[9] 截止 2022 年 8 月由英特尔开展的测试。测试配置 2：单节点，双路英特尔® 至强® 铂金 8480+ 处理器，56 核，启用超线程，启用睿频加速技术，256 GB 总内存（16 插槽/32 GB/4800 MHz），EGSDCRB1.SYS.0085.D15.2207241333，0x2b000070，Ubuntu 22.04.1 LTS，5.19.0-051900-generic，gcc 11.2，Yonyou OCR v1，OpenVINO 2022.2.0-custom_onednn2.6_9a3a3181e7056dcf7ccd3a16e599e6882a4edc23，Yonyou OCR v1，OneDNN 2.6。测试配置 3：单节点，双路英特尔® 至强® 铂金 8480+ 处理器，56 核，启用超线程，启用睿频加速技术，256 GB 总内存（16 插槽/32 GB/4800 MHz），EGSDCRB1.SYS.0085.D15.2207241333，0x2b000070，Ubuntu 22.04.1 LTS，5.19.0-051900-generic，gcc 11.2，Yonyou OCR v1，OpenVINO 2022.2.0-custom_onednn2.6_9a3a3181e7056dcf7ccd3a16e599e6882a4edc23，Yonyou OCR v1，OneDNN 2.6。实际性能受使用情况、配置和其他因素的差异影响。更多信息请见 www.Intel.com/PerformanceIndex