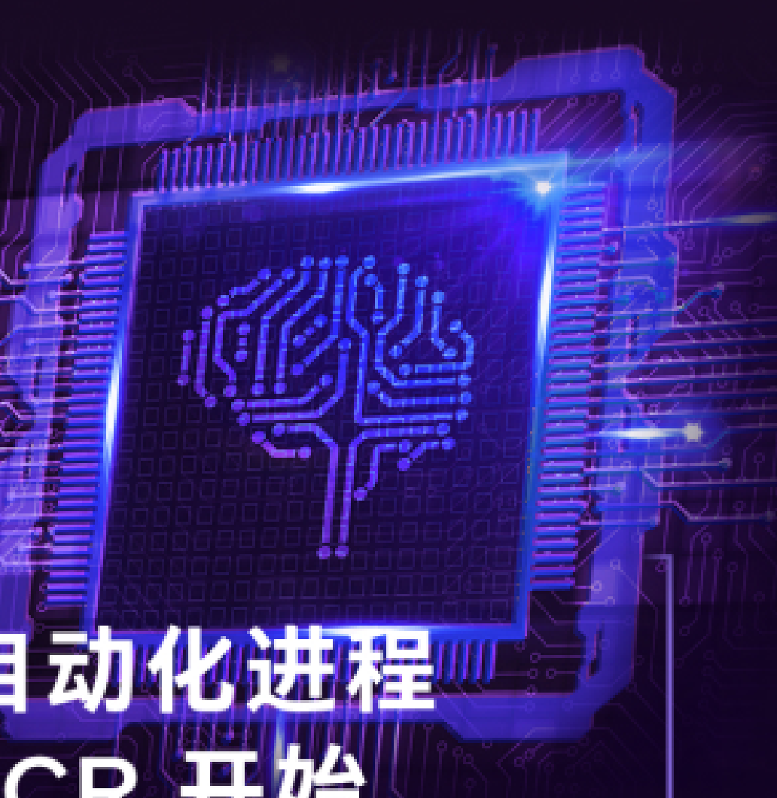
intel
XEON



用 AI 激活业务自动化进程
从至强加速的 OCR 开始

第四代英特尔® 至强® 可扩展处理器

实战智能 OCR，三大挑战要破解

加速人工智能（AI）技术的落地，加速企业业务的自动化进程，以及推动业务流程的数智化重塑或升级，现成为企业提升工作效率、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。

如今智能 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，从而帮助企业释放人力资源，提升工作效率，为广泛的数智化应用提供基础能力支撑。

基于深度学习的 OCR 应用在算法、算力等层面面临严峻挑战：

庞大性能开销

在智能质检、智能化采集等典型应用场景中，OCR 应用常需以模型推理的方式，对海量单据、文档等图片进行处理，产生庞大的 AI 算法开销。

沉重 TCO 压力

投资回报率是用户在部署 OCR 应用时重要的考虑因素。独立 GPU 等专用加速器虽能满足 OCR 应用的算力需求，但通常有较高的硬件部署和开发成本，大量用户倾向于充分利用 CPU 资源，选择基于 CPU 的推理方案。

异构化扩展难以实现

鉴于客户的部署情况千差万别，AI 算法需既能在异构化平台上进行移植，又能保证满足性能要求，这常导致企业将大量资金耗费在应用开发、性能优化等工作中。

至强 CPU 内置的 AI 加速“神器”：英特尔® 高级矩阵扩展（Intel® AMX）

第四代英特尔® 至强® 可扩展处理器帮助企业在 AI 性能上更进一步。其内置的英特尔® AMX 加速引擎可针对广泛的硬件和软件优化，能将深度学习训练和推理性能提升**高达 10 倍**，非常适合自然语言处理、推荐系统和图像识别等工作负载。

英特尔® AMX 采用全新的指令集与电路设计，在实际工作负载中，能同时支持 BF16 和 INT8 数据类型，其每个物理核在每个时钟周期可实现 **2,048 次** INT8 运算和 **1,024 次** BF16 运算，AI 工作负载的运行效率大幅提升。

同时这款处理器也可通过与 OpenVINO™ 工具套件的结合，进一步提升 AI 推理性能，可在包括英特尔 CPU、iGPU 和 FPGA 在内的英特尔硬件平台（包括加速器）上部署并加速神经网络模型，在保持精度的同时提高推理速度。

多种场景火炼真金，AMX 加速无往不利

媒体娱乐

助个性化内容推荐速度提升高达 **6.3** 倍[1]，带来更顺畅的用户体验

零售

助视频分析速度提升高达 **2.3** 倍[2]，打造更好的消费者体验，更快触达目标客户

制造

可用于缺陷检测等工厂自动化应用的机器视觉解决方案，让工厂更智能、更高效

医疗

可支持工作负载整合以及快速的 AI 推理和模型训练，让医疗服务更高效

• • •

轻松 hold 住 OCR 多场景应用
助更多企业拥抱自动化进程

用友商业创新平台（BIP）：
实现高速 AI 推理，提升 OCR 应用投资回报率

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、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，服务企业与产业商业创新的平台型、生态化云服务群。

该平台借助第四代英特尔® 至强® 可扩展处理器及内置的 AMX 技术，不仅避免了独立 GPU 所带来的高昂支出，而且充分利用现有的 CPU 资源，实现了更高的灵活性和敏捷性。

- OCR 模型的推理性能提升**3.42 倍**[3]
- 将模型从 FP32 量化到 INT8/BF16，性能提升 **7.3 倍**[4]
- 降低如专用 GPU 等硬件的采购成本，显著降低空间、功耗、软硬件调优等成本，同时提升基础设施灵活性

> “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

点击下方链接查阅 AMX 实战方案

基于第四代英特尔® 至强® 可扩展处理器的用友智能 OCR 服务

亚信科技 AIRPA 平台：
提升 OCR 效能实现数据共享与流程自动办理

亚信科技机器人流程自动化平台（AISWare AIRPA）是通过模拟并增强人类与计算机的交互过程，实现工作流程自动化的平台。其在 RPA 功能的基础上，提供了强大的智能 OCR 能力支持。

在智能 OCR 应用中，亚信科技通过英特尔® AMX 支持实现从 FP32 到 INT8/BF16 的量化，在可接受的精度损失下，增加吞吐量并加速推理，从而推动降本增效、助力人力解放，打破业务烟囱，联通数据孤岛，最终加速企业数智化转型。

- OCR 算法推理性能提升 **3.38 倍**[5]
- 面向重复、高耗时的业务，人工成本降至原来的 1/5 到 1/9，效率提升约 **5-10 倍**[6]

> “流程自动化已经成为企业数字化转型战略的关键部分，亚信科技智能 RPA 平台致力于数字员工的快速构建，提供简单、高效、灵活、智能的机器人流程自动化解决方案。通过采用最新的第四代英特尔® 至强® 可扩展处理器，我们进一步提升了 RPA 平台中智能 OCR 等应用的推理性能表现，帮助用户实现从流程自动化向流程智能化的转变。”
>
> **赵一鸣**
> 亚信科技平台产品研发中心总经理

点击下方链接查阅 AMX 实战方案

第四代英特尔® 至强® 可扩展处理器助力亚信科技提升 OCR 应用性能

东软医保 OCR 票据识别解决方案：
加快业务流程，实现高性能、高性价比的 AI 推理

东软医保 OCR 票据识别方案可通过纸质单据电子化、OCR 文字识别等流程，形成符合业务系统报销要求的医保电子结构化数据，优化医保经办工作流程，保障医保基金安全。

基于第四代英特尔® 至强® 可扩展处理器和内置的 AMX 加速器，同时采用了 OpenVINO™ 工具套件作为 AI 框架，东软医保 OCR 票据识别方案帮助用户在显著节约 IT 基础设施投入的前提下，充分挖掘硬件潜力，显著提升了 OCR 的识别性能。

- 数据精度为 FP32 时，OCR 模型推理性能提升 **4.66 倍**[7]
- 数据精度为 INT8 时，OCR 模型推理性能提升 **2.29 倍**[7]
- OCR 识别准确度可达 **95%** 以上，缩短业务办理周期

> “医疗票据录入是医保机构在办理费用报销业务时经常面临的场景之一，通过应用智能 OCR 产品，东软能够将手动录入流程转化为自动流程，从而提高医保报销效率，实现医保业务的智能化精细化管理。第四代英特尔® 至强® 可扩展处理器的应用让我们能够显著提升智能 OCR 的推理效率，帮助客户打造更加现代化的智能医保平台，助力客户践行服务型政府和数字化政府的建设目标。”
>
> **刘 兵**
> 东软集团医疗保障事业部总经理

点击下方链接查阅 AMX 实战方案

第四代英特尔® 至强® 可扩展处理器助力东软医保 OCR 加速 AI 推理

即刻添加英特尔商用小助手
免费获得技术干货
更有不定期福利拿到手软

扫码进入
@英特尔商用会员中心 小程序
免费获取干货信息与系列课程

---

[1] 与上一代产品（FP32）相比，启用英特尔® AMX（BF16）时可将批量推荐系统推理性能（针对 DLRM 模型）提升高达 6.3 倍。详情请见以下网址的 [A21]：https://edc.intel.com/content/www/cn/zh/products/performance/benchmarks/processors/（第四代英特尔® 至强® 可扩展处理器）。结果可能不同。

[2] 与上一代产品（FP32）相比，使用内置英特尔® AMX（BF16）的第四代英特尔® 至强® 可扩展处理器在处理端到端视频流媒体时，帧率提升高达 2.3 倍。

[3] 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_9a3a3181e7056dcf7ccd3a16 e599e6882a4edc23，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_9a3a318 1e7056dcf7ccd3a16e599e6882a4edc23，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

[4] 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，YonyouOCR 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

[5] 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，AsiaInfo OCR v1，OpenVINO 2022.2.0-custom_onednn2.6_9a3a3181e7 056dcf7ccd3a16e599e6882a4edc23，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，AsiaInfo 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

[6] 数据援引自亚信科技内部测试结果。英特尔并不控制或审计第三方数据。请您审查该内容，咨询其他来源，并确认提及数据是否准确。

[7] 截止 2022 年 8 月东软联合英特尔开展的测试。测试配置：基准配置/新配置 3 — 单节点，双路英特尔® 至强® 铂金 8380 处理器，40 核，开启超线程，开启睿频加速技术，256 GB 总内存（16 插槽/16 GB/3200 MHz），<SE5C620.86B.01.01.0005.2202160810>，<0xd000375>，<Ubuntu 22.04.1 LTS>，<5.19.0-051900-generic>，<gcc 11.2>，<Neusoft OCR>，<OpenVINO 2022.2.0-custom_onednn2.6_9a3a3181e7056dcf7ccd3a16e599e6882a4edc23>，<Neusoft OCR>，<OneDNN 2.6>；新配置 1/2 — 单节点，双路英特尔® 至强® 铂金 8480+ 处理器，56 核，开启超线程，开启睿频加速技术，256 GB 总内存（16 插槽/16 GB/4800 MHz），<EGSDCRB1.SYS.0085.D15.2207241333>，<0x2b000070>，<Ubuntu 22.04.1 LTS>，<5.19.0-051900-generic>，<gcc 11.2>，<Neusoft OCR>，<OpenVINO 2022.2.0-custom_onednn2.6_9a3a3181e7056dcf7ccd3a16e599e6882a4edc23>，<Neusoft OCR>，<OneDNN 2.6>。