解决方案简介

基于第四代英特尔® 至强® 可扩展处理器的海鑫智圣云边端一体化方案加速 AI 推理

“人工智能是智慧金融系统构建的核心技术支撑之一，通过采用第四代英特尔® 至强® 可扩展处理器，我们得以在保证推理精度的前提下，大幅提升生物特征识别等 AI 模型的推理速度。我们还将与英特尔进行持续合作，以云边端一体化的金融 AI 应用，推动智慧金融系统的进化，助力金融业务价值的升级。”

— 孟凡军
海鑫智圣总经理

挑战

面向智慧金融场景的人工智能应用需要对于海量的结构化、非结构化数据（如图像数据、音频数据、文字数据等）进行处理，在性能、时延、成本等方面面临着巨大的挑战：

- **如何缩短数据处理路径、降低时延：**传统方案在处理 AI 推理等方案时，往往涉及到较长的网络路径，而且可能会涉及反复的内存拷贝等问题，数据处理的时延较高，难以支撑即时性较高的智慧金融场景。
- **如何提供更高的算力，以支持数据处理、AI 推理等性能要求：**金融机构 AI 业务规模庞大且对效率、精度要求较高，因此对算力有着较高的要求，单纯依靠后端服务平台进行数据结构化分析处理会导致数据处理效率和精准度都出现严重问题。
- **如何降低云、网络等关键资源的成本，提升经济性：**如果金融机构将业务所产生的海量非结构化数据接入云端进行处理，将会带来较高的云平台和网络资源成本，延缓智慧金融的转型速度。同时，AI 推理基础设施通常会带来较高的算力成本支出，通过采用 CPU 进行 AI 推理，将有助于降低 AI 加速器成本，并敏捷扩展到多种应用场景。
- **如何基于深度学习构建卓越的算法模型：**基于深度学习构建的算法模型面临不同训练框架训练出的模型彼此不能互相调用、推理性能优化难以为继、算法在云边端的推理精度不一致等问题。

解决方案概述

得益于人工智能（AI）、大数据等技术的应用，金融机构正在日趋智能化，不仅能够提升业务处理效率和金融安全性，同时为客户提供更加人性化的金融服务，让客户体会到金融服务的“温度”。例如，通过 AI 系统对于客户的业务数据进行智能化分析，金融机构将能够清晰地洞察客户的偏好，让客户服务过程更加精准。但同时，海量数据的分析与 AI 应用也将给基础设施带来巨大的压力。

北京海鑫智圣技术有限公司（以下简称：海鑫智圣）推出了集成“海鑫智圣边缘计算终端 + 云端 Cell 推理引擎”的云边端一体化解决方案，并搭载了全新的第四代英特尔® 至强® 可扩展处理器。借助内置的英特尔® 高级矩阵扩展（英特尔® AMX）加速引擎，第四代英特尔® 至强® 可扩展处理器能够帮助海鑫智圣进一步优化端到端推理性能，提升 AI 平台的整体算力和可扩展性，满足智慧金融在生物特征识别、智能 OCR 识别等领域的性能要求，加速智慧金融变革。

海鑫智圣云边端一体化平台加速 AI 推理

“海鑫智圣边缘计算终端 + 云端 Cell 推理引擎” 的云边端一体化平台能够支持云边推理的异构化数据分析流程，可由边缘计算终端完成对各类原始数据的目标检测、目标特征标定和特征提取，将非结构化的原始数据转化为包含各类目标特征信息的数据，再传输至后端推理引擎进行推理运算和融合。

在云端，该方案将多种学习框架训练的模型封装到跨平台算法引擎中。同时，方案在云端运行基于算法模型开发的中心平台系统，与边端子系统协同工作，完成如目标特征提取（用于建立目标特征库，包括目标人像特征库、目标车辆库等）和数据关联性分析等任务。

在边缘端，该方案可以运行基于算法模型开发的软件或系统，既可以独立工作，也可以作为智能化系统的前端子模块，与中心系统协调工作，负责诸如数据结构化，特征比对等工作，以满足用户在数据时延等方面的更高需求。在云端与边缘端，该方案都可以通过基于第四代英特尔® 至强® 可扩展处理器的服务器，提供强大的 AI 算力。

海鑫智圣云端 Cell 推理引擎为该方案奠定了技术基础，海鑫智圣 Cell 推理引擎拥有自定义网络结构及参数定义文件格式，全面支持 OpenVINO™ 工具套件等加速库。针对没有加速库的情况，Cell 推理引擎还使用 C++ 开发了一套完全可控的推理程序 (Raw)，作为各个平台实现一致性的基准。此外，Cell 推理引擎还拥有自有深度算法算子，并使用工具与人工组合的方式转换训练出的深度模型。

通过将 OpenVINO™ 工具套件作为高效的加速库，Cell 推理引擎支持用户进行推理性能加速。OpenVINO™ 工具套件支持从边缘到云的深度学习推理，可支持开发人员使用行业标准人工智能框架、标准或自定义层，将深度学习推理轻松集成到应用中，在英特尔® 硬件（包括加速器）中扩展工作负载并改善性能。

Cell 推理引擎具备如下优势：

- 各个平台使用相同的模型文件，管理维护简单；
- 得益于 OpenVINO™ 工具套件的加持，推理引擎有着卓越的性能，而且接口一致，对用户透明；
- Cell 推理引擎集成了多种深度学习算法框架，同时支持自定义网络结构及参数定义文件格式，从而使得通过 Cell 推理引擎训练得出的算法模型，能够兼容云端及边端设备，保证推理性能的一致性；
- 运行结果高度一致，只需对 Raw 做精度测试即可推广到所有平台。

由于实现了对非结构化数据的前置化处理、分析和响应，该方案提升了 AI 平台系统整体运算能力，能够支持 AI 平台快速、高效地应对海量数据分析场景。

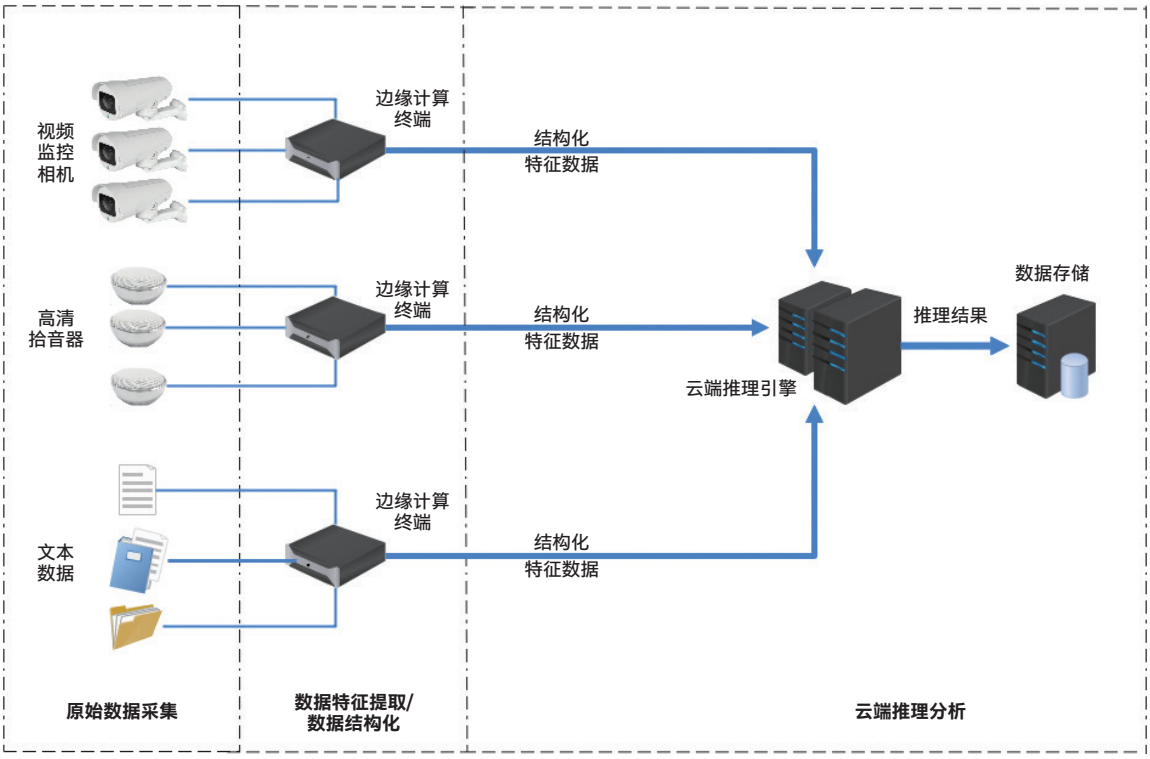

图 1. 海鑫智圣一体化解决方案数据处理流程

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，有效提升了内存带宽与速度，并通过 PCIe 5.0（80 个通道）实现了更高的 PCIe 带宽提升。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。

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上提供显著的性能提升。

第四代英特尔® 至强® 可扩展处理器与 OpenVINO™ 工具套件的结合可以进一步提升 AI 推理性能。OpenVINO™ 工具套件支持从边缘到云的深度学习推理，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，能够在保持精度的同时提高视频推理速度。OpenVINO™ 工具套件支持开发人员使用行业标准人工智能框架、标准或自定义层，将深度学习推理轻松集成到应用中。

第四代英特尔® 至强® 可扩展处理器提升 AI 模型推理性能

为了加速 AI 推理、数据处理等负载的运行速度，海鑫智圣云边端一体化平台采用了第四代英特尔® 至强® 可扩展处理器。

在使用第四代英特尔® 至强® 可扩展处理器进行推理并通过 OpenVINO™ 工具套件进行性能优化的场景中，海鑫智圣一体化解决方案能够支持金融机构获得强大的推理性能，同时增强基础设施调度的灵活性，降低 TCO。为了验证第四代英特尔® 至强® 可扩展处理器带来的推理性能提升，海鑫智圣与英特尔在生物特征识别应用中，进行了联合测试。测试系统配置如表 1 所示。

表 1. 测试系统配置

	配置 1	配置 2
处理器	2*英特尔® 至强® 铂金 8380 处理器 @ 2.30 GHz，40 核	2*英特尔® 至强® 铂金 8480+ 处理器 @ 2.30 GHz，56 核
内存	256 GB（16 插槽/16 GB/3200 MHz）	256 GB（16 插槽/16 GB/4800 MHz）
操作系统	Ubuntu 22.04.1 LTS	Ubuntu 22.04.1 LTS

首先，双方对比了在模型数据类型同为 FP32 时，生物特征识别模型在第三代/第四代英特尔® 至强® 可扩展处理器上的推理性能差异。测试结果如图 3 所示，第四代英特尔® 至强® 可扩展处理器获得了 1.58 倍的代际性能提升[1]。

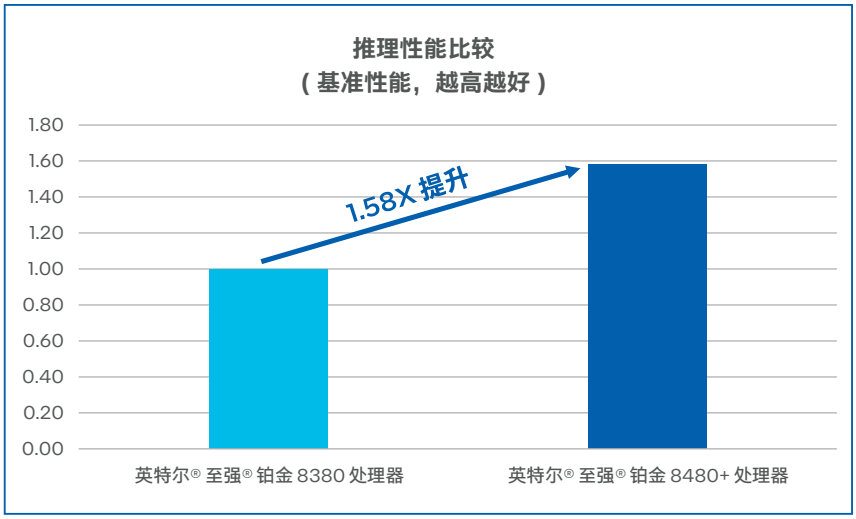

图 3. 生物特征识别模型在第三代/第四代英特尔® 至强® 可扩展处理器上的推理性能对比[2]

[1,2] 截止 2022 年 8 月由海鑫智圣联合英特尔开展的测试。测试配置：基准配置 — 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>，<Hisign CELL>，<OpenVINO 2022.2.0-custom_onednn2.6_9a3a3181e7056dcf7ccd3a16e599e6882a4edc23>，<Hisign CELL>，<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>，<Hisign CELL>，<OpenVINO 2022.2.0-custom_onednn2.6_9a3a3181e7056dcf7ccd3a16e599e6882a4edc23>，<Hisign CELL>，<OneDNN 2.6>。

随后，双方在基于第四代英特尔® 至强® 可扩展处理器的平台上，采用英特尔® AMX 将模型从 FP32 转化为 BF16，并测试了两者的性能差异。测试数据如图 4 显示，优化后可以实现 4.55 倍的性能提升[3]。

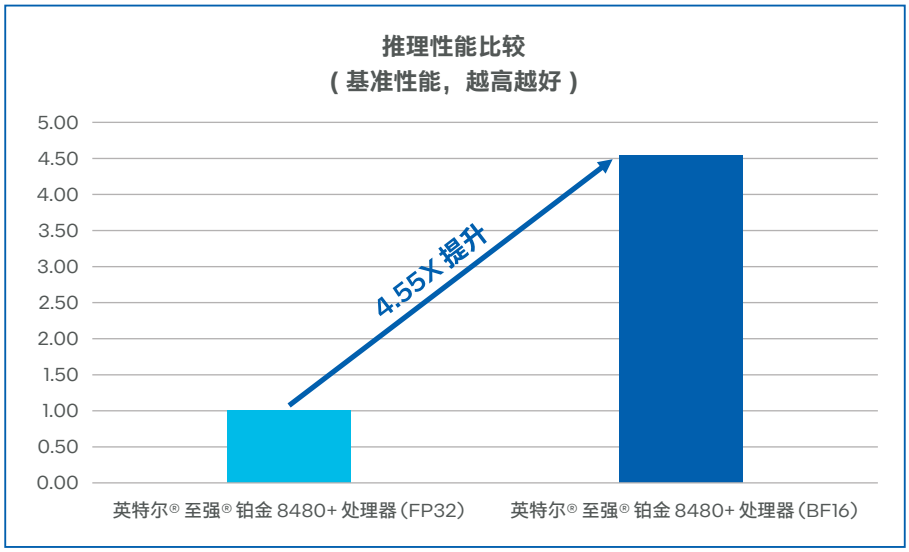

图 4. 不同数据类型在第四代英特尔® 至强® 可扩展处理器上的推理性能比较[4]

收益

显著的推理性能提升

相较于第三代英特尔® 至强® 可扩展处理器，内置英特尔® AMX 加速引擎的第四代英特尔® 至强® 可扩展处理器能够大幅提升推理性能，加快 AI 应用响应速度，降低性能压力。

较低的精度损失

测试数据显示，通过将数据类型从 FP32 转化为 BF16，英特尔® AMX 在大幅提升性能的同时，精度损失仅为 0.08%[5]，完全可以满足实际应用所需。

实现算力在云边端的合理布局

通过算力在云边端的分配，降低了云端的数据处理以及 AI 推理压力，缩短了处理时间，有效提升了数据处理效率。

快速适配各种应用场景

解决深度学习框架不统一、生产环境不稳定等问题，提升人工智能平台整体算力和可扩展性能，满足更多场景所需。

支持边缘端高效 AI 推理

即使在轻量化的边缘设备侧，也具备较高的 AI 推理性能，提升 AI 应用在边缘端的运行效率，降低数据处理延迟。

目前，该方案已经在金融机构的多个场景得到应用。例如在银行开户验证/支付等场景中，银行能够通过生物特征识别技术验证是否实现了“人证合一”，有效解决在线无人值守环境中的各类假体攻击，构建安全、高效的金融服务体系，减少流程时间以提高转化率。

[3,4,5] 截止 2022 年 8 月由海鑫智圣联合英特尔开展的测试。测试配置：基准配置 — 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>，<Hisign CELL>，<OpenVINO 2022.2.0-custom_onednn2.6_9a3a3181e7056dcf7ccd3a16e599e6882a4edc23>，<Hisign CELL>，<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>，<Hisign CELL>，<OpenVINO 2022.2.0-custom_onednn2.6_9a3a3181e7056dcf7ccd3a16e599e6882a4edc23>，<Hisign CELL>，<OneDNN 2.6>。

展望

在第四代英特尔® 至强® 可扩展处理器的助力下，海鑫智圣云边端一体化平台能够在保证推理精度的同时，显著提升推理性能，帮助金融机构提升业务处理效率和金融安全性，为客户提供更加个性化的金融服务。同时，得益于云边端的算力合理分配，该平台可以将大量基础原始数据的预处理和数据结构化工作交给边缘计算终端处理，有效减轻基础设施的工作压力。

依托于该平台，金融机构能够高效、敏捷扩展目标检测、目标检索、特征分析等 AI 技术，赋能智慧认证、智慧支付、智慧营销、智慧接待等各种应用场景，帮助金融机构等行业用户构建高性能、高扩展、高稳定性的 AI 基础设施，并加速智慧能力在不同行业的深度落地，为数字化创新赋能。