# 案例研究

第二代英特尔® 至强® 可扩展处理器
英特尔® 傲腾™ 持久内存
英特尔® FPGA
SageOne Advanced 企业级 AI 集成系统

# 软硬优化一体 助企业跨过 AI 落地门槛

第四范式 The Fourth Paradigm

虽然日渐成熟的人工智能（Artificial Intelligence, AI）正成为更多企业实施数字化、智能化转型的核心引擎，但不是所有企业和机构都有充足的技术积淀和专业人才来实现 AI 能力与行业需求的快速融合，他们若要跨过 AI 应用落地的门槛，普遍需要两个层面的帮助和支持：构建坚实的基础设施能力和 AI 技术储备，以及行之有效的 AI 方法论和支持。

为更好地推进 AI 与产业的融合，解决 AI 落地中存在的性能、部署及总拥有成本（Total Cost of Ownership, TCO）等问题，来自 AI 领域的“独角兽”——第四范式，创新地提出了企业智能化转型的“1+N”模式，既要致力于实现更优的 AI 应用效果，助推企业核心业务发展，也要寻求通过 AI 规模化落地，来实现 AI 价值的最大化。

基于这一理念，第四范式与英特尔公司等合作伙伴展开了紧密的创新协作。通过引入第二代英特尔® 至强® 可扩展处理器、英特尔® 傲腾™ 持久内存以及基于英特尔® FGPA 的自研计算加速卡等高性能产品，第四范式打造了新一代 SageOne Advanced 企业级 AI 集成系统产品，以软、硬一体优化的方式，为企业提供开箱即用的一站式 AI 能力，不仅可最大程度地发挥先知平台提供的 AI 应用能力，也能够激发出高性能硬件蕴含的磅礴潜力，为企业的全方位 AI 转型提供可靠的基石。

“我们以‘1’来结合企业核心业务，通过更优的 AI 效能全力助推其发展；用‘N’让更多的 AI 规模化落地，使应用场景的总体价值最大化。基于此，我们通过采用第二代英特尔® 至强® 可扩展处理器等先进产品，为企业提供开箱即用的 SageOne Advanced 企业级 AI 集成系统产品，为他们实施全方位 AI 转型提供了更强劲也更为可靠的助力。”

**郑婴**
**软硬一体产品首席架构师**
**副总裁**
**第四范式**

**第四范式实现的解决方案优势：**

- **更强性能表现：** 英特尔® 架构先进硬件设备的助力，让 SageOne Advanced 企业级 AI 集成系统具备了强劲性能，与上一代 Sage on x86 服务器相比，性能提升达 4-10 倍[1]。
- **更短项目周期：** 整机交付的方式，让用户获得从服务器采购到网络设备安装、从软件部署到功能测试的全流程交付服务，大大缩短交付时间，获得开箱即用的 AI 能力。
- **更大调优空间：** 通过硬件升级，采用英特尔针对高性能计算和 AI 应用优化的至强® 铂金® 9200 系列处理器，引入 NUMA (Non-Uniform Memory Access Architecture，非一致性内存架构) 绑定，并针对应用并行度和内存利用率进行优化，SageOne Advanced企业级 AI 集成系统可进一步实现多达 4.76 倍的性能提升[2]。

以 AI 促进业务发展，提升业务创新和经营效率，已开始被越来越多的企业管理者奉为数字化时代的致胜法宝。在第四范式看来，优秀的 AI 方案既能为企业提供出色的预测、推荐效果，也能够让全面智能化给企业带来业务价值的聚变。为此，第四范式创新地提出了企业智能化转型“1+N”模式。

以某零售企业为例，利用 AI 技术提供个性化推荐能力，改善供应链效率是“1”，通过 AI 技术实现多场景下的语音识别、单据 OCR、智能客服等则是“N”。通过“1+N”式的协作，企业可以更高效地实现整体经营效率的优化。

## 用技术实现企业 AI 的“1+N”

第四范式“1+N”理念中的 1，更具体来说，就是以更优的 AI 效果，有效推动企业的一个或多个核心业务，以点带面，拉动整个企业的业务发展。如图一所示，这种更优的效果，源于第四范式 SageOne Advanced 企业级 AI 集成系统所具备的高维、实时和闭环三大优势。

传统上，机器学习专家模型的维度在数个至数千个不等，即便是一般意义上的高维模型，往往也局限在一万以下。通过 SageOne Advanced 企业级 AI 集成系统内置先知平台中的高维度模型算法和特征工程算法，第四范式对业务数据实施切片，巧妙地将数据“升维”，并借助基于英特尔® 架构高性能硬件产品提供的强大算力，使方案获得远超传统模型的精准性。


图一 基于“1+N” 理念的第四范式 AI 解决方案

由事后分析转为实时决策，是 AI 应用产生业务效果的关键。为保障核心业务的实时决策能力，第四范式将自主研发的高性能内存时序数据库 RTIDB 与 SageOne Advanced 企业级 AI 集成系统相融合，凭借毫秒级海量特征数据供给、InfiniCache 无限缓存、特征存储介质自动分级等技术，使万亿级维度模型在百万级吞吐量下，依然可实现毫秒级响应的精准决策。相比主流的开源方案，第四范式 AI 引擎内嵌自主研发的分布式参数服务器，设计并优化张量分片、NUMA 感知本地通信、流水线、RDMA 网络优化等特性，在稠密/稀疏场景下支持同步和异步两种训练模式，具备线上高可用、多副本负载均衡等企业级特性，为线上预估与线下训练构建了高效的桥梁，可做到模型的秒级更新上线。

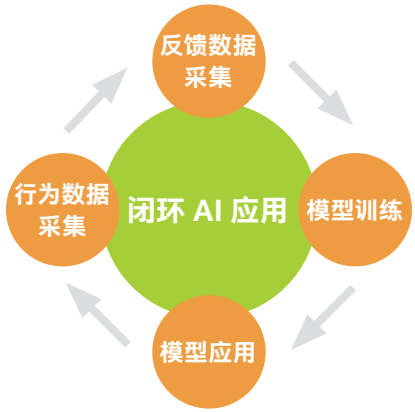
基于创新 3D Xpoint™ 存储介质的英特尔® 傲腾™ 持久内存，为 SageOne Advanced 企业级 AI 集成系统提供了其实时优势所需的可靠性能支撑。其非易失性特性，以及接近于 DRAM 内存的“读”性能，让 SageOne Advanced 企业级 AI 集成系统的数据读取性能提升高达数倍。

来自第四范式的一项测评数据显示，在某企业核心场景的海量实时业务决策中，SageOne Advanced 企业级 AI 集成系统的 15 个推理引擎节点，能够支撑 20 亿维特征模型进行 2 万次实时并发请求，平均响应延迟在 5ms 以内，99% 请求在 20ms 内响应，大幅提升了实时决策在核心业务场景中的性能表现和关键价值[3]。

而 SageOne Advanced 企业级 AI 集成系统具备的 AutoML、企业大数据平台兼容性、低门槛 AI 平台等核心能力，也能够帮助企业一站式地构建从数据引入、模型训练到模型上线的闭环流程。同时，还可实现自我学习、快速迭代，使 AI 应用能够紧扣企业业务的飞速变化，真正做到 AI 与业务无缝和充分的融合。

第四范式“1+N”理念中的 N，则是针对广泛的应用场景实现 AI 的规模化落地，进而提升企业整体效率。对此，第四范式给出了统一的方法论、面向 AI 的数据治理以及充沛的 AI 算力这三种支持。

面对数以千计的 AI 应用场景，分别构建不同的 AI 方法无疑会事倍功半。第四范式通过建立统一的方法论来规模化“生产” AI。如图二所示，SageOne Advanced 企业级 AI 集成系统内置的先知平台遵循“库伯学习圈”理论，将 AI 开发分为“行为数据采集、反馈数据采集、模型训练、模型应用”四个步骤，并引导用户顺序推进 AI 的开发。




图二 基于“库伯学习圈”理论的闭环 AI 应用

同时，借助与现有企业级大数据平台之间良好的兼容性，配合先进的英特尔® 固态盘产品带来的高性能存储能力，SageOne Advanced 企业级 AI 集成系统也能够帮助企业获得完善的数据治理能力。它能够支持多源异构数据实时引入、离线/在线数据一致性管理、回流数据自动标注等功能，并支持存取 PB 级甚至更大量的日志，具备支持实时存储，并形成线上数据采集和处理的闭环能力。

最后，对于 AI 规模化落地所需的强大算力，第四范式亦将其视为完整的体系架构进行构建，不仅以高性能硬件产品作为基石，更结合 AI 算法的运算架构与逻辑，针对软、硬件进行深层次优化，从而确保充沛 AI 算力的供应。

## 至强® 可扩展平台：强劲 AI 算力之源

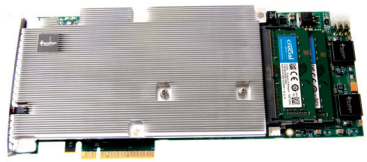
企业实施全方位 AI 转型过程中，往往会面对既有 IT 基础设施改造困难等问题。为此，第四范式在全新一代的 SageOne Advanced 企业级 AI 集成系统中，以高性能处理器、100Gb 以太网、自研计算加速卡以及高性能存储设备为核心，融合先知机器学习平台与主流 AI 计算架构，助力企业解决这些难题。

首先，针对高维特征计算所需的高频、多核算力，SageOne Advanced 企业级 AI 集成系统引入了第二代英特尔® 至强® 可扩展处理器作为其核心处理引擎，其可集成 28 核心/56 线程，并可使用英特尔® 睿频加速技术 2.0 将频率调整至最高 4.4GHz。该处理器具有更为优化的内核微架构、核内互联和内存控制器技术以及独特的频率锁定技术，搭配前文提及的英特尔® 傲腾™ 持久内存，使 SageOne Advanced 企业级 AI 集成系统在建模、仿真、机器学习和高性能计算等一系列工作负载中都能尽显其能。

同时，为了追求更优的性能表现，SageOne Advanced 企业级 AI 集成系统也可升级使用英特尔® 至强® 铂金 9200 系列处理器，这款集成更多内核，针对高性能计算、AI 应用优化的处理器能够带来明显强化的基础算力支撑，在性能调优上也给予了用户更广阔的空间，以进一步释放其性能潜力。在实践中，用户可以采用 NUMA 绑定的方式，消除处理器对远程内存的访问，实现各节点对内存访问效率的优化；用户还可以进一步引入英特尔® VTune™ 可视化性能分析器 (英特尔® VTune™ Amplifier)，寻找应用代码中制约并行度和内存利用率的“热点”，进而有的放矢地进行优化，以充分利用处理器多核多线程的能力，扩展应用的并行度并缓解内存受限的状况。

其次，针对分布式机器学习在特征工程、模型训练等过程中涉及的大量任务和数据交互过程，SageOne Advanced 企业级 AI 集成系统创新地采用软件定义通信协议，来避免各服务器实例间的网络 I/O 成为性能瓶颈。它配备了 100Gb 以太网，并在软件中针对协议优化配置代码，使系统可通过以太网使用远程直接内存访问 (Remote Direct Memory Access，RDMA) 技术，并利用高速网络来减少大批量数据传输时处理器的开销，从而更好地破解 AI 开发过程中的 I/O 瓶颈问题。

此外，SageOne Advanced 企业级 AI 集成系统还导入了数据定义存储的理念，可根据企业构建 AI 应用过程中的数据特性来选择更适合的存储设备。在这一过程中，英特尔® 固态盘 DC P4600 以其高吞吐量、低延迟、高服务质量和高耐用性的特性，满足了该企业级 AI 集成系统在特征处理、数据落盘时的定制化调节需求，以及 Spark 等组件在先知平台上的升级和调优需求。



图三 第四范式计算加速卡-ATX800

最后，针对先知平台中耗时最多的任务，SageOne Advanced 企业级 AI 集成系统选择以硬件加速的方式来提升效率。第四范式基于英特尔® Arria® 10 FPGA 构建的计算加速卡 ATX800，可在无损数据压缩加速和 GBDT 训练加速两个关键过程中发挥作用，令高维特征计算过程中的 I/O 获得最高达 10 倍的加速[4]。

## 性能测试：彰显软硬一体实战优势

为验证 SageOne Advanced 企业级 AI 集成系统在性能和 TCO 等方面的表现，第四范式设计了多种测试方案。

其中一项非常重要的测试，是进行全面软硬件一体优化升级后的 SageOne Advanced 企业级 AI 集成系统与上一代 Sage on x86 服务器进行的性能对比测试。如表一所示，测试分别设计了 2 种应用场景和 2 种机器学习模式，通过交叉组合后得到 3 个测试用例[5]。

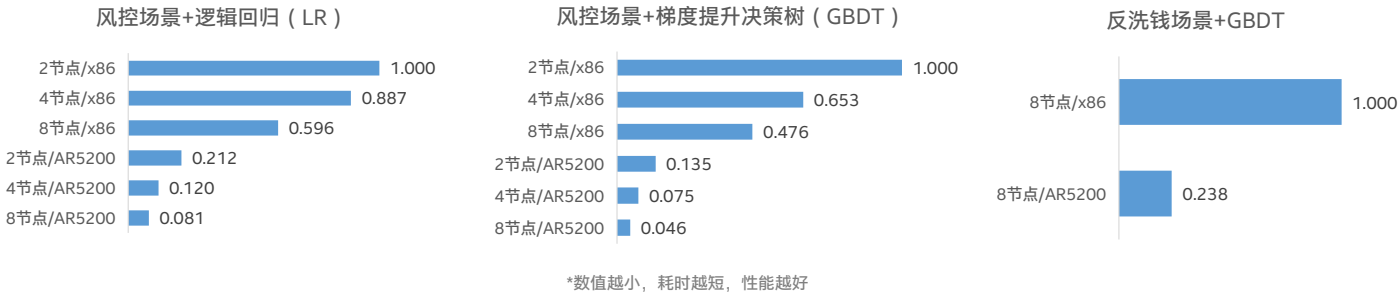
| | | 机器学习模型 | |
|---|---|---|---|
| | | 逻辑回归(LR) | 梯度提升决策树(GBDT) |
| 应用场景 | 风控场景 | 测试用例一 | 测试用例二 |
| | 反洗钱场景 | | 测试用例三 |

| 测试平台 | |
|---|---|
| SageOne Advanced (AR5200) 集群 | Sage on x86 集群 |
| 节点数: 8 | 节点数: 8 |
| 单机配置 | 单机配置 |
| 处理器: 英特尔® 至强® 金牌 6230 处理器*2<br>核心/线程: 40 核心/ 80 线程<br>主频: 2.8GHz<br>内存: 32GB*12<br>存储: 英特尔® 固态盘(2T NVMe)<br>网络适配器: 100GbE<br>加速卡: ATX800*1 | 处理器: 英特尔® 至强® 金牌 5218 处理器*2<br>核心/线程: 32 核心/64 线程<br>主频: 2.1GHz<br>内存: 32GB*12<br>网络适配器: 1GbE |

表一 测试用例设计

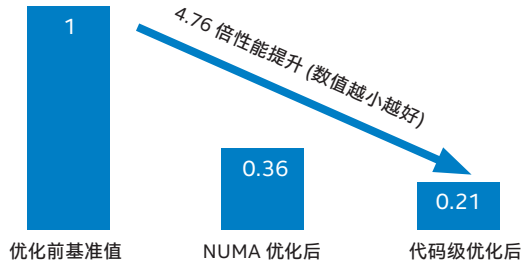
测试结果如图四所示，与 Sage on x86 服务器相比，SageOne Advanced 企业级 AI 集成系统在各个场景和机器学习模型中的性能，均获得了非常明显的提升，提升幅度达到 4-10 倍左右[6]。

同时，第四范式还对升级到英特尔® 至强® 铂金 9200 系列处理器的 SageOne Advanced 企业级 AI 集成系统进行了性能验证。测试结果如图五所示，在反洗钱场景中，系统经过充分的代码级调




*数值越小，耗时越短，性能越好

**图四** SageOne Advanced 企业级 AI 集成系统与 Sage on x86 服务器性能对比




**图五** SageOne Advanced 企业级 AI 集成系统采用至强® 铂金 9200 系列处理器，以及多种调优方案之后的归一化性能表现

优 (NUMA 优化+并行度和内存利用率的优化) 后，可为用户带来高达 4.76 倍的性能增益[7]。

今天，第四范式已通过一体化的“1+N”企业 AI 解决方案，为金融、医疗、零售、能源等领域中数以千计的客户提供服务，并获得了良好的应用效果和客户反馈，被证明其能有效地帮助用户从 AI 中获益。在医疗领域，以慢性病预测为例，其预测三年后糖尿病患病概率的准确率比专业医生高 2-3 倍，并已在 30 多个省市的 400 多家医院投入使用[8]。

2020 年初新型冠状病毒疫情出现后，第四范式也将 AI 技术投入到科技战“疫”一线，与南京大学周志华教授领导的 LAMDA 研究所 (机器学习与数据挖掘研究所，Learning And Mining from DatA) 团队，以及某地方医院的一线临床专家在第一时间成立了联合团队，研发出了基于 AI 的精准防控、疫情推演及病毒溯源方案，推进了以下关键举措:

- 基于第四范式提供的高维机器学习防控筛查模型，提升人群的覆盖面及筛查的召回率和准确率，并利用基于 AI 的自学习能力及临床专家的专业经验，构建了数据及系统闭环，进行持续迭代，保证了在有限的时间内基于动态环境变化 (包括病毒变异和人群动态复杂性)，为各有关部门提供高效、准确的判断支持，进而帮助制定科学有效的措施和行动计划;
- 利用高维机器学习技术及多维度的数据，构建更细粒度、更接近实际情况的省市区县级数字孪生系统，可充分考虑各种突发因素，如交通管制、复工时间和药物投放等突发因素对疫情的影响，实时通过模拟功能预演分析疫情的发展，进而为防控政策的制定和调整提供依据;
- 利用机器学习技术构建数据驱动的新型冠状病毒传播数字孪生系统，构建潜在的传染关系网，并结合病患信息，在关系网中找到可能的传播源与潜在超级传染者，并快速追溯传染路径，定位可能引起疫情传播的人群，帮助防疫部门尽快切断疫情蔓延的源头。

实战证明，联合团队推出的这套创新型方案比传统疫情传播模型表现更优，它让 AI 在疫情防控中发挥了不可替代的作用，为疫情的防控和治理做出了重要贡献。

## 展望

这些出色实践战绩的背后，包括第二代英特尔® 至强® 可扩展处理器在内的一系列英特尔先进产品与技术，无疑是第四范式“1+N”企业 AI 解决方案及 SageOne Advanced 企业级 AI 集成系统为企业提供高性能、易部署和低 TCO AI 能力的动力之源。以目前的成功合作为基础，第四范式和英特尔未来还计划在 AI 等领域继续深入开展协同创新，为实现 AI 技术更大规模的落地应用，帮助更多企业顺利实现智能化转型持续献“技”献力。

[1, 4, 5, 6] 数据援引自 https://www.4paradigm.com/images/pdf/sageone-product-performance-white-paper.pdf
[2, 7] 调优测试采用 4 节点、8 路配置的英特尔® 至强® 铂金 9242 处理器平台，48 核心 / 96 线程，开启超线程和睿频加速技术，采用 24 x 32 GB 2666 MHz DDR4 DRAM 内存，操作系统为 CentOS 7.6, Linux内核 3.10.0-957，测试工作负载为反洗钱场景，GCC 版本 4.8.5，glibc 库版本 2.17，Hadoop 版本 2.7.7。
[3] 数据援引自第四范式胡时伟的公开演讲稿《开启软件定义算力新时代》。
[8] 以上数据援引自第四范式提供的宣传折页。
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。