白皮书

英特尔® 至强® 可扩展处理器
英特尔® 高级矩阵扩展（英特尔® AMX）

基于第五代英特尔® 至强® 可扩展处理器的新一代腾讯云服务器加速乐元素游戏 AI 推理

“在游戏开发与运营中采用 AI 已经成为我们的一项关键步骤，但这也带来了显著的算力需求。基于第五代英特尔® 至强® 可扩展处理器的新一代腾讯云服务器在性能方面展现了明显的优势，特别是在游戏模型推理性能上的表现，让我们感到非常满意。我们计划未来在更多线上环境中部署和使用这款服务器，同时也期待能够与英特尔共同探索更多的技术创新，以便为各种使用场景带来更强的性能提升体验。”

— 钱晓东
乐元素开心消消乐制作人

“伴随着人工智能的快速发展，越来越多的玩家希望能体验到更创新的游戏体验。腾讯云依托腾讯内部多个人工智能实验室，将 AI 与云深度融合，通过公有云的方式开放给行业，让客户实现高效接入、灵活使用，推动业务的智能化再升级。”

— 许华彬
腾讯云副总裁

“AI 正在为游戏产业带来巨大的革新，并已经在众多流行游戏作品中得到成功应用。要想在这场技术浪潮中获得先机，游戏企业需要更加重视 AI 战略的执行，并在 AI 算力基础设施上投入更多的精力。英特尔为游戏企业提供了涵盖多种软硬件的 AI 全栈解决方案，能够帮助游戏企业解决 AI 应用的算力困扰，加速拥抱 AI 创新。”

— 陈葆立
英特尔数据中心与人工智能集团副总裁
中国区总经理

概述

针对关卡上线流程长、难度不易预测、玩家离线数据真假难辨、新玩法兼容旧关卡等问题，移动网络游戏研发及运营商：乐元素科技（北京）股份有限公司（以下简称：乐元素）创新地在关卡设计等流程中引入了人工智能（AI）技术，从而加快关卡设计质量与效率。但同时，AI 技术的应用也带来了 AI 算力挑战，如何构建高性能、低成本、高灵活性的 AI 算力平台成为乐元素需要考虑的重要问题。

为进一步给用户提供更流畅、优质的游戏体验，乐元素引入了基于第五代英特尔® 至强® 可扩展处理器的新一代腾讯云实例 S8，比上一代服务器的整体性能提升了 23%[1]。除此之外，乐元素与英特尔紧密合作，采用处理器内置的英特尔® 高级矩阵扩展（英特尔® AMX）等高级硬件能力，以及英特尔® oneAPI 和英特尔® 深度神经网络库（英特尔® oneDNN）等软件技术，提升 AI 推理等方面的性能表现，并在自研打关模型/ ResNet-50 等模型推理中得到成功验证。

挑战：AI 模型推理需要强大的算力作为支撑

近年来，乐元素在旗下热门游戏中，强化了 AI 技术的应用。以《开心消消乐》为例，该游戏是一款消除类休闲游戏，画面精美、上手简单、轻松有趣之余，又充满惊喜与挑战。游戏拥有 9 大关卡类型、60 余种障碍设计、8000 多个精心设计的关卡，关卡制作是这款游戏运营工作的主要内容之一。在《开心消消乐》中，用户每日都会进行游戏关卡挑战，而关卡的质量对于游戏的收入和用户留存起着至关重要的作用。

乐元素的游戏团队不断推出新关卡和玩法，并持续调整线上关卡的体验和难度，以提供持续新鲜的游戏体验。其中 AI 在关卡制作和优化中扮演了重要角色。对于新增和调整的关卡，AI 通过大量自动打关任务，确保关卡配置无错误，难度符合预期，并快速验证关卡。对于新开发的玩法，AI 也通过大量自动打关任务，确保逻辑无错误。每天平均运行超过 1 亿次打关任务，推理次数更是超过 30 亿次[2]。

但同时，由于用户群体不断增长，以及游戏内容持续更新，乐元素 AI 模型推理面临着性能、成本和灵活性等方面的挑战。

性能挑战

随着游戏用户数量的增加和游戏内容的扩充，服务器需要处理大量的游戏数据和用户请求。这意味着游戏服务器需要具备足够的算力来支持游戏的顺畅运行。要提升模型推理性能，一方面可以采用具备更高算力的硬件基础设施，另一方面也可以利用将模型转化为低精度格式、采用加速指令集等方式，以更好地释放算力。

成本挑战

游戏运营成本随着用户数量和游戏内容的增加而增加，特别是当部署专用的模型服务器时。乐元素希望在满足模型推理的性能需求时，能够尽可能地降低模型推理的单位成本，从而寻找更适合推理的算力选项。

灵活性挑战

游戏服务器需要具备足够的灵活性，以适应不断变化的游戏内容和用户需求。特别是在处理不同的模型推理需求时，需要具备灵活的基础设施和工作负载切换支持，以满足游戏运营的需求。

解决方案：基于第五代英特尔® 至强® 可扩展处理器的腾讯云实例 S8

新一代腾讯云实例 S8 基于全新优化虚拟化平台，提供了平衡、稳定的计算、内存和网络资源，是众多应用程序的卓越选择。其中，标准型实例采用第五代英特尔® 至强® 可扩展处理器，内存采用最新 DDR5，默认网络优化，最高内网收发能力达 4500 万 pps，最高内网带宽可支持 120Gbps[3]。

腾讯云实例 S8 搭载的第五代英特尔® 至强® 可扩展处理器凭借内置加速器实现单核性能提升，能够轻松应对要求严苛的工作负载。第五代英特尔® 至强® 可扩展处理器拥有更可靠的性能，更出色的能效。它在运行各种工作负载时均可实现显著的每瓦性能增益，在 AI、数据中心、网络和科学计算的性能和总体拥有成本（TCO）方面亦有更出色的表现。相较上一代产品，第五代英特尔® 至强® 可扩展处理器可在相同功耗范围内提供更高的算力和更快的内存。此外，它与上一代产品的软件和平台兼容，因此部署新系统时可大大减少测试和验证工作。

[1] 乐元素截止至 2024 年 2 月的内部测试结果，通过比较腾讯云 S8 与 S6 服务器得出。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。

[2] 数据援引自乐元素内部数据。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。

[3] https://cloud.tencent.com/document/product/213/11518#s8，2024 年 3 月访问。

第五代英特尔® 至强® 可扩展处理器每个内核都具备 AI 加速功能，无需添加独立加速器，就可处理要求严苛的 AI 工作负载，包括对参数量多达 200 亿的模型进行推理和调优[4]。

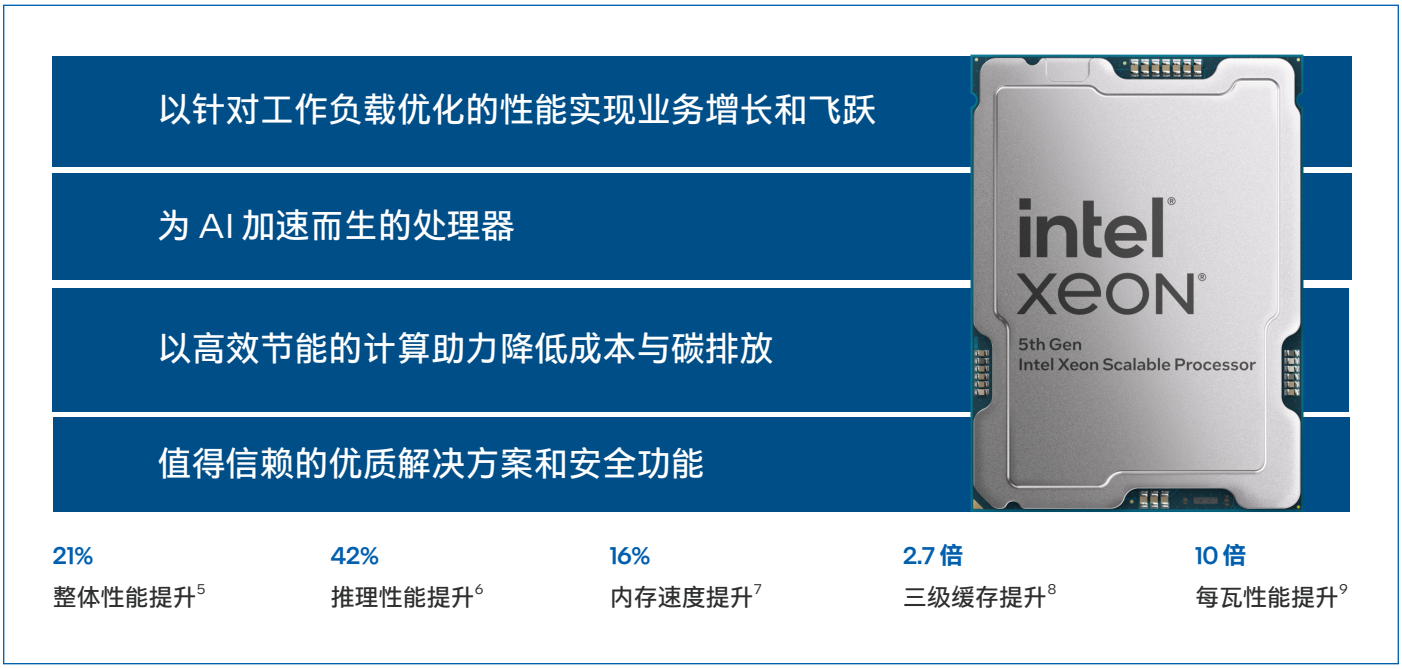

图 1. 第五代英特尔® 至强® 可扩展处理器具备更强大性能

为了进一步提升新一代腾讯云实例 S8 在模型推理等负载中的性能表现，乐元素与英特尔深度利用英特尔® AMX 以及英特尔® oneAPI、英特尔® oneDNN 来进行性能优化。英特尔® AMX 作为英特尔® 至强® 可扩展处理器内置的加速器，可加速基于 CPU 的深度学习推理，避免了使用独立加速器带来的成本和复杂性。英特尔® AMX 在迁移学习和再训练方面同样出色，用户无需额外添置硬件即可使模型保持最新状态。

英特尔® AMX 引入了一种用于矩阵处理的新框架（包括了两个新的组件，一个二维寄存器文件，其中包含称为 “tile” 的寄存器，以及一组能在这些 tile 上操作的加速器），从而能高效地处理各类 AI 任务所需的大量矩阵乘法运算，提升其在训练和推理时的工作效能。例如在向量检索的过程中，如存在 n 个 batch 任务，进行相似度计算时就需要对 n 个输入向量 x 和 n 个数据库中向量 y 进行比对，这其中的距离计算会产生大量的矩阵乘法，而英特尔® AMX 能够针对这一场景实现有效加速。

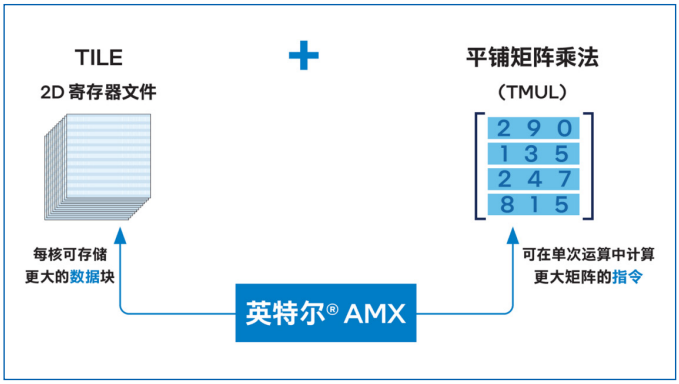

图 2. 英特尔® AMX 架构

[4] 基于截至 2023 年 12 月英特尔的内部建模。

[5] 与第四代英特尔® 至强® 处理器相比的平均性能提升，以 SPEC CPU rate、STREAM Triad 和 LINPACK 的几何平均值为衡量标准。请参阅 intel.com/processorclaims 上的 [G1]：第五代英特尔® 至强® 可扩展处理器。结果可能有所差异。

[6] 与第四代英特尔® 至强® 处理器相比，取得 1.19 倍到 1.42 倍的性能提升（ResNet50v1.5、BERT-Large、SSD-ResNet34、RNN-T（仅 BF16）、Resnext101 32x16d、MaskRCNN（仅 BF16）、DistilBERT）。请参阅 intel.com/processorclaims 上的 [A15-A16]：第五代英特尔® 至强® 可扩展处理器。结果可能有所差异。

[7] 请参阅 intel.com/processorclaims 上的 [G12]：第五代英特尔® 至强® 可扩展处理器。结果可能有所差异。

[8] 请参阅 intel.com/processorclaims 上的 [G11]：第五代英特尔® 至强® 可扩展处理器。结果可能有所差异。

[9] 使用内置加速器在 AI、数据和网络工作负载上进行测量，取得 1.46 到 10.6 倍的每瓦性能提升。请参阅 intel.com/processorclaims 上的 [A19-A25]、[D1]、[D2]、[D5] 和 [N16]：第五代英特尔® 至强® 可扩展处理器。结果可能有所差异。

通过采用英特尔® AMX 技术，乐元素能够提升 AI 性能，满足包括以下场景在内的众多场景的需求：

个性化体验

AI 可以分析玩家的行为和偏好，为每个玩家提供个性化的游戏体验。英特尔® AMX 技术可以助力快速处理玩家数据，以实现快速的游戏元素调整，例如游戏难度、小动物掉落等。

升级的三消对战游戏体验

AI 控制的多人游戏系统可以创造更加真实和吸引人的在线互动，例如 AI 控制的对手棋盘实现 PVP 的游戏体验。英特尔® AMX 可以快速处理大量数据，以提供更加平滑和快速的在线游戏体验。

英特尔® oneDNN 则提供了深度学习构建块的高度优化实现。借助这一开源、跨平台的库，深度学习应用程序和框架开发人员可以对 CPU、GPU 或两者使用相同的 API，从而抽象出指令集和其他复杂的性能优化。在模型性能优化中，通过使用英特尔® oneDNN，操作者只需要调用包含一些后期操作的 MatMul 基元并传递几个参数，oneDNN 即可完成其余的工作，例如配置块寄存器文件、从内存加载数据、使用后期操作执行矩阵乘法计算、将结果存储回内存中，最后释放块寄存器文件。通过使用英特尔® oneDNN，编程人员可大大降低编程的难度。

通过上述优化措施，腾讯云能够在游戏业务中使用 AI 加快关卡学习及迭代速度等场景里提供满足客户性能和服务质量（QoS）需求的解决方案。

性能验证：实现显著的代际性能提升

为了验证在典型的模型推理负载中，基于第五代英特尔® 至强® 可扩展处理器的新一代腾讯云服务器带来的代际性能提升进行了性能测试。

自研打关模型

乐元素自研 AI 打关模型用于自动打关任务，以确保关卡配置无错误，难度符合预期，并快速验证关卡。推理性能测试数据如图 3 所示，对比腾讯云与英特尔联合定制优化的第三代英特尔® 至强® 可扩展处理器，在相同的数据精度下，第五代英特尔® 至强® 可扩展处理器的代际性能提升 1.37 倍，而在启用了英特尔® AMX 将模型从 FP32 转化为 BF16 后，第五代英特尔® 至强® 可扩展处理器的推理性能提升 3.44 倍[10]。

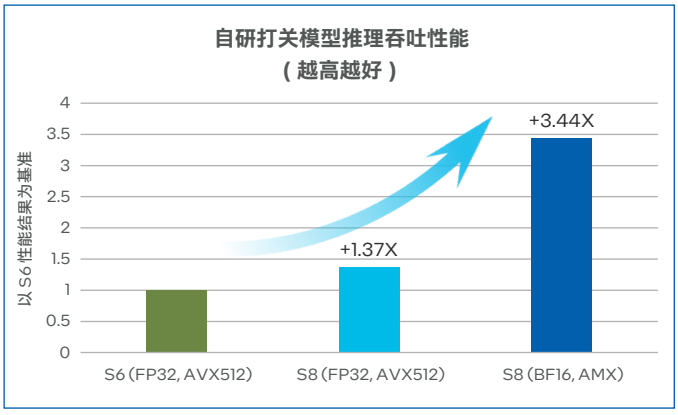

图 3. 自研打关模型推理性能测试数据

[10] 乐元素截止至 2024 年 2 月的内部测试结果，通过比较腾讯云 S8 与 S6 服务器得出。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。

消消乐新春扫龙字活动

乐元素在《开心消消乐》中引入了新春扫龙字活动，在玩家上传扫描的图片后，乐元素会通过 ResNet-50 模型进行图片识别，并返回识别的结果。ResNet（Residual Network）是一种深度学习模型架构，被广泛应用于处理视觉和文本数据的深度学习问题中。ResNet-50 作为 ResNet 的一个变种，在大规模数据集上表现出色，并且在图像分类、目标检测和语义分割等视觉任务中取得了显著的性能。作为一个中等规模的深度学习模型，ResNet-50 在计算资源有限的情况下，仍能够提供优异的性能表现。

《开心消消乐》新春扫龙字活动的模型推理性能的测试数据如图 4 所示，在相同的数据精度下，第五代英特尔® 至强® 可扩展处理器的代际性能提升 1.19 倍，而在启用了英特尔® AMX 后第五代英特尔® 至强® 可扩展处理器的推理性能提升 5.19 倍[11]。

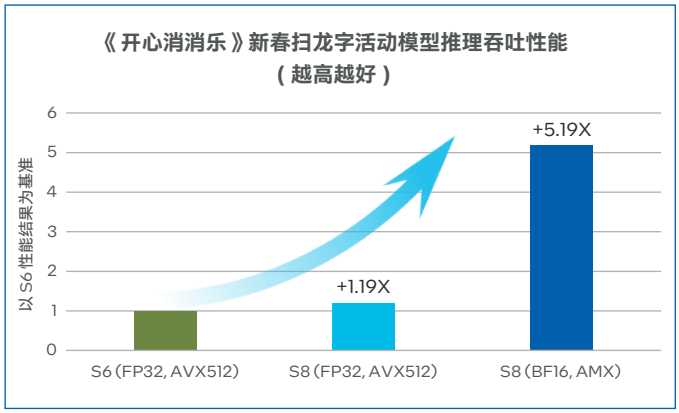

图 4.《开心消消乐》新春扫龙字活动模型测试数据

收益

通过硬件升级以及软件优化，基于第五代英特尔® 至强® 可扩展处理器的腾讯云服务器能够显著提升乐元素在模型推理等负载中的性能表现，为其带来以下收益：

- **满足关卡设计的 AI 算力需求，提升游戏开发与运营效率：**通过高性能的第五代英特尔® 至强® 可扩展处理器，乐元素获得了充沛的 AI 算力支撑，能够游刃有余地应对自动打关等模型的推理性能需求，提升游戏开发与运营效率。
- **降低成本，实现效益化游戏运营：**通过部署基于第五代英特尔® 至强® 可扩展处理器的腾讯云实例，乐元素无需采用昂贵的专用 AI 服务器，而且能够按需进行扩展，有助于乐元素在 AI 战略中获得更高的投资回报率。
- **灵活应对其它 AI 扩展应用：**除了关卡设计之外，乐元素还积极在游戏开发与运营的其它环节中引入 AI 技术，基于第五代英特尔® 至强® 可扩展处理器的腾讯云可满足这些 AI 模型推理的算力需求。

展望

AI 技术已经成为游戏产业发展的热门技术方向，采用生成式人工智能（AIGC）等应用辅助原画设计、角色生成、脚本设计已经广泛盛行，并被应用到大量成功的游戏作品中。一份研究报告预计，2024 年 AI 技术应用将为游戏公司带来约 21% 的人力成本下降[12]，释放生产力的同时，人才布局重心将向创新力转移。在此背景下，构建面向游戏开发与运营的 AI 算力平台，推动 AI + 游戏应用的创新，成为影响游戏公司竞争力的关键因素。

乐元素的实践证实，基于第五代英特尔® 至强® 可扩展处理器的腾讯云实例 S8 能够满足典型 AI 模型在推理算力上的需求，同时具备更高的经济性与灵活性，能够成为游戏企业拓展 AI 应用的理想选择。在当前合作成果的基础上，英特尔将与腾讯云和乐元素展开更多合作，加快将 AI 融入到游戏开发与运营的整体流程之中，为玩家带来更加卓越的游戏体验。

[11] 乐元素截止至 2024 年 2 月的内部测试结果，通过比较腾讯云 S8 与 S6 服务器得出。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。

[12] 数据援引自：天风证券：《数据研究·科技行业专题：游戏产业人士看 AI 应用前景》。

关于乐元素

乐元素成立于 2009 年，从事移动网络游戏的研发及运营，同时开发原创 IP，并提供基于原创 IP 的演出、音乐、动画片、授权业务和周边商品等产品及服务。乐元素在北京、上海、京都、东京、广州等城市均设有游戏工作室和研究机构，至今已吸纳来自世界各地的人才 1500 余人。乐元素旗下拥有《开心消消乐》、《开心水族箱》、《海滨消消乐》、《松松总动员》等多款畅销产品，同时也拥有全世界最具影响力的虚拟偶像组合 Ensemble Star!《偶像梦幻祭》。乐元素致力于把产品和快乐传递到全世界的每一个角落，「创造更好的娱乐文化体验、让更多人感受到美好与欢乐」。

关于腾讯云

腾讯云是中国领先的互联网综合服务提供商腾讯集团旗下的云计算品牌，面向全世界各个国家和地区的企业、组织、机构和个人开发者，提供全球领先的云计算、人工智能、大数据等技术产品与服务。作为产业互联网的基础设施，腾讯云以卓越的技术能力打造丰富的行业解决方案，构建开放共赢的云端生态，助力各行各业实现数字化升级。腾讯云的基础设施覆盖全球五大洲 26 个地理区域，运营 70 个可用区，全球各地服务器数量超过 100 万台，是中国首家服务器总量超过百万的公司，也是全球五家服务器数量过百万的公司之一；在全球范围内部署超过 2800 个加速节点，带宽储备达 200T，为更多企业提供强有力的技术支持，助力业务飞速拓展。

关于英特尔

英特尔（NASDAQ: INTC）作为行业引领者，创造改变世界的技术，推动全球进步并让生活丰富多彩。在摩尔定律的启迪下，我们不断致力于推进半导体设计与制造，帮助我们的客户应对最重大的挑战。通过将智能融入云、网络、边缘和各种计算设备，我们释放数据潜能，助力商业和社会变得更美好。如需了解英特尔创新的更多信息，请访问英特尔中国新闻中心 newsroom.intel.cn 以及官方网站 intel.cn。

intel

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

性能测试结果基于配置信息中显示的日期进行测试，且可能并未反映所有公开可用的安全更新。详情请参阅配置信息披露。没有任何产品或组件是绝对安全的。

具体成本和结果可能不同。

英特尔技术可能需要启用硬件、软件或激活服务。

英特尔未做出任何明示和默示的保证，包括但不限于，关于适销性、适合特定目的及不侵权的默示保证，以及在履约过程、交易过程或贸易惯例中引起的任何保证。

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。