案例研究

第二代英特尔® 至强® 可扩展处理器
英特尔® 傲腾™ 数据中心级持久内存
英特尔® 深度学习加速技术
数据库云服务
AI 即服务

腾讯云：基于至强® 可扩展平台打造多元、高效云服务

经过十余年高速发展，云服务不仅成为了众多企业和机构应对数据处理、支撑业务发展的基石，其自身所覆盖的服务类型，以及在服务能力的完善上，也更趋广泛和深入。尤其当基于云的人工智能（Artificial Intelligence，以下简称 AI）、大数据分析等创新应用逐渐走向成熟，以数据库或数据分析即服务、AI 即服务为代表的新一代多样化云服务，更是成为企业实施数字化转型的良好臂助。作为中国云服务行业的翘楚，腾讯云* 正以创新为使命，致力于为用户提供这些更为敏捷、高效、可靠和多元化的云服务。

如此高效能的云服务，离不开强有力的 IT 基础设施作为支撑，作为腾讯云长久以来的创新伙伴，“以数据为中心”的英特尔一直以提供领先的计算、存储、网络连接用硬件产品与技术为使命，希望能通过输出更全的数据计算、更强的数据存储以及更快的数据传输能力，助腾讯云进一步优化其应用和技术框架，从而达到升级、拓展和优化云服务的目标。

基于这一合作体系，腾讯云第一时间引入了英特尔全新一代至强® 可扩展平台中的核心产品和技术，包括集成有英特尔® 深度学习加速技术的为腾讯定制的第二代英特尔® 至强® 可扩展处理器，以及英特尔® 傲腾™ 数据中心级持久内存。这些技术的搭配组合，不仅让腾讯云基于 AI 的智能视频分析、视频鉴黄等创新云服务实现了显著的效率提升，还使它能够为用户提供更具性价比的 Redis* 云数据库服务。

作为中国主要的云服务提供商之一，腾讯云正以其深厚的技术积淀、丰富的行业经验以及良好的基础设施能力，不断引领创新，为众多用户提供敏捷、高效、可靠和多样化的云服务。

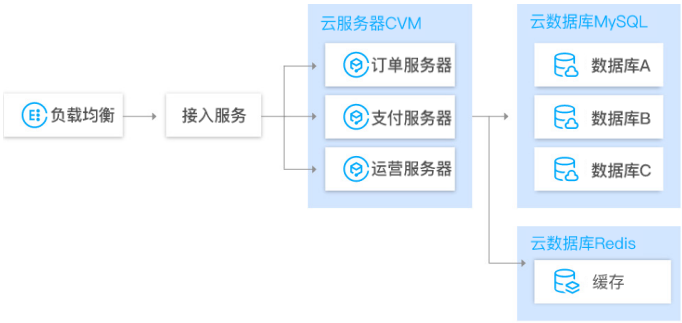
所谓“工欲善其事，必先利其器”，高效可靠的云服务能力离不开强有力的 IT 基础设施能力作为支撑。为助力腾讯云进一步升级、完善或革新其云服务的性能和体验，英特尔为它提供了全新的英特尔® 至强® 可扩展平台。该平台以第二代英特尔® 至强® 可扩展处理器为核心，凭借集成在处理器内的英特尔® 深度学习加速技术，及该处理器对英特尔® 傲腾™ 数据中心级持久内存的良好支持，使腾讯云基于内存数据库技术的 Redis 云数据库服务，以及基于 AI 的智能视频分析和视频鉴黄等多项创新云服务，实现了更为出色的性能和可用性。

“基于行业领先的硬件架构，致力于为用户提供更出色的服务体验，是腾讯云一直以来的愿景与使命。通过与英特尔的深入合作，特别是第二代英特尔® 至强® 可扩展处理器和英特尔® 傲腾™ 数据中心级持久内存的引入，腾讯云得以在 AI 云服务、数据库云服务等领域持续占据创新先机，为用户提供更为多元化、差异化和更加高效的云服务。”

刘颖
副总裁
腾讯云

更具性价比的 Redis 云数据库服务

以内存数据库技术为基石的 Redis 云数据库，一直以高性能、高灵活、低响应时延以及丰富的数据结构类型等特性而备受用户青睐，在业务缓存、会话存储、消息队列和信息发布等应用场景中发挥着关键作用。Redis 云数据库服务作为腾讯云的核心业务之一，通过与其他云服务能力的配合，能有效地帮助用户优化业务流程，提升经营效率。



图一 腾讯 Redis 云数据库服务在电子商务场景中的应用

以电子商务行业为例，如图一所示，线上商家的商品展示、购物推荐等对时延要求较高的关键数据，可以通过存储在 Redis 云数据库构建的缓存中，来实现更为快速的访问。目前，腾讯 Redis 云数据库已能够提供 10 万级的每秒查询量（Query Per Second，QPS），足以轻松应对大型促销秒杀活动中的高并发访问需求[1]。

但动态随机存取存储器（Dynamic Random Access Memory，DRAM）内存昂贵的价格，大大增加了腾讯 Redis 云数据库的成本，进而也限制了用户的使用规模。为帮助用户获得更多高可用的 Redis 云数据库服务，腾讯云通过与英特尔合作，引入第二代英特尔® 至强® 可扩展处理器，并搭配了英特尔® 傲腾™ 数据中心级持久内存，让腾讯 Redis 云数据库服务获得了更为“物美价廉”的内存容量扩展方案。

基于英特尔® 3D XPoint™ 技术打造的英特尔® 傲腾™ 数据中心级持久内存，可为用户带来更大容量和数据持久性支持的全新内存应用体验。一方面，它能够提供最高达 512GB 的内存容量，另一方面，数据持久性可确保设备断电后，它仍能够保存数据，从而提升数据安全性。

在提供大容量和存储持久性的同时，英特尔® 傲腾™ 数据中心级持久内存也为腾讯 Redis 云数据库服务带来了更高的性价比。源自腾讯云的对比测试数据显示，在同一服务水平协议(SLA)等级上，搭载傲腾™ 数据中心级持久内存的平台所能提供的单实例内存容量，可扩展至只配置有 DRAM 内存的平台的 1.34 倍[2]，这能让最终用户在同等成本下，获取更优质的 Redis 云数据库服务能力。

腾讯在这一对比测试中使用了两个平台，它们均基于双路英特尔® 至强® 铂金 8260 处理器。该款处理器具有 24 核心/48 线程，主频为 2.4GHZ。在总拥有成本（Total Cost of Ownership，TCO）近似的情况下，对比组每路处理器只配置了 384GB DRAM 内存（32GB*12）；而测试组则采用了 DRAM 内存和英特尔® 傲腾™ 数据中心级持久内存的混合配置，配比为 1:5.3，每路处理器配备 96GB DRAM 内存和 512GB 傲腾™ 数据中心级持久内存。测试中，两组平台均启动了 88 个实例。

模型	测试组 DRAM 物理内存与英特尔® 傲腾™ 数据中心级持久内存混合配置	对比组 仅配置 DRAM 物理内存
单实例内存容量 (GB)	11.27	8.36
全量写吞吐量 (TPS/实例)/ 1K 数据下 P99 延迟（毫秒）	51k/1.86	58k/1.25
全量读吞吐量 (TPS/实例)/ 1K 数据下 P99 延迟（毫秒）	63k/0.82	61.5k/0.71
TCO	0.986	1
单实例内存容量/总拥有成本	1.36	1

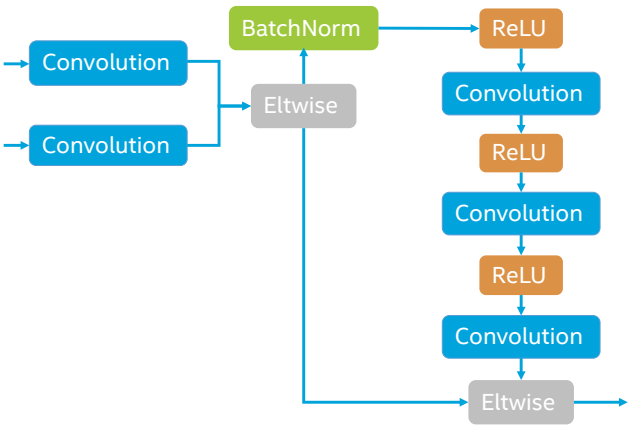
表一 腾讯云分别配置 DRAM 物理内存与 DRAM 内存+英特尔® 傲腾™ 数据中心级持久内存后的对比测试结果

测试结果如表一所示，分别配置 DRAM 内存与英特尔® 傲腾™ 数据中心级持久内存的两个平台，在全量读写吞吐量、1K 数据下的 P99 延迟等性能指标上均表现相近。而在 TCO 相近的情况下，采用 DRAM 内存和傲腾™ 数据中心级持久内存混合配置的平台，在单实例内存容量上明显优于仅配置 DRAM 内存的平台，前者的单实例内存容量是后者的 1.34 倍，比较单实例内存容量 /TCO 时，则是后者的 1.36 倍[3]。

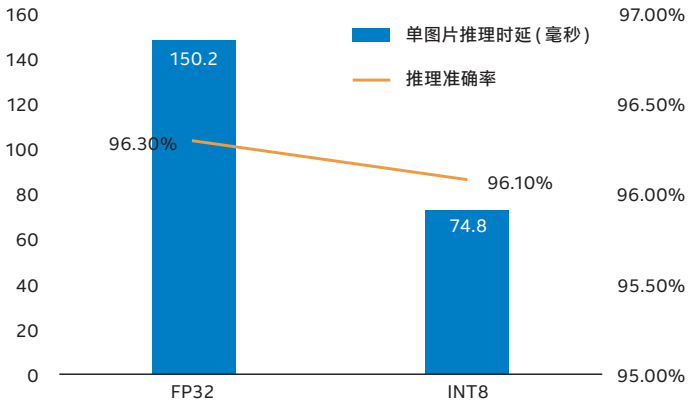
效率翻番的视频鉴黄能力

CDN 服务是目前互联网服务和云服务的重要组成部分，它可为各类互联网应用提供强大的内容分发能力，尤其是在游戏直播、短视频等视频类应用上，CDN 服务可明显改善或优化用户的收视体验，提升用户的满意度。但与此同时，混杂在视频中的不良内容，也成为互联网视频应用发展的毒瘤。虽然在 CDN 服务中加入视频鉴黄模块可有效应对这一问题，但随之又可能会出现因鉴黄效率不高导致的视频延迟问题，造成抖动、拖影等，进而影响用户体验的情况。

通过与腾讯优图* AI 开发平台提供的深度学习解决方案相结合，腾讯云在其 CDN 云服务中添加了更为可靠、高效的视频鉴黄能力。它可在实时视频流中提取图像帧，并采用深度神经网络(Deep Neural Networks，DNN）模型对图像帧实施AI推理，从而判别视频流是否涉黄。



图二 腾讯云深度神经网络拓扑示意



图三 腾讯云视频鉴黄深度学习模型在不同计算精度下的推理准确率和速度比较

如图二所示，腾讯云视频鉴黄采用了 ResNet v2 变体深度神经网络模型。为提升其工作效率和速度，腾讯云一方面引入面向深度神经网络的英特尔® 数学核心函数库（Intel® Math Kernel Library for Deep Neural Networks，英特尔® MKL-DNN）来优化模型推理效能；另一方面，第二代英特尔® 至强® 可扩展处理器的加入，特别是其英特尔® 深度学习加速技术的加持，也令其鉴黄效率实现了大幅提升。

传统上，AI 推理多是基于 32 位浮点计算开展，32 位浮点计算在保证推理准确率的同时，也带来了巨大的计算量和部署复杂度。而在图像识别、图像分类等应用场景中，INT8 等低精度定点计算，在推理准确率上完全能与 32 位浮点计算相媲美，同时还能带来显著的推理速度提升。为腾讯定制的第二代英特尔® 至强® 可扩展处理器不仅可凭借更优的微架构、更多的内核和更快、更大容量的内存支持能力为腾讯云带来更强劲的基础算力，其集成的全新英特尔® 深度学习加速技术，对于基于 INT8 的深度学习模型推理，也有着非常出色的加速效果，再通过与英特尔® MKL-DNN 的配合，可在不影响推理准确率的情况下，大大提升深度学习模型的推理速度。

英特尔® 深度学习加速技术能有如此效能，因为其 VNNI 指令集可为深度学习模型提供多条全新的宽融合乘加 (FMA) 内核指令，可用于 8 位或 16 位低精度数值相乘，这对于需要执行大量矩阵乘法的推理过程而言尤为重要。该技术的引入，使深度学习系统在执行 INT8 推理时，对系统内存要求最大可减少 75%[4]，而在内存和带宽需求上的减少，则能大大加快低精度数值运算的速度。

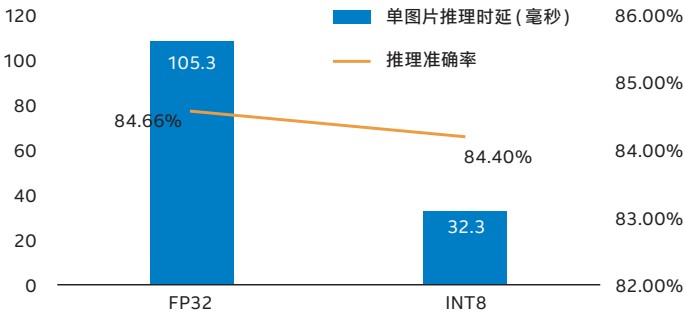
由腾讯云进行的实际测试，也有力地证明了这一点。该测评基于为腾讯定制的第二代英特尔® 至强® 可扩展处理器平台展开。该款处理器具有 24核心/48 线程，主频为 2.5GHZ，并搭载了 192GB 内存；同时在测试中，系统使用了英特尔® MKL-DNN v0.17 版来实现软件优化。测试对比了在使用单处理器核心，Minibatch=1 配置的情况下，分别使用 32 位浮点计算和基于英特尔® 深度学习加速技术的 INT8 定点计算，对腾讯私有标签样本数据集（约 1 万张图片）进行视频鉴黄深度学习模型推理时的准确率和速度。测试结果如图三所示，在使用基于英特尔® 深度学习加速技术的 INT8 进行推理时，推理准确率仅比采用 32 位浮点计算时下降 0.2%，即两者准确率基本一致。但同时，使用 INT8 进行推理，其单图片推理时延为 74.8 毫秒，与采用 32 位浮点计算时的 150.2 毫秒相比，下降了 50.2%。这就意味着，新一代英特尔® 至强® 可扩展平台的导入，可令腾讯云视频鉴黄效率提升达 2 倍以上[5]。

更高性能的智能视频分析能力

同样，在腾讯云提供的其他基于 AI 的云服务中，英特尔® 深度学习加速技术也大有用武之地。例如腾讯云在音、视频解决方案中新增的视频智能化分析服务能力，它能够允许用户在网络游戏直播、美容直播、足球比赛直播等 25 种场景中，对视频实施分类、打标签以及抽取精彩画面等操作，从而帮助用户更为便捷地根据视频内容，衍伸出丰富多样的应用与服务。

腾讯云智能视频分析能力同样也采用了深度学习方法（Inception V3 模型，并通过针对英特尔® 架构优化的 Caffe 框架进行部分修改）构建。因此，在引入第二代英特尔® 至强® 可扩展处理器，利用其集成的英特尔® 深度学习加速技术之后，后者基于 INT8 定点计算的更高推理效率，也使智能视频分析的整体效率获得了飞跃式的提升。

为验证英特尔这项技术为智能视频分析带来的功效，腾讯云同样基于为腾讯定制的第二代英特尔® 至强® 可扩展处理器平台进行了测试。该处理器具有 24 核心/48 线程，主频为 2.5GHz。平台搭载了 192GB 内存，并使用了包含针对英特尔® 架构优化的 Caffe v1.1.3 版本的英特尔® MKL-DNN。



图四 腾讯云智能视频分析深度学习模型在不同计算精度下的推理准确率和速度比较

这项测试对比了在使用单处理器核心，Minibatch=1 配置的情况下，分别使用 32 位浮点计算和基于英特尔® 深度学习加速技术的 INT8 定点计算，对腾讯私有标签样本数据集（约 4 万张图片）进行视频智能化分析深度学习模型推理时的准确率和速度。测试结果如图四所示，在使用基于英特尔® 深度学习加速技术的 INT8 进行推理时，推理准确率仅比使用 32 位浮点计算时下降 0.26%，两者准确率基本保持一致。但同时，在使用 INT8 进行推理时，其单图片推理时延为 32.3 毫秒，仅为使用 32 位浮点计算时的推理时延——105.3 毫秒的 30.7%。由此可见，在采用相同硬件和算法的情况下，引入英特尔® 深度学习加速技术的新方案，使腾讯云智能视频分析的效率提升到了原先的 3.26 倍[6]。

展望

上述一系列基于实际云服务环境的测试表明：新一代英特尔® 至强® 可扩展平台中的几项核心产品技术，包括第二代英特尔® 至强® 可扩展处理器、英特尔® 深度学习加速技术以及英特尔® 傲腾™ 数据中心级持久内存的引入，切切实实地为腾讯云带来了更全面、更优秀也更为均衡的 IT 基础设施能力。腾讯云也将以此为基础，在未来与英特尔继续携手，在不断优化现有平台能力的同时，围绕高性能基础设施之上的高品质云服务，开展更为广泛且深入的合作。

腾讯云实现的全新云服务优势：

- 第二代英特尔® 至强® 可扩展处理器对英特尔® 傲腾™ 数据中心级持久内存有着良好的支持，这帮助腾讯云在同样的 SLA 及相应的成本前提下，将 Redis 云数据库服务的单实例内存容量提升达 1.34 倍[7]；
- 为腾讯定制的第二代英特尔® 至强® 可扩展处理器所集成的英特尔® 深度学习加速技术，可为腾讯云智能视频分析和视频鉴黄的深度学习模型提供出色的加速能力，在采用相同硬件、算法并取得相近准确率的情况下，可助视频分析效率提升达 3.26 倍[8]，视频鉴黄速度提升达 2 倍[9]。

[1] 数据源引自腾讯云官网介绍：https://cloud.tencent.com/act/pro/redis?fromSource=gwzcw.1345398.1345398.1345398

[2, 3, 7] 数据援引自腾讯云基于第二代英特尔® 至强® 可扩展处理器开展的 Redis 云服务测试，测试配置为：测试组采用双路英特尔® 至强® 铂金 8260 处理器，24 核心/48 线程，启用 HT/Turbo 技术，BIOS 版本为 1.018，每路处理器搭载 96GB DRAM 物理内存和 512GB 英特尔® 傲腾™ 数据中心级持久内存，单块 25GbE 网络适配器，Linux Kernel 版本为 4.14.68-1-tlinux3-nvdimm-0005，Redis 版本为 4.10，原始数据量为 11.27GB，测试组共启动 88 个虚拟机实例；对比组采用双路英特尔® 至强® 铂金 8260 处理器，24 核心/48 线程，启用 HT/Turbo 技术，BIOS 版本为 1.018，每路处理器搭载 384GBDRAM 物理内存，单块 25GbE 网络适配器，Linux Kernel 版本为 4.14.68-1-tlinux3-nvdimm-0005，Redis 版本为 4.10，原始数据量为 8.36GB，对比组共启动 88 个虚拟机实例。

[4] 数据源引自https://software.intel.com/en-us/articles/lower-numerical-precision-deep-learning-inference-and-training

[5, 9] 数据援引自腾讯云基于第二代英特尔® 至强® 可扩展处理器开展的视频鉴黄云服务测试，测试配置：处理器采用为腾讯定制的第二代英特尔® 至强® 可扩展处理器，主频为 2.5GHz，24 核心/48 线程，启用 HT/Turbo 技术，搭载 192GB 内存，操作系统为 CentOS 7.6，Kernel 版本为 3.10.0-957.el7.x86_64，编译器版本为 GCC4.8.5，测试中工作负载基于腾讯云 CDN 非法图像过滤展开，使用英特尔® MKL-DNN 库 V0.17 版，FP32 和INT8 数据对比测试均在单处理器核心，Minibatch=1 配置下完成。

[6, 8] 数据援引自腾讯云基于第二代英特尔® 至强® 可扩展处理器开展的智能视频分析云服务测试，测试配置：处理器采用为腾讯定制的第二代英特尔® 至强® 可扩展处理器，主频为 2.5GHz，24 核心/48 线程，启用 HT/Turbo 技术，搭载 192GB 内存，操作系统为 CentOS 7.6，Kernel 版本为 3.10.0-957.el7.x86_64，编译器版本为 GCC4.8.5，测试中工作负载基于腾讯云视频分析展开，使用包含有针对英特尔® 架构优化的Caffe V1.1.3 版本的英特尔® MKL-DNN 库，FP32 和 INT8 数据对比测试均在单处理器核心，Minibatch=1 配置下完成。