
intel
XEON
英特尔互联网行业
音视频创新实践
高效 灵活 可靠
构建云创新基石

Contents
目录

概述篇

音视频能力正成为互联网新基建的必选项

从短视频到网红直播，从云游戏到 VR 电竞，从视频会议到虚拟机器人，各种纷繁复杂、绚丽颖奇的新应用，在为人们带来多姿多彩的互联网生活体验之余，也悄然改变着互联网世界的流量版图。在这一进程中，传统以文本、图片等为主的互联网信息流也逐渐被以音视频为代表的多媒体流所取代。以目前火热的短视频服务为例，至 2021 年末，互联网用户对短视频服务的粘性已成为各互联网服务之首，使用总时长达 25.7%。[1]

这一变化从近来年通信技术发展带来的属性优势、各互联网热门应用产品设计和市场趋势亦可见端倪：

- 通信与网络技术的迅猛发展，尤其是 5G 网络、WiFi6 等新一代高速无线通信技术的普及，让更大的网络带宽和更快的接入速度为音视频与网络应用更顺畅地融合提供了基础；
- 人工智能（Artificial Intelligence，AI）、增强现实（Augmented Reality，AR）/ 虚拟现实（Virtual Reality，VR）以及 4K/8K 高清等技术的出现，让音视频在各类应用中的核心作用愈发明显；
- 音视频市场规模突飞猛进，催化着音视频技术与产品的快速发酵，让整个互联网行业内外对音视频领域都更为关注。

事实上，音视频能力已不仅在那些为人们所熟知的视频和音乐等类别的应用中至为关键，在游戏电竞、社交娱乐、互动电商以及在线会议等场景中，音视频数据流同样也是其重要基石。例如，在日前火热的“元宇宙（Metaverse）”概念中，音视频数据流的处理效能就是打造高品质沉浸式体验，乃至关乎其产品能否成功的一大核心要素。

同样，近两年来的疫情也使更多的办公交流、贸易推介、技术研讨等商务学术活动转移到了线上，也对音视频市场的爆发式增长产生了极大助推作用。以新一代网页即时通信（Web Real-Time Communication，WebRTC）为例，其相关的技术与服务，包括视频会议、语音会议、音视频服务等在未来的数年中将以 43.6% 的复合年均增长率高速发展[2]。因此，构筑高品质的音视频能力无疑已成为互联网产业未来发展的致胜要决之一，甚至有观察家预言：“音视频能力的构建，将成为互联网新基建的必选项。”

[1] 数据援引自公开媒体报道《QuestMobile2021 中国移动互联网年度大报告》：http://news.lznews.cn/guonei/202202/t20220222_9854580.html

[2] 数据援引自公开媒体报道：https://www.sohu.com/a/298557136_458408

基于音视频新赛道的舞台视角：更多挑战与收益

随着新一代互联网应用拉开序幕，全新音视频技术与能力在为更多技术融合、更多产品形态、更多应用场景提供秀场的同时，也正因各行业的差异化需求而遭遇更多的挑战。这些挑战包括：

- 在全球云计算渗透率不断上升的今天，各类云服务（包括公有云、私有云和混合云等）已逐渐成为各类互联网应用的重要基础设施，音视频能力与云平台的“无缝”对接也成为值得关注的重点；
- 不断“扩容”的音视频数据流正对各类互联网应用的产品品质、工作效率和运营成本提出挑战；要在保证品质的同时提升系统效率，高可用的音视频编解码、转码、格式转换和解压缩能力必不可少；
- 许多热门互联网应用，如远程辅助医疗、智能家居、元宇宙等，都体现出音视频能力与 AI 等技术的深度融合，如何为音视频能力提供高效的 AI 框架和加速能力是各应用厂商能否赢得竞争力的关键；
- 愈加复杂的产品设计与功能呈现也对承载音视频能力的硬件基础设施提出更高要求，这些要求包括更快的计算处理能力、更灵活的算力分布方案、更优的数据存储性能以及更强的网络吞吐能力等。

以 5G 为代表的新一代网络技术、以深度学习（Deep Learning）为引领的前沿 AI 应用以及更多样化的云服务能力（包括在基础设施即服务（Infrastructure as a Service，IaaS）、平台即服务（Platform as a Service，PaaS）和软件即服务（Software as a Service，SaaS）能力上的分别演进），通过与来自英特尔等提供的先进软硬件产品与技术相融合，正为上述挑战提供更佳解决方案。如今，由这些全新解决方案推动着高速发展的互联网应用包括：

- **基于云或数据中心的音视频处理能力：**如云游戏、多方视频会议等；
- **专业高效的编解码 / 转码 / 格式转换能力：**如视频点播 / 直播、短视频、视频电商等；
- **拥抱 AI 能力的各类音视频新应用：**如智能语音、元宇宙应用等。

与传统音视频方案相比，这些新应用、新模式与新场景无疑能让企业与用户分别从中获益。对普通用户而言，其意味着可通过网络连接获得使用更流畅、交互更简捷、价格更实惠的视听体验。以云游戏为例，其可帮助玩家在免去购置高配置终端之余，还能随时随地轻松玩到各类游戏大作。而对企业而言，编解码、压缩等新技术能力的出现，也使企业能更有效地推动系统优化，降低运营成本。

英特尔先进软硬件产品与技术助力音视频能力构建

在各类音视频能力的落地实施上，英特尔正通过一系列先进产品与技术方案，提供和优化算力、存储、网络以及软件能力，来满足上述各应用在软硬件、系统架构以及生态构建上的独特需求。

一方面，在硬件基础设施上，英特尔通过英特尔® 至强® 可扩展处理器、英特尔® FPGA 产品、英特尔® 傲腾™ 持久内存，以及英特尔® 以太网网络适配器、英特尔® 视觉云媒体分析加速卡等产品，为各类基于音视频能力的创新方案提供强劲的计算、存储和网络处理能力；另一方面，在软件优化加速上，来自英特尔的 Media SDK、SVT、英特尔® oneAPI 等，在不同应用场景中以完整的软件栈来加速音视频能力的工作效能。

基于这些产品和技术方案，英特尔正与众多合作伙伴一起，以灵活可扩展的生态和各类成熟的解决方案，满足更多互联网应用对音视频能力的需求，共同推动互联网迈向动态、丰富、多维和可交互的新纪元。

应用篇

直播场景

作为互联网应用与文化传媒相融合的产物，直播行业在近年来无疑获得了巨大的关注与高速的发展。数据表明，截至 2021 年底，我国网络直播用户规模已达 7.03 亿。[3]

图 1 典型直播场景

随着行业发展的逐渐深入及竞争的日趋激烈，直播形式也从单一的内容播放、用户互动逐渐转向更多元化的模式，包括：

- **电商直播发展迅猛：**直播逐渐成为电商运营中的“标配”且用户量可观，2021 年直播电商用户平均年消费超过 2,500 元，平均增长达 12%[4]；
- **“直播 + 行业”紧密融合：**通过直播为行业发展赋能成为众多企业的共识，企业直播市场发展迅速，市场规模已达数十亿元[5]；
- **大幅引入 AI 能力：**AI 能力的发展驱动虚拟技术变革传统视频直播形式；除耳熟能详的换肤等特效外，AR 及 VR 技术也在打造全新的视听直播形式。

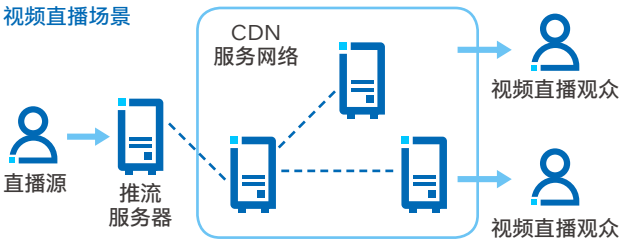


图 2 传统的直播方案

[3] 数据援引自公开媒体报道：https://www.askci.com/news/chanye/20220318/14163217746317.shtml

[4] 数据援引自益普索发布的《2021 直播电商趋势报告》，http://www.ec100.cn/detail--6611672.html

[5] 数据援引自艾瑞咨询发布的《2022 年中国企业直播行业发展趋势研究报告》，https://www.thepaper.cn/newsDetail_forward_17480554

不可避免地，这些多元化的直播趋势也正推动着直播技术方案发生变革。与传统方案相比，新直播场景的变化以及所需的技术包括：

新直播场景的变化	新技术需求
各类高清视听设备的普及，需要方案能更快应对更高清晰度的视频编解码和转码。	可高速应对 1080P、2K 乃至 4K 视频的转码与编解码能力。
与电商平台、业务平台等更多服务能力的连接，需提升数据存储能力。	打造具有高性能数据读写、存储能力的分布式数据库、缓存数据库。
更多 AI 能力在直播中的加入，如产品推荐、虚拟形象以及 AR/VR 功能等。	能够对各类 AI 框架提供有效加速。

面对新直播场景下的 IT 需求，英特尔正以全面的、功能丰富且性能强劲的产品与技术，为其提供一系列的支持，具体包括：

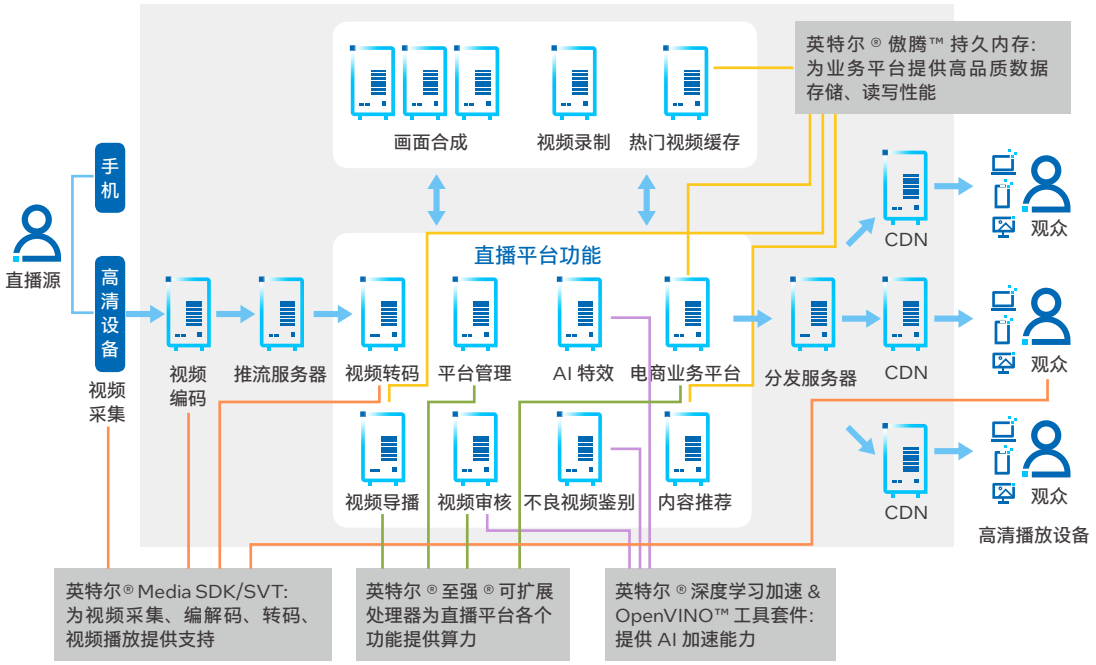


图 3 变革中的视频直播场景

- **英特尔® 至强® 可扩展处理器：**大幅提升的基础性能与多种内置先进特性，为直播场景中的密集工作负载提供更佳算力支撑；
- **英特尔® AVX-512：**具有更大寄存器的SIMD指令，为视频直播提供具有优势的音视频编解码、转码以及特效计算能力；
- **英特尔® Media SDK：**可为直播中的视频播放、编解码、转码和媒体格式转换提供性能增强；
- **英特尔® SVT：**以灵活的高性能软件编码器核心库提升直播中的视频处理效率；
- **英特尔® 傲腾™ 持久内存：**通过创新内存技术，为直播场景中的业务数据提供高品质数据存储、读写性能；
- **英特尔® 深度学习加速 & OpenVINO™ 工具套件：**为AI应用中的深度学习训练和推理过程提供加速，推动AI能力在直播场景的落地。

面对新直播场景下的不同 IT 需求，英特尔全栈的产品与技术，为其提供了广泛支持并在一系列客户场景中取得了良好的实践反馈。在下文案例中，将围绕英特尔产品在金山云集智高清产品和腾讯视频云平台中的应用及客户获益展开介绍。

英特尔助金山云实现集智高清产品 显著节省传输带宽

通过对 AI、编码、图像处理等多种技术的整合，金山云正借助深度神经网络，以“集智高清”产品为直播服务提供效能增强。如图 4 所示，在视频场景分类上，“集智高清”通过深度学习应用的加入，形成了十余大类、几十种小类视频场景模型库。客户在使用集智高清服务时，可实时分析直播流，匹配相应的视频场景模型，根据客户对视频画质的要求对画质进行实时处理。

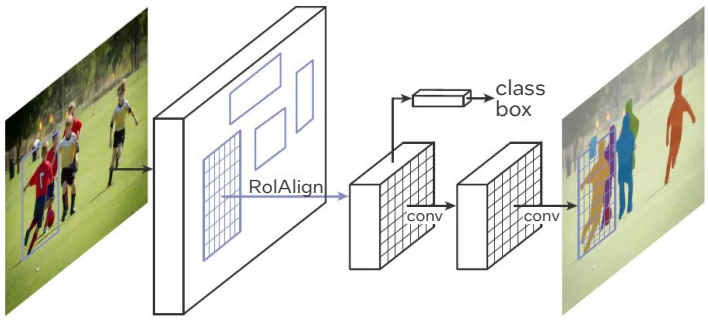


图 4 集智高清中应用的视觉重点分割技术

另外，“集智高清”还使用了视觉重点分割等技术，在处理视频时可对不同区域进行分层，针对每一层进行处理，保护人眼视觉核心区域，对此区域内的宏块编码进行锐化、亮度增强等技术处理，降低不重要的区域的码率，将视域进行精细化处理，加强用户的视觉体验。

另一方面，“集智高清”也借助金山云打造的块级智能决策、AI 修复等创新功能，通过转码能力来实现码率降低的效果，有效提升直播服务质量。其中在主观增益上，一方面块级智能决策可以很好地避免块效应，减少低清视频比率，降低产生用户反感的可能性；另一方面利用 AI 的修复能力，能够有效避免第一次编码可能造成的压缩噪声、ROI 修复、去除运动模糊，还能够利用帧间信息修复对焦失真产生的模糊。

为了在帮助用户提升画质、降低带宽压力的同时，提升云转码的性能表现，实现更高的性能密度，金山云在“集智高清”的云服务器中搭载了英特尔® 至强® 铂金 8358P 处理器，该处理器属于第三代英特尔® 至强® 可扩展处理器家族，基于新一代 Sunny Cove 微架构而开发，针对公有云应用负载的特点进行了定制与优化，提供了高达 32 个物理核心，运行频率达到 2.6GHz。与第二代英特尔® 至强® 可扩展处理器相比，该处理器在单核性能、核心数量等方面都实现了显著提升，为“集智高清”系统的云转码性能提升奠定了坚实基础。

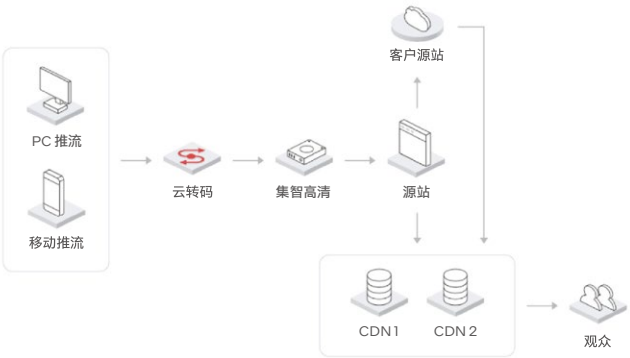


图 5 与云转码能力协同的金山云集智高清服务

在使用英特尔® 至强® 铂金 8358P 处理器的基础上，金山云使用了英特尔® AVX-512 来优化卷积的重载。英特尔® AVX-512 是最新的 x86 矢量指令集，用以提升要求苛刻的计算任务的性能，能够显著提升媒体处理集群的计算能力，提供各种高性能图像处理解决方案，有效减小在线图像处理时延和带宽问题。金山云直接使用英特尔® AVX-512 指令集对固有的卷积进行处理，与传统卷积实现相比，这种方式具备无需对源图像进行扩边、无需填充过滤器、无需传输整个过滤缓冲区、无需旋转源图像/滤镜/输出等优势，能够加速性能表现。

如图 6 所示，在金山云环境下进行的测试数据显示 [6]，通过使用英特尔® AVX-512 指令集，ERJND 模块能够实现 48-103 倍的性能提升。而且随着分辨率的提升，性能的领先幅度还在不断扩大。

英特尔® 集成性能原件（Intel® Integrated Performance Primitives，英特尔® IPP）也为新方案提供了性能优化。英特尔® IPP 能够在一个函数调用中快速实现离散余弦变换（DCT），在提升运算效率的同时，极大精简了书写代码量。如图 7 所示，通过使用英特尔® IPP 库函数、优化算法以最小化内存占用、英特尔® AVX-512 指令集优化等方式，金山云能够在 DCT 计算等方面实现 3 倍左右的性能提升 [7]。

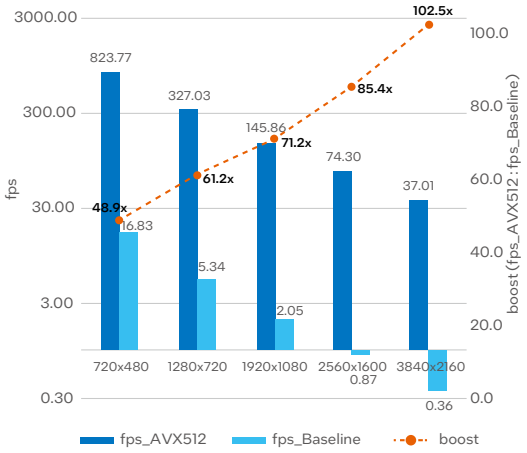


图 6 英特尔® AVX-512 指令集使用前后 ERJND 性能对比
（越高越好，英特尔® 至强® 铂金 8358P 处理器 @32 instances）

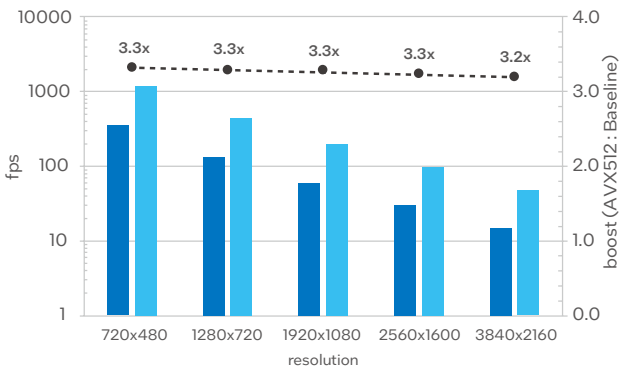


图 7 优化前后 DCT 性能对比
（越高越好，英特尔® 至强® 铂金 8358P 处理器 @32 instances）

此外，金山云还依托英特尔® 至强® 铂金 8358P 处理器集成的英特尔® 深度学习加速（英特尔® DL Boost）技术，将 “集智高清” 深度学习应用中数值精度为 FP32 的模型转换成为使用 VNNI 指令集进行优化的 INT8 数值精度。采用 INT8 等较低精度的数值可以更好地使用高速缓存，增加内存数据传输效率，减少带宽瓶颈，从而能够更为充分地利用计算和存储资源，并降低系统功率。这意味着，在同样资源的支持下，INT8 可为深度学习的推理带来更多的每秒操作数（Operations Per Second，OPS）。通过该方式，金山云能够在精度符合需求的前提下，大幅提升深度学习的性能。

借助上述英特尔软硬件产品与技术，金山云基于集智高清的新方案获得了以下显著增益：

- **大幅节省带宽：** 视频画质增强以更小的视频文件实现更清晰画质和更低码率，可以帮助视频服务提供商极大地降低带宽成本，为更广泛的用户提供高质量的视频服务；
- **视频体验更优：** 视频在编码前进行前处理工作，采用机器学习加图像算法对视频中出现的模糊、噪点、色块等问题进行处理修复；可明显提升视频的主观效果，处理后的视频 MOS 和 VMAF 可远高于原视频；
- **画面品质提升：** 通过 ROI 区域检测，可将视频内容画质增强处理的更加精细化，先将每帧的视频内容分层，再将视频画面的主体和背景根据人眼视觉特性做差异化处理，使得主体更加突出，背景更加纯净。

[6、7] 有关更多案例性能详情，请参阅 https://www.intel.cn/content/www/cn/zh/cloud-computing/kingsoft-cloud-smart-high-definition-transmission.html

英特尔助力腾讯云提供创新、高效的视频云直播平台解决方案

直播业务的快速发展和用户数量的不断增长，正推动腾讯视频云以更先进的平台架构、更强劲的技术支撑能力以及更新颖的产品服务来解决其在发展中遇到的挑战。

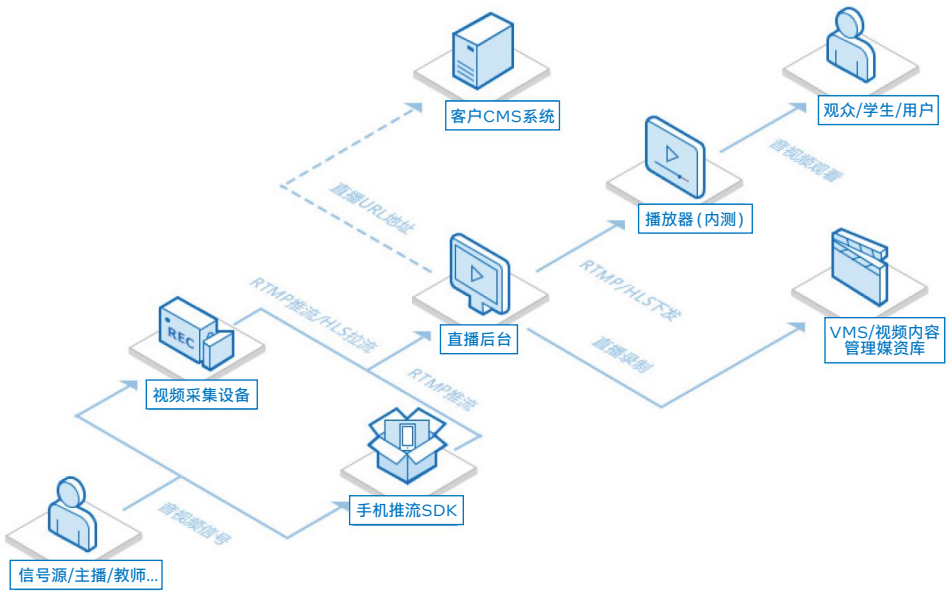


图 8 腾讯视频云直播方案架构图

不同的用户需求对平台的能力有着不同的需求，以游戏直播为例，像《王者荣耀》这样的重磅游戏，在重要赛事直播时观众数量往往会出现浪涌现象，这对直播平台的高并发、低延迟能力提出较高要求。为此，腾讯视频云基于不同用户需求，设计了高效且完整的音/视频采集、编码、封装、转码和分发流程，对各个流程环节都实施了优化和革新，来保证高质量的直播过程。

这些优化革新过程离不开底层IT基础设施提供的有力支撑。为给视频直播平台配备强劲有力的“性能发动机”，腾讯视频云引入性能卓越的英特尔®至强®可扩展处理器等产品，在提升视频直播用户体验上发挥了关键作用。

以直播后台为例，腾讯视频云借助高效能的英特尔®至强®可扩展处理器带来的强大计算性能，保证了转码过程的高效，而转码性能是视频直播平台的核心能力。H.265/HEVC作为新一代的视频编解码标准，在高压缩比下依然能确保高清晰的视频质量，非常适合在移动互联网环境下使用，也因此受到各大视频直播平台的青睐。目前，包括腾讯视频云在内的很多云媒体服务提供商都在开发和部署基于H.265/HEVC的编解码器，但因其算法和数据结构的复杂性，使其编解码的复杂度4倍于上一代H.264编解码器，因此对执行转码工作的处理器的性能要求，就显得格外突出。

为了提升平台转码效率，腾讯视频云一方面部署了万台量级的转码集群，另一方面引入了英特尔®至强®可扩展处理器，为包括转码在内的视频直播流程提供性能支撑。同时，英特尔®至强®可扩展处理器内置的英特尔®AVX-512则进一步强化了它的表现。

强大的转码性能保证了腾讯视频云可以让视频在H.264、wma、rmvb、avi等格式间快速转换，并通过灵活、简洁的播放器，帮助游戏直播运营者通过移动应用、网页等各种展示方式迅速且全方位地覆盖观众。

高性能的游戏直播平台配合腾讯视频云创新的低延时 P2P 直播方案，使游戏直播方案在用户侧的部署中，各项业务指标均令人满意。以最常见的 Flash 直播形式来说，视频卡顿问题与传统直播平台相比有大幅下降，卡顿率可低至 2%，而在移动端，播放启动时延仅为 400 毫秒，视频卡顿率也低至 5%，同时服务带宽峰值提升至 2.3Tbps。[8]

除在性能上保证平台高效运转，腾讯视频云还和英特尔一起，针对视频直播行业不断涌现的互动、趣味、安全、防盗链等需求，推出创新方案供用户选择，并根据用户的特定需求，为其量身定制方案。

例如，在直播过程中对直播效果进行实时监控，不断优化用户接入和收视效果是视频直播平台实现良好服务体验的重要手段。腾讯视频云自研 GSLB 调度体系，利用遍布全国的监测点，定时访问监测文件，可对来自全国各地域、各运营商的用户访问状态进行监控、分析，进而对风险节点及时进行切换和恢复，提升系统稳定性，同时结合腾讯云大禹分布式攻击防御系统，还能帮助用户有效抵御 DDoS 攻击、CC 攻击以及 HTTP DNS 域名劫持等恶意攻击行为。

另一方面，腾讯视频云还利用先进技术手段，例如通过智能视频分析结合智能人脸识别来推出趣味视频方案，这一吸引眼球的功能背后，蕴含着人脸识别、人脸追踪、区域拟合、智能 P 图等一系列人工智能技术，英特尔® 至强® 可扩展处理器所配备的英特尔® AVX-512 为这些 AI 应用所需的并行计算能力提供了更优的支持，这让趣味视频方案在运行时毫无迟滞感，让最终用户真正体验到了在顷刻间 “改头换面” 。

多种类型的 AI 技术已悄然在视频直播领域投入实用，但要使其展现更佳效果，离不开处理器对 AI 所需的并行计算能力的良好支持，英特尔® 至强® 可扩展处理器内置的英特尔® AVX-512 正在这些应用中尽显其所能 。

[8] 有关更多案例性能详情，请参阅 https://www.intel.cn/content/www/cn/zh/cloud-computing/cloud-service-provider-resources-intel-offer-tencent-cloud-innovative-and-efficient-live-video-cloud-platform-solutions.html

视频点播场景

作为互联网时代面向大众娱乐的重要应用之一，视频点播（Video on Demand, VOD）可根据观众的需要播放相应的视频节目，在用户点击或选择视频后，将对应的内容传送至所请求的客户端，其互联网流量和市场关注都处于 “C 位”。有数据表明，在全球 TOP 100 的娱乐应用程序中，VOD 具有极强吸金力，市场贡献高达 38%[9]。

图 9 典型视频点播场景

近年来，VOD 技术已经在更多行业有了全新的落地应用，其不仅可为终端用户提供多样化的媒体信息流，也在广告及娱乐领域有了更多的发展。在这一过程中，其客户需求和业务运营模式也在悄然发生变化，具体包括：

- **更大更高清的观影体验：**许多新增 VOD 用户是从线下转为线上，对大屏、高清视听体验有着较高要求，同时还需 VOD 提供一定的社交功能；
- **精准的视频内容推荐：**随着更多视频节目迁入 VOD 应用，如何让观众在海量内容中寻获满意的视频，是 VOD 运营商提升竞争力的重要砝码；
- **逐渐放大的多平台需求：**更多 VOD 观众不再局限于家中电视机观看视频，而是希望利用碎片时间在不同平台，包括手机、平板或者 Web 页面上进行收视。

这些变化也推动着 VOD 系统提供商寻求引入更灵活多变的方案设计，采用更多先进产品与技术，来应对观众的喜好变化与市场的发展趋势。

[9] 数据援引自 Sensor Tower：2021 年娱乐应用市场洞察报告：http://www.199it.com/archives/1375107.html

点播场景变化	技术需求
大屏设备与高清视频内容的普及，令 VOD 系统面临高清视频编解码和播放双重压力。	解决播放 1080P、2K 乃至 4K 高清视频时出现的卡顿、时延等问题。
面对 VOD 平台上数以万计的视频内容，观众希望在搜索和选择上消耗更少时间。	平台可以根据精准推荐算法，快速精准地向观众推荐内容。
不同平台在提供 VOD 服务时，需要根据使用场景，提供相应的视频格式。	平台需要具有高效的内容分发网络（CDN）和视频转码能力。

适应市场高速变化的需求，英特尔正以一系列功能丰富且性能强劲的产品与技术为以上 VOD 新场景提供支持，例如：

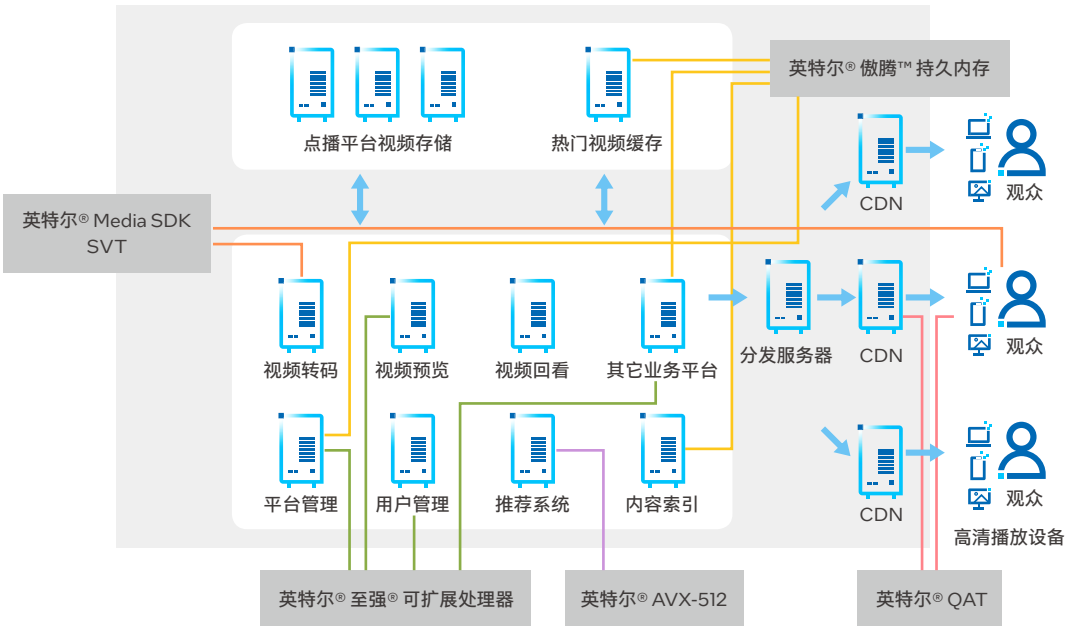


图 10 变革中的 VOD 场景

- **英特尔® 至强® 可扩展处理器：**新一代处理器平台拥有更多内核数量、更强内核性能，以及对大内存和 I/O 带宽的支持，能为 VOD 场景提供强劲的算力支撑；
- **英特尔® AVX-512：**英特尔® 至强® 可扩展处理器内置的英特尔® AVX-512 可以提供高效的视频转码能力，同时提升 VOD 平台各 AI 场景的性能，例如为用户提供更为精准的内容推荐；
- **英特尔® Media SDK：**可为 VOD 场景中的视频播放、编解码、转码和媒体格式转换提供显著的性能增强；
- **英特尔® 傲腾™ 持久内存：**通过创新内存技术，为 VOD 系统中不可或缺的 CDN 网络提供高品质数据存储、读写性能，同时也能为高性能推荐系统提供硬件基座；
- **英特尔® QAT：**作为英特尔推出的数据加解密、压缩和安全技术，英特尔® QAT 能有效助力 VOD 平台增强各环节的数据安全和传输性能。

目前，英特尔产品与技术正助力广大 VOD 系统提供商为用户提供更加流畅和高清的视频点播服务。在下文案例里，就将围绕英特尔产品及技术在金山云 CDN 平台中的应用以及客户获益展开具体介绍。

金山云采用英特尔® QAT 加速 CDN 业务 显著提升性能

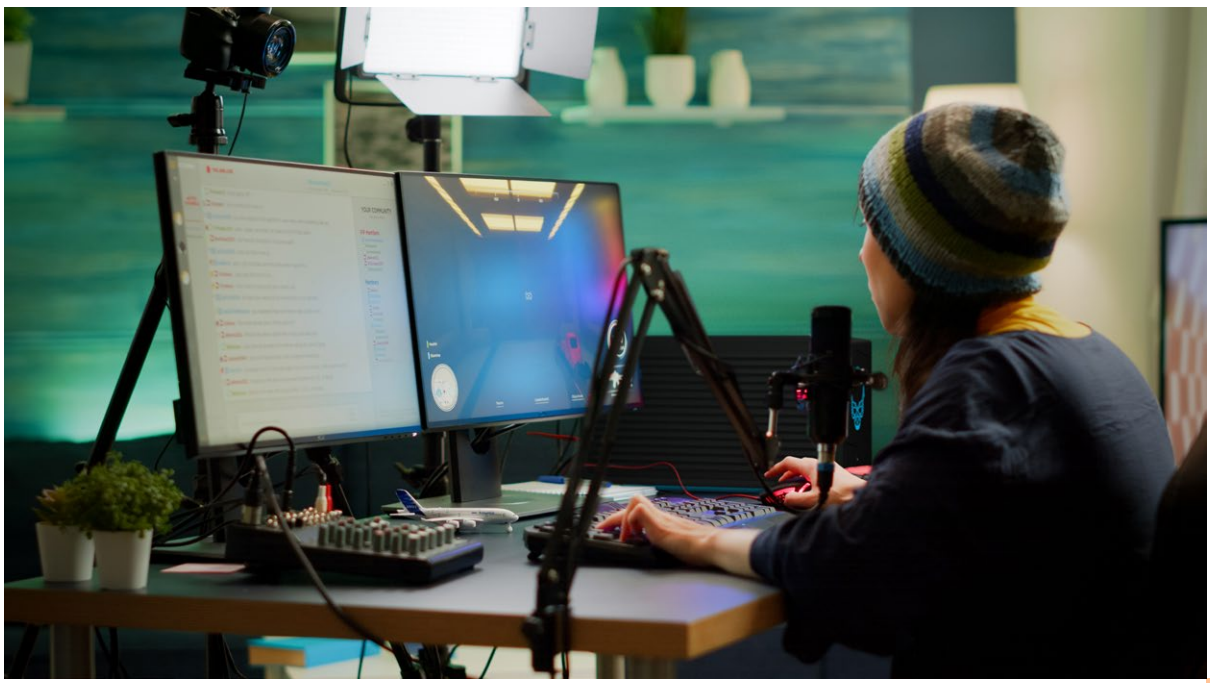
金山云 CDN（KCDN）是由分布在不同区域的边缘节点服务器集群组成的分布式网络。它能够将用户内容分发到边缘节点，从而有效解决互联网网络拥塞问题，提高用户访问网站的响应速度与网站的可用性。KCDN 包含下载类加速服务（支持页面加速、点播加速、下载加速等服务）和直播类加速服务（支持事件直播、社交直播、手游直播、秀场直播等服务）。

由于网络安全重要性日趋凸显，加之网络安全环境渐趋复杂化，HTTPS（Hyper Text Transfer Protocol over Secure Socket Layer）、HTTP2.0、QUIC（Quick UDP Internet Connection）等使用安全传输层（Transport Layer Security，TLS）协议的请求在Web服务中占比越来越高。TLS协议利用非对称加密算法实现身份认证和密钥协商，以及对称加密算法采用协商的密钥对数据加密，和基于散列函数验证信息的完整性，使连接安全性具备了私有、可靠两大特点。

在网络攻击技术不断变化的背景下，互联网服务提供商倾向采用更高级的加解密方法，但这在显著提升了破解难度的同时，也会导致服务器端的加解密计算量大增。对此，金山云表示，在当前HTTPS、QUIC相关业务中，超过 50%的业务是密文传输，需要进行大量的对称加解密与非对称加解密计算，这将消耗大量的处理器资源，特别是非对称加解密对于处理器资源的消耗更为巨大。[10]

为化解面向HTTPS、HTTP2.0、QUIC的CDN业务中大量加解密以及压缩计算带来的性能瓶颈，金山云使用了英特尔® QAT来进行加速。

英特尔® QAT是英特尔针对网络安全和数据存储推出的硬件加速技术。在本方案采用的英特尔® QAT加速卡设备中，英特尔® QAT针对Nginx进行了专门适配，使其可以用异步方式调用加速卡。Nginx是一个高性能的HTTP和反向代理Web服务器，同时也提供了IMAP/POP3/SMTP服务，通过启用异步模式，Nginx能够通过并行处理减少等待，在消耗很少系统资源的前提下达到所需的性能，进而缩短应用响应时间。



[10] 有关更多案例性能详情，请参阅 https://www.intel.cn/content/www/cn/zh/cloud-computing/kingsoft-cloud-adopts-intel-qat-cdn.html

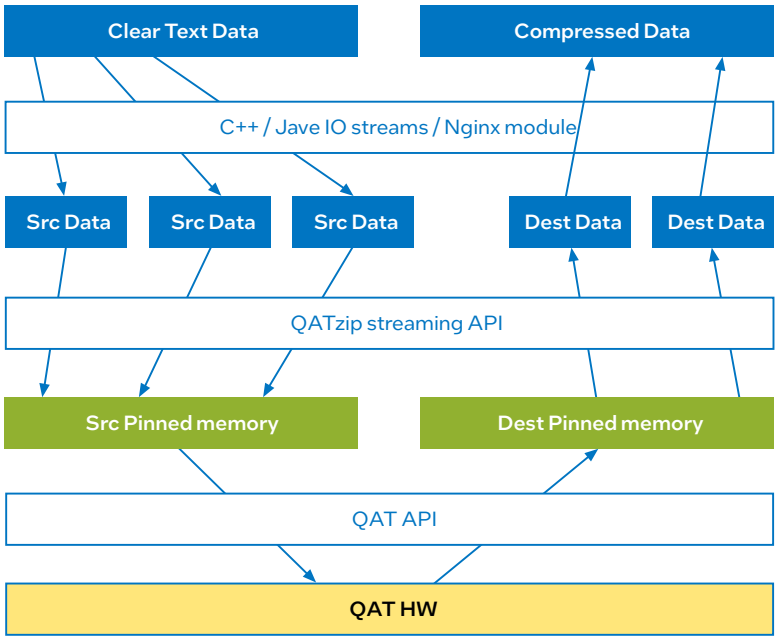


图 11 英特尔® QAT 设备技术架构

英特尔® QAT 还具备强大的压缩加速能力，如图 11 所示，其提供了由 QAT 设备提供加速能力的同步压缩 API，支持无状态并发压缩和解压模式、基于 QAT 异步 API 的流处理模式、线程安全压缩 API，以及基于物理连续地址内存的零拷贝模式，能够将多个小数据压缩和解压请求整合到一个 QAT 硬件请求中，来降低处理器使用率和提高吞吐量。

金山云的CDN业务基于Nginx环境而构建，能够与英特尔® QAT加速卡实现极佳适配。金山云在该业务中以异步模式调用英特尔® QAT加速卡，从而降低了线程间上下文切换的开销，减少负载转移成本，显著提升性能表现。

通过引入英特尔® QAT加速卡，金山云得以将业务中的非对称和对称加解密计算从CPU卸载到硬件加速卡，从而大大降低了CPU的压力。具体而言，英特尔® QAT加速卡实现了如下效果[11]：

- **性能显著提升：**在测试中，金山云发现，由于英特尔® QAT从硬件层面提供了加解密计算卸载方案，不仅CPU负载降低，而且系统的整体性能也得到了提升。此外，相较于使用CPU运行的软件方案，引入英特尔® QAT加速卡还使得系统单节点的每秒查询率（QPS）显著提升；
- **TCO获得有效控制：**由于在现有IT架构的基础上实现了单节点处理性能的提升，金山云得以在不增加服务器集群、不对现有系统环境进行颠覆式修改的前提下，满足更多CDN业务对于加解密性能的要求， TCO也得到了有效控制；
- **支撑更多业务：**英特尔® QAT技术不仅可以应用于HTTPS、HTTP2.0、QUIC等应用的加解密场景，还能够满足更广泛的性能提升需求，有助于金山云进一步提升性能，为用户提供更具成本竞争力的服务。

[11] 有关更多案例详情，请参阅：https://www.intel.cn/content/www/cn/zh/cloud-computing/kingsoft-cloud-adopts-intel-qat-cdn.html

云游戏场景

“云游戏”可谓是近年来各大厂商追逐的风口，与各类传统游戏形式相比，云游戏（Cloud Gaming）有着免安装、硬件配置要求低、随开随玩等优势，玩家可将自己的手机、电视或者是电脑连接至服务器，无需下载游戏，所操作的一切指令都会实时通过网络传输到服务器中，并由服务器运算后进行实时反馈，从而使玩家获得流畅的游戏体验。随着 5G、云计算以及移动互联网等技术的高速发展，云游戏在运行基础上逐渐摆脱束缚而迎来快速发展期，成为数字娱乐产业的重要一环。预测数据指出，2022 年中国云游戏市场收入将增至 79.2 亿元，同比增长达 95.1%。[12]

图 12 典型云游戏场景

高速发展的市场规模与不断丰富的游戏品类，也正对云游戏的承载平台提出更多挑战，具体包括：

- **高品质游戏体验需求：**激烈的竞争使游戏高画质和低延时响应的需求水涨船高，云端游戏平台需要对游戏数据、游戏画面开展更高效的计算处理、音视频编解码及图像渲染，并通过网络快速反馈到玩家的终端；
- **高并发用户接入压力：**优秀的游戏大作总会吸引成千上万玩家同时接入。当海量玩家涌入时，迅速增加的游戏交互和数据处理需求，对云游戏平台的并行接入能力提出巨大挑战；
- **大规模平台部署成本：**成熟游戏项目的商业化落地，不仅需要足够的数据算力及图形视频处理能力予以支撑，也要求每路游戏的成本能获得精细化控制；而传统 GPU 等产品的价格昂贵，给云游戏厂商带来很大的 TCO 压力。

[12] 数据援引自中国信通院发布的《全球云游戏产业深度观察及趋势研判 2022 年》：https://www.sohu.com/a/531448976_120189950

为应对这些挑战，云游戏厂商在构建云游戏平台时，也力求寻找更高性能表现和更有性价比的产品与技术方案。与传统游戏后台系统相比，云游戏平台场景的变化以及所需的技术包括：

云游戏场景的变化	新技术需求
大规模音视频流和游戏数据基于云平台处理并使用网络高速传输。	▪ 云游戏平台需要配备强大算力和图形处理能力； ▪ 平台能够基于 5G MEC（Mobile Edge Computing）等前沿技术方案部署。
云平台需要并行接入大量玩家的处理请求，任何延迟都会带来游戏体验的下降。	▪ 云游戏平台需能根据玩家数量弹性地扩展处理能力。
海量基础设备部署的需求，使云游戏平台对 TCO 控制更为敏感。	▪ 云游戏平台需要选择更具性价比的硬件产品来搭建基础能力平台。

面对全新云游戏平台解决方案中对于基础设施的需求，英特尔正为之提供一系列具有针对性的产品与技术方案，具体包括：

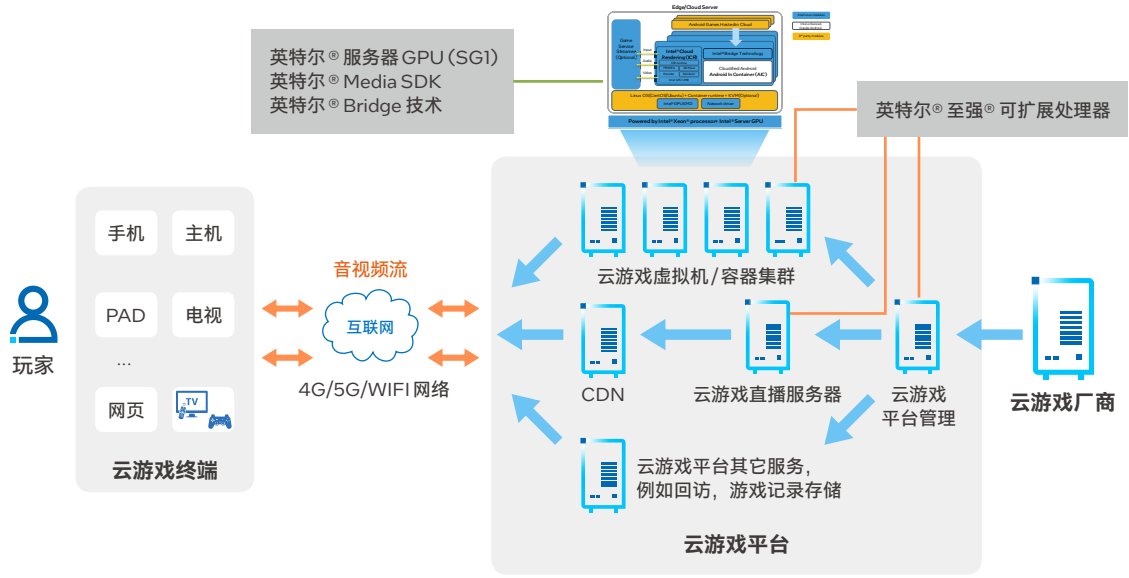


图 13 变革中的云游戏场景

- **英特尔® 至强® 可扩展处理器：**大幅提升的基础性能与多种内置先进特性，为云游戏平台提供更佳算力支撑，并提供强有力的并行处理能力；
- **英特尔® 服务器 GPU：**为云游戏平台所需的音视频提供强有力的编解码和渲染能力，并具有良好的可扩展性；
- **英特尔® Media SDK：**可为云游戏的视频流编解码转换提供性能增强；
- **英特尔® Bridge 技术（Houdini）：**作为一种可集成于安卓容器（Android in Container，AIC）的运行时后期编译器（Run-time Post Compiler），其能让一些并非使用 Java 环境开发或编译的安卓应用也能在基于英特尔® 架构的设备上以“原生应用”形态运行，助力云游戏产品有效扩展应用生态。

英特尔先进软硬件产品已在众多云游戏真实场景中取得了良好的实践反馈。在下文案例中，将围绕英特尔产品与技术在 OPPO 云游戏平台中的应用实践展开详细介绍。

OPPO 基于英特尔® 服务器 GPU 打造云游戏平台 全面提升游戏体验

结合实际业务和基础架构，OPPO 推出了搭载英特尔® 服务器 GPU 、英特尔® 至强® 可扩展处理器，以及英特尔® Android Cloud Gaming Software Stack（ACGSS）软件栈的 OPPO 云游戏平台，将大部分的数据运算和画面处理工作从本地转到了云端服务器。云端将游戏内容编码为视频流，通过高速网络将游戏流画面快速反馈给玩家，玩家的操作也会实时和云端回传。云游戏平台的出现，使更多游戏体验不再局限于终端设备，玩家可以在 PC、平板、手机、电视等设备上玩游戏，实现跨平台的无缝连接，一触即达、即点即玩。

图 14 OPPO 云游戏平台

为了向玩家提供更加卓越的云游戏体验，新的云游戏平台在架构设计上依托于强大的云数据中心服务器集群，打造了庞大、弹性的游戏容器池，能够在云端完成游戏的渲染、编码，并通过容器云的方式交付不同的安卓游戏实例。

在核心基础设施能力建设上，OPPO 云游戏平台包含具有出色计算、存储和网络性能的服务器。该服务器配备了英特尔® 服务器 GPU、第二代英特尔® 至强® 可扩展处理器，以及能够出色运行专门针对英特尔® 架构优化的英特尔® 云游戏参考软件。

其中，英特尔® 服务器 GPU 是基于全新英特尔® Xe 架构的第一款数据中心独立显卡处理单元，专为加速各种视觉云工作负载的渲染和媒体处理而构建。英特尔® 服务器 GPU 基于 23 W 独立片上系统（SoC）设计，具有 96 个独立执行单元、128 位宽的流水线和 8 GB 的专用低功耗 DDR4 内存。英特尔® 服务器 GPU 低廉的每流成本，有助于以更少的基础架构为更多玩家带来安卓游戏和媒体直播，从而降低 TCO。[13]

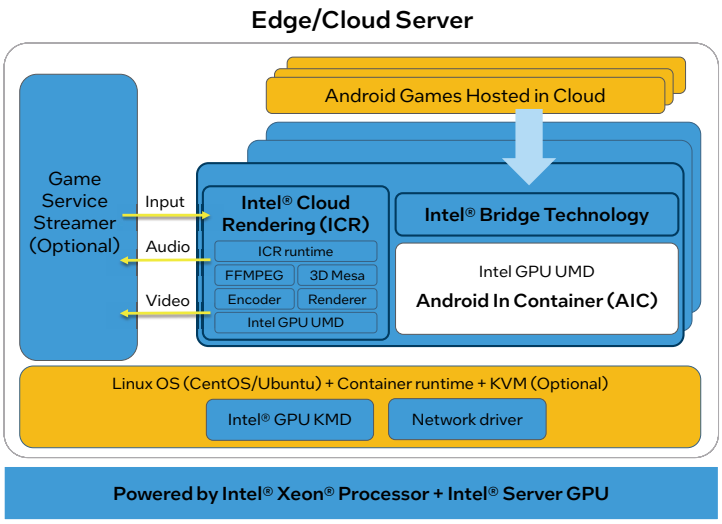


图 15 基于英特尔® 服务器 GPU 的云游戏平台架构

OPPO 云游戏平台的服务器中使用了包含四个英特尔® 服务器 GPU 核心的 H3C XG310 PCIe 显卡，在一个典型的双显卡系统中可并发支持 100 个以上的用户[14]。该服务器还搭载了第二代英特尔® 至强® 可扩展处理器，提供了强大的通用算力。OPPO 还计划在未来将服务器处理器升级到第三代英特尔® 至强® 可扩展处理器，以进一步提升服务器算力。

[13] 如欲了解更多详情，请参阅：https://www.intel.cn/content/www/cn/zh/architecture-and-technology/oppo-gpu-create-cloud-game-platform.html

[14] 如欲了解更多详情，请参阅：https://www.intel.cn/content/www/cn/zh/benchmarks/server/graphics/intelservergpu.html

在硬件之外，OPPO 云游戏平台还使用了英特尔® ACGSS 软件栈。该软件栈包含驱动程序、API 和开发人员工具等内容，如英特尔® ICR 运行时、英特尔® Media SDK、FFmpeg 插件和 Mesa 3D 图形库等，能够有效提升服务器游戏直播密度、游戏性能以及可运行的游戏种类。

其中，英特尔® ICR 运行时提供了核心云渲染技术，并利用英特尔优化的 Mesa 3D 图形库来优化英特尔® 服务器 GPU 利用率。云游戏软件堆栈可以利用英特尔® Bridge 技术，使某些非 Java 编写或编译的安卓应用程序能够在这些设备上运行。该软件与基于英特尔® 架构处理器的服务器配置相结合，为部署安卓游戏服务奠定了坚实的基础。

在虚拟化管理方面，系统基于英特尔提供的英特尔® Bridge 技术，实现了基于容器的安卓虚拟化功能，并以软件开发工具包（SDK）的方式对外提供云游戏服务。一方面，与传统虚拟化方式相比，容器对处理器、内存的利用率更高，能帮助游戏运营商有效地降低云游戏的硬件部署成本；另一方面，SDK 的方式也让玩家接入云游戏时变得更为便捷。

在此基础上，OPPO 结合云游戏的实际业务和基础架构情况，使用自研的 ORTC 对云游戏平台进行了串流优化。终端设备中可以通过采样、去重编码等形式实现游戏指令流的优化，并将优化后的指令流发送给云平台，云平台对指令流进行高效响应与处理，并输出编码的视频，结合云平台的内容分发网络能力，能够进一步缩短云游戏的时延，为用户提供良好的游戏体验。

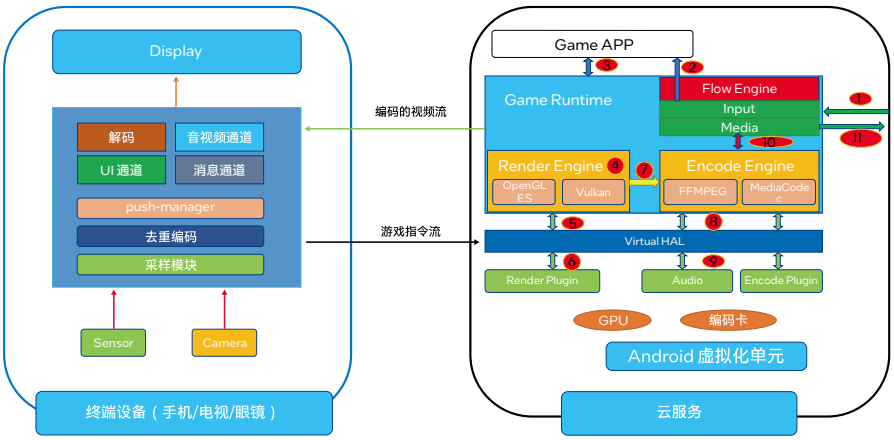


图 16 OPPO 云游戏优化串流优化

OPPO 还在该云游戏平台中集成了 OPPO Cloud OS，提供了容器服务，能够对于游戏容器、在线服务容器、大数据计算容器等进行管理与调度，从而根据实际的玩家量进行弹性的游戏容器池调整，敏捷地满足不同云游戏负载的处理要求，同时实现更低的资源消耗。

卓越的渲染能力、创新的容器化安卓虚拟技术，结合面向 5G 网络的边缘云所提供的高带宽、低时延数据传输能力，再加上 OPPO 云游戏团队在容器云游戏管理、容器调度方案及容器针对游戏的兼容性优化，以及对于 OPPO Cloud OS 的集成，使玩家能够获得更出色的云游戏体验。包括：

- **单卡支持 60 路典型游戏实例** [15]：英特尔® 服务器 GPU 具备强大性能，结合英特尔高度优化的云游戏参考软件，能够实现性能的显著改善。测试数据显示，单块英特尔® 服务器 GPU 可以支持 60 路典型的高负载游戏实例（不同游戏表现有所差异）；
- **实现流畅、稳定的游戏体验：** OPPO 云游戏平台使用门槛低，玩家无需下载游戏安装包，点击对应游戏即可畅玩；此外，由于云游戏的特性，玩家即便使用低端手机，同样也可以享受到高画质游戏大作带来的精彩体验；
- **有效控制 TCO：** 与传统云游戏平台的方案相比，新平台使用了英特尔® 服务器 GPU 方案，具备更高的每瓦性能、更小的数据中心占用空间且能够支持更多的用户，从而有助于降低与设备采购、数据中心运营等有关的资本支出。

[15] 如欲了解更多案例性能详情，请参阅 https://www.intel.cn/content/www/cn/zh/architecture-and-technology/oppo-gpu-create-cloud-game-platform.html

AR/VR 场景

元宇宙（Metaverse）概念风起，使 AR/VR 技术再一次成为人们关注的焦点。作为全新的视听交互模式，AR/VR 在游戏娱乐、工业设计、培训模拟等领域也正获得广泛的应用并持续快速增长。数据表明，2021 年全年全球 AR/VR 头显出货量达 1,123 万台，同比增长 92.1%。[16]

图 17 典型 AR/VR 场景

得益于 AI、5G 以及边缘计算技术等的驱动，AR/VR 也正在更多行业、更多领域的应用中呈现出更多变化与可能性，具体体现在：

- **基于边缘计算的解决方案：**借助 5G 网络和 MEC 节点带来的计算和网络优势，AR/VR 应用可在场景周边就近部署，并在应用效果上获得飞跃；
- **更为精细逼真的虚拟场景：**各类 AR/VR 应用，尤其是游戏应用，为用户提供的交互性、沉浸感更强，整体操控时延已被控制在毫秒级；
- **更深度与行业应用融合：**制造、医疗、电力等更多行业正通过引入 AR/VR 设备，结合智能化识别、检测、通讯和交互能力，有效开展排障、巡检、培训等方面的运用；另外，在红色党建以及教育等领域，通过 AR/VR 技术，也可让用户获得身临其境的体验，领受深刻的历史教育和思想认知。

[16] 数据援引自 IDC 发布的《全球 AR/VR 头显市场季度跟踪报告，2021 年第四季度》，https://baijiahao.baidu.com/s?id=1728720832687680875&wfr=spider&for=pc

而这些也正推动着 AR/VR 技术方案发生变革，让技术创新需求更加广泛，包括：

AR/VR 场景的变化	新技术需求
各类 AR/VR 应用场景在画面品质、操控流畅度和应用范围上的升级换代，对方案的计算处理能力、网络覆盖传输能力以及视觉 AI 能力提出更高要求。	借助 5G 网络具备的高带宽、低延时特性，以及 MEC 技术在算力下沉上的优势，为 AR/VR 技术方案提供强劲边缘计算能力。
由 AR/VR 应用提供的虚拟场景变得更为立体和逼真，用户的交互性、参与感也更强。	后台服务器图形处理能力，包括编解码和渲染能力面临挑战；同时也需要更强的实时 AI 算力。
AR/VR 应用与行业应用绑定更紧，需要形成有效的云网协同能力、快捷的部署推广能力，并对行业效率提升形成强助力。	更强的视频、图像和应用数据处理、存储能力，以及统一的应用开发、托管环境。

适应应用场景快速变化趋势，英特尔密切关注 AR/VR 创新发展，正以先进的软硬协同能力，促进其应对新场景下所需的 IT 基础设施建设，具体包括：

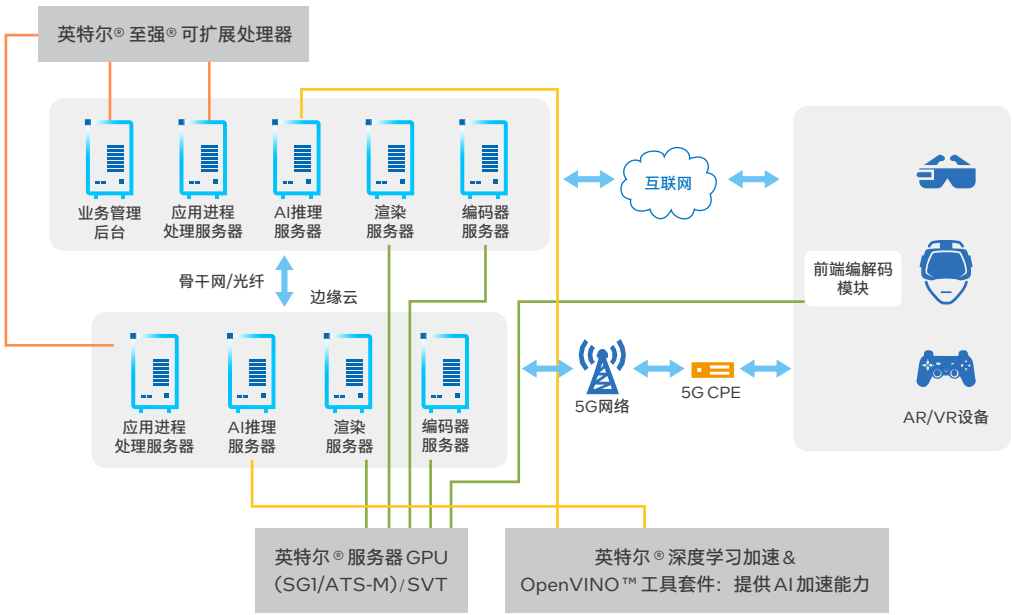


图 18 应对变革中的 AR/VR 场景所需的基础设施架构

- **英特尔® 服务器 GPU：** 为 AR/VR 应用所需的音视频提供强有力的编解码和渲染能力；
- **英特尔® 至强® 可扩展处理器：** 大幅提升的基础性能与多种内置先进特性，为 AR/VR 应用与系统后台提供更佳算力支撑；
- **SVT：** 以灵活的高性能软件编码器核心库，提升 AR/VR 应用中的视频处理效率；
- **英特尔® 深度学习加速 & OpenVINO™ 工具套件：** 为 AR/VR 应用中的深度学习训练和推理过程提供加速。

面对 AR/VR 应用中的不同基础设施需求，英特尔正以功能丰富且性能强劲的产品与技术方案，为之提供广泛支持并在一系列真实场景中取得了良好的实践反馈。在下文案例中，将围绕北京移动与当红齐天联手打造的 5G VR 电竞新体验项目，对英特尔产品与技术在其中的应用实践展开介绍。

携手英特尔与北京移动，当红齐天打造 5G VR 电竞新体验

与传统电竞相比，VR 电竞在真实性上独具特色。随着 VR 电竞获得越来越多关注，选手和观众也在画面品质、操控流畅度、游戏特效、随身装备以及赛场范围等维度上有着更高需求，让承载系统的计算处理能力、网络覆盖传输能力、视觉 AI 能力以及装备面临新挑战。

随着 5G 网络优势的进一步显现，其在高带宽、低延时上的潜能被逐渐释放，大幅优化 VR 电竞网络时延性能，强化了其稳定性和覆盖能力。而 MEC 技术的日益成熟，也让先进英特尔产品与技术所具有的强大计算处理能力，与 AI 加速和 MEC 平台管理能力一起为 VR 电竞场景多方位提供更有力的支撑。

基于多项创新产品与技术，当红齐天集团全新的 5G VR 边缘计算解决方案就采用了“云 - 边 - 端”架构，如图 19 所示，其由中心云、边缘云以及在 5G VR 电竞场景中使用的终端（头显设备 + 手柄等控制设备）组成。其中，中心云主要由内容分发数据中心、游戏同步服务器以及会员主服务器等组建，用于进行 VR 电竞时，支持游戏内容的分发、选手信息的管理以及游戏数据的同步等功能。

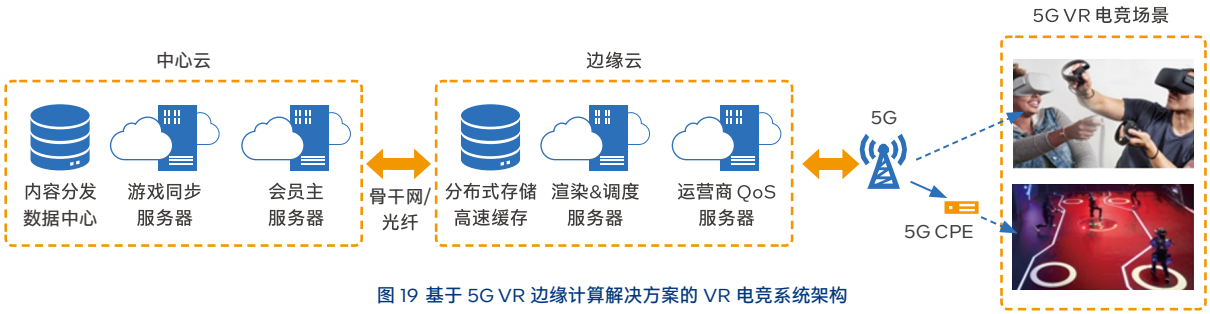


图 19 基于 5G VR 边缘计算解决方案的 VR 电竞系统架构

其中，由分布式边缘服务器集群构建的边缘云是 5G VR 电竞高效运行的核心。基于当红齐天集团自研的交互式云 VR 技术，不同的边缘服务器将分别承载游戏画面渲染、视频串流调度、运营商网络 QoS 管理以及游戏数据的分布式存储和高速缓存等工作负载。而在终端侧 VR 设备的设计上，当红齐天集团融合“一体机边缘化计算 + 全感大空间”创新技术，帮助选手甩掉了沉重的 VR 背包，极大提升了移动性和敏捷性。

当 VR 电竞赛事开始时，如图 20 所示，游戏首先会从中心云同步获取选手的会员数据，然后边缘云中的渲染服务器会通过串流软件服务端，对游戏内容进行渲染并对渲染后的游戏画面进行编码压缩，再借助北京移动提供的 5G 网络覆盖将数据透传到选手的头显端。

通过头显端所部署的串流软件客户端，数据包能被高速接受并解码还原成选手所看见的游戏画面。而当选手的位置、姿态、动作发生变化时，客户端也会实时采集头显、手柄等设备的 6dof 数据（即 6 自由度数据，包括 3 种类型的平移自由度与 3 种类型的旋转自由度）回传给边缘渲染服务器，并在完成游戏内容同步后开始新一轮的渲染和编解码。

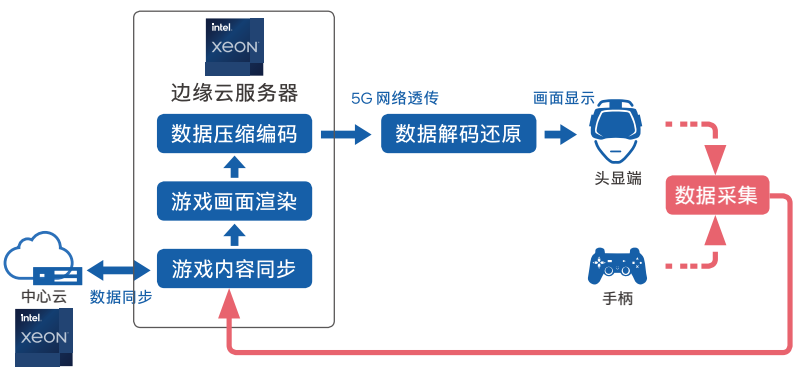


图 20 VR 电竞基本工作流程

为保证 5G VR 电竞方案发挥更佳水准，选手与观众不断提升体验，当红齐天集团敏锐地把握方案对计算处理、AI 加速、MEC 应用管理以及 5G 网络的需求，为方案引入一系列英特尔先进产品与技术。

第三代英特尔® 至强® 可扩展处理器

在新方案中，中心云和边缘云服务器都采用了以第三代英特尔® 至强® 可扩展处理器为核心的灵活可扩展架构，并基于不同VR电竞中的计算密度需求部署为可堆叠的服务器集群。第三代英特尔® 至强® 可扩展处理器为 5G VR 电竞方案带来多方位性能增强，包括：

- 更多的内核、更优化的架构带来处理性能的大幅提升，使方案能更有效应对激烈游戏中的密集计算需求；
- 支持 PCIe Gen4，同时内存带宽也提升为上一代的 1.6 倍[17]，不仅为 VR 游戏各项计算带来所需的高带宽性能，也提升了 5G 客户平面功能（UPF）的性能，使多路选手数据的实时处理更高效；
- 增强的 AI 加速能力（英特尔® 深度学习加速）和安全技术（英特尔® SGX），加强了游戏的智能辅助能力和数据安全性。

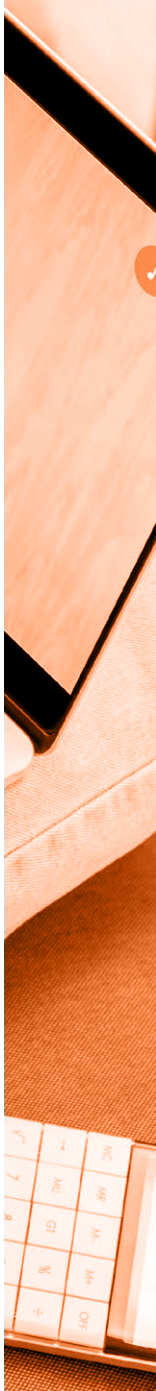
OpenVINO™ 工具套件

当红齐天集团在方案中创新地加入了 AI 辅助裁判，实时对选手进行积分和动作监控，令比赛更为公平公正、更具体育竞技精神，赛事吸引力也得以大幅提升。为了让 AI 能力跟上 VR 电竞的快节奏，新方案引入由英特尔开源的 OpenVINO™ 工具套件，其可通过模型优化器（Model Optimizer）和推理引擎（Inference Engine）两个核心组件，提供 AI 模型优化和推理加速。

英特尔® Smart Edge Open 软件平台[18]

为了帮助当红齐天集团在方案中实现更灵活的 5G MEC 平台VR应用管理，英特尔为之提供了开源的英特尔® Smart Edge Open软件平台。作为一种良好的云网和边缘“黏合剂”，英特尔® Smart Edge Open软件平台能在方案的“云-边-端”架构中打造统一的应用开发、托管环境，提升VR游戏以及其它能力在 5G MEC 平台部署应用时的敏捷性。

技术革新带来了新方案的全面成功。在火热开展的 2021 年VRES大赛中，当红齐天集团与赛事合作伙伴通过 5G VR边缘计算解决方案对赛事方案进行了全面升级。得益于全新第三代英特尔® 至强® 可扩展处理器、OpenVINO™ 工具套件等带来的性能优势，VR游戏所需的高强度计算负载获得了充分满足。通过与 5G MEC 平台相结合，所带来的边缘计算优势使游戏画面更流畅逼真，游戏操作更实时生动，让选手与观众遗忘了真实环境与虚拟世界的边界，充分感受到VR游戏独有的乐趣。同时，借助第三代英特尔® 至强® 可扩展处理器内置的英特尔® SGX，新方案还实现了从边缘到云端的实时数据和应用代码保护，为VR电竞赛事提供了更安全的保障。



[17] 内存带宽比较组为第三代英特尔® 至强® 铂金 8380 处理器：8 通道，3200 MT/s（2 DPC）
与之相比的第二代英特尔® 至强® 铂金 8280 处理器：6 通道，2666 MT/S（2 DPC）

[18] 英特尔® Smart Edge Open 为原 OpenNESS

智能语音场景

得益于 AI 技术的快速发展，各类 AI 和声学模型的构建正使语音与信息之间的高质量转化，包括语音合成、语音转换等得以实现，并推动智能语音技术在语音导航、智能会议系统、智能客服以及智能语音输入与识别等丰富的应用场景中获得广泛的商用化落地。预测表明，中国智能语音市场规模将不断扩大，预计在 2030 年将达到 1,452 亿元 [19]。

图 21 智能语音应用场景

智能语音技术正在更多行业得到应用，为人们的生活带来便捷。而要满足实际应用中对语音交互实时性、准确性的高要求，各类智能语音系统在构建时也面临诸多挑战，包括：

- **更快响应速度：**伴随更多的商用化落地，智能语音应用在实时性上也面临更高要求；例如在智能导航场景中，响应时间需控制在毫秒级，因此系统要在保证质量的同时，有效提升语音处理速度；
- **更高性能表现：**互联网化的应用模式，意味着智能语音系统会面临海量的并行接入；例如在智能客服系统中，要求能稳定可靠地应对大负载的处理需求，从而为用户提供高品质服务；
- **精准转化能力：**在许多复杂且多样化的部署场景，例如多方智能会议系统中，需要智能语音系统能将参会者的信息精准地表达出来，这对相关声学模型的准确率是很大的挑战。

[19] 数据援引自公开媒体报道《德勤报告丨未来的语音世界 - 中国智能语音市场分析》：https://new.qq.com/rain/a/20211228A09UNN00

因此智能语音系统在构建时，需要针对各类应用场景的具体特性引入更多新技术与新产品。与传统语音系统方案相比，智能语音系统应用场景的变化以及所需的技术包括：

智能语音场景的变化	新技术需求
商用化智能语音应用的落地，需要系统为之提供足够敏捷的响应能力。	系统需要具备强劲的算力以及面向声学模型的加速能力。
在一些场景中，智能语音应用需要面临海量并行接入需求。	系统对高密度工作负载具有应对能力，且具有弹性扩展能力。
在企业级应用中，智能语音应用需要能提供足够的信息转化精准度。	系统所采用的声学模型具有足够的精准性。

英特尔一直致力于以科技便捷人们的生活，面对智能语音系统构建中对于基础设施的需求，英特尔以先进的软硬件产品及技术为其提供支持，具体包括：

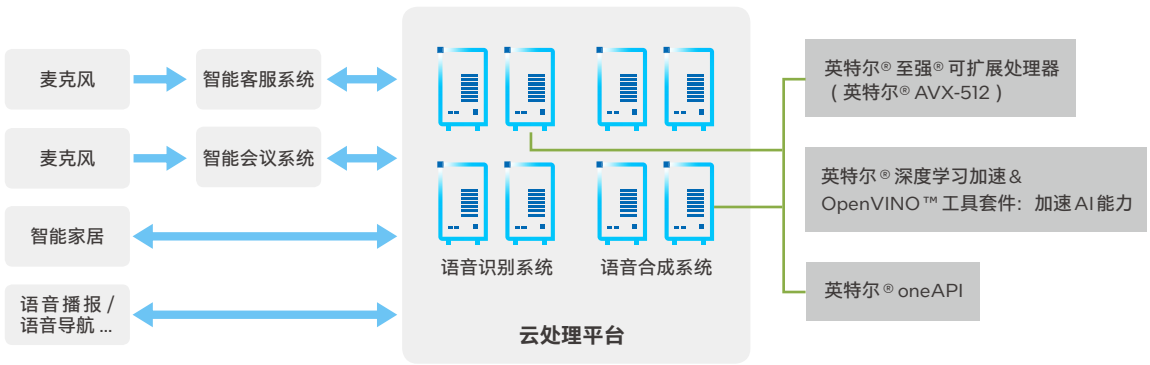


图 22 智能语音系统架构

- **英特尔® 至强® 可扩展处理器（包含英特尔® AVX-512）：** 强劲的基础性能表现可为智能语音系统提供更佳算力支撑；处理器内置的英特尔® AVX-512 指令集能够提供强有力的并行处理能力，支撑高密度计算负载所需；
- **英特尔® 深度学习加速：** 可以为智能语音应用中的模型训练和推理过程提供优化和加速能力，推动商用化落地进程；
- **英特尔® oneAPI 工具套件：** 作为统一编程模型，英特尔® oneAPI 工具套件在面向智能语音应用的开发中提供一系列优化工具和框架，并提供了基于英特尔® 架构硬件基础设施的加速能力。

英特尔提供的系列产品与技术方案已经在众多智能语音场景得到了应用，并在一系列真实场景中取得良好的实践反馈。在下文案例中，将围绕英特尔产品与技术在腾讯云小微智能语音与视频服务接入平台中的应用实践展开详细介绍。

腾讯云小微定制化声码器优化方案，提升实时语音合成性能

腾讯推出的云小微智能语音与视频服务接入平台，通过与全栈语音语义 AI 能力和腾讯云服务的结合，在为用户输出高品质 AI 平台能力的同时，可依托腾讯丰富的产品线和大数据能力，帮助用户获得整合腾讯中台能力的丰富场景应用方案。

为了寻求更高效的语音合成方案，应对传统方案（例如采用 WaveNet 模型的方案）在实时性与吞吐量上遇到的挑战，腾讯与英特尔强强联手，共同构建了定制化 pWaveNet 声码器以及定制化 WaveRNN 声码器这两套语音合成解决方案，将平台性能推向更优。

定制化 Parallel WaveNet 与 WaveRNN 声码器解决方案

pWaveNet 声码器解决方案在传统 WaveNet 模型的基础上，引入了 “概率密度蒸馏” 技术，即用一个提前训练好的 WaveNet 模型作为 “老师”，来指导真正实施生产的“学生”网络进行预测。其不依赖于 “学生” 自身网络任何先前的输出节点，使并行计算成为可能，可以一次性生成整个序列的输出采样点，大幅减少语音合成时间。

但 pWaveNet 模型中的 “学生” 网络依旧是以卷积神经网络为基础的网络架构，虽然规模较小，但卷积操作相较于普通的加减乘除运算要耗费更大的计算量。为此，腾讯在 pWaveNet 模型的基础上进行定制化开发，将网络中一维卷积运算转换为几个通用矩阵相乘的操作，以简化网络拓扑并减少计算量，同时引入 Open-MP 并行机制，充分发挥 pWaveNet 模型中的并行计算优势，使得该定制化模型在不影响语音质量的同时，有效提高了语音合成速度。

除了对语音合成速度的不断追求以外，云小微平台还面对着越来越多设备的接入压力，随之而来的是对整体吞吐量的严苛要求。即在面对大量的实例运算时，单核心所服务的实例数越多越好，而提升单核吞吐量最直接的方法是进一步降低计算量。

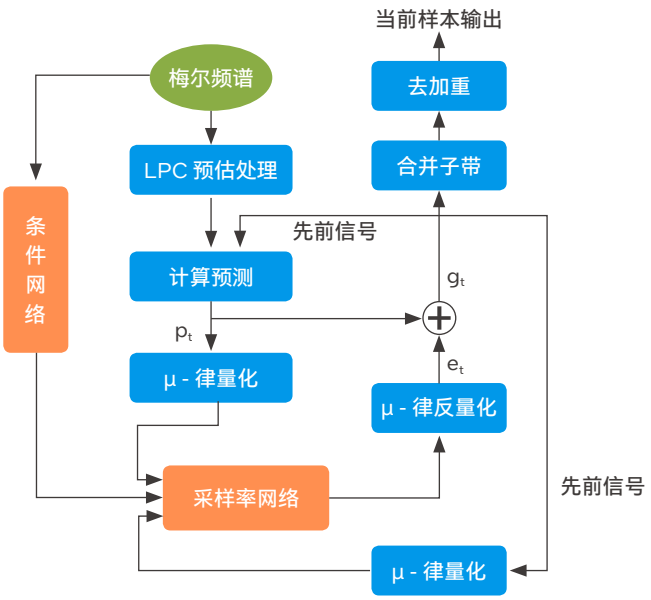


图 23 定制化 WaveRNN 声码器模型架构图

针对这一问题，腾讯在充分利用 WaveRNN 模型优势的基础上，进一步开发定制化 WaveRNN 声码器模型。定制化 WaveRNN 模型的主体部分—采样率网络，是一个具有双 softmax 层的单循环网络。与 WaveRNN 模型相比，其将该网络原始输入中的线性部分分离出来，预先进行了 LPC 预估处理，以大幅降低网络处理难度，并将样本序列划分成多个子带，在前一个子带生成开始不久后即启动下一个子带的计算，有效提高整体计算速度，同时方案还引入了稀疏化技术，减少带宽占用，降低网络整体计算时间，并且在多核环境中，大型稀疏模型能更好地平衡计算力，比小型密集模型性能更好。

英特尔助力语音合成解决方案大幅提升性能

在确定模型架构之后，腾讯选择采用先进的基于英特尔® 架构的硬件作为底层支撑，为整个方案达到更佳性能增光添彩。定制化 pWaveNet 声码器模型与定制化 WaveRNN 模型解决方案都采用了面向四路和八路的第三代英特尔® 至强® 可扩展处理器，该处理器具有高达 28 核的强劲内核，在提升计算力的同时，也很好满足了云小微平台对吞吐量的需求，内置的 BF16 指令集在整个方案中起到了十分关键的作用，可有效提升内存利用率，同时与英特尔® AVX-512 指令一起，在英特尔® oneAPI 深度神经网络库的配合下，加速硬件效率，配合以新一代处理器的超大缓存，能够有效提升处理性能，为语音合成速度的提升做出卓越贡献。

BF16 指令减少内存读写时间

BFloat16 浮点数是一种新型数据格式，由 1 位符号位、8 位指数位与 7 位尾数位组成，相当于在 FP32 浮点数据基础上截断后 16 位尾数位。此种格式同 FP32 浮点数格式具有相同的指数位，即具有近似的动态范围，从而可达到与 FP32 格式相似的模型精度，但由于尾数位的减少，大大降低了计算量并提升了内存存储与读取性能。在以上模型优化方案中采用 BF16 数据格式可达到与 FP32 格式同等的语音质量，却可大大缩短语音合成时间。

英特尔® AVX-512 指令提高执行效率

英特尔® AVX-512 是在处理器上执行 SIMD 运算的指令集，即通过一个处理器控制多个处理微元实现并行操作多段数据以提升处理效率。该指令集将指令宽度扩展到 512 位，使每个时钟周期内可打包更多运算。且该指令集支持三操作 (3-Operand)，即通过创建复杂高级指令来代替多个简单单独指令，增强指令灵活性，减少内存访问数量，从而实现单核执行效率最大化。

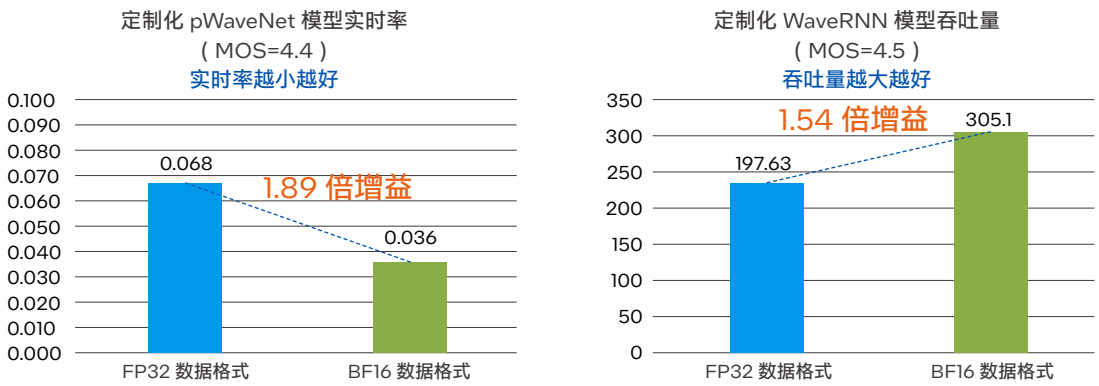


图 24 定制化解决方案性能增益 [20]

超大处理器缓存提升处理性能

介于处理器与内存之间的缓存被用来存储需要经常访问的数据内容。由于处理器的处理速度远远大于内存读写速度，缓存的重要性就在于提供一个比内存更快的临时中转存储，以减少处理器等待数据的时间。当处理器读取数据时，首先从位置更近的缓存中查找，若没查到再去内存中查找，英特尔超大处理器缓存可有效提升缓存命中率，从而提升处理器性能。

第三代英特尔® 至强® 可扩展处理器内置的 BF16 指令和英特尔® AVX-512 指令集帮助腾讯平台的定制化模型有效提升语音合成速度，平台中的定制化 pWaveNet 声码器在 MOS 值为 4.4 的条件下，实现 0.036 的语音合成实时率；而定制化 WaveRNN 声码器也在提升语音合成速度的同时，具备了更强的工作负载处理能力。[21]

[20、21] 如欲了解更多案例性能详情，请查阅：https://www.intel.cn/content/www/cn/zh/now/csp-abm-tencent/optimization-of-customized-vocoder.html

影视制作场景

视频，如今正成为人们生活和学习中非常重要的一部分，从电影创造的幻想世界，到新闻关注的热点时事，再到随处可见的商业广告，视频正深刻地影响着我们的世界。而今天的影视制作已逐渐基于数字化流程展开，无论是前期拍摄，还是包括剪辑、特效和音效合成在内的后期制作，都伴随着对海量数据的传输、处理和存储过程。尤其随着各类特效在影视作品中有了更多的运用，影视制作过程中的数据量也呈几何级数增长。

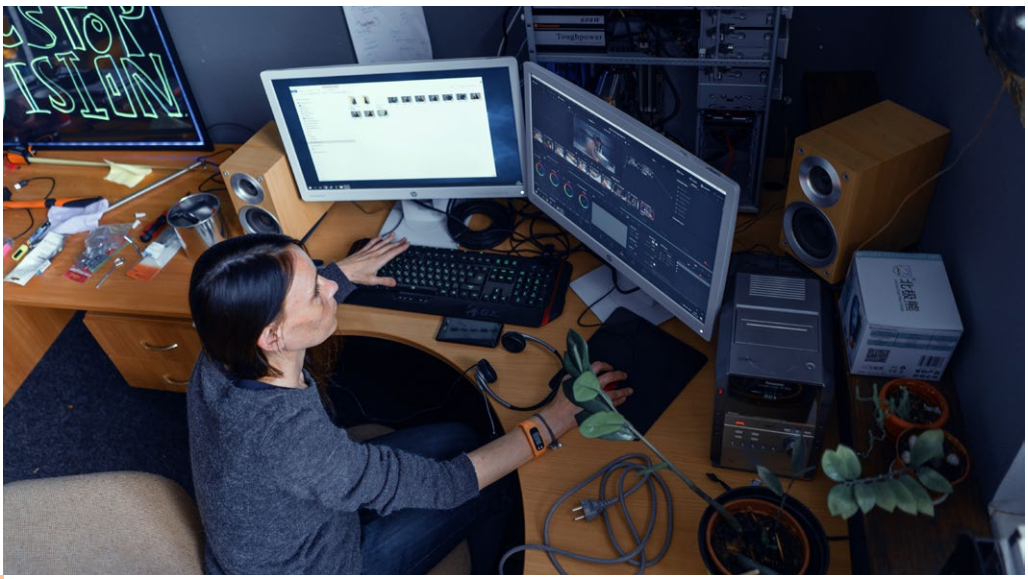


图 25 典型视频后期场景

影视制作数字化变革的推进，也使影视行业在后期制作解决方案的构建上有了更丰富、更细化的需求，包括：

- **大规模算力的部署：**在影视后期制作中，需要使用大量专业软件进行调色、渲染、剪辑、特效制作以及预览；部署足够的算力对于缩短制作时间、提升制作效果至关重要；
- **数据快速读写能力：**在高清模式、或复杂特效中，即便单一镜头都意味着巨大的数据量（数 GB 乃至数十 GB），要快速处理这些数据，数据快速读写能力显然不可或缺；
- **AI 加速能力的引入：**更多影视作品的后期制作已由 AI 应用来担当主角，例如智能剪辑、智能合成及智能配音等，在提升制作效率的同时也大幅降低了制作者的工作压力。

更加细致的需求正推动着影视后期制作解决方案的技术特性发生改变。与传统方案相比，影视制作解决方案技术特性的变化以及所需的 IT 技术包括：

影视制作场景的变化	新技术需求
大量对算力有着高要求的应用软件正参与到影视后期制作的全流程中。	▪ 具有更强算力的多核心、高主频处理器； ▪ 对高密度计算负载的加速能力。
影视后期制作所需处理的数据规模越来越大，实时性要求也越来越高。	▪ 提供大容量的非易失性内存； ▪ 配备更强存储性能的存储设备。
更多 AI 技术被引入影视后期制作过程，场景贯穿剪辑、特效、配音等不同环节。	▪ 系统能对 AI 模型开展有效加速。

面对影视后期制作解决方案中对于基础设施的需求，英特尔正为之提供功能丰富且性能强劲的产品与技术方案，具体包括：

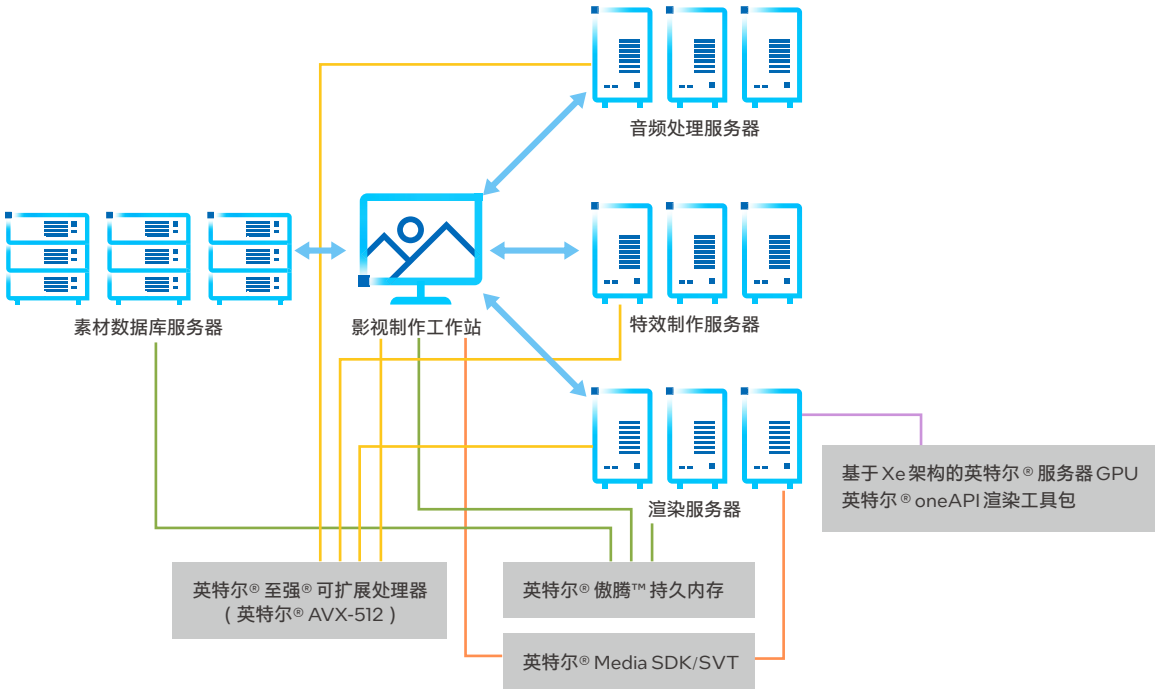


图 26 变革中的影视制作场景

- **英特尔 ® 至强 ® 可扩展处理器（包含英特尔 ® AVX-512）：** 大幅提升的基础性能与多种内置先进特性，为影视后期制作解决方案提供更佳算力支撑，并为高密度计算负载提供加速；
- **英特尔 ® oneAPI 渲染工具包：** 该工具包通过一系列开源的渲染和光线追踪库，能够实现高性能、高保真度的视觉体验，提供可扩展、性价比高的平台，将渲染能力提升至新的台阶，并对包含大型数据集、高度复杂、需内置人工智能的工作负荷进行加速。英特尔 ® oneAPI 渲染工具包可以大显身手的场景包括：数字内容创作、专业渲染、动画制作、计算机辅助设计、建筑工程、科学可视化以及游戏 VR 和 AR 开发等。通过英特尔 XPU 硬件、英特尔 ® 傲腾 ™ 持久内存、网络连接解决方案以及 oneAPI 软件解决方案，内容创作者和开发者可以轻松将创意以更加真实的方式呈现，并能享受到现在以及未来系统和加速器的高性能、高效及灵活性；
- **英特尔 ® 傲腾 ™ 持久内存：** 通过创新内存技术，为数据提供基于大内存、且具有持久性的高品质数据存储、读写性能；
- **英特尔 ® Media SDK：** 可为视频播放、编解码、转码和媒体格式转换提供性能增强；
- **英特尔 ® SVT：** 以灵活的高性能软件编码器核心库提升视频处理效率。

英特尔先进产品及技术已在众多影视制作场景中得到运用，并在一系列真实场景中取得了良好的实践反馈。在下文案例中，将围绕英特尔产品与技术在 Chapeau Studios 公司使用 Memory Machine（一款由 MemVerge 公司出品的大内存软件）进行影视后期制作的应用实践展开详细介绍。

用持久内存，让内容创作远离数据丢失（英特尔 xMemVerge《守护者联盟》）

数据规模的迅速增长在影视制作过程中正变得尤为明显，根据 TechRadar 提供的数据，制作一部像《守护者联盟》这样的动画电影可能需要长达 6,500 万小时的素材，最终才能完成一部 90 分钟的电影成片[22]。

传统的存储设备在应对大规模数据时显然有些力不从心。例如在一些环节中，制作者将场景工程文件加载到工作站通常就需要 10 分钟或更长时间（这具体取决于场景的复杂度以及视觉效果和动画艺术家所使用的不同应用软件）。而如果在加载或保存数据时遇到瓶颈，或者应用崩溃死机，那么花费的时间可能会延长到数小时，从而拖累生产力和进度。

因此，在影视制作过程中部署大量算力和内存设备变得极为重要。但传统DRAM内存具有数据易失性特性且价格昂贵，并难以提供所需的大容量。为此，MemVerge正通过在其Memory Machine软件中引入英特尔® 傲腾™ 持久内存这一全新内存产品来应对上述挑战。

英特尔® 傲腾™ 持久内存为 Memory Machine 软件带来创新优势

作为一款新型大内存软件，Memory Machine 可将传统 DRAM 内存和英特尔® 傲腾™ 持久内存一起虚拟化，从而在不修改应用代码的情况下进行内存访问。与现有应用兼容的 Memory Machine 对于应用而言就像 DRAM 内存一样，可以百分百地利用 DRAM 内存和持久内存。

这就使Chapeau Studios这样的媒体和娱乐公司在运行Maya等资源需求密集型软件应用时能够获得足够的内存容量。同时，通过“内存快照”功能的运用，应用在崩溃时也能保存相关数据。此时制作者能在几秒内就从持久内存中恢复TB级的数据而无需耗费几分钟甚至几小时的时间等待数据从传统存储设备中恢复。

“我们打造了一款出色的大内存软件，可以在无需修改应用的情况下为应用提供软件定义内存服务。‘内存快照’功能会记录任何给定时间内正在进行的事务，并在应用崩溃时恢复数据。各大电影工作室往往有成千上万的艺术家在全世界范围内进行剪辑、灯光、布局、动画和特效方面的工作，因此该软件对电影制作行业尤其有用。”来自MemVerge公司的联合创始人兼首席执行官范承工博士介绍道。

他补充说：“无论是出于剪辑需要还是因为应用崩溃，能让应用回滚到之前的状态都是一个很大的卖点。如果没有大内存容量和快照功能，一旦应用崩溃，所有的工作都会前功尽弃。应用崩溃时要恢复数据真的相当痛苦。”

而英特尔® 傲腾™ 持久内存的使用也非常便捷。其只需和传统的DRAM内存一样插入，就能让更多数据在更靠近处理器的位置进行读写，实现与DRAM内存相近的性能表现。

“人们谈论持久内存已经有 20 年了，”范承工博士谈道：“我第一次听说持久内存的时候，也觉得它很神奇，但我还是持怀疑态度，因为没有人真的发布这样的产品。后来英特尔确实做到了。”

[22] 如欲了解更多详情，请参阅：https://www.techradar.com/news/world-of-tech/inside-dreamworks-how-animated-movies-are-rendered-1127122#:~:text=Most%20films%20take%203%2D5,make%20more%20films%20per%20year

图 27 采用英特尔 ® 傲腾™ 持久内存助力影视创作 [23]

Memory Machine 和英特尔 ® 傲腾™ 持久内存相结合的实践

作为一家有着出色技术且善于协作的视觉效果和设计公司，Chapeau Studios 致力于在传统电影工艺中融入现代技术、用户体验和设计。其创始合伙人兼创意主管 Ben Looram 说：“视觉效果是 Chapeau 的基石。在电影制作的众多领域中，我们以擅长将以假乱真的 CG 动画融入实景场景而闻名，以技术追求效率是我们与生俱来的基因。”

Chapeau Studios 公司的技术经理 Mark Wright 详细介绍了 Memory Machine 和英特尔 ® 傲腾™ 持久内存相结合的优势。他这样说道：“举个例子，假设有位艺术家要去吃午饭，而此时有人要用他的电脑。这时候他就可以用 Memory Machine 为当前的工作拍一张快照，然后去吃午饭。回来的时候，Memory Machine 可以让他直接回到离开前所在的地方而无需再次打开场景工程，而是让 Maya 完完全全重现离开前的进度。你不是‘回到’刚才的工作，而是你就在那里。”

“有了这样的硬件，你能够以一种全新的方式工作，” Wright 补充说：“将英特尔 ® 傲腾™ 持久内存添加到每位艺术家的工作站中，不仅性价比高，而且还为我们能够按时交付工作添加了一份保障。这项技术使得我们对于编辑场景工程以及保存文件有了全新的思考。随着 MemVerge 进一步优化他们的软件以适应 Windows 系统，我想整个行业都会逐渐采用它。”

在技术和视觉效果领域极富远见的 Hank Driskill 这样看待英特尔 ® 傲腾™ 持久内存与 Memory Machine 软件的配合，“场景加载有时需要花费 45 分钟到一个小时，而保存一下可能也需要好几分钟的时间，所以艺术家往往不会经常保存他们的工作文件。一旦系统崩溃，他们就会丢失工作进度，创作的思路和状态也会受到干扰。如果他们丢失了之前的工作， 可能需要半天时间才能恢复，这对艺术家的工作状态和生产进度来说都可能是灾难性的。英特尔 ® 傲腾™ 持久内存的出现则意味着即便系统崩溃，艺术家之前的工作进度也不会消失。”

“在每一部新电影中，艺术家们都在追求精益求精，探索无限可能。他们会快速采纳新技术，然后提高对于他们正在创作的艺术的标准。这意味着借助英特尔 ® 傲腾™ 持久内存和 MemVerge 的 Memory Machine 软件，艺术家能够实现事半功倍。我认为这对他们而言是一件天大的好事。” [24]

[23、24] 如欲了解更多性能和案例详情，请参阅 https://www.intel.cn/content/www/cn/zh/customer-spotlight/cases/memverge-application-crashes-no-starring-role.html

产品篇

硬件

第三代英特尔® 至强® 可扩展处理器

得益于英特尔的持续创新，包括 Cooper Lake 和 Ice Lake 在内的第三代英特尔® 至强® 可扩展处理器已针对多样化的工作负载类型和性能需求，对包括音视频领域在内的高密度计算需求实现了优化，并通过平衡的架构以及多种内置加速和先进的安全功能，帮助用户将各类工作负载安全地放置在从边缘到云的更佳性能位置，展现出多种优势，包括：

图 28 第三代英特尔® 至强® 可扩展处理器

基础性能大幅提升：

- 更多内核数量和更强的内核性能，八路配置下，单平台支持多达 224 个内核；
- 更多内存和 I/O 带宽支持，支持 8 条 DDR4 内存通道（Ice Lake）或 6 条 DDR4 内存通道（Cooper Lake）。

更多 AI 加速与安全功能：

- 加入增强型英特尔® 深度学习加速技术，同时支持 16 位 Brain Floating Point（BF16）和矢量神经网络指令（VNNI）；
- 单路和双路配置的第三代英特尔® 至强® 可扩展处理器（Ice Lake）对英特尔® 软件防护扩展（英特尔® Software Guard Extensions，英特尔® SGX）提供支持。

为多种工作负载提供增强能力：

- 内置英特尔® 高级矢量扩展 512（英特尔® AVX-512），可以加速工作负载和用例的性能，如科学模拟、金融分析、人工智能/深度学习、3D 建模和分析、图像和音频/视频处理、密码学和数据压缩等。借助超宽 512 位矢量运算功能和多达两个 512 位融合乘加（FMA）单元，应用程序在 512 位矢量内的每个时钟周期每秒可打包 32 次双精度和 64 次单精度浮点运算，以及八个 64 位和十六个 32 位整数。因此，与英特尔® 高级矢量扩展 2（英特尔® AVX2）相比，数据寄存器的宽度、数量以及 FMA 单元的宽度都增加了一倍。
- 增强的英特尔® SST（英特尔® Speed Select 技术），可对处理器性能实施精细控制。

英特尔® 服务器 GPU

图 29 英特尔® 服务器 GPU

作为英特尔首款数据中心独立图形显卡产品，英特尔® 服务器 GPU 正在音视频处理领域中扮演越来越重要的角色。这一产品基于英特尔 Xe 微架构构建，配备低功耗独立片上系统（System on Chip，SoC）设计、128 比特管道以及 8GB 专用板载低功耗 DDR4 内存，能从容应对各类具有高密度、低时延处理需求的应用场景。

针对编解码和渲染这两个音视频处理的核心需求，英特尔® 服务器 GPU 具备良好的解决方案。首先在音视频编解码能力上，针对不同的视频格式，英特尔® 服务器 GPU 在编解码场景中都有着良好的性能表现；同时，这一 GPU 产品所具备的强劲渲染能力，能轻松帮助用户实施“云渲染”，从而为云游戏等应用的加速落地奠定基础。

而在可扩展性上，英特尔® 服务器 GPU 可与英特尔® 至强® 可扩展处理器相配合，在相应英特尔软件（如英特尔® Media SDK 和英特尔® oneVPL）的配合下，可助力用户在保持服务器数量不变的情况下，灵活扩展 GPU 显卡的数量和容量。目前，一个典型的双卡系统已经能扩展接入 160 个云游戏的并发用户[25]，有助于用户降低成本（Total Cost of Ownership, TCO）。

英特尔® 傲腾™ 持久内存

英特尔® 傲腾™ 持久内存采用创新的内存技术，将高性价比的大容量内存（单条容量 128 GB、256 GB 和 512 GB）与对数据持久性的支持巧妙结合，具备高耐用性、一致性和低时延等高性能特性，是从云到数据库，再到内存分析、虚拟化基础设施、内容分发网络等数据密集型和计算密集型工作负载所需大规模持久内存的理想选择。

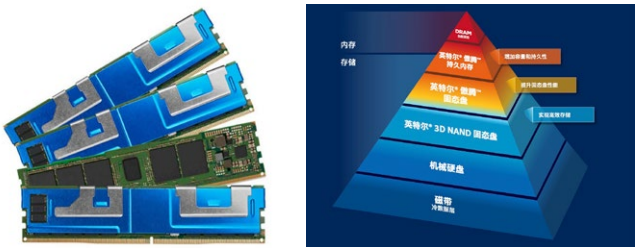


图 30 英特尔® 傲腾™ 持久内存支持分层架构，从而实现高性能、大内存计算

作为具有突破意义的内存技术创新产品，英特尔® 傲腾™ 持久内存通过与第三代英特尔® 至强® 可扩展处理器组合，可创建两层内存和存储分层架构，打造出色的低时延、高带宽、高服务质量（QoS）和耐用性，尤其是新推出的 200 系列较 100 系列带宽平均提升 32%[26]，且总内存每路高达 6 TB，可进一步优化工作负载的性能与成本。

在使用上，英特尔® 傲腾™ 持久内存可通过灵活配置，为用户提供内存模式和 App Direct 模式 2 种不同模式，以应对不同场景需求：

- **内存模式：**无需更改应用即可提供大内存容量，且性能接近 DRAM（具体视工作负载而定）；
- **App Direct 模式：**能够实现大内存容量和数据持久性，在该模式下软件可以将 DRAM 和持久内存作为两个独立的内存池进行访问。支持行业标准持久内存编程模型的软件和应用可以直接与持久内存通信，从而实现低延迟，为处理更大的数据集提供支持。

[25] 如欲了解更多详情，请访问英特尔官网：https://www.intel.cn/content/www/cn/zh/benchmarks/server/graphics/intelservergpu.html
[26] 如欲了解更多详情，请访问英特尔官网：https://www.intel.cn/content/www/cn/zh/customer-spotlight/cases/achieve-insight-data-intel-optane.html

英特尔® 以太网产品

随着各类音视频应用场景中出现越来越大的数据传输需求，如何帮助用户更快地在云端、数据中心、边缘和应用终端之间移动数据正成为人们所关注的问题。为此，英特尔推出一系列以太网产品（包括各个型号的英特尔®以太网控制器、适配器和配件），来满足不同工作负载上所需的网络传输性能以及复杂网络交互所需的灵活性，所支持速度从 1GbE 至 100GbE 不等。

图 31 英特尔®以太网产品

广泛的网络互操作性、可靠性和性能等测试表明，这些产品能以更强的通用性和灵活性能力，来帮助用户加速高优先级应用程序、分组处理延迟敏感型工作负载。为了优化性能，英特尔还在其中的一些产品中加入了诸多改进功能，来满足 NFV 系统、分布式存储、HPC 以及混合云中音视频数据的处理需求，包括：

- **应用程序设备队列（ADQ）：** 是一种先进的流量导向技术，可提高应用程序响应时间的可预测性和可扩展性；
- **动态设备个性化（DDP）：** 可为云、通信和网络边缘负载减少时延，并且提高分组处理的性能和效率；
- **远程直接内存存取（RDMA）：** 支持 iWARP 和 RoCE v2，可提高网络性能，以应对低延迟、高吞吐量的工作负载。

英特尔® FPGA和SoC FPGA

各类音视频处理工作负载正对用户处理能力储备带来巨大挑战，而具有灵活高效、可重复编程特性的 FPGA 产品能满足用户在性能定制、功耗定制、高吞吐量和低延迟上的需求，在音视频领域获得越来越多的青睐和运用。

英特尔推出的不同类型和配置的系列 FPGA 产品，可用以帮助包括音视频领域在内的用户加速其关键工作负载。这些产品包括：

图 32 英特尔®FPGA 产品家族

- **英特尔® Agilex™ FPGA 和 SoC FPGA：** 作为首款基于 10 纳米 SuperFin 工艺技术构建的 FPGA 结构，能为需要灵活性、敏捷性和高性能的应用程序，如边缘节点视频处理提供性能加速；
- **英特尔® Stratix® 系列：** 融合了高密度、高性能和丰富特性，可实现更多功能并提高系统带宽，从而支持客户更快地向市场推出一流的高性能产品，并且降低风险；
- **英特尔® Arria® 系列：** 拥有丰富的内存、逻辑和数字信号处理（DSP）模块特性集，以及高达 25.78 Gbps 收发器的卓越信号完整性，可为终端市场用户提供更佳性能和能效输出；
- **英特尔® MAX® 系列：** 在小尺寸设备中提供了高级处理功能，实现了非易失性集成的革新，并针对各种成本敏感型大体量应用进行了优化；
- **英特尔® Cyclone® 系列：** 可帮助用户提高集成度、提升性能、降低功耗并缩短产品上市时间。

硬件加速

英特尔® QuickAssist 技术

针对不断变化的网络安全和数据存储需求，英特尔推出英特尔® QuickAssist 技术（英特尔® QAT），加快计算密集型操作，增强云、网络、大数据和存储应用中动态数据和静态数据的安全性和压缩性能，助力应用程序和平台提升性能。

英特尔® QAT 为用户提供了以下两个方面的核心能力：

- **压缩&加速：**提供了同步压缩 API、线程安全压缩 API、支持无状态并发压缩/解压模式、基于 QAT 异步 API 的流处理模式，以及基于物理连续地址内存的零拷贝模式等能力。同时能通过将多个小数据压缩/解压请求整合到一个 QAT 硬件请求中，来降低处理器使用率和提高吞吐量；
- **加密&认证：**在网络安全应用方面，英特尔® QAT 支持多种对称数据加密（如 AES）、非对称公钥加密（如 RSA、椭圆曲线等）和数据完整性（SHA1/2/3 等）算法，能够加速数据的加解密和数字签名等操作。

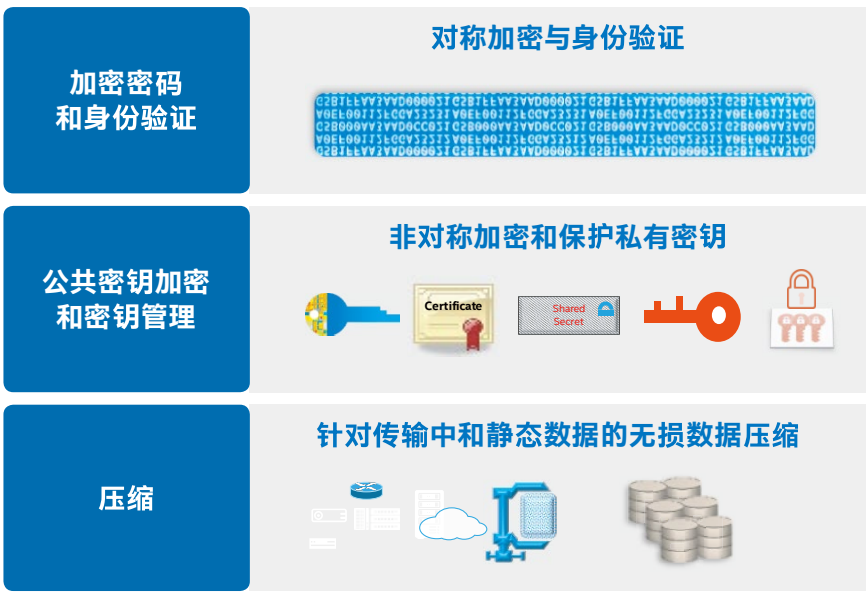


图 33 英特尔® QAT 可提供的核心能力

基于以上能力，英特尔® QAT 可在云服务、网络安全、大数据处理以及高性能存储等不同应用场景中发挥作用，包括：

- **云服务：**通过支持更高性能的安全隧道、更大数量的认证客户端以及更大化处理器利用率，帮助用户提升云服务中的应用程序吞吐量；
- **网络安全：**通过提供更高性能的加密通信，加速对称和非对称加密所需的密集计算，提高网络性能和网络安全水平；
- **大数据处理：**帮助用户减少总体数据大小，降低处理器负载，提升整体性能；
- **高性能存储：**通过实时压缩数据实现高性能存储。

CPU 指令集加速

英特尔® 高级矢量扩展 512

随着计算需求的不断提升，基于单指令单数据流（Single Instruction Single Datastream, SISD）指令的计算方式在许多场景中效率有限。为此，处理器引入新的单指令多数据流（Single Instruction Multiple Data，SIMD）指令来提升效率，这种指令能让一个指令单次操作多条数据，使效率大为提升。

英特尔在 1996 年推出的 MMX 指令集中率先加入了对 SIMD 指令的支持，并在后续不断升级优化；同时英特尔还在指令集中增加了 FMA（融合乘加）指令集，让处理器一次能同时完成加法和乘法两种基本操作。

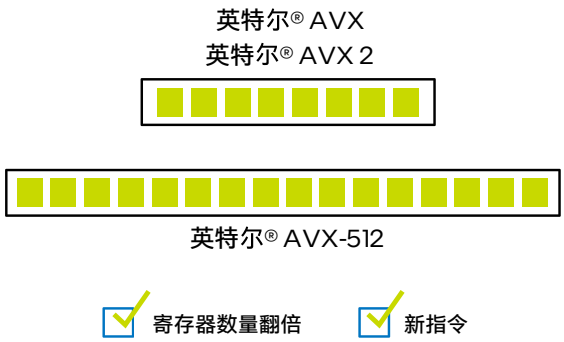


图 34 英特尔® 高级矢量扩展 512

目前，在英特尔® 至强® 可扩展处理器家族中集成的 AVX-512 指令集，寄存器已由最初的 64 位升级到了 512 位，且具备两个 512 位的 FMA 单元。这意味着应用程序可同时执行 32 次双精度、64 次单精度浮点运算，或操作八个 64 位和十六个 32 位整数。

AVX-512 指令集的加入，让英特尔® 至强® 可扩展处理器家族在音视频处理、游戏、科学计算、数据加密压缩以及深度学习等场景中拥有了更加出色的表现。例如，在视频编解码、转码等流程中，应用程序需要执行大规模的整型和浮点计算，而 AVX-512 指令集正可在其中发挥所长。在一些视频服务场景中，集成 AVX-512 指令集的处理器平台的转码性能可获得大幅提升。

英特尔® 深度学习加速

对算力的巨大需求影响着深度学习方法在各音视频场景中的落地。大多数深度学习方法的训练和推理过程采用了 32 位浮点精度（FP32），这种高精度数据格式虽然能带来更精确的结果，但由于系统内存带宽等限制，训练和推理过程往往易陷入内存瓶颈而影响效率。

近年来许多研究和实践已表明，以较低精度的数据格式进行深度学习训练和推理并不会对结果的准确性带来太多影响。而低精度数据格式带来的优势不仅能提升内存的利用效率，同时也可减少处理器资源消耗，并实现更高的处理速度（OPS）和性能。

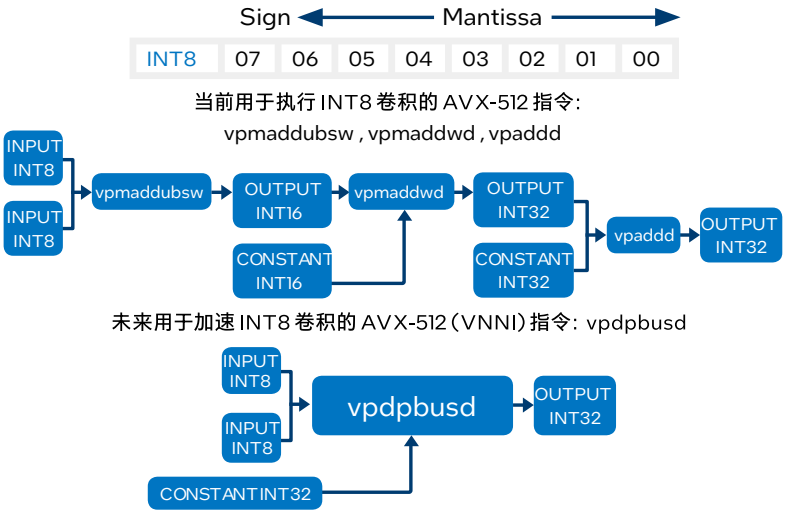


图 35 英特尔® 深度学习加速

英特尔® 深度学习加速（英特尔® DL Boost）的精髓，就是把对低精度数据格式的操作指令融入到了 AVX-512 指令集中，即 AVX-512_VNNI（矢量神经网络指令）和 AVX-512_BF16（bfloat16），分别支持 INT8（主要用于推理过程）和 BF16（兼顾推理和训练过程）。目前，第三代英特尔® 至强® 可扩展处理器家族已集成了英特尔® 深度学习加速的两种 AI 加速指令集，并被广泛运用于各种场景的深度学习训练和推理。

软件

英特尔® oneAPI

作为一种跨行业、开放、基于标准的统一编程模型，英特尔® oneAPI 能提供跨加速器架构的通用开发体验，实现更快的应用程序性能、更高的生产力和更大的创新。这一由各种开发工具和高性能库组成的工具套件在提供友好的编程环境之外，还可为用户的方案提供 DPC++、oneDPL、oneDNN、oneCCL 等众多核心元素，能够助力用户快速建立从数据到应用的端到端整体方案，并实现更优硬件性能。

在面向音视频领域的应用中，其中的英特尔® oneAPI 渲染工具包（英特尔® oneAPI Rendering Toolkit）可借助有效的渲染和光线追踪库，以及英特尔® Embree、英特尔® IPSC、英特尔® Open Image Denoise、英特尔® OSPRay 等软件工具，来加速相应的工作负载，高效创建高性能、高保真视觉体验。其优势包括：

- **跨并行处理架构和平台的高效部署；**
- **访问所有系统内存空间，即使是大规模数据集；**
- **通过具有全局照明的光线追踪提高视觉保真度；**
- **适用于任何数据大小的经济高效的交互式性能；**
- **高性能、基于深度学习的去噪。**

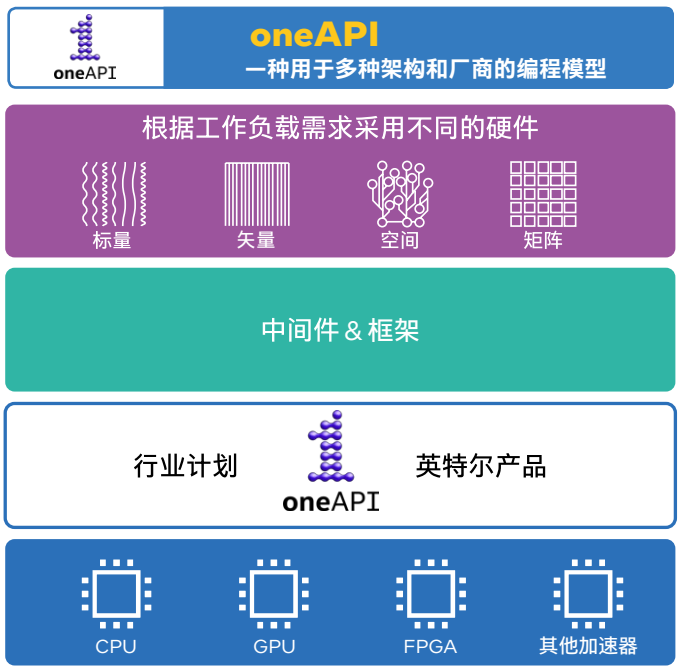


图 36 英特尔® oneAPI 工具套件组成

借助这一工具包，用户可以对通用处理器（例如英特尔® 至强® 可扩展处理器）、图形处理器（例如基于 Xe 架构的英特尔® 服务器 GPU）以及 FPGA 产品等进行优化，在包括数字内容创作、专业影视渲染、AR 游戏等不同的工作负载中突破可视化界限，构建出各类绚丽夺目的动画和视觉效果。

英特尔® Media SDK

作为视频处理的综合型 API，英特尔® Media SDK 能面向不同平台的视频和媒体应用，例如数字监控、零售、云游戏、视频会议等，提供丰富的基于硬件加速的视频编码、解码和处理能力。

英特尔® Media SDK 为用户提供了高性能软件开发工具、库和基础设施，有助于用户在基于英特尔® 架构的硬件基础设施上创建、开发、调试、测试和部署企业级媒体解决方案。其提供的优势特性包括：

- 可在应用场景中为用户提供快速、高质量、实时的视频转码能力，例如广播、OTT 交付、实时视频和视频点播（VOD），以及云游戏和远程桌面解决方案；
- 加速用户的媒体功能，包括视频播放、编码、处理和媒体格式转换；
- 支持 60 FPS 的 HEVC（High Efficiency Video Coding，高效视频编码）编解码能力实现实时 4K 视频通信。在一些基于英特尔® 架构的处理器平台上，能帮助用户以 30 FPS 的速度获得多达 18 路的 AVC 全高清视频[27]；
- 可助力用户快速开展媒体功能的原型设计、优化和产品化，并加速采用更高效编解码器以提升 AVC、HEVC 和 MPEG-2 视频的速度、压缩率和质量；
- 助力用户快速调试和定制产品，查找并解决应用程序中的错误，并快速将应用程序移植到新平台中去。

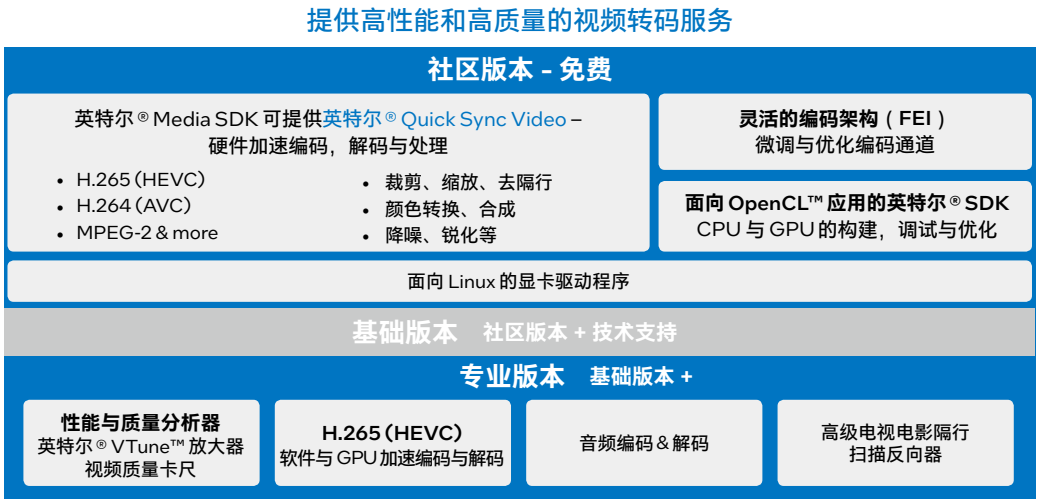


图 37 英特尔® Media SDK 构成

[27] 如欲了解更多详情，请访问英特尔官网：https://www.intel.cn/content/www/cn/zh/develop/tools/media-sdk.html

英特尔® oneVPL

英特尔® oneVPL（英特尔® oneAPI Video Processing Library）是继英特尔® Media SDK 推出的下一代视频处理软件，其为视频编解码及其它通用视频处理提供了统一的、以视频为中心的 API 接口，并支持跨各种硬件加速器工作，可帮助用户在更多硬件加速器和更广泛的应用场景中获得性能提升和编程灵活性，非常适用于视频广播、直播流媒体、视频点播、云游戏和远程桌面解决方案等场景，且广泛继承了英特尔® Media SDK 的优势，包括：

- 提供了与英特尔® Media SDK 核心 API 的兼容性；
- 具备与英特尔® Media SDK 相同的视频编解码器和滤波器；
- 支持在通用处理器、集成显卡 GPU、独立显卡 GPU 以及其他硬件加速器中的部署。

相应地，英特尔® oneVPL 也舍弃了英特尔® Media SDK 的一些特性，不再支持音频处理相关 API、灵活编码功能模块（Flexible Encode Infrastructure，FEI）和 Opaque Memory，增添了以下特性：

- 改进了视频处理初始化模式，可用于支持更广泛的视频处理实现方式；
- 提供了新的内存抽象和优化方式，以及对解码性能的优化。

同时，与英特尔® Media SDK 良好的兼容性，也使用户通过英特尔® oneVPL 对现有视频处理代码实施灵活迁移，以及利用英特尔® oneVPL 中新的软件实现方式以及即将推出的新硬件特性，例如扩展的 AV1 编码、更方便的设备枚举和视频处理初始化等，来获得更好的视频处理性能。

可扩展视频技术（SVT）

作为开源项目 Open Visual Cloud 的重要组成部分，可扩展视频技术（Scalable Video Technology，SVT）是一种基于软件的视频编码技术，可为媒体和视觉云（Visual Cloud）开发人员提供灵活的高性能软件编码器核心库。

通过独特的架构设计与算法特性，SVT 不仅具有良好的运行效率和可扩展性，并能让不同编解码器在性能、延迟和视觉质量之间实现更佳平衡。同时，SVT 也针对多个英特尔® 处理器平台（例如英特尔® 至强® 可扩展处理器）进行了高度优化，使处理器的多内核资源可为 SVT 编解码器的性能提升提供助力。因此，用户可以使用 SVT 视频编解码器，组合基于英特尔® 架构的处理器，来实现视频处理效率的大幅提升。

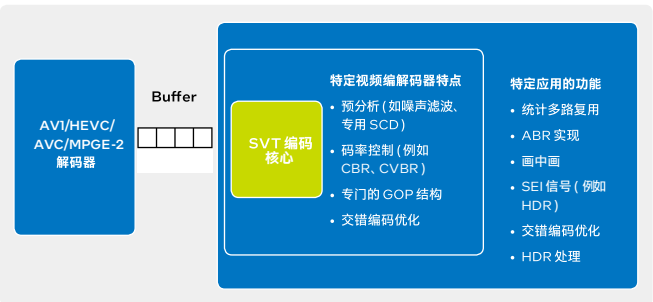


图 38 可扩展视频技术

得益于 SVT 技术，开发人员可以加速创新视觉云服务的开发过程，构建更快、更高质量的全功能编解码器，并缩短云视频转码解决方案的产品开发周期。

SVT 技术可用于开发符合不同标准的编解码器，其中英特尔创建了开源 SVT-HEVC 和 SVT-VP9 编码器内核供开发人员用于创建自己的产品和服务。另外，SVT-AV1 和 SVT-AVS3 等其它 SVT 编解码器同样也具有显著的架构优势。

结语

伴随通信与网络技术的高速发展以及AI、AR/VR等前沿技术的不断推进，更多基于音视频能力的创新方案正在不同行业和领域中获得广泛应用与商用化落地。无论是对于直播、点播或游戏等数字娱乐产业的“当红花旦”，还是对于影视制作、智能语音处理等行业应用背后的“无名英雄”而言，是否构建高品质音视频能力已成为其能否致胜未来发展的要决之一。

但在高品质音视频能力建设过程中，由高清技术发展带来的视频编解码和转码压力，云化解决方案带来的算力分布式部署难题，以及各类智能应用融合带来的AI加速等纷繁复杂的需求，再加之以不断“扩容”的市场规模带来的数据处理风暴，都对音视频技术方案的构建提出了严峻挑战。

英特尔长期以来打造的广泛而丰富的生态系统，以及一系列先进的英特尔产品与技术，正在帮助用户以更优的效率、更好的可用性以及更低的TCO，在从边缘到云的不同位置上构建不同类型的音视频解决方案。这些产品与技术既包括英特尔® 至强® 可扩展处理器、英特尔® 服务器GPU、英特尔® 傲腾™ 持久内存以及英特尔® 视觉云媒体分析加速卡等硬件，其已在多个实践案例中被证明可为方案提供强劲的计算、存储和网络处理能力；也包括英特尔® Media SDK、SVT、英特尔® oneVPL、英特尔® oneAPI等软件，同样经实践验证可为音视频能力的工作效能提供有力加速。

这些基于英特尔产品与技术的方案涵盖了视频直播、视频点播、影视制作、AR/VR、云游戏以及智能语音等不同层面的领域与行业，帮助客户能基于高品质的音视频能力推动行业创新、提升企业运营效率并实现项目的快速落地。

同时，由英特尔提供的一系列相关服务和计划，也正为各个行业用户提供强大的生态系统支持，以及活动、培训、推广和其他参与机会，旨在与合作伙伴协同推动各基于音视频能力的前沿技术与应用在更广泛的企业级场景中的落地实践，让互联网创新和应用更为绚丽多彩，让人们的生活更加幸福欢乐。

扫码了解更多英特尔云计算创新实践

法律声明

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。

描述的成本降低情景均旨在在特定情况和配置中举例说明特定英特尔产品如何影响未来成本并提供成本节约。情况均不同。英特尔不保证任何成本或成本降低。

英特尔技术特性和优势取决于系统配置，并可能需要支持的硬件、软件或服务得以激活。产品性能会基于系统配置有所变化。没有任何产品或组件是绝对安全的。更多信息请从原始设备制造商或零售商处获得，或请见 intel.com。

英特尔技术可能需要支持的硬件、软件或服务得以激活。请从原始设备制造商或零售商处获得更多信息。