intel XEON

白皮书

第五代英特尔® 至强® 可扩展处理器
英特尔® AVX-512
英特尔® AMX
腾讯云向量数据库 VectorDB

软硬件并行优化，第五代英特尔® 至强® 可扩展处理器助腾讯云向量数据库成为大模型时代数据中枢

腾讯云

“大语言模型以及 AGI、AIGC 应用等新一代 AI 技术与能力正在各行各业崭露头角，向量数据库无疑是为其提供强大知识库后援的数据中枢。腾讯云向量数据库希望成为这个数据中枢，并联合腾讯云的其它能力，为用户打造下一代 AI 能力的坚实基座。来自英特尔的第五代至强® 可扩展处理器、英特尔® AVX-512 指令集以及英特尔® AMX 加速引擎等产品与技术，为腾讯云向量数据库性能的提升提供了更加强劲的助力。”

罗云
腾讯云数据库副总经理

概述

以大语言模型（Large Language Model，LLM）、人工通用智能（Artificial General Intelligence，AGI）以及生成式人工智能（Artificial Intelligence Generated Content，AIGC）为代表的新一代人工智能（Artificial Intelligence，AI）技术与应用正对数据提出更高的要求，而向量数据库能以全新的构建模式，作为 AI 知识库来为 LLM 模型等提供补充，有效降低模型训练的成本，并提高 AGI、AIGC 等应用输出结果的及时性和准确度。

作为 AI 领域的领先者，腾讯依托旗下腾讯云推出向量数据库（Tencent Cloud VectorDB），其不仅支持多种索引类型和相似度计算方法，还具有单索引支持千亿级向量规模、百万级每秒查询率（Queries-per-second，QPS）及毫秒级查询时延等优势。

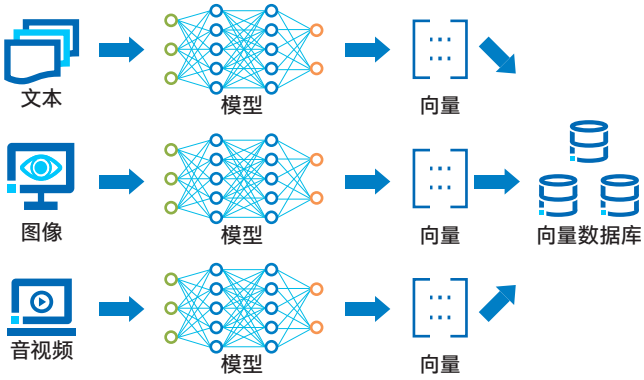
为进一步提升向量数据库的性能，腾讯云与英特尔展开合作，引入第五代英特尔® 至强® 可扩展处理器、英特尔® AVX-512 指令集以及英特尔® AMX 加速引擎等最新产品与技术，以软、硬件两方面的并行优化，为腾讯云向量数据库提供显著的性能加速，这在双方随后开展的验证测试中也得到了证实。

方案背景：腾讯云向量数据库为新一代 AI 提供知识库

在 AI 技术发展和应用落地的过程中，数据无疑扮演着极为重要的角色。借助文本、图像、音视频等不同类型数据所开展的模型训练，能让 AI 性能不断得以优化，应用能力也能获得持续更新。而随着 LLM 模型等新一代 AI 技术的普及，爆炸性增长的模型体量与复杂度也对数据提出了更高要求，AI 能力的构建中亟需引入规模更大、质量更优的数据。

但对用户而言，准备数量充足、且符合安全等各方面要求的数据无疑是一项巨大的挑战。很多场景下用户会选择公共数据集来开展模型预训练等工作，但这会带来新的问题，包括预训练得到的 LLM 模型无法借助用户获得的实时数据进一步迭代优化，且用户庞大的私域数据也无法为 LLM 模型提供更新，而由此带来的知识库滞后会造成应用效果的不佳，例如在 AGI、AIGC 等应用中出现内容谬误、AI 幻觉等情况。这一趋势下，以内容丰富且更新及时的知识库来作为 LLM 模型等新一代 AI 技术运行的数据补充已势在必行。

AI 任务中所用的各类数据通常会以向量 (Vector) 形式表示。向量是一种具有大小和方向的数据表示，其包含多个维度的信息，且每个维度都会用来表示一个特征或属性，以向量形式来构建的知识库无疑将更具效率。例如在图像处理任务中，图像可表示为像素值的向量；而自然语言处理任务中，文本可表示为词向量或句子向量。



图一 各类数据转化为向量后存入向量数据库

近年来备受瞩目的向量数据库，正是业界为实现高效能的新一代 AI 知识库而推动的全新数据库类型。作为一种以向量空间模型为基础的数据库，向量数据库可对向量化后的数据进行高效的存储、处理与管理。如图一所示，数据的向量化是借助词向量模型、卷积神经网络等 AI 模型，经过嵌入 (Embedding) 环节将文本、图像、音视频等不同数据转换为向量后存入向量数据库中，而向量查询则是通过对向量之间的相似度计算来完成。

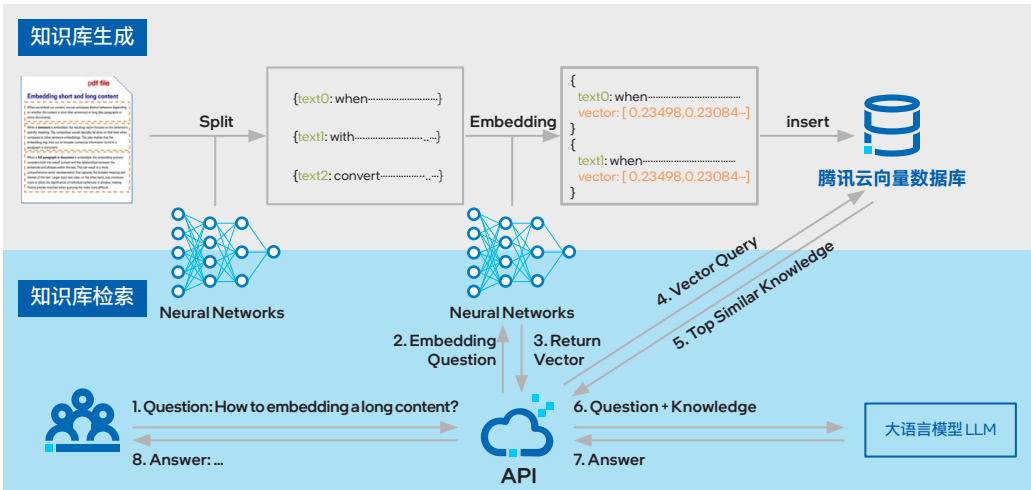
在实践中，向量数据库可作为知识库，通过与 LLM 模型等的结合来有效降低用户的模型训练成本，提升 AIGC 等应用的信息输出准确度和及时性。同时向量数据库还可用于 LLM 模型预训练数据的分类、去重和清洗，相比传统方式实现效率的大幅提升，这些优势也使向量数据库获得了市场的巨大青睐。

而一直站在 AI 技术潮头的腾讯，也依托其旗下腾讯云推出了全新的向量数据库产品[1]。作为一款全托管的自研企业级分布式数据库服务，其能为多维向量数据提供高效的存储、检索和分析能力，其优势包括：

具备完善的嵌入功能	高性能、高可用、稳定可靠
支持多种索引类型 和相似度计算方法	单索引支持千亿级向量规模
简单易用、低成本	可支持百万级 QPS 及毫秒级查询延迟

借助上述优势与特性，腾讯云向量数据库正成为用户构建 AI 能力时的强力“后援团”。如图二所示，数据库能与 LLM 模型配合使用。用户的私域数据在经过文本分割、向量化后存储在腾讯云向量数据库中，构建起专属的外部知识库，并在后续的检索任务中，为 LLM 模型提供提示信息，辅助 AGI、AIGC 等应用生成更加精准的结果。此外，这一产品还可广泛应用于推荐系统、计算机视觉以及智能客服等 AI 领域。

为助力腾讯云向量数据库进一步提升效能，英特尔也以其先进产品与技术为抓手，为这一数据库产品提供软硬件的双向驱动。在双方的合作中，全新第五代英特尔® 至强® 可扩展处理器与英特尔® AVX-512 指令集，英特尔® AMX 加速引擎等的加入，让向量检索的计算性能得到了显著提升，双方随后开展的验证测试也有力地证明了这一点。



图二 腾讯云向量数据库为 LLM 模型提供知识库[2]

解决方案：英特尔产品与技术从软硬件两方面为腾讯云向量数据库提供加速

不同于传统数据库通过关键词匹配、筛选或过滤等方式来实现数据检索，向量数据库中的向量检索主要依靠相似度度量方法来完成。相似性度量是通过计算来确定向量数据库中两个向量的相似程度，并最终找出与给定查询向量最相似的向量。常用的相似度度量计算方法包括内积 (Inner Product)、欧氏距离 (Cosine Similarity) 以及余弦相似度 (Euclidean Distance) 等：

- **内积：** 计算两个向量之间的点积 (内积)，所得值越大与搜索值越相似；
- **欧氏距离：** 计算向量之间的直线距离，所得的值越小与搜索值越相似；
- **余弦相似度：** 计算两个向量在多维空间中的夹角余弦值，从而衡量它们的相似程度，所得值越大与搜索值越相似。

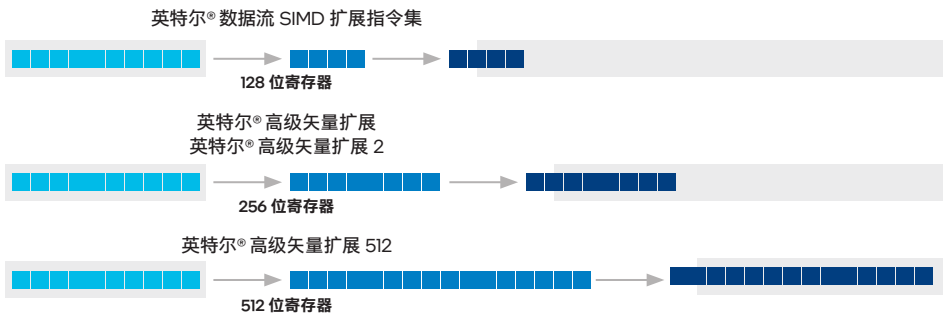
由此可知，在向量数据库的检索中需要完成大量的相似性度量计算任务，计算性能越强，数据库性能表现也就越佳。因此，英特尔需要帮助腾讯云向量数据库在更好的性价比前提下构建更高密度的算力输出，而全新第五代英特尔® 至强® 可扩展处理器和英特尔® AVX-512 指令集、英特尔® AMX 加速引擎的加入，无疑可有效地应对这一需求。

作为至强® 可扩展处理器家族的最新一代，第五代英特尔® 至强® 可扩展处理器不仅以全新的架构设计、更多的内核、更强的单核性能以及性能更优越的内存子系统带来更强的性能输出，同时也为多样化的 AI 任务 (例如基于 LLM 模型的 AI 应用) 提供了算力密度、总拥有成本 (Total Cost of Ownership，TCO) 等方面的优化。同时，新处理器还内置了一系列强劲的指令集和 AI 加速引擎，包括英特尔® AVX-512、英特尔® AMX 等，来为 AI 任务提供强劲助力。这些技术更新都有效地帮助腾讯云向量数据库深度优化其向量检索算法的执行效率，大幅提升了检索性能。

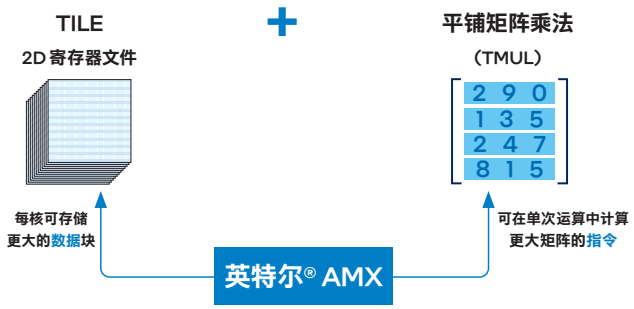
作为一种单指令多数据 (Single Instruction Multiple Data，SIMD) 指令集，英特尔® AVX-512 在密集型计算负载中有着得天独厚的优势。得益于其 512 位的寄存器宽度和两个 512 位的融合乘加 (Fused Multiply Add，FMA) 单元，指令集能并行地执行 32 次双精度、64 次单精度浮点运算，或操作 8 个 64 位和 16 个 32 位整数。

在腾讯云向量数据库所需的向量相似度计算中，假设数据类型是 FP32，输入向量 x 中的 16 个维度数据和数据库中向量 y 的 16 个维度数据，都可以一次性被加载到英特尔® AVX-512 的寄存器中，从而实现一次处理 16 个维度的并行运算，效率提升极为可观。由于在各类向量检索算法中 (典型如使用 IVF-PQFastScan 算法)，类似计算需求的比重往往很高，因此数据库的性能可获得巨大提升。

另一项可为腾讯云向量数据库带来巨大性能提升的是英特尔® AMX 加速引擎。这一技术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 能针对这一场景实现有效加速。



图三 英特尔® SSE、英特尔® AVX2 和英特尔® AVX-512 之间的寄存器大小和计算效率的差异说明



图四 英特尔® AMX 架构由 2D 寄存器文件（TILE）和 TMUL 组成

基于英特尔® AVX-512 和英特尔® AMX，腾讯云与英特尔一起，针对腾讯云向量数据库常用的一些计算库进行了专门的优化。包括：

- **FAISS：** 方案中针对其不同的索引提出了不同的优化方案，包括面向 IVF-FLAT 算法的 ReadOnce（单次读取）和 Discretization（离散化）两种优化思路，以及借助英特尔® AVX-512 加速 IVF-PQFastScan 算法和 IVF-SQ 索引的优化方案；
- **HNSWlib：** 方案借助英特尔® AVX-512，对 HNSWlib 的向量检索性能进行了加速。同时方案也针对增删数据后的性能和召回率抖动的问题进行了专向优化，使 HNSWlib 的性能和召回率可以保持较平稳状态。

此外，英特尔还为腾讯云向量数据库提供了英特尔® FMAL 加速库（Intel® Feature Matching Acceleration Library，英特尔® 特征匹配加速库）。在面临海量向量数据时，暴力搜索有着非常多的使用，但这一场景对算力需求非常高，因此性能优化极为必要。作为针对向量暴力搜索场景开发的算法库，英特尔® FMAL 在英特尔® AVX-512 和英特尔® AMX 的加持下，能对相似度计算进行加速并提供了相似度计算和 top-K 查询的 API 接口。值得一提的是，英特尔® FMAL 能与英特尔® AMX 结合，对 INT8 数据类型的性能实现进一步优化。

同时，英特尔® FMAL 还能在多线程并发下对处理器资源进行合理地调配，以便让用户充分挖掘最新处理器所具备的多核心优势。除此之外，加速库也提供了对内存的非一致内存访问架构（Non Uniform Memory Access，NUMA）优化和缓存数据对齐功能，这些都进一步提升了腾讯云向量数据库的性能。

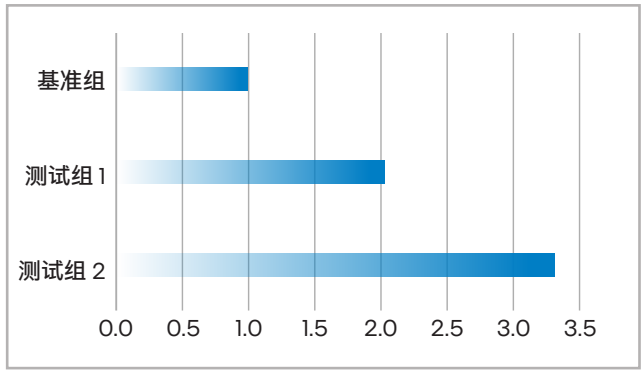
效果评估：经英特尔产品与技术加速的腾讯云向量数据库获得显著性能提升

为验证第五代英特尔® 至强® 可扩展处理器、英特尔® AVX-512 及英特尔® AMX 的引入，对腾讯云向量数据库中向量检索任务提供的助力，腾讯云与英特尔携手开展了验证测试，测试分为两个场景：

场景 1：英特尔® AVX-512 优化效果与代际性能提升测试[3]

测试分为以下三个对比组：

- **基准组：** 基于第三代至强® 可扩展处理器的腾讯云 S6 服务器，实例规格：16 虚拟核；测试中使用 Faiss 计算库的 IVF-PQFastScan 算法进行检索，向量维度 768，数据集数据量 10 万，nprobe=10；
- **测试组 1：** 基于第三代至强® 可扩展处理器的腾讯云 S6 服务器，实例规格：16 虚拟核；使用英特尔® AVX-512 对 Faiss 计算库的 IVF-PQFastScan 算法进行优化，向量维度 768，数据集数据量 10 万，nprobe=10；
- **测试组 2：** 基于第五代至强® 可扩展处理器的服务器，实例规格：16 虚拟核；使用英特尔® AVX-512 对 Faiss 计算库的 IVF-PQFastScan 算法进行优化，向量维度 768，数据集数据量 10 万，nprobe=10。



图五 英特尔软硬件产品与技术带来的性能提升（归一化）

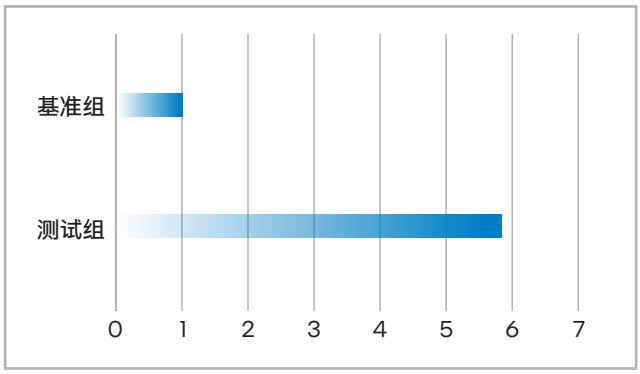
测试结果如图五所示，经数据归一化对比后，在同样使用腾讯云 S6 服务器（基于第三代至强® 可扩展处理器）的情况下，使用英特尔® AVX-512 优化后，使用 IVF-PQFastScan 算法执行向量检索时的 QPS 性能提升了约 100%，而将算力设备升级为

第五代至强® 可扩展处理器后，性能相比基准组提升了约 230%。这表明，第五代至强® 可扩展处理器与英特尔® AVX-512 指令集能显著提升向量数据库的向量检索效率。

场景 2：英特尔® AMX 进一步提升向量检索性能[4]

测试分为以下两个对比组：

- **基准组：** 基于第五代至强® 可扩展处理器的服务器，规格：16 虚拟核，使用经英特尔® AVX-512 优化的英特尔® FMAL 执行暴力搜索算法，数据量为 1,000 万，batch size 为 16，数据格式为 FP32，线程数为 16；
- **测试组：** 基于第五代至强® 可扩展处理器的服务器，规格：16 虚拟核，使用经英特尔® AMX 优化的英特尔® FMAL 执行暴力搜索算法，数据量为 1,000 万， batch size 为 16，数据格式为 INT8，block size 为 320，线程数为 16。



图六 英特尔® AMX 优化加速暴力检索的吞吐性能（归一化）

测试结果如图六所示，经数据归一化对比后，在同样使用第五代至强® 可扩展处理器的算力平台上，使用英特尔® AMX 加速数据格式为 INT8 的测试场景相比使用英特尔® AVX-512 加速数据格式为 FP32 的测试场景，性能提升达约 5.8 倍。这表明，第五代至强® 可扩展处理器与英特尔® AMX 的配合，能进一步帮助腾讯云向量数据库提升向量检索效率。

展望

随着 LLM 等新一代 AI 技术与能力在更多行业获得实践与落地，以腾讯云向量数据库为代表的 AI 知识库，也正在更多 AI 任务和场景中发挥越来越重要的作用。这一过程中，计算性能的持续提升，势必能为这一趋势提供更强助力。面向未来，腾讯云还将与英特尔展开更多合作，将更多先进计算产品与技术应用到该领域中，在竞争激烈的 AI 时代取得更为领先的优势。

[1] 更多腾讯云向量数据库特性，请参阅腾讯云官网介绍：https://cloud.tencent.com/product/vdb
[2] 图片及相关介绍援引自腾讯云官网介绍，详情请参阅：https://cloud.tencent.com/product/vdb
[3, 4] 数据来源于腾讯未公开的内部测试，如欲了解更多详情，请访问：https://www.tencent.com/

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。
性能测试结果基于【2023 年 11 月】进行的测试，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。
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。
本文中提供的所有信息可在不通知的情况下随时发生变更。关于英特尔最新的产品规格和路线图，请联系您的英特尔代表。
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、英特尔标识以及其他英特尔商标是英特尔公司或其子公司在美国和/或其他国家的商标。
©英特尔公司版权所有