解决方案

超大规模数据治理

格灵深瞳基于第二代英特尔® 至强® 可扩展处理器 + 英特尔® 傲腾™ 数据中心级持久内存的以图搜图引擎解决方案

企业介绍

北京格灵深瞳信息技术有限公司（以下简称：格灵深瞳）是一家专注于计算机视觉以及人工智能的科技公司，致力于推出低成本、具备大规模部署能力的人工智能产品和服务。作为最早一批在中国把人工智能投入到商业应用的创业企业，格灵深瞳在智能交通、智能零售、智慧安防、智慧银行等领域持续深耕，为遍布全国和全世界的客户提供包含智能传感器、智能识别、智能云计算和服务机器人的综合智能解决方案和服务。

背景

随着计算机视觉和深度学习技术的迅速发展，通过计算机视觉能力检测、识别图像如今已经被零售、安防监控、公共安全、智能交通等各个领域接受并广泛应用。例如，在智能交通领域，交通管理部门常常需要通过一张机动车图片检索目标信息，并在省市级的范围内追踪目标轨迹。针对目标特征应用，格灵深瞳推出“以图搜图”引擎产品，该产品运用领先的深度学习和高性能运算技术，可检测并识别图片中的全目标（人脸、机动车、非机动车、行人等）结构化信息，进行大规模特征比对、以图搜图等服务。

解决方案

针对不同的场景需求，格灵深瞳以图搜图引擎提供不同的解决方案：

- 对于规模较小，1 亿内特征存储、超高频率的特征比对场景（例如：动态人脸布控），提供基于图形处理器（GPU）的解决方案，满足用户的高频比对需求：
- 对于超大规模特征存储、低频比对场景（例如：抓拍库的以图搜图），提供基于第二代英特尔® 至强® 可扩展处理器 + 英特尔® 傲腾™ 固态盘/数据中心级持久内存的解决方案，满足用户的超大规模特征存储比对需求：
- 同时引擎支持横向扩展，支持最大 100 台机器的横向扩展，最大支持 3000 亿目标（人体/车辆/非机动车）或 1000 亿人脸的以图搜图在 5 秒内返回结果。
- 下图为格灵深瞳基于 Intel 技术的分布式以图搜图引擎架构图：

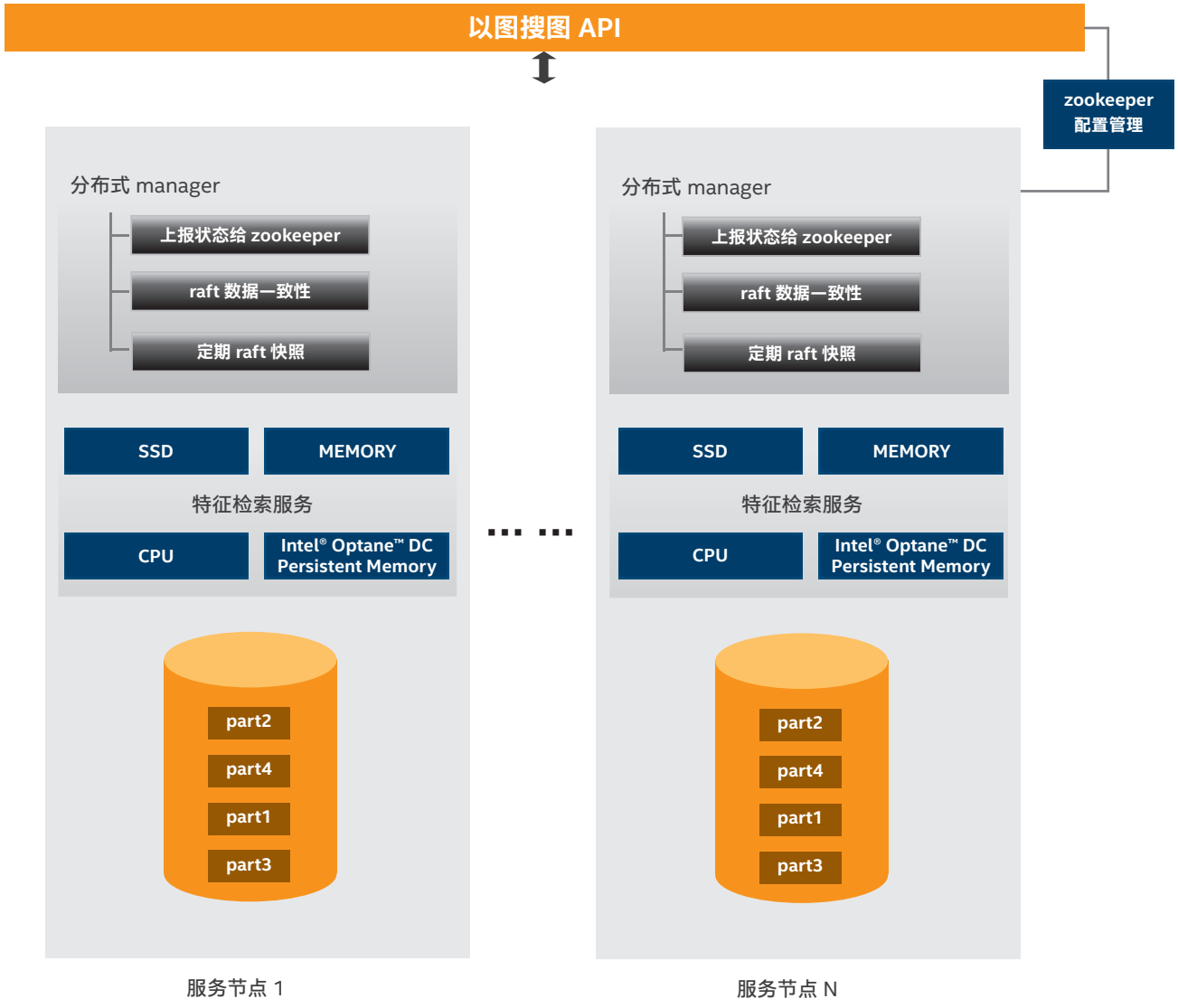

图 1. 基于 Intel 技术的格灵深瞳以图搜图引擎分布式架构

格灵深瞳以图搜图引擎采用第二代英特尔® 至强® 金牌处理器，该处理器性能优异，同时支持 AVX2，AVX-512 指令集。每处理器六通道的 DRAM 提供了充足的内存带宽，同时经过高度优化的英特尔® 数学核心库（英特尔® MKL）提供了优越的函数计算能力和多线程支持。

效果

下文以人脸特征描述性能参数，人/车/非特征为人脸性能的 3 倍，针对 1 亿条或者更少的数据，所有目标特征都可以保存在同一台服务器的 DDR4 内存中（占用 50G 内存），达到 1 亿条特征比对 1 秒返回的性能指标。

在智慧城市等应用场景中，1 亿条的特征比对上限已经不能满足用户的需求。为了支持更多的特征数量，同时克服内存空间较小、价格较贵、普通固态盘读取速度慢的缺点，格灵深瞳在超亿级特征对比产品中采用英特尔® 傲腾™ 固态盘，组建虚拟内存方案。

英特尔® 傲腾™ 固态盘比内存价格低，但虚拟内存读取性能远高于普通固态盘。经过测试，该解决方案达到了单机 5 亿人脸特征的比对能力，以图搜图结果可以在 3 秒内返回。

当目标数量超过 5 亿时，传统内存 + 固态盘/虚拟内存的方案，都存在数据搬运的瓶颈问题。格灵深瞳采用英特尔® 傲腾™ 数据中心级持久内存，在单一服务器上实现了大容量、高带宽和低延迟的持久存储。测试结果表明（见表 2）：在 10 亿目标的情况下，以图搜图可以在 10 秒内返回，单机可支持 40 亿热数据搜索。

	软件和硬件解决方案	优势	局限性
1	第二代英特尔® 至强® 可扩展处理器； DDR4 内存，英特尔® MKL； 英特尔® 深度学习加速技术；	英特尔® 至强® 可扩展处理器可提供卓越的性能； CPU 支持英特尔® AVX2、AVX-512 指令集， 英特尔® MKL 可显著提升计算效率。	普通固态盘存在内存容量小、成本高、读取速度低等劣势，这意味着该解决方案不适用于大规模特征以图搜图场景。
2	第二代英特尔® 至强® 可扩展处理器； 英特尔® 傲腾™ 固态盘虚拟内存解决方案； 英特尔® MKL； 英特尔® 深度学习加速技术；	英特尔® 傲腾™ 固态盘的成本低于内存，但其存储容量和读取性能远高于普通固态盘。	数据块数量超过 5 亿时，搜索结果将延迟 10 秒。
3	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® MKL； 英特尔® 深度学习加速技术；	一台机器支持更多特征块，具有较为出色的可扩展性。数据读取性能远高于普通固态盘。	

表 1. 不同以图搜图引擎解决方案对比

下表展示了三款不同内存解决方案的以图搜图性能。

数据大小	DRAM 解决方案*1		英特尔® 傲腾™ DC 固态盘 解决方案（iMDT）*2		英特尔® 傲腾™ DC 持久内存 解决方案（App Direct 模式）*3	
	首次搜索	后续搜索	首次搜索	后续搜索	首次搜索	后续搜索
1 亿	460 毫秒	460 毫秒	1623 毫秒	590 毫秒	403 毫秒	403 毫秒
2 亿	980 毫秒	980 毫秒	14.8 秒	1.2 秒	786 毫秒	786 毫秒
3 亿	-	-	17.3 秒	1.8 秒	1.9 秒	1.9 秒
4 亿	-	-	20.1 秒	2.4 秒	3.1 秒	3.1 秒
5 亿	-	-	20.5 秒	3.0 秒	4.2 秒	4.2 秒
10 亿	-	-	-	-	10.0 秒	10.0 秒

表 2： 3 款内存解决方案的图片搜索性能

注 1： 配置 1 — 英特尔® 至强® 金牌 6140 处理器，DDR4 DRAM 32GB x6 @ 2666 MT/秒

注 2： 配置 2 — 英特尔® 至强® 金牌 6140 处理器，英特尔® 傲腾™ 固态盘，DDR4 DRAM 32GB x6 @ 2666 MT/秒

注 3： 配置 3 — 英特尔® 至强® 金牌 6240 处理器（测试时使用 18 核），英特尔® 傲腾™ DC 持久内存 128GB x6 @ 2666 MT/秒，配置成 2-2-2 模式，DDR4 DRAM 32GB x6 @ 2666 MT/秒

注 4： 随着更多测试的进行，性能指标评测结果可能进行修改。结果依赖于在测试中使用的特定平台配置与工作负载，可能不适用于具体的用户组件、计算机系统或工作负载。结果并不代表其他性能指标评测，其他性能指标评测结果可能或多或少受到限制。

性能结果依据 2019 年 5 月 30 日的测试，可能无法反映所有发布的安全更新。请查看配置声明，了解详情。
任何产品都无法提供绝对的安全。

相比 DRAM 解决方案，英特尔® 傲腾™ DC 持久内存解决方案还提供了更快的应用启动时间。对于 DRAM 解决方案，将 1 亿个数据片段从硬盘加载至内存以及为这些数据构建关系所需的时间约为 30 分钟，这意味着 10 亿个数据片段约花费 5 小时。通过将构建的数据直接放在持久内存，每 1 亿个数据片段花费的时间将缩短为 3 分钟，10 亿个数据片段仅需 30 分钟（参见图 2）。

更高的性能意味着更低的基础设施成本，相比采用 CPU 和传统 DRAM 内存的多机解决方案，采用英特尔® 傲腾™ DC 持久内存的第二代英特尔® 至强® 可扩展处理器可显著降低成本。

格灵深瞳与英特尔合作，利用最新款英特尔® 至强® 可扩展处理器、英特尔® 深度学习加速技术和创新的英特尔® 傲腾™ 存储技术将人工智能解决方案的性能提升至全新水平。

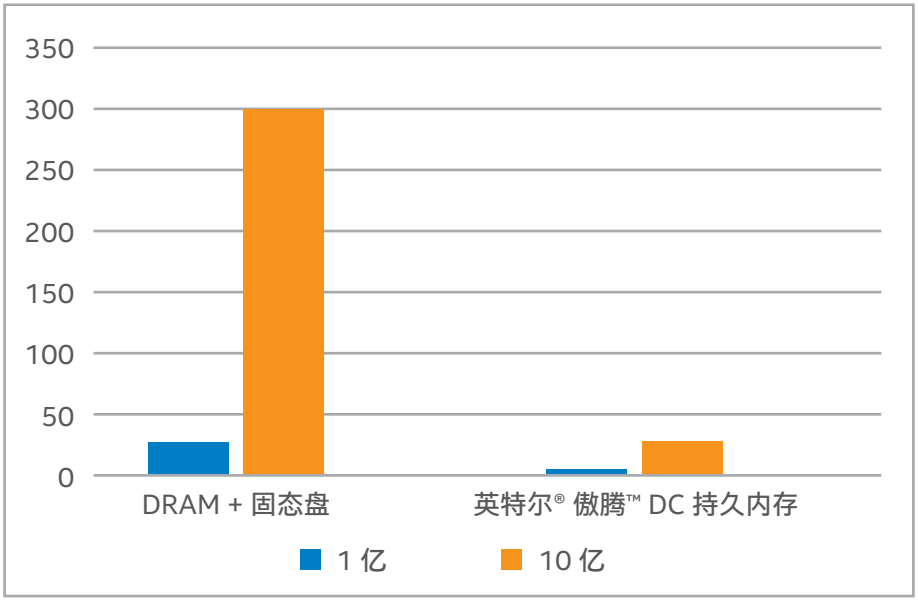

图 2：数据加载和构建时间（单位：分钟，越低越好）

注： 随着更多测试的进行，性能指标评测结果可能进行修改。结果依赖于在测试中使用的特定平台配置与工作负载，可能不适用于具体的用户组件、计算机系统或工作负载。结果并不代表其他性能指标评测，其他性能指标评测结果可能或多或少受到限制。

性能结果依据 2019 年 5 月 30 日的测试，可能无法反映所有发布的安全更新。任何产品都无法提供绝对的安全性。

在性能检测过程中涉及的软件及其性能只有在英特尔微处理器的架构下方能得到优化。诸如 SYSmark 和 MobileMark 等测试均系基于特定计算机系统、硬件、软件、操作系统及功能，上述任何要素的变动都有可能导致测试结果的变化。请参考其他信息及性能测试（包括结合其他产品使用时的运行性能）以对目标产品进行全面评估。更多信息敬请登陆：http://www.intel.cn/content/www/cn/zh/benchmarks/intel-product-performance.html

此处提供的信息可随时改变而毋需通知。如欲获得英特尔最新的产品规格和发展蓝图，请联系您的英特尔代表。

英特尔技术的特性和优势取决于系统配置、硬件、软件和服务。没有计算机系统是绝对安全的。如欲了解更多信息，请访问：Intel.cn