英特尔中国
AI应用案例集锦


腾讯云小微
飞桨
PaddlePaddle
百度智能云 BML
韵达
EXPRESS
ZTO 中通快递
ZTO EXPRESS
Bai du 百度
美团
云创大数据
www.cStor.cn
JD.京东
.COM
一起
南京大学
NANJING UNIVERSITY
平安科技
PING AN TECHNOLOGY


GREENET
INFORMATION SERVICE
AsiaInfo
亚信科技
当虹科技
Arcvideo Tech
China unicom中国联通
iQIYI 爱奇艺
悦 享 品 质
搜狐
SOHU.com
CATL 宁德时代
UnionPay 银联
BOE
Midea
阿里云
GOLDWIND SE
金风慧能
Smart Energy Services
智慧能源服务
首云
火山引擎
百度智能云
腾讯云
金山云

让网格仓智能分拣更快、更准

英特尔® GPU +OpenVINO™ 助力韵达提升业务效率并降低成本

视觉 AI 推理异构加速

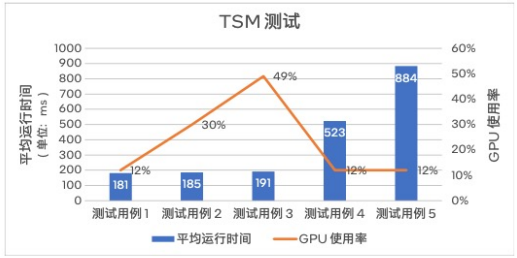
智能共配分拣系统

2023 年

114ms[1]
在三段码 OCR 测试中，相对于 130ms 的期望标准，测试结果平均运行时间为

98%[2]
在三段码 OCR 测试中，相对于 95% 的期望标准，测试结果准确度达

带来两大业务优势
借智能分拣节支增效，用分拨视频流分析优化业务管理和决策



测试用例 1：1 实例，批量大小=1
测试用例 2：2 实例，批量大小=1
测试用例 3：3 实例，批量大小=1
测试用例 4：1 实例，批量大小=3
测试用例 5：1 实例，批量大小=5

基于英特尔® 数据中心 GPU Flex 系列 170 的 TSM 性能测试结果[3]

优化后的视觉 AI 方案在算力时延、准确度、并发能力、稳定性和散热等多个方面都能够很好地满足韵达应用需求

OpenVINO™ 工具套件中的模型优化器和推理引擎两大组件，可优化模型性能，并为计算机视觉异构计算提供加速支持

引入英特尔® 数据中心 GPU Flex 系列，为体量庞大的视频流分析提供强劲的算力资源基础

intel XEON GOLD

intel DATA CENTER GPU FLEX SERIES

OpenVINO™

英特尔® oneAPI 工具套件

为应对分拨视频分析平台推进过程中遇到的快件数量庞大、智能分拣高准确度和低时延需求、视频流分析体量庞大等挑战，韵达引入英特尔® 数据中心 GPU Flex 系列 170 和 OpenVINO™ 工具套件来优化其三段码 OCR 检测系统和分拨视频分析平台性能，有效助力韵达优化视觉 AI 方案，提升业务效率并降低成本，同时优化业务管理和决策。

获取白皮书

1.2.3 有关性能和基准测试结果的更完整信息，请访问 https://www.intel.cn/content/www/cn/zh/customer-spotlight/cases/yunda-smart-logistics-strategy-visual-ai-solution.html

软硬协同+人机协作助物流作业持续优化

英特尔® XPU + OpenVINO™ + oneAPI 助中通快递加速边缘视觉 AI 方案

AI 推理
异构加速

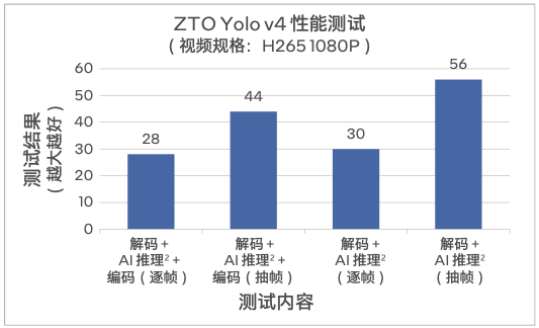
物流边缘
视觉 AI 应用

2023 年

34.8%[1]
基于英特尔® XPU 且软硬协同的新方案可助中通节约成本约

更简洁更高效
模型开发与维护

更好满足场景需求
提升工作效率 + 防范违规作业



基于英特尔® 数据中心 GPU Flex 系列 170 的 ZTO Yolo v4 性能测试结果[2]

基于 OpenVINO™ 工具套件和英特尔® oneAPI 工具套件，中通可大大简化 AI 应用开发，并实现应用跨 XPU 的无缝切换

OpenVINO™ 工具套件中的模型优化器和 Open Model Zoo 等可有效降低模型优化与开发难度，并缩短应用开发时间

XPU 加速带来更强灵活性：至强® CPU 加速轻量级 AI 场景，数据中心 GPU Flex 系列则负责对实时性要求较高或多并发的场景

intel XEON GOLD

intel DATA CENTER GPU FLEX SERIES

OpenVINO™
英特尔® oneAPI 工具套件

中通众多中心或网点都配备了 x86 服务器，同时也部署了英特尔® 数据中心 GPU Flex 系列。在英特尔工程师的协助下，中通只需在相同模型上进行开发，即可实现基于 XPU 的 AI 推理加速，同时加持 OpenVINO™ 工具套件和英特尔® oneAPI 工具套件，中通边缘视觉 AI 方案可有效满足业务端的更多需求，降低成本，在实际场景中实现更高性价比。

获取白皮书

1.2 有关性能和基准测试结果的更完整信息，请访问 https://www.intel.cn/content/www/cn/zh/customer-spotlight/cases/zto-express-expand-edge-vision-ai-application.html

用 CPU 担起 AI for Science 重任

英特尔® 至强® 可扩展平台五步助力 AlphaFold2 端到端优化

intel

2022 年

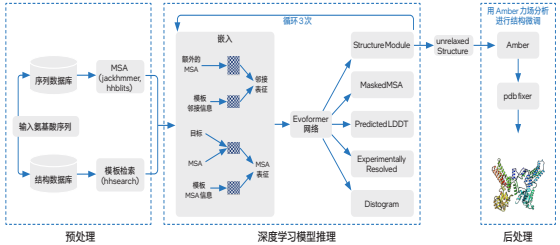
高通量优化/推理优化

蛋白质结构预测

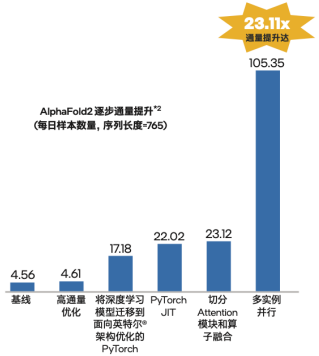
23.11倍[1]
每个优化步骤获得的提升累积后，端到端通量提升可达

5.05倍[2]
模型本身优化带来的提升达

4.56倍[3]
傲腾™ 持久内存 TB 级内存支持带来的提升达



AlphaFold2 基本架构



推理过程中多种优化措施带来的累积性能提升4

引入英特尔® 傲腾™ 持久内存的 TB 级内存支持，进一步打破"内存墙"瓶颈，并实现长序列高通量并行推理优化

使用 oneAPI 工具套件实施算子融合等优化，解决 AlphaFold2 计算效率低和处理器利用率不足等难题，缓解内存瓶颈等问题

利用第三代英特尔® 至强® 可扩展处理器在算力上的整体优势及其内置的 AI 加速技术进行并行计算优化

intel XEON PLATINUM

intel OPTANE PERSISTENT MEMORY

oneAPI

英特尔与专攻医药和生命科学研究和创新的产、学、研领域用户及合作伙伴们积极开展广泛及深入的协作，基于英特尔® 至强® 可扩展平台开展了 AlphaFold2 端到端优化，包括一系列并行计算优化举措和英特尔® 傲腾™ 持久内存产品的引入，使得整个 AlphaFold2 端到端处理过程的性能获得了质的提升。英特尔将继续推进更多创新产品、技术与 AlphaFold2 等 AI for Science 类应用的交互与融合，为各类前沿科学研究和探索带来更多加速、助力与收获。

杨威
人工智能架构师
英特尔

获取白皮书

1.2.3.4 有关性能和基准测试结果的更完整信息，请访问 https://www.intel.cn/content/www/cn/zh/customer-spotlight/cases/alphafold2-throughput-improvement-optimization.html

内铸智能运维 外促网络+AI 融合创新

至强® 可扩展平台助力亚信科技，基于网络大数据探索智能网络应用

AsiaInfo 亚信科技 | intel

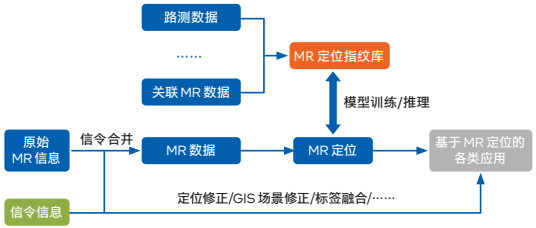
2022 年

算法/模型优化与部署 | **Network AI**

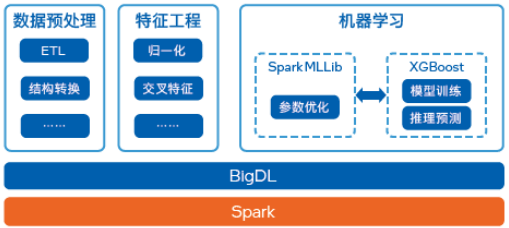
10%-15%[1]
在精准轨迹预测模型中，新方案相较于传统算法，定位精度提升了

50米[2]
在精准轨迹预测模型中，新方案相较于传统算法，模型预测误差可小于

全面提升
多个维度上的优化使方案在可用性、时效性和准确率上实现



亚信科技基于 MR 数据的应用解决方案



基于 BigDL 部署的 Spark + XGBoost 的方案架构

引入英特尔® oneAPI 工具套件，实现从数据预处理到数据特征工程、数据建模和部署的整体端到端优化

利用 BigDL 将面向 MR 定位数据的完整模型训练和预测方案无缝部署在 Spark 平台上，实现 XGBoost 单机方案的分布式部署

导入英特尔® 至强® 可扩展平台作为方案基础设施核心，为方案的高密度计算负荷提供高效算力支持

intel XEON | oneAPI | BigDL

基于 MR 数据等丰富的基础网络信息资源，我们的方案可帮助客户在市场经营、网络管理等领域开展广泛的智能应用探索与落地，成为企业效率倍增、提升用户体验的重要手段。英特尔® 至强® 可扩展平台、英特尔® oneAPI 工具套件和 BigDL 等产品与技术，帮助新方案实现了更高效的开发和更优异的性能。

鹿岩
无线网络规划优化产品研发总监
亚信科技

获取白皮书

1.2 有关性能和基准测试结果的更完整信息，请访问 https://www.intel.cn/content/www/cn/zh/customer-spotlight/cases/asiainfo-taps-intelligent-network-applications.html

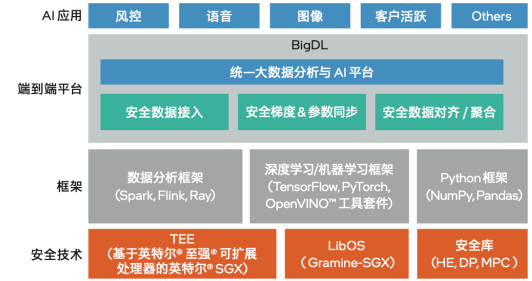
验证多源金融大数据隐私计算

英特尔® SGX 与 BigDL 助力打造基于 TEE 的联邦学习与实时预测方案

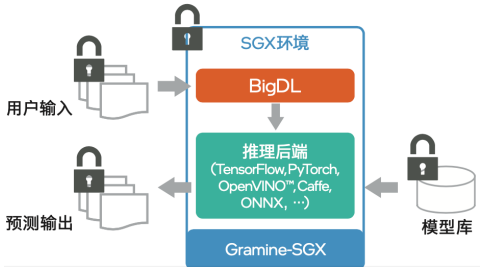
intel

2022 年

可信执行环境　联邦学习



基于英特尔® SGX 的 BigDL 平台架构



典型的基于英特尔® SGX –BigDL 平台的预测方案

多方金融大数据安全性 **大幅提升**

大数据平台与联邦学习 **无缝对接**

新方案在强化端到端安全的同时，性能 **符合预期**

融合英特尔® SGX 技术的 BigDL 平台能让用户便捷地在既有大数据平台上部署联邦学习方案

英特尔® SGX 在特定硬件环境中构建“飞地”，可为敏感数据和代码提供独立于操作系统和硬件配置的、更为强化的安全防护

英特尔® 至强® 可扩展处理器内置英特尔® SGX 及 AI 加速技术，能在可信执行环境（TEE）中助力 AI 模型训练与推理的扩展

intel Xeon PLATINUM　英特尔® 安全防护扩展 (SGX)技术　BigDL

以往出于数据安全考虑，AI 训练与推理所利用的数据往往被局限在部门或公司内部，很难进一步发挥 AI 应用的潜能。英特尔® SGX 与开源的统一大数据分析和 AI 平台 – BigDL 的引入，让广大金融机构能够更加安全、高效地开展基于 TEE 的联邦学习方法的探索与部署，并无缝扩展到多个大数据平台上，实现更多源数据的协同训练与推理。

获取白皮书

有关性能和基准测试结果的更完整信息，请访问 https://www.intel.cn/content/www/cn/zh/now/data-centric/sgx-bigdl-financial-big-data.html

为内外业务开展深度优化 AI 推荐引擎

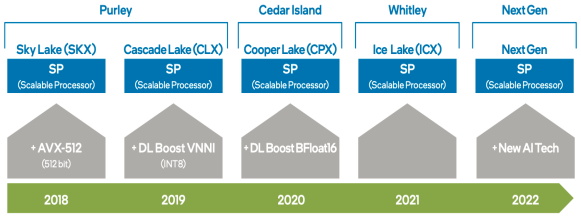
阿里云 | intel

英特尔® CPU+FPGA 助阿里巴巴 PAI 团队加速 DeepRec 推荐引擎

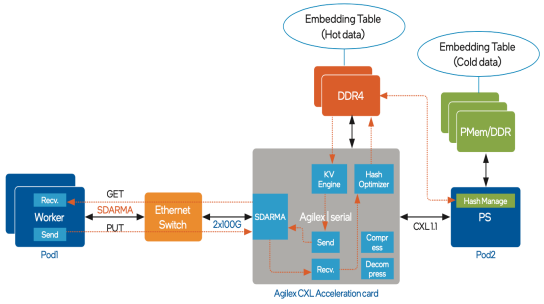
预测/推荐引擎优化

超大规模稀疏训练

2022 年



英特尔® 架构平台内置 AI 加速能力的演进



引入英特尔® Agilex™ I 系列 FPGA 实施优化

千亿特征 万亿样本

稀疏模型 场景定制 优化

端到端 性能加 速明显

基于 Agilex™ I 系列 FPGA 加速 Embedding Lookup，用一个平台支持多个场景，显著提升流通量并提供较低的访问时延

使用具备更高存储密度和数据持久化优势的傲腾™持久内存，来满足超大规模稀疏训练和预测对性能与容量的要求

引入至强® 可扩展处理器提供可靠算力，并以其内置的英特尔® DL Boost 进行框架、算子、子图和模型四个层面的优化

intel XEON | intel OPTANE PERSISTENT MEMORY | intel AGILEX

英特尔® DL Boost

英特尔® AVX-512

阿里巴巴 PAI 团队自 2019 年以来，用英特尔 AI 产品技术针对 DeepRec 进行了算子、子图、runtime、框架层和模型等多层面优化，以助力阿里巴巴加速内外部 AI 业务性能。针对 DeepRec 在 CPU、PMem 和 FPGA 不同硬件的优化实现方案，已成功部署到阿里巴巴多个内部和外部业务场景，并在实际业务中获得了明显的端到端性能加速，从不同角度解决了超大规模稀疏场景面临的问题和挑战。

获取白皮书

有关性能和基准测试结果的更完整信息，请访问 https://www.intel.cn/content/www/cn/zh/customer-spotlight/cases/sparse-model-training-prediction-engine-deeprec.html

5G 网络要借智能化实现动态节能

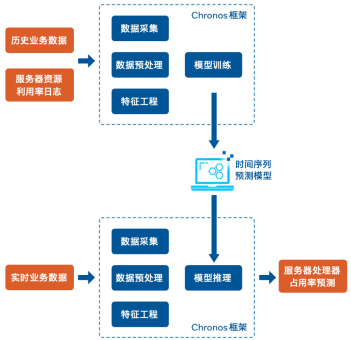
BigDL Chronos 框架助中国联通打造 5G 网元资源占用率预测方案

China unicom中国联通 | intel

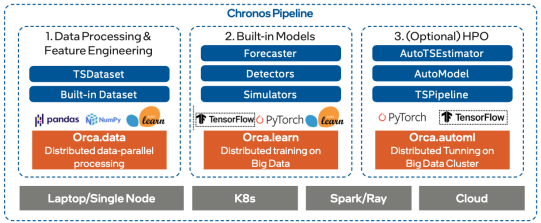
时间序列模型

5GC 网元资源占用率预测

2022 年



中国联通基于 Chronos 框架的时间序列预测方案基本流程



Chronos 框架基本架构

1.71[1]

处理器占用率预测值与实际值对比的最终 MSE 结果仅为

15%[2]

新方案预计可使单台服务器能耗降低超过

4,600 万度[3]

新方案可使整个云资源池预计每年节电

基于 BigDL Chronos 框架丰富的组件和集成的优化策略，方案实现了更优的预测效果和更快的预测速度

在 BigDL 超参数优化组件帮助下，开展从数据预处理、特征工程到模型训练等全栈的自动化机器学习过程

方案使用至强® 可扩展处理器为5GC (5G Core Network) 网元资源占用率预测方案提供通用算力支持

intel XEON GOLD

BigDL Chronos 框架

作为承载各类通信业务的基座，通信云资源池的能耗管理水平将直接影响中国联通实现节能减排的总体目标。为此，我们借助 AI 技术，以时间序列预测方案来助力降低服务器能耗。在这一过程中，来自英特尔的 Chronos 框架帮助我们更快、更好地完成了新方案的搭建，并获得了良好的预测准确率。

康凯
通信云项目经理
云网运营中心
中国联通

获取白皮书

1.2.3 有关性能和基准测试结果的更完整信息，请访问 https://www.intel.cn/content/www/cn/zh/customer-spotlight/cases/china-unicom-bigdl-chronos-framework-5gc.html

深挖 CPU 潜力提升训练性能及扩展性

至强® 可扩展平台助美团推进 TensorFlow 优化实践

美团 | intel

TensorFlow 框架优化

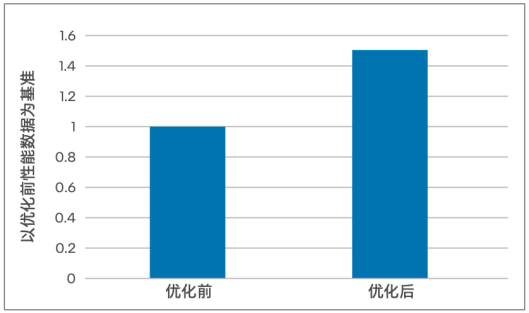
推荐系统

2022 年

1天内[1]
优化过后，全年样本完成训练的时间可控制在

10倍[2]
在推荐系统场景中，分布式扩展性提升超过

近线性加速
可做到千亿参数模型、上千 Worker 分布式训练的



Unique 单算子性能优化前后对比[3]

（越高越好）

与英特尔一同通过分布式负载均衡优化、通信优化和时延优化等举措，实现更优的分布式计算优化

在英特尔支持下围绕 Embedding ID 的 Unique 和 Partition 环节进行算子合并，取得 51% 的加速效果，实现单节点算力吞吐优化

在 TensorFlow 系统中应用英特尔® oneDNN，能更充分利用至强® 可扩展处理器的算力和内置 AI 加速能力来执行深度学习应用

intel XEON

英特尔® AVX-512　英特尔® oneDNN

英特尔® DL Boost　英特尔® VTune™ Profiler

为化解用户快速增长、智能业务不断创新升级、以及 AI 模型规模与复杂度持续上升带来的挑战，美团基于英特尔® 至强® 可扩展处理器，针对深度学习框架TensorFlow，从大规模稀疏参数的支持、训练模式、分布式通信优化、流水线优化、算子优化融合等多维度进行了深度优化。通过单位算力吞吐和分布式计算优化，能在不对硬件进行大规模投入的前提下，显著提升TensorFlow 模型的训练性能。

获取白皮书

1.2.3 有关性能和基准测试结果的更完整信息，请访问 https://www.intel.cn/content/www/cn/zh/cloud-computing/optimized-practice-meituan-tensorflow-xeon.html

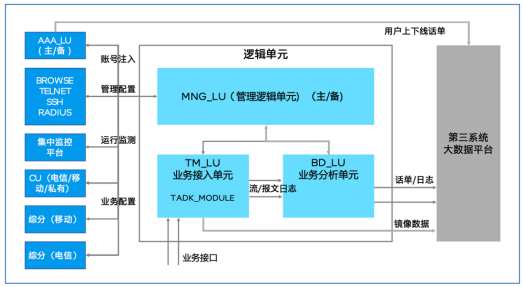
一站打通加密流量提取、建模与推断

英特尔® TADK 助绿网打造智能 DPI 检测方案，保障高性能报文处理

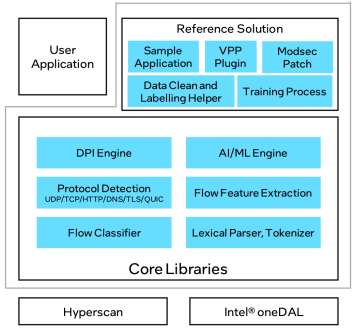
2022 年

加密流量分析优化

智能深度报文检测



绿网固网 DPI 处理流程图



英特尔® 流量分析开发工具套件(TADK) 架构

96%[1]

整体分类准确率超过

TCO 有效降低

高效挖掘流量数据价值

在线推理阶段，TADK 的在线处理模块与在线推理模块会结合初始化时加载的 ML 模型，对每个流进行推断并按流输出推断结果

离线训练层面，英特尔® TADK 支持基于流的业务分类，对每个流打上业务标签后，可利用其提供的离线训练工具生成模型

利用至强® 可扩展处理器的强大通用算力，及其内置的英特尔® AVX-512 等创新技术，实现对并行计算和 AI 应用的兼顾

intel XEON PLATINUM

英特尔® AVX-512

英特尔® oneDAL

英特尔® 流量分析开发工具套件（英特尔® TADK）

在 DPI 产品中融合 AI 能力之后，我们很好地解决了加密流量的分析问题，但随之而来的是 AI 应用带来的巨大性能消耗。为此，我们采用了英特尔的 TADK 方案，该方案支持我们一站式解决加密流量的特征提取、学习建模、在线推断等问题，加速 AI 性能，这让我们能够为用户提供更高效的 DPI 分析能力支撑。

程波
研发经理
绿网

获取白皮书

1 有关性能和基准测试结果的更完整信息，请访问 https://www.intel.cn/content/www/cn/zh/customer-spotlight/cases/greenet-tadk-intelligent-dpi-detection-solution.html

基础设施小变化带来推荐系统大爆发

至强® 可扩展平台助搜狐提升推荐业务系统 AI 推理性能与投资回报

AI 推理性能优化

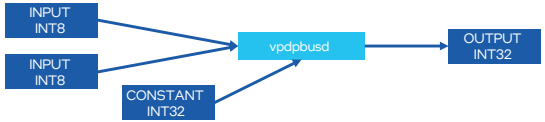
推荐系统

2022 年

50%[1]
相较至强® E5-2650v4，运行在至强® 金牌 6330的 DeepFM 模型时延降低接近

3.4倍[2]
运行在至强® 金牌 6330 + OpenVINO™ 工具套件下的 ResNet50 模型带宽提升高达

3.4倍[3]
运行在至强® 金牌 6330 + OpenVINO™ 工具套件下的 gRPC 模型性能提升高达

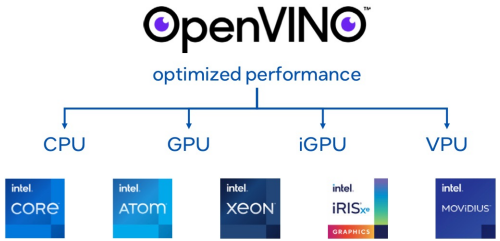


基于英特尔® DL Boost 的 INT8 卷积运算流程

跨英特尔® 架构硬件扩展计算机视觉和非视觉工作负载的 OpenVINO™ 工具套件，可进一步优化 AI 推荐系统的性能表现

利用该处理器内置的英特尔® DL Boost 技术，能更充分地发挥其计算潜能，提升 INT8 模型的推理性能

第三代英特尔® 至强® 可扩展处理器在性能和内存容量上均有显著提升，可为搜狐推荐系统提供可靠的算力支持



OpenVINO™ 工具套件可支持的 XPU 硬件组合

OpenVINO™
英特尔® DL Boost

第三代英特尔® 至强® 可扩展处理器带来了可观的性能提升，以及领先的技术特性，这支持我们在不大幅改变现有基础设施架构的前提下，显著提升应用系统的性能，同时提高投资回报。我们将与英特尔进一步探索在更多领域的合作，以融合双方的创新能力，构建下一代的数据中心基础设施，在性能、敏捷性、扩展性等方面实现更大的优化。

王帅
大数据中心总经理
搜狐

获取白皮书

1.2.3 有关性能和基准测试结果的更完整信息，请访问 https://www.intel.cn/content/www/cn/zh/customer-spotlight/cases/sohu-acclte-load-balancing-encrypt-decrypt-prcss.html

破解人工老片修复三大挑战

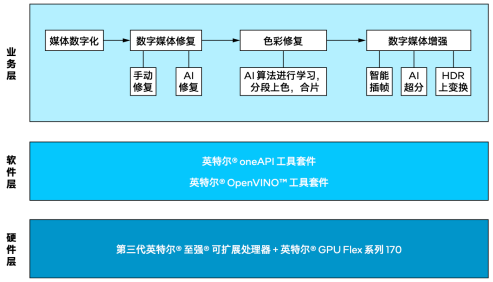
英特尔® 架构 CPU 与 GPU 助当虹科技打造一站式、全流程 AI 老片修复系统

当虹科技 Arcvideo Tech | intel

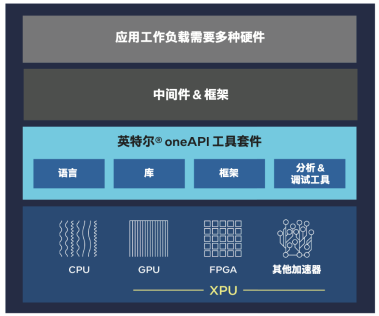
2022 年

GPU AI 推理加速

AI 老片修复



当虹科技基于英特尔® 数据中心 GPU Flex 系列 170 的 AI 老片修复系统



英特尔® oneAPI 工具套件：统一的跨架构编程模型

10-20 小时[1]
一部老片完整的修复时间从 2-3 个月缩短至

上百倍[2]
老片修复效率提升

高效智能一站式全流程

使用英特尔® 分发版 OpenVINO™ 工具套件进行 AI 模型算子优化，加速 AI 推理，进一步提升老片修复效率

用英特尔® oneAPI 工具套件中的 DPC++ 兼容性工具，将多个内核函数从 CUDA 迁移到 SYCL，使代码能在 CPU 和 GPU 间无缝切换

引入英特尔® 至强® 可扩展处理器及英特尔® 数据中心 GPU Flex 系列 170，为 AI 老片修复系统提供算力支持

intel XEON | intel DATA CENTER GPU FLEX SERIES | oneAPI | OpenVINO™

为满足高速发展的传媒行业对老片修复的需求，我们与英特尔合作，在第三代英特尔® 至强® 可扩展处理器、英特尔® 数据中心 GPU Flex 系列 170、英特尔® oneAPI 工具套件和英特尔® 分发版 OpenVINO™ 工具套件的软硬件协同支持下，打造了智能高效的一站式和全流程 AI 老片修复系统。我们的系统广泛适用于老片修复场景，包括老电影、历史资料纪录片、国家文化数据音视频素材等老片资源的高效智能翻新与修复。

黄进
首席技术执行官（CTO）
当虹科技

获取白皮书

1.2 有关性能和基准测试结果的更完整信息，请访问 https://www.intel.cn/content/www/cn/zh/customer-spotlight/cases/arcvideo-tech-efficient-ai-system-old-film-repair.html

让作业批改从“人工”走向“人工智能”

一起 | intel

至强® 可扩展平台助一起教育打造高质量作业 AI 解决方案

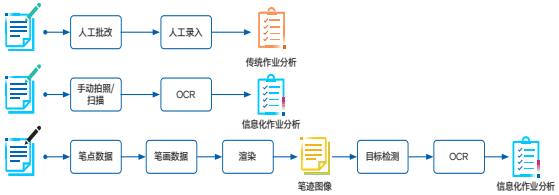
AI 推理加速　OCR

2022 年

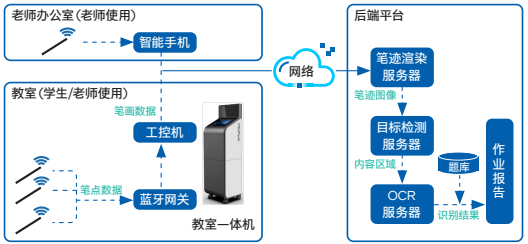
4%[1]
与原生框架相比，OpenVINO™（FP32）让方案推理效果提升约

12.5倍[2]
将数据类型做 INT8 量化后，方案推理性能与原生框架相比提升高达约

0.32%[3]
在召回率没有损失的情况下，精度损失仅为



从人工到人工智能——作业批改模式的演进



一起教育科技高质量作业解决方案架构示意图

利用OpenVINO™工具套件，借助算子融合等操作，进一步加速在至强® 可扩展处理器上的推理性能

通过至强® 可扩展处理器内置的英特尔® 深度学习加速技术对 INT8 低精度数据格式的良好支持，实现推理加速

引入英特尔® 至强® 可扩展处理器，为新方案提供可靠的算力支持

intel Xeon GOLD　英特尔® DL Boost　英特尔® AVX-512　OpenVINO™

通过对作业内容、老师批改痕迹等信息的高效采集与转化，新的作业解决方案能帮助老师更好地掌握学生的学习情况，并以此制定合理的教学计划，提升教学质量。 英特尔® 至强® 可扩展处理器与 OpenVINO™ 工具套件这一组合为方案提供了强有力的计算与 AI 加速能力，使方案在精度和实时性上的表现均达到甚至超越传统技术方案。

吕俊
AI 技术总监
一起教育科技

获取白皮书

1.2.3 有关性能和基准测试结果的更完整信息，请访问 https://www.intel.cn/content/www/cn/zh/customer-spotlight/cases/17-education-technology-high-quality-homework.html

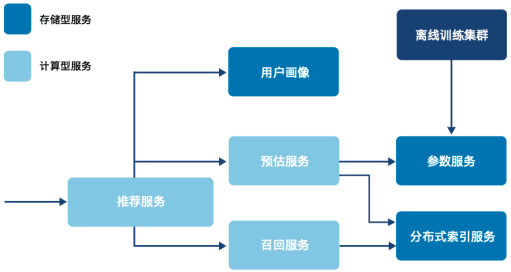
打破存储架构与介质局限

至强® 可扩展处理器 + 傲腾™ 持久内存助火山引擎优化推荐系统存储架构

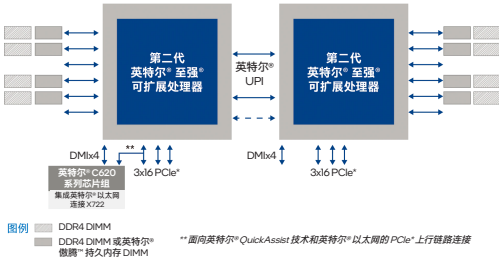
火山引擎 | intel

存储架构优化 | 推荐系统

2021 年



计算与存储分离的推荐系统架构



基于英特尔® 傲腾™ 持久内存的新方案配置

16%[1]
达到相同 TPS 性能的情况下，服务器总成本降低了

20%[2]
在原有成本不变的情况下，性能提升了

4.5TiB[3]
原先硬件架构最大支持1TiB的存储，现在最高可支持

使用持久内存开发套件，帮助应用直接访问持久内存，
无需经过文件系统的页高速缓存系统、 系统调用和驱动

引入英特尔® 傲腾™ 持久内存，以 App Direct 模式，助力
火山引擎推荐系统实现高性能、低延迟和持久存储

利用第二代英特尔® 至强® 可扩展平台
为火山引擎全新存储方案提供底层算力支撑

intel XEON PLATINUM | intel OPTANE PERSISTENT MEMORY

持久内存开发套件（PMDK）

推荐系统对我们的业务至关重要，英特尔® 傲腾™ 持久内存的加持为我们推荐系统存储架构的优化带来了新的思路，让我们在提升性能的同时实现了成本节约目标。

柴君钧
高级工程师
火山引擎

获取白皮书

1.2.3 有关性能和基准测试结果的更完整信息，请访问 https://www.intel.cn/content/www/cn/zh/analytics/optane-memory-volcano-optimize.html

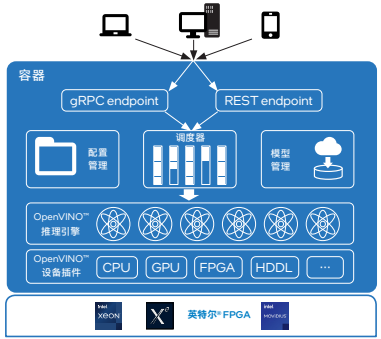
AI 云服务部署+优化一步到位

CDS 首云在 K8S 容器平台上导入 OpenVINO™ Model Server

CDS 首云 | intel

深度学习推理 | AI 云服务 | 2021 年

CDS 首云 AI 云服务方案架构



OpenVINO™ Model Server 架构

2.4[1]倍

与 TF Serving 相比，AI 业务并发接入能力提升了

30[2]毫秒

不良视频内容并发接入检测时延均低于

高可用性

摆脱单一框架限制，并帮助用户降低 TCO

借助 OpenVINO™ Model Server 对主流深度学习框架及对英特尔® 架构基础设施的优化，强化 AI 云服务能力

利用 OpenVINO™ Model Server 与 K8S 的良好集成，实现 AI 模型的快速部署、维护和扩展

基于至强® 可扩展平台提供的云和 AI 负载加速能力

intel XEON GOLD

OpenVINO™ Model Server

OpenVINO™ Model Server 的引入，在帮助我们进一步简化部署流程、提升最终用户使用便捷性的同时，也使我们 AI 云服务方案的生产性能获得了显著提升。

赵二城
架构师
CDS 首云

获取白皮书

1.2 有关性能和基准测试结果的更完整信息，请访问 https://www.intel.cn/content/www/cn/zh/analytics/cds-accelerates-aiaas-deployment.html

深度学习落地要兼顾性能与安全

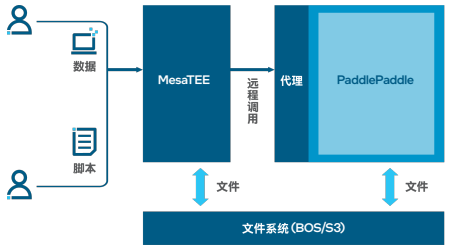
至强® 可扩展平台助百度飞桨加速量化模型并强化隐私数据保护

飞桨 PaddlePaddle | intel

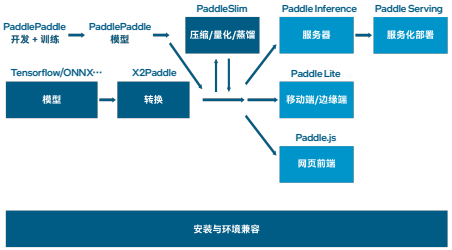
模型量化/多源数据聚合

深度学习

2021 年



PaddlePaddle 与 MesaTEE 的联动



产出量化模型并在英特尔® 架构平台上部署

4倍[1] 算力提升

25%[2] 内存要求降至

至强® 可扩展处理器的嵌入式加速器使系统

3.56倍[3]

飞桨搭载至强® 铂金 8358 处理器后，ResNet50 INT8 推理吞吐量是 FP32 的

更加完善

借助英特尔® SGX，MesaTEE 为机密深度计算提供的保护

借助英特尔® SGX 技术，将敏感的程序代码和数据加载到受 CPU 保护的内存“飞地”中，从而强化敏感数据的安全防护

采用英特尔® oneAPI 工具套件集成的多平台下算子的 JIT 编码库，助开发者在不同架构上灵活调用 oneAPI 算子的即时代码通用接口

借力至强® 可扩展处理器内置 AI 加速及模型量化，显著提升模型推理速度，助开发者高效轻松部署深度学习模型

intel Xeon PLATINUM

oneAPI

英特尔® SGX
英特尔® DL Boost
英特尔® AVX-512

百度飞桨结合第三代英特尔® 至强® 可扩展处理器强劲的算力和内置 AI 加速功能，在英特尔® oneAPI 工具套件的助力下，通过量化模型和加速，显著提升深度学习推理效率；同时，基于英特尔® SGX 的支持，飞桨实现了与 MesaTEE 的对接，引入了机密计算能力，从而帮助更多企业以安全可信的方式为深度学习模型提供更多源的数据。

获取白皮书

1.2.3 有关性能和基准测试结果的更完整信息，请访问 https://www.intel.cn/content/www/cn/zh/now/data-centric/creating-speed-security-leverage-points-baidu.html

构建一站式全功能云上AI开发平台

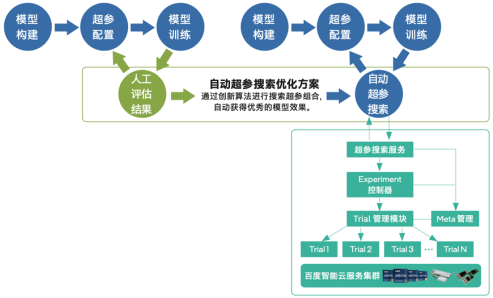
百度智能云 BML | intel

至强® 可扩展处理器内置 AI 加速，助百度 BML 优化用户开发体验

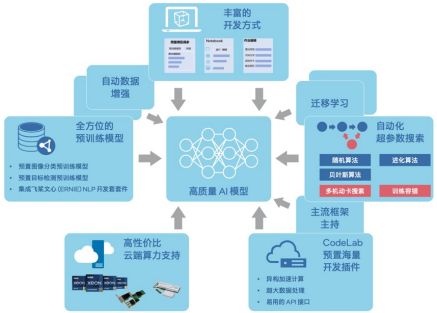
2021 年

算法/模型优化与部署

定制化 AI 开发



BML平台自动超参搜索优化方案



多种创新特性助力开发者构建高质量 AI 模型

- 6倍+[1] 与其他引擎相比，性能提升最高达
- 小时级[2] 数据采集
- 70%[3] 数据标注量降低
- 80%[4] 自动数据清洗/扩充的人工处理成本降低
- 90%[5] 数据需求量降低

采用英特尔® AVX-512 指令集，通过强大的 SIMD 指令大幅提升处理器在密集计算任务中的性能表现

引入英特尔® DL Boost，通过对 INT8 等低精度数据格式的出色支持，加速 BML 平台 AI 推理等任务的执行效率

采用英特尔® 至强® 可扩展平台为百度机器学习(BML)平台提供坚实的算力支持

intel XEON GOLD

英特尔® AVX-512

英特尔® DL Boost

第三代英特尔® 至强® 可扩展处理器的引入，及英特尔® 深度学习加速等带来的 AI 加速能力，有效提升了 BML 平台的整体性能，让一站式 AI 开发专业化、定制化服务更易用、更流畅、更灵活，并通过平台提供的优化能力，使 AI 应用在实际业务场景中也获得了更佳表现。

获取白皮书

1.2.3.4.5 有关性能和基准测试结果的更完整信息，请访问 https://www.intel.cn/content/www/cn/zh/analytics/artificial-intelligence/baidu-bml-ai-development.html

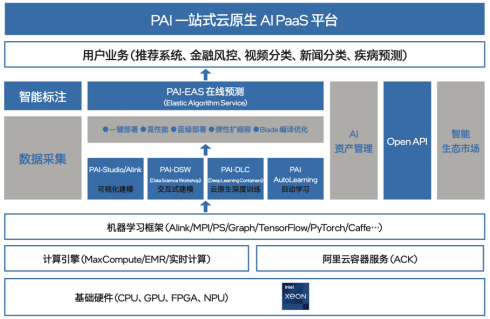
构建一站式云原生 AI PaaS 平台

至强® 可扩展处理器助力阿里云 PAI 平台提升训练和推理效率

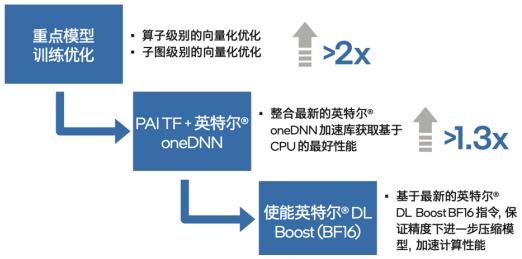
阿里云 | intel

2021 年

模型训练&推理加速 | **机器学习**



使用英特尔® DL Boost（BF16）可大幅提升推理效率



PAI 平台训练加速[4]

1.83倍[1]
PAI 平台 Bert 模型推理性能提升达

1.58倍[2]
PAI 平台 TextCNN 模型推理性能提升达

1.3倍[3]
在保证精度的情况下进一步压缩模型，计算性能提升达

应用英特尔® one DNN 大幅提升 PAI 平台的集成能及性价比，降低优化和使用的门槛

引入英特尔® DL Boost 大幅提升 PAI 平台推理能力，结合英特尔® AVX-512 提供的通用底层算子，加速 NLP 模型推理

采用第三代英特尔® 至强® 可扩展处理器为阿里云第七代高主频云服务器提供强劲算力和高可扩展性

intel XEON PLATINUM

英特尔® AVX-512
英特尔® DL Boost
英特尔® oneDNN

阿里云机器学习平台 PAI 是一站式云原生 AI PaaS 平台。PAI 与英特尔的合作贯穿了从硬件到驱动到软件，第三代英特尔® 至强® 可扩展处理器为 PAI 平台提供了强大的算力和优秀的可扩展性，英特尔® AVX-512 和英特尔® DL Boost 为我们的机器学习平台提供了非常强的优化空间，大幅提升了公有云和专有云客户使用阿里云智能机器学习平台 PAI 的性价比。

黄博远
产品负责人
阿里云智能机器学习平台PAI

获取白皮书

1.2.3.4 有关性能和基准测试结果的更完整信息，请访问 https://www.intel.cn/content/www/cn/zh/cloud-computing/xeon-ai-alibaba-cloud-pai-optimize.html

用 AI 缺陷检测实现产能与品质双赢

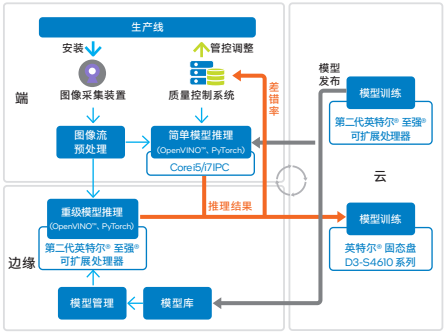
CATL 宁德时代 | intel

宁德时代导入至强® 可扩展平台构建“云-边-端” AI 缺陷检测方案

CV+DL+ML

电池缺陷检测

2020 年



全新工业视觉平台系统架构图

	内容相似度高	内容相似度低
数据集少	修改预先训练的源模型中最后几层或者全连接层（FC 层）的输出类别，训练自己的目标模型 宁德时代选用方案	冻结预训练源模型的初始层（比如 K 层），然后训练剩余的 N-K 层，主要对较高层进行重新训练，得到目标模型
数据集大	保留预训练源模型的结构和初始权重，重新训练自己的目标模型	根据自己的数据从头开始训练神经网络，得到自己的目标模型

不同条件下的模型训练选用方案建议

零漏检
达到预定目标

单工序 400FPS
图像处理速度达到预定目标

更优
推理性能、训练准确率与检出率

利用面向英特尔® 架构优化的 PyTorch 及其内置的英特尔® one-DNN，提升深度学习框架性能

以 OpenVINO™ 进行模型转换，进一步提升AI 推理性能

借英特尔® 至强® 可扩展处理器内置 AI 加速能力，为方案提供强劲算力支持和更优 AI 推理能力

intel XEON | intel CORE i5 | intel CORE i7

OpenVINO™
英特尔® oneDNN
面向英特尔® 架构优化的 PyTorch

基于 AI 技术的缺陷检测方案，是我们用以提升动力电池产能和质量的重要平台。英特尔内置了 AI 加速能力的至强® 可扩展平台等一系列软硬件产品的引入，以及来自英特尔的全方位技术支持，为整个方案实现统一部署和管控，并达成快速、准确的检测处理能力做出了重要的贡献。

潘伟伟
人工智能高级工程师
宁德时代

获取白皮书

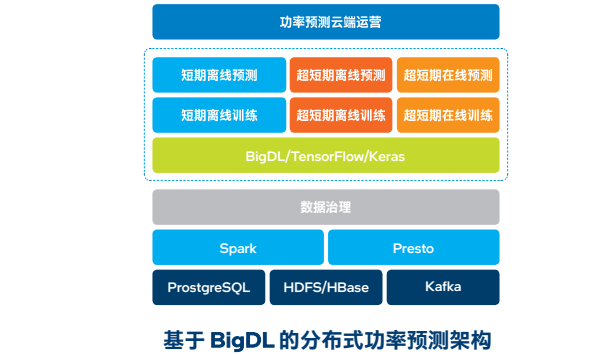
借力 AI 领跑绿色能源发展之路

GOLDWIND SE 金风慧能 | Smart Energy Services 智慧能源服务 | intel

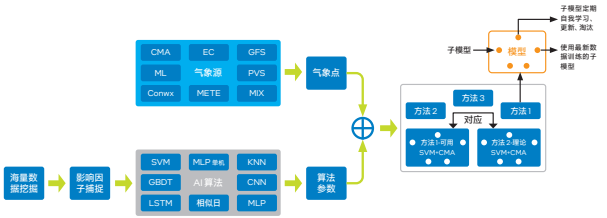
英特尔统一大数据分析 + AI 平台助金风慧能打造新能源智能功率预测方案

时序数据分析　智能功率预测

2020 年



基于 BigDL 的分布式功率预测架构



结合气象预报数据的多模型组合预测方案架构

20%[1]
新方案将预测准确率提升超过

100个[2]
新方案验证将扩大到更多光伏场站

显著提升
在风电场中引入新方案，发电效率

构建分布式、结合气象预报的多模型组合功率预测系统，有效提升新方案部署效率及可扩展性，减少成本并提高预测准确率和稳定性

引入 BigDL 打造分布式架构，并针对预测数据的时序特性进行有针对性的优化

采用英特尔® 至强® 可扩展处理器，为智能功率预测方案提供强劲算力支撑

intel XEON GOLD　BigDL

BigDL 所提供的统一端到端架构，及其在时序数据分析方面的优势，不仅使我们结合气象预报数据的多模型组合功率预测方案具有更敏捷的部署效率，而且在预测准确率、稳定性方面也能获得较大提升。

张利
首席架构师
金风慧能

获取白皮书

* BigDL：指 BigDL 2.0，其合并原始 BigDL 和 Analytics Zoo。

1.2 有关性能和基准测试结果的更完整信息，请访问 https://www.intel.cn/content/www/cn/zh/analytics/artificial-intelligence/create-power-forecasting-solutions.html

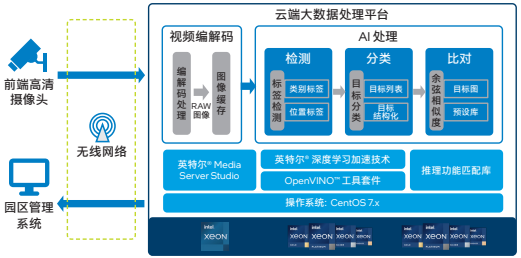
用“慧眼”推进园区管理智能化

云创大数据基于至强® 平台打造端到端智慧园区视频监控方案

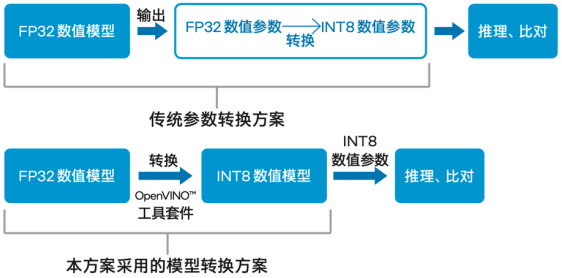
云创大数据 www.cStor.cn | intel

深度学习推理 | 图像识别

2020 年



云创大数据智慧园区视频监控方案采用的端到端架构



传统参数转换方式与 OpenVINO™ 工具套件模型转换方式对比

100%[1]
方案在主要监控应用场景中的准确率已近

减半[2]
方案开发周期从原本预估 6 个月缩短至实际 3 个月

秒级
百万张图像对比和匹配速度，及园区灾情告警速度均达

利用第二代英特尔® 至强® 可扩展处理器内置的深度学习加速技术，显著提升转换后的 INT8 模型推理速度

借助 OpenVINO™ 提升系统模型转换和优化能力

导入英特尔® 视频分析参考设计方案，提供视频编解码加速

intel XEON PLATINUM

英特尔® DL Boost

英特尔® Media SDK

OpenVINO™

英特尔® 视频分析参考设计方案

借助英特尔® 至强® 可扩展平台集成的 AI 加速能力，及英特尔® 视频分析参考设计方案，我们的方案不但在图像识别速度和精度上表现卓越，还让我们得以迅速抢占市场先机。

曹骝
人工智能产品研发总监
云创大数据

获取白皮书

1.2 有关性能和基准测试结果的更完整信息，请访问 https://www.intel.cn/content/www/cn/zh/artificial-intelligence/video-surveillance-solution-for-intelligent-industrial-park.html

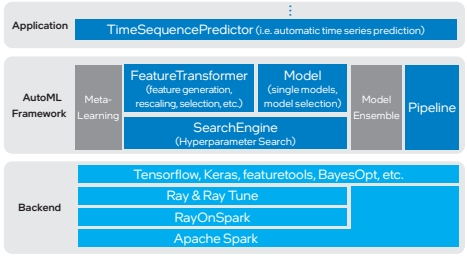
打造高效易用一站式云上机器学习平台

腾讯云 | intel

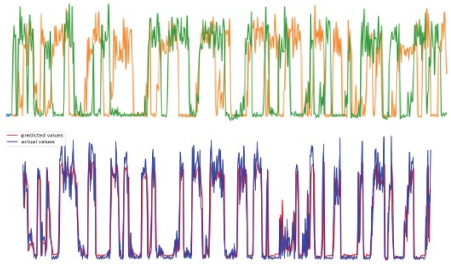
BigDL 助腾讯云智能钛机器学习平台强化 AutoML 特性

2020 年

AutoML　时序预测



BigDL 中的 AutoML 框架



基于传统方法与 BigDL AutoML 的时序预测对比

更加容易
BigDL AutoML 框架使训练时序分析模型的过程

吻合度更高
基于BigDL AutoML 的预测值和实际值

便捷轻松
用户创建端到端 AI 应用

基于 BigDL 强化平台的 AutoML 特性，提供方便易用的接口，让用户能轻松进行时序数据分析及机器学习建模

利用英特尔® oneDNN 的优化和加速，充分释放至强® 可扩展处理器的模型训练和推理性能

采用第二代英特尔® 至强® 可扩展平台及其内置的英特尔® DL Boost，提升平台的深度学习负载性能

intel XEON　BigDL　英特尔® oneDNN　英特尔® DL Boost

英特尔与腾讯开展深度合作，将 BigDL 集成到腾讯云智能钛机器学习平台，使平台获得了 AutoML 高级特性，让 AI 初学者也能轻松使用。利用已整合 BigDL 的智能钛机器学习平台，用户可对各种数据源、组件、算法、模型和评估模块进行组合，使得算法工程师和数据科学家在其之上能够方便地进行模型训练、评估和预测。

获取白皮书

- BigDL：指 BigDL 2.0，其合并原始 BigDL 和 Analytics Zoo。
- 英特尔® MKL-DNN已改名为英特尔® oneDNN。

有关性能和基准测试结果的更完整信息，请访问 https://www.intel.cn/content/www/cn/zh/service-providers/analytics-zoo-helps-tencent-cloud-improve-ti-ml-platform-performance.html

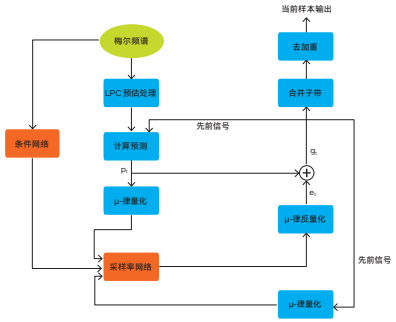
定制化模型提升实时语音合成性能

腾讯云小微 | intel

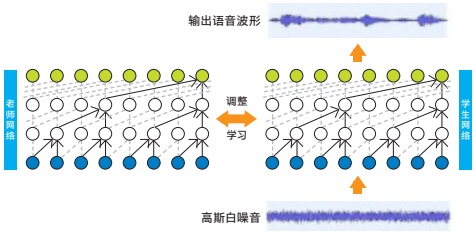
第三代英特尔® 至强® 可扩展处理器助腾讯云小微平台升级

文本转语音 | 智能语音服务

2020 年



定制化 WaveRNN 声码器模型架构图



Parallel WaveNet 模型架构图

1.89[1]倍
经 BF16 优化，在 MOS为4.4时，定制化 pWaveNet 模型实现性能增益

1.54[2]倍
经 BF16 优化，在 MOS为 4.5 时，定制化 WaveRNN 模型实现性能增益

显著提升
云小微平台语音合成的实时率与吞吐量

配合英特尔® oneAPI 深度神经网络库，
进一步释放硬件加速能力，提升处理性能

利用该处理器内置的英特尔® AVX-512
及英特尔® DL Boost（BF16）减少内存访问量，提升语音合成速度

采用面向四路和八路服务器的
第三代英特尔® 至强® 可扩展处理器，为平台提供算力支持

intel XEON

英特尔® AVX-512
英特尔® DL Boost
（BF16）

英特尔® oneAPI 深度神经网络库（oneDNN）

通过不断提升的吞吐量以及更高的实时性，云小微平台能够为企业级应用提供高质量智能化语音服务。得益于英特尔先进的软硬件技术支撑，基于第三代英特尔® 至强® 可扩展处理器的定制化解决方案，使云小微平台的语音合成性能得以更充分释放。

田乔
高级研究员
腾讯云

获取白皮书

1.2 有关性能和基准测试结果的更完整信息，请访问 https://www.intel.cn/content/www/cn/zh/now/csp-abm-tencent/optimization-of-customized-vocoder.html

兼顾产品缺陷检测与不良根因分析

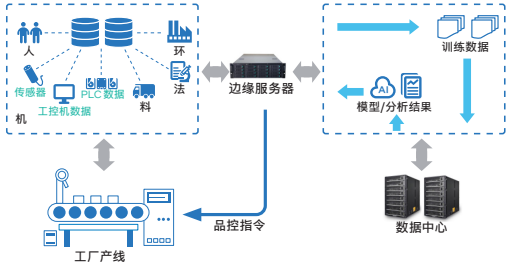
京东方基于至强® 可扩展平台打造云边协同智能化品质管控方案

BOE | intel

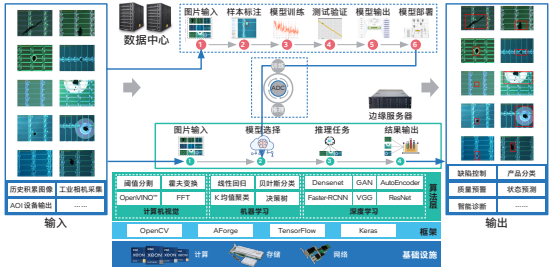
2020 年

云边协同

AI 瑕疵检测/不良根因分析



基于云边协同的品质控制解决方案



京东方 AI 缺陷检测系统架构及工作流程示意图

96%[1] 检测准确率

70%[2] 人工替代率

AI 缺陷检测系统

56%[3] 大数据不良根因分析系统使产线效率提升

>60%[4] 检测工艺和维修工艺的操作人员也可因此减少

用 OpenVINO™ 工具套件为缺陷检测提供软件调优支持

采用至强® 可扩展处理器内置的英特尔® AVX-512 技术
为深度学习推理任务中的密集计算提供特定硬件加速支持

引入英特尔® 至强® 可扩展处理器作为边缘服务器的核心计算引擎

intel XEON GOLD

intel XEON SILVER

英特尔® AVX-512

OpenVINO™

利用边缘计算和云边协同，让我们工业互联网解决方案中的数据预处理、缺陷检测等应用在生产一线的降本增效中发挥了更大作用。这其中，来自英特尔的高性能处理器平台和软件加速库，为新方案、新应用的部署和运行提供了可靠的支持。

李昭月
智能工厂解决方案技术专家
京东方

获取白皮书

1.2.3.4 有关性能和基准测试结果的更完整信息，请访问 https://www.intel.cn/content/www/cn/zh/customer-spotlight/cases/edge-computing-capability.html

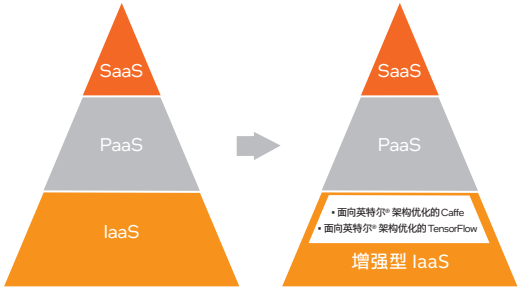
AI 开发上云一步到位

金山云云计算进阶：更优 AI 软硬件“打包出售”

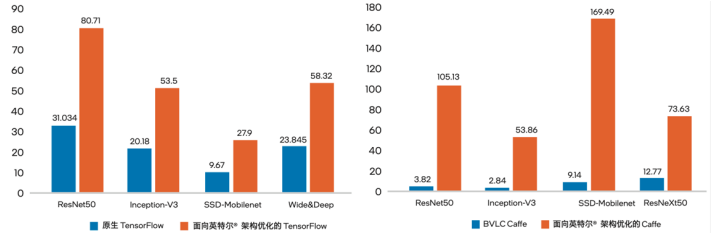
金山云 | intel

2019 年

增强型 IaaS | **AI 即服务**



面向 AI 研发的增强型 IaaS 云服务



在不同深度神经网络的 AI 前向传播中使用原生 TensorFlow 和 Caffe，和面向英特尔® 架构优化的 TensorFlow、Caffe 的性能对比

2.89[1] 倍
优化版 TensorFlow 与原生版相比性能提升高达

27.5[2] 倍
优化版 Caffe 与原生版比，性能提升高达

成本降低
减少用户在 AI 系统部署和优化上的投入

将彼此优化的软硬件打包，构建面向 AI 开发者的增强型 IaaS

引入内含 MKL-DNN 的、面向英特尔® 架构优化的 TensorFlow 和 Caffe 框架，充分释放硬件潜能

引入至强® 可扩展平台产品组合，提升云实例基础算力

intel XEON PLATINUM

intel OPTANE PERSISTENT MEMORY

面向英特尔® 架构优化的 Caffe TensorFlow

英特尔® oneDNN

至强® 可扩展平台的部署，以及面向英特尔® 架构优化的 Caffe、TensorFlow 等框架的引入，帮助我们打造了增强型 IaaS，大幅降低用户在系统部署和优化上的投入，让他们能更多关注 AI 业务本身。

杨峰
云计算研发总监
金山云

获取白皮书

1.2 有关性能和基准测试结果的更完整信息，请访问 https://www.intel.cn/content/www/cn/zh/service-providers/kingsoft-cloud-upgrades-cloud-service-for-ai-developers.html

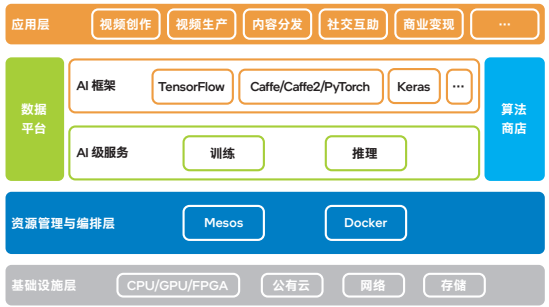
让视频服务实现全流程智能化

爱奇艺导入英特尔® 架构软硬件组合优化深度学习云平台

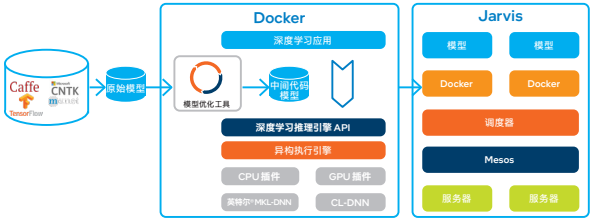
爱奇艺 iQIYI 悦享品质 | intel

深度学习训练+推理　智能视频服务

2019 年



爱奇艺 Jarvis 深度学习云平台的架构图解



基于 OpenVINO™ 工具套件的 Jarvis 平台推理优化过程

1 **5倍左右** 实时弹屏显示的推理速度提升达

2 **6倍左右** 涉黄内容检测的推理效率提升达

3 **11倍** 文本检测应用中的推理性能提升达

引入 OpenVINO™ 工具套件，提升 Jarvis 平台的推理效率

导入英特尔® 至强® 可扩展处理器，
并使用 MKL/MKL-DNN 对云平台进行系统级优化

将 AI 与云计算结合，构建基于云的深度学习平台 Jarvis

intel XEON GOLD

OpenVINO™ 英特尔® one-DNN

英特尔® 至强® 可扩展处理器及 OpenVINO™ 工具套件不仅使我们的深度学习云平台获得了更强的算力，也使深度学习推理效率得到了显著的提升，进而让视频服务变得更为高效和智能。

张磊
研究员
爱奇艺

获取白皮书

1.2.3 有关性能和基准测试结果的更完整信息，请访问 https://www.intel.cn/content/www/cn/zh/analytics/artificial-intelligence/optimize-the-efficiency-of-dl-inference-and-create-smarter-video-services.html

为飞桨增添“加力推进器”

百度飞桨 INT8 方案借英特尔® 深度学习加速技术提升推理效率

深度学习推理　智能图像分析

2019 年

2-3倍[1]
使用 INT8 时的推理速度是 FP32 时的

1%以内[2]
INT8 与 FP32 的深度学习模型推理准确度差值

能效更优
提升推理效率、降低功耗和部署复杂度

模型	数据集	FP32 准确率	INT8 准确率	准确率差值
ResNet-50	Full ImageNet Val	76.63%	76.23%	0.40%
MobileNet-V1	Full ImageNet Val	70.78%	70.47%	0.31%

FP32 和 INT8 推理准确度结果比较

模型	数据集	FP32 吞吐量	INT8 吞吐量	INT8/FP32 吞吐量比比率
ResNet-50	Full ImageNet Val	11.54 images/s	32.2 images/s	2.79
MobileNet-V1	Full ImageNet Val	49.21 images/s	108.37 images/s	2.2

FP32 和 INT8 推理吞吐量结果比较

- 在图像识别与分类等场景的深度学习中发挥 INT8 的优势
- 基于 MKL/MKL-DNN 对不同深度学习模型进行特定优化
- 利用第二代英特尔® 至强® 可扩展处理器内置的英特尔® 深度学习加速技术对 INT8 提供更优的支持

intel XEON GOLD

英特尔® 深度学习加速

英特尔® oneDNN

第二代英特尔® 至强® 可扩展处理器的强劲算力及英特尔® 深度学习加速，让飞桨 INT8 方案在不影响推理准确度的情况下，推理速度得以显著提升。

高铁柱
高级经理
百度深度学习平台部

获取白皮书

1.2 有关性能和基准测试结果的更完整信息，请访问 https://www.intel.cn/content/www/cn/zh/analytics/artificial-intelligence/boost-technology-baidu-paddlepaddle-int8-solution-dramatically-speeds-up-deep-learning-inference.html

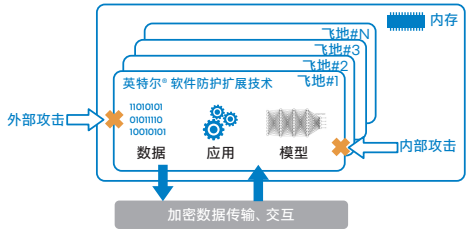
打破数据孤岛 助联邦学习落地

平安科技 PING AN TECHNOLOGY | intel

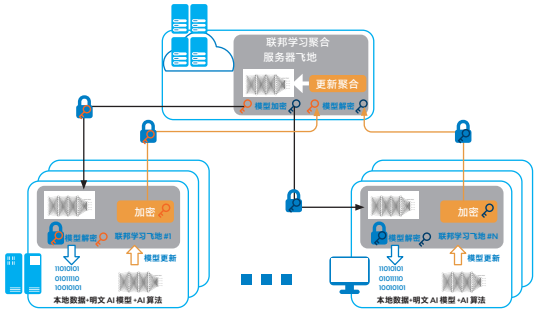
平安科技借助英特尔硬件增强型安全技术保护训练数据和结果

2019 年

联邦学习训练　数据信息安全



英特尔® 软件防护扩展以可信“飞地”来增强数据安全防护



使用英特尔® 软件防护扩展的联邦学习方案

更安全的协作

提供加密方式，支持同态加密等多方安全计算机制

更安全的建模

多方数据不离开本地即可联合建模，中间结果也得到充分保护

多框架支持

支持多种深度学习框架，如TensorFlow，Keras，PyTorch，MXNet等

采用英特尔® 软件防护扩展，在参与各方的系统中构建更为安全可信的“飞地”，以强化训练数据和中间结果的安全

在联邦学习的框架下，吸引多方参与建模和训练，为 AI 应用提供更多来源、更多维度、更为优质的训练数据集

intel XEON

英特尔® 软件防护扩展
（英特尔® SGX）

英特尔® 软件防护扩展，是联邦学习方案中构建硬件可信执行环境的理想之选。它能帮助我们打破数据孤岛，在强化数据安全的条件下，利用更多来源、更多维度的数据，来提升 AI 模型的训练效果。

王健宗 博士
副总工程师
平安科技

获取白皮书

为 AI 训练和超算定制高性能存储

百度智能云全闪对象存储方案导入 QLC 固态盘+傲腾™ 固态盘组合

百度智能云 | intel

高性能存储

AI, 大数据, 高性能计算

2019 年

由 ABC Storage 高性能存储解决方案支持的 AI 训练流程

1 5%以内
文件数据增加 10 倍时，QPS 和时延波动保持在

2 60%
TCO 降低

3 1-3倍
用户业务效率提高

可部署于私有云，专攻 AI 训练、大数据和高性能计算

将久经考验的高性能对象存储引擎引入方案

“全闪” = 傲腾™ 固态盘 + QLC 3D NAND 固态盘

intel XEON GOLD

intel OPTANE SSD

intel 3D NAND SSD

来自英特尔的全闪产品组合，及其与英特尔® 至强® 可扩展处理器的配合，帮助我们的方案在稳定性、IOPS 等方面实现了更优表现，成为应对海量非结构化小文件的得力手段。

百度智能云私有云存储团队

获取白皮书

1.2.3 有关性能和基准测试结果的更完整信息，请访问 https://www.intel.cn/content/www/cn/zh/cloud-computing/baidu-private-cloud-abc-storage-cooperates-with-intel-to-launch-high-performance-all-flash-object-storage-solution.html

用 CPU 替代 GPU，加速图像分析

JD.京东.COM | intel

至强® 平台 + BigDL 助京东实现商品图像数据库高效扩展及分析

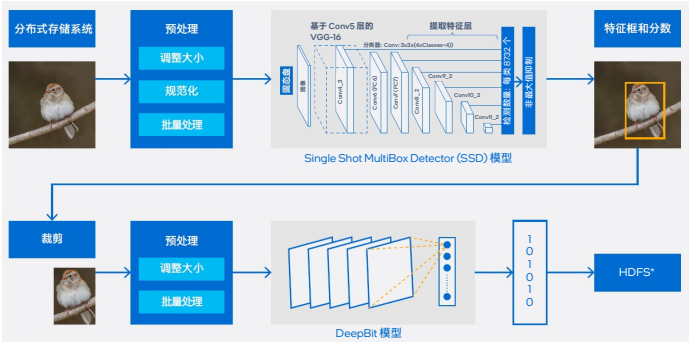
深度学习推理　智能零售

2018 年

3.83[1]倍
与 GPGPU 相比，性能提升高达

横向扩展
基于至强® 平台，技术设施更易扩展

更利于应用开发
基于 GPGPU 的加速和生产部署则难以实现



京东的特征提取工作流程，使用 BigDL 管理用于目标检测的 SSD 模型以及用于特征提取的 DeepBit 模型

- 在 Spark 集群上使用 BigDL 库运行现有 Caffe 模型
- 可通过基于至强® 处理器的服务器进行横向扩展
- 基于英特尔® 架构硬件基础设施，优化算法与模型

intel XEON　BigDL　英特尔® one-DNN

应用英特尔® 架构产品，帮助我们成功地应对了在大数据集群上构建大规模深度学习应用的挑战，与运行独立的 GPU 集群比，提升了性能，降低了成本。

汪振华
京东高级软件工程师

获取白皮书

1 有关性能和基准测试结果的更完整信息，请访问 https://www.intel.cn/content/www/cn/zh/analytics/artificial-intelligence/jd-replacing-gpus-with-cpus-case-study-v2.html

为制造增添“眼”和“脑”

英特尔助力美的构建工业视觉检测云平台

Midea | intel

2018 年

深度学习训练+推理

智能制造

57%[1]
项目部署周期缩短

70%[2]
人工成本减少

124毫秒[3]
推理时间从2秒缩短至

基于“大数据+AI”端到端解决方案，打造工业视觉检测系统，实现敏捷、高性能通用化缺陷检测能力

intel XEON

intel MOVIDIUS

BigDL

实践表明，我们工厂非常欢迎这种务实的工业视觉检测解决方案，它的部署速度快，对产线的生产影响非常低，未来我们会把它复制到我们所有的产线。

黄姚根
设备负责人
美的微波炉工程部

获取白皮书

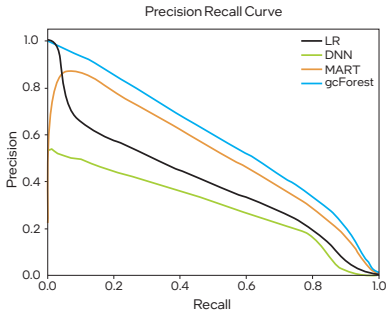
* BigDL：指 BigDL 2.0，其合并原始 BigDL 和 Analytics Zoo。

AI≠深度神经网络

南京大学 LAMDA 携手英特尔，推进深度森林的探索与应用

深度学习　前沿研究

2018 年



深度森林与其他 AI 方法的 PR 曲线对比[2]

非常显著

至强® 处理器使决策树训练速度提升

表现上佳[1]

经海量样本训练对比验证，召回率、准确率

深化合作

成立面向 AI 领域的英特尔® 并行计算中心

利用面向英特尔® 架构的优化，提升训练速度和质量

基于英特尔® 架构硬件基础设施，优化算法与模型

intel XEON 英特尔® 编译器

深度森林更多是需要并行地在多棵树上进行决策计算，在这方面，GPGPU 很难发挥其长处，而这恰恰是 CPU，尤其是拥有多核、高频能力的英特尔® 架构处理器所擅长的地方。

冯霁博士
LAMDA
南京大学

获取白皮书

1.2 有关性能和基准测试结果的更完整信息，请访问 https://www.intel.cn/content/www/cn/zh/analytics/artificial-intelligence/nanjing-university-advance-deep-forest-exploration-and-application-case-study.html

加速快递行业智能化变革

英特尔® 架构 AI 核心软硬件助力韵达提升物流系统运转效率

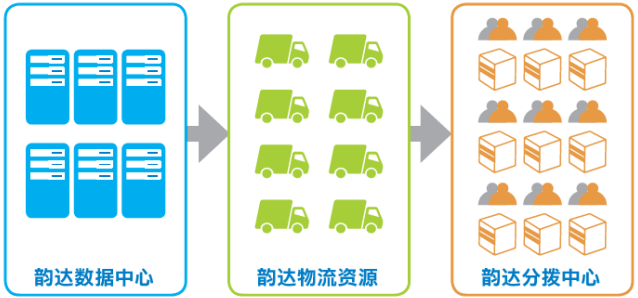
深度学习训练+推理　智能物流

2018 年

行业智能化
提升效率、优化资源、降低成本

大规模部署
京、沪、广、深等一线城市逐步展开

未来合作
更多 AI 应用开发与部署



韵达物流快递三大关键业务环节

运用英特尔® 架构软硬件产品组合，助力三大关键业务环节：

- 在分拨中心实现大小件测量的全流程 AI 处理
- 在资源配置上实施更精准的 AI 件量预测
- 在数据中心内实现基于 AI 分析的数据中心异常检测

intel XEON　BigDL

依托英特尔® 架构 AI 核心软硬件强大的算力和优化，韵达发挥数据优势，在三大关键业务环节构建 AI 应用，极大提升全产业链的效率，降低成本，推动行业智能化变革。

李培吉
韵达快递首席架构师
韵达股份

获取白皮书

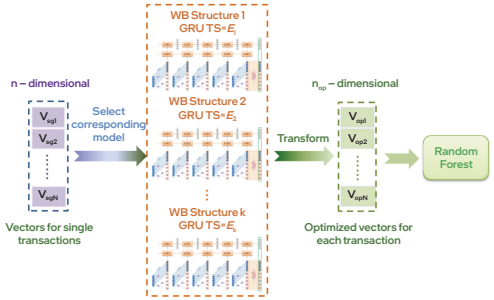
* BigDL：指 BigDL 2.0，其合并原始 BigDL 和 Analytics Zoo。

用“三明治”来反欺诈

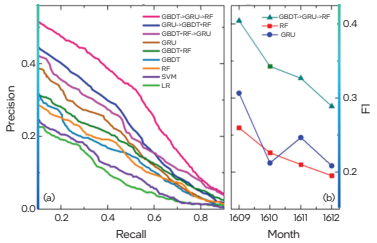
英特尔助力中国银联电子商务与电子支付国家工程实验室
金融反欺诈模型研究与应用

机器学习训练 | 金融反欺诈

2017 年



GBDT→GRU→RF 三明治结构欺诈侦测模型的架构



GBDT→GRU→RF 三明治结构欺诈侦测模型评估效果

创新的反欺诈模型，树立了行业标杆

验证英特尔技术在金融行业的正向作用

将引进更多英特尔技术和产品

对模型中的三层方法分别提供针对性的技术和工具优化

构建 GBDT→GRU→RF 三明治结构欺诈侦测模型

面向英特尔® 架构优化的

英特尔® Python发行版

英特尔为反欺诈模型提供了高性能处理器，为模型各层面提供了优化手段，帮助系统实现了高功效，为人工智能在该领域的应用提供了经验，为深度学习在金融领域场景的落地铺平了道路。

中国银联电子商务与电子支付国家工程实验室

获取白皮书

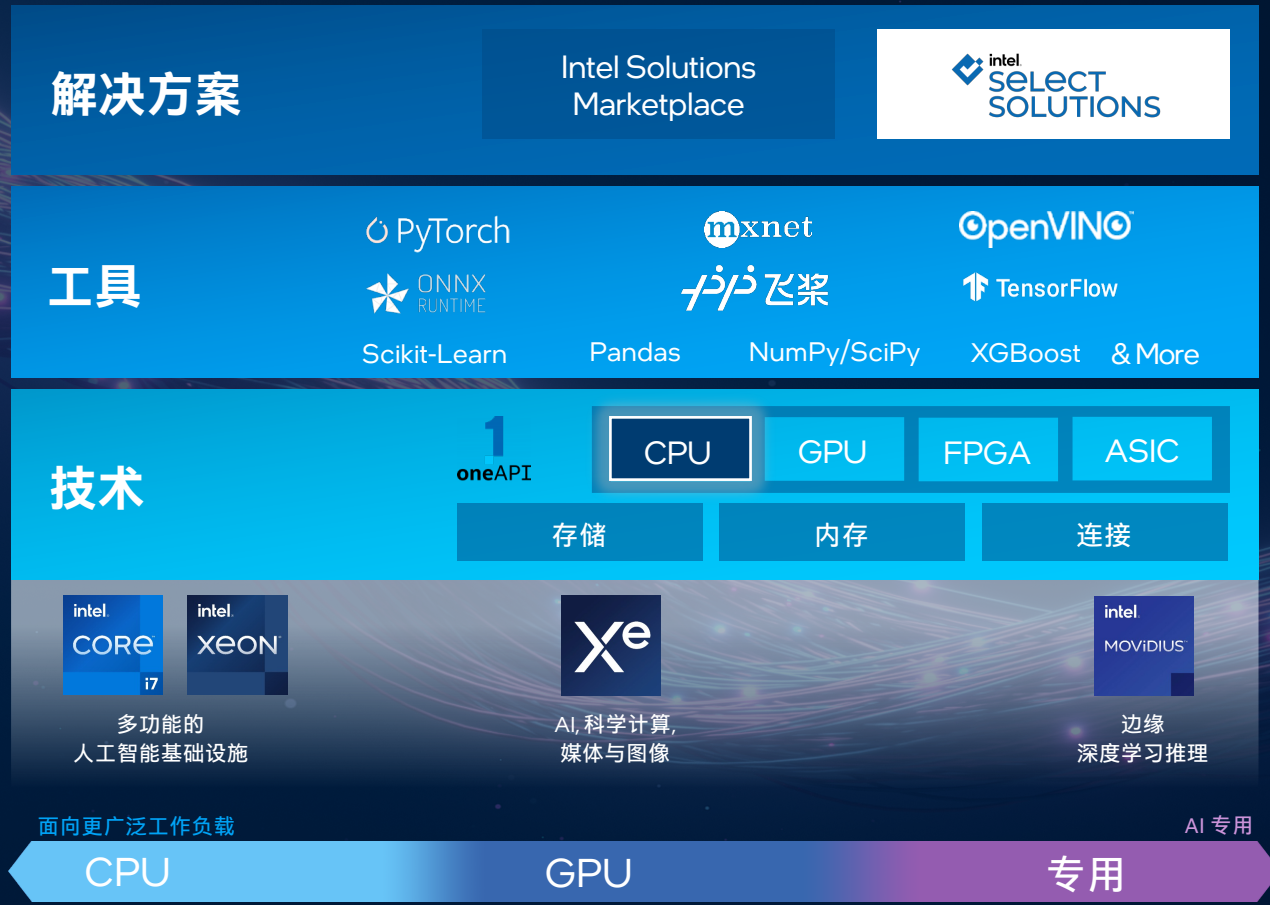
有关性能和基准测试结果的更完整信息，请访问 https://www.intel.cn/content/www/cn/zh/analytics/artificial-intelligence/deep-learning-framework-based-on-the-sandwich-structure-of-financial-fraud-model-research-and-application.html

技术篇：“AI 无处不在”背后的英特尔® 架构基石

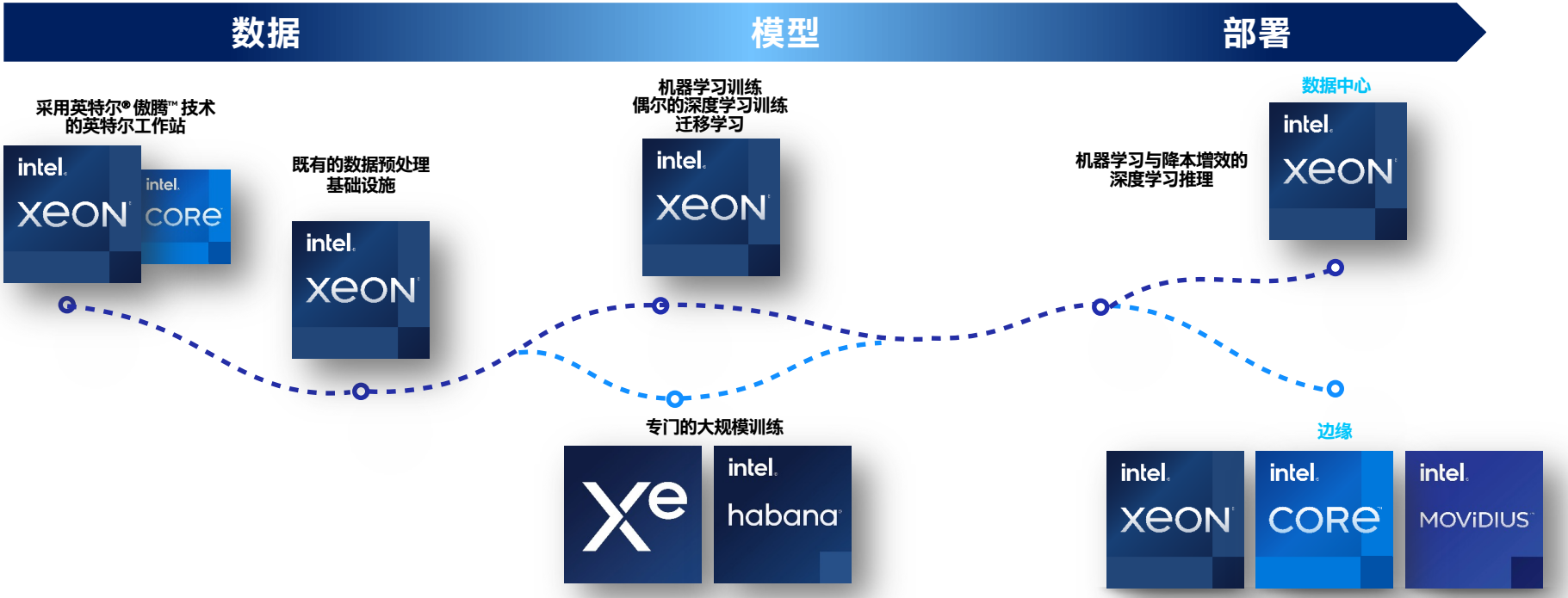
加速 AI 落地
200+ 一站式 AI 方案可选
助力应用快速落地

加速 AI 应用构建
150+ 容器镜像帮助用户
快速构建端到端 AI 数据应用

加速 AI 性能
在 20+ 典型 AI 负载下提供
卓越性能表现



英特尔® XPU 平台：满足 AI 之旅各阶段需求



第四代英特尔® 至强® 可扩展处理器

英特尔® 高级矩阵扩展（英特尔® AMX）

英特尔® 存内分析加速器（英特尔® IAA）

英特尔® 数据流加速器（英特尔® DSA）

英特尔® 动态负载均衡器（英特尔® DLB）

高达1.53倍
平均性能增益[1]
（与上一代产品比较）

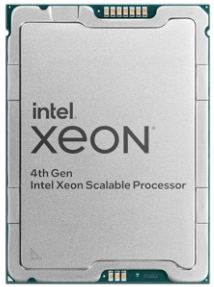
全新内置加速器

80 条 PCIe 5.0 通道

支持 1 至 8 路配置

更高的单核性能
每路多达 60 个内核

英特尔® UPI 2.0
（高达 16 GT/s）



增加三级缓存
（LLC）共享容量

Compute Express Link（CXL）1.1

8 通道 DDR5
传输速率高达 4,800 MT/s（1DPC）
传输速率高达 4,400 MT/s（2DPC）
每路 16 个 DIMM
全新 RAS 功能（增强型ECC、ECS）

高带宽内存（HBM）
（64GB/每路）

经优化的电源模式

英特尔® 数据保护与压缩加速技术（英特尔® QAT）

英特尔® 安全引擎
英特尔® SGX
英特尔® TDX
...

面向 vRAN 的英特尔® 高级矢量扩展

英特尔® 至强® CPU Max 系列（配备 HBM）

高达10倍
PyTorch 实时推理和训练性能提升[2]
启动英特尔® AMX（BF16）时与上一代产品（FP32）的比较结果

创新、设计和交付均坚持工作负载优先

CPU 内置多种加速器

更出色的性能与能效，更好满足客户所需

更全面的机密计算产品组合

1.2 如欲了解更多详情，请访问： https://www.intel.cn/content/www/cn/zh/products/docs/processors/xeon-accelerated/4th-gen-xeon-scalable-processors-product-brief.html

第三代英特尔® 至强® 可扩展平台

intel XEON PLATINUM

10 - 100x
英特尔优化使 TensorFlow 和 Scikit-Learn 性能提升达[1]
(图像识别) (SVC & kNN 预测)

74%
与上一代相比，推理性能提升了[2]
(自然语言处理)

32%
英特尔® 傲腾™ 持久内存 200 系列内存带宽平均提升了[3]
(相较于上一代产品)

intel OPTANE PERSISTENT MEMORY

跨越广泛的机器学习和深度学习模型...

英特尔® 高级矢量扩展 512
(英特尔® AVX-512)

英特尔® 深度学习加速
(英特尔® DL Boost)

英特尔® 软件防护扩展
(英特尔® SGX)

Data Processing
Train-test-split†

Classification
brute force knn†, SVC

Neural Networks
Resnet50-v1.5, SSD-ResNet34, BERT large†

Gradient Boosting Machines
XGBoost

Clustering
Kmeans, dbscan†

Regression
Linear, Logistic, Ridge, elastic-net†

1.2 如欲了解更多详情，请访问：https://www.intel.cn/content/www/cn/zh/artificial-intelligence/ai-acceleration-on-technology.html
3 如欲了解更多详情，请访问：https://www.intel.cn/content/www/cn/zh/customer-spotlight/cases/achieve-greater-insight-from-your-data-optane.html

英特尔® 至强® 可扩展处理器内置 AI 加速能力的演进

内置 AI 加速能力的数据中心级 CPU

第二代至强® 可扩展处理器 (Cascade Lake)
英特尔® DL Boost (AVX-512_VNNI)
全新内存存储层次结构

第三代至强® 可扩展处理器 (Cooper Lake)
英特尔® DL Boost (AVX-512_BF16)

第三代至强® 可扩展处理器 (Ice Lake)
英特尔® DL Boost (AVX-512_VNNI) 和英特尔® Software Guard Extensions (英特尔® SGX)，支持领先 AI 应用，如联邦学习

第四代至强® 可扩展处理器 (Sapphire Rapids)
英特尔® Advanced Matrix Extensions (AMX) 进一步扩展了至强® 可扩展处理器上的内置 AI 加速功能

英特尔® AVX-512

VPMADDUBSW
VPMADDWD
VPADDD

第一代至强® 可扩展处理器

英特尔® DL Boost (VNNI)

VPDPBUSD
(8-bit new instruction)

更高效的推理加速

第二代和第三代至强® 可扩展处理器

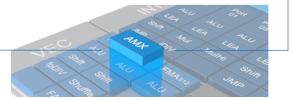
"Tiles"
2D Register Files

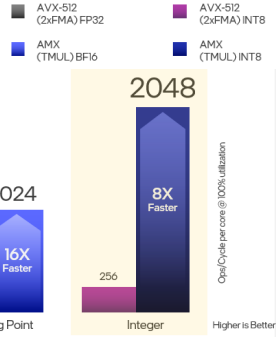
+ Intel® AMX

"TMUL"
Tile Matrix Multiply

INSTRUCTIONS that compute larger matrices in a single operation

Store bigger chunks of DATA in each core





将三条指令合而为一，可最大限度地利用计算资源，提高缓存利用率

1.74x
推理表现速度提升[1]
(BERT, 第三代vs 第二代)

相比英特尔® AVX-512，英特尔® AMX 可提供超过
8x operations/clock/core

领先性能

1 如欲了解更多详情，请访问：https://www.intel.cn/content/www/cn/zh/now/data-centric/3rd-gen-intel-xeon-scalable-processors.html

英特尔® 高级矩阵扩展（英特尔® AMX）

功能

- 提供广泛的软硬件优化，提升 AI 加速能力
- 同时支持 INT8 和 BF16 数据类型

软件支持

- 市场上的主流框架、工具套件和库（PyTorch、TensorFlow），英特尔® oneAPI 深度神经网络库（英特尔® oneDNN）

用例

- 图像识别、推荐系统、机器/语言翻译、自然语言处理（NLP）、媒体处理和分发

商业价值

- 为 AI/深度学习推理和训练工作负载带来显著性能提升
- 通过硬件加速使常见应用更快交付

高达 **8.6 倍**

语音识别推理性能提升[1]

启用英特尔® AMX（BF16）时
与上一代产品（FP32）的比较结果

高达 **3.5-10 倍**

PyTorch 训练性能提升[2]

高达 **5.7-10 倍**

PyTorch 实时推理性能提升[3]

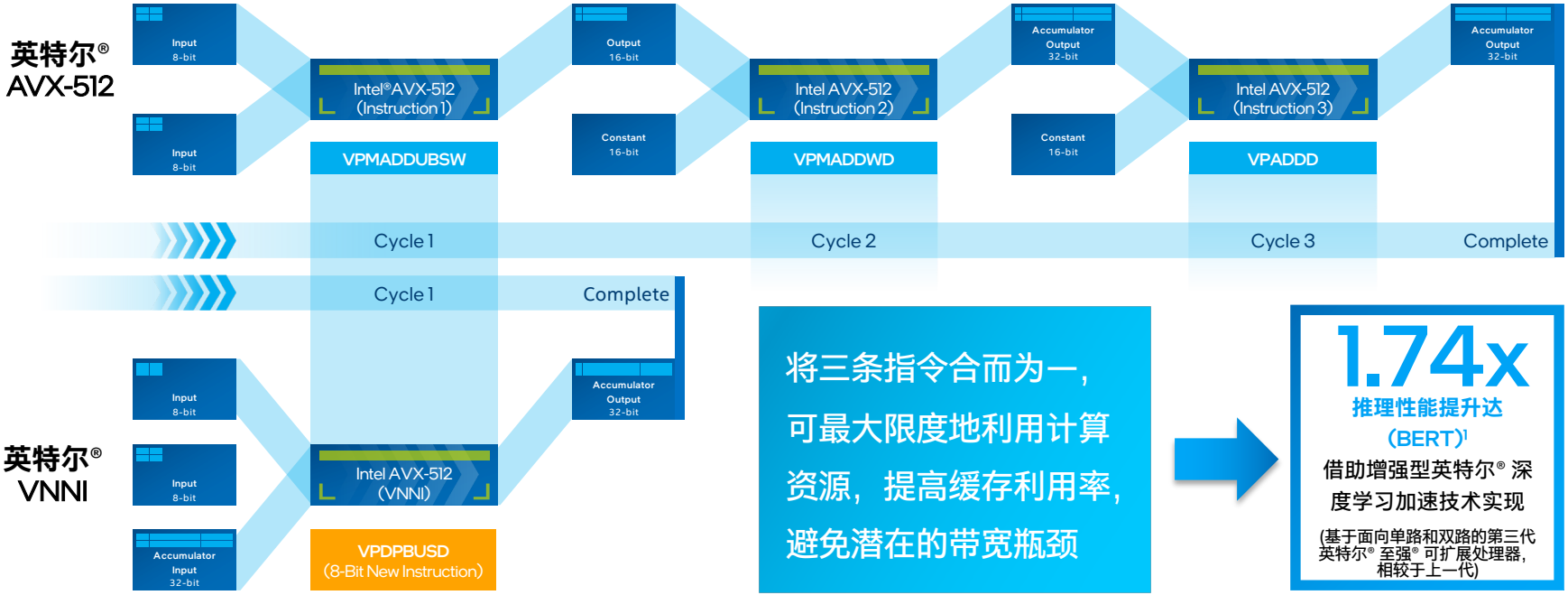
启用英特尔® AMX（BF16）时与
上一代产品（FP32）的比较结果

1.2.3 如欲了解更多详情，请访问：https://edc.intel.com/content/www/cn/zh/products/performance/benchmarks/processors/（第四代英特尔® 至强® 可扩展处理器）
https://edc.intel.com/content/www/us/en/products/performance/benchmarks/4th-generation-intel-xeon-scalable-processors/

结果可能不同。

英特尔® 深度学习加速

矢量神经网络指令（VNNI）
扩展英特尔® AVX-512 以加速 CPU 平台上的 AI / 深度学习推理



将三条指令合而为一，可最大限度地利用计算资源，提高缓存利用率，避免潜在的带宽瓶颈

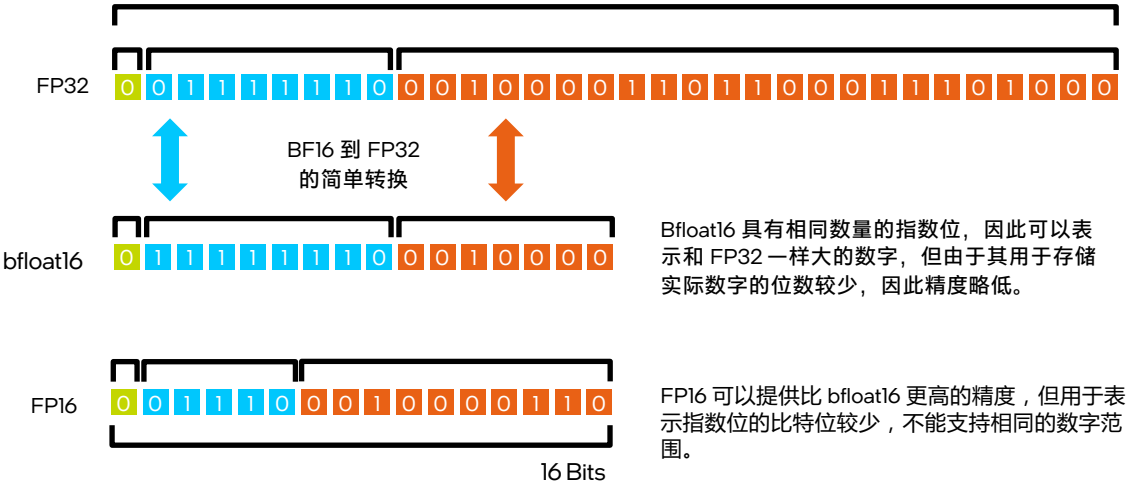
1.74x
推理性能提升达（BERT）[1]
借助增强型英特尔® 深度学习加速技术实现
(基于面向单路和双路的第三代英特尔® 至强® 可扩展处理器，相较于上一代)

1 如欲了解更多详情，请访问： https://www.intel.cn/content/www/cn/zh/now/data-centric/3rd-gen-intel-xeon-scalable-processors.html

脑浮点数 (bfloat16)

- 依据表示数字的比特位数，FP32 可提供更高的精度
- 许多 AI 功能并不需要 FP32 提供的精度水平
- bfloat16 支持基于相同指数域的相同范围的数字，但精度略低
- 从 FP32 转换到 bfloat16 比转换到 FP16 更简单
- 与 FP32 相比，使用 bfloat16 可实现每周期两倍的吞吐量

示例:
Number: 0.5658097267150878906 2596
As FP32: 0.565809726715087890625



Bfloat16 具有相同数量的指数位，因此可以表示和 FP32 一样大的数字，但由于其用于存储实际数字的位数较少，因此精度略低。

FP16 可以提供比 bfloat16 更高的精度，但用于表示指数位的比特位较少，不能支持相同的数字范围。

符号 – 表示正数或负数

指数 – 表示小数点在分数/尾数中的位置

分数/尾数 – 用于存储“数字”的比特位数

英特尔® SGX

采用英特尔® SGX 的机密人工智能应用，保护使用中的数据和代码

基于硬件的
可信执行环境 (Trusted Execution Environment)

加密敏感数据

隔离

保护敏感数据和代码，不受所有其他软件、云租户或管理员的影响，即使是恶意的访问

加密或匿名的结果

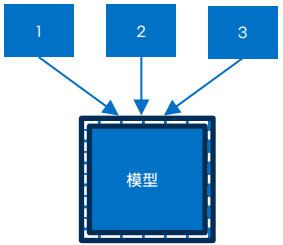
鉴证

加密验证 TEE 是否真实、配置正确且仅运行预期的软件负载

人工智能应用场景

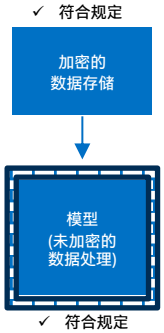
集中式多方

示例：多家医院汇集受监管的患者数据，以进行诊断模型训练



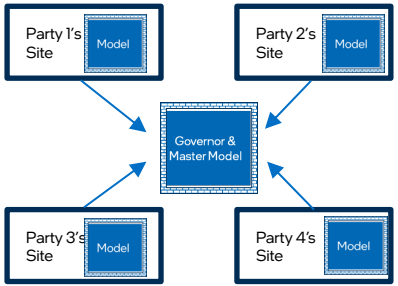
受监管的数据

示例：智慧城市摄像头捕获的受严格数据处理法规约束的个人身份信息 (PII)



联邦学习

示例：银行合作进行反洗钱，但数据太大且敏感，无法移动



英特尔® 数据中心 GPU Flex 系列

AV1
编解码

ECC
内存

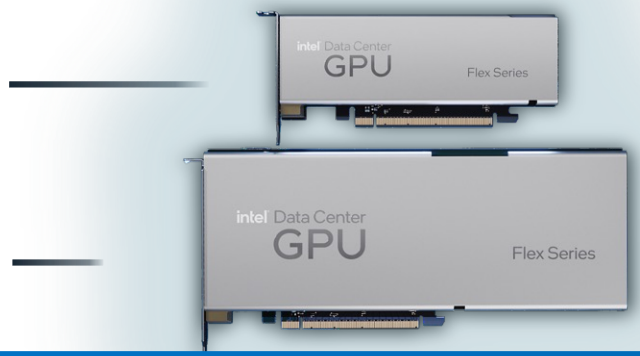
高达
256TOPS
(INT8)

Xe HPG
架构

基于硬件的
SR-IOV

英特尔
Xe 矩阵扩展

Flex 140	16 Xe 内核 16 光线追踪单元
Flex 170	32 Xe 内核 32 光线追踪单元



75W
半高 PCIe

150W
全高 PCIe

多达
4 个 Xe
媒体引擎

面向智能视觉云的 GPU 解决方案，支持基于标准的开放式软件堆栈，针对密度和质量进行了优化，具有关键的服务器功能，可实现高可靠性、可用性和可扩展性，有助于减少数据中心使用不同解决方案并管理异构或专有环境的需求，支持的工作负载包括：

AI 视觉推理 **媒体处理和交付** **云游戏** **虚拟桌面基础设施**

如欲了解更多详情，请访问： https://www.intel.cn/content/www/cn/zh/products/docs/discrete-gpus/data-center-gpu/flex-series/overview.html

基于至强® 可扩展平台的英特尔全栈 AI 生态支持

对数据进行工程处理 | 创建机器学习和深度学习模型 | 部署

容器存储库 oneContainer | MLOps Cnvrg.io | 开发人员沙盒 DevCloud | 标注/训练/优化平台 Sonoma Creek

打通 AI 与大数据的连接 — Spark — BigDL（以前的“Analytics Zoo”）

加速端到端数据科学和 AI | AI 分析工具套件

数据分析扩展

MODIN | SciPy | pandas | NumPy

经优化的框架和中间件

TensorFlow | PyTorch | mxnet | PaddlePaddle | scikit-learn | ONNX | LightGBM | XGBoost | CatBoost

优化模型

自动模型调优 AutoML | 自动低精度优化 | SigOpt | 英特尔® Neural Compressor

OpenVINO™ 工具套件

只需编写一次

自动部署

随处可用

经过英特尔调优

oneAPI | oneDAL | oneDNN | oneCCL | oneMKL

通用计算 #内核，#频率

矢量加速 AVX2、AVX-512、VNNI

矩阵加速 AMX

内存 高速缓存、DDR5、HBM、傲腾、频率

英特尔® 至强® 可扩展处理器全栈 AI 软件支持情况

类别	软件产品	是否开源	优化已提交给上游	英特尔® 扩展技术/工具	英特尔® 分发版	英特尔® 工具/套件
编排	Cnvrg.io	否				■
封装的软件工具套件	AI 套件	是				■
	BigDL	是				■
	OpenVINO™ 工具套件	是			■	■
优化	Neural Compressor	是				■
	SigOpt	否				■
深度学习框架	TensorFlow	是	■		■	
	PyTorch	是	■	■		
	ONNX	是	■			
	PDPD	是	■			
	MxNet	是	■			
机器学习框架	XGBoost	是	■			
	scikit-learn	是		■		
	CatBoost	是	■			
	LightGBM	是	■			
数据准备	Modin	是	■		■	
	Spark	是	■	■		

英特尔® oneAPI AI Analytics 工具套件

利用面向英特尔® 架构优化的库加速端到端人工智能和数据分析管道

显著优势

- 利用面向英特尔® 架构优化的深度学习框架和工具提升训练和推理性能
- 使用计算密集型 Python 包为数据分析和机器学习工作流提供落地加速

性能加速

简化端到端工作流程

提高生产力

加快开发

深度学习

- 面向英特尔® 架构优化的 TensorFlow
- 面向英特尔® 架构优化的 PyTorch
- 英特尔® 低精度优化工具（英特尔® LPOT）
- 面向英特尔® 架构优化的 Model Zoo

数据分析 & 机器学习

加速数据库

- 面向英特尔® 架构优化的 Modin
- HEAVY.AI Backend (formerly OmniSci)

面向英特尔® 架构优化的 Python

- XGBoost
- Scikit-learn
- Daal-4Py
- NumPy
- SciPy
- Pandas

示例及端到端工作负载

CPU　GPU　**支持的硬件架构**

硬件支持因个别工具而异。架构支持将随着时间的推移而扩大。

点击或通过如下链接获取工具包

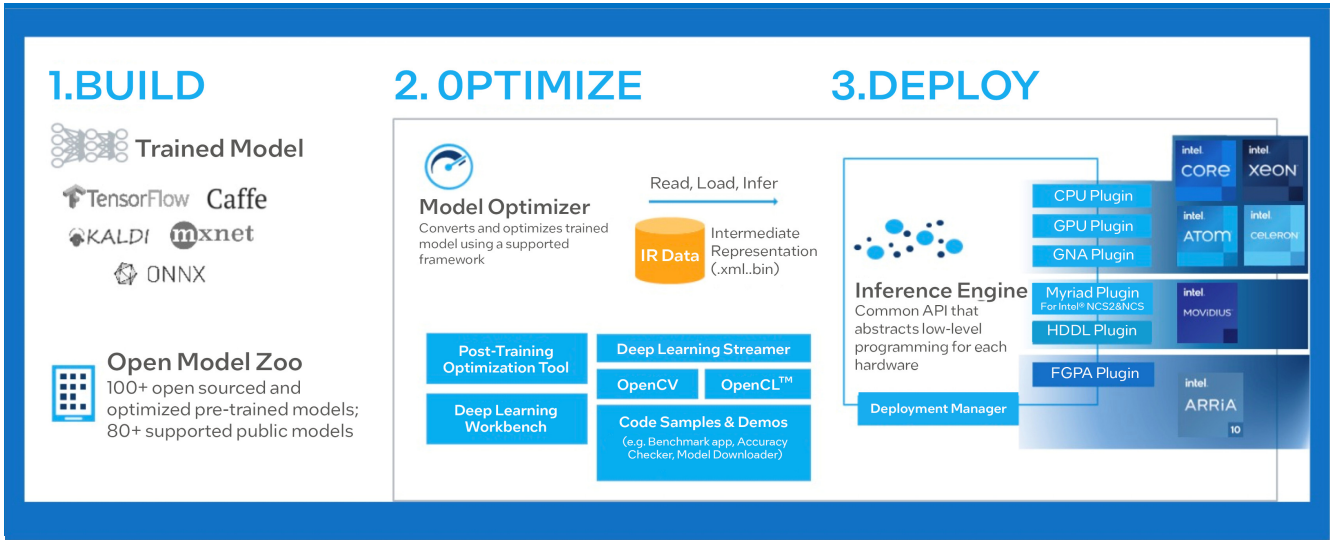
Intel Installer | Docker | Apt, Yum | Conda | Intel® DevCloud

了解更多：software.intel.com/oneapi/ai-kit

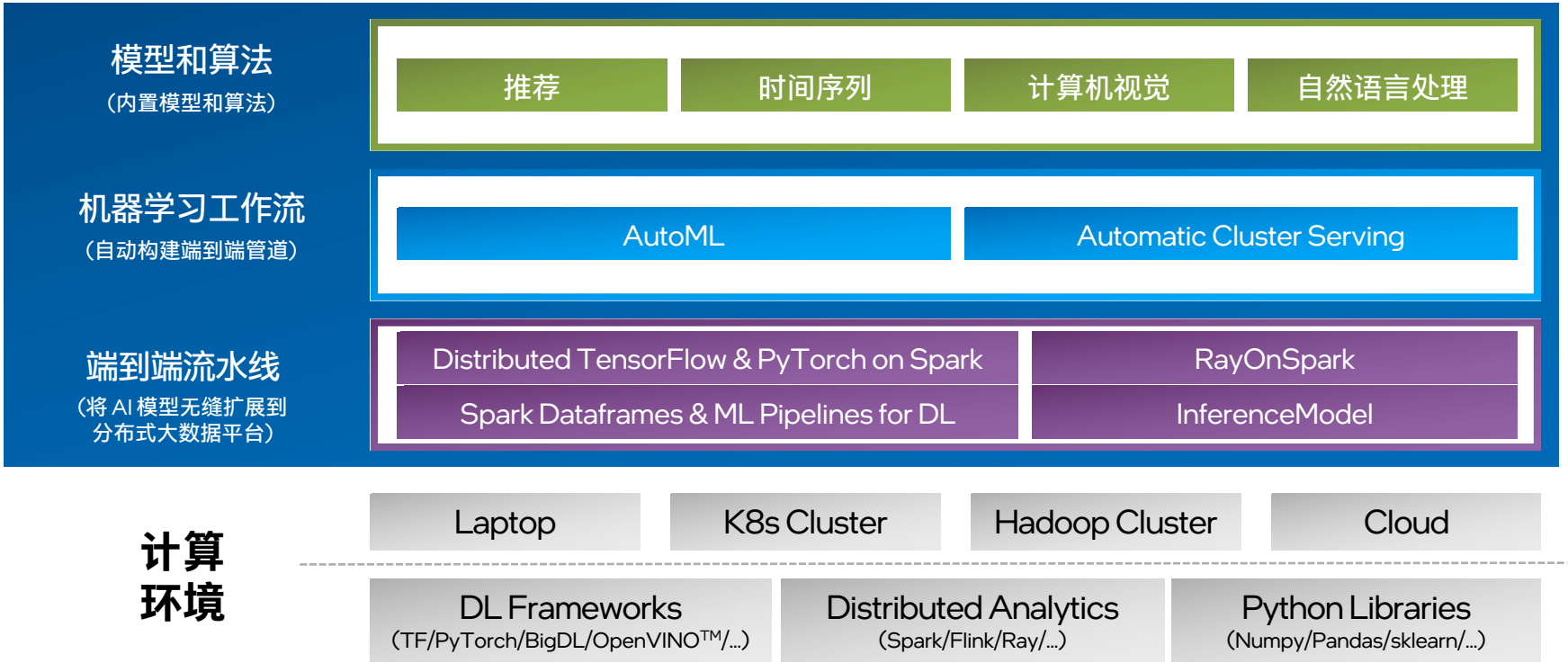
OpenVINO™ 工具套件 - 由 oneAPI 提供支持

旨在使用高性能人工智能和计算机视觉推理实现更加快速和准确的实际结果，部署在从边缘到云的、基于英特尔® XPU 架构（CPU、GPU、FPGA、VPU）的生产环境中

- **高性能、深度学习推理部署**
- **简化开发、易于使用**
- **一次编写、随处部署**



BigDL*：统一的大数据分析和 AI 平台



由英特尔® oneAPI 工具套件提供支持

* 指 BigDL 2.0 已包含 BigDL 和 Analytics Zoo。

了解更多: https://github.com/intel-analytics/analytics-zoo

大数据分析+人工智能端到端流水线

从笔记本电脑无缝扩展到分布式大数据平台

使用样本数据在
笔记本电脑上制作原型

在承载历史数据的
集群上进行试验

使用分布式数据流水线
进行**生产**部署

- 轻松构建将 AI 模型与大数据融合对接的端到端流水线原型
- 从笔记本电脑到分布式集群的**“零”**代码更改
- 可在**生产环境中的** Hadoop/K8s 集群上无缝部署
- 实现从机器学习到大数据应用的**流程自动化**

intel

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。

在特定系统的特殊测试中测试组件性能。硬件、软件或配置的差异将影响实际性能。当您考虑采购时，请查阅其他信息来源评估性能。关于性能和基准测试程序结果的更多信息，请访问www.intel.com/benchmarks。

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。

英特尔编译器针对英特尔微处理器的优化程度可能与针对非英特尔微处理器的优化程度不同。这些优化包括 SSE2、SSE3 和 SSSE3 指令集和其他优化。对于非英特尔微处理器上的任何优化是否存在、其功能或效力，英特尔不做任何保证。本产品中取决于微处理器的优化是针对英特尔微处理器。不具体针对英特尔微架构的特定优化为英特尔微处理器保留。请参考适用的产品用户与参考指南，获取有关本声明中具体指令集的更多信息。

声明版本：#20110804

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。

此处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。

intel®

加速 AI 实践，请访问：

官网
Intel.cn/ai

微博
@ Intel Business

微信
英特尔数据中心