解决方案简介

英特尔® 精选解决方案 | 版本 2
人工智能
第二代智能英特尔® 至强® 可扩展处理器
2019 年 11 月

基于人工智能推理的英特尔® 精选解决方案

利用基于行业标准英特尔® 技术且经过优化和验证的基础架构，加速人工智能（AI）推理和部署。

企业日益希望借助人工智能（AI）增加收入，提高效率和推动产品创新。尤其需要指出的是，基于深度学习（DL）的 AI 用例带来了最实用、最深刻的洞察，其中一些用例可推动多个行业的进步，如：

- **图像分类**，可用于图像所属类别分类（如面部表情分类）
- **对象检测**，可被自动驾驶汽车用于对象定位
- **图像分割**，可以在患者的磁共振成像（MRI）中勾勒出器官轮廓
- **自然语言处理**，可进行文本分析或翻译
- **推荐系统**，可被在线商店用于预测客户偏好或推荐向上销售选项

这些用例仅仅是个开始。通过将 AI 融入公司运营中，企业将发现应用 AI 的新方法。然而，所有 AI 用例的商业价值在很大程度上取决于通过深度神经网络训练的模型推理出答案的速度。在 DL 模型上实施推理需要大量资源，通常需要企业更新硬件以获得所需的性能和速度。然而，许多客户希望扩展其现有的基础设施，而不是购买新的单一用途硬件。您的 IT 部门已经非常熟悉英特尔® 硬件架构，其灵活性优势可帮助您保护 IT 投资。基于人工智能推理的英特尔® 精选解决方案是“交钥匙平台”，经过了预配置、验证和优化，可在 CPU 而非单独的加速卡上执行低延迟、高吞吐量推理。

基于人工智能推理的英特尔® 精选解决方案

基于人工智能推理的英特尔® 精选解决方案可帮助您在基于已验证过的英特尔架构解决方案上，快速部署高效的 AI 推理算法，从而加快创新和进入市场的步伐。为加快 AI 应用的推理和进入市场的速度，基于人工智能推理的英特尔® 精选解决方案将多种英特尔及第三方软件和硬件技术相结合。

软件选择

基于人工智能推理的英特尔® 精选解决方案使用的软件包括开发者工具和管理工具，可帮助您在生产环境中进行 AI 推理。

英特尔® OpenVINO™ 工具套件

英特尔® 开放视觉推理和神经网络优化工具套件（英特尔® OpenVINO™ 工具套件）是开发人员套件，可加速高性能 AI 和 DL 推理部署。该工具套件对在不同框架中训练的模型进行优化，以支持多种英特尔硬件选项，从而实现最佳性能的部署。使用工具套件的深度学习工作台可将模型量化到较低的精度，在此过程中，工具套件可以将模型从使用较大的高精度 32 位浮点数（通常用于训练且占用较多内存），转变为使用 8 位整数，以优化内存占用和性能。将浮点数转换为整数可以显著提高 AI 推理速度，同时实现几乎相同的精度。[1] 该工具套件可以转换和执行在多种框架中构建的模型，包括 TensorFlow*、MXNet*、PyTorch*、Kaldi* 以及开放式神经网络交换（ONNX）生态系统支持的任何框架。此外，还提供已预训练过的公共模型，使用户不用自己去搜索和训练模型，从而加快基于英特尔处理器的开发和图像处理相关工作。

深度学习参考栈

基于人工智能推理的英特尔® 精选解决方案配有深度学习参考堆栈（DLRS），这是一种集成的高性能开源软件堆栈，针对英特尔® 至强® 可扩展处理器进行了优化，封装在一个便捷的 Docker 容器中。DLRS 是一个预先验证和配置完备的堆栈，包括所需的库和软件组件，有助于降低与在 AI 生产环境中集成多个软件组件相关的复杂性。该堆栈还包括针对流行 DL 框架 TensorFlow 和 PyTorch 高度调优的容器，及英特尔 OpenVINO 工具套件。此开源社区版本有助于确保 AI 开发人员轻松访问英特尔平台所有特性和功能。

Kubeflow 和 Seldon Core

随着企业积累了在生产环境中部署推理模型的经验，业界在一组统称为 “MLOps” 的最佳实践方面达成了共识，这些实践类似于 “DevOps” 软件开发实践。为帮助团队应用 MLOps，基于人工智能推理的英特尔® 精选解决方案使用了 Kubeflow*。借助 Kubeflow，团队可以在零停机的情况下顺利推出其模型的新版本。Kubeflow 使用受支持的模型服务后端（例如 TensorFlow Serving），将经过训练的模型导出至 Kubernetes。模型部署可使用金丝雀测试或影子部署来实现新版与旧版的并行验证。如果检测到问题，除跟踪外，团队还可使用模型和数据版本控制来简化原因分析。

为确保服务可满足人工智能推理不断增长的需求，基于人工智能推理的英特尔® 精选解决方案还提供了负载均衡功能，可自动将推理分片到节点中可服务对象的可用实例中。多租户支持提供不同的模型，从而提高硬件利用率。

最后，为加快运行 AI 推理的服务器与需要 AI 洞察的端点之间的推理请求处理速度，基于人工智能推理的英特尔® 精选解决方案可使用 Seldon Core 帮助管理推理工作流。Kubeflow 还与 Seldon Core 集成，以便在 Kubernetes 上部署 DL 模型，并使用 Kubernetes API 管理在推理工作流中部署的容器。

硬件选择

基于人工智能推理的英特尔® 精选解决方案组合了第二代智能英特尔® 至强® 可扩展处理器、英特尔® 傲腾™ 数据中心级固态盘、英特尔® 3D NAND 固态盘和英特尔® 以太网 700 系列，可帮助您的企业快速部署构建于性能优化平台上的生产级 AI 基础设施，使用大容量内存处理要求最苛刻的应用和工作负载。

第二代智能英特尔® 至强® 可扩展处理器

基于人工智能推理的英特尔® 精选解决方案具有第二代智能英特尔® 至强® 可扩展处理器的性能和功能。对于 “Base” 配置，英特尔® 至强® 金牌 6248 处理器在价格、性能和内置技术之间实现了最佳平衡，可增强在 AI 模型上进行推理的性能和效率。“Plus” 配置推荐搭载英特尔® 至强® 铂金 8268 处理器，以实现更快的 AI 推理。这两种配置也可以使用更多的处理器数量。第二代智能英特尔® 至强® 可扩展处理器包括英特尔® 深度学习加速，该系列加速特性可通过使用专门的矢量神经网络指令（VNNI）集来提高 AI 推理性能，该指令集使用单指令完成以往需要三个单独指令的 DL 计算。

英特尔® 傲腾™ 数据中心级技术

英特尔® 傲腾™ 数据中心级技术填补了存储和内存层级中的关键空白，可帮助数据中心更快速访问数据。这项技术还颠覆了内存和存储层，能够为各种产品和解决方案提供持久内存、大内存池、高速缓存和存储。

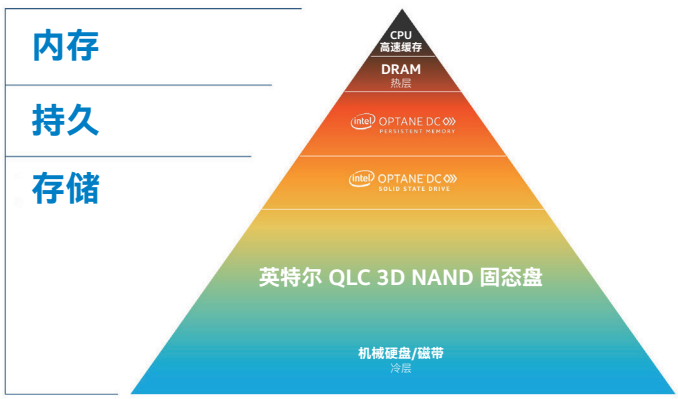

图 1. 英特尔傲腾技术填补了数据中心领域的内存和存储性能空白

英特尔® 傲腾™ 数据中心级固态盘和英特尔® 3D NAND 固态盘

当高速缓存层运行在低延迟和高耐用性的快速固态盘上时，AI 推理性能最佳。如果在高速缓存层采用最高性能的固态盘代替主流串行 ATA（SATA）固态盘，需要高性能的工作负载将受益匪浅。在这些英特尔® 精选解决方案中，英特尔® 傲腾™ 数据中心级固态盘用于驱动高速缓存层。英特尔® 傲腾™ 数据中心级固态盘可提供较高的每秒数据输入输出操作（IOPS）、性价比、低延迟和每天 30 次全盘写入的耐用性，因此它们非常适合承载写入负载较大的高速缓存功能。[2] 解决方案的容量层由英特尔® 3D NAND 固态盘支持，同时具备出色的数据完整性、性能一致性和磁盘可靠性，可提供优化的读取性能。

25Gb 以太网

25Gb 英特尔® 以太网 700 系列网络适配器可提高基于人工智能推理的英特尔® 精选解决方案的性能。搭配第二代英特尔® 至强® 铂金处理器和英特尔® 固态盘 DC P4600，它们相比 1Gb 以太网（GbE）适配器和英特尔® 固态盘 DC S4500 可提供高达 2.5 倍的性能。[3,4] 英特尔® 以太网 700 系列提供了经过验证的性能，可通过广泛的互操作性达到数据弹性和服务可靠性的高质量阈值。[5] 所有英特尔® 以太网产品均享受全球售前和售后支持，并提供有限的终身保修。

通过基准测试验证性能

所有英特尔® 精选解决方案均经过基准测试的验证，达到工作负载优化性能的预先指定最低性能水平。在该解决方案中，英特尔选择使用标准的 DL 基准测试方法并模拟真实场景进行测量和基准测试。

对于标准基准测试，每秒可处理的图像数量（吞吐量）是在预先训练的深度残差神经网络（ResNet 50 v1）上进行测量的，该网络与使用合成数据的 TensorFlow、PyTorch 和 OpenVINO 工具套件上广泛使用的 DL 用例（图像分类、定位和检测）紧密关联。

为模拟实际场景，该测试启动了代表多个请求流的多个客户端。这些客户端将图像从外部客户端系统发送到服务器以进行推理。在服务器端，入站请求由 Istio 进行负载平衡。然后，请求被发送到一个可服务对象的多个实例，该对象包含通过 Seldon Core 运行的预处理、预测和后处理步骤工作流。预测是使用 OpenVINO 工具套件模型服务器的优化 DLRS 容器映像完成的。一旦请求通过工作流，推理将被发回请求客户端。测量吞吐量和延迟可以帮助确保测试配置支持生产环境中的推理规模。

Base 和 Plus 配置

如表 1 所示，基于人工智能推理的英特尔® 精选解决方案提供两种配置。Base 配置指定了解决方案的最低要求性能，Plus 配置旨在展示系统构建商、系统集成商以及解决方案和服务提供商可如何进一步优化基于人工智能推理的英特尔® 精选解决方案，进而实现更高的性能。

客户可以升级或扩展其中一项配置，以提高容量或性能。Plus 配置利用性能更高的第二代智能英特尔® 至强® 可扩展处理器和更多内存，与 Base 配置相比可将 AI 推理速度加快高达 39%。[6]

如需满足基于人工智能推理的英特尔® 精选解决方案的要求，解决方案提供商必须达到或超过定义的最低配置要素，并达到以下所列的最低基准性能阈值。

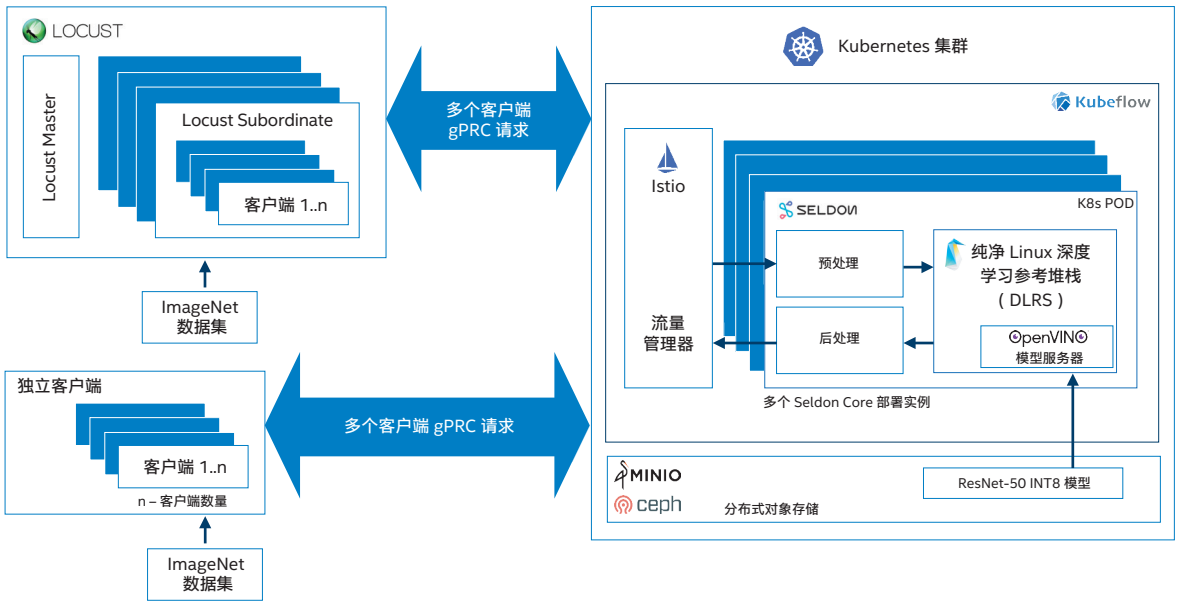

图 2. 在基于人工智能推理的英特尔® 精选解决方案上进行的真实场景基准测试的架构图

表 1. 基于人工智能推理的英特尔® 精选解决方案版本 2 的 Base 和 Plus 配置

要素	基于人工智能推理的英特尔® 精选解决方案的 **Base 配置**	基于人工智能推理的英特尔® 精选解决方案的 **PLUS 配置**
节点数	单节点配置	单节点配置
处理器	2 x 英特尔® 至强® 金牌 6248 处理器（2.50 GHz，20 个内核，40 个线程），或更多数量的英特尔® 至强® 可扩展处理器	2 x 英特尔® 至强® 金牌 8268 处理器（2.90 GHz，24 个内核，48 个线程），或更多数量的英特尔® 至强® 可扩展处理器
内存	192 GB 或更高（12 个 16 GB 2,666 MHz DDR4 ECC RDIMM）	384 GB（12 个 32 GB 2,934 MHz DDR4 ECC RDIMM）
启动盘	1 x 256 GB 英特尔固态盘 DC P4101（M.2 80 mm PCIe 3.0 x4，3D2，TLC）或更多	1 x 256 GB 英特尔固态盘 DC P4101（M.2 80 mm PCIe 3.0 x4，3D2，TLC）或更多
存储	**数据盘：** 1.6 TB NVM Express（NVMe）英特尔固态盘 DC P4510 **高速缓存盘：** 375 GB 英特尔傲腾固态盘 DC P4800X U.2 NVMe 固态盘	**数据盘：** 1.6 TB NVMe 英特尔固态盘 DC P4510 **高速缓存盘：** 375 GB 英特尔傲腾固态盘 DC P4800X U.2 NVMe 固态盘
数据网络	1 x 英特尔® 以太网融合网络适配器 XXV710-DA2 SFP28 DA Copper PCIe x 8 双端口 25/10/1 GbE	1 x 英特尔® 以太网融合网络适配器 XXV710-DA2 SFP28 DA Copper PCIe x 8 双端口 25/10/1 GbE
管理网络	集成的 1 GbE 端口 0/RMM 端口	集成的 1 GbE 端口 0/RMM 端口
软件		
Linux 操作系统	CentOS Linux 版本 7.6.1810/Red Hat Enterprise Linux（RHEL）7	CentOS Linux 版本 7.6.1810/Red Hat Enterprise Linux（RHEL）7
英特尔数学核心函数库（英特尔 MLK）	英特尔 MKL 版本 2019 Update 4	英特尔 MKL 版本 2019 Update 4
英特尔 OpenVINO 工具套件分发版	2019 R1.0.1	2019 R1.0.1
OPENVINO 模型服务器	0.4	0.4
TENSORFLOW	1.14	1.14
PYTORCH	1.2.0	1.2.0
MXNET	1.3.1	1.3.1
英特尔® Python 分发版	2019 Update 1	2019 Update 1
基于深度神经网络的英特尔® 数学核心函数库（英特尔® MKL-DNN）	0.19（表示 OpenVINO 工具套件）	0.19（表示 OpenVINO 工具套件）
深层学习参考堆栈（DLRS）	v4.0	v4.0
源到图像	1.1.14	1.1.14
DOCKER	18.09	18.09
KUBERNETES	v1.15.3	v1.15.3
KUBEFLOW	v0.6.1	v0.6.1
HELM	2.14.3	2.14.3
SELDON CORE	0.3.2	0.3.2
CEPH	v14.2.1	v14.2.1
MIN.IO（ROOK V1.0）	RELEASE.2019-04-23T23-50-36Z	RELEASE.2019-04-23T23-50-36Z
ROOK	1.0.5	1.0.5
其它		
可信平台模块（TPM）	TPM 2.0	TPM 2.0

最低性能标准		
经过验证，可实现或超过以下最低性能和功能：		
在 OPENVINO 工具套件上使用 RESNET-50 进行分类	每秒 1,900 张图像（精度为 91%，位列前五）[6]	每秒 2,650 张图像（精度为 91%，位列前五）[6]
在 1 个节点到 2 个节点的模拟真实场景中进行扩展	高达 1.91x[7]	高达 1.91x[8]
Plus 配置的业务价值（相比 Base 配置）	Plus 配置可将推理性能提高多达 39%。[6]	

**推荐但不要求

基于人工智能推理的英特尔® 精选解决方案的技术选择

除了基于人工智能推理的英特尔® 精选解决方案使用的英特尔硬件基础外，英特尔技术进一步提升了性能与可靠性：

- **英特尔® 高级矢量扩展指令集 512（英特尔® AVX-512）：** 一个 512 位指令集，可提高要求苛刻的工作负载和用例（如 AI 推理）的性能。
- **英特尔® 深度学习加速：** 第二代智能英特尔® 至强® 可扩展处理器中引入了一组加速功能，可大幅提高使用领先 DL 框架（例如 PyTorch、TensorFlow、MXNet、PaddlePaddle 和 Caffe）构建的推理应用的性能。英特尔深度学习加速技术的基础是 VNNI，这是一种专用指令集，使用单一指令进行 DL 计算，该任务以前需要三条单独指令。
- **英特尔® OpenVINO™ 工具套件分发版：** 一款免费的软件套件，可帮助开发人员和数据科学家加速 AI 工作负载，并简化从网络边缘到云端的 DL 推理和部署。
- **英特尔® 数学核心函数库（英特尔® MLK）：** 该函数库实施了流行的数学运算，这些运算已针对英特尔硬件进行了优化，以帮助应用充分利用英特尔 AVX-512 指令集。它兼容广泛的编译器、语言、操作系统以及链接和线程模型。
- **基于深度神经网络的英特尔® 数学核心函数库（英特尔® MKL-DNN）：** 一个开源性能增强库，用于在英特尔硬件上加速 DL 框架。

什么是英特尔® 精选解决方案？

英特尔® 精选解决方案是针对工作负载优化的预定义解决方案，旨在最大限度地克服基础设施评估和部署方面的挑战。这些解决方案经过 OEM/ODM 的验证，以及 ISV 和英特尔认证。英特尔与硬件、软件、操作系统厂商以及全球领先的数据中心和服务提供商广泛合作，共同开发这些解决方案。每款英特尔® 精选解决方案均是定制的英特尔数据中心计算、内存、存储和网络技术组合，提供了可预测、可靠的超凡性能。

如欲满足英特尔® 精选解决方案的条件，厂商必须：

1. 满足解决方案参考设计规范规定的软件和硬件堆栈要求
2. 达到或超过既定的参考性能指标评测测试结果
3. 发布解决方案简介和详细实施指南，帮助客户顺利部署

解决方案提供商还可以开发自己的优化解决方案，为最终客户提供更简单、更一致的部署体验。

英特尔® 至强® 可扩展处理器

第二代智能英特尔® 至强® 可扩展处理器：

- 以经济高效、灵活的方式提供较高的可扩展性，涵盖多云到智能边缘
- 建立无缝的性能基础，进一步加快数据带来的深远影响
  - 支持突破性的英特尔傲腾数据中心级持久内存技术
  - 加快人工智能（AI）性能并帮助整个数据中心为人工智能做好准备
  - 提供硬件增强平台保护和威胁监控

解决方案基于：

- **英特尔® Python 分发版：**借助集成的英特尔® 性能库（例如英特尔® MKL），加速与 AI 有关的 Python 库（例如 NumPy*、SciPy* 和 scikit-learn*），以实现更快的 AI 推理。
- **框架优化：**英特尔与 Google* 开展 TensorFlow 合作，与 Apache* 开展 MXNet 合作，与百度* 开展 PaddlePaddle* 、Caffe* 和 PyTorch 合作，在数据中心内使用基于英特尔® 至强® 可扩展处理器的软件优化来增强 DL 性能，并且继续增加其他行业领导者的框架。

在行业标准硬件上部署优化的快速 AI 推理

借助为英特尔® 至强® 可扩展处理器验证的工作负载优化配置，英特尔® 精选解决方案可帮助企业快速推进数据中心转型。选择基于人工智能推理的英特尔® 精选解决方案，意味着企业可获得经过优化、测试、可扩展性验证及预先调优的配置，从而帮助 IT 部门在生产环境中快速、高效地部署 AI 推理。此外，通过选择基于人工智能推理的英特尔® 精选解决方案，IT 部门可在惯常部署和管理的硬件上实现高速 AI 推理。

访问：https://www.intel.cn/content/www/cn/zh/architecture-and-technology/intel-select-solutions-overview.html，了解更多信息，并向您的基础设施厂商咨询英特尔® 精选解决方案的信息。

了解详细信息

英特尔® 精选解决方案网页：**intel.com/content/www/us/en/architecture-and-technology/intel-select-solutions-overview.html**

英特尔® 至强® 可扩展处理器：**www.intel.cn/content/www/cn/zh/products/processors/xeon/scalable.html**

英特尔傲腾数据中心级技术：**intel.com/optane**

英特尔固态盘数据中心产品家族：**www.intel.cn/content/www/cn/zh/products/memory-storage/solid-state-drives/data-center-ssds.html**

英特尔以太网产品：**https://www.intel.cn/content/www/cn/zh/products/network-io/ethernet.html**

英特尔以太网 700 系列：**https://www.intel.cn/content/www/cn/zh/products/network-io/ethernet.html**

英特尔® OpenVINO 工具套件分发版：**https://software.intel.com/en-us/openvino-toolkit**

英特尔深度学习加速：**intel.ai/increasing-ai-performance-intel-dlboost**

英特尔框架优化：**intel.ai/framework-optimizations**

英特尔深度学习参考堆栈：**https://software.intel.com/en-us/blogs/2018/12/07/intel-introduces-the-deep-learning-reference-stack**

英特尔® 精选解决方案由英特尔® Builders 提供支持：**builders.intel.com**。在 Twitter 上关注我们：**#IntelBuilders**

Kubeflow：**kubeflow.org**

Seldon Core：**seldon.io/tech/products/core/**

Seldon 部署：**seldon.io/tech/products/deploy/**

[1] 英特尔。“低数值精度深度学习推理与训练。” 2018 年 10 月。https://software.intel.com/en-us/articles/lower-numerical-precision-deep-learning-inference-and-training

[2] 基于英特尔内部测试。资料来源：英特尔。“产品简介：英特尔傲腾固态盘 DC P4800X 系列。” intel.com/content/www/us/en/solid-state-drives/optane-ssd-dc-p4800x-brief.html

[3] 测试基于第二代智能英特尔® 至强® 铂金 8260 处理器，从 1Gb 升级到 25Gb 英特尔® 以太网网络适配器 XXV710，从串行 ATA（SATA）固态盘升级到基于 NVM Express（NVMe）的 PCIe 英特尔固态盘 DC P4600。

[4] HeadGear Strategic Communications 提供的性能结果基于截至 2019 年 2 月 12 日的测试。本文中的对比分析由英特尔委托 HeadGear Strategic Communications 完成。配置详情：**虚拟机（VM）主机服务器：** 英特尔® 至强® 铂金 8160 处理器、英特尔® 至强® 铂金 8160F 处理器（CPUID 50654，microcode revision 0x200004D）和英特尔® 至强® 铂金 8260 处理器（CPUID 50656，microcode revision 04000014）；英特尔® 服务器主板 S2600WFT（主板型号 H48104-850，BIOS ID SE5C620.86B.0D.01.0299.122420180146，底板管理控制器 [BMC] 版本 1.88.7a4eac9e；英特尔® 管理引擎 [英特尔® ME] 版本 04.01.03.239；SDR 软件包版本 1.88）；576 GB DDR4 2,133 MHz 寄存器内存，一个英特尔以太网物理适配 XXV710-DA2，一个英特尔® 以太网融合网络适配器 X710-DA2；操作系统盘配置：2 个英特尔固态盘 DC S3500，采用英特尔® 快速存储技术企业版 [英特尔® RSTe] RAID1 配置。Windows Server 2016 Datacenter 版本 10.0.14393 build 14393，Hyper-V 版本 10.0.14393.0，Hyper-V scheduler type 0x3，安装的更新 KB4457131、KB4091664、KB1322316、KB3211320 和 KB3192137。**电子邮件虚拟机配置：** Windows Server 2012 Datacenter 版本 6.2.9200 build 9200；4 个 vCPU；12 GB 系统内存，BIOS 版本/日期：Hyper-V 版本 v1.0，2012 年 11 月 26 日），SMBIOS 版本 2.4；Microsoft Exchange Server 2013，通过运行 Microsoft Exchange Load Generator 2013（应用版本 15.00.0805.000）的虚拟机客户端生成工作负载。**数据库虚拟机配置：** Windows Server 2016 Datacenter 版本 10.0.14393 build 14393，2 个 vCPU 7.5 GB 系统内存；BIOS 版本/日期：Hyper-V 版本 v1.0，2012 年 11 月 26 日），SMBIOS 版本 2.4，Microsoft SQL Server 2016 工作负载生成 DVD 存储应用（dell.com/downloads/global/power/ps3q05-20050217-Jaffe-OE.pdf）。**存储服务器：** 英特尔® 服务器系统 R2224WFTZS；英特尔服务器主板 S2600WFT（主板型号 H48104-850，BIOS ID SE5C620.86B.00.01.0014.070920180847，BMC 版本 1.60.56383bef；英特尔 ME 版本 04.00.04.340；SDR 软件包修订版 1.60）；96 GB DDR4 2,666 Mhz 寄存器内存，1 个英特尔以太网网络适配器 XXV710-DA2，1 个英特尔以太网融合网络适配器 X710-DA2；操作系统盘配置：2 个英特尔固态盘 DC S3500 — 英特尔 RSTe RAID1 配置。**存储配置：** 8 个英特尔固态盘 DC P4600（2.0 Tb），使用英特尔® Virtual RAID on CPU（英特尔® VROC）配置为 RAID 5 卷，8 个英特尔固态盘 DC S4500（480 GB）— RAID5 配置，使用英特尔® RAID 模块 RMSP3AD160F，8 个英特尔固态盘 DC P4510 - RAID5 配置，使用基于虚拟机操作系统存储的英特尔® RAID 模块 VROC，Windows Server 2016 Datacenter 版本 10.0.14393 build 14393，Hyper-V 版本 10.0.14393.0，Hyper-V scheduler type 0x3，安装的更新 KB4457131、KB4091664、KB1322316、KB3211320 和 KB3192137。**使用英特尔® 至强® 铂金 8160 和英特尔® 至强® 铂金 8160F 处理器配置的 Windows Server 2016 Datacenter 和 Windows Server 2012 Datacenter：** CVE-2017-5715（分支目标注入）的推测控制设置 — 为缓解分支目标注入提供硬件支持：正确；提供 Windows 操作系统对缓解分支目标注入的支持：正确；启用 Windows 操作系统对缓解分支目标注入的支持：正确；系统策略禁用 Windows 操作系统对缓解分支目标注入的支持：错误；由于缺少硬件支持，Windows 操作系统对缓解分支目标注入的支持被禁用：错误。CVE-2017-5754（流氓数据高速缓存加载）的推测控制设置 — 硬件需要内核 VA 阴影：正确；提供 Windows 操作系统对内核 VA 阴影的支持：正确；启用 Windows 操作系统对内核 VA 阴影的支持：正确。CVE-2018-3639（推测性存储旁路）的推测控制设置 — 硬件容易受到推测性存储旁路的影响：正确；提供对推测性存储旁路禁用的硬件支持：正确；提供 Windows 操作系统对推测性存储旁路禁用的支持：正确；在整个系统中启用 Windows 操作系统对推测性存储旁路禁用的支持：正确。CVE-2018-3620（L1 终端故障）的推测控制设置 — 硬件容易受到 L1 终端故障的影响：正确；提供 Windows 操作系统对 L1 终端故障缓解的支持：正确；启用 Windows 操作系统对 L1 终端故障缓解的支持：正确。**使用英特尔® 至强® 铂金 8160 和英特尔® 至强® 8160F 处理器配置的 Windows Server 2016 Datacenter 和 Windows Server 2012 Datacenter：** CVE-2017-5715（分支目标注入）的推测控制设置 — 为缓解分支目标注入提供硬件支持：正确；提供 Windows 操作系统对缓解分支目标注入的支持：正确；启用 Windows 操作系统对缓解分支目标注入的支持：正确。CVE-2017-5754（流氓数据高速缓存加载）的推测控制设置 — 硬件需要内核 VA 阴影：错误。CVE-2018-3639（推测性存储旁路）的推测控制设置 — 硬件容易受到推测性存储旁路的影响：正确；提供对推测性存储旁路禁用的硬件支持：正确；提供 Windows 操作系统对推测性存储旁路禁用的支持：正确；在整个系统中启用 Windows 操作系统对推测性存储旁路禁用的支持：正确。CVE-2018-3620（L1 终端故障）的推测控制设置 — 硬件容易受到 L1 终端故障的影响：错误。**网络交换机：** 1/10GbE Supermicro SSE-X3348S，硬件版本 P4-01，固件版本 1.0.7.15；10/25GbE Arista DCS-7160-48YC6，EOS 4.18.2-REV2-FX。

[5] 除广泛的操作系统支持外，英特尔以太网 700 系列还包括经过全面测试的网络适配器、附件（光学和电缆）、硬件和软件。产品组合解决方案的完整列表发布在 https://www.intel.cn/content/www/cn/zh/products/network-io/ethernet.html 上。在英特尔® 至强® 可扩展处理器和网络生态系统中对软硬件进行了全面验证。这些产品针对英特尔架构和广泛的操作系统生态系统进行了优化：Windows、Linux 内核、FreeBSD、Red Hat Enterprise Linux（RHEL）、SUSE、Ubuntu、Oracle Solaris 和 VMware ESXi。英特尔以太网 700 系列支持的连接和介质类型为：直接连接铜缆和光纤 SR/LR（QSFP +、SFP+、SFP28、XLPPI/CR4、25G-CA/25G-SR/25G-LR），双绞铜线（1000BASE-T/10GBASE-T）、背板（XLAUI/XAUI/SFI/KR/KR4/KX/SGMII）。请注意，英特尔是唯一提供 QSFP+ 介质类型的厂商。英特尔以太网 700 系列支持的速度包括 10GbE、25GbE、40GbE。

[6] Plus 配置相对于 Base 配置实现了 39.47% 的性能提升，基于英特尔于 2019 年 5 月 23 日使用 ImageNet 数据集分类和 ResNet-50 在 OpenVINO 工具套件上实施的基准测试。**Base 配置：** 单节点，2 x 英特尔® 至强® 金牌 6248 处理器（2.50 Ghz，20 核，40 线程），12 x 16 GB 2,666 MHz DDR4 ECC RDIMM（总内存 192 GB），启动盘：1 x 256 GB 英特尔固态盘 DC P4101（M.2 80 mm PCIe 3.0 x4，3D2，TLC），数据盘：1.6 TB NVM Express（NVMe）英特尔固态盘 P4510，高速缓存盘：375 GB 英特尔傲腾固态盘 DC P4800X U.2 NVMe 固态盘，数据网络：1 x 10 Gb 英特尔以太网融合网络适配器 XXV710-DA2，管理网络：集成的 1 Gb 以太网（GbE）端口 0/RMM 端口。软件：CentOS Linux 版本 7.5.1804/Red Hat Enterprise Linux（RHEL）7，英特尔数学核心函数库（英特尔 MKL）版本 2018 update 3，英特尔 OpenVINO 工具套件分发版 2019 R1 运行时，OpenVINO Model Server 0.4，TensorFlow 1.14，PyTorch 1.01，MXNet 1.31，英特尔 Python 2019 update 1 分发版，基于深度神经网络的英特尔数学核心函数库（英特尔 MKL-DNN）0.18（由 OpenVINO 确定），深度学习参考堆栈（DLRS）v4.0，Kubernetes v1.15.1，Kubeflow v0.6.1，Seldon Core，Ceph v14.2.1，Min.io（Rook v1.0）RELEASE.2019-04-23T23-50-36Z。在 OpenVINO 工具套件上使用 ResNet-50 进行 ImageNet 数据集分类：每秒 1,880 张图像（精度为 91%，位列前五）**Plus 配置：** 单节点，2 x 英特尔® 至强® 铂金 8268 处理器（2.90 Ghz，24 核，48 线程），12 x 32 GB 2,934 MHz DDR4 ECC RDIMM（总内存 384 GB），启动盘：1 x 256 GB 英特尔固态盘 DC P4101（M.2 80 mm PCIe 3.0 x4，3D2，TLC），数据盘：1.6 TB NVM 英特尔固态盘 P4510，高速缓存盘：375 GB 英特尔傲腾固态盘 DC P4800X U.2 NVMe 固态盘，数据网络：1 x 10 Gb 英特尔以太网融合网络适配器 XXV710-DA2，管理网络：集成的 1 GbE 端口 0/RMM 端口。软件：CentOS Linux 版本 7.5.1804/RHEL 7，英特尔 MKL 版本 2018 update 3，英特尔 OpenVINO 工具套件分发版 2019 R1 运行时，OpenVINO Model Server 0.4，TensorFlow 1.14，PyTorch 1.01，MXNet 1.31，英特尔 Python 2019 update 1 分发版，英特尔 MKL-DNN 0.18（由 OpenVINO 确定），DLRS v4.0，Kubernetes v1.15.1，Kubeflow v0.6.1，Seldon Core，Ceph v14.2.1，Min.io（Rook v1.0）RELEASE.2019-04-23T23-50-36Z。在 OpenVINO 工具套件上使用 ResNet-50 进行 ImageNet 数据集分类：每秒 2,650 张图像（精度为 91%，位列前五）。

[7] 英特尔于 2019 年 10 月 9 日进行了测试。测试配置：两个节点，2 x 英特尔® 至强® 金牌 6248 处理器（2.50 Ghz，20 核，40 线程），12 x 16 GB 2,666 MHz DDR4 ECC RDIMM（总内存 192 GB），启动盘：1 x 256 GB 英特尔固态盘 DC P4101（M.2 80 mm PCIe 3.0 x4，3D2，TLC），数据盘：1.6 TB NVM Express（NVMe）英特尔固态盘 P4510，高速缓存盘：375 GB 英特尔傲腾固态盘 DC P4800X U.2 NVMe 固态盘，数据网络：1 x 10 Gb 英特尔以太网融合网络适配器 X722，管理网络：集成的 1 Gb 以太网（GbE）端口 0/RMM 端口。软件：CentOS Linux 版本 7.6.1810/Red Hat Enterprise Linux（RHEL）7，英特尔数学核心函数库（英特尔 MKL）版本 2019 update 4，英特尔 OpenVINO 工具套件分发版 2019 R1，OpenVINO Model Server 0.4，英特尔 Python 2019 update 1 分发版，基于深度神经网络的英特尔数学核心函数库（英特尔 MKL-DNN）0.19，深度学习参考堆栈（DLRS）v4.0，Docker v18.09，Helm v2.14.3，Kubernetes v1.15.3，Kubeflow v0.6.1，Seldon Core v0.3.2，Rook v1.0.5，Ceph v14.2.1，Min.io（Rook v1.0）RELEASE.2019-04-23T23-50-36Z。在模拟的真实场景中进行扩展 — 吞吐量测试：标准化性能 1（英特尔® 超线程技术：关闭）。

[8] 英特尔于 2019 年 10 月 9 日进行了测试。测试配置：两个节点，2 x 英特尔® 至强® 铂金 8268 处理器（2.90 GHz，24 核，48 线程），12 x 16 GB 2,666 MHz DDR4 ECC RDIMM（总内存 192 GB），启动盘：1 x 256 GB 英特尔固态盘 DC P4101（M.2 80 mm PCIe 3.0 x4，3D2，TLC），数据盘：1.6 TB NVM Express（NVMe）英特尔固态盘 P4510，高速缓存盘：375 GB 英特尔傲腾固态盘 DC P4800X U.2 NVMe 固态盘，数据网络：1 x 10 Gb 英特尔以太网融合网络适配器 X722，管理网络：集成的 1 Gb 以太网（GbE）端口 0/RMM 端口。软件：CentOS Linux 版本 7.6.1810/Red Hat Enterprise Linux（RHEL）7，英特尔数学核心函数库（英特尔 MKL）版本 2019 update 4，英特尔 OpenVINO 工具套件分发版 2019 R1，OpenVINO Model Server 0.4，英特尔 Python 2019 update 1 分发版，基于深度神经网络的英特尔数学核心函数库（英特尔 MKL-DNN）0.19，深度学习参考堆栈（DLRS）v4.0，Docker v18.09，Helm v2.14.3，Kubernetes v1.15.3，Kubeflow v0.6.1，Seldon Core v0.3.2，Rook v1.0.5，Ceph v14.2.1，Min.io（Rook v1.0）RELEASE.2019-04-23T23-50-36Z。在模拟的真实场景中进行扩展 — 吞吐量测试：标准化性能 1.91（英特尔超线程技术：关闭）。

性能测试结果基于截止到配置中所示日期的测试，且可能并未反映所有公开可用的安全更新。请参阅配置披露了解详细信息。没有任何产品或组件能保证绝对安全。

在性能测试过程中使用的软件及工作负载可能仅针对英特尔微处理器进行了性能优化。SYSmark 和 MobileMark 等性能测试采用特定的计算机系统、组件、软件、操作和功能进行测量。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。更多信息请访问：https://www.intel.cn/content/www/cn/zh/benchmarks/benchmark.html

所描述的降低成本方案仅用作示例，表明某些基于英特尔的产品在特定环境和配置下会如何影响未来的成本，并节约成本。环境各不相同。英特尔不保证任何成本和成本的节约。

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，及履约过程、交易过程或贸易惯例引起的任何保证。

中国印制　1119/MM/PRW/PDF　请注意环保　341602-001CN