解决方案简介

2021 年 12 月

基于 VMware Cloud Foundation 的多云分析解决方案的优势

利用针对英特尔® 架构优化的高性能第三代英特尔® 至强® 可扩展处理器和软件，部署和管理从边缘到云的数据密集型工作负载

要点综述

现代化计算环境是保持竞争力的关键。传统的应用和服务部署方法无法满足如今企业的创新需求。此外，随着数据量的增长，企业难以从数据中获得更多价值。逐渐增多的数据孤岛，繁琐的数据管理与分析流程，这使得企业难以发掘有助于增强竞争优势的重要业务洞察。更重要的是，随着零售等行业的应用向边缘过渡，在核心数据中心、云和边缘之间建立安全连接对于成功至关重要。

若要解决这些挑战，企业需要用支持多云的现代解决方案替换旧硬件和软件，这些解决方案可以加速和简化整个软件和硬件配置、部署和维护生命周期。同时，企业需要一个原生支持容器化的平台，以便高效处理数据密集型工作负载，如 AI 和机器学习。

英特尔灵活的多云分析解决方案基于 VMware Cloud Foundation，为管理虚拟机和编排容器提供了一个易于部署的平台。该解决方案有助于消除数据孤岛，并在私有云、公有云和边缘提供安全的基础实施、操作和连接。该解决方案采用英特尔的创新技术（如第三代英特尔® 至强® 可扩展处理器和英特尔® 傲腾™ 技术），可提供卓越的性能和可靠性。例如，使用针对英特尔架构优化的 TensorFlow 框架，深度学习推理吞吐量最多可提高 6 倍。[1]

业务挑战

当今的企业希望在最合理的位置灵活地运行分析工作负载，例如核心数据中心、一个或多个公有云（多云）和/或边缘。但是，为了确保这种灵活性在操作上可行，必须要通过有一种方法来高效管理任何位置的所有分析工作负载。如果没有单一管理平台，管理成本会螺旋上升并失去控制，应用开发变得不一致，性能可能会受到影响。

企业希望分析基础设施能够帮助他们缩短停机时间和设置时间，简化维护并降低管理成本，同时不影响性能。传统数据中心无法利用多云分析环境中的成本优势和新技术。这样的数据中心也无法快速灵活地满足不断变化的工作负载需求。

对于拥有过时数据中心技术的公司而言，应对这些挑战需要用支持混合云的现代分析解决方案替换传统的硬件和软件。这些解决方案可以加速整个软件和硬件配置、部署和维护生命周期以及应用开发、测试和交付。但是，无论是本地机器学习集群还是远程分支机构分析集群，公司都可能会发现组装和维护多云基础设施是一项艰巨的任务。

解决方案价值

英特尔和 VMware 联手提供多云分析解决方案，旨在帮助企业信心十足地构建多云和边缘分析解决方案。多云分析解决方案将 VMware Cloud Foundation 与创新的英特尔技术相结合，为运行和管理私有云、多云和边缘容器化分析工作负载提供了一个统一的软件定义数据中心（SDDC）平台。

VMware Cloud Foundation 是一种全栈超融合基础架构（HCI）解决方案，可简化混合/多云分析环境的部署，并帮助加速采用。它提供一整套软件定义的计算、内存、存储、网络和安全服务，以及以应用为中心的云管理功能。与英特尔技术结合使用时，VMware Cloud Foundation 可提供始终如一的高性能分析能力，减少数据中心占用空间并实现高效的运营管理。

企业可以使用端到端多云分析解决方案快速启动数据库处理和人工智能，并扩展工作负载以满足未来需求。本解决方案简介中介绍的统一云解决方案可以运行位于本地数据中心以及公有云（例如 Amazon Web Services（AWS）和 Microsoft Azure）中的容器化应用和传统虚拟机。

简言之，多云分析解决方案是一种简单、安全且敏捷的云基础设施，适用于本地、服务化公有云和边缘分析工作负载。

解决方案优势

- 为跨私有云、多云和边缘环境运行、管理和无缝连接虚拟机和容器提供一个**统一的平台**
- 通过经过验证的端到端解决方案处理各种工作负载，**加速分析部署**
- 敏捷、可扩展且**安全的基础设施**，具有出色的分析性能
- 通过针对英特尔架构优化的深度学习框架**提高吞吐量**[1]

解决方案架构亮点

英特尔的多云分析解决方案参考架构包含几个主要的 VMware 组件：VMware vSphere with Kubernetes、VMware Secure Access Service Edge（SASE）with VMware Software-Defined WAN（SD-WAN）、VMware Tanzu Mission Control、VMware vSAN、VMware NSX-T、VMware SDDC Manager 和具有基础设施即服务功能的 VMware vRealize Suite。它还包括公有云环境中的 VMware 服务 — VMware Cloud on AWS（VMC）和 Azure VMware Solution（AVS）。VMware Tanzu Kubernetes Grid（TKG）提供容器配置和生命周期管理功能。

该解决方案采用混合/多云结构，允许企业扩展可用资源，并在本地、公共云和边缘之间轻松分配分析工作负载。VMware SD-WAN 可通过公共网络从任何位置提供可靠且安全的网络连接（从本地到边缘再到公有云，反之亦然）。

VMware Cloud Foundation 包括对 Tanzu 应用目录的访问权限，该目录包含来自 Bitnami 集合的 70 多个 Kubernetes 应用和组件，这些应用和组件经过维护和验证，可在生产环境中使用。这些应用中有一些常用的分析工具，如 TensorFlow、MxNet、PyTorch 等。

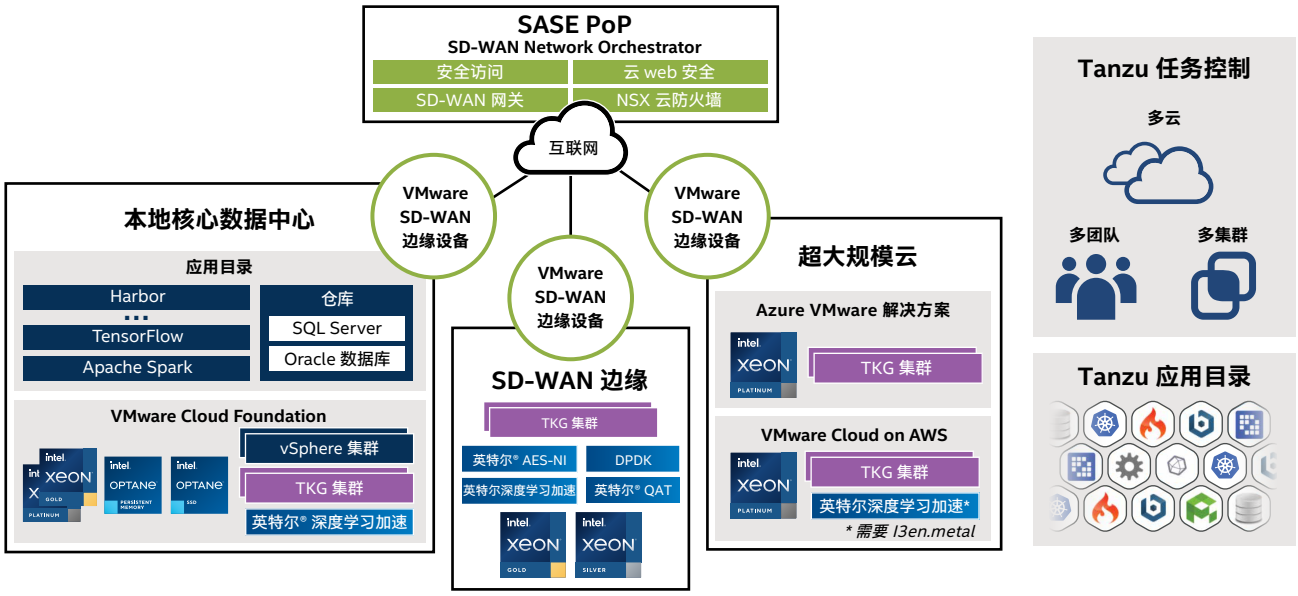

图 1. VMware 和英特尔提供面向多云分析解决方案的构建模块。

VMware Cloud Foundation 在本地核心数据中心的基础软件组件包括第三代英特尔® 至强® 可扩展处理器、英特尔® 傲腾™ 持久内存（PMem）、英特尔傲腾 SSD、英特尔® SSD D7 和 D5 系列以及英特尔® 以太网产品（见图 1）。

通过使用英特尔傲腾技术，企业可以将数据放置在更靠近 CPU 的位置，从而提高其 VMware Cloud Foundation 工作负载性能。该技术是一种非易失性内存和存储介质，填补了高性能易失性内存与低性能 NAND 存储和机械硬盘之间的空白。通过将数据放置在更靠近 CPU 的位置，英特尔傲腾技术可帮助架构师自信地部署敏捷、高性能的基础设施，进而帮助企业构建创新的分析服务并优化其基础设施投资。

英特尔傲腾技术可以通过两种方式进行部署（见图 2）：

- **英特尔傲腾持久内存**支持企业通过 DIMM 的形式获得原生持久性。企业可以近乎实时地访问、处理和分析数据，从而获取深刻洞察，改善运营并创造新的收入来源。
- **英特尔傲腾固态盘**可帮助消除数据瓶颈，加快交易速度并缩短洞察时间，让用户在需要时获得所需数据。凭借高服务质量和低队列深度下至少比 NAND 固态盘高 6 倍的性能，英特尔傲腾固态盘即使在最苛刻的环境中也能提供快速、可预测的性能。[2] 对于 vSAN 等分层存储，建议在缓存层使用英特尔傲腾固态盘，在容量层使用英特尔固态盘 D7 或 D5 系列。

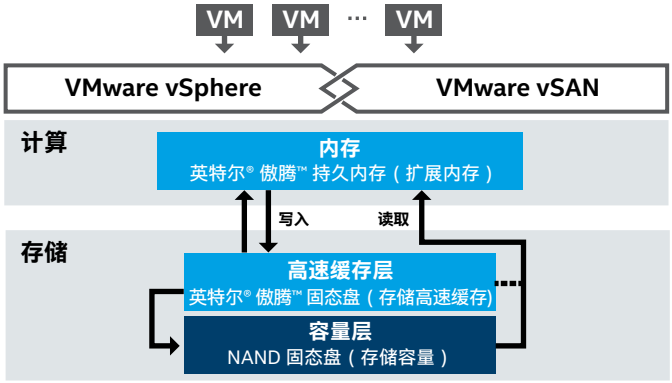

图 2. 英特尔® 傲腾™ 固态盘和英特尔® 傲腾™ 持久内存在架构中的位置。

深入了解 VMware Cloud Foundation 4.3

VMware Cloud Foundation 4.3 引入了多项新功能和增强特性，可帮助客户部署可扩展、灵活的基础设施：

- **增强的工作负载域部署和生命周期管理**，能够支持大规模虚拟机和容器架构。
- **与 VMware vSphere 7 Update 2 集成**，提供面向开发人员、支持 AI 的基础设施，提高数据安全性并帮助简化操作。
- **与 VMware vSAN 7 Update 2 集成**，增强了 vSAN 数据持久性平台，改进云原生存储和持久性服务支持。
- **增强的网络自动化**，加速和优化 NSX-T Edge 集群的扩展。
- **增强的安全操作**，包括更强大的安全机制，以改进 VMware Cloud Foundation 中安全设置的管理。

有关 VMware Cloud Foundation 4.3 新增功能的更多详细信息，请参见发布公告。

用例

通过将 VMware Cloud Foundation 与在虚拟机或容器中运行的英特尔技术相结合，企业可以支持各种用例：

深度学习推理

推理是一种计算密集型操作，可以从带有矢量神经网络指令（VNNI）的英特尔深度学习加速（英特尔® DL Boost）等创新技术中获益。矢量神经网络指令是一种能够加速推理的特殊指令集，是 VMware Cloud Foundation 4.3 平台的基础组件（从 vSphere 7 和 ESXi 7.0 开始支持该指令）。

企业需要高性能数据分析和人工智能来保持竞争力。他们需要可运行传统数据分析和人工智能应用的灵活解决方案。Vmware 多云平台包括利用英特尔硬件性能优化的组件。英特尔支持在多个解决方案堆栈层上开发机器学习工作负载。这些构建模块使企业能够快速实施分析、人工智能和机器学习工作负载，因为它们已经针对英特尔架构进行了优化，并且已经过多个生产部署的验证。因此，企业可以立即开始使用这些模块。

建议的用例展示了一种可以提高深度学习推理工作负载性能的解决方案。此用例展示了针对英特尔架构进行优化的容器堆栈所带来的推理性能提升，该容器堆栈使用特殊的 VNNI 指令集并释放了第三代英特尔至强可扩展处理器的全部潜能。

请参阅“结果”部分，了解英特尔技术和针对该技术的软件优化如何显著提升深度学习推理吞吐量。

边缘零售

对于零售商店、医疗和智能行业，在更靠近客户和数据源的位置运行工作负载可以改进性能，从而提高客户满意度。VMware Cloud Foundation 使用与公有云和私有云工作负载相同的技术，可以轻松部署和管理远程工作负载。

建议的用例展示了一种可以提高零售客户参与度并改善购物体验的解决方案。我们将介绍三个场景：

- **产品推荐。**当顾客对特定区域或部分表现出兴趣时，我们可以使用机器学习算法发送个性化的产品推荐。该算法将根据其他顾客的类似选择和顾客在商店中的位置，创建一个最相关产品的列表。顾客会收到通知，并可以使用移动应用查看个性化推荐信息。每当系统发现顾客的新兴趣时，都会向顾客发送通知。
- **存在检测。**我们使用深度学习技术和图像识别算法来检测“顾客服务位置”的顾客。店内安装的摄像头会将图像发送到深度学习管道。然后商店工作人员便会收到通知。
- **犹豫检测。**当顾客在店中漫无目的地闲逛时，业务规则引擎会认为客户正在寻找某样产品、迷路或者可能需要帮助。此时商店工作人员将收到通知（包括顾客姓名、年龄、性别和在店内的位置），以便他们快速找到并确定需要帮助的顾客。

数据仓库和分析

数据仓库被视为商业智能的一个核心组件。它们处于数据存储的中心位置，可存储来自一个或多个不同来源的数据以及当前和历史数据。VMware 混合/多云平台支持数据仓库，包括经过行业验证的基于 Microsoft SQL Server 2019 或 Oracle Database 19c 的解决方案。

结果：深度学习推理

图像分类是深度学习最热门的用例之一。我们使用来自英特尔架构优化型容器堆栈的 TensorFlow 分布和英特尔的 Model Zoo 预训练模型，对具有 int8 和 fp32 精度的 ResNet50 v1.5 拓扑进行了基准测试。

我们运行了两个测试（测试中使用的软件请见附录 A）：

- 默认 TensorFlow 容器与针对英特尔架构优化的 TensorFlow 容器的性能比较
- fp32 精度与 int8 精度的性能比较（均使用带有 VNNI 的英特尔深度学习加速和针对英特尔架构优化的 TensorFlow 容器）

如下图所示，针对推理的软硬件优化对推理性能的提升有重大影响。在这个用例中，这些优化推动了吞吐量的显著提升（每秒帧数）。VMware Cloud Foundation 4.3 平台充分说明了软件如何利用英特尔深度学习加速和 VNNI 等硬件创新技术来加速获取洞察。

通过针对英特尔架构优化 TensorFlow，吞吐量提升多达 6 倍

在此基准测试中，我们将默认 TensorFlow 容器的吞吐量性能与使用英特尔® Optimization for TensorFlow 的容器（经过优化以充分利用英特尔深度学习加速和 VNNI）进行了比较。两个容器都使用 fp32 精度。如图 3 所示，英特尔 Optimization for TensorFlow 的框架优化可为 Base 设计提供高达 5.56 倍的吞吐量提升，并为 Plus 设计提供高达 6.14 倍的吞吐量提升。[3]

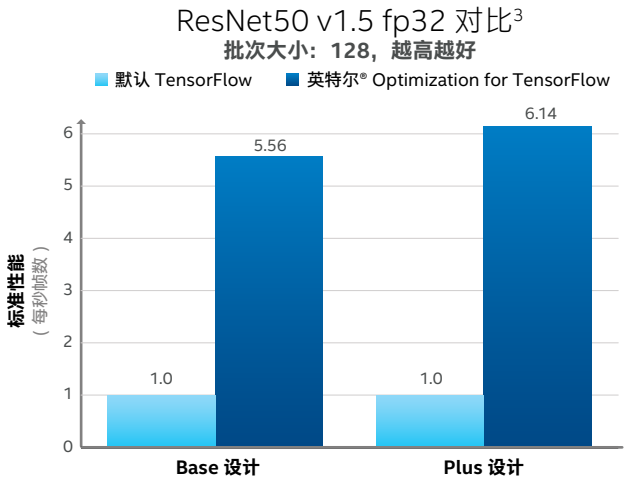

图 3. 与默认 TensorFlow 框架相比，英特尔® Optimization for TensorFlow 可提供高达 6.14 倍的吞吐量提升。

在 int8 精度下将吞吐量提高多达 4 倍

在此基准测试中，我们使用 int8 精度与 fp32 精度比较了英特尔深度学习加速与 VNNI 的吞吐量性能。两个容器都使用了英特尔 Optimization for TensorFlow。如图 4 所示，精度的小幅降低使 Base 设计的吞吐量提高了 3.44 倍，Plus 设计的吞吐量提高了 4 倍。[4]

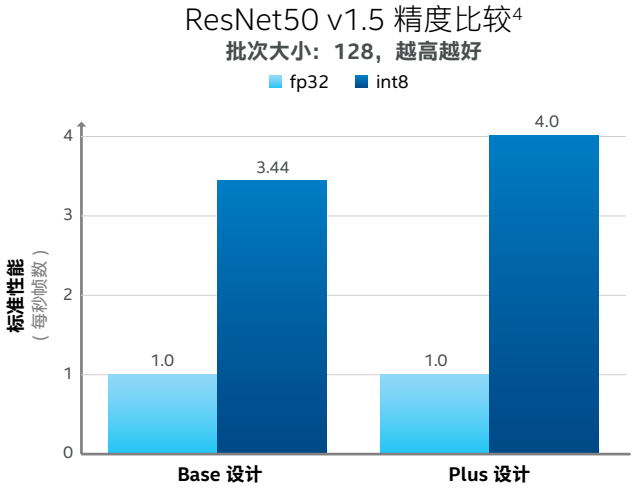

图 4. 与 fp32 精度相比，int8 精度的英特尔® Optimization for TensorFlow 将 Base 设计的吞吐量提高了 3.44 倍，并将 Plus 设计的吞吐量提高了 4 倍。

了解更多信息

- 第三代英特尔® 至强® 可扩展处理器
- 英特尔® 以太网产品
- 英特尔® 傲腾™ 持久内存
- 英特尔® 傲腾™ 固态盘
- VMware Cloud Foundation

请联系您的英特尔代表或访问英特尔和 VMware 合作伙伴关系网站。

附录 A：测试软件

表 A1. 用于测试深度学习推理的软件版本

	BASE	PLUS
客户机操作系统	Ubuntu Server 20.04.3 LTS	
客户机操作系统内核	5.4.0-88-generic	
容器	intel/intel-optimized-tensorflow: 2.5.0-ubuntu-18.04 tensorflow/tensorflow: 2.5.0	
人工智能精度	int8，fp32	
其他软件	VMware Cloud Foundation 4.3；VMware vSAN 7.0 U2a；VMware vSAN 7.0 U2a； VMware vCenter Server 7.0 U2c；VMware NSX-T 3.1.3	
其他软件（管理程序）	VMware ESXi 7.0 U2a (build 17867351)	
VM vCPU	42	56
VM vRAM	256 GB	256 GB
包含框架/工具套件的版本	TensorFlow	
Framework URL	使用的 TensorFlow Docker 映像：intel/intel-optimized-tensorflow: 2.5.0-ubuntu-18.04 和 tensorflow/tensorflow: 2.5.0	
拓扑或机器学习算法	ResNet50v1.5	
编译器	未编译，使用 Docker 镜像	
库	使用 oneAPI 优化的 TensorFlow 容器 深度神经网络库（oneDNN）	
数据集	合成数据（自动生成，--benchmark-only 参数）	
精度	int8，fp32	
Build Flag	未编译，使用 Docker 镜像	
KMP AFFINITY	granularity=fine, verbose, compact, 1, 0 'verbose, warnings, respect, granularity=fine, compact, 1, 0'	
NUMACTL	未使用	
OMP_NUM_THREADS	42/56	
使用的命令行	`python3 /tf/intel-models/benchmarks/launch_benchmark.py --in-graph ${IN_GRAPH} --model-name ${MODEL_NAME} --framework tensorflow --precision ${PRECISION} --mode inference --batch-size ${BATCH_SIZE} --benchmark-only`	

[1] 英特尔测试，截至 2021 年 10 月 11 日。结果可能不同。关于测试所用软件的详细信息，请参见附录 A。

基础配置： 单节点，2 路英特尔® 至强® 金牌 6342 处理器，1 个英特尔® 服务器主板 M50CYP2UR，总内存 = 512 GB（16 个插槽/32 GB/3200 MHz），英特尔® 超线程技术 = 关闭，英特尔® 睿频加速技术 = 开启，BIOS：SE5C6200.86B.0022.D64.2105220049（ucode: 0x0d0002b1），存储（启动）：1 个英特尔® 傲腾™ P1600X 118 GB，存储（高速缓存）：2 个英特尔® 傲腾™ 固态盘 P5800X 400 GB，存储（容量）：4 个英特尔® 固态盘 D7-P5510 3.84 TB，网络设备：1 个英特尔® 以太网适配器 E810-CQDA2（100 GbE），管理程序：VMware ESXi 7.0 U2a（build 17867351）和 VMware Cloud Foundation 4.3，操作系统/软件：Ubuntu Server 20.04.3 LTS，5.4.0-88-generic，42 vCPU，256 GB vRAM，容器：intel/intel-optimized-tensorflow: 2.5.0-ubuntu-18.04，tensorflow/tensorflow: 2.5.0，英特尔的 Model Zoo：https://github.com/IntelAI/models/tree/v2.4.0，ResNet50 v1.5，fp32，批次大小 = 128

Plus 配置： 单节点，2 路英特尔® 至强® 铂金 8362 处理器，1 个英特尔® 服务器主板 M50CYP2UR，总内存 = 1024 GB（2LM）- 256 GB（16 个插槽/16 GB/3200 MHz）+ 1024 GB 英特尔® 傲腾™ 持久内存 200 系列（8 个插槽/128 GB/3200 MHz），英特尔超线程技术 = 关闭，英特尔® 睿频加速技术 = 开启，BIOS：SE5C6200.86B.0022.D64.2105220049（ucode: 0x0d0002b1），存储（启动）：1 个英特尔傲腾 P1600X 118 GB，存储（高速缓存）：2 个英特尔傲腾固态盘 P5800X 800 GB，存储（容量）：6 个英特尔固态盘 D7-P5510 3.84 TB，网络设备：2 个英特尔以太网适配器 E810-CQDA2（100 GbE），管理程序：VMware ESXi 7.0 U2a（build 17867351）和 VMware Cloud Foundation 4.3，操作系统/软件：Ubuntu Server 20.04.3 LTS，5.4.0-88-generic，56 vCPU，256 GB vRAM，容器：intel/intel-optimized-tensorflow: 2.5.0-ubuntu-18.04，tensorflow/tensorflow: 2.5.0，英特尔的 Model Zoo：https://github.com/IntelAI/models/tree/v2.4.0，ResNet50 v1.5，fp32，批次大小 = 128

[2] 英特尔，https://www.intel.com/content/www/us/en/products/docs/memory-storage/solid-state-drives/data-center-ssds/optane-ssd-p5800x-p5801x-brief.html

[3] 请参见尾注 1。

[4] 英特尔测试，截至 2021 年 10 月 10 日。结果可能不同。关于测试所用软件的详细信息，请参见附录 A。

Base 配置： 单节点，2 路英特尔® 至强® 金牌 6342 处理器，1 个英特尔® 服务器主板 M50CYP2UR，总内存 = 512 GB（16 个插槽/32 GB/3200 MHz），英特尔® 超线程技术 = 关闭，英特尔® 睿频加速技术 = 开启，BIOS：SE5C6200.86B.0022.D64.2105220049（ucode: 0x0d0002b1），存储（启动）：1 个英特尔® 傲腾™ P1600X 118 GB，存储（高速缓存）：2 个英特尔® 傲腾™ 固态盘 P5800X 400 GB，存储（容量）：4 个英特尔® 固态盘 D7-P5510 3.84 TB，网络设备：1 个英特尔® 以太网适配器 E810-CQDA2（100 GbE），管理程序：VMware ESXi 7.0 U2a（build 17867351）和 VMware Cloud Foundation 4.3，操作系统/软件：Ubuntu Server 20.04.3 LTS，5.4.0-88-generic，42 vCPU，256 GB vRAM，容器：intel/intel-optimized-tensorflow: 2.5.0-ubuntu-18.04，英特尔的 Model Zoo：https://github.com/IntelAI/models/tree/v2.4.0，ResNet50 v1.5，int8 和 fp32，批次大小 = 128

Plus 配置： 单节点，2 路英特尔® 至强® 铂金 8362 处理器，1 个英特尔® 服务器主板 M50CYP2UR，总内存 = 1024 GB（2LM）- 256 GB（16 个插槽/16 GB/3200 MHz）+ 1024 GB 英特尔® 傲腾™ 持久内存 200 系列（8 个插槽/128 GB/3200 MHz），英特尔超线程技术 = 关闭，英特尔® 睿频加速技术 = 开启，BIOS：SE5C6200.86B.0022.D64.2105220049（ucode: 0x0d0002b1），存储（启动）：1 个英特尔傲腾 P1600X 118 GB，存储（高速缓存）：2 个英特尔傲腾固态盘 P5800X 800 GB，存储（容量）：6 个英特尔固态盘 D7-P5510 3.84 TB，网络设备：2 个英特尔以太网适配器 E810-CQDA2（100 GbE），管理程序：VMware ESXi 7.0 U2a（build 17867351）和 VMware Cloud Foundation 4.3，操作系统/软件：Ubuntu Server 20.04.3 LTS，5.4.0-88-generic，56 vCPU，256 GB vRAM，容器：intel/intel-optimized-tensorflow: 2.5.0-ubuntu-18.04，英特尔的 Model Zoo：https://github.com/IntelAI/models/tree/v2.4.0，ResNet50 v1.5，int8 和 fp32，批次大小 = 128