英特尔 On 技术创新峰会声明 - 性能指标

内容 ID: 615781

最后更新日期:

创新活动声明

主题演讲

课程代码、演讲人姓名和/或幻灯片编号	声明	声明详情
KEY100 UC/ GB	使用 Adobe Lightroom 执行从 Adobe Premiere Pro 导出视频并批量处理照片导入+色彩校正+导出，相比使用 AMD Ryzen 9 5950x 和 Nvidia RTX 3080，使用第 12 代智能英特尔酷睿 i9-12900k 和 Nvidia RTX 3080 的速度快 30%	性能结果基于 2021 年 10 月 14 日的英特尔测试，可能没有反映所有公开的更新。 使用以下操作进行测量，对比了第 12 代智能英特尔酷睿 i9- 12900K 和 AMD Ryzen 9 5950X: 处理器：第 12 代智能英特尔® 酷睿™ i9-12900K 处理器 (ADL-S) Pl1，设置为 241W TDP，16C24T (8P + 8E)；主板：预生产 Asus ROG Strix-E Z690，内存：Sk Hynix DDR5 CL 36-36-36-70，2X 32GB DDR5-4400MHz；存储：Samsung 980 Pro 1TB，显示屏分辨率：1920x1080；操作系统：Microsoft Windows 11 Pro 22000.9；显卡：NVIDIA RTX 3090 (FTW3)；显卡驱动程序：471.68；主板 BIOS 版本：0007. 处理器：AMD Ryzen 9 5950X 处理器 PL1=105W TDP，16C32T，主板：Asus ROG Crosshair Hero VIII；内存：G. Skill DDR4 CL 14-14-14-34，4X 16GB DDR4-3200 MHz；存储：Samsung 980 Pro 1TB，显示屏分辨率：1920x1080；操作系统：Microsoft Windows 11 Pro 22000.9；显卡：NVIDIA RTX 3090 (FTW3)；显卡驱动程序：471.68；主板 BIOS 版本：3801. 处理器：第 12 代智能英特尔® 酷睿™ i9-12900K 处理器 (ADL-S) Pl1，设置为 241W TDP，16C24T (8P + 8E)；主板：预生产 Asus ROG Strix-E Z690，内存：Sk Hynix DDR5 CL 36-36-36-70，2X 32GB DDR5-4400MHz；存储：Samsung 980 Pro 1TB，显示屏分辨率：1920x1080；操作系统：Microsoft Windows 11 Pro 22000.9；显卡：NVIDIA RTX 3090 (FTW3)；显卡驱动程序：471.68；主板 BIOS 版本：0007.

		处理器：AMD Ryzen 9 5950X 处理器 PL1=105W TDP，16C32T，主板：Asus ROG Crosshair Hero VIII；内存：G. Skill DDR4 CL 14-14-14-34，4X 16GB DDR4-3200 MHz；存储：Samsung 980 Pro 1TB，显示屏分辨率：1920x1080；操作系统：Microsoft Windows 11 Pro 22000.9； 显卡：NVIDIA RTX 3090 (FTW3)；显卡驱动程序：471.68；主板 BIOS 版本：3801.
KEY100 UC/GB	采用英特尔傲腾内存，基于 Linux 的工作站的内存是同类产品的 3 倍	基于技术规格，对比了 Nvidia RTX A6000（每个 PCI Express Gen 4x16 显卡的内存为 48GB）和使用英特尔傲腾持久内存（最大内存 4.5TB）的英特尔至强金牌 6238L 处理器。相比采用高达 4.5TB 英特尔傲腾持久内存和英特尔至强金牌 6238L 处理器的一流数据科学工作站，采用 Nvidia RTX A6000 的一流数据科学工作站可容纳高达 96GB 内存。
KEY100 UC/ GB	基于 Linux 的工作站可将硬件加速提升高达 100 倍。	请访问 https://medium.com/intel-analytics-software，查看相关性能指标评测详情，包括对比 Stock Scikit-Learn 测试各种数据集时的性能结果：https://medium.com/intel-analytics-software/leverage-intel-optimizations-in-scikit-learn-f562cb9d5544。
KEY100 UC/ GB	全球性能出众的游戏处理器	（截至 2021 年 10 月 1 日）采用独特特性和游戏内基准模式性能为评测标准（得分或帧率）测试大多数（共 31 款）游戏，包括对比 AMD Ryzen 5950X。其他详情请访问 http://www.intel.com/PerformanceIndex（第 12 代智能英特尔酷睿台式机处理器）。结果可能不同。
KEY100 UC/GB	极致的超频体验	基于英特尔综合工具和独特架构调优功能实现的增强超频能力。超频可能会使保修失效，也可能会影响系统状态。更多信息请见 intel.com/overclocking。结果可能不同。
KEY100 UC/ GB	高性能混合架构	高性能混合架构在单个处理器芯片上集成两种全新内核微架构：性能核（P 核）和能效核（E 核）。部分第 12 代智能英特尔® 酷睿™ 处理器（部分第 12 代智能英特尔酷睿 i5 处理器及更低版本）未采用高性能混合架构，只采用了 P 核。
KEY100 UC/ GB	英特尔硬件线程调度器	英特尔® 硬件线程调度器内置在硬件中，仅高性能混合架构配置的第 12 代智能英特尔® 酷睿™ 处理器提供该功能；需要操作系统支持。操作系统不同，可用的特性和功能会有所差异。

KEY100 UC/戴尔	相比前代产品，戴尔 Alienware Aurora R13 的性能提升了 87%，噪音降低了 16%，散热功能也得以加强，因此，客户在该机箱中运行相同显卡时，可将性能提升 5%。	2021 年 10 月采用戴尔分析进行测量，基于支持 CPU 超频至 175W 和支持 160W CPU 超频的 Aurora R12 进行的多核测试；相比 legend 2.0 工业设计，采用相同显卡时，Alienware Aurora R13 的图形性能提升了 5%。
KEY100 UC/ GB	超频创纪录演示	如欲了解更多关于创造超频记录的信息，请访问 https://hwbot.org/benchmarks/world_records。
KEY100 Greg Lavender	DeepRec 采用 OneAPI，并支持 AVX-512、VNNI 和 BF16 加速，将性能提升了 7 倍	基于截至 2021 年 10 月阿里巴巴使用定制 SKU 和软件堆栈进行的内部测量。
KEY100 Greg Lavender	英特尔 DPC++/C++ 编译器的速度比面向 SPECspeed 2017 浮点工具套件的 GCC* 快 1.8 倍	声明 ID： 07-09-2021-04 配置：截至 2021 年 6 月 10 日的英特尔测试。英特尔® 至强® 奔腾 8380 CPU @ 2.30GHz，双插槽，超线程启用，睿频加速启用，32G x16 DDR4 3200 (1DPC)。Red Hat Enterprise Linux release 8.2 (Ootpa)，4.18.0-193.el8.x86_64。软件：面向在英特尔® 64 上运行的应用的英特尔® oneAPI DPC++/C++ 编译器，版本 2021.3.0 Build 20210604。面向在英特尔® 64 上运行的应用的英特尔® C++ 英特尔® 64 编译器 Classic，版本 2021.3.0 Build 20210604_000000，GCC 11.1，Clang/LLVM 12.0.0。SPECint®_speed_base_2017 编译器开关：英特尔® oneAPI DPC++/C++ 编译器：-xCORE-AVX512 -O3 -ffast-math -flto -mfpmath=sse -funroll-loops -qopt-mem-layout-trans=4 -fiopenmp。英特尔® C++ 英特尔® 64 编译器 Classic：-xCORE-AVX512 -ipo -O3 -no-prec-div -qopt-mem-layout-trans=4 -qopt-GCC：-march=skylake-avx512 -mfpmath=sse -Ofast -funroll-loops -flto –fopenmp。LLVM：-march=skylake-avx512 -mfpmath=sse -Ofast -funroll-loops -flto -fopenmp=libomp. multiple-gather-scatter-by-shuffles -qopenmp. jemalloc 5.0.1 用于英特尔编译器、gcc 和 llvm。SPECfp®_speed_base_2017 编译器开关：英特尔® oneAPI DPC++/C++ 编译器：-xCORE-AVX512 -Ofast -ffast-math -flto -mfpmath=sse -funroll-loops -qopt-mem-layout-trans=4 -fiopenmp。英特尔® C++ 英特尔® 64 编译器

		Classic：-xCORE-AVX512 -ipo -O3 -no-prec-div -qopt-prefetch -ffinite-math-only -qopt-multiple-gather-scatter-by-shuffles -qopenmp。GCC：-march=skylake-avx512 -mfpmath=sse -Ofast -fno-associative-math -funroll-loops -flto –fopenmp。LLVM：-march=skylake-avx512 -mfpmath=sse -Ofast -funroll-loops -flto -fopenmp=libomp. jemalloc 5.0.1 用于英特尔编译器、gcc 和 llvm。
KEY100 Greg Lavender / Sandra Rivera	运行 Tensorflow 时英特尔 oneAPI 深度神经网络库的速度提高 1.5 倍	单节点，2 颗第三代英特尔至强铂金 8380，共 512 GB（16 插槽/32GB/3200）DDR4 内存，microcode 0x8d9522d4，超线程启用，睿频加速启用，Ubuntu 20.04.2 LTS(docker)，5.4.0-77-通用，TensorFlow v2.5.0（不支持 oneDNN），TensorFlow v2.6.0（支持 oneDNN），测试由英特尔 2021 年 9 月 28 日进行
KEY100 Greg Lavender	在第三代至强可扩展处理器（代号“Ice Lake”）上进行加密算法多缓冲实施时，IPP 密码库的运行速度提升高达 5.63 倍。	OpenSSL RSA Sign 2048 性能提升 5.63 倍，OpenSSL ECDSA Sign p256 性能提升 1.90 倍，OpenSSL ECDHE x25519 性能提升 4.12 倍，OpenSSL ECDHE p256 性能提升 2.73 倍，8280M: 单节点，S2600WFT 上配置 2 颗英特尔® 至强® 铂金 8280M 处理器，共 384 GB（12 插槽/ 32GB/ 2933）DDR4 内存，ucode 0x5003003，超线程启用，睿频加速禁用，Ubuntu 20.04.1 LTS，5.4.0-65-通用内核，1x INTEL_SSDSC2KG01，OpenSSL 1.1.1j，GCC 9.3.0，测试由英特尔 2021 年 3 月 5 日进行。8380: 单节点，M50CYP2SB2U 上配置 2 颗英特尔® 至强® 铂金 8380 处理器，共 512 GB（16 插槽/ 32GB/ 3200）DDR4 内存，ucode 0xd000270，超线程启用，睿频加速禁用，Ubuntu 20.04.1 LTS，5.4.0-65-通用内核，1x INTEL_SSDSC2KG01，OpenSSL 1.1.1j，GCC 9.3.0，QAT 引擎 v0.6.4，测试由英特尔 2021 年 3 月 24 日进行。8380: 单节点，M50CYP2SB2U 上配置 2 颗英特尔® 至强® 铂金 8380 处理器，共 512 GB（16 插槽/ 32GB/ 3200）DDR4 内存，ucode 0xd000270，超线程启用，睿频加速禁用，Ubuntu 20.04.1 LTS，5.4.0-65-通用内核，1x INTEL_SSDSC2KG01，OpenSSL 1.1.1j，GCC 9.3.0，QAT 引擎 v0.6.5，测试由英特尔 2021 年 3 月 24 日进行。
KEY100 Greg Lavender AITI001	Modin 是一个开源库，可将 Pandas 应用速度提高 20 倍，还可以使用 Jupyter notebook 实现从 PC	性能结果基于 2020 年 10 月 16 日的英特尔测试，可能没有反映所有公开的更新。配置详情和工作负载设置：2 颗英特尔® 至强® 铂金 8280 @ 28 核，操作系统：Ubuntu 19.10.5.3.0-64-通用 Mitigated 384GB RAM (192 GB RAM (12x 32GB 2933)。软件：Modin 0.81。

Pradeep Dubey，第 20 页幻灯片	到云端近乎无限的可扩展性，修改一行代码即可完成这一切。	Scikit-learn 0.22.2。Pandas 1.01，Python 3.8.5，DAL(DAAL4Py) 2020.2，Census Data，(21721922.45) 数据集来自 IPUMS USA，明尼苏达大学，www.ipums.org [Steven Ruggles，Sarah Flood，Ronald Goeken，Josiah Grover，Erin Meyer，Jose Pacas 和 Matthew Sobek。IPUMS USA：版本 10.0 [数据集]，明尼苏达州明尼阿波里斯市。IPUMS, 2020. https//doc.org/10.18128/D010.V10.0]，测试由英特尔 2020 年 10 月 16 日进行
KEY100 Sandra Rivera AITI001 Pradeep Dubey 第 21 页	Scikit-learn 是非常流行的机器学习库，英特尔优化可将部分模型的速度提升高达 100 倍，而且只需修改一行代码就可支持英特尔扩展指令。	性能结果基于 2020 年 10 月 23 日的英特尔测试，可能没有反映所有公开的安全更新。配置详情和工作负载设置：英特尔® oneAPI 数据分析库 2021.1 (oneDAL)。Scikit-learn 0.23.1，英特尔® Python 分发版 3.8；英特尔® 至强® 铂金 8280LCPU @ 270Ghz，双插槽，每插槽 28 核，10M 样本，10 项特性，100 个集群，100 次迭代，float32。
KEY100 Sandra Rivera AITI001 Pradeep Dubey 第 37 页	采用下一代通用 CPU 的双路服务器每秒推理超过 2.4 万张图像，相比之下，采用 Nvidia A30 GPU 的服务器每秒推理 1.6 万张。这意味着 2022 年我们提供的性能比 Nvidia 主流推理 GPU 的性能高 1.5 倍。	基准：英特尔预生产版平台采用单节点，2 颗下一代英特尔至强可扩展处理器（代号 Sapphire Rapids，超过 40 核），512 GB DDR 内存 (8(1DPC)/64GB/4800 MT/s)，超线程启用，睿频加速启用，Ubuntu Linux 18.04.4 LTS，内部预生产 BIOS 和软件运行 ResNet50-v1.5，BS=504，INT8（支持英特尔内部优化），测试由英特尔 2021 年 10 月 11 日进行 同类产品：EPYC 7742@2.25Ghz，采用 1x NVIDIA A30，GIGABYTE G482-Z52-00，TensorRT 8.0，批次大小 = 128，21.08-py3，INT8，数据集：Synthetic。上次更新日期：2021 年 9 月 27 日，资料来源：https://developer.nvidia.com/deep-learning-performance-training-inference。
KEY100 Sandra Rivera	将模型从 FP32 量化至 INT8 数值格式时，开发人员使用 INC 可将工作效率提升高达 18 倍（相比 NV TensorRT）。	基准： 英特尔预生产版平台采用单节点，1 颗下一代英特尔至强可扩展处理器（代号 Sapphire Rapids，超过 40 核），512 GB DDR 内存 (8(1DPC)/64GB/4800 MT/s)，超线程启用，睿频加速禁用，CentOS Linux 8.4，内部预生产 BIOS 和软件使用 TensorFlow 2.6 运行 SSD-ResNet34 BS=1（支持英特尔内部优化），测试由英特尔 2021 年 10 月 25 日进行。 同类产品： 单节点，1 颗 AMD EPYC 7742 64 核处理器，A100-40GB-PCIE，8x32GB DDR4-3200 内存，超线程启用，

		睿频加速禁用，Ubuntu 20.04，TensorRT 8.0.1.6 使用 CUDA 11.3，CUDNN 8.2.1.32 运行 SSD-ResNet34，测试由英特尔 2021 年 10 月 18 日进行。
KEY100 Lisa Pearce	在本示例中，DeepLink 将转码速度提升了高达 40%。	相比仅使用英特尔 Arc 显卡，通过集成了英特尔 Xᵉ 显卡 + 独立英特尔 Arc 显卡的内部版本 HandBrakeon，将视频编码过程中的 FPS 提升了高达 40%。Handbrake 在 Alchemist 预生产芯片上运行。截止到 2021 年 10 月。

Tech Insights

AITI001 Pradeep Dubey 第 13 页	第三代至强对比面向 Census 的 Nvidia Ampere A100 的端到端 AI 性能	第三代英特尔至强铂金 8380 CPU：2 颗第三代英特尔至强铂金 8380，共 512GB（16 插槽/ 32GB/ 3200MHz）DDR4 内存，microcode 0x8d055260，超线程启用，睿频加速启用，Ubuntu 20.04.2 LTS，5.4.0-65-通用内核，1 块英特尔 SSDSC2KG960G8，Python 3.7.9，Modin 0.8.3，Omniscidbe v5.4.1，scikit-learn v0.24.1（通过 daal4py v2021.2 加速），测试由英特尔 2021 年 3 月 15 日进行。 Nvidia Ampere A100 GPU：托管在 2x AMD EPYC 7742 CPU 的 Nvidia Ampere A100 GPU，共 512GB（16 插槽/ 32GB/ 3200MHz）DDR4 内存，microcode 0x8301034，超线程启用，睿频加速启用，Ubuntu 18.04.5 LTS, 5.4.0-42-通用内核，1 块 SAMSUNG 3.5TB SSD，Python 3.7.9，RAPIDS0.17，cuDF 0.17，cuML 0.17，scikit-learn v0.24.1，CUDA 11.0.221，测试由英特尔 2021 年 2 月 4 日进行。Census 数据 [21721922, 45]：数据集来自 IPUMS USA，明尼苏达大学，www.ipums.org [Steven Ruggles，Sarah Flood，Ronald Goeken，Josiah Grover，Erin Meyer，Jose Pacas 和 Matthew Sobek。IPUMS USA：版本 10.0 [数据集]。明尼苏达州明尼阿波里斯市：NIPS，2020 年。https://doi.org/10.18128/D010.V10.0]
AITI001 Pradeep Dubey 第 13 页	第三代至强对比面向 PlasticC 的 Nvidia Ampere A100 的端到端 AI 性能	第三代英特尔至强铂金 8380 CPU：2 颗第三代英特尔至强铂金 8380，共 512GB（16 插槽/ 32GB/ 3200MHz）DDR4 内存，microcode 0x8d055260，超线程启用，睿频加速启用，Ubuntu 20.04.2 LTS，5.4.0-65-通用内核，1 块英特尔 SSDSC2KG960G8，Python 3.7.9，Modin 0.8.3，Omniscidbe v5.4.1，XGBoost 1.3.3，测试由英特尔 2021 年 3 月 15 日进行。

		Nvidia Ampere A100 GPU：托管在 2x AMD EPYC 7742 CPU 的 Nvidia Ampere A100 GPU，共 512GB（16 插槽/ 32GB/ 3200MHz）DDR4 内存，microcode 0x8301034，超线程启用，睿频加速启用，Ubuntu 18.04.5 LTS，5.4.0-42-通用内核，1 块 SAMSUNG 3.5TB SSD，Python 3.7.9，RAPIDS0.17，cuDF 0.17，XGBoost 1.3.0dev.rapidsai0.17，CUDA 11.0.221，测试由英特尔 2021 年 2 月 4 日进行。 PLAsTiCC 数据训练集：(1421705, 6)；测试集: (189022127, 6)。数据集来自 Kaggle challenge“PLAsTiCC Astronomical Classification”https://www.kaggle.com/c/PLAsTiCC-2018/data
AITI001 Pradeep Dubey 第 13 页	第三代至强对比面向 DLSA 的 Nvidia Ampere A100 的端到端 AI 性能	第三代英特尔至强铂金 8380 CPU：2 颗第三代英特尔至强铂金 8380，共 512GB（16 插槽/ 32GB/ 3200MHz）DDR4 内存，microcode 0xd0002b1，超线程禁用，睿频加速启用，Ubuntu 20.04 LTS，5.4.0-84-通用内核，1 块英特尔 960GB SSD，英特尔® Extension for PyTorch v1.8.1，Transformers 4.6.1，MKL 2021.3.0，Bert-large-uncased (https://huggingface.co/bert-large-uncased) 模型，每实例 BS=1 ，20 实例/节点，4 内核/实例，测试由英特尔 2021 年 9 月 17 日进行。 Nvidia Ampere A100 GPU：托管在 2x AMD EPYC 7742 CPU 的 Nvidia Ampere A100 GPU，共 1024GB（16 插槽/ 64GB/ 3200MHz）DDR4 内存，microcode 0x8301034，超线程禁用，睿频加速启用，Ubuntu 20.04 LTS，5.4.0-80-通用内核，1 块 SAMSUNG 3.5TB SSD，PyTorch 1.8.1，Transformers 4.6.1，CUDA 11.1，Bert-large-uncased (https://huggingface.co/bert-large-uncased) 模型，每实例 BS=1，共 7 个实例，且 MIG 已启用，测试由英特尔 2021 年 9 月 22 日进行。
AITI001 Pradeep Dubey 第 13 页	第三代至强对比面向 DIEN 的 Nvidia Ampere A100 的端到端 AI 性能	第三代英特尔至强铂金 8380 CPU：Coyote Pass 上的单节点，2 颗第三代英特尔至强铂金 8380，共 512 GB（16 插槽/ 32GB/ 3200）DDR4 内存，microcode 0xd0002b1，超线程禁用，睿频加速启用，Ubuntu 20.04 LTS,5.4.0-84-通用内核，1 块英特尔 960GB SSD 操作系统盘，Modin 0.10.2，Intel-tensorflow-avx512 2.6.0，oneDNN v2.3，测试由英特尔 2021 年 9 月 29 日进行 Nvidia Ampere A100 GPU：Nvidia DGXA100 920-23687-2530-000 上的单节点，2 颗 AMD EPYC 7742，采用 1x A100 GPU，共 1024 GB（16 插槽/ 64GB/ 3200）DDR4 内存，microcode 0x8301034，超

		线程禁用，睿频加速启用，Ubuntu 20.04 LTS，5.4.0-84-通用内核，1 块 SAMSUNG 3.5TB SSD 操作系统盘，Modin 0.10.2，tensorflow 2.6.0+nv，CUDA 11.4，测试由英特尔 2021 年 9 月 29 日进行
AITI001 Pradeep Dubey Slide 20	英特尔使用 oneAPI 内核构建模块（如 oneMKL）优化这些库（NumPy 和 SciPy），以将速度提升高达 100 倍。	**新特性：** Coyote Pass 平台上的单节点，2 颗第三代英特尔至强 8368Q，共 512GB（16 插槽/ 32GB/ 3200[运行速度 2933]）DDR4 内存，microcode 0xd0002a0，超线程启用，睿频加速启用，CentOS Linux 7，3.10.0-1160.36.2.el7.x86_64，1 块 1TB 固态盘，iBench https://github.com/IntelPython/ibench，英特尔 Python 分发版 2021.4，测试由英特尔 2021 年 10 月 13 日进行 **基准：** Coyote Pass 平台上的单节点，2 颗第三代英特尔至强 8368Q，共 512GB（16 插槽/ 32GB/ 3200[运行速度 2933]）DDR4 内存，microcode 0xd0002a0，超线程启用，睿频加速启用，CentOS Linux 7，3.10.0-1160.36.2.el7.x86_64，1 块 1TB 固态盘，iBench https://github.com/IntelPython/ibench，NumPy 1.20.3 (pypi)，SciPy 1.7.1 (pypi)，测试由英特尔 2021 年 10 月 13 日进行
AITI001 Pradeep Dubey 第 36 页	正如大家在开幕主题演讲中所看到的，我们演示了如何使用 oneDNN 优化和神经压缩器提升性能，加上推理优化自动化实现的工作效率提升和 Sapphire Rapids 上实现的提升，性能共提升了高达 30 倍！	基准：单节点，2 颗第三代英特尔至强铂金 8380，共 512 GB（16 插槽/32GB/3200）DDR4 内存，microcode 0x8d9522d4，超线程启用，睿频加速启用，Ubuntu 20.04.2 LTS(docker)，5.4.0-77-通用内核，TensorFlow v2.5.0（不支持 oneDNN），TensorFlow v2.6.0（支持 oneDNN），测试由英特尔 2021 年 9 月 28 日进行 新特性： 英特尔预生产版平台采用单节点，2 颗下一代英特尔至强可扩展处理器（代号 Sapphire Rapids，超过 40 核），512 GB DDR 内存 (8(1DPC)/64GB/4800 MT/s)，超线程启用，睿频加速启用，CentOS Linux 8.4，内部预生产 BIOS 和软件使用 TensorFlow 2.6 运行 SSD-ResNet34 BS=1（支持英特尔内部优化），测试由英特尔 2021 年 9 月 28 日进行
AITI001 Pradeep Dubey 第 41 页	Gaudi 加速器将这种高效性带给了 Amazon EC2 实例训练，相比当前基于 GPU 的实例，其性	性价声明由 AWS 做出，基于 AWS 内部性能测试和 https://aws.amazon.com/ec2/pricing/on-demand/ 上的公开定价。英特尔子公司 Habana Labs 不对第三方资料进行控制或审计；实际性价比可能有所差异。

	价比提升了高达 40%，因此 AWS 客户可显著提高训练效率，降低成本。	
CLDTI002 Kamhout/Weekly 第 29 页 CLD005 Arijit Biswas 第 13 页	卸载 DSA 后 CPU 内核生命周期延长 39%。在包含 4 个 Data Streaming Accelerator (DSA) 实例的开放虚拟交换机用例中，CPU 占用率降低了近 40%，数据传输性能提升了 2.5 倍。	结果经过估算或模拟得出，基于截至 2021 年 7 月在预生产软硬件上进行的测试。 平台：Archer City SDV；CPU：Sapphire Rapids C1 预生产版，内存：128GB DDR5 (16GB PC5-4800E)；BIOS：EGSDCRB1.86B.0056.D18.2104081151；操作系统：Ubuntu 20.04；NIC：2x100Gb/s E810 (CVL)；虚拟交换机：OVS 2.15.9；数据平面：DPDK 21.08-rc0。
CLD005 Arijit Biswas 第 17 页 CLD001 Giri 第 19 页 CLDTI002, Kamhout/Weekly 第 25 页	在微服务性能方面，我们展示了每内核吞吐量（在 p99 的时延 SLA <30 毫秒的情况下）提升： 高达 24%（Ice Lake 第三代至强对比第二代至强）； 高达 69%（下一代至强 (Sapphire Rapids) 对比第二代至强）。	工作负载：DeathStarBench 'hotelReservation', 'socialNetwork' (https://github.com/delimitrou/DeathStarBench) 和谷歌微服务演示 (https://github.com/GoogleCloudPlatform/microservices-demo) 操作系统：Ubuntu 20.04，内核版本 v5.10，Kubernetes v1.21.0；截至 2021 年 7 月的测试。 在 us-west2b 中的 AWS M5.metal 实例（双路，24 核，8259CL，采用 384GB DDR4 RAM 和 25Gbps 网络）的 3 节点 Kubernetes 设置下测量第二代至强的结果。 在 3 节点双路系统（32 核，2.5GHz，300W TDP SKU，采用 512GB DDR4 RAM 和 40Gbps 网络）上测量第三代至强（代号“Ice Lake”）的结果。
CLD005 Arijit Biswas 第 15 页 CLDTI002 Kamhout/Weekly，第 16 页	使用 Zlib L9 压缩算法，CPU 占用率下降了高达 50 倍（例如，预期内核占用率降低 98%）	英特尔截至 2021 年 7 月对结果进行了评估或模拟。Sapphire Rapids 估算结果基于架构模型对在第三代至强上得出的基准测量结果的缩放。 使用第三代至强和英特尔 QAT 进行基准测试：平台：Ice Lake XCC，SKU：8380，内核：40，频率：2.3 GHz，TDP：270W，LLC：60MB，主板：Coyote Pass，RAM：16 x 32GB DDR4 3200，Hynix HMA84GR7CJR4N-XN，BIOS：SE5C6200.86B.3021.D40.2103160200，Microcode：0x8d05a260 (03/16)

		操作系统：Ubuntu 20.04.2，内核：5.4.0-65-通用内核，GCC：9.3.0，yasm：1.3.0，nasm：2.14.02，ISA-L：2.3，ISA-L Crypto：2.23，OpenSSL：1.1.1i，zlib：1.2.11，lzbench：1.7.3
CLDTI002，第 14 页的 Kamhout/Weekly Verbal	加密加速示例：NGINX TLS Webserver 加速 7 倍	NGNIX 1.20.1 由英特尔 2021 年 9 月 13 日使用基于第三代至强可扩展处理器 (Ice Lake) 的 n2-standard-1 和 -16 16 (us-central1-a region) 实例，使用 IFMA 加密指令，对比启用或不启用 qatengine 时的加密性能。使用 TLS version 1.2 测量，使用的 ECDH Curves：secp384r1，Cipher： ECDHE-RSA-AES128-GCM-SHA256
CLITI001 Chris Kelly 展示演示	无声明	**平台：** ADL Whitebox **处理器：** 英特尔® 酷睿™ i9-12900K 处理器 (16C/24T) **内存：** 2x16GB DDR5 4800 MHz SDRAM **存储：** 三星 SSD 980 PRO (1 TB) **显示屏分辨率：** 1920 x 1080 显示屏 **操作系统：** Microsoft Windows 11 Pro，Build 22000.120 **显卡：** RTX 3090
CLITI001 Chris Kelly	混合架构结合了性能核（P 核）和能效核（E 核），可在相同功耗下提供更出色的多线程性能，因此是一种更高效的方法	估算值基于预生产英特尔内部 Alder Lake 验证平台（采用 8 个 P 核和 8 个 E 核），对比在 P 核和 E 核上运行 Spec Int 544.nab_r。截至 2021 年 7 月。所示图表仅用于举例说明，不按比例尺绘图。估算值基于预生产英特尔内部 Alder Lake 验证平台，在 8 个 P 核/8 个 E 核平台（配置为运行 4 个 P 核、2 个 P 核和 8 个 E 核）上运行 Spec Int 544.nab_r。截至 2021 年 7 月。所示图表仅用于举例说明，不按比例尺绘图。
CLITI001 Chris Kelly	英特尔硬件线程调度器可帮助操作系统避免让可扩展性低的线程消耗宝贵的能源。	英特尔® 硬件线程调度器需要第 12 代智能英特尔® 酷睿™ 高性能混合架构和操作系统支持。操作系统不同，可用的特性和功能会是否有所差异。

CLITI001 Chris Kelly	英特尔® 硬件线程调度器演示	**平台:** ADL Whitebox **处理器:** 英特尔® 酷睿™ i9-12900K 处理器 (16C/24T) **内存:** 2x16GB DDR5 4800 MHz SDRAM **存储:** 三星 SSD 980 PRO (1 TB) **显示屏分辨率:** 1920 x 1080 显示屏 **操作系统:** Microsoft Windows 11 Pro，Build 22000.120 **显卡:** RTX 3090
CLITI001 Chris Kelly	英特尔® Evo™ - 旨在为消费者带来极致的整体笔记本体验	使用搭载第 11 代智能英特尔® 酷睿™ i7 处理器（轻薄型设备的理想之选）的英特尔® Evo™ 平台笔记本电脑测量，使用行业性能指标评测和典型使用指南测试测量了搭载第十一代智能英特尔® 酷睿™ i7-1185G7 处理器的系统，评估了独特特性，与 AMD Ryzen 7 4800U 进行了比较。如欲了解英特尔® 酷睿™ i7-1185G7 处理器为何是支持在轻薄型笔记本上办公、创作、游戏、协作和娱乐的理想处理器，请点击此处。 英特尔面向笔记本电脑的综合型雅典娜创新计划将使用高级规范和关键体验指标，测试、测量和验证所有带有英特尔 Evo 品牌徽章的设计。测试结果截至 2020 年 8 月，不保证单台笔记本电脑的性能。功耗和性能因使用、配置和其他因素而异。
CLITI001 Chris Kelly	英特尔桥接技术	手机：型号：三星 Galaxy S21 5G；型号：SM-G991N；Android 版本 11；Build RP1A。200720.012.G991NKSU3AUF6。 Chromebook：Acer Spin 713；ChromeOSversion：93.0.4577.95 (Official Build)（64 位）；CPU：第 11 代智能英特尔酷睿 i5-1135G7。
CLITI001 Chris Kelly	我们与 W3C 的合作伙伴密切合作以及对 Web 标准的开发不断丰富着 Web 平台的体验，并持续推	更多关于万维网联盟 (W3C) 与所提标准的信息，请参阅此处

<table>
<tr><td></td><td>出全新的硬件平台功能。</td><td></td></tr>
<tr><td>E5GTI001</td><td>时序推理性能</td><td><u>适用于两种示例的 Algo:</u>
标准 80/20 训练和测试分割

在两种示例中通过 nTrees=500 的标准 RandomForestClassfier 运行该训练 dataframe

注: 在英特尔示例中，使用的是标准 pandas dataframe。在 Nvidia 示例中，使用的是 cudf (Nvidia dataframe)。

对于每个测试点（1，10，100，200，500），使用 dataframe 的标头，运行推理函数 3 次，并报告平均耗时（秒）

<table><tr><td></td><td>英特尔</td><td>Nvidia</td></tr><tr><td>操作系统</td><td>Ubuntu 18.04</td><td>Ubuntu 18.04</td></tr><tr><td>Python 环境</td><td>使用英特尔 sklearn 扩展库，通过公共 conda 'intel' 渠道安装的英特尔 Python 分发版</td><td>Conda 稳定版来自 Nvidia 'rapidsai'渠道。</td></tr><tr><td>软件版本</td><td>python 3.7.9 和 scikit-learn-intelex 2021.2.2</td><td>Python 3.7，cudatoolkit=10.1，rapids 0.19</td></tr></table>
<h2>至强® SP 系统配置</h2>
<table><tr><td>CPU</td><td>产品</td><td>英特尔® 至强™ SP - 6242</td></tr><tr><td>频率</td><td>2.8GHz</td><td></td></tr><tr><td>内核数/线程数</td><td>16 核/32 线程</td><td></td></tr></table></td></tr>
</table>

高速缓存 (MB)	22	
显卡	频率	
显卡	Nividia T4	
内存		
类型	DDR4 DIMM @3200 MHz	
容量 (GB)	16x32	

注:

ICPS = 每秒推理周期

英特尔® Edge Insights for Industrial v2.3

算法：随机森林

软件环境：Ubuntu 18.04

英特尔® 分发版：英特尔 Python 分发版（通过公共 Conda 'intel'渠道）

Nvidia 分发版：Conda 稳定版（来自 NVIDIA 'rapids ai'渠道）

数据集 = Bosch Manufacturing Data dataset.shape = (406879, 102)，大小：230MB（已完成全部预处理，可用于建模）。

测试日期：2021 年 5 月。

其他的名称和品牌可能是其他所有者的资产。测试日期：2021 年 5 月。工作负载和配置请参阅备用页。结果可能不同。

GAMTI001 Roger Chandler	性能连续两年翻了一番。首先是从第九代到第 11 代，然后是从第 11 代到 Xe LP。	根据运行《杀手 2》时进行的游戏内基准测试，图形性能提升了 4 倍。基于 2020 年 11 月在 1080p 和低设置下进行的测试。测试系统: 酷睿 i7-8565U (PL1=15W)，在 HP Spectre x360 上测量：512 GB 英特尔固态盘 660p 系列，2 x 8 GB DDR4-

		2400，Windows 10 Home 19042.63，英特尔超核芯显卡 620 27.20.100.8935，F.34。 酷睿 i7-1065G7 (PL1=25W)，在 Razer Blade Stealth 13 (2019) 上测量：256GB 三星 MZVLB256HAHQ (PM981)，2 x 8 GB LPDDR4X-3733，Windows 10 Home 19042.63，英特尔锐炬 Plus 27.20.100.8935，1.02。 酷睿 i7-1185G7 (B0) (PL1=28W)，在英特尔参考平台上测量：三星 MZVLB512HBJQ，2 x 8 GB LPDDR4X-4266，Windows 10 Pro 19042.63，英特尔锐炬 Xe 27.20.100.8935，92A。

演示

展示演示 - **游戏、直播、录制和内容创作，以及多任务处理**	运行《骑马与砍杀 2:领主》时 FPS 提升高达 19% 基于在第 12 代智能英特尔® 酷睿™ i9-12900K 与第 11 代智能英特尔® 酷睿™ i9-11900K 上运行《骑马与砍杀 2:领主》时的测量结果	**游戏、直播和录制演示** 面向游戏 + 直播工作负载的 Open Broadcaster Software (OBS) - OBS 版本 27.1.2。在“Medium”编码预置前提下，使用 1920x1080 分辨率以 60FPS 和 6000kbps 编码速度进行编码和直播。视频编解码器设为默认值。在“High”图形预置前提下，使用游戏内基准测试运行《骑马与砍杀 2:领主》 性能结果基于英特尔截至 2021 年 10 月 1 日的测试，可能无法反映所有公开可用的更新。 CPU：英特尔® 酷睿 ™ i9-12900K 处理器，主板：MSI MPG Z690 CARBON WIFI (MS-7D30)，内存：32GB DDR5-44000 DDR SDRAM，显卡驱动程序版本：472.12，显卡：NVIDIA GeForce RTX3090 Founders 版本，存储固态盘：三星 SSD 980 PRO 1TB，操作系统版本：Microsoft Windows 11 Professional (x64) Build 22000.194，OBS 版本：27.0.1，《骑马与砍杀 2:领主》e1.6.2.284832 CPU：英特尔® 酷睿 ™ i9-11900K 处理器，主板：MSI MPG Z590 CARBON WIFI (MS-7D06)，内存：32GB DDR4-32000 DDR SDRAM，显卡驱动程序版本：472.12，显卡：NVIDIA GeForce RTX3090 Founders 版本，存储固态盘：三星 SSD 980 PRO 1TB，操作系统版本：Microsoft Windows 11 Professional (x64) Build 22000.194，OBS 版本：27.0.1，《骑马与砍杀 2:领主》e1.6.2.284832
展示演示 - **游戏、直播、录制和内容**	使用第 12 代智能英特尔® 酷睿™ i9-12900K	**游戏、直播和录制演示**

创作，以及多任务处理	处理器运行《骑马与砍杀 2:领主》，同时使用 OBS 进行游戏、直播和录制时，FPS 提升了高达 84%	面向游戏 + 直播工作负载的 Open Broadcaster Software (OBS) - OBS 版本 27.1.2。在“Medium”编码预置前提下，使用 1920x1080 分辨率以 60FPS 和 6000kbps 编码速度进行编码和直播。视频编解码器设为默认值。在“High”图形预置前提下，使用游戏内基准测试运行《骑马与砍杀 2:领主》 性能结果基于英特尔截至 2021 年 10 月 1 日的测试，可能无法反映所有公开可用的更新。 CPU：英特尔® 酷睿 ™ i9-12900K 处理器，主板：MSI MPG Z690 CARBON WIFI (MS-7D30)，内存：32GB DDR5-44000 DDR SDRAM，显卡驱动程序版本：472.12，显卡：NVIDIA GeForce RTX3090 Founders 版本，存储固态盘：三星 SSD 980 PRO 1TB，操作系统版本：Microsoft Windows 11 Professional (x64) Build 22000.194，OBS 版本：27.0.1，《骑马与砍杀 2:领主》e1.6.2.284832 CPU：英特尔® 酷睿 ™ i9-11900K 处理器，主板：MSI MPG Z590 CARBON WIFI (MS-7D06)，内存：32GB DDR4-32000 DDR SDRAM，显卡驱动程序版本：472.12，显卡：NVIDIA GeForce RTX3090 Founders 版本，存储固态盘：三星 SSD 980 PRO 1TB，操作系统版本：Microsoft Windows 11 Professional (x64) Build 22000.194，OBS 版本：27.0.1，《骑马与砍杀 2:领主》e1.6.2.284832
展示演示 - **游戏、直播、录制和内容创作，以及多任务处理**	顺序内容创作性能提升高达 29% 使用顺序多任务处理内容创作工作流测量，对比第 12 代智能英特尔® 酷睿™ i9-12900K 和第 11 代智能英特尔® 酷睿™ i9-11900K	**内容创作演示** 2 个工作负载同时运行，都进行了测量。第一个工作负载在 Adobe Premiere Pro 中。该工作负载测量 Adobe Premiere Pro 导出带有（2 个）音轨和 Lumetri 色彩效果的 11:19 序列 AVC 剪辑所需的时间。该视频通过硬件加速导出。第二个工作负载在 Adobe Lightroom Classic 中。该工作负载测量 Adobe Lightroom Classic 将 100 张图像导入目录（导入期间应用预设），然后将文件导出为 60% 质量的 JPEG 所需的时间。 该视频源为一组以大约 90 Mb/s 的比特率录制的 4K (3840X2160)，H.264，29.97 FPS，MP4 视频片段。该序列包含 .MP3 音频。视频剪辑应用了 Lumetri 色彩效果。序列长度为 11 分 19 秒。导出的预设基于名为“YouTube 2160p 4K 超高清”的 Premiere Preset，但‘Time Interpolation'设为“帧混合”。

		性能结果基于英特尔截至 2021 年 10 月 1 日的测试，可能无法反映所有公开可用的更新。 CPU：英特尔® 酷睿 ™ i9-12900K 处理器，主板：Gigabyte Aorus Z690 Sabre Master，内存：64GB DDR5-44000 DDR SDRAM，显卡驱动程序版本：472.12，显卡：NVIDIA GeForce RTX3080，存储 SSD：PCIe Gen4 NVMe SSD，操作系统版本：Microsoft Windows 11 Professional (x64)，Adobe Lightroom Classic ver 10.4，Adobe Premiere Pro ver 15.4.1 CPU：英特尔® 酷睿 ™ i9-11900K 处理器，主板：ROG Maximus XIII Hero，内存：64GB DDR4-32000 DDR SDRAM，显卡驱动程序版本：472.12，显卡：NVIDIA GeForce RTX3080，存储 SSD：PCIe Gen4 NVMe SSD，操作系统版本：Microsoft Windows 11 Professional (x64)，Adobe Lightroom Classic ver 10.4，Adobe Premiere Pro ver 15.4.1
展示演示 - **游戏、直播、录制和内容创作，以及多任务处理**	并发内容创作性能提升高达 47% 采用多任务处理内容创作工作流测量，对比了第 12 代智能英特尔® 酷睿™ i9-12900K 和第 11 代智能英特尔® 酷睿™ i9-11900K	**内容创作演示** 两个工作负载同时运行，均进行了测量。第一个工作负载在 Adobe Premiere Pro 中。该工作负载测量 Adobe Premiere Pro 导出带有（2 个）音轨和 Lumetri 色彩效果的 11:19 序列 AVC 剪辑所需的时间。该视频通过硬件加速导出。第二个工作负载在 Adobe Lightroom Classic 中。该工作负载测量 Adobe Lightroom Classic 将 100 张图像导入目录（导入期间应用预设），然后将文件导出为 60% 质量的 JPEG 所需的时间。该视频源为一组以大约 90 Mb/s 的比特率录制的 4K (3840X2160), H.264，29.97 FPS，MP4 视频片段。该序列包含 .MP3 音频。视频剪辑应用了 Lumetri 色彩效果。序列长度为 11 分 19 秒。导出的预设基于名为“YouTube 2160p 4K 超高清”的 Premiere Preset，但 'Time Interpolation'设为“帧混合”。 CPU：英特尔® 酷睿 ™ i9-12900K 处理器，主板：Gigabyte Aorus Z690 Sabre Master，内存：64GB DDR5-44000 DDR SDRAM，显卡驱动程序版本：472.12，显卡：NVIDIA GeForce RTX3080，存储 SSD：PCIe Gen4 NVMe SSD，操作系统版本：Microsoft Windows 11 Professional (x64)，Adobe Lightroom Classic ver 10.4，Adobe Premiere Pro ver 15.4.1 CPU：英特尔® 酷睿 ™ i9-11900K 处理器，主板：ROG Maximus XIII Hero，内存：64GB DDR4-32000 DDR SDRAM，显卡驱动程序版本：472.12，显卡：NVIDIA GeForce RTX3080，存储 SSD：PCIe Gen4

		NVMe SSD，操作系统版本：Microsoft Windows 11 Professional (x64)，Adobe Lightroom Classic ver 10.4，Adobe Premiere Pro ver 15.4.1
展示演示 - **采用英特尔® XTU 对第 12 代智能英特尔® 酷睿™ 处理器实施超频**	“极致的超频体验”基于英特尔综合工具和独特架构调优功能实现的增强超频能力。结果可能会有所不同。超频可能会使保修失效，也可能会影响系统状态。详情请访问 intel.com/overclocking	探索英特尔® Extreme Tuning Utility（英特尔® XTU）的新特性，该实用程序支持超频新手和专业人士对系统实施超频、监测和施压。XTU 支持我们对全新第 12 代台式机处理器实现终极调优和控制。 CPU：第 12 代智能英特尔® 酷睿™ i9-12900KF 处理器，主板：Asus Strix-E Z690 主板，内存：32GB (2x16GB) DDR5-6000MHz XMP3.0，显卡驱动程序版本：471.68，显卡：EVGA NVIDIA GeForce RTX3090 FTW3，存储 SSD：Samsung 980 Pro 1TB PCIe Gen4 NVMe SSD，操作系统版本：Windows 11 Pro Version 22000.194 性能结果基于截至 2021 年 10 月 1 日的英特尔测试，可能无法反映所有公开可用的更新。
Python 数据科学演示	出租车演示部分	对于 Python 数据科学 NYC 出租车演示，请注意，为了运行演示所示的 Nvidia RAPIDS 笔记本电脑，我们分别尝试了 3 次，每次都导致错误。以下尝试为： 1) 在 AWS（只在演示中提到，未显示）上，在 p3.16x 大型 Ubuntu 20.04 实例，其中 CUDA（NVidia 显卡驱动程序和其他与系统相关的库）由 Amazon 预装。按照 CUDF 存储库的指令安装 Anaconda 程序包版本 11.2。笔记本电脑无法执行复杂的“查询”语句，显示“未知错误”。由于该错误含糊不清，因此后续无法成功解决该错误。 2) 尝试按照第一次的指令，使用相同的 README 安装夜间构建（未在演示中显示），但是遗憾的是，另一个奇怪的错误导致 dask_cuda 集群甚至无法启动。请注意，相比稳定版本，安装夜间环境的难度更大，因为 RAPIDS 笔记本电脑有众多关联组件和夜间通道，引起很多冲突版本。 3) 在联想工作站（演示视频中显示）上，按照 NVidia 网页的指令安装面向 Ubuntu 20.04 的显卡驱动程序和 CUDA 库，然后使用 CUDF 存储库页面的指令来安装面向 CUDF 11.2 的 Anaconda 软件包。这一次，“查询”语句成功执行，但是在 XGBoost 训练（在演示中显示）时出现错误。错误说明同样无法明确解释哪里出错以及如何改进以避免错误。对此，请勿尝试在第 2 次尝试后进行夜间安装，因为不稳定性导致几乎不可能为 Anaconda 环境配备所有必需的软件包

演示代码-TBC 标题 微服务性能优化	POC web 服务器使用动态负载均衡器 (DLB)，在软件内核上进行负载均衡。声明: 1.时延降低高达 - 22-42% 2.周期利用率降低高达-30-60%	硬件：1 节点，英特尔预生产平台上的双路英特尔至强可扩展处理器（预生产 Sapphire Rapids），256 GB DDR5 内存，0x8c0003b0，超线程启用，睿频启用，2 x P4510 2TB，Cent OS 8.4, 5.12.0-0507.intel_next，英特尔于 2021 年 9 月 24 日测试。基准： Nginx 1.18.0 新配置 1：内部预生产 web 服务器（类似于 NGINX，规模更小），具有软件负载均衡 新配置 2：内部预生产 web 服务器（类似于 NGINX，规模更小），具有动态负载均衡硬件加速器

课程

AI001，Meena Arunachalam，#13	MacBook Pro 的性能比 Spark 集群（AWS 上的 m3.xlarge 实例）高 1.7 倍，双插槽单节点 Cascade Lake 的性能比 Spark 集群高 15.7 倍。	**由 OmniSci 测试** **第二代英特尔至强金牌 6226R CPU**：双路第二代英特尔至强金牌 6226R，384GB（12 个插槽/ 32GB/ 2933MHz）DDR4 总内存，微代码 0x5003003，超线程启用，睿频启用，Ubuntu 20.04.2 LTS, 5.4.0-65-generic 内核，1x 英特尔 SSDSC2KG960G8，Python 3.7.9，Modin 0.8.3，Omniscidbe v5.4.1，由 daal4py v2021.2 加速的 scikit-learn v0.24.1，OmniSci 于 2021 年 3 月 15 日执行测试。 **MacBook Pro**：第九代 i9-9880H CPU，64GB (2666MHz) DDR4 总内存，超线程启用，睿频启用，Mac OS X ver 11，1x4TB 固态盘，Python 3.7.9，Omnisci DB v5.4.1，OmniSci 于 2021 年 3 月 15 日执行测试。
AI001，Meena Arunachalam，#14	MacBook Pro 的性能比 Spark 集群（AWS 上的 m3.xlarge 实例）高 1.7 倍，双插槽单节点 Cascade Lake 的性能比 Spark 集群 15.7 倍。 采用傲腾内存的至强可在内存中处理 1.2 Bn NYC Taxi 基准测试数据。	**由 OmniSci 测试** **第二代英特尔至强金牌 6226R CPU**：双路第二代英特尔至强金牌 6226R，384GB（12 个插槽/ 32GB/ 2933MHz）DDR4 总内存，微代码 0x5003003，超线程启用，睿频启用，Ubuntu 20.04.2 LTS, 5.4.0-65-generic 内核，1x 英特尔 SSDSC2KG960G8，Python 3.7.9，Modin 0.8.3，Omniscidbe v5.4.1，由 daal4py v2021.2 加速的 scikit-learn v0.24.1，OmniSci 于 2021 年 3 月 15 日执行测试。 **MacBook Pro**：第九代 i9-9880H CPU，64GB (2666MHz) DDR4 总内存，超线程启用，睿频启用，Mac OS X ver 11，1x4TB 固态盘，Python 3.7.9，Omnisci DB v5.4.1，OmniSci 于 2021 年 3 月 15 日执行测试。

		第二代英特尔至强铂金 8276L CPU：双路第二代英特尔至强铂金 8276L，DDR4 总内存为 384GB（12 个插槽/ 32GB/ 2666Mhz）和 4TB 第一代傲腾内存 (2666Mhz)，微代码 0x5003003，超线程启用，睿频启用，Ubuntu 20.04.2 LTS，5.4.0-65-generic 内核，1x 英特尔 SSDSC2KG960G8，Python 3.7.9，Modin 0.8.3，Omniscidbe v5.4.1，由 daal4py v2021.2 加速的 scikit-learn v0.24.1，英特尔于 2021 年 3 月 15 日执行测试。
AI001，Meena Arunachalam，#24	在 Census 上执行训练与推理时，相比 DGX A100（利用 1xA100），至强 8380 的机器学习性能提升 5 倍以上。	**第三代英特尔至强铂金 8380 CPU**：双路第三代英特尔至强铂金 8380，512GB（16 个插槽/ 32GB/ 3200MHz）DDR4 总内存，微代码 0x8d055260，超线程启用，睿频启用，Ubuntu 20.04.2 LTS，5.4.0-65-generic 内核，1x 英特尔 SSDSC2KG960G8，Python 3.7.9，Modin 0.8.3，Omniscidbe v5.4.1，由 daal4py v2021.2 加速的 scikit-learn v0.24.1，英特尔于 2021 年 3 月 15 日执行测试。 **Nvidia Ampere A100 GPU**：在 2x AMD EPYC 7742 CPU 上托管的 Nvidia Ampere A100 GPU，512GB（16 个插槽/ 32GB/ 3200MHz） DDR4 总内存，微代码 0x8301034，超线程启用，睿频启用，Ubuntu 18.04.5 LTS，5.4.0-42-generic 内核，1x SAMSUNG 3.5TB 固态盘，Python 3.7.9，RAPIDS0.17，cuDF 0.17，cuML 0.17，scikit-learn v0.24.1，CUDA 11.0.221，英特尔于 2021 年 2 月 4 日执行测试。调查数据 [21721922, 45]：数据集来自 IPUMS USA、明尼苏达大学、www.ipums.org [Steven Ruggles，Sarah Flood，Ronald Goeken，Josiah Grover，Erin Meyer，Jose Pacas 和 Matthew Sobek。IPUMS USA：版本 10.0 [数据集]。明尼苏达州明尼阿波里斯市：IPUMS，2020 年。 https://doi.org/10.18128/D010.V10.0]
AI001，Meena Arunachalam，#25	对 E2E 文档级情绪分析 (DLSA) Huggingface SST 数据集 WL 进行深度学习推理，Nvidia A100 的性能比至强 8380 高两倍。	**第三代英特尔至强铂金 8380 CPU**：双路第三代英特尔至强铂金 8380，512GB（16 个插槽/ 32GB/ 3200MHz）DDR4 总内存，微代码 0xd0002b1，超线程禁用，睿频启用，Ubuntu 20.04 LTS，5.4.0-84-generic 内核，1x 英特尔 960GB 固态盘，面向 PyTorch v1.8.1 的英特尔® 扩展，Transformers 4.6.1，MKL 2021.3.0，Bert-large-uncased (https://huggingface.co/bert-large-uncased) 模型，BS=每实例 1 ，20 个实例/节点，4 个内核/实例，英特尔于 2021 年 9 月 17 日执行测试。 **Nvidia Ampere A100 GPU**：在双路 AMD EPYC 7742 CPU 上托管的 Nvidia Ampere A100 GPU，1024GB（16 个插槽/ 64GB/ 3200MHz）DDR4 总内存，微代码 0x8301034，超线程禁用，睿频启用，Ubuntu 20.04 LTS，5.4.0-80-generic 内核，1x SAMSUNG 3.5TB 固态盘，PyTorch 1.8.1，Transformers 4.6.1，CUDA 11.1，Bert-large-

		uncased (https://huggingface.co/bert-large-uncased) 模型，BS=每实例 1，总共 7 个实例，MIG 已启用，英特尔于 2021 年 9 月 22 日执行测试。
AI001，Meena Arunachalam，#24	所有阶段的端到端调查数据，至强 8380 的性能比 DGX A100（利用 1xA100）低 7%。	**第三代英特尔至强铂金 8380 CPU**：双路第三代英特尔至强铂金 8380，512GB（16 个插槽/ 32GB/ 3200MHz）DDR4 总内存，微代码 0x8d055260，超线程启用，睿频启用，Ubuntu 20.04.2 LTS，5.4.0-65-generic 内核，1x 英特尔 SSDSC2KG960G8，Python 3.7.9，Modin 0.8.3，Omniscidbe v5.4.1，由 daal4py v2021.2 加速的 scikit-learn v0.24.1，英特尔于 2021 年 3 月 15 日执行测试 **Nvidia Ampere A100 GPU**：在双路 AMD EPYC 7742 CPU 上托管的 Nvidia Ampere A100 GPU，512GB（16 个插槽/ 32GB/ 3200MHz） DDR4 总内存，微代码 0x8301034，超线程启用，睿频启用，Ubuntu 18.04.5 LTS，5.4.0-42-generic 内核，1x SAMSUNG 3.5TB 固态盘，Python 3.7.9，RAPIDS0.17，cuDF 0.17，cuML 0.17，scikit-learn v0.24.1，CUDA 11.0.221，英特尔于 2021 年 2 月 4 日执行测试。Census 数据 [21721922, 45]：数据集来自 IPUMS USA、明尼苏达大学、www.ipums.org [Steven Ruggles，Sarah Flood，Ronald Goeken，Josiah Grover，Erin Meyer，Jose Pacas 和 Matthew Sobek。IPUMS USA：版本 10.0 [数据集]。明尼苏达州明尼阿波里斯市：IPUMS，2020 年。 https://doi.org/10.18128/D010.V10.0]
AI001，Meena Arunachalam，#25	所有阶段的端到端 DLSA，至强 8380 的性能比 DGX A100（利用 1xA100）高 13%。	**第三代英特尔至强铂金 8380 CPU**：双路第三代英特尔至强铂金 8380，512GB（16 个插槽/ 32GB/ 3200MHz）DDR4 总内存，微代码 0xd0002b1，超线程禁用，睿频启用，Ubuntu 20.04 LTS，5.4.0-84-generic 内核，1x 英特尔 960GB 固态盘，面向 PyTorch v1.8.1 的英特尔® 扩展，Transformers 4.6.1，MKL 2021.3.0，Bert-large-uncased (https://huggingface.co/bert-large-uncased) 模型，BS=每实例 1，20 个实例/节点，4 个内核/实例，英特尔于 2021 年 9 月 17 日执行测试。 **Nvidia Ampere A100 GPU**：在双路 AMD EPYC 7742 CPU 上托管的 Nvidia Ampere A100 GPU，1024GB（16 个插槽/ 64GB/ 3200MHz）DDR4 总内存，微代码 0x8301034，超线程禁用，睿频启用，Ubuntu 20.04 LTS，5.4.0-80-generic 内核，1x SAMSUNG 3.5TB 固态盘，PyTorch 1.8.1，Transformers 4.6.1，CUDA 11.1，Bert-large-uncased (https://huggingface.co/bert-large-uncased) 模型，BS=每实例 1，总共 7 个实例，MIG 已启用，英特尔于 2021 年 9 月 22 日执行测试。

AI001，Meena Arunachalam，#30	DLSA 多实例配置（每插槽 10 个实例可确保卓越性能）。	**第三代英特尔至强铂金 8380 CPU：**双路第三代英特尔至强铂金 8380，512GB（16 个插槽/ 32GB/ 3200MHz）DDR4 总内存，微代码 0xd0002b1，超线程禁用，睿频启用，Ubuntu 20.04 LTS，5.4.0-84-generic 内核，1x 英特尔 960GB 固态盘，面向 PyTorch v1.8.1 的英特尔® 扩展，Transformers 4.6.1，MKL 2021.3.0，Bert-large-uncased (https://huggingface.co/bert-large-uncased) 模型，BS=每实例 1 ，20 个实例/节点，4 个内核/实例，英特尔于 2021 年 9 月 17 日执行测试。
AI001，Meena Arunachalam，#31；另请参见 AI 效率与性能演示	结果：更高的性价比	**第三代英特尔至强铂金 8380 CPU：**双路第三代英特尔至强铂金 8380，512GB（16 个插槽/ 32GB/ 3200MHz）DDR4 总内存，微代码 0x8d055260，超线程启用，睿频启用，Ubuntu 20.04.2 LTS，5.4.0-65-generic 内核，1x 英特尔 SSDSC2KG960G8，Python 3.7.9，Modin 0.8.3，Omniscidbe v5.4.1，由 daal4py v2021.2 加速的 scikit-learn v0.24.1，英特尔于 2021 年 3 月 15 日执行测试。 **Nvidia Ampere A100 GPU：**在双路 AMD EPYC 7742 CPU 上托管的 Nvidia Ampere A100 GPU，512GB（16 个插槽/ 32GB/ 3200MHz） DDR4 总内存，微代码 0x8301034，超线程启用，睿频启用，Ubuntu 18.04.5 LTS，5.4.0-42-generic 内核，1x SAMSUNG 3.5TB 固态盘，Python 3.7.9，RAPIDS0.17，cuDF 0.17，cuML 0.17，scikit-learn v0.24.1，CUDA 11.0.221，英特尔于 2021 年 2 月 4 日执行测试。Census 数据 [21721922, 45]：数据集来自 IPUMS USA、明尼苏达大学、www.ipums.org [Steven Ruggles，Sarah Flood，Ronald Goeken，Josiah Grover，Erin Meyer，Jose Pacas 和 Matthew Sobek。IPUMS USA：版本 10.0 [数据集]。明尼苏达州明尼阿波里斯市：IPUMS，2020 年。https://doi.org/10.18128/D010.V10.0] 免责声明（仅供定价）：系统价格基于与测试系统相似的系统的平均值，即 2021 年 9 月 20 日 www.colfax-intl.com 和 www.thinkmate.com 上的价格。 用于第三代英特尔® 至强® 可扩展 8380 处理器的 4U 机架安装式系统：Thinkmate GPX XN6-24S3-10GPU 和 Colfax CX41060s-XK8。 用于 AMD EPYC 7742（采用 Nvidia A100 GPU）的 4U 机架安装式服务器： Thinkmate GPX QT24-24E2-8GPU 和 Colfax CX4860s-EK8。 更多详情请参见 www.colfax-intl.com 和 www.thinkmate.com。
AI001，Meena Arunachalam，#21	在 TPC-DS 基准测试，缩短处理 3TB 数据集所需的时间	基准：Azure-US-East-2，Standard_E16s_v3，16 vCPU，10 个实例，铂金 8171M @ 2.60GHz，128GB RAM 内存/实例，256GB 存储/实例（NW 或直连），NW 带宽/实例 32000 (IOPS)/256 (MBps)/400（缓存），8000 存储 BW/实例 新配置：Azure-US-East-2，Standard_E16s_v4，

		16vCPU，10 个实例，铂金 8272CL @ 2.60GHz，128GB RAM 内存/实例，600GB 存储/实例（NW 或直连），NW 带宽/实例 154000 (IOPS)/968 (MBps)/400（缓存），8000 存储 BW/实例；18.04.1-Ubuntu，5.4.0-1046-azure，Databricks 7.5（包括 Apache Spark 3.0.1，Scala 2.12），英特尔于 2021 年 7 月 5 日执行测试 新配置：Azure-US-East-2，Standard_E8s_v4, 8vCPU，20 个实例，铂金 8272CL @ 2.60GHz，256GB RAM 内存/实例，1200GB 存储/实例（NW 或直连），NW 带宽/实例 308000 (IOPS)/1936 (MBps)/800（缓存），16000 存储 BW/实例；18.04.1-Ubuntu，5.4.0-1046-azure，Databricks 7.5（包括 Apache Spark 3.0.1，Scala 2.12），英特尔于 2021 年 7 月 5 日执行测试
AI001，Meena Arunachalam，#21	在 TPC-DS 基准测试，缩短处理 10TB 数据集所需的时间	基准：Azure-US-East-2，Standard_E16s_v3，16 vCPU，10 个实例，铂金 8171M @ 2.60GHz，128GB RAM 内存/实例，256GB 存储/实例（NW 或直连），NW 带宽/实例 32000 (IOPS)/256 (MBps)/400（缓存），8000 存储 BW/实例 新配置：Azure-US-East-2，Standard_E16s_v4，16vCPU，10 个实例，铂金 8272CL @ 2.60GHz，128GB RAM 内存/实例，600GB 存储/实例（NW 或直连），NW 带宽/实例 154000 (IOPS)/968 (MBps)/400（缓存），8000 存储 BW/实例；18.04.1-Ubuntu，5.4.0-1046-azure，Databricks 7.5（包括 Apache Spark 3.0.1，Scala 2.12）。新配置：Azure-US-East-2，Standard_E8s_v4, 8vCPU，20 个实例，铂金 8272CL @ 2.60GHz，256GB RAM 内存/实例，1200GB 存储/实例（NW 或直连），NW 带宽/实例 308000 (IOPS)/1936 (MBps)/800（缓存），16000 存储 BW/实例；18.04.1-Ubuntu，5.4.0-1046-azure，Databricks 7.5（包括 Apache Spark 3.0.1，Scala 2.12），英特尔于 2021 年 7 月 5 日执行测试
AI001，Meena Arunachalam, #22	使用经过英特尔优化的 scikit-learn 加快处理速度。	使用经过英特尔优化的 Scikit-learn 加快处理速度：Azure-US-West，Standard_F16s_V2，16 vCPU，1 个实例，铂金 8168 @ 2.70 GHz / 铂金 8272CL @ 2.60 GHz，32GB 内存容量/实例，直连存储，Ubuntu 18.04.5 LTS，5.4.0-1051-azure，Databricks 9.0 ML 运行时，Stock scikit-learn-0.22.1 与英特尔 scikit-learn-0.24.2，英特尔于 2021 年 9 月 23 日执行测试
AI001，Meena Arunachalam, #22	使用经过英特尔优化的 TensorFlow/BERT-large 加快处理速度。	基准：处理时间加速：经过英特尔优化的 TensorFlow/BERT-large：Azure-US-West，Standard_F32s_V2，32 vCPU，1 个实例，铂金 8168 @ 2.70 GHz / 铂金 8272CL @ 2.60 GHz，64GB 内存容量/实例，直连存储，Ubuntu 18.04.5 LTS，5.4.0-1051-azure，Databricks

		9.0 ML 运行时，Stock TensorFlow 2.3.1 与英特尔 TensorFlow 2.3.0，英特尔于 2021 年 9 月 23 日执行测试 新配置：处理时间加速：经过英特尔优化的 TensorFlow/BERT-large：Azure-US-West，Standard_F64s_V2，64 vCPU，1 个实例，铂金 8168 @ 2.70 GHz / 铂金 8272CL @ 2.60 GHz，128GB 内存容量/实例，直连存储，Ubuntu 18.04.5 LTS，5.4.0-1051-azure，Databricks 9.0 ML 运行时，Stock TensorFlow 2.3.1 与英特尔 TensorFlow 2.3.0，新配置：处理时间加速：经过英特尔优化的 TensorFlow/BERT-large：Azure-US-West，Standard_F72s_V2，72 vCPU，1 个实例，铂金 8168 @ 2.70 GHz / 铂金 8272CL @ 2.60 GHz，144GB 内存容量/实例，直连存储，Ubuntu 18.04.5 LTS，5.4.0-1051-azure，Databricks 9.0 ML 运行时，Stock TensorFlow 2.3.1 与英特尔 TensorFlow 2.3.0，于 2021 年 9 月 23 日执行测试
AI002，Rachel Oberman，#9	1. 相比 stock scikit-Learn，性能提升高达 100 倍 2. PyTorch 优化可将 DLRM 训练（FP32 与 BF16）速度提升高达 1.55 倍，将 DLRM 推理（FP32 与 Int8）速度提升高达 2.8 倍 3. TF 优化和 LPOT 可将量化推理（FP32 与 Int8）提升高达 2.8 倍 4. SKLearn 拟合/预测：英特尔 SKLearnEx 的性能比 Nvidia 和	请参见全部基准测试和配置： https://software.intel.com/content/www/us/en/develop/articles/blazing-fast-python-data-science-ai-performance.html. 每项性能声明和配置数据均源于第 1、2、3、4 和 5 小节中列出的文章正文。另请访问本页面，以获取有关所有分数和测量结果的更多详情。 **测试日期**：性能结果基于 2020 年 10 月 16 日的英特尔测试，可能没有反映所有公开的更新。**配置详细信息和工作负载设置**：2 双路英特尔® 至强® 铂金 8280 @ 28 核，操作系统：Ubuntu 19.10.5.3.0-64-generic Mitigated 384GB RAM (192 GB RAM (12x 32GB 2933)。软件：Modin 0.81。Scikit-learn 0.22.2。Pandas 1.01，Python 3.8.5，DAL(DAAL4Py) 2020.2，Census Data，(21721922.45) 数据集来自 IPUMS USA、明尼苏达大学、www.ipums.org [Steven Ruggles，Sarah Flood，Ronald Goeken，Josiah Grover，Erin Meyer，Jose Pacas 和 Matthew Sobek。IPUMS USA：版本 10.0 [数据集]，明尼苏达州明尼阿波里斯市。IPUMS，2020. https//doc.org/10.18128/D010.V10.0] **测试日期**：性能结果基于 2020 年 10 月 23 日的英特尔® 测试，可能没有反映所有公开的更新。**配置详细信息和工作负载设置**：英特尔® oneAPI 数据分析库 2021.1 (oneDAL)。Scikit-learn 0.23.1，英特尔® Python 分发版 3.8；英特尔® 至强® 铂金 Platinum 8280LCPU @ 270GHz，双插槽，每插槽 28 核，10M 样本，10 项特性，100 个集群，100 次迭代，float32.。 **测试日期**：性能结果基于 2020 年 10 月 23 日的英特尔® 测试，可能没有反映所有公开的更新。**配置详细信息和工作负**

	AMD CPU 分别高 10 倍与 5 倍 5. 相比采用 XGBoost 推理优化的 Nvidia GPU，推理速度提升高达 4.5 倍 6. 在 Census 2020 上，使用 Modin 的 ETL 速度提升了 38 倍，SKlearn 中使用 ML 选项进行拟合与预测的速度提升了 21 倍 英特尔优化 Modin 和 SKLearnEx 可将 census 工作负载的整体性能提升 40%	**载设置**：英特尔® oneAPI AI 分析工具套件 v2021.1；英特尔® oneAPI 数据分析库 (oneDAL) beta10，Scikit-learn 0.23.1，英特尔® Python 分发版 3.7，英特尔® 至强® 铂金 8280 CPU @ 2.70GHz，双插槽，每插槽 28 颗内核，微代码：0x4003003，376 GB 总可用内存，12X32GB 模块，DDR4。**AMD 配置**：AMD Rome 7742 @2.25 GHz，双插槽，每插槽 64 颗内核，微代码：0x8301038，512 GB 总可用内存，16X32GB 模块，DDR4，oneDAL beta10，Scikit-learn 0.23.1，英特尔® Python 分发版 3.7。**NVIDIA 配置**：NVIDIA Tesla V100 - 16 Gb，376 GB 总可用内存，12X32GB 模块，DDR4，英特尔® 至强 ®铂金 8280 CPU @ 2.70GHz，双插槽，每插槽 28 颗内核，微代码：0x5003003，cuDF 0.15，cuML 0.15，CUDA 10.2.89，驱动程序 440.33.01，操作系统：CentOS Linux 7（内核），Linux 4.19.36 内核。 **测试日期：**性能结果基于 2020 年 10 月 13 日的英特尔® 测试，可能没有反映所有公开的更新。**配置详细信息和工作负载设置**： CPU：c5.18xlarge AWS 实例（2 x 英特尔® 至强® 铂金 8124M @ 18 个内核。操作系统：Ubuntu 20.04.2 LTS，193 GB RAM。GPU：p3.2xlarge AWS 实例（GPU：NVIDIA Tesla V100 16GB，8 vCPU，操作系统：Ubuntu 18.04.2LTS，61 GB RAM。软件：XGBoost 1.1：从源代码构建，编译器 - G++ 7.4，nvcc 9.1 英特尔® DAAL：2019.4 版本：Python 环境：Python 3.6，Numpy 1.16.4，Pandas 0.25 Scikit-learn 0.21.2。 **测试日期**：性能结果基于 2020 年 10 月 26 日的英特尔测试，可能没有反映所有公开的更新。**配置详细信息和工作负载设置**：面向 Tensorflow v2.2.0 的英特尔® 优化；oneDNN v1.2.0；Intel® 低精度优化工具 v1.0；平台；英特尔® 至强® 铂金 8280 CPU；节点 1；插槽：2；内核/插槽：28；线程/插槽：56；超线程：启用，睿频：启用；BIOS 版本：SE5C620.86B.02.01.0010.010620200716；系统 DDR 内存配置：12 插槽/16GB/2933；操作系统：CentOS Linux 7.8；内核：4.4.240-1.el7.elrepo x86_64。 **测试日期**：性能结果基于 2020 年 2 月 3 日的英特尔测试，可能没有反映所有公开的更新。**配置详细信息和工作负载设置：**面向 PyTorch v1.5.0 的英特尔® 优化；面向 PyTorch (IPEX) 1.1.0 的英特尔® 扩展；oneDNN 版本：v1.5；DLRM：训练批次大小 (FP32/BF16)：2K/实例，1 个实例；DLRM 数据集 (FP32/BF16)：Criteo Terabyte 数据集；BERT-Large：训练批次大小 (FP32/BF16)：24/实例。单个 CPU 插槽上 1 个实例。数据集 (FP32/BF16)：WikiText-2 [https://www.salesforce.com/products/einstein/ai-

		research/the-wiktext-dependency-language-modeling-dataset/]: ResNext101-32x4d：训练批次大小(FP32/BF16)：128/实例，单个 CPU 插槽上 1 个实例，数据集 (FP32/BF16)：ILSVRC2012；DLRM：推理批次大小(INT8)：16/实例，28 个实例，空数据。英特尔® 至强® 铂金 8380H 处理器，4 插槽，28 核，超线程启用，睿频启用，总内存 768 GB（24 插槽/32GB/3200 MHz），BIOS；WLYDCRBLSYS.0015.P96.2005070242（ucode:：OX 700001b0，Ubuntu 20.04 LTS，内核 5.4.0-29-genen：ResNet50：[https://github.com/Intel/optimized-models/tree/master/pytorch/ResNet50]：ResNext101 32x4d：[https://github.com/intel/optimized-models/tree/master/pytorch/ResNext101_32x4ct：DLRM：https//github.com/intel/optimized-models/tree/master/pytorch/dlrm].
AI002，Rachel Oberman，演示（部分演讲）	1. Modin 交互始终小于 1 秒，大数据也不例外 2. Xgboost 优化前为 2 小时，优化后为 11 分钟 3. Sklearn 优化前为 50 秒，优化后为 7 秒 4. 展示与 RAPIDS 的交互需要更改代码以及解决未解决的错误(如定性，而非性能比较）	**测试日期：**性能结果基于 2021 年 10 月 4 日的英特尔测试，可能没有反映所有公开的更新。**配置详细信息和工作负载设置：硬件（所有配置均相同）：**1 节点，双路第二代英特尔® 至强® 金牌 6258R，联想 30BC003DUS，768GB（12 插槽/ 64GB/ 2666）DDR4 总内存和 2TB（4 插槽/ 512GB/ 2666）DCPMM 内存，微代码 0x5003102，超线程启用，睿频启用，Ubuntu 20.04.3 LTS，5.10.0-1049-oem，1x Samsung 1TB 固态盘操作系统硬盘，RAID0 数据硬盘中的 4x Samsung 2TB 固态盘，3x NVIDIA Quadro RTX 8000。**3 个月的 NYCTaxi 数据，Stock 软件配置：**Python 3.9.7，Pandas 1.3.3，Scikit-Learn 1.0，XGBoost 0.81，IPython 7.28.0，IPKernel 6.4.1.**满 30 个月的 NYCTaxi 数据，Nvidia RAPIDS 软件配置：**Python 3.7.10，Pandas 1.2.5，XGBoost 1.4.2，cuDF 21.08.03，cudatoolkit 11.2.72，dask-cudf 21.08.03，dask-cuda 21.08.00，IPython 7.28.0，IPKernel 6.4.1。**满 30 个月的 NYCTaxi 数据，经过英特尔优化的软件配置：**Python 3.9.7，Pandas 1.3.3，Modin 0.11.0，OmniSci 5.7.0，Scikit-:earn 1.0，英特尔® Extension for Scikit-Learn* 2021.3.0，XGBoost 1.4.2，IPython 7.28.0，IPKernel 6.4.1。来自纽约市的 NYCTaxi 数据集 (nyc.gov)：[https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page]
AI003 AG Ramesh Andres Rodriguez #18	从第二代英特尔至强（代号为 Icelake）到第三代英特尔至强（代号为 Sapphire Rapids），SSD-RN34 工作负载的总	基准：1 节点，双路第三代英特尔至强铂金 8380，512 GB（16 插槽/ 32GB/ 3200）DDR4 总内存，微代码 0x8d9522d4，超线程启用，睿频启用，Ubuntu 20.04.2 LTS(docker)，5.4.0-77-generic，TensorFlow v2.5.0（无

	体性能提升高达 30 倍。	oneDNN），TensorFlow v2.6.0（带 oneDNN），英特尔于 2021 年 9 月 28 日执行测试 新配置：1 节点，英特尔预生产平台上的双路下一代英特尔至强可扩展处理器（代号为 Sapphire Rapids, > 40 内核），512 GB DDR 内存（8(1DPC)/32GB/4800 MT/秒），超线程启用，睿频启用，CentOS Linux 8.4，内部预生产 bios 和软件运行 SSD-ResNet34 BS=1 使用 TensorFlow 2.6，具有英特尔内部优化，英特尔于 2021 年 9 月 28 日执行测试
AI004，Susan Lansing，第 9 页幻灯片	相比新的 GPU 实例，性价比提升高达 40%	发布时，AWS EC2 DL1 的价格请参见：https://aws.amazon.com/ec2/pricing/ondemand/ 性价比声明是 AWS 根据 AWS 内部性能测试提出的， 基于 di 的实例、基于 Nvidia A100 和 V100 的实例的性能与各个实例的 AWS EC2 价格之比即为性价比。净值：表示客户可以从成本中获得多少训练性能。 Gaudi 性能指标的计算方法为：**AI 处理器：**对于 DL1 实例，AWS 定制服务器上的 Habana Gaudi HL-205 处理器 1 卡和 8 卡配置，**CPU：**AWS 定制第二代英特尔® 至强® 可扩展处理器 **计算机视觉指标基于 ResNet-50 模型：**https://github.com/HabanaAI/Model-References/tree/master/TensorFlow/computer_vision/Resnets/resnet_keras 和 Habana 测试容器：http://vault.habana.ai/gaudi-docker/0.15.0/ubuntu18.04/habanalabs/tensorflow-installer-tf-cpu-2.5.1:1.0.1-81 A100 / V100 性能基于软件构建：https://ngc.nvidia.com/catalog/resources/nvidia:resnet_50_v1_5_for_tensorflow/performance 结果发布在 DGXA100 和 DGX-1 上； **批次大小：**所有加速器均为 256。 **自然语言处理指标基于 BERT-Large，预训练，第一阶段，** DL1 性能基于 Habana BERT 模型：https://github.com/HabanaAI/Model-References/tree/master/TensorFlow/nlp/bert Habana 测试容器：https://vault.habana.ai/artifactory/gaudi-docker/1.0.1/ubuntu18.04/habanalabs/tensorflow-installer-tf-cpu-2.5.1/1.0.1-81

		A100 / V100 基准测试来源*: https://ngc.nvidia.com/catalog/resources/nvidia:bert_for_tensorflow/performance * 结果发布在 DGXA100 和 DGX-1 上 **批次大小：**A100 和 Gaudi 为 64，V100 为 16。
AI004，Susan Lansing，第 11 页幻灯片	400 Gbps 网络：为所有 EC2 实例提供极快的网络带宽，以支持具有高吞吐量节点间连接的大型集群	AWS 声称，在 EC2 DL1 上为客户提供的网络速度是所有 EC2 实例中较快的。 该信息及价格公开发布在该链接中: https://aws.amazon.com/ec2/pricing/ondemand/
AI004，Susan Lansing，第 11 页幻灯片	面向 Lustre 的 Amazon FSx：高性能、可扩展存储提供亚毫秒级时延和数百 GB/秒的吞吐量	AWS 有关其 EC2 服务、面向 Lustre 的 FSx 的存储速度和吞吐量声明 如欲了解有关服务的更多信息，请访问 https://aws.amazon.com/fsx/lustre/?nc2=type_
AI004，Susan Lansing，第 13 页幻灯片	图表包含 EC2 实例的计算机视觉/ResNet-50 性能指标：基于 Gaudi 的 DL1，A100/P4d 和 V100/p3，面向 1 卡和 8 卡配置。	模型：ResNet-50 混合 精度： BF16/FP16 框架：TensorFlow 配置：所有测试配置均为 1 卡和 8 卡 批次大小： 256 基于 Habana 的 DL1 指标和基于 ResNet50 模型的 AWS 内部性能测试：Habana 软件: https://github.com/HabanaAI/Model-References/tree/master/TensorFlow/computer_vision/Resnets/resnet_keras Habana 测试容器：http://vault.habana.ai/gaudi-docker/0.15.0/ubuntu18.04/habanalabs/tensorflow-installer-tf-cpu-2.5.1:1.0.1-81 A100/p4d 和 V100/p3 性能 来源*: https://ngc.nvidia.com/catalog/resources/nvidia:resnet_50_v1_5_for_tensorflow/performance 结果发布在 DGXA100 和 DGX-1 上
AI004，Susan Lansing，第 14 页幻灯片	图表包含 EC2 实例的 NLP 性能指标：基于 Gaudi 的	模型：BERT-Large；预训练（第一阶段） 混合精度： B16/FP16

	DL1、A100/P4d 和 V100/p3，面向 1 卡和 8 卡配置。	批次大小：A100 和 Gaudi 为 64；V100 为 16。 基于 Habana 的 DL1 指标和基于 BERT 模型的 AWS 性能测试：https://github.com/HabanaAI/Model-References/tree/master/TensorFlow/nlp/bert Habana 测试容器：https://vault.habana.ai/artifactory/gaudi-docker/1.0.1/ubuntu18.04/habanalabs/tensorflow-installer-tf-cpu-2.5.1/1.0.1-81 A100 / V100 基准测试来源*：https://ngc.nvidia.com/catalog/resources/nvidia:bert_for_tensorflow/performance 结果发布在 DGXA100 和 DGX-1 上
AI004，Susan Lansing，第 15 页幻灯片	图表显示 ResNet-50 和 BERT 模型上影响 EC2 实例价格与性能的多重因素，以展示性价比（实现的训练量与特定时间内的实例价格之比）。	Habana 和 AWS 性能测试基于： Habana BERT 模型：https://github.com/HabanaAI/Model-References/tree/master/TensorFlow/nlp/bert Habana ResNet50 模型：https://github.com/HabanaAI/Model-References/tree/master/TensorFlow/computer_vision/Resnets/resnet_keras Habana 测试容器：https://vault.habana.ai/artifactory/gaudi-docker/1.0.1/ubuntu18.04/habanalabs/tensorflow-installer-tf-cpu-2.5.1/1.0.1-81 A100 / V100 基准测试来源*： https://ngc.nvidia.com/catalog/resources/nvidia:bert_for_tensorflow/performance 和 https://ngc.nvidia.com/catalog/resources/nvidia:resnet_50_v1_5_for_tensorflow/performance * 结果发布在 DGXA100 和 DGX-1 上。价格发布在 https://aws.amazon.com/ec2/pricing/on-demand/ 上
AI004，Susan Lansing，第 18 页幻灯片	图表显示性价比指标背后的详细信息，相比在基于 Nvidia 的实例上执行 EC2 训练，AWS 客户在 Gaudi/EC2 DL1 上训练 ResNet-50 可获得更高的性价比。	评估方法： **模型：** ResNet-50 **框架：** TensorFlow **混合精度配置：** 所有实例均为 8 卡 6 月 28 日，Habana 使用 Nvidia 深度学习 AMI (Ubuntu 18.04) + Docker 21.06-tf1-py3（可通过

	比较：-ResNet-50 性能 -每种实例类型每小时按需定价 - 每美元训练数百万张图像 -以 A100 40 GB 的性价比为基准 -与 A100 40 GB 的性价比进行比较	https://ngc.nvidia.com/catalog/containers/nvidia.tensorflow/tags 获得），在基于 AWS EC2 GPU 的实例上测量基于 A100 和 V100 的实例， 模型训练：https://github.com/NVIDIA/DeepLearningExamples/tree/master/TensorFlow/Classification/ConvNets/resnet50v1.5 Habana 在 AWS EC2 DL1.24xlarge 实例上使用位于 Habana Vault 的 DLAMI 集成 SynapseAI 1.0.1-81 Tensorflow 2.5.1 容器测量 Gaudi，模型：https://github.com/HabanaAI/Model-References/tree/master/TensorFlow/computer_vision/Resnets/resnet_keras AWS 价格发布在 https://aws.amazon.com/ec2/pricing/on-demand/ 上。结果可能不同。
AI004，Susan Lansing，第 19 页幻灯片	图表显示性价比指标背后的详细信息，相比在基于 Nvidia 的实例上执行 EC2 训练，AWS 客户在 Gaudi/EC2 DL1 上训练 BERT 可获得更高的性价比。比较：-ResNet-50 性能 -每种实例类型每小时按需定价 - 每美元训练数百万张图像 -以 A100 40 GB 的性价比为基准 -与 A100 40 GB 的性价比进行比较	**评估方法：** **模型：**BERT-Large，预训练，第一阶段 **框架：**TensorFlow **混合精度配置：**所有实例均为 8 卡 6 月 28 日，Habana 使用 Nvidia 深度学习 AMI (Ubuntu 18.04) + Docker 21.06-tf1-py3（可通过 https://ngc.nvidia.com/catalog/containers/nvidia.tensorflow/tags 获得），在基于 AWS EC2 GPU 的实例上测量 A100 和 V100 实例 模型：https://github.com/NVIDIA/DeepLearningExamples/tree/master/TensorFlow/LanguageModeling/BERT Habana 在 AWS EC2 DL1.24xlarge 实例上使用位于 Habana Vault 的 DLAMI 集成 SynapseAI 1.0.1-81 Tensorflow 2.5.1 容器测量 Gaudi 实例，模型：https://github.com/HabanaAI/Model-References/tree/master/TensorFlow/ nlp/bert。EC2 实例的价格发布在 https://aws.amazon.com/ec2/pricing/on-demand/ 上。结果可能不同。
AI004，Susan Lansing，第 20 页幻灯片	MLPerf 性能基于相应厂商提交的指标。MLPerf 指标之间的比较针对 Nvidia A100，在 MxNet 和 Pytorch（而非	MLPerf 训练时间指标发布在 https://mlcommons.org/en/training-normal-10/ 上 Habana 在 Nvidia GPU 上使用基于公有软件的 TensorFlow 框架测量性

	TensorFlow）上报告。 下图展示 ResNet 和 BERT 的性能，基于“TensorFlow 上公开可用的软件（NGC 和 Habana 容器）的训练吞吐量。” 然后将测量结果与 EC2 实例的客户定价进行比较。	能 https://ngc.nvidia.com/catalog/resources/nvidia:bert_for_tensorflow/performance； Habana 在 AWS EC2 DL1.24xlarge 实例上使用位于 Habana Vault 的 DLAMI 集成 SynapseAI 1.0.1-81 Tensorflow 2.5.1 容器测量 DL1 性能。 **模型：** https://github.com/HabanaAI/Model-References/tree/master/TensorFlow/nlp/bert。EC2 实例价格发布在此处：https://aws.amazon.com/ec2/pricing/on-demand/ 测量的性能可能有所差异。
CLD005，Arijit Biswas，#17 CLD001， Giri，#19 CLDTI002，Kamhout/Weekly，#25	在微服务性能方面，提升了每核吞吐量（p99 的时延 SLA 低于 30 毫秒）： 第三代至强比第二代至强高 24% 下一代至强 (Sapphire Rapids) 比第二代至强高 69%	工作负载：DeathStarBench 'hotelReservation', 'socialNetwork' (https://github.com/delimitrou/DeathStarBench）和 Google 微服务演示 (https://github.com/GoogleCloudPlatform/microservices-demo) 操作系统：Ubuntu 20.04，内核版本 v5.10，Kubernetes v1.21.0；测试截至 2021 年 7 月。 在 3 节点 Kubernetes 设置上测量 Cascade Lake，在 us-west2b 中部署 AWS M5.metal 实例（2S 24 核 8259CL，384GB DDR4 RAM 和 25Gbps 网络） 在 3 节点 2S 32 核、2.5GHz, 300W TDP SKU（具有 512GB DDR4 RAM 和 40Gbps 网络）上测量 Ice Lake
CLD001，Giri，#16	通过来自 Pathlength 和 CPU 停滞减少优化在 Wordpress v4.2 中实现高达 83% 的性能提升	在 Wordpress 4.2 HTTPS 中比较了经过优化的第三代英特尔至强可扩展平台 (ICX) 以及第二代英特尔至强可扩展平台 (CLX)。 CLX 基准：1 节点，S2600WFT 上的双路英特尔至强金牌 6238R（28 核），384GB（12 插槽/ 32 GB / 2933）DDR4 总内存，微代码 0x5003003，超线程启用，睿频启用，Ubuntu 20.04，Linux 5.4.0-65-generic，1x 英特尔 1.8T SSDSC2KG01，wordpress 4.2.0，PHP 7.4，gcc (Ubuntu 9.3.0-17ubuntu1~20.04) 9.3.0，GNU C 库 (Ubuntu GLIBC 2.31-0ubuntu9.1)，mysqldVer: 10.3.25-MariaDB-0ubuntu0.20.04.1，1x 英特尔 X722，TLSv1.3 - TLS_AES_256_GCM_SHA384，英特尔于 2021 年 5 月 19 日执行测试。 ICX：1 节点，Coyote Pass 上的双路英特尔至强金牌 6348（28 核），512GB（16 插槽/ 32 GB / 3200）DDR4 总内

		存，微代码 0xd000270，超线程启用，睿频启用，Ubuntu 20.04，Linux 5.4.0-72-generic，1x 英特尔 895GB SSDSC2KG96，wordpress 4.2.0，PHP 7.4，gcc (Ubuntu 9.3.0-17ubuntu1~20.04) 9.3.0，GNU C 库 (Ubuntu GLIBC 2.31-0ubuntu9.1)，mysqldVer: 10.3.25-MariaDB-0ubuntu0.20.04.1，1x XL710-Q2，TLSv1.3 - TLS_AES_256_GCM_SHA384，英特尔于 2021 年 5 月 19 日执行测试。
CLD005，Arijit Biswas，#18	我们引入了 AMX 功能，该功能可显著加速深度学习算法的核心 - 张量处理。借助 AMX，我们可以每周期执行 2048 次 int8 运算（如果不采用 AMX，仅为 256 次）以及 1024 次 bfloat16 运算（如果不采用 AMX，仅为 64 次）。	基于每核每周期矩阵乘法 + 累加的峰值架构功能，假设 CPU 利用率为 100%。截至 2021 年 8 月的测试。
CLD011，第 9 页幻灯片，Pranav KalavadeRobbie Frickie	每单元比特技术进一步降低了 QLC 的成本	使用 FG 和 CTF 技术在 QLC 固态盘的组件上进行测量。所使用的测量平台为 Teradyne Magnum 2 Memory 测试系统，使用客户命令对使用随机模式和裕度的编程进行了量化。2019 年 8 月测量的数据。结果可能会有所不同。
CLD011，第 7 页幻灯片 Pranav Kalavade，Robbie Frickie	浮栅技术提供了领先的面密度。	资料来源：ISSCC 2015，J.Im；ISCC 2017 R Yamashita；ISSCC 2017 C Kim；ISCC 2018；H. Maejima；ISSCC 2019 C. Siau
CLD011，第 12 页幻灯片，Pranav Kalavade, Robbie Frickie	相比 TLC 固态盘和大容量机械硬盘，英特尔固态盘 D5-P5316 可提供更卓越的生命周期写入性能。	**大多数固态盘无法达到 DWPD 评级**：UoT 对部署于企业存储中的 140 万行业固态盘展开调查研究
CLD011，第 13 页幻灯片，Pranav Kalavade，Robbie Frickie	过去，10TB 技术执行 1 DWPD，而全新的 20TB 技术写入的 PB 数相当于	**理论 TBW 比较**：10TB 硬盘 PBW = 10TB x 1DWPD x 5 年 x 365 天/年= 18.25 写入 PB。理论 20TB 硬盘 PBW = 20TB x 0.5DWPD x 5 年 x 365 天/年= 18.25 写入 PB。

	5 DWPD	
CLD011，第 13 页幻灯片，Pranav Kalavade，Robbie Frickie	固态盘比 HDD 更耐用。HDD 会随着时间推移不断磨损。	**HDD 耐用性。**20TB WD HC650 和 18TB Seagate EXOS X18 的耐用性。
CLD011，第 13 页幻灯片，Pranav Kalavade，Robbie Frickie	QLC 固态盘具有非常高的耐用性。QLC 技术不到 1DWPD，足以满足市场需求。	**QLC 固态盘耐用性。**英特尔® 固态盘 D5-P5316 规格。
CLI006，Ajith Illendula，课程与展示演示 - **使用英特尔® 深度学习加速在第 12 代智能英特尔® 酷睿™ 处理器上对图像降噪**	相比 Int-8，使用 FP32 时 ADL 上的降噪时间缩短了高达 1.28 倍	**测试日期：**性能结果基于 2021 年 9 月 23 日的英特尔测试，可能没有反映所有公开的更新。 **平台：**Alderlake 台式机 **处理器：**英特尔® 酷睿™ i9-12900K 处理器 (16C/24T) **内存：** 4x8GB DDR5 4800 MHz SDRAM **存储：**三星固态盘 980 Pro 1TB **显示屏分辨率：**1920x1080 显示屏 **操作系统：**Microsoft Windows 11，Build 10.0.22000.132 **显卡：** NVIDIA GeForce RTX™ 3080
CLI006，Ajith Illendula，#20	通过使用 OpenVINO 基准测试工具，Int-8 的速度比 FP32 高 1.38 倍	**测试日期：**性能结果基于 2021 年 9 月 23 日的英特尔测试，可能没有反映所有公开的更新。 **平台：**Alderlake 台式机 **处理器：**英特尔® 酷睿™ i9-12900K 处理器 (16C/24T) **内存：**4x8GB DDR5 4800 MHz SDRAM **存储：**三星固态盘 980 Pro 1TB **显示屏分辨率：**1920x1080 显示屏 **操作系统：**Microsoft Windows 11，Build 10.0.22000.132 **显卡：**NVIDIA GeForce RTX™ 3080
CLI006，Ajith Illendula，#12	相比 Tensor Flow CPU 实施，OpenVINO (DeNoise 1.3) 可将性能提升高达 4.5	**测试日期：**性能结果基于 2021 年 9 月 23 日的英特尔测试，可能没有反映所有公开的更新。

	倍	**平台：**Alderlake 台式机**处理器：**英特尔® 酷睿™ i9-12900K 处理器 (16C/24T) **内存**：4x8GB DDR5 4800 MHz SDRAM **存储：**三星固态盘 980 Pro 1TB **显示屏分辨率：**1920x1080 显示屏 **操作系统：**Microsoft Windows 11，Build 10.0.22000.132 **显卡：**NVIDIA GeForce RTX™ 3080
CLI006，Ajith Illendula，#22	Int8 输出质量类似于 FP32 输出质量	**测试日期：**性能结果基于 2021 年 9 月 23 日的英特尔测试，可能没有反映所有公开的更新。 **输出质量：0.19 LPIPS 指标，比较 Int8 输出与 FP32 输出。平台：**Alderlake 台式机 **处理器：**英特尔® 酷睿™ i9-12900K 处理器 (16C/24T) **内存**：4x8GB DDR5 4800 MHz SDRAM **存储：**三星固态盘 980 Pro 1TB **显示屏分辨率：**1920x1080 显示屏 **操作系统：**Microsoft Windows 11，Build 10.0.22000.132 **显卡：**NVIDIA GeForce RTX™ 3080
CLI011，Erik Niemeyer，课程与展示演示 - **英特尔和 Autodesk 加速光线追踪工作流程**	使用高质量 Maya 场景或 +4AA 比较渲染时间，12900K 的性能比 11900K 提升高达 1.9 倍。	**平台：**ADL 白盒 **处理器：**英特尔® 酷睿™ i9-12900K 处理器 (16C/24T) **内存：** 2x16GB DDR5 4800 MHz SDRAM **存储：** Samsung SSD 980 PRO (1 TB) **显示屏分辨率：**1920 x 1080 显示屏 **操作系统：**Microsoft Windows 11 专业版，Build 22000.120 **显卡：**NVIDIA GeForce RTX™ 3090 **平台：**RKL 白盒 **处理器：**英特尔® 酷睿™ i9-11900K 处理器 (8C/16T) **内存：**2x16GB DDR4 3200 MHz SDRAM **存储：** WDS400T3X0C-00SJG0 (4 TB) **显示屏分辨率：**1920x1080 显示屏

		操作系统： Microsoft Windows 11 专业版，Build 22000.120 **显卡：** NVIDIA GeForce RTX™ 3090 性能结果基于 2020 年 9 月 23 日的英特尔测试，可能没有反映所有公开的更新。
CLI011，Erik Niemeyer，课程与展示演示 - **英特尔和 Autodesk 加速光线追踪工作流程**	在 IPR 场景中，第 12 代酷睿的性能比前代提升两倍（以上），11900K = 11 秒，12900K = 4 秒	**平台：** ADL 白盒 **处理器：** 英特尔® 酷睿™ i9-12900K 处理器 (16C/24T) **内存：** 2x16GB DDR5 4800 MHz SDRAM **存储：** 三星固态盘 980 PRO (1 TB) **显示屏分辨率：** 1920 x 1080 显示屏 **操作系统：** Microsoft Windows 11 专业版，Build 22000.120 **显卡：** NVIDIA GeForce RTX™ 3090 **平台：** RKL 白盒**处理器：** 英特尔® 酷睿™ i9-11900K 处理器 (8C/16T) **内存：** 2x16GB DDR4 3200 MHz SDRAM **存储：** WDS400T3X0C-00SJG0 (4 TB) **显示屏分辨率：** 1920x1080 显示屏 **操作系统：** Microsoft Windows 11 专业版，Build 22000.120 **显卡：** NVIDIA GeForce RTX™ 3090 性能结果基于 2020 年 9 月 23 日的英特尔测试，可能没有反映所有公开的更新。
E5G005，Deepthi Karkada，Ryan Loney，GE,	第 11 代智能英特尔® 酷睿联手 OpenVINO，可提升可扩展性能（第 36 页幻灯片）	<table><tr><td>CPU</td><td>第 11 代智能英特尔® 酷睿™ i7-1185G7E @ 2.80GHz（**在图表中被称为 TGL i7-1185G7E**）</td><td>第 8 代智能英特尔® 酷睿™ i7-8700T @ 2.40GHz（**在图表中被称为 CFL i7-8700T**）</td></tr><tr><td>主板</td><td>英特尔参考验证平台</td><td>英特尔参考验证平台</td></tr><tr><td>超线程</td><td>启用</td><td>启用</td></tr></table>

睿频设置	启用	启用
内存	2 x 8 GB DDR4 3200MHz	16 GB DDR4 2666 MT/秒
BIOS 厂商	英特尔公司	美国安迈科技股份有限公司
BIOS 版本	TGLSFWI1.R00.3373.A00.2009091720	IG1b IA
BIOS 版本	09/09/2020	09/17/2019
微代码	0x88	0xea
GPU	锐炬 Xᵉ 96EU	英特尔® 超核芯显卡 630
批次大小	1	1
精度	FP32	FP32
并发推理请求数量	CPU & GPU 插件延迟和吞吐量结果为 1，多个插件的结果：CPU 为 4 个数据流，GPU 为 2 个数据流	CPU & GPU 插件延迟和吞吐量结果为 1，多个插件的结果：CPU 为 4 个数据流，GPU 为 2 个数据流，TF 结果为为 1 个数据流
OS 名称	Ubuntu 18.04.3 LTS	Ubuntu 18.04.3 LTS
操作系统内核	Linux 5.9.0-050900-generic	Linux 5.9.0-050900-generic

软件	OV-2021.4.1	OV-2021.4.1, Stock TF 2.0.0, 不包含 oneDNN 优化, TensorFlow 不支持在 iGPU 上执行推理 [1]
测试日期	9/15/2021	9/21/2021
功耗, TDP（瓦）	28	35
成本	431	303
性能功耗比	OV 的吞吐量/TDP（瓦） = 7.63/28 = 0.2725	OV 的吞吐量/TDP（瓦） = 4.02/35 = 0.114
性价比	OV 的吞吐量/成本 = 7.63/431 = 0.0177	OV 的吞吐量/成本 = 4.02/303 = 0.0132

IOF005 Randi Rost 第 14 页	• 2-20X 加速深度学习预测 • 节省 30-90% 的深度学习推理成本	模型： Squeezenet 1.1 操作系统：Ubuntu 20.04 LTS；Linux 版本： Linux ip-172-31-30-130 5.11.0-1019-aws #20~20.04.1-Ubuntu SMP Tue Sep 21 10:40:39 UTC 2021 硬件平台：AWS c5.12xlarge；英特尔(R) 至强(R) 铂金 8275CL CPU @ 3.00GHz；内核数=24，启用的内核数=24，线程数=48 内存：96GB DIMM DDR4 静态列伪静态同步窗口 DRAM 2933 MHz（0.3 纳秒） 基准测试：Pytorch v1.7.1 标准版；测试日期： 2021-10-22T23:23:14Z 基准测试结果：46.9 微秒 优化测试：Inteon 平台 v.JF0.5；测试日期： 2021-10-05T03:28:12Z

		优化结果：1.99 微秒 基准测试框架： MLperf 报告的结果：90% 批次大小：1 基准/优化= 46.9/1.99 = 速度提升 23.6 倍 AWS 计算时间按小时购买，因此预计 c5.12xlarge 实例将节省 = 1 – 1.99/46.9 = 95.8% 的成本
IOF005 Randi Rost 第 33 页幻灯片	使用 Inteon 优化时，Resnet 50 v2.7 上的推理性能提升高达 7.76 倍	模型： Resnet 50 v2.7 操作系统：Ubuntu 20.04 LTS；Linux 版本：Linux ip-172-31-30-130 5.11.0-1019-aws #20~20.04.1-Ubuntu SMP Tue Sep 21 10:40:39 UTC 2021 x86_64 x86_64 x86_64 GNU/Linux；Linux 内核详细信息：Linux 版本 5.11.0-1019-aws (buildd@lgw01-amd64-037) (gcc (Ubuntu 9.3.0-17ubuntu1~20.04) 9.3.0, GNU ld (GNU Binutils for Ubuntu) 2.34) #20~20.04.1-Ubuntu SMP Tue Sep 21 10:40:39 UTC 2021 硬件平台：AWS c5.12xlarge；英特尔(R) 至强(R) 铂金 8275CL CPU @ 3.00GHz；内核数=24，启用的内核数=24，线程数=48 内存：96GB DIMM DDR4 静态列伪静态同步窗口 DRAM 2933 MHz（0.3 纳秒） 基准测试：Pytorch v1.8.1 标准版；测试日期：2020-10-02T13:32:54Z 优化测试：Inteon 平台 v.JF0.5；测试日期：2021-09-30T15:05:18Z 基准测试框架：MLperf 报告的结果：90%，批次大小：1
IOF005 Randi Rost 第 34 页幻灯片	使用 Inteon 优化时，VGG16 上的推理性能提升高达 4.22 倍（第 34 页幻灯片）	模型：VGG16 操作系统：Ubuntu 20.04 LTS；Linux 版本：Linux ip-172-31-30-130 5.11.0-1019-aws #20~20.04.1-Ubuntu SMP Tue Sep 21 10:40:39 UTC 2021 x86_64 x86_64 x86_64 GNU/Linux；Linux 内核详细信息：Linux 版本 5.11.0-1019-aws (buildd@lgw01-amd64-037) (gcc (Ubuntu 9.3.0-17ubuntu1~20.04) 9.3.0, GNU ld (GNU Binutils for Ubuntu) 2.34) #20~20.04.1-Ubuntu SMP Tue Sep 21 10:40:39 UTC 2021 硬件平台：AWS c5.12xlarge；英特尔(R) 至强(R) 铂金 8275CL CPU @ 3.00GHz；内核数=24，启用的内核数=24，线程数=48 内存：96GB DIMM DDR4 静态列伪静态同步窗口 DRAM 2933 MHz（0.3 纳秒） 基准测试：TensorFlow 2.4.1；测试日期：2021-04-12T18:57:16Z 优化测试：Inteon 平台 v.JF0.5；测试日

		期：2021-09-30T15:02:45Z 基准测试框架：MLperf 报告的结果：90%，批次大小：1
IOF005 Randi Rost 第 35 页幻灯片	使用 Inteon 优化时，Squeezenet 1.1 上的推理性能提升高达 9.25 倍（第 35 页幻灯片）	模型：Squeezenet 1.1 操作系统：Ubuntu 20.04 LTS；Linux 版本：Linux ip-172-31-30-130 5.11.0-1019-aws #20~20.04.1-Ubuntu SMP Tue Sep 21 10:40:39 UTC 2021 x86_64 x86_64 x86_64 GNU/Linux；Linux 内核详细信息：Linux 版本 5.11.0-1019-aws (buildd@lgw01-amd64-037) (gcc (Ubuntu 9.3.0-17ubuntu1~20.04) 9.3.0, GNU ld (GNU Binutils for Ubuntu) 2.34) #20~20.04.1-Ubuntu SMP Tue Sep 21 10:40:39 UTC 2021 硬件平台：AWS c5.12xlarge；英特尔(R) 至强(R) 铂金 8275CL CPU @ 3.00GHz；内核数=24，启用的内核数=24，线程数=48 内存：96GB DIMM DDR4 静态列伪静态同步窗口 DRAM 2933 MHz（0.3 纳秒） 基准测试：Pytorch v1.8.1 标准版；测试日期：2020-09-29T17:30:36Z 优化测试：Inteon 平台 v.JF0.5；测试日期：2021-10-05T03:28:12Z 基准测试框架：MLperf 报告的结果：90%，批次大小：1