

基于英特尔®CPU MAX处理器加速 AI4Science

姓名 邓会鹏

职务 英特尔AI解决方案架构师

议程

基于英特尔® oneAPI和TPP张量函数库加速AlphaFold2的蛋白结构预测分析

基于开源组学加速框架的scRNA聚类、降维和可视化

基于英特尔® OpenVINO™ Model Server 部署生物影像目标检测和图像分割

基于英特尔® 实现CPU上LLM大语言模型推理量化加速

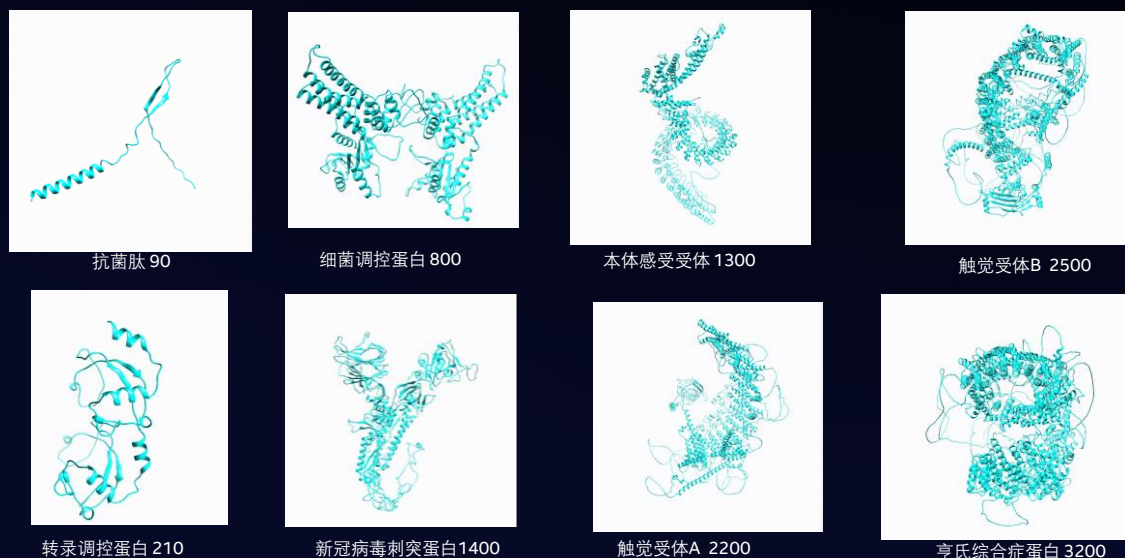
英特尔® CPU MAX 计算加速的新特性

AlphaFold2 的意义和挑战

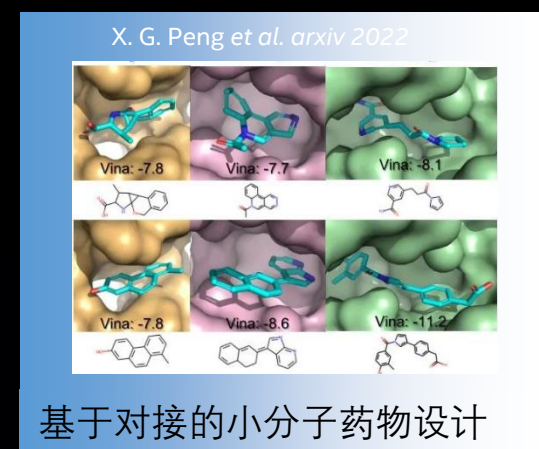
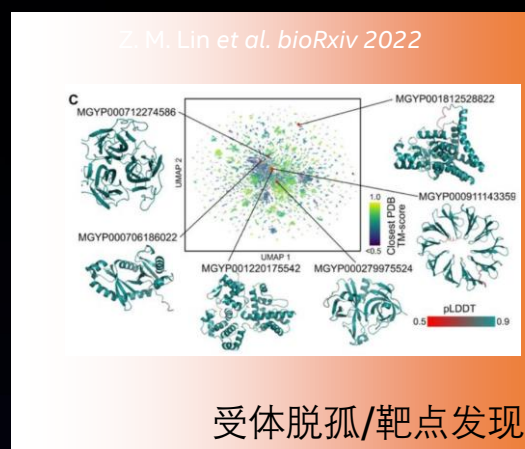
1. AlphaFold2 加速结构解析:



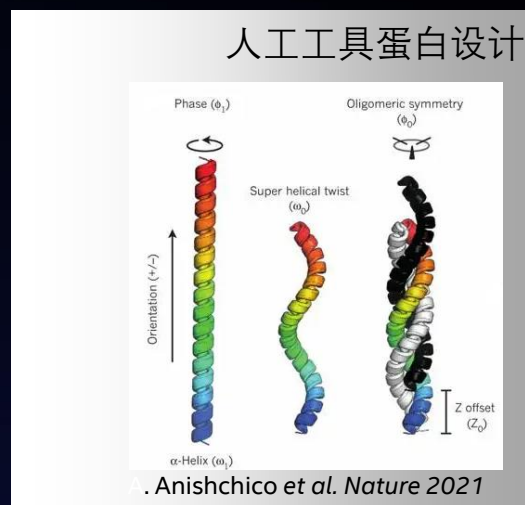
3. Intel AlphaFold2 在阿里云上的结果样例:



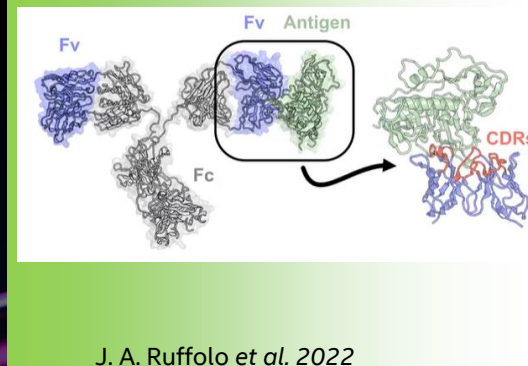
2. AlphaFold2 已成为AI制药新的基础架构



AlphaFold2

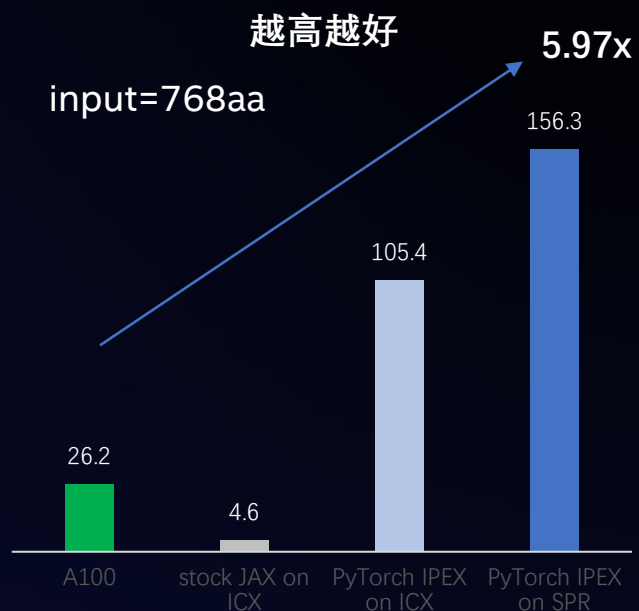


单克隆抗体设计



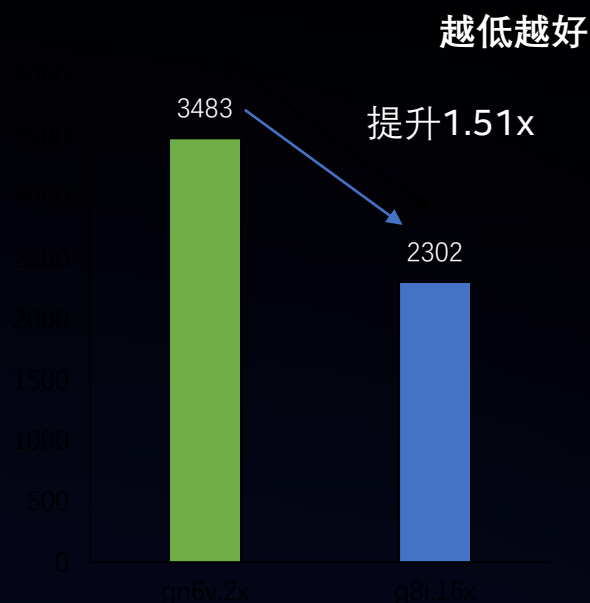
结果总结

高通量场景 (非阿里云数据)



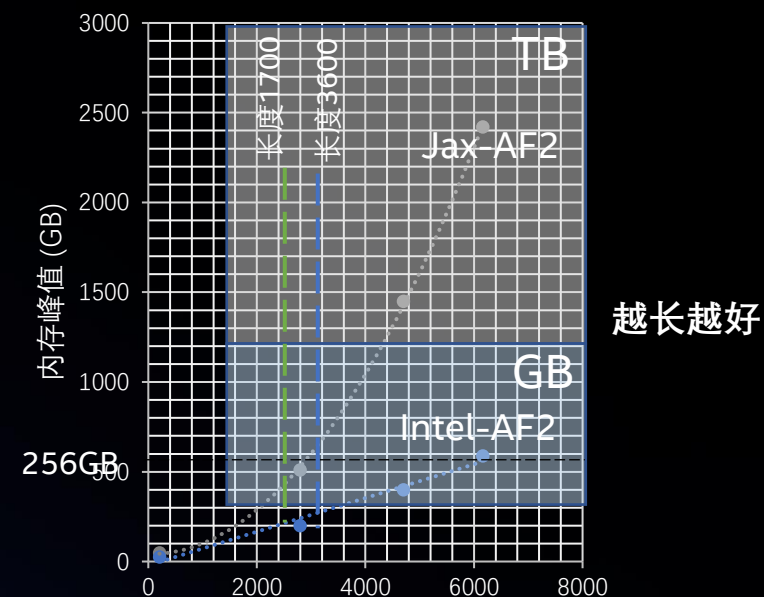
- 多实例预测在SPR单节点上总吞吐量达到单卡A100的 **5.97x**
- 且4台g8i-2x实例月租才与gn6v.2x实例的相当 (~95%)

单实例场景(阿里云数据)



- 单实例预测在g8i-16x SPR实例上达到gn6v.2x 单卡V100实例的 **1.51x**
- 且g8i-16x实例月租与gn6v.2x相当的 **92%**
- 前者性价比(吞吐量/RMB)是后者的 **1.64x**

长序列场景(阿里云数据)



- 最大内存/显存的实例上, 在SPR上用Intel-AF2能预测的样本长度翻了 **2.1x**
- 且SPR实例月租只有V100实例的 **23.2%**

议程

基于英特尔® oneAPI和TPP张量函数库加速AlphaFold2的蛋白结构预测分析

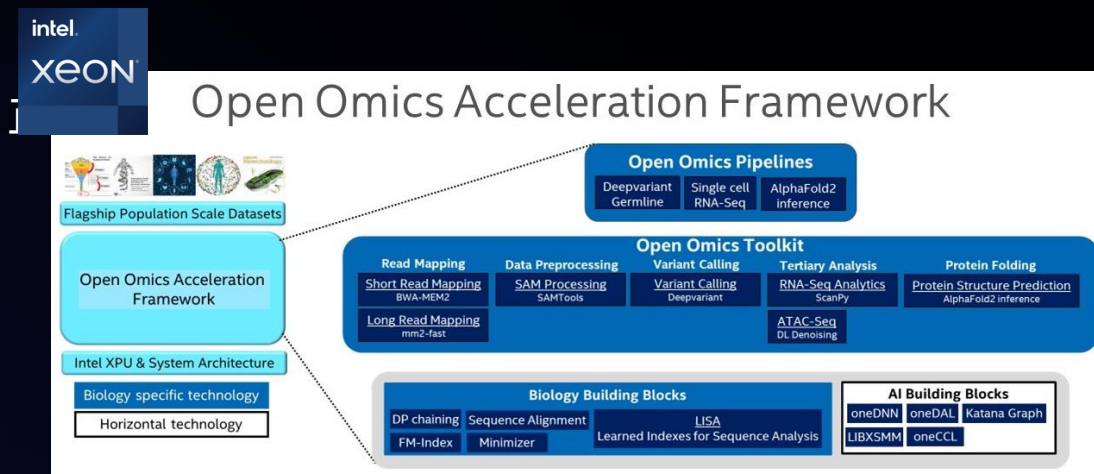
基于开源组学加速框架的scRNA聚类、降维和可视化

基于英特尔® OpenVINO™ Model Server 部署生物影像目标检测和图像分割

基于英特尔® 实现CPU上LLM大语言模型推理量化加速

英特尔® CPU MAX 计算加速的新特性

基于英特尔®至强®的scRNA单细胞测序后分析



<https://github.com/IntelLabs/Open-Omics-Acceleration-Framework>

Single cell analysis of 1.3 Million mouse brain cells (RNA_Seq)

```
In [1]: import time
import dask4py
from sklearnex import patch_sklearn
patch_sklearn()

import numpy as np
import scanpy as sc
from sklearn.cluster import KMeans

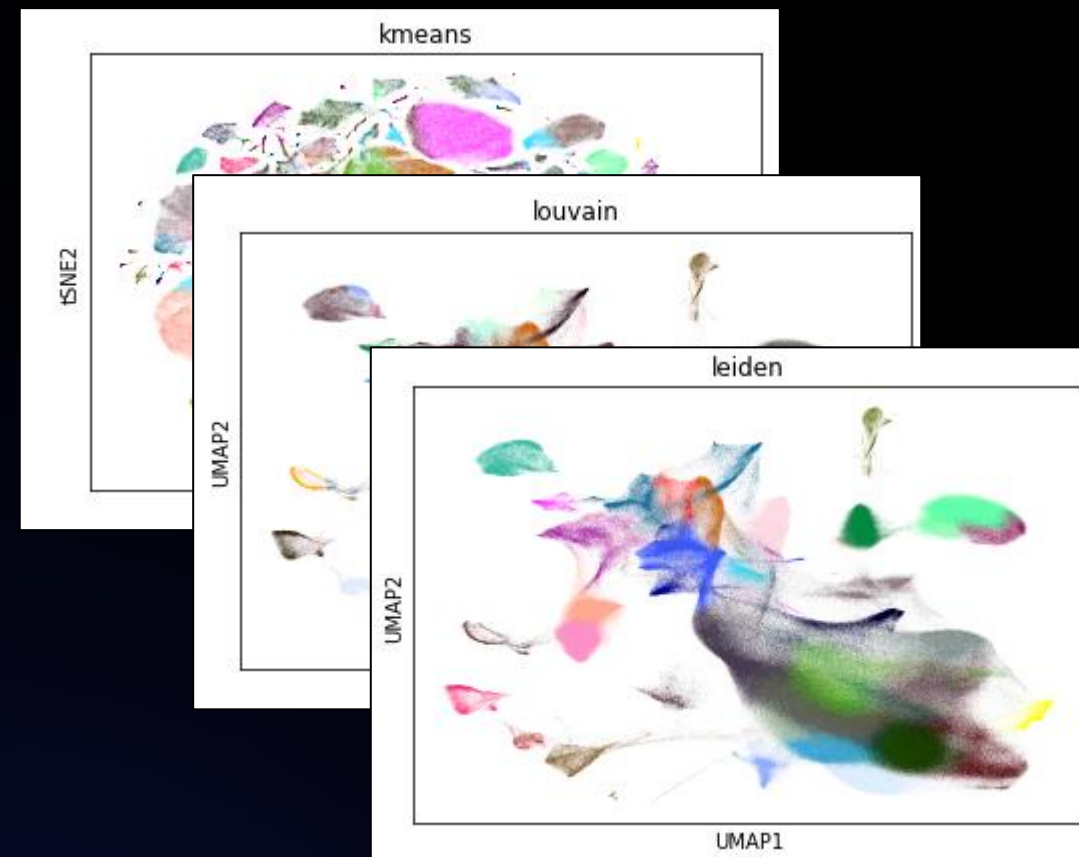
import os, wget
import mkl
import math

import fastpp
import sc_nbrs

import warnings
warnings.filterwarnings('ignore', 'Expected ')
warnings.simplefilter('ignore')
```

Intel(R) Extension for Scikit-learn* enabled (https://github.com/intel/scikit-learn-intelx)

提供优化的测序或结构分析工具及教程



scRNA分析支持多种聚类方式和特征维度

基于英特尔®至强®的scRNA单细胞测序后分析

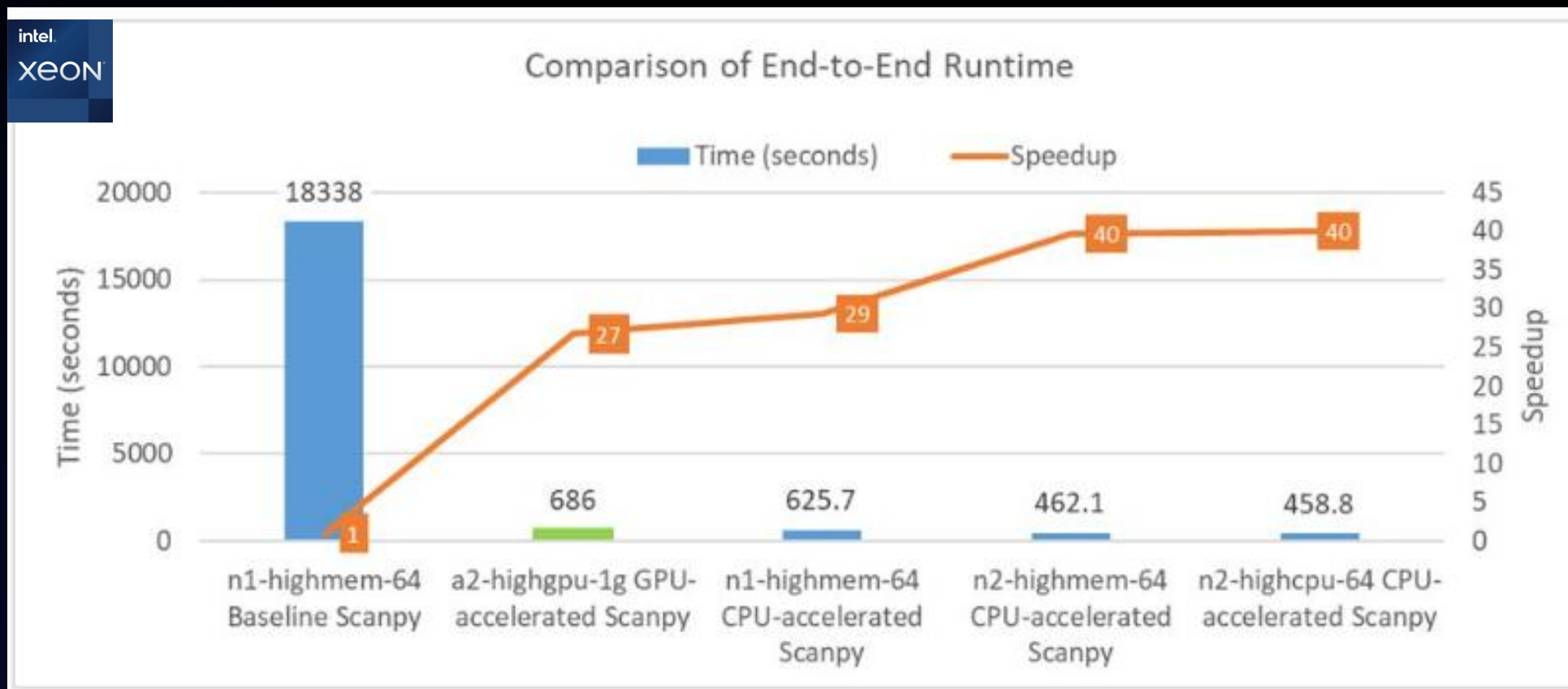
intel
XEON



scRNA后分析的管线示意图: 其中降维、聚类和可视化处理是计算耗时的主要瓶颈

<https://community.intel.com/t5/Blogs/Tech-Innovation/Artificial-Intelligence-AI/Intel-Labs-Accelerates-Single-cell-RNA-Seq-Analysis/post/1390715>

基于英特尔®至强®的scRNA单细胞测序后分析



scRNA后分析在英特尔®至强® CPU上的速度优化到原来的40x，比单卡参考GPU速度还快

<https://community.intel.com/t5/Blogs/Tech-Innovation/Artificial-Intelligence-AI/Intel-Labs-Accelerates-Single-cell-RNA-Seq-Analysis/post/1390715>

基于开源组学加速框架的降维、聚类 and 可视化流程

基于Katana 和 PyScan 的 Louvain的图聚类分析和UMAP可视化

基于Kmeans 聚类 and tSNE可视化

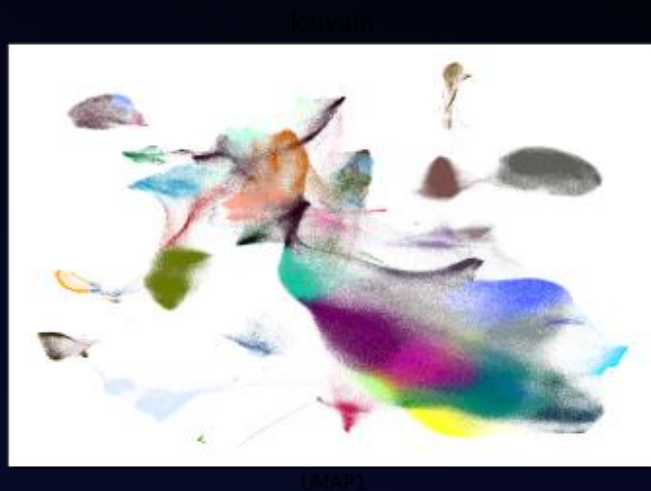
```
louvain_plan = LouvainClusteringPlan.do_all(enable_vf, modularity_threshold_per_round, modularity_threshold_total, max_iter)
# Louvain_plan = LouvainClusteringPlan.deterministic(enable_vf, modularity_threshold_per_round, modularity_threshold_total,

louvain_clustering(graph, "value", "output", plan=louvain_plan)
stats = LouvainClusteringStatistics(graph, "value", "output")
print(stats)
groups = graph.get_node_property("output").to_numpy().astype('int')

adata.obs['louvain'] = pd.Categorical(values=groups.astype('U'), categories=natsorted(map(str, np.unique(groups))),)
```

```
In [18]: %time
sc.pl.umap(adata, color=["louvain"])
```

intel.
XEON



```
In [11]: %time
daal4py.daalinit(sc.settings.n_jobs)
sc.tl.tsne(adata, n_pcs=tsne_n_pcs)

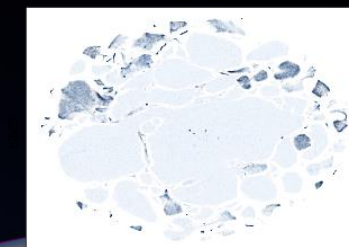
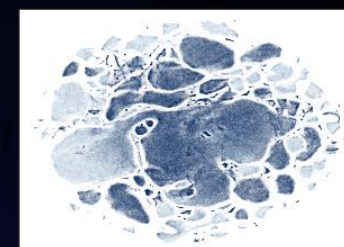
CPU times: user 2h 1min 28s, sys: 4min 25s, total: 2h 5min 54s
Wall time: 3min 37s

In [12]: %time
kmeans = KMeans(n_clusters=k, random_state=0).fit(adata.obsm['X_pca'])
adata.obs['kmeans'] = kmeans.labels_.astype(str)

CPU times: user 5min 47s, sys: 41 s, total: 6min 28s
Wall time: 8.4 s

In [13]: %time
sc.pl.tsne(adata, color=["kmeans"])
sc.pl.tsne(adata, color=["Stm2_raw"], color_map="Blues", vmax=1, vmin=-0.05)
sc.pl.tsne(adata, color=["Hes1_raw"], color_map="Blues", vmax=1, vmin=-0.05)
```

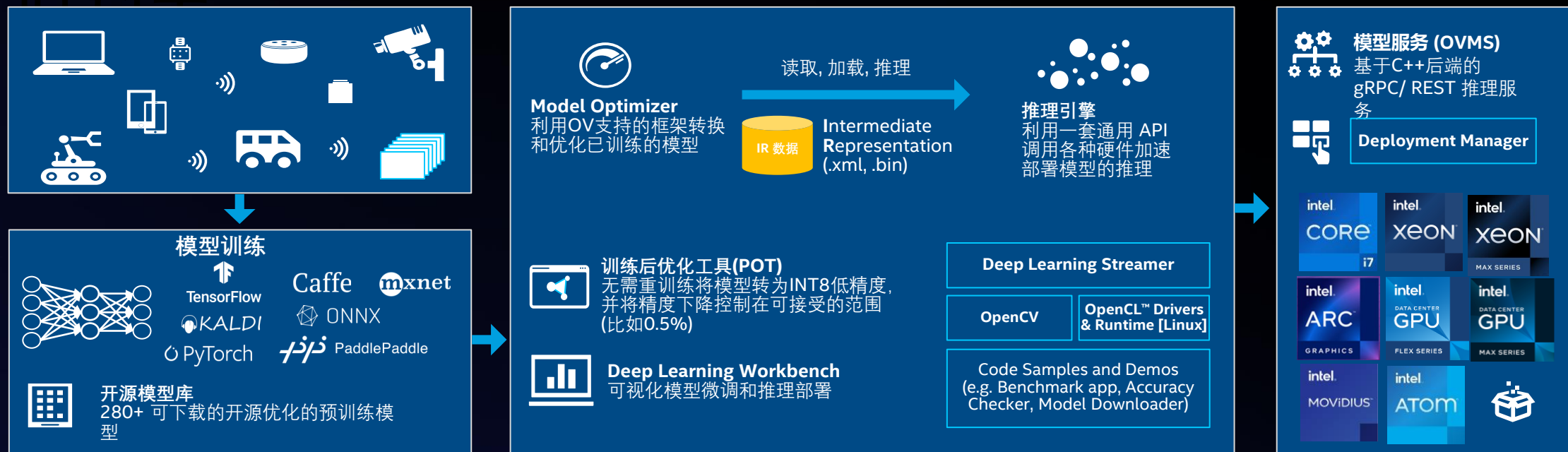
intel.
XEON



https://github.com/IntelLabs/Open-Omics-Acceleration-Framework/blob/main/pipelines/single_cell_pipeline/notebooks/1.3_million_single_cell_analysis.ipynb

OpenVINO™ 工具使用路径

1 构建 2 优化 3 部署



议程

基于英特尔® oneAPI和TPP张量函数库加速AlphaFold2的蛋白结构预测分析

基于开源组学加速框架的scRNA聚类、降维和可视化

基于英特尔® OpenVINO™ Model Server 部署生物影像目标检测和图像分割

基于英特尔® 实现CPU上LLM大语言模型推理量化加速

英特尔® CPU MAX 计算加速的新特性

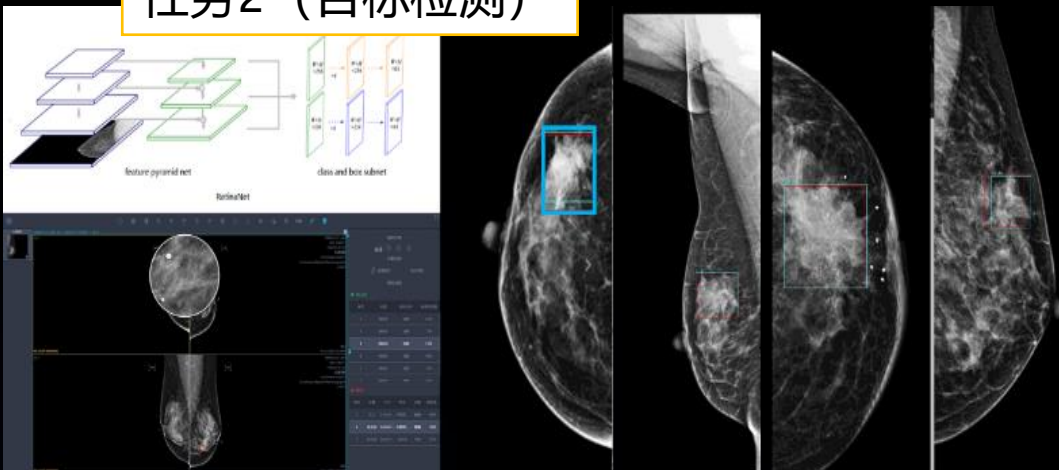


汇医慧影AI医学影像解决方案



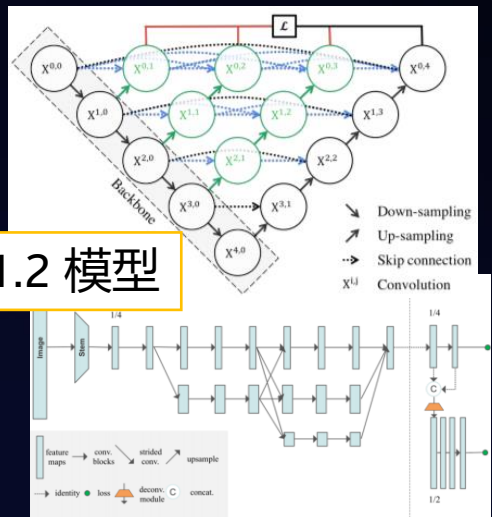
基于汇医慧影的Dr.Turing AI平台的新冠肺炎筛查

任务2 (目标检测)



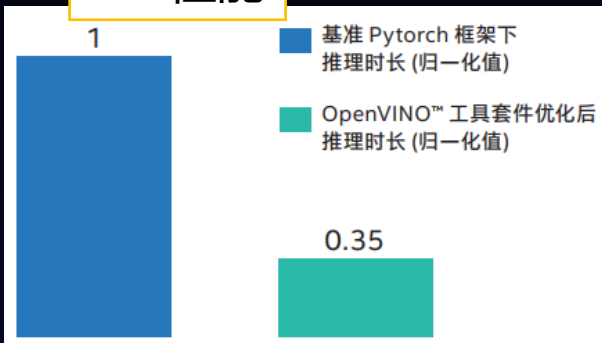
基于ResNet50-RetinaNet模型的乳腺癌筛查

1.2 模型



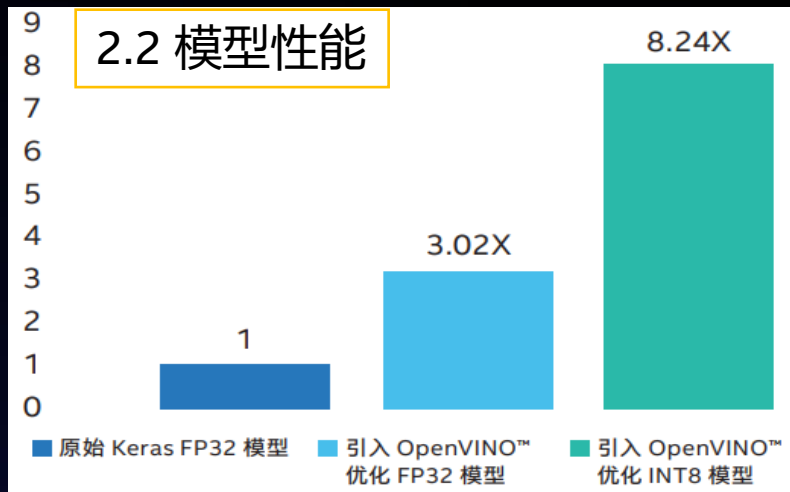
基于U-Net++和HRNet的病灶分割

1.3 性能



使用OpenVINO工具套件对新冠肺炎CT检测模型优化前后效果对比

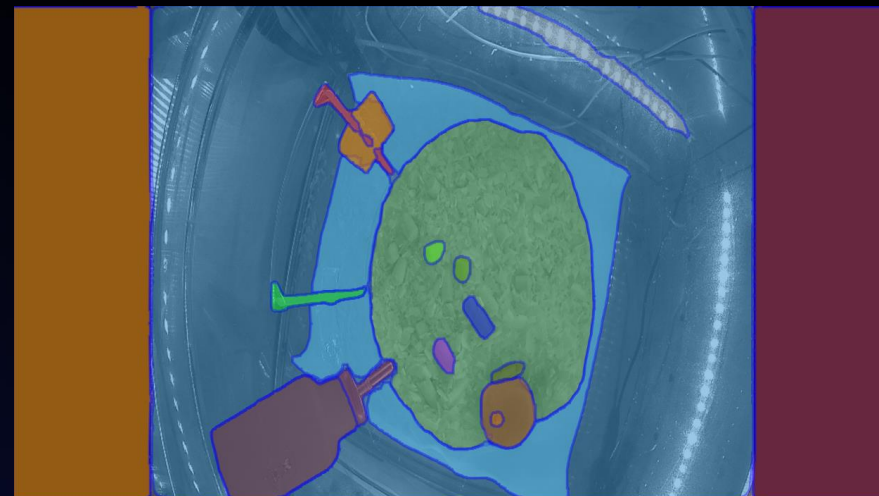
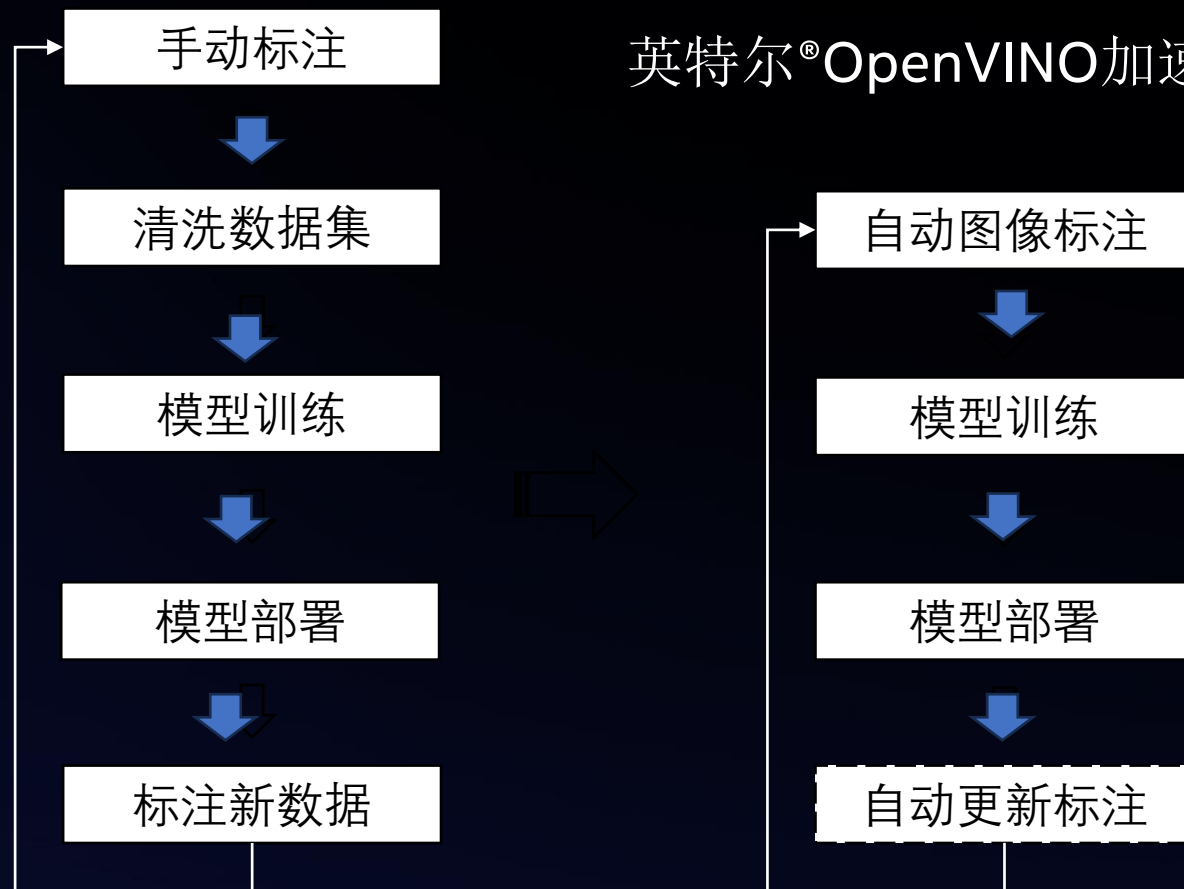
2.2 模型性能



OpenVINO 低精度加速INT8推理速度提升

Segment-Anything 革新了生物图像分析流程

英特尔®OpenVINO加速的SAM



议程

基于英特尔® oneAPI和TPP张量函数库加速AlphaFold2的蛋白结构预测分析

基于开源组学加速框架的scRNA聚类、降维和可视化

基于英特尔® OpenVINO™ Model Server 部署生物影像目标检测和图像分割

基于英特尔® 实现CPU上LLM大语言模型推理量化加速

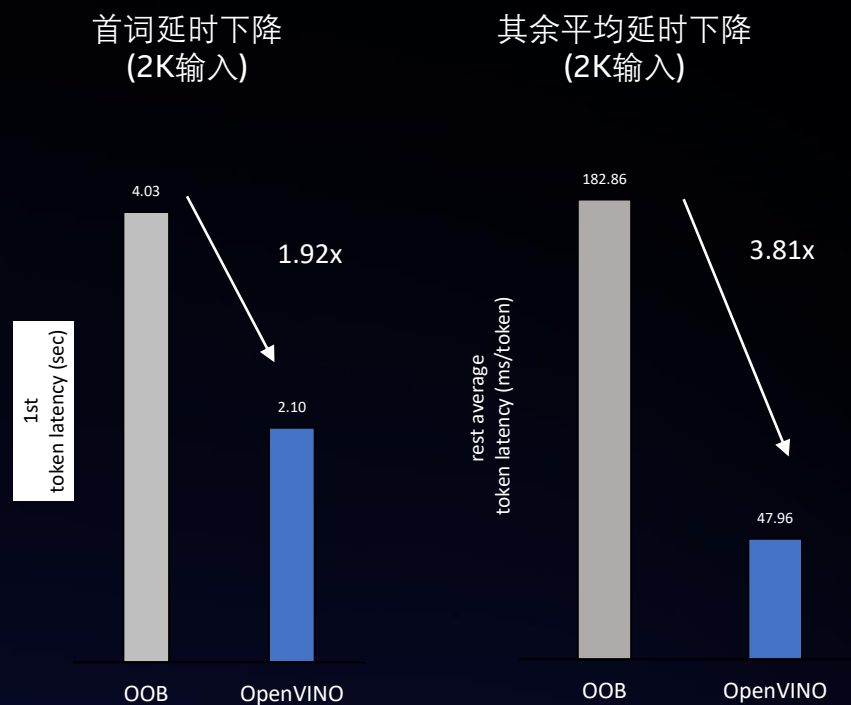
英特尔® CPU MAX 计算加速的新特性

ChatGLM-6B: 医疗大模型BF16非量化推理加速



HuiMei

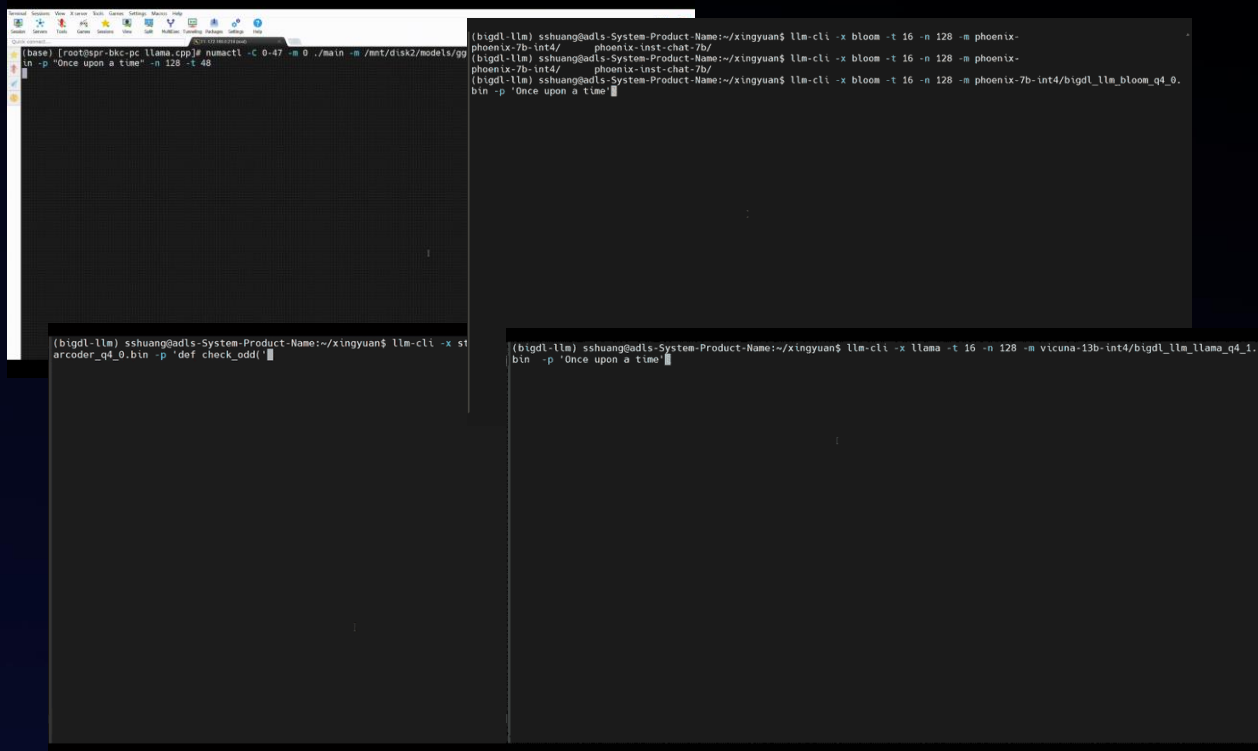
英特尔® OpenVINO™-LLM非量化方案
加速英特尔®至强®MAX实例加速性能(48核)



- 1S-SPR9468 Quadrant/HBM-Cache

低精度 (INT8) 加速 LLM 在 SPR 上的推理性能

基于Sapphire Rapids BigDL Nano 的 INT4 实时 Demo



```
(bigdl-llm) sshuang@adls-System-Product-Name:~/xingyuan$ llm-cli -x bloom -t 16 -n 128 -m phoenix-7b-int4/ phoenix-inst-chat-7b/
(bigdl-llm) sshuang@adls-System-Product-Name:~/xingyuan$ llm-cli -x bloom -t 16 -n 128 -m phoenix-7b-int4/ phoenix-inst-chat-7b/
(bigdl-llm) sshuang@adls-System-Product-Name:~/xingyuan$ llm-cli -x bloom -t 16 -n 128 -m phoenix-7b-int4/bigdl_llm_bloom_q4_0.
bin -p 'Once upon a time'

(bigdl-llm) sshuang@adls-System-Product-Name:~/xingyuan$ llm-cli -x llama -t 16 -n 128 -m vicuna-13b-int4/bigdl_llm_llama_q4_1.
bin -p 'Once upon a time'

(bigdl-llm) sshuang@adls-System-Product-Name:~/xingyuan$ llm-cli -x startcoder_q4_0.bin -p 'def check_odd()'
bin -p 'Once upon a time'
```

- 支持所有主流huggingface或ggml的LLM推理
- 将 FP16/BF16 转为 INT8/INT4, 以便简化计算复杂度

<https://github.com/intel-analytics/BigDL/tree/main/python/llm>

LLaMA-7B, BLOOM-7B1, Startcoder-15B, Pheonix on ggml + BigDL

llama.cpp ggml OpenBLAS AVX2 -> intrinsics AVX512-VNNI (INT8/4)

议程

基于英特尔® oneAPI和TPP张量函数库加速AlphaFold2的蛋白结构预测分析

基于开源组学加速框架的scRNA聚类、降维和可视化

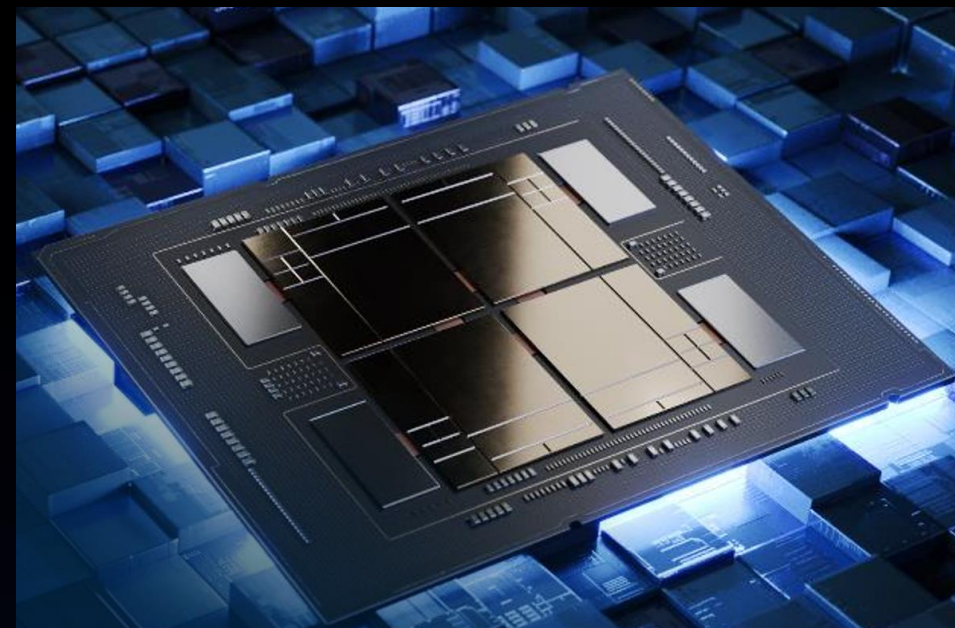
基于英特尔® OpenVINO™ Model Server 部署生物影像目标检测和图像分割

基于英特尔® 实现CPU上LLM大语言模型推理量化加速

英特尔® CPU MAX 计算加速的新特性

英特尔® CPU Max处理器

>内存带宽受限的应用中有**2x***性能提升



Compute	 up to 56 cores	AMX Advanced Matrix Extensions	Integrated Acceleration Engines	up to 350W
Memory	 64GB HBM2e	up to 112.5MB shared LLC	6TB Max mem size	supports 1DPC - 4800 MT/s 16 DIMMs per socket

I/O	CXL 1.1 Compute Express Link	PCIe Gen 5	UPI 2.0 Improved Multi-socket Scaling	Security & Reliability Technologies
Technology	 EMIB	Multi-Tile	Logically Monolithic	Scalable up to 2S

英特尔® 高级矩阵指令扩展集 (AMX)



AMX 架构包含两个部分:

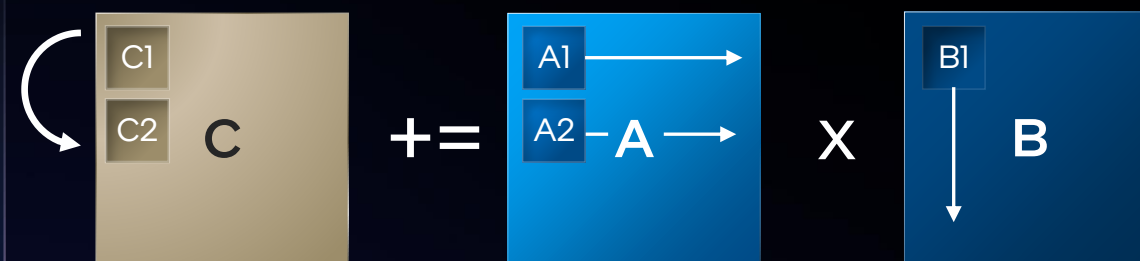
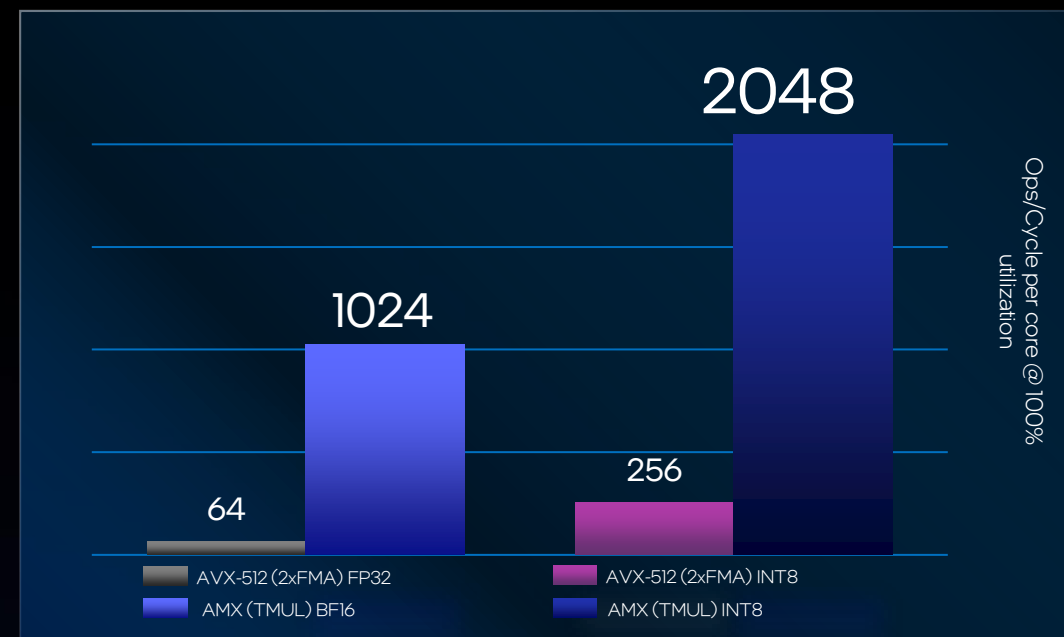
Tiles 数据块

- 一种全新的2D可扩展寄存器组 – 8个新的寄存器, 每个1Kb: T0-T7
- 该寄存器组支持基本操作包括 – load/store, clear, set to constant等
- TILES 支持通过XSAVE架构保存状态并接受系统管理

TMUL 脉动阵列

- 矩阵乘法指令集, 是首个基于TILES的指令集
- 包括一组 MAC 计算网格用于计算“成块”的数据
- TMUL – 使用3组Tile寄存器(T2=+T1*T0)执行矩阵的乘加运算 ($C = +A * C$)
- TMUL 计算依赖于 TILE 寄存器

Express more work per instruction and per μop – save power for fetch/decode/OOO



服务

一体机

私有服务器

HPC集群

私有云

公有云

任务

分子对接 分子动力学 小分子生成 AGME属性预测 蛋白互作 大分子结构预测 抗体-抗原-TCR 界面设计

模型

Regression Clustering PCA Bayesian XGBoost SVM MLP CNN RNN Transformers GPT GCN/GRN RL

框架

ScanPy Scikit-learn PyTorch TensorFlow PaddlePaddle ONNX Caffe mxnet KALDI OpenVINO

硬件

CPU

易获取的AI算力



- 矩阵加速: AMX-BF16/INT8
- 内存扩展: DDR5, HBM2e
- 隐私计算: SGD/TDX, PPML

GPU

易扩展的异构算力



- XMX矩阵加速引擎
- VectorEngine矢量加速引擎

TPU

高密度的AI算力



- 高性能AI训练/推理: LLM等
- 多卡分布式: Gaudi Link

总结

- 英特尔®提供包括oneAPI, TPP/libxmm, OpenVINO, xFasterTransformer等加速阿里云上的8代ECS实例(SPR-SP)和SPR-HBM实例
- 在蛋白结构预测和大分子设计领域，英特尔®实现了AlphaFold2管线端到端优化，并在阿里云8代ECS实例完成了部署和TCO及性能验证
- 在医疗影像领域，英特尔®全面实现了通用CV模型的适配和加速
- 英特尔®联合多家partner，可以为医疗计算提供更完整的生态

谢 谢