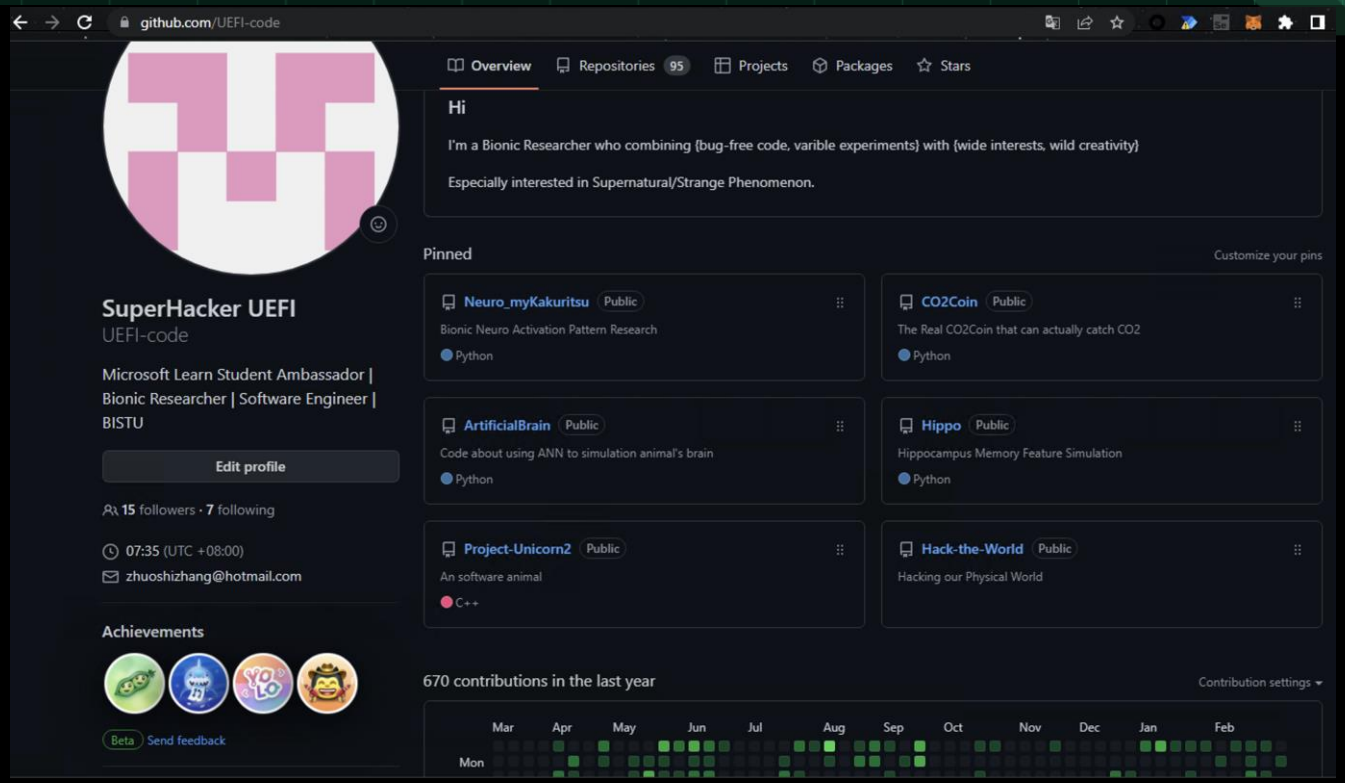
Microsoft Azure intel CLOUD NATIVE COMPUTING FOUNDATION

高性能向量计算编程实践

提高深度学习业务响应速度，降低部署成本

张世卓
微软学生大使 & 开源爱好者
2023/04/22

讲师介绍



微软学生大使、中国发明协会会员、北信科在读20级本科生

热爱发明创新、开源、Coding、理化生实验&机器人仿生学综合研究

已获发明专利2项、实用新型专利多项

目录

CATALOGUE

经典深度学习云推理管道部署模式

云原生部署：构建Docker镜像，部署到K8S，使用CPU推理

使用CPU通常具有如下优势：

CPU节点比GPU节点租金便宜很多

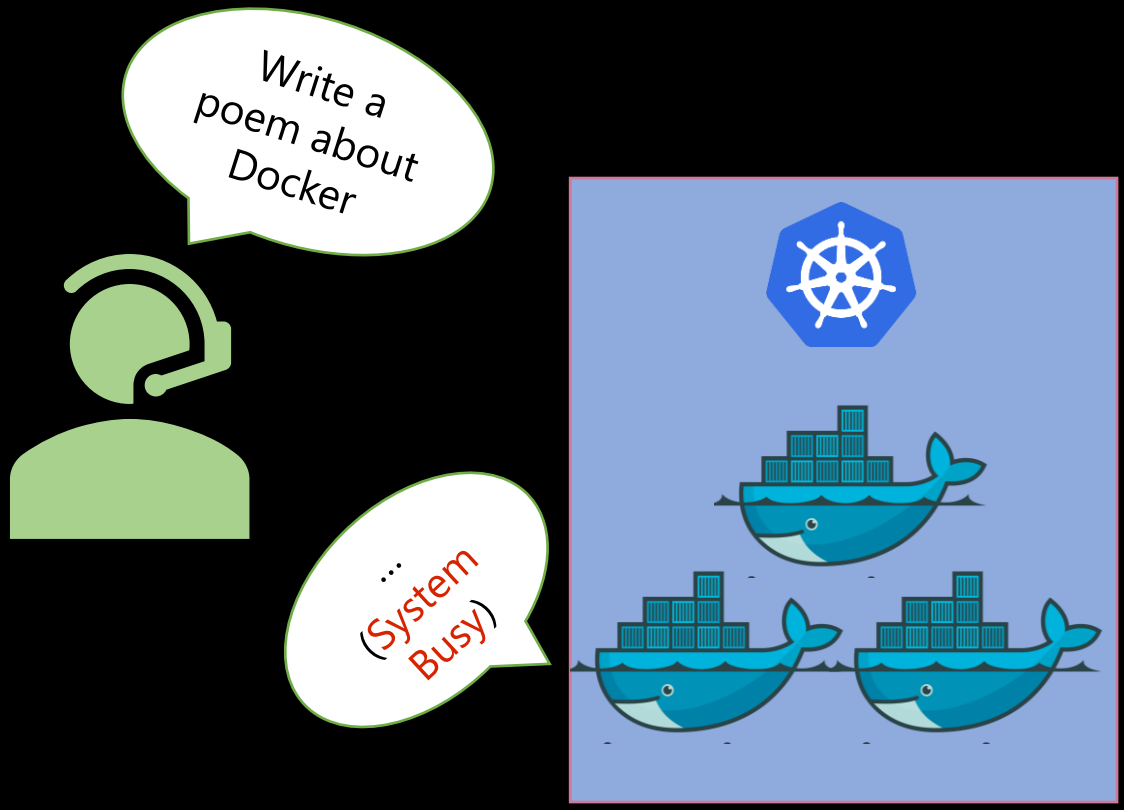
避免数据从RAM到VRAM的拷贝延迟

现代CPU指令集变化较GPU平缓

在多样化环境兼容性好

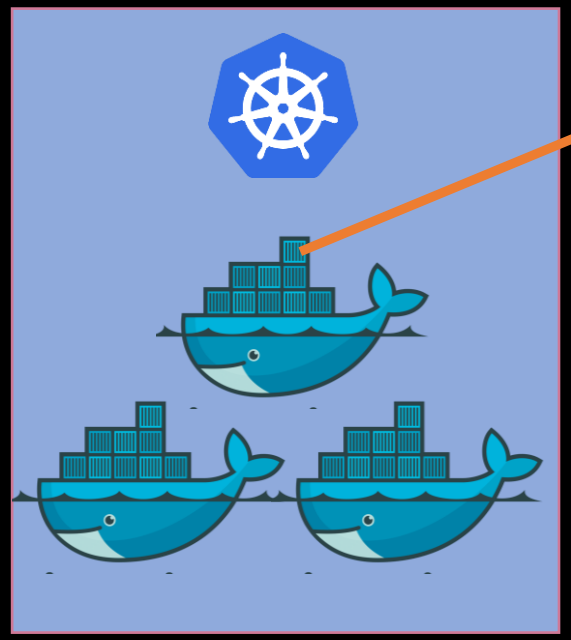
使用CPU通常具有如下缺点：

并行处理能力比GPU弱，模型参数量大时只能串行遍历



深度学习推理框架性能直接影响集群响应时间

…
(System Busy)



阻塞: 前端Wait:
推理线程x

当前Stack:
嵌入层1->矩阵乘1->For Loop

如果使用传统循环遍历神经网络参数，
将是非常低效的做法

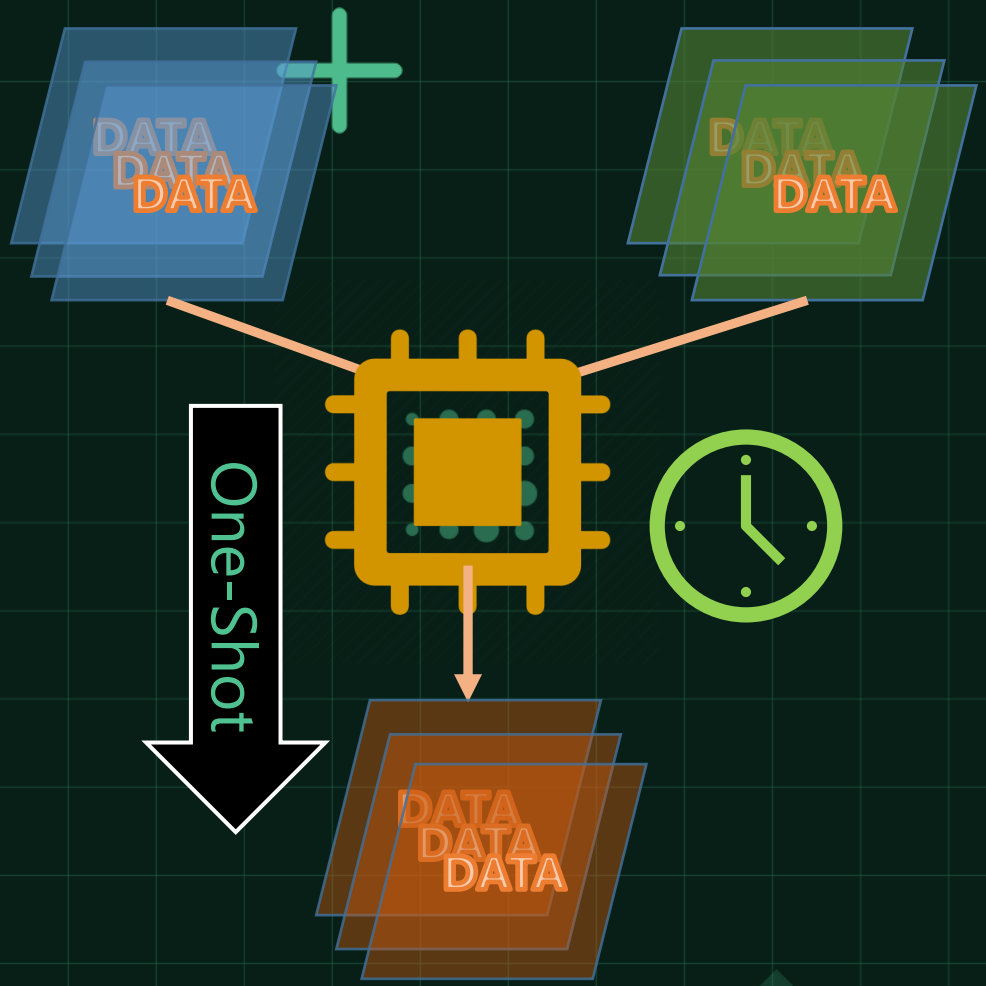
编译器自动优化
经常不可靠

改善编程
习惯/使用
优质框架

CPU其实也有针对大向量计算的优化：
向量化计算指令集-单指令多数据处理

$$$: 按CPU时间计费

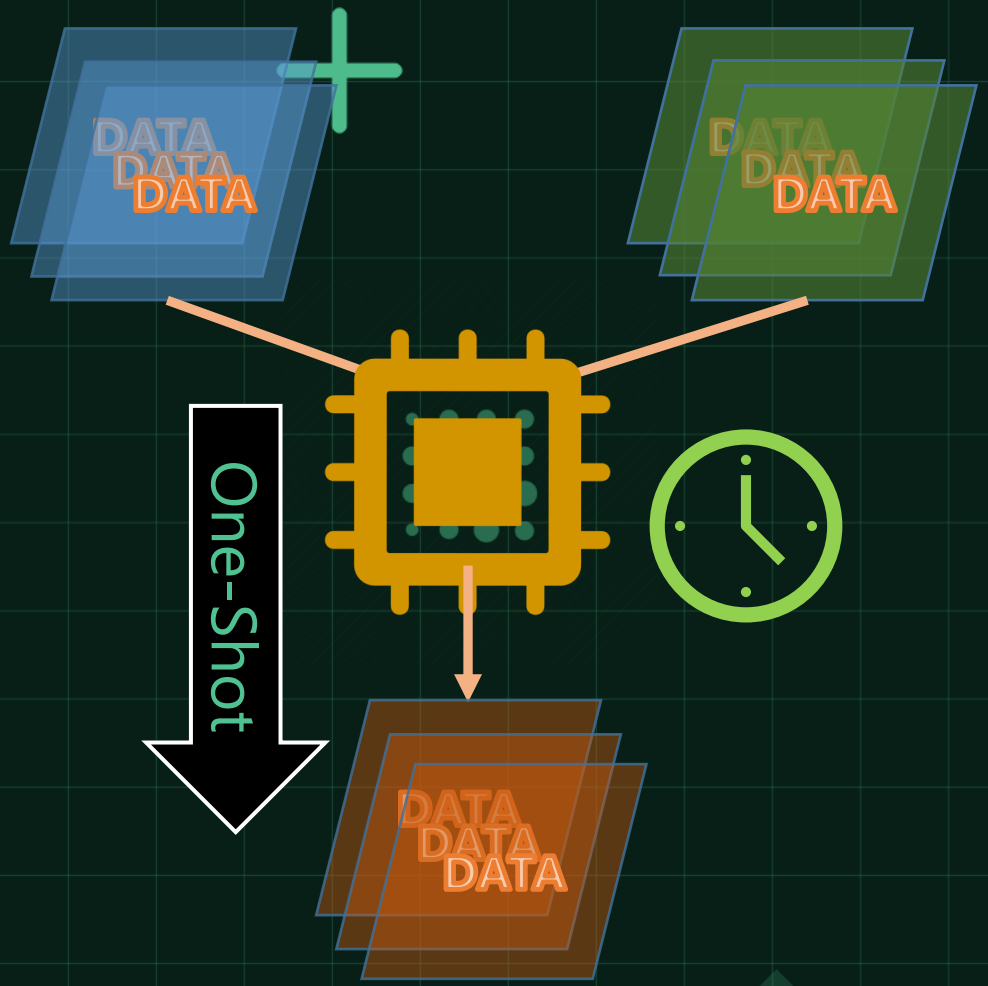
向量化计算概念



向量化计算是一种利用处理器的SIMD（单指令多数据）能力来加速数学运算的技术。在向量化计算中，数值数据被存储在连续的内存地址中，处理器可以一次性地对整个数据向量执行相同的操作。

这种方法能够提高运算效率和吞吐量，因为它可以减少内存访问和操作指令的数量，从而降低了处理器的负载和延迟。向量化计算在许多科学计算、数据分析和机器学习应用中得到广泛应用。

向量化计算应用场景



需求实例：两个向量求和

```
__declspec(align(16)) float vecA[] = {1.1, 2.2, 3.3, 4.4};
__declspec(align(16)) float vecB[] = {5.5, 6.6 3.3, 4.4};
__declspec(align(16)) float res[4] = {0.0}; //res = vecA + vecB
```

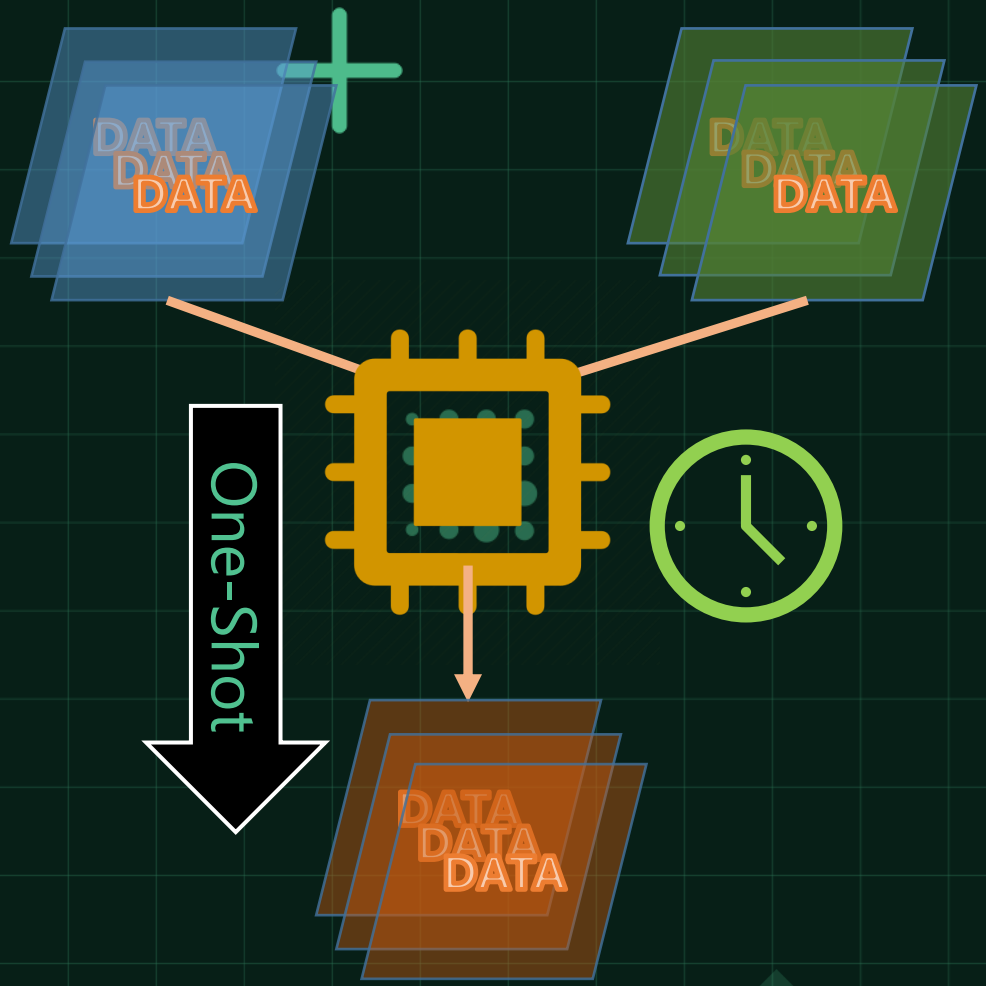
```
;Code Block A
LEA RAX, vecA;
LEA RBX, vecB;
LEA, RCX, vecC;
MOV RDX, 0;
LOOOP:
MOVSS XMM0, [RAX];
MOVSS XMM1, [RBX];
ADDSS XMM0, XMM1;
MOVSS [RCX], XMM0
ADD RAX, 4;
ADD RBX, 4;
ADD RCX, 4;
ADD DL, 1;
CMP DL, 4;
JB LOOOP;
...;END
```

SLOW

```
;Code Block B
LEA RAX, vecA;
LEA RBX, vecB;
LEA, RCX, vecC;
MOVAPS XMM0, [RAX]
MOVAPS XMM1, [RBX]
ADDPS XMM0, XMM1;
MOVAPS [RCX], XMM0
...;END
```

FAST

使用向量化计算的好处

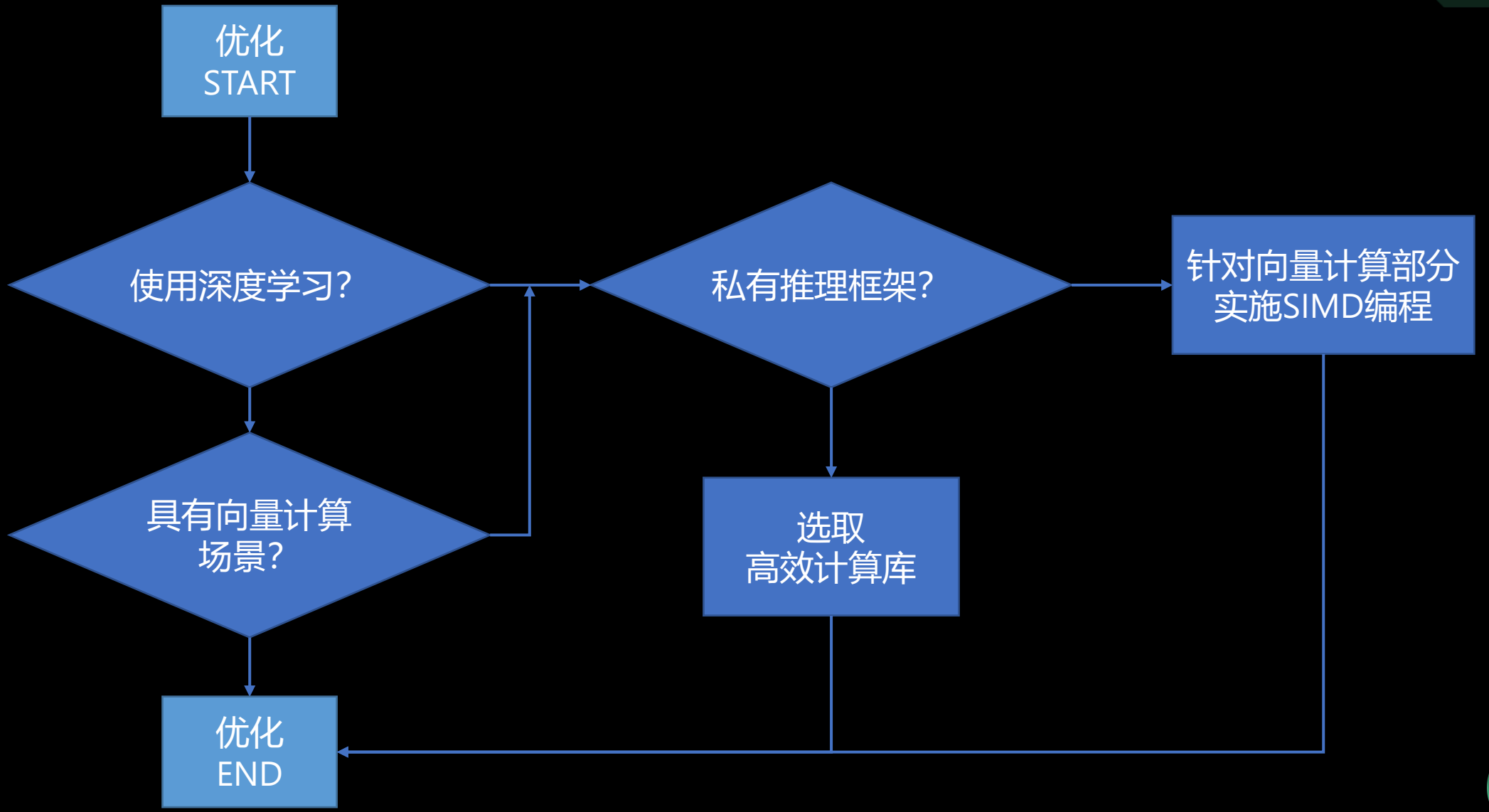


1. 编译后的程序Bin体积小
2. 程序加载快、运行快
4. 操作系统无关性（兼容性好）
5. 对深度学习类应用效果显著

向量化计算编程注意事项

0. 不要指望编译器能优化的多好
1. 处理好内存对齐、数据格式
2. 手写汇编封装一个函数供上层调运行效率最高
3. 注意处理器硬件架构，尤其是使用较新的指令集时，有必要检查处理器FLAG

TAKE ACTION

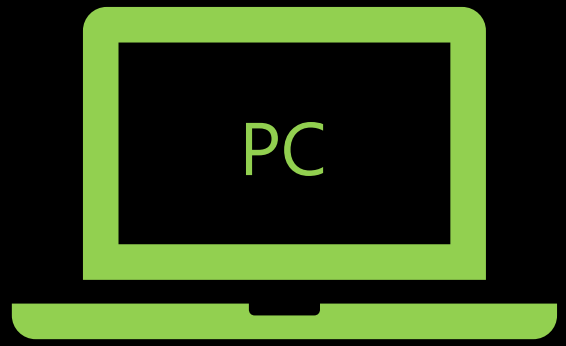

优化
START
使用深度学习?
具有向量计算
场景?
私有推理框架?
选取
高效计算库
针对向量计算部分
实施SIMD编程
优化
END

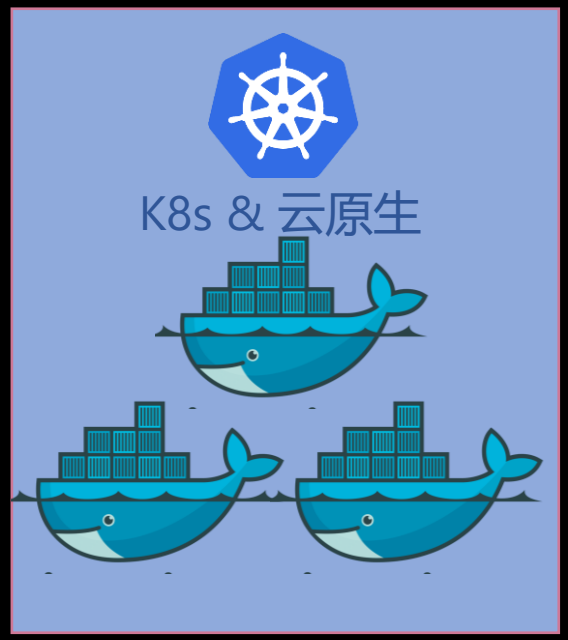
OpenVINO介绍

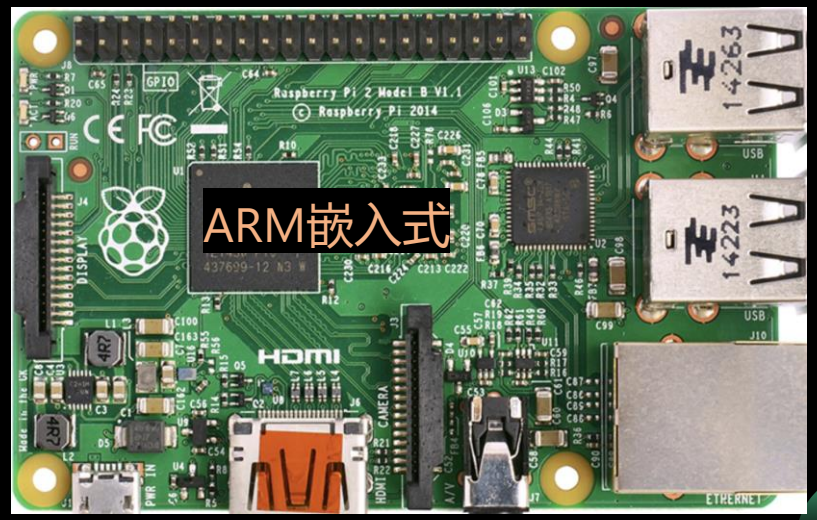
如果你不想自己编框架，
可用现成的，
SIMD Optimized

OpenVINO介绍

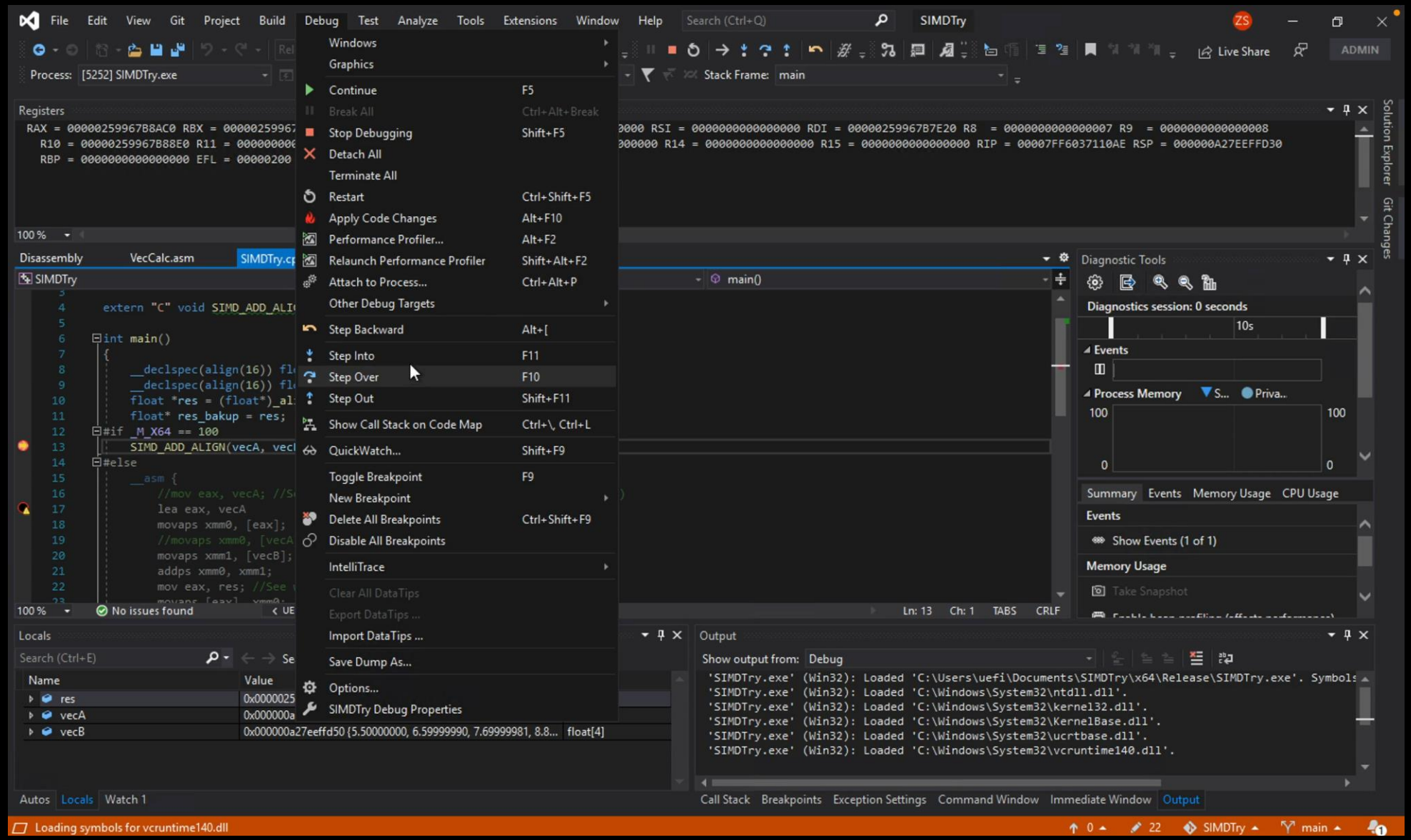
- OpenVINO是一个高性能的深度学习推理框架（你可用C/Cpp/Python调API)
- 诞生之初，OpenVINO主要对Intel自家产品优化，以发挥X86 CPU、核心显卡、VPU（加速棒）、FPGA性能为主。
- 现在已经支持ARM CPU，也能通过OpenCL使用NVIDIA GPU





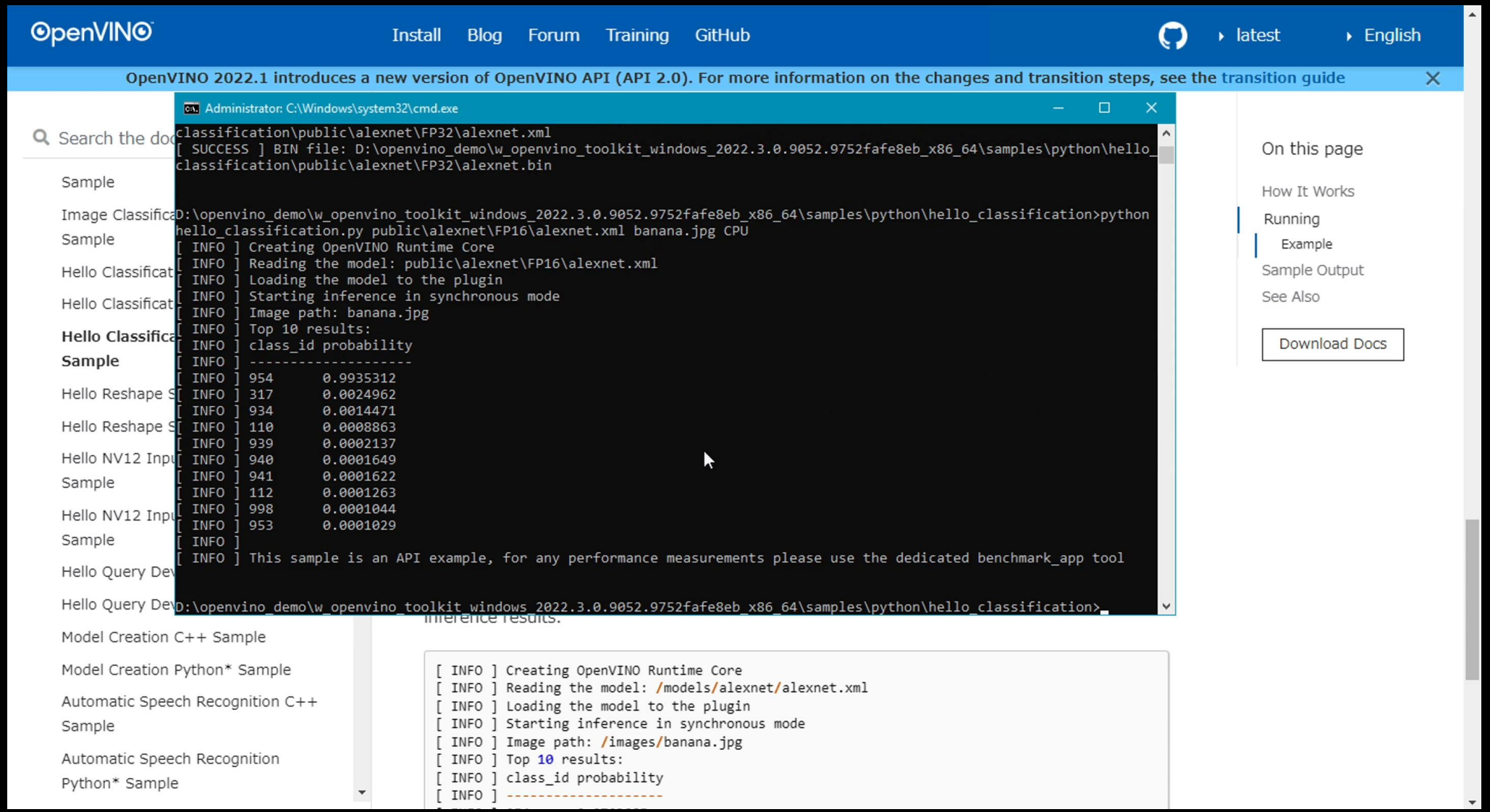


Demo-SIMD



Demo-Repo

Demo-VINO



VINO-Video

Thanks