



58同城深度学习平台资源使用率优化实践

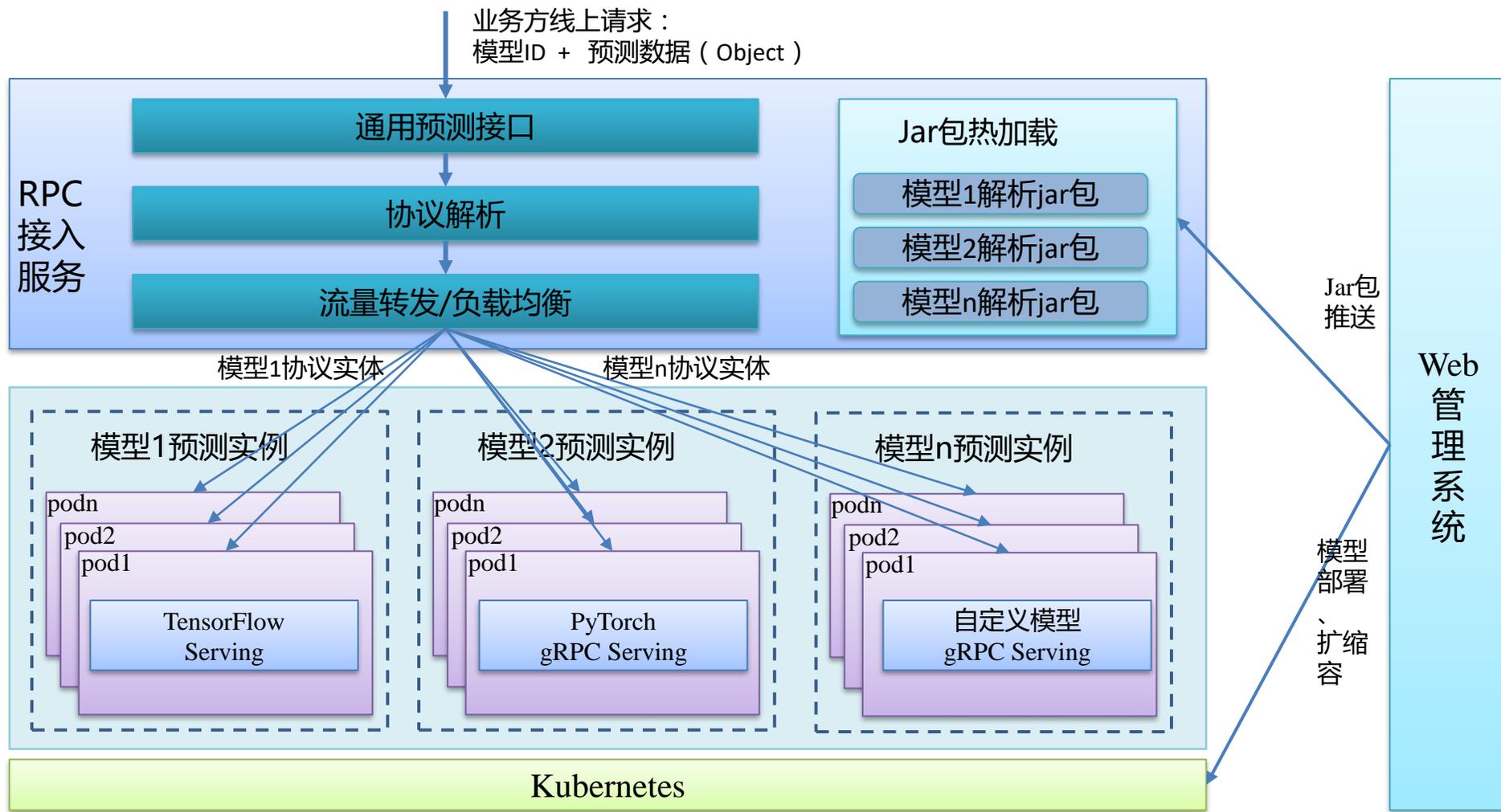


- **深度学习平台简介**
- **资源使用率优化背景介绍**
- **提升模型推理性能**
 - Intel MKL库应用
 - Intel OpenVINO推理引擎集成
- **优化模型推理资源分配和调度**
 - TensorFlow模型混合部署
 - GPU虚拟化技术应用
 - 模型推理资源监控告警
- **总结**



深度学习平台总体架构





- 深度学习平台简介
- 资源使用率优化背景介绍
- 提升模型推理性能
 - Intel MKL库应用
 - Intel OpenVINO推理引擎集成
- 优化模型推理资源分配和调度
 - TensorFlow模型混合部署
 - GPU虚拟化技术应用
 - 模型推理资源监控告警
- 总结



- 部分模型推理时在CPU上latency达不到业务要求而必须使用GPU。 —— 解决方法：**使用Intel MKL库降低latency，使得部门原来必须跑在GPU上的模型可以迁移CPU，节省GPU资源。**
- 提升模型GPU上推理throughput，降低模型GPU资源使用量。 ——解决方法：**应用OpenVINO**
- k8s对GPU卡是按整数进行调度，开发环境、推理流量小模型、GPU使用率占用有限模型分配整张GPU卡存在资源浪费。 ——解决方法：**TensorFlow模型混合部署、GPU虚拟化技术应用**
- 线上模型资源配置不合理，资源存在浪费。 —— 解决方法：**推理资源监控告警，及时进行调整**



- 深度学习平台简介
- 资源使用率优化背景介绍
- **提升模型推理性能**
 - Intel MKL库应用
 - Intel OpenVINO推理引擎集成
- **优化模型推理资源分配和调度**
 - TensorFlow模型混合部署
 - GPU虚拟化技术应用
 - 模型推理资源监控告警
- **总结**



Intel® Math Kernel Library (Intel® MKL) 是一套高度优化、线程安全的数学例程、函数，面向高性能的工程、科学与财务应用。英特尔 MKL 的集群版本包括 ScaLAPACK 与分布式内存快速傅立叶转换，并提供了线性代数 (BLAS、LAPACK 和 Sparse Solver)、快速傅立叶转换、矢量数学 (Vector Math) 与随机号码生成器支持。

主要包括：

- LAPACK (线形代数工具 linear algebra package)
- DFTs (离散傅立叶变换 Discrete Fourier transforms)
- VML (矢量数学库 Vector Math Library)
- VSL (矢量统计库 Vector Statistical Library)



- **接口支持**

Intel MKL是一套经过高度优化和线程化的函数库，提供C和Fortran接口。

- **处理器支持**

可以为当前以及下一代处理器提供性能优化，其支持全部兼容英特尔处理器的处理器。说明：MKL会更加运行的处理器环境，自动运行时处理器检测，从而对不同的处理器运行不同的优化版本的程序，从而保证其在所运行的处理器上都能获得较好的性能，所以，有可能同一个使用了MKL的程序，在不同的处理器上运行的性能不同，因为MKL会针对不同的处理器进行检测，对其进行尽可能的最大化优化。

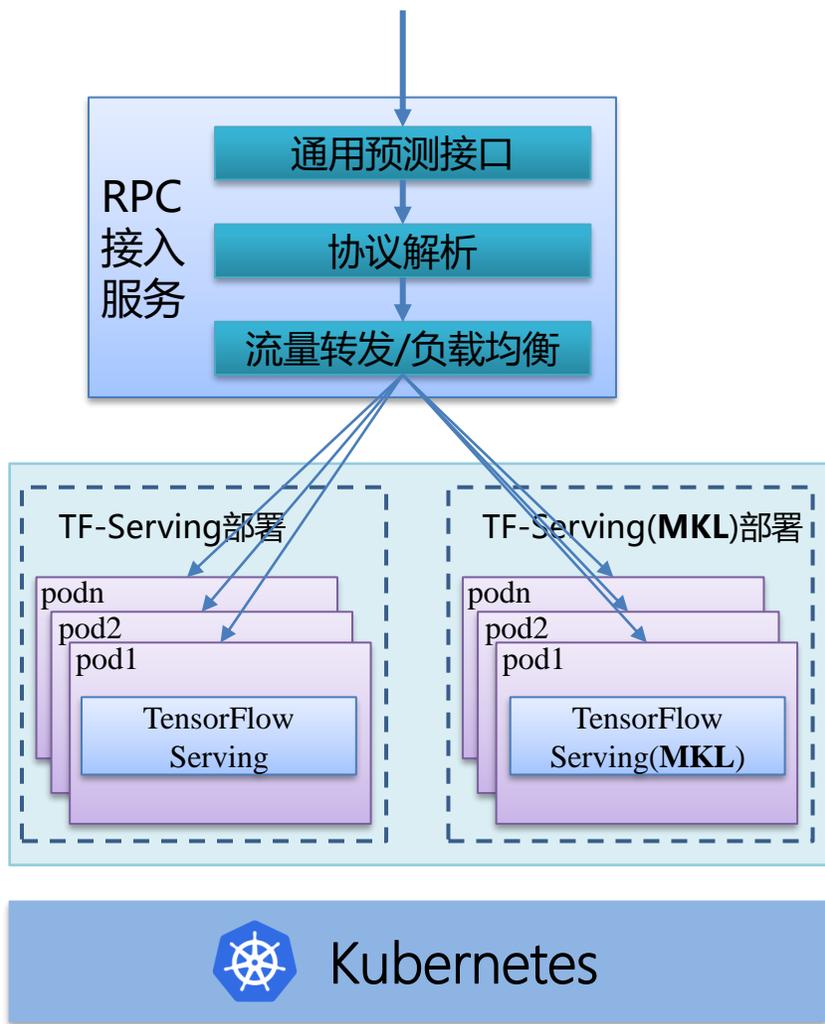
- **平台和工具支持**

支持主流的操作系统（Windows、linux、Mac OS等），与主流的开发工具（VS、Eclipse、Xcode、GCC等）集成。

- **多核多线程扩充性能和线程安全性**

内置并行处理机制，在多核和多处理器上自动获取出色的扩充性能。所有的MKL函数都是线程安全的。同时提供非线程化顺序型MKL。





OCR模型 Intel Xeon E5-2630 v4

mkl版本

预测耗时有效**减少62.07%**

CPU资源使用**增加177.3%**

TF版本	使用CPU资源/核	平均耗时/ms	QPS
1.12	1.19	150.0	6.6
1.12-mkl	3.30	60.6	16.4

- 深度学习平台简介
- 资源使用率优化背景介绍
- **提升模型推理性能**
 - Intel MKL库应用
 - **Intel OpenVINO推理引擎集成**
- **优化模型推理资源分配和调度**
 - TensorFlow模型混合部署
 - GPU虚拟化技术应用
 - 模型推理资源监控告警
- 总结

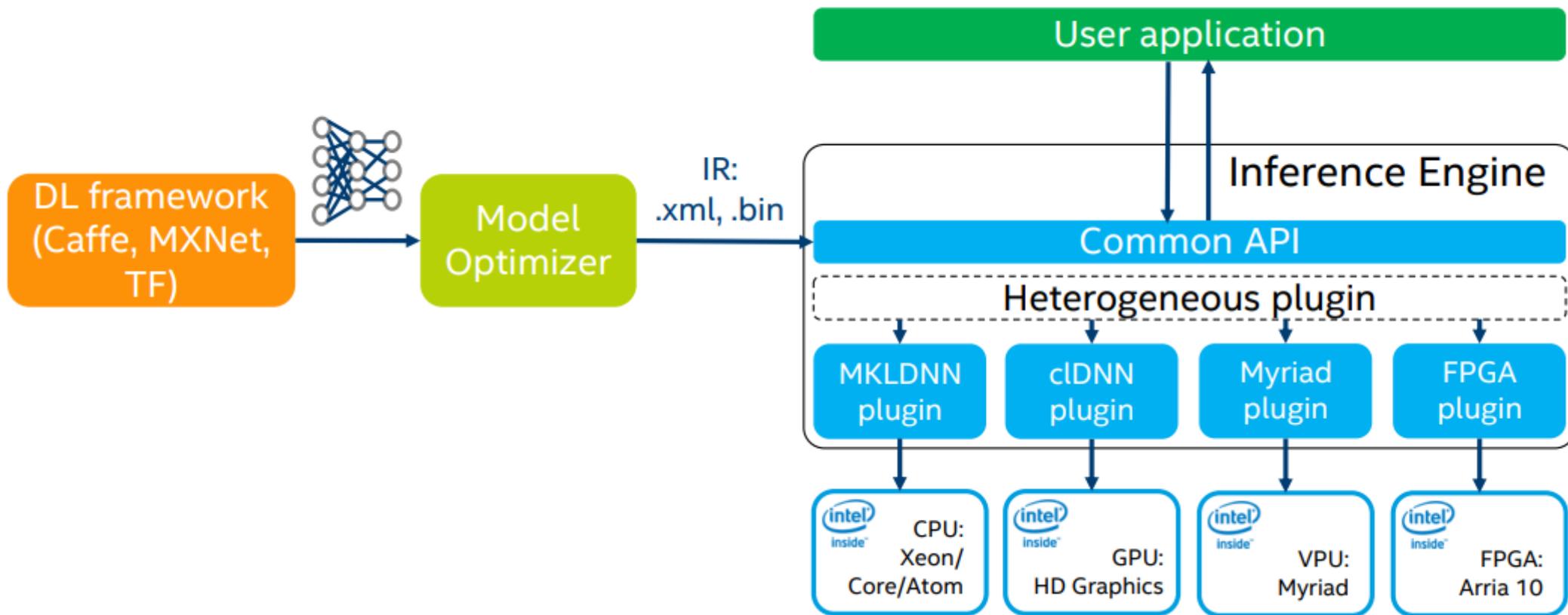


OpenVINO ToolKit是英特尔发布的一套深度学习推断引擎，支持各种网络框架，支持100多种网络训练导出的模型

工具包的主要特点主要包括：

- 在Intel平台上提升计算机视觉相关深度学习性能达19倍以上
- 解除CNN-based的网络在边缘设备的性能瓶颈
- 对OpenCV，OpenXV*视觉库的传统API实现加速与优化
- 基于通用API接口在CPU、GPU、FPGA等设备上运行





- 深度学习平台简介
- 资源使用率优化背景介绍
- 提升模型推理性能
 - Intel MKL库应用
 - Intel OpenVINO推理引擎集成
- 优化模型推理资源分配和调度
 - TensorFlow模型混合部署
 - GPU虚拟化技术应用
 - 模型推理资源监控告警
- 总结



存在的问题



在线推理任务流量较小，
存在GPU资源浪费

资源浪费

开发实验/训练/推理任务
极限使用情况仍无法
占满一张GPU卡



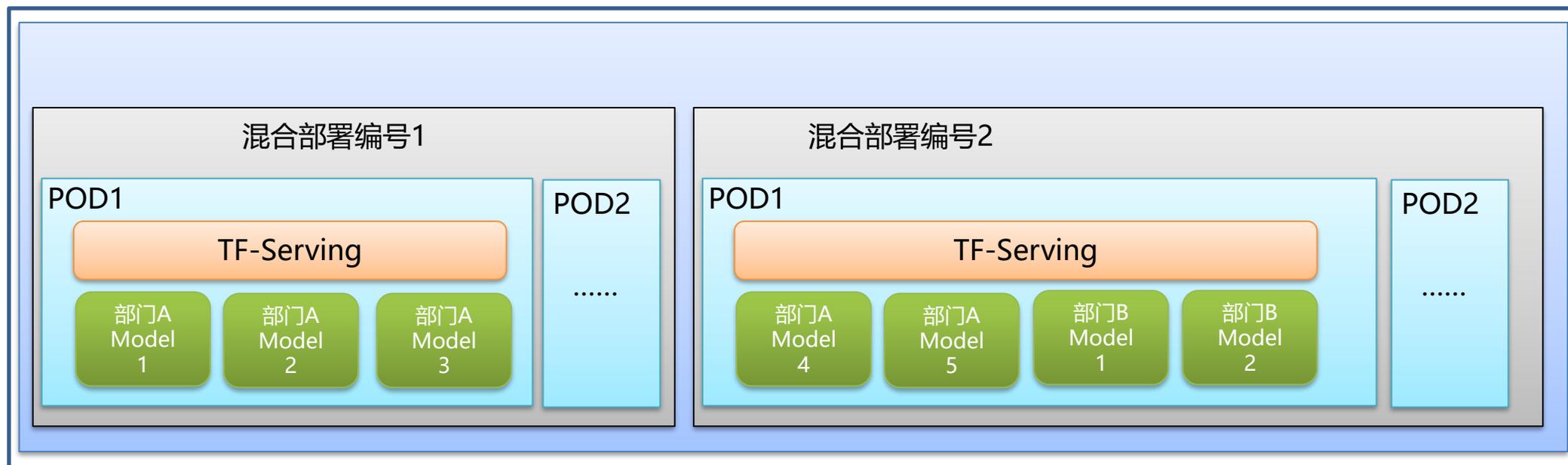
- 应用场景：模型需要使用GPU但流量小，GPU使用率低
- TensorFlow-Serving支持多模型部署
- Kubernetes实现混合部署资源调度

```
model_config_list:{
  config:{
    name:"model1",
    base_path:"/models/multiModel/model1",
    model_platform:"tensorflow"
  },
  config:{
    name:"model2",
    base_path:"/models/multiModel/model2",
    model_platform:"tensorflow"
  },
  config:{
    name:"model3",
    base_path:"/models/multiModel/model3",
    model_platform:"tensorflow"
  }
}
```



TensorFlow模型混合部署

推理请求接入RPC服务(请求转发+负载均衡)



Kubernetes

- 平台统一进行GPU资源调度和部署
- 混合部署时申请模型所需要的GPU资源 0.01~0.5

资源配置

是否需要GPU资源:

GPU

GPU型号:

P40

模型部署方式:

单卡混合部署

* 申请GPU资源:

0.05

高级设置

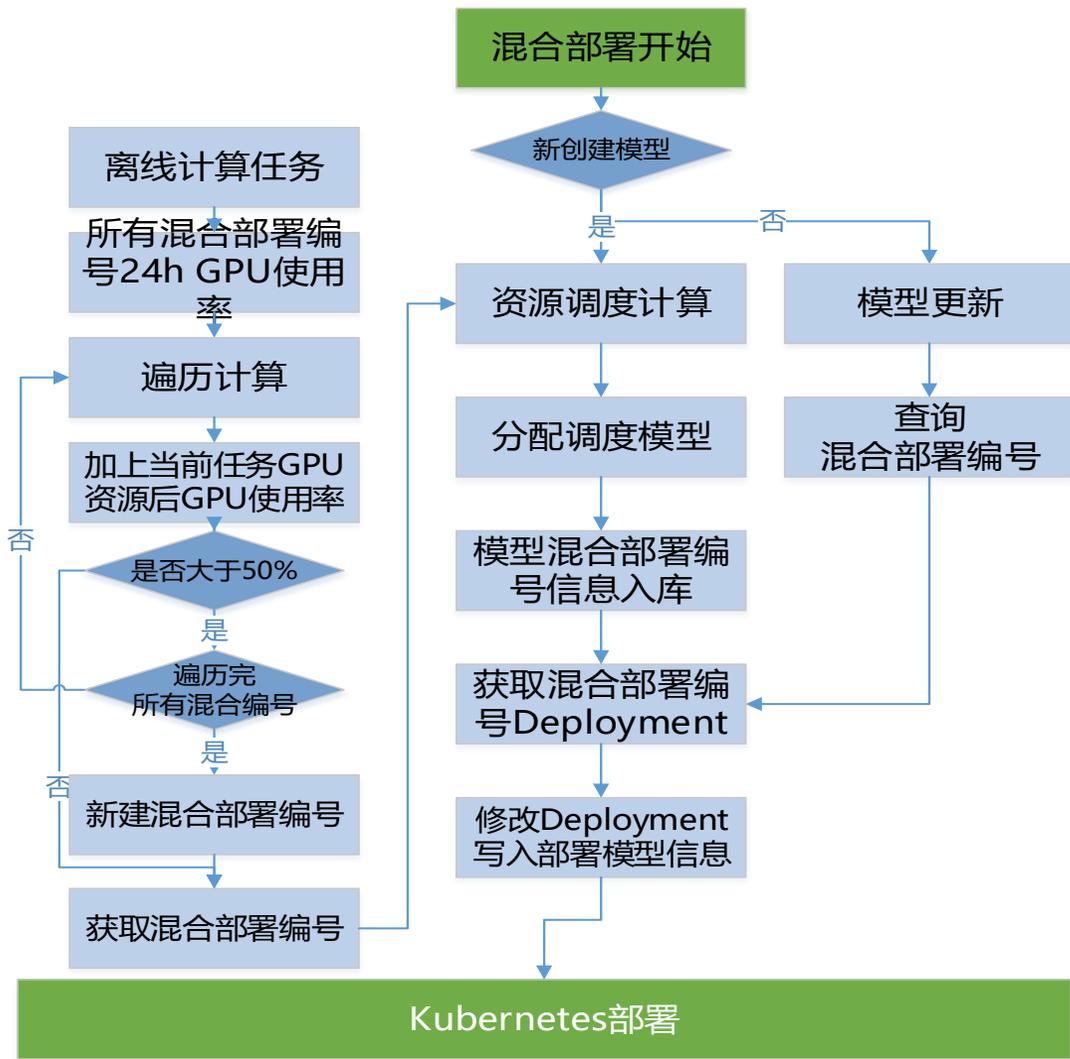
单卡独立部署

单卡混合部署

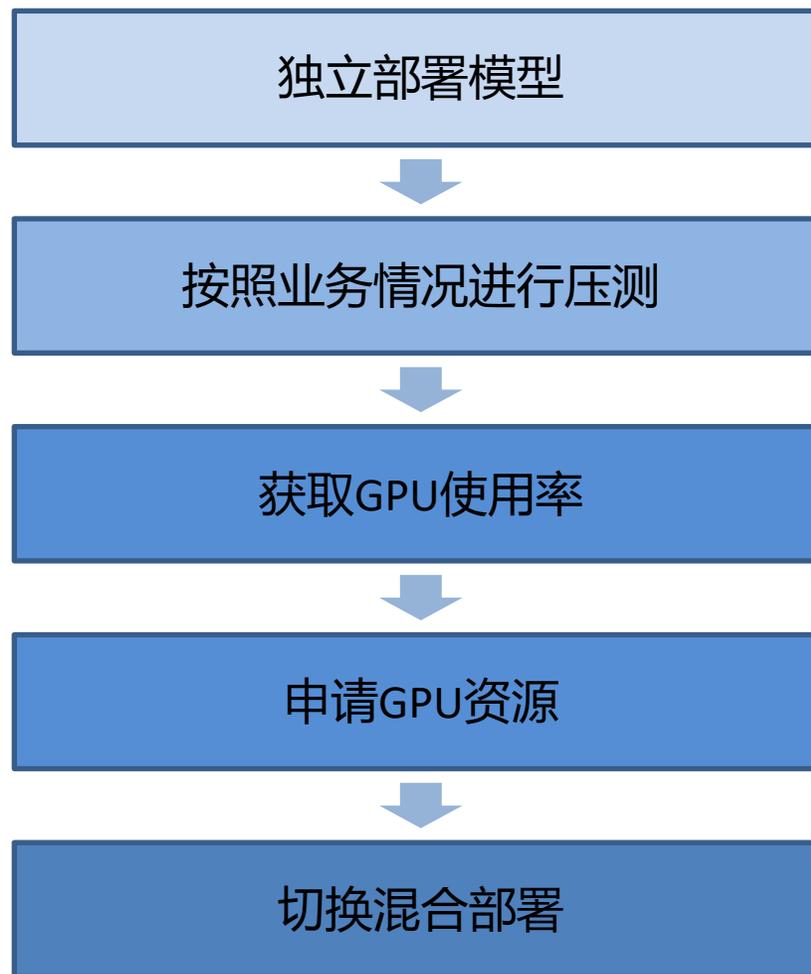


TensorFlow模型混合部署

资源统一调度



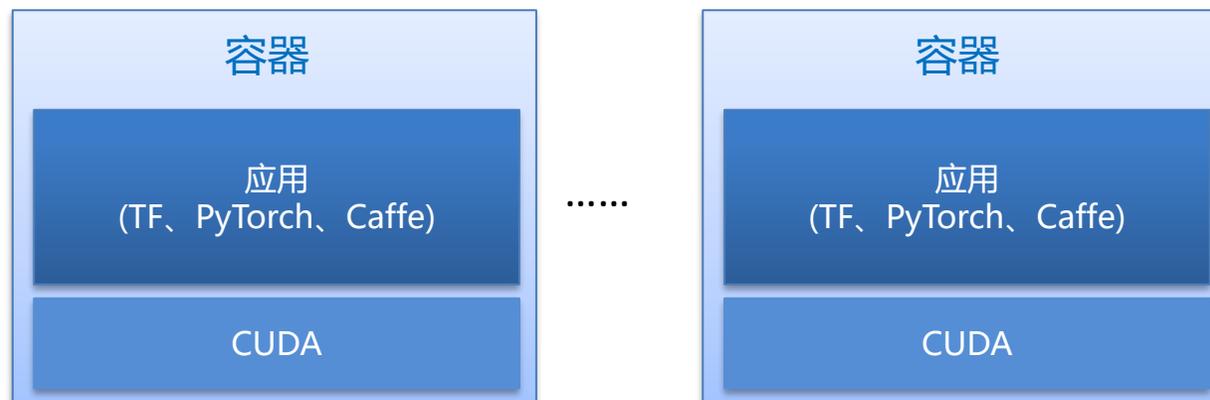
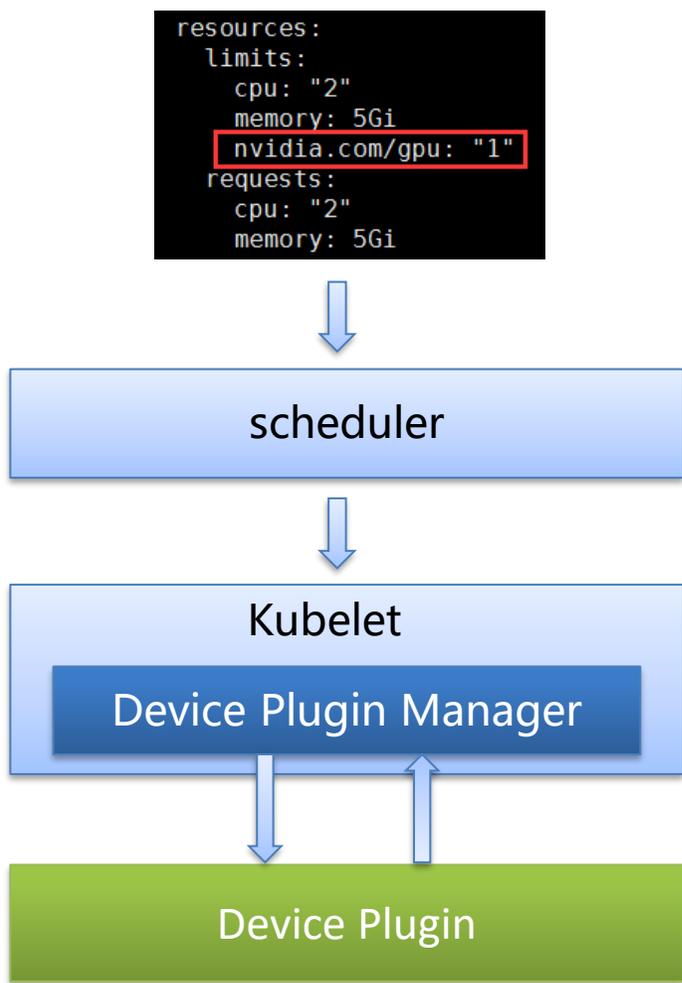
混合部署操作流程



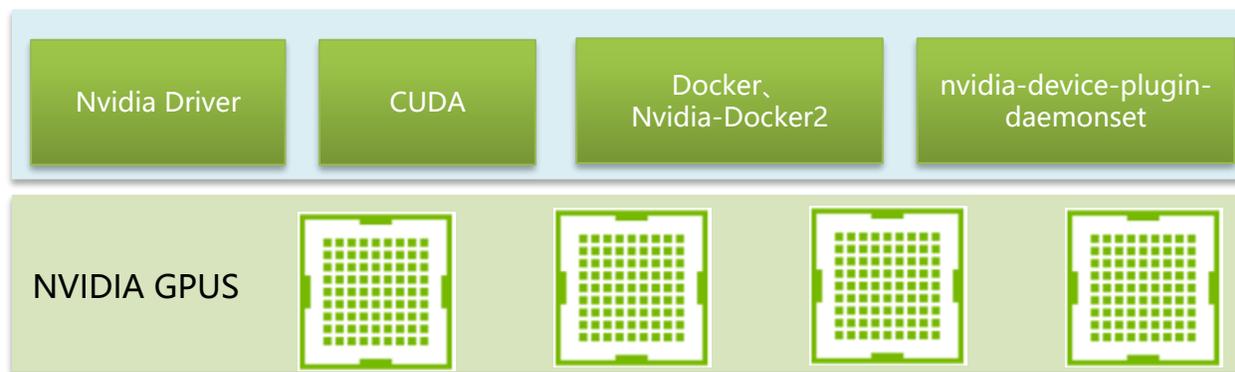
- 深度学习平台简介
- 资源使用率优化背景介绍
- 提升模型推理性能
 - Intel MKL库应用
 - Intel OpenVINO推理引擎集成
- **优化模型推理资源分配和调度**
 - TensorFlow模型混合部署
 - **GPU虚拟化技术应用**
 - 模型推理资源监控告警
- 总结



GPU容器调度和运行



物理机



存在的问题

Kubernetes GPU只能按整数调度



```
resources:
  limits:
    cpu: "6"
    memory: 10Gi
    nvidia.com/gpu: "1"
  requests:
    cpu: "6"
    memory: 10Gi
```



TensorFlow模型混合部署

用户自定义模型

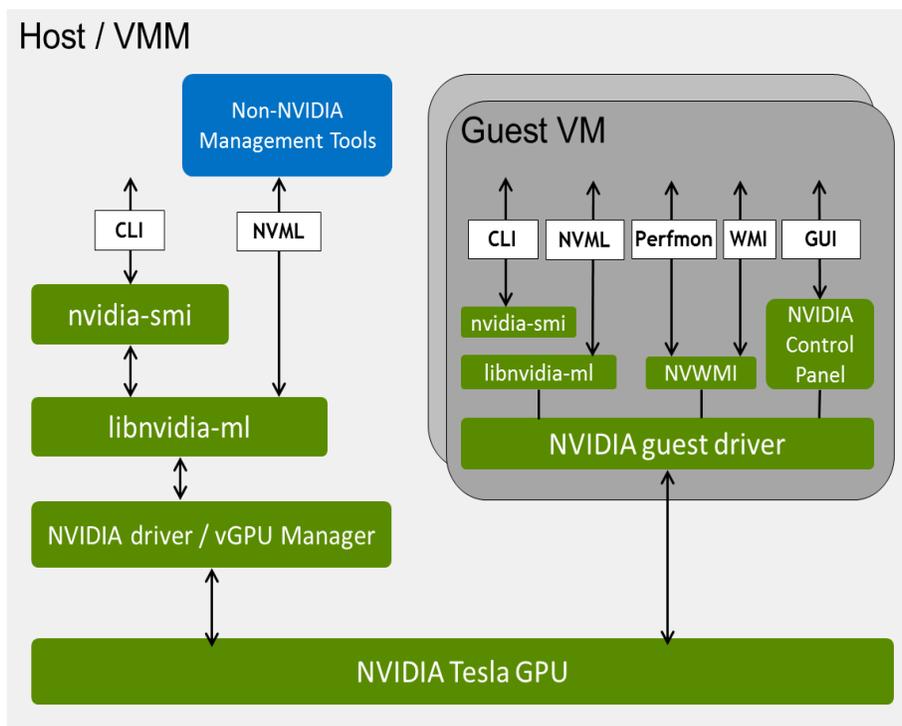
Caffe

PYTORCH

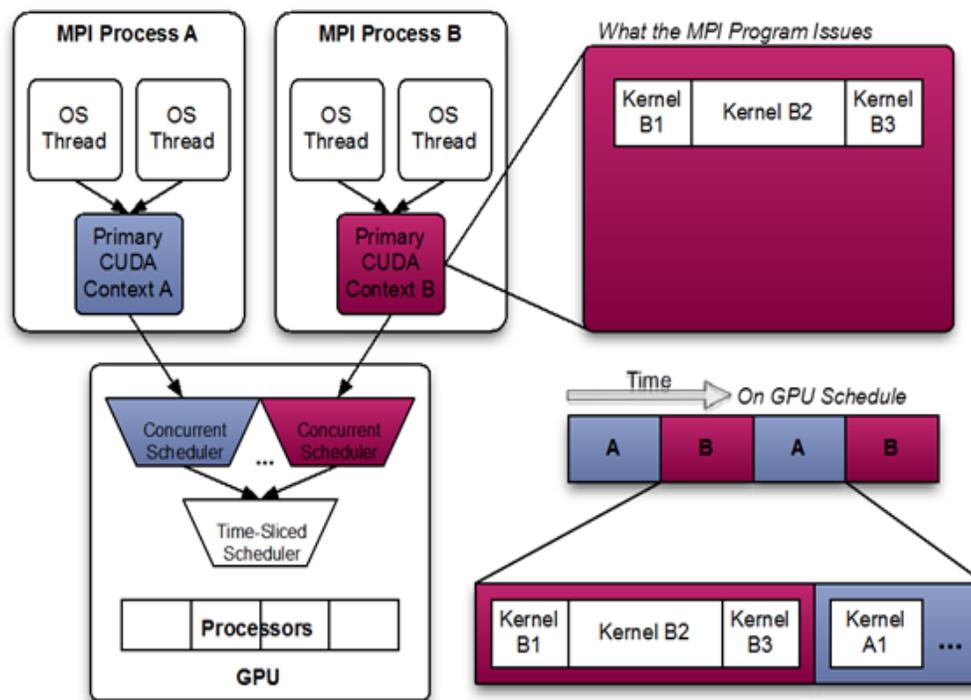


虚拟化方式	简介	优点	缺点
API转发	接收VM的显示请求，并转发给显卡驱动	VM侧实现简单 多个VM比较容易共享显卡资源	VM不具备真正的显卡，无法使用硬件加速能力，部分程序会存在兼容性问题
设备仿真	为客户机提供一个仿真虚拟的GPU	实现简单	模拟效率低
显卡透传	通过IO映射方式，将显卡分配给特定的VM	得到显卡的VM，表现近似真实主机，具备良好的显示性能和兼容性	显卡无法在多个VM间共享
全虚拟化	显卡驱动提供动态资源调度层，动态地分配显卡资源到各个VM	良好的性能 优秀的兼容性 显卡可在多个VM间共享	

GPU虚拟化-Nvidia



Nvidia Grid



Nvidia MPS



Quadro vDWS :

适用于专业级图形应用程序；内置 NVIDIA Quadro 驱动程序。虚拟工作站可通过数据中心提供GPU资源让用户随时随地在任何设备上安全访问数据，用户不再受物理位置的限制，通过数据中心虚拟化应用程序为终端用户如建筑师、工程师和设计师提供专业工作站级别的用户体验。常见应用程序有：Ansys Discovery Live、ESRI ArcGIS、Autodesk AutoCAD、Autodesk Maya、Autodesk Revit、CATIA、Petrel、Siemens NX、SOLIDWORKS等。

GRID vPC :

适用于拥有标准 PC 应用程序、浏览器和多媒体的虚拟桌面。

GRID vApps :

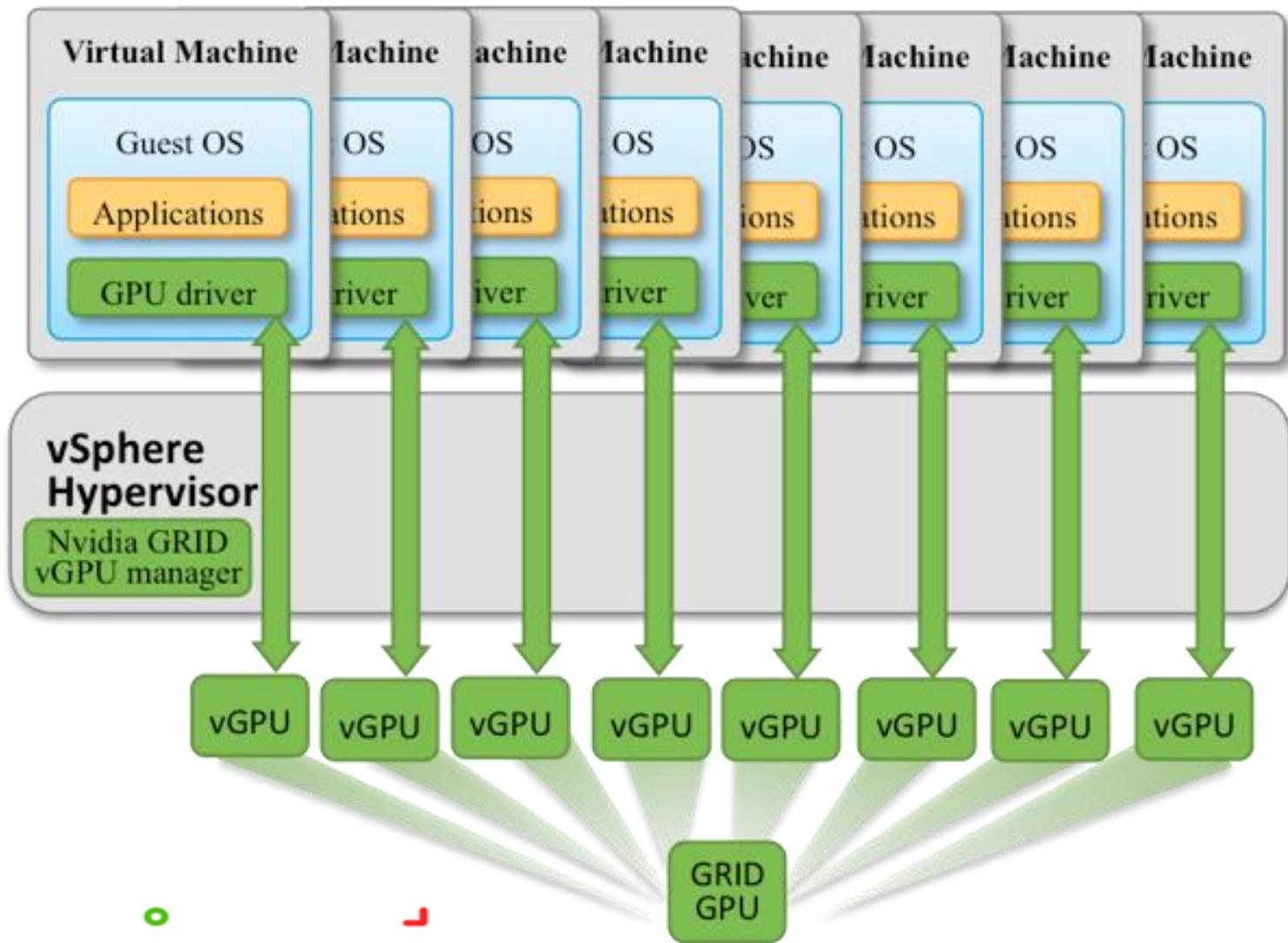
与 Citrix XenApp 或其他 RDSH 解决方案配合使用，例如 VMware Horizon 应用程序

vComputeServer :

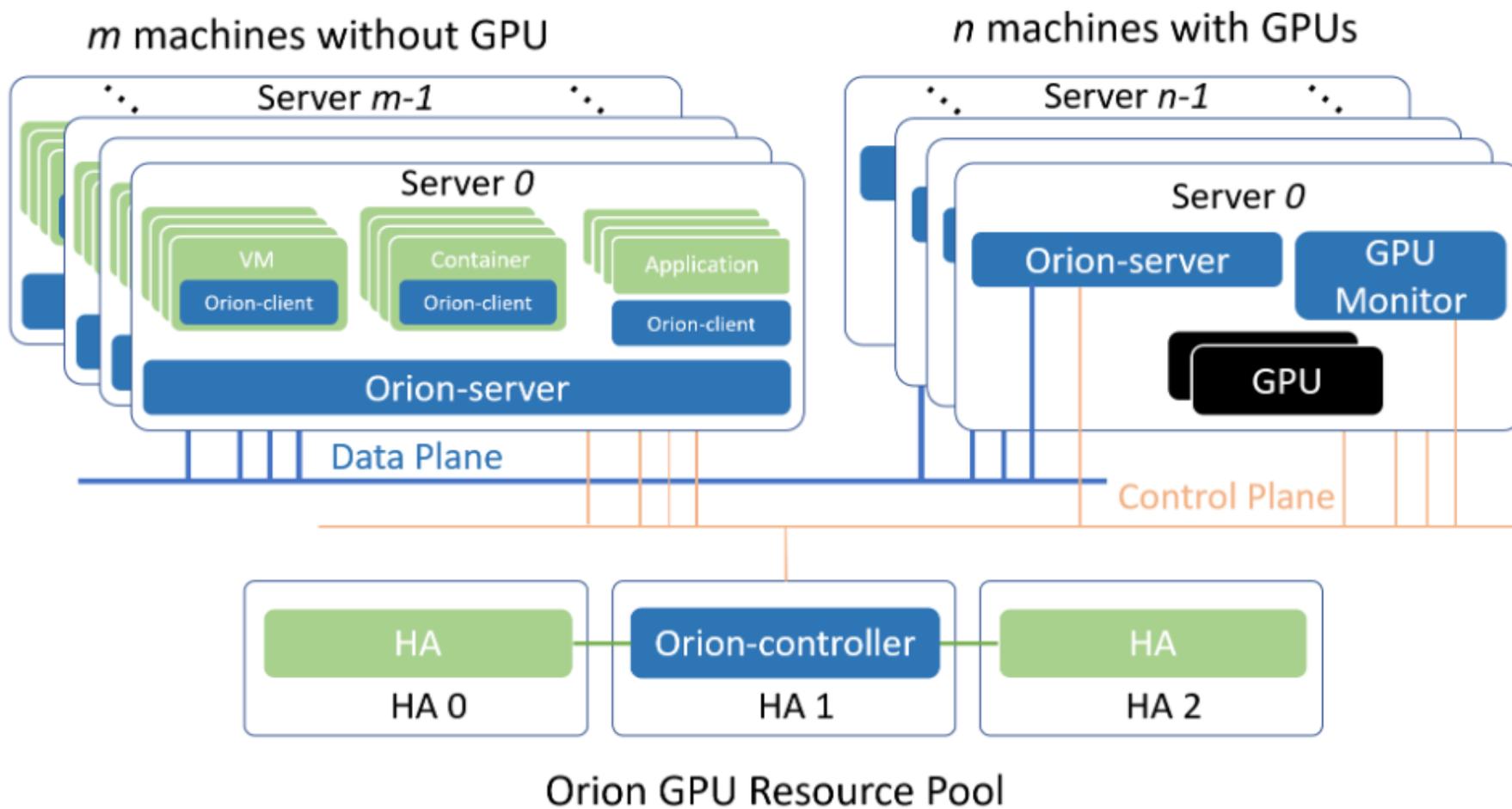
vComputeServer 让数据中心管理员可以在虚拟环境中在 GPU 服务器上运行 AI 工作负载，进一步提升了计算的安全性、利用率和可管理性。IT 管理员可以使用 VMware vSphere (包括 vCenter 和 vMotion) 等 hypervisor 虚拟化工具来管理所有数据中心应用，包括运行于 NVIDIA GPU 之上的 AI 应用。



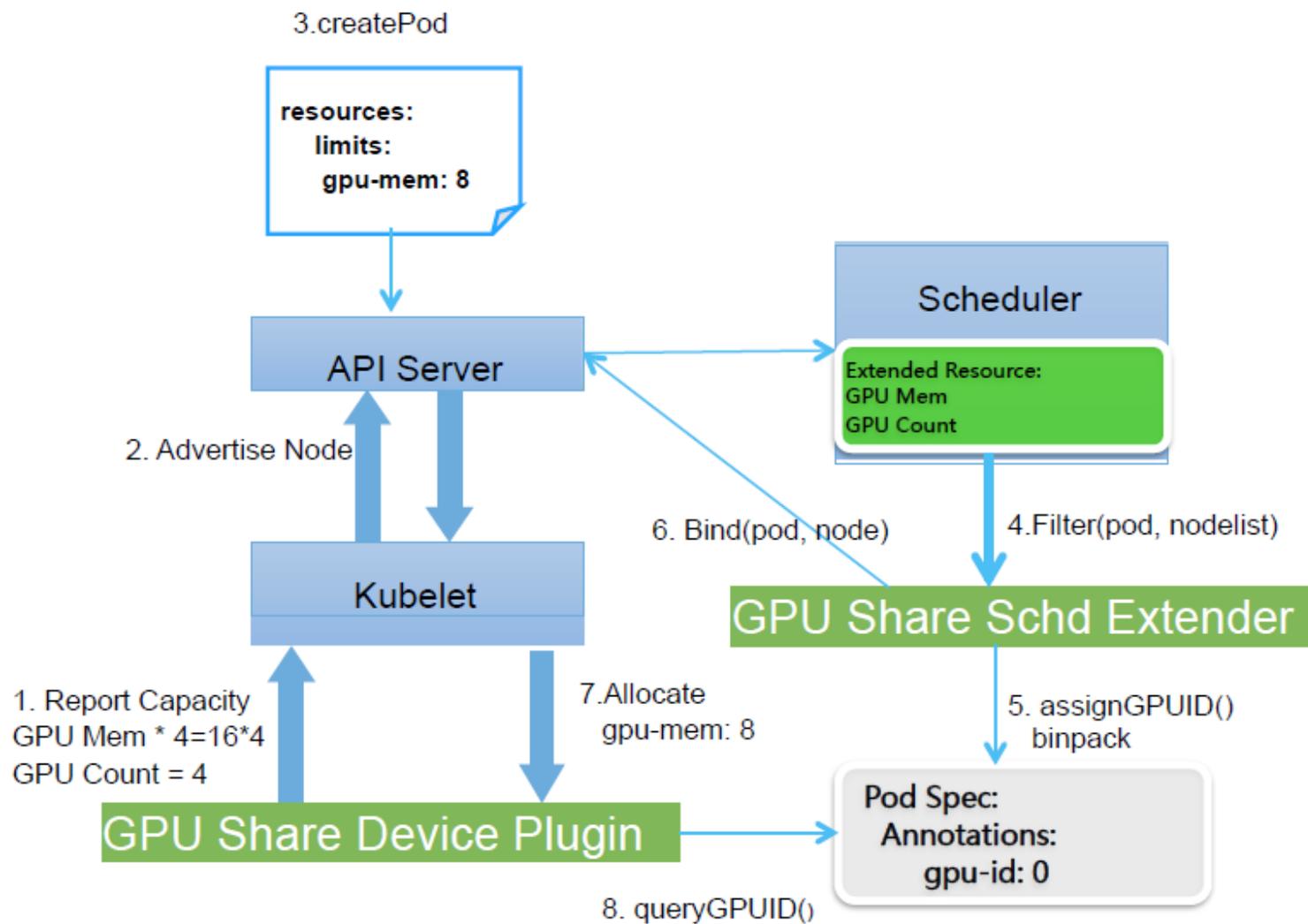
GPU虚拟化-Vmware vSphere



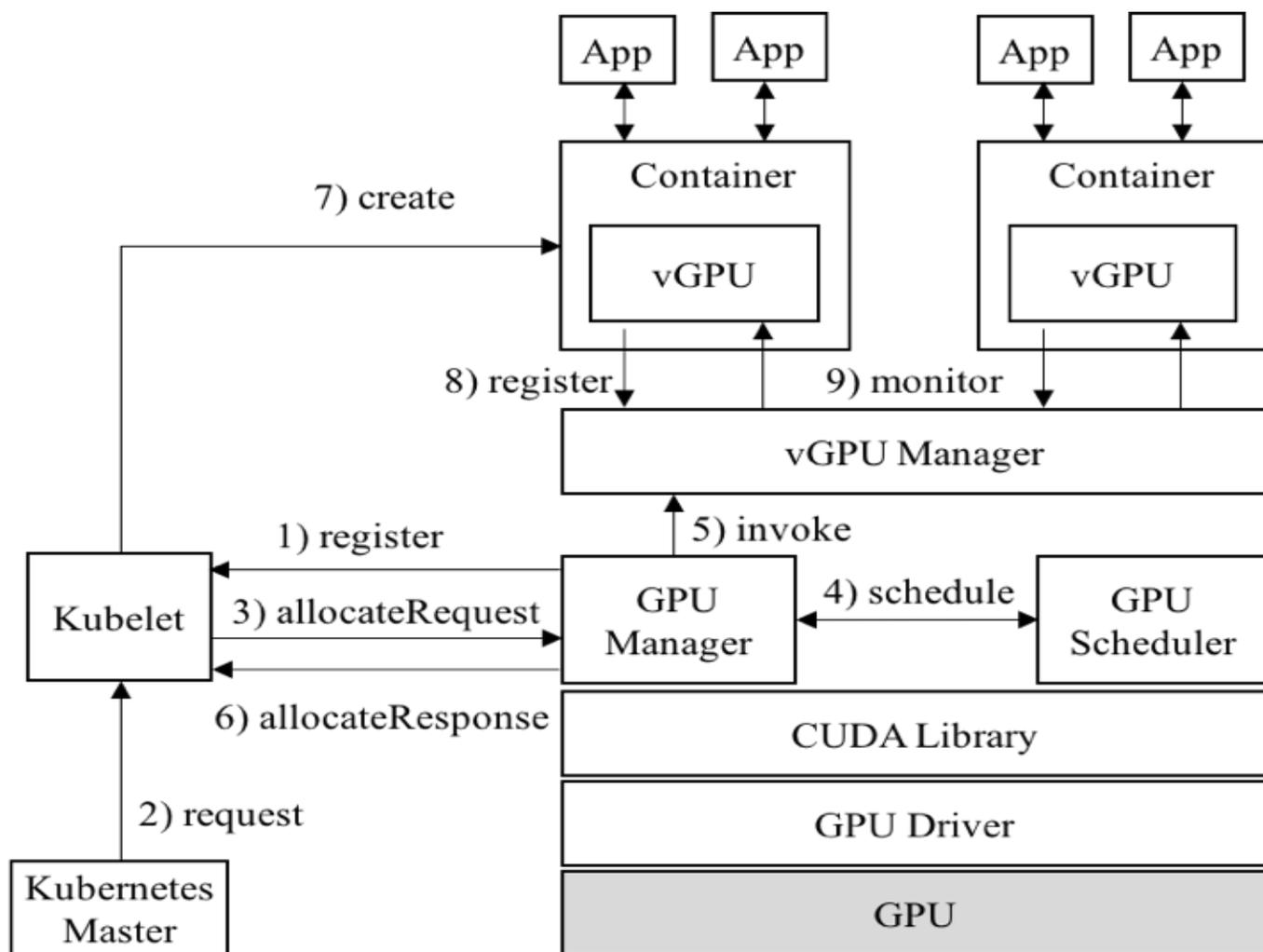
GPU虚拟化-OrionX



GPU虚拟化-GPU Sharing



GPU虚拟化-GPU Manager



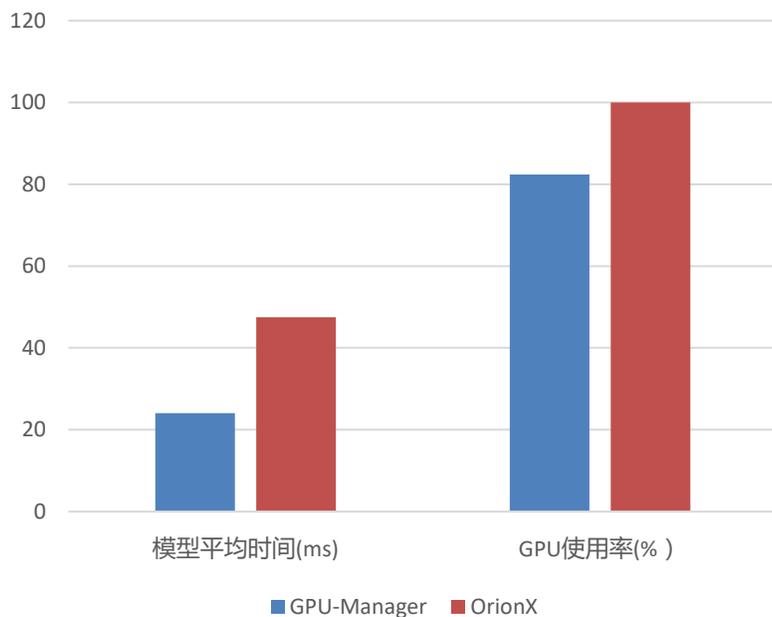
GPU虚拟化-方案对比

虚拟化方案	优点	缺点
Nvidia vComputeServer	GPU硬件兼容更好	收费较高，使用复杂
Vmware vSphere	GPU硬件兼容更好	收费较高，使用复杂
OrionX	提供资源监控，可以基于算力和显存分配资源	显卡无法在多个VM间共享
GPU Sharing	开源，使用简便	仅基于显存分配
GPU Manager	开源，可以基于算力和显存分配资源，使用简便	没有资源监控

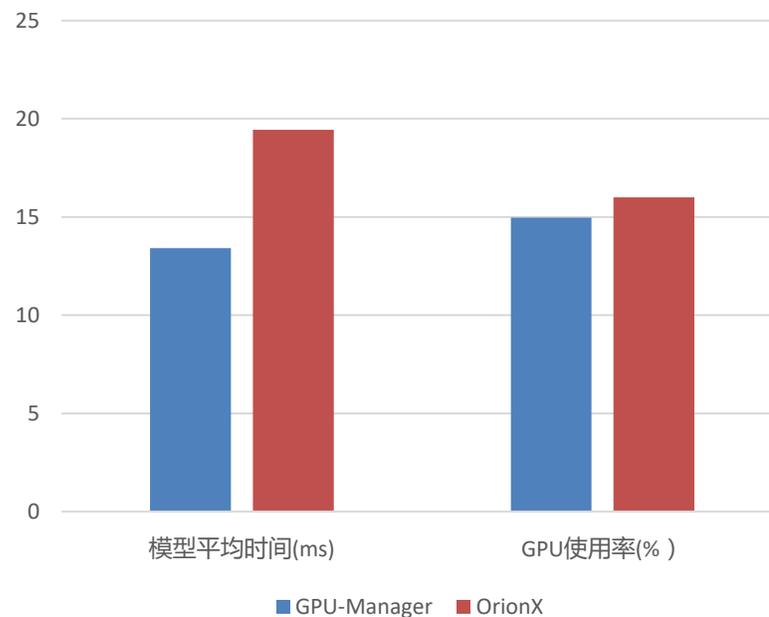


趋动科技的GPU虚拟化工具OrionX以及开源工具GPU-Manager进行了测试，主要测试GPU资源划分后的推理性能

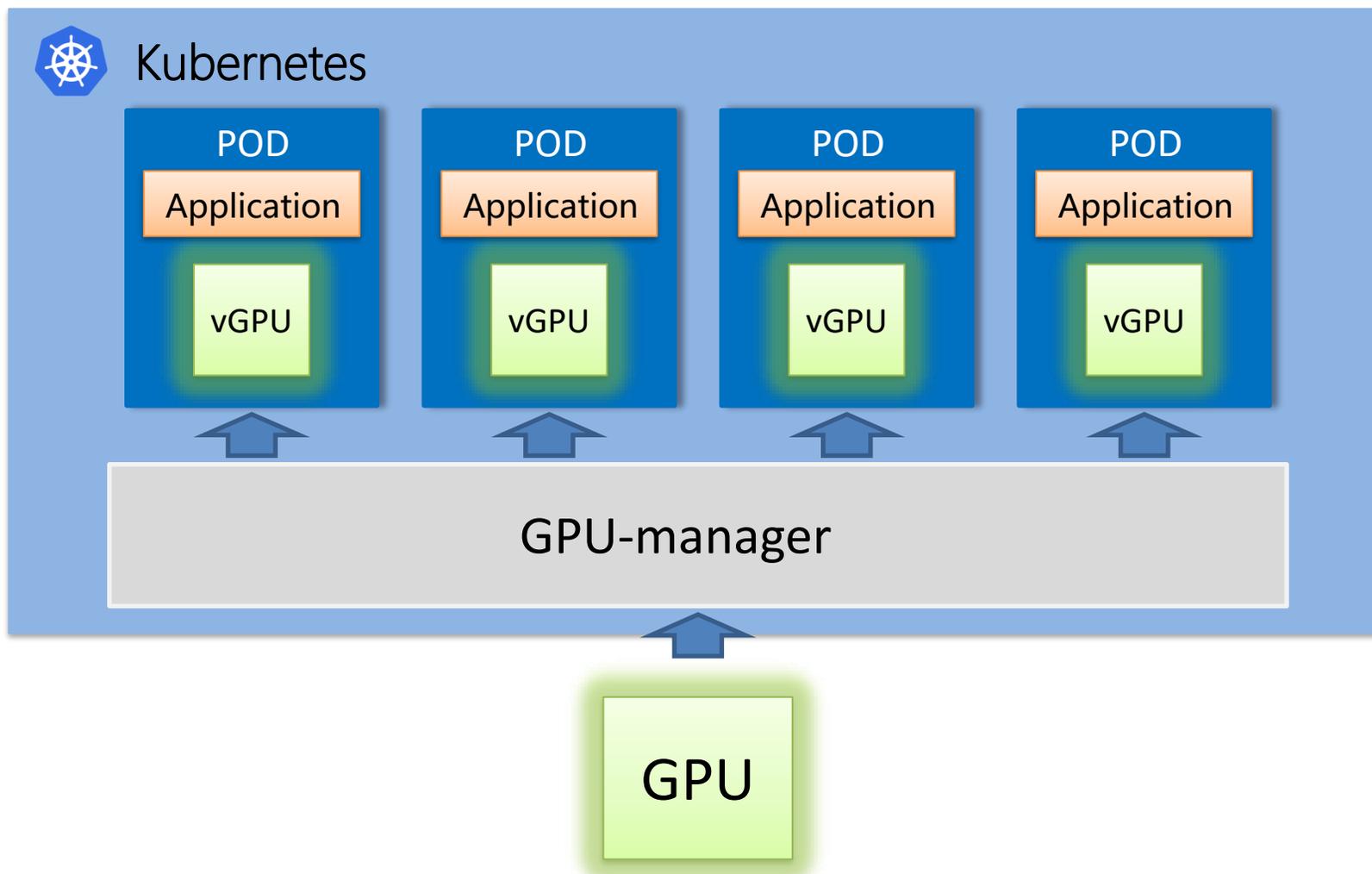
单卡部署4个推理节点



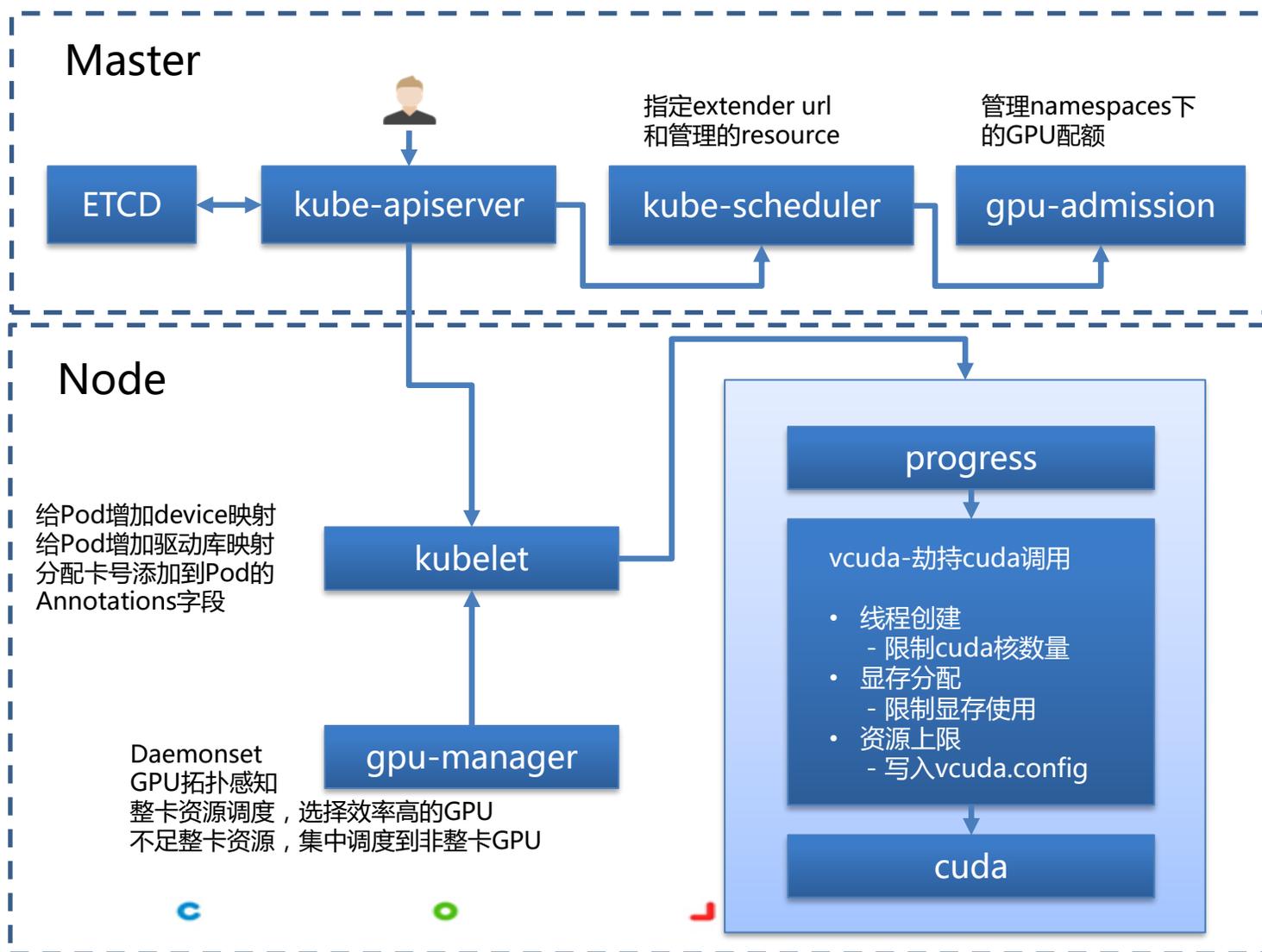
单卡部署1个推理节点



GPU虚拟化-整体架构



GPU虚拟化-线上部署

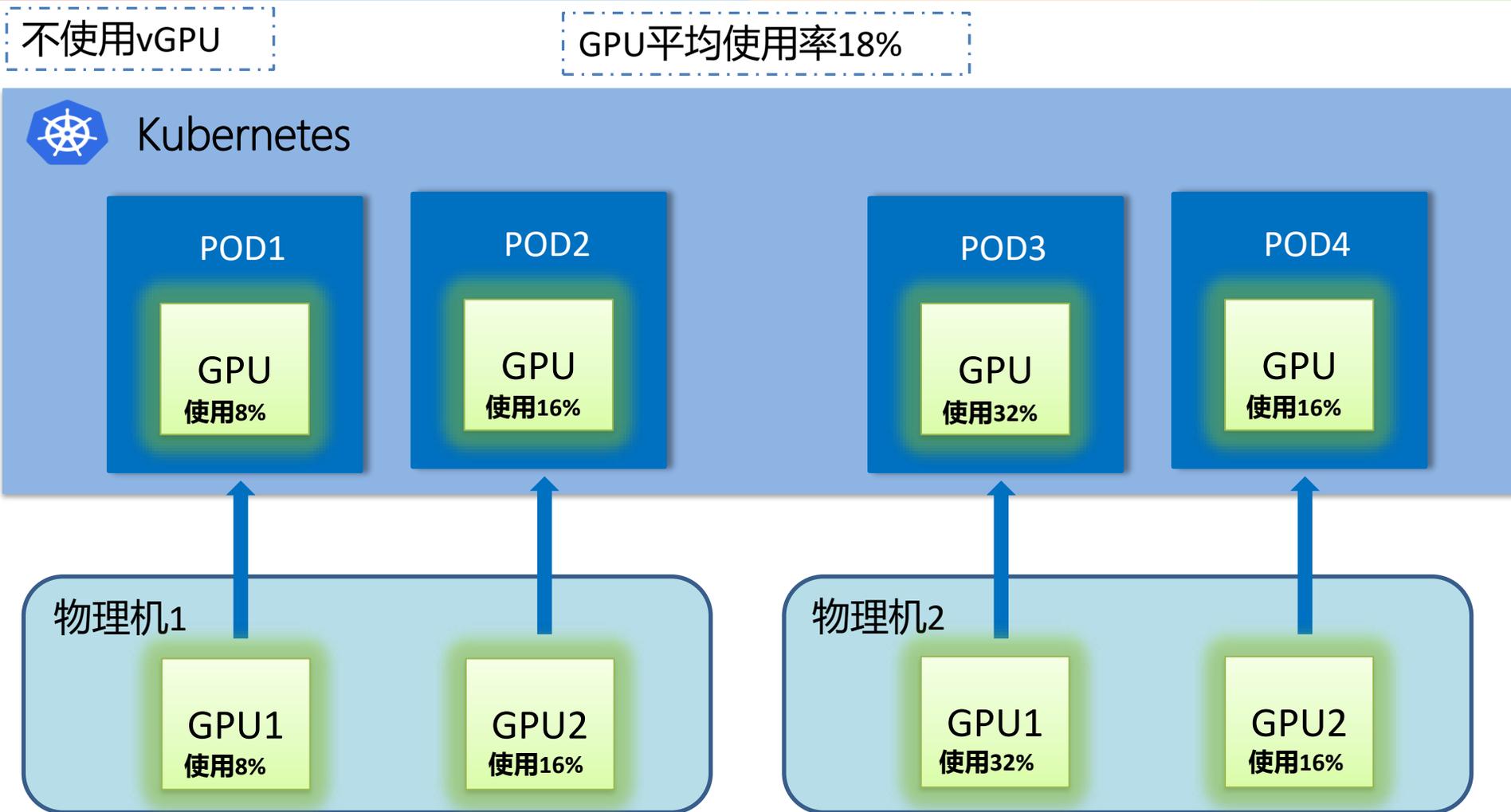


```
resources:
  limits:
    cpu: "4"
    memory: 20Gi
   .tencent.com/vcuda-core: "20"
   .tencent.com/vcuda-memory: "11"
  requests:
    cpu: "4"
    memory: 20Gi
   .tencent.com/vcuda-core: "20"
   .tencent.com/vcuda-memory: "11"
```

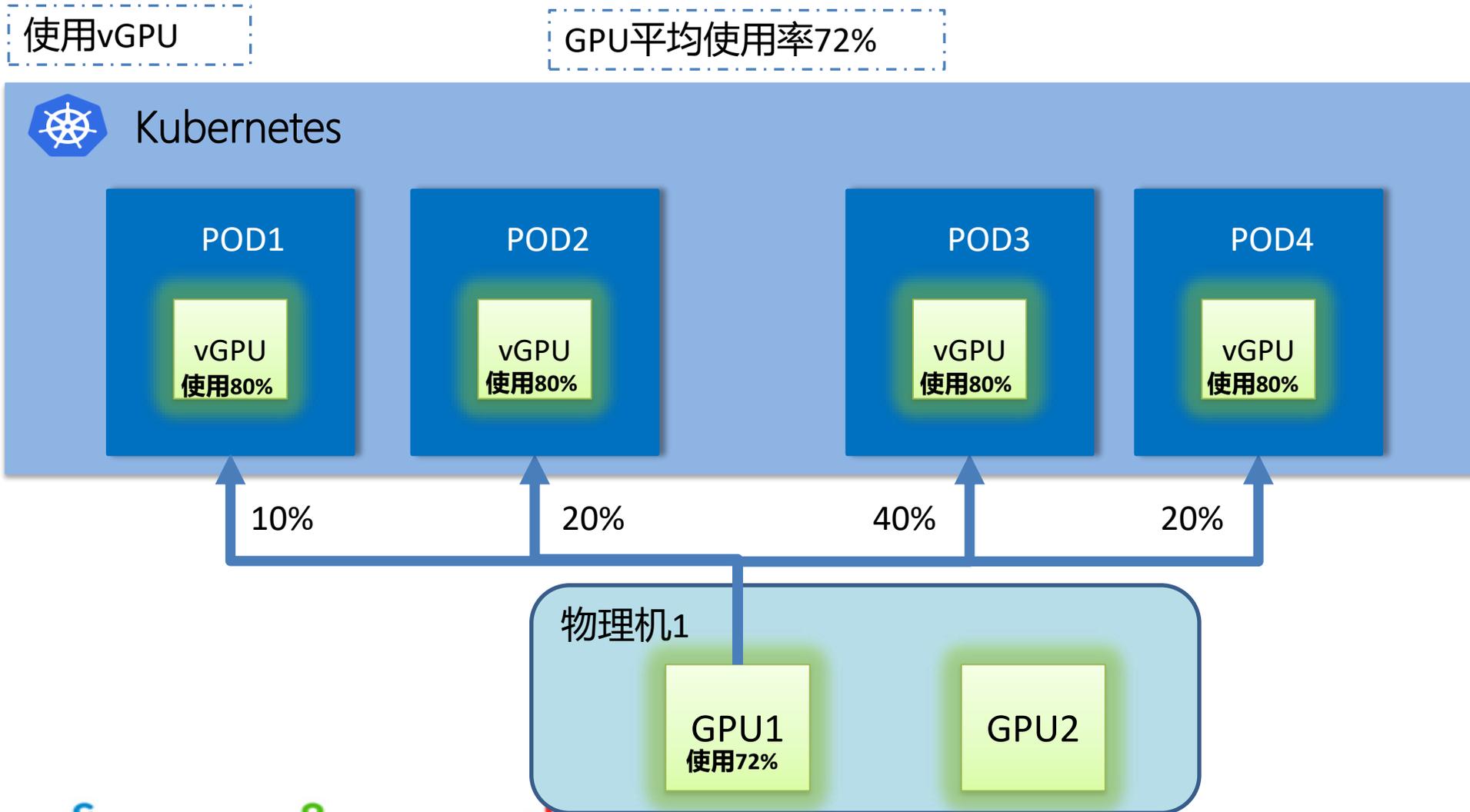
```
apiVersion: v1
kind: Pod
metadata:
  annotations:
   .tencent.com/gpu-assigned: "true"
   .tencent.com/predicate-gpu-idx-0: "1"
```

```
resources:
  limits:
    cpu: "4"
    memory: 20Gi
   .tencent.com/vcuda-core: "20"
   .tencent.com/vcuda-memory: "11"
  requests:
    cpu: "4"
    memory: 20Gi
   .tencent.com/vcuda-core: "20"
   .tencent.com/vcuda-memory: "11"
```

线上部署-示例



线上部署-示例



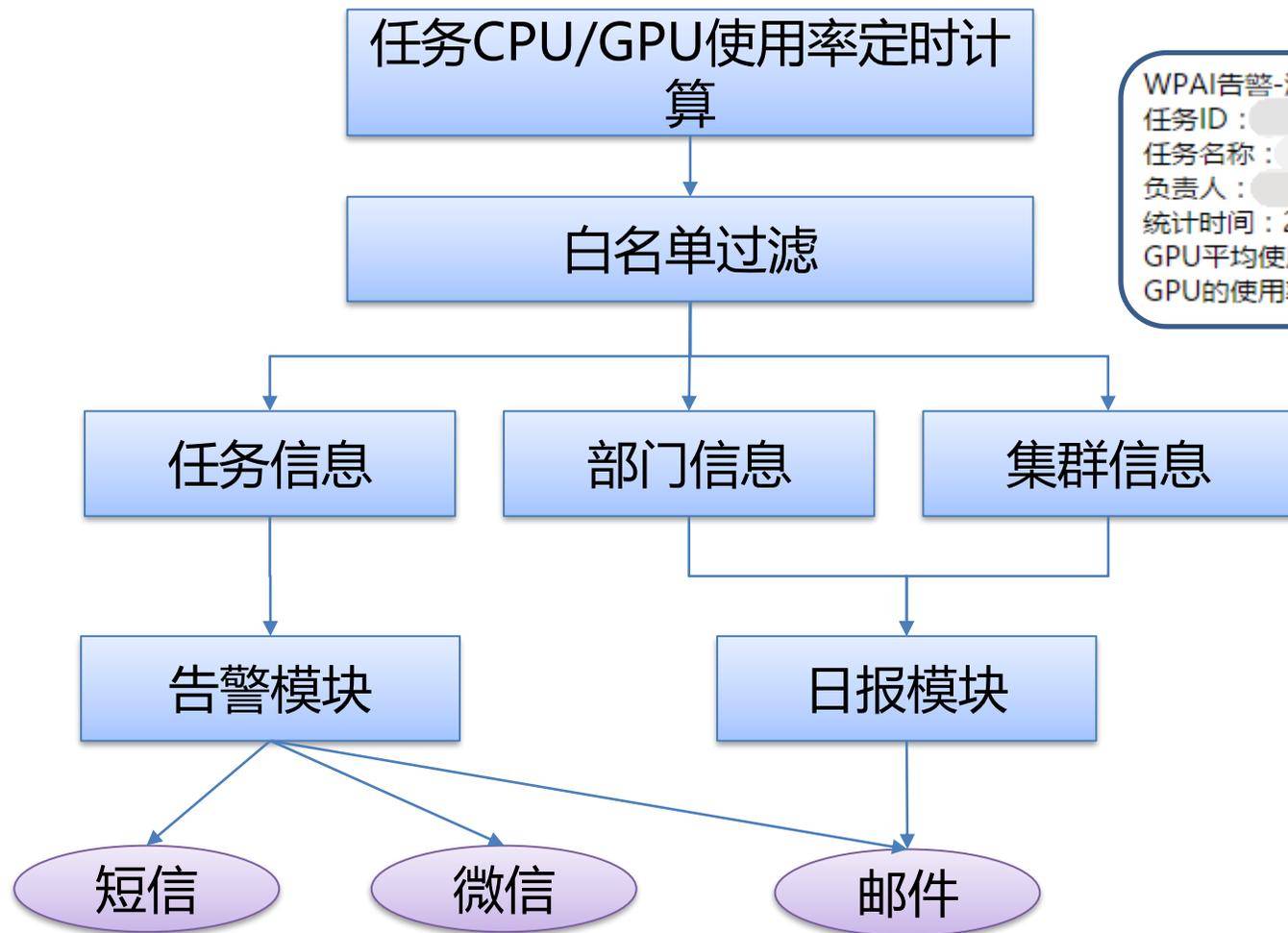
- 深度学习平台简介
- 资源使用率优化背景介绍
- 提升模型推理性能
 - Intel MKL库应用
 - Intel OpenVINO推理引擎集成
- **优化模型推理资源分配和调度**
 - TensorFlow模型混合部署
 - GPU虚拟化技术应用
 - **模型推理资源监控告警**
- 总结



问题：

- 任务资源配置后，用户不再进行关注
- 没有自动扩缩容和模型部署方式自动切换
- 线上业务流量减小，资源产生浪费





WPAI告警-深度学习在线预测任务资源配置不合理
任务ID: [模糊]
任务名称: [模糊]
负责人: [模糊]
统计时间: 2020-05-11
GPU平均使用率: 13.0%, 低于GPU告警阈值50%
GPU的使用率过低, 请重新配置资源或进行GPU混合部署



- 深度学习平台简介
- 资源使用率优化背景介绍
- 提升模型推理性能
 - Intel MKL库应用
 - Intel OpenVINO推理引擎集成
- 优化模型推理资源分配和调度
 - TensorFlow模型混合部署
 - GPU虚拟化技术应用
 - 模型推理资源监控告警
- 总结



- MKL及OpenVINO的使用，提高了单节点推理能力，减少任务节点部署
- TensorFlow混合部署实现了单Pod多模型部署
- GPU虚拟化实现了GPU细粒度调用，可以实现一张卡部署多个Pod
- 模型资源监控有助于及时发现不合理的资源配置，为重新配置提供合理的建议
- 平台模型推理占用GPU卡**减少37%**，在用GPU卡平均使用率**提升146%**



Thanks!

让生活更简单

