Project Supernova

高效、快捷通用ML推理服务设计

张华乔

huaqiaoz@vmware.com

vmware

Agenda

- Supernova 设计理念
- Supernova 体系结构简介
- 通用推理服务简介
- 推理服务的创建流程
- RESTful API示例

Supernova设计理念


降低各类机器学习推理toolkit使用门槛


基于RESTful API简洁易懂、零术语


操作步骤简单，5步完成一个数据推理任务


自身尺寸足够小，资源消耗足够低，用户无感知


基于容器化发布，易部署


节约企业人力开发运维成本

A44
A45
A46
A47

Slide 3

A44 为甚要减低推理toolkit使用门槛

Author, 12/13/2019

A45 配置环境总是一件麻烦的事，鬼知道会缺失什么依赖包，到头来没装好不说，还可能影响已有的环境，这在生产线上是非常危险的事情！！！

Author, 12/13/2019

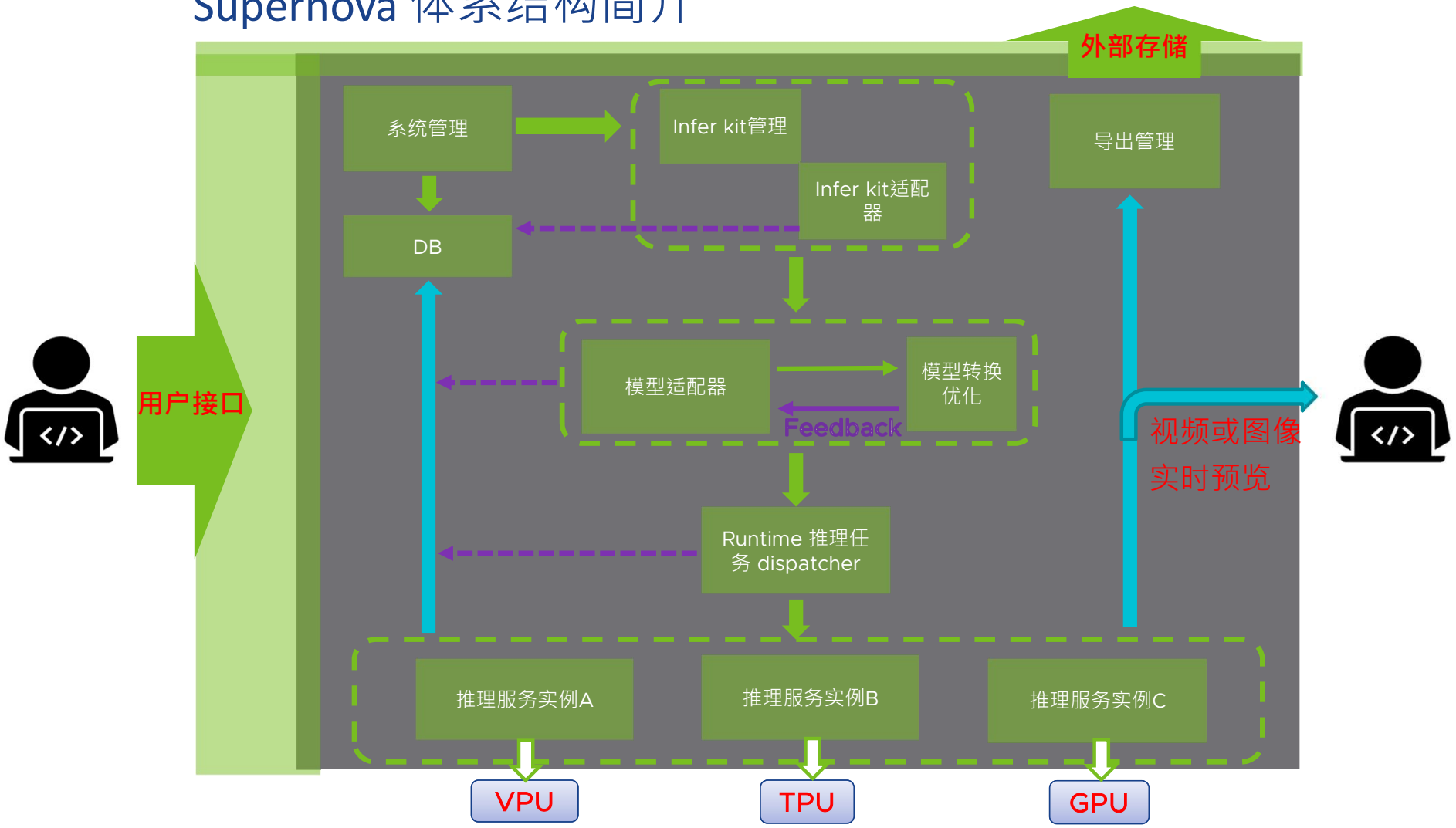
A46 爱要学习各类inference toolkit的SDK，各种不同的语言，配置环境等

Author, 12/13/2019

A47 所以我们不提供什么sdk让用户二次开发。

Author, 12/13/2019

Supernova 体系结构简介



通用推理服务介绍

```python
def image_preprocess(self, image,):

    blob = cv2.resize(image, (300, 300))
    blob = blob[np.newaxis, :, :, :]
    blob = blob.transpose((0, 3, 1, 2))
    return blob
```

```python
def image_preprocess(self, image):
    try:
        prepimg = cv2.resize(image, (64, 64))
    except:
        prepimg = np.full((64, 64, 3), 128)
    prepimg = prepimg[np.newaxis, :, :, :]
    prepimg = prepimg.transpose((0, 3, 1, 2))
    return prepimg
```

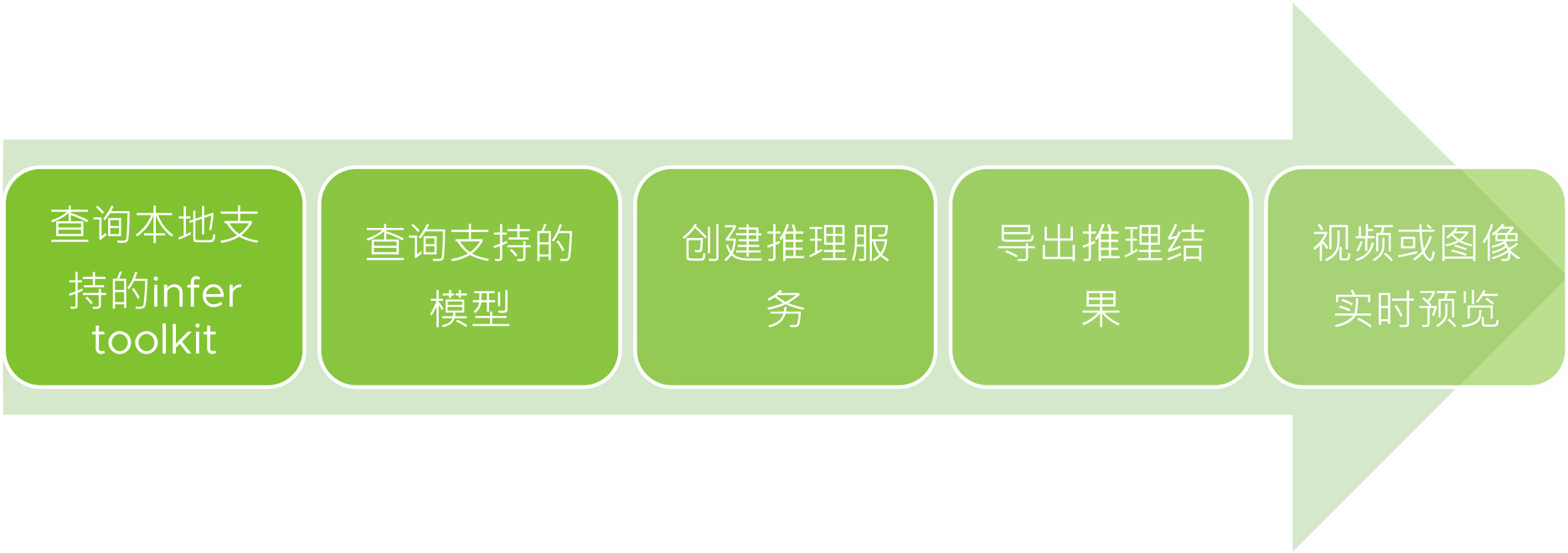
模型输入源的差异性

```python
for object_info in object_infos:
        if object_info[2] == 0.0:
            break
        if (not np.isfinite(object_info[0]) or
            not np.isfinite(object_info[1]) or
            not np.isfinite(object_info[6])):
            continue
        min_score_percent = 60
        source_image_width = width
        source_image_height = height
        percentage = int(object_info[2] * 100)
```

```python
for output in outputs.values():
        objects = ParseYOLOV3Output(output, new_h, new_w,
                                    camera_height,
                                    camera_width,
                                    0.4, objects)
        objlen = len(objects)
        for i in range(objlen):
            if (objects[i].confidence == 0.0):
                continue
            for j in range(i + 1, objlen):
                if (IntersectionOverUnion(objects[i], objects[j]) >= 0.4):
                    if objects[i].confidence < objects[j].confidence:
                        objects[i], objects[j] = objects[j], objects[i]
                    objects[j].confidence = 0.0
```

推理结果处理差异性

推理服务创建流程



RESTful API操作示例

POST /api/v1/task

```
{
    "name":"my-inference",
    "device":{
        "name":"dev-intel-vpu-1"
    },
    "auto":true,
    "inputDataSource":"0",
    "inferenceModel":{
        "name":"object_detect_classification",
        "type":"Classification"
    },
    "inference":{
        "name":""
    },
    "inferKit":{
        "name":"OpenVINO",
        "version":"0.1.0"
    }
}
```

创建一个推理任务：

1. 选择本地支持的infer kit
2. 选择本地已有model
3. 选择硬件accelerator
4. 指定数据源（有界源/无界源）

Thank You

vmware®