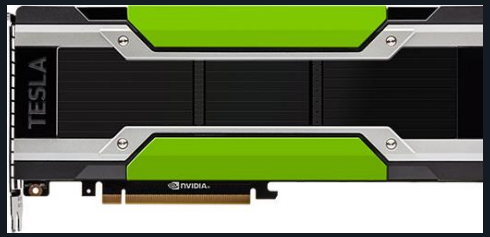
阿里云异构计算平台

——加速AI视觉智能创新

刘令飞　阿里云GPU云计算专家

异构计算

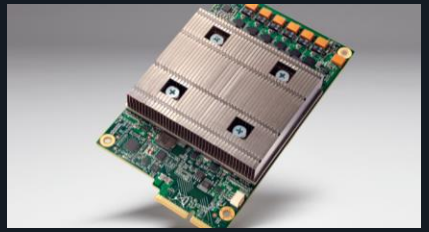
CPU



GPU

GPU

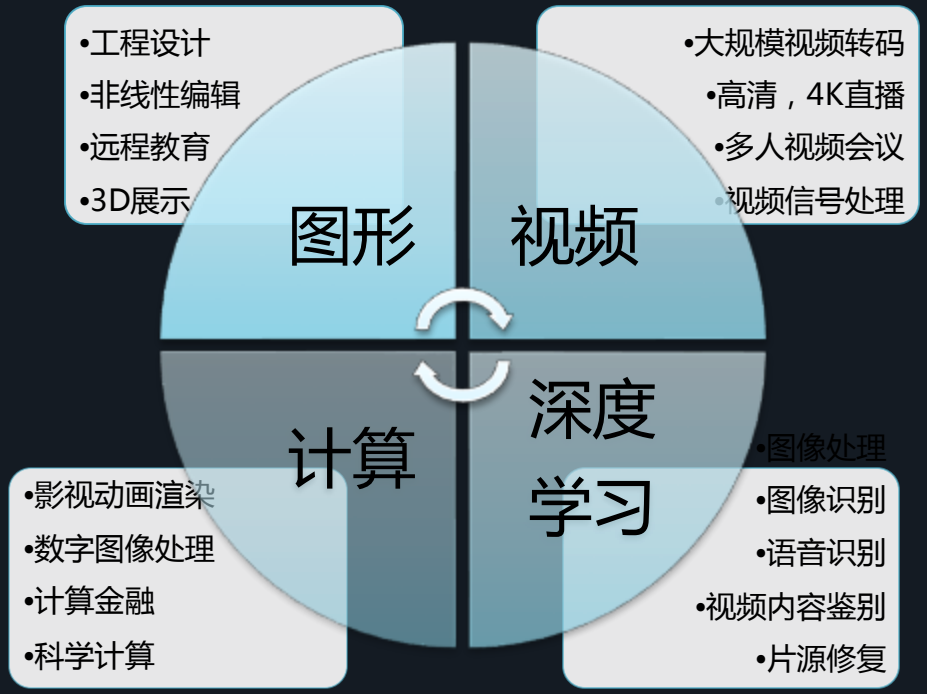
FPGA



ASIC

异构计算主要是指使用不同类型指令集和体系架构的计算单元组成系统的计算方式，常见的计算单元类别包括CPU、GPU、FPG、ASIC等。

GPU适用的领域及业务场景



3D渲染

- Direct X
- OpenGL
- Vulcan

视频编解码

- DXVA/LibVA
- NVEnc/VCE

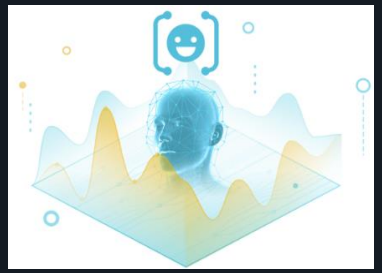
计算

- OpenCL
- CUDA

GPU的特点：实时高速、并行计算、浮点计算能力强

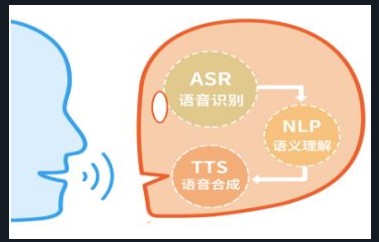
AI深度学习催生GPU服务需求

人脸识别

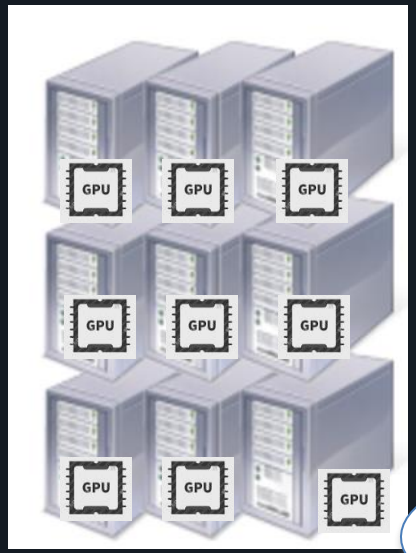


OCR文字识别

语音识别与合成



自然语言理解与交互



GPU资源
如何快速
扩容？

哪
的
GPU资源？

GPU如何满
足训练和推
理的需求？

如何多
地域线
上服务？

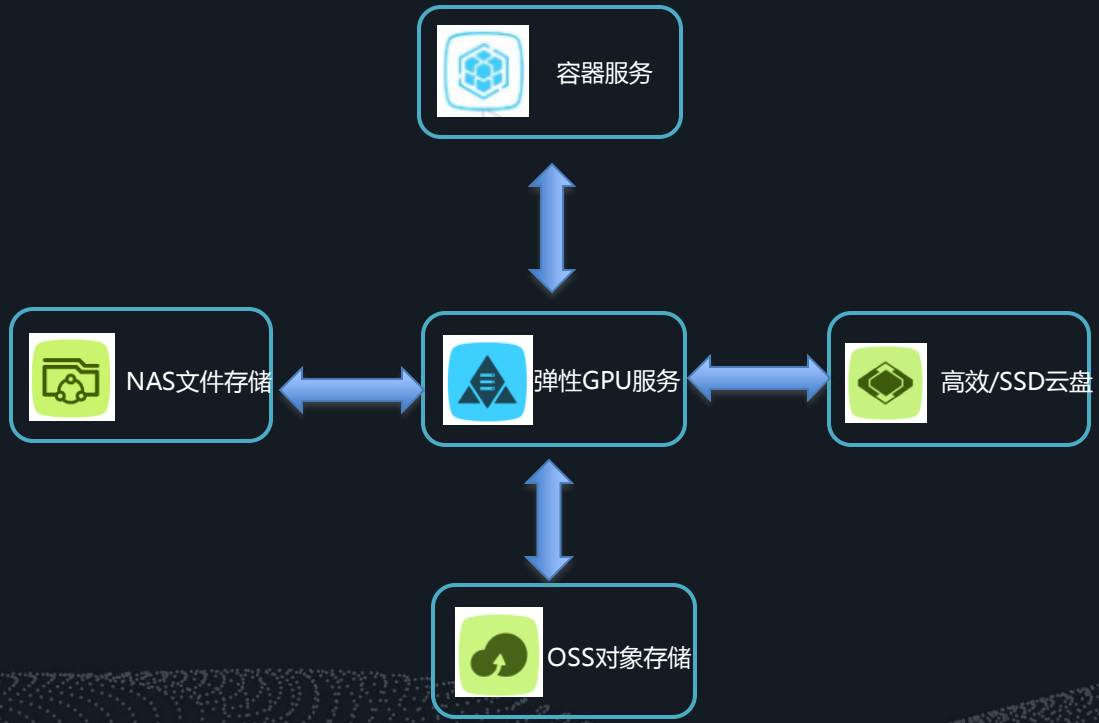
GPU还正
常工作吗？

弹性GPU服务（Elastic GPU Service – EGS）

EGS是基于GPU应用的计算服务，适用于深度学习、视频解码、图形渲染、科学计算等应用场景，具有实时高速，并行计算跟浮点计算能力强等特点。

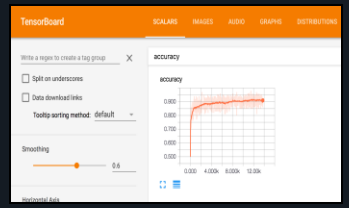
EGS具备与阿里云生态深度整合能力

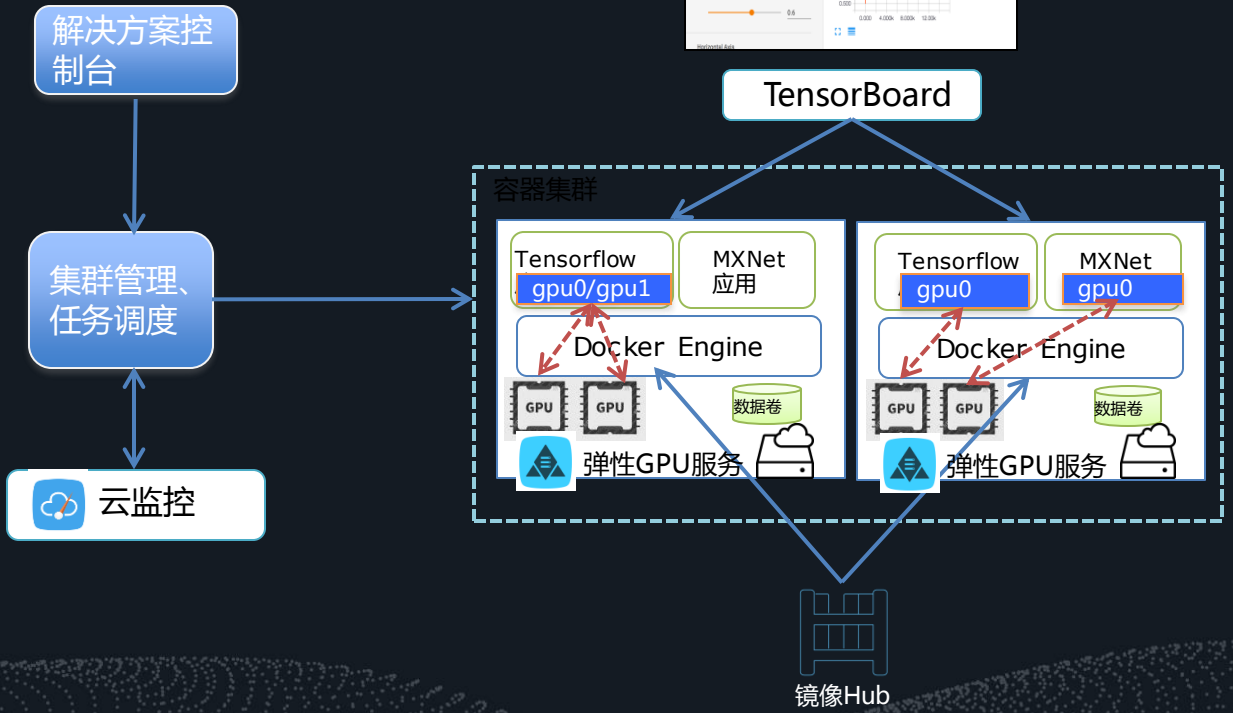
EGS具备与阿里云生态深度整合能力——容器为例



基于容器的弹性GPU服务一键式部署

- 一键部署集群
- 支持GPU资源调度
- 挂载共享存储
- 负载均衡
- 弹性伸缩
- CPU、GPU监控
- 日志管理





基于容器的EGS工作方式

	工作	EGS部署	EGS+容器部署
深度学习环境搭建	安装配置	Make，Bazel或者pip安装，或者通过容器镜像	无需安装，支持TensorFlow，Keras, MXNet
	分布式环境	通过SSH登录到每台机器上手工部署	一键完成整个集群的构建
	GPU资源调度	手动记录管理，整机分配，使用效率低	容器化隔离，自动化统一管理和度
数据准备	数据集存储、共享	手动拷贝数据到每台机器上	利用对象存储(OSS)， HDFS等分布式存储保存数据，并通过数据卷挂载到每个容器的本地目录
模型开发	开发、调试模型代码	手动安装Jupyter、Tensorboard	自动部署Jupyter+Tensorboard
模型å训练	训练	通过SSH登录到每台机器上手工执行训练任务	通过图形界面设置参数，一键启动训练
	监控	GPU监控，需要登录到每台机器不断地执行nvidia-smi； 训练过程监控, 手动配置TensorBoard	自动集成GPU资源监控服务；自动启动Tensorboard
	Checkpoint保存和模型导出	手动保存checkpoint和导出模型	自动将模型导出到分布式存储，支持checkpoint自动恢复
模型预测	模型预测	用户需自定义实现	提供Serving服务，自动支持负载均衡和弹性伸缩

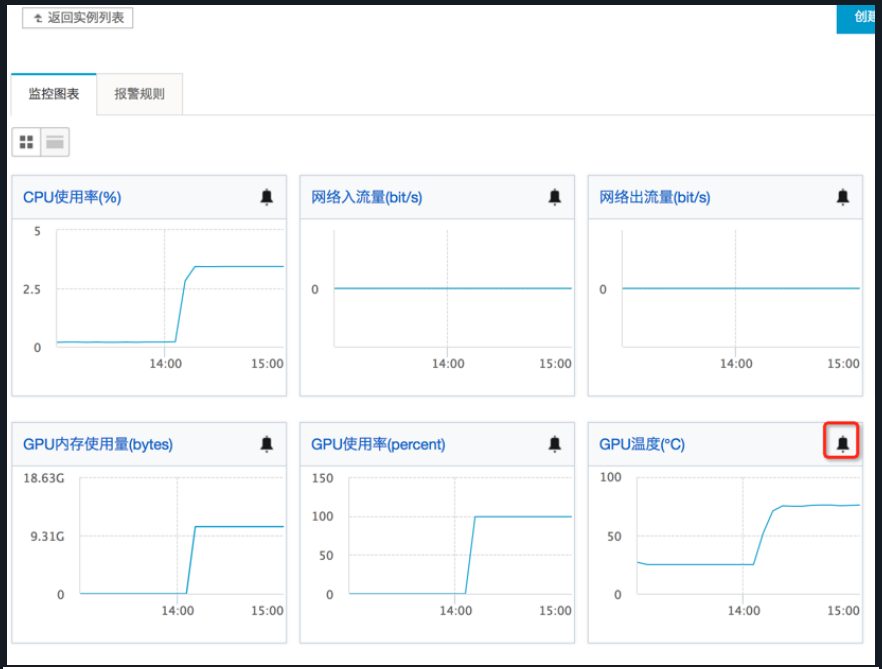
配置弹性GPU服务监控

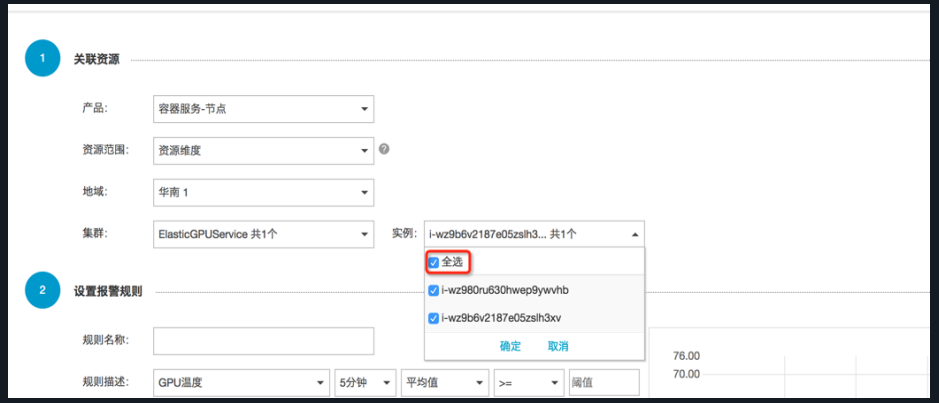
❶ 创建弹性GPU服务集群

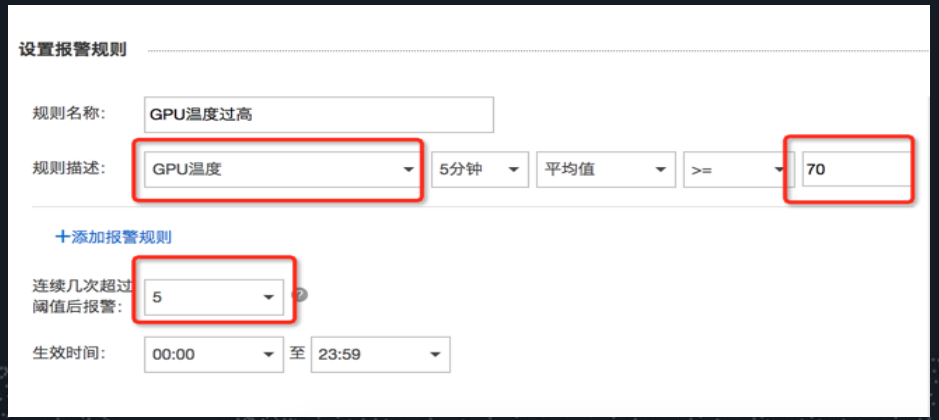
❷ 登录云监控查看节点

❸ 选择节点监控图标

配置弹性GPU服务监控

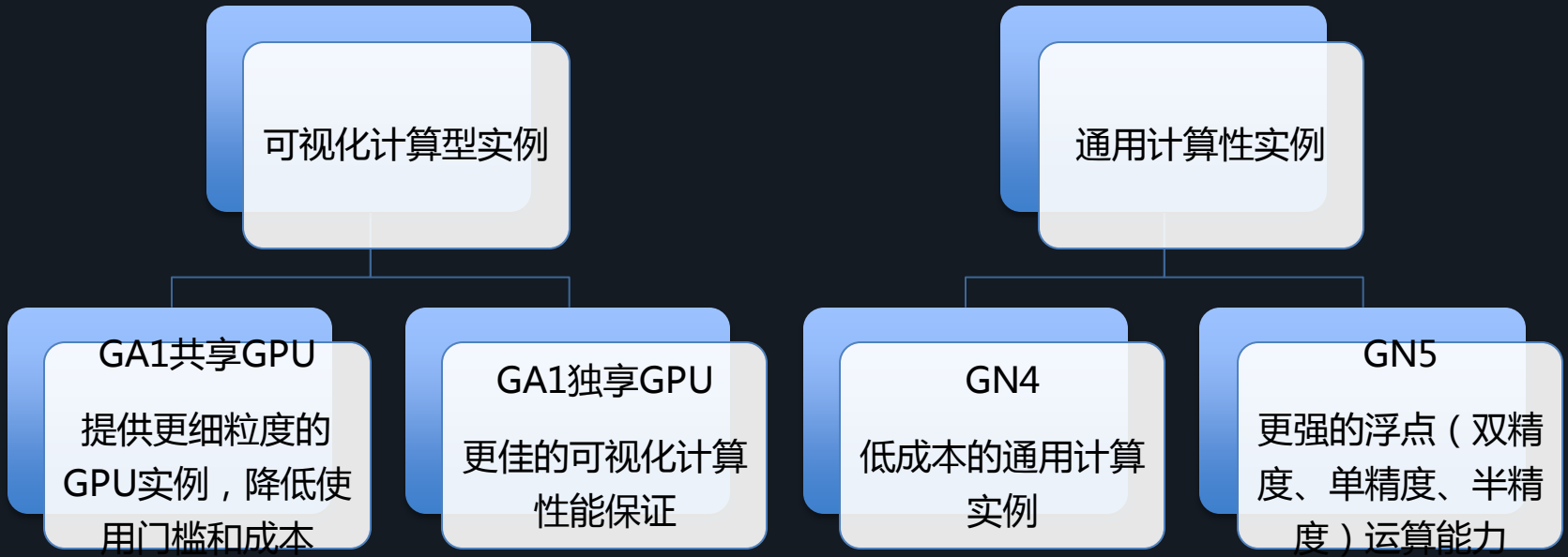






提供节点级别资源监控，其中包括CPU使用率、网络流量、GPU使用率、GPU显存使用率和温度等监控信息

EGS产品家族



GA1 – 可视化计算型

- GA1实例规格族是企业级异构计算ECS，提供了高性价比的渲染和视频处理能力
- 特性：
  - AMD S7150 GPU计算卡
  - Intel Xeon E5-2682v4 (Broadwell), 2.5GHz
  - DDR4 内存
  - 包含一块NVMe SSD本地盘存储
  - 共计 32GB 的 GPU显存、总计提供8192个并行处理核心、15 TFLOPS（单精度浮点运算处理能力）和1 (4x250G)TFLOPS（双精度峰值浮点性能）
- 用户场景：
  - 3D图形渲染，云游戏，电影、动画渲染
  - 视频处理
  - 视频编解码等场景

实例规格	vCPU	MEM (GiB)	GPU	显存 (GiB)	本地盘存储 (GiB)	Max pps ***	Max BW *** Gbps
ecs.ga1.2xlarge	8	20	AMD S7150 x ½	4	1 x 175	15万	1.5
ecs.ga1.4xlarge	16	40	AMD S7150 x 1	8	1 x 350	40万 ****	3
ecs.ga1.8xlarge	32	80	AMD S7150 x 2	16	1 x 700	80万**	6
ecs.ga1.14xlarge	56	160	AMD S7150 x 4	32	1x1400	120万*	10

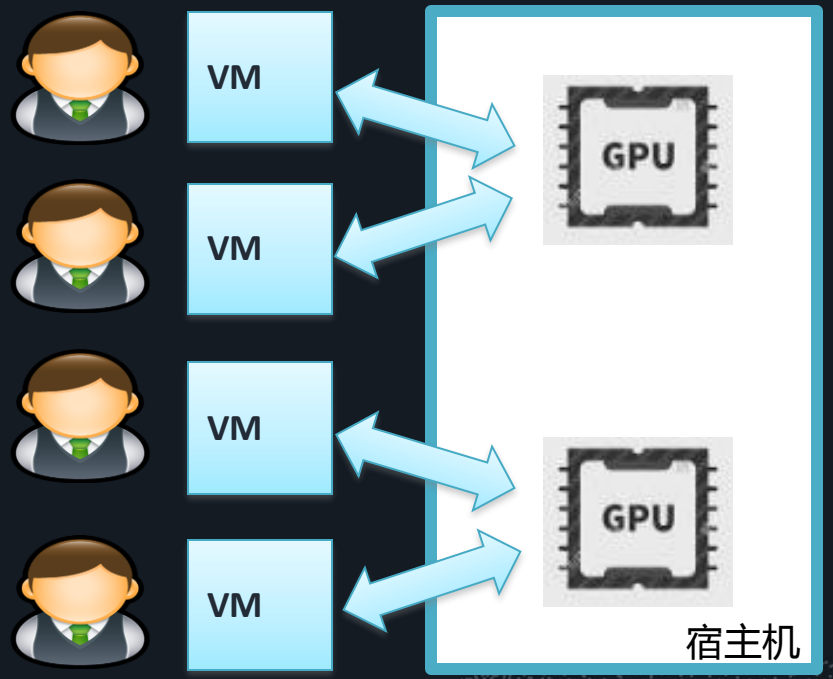
* 需要开启4队列，操作系统（镜像） CentOS 7.3，调整队列可能需要重启实例
** 需要开启3队列，操作系统（镜像） CentOS 7.3，调整队列可能需要重启实例
*** 网络性能持续提升中
**** 需要开启2队列，操作系统（镜像） CentOS 7.3，调整队列可能需要重启实例

可视化计算共享GPU实例

共享GPU实例

- 可以更加灵活的选择和自身业务需求匹配的GPU实例规格
- 将一块物理GPU同时共享给多个用户使用，避免不必要的资源浪费
- 基于硬件虚拟化技术，可以完美实现高性能GPU在硬件安全隔离下为多用户共享
- GPU虚拟化损耗可忽略，每用户分配到的GPU能力严格保证，相互之间无干扰

率先在公有云上推出共享GPU实例
可视化计算成本大降50%



GN4 – Nvidia Tesla M40通用计算加速型

- GN4实例规格族是企业级异构计算ECS，提供了高性价比深度学习和视频处理能力
- 特性：
  - Nvidia M40 GPU卡，最大支持2块
  - Intel Xeon E5-2682v4 (Broadwell), 2.5GHz
  - DDR4 内存
  - 共计 24GB 的 GPU显存、总计提供6000个并行处理核心、最高14 TFLOPS的单精度浮点运算处理能力
- 用户场景：
  - 深度学习
  - 科学计算
  - 基因测序
  - 电影、动画渲染
  - 视频处理
  - 视频编解码等

实例规格	vCPU	MEM (GiB)	GPU (M40)	Max pps ***	Max BW ***
gn4-c4g1.xlarge	4	30	1	10万	800 Mbps
gn4-c8g1.2xlarge	8	30	1	20万	1.5 Gbps
gn4.8xlarge	32	48	1	80万**	6Gbps
gn4-c4g1.2xlarge	8	60	2	20万	1.5 Gbps
gn4-c8g1.4xlarge	16	60	2	40万	3 Gbps
ecs.gn4.14xlarge	56	96	2	120万*	10Gbps

* 需要开启4队列，操作系统（镜像） CentOS 7.3，调整队列可能需要重启实例
** 需要开启3队列，操作系统（镜像） CentOS 7.3，调整队列可能需要重启实例
*** 网络性能持续提升中
**** 需要开启2队列，操作系统（镜像） CentOS 7.3，调整队列可能需要重启实例

GN5 –Nvidia Tesla P100通用计算加速型

- GN5实例规格族是企业级异构计算ECS，提供了超高性能深度学习和视频处理能力
- 特性：
  - Nvidia P100 GPU卡，最大可支持8块
  - Intel Xeon E5-2682v4 (Broadwell), 2.5GHz
  - DDR4 内存
  - 最大 96GB 的 GPU显存、总计提供28672个并行处理核心、最高84.8TFLOPS的单精度浮点运算处理能力
- 用户场景：
  - 深度学习
  - 科学计算
  - 基因测序
  - 电影、动画渲染
  - 视频处理
  - 视频编解码等

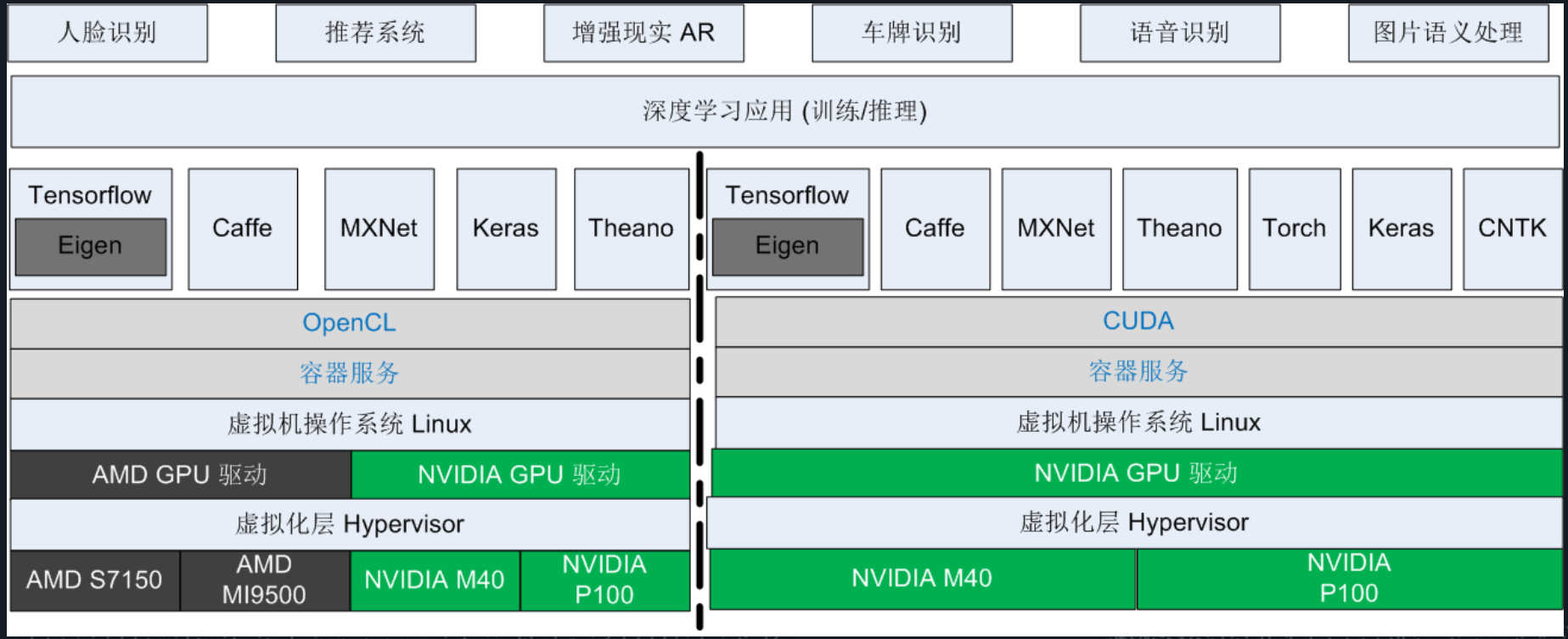
实例规格	vCPU	MEM (GiB)	GPU (P100)	SSD (GiB)	Max pps ***	Max BW ***
gn5-c4g1.xlarge	4	30	1	440	10万	800 Mbps
gn5-c8g1.2xlarge	8	60	1	440	20万	1.5 Gbps
gn5-c4g1.2xlarge	8	60	2	880	20万	1.5 Gbps
gn5-c8g1.4xlarge	16	120	2	880	40万	3 Gbps
gn5-c8g1.8xlarge	32	240	4	1760	80万**	6 Gbps
gn5-c8g1.14xlarge	56	480	8	3520	120万*	10Gbps

* 需要开启4队列，操作系统（镜像） CentOS 7.3，调整队列可能需要重启实例
** 需要开启3队列，操作系统（镜像） CentOS 7.3，调整队列可能需要重启实例
*** 网络性能持续提升中
**** 需要开启2队列，操作系统（镜像） CentOS 7.3，调整队列可能需要重启实例

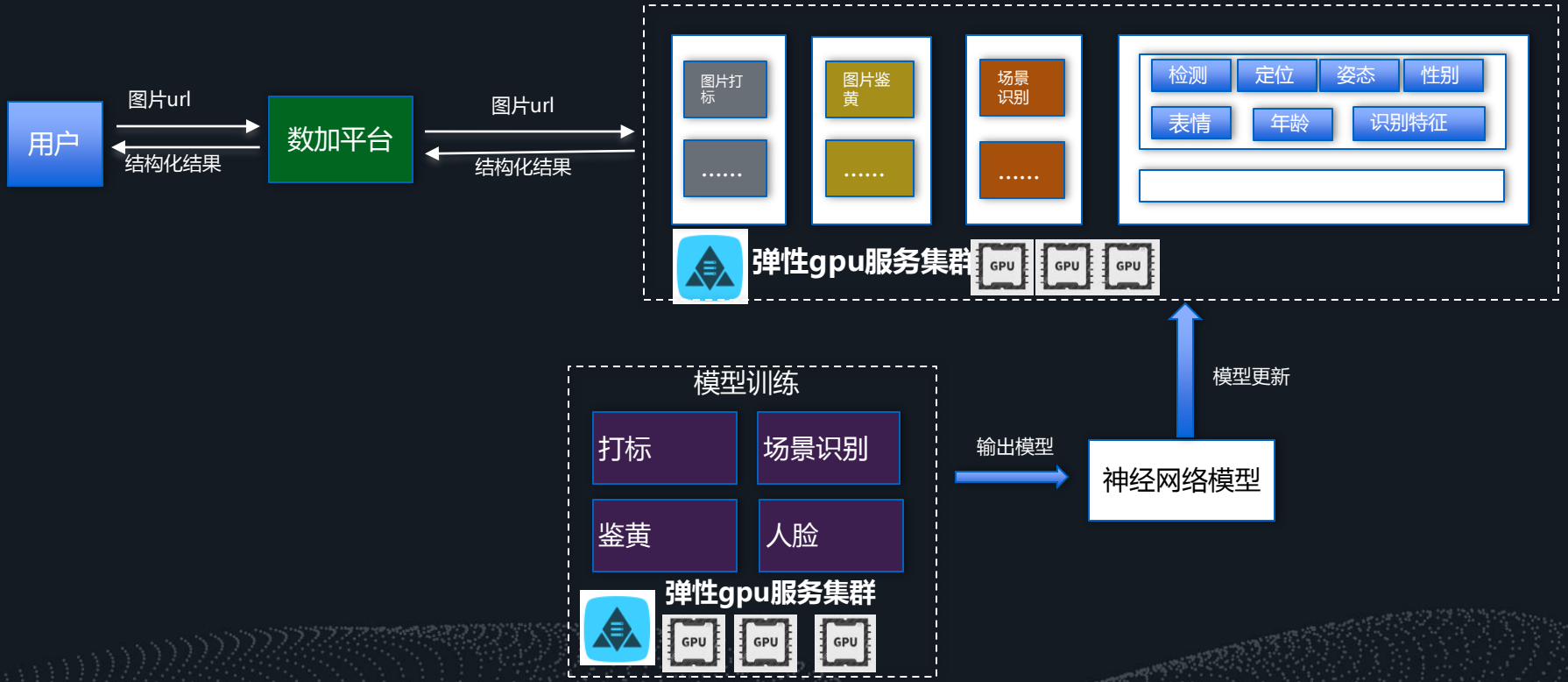
视觉智能场景规格族推荐

一级分类	二级分类	最新一代适用规格族	适用场景	计算			网络		存储
				CPU	CPU：内存	GPU	最大带宽(Gbps)	收发包能力(万PPS)	数据盘
机器学习	训练	gn5	离线训练	E5-2682v4 2.5GHz	1:7.5	Nvidia P100	10	120	SSD实例存储、高效云盘、SSD云盘
		gn4		E5-2682v4 2.5GHz	1:7.5	Nvidia M40	10	120	高效云盘、SSD云盘
	推理	gn4	在线推理	E5-2682v4 2.5GHz	1:7.5	Nvidia M40	10	120	高效云盘、SSD云盘
多媒体	视频编码	c4	基于CPU进行编解码	E5-2667v4 3.2GHz	1:2	/	10	120	高效云盘、SSD云盘
		gn4	NVEnc(H.264)	E5-2682v4 2.5GHz	1:7.5	Nvidia M40	10	120	高效云盘、SSD云盘
		ga1	VCE(H.264)	E5-2682v4 2.5GHz	1:2.5	AMD S7150	10	120	SSD实例存储、高效云盘、SSD云盘
	渲染	ga1	AMD S7150图形卡	E5-2682v4 2.5GHz	1:2.5	AMD S7150	10	120	SSD实例存储、高效云盘、SSD云盘
		gn4	Nvidia M40卡	E5-2682v4 2.5GHz	1:7.5	Nvidia M40	10	120	高效云盘、SSD云盘

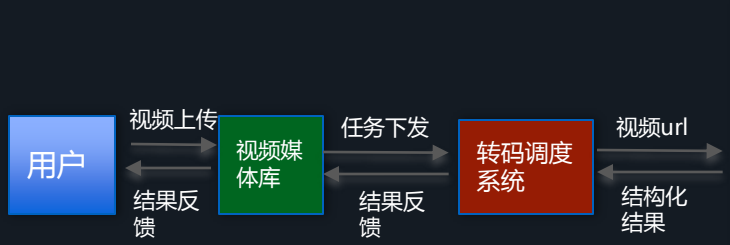
弹性GPU服务支撑视觉智能创新



典型应用——弹性GPU实例加速阿里云图像识别服务



典型应用——弹性GPU实例加速阿里云视频转码服务



弹性GPU服务

售卖区域

- 国内
  - 华北2
  - 华东2
  - 华南1
- 国外
  - 美国东部1

*更多区域即将上线，请留意官网通知

阿里云弹性GPU服务全球化部署，加速客户AI智能创新

飞天·智能

APSARA INTELLIGENCE