ALIBABA CLOUD

阿里云

FPGA云服务器

产品简介

文档版本：20220602

阿里云

法律声明

阿里云提醒您在阅读或使用本文档之前仔细阅读、充分理解本法律声明各条款的内容。如果您阅读或使用本文档，您的阅读或使用行为将被视为对本声明全部内容的认可。

1. 您应当通过阿里云网站或阿里云提供的其他授权通道下载、获取本文档，且仅能用于自身的合法合规的业务活动。本文档的内容视为阿里云的保密信息，您应当严格遵守保密义务；未经阿里云事先书面同意，您不得向任何第三方披露本手册内容或提供给任何第三方使用。

2. 未经阿里云事先书面许可，任何单位、公司或个人不得擅自摘抄、翻译、复制本文档内容的部分或全部，不得以任何方式或途径进行传播和宣传。

3. 由于产品版本升级、调整或其他原因，本文档内容有可能变更。阿里云保留在没有任何通知或者提示下对本文档的内容进行修改的权利，并在阿里云授权通道中不时发布更新后的用户文档。您应当实时关注用户文档的版本变更并通过阿里云授权渠道下载、获取最新版的用户文档。

4. 本文档仅作为用户使用阿里云产品及服务的参考性指引，阿里云以产品及服务的“现状”、“有缺陷”和“当前功能”的状态提供本文档。阿里云在现有技术的基础上尽最大努力提供相应的介绍及操作指引，但阿里云在此明确声明对本文档内容的准确性、完整性、适用性、可靠性等不作任何明示或暗示的保证。任何单位、公司或个人因为下载、使用或信赖本文档而发生任何差错或经济损失的，阿里云不承担任何法律责任。在任何情况下，阿里云均不对任何间接性、后果性、惩戒性、偶然性、特殊性或刑罚性的损害，包括用户使用或信赖本文档而遭受的利润损失，承担责任（即使阿里云已被告知该等损失的可能性）。

5. 阿里云网站上所有内容，包括但不限于著作、产品、图片、档案、资讯、资料、网站架构、网站画面的安排、网页设计，均由阿里云和/或其关联公司依法拥有其知识产权，包括但不限于商标权、专利权、著作权、商业秘密等。非经阿里云和/或其关联公司书面同意，任何人不得擅自使用、修改、复制、公开传播、改变、散布、发行或公开发表阿里云网站、产品程序或内容。此外，未经阿里云事先书面同意，任何人不得为了任何营销、广告、促销或其他目的使用、公布或复制阿里云的名称（包括但不限于单独为或以组合形式包含“阿里云”、“Aliyun”、“万网”等阿里云和/或其关联公司品牌，上述品牌的附属标志及图案或任何类似公司名称、商号、商标、产品或服务名称、域名、图案标示、标志、标识或通过特定描述使第三方能够识别阿里云和/或其关联公司）。

6. 如若发现本文档存在任何错误，请与阿里云取得直接联系。

通用约定

格式	说明	样例
危险	该类警示信息将导致系统重大变更甚至故障，或者导致人身伤害等结果。	危险 重置操作将丢失用户配置数据。
警告	该类警示信息可能会导致系统重大变更甚至故障，或者导致人身伤害等结果。	警告 重启操作将导致业务中断，恢复业务时间约十分钟。
注意	用于警示信息、补充说明等，是用户必须了解的内容。	注意 权重设置为0，该服务器不会再接受新请求。
说明	用于补充说明、最佳实践、窍门等，不是用户必须了解的内容。	说明 您也可以通过按Ctrl+A选中全部文件。
>	多级菜单递进。	单击**设置**> **网络**> **设置网络类型**。
粗体	表示按键、菜单、页面名称等UI元素。	在**结果确认**页面，单击**确定**。
`Courier字体`	命令或代码。	执行 `cd /d C:/window` 命令，进入Windows系统文件夹。
斜体	表示参数、变量。	`bae log list --instanceid` *Instance_ID*
[] 或者 [a|b]	表示可选项，至多选择一个。	`ipconfig [-all|-t]`
{} 或者 {a|b}	表示必选项，至多选择一个。	`switch {active|stand}`

目录

1.什么是FPGA云服务器

FPGA云服务器是一类提供了现场可编程门阵列（FPGA）的实例规格。由于FPGA硬件的可重配特性，您可以快速擦写和重配已创建的FPGA硬件加速应用，同时拥有低时延硬件与资源弹性。

FaaS平台介绍

传统FPGA开发硬件周期长，开发难度大，硬件加速算法的发布和部署保护要求也非常高。FPGA云服务器平台FaaS（FPGA as a Service）在云端提供统一硬件平台与中间件，可大大降低加速器的开发与部署成本。您无需了解底层硬件即可快速开发和部署自己的定制加速器，也可以直接使用加速器提供商提供的加速服务。

FaaS包括三个组件：

- 硬件基础设施：FPGA云服务器、硬件加速开发和部署平台（Intel、Xilinx）。
- 云上配套开发环境：厂商配套软件（Quartus、Vivado）、第三方EDA软件（仿真、模拟）。
- FPGA IP开发生态：图片转码、基因计算、数据加密、视频压缩、硬件仿真设计、深度学习（预测/训练）等。

阿里云基于FaaS平台推出了FPGA云服务器，在提供FPGA加速能力的同时，保留了与普通ECS实例一致的使用体验。您在创建ECS实例时，选择企业级异构计算规格即可。实例规格的详细信息，请参见实例规格族。

功能特性

- 统一性

  兼容多种FPGA器件（Intel、Xilinx），支持Multi-boot的Shell烧写，更可靠，易移植开发。

- FPGA虚拟化

  自主研发的FPGA软硬件虚拟化方案，实现上云安全隔离要求。支持热升级功能，在不中断业务的前提下，对部分用户逻辑进行在线重配置，以实现新的功能。

- 联合仿真平台

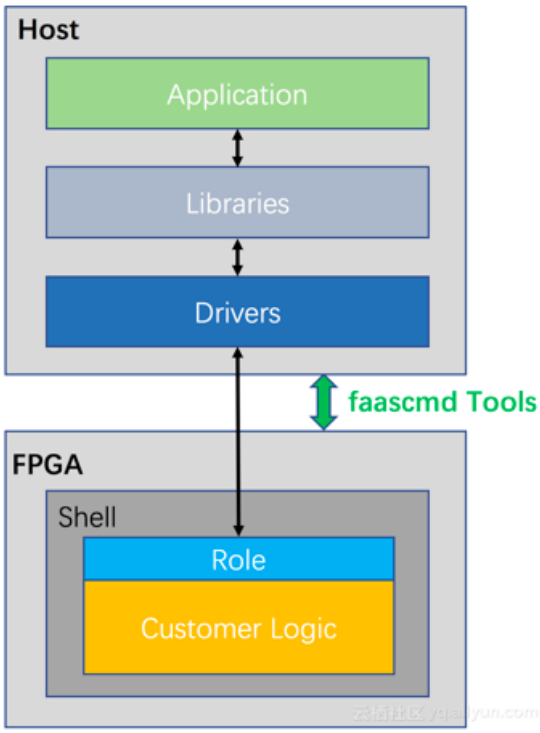
  支持Intel和Xilinx器件，您无须更改原有设计即可进行软硬件联合仿真，降低输出FPGA高性价比算力的复杂度。

- 互联拓扑动态可配置

  支持1片、2片、4片FPGA互联拓扑，可动态配置拓扑，实现最高性价比。同卡FPGA之间使用高速互联通道，应用实现两片FPGA之间实时、大批量数据搬运时，不存在带宽瓶颈。

工具套件

FaaS平台提供HDK和SDK套件，搭建更加高效、统一的开发及部署平台。



- HDK采用Shell+Role的组合方式，保证Shell的最轻量化和稳定性，同时兼顾便捷性和灵活性。
- SDK包括两部分：
  - HDK对应的主机端驱动（Drivers）与软件库（Libraries），和HDK的Shell、Role相对应，一起为您提供统一灵活的软件支持。
  - FPGA管理工具faascmd套件，为您提供云上FPGA管理服务，包括BIT/DCP文件安全校验、FPGA镜像生成、下载及管理、FPGA加速卡状态查询反馈等功能。

  FaaS的镜像相关操作依赖于阿里云OSS存储，因此使用FPGA云服务器时必须开通OSS服务，具体操作请参见开通OSS服务。

产品计费

FPGA云服务器的计费相关功能和云服务器ECS一致，其中，计算资源（vCPU和内存）、镜像、块存储、公网带宽以及快照等资源涉及计费。

常见的计费方式如下所示：

- 包年包月：按一定时长购买资源，先付费后使用。
- 按量付费：按需开通和释放资源，先使用后付费。
- 抢占式实例：通过竞价模式抢占库存充足的计算资源，相对按量付费实例有一定的折扣，但是存在回收机制。
- 预留实例券：搭配按量付费实例使用的抵扣券，承诺使用指定配置的实例（包括实例规格、地域可用区等），以折扣价抵扣计算资源的账单。
- 节省计划：搭配按量付费实例使用的折扣权益计划，承诺使用稳定数量的资源（以元/小时为单位衡量），以折扣价抵扣计算资源、系统盘等资源的账单。
- 存储容量单位包：搭配按量付费存储产品使用的资源包，承诺使用指定容量的存储资源，以折扣价抵扣块存储、NAS、OSS等资源的账单。

更多云服务器ECS计费的介绍，请参见计费概述和云产品定价页。

2.FaaS f3云服务器简介

本文为您详细介绍FPGA云服务器的推荐主售规格族f3型FPGA云服务器的相关概念、组成架构以及HDK介绍。

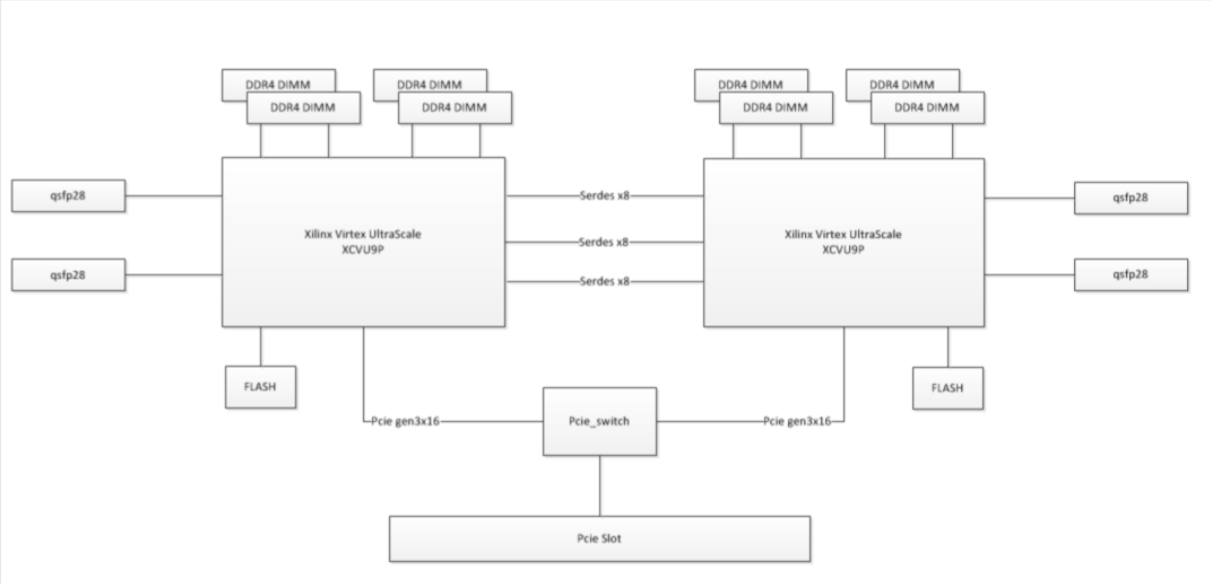
概述

随着云计算和数据中心业务的迅速发展，对于算力的需求呈现上升趋势。在越来越多对算力有高度要求的场景中，仅靠CPU提供的算力已无法满足需求。实践表明，在特定的应用场景中，FPGA相比CPU能够取得几十、乃至上百倍的加速效果，而延时则可以降低两个数量级。因此，在2019年9月的云栖大会上，阿里云发布了基于Xilinx 16nm工艺器件的f3实例，该实例首推了单卡双芯片设计，且计算密度处于领先地位。

阿里云f3 FPGA云服务器（FPGA as a Server, FaaS）为开发者提供了云上FPGA开发和使用的工具及环境，具有易用、经济、敏捷和安全的优势，能够让您轻松地进行FPGA加速器的开发以及基于FPGA加速的业务部署。

硬件架构

FaaS f3的基本结构为单卡双芯片，实现了片间互联以及卡间互联。其硬件架构如下图所示：



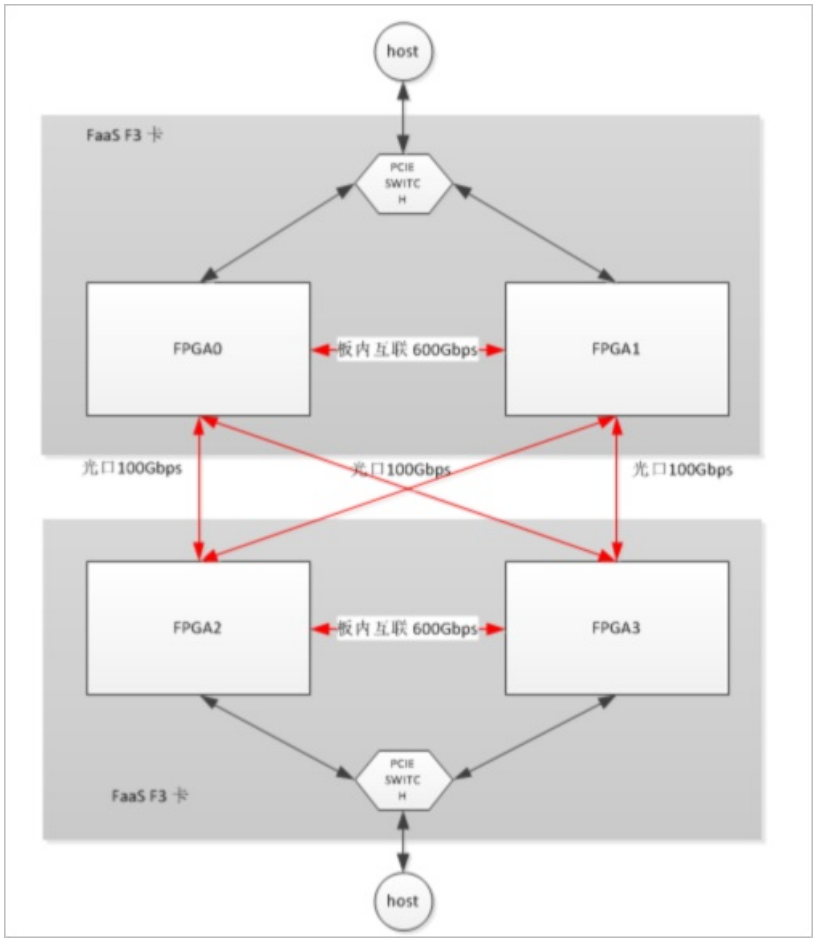
规格说明

规格	说明（单VU9P）
尺寸	全高全长
FPGA型号	XCVU9P
PCIe接口	PCIe GEN3 X16
内存	4 x DDR4 2133MHz，容量为4 x 16GB
片间互联	200 Gbps x 3
Ethernet接口	100 Gbps x 2

规格	说明（单VU9P）
时钟模块	时钟可动态配置

拓扑结构

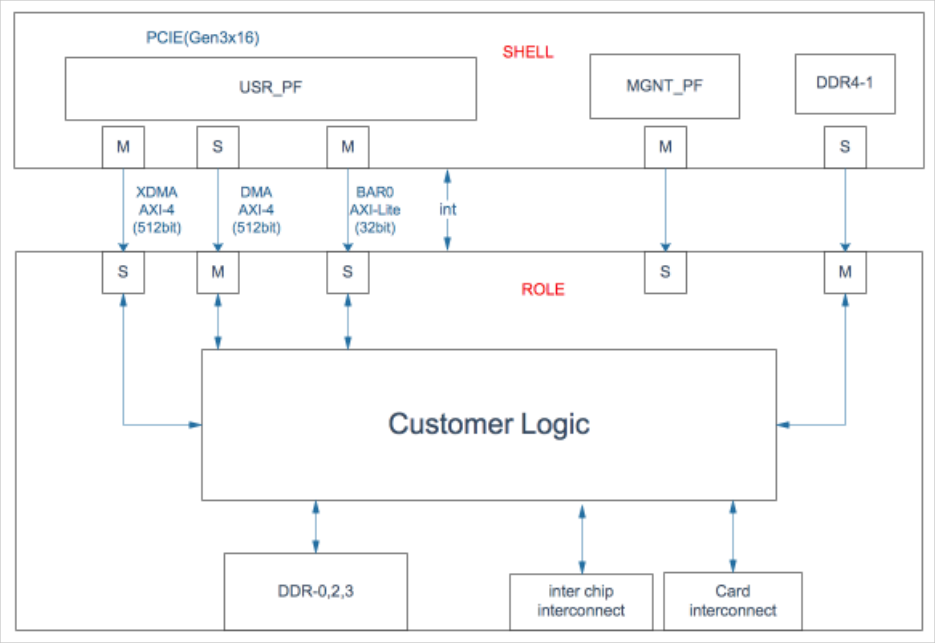
FaaS f3采用双卡互联拓扑结构，能够使每个FPGA之间相互通信，且最小通信带宽为100 Gbps。其拓扑结构如下图所示：



开发环境

平台	说明
开发工具	Vivado 2018.2
芯片	XCVU9P
开发环境	Linux CentOS 7.4
内核版本	3.10.0-693.el7.x86_64

逻辑结构



如上图所示，VU9P芯片内包含以下三部分：

- SHELL

  静态区，其包含PCIE DMA/XDMA、寄存器通路、DDR1和其他管控逻辑。

- ROLE

  动态区域，包含了三路DDR控制器（DDR0、DDR2、DDR3）、DMA交互通路、serdes（片间互联和板卡互联）。

- Customer Logic

  包含在Role内部，根据提供的固定接口逻辑进行自己的逻辑定制。

用户接口描述

信号名	I/O	位宽	描述	
	sys_alite_aclk	I	1	寄存器时钟域时钟，该时钟为50 MHz。
	sys_alite_aresetn	I	1	寄存器时钟域复位信号。
	sys_clk_200m	I	1	用户时钟，200 MHz，您可以利用该时钟连接MMCM来扩展时钟。
	sys_clk_rstn	I	1	全局复位信号。

信号名	I/O	位宽	描述	
clock&reset	kernel_clk_300m	I	1	用户时钟，300 Mhz，该时钟固定可配置，一般建议您采用该时钟作为主时钟。
	kernel_clk_rstn	I	1	用户时钟复位信号。
	kernel2_clk_500m	I	1	用户时钟，500 Mhz，该时钟固定可配置。
	kernel2_clk_rstn	I	1	用户时钟复位信号。
	pcie_axi_aclk	I	1	pcie axi clock: PCIEcore时钟域，xdma/dma/int接口时钟域。
	pcie_axi_arstn	I	1	pcie core rstn。
	c0_ddr4_ui_clk	I	1	ddr0 channel时钟域。
	c0_ddr4_rstn	I	1	ddr0 channel时钟域复位信号。
	c1_ddr4_ui_clk	I	1	ddr1 channel时钟域。
	c1_ddr4_rstn	I	1	ddr1 channel时钟域复位信号。
	c2_ddr4_ui_clk	I	1	ddr2 channel时钟域。
	c2_ddr4_rstn	I	1	ddr2 channel时钟域复位信号。
	c3_ddr4_ui_clk	I	1	ddr3 channel时钟域。
	c3_ddr4_rstn	I	1	ddr3 channel时钟域复位信号。

信号名	I/O	位宽	描述	
AXI-MM	XDMA	-	-	AXI MM标准接口。 • 接口的具体信息，您可以参考AXI4_specification规格说明书。 • XDMA：您可以前往Xilinx官网查看PG195的相关信息。 • DMA：您可以前往Xilinx官网查看PG194的相关信息。 • AXI- Lite：提供给用户接口8 MB的寄存器访问空间。 注意 在接口使用中注意各自的时钟域。
	DMA	-	-	
	AXI-Lite	-	-	
	DDR0/1/2/3	-	-	
	int	-	16	您可以单独发送16个中断上报，其时钟域为pcie_axi_aclk。
AXI_STREAM	inter chipinterconnect	-	-	轻量级互联接口采用Xilinx aurora协议标准。您可以前往Xilinx官网查看PG074的相关信息。
	Card interconncet	-	-	

术语

术语	描述
FaaS	FPGA as a Server，FaaS
HDK	Hardware Develop Kit，硬件开发套件
SDK	Software Develop Kit，软件开发套件
SHELL	静态逻辑，包括PCIe、DDR4等外部接口
Role	动态逻辑，PR区域

术语	描述
CL	Customer Logic，客户逻辑，由开发者提供
PR	Partial Reconfigration，部分重加载技术
MGNT PF	Mangement Phsical Function
USRPF	User Phsical Function
OpenCL	Open Computing Language
HAL	Hardware Abstraction Layer

3.产品优势

FPGA云服务器具有独特优越的加速性能和经济实惠的性价比，并且易于复用已有的FPGA设计。

- 分钟级交付

  基于阿里云弹性计算框架，您可以几分钟内轻松创建FPGA实例，创建自定义的专用硬件加速器。

- 独特优越的加速性能

  FPGA器件通过PCIe 3.0接口与上层应用程序通信。对应用中消耗大量的CPU计算，系统可以选择性的交给FPGA专用硬件进行加速计算。释放CPU资源用于支持更大的应用访问量和吞吐量。您可选高配FPGA特性，共同使用一个地址空间相互通信速度高达48 Gbit/s。

- 经济实惠的性价比

  购买FPGA实例时，无需单独购买FPGA器件和板卡。FPGA实例支持按量付费，可随时释放，实现轻资产开发，降低项目研发期间的投入成本。

- 可复用已有的设计

  如果您已经有现成的FPGA工程，可以利用阿里云提供的开发环境和硬件支持包（BSP）轻松地将已有工程导入云端，并在FPGA实例中使用。阿里云提供了Intel、Xilinx主流的开发软件工具链，FPGA实例中的FPGA器件也兼顾两家，方便您根据原设计的具体情况进行合理选择，无缝迁移。

4.应用场景

本章节介绍FPGA云服务器的典型应用场景。

直播实时视频转码

阿里云异构GPU/FPGA服务器重点支持2019年双11猫晚直播的实时视频转码，以高画质、低带宽、高分辨率、实时的综合优势服务猫晚当天直播业务4K、2K、1080P等各个分辨率的转码。其中FPGA H.265高清编码、720P节省带宽21.6%，GPU云服务器支持高并发实时视频流5000路以上，并逐步上升到峰值6200路每分钟，且顺利度过洪峰。异构GPU云服务器还参与实时家居渲染图片生成等业务，首次提供了大量算力强劲的ebmgn6v裸金属实例，支持淘宝渲染方提升几十倍的渲染性能，第一次实现秒级实时渲染，完成总计超过5000张大型家居渲染图。异构FPGA图片转码业务则以3K+片的超大集群，为淘宝图片空间提供高达数百万QPS的处理能力，承担了双十一淘宝图片85%的流量，预计节省计算成本3亿。

人工智能

目前，GPU是人工智能技术方案的首选，原因在于两个方面：

- GPU具有完善的生态和高并行度的计算力，能很好地帮助您实现方案和部署上线。
- 人工智能的发展仍处于早期阶段，各个行业正在从算法层面尝试寻找商业落地的可能性，是一个从0到1的过程。

可以预见在未来几年，人工智能落地应用越来越多，大规模商业部署渐渐成为可能。这时对于更低功耗、更低成本、更低处理延时、更多定制化等方面的需求将会逐渐凸显。在人工智能大规模商业部署（推理应用）中，f3实例将具备独特的性能优势和广阔的潜在空间。

GPU计算的处理优势在于拥有众多专用的并行计算单元以及超高的显存带宽，让多路大规模数据搬移和快速并行计算成为典型的计算模式，但该模式也导致了每路数据的处理延迟增加。在具有低延迟需求的在线业务场景中，例如语音识别等，在Batch值较小的情况下，f3实例的处理延时仅为GPU的1/10。

深度神经网络计算的一个发展趋势是降低数据表示的精度，降低网络对于计算力的需求，以提高计算吞吐量。从双精度浮点到单精度浮点，再到定点处理，而定点运算是FPGA的传统优势。与GPU相比，FPGA内部配备了众多的定点处理单元，甚至可以将整个FPGA芯片的内部逻辑资源配置成定点处理单元，进而具备了超高的定点运算能力。

基因测序

基因测序是一种新型基因检测技术，能够从血液或唾液中分析测定基因全序列，预测罹患多种疾病的可能性。基因测序技术能锁定个人病变基因，提前预防和治疗，目前一个广为人知的用途是针对唐氏综合征的无创产前基因检测。随着基因测序技术的快速发展，基因数据的生成呈现指数级增长，应用也越来越广泛，对分析能力提出更高要求。

传统的计算系统通过采用多个高端CPU搭建HPC系统实现了缩短时间的目的，但导致成本增加，行业应用规模以及基因企业发展规模受限。目前中国内地基因企业面临基因计算成本偏高，但业务需求旺盛的行业困境，急需高性价比的算力资源。

以人类全基因组（WGS）分析为例，使用一台16 vCPU、64 GiB的CPU实例，完成单个WGS分析耗时近100小时，而f3实例可以在30分钟以内完成，极大地缩减了计算时间，降低了成本。

IC设计原型验证

在传统的数字IC设计流程中，使用FPGA搭建芯片原型验证平台测试功能是一个重要环节，验证过程需要大量的FPGA逻辑单元。而对于传统数字芯片设计公司，购买或者自研复杂的FPGA验证单板或平台，不仅耗时耗力，而且本不是公司主要业务方案。加之FPGA平台升级换代速度快过芯片设计周期，大型数字芯片设计中追求更大逻辑量FGPA板卡，需要不断研制最新fpga板卡一直是个痛点。

f3实例选用单芯片逻辑单元达250万个的VU9P，支持双芯片600 Gbit/s互联以及多板卡间的100 Gbit/s互联。f3实例最多支持16个VU9P芯片，充分满足了数字芯片原型验证阶段对于大逻辑量的需求。同时选择f3实例还可以避免维护复杂FPGA板卡，缩减了验证平台的维护成本。

云端压缩的计算加速

云上用户在进行大数据存储、高速网络传输时，常常因为实例性能需要在效率和成本之间做出取舍。gzip是一种广泛用于互联网服务的压缩工具，但传统的CPU实现gzip效率低、耗时长、难以支撑较大流量。使用带有FPGA的计算实例进行gzip压缩，性能比仅用CPU的普通实例提升8~10倍，充分满足用户的数据压缩需求。

在后台服务日志压缩、网站静态资源文件压缩、批量计算任务压缩、分布式存储压缩等方面均可使用FPGA进行加速。

数据库加速

以大型互联网公司为例，每天处理的数据量级都在PB，每天更新的网页以亿计，每24小时更新的日志超过PB，因此需要大型的集群处理大规模的数据。在处理大规模数据时，数据仓库的性能直接影响数据本身的处理能力。

f3实例得益于FPGA细颗粒度的数据处理能力、高并发度的并行计算能力，能够大幅提升数据库产品的性能。

- 以数据库处理中的排序单元为例，在PostgreSQL的核心处理单元加速中，f3实例的性能比只使用CPU提升10倍以上。
- 以时序数据处理为例，时序数据广泛应用于物联网（IoT）设备监控系统、企业能源管理系统（EMS）、生产安全监控系统、电力检测系统等行业场景，f3实例单路数据吞吐性能是单核CPU的30倍以上。

5.基本概念

本文汇总使用FPGA云服务器过程中涉及的基本概念，方便您查询和了解相关概念。

FPGA云服务器概念

名词	说明	相关文档
FPGA	现场可编程门阵列（Field Programmable Gate Array），具有硬件并行加速能力和可编程特性。	什么是FPGA云服务器
FaaS	FPGA即服务（FPGA as a Service），阿里云在云端提供的统一硬件平台与中间件，可大大降低加速器的开发与部署成本。	FaaS平台介绍
faasutil	新一代命令行工具，用简单的命令完成配置环境、生成FPGA镜像、加载FPGA镜像等操作。	获取faasutil
faascmd	FPGA云服务器提供的一个命令行工具，是基于Python SDK开发的脚本，提供云上FPGA管理服务。	faascmd工具概述
FPGA镜像	FPGA设备使用的镜像，用于安全开放加速能力，faascmd提供全套镜像管理流程。	无
OSS	阿里云对象存储服务（Object Storage Service），是阿里云提供的海量、安全、低成本、高可靠的云存储服务。使用FaaS服务时需要创建OSS Bucket存放相关文件。	什么是对象存储OSS

ECS通用概念

概念	说明
ECS实例	等同于一台虚拟服务器，内含CPU、内存、操作系统、网络配置、磁盘等基础的组件。
ECS实例规格	定义了ECS实例在计算性能、存储性能、网络性能等方面的基本属性，但需要同时配合镜像、块存储、网络等配置才能确定一台ECS实例的具体服务形态。
镜像	提供了运行实例所需的信息，包括操作系统、初始化应用数据等。
公共镜像	阿里云官方提供的基础镜像，均已获得正版授权，涵盖Windows Server系统镜像和主流的Linux系统镜像。 | • Alibaba Cloud Linux 3 • Alibaba Cloud Linux 2	阿里云官方操作系统，为云上应用程序提供安全、稳定、高性能的定制化运行环境，并进行了针对性深度优化，更加适合阿里云基础设施。
自定义镜像	您自行创建或导入的镜像，包含了初始系统环境、应用环境、软件配置等信息，可以节省重复配置的时间。
块存储	高性能、低时延的块设备，像物理硬盘一样分区格式化并创建文件系统后使用，满足大部分通用业务场景下的数据存储需求。

概念	说明
云盘	数据块级别的块存储产品，采用分布式三副本机制，为ECS实例提供99.9999999%的数据可靠性保证。
本地盘	ECS实例所在物理机上的本地硬盘设备，存储I/O性能、海量存储的性价比极高，但数据可靠性取决于物理机的可靠性，存在单点故障风险。
快照	某一时间点云盘数据状态的备份文件，用于备份或者恢复整个云盘。
安全组	一种虚拟防火墙，您可以基于安全组控制实例的入流量和出流量。
SSH密钥对	一种安全便捷的登录认证方式，由公钥和私钥组成，仅支持Linux实例。
实例RAM角色	ECS实例通过实例RAM角色获得该角色拥有的权限，可以基于临时安全令牌STS（Security Token Service）访问指定云服务的API和操作指定的云资源，安全性更高。
专有网络	您基于阿里云创建的自定义私有网络，不同专有网络之间通过隧道在逻辑上彻底隔离。您可以完全掌控自己的专有网络，例如选择IP地址范围、配置路由表和网关等。
弹性网卡	一种独立的虚拟网卡，可以绑定到ECS实例或从ECS实例解绑，实现业务的灵活扩展和迁移。
实例启动模板	包含了ECS实例的配置信息，使用实例启动模板创建ECS实例可以免去重复配置的操作。
部署集	部署集支持高可用策略，部署集内实例会严格分散在不同的物理服务器上，保证业务的高可用性和底层容灾能力。
专有宿主机	一台由单租户独享物理资源的云主机，具有满足严格的安全合规要求、允许自带许可证（BYOL）上云等优势。
弹性供应组	用于快速部署多可用区、多实例规格的ECS实例集群，通过多种供应策略组合使用抢占式实例和按量付费实例，满足对低成本和高稳定性的要求。
标签	由一对键值（Key-Value）组成。使用标签标识具有相同特征的资源后，例如所属组织或用途相同的资源，您可以基于标签方便地检索和管理资源。
资源组	供您从业务角度管理跨地域、跨产品的资源，并支持针对资源组的权限管理。
云助手	阿里云提供的自动化运维工具，无需登录即可完成在ECS实例上执行命令、向ECS实例发送文件等操作。
系统事件	影响ECS实例运行状态的计划底层运维事件或非预期维修事件，需要进行重启、停止或释放ECS实例等操作。系统事件会及时发送通知、应对措施和事件周期等信息，方便您提前完成备份数据等准备工作。

6.使用限制

FPGA实例作为云服务器ECS的一类实例规格，保持了与ECS实例相同的使用限制。本文介绍云服务器ECS在产品功能和服务性能上的不同限制，以及如何申请更高配额。

限制概述

使用云服务器ECS有下列限制：

- 仅弹性裸金属服务器和超级计算集群支持二次虚拟化，其他规格族不支持安装虚拟化软件和二次虚拟化。
- 不支持声卡应用。
- 不支持直接加载外接硬件设备（如硬件加密狗、U盘、外接硬盘、银行U key等），您可以尝试软件加密狗或者动态口令二次验证等。
- 不支持多播协议。如果需要使用多播，建议改为使用单播点对点方式。
- 日志服务不支持32位Linux系统云服务器。

  如何查看日志服务支持的云服务器系统，请参见Logtail采集概述。
- 如果云服务器需要备案，则云服务器有购买要求，且每台ECS实例可申请的备案服务号数量有限。详情请参见ICP备案服务器（接入信息）准备与检查。备案流程请参见ICP备案流程概述。
- 部分软件或应用的许可证（License）需要与云服务器的硬件信息绑定。当云服务器进行迁移操作时可能会引起硬件信息的变更，进而导致License失效。

实例

限制项	限制	提升限额方式
创建ECS实例的用户限制	完成实名认证。	无
创建按量付费资源的限制	开通按量付费ECS资源时，您的阿里云账户余额（即现金余额）和代金券的总值不得小于100.00元人民币。	提交工单。
可以创建按量付费实例的规格	vCPU核数少于16（不含16）的实例规格。	提交工单。
指定地域可用区、实例规格、付费类型、网络类型的实例配额	在ECS管理控制台查看实例配额。具体操作，请参见查看和提升实例配额。	在ECS管理控制台申请提升实例配额。具体操作，请参见查看和提升实例配额。
单次可购买的包年包月实例的最大数量	在ECS管理控制台查看资源配额。具体操作，请参见查看和提升资源配额。	无
一个账号在每个地域的实例启动模板数量	30	无
一个实例启动模板中的版本数量	30	无
按量付费转包年包月	已停售的实例规格不支持按量付费转包年包月。更多信息，请参见 已停售的实例规格。	无

限制项	限制	提升限额方式
包年包月转按量付费	• 是否支持此功能根据您的云服务器使用情况而定。 • 每月有最大退款额度限制，额度以转换页面显示为准。	无

预留实例券

限制项	限制	提升限额方式
一个账号的地域级预留实例券数量	20	提交工单。
一个账号在一个可用区的可用区级预留实例券数量	20	提交工单。
预留实例券支持的实例规格	支持使用预留实例券的规格族包括： • 通用型：g7、g6e、g6、g5、g5ne、sn2ne • 计算型：c7、c6e、c6、c5、ic5、sn1ne • 内存型：r7、r6e、r6、r5、re6、re4、se1ne • 大数据型：d2s • 本地SSD型：i3、i3g、i2、i2g、i2gne • 高主频型：hfg7、hfc7、hfr7、hfg6、hfc6、hfr6、hfg5、hfc5 • GPU计算型：gn7、gn6i、gn6e、gn6v、gn5、gn5i • 弹性裸金属服务器：ebmgn7、ebmgn6i、ebmgn6e、ebmg6、ebmc6、ebmr6、ebmhfg6、ebmhfc6、ebmhfr6 • 突发型：t6、t5	无

> **说明** 更多详情，请参见预留实例券使用限制。

节省计划

限制项	限制	提升限额方式
一个账号的节省计划数量	40	无
节省计划支持的实例规格	已停售的系列I实例规格不支持节省计划，包括：t1、s1、s2、s3、m1、m2、c1、c2。	无

块存储

限制项	限制	提升限额方式
创建按量付费云盘的用户限制	• 账号必须实名认证。 • 开通按量付费ECS资源时，您的阿里云账户余额（即现金余额）和代金券的总值不得小于100.00元人民币。	无
按量云盘的总数量	在ECS管理控制台查看资源配额。具体操作，请参见查看和提升资源配额。	无
单实例系统盘数量	1块	无
单实例数据盘数量	64块 **说明** 创建实例时最多可挂载16块数据盘，如果实例需要更多数据盘，需要在创建实例后继续挂载。不同实例规格支持挂载的最多云盘数量不同，更多信息，请参见实例规格族。	无
按量高效云盘容量限额	在ECS管理控制台查看资源配额。具体操作，请参见查看和提升资源配额。	无
按量SSD云盘容量限额	在ECS管理控制台查看资源配额。具体操作，请参见查看和提升资源配额。	无
按量ESSD云盘容量限额	在ECS管理控制台查看资源配额。具体操作，请参见查看和提升资源配额。	无
单块普通云盘容量	5 GiB~2,000 GiB	无
单块SSD云盘容量	20 GiB~32,768 GiB	无
单块高效云盘容量	20 GiB~32,768 GiB	无
单块ESSD云盘容量	• PL0：40 GiB~32,768 GiB • PL1：20 GiB~32,768 GiB • PL2：461 GiB~32,768 GiB • PL3：1,261 GiB~32,768 GiB	无
单块ESSD AutoPL云盘容量	40 GiB~32,768 GiB	无
单块ESSD PL-X云盘容量	40 GiB~32,768 GiB	无
单块SSD本地盘容量	5 GiB~800 GiB	无
单实例SSD本地盘总容量	1,024 GiB	无

限制项	限制	提升限额方式
系统盘单盘容量限制	• Windows Server：40 GiB~500 GiB • Red Hat：40 GiB~500 GiB • CoreOS与FreeBSD：30 GiB~500 GiB • 其他Linux：20 GiB~500 GiB	无
本地盘实例是否可以自行挂载新的本地盘	不允许。	无
本地盘实例是否支持变更配置	仅允许变更带宽。	无
系统盘挂载点范围	/dev/vda	无
数据盘挂载点范围	/dev/vd[b-z]	无

说明 块存储按照二进制单位计算。二进制单位用于表示1024进位的数据大小。例如，1 GiB=1024 MiB。

存储容量单位包

限制项	限制	提升配额方式
一次可以购买的存储容量单位包最大容量	50 TiB	提交工单。
同一地域内最大支持购买SCU数量	100个	无
存储容量单位包支持的产品类型	• ESSD云盘、SSD云盘、高效云盘和普通云盘 • 容量型NAS和性能型NAS • 普通快照 • 标准型OSS、低频型OSS和归档型OSS • 混合云存储HBR的备份库存储容量	无

快照

限制项	限制	提升限额方式
每块云盘可以保留的手动快照个数	256	无
每块云盘可以保留的自动快照个数	1000	无
一个账号在一个地域可以保留的自动快照策略数量	100	无

镜像

限制项	限制	提升限额方式
当前账户镜像数量	在ECS管理控制台查看资源配额。具体操作，请参见查看和提升资源配额。	在ECS管理控制台申请提升资源配额。具体操作，请参见查看和提升资源配额。
单个镜像最多可共享的用户数量	50	提交工单。
镜像与实例规格的限制	4 GiB及以上内存的实例规格不能使用32位镜像。	无

SSH密钥对

限制项	限制	提升限额方式
一个账号在每个地域的SSH密钥对配额	500	无
支持SSH密钥对的实例规格	不支持系列I的非I/O优化实例	无
支持SSH密钥对的镜像类型	仅支持Linux系统	无

公网带宽

自2020年11月27日起，创建和变配ECS实例时带宽峰值受账户限速策略影响。如需更大带宽峰值，请提交工单。具体限速策略如下：

- 单个地域下，所有按使用流量计费ECS实例的实际运行带宽峰值总和不大于5 Gbit/s。
- 单个地域下，所有按固定带宽计费ECS实例的实际运行带宽峰值总和不大于50 Gbit/s。

限制项	限制	提升限额方式
单实例入带宽峰值	• 当所购出带宽峰值小于等于10 Mbit/s时，阿里云会分配10 Mbit/s入方向带宽。 • 当所购出带宽峰值大于10 Mbit/s时，阿里云会分配与购买的出带宽峰值相等的入方向带宽。	无

单实例出带宽峰值	• 按使用流量计费：100 Mbit/s
	• 按固定带宽计费：
	  ◦ 包年包月实例：200 Mbit/s
	  ◦ 按量付费实例：100 Mbit/s
		无

单实例更换分配的公网IP地址的限制	新建实例六小时内可以更换公网IP地址，一台实例最多可以更换三次。	无

注意 **按使用流量计费**模式下的出入带宽峰值都是带宽上限，不作为业务承诺指标。当出现资源争抢时，带宽峰值可能会受到限制。如果您的业务需要有带宽的保障，请使用**按固定带宽计费**模式。

安全组

限制项	普通安全组限制	企业安全组限制
安全组总数量上限	在ECS管理控制台查看资源配额 ①。具体操作，请参见查看和提升资源配额。	与普通安全组相同
一个经典网络类型的安全组能容纳的经典网络类型ECS实例数量	1000 ②	不支持经典网络
一个专有网络VPC类型的安全组能容纳的VPC类型ECS实例数量	不固定，受安全组能容纳的私网IP地址数量影响。	无限制
一台ECS实例可以加入的安全组数量	5 如需提高上限，请提交工单，可以增加到10个或者16个安全组。	与普通安全组相同
一台ECS实例的每张弹性网卡可以加入的安全组数量		
一个安全组最大规则数量（包括入方向规则与出方向规则）	200 ③	与普通安全组相同
一张弹性网卡在所有已加入的安全组中的最大规则数量（包括入方向规则与出方向规则）	1000	与普通安全组相同
一个专有网络VPC类型的安全组能容纳的私网IP地址数量	2000 ④	65536
公网访问端口	出方向的SMTP默认端口25默认受限，而且不能通过安全组规则打开。关于如何申请解封，请参见TCP 25端口控制台解封申请。	与普通安全组相同

- ① 华东 1（杭州）、华东 2（上海）、华北 1（青岛）、华北 2（北京）、华南 1（深圳）、中国（香港）、美国（硅谷）、新加坡，以上地域共享可以创建的安全组数量限制。即一个账号在这些地域创建安全组数量的总上限为100。
- ② 如果您有超过1000台经典网络类型ECS实例需要内网互访，可以将ECS实例分配到多个安全组内，并通过互相授权的方式允许互访。
- ③ 如果您提高了一台ECS实例的可以加入的安全组数量限制，相应的安全组最大规则数量会下降。该实例可加入的安全组数量乘以每个安全组入方向和出方向规则最大数量必须小于等于1000。即5*200=1000、10*100=1000、16*60≤1000。

  如果安全组规则中引用前缀列表，则前缀列表的最大条目容量会占用安全组规则数量。例如，某前缀列表的最大条目容量设置为100个，如果安全组规则中引用该前缀列表，会占用该安全组100个规则额度，与该前缀列表中已有的条目数无关。
- ④ 如果您有超过2000个私网IP需要内网互访，可以将这些私网IP的ECS实例分配到多个安全组内，并通过互相授权的方式允许互访。

前缀列表

限制项	限制	提升额度方式
一个账号在每个地域的前缀列表数量上限	100	无
一个前缀列表中设置的条目数量上限	200	无
一个前缀列表的关联资源数量上限	1000	无

网络连通性诊断

限制项	限制	提升限额方式
单个地域下诊断线路的最大数量	100	无
单个地域下诊断任务的最大数量	1000	无
单个地域下同时执行的诊断任务的最大数量	5	无

部署集

限制项	限制	提升限额方式
部署集数量	在ECS管理控制台查看资源配额。具体操作，请参见查看和提升资源配额。	在ECS管理控制台申请提升资源配额。具体操作，请参见查看和提升资源配额。
一个部署集内能容纳的实例数量	一个可用区内最多允许20台实例，一个地域内允许20*（可用区数量）的实例数量。	无
部署集内能创建的实例规格	• c7、g7、r7、c6、g6、r6、c5、g5、r5、c6e、g6e、r6e、c7se、g7se、r7se、r6se、c7t、g7t、r7t、c7a、g7a、r7a、c6a、g6a、r6a、g5ne、re6、re4、ic5 • hfc7、hfg7、hfr7、hfc6、hfg6、hfr6、hfc5、hfg5 • d2s、d2c、d1、d1ne、d1-c14d3、d1-c8d3 • i3、i3g、i2、i2g、i2ne、i2gne、i1 • se1ne、sn1ne、sn2ne、se1 • ebmg5、sccgn6、scch5、sccg5、scch5s、sccg5s • s6、t6、xn4、mn4、n4 • gn6i	无

云助手

限制项	限制	提升限额方式
云助手命令总数量上限	在ECS管理控制台查看资源配额。具体操作，请参见查看和提升资源配额。	无
云助手每天可执行调用的次数	在ECS管理控制台查看资源配额。具体操作，请参见查看和提升资源配额。	无
云助手任务执行记录保留时长	14天	无
云助手任务执行记录保留数量上限	100000	无

弹性网卡

限制项	限制	提升限额方式
弹性网卡（辅助网卡）创建限额	在ECS管理控制台查看资源配额。具体操作，请参见查看和提升资源配额。	无

标签

限制项	限制	提升限额方式
单台实例允许绑定的标签数量	20	无

API

限制项	限制	提升限额方式
CreateInstance调用次数	一分钟内最多200次	提交工单。

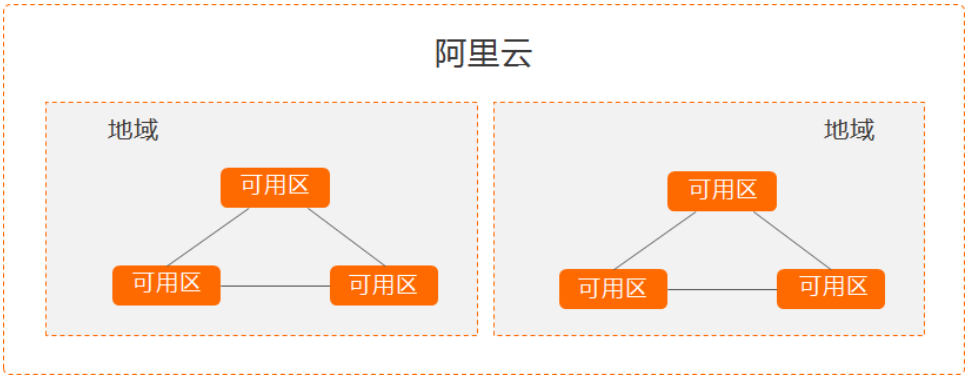
说明 专有网络VPC的产品限制请参见限制与配额。

7.地域和可用区

本文提供完整的阿里云地域和可用区列表。

每个地域完全独立，不同地域的可用区完全隔离，但同一个地域内的可用区之间使用低时延链路相连。

地域和可用区之间的关系如下图所示。



有关地域和可用区的更多信息，请参见阿里云全球基础设施。

地域

地域是指物理的数据中心。资源创建成功后不能更换地域。当前所有的地域、地域所在城市和Region ID的对照关系如下表所示。

> **说明** 不同产品可选择的地域有所不同，请以各产品实际支持的地域为准。

- 中国内地

地域名称	所在城市	Region ID	可用区数量
华北 1	青岛	cn-qingdao	2
华北 2	北京	cn-beijing	12
华北 3	张家口	cn-zhangjiakou	3
华北 5	呼和浩特	cn-huhehaote	2
华北 6	乌兰察布	cn-wulanchabu	3
华东 1	杭州	cn-hangzhou	8
华东 2	上海	cn-shanghai	11
华南 1	深圳	cn-shenzhen	6
华南 2	河源	cn-heyuan	2
华南 3	广州	cn-guangzhou	2
西南 1	成都	cn-chengdu	2

地域名称	所在城市	Region ID	可用区数量
华东 5	南京（属于本地地域，邀测中）	cn-nanjing	1

- 其他国家和地区

所在国家	所在城市	Region ID	可用区数量
中国	香港	cn-hongkong	3
新加坡	新加坡	ap-southeast-1	3
澳大利亚	悉尼	ap-southeast-2	2
马来西亚	吉隆坡	ap-southeast-3	2
印度尼西亚	雅加达	ap-southeast-5	3
菲律宾	马尼拉	ap-southeast-6	1
泰国	曼谷	ap-southeast-7	1
印度	孟买	ap-south-1	2
日本	东京	ap-northeast-1	2
韩国	首尔	ap-northeast-2	1
美国	硅谷	us-west-1	2
美国	弗吉尼亚	us-east-1	2
德国	法兰克福	eu-central-1	3
英国	伦敦	eu-west-1	2
阿联酋	迪拜	me-east-1	1

选择地域时，您需要考虑以下几个因素：

- 地理位置

  请根据您以及您目标用户所在的地理位置选择地域

  - 中国内地

    一般情况下建议选择和您目标用户所在地域最为接近的数据中心，可以进一步提升用户访问速度。不过，在基础设施、BGP网络品质、服务质量、云服务器操作使用与配置等方面，阿里云中国内地地域没有太大区别。BGP网络可以保证中国内地全部地域的快速访问。

  - 其他国家及地区

    其他国家及地区提供的带宽主要面向非中国内地的用户。如果您在中国内地，使用这些地域会有较长的访问延迟，不建议您使用。

- 阿里云产品之间的关系

  如果多个阿里云产品一起搭配使用，需要注意：

  - 不同地域的云服务器ECS、关系型数据库RDS、对象存储服务OSS内网不互通。
  - 不同地域之间的云服务器ECS不能跨地域部署负载均衡，即在不同的地域购买的ECS实例不支持跨地域部署在同一负载均衡实例下。

- 资源的价格

  不同地域的资源价格可能有差异，请参见阿里云产品定价页面。

- 关于经营许可证备案

  选择地域时您需要考虑某些地区的特殊要求。如您在中国内地地域购买了ECS实例，并用于Web服务器，您需要完成经营许可证备案。

  如您有办理经营许可证备案的需求，请您重点关注：

  - 北京地区企业，请选择购买的地域为**华北 2**。
  - 广东地区企业，请选择购买的地域为**华南 1**。

  > **说明** 各省通信管理局对经营性备案的审批要求不同，如有变化，请以当地管局经营性备案网站公示内容为准。

可用区

可用区（Availability Zone，简称AZ）是指在同一地域内，电力和网络互相独立的物理区域。同一可用区内实例之间的网络延时更小。

在同一地域内可用区与可用区之间内网互通，可用区之间能做到故障隔离。是否将实例放在同一可用区内，主要取决于对容灾能力和网络延时的要求。

- 如果您的应用需要较高的容灾能力，建议您将实例部署在同一地域的不同可用区内。
- 如果您的应用要求实例之间的网络延时较低，建议您将实例创建在同一可用区内。

8.FPGA实例规格族

8.1. FPGA计算型

本章节介绍云服务器ECS FPGA计算型实例规格族的特点，并列出了具体的实例规格。

- 主售（推荐类型）

  FPGA计算型实例规格族f3

- 在售（如果售罄，建议使用主售的规格族）

  FPGA计算型实例规格族f1

FPGA计算型实例规格族f3

f3的特点如下：

- 采用Xilinx 16nm Virtex UltraScale+器件VU9P
- 计算：
  - 处理器与内存配比为 1:4
  - 处理器：2.5 GHz主频的Intel® Xeon® Platinum 8163（Skylake）
- 存储：
  - I/O优化实例
  - 仅支持SSD云盘和高效云盘
- 网络：
  - 实例网络性能与计算规格对应（规格越高网络性能越强）
- 适用场景：
  - 深度学习推理
  - 基因组学研究
  - 数据库加速
  - 图片转码，例如JPEG转WebP
  - 实时视频处理，例如H.265视频压缩

f3包括的实例规格及指标数据如下表所示。

实例规格	vCPU	内存（GiB）	FPGA	网络带宽（Gbit/s）	网络收发包PPS（万）	多队列	弹性网卡	单网卡私有IP
ecs.f3-c4f1.xlarge	4	16.0	1 * Xilinx VU9P	1.5	30	2	3	10
ecs.f3-c8f1.2xlarge	8	32.0	1 * Xilinx VU9P	2.5	50	4	4	10

实例规格	vCPU	内存（GiB）	FPGA	网络带宽（Gbit/s）	网络收发包PPS（万）	多队列	弹性网卡	单网卡私有IP
ecs.f3-c16f1.4xlarge	16	64.0	1 * Xilinx VU9P	5.0	100	4	8	20
ecs.f3-c16f1.8xlarge	32	128.0	2 * Xilinx VU9P	10.0	200	8	8	20
ecs.f3-c16f1.16xlarge	64	256.0	4 * Xilinx VU9P	20.0	250	16	8	20
ecs.f3-c22f1.22xlarge	88	336.0	4 * Xilinx VU9P	30.0	450	16	8	20

说明

- 您可以前往ECS实例可购买地域，查看实例在各地域的可购情况。
- 指标的含义请参见实例规格族。

FPGA计算型实例规格族f1

f1的特点如下：

- 采用Intel® ARRIA® 10 GX 1150计算卡
- 计算：
  - 处理器：2.5 GHz主频的Intel® Xeon® E5-2682 v4（Broadwell）
  - 处理器与内存配比为 1:7.5
- 存储：
  - I/O优化实例
  - 仅支持SSD云盘和高效云盘
- 网络：
  - 支持IPv6
  - 实例网络性能与计算规格对应（规格越高网络性能越强）
- 适用场景：
  - 深度学习推理
  - 基因组学研究
  - 金融分析
  - 图片转码
  - 实时视频处理及安全等计算工作负载

f1包括的实例规格及指标数据如下表所示。

实例规格	vCPU	内存（GiB）	FPGA	网络带宽（Gbit/s）	网络收发包PPS（万）	多队列	弹性网卡	单网卡私有IP
ecs.f1-c8f1.2xlarge	8	60.0	Intel ARRIA 10 GX 1150	3.0	40	4	4	10
ecs.f1-c8f1.4xlarge	16	120.0	2 * Intel ARRIA 10 GX 1150	5.0	100	4	8	20
ecs.f1-c28f1.7xlarge	28	112.0	Intel ARRIA 10 GX 1150	5.0	200	8	8	20
ecs.f1-c28f1.14xlarge	56	224.0	2 * Intel ARRIA 10 GX 1150	10.0	200	14	8	20

说明

- 您可以前往ECS实例可购买地域，查看实例在各地域的可购情况。
- 指标的含义请参见实例规格族。