

高性能计算集群







【版权声明】

©2013-2025 腾讯云版权所有

本文档(含所有文字、数据、图片等内容)完整的著作权归腾讯云计算(北京)有限责任公司单独所有,未经腾讯云 事先明确书面许可,任何主体不得以任何形式复制、修改、使用、抄袭、传播本文档全部或部分内容。前述行为构成 对腾讯云著作权的侵犯,腾讯云将依法采取措施追究法律责任。

【商标声明】

🕗 腾讯云

及其它腾讯云服务相关的商标均为腾讯云计算(北京)有限责任公司及其关联公司所有。本文档涉及的第三方主体的 商标,依法由权利人所有。未经腾讯云及有关权利人书面许可,任何主体不得以任何方式对前述商标进行使用、复 制、修改、传播、抄录等行为,否则将构成对腾讯云及有关权利人商标权的侵犯,腾讯云将依法采取措施追究法律责 任。

【服务声明】

本文档意在向您介绍腾讯云全部或部分产品、服务的当时的相关概况,部分产品、服务的内容可能不时有所调整。 您所购买的腾讯云产品、服务的种类、服务标准等应由您与腾讯云之间的商业合同约定,除非双方另有约定,否则, 腾讯云对本文档内容不做任何明示或默示的承诺或保证。

【联系我们】

我们致力于为您提供个性化的售前购买咨询服务,及相应的技术售后服务,任何问题请联系 4009100100或 95716。



文档目录

操作指南

管理高性能计算集群

- 集群自助检测
- GPU 型实例安装 nvidia-fabricmanager 服务

GPU 型实例安装 TCCL

- GPU 型实例安装 RDMA 毫秒级监控组件
- GPU 型实例监控和告警



操作指南 管理高性能计算集群

最近更新时间: 2025-05-07 16:54:12

操作场景

高性能计算集群用于实现高性能计算实例的 RDMA 网络隔离管理。

- 同集群内,实例 RDMA 网络互联互通。
- 跨集群间,实例 RDMA 网络相互隔离。

在创建高性能计算**实例**前,您需要首先创建高性能计算**集群。**后续在创建实例时通过选择已有的高性能计算集群,可 实现集群内节点高速计算网络互通。

本文介绍高性能计算集群常见的相关操作,例如对集群的创建、修改、扩容、删除等,以下是具体操作步骤:

创建高性能计算集群

- 1. 登录 云服务器控制台,在左侧导航栏选择高性能计算集群。
- 2. 在高性能计算集群列表页面中,按需选择地域。
- 3. 单击新建。

新建 删除					集群ID		Q
集群ID/名称	描述	可用区 了	实例数	标签(key:valu	e)	操作	
hpc-	-	上海五区	1	\bigcirc		编辑标签 扩容 查看详情	

4. 在弹出的创建集群窗口中,选择填写可用区、集群名称、集群描述信息。



创建集群					
可用区 *	上海二区	上海三区	上海五区	上海八区	
集群名称 *					
	您还可以输入60-	个字符			
集群描述					
	您还可以输入250	6个字符			
标签(选填)	标签键	~	标签值	``	0
	+ 添加 🕥 🕅	圭值粘贴板			
			确定	取消	
			HUN	-97/13	

5. 确认信息无误后,单击确定,等待集群创建完成。

修改高性能计算集群信息

1. 登录 云服务器控制台,在左侧导航栏选择高性能计算集群。

2. 在高性能计算集群页面,选择需要修改的集群名称或描述右侧的 🞤 ,如下图所示。

新建 删除				集群ID	Q
集群ID/名称	描述	可用区 了	实例数	标篮(key:value)	操作
hpc- A100-sl52	-0	上海五区	1	\bigtriangledown	编辑标签 扩容 查看详情

3. 在弹出的修改名称或修改描述窗口中,输入新的集群名称和集群描述,单击确定,完成操作。



收起				
ID/쉮	名称		可用区	
hpc	•		上海五区	
名称				
<u>ы</u> 137	您还可以输入52个字符			
		确定	取消	

扩容高性能计算集群

- 1. 登录 云服务器控制台,在左侧导航栏选择高性能计算集群。
- 2. 在高性能计算集群页面,选择需要扩容的集群单击**扩容**,进入实例购买页。

新建制除					集群ID	Q
集群ID/名称	描述	可用区 了	实例数	标签(key:val	ue)	操作
hpc-	-	上海五区	1	\triangleleft		编辑标签扩容查看详情

3. 参见 购买方式 完成扩容操作。

删除高性能计算集群

- ⚠ 注意: 若高性能计算集群已部署实例,则该集群无法删除。需销毁集群内全部实例后,才可删除集群。
- 1. 登录 云服务器控制台,在左侧导航栏选择高性能计算集群。
- 2. 在高性能计算集群页面,按需勾选一个或多个集群后,单击删除。

新建 删除				集群ID	Q
■ 集群ID/名称	描述	可用区 ℃	实例数	标签(key:value)	操作
► hpc	-	上海五区	1	\bigcirc	编辑标签 扩容 查看详情
hpc-		上海三区	0	Ø	编辑标签 扩容 查看详情



3. 在弹出的窗口中确认信息,单击确定,完成操作。

集群自助检测

最近更新时间: 2025-01-06 10:56:32

概述

高性能计算集群的自助检测(即集群一致性检测)功能提供集群维度的实例检测,您可以检测集群中所有实例的硬件 和软件状态。您可通过该功能及时发现并解决集群实例的相关问题。

操作场景

以下两种场景推荐使用集群自助检测:

- **集群故障排查**:在日常运维过程中,您可以使用集群自助检测功能,检测集群中所有实例的硬件和软件状态,并 根据相应建议对异常情况进行处理。
- 大规模 AI 模型训练环境检测:集群训练需要保障硬件、GPU 驱动、CUDA、NCCL 和 RDMA 等配置状态的可用性和一致性,自助检测功能提供集群维度的实例检测能力,保障训练正常运行。

检测项说明

集群自助检测功能支持诊断硬件和软件配置的一致性并提供诊断报告。

集群训练环境

检测项	检测说明	风险等 级	解决方案
集群训练环境 GPU 型号	实例的 GPU 型号	异常	更换为同种型号 GPU 机型。
集群训练环境 CPU 型号	实例的 CPU 型号	异常	更换为同种型号 CPU 机型。
集群训练环境 GPU 卡数	实例的 GPU 卡总数	异常	实例可能存在 GPU 卡异常, 请您关注 CVM 控制台维修 任务中是否存在 维修任务并 授权腾讯云 进行异常显卡的 硬件维护。
集群训练环境内核版 本	实例的操作系统内核版本	酸牛	操作系统内核版本不一致,建 议更换同操作系统保持内核一 致。
集群训练环境 OS 版 本	实例的操作系统版本	异常	建议更换操作系统保持版本一 致。



集群训练环境 OFED 版本	实例的网卡驱动版本	异常	建议使用 HCC 公有镜像,已 安装 OFED 驱动。
集群训练环境 GPU 驱动版本	实例的 GPU 驱动版本	异常	建议保持 GPU 驱动版本一 致,您可以使用 <mark>自动安装驱</mark> 动功能 在创建或重装实例时 指定 GPU 驱动版本。
集群训练环境 CUDA Runtime 版本	实例的 CUDA Runtime 版本	异常	建议保持 CUDA 版本一致, 您可以使用 <mark>自动安装驱动功</mark> <mark>能</mark> 在创建或重装实例时指定 CUDA 版本。
集群训练环境 CUDA Driver 版本	实例的 CUDA 驱动版本	异常	建议保持 CUDA 版本一致, 您可以使用 <mark>自动安装驱动功</mark> <mark>能</mark> 在创建或重装实例时指定 CUDA 版本。
集群训练环境 GDR 状态	实例是否加载 nvidia_peermem 模 块。 nvidia_peermem 是 NVIDIA 提供的一个内核 模块,通常用于支持设备 间直接内存访问(peer- to-peer memory access),在高性能计 算(HPC)和深度学习 应用中用于优化 GPU 间 的数据传输效率,对多机 多卡训练性能有影响。	异常	<pre>默认自动安装驱动后自动加载 nvidia_peermem 模块, 若检测未加载 nvidia_peermem 模块, 请确保 GDR 挂载状态,参考 如下步骤处理: modprobe nvidia_peermem && echo "modprobe nvidia_peermem" >>/etc/rc.local</pre>
集群训练环境 NCCL 版本	实例的 NCCL 版本。 NCCL (NVIDIA Collective Communications Library) 是 一个 NVIDIA GPU 集 合通信库,用于多 GPU 或多个节点通信,是 AI 分布式训练的必选软件。	异常	建议所有实例 NCCL 版本保 持一致,安装方法请参见 NVIDIA NCCL 官方说明 。
集群训练环境 NCCL-plugin 版 本	实例的 NCCL−plugin 版本。NCCL−plugin 是一款针对腾讯云星脉网 络架构的高性能定制加速	异常	建议使用统一版本 NCCL– plugin,详情请参见 GPU 型实例安装 TCCL 说明 。



通信库插件,依托星脉网 络硬件架构,为 AI 大模 型训练提供更高效的网络 通信性能,详情请参见 GPU 型实例安装 TCCL 说明 。	
---	--

集群 RDMA

检测项	检测说明	风险 等级	解决方案
集群 RDMA Bonding 状态	检查特定的网卡绑定接 口(bonding interface)为 up 状 态的网卡数量。	异常	若存在网卡绑定状态未开启,请执行 ifup bondX (bondX 是要重启 的网卡的名称,例如 bond0) 将网 卡开启,以便 RDMA 网络正常使 用。
集群 RDMA MTU 配置	当前网络接口的最大传 输单元(MTU)。 MTU 是指在一个网络 数据包中,数据部分的 最大字节数。不同的网 络接口和网络环境可能 有不同的最佳 MTU 值。	异常	若检测 MTU 值异常,建议调整 MTU 值为9100。
集群 RDMA 网卡 速率	了解网络接口的速度, 以确保网络设备和连接 正常工作。	异常	若检测网卡速率不符合预期,请关注 CVM 控制台维修任务中是否存在 <mark>维</mark> 修任务并授权腾讯云 进行异常网卡的 维护。
集群 RDMA rp_filter 状态	读取或设置 Linux 系 统中某个网络接口的反 向路径过滤 (Reverse Path Filtering)。	异常	若检测到 rp_filter 配置异常,请联 系腾讯云处理。
集群 RDMA QoS 配置	读取 QoS 配置信息, 配置影响数据包的处理 优先级和质量,可能具 有 0 到 255 之间的数 值,用来表示不同的优 先级或服务质量,一般	异常	若检测到您的端口 QoS 配置问题, 请执行如下命令修复: bash /usr/local/qcloud/rdma/ds cp.sh

腾讯云			7
	是 160。所有实例配置 需要保持一致。		
集群 RDMA CRC 异常	查询并过滤特定网络接 口的统计信息,特别是 接收数据包时的 CRC 错误统计。CRC 是一 种用于检测错误的校验 方法,如果一个数据包 的 CRC 校验失败,表 明该数据包在传输过程 中可能被损坏。	警告	若该计数持续有增加,请联系腾讯云 处理。
集群 RDMA ARP 双发检查	用于配置和管理网络接 口绑定(bonding) 设置中的广播 ARP (Address Resolution Protocol)检测功 能。广播 ARP 检测是 网络绑定中的一种链路 监控方法。通过发送 ARP 请求并监听响 应,来检测链路的可用 性。	异常	若检测到 ARP 配置问题,请联系腾 讯云处理。
集群 RDMA rpg_time_reset 拥塞控制参数	RoCE (RDMA over Converged Ethernet)网络接口 的 RPG (Rate- Parity Group)时间 重置相关的设置。 RPG 时间重置是 RoCE 网络接口的一 种功能,用于在网络链 路出现异常情况时执行 重置操作,以确保网络 的稳定性和可靠性。重 置操作可能涉及重新初 始化或重新同步网络接 口的一些参数或状态。	警告	拥塞控制参数不一致,如观测到业务 训练速度变慢,请联系腾讯云处理。
集群 RDMA rpg_ai_rate 拥 塞控制参数	用于配置 RoCE (RDMA over Converged Ethernet)网络接口	酸牛	拥塞控制参数不一致,如观测到业务 训练速度变慢,请联系腾讯云处理。

	中 RPG (Rate- Parity Group)的自 适应传输速率。自适应 传输速率是 RoCE 网 络接口的一种功能,允 许在传输数据时动态调 整传输速率,以根据网 络负载和性能需求实现 最佳的数据传输效率。		
集群 RDMA rpg_byte_reset 拥塞控制参数	RoCE (RDMA over Converged Ethernet) 网络接口 的 Rate-Parity Group (RPG) 的字 节重置设置。	警告	拥塞控制参数不一致,如观测到业务 训练速度变慢,请联系腾讯云处理。
集群 RDMA rate_reduce_m onitor_period拥 塞控制参数	RoCE 网络接口的 Rate-Parity Group (RPG)的速率降低 监控周期(Rate Reduce Monitor Period)。	警告言口	拥塞控制参数不一致,如观测到业务 训练速度变慢,请联系腾讯云处理。

集群单机检测

检测项	检测说明	风险 等级	解决方案
集群单机检测 GPU 硬件降速	检测硬件热量减速功 能当前是否处于激活 状态	异常	若检测到实例 GPU 被降速,请您关 注 CVM 控制台维修任务中是否存在 <mark>维修任务并授权腾讯云</mark> 进行异常显 卡的硬件维护。
集群单机检测 GPU 软件降速	检测软件热量减速功 能当前是否处于激活 状态	异常	若检测到实例存在 GPU 运行时 被降 速,请您关注CVM 控制台维修任务 中是否存在 <mark>维修任务并授权腾讯云</mark> 进行异常显卡的硬件维护。
集群单机检测 GPU XID 异常	GPU 是否存在 XID 报错	异常	若检测到实例存在 GPU XID 错误, 请您关注 CVM 控制台维修任务中是 否存在 <mark>维修任务并授权腾讯云</mark> 进行 异常显卡的硬件维护。



集群单机检测 NVLink 激活失败	检测 NVLink 激活 失败个数	异常	若检测失败,请确认 nvidia- fabricmanager 服务是否正常。 若服务异常,请启动服务再验证是否 正常;若服务正常,请重启实例再验 证是否正常。
集群单机检测 NVSwitch 初始化 失败	检测 NVSwitch 初 始化失败	异常	请重启机器再检测是否可以恢复,若 无法恢复请联系腾讯云处理。
集群单机检测 FabricManager 服务报错	检测 FabricManager 服务报错个数	异常	若实例没有启动 nvidia− fabricmanager 服务,可参见 安 装 nvidia−fabricmanager 服 务 。
集群单机检测 NVLink P2P 状态 异常	检测 NVLink P2P 状态异常	异常	若检测到实例存在 GPU 异常,请您 关注 CVM 控制台维修任务中是否存 在 <mark>维修任务并授权腾讯云</mark> 进行异常 显卡的硬件维护。
集群单机检测 PCle 链路故障	检测 PCle 链路故障 个数	异常	若检测到实例存在 GPU 异常,请您 关注 CVM 控制台维修任务中是否存 在 <mark>维修任务并授权腾讯云</mark> 进行异常 显卡的硬件维护。
集群单机检测 Row remapper 状态	检测 Row remapper 失败的 GPU 故障的个数	异常	若检测到实例存在 GPU 异常,请您 关注 CVM 控制台维修任务中是否存 在 <mark>维修任务并授权腾讯云</mark> 进行异常 显卡的硬件维护。
集群单机检测 SRAM UCE 超标 的 GPU 故障	检测 SRAM UCE 超标的 GPU故障个 数	异常	若检测到实例存在 GPU 异常,请您 关注 CVM 控制台维修任务中是否存 在 <mark>维修任务并授权腾讯云</mark> 进行异常 显卡的硬件维护。
集群单机检测驱动状 态	检测驱动状态	异常	若检测到实例存在 GPU 异常,请您 关注 CVM 控制台维修任务中是否存 在 <mark>维修任务并授权腾讯云</mark> 进行异常 显卡的硬件维护。
集群单机检测 GPU 初始化状态	检测 GPU 初始化状 态	异常	若检测到实例存在 GPU 异常,请您 关注 CVM 控制台维修任务中是否存 在 <mark>维修任务并授权腾讯云</mark> 进行异常 显卡的硬件维护。



集群单机检测 InfoROM 状态	检测 InfoROM 状态	异常	若检测到实例存在 GPU 异常,请您 关注 CVM 控制台维修任务中是否存 在 <mark>维修任务并授权腾讯云</mark> 进行异常 显卡的硬件维护。	
----------------------	---------------	----	--	--

前提条件

实例需安装自动化助手后才可以使用集群自助检测能力。如未安装,可参见 安装自动化助手客户端 进行安装。

相关操作

集群自助检测功能支持诊断硬件和软件配置一致性并提供诊断报告。

启动检测

- 1. 登录 云服务器控制台, 在左侧导航栏选择高性能计算集群。
- 2. 在高性能计算集群列表页面中,选择集群所在地域。
- 3. 单击集群 ID,进入集群详情页面。

高性能计算集群	S 北京 其它	2地域 ▼		
Refit				
新建 删除				集群ID
集群ID/名称		描述	可用区 🔻	实例数
hpc-attaches		zp_test	北京六区	1, contraction
hpc-rp= P Fal		hcctool	北京六区	3

4. 在**自助检测**页签,单击**集群自助检测**。



- hpc-speeling	🕯 (default)				
💊 default Ø	same and a second				
hpc	• 例数 3 创建时间	2023-11-28 15:27:22	描述 hcctool ,	C and a start of the	
集群 资 源 目 集群自助检测					
报告ID		检测时间			检测状态 ▼
				 	昏无数据

5. 选择本次检测的实例,单击**开始检测**。

脸测类型						
集群一致性检测						
E分布式训练任务启动前	前,对集群环境进行一	-致性检测,分析可能	能影响训	练性能或者导致训	练异常的因素,方便故	章定位
择实例				已选择 (3)		
		Q	01180	ID/实例名	。 IP地址	
✓ ID/实例名	实例类型	IP地址		5	10 ^{ml}	10010
LeonardTest1	HCCPNV5v.43			LeonardTest1		8
APR.	XLARGE1939		77130			
				\$		
LeonardTest2	HCCPNV5v.43		↔	LeonardTest2		
	XLARGE1939					
✓ hubery-test	HCCPNV5v.43			hubery-test		
	XLARGE1939					



← hpc-	🍽 🎫 / 集群自助检测			
٢	正在检测 关闭页面不会终止检测		普通警告 0	严重警告 O
fantart/100				
检测结果 检测详情 ~				

检测完成,页面会显示检测结果。

← hpc					
存在异常 部分检测项配置错误,影	响集群任务运行!		普通警告 1	严重警 50 ⁻¹⁰⁻¹⁰⁻¹⁰ 3	音 Contractionの
检测结果					
4个优化项 ~ 集群训练环境OS版本 检测到您集群实例的OS版本不一致,请	使用同OS版本实例进行训	练			
集群训练环境GPU驱动版本 检测到您集群实例的GPU驱动版本不一到	牧,请确保训练环境驱动 版	反本一致			
集群训练环境CUDA Runtime版本 检测到您集群实例的Cuda Runtime版本	不一致,请确保训练环境	版本一致			
集群训练环境内核版本					

检测到您集群实例的内核版本不一致,请使用同内核版本实例进行训练

查看历史报告

- 1. 登录 云服务器控制台,在左侧导航栏选择**高性能计算集群。**
- 2. 在**高性能计算集群**列表页面中,选择集群所在**地域**。



3. 单击集群 ID,进入集群详情页面。

高性能计算集群	北京 其它 其	3地域 ▼		
40 ¹⁶ 0				
新建删除				集群ID
集群ID/名称		描述	可用区 ▼	实例数
hpc-stanting Maint		zp_test	北京六区	1, and the second
hpc	Fartlanding	hcctool	北京六区	3

4. 单击**查看报告**,即可查看历史检测报告。

+ hpc-mail@wak	(default)					
default a						
可用区 北京六区 实例数	3 创建时间 2023-	11-28 15:27:22 描述	hcctool 🖉			
集群资源 自助检测	And the other					
集群自助检测 报告ID		检测时间		检测状态 ▼		操作
cl-cb745a32-	ever Jahrend Sinte	2024-12-12 19:43:	39	检测完成	dentheo	查看报告
cl-605795c3	0/00-1/10210110-000	2024-12-12 19:39:	49	检测完成		查看报告



GPU 型实例安装 nvidia-fabricmanager 服务

最近更新时间: 2025-05-06 10:26:22

操作场景

高性能计算集群实例搭载了 A100/A800/H800 GPU 并支持 NVLink & NVSwitch,需额外安装与驱动版本对 应的 nvidia-fabricmanager 服务使 GPU 卡间能够互联。如果您使用该实例,请参考本文安装 nvidiafabricmanager 服务,否则可能无法正常使用 GPU 实例。

操作步骤

本文以驱动版本 535.216.01 为例,您可以参考以下步骤进行安装,您也可以根据实际情况替换其他版本。

安装 nvidia-fabricmanager 服务

- 1. 使用标准登录方式登录 Linux 实例。
- 不同操作系统版本安装方法不同,请您参考以下方式,替换对应安装包路径,执行命令进行安装。不同镜像和驱 动版本匹配不同版本的安装包,更多版本选择可进入 NV 官网 查看。

🕛 说明:

- NVIDIA GPU 型实例升级 GPU 驱动的同时,还需同步升级 Fabric Manager,否则无法正常使用 GPU 实例。
- 使用 购买页自动安装驱动 功能将自动安装 Fabric Manager。



```
wget
https://developer.download.nvidia.cn/compute/cuda/repos/rhel8/x86_64
/nvidia-fabric-manager-535.216.01-1.x86_64.rpm
wget
https://developer.download.nvidia.cn/compute/cuda/repos/rhel8/x86_64
/nvidia-fabric-manager-devel-535.216.01-1.x86_64.rpm
rpm -ivh nvidia-fabric-manager-535.216.01-1.x86_64.rpm
rpm -ivh nvidia-fabric-manager-devel-535.216.01-1.x86_64.rpm
```



Ubuntu 22.04 镜像

wget

```
https://developer.download.nvidia.cn/compute/cuda/repos/ubuntu2204/x
86_64/nvidia-fabricmanager-535_535.216.01-1_amd64.deb
wget
https://developer.download.nvidia.cn/compute/cuda/repos/ubuntu2204/x
86_64/nvidia-fabricmanager-dev-535_535.216.01-1_amd64.deb
sudo dpkg -i nvidia-fabricmanager-dev-535_535.216.01-1_amd64.deb
sudo dpkg -i nvidia-fabricmanager-dev-535_535.216.01-1_amd64.deb
```

CentOS 7.x 镜像

wget

```
https://developer.download.nvidia.cn/compute/cuda/repos/rhel7/x86_64
/nvidia-fabric-manager-535.216.01-1.x86_64.rpm
wget
https://developer.download.nvidia.cn/compute/cuda/repos/rhel7/x86_64
/nvidia-fabric-manager-devel-535.216.01-1.x86_64.rpm
rpm -ivh nvidia-fabric-manager-devel-535.216.01-1.x86_64.rpm
rpm -ivh nvidia-fabric-manager-devel-535.216.01-1.x86_64.rpm
```

启动 nvidia-fabricmanager 服务

依次执行以下命令,启动服务。

systemctl enable nvidia-fabricmanager

systemctl start nvidia-fabricmanager

查看 nvidia-fabricmanager 服务状态

执行以下命令,查看服务状态。



systemctl status nvidia-fabricmanager

若输出信息如下,则表示服务安装成功。

[root@VM-17-134-tencentos ~]# systemctl status nvidia-fabricmanager
• nvidia-fabricmanager.service - NVIDIA fabric manager service
Loaded: loaded (/usr/lib/systemd/system/nvidia-fabricmanager.service; enabled; vendor preset: disabled)
Active: active (running) since Mon 2025-04-28 11:58:43 CST; 3h 57min ago
Main PID: 90357 (nv-fabricmanage)
Tasks: 17
Memory: 14.7M
CGroup: /system.slice/nvidia-fabricmanager.service
└─90357 /usr/bin/nv-fabricmanager -c /usr/share/nvidia/nvswitch/fabricmanager.cfg
Apr 28 11:58:42 systemd[1]: Starting NVIDIA fabric manager service
Apr 28 11:58:43 nv-fabricmanager[90357]: Connected to 1 node.
Apr 28 11:58:43 nv-fabricmanager[90357]: Successfully configured all the available NVSwitches to route GPU NVLink traffic.
Apr 28 11:58:43 systemd[1]: Started NVIDIA fabric manager service.



GPU 型实例安装 TCCL

最近更新时间: 2024-12-16 15:43:02

TCCL简介

TCCL(Tencent Collective Communication Library)是一款针对腾讯云星脉网络架构的高性能定制加速 通信库。主要功能是依托星脉网络硬件架构,为 AI 大模型训练提供更高效的网络通信性能,同时具备网络故障快速 感知与自愈的智能运维能力。TCCL 基于开源的 NCCL 代码做了扩展优化,完全兼容 NCCL 的功能与使用方法。 TCCL 目前支持主要特性包括:

- 双网口动态聚合优化,发挥 bonding 设备的性能极限。
- 全局 Hash 路由(Global Hash Routing),负载均衡,避免拥塞。
- 拓扑亲和性流量调度,最小化流量绕行。

操作场景

本文介绍如何在腾讯云环境中配置 TCCL 加速通信库,实现您在腾讯云 RDMA 环境中多机多卡通信的性能提升。 在大模型训练场景,对比开源的 NCCL 方案,TCCL 预计约可以提升 50% 带宽利用率。

操作步骤

准备环境

- 1. 创建 GPU 型 HCCPNV4sne 或 GPU 型 HCCPNV4sn 高性能计算集群实例,分别支持 1.6Tbps 和 800Gbps RDMA 网络。
- 2. 为 GPU 型实例 安装 GPU 驱动 和 nvidia-fabricmanager 服务。

▲ 注意: TCCL 运行软件环境要求 glibc 版本 2.17 以上, CUDA 版本 10.0 以上。

选择安装方式

TCCL目前支持三种使用方式安装,您可以根据需要选择适合业务场景的安装方式使用。由于当前大模型训练基本都 基于 Pytorch 框架,所以主要以 Pytorch 为例进行说明。

- NCCL 插件 + 排序的 IP 列表
- TCCL 通信库 + 编译安装 Pytorch
- TCCL 通信库 + Pytorch 通信插件

TCCL 的三种接入方案对比如下表:

┍╪┑┧╪╴╤╸╼┶	(推荐)方法一:安装	方法二:编译安装	方法三:安装 Pytorch 通
女表力式	NCCL 通信插件	Pytorch	信插件

🔗 腾讯云	
-------	--

适用场景	资源和模型训练场景相对比 较固定,推荐使用该方法, 兼容不同的 NCCL 版本和 CUDA 版本,安装使用方 便,不需要修改业务代码或 者重新编译 Pytorch。	 资源需要提供给不同的业务团队,或者经常有扩容的需求,不需要算法人员或者调度框架刻意去感知机器的网络拓扑信息。 不希望对业务代码做适配,只需要重新编译 Pytorch 框架。 	资源需要提供给不同的 业务团队,或者经常有 扩容的需求,不需要算 法人员或者调度框架刻 意去感知机器的网络拓 扑信息。
使用步骤	● 安装 NCCL 插件● 修改启动脚本	 安装 TCCL 重新编译安装 Pytorch 	 安装 Pytorch 通信插件 修改分布式通信后端
优点	安装方便	对业务代码无入侵	安装方便
缺点	集群节点扩充之后,需要更 新排序列表	需要重新编译安装 Pytorch 对软件环境有要求	需要修改业务代码 对软件环境有要求
软件环境依 赖	安装 NCCL 即可	 NCCL 版本 2.12 glibc 版本 2.17 以上 CUDA 版本 10.0以上 	 Pytorch 仅支持 1.12 glibc 版本 2.17 以上 CUDA 版本 10.0 以上

配置 TCCL 环境并验证

(推荐)方法一:安装 NCCL 通信插件

如果您已经安装了 NCCL,也可以使用 NCCL 插件的方式使用 TCCL 加速能力。

1. 安装 NCCL 插件。

○ 以 Ubuntu 20.04 为例,您可以使用以下命令安装插件。

```
# 卸载现有的 tool 和 nool 插件
dpkg -r tool && dpkg -r nool-rdma-sharp-plugins
# 下载安装 nool 1.2 插件
wget https://taco-1251783334.cos.ap-
shanghai.myqcloud.com/nool/nool-rdma-sharp-
plugins_1.2_amd64.deb && dpkg -i nool-rdma-sharp-
plugins_1.2_amd64.deb
```



请确保集群内使用的 nccl 插件版本一致,以下为 nccl 1.0 版本下载安装命 令,推荐使用稳定性更优的 nccl 1.2 版本 # wget https://taco-1251783334.cos.apshanghai.myqcloud.com/nccl/nccl-rdma-sharpplugins_1.0_amd64.deb && dpkg -i nccl-rdma-sharpplugins_1.0_amd64.deb && rm -f nccl-rdma-sharpplugins_1.0_amd64.deb

○ 如果您使用 CentOS 或 TencentOS,参考以下步骤安装:

```
# 卸载现有的 nccl 插件
rpm -e nccl-rdma-sharp-plugins-1.0-1.x86_64
# 下载安装 nccl 1.2 插件
wget https://taco-1251783334.cos.ap-
shanghai.myqcloud.com/nccl/nccl-rdma-sharp-plugins-1.2-
1.x86_64.rpm && rpm -ivh --nodeps --force nccl-rdma-sharp-
plugins-1.2-1.x86_64.rpm
# 请确保集群内使用 nccl 插件版本一致,以下为 nccl 1.0 版本下载安装命
令,推荐使用稳定性更优的 nccl 1.2 版本
# wget https://taco-1251783334.cos.ap-
shanghai.myqcloud.com/nccl/nccl-rdma-sharp-plugins-1.0-
1.x86_64.rpm && rpm -ivh --nodeps --force nccl-rdma-sharp-
plugins-1.0-1.x86_64.rpm && rm -f nccl-rdma-sharp-plugins-1.0-
1.x86_64.rpm
```

2. 获取拓扑排序的 IP 列表。

NCCL 插件不需要依赖文件可提供 bonding 口动态聚合和全局 hash 路由两种优化能力。如果需要支持 网络拓扑的亲和性感知,用户可以通过排序的 IP 列表来实现。

IP 排序可以按照如下方式完成:

○ 准备 IP 列表文件

VPC IP 地址通过 if config eth0 获取,每行1个节点 IP,格式如下:

```
root@VM-125-10-tencentos:/workspace# cat ip_eth0.txt
172.16.177.28
172.16.176.11
172.16.177.25
172.16.177.12
```

🔗 腾讯云

○ 执行排序

```
wget https://taco-1251783334.cos.ap-
shanghai.myqcloud.com/tccl/get_rdma_order_by_ip.sh && bash
get_rdma_order_by_ip.sh ip_eth0.txt
```

▲ 注意:

- 所有节点都安装了 curl 工具(例如对于 Ubuntu,通过 apt install curl 安装)。
- 执行脚本的节点可以 SSH 免密访问其他所有节点。

```
○ 查看排序后的 IP 列表文件
```

```
root@VM-125-10-tencentos:/workspace# cat hostfile.txt
172.16.176.11
172.16.177.12
172.16.177.25
172.16.177.28
```

3. 配置 TCCL 环境变量。

```
export NCCL_DEBUG=INFO
export NCCL_SOCKET_IFNAME=eth0
export NCCL_IB_GID_INDEX=3
export NCCL_IB_DISABLE=0
export
NCCL_IB_HCA=mlx5_bond_0,mlx5_bond_1,mlx5_bond_2,mlx5_bond_3,mlx5_b
ond_4,mlx5_bond_5,mlx5_bond_6,mlx5_bond_7
export NCCL_NET_GDR_LEVEL=2
export NCCL_IB_QPS_PER_CONNECTION=4
export NCCL_IB_TC=160
export NCCL_IB_TIMEOUT=22
export NCCL_PXN_DISABLE=0
# 机器IP手动排序之后,就不需要添加如下变量了
# export TCCL_TOPO_AFFINITY=4
```

4. 修改启动脚本。

您需要在启动分布式训练时修改启动脚本。例如,如果使用 deepspeed launcher 启动训练进程,拿到 排序后的 IP 列表之后,将对应的 IP 列表写入 hostfile,再启动训练进程。



```
root@vm-3-17-centos:/workspace/ptm/gpt# cat hostfile
172.16.176.11 slots=8
172.16.177.12 slots=8
172.16.177.25 slots=8
172.16.177.28 slots=8
deepspeed --hostfile ./hostfile --master_addr 172.16.176.11
train.py
```

如果使用 torchrun 启动训练进程,通过 --node_rank 指定对应的节点顺序,

```
// on 172.16.176.11
torchrun --nnodes=4 --nproc_per_node=8 --node_rank=0 --
master_addr=172.16.176.11 train.py ...
// on 172.16.176.12
torchrun --nnodes=4 --nproc_per_node=8 --node_rank=1 --
master_addr=172.16.176.11 train.py ...
// on 172.16.176.25
torchrun --nnodes=4 --nproc_per_node=8 --node_rank=2 --
master_addr=172.16.176.11 train.py ...
// on 172.16.176.28
torchrun --nnodes=4 --nproc_per_node=8 --node_rank=3 --
master_addr=172.16.176.11 train.py ...
```

如果使用 mpirun 启动训练进程,按照顺序排列 IP 即可。

```
mpirun \
-np 64 \
-H 172.16.176.11:8,172.16.177.12:8,172.16.177.25:8,172.16.177.28:8
\
--allow-run-as-root \
-bind-to none -map-by slot \
-x NCCL_DEBUG=INFO
-x NCCL_IB_GID_INDEX=3 \
-x NCCL_IB_DISABLE=0 \
-x NCCL_SOCKET_IFNAME=eth0 \
-x
NCCL_IB_HCA=mlx5_bond_0,mlx5_bond_1,mlx5_bond_2,mlx5_bond_3,mlx5_b
ond_4,mlx5_bond_5,mlx5_bond_6,mlx5_bond_7 \
-x NCCL_NET_GDR_LEVEL=2 \
-x NCCL_IB_QPS_PER_CONNECTION=4 \
```



-x NCCL_IB_TC=160 \
-x NCCL_IB_TIMEOUT=22 \
-x NCCL_PXN_DISABLE=0 \
-x LD_LIBRARY_PATH -x PATH \setminus
-mca coll_hcoll_enable 0 \
-mca pml obl \
-mca <code>btl_tcp_if_include eth0</code> \setminus
-mca btl ^openib \
all_reduce_perf -b 1G -e 1G -n 1000 -g 1

方法二:编译安装 Pytorch

由于社区 Pytorch 默认采用静态方式连接 NCCL 通信库,所以无法通过替换共享库的方式使用 TCCL。 1. 安装 TCCL。

以 Ubuntu 20.04 为例,您可以使用以下命令安装,安装之后TCCL位于 /opt/tencent/tccl 目录。



○ 如果您使用 CentOS 或 TencentOS,参考以下步骤安装:

```
# 卸载已有tccl版本和nccl插件
rpm -e tccl && rpm -e nccl-rdma-sharp-plugins-1.0-1.x86_64
# 下载tccl v1.5版本
wget https://taco-1251783334.cos.ap-
shanghai.myqcloud.com/tccl/tccl-1.5-1.tl2.x86_64.rpm && rpm -
ivh --nodeps --force tccl-1.5-1.tl2.x86_64.rpm && rm -f tccl-
1.5-1.tl2.x86_64.rpm
```

2. 重新编译安装 Pytorch。

以下为 Pytorch 源码安装示例,详情请参见 官网 Pytorch 安装说明。



```
#!/bin/bash
# 卸载当前版本
pip uninstall -y torch
# 下载pytorch源码
git clone --recursive https://github.com/pytorch/pytorch
cd pytorch
# <!重要> 配置TCCL的安装路径
export USE_SYSTEM_NCCL=1
export NCCL_INCLUDE_DIR="/opt/tencent/tccl/include"
export NCCL_LIB_DIR="/opt/tencent/tccl/lib"
# 参考官网添加其他编译选项图
# 安装开发环境
python setup.py develop
```

3. 配置 TCCL 环境变量。

```
export NCCL_DEBUG=INFO
export NCCL_SOCKET_IFNAME=eth0
export NCCL_IB_GID_INDEX=3
export NCCL_IB_DISABLE=0
export
NCCL_IB_HCA=mlx5_bond_0,mlx5_bond_1,mlx5_bond_2,mlx5_bond_3,mlx5_b
ond_4,mlx5_bond_5,mlx5_bond_6,mlx5_bond_7
export NCCL_NET_GDR_LEVEL=2
export NCCL_IB_QPS_PER_CONNECTION=4
export NCCL_IB_TC=160
export NCCL_IB_TIMEOUT=22
export NCCL_PXN_DISABLE=0
export TCCL_TOPO_AFFINITY=4
```

▲ 注意:

需要通过 TCCL_TOPO_AFFINITY=4 开启网络拓扑感知特性。

4. 运行 Pytorch 单机多卡或者多机多卡训练过程中有如下打印,说明安装成功:



VM-3-17-Centos:74350:74350 [0] NCCL INFO BOOTSTRAP : USING eth0:10.100.3.17<0>
vm-3-17-centos:74350:74350 [0] NCCL INFO NET/Plugin : No plugin found (libnccl-net.so), using internal implementation
vm-3-17-centos:74350:74350 [0] NCCL INFO NET/IB : Using [0]mlx5_bond_0:1/RoCE [R0]; 00B eth0:10.100.3.17<0>
<u>vm-3-17-centos:74350:74350 [0] NCCL INFO U</u> sing network IB
NCCL version 2.12.12_TCCL_v1.5+cuda11.6
vm-3-17-centos:74352:74352 [2] NCCL INFO Bootstrap : Using eth0:10.100.3.17<0>
vm-3-17-centos:74352:74352 [2] NCCL INFO NET/Plugin : No plugin found (libnccl-net.so), using internal implementation
vm-3-17-centos:74352:74352 [2] NCCL INFO NET/IB : Using [0]mlx5_bond_0:1/RoCE [R0]; 00B eth0:10.100.3.17<0>
vm-3-17-centos:74352:74352 [2] NCCL INFO Using network IB

5. 如果运行 nccl-tests,需要执行以下命令设置 TCCL 路径:

export LD_LIBRARY_PATH=/opt/tencent/tccl/lib:\$LD_LIBRARY_PATH

方法三:安装 Pytorch 通信插件

Pytorch 支持通过插件的方式接入第三方通信后端,所以在不重新编译 Pytorch 的前提下,用户可以使用 TCCL 通信后端,API 与 NCCL 完全兼容。详情可参见 Pytorch 现有通信后端介绍。

1. 安装 Pytorch 通信插件。

```
# 卸载现有的tccl和NCCL插件
dpkg -r tccl && dpkg -r nccl-rdma-sharp-plugins
# 卸载torch_tccl
pip uninstall -y torch-tccl
# 安装torch_tccl 0.0.2版本
wget https://taco-1251783334.cos.ap-
shanghai.myqcloud.com/tccl/torch_tccl-0.0.2_pt1.12-py3-none-
any.whl && pip install torch_tccl-0.0.2_pt1.12-py3-none-any.whl &&
rm -f torch_tccl-0.0.2_pt1.12-py3-none-any.whl
```

2. 修改业务代码。

```
import torch_tccl
#args.dist_backend = "nccl"
args.dist_backend = "tccl"
torch.distributed.init_process_group(
    backend=args.dist_backend,
    init_method=args.dist_url,
    world_size=args.world_size, rank=args.rank
```



)

3. 配置 TCCL 环境变量。

```
export NCCL_DEBUG=INFO
export NCCL_SOCKET_IFNAME=eth0
export NCCL_IB_GID_INDEX=3
export NCCL_IB_DISABLE=0
export
NCCL_IB_HCA=mlx5_bond_0,mlx5_bond_1,mlx5_bond_2,mlx5_bond_3,mlx5_b
ond_4,mlx5_bond_5,mlx5_bond_6,mlx5_bond_7
export NCCL_NET_GDR_LEVEL=2
export NCCL_IB_QPS_PER_CONNECTION=4
export NCCL_IB_TC=160
export NCCL_IB_TIMEOUT=22
export NCCL_IB_TIMEOUT=22
export NCCL_PXN_DISABLE=0
export TCCL_TOPO_AFFINITY=4
```

▲ 注意: 需要通过 TCCL_TOPO_AFFINITY=4 开启网络拓扑感知特性。

4. 在执行 Pytorch 分布式训练业务时,出现如下提示可确认通信后端被正确加载。

vm-3-17-centos:35915:35915 [0] NCCL INFO	NET/Plugin : No plugin found (libnccl-net.so), using internal implementation
vm-3-17-centos:35915:35915 [0] NCCL INFO	NET/IB : Using [0]mlx5_bond_0:1/RoCE [R0]; 00B eth0:10.100.3.17<0>
vm-3-17-centos:35915:35915 [0] NCCL INFO	Using network IB
NCCL version 2.12.12_TCCL_v1.5+cuda11.6	
vm-3-17-centos:35919:35919 [4] NCCL INFO	Bootstrap : Using eth0:10.100.3.17<0>
vm-3-17-centos:35919:35919 [4] NCCL INFO	NET/Plugin : No plugin found (libnccl-net.so), using internal implementation
vm-3-17-centos:35921:35921 [6] NCCL INFO	Bootstrap : Using eth0:10.100.3.17<0>
Vm-3-17-centos:35920:35920 [5] NCCL TNEO	Rootstran · Using eth0·10 100 3 17/05

🕛 说明:

如果运行 nccl-tests 或者其他需要动态链接通信库的场景,请使用方法一安装 TCCL。



GPU 型实例安装 RDMA 毫秒级监控组件

最近更新时间: 2025-01-06 10:56:32

功能简介

高性能计算集群具备在 RDMA 网络环境下实现毫秒级监控的能力,这使得您能够实时监测和分析瞬时的网络数据, 帮助您深入分析网络流量模式,进行网络优化和性能提升,为业务提供有力支持。

操作场景

本文介绍如何在腾讯云高性能计算集群环境中安装毫秒级监控组件,实现您在腾讯云 RDMA 环境中毫秒级的性能监 控。腾讯云提供两种监控数据的查看方式,您可以选择在云产品监控上查看毫秒级监控的统计数据或在实例本地查看 保存的监控日志。

▲ 注意:

RDMA 毫秒级监控启动后约占用小于 0.05 个核资源,可根据业务需要判断是否使用。

操作步骤

准备环境

- 1. 创建 GPU 型 HCCPNV4sne、GPU 型 HCCPNV4sn 或 GPU 型 HCCPNV5v 高性能计算集群实例, 镜像建议选择 TencentOS Server 2.4 (TK4)。
- 2. 为 GPU 型实例 安装 GPU 驱动 和 nvidia-fabricmanager 服务。

安装验证

1. 在 TencentOS Server 2.4 (TK4)环境下,您可以使用以下命令安装:

```
# 卸载已有增强型监控软件包
rpm -e rdma_monitor-1.0-1.tl2.x86_64
# 下载并安装毫秒级监控组件,
# 安装好软件包后,会自动注册系统服务来启动增强型监控并保活,无需手动启动
wget http://mirrors.tencentyun.com/install/GPU/rdma_monitor-1.0-
1.tl2.x86_64.rpm && rpm -ivh rdma_monitor-1.0-1.tl2.x86_64.rpm
```

2. 使用以下命令,验证是否安装成功:

ps -aux | grep monitor_server

执行命令,如果红字所示字段,代表增强型监控成功安装启动。

🔗 腾讯云

[root@VM-16-2-tencentos ~]# ps -aux | grep monitor_server root 3719 0.0 0.0 112824 2260 pts/5 S+ 15:26 0:00 grep --color=auto monitor_server root 230127 1.8 0.0 1587160 25080 pts/0 Sl+ 15:20 0:06 monitor_server -m 3

配置云产品监控

RDMA 毫秒级监控可在云产品监控查看统计数据,您可以在云产品监控--DashBoard 中配置您需要的监控指标, 操作步骤如下:

1. 新建 DashBoard,指标选择**云服务器-RDMA 监控**。

云服务器-RDMADash	Board: RDMA网卡友达带宽			1小时		
1MBps						
0.8MBps						
0.6MBps		ᆂᇊᆂᄴᇊ				
0.4MBps		打力数3 1	店			
0.2MBps	云服务器	• U =	基础监控			
0MBno	轻量应用服务器	•	存储监控			
19:15 19:19	云数据库	· · · [RDMA监控 1	19:55 19:59 20	:03 20:07	20:11 20:15
	容器服务(2.0)	•	,			
▼ 云产品监控	消息服务CKafka				① 左Y轴	▼ G
	各					
指标 🛈	云服务器 / RDMA监控 ▼	RDMA监控 / RD	MA网卡发送带宽 🔻			
筛选 ()		对象				
group by	实例 😮					
5						

2. 选择您需要监控的 RDMA 毫秒级统计指标。

▼ 云产品监控	New 应用性能监控 前端性能监	New New 空控 云拨测 告警数据源
指标 🛈	云服务器 / RDMA监控 ▼	RDMA监控 / RDMA网卡接收带宽 ▼
筛选 🛈	实例 ▼ 请选择	RX_PFC统计量(个/秒)
group by (•)	头例 🕹	毫秒级_RDMA网卡接收最小带宽(Mbps)
对比	小环比(昨天同时段) 同比	(_ 毫秒级_RDMA网卡接收带宽50百分位值(Mbps)
▼ 更多配置		毫秒级_RDMA网卡接收带宽90百分位值(Mbps)
别名	请输入图例别名	



腾讯云

指标英 文名	指标中文名	指标说明(非必填)	单 位	维度	统计粒度
RxHpb wAvg	毫秒级 RDMA 网卡 接收带宽平均 值	10秒内 RDMA 网卡接 收带宽的毫秒级统计粒 度平均值	M b p s	Instan celd	10s、60s、 300s、3600s
RxHpb wMax	毫秒级 RDMA 网卡 接收带宽最大 值	10秒内 RDMA 网卡接 收带宽的毫秒级统计粒 度最大值	M b p s	Instan celd	10s、60s、 300s、3600s
RxHpb wMin	毫秒级 RDMA 网卡 接收带宽最小 值	10秒内 RDMA 网卡接 收带宽的毫秒级统计粒 度最小值	M b p s	Instan celd	10s、60s、 300s、 3600s
RxHpb wP50	毫秒级 RDMA 网卡 接收带宽50百 分位值	10秒内从小到大 RDMA 网卡接收带宽的 毫秒级统计粒度前50百 分位数	M b p s	Instan celd	10s、60s、 300s、3600s、 86400s
RxHpb wP90	毫秒级 RDMA 网卡 接收带宽90百 分位值	10秒内从小到大 RDMA 网卡接收带宽的 毫秒级统计粒度前90百 分位数	M b p s	Instan celd	10s、60s、 300s、 3600s
TxHpb wAvg	毫秒级 RDMA 网卡 发送带宽平均 值	10秒内 RDMA 网卡发 送带宽的毫秒级统计粒 度平均值	M b p s	Instan celd	10s、60s、 300s、 3600s
TxHpb wMax	毫秒级 RDMA 网卡 发送带宽最大 值	10秒内 RDMA 网卡发 送带宽的毫秒级统计粒 度最大值	M b p s	Instan celd	10s、60s、 300s、3600s
TxHpb wMin	毫秒级 RDMA 网卡 发送带宽最小 值	10秒内 RDMA 网卡发 送带宽的毫秒级统计粒 度最小值	M b p s	Instan celd	10s、60s、 300s、3600s
TxHpb wP50	毫秒级 RDMA 网卡	10秒内从小到大 RDMA 网卡发送带宽毫	M b	Instan celd	10s、60s、 300s、3600s



	发送带宽50百 分位	秒级统计粒度前50百分 位数	p s		
TxHpb wP90	毫秒级 RDMA 网卡 发送带宽90百 分位	10秒内从小到大 RDMA 网卡发送带宽毫 秒级统计粒度前90百分 位数	M b p s	Instan celd	10s、60s、 300s、3600s

3. 选择需要监控的高性能计算集群实例 ID。

			_	
筛选 🛈	实例	请选择对象		

4. 单击确定即可快速创建 DashBoard。

新的 Dashboard		×
Dashboard 名称	新的 Dashboard(3)	
所属文件夹	公共Dashboard文件夹 ▼	
	确定取消	

查看本地监控

RDMA 毫秒级监控可查看最小 10ms 粒度级别的带宽数据监控,但云产品监控只支持最小粒度为 10s 的数据上报。如果用户想获取更精确的网卡监控数据,可以使用如下命令,保存毫秒级的数据在本地查看。



查看记录的监控数据,您可以根据需要分析监控记录,监控记录的格式如下:



Device	e: bon	nd3,	Т	ransmitt	ed	data	points:	1000,	Timestamp:	1697616573
Data P	Point	0:	0							
Data P	Point	1:	0							
Data P	Point	2:	0							
Data P	Point	3:	0							
Data P	Point	4:	0							
Data P	Point	5:	0							
Data P	Point	6:	0							
Data P	oint	7:	0							
Data P	oint	8:	0							
Data P	oint	9:	0							
Data P	oint	10:	0							
Data P	Point	11:	0							
Data P	oint	12:	0							
Data P	oint	13:	0							
Data P	oint	14:	0							
Data P	oint	15:	0							
Data P	oint	16:	0							

图中部分参数含义解释如下:

- Device: RDMA 网卡的名称。
- Transmitted data points: 接收侧 10s 内采集到的数据点数,这里是 10s 内采集了1000个点,也就是每 10ms 采集一次数据点,每个点的数据为对应 10ms 的接收带宽。
- Timestamp: 采集时的时间戳。
- Data Point n: 自时间戳 n × 10ms 后采集到的接收带宽。每个点的采样时间与前后的点均间隔 10ms。



GPU 型实例监控和告警

最近更新时间: 2025-04-16 15:16:32

监控与告警是保证高性能计算集群 GPU 型实例高可靠性、高可用性和高性能运行的重要部分。创建实例时,默认免 费开通腾讯云可观测平台。您可以通过 云服务器控制台 查看监控指标,详细说明请参见 云服务器监控内容。 NVIDIA GPU 系列实例另外提供了监控 GPU 使用率,显存使用量,功耗以及温度等参数的能力。 您也可以在 腾讯云可观测平台 分析监控指标和实施告警,更多详细内容可参见 腾讯云可观测平台告警管理。

() 说明:

监控功能是通过在 GPU 型实例上部署安装相关 GPU 驱动 、nvidia-fabricmanager 服务 和 云服务器 监控组件 来实现的,公共镜像默认包含云服务器监控组件,只需安装 GPU 驱动。如果您使用自定义镜 像,需手动安装云服务器监控组件和 GPU 驱动。

在控制台查看 GPU 监控指标

单击 GPU 列表中的 🕕 监控图标, 访问 控制台 GPU 实例的监控页面,查看 GPU 监控,移动鼠标到指标曲线 上将显示对应 GPU 设备的 BDF(Bus、Device、Function,设备唯一地址)和监控数据。如下图所示:

PU使用率(%) ()	♣ D ···	GPU显存使用量(MB) ()	▲ □ ····	GPU显存使用率(%) (ì	♣ D ··
2 9 6 3		600 400 200	02:37 431.00	3.2 2.4 1.6 0.8	02:37 2.81
0.38 01:44 01:49 01:54 02:00 02:05 02:10 02:1	6 02:21 02:26 02:32 02:37 0 最大值: 0.00 最小值: 0.00	0 01:38 01:44 01:49 01:54 02:00 02:05 0	2:10 02:16 02:21 02:26 02:32 02:37 0 最大值: 431.00 最小值: 4	0 01:38 01:44 01:49 01:54 0 <u>2:00 02:05</u> 02:10	02:16 02:21 02:26 02:32 02
Put)時後用量 (W) ① 2	02:32 11:00	GPU道度(攝氏度) ① 36 27 18 9 0	▲ 53 ···· 0237 36.00	GPU编码器使用率(%) ① 12 0.9 0.6 0.3	▲ □ +
11:38 01:44 01:49 01:54 02:00 02:05 02:10 02:1	6 02:21 02:26 02:32 02:37		:10 02:16 02:21 02:26 02:32 02:37 0 最大值: 36.00 最小值: 36	01:38 01:44 01:49 01:54 02:00 02:05 02:10	02:16 02:21 02:26 02:32 02 电,,,,,)最大值: 0.00最小值: 0.0
Pu#R48使用車(%) ① 2 2 9 6 6 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	* D ····				
0 01:38 01:44 01:49 01:54 02:00 02:05 02:10 02:1	6 02:21 02:26 02:32 02:37				

参数说明:

指标名称	含义	单位	维度
GPU 使用率	评估负载所消耗的计算能力,非空闲状态百分比	%	per-GPU
GPU 显存使用量	评估负载对显存占用	MB	per-GPU
GPU 显存使用率	评估负载对显存占用百分比	%	per-GPU



GPU 功耗使用量	评估 GPU 耗电情况	W	per-GPU
GPU 温度	评估 GPU 散热状态	摄氏度	per-GPU
GPU 编码器使用 率	评估编码器使用百分比	%	per-GPU
GPU 解码器使用 率	评估解码器使用百分比	%	per-GPU

在控制台查看 RDMA 监控指标

1. 在 高性能计算集群 中选择单击集群 ID 或查看详情, 可看到集群的 GPU 服务器实例详情。

三 ◇ 腾讯云 ∩ 🛙	211台				田黙号 备案) Q
云服务器	EIEE 遊怒免费试用CDN,全面加速您的图文/视频/下载等业务内容	8, 提供绝佳用户体验						
- 概览	高性能计算集群 🔇 上海 其它地域 🔻							
③ OrcaTerm	afite mile					集群ID		
② 实例自助检测	集群ID/名称	描述	可用区 〒	实例数			操作	
 ◎ 自动化助手 ~ ○ 配额管理 ~ 	hpo		上海八区	1			扩容 查看详情	I
11月1日日間 〇 11月1月日 11月11日 11月11日 11111 11月111 11月111 11月11111 11月11111 111111		-	上海八区	1			扩容 查看详情	ł
🖽 实例启动模板	hp hp		上海八区	1			扩容 查看详情)
 ◇ 专用宿主机 ◇ 高性能计算集群 ○ 雪放群组 	hoc		上海八区	3			扩容 查看详情	I

2. 单击 GPU 实例 列表中的 通 监控图标,访问 腾讯云控制台 GPU 实例的监控页面,查看 RDMA监控(仅支持分钟级以上粒度),移动鼠标到指标曲线上将显示实例对应 bond 的监控数据。如下图所示,当前支持 RDMA网卡发送带宽,RDMA 网卡接收带宽,RDMA 网卡出包量,RDMA 网卡入包量显示。

云服务器	大促 11.1-11.30会员消费金额TOP5	0赢万元大礼					监控详情	配置告答
## 概览	÷						1小时 📋 🚫 时间程度: 1分钟 🔻 🗘 关闭 💌 🚥	✔ 显示图例 仅查看有告警的指标
OrcaTerm 2							CPU监控 内存监控 磁盘监控 网络监控 存储监控 GPU监控 NVME	磁盘监控 网络相关指标 RDMA监控
④ 实例自助检测	3							
	可用区 上海三区 实例数 5	创建时间 20	23-11-28 15:27:22	描述 - 0			网络相关指标	
① 配额管理 ~							子机连接数(个) ①	
实例与镜像	使群次语 白肺经测						6004:37 55.00	
◎ 实例	果研資源目的性別	0.10 M					40	
三 实例启动模板	新建 整控 更	多操作 ▼					20	
◎ 专用宿主机	- ID/名称	监控	状态	可用区	实例类型	实例配置	0 03:40 03:47 03:54 04:01 04:08 04:15 04:22 04:29 04:36	
8 高性能计算集群		di	🖂 运行中	上海三区	HCCPNV6 🧿	384核 2304GB 21Mbps	量 最大值: 55.00 最小值: 42.00 平均值: 47.92	
③ 置放群组						系统盘:增强型SSI 硬盘	RDMA监控	
③服务器迁移 ~						网络: Default-VPC	RDMA网卡发送带宽(Mbps) ①	ps) 🗓 🌲 🛄 🚥
◎ 镜像							1	
⑥ SSH密钥		di	🖂 运行中	上海三区	HCCPNV6 🚹	384核 2304GB 20Mbns	0.5	
① 回收站 🛛 🗸	_					系統盘:增强型SSE 硬盘	03:40 03:47 03:54 04:01 04:08 04:15 04:22 04:29 04:36 03:40 03:47 03:54	04:01 04:08 04:15 04:22 04:29 04:36
网络与安全						网络: Default-VPC	jbondu ka. ♥	bond1 뮶7 bond1 뮶7
IP 公网ⅠP							Doug' we	DONG2 MR.
		di	🖂 运行中	上海三区	HCCPNV6 🔁	384核 2304GB	RDMA网卡出包量(个/秒) ①	(i)
存储与快照	_					ZUMBDS 系统盘:增强型SSI 项曲	1	
□ 云硬盘						∞篇 网络:Default-VPC	0.5	
 回 快照							03:40 03:47 03:54 04:01 04:08 04:15 04:22 04:29 04:36 03:40 03:47 03:54	04:01 04:08 04:15 04:22 04:29 04:36
运维与监控		di	🛞 运行中	上海三区	HCCPNV6 🚺	384核 2304GB 20Mbps	i bond 윤: ♥ i bond 윤: i bond2 윤:	bond0 最: bond1 最₂ bond2 最;



参数说明:

指标中文名	含义	单位	维度
RDMA 网卡接 收带宽	RDMA 网卡接收带宽	MBit/ s	Instanceld
RDMA 网卡发 送带宽	RDMA 网卡发送带宽	MBit/ s	Instanceld
RDMA 网卡入 包量	RDMA 网卡入包量	个/秒	Instanceld
RDMA 网卡出 包量	RDMA 网卡出包量	个/秒	Instanceld

() 说明:

腾讯云也提供 RDMA 网络毫秒级监控的能力,需要安装毫秒级监控组件,实现在腾讯云 RDMA 环境 中毫秒级的性能监控。操作步骤可参见 GPU 型实例安装 RDMA 毫秒级监控组件 。

在腾讯云可观测平台查看监控指标

腾讯云可观测平台 支持分析更丰富的 GPU 监控指标。

- 1. 登录 腾讯云可观测平台,左侧导航栏中选择 Dashboard ,进入 Dashboard 列表页。
- 2. 在 Dashboard 列表中,单击新建 Dashboard,在新的 Dashboard 选择新建图表。

查看 GPU 监控指标

在**指标**处选择 GPU / 云服务器 / GPU 监控,单击您关注的**指标**,自定义监控面板进行多实例展示,如下图所示:

~	云产品监控	New 应用性能监控	<mark>New</mark> 前端性能监控	New 云拨测	告警数据源	Prometheus	
	指标 🛈	GPU / 云服务器 / GPU	监控 🔻 GPL	J监控 / GPU编码	器使用率(% 🔻		
	筛选 🛈	实例	▼ 16个(ins-(•		
	group by i	实例 😣					
参数详	情可参见腾讯	云可观测平台 GP	U 云服务器监持	空指标,提供	以下监控指标	示:	
指标	英文名	指标中文名	指标说明			单位	维度



Gpumem usage	GPU 显存使用 率	GPU 显存使用率	%	per- GPU
GpuMem Used	GPU 显存使用 量	评估负载对显存占用	MB	per- GPU
Gpupowd raw	GPU 功耗使用 量	GPU 功耗使用量	W	per- GPU
Gpupowu sage	GPU 功耗使用 率	GPU 功耗使用率	%	per- GPU
Gputemp	GPU 温度	评估 GPU 散热状态	摄氏度	per- GPU
Gpuutil	GPU 使用率	评估负载所消耗的计算能力,非 空闲状态百分比	%	per- GPU
GpuEncU til	GPU 编码器使 用率	GPU 编码器使用率	%	per- GPU
GpuDecU til	GPU 解码器使 用率	GPU 解码器使用率	%	per- GPU

查看 RDMA 监控指标

在指标处选择 云服务器 / RDMA 监控,单击您关注的指标,自定义监控面板进行多实例展示,如下图所示:

云产品监控	New 应用性能监控	前端性	New 能监控	New 云拨测	告警数据源	Prometheus
指标 🛈	云服务器 / RDMA监控		▼ RD	DMA监控 / RDN	/IA网卡发送带货 ▼	
筛选 🕕	实例	▼ 请	RD 选择》	DMA监控 ▶	RDMA网卡发送带	宽(Mbps)
group by 🚺	实例 😢				RDMA网卡接收带到 RDMA网卡出包量(宽(Mbps) (个/秒)
对比	环比(昨天同时段)	同	比 (.		RDMA网卡入包量	(个/秒)

参数详情可参见腾讯云可观测平台 RDMA 监控指标,提供以下监控指标:

指标英文名	指标中文名	指标说明	单位	维度
RdmaIntraffic	RDMA 网卡接收 带宽	RDMA 网卡接收带宽	MBit/s	Instan celd



RdmaOuttraffi c	RDMA 网卡发送 带宽	RDMA 网卡发送带宽	MBit/s	Instan celd
Rdmalnpkt	RDMA 网卡入包 量	RDMA 网卡入包量	个/秒	Instan celd
RdmaOutpkt	RDMA 网卡出包 量	RDMA 网卡出包量	个/秒	Instan celd
CnpCount	CNP 统计量	拥塞通知报文统计	个/秒	Instan celd
EcnCount	ECN 统计量	显示拥塞通知统计	个/秒	Instan celd
RdmaPktDisc ard	端测丢包量	端测丟包量	个/秒	Instan celd
RdmaOutOfS equence	接收方乱序错误量	接收方乱序错误量	个/秒	Instan celd
RdmaTimeout Count	发送方超时错误量	发送方超时错误量	个/秒	Instan celd
TxPfcCount	TX PFC 统计量	TX PFC 统计量	个/秒	Instan celd
RxPfcCount	RX PFC 统计量	RX PFC 统计量	个/秒	Instan celd

监控指标告警配置

配置 GPU 监控指标告警

- 1. 登录 腾讯云可观测平台, 在左侧导航栏中, 选择告警管理 > 告警配置。
- 2. 单击 新建告警策略,在监控类型选择云产品监控,策略类型中选择云服务器/GPU 监控,选择您希望接 收告警的 GPU 实例对象,触发条件选择**手动配置**。



腾讯云可观测平台	÷	新建告警策略	ł			
G.告警管理 ^		1 配置告警	> 2 21	霍告警通知		
 告答配置 		基本信息				
· 告警治理		策略名称	GPU告警配置参考示	例		
· 统计大盘		备注	最多100个字符			
Dashboard						
◎ 接入中心						
□ 报表管理		配置告警规则				
全景监控		监控类型	云产品监控	HO 应用性能监控	HOT 前端性能监控	云拔测
△ 云产品监控 🗸 🗸		策略类型	云服务器 / GPU监控	~		
♀ Prometheus 监控						
Grafana 服务		門鵰恢盘	标签键	~	标签值	~ 6
④ 应用性能监控 🛛 🗸			+添加 ③ 键值粘	貼板		
™ 前端性能监控 ~		告警对象(与) 🛈	地域	~	上海	Ý
④ 终端性能监控 ~			实例	~	已选择1个	. *
日 日志服务 ~			将对上述维度筛选条件	的所有gpu做告	醫检測	

3. GPU 云服务器监控支持以下指标告警: GPU 内存使用率、GPU 功耗使用率、GPU 使用率、GPU 温度、GPU 是否存在显存页需隔离、GPU 显存是否发生 UCE 等。您可以参考下图进行配置告警。告警通知的配置可参见新建通知模板,支持通过多种渠道进行通知。

腾讯云可观测平台	策略类型	元極务器 / GPU监控 ✓
■ 监控概览	所属标签	
♪」 告警管理 ^		+ 流加 ② 建催枯枯板
・ 告警配置	告誓对象(与) ①	地域 > 上海 >
· 告警治理		実例 ✓ 已逃排 ✓ +
• 统计大盘		将对上述地度筛选条件的所有gpug音量检测
Dashboard	触发条件	○ 选择模板 ◎ 手动配置
⑦ 接入中心		
── 报表管理		指标告册
全景篮控		溝足以下 任意 > 指标判断条件时,触发告誓 启用告誓分级功能
△ 云产品监控 🛛 🗸		
♀ Prometheus 监控		If GPU助発使用率 体計粒度1分钟 <=
Grafana 服务		
② 应用性能监控 ~		if GPU温度 > 统计构度5分钟 > > > 0 80 °C 持续1个数据点 > then 每小时告誓一次 > 0 位
Raw 前端性能监控 ~		
⑦ 终端性能监控 ∨		ff CDI直示方左回 ∨ 併计指面1分钟 ∨ = ∨ ○ 1 kon 指述1个数据点 ∨ then 毎1小別生態→次 ∨ ○ □
□ 日志服务 ~		
(*) 云拨测 ~		
🖸 云压测 🗸 🗸		if GPU显存是否发 > 統计构度1分钟 > = > 〇 1 None 持续1个数量点 > then 每U小时告替一次 > 〇 首
⊗ 事件总线 ~		X-HEL
		192021179
	上一步	下一步:配置告望透知

常用告警指标参考如下:

|--|



GPU 功耗 使用率	<=0	功耗小于0时可能 功率出现 Unknown Error 了,会影响 GPU 的正常使用。	执行 nvidia-smi 命令查看 GPU 的功率 是否有 ERR 或 nvidia-smi -i <target gpu> -q grep "Power Draw" 是否为 Unknown Error,若存在该现象则尝试重 启机器恢复及更新驱动观察。若重启无法恢 复 提交工单 联系腾讯云支持。</target
GPU 温度	持续5分 钟>80	当 GPU 温度过高 时可能会导致 GPU SlowDown,影 响业务性能。	可能负载过高导致 GPU 温度过高,可尝试 重启实例恢复,若无法恢复 <mark>提交工单</mark> 联系 腾讯云支持。
GPU 是否 存在显存页 需隔离	=1	安培以下架构 GPU 出现了 ECC ERROR, 应用进程被 kill, GPU卡处于 pending 状态。	执行 nvidia-smi -i <target gpu=""> -q -d PAGE_RETIREMENT 命令查看是否有 GPU 卡处于 pending 状 态,重置 GPU 卡或重启实例恢复。若重启 无法恢复 提交工单 联系腾讯云支持。</target>
GPU 显存 是否发生 UCE	=1	安培及以上架构 GPU 出现了 ECC ERROR, 应用进程被 kill, GPU卡处于 pending 状态。	执行 nvidia-smi -i <target gpu=""> -q -d ROW_REMAPPER 命令查看是否有 GPU 卡处于 Pending 状 态,重置 GPU 卡或重启实例恢复。若重启 无法恢复 提交工单 联系腾讯云支持。</target>
GPU 内存 使用率	仅保持 观察	-	评估负载对显存占用。
GPU 使用 率	仅保持 观察	_	评估负载对 GPU 流处理器占用。

配置 RDMA 监控指标告警

1. 登录 腾讯云可观测平台,左侧导航栏中选择**告警配置**,新建告警策略,监控类型选择**云产品监控**,策略类型选择云服务器 /RDMA 监控,选择告警对象。



腾讯云可观测平台		← 新建告警策略			
器 监控概览					
□ 告警管理	~	1 配置告警	> (2	配置告警通知	
 告警配置 		基本信息	重新4021年4月		
 告營治理 		東町白柳	同证用印计算集	eeromamiti	
· 统计大盘		备汪	最多100个字	10	
G Bashboard					
□ 报表管理		配置告警规则			
全景监控		监控类型	云产品监控	(H) 应用性能监控	前端性能监控
🛆 云产品监控	~	策略类型	云服务器 / RI	DMA监控 ~	
♀ Prometheus 监控		所属标签	1-22.09	U.	1=20.48
Grafana 服务			+ 添加 ③	键值粘贴板	797-386 BB
 应用性能监控 前端体体体性 	Ň	生態対象(生) 〇	-		r-w
······利诺性能监控 ····································	ý.	百重对象(与)()	空侧	~	「加」
	~		将对上述维度领	^亲 选条件的所有rdma做	告警检测
(*) 云拔测	~	触发条件	选择模板	● 手动配置	
🖂 云压测	~		也存在會		
◎ 事件总线	Ý		加你告望	<u> </u>	
			满足以下	任意 ~	皆标判断条件时,触发

指标告警参考如下配置:

指标告警 满足以下	任第	意 ▼ 指标判	川断条	件时,触发告警										
Þ	if	接线状况监测	¥	统计粒度1分钟 🔻	!=	•	1	None	持续1个数据点 ▼	then	每1小时告警一次	Ŧ	(j)	Ū
Þ	if	ACS开关	¥	统计粒度1分钟 🔻	!=	•	0	None	持续1个数据点 ▼	then	每1小时告警一次	Ŧ	(j)	Ū
Þ	if	RDMA_MTU大小	•	统计粒度1分钟 🔻	!=	•	9100	Bytes	持续1个数据点 ▼	then	每1小时告警一次	•	(j)	Ū
Þ	if	ATS开关	•	统计粒度1分钟 🔻	!=	•	0	None	持续1个数据点 ▼	then	每1小时告警一次	•	<u>(</u>)	Ū
Þ	if	bonding模式	•	统计粒度1分钟 🔻	!=	•	4	None	持续1个数据点 ▼	then	每1小时告警一次	Ŧ	(j)	Ū
Þ	if	dcqcn使能	¥	统计粒度1分钟 🔻	!=	▼	11	None	持续1个数据点 ▼	then	每1小时告警一次	Ŧ	(j)	Ū
Þ	if	网卡混杂模式	¥	统计粒度1分钟 🔻	!=	¥	0	None	持续1个数据点 ▼	then	每1小时告警一次	Ŧ	(j)	Ū
Þ	if	流量类别	•	统计粒度1分钟 🔻	!=	•	160	None	持续1个数据点 ▼	then	每1小时告警一次	•	i	Ū
Þ	if	q5PFC配置	•	统计粒度1分钟 🔻	!=	•	1	None	持续1个数据点 ▼	then	每1小时告警一次	•	i	Ū
Þ	if	优先级信任状态	•	统计粒度1分钟 🔻	!=	•	1	None	持续1个数据点 ▼	then	每1小时告警一次	•	i	Ū
Þ	if	PCIE速率	Ŧ	统计粒度1分钟 🔻	!=	•	16	None	持续1个数据点 ▼	then	每1小时告警一次	Ŧ	i	Ū
Þ	if	PCIE宽度	¥	统计粒度1分钟 🔻	!=	Ţ	16	None	持续1个数据点 ▼	then	每1小时告警一次	Ŧ	i	Ū
Þ	if	IB设备状态	¥	统计粒度1分钟 🔻	!=	Ţ	1	None	持续1个数据点 ▼	then	每1小时告警一次	Ŧ	i	Ū
Þ	if	PCIE最大读取长度	•	统计粒度1分钟 🔻	!=	v	4096	Bytes	持续1个数据点 ▼	then	每1小时告警一次	•	(j)	Ū
Þ	if	NV_PEER_MEM	•	统计粒度1分钟 🔻	!=	•	1	None	持续1个数据点 ▼	then	每1小时告警一次	•	i	Ū
添加指标														

2. 告警通知可参见 新建通知模板 配置,支持多渠道通知。

配置完成后策略查看截图如下:

分 腾讯云



告警管理									
告警大盘 告警历史	1188119727 告愍屈蔽 词	知模板 鲀发条件模板							
① 如有任何问题或建议、 请	扫码加技术交流群。 我们将诸说为您服务	5.							
新建筑路	更多操作 👻						高级铸造 策略名称/ID: rdf	na 📀	φ \$
							多个关键字用竖线	"广分隔,多个过滤标签用回车键分	M 🕲 🛈 Q
策略名称	监控类型	策略类型	告誓规则	策略所属项目 ¥	关联实例数	通知模板 ▼	最后修改 ↓	告誓启停 T	操作
				找到1条结果 🕽	青睐筛选条件				
	元产品监控	云服务器-RDMA监控	地域 = 广州、实例 = 撞线状况监测 I= 1 None,统计和度1分钟,连续1次满足… AGS开关 I= 0 None,统计和度1分钟,连续1次满足条件… RDMA_MTU大小 I= 9100Bytes,统计和度1分钟,连续1…		·	系统预设通知模板	3248661176 2023/05/08 11:59:44		复制 删除 告警历史
H1.5		構成 構成 構成 構成 日 日 日 日 日 日 日 日 日 日 日 日 日	(4) (1) (2) (2) (3) (3) (4) (4) (4) (4) (4) (4) (4) (4) (4) (4) ー次 ・ ・ ・ ・ ・ ・ ・ ・ ・ ・ ・ ・ ・				20 -	8/R X 1 /1R x

告警示例截图如下:

← 管理告警策略										伊用 熊稔
策略详情 告警历史										
今天 昨天 近7天	近30天 2023-05-02 00:0	0:00 ~ 2023-05-08 23:59:59 🗄							高级铸造 请输入告誓对象	Q \$
发生时间 \$	监控类型	策略类型	告誓对象	告誓内容	地域	持续时长	告豐状态 🕇	策略名称	结束时间	告警类型
2023-05-08 11:59:00	云产品监控	云服务器-RDMA监控		接线状况监测 !=1None	广州	2分钟	未恢复		-	指标

高性能计算集群 GPU 型实例 RDMA 告警处理建议如下:

监测指 标	指标 名	错误描 述	正确配 置值	处理策略	客户更正方式	设 备
接线状 况监测	link _de tect ed	链路 down	1(1代 表端口 up)	客户尝试 软件恢 复,如无 法恢复, 授权运维 维修	ifconfig \$ethname up	et h
ACS 开关	acs	ACS 开关配 置错误	0(0代 表关闭 ACS)	客户改正 配置(需 要重启机 器)	bash /etc/acsctl_online.sh disable_acsctl	et h
RDM A_MT U 大小	acti ve_ mtu	RDM A 卡的 MTU 配置错 误(影 响性 能)	9100	客户更正 配置即可	ifconfig \$ethname mtu 9100	bo nd

ATS 开关	ats _en able d	ATS 开关配 置错误	0(0代 表关闭 ATS)	客户改正 配置(需 要重启机 器)	<pre>// 关闭 ATS for i in `lspci -d 15b3: awk '{print \$1}'`; do echo \$i; mlxconfig -d \$i -y s ATS_ENABLED=0; done // 重启之后确认状态 for i in `lspci -d 15b3: awk '{print \$1}'`; do echo \$i; mlxconfig -d \$i q grep ATS_ENABLED; done</pre>	et h
bondi ng 模 式	bon ding _m ode	bondi ng 模 式配置 错误	4(4代 表双发 模式)	客户更正 配置即可	cd /usr/local/qcloud/rdm a/; sh set_bonding.sh; sh dscp.sh	bo nd
dcqc n 使能	dcq cn_ ena ble	dcqc n 未使 能	11(两 个1分别 代表 rp 和 np 的状 态)	客户更正 配置即可	echo 1 > /sys/class/net/\$ethna me/ecn/roce_rp/enab le/5 echo 1 > /sys/class/net/\$ethna me/ecn/roce_np/ena ble/5	et h
网卡混 杂模式	eth _pr omi sc	网卡误 配为混 杂模式	0(0代 表非混 杂模 式)	客户更正 配置即可	ifconfig \$ethname – promisc	et h
流量类 别	traff ic_c lass	流量类 别配置 错误	160	客户更正 配置即可	echo 160 > /sys/class/infiniband/ \$RDMA_name/tc/1/tr affic_class	bo nd
q5 PFC 配置	q5_ pfc _en able d	PFC 未使 能,存 在 QOS	1(1代 表 PFC 使能)	客户更正 配置即可	mlnx_qos −i \$ethname −f 0,0,0,0,0,1,0,0	et h



🔗 腾讯云

		ERRO R				
优先级 信任状 态	prio _tru st_s tate	优先级 信任状 态配置 错误	1(1代 表 dscp)	客户更正 配置即可	mlnx_qos –i \$ethname –– trust=dscp	et h
pcie 速率	max _lin k_s pee d	PCIE GEN 配置错 误	16	客户更正 配置即可	尝试重启实例恢复,若无 法恢复 <mark>提交工单</mark> 腾讯云支 持	et h
pcie 宽度	max _lin k_w idth	PCIE width 配置错 误	16	客户更正 配置即可	尝试重启实例恢复,若无 法恢复 <mark>提交工单</mark> 腾讯云支 持	et h
IB 设 备状态	link _st ate	bond 口下两 个 eth 口全部 down	1(1代 表 bond 口up)	客户尝试 软件恢 复,如无 法恢复, 授权运维 维修	ifconfig \$ethname up	bo nd
MRS S PCIE 最大读 取长度	mrs s	MRS S 配置 错误	4096	客户更正 配置即可	lspci −D −nn grep 15b3 awk −F' ' '{print \$1}' xargs −I {} setpci −s {} 68.w=5936	et h
NV_ MEM _PEE R 是否 安装	nv_ pee r_m em _st ate	nvidia _peer mem 模块未 加载	1(1代 表模块 已加 载)	客户加载 模块即可	modprobe nvidia_peermem	整机