



曦云®系列通用计算 GPU

Warm Reset 使用指南

CSOG-24025-020-F3_V02

2024-08-05

沐曦专有和三级保密信息

本文档受 NDA 管控

更新记录

版本	日期	更新说明
V02	2024-08-05	更新以下章节： 5.2 虚拟机部署
V01	2024-06-14	正式版本首次发布

目录

1. 概述.....	1
1.1 应用场景.....	1
2. 软件栈版本依赖	2
3. 适配服务器	3
4. 物理机部署	4
4.1 基本原理.....	4
4.2 物理机部署	5
5. 虚拟机部署	7
5.1 基本原理.....	7
5.2 虚拟机部署	7
6. 已知问题和限制	9

图目录

图 4-1 单主机无互联	4
图 4-2 单主机多卡互联.....	4
图 4-3 多主机多卡互联.....	5

飞腾信息 MetaX Confidential
2024-11-18 15:00:00

表目录

表 3-1 推荐使用的服务器列表..... 3

飞腾信息 MetaX Confidential
2024-11-18 15:00:00

1. 概述

本文档描述了如何通过 warm reset 复位 GPU 卡来恢复业务。

1.1 应用场景

当服务器上的一张或者多张 GPU 卡出现不可通过软件恢复的问题，且服务器主机及操作系统没有明显异常的情况下，可以尝试 warm reset 复位出现问题的 GPU 卡来恢复业务，以此避免重启服务器主机，节省运维的时间和人力成本。

一般情况下，warm reset 适用以下场景：

- 非显存坏页的不可自动纠正的 RAS 错误
- VBIOS 更新后立即生效
- GPU 卡之间 MetaXLink 出现异常，比如连接断开或降速等
- 其它未知问题引起的 GPU 卡异常，hang 机或不响应等

说明

由于 warm reset 在软硬件实现上和服务器型号、操作系统以及出现问题时的上下文都紧密相关，无法保证在所有实际场景下都能正常工作，建议 warm reset 后对所有 GPU 卡都进行基本的测试后再运行业务。如果 warm reset 后有任何系统异常，请立即冷重启服务器。

2. 软件栈版本依赖

使用 warm reset 需要软件栈版本满足以下条件：

- VBIOS $\geq$ 1.12.0.0
- KMD $\geq$ 2.4.1
- mx-smi $\geq$ 2.1.5
- libvirt $\geq$ 0.8.0（虚拟机部署下的 host 端环境）

3. 适配服务器

由于 warm reset 在 PCIe 接口上和服务器紧密相关，并不能保证在所有服务器上都能正常工作，有些服务器存在 warm reset 后 GPU 卡无法识别或者无法正常工作等问题，所以对于不同的服务器，有可能需要特殊的适配才能使用，以下是经过测试，推荐使用的服务器列表：

表 3-1 推荐使用的服务器列表

服务器型号	Warm Reset 功能	已知问题
H3C UniServer R5300 G6	支持	无

4. 物理机部署

4.1 基本原理

MetaX GPU 卡的 warm reset 是指在不重启主机电源或操作系统的情况下进行 GPU 卡的复位，管理人员确认 GPU 卡空闲的情况下，通过 mx-smi 接口下发 warm reset 命令来触发 GPU 卡的复位，目前支持以下使用部署场景：

- 单主机无 GPU 卡互联环境，如图 4-1，即 GPU 卡之间没有通过 MetaXLink 互联。

可以直接使用 mx-smi 对出现问题的 GPU 卡进行 warm reset 复位。



图 4-1 单主机无互联

- 单主机多卡互联的 GPU 卡环境，如图 4-2，即 GPU 卡之间通过 MetaXLink 互联。

因为 warm reset 在执行过程中会将 MetaXLink 硬件复位，板卡内的 MetaXLink 会和对端 GPU 卡断开连接，所以在有 MetaXLink 互联的多卡环境中如果有任何板卡需要 warm reset，则互联群组内所有 GPU 卡都需要进行 warm reset。

使用 mx-smi 下发 warm reset 命令时，KMD 驱动会判断 GPU 卡是否存在 MetaXLink 互联并自动完成互联群组内所有 GPU 卡的 warm reset，且在 warm reset 完成后，KMD 会自动恢复群组内所有 GPU 卡的 MetaXLink 互联。

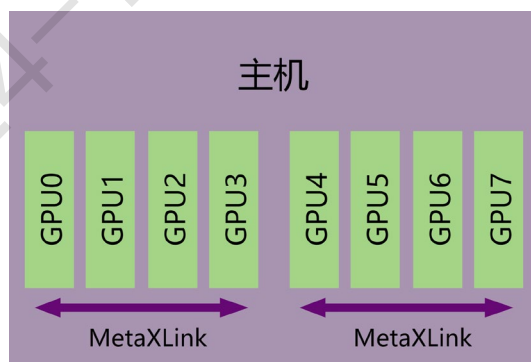


图 4-2 单主机多卡互联

- 多主机多卡互联的 GPU 卡环境，如图 4-3，即整个 GPU 卡的互联群组分布在多个主机上，单主机内的 GPU 卡之间有 MetaXLink 互联，不同主机上的 GPU 卡之间也存在 MetaXLink 互联。

因为 warm reset 在执行过程中会将 MetaXLink 硬件复位，板卡内的 MetaXLink 会和对端 GPU 卡断开连接，所以在有 MetaXLink 互联的多卡环境中如果有任何板卡需要 warm reset，则互联群组内所有 GPU 卡都需要进行 warm reset。

由于这个互联群组分布在多个主机上，所以对这种系统进行 warm reset 的时候需要分别对每台主机下的 GPU 卡下发 warm reset 命令。在每个主机上，只需要选一张卡下发 warm reset 命令，同一主机内有 MetaXLink 互联的其它 GPU 卡也会自动进行 warm reset，等所有主机都完成 warm reset 之后，各主机内的 MetaXLink 会自动恢复连接，跨主机之间的 MetaXLink 则按需下发连接命令来实现跨主机之间的 MetaXLink 互联。

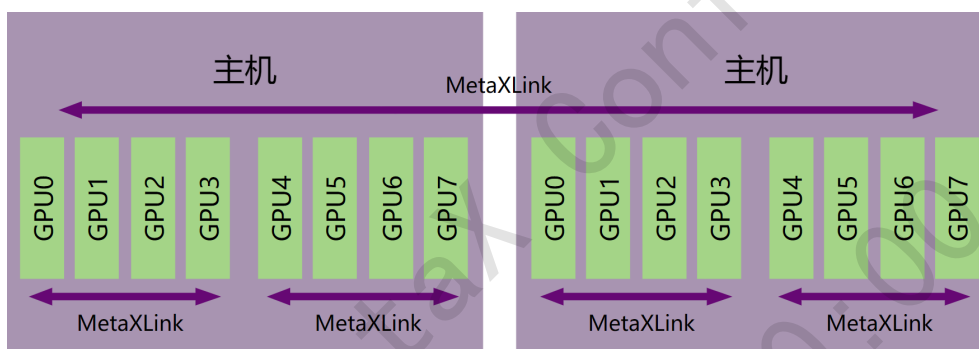


图 4-3 多主机多卡互联

4.2 物理机部署

单主机下，带 MetaXLink 互联与不带 MetaXLink 互联的 GPU 卡在执行 warm reset 时的操作步骤相同。多主机下，若不同主机上的 GPU 卡间存在 MetaXLink 互联，则需要所有机器上同时执行 warm reset。

操作步骤

- 确认当前环境下没有业务在使用 metax 驱动。执行以下命令，确认 metax 引用行的最后一列显示为 0，若不为 0，则表示还有业务在引用 metax 驱动，需要将业务强制退出。

```
lsmod | grep metax
```

- 通过错误日志或者 mx-smi 定位出错 GPU 卡的 node ID。

```
mx-smi -L
```

- 通过 mx-smi 工具，找出需要 warm reset 的 GPU 卡的 node ID（有 MetaXLink 互联的 GPU 卡群组可以指定群组内的任意 node ID），然后执行以下命令进行 warm reset。

```
mx-smi -r -i <node ID>
```

说明

- 如果执行 warm reset 命令后 console 显示了错误信息，或者该命令无法结束，请重启主机。
- 如果不知道 node ID，可直接使用 `mx-smi -r -i all` 命令对所有 GPU 卡进行 warm reset。

- 等待大约 90 秒钟后查看 GPU 卡是否复位成功，执行以下命令查看所有 GPU 卡和拓扑连接是否恢复，同时确认 dmesg 没有错误日志。

```
mx-smi topo -m
```

说明

- 执行 warm reset 命令后的 90 秒等待时间内，不要对 GPU 再下发任何其他操作，等检查到 GPU 复位完成后再进行操作。
- 如果 warm reset 完成后，通过 mx-smi 命令查看的拓扑连接不正常，或者 dmesg 有错误日志，请冷重启主机。

- 多主机 MetaXLink 互联环境下完成 warm reset 后，按需进行 MetaXLink 的 training，操作步骤和冷重启主机一样。

5. 虚拟机部署

5.1 基本原理

管理员在 host 端部署一个 warm reset 监控工具，该工具会在管理平台对虚拟机开机和关机操作时进行 warm reset 相关操作。

当用户在虚拟机测试过程中遇到 GPU 卡严重故障问题，并提示需要 warm reset 操作恢复后，用户在虚拟机里面发起 warm reset 操作（实际为 warm reset 请求），然后将虚拟机关机并重新开机。

在 host 端启动虚拟机前，warm reset 监控工具检测到 GPU 卡有 warm reset 请求标记，会对 GPU 卡进行 warm reset 复位。

全部 GPU 卡 warm reset 完成后，在 host 端正式启动虚拟机，用户继续执行业务。

说明

host 端的 warm reset 监控工具无法获取到完整 MetaXLink 互联信息，在 host 启动虚拟机前，warm reset 监控工具会对属于该虚拟机下的所有 GPU 卡进行必要的状态检查，只有当前状态满足 warm reset 条件了，监控工具才会对 GPU 卡进行 warm reset 操作。

5.2 虚拟机部署

操作步骤

1. 第一次部署时，管理员在 host 安装 mxhv 工具。

```
rpm: rpm -i mxhv-xxx.rpm
```

或

```
deb: dpkg -i mxhv-xxx.deb
```

说明

- 必须在 host 安装 mxhv 工具，因为在虚拟机里触发 warm reset 命令后，都是由 host 端的 mxhv 工具执行 warm reset 复位。
- 不要在 host 安装 GPU KMD 驱动或其他 MXMACA 工具，防止 host 端 GPU KMD 驱动对透传 GPU 卡有影响。

2. 管理员将用户需要的 GPU 卡透传配置到虚拟机。
3. 用户通过虚拟机管理平台启动虚拟机。
4. 用户需要 warm reset 时，通过 mx-smi 工具，找出需要 warm reset 的 GPU 卡的 node ID（有 MetaXLink 互联的 GPU 卡群组可以指定群组内的任意 node ID），然后在虚拟机下执行如下命令：

```
mx-smi -r -i <node ID>
```

说明

- 如果执行 warm reset 命令后 console 显示了错误信息，或者该命令无法结束，请重启主机。
- 如果不知道 node ID，可直接使用 `mx-smi -r -i all` 命令对所有 GPU 卡进行 warm reset。

5. 用户在虚拟机里面执行 `poweroff` 命令，关闭虚拟机。

```
poweroff
```

说明

必须使用 `poweroff` 命令，不能用 `reboot` 等其它重启命令。

6. 用户通过虚拟机管理平台再次启动虚拟机。

说明

在 host 端启动虚拟机前，warm reset 监控工具会对 GPU 卡进行 warm reset 复位，完成后才会正式启动虚拟机，如果 5 分钟后虚拟机启动没有成功，请重启主机。

7. 虚拟机启动成功后，执行以下命令查看所有 GPU 卡和拓扑连接是否恢复，同时确认 `dmesg` 没有错误日志。

```
mx-smi topo -m
```

说明

如果 warm reset 完成后，通过 `mx-smi` 命令查看的拓扑连接不正常，或者 `dmesg` 有错误日志，请冷重启主机。

8. 用户继续执行业务。

6. 已知问题和限制

1. warm reset 不能解决服务器硬件故障问题。
2. warm reset 不能解决操作系统故障问题。
3. 未经测试的服务器可能存在 warm reset 之后丢卡现象（参见表 3-1 推荐使用的服务器列表）。
4. 未经测试的服务器可能 RC 在 warm reset 过程收到 AER 会出现服务器直接重启现象（参见表 3-1 推荐使用的服务器列表）。
5. 在同一个互联的 GPU 卡群组上存在老版本 VBIOS (< 1.12.0.0) 和新版本 VBIOS (≥ 1.12.0.0) 的 GPU 卡混插，会导致老版本 GPU 卡启动失败，且也无法升级 VBIOS，需将 MetaXLink 断开后才能正常工作。
6. 在同一个互联的 GPU 卡群组上老版本 VBIOS (< 1.12.0.0) 升级到新版本 VBIOS (≥ 1.12.0.0) 时，必须将所有老版本 VBIOS 的 GPU 卡的 VBIOS 全部升级后，才能进行 warm reset 或重启主机的操作，避免出现上述混插场景。
7. 如果通过 warm reset 复位 GPU 卡方式让新 VBIOS 版本生效，必须保证新版本 VBIOS 的 BAR size 配置和旧版本 VBIOS 的 BAR size 配置一致。
8. 支持 PF 卡 warm reset，不支持 VF 卡的 warm reset。

声明

版权所有 ©2024 沐曦集成电路（上海）有限公司。保留所有权利。

本文档中呈现的信息属于沐曦集成电路（上海）有限公司和/或其附属公司（以下统称为“沐曦”），非经沐曦事先书面许可，任何实体或个人均不得获得本文档的副本，且无权以任何方式处理本文档，包括但不限于使用、复制、修改、合并、出版、发行、销售或传播本文档的部分或全部。

本文档内容仅供参考，不提供任何形式的、明示或暗示的保证，包括但不限于对适销性、适用于任何目的和/或不侵权的保证。在任何情况下，沐曦均不对因本文档引起的、由本文档造成的、或与之相关的任何索赔、损害或其他责任负责。

沐曦保留自行决定随时更改、修改、添加或删除本文档的部分或全部的权利。沐曦保留最终解释权。

沐曦、MetaX 和其他沐曦图标是沐曦的商标。本文档中提及的所有其他商标和商品名称均为其各自所有者的财产。