NVIDIA GPU Memory Error Management

Application Note

DA-09826-002_vR520 | December 2022

Table of Contents

Chapter 1. Overview

NVIDIA® Ampere architecture introduces new memory error recovery features that improve resilience and avoids impacting unaffected applications. The new features improve various aspects of the graphics processing units' (GPU) response to memory errors, which improves the overall robustness of the error handling and recovery process.

The error handling and response features are:

- Error-containment.
- Dynamic page offlining.
- Row-remapping.
- Uncorrectable error to correctable error coverage improved by 10%.
- ECC correction of single bit errors (SBEs).

When referring to ECC errors in this application note, we focus on uncorrectable high bandwidth memory (HBM) memory errors. SRAM errors and correctable HBM errors are outside the scope of this application note.

Chapter 2. Supported GPUs

Error-containment and row-remapping are two separate features in the GPU architecture. The following table shows the GPUs that support these features.

Table 1. Supported GPUs

	Ampere GA100	Ampere GA10x	Ada AD10x	Hopper GH100
Error Containment	X			X
Row remapping	X	X	X	X
Dynamic Page Offlining	X			X

Chapter 3. Error Containment

The NVIDIA Ampere architecture introduces the concept of error containment to NVIDIA GPUs. The benefit of error containment is being able to limit the impact of uncorrectable ECC errors on GPU applications. Uncorrectable ECC errors on prior architectures such as NVIDIA Volta™ impacted all of the currently executing GPU workloads. On NVIDIA data center-class GPUS, such as NVIDIA A100 and NVIDIA H100, the impact will be limited to the applications that encounter the error. All other workloads will continue running unaffected both in terms of accuracy and performance, and new workloads can be launched. Unlike earlier GPU architectures, NVIDIA 100-class GPUs do not require a GPU reset when memory errors occur.

Note: While most frequently occurring classes of uncorrectable errors are contained, there can be rare cases where uncorrectable errors are still uncontained and might impact all the workloads being processed in the GPU.

Chapter 4. Dynamic Page Offlining

Dynamic Page Offlining improves resiliency and availability of NVIDIA 100-class GPUs to uncorrectable ECC errors. Once the NVIDIA driver identifies the location of an uncorrectable error in the frame buffer memory, it marks the page containing the error as unusable. Once the page is marked unusable, any of the currently executing or newly launched workloads will not be allocating this page in question.

Dynamic Page Offlining exists on NVIDIA 100-class starting with the NVIDIA Ampere architecture. It is not available on previous generations of NVIDIA GPUs that do not support error containment.

GPUs that support dynamic page offlining do not require a GPU reset to recover from most uncorrectable ECC errors.

After the page is marked as unusable, it will not be mapped to the address space of any currently running or newly launched CUDA kernels.

Chapter 5. Row-Remapping

Row-remapping is a hardware mechanism to improve the reliability of frame buffer memory on GPUs starting with the NVIDIA Ampere architecture. This feature is used to prevent known degraded memory locations from being used. The row-remapping feature is a replacement for the page retirement scheme used in prior generation GPUs. Every bank in HBM is equipped with spare rows in hardware. As opposed to traditional page retirement, the row-remapper replaces degrading memory cells with spare ones to avoid offlining regions of memory in software. This differs from dynamic page offlining in that the memory is fixed at a hardware level and does not leave software visible holes in the address space. The process of row-remapping requires a GPU reset to take effect and will remain persistent throughout the life of the life of the GPU.

The following table describes the differences between page retirement and row-remapping.

Table 2. Page Retirement vs. Row-Remapping

Feature	Page Retirement for Legacy GPUs	Row-Remapping for A100/ H100
Available remappings/ retirements	Supported a maximum of 64 retirements for the frame buffer	Supports up to 512 remapping for the frame buffer.
Policy changes	Once a retirement takes effect, the page can never be unretired, regardless of correctable or uncorrectable errors	Remapping due to correctable errors can be replaced by uncorrectable error remapping when the memory bank's reserved rows are exhausted.
RMA criteria	A threshold of page retirements on a GPU usually resulted in investigation of whether the GPU was worthy of an RMA	See RMA Policy Thresholds for Row-Remapping.
Application of pending changes	Needed a kernel module reload or driver re-initialization or GPU reset	GPU reset is required.

Chapter 6. Response to Uncorrectable Contained ECC Errors

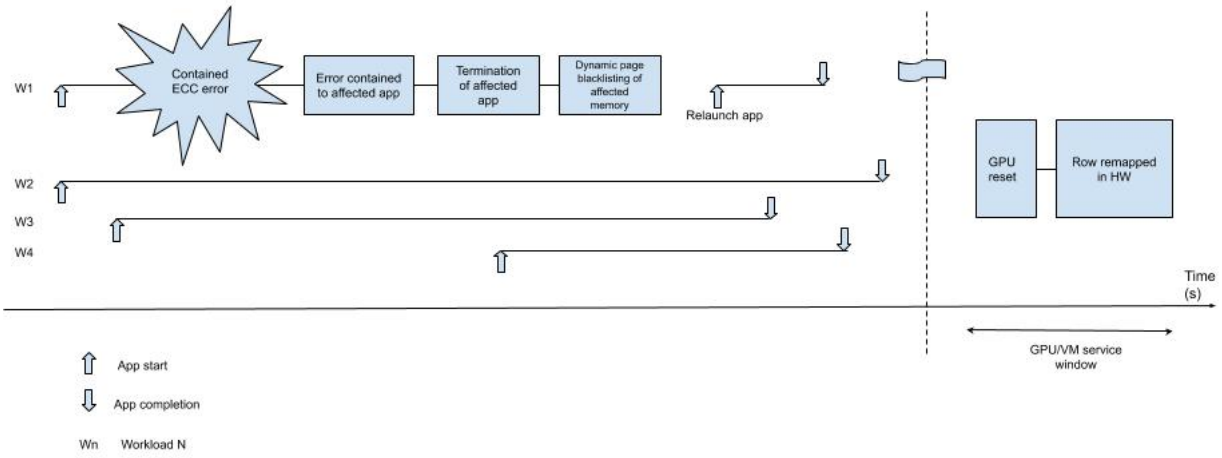
Like previous GPU architectures, when an uncorrectable ECC error is detected, the NVIDIA driver software will perform error recovery. Error containment ensures that erroneous data does not continue to propagate, and the affected application is terminated.

- Uncorrectable **contained** ECC error are uncorrectable ECC errors where error containment process **was** successful.
- Uncorrectable **uncontained** ECC error are uncorrectable ECC errors where error containment process **was not** successful.

Dynamic page offlining marks the page containing the faulty memory as unusable. This ensures that new allocations do not land on the page that contains the faulty memory. Unaffected applications will continue to run and additional workloads can be launched on this GPU without requiring a GPU reset.

When GPU reset occurs as a part of the regular GPU/VM service window, row remapping fixes the memory in hardware without creating any holes in the address space and the offlined page is reclaimed.

Figure 1. NVIDIA A100/H100 Response to Uncorrectable Contained ECC Error



Chapter 7. Error Recovery and Response Flags

Here is a list of flags that are required for client error recovery and response:

- Reset pending flag
  - When set, this flag indicates that the GPU has encountered an uncorrectable, uncontained error that requires a GPU reset to recover.

    To restore operation, the GPU should be reset as soon as possible.
  - This flag is also present on previous generation products (NVIDIA V100) and expects the same client response.
- Row-remapping pending flag
  - This flag indicates that row-remapping will happen at the next GPU reset.
  - Even with the flag set, unaffected applications can continue running without affecting accuracy and performance, and new workloads can be launched.
  - This is useful to identify readiness for live virtual machine (VM) migrations into this GPU, and this GPU should be reset if a live VM will be migrated to it.
- Row-remapping failure flag
  - For definition of row-remapping failure, see RMA Policy Thresholds for Row-Remapping.
- Drain and reset flag
  - NVIDIA A100/H100 GPU supports GPU partitioning feature called Multi Instance GPU (MIG).
  - With MIG enabled, this flag indicates that at least one instance is affected.

    Any work on the other GPU instances should be drained, and the GPU should go through reset at the earliest opportunity for full recovery.
  - Even with this flag set, the applications on unaffected GPU partitions can continue running without any effects on accuracy and performance.

Note: These client flags are currently exposed through SMBPBI.

Chapter 8. User Visible Statistics

Previously, end-users or sysadmins could use the page retirement count to monitor the health of the GPU (and possible RMA conditions) and determine whether to reset the GPU or reload the module for page offlining to take effect. For the same purpose, row-remapping statistics will be exposed to users to give an indication of the health of the GPU memory. This section describes the row-remapping statistics that are available via in-band and out-of-band reporting mechanisms.

- In-band reporting
  - XID error log (see *Table 2* for a list of XID log examples for uncorrectable ECC errors)
    - XID 94: This XID indicates a contained ECC error has occurred
    - XID 95: This XID indicates an uncontained ECC error has occurred
    - XID 63: This XID indicates successful recording of a row-remapping entry to the InfoROM
    - XID 64: This XID indicates a failure in recording a row-remapping entry to the InfoROM
  - NVML/nvidia-smi
    - Number of remapped rows (correctable and uncorrectable)

      This is the number of entries recorded in the InfoROM, not the ones remapped in hardware.
    - Row remapping pending boolean
    - Row remapping failure boolean [1]
    - Bucketized counts
    - Refer to https://docs.nvidia.com/deploy/nvml-api/index.html for more information about NVML.
    - Refer to https://developer.nvidia.com/nvidia-system-management-interface for more information about nvidia-smi.
- Out-of-band reporting (SMBPBI)
  - Number of remapped rows (correctable and uncorrectable), and this is the number of entries recorded in the InfoROM, not the ones remapped in hardware.
  - Row remapping pending boolean

[1] Support will be added with the first NVIDIA Tesla recommended driver (TRD) for A100

- Row remapping failure Boolean
- Bucketized counts
- *Table 3* lists the new SMBPBI APIs for error reporting. For additional details refer to the *NVIDIA SMBus Post-Box Interface (SMBPBI) Software Design Guide* (DG-06034-002).

Table 3. Uncorrectable ECC Errors XID Log Examples

Error Type	XID Log
Contained error with MIG enabled	NVRM: Xid (PCI:0000:01:00 GPU-I:05): 94, pid=7194, Contained: CE User Channel (0x9). RST: No, D-RST: No
Contained error with MIG disabled	NVRM: Xid (PCI:0000:01:00): 94, pid=7062, Contained: CE User Channel (0x9). RST: No, D-RST: No
Uncontained error	NVRM: Xid (PCI:0000:01:00): 95, pid=7062, Uncontained: LTC TAG (0x2,0x0). RST: Yes, D-RST: No

Table 4. SMBPBI APIs for NVIDIA A100 Memory Error Reporting

Opcode	Description
0x1E	Request ECC statistics (format V6)
0x20	Request row-remapping related statistics

Table 5 shows an example of bucketized count where all remapping resources are available.

Table 5. Bucketized Count Example #1

Bucket Name	Bank Remap Availability(shown as reference only)	Number of Banks per Bucket
`Max`	8	`640`
`High`	7	`0`
`Partial`	2 to 6	`0`
`Low`	1	`0`
`None`	0	`0`

Note: The **Bank Remap Availability** column is shown only for reference purposes. The information is not available through APIs or nvidia-smi.

Table 6 shows an example of bucketized count where 635 banks have 8, 3 banks have 7, and 2 banks have one remaining remapping resource available.

Table 6. Bucketized Count Example #2

Bucket Name	Bank Remap Availability (shown as reference only)	Number of Banks per Bucket
Max	8	635
High	7	3
Partial	2 to 6	0
Low	1	2
None	0	0

Note: The **Bank Remap Availability** column is shown only for reference purposes. The information is not available through APIs or nvidia-smi.

Chapter 9. RMA Policy Thresholds for Row-Remapping

The NVIDIA Field Diagnostic tool determines whether a GPU qualifies for RMA. Regarding row-remapping failures, the RMA criteria is met when the row-remapping failure flag is set and validated by the field diagnostic. Any of the following events will trigger a row-remapping failure flag:

- A remapping attempt for an uncorrectable memory error on a bank that already has eight uncorrectable error rows remapped.
- A remapping attempt for an uncorrectable memory error on a row that was already remapped and can occur with less than eight total remaps to the same bank.
- After 512 total remappings for an uncorrectable memory error have occurred.

The row-remapping failure flag is available through in-band (NVML/nvidia-smi) and out-of-band (SMBPBI) tools.

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051
www.nvidia.com