NVIDIA A100 GPU Memory Error Management

Application Note
Table of Contents

Chapter 1. Overview..............................................................................................................1
Chapter 2. Error Containment............................................................................................. 2
Chapter 3. Dynamic Page Blacklisting................................................................................ 3
Chapter 4. Row-Remapping................................................................................................. 4
Chapter 5. Response to Uncorrectable Contained ECC Errors...........................................5
Chapter 6. Error Recovery and Response Flags.................................................................... 6
Chapter 7. User Visible Statistics........................................................................................ 7
Chapter 8. RMA Policy Thresholds for Row-Remapping.................................................... 9
Chapter 1. Overview

NVIDIA® A100 introduces new memory error recovery features that improve resilience and avoids impacting unaffected applications. The new features improve various aspects of the graphics processing unit’s (GPU’s) response to memory errors and thereby improve the overall robustness of the error handling and recovery process.

**Note:** NVIDIA A100 has dedicated resources for ECC and does not incur any overhead in capacity or DRAM interface performance. ECC is enabled by default and disabling ECC is not supported on A100 GPUs.

The error handling and response features are:

- Error-containment
- Dynamic page blacklisting
- Row-remapping
- Uncorrectable error to correctable error coverage improved by 10%

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. Error Containment

NVIDIA A100 GPU 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 A100 GPU 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 prior GPU architectures, NVIDIA A100 does not require a GPU reset when memory errors occur. Error containment also applies to contained errors over from peer NVIDIA A100 GPUs over NVIDIA® NVLink® as well.

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 3. Dynamic Page Blacklisting

Dynamic Page Blacklisting improves resiliency and availability of NVIDIA A100 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 Blacklisting exists on NVIDIA A100 GPUs. It is not available on prior generations of NVIDIA GPUs that do not support error containment.

The NVIDIA A100 GPU does not require a GPU reset to recover from most uncorrectable ECC errors.
Chapter 4. Row-Remapping

Row-remapping is a hardware mechanism to improve the reliability of frame buffer memory on NVIDIA A100. 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 NVIDIA A100 HBM is equipped with spare rows in hardware. As opposed to traditional page retirement, the NVIDIA A100 row-remapper replaces degrading memory cells with spare ones to avoid blacklisting regions of memory in software. This differs from dynamic page blacklisting in that the memory is fixed at a hardware level and thus doesn’t 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 GPU.

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

### Table 1. Page Retirement vs. Row-Remapping

<table>
<thead>
<tr>
<th>Feature</th>
<th>Page Retirement for Legacy GPUs</th>
<th>Row-Remapping for A100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Available remappings/retirements</td>
<td>Supported a maximum of 64 retirements for the frame buffer</td>
<td>Supports up to 512 remapping for the frame buffer.</td>
</tr>
<tr>
<td>Policy changes</td>
<td>Once a retirement takes effect, the page can never be unretired, regardless of correctable or uncorrectable errors</td>
<td>Remapping due to correctable errors can be replaced by uncorrectable error remapping when the memory bank’s reserved rows are exhausted.</td>
</tr>
<tr>
<td>RMA criteria</td>
<td>A threshold of page retirements on a GPU usually resulted in investigation of whether the GPU was worthy of an RMA</td>
<td>See “RMA Policy Thresholds for Row-Remapping.”</td>
</tr>
<tr>
<td>Application of pending changes</td>
<td>Needed a kernel module reload or driver re-initialization or GPU reset</td>
<td>GPU reset is required.</td>
</tr>
</tbody>
</table>
Chapter 5. Response to Uncorrectable Contained ECC Errors

Similar to prior GPU architectures, when an uncorrectable ECC error is detected, the NVIDIA driver software will perform error recovery. Error containment ensures that erroneous data doesn’t propagate further and only 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 blacklisting marks the page containing the faulty memory as unusable. This ensures that new allocations do not land on the page containing the faulty memory. Unaffected applications will continue to run and further 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 blacklisted page is reclaimed.

Figure 1. NVIDIA A100 Response to Uncorrectable Contained ECC Error
Chapter 6. Error Recovery and Response Flags

The following is a list of flags required for client error recovery and response.

- **Reset pending flag**
  - This flag, when set, indicates the GPU has encountered an uncorrectable, uncontained error requiring a GPU reset to recover. The GPU should be reset as soon as possible in order to restore operation.
  - This flag is also present on prior generation products (NVIDIA V100) and expects the same client response.

- **Row-remapping pending flag**
  - This flag indicates row-remapping will happen at the next GPU reset.
  - Even with the flag set, unaffected applications can continue running without any effects on accuracy and performance and new workloads can be launched.
  - This is useful to identify readiness for live virtual machine (VM) migrations into this GPU - 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 Section “RMA Policy Thresholds for Row-Remapping” for details.

- **Drain and reset flag**
  - NVIDIA A100 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.
Previously end-users or sysadmins can use the page retirement count to monitor the health of the GPU (and possible RMA conditions) as well as whether there is a need to reset the GPU or module reload for page blacklisting 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 \(^2\)
    - NVML documentation can be found at https://docs.nvidia.com/deploy/nvml-api/index.html.
    - nvidia-smi documentation can be found at https://developer.nvidia.com/nvidia-system-management-interface

- **Out-of-band reporting (SMBPBI)**
  - Number of remapped rows - correctable and uncorrectable
    - This is the number of entries recorded in the InfoROM, not the ones remapped in hardware.

\(^1\) Support will be added with the first NVIDIA Tesla recommended driver (TRD) for A100
\(^2\) Support will be added with a later driver release
- Row remapping pending boolean
- 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 2. Uncorrectable ECC Errors XID Log Examples

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

### Table 3. SMBPBI APIs for NVIDIA A100 Memory Error Reporting

<table>
<thead>
<tr>
<th>Opcode</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0x1E</td>
<td>Request ECC statistics (format V6)</td>
</tr>
<tr>
<td>0x20</td>
<td>Request row-remapping related statistics</td>
</tr>
</tbody>
</table>

The following table describes the bucketized counts data format. The data in the “Number of Banks” column is an example and for illustration purpose only.

### Table 4. Bucketized Counts Example

<table>
<thead>
<tr>
<th>Remapping Availability per Bank</th>
<th>Basis (information provided here for reference)</th>
<th>Number of Banks (Examples)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max</td>
<td>$N$ [the maximum]</td>
<td>1516 banks</td>
</tr>
<tr>
<td>High</td>
<td>$N - 1$</td>
<td>15 banks</td>
</tr>
<tr>
<td>Partial</td>
<td>range of 2 to $N-2$</td>
<td>5 banks</td>
</tr>
<tr>
<td>Low</td>
<td>1</td>
<td>0 banks</td>
</tr>
<tr>
<td>None</td>
<td>0</td>
<td>0 banks</td>
</tr>
</tbody>
</table>

---

3 Support will be added with a later driver release
Chapter 8. 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. 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 8 uncorrectable error rows remapped
- A remapping attempt for an uncorrectable memory error on a row that was already remapped (which can occur with less than 8 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.
ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, "MATERIALS") ARE BEING PROVIDED "AS IS." NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE.

Information furnished is believed to be accurate and reliable. However, NVIDIA Corporation assumes no responsibility for the consequences of use of such information or for any infringement of patents or other rights of third parties that may result from its use. No license is granted by implication of otherwise under any patent rights of NVIDIA Corporation. Specifications mentioned in this publication are subject to change without notice. This publication supersedes and replaces all other information previously supplied. NVIDIA Corporation products are not authorized as critical components in life support devices or systems without express written approval of NVIDIA Corporation.

NVIDIA and the NVIDIA logo are trademarks or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

© 2013-2020 NVIDIA Corporation. All rights reserved.