CUDA-MEMCHECK

DU-05355-001_v11.0 | August 2020

User Manual

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1. INTRODUCTION

1.1. About CUDA-MEMCHECK

CUDA-MEMCHECK is a functional correctness checking suite included in the CUDA toolkit. This suite contains multiple tools that can perform different types of checks. The *memcheck* tool is capable of precisely detecting and attributing out of bounds and misaligned memory access errors in CUDA applications. The tool also reports hardware exceptions encountered by the GPU. The *racecheck* tool can report shared memory data access hazards that can cause data races. The *initcheck* tool can report cases where the GPU performs uninitialized accesses to global memory. The *synccheck* tool can report cases where the application is attempting invalid usages of synchronization primitives. This document describes the usage of these tools.

CUDA-MEMCHECK can be run in *standalone mode* where the user's application is started under CUDA-MEMCHECK. The *memcheck* tool can also be enabled in *integrated mode* inside CUDA-GDB.

1.2. Why CUDA-MEMCHECK?

NVIDIA allows developers to easily harness the power of GPUs to solve problems in parallel using CUDA. CUDA applications often run thousands of threads in parallel. Every programmer invariably encounters memory access errors and thread ordering hazards that are hard to detect and time consuming to debug. The number of such errors increases substantially when dealing with thousands of threads. The CUDA-MEMCHECK suite is designed to detect those problems in your CUDA application.

1.3. How to Get CUDA-MEMCHECK

CUDA-MEMCHECK is installed as part of the CUDA toolkit.

1.4. CUDA-MEMCHECK tools

Tools allow use the basic CUDA-MEMCHECK infrastructure to provide different checking mechanisms. Currently, the supported tools are :

- *Memcheck* - The memory access error and leak detection tool. See Memcheck Tool
- *Racecheck* - The shared memory data access hazard detection tool. See Racecheck Tool
- *Initcheck* - The unitialized device global memory access detection tool. See Initcheck Tool
- *Synccheck* - The thread synchronization hazard detection tool. See Synccheck Tool

Table 1 Supported Modes by CUDA-MEMCHECK tool

Tool Name	Standalone Mode	Integrated Mode
Memcheck	Yes	Yes
Racecheck	Yes	No
Initcheck	Yes	No
Synccheck	Yes	No

Chapter 2. USING CUDA-MEMCHECK

CUDA-MEMCHECK tools can be invoked by running the **cuda-memcheck** executable as follows:

```
cuda-memcheck [options] app_name [app_options]
```

For a full list of options that can be specified to memcheck and their default values, see Command Line Options.

2.1. Command Line Options

Command line options can be specified to **cuda-memcheck**. With some exceptions, the options to memcheck are usually of the form **--option value**. The option list can be terminated by specifying **--**. All subsequent words on the command line are treated as the application being run and its arguments.

The table below describes the supported options in detail. The first column is the option name as passed to CUDA-MEMCHECK. Some options have a one character short form, which is given in parentheses. These options can be invoked using a single hyphen. For example, the help option can be invoked as **-h**. The options that have a short form do not take a value.

The second column contains the permissible values for the option. In case the value is user defined, this is shown below in braces {}. An option that can accept any numerical value is represented as *{number}* .

The third column contains the default value of the option. Some options have different default values depending on the architecture they are being run on.

Table 2 CUDA-MEMCHECK Command line options

Option	Values	Default	Description
binary-patching	yes, no	yes	Controls whether CUDA-MEMCHECK should modify the application binary at runtime. This option is enabled by default. Setting this to "no" will reduce

Option	Values	Default	Description
			the precision of errors reported by the tool. Normal users will not need to modify this flag.
check-deprecated-instr	yes, no	no	When enabled, CUDA-MEMCHECK will report errors when deprecated instructions are detected in executed kernels. Which instructions are reported depends on the selected tool. This option is disabled by default.
demangle	full, simple, no	full	Enables demangling of device function names. For more information, see Name Demangling.
destroy-on-device-error	context,kernel	context	This controls how the application proceeds on hitting a memory access error. For more information, see Error Actions.
error-exitcode	{number}	0	The exit code CUDA-MEMCHECK will return if the original application succeeded but memcheck detected errors were present. This is meant to allow CUDA-MEMCHECK to be integrated into automated test suites
filter	{key1=val1} [{,key2=val2}]	N/A	Controls which application kernels will be checked by the running CUDA-MEMCHECK tool. For more information, see Specifying Filters.
flush-to-disk	yes,no	no	Forces every disk write to be flushed to disk. When enabled, this will make CUDA-MEMCHECK tools much slower.
force-blocking-launches	yes,no	no	This forces all host kernel launches to be sequential. When enabled, the number and precision of memcheck reported errors will decrease.
help (h)	N/A	N/A	Displays the help message
language	c,fortran	c	This controls application source language specific behavior in CUDA-MEMCHECK tools. For fortan specific behavior, see CUDA Fortran Specific Behavior.
log-file	{filename}	N/A	This is the file CUDA-MEMCHECK will write all of its text output to. By default, CUDA-MEMCHECK will print all output to stdout. For more information, see Escape Sequences.
prefix	{string}	========	The string prepended to CUDA-MEMCHECK output lines
print-level	info,warn,error,fatal	warn	The minimum level print level of messages from CUDA-MEMCHECK.

Option	Values	Default	Description
print-limit	{number}	10000	When this option is set, memcheck will stop printing errors after reaching the given number of errors. Use 0 for unlimited printing.
read	{filename}	N/A	The input CUDA-MEMCHECK file to read data from. This can be used in conjunction with the --save option to allow processing records after a run.
save	{filename}	N/A	Filename where CUDA-MEMCHECK will save the output from the current run. For more information, see Escape Sequences.
show-backtrace	yes,host,device,no	yes	Displays a backtrace for most types of errors. No disables all backtraces, Yes enables all backtraces. Host enables only host side backtraces. Device enables only device side backtraces. For more information, see Stack Backtraces.
tool	memcheck, racecheck, initcheck, synccheck	memcheck	Controls which CUDA-MEMCHECK tool is actively running
version (V)	N/A	N/A	Prints the version of cuda-memcheck

Table 3 *Memcheck* Tool Command line options

Option	Values	Default	Description
check-api-memory-access	yes,no	yes	Enable checking of cudaMemcpy/ cudaMemset
check-device-heap	yes,no	yes	Enable checking of device heap allocations. This applies to both error checking and leak checking.
leak-check	full,no	no	Prints information about all allocations that have not been freed via cudaFree at the point when the context was destroyed. For more information, see Leak Checking.
report-api-errors	all, explicit, no	explicit	Report errors if any CUDA API call fails. For more information, see CUDA API Error Checking.

Table 4 *Racecheck* Tool Command line options

Option	Values	Default	Description
racecheck-report	hazard,analysis,all	analysis	Controls how racecheck reports information. For more information, see Racecheck report modes.

2.2. Supported Operating Systems

The standalone CUDA-MEMCHECK binary is supported on all CUDA supported platforms i.e. Windows, supported Linux distributions and Android. CUDA-MEMCHECK can interoperate with CUDA-GDB on Android and Linux.

2.3. Supported Devices

The CUDA-MEMCHECK tool suite is supported on all CUDA capable GPUs with SM versions 3.5 and above.

Virtual GPUs (such as NVIDIA GRID) are not supported by CUDA-MEMCHECK.

CUDA-MEMCHECK tools are not supported when Windows Hardware-accelerated GPU scheduling is enabled. For such cases the compute-sanitizer tool should be used as a replacement for CUDA-MEMCHECK.

2.4. Compilation Options

The CUDA-MEMCHECK tools do not need any special compilation flags to function.

The output displayed by the CUDA-MEMCHECK tools is more useful with some extra compiler flags. The **`-G`** option to nvcc forces the compiler to generate debug information for the CUDA application. To generate line number information for applications without affecting the optimization level of the output, the **`-lineinfo`** option to nvcc can be used. The CUDA-MEMCHECK tools fully support both of these options and can display source attribution of errors for applications compiled with line information.

The stack backtrace feature of the CUDA-MEMCHECK tools is more useful when the application contains function symbol names. For the host backtrace, this varies based on the host OS. On Linux, the host compiler must be given the **`-rdynamic`** option to retain function symbols. On Windows, the application must be compiled for debugging, i.e. the **`/Zi`** option. When using nvcc, flags to the host compiler can be specified using the **`-Xcompiler`** option. For the device backtrace, the full frame information is only available when the application is compiled with device debug information. The compiler can skip generation of frame information when building with optimizations.

Sample command line to build with function symbols and device side line information on linux:

```
nvcc -Xcompiler -rdynamic -lineinfo  -o out in.cu
```

2.5. Environment variables

The CUDA-MEMCHECK tools can fail to initialize when there are a lot of CUDA functions in the target app. This is due to CUDA-MEMCHECK trying to find a

subset of functions to patch and running out of memory. The environment variable CUDA_MEMCHECK_PATCH_MODULE can be set to 1 in order to bypass this behavior, thus resolving the initialization error.

Chapter 3. MEMCHECK TOOL

3.1. What is Memcheck ?

The *memcheck* tool is a run time error detection tool for CUDA applications. The tool can precisely detect and report out of bounds and misaligned memory accesses to global, local, shared and global atomic instructions in CUDA applications. It can also detect and report hardware reported error information. In addition, the memcheck tool can detect and report memory leaks in the user application.

3.2. Supported Error Detection

The errors that can be reported by the memcheck tool are summarized in the table below. The location column indicates whether the report originates from the host or from the device. The precision of an error is explained in the paragraph below.

Table 5 Memcheck reported error types

Name	Description	Location	Precision	See also
Memory access error	Errors due to out of bounds or misaligned accesses to memory by a global, local, shared or global atomic access.	Device	Precise	Memory Access Error Reporting
Hardware exception	Errors that are reported by the hardware error reporting mechanism.	Device	Imprecise	Hardware Exception Reporting
Malloc/Free errors	Errors that occur due to incorrect use of `malloc()`/`free()` in CUDA kernels.	Device	Precise	Device Side Allocation Checking
CUDA API errors	Reported when a CUDA API call in the application returns a failure.	Host	Precise	CUDA API Error Checking

Name	Description	Location	Precision	See also
cudaMalloc memory leaks	Allocations of device memory using `cudaMalloc()` that have not been freed by the application.	Host	Precise	Leak Checking
Device Heap Memory Leaks	Allocations of device memory using `malloc()` in device code that have not been freed by the application.	Device	Imprecise	Device Side Allocation Checking

The memcheck tool reports two classes of errors *precise* and *imprecise*.

Precise errors in memcheck are those that the tool can uniquely identify and gather all information for. For these errors, memcheck can report the block and thread coordinates of the thread causing the failure, the program counter (PC) of the instruction performing the access, as well as the address being accessed and its size and type. If the CUDA application contains line number information (by either being compiled with device side debugging information, or with line information), then the tool will also print the source file and line number of the erroneous access.

Imprecise errors are errors reported by the hardware error reporting mechanism that could not be precisely attributed to a particular thread. The precision of the error varies based on the type of the error and in many cases, memcheck may not be able to attribute the cause of the error back to the source file and line.

3.3. Using Memcheck

The memcheck tool is enabled by default when running the CUDA-MEMCHECK application. It can also be explicitly enabled by using the **`--tool memcheck`** option.

```
cuda-memcheck [memcheck_options] app_name [app_options]
```

When run in this way, the memcheck tool will look for precise, imprecise, malloc/free and CUDA API errors. The reporting of device leaks must be explictly enabled. Errors identified by the memcheck tool are displayed on the screen after the application has completed execution. See Understanding Memcheck Errors for more information about how to interpret the messages printed by the tool.

3.4. Understanding Memcheck Errors

The memcheck tool can produce a variety of different errors. This is a short guide showing some samples of errors and explaining how the information in each error report can be interpreted.

1. *Memory access error*: Memory access errors are generated for errors that the memcheck tool can correctly attribute and identify the erroneous instruction. Below is an example of a precise memory access error

```
========= Invalid __global__ write of size 4
=========     at 0x00000060 in memcheck_demo.cu:6:unaligned_kernel(void)
=========     by thread (0,0,0) in block (0,0,0)
=========     Address 0x400100001 is misaligned
```

Let us examine this error line by line :

```
Invalid __global__ write of size 4
```

The first line shows the memory segment, type and size being accessed. The memory segment is one of :

- __global__ : for device global memory
- __shared__ : for per block shared memory
- __local__ : for per thread local memory

In this case, the access was to device global memory. The next field contains information about the type of access, whether it was a read or a write. In this case, the access is a write. Finally, the last item is the size of the access in bytes. In this example, the access was 4 bytes in size.

```
at 0x00000060 in memcheck_demo.cu:6:unaligned_kernel(void)
```

The second line contains the PC of the instruction, the source file and line number (if available) and the CUDA kernel name. In this example, the instruction causing the access was at PC 0x60 inside the **`unaligned_kernel`** CUDA kernel. Additionally, since the application was compiled with line number information, this instruction corresponds to line 6 in the memcheck_demo.cu source file.

```
by thread (0,0,0) in block (0,0,0)
```

The third line contains the thread indices and block indices of the thread on which the error was hit. In this example, the thread doing the erroneous access belonged to the first thread in the first block.

```
Address 0x400100001 is misaligned
```

The fourth line contains the memory address being accessed and the type of of access error. The type of access error can either be out of bounds access or misaligned access. In this example, the access was to address 0x400100001 and the access error was because this address was not aligned correctly.

2. *Hardware exception*: Imprecise errors are generated for errors that the hardware reports to the memcheck tool. Hardware exceptions have a variety of formats and messages. Typically, the first line will provide some information about the type of error encountered.

3. *Malloc/free error*: Malloc/free errors refer to the errors in the invocation of device side **`malloc()/free()`** in CUDA kernels. An example of a malloc/free error :

```
========= Malloc/Free error encountered : Double free
=========     at 0x000079d8
=========     by thread (0,0,0) in block (0,0,0)
=========     Address 0x400aff920
```

We can examine this line by line.

```
Malloc/Free error encountered : Double free
```

The first line indicates that this is a malloc/free error, and contains the type of error. This type can be :

- Double free : This indicates that the thread called **`free()`** on an allocation that has already been freed.
- Invalid pointer to free : This indicates that **`free`** was called on a pointer that was not returned by **`malloc()`**
- Heap corruption : This indicates generalized heap corruption, or cases where the state of the heap was modified in a way that memcheck did not expect

In this example, the error is due to calling **`free()`** on a pointer which had already been freed.

```
at 0x000079d8
```

The second line gives the PC on GPU where the error was reported. This PC is usually inside of system code, and is not interesting to the user. The device frame backtrace will contain the location in user code where the **`malloc()/free()`** call was made.

```
by thread (0,0,0) in block (0,0,0)
```

The third line contains the thread and block indices of the thread that caused this error. In this example, the thread has threadIdx = (0,0,0) and blockIdx = (0,0,0)

```
Address 0x400aff920
```

This line contains the value of the pointer passed to **`free()`** or returned by **`malloc()`**

4. *Leak errors*: Errors are reported for allocations created using cudaMalloc and for allocations on the device heap that were not freed when the CUDA context was destroyed. An example of a cudaMalloc allocation leak report follows :

   ```
   ========= Leaked 64 bytes at 0x400200200
   ```

   The error message reports information about the size of the allocation that was leaked as well as the address of the allocation on the device.

   A device heap leak message will be explicitly identified as such:

   ```
   ========= Leaked 16 bytes at 0x4012ffff6 on the device heap
   ```

5. *CUDA API error*: CUDA API errors are reported for CUDA API calls that return an error value. An example of a CUDA API error:

   ```
   ========= Program hit error 11 on CUDA API call to cudaMemset
   ```

   The message contains the returned value of the CUDA API call, as well as the name of the API function that was called.

3.5. Integrated Mode

You can execute the memcheck tool from within CUDA-GDB by using the following option before running the application:

```
(cuda-gdb) set cuda memcheck on
```

In integrated mode, the memcheck tool improves the precision of error reporting by CUDA-GDB. The memory access checks are enabled, allowing identification of the thread that may be causing a warp or device level exception.

3.6. CUDA API Error Checking

The memcheck tool supports reporting an error if a CUDA API call made by the user program returned an error. The tool supports this detection for both CUDA run time and CUDA driver API calls. In all cases, if the API function call has a nonzero return value, CUDA-MEMCHECK will print an error message containing the name of the API call that failed and the return value of the API call.

CUDA API error reports do not terminate the application, they merely provide extra information. It is up to the application to check the return status of CUDA API calls and handle error conditions appropriately.

3.7. Device Side Allocation Checking

The *memcheck* tool checks accesses to allocations in the device heap.

These allocations are created by calling **`malloc()`** inside a kernel. This feature is implicitly enabled and can be disabled by specifying the **`--check-device-heap no`** option. This feature is only activated for kernels in the application that call **`malloc()`**.

The current implementation does not require space on the device heap, and so the heap allocation behavior of the program with and without memcheck should remain similar. The *memcheck* tool does require space in device global memory to track these heap allocations and will print an internal error message if it is not able to allocate this space in device global memory.

In addition to access checks, the *memcheck* tool can now perform libc style checks on the **`malloc()/free()`** calls. The tool will report an error if the application calls a **`free()`** twice on a kernel, or if it calls **`free()`** on an invalid pointer.

Make sure to look at the device side backtrace to find the location in the application where the **`malloc()/free()`** call was made

3.8. Leak Checking

The *memcheck* tool can detect leaks of allocated memory.

Memory leaks are device side allocations that have not been freed by the time the context is destroyed. The *memcheck* tool tracks device memory allocations created using the CUDA driver or runtime APIs. Starting in CUDA 5, allocations that are created dynamically on the device heap by calling **`malloc()`** inside a kernel are also tracked.

For an accurate leak checking summary to be generated, the application's CUDA context must be destroyed at the end. This can be done explicitly by calling **`cuCtxDestroy()`** in applications using the CUDA driver API, or by calling **`cudaDeviceReset()`** in applications programmed against the CUDA run time API.

The **`--leak-check full`** option must be specified to enable leak checking.

Chapter 4. RACECHECK TOOL

4.1. What is Racecheck ?

The *racecheck* tool is a run time shared memory data access hazard detector. The primary use of this tool is to help identify memory access race conditions in CUDA applications that use shared memory.

In CUDA applications, storage declared with the **__shared__** qualifier is placed in on chip *shared memory*. All threads in a thread block can access this per block shared memory. Shared memory goes out of scope when the thread block completes execution. As shared memory is on chip, it is frequently used for inter thread communication and as a temporary buffer to hold data being processed. As this data is being accessed by multiple threads in parallel, incorrect program assumptions may result in data races. Racecheck is a tool built to identify these hazards and help users write programs free of shared memory races.

Currently, this tool only supports detecting accesses to on-chip shared memory. For supported architectures, see Supported Devices.

4.2. What are Hazards?

A *data access hazard* is a case where two threads attempt to access the same location in memory resulting in nondeterministic behavior, based on the relative order of the two accesses. These hazards cause *data races* where the behavior or the output of the application depends on the order in which all parallel threads are executed by the hardware. Race conditions manifest as intermittent application failures or as failures when attempting to run a working application on a different GPU.

The racecheck tool identifies three types of canonical hazards in a program. These are :

- Write-After-Write (*WAW*) hazards

This hazard occurs when two threads attempt to write data to the same memory location. The resulting value in that location depends on the relative order of the two accesses.

- Write-After-Read (*WAR*) hazards

  This hazard occurs when two threads access the same memory location, with one thread performing a read and another a write. In this case, the writing thread is ordered before the reading thread and the value returned to the reading thread is not the original value at the memory location.
- Read-After-Write (*RAW*) hazards

  This hazard occurs when two threads access the same memory location, with one thread performing a read and the other a write. In this case, the reading thread reads the value before the writing thread commits it.

4.3. Using Racecheck

The racecheck tool is enabled by running the CUDA-MEMCHECK application with the **`--tool racecheck`** option.

```
cuda-memcheck --tool racecheck [memcheck_options] app_name [app_options]
```

Once racecheck has identified a hazard, the user can make program modifications to ensure this hazard is no longer present. In the case of Write-After-Write hazards, the program should be modified so that multiple writes are not happening to the same location. In the case of Read-After-Write and Write-After-Read hazards, the reading and writing locations should be deterministically ordered. In CUDA kernels, this can be achieved by inserting a **`__syncthreads()`** call between the two accesses. To avoid races between threads within a single warp, **`__syncwarp()`** can be used.

The racecheck tool does not perform any memory access error checking. It is recommended that users first run the memcheck tool to ensure the application is free of errors

4.4. Racecheck report modes

The racecheck tool can produce two types of output :

- *Hazard* reports

  These reports contain detailed information about one particular hazard. Each hazard report is byte accurate and represents information about conflicting accesses between two threads that affect this byte of shared memory.
- *Analysis* reports

  These reports contain a post analysis set of reports. These reports are produced by the racecheck tool by analysing multiple hazard reports and examining active device

state. For example usage of analysis reports, see Understanding Racecheck Analysis Reports.

4.5. Understanding Racecheck Analysis Reports

In *analysis* reports, the racecheck tool produces a series of high level messages that identify the source locations of a particular race, based on observed hazards and other machine state

A sample racecheck analysis report is below:

```
========= ERROR: Race reported between Write access at 0x00000050 in
 raceGroupBasic.cu:53:WAW(void)
=========     and Write access at 0x00000050 in raceGroupBasic.cu:53:WAW(void)
```

The analysis record contains high level information about the hazard that is conveyed to the end user. Each line contains information about a unique location in the application which is participating in the race.

The first word on the first line indicates the severity of this report. In this case, the message is at the ERROR level of severity. For more information on the different severity levels, see Racecheck Severity Levels. Analysis reports are composed of one or more racecheck hazards, and the severity level of the report is that of the hazard with the highest severity.

The first line additionally contains the type of access. The access can be either:

- Read
- Write

The next item on the line is the PC of the location where the access happened from. In this case, the PC is 0x50. If the application was compiled was compiled with line number information, this line would also contain the file name and line number of the access. Finally, the line contains the kernel name of the kernel issuing the access.

A given analysis report will always contain at least one line which is performing a write access. A common strategy to eliminate races which contain only write accesses is to ensure that the write access is performed by only one thread. In the case of races with multiple readers and one writer, introducing explicit program ordering via a **`__syncthreads()`** call can avoid the race condition. For races between threads within the same warp, the **`__syncwarp()`** intrinsic can be used to avoid the hazard.

4.6. Understanding Racecheck Hazard Reports

In *hazard* reporting mode, the racecheck tool produces a series of messages detailing information about hazards in the application. The tool is byte accurate and produces a message for each byte on which a hazard was detected. Additionally, when enabled, the host backtrace for the launch of the kernel will also be displayed.

A sample racecheck hazard is below:

```
========= ERROR: Potential WAW hazard detected at __shared__ 0x0 in block (0, 0,
 0) :
=========     Write Thread (0, 0, 0) at 0x00000088 in raceWAW.cu:18:WAW(void)
=========     Write Thread (1, 0, 0) at 0x00000088 in raceWAW.cu:18:WAW(void)
=========     Current Value : 0, Incoming Value : 2
```

The hazard records are dense and capture a lot of interesting information. In general terms, the first line contains information about the hazard severity, type and address, as well as information about the thread block where it occurred. The next 2 lines contain detailed information about the two threads that were in contention. These two lines are ordered chronologically, so the first entry is for the access that occurred earlier and the second for the access that occurred later. The final line is printed for some hazard types and captures the actual data that was being written.

Examining this line by line, we have :

```
ERROR: Potential WAW hazard detected at __shared__ 0x0 in block (0, 0, 0)
```

The first word on this line indicates the severity of this hazard. In this case, the message is at the ERROR level of severity. For more information on the different severity levels, see Racecheck Severity Levels.

The next piece of information here is the type of hazard. The racecheck tool detects three types of hazards:

- WAW or Write-After-Write hazards
- WAR or Write-After-Read hazards
- RAW or Read-After-Write hazards

The type of hazard indicates the accesses types of the two threads that were in contention. In this example, the hazard is of Write-After-Write type.

The next piece of information is the address in shared memory that was being accessed. This is the offset in per block shared memory that was being accessed by both threads. Since the racecheck tool is byte accurate, the message is only for the byte of memory at given address. In this example, the byte being accessed is byte 0x0 in shared memory.

Finally, the first line contains the block index of the thread block to which the two racing threads belong.

The second line contains information about the first thread to write to this location.

```
Write Thread (0, 0, 0) at 0x00000088 in raceWAW.cu:18:WAW(void)
```

The first item on this line indicates the type of access being performed by this thread to the shared memory address. In this example, the thread was writing to the location. The next component is the index of the thread the thread block. In this case, the thread is at index (0,0,0). Following this, we have the byte offset of the instruction which did the access in the kernel. In this example, the offset is 0x88. This is followed by the source file and line number (if line number information is available). The final item on this line is the name of the kernel that was being executed.

The third line contains similar information about the second thread which was causing this hazard. This line has an identical format to the previous line.

The fourth line contains information about the data in the two accesses.

```
Current Value : 0, Incoming Value : 2
```

If the second thread in the hazard was performing a write access, i.e. the hazard is a Write-After-Write (WAW) or a Write-After-Read (WAR) this line contains the value after the access by the first thread as the *Current Value* and the value that will be written by the second access as the *Incoming Value*. In this case, the first thread wrote the value 0 to the shared memory location. The second thread is attempting to write the value 2.

4.7. Racecheck Severity Levels

Problems reported by racecheck can be of different severity levels. Depending on the level, different actions are required from developers. By default, only issues of severity level WARNING and ERROR are shown. The command line option **`--print-level`** can be used to set the lowest severity level that should be reprted.

Racecheck reports have one of the following severity levels:

- *INFO* : The lowest level of severity. This is for hazards that have no impact on program execution and hence are not contributing to data access hazards. It is still a good idea to find and eliminate such hazards
- *WARNING* : Hazards at this level of severity are determined to be programming model hazards, however may be intentionally created by the programmer.
  An example of this are hazards due to warp level programming that make the assumption that threads are proceeding in groups. Such hazards are typically only encountered by advanced programmers. In cases where a beginner programmer encounters such errors, he should treat them as sources of hazards.

  Starting with the Volta architecture, programmers cannot rely anymore on the assumption that threads within a warp execute in lock-step unconditionally.
  As a result, warnings due to warp-synchronous programming without explicit synchronization must be fixed when developing or porting applications from earlier architectures to Volta and above. Developers can use the **`__syncwarp()`** intrinsic or the Cooperative Groups API.
- ERROR : The highest level of severity. Correspond to hazards that are very likely candidates for causing data access races. Programmers would be well advised to examine errors at this level of severity.

Chapter 5. INITCHECK TOOL

5.1. What is Initcheck ?

The *initcheck* tool is a run time uninitialized device global memory access detector. This tool can identify when device global memory is accessed without it being initialized via device side writes, or via CUDA memcpy and memset API calls.

Currently, this tool only supports detecting accesses to device global memory. For supported architectures, see Supported Devices.

5.2. Using Initcheck

The initcheck tool is enabled by running the CUDA-MEMCHECK application with the **`--tool initcheck`** option.

```
cuda-memcheck --tool initcheck [memcheck_options] app_name [app_options]
```

The initcheck tool does not perform any memory access error checking. It is recommended that users first run the memcheck tool to ensure the application is free of errors

Chapter 6. SYNCCHECK TOOL

6.1. What is Synccheck ?

The *synccheck* tool is a runtime tool that can identify whether a CUDA application is correctly using synchronization primitives, specifically **`__syncthreads()`** and **`__syncwarp()`** intrinsics and their Cooperative Groups API counterparts.

For supported architectures, see Supported Devices.

6.2. Using Synccheck

The synccheck tool is enabled by running the CUDA-MEMCHECK application with the **`--tool synccheck`** option.

```
cuda-memcheck --tool synccheck [memcheck_options] app_name [app_options]
```

The synccheck tool does not perform any memory access error checking. It is recommended that users first run the memcheck tool to ensure the application is free of errors

6.3. Understanding Synccheck Reports

For each violation, the synccheck tool produces a report message that identifies the source location of the violation and its classification.

A sample synccheck report is below:

```
========= Barrier error detected. Divergent thread(s) in block
=========     at 0x00000130 in divergence.cu:61:threadDivergence(int*)
=========     by thread (37,0,0) in block (0,0,0)
=========
========= ERROR SUMMARY: 1 error
```

Each report starts with "Barrier error detected". In most cases, this is followed by a classification of the detected barrier error. In this message, a CUDA block with divergent threads was found. The following error classes can be reported:

- *Divergent thread(s) in block* : Divergence between threads within a block was detected for a barrier that does not support this on the current architecture. For example, this occurs when **`__syncthreads()`** is used within conditional code but the conditional does not evaluate equally across all threads in the block.
- *Divergent thread(s) in warp* : Divergence between threads within a single warp was detected for a barrier that does not support this on the current architecture.
- *Invalid arguments* : A barrier instruction or primitive was used with invalid arguments. This can occur for example if not all threads reaching a **`__syncwarp()`** declare themselves in the mask parameter.
- *Deprecated instruction* : A deprecated instruction was detected. Synccheck currently reports the following deprecated instructions:
  - __shfl(), __shfl_up(), __shfl_down(), __shfl_xor()
  - __any(), __all()
  - __ballot()
- *Unknown error* : synccheck does not recognize this particular error class. This can occur if the CUDA driver is newer than the CUDA-MEMCHECK utility.

The next line states the PC of the location where the access happened. In this case, the PC is 0x130. If the application was compiled with line number information, this line would also contain the file name and line number of the access, followed by the name of the kernel issuing the access.

The third line contains information on the thread and block for which this violation was detected. In this case, it is thread 37 in block 0.

Chapter 7. CUDA-MEMCHECK FEATURES

7.1. Nonblocking Mode

By default, the standalone CUDA-MEMCHECK tool will launch kernels in nonblocking mode. This allows the tool to support error reporting in applications running concurrent kernels

To force kernels to execute serially, a user can use the **`--force-blocking-launches yes`** option. One side effect is that when in blocking mode, only the first thread to hit an error in a kernel will be reported.

7.2. Stack Backtraces

In standalone mode, CUDA-MEMCHECK can generate backtraces when given **`--show-backtrace`** option. Backtraces usually consist of two sections - a saved host backtrace that leads upto the CUDA driver call site, and a device backtrace at the time of the error. Each backtrace contains a list of frames showing the state of the stack at the time the backtrace was created.

To get function names in the host backtraces, the user application must be built with support for symbol information in the host application. For more information, see Compilation Options

In CUDA 5, the host stack backtrace will show a maximum of 61 frames. Some device frames are internal and will not be shown in the backtrace. Instead, a placeholder message like the following will be inserted:

```
=========     Device Frame:<1 frames were hidden>
```

Backtraces are printed for most CUDA-MEMCHECK tool outputs, and the information generated varies depending on the type of output. The table below explains the kind of host and device backtrace seen under different conditions.

Table 6 CUDA-MEMCHECK Stack Backtrace Information

Output Type	Host Backtrace	Device Backtrace
Memory access error	Kernel launch on host	Precise backtrace on device
Hardware exception	Kernel launch on host	Imprecise backtrace on device [1]
Malloc/Free error	Kernel launch on host	Precise backtrace on device
cudaMalloc allocation leak	Callsite of cudaMalloc	N/A
CUDA API error	Callsite of CUDA API call	N/A
CUDA-MEMCHECK internal error	Callsite leading to internal error	N/A
Device heap allocation leak	N/A	N/A
Shared memory hazard	Kernel launch on host	N/A

7.3. Name Demangling

The CUDA-MEMCHECK suite now supports displaying mangled and demangled names for CUDA kernels and CUDA device functions. By default, tools display the fully demangled name, which contains the name of the kernel as well as its prototype information. In the simple demangle mode, the tools will only display the first part of the name. If demangling is disabled, tools will display the complete mangled name of the kernel.

7.4. Dynamic Parallelism

The CUDA-MEMCHECK tool suite supports dynamic parallelism. The *memcheck* tool supports precise error reporting of out of bounds and misaligned accesses on global, local and shared memory accesses as well as on global atomic instructions for applications using dynamic parallelism. In addition, the imprecise hardware exception reporting mechanism is also fully supported. Error detection on applications using dynamic parallelism requires significantly more memory on the device and as a result, in memory constrained environments, *memcheck* may fail to initialize with an internal out of memory error.

For limitations, see Known Issues.

7.5. Error Actions

On encountering an error, CUDA-MEMCHECK behavior depends on the type of error. The default behavior of CUDA-MEMCHECK is to continue execution on purely host

[1] In some cases, there may be no device backtrace

side errors. Hardware exceptions detected by the memcheck tool cause the CUDA context to be destroyed. Precise errors (such as memory access and malloc/free errors) detected by the memcheck tool cause the kernel to be terminated. This terminates the kernel without running any subsequent instructions and the application continues launching other kernels in the CUDA context. The handling of memory access and malloc/free errors detected by the memcheck tool can be changed using the **`--destroy-on-device-error`** option.

For racecheck detected hazards, the hazard is reported, but execution is not affected.

For a full summary of error action, based on the type of the error see the table below. The error action *terminate kernel* refers to the cases where the kernel is terminated early, and no subsequent instructions are run. In such cases, the CUDA context is not destroyed and other kernels continue execution and CUDA API calls can still be made.

When kernel execution is terminated early, the application may not have completed its computations on data. Any subsequent kernels that depend on this data will have undefined behavior.

The action *terminate CUDA context* refers to the cases where the CUDA context is forcibly terminated. In such cases, all outstanding work for the context is terminated and subsequent CUDA API calls will fail. The action *continue application* refers to cases where the application execution is not impacted, and the kernel continues executing instructions.

Table 7 CUDA-MEMCHECK Error Actions

Error Type	Location	Action	Comments
Memory access error	Device	Terminate CUDA context	User can choose to instead terminate the kernel
Hardware exception	Device	Terminate CUDA context	Subsequent calls on the CUDA context will fail
Malloc/Free error	Device	Terminate CUDA context	User can choose to instead terminate the kernel
cudaMalloc allocation leak	Host	Continue application	Error reported. No other action taken.
CUDA API error	Host	Continue application	Error reported. No other action taken.
Device heap allocation leak	Device	Continue application	Error reported. No other action taken.
Shared memory hazard	Device	Continue application	Error reported. No other action taken.
Synchronization error	Device	Terminate CUDA context	User can choose to instead terminate the kernel
CUDA-MEMCHECK internal error	Host	Undefined	The application may behave in an undefined fashion

7.6. Escape Sequences

The **--save** and **--log-file** options to CUDA-MEMCHECK accept the following escape sequences in the file name.

- %% : Replaced with a literal %
- %p : Replaced with the PID of the CUDA-MEMCHECK frontend application.
- %q{ENVVAR} : Replaced with the contents of the environment variable 'ENVVAR'. If the variable does not exist, this is replaced with an empty string.
- Any other character following the % causes an error.

7.7. Specifying Filters

CUDA-MEMCHECK tools support filtering the choice of kernels which should be checked. When a filter is specified, only kernels matching the filter will be checked. Filters are specified using the **--filter** option. By default, CUDA-MEMCHECK tools will check all kernels in the application.

The **--filter** option can be specified multiple times. If a kernel satisfies any filter, it will be checked by the running CUDA-MEMCHECK tool.

The **--filter** takes a filter specification consisting of a list of comma separated key value pairs, specified as **key=value**. In order for a filter to be matched, all components of the filter specification must be satisfied. If a filter is incorrectly specified in any component, the entire filter is ignored. For a full summary of valid key values, see the table below. If a key has multiple strings, any of the strings can be used to specify that filter component.

Table 8 CUDA-MEMCHECK Filter Keys

Name	Key String	Value	Comments
Kernel Name	kernel_name, kne	Complete mangled kernel name	User specifies the complete mangled kernel name. Cannot be included in same filter specification as kernel_substring
Kernel Substring	kernel_substring, kns	Any substring in mangled kernel name	User specifies a substring in the mangled kernel name. Cannot be included in same filter specification as kernel_name.

When using the **kernel_name** or **kernel_substring** filters, CUDA-MEMCHECK tools will check all **device** function calls made by the kernel. When using CUDA Dynamic Parallelism (CDP), CUDA-MEMCHECK tools will not check child kernels launched from a checked kernel unless the child kernel matches a filter. If a GPU launched kernel that does not match a filter calls a device function that is reachable from a kernel that

does match a filter, the device function will behave as though it was checked. In the case of some tools, this can result in undefined behavior.

Chapter 8. OPERATING SYSTEM SPECIFIC BEHAVIOR

This section describes operating system specific behavior.

8.1. Windows Specific Behavior

- Timeout Detection and Recovery (TDR)

  On Windows Vista and above, GPUs have a timeout associated with them. GPU applications that take longer than the threshold (default of 2 seconds) will be killed by the operating system. Since CUDA-MEMCHECK tools increase the runtime of kernels, it is possible for a CUDA kernel to exceed the timeout and therefore be terminated due to the TDR mechanism.

  For the purposes of debugging, the number of seconds before which the timeout is hit can be modified by setting the the timeout value in seconds in the DWORD registry key **`TdrDelay`** at **`HKEY_LOCAL_MACHINE\System\CurrentControlSet\Control\GraphicsDrivers`**

  More information about the registry keys to control the Timeout Detection and Recovery mechanism is available from MSDN at http://msdn.microsoft.com/en-us/library/windows/hardware/ff569918%28v=vs.85%29.aspx

8.2. Android Specific Behavior

- TMPDIR environment variable

  On Android, CUDA-MEMCHECK requires the TMPDIR environment variable to be set to a directory that is readable and writable by the current user.
- Host stack backtraces

  Host side function call stack backtraces are disabled on Android.
- Andoid GUI

  To ensure the GPU kernel is not terminated unexpectedly, the Android UI can be stopped by using the "stop" command in the **`adb`** shell.
- CUDA-MEMCHECK tool cannot be used with APK binaries.

8.3. QNX Specific Behavior

- TMPDIR environment variable

  On QNX, CUDA-MEMCHECK requires the TMPDIR environment variable to be set to a directory that is readable and writable by the current user.
- Host stack backtraces

  Host side function call stack backtraces are disabled on QNX.

Chapter 9. CUDA FORTRAN SUPPORT

This section describes support for CUDA Fortran.

9.1. CUDA Fortran Specific Behavior

- By default, error reports printed by CUDA-MEMCHECK contain 0-based C style values for thread index (threadIdx) and block index (blockIdx). For CUDA-MEMCHECK tools to use Fortran style 1-based offsets, use the **`--language fortran`** option.
- The CUDA Fortran compiler may insert extra padding in shared memory. Accesses hitting this extra padding may not be reported as an error.

Chapter 10. CUDA-MEMCHECK TOOL EXAMPLES

10.1. Example Use of Memcheck

This section presents a walk-through of running the memcheck tool from CUDA-MEMCHECK on a simple application called **`memcheck_demo`**.

Depending on the SM type of your GPU, your system output may vary.

memcheck_demo.cu source code

```
#include <stdio.h>

__device__ int x;

__global__ void unaligned_kernel(void) {
    *(int*) ((char*)&x + 1) = 42;
}

__device__ void out_of_bounds_function(void) {
    *(int*) 0x87654320 = 42;
}

__global__ void out_of_bounds_kernel(void) {
    out_of_bounds_function();
}

void run_unaligned(void) {
    printf("Running unaligned_kernel\n");
    unaligned_kernel<<<1,1>>>();
    printf("Ran unaligned_kernel: %s\n",
    cudaGetErrorString(cudaGetLastError()));
    printf("Sync: %s\n", cudaGetErrorString(cudaDeviceSynchronize()));
}

void run_out_of_bounds(void) {
    printf("Running out_of_bounds_kernel\n");
    out_of_bounds_kernel<<<1,1>>>();
    printf("Ran out_of_bounds_kernel: %s\n",
    cudaGetErrorString(cudaGetLastError()));
    printf("Sync: %s\n", cudaGetErrorString(cudaDeviceSynchronize()));
}

int main() {
    int *devMem;

    printf("Mallocing memory\n");
    cudaMalloc((void**)&devMem, 1024);

    run_unaligned();
    run_out_of_bounds();

    cudaDeviceReset();
    cudaFree(devMem);

    return 0;
}
```

This application is compiled for release builds as :

```
nvcc -o memcheck_demo memcheck_demo.cu
```

10.1.1. memcheck_demo Output

When a CUDA application causes access violations, the kernel launch may terminate with an error code of unspecified launch failure or a subsequent cudaDeviceSynchronize call which will fail with an error code of unspecified launch failure.

This sample application is causing two failures but there is no way to detect where these kernels are causing the access violations, as illustrated in the following output:

```
$ ./memcheck_demo
Mallocing memory
Running unaligned_kernel
Ran unaligned_kernel: no error
Sync: unspecified launch failure
Running out_of_bounds_kernel
Ran out_of_bounds_kernel: unspecified launch failure
Sync: unspecified launch failure
```

10.1.2. `memcheck_demo` Output with Memcheck (Release Build)

In this case, since the application is built in release mode, the CUDA-MEMCHECK output contains only the kernel names from the application causing the access violation. Though the kernel name and error type are detected, there is no line number information on the failing kernel. Also included in the output are the host and device backtraces for the call sites where the functions were launched. In addition, CUDA API errors are reported, such as the invalid **`cudaFree()`** call in the application.

```
$ cuda-memcheck ./memcheck_demo
========= CUDA-MEMCHECK
Mallocing memory
Running unaligned_kernel
Ran unaligned_kernel: no error
Sync: no error
Running out_of_bounds_kernel
Ran out_of_bounds_kernel: no error
Sync: no error
========= Invalid __global__ write of size 4
=========     at 0x00000028 in unaligned_kernel(void)
=========     by thread (0,0,0) in block (0,0,0)
=========     Address 0x400100001 is misaligned
=========     Saved host backtrace up to driver entry point at kernel launch
 time
=========     Host Frame:/usr/local/lib/libcuda.so (cuLaunchKernel + 0x3ae)
 [0xddbee]
=========     Host Frame:/usr/local/lib/libcudart.so.5.0 [0xcd27]
=========     Host Frame:/usr/local/lib/libcudart.so.5.0 (cudaLaunch + 0x1bb)
 [0x3778b]
=========     Host Frame:memcheck_demo [0xdfc]
=========     Host Frame:memcheck_demo [0xc76]
=========     Host Frame:memcheck_demo [0xc81]
=========     Host Frame:memcheck_demo [0xb03]
=========     Host Frame:memcheck_demo [0xc27]
=========     Host Frame:/lib64/libc.so.6 (__libc_start_main + 0xfd) [0x1eb1d]
=========     Host Frame:memcheck_demo [0x9b9]
=========
========= Invalid __global__ write of size 4
=========     at 0x00000010 in out_of_bounds_kernel(void)
=========     by thread (0,0,0) in block (0,0,0)
=========     Address 0xffffffff87654320 is out of bounds
=========     Saved host backtrace up to driver entry point at kernel launch
 time
=========     Host Frame:/usr/local/lib/libcuda.so (cuLaunchKernel + 0x3ae)
 [0xddbee]
=========     Host Frame:/usr/local/lib/libcudart.so.5.0 [0xcd27]
=========     Host Frame:/usr/local/lib/libcudart.so.5.0 (cudaLaunch + 0x1bb)
 [0x3778b]
=========     Host Frame:memcheck_demo [0xdfc]
=========     Host Frame:memcheck_demo [0xca0]
=========     Host Frame:memcheck_demo [0xcab]
=========     Host Frame:memcheck_demo [0xbbc]
=========     Host Frame:memcheck_demo [0xc2c]
=========     Host Frame:/lib64/libc.so.6 (__libc_start_main + 0xfd) [0x1eb1d]
=========     Host Frame:memcheck_demo [0x9b9]
=========
========= Program hit error 17 on CUDA API call to cudaFree
=========     Saved host backtrace up to driver entry point at error
=========     Host Frame:/usr/local/lib/libcuda.so [0x28f850]
=========     Host Frame:/usr/local/lib/libcudart.so.5.0 (cudaFree + 0x20d)
 [0x364ed]
=========     Host Frame:memcheck_demo [0xc3d]
=========     Host Frame:/lib64/libc.so.6 (__libc_start_main + 0xfd) [0x1eb1d]
=========     Host Frame:memcheck_demo [0x9b9]
=========
========= ERROR SUMMARY: 3 errors
```

10.1.3. `memcheck_demo` Output with Memcheck (Debug Build)

The application is now built with device side debug information and function symbols as :

```
nvcc -G -Xcompiler -rdynamic -o memcheck_demo memcheck_demo.cu
```

Now run this application with CUDA-MEMCHECK and check the output. By default, the application will run so that the kernel is terminated on memory access errors but other work in the CUDA context can still proceed.

In the output below the first kernel no longer reports an unspecified launch failure as its execution has been terminated early after CUDA-MEMCHECK detected the error. The application continued to run the second kernel. The error detected in the second kernel causes it to terminate early. Finally, the application calls **`cudaDeviceReset()`**, which destroys the CUDA context and then attempts to call **`cudaFree()`**. This call returns an API error that is caught and displayed by memcheck.

```
$ cuda-memcheck ./memcheck_demo
========= CUDA-MEMCHECK
Mallocing memory
Running unaligned_kernel
Ran unaligned_kernel: no error
Sync: no error
Running out_of_bounds_kernel
Ran out_of_bounds_kernel: no error
Sync: no error
========= Invalid __global__ write of size 4
=========     at 0x00000028 in memcheck_demo.cu:6:unaligned_kernel(void)
=========     by thread (0,0,0) in block (0,0,0)
=========     Address 0x400100001 is misaligned
=========     Saved host backtrace up to driver entry point at kernel launch
 time
=========     Host Frame:/usr/local/lib/libcuda.so (cuLaunchKernel + 0x3ae)
 [0xddbee]
=========     Host Frame:/usr/local/lib/libcudart.so.5.0 [0xcd27]
=========     Host Frame:/usr/local/lib/libcudart.so.5.0 (cudaLaunch + 0x1bb)
 [0x3778b]
=========     Host Frame:memcheck_demo (_Z10cudaLaunchIcE9cudaErrorPT_ + 0x18)
 [0x11a4]
=========     Host Frame:memcheck_demo (_Z35__device_stub__Z16unaligned_kernelvv
 + 0x1d) [0x101d]
=========     Host Frame:memcheck_demo (_Z16unaligned_kernelv + 0x9) [0x1028]
=========     Host Frame:memcheck_demo (_Z13run_unalignedv + 0x76) [0xeaa]
=========     Host Frame:memcheck_demo (main + 0x28) [0xfce]
=========     Host Frame:/lib64/libc.so.6 (__libc_start_main + 0xfd) [0x1eb1d]
=========     Host Frame:memcheck_demo [0xd79]
=========
========= Invalid __global__ write of size 4
=========     at 0x00000028 in memcheck_demo.cu:10:out_of_bounds_function(void)
=========     by thread (0,0,0) in block (0,0,0)
=========     Address 0x87654320 is out of bounds
=========     Device Frame:memcheck_demo.cu:15:out_of_bounds_kernel(void)
 (out_of_bounds_kernel(void) : 0x10)
=========     Saved host backtrace up to driver entry point at kernel launch
 time
=========     Host Frame:/usr/local/lib/libcuda.so (cuLaunchKernel + 0x3ae)
 [0xddbee]
=========     Host Frame:/usr/local/lib/libcudart.so.5.0 [0xcd27]
=========     Host Frame:/usr/local/lib/libcudart.so.5.0 (cudaLaunch + 0x1bb)
 [0x3778b]
=========     Host Frame:memcheck_demo (_Z10cudaLaunchIcE9cudaErrorPT_ + 0x18)
 [0x11a4]
=========     Host Frame:memcheck_demo
 (_Z39__device_stub__Z20out_of_bounds_kernelvv + 0x1d) [0x1047]
=========     Host Frame:memcheck_demo (_Z20out_of_bounds_kernelv + 0x9)
 [0x1052]
=========     Host Frame:memcheck_demo (_Z17run_out_of_boundsv + 0x76) [0xf63]
=========     Host Frame:memcheck_demo (main + 0x2d) [0xfd3]
=========     Host Frame:/lib64/libc.so.6 (__libc_start_main + 0xfd) [0x1eb1d]
=========     Host Frame:memcheck_demo [0xd79]
=========
========= Program hit error 17 on CUDA API call to cudaFree
=========     Saved host backtrace up to driver entry point at error
=========     Host Frame:/usr/local/lib/libcuda.so [0x28f850]
=========     Host Frame:/usr/local/lib/libcudart.so.5.0 (cudaFree + 0x20d)
 [0x364ed]
=========     Host Frame:memcheck_demo (main + 0x3e) [0xfe4]
=========     Host Frame:/lib64/libc.so.6 (__libc_start_main + 0xfd) [0x1eb1d]
=========     Host Frame:memcheck_demo [0xd79]
=========
========= ERROR SUMMARY: 3 errors
```

10.1.4. Leak Checking in CUDA-MEMCHECK

To print information about the allocations that have not been freed at the time the CUDA context is destroyed, we can specify the **`--leak-check full`** option to CUDA-MEMCHECK.

When running the program with the leak check option, the user is presented with a list of allocations that were not destroyed, along with the size of the allocation and the address on the device of the allocation. For allocations made on the host, each leak report will also print a backtrace corresponding to the saved host stack at the time the allocation was first made. Also presented is a summary of the total number of bytes leaked and the corresponding number of allocations.

In this example, the program created an allocation using **`cudaMalloc()`** and has not called **`cudaFree()`** to release it, leaking memory. Notice that CUDA-MEMCHECK still prints errors it encountered while running the application.

```
$ cuda-memcheck --leak-check full memcheck_demo
========= CUDA-MEMCHECK
Mallocing memory
Running unaligned_kernel
Ran unaligned_kernel: no error
Sync: no error
Running out_of_bounds_kernel
Ran out_of_bounds_kernel: no error
Sync: no error
========= Invalid __global__ write of size 4
=========     at 0x00000060 in memcheck_demo.cu:6:unaligned_kernel(void)
=========     by thread (0,0,0) in block (0,0,0)
=========     Address 0x400100001 is misaligned
=========     Saved host backtrace up to driver entry point at kernel launch
 time
=========     Host Frame:/usr/local/lib/libcuda.so (cuLaunchKernel + 0x3ae)
 [0xddbee]
=========     Host Frame:/usr/local/lib/libcudart.so.5.0 [0xcd27]
=========     Host Frame:/usr/local/lib/libcudart.so.5.0 (cudaLaunch + 0x1bb)
 [0x3778b]
=========     Host Frame:memcheck_demo (_Z10cudaLaunchIcE9cudaErrorPT_ + 0x18)
 [0x122c]
=========     Host Frame:memcheck_demo (_Z35__device_stub__Z16unaligned_kernelvv
 + 0x1d) [0x10a6]
=========     Host Frame:memcheck_demo (_Z16unaligned_kernelv + 0x9) [0x10b1]
=========     Host Frame:memcheck_demo (_Z13run_unalignedv + 0x76) [0xf33]
=========     Host Frame:memcheck_demo (main + 0x28) [0x1057]
=========     Host Frame:/lib64/libc.so.6 (__libc_start_main + 0xfd) [0x1eb1d]
=========     Host Frame:memcheck_demo [0xde9]
=========
========= Invalid __global__ write of size 4
=========     at 0x00000028 in memcheck_demo.cu:10:out_of_bounds_function(void)
=========     by thread (0,0,0) in block (0,0,0)
=========     Address 0x87654320 is out of bounds
=========     Device Frame:memcheck_demo.cu:15:out_of_bounds_kernel(void)
 (out_of_bounds_kernel(void) : 0x10)
=========     Saved host backtrace up to driver entry point at kernel launch
 time
=========     Host Frame:/usr/local/lib/libcuda.so (cuLaunchKernel + 0x3ae)
 [0xddbee]
=========     Host Frame:/usr/local/lib/libcudart.so.5.0 [0xcd27]
=========     Host Frame:/usr/local/lib/libcudart.so.5.0 (cudaLaunch + 0x1bb)
 [0x3778b]
=========     Host Frame:memcheck_demo (_Z10cudaLaunchIcE9cudaErrorPT_ + 0x18)
 [0x122c]
=========     Host Frame:memcheck_demo
 (_Z39__device_stub__Z20out_of_bounds_kernelvv + 0x1d) [0x10d0]
=========     Host Frame:memcheck_demo (_Z20out_of_bounds_kernelv + 0x9)
 [0x10db]
=========     Host Frame:memcheck_demo (_Z17run_out_of_boundsv + 0x76) [0xfec]
=========     Host Frame:memcheck_demo (main + 0x2d) [0x105c]
=========     Host Frame:/lib64/libc.so.6 (__libc_start_main + 0xfd) [0x1eb1d]
=========     Host Frame:memcheck_demo [0xde9]
=========
========= Leaked 1024 bytes at 0x400200000
=========     Saved host backtrace up to driver entry point at cudaMalloc time
=========     Host Frame:/usr/local/lib/libcuda.so (cuMemAlloc_v2 + 0x236)
 [0xe9746]
=========     Host Frame:/usr/local/lib/libcudart.so.5.0 [0x26dd7]
=========     Host Frame:/usr/local/lib/libcudart.so.5.0 [0xb37b]
=========     Host Frame:/usr/local/lib/libcudart.so.5.0 (cudaMalloc + 0x17a)
 [0x36e6a]
=========     Host Frame:memcheck_demo (main + 0x23) [0x1052]
=========     Host Frame:/lib64/libc.so.6 (__libc_start_main + 0xfd) [0x1eb1d]
=========     Host Frame:memcheck_demo [0xde9]
=========
========= Program hit error 17 on CUDA API call to cudaFree
=========     Saved host backtrace up to driver entry point at error
=========     Host Frame:/usr/local/lib/libcuda.so [0x28f850]
=========     Host Frame:/usr/local/lib/libcudart.so.5.0 (cudaFree + 0x20d)
 [0x364ed]
=========     Host Frame:memcheck_demo (main + 0x3e) [0x106d]
=========     Host Frame:/lib64/libc.so.6 (__libc_start_main + 0xfd) [0x1eb1d]
=========     Host Frame:memcheck_demo [0xde9]
=========
```

10.2. Integrated CUDA-MEMCHECK Example

This example shows how to enable CUDA-MEMCHECK from within CUDA-GDB and how to detect errors within the debugger so you can access the line number information and check the state of the variables

In this example the unaligned kernel has a misaligned memory access in block 1 lane 1, which gets trapped as an illegal lane address at line 6 from within CUDA-GDB. Note that CUDA-GDB displays the address and that caused the bad access.

```
(cuda-gdb) set cuda memcheck on
(cuda-gdb) run
Starting program: memcheck_demo
[Thread debugging using libthread_db enabled]
Mallocing memory
[New Thread 0x7ffff6fe1710 (LWP 7783)]
[Context Create of context 0x6218a0 on Device 0]
[Launch of CUDA Kernel 0 (memset32_post<<<(1,1,1),(64,1,1)>>>) on Device 0]
Running unaligned_kernel
[Launch of CUDA Kernel 1 (unaligned_kernel<<<(1,1,1),(1,1,1)>>>) on Device 0]
Memcheck detected an illegal access to address (@global)0x400100001

Program received signal CUDA_EXCEPTION_1, Lane Illegal Address.
[Switching focus to CUDA kernel 1, grid 2, block (0,0,0), thread (0,0,0), device
 0, sm 0, warp 0, lane 0]
0x0000000000078b8b0 in unaligned_kernel<<<(1,1,1),(1,1,1)>>> () at
 memcheck_demo.cu:6
6           *(int*) ((char*)&x + 1) = 42;
(cuda-gdb) print &x
$1 = (@global int *) 0x400100000
(cuda-gdb) continue
Continuing.
[Termination of CUDA Kernel 1 (unaligned_kernel<<<(1,1,1),(1,1,1)>>>) on Device
 0]
[Termination of CUDA Kernel 0 (memset32_post<<<(1,1,1),(64,1,1)>>>) on Device 0]

Program terminated with signal CUDA_EXCEPTION_1, Lane Illegal Address.
The program no longer exists.
(cuda-gdb)
```

10.3. Example Use of Racecheck

This section presents two example usages of the racecheck tool from CUDA-MEMCHECK. The first example uses an application called **`block_error`**, which has shared memory hazards on the block level. The second example uses an application called **`warp_error`**, which has shared memory hazards on the warp level.

Depending on the SM type of your GPU, your system output may vary.

10.3.1. Block-level Hazards

block_error.cu source code

```
#define THREADS 128

__shared__ int smem[THREADS];

__global__
void sumKernel(int *data_in, int *sum_out)
{
    int tx = threadIdx.x;
    smem[tx] = data_in[tx] + tx;

    if (tx == 0) {
        *sum_out = 0;
        for (int i = 0; i < THREADS; ++i)
            *sum_out += smem[i];
    }
}

int main(int argc, char **argv)
{
    int *data_in = NULL;
    int *sum_out = NULL;

    cudaMalloc((void**)&data_in, sizeof(int) * THREADS);
    cudaMalloc((void**)&sum_out, sizeof(int));
    cudaMemset(data_in, 0, sizeof(int) * THREADS);

    sumKernel<<<1, THREADS>>>(data_in, sum_out);
    cudaDeviceSynchronize();

    cudaFree(data_in);
    cudaFree(sum_out);
    return 0;
}
```

Each kernel thread write some element in shared memory. Afterwards, thread 0 computes the sum of all elements in shared memory and stores the result in global memory variable **sum_out**.

Running this application under the racecheck tool with the **--racecheck-report analysis** option, the following error is reported:

```
========= CUDA-MEMCHECK
========= ERROR: Race reported between Write access at 0x00000068 in
 block_error.cu:9:sumKernel(int*, int*)
=========     and Read access at 0x000000e8 in block_error.cu:14:sumKernel(int*,
 int*) [128 hazards]
=========     and Read access at 0x00000130 in block_error.cu:14:sumKernel(int*,
 int*) [128 hazards]
=========     and Read access at 0x000000d0 in block_error.cu:14:sumKernel(int*,
 int*) [124 hazards]
=========     and Read access at 0x00000188 in block_error.cu:14:sumKernel(int*,
 int*) [128 hazards]
```

Racecheck reports races between thread 0 reading all shared memory elements in line 14 and each individual thread writing its shared memory entry in line 9. Accesses to shared memory between multiple threads, where at least one access is a write, can potentially

race with each other. Since the races are between threads of different warps, the block-level synchronization barrier **`__syncthreads()`** is required in line 10.

Note that a total of 508 hazards are reported: the kernel uses a single block of 128 threads. The data size written or read, respectively, by each thread is four bytes (one **`int`**) and hazards are reported at the byte level. The writes by all threads race with the reads by thread 0, except for the four writes by thread 0 itself.

10.3.2. Warp-level Hazards

warp_error.cu source code

```
#define WARPS 2
#define WARP_SIZE 32
#define THREADS (WARPS * WARP_SIZE)

__shared__ int smem_first[THREADS];
__shared__ int smem_second[WARPS];

__global__
void sumKernel(int *data_in, int *sum_out)
{
    int tx = threadIdx.x;
    smem_first[tx] = data_in[tx] + tx;

    if (tx % WARP_SIZE == 0) {
        int wx = tx / WARP_SIZE;

        smem_second[wx] = 0;
        for (int i = 0; i < WARP_SIZE; ++i)
            smem_second[wx] += smem_first[wx * WARP_SIZE + i];
    }

    __syncthreads();

    if (tx == 0) {
        *sum_out = 0;
        for (int i = 0; i < WARPS; ++i)
            *sum_out += smem_second[i];
    }
}

int main(int argc, char **argv)
{
    int *data_in = NULL;
    int *sum_out = NULL;

    cudaMalloc((void**)&data_in, sizeof(int) * THREADS);
    cudaMalloc((void**)&sum_out, sizeof(int));
    cudaMemset(data_in, 0, sizeof(int) * THREADS);

    sumKernel<<<1, THREADS>>>(data_in, sum_out);
    cudaDeviceSynchronize();

    cudaFree(data_in);
    cudaFree(sum_out);
    return 0;
}
```

The kernel computes the some of all individual elements in shared memory two stages. First, each thread computes its local shared memory value in **`smem_first`**. Second, a

single thread of each warp is chosen with **`if (tx % WARP_SIZE == 0)`** to sum all elements written by its warp, indexed **`wx`**, and store the result in **`smem_second`**. Finally, thread 0 of the kernel computes the sum of elements in **`smem_second`** and writes the value into global memory.

Running this application under the racecheck tool with the **`--racecheck-report hazard`** option, multiple hazards with WARNING severity are reported:

```
========= WARN:(Warp Level Programming) Potential RAW hazard detected at
 __shared__ 0x7 in block (0, 0, 0) :
=========     Write Thread (1, 0, 0) at 0x00000070 in
 warp_error.cu:12:sumKernel(int*, int*)
=========     Read Thread (0, 0, 0) at 0x000000b0 in
 warp_error.cu:19:sumKernel(int*, int*)
=========     Current Value : 0
```

The avoid the errors demonstrated in the Block-level Hazards example, the kernel uses the block-level barrier **`__syncthreads()`** in line 22. However, racecheck still reports read-after-write (RAW) hazards between threads within the same warp, with severity WARNING. On architectures prior to SM 7.0 (Volta), programmers commonly relied on the assumption that threads within a warp execute code in lock-step (warp-level programming). Starting with CUDA 9.0, programmers can use the new **`__syncwarp()`** warp-wide barrier (instead of only **`__syncthreads()`** beforehand) to avoid such hazards. This barrier should be inserted at line 13.

10.4. Example Use of Initcheck

This section presents the usage of the initcheck tool from CUDA-MEMCHECK. The example uses an application called **`memset_error`**.

10.4.1. Memset Error

memset_error.cu source code

```
#define THREADS 128
#define BLOCKS 2

__global__
void vectorAdd(int *v)
{
    int tx = threadIdx.x + blockDim.x * blockIdx.x;

    v[tx] += tx;
}

int main(int argc, char **argv)
{
    int *d_vec = NULL;

    cudaMalloc((void**)&d_vec, sizeof(int) * BLOCKS * THREADS);
    cudaMemset(d_vec, 0, BLOCKS * THREADS);

    vectorAdd<<<BLOCKS, THREADS>>>(d_vec);
    cudaDeviceSynchronize();

    cudaFree(d_vec);
    return 0;
}
```

The example implements a very simple vector addition, where the thread index is added to each vector element. The vector contains **`BLOCKS * THREADS`** elements of type **`int`**. The vector is allocated on the device and then initialized to 0 using **`cudaMemset`** before the kernel is launched.

Running this application under the initcheck tool reports multiple errors like the following:

```
========= Uninitialized __global__ memory read of size 4
=========     at 0x00000070 in /home/user/memset_error.cu:9:vectorAdd(int*)
=========     by thread (65,0,0) in block (0,0,0)
=========     Address 0x10208e00104
=========
```

The problem is that the call to **`cudaMemset`** expects the size of the to-be set memory in bytes. However, the size is given in elements, as a factor of **`sizeof(int)`** is missing while computing the parameter. As a result, 3/4 of the memory will have undefined values during the vector addition.

10.5. Example Use of Synccheck

This section presents two example usages of the synccheck tool from CUDA-MEMCHECK. The first example uses an application called **`divergent_threads`**. The second example uses an application called **`illegal_syncwarp`**.

Depending on the SM type of your GPU, your system output may vary.

10.5.1. Divergent Threads

divergent_threads.cu source code

```
#define THREADS 64
#define DATA_BLOCKS 16

__shared__ int smem[THREADS];

__global__ void
myKernel(int *data_in, int *sum_out, const int size)
{
    int tx = threadIdx.x;

    smem[tx] = 0;

    __syncthreads();

    for (int b = 0; b < DATA_BLOCKS; ++b) {
        const int offset = THREADS * b + tx;
        if (offset < size) {
            smem[tx] += data_in[offset];
            __syncthreads();
        }
    }

    if (tx == 0) {
        *sum_out = 0;
        for (int i = 0; i < THREADS; ++i)
            *sum_out += smem[i];
    }
}

int main(int argc, char *argv[])
{
    const int SIZE = (THREADS * DATA_BLOCKS) - 16;
    int *data_in = NULL;
    int *sum_out = NULL;

    cudaMalloc((void**)&data_in, SIZE * sizeof(int));
    cudaMalloc((void**)&sum_out, sizeof(int));

    myKernel<<<1,THREADS>>>(data_in, sum_out, SIZE);

    cudaDeviceSynchronize();
    cudaFree(data_in);
    cudaFree(sum_out);

    return 0;
}
```

In this example, we launch a kernel with a single block of 64 threads. The kernels loops over **DATA_BLOCKS** blocks of input data **data_in**. In each iteration, **THREADS** elements are added concurrently in shared memory. Finally, a single thread 0 computes the sum of all values in shared memory and writes it to **sum_out**.

Running this application under the synccheck tool, 16 errors like the following are reported:

```
========= Barrier error detected. Divergent thread(s) in block
=========     at 0x000006c8 in divergent_threads.cu:20:myKernel(int*, int*, int)
=========     by thread (32,0,0) in block (0,0,0)
```

The issue is with the **__syncthreads()** in line 20 when reading the last data block into shared memory. Note that the last data block only has 48 elements (compared to 64 elements for all other blocks). As a result, not all threads of the second warp execute this statement in convergence as required.

10.5.2. Illegal Syncwarp

illegal_syncwarp.cu source code

```
#define THREADS 32

__shared__ int smem[THREADS];

__global__ void
myKernel(int *sum_out)
{
    int tx = threadIdx.x;

    unsigned int mask = __ballot_sync(0xffffffff, tx < (THREADS / 2));

    if (tx <= (THREADS / 2)) {
        smem[tx] = tx;

        __syncwarp(mask);

        *sum_out = 0;
        for (int i = 0; i < (THREADS / 2); ++i)
            *sum_out += smem[i];
    }
}

int main(int argc, char *argv[])
{
    int *sum_out = NULL;

    cudaMalloc((void**)&sum_out, sizeof(int));

    myKernel<<<1,THREADS>>>(sum_out);

    cudaDeviceSynchronize();
    cudaFree(sum_out);

    return 0;
}
```

This example only applies to devices of compute capability 7.0 (Volta) and above. The kernel is launched with a single warp (32 threads), but only thread 0-15 are part of the computation. Each of these threads initializes one shared memory element with its thread index. After the assignment, **__syncwarp()** is used to ensure that the warp is converged and all writes are visible to other threads. The mask passed to **__syncwarp()** is computed using **__ballot_sync()**, which enables the bits for the first 16 threads in **mask**. Finally, the first thread (index 0) computes the sum over all initialized shared memory elements and writes it to global memory.

Building the application with **-G** to enable debug information and running it under the synccheck tool on SM 7.0 and above, multiple errors like the following are reported:

```
========= Barrier error detected. Invalid arguments
=========     at 0x00000040 in __cuda_sm70_warpsync
=========     by thread (0,0,0) in block (0,0,0)
=========     Device Frame:__cuda_sm70_warpsync (__cuda_sm70_warpsync : 0x40)
=========     Device Frame:/usr/local/cuda/include/
sm_30_intrinsics.hpp:112:__syncwarp(unsigned int) (__syncwarp(unsigned int) :
 0x110)
=========     Device Frame:/home/user/illegal_synwarp.cu:15:myKernel(int*)
 (myKernel(int*) : 0x460
```

The issue is with the **__syncwarp(mask)** in line 15. All threads for which **tx < (THREADS / 2)** holds true are enabled in the mask, which are threads 0-15. However, the if condition evaluates true for threads 0-16. As a result, thread 16 executes the **__syncwarp(mask)** but does not declare itself in the mask parameter as required.

Appendix A.
MEMORY ACCESS ERROR REPORTING

The memcheck tool will report memory access errors when run standalone or in integrated mode with CUDA-GDB. The table below describes the types of accesses that are checked and the SM version where such checks happen

Table 9 Memcheck memory access error detection support

Error Type	SM 3.5	SM 5.x	SM 6.x	SM 7.x	SM 8.0
Global	Yes	Yes	Yes	Yes	Yes
Shared	Yes	Yes	Yes	Yes	Yes
Local	Yes	Yes	Yes	Yes	Yes
Global Atomic	Yes	Yes	Yes	Yes	Yes
Load through texture	Yes	N/A	N/A	N/A	N/A
System-scoped Atomics	N/A	N/A	Yes	Yes	Yes

Appendix B. HARDWARE EXCEPTION REPORTING

The CUDA-MEMCHECK tool will report hardware exceptions when run as a standalone or as part of CUDA-GDB. The table below enumerates the supported exceptions, their precision and scope, as well as a brief description of their cause. For more detailed information, see the documentation for CUDA-GDB.

Table 10 CUDA Exception Codes

Exception code	Precision of the Error	Scope of the Error	Description
CUDA_EXCEPTION_1 : "Lane Illegal Address"	Precise	Per lane/thread error	This occurs when a thread accesses an illegal (out of bounds) global address.
CUDA_EXCEPTION_2 : "Lane User StackOverflow"	Precise	Per lane/thread error	This occurs when a thread exceeds its stack memory limit.
CUDA_EXCEPTION_3: "Device Hardware Stack Overflow"	Not precise	Global error on the GPU	This occurs when the application triggers a global hardware stack overflow. The main cause of this error is large amounts of divergence in the presence of function calls.
CUDA_EXCEPTION_4: "Warp Illegal Instruction"	Not precise	Warp error	This occurs when any thread within a warp has executed an illegal instruction.
CUDA_EXCEPTION_5: "Warp Out-of-range Address"	Not precise	Warp error	This occurs when any thread within a warp accesses an address that is outside the valid range of local or shared memory regions.
CUDA_EXCEPTION_6: "Warp Misaligned Address"	Not precise	Warp error	This occurs when any thread within a warp accesses an address in the local or shared memory segments that is not correctly aligned.
CUDA_EXCEPTION_7: "Warp Invalid Address Space"	Not precise	Warp error	This occurs when any thread within a warp executes an instruction

Exception code	Precision of the Error	Scope of the Error	Description
			that accesses a memory space not permitted for that instruction.
CUDA_EXCEPTION_8: "Warp Invalid PC"	Not precise	Warp error	This occurs when any thread within a warp advances its PC beyond the 40-bit address space.
CUDA_EXCEPTION_9: "Warp Hardware Stack Overflow"	Not precise	Warp error	This occurs when any thread in a warp triggers a hardware stack overflow. This should be a rare occurrence.
CUDA_EXCEPTION_10: "Device Illegal Address"	Not precise	Global error	This occurs when a thread accesses an illegal (out of bounds) global address.
CUDA_EXCEPTION_11: "Lane Misaligned Address"	Precise	Per lane/thread error	This occurs when a thread accesses a global address that is not correctly aligned.
CUDA_EXCEPTION_12: "Warp Assert"	Precise	Per warp	This occurs when any thread in the warp hits a device side assertion.
CUDA_EXCEPTION_13: "Lane Syscall Error"	Precise	Per lane	This occurs when a particular thread causes an syscall error, such as calling `free()` in a kernel on an already free'd pointer.
"Unknown Exception"	Not precise	Global error	The precise cause of the exception is unknown. Potentially, this may be due to Device Hardware Stack overflows or a kernel generating an exception very close to its termination.

Appendix C. RELEASE NOTES

C.1. New Features in 11.0

- Support for SM 8.0

C.2. New Features in 10.2

- Support for CUDA graphs

C.3. New Features in 10.1

- Support for atomic instructions on 16-bit __half floating point type

C.4. New Features in 10.0

- Support for SM 7.5

C.5. New Features in 9.1

- On Volta, the synccheck tool will report an error if a deprecated variant of __shfl() is used in divergent code.
- Added a command line option to report deprecated instructions even when they are used in safe execution paths. For more information, see Command Line Options.

C.6. New Features in 9.0

- Support for host API functions with pitch parameter. For more information see Initcheck Tool.

- Initial support for the Cooperative Groups programming model.
- Support for shared memory atomic instructions. For more information see Memcheck Tool.
- Support for detecting invalid accesses to global memory on Pascal and later architectures that extend beyond the end of an allocation. For more information see Memcheck Tool.
- Support for limiting the numbers of errors printed by cuda-memcheck. For more information see Command Line Options.
- Racecheck analysis reports are assigned a severity level. For more information see Understanding Racecheck Analysis Reports.
- Default print level changed from INFO to WARN. For more information see Command Line Options.
- Support for SM 7.0 and 7.2

C.7. New Features in 8.0

- Support for non-migratable system-scoped atomics checking on SM 6.x. For more information see Memcheck Tool.
- Support for reporting fatal CPU-side faults when Unified Memory is enabled. For more information see Memcheck Tool.
- Support for correctly determining the expected set of threads at a barrier in the presence of exited threads in Synccheck Tool.
- Support for SM 6.x

C.8. New Features in 7.0

- Support for uninitialized global memory access checking. For more information see Initcheck Tool.
- Support for divergent block synchronization checking. For more information see Synccheck Tool.
- Support for SM 5.2

C.9. New Features in 6.5

- More information printed for API errors
- Support for escape sequences in file name to **`--log-file`** and **`--save`**.
- Support for controlling which kernels are checked using **`--filter`**. For more information see Specifying Filters.

C.10. New Features in 6.0

- Support for Unified Memory
- Support for CUDA Multi Process Service (MPS)

- Support for additional error detection with cudaMemcpy and cudaMemset

C.11. New Features in 5.5

- Analysis mode in racecheck tool. For more information, see Racecheck Tool
- Support for racecheck on SM 3.5 GPUs.

C.12. New Features in 5.0

- Reporting of data access hazards in shared memory accesses. This is supported on Fermi SM 2.x and Kepler SM 3.0 GPUs. This functionality is not supported on Windows XP. For more information, see Racecheck Tool.
- Support for SM 3.0 and SM 3.5 GPUs. For more information, see Supported Devices.
- Support for dynamic parallelism. All memory access error detection is supported for applications using dynamic parallelism. For more information, see Dynamic Parallelism.
- Precise error detection for local loads/stores, shared loads/stores, global atomics/reductions. On SM 3.5, added precise memory access error detection for noncoherent global loads through the texture unit. For more information, see Memory Access Error Reporting.
- Error detection in device side malloc()/free(), such as double free() or invalid free() on the GPU. For more information, see Device Side Allocation Checking.
- Leak checking for allocations on the device heap. For more information, see Leak Checking.
- Display of a saved stack backtrace on the host and captured backtrace on the device for different errors. For more information, see Stack Backtraces.
- Reporting of CUDA API errors in the user's application. For more information, see CUDA API Error Checking.
- Added display of mangled, demangled, and full prototype of the kernel. For more information, see Name Demangling.
- Increased functionality in integrated mode with CUDA-GDB. Added reporting of the address and address space being accessed that caused a precise exception. Added checking of device side malloc() and free() when in integrated mode. For more information, see Integrated Mode.
- Support for applications compiled separately that use the device side linker.
- Support for applications compiled with the -lineinfo flag.
- New style of command line options. For more information, see Command Line Options.
- Changed default behavior. CUDA-MEMCHECK will display backtraces by default and will report API errors by default. For more information, see Command Line Options.

Appendix D. KNOWN ISSUES

The following are known issues with the current release.

- Applications run much slower under CUDA-MEMCHECK tools. This may cause some kernel launches to fail with a launch timeout error when running with CUDA-MEMCHECK enabled.
- When running CUDA-MEMCHECK tools in integrated mode with CUDA-GDB, only the *memcheck* tool is enabled. Also, the following features are disabled:
  - Nonblocking launches
  - Leak checking
  - API error checking
- CUDA-MEMCHECK tools do not support CUDA/Direct3D interop.
- CUDA-MEMCHECK tools do not fully support CUDA concurrent streams. Applications relying on kernels running concurrently in different streams may hang.
- The memcheck tool does not support CUDA API error checking for API calls made on the GPU using dynamic parallelism.
- The racecheck, synccheck and initcheck tools do not support CUDA dynamic parallelism.
- CUDA-MEMCHECK tools do not support OptiX.
- In cases where a CUDA application spawns child processes that in turn use CUDA, CUDA-MEMCHECK tools may not report errors from the child processes.
- Tools in the CUDA-MEMCHECK suite cannot interoperate with the following applications:
  - Nvidia legacy command line profiler (CUDA_PROFILE/COMPUTE_PROFILE)
  - nvprof
  - Nvidia Visual Profiler
  - Nvidia Nsight Visual Studio Edition

  If such tools are detected, CUDA-MEMCHECK will terminate with an internal error that initialization failed. Please make sure that the tools listed above are not running. In case the message persists, make sure the following environment variables are not set :
  - COMPUTE_PROFILE

  - CUDA_PROFILE
  - CUDA_INJECTION32_DLL
  - CUDA_INJECTION64_DLL
  - CUDA_INJECTION32_PATH
  - CUDA_INJECTION64_PATH
- On SM 7.0 and above, the racecheck tool does not fully support warp synchronization instructions with a partial thread mask. If such an instruction is encountered, it is handled as if the mask would have been full (i.e. 0xffffffff). As a result, checking can be too conservative at times and some potential intra-warp hazards will not be detected.
- The memcheck tool terminates threads which are caught performing double free. On SM 7.0 and above, this might also cause other threads in the same block to exit when a double free is detected.
- On Windows platforms, device call stack backtraces only report the current frame for GPUs with SM versions 7.0 and above in WDDM mode. All frames are shown if the device is put in TCC mode.
- CUDA-MEMCHECK tools do not support CUDA AsyncCopy and CUDA AWBarrier features.
- CUDA-MEMCHECK tools do not fully support CUDA graphs, which can result in kernel launches failures. For such cases, the compute-sanitizer tool should be used as a replacement for CUDA-MEMCHECK.
- CUDA-MEMCHECK tools may not report errors or report false positives when the application is using CUBLAS. For such cases, the compute-sanitizer tool should be used as a replacement for CUDA-MEMCHECK.
- CUDA-MEMCHECK tools are not supported when Windows Hardware-accelerated GPU scheduling is enabled. For such cases the compute-sanitizer tool should be used as a replacement for CUDA-MEMCHECK.

Notice

Trademarks

Copyright