



SHARED MEMORY BANK CONFLICTS SAMPLE

v2023.1.0 | March 2023



TABLE OF CONTENTS

Chapter 1. Introduction..... 1

Chapter 2. Application..... 2

Chapter 3. Configuration..... 3

Chapter 4. Initial version of the kernel.....4

Chapter 5. Updated version of the kernel..... 10

Chapter 6. Resources..... 13

Chapter 1.

INTRODUCTION

This sample profiles a CUDA kernel which transposes an $N \times N$ square matrix of float elements in global memory using the Nsight Compute profiler. To avoid uncoalesced global memory accesses this kernel reads the data into shared memory. The profiler is used to analyze and identify the shared memory bank conflicts which result in inefficient shared memory accesses.

Shared memory accesses on a GPU

Shared memory is located on-chip, so it has much higher bandwidth and much lower latency than either local or global memory. Shared memory can be shared across a compute Cooperative Thread Array (CTA). In CUDA, CTAs are referred to as Thread Blocks. Compute CTAs attempting to share data across threads via shared memory must use synchronization operations (such as `__syncthreads()`) between stores and loads to ensure data written by any one thread is visible to other threads in the CTA.

Shared memory has 32 banks that are organized such that successive 32-bit words map to successive banks that can be accessed simultaneously. Any 32-bit memory read or write request made of 32 addresses that fall in 32 distinct memory banks can therefore be serviced simultaneously, yielding an overall bandwidth that is 32 times as high as the bandwidth of a single request. However, if two addresses of a memory request fall in the same memory bank, there is a bank conflict and the access has to be serialized. The exception to this rule is when all threads read the same shared memory address, which results in a broadcast where the data at that address is sent to all threads in one transaction.

To get maximum performance, it is therefore important to understand how memory addresses map to memory banks in order to schedule the memory requests so as to minimize bank conflicts.

Chapter 2.

APPLICATION

The sample CUDA application transposes a matrix of floats. The input and output matrices are at separate memory locations. For simplicity it only handles square matrices whose dimensions are integral multiples of 32, the tile size.

The sharedBankConflicts sample is available with Nsight Compute under `<nsight-compute-install-directory>/extras/samples/sharedBankConflicts`.

Chapter 3.

CONFIGURATION

The profiling results included in this document were collected on the following configuration:

- ▶ Target system: Linux (x86_64) with a NVIDIA RTX A2000 (Ampere GA106) GPU
- ▶ Nsight Compute version: 2022.3.0

The Nsight Compute UI screen shots in the document are taken by opening the profiling reports on a Windows 10 system.

Chapter 4.

INITIAL VERSION OF THE KERNEL

The initial version of the kernel `transposeCoalesced` uses shared memory to ensure that global memory accesses for loading data from the input matrix **`idata`** and storing data in the output matrix **`odata`** are coalesced. The matrix is sub-divided into tiles of size 32 x 32. The tile size is defined as:

```
#define TILE_DIM    32
```

For simplicity the code only handles square matrices whose dimensions are integral multiples of 32, the tile size. Each block transposes a tile of 32 x 32 elements. Each thread in the block transposes `TILE_DIM/BLOCK_ROWS` i.e. 4 elements, where `BLOCK_ROWS` is defined as:

```
#define BLOCK_ROWS  8
```

`TILE_DIM` must be an integral multiple of `BLOCK_ROWS`.

The way to avoid uncoalesced global memory access is to read the data into shared memory, and have each warp access noncontiguous locations in shared memory in order to write contiguous data to **`odata`**. The above procedure requires that each element in a tile be accessed by different threads, so a `__syncthreads()` call is required to ensure

that all reads from **idata** to shared memory have completed before writes from shared memory to **odata** commence.

```
__global__ void transposeCoalesced(float* odata, float* idata, int width, int
height)
{
    __shared__ float tile[TILE_DIM][TILE_DIM];

    int xIndex = blockIdx.x * TILE_DIM + threadIdx.x;
    int yIndex = blockIdx.y * TILE_DIM + threadIdx.y;
    int indexIn = xIndex + yIndex*width;

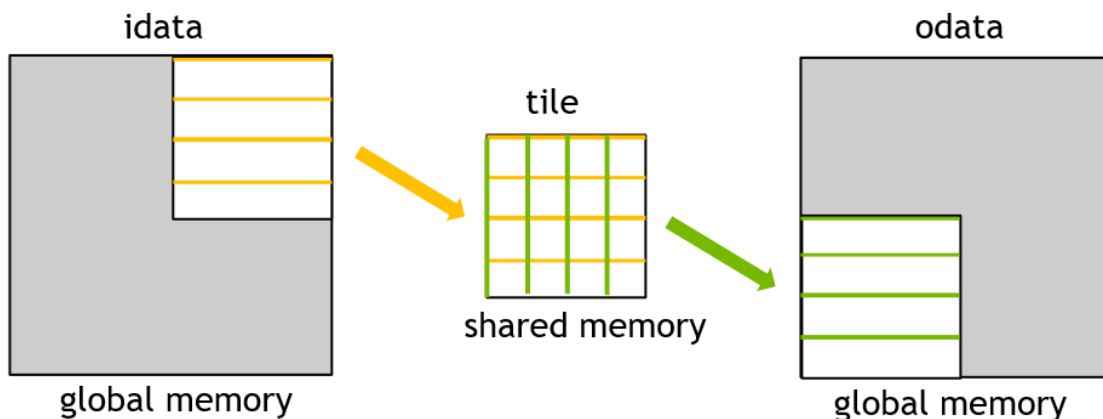
    xIndex = blockIdx.y * TILE_DIM + threadIdx.x;
    yIndex = blockIdx.x * TILE_DIM + threadIdx.y;
    int indexOut = xIndex + yIndex*height;

    for (int i = 0; i < TILE_DIM; i += BLOCK_ROWS)
    {
        tile[threadIdx.y + i][threadIdx.x] = idata[indexIn + i * width];
    }

    __syncthreads();

    for (int i = 0; i < TILE_DIM; i += BLOCK_ROWS)
    {
        odata[indexOut + i * height] = tile[threadIdx.x][threadIdx.y + i];
    }
}
```

A depiction of the data flow of a warp in the coalesced transpose kernel is given below. The warp writes four rows of the **idata** matrix tile to the shared memory 32x32 array "tile" indicated by the yellow line segments. After a **__syncthreads()** call to ensure all writes to tile are completed, the warp writes four columns of tile to four rows of an **odata** matrix tile, indicated by the green line segments.



Profile the initial version of the kernel

There are multiple ways to profile kernels with Nsight Compute. For full details see the [Nsight Compute Documentation](#). One example workflow to follow for this sample:

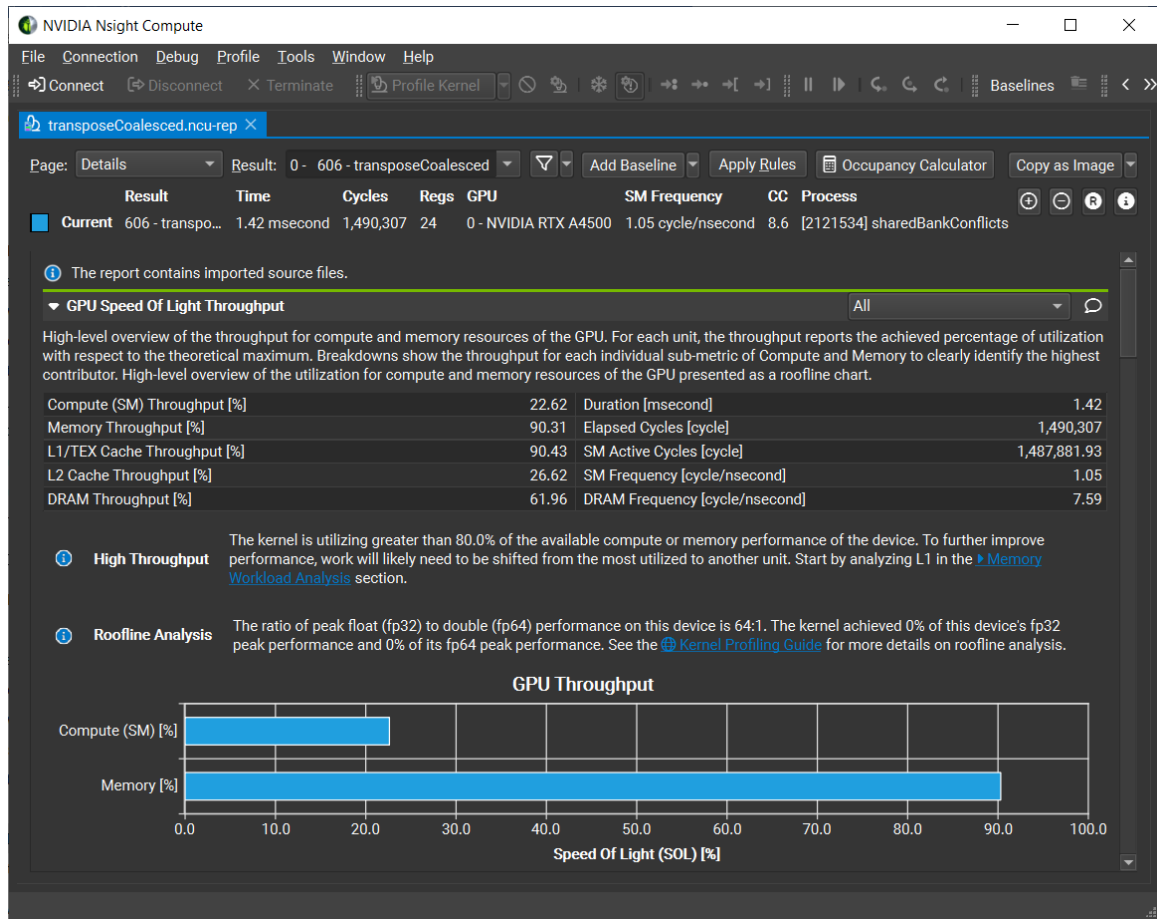
- ▶ Refer to the README distributed with the sample on how to build the application
- ▶ Run `ncu-ui` on the host system

- ▶ Use a local connection if the GPU is on the host system. If the GPU is on a remote system, set up a remote connection to the target system
- ▶ Use the "Profile" activity to profile the sample application
- ▶ Choose the "full" section set
- ▶ Use defaults for all other options

Details page - GPU Speed Of Light Throughput

The details page "GPU Speed Of Light Throughput" section provides a high-level overview of the throughput for compute and memory resources of the GPU used by the kernel.

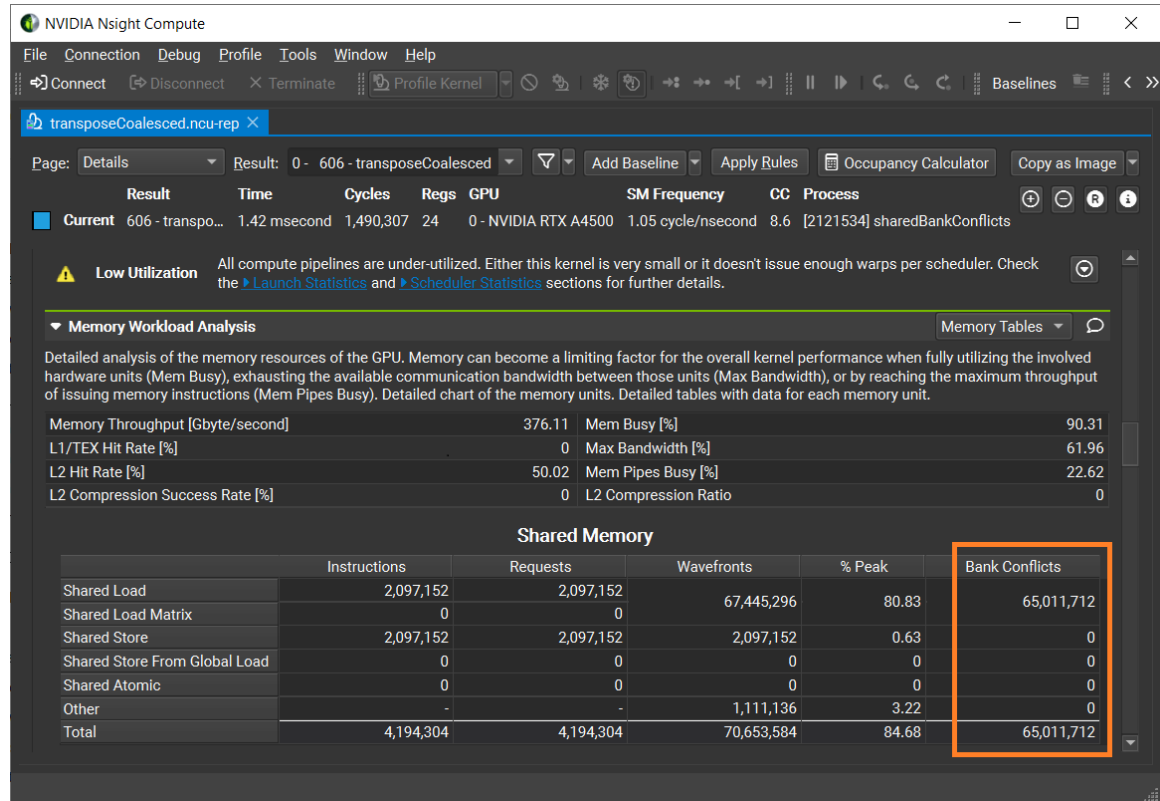
The duration for this initial version of the kernel is 1.42 milliseconds and this is used as the baseline for further optimizations.



For this kernel it shows a hint for High Memory Throughput and suggests looking at the memory workload analysis section. Click on Memory Workload Analysis.

Details page - Memory Workload Analysis section

The Shared Memory table under Memory Workload Analysis shows a high count of bank conflicts. We next check the Source Counters section to identify the source line resulting in bank conflicts.

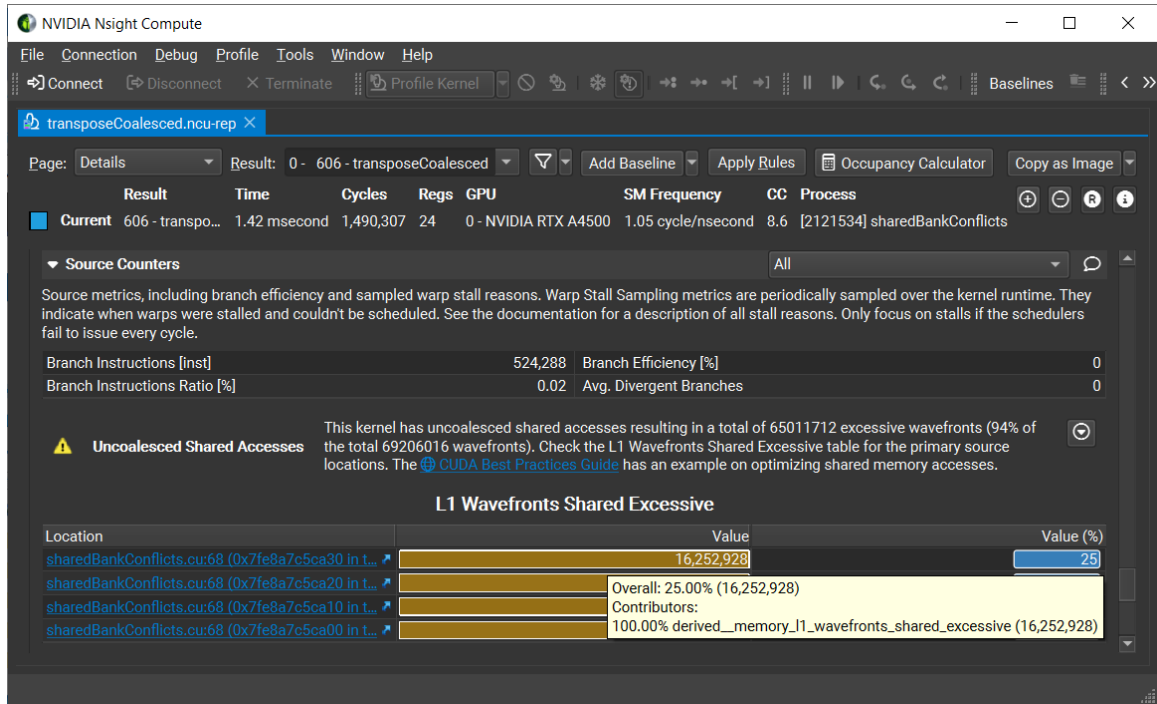


The screenshot shows the NVIDIA Nsight Compute interface. The 'Memory Workload Analysis' section is expanded, showing a 'Shared Memory' table. The table has columns: Instructions, Requests, Wavefronts, % Peak, and Bank Conflicts. The 'Bank Conflicts' column is circled in orange. The 'Shared Load' row shows a high count of bank conflicts (65,011,712).

	Instructions	Requests	Wavefronts	% Peak	Bank Conflicts
Shared Load	2,097,152	2,097,152	67,445,296	80.83	65,011,712
Shared Load Matrix	0	0			0
Shared Store	2,097,152	2,097,152	2,097,152	0.63	0
Shared Store From Global Load	0	0	0	0	0
Shared Atomic	0	0	0	0	0
Other	-	-	1,111,136	3.22	0
Total	4,194,304	4,194,304	70,653,584	84.68	65,011,712

Details page - Source Counters section

The Source Counters section table for the metric "L1 Wavefronts Shared Excessive" which is the indicator for shared memory bank conflicts lists the source lines with the highest value. Click on one of the source lines to view the kernel source at which the bottleneck occurs.



Source page

The CUDA source for the kernel is shown. When opening the Source page from Source Counters section, the Navigation metric is automatically filled in to match, in this case the "L1 Wavefronts Shared Excessive" metric. You can see this by the bolding in the column header. The source line at which the bottleneck occurs is highlighted.

It shows shared memory bank conflicts at line #68:

```
odata[indexOut + i * height] = tile[threadIdx.x][threadIdx.y + i];
```

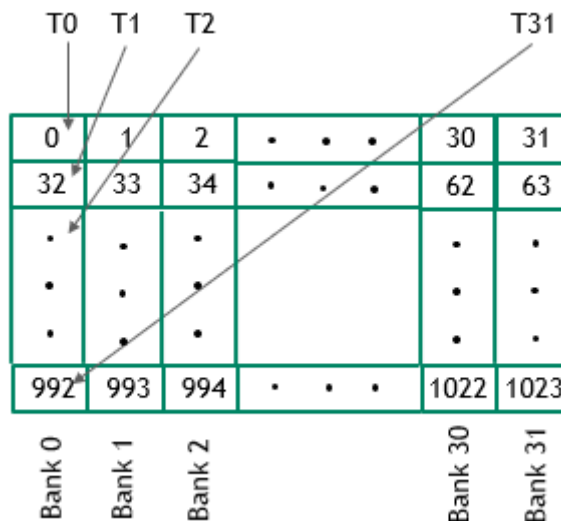
The screenshot shows the NVIDIA Nsight Compute interface. The top menu bar includes File, Connection, Debug, Profile, Tools, Window, and Help. The toolbar contains buttons for Connect, Disconnect, Terminate, Profile Kernel, and various execution controls. The main window displays the source code for the kernel `transposeCoalesced.ncu-rep`. The code is a CUDA kernel that processes a 2D array of floats, performing a transpose operation. The performance metrics table on the right shows the following data:

Instructions Executed	L1 Wavefronts Shared	L1 Wavefronts Excessive	L1 Wavefronts Shared
524,288			
1,048,576			
2,097,152			
524,288			
9,437,184	0	2,097,152	
1,048,576			
8,388,608	65,011,712	67,108,864	
524,288			

Chapter 5.

UPDATED VERSION OF THE KERNEL

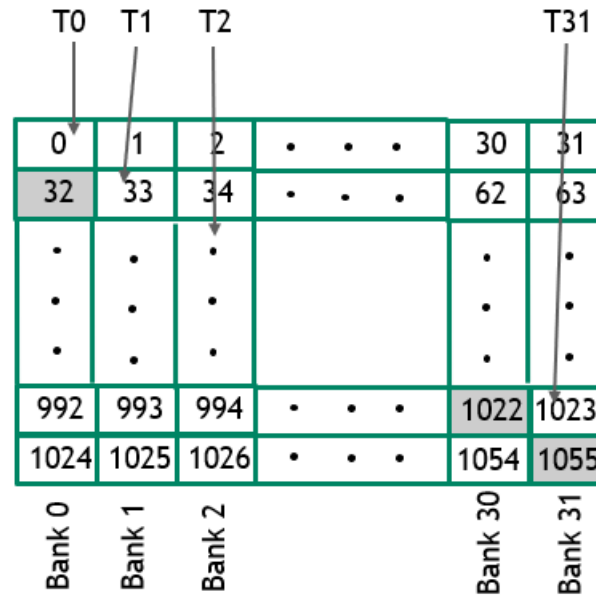
Considering the shared memory bank conflicts reported by the profiler we analyze the shared memory access pattern. The coalesced transpose uses a 32x32 shared memory array of floats. For this sized array, all data in each column is mapped to the same shared memory bank. As a result, when writing columns from the tile in shared memory to rows in **odata** the warp experiences a 32-way bank conflict and serializes the request.



A simple way to avoid this conflict is to pad the shared memory array by one column:

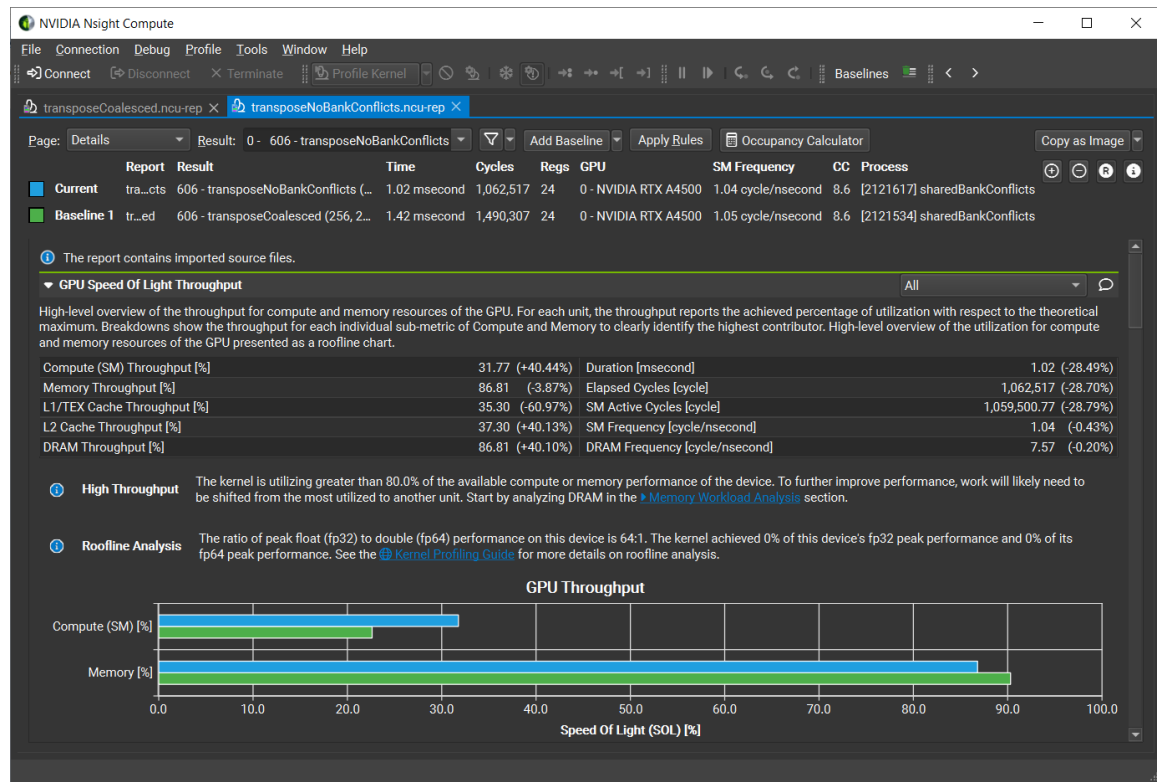
```
shared float tile[TILE_DIM][TILE_DIM+1];
```

The padding does not affect shared memory bank access pattern when a warp writes data to shared memory, which remains conflict free. But by adding a single column now the access of data in a column from shared memory by a warp is also conflict free. In the diagram below the elements in the extra column added for padding are shown with grey background.



Profile the updated kernel

The kernel duration has reduced from 1.42 milliseconds to 1.02 milliseconds. We can set a baseline to the initial version of the kernel and compare the profiling results.



We can confirm that there are no shared memory bank conflicts by looking at the Shared Memory metrics table under the Memory workload analysis section.

The screenshot displays the NVIDIA Nsight Compute interface. The top bar shows the application name and various toolbars. The main window is divided into several sections. The 'Memory Workload Analysis' section is expanded, showing a detailed analysis of memory resources. Below this, the 'Shared Memory' table is highlighted with an orange box. This table lists various memory access patterns and their corresponding metrics, including Instructions, Requests, Wavefronts, % Peak, and Bank Conflicts. The 'Bank Conflicts' column shows zero conflicts for all categories, indicating no shared memory bank conflicts.

	Instructions	Requests	Wavefronts	% Peak	Bank Conflicts
Shared Load	2,097,152 (+0.00%)	2,097,152 (+0.00%)	2,098,820 (-96.89%)	3.53 (-95.63%)	0 (-100.00%)
Shared Load Matrix	0 (+0.00%)	0 (+0.00%)			0 (+0.00%)
Shared Store	2,097,152 (+0.00%)	2,097,152 (+0.00%)	2,097,152 (+0.00%)	0.88 (+40.44%)	0 (+0.00%)
Shared Store From Global Load	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)
Shared Atomic	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)
Other	-	-	1,113,987 (+0.26%)	4.52 (+40.59%)	0 (+0.00%)
Total	4,194,304 (+0.00%)	4,194,304 (+0.00%)	5,309,959 (-92.48%)	8.94 (-89.45%)	0 (-100.00%)

Chapter 6.

RESOURCES

- ▶ GPU Technology Conference 2022 talk S41723: [How to Understand and Optimize Shared Memory Accesses using Nsight Compute](#)
- ▶ NVIDIA CUDA Sample transpose document - Optimizing Matrix Transpose in CUDA https://github.com/NVIDIA/cuda-samples/blob/master/Samples/6_Performance/transpose/doc/MatrixTranspose.pdf
- ▶ NVIDIA CUDA Sample transpose source code [transpose.cu](#)
- ▶ [Nsight Compute Documentation](#)

Notice

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.

Trademarks

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.

Copyright

© 2023-2023 NVIDIA Corporation and affiliates. All rights reserved.

This product includes software developed by the Syncro Soft SRL (<http://www.sync.ro/>).