



TUNING CUDA APPLICATIONS FOR VOLTA

DA-08647-001_v9.1 | April 2018

Application Note



TABLE OF CONTENTS

Chapter 1. Volta Tuning Guide	1
1.1. NVIDIA Volta Compute Architecture.....	1
1.2. CUDA Best Practices.....	1
1.3. Application Compatibility.....	2
1.4. Volta Tuning.....	2
1.4.1. Streaming Multiprocessor.....	2
1.4.1.1. Instruction Scheduling.....	2
1.4.1.2. Occupancy.....	2
1.4.1.3. Integer Arithmetic.....	2
1.4.2. Tensor Core Operations.....	3
1.4.3. Memory Throughput.....	3
1.4.3.1. High Bandwidth Memory 2 DRAM.....	3
1.4.3.2. Unified Shared Memory/L1/Texture Cache.....	3
1.4.4. NVLink Interconnect.....	4
Appendix A. Revision History	5

Chapter 1.

VOLTA TUNING GUIDE

1.1. NVIDIA Volta Compute Architecture

Volta is NVIDIA's latest architecture for CUDA compute applications. Volta retains and extends the same CUDA programming model provided by previous NVIDIA architectures such as Maxwell and Pascal, and applications that follow the best practices for those architectures should typically see speedups on the Volta architecture without any code changes. This guide summarizes the ways that an application can be fine-tuned to gain additional speedups by leveraging Volta architectural features.¹

Volta architecture comprises a single variant: GV100. A detailed overview of the major improvements in GV100 over earlier NVIDIA architectures is provided in a white paper entitled [NVIDIA Tesla V100 GPU Architecture: The World's Most Advanced Datacenter GPU](#).

For further details on the programming features discussed in this guide, please refer to the [CUDA C Programming Guide](#).

1.2. CUDA Best Practices

The performance guidelines and best practices described in the [CUDA C Programming Guide](#) and the [CUDA C Best Practices Guide](#) apply to all CUDA-capable GPU architectures. Programmers must primarily focus on following those recommendations to achieve the best performance.

The high-priority recommendations from those guides are as follows:

- ▶ Find ways to parallelize sequential code,
- ▶ Minimize data transfers between the host and the device,
- ▶ Adjust kernel launch configuration to maximize device utilization,
- ▶ Ensure global memory accesses are coalesced,

¹ Throughout this guide, *Kepler* refers to devices of compute capability 3.x, *Maxwell* refers to devices of compute capability 5.x, *Pascal* refers to device of compute capability 6.x, and *Volta* refers to devices of compute capability 7.x.

- ▶ Minimize redundant accesses to global memory whenever possible,
- ▶ Avoid long sequences of diverged execution by threads within the same warp.

1.3. Application Compatibility

Before addressing specific performance tuning issues covered in this guide, refer to the [Volta Compatibility Guide for CUDA Applications](#) to ensure that your application is compiled in a way that is compatible with Volta.

1.4. Volta Tuning

1.4.1. Streaming Multiprocessor

The Volta Streaming Multiprocessor (SM) provides the following improvements over Pascal.

1.4.1.1. Instruction Scheduling

Volta further simplifies Pascal's instruction scheduling. Each SM includes 4 warp-scheduler units. Each scheduler handles a static set of warps and issues to a dedicated set of instruction units. Warps are serviced over two cycles, and the schedulers can issue independent instructions every cycle. Dependent instruction issue latency for core FMA math operations are also reduced to a mere four clock cycles, compared to six cycles on Pascal.

Similar to GP100, the GV100 SM provides 64 FP32 cores and 32 FP64 cores. The GV100 additionally includes 64 INT32 cores and 8 mixed-precision Tensor Cores. GV100 provides up to 84 SMs.

1.4.1.2. Occupancy

The maximum number of concurrent warps per SM remains the same as in Pascal (i.e., 64), and other [factors influencing warp occupancy](#) remain similar as well:

- ▶ The register file size is 64k 32-bit registers per SM.
- ▶ The maximum registers per thread is 255.
- ▶ The maximum number of thread blocks per SM is 32.
- ▶ Shared memory capacity per SM is 96KB, similar to GP104, and a 50% increase compared to GP100.

Overall, developers can expect similar occupancy as on Pascal without changes to their application.

1.4.1.3. Integer Arithmetic

Unlike Pascal GPUs, the GV100 SM includes dedicated FP32 and INT32 cores. This enables simultaneous execution of FP32 and INT32 operations. Applications can now interleave pointer arithmetic with floating-point computations. For example, each

iteration of a pipelined loop could update addresses and load data for the next iteration while simultaneously processing the current iteration at full FP32 throughput.

1.4.2. Tensor Core Operations

Each Tensor Core performs the following operation: $D = AxB + C$, where A, B, C, and D are 4x4 matrices. The matrix multiply inputs A and B are FP16 matrices, while the accumulation matrices C and D may be FP16 or FP32 matrices.

When accumulating in FP32, the FP16 multiply results in a full precision product that is then accumulated using FP32 addition with the other intermediate products for a 4x4x4 matrix multiply. In practice, Tensor Cores are used to perform much larger 2D or higher dimensional matrix operations, built up from these smaller elements.

The Volta tensor cores are exposed as Warp-Level Matrix Operations in the CUDA 9 C++ API. The API exposes specialized matrix load, matrix multiply and accumulate, and matrix store operations to efficiently use Tensor Cores from a CUDA-C++ program. At the CUDA level, the warp-level interface assumes 16x16 size matrices spanning all 32 threads of the warp. See the *CUDA C Programming Guide* for more information.

1.4.3. Memory Throughput

1.4.3.1. High Bandwidth Memory 2 DRAM

Like GP100, GV100 uses four memory dies per HBM2 stack, and four stacks, with a maximum of 16 GB of GPU memory. A faster and more efficient HBM2 implementation delivers up to 900 GB/s of peak memory bandwidth, compared to 732 GB/s for GP100. The HBM2 efficiency of the GV100 memory controller has been significantly improved as well. The combination of both a new generation HBM2 memory, and a new generation memory controller in Volta, provides 1.5x delivered memory bandwidth versus Pascal GP100, and greater than 95% memory bandwidth efficiency running many workloads.

In order to hide DRAM latencies at full HBM2 bandwidth, more memory accesses must be kept in flight compared to GPUs equipped with traditional GDDR5. Helpfully, the large complement of SMs in GV100 will typically boost the number of concurrent threads (and thus reads-in-flight) compared to previous architectures. Resource constrained kernels that are limited to low occupancy may benefit from increasing the number of concurrent memory accesses per thread.

1.4.3.2. Unified Shared Memory/L1/Texture Cache

In Volta the L1 cache, texture cache, and shared memory are backed by a combined 128 KB data cache. As in previous architectures, such as Kepler, the portion of the cache dedicated to shared memory (known as the *carveout*) can be selected at runtime using `cudaFuncSetAttribute()` with the attribute `cudaFuncAttributePreferredSharedMemoryCarveout`. Volta supports shared memory capacities of 0, 8, 16, 32, 64, or 96 KB per SM.

A new feature, Volta enables a single thread block to address the full 96 KB of shared memory. To maintain architectural compatibility, static shared memory allocations

remain limited to 48 KB, and an explicit opt-in is also required to enable dynamic allocations above this limit. See the *CUDA C Programming Guide* for details.

Like Pascal, Volta combines the functionality of the L1 and texture caches into a unified L1/Texture cache which acts as a coalescing buffer for memory accesses, gathering up the data requested by the threads of a warp prior to delivery of that data to the warp.

Volta increases the maximum capacity of the L1 cache to 128 KB, more than 7x larger than the GP100 L1. Another benefit of its union with shared memory, the Volta L1 improves in terms of both latency and bandwidth compared to Pascal. The result is that for many applications Volta narrows the performance gap between explicitly managed shared memory and direct access to device memory. Also, the cost of register spills is lowered compared to Pascal, and the balance of occupancy versus spilling should be re-evaluated to ensure best performance.

1.4.4. NVLink Interconnect

NVLink is NVIDIA's high-speed data interconnect. NVLink can be used to significantly increase performance for both GPU-to-GPU communication and for GPU access to system memory. GV100 supports up to six NVLink connections with each connection carrying up to 50 GB/s of bi-directional bandwidth.

NVLink operates transparently within the existing CUDA model. Transfers between NVLink-connected endpoints are automatically routed through NVLink, rather than PCIe. The `cudaDeviceEnablePeerAccess()` API call remains necessary to enable direct transfers (over either PCIe or NVLink) between GPUs. The `cudaDeviceCanAccessPeer()` can be used to determine if peer access is possible between any pair of GPUs.

Appendix A. REVISION HISTORY

Version 1.0

- ▶ Initial Public Release

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

© 2012-2018 NVIDIA Corporation. All rights reserved.