An Instruction Roofline Model for GPUs

Nan Ding, Samuel Williams
Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
{*NanDing, SWWilliams*}*@lbl.gov*

***Abstract*—The Roofline performance model provides an intuitive approach to identify performance bottlenecks and guide performance optimization. However, the classic FLOP-centric approach is inappropriate for the emerging applications that perform more integer operations than floating point operations. In this paper, we propose an Instruction Roofline Model on NVIDIA GPUs. The Instruction Roofline incorporates instructions and memory transactions across all memory hierarchies together and provides more performance insights than the FLOP-oriented Roofline Model, i.e., instruction throughput, stride memory access patterns, bank conflicts, and thread predication. We use our Instruction Roofline methodology to analyze five proxy applications: HPGMG from AMReX, BatchSW from merAligner, Matrix Transpose benchmarks, cudaTensorCoreGemm, and cuBLAS. We demonstrate the ability of our methodology to understand various aspects of performance and performance bottlenecks on NVIDIA GPUs and motivate code optimizations.**

***Keywords*-Instruction Roofline Model, NVIDIA GPUs, memory patterns**

I. Introduction

Migrating an application to a new architecture is a challenge, not only in porting the code but also in understanding and tuning the performance. Rather than manually performing the analysis, developers tend to use tools to motivate the optimization. Therefore, performance analysis tools are becoming one of the most critical components for modern architectures. Performance modeling, the critical technology to quantify performance characteristics and identify potential performance bottlenecks associated with machine capabilities, is becoming an indispensable tool understanding performance behaviors and guiding performance optimization.

The Roofline model [1] is a visually-intuitive method for users to understand performance by coupling together floating-point performance, data locality (arithmetic intensity), and memory performance into a two-dimensional graph. The Roofline model [2–4] can tell whether the code is either memory-bound across the full memory hierarchy or compute-bound. Unfortunately, even with sufficient data locality, one cannot guarantee high performance. Many applications perform more integer operations than floating-point, and there are applications in emerging domains, e.g., graph analytics, genomics, etc.. that perform no floating-point operations at all. The classic, FLOP-centric Roofline model is inappropriate for such domains. To that end, we develop an Instruction Roofline Model for GPUs to affect performance analysis of integer-heavy computations.

The contributions in this paper include:

- Definition of instruction Roofline ceilings that characterize the peak machine instruction throughput.
- Creation of memory walls that define a range of efficiency for memory access to global or shared memory.
- Development of an application characterization methodology that incorporates instruction throughput and thread predication into the Roofline model.
- Development of an application execution characterization methodology that allows easy visualization of global and shared memory access patterns.
- Evaluation of the Instruction Roofline Model and methodology on NVIDIA's latest V100 GPUs using five proxy applications: HPGMG (mix of floating-point and integer instructions), BatchSW (integer-only), MatrixTranspose (load/store), and cudaTensorCoreGemm (tensor core Warp-Matrix-Multiply-Accumulate operations) and cuBLAS (cublasGemmEx API).

II. The Classic Roofline Model

The Roofline model characterizes a kernel's performance in GigaFLOPs per second (GFLOP/s) as a function of its arithmetic intensity (AI), as described as Eq.(1). The AI is expressed as the ratio of floating-point operations performed to data movement (FLOPs/Bytes). For a given kernel, we can find a point on the X-axis based on its AI. The Y-axis represents the measured GFLOP/s. This performance number can be compared against the bounds set by the peak compute performance (Peak GFLOPs) and the memory bandwidth of the system (Peak GB/s) to determine what is limiting performance: memory or compute.

$$GFLOP/s \leq \min \begin{cases} Peak\ GFLOP/s \\ Peak\ GB/s \times Arithmetic\ Intensity \end{cases} \quad (1)$$

The classic Roofline model has been successfully used for performance analysis on different architectures. In prior work, researchers created additional compute ceilings (e.g. "no FMA" peak) and memory ceilings (e.g., cache levels) [3, 5, 6]. However, such refinement is misplaced when the bottleneck is not floating-point in nature but pertains to integer instruction throughput or memory access.

III. Instruction Roofline Model for GPUs

Intel Advisor [7] introduced the ability to analyze integer-heavy applications and generate scalar and vector integer *operation* (IntOP) ceilings. At its core, the Roofline model in Intel Advisor is based on *operations* (floating-point or integer). In either case, the vertical axis remains performance

(FLOP/s, *int*/s, or *int/s+float/s*) while the horizontal axis remains Arithmetic Intensity (operations per byte).

Whereas Intel Advisor can be quite effective in this regard, it suffers from two aspects. First, it only captures pipeline throughput and may not detect instruction fetch-decode-issue bottlenecks. Second, it is an x86 CPU-only solution. The latter is particularly troublesome given the ascendancy of accelerated computing.

In order to affect the Instruction Roofline analysis for GPU-accelerated applications, we need to target a different set of metrics. First, rather than counting floating-point and/or integer operations, we count instructions. Counting instructions allows us to both identify fetch-decode-issue bottlenecks, and, when categorized by types, pipeline utilization. Thread predication is another critical performance factor on GPUs. When a branch is executed, threads that don't take the branch are predicated (masked in vector parlance) so that they do not execute subsequent operations. When predication is frequent, one may observe poor kernel performance as very few threads execute work on any given cycle. Finally, the nature of GPU computing makes efficient data movement a critical factor in application execution time. As such, it is essential that we also characterize the global and shared memory access patterns to assess the efficiency of data motion and motivate future code optimization. Although developers can use nvprof [8] and nvvp [9] to diagnose the performance bottlenecks discussed above, the Instruction Roofline Model provides an approachable means of characterizing performance bottlenecks in a single figure.

A. Architectural Characterization

First, we describe how we define the Instruction Roofline ceilings and memory pattern walls. Here we use NVIDIA's latest V100 GPU (GV100) [10] to describe the methodology, but it is applicable to any GPU architecture.

Instructions and Bandwidth Ceilings: Each GV100 Streaming Multiprocessor (SM) consists of four processing blocks (warp schedulers), and each warp scheduler can dispatch one instruction per cycle. As such, the theoretical maximum (warp-based) instruction/s is $80(SM) \times 4(warp\ scheduler) \times 1(instruction/cycle) \times 1.53(GHz) = 489.6$ GIPS. Memory access is coalesced into transactions. The transaction size for global/local memory, the L2 cache, and HBM are 32 bytes. The shared memory transaction size is 128 bytes. In practice, a warp-level load may generate anywhere from 1 to 32 transactions depending on memory patterns. This makes the "transaction" the natural unit when analyzing memory access. We leverage Yang et al.'s methodology [2] for measuring GPU bandwidths but rescale into billions of transactions per second (GTXN/s) based on the transaction size.

The Instruction Roofline Model is described in Eq.(2). A kernel's performance, characterized in billions of instructions per second (GIPS), is a function of peak machine bandwidth (GTXN/s), Instruction Intensity, and machine peak GIPS. "Instruction Intensity" on the GPU is defined as warp-based instructions per transaction. Figure 1 shows the resultant Instruction Roofline ceilings for the GV100. L1, L2, and HBM bandwidths of 14000, 2996, and 828 GB/s sustain 437, 93.6, and 25.9 GTXN/s when normalized to the typical 32-byte transaction size.

$$GIPS \leq \min \begin{cases} Peak\ GIPS \\ Peak\ GTXN/s \times Instruction\ Intensity \end{cases} \tag{2}$$

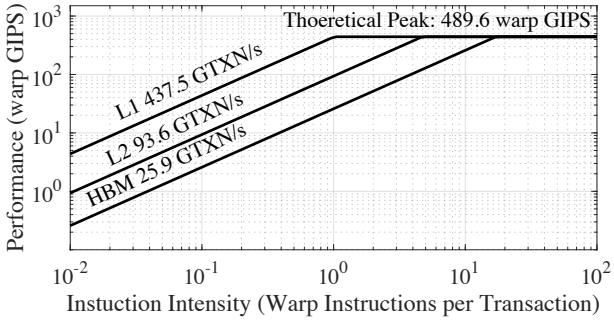


Figure 1: Instruction Roofline Model for the GV100. 1 GTXN/s is 10^9 transactions per second.

Global Memory Walls: When a warp executes an instruction to access global memory, it is crucial to consider the access pattern of threads in that warp because inefficient memory access can lower the performance by generating superfluous transactions.

As discussed, a warp-level load instruction can generate anywhere from 1 to 32 transactions. We can recast such a ratio into two key instruction intensities: 1 and $1/32$ warp-level *global* loads per *global* transaction. The former arises when all threads in a warp reference the same memory location and only a single transaction is generated. We call this "stride-0". Conversely, the other extreme can occur for a number of scenarios including random access, and striding by over 32 bytes ("stride-8" if FP32/INT32 and "stride-4" if FP64). Unit-stride ("stride-1"), memory access provides a global LD/ST intensity of $1/8$ (FP64) and $1/4$ (FP32, INT32). Thus, on the Instruction Roofline, we may plot three intensity "walls" representing stride-0, stride-1 (unit-stride), and stride-8.

Shared Memory Walls: GPU shared memory is a highly-banked structure within each SM providing 4-byte access. In theory, this allows for all 32 threads in a warp to make concurrent random access to shared memory in a single 128-byte transaction. However, as there are only 32 banks on the GV100, two threads contending for the same bank but different 4-byte words will cause a "bank conflict" and multiple transactions will be generated. In the worst case, all 32 threads hit different 4-byte words in the same bank, and 32 transactions are generated. As with

global/local memory walls, we can visualize bank conflicts on the Instruction Roofline Model. Note, there are two key instruction intensities: 1 and 1/32 warp-level *shared* loads per *shared* transaction.

B. Application Characterization

Mirroring the previous section that characterizes GPU performance capabilities, in this section, we describe the methodology we employ to characterize application execution in terms of the Instruction Roofline Model.

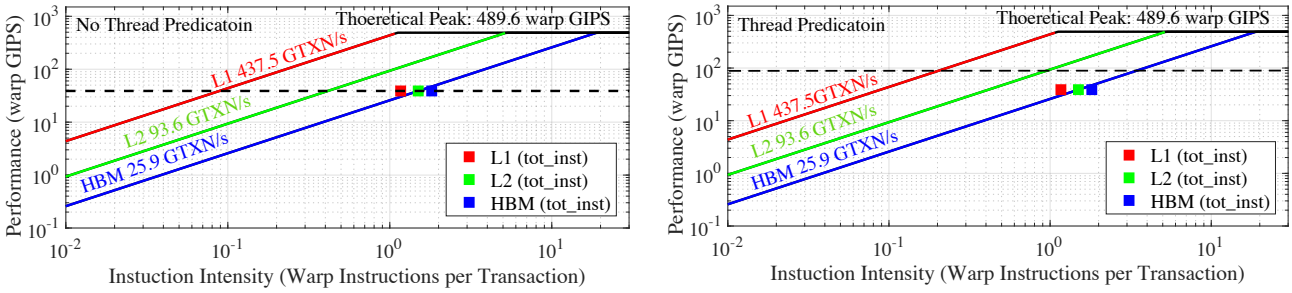
Instruction Intensity and Performance: The instruction Roofline requires we measure three terms. We use `inst_executed_thread`/32 to record the number of instructions executed by each kernel (scaled to warp-level), and `nvprof -print-gpu-summary` to extract kernel run time. We use the sum of `gld_transactions` and `gst_transactions` to record the total number of global transactions (for L1) and the sum of `shared_load_transactions` and `shared_store_transactions` to record the total number of shared transactions (for L1). Similarly, we use the sum of `l2_read_transactions` and `l2_write_transactions` to record the total number of L2 transactions, and the sum of `dram_read_transactions` and `dram_write_transactions` to record the total number of HBM transactions. The ratio $\frac{inst\_executed\_thread/32}{HBM\ transactions}$ is the HBM Instruction Intensity while $\frac{inst\_executed\_thread/32}{1e9 * run\ time}$ is instruction performance in GIPS.

Figure 2 visualizes the resultant Instruction Roofline Model for an arbitrary kernel. As with the traditional Roofline, one may infer cache reuse based on the distance between points (no reuse) and performance bounds based on how closes each colored point is to the associated colored bandwidth ceiling or peak performance.

Tensor Cores: Tensor Cores in GV100 are programmable matrix-multiply-and-accumulate units that can deliver up to 125 TFLOP/s. Whereas the traditional Roofline is premised on "operations", as discussed in the previous section, we can recast intensity in terms of instructions. We can take that one step further and use floating-point instructions or tensor instruction. Such metrics are particularly useful in understanding why pipeline utilization can be high even if FLOP/s is low. We use *smsp_inst_executed_pipe_tensor.sum* in NSight Compute [11] to collect the total number of warp-based instructions (HMMA) executed on tensor cores. Similar to how previous work would plot floating-point performance relative to peak and a "no FMA" peak [2], we plot sustained HMMA GIPS relative to either peak GIPS or peak HMMA GIPS.

Thread Predication: Quantifying the performance impact of thread predication relative to bandwidth bottlenecks can be critical in determining the overall kernel performance impact. Moreover, thread predication is an intuitive way to understand resource utilization — the fraction of threads in one warp that are active during the execution. Recall that regardless of whether one thread in a warp is executing an instruction, or 32 threads in a warp are executing 32 instructions, only one warp-level instruction is executed. Therefore, the ratio of warp-based instructions (`inst_executed`) to thread-based instructions (`inst_thread_executed/32`) is the degree of predication. When the number is close to 1.0, there is little predication, and the performance impact is small.

Figure 2(a) and (b) highlight two cases of thread predication. In both figures, the dotted line represents the observed warp-level instruction performance. A dotted line close to the theoretical peak indicates instruction issue rates are bottlenecked by hardware issue rates. Thread-level instruction performance is constrained to be less than or equal to this bound (ceiling). Conversely, the dots represent thread instruction throughput normalized by 32 (warp size) for each level of memory. By definition, (normalized) thread-level instruction rates must be less than warp instruction rates.



(a) Instruction Roofline with no thread predication.

(b) Instruction Roofline with thread predication.

Figure 2: Thread predication in the context of the Instruction Roofline Model. A kernel without predication (left) has thread instruction throughput (dots) matched to the warp instruction throughput (dotted line) while a kernel with moderate predication (right) has thread instruction throughput substantially below the warp instruction throughput.

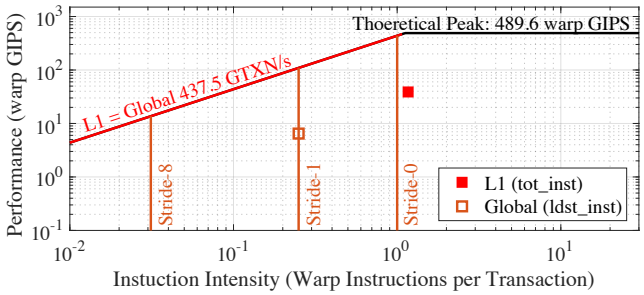


Figure 3: Global memory walls on the Instruction Roofline plot for the GV100. A generic kernel has been plotted on the Roofline for pedagogical purposes. The position of the open dot visualizes the stride of the memory access pattern.

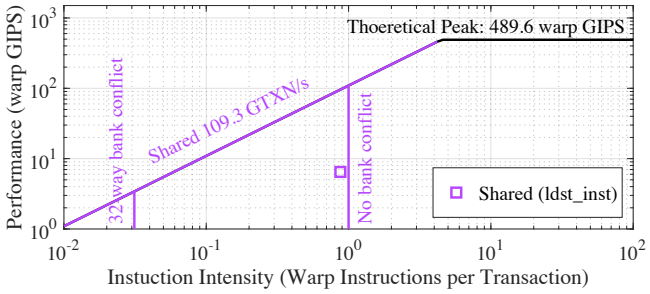


Figure 4: Shared memory walls on the Instruction Roofline plot on GV100. A generic kernel has been plotted on the Roofline for pedagogical purposes. The position of the open dot visualizes the number and impact of bank conflicts.

In Figure 2(a), the dots (thread-level instruction throughput) fall on the dotted line (warp-level instruction throughput) indicating no thread predication. Conversely, in Figure 2(b), the dots are well below the dotted line indicating a 2× loss in performance due to thread predication.

Global Memory Pattern Walls: Figure 3 shows the Instruction Roofline Model with Global Memory Walls. The solid red dot's instruction intensity is based on total instructions and L1 transactions that include global, local, and shared memory. Whereas global/local transactions are 32 bytes, shared memory transactions are 128 bytes. Thus, in order to calculate the number of equivalent 32-byte transactions and create a common denominator, one must scale the number of shared memory transactions by four. The resultant denominator, $1 \times \textit{global transactions} + 4 \times \textit{shared transactions}$, allows for direct comparisons to the L1 ceiling and allows us to determine whether the combined effect of global, local, and shared transactions has made the code L1-bound. We may refine it (open orange dot) to include only warp-level global load/store instructions (e.g. `inst_executed_global_loads`) and transactions (e.g. `gld_transactions`). This resultant dot will have both a different performance (GIPS) and a different instruction intensity as there are fewer load and store instructions than total instructions. The distance between the open and solid points is the fraction of load/store instructions constitute the dynamic instruction mix (close points indicate code that are mostly load/store).

The position of the load/store instruction intensity relative to the memory walls visualizes the average memory access pattern. In this generic example, we see that the dot lies on the unit-stride wall indicating the kernel accesses memory in a unit-stride manner. If it were to move to the left, one would conclude strided or gather memory access while if it were to move to the right, one would conclude multiple threads access the same word in memory.

Shared Memory Walls: Recasting our previous discussion from global memory and the L1 cache to shared memory, Figure 4 shows the Instruction Roofline Model with shared memory bandwidth and shared Memory Walls. Note, shared memory GTXN/s is less than global memory (L1 cache) GTXN/s in as the shared memory transaction size is 128 bytes while the global memory size is 32 bytes. Here, striding effects have been juxtaposed with bank conflicts. We measure the number of warp-level shared load or store instructions a kernel executes with `inst_executed_shared_loads` and `inst_executed_shared_stores` and the number of shared load and store transactions with `shared_load_transactions` and `shared_store_transactions`. We may use these terms to calculated shared memory instruction intensity.

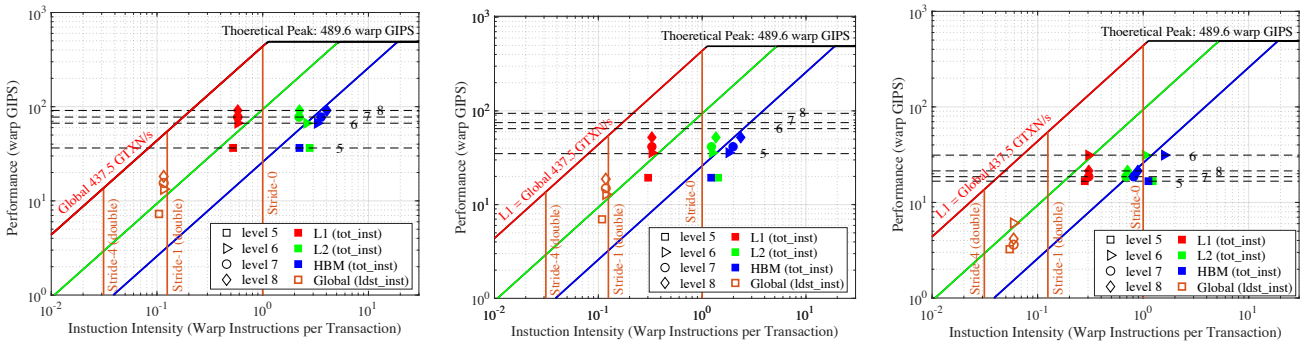
Consider the exemplar kernel in Figure 4. We may plot its shared memory instruction intensity and performance (open dot) relative to the shared memory bandwidth and walls. We observe that the kernel is efficiently accessing shared memory as it is close to the "no bank conflict" wall. Conversely, kernels that generate large numbers of bank conflicts will move to the left towards the "32-way bank conflict" wall. Similarly, we can compare the shared open dot to the shared memory ceiling to tell whether the code is bound by the shared memory.

IV. Results

In this section, we describe our test machine and profiling tool, and use the Instruction Roofline Model to evaluate and analyze several GPU-accelerated applications.

A. Experimental Setup

Results presented in this paper were obtained on the GPU-accelerated partition on Cori (Cori-GPU) at NERSC and Summit at OLCF. Cori-GPU is comprised of nodes with two Intel Skylake CPUs and eight NVIDIA V100 GPUs, while each compute node on Summit contains two IBM POWER9 processors and six NVIDIA V100 accelerators. Nevertheless, in all experiments, we use only a single process running on one GPU and thus mitigate NVLink, PCIe, and host



(a) Instruction Roofline for GSRB_FP. (b) Instruction Roofline for GSRB_BRANCH. (c) Instruction Roofline for GSRB_STRIDE2.

Figure 5: Instruction Rooflines on GV100 for the three implementations of HPGMG's GSRB. Solid dots are the total number of instructions (tot_inst) executed by non-predicated threads, open dots refer to the global memory access patterns, and dotted lines are warp-level instructions. Note, "level #" refers to the level in the multigrid v-cycle.

processor performance. As a result, Cori-GPU and Summit benchmark performance will be the same for purposes of this paper. On Both machines, we use CUDA 10, `nvprof` [8], and NSight Compute [11].

We evaluate the Instruction Roofline methodology for NVIDIA GPUs using five proxy applications: HPGMG [12–14] from AMRex [15], BatchSW [16] from merAligner [17], Matrix Transpose benchmarks [18], cudaTensorCoreGemm [19], and cuBLAS [20]. These applications exhibit a range of computational characteristics including data types, data locality, and thread predication properties. Specifically, HPGMG executes roughly 50% integer instructions and 30% floating-point instructions, matrix transpose performs almost entirely load/store instructions, BatchSW performs entirely integer instructions, and both cudaTensorCoreGemm and cuBLAS perform HMMA instructions but use different implementations. We show that unlike the classical FLOP-centric Roofline model, the Instruction Roofline Model and our methodology can effectively analyze such a wide range of applications.

B. HPGMG

HPGMG is a geometric multigrid benchmark to proxy the multigrid solves in block structured AMR (Adaptive Mesh Refinement) applications that use the AMReX framework. HPGMG solves the 4^{th} order, variable-coefficient Laplacian on a unit-cube with Dirichlet boundary conditions using a multigrid F-cycle. The Gauss-Seidel, Red-Black (GSRB) smoother dominates HPGMG's run time. GSRB smoothers perform two stencil kernel invocations per smooth (red and black). Cells are marked as either red or black in a 3D checkerboard pattern. Cells matching the sweep color are updated, while the others are copied to the result array.

HPGMG includes three different implementations of its GSRB smoother: GSRB_FP, GSRB_BRANCH, and GSRB_STRIDE2. All three perform the same computation and touch the same data over the course of a thread blocks execution, but vary memory access and predication. As such, they are ideal cases to demonstrate the subtle performance differences using the Instruction Roofline Model. GSRB_BRANCH is conceptually the simplest implementation. In it, a thread block operates on a 2D slice or a 3D cache block. Within the slice, red-black execution is affected through a branch. This branch reduces the number of non-predicated threads without reducing the number of warps. The branch is eliminated in GSRB_FP through multiplication by a precomputed array of 1's and 0's. As such, predication is eliminated at the cost of doubling nominal computation whilst maintaining the same number of warp-level instructions. GSRB_STRIDE2 is similar to GSRB_BRANCH with the caveat that the 2D thread block's x-dimension is half the 2D tile's x-dimension. Thus, each thread is responsible for updating two adjacent points. Through clever address calculation, FLOPs are reduced, predication is eliminated, and the number of warp-level instructions is minimized. Note, none of these implementations use shared memory. As such, only global memory stride walls are relevant. We show how the performance insights gained from the Instruction Roofline correlate with the performance observations.

Figure 5 shows the Instruction Roofline on GV100 for the three implementations of GSRB as a function of multigrid level using eight 128^3 boxes on the finest level (largest arrays). On each of these figures, we overlay both total instruction intensity/performance as well as global memory access pattern (global load/store intensity/performance). The former is comparable to the nominal instruction Roofline, while the latter is shown relative to the memory stride walls. GIPS generally increases from level 5 (smallest arrays) to level 8 (largest arrays) as the overhead:surface:volume ratio improves. Figure 5a (GSRB_FP) makes it immediately obvious that: (1) memory access is unit-stride (open brown

dots are very close to the stride-1 wall), (2) data reuse is captured by the L1, but not the L2 (red solid are far from green solid, but green solid are close to blue solid), and (3) the blue dots are very close to the HBM ceiling indicating GSRB_FP is HBM-bound.

Figure 5b presents the Instruction Roofline Model for GSRB_BRANCH. The thread predication impact from the branch in GSRB_BRANCH is immediately obvious in this figure as the dots (scaled thread GIPS) are well below the dashed black lines (warp-based GIPS). Although the thread-level instruction throughput (dots) of GSRB_FP (Figure 5a) and GSRB_BRANCH (Figure 5b) differs by a factor of two, the Instruction Roofline Model for GPUs makes it clear that both implementations stress the architecture's issue bandwidth to the same degree (equal dotted lines imply equal warp-based GIPS). In terms of memory access pattern, we see no difference between GSRB_FP and GSRB_BRANCH. This should come as no surprise as the both implementations generate the same number of warp-level load/store instructions and access global memory in the same manner (same number of transactions).

Figure 5c shows the Instruction Roofline for GSRB_STRIDE2. Recall, in this implementation, there is no predication and redundant computation is minimized. Thus, the requisite number of thread and warp instructions per transaction should be further reduced which we see in reduced instruction intensity compared to Figure 5a. However, as L1 and L2 data locality crash for levels 7 and 8 (lower L2 and HBM intensity), data movement increases and the net effect is decreased performance (HBM-bound with superfluous data movement). This results in a decrease in GIPS (solid dots) with increasing level. On the converse, the Instruction Roofline shows GSRB_STRIDE2 presents a different memory access pattern from GSRB_BRANCH or GSRB_BRANCH as its load intensity is lower than the stride-1 wall (open dots).

We use Figure 6 to further demonstrate the capability of the Instruction Roofline Model. First, we can see the execution time of level 7 and 8 in GSRB_STRIDE2 implementation is twice that of GSRB_FP and GSRB_BRANCH. This fact can be inferred from the Figure 5c and Figure 5b: GSRB_STRIDE2 and GSRB_BRANCH have a very similar number of thread instructions. As such, GSRB_STRIDE2's lower GIPS is indicative of a longer execution time. Second, GSRB_FP and GSRB_BRANCH implementations have the same execution time but very different GFLOP/s. GFLOP/s of GSRB_FP is double that of GSRB_BRANCH due to the redundant computation. The performance of levels 7-8 of GSRB_STRIDE2 is lower than GSRB_BRANCH. This is because the longer execution time of level 7-8 in GSRB_BRANCH. Resource utilization (the fraction of threads in one warp that are active during the execution) of GSRB_BRANCH is 50% while the other two implementations are 100% due to the thread predication in

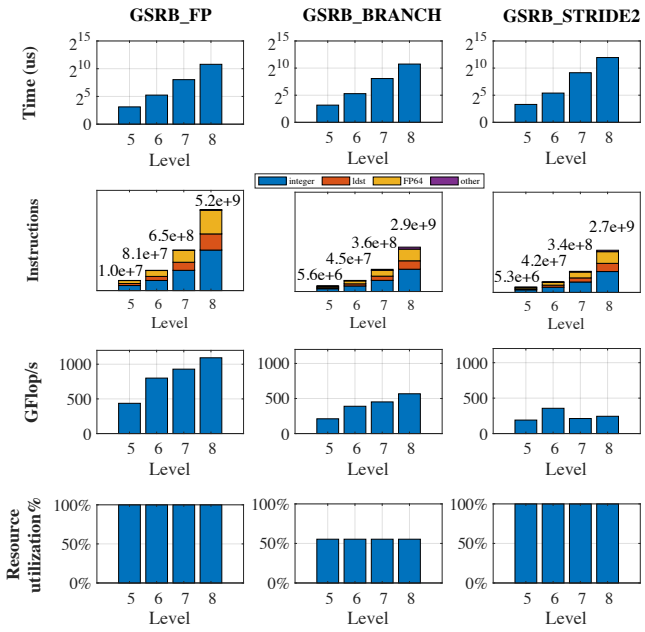


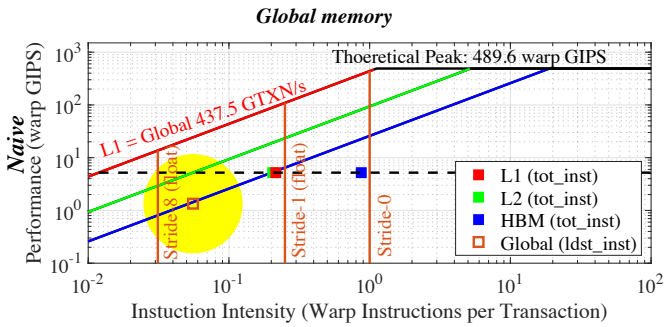
Figure 6: Performance insights breakdown on GV100 for the three different implementations of HPGMG as a function of multigrid level (note, level 8 represents the largest arrays). Resource utilization refers to the ratio of thread predication.

GSRB_BRANCH. All of these facts can be captured by the Instruction Roofline in Figure 5. The third observation in Figure 6 is that the integer instructions take nearly 50% of the total number of instructions, load/store instructions take about 20%, and floating-point instructions take about 30%. The integer instructions are performing *IMAD*, *IADD* and *ISETP* for indexing the preparing the offsets of stencil computations. This indicates the code performs 1.6 integer instructions and 0.6 load/store instructions for every floating-point instruction. "Others" in the figure refers to the instructions like *ctv, etc.* which are not counted in any specific metric in *nvprof*.

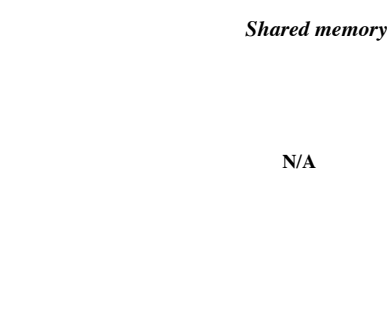
C. Matrix Transpose

Matrix transpose flips a matrix A over its diagonal, that is, it swaps the row and column indices of the matrix by producing another matrix A^T. We use three different single-precision implementations (Naive, Coalesced, and Coalesced_NoBankConflict) to illustrate the different performance of both global and shared memory access patterns. In all cases, we use a 1024×1024 matrix and all three implementations using 32×8 thread blocks operating on 32×32 matrix tiles.

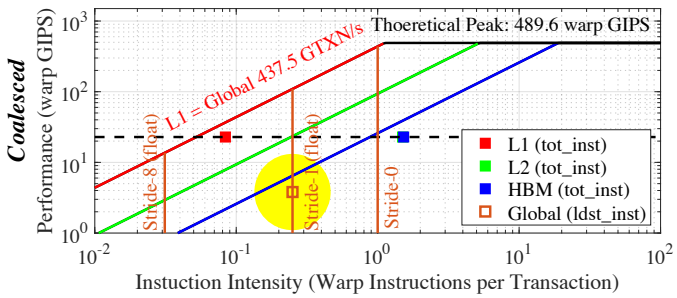
The Naive implementation uses the array index to access elements in both input arrays and output arrays. Each thread reads four elements from one column of the input matrix and writes them to their transposed locations in one row of the output matrix. As the matrices are column-major,



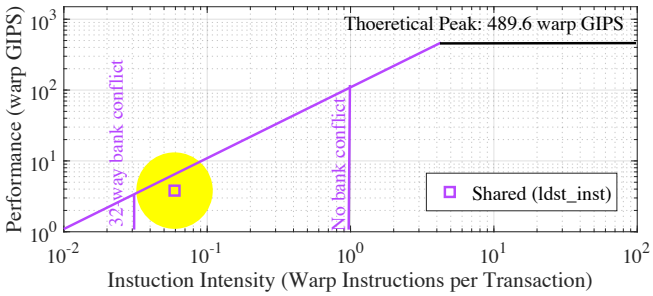
(a) Matrix Transpose Naive: global memory pattern.



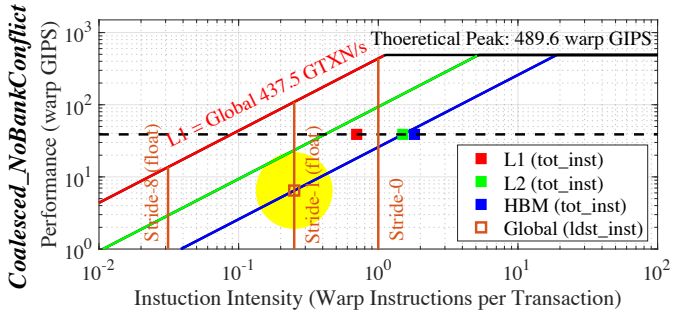
(b) Matrix Transpose Naive: shared memory pattern.



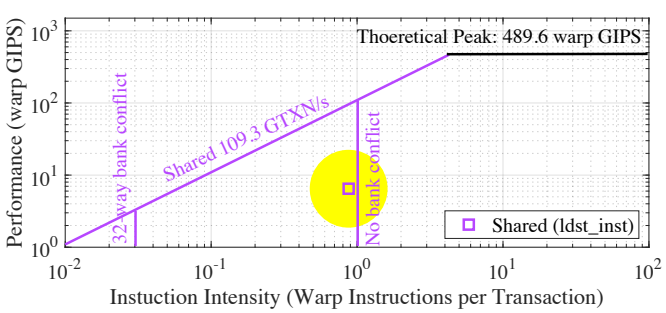
(c) Matrix Transpose Coalesced: global memory pattern.



(d) Matrix Transpose Coalesced: shared memory pattern.



(e) Matrix Transpose Coalesced_NoBankConflict: global memory pattern.



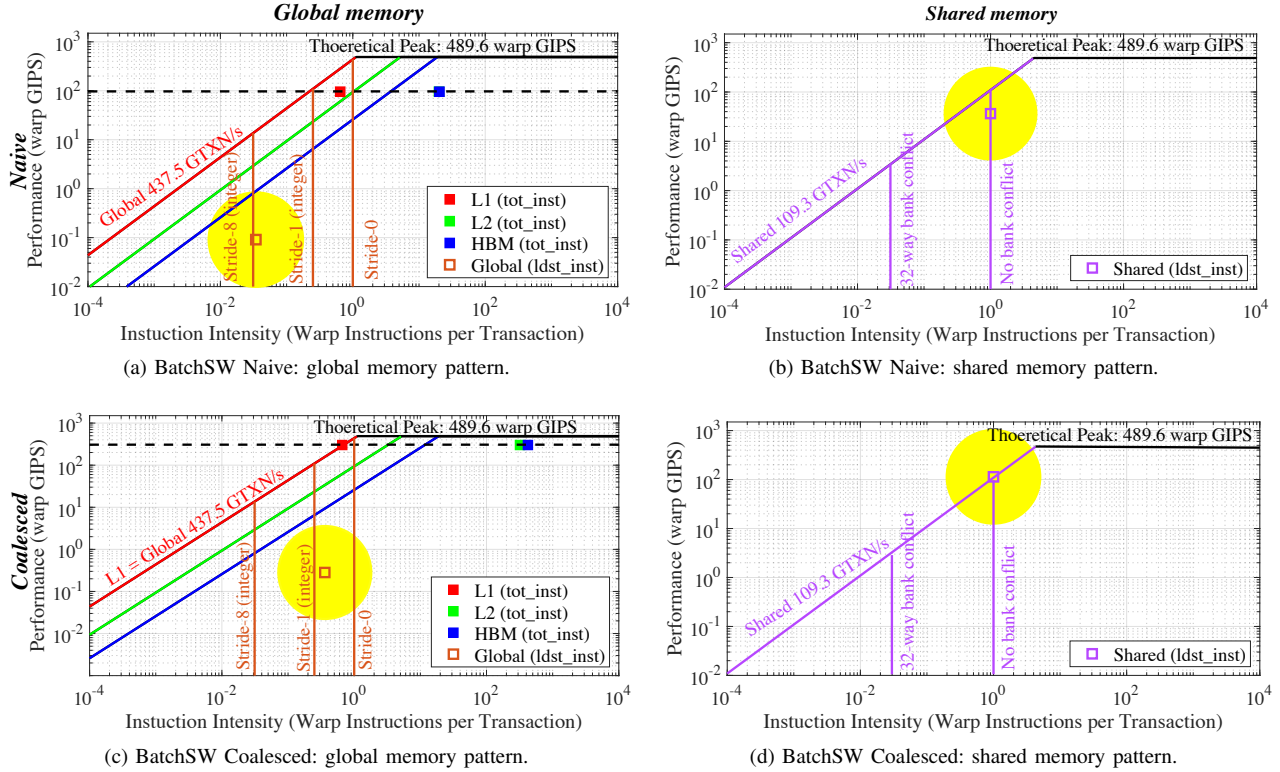
(f) Matrix Transpose Coalesced_NoBankConflict: shared memory pattern.

Figure 7: Instruction Roofline on GV100 for the three implementations in Matrix Transpose. The solid dots are the total number of instructions (tot_inst) executed by non-predicated threads. The open dots refer to memory access patterns.

the reads from the input matrix are coalesced while the writes to the output matrix have a stride of 4096 Bytes (1024 floats) between successive threads. This results in the worst case of 32 separate memory transactions per warp-based load/store instruction. Figure 7a plots the instruction Roofline and memory walls for the Naive implementation. Clearly, memory access (open symbol) is far from unit-stride on average. Moreover, we see poor L1 data locality but good L2 locality (green and red dots are close but red and blue are widely separated). Note, as the Naive implementation doesn't use shared memory, there is no corresponding shared memory figure in Figure 7a.

The Coalesced implementation reads contiguous data from matrix A into rows of a shared memory buffer. After re-calculating the array index, a column of the shared memory buffer is written to contiguous addresses in the matrix A^T. As such, the global memory access pattern of the Coalesced implementation is unit-stride; Figure 7c shows exactly this (open dot). We can also see from Figure 7c that that L1 cache is better utilized than the Naive implementation as the L2 intensity has moved to the right (green dot). Importantly, the number of transactions in the red solid dot (L1) is a linear combination of global and shared memory transactions. The combined effect of these has made the code nearly L1-bound (L1 solid red dot is close to the L1 ceiling).

Although the use of shared memory has substantially improved performance, Figure 7d shows this implementation produces a large number of shared memory bank conflicts



(a) BatchSW Naive: global memory pattern.

(b) BatchSW Naive: shared memory pattern.

(c) BatchSW Coalesced: global memory pattern.

(d) BatchSW Coalesced: shared memory pattern.

Figure 8: Instruction Roofline on GV100 for the two implementations in BatchSW with visualized global memory pattern. The solid dots are the total number of instructions (tot_inst) executed by non-predicated threads. The open dots refer to the memory access patterns.

— shared load intensity (open dot) is close to the "32-way bank conflict" wall. This is easily understood as the Coalesced implementation uses a 32×32 shared memory array of floats but all data in columns k and $k + 16$ are mapped to the same bank. As a result, when writing partial columns from buffer in the shared memory to rows in the output matrix, there is a 16-way bank conflict as Figure 7d shows. Concurrently, the shared intensity (open dot) is very close to the shared memory ceiling indicating that shared transactions are dominating all L1 transactions and is the cause of the L1 bottleneck.

Proximity to the shared memory wall and ceiling can be used to motivate software optimization. The Coalesced_NoBankConflict implementation pads the shared memory array by one column $32 \times (32 + 1)$ to avoid bank conflicts. Figure 7f shows that shared load intensity has moved to the right and is now near the "no bank conflict" wall. Note, the Coalesced_NoBankConflict version has same global load intensity (Figure 7e) as the Coalesced version (Figure 7c) because they have the same implementation for global memory access.

As a summary for Matrix Transpose, by comparing Figure 7a, Figure 7c and Figure 7e, we can see how the global memory access pattern improved from the Naive to Coalesced and Coalesced_NoBankConflict implementations. By comparing Figure 7d and Figure 7f, we can tell how the shared memory access pattern improved from the Coalesced to Coalesced_NoBankConflict implementations.

D. BatchSW

BatchSW [16] proxies merAligner's sequence alignment phase [17]. merAligner is a parallel sequence aligner that implements a *seed-and-extend* algorithm and employs parallelism in all of its components. merAligner spends a significant portion of its run time using the Smith-Waterman (SW) algorithm [21] which is based on dynamic programming. Within SW, a $M \times N$ matrix A is constructed, where M and N are the lengths of the two sequences. Matrix element $A(i, j)$ is calculated as a score for aligning the i^{th} element in the first sequence with the j^{th} element in the second sequence. The score of $A(i, j)$ depends on $A(i, j - 1)$, $A(i - 1, j)$ and $A(i - 1, j - 1)$. Next, the optimal local alignment is defined as tracing back a path connecting the starting point (i_0, j_0) in the matrix to any point $(i, j), i > i_0, j > j_0$, with the highest sum of scores along this path. In this paper, we set $M = 126, N = 1066$

with a total number of 15,000 pairs of sequences — a typical case in merAligner.

BatchSW is a GPU version of the Smith-Waterman algorithm. BatchSW has two implementations: Naive and Coalesced. Both implementations have one kernel of two phases: scoring and tracing. The scoring phase of the two implementations is the same. The execution flow the matrix is diagonal-major: $A(0,0) \rightarrow (A(1,0), A(0,1)) \rightarrow (A(2,0), A(1,1), A(0,2))$, and successive diagonals can be done in the same manner. The three most recent diagonals are stored in three different shared memory arrays. Each thread in one warp writes the scores to a unique cell of the shared memory. The tracing phases of the two implementations are different. The Naive implementation uses a row-major data layout to store the scores for the whole matrix, whereas the Coalesced version uses a diagonal-major layout.

Figure 8 shows the Instruction Roofline on GV100 for the two implementations. When examining total instruction throughput, it is clear the Naive implementation underperforms attaining roughly 25% of the peak instruction Roofline GIPS. Comparing Figures 8a and 8c, it is immediately obvious how replacing a poor memory access pattern (every N^{th} element) in the row-major data layout with a diagonal-major data layout improves load intensity (open dots) to the point where it is bound by the stride-1 wall.

Shared memory is used in the scoring phase in both implementations. Recall that the scoring phase of the two implementation is the same but are in the same kernel as the tracing phase. Each thread in one warp writes the scores to a unique cell of shared memory with the three most recent diagonals stored in shared memory. Thus, the 32 threads in a warp access 32 consecutive elements of shared memory and thus 32 different banks. Figures 8b and 8d show that both implementations attain a perfect shared intensity of 1.0 (no bank conflicts). Despite having the same intensity, the coalesced implementation attains a higher GIPS due to the shortened kernel execution time. In fact, coalesced shared load/store GIPS is very close to the shared memory roofline and thus indicative of the ultimate performance bound.

E. Matrix Multiplication using Tensor Cores

Each tensor core can complete a single 4×4 FP16 matrix multiplication and FP32 accumulation per cycle, i.e. $D = A \times B + C$, where A, B, C are 4×4 matrices [10]. Currently, the lowest level interface to program Tensor Cores is CUDA's Warp Matrix Multiply and Accumulation (WMMA) API [22]. The WMMA API provide warp-wide operations for performing the computation of $D = A \times B + C$, where A (FP16), B (FP16), C (FP32) and D (FP32) can be tiles of larger matrices. All threads in a warp cooperatively work together to perform a matrix-multiply and accumulate operation on these tiles using the WMMA API. Matrix A($M \times N$) and C($M \times K$) are row-major and matrix B($N \times K$) is column-major.

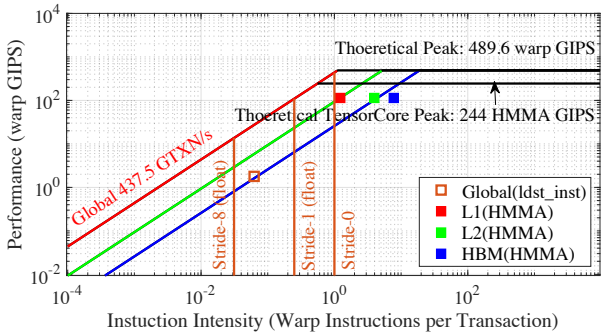
Another popular method to perform matrix multiplications is to use cuBLAS [23]. The cuBLAS library is an highly-tuned implementation of BLAS (Basic Linear Algebra Subprograms) on top of the NVIDIA CUDA runtime. We use the cuBLASLt API supported by CUDA 10.1. The cuBLASLt is a new lightweight library dedicated to GEneral Matrix-matrix Multiply (GEMM) operations.

Unfortunately, these two methods can have very different performance when performing the same matrix multiplications — cuBLAS can attain 104 TFLOP/s while WMMA can only attain 58.23 TFLOP/s. As such, it is essential to understand the performance nuances of these two methods in order to motivate the future code and hardware optimization. To that end, in this section we use the Instruction Roofline Model to evaluate both approaches using test matrices of size $M = N = K = 32,768$.

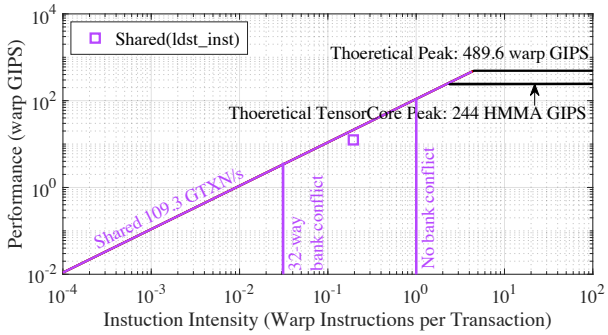
WMMA Roofline: There are three WMMA instructions in cudaTensorCoreGemm: `wmma.load`, `wmma.store`, and `wmma.mma`. `wmma.load` and `wmma.store` are broken into a group of normal load and store instructions. Each `wmma.mma` instruction is broken into 16 HMMA instructions for mixed precision. As such, there are $16 \times \frac{M \times N \times K}{16 \times 16 \times 16}$ HMMA instructions, where $16 \times 16 \times 16$ matrix operations can be performed by one WMMA instruction. In other words, there are 512 floating-point operations per HMMA instruction. Thus, we may define tensor core HMMA peak instruction ceiling by tensor core peak TFLOP/s by 512: $\frac{125 \times 1e3 \ GFLOP/s}{512} = 244$ HMMA GIPS.

In cudaTensorCoreGemm, a 128×128 tile is computed each iteration. Within each tile, each warp computes eight 16×16 sub-tiles using `wmma.mma` operations by iterating through the full A and B matrices and accumulating the intermediate result in the local thread state. At the beginning of each iteration, a CUDA block of eight warps copies a 128×128 tile of the two input matrices from the global memory to shared memory with warps 0-3 copying matrix A and warps 4-7 copying matrix B. Each warp is assigned 64 16×16 sub-matrices and works on one at a time. Threads in one warp are organized as a 2×16 block, i.e., thread 0 loads element (0,0) to element (0,7), thread 1 loads element (1,0) to element (1,7), thread 15 loads element (0,8) to element (0,15), etc.. Thus, each warp loads the corresponding data with a stride of eight elements using 16 transactions. As such, we can see the global memory pattern in Figure 9a is between the stride-8 and the stride-1 walls.

Before performing the `wmma.mma` operation, each warp loads data using the `wmma.load`. The shared memory access pattern is not explicitly specified, but each thread in the warp can read one or multiple matrix elements from different matrix rows or columns. The worst case is that each thread accesses a different row mapped into the same bank (32-way bank conflict). As such, portions of the A and B matrices are stored in shared memory with an additional padding to reduce the number of shared memory

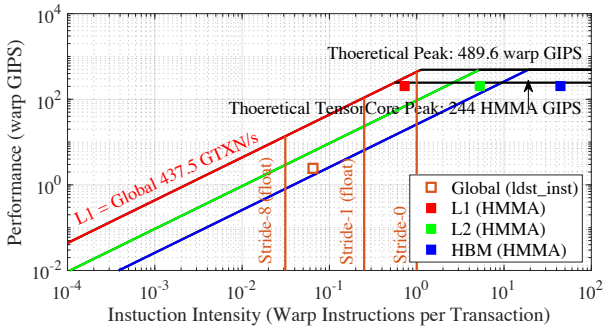


(a) Visualized global memory pattern of cudaTensorCoreGemm in Instruction Roofline on GV100.

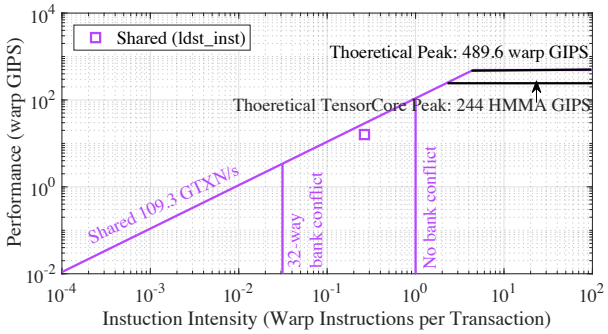


(b) Visualized shared memory pattern of cudaTensorCoreGemm in Instruction Roofline on GV100.

Figure 9: Instruction Roofline on GV100 for the cudaTensorCoreGemm (WMMA interface). The solid dots are HMMA instructions while the open dots are global/shared loads/stores.



(a) Visualized global memory pattern of in Instruction Roofline on GV100.



(b) Visualized shared memory pattern in Instruction Roofline on GV100.

Figure 10: Instruction Roofline on GV100 for the cuBLAS. The solid dots are just HMMA instructions while the open dots are global/share load/store.

access bank conflicts. The number of eight FP16 elements is chosen as the minimum shift in cudaTensorCoreGemm. This is because `wmma.load` requires 128-bit alignment. As a result, Figure 9b shows that bank conflicts are reduced to only 6-way. As Figure 9b also shows that the WMMA code is bound by the shared memory bandwidth, further reductions in bank conflicts can improve performance.

cuBLAS Roofline: Our cuBLAS benchmark simply calls *cublasGemmEx* to perform the matrix-to-matrix Multiply. Figure 10 shows the Instruction Roofline plot for it. Observe, cuBLAS has the same global and shared memory access pattern with cudaTensorCoreGemm (open dot). However, cuBLAS performance (solid dot) has increased with increased instruction intensity compared to cudaTensorCoreGemm in Figure 9. As the two implementations perform the same number of HMMA instructions, one can infer that cuBLAS requires fewer transactions than cudaTensorCoreGemm (better register and L2 locality) and has a higher performance as it is L1-bound.

V. Conclusions and Future Work

We developed and applied a methodology for analyzing instruction throughput on a GPU using the Roofline model. This allows us to analyze both total instruction throughput (fetch-decode-issue) as well as function unit utilization (FPU, tensor, integer, etc...) thereby expanding the applicability of roofline to several emerging computational domains. With the insight that load/store instructions coupled with transaction categorization allows us to quantify both the memory access pattern (e.g. unit-stride vs. gather/scatter) as well as the frequency of shared memory bank conflicts, we were able to incorporate both aspects into the Roofline model as walls that denote the efficiency of memory access. The unified visualization of bandwidth and access efficiency endows users with far greater insights as to how different aspects of modern GPU architectures constrain performance.

In the future, we will apply our methodology to other accelerated architectures and extend the access efficiency concept to networking, I/O, and lustre file systems.

ACKNOWLEDGEMENTS

This material is based upon work supported by the Advanced Scientific Computing Research Program in the U.S. Department of Energy, Office of Science, under Award Number DE-AC02-05CH11231. This research used resources of the National Energy Research Scientific Computing Center (NERSC) which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 and the Oak Ridge Leadership Facility which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. We thank NVIDIA Corporation for their willingness to answer our myriad of questions on nvprof metrics.

REFERENCES

[1] S. Williams, A. Waterman, and D. Patterson, "Roofline: An Insightful Visual Performance Model for Multicore Architectures," *Commun. ACM*, vol. 52, no. 4, 2009.

[2] C. Yang, T. Kurth, and S. Williams, "Hierarchical roofline analysis for gpus: Accelerating performance optimization for the nersc-9 perlmutter system."

[3] T. Koskela, Z. Matveev, C. Yang, A. Adedoyin, R. Belenov, P. Thierry, Z. Zhao, R. Gayatri, H. Shan, L. Oliker *et al.*, "A novel multi-level integrated roofline model approach for performance characterization," in *International Conference on High Performance Computing*. Springer, 2018, pp. 226–245.

[4] S. W. Williams, *Auto-tuning performance on multicore computers*. University of California, Berkeley, 2008.

[5] D. Doerfler, J. Deslippe, S. Williams, L. Oliker, B. Cook, T. Kurth, M. Lobet, T. Malas, J.-L. Vay, and H. Vincenti, "Applying the roofline performance model to the intel xeon phi knights landing processor," in *International Conference on High Performance Computing*. Springer, 2016, pp. 339–353.

[6] A. Ilic, F. Pratas, and L. Sousa, "Cache-aware Roofline model: Upgrading the loft," *IEEE Computer Architecture Letters*, vol. 13, no. 1, pp. 21–24, 2014.

[7] "Integer Roofline Modeling in Intel Advisor." https://software.intel.com/en-us/articles/a-brief-overview-of-integer-roofline-modeling-in-intel-advisor#ref_releasenotes.

[8] "Nvidia Profiler User's Guide." https://docs.nvidia.com/cuda/profiler-users-guide/.

[9] "NVIDIA Visual Profiler." https://developer.nvidia.com/nvidia-visual-profiler.

[10] "NVIDIA Tesla V100 GPU Architecture." http://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf.

[11] "Nsight Compute Command Line Interface." https://docs.nvidia.com/nsight-compute/pdf/NsightComputeCli.pdf.

[12] "HPGMG CUDA Code." https://bitbucket.org/nsakharnykh/hpgmg-cuda.

[13] "HPGMG Website." https://hpgmg.org/.

[14] "HPGMG-FV Documentation." http://crd.lbl.gov/departments/computer-science/PAR/research/hpgmg.

[15] "AMReX Documentation." https://amrex-codes.github.io/amrex/.

[16] "BatchSW CUDA Code." https://bitbucket.org/mgawan/batch-sw/src/master/.

[17] E. Georganas, A. Buluç, J. Chapman, L. Oliker, D. Rokhsar, and K. Yelick, "meraligner: A fully parallel sequence aligner," in *2015 IEEE International Parallel and Distributed Processing Symposium*. IEEE, 2015, pp. 561–570.

[18] "Matrix Transpose Code." https://github.com/NVIDIA-developer-blog/code-samples/.

[19] "CUDA Samples," http://hpc-simulations.utp.ac.pa/wp-content/uploads/2018/04/CUDA_Samples-min.pdf.

[20] "Tensor Core Miniapp," https://github.com/PointKernel/tensor-core-miniapp.

[21] M. Zhao, W.-P. Lee, E. P. Garrison, and G. T. Marth, "Ssw library: an simd smith-waterman c/c++ library for use in genomic applications," *PloS one*, vol. 8, no. 12, p. e82138, 2013.

[22] "CUDA C Programming Guide (CUDA 9.0)." https://docs.nvidia.com/cuda/archive/9.0/cuda-c-programming-guide/.

[23] "cuBLAS," https://docs.nvidia.com/cuda/cublas/.

APPENDIX
ARTIFACT DESCRIPTION

A. Abstract

The key contribution of this paper is the methodology of the Instruction Roofline Model for GPUs. The hardware and software environment used in this paper are all publicly available as described below.

B. Description

Machines: Results presented in this paper were obtained on the GPU-accelerated partition on Cori (Cori-GPU) at NERSC and Summit at OLCF. Cori-GPU is comprised of nodes with two Intel Skylake CPUs and eight NVIDIA V100 GPUs, while each compute node on Summit contains two IBM POWER9 processors and six NVIDIA V100 accelerators. On Both machines, we use CUDA 10, `nvprof`, and NSight Compute. In all experiments, we use only a single process running on one GPU and thus mitigate NVLink, PCIe, and host processor performance. As a result, a benchmark running on a V100 on Summit performs the same as it would on a V100 on Cori-GPU.

Metrics for data collection: Table I lists the nvprof metrics used to measure instructions and data movement on GV100 CUDA core and NSight Compute metrics for tensor core HMMA instructions.

Table I: Metrics for Instruction Roofline Model

	Metrics	Descriptions
thread-based	inst_thread_executed	non-predicated
warp-based	inst_executed	total instructions
	inst_executed_global_loads inst_executed_global_stores inst_executed_local_loads inst_executed_local_stores inst_executed_shared_loads inst_executed_shared_stores	L1 Cache Instructions
	gld_transactions gst_transactions local_load_transactions local_store_transactions shared_load_transactions shared_store_transactions	L1 Cache Transactions
	l2_read_transactions l2_write_transactions	L2 Cache
	dram_read_transactions dram_write_transactions	HBM Memory
	sysmem_read_transactions sysmem_write_transactions	PCIe/NVLINK
	smsp_inst_executed_pipe_tensor.sum	tensor core
kernel-based	nvprof –print-gpu-summary	execution time

Profiling command lines: We use `nvprof --print-gpu-summary` to collect the kernel execution time. The examples of profiling command line for timing collection, cuda core metrics and tensor core metrics described below are for Cori-GPU. One can replace `srun` with `jsrun` on Summit.

For timing collection: `srun -n 1 nvprof --print-gpu-summary ./transpose`

For cuda code: `srun -n 1 nvprof --kernels ''transposeNaive'' --csv --metrics inst_executed ./transpose`

For tensor core: `srun -n 1 nv-nsight-cu-cli -k ''compute_gemm'' --metrics smsp__inst_executed_pipe_tensor.sum ./cudaTensorCoreGemm`

Applications: The five proxy applications we evaluated in the paper are described below. All application are built with `arch=compute_70`.

1) HPGMG from AMRex, the source code can be found https://bitbucket.org/nsakharnykh/hpgmg-cuda. We run HPGMG with eight 128^3 boxes for the three implementations. This results in multigrid levels 5-8 running on GPU and levels 1-4 on CPU. As this paper is focused on Roofline on GPUs, we only examine levels 5-8.
2) BatchSW from merAligner, the source code can be found https://bitbucket.org/mgawan/batch-sw/src/master/. We run BatchSW using two sequences length of 126 and 1066 with a total number of 15,000 pairs of sequences which is a typical case in merAligner.
3) Matrix Transpose from cuda sample, the source code can be found https://github.com/NVIDIA-developer-blog/code-samples/. The matrix size we use is 1024×1024, and all three implementations using 32×8 thread blocks operating on 32×32 matrix tiles.
4) cudaTensorCoreGemm from cude sample, the source code can be found https://github.com/NVIDIA/cuda-samples/tree/master/Samples/cudaTensorCoreGemm. The matrix size of the three metrices are all 32,768×32,768.
5) cuBLAS, the source code can be found https://github.com/PointKernel/tensor-core-miniapp. The matrix size of the three metrices are all 32,768×32,768.