

Update on Triangle Counting on GPU

Carl Pearson*, Mohammad Almasri*, Omer Anjum*, Vikram S. Malthody*, Zaid Qureshi[§],

Rakesh Nagi[‡], Jinjun Xiong[†], and Wen-mei Hwu*

*ECE, [§]CS, [‡]ISE, University of Illinois at Urbana-Champaign, Urbana, IL 61801

[†]Cognitive Computing Systems Research, IBM Thomas J. Watson Research Center, Yorktown Heights, NY, 10598

{pearson, almasri3, oanjum, vsm2, zaidq2}@illinois.edu,
nagi@illinois.edu, jinjun@us.ibm.com, w-hwu@illinois.edu

Abstract—This work presents an update to the triangle-counting portion of the subgraph isomorphism static graph challenge. This work is motivated by a desire to understand the impact of CUDA unified memory on the triangle-counting problem. First, CUDA unified memory is used to overlap reading large graph data from disk with graph data structures in GPU memory. Second, we use CUDA unified memory hints to solve multi-GPU performance scaling challenges present in our last submission. Finally, we improve the single-GPU kernel performance from our past submission by introducing a work-stealing dynamic algorithm GPU kernel with persistent threads, which makes performance adaptive for large graphs without requiring a graph analysis phase.

Index Terms—GPU, graph algorithms, triangle counting

I. INTRODUCTION

Triangles are a simple non-trivial structure of many graphs, and are used as foundational elements for graph analyses and transformations. The widespread adoption of GPUs in commodity, cloud-computing, and specialized supercomputing systems, provides substantial motivation for accelerating graph operations on GPUs, whether for stand-alone operations or as part of an extended pipeline. Of particular interest is whether large-scale graph operations can be effectively parallelized to multiple tightly-connected GPUs.

A key challenge of using GPUs for large-scale graph operations is the low compute intensity of those operations and the relatively low bandwidth between GPU memory and system memory. At the same time, GPU system integrators like IBM have developed and released programmer-friendly and tightly-integrated heterogeneous CPU/GPU systems, where GPUs have full access to expansive host memory and interconnects that have an order of magnitude higher bandwidth than previously available. In previous submissions, we presented our collaborative CPU/GPU approach to triangle counting and k -truss decomposition based on CUDA’s zero-copy and unified memory for optimizations to our static graph challenge algorithms.

In this work, we examine the impact of programmer-friendly and tightly-integrated GPU programming APIs on triangle counting. We make the following contributions:

- 1) Triangle counting based on persistent-kernel dynamic algorithm selection;
- 2) Evaluation of performance impact of using unified memory for in-GPU-memory triangle counting;

- 3) Evaluation of performance impact of using unified memory for computation support, implicit graph partitioning, and multi-GPU scaling;
- 4) Unified memory for overlap of graph construction in GPU memory with disk I/O.

II. DESIGN

The algorithm operates on an unweighted directed acyclic graph $G = (E, V)$, where each edge E is from a source vertex V_s to a destination vertex V_d . Each triangle is uniquely associated with a particular “base” edge. Each edge in the graph may be the base edge of zero or more triangles, and the triangle count for the graph is the sum of the triangle count of each base edge. Most datasets are provided by the GraphChallenge in a sorted edge-list file format. Similar to our previous submission, we follow an edge-oriented approach.

A. Graph Storage

In previous submissions, we used a hybrid “COO+CSR” (coordinate + compressed-sparse-row) format, where the CSR *colInd* array is paired with an equal-length *rowInd* array containing the non-zero row index. In a traditional CSR, each parallel task is typically assigned to a row, and then while processing that row additional parallelism may optionally be leveraged to access each entry in that row. Parallel tasks are not directly mapped to row entries, as the length of each row is not known until the CSR is accessed. The inclusion of the *rowInd* array makes it straightforward to parallelize across edges instead of rows in the CSR, as each parallel task can look up the source vertex of an edge without traversing the *rowPtr* array. We refer to the **rowPtr**, **rowInd**, and **colInd** arrays in the algorithms in Section II-D.

B. CUDA Unified Memory and Hints

CUDA unified memory provides a single coherent memory image available to all participating devices (in this case, CPUs and GPUs). At page granularity data is migrated to the memory of the device that has used it most recently. Since data access patterns for graphs are data-dependant, we hope to use this fine-grained on-demand data transfer system as a platform to implicitly partition graph data across GPUs and prevent unnecessary data transfers.

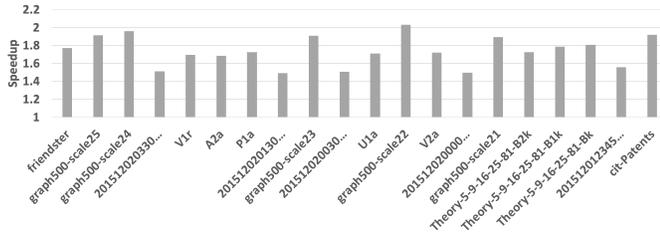


Figure 1. Speedup achieved by overlapping disk I/O with COO+CSR construction, over first reading the entire edge list and then constructing the COO+CSR representation.

Unified memory comes with a large performance cost of handling page faults as GPU threads access non-local pages. When a thread makes an access, the warp is stalled while the page is transferred to the demanding device and the device’s page table is updated. Furthermore, if multiple devices all access the same page, the page may thrash between devices, bringing forward process to a near stand-still. A programmer with some knowledge of application data-access patterns can use hints to the system to affect the behavior of the page migration and reduce the performance costs.

- *cudaMemAdviseSetReadMostly*: a memory region is marked as read-mostly, and devices which read from this region get a read-only copy of the page. This prevents multiple devices that only read data from thrashing.
- *cudaMemPrefetchAsync*: This prefetches a memory region to a device, preventing the device from stalling while during a demand page migration, as the page will already be on the device.

C. COO+CSR Construction and Disk I/O Overlap

Thanks to unified memory, we are able to overlap CSR construction in the GPU address with the disk I/O for reading the data. Two threads are synchronized by a shared double-buffer, which one thread fills from the edge list file and the other consumes to create the COO+CSR in CUDA unified memory. **As such total_time includes the cost of disk I/O; gpu_time does not.** Figure 1 shows the speedup of overlapping these operations, compared with disk I/O alone followed by CSR construction. Both operations take approximately the same amount of time, leading to nearly a 2x speedup. For these measurements, the disk I/O is accelerated by the edge list file contents being in the operating system disk cache, as well as being in a binary format.

D. Triangle-Counting Kernel

We continue to approach the triangle-counting problem in terms of a per-edge count of common elements between source and destination node neighbor lists. In our previous submission, we used a single algorithm regardless of the incoming graph. Like many submissions, we continue to take advantage of the observation that converting the undirected graph into a DAG (directed acyclic graph) can cut the triangle counting work in half. By directing edges along a total ordering of nodes, the graph can be converted to a DAG. In this work, we direct all edges from lower to higher vertex id. Other works

take a different approach, such as [1], [2] who order by degree. We did not take this approach, as we feel it requires a pass over the entire graph before the orientation of each edge can be known.

Algorithm 1 Two-pointer intersection count

Input: a, aSz, b, bSz: sorted arrays a, b and sizes
Output: c: number of common elements in a and b

```

1: procedure THSEQCNT(a, aSz, b, bSz)
2:   c = 0
3:   ap = a; loadA = true
4:   bp = b; loadB = true
5:   while ap < a + aSz and bp < b + bSz do
6:     if loadA then
7:       aVal = *ap; loadA = false
8:     end if
9:     if loadB then
10:      bVal = *bp; loadB = false
11:    end if
12:    if aVal == bVal then
13:      c += 1
14:      ap += 1; loadA = true
15:      bp += 1; loadB = true
16:    else if aVal < bVal then
17:      ap += 1; loadA = true
18:    else
19:      bp += 1; loadB = true
20:    end if
21:  end while
22:  return c
23: end procedure

```

1) *Two-Pointer Intersection Count*: We continue to use a two-pointer intersection count. This approach is similar to the ones described by other finalists. Algorithm 1 shows our current two-pointer intersection count implementation. The expected number of accesses for each thread is $aSz + bSz$, where $aSz = aEnd - aBegin$ and $bSz = bEnd - bBegin$. Compared to our submission last year, this implementation prevents generating a load instruction when a pointer has not moved, and fixes an out-of-bounds access bug.

2) *Binary Search Intersection Count*: In addition to the linear intersection count, we considered a binary-search approach similar to 2017 finalist/2018 champion Hu[1] and 2018 finalist Fox[3], who motivate it in more detail. Each value in one array may be sought in the other array using a binary search. To summarize, the expected number of global memory accesses is $cSz \times \log_2(dSz)$, where cSz is the shorter of the two arrays and dSz is the longer. When the lengths of the arrays are severely mismatched, the binary search will require fewer memory accesses than the two-pointer search. Furthermore, every element of the shorter array can be sought in the longer array in parallel. This yields good cache performance, as adjacent threads’ accesses in the binary search will all be to the same or similar locations.

E. Static Kernel

In our static kernels, we assign each thread to a single edge. Each thread loads the source and destination vertex from the rowInd and colInd arrays (the COO), then traverses the

CSR part of the COO+CSR to determine which entries in the `colInd` array correspond to the neighbor list of each vertex.

The number of common elements (triangles) between the two neighbor lists is counted using the two-pointer sequential intersection. For the linear kernel, Algorithm 1 is used for the “linear” kernel.

The binary kernel uses a single thread, which searches for values from the shorter neighbor list in the longer neighbor list using a binary search. At the end of the kernel execution, each thread block uses a CUB [4] `BLOCKSUM` to aggregate the count into thread 0, which uses a single `ATOMICADD` to contribute its final count to memory.

F. Dynamic Algorithm Selection

Fox et al. [3] presented a logarithmic radix binning approach, where different amounts of parallelism were provided to different edges based on an estimate of their workload. Their results showed that the performance overhead of the binning was non-trivial, so we attempt to capture a similar idea without the associated overhead.

Instead of providing a different amount of parallelism to each edge, we apply a different intersection-count approach to each edge based on the expected number of loads described in Sections II-D1 and II-D2. This should give the benefit of fast $\mathcal{O}(\log_2)$ search through very long rows, or a fast sequential search through shorter rows.

Furthermore, instead of paying the price of binning edges by expected workload ahead of time, we implement a work-stealing approach at the granularity of 32-thread warps. In our static kernels, where the grid size is determined by the number of edges, a single long-running thread could prevent an otherwise idle block from finishing and freeing resources for other blocks to begin their work.

Instead of statically assigning threads to edges, we create enough GPU warps to fully occupy the device, and then allow these persistent warps to work-steal edges from each other until the entire triangle count has been completed. Modern GPUs still broadly follow the lockstep execution model for threads in a warp, so any warp that gets stuck with a long-running thread will still block the other threads in that warp from doing useful work. However, the other warps in the block will be free to claim additional edges to count. This is analogous to the work-stealing runtime used by [5] except at the granularity of edges instead of partitions. We choose groups of 32 edges because if the two-pointer algorithm is selected, the 32 threads in each warp can execute independently on those edges. If the binary algorithm is selected, the warp can collaboratively count triangles without synchronizing the entire block.

Algorithm 2 shows pseudocode for our implementation. `WARPBCAST` is a function that transmits the first parameter from the thread with the lane index of the second parameter to every thread in the warp. `WARPSUM` similarly collaboratively sums the values in the first parameter of every warp and returns the result to the thread with the lane index of the second parameter. `WARPBINCNT` is a warp-collaborative binary search algorithm, where each thread in the warp searches for an

Algorithm 2 Edge-oriented Dynamic Kernel

Input: `sf`: a tuning parameter for the selection heuristic

Input: `ei`: the starting edge index (0)

Output: `c`: the total count

```

1: lx = THREADIDX.X % 32                                     ▷ lane in warp
2: thCnt = 0                                               ▷ this thread's count
3: while true do
4:   if 0 then == lx
5:     wi = ATOMICADD(ei, 32)                               ▷ first edge in 32
6:   end if
7:   wi = WARPBCAST(wi, 0)
8:   if wi > nnz + 32 then
9:     break
10:  end if
11:  src = colInd[rowPtr[src]]
12:  srcSz = colInd[rowPtr[src+1]] - src
13:  dst = colInd[rowPtr[dst]]
14:  dstSz = colInd[rowPtr[dst+1]] - dst
15:  lCost = srcSz + dstSz
16:  lcost = WARPSUM(lcost, 0)
17:  lcost = WARPBCAST(lcost, 0)
18:  if srcSz < dstSz then
19:    bincost = sf × srcSz × log2(dstSz)
20:  else
21:    bincost = sf × dstSz × log2(srcSz)
22:  end if
23:  bincost = WARPSUM(bincost, 0)
24:  bincost = WARPBCAST(bincost, 0)
25:  if lcost < bcost then
26:    thCnt = THSEQCNT(src, srcSz, dst, dstSz)
27:  else
28:    for j = wi; j < iw + 32 and j < nnz; ++j do
29:      eSrc = WARPBCAST(src, j - wi)
30:      eSrcSz = WARPBCAST(srcSz, j - wi)
31:      eDst = WARPBCAST(dst, j - wi)
32:      eDstSz = WARPBCAST(dstSz, j - wi)
33:      thCnt = WARPBINCNT(eSrc, eSrcSz, eDst, eDstSz)
34:    end for
35:  end if
36: end while
37: thCnt = BLOCKSUM(THCNT)
38: if ththreadIdx.X == 0
39:   ATOMICADD(c, thCnt)
40: end if

```

element of the shorter list in the longer list using a binary search. In line 1, each thread computes its lane index within the warp. Lines 4-7 lane 0 uses an atomic operation to work-steal a group of 32 edges and broadcasts the starting edge to every other warp in the group. Lines 8-10 terminates the warp if no thread has a valid edge to work on. Lines 11-14 traverse the CSR data to get the offset and size of the source and destination vertex neighbor lists. Lines 15-25 compute an aggregate cost of using either the linear or binary intersection algorithm for those 32 threads. The binary cost is scaled by a user-tunable `sf` factor. The effects of that factor are explored in Figure 2. Lines 26-36 use the threads in the warp to either count triangles for all 32 edges in parallel with the sequential approach, or collaboratively count triangles using the binary search approach for one edge at a time. Finally, lines 38-41 do a collaborative summation of all of the triangles discovered by this block and contribute them to the total count with an

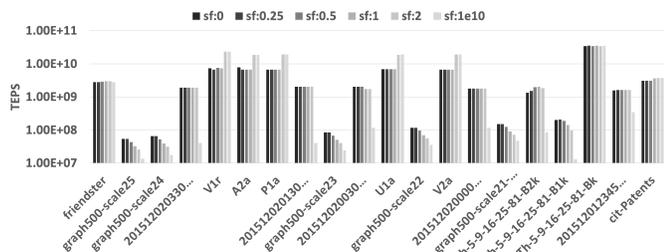


Figure 2. Performance of the dynamic selection kernel on the twenty largest graphs with a varying scale factor in the cost heuristic. A larger scale factor favors the two-pointer intersection algorithm over the binary intersection algorithm.

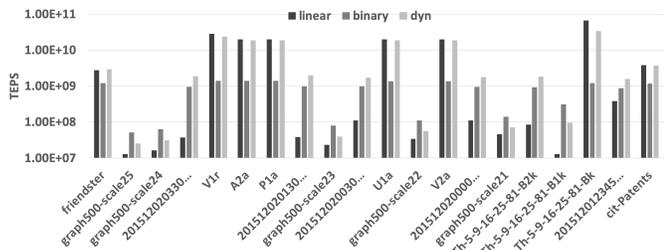


Figure 3. Achieved TEPS of various kernel implementations on the twenty largest graphs with the largest edge count.

ATOMICADD.

Figure 2 shows how the performance of the dynamic kernel varies with the tunable scale factor. The larger the scale factor, the more expensive the binary search is considered and the more likely that the linear approach is selected for any group of 32 edges. We can see that some graphs greatly favor the linear approach, particularly graphs where the average out degree is low and the variance in the out degree is low. The internet topology graphs performance drops sharply for the linear method, because they feature several extremely populous rows, with mostly empty rows. With the linear kernel, an entire warp may be blocked for a long time while a single thread works through the lists. When any binary search is allowed, binary search into the long rows prevents the performance degradation. The `graph500` family of graphs exhibits gradually decreasing performance as the linear approach is favored. This suggests the performance model is not quite able to discriminate successfully in all cases. Based on these results, we select a scale factor of 2 for the final evaluation.

Figure 3 shows the achieved kernel TEPS (traversed edges per second, or edges divided by time), for the twenty largest graphs in the GraphChallenge dataset (by edge count). In most cases, the dynamic kernel matches the performance of either the linear or binary approach alone, without having to make a heuristic decision based on the graph properties before counting begins. This style of approach is promising for translating performance improvements to never-before-seen graphs. Only in the case of the `graph500` family does the dynamic kernel not match the performance of either other approach. This is consistent with the reduced performance for scale factor 2 in those graphs.

TABLE I
IBM AC922 “NEWELL” ARCHITECTURE SUMMARY.

CPU	2 × POWER9
System RAM	512 GB
GPU	4x NVidia V100 (16GB) [7]
CPU-CPU Interconnect	64 GB/s X-bus
NVLink Triad Interconnect	V2.0 x3 (150 GB/s)
CUDA Release	9.2.148
Nvidia Driver	396.44
Linux	4.14.0-49.13.1.el7a.ppc64le
Page Size	64 KB

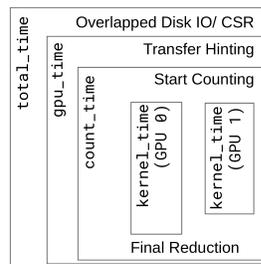


Figure 4. Diagram of what operations are included in measured `total_time`, `gpu_time`, `count_time`, and `kernel_time`. When multiple triangle kernels run in a multi-GPU execution, `kernel_time` is not reported.

G. Multi-GPU Approach

For the multi-GPU approach, we do a simple partition of the edges to each GPU. This causes substantial load imbalance, since our edge orientation approach causes low-index rows to be on average much longer than high-index rows. The CSR structures are marked with the `cudaMemAdviseReadMostly` hint (Section II-B to prevent thrashing when multiple GPUs access the same page (for example, if they read the same row).

III. RESULTS

A. Newell Evaluation Hardware

Triangle counting was evaluated on the same “Newell” platform [6] used in our previous submission. Table I summarizes the hardware and software environment of the evaluation. The systems comprise two triads, CPU0-GPU0-GPU1 and CPU1-GPU2-GPU3. Components within a triads are connected by the listed “triad interconnect,” and the CPUs in each triad are connected by the “CPU-CPU interconnect.”

B. Timing Methodology

Figure 4 summarizes what is included in the four reported times for each graph. In all cases, the clock is stopped once the final triangle count is available on the CPU. In `total_time`, the clock is started after the disk I/O is completed. In `gpu_time`, the clock is started after the CSR is constructed, before unified-memory hints are provided to the system. In `process_time`, the clock is started after unified memory hints have completed, but before counting operations begin. In `kernel_time`, only the actual triangle counting kernel is timed. The `process_time` is the time that includes all triangle counting operations,

but excludes all other operations. Table II summarizes the performance of our approach on all graphs.

C. Twitter Graph

We also evaluate performance on `twitter` [8], a follower-follower relationship graph from 2014. Since the `twitter` dataset is not provided by the Graph Challenge, we convert it to take a similar form.

D. Comparison with Past Submissions

1) *Mailthody et al. (2018 Finalist)*: In our 2018 submission [9], we showed substantial performance degradation in some multi-GPU runs, up to several orders of magnitude in some cases. For the unified-memory case, this was due to unified memory thrashing. For the zero-copy case, zero-copy causes redundant host-to-device transfers because data is not cached in GPU memory. Through the use of unified-memory hints, we have brought the multi-GPU performance up to a more reasonable baseline, where performance is actually improved once the data is in the GPU. In addition, improvements to the sequential kernel code as well as the dynamic algorithmic selection have improved kernel performance.

2) *Bisson & Fatica (2018 Champion)*: Due to differences in procedure, there is no good direct comparison between [10] (B&F) timing and ours. On the surface it seems natural to compare B&F's `total_t` with our `total_time`, but B&F suggest that CSR construction is done on the GPU, as total time starts after data has been copied "to device memory."

An ideal comparison would be between our `gpu_time` with their `process_t` plus matrix compaction. Such a comparison would ignore the differences in CSR construction time. B&F do not provide results cut that way. Since their kernel is typically substantially accelerated by their matrix processing, a direct kernel-only comparison may be interesting but not ultimately useful.

Finally, we designed our approach ultimately with scalability to large graphs in mind, and it contains no steps (other than the triangle counting) which require a full pass over the graph data. B&F require processing the entire graph for compaction, and again for triangle counting. If the graph does not fit in GPU memory, this approach may not be feasible. They reported requiring a GPU with 32GB instead of 16GB of memory in order to process the friendster graph.

3) *Hu, Liu, & Huang (2018 Champion)*: Our `gpu_times` and `count_times` are substantially faster than [11]'s single-GPU times on the larger graphs where they report results, though we evaluate on Nvidia V100s instead of P100s. For the large graphs, they do not report single GPU results, though their multi-GPU times are better than those achieved here.

4) *Yacsar et al. (2018 Champion)*: Cilk gives [12] a work-stealing runtime, an idea analogous to the work-stealing done between warps in our dynamic kernel implementation. Many of their peak rates are similar to ours for larger graphs, though typically once a graph in the GPU memory our rates are much higher.

IV. EXTRA CONTRIBUTIONS TO THE COMMUNITY

Our graph dataset handling code is available at <https://github.com/cwpearson/graph-datasets2>.

Some analysis of the Graph Challenge graph datasets has been made available online at <https://graphchallenge-datasets.netlify.com>.

Our triangle counting code is in the Pangolin library, currently available at <https://github.com/c3sr/pangolin>

V. CONCLUSION

We present an update to our 2018 submission, motivated by understanding whether CUDA unified memory can be used as a foundation for processing large graphs on high-bandwidth heterogeneous systems. We introduce a work-stealing dynamic algorithm GPU kernel with persistent threads, to make performance adaptive for large graphs without requiring pre-counting graph analysis. We utilize CUDA unified memory to allow simultaneous access of large graph data on disk with construction of graph data structures in GPU memory. We utilize CUDA unified memory hints to tackle some performance challenges when scaling to multiple GPUs. Finally, we improve the single-GPU performance from our 2018 submission, and greatly improve the multi-GPU performance. Based on these results, we believe this approach is a solid foundation for scaling performance to large graphs. Moving forward, we intend to look at techniques to improve the multi-GPU load balancing, improving the dynamic algorithm kernels to reflect advances in kernel performance achieved by other teams, and extend our work to graphs that do not fit in GPU memory.

ACKNOWLEDGMENTS

This work is supported by IBM-ILLINOIS Center for Cognitive Computing Systems Research (C3SR) - a research collaboration as part of the IBM AI Horizons Network. This work utilizes resources supported by NSF-MRI program, grant #1725729, and University of Illinois at Urbana-Champaign

REFERENCES

- [1] Y. Hu, P. Kumar, G. Swope, and H. H. Huang, "Trix: Triangle counting at extreme scale," in *2017 IEEE High Performance Extreme Computing Conference (HPEC)*. IEEE, 2017, pp. 1–7.
- [2] R. Pearce, "Triangle counting for scale-free graphs at scale in distributed memory," in *2017 IEEE High Performance Extreme Computing Conference (HPEC)*, Sep. 2017, pp. 1–4.
- [3] J. Fox, O. Green, K. Gabert, X. An, and D. A. Bader, "Fast and adaptive list intersections on the gpu," in *2018 IEEE High Performance extreme Computing Conference (HPEC)*. IEEE, 2018, pp. 1–7.
- [4] CUB library. [Online]. Available: <https://nvlabs.github.io/cub>
- [5] A. Yaar, S. Rajamanickam, M. Wolf, J. Berry, and V. atalyrek, "Fast triangle counting using cilk," in *2018 IEEE High Performance extreme Computing Conference (HPEC)*. IEEE, 2018, pp. 1–7.
- [6] A. B. Caldeira, "IBM power system AC922 introduction and technical overview," *IBM Redbooks*, 2018.
- [7] (2017) NVIDIA Tesla V100 GPU architecture. Nvidia. [Online]. Available: <http://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf>
- [8] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *WWW '10: Proceedings of the 19th international conference on World wide web*. New York, NY, USA: ACM, 2010, pp. 591–600.

TABLE II

TRIANGLE COUNTING PERFORMANCE SUMMARY. TOTAL_TIME INCLUDES THE OVERLAPPED DATA I/O AND CSR CONSTRUCTION. GPU_TIME INCLUDES COST OF TRANSFERRING DATA TO THE GPU. COUNT_TIME INCLUDES THE COST OF ALL COUNTING OPERATIONS. KERNEL_TIME INCLUDES ONLY THE TRIANGLE COUNTING KERNEL TIME.

graph	edges	nodes	tris	Elapsed time (s)			TFPS			2-GPU Speedup			4-GPU Speedup		
				kernel	count	gpu	total	kernel	count	gpu	total	kernel	count	gpu	total
friendster_adj	179999986	119432958	19176	6.11E-01	6.14E-01	1.37E+00	5.72E+01	2.94E+09	2.93E+09	1.31E+09	3.15E+07	1.33	0.97	2.02	0.62
twitter	120251306	41629230	34824916864	2.97E+01	2.97E+01	3.02E+01	7.60E+01	4.03E+07	4.03E+07	3.98E+07	1.38E+07	1.11	1.10	1.78	1.30
graph500-scale25-e16_adj	523464748	17043781	21575375802	2.00E+01	2.06E+01	2.08E+01	3.67E+01	2.54E+07	2.54E+07	2.51E+07	1.43E+07	1.52	1.51	2.75	2.61
graph500-scale24-e16_adj	260261843	8860451	9936161560	8.34E+00	8.34E+00	8.46E+00	1.64E+01	3.12E+07	3.12E+07	3.08E+07	1.59E+07	1.58	1.56	2.87	2.63
201512020350_v226196185_4880407894	240023945	226196186	26	1.20E-01	1.27E-01	2.72E-01	8.09E+00	1.91E+09	1.88E+09	8.85E+08	2.97E+07	1.03	0.88	1.76	0.56
Aza	232704582	214008018	49	9.80E-03	1.17E-02	1.57E-02	7.70E+00	2.38E+10	2.00E+10	1.50E+10	3.13E+07	1.62	0.84	1.53	0.38
PIA	148914992	139353212	3412	7.85E-03	9.61E-03	1.02E-01	5.11E+00	1.90E+10	1.35E+10	1.46E+09	2.98E+07	1.34	0.83	1.76	0.37
201512020130_v128568730_c270232840	153117420	128568731	10	6.71E-02	6.85E-02	1.55E-01	4.86E+00	2.01E+09	1.97E+09	8.73E+08	2.78E+07	1.01	0.92	1.43	0.55
graph500-scale23-e16_adj	19250705	4608135	4500133002	3.23E+00	3.23E+00	3.29E+00	7.39E+01	4.00E+05	3.93E+05	1.37E+07	1.44	1.46	2.52	2.30	
Uta	69389281	67716232	325	3.66E-03	5.11E-03	4.97E-02	2.39E+00	1.89E+10	1.36E+10	1.40E+09	2.90E+07	1.15	0.87	0.73	0.38
graph500-scale22-e16_adj	64097004	2391286	2067392370	1.15E+00	1.15E+00	1.18E+00	3.12E+00	5.59E+07	5.58E+07	5.45E+07	2.03E+07	1.49	1.46	2.58	2.25
graph500-scale22-e16_adj	58668800	55012370	1443	3.06E-03	4.08E-03	4.20E-02	1.90E+00	1.91E+10	1.43E+10	1.39E+09	3.09E+07	0.82	0.80	0.74	0.40
201512020000_v35991442_c744842420	37242710	35991343	2	2.10E-02	2.38E-02	4.75E-02	1.35E+00	1.77E+09	1.64E+09	7.84E+08	2.75E+07	0.97	0.87	1.30	0.57
graph500-scale21-e16_adj	31731650	1243073	935100883	4.38E-01	4.40E-01	4.55E-01	1.58E+00	7.24E+07	7.21E+07	6.98E+07	2.00E+07	1.52	1.48	2.62	2.22
Theory-5-9-16-25-B1k	2554641	6657899	0	2.99E-03	3.01E-03	3.14E-03	1.24E+00	9.51E+07	9.32E+07	9.13E+07	2.31E+07	1.11	1.14	1.74	1.55
Theory-5-9-16-25-B1k	28667380	2176441	155	1.36E-02	1.51E-02	2.88E-02	9.46E-01	2.10E+09	1.90E+09	9.97E+08	3.03E+07	0.96	0.91	0.97	0.58
Theory-5-9-16-25-B1k	23338000	2176441	0	6.67E-04	2.39E-03	1.35E-02	7.64E-01	3.50E+10	9.78E+09	1.73E+09	3.05E+07	0.90	0.87	0.36	0.42
201512012345_v18571154_c83040230	19030160	18571155	2	1.19E-02	1.34E-02	2.71E-02	6.92E-01	1.60E+09	1.42E+09	7.01E+08	2.75E+07	1.00	0.88	1.02	0.38
cit-Patents_adj	3747490	7515023	193	4.41E-03	6.11E-03	1.51E-02	5.29E-01	3.73E+09	2.76E+09	1.09E+09	3.06E+07	1.13	0.97	0.88	0.54
graph500-scale19-e16_adj	15680861	645821	419349784	1.69E-01	1.71E-01	1.78E-01	6.78E-01	9.28E+07	9.19E+07	8.80E+07	2.31E+07	1.41	1.36	2.37	1.96
Theory-3-4-5-9-16-25-B1k	11080030	530401	35822427	9.61E-02	9.79E-02	1.04E-01	4.69E-01	1.13E+08	1.06E+08	2.36E+05	1.05	1.05	1.16	1.05	
Theory-3-4-5-9-16-25-B1k	11080030	530401	651	6.99E-03	8.37E-03	1.49E-02	3.75E-01	1.59E+09	1.32E+09	4.71E+08	2.90E+07	0.87	0.91	0.73	0.58
graph500-scale19-e16_adj	722897	339119	186288972	6.52E-02	6.69E-02	7.82E-02	3.16E-03	1.13E+08	1.13E+08	1.08E+08	2.85E+07	1.36	1.11	0.78	0.70
Theory-3-4-5-9-16-25-B1k	6912000	530401	0	2.06E-04	1.90E-03	6.10E-03	2.28E-01	3.36E+10	3.64E+09	1.13E+09	3.03E+07	0.86	0.86	0.29	0.38
graph500-scale18-e16_adj	880348	174148	82287285	2.04E-02	2.21E-02	2.56E-02	1.48E-01	1.86E+08	1.72E+08	1.48E+08	2.57E+07	1.30	1.13	1.46	1.26
roadNet-CA_adj	2766607	1965207	120676	1.21E-04	1.73E-03	5.21E-03	9.55E-02	2.28E+10	1.86E+09	5.31E+08	2.90E+07	0.53	0.84	0.51	0.62
Theory-9-16-25-81-Bk	2606125	362441	0	6.75E-04	6.11E-03	2.46E-02	3.46E-01	1.74E+08	1.74E+08	1.73E+08	4.84E+07	1.13	0.87	0.54	0.44
Theory-9-16-25-81-Bk	2606125	362441	35	6.75E-04	2.23E-03	4.11E-03	9.34E-02	3.86E+09	1.17E+09	6.35E+08	2.79E+07	0.59	0.64	0.34	0.38
amazon0101_adj	2443408	403395	3986507	6.97E-04	2.26E-03	4.10E-03	8.79E-02	3.50E+09	1.08E+09	5.97E+08	2.78E+07	0.72	0.66	0.42	0.42
amazon0105_adj	2439347	410237	3951063	5.62E-04	2.14E-03	5.06E-03	8.84E-02	4.34E+09	1.14E+09	4.82E+08	2.76E+07	0.54	0.72	0.60	0.62
amazon0112_adj	2348966	407228	4218459	6.24E-04	7.80E-04	4.13E-03	7.47E-02	3.49E+09	1.09E+09	4.97E+08	2.76E+07	0.76	0.71	0.49	0.44
Theory-9-16-25-81-Bk	23338000	362441	0	8.90E-05	1.65E-03	4.15E-03	8.56E-02	2.60E+10	1.41E+09	3.62E+08	2.73E+07	0.46	0.60	0.25	0.38
RickEgges_adj	2314948	105939	107987357	3.81E-03	4.77E-03	6.32E-03	8.66E-02	6.08E+08	4.80E+08	3.55E+08	2.67E+07	0.89	0.78	0.67	0.59
Theory-25-81-256-Bk	2132284	547925	2102761	5.43E-03	6.11E-03	7.82E-03	1.03E-01	9.32E+07	8.88E+07	8.09E+07	2.16E+07	0.91	0.89	0.78	0.70
Theory-25-81-256-Bk	2132284	547925	0	6.82E-04	2.26E-03	5.16E-03	8.14E-02	3.31E+09	9.42E+08	4.13E+08	2.62E+07	0.67	0.88	0.62	0.70
Theory-25-81-256-Bk	2073600	547925	0	7.73E-05	1.63E-03	4.43E-03	7.47E-02	2.68E+10	1.27E+09	4.68E+08	2.78E+07	0.93	1.06	0.48	0.59
roadNet-TX_adj	1921660	1739918	82869	9.43E-05	8.61E-04	2.68E-03	6.50E-02	2.04E+10	2.23E+09	7.18E+08	2.90E+07	0.28	0.42	0.25	0.29
Theory-4-5-9-16-25-Bk	1832861	132601	1254643	7.93E-03	9.53E-03	1.21E-02	8.83E-02	2.66E+08	2.24E+08	1.76E+08	2.31E+07	1.08	0.88	0.54	0.44
Theory-4-5-9-16-25-Bk	1582861	132601	155	1.73E-03	2.50E-03	4.10E-03	5.85E-02	9.16E+08	6.32E+08	3.86E+08	2.71E+07	0.45	0.55	0.31	0.36
roadNet-PA_adj	1541898	1088093	67150	7.39E-05	1.62E-03	3.33E-03	5.58E-02	2.09E+10	9.53E+08	4.62E+08	2.76E+07	0.66	0.66	0.49	0.46
Theory-4-5-9-16-25-Bk	1152000	132601	0	4.60E-05	1.58E-03	3.72E-03	4.50E-02	2.51E+10	7.29E+08	3.10E+08	2.50E+07	0.52	0.71	0.24	0.29
loc-brightkite_edges_adj	755737	26521	1829927	1.43E-03	2.21E-03	4.22E-03	3.37E-02	1.43E+09	8.20E+08	4.20E+08	2.46E+07	0.84	0.83	0.75	0.65
amazon0102_adj	899992	262172	171719	1.16E-04	1.66E-03	3.94E-03	3.60E-02	7.78E+09	5.43E+08	2.38E+08	2.50E+07	0.67	0.86	0.50	0.65
soc-Slashdot0902_adj	504230	82169	602592	6.61E-04	1.80E-03	2.59E-03	2.26E-02	7.63E+08	2.30E+08	1.94E+08	2.24E+07	0.87	0.65	0.52	0.45
soc-Slashdot0111_adj	469180	77561	551724	6.30E-04	1.75E-03	2.98E-03	2.24E-02	7.44E+08	2.69E+08	1.57E+08	2.10E+07	0.84	0.95	0.51	0.47
cit-HepTh_adj	428077	34547	235840	3.58E-04	7.86E-04	1.44E-03	3.58E-02	1.80E+09	1.18E+09	8.14E+08	2.82E+07	1.13	0.97	0.54	0.44
soc-Epinions1_adj	405740	75880	1624481	1.44E-03	3.20E-03	1.98E-02	5.72E+08	2.81E+08	1.27E+08	2.04E+07	0.44	0.69	0.42	0.57	
email-EuAll_adj	364481	265215	267131	1.17E-03	2.24E-03	4.17E-03	1.95E-02	3.14E+08	1.63E+08	8.75E+07	1.86E+07	0.58	0.80	0.37	0.67
cit-HepTh_adj	352285	27771	1748735	4.74E-04	1.95E-03	3.62E-03	1.87E-02	7.38E+08	1.81E+08	9.72E+07	1.88E+07	1.05	0.92	0.56	0.65
Theory-256-625-B1k	160883	160883	16000	9.43E-04	1.46E-03	3.69E-03	6.68E-02	3.69E+08	3.69E+08	3.69E+08	1.81E+07	0.69	0.69	0.40	0.31
Theory-256-625-B1k	230881	160883	1	1.56E-04	1.25E-03	2.74E-03	1.78E-02	2.06E+09	2.56E+08	1.17E+08	1.81E+07	0.69	0.69	0.40	0.31
Theory-256-625-B1k	320000	160883	0	2.34E-05	1.53E-03	2.73E-03	1.63E-02	1.37E+10	2.09E+08	1.57E+08	1.90E+07	0.43	0.44	0.48	0.37
Theory-9-16-25-81-Bk	217255	20401	2146227	3.09E-04	1.18E-03	2.38E-03	1.31E-02	7.03E+08	1.84E+08	9.12E+07	1.66E+07	0.42	0.64	0.35	0.44
Theory-9-16-25-81-Bk	217255	20401	155	3.09E-04	1.18E-03	2.38E-03	1.31E-02	7.03E+08	1.84E+08	9.12E+07	1.66E+07	0.42	0.64	0.35	0.44
loc-brightkite_edges_adj	214078	58229	404728	2.37E-04	1.76E-03	2.88E-03	1.31E-02	9.03E+08	1.22E+08	7.43E+07	1.63E+07	0.58	0.57	0.54	0.53
email-Eronn_adj	183831	36693	372994	3.94E-04	1.89E-03	3.05E-03	1.25E-02	5.05E+08	9.71E+07	6.02E+07	1.47E+07	0.88	0.99	0.58	0.56
Theory-5-9-16-25-Bk	175873	26521	1254643	7.25E-03	8.55E-03	1.03E-02	7.80E-02	1.46E+09	1.						

- [9] V. S. Malthody, K. Date, Z. Qureshi, C. Pearson, R. Nagi, J. Xiong, and W.-m. Hwu, "Collaborative (cpu+ gpu) algorithms for triangle counting and truss decomposition," in *2018 IEEE High Performance extreme Computing Conference (HPEC)*. IEEE, 2018, pp. 1–7.
- [10] M. Bisson and M. Fatica, "Update on static graph challenge on gpu," in *2018 IEEE High Performance extreme Computing Conference (HPEC)*. IEEE, 2018, pp. 1–8.
- [11] Y. Hu, H. Liu, and H. H. Huang, "High-performance triangle counting on gpus," in *2018 IEEE High Performance extreme Computing Conference (HPEC)*. IEEE, 2018, pp. 1–5.
- [12] A. Yaşar, S. Rajamanickam, M. Wolf, J. Berry, and Ü. V. Çatalyürek, "Fast triangle counting using cilk," in *2018 IEEE High Performance extreme Computing Conference (HPEC)*. IEEE, 2018, pp. 1–7.