Taming the Zoo: The Unified GraphIt Compiler Framework for Novel Architectures

Ajay Brahmakshatriya
MIT CSAIL
ajaybr@mit.edu

Emily Furst
University of Washington
eafurst@cs.washington.edu

Victor A. Ying
MIT CSAIL
victory@csail.mit.edu

Claire Hsu
MIT CSAIL
clhsu@mit.edu

Changwan Hong
MIT CSAIL
changwan@mit.edu

Max Ruttenberg
University of Washington
mrrt@cs.washington.edu

Yuning Zhang
MIT CSAIL
yunning@mit.edu

Dai Cheol Jung
University of Washington
dcjung@uw.edu

Dustin Richmond
University of Washington
dustinar@uw.edu

Michael B. Taylor
University of Washington
prof.taylor@gmail.com

Julian Shun
MIT CSAIL
jshun@mit.edu

Mark Oskin
University of Washington
oskin@cs.washington.edu

Daniel Sanchez
MIT CSAIL
sanchez@csail.mit.edu

Saman Amarasinghe
MIT CSAIL
saman@csail.mit.edu

Abstract—We live in a new Cambrian Explosion of hardware devices. The end of conventional processor scaling has driven research and industry practice to explore a new generation of approaches. The old DNA of architecture design, including vectors, threads, shared or private memories, coherence or message passing, dataflow or von Neumann execution, are hybridized together in new and exciting ways. Each new architecture exposes a unique hardware-level API. Performance and energy efficiency are critically dependent on how well programs can use these APIs. One approach is to implement custom libraries for each new hardware architecture and application domain. A more scalable approach is to utilize a portable compiler infrastructure tailored to the application domain that makes it easy to generate efficient code for a diverse set of architectures with minimal porting effort.

We propose the Unified GraphIt Compiler framework (UGC), which does exactly this for graph applications. UGC achieves portability with reasonable effort by decoupling the architecture-independent algorithm from the architecture-specific schedules and backends. We introduce a new domain-specific intermediate representation, GraphIR, that is key to this decoupling. GraphIR encodes high-level algorithm and optimization information needed for hardware-specific code generation, making it easy to develop different backends (GraphVMs) for diverse architectures, including CPUs, GPUs, and next-generation hardware such as Swarm and the HammerBlade manycore. We also build scheduling language extensions that make it easy to expose optimization decisions like load balancing strategies, blocking for locality, and other data structure choices. We evaluate UGC on five algorithms and 10 input graphs on these 4 distinct architectures and show that UGC enables implementing optimizations that can provide up to 53x speedup over programmer-generated straightforward implementations.

Index Terms—Compilers for Novel Architectures, Domain-Specific Languages, Graphs, Intermediate Representations

I. INTRODUCTION

As we enter the twilight of Moore’s Law, architectural diversity is rapidly exploding. New designs from generic parallel substrates such as manycores and dataflow engines, to highly domain-specific engines such as machine learning and graph accelerators are being researched and commercially deployed. Many of these architectures are programmable and combine well-understood techniques such as vectorization, threading, and explicit data movement in novel ways. Oftentimes, the difference between lackluster performance and dramatic speedup hinges on correctly using the particular combination of features an architecture provides. Yet it is a daunting task to do so for the wide range of hardware architectures and application domains targeted by general-purpose systems.

In this paper, we advocate for a novel compiler and software stack that can support this explosion in architectural diversity. We pursue a domain-specific approach, focusing on graph analytics, to enable the compiler to capture programmer intent and produce optimized implementations. We present a compiler toolchain, the Unified GraphIt Compiler framework (UGC), that targets diverse architectures while making it easy to write and compose optimizations that make use of each architecture’s unique features. Recent work has developed domain-specific toolchains for deep learning and image processing [22, 70] that target CPUs, GPUs, and accelerators, showing the potential of this approach. But graphs, due to their irregularity, present a unique set of challenges for both hardware and software.

Graph processing is a crucial application domain that can benefit from hardware acceleration [6, 23, 34, 43, 87]. Graph algorithms are at the heart of many applications [16, 29, 68, 75, 90], but are notoriously difficult to optimize [59, 95]. Graph programs exhibit irregular memory access patterns that often
A compiler framework with a novel and carefully designed evaluation of code generated by the four GraphVMs, which implementations of four GraphVMs that can generate efficient code for CPUs, GPUs, Swarm, and the HammerBlade Manycore. This framework allows programmers to explore the optimization spaces of different hardware platforms.

This paper makes the following contributions:

- A compiler framework with a novel and carefully designed intermediate representation, GraphIR (Section III-B); hardware-independent passes; and hardware-specific GraphVMs to generate fast code on diverse architectures.
- A novel extensible scheduling language (Section III-D) that allows programmers to explore the optimization spaces of different hardware platforms.
- Implementations of four GraphVMs that can generate efficient code for CPUs, GPUs, Swarm, and the HammerBlade Manycore (Section III-C).
- Evaluation of code generated by the four GraphVMs, which shows up to 53× speedup over user-supplied baseline code.

This paper also provides insights on techniques for building portable compilers targeting very different architectures for a specific application domain.

II. BACKGROUND

The performance of graph processing depends on optimizations to mine for locality within sparse data structures, to minimize high-cost memory accesses and synchronization, and to balance load across parallel threads. Unfortunately, the structure of graphs varies widely, as does the work across iterations of an algorithm. As a result, the optimal approach can change not only across graphs, but also across iterations. This makes graph programs notoriously difficult to optimize on any architecture.

To make matters worse, modern architectures employ a wide range of hardware features to exploit parallelism and achieve high throughput: threads, vectors and warps, tasks, task or instruction speculation, memory consistency models, cache coherence, variants of atomic operations, data movement engines, and scratchpads (to name a few). Combinations of these features produce exponentially many architectural variations, each with different performance characteristics. Each architecture has a different low-level language, compiler, and runtime that exposes these features, and implementations must be cognizant of these features and their implications.

Domain-specific languages (DSLs) for graph processing abstract the complexity of modern architectures and the dynamic challenges of graph structure. DSLs have been used to abstract hardware in domains like machine learning, image processing, networking, tensor algebra, and bioinformatics, or sometimes combining a few domains. An ideal DSL for graph processing would facilitate algorithm expression and abstract away architectural details to provide good performance on a wide range of applications and architectures.

A. GraphIt Domain-Specific Language (DSL)

GraphIt is a domain-specific language for graph applications that decouples the algorithm specification and computation schedule. This enables GraphIt to generate high-performance code with optimizations tailored for diverse graph inputs from a single portable algorithm specification.

```
1  func toFilter(v : Vertex) -> output : bool
2  output = (parent[v] == -1);
3  end
4  func updateEdge(src : Vertex, dst : Vertex)
5  toFilter(v : Vertex) -> output : bool
6  output = (parent[v] == -1);
7  end
8  func main()
9  end
10  while(frontier.getVertexSetSize() != 0)
11     applyModified(updateEdge, parent, true);
12     delete frontier;
13   end
```

Fig. 2: Algorithm specification for Breadth-First Search (BFS) in GraphIt.
calls the \texttt{edgeset.applyModified} operator, which uses these functions to specify which edges are to be processed and what computation is to be performed on each edge. The algorithm specification does not specify the loop nests or the iteration order; this is specified in the schedule. This separation makes it possible to generate different implementations suitable to the algorithm and graph input.

UGC uses exactly the same algorithm language as GraphIt, enabling us to reuse the source code written for various applications. The high-level design of the operators also makes it easy for UGC to target very different architectures. For example, the \texttt{edgeset.applyModified} operator can be easily mapped to architectures that have specialized units for traversing edges in parallel. We extend the scheduling language to fit the optimizations of different architectures. Examples are shown in Figure 6.

**B. Parallel Architectures**

In this work, we target the four parallel architectures shown in Table I. These architectures are built with a diverse set of hardware features that expose different forms of parallelism, latency hiding techniques, and synchronization. These architectures require significantly different optimization strategies and pose unique challenges for UGC. We briefly explain each architecture below, and the challenges that they present when compiling graph programs.

1) **Multicore CPU**

A multicore CPU has cores optimized for single-thread performance, with prefetching and a multi-level cache hierarchy. To hide latency, each core supports speculative out-of-order execution as well as simultaneous multithreading. CPU software exposes explicit parallelism through threads provided by a multithreaded runtime [13, 19]. CPUs perform well on graph applications that provide limited parallelism, high locality, or predictable memory access patterns [11]. The large memory capacity also means that CPUs outperform other systems on multi-terabyte graphs [28].

2) **GPU**

GPUs provide massive parallelism through a SIMT programming (SIMD execution) model, where arithmetic units are vectorized and use predication to handle divergent control flow. GPUs use multithreading with many hardware thread contexts to hide memory latency. GPUs are suitable for graph applications with massive parallelism that exhibit regularity in graph structure, memory access pattern, and limited control flow. Applications such as PageRank [66] or less-sparse graph workloads that map well to existing linear algebra libraries [45, 89] can exploit massive memory bandwidth with coalesced memory accesses. GPUs perform poorly on applications that suffer from control divergence [88], load-imbalance [77], or too many sparse memory accesses [1]. Finally, GPUs function best on graphs that fit in device global memory.

3) **Swarm**

Swarm augments a CPU with support for fine-grained task parallelism [43]. Swarm can achieve order of magnitude

<table>
<thead>
<tr>
<th>Hardware features</th>
<th>CPU</th>
<th>Swarm</th>
<th>GPU</th>
<th>HammerBlade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parallel Execution Model</td>
<td>Threads</td>
<td>Ordered Tasks</td>
<td>SIMT / SIMD</td>
<td>SPMD</td>
</tr>
<tr>
<td>Number of Processors</td>
<td>~ 100</td>
<td>100</td>
<td>~ 100K (Threads)</td>
<td>~ 1000</td>
</tr>
<tr>
<td>Speculation</td>
<td>Instruction-Level</td>
<td>Instruction- &amp; Task-Level</td>
<td>No</td>
<td>Non-Blocking Memory Ops.</td>
</tr>
<tr>
<td>Memory Latency Hiding</td>
<td>OoO execution, SMT</td>
<td>OoO execution, SMT</td>
<td>Multithreading</td>
<td>Atomics</td>
</tr>
<tr>
<td>Synchronization support</td>
<td>Atomics</td>
<td>Coherence-Enforced Ordering</td>
<td>~ 40</td>
<td>Atomics</td>
</tr>
<tr>
<td>Addressable Memory (GB)</td>
<td>~ 1000</td>
<td>~ 1000</td>
<td>L2</td>
<td>~ 50</td>
</tr>
<tr>
<td>Shared on-chip Memory</td>
<td>Coherent L3</td>
<td>Coherent L3</td>
<td>L1 (No Coherence)</td>
<td>Globally-Partitioned LLC</td>
</tr>
<tr>
<td>Core-local data Memory</td>
<td>Coherent L1, L2</td>
<td>~ 1MB</td>
<td>Scratchpad (SW Coherence)</td>
<td>4KB</td>
</tr>
<tr>
<td>On-Chip Storage per Thread</td>
<td>~ 1MB</td>
<td>~ 1MB</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**TABLE I: Summary of parallel architectures studied in this paper. Two of the four architectures are shown in Figure 3.**
improvements in scalability over conventional CPUs and GPUs on some graph algorithms by using dedicated hardware task queues and speculative execution to distribute tasks across hundreds of cores [42, 44, 81].

Swarm’s execution model uses order as the main synchronization primitive. Swarm programs consist of tasks. Each task can read and write arbitrary memory and spawn child tasks. Each task is given a timestamp when it is spawned, which must be greater than or equal to its parent’s timestamp. Swarm guarantees that tasks appear to run atomically and in timestamp order, hiding the effects of concurrency from software. Under the hood, Swarm hardware executes tasks in parallel and out of order. To preserve ordered semantics, tasks execute speculatively and the coherence protocol is extended to detect order violations. Upon a violation, the offending task is aborted and re-executed. Fig. 3a shows how Swarm adds a task unit near the cores of each chip tile. These distributed task units perform asynchronous, high-throughput task dispatch and task commit, efficiently supporting tasks as short as a few instructions. These tiny tasks can be selectively aborted or serialized with compiler-generated hints, exposing unique tradeoffs as optimizations must balance task overheads, parallelism, and the costs of aborts and re-executions.

This execution model is a natural fit for priority-based or iterative algorithms, where each task can be assigned a timestamp based on its priority or iteration number. Swarm’s speculative execution uncovers more parallelism than CPUs and GPUs by executing tasks with different timestamps in parallel. Swarm can be programmed in C++ using the T4 compiler [91], but the key challenge is in appropriately dividing the computation into tiny tasks to exploit parallelism and minimize abort costs.

4) HammerBlade Manycore

Manycore architectures provide thread-level parallelism and flexibility with hundreds to thousands of general-purpose cores [5, 25, 33, 71, 84]. We target the HammerBlade Manycore with hundreds of independent cores. The cores have a scalar pipeline, low-latency software-managed scratchpad memory, and support integer, floating-point, and atomic instructions. The cores communicate over the memory-mapped 2-D mesh Network-on-Chip. Cores can issue non-blocking loads to hide latency, to exploit thread-level and memory level parallelism, and to balance work between independent threads of execution.

III. Compiler Design

Choosing the right abstraction for the intermediate representation is critical to simplify code generation for diverse architectures and to expose optimization opportunities. To achieve these goals, we designed GraphIR, a novel intermediate representation. We studied the features of varied architectures to identify the right level of abstraction with enough expressiveness to capture algorithmic details from the graph domain. For example, instead of low-level loop nests, GraphIR has operators for iterating over a set of vertices, or edges incident to a set of vertices. These operators can be directly mapped to thread hierarchies or manycore tiles on architectures such as GPUs and the HammerBlade Manycore, without needing to lift computations from loop nests. GraphIR also avoids making assumptions about the concrete representation of data structures. This allows different architectures to choose various implementations for vertex sets depending on the available memory, bandwidth, and other tradeoffs. For Swarm, the compiler can even eliminate the use of software work queues for vertex sets, instead mapping the operations on vertex sets to hardware tasks. This section explains how GraphIR enables building GraphVMs with these specialized optimizations.

A. Hardware-Independent Transformations

GraphIR has a dual goal of offering flexibility while allowing for maximum code reuse. Even though GraphIR’s main goal is to support specialization and optimizations unique to each hardware backend, a large part of the compiler infrastructure, including analysis and transformation passes, is target-agnostic. Specifically, UGC adapts the domain-specific transformations from the GraphIt DSL compiler [15, 93, 95], such as dependence analysis to insert atomics in the user-defined functions (UDFs), liveness analysis to find frontier memory reuse opportunities,
and other transformations to UDFs for traversal direction, parallelism, and data structure choices. These hardware-independent transformations and analyses are performed on the GraphIR before it is passed to the GraphVMs for code generation. These passes can access scheduling language inputs (Section III-D). These passes also add metadata to the GraphIR for the GraphVMs to use during code generation. Section III-C shows how the bulk of the frontend and the hardware-independent compiler are reused by all four GraphVMs that we implement.

B. GraphIR

One of the main contributions of this paper is the GraphIR intermediate representation that decouples the algorithm specification and hardware-independent optimizations from hardware-specific optimizations. Like LLVM IR, GraphIR is an in-memory representation of a program that allows optimizations through IR-to-IR transformations before final code generation. This design enables us to build reusable program analyses, transformations, and lowering passes shared across different hardware platforms, reducing the effort needed to support a new backend (GraphVM) in UGC. However, unlike LLVM IR, GraphIR uses a high-level domain-specific representation that facilitates more powerful and flexible optimizations.

GraphIR is composed of variables, functions, and instructions. Each variable, function, or instruction carries both arguments and metadata, as shown in Table II. Arguments capture all of the information derived from the algorithm specification and is required for correctness of the generated code. Metadata captures information related to the performance optimizations, and hardware backends can choose to ignore these or add new ones specific to their hardware. GraphIR’s metadata can be manipulated with an API that includes two functions: setMetadata<T>(std::string label, T val) and T getMetadata<T>(std::string label), where T is any C++ type (including other GraphIR nodes). Because this API allows arbitrarily many string labels, metadata can easily stack without having to change GraphIR base class definitions. This metadata API is the primary way in which GraphVMs extend GraphIR nodes for hardware-specific optimizations.

To perform hardware-specific transformations and code generation, each backend implements an abstract machine (GraphVM) to optimize and run GraphIR, similar to the Java VM or LLVM. Section III-C provides details on GraphVMs.

Operators and data types are designed in an implementation-agnostic way to make it easy for the GraphVM developer to pick the right data structure and choice of mapping computations to various hardware units. The two most important instructions in GraphIR are the EdgeSetIterator and VertexSetIterator instructions, shown in Table II. EdgeSetIterator iterates through all or a subset of the edges of a graph and invokes a function on each edge. The arguments of EdgeSetIterator specify the graph (input_graph), input frontier vertexset (input_vset), output frontier vertexset (output_vset), and the user-defined function that works on the edges (apply_f). These arguments are derived from the operators in the algorithm specification. The instruction also has metadata to generate optimized implementations, such as choosing the input/output frontier representations, edge traversal direction, deduplicating the output frontiers, or generating specialized code if the edge set representation is dense. VertexSetIterator iterates over the vertices in a frontier, and similarly has arguments and metadata for optimizations. Apart from these key instructions, GraphIR has instructions for data structure allocation both on the host and on the device, general arithmetic and reductions, and program control flow.

Architectures with these features make use of the metadata attached to the instructions to implement various optimizations. For example, GPUs, which have a hierarchy of threads, can implement different load-balancing strategies to efficiently process vertices with varying degrees. CPUs and GPUs both have multiple levels of memory, which enables blocking of edges for better cache utilization.

Figure 4 shows the pretty-printed GraphIR for the BFS algorithm input from Figure 2. Table II explains each of the GraphIR operators and types used in this example. Line 11 shows the key EdgeSetIterator GraphIR node. This node contains arguments such as the graph to iterate on, the input and output frontiers, the function to apply on each edge, and the source and destination filters. This operator also has metadata attached to it (shown in <>). For example, the can_reuse_frontiers is the result of the frontier reuse analysis pass. As shown in Table III, the result of this analysis is used by the GPU, Swarm, and HammerBlade Manycore GraphVMs. The EnqueueVertex node is another GraphIR node that has metadata, in this case for the representation of the frontier to enqueue to (Line 5). The code in Figure 4 is just the pretty-printed version of the in-memory GraphIR data structure.

The BFS example also shows the updateEdge user-defined function (UDF) that EdgeSetIterator applies to each edge. Line 3 shows that the high-level compiler inserted
The Graph Virtual Machine (or GraphVM) is an abstract machine that executes the target-independent GraphIR. Each backend developer implements a GraphVM tailored to their specific pass to indicate that the schedule has prescribed fusing all of the operator calls inside the loop into a single kernel launch, as Section III-C2 will explain. The GPU GraphVM extends the metadata of WhileLoopStmt with a needs_fusion flag, and sets it to true in a hardware-specific pass to indicate that the schedule has prescribed fusing all of the operator calls inside the loop into a single kernel.

The right level of abstraction and the support for extending GraphIR with metadata makes GraphIR an ideal representation for accommodating hardware-specific optimizations in UGC.

C. GraphVM

The Graph Virtual Machine (or GraphVM) is an abstract machine that executes the target-independent GraphIR. Each backend developer implements a GraphVM tailored to their architecture that includes hardware-specific passes and code generation. The UGC framework provides all of the required tools to build diverse optimization passes including APIs to access GraphIR nodes and scheduling objects attached to them, a set of reusable passes that can be enabled depending on whether the hardware benefits from it, and common routines to aid code generation. GraphVMs for different architectures can be very diverse. Each GraphVM developer can implement it as an interpreter that directly consumes and executes GraphIR or as a combination of transformation and code generation passes. The developer can also choose to move complexity between the generated code or the runtime library, as we discuss next.

As Figure 1 shows, a typical GraphVM has the following parts:
- Hardware-dependent analyses and transformation on GraphIR using hardware-specific scheduling information.
- Code generation for the target device and host (if applicable).
- Runtime library and backend compiler infrastructure to execute the generated code.

As shown in Table III, UGC provides a library of analysis and transformation passes that GraphVMs reuse or specialize, easing the development of new backends. We now discuss our our GraphVMs and their hardware-specific optimizations.
Kernel fusion is an important optimization for road graphs because it amortizes kernel launch overheads for applications where there is very little work in each iteration [67]. The kernel fusion optimization is implemented entirely in the GPU GraphVM as a series of passes. A preliminary pass identifies the loops to be fused and all of the variables that the body of the loop uses from the main function. The first pass in the code generation then generates the actual __global__ kernel to be launched on the GPU. The code generation pass inserts appropriate CUDA API calls to copy state between the host and device. Since the fused kernel has a fixed number of threads, the code generator also generates some outer loops to simulate the work of more threads and inserts __grid_sync() calls for synchronization. Finally, a pass generates appropriate calls to launch a single GPU kernel instead of a separate kernel for each step in each iteration of the loop. Table III shows how this GPU-specific pass is a very small fraction of the total lines of code. This demonstrates that the design choices in GraphIR and GraphVMs significantly reduce the effort required to support unique hardware optimizations.

The GPU GraphVM also implements other optimizations, such as EdgeBlocking and fused vs. unfused frontier creation.

3) Swarm GraphVM

The Swarm architecture relies on speculative execution of tasks to extract parallelism and make applications scale to a large number of cores. Tasks are executed out of order but are aborted when memory dependencies are violated, thus ensuring correctness. However, repeated aborts are undesirable because they result in wasted work. Therefore, the Swarm GraphVM focuses a great deal on eliminating false dependencies between memory accesses. Figure 5 shows the code generated by the Swarm GraphVM for the BFS algorithm.

### Table III: Lines of code for modules of UGC

<table>
<thead>
<tr>
<th>Module</th>
<th>Base version</th>
<th>CPU</th>
<th>GPU</th>
<th>Swarm</th>
<th>HammerBlade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frontend</td>
<td>Algorithm parser &amp; AST definitions</td>
<td>10,900</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Scheduled language functions</td>
<td>136</td>
<td>306</td>
<td>383</td>
<td>524</td>
</tr>
<tr>
<td>Hardware-Independent Compiler</td>
<td>Frontier Reuse Analysis</td>
<td>125</td>
<td>Not used</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Property analysis/Atomic insertion</td>
<td>5.56</td>
<td>0</td>
<td>0</td>
<td>Not used</td>
</tr>
<tr>
<td></td>
<td>Ordered Processing Lowering</td>
<td>286</td>
<td>0</td>
<td>120</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Other Lowering Passes</td>
<td>4,171</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GraphVM</td>
<td>Ordered Processing Specialization</td>
<td>104</td>
<td>0</td>
<td>Not used</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Kernel Fusion</td>
<td>276</td>
<td>Not used</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Code Generator</td>
<td>-</td>
<td>3,843</td>
<td>1,874</td>
<td>959</td>
</tr>
<tr>
<td></td>
<td>Runtime Library</td>
<td>-</td>
<td>10,385</td>
<td>2,470</td>
<td>156</td>
</tr>
</tbody>
</table>

From vertex sets to tasks. One of GraphIR’s main data types is the VertexSet, which holds the current set of active vertices. This active set is read from on every round and written to for the next round. Storing the active vertex set in memory introduces data dependencies (e.g., reuse of memory across rounds, or between updates to in-memory tail pointers or size variables) that prevent Swarm from obtaining more parallelism by speculating across many rounds. These data dependencies are spurious, because the insertion of distinct vertices should actually be independent. We solve this problem through a pass in the Swarm GraphVM that replaces the enqueuing of a vertex ID to a VertexSet with a task spawn. The body of this task is the operation that we would perform with the vertex after dequeuing it. The timestamp of the task is set based on the round in which the vertex would be dequeued. This way, while Swarm’s execution model guarantees tasks from one round appear to execute before any task from the next round, tasks from different rounds can execute speculatively in parallel without false dependencies arising from storing the VertexSet in memory.

Figure 5 shows a lambda passed to `for_each_prio` to indicate what action should be taken per element in the frontier. The body of the lambda calls `push` to spawn tasks that will execute the lambda on vertices at later timestamps.
This approach generalizes to priority-based algorithms like $\Delta$-stepping as well, where task timestamps are set based on priorities.

**From shared to private state.** Some applications have shared variables that are updated periodically. For example, in the forward pass of BC there is a variable updated once per round to track the region of the output data structure that visited vertices are recorded to. If all parallel tasks access this single variable, data dependencies on the updates to this variable would prevent speculation across rounds. To address this, the Swarm GraphVM passes a private copy of this value to each task that needs it, and updates are performed in a functional style before passing these values to any task spawned for the next round. By avoiding updates to any copy of the variable shared by multiple parallel tasks, this pass eliminates unnecessary dependencies and unlocks more speculative parallelism.

**Fine-grained splitting and spatial hints.** When a dependence is violated, the Swarm hardware must roll back and re-execute the work done by the offending speculative task. It is important to minimize this wasted work. We add a pass in the Swarm GraphVM that helps the hardware schedule tasks in a way that reduces both the number of aborts and the cost of each abort. Swarm’s T4 compiler tries to assign spatial hints to each task based on the memory locations that it accesses, but it can do this only for tasks that do not access disparate memory addresses [91]. Line 4 in Figure 5 shows how the GraphVM adds an annotation to instruct T4 to split the subsequent block of code into a subtask that accesses a single memory address. This lets T4 dispatch these subtasks to chip tiles according to the cache line that they access. As a result, accesses to a given cache line are all executed within one chip tile, where hardware can selectively serialize tasks that access the same cache line, reducing the likelihood of aborts [42]. These fine-grained subtasks are also cheaper to roll back and re-execute if they are aborted, reducing the cost of aborts. Additionally, the GraphVM exploits domain knowledge about the loops iterating over constant edge arrays to strike a balance between the cost of aborts and the cost of spawning additional tasks, by generating annotations that help the backend compiler schedule memory access instructions.

4) **HammerBlade Manycore GraphVM**

The HammerBlade Manycore GraphVM produces parallel C++ code targeting the HammerBlade Manycore architecture described in Section II-B4. The code produced by this GraphVM is separated into sequential host code and parallel device code. The sequential host code handles initialization and coordination, while the parallel device code executes the body of the graph algorithm. The HammerBlade Manycore GraphVM implements optimizations and GraphIR transformations that target the manycore architecture and its memory hierarchy. Similar to the GPU GraphVM, the HammerBlade Manycore GraphVM also provides extensive host and device runtime libraries to simplify code generation.

**Atomics.** Similar to a GPU, atomics on the HammerBlade

```c
1    frontier_for_each_prio([](int round, int src) {
2        for (int edgeID: neighbors(src)) {
3            int dst = edgeDest[edgeID];
4            #pragma task hint(&(parent[dst]))
5            if (parent[dst] == -1) {
6                parent[dst] = src;
7                push(round + 1, dst);
8            }
9        }
10    });
```

**Fig. 5:** Code generated by the Swarm GraphVM for BFS.

Manycore are expensive operations. On the manycore, atomic operations in global memory are implemented using lock data structures. The HammerBlade Manycore GraphVM leverages the atomic operations pass from the GPU GraphVM to determine where atomics are necessary. If an atomic operation is determined to be necessary within a kernel, initialization code for the necessary locks is also inserted into the host code.

**Blocked access optimization.** The HammerBlade Manycore GraphVM implements an optimization that utilizes the software managed scratchpads on each core. We call this optimization the blocked access method. This method aims to reschedule long-latency requests to main memory in parallel by partitioning work items into blocks. Cores iterate over their assigned blocks, prefetching the entire block at once. Block data is stored in the core’s scratchpad memory, repurposing it as a software-managed L1 cache. The HammerBlade Manycore GraphVM determines which elements can be safely read into scratchpad memory without requiring synchronization between cores at the end of processing.

**Alignment-based partitioning.** Memory-level parallelism and workload partitioning are very important to achieve high performance on the HammerBlade Manycore. We propose an alignment-based partitioning method that aims to improve memory system performance. In this method, cores work on smaller work blocks of vertices that better align with cache lines in the LLC, increasing effective memory access bandwidth. This optimization utilizes a similar partitioning scheme to the blocked access method, but does not use the core’s scratchpad memory. This is due to the observation that, for some graph workloads, the cost of loading data into scratchpads outweighs the benefit of low-latency scratchpad accesses. Graph vertices are split into $V/b$ work blocks, where $b$ is the number of vertices contained in each work block, and $V$ is the total number of vertices in the graph. We select $b$ to be a multiple of the cache line size. Cores work on these blocks until all work blocks have been processed. Because this reduces the size of the active vertex set that each core is working on, we are able to increase the cache hit rate and reduce cache line contention.

**D. Extensible Scheduling**

One of the main features of UGC is that it decouples the algorithm input from the optimization schedules. This way, the programmer or an autotuner [7] can generate different variants of the same algorithm tailored to specific graph inputs simply bysupplying different schedules. Since different GraphVMs support different optimizations, we build a new scheduling
Abstract Class | Description
--- | ---
SimpleSchedule | Hardware-independent abstract class for simple schedule objects.

Function | Description
--- | ---
getParallelization | Get the parallelization scheme of the schedule (VERTEX_BASED or EDGE_BASED).
getDirection | Get the direction of traversal of edges. Can be PUSH or PULL.
getPullFrontier | Get how the next frontier will be created. Can be BOOLMAP or BITMAP.
getDuplication | Get whether explicit deduplication should be performed on the output frontier.
getDelta | Get the delta value to use when creating buckets in PriorityQueue.

TABLE IV: Description of the SimpleSchedule type and some associated virtual functions.

Abstract Class | Description
--- | ---
CompositeSchedule | Hardware-independent abstract class for hybrid schedule objects (schedule that changes based on runtime value).

Function | Description
--- | ---
getFirstSchedule | Get the first schedule object within this hybrid schedule.
getSecondSchedule | Get the second schedule object within this hybrid schedule.

TABLE V: Description of the CompositeSchedule type and some associated virtual functions.

language for each target. These scheduling languages have essential features for optimizations on their respective targets.

One of the challenges with this approach is that the hardware-independent part of UGC now has to deal with different scheduling languages for the parameters that it needs. For example, the dependence analysis to insert atomics in the UDFs at least needs to know if the parallelization is vertex based or edge based and if the traversal direction is PUSH or PULL.

To address this problem, we use object-oriented programming techniques to enable the hardware-independent part of UGC to query the information that it needs from various scheduling representations. The scheduling language input is stored internally as scheduling objects attached to program nodes. UGC creates an abstract interface with virtual functions for all of the information that the hardware-independent compiler needs. We implement new scheduling object classes for each GraphVM by inheriting from this abstract interface. These new classes have members and functions to configure various scheduling options specific to optimizations supported for their GraphVMs. These classes implement the virtual functions to provide the hardware-independent part of UGC with the information that it needs. Tables IV and V describe these abstract scheduling classes with the virtual functions to query information, such as direction and parallelization type.

Figure 6 shows example scheduling inputs for the BFS algorithm for different GraphVMs. The HammerBlade schedule example shows hybrid traversal with cache-aligned load balancing, while the Swarm example enables transformations for consecutive frontiers into a priority queue and breaks down updates into smaller tasks.

Figure 6a shows a use of the CompositeGPUSchedule class, which inherits from the CompositeSchedule class shown in Table V. The CompositeGPUSchedule object is a hybrid schedule combining two AbstractSchedule objects (which could be other CompositeSchedule objects). The user also specifies the runtime criteria and its associated parameters. Here, the INPUT_SET_SIZE criteria is used with 0.15 as the criteria. This tells the compiler to generate code that chooses between sched1 and sched2, based on whether the input vertex set is above 15% of the total vertices in the graph. Figure 7 shows the generated code. The conditions and copies of the EdgeSetIterator with schedules sched1 and sched2 attached are created by the hardware-independent compiler and GraphVMs need not be aware of it. The compiler generates a nested if-then-else statement if multiple CompositeSchedule objects are combined.

IV. Evaluation

In this section, we demonstrate that UGC supports implementing optimizations that are critical for performance on the four architectures we target: CPUs, GPUs, Swarm, and the HammerBlade Manycore. We compare the performance of optimized code generated by the GraphVMs for each of the architectures with baseline, unoptimized code on 5 graph algorithms and up to 10 different graph inputs. Baseline code is generated by applying the default schedule for each GraphVM to the algorithm. For the optimized version, we tune the schedules for each application and graph pair, but always compile from exactly the same algorithm specification.
We evaluate the GPU GraphVM on a system with 128 cores in 16 tiles (4 cores/tile), 3.5 GHz, x86-64 ISA, 64 MB per-core, 8-way, inclusive, 9-cycle latency. The GPU has 4 KB Instruction Cache, 4 KB Data Scratchpad, and 4 KB Direct-mapped L1 cache. We evaluate the Swarm GraphVM by COHERENCE MESI, 64 B lines, in-cache directories.

Table VIII lists the input graphs used in the evaluation. Each undirected edge is counted twice, once per direction. CC [8, 80] are topology-driven algorithms where all the edges are traversed in each iteration. These applications have massive parallelism each round. BC [9] and BFS are data-driven algorithms where only a set of active vertices are processed each round. SSSP with Δ-stepping is a priority-based algorithm where the vertices are processed in a priority order for greater work efficiency. UGC compiles a single source code specification for each algorithm, reusing the same application code for all different architectures. In real-world applications, these algorithms could be run many times on one graph or class of graph (e.g., one runs many iterations of PageRank, while BFS, BC, and SSSP may be rerun from different starting vertices), necessitating tuning the implementation to the characteristics of the graph and architecture for high efficiency.

Schedules. The performance of the GraphVMs heavily depends on the schedules specified. We manually wrote schedules to tune the implementation of each algorithm to the graph type (e.g., road graphs vs. social graphs). Schedule parameters were further tuned by sweeping the parameter space. Prior work [15, 93, 95] has also shown that techniques like autotuning can find high-performance schedules in relatively little time.

B. Performance of Optimized Code

Figure 8 shows the performance improvements produced by optimization passes in each of our four GraphVMs. The speedups reported here are over the baseline code generated by applying the default schedule. Both the baseline and optimized code are parallel, and all generated C++ is compiled with optimizations enabled in the backend compiler.

We now discuss how the hardware-specific optimizations in the GraphVMs produce these speedups.

C. CPU and GPU

The baseline schedule for CPUs and GPUs uses push-based traversal with vertex-based parallelism. UGC achieves large speedups (up to 53×) on both of the architectures on BFS and BC by using Hybrid (Push+Pull) traversals and tuning the input frontier representation. PageRank greatly benefits from EdgeBlocking and NUMA optimizations, which improve locality of random accesses by tiling for the last-level cache. SSSP on CPUs benefits from the bucket fusion optimization for road graphs. This is consistent with the speedups of the GraphIt compiler [93]. Finally, CC benefits from better load balancing techniques (ETWC) on GPUs, and from edge-aware

<table>
<thead>
<tr>
<th>TABLE VI: Configuration of the 64-core Swarm system.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cores</td>
</tr>
<tr>
<td>L1 Cache</td>
</tr>
<tr>
<td>L2 Cache</td>
</tr>
<tr>
<td>L3 Cache</td>
</tr>
<tr>
<td>Coherence</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE VII: HammerBlade Manycore configuration.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cores</td>
</tr>
<tr>
<td>RISC-V 32-bit IMAF ISA</td>
</tr>
<tr>
<td>4 KB Instruction Cache</td>
</tr>
<tr>
<td>4 KB Data Scratchpad</td>
</tr>
<tr>
<td>Cache</td>
</tr>
<tr>
<td>32 Independent Banks</td>
</tr>
<tr>
<td>8-way Set Associative</td>
</tr>
<tr>
<td>NoC</td>
</tr>
<tr>
<td>32-bit data, 64-bit addr</td>
</tr>
<tr>
<td>Memory</td>
</tr>
<tr>
<td>32 GB/s per channel</td>
</tr>
<tr>
<td>512 MB per channel</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE VIII: Graph inputs used for evaluation.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graph</td>
</tr>
<tr>
<td>------------------------------------------------</td>
</tr>
<tr>
<td>RN [54]</td>
</tr>
<tr>
<td>RC [27]</td>
</tr>
<tr>
<td>RU [27]</td>
</tr>
<tr>
<td>PK [54]</td>
</tr>
<tr>
<td>HW [26]</td>
</tr>
<tr>
<td>LJ [54]</td>
</tr>
<tr>
<td>OK [73]</td>
</tr>
<tr>
<td>IC [26]</td>
</tr>
<tr>
<td>TW [73]</td>
</tr>
<tr>
<td>SW [73]</td>
</tr>
</tbody>
</table>

A. Methodology

CPU and GPU. We evaluate the GPU GraphVM on a system with an NVIDIA Tesla V100 GPU with 32 GB of GDDR5 main memory, 6 MB of L2 cache, and 128 KB of L1 cache per SM, with a total of 80 SMs. This is a Volta-generation GPU. We evaluate the CPU GraphVM on a dual-socket system with Intel Xeon E5-2695 v3 12-core CPUs, for a total of 24 cores and 48 hardware thread contexts. The machine has 128 GB of DDR3-1600 main memory and a 30 MB last-level cache per socket, and has Transparent Huge Pages (THP) enabled.

Swarm Simulation. We evaluate the Swarm GraphVM by running each algorithm’s compiled code in full on the open-source Swarm architectural simulator [62, 91]. We model a 64-core Swarm CPU with parameters shown in Table VI, similar to prior work [4, 91]. We model wide out-of-order cores similar to the Haswell cores in the Xeon E5-2695 v3 used for the CPU GraphVM. We perform cycle-level simulation of Swarm with detailed core, network, and memory system models, and model task and speculation overheads in detail [4, 91].

HammerBlade Manycore Simulation. We model a HammerBlade Manycore system running at 1GHz with 16 columns and 8 rows of core tiles, with parameters shown in Table VII. We use detailed, cycle-accurate RTL simulation to model the RISC-V cores, network on chip, and LLC. The RTL for this manycore has been validated in silicon, and this configuration occupies approximately 3.5 mm² of die area. We model the HBM2 memory system with DRAMSim3 [56], a timing accurate simulator. Generated host code runs natively on an Intel Xeon Gold 6254 CPU, and host libraries interface directly with the simulator environment using SystemVerilog DPI.

Datasets. Table VIII lists the input graphs used in the evaluation, along with their sizes in vertices and edges. Out of the 10 graphs, Orkut (OK), Twitter (TW), LiveJournal (LJ), SinaWeibo (SW), Hollywood (HW), Pokerc (PK), and Indochina (IC) have power-law degree distributions, while RoadUSA (RU), RoadNetCA (RN), and RoadCentral (RC) have bounded degree distributions. These datasets include social graphs, web graphs, and road graphs.

Algorithms. We evaluate all GraphVMs on five algorithms: PageRank, BFS, SSSP with Δ-stepping, connected components (CC) and betweenness centrality (BC). PageRank [66] and
Fig. 8: Heatmap of speedups for the four evaluated architectures. Each cell reports the speedup of the optimized code over the baseline unoptimized version, with larger speedups in darker green. Columns correspond to algorithms, and rows correspond to graph inputs. Some graphs were not run on HammerBlade Manycore due to simulation time constraints.

Fig. 9: Speedups of the GPU GraphVM over the next-best framework from Gunrock, GSwitch, or SEP-Graph. Figure 9 compares the performance of the GPU GraphVM with three state-of-the-art graph libraries that specifically target GPUs: Gunrock [87], GSwitch [60], and SEP-Graph [86]. These speedups are consistent with those of the GPU code generated from GraphIt [15]. UGC is consistently outperformed by SEP-Graph on SSSP when run on road graphs. SEP-Graph implements asynchronous execution to remove barriers between successive rounds of SSSP. UGC does not currently implement this optimization because it is very algorithm specific and cannot be generalized.

D. HammerBlade Manycore

Due to the costs of RTL simulation, we evaluate the HammerBlade Manycore GraphVM on 6 of the 10 input graphs and a subset of the total iterations for each application. For PR we simulate one iteration, and for the remaining applications, we simulate five representative iterations that cover a range of frontier densities and execution behavior. We use hybrid traversal in the baseline code of BFS, BC, and SSSP to decrease simulation times. The speedups reported in Figure 8 come from applying the HammerBlade Manycore-specific optimizations described in Section III-C.4. BC, CC, and BFS benefit from alignment-based partitioning, while PR and SSSP use the blocking optimization due to their more compute-intensive nature. These optimizations better utilize the memory hierarchy and provide up to $4.97 \times$ speedup over unoptimized code.

Figure 10a shows how performance scales on the HammerBlade Manycore. We ran our optimized BFS code on four different machine configurations: we hold the LLC capacity and number of columns (16) constant and vary the number of rows (2, 4, 8, and 16) to vary the total number of cores. The strong scaling indicates that the HammerBlade Manycore GraphVM can successfully exploit parallelism. We highlight BFS for this scaling study due to its high memory access to vertex-based parallelism on CPUs.

Figure 10 compares the scalability of BFS on five graphs, shown across four sizes of HammerBlade Manycore and Swarm machines.

<table>
<thead>
<tr>
<th>Graph</th>
<th>DRAM Stalls</th>
<th>Bandwidth</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>LJ</td>
<td>0.78</td>
<td>3.03</td>
<td>1.19</td>
</tr>
<tr>
<td>HW</td>
<td>0.79</td>
<td>2.17</td>
<td>1.53</td>
</tr>
<tr>
<td>PC</td>
<td>0.83</td>
<td>3.02</td>
<td>1.49</td>
</tr>
</tbody>
</table>

TABLE IX: Impact of the HammerBlade blocked access optimization on SSSP. Reduction in DRAM stalls, improvement in memory bandwidth utilization, and overall speedup.
compute ratio.

Table IX demonstrates performance improvements for SSSP with Δ-stepping when the blocked-access optimization is applied on three selected input graphs. This optimization exploits memory parallelism to hide DRAM access latency in exchange for loading unused data and reducing effective bandwidth. For SSSP, we observe that this optimization decreases DRAM stalls, increases memory bandwidth utilization, and improves overall application performance.

E. Swarm

Figure 8 shows the speedup achieved by choosing an appropriate schedule for each algorithm and graph input, compared the Swarm GraphVM’s default schedule. Swarm’s T4 compiler [91] already applies many optimizations to uncover parallelism in serial code, and achieves good baseline performance in many cases. However, the Swarm GraphVM improves performance further by exploiting domain knowledge to choose optimizations.

On BFS and SSSP, converting VertexSets to tasks is responsible for the majority of the improvement on road graphs. This optimization avoids synchronization overheads between distance levels, by allowing tasks from different levels to execute speculatively in parallel. Additionally, all algorithms benefit from the Swarm GraphVM’s diverse schedule options for task granularity and spatial hints. Fine-grained splitting with spatial hints allows trading increased task overheads for reduced cache line ping-ponging and abort costs. Finally, on CC and PageRank, some graphs featuring many high in-degree nodes benefit from a schedule that shuffles the order in which edges are processed, thus trading off locality to reduce aborts.

This reordering is enabled by the Swarm GraphVM’s domain knowledge that a valid result will still be produced if edges are visited in a different order within one round.

Table X compares the performance of optimized code generated by the CPU and Swarm GraphVMs. Since Swarm offers a superset of a CPU’s features, the CPU code runs on the same Swarm hardware. On road graphs, the Swarm GraphVM consistently outperforms the CPU GraphVM using Swarm’s speculative parallel execution of fine-grained tasks.

![Graph cycle time comparison](image)

Fig. 11: Breakdown of how cores spend time for Swarm.

Figure 11 breaks down how cores spend time with the optimized schedules, averaged across the 64 cores: cores may spend time executing tasks that commit or abort; or idle due to the speculation-throttling heuristic [91], exhaustion of the commit queue, or lack of tasks to run. In some cases, cores also spend time spilling contents of overfull task queues to memory. We see that across all five algorithms, cores spend most of their time executing useful work that commits, demonstrating that the Swarm GraphVM exposes enough fine-grained parallelism through task spawns to utilize tens of cores. This is reflected in Swarm’s strong scalability in Figure 10b. (Adding tiles to Swarm increases aggregate cache and queue capacity, sometimes yielding superlinear speedups.)

Prior work on Swarm has developed hand-tuned versions of BFS and SSSP [42, 43]. Figure 12 shows that the Swarm GraphVM versions are competitive with the manually tuned ones, especially on larger social graphs like TW and SW where the algorithms are memory-bound. The hand-tuned versions were tailored to work well on road graphs, which have low vertex degrees. As a result, the hand-tuned code for SSSP performs poorly on social graphs, where the Swarm GraphVM achieves much better performance by being selective in spawning tasks for the possibly many neighbors of each visited node. UGC makes it easy to bring a wide set of algorithms to port applications to different architectures.

V. RELATED WORK

There has been a large amount of work on both graph processing frameworks and on leveraging of common IRs to port applications to different architectures.

**Common IRs for diverse architectures.** Delite [18] introduces a new IR for parallel programs to target heterogeneous architectures. However, Delite’s IR is generic rather than specific to a particular domain. By customizing the IR specifically for the graph domain, UGC can perform optimizations that are otherwise infeasible in general C++ programs. Furthermore, Delite does not have an extensible scheduling language that allows users to specify optimizations for different targets. MLIR [53] is another proposed IR that is generic and not specific to a domain. Tensorflow [2] and TVM [22] have shown how a common IR can be used to apply machine learning optimizations across different architectures.
Graph processing frameworks. There has been a large body of work on graph processing for shared-memory [3, 31, 32, 39, 52, 69, 78, 79, 82, 83, 85, 92, 94]. GPUs [12, 20, 24, 30, 36, 37, 38, 40, 46, 47, 49, 57, 58, 61, 63, 64, 67, 76, 80, 86, 87], and manycore architectures [21, 55, 69]. These frameworks support a limited set of optimizations, do not achieve consistently high performance across algorithms and graphs [15, 93, 95], and do not offer portability across architectures. Abelian [31] uses the Galois framework as an interface for shared-memory CPU, distributed-memory CPU, and GPU platforms. However Abelian is not extensible enough to support new architectures. In contrast, UGC demonstrates state-of-the-art performance across different platforms.

Compilers for graph applications. IrGL [67] is a compiler framework that creates an intermediate representation specifically for graph applications on GPUs. IrGL introduces several optimizations for GPUs, but does not achieve state-of-the-art GPU performance [15, 86]. GraphIt [15, 93, 95] is a domain-specific language that expands the optimization space to outperform other CPU and GPU frameworks by decoupling algorithm from optimizations. UGC extends GraphIt by decoupling algorithms, optimizations, and hardware backends to enable efficient implementations across different platforms.

VI. CONCLUSION

This paper has presented UGC, a novel graph processing framework that makes it easy to create compiler backends across diverse architectures. We introduced a new IR for graph processing. GraphIR, and showed how it can be used to implement GraphVMs for four different architectures. We demonstrated how UGC can reason about algorithmic and hardware-specific optimizations to generate high-performance code on all four architectures, and find that these optimizations can provide up to 53× speedup over programmer-generated baseline implementations.

ACKNOWLEDGEMENTS

We thank Mark C. Jeffrey, Quan M. Nguyen, Hyun Ryong Lee and the anonymous reviewers for helpful discussions and feedback. This work was partially supported by Air Force Research Laboratory (AFRL) and Defense Advanced Research Projects Agency (DARPA) under agreement numbers FA8650-18-2-7863, FA8650-18-2-7856; DARPA SDH under contract HR0011-18-3-0007; NSF grants SaTC-1563767, SaTC-1565446, SHF-1814969, and CAREER-1845763; DOE Early Career Award DE-SC0018947; a Sony research grant; and the DARPA/SRC JUMP ADA Center. This research was, in part, funded by the U.S. Government. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

REFERENCES


W. Han, D. Mawhirter, B. Wu, and M. Buland, "Graphie: Large-scale asynchronous graph traversals just on a gpu," in Proc. PACT-26, 2017.


C. Hong, A. Sukumaran-Rajam, J. Kim, and P. Sadayappan, "MultiGraph: Efficient graph processing on GPUs," in Proc. PACT-26, 2017.


D. Li, S. Chakradhar, and M. Becchi, "GRapids: A compilation and runtime framework for rapid prototyping of graph applications on many-core processors," in Proc. ICPADS, 2014.


Q. Xu, H. Jeon, and M. Annamraju, "Graph processing on GPUs: Where are the bottlenecks?" in Proc. IISWC, 2014.


