Parallel Breadth First Search on GPU Clusters using MPI and GPUDirect

Speaker: Harish Kumar Dasari,
Scientific Computing and Imaging Institute, University of Utah

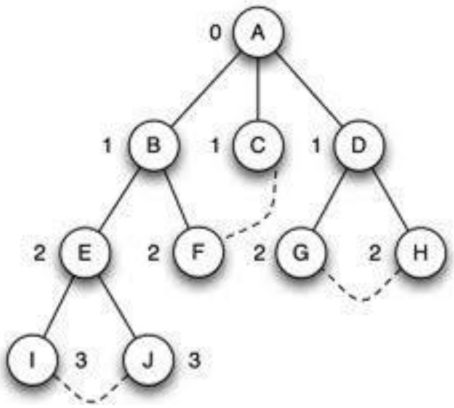
Advisor: Dr. Martin Berzins, SCI, University of Utah

In collaboration with Dr. Zhisong Fu, Bryan Thompson, Systap, LLC.

http://sourceforge.net/projects/mpgraph/

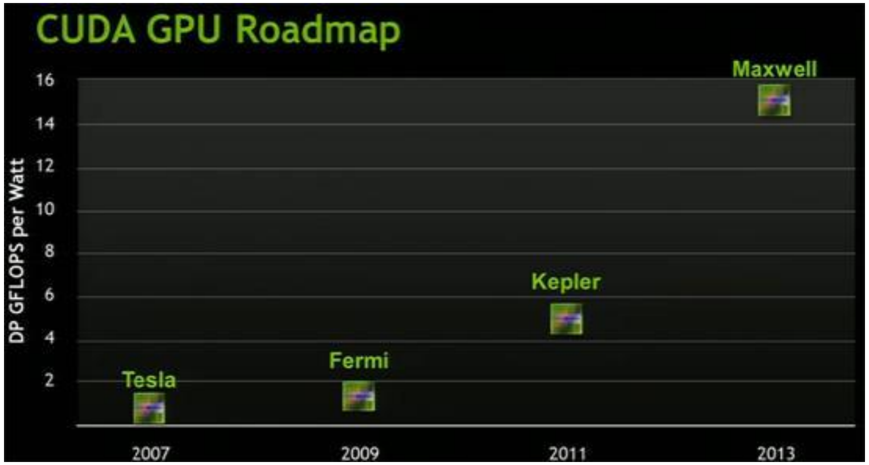
Introduction

- Breadth First Search: It is a graph search algorithm that begins at the root vertex and explores all the connected vertices, traversing all vertices of a particular level before traversing the vertices of the next level
- At the end of the BFS we can find out the level of a vertex if it is connected to the root element and also its predecessor
- Useful in social media, logistics and supply chains, e-commerce, counter-terrorism, fraud detection etc.



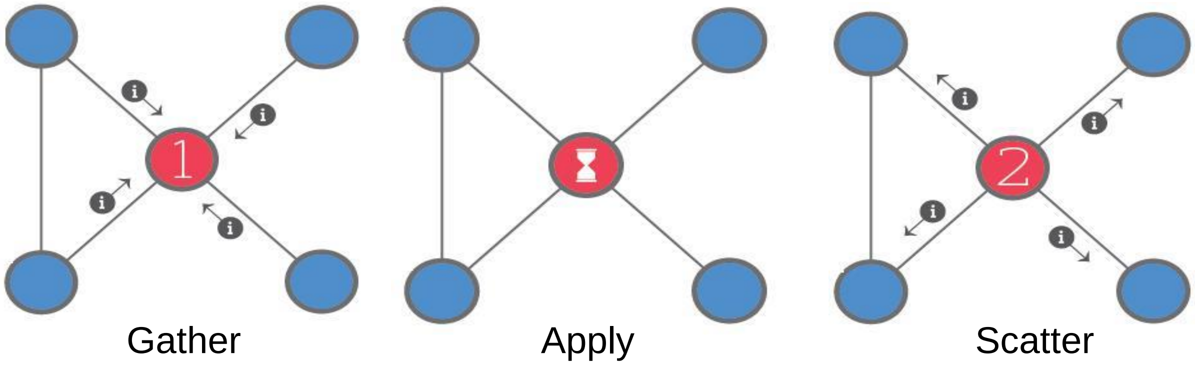
Introduction

- Why BFS?
  - Least work/byte of the graph algorithms
  - Building blocks for many other graph problems
- Why GPUs?
  - High Performance: NVIDIA K40 peak performance: 1.43 Tflops
  - High Energy Efficiency
  - Central for next generation of architectures



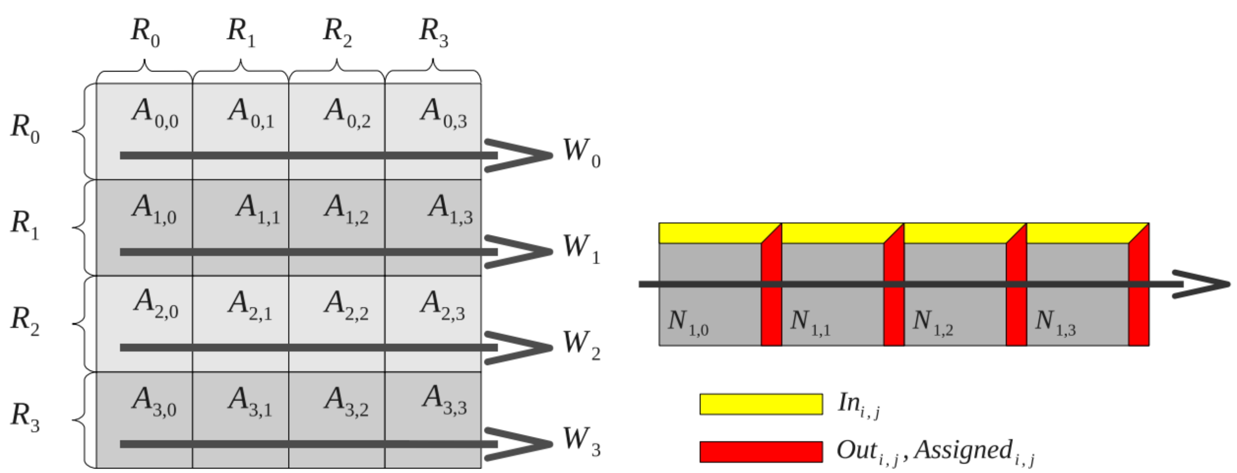
Related Work

- Scalable GPU Graph Traversal - Single node multi-GPU, Merrill, Garland et al.
    - Around 12x speedup over idealized multi-core CPU
    - 3 GTEPS on single node
- MapGraph, Fu, Thompson et al.
    - Generalized for many graph algorithms using Gather Apply Scatter (GAS) abstraction
    - Provides an easy framework for the developer to develop solutions to other graph problems like SSSP(Single Source Shortest Path), PageRank etc.



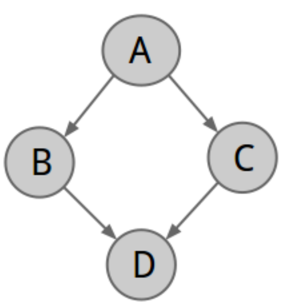
Related Work

- Breaking the Speed and Scalability barriers for graph exploration on distributed-memory machines by Checconi, Petrini et al from IBM
  - BFS on Bluegene supercomputers, uses CPUs
  - On Graph500 data sets, on the order of 2^{40} edges
  - 254 billion edges/sec with 64k cores
  - Uses 2D partitioning and waves for communication



Partitioning of the Graph

- RMAT graph generated using the Graph500 generator
  - Scale Free
  - Follows power law, at least asymptotically
  - undirected edges are converted to directed edges
- 2-D Partitioning of directed edges with a square layout
- Each subgraph resides in GPU memory
- Bitmaps used to represent the frontiers
  - Bit is set to 1 to represent active vertex



	A	B	C	D
A	0	0	0	0
B	1	0	0	0
C	1	0	0	0
D	0	1	1	0

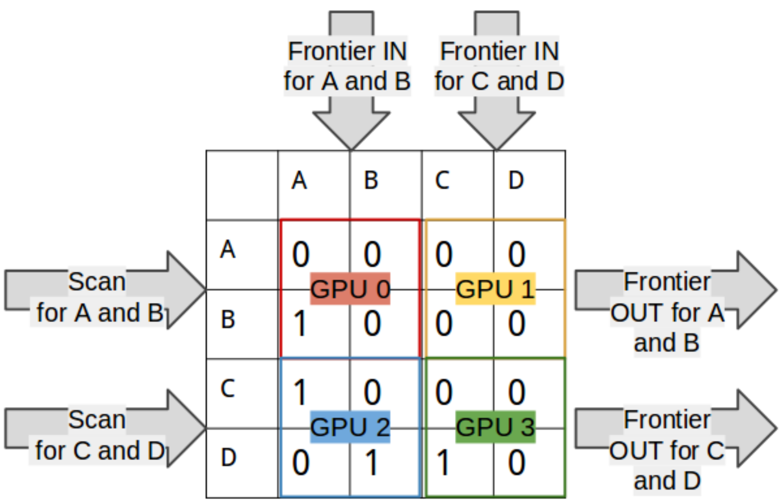
GPU 0 (rows A–B, columns A–B), GPU 1 (rows A–B, columns C–D), GPU 2 (rows C–D, columns A–B), GPU 3 (rows C–D, columns C–D)

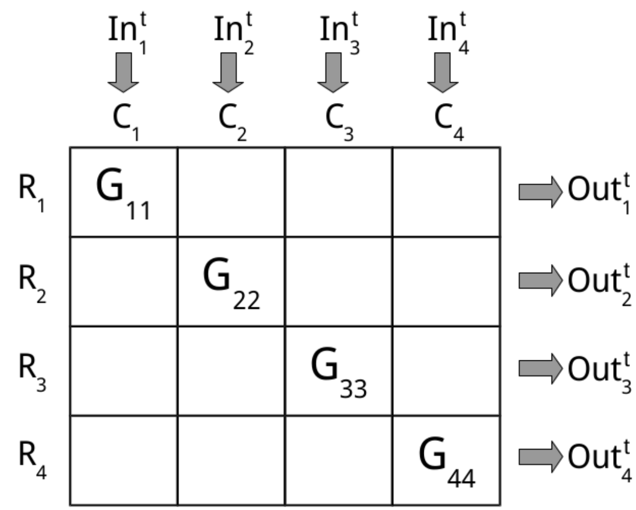
Bitmap

A	B
0	1

The Algorithm and Communication

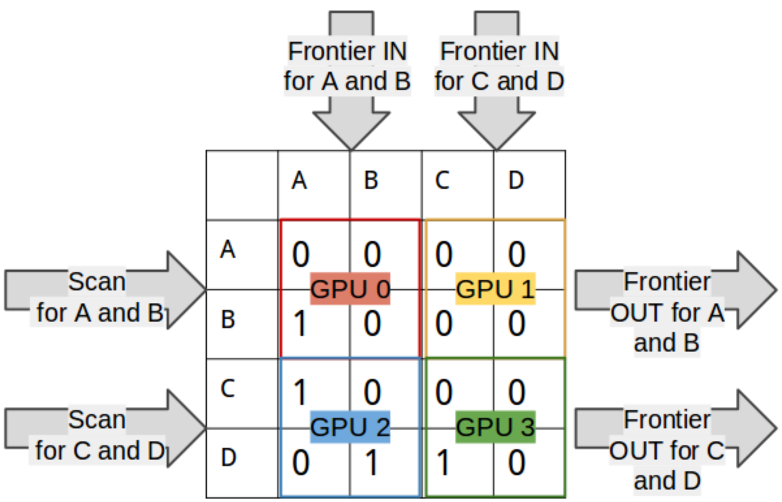
- Each GPU G_{ij} takes in its input frontier bitmap In_i^t and perform BFS on its subgraph to produce Out_{ij}^t
- Parallel Scan for bitmaps along the row R_i to produce prefix sum $Prefix_{ij}$ in Bitwise-OR

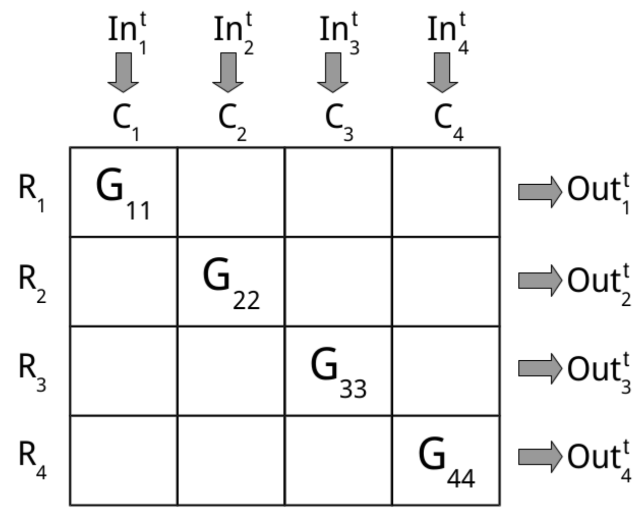




The Algorithm and Communication

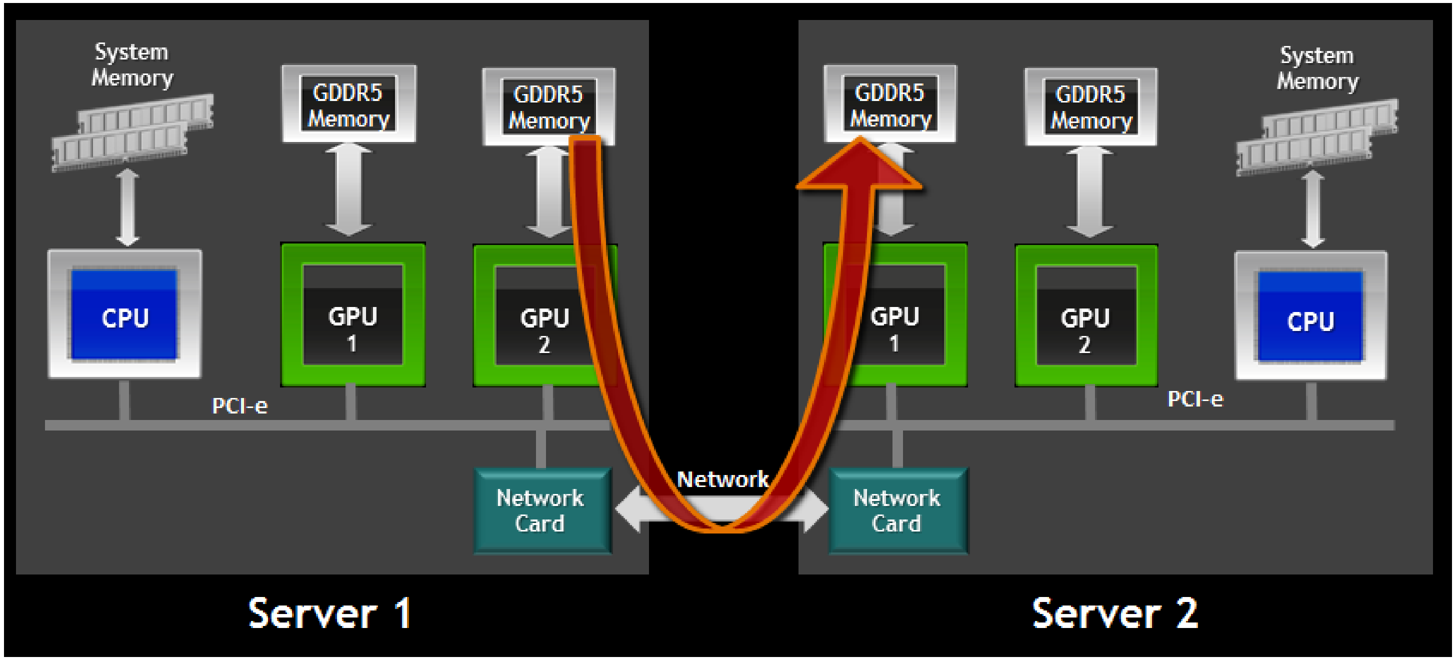
- The Prefix is used to determine the vertices the GPU is assigned for predecessor updates
- Out_i^t is broadcast across row R_i and also as In_i^{t+1} across column C_i





Experimental Setup

- 32 nodes and 64 NVIDIA K20c GPUs with 5GB DDR5 memory
- Two Mellanox InfiniBand SX6025 cards per node
- CUDA 5.5 used for these results
- Used GPUDirect support in MVAPICH2-GDR to avoid explicit copy of messages to host memory

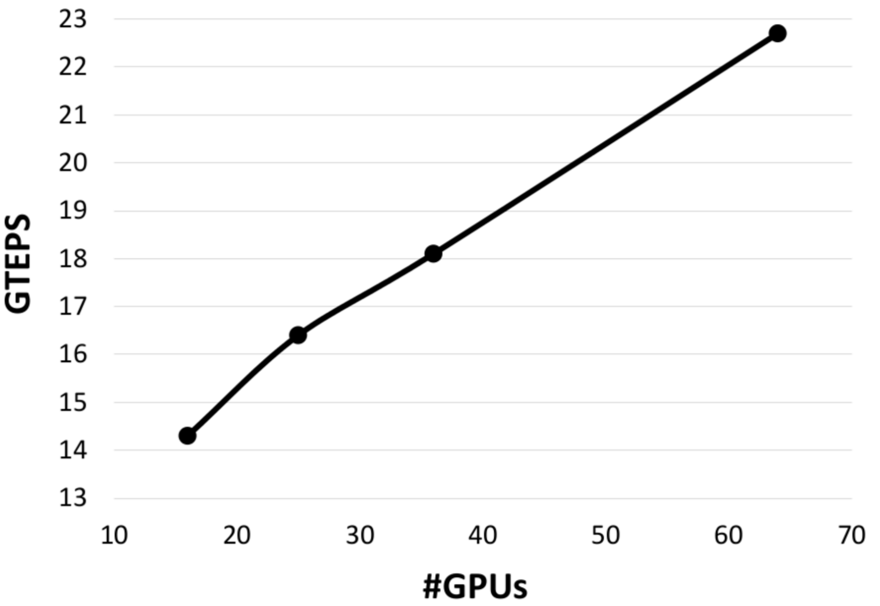


Results - Strong Scaling

- The scale of the problem remains the same as we increase the computational resources (GPUs)
- GTEPS= Giga(Billion) Traversed Edges Per Second = 10^9 edges per second

GPUs	Scale	Time	GTEPS
16	25	0.075	2.5
25	25	0.066	6.3
36	25	0.059	15.0
64	25	0.047	29.1

Number of Vertices in graph = 2^{SCALE}
Number of Directed Edges in graph = $32*2^{SCALE}$

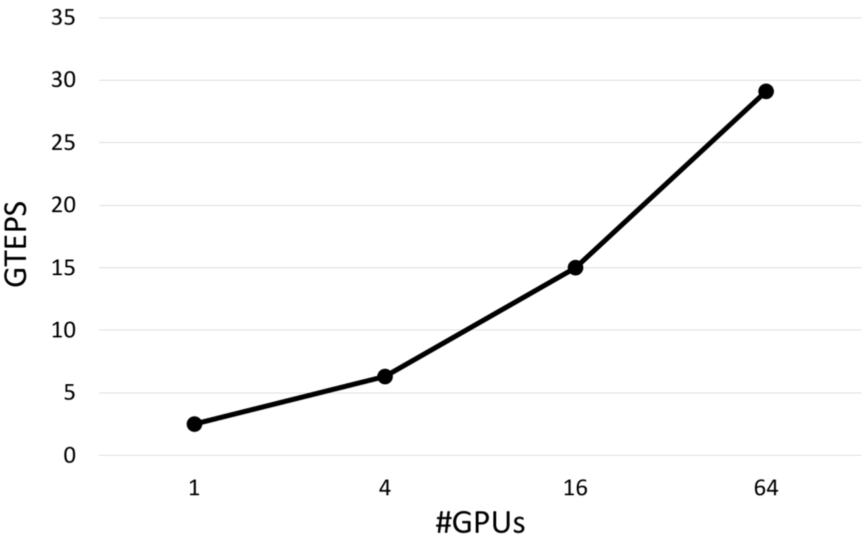


Results - Weak Scaling

- Problem size grows proportional to the growth in computational resources (GPUs)
- Each GPU has same amount of work?

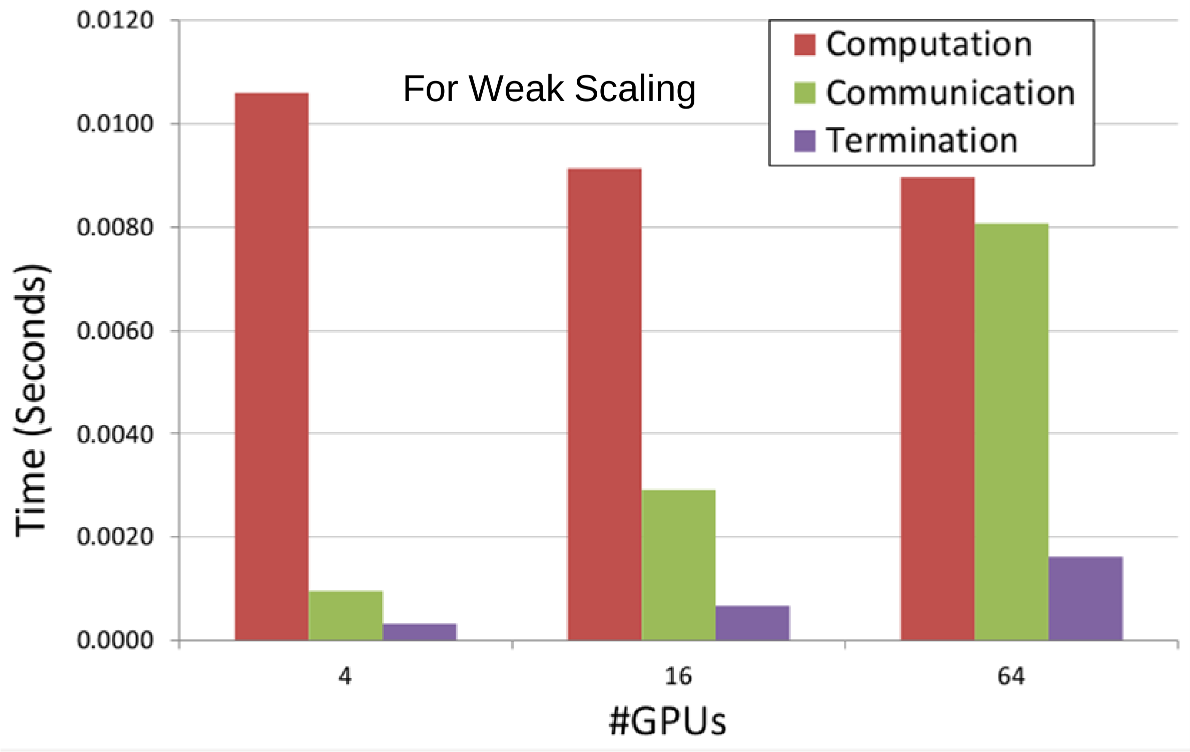
GPUs	Scale	Time	GTEPS
1	21	0.0254	14.3
4	23	0.0429	16.4
16	25	0.0715	18.1
64	27	0.1478	22.7

Number of Vertices in graph = 2^{SCALE}
Number of Directed Edges in graph = $32*2^{SCALE}$



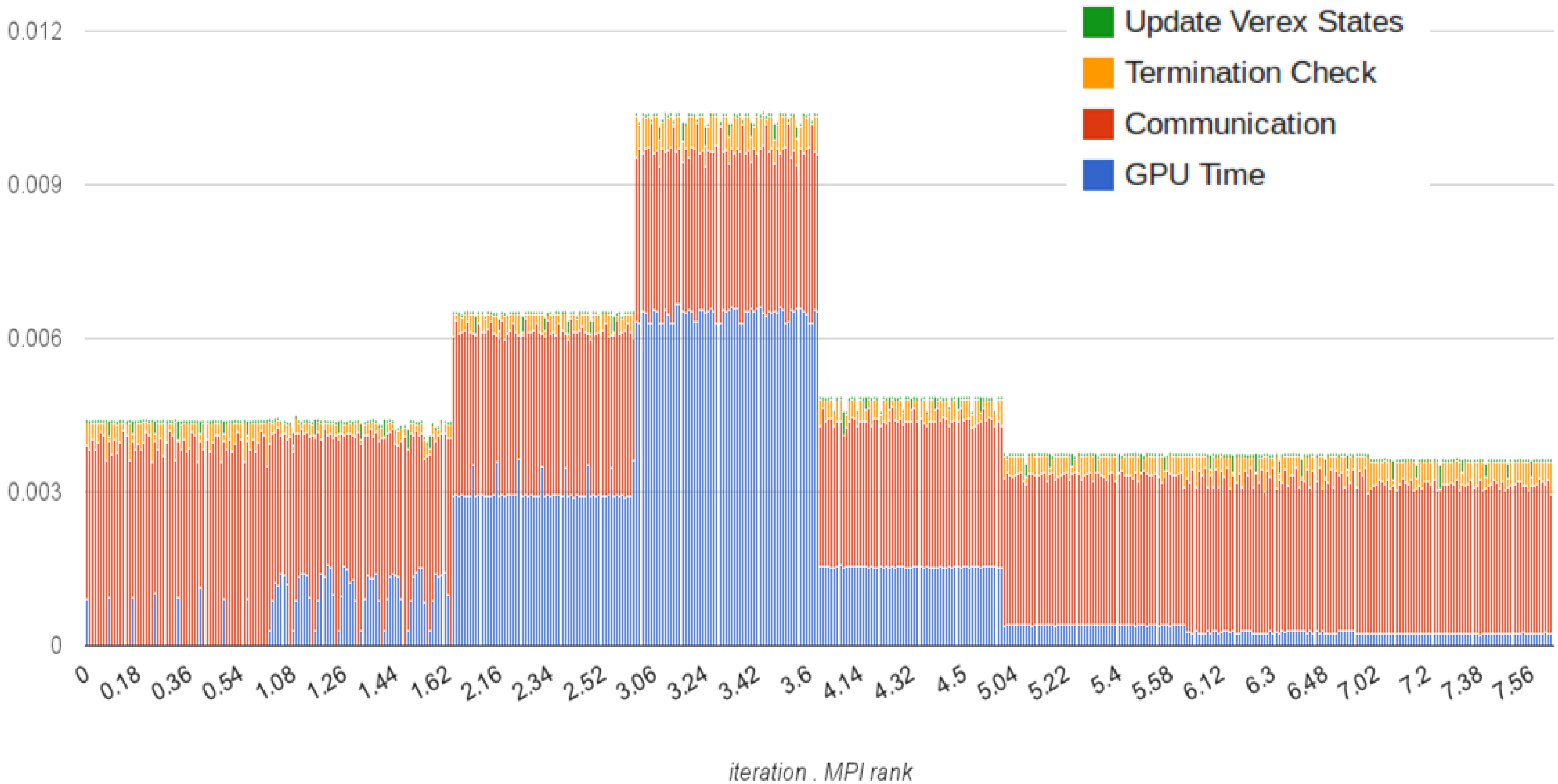
Communication vs Computation

- Even if the work per GPU remains the same, the communication costs grow
- Impacts weak Scalability



Breakdown of Timings

- Near constant communication times across iterations
- Load Imbalance in the first iterations

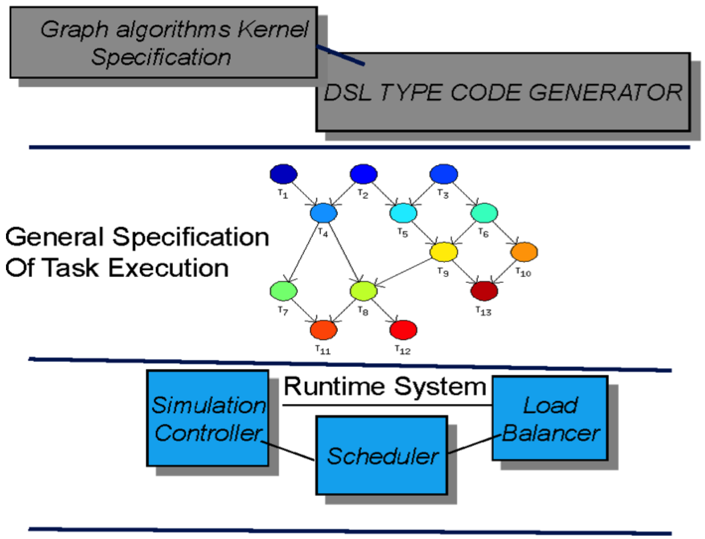


Challenges faced

- Working with cutting edge Hardware and Software stack
- Lack of expertise on the new software
- Crashes due to the old GPU Drivers
- Thanks for the OSU MVAPICH team for the excellent support

Future Scope

- GAS abstraction for the distributed graph processing platform
- Message compression in iterations with small frontiers
- Graph compression to allow us to load bigger graphs to GPU memory
- Runtime engine to overlap computation and Communication, similar to the one in Uintah (www.uintah.utah.edu)



Conclusion

- Implemented a high Performance distributed parallel Breadth First Search (BFS) on GPU cluster
- Implemented a parallel scan for Bitwise-OR reduction of bitmaps
- Implemented compression methods for messages and found that it increases wait times
- This work has been funded by DARPA STTR Phase-1

Queries?

Thank You