Accelerating Force-Directed Graph Drawing with RT Cores

Stefan Zellmann
University of Cologne

Martin Weier
Hochschule Bonn-Rhein-Sieg

Ingo Wald
NVIDIA

ABSTRACT

Graph drawing with spring embedders employs a $V \times V$ computation phase over the graph's vertex set to compute repulsive forces. Here, the efficacy of forces diminishes with distance: a vertex can effectively only influence other vertices in a certain radius around its position. Therefore, the algorithm lends itself to an implementation using search data structures to reduce the runtime complexity. NVIDIA RT cores implement hierarchical tree traversal in hardware. We show how to map the problem of finding graph layouts with force-directed methods to a ray tracing problem that can subsequently be implemented with dedicated ray tracing hardware. With that, we observe speedups of $4 \times$ to $13 \times$ over a CUDA software implementation.

Index Terms: Human-centered computing—Visualization—Visualization techniques—Graph drawings; Computing methodologies—Computer graphics—Rendering—Ray tracing;

1 INTRODUCTION

Graph drawing is concerned with finding layouts for graphs and networks while adhering to particular aesthetic criteria. These can, for example, be minimal edge crossings, grouping by connected components or clusters, and obtaining a uniform edge length. Force-directed algorithms associate forces with the vertices and edges and iteratively apply those to the layout until equilibrium is reached and the layout becomes stationary.

Spring embedders, as one representative of force-directed algorithms, iteratively apply repulsive and attractive forces to the graph layout. The repulsive force computation phase requires $O(|V|^2)$ time over the graph's vertex set $V$. This phase can be optimized using data structures like grids or quadtrees, as the mutually applied forces effectively only affect vertices within a certain radius.

$F_{\text{rep}}(\Delta, k) = \frac{\Delta}{|\Delta|} \frac{k^2}{|A|}$

and

$F_{\text{att}}(\Delta, k) = \frac{\Delta}{|\Delta|} \frac{\Delta^2}{|k|}$

where $\Delta = p(v) - p(u)$ is the vector between the two vertices acting forces upon each other. $k$ is computed as $\sqrt{A/|V|}$, where $A$ is the area of the axis-aligned bounding rectangle of $V$.

As the complexity of the first nested for loop per iteration is $O(|V|^2)$, and by observing that the pairwise forces diminish with increasing distance between vertices, the authors propose to adapt the computation of the repulsive force using:

$F_{\text{rep}}(\Delta, k) = \frac{\Delta}{|\Delta|} \frac{k^2}{|A|} u(2k - |\Delta|)$.
Algorithm 1 Fruchterman-Reingold spring embedder algorithm.

\begin{algorithm}
\begin{algorithmic}
\For{$i := 1$ to $\text{Iterations}$} \Comment{dispersion to displace vertices}
\State $D \leftarrow |V|$ \Comment{calculate repulsive forces ($V \times V$)}
\ForAll{$v \in V$}
\State $D(v) := 0$
\EndFor
\ForAll{$u \in V$}
\State $D(v) := D(v) + F_{\text{rep}}(p(v) - p(u), k)$
\EndFor
\EndFor
\ForAll{$e \in E$} \Comment{calculate attractive forces}
\State $D(v) := D(v) - F_{\text{at}}(p(v) - p(u), k)$
\State $D(u) := D(u) + F_{\text{at}}(p(u) - p(v), k)$
\EndFor
\ForAll{$v \in V$} \Comment{displace vertices according to forces}
\State $\text{DISPLACE}(v, D(v), t)$ \Comment{$t$ is a damping factor}
\EndFor
\State $t := \text{COOL}(t)$ \Comment{Decrease damping factor}
\end{algorithmic}
\end{algorithm}

where $u(x)$ is 1 if $x > 0$ and 0 otherwise. With that, only vertices inside a radius $2k$ will have a non-zero contribution, which in turn allows for employing acceleration data structures to focus computations on only vertices within the neighborhood of $p(v)$.

The FR algorithm is a good match for GPUs as the three phases—repulsive force computation, attractive force computation, and vertex displacement—are highly parallel. The most apparent parallelization described by Klapka and Slaby\cite{25} devotes one GPU kernel to each phase. The outer dimension of the nested for-loop over $v \in V$ is executed in parallel, but each GPU thread runs the full inner loop over $u \in V$ in Alg. 1. This reduces the time complexity to $\Theta(|V|)$, whereas the work complexity remains $\Theta(|V|^2)$. Force-directed algorithms—and in general graph drawing algorithms based on nearest neighbor search—lend themselves well to massive parallelization on distributed systems\cite{10} or on many-core systems and GPUs using nearest neighbor queries\cite{18,31,33}.

Gajdos et al.\cite{10} accelerate the repulsive force computation phase by initially sorting the $v \in V$ on a Morton curve. This order is subdivided into individual blocks to be processed in parallel in separate CUDA kernels. However, this process is inaccurate, as forces will only affect vertices from the same block. The authors try to account for that by randomly jittering vertex positions so that some of them spill over to neighboring blocks. Mi et al.\cite{29} use a similar approximation but motivate that by imbalances originating from the multi-level approach described in\cite{18} that they use in combination with FR. Our approach does not use approximations but is equivalent to the FR algorithm using the grid optimization that was proposed in the original work.

General nearest neighbor queries have been accelerated on the GPU with $k$-d trees\cite{37}, as in the work of Hu et al.\cite{22} and by Wehr and Radkowski\cite{37}. For dense graphs with $\Theta(|E|) = \Theta(|V|^2)$, the attractive force phase can also become a bottleneck. The works by Brandes and Pich\cite{5} and by Gove\cite{15} propose to choose only a subset of $E$ using sampling to compute the attractive forces. Gove also suggests using sampling for the graph’s vertex set $V$ to improve the complexity of the repulsive force phase\cite{16}. Other modifications to the stress model exist. The COAST algorithm by Ganser et al.\cite{12} extends force-directed algorithms to support given, non-uniform edge lengths. They reformulate the stress function based on those edge lengths so that it can be solved using semi-definite programming. The maxstress-extension by Ganser et al.\cite{15} initially solves the model only for the edge lengths and later resolves the remaining degrees of freedom via an entropy maximization model. The repulsive force computation in this work is based on the classical N-body model by Barnes and Hut\cite{3} and uses a quadtree data structure for the all-pairs comparison. Hachul and Jünger\cite{20} gave a survey of force-directed algorithms for large graphs. For a general overview of force-directed graph drawing algorithms, we refer the reader to the book chapter\cite{26} by Kobourov.

2.2 RTX ray tracing

NVIDIA RTX APIs allow the user to test for intersections of rays and arbitrary geometric primitives. This technique is often used to generate raster images. Here, Bounding volume hierarchies (BVHs) help reduce the complexity of this test, which is otherwise proportional to the number of rays times the number of primitives. The user supplies a bounds program so that RTX can generate axis-aligned bounding boxes (AABBs) for the user geometry and build a BVH. Now, a ray generation program can be executed on the GPU’s programmable shader cores that will trace rays through the BVH using an API call. In the intersection program, called when rays hit the AABBs, the user can test for and potentially report an intersection with the geometry. A reported intersection will then be available in potential closest-hit or any-hit. RTX GPUs perform BVH traversal in hardware. When RTX calls an intersection program, hardware traversal is interrupted and a context switch occurs that switches execution to the shader cores.

RTX was recently used to accelerate visualization algorithms like direct volume rendering\cite{30} or glyph rendering\cite{39}. RT cores have, however, also been used for non-rendering applications, such as the point location method on tetrahedral elements presented by Wald et al.\cite{50}.

3 Method Overview

We propose to reformulate the FR algorithm as a ray tracing problem. That way, we can use an RTX BVH to accelerate the nearest neighbor query during the repulsive force computation phase. The queries and data structures used by the two algorithms differ substantially: force-directed algorithms use spatial subdivision data structures, whereas RTX uses object subdivision. Nearest neighbor queries do not directly map to the ray/primitive intersection query supported by RTX. However, we present a mapping from one approach to the other and demonstrate its effectiveness using an FR implementation with the CUDA GPU programming interface.

3.1 Mapping the force-directed graph drawing problem to a ray tracing problem

We present a high-level overview of our approach in Fig. 2. A nearest neighbor query can be performed by expanding a circle around the position $p(v)$ of the vertex $v \in V$ that we are interested in and gathering all $u \in V, u \neq v$ inside that circle. To compute forces, we would perform a search query for all $v \in V$ and integrate the accumulation of the forces directly into the query.

By observing that the circle we expand around $v$ always has a radius $2k$, we can reverse the problem: instead of expanding a circle around $v$, we instead expand circles around all $v \in V$. We then trace an epsilon ray with infinitesimal length and origin at $p(v)$ against this set of circles and accumulate the forces whenever $p(v)$ is inside the circle associated with $u \in V$, given that $u \neq v$. The intersection routine of the ray tracer only has to compute the length of the vector between the ray origin and the center of the circle and report an intersection whenever that length is less than $2k$. Geometrically, one can think of this as splatting, where the splats whose footprints overlap $p(v)$ act a repulsive force upon $v$.

The runtime complexity of the repulsive force computation phase using nearest neighbor queries can be reduced from $\Theta(|V|^2)$ to $\Theta(|V| \log(|V|))$ using spatial indices like quadtrees\cite{18} or binary space partitioning trees\cite{28} built over $V$. The spatial index would have to be rebuilt on each iteration. Likewise, the ray tracing query complexity can be reduced in the same manner using a BVH.
We implemented the FR algorithm with CUDA. We use separate CUDA kernels for the repulsive and attractive forces and for the vertex dispersion phase. Those kernels are called sequentially in a loop over all iterations. The dispersion that is computed during the force phases is stored and updated in a global GPU array.

The parallel attractive force phase uses atomic operations to update the dispersion array. The repulsive phase is implemented using OptiX 7 and the OptiX Wrapper Library (OWL) [35]. Since the number of vertices will never change, we use a global, fixed-size GPU array for the 2-d positions that is shared between CUDA kernels and OptiX programs. Initial vertex placement is at random and in a square. RTX does not support 2-d primitives, so that we construct the BVH from discs with infinitesimal thickness.

The ray generation program spawns one infinitesimal ray per vertex \( v \) originating at \( p(v) \); we again account for RTX being a 3-d API by setting the \( z \) coordinates of the ray origin and direction vector to 0 and 1, respectively. In this way, we can directly accumulate the dispersion inside the intersection program and do not even have to report an intersection that would otherwise be passed along to a potential closest-hit or any-hit program.

### 4 Evaluation

For a comparison with a fairly optimized, GPU-based nearest neighbor query, we use a 2-d spatial data structure based on the LBVH algorithm [27, 40]. As the vertices have no area, we obtain a 2-d BSP tree with axis-aligned split planes that subdivide parent nodes into two same-sized halves (middle split). With the restriction being relaxed that two split planes need to be placed at once, we should outperform the commonly used grid or quadtree implementations [6, 16].

Using Karras’ construction algorithm [24], the build complexity is \( O(n) \) in the number of primitives. Our motivation to use a data structure with superior construction performance is that is must be rebuilt after each iteration. We use a full traversal stack in local GPU memory and perform nearest neighbor queries by gathering all vertices within a 2\( r \) radius around the current vertex position at the leaves. We have a slight advantage over RTX as our data structure is tailored for 2-d. At the same time we cannot possibly optimize our data structure in the same way that NVIDIA probably has done with RTX, and neither that this is our goal with this comparison.

Note that the LBVH and RTX implementations and grid-based FR result in identical graph layouts. In comparison to state-of-the-art implementations in graph drawing libraries such as OGDF [6], Tulip [2], or Gephi [4]—all of which provide sequential CPU implementations of FR—both our RTX and LBVH solutions are magnitudes faster. In order to put both our GPU results into perspective, we also implemented the naive GPU parallelization from [25] over just the outer loop of the repulsive force phase.

We report execution times for the four data sets depicted in Table 1. Two artificial data sets consist of many fully connected \( K_5 \): 10 graphs (five vertices, ten edges). In one case we use 50K of those and sequentially connect pairs of them with a single edge. In the second case we use 50K of them as individual connected components. We also test using a complete binary tree with depth 16, as well as the graph representing twitter feed data that is also depicted in Fig. 1. For the results reported in Table 1, we used an NVIDIA GTX 1080 Ti (no RT cores), an RTX 2070, and a Quadro RTX 8000. The scalability study from Fig. 3 and the evaluation of the repulsive phase in Table 2 were conducted solely on the Quadro GPU.

### 5 Discussion

Our evaluation suggests speedups of \( 4 \times \) to \( 13 \times \) over LBVH. From the difference between the mean iteration times in Table 1 and the mean times for only the repulsive phase in Table 2, we see that the algorithm is dominated by the latter. The other phases plus overhead account for less than 1 % of the execution time. While Fig. 3 shows that our method’s performance overhead for small graphs can be neglected—because it is on the order of about 1 ms—we observe dramatic speedups that increase asymptotically with \( |V| \).

Interestingly, we see about the same relative speedups on the GeForce GTX GPU and on the RTX 2070 GPU with hardware acceleration. At the same time, we observe that the absolute runtimes differ substantially, which we cannot intuitively explain, as neither the peak performance in FLOPS, nor the memory performance of the two GPUs, differ that much. Profiling our handwritten CUDA nearest neighbor query, we find tree traversal to be limited by L2 cache hit rate, which is about 20 %. For RTX, such an analysis is impossible and we can only speculate about the results. It is conceivable that the RTX BVH has an optimized memory layout such as the one by Ylitie et al. [38]. Assuming that we are bound by memory access latency, the speedups we observe might stem from better utilization of the GPU’s memory subsystem rather than hardware acceleration. Switching between hardware and software execution on RTX GPUs incurs an expensive context switch. Hardware traversal is interrupted whenever the intersection program is called. For our test data sets, we consistently found the average number of intersection program instances called to be in the hundreds. We might see an adversarial effect where we, on the one hand, benefit from hardware acceleration, but on the other hand suffer from expensive context switches and that the two effects in the end cancel. We find the speedups that we observe reassuring, especially because using RTX...
Table 1: Statistics and average execution times on different GPUs. We use three artificial graphs with different connectivity and edge degrees, and a twitter feed graph. $c \in C$ denote connected components. Execution times reported are per full iteration including all phases.

| Data Set | Mode | Mem | Build | Traversal | $\Sigma F_{rep}$ | Speedup |
|----------|------|-----|-------|-----------|-----------------|--------|
| 5K $\times$ K_5; 10 (connected) | LBVH | 1.53 | 10.92 (8.37%) | 10.0 (91.6%) | 10.9 | 4.27 x |
| | RTX | 1.18 | 11.16 (45.5%) | 1.39 (54.5%) | 2.55 | 4.44 x |
| Twitter | LBVH | 4.16 | 11.94 (7.94%) | 22.5 (92.1%) | 24.4 | 11.2 x |
| | RTX | 3.22 | 21.38 (39.7%) | 3.31 (60.3%) | 5.49 | 13.0 x |
| Binary Tree (Depth=16) | LBVH | 8.00 | 25.3 (8.34%) | 62.3 (96.2%) | 65.8 | 7.77 x |
| | RTX | 6.19 | 23.26 (40.3%) | 3.30 (59.7%) | 5.87 | 8.33 x |
| 50K $\times$ K_5; 10 (unconnected) | LBVH | 15.3 | 28.7 (3.26%) | 85.4 (96.7%) | 88.3 | 6.77 x |
| | RTX | 11.8 | 28.2 (41.6%) | 3.95 (58.4%) | 7.12 | 8.88 x |

Table 2: Acceleration data structure statistics on RTX 8000, for the repulsive force computation phases. Execution times per iteration are given in milliseconds and the ratio of build vs. traversal times in percent. We also report total BVH memory consumption in MB.

| Data Set | Mode | Mem | Build | Traversal | $\Sigma F_{rep}$ | Speedup |
|----------|------|-----|-------|-----------|-----------------|--------|
| 5K $\times$ K_5; 10 (connected) | Naive LBVH | 1.18 | 10.92 (8.37%) | 10.0 (91.6%) | 10.9 | 4.27 x |
| | RTX | 1.18 | 11.16 (45.5%) | 1.39 (54.5%) | 2.55 | 4.44 x |
| Twitter | LBVH | 4.16 | 11.94 (7.94%) | 22.5 (92.1%) | 24.4 | 11.2 x |
| | RTX | 3.22 | 21.38 (39.7%) | 3.31 (60.3%) | 5.49 | 13.0 x |
| Binary Tree (Depth=16) | LBVH | 8.00 | 25.3 (8.34%) | 62.3 (96.2%) | 65.8 | 7.77 x |
| | RTX | 6.19 | 23.26 (40.3%) | 3.30 (59.7%) | 5.87 | 8.33 x |
| 50K $\times$ K_5; 10 (unconnected) | LBVH | 15.3 | 28.7 (3.26%) | 85.4 (96.7%) | 88.3 | 6.77 x |
| | RTX | 11.8 | 28.2 (41.6%) | 3.95 (58.4%) | 7.12 | 8.88 x |

Figure 3: Scalability study where we build complete binary trees with depth $D = 4, 5, \ldots, 18$. Left: linear scale, right: logarithmic scale. We report mean times for only the repulsive force phase.

The method by Gajer and Kobourov [11] employs a refinement phase that uses FR to compute local displacement vectors. Although we assume that our approach will complement state-of-the-art algorithms with better convergence rates, a thorough comparison is outside of this paper’s scope and presents a compelling direction for future work.

6 LIMITATIONS OF OUR STUDY

We acknowledge that force-directed methods for large graphs exist that require fewer iterations to arrive at a converged layout and outperform FR by far in this regard [20] and are often based on multilevel optimizations [34]. We chose FR as a most simple force-directed algorithm to reason about the speedup and practicability of our approach. Algorithms that perform a nearest neighbor search to compute forces will generally benefit from the proposed techniques. The Fast Multiple Multilevel Method (FM^3) [19] employs such a nearest neighbor search and uses a coarsening phase in-between iterations. Similar to our method, the GPU multiple algorithm by Godiyal et al. [14] employs a k-d tree that is rebuilt per iteration, uses stackless traversal, and would likely benefit from RTX. The GRIP

7 CONCLUSIONS

We presented a GPU-based optimization to the force-directed Fruchterman-Reingold graph drawing algorithm by mapping the nearest neighbor query performed during the repulsive force computation phase to a ray tracing problem that can be solved with RT core hardware. The speedup over a nearest neighbor query with a state-of-the-art data structure that we observe is encouraging. Force-directed algorithms lend themselves to a parallelization with GPUs. We found that those algorithms can be optimized even further by using RT cores and hope that our work raises awareness for this hardware feature even outside the typical graphics and rendering communities.
