Abstract

Previous work has used entry point search (EP search) to accelerate raytracing by avoiding redundant traversal of nodes in a KD tree. EP search works by identifying a node which may be used as a starting point for tracing a particular group of coherent rays, thus bypassing all nodes above it. This is most effective when all rays can be shown to be occluded, since a leaf or low inner node can often be used as the entry point. Unfortunately, occlusion detection is difficult in scenes with complex geometry, and is particularly difficult for BVH data structures, for which ordered traversal cannot be guaranteed. If occlusion cannot be considered, EP search alone becomes considerably less effective. We present an alternative to EP search which eliminates traversals above and below the entry point node. Compared to EP search, our method has a similar run-time cost, generally eliminates more nodes, never eliminates fewer nodes, and is more effective in the absence of occlusion. Although designed for BVHs, our algorithm can be easily applied to other tree structures.


Keywords: Ray Tracing, Frustum Culling, Entry Point Search, Path Compression

1 Introduction

Raytracing is a very important technique in computer graphics, which has seen a surge of exciting research activity in recent years. It provides a framework for simulating many diverse visual phenomena, and it can easily accommodate a wide variety of surface representations. The fundamental operation in raytracing is determining the first intersection point between a ray and a geometric primitive in a 3D scene. This is often done by traversing a hierarchical data structure containing the scene geometry. Raytracing in this manner scales logarithmically with scene complexity, making it a favorable choice for rendering scenes containing large numbers of primitives.

Wald et al. [2001; 2003] introduced the use of SIMD instructions to simultaneously trace multiple rays through a KD tree, which made interactive raytracing feasible on multi-processor desktop machines. Subsequent work [Dmitriev et al. 2004; Reshetov et al. 2005; Boulos et al. 2006; Wald et al. 2007] yielded a rich class of packet tracing techniques, whereby coherent groups of rays are simultaneously traversed through an acceleration structure. In addition to the use of SIMD instructions to vectorize calculations, current techniques employ packet-level tests to eliminate calculations, which are often most effective when applied to packets much larger than the SIMD width.

Whereas packet tracing seeks to amortize or cull ray-level operations, a more recent idea is to cull packet-level calculations using a pre-pass through the tree. This was introduced in the form of entry point search, (EP search) [Reshetov et al. 2005], and the basic idea is to start traversal at a node other than the root, thus traversing fewer nodes. Our work began as an attempt to apply this idea to BVH data structures, but has resulted in an entirely different, and more effective algorithm.

2 Prior work

Although KD-trees [Fussell et al. 1988] were the preferred raytracing data structure in earlier times, the bounding volume hierarchy or BVH [Rubin and Whitted 1980] has become increasingly popular. Recent results demonstrate that BVH packet tracing achieves comparable performance to a KD-tree, and is much more flexible in its ability to handle dynamic scenes [Wald et al. 2007; Lauterbach et al. 2006; Yoon et al. 2007]. Packet traversal schemes for BVHs were presented by Wald et al. [2007] and Lauterbach et al. [2006] at roughly the same time. Although BVH traversal requires more computation at each node, the trees generally are shorter and contain fewer nodes than the equivalent KD trees. A key drawback of a BVH is that the possibility of overlapping bounding volumes makes it impossible to guarantee that the first intersection point that is found is the first one struck by the ray.

Reshetov et al. [2005] introduced entry point (EP) search, which is a very important precursor to our work. EP search uses the bounding frusta for large groups of rays in order to locate an entry point node, which is then used as a starting point for tracing the rays in the tile. The node selected as the entry point is the lowest common ancestor of all (non-empty) leaf nodes which are intersected by the bounding frustum. By starting traversal at this node, traversal calculations on its ancestors can be avoided.

Quite recently, Fowler et al. [2009] proposed a modification to EP search in which the candidate entry points are collected into a queue and then tested in top-down, rather than bottom-up order. This re-
At each node, we first perform a frustum test against the node’s candidate evaluation inspired by Fowler [2009]. EP search, for reasons which will soon be made clear. (For actual practice, we are deliberately describing a suboptimal variant of the formulation of EP search. Besides being recursive, this formulation is natural to begin by making it worse. Figure 3 shows a recursive traversal, and which do not explicitly represent the empty space inside of hollow objects.

In the absence of reliable occlusion detection, all nodes which intersect the frustum must be considered when choosing the entry point. This severely compromises the effectiveness of EP search, because it tends to confine the set of possible entry points to the uppermost levels of the tree. Previous work [Benthin 2006; Fowler et al. 2009] has observed this effect at work in complex scenes. In the worst case, any frustum which includes geometry on both sides of the topmost splitting plane will never benefit from EP search, despite incurring the overhead of a tree walk. Unfortunately, this worst case can be very common, as it occurs whenever the viewer looks towards the center of the model.

Even when occlusion can be considered, there will still exist reasonable scenarios in which an EP search yields no benefit. This can happen if rays in the frustum strike geometry on opposite sides of the root splitting plane. A hypothetical example is shown in Figure 2. In this example, both sides of the root node contain intersected geometry, which forces the root to be chosen as the entry point.

3.2 Re-Thinking Entry Point Search

Paradoxically, in our effort to improve upon EP search, it is most natural to begin by making it worse. Figure 3 shows a recursive formulation of EP search. Besides being recursive, this formulation will visit many more nodes than necessary, and should NOT be used in practice. We are deliberately describing a suboptimal variant of EP search, for reasons which will soon be made clear. (For actual implementation, we follow Reshetov et al. [2005] with a top-down candidate evaluation inspired by Fowler [2009]).

At each node, we first perform a frustum test against the node’s bounding box, returning a NULL in the event of a miss. If the node is hit, and it is a leaf, it is returned as an entry point. Otherwise, a search is performed recursively on each subtree. If neither subtree contains an entry point, a NULL will be returned. If only one subtree contains an entry point, then it is returned. If both subtrees contain entry points, then the parent node is returned, as it is their lowest common ancestor. Rather than first filling a bifurcation stack and evaluating the nodes it contains, this approach simply traverses the tree and propagates entry point nodes up the call stack.

As an aside, we note that one can stop traversing immediately if a node’s bounding volume is completely contained within the frustum. In this case, it is not necessary to traverse all the way to the leaves, because we know that every leaf node will eventually intersect the frustum. A test for this case is given by [Assarsson and Moller 2000] and it may be vectorized as in [Reshetov et al. 2005]. We have not seen this test used in prior work, but we found that it gives a notable improvement in our EP search implementation. Consider what happens when only one subtree contains an entry point. This entry point is propagated up the call stack, bypassing its ancestors. This will continue to occur until another entry point is located in a different subtree, at which time their common ancestor will replace them as the entry point. These common ancestors are a superset of the bifurcation nodes used in iterative EP search. If we examine the set of nodes visited by a traversal of the tree (as illustrated in Figure 4), what we will generally find is a small number of bifurcating nodes which contain reachable leaves in both
subtrees, separated by groups of bypassed nodes, which have reachable leaves in only one subtree. Each bypassed node has at most one bifurcating ancestor and descendant, and there are generally many more bypassed nodes than bifurcating ones.

### 3.3 Path Compression

The key insight behind our work is that the bypassed nodes are essentially dead weight. For rays in the frustum, only one child of a bypassed node will ever contain any intersections. Furthermore, any ray which hits the bypassed node will also hit its bifurcating parent, and any ray which misses it will also miss its bifurcating descendant. For rays in the frustum, it is therefore possible to traverse directly from the node’s bifurcating ancestor to its bifurcating descendant without compromising correctness. The bypassed nodes in the middle do not need to be tested. In fact, they do not even need to be accessed.

This observation leads us to our new algorithm. Rather than simply locating the highest possible entry point, we instead construct a temporary BVH on the fly, optimized for a particular frustum. This BVH is a subset of the original which stores only the bifurcating nodes, with direct links to their bifurcating descendants. The root of this BVH is the same node that would have been returned by an EP search, and its leaves correspond either to leaves in the original tree, or to inner nodes whose bounding volumes are fully contained by the frustum. The paths from root to leaves in this tree are shorter versions of the corresponding paths in the original tree. We call this technique path compression (overloading the name of a classical optimization for union-find data structures [Cormen et al. 2001]). We refer to the resulting tree as a path compression tree (PC tree).

PC tree construction is surprisingly easy to implement. It can be done using a simple modification to the recursive EP search, as shown in Figure 5. At each bifurcating node, rather than discarding the entry points found in the subtrees, we instead return a new PC node, with the two entry points as its children. For bypassed nodes, there is at most one entry point, so we simply propagate the result upwards. An iterative implementation of path compression is also possible, and is considerably more complex. We briefly outline the iterative implementation in Appendix A.

Once we have constructed a PC tree for a given frustum, we can raytrace against it by traversing it in the same manner as an ordinary BVH, using any traversal algorithm. When a leaf is reached in the PC tree, we begin traversing the original data structure, starting with the node indicated by the PC leaf. Note that path compression can, in principle be applied to other data structures than BVHs. In this case, the original data structure is simply overlaid with a BVH.

### 3.4 Memory Layout and Management

If the original data structure is a BVH, the PC nodes need only store a pointer to the original BVH node, along with a pair of child links. The bounding volume of the original node can be re-used for the PC node. For other data structures such as KD trees, path compression can be applied, but the bounding boxes of the bifurcation nodes must be tracked during PC tree construction and stored explicitly in the PC nodes.

For high performance, it is critical to use an efficient memory management strategy for PC tree nodes. All nodes in our implementation are stored in one contiguous array, and are assigned in sequential order, as needed. A count is maintained of the number of nodes in use. Because each tree is assigned a contiguous range of the array, deallocation of an entire tree can be done by simply decrementing the node count (provided the trees are destroyed in reverse order of construction). If the array becomes full during PC tree construction, a new one is allocated, and used nodes are copied. The array size is doubled with each re-allocation, and eventually stabilizes. In order to enable this reallocation, it is best to store all node references as relative offsets, instead of pointers. This memory management scheme all but eliminates allocation overhead, and it also ensures good cache behavior for the PC nodes.

### 3.5 Hierarchical Path Compression

Once a PC tree has been built for a particular bounding frustum, it is possible to refine it further to suit a subset of the frustum. To refine an existing PC tree, we can simply apply the same algorithm to the PC tree, creating a new tree which is a subset of the previous one. If a leaf is reached during PC refinement, path compression should be re-run on the original BVH, beginning with the node indicated by the leaf. Note that the test for fully intersected nodes can be omitted when testing inner nodes in the PC tree. Since these nodes

![Figure 4: Sample tree topology. Thin lines indicate parent-child links in the BVH. Grey circles indicate visited leaves. Grey squares indicate bifurcation nodes. Hollow circles indicate bypassed nodes. Bold arrows indicate parent-child links resulting from path compression.](image-url)
were not fully contained in the original frustum, they cannot be fully contained in a subset of it.

Refinement can be done as often as desired, but it eventually becomes counter-productive. In our experiments with primary rays, we tried numerous permutations of initial tile count and splitting factor. We achieved the best results for 1024x1024 pixel images by building one PC tree for the entire scene, and refining it by splitting into an 8x8 tile grid (256x256 pixels per tile). Each of the 8x8 “super-tiles” are then split directly into 16x16 pixel ray packets (which we found to be optimal in our implementation). Different image resolutions or traversal schemes will naturally benefit from different tiling strategies, but an 8x8 grid of super-tiles seems to be a good baseline.

4.1 Traversal Reduction

We begin by examining the number of node traversals required to trace primary rays in each scene (Table 1). Here we count the number of times a node is visited when tracing a ray packet, as well as the number of nodes accessed by EP search or path compression. As can be seen, PC tree construction will tend to visit many more nodes than a corresponding EP search would, but this is more than paid for by the reduction in node visits during raytracing. Neither path compression nor EP search is particularly effective for the dragon scene, since the camera is often viewing the entire model from a distance, and most leaves are reachable. Note that for the dragon, EP search is counter-productive.

4.2 Performance

Table 2 gives raw performance measurements. The numbers presented are pure raycasting rates only, and do not include shading or display. We see that path compression can yield speedups as high as 88 percent, and that speedups of over 20 percent are routine. In contrast, we see that simple EP search, while it is occasionally beneficial, is generally of little value.

While our speedups are correlated with the reductions in node traversals, their magnitudes are not very high. This is due to the effect of Amdahl’s law. The maximum possible speedup is limited by the fraction of time spent in BVH traversal, which in our case is only about 50%. When path compression is used, this figure drops to about 34%.

Profiling indicates that the vast majority of our time is still spent on ray packet traversal and intersection testing. PC construction and EP search are done quite infrequently, and thus add very little overhead. In more distant viewpoints, where PC provides no benefit, it also tends to terminate very quickly. Our implementation of EP search exhibits the same characteristics.

We conducted our experiments using both recursive and iterative implementations of path compression. We found that while the iterative variant executes about 20% faster than the recursive, the relative cost of path compression is so small that there is no significant change in frame time.

4.3 Memory Consumption

An important question for our work is the memory footprint of the PC tree. In Table 3, we present the final sizes of the PC node pools for each scene, as well as the maximum and average memory requirements (that is, the amount of memory in the pool that was actually in use on any particular frame). The data are presented relative to the size of each scene’s BVH, to help place our overhead in context. We find that the amount of memory consumed by the PC tree is typically very small, relative to the size of the main data structure. In addition, memory consumption remains stable across a variety of scene sizes.

There are two principle reasons why our memory consumption is so low. The first is the relative scarcity of bifurcating nodes, and the second is the fact that we terminate PC tree construction when an entire node is contained in the frustum. This early termination allows us to avoid the densely populated lower levels, and is important in keeping our memory consumption under control.

4.4 Partition Traversal

Overbeck et al. [2008] presented an alternative ray traversal technique for BVHs, which is more effective for incoherent rays. This approach divides the rays in a packet into subsets at each visited node, rather than using early-descent. This trades additional work at the top portion of the tree for a reduction in work at the lower levels, because missed rays are prevented from being ‘dragged’ to lower levels of the tree by rays which hit in the higher levels. We experimented briefly with partition traversal, and found that path compression yields a much higher performance gain in this case. Average performance and speedup are shown in Table 4.

Interestingly, we found that because partition traversal does so much more work per node, it favors a more extreme approach. In addition to the hierarchical tiling used in previous experiments, we found that it was also helpful to construct a PC tree for each individual ray packet prior to tracing it, and to omit the frustum test during packet tracing. We can reasonably expect to see similar results in a
Table 1: Average traversals, in thousands. Numbers in parenthesis are nodes visited during EP search or PC tree construction.

<table>
<thead>
<tr>
<th></th>
<th>Traversals</th>
<th>Traversals(EP)</th>
<th>Traversals(PC)</th>
<th>EP Ratio</th>
<th>PC Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dragon</td>
<td>357</td>
<td>356 (2.9)</td>
<td>277 (55)</td>
<td>1.004</td>
<td>0.94</td>
</tr>
<tr>
<td>City</td>
<td>199</td>
<td>177 (2.5)</td>
<td>56 (5.7)</td>
<td>0.99</td>
<td>0.31</td>
</tr>
<tr>
<td>Kitchen</td>
<td>273</td>
<td>267 (2.2)</td>
<td>88 (12.9)</td>
<td>0.98</td>
<td>0.36</td>
</tr>
<tr>
<td>Outbound</td>
<td>394</td>
<td>374 (5.3)</td>
<td>147 (29.2)</td>
<td>0.96</td>
<td>0.44</td>
</tr>
</tbody>
</table>

Table 2: Average packet tracing rates (MRays/s), and speedups. Speedups in parenthesis are the highest observed for a given scene.

<table>
<thead>
<tr>
<th></th>
<th>Reference</th>
<th>EP</th>
<th>Speedup (Max)</th>
<th>PC</th>
<th>Speedup (Max)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dragon</td>
<td>5.43</td>
<td>5.44</td>
<td>5.30 (1.06)</td>
<td>5.44</td>
<td>5.50 (1.08)</td>
</tr>
<tr>
<td>City</td>
<td>8.81</td>
<td>9.01</td>
<td>11.25 (1.08)</td>
<td>9.01</td>
<td>12.8 (1.46)</td>
</tr>
<tr>
<td>Kitchen</td>
<td>7.78</td>
<td>7.91</td>
<td>9.57 (1.23)</td>
<td>7.91</td>
<td>12.3 (1.51)</td>
</tr>
<tr>
<td>Outbound</td>
<td>6.02</td>
<td>6.04</td>
<td>8.30 (1.05)</td>
<td>6.04</td>
<td>13.8 (1.88)</td>
</tr>
</tbody>
</table>

Table 3: Memory consumption. PC Tree sizes are given as percentages of BVH size

<table>
<thead>
<tr>
<th></th>
<th>Triangles(Thousands)</th>
<th>BVH Size (MB)</th>
<th>Avg. PC Bytes (%)</th>
<th>Max PC Bytes (%)</th>
<th>Pool Size (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dragon</td>
<td>871</td>
<td>12.97</td>
<td>0.52</td>
<td>0.98</td>
<td>1.44</td>
</tr>
<tr>
<td>City</td>
<td>9.2</td>
<td>0.12</td>
<td>4.11</td>
<td>8.72</td>
<td>9.85</td>
</tr>
<tr>
<td>Kitchen</td>
<td>103</td>
<td>1.38</td>
<td>1.30</td>
<td>3.23</td>
<td>3.39</td>
</tr>
<tr>
<td>Outbound</td>
<td>644</td>
<td>7.97</td>
<td>0.35</td>
<td>1.22</td>
<td>2.35</td>
</tr>
</tbody>
</table>

Table 4: Average packet tracing rates (MRays/s), and speedups for partition traversal.

<table>
<thead>
<tr>
<th></th>
<th>Partition</th>
<th>Partition(PC)</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dragon</td>
<td>3.63</td>
<td>4.66</td>
<td>1.23</td>
</tr>
<tr>
<td>City</td>
<td>4.18</td>
<td>8.58</td>
<td>2.05</td>
</tr>
<tr>
<td>Kitchen</td>
<td>3.49</td>
<td>7.53</td>
<td>2.16</td>
</tr>
<tr>
<td>Outbound</td>
<td>2.44</td>
<td>5.86</td>
<td>2.40</td>
</tr>
</tbody>
</table>

5 Future Work

5.1 Other Data Structures

As we have already indicated, path compression, suitably modified, is compatible with other raytracing data structures than BVHs. KD trees or bounding interval hierarchies [Wachter and Keller 2006] are both binary trees in which there is an implicit bounding box for every node. Path compression can thus be applied to these data structures as well, and a comparison across the various data structures would be interesting to study.

5.2 Shadow Rays

While we have shown PC to be beneficial for primary rays, this is arguably an unconvincing case. It would be desirable to examine ways that this approach could be applied for secondary rays. We conducted brief experiments with area light sources in the kitchen scene, firing one large ray packet at the area light for each rendered pixel. We had expected to see a performance benefit by performing a PC pass prior to shooting each ray packet, but this turned out to be counterproductive. It is clear from these results that path compression or EP search must be amortized over multiple traversals of the tree in order to be effective, which concurs with a prediction made in [Overbeck et al. 2007].

Despite these negative results, there are certain scenarios in which path compression might be beneficial for shadow rays. For very localized light sources, such as spot lights or point lights with falloff, it might be useful to perform path compression using the bounding volume of the light’s region of influence (compressing away any nodes which do not lie inside). In this way, a tree could be computed once per frame and re-used for all shadow rays cast from the light. For shadows which are directly visible from the camera, the view-frustum partitioning methods which are used for shadow maps [Zhang et al. 2006] could also be used to build path compression trees.

5.3 Applications Beyond Raytracing

We find it exciting to consider that our ideas could be generalized to other algorithms which perform searches on tree structures. Any problem which involves repeated, coherent queries against a spatial data structure might benefit from a variant of our path compression technique. For example, one could imagine a photon mapping implementation which batches photon lookups into spatially coherent groups, and uses a loose bounding volume of the query points to compress a KD tree of photon positions. Applications could also be found in areas outside rendering, such as collision detection or database search.

6 Acknowledgements

The dragon model was obtained from the Stanford 3D Scanning Repository. The kitchen and city are from [Lext et al. 2001]. Outbound was created by students from NHTV Breda University of Applied Sciences, and is available at: http://igad.nhtv.nl/bikker/downloads.htm.
References


A Iterative Path Compression

Here we briefly outline an iterative algorithm for building path compression trees. The iterative algorithm uses a bifurcation stack similar to that used in EP search. Each entry in the stack contains a pointer to a BVH node, and space for a pair of PC node references. The algorithm possesses two distinct phases: A traversal phase, which walks downwards through a section of the tree, and a stack contraction phase, which incrementally builds the PC tree from the bottom up. The algorithm begins in the traversal phase, and switches to stack contraction whenever a change is made to the PC tree.

During the traversal phase, stack entries may be added or removed, and PC nodes may be created and linked to the topmost node in the stack. There are several cases to consider:

- When a leaf node or fully enclosed inner node is reached, a PC node is created. If the stack is empty, then this PC node is the only one in the PC tree, and it is returned. Otherwise, the PC node is linked to the topmost stack entry, and stack contraction occurs.
- Otherwise, if exactly one child node is hit by the frustum, traversal proceeds to that child.
- Otherwise, if both child nodes are hit by the frustum, the node is tentatively considered a bifurcator, and a new entry is pushed onto the stack. Traversal proceeds to the left child.
- Otherwise, if neither child is hit, it means that the subtree being traversed does not contain reachable leaves. This means that the node at the top of the stack is not a "true" bifurcator, so it is popped from the stack, and its linked PC tree (if present) is moved up to the next stack entry. If the stack becomes empty, this linked PC tree is returned immediately, otherwise, stack contraction occurs.

Following any change to the stack, the algorithm enters a stack contraction phase. If the topmost entry in the bifurcation stack possesses two linked PC trees, then these trees are parented to a new PC node containing the BVH node from atop the stack. The stack is then popped, and the resulting PC tree is linked to the new stack top. This continues until one of the following conditions occurs:

- **The stack is empty.** In this case, the completed PC tree is returned.
- **There is only one PC tree.** In this case, the algorithm returns to the traversal phase, beginning at the right child of the node atop the stack.