An Architecture for Hierarchical Collision Detection

Gabriel Zachmann
Computer Graphics, Informatik II
University of Bonn
e-mail: zach@cs.uni-bonn.de

Günter Knittel
WSI/GRIS
University of Tübingen
e-mail: knittel@gris.uni-tuebingen.de

Abstract

We present novel algorithms for efficient hierarchical collision detection and propose a hardware architecture for a single-chip accelerator. We use a hierarchy of bounding volumes defined by k-DOPs for maximum performance. A new hierarchy traversal algorithm and an optimized triangle-triangle intersection test reduce bandwidth and computation costs. The resulting hardware architecture can process two object hierarchies and identify intersecting triangles autonomously at high speed. Real-time collision detection of complex objects at rates required by force-feedback and physically-based simulations can be achieved.

Keywords: graphics hardware, computer animation, virtual reality, hierarchical algorithms, triangle intersection.

1 Introduction

Collision detection is an elementary task in areas like animation systems, virtual reality, games, physically-based simulation, automatic path finding, virtual assembly simulation, and medical training and planning systems.

In many of these systems, collision avoidance is the ultimate goal. Since algorithms for computing the exact time of collision are still too slow or too restrictive, most approaches are “reactive” in that they first try to place objects at a new position, then check for collision, and then try other positions, based on physical laws or constraints [14, 21]. This poses very high demands on collision detection performance, because they must do many collision checks per simulation cycle. Another very demanding application is rendering force-feedback, where collisions of an (invisible) surface contact object must be checked at about 1000Hz in order to achieve stable force computations.

Since collision detection is such a fundamental task, it would be highly desirable to have hardware acceleration available just like 3D graphics accelerators. Using specialized hardware, general-purpose processors can be freed from computing collisions.

In this paper, we propose an architecture which implements hierarchical collision detection for rigid objects in hardware. We have concentrated on hierarchical algorithms, because they have offered the best performance for so-called “polygon soups”. Such a collision detection hardware will comprise the last stage of a collision detection pipeline [20]. This is where the bulk of the work is done in typical scenarios involving a modest number of objects with large polygon counts. We assume the hierarchies have already been computed. This is not a time-critical task, and can be done in software when the application loads objects at startup time.

The next section describes related work, while Section 3 describes novel algorithms that are suitable for hardware implementation. Section 4 describes the hardware design in detail. Finally, Section 5 presents some benchmarks and considerations about the performance of the envisioned architecture. An extended version of this paper is available at www.gabrielzachmann.org.

2 Related Work

Considerable work has been done on hierarchical collision detection in software [4–6, 17, 19]. Some of the bounding volumes (BVs) utilized are spheres, axis-aligned bounding boxes (AABB), oriented bounding boxes (OBB), and discretely oriented polytopes (DOP). However, all traversal schemes proposed so far are inefficient in that they possibly visit the same nodes many times.

There is basically no literature about the design of hardware architectures dedicated to collision detection. All research so far has tried to utilize existing
3 The Algorithm

3.1 DOP Trees

The basic operation of any hierarchical collision detection algorithm is the overlap check of two nodes from different objects. In this section, we briefly recall the calculations necessary for collision detection using DOP trees. The derivation of the following formulas can be found in [19].

DOPs are bounding volumes that are a generalization of axis-aligned bounding boxes. They have been introduced into computer graphics by [8]. DOP trees are a hierarchical representation of objects [9, 19]. Each inner node stores a DOP and pointers to its children which it encloses; leaves store polygons (or other graphical primitives). A DOP is described by \( k \) numbers (hence \( k \)-DOP), usually represented by a vector of \( k \) floats. Extensive benchmarks have shown \( k = 24 \) to be optimal.

Given two objects \( O_A \) and \( O_B \), and two DOPs \( d, e \in \mathbb{R}^k \) from \( O_A \) and \( O_B \)'s DOP trees, resp., the overlap test proceeds in two steps: first, DOP \( d \) from \( O_A \)'s hierarchy is transformed into \( d' \) in the coordinate frame of \( O_B \) by

\[
d' = C \times d + c, \tag{1}
\]

where

\[
C = \begin{pmatrix}
\ldots & c_{0,0} & \ldots & c_{0,1} & \ldots & c_{0,2} & \ldots \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\
\ldots & c_{k-1,0} & \ldots & c_{k-1,1} & \ldots & c_{k-1,2} & \ldots \\
\end{pmatrix}
\]

in matrix \( C \) exactly three entries per row are non-zero. Second, \( d' \) is compared componentwise with DOP \( e \) according to

\[
\exists i \leq \frac{k}{2} : d'_{i} > e_{\frac{k}{2}+i} \lor e_{i} > d'_{\frac{k}{2}+i} \iff d \text{ and } e \text{ do not overlap} \tag{2}
\]

where \( d'_{i} < d'_{\frac{k}{2}+i} \) define a slab (analogously for all DOPs).

Matrix \( C \) and vector \( c \) depend only on the position of the two objects relative to each other. They are computed during the set-up by the software API of the collision detection hardware.

Since the \( k \times k \)-matrix \( C \) in Equation 1 has exactly 3 coefficients per row that are not 0, we can compute \( d' \) more efficiently by

\[
d'_{j} = c_{j} + \begin{pmatrix} d_{i,0} & d_{i,1} & d_{i,2} \end{pmatrix} \tag{3}
\]

where correspondence \( j \) stores the place of those coefficients which are not zero. So, by introducing a \( k \times 3 \) correspondence matrix \( i \), we can reduce the size of the transformation matrix \( C \) to \( k \times 3 \). Consequently, the number of multiplications is \( 3k \).

3.2 Hierarchy Traversal

The general, traditional scheme for hierarchical collision detection is a simultaneous, recursive traversal of two BV hierarchies. However, this procedure incurs several penalties:

1. Nodes in both trees are usually visited several times; this is a general problem of all hierarchical collision detection algorithms (see Figure 1).

2. If the nodes have to be transformed (or other computations specific to individual nodes have to be performed), then this will be done several times for the same node.

The second penalty is a consequence of the first one; it could be alleviated by storing the result of the node transformation back into the node. Unfortunately, this has other disadvantages: first, the BV hierarchy occupies more memory (in the case of DOP trees, this would increase the memory usage by a factor 2); second, more importantly, the algorithm would no longer be thread-safe, so that multiple pairs of objects could no longer be checked in parallel.
In contrast, our novel traversal scheme reduces the number of nodes visited, transfer volume from memory, and number of node transformations dramatically. Our traversal scheme only needs an additional small stack.

The idea is to avoid simultaneous traversal of two BV hierarchies. Instead, we traverse only one hierarchy and compare each node of that one with a list of nodes from the other hierarchy. Let us call nodes that need to be transformed tumbled nodes, the other ones aligned nodes (see Figure 1). Assume that we are visiting a tumbled node A, and that a list \( \mathcal{L} \) contains all aligned nodes with which A needs to be checked for overlap. So we check all pairs \((A, \mathcal{L}^1)\); whenever such a pair overlaps, we append the two children \( \mathcal{L}_i^1, j \in [1, 2] \), to a new list \( \mathcal{L}' \). After \( \mathcal{L} \) has been completely processed, \( \mathcal{L}' \) contains all aligned nodes that need to be checked with \( A_1 \) and \( A_2 \), the two children of A. It is obvious that with this traversal we visit each tumbled node only once, and thus we transform the DOP stored with it exactly once.

This scheme works for all kinds of hierarchical collision detection, not just DOP trees. Depending on how much work per node-node overlap test can be factored out into one of the two nodes, the benefit of our new method can be dramatic.

A hardware implementation allows us to improve the algorithm further by performing DOP overlap tests in parallel. By the nature of the binary tree, performing two overlap tests in parallel yields the greatest cost/performance benefit. To this end, we load a sibling pair of tumbled DOPs \((A, B)\), transform them sequentially, and compare the two in parallel with each DOP from \( \mathcal{L} \). This results in two new lists, one for child pair \((A_1, A_2)\) and one for \((B_1, B_2)\). In the sequential version described in the previous paragraph, we produced these two lists at very different times during the traversal, and we processed each of them twice; now, we produce those two lists simultaneously, and then we process each of them only once. The benefit of this is that the time needed for overlap tests and the number of times an axis-aligned DOP needs to be transferred from memory is cut by a factor of two. Note that, for clarity, we have omitted the “mixed” cases.

In a hardware implementation, we have to maintain the stack and the lists ourselves. This can be done by a stack of lists (see Figure 2). On the same stack, we keep pointers to pairs of tumbled nodes. Going down from node pair \((A, B)\) to \((A_1, A_2)\), we push the pointer to \((A, B)\) onto the stack. Later, when the recursion returns to this node pair, we need to decide whether to go down into node pair \((B_1, B_2)\) or to make a step upwards. This information can be kept in an additional bit on the stack: when the pointer is pushed onto the stack, the corresponding bit is reset; when we return to this node, we go down into the other branch and flip the bit to 1. When we return the next time, the algorithm knows to make another step upwards.

### 3.3 Polygon Intersection Test

In the case of collision, the traversal reaches pairs of leaves containing triangles, which have to be checked for intersection. Assume triangle A is given by vertices \( V^1, V^2, V^3 \) and triangle B is given by vertices \( W^1, W^2, W^3 \), both in their object’s reference frame. Assume triangle A is part of object \( O_A \), and B is part of \( O_B \).

The approach in our algorithm is to check (conceptually) each edge of A against B, and vice versa. First, A’s vertices are transformed into the reference frame of \( O_B \). Assume further a \( 3 \times 3 \) transformation \( M_B \) for triangle B such that \( M_B \cdot (W^i - W^j) \) maps onto the unit triangle \( (0, 0, 0), (1, 0, 0), (0, 1, 0) \). Then, we transform A by \( M_B \cdot W^j \) (see Figure 3).

---

1. This scheme can be generalized straight-forward to process \( 2^m \) tumbled nodes simultaneously.
For sake of simplicity, we will call the new vertices $V_i$ again.

Consider each edge $PQ := V_iV_{i+1}$. If both $P_z$ and $Q_z \geq 0$ or $\leq 0$, then we skip this edge. Now we compute (conceptually) the intersection $S$ of the supporting line $X = P + tr$, $r = Q - P$, with the plane $z = 0$, which is defined by $t = -P_z/r_z$ as $S = P - r_Pz$ (we know $r_z \neq 0$). We know that $0 \leq t \leq 1$. We also know that $S_x = 0$, so we need to compute only $S_x = P_x - r_xP_z$, and similarly $S_y$, which are, basically, the barycentric coordinates of the intersection point. Finally, we just check whether or not $S_x \geq 0 \land S_y \geq 0 \land S_x + S_y \leq 1$. If so, $A$ and $B$ do intersect; otherwise, we check the other edges, and, in case of no intersection, we check $B$ against $A$.

In order to save the division and the vector subtraction (for $r$), we reformulate the condition as follows (assuming $r_z > 0$):

$$P_xQ_z \geq Q_xP_z \quad \land \quad P_yQ_z \geq Q_yP_z \quad \land \quad (4)$$

If $r_z < 0$, then we must compare with $\leq 0$, and $\geq 0$, respectively.

The algorithm gains its special efficiency because we can precompute the matrices $M_A$ and $M_B$ (they can be obtained from a simple linear equation system), and because we do not need to compute the exact intersection point.

In our case of collision detection using DOP trees, we can store these matrices in the leaves instead of the DOPs. We do not need to check pairs of leaf DOPs, because the immediate check of triangles is faster. Storing the triangle matrix $M_B$ and 3 vertices needs $3 \times 4 + 3 \times 3 = 21$ floats, which fit well into the nodes of a 24-DOP tree.

4 Hardware Design

The target design is a PCI-board with one ASIC, a large on-board memory for the hierarchy, and a small SRAM as dedicated stack memory. Crucial for the performance is the bandwidth towards the local memory, and so a four-bank SDRAM configuration with a 256-bit bus was chosen (see Figure 4).

4.1 The CollisionChip

Figure 5 shows all functional units of the CollisionChip. It consists of a number of large register files grouped around an arithmetic unit for floating-point dot-products (the DOTADD-unit), a Triangle Intersection Test Unit (IT-Unit), register banks connected to comparators, interfaces to the PCI-bus and to the local memories, the Stack Engine and control units as well as address generators. Although the processing of bounding volumes and triangles differ quite substantially, a common architecture was found with only low redundancy.

**The DOTADD-Unit.** The DOTADD-unit is similar to transform units as found in modern graphics accelerators. Its basic function is to perform

$$d'_i = d_k \times C_{i,0} + d_m \times C_{i,1} + d_n \times C_{i,2} + c_i$$

on 32-bit floating-point numbers. The indices refer to the location in the register files. Due to the absence of data dependencies in the control flow, it can be pipelined for high clock frequency and throughput.
Figure 5: All functional units of the CollisionChip.
Processing of Bounding Volumes. Prior to the processing of two hierarchies, the matrix C and the coefficients c must be loaded into the Matrix Register File. Also, the correspondence indices i, k, m and n must be stored in the Correspondence Register File. This happens via the PCI-bus under software control, and occupies 24 lines in both register files. The software also transmits the pointers (local memory addresses) of the two root nodes as starting point to the Master Controller.

A DOP from the tumbled object is loaded from memory and sequentially stored in the DOP Register File under control of Address Generator 1. After some constant delay (again predetermined by software) a sufficiently large subset of DOP-elements d are or will become available in time for continuous evaluation of Equation 1. At this point in time, Address Generator 2 is triggered and the operands are fed into the DOTADD-unit. Note that for maximum performance, processing of the lines in matrix C occurs out-of-order, depending on the earliest availability of the required d-elements. Also note that a specific d may be used for more than one line. The transformed DOP-elements d' are then stored into the Results Register Bank under control of Address Generator 3, which basically delays an index i for the duration of the pipeline delay of the DOTADD-unit. The same processing is applied to the sibling of the tumbled node.

The DOP-elements e of the aligned node are then loaded from memory in three transfers and stored in the registers labeled e0 . . . e23 under control of the Master Controller. The bank of comparators determines overlap in parallel and signals this condition to the Master Controller. The lists of child nodes to be checked for overlap are constructed according to this condition by the Stack Engine.

Hierarchy Traversal. The novel traversal algorithm as described in Section 3.2 is implemented using a dedicated and fast external SRAM to store the lists and a suitably designed Stack Engine. Its basic task is to hand node pointers from the current list to the Master Controller and to receive child node pointers to construct new lists. Internally, it maintains a stack of list pointers and a register containing the actual level.

Processing of Triangles. As described in Sect. 3.3, testing triangle A from object O_A against triangle B from object O_B requires transforming A into the coordinate system of O_B using a rotation and a translation. These are constant for two objects, and can therefore be precomputed. The reverse test B against A may also be necessary, which requires the inverse transform. For maximum performance, all coefficients are kept on chip in the Matrix Register File in lines 24 through 29. After this first transformation, the triangle must then be transformed using the matrix stored in the leaf of the other triangle, whose coefficients are loaded into lines 30 through 32 of the Matrix Register File. The Correspondence Register File has 27 lines dedicated to these transforms, and so they can be performed using the DOTADD-unit by properly setting up the Address Generators.

The concrete sequence of operations is as follows. The vertices of A are loaded into the DOP Registers d_0 . . . d_8, transformed and written back into DOP Registers d_9 . . . d_17. Meanwhile, the coefficients from the other leaf are loaded from memory. The second transformation of A leaves the vertices in registers d_0 . . . d_8 again.

Further processing according to Section 3.3 and Equation 4 is then done in a separate unit called IT-Unit. This unit can have low-performance and thus low-cost arithmetic units, since triangle tests are not performed very frequently (see Table 1). This unit is controlled by the Master Controller via a 7-bit bus and supplied with operands from the DOP Register File using indices from the Correspondence Register File. Five pairs of operands need to be read per edge, which requires additional 15 entries for a total of 66 lines in the Correspondence Register File. The accumulation circuitry consisting of the ADD/SUB unit and the register ACC compute the left side of the third condition in Equation 4. The combinational circuitry to the left determines whether the edge cuts the z = 0 plane at all.

An intersection is reported to the software; otherwise, the above procedure is repeated with operands reversed.

5 Performance Estimations

Processing of a given list involves reading and transforming two tumbled nodes, and reading and comparing the appropriate number of aligned node pairs. We assume that throughputs are limited by transformation performance and memory bandwidth; the stack engine is assumed to be always fast enough. We also don't consider triangle-triangle-tests here since they don't occur very often.

Further assumptions are as follows: nodes are defined by 24 single-precision floating-point numbers plus auxiliary data, placed in memory on 128-byte boundaries. The memory is build from DDR-SDRAM chips with a 2-2-2 access characteristic (2 cycles each for the precharge time, RAS-CAS-delay, and CAS-latency). The CollisionChip is assumed to run at the data burst frequency, e.g. 266MHz for PC133 memory chips. A cycle of the CollisionChip equals one half of a memory cycle. The SDRAM Interface can buffer an entire node pair (256 bytes) and thus allows a burst length of eight to be used.
Table 1: This table shows the performance of our hardware architecture for a number of objects that are placed in close proximity. All times are in microseconds. The average numbers have been obtained by rotating one of the objects relative to the other. The worst-case numbers are the respective maxima observed during that rotation. The collision detection times have been calculated with Equation 5, assuming 3.76 nsec/cycle ($\alpha = \text{num. aligned nodes}/\text{num. tumbled nodes}$, $\tau = \text{num. tumbled nodes}/2$). Columns marked by (1) count multiple visits of aligned nodes, while those marked by (2) count the number of unique tumbled nodes (which are, unlike traditional traversal schemes, visited only once). The software performance has been measured using the traditional recursive hierarchy traversal.

<table>
<thead>
<tr>
<th>Object</th>
<th>num nodes</th>
<th>aligned nodes visited$^1$</th>
<th>tumbled nodes visited$^2$</th>
<th>time in SW</th>
<th>aligned nodes visited$^1$</th>
<th>tumbled nodes visited$^2$</th>
<th>time in HW</th>
<th>speedup avg/worst-case</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filter</td>
<td>19326</td>
<td>12474</td>
<td>240</td>
<td>1660</td>
<td>153</td>
<td>14936</td>
<td>542888</td>
<td>717164 / 98 / 127</td>
</tr>
<tr>
<td>Frontlight</td>
<td>30705</td>
<td>389</td>
<td>73</td>
<td>68</td>
<td>15</td>
<td>524</td>
<td>7065</td>
<td>937 / 207 / 36 / 44</td>
</tr>
<tr>
<td>Lock</td>
<td>62023</td>
<td>279</td>
<td>81</td>
<td>9</td>
<td>15</td>
<td>401</td>
<td>4854</td>
<td>877 / 178 / 27 / 38</td>
</tr>
<tr>
<td>Car body</td>
<td>60755</td>
<td>259</td>
<td>66</td>
<td>55</td>
<td>12</td>
<td>383</td>
<td>3076</td>
<td>538 / 110 / 31 / 49</td>
</tr>
<tr>
<td>Buddha</td>
<td>125000</td>
<td>139</td>
<td>30</td>
<td>7</td>
<td>9</td>
<td>240</td>
<td>3345</td>
<td>301 / 77 / 40 / 74</td>
</tr>
</tbody>
</table>

In the following, cycles refer to chip cycles. Then, a random access to a node pair takes 16 cycles to complete.

The first d-parameter of a tumbled node can be written in the DOP Register File 10 cycles after the (random) read was initiated. On average, continuous evaluation of Equation 1 can start after additional 12 cycles, when the first half of all d’s are available. The DOTADD-unit is assumed to have 6 pipeline stages. The first result will be clocked into the Results Register Bank after a total of 28 cycles, the last one after 52 cycles.

Last access to the DOP Register File for the processing of the first tumbled node sibling occurs in cycle 46. The other sibling can then be transferred sequentially from the SDRAM Interface Unit into the DOP Register File and processed in the same way. The transformed sibling will be ready in the Results Register Bank after 88 cycles.

By that time, the first pair of aligned nodes in the list has been fetched from memory, with one of the nodes being present in ”e”-register bank. The other node will be processed four cycles later. The load of the second node pair has been initiated such that processing can continue uninterrupted throughout cycle 100.

For all further memory reads, since we assume page faults for practically all memory reads, a delay will occur between read cycles. On memory chips with four internal banks, this delay will be two cycles on average due to bank interleaving, giving a total read time of 10 cycles for a node pair. Thus, the performance can be estimated as

$$T_L = 100 + (\alpha - 2) \times 10,$$

where $T_L$ is the number of cycles needed to process a list, and $\alpha$ is the number of aligned node pairs in the list. If for a given collision test for two objects there are $\tau$ lists to process, each with $\alpha$ node pairs on average, the total performance can be characterized as

$$T_T = (100 + (\alpha - 2) \times 10) \times \tau$$ (5)

The number of lists $\tau$ is given by the number of visited tumbled node pairs.

This estimation is compared to a software implementation on a PC with a Pentium-III CPU running at 1GHz. The results are summarized in Table 1.

6 Conclusions and Future Work

In this paper, we have presented novel algorithms and a hardware architecture for performing hierarchical collision detection. It is arguably the first special-purpose hardware architecture dedicated to this task. We lay special emphasis on the fact that this architecture is suitable for “polygon soups” in general, as opposed to previously reported methods utilizing graphics hardware.

As can be seen in Table 1, the speedup ranges between about 25 and 125 for our benchmarks. It is generally higher in worst-case scenarios, which is an important result, because interactivity is limited most severely by these cases. Thus, a chip design is very well justified.

A good part of the speedup can be attributed to our novel hierarchy traversal scheme, which can be applied to all kinds of bounding volume hierarchies.

Our near-term goal will be to implement a VHDL model of the CollisionChip, identify potential bottlenecks, and further optimize the architecture towards even higher processing speeds. Our long-term goal will be to integrate this project into an industrial virtual prototyping application.
Figure 6: Some of the objects of our test suite (car body, front light, cooling filter, door lock; courtesy VW and BMW).

References


