Compiling Geometric Algebra Computations into Reconfigurable Hardware Accelerators

Jens Huthmann\textsuperscript{1}, Peter Müller\textsuperscript{1}, Florian Stock\textsuperscript{1}, Dietmar Hildenbrand\textsuperscript{2}, and Andreas Koch\textsuperscript{1}

\textsuperscript{1} Embedded Systems and Applications Group
Technische Universität Darmstadt, Germany
Email: \{huthmann|mueller|stock|koch\}@esa.cs.tu-darmstadt.de

\textsuperscript{2} Computer Science Department
Technische Universität Darmstadt, Germany
Email: dhilden@gris.informatik.tu-darmstadt

Abstract. Geometric Algebra (GA), a generalization of quaternions and complex numbers, is a very powerful framework for intuitively expressing and manipulating the complex geometric relationships common to engineering problems. However, actual processing of GA expressions is very compute intensive, and acceleration is generally required for practical use. GPUs and FPGAs offer such acceleration, while requiring only low-power per operation. In this paper, we present key components of a proof-of-concept compile flow combining symbolic and hardware optimization techniques to automatically generate hardware accelerators from the abstract GA descriptions that are suitable for high-performance embedded computing.

1 Introduction

1.1 Compiling for Reconfigurable Computing

Reconfigurable computers have successfully been used to accelerate a wide spectrum of high-performance embedded applications, while requiring a power budget far below that of Graphics Processing Units (GPUs) with comparable throughput. However, the use of reconfigurable technology often required significant manual implementation effort and knowledge not only of the application, but also of digital design and computer architecture.

As in ASICs, the productivity gap between the hardware description languages (HDLs) traditionally used for digital design and the ever-increasing FPGA capacities has widened. On one side, this has been addressed by growing the synthesizable subsets of HDLs. Today, some tools can already synthesize variable operand multiplication and division into hardware and infer various kinds of memories directly from the HDL code. On the other side, many attempts have been made to compile from higher-level software programming languages (HLL) into hardware, e.g. [1–4].

Despite the progress in that area, translating HLLs into hardware is complex. In many cases, only a limited subset of language constructs can be translated.
Restrictions often exist with regard to control flow, data types, and pointer handling. All language features, that software developers expect to be available. Their lack one more complicates the use of hardware acceleration by non-experts.

A different approach to compiling to hardware lies in using more abstract domain-specific languages (DSL) instead of generic HLLs as input. These DSLs often pose less difficulty for automatic compilation since, e.g., difficult-to-translate constructs pointers or irregular control flow are not part of the language at all. This has already been done successfully for signal processing applications from MATLAB and Simulink ([5,6]). Our work also takes this route of compiling from Geometric Algebra, a powerful DSL much better suited to hardware mapping than a full HLL.

1.2 Geometric Algebra

The input language for our compiler are expressions formulated in Geometric Algebra (GA). GA is a very powerful mathematical framework for intuitively expressing and manipulating complex geometric relationships. In many cases, GA descriptions require only a fraction of the space of that conventional formulations (e.g., half a page instead of dozens of pages).

GA generalizes projective geometry, imaginary numbers, and quaternions to provide a powerful and flexible mathematical framework. It describes the manipulation of multi-vectors, which are linear combinations of simple vectors (called blades in this context). In addition to standard operators such as addition and subtraction, GA also encompasses special operators such as geometric product, inner product, outer product, inverse and division, dual and reverse operators (see [7] for an introduction).

The current form of GA has its roots in work by Grassmann [8] and Clifford [9] from the 19th century. However, its usefulness and wide practical applicability have only recently been discovered. Initially, it became popular to solve physics problems [10–12].

With the invention of conformal geometric algebra [13] by David Hestenes, this has also been extended to engineering applications such as robotics, computer graphics and computer vision. In conformal geometric algebras, high-level geometric objects such as points, lines, planes and spheres, as well as operations on them (e.g., intersection) can all be concisely expressed using GA operators.

However, due to the significant computation effort necessary to evaluate the multi-dimensional GA expressions, practical adoption has only been limited so far. While modern GPUs do have sufficient compute capacity [14], their long latencies (40 µs for a single computation) and high power requirements (170 W+) make them infeasible for many embedded control scenarios. Most FPGA-based reconfigurable computers do not quite reach the throughput of GPUs, but achieve much shorter latencies (for this example, 2 µs) and a much reduced power draw (here just 7 W). This will be discussed in greater detail in Sec. 4.
2 Related Work

This work is an extension of research previously presented as [15], now updated with further details and additional information regarding the hardware module library. Since [15] already contains an exhaustive survey of prior work, especially with regards to hardware compilation, the current discussion will just examine work directly relevant to our approach.

2.1 Tools

A number of pure software tools exist for working with GA expressions. Some of these, specifically CLUCalc, CLIFFORD, and Gaalop are used in our hardware compile flow.

CLUCalc is a software environment [7] for developing GA algorithms in the CLUCalc-script DSL.

CLIFFORD [16] is a Maple library allowing symbolic computations with GA operations.

Gaalop (Geometric Algebra Algorithms Optimizer) [17] is our plugin-based source-to-source compiler framework. It reads CLUCalc-script programs into a flexible intermediate representation, which can then be optimized in both target independent and dependent manner (internally using Maple and CLIFFORD for symbolic transformations). Output code can be generated C, LATEX, dot-graphs or CLUCalc-script (to visualize and verify the output). We use Gaalop as the base for our hardware compiler.

The main contribution of this work is the extension of Gaalop with numerous hardware-specific optimizations and a code generator for Verilog HDL.

2.2 Hardware Accelerators

With the high computing requirements for actually evaluating a GA model, much effort has been expended on special-purpose processor architectures.

Different attempts include [18], [19], [20], [21], [22], and [23]. However, in general, these approaches have not resulted in actual speed-ups over current CPUs.

The main reason for the poor performance of these prior attempts appears to be the fixation on software programmable, instruction-fetching processor architectures, even when targeting reconfigurable logic.

2.3 Benchmark Application

As the focus is on the compiler and not on the application, we chose as benchmark application an algorithm,

1. whose GA description superior to the standard formulation, and
2. where we have a both manual FPGA design and a GPU implementation as reference solutions.
A typical engineering application useful for evaluating the performance of our proof-of-concept compiler is an inverse kinematic computation: given a target point and a kinematic chain (e.g., shoulder, upper arm, elbow, forearm, wrist, hand), the algorithm computes the angles of all joints so the target point can be reached.

This specific inverse kinematics algorithm is used in a virtual reality (VR) applications [24]. As shown in [25], a formulation in five-dimensional conformal GA was 3x faster in software and much more concise (a page of formulas instead of many pages) than an algorithm using conventional mathematics.

Also, we have a manually created highly optimized versions for FPGA, GPU and multi-core CPU targets [14, 26]. Each implementation has been carefully hand-tuned for each platform (including, e.g., optimal bit width determination of FPGA operators and multi-threading for the GPU and CPU targets). We can thus judge the performance of the GA-to-hardware compiler using the manual design and a GPU implementation as reference.

3 Extending Gaalop

In this section we will give an overview over the entire compile flow of the Gaalop compiler framework and how its back-end is extended to translate the intermediate representation into high performance pipelined hardware data paths.

3.1 Gaalop Introduction

While CLUCalc is a very productive environment for the interactive development and debugging of GA algorithms, it does not allow the export of the completed models for execution outside of the tool. Gaalop aims to close this gap and export GA models into a variety of external formats (both executable and graphical).

As shown in Fig. 1, Gaalop reads a description for five-dimensional conformal GA algorithms as developed in CLUCalc. The CLUCalc-script is parsed into an intermediate representation (IR), specifically a control flow graph (CFG) of basic blocks holding the actual GA expression. Gaalop does not fully support control flow in CLUCalc-script yet, but that did not hinder the implementation of the inverse kinematics algorithm. A CFG was chosen for future extensibility to full control-flow support, which is currently under development.

Each of the basic blocks is stored as a data flow graph (DFG). The DFG represents the linear combinations of five blades. Each blade itself is represented as the outer product of five basis vectors \( (e_0, \ldots, e_3, e_\infty) \), with the grade of the blade being the number of different basis vectors combined. At this stage, the multi-vectors in the DFG may be fed to high-level GA operators.

5D conformal GA has multi-vectors which are linear combinations of at most 32 independent blades. For efficient compilation to a language without GA operators (e.g., C or fully spatial hardware), both the multi-vectors as well as the GA operators combining them have to be translated into their underlying primitive scalar representations and computations.
This is achieved symbolically using the Maple computer algebra system with the CLIFFORD library. With the library, Maple can now symbolically evaluate GA expressions in each DFG, simplifying them. Next, we also symbolically transform the remaining GA operators in the simplified GA expressions into their scalar equivalents, now operating on the individual scalar components of the basis vectors making up the blades. The results are scalar computations, amenable to both parallel as well as pipelined computation. Note that these scalar computations may well include trigonometric and similar functions as operators.

Fig. 2 shows this process of lowering a set of GA expressions into an expression solely consisting of primitive operations.

From this lowered DFG, the various back-ends can then generate code in the desired format. In addition to various textual and graphical formats (for documentation and debugging purposes), we have so far generated executable code in C/C++ and CLUCalc-script. In the next Section, we describe the flow from the lowered DFG to efficient hardware.

3.2 IR for Hardware Generation

Several standard optimization techniques are performed on the Gaalop-DFG, i.e. constant folding and common subexpression elimination. Afterward it is translated into an expanded form better suited to hardware generation. Still a DFG, it now holds only primitive operations acting on scalar data and also carries additional attributes such as data types (floating or fixed-point), format
DefVarsN3();

// Generic example:
// inputs: two points (x1, x2, x3),(p1,p2,p3)
// two spheres are intersected, and the
// resulting circle is intersected with a plane
// the end result is a pair of points Pp
Pw := x1*e1+x2*e2+x3*e3;
s1 := Pw-0.5*d2*d2*einf;
s2 := e0-0.5*d1*d1*einf;
Ze := s1^s2;
Plane := p1*e1+p2*e2+p3*e3;
?Pp = Ze ^ Plane;

(a)
⇓

Pw := ((subs(Id=1,(x1 &c e1)) + subs(Id=1,(x2 &c e2))) + subs(Id=1,(x3 &c e3)));
s1 := (Pw - subs(Id=1,(subs(Id=1,(subs(Id=1,(0.5 &c d2)) &c d2)) &c einf)));
s2 := (e0 - subs(Id=1,(subs(Id=1,(subs(Id=1,(0.5 &c d1)) &c d1)) &c einf)));
Ze := (s1 &w s2);
Plane := ((subs(Id=1,(p1 &c e1)) + subs(Id=1,(p2 &c e2))) + subs(Id=1,(p3 &c e3)));
Pp := (Ze &w Plane);
gaalop(Pp);

(b)
⇓

Pw := x1*e1+x2*e2+x3*e3
s1 := x1*e1+x2*e2+x3*e3-.5*d2^2*e4-.5*d2^2*e5
s2 := -1/2*e4+1/2*e5-.5*d1^2*e4-.5*d1^2*e5
...

(c)
⇓

Fig. 2. Converting a geometric algebra expression into primitive scalar operations. (a) GA computation in CLUCalc-script. (b) GA expressions as given to Maple. (c) CliffordLib results in GA, containing only primitive scalar operations (shortened). (d) Data flow graph used for hardware generation. For brevity, we just show the computation of blade 23 of the result Pp.

(bit-widths of integer and fractional parts of fixed-point representations), latencies and scheduling cycles. As we also pursue other HLL-to-hardware compiler
projects (e.g., [27]), which also require a hardware generation back-end, this new IR works as interface to a reusable back-end. The specific hardware generator used is called Nymble, accordingly the new IR is named Nymble-IR. Fig. 3 shows such an Nymble-IR graph (including control flow - the Gaalop back-end is not yet able to handle control flow, but the used hardware generator Nymble is).

### 3.3 Word Length Optimization

The area and speed of fully spatial compute units can be improved significantly by matching the width of the hardware operators to the data types processed at
this point in the calculation. This optimization must be assisted by the developer by specifying the value ranges and precisions of input and output data.

Word length optimization is performed by forward and backward propagation of the desired value ranges and precision. In the forward phase, the incoming value ranges (integer and fractional parts) determine the required width of the operator and its result. In the backward phase, unnecessarily precise (and thus too wide) operators can be narrowed and this truncation also propagated back toward the operator inputs.

For addition, subtraction, and multiplication the forward propagation is quite simple. However, division or functions such as square root, sine or cosine have more complex behavior. In this proof-of-concept implementation, we currently assume a default value (32b, with 16b fraction) for these functions, but this will be refined in future work. Similarly, we can set a global limit on the width of intermediate values. Note that the operators themselves are not affected by this and compute at the full required precision. Only the result is then clipped to the global limit.

In addition to these established techniques, we can also do word-width optimization based on the original higher-level Gaalop DFG-representation containing GA operators. For the proof-of-concept compiler, e.g., we recognize the normalization of vectors at the GA level, and restrict the output value range of the corresponding scalar operator to $[-1, \ldots, 1]$.

Good examples for operators that profit from backward propagation are inverse trigonometric functions (which will restrict the input value range to $[-1, \ldots, 1]$), or the square root (which limit the input value to be positive). If we cannot determine a narrow value range for an operator analytically, we then perform an automatic Monte-Carlo-Simulation of the entire data path to achieve a better fit. This Monte-Carlo-Simulation runs in parallel using both floating-point and fixed-point formats to also perform error estimation for all operator nodes.

While we can also directly generate data paths using single or double-precision floating-point operators, this is currently not practical: The proof-of-concept compiler presented here aims for a fully-spatial implementation (no operator sharing, but higher throughput). Even very simple GA algorithms will quickly lead to hardware exceeding the capacities of even the largest FPGAs. Area optimization of floating-point computations will be one topic for future research (see Sec. 5).

3.4 Scheduling and Balancing

After word-length optimization, the latency of the hardware operations can be determined and the computation actually scheduled. Since we aim for fully spatial operation without operator sharing, we use a simple greedy ASAP (as-soon-as-possible) approach: An operation $op_i$ with latency $l_i$ is scheduled at time $t_i = \max_{op_j \in \text{Predecessor}(op_i)} \{t_j + l_j\}$, i.e. it is scheduled after the latest predecessor operation has finished its computation.
For maximum pipeline throughput of one result per clock cycle, we then need to balance converging paths with unequal latencies by inserting registers. Also, all paths from all inputs to all outputs need to be brought to equal latency.

let \( D_{ij} \) the distance of the current node \( i \) to the successor \( j \)
sort \( D_{ij} \) by ascending distance

if \( D_{\text{first}} \neq D_{\text{last}} \) then
create new register node NOP \( n \)
for all successors \( j \) of \( i \) with \( D_k > D_{\text{first}} \) do
remove edge \((i,j)\)
insert edge \((n,j)\)
end for
add edge \((i,n)\)
execute algorithm for \( n \)
end if

Fig. 4. Balance successors of a node \( i \)

Fig. 4 shows the balancing algorithm. The successors \( j \) of the current node \( i \) in the DFG are sorted by their latency distance. The latter is defined as 
\[
\text{dist}(op_i, op_j) := t_j - t_i - l_j, \quad \text{with} \quad op_i \in \text{Predecessor}(op_j),
\]
\( t \) being the scheduled start cycles, and \( l \) the latency in cycles. If the minimal and maximal and distances are different, a register node is inserted in all paths from the current node that are longer than the shortest path. The register node itself is scheduled at cycle \( t_i + D_{\text{first}} \). The algorithm is then restarted on the new register node. The result is a data path with balanced path lengths.

### 3.5 Hardware Generation

Hardware is generated by our Nymble hardware back-end. Since the data path is a fully spatial, perfectly balanced pipeline, no additional control logic is required beyond markers indicating if and when results are available in the output (a simple shift register). For control flow implementation Nymble can generate lightweight controllers to (partially) pause the data path. This is necessary, e.g., in case of operations with variable latency (such as memory accesses) or (nested) loops.

We support chaining of some computations within the same clock cycle. At the moment, these are constant shifts, sign/bitwidth extension and bit-select operations that reduce to simple wires.

If the sinks of an operator are scheduled one or more cycles later, the source operator is fitted with a shift register to delay results over that time. Note that the balancing algorithm in Fig. 4 ensured that all sinks (possibly NOP nodes inserted for that very purpose) have the same latency distance from the source operator. Thus, many paths can share the balancing shift register.
Dedicated input registers accept input values for the computation, either as slave-writes from the CPU, or via a streaming mechanism directly from memory. Output registers can also be read from the CPU or be streamed back into memory.

The operators themselves are implemented using the flexible target-independent module library Modlib [28]. Internally, it expresses simple operators (e.g., addition, etc.) as synthesizable Verilog HDL operators. More complex operators (e.g., multiplication, division, square root, trigonometric functions) are realized internally using the Xilinx CoreGen IP blocks, using fully pipelined implementations with maximum throughput. The operators are generated on-the-fly for the specific bit-widths and data types required, caching generated modules for re-use if an operator with the same characteristics occurs again. As our other compiler projects also exploit a more complex, token-based hardware generation scheme [27], the Modlib operators support not only the static scheduling as used in Nymble, but also dynamic scheduling. This allows control speculation and canceling of mispredicted operators. To estimate the final area/timing data, Modlib can provide meta-information for the single operators. It determines this information empirically by actually automatically mapping the required operator once to the selected device, analyzing it, and then caching this information for further re-use.

Supported operators are add, sub, mul, div, mod, and type conversions for integer and float types, bitwise and logical operations, memory reads and writes, and I/O and IRQ registers. The parameters include the bit-width of the operands and the result, the signedness, the length of an output shift register (used for balancing), a queue depth for decoupling when used for dynamic scheduling, and in- and outputs for the speculation cancellation.

4 Evaluation and Results

As described in Sec. 2.3, we use an inverse kinematics application to evaluate the compiler prototype. Specifically, we compare the compiler-generated hardware with an implementation very carefully manually optimized by two experienced designers. In both cases, we target the Xilinx Virtex 5 devices using Synplify Premier for synthesis and Xilinx ISE for mapping.

For a fair comparison of the different platforms, our performance numbers assume that the input and output data is fetched from/stored to memory local to the computing device: The CPU has the data in its node-local memory accessed via FSB, the FPGA uses directly attached DRAM, and the GPU processes data in its on-board device memory.

Tab. 1 compares the area requirements and the performance for both solutions. Obviously, the compiler-generated datapath requires significantly more space than the manually optimized one, specifically a high number of DSP blocks. But with its deeper pipeline, it can be mapped to a Virtex 5 SX 240T device with a slightly higher clock frequency than the manual design.
Table 1. Comparison of manual design ([14]) vs. compiler-generated data path (compute kernels only, disregarding communication interface).

<table>
<thead>
<tr>
<th></th>
<th>manual</th>
<th>compiler</th>
</tr>
</thead>
<tbody>
<tr>
<td># operations</td>
<td>140</td>
<td>258</td>
</tr>
<tr>
<td>resources # FFs</td>
<td>49938</td>
<td>71173</td>
</tr>
<tr>
<td>resources # LUTs</td>
<td>34912</td>
<td>72664</td>
</tr>
<tr>
<td>resources # DSPs</td>
<td>74</td>
<td>817</td>
</tr>
<tr>
<td>pipeline length</td>
<td>365</td>
<td>447</td>
</tr>
<tr>
<td>max. frequency [MHz]</td>
<td>170</td>
<td>180</td>
</tr>
<tr>
<td>throughput [10⁶eval/s]</td>
<td>170</td>
<td>180</td>
</tr>
<tr>
<td>latency [µs]</td>
<td>2.147</td>
<td>2.483</td>
</tr>
<tr>
<td>speed-up to CPU</td>
<td>6.9x</td>
<td>7.3x</td>
</tr>
<tr>
<td>average word-length</td>
<td>38</td>
<td>45</td>
</tr>
<tr>
<td>average fraction-length</td>
<td>23</td>
<td>41</td>
</tr>
<tr>
<td>implementation time</td>
<td>80</td>
<td>&lt;&lt; 1</td>
</tr>
</tbody>
</table>

It is clear that our future work needs to concentrate on area optimization. The human designers exploited a number of high-level algebraic simplifications that are not yet performed automatically using the Maple computer algebra system in the Gaalop flow. This also affects the fixed-point conversion: The manually optimized design contains significantly fewer operators that are infeasible for analytical value range determination. Instead, the compiler has to rely on the Monte-Carlo-Pass to tighten the constraints. That approach, however, suffers from the nature of the Monte-Carlo test data generation: Since we aimed for a general-purpose solution, we generate streams of completely random input vectors. Not all of these will actually be valid inputs for this specific problem (e.g., a kinematic chain anchored at the origin can obviously not reach the origin and other points very close to it). Thus, we have to extend operator value ranges

---

**Fig. 5.** Throughput
to handle values that will actually never appear in practice, leading to wider operators. This explains that the average word-length in the compiler-generated design is 1.2x larger than the one in the manual design.

Performance-wise, though, the compiler-generated design performs quite satisfactorily (see Fig. 5): It slightly exceeds the throughput of the manual design (measured as million function evaluations per second) and has similar latency. It is still significantly better in terms of throughput than a four-threaded software implementation running on a 2.4 GHz Intel Core 2 Quad Q6600 CPU (which would draw 4.6x the power of the FPGA), yielding a real wall-clock speed-up compared to most of the prior approaches outlined in Sec. 2. While a GPU under optimum conditions could be even faster (1366M evaluations/s), it also incurs a latency of more than 40 µs on a NVidia GTX 280 card, which also would draw more than 24x the power of the FPGA. Fig. 6 shows that the FPGA is much more efficient in terms of required energy per computation. [14] gives greater details on these alternate implementations.

Apart from the area and performance issues, however, an automatic tool must be rated by its effect on designer productivity. This is the area where even the proof-of-concept compiler shines: The manual implementation required a total of approx. 80 h of determined effort by two experienced designers, familiar with both digital design/computer architecture as well as the maths underlying GA (which they exploited for the operator-reducing high-level simplifications). The compiler itself takes less than a minute to execute, with the bulk of the total implementation time taken by the Xilinx ISE mapping tools. Now, a domain expert proficient in GA can use a familiar notation to describe an algorithm, with no hardware design knowledge required.
5 Conclusion and Future Work

Even in its proof-of-concept stage, the compiler generates compute pipelines for the GA descriptions with a throughput significantly higher than the carefully tuned software version on a quad-core CPU.

The compiled compute pipeline does not yet reach the performance of the manual reference implementation, but was created in a fraction of the design time (minutes vs. days). Gaalop can already be used to quickly perform experiments with other GA algorithms, something simply not possible if a manual hardware design would be required for each problem.

Ongoing research also tackles going from the fully spatial design presented here to one with a flexible degree of operator sharing. This not only will allow the implementation of even more complex GA applications without using excessive amounts of reconfigurable area, but also the use of smaller reconfigurable devices for less extreme application performance requirements.

The compiler does not yet perform all of the optimizations that were undertaken for the manual design. Specifically, tree height-reduction would have been advantageous. Also, when implementing the CLUCalc-script control flow constructs, our very simple word-length optimization has to be replaced with a more precise algorithm, e.g., [29] or [30]. All of these classical techniques will need to be extended to exploit the underlying structure of the high-level GA operators to achieve even tighter word-length fittings. These issues are also the subject of current research in our group.

5.1 Dynamic Reconfiguration

As good as the inverse kinematic is for demonstrating the capabilities of our proof-of-concept compiler and the potential of GA, it is not a representative choice for general GA algorithms: Usually, these consist of much smaller sub-models, that are called at different times from the host program. These sub-models could be mapped onto the FPGA one at a time using dynamic partial reconfiguration.

An example of an application with these characteristics is a molecular dynamic simulation, which consists of a number of different kernels. Since a fully spatial, high-performance implementation of each kernel fills a large FPGA all by itself, putting the entire application onto a chip is not feasible using current devices.

Dynamic reconfiguration, however, would allow to execute all relevant parts using appropriate hardware accelerators. This approach is subject of ongoing research work and will eventually also be added to the compiler.

References


