Advances in Integrated Software Sciences

Proceedings of the Master and PhD Seminar Summer term 2011, part 1

Softwarepark Hagenberg
SCCH, Room 0/2
15 April 2011
Program

Session 1. Chair: Bernhard Moser

9:00  Volkmar Wieser:
      Combining High Productivity and High Performance in Image Processing Using Single Assignment C

9:30  Dominic Looser:
      Semantics — Seeking the Big Picture. Part 1.
Combining High Productivity and High Performance in Image Processing Using Single Assignment C

Volkmar Wieser\textsuperscript{a} and Bernhard Moser\textsuperscript{a} and Sven-Bodo Scholz\textsuperscript{b} and Stephan Herhut\textsuperscript{b} and Jing Guo\textsuperscript{b}

\textsuperscript{a}Software Competence Center Hagenberg, Software Park 21, Hagenberg, Austria
\textsuperscript{b}University of Hertfordshire, College Lane, Hatfield, AL10 9AB, UK

ABSTRACT

In this paper the problem of high performance software engineering is addressed in the context of image processing regarding productivity and optimized exploitation of hardware resources. Therefore, we introduce the functional array processing language Single Assignment C (\textsc{SaC}), which relies on a hardware virtualization concept for automated, parallel machine code generation. An illustrative benchmarking example proves both utility and adequacy of \textsc{SaC} for image processing.

Keywords: Functional Programming, Parallel Programming, Hardware Virtualization, Anisotropic Diffusion

1. INTRODUCTION

The landscape of parallel computing has substantially changed in the last years. It is not only obvious that “the future is parallel” but also current trends confirm that computing power through parallelism will be provided by many-core architectures.\textsuperscript{1} General-purpose many-core architectures must conveniently support a wide range of programming styles and languages. If architectures prefer a particular model of parallel programming, they are not likely to become widely accepted, especially if such architectures require programming skills that probably overstrain the average programmer. Furthermore, the development of high performance applications on novel and ever-changing hardware environments like multi- and many-core systems, Graphics Processing Units (GPUs) or Field Programmable Gate Arrays (FPGAs) is cost- and time-intensive. What is needed is a convenient abstract language that supports automatic parallelization on different architectures without changing source code and robust performance benefits. Software engineers neither should need to design and develop in unintuitive ways nor to deal with a variety of hardware and language details just to avoid design mistakes or bottlenecks or just to achieve an attractive speedup. From the economical point of view, an approach that yields the desired performance with minimal effort will be preferred, particularly for real-time performance applications in industry as well as for less efficient hardware.

In the field of quality inspection of textured surfaces, e.g. for the production of foils or industrial woven fabrics, we have to cope with a scanning speed up to 300m/min, i.e. about 80MB/sec per camera systems, and a complex phenomenology of textures and defects. This requires the application of advanced cost-intensive algorithms of image processing as well as machine learning, the use of high-performance computational hardware like GPUs or multi-core systems and the exploitation of parallelization potentials. The analysis of the whole processing pipeline (image acquisition, preprocessing, feature extraction, registration, defect detection and classification) with standard languages regarding performance is a resource- and time-intensive challenge.

While image acquisition as well as major parts of preprocessing and feature extraction can be computed in a well-predictable way due to the pre-fixed size of filter operations and amount of data transfer known in advance, this is usually not the case for the high-level processing steps of pattern recognition and classification. With increasing semantics the processing steps involve operations of increasing complexity and computational costs. For example at the beginning of the processing pipe a typical scenario is to use local filter operations acting e.g. on 9x9 matrices while later on also global operations on the whole image data are applied like a registration with

Further author information: (Send correspondence to Volkmar Wieser)
Volkmar Wieser: E-mail: volkmar.wieser@scch.at, Telephone: +43 7236 3343 844
a reference model based on thousands of feature points. So far the processing steps are acting on a physical, appearance-based level which only depends on the image intensity values. Finally, defect candidates have to be identified, located and classified. This final step heavily depends on parameters that are not coded within the image, e.g. the customer’s judgment whether some product quality aspects can be accepted or have to be rejected. The complexity of the classification step correlates with the complexity of the defect taxonomy. For industrial textile fabrics there are about 50 major categories of defects like wrong webbing, swelling, tension of yarn, surface wrinkling or staining problems, which require appropriate and sophisticated strategies for defect classification.

Therefore, from the economical point of view the main question is how to find the optimal trade-off between performance and economical resources. Single Assignment C (SAC)\textsuperscript{2} provides a design scope which definitely reduces the period for design and development and hence saves resources by increasing productivity. In this paper we demonstrate this benefit on the basis of an application-oriented example (Perona-Malik Anisotropic Diffusion\textsuperscript{1}) which has poor performance characteristics by default. First, we introduce the functional array language SAC in Section 2. Afterwards, we compare the SAC optimization strategies against those of the OpenCV\textsuperscript{2} in Section 3. Finally, we illustrate benchmarking results of both implementations on a many-core system in Section 4.

2. SINGLE ASSIGNMENT C

SAC is a purely functional and data-parallel programming language. All its basic language constructs are identical to those of C, not only with respect to their syntax but also with respect to their semantics. Despite this rather imperative look and feel, a side-effect free semantics is enforced by the exclusion of a few features of C, most notably the absence of pointers. As a replacement, extensive support for compiler-managed n-dimensional arrays as well as a very powerful language construct for expressing data-parallel operations called with-loop have been added. Based on this language construct, various MATLAB-like operations have been defined. Similar to the STL of C++, these operations are provided within a library and come as a part of the SAC distribution\textsuperscript{9}.

Various application studies demonstrate that this setting enables (i) sequential runtimes competitive with those of hand-written C and FORTRAN codes, and (ii) almost linear speedups from auto-parallelization for shared memory systems.\textsuperscript{4}

A first introduction of SAC is presented in "Single Assignment C - efficient support for high-level array operations in a functional setting".\textsuperscript{5} This paper discusses the general approach for integrating arrays with access time $O(1)$ into functional languages as well as the shape-invariant language construct called with-loop. Finally, a runtime comparison between SAC and High Performance Fortran (HPF)\textsuperscript{6} implementation is done. Further experiments with respect to runtime performance are done in "SAC - from high-level programming with arrays to efficient parallel execution".\textsuperscript{2} Beside a general introduction this paper demonstrate the major steps in compiling rank- and shape-invariant high-level SAC specifications into efficiently, executable multithreaded code. The measurement of runtime performance is taken in SAC, HPF\textsuperscript{6} and ZPL\textsuperscript{7} (A Machine Independent Programming Language for Parallel Computers). More detailed introductions of SAC can be found in further readings (e.g. design\textsuperscript{8} and compilation\textsuperscript{2} aspects, shape-invariant programming facilities,\textsuperscript{9} accelerating APL\textsuperscript{1} programs,\textsuperscript{10} generic programs on arrays,\textsuperscript{11} and with-loop-fusion\textsuperscript{12}).

At present, in the scope of the EU-funded project ADVANCE,\textsuperscript{13} further benchmarks with industrial applications will transfer to SAC and the stream processing language S-Net.\textsuperscript{14} These applications are benchmarked regarding throughput, scalability, and runtime performance on UNIBENCH,\textsuperscript{15} where first results are available. To demonstrate the combination of SAC’s array support with standard C code, we present a small example (see Algorithm 1).

Algorithm 1 shows a program fragment that is typical for many scientific codes. It contains the definition of a function relax which takes two arguments: a scalar double $\epsilon$ and a two-dimensional array (matrix) of double elements $\mathbf{a}$. In the body of the function the array $\mathbf{a}$ is repeatedly modified within a do-loop. Each modification is expressed by means of the with-loop in lines 5-10. It specifies that all inner elements of the new version of $\mathbf{a}$ are

\*available from www.sac-home.org

\textsuperscript{1}APL (A Programming Language) is an interactive array-oriented language
Algorithm 1 Program fragment for combination of SaC and C code

default[...\] relax(double eps, double[...\] a) {
1:  do
2:    old\_\_a = a;
3:    a = with {
4:      \([(1, 1) <= iv < shape(a) - 1]\} 
5:      res = 0.25 * (a[iv - [0, 1]] + a[iv + [0, 1]]
6:                + a[iv - [1, 0]] + a[iv + [1, 0]]);
7:    }
8:    :modarray(a);
9:  while (maxval(abs(a - old\_\_a)) > eps);
10: return (a);
}

computed as a mean of the four direct neighbor elements of the old version of a. Note that indexing into arrays in SaC is done by means of index vectors rather than sequences of scalar indices. The function shape used in line 6 returns a vector containing the number of elements of the given array; the subtraction of 1 is applied to all elements of that vector, indexing starts at 0. As a consequence, all indices between [1,1] (inclusive) and shape(a)-1 (exclusive) denote exactly all inner elements of a. Furthermore, it is important to note that the semantics of the with-loop ensures that all array modifications, conceptually, happen at the same time, i.e., they can be executed in parallel.

In line 12 we see the termination condition of the do-loop. As for the description of the indexing range in line 6, we make extensive use of the array operations from the SaC standard library. The expression (a - old\_\_a) denotes a two-dimensional array containing all the differences between the old version of a and the current one. The function abs computes their absolute values and maxval determines their maximum value. Only if this value exceeds a given epsilon eps, computation proceeds.

Note here that the nature of the termination condition always requires two generations of the array a (a and old\_\_a) to be available. This is achieved by the assignment in line 4. It conceptually creates a copy of the current array a. However, the compiler/runtime system of SaC ensures that these copies in fact do not materialize at runtime. More details on the technology that facilitates this kind of optimization can be found in.

3. A BENCHMARKING EXPERIMENT

In this paper we claim that SaC allows high-productivity and high-performance. In order to underpin this claim we want to work out an example by showing its style of implementations and performance behavior in comparison with an OpenCV implementation. As example we choose the anisotropic diffusion filter operation of Perona-Malik\(^1\) which is widely-used within the image processing community. The execution time of the filter only depends on the dimension of the input image and not on the content. This characteristic allows an objective analysis of the scalability on multi/many-core architectures as well as a good demonstration of the automatic data-parallelism feature of SaC.

3.1 The Algorithmic Background of Perona-Malik Anisotropic Diffusion

First, the idea of Perona-Malik anisotropic diffusion is to locally steer the diffusion in a specific direction by introducing \( D(\ldots) \)
\[
\frac{\partial}{\partial t} \varphi = \text{div}(D \Delta \varphi),
\]  
(1)

with boundary condition
\[
\varphi(\ldots, 0) = \varphi,
\]  
(2)
where \( D \) depends on the local derivatives. Perona-Malik propose two different derivates
\[
D = \frac{1}{1 + c \|\Delta \varphi\|^2}, \tag{3}
\]
and
\[
D = \exp(-\|\Delta \varphi\|/c). \tag{4}
\]
where Equation (3) acts like a smoothing filter that suppresses fine (noisy) structures, while Equation (4) strengthens high contrast edges. For an illustration, see Figure 1 to Figure 3.

Algorithm 2 outlines a pseudo-code for an implementation of Perona-Malik anisotropic diffusion, which contains the degree of smoothing by the number of iteration steps in the for-loop (line 3 till 14). This means that the result of the first iteration is the input of the next iteration and so on, which limits the scope of parallelization.

Algorithm 2  Pseudo code for Perona-Malik anisotropic diffusion

1: \textit{diff}f \leftarrow \textit{create px1} border around image
2: \textit{diff} \leftarrow \textit{image}
3: \textbf{for} \textit{i} = 1 to \textit{NITER} \textbf{do}
4: \textit{shift}(\textit{diff}, px1 to North, South, East and West)
5: \textit{delta} \leftarrow \textit{substract}(\textit{diff} from shifted image)
6: \textbf{if} \textit{type} == 1 \textbf{then}
7: \textit{else}
8: \textit{end if}
9: \textit{c} \leftarrow 1/(1 + \textit{kappa} \times \textit{delta}^2)
10: \textit{c} \leftarrow \textit{c} \times \textit{delta}
11: \textit{c} \leftarrow \textit{add}(\textit{allshiftedimages}) \times \textit{LAMBDA}
12: \textit{diff} \leftarrow \textit{diff} + \textit{c}
13: \textbf{end for}
14: \textit{result} \leftarrow \textit{diff}

Figure 2 illustrates the use of the deviation in Equation (3), where we can see that only the connected wide regions are left, whereas noise structure is largely removed. The use of the deviation in Equation (4) in Figure 3 shows us that beside the big deep scratch in the middle also fine, noisy, high contrast edges are left.

Figure 1. stainless steel plate with noise surface and scratch
Figure 2. Application of Equation (3) on image of Figure 1; NITER=5; \( delta = 1/3; \) kappa=10;
Figure 3. Application of Equation (4) on image of Figure 1; NITER=5; \( delta = 1/3; \) kappa=10;

In Chapter 4 we evaluate the execution time and the scalability of the anisotropic diffusion, which is implemented as shown in Appendix A.1 for OpenCV and Appendix A.2 for SaC.
4. BENCHMARKING RESULTS

For our test scenario we use three different hardware environments, i.e., a DELL Precision™ 690 for CPU performance tests, a NVIDIA GeForce GTX 480 graphic card for GPU computing, and a benchmarking environment named unibench™ to demonstrate scalability. The Dell Precision™ 690 has two separate Intel® Xeon® 5060 with 3.2GHz, i.e., 8 cores in total, and 2GB full buffered DDR2 memory. The NVIDIA GeForce GTX 480 has a core frequency of 1.4 GHz and in total 480 cores, i.e., the graphic card has a bandwidth of about 177.4 GB per second. unibench has a Symmetric Multiprocessing (SMP) architecture, where four AMD Opteron TM 6174 processors are connected to a single shared main memory. Each processor has 12 cores and a clock speed of 2.2 GHz, i.e., that are in total 48 cores. The advantage of SMP machines is the large amount of physical memory of about 256 GB, visible to all 48 cores, which allows multi-threaded applications. The input data of the anisotropic filter range from 256 × 256 pixels to 4096 × 4096 pixels with pseudo-randomly generated 8-bit values between 0 and 255 since in this example only the dimension of the data affects the execution time.

Therefore, in Table 1 we present five different input sizes and propagate for each of them Equation (3) ten times. In comparison to OpenCV, with the input size of 4096 × 4096 pixels, using the multithreaded SAC-MT compilation we can reduce the execution time of the filter and, consequently, the execution time of the application on 8 cores by 27.72 seconds. This reduction is realized with a similar effort of development. This improvement can be a time buffer for using more complex algorithms or having a significant competitive advantage against other applications.

<table>
<thead>
<tr>
<th>Input Size</th>
<th>OpenCV</th>
<th>SAC-SEQ</th>
<th>SAC-MT</th>
<th>SAC-MT vs. OpenCV</th>
</tr>
</thead>
<tbody>
<tr>
<td>256 × 256</td>
<td>0.09 sec</td>
<td>0.16 sec</td>
<td>0.15 sec</td>
<td>0.6x</td>
</tr>
<tr>
<td>512 × 512</td>
<td>0.51 sec</td>
<td>0.86 sec</td>
<td>0.21 sec</td>
<td>2.4x</td>
</tr>
<tr>
<td>1024 × 1024</td>
<td>2.31 sec</td>
<td>2.67 sec</td>
<td>0.69 sec</td>
<td>3.3x</td>
</tr>
<tr>
<td>2048 × 2048</td>
<td>9.21 sec</td>
<td>10.65 sec</td>
<td>2.41 sec</td>
<td>3.8x</td>
</tr>
<tr>
<td>4096 × 4096</td>
<td>37.43 sec</td>
<td>42.75 sec</td>
<td>9.71 sec</td>
<td>3.8x</td>
</tr>
</tbody>
</table>

Table 1. Comparison of OpenCV and SAC implementation of anisotropic filter using DELL Precision™ 690

The reason of the performance gain of SAC is the optimal processor load on all 8 cores and the functional programming paradigm, which results in a low memory usage of about 350 MB on average. The standard OpenCV implementation does not support the distribution on the available cores. Therefore, the OpenCV anisotropic filter is executed only on one core and needs many times more memory of about 1.9 GB on average due to data materialization. For further analysis of SAC we restricted the execution of the anisotropic filter to a single-core. As we can see in Table 1, the sequential execution time of SAC-SEQ is close to OpenCV but with slightly higher communication overhead.

If we need more performance for the application scenario, we have the possibility either to re-implement the whole algorithm with NVIDIA’s CUDA framework or automatically generate executable GPU code with the SAC-CUDA backend. A re-implementation definitely requires higher development costs and programming know-how from experts, where SAC-CUDA allows a flexible time and cost efficient development. For the moment being, it is not possible to outperform a hand coded implementation by means of SAC-CUDA. While for the automated SAC-CUDA code we achieve a speedup about 40× (with respect to OpenCV implementation with input size of 4096 × 4096), a manually optimized CUDA implementation allows a speedup of 197× for the same example. However, note that SAC-CUDA is still under development, which in the near future will bring further improvements and speedups by means of SAC-CUDA.

<table>
<thead>
<tr>
<th>Input Size</th>
<th>OpenCV vs. SAC-CUDA</th>
<th>OpenCV vs. CUDA-hand</th>
</tr>
</thead>
<tbody>
<tr>
<td>256 × 256</td>
<td>9×</td>
<td>18×</td>
</tr>
<tr>
<td>512 × 512</td>
<td>25.5×</td>
<td>72×</td>
</tr>
<tr>
<td>1024 × 1024</td>
<td>38.5×</td>
<td>144×</td>
</tr>
<tr>
<td>2048 × 2048</td>
<td>38.5×</td>
<td>184×</td>
</tr>
<tr>
<td>4096 × 4096</td>
<td>40×</td>
<td>197×</td>
</tr>
</tbody>
</table>

Table 2. Comparison of hand-coded CUDA code and SAC-CUDA produced code on NVIDIA GeForce GTX 480

To analyse the scalability of SAC compiled code, we can see in Table 3 that the scalability of the anisotropic diffusion using the SAC implementation is nearly linear by increasing the number of cores.
It is interesting to observe that the OpenCV and SAC implementations have a similar programming style. Particularly, the handling of matrix operations is comparably comfortable in both languages.

Concerning the optimization strategies, we can say that OpenCV offers an optimized Streaming SIMD Extensions 2 (SSE2) code. SSE2 is a processor supplementary instruction set for modern 32-bit x86 and 64-bit x64 Single Instruction, Multiply Data (SIMD) architectures, where many of the basic arithmetic functions can run significantly faster. OpenCV also contains Intel® Threading Building Blocks (TBB)\textsuperscript{16} support for several functions. TBB is a C++ template library which offers a complete threading mechanism on modern multi-core processors. The advantages of this library are easy and efficient handling (software engineers do not need to be threading experts), scalable performance and a higher-level, task-based parallelism. In our example TBB is irrelevant as we do not use OpenCV functions that support this library. OpenCV applies the TBB only to OpenCV applications, e.g., haartraining, traincascade, and not to basic arithmetic/filter operations.

The optimization strategy of SAC is different. One of the major design principles of SAC is the with-loop construct, which supports the specification of shape-invariant array operations. All primitive array operations of SAC can be defined as with-loops within a standard library rather than being implemented as part of the compiler. The basic idea is to use with-loops as a universal representation for array operations and to develop a general transformation scheme that allows the transforming of two consecutive with-loops into a single one. Using this mechanism called with-loop-folding,\textsuperscript{17} any nesting of primitive array operations can be stepwise transformed into a single loop construct that contains an element-wise specification of the resulting array. During the compilation process of the SAC-compiler various conventional optimization techniques,\textsuperscript{18} such as function inlining, constant folding, constant propagation, loop unrolling, and dead code removal, are applied to produce efficiently executable, multithreaded C code with Posix\textsuperscript{TM} threads.\textsuperscript{19}

5. CONCLUSION

In this paper we showed the advantage of the functional array language Single Assignment C (SAC) in the field of image processing, especially for the anisotropic diffusion filter. Therefore, we conducted a benchmarking experiment in which we compared the effort of the development using the common image library opencv2.0. Compared to the language syntax, SAC is similar to MATLAB because of the definition of various MATLAB-like operations. Furthermore, the period of development can be reduced by the well-known C/C++ semantics, however, by applying a side-effect free semantics most notably due to the absence of pointers and hardware virtualization. The hardware virtualization allows flexible and fast development on architectures corresponding to CPUs, GPUs or FPGAs using the same language and the same implementation. Moreover, on multi/many-core architectures we achieve a nearly linear speedup due to the auto-parallelization of SAC. This is indispensable for inline quality inspection applications due to a scanning speed up to 300m/min and about 80MB/sec per camera systems, and the complex phenomenology of textures and defects. Additionally, from the economical point of view, SAC provides us with an extremely good balance between time of development and performance.

Although SAC is well suited for image processing because of the data-parallelism and the n-dimensional array support, there does not exist an image library like opencv. In addition, during the development of SAC applications the support of debugging tools is limited. This will be changed in the future by an intensive enhancement of SAC and community building.
REFERENCES

APPENDIX A. IMPLEMENTATION DETAILS OF ANISOTROPIC DIFFUSION

A.1 "OpenCV 2.0" implementation

```c
void executeAnisotropicFilter(const cv::Mat& image, cv::Mat& result, int NITER, int type)
{
    // allocate memory
    cv::Mat diff(image.rows + 2, image.cols + 2, CV_64FC1);
    ... // code
    cv::Mat roi = diff(rectl); diff.copyTo(roi);
    cv::Mat roiN = diff(rectN);
    cv::Mat roiS = diff(rectS);
    cv::Mat roiE = diff(rectE);
    cv::Mat roiW = diff(rectW);
    deltaN = roiN - diff;
    deltaS = roiS - diff;
    deltaE = roiE - diff;
    deltaW = roiW - diff;
    conduction(cN, deltaN, type);
    conduction(cS, deltaS, type);
    conduction(cE, deltaE, type);
    conduction(cW, deltaW, type);
    cN = cN.mul(deltaN);
    cS = cS.mul(deltaS);
    cE = cE.mul(deltaE);
    cW = cW.mul(deltaW);
    diff = diff + (cN + cS + cE + cW) * LAMBDAB;
    } /* convert image back to int */
    diff.convertTo(result, image.type(), 1, 0);
}

inline void conduction(cv::Mat& c, cv::Mat& delta, int type)
{
    if(type == 1){
        c = delta / KAPPA;
        c = c.mul(c, -1);
        cv::exp(c, c);
    } else{
        c = pow(delta, 2.0);
        c = c.mul(KAPPA);
        cv::add(c, 1, c);
        c = 1 / c;
    }
}
```

Figure 4. opencv 2.0 implementation of Perona-Malik Anisotropic filter
A.2 "Single Assignment C" implementation

```c
int[...] executeAnisotropicFilter(int[...] image, int niter, int type)
{
    /*convert image to double*/
    result = tod(image);
    for (i=0; i<niter; i++)
    {
        imageN = shift(0,-1,0.0,result);
        imageS = shift(0,1,0.0,result);
        imageE = shift(1,-1,0.0,result);
        imageW = shift(1,1,0.0,result);
        deltaN = result - imageN;
        deltaS = result - imageS;
        deltaE = result - imageE;
        deltaW = result - imageW;
        cN = conduction(deltaN, kappa, type);
        cS = conduction(deltaS, kappa, type);
        cE = conduction(deltaE, kappa, type);
        cW = conduction(deltaW, kappa, type);
        mN = cN * deltaN;
        mS = cS * deltaS;
        mE = cE * deltaE;
        mW = cW * deltaW;
        result = result + (lambda*(mN + mS + mE + mW));
    }
    /*convert image back to int and return*/
    return (toi(result));
}
```

```c
inline double[...] conduction(double[...] delta, double kappa, int type)
{
    /*privileges high-contrast edges over low-contrast ones.*/
    if (type == 1)
        c = exp(-(delta / kappa));
    /*privileges wide regions over smaller ones.*/
    else
        c = 1.0 / (1.0 + pow((delta / kappa),2.0));
    return(c);
}
```

Figure 5. SaC implementation of Perona-Malik Anisotropic filter
Semantics: Seeking the Big Picture. Part 1

Dominic Looser

Abstract

There is a recent surge of use of the term *semantic* in computer science. This is due to developments that can be subsumed under the notion of the semantic web. But the field of semantics also has a rich history and is highly connected to various scientific disciplines. Its study quickly leads to general questions concerning the relations of humans, mind, signs, and reality. This series of talks seeks to approximate the big picture of semantics, leading from semiotics to the semantic web.