Memory Addressing Organization for Stream-based Reconfigurable Computing

(invited paper)

Michael Herz,
Agilent Technologies,
D-71034 Böblingen, Germany
herz@computer.org

Reiner Hartenstein,
Dept. of Computer Science (Informatik)
University of Kaiserslautern,
D-67653 Kaiserslautern, Germany
reiner@hartenstein.de

Miguel Miranda, Erik Brockmeyer, IMEC,
Francky Catthoor, IMEC, and,
part-time professor at K.U.Leuven
B-3001 Leuven-Heverlee, Belgium,
{miranda}@imec.be

Abstract. The paper gives a survey on methods to cope with the memory bandwidth problem, especially in the context of data-stream-based computing systems, like reconfigurable (and hardened) datapath arrays, introduced in [1] and elsewhere.

1. Introduction

Being more area-efficient than reconfigurable logic platforms like FPGAs [2], coarse grain Reconfigurable Arrays (cGRA) [1] are preferred hardware platforms for Reconfigurable Computing (RC) applications like multimedia and next generation cellular wireless [2] [3] [4]. In a similar way as already known from systolic arrays, such RAs are running in a data-transport-triggered execution mode, with many data streams flowing into and out of the RA. The more data stream traffic is meeting the RA the more the efficiency and performance depends on the RA / memory communication architecture. This paper deals with the requirements which are much different from traditional parallel computing systems.

2. Smart Address Generators

Latency of data accesses often determines the computation throughput (a study e. g. in [5]). Address generators have the potential to reduce computation time significantly. In a grid-based design rule check [6] implemented on the MoM-1 (see section "The GAG") a speed-up of more than 2000 has been achieved, compared to a VAX-11/750, where a dedicated address generator contributed a factor of about 10 by avoiding memory cycles otherwise needed for address computation overhead. This section gives an overview on address generators and focuses on: simultaneous address generation and data manipulation, reduction of the time needed for address generation, and, reducing or avoiding memory cycles by software to hardware migration in address computation.

The Structured Memory Access (SMA) Machine. [5] is a von Neumann machine reducing address calculation overhead, by special hardware to generate data structure references. SMA consists of a computational processor (CP) and an independent memory access processor (MAP). For computations, the CP buffers an entire instruction block from where it may loop over one or more blocks of instructions.

MAP supports absolute accesses always accessing the same data and accesses relative to some index value. Index register functions use a hardware stack which tracks the active indices of inner loops, and all data structure references are made by using a subset of the index values. Tables in the SMA processor store the base address of each data structure and other information needed for an entire address from indices. These tables are loaded once at the beginning of a program execution. MAP has an internal Operand-Instruction Buffer (OIB) and immediately generates operand addresses and places them on the read or write queue of the outstanding memory requests. Index instructions are especially supported by sophisticated hardware features.

The GAG (generic address generator). The MoM-1 (Map-oriented Machine 1) was originally called PISA machine (pixel-oriented system for image analysis) [7] [8] [9], an image processing machine with 2-D memory organization, primarily designed to implement a pattern matching approach to grid-based design rule check (DRC [10] [11]), which achieved a speed-up of more than 2000 compared to a VAX-11/750. This acceleration stems from avoiding address calculation overhead (90% of the VAX CPU time) and fully parallelized pattern matching by a dynamically reconfigurable PLA (DPLA, designed at Kaiserslautern and manufactured via the German multi university E.I.S. project).

The MoM-1 address generator called MCU (move control unit) is an application specific generic address generator (later called GAG) is configured before execution time and needs no memory cycles at runtime. It generates independent x- and y-addresses to move a 3-by-3 word “scan window” with a limited set of operations needed for a video scan pattern over a 2-D pixel file. A local 2-D shift register file avoids multiple reading of the same data due to overlapping scan window locations.

Application-specific Address Generator (ASAG). Grant, Denyer and Finlay [12] use a synthesis method for application specific address generators (ASAG), dedicated for multimedia applications, where the sequence of storage and retrieval of particular blocks of data is strongly patterned. The needed address patterns are generated by a dedicated counter, or via circuit transformations (bit shuffling and combinatorial logic operations) applied to a counter output. Details of the synthesis method and an example can be found in [12].

GAG of MoM-2. The Map-oriented Machine 2 [13] uses a 2-level address generator [14] based on the extremely flexible slider method [10], supporting vari-size scan windows (MoM-2 supports only up to 4-by-4 size scan windows) and their navigation through 2-D memory space. The data sequencer's address generator is configured by parameters and needs no memory cycles at run time. It consists of a Task Manager, a JumpGenerator [15], and a Single Step Control Unit (SSCU), which operate in a pipelined fashion.

The JumpGenerator which computes x- and y-address of the scan window base location [13] on the basis of a few buffered parameters, where each 1-D address generator holds three identical steppers (fig. 2 a). Each address stepper is programmed by an initial base value (delivered by the base stepper, e. g. B0) which is iteratively modified by a step width (e. g. Δx) until limit value (delivered by the limit stepper, e. g.
The SSCU controls the updates of a scan window cache on the basis of its “handle” base address provided by the JumpGenerator. The updates of the Scan Cache between two the basis of its “handle” base address provided by the JumpGenerator. The updates of the Scan Cache between two.

For complex access pattern the control of the Task Manager is required. For instance an access pattern like the JPEG zig-zag scan requires frequent exchange of the parameters.

The AGU provides seventeen addressing modes: a 2-dimensional raster scan mode, a block scan mode is suited for spatial filtering, eight variations of a neighborhood search mode, a 2-dimensional indirect access mode to address external image memory, a FFT mode (1-D and 2-D), and, an address buses), with a CPU with a powerful set of arithmetic, logic, and single cycle bit-manipulation operations and with application specific hardware logic, on-chip memory, on-chip peripherals, and a highly specialized instruction set, a 40-bit ALU, including a 40-bit barrel shifter and two independent 40-bit accumulators. Separate program and data memory allow a high degree of parallelism, like, for example, three reads and one write in a single cycle. Additionally a 17-bit x 17-bit parallel multiplier is coupled to a 40-bit dedicated adder to perform a non pipelined single-cycle multiply/accumulate (MAC) operation. Furthermore the CPU includes a compare, select, store unit (CSSU) to perform add/compare selection operations as needed in the Viterbi operator. An exponent encoder is integrated to compute the exponent of a 40-bit accumulator value in a single cycle.

The Texas Instruments TMS320C54x DSP, is a fixed-point DSPs in the TMS320 family [20]. It provides hardware support to avoid too much address computation overhead including two dedicated address generators, eight auxiliary registers and two auxiliary register arithmetic units.

It combines an advanced Harvard architecture (with one program memory bus, three data memory buses, and four address buses), with a CPU with a powerful set of arithmetic, logic, and single cycle bit-manipulation operations and with application specific hardware logic, on-chip memory, on-chip peripherals, and a highly specialized instruction set, a 40-bit ALU, including a 40-bit barrel shifter and two independent 40-bit accumulators. Separate program and data memory allow a high degree of parallelism, like, for example, three reads and one write in a single cycle. Additionally a 17-bit x 17-bit parallel multiplier is coupled to a 40-bit dedicated adder to perform a non pipelined single-cycle multiply/accumulate (MAC) operation. Furthermore the CPU includes a compare, select, store unit (CSSU) to perform add/compare selection operations as needed in the Viterbi operator. An exponent encoder is integrated to compute the exponent of a 40-bit accumulator value in a single cycle.

The internal RAMs can be configured as data memory or as program/data memory. The DARAM can be accessed twice per machine cycle, the CPU can read and write to a single block in the same cycle. The TMS320C54x has an independent data-address generation logic (DAGEN, [20]) offering seven basic data addressing modes: immediate addressing, absolute addressing, accumulator addressing,
Adopt, register addressing, stack addressing, direct addressing, indirect addressing, memory-mapped small application-specific Address Calculation Units (ACUs) optimizations, amongst others.

Adopt includes target architecture selection, address suitable for custom processors in partitioned architectures. memories to cope with the increasing bandwidth requirements applications often making use of embedded distributed realizations environment [21] to minimize the cycle, area and memory port connects are routed through application DPUs). interconnection overhead in terms of delay, power and area associated with routing between underlying (hierarchical) memory architecture and the MMU, a distributed address unit architecture is then selected.

For indexed addressing, a fixed storage ordering is selected when multidimensional signals are presented. Typically this order is first optimized for size reduction by so called in-place mapping [23]. Then, an address expression (AE) for every manifest index expression of a signal located in a linearly addressed RAM is generated. In case of nested loops, the AEs are indexed by common loop iterators. This creates opportunities for cost minimization by using advanced algebraic transformations. In Adopt this is done by applying aggressive transformations on a global scale and over word-level arithmetic [24].

Two clearly differing target architecture styles for the address calculation unit can be identified: incremental address generation unit (iAGU) and custom address calculation unit (cACU). In the iAGU the Address Sequences (ASs) are generated and realized as a counter modified by a two- or multi-level logic filter, under control by the global/master cMMU controller. That is based on the original approach of [25] but extended for more general filters. The least area-and power-expensive solution can be to map the AS onto an iAGU rather than a cACU. This decision depends heavily on regularity properties of the target AS [26]. In the cACU case, the AEs are realized as Application-Specific Units (ASUs) with custom arithmetic building blocks selected from a library being a subset of the Cathedral-3 methodology for lowly multiplexed custom data-path synthesis [21]. This is the first style proposed as an alternative to the iAGU style for the synthesis of address generators [21] [26]. Many other ASs with irregular access patterns are very area hungry, depending on sequence length. Here the AE can be much more efficiently implemented on cACU - a trade-off for power and area since a regular access pattern implemented on a counter has potentially much less toggle activity than with a cACU.

For the ACU style it is better to first identify similar AE clusters (instead of scheduling first, followed by function, mux and register allocation) with the goal to obtain optimal cACU partitioning of the in ASU [27]. After cACU' partitioning, further regularity improvement between the shared clusters [28] aiming to decrease the multiplexing overhead is possible by applying local scope algebraic transformations. After that it is mapped by a commercial behavioral synthesis tool.

For iAGU synthesis also AE multiplexing possibilities are provided (at sequence level by expanding the AE at compile time). This way an interleaved AS version of the multiplexed AEs can be obtained. Heuristic measures based on first and second order differences between pairs of AS values are sufficient [29]. The result is a partitioned iAGU architecture in terms of counters and look-up tables. After counter sequence sharing decisions, the characteristics of the counter (i.e., counter modulo value, reset and increment state values, etc.) must be defined still for an optimal synthesis of the iAGU unit. Analysis of the (pseudo) periodicity properties of the target AS [21] [26] is exploited. Also, the optimal assignment of counter to address port-bit, detected during the architecture exploration phase is finally incorporated [26]. Once the exact definition of the different iAGUs has been obtained, RT-level merging possibilities amongst the resulting iAGUs can still be
3. Customized memory organization and access scheduling issues.

Also address generators having been surveyed in previous section are an important methodological ingredient for memory organization and access scheduling issues. Data Transfer and Storage Exploration (DTSE) [33] is a systematic methodology to optimize data dominated applications for memory size, energy consumption and bus load reduction. The step discussed in this paper involves the design of a (partly) custom memory organization. We believe that this can be directly matched to the opportunities that modern FPGAs like the Xilinx Virtex are providing. The final memory organization should be as power-efficient as possible, but on the other hand its performance must (just) meet the real-time constraints and the memory size should fit in the available memory space.

The design of a custom memory organization consists of two main tasks: introducing and exploiting parallelism, and designing the memory architecture. In the DTSE methodology, this results in two steps. First, the Storage Cycle Budget Distribution (SCBD) step trades off the costly memory bandwidth (i.e. the amount of parallelism of the memory architecture) with the speed that can be achieved with it [34]. This results in a set of constraints on the memory architecture in terms of required parallelism. Second, the Memory Allocation and Assignment (MAA) step constructs a low-cost memory architecture which satisfies these constraints. This detailed memory architecture allows to derive quite accurate data on transfer and storage related costs [35].

The goal of inserting the SCBD step before the actual Memory Allocation and Assignment is to explore different solutions in the speed versus cost (energy or area) trade-off space. This results in a set of so-called Pareto trade-off points: each point represents a solution for which it is not possible to find an alternative which is both faster and less costly. So for each of these points, it is possible to execute the algorithm in more/less time, but at the cost of a less/more expensive (energy-consuming) memory architecture. This freedom can be used in various ways to improve the quality of the design [36]. For instance, when combining several components into a system, the global time and power budget has to be distributed over these. Or even for a single component, the energy versus cycles trade-off can be used to optimize the power supply voltage selection and operating frequency.

Several other approaches have looked at custom memory architectures for a given execution order [37][38][39]. Only [40] takes into account the constraints (access conflict graph related) due to available memory bandwidth and the concurrent accesses imposed by the scheduling. It is directly based on the output of the SCBD step [34].

Many other steps related to background memory management are international Dubrovnik Conference on reconfigurable architectures. The reconfigurable data-procedural Xputer [10][11] machine architecture MoM (Map-oriented machine [41][42][43][44]) supports optimized address sequences and optimum storage schemes [45][46]. A broad overview of memory organization related issues can be found in [47].

4. Conclusions

The paper has shown that the requirements for memory communication architectures of data-stream-driven reconfigurable computing systems are different from those for traditional parallel computing systems, and, has given a survey on architectural resources and optimization methods to cope with the memory bandwidth problem, especially for (data-)stream-based computing systems, like reconfigurable (and hardened) datapath arrays.

5. Literature
