

Hardware Implementation of Generalized Profile Search on the GENSTORM Machine

Emeka Mosanya Jean-Michel Puiatti Eduardo Sanchez

Logic Systems Laboratory
Swiss Federal Institute of Technology
CH-1015 Lausanne, Switzerland

Emeka.Mosanya@epfl.ch

Abstract

We describe the hardware implementation of the ProfileScan algorithm, a very sensitive method to discover distant biomolecular sequence relationships. This is part of the GENSTORM project, aimed at providing a dedicated computer for biological sequences processing based on FPGAs.

1 Introduction

We focus on the *Generalized Profile Search* designed by the bioinformatic team at the Swiss Institute for Experimental Cancer Research [1][2]. A Generalized Profile can model a family of sequences with common biological properties, it is described using a well defined syntax [3], and the process to evaluate if a given sequence contains a subsequence corresponding to a given profile has also been defined. The results are the *pfscan* utility freely available by ftp¹ and the *ProfileScan WWW Server*².

We propose a hardware implementation of the *pfscan* utility in order to improve the response time offered to the user. This implementation targets the GENSTORM³ machine [4], an FPGA-based accelerator for biological sequence processing.

2 Architecture of the GENSTORM machine

Figure 1 shows an overview of the GENSTORM architecture. It is composed by several cards linked

¹<http://ulrec3.unil.ch/ftp-server/pftools>

²http://ulrec3.unil.ch/software/PFSCAN_form.html

³<http://lslwww.epfl.ch/pages/research/papers/genstorm>

together by a VME bus : a SPARC VME card, a dedicated disk controller, and one or several computing card(s).

The GENOME card (Figure 2) implements the scanning process of a *Generalized Profile* against a sequence with a systolic architecture [5] [6]. Its general structure is similar to a lot of other existing cards since we focus on the target application part and use a standard architecture. This card contains nine FPGAs of the Xilinx XC4000 family, eight processing units (PUs) and one controller unit (CU), a VME interface, and memory blocks. Each PU is connected to a local memory (512KB) and to its neighbors with a 30 bit-wide local connection. The CU is connected to two of the PUs and to the memory blocks. It is responsible for the data transfer and the generation of interruptions. The card is designed to manage the subdivision of large query entirely in hardware without communication with the host. It is its main originality.

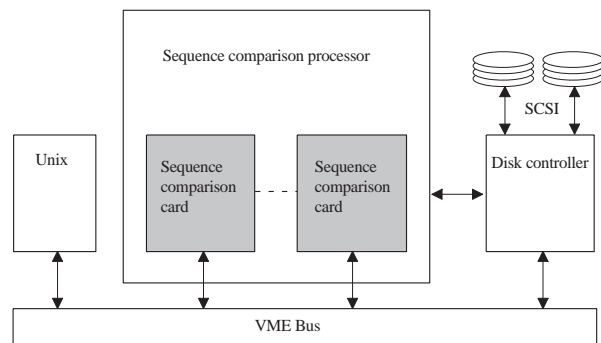


Figure 1: The architecture of the GENSTORM machine.

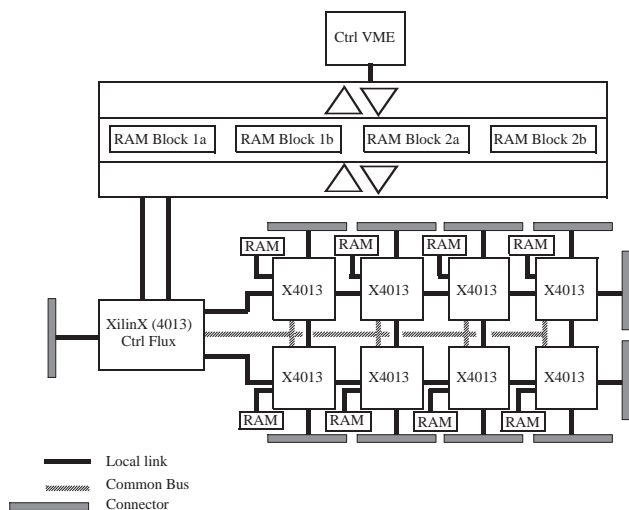


Figure 2: The architecture of the GENOME card.

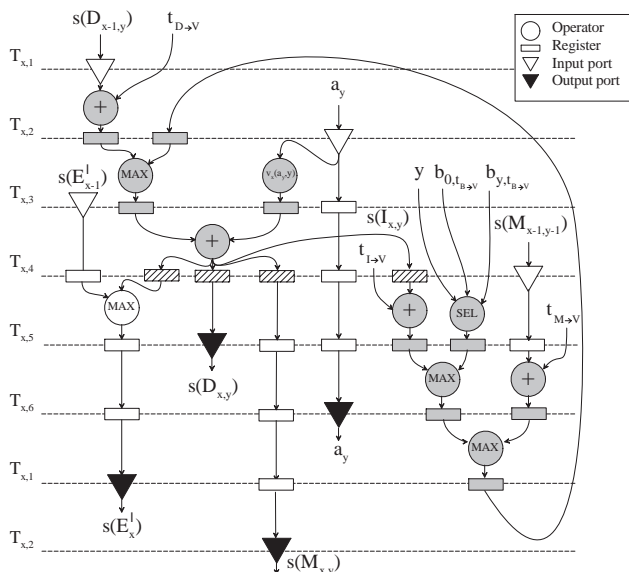


Figure 3: Scheduling diagram of score computation.

3 Parallel implementation of the score calculation

The scheduling shown in Figure 3 is a representation of the computation and data dependency carried on in each processing element. We observe that 6 clock cycles are needed.

The computation has been implemented in VHDL following the schedule presented above. It has been compiled with the Synopsys synthesizer with an input/output port size of 28 bits, a score size of 16 bits, a profile parameter size of 8 bits, and a sequence length

size of 10 bits. All values are integer. This configuration is realistic and can be used in the everyday life of a biology laboratory. The target is the Xilinx XC4000 family, since they are used on the GENSTORM machine.

4 Conclusion and perspectives

The *Generalized Profile Search* can be implemented with a systolic architecture on GENSTORM, a FPGA-based computer dedicated to sequence processing. The resulting design can compute a score value in 6 clock cycles, which, considering the length of each cycle, results in more than 2.8 millions profile positions by second for each computing card. These performance have been reached even with very slow devices.

A Profile Search Engine running on a SPARC 20 machine can be successfully replaced by GENSTORM with two computing cards to obtain a performance improvement of an order of magnitude. The machine, being based on reconfigurable FPGAs, can be easily reprogrammed online for other types of algorithm such as the well known Smith and Waterman sequence alignment [7].

References

- [1] Philipp Bucher and Amos Bairoch. A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation. In *Proceedings of the 2nd ISMD Conference*. AAAI Press, 1994.
- [2] Philipp Bucher, Kevin Karplus, Nicolas Moeri, and Kay Hofmann. A flexible search technique based on generalized profiles. *Computers and Chemistry*, 20, 1996.
- [3] Philipp Bucher. A generalised profile syntax for protein and nucleic acid sequence motifs. Technical report, Swiss Institute for Experimental Cancer Research, 1066 Epalinges s/Lausanne, 1997.
- [4] Jean-Michel Puiatti. Genome : documentation de la carte. Technical report, EPFL-Logic Systems Laboratory, INN Ecublens, 1015 Lausanne, 1995.
- [5] Patrice Quinton et Yves Robert. *Algorithmes et architectures systoliques*. Masson, 1989.
- [6] Daniel Philip Lopresti. *Discounts for dynamic programming with applications in VLSI processor arrays*. PhD thesis, Faculty of Princeton University, 1987.
- [7] T.F. Smith and M.S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147, 1981.