Accelerating Virtual Screening of Compund Libraries

Péter Szántó, Béla Fehér
Dept. of Measurement and Information Systems
Budapest University of Technology and Economics
Budapest, Hungary
{szanto, feher}@mit.bme.hu

Attila Bérces
Chemistry Logic Ltd.
Nyúl, Hungary
attila.berces@chemistrylogic.com

***Abstract*—The aim of virtual screening is to find compounds in libraries which exhibit the required properties. These properties are represented in fingerprints; during the screening process the descriptors of the compounds are compared to each other and a dissimilarity score is calculated. As compound libraries typically contain a large amount of fingerprints, comparison using CPUs takes a very long time. This paper presents hardware architecture for FPGAs which can drastically shorten the time required for screening.**

I. INTRODUCTION

The virtual screening process uses two large compound databases which contain molecular descriptors. The aim of the process is to find compounds in the *search database* which are dissimilar to the compounds in the *reference database*. Thus, each descriptor in the *search database* is compared to the descriptors in the *reference database* and a dissimilarity score is computed. The screening process accepts compounds in the *search database* which have dissimilarity score larger than a predefined threshold value; other compounds are rejected during the calculation.

As a general definition, a molecular descriptor is nothing more than a set of values associated with the 2D or 3D structure of the molecule. The presented hardware architectures support the most widely used descriptors, namely binary and pharmacophore fingerprints. There are numerous methods to generate the dissimilarity score for two molecule descriptors (e.g. Euclidean distance, Tanimoto metric), the accelerator employs the most popular method, the Tanimoto coefficient.

A. Tanimoto distance for binary fingerprints

Binary fingerprints are a set of binary values. Each binary value represents the presence or the absence of given chemical properties. During the fingerprint generation linear paths consisting bonds and atoms are detected; for each different paths (patterns) a set of bits in the fingerprint is set. It is allowed that different patterns set the same bit (this is often referred as bit collision), therefore fingerprint generation is a non-reversible process. Increasing the number of bits in the molecular fingerprint increases the information storing capability of the descriptor and reduces the possibility of bit collisions. However, too long fingerprints decrease the information storage efficiency and increases the time required for the screening process. Generally, 512 bit fingerprints offer a good compromise, but for similarity calculations 1024 bit descriptors often considerably improves the selectivity.

For binary fingerprints, the Tanimoto similarity coefficient is defined with the following equation.

$$S_{a,b} = \frac{c}{a+b-c}, \tag{1}$$

Where $S_{a,b}$ is the similarity coefficient, a is the number of '1' bits in the *reference fingerprint*, b is the number of '1' bits in the *search fingerprint* and c is the number of '1' bits which can be found in both fingerprints. $S_{a,b}$ results in value of 1 for similar fingerprints and in value of 0 for different fingerprints. During the screening process, dissimilarity is the useful information, therefore in this application the Tanimoto dissimilarity metric is used, which can be generated from the similarity metric straightforwardly:

$$D_{a,b} = 1 - S_{a,b} = 1 - \frac{c}{a+b-c}, \tag{2}$$

For every *reference molecule*, the screening process results in a set of *search database* molecules, where the dissimilarity metric is larger than a pre-defined threshold value, that is:

$$D_{a,b} = 1 - S_{a,b} = 1 - \frac{c}{a+b-c} > D_{\text{limit}}, \tag{3}$$

Rearranging Eq. 3.:

$$\frac{a+b-2c}{a+b-c} > D_{\text{limit}}, \tag{4}$$

Furthermore:

$$a+b > c * \frac{2 - D_{\text{limit}}}{1 - D_{\text{limit}}} \tag{5}$$

The advantage of Eq. 5. compared to the original expression is that for a defined threshold value the factor of c is constant.

B. *Tanimoto distance for pharmacophore fingerprints*

Pharmcophore fingerprints are based on the topological properties of a molecule, more exactly on the shortest path between atom-pairs. The fingerprint consists of several histograms. One histogram is associated with each pharmacophore feature pairs, e.g. acceptor-acceptor, donor-positive, etc. The histogram shows the distribution of the distances between the feature pairs. The total number of histograms stored in the fingerprint is $f*(f+1)/2$, where f denotes the number of feature pairs used during the fingerprint generation. Typically, 6 feature pairs are used with 10 bin histograms and a single bin is represented on a single byte. Therefore, the total number of bits stored in a phamacophore fingerprint equals to $(6*7/2)*10*8=1680$. The Tanimoto dissimilarity metric can be computed with Eq. 5. if binary histogram values are represented with unary numbers (that is, each 8 bit binary value is coded on 255 bits and the number of right justified '1' bits equals to the binary value). Unfortunately unary representation increases fingerprint size by a huge amount, therefore – although different type of computation is required – it is much more efficient to use Eq. 6.

$$D_{a,b} = \frac{\sum_i \min(a_i, b_i)}{\sum_i a_i + \sum_i b_i - \sum_i \min(a_i, b_i)} = \frac{\sum_i \min(a_i, b_i)}{\sum_i \max(a_i, b_i)} > D_{\text{limit}} \quad (6)$$

Where a_i is a single histogram bin in the *reference fingerprint* and b_i is the histogram bin of the *search fingerprint* in the same position.

Rearranging Eq. 6. leads to Eq. 7.

$$0 > D_{\text{limit}} * \sum_i \max(a_i, b_i) - \sum_i \min(a_i, b_i) \quad (7)$$

II. TARGET PLATFORM

The properties of the target platform define several aspects of the hardware architecture of the accelerator. In our case, the selected technology was the Silicon Graphics RC100 blade integrated into a SGI Altix server. The RC100 accelerator contains two Xilinx Virtex-4 FPGAs (XC4VLX200) which are connected to the system via SGI's proprietary, high speed NUMALink interface. For local data storage five banks of high speed QDR SRAMs are available. Figure 1. shows the high level block diagram of a single FPGA in the RC100 blade. The FPGA connection to the SRAMs offers 1.6 GB/sec bandwidth per bank per direction; the host (via the TIO chip) can access the FPGA with a maximal theoretical bandwidth of 3.2 GB/sec/direction.

The FPGA's internal SRAM interface (provided by SGI) concatenates two SRAM banks together and provides two 128-bit wide interfaces operating at 200 MHz. That is, the data width the accelerator architecture should process in a single clock cycle is 128 bit. To allow parallel host and FPGA accesses to the SRAMs, a double buffered operating scheme is employed.

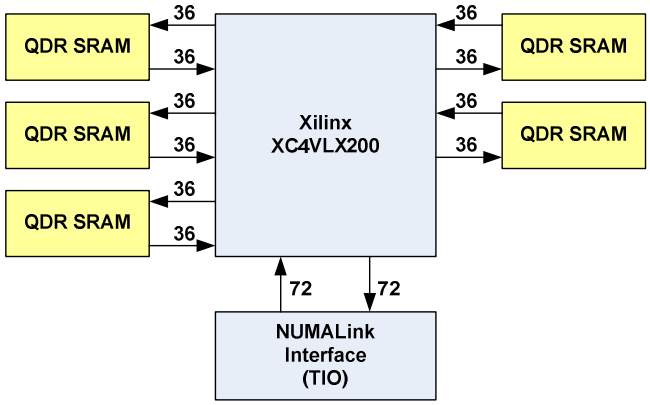

Figure 1. RASC100 FPGA architecture

From the FPGA point of view, one of the 128-bit SRAM interfaces (each of them consists of two SRAM banks) acts as a source and the other one acts as a sink. The source banks are written by the host and read by the FPGA, while the sink banks are written by the FPGA and read by the host. At any given time, half of the source memory is associated to the host and the other half is associated to the FPGA – so data writing by the host and data reading by the FPGA can happen in parallel (the same is true for the output memory).

III. ACCELERATOR ARCHITECTURE

The configurations for processing binary and pharmacophore fingerprints share a large part of the architecture, which is shown on Figure 2.

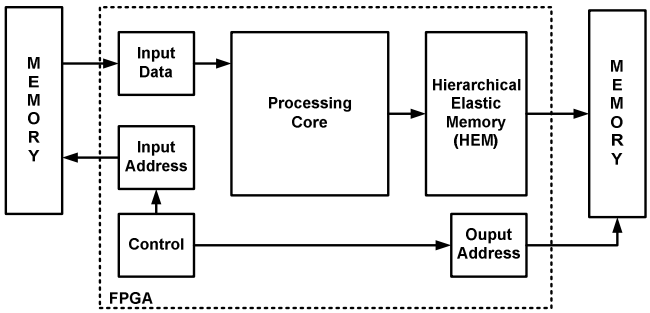

Figure 2. General FPGA architecture

The Control block (which handles host and memory interface communication, and controls the Processing Core), the memory interface and the output buffer architecture is shared between the two configurations; only the Processing Core is changed for different type of fingerprints.

Both the input data and output data are stored in the external memories. The Processing Core (PC) contains several Processing Units (PU), which operate on a single input data in parallel – data processing rate is one 128 bit input word in a single clock cycle. The Processing Core is responsible for computing the Tanimoto dissimilarity coefficient and accepting or rejecting molecules. Accepted molecules are presented on the outputs of the PC. The output data rate of the Processing Units is completely data and setting (Tanimoto threshold) dependant; therefore no preliminary assumptions can be made. Moreover, the different PUs can generate outputs at a very different rate – the Hierarchical Elastic Memory is responsible for load balancing between the PUs.

Irrespectively of the fingerprint type, comparing the reference database to the search database involves three steps:

1. Load N *reference fingerprints* into the N Processing Units.
2. Load the whole *search database* and compare the fingerprints with the N *reference fingerprints*.
3. If the *reference database* contains unprocessed fingerprints, go to step 1.

The *reference fingerprints* in the PUs are stored in 128 bit wide shift registers which are built up from SRL primitives, thus allow addressable output. During Step 1. the input data is shifted into the appropriate shift register; during Step 2. the SRL output is compared to the input data read from the external memory.

A. Processing Core for binary fingerprints

The Processing Core consists of 128 similar Processing Units and a Primary Processing Unit (PPU). Unlike PUs, the PPU does not store any fingerprints; it has two functions:

- During Step 1. the PPU computes the number of '1' bits in the incoming *reference fingerprints* and passes this information to the appropriate PU, which stores this value (*a* in Eq. 5.).
- During Step 2., the PPU computes the number of '1' bits in the incoming *search fingerprints* and passes the information to all PUs (*b* variable in Eq. 5.)

As the *a* value is stored in every PU, and *b* is computed by the PPU, the PUs only have to generate the *c* value; that is *AND* the two fingerprint-parts and compute the '1' bits in the result.

The division in Eq. 5. is not directly computed for two reasons:

- Division in hardware requires a lot of hardware resources
- Fixed point calculation would result in precision problems

Instead, a local memory stores the threshold-dependent portion of Eq. 5., that is memory location *c* is loaded with the expression in Eq. 8.

$$MEM[c] = c * \frac{2 - D_{\text{limit}}}{1 - D_{\text{limit}}}, \tag{8}$$

During the threshold computation, this memory is addressed with *c*, and the memory output is compared to (*a*+*b*).

The input memory interface is 128 bit wide, therefore processing a single 1024 bit fingerprints requires at least 8 clock cycles. Thus, to spare resources, 8 PUs are grouped to form an Octal Core, as Figure 3. shows.

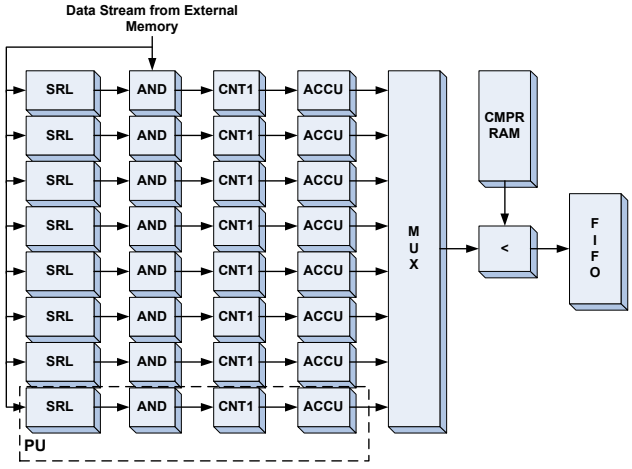

Figure 3. Block diagram of an Octal Core

A single PU consists of the previously mentioned blocks: a shift register (SRL), 128 two-input AND gates (AND), a bit summarizer unit (CNT1) and an accumulator (ACCU).

The CNT1 block is a pipelined structure to sum 128 bits. At the first stage K input bits are summed, while the following stages contain two-input adders.

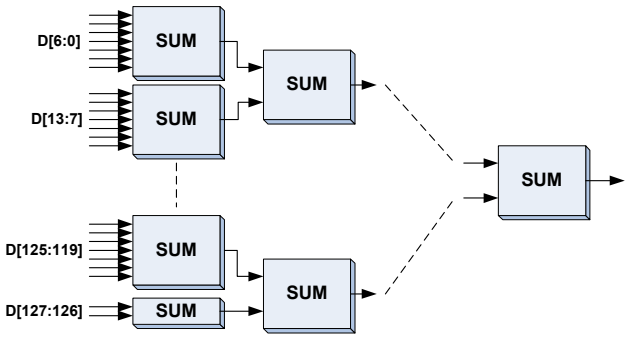

Figure 4. CNT1 architecure

To fulfill the performance requirements (300 MHz for sub-modules after post synthesis), yet spare resources K=7 was selected – that is, the first stage contains 19 bit-summarizers and the pipeline is 6 stages deep.

The accumulator accumulates 8 consecutive results of the CNT1 block generating the number of '1' bits in the 1024 bit fingerprint.

Threshold comparison for the 8 PUs is done in a time divided manner during 8 clock cycles, sharing a single memory (CMPR RAM) and comparator. If the comparison result is true, the fingerprint indexes and the *(a+b)* and *c* values are written into a 64-bit wide, 1024 word deep FIFO. The read port of this FIFO acts as the output port of the Octal Core, and connects directly to the Hierarchical Elastic Memory.

B. Processing Core for pharmacophore fingerprints

In the hardware, pharmacophore fingerprints are represented on 1792 bits, thus the 1680 bit real descriptor is extended with zeros by the host software. Just in the case of binary fingerprints, each PU in the Processing Core stores a reference fingerprint which is compared to the incoming search fingerprint in Step 2. From the two fingerprints (reference and search) each 8 bit histogram bin is compared to find the minimum and maximum values and then these values are summarized as Figure 5. shows (*r()* and *s()* denote histogram bins from the reference and search fingerprints).

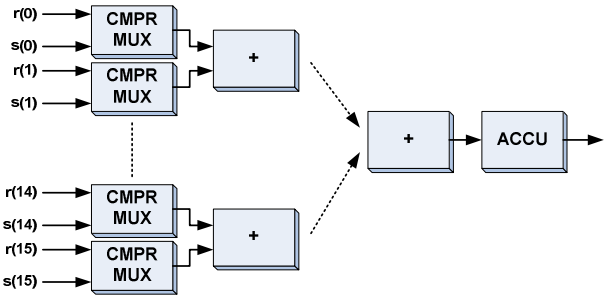

Figure 5. Adder tree

Minimum/maximum values are selected with a multiplexer which select input is driven by a comparator (CMPR MUX), and then the appropriate 8 bit values are summarized with an adder tree. The output of the adder tree drives an accumulator which accumulates 1792/128=14 consecutive outputs to form the final values in Eq. 6.

As the summarized bin values can be as large as 53550, the division cannot be avoided the way it was done with the binary fingerprints. However, as Eq. 7. shows, the division itself can be replaced with a multiply-subtract-compare operation. This operation fits the DSP48 blocks in Xilinx Virtex-4 FPGAs well; as these contain a multiplier and an adder/subtractor (and comparing to 0 is nothing more then inspecting the MSB bit of the result).

C. Hierarchical Elastic Memory

As it was mentioned earlier, the output data rate of the Processing Units, and hence the Octal Cores, depends largely on the actual fingerprint data and the set threshold value. If any of the Octal Core output FIFOs becomes full, processing new inputs should be hung to avoid data loss – therefore it is critical to balance the number of data residing in these FIFOs. This load balancing is achieved with a hierarchical, dynamic priority based FIFO structure. The block diagram of the Hierarchical Elastic Memory is shown on Figure 6.

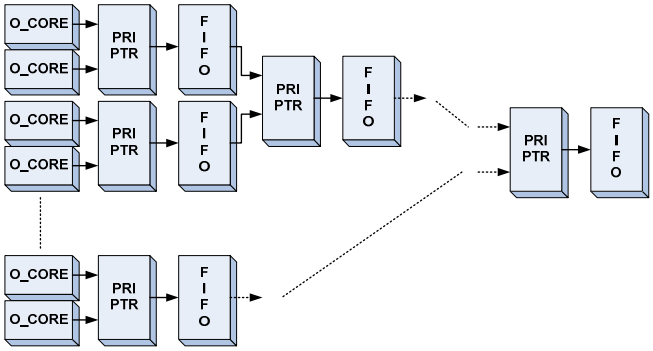

Figure 6. Hierarchical Elastic Memory

For every two input data, each level of the structure contains a priority based input selector (PRI PTR) and a 64 bit wide, 1024 word deep FIFO. The input priority is determined by the FIFO status of the previous level – the input which has more data in the FIFO gets the higher priority.

IV. CONCLUSION

For both fingerprint types, 128 parallel Processing Units can be fitted into a single Xilinx Virtex-4 LX200 FPGA. In this configuration, 64% of the logic resources and 44% of the internal memory resources are used. Although all hierarchical sub-modules was designed to be able to operate well over 200 MHz (when implemented alone), Xilinx place-and-route tool were unable to implement the full design and meet timing for 200 MHz operation. Therefore, final operating frequency was limited to 100 MHz.

Performance was compared to a well known industrial software. Using a single FPGA device, speedup was measured to be in the range of 100x-200x, depending on the database and the threshold setting.

REFERENCES

[1] Schneider, G.; Clement-Chomienne, O.; Hilfiger, L.; Schneider, P.;Kirsch, S.; Bohm, H-J. and Neihart, W. Virtual Screening for Bioactive Molecules by Evolutionary De Novo Design

[2] Osman F. Güner, Pharmacophore - Perception, Development, and use in Drug Design, International University Line, La Jolla, California, 2000

[3] Daylight Theory Manual, Daylight Chemical Information Systems, Inc.

[4] Reconfigurable Application-Specific Computing User's Guide, Silicon Graphics Inc.