Spiking Neural P system without delay simulator implementation using GPGPUs

Francis Cabarle
Algorithms and Complexity Laboratory
Department of Computer Science
University of the Philippines Diliman
fccabarle@up.edu.ph

Henry Adorna
Algorithms and Complexity Laboratory
Department of Computer Science
University of the Philippines Diliman
hnadorna@dcs.upd.edu.ph

ABSTRACT

This paper presents a parallel simulator for a type of P system known as spiking neural P system (SNP system) using general purpose graphics processing units (GPGPUs). GPGPUs, unlike the more conventional and general purpose, multi-core CPUs, are used for parallelizable problems due to their architectural optimization for parallel computations.

Membrane computing or P systems on the other hand, are cell-inspired computational models which compute in a maximally parallel and non-deterministic manner. SNP systems, w/c compute via time separated spikes and whose inspiration was taken from the way neurons operate in living organisms, have been represented as matrices.

The matrix representation of SNP systems provides a crucial step into their simulation on parallel devices such as GPGPUs. Simulating the highly parallel nature of SNP systems necessitates the use of hardware intended for parallel computations. The simulator algorithms, design considerations, and implementation are presented. Finally, simulation results, observations, and analyses using an SNP system that generates all numbers in $\{\mathbb{N} - 1\}$ are discussed.

Keywords

Membrane computing, Parallel computing, GPU computing

1. INTRODUCTION

1.1 Parallel computing: Via graphics processing units (GPUs)

The trend for massively parallel computation is moving from the more common multi-core CPUs towards GPGPUs for several significant reasons [7][8]. One important reason for such a trend in recent years include the low consumption in terms of power of GPGPUs compared to setting up machines and infrastructure which will utilize multiple CPUs in order to obtain the same level of parallelization and performance[9]. Another more important reason is that GPGPUs are architectured for massively parallel computations since unlike the architectures of most general purpose CPUs, a large part of GPGPUs are devoted for arithmetic operations and not on control and caching [7][8]. Arithmetic operations are at the heart of many basic operations as well as scientific computations, and these are performed with larger speedups when done in parallel, by GPGPUs over CPUs.

1.2 Paralel computing: Via Membranes

Membrane computing or its more specific counterpart, a P system, are Turing complete computing models that perform computations nondeterministically, exhausting all possible computations at any given time. This type of unconventional model of computation was introduced by Gheorghe Păun in 1998 and takes inspiration, similar to other members of natural computing, from nature[4][5]. Specifically, P systems try to mimic the constitution and dynamics of the living cell: the multitude of elements inside it, and their interactions within themselves and their environment, or outside the cell's skin membrane.

SN P systems differ from other types of P systems precisely because they are mono-membranar and only use one type of object in its computation. These characteristics, among others, are meant to capture the workings of a special type of cell known as the neuron. Neurons, such as those in the human brain, communicate or 'compute' by sending indistinct electro-chemical signals more commonly known as spikes. Information is then communicated and encoded not by the spikes themselves, since the spikes are unrecognizable from one another, by means of time duration as well as the number of spikes sent/received from one neuron to another, oftentimes under a certain time interval[1]. The time duration between two spikes, or several successive spikes, transmit information from one cell to another.

It has been shown that SN P systems, given their nature, are representable by matrices[2][3]. This representation allows design and implementation of an SN P system simulator using parallel computing machines such as GPGPUs.

1.3 Simulating SNP systems in GPGPUs

Matrix operations and their algorithms have been studied and efficiently implemented in GPGPUs [10][11]. Thus the matrix representation of SNP systems bridges the gap between the highly theoretical yet still computationally pow-

erful SNP systems and the applicative and more tangible GPGPUs, via an SNP system simulator. The design of the simulator, including the algorithms deviced, architectural considerations, are then implemented using a particular type of GPGPU, namely NVIDIA CUDA (compute unified device architecture). NVIDIA CUDA extends the widely known ANSI C programming language and makes parallel computations, via GPGPUs manufactured by NVIDIA.

This paper starts out by introducing and defining the type of SNP system that will be simulated. Afterwards the NVIDIA CUDA model and architecture are discussed, baring the scalability and parallelization CUDA offers. Next, the design of the simulator, constraints and considerations, as well as the details of the algorithms used to realize the SNP system are discussed. The simulation results are presented next, as well as observations and analysis of these results. The paper ends by providing the conclusions and future work.

2. SPIKING NEURAL P SYSTEMS

2.1 Computing with SNP systems

The type of SNP systems focused on by this paper are those without delays i.e. those that spike or transmit signals the moment they are able to do so [2][3]. A variant, which allows for delays before a neuron produces a spike, are also available [1]. An SNP system without delay is of the form:

$$\Pi = (O, \sigma_1, \ldots, \sigma_m, syn, in, out),$$

where:

1. $O = \{a\}$ is the alphabet made up of only one object, the system spike a.
2. $\sigma_1, \ldots, \sigma_m$ are m number of neurons of the form
$$\sigma_i = (n_i, R_i), 1 \leq i \leq m,$$
where:
   a) $n_i \geq 0$ gives the initial number of as i.e. spikes contained in neuron σ_i
   b) R_i is a finite set of rules of with two forms:
      (b-1) $E/a^c \rightarrow a$, are known as *Spiking rules*, where E is a regular expression over a, and $c \geq 1$, such that $c \geq 1$.
      (b-2) $a^s \rightarrow \lambda$, are known as *Forgetting rules*, for $s \geq 1$, such that for each rule $E/a^c \rightarrow a$ of type (b-1) from R_i, $a^s \notin L(E)$.
3. $syn = \{(i,j) \,|\, 1 \leq i,j \leq m,\ i \neq j\}$ are the synapses i.e. connection between neurons.
4. $in, out \in \{1, 2, \ldots, m\}$ are the input and output neurons, respectively.

Furthermore, rules of type (b-1) are applied if σ_i contains k spikes, $a^k \in L(E)$ and $k \geq c$. Using this type of rule uses up or consumes k spikes from the neuron, producing a spike to each of the neurons connected to it via a forward pointing arrow i.e. away from the neuron. In this manner, for rules of type (b-2) if σ_i contains s spikes, then s spikes are forgotten or removed once the rule is used. Rules of type (b-1) can be simplified with the notation

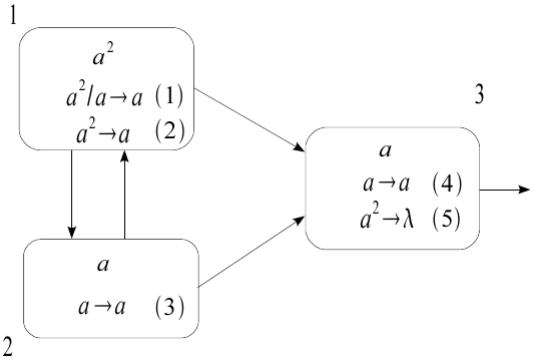

Figure 1: An SNP P system Π, generating all numbers in {ℕ - 1}, from [3].

(b-3) $a^k \rightarrow a$

where the regular expression $E = a^k$, again consuming k spikes and producing a spike.

The non-determinism of SNP systems comes with the fact that more than one rule of the several types are applicable at a given time, given enough spikes. The rule to be used is chosen non-deterministically in the neuron. However, only one rule can be applied or used at a given time [1][2][3]. The neurons in an SN P system operate in parallel and in unison, under a global clock [1]. For Figure 1 no input neuron is present, but neuron 3 is the output neuron, hence the arrow pointing towards the environment, outside the SNP system. The SNP system in Figure 1 is a 3 neuron system whose neurons are labeled (neuron/σ_1 to 3) and whose rules have a total system ordering from (1) to (5)

2.2 Matrix representation of SNP systems

A matrix representation of an SN P system makes use of the following vectors and matrix definitions [2][3] . It is important to note that, just as in Figure 1, a total ordering of rules is in order.

Configuration vector C_k is the vector containing all spikes in every neuron on the kth computation step/time, where C_0 is the initial vector containing all spikes in the system at the beginning of the computation. For Π (in Figure 1) the initial configuration vector is $C_0 =< 2, 1, 1 >$.

Spiking vector which shows, at a given configuration C_k, if a rule is applicable (has value 1) or not (has value 0 instead). For Π we have the spiking vector $S_k =< 1, 0, 1, 1, 0 >$ given C_0. Note that a 2nd spiking vector, $S_k =< 1, 0, 1, 1, 0 >$, is possible if we use rule (2) over rule (1) instead (but not both at the same time, hence we cannot have a vector equal to < 1, 1, 1, 1, 0>).

Spiking transition matrix M_Π is a matrix comprised of a_{ij} elements where a_{ij} is given as

$$a_{ij} = \begin{cases} -c, & \text{if rule } r_i \text{ is in } \sigma_j \text{ and it is applied consuming } c \text{ spikes;} \\ p, & \text{if rule } r_i \text{ is in } \sigma_s \ (s \neq j \text{ and } (s,j) \in syn) \\ & \text{and it is applied producing } p \text{ spikes in total;} \\ 0, & \text{if rule } r_i \text{ is in } \sigma_s \ (s \neq j \text{ and } (s,j) \notin syn). \end{cases}$$

For Π, the M_Π is as follows:

$$M_\Pi = \begin{pmatrix} -1 & 1 & 1 \\ -2 & 1 & 1 \\ 1 & -1 & 1 \\ 0 & 0 & -1 \\ 0 & 0 & -2 \end{pmatrix} \quad (1)$$

In such a scheme, rows represent rules and columns represent neurons.

Finally, the following equation provides the configuration vector at the $(k+1)th$ step, given the configuration vector and spiking vector at the kth step, and M_Π:

$$C_{k+1} = C_k \ + \ S_k \cdot M_\Pi. \quad (2)$$

3. THE NVIDIA CUDA ARCHITECTURE

NVIDIA, a well known manufacturer of GPUs, released in 2006 the CUDA programming model and architecture [9]. Using extensions of the widely known C language, a programmer can write parallel code which will then execute in multiple threads within multiple thread blocks, each contained within a grid of (thread) blocks. These grids belong to a single device i.e. a single GPGPU. Each device/GPGPU has multiple cores, each capable of running its own grids. The program run in the CUDA model scales up or down, depending on the number of cores the programmer currently has in a device. This scaling is done in a manner that is abstracted from the user, and is efficiently as well. Automatic and efficient scaling is shown in Figure 2 Parallelized code will run faster with more cores than with fewer ones [8].

Figure 3 shows another important feature of the CUDA model: the host and the device parts. As mentioned earlier, device pertains to the GPGPU/s of the system, while the host pertains to the CPU/s. A function known as a kernel function, is a function called from the host but executed in the device.

A general model for creating a CUDA enabled program is shown in Listing 1.

Listing 1: General code flow for CUDA programming

```
1 //allocate memory on GPU e.g.
2 cudaMalloc( (void**)&dev_a, N * sizeof(int)
3
4 //populate arrays
5  . . .
6
7 //copy arrays from host to device e.g.
```

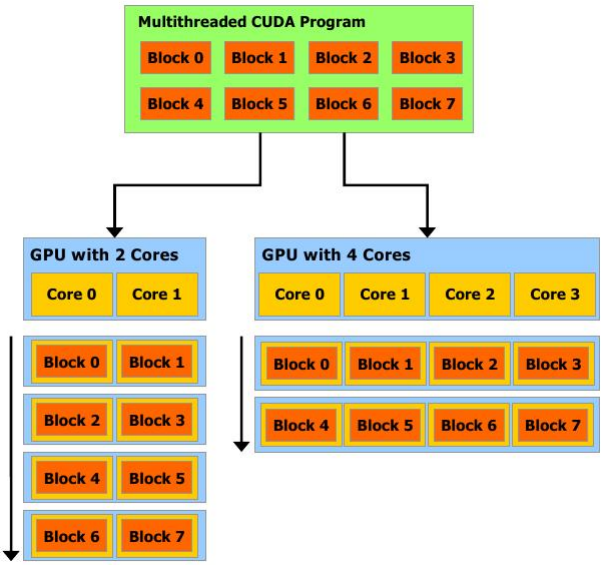

Figure 2: NVIDIA CUDA automatic scaling, hence more cores result to faster execution, from [8].

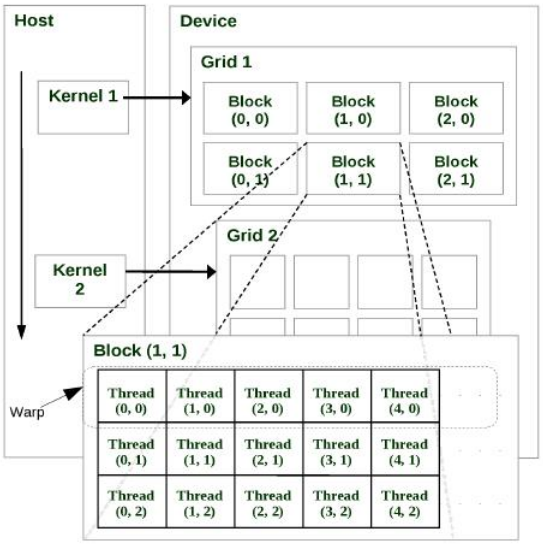

Figure 3: NVIDIA CUDA programming model showing the sequential execution of the *host* code alongside the parallel execution of the *kernel* function on the *device* side, from [6].

```
8  cudaMemcpy( dev_a, a, N * sizeof(int),
9  cudaMemcpyHostToDevice)
10
11 //call kernel (GPU) function e.g.
12 add<<<N, 1>>>( dev_a, dev_b, dev_c );
13
14 // copy arrays from device to host e.g.
15 cudaMemcpy( c, dev_c, N * sizeof( int),
16 cudaMemcpyDeviceToHost )
17
18 //display results
19
20 //free memory e.g.
21 cudaFree( dev_a );
```

Lines 2 and 19, implement CUDA versions of the standard C language functions e.g. the standard C function *malloc* has the CUDA C function counterpart being *cudaMalloc*, and the standard C function $free$ has *cudaFree* as its CUDA C counterpart.

Lines 8 and 15 show a CUDA C specific function, namely *cudaMemcpy*, which, given an input of pointers (from Listing 1 host code pointers are single letter variables such as *a* and *c*,while device code variable counterparts are prefixed by *dev_* such as *dev_a* and *dev_c*) and the size to copy (as computed by the *sizeof* function), moves data from host to device (parameter *cudaMemcpyHostToDevoce*) or device to host (parameter *cudaMemcpyDeviceToHost*).

A kernel function call uses the double < and > operator, in this case the kernel function

$add << N, 1 >> (dev\_a, dev\_b, dev\_c)$.

This function adds the values, per element (and each element is associated to 1 thread), of the variables dev_a and dev_b sent to the device, collected in variable dev_c before being sent back to the host/CPU. The variable N in this case allows the programmer to specify N number of threads which will execute the *add* kernel function in parallel, with 1 specifying only one block for all N threads.

3.1 Citations

Citations to articles [**?**, **?**, **?**, **?**], conference proceedings [**?**] or books [**?**, **?**] listed in the Bibliography section of your article will occur throughout the text of your article. You should use BibTeX to automatically produce this bibliography; you simply need to insert one of several citation commands with a key of the item cited in the proper location in the `.tex` file [**?**]. The key is a short reference you invent to uniquely identify each work; in this sample document, the key is the first author's surname and a word from the title. This identifying key is included with each item in the `.bib` file for your article.

The details of the construction of the `.bib` file are beyond the scope of this sample document, but more information can be found in the *Author's Guide*, and exhaustive details in the *LaTeX User's Guide*[**?**].

This article shows only the plainest form of the citation command, using `\cite`. This is what is stipulated in the SIGS style specifications. No other citation format is endorsed.

Figure 4: A sample black and white graphic (.eps format).

Figure 5: A sample black and white graphic (.eps format) that has been resized with the `epsfig` command.

3.2 Tables

Because tables cannot be split across pages, the best placement for them is typically the top of the page nearest their initial cite. To ensure this proper "floating" placement of tables, use the environment **table** to enclose the table's contents and the table caption. The contents of the table itself must go in the **tabular** environment, to be aligned properly in rows and columns, with the desired horizontal and vertical rules. Again, detailed instructions on **tabular** material is found in the *LaTeX User's Guide*.

Immediately following this sentence is the point at which Table 1 is included in the input file; compare the placement of the table here with the table in the printed dvi output of this document.

To set a wider table, which takes up the whole width of the page's live area, use the environment **table*** to enclose the table's contents and the table caption. As with a single-column table, this wide table will "float" to a location deemed more desirable. Immediately following this sentence is the point at which Table 2 is included in the input file; again, it is instructive to compare the placement of the table here with the table in the printed dvi output of this document.

3.3 Figures

Like tables, figures cannot be split across pages; the best placement for them is typically the top or the bottom of the page nearest their initial cite. To ensure this proper "floating" placement of figures, use the environment **figure** to enclose the figure and its caption.

This sample document contains examples of **.eps** and **.ps** files to be displayable with LaTeX. More details on each of these is found in the *Author's Guide*.

As was the case with tables, you may want a figure that spans two columns. To do this, and still to ensure proper "floating" placement of tables, use the environment **figure*** to enclose the figure and its caption.

Note that either **.ps** or **.eps** formats are used; use the `\epsfig` or `\psfig` commands as appropriate for the different file types.

3.4 Theorem-like Constructs

Other common constructs that may occur in your article are the forms for logical constructs like theorems, axioms, corollaries and proofs. There are two forms, one produced by

Table 1: Some Typical Commands

Command	A Number	Comments
`\alignauthor`	100	Author alignment
`\numberofauthors`	200	Author enumeration
`\table`	300	For tables
`\table*`	400	For wider tables

Figure 6: A sample black and white graphic (.ps format) that has been resized with the `psfig` command.

the command `\newtheorem` and the other by the command `\newdef`; perhaps the clearest and easiest way to distinguish them is to compare the two in the output of this sample document:

This uses the **theorem** environment, created by the `\newtheorem` command:

THEOREM 1. *Let f be continuous on $[a, b]$. If G is an antiderivative for f on $[a, b]$, then*

$$\int_a^b f(t)dt = G(b) - G(a).$$

The other uses the **definition** environment, created by the `\newdef` command:

Definition 1. If z is irrational, then by e^z we mean the unique number which has logarithm z:

$$\log e^z = z$$

Two lists of constructs that use one of these forms is given in the *Author's Guidelines*.

and don't forget to end the environment with figure*, not figure!

There is one other similar construct environment, which is already set up for you; i.e. you must *not* use a `\newdef` command to create it: the **proof** environment. Here is a example of its use:

PROOF. Suppose on the contrary there exists a real number L such that

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = L.$$

Then

$$l = \lim_{x\to c} f(x) = \lim_{x\to c} \left[gx \cdot \frac{f(x)}{g(x)} \right] = \lim_{x\to c} g(x) \cdot \lim_{x\to c} \frac{f(x)}{g(x)} = 0 \cdot L = 0,$$

which contradicts our assumption that $l \neq 0$. □

Complete rules about using these environments and using the two different creation commands are in the *Author's Guide*; please consult it for more detailed instructions. If you need to use another construct, not listed therein, which you want to have the same formatting as the Theorem or the Definition[**?**] shown above, use the `\newtheorem` or the `\newdef` command, respectively, to create it.

A *Caveat* for the TEX Expert

Because you have just been given permission to use the `\newdef` command to create a new form, you might think you can use TEX's `\def` to create a new command: *Please refrain from doing this!* Remember that your LATEX source code is primarily intended to create camera-ready copy, but may be converted to other forms – e.g. HTML. If you inadvertently omit some or all of the `\def`s recompilation will be, to say the least, problematic.

4. CONCLUSIONS AND FUTURE WORK

Using a highly parallel computing device such as a GPGPU, particularly NVIDIA CUDA, an SNP system simulator was successfully designed and implemented. The use of a high level programming language such as Python for host tasks, mainly for logic and string representation and manipulation of values (vector/matrix elements) provided the necessary expressivity to implement the algorithms created to produce and exhaust all possible and valid configuration and spiking vectors. For the device tasks, CUDA C allowed the manipulation of the NVIDIA CUDA enabled GPGPU which took care of repetitive and highly parallel computations (addition and multiplication essentially).

Future versions of the SNP system simulator will focus on several improvements. These improvements include the use of an algorithm for matrix computations without requiring the input matrix to be turned into a square matrix (this is currently handled by the simulator by padding zeros to an otherwise non-square matrix input). Another improvement would be the simulation of systems not of the form b-3). Byte-compiling the Python/host part of the code to improve performance as well as metrics to further enhance and measure execution time are desirable as well. Finally, deeper understanding of the CUDA architecture, such as inter- thread/block communication, for extremely large systems with equally large matrices, is required. These improvements as well as the current version of the simulator should also be run in a machine with higher versions of GPGPUs running NVIDIA CUDA.

5. ACKNOWLEDGMENTS

The authors are supported by the *ERDT Project*. They also wish to acknowledge the *Algorithms and Complexity laboratory* of *UP Diliman Department of Computer Science* for the use of Apple iMacs with NVIDIA CUDA enabled GPUs, which provided the proper environment for the simulations, their development, design and tests.

Figure 7: A sample black and white graphic (.eps format) that needs to span two columns of text.

6. REFERENCES

[1] M. Ionescu, Gh. Păun, T. Yokomori, "Spiking Neural P Systems", *Journal Fundamenta Informaticae* , vol. 71, issue 2,3 pp. 279-308, Feb. 2006.

[2] X. Zeng, H. Adorna, M. A. Martinez-del-Amor, L. Pan, "When Matrices Meet Brains", *Proceedings of the Eighth Brainstorming Week on Membrane Computing* , Sevilla, Spain, Feb. 2010.

[3] X. Zeng, H. Adorna, M. A. Martinez-del-Amor, L. Pan, M. Pérez-Jiménez, "Matrix Representation of Spiking Neural P Systems", *11th International Conference on Membrane Computing* , Jena, Germany, Aug. 2010.

[4] Gh. Păun, G. Ciobanu, M. Pérez-Jiménez (Eds), *"Applications of Membrane Computing"* , Natural Computing Series, Springer, 2006.

[5] P systems resource website. (2010, Jan) [Online]. Available: `www.ppage.psystems.eu`.

[6] J. Cecilia, J. Garcia, G. Guerrero, M. Martinez-del-Amor, I. Perez-Jurtado, M.J. Pérez-Jiménez, "Simulating a P system based efficient solution to SAT by using GPUs", *Journal of Logic and Algebraic Programming* , Vol 79, issue 6, pp. 317-325, Apr. 2010.

[7] D. Kirk, W. Hwu, *"Programming Massively Parallel Processors: A Hands On Approach"* , 1st ed. MA, USA: Morgan Kaufmann, 2010.

[8] NVIDIA corporation, *"NVIDIA CUDA C programming guide"* , version 3.0, CA, USA: NVIDIA, 2010.

[9] NVIDIA CUDA developers resources page: tools, presentations, whitepapers. (2010, Jan) [Online]. Available: `http://developer.nvidia.com/page/home.html` .

[10] V. Volkov, J. Demmel, "Benchmarking GPUs to tune dense linear algebra", *Proceedings of the 2008 ACM/IEEE conference on Supercomputing*, NJ, USA, 2008.

[11] K. Fatahalian, J. Sugerman, P. Hanrahan, "Understanding the efficiency of GPU algorithms for matrix-matrix multiplication", *In Proceedings of the ACM SIGGRAPH/EUROGRAPHICS conference on Graphics hardware (HWWS '04)* , ACM, NY, USA, pp. 133-137, 2004