An Overview of Hardware Implementation of Membrane Computing Models

Gexiang Zhang, Zeyi Shang, Sergey Verlan, Miguel Á. Martínez-Del-Amor, Chengxun Yuan, Luis Valencia-Cabrera, Mario Pérez-Jiménez

To cite this version:


HAL Id: hal-03550487
https://hal.science/hal-03550487
Submitted on 1 Feb 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
An Overview of Hardware Implementation of Membrane Computing Models

GEXIANG ZHANG*, Chengdu University of Technology, China and Southwest Jiaotong University, China
ZEYI SHANG, Southwest Jiaotong University, China and Université Paris-Est Créteil Val de Marne, France
SERGEY VERLAN, Université Paris-Est Créteil Val de Marne, France
MIGUEL Á. MARTÍNEZ-DEL-AMOR, Universidad de Sevilla, Spain
CHENGXUN YUAN, Southwest Jiaotong University, China
LUIS VALENCIA-CABRERA, Universidad de Sevilla, Spain
MARIO J. PÉREZ-JIMÉNEZ, Universidad de Sevilla, Spain

The model of membrane computing also known under the name of P systems is a bio-inspired large-scale parallel computing paradigm having a good potential for the design of massively parallel algorithms. For its implementation it is very natural to choose hardware platforms that have important inherent parallelism, like field-programmable gate arrays (FPGAs) or compute unified device architecture (CUDA)-enabled graphic processing units (GPUs). This paper performs an overview of all existing approaches of hardware implementation in the area of P systems. The quantitative and qualitative attributes of FPGA-based implementations and CUDA-enabled GPU-based simulations are compared to evaluate the two methodologies.

CCS Concepts: • Hardware → Simulation and emulation; Transaction-level verification; Semi-formal verification.

Additional Key Words and Phrases: Membrane Computing, P Systems, Hardware Implementation, Field Programmable Gate Array (FPGA), Compute Unified Device Architecture (CUDA), Graphic Processing Unit (GPU)

ACM Reference Format:

*Corresponding author

Authors’ addresses: Gexiang Zhang, Chengdu University of Technology, No.1, Dongsan Road, Erxianqiao, Chenghua District, Chengdu, 610059, China, Southwest Jiaotong University, 999 Xian Rd, Chengdu, 611756, China, zhgxdylan@126.com; Zeyi Shang, Southwest Jiaotong University, 999 Xian Rd, Chengdu, 611756, China, Zeyi.shang@lacl.fr; Sergey Verlan, Université Paris-Est Créteil Val de Marne, 61 av. du général de Gaulle, Créteil, 94010, France, sergey.verlan@u-pec.fr; Miguel Á. Martínez-del-Amor, Universidad de Sevilla, Avda. Reina Mercedes S/N, Seville, 41012, Spain, mdelamor@us.es; Chengxun Yuan, Southwest Jiaotong University, 999 Xián Rd, Chengdu, 611756, China, 502220232@qq.com; Luis Valencia-Cabrera, Universidad de Sevilla, Avda. Reina Mercedes S/N, Seville, 41012, Spain, lvalencia@us.es; Mario J. Pérez-Jiménez, Universidad de Sevilla, Avda. Reina Mercedes S/N, Seville, 41012, Spain, marper@us.es.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.

0360-0300/2020/6-ART $15.00
https://doi.org/10.1145/nnnnnn.nnnnnnn
1 INTRODUCTION

With the arising of protocells 3.84 billion years ago in hydrothermal vent precipitates [26], the evolution of unicellular organisms led to the apparition of multicellularity [83]. The biological cell structure defined by the membranes has evolved and was optimized for billions of years. As a consequence, a biological cell is a powerful parallel processing unit that can perform sophisticated biologic behaviors. Inspired by the structure of biological membranes and by internal bio-chemical reactions, Gheorghe Păun initiated the area of membrane computing in 1998 as a theoretical computer science model borrowing many concepts from the cell biology [84]. The model, also called P systems, features a nested membrane-like structure delimiting regions that host objects (modeling cell chemicals) which are transformed locally or communicated to other regions by different types of rules that mimic cell chemical transformations. The evolution of the state of the system is performed by transition steps doing a synchronous parallel execution of all rules. Since a typical model contains hundreds and even thousands of such parallel executions, P systems feature an inherent massive parallelism and the global behavior of the system is emerging from many simple interactions provided by rules application. There exist numerous variants of the model depending on the manipulated objects, used types of rules and the parallel execution strategy.

Membrane Computing is a computing paradigm inspired from the structure and functioning of living cells, and the organization of cells in tissues and other structures, including the brain. It provides distributed parallel computing devices called, generically, P systems. Membrane computing models are an extension of DNA computing, a possible source of novel/useful/interesting computing models. If one wants to model (1) discrete, (2) distributed, (3) parallel, (4) cell-like or tissue-like system, (5) dealing with multisets, (6) evolving by rewriting-like rules, then P systems are obligatory/unavoidable and the unique models dealing with all these features [85]. As unconventional computing devices within natural computing, P systems have proved to overcome the well-known limitations imposed by the conventional techniques based on electronic technology. More specifically, the relevance of these computing models can be non-exhaustively summarized in the following points. From a theoretical view, a novel methodology to tackle the famous P versus NP problem has been given [76]. The main focus in the area was the introduction of (biologically-inspired) variants of P systems and the further study of their computational power, in particular of their computational completeness, see [87] resuming hundreds of articles on this topic. From a practical view, the theoretical investigation led to a series of successful applications in different areas ranging from image processing [25, 114], meta-heuristic algorithms for optimization problems [116, 117, 119], robot controllers [14, 109] and path planning [79, 81, 110] and fault diagnosis for electrical power systems [90, 107, 108]. Due to its historical background, membrane computing was also used as a modeling framework for biological and ecological subjects such as artificial life [94], photosynthesis [75], p53 protein signaling pathway [95], myxobacterial colony [68], biopolymer duplication [56]. Please see a recent overview in [118].

For twenty years many applications of P systems have arisen, not only in Theoretical Computer Science, but also in Computational Modelling, Robotics, Optimization, etc., as mentioned above. The interest in this area is increasing, given that this unconventional, massively-parallel approach has been demonstrated to provide a powerful, flexible and expressive framework. One example is the 2016 report of the National Research Council of Canada where Membrane Computing appears to be mentioned several times as prominent parts of bio-computing [112]. Therefore, there is a need for simulators in order to build model validation tools, assistants for formal verification, environments for virtual experimentation, model calibration tools, etc. So far, most of the simulation software aim at reproducing only the computation of the devices. However, there have been also a research line concerning the implementation of P system parallelism in real parallel platforms. The main
hardware employed for this is FPGAs and GPUs, given their efficiency, fast shared memory system and scalability. The objective is twofold: on the one hand, provide efficient simulation tools where the massive, natural P system parallelism do not get serialized and is instead harnessed in order to speed up the simulations; and on the other hand, to explore how this special parallelism can be mapped in current modern computing architectures. It is known that the future of computer architecture will go through non-Von Neumann architectures. Natural Computing models can provide solutions where the instructions are closed to, or along with, the data. P systems are just an alternative, like artificial neural networks, for this. We want to shed a light into this alternative from the perspective of efficient implementation, analysing all the challenges and current solutions. In addition, a number of these applications mentioned above only theoretically profit from the potential speed-up promised by the model as there are no truly parallel implementations available. On the other hand, the inherent large scale parallelism of the model has the profound potential for the progress of extreme data processing, especially taking into account that there are not so many widely investigated massively parallel computational paradigms. Thus, an interesting topic is the implementation of membrane computing models on contemporary silicon integrated circuits. This allows to exploit the desirable parallel computational capability of P systems to explore a new orientation for high performance computing (HPC). As pointed out in "some unmistakable trends in hardware design indicate that uniprocessor (or implicitly parallel) architectures may not be able to sustain the rate of realizable performance increments in the future". The augmentation of electronic ingredients’ density had been subject to the well-known Moore’s law for decades. After extraordinary exponential growth of many years, the number of transistor in chips cannot follow this law, at least it cannot be doubled within two years. With transistors shrunk to nanoscale, quantum effects stand out and the behavior of circuits is not up to expectations. Moreover, even with the further increase of the density, the computational capability growth is not linearly proportional to it. Another knotty problem is the heat dissipation, which would melt the silicon substrate with density level increasing. Traditional semiconductor scaling is predicted to reach an end by about 2024 on the foundation of prior arts. Parallel computing has the potential to further uplift computing power provided that the density of transistor is constant with multicore and multithread architecture, although heat dissipation and interconnect issues would be challenges.

All these observations give strong arguments in favor of investigation of parallel computing platforms. Before claiming that the era of parallel computing has dawned, one essential question should be clarified: what does parallel computing mean? Though not rigorous, parallel computing implies computing based on the decomposition of the task into a set of concurrently executable operations and the assignment of these operations to multiple parallel processing nodes. The evolution of computer processor scheme, from single-core single central processing unit (CPU) to multiple-core single CPU and multiple-core multiple-CPU frame is an instance of parallel computing. The inherent parallelism of P systems place them in the class of multiple processing nodes computing devices. The mapping of constituents to processing nodes gives rise to different implementation strategies. A common practice in the area of natural computing is to observe natural processes and somehow mimic the corresponding behavior. The way how living cells allocate, organize and coordinate processing nodes has evolved for billions of years. Investigating this magnificent course might help us to handle the multiple cores computing, which has cut a striking figure in the contemporary parallel computing realm.

The large scale distributed parallel processes occurring in vesicle compartments and the vesicle division functionality enlightened from mitosis of living cells are two of the most outstanding advantages of membrane computing. They allow to underlie the foundation for the construction of highly parallel computation platform whose performance, flexibility and scalability outperforms...
traditional sequential counterparts substantially [118]. As a parallel computing paradigm inspired by the structural and functional features of biological membranes, only parallel computing platforms are suitable for the implementation of P systems. More precisely, the limited parallelism of general computers realized by the communication mechanism among the multiple cores of CPU and GPU cannot make full use of the large scale parallelism, non-determinism and other particular attributes that impart an enormous computing potential like creation and dissolution of inner membranes, the self-replication or *autopoiesis* [29] of the whole cell-like entirety that works as a computing unit, the communication by *symport* and *antiport* of objects [4], etc. We would like to remark that programming membrane computing algorithms with high level general purpose languages and executing them on the computer represents just a simulation and not a real implementation of P systems [85].

Several software-based and hardware-based parallel computing platforms have been developed to implement P systems. The first software-based parallel computing platform was constructed using a cluster of computers [22]. This platform achieves good performance and flexibility. Nonetheless, when the size of target P system is increased, the consumption of CPU time and resources caused by the communication between different computers rises dramatically. Moreover, the underlying hardware (a cluster of computers) of this platform cannot be miniaturized closing the way to corresponding membrane computing algorithms to be used in embedded chips and compact controllers which can be employed in robots, automobiles, machine tools, etc. This disadvantage limits the range of applications of such platforms for membrane computing.

Hence, it is important to propose hardware implementations of P systems as specific architectures that do not have the drawbacks related to the traditional ways of implementation. There are two main directions for such research using (1) Field-programmable gate arrays (FPGA) and (2) graphical processing units (GPUs) relying on Compute unified device architecture (CUDA) platform. In the first case a completely new parallel circuit is specially designed to implement some variants of P systems. In the second case the pre-defined CUDA parallel platform is used to simulate P systems. The achieved performance and correspondence is smaller in this case, but the development effort is much lower, so finally it becomes an interesting compromise between traditional computers implementations and a highly parallel one using specialized circuits. Also, when implementing membrane computing models in hardware, the main difficulty comes from the fact that there exists a big number of variations of the basic model of P systems having quite distinct characteristics [88]. This poses a great challenge for the conception of a general computational architecture to implement these various models.

The computation in a P system is a sequence of transitions between configurations. The core problem for implementations is the *object distribution problem* (ODP) that computes one of the multisets of applicable rules to the current configuration and whose application permits to reach the next configuration. This problem is a particular variant of a more general problem that computes the whole applicable set of multisets of rules for a given configuration and it is known to be NP-complete [21]. Known algorithms and heuristics do not parallelize well, so special heuristics were developed in order to quickly compute the desired multiset of rules. We decided to present these heuristics uniformly in terms of multi-criteria optimization, see Sections 4, 5. Another problem for implementations is that in the general case the model is non-deterministic, so an equitable choice among different possibilities should be provided. However, this is very difficult to achieve and in most of the cases this property is not satisfied.

This paper is organized as follows: Section 2 gives a brief description of the capabilities and of the architecture of hardware used, Section 3 recalls the definition of the most general model of P systems, based on the formal framework [35]. Next Section 4 expresses object distribution problem (ODP) in terms of multi-criteria optimization and integer linear programming. Section 5
presents an overview of different simulation approaches, including the Direct Non-deterministic Distribution algorithm (DND) algorithm which is the base for handling non-determinism. Next Sections 6, 7 and 8 give more details on existing simulation approaches. At the end, conclusions and future research directions are discussed.

2 THE SELECTION OF HARDWARE

2.1 FPGA hardware

The FPGA is a reconfigurable hardware allowing to prototype digital circuits. The modification of circuits for FPGA is performed by altering the interconnections between circuit elements using a hardware description language, the most common ones being VHSIC Hardware Description Language (VHDL) and Verilog. The design can be performed on several levels of abstraction, ranging from the switch/transistor level until the behavioral level (corresponding to a Mealy machine [67]).

From structural point of view, an FPGA is an array of configurable logic blocks (CLB) inlaid in the matrix of interconnects. The CLBs work as containers for slice-organized logic cells, which are composed of look-up tables (LUTs) and flip-flops (FFs). The LUTs are used to implement different combinatorial circuits such as basic gates, decoders, encoders and multiplexers. The FFs are the sequential logic components acting as a memory element in the circuit. In order to produce a circuit, the interconnects of CLBs should be reconfigured. The interconnection in FPGA is performed by switch boxes routing signals between its logic blocks. Modern FPGAs use one of the following interconnect technologies: static RAM, flash memory and anti-fuse. The first one dominates the current FPGAs implementations. With the re-programmability, FPGAs comply with the model-oriented hardware implementation quite well.

2.2 CUDA-enabled GPU hardware

The multiple computing cores architecture originates from mainframe computers where it was used to improve the clock speed. Nowadays, it is hard to find a PC equipped with only one core. On the other side, the component integration scale of some High-end GPUs has outpaced the CPUs for the booming demand of graphics processing (advanced rendering and 3D vision) [52]. Currently, the GPU is a computing element as powerful as the CPU. Different from FPGAs, there are manufactured parallel architectures in GPUs. The advantage is that developers should just concern about the efficient utilization of these architectures and the drawback is that these frameworks are un-reconfigurable.

Nevertheless, the GPU is not a general processing unit which can handle other computing assignments except for graphics processing. The predicament has changed for the arise of compute unified device architecture, known as CUDA, from the leading chip vendor-NVIDIA corporation. CUDA is a technology that enables general-purpose computing on graphics processing units (GPGPU). Usually, when referring to CUDA, what is referred is not the parallel computing framework but a GPU supporting CUDA. A CUDA-enabled graphics processing unit is a universal parallel computing device which is suitable for the implementing of parallel algorithm models. The parallel computing behavior of CUDA is based on the execution of multiple compute kernels on the GPU. These compute kernels are without physical construction, but based on an abstract parallel programming model. In other words, CUDA does not alter the physical structure of GPU. CUDA programming model is based on heterogeneous computing, where the CPU (host) is the master node that controls the execution flow and launches kernels on the GPU (device) when massive parallelism is required [52]. A kernel is executed by a grid of (thousands of) threads. The grid is a two-level hierarchy, where threads are arranged into thread blocks of equal size. Each block and each thread is unequivocally identified by an identifier. In this way, threads and blocks can be distributed easily.
to different portions of data, or to compute different instructions. Threads from the same block can be synchronized using barriers, while those belonging to different blocks can only be synchronized by the end of the execution of the kernel.

A GPU contains a global memory, which has the biggest size, but has the longest access time and a shared memory, which is smaller but faster [52]. Although current GPUs contain cache memories, in order to accelerate memory accesses, best performance is achieved when doing it manually. Global memory is accessed by all threads launched in all grids, and also by the host, but shared memory is only accessible by threads in a block. Threads also have fast access to their own registers for single variables, and local memory (which is normally outsourced to global memory). Accesses to memory have to be carefully programmed, so that contiguous portion of data is read by consecutive threads (providing so called coalesced access), since this increases the memory bandwidth utilization.

Nowadays the architecture of GPUs is upgraded to Streaming Multiprocessors (SMs) which are composed of an array of Streaming Processors (SPs), working as computing cores. A thread set consisting of 32 threads named warp is the basic unit which a SM fulfills in its executions. A SM can manage multiple warps which are based on Single-Instruction Multiple-Thread (SIMT) model in effect. Each thread in a warp should commence its processing at the identical program address concurrently, although after beginning, threads can execute independently abiding by a sequential manner. The parallelism of CUDA is terminated when a warp branches or the memory stalls [61].

2.3 Other hardware

There were attempts to simulate certain types of P systems on micro-processor based architectures [47]. Although the performance of micro-processors is low, they are economical alternatives suitable for the developing of prototypes and verifying the design methods. The drawback is that in most of the cases the parallel nature of P systems is not exploited at all, often leading to inefficient implementations. It could be interesting to use out-of-order (OoO) execution [93] processors to tackle this problem, but this possibility was not investigated yet.

Custom application specific integrated circuits (ASIC) can also be employed to implement P systems. However, no such attempts exit at the moment of the writing of this paper. One of the main reasons is that the flexibility of the hardware platform constructed on ASIC is quite insufficient to adapt to the different variants of P systems. However, such attempts could still be meaningful as they would allow to simulate features of P systems on some tailored circuits with different properties, like low power consumption. Another possible direction is to use ASIC for particular commonly occurring computational cores and combine them with software execution [28, 103].

3 THE MODEL OF P SYSTEMS

We suppose the reader to have a knowledge of basic notions from formal language theory and membrane computing. We refer to [37, 87] and [91] for missing details. We will also closely follow [35, 105] for the definitions.

We recall that a multiset can be seen as a set whose elements can have greater than one multiplicity. We will use the string notation for multisets, i.e., a multiset $M$ will be represented by a string where the number of occurrences of each letter, corresponds to its multiplicity in $M$. We will denote by $|M|$ the size of the multiset $M$ and by $|M|_a$ the number of elements $a$ in $M$.

There exist many variants of P systems, see e.g. [85]. In this paper the presentation will be given in terms of the formal framework for P systems introduced in [35], see also [104, 105]. This framework is based on the model of network of cells specifying the structure and rules to be executed as well as on a set of 4 functions giving the semantics of the system. Different combinations of these parameters allow in most of the cases to construct for a given P system a network of cells that strongly bisimulates it (i.e. one step in the one system is simulated by one step in the other.
one). Moreover, in many cases there is a one-to-one correspondence between rules applied in each system, meaning that they are mostly indistinguishable. Hence, the framework for P systems can be seen as a common language to compare different P systems as well as to express notions related to this area. Moreover, other multiset rewriting-based models (like Petri nets) can be easily expressed into this framework giving a possibility to compare corresponding models. Finally, we would like to remark that in what follows we will use the variant of the framework supposing that the system structure does not change in time. An extension of the framework that permits to take into account notions related to P systems with dynamically evolving structure is given in [34], but corresponding definitions are too complex for the purpose of this paper.

3.1 Network of Cells

As pointed out in [33, 35, 104, 105], most types of (static structure) P systems can be seen as variants of parallel multiset rewriting (by using the algorithm called flattening). Since multiset rewriting level is not practical for system description and understanding, a higher-level concept called network of cells was introduced in [35]. This model augments multiset rewriting with the notion of spatial locations (cells) as well as the corresponding operations and it can be seen as a particular interpretation of the symbols. The works [35, 104, 105] define some basic building blocks in terms of network of cells and give several examples of the construction of widespread notions and types of rules in membrane computing using these blocks.

Below, we provide the definition of network of cells, taken from in [35]. We remark that the definition from [34] is slightly different, however both models coincide when the structure of the system does not evolve.

Definition 3.1 ([35]). A network of cells of degree \( n \geq 1 \) is a construct

\[
\Pi = (n, V, w, Inf, R)
\]

where

1. \( n \) is the number of cells;
2. \( V \) is an alphabet;
3. \( w = (w_1, \ldots, w_n) \) where \( w_i \in V^\circ \), for all \( 1 \leq i \leq n \), is the finite multiset initially associated to cell \( i \);
4. \( Inf = (Inf_1, \ldots, Inf_n) \), where \( Inf_i \subseteq V \), for all \( 1 \leq i \leq n \), is the set of symbols occurring infinitely often in cell \( i \) (in most of the cases, only one cell, called the environment, will contain symbols occurring with infinite multiplicity);
5. \( R \) is a finite set of rules of the form

\[
(X \rightarrow Y; P, Q)
\]

where \( X = (x_1, \ldots, x_n), Y = (y_1, \ldots, y_n), x_i, y_i \in V^\circ, 1 \leq i \leq n \), are vectors of multisets over \( V \) and \( P = (p_1, \ldots, p_n), Q = (q_1, \ldots, q_n), p_i, q_i, 1 \leq i \leq n \) are finite sets of multisets over \( V \).

We will also use the notation (omitting \( p_i, q_i, x_i \) or \( y_i \) if they are empty)

\[
(1, x_1) \ldots (n, x_n) \rightarrow (1, y_1) \ldots (n, y_n); [(1, p_1) \ldots (1, p_n)]; [(1, q_1) \ldots (n, q_n)].
\]

The above rule is applied as follows: objects \( x_i \) from cells \( i \) are rewritten into objects \( y_j \) produced in cells \( j, 1 \leq i, j \leq n \), if every cell \( k, 1 \leq k \leq n \), contains all multisets from \( p_k \) and does not contain any multiset from \( q_k \).

By taking for each rule the set of cells that are involved a hypergraph relation, called the structure of the system, is induced. Commonly, tree-like (for P systems) or graph-like (for tissue P systems) relations are considered.

ACM Comput. Surv., Vol. 1, No. 1, Article . Publication date: June 2020.
The configuration \( C \) of \( \Pi \) is defined as an \( n \)-tuple of multisets over \( V (u_1, \ldots, u_n) \) satisfying \( u_i \cap \text{Inf}_i = \emptyset, 1 \leq i \leq n \).

In order to define the \textit{computation} in network of cells according to some derivation mode \( \delta \) the following functions should be specified:

- **Applicable(\( \Pi, C, \delta \))** – the function taking a system \( \Pi \), a configuration \( C \) and a derivation mode \( \delta \) and yielding the set of multisets of rules of \( \Pi \) that can be applied to \( C \).
- **Apply(\( \Pi, C, R \))** – the function allowing to compute the configuration obtained by the parallel application of the multiset of rules \( R \) to the configuration \( C \).
- **\textit{Halt}(\( \Pi, C, \delta \))** – a predicate that yields true if \( C \) is a halting configuration of the system \( \Pi \) (in some derivation mode \( \delta \)).
- **\textit{Result}(\( \Pi, C \))** – a function giving the result of the computation of the \( P \) system \( \Pi \) when the halting configuration \( C \) has been reached.

Then the computation is a sequence of transitions where each transition step \( C \Rightarrow C' \) is defined as \( C' = \text{Apply}(\Pi, C, R) \), for some \( R \in \text{Applicable}(\Pi, C, \delta) \). As usual, this sequence starts with the initial configuration and ends with the final configuration for which the halting predicate \( \text{Halt} \) yields true. In more formal terms, the result of the computation of a network of cells \( \Pi = (n, V, w, \text{Inf}, R) \) working in the derivation mode \( \delta \) is defined as follows (we refer to [35] for more technical details):

\[
\text{Result}(\Pi) = \{\text{Result}(\Pi, z) : w \Rightarrow^* z, \text{Halt}(\Pi, z, \delta) = \text{true} \text{ and } \text{Halt}(\Pi, x, \delta) = \text{false for any } x : w \Rightarrow^* x \Rightarrow^+ z\}
\]

We remark that [35] gives an algorithm to compute the set of multisets of applicable rules, denoted as \( \text{Applicable}(\Pi, C, \text{asyn}) \), the algorithm to compute \( \text{Apply}(\Pi, C, R) \) and several definitions for \( \text{Halt} \) and \( \text{Result} \) functions. The most common way of halting is the total halting, which means that there are no more applicable rules and the most common way of getting the result is to consider the multiset of objects present in the halting configuration at some predefined cell.

At each step, the set of multisets of applicable rules \( \text{Applicable}(\Pi, C, \text{asyn}) \) can be restricted using a \textit{derivation mode}, denoted by \( \delta \), which specifies which sub-multisets are chosen for the next step application. The most common example is the maximally parallel derivation mode (\text{max}) which is defined as follows:

\[
\text{Applicable}(\Pi, C, \text{max}) = \{R \subseteq \text{Applicable}(\Pi, C, \text{asyn}) \mid \exists R' \in \text{Applicable}(\Pi, C, \text{asyn}) : R' \supset R\}.
\]

The filter condition states that only non-extensible multisets are considered in this derivation mode. We remark that there might be several such multisets, hence for the application a non-deterministic choice is used to take one of them. We also refer to [105] for a description of several other derivation modes.

**Example 3.2.** Consider the system \( \Pi = (O, w_1, R) \), having the alphabet \( O = \{a, b, c\} \) and the set of rules \( R = \{r_1 : (1, ab) \rightarrow (1, abc); r_2 : (1, bbc) \rightarrow (1, abb)\} \). Consider the configuration \( C = (1, a^2b^4c^2) \). Then, \( \text{Applicable}(\Pi, C, \text{asyn}) = \{r_1, r_1^2, r_2, r_2^2, r_1r_2, r_1^2r_2\} \). The maximally-parallel set of multisets of rules is the following:

\[
\text{Applicable}(\Pi, C, \text{max}) = \{r_1^2, r_2^2, r_1^2r_2\}
\]

The application of the multiset of rules \( r_1r_2 \) on \( C \) yields \( (1, a^4b^4c^3) \).

**Example 3.3.** Consider the system \( \Pi = (O, w_1, w_2, R) \), with \( O = \{a, b, c\} \) and \( R = \{r_1 : (1, a)(2, b) \rightarrow (1, b)(2, a); r_2 : (1, b) \rightarrow (2, b)\} \). Consider the configuration \( C = (1, a^3)(2, b) \). It is easy to observe that at each step only a single rule is applicable \( (r_1 \text{ or } r_2 \text{ alternatively}) \). The only possible sequence of rule applications is \( (r_1r_2)^4 \) after which no rule is applicable anymore. This sequence moves all
symbols a from cell 1 to cell 2. We remark that the above rules do not rewrite objects, but only move them in the structure.

3.2 Examples of P systems

Here we will give the formal framework description of some main variants of P systems that were targeted for an implementation.

Transitional (cell-like) P systems [77]. This model considers that cells are organized in a tree structure and uses rules of following type: (i, u) → (i, u')(j, u'')(k1, v1) ... (km, vm), where j is the parent of i and i is the parent of k1, ..., km, m ≥ 0. When |u| > 1 corresponding rules are called cooperative.

Symport/antiport P systems. This variant considers that objects are not transformed but rather moved in a tree-like or graph-like structure. This corresponds to rules of form (i, u)(j, v) → (i, v)(j, u) or (i, u) → (j, u).

Populational Dynamics P (PDP) systems [13, 92]. The structure of this model corresponds to several trees that have their roots linked together. There are two types of rules: (1) working inside some tree: (i, u)(j, v) → (i, u')(j, v'), where i is parent of j and (2) communication between tree roots: (r1, x) → (r1, x1) ... (rk, xk), where r1, ..., rk are the numbers of the corresponding tree roots. Moreover, each rule has a probability associated to it, so the derivation mode is maximally parallel, followed by a probabilistic choice between rules having the same left-hand-side.

Spiking Neural P systems [24, 49]. This model has a graph structure and restricts the alphabet of the system to be a single letter, however permitting and forbidding conditions are replaced by a regular expression check. The rules are of form (i, a) → (k1, a1) ... (km, am); (i, E), where E is the regular expression checking for the contents of cell i, a is the single symbol used in the alphabet and k1, ..., km are the cells linked to cell i. The derivation mode is sequential at the level of each cell (only one rule per cell is applied) and maximally parallel at the level of all cells.

(Enzymatic) Numerical P Systems [78, 86]. A multiset M can also be seen as a function M : V → N. The model of numerical P systems extends the notion of multiset to M : V → R^n. Hence, the configuration is a vector of real-valued variables and rules are equations describing how to update each variable during each step.

P Systems with Active Membranes [36]. This model corresponds to transitional P systems with an additional rule allowing to create new child cells.

4 REDUCTION OF THE COMPUTATION STEP TO MULTI-CRITERIA OPTIMIZATION

The computation in P systems corresponds to a finite sequence of transitions between configurations. Each transition C → C′ can be seen as a sequence of 4 steps: (1) computing the set of multisets of applicable rules (Applicable(Π, C, asyn)), (2) restricting it according to the derivation mode δ (Applicable(Π, C, δ)), (3) choosing one element from it R and (4) applying R to the configuration C yielding C′ = Apply(Π, C, R). The first step is the most time consuming, as the corresponding problem is generally NP-hard. As we show below that in some cases it is possible to reduce the first 3 steps (the computation of an element from Applicable(Π, C, δ)) to a multi-criteria optimization problem. This will allow us to express different algorithms used to compute this step by different authors in a common language for a better comparison. Another theoretical advantage of such reduction is an extensive standard vocabulary and a plethora of solving methods existing in the optimization area, that could be possibly applied for large-scale P systems. We would like to remark
one more time that the presented reductions are not possible in the general case, however they are possible for many concrete cases.

4.1 Preliminaries

A multi-criteria optimization problem (MCOP) is an optimization problem that involves multiple objective functions. In mathematical terms it can be stated as

$$\max(f_1(x), f_2(x), \ldots, f_m(x))$$

subject to $x \in X$

where $m \geq 2$ and $X$ is the set of feasible vectors (or solutions). This set is usually defined by some constraint functions. We can also consider the objective function as a vector: $f : X \rightarrow \mathbb{R}^m$, $f = (f_1, \ldots, f_m)$. For a feasible solution $x$, the vector $z = f(x)$ is called an objective vector or an outcome [50, 53].

When corresponding functions as well as $f_i$, $1 \leq i \leq m$ are linear, we speak about a multi-criteria linear optimization problem. We also remark that as for classical optimization problems the objective functions are minimized; the other cases like maximization or hybrid min/max can be easily reduced to the minimization one. When further $X \subseteq \mathbb{N}^k$, $k > 0$ and $f : X \rightarrow \mathbb{N}^m$ we speak about an integer multi-criteria linear optimization problem (IMCLOP).

In multi-criteria optimization, typically there is no feasible solution that minimizes all objective functions simultaneously. Hence, the main attention is focused on solutions that cannot be improved in any of the objectives without degrading some other objective(s). Such solutions are called Pareto-optimal. Formally, they are defined as preimages of maximal elements of the outcomes, which are also called Pareto front.

**Definition 4.1.** A vector $x \in X$ is Pareto-optimal for a MCOP (defined as above) iff there is no other vector $y \in X$ for which $f(x) < f(y)$, where $u < v$ iff $u_i \leq v_i$, $1 \leq i \leq k$, and $\exists j$, $1 \leq j \leq m$ such that $u_j < v_j$.

**Example 4.2.** Consider the following problem:

$$\max(r_1, r_2) \text{ subject to }\begin{align*}
    r_1 &\leq 5 \\
r_1 + 2r_2 &\leq 6 \\
r_2 &\leq 3 \\
r_1 &\in \mathbb{N}, \quad r_2 \in \mathbb{N}
\end{align*}$$

The corresponding feasible solutions and Pareto-optimal solutions are shown in Fig. 1. We note that Pareto-optimal solutions are $(5,0)$, $(4,1)$, $(2,2)$ and $(0,3)$.

![Fig. 1. Feasible and Pareto-optimal solutions from Example 4.2.](image)
4.2 Rule choice as IMCLOP

It is not difficult to see that the problem of the computation of elements from \( Applicable(\Pi, C, \max) \) can be reduced to IMCLOP. For the first time it was noticed in [27], but without any further development. For simplicity, we consider that \( \Pi \) is a transitional P system and has only one membrane (and no environment). If this is not the case, we can apply the flattening procedure reducing it to one membrane [33, 35]. So \( \Pi = (O, w_1, R) \).

Let \( R = \{ r_1, \ldots, r_m \} \) and \( O = \{ a_1, \ldots, a_n \} \). Consider that \( r_i : (1, u_i) \rightarrow (1, v_i), 1 \leq i \leq m \). Let \( C \) be the current configuration and let \( C_a = |C|_a, a \in O \).

We will consider a set of variables \( x_i, 1 \leq i \leq m \), that indicate the cardinality of corresponding rules in some parallel multiset \( M \in Applicable(\Pi, C, \text{asyn}) \), \( |M|_{r_i} = x_i, 1 \leq i \leq m \). Then the feasible set \( X \) of (asynchronous) solutions is defined by the following inequalities:

\[
\sum_{i=1}^{m} |u_i|_a x_i \leq C_a, \quad \forall a \in O,
\]

\[
x_i \in \mathbb{N}, \quad 1 \leq i \leq m,
\]

Inequalities (1a) state that the sum of all consumed objects is included in \( C \). Technically, for each object \( a \in O \) it is verified that the weighted sum (by the number of symbols \( a \) in the left-hand-side) of rule cardinalities is smaller or equal than the number of objects \( a \) in \( C \). We remark that system (1) can be also seen as a system of Diophantine equations and corresponding solutions are exactly describing the set \( X \) of feasible solutions.

The IMCLOP corresponding to the computation of \( Applicable(\Pi, C, \max) \) is defined by:

\[
\max(x_1, \ldots, x_m)
\]

subject to \((x_1, \ldots, x_m) \in X\).

It should be clear that Pareto-optimal solutions represent exactly the multiplicities of rules for some maximally parallel solution.

Example 4.3. Consider the system \( \Pi = (O, w_1, R) \), with \( O = \{ a, b, c \} \) and \( R = \{ r_1 : (1, ab) \rightarrow (1, abc); r_2 : (1, bbc) \rightarrow (1, abb) \} \). Consider the configuration \( C = (1, a^2 b^6 c^3) \). Then, the constructed IMCLOP corresponds to the one given in Example 4.2.

So this system has 4 Pareto-optimal solutions: \((5, 0)\), \((4, 1)\), \((3, 3)\) and \((0, 3)\), corresponding to multisets of rules \( r_1^5 r_2^1 \), \( r_1^4 r_2^2 \), \( r_2^5 r_2^2 \) and \( r_2^6 \), which are exactly the maximal multisets of rules applicable to \( C \).

4.3 Tentative solutions

One of the traditional approaches to solve multi-criteria optimization problems is called scalarization. It consists in reducing the corresponding MCOP to a single objective optimization problem by using a real-valued scalarizing function involving the objective functions and additional scalar or vector parameters and variables. This can also imply additional restrictions to the feasible set based on the newly introduced variables.

One of the “simplest” methods to solve multi-criteria problems is the weighted sum method, where we solve the following single objective optimization problem

\[
\max_{x \in X} \sum_{k=1}^{m} \lambda_k f_k(x)
\]

The weighted sum problem (3) is constructed using the scalar vector product of the vector of objective functions \( f \) and the vector of non-negative weights \( \lambda \in \mathbb{R}^m \) as a parameter. It is known that it allows to compute all Pareto-optimal solutions for convex problems by varying \( \lambda \) [27].

ACM Comput. Surv., Vol. 1, No. 1, Article . Publication date: June 2020.
In the literature on P systems some simple variants of the weighted sum method can be found. In [3, 89] the vector \( \lambda = (1, \ldots, 1) \) is considered (so the objective function is the sum of all variables). However, in this case only maximally parallel rule sets having a maximal number of rules are obtained. In terms of the formal framework [35] this corresponds to \( \text{max rules } \text{max} \) derivation mode.

Another attempt was done in [20, 21] where the parameters \( \lambda_k \) correspond to the size of the left-hand-side of rules \( (\lambda_k = |u_k|, r_k : u_k \rightarrow v_k) \). Such optimization problem finds only maximally parallel solutions involving the maximal number of objects, corresponding to \( \text{max objects } \text{max} \) mode in terms of the formal framework.

In [71] the set of maximally parallel multisets of rules is expressed as solutions of a system of Diophantine equations (roughly equations (1a)) with an additional constraint to be satisfied on a solution, expressed as another system of Diophantine equations.

Finally, in [9] the set of maximally parallel multisets of rules can be expressed as solutions to a system of equations defining some Diophantine sets. While the construction is similar to the one we give below, it is not trivial to manipulate Diophantine sets and it is not clear how to express the constraints as a single system of equations. Below, we show how to handle this problem by using same construction as for handling multiple "either-or" constraints in integer linear programming.

As above we consider that \( \Pi \) has only one membrane (and no environment). So, \( \Pi = (O, v_1, R) \).

Let \( R = \{r_1, \ldots, r_m\} \) and \( O = \{a_1, \ldots, a_n\} \). Consider that \( r_i : (1, u_i) \rightarrow (1, v_i) \), \( 1 \leq i \leq m \). Let \( C \) be the current configuration and let \( C_a = |C|_a, a \in O \).

In addition to the \( \lambda \) vector above we introduce an integer parameter \( M \in \mathbb{N} \) having a value that is sufficiently big (in fact it should be greater than the maximal possible value of any variable \( x_i \), \( 1 \leq i \leq m \) of any feasible solution). We remark that in general case of maximal parallelism it is impossible to limit the maximal values of variables, however, for practical applications, it is often possible to find such a bound.

We will modify the system (1) to construct a system of inequalities whose integer solutions will be maximally parallel multisets of rules for the configuration \( C \).

\[
\sum_{i=1}^{m} u_i a_i x_i \leq C_a, \quad \forall a \in O, \tag{4a}
\]

\[
\sum_{i=1}^{m} u_i a_i x_i + |u_k|_a + M z_a^k \geq C_a + 1, \quad 1 \leq k \leq m, a \in O, |u_k|_a > 0, \tag{4b}
\]

\[
\sum_{a \in O} z_a^i = N_i - 1, \quad 1 \leq i \leq m, \quad N_i = \sum_{a \in O} \text{sgn}(z_a^i) \tag{4c}
\]

\[
x_i \in \mathbb{N}, \quad 1 \leq i \leq m, \tag{4d}
\]

\[
z_a^i \in \{0, 1\}, \quad 1 \leq i \leq m, a \in O. \tag{4e}
\]

Inequalities (4a) are the same as (1a) and they state that the sum of all consumed objects is included in \( C \).

Inequalities (4b) state the maximality property of the rule set defined by \( x_1, \ldots, x_m \). It verifies that for each rule there exist at least one object whose remaining quantity is not sufficient to apply this rule. They are based on multiple "either-or" constraints representation in ILP. The big value of \( M \) and the inequalities (4c) ensure that only one constraint from (4b) will be considered (the other ones will be satisfied because of the big value of \( M \)).

Hence, any solution \( x_1, \ldots, x_m \) satisfying the system of inequalities (4) corresponds to a Pareto-optimal solution of (3), hence to a maximally parallel rule set \( M = r_1^{x_1} \ldots r_m^{x_m} \) applicable to configuration \( C \). We also remark that from the construction given above, it immediately follows that system (4) is Diophantine.
Example 4.4. Consider the system \( \Pi = (O, w_1, R) \), with \( O = \{a, b, c\} \) and \( R = \{r_1 : (1, abc) \rightarrow (1, ab); r_2 : (1, a) \rightarrow (1, bb); r_3 : (1, b) \rightarrow (1, cb)\} \). Consider the configuration \( C = (1, a^2b^3c^2) \). Then, we construct an ILP according to the rules above (we recall that \( M \) is a big integer number):

Derived from (4a):

\[
\begin{align*}
    x_1 + x_2 & \leq 2 \\
    x_1 + x_3 & \leq 3 \\
    x_1 & \leq 2
\end{align*}
\]

Derived from (4b) for \( r_1 \):

\[
\begin{align*}
    x_1 + x_2 + 1 + Mz_a^1 & \geq 3 \\
    x_1 + x_3 + 1 + Mz_b^1 & \geq 4 \\
    x_1 + 1 + Mz_c^1 & \geq 3
\end{align*}
\]

Derived from (4c) for \( r_1 \):

\[
z_a^1 + z_b^1 + z_c^1 = 2
\]

Derived from (4b) for \( r_3 \):

\[
x_1 + x_2 + 1 \geq 3
\]

Derived from (4b) for \( r_3 \):

\[
x_1 + x_3 + 1 \geq 4
\]

It is not difficult to see that this system has 3 solutions: \((2, 0, 1)\), \((1, 1, 2)\) and \((0, 2, 3)\), corresponding to multisets of rules \( r_1^2r_3, r_1r_2r_3^2 \) and \( r_2^2r_3^3 \), which are exactly the maximal multisets of rules applicable to \( C \).

We remark that all solutions of (4),(3) might be tedious to obtain. For the simulation purposes, only one such solution is necessary.

As previously mentioned, using the weighted sum scalarization technique it is theoretically possible to reach any single point from the Pareto front by choosing appropriate values of the parameter vector \( \lambda \). However, by using equations (4b) it becomes much easier to choose corresponding parameters as the feasible set is restricted only to Pareto-optimal values. For example, using \( \max x_i \) as objective function with the constraints from Example 4.4, would yield the solution \((2, 0, 1)\). By using \( \min (x_1 + x_2 + x_3 - 5) \) as the objective function, the solution \((0, 2, 3)\) is obtained (the above constraint allows to consider only multisets of rules of size 5).

In the literature on P systems, there are other examples of the construction of a single Pareto-optimal solution having certain properties. One of them is the DND algorithm introduced in [71] for the FPGA simulations that we discuss below.

In what follows, we explain briefly the functioning of the DND algorithm in the view of equations (1). First a random rule permutation is computed (corresponding to a random permutation of variable indices). Next, during the forward step a random value (bounded by the number of possible applications) for the number of each rule applications is taken. This corresponds to finding the values of \( x_i \), satisfying the constraints (1a). Finally, during the backward step, the frequency of each rule is increased until it cannot be applied anymore. This step corresponds to the elimination of the dominated, i.e. smaller, solutions for the system (1) yielding only Pareto-optimal ones.
The described procedure can also be seen as a combination of scalarizing and of the lexicographic method, which is a method from the family of a priori methods for multi-criteria optimization. More precisely, it corresponds to a solution of a series of optimization problems, each of them bounded by the parameter corresponding to the choice of the maximal rule multiplicity for the forward step. The backward step also corresponds to a series of optimization problems that just reach the maximum for each component (like in the lexicographical method).

A simpler variant of the DND algorithm that does not perform the initial rule permutation can be found in [60, 98, 100]. Other variations of DND algorithm were used for CUDA-based simulations and are discussed later. A different approach is used in [106]. It supposes that for a P system $\Pi$ working in the derivation mode $\delta$ there exists an easily computable function $NBVariants(\Pi, C, \delta)$ that for any configuration $C$ gives the number of solutions of (4). Next, it also supposes that there exists an easily computable function $Variant(\Pi, C, \delta, n)$ that for each integer $n$ (up to the corresponding value) yields the corresponding solution (the used method is similar to the decoding of a number in the combinatorial number system). It is clear that such functions do not exist for any variant of P system. The papers [88, 106] give some sufficient relations between rules that allow to construct the above functions for a P system working in set-maximal $(s_{max})$ derivation mode.

5 IMPLEMENTATION OF P SYSTEMS USING HARDWARE

The discussion about the hardware implementation of membrane computing concerns mainly the simulation of concrete variants (sometimes even examples) of P systems using a dedicated hardware (for ASIC and FPGA) or firmware (for CUDA). In most of the cases a (single) particular class of P systems is simulated. The simulator is composed from two parts: (1) the hardware simulator for a concrete system and (2) the software generator of hardware simulators, which based on input parameters (rules, membranes, initial configuration etc.) generates the code for the dedicated hardware simulator. In this section we will mainly discuss the structure of the corresponding hardware simulators, as the generator part is more-or-less following standard compiler construction techniques. We will concentrate on three points, which are the most important for a hardware implementation: (a) the representation of the configuration (multisets of objects and the membranes), (b) the representation of rules and their parallel application and (c) handling of the non-determinism.

5.1 Data organization

Membranes. We remark that there are no compartments in silicon circuits, hence the notion of membranes is relatively difficult to represent directly. Nevertheless, according to [85], a membrane is just an idealized concept without internal structures. The main functionality of membranes is to perform a topological division of the space allowing P systems to compute in a distributed manner (based on an correspondence between the membrane and its contents). Hence, the spatial placement and size of membrane are not important, only the inter-relationship among them matters. Moreover, it is known that any membrane structure can be reduced (flattened) to just a single membrane, see [33, 35] for more details. The existing hardware simulators adapt in most cases this last point of view, where the membrane structure is not physically implemented on the device. As exceptions from this rule we cite [82] and [73] (region-based), which implicitly implement the membrane topology by using dedicated buses and message passing in the corresponding circuits.

Configuration. The representation of the configuration in all cases is done as a vector of non-negative integers (stored in the memory/registers of the device). We remark that this vector corresponds to a flattened system, so it is relatively big and sparse, as it encodes each object/membrane pair. For performance reasons, it is physically split in several places to be closer to the processing units (as routing is relatively expensive). In the case of [73, 82] it can be argued that corresponding
parts are internalized into the corresponding region circuit, as the access to corresponding values is not direct and it is done by message passing. In CUDA, only the portions of the array representing the configuration in the corresponding membranes to be processed are loaded by the cores.

**Evolution Rules.** As it can be seen from definition 3.1 in the simple case (without permitting and forbidding) rules can be defined by 2 integer vectors indicating the multiplicities of corresponding objects in the left-hand-side and right-hand-side of each rule. This gives a natural rule representation as two vectors stored in the memory/registers. Then a specific circuit/module verifies the applicability of rules and performs their application. Most variants of hardware implementations use this idea, however in [70, 72, 82] each rule is encoded in hardware as a specific circuit that verifies the needed resources and performs the rule application. In the case of P systems with active membranes, there might be several cells with the same label, so the corresponding CUDA simulator uses a split representation: 2 integer vectors, as above, for all rules, and an index array with one position per rule giving for each cell the rules it contains.

### 5.2 Object Distribution Problem and Non-Determinism

The hardware implementation of P systems faces the problem of the computation of the applicable rule set according to some derivation mode (usually maximally parallel). More precisely, an efficient way to compute and represent an element from \( \text{Applicable}(\Pi, C, \delta) \) is required [88, 106]. The difficulty of the problem is that rules can compete for same objects, so increasing the number of occurrences for one rule, may decrease the application possibilities for another one. Another important problem is to ensure that a non-deterministic choice among all possibilities is performed. In [41, 44, 57, 60], hardware architectures aiming at parallel processing and communication, and the application of rules are developed. In [11], a formal exposition of non-deterministic evolution in transition P systems was suggested.

We call the first problem as **object distribution problem** (ODP). It consists in the computation of the set \( \text{Applicable}(\Pi, C, \delta) \) (or of an element from this set). As discussed in Section 4 in terms of multi-criteria optimization this corresponds to the computation of the corresponding Pareto front (or an element of it).

In [71] different algorithms solving ODP are classified in **direct** and **indirect** ones. In the direct approach, the corresponding multiset is directly constructed by the algorithm. In terms of MCOP this corresponds to a particular fixed scalarization. The indirect approaches are based on the observation that the solution number is finite because the solution space is bounded by the size of the configuration. Hence, a heuristic or brute-force approach can be used to explore this bounded space. However, since it is an overestimation, there might be visited elements that are not valid solutions. Hence, the algorithms are iterative and explore the whole space until a valid solution is encountered. In terms of MCOP this corresponds to different searches through the space limited only by the maximal values for each axis.

Sometimes it is not easy to classify an algorithm in one of these categories. We will classify an algorithm as a direct approach if its main goal is to construct a valid multiset of rules. Otherwise, if an algorithm is exploring different solutions until it reaches a valid one, it will be classified as indirect. We will use this classification to overview different strategies for ODP solution known in the literature.

**Indirect approaches.** Generally, the enumeration of all possible solutions and their verification one by one until a correct solution is obtained is the simplest method for the indirect approach [31]. Before the first correct solution is obtained, some invalid solutions should be rejected. This approach is called **indirect straightforward approach** [71]. Taking into account that it is not viable to enumerate all possible solutions for many problems, the feasibility of the approach is low. However,
the performance of the algorithm suggests its use as to compute the floor values for the object distribution problem. Another indirect approach discussed in [71, 73] called indirect incremental approach investigates a strategy generating possible solutions in rounds. Other attempts based on a similar idea but with different rule elimination strategies were done in [30, 39, 46, 48, 98, 99]

**Direct approaches.** In contrast to indirect approaches, the direct approach fabricates a solution straightforwardly rather than identifying a number of possible solutions before a solution is confirmed.

The simplest approach is the direct straightforward approach. As defined in the paper [71], in this approach “all the solutions to the object distribution problem are given as input, and one of these solutions is simply selected at random”. While in the same paper it is argued that such approach is infeasible for an arbitrary configuration and rule types, it can still be applied in a big number of cases. As shown in [88, 106], if at each step the number of solutions can be expressed as a the number of words of some length in a regular language, then it becomes possible to compute the solution only based on its number. In [106] it is shown that the corresponding class of P systems is quite large and also that this method is particularly interesting for bounded derivation modes like the set-maximal derivation mode (called also flat mode) where the rules are chosen in a set-maximal way (instead of the multiset maximal way).

Another variant of the direct approach is the Direct Non-deterministic Distribution algorithm (DND) proposed in [71]. A similar algorithm can also be found in [38, 40]. This algorithm works in 2 phases. At the first phase all rules (initially randomly shuffled) except one are selected to be applied a random number of times below its maximal applicability value. In the second phase, all the rules are taken in the converse order (with the first one being the rule excluded at the previous step) and their applicability is increased up to the maximal still possible value. A variant of DND, named DND-P, became popular in the simulation of Population Dynamics P (PDP) systems [65]. Together with another algorithm Direct distribution based on Consistent Blocks Algorithm (DCBA) [63] it was employed for the engine of the PDP system simulator on CUDA [62].

**Non-determinism.** One of the difficulties of the above approaches is the handling of the non-determinism. From the formal point of view, the non-determinism corresponds to a random equiprobable choice of an element from the set of all applicable multisets of rules (Applicable(Π, C, δ)). In the case of indirect approaches, due to the iterative nature of the algorithms, it is not easy to argue that each possibility has the same probability to occur. We would state that solutions containing a smaller number of different rules have a higher chance to be selected. In the case of DND algorithm and related variants, it looks like the obtained solution tends to be an equiprobable choice. However, the corresponding articles do not give such a proof and there are some unclear points, which do not allow us to affirm this fact. Up to now, the only algorithm that is performing a truly non-deterministic choice is the one described in [88, 106]. However the corresponding implementation is limited to some particular derivation modes and particular classes of P systems.

6 AN OVERVIEW OF EXISTING FPGA SIMULATIONS

With the advent of reconfigurable hardware which realizes the idea of modifying the hardware circuits by programming, conceiving a novel circuit simulating an innovative processing paradigm is no longer an exceedingly hard task. The first attempt to use FPGA reconfigurable hardware to simulate P systems dates back to 2003 [82]. Since then two simulation approaches emerged, considering regions or rules as basic processing units.
6.1 Region-based Simulations

In the region-based simulation approach rules and objects from different membranes are physically located in different places of the circuit, while those from the same membrane are physically close and well connected. The biggest problem is to ensure the correct communication of objects between membranes as this requires a global level synchronization. As advantage, the obtained system is highly scalable and robust. Below, we give two examples of region-based simulations.

In contrast, the rule-based implementation approach discussed later explicitly represents the evolution rules as processing units and multisets of objects as register arrays, while membranes and regions are represented implicitly as logical constructions existing between those processing units and data structures.

6.1.1 Petreska and Teuscher simulation. Petreska and Teuscher designed the first FPGA circuit simulating membrane computing (more precisely transitional P systems) [82]. They could realize several interesting features like communication to inner membranes, priorities between rules and proposed ideas how to simulate membrane creation and dissolving mechanism in integrated circuits. This work inspired the successors to engage in this challenging and breathtaking field to advance the development of hardware simulation of P systems. In their simulation they made two strong assumptions: the application of evolution rules in each membrane is not done in a maximally parallel but in a sequential manner (but still keeping a parallelism at the system level); the non-deterministic evolution of configuration is substituted by a deterministic transition following a predetermined order. Such computational step corresponds to an ILP with the subject function as a weighted sum of variables with predefined fixed weights.

In theory, membranes are borders without internal structures and material consistence. In this implementation, the membrane structures represent regions on the circuit containing the enclosed substances, i.e., the multisets of objects, evolution rules and children membrane architectures. The objects exchange among membranes is a kind of bi-directional traversing behavior. In case of the possible objects exchange between regions, the communications are made by data buses connecting to different parts of hardware representing inner membranes. To avoid the multiple buses used to connect the parent membrane to its plural children membranes, a single bus links all the children membranes before it connects to the parent membrane. Hence, the communication is limited to parent-child membranes and there is no object exchange among children membranes or non-immediate contained membranes.

The representation of the multisets of objects is implemented by using registers. Different registers just preserve different multiplicities of objects. A register does not store the objects but only a number indicating the multiplicity of each object. The order of these registers is in accordance with the lexicographic order of the alphabet of objects. The recognition of an object is indirectly realized by examining the position of the register storing the multiplicity of this object. An evolution rule defined here is in the form of \( u \rightarrow v(v_1, in_i)(v_2, out) \), where \( v_1 \) is the string to be sent into lower-immediate membrane labeled \( i \), \( v_2 \) will be sent to upper-immediate membrane. The treatment employed to deal with the formulation of evolution rules is storing the rule’s left-hand side and right-hand side into different registers separately. A particular module is designed to determine whether a rule is applicable. This module compares the left-hand side of a rule \( u \) with the multiset of objects \( w \) present in the current membrane. If and only if \( u \leq w \), this rule is applicable and this module will generate a signal Applicable = 1. Input all the Applicable signals to an OR gate, the result of this logical gate can used as a monitor to identify whether the evolution reaches halt configuration.

The transition of configurations of P system is realized deterministically and sequentially, which is different from the general model. The consecutive transformation of configurations is regarded as
the evolution process. This evolution process is decomposed into micro-steps and macro-steps. The application of rules enclosed by membranes is performed in terms of a predefined sequential order. This deterministic execution of rules is conducted in micro-steps sequentially. If a selected rule is applicable, the left-hand side of the rule $u$ will be removed. Then the right-hand sides $v, v_1$ and $v_2$ is stored in corresponding registers. The objects from the upper immediate membrane will be preserved in another register. Although the micro-steps are carried out deterministically, they are performed simultaneously in all membranes, until there are no applicable rules. The micro-steps terminates when there are no applicable rules, i.e. the halt condition is reached. All the registers are updated in line with associated rules in macro-steps.

This implementation considered and respected the priorities of applicable rules at the beginning of each micro-step. By labeling the applicable rules with higher priorities and storing the corresponding labels, applicable rules are executed in accordance with their respective priorities. Besides, two additional features of P system, the dissolution and creation of membranes, are simulated. When a rule with membrane dissolving function is applied, its contents are owned by its upper immediate membrane, setting the membrane Enable signal of the relevant membrane to "0". However, the connections and registers defining the dissolved membrane still exist. This scheme gives rise to a disadvantage that the hardware resources cannot be released. The creation of new membranes is executed in the initialization process of the P system since all the information about new membrane is known from the specification of the system. The created membranes are inactive until membrane creating rules invoke them.

6.1.2 Nguyen simulation. In this implementation, a parallel computing platform simulating membrane computing based on FPGA named Reconfig-P is developed [69, 74]. Reconfig-P is fabricated on the basis of the region-oriented idea that regions work as the computational entities communicating objects through message passing. The functionality of these regions is extended by the included set of evolution rules. P Builder, the software component of Reconfig-P, specifies the P system concerned in software, converts the specification of P system written in Java to Handel-C (a hardware description language) source code. Software simulation of the circuits to be constructed is supported by P Builder to test the functionality of circuits before mapping the code to hardware circuits.

The execution of a evolution step is divided into two phases: object assignment phase and object production phase [69]. The maximal instance of each rule in a region is determined in the object assignment phase. The update of multiplicity of objects is accomplished in the object production phase. The maximal instance of the rules with higher priorities is computed before the rules with lower priorities. Note that the consumption of objects for rules with higher priorities performed during the object assignment phase to save clock cycles. It is assumed that all rules are assigned relative priorities. The priority between rules is implemented as the temporal order which should be respected by region processing units in the assignment phase. Rules with same priority are executed concurrently. The temporal order is determined at compile-time. The rules are applied according to their priorities in rounds until no rules are applicable using the indirect iterative approach. Under this circumstance, the applicability of each rule is non-stationary because of the existence of priorities. To avoid processing inapplicable rules, the applicability status of each rule is checked at the outset of the assignment phase and immediate after a applicable is applied to consume some objects.

The objects traversing behavior is the origin of communication between regions. The update of multiplicity of objects caused by rules with and without traversing behavior is completed in the object production phase. When different region processing units update the multiplicity value of the same object at the same time, a conflict occurs. To handle this conflict, in [69] two solution
strategies, the \textit{space-oriented strategy} and the \textit{time-oriented strategy} are proposed. Table 1 and 2 summarize the different strategies of two resource conflict resolutions and their modifications in the rule-based and region-based design [70, 72]. In order to simplify the exposition of processes of rule-based and region-based implementations of P systems, tables are designed to delineate the relevant details, which will be given below.

Table 1. The comparison of the time-oriented and space-oriented conflict resolution

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Resource conflict resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time-oriented</td>
<td>(a) construct a conflict matrix in which each row is a quadruple ((p, q, r, s)). (p) is the object competed by multiple rules. (q) is the region where (p) is produced or consumed. (r) is the set of the conflicting rules, (s) is the size of set (r). (b) insert delay statement among conflict rules such that the updating operations of multiplicity of (p) can be executed in distinct clock cycles. The number of delays is equal to (s).</td>
</tr>
<tr>
<td>Space-oriented</td>
<td>(a) construct the conflict matrix as in the Time-oriented strategy. (b) the register storing the object accessed by multiple rules concurrently is replicated (s) times. These copy registers are assigned to each conflicting rules to write. After the updating process for all the copy registers, the corresponding values are joined to the original register.</td>
</tr>
</tbody>
</table>

Table 2. The differences of the conflict resolutions adopted in two design modes

<table>
<thead>
<tr>
<th>Item</th>
<th>Time-oriented strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule-oriented</td>
<td>The interleaving operation can be determined at compile-time and it can be hard-coded into the HDL source. Need a multiset replication coordinator to coordinate the multiplicities stored in the copy registers.</td>
</tr>
<tr>
<td>Region-oriented</td>
<td>The objects received from other regions are regarded as external objects, otherwise the objects are internal objects. The interleaving merely caused by the production of internal objects can be identified at compile-time. To retain the independence of region processing units, the interleaving induced completely or partially by the receipt of external objects can only be calculated at runtime. The role of the multiset replication coordinator in rule-oriented design is played by the region processing units. The existing register storing the multiplicity received from the associated communication channels in the considered region can be assigned to those processing units which sending objects to the considered region to write the new values of the objects competed by multiple rules.</td>
</tr>
</tbody>
</table>

The extensibility of the region-based design is the consequence of the representation of membranes as processing units interacting with two region processing units corresponding to inner and outer regions. This allows to achieve a strong separation of the processing logic inside different membranes and the independence of the communication. Thus, adding additional elements to the system does not lead to the redesign of the remaining part of the system.

6.2 Rule-based simulations

Rule-based approaches consider evolution rules as processing units performing the update of multiplicities and of the membrane structure.

6.2.1 Nguyen simulation. Every rule in all regions of the P system is represented as a processing unit synchronized by a global clock that implements the parallel processing. However, a processing unit does not correspond a concrete hardware component but rather to a potential infinite \texttt{while loop} which contains the code related to the rule application using Handel-C. The information associated to execution and synchronization is contained in processing units as well. Each rule processing unit in a region is linked to the array of registers containing multisets of objects. The membrane inclusion relationships can be described with the connections between processing units and arrays. Generally speaking, a rule processing unit in a region is linked to the objects array located in the same region with the rule processing unit. If there are objects traversing regions which imply the containment, connecting the rule processing unit to the object array to which
the rule will send objects contained in different region. This procedure permits to represent the membrane inclusion. The rule execution is split into preparation phase and updating phase. We compare the two designs (the rule-based and the region-based) of Nguyen’s implementation in Table 3.

Table 3. The comparison of the rule-based and region-based design of Nguyen’s implementation

<table>
<thead>
<tr>
<th>Object</th>
<th>Rule-based Design</th>
<th>Region-based Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region and their containment relationships</td>
<td>Regions are realized in hardware implicitly by the content they included. For the containment relationships including the traversing of objects between regions, they are implemented by imparting the corresponding rules with ‘in’ or ‘out’ target directives the abilities to access the multiset of objects in the destination region of the objects traversed.</td>
<td>Regions are represented as parallel processing units. The traversing of objects among regions is realized as message passing through channels connecting different region processing units.</td>
</tr>
<tr>
<td>Multiset of objects</td>
<td>An array of registers contains each type of object in the alphabet for a region.</td>
<td>The same strategy adopted as the rule-oriented design.</td>
</tr>
<tr>
<td>Evolution rule</td>
<td>Potentially infinite while loop in which contain procedures representing the operations of the relevant evolution rules in HDL language.</td>
<td>Implicitly expounded through integrating them into the region processing units.</td>
</tr>
<tr>
<td>Operation process</td>
<td>Preparation phase and updating phase.</td>
<td>Object assignment phase and production phase.</td>
</tr>
<tr>
<td>Synchronization</td>
<td>An array of registers composed of three 1-bit registers is associated to each rule processing unit. The values in the array indicate whether the rule processing unit has complete the preparation and updating phase, or is applicable. Compute the logic AND or OR values of all the values stored in the 1-bit register that indicates the same status of the processing unit and store the three results into three sentinel registers. The coordinating processing unit read the sentinel values to synchronize the whole procedures.</td>
<td>For the synchronization of object assignment phase, it is realized when all the region processing units communicate with each other on channels at the beginning of object production phase. For the object production phase, a region execution coordinator connecting to each region processing unit via dedicated channels is designed to perform the synchronization of operations. After the region execution coordinator received the signals denoting the completion of all the operations of every region processing unit, the current transition is done and the next one is carried out.</td>
</tr>
</tbody>
</table>

6.2.2 Verlan and Quiros simulation. The target model is a static P system. The system is considered flattened, so only one skin membrane is present. A special strategy was elaborated in order to not compute the complete solution, i.e. Applicable(Π, C, δ) but the cardinality of its elements. Then a random value between 1 and this cardinality is taken. Finally, this number is decoded to the corresponding solution [88, 106]

Devising an algorithm which carries out the computation of the cardinality and of all elements of the solution set in constant time on FPGA is the key issue of the approach. A remarkable characteristic of FPGA is that the time consumed for executions of functions that do not exceed the cycle of the global clock is done in one cycle of FPGA, hence in constant time. The computation of the cardinality and the decoding of solutions are accomplished by two functions hardwired into the circuit: NBVvariants(Π, C, δ), which gives the cardinality of Applicable(Π, C, δ), and Variant(n, Π, C, δ), which returns the n-th element of Applicable(Π, C, δ).

A concept named rules’ dependency graph is introduced to compute the two functions above. It is a bipartite graph that contains as nodes rules and objects and there is an edge between two nodes if a rule contains the corresponding objects in its left-hand-side. The picture below depicts the rules’ dependency graph for rules r₁ : ab → u and r₂ : bc → v.

```
  a   b ─ c
     |   |
     r₁ r₂
```

Assume that the derivation mode is maximal parallelism (max). Suppose that Nₐ, N₈ and Nᵥ represent the number of objects a, b and c in C. Let N₁ = min(Nₐ, N₈), N₂ = min(N₈, Nᵥ), N = min(N₁, N₂), kᵢ = Nᵢ ⊕ N, 1 ≤ i ≤ 2, where ⊕ denotes the positive subtraction. Let also p, q =
0, 1, 2, …, N. From the dependency graph we can deduce the following:

\[ \text{Applicable}(\Pi, C, \text{max}) = \bigcup_{p+q=N} \{ r_1^{p+k_1} r_2^{q+k_2} \} \]

\[ \text{NBVariants}(\Pi, C, \text{max}) = N + 1 \]

\[ \text{Variant}(n, \Pi, C, \text{max}) = r_1^{N-n+1+k_1} r_2^{-n-1+k_2}. \]

Example 6.1. Consider a configuration where \( N_a = 5, N_b = 5 \) and \( N_c = 3 \). It can be easily verified that \( N_1 = \min(5, 5) = 5, N_2 = \min(5, 3) = 3, N = \min(5, 3) = 3, k_1 = N_1 \cap N = 5 - 3 = 2, k_2 = N_2 \cap N = 3 - 3 = 0 \). Hence, we can enumerate the elements of \( \text{Applicable}(\Pi, C, \text{max}) \) as follows:

\[ \text{Applicable}(\Pi, C, \text{max})_1 = \{ r_1^{3+2} r_2^{0+0}, r_1^{2+2} r_2^{1+0}, r_1^{1+2} r_2^{2+0}, r_1^{0+2} r_2^{3+0} \} = \{ r_1^5, r_1^3 r_2, r_1^2 r_2, r_1^2 r_2^2 \}. \]

The same result can be easily obtained by using formal power series associated to context-free languages. In this case, any maximal rule combination is a part of the language \( L_N = \{ r_1^p r_2^q | p + q = N \} \). It is quite easy to observe that the number of words of length \( N \) in \( L_N \) is exactly the same as the number of words of same length in the language \( L = \{ r_1^* r_2^* \} \). This last language is regular and its generating function is \( q_0(x) = 1/(1 - x)^2 \). The \( n \)-th coefficient of the expansion of \( q_0(x) \) is equal to \( n + 1 ([x^n]q_0 = n + 1) \), which immediately gives \( \text{NBVariants}(\Pi, C, \text{max}) = N + 1 \). The \( \text{Variant}(n, \Pi, C, \text{max}) \) is computed using an algorithm that performs a weighted breadth-first search of the decomposition of \( n \) with respect to the number of variants found on each branch of the execution of automaton for \( L \).

Such a process can be easily repeated for any regular language, yielding a constant time simulation of a computational step. The reason for such performance is that any generating function is equivalent to a recurrence relation and such relations can be computed in one synchronous time unit using asynchronous operations. The described algorithm functions for any \( \Pi \) system where the rule choice can be expressed as words of certain length in a regular language. The corresponding class is quite large (containing even computationally complete models), thus allowing an extremely fast execution. Examples from [106] exhibited a speed-up of order \( 10^5 \).

Another important point is that this approach allows to handle the non-determinism in a natural way, by performing a uniform random choice between all possible rule applications at each step.

Technically, the implementation represents only objects by registers and rules by layered logic. Each rule implementation is modularized and contains an own copy of processing instructions needed to compute the two above functions, based on asynchronous operations. Consequently, five clock cycles are required to compute the \( \text{NBVariants}(\Pi, C, \text{max}), \text{Variant}(n, \Pi, C, \text{max}) \) and to apply the corresponding rules. The entire process of the implementation is split into several consecutive stages, which take charge of different operations associated to phases of evolutions of configurations.

Persistence stage stores the states that the hardware system goes through. A independent stage computes the maximal instance of each rule by means of the dividing operation and MIN logic operation. Assignment stage in charge of selecting a rule to be applied non-deterministically, and determines its instances. Updating stage is responsible for updating the current configuration with the values from the previous stage. During Halting stage, the system inspects whether the halting condition is reached, and once reached, stops the system.

According to the elaborations stated above, the comparisons of the FPGA implementations of \( \Pi \) systems is listed in table 4. The details of each design are summarized. We can check the characteristics of the three main prototypes.
Table 4. The quantitative attributes of FPGA implementation

<table>
<thead>
<tr>
<th>Item</th>
<th>Biljana Petreska</th>
<th>Van Nguyen</th>
<th>Juan Quiros</th>
</tr>
</thead>
<tbody>
<tr>
<td>Period</td>
<td>2003 Swiss Federal Institute of Technology</td>
<td>2007-2010 University of South Australia</td>
<td>2012-2015 University of Seville</td>
</tr>
<tr>
<td>Institute</td>
<td>Lausanne (EPFL)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Target FPGA</td>
<td>Xilinx Virtex-II Pro 2VP50ff1517-7</td>
<td>Xilinx Virtex-II XC2V6000-f1152-4 (rule-oriented) and Virtex-II RC2000 (region-oriented)</td>
<td>Xilinx Virtex-V XC5VFX70T and Virtex-VII XC7VX485T</td>
</tr>
<tr>
<td>Host processing platform</td>
<td>Not given.</td>
<td>1.73GHz Intel Pentium M processor with 2GB of memory</td>
<td>Intel Core i5 – 5220 at 3 GHz, with 8 GB of RAM.</td>
</tr>
<tr>
<td>HDL</td>
<td>VHDL</td>
<td>VHDL</td>
<td>VHDL</td>
</tr>
<tr>
<td>Experiment Subjects</td>
<td>Cell-like P systems with following characteristics: membrane dissolution and creation, objects exchange between upper- and lower-immediate membranes, cooperative P systems with priorities. The range (non-continuous) for object number is 6 to 12 and for membrane is 10 to 20.</td>
<td>For rule-oriented design, the subjects are cell-like P systems cascaded in vertical, horizontal, vertical and horizontal structures. For region-oriented design, the objects are cell-like P systems containing hierarchical regions and tissue-like connected regions. The rule range is [10,50], [1,25] for regions, [3,200] for objects. The extent of the inter-region communication is [80,319].</td>
<td>The subject P system is simplified according to the multiset rewriting point of view, which it has a skin membrane, no inner regions. The 4 subjects differ in the rule dependencies which form chains: circular, 2-circular, linear, opposite. The object number range is [10,200].</td>
</tr>
<tr>
<td>Experiment results</td>
<td>The hardware consumption ranges from 4.2% to 33% (CLB). The extent of clock rate is 27 to 198 MHz.</td>
<td>For rule-oriented design, the number of rules applied per second ranges from 2.7×10^5 to 1000×10^5. The hardware consumption extent is 1.55% to 21.43% (LUTs). For region-oriented design, the hardware usage ranges 1.82% from 16.79%. The clock rate fluctuates from 52.63 MHz to 81.77 MHz. The biggest size which can be executed is a P system with 550 rules, 1280 communication channels, 1100 objects conflicts.</td>
<td>The hardware consumption ranges from nearly 1% to 48% (LUTs) or nearly 1.1% to 11% (slice). The period needed to perform a computation step fluctuates from 5.46 ns to 9.14 ns. The highest frequency exceeds 100 MHz, permitting 2×10^7 computational steps per second. The run-time for each experiment subject ranges from 3.017×10^-5 s to 4.174×10^-5 s.</td>
</tr>
<tr>
<td>Parallelism</td>
<td>System-level parallelism</td>
<td>Region-level and system-level parallelism</td>
<td>System-level (there is only a skin membrane, so it is also region-level parallelism).</td>
</tr>
<tr>
<td>Non-determinism</td>
<td>No non-determinism</td>
<td>DND algorithm</td>
<td>True non-determinism.</td>
</tr>
</tbody>
</table>

7 AN OVERVIEW OF EXISTING CUDA SIMULATIONS

In [61], it is concluded that the GPU is a suitable platform to accelerate the simulation of P systems because of the following features:

- **Good performance**: for example, the NVIDIA Tesla K40 delivers 1.43 TeraFLOPS double-precision peak floating point performance, 4.29 TeraFLOPS of single-precision, and 288 GBytes/s of global memory bandwidth;
- **An efficiently synchronized platform**: GPUs implement a shared memory system, avoiding communication overload;
- **A medium scalability degree**: the amount of resources depend on the GPU model, e.g. a K40 includes 2880 cores and 12 GBytes of memory. If the resources of a GPU is not enough, there are more scalable solutions such as multi-GPU systems, but they then require communication among nodes;
- **Low-medium flexibility**: although CUDA programming is based on C++, and hence programmers are free to use the same data structures than in CPU, both the algorithm and the data structure have to be adapted for best performance on GPUs.

Simulation algorithms implemented on CUDA have a common structure, in which for each simulated computation step, first the rules are selected (obtaining a multiset of rules, while consuming the left-hand sides), and secondly the rules are executed (based on the obtained multiset of rules,
## Table 5. The qualitative attributes of FPGA implementations

<table>
<thead>
<tr>
<th>Item</th>
<th>Biljana Petreska</th>
<th>Van Nguyen</th>
<th>Juan Quiros</th>
</tr>
</thead>
<tbody>
<tr>
<td>Membranes (regions) and their</td>
<td>Implicitly represented by the contents enclosed by the membranes. The objects exchange is interpreted as transferring objects with communication buses by which connect the origin region to the destination region. For the rule-oriented design, the treatment is similar with Petreska’s. For the region-oriented design, regions are represented as parallel processing units. The objects exchange among regions is realized as message passing through channels connecting different region processing units. According to the multiset rewriting system framework, the topology structures of membranes are not important. What concerned is the rule dependency graph.</td>
<td>For the rule-oriented design, the treatment is similar with Petreska’s. For the region-oriented design, regions are represented as parallel processing units. The objects exchange among regions is realized as message passing through channels connecting different region processing units. According to the multiset rewriting system framework, the topology structures of membranes are not important. What concerned is the rule dependency graph.</td>
<td></td>
</tr>
<tr>
<td>containment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiset of objects</td>
<td>Store the multiplicity value of each type of object in different registers whose positions indicate the type. The same method. The same method.</td>
<td>The same method.</td>
<td></td>
</tr>
<tr>
<td>Evolution rule</td>
<td>Store the left-hand-side and right-hand-side of a rule in different registers. In rule-oriented design, they are characterized as potentially infinite while loop in which contain procedures representing the operations of the relevant evolution rules in HDL. In region-oriented design, they are implicitly expounded through integrating them into the region processing units.</td>
<td>In rule-oriented design, perform preparation phase and updating phase. In region-oriented design, perform object assignment phase and production phase. Because of the region-oriented design, Membrane mediated features such as symport and antiport functions can be extended.</td>
<td>1. Persistence stage 2. Independent stage 3. Assignment stage 4. Application stage 5. Updating stage 6. Halting stage. This implementation is designed only for P systems whose applicable multisets of rules can be represented as regular language. So its extensibility is limited.</td>
</tr>
<tr>
<td>Operation Process</td>
<td>Micro-step: only one applicable rule is applied with respect to the instance number in a region. Macro-step: execute a micro-step concurrently in every region. In rule-oriented design, perform preparation phase and updating phase. In region-oriented design, perform object assignment phase and production phase.</td>
<td>In rule-oriented design, perform preparation phase and updating phase. In region-oriented design, perform object assignment phase and production phase. Because of the region-oriented design, Membrane mediated features such as symport and antiport functions can be extended.</td>
<td>1. Persistence stage 2. Independent stage 3. Assignment stage 4. Application stage 5. Updating stage 6. Halting stage. This implementation is designed only for P systems whose applicable multisets of rules can be represented as regular language. So its extensibility is limited.</td>
</tr>
<tr>
<td>Extensibility</td>
<td>Membrane mediated features cannot be added in since membranes are implicitly represented by their contents.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scalability</td>
<td>With the increase of the rules, the hardware usage rise approximately proportional. The clock rates decline a small amount. The membrane creation ability cripples clock rate significantly. The hardware consumption scales linearly with respect to the size of the P system executed in both rule-oriented and region-oriented designs. The performances grow linearly when the number of rules increase.</td>
<td>The hardware consumption scales linearly with respect to the size of the P system executed in both rule-oriented and region-oriented designs. The performances grow linearly when the number of rules increase.</td>
<td>The hardware usage is scalable, but the rate of increase in the performance is not always linear for different systems with distinct structures.</td>
</tr>
<tr>
<td>Contributions</td>
<td>Membrane creation and dissolution functionality</td>
<td>Two kinds of methodology for P system characterization and resource conflict resolution, DND algorithm</td>
<td>Put up with a new methodology with absolute equiprobability to implement non-determinism.</td>
</tr>
<tr>
<td>Drawback</td>
<td>Partially parallelism, no non-determinism</td>
<td>The equiprobability of DND algorithm has not been proven theoretically.</td>
<td>The types of P systems which can be implemented are confined to those whose rules can be represented as regular language.</td>
</tr>
</tbody>
</table>

and generating the right-hand side. This strategy is necessary in order to synchronize which rules get executed in the corresponding computation step. For some models, for example for those ad-hoc simulators, the selection step is done in micro-stages.

In the following subsections, the existing CUDA simulations are summarized, by organizing into P system models.

### 7.1 Cell-like P systems

The first test of concept for simulating P systems on GPUs was applied to P systems with active membranes using CUDA [18]. This simulator performs only one computation out of the whole tree, in order to avoid non-determinism, by requiring the confluence property to the simulated P systems. Bearing this in mind, the "lowest-cost computation path" is selected: the one in which least membranes and communication are required. This is achieved by giving preferences to rules that lead to least membranes (e.g. dissolution over division rules). The simulation algorithm is composed of two main stages: selection and execution of rules. Selection is where the semantics of the model is...
actually simulated. Rules are chosen by following the defined constrains, altogether with a number of applications. The result of this stage is used for the next one, which is the execution of the rules; that is, updating the P system configuration. This two-staged strategy allows to synchronize the application of rules within and among membranes.

Both P systems and GPUs have a double-parallel nature \cite{18}, and this is harnessed for implementing a mapping: (elementary) membranes are assigned to thread blocks, and a subset of rules to threads. Each thread is in charge of selecting rules for a portion of the defined objects in the alphabet. Note that this is enough given that in P systems with active membranes, rules have no cooperation. However, this mapping of parallelism is naive, since it assumes that all the objects in the alphabet can be present within each membrane. This require allocating memory space and assigning resources (threads) to all of them. This, in fact, does not take place in the majority of P systems to be simulated, but turns out to be the smallest worst case to handle. Thus, the performance of the simulator completely depends on the P system being simulated, and drops as long as the variety of different objects appearing in membranes decreases.

Non-determinism is handled by imposing the simulated P systems to be confluent; that is, all computations halt and they generate the same result \cite{18}. This way, the simulator can choose any path in the computation tree, since the aim is to find a halting configuration. The GPU simulator take advantage of this by preferring to select evolution rules rather than division or dissolution.

The performance of the simulator was analyzed on a GPU Tesla C1060 (240 cores, 4GB memory) by using two benchmarks \cite{61}: a simple test P system designed in a convenient manner (up to 7x of speedup), and a family of P systems designed to solve different instances of the SAT problem (1.67x of speedup).

Improvements to this design have followed \cite{61}, reporting up to 38x of acceleration when taking advantage of shared memory and data transfer minimization. By constructing a dependency graph, the set of rules of the input model are arranged so that those having common objects in the left-hand side and in the right-hand side (respectively) are more likely to be in a node. This reduces communication given that thread blocks are assigned to nodes.

Another approach was to implement ad-hoc simulators for a specific family of P system with active membranes allowing to solve SAT in linear time \cite{19}. In this way, the worst case assumed before (all objects defined can appear in every membrane) is further reduced by analyzing the upper bound to the number of existing objects in membranes. In this way, the work done by threads is maximized, and the design remains similar: each elementary membrane is implemented by a thread block, and each object of the input multiset is implemented by a thread. The experiments carried out on a NVIDIA Tesla C1060 GPU reported up to 63x of speedup. Further developments took place focusing on this family of simulators, with the aim at being better tailored to newer GPU architectures, and also to enable multi-GPU systems and supercomputers \cite{19}.

A related work was to explore which P system ingredients are better suited to be handled by GPUs. To this aim, another solution to SAT based on a family of tissue P systems having the operation of cell division was simulated \cite{61}. The design is similar to the previous case: each cell is simulated by a thread block, however the constructed solution required a higher number of objects to be placed inside each cell. At the same time, this simulator does not need to store nor handle charges associated to membranes. Experiments on a NVIDIA Tesla C1060 GPU leaded to speedup by 10x. This showed that using charges associated to membranes helped to save instantiation of objects, and so, they entail a lightweight ingredient to be processed by threads.

### 7.2 Population Dynamics P systems

Population Dynamics P (PDP) systems is a multi-environmental model successfully used for modeling real ecosystems. Thus, their efficient simulation was critical for experimental validation.
Simulators for PDP systems typically run several simulations in order to extract statistical information from the models.

A CUDA simulator for PDP systems was presented in [62]. The selected simulation algorithm was the DCBA [63], since it provides better accuracy in the simulation results. The selection of rules in DCBA consists of three phases: phase 1 (distribution of objects), phase 2 (maximality) and phase 3 (probability). This algorithm uses a distribution table in order to distribute the objects in a proportional way between competing rules, i.e. with overlapping left-hand sides. Moreover, rules are grouped into rule blocks, when having the same left-hand side.

It can be noticed that the approach of DCBA is to handle non-determinism by first using an uniform distribution of objects to competing rules. This requires an extra phase (2) to avoid rounding errors, in this way a random order is employed to control the remaining rules and which one takes all objects. This phase is based on the procedure of DND.

The CUDA simulator for PDP systems [61, 62, 64] uses the following design: the contents of environments and rule simulations are fractioned and distributed all-over thread blocks, while individual rules are fractioned over threads. Phases 1, 3 and 4 are efficiently managed by the GPU, but Phase 2 can become a bottleneck. For phase 3, a random binomial variate generation library was developed, so that binomial and multinomial distributions were supported for random number generation. In the benchmark using a set of randomly generated PDP systems (having no biological meaning), speedups of up to 7x were achieved on a Tesla C1060 with respect to a multi-core version. The simulator was validated and tested by using a known ecosystem model of the Bearded Vulture in the Catalan Pyrenees, leading to speedups of up to 4.9x with a C1060, and 18.1x using a Tesla K40 GPU (2880 cores). Finally, the simulator was extended in [64] in such a way that it introduces high-level information provided by the model designer into the code. This is accomplished through a new syntactical ingredient in P-Lingua 5 [80], called feature. This ingredient was employed to define the usual algorithmic scheme based on rule modules when designing biological models. This new version is called adaptative simulator, and helped to obtain an extra speedup of 2.5x on a Tesla K40, and 1.8x on a Tesla P100 GPU (3584 cores).

Table 6. The summing up of CUDA-GPU implementations I. The target CUDA-GPU is NVIDIA C1060 (240 cores) and the host platform is a computer with two Intel i5 Nehalem processors (8 cores).

<table>
<thead>
<tr>
<th>Item</th>
<th>PCUDA</th>
<th>PCUDASAT</th>
<th>TPCUDASAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>P system class</td>
<td>Cell-like (recognizer P system with active membranes)</td>
<td>Cell-like P system with active membranes (resolve SAT in linear time)</td>
<td>Tissue-like P system (resolve SAT in linear time)</td>
</tr>
<tr>
<td></td>
<td>Elementary membranes are represented by thread blocks. Threads are assigned to objects.</td>
<td>Each elementary is assigned a thread block. Only the objects appearing in the input multisets will be assigned a thread.</td>
<td>Every Cell is assigned a thread block. More objects are requested by every cell than the PCUDASAT simulator.</td>
</tr>
<tr>
<td>Experiment models</td>
<td>512 objects and 1024 membranes for (a) and 914 objects and 4096 membranes for (b).</td>
<td>256 objects and 4096 membranes.</td>
<td>256 objects and 4096 membranes.</td>
</tr>
<tr>
<td>Size of the simulated P system</td>
<td>7 times for (a) and 1.67 times for (b).</td>
<td>The CUDA simulator is 63 times faster than the sequential simulator and the hybrid CUDA simulator is 9.3 times faster than the normal one.</td>
<td>10 times higher than the sequential simulator.</td>
</tr>
</tbody>
</table>

ACM Comput. Surv., Vol. 1, No. 1, Article . Publication date: June 2020.
7.3 Spiking Neural P systems

Parallel simulation of Spiking Neural P (SNP) systems has been based on a matrix representation so far [115]. The simulation algorithm uses the following vectors and matrices:

- **Spiking transition matrix**: stores information about rules, and is employed for computing transitions. It assigns a row per rule and a column per neuron. The values coded correspond to the left-hand side of rules (as negative numbers) and to the right-hand sides (as positive numbers).
- **Spiking vector**: defines a selection of rules to be fired in a transition step, using a position per rule. Given the non-deterministic nature of SNP systems, there are more than one valid spiking vector for a given configuration.
- **Configuration vector**: defines the number of spikes per neuron, that is, the configuration in a given time.

The algorithm is then used to compute the next configuration from a given one, by performing only vector-matrix operations. These operations are offloaded to the GPU, which is optimized to handle matrices [15]. CuSNP is a simulator for SNP systems which is written in Python and the CUDA kernels were launched by using the binding library PyCUDA. For the first approach, SNP systems without delays were simulated by covering each computation path sequentially, leading to speedups of up to 2.31x [61] [15].

Further extensions have followed, enabling delays, support of more types of regular expressions and input of P-Lingua files [17]. The support of delays is done by an extension of the matrix representation, introducing vectors to control when rules fire after the delay is met, and if neurons are open. These leaded to up 50x of acceleration on a GTX750 GPU [17]. Moreover, non-deterministic SNP systems with delays are supported by generating all computation paths sequentially and using GPU to speedup the simulation [16] (with up to 2x on a GTX1070 for a non-uniform solution to Subset Sum).

Furthermore, the simulation of Fuzzy Reasoning Spiking Neural (FRSN) P systems on the GPU was explored [55]. FRSNP systems allows modeling fuzzy diagnosis knowledge and reasoning for fault diagnosis applications. The simulation algorithm is also based on a matrix representation and vector-matrix operations. The employed simulation framework was pLinguaCore, so the CUDA kernels were launched by using the binding library JCUDA.

7.4 Other models

Enzymatic Numerical P systems has been employed for modeling robot controllers and path planning, being significant for the Artificial Intelligence. A CUDA simulator was developed [61], where the selection of applicable programs was first applied, second, the calculation of production functions, and third, distribution of production function results according to repartition protocols. Production functions are computed using a recursive solution. Simulators were implemented in Java (inside pLinguaCore) and C programming languages as standalone tools. On a GeForce GTX 460M, the achieved speedup was of up to 11x. Moreover, a model for RRT and RRT* algorithms for robotic motion planning using ENPS was also simulated on CUDA [79] 24x faster than in multicore CPU using an RTX2080.

Evolution-Communication P systems with Energy (ECPE) [61] were also target for a CUDA simulation. The employed simulation algorithm used a matrix representation and linear-algebra based algorithm, similarly as for the spiking neural P systems simulator: a configuration vector, a trigger matrix, an application vector and a transition vector.

The summary of CUDA-GPU implementations is concluded in Table 6 and 7, providing an easy checking for the CUDA-based simulations for academic communities. For the sake of presenting
Table 7. The summing up of CUDA-GPU implementations II

<table>
<thead>
<tr>
<th>Item</th>
<th>ABCD-GPU</th>
<th>ENPS-GPU</th>
<th>snpgpu and CuSNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>P system class</td>
<td>Population Dynamics P system</td>
<td>Enzymatic numerical P system (cell-like)</td>
<td>Spiking Neural P system</td>
</tr>
<tr>
<td>P system ingredients</td>
<td>Environments and simulations are distributed through thread blocks.</td>
<td>Each production function and each repartition protocol element are associated with a thread respectively.</td>
<td>Spiking transition matrices, spiking vectors and configuration vectors are employed to model the target systems.</td>
</tr>
<tr>
<td>Operation Process</td>
<td>(DCBA algorithm) Selection stage (1. Distribution 2. Maximalcy 3. Probability) and Execution stage.</td>
<td>1. Select the applicable programs 2. Compute production functions 3. Distribute the results of the precede step according to repartition protocols.</td>
<td>Generate the matrix representation for the model, and iterate the computation of next configuration until a stopping criterion is met.</td>
</tr>
<tr>
<td>Experiment models</td>
<td>A set of randomly generated PDP systems (without biological meaning), the ecosystem model of the Bearded Vulture and a tritrophic example model with modules defined as features.</td>
<td>A dummy model and an exponential function approximation model.</td>
<td>(a) SNP systems covering each computation path sequentially, without delays. (b) Extensions enabling delays, support of more types of regular expressions and input of P-Lingua files.</td>
</tr>
<tr>
<td>Size of the simulated P system</td>
<td>Running 50 simulations, 20 environments and more than 20000 rule blocks.</td>
<td>2 membranes, range from 1 to 120000 programs.</td>
<td>(a) From 1 to 16 neurons. (b) Generalized sorting networks with up to 512 neurons.</td>
</tr>
<tr>
<td>Speedup</td>
<td>7 times (Tesla C1060) compared with the sequential model, 3 times contrasted to the 4 cores CPU. The speedup for the real ecosystem, 4.9x with a C1060, and 18.1x using a Tesla K40 GPU (2880 cores). Using the adaptive simulator with the tritrophic model using modules achieved an extra 2.5x on a K40 GPU and 1.8x on a P100 GPU (3584 cores).</td>
<td>6.5 times for dummy model and 10 times for exponential model. 11 times on GeForce GTX 460M.</td>
<td>(a) 2.31 times for the Python version simulator. (b) 50 times on CUDA-based version (GTX750 GPU).</td>
</tr>
</tbody>
</table>

A overview contrast of FPGA and CUDA-GPU implementation, we summarize the quantitative and qualitative attributes and conclude their methodologies in the Table 8, 9 and 10. This overall contrast present a recapitalized conclusion about hardware implementations of P system.

Table 8. The comparison of the quantitative attributes of the implementations between FPGA and CUDA-GPU

<table>
<thead>
<tr>
<th>Item</th>
<th>FPGA</th>
<th>CUDA-GPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inchoate year</td>
<td>2003</td>
<td>2009</td>
</tr>
<tr>
<td>Hardware Price</td>
<td>50-4500 $</td>
<td>150-2000 $</td>
</tr>
<tr>
<td>Concerned institute</td>
<td>Southwest Jiaotong University, Université Paris Est Créteil Val de Marne, Xihua University</td>
<td>Universidad de Sevilla, University of the Philippines Diliman</td>
</tr>
<tr>
<td>Number of researchers</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>Implemented P system type</td>
<td>Cell-like P system</td>
<td>P systems with active membranes, tissue P systems with cell division for a SAT solution, Population dynamics P system, spiking Neural P system, enzymatic Numerical P system, evolution communication P system.</td>
</tr>
<tr>
<td>Size of the simulated P systems</td>
<td>No limit for objects and membranes, but less than 1000 rules</td>
<td>No limit as long as GPU resources (mainly memory) are available.</td>
</tr>
<tr>
<td>Approximate time needed to obtain a hardware system</td>
<td>2 months</td>
<td>1 month.</td>
</tr>
<tr>
<td>Speed-up effect</td>
<td>2×10^7 computational steps per second at present.</td>
<td>Depends on model, implementation or GPU device. Reported speedups from 2 to 90x compared to a sequential counterpart.</td>
</tr>
<tr>
<td>Handling of non-determinism</td>
<td>DND algorithm and formal power serial theory.</td>
<td>DCBA, or ad-hoc strategies (e.g. assuming confluent P systems).</td>
</tr>
</tbody>
</table>
Table 9. The comparison of qualitative attributes of the implementations between FPGA and CUDA-GPU

<table>
<thead>
<tr>
<th>Item</th>
<th>FPGA</th>
<th>CUDA-GPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complexity of the parallel framework</td>
<td>Both region-level and system-level parallelism are realized, as well as the non-determinism (the equiprobability of the rules are guaranteed in Juan’s contribution)</td>
<td>Both region-level and system-level parallelism are realized, several strategies are carried out to sort out non-determinism (for PDP systems, based on DND algorithm in which the equiprobability is not proven mathematically).</td>
</tr>
<tr>
<td>Extensibility</td>
<td>Although the hardware consumptions are no more than 30% when rules are no more than 300, which means there are amount of space to accommodate new features of the P system, extending the existing type of P system to additional types needs significant revisions.</td>
<td>The programming framework is flexible enough to support other ingredients, but it might require a significant revision to the parallel design.</td>
</tr>
<tr>
<td>Scalability</td>
<td>With the size of the subject P systems growing, the computation clock rates decrease not too much and the hardware usages are approximately proportional to the size.</td>
<td>Newer GPUs showed to reach higher performance, but the simulations have been demonstrated to be memory bandwidth bounded. It requires to run large models to see a positive speedup.</td>
</tr>
<tr>
<td>Experiment subject</td>
<td>Cell-like P systems with 10-1000 rules, 1 to 55 regions, 3-165 objects.</td>
<td>(a) P systems with active membranes with 914 objects/rules and 4096 membranes, (b) SAT solutions in active membranes and cell division 256 objects/rules and 4096 membranes, (c) PDP systems up to 200000 rules on 50 simulations and 20 environments, (d) SN P systems with 16 to 512 neurons, (e) ENP systems with 2 membranes and up to 120000 programs.</td>
</tr>
<tr>
<td>Difficulties</td>
<td>Programming with HDL is difficult contrasting with its software counterpart, and the synchronization of intricate circuits. It usually cost 8-10 months to be proficient with HDL programming and to comprehend the parallel architecture.</td>
<td>Work assignment for a parallel design can be the most difficult task when developing GPU simulators, since the restrictions of GPU computing meets with P system semantics. Need to think on GPU architecture for high performance: memory accesses, memory transfers, thread synchronization, etc.</td>
</tr>
<tr>
<td>Application</td>
<td>The prototypes are developed but no applications for concrete problems to date.</td>
<td>several application for biological simulations have to executed. Solutions to computationally-hard problems.</td>
</tr>
</tbody>
</table>

Table 10. The conclusion of the implementations for FPGA and CUDA-GPU

<table>
<thead>
<tr>
<th>Item</th>
<th>FPGA</th>
<th>CUDA-GPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advantages</td>
<td>The computation speed is fast and the hardware framework is adjustable. The compromises between implementing the P system model and constructing the most proximate hardware are relatively ideal.</td>
<td>Relatively convenient way to implement P systems for the established parallel framework in the GPU. The programming is similar with the classic software developing process.</td>
</tr>
<tr>
<td>Disadvantages</td>
<td>Building the parallel architecture from scratch is a laborious and challenging adventure</td>
<td>Slower than FPGA. Sometimes big concessions are indispensable as the architectures are unchangeable.</td>
</tr>
<tr>
<td>Contributions</td>
<td>Introducing the reconfigurable hardware to develop the real parallel architectures which can exploit the maximal parallelism of the P systems substantially.</td>
<td>The tremendous potential of the emergent universal computing GPU is concentrated, which the ready-made parallel framework is off-the-shelf, to establish a P system computing platform conveniently.</td>
</tr>
<tr>
<td>Conclusions</td>
<td>We can develop sophisticated P systems hardware circuits which implement the target models as approximate as possible on FPGA devices. The expense we should pay is also considerable high taking the painstaking effort into account.</td>
<td>CUDA-GPU provides a relatively comfortable and alternative choice to implement P systems, which the simulation results are acceptable, nevertheless at a price of the modification of the source models. GPUs worth when simulating large P system models.</td>
</tr>
</tbody>
</table>

8 OTHER APPROACHES

Besides FPGA-based and CUDA-enabled GPU-based hardware platforms developed for the implementation of P systems, there are several attempts to implement P systems on micro-controllers. Serial algorithms and their hardware circuits designs in terms of the exhaustive investigation line focusing on implementing the transition of configurations of P systems were developed. The
corresponding research does not precise concrete hardware devices, it is just designing the circuits aiming at simulating certain operations of particular P systems with registers, logical gates, magnitude comparators, and data buses.

In [32], a digital circuit is presented to select active rules in the current configuration. Each evolution rule is represented by 2 hardware registers. The first register characterizes the left-hand side (antecedent) of a rule, and the other specifies the right-hand side of the rule, which also determines whether the rewritten objects go out from the current membrane or stay where they are, or go into inner membranes. By comparing the left-hand side of each rule with the multiset of objects in a region, the applicability of every rule can be determined. Next, an algorithm computes the number of applications of active rules given the multiset of objects and evolution rules [58]. The corresponding circuit is composed of logical gates, registers, multiplexers and sequential elements. The computation process is bounded. In [47], a P system circuit is constructed by means of PIC16F88 micro-processor enriched with the storage component 24LC1025, connected by an I2C bus. Thus, the shortcoming of insufficient storage capacity of the micro-processor is solved by the introduced external memory. The flexibility of the circuit is acceptable as the modification of the structure is not necessary.

Continuing the research from [58], the paper [59] presents an improved algorithm and the corresponding circuit calculating the application times of active rules. The computing process can be completed in minor steps and the theoretical performance is optimized. In [6, 7], a draft of a circuit implementing the inherent parallelism of P system is presented. Towards the parallelism, the rule treatment is similar with the region-based solution defined in Section 6, namely, what applied is a multiset of rules instead of a single rule. An operating environment is elaborated in [42] which performs the automatic transformation of tasks involved in the hardware simulation of P systems, including loading, execution and interpretation, into a distributed framework constructed on micro-controllers. The execution results of the circuit, which are in the form of binary data, can be interpreted in a transparent manner. What should be emphasized here is that all the hardware circuits introduced are just on the blueprints which were never implemented in practice. They remain mostly theoretical and the actual functionality and performances are unknown, unlike for FPGA-based and CUDA-based hardware implementations/simulations which are carried out practically. In [45], micro-controllers are also chosen as target hardware to implement communication architectures of P systems. A digital circuit carrying out massive parallelism in transition P Systems is established in [6].

As a new attempt for implementing neural P systems on different hardware, DRAM-based CMOS circuits are adopted to construct elementary spiking neural P systems. We do not carry out an in-depth discussion about this topic and refer to [113] for more details.

9 CONCLUSIONS

In this article we gave an overview of different existing hardware implementations of P systems. As expected the FPGA-based approach is very promising, yielding truly parallel implementations achieving a speed-up of order $10^5$. However, it requires a considerable amount of work as the corresponding hardware architecture should be created from scratch. Moreover, the hardware description languages like VHDL are very low-level and programming in such languages tends to be extremely tedious. Also, as it was shown from the existing implementations, the most promising results are obtained using rule-based architectures, because of the smaller communication cost.

By contrast, CUDA-based approach features a powerful unique hardware platform that can be programmed using standard C/C++ language. While it gives less freedom and accuracy, this approach gives a good trade-off between complexity, scalability and maintenance. The achieved
speed-ups are relatively small (between 4 and 50 times), however, due to the programming simplicity, more types of P systems were implemented.

The central problem in both approaches – the object distribution problem – was tackled from different points of view; the most fruitful attempts being the variations of the DND algorithm and direct approaches using mathematical properties of the dependency relation between rules. Future research can be done in this direction by providing new classes of P systems suitable for the precomputation of possible rule applications. Another research direction related to software implementation of P systems is given by the construction from Section 4, which features a novel reduction of the object distribution problem to ILP. This reduction handles well the maximal parallelism and allows to define criteria for the choice of the solution. Then, it would be possible to design a software framework that would use existing solvers and algorithms to quickly obtain the desired solution, including those optimized for parallel hardware [10, 23, 96, 97, 102, 120].

Finally, as a future research interest we mention a robust implementation of different features of P systems like membrane creation/dissolution, as well as non-classical variants of P systems like numerical P systems. The motivation for such research is that the corresponding models feature numerous applications, so their scalable implementation would immediately allow to test the acceleration of corresponding algorithms and their practical usage.

ACKNOWLEDGMENTS

The work of G.Z., Z. S. and S. V. is supported by the National Natural Science Foundation of China (61972324, 61672437, 61702428), Beijing Advanced Innovation Center for Intelligent Robots and Systems (2019IRS14), Artificial Intelligence Key Laboratory of Sichuan Province (2019RY06), the New Generation Artificial Intelligence Science and Technology Major Project of Sichuan Province (2018GZDZX0043) and the Sichuan Science and Technology Program (2018GZ0086). M.A.M-d-A., LV-C. and M.J.P-J. also acknowledge the support of the research project TIN2017-89842-P (MABICAP), co-financed by Ministerio de Economía, Industria y Competitividad (MINECO) of Spain, through the Agencia Estatal de Investigación (AEI), and by Fondo Europeo de Desarrollo Regional (FEDER) of the European Union.

The authors are grateful to the Editor-in-Chief, Prof. Sartaj Sahni, the anonymous handling editor and all reviewers for their insightful and detailed comments on this manuscript, and are also indebted to Academician Gheorghe Păun for his useful discussions and valuable suggestions.

REFERENCES

An Overview of Hardware Implementation of Membrane Computing Models

Luis Fernández, Fernando Arroyo, Jorge A Tejedo, and Juan Castellanos. 2006. Massively Parallel Algorithm for


Matthew S. Dodd, Dominic Papineau, Tor Grenne, John F. Slack, Martin Rittner, Franco Pirajno, Jonathan O’Neil, and

Daniel Díaz-Pernil, Miguel A. Gutiérrez-Naranjo, and Hong Peng. 2019. Membrane computing and image processing:


Gabriel Ciobanu and Andreas Resios. 2009. Complexity of evolution in maximum cooperative P systems. Natural

Jambhlekar P. Arun, Manoj Mishra, and Sheshasayee V. Subramaniam. 2011. Parallel implementation of MOPSO on

Angel V. Baranda, Fernando Arroyo, Juan Castellanos, and Rafael Gonzalo. 2001. Towards an Electronic Implementation

Angel V. Baranda, Fernando Arroyo, Juan Castellanos, and Rafael Gonzalo. 2001. Towards an Electronic Implementation


An Overview of Hardware Implementation of Membrane Computing Models

35

Weihang Zhu, Ashraf Yaseen, and Yaohang Li. 2011. DEMCMC-GPU: An Efficient Multi-Objective Optimization

Zihan Xu, Matteo Cavaliere, Pei An, Sarma Vrudhula, and Yu Cao. 2014. The stochastic loss of spikes in spiking neural


Xueyuan Wang, Gexiang Zhang, Junbo Zhao, Haina Rong, Florentin Ipate, and Raluca Lefticaru. 2015. A modified


