CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	A GPU implementation of parallel constraint-based local search
Authors	Arbelaez, Alejandro;Codognet, Philippe
Publication date	2014-04-14
Original Citation	Arbelaez, A. and Codognet, P. (2014) 'A GPU implementation of parallel constraint-based local search', 2014 22nd Euromicro International Conference on Parallel, Distributed, and Network-Based Processing, Turin, Italy, 12-14 February, pp. 648-655, doi: https://doi.org/10.1109/PDP.2014.28
Type of publication	Conference item
Link to publisher's version	https://doi.org/10.1109/PDP.2014.28
Rights	© 2014, IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
Download date	2024-05-17 04:02:14
Item downloaded from	https://hdl.handle.net/10468/15487

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

A GPU Implementation of Parallel Constraint-based Local Search

Alejandro Arbelaez
INSIGHT Centre for Data Analytics
University College Cork, Ireland
a.arbelaez@4c.ucc.ie

Philippe Codognet
JFLI - CNRS / UPMC / University of Tokyo
University of Tokyo, Dept. of Computer Science
codognet@is.s.u-tokyo.ac.jp

***Abstract*—In this paper we study the performance of constraint-based local search solvers on a GPU. The massively parallel architecture of the GPU makes it possible to explore parallelism at two different levels inside the local search algorithm. First, by executing multiple copies of the algorithm in a multi-walk manner and, second, by evaluating large neighborhoods in parallel in a single-walk manner. Experiments on three well-known problem benchmarks indicate that the current GPU implementation is up to 17 times faster than a well-tuned sequential algorithm implemented on a desktop computer.**

I. Introduction

In the last decade, the interest for the family of Local Search methods and Metaheuristics for solving large combinatorial problems has been growing and has attracted much attention from both the Operations Research and the Artificial Intelligence communities for solving real-life problems [1]. Theses methods have been used in Combinatorial Optimization for finding optimal or near-optimal solutions for several decades and they are now widely used to solve real-life problems when the search space is too large to be explored by complete search algorithm, such as Mixed Integer Programming or Constraint Solving. There also exist some efficient general-purpose systems for Local Search, such as for instance the commercial system Comet [2].

In general there exists two approaches to devise a parallel local search solver: multi-walk and single-walk [3]. Multi-walk methods consist in developing concurrent explorations of the search space, either independently or cooperatively with some communication between the processes. Sophisticated cooperative strategies for multi-walk methods can be devised by using solution pools [4], but require shared-memory or emulation of central memory in distributed clusters, thus impacting on performance. On the other hand, Single-walk methods consist in using parallelism inside a single search process, *e.g.,* for parallelizing the exploration of the neighborhood.

Graphic Processing Units (GPUs) are nowadays available in nearly all personal computers and they offer a potential reduction in the computational time of local search solvers. The *thread hierarchy* in a GPU consists of threads, blocks, and grids. A block is a batch of threads (all executing the same code) and blocks are grouped in a grid (blocks are independent). This hierarchy matches the architecture of the local search process by allowing the combination of multi-walk and single-walk. Taking this into account, independent local search algorithms can be allocated in different blocks, each block exploiting parallelism by evaluating neighbors in parallel.

The rest of the paper is structured as follows. Section II presents a generic description of local search and the *Adaptive Search* method; a generic, domain-independent constraint-based local search algorithm. Section III presents a description of parallel local search and the most relevant CPU and GPU implementations in the area. Section IV details the architecture of modern GPUs. Section V highlights the main considerations to take into account when designing a local search algorithm targeting a GPU. Section VI details the problems benchmarks used in this paper. Section VII presents experiments of the CPU and GPU implementations. A conclusion and discussion of future work end the paper.

II. Local Search

Local Search algorithms start from an initial assignment for the variables (usually random) and try to improve this configuration, little by little, by small changes in the values of the problem variables. Hence the term "local search" as, at each time step, only new configurations that are "neighbors" of the current configuration are explored. The definition of what constitutes a neighborhood will of course be problem-dependent, but basically it consists in changing the value of a few variables only (usually one or two). The advantage of Local Search methods is that they will usually quickly converge towards a solution (if the optimality criterion and the notion of neighborhood are defined correctly) and not exhaustively explore the entire search space. This is however at the loss of completeness of the search. These methods naturally lead to concurrent execution, by considering the development of several configurations at the same time. This can be done sequentially by maintaining a pool of candidate configurations (as in genetic algorithms) or in parallel if the adequate hardware is available.

The domain of Constraint-based local search, that is, applying local search techniques to solve Constraint Satisfaction Problems has been attracting some interest for about a decade [5], [6], [2]. Constraint Satisfaction can obviously be seen as a branch of Combinatorial Optimization in which the objective function to minimize is the number of violated constraints: a solution is therefore obtained when the function has value zero. Constraint-based local search can tackle large size problem instances, far beyond the reach of classical propagation-based constraint solvers.

A. The Adaptive Search Method

Algorithm 1 presents the general scheme of the *Adaptive Search* (AS) model, a generic constraint-based local search

method proposed nearly a decade ago [5], [7], [8]. This meta-heuristic takes advantage of the structure of the problem in terms of constraints and variables and can guide the search more precisely than a single global cost function to optimize, such as for instance the number of violated constraints. The algorithm also uses a short-term adaptive memory in the spirit of Tabu Search in order to prevent stagnation in local minima and loops. Moreover it intrinsically copes with over-constrained problems.

The input of the method is a Constraint Satisfaction Problem (CSP), which is defined as a triple (X;D;C), where X is a set of variables, D is a set of domains, *i.e.,* finite sets of possible values (one domain for each variable), and C a set of constraints restricting the values that the variables can simultaneously take. For each constraint, an *error function* needs to be defined; it gives, for each tuple of variable values, an indication of how much the constraint is violated. This idea has also been proposed independently by [6], where it is called "penalty functions", and then reused by the Comet system [2], where it is called "violations". For example, the error function associated with an arithmetic constraint $|X - Y| < c$, for a given constant $c > 0$, can be $max(0, |X - Y| - c)$.

AS relies on iterative repair, based on variable and constraint error information, seeking to reduce the error on the variable with the highest error value. The basic idea is to compute the error function for each constraint, then combine for each variable the errors of all constraints in which it appears, thereby projecting constraint errors onto the relevant variables. This combination of errors is problem-dependent, see [5] for details and examples, but it is usually a simple sum or a sum of absolute values, although it might also be a weighted sum if constraints are given different priorities. Finally, the variable with the highest error is designated as the "culprit" and its value is modified. In this second step, the well known min-conflict heuristic [9] is used to select the value in the variable domain which is the most promising, that is, the value for which the total error in the next configuration is minimal.

In order to prevent being trapped in local minima, the AS method also includes a short-term memory mechanism to store configurations to avoid (variables can be marked Tabu and "frozen" for a number of iterations). It also integrates reset transitions to escape stagnation around local minima. A reset consists in assigning fresh random values to some variables (also randomly chosen). A reset is guided by the number of variables being marked Tabu. It is also possible to restart from scratch when the number of iterations becomes too large (this can be viewed as a reset of all variables but it is guided by the number of iterations). The core ideas of adaptive search can be summarized as follow:

- to consider for each constraint a heuristic function that is able to compute an approximated degree of satisfaction of the goals (the current *error* on the constraint);
- to aggregate constraints on each variable and project the error on variables thus trying to repair the *worst* variable with the most promising value;
- to keep a short-term memory of bad configurations to avoid looping (*i.e.* some sort of *tabu list*) together with a reset mechanism.

A few improvements to the basic algorithm are important and worth mentioning here, let us just briefly detail the two main ones. Firstly, when following a *plateau*, a simple stochastic scheme is used for deciding either to continue on the plateau or to escape, by adding a probability p for doing this. With good tuning (*e.g.,* probability of 90% to 95% of following a plateau) this boosts the performance of the algorithm by an order of magnitude on some problems such as MAGIC-SQUARE. Secondly, when too many variables become Tabu, there is a risk of "freezing" the configuration and of getting trapped around a local minimum. Diversification is then performed by a (partial) reset, *i.e.,* by assigning fresh values to a given percentage of the problem variables (parameter RP of the algorithm). A reset is triggered by the total number of variables being marked Tabu at a given iteration (parameter RL of the algorithm).

AS is a simple algorithm but it turns out to be quite efficient in practice. A C-based implementation of AS for permutation problems is available at http://cri-dist.univ-paris1.fr/diaz/adaptive. We used this version as a baseline of the GPU implementation. In [10] AS is briefly compared with Comet [2] on some CSP benchmarks included in the distribution of Comet such as the N-QUEENS problem and the MAGIC-SQUARE problem. Adaptive Search is one or two order of magnitude faster than Comet. Of course, it should be noticed that Comet is a complete and very versatile system while AS is just a C-based library. AS is also compared on the COSTAS ARRAY problem in [11] with Dialectic Search [12] and it is nearly one order of magnitude faster.

III. PARALLEL LOCAL SEARCH

As pointed out above, there are two main parallel techniques to build parallel local search solvers. On one hand, the parallel speedup of the single-walk approach is limited by the Ambahl's law. We recall that the Amdahl's low indicates that the parallel speedup of a given algorithm is bounded by the sequential portions of the code. For instance, an algorithm with 5% of sequential code has a maximum speedup of factor 20. On the other hand, the speedup of the multi-walk method highly depends on the sequential behavior of the algorithm. It is well-known that if the runtime distribution of a local search algorithm follows a pure exponential law (i.e., non-shifted), its parallel multi-walk version will have a linear speedup for an unbounded number of cores [3]. However, in practice, only very few examples have a runtime distribution following a such an ideal case and the speedup is far from linear for a large number of cores [13], [14].

A. Constraint-based Parallel Local Search

Most of the work in parallel constraint-based parallel local search for traditional architectures (*i.e.,* PC clusters) has been devoted to using the multi-walk method, mainly because this method provides two general advantages. First, it requires no extra work to implement, and second it has been theoretically and practically proven to be powerful in a wide range of

Algorithm 1 Adaptive Search Base Algorithm

Input: problem given in CSP format:
- variables X_i with their domains
- constraints C_j w/error functions
- function to project errors on vars
- cost function to minimize

some tuning parameters:
- *TT*: # iterations a variable is frozen
- *RL*: # frozen variables triggering a reset
- *RP*: % of variables to reset
- *MI*: max. # iterations before restart
- *MR*: maximal # of restarts

Output: a solution if the CSP is satisfied or a quasi-solution of minimal cost otherwise.

```
Restart ← 0
repeat
    Restart ← Restart + 1
    Iteration ← 0
    Compute a random assignment A of variables in V
    Opt_Sol ← A
    Opt_Cost ← cost(A)
    repeat
        Iteration ← Iteration + 1
        Compute errors of all constraints in C and combine errors on each variable
                                    ▷ (by considering only the constraints in which a variable appears)
        Select the variable X (not marked Tabu) with highest error
        Evaluate costs of possible moves from X
        if no improvement move exists then
            mark X as Tabu until iteration number: Iteration + TT
            if the number of variables marked Tabu ≥ RL then
                randomly reset RP % variables in V (and unmark those Tabu)
            end if
        else
            Select best move and change X, yielding the next configuration A'
            if cost(A') < Opt_Cost then
                Opt_Sol ← A ← A'
                Opt_Cost ← cost(A')
            end if
        end if
    until Opt_Cost = 0 (a solution is found) or Iteration ≥ MI
until Opt_Cost = 0 (a solution is found) or Restart ≥ MR
output(Opt_Sol, Opt_Cost)
```

domains; moreover, this technique is not affected by the Amdahl's law.

A multi-walk adaptation of AS has been proposed recently and has been implemented on massively parallel computers and grid systems [10], Cell Broadband Engine [15], and devices with Partitioned Global Address Space in [16]. Although for a few problems such as the COSTAS ARRAY parallel speedup can be linear (experiments on the BlueGene supercomputer show linear speedup up to 8000 cores, see [17]), for many problems speedup are good but suboptimal. There is therefore a need for more complex parallelization schemes, for example that would mix single-walks and multi-walks, *i.e.,* parallelizing both neighborhood search and multi-starts.

B. Local Search on GPUs

In [18] the authors use the GPU to accelerate the resolution of a variety of problems such as Quadratic Assignment Problem and Traveling Salesman Problem. The authors propose the use of texture memory to store the definition of the problem and perform the execution of the local search algorithm on both the CPU and the GPU. The CPU is in charge of the control of the algorithm, and the GPU evaluates a rather expensive objective function for a large number of neighbors. This approach allows the exploration of the multi-walk parallel algorithm by executing different copies of the hybrid algorithm on multiple GPUs, *i.e.,* one local search process per GPU. The algorithm reports speedups (w.r.t. to the sequential CPU implementation) of up to 50 for traditional benchmark instances and up to 240 for benchmark instances that required vector of real values.

In the same direction as [18], other authors also use the GPU to accelerate the performance of local search algorithms by means of exploring very large neighborhoods (see [19] for a recent survey). For instance, [20] indicate that the GPU+CPU algorithm reports speedups up to a factor of 27 w.r.t. the sequential implementation for the unrelated parallel machine scheduling problem. [21] uses the GPU to efficiently explore neighborhoods with millions of elements for the vehicle routing problem.

In [22] the authors proposes a framework for heterogeneous systems with multiple CPUs and GPUs interconnected through a network. The framework executes several copies of the local search algorithm and allows cooperation by sharing the best solution found so far for each algorithm. This information is then exploited at each restart to properly craft a new assignment for the variables to start with. The authors present extensive experimental results targeting the TSP problem which indicate that their approach scale up to thousands of cores (including 256 CPUs and 384 GPUs).

In a different context, [23] uses the GPU to solve SAT instances using a tree-based search algorithm, the algorithm

combines backtracking search and clauses learning. Also in the context of SAT solving, in [24] the authors implement the survey propagation algorithm on a GPU and showed that the GPU version is up to 9 times faster than the sequential algorithm implemented in the CPU.

IV. The GPU Architecture

Figure 1 describes the architecture of the GPU. This architecture offers an interesting opportunity to improve the performance of parallel local search solvers. Unlike a CPU which usually consists of few processors units (usually dual or quad cores), a GPU consists of a set of Stream Multiprocessors (SMs), each one features multiple processing units with the ability of executing thousands of operations concurrently. Inside a SM, threads are grouped into warps. A warp executes 32 threads in a SIMD (Single Instruction, Multiple Data) manner, that is all threads within a warp execute the same operation on multiple data points. Furthermore, warps are grouped into a block and blocks are grouped into a grid. NVIDIA offers the Cuda Occupacy Calculator tool which allows to determine the maximum number of active blocks per grid, this number varies for a number of reasons including: memory limitations, maximum number of threads per block, and hardware limitations.

Generally speaking, a GPU uses two types of memory: global and shared. Shared memory is usually limited to few KiloBytes (*e.g.,* 16KB or 48KB) per SM which is equally divided into all blocks allocated for a given SM; an important feature of shared memory is that it has low latency requiring about 2 clock cycles to issue a memory instruction. On the other hand, global memory allows to store a large amount of data (*e.g.,* 6GB), however, it has high latency and requires between 400 and 600 clock cycles to issue a memory instruction. Moreover, GPUs are also equipped with texture and constant memory sections for read-only operations; accessing these two memories is usually faster than the global memory as they include on-chip cache.

Branch divergence and coalesced memory are among the most important aspects to take into consideration when designing algorithms in a GPU. First, branch divergence refers to the fact that threads within the same warp can take different paths (*e.g., if-else* instructions), in this case different execution paths are serialized. In the worst case-scenario a warp executes one thread a time until completing the instruction. Second, in order to obtain peak performance, accesses to global memory should be coalesced, that is, individual threads must access independent and consecutive elements from the global memory; non-coalesced accesses to global memory might result in an important degradation in the performance of the algorithm.

V. Constraint-based Local Search on a GPU

The thread hierarchy of the GPU matches the architecture of the local search algorithm depicted in Section II and allows to exploit parallelism at two different levels. Firstly, multi-walk parallelism can be achieved by executing multiple copies of Algorithm 1 in parallel using different blocks. This form of parallelization is probably the most widely used one due to its simplicity and it often provides a great speedup up to few hundreds of cores. Secondly, the parallel nature of the

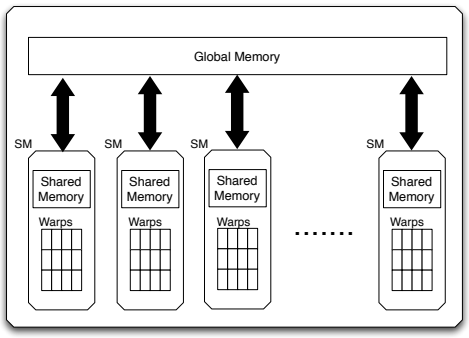


Figure 1. GPU Architecture

GPU also allows to speedup the search process by including parallelism inside the blocks in a single-walk manner; this form of parallelism is also known as parallel neighborhood evaluation. In the context of AS, it would consist in selecting the two most suitable variables for swapping.

Broadly speaking, AS employs two strategies for selecting the variables to swap. The first one (*exhaustive-search*), evaluates all possible swaps at a given state of the search, and selects uniformly at random, the best one. This method explores a large neighborhood, which usually grows quadratically with the size of the problem. The second one (*heuristic-search*), employs a heuristic strategy to select the best move, it starts by selecting the variable X_i with the highest error and then selects another variable X_j that when swapping X_i and X_j minimizes the global error; this method requires a linear complexity $O(n)$, where n is the number of variables of the problem. In both situations, the goal is to select the best move in a considerably large space.

At a first glance, one could consider applying the parallel reduction technique [25], which reduces the complexity to $O(log\ n)$ by applying a tree decomposition of the problem; and returns the best candidate breaking ties in a rather deterministic way. However, a key point to allow diversification in the local search procedure consists in breaking ties with an uniform distribution, then all candidates are equally likely to be selected. Taking this into account, we decided to use the traditional divide-and-conquer approach, where the problem space is divided into several independent subspaces and select, in parallel, the best candidate for each sub-space at a cost of reconciling partial information in order to decide the best action. This simple approach reduces the complexity of an iteration to $O(n/m + m)$, where m is the number of threads per block.

As pointed out in the previous section, an important aspect to take into consideration when designing algorithms in a GPU is the administration of the global and shared memory. Ideally, it would be desirable to store all relevant information in shared memory. However, a CSP might contain thousands of variables and constraints, and it would require too much memory. In a nutshell, the main aspects to consider in the *Adaptive Search* framework are the following:

- Current assignment: requires a linear space complexity and represents the current value for the variables.

- Tabu list: a common way to implement a Tabu list consists in maintaining a list with the age of the variables (indicating the last time it was used). Thus, a variable is Tabu if it has been used within the last $tabuTenure$ iterations.
- Incremental update: a critical aspect when implementing efficient local search algorithms is to incrementally update the objective function. As will be described in Section VI for a few examples, AS employs a set of extra variables to maintain an efficient incremental evaluation of the error function.
- List of best candidate variables: when selecting the swapping variables, it is necessary to maintain a list with the variables that maximizes (or minimizes) the objective function, then a variable from the list is chosen with an uniform distribution.

The decision of using the global or shared memory is entirely up to the problem. In a few cases, it would be possible to only use the shared memory; however, as will be detailed in Section VI, to tackle large and complex instances it would necessary to extensively use the global memory.

VI. Problem Benchmarks

In this section, we detail the benchmarks used for comparison and provide a brief description of how they are implemented in the GPU. One key issue is to reduce accesses to global memory in order to avoid degrading performance of the GPU implementation.

Magic Square: This is problem 19 of CSPLib[1]. It consists in completing a $N \times N$ matrix with numbers $1, 2, \ldots, N^2$, where each column, row, and the two main diagonals sum the same objective number b. As described in [5], the *Adaptive Search* model of this problem is an instance of magic square in which b=$N(N^2+1)/2$. The error function for each constraint is defined as $f(X)$=X_1+X_1+...+X_k - b, and the global error is the sum of the absolute value of the error for all constraints.

For an efficient incremental evaluation of the global error, the algorithm maintains a list of $2n$+2 elements with the error of the columns, rows, and the two diagonals. This way, obtaining the error of swapping two variables is then done in constant time. AS uses *exhaustive-search* to identify the most suitable variables to swap at each iteration of the algorithm; therefore, the algorithm explore a neighborhood of N^2 elements, where N is the input of the problem.

We decided to keep the list variables with the errors for columns, rows, and the two diagonals in shared memory, and the remaining variables in global memory. The GPU implementation of the magic square problem uses the same parameter configuration as the CPU version, except that in order to reduce accesses to global memory, the GPU implementation does not clean the Tabu list after performing a reset.

Number Partition: This is problem 49 of CSPLib. It consists in dividing a set of numbers from 1 to N into two subsets, where both subsets have the same cardinality, the same sum of numbers, and the same sum of squares of numbers. The model of this problem involves two set of $N/2$ variables each one, and the global error is obtained as the sum the absolute value of the errors of the two following constraints:

$$\sum_{i=1}^{N/2} X_i = \sum_{i=N/2+1}^{N/2} X_i = N(N+1)/4$$
$$\sum_{i=1}^{N/2} X_i^2 = \sum_{i=N/2+1}^{N/2} X_i^2 = N(N+1)(2N+1)/12$$

At each iteration of the local search process the algorithm uses *exhaustive-search* (or exhaustive search), and explores all possible swaps by exchanging values from the two subsets. Thus, the algorithm needs to explore a neighborhood of $(N/2)^2$ elements. For this problem, the shared memory is used for the list of variables representing the current configuration, and the global memory stores the remaining variables. The GPU version of the algorithm uses the same parameters as the CPU version, except that the CPU uses a Tabu list of size one, and the GPU algorithm omits the Tabu list in order to reduce accesses to global memory.

Costas Array: This is a hard combinatorial problem abstracted from radar and sonar applications [26]. It consists in placing N marks in a $N \times N$ grid, such that there is exactly one mark per row, column, and the $N(N-1)/2$ vectors joining the marks are all different. This problem is modeled using the difference triangle, in which N variables form a permutation of numbers from 1 to N, and the $j-th$ row of the triangle contains the differences between the variables X_{i+j} - X_i, where $j \in [1, N-1]$ and $i \in [1, N-j]$. In order to obtain the cost of a given assignment for the variables the algorithm computes the global error by checking the first $(N-1)/2$ rows of the triangle. In this case, AS uses *heurisitc-search* to identify the best move at each iteration of the local search process. Since this problem requires a rather small number of variables and constraints, we decided to store all variables in shared memory.

VII. Experiments

This section compares the performance of the CPU and GPU implementations of AS. All the experiments were performed using a 6-core processor (Xeon W3670 at 3.20 Ghz) and 12MB of RAM, a GPU NVIDIA Tesla C2075 which features 14 SMs, 6GB of global memory, 1.15 Ghz clock rate and up to 48KB of shared memory per SM, and a GPU GeForce GTX 760 which features 6 SMs, 2GB of global memory, 1.15 GHz clock rate and up to 48KB of shared memory per SM. Although both GPUs allow more than a thousand threads per block, the following experimentation is limited to up to 512 threads per block due to hardware limitations. *e.g.,* shared memory and registers per SM. We implemented the GPU algorithm using CUDA version 5. For all the experiments we used the default reset function described in Section II, and in the following results we report the median value of 20 executions unless otherwise stated.

At the time this article was written, the price of the baseline processor was US$ 824.50, the Tesla GPU was US$ 1784.99 and the Geforce GPU was US$ 288.52 as indicated at www.amazon.com. Tesla GPUs are usually more expensive than the GeForce ones as they support double precision, error-correcting code, and usually provide more dedicated memory; we refer the reader to www.nvidia.com for a complete

[1] www.csplib.org

Table I. GPU (TESLA) PERFORMANCE OF MAGIC SQUARE 200 (TIME IN SECS)

Blocks \ Threads	32	64	128	256	512
1	492	280	96	60	79
7	116	56	48	26	17
14	59	35	28	20	**16**
28	64	64	23	27	18
56	52	24	37	39	25

Table II. GPU (TESLA) PERFORMANCE OF MAGIC SQUARE 300 AND 400 (TIME IN SECS)

Size	Blocks \ Threads	128	256	512
300	14	107	75	**54**
	28	76	80	79
400	14	298	**160**	273
	28	191	279	209

Table III. CPU PERFORMANCE OF MAGIC SQUARE (TIME IN SECS)

Cores \ Size	200	300	400
1	72	970	2572
3	35	115	710
6	**24**	**98**	**483**

description of both GPUs. Hereafter, we will observe that in addition of the speedup gained with the GPU w.r.t. the sequential CPU version, the GeForce GPU also provides an important price/performance benefit, that is, a lower cost and higher performance.

We start our analysis with Figure 2, which depicts the quantile-quantile plot (Q-Q plot) of the number of iterations required to solve three small size instances of Magic Square (20), Partition (2600), and Costas Array (15) for the CPU and GPU implementations of AS. The Q-Q plot displays the quantiles of the data of the two samples of 500 runs. A given *quantile* is a point in a sample for which a certain fraction of the sample lies, *e.g.,* the 0.3 quantile is the point for which 30% falls below and 70% above. For example the well-known *median value* of a sample is the 0.5 quantile (also noted 50% quantile). A *quantile-quantile plot* will depict, for the same quantile, the value in one sample as abscissa and the value of the other one as ordinate. Thus, checking if both samples match is straightforward: if points are on the diagonal ($y = x$), then both samples perfectly fit. The dimension of both axes are the number of iterations.

Although the GPU implementation reuses large portions of the original algorithm, the figure shows that the two versions behave similarly but are not perfectly identical. This is due the slightly differences indicated the previous section to provide an efficient GPU algorithm.

Tables I & II report the runtime in seconds to solve Magic Square problem (N=200, 300 and 400) on the GPU (Tesla) using different configurations for the grid, *i.e.,* varying the ratio threads/block. We set the minimum number of threads per block to 32 as this is minimum allowed by CUDA and the number of blocks varies from 1, 7, 14, 28, and 56. Bold number indicate the best performance obtained in this table. As expected, we observe a reduction in the runtime by only using the multi-walk parallel scheme (columns), however, the speedup changes when increasing the number of threads per block. For instance, when using 32 threads per block, we observe a maximum speedup of 9.46 (1 block vs. 56 blocks), but when using 512 threads per block, we observe a maximum speedup of 3.16. There are several aspects that influence the speedup of the multi-walk algorithm in the GPU, such as divergent branch, non-coalesced memory, bandwidth occupancy, warp synchronization, etc. We recall that such speedup differences should not be observed in homogeneous cluster with nodes that have similar computation capabilities. For these reasons, increasing the number of threads per block might degrade the performance of the multi-walk parallel algorithm.

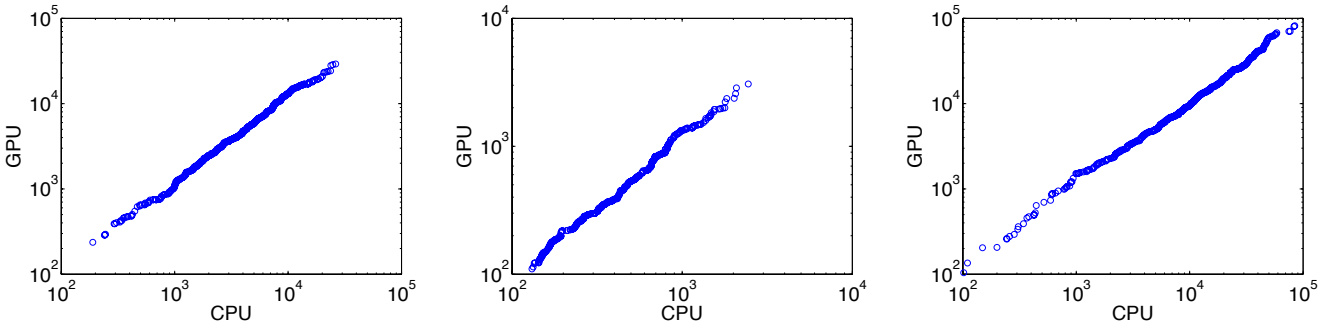
The best performance with the GPU (Tesla) for Magic Square 200 is 16 secs, this value is obtained when using 14 blocks and 512 threads per block. We attribute this optimal value to the fact that the reference GPU uses 14 SM. On the other hand, it also important to notice that this optimal performance obtained with the GPU is 4.5 times faster than the sequential CPU version (see Table III) and 1.5 times faster than the best runtime obtained with the CPU using 6 cores. Furthermore, the best result obtained for Magic Square 300 (54 secs) is 17 times faster than the sequential algorithm, and 1.8 times faster than the best CPU multi-core implementation (with 6 cores). And for Magic Square 400 the best result of the GPU (160 secs) is 16 times faster than the sequential CPU version[2], and 3 times faster than the best CPU multi-core performance (with 6 cores). We believe that increasing the size of the problem might also increase the difference in the performance between the GPU and the CPU. At this point it is worth noticing that parallel versions of local search algorithms can achieve super-linear speedups w.r.t. the sequential algorithm (*e.g.,* CPU implementation of Magic Square 300). This is the case for instance when the sequential runtime distribution follows a lognormal distribution, which is heavy-tailed. This is because the sequential algorithm exhibits the heavy tail phenomena, this phenomena is less relevant when executing the algorithm in parallel [14].

Interestingly, the performance using a single GPU outperforms a multiple-walk implementation of AS for the Magic Square problem (size 400) on a supercomputer using 64 cores, *i.e.,* 198 secs (Hitachi HA8000 Supercomputer with 64 cores) vs. 160 secs (GPU with 14 blocks and 256 threads), see [10] for complete details of the performance of AS on the supercomputer.

Let us now switch our attention to the Partition problem (size 9600). Table IV, this table depicts the experimental results of the GPU (Tesla). In this case, we observe that the best performance is 6 secs, obtained when using 14 blocks and 512 cores. On the other hand, the GPU reports a speedup of 17 w.r.t. the sequential implementation, and a speedup of 5 against the best parallel CPU implementation (with 6 cores).

Tables VII and VI show the results of two instances of the Costas array problem (*i.e.,* N=18 and N=19). Unlike Magic Square and partition, which involve the exploration of large neighborhoods (quadratic w.r.t. the size of the problem), the Costas array problem requires to explore a small neighborhood

[2]The runtime of the CPU sequential version of Magic Square (N=400) was computed as the median value of 11 executions



(a) Magic Square 20 (b) Partition 2600 (c) Costas 15

Figure 2. Q-Q plot of the CPU and GPU implementations for three different CSPs based on 500 runs per algorithm.

Table IV. GPU(TESLA) PERFORMANCE OF PARTITION 9600 (TIME IN SECS)

Blocks \ Threads	32	64	128	256	512
1	521	279	104	123	84
7	68	45	32	9	9
14	37	28	16	9	**6**
28	37	30	14	10	9
56	43	30	12	8	9

Table V. CPU PERFORMANCE OF PARTITION (TIME IN SECS)

Cores \ Size	9600
1	102
3	49
6	30

of N elements, where N is usually a small number. Therefore, for this problem we increase the number of blocks up to 112 (being 144 the maximum number of blocks allowed for this GPU, *i.e.,* 12 blocks per SM). In this case, the best performance observed for the GPU is about 2 times faster than the sequential CPU implementation for Costas 18, and about 3 times faster for Costas 19; however, due to the great speedup capabilities of the AS to solve the Costas array problem the best parallel CPU implementation (with 6 cores) is faster than the GPU one.

Even though the parallel multi-core implementation for Costas is better than the GPU (Tesla) one, these results are encouraging due to the fact that for this problem the GPU is being underused. Notice that for this problem the algorithm does not exploit parallelization in a single-walk manner. Indeed, part of our future work will be to support the multi-walk method at a warp level, *i.e.,* executing multiple local search processes inside a single block.

Finally, Table VIII shows the results using the GeForce GPU with the best two configurations obtained in the previous experiments, *i.e.,* 14 and 28 blocks, and 512 threads per block

Table VI. GPU (TESLA) PERFORMANCE OF COSTAS ARRAY (TIME IN SECS)

Size \ Blocks	1	7	14	28	56	112
18	204	34	22	10	8	**5**
19	1453	189	128	72	33	**21**

Table VII. CPU PERFORMANCE OF COSTAS ARRAY (TIME IN SECS)

Size \ Cores	1	3	6
18	10	4	**0.3**
19	66	24	**19**

Table VIII. GPU (GEFORCE) PERFORMANCE FOR MAGIC SQUARE,PARTITION AND COSTAS (TIME IN SECS)

Blocks	512 threads/block			
	Magic Square			Partition
	200	300	400	9600
14	14	76	248	7
28	17	59	292	6

Blocks	1 thread/block	
	Costas	
	18	19
112	1	8

(Magic Square and Partition) and 1 thread block (Costas). We recall that Costas does not include single walk parallelization. In this table we observe that the GeForce GPU reports nearly the same results as the Tesla one for Magic Square and Partition. For instance, the best results obtained in the Tesla GPU for Magic Square (N=400) is 209 seconds vs. 248 seconds for the GeForce GPU; for the Partition problem both GPUs reported 6 seconds as best result; and for Costas the GeForce GPU is considerably faster than the Tesla one, and this GPU is also faster than the multi-core version for Costas-19 (8 seconds vs. 19 seconds). We attribute the important performance difference in the Costas problem to the fact that for this problem all relevant information is stored in the local memory and the global memory is not used, therefore factors such as coalesced memory and bandwidth are not relevant in this particular example.

One of the main advantages of the Tesla over the GeForce GPU is the double precision, however, AS does not require intensive double precision operations. Therefore, even thought the GeForce GPU is considerably cheaper (price-wise) than the Tesla GPU, the performance is comparable and sometimes better. In addition, we observe that the GeForce GPU is a promising alternative to traditional CPUs, obtaining a higher performance for a lower cost.

VIII. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented a GPU implementation of the *Adaptive Search* framework. This new version of *Adaptive Search* exploits the architecture of the GPU to device a parallel local search algorithm. Thus, the algorithm exploits parallelism by executing multiple local search processes and at the same

time allows the exploration of the large neighborhoods in parallel. Unlike other local search implementations which are problem dependent and combine the execution of the CPU and the GPU, in this work the entire local search process is executed on the GPU; therefore, the performance is not bounded to a robust CPU. Extensive experiments using three well-known problem benchmarks indicate that the GPU (Tesla C2075) provides speedups up to 17 w.r.t. a well tuned sequential version of the Magic Square and Partition problems, and up to 3 for the sequential CPU version of the Costas Array problem. Furthermore, we believe that increasing the input size for the problems might also increase the difference between the CPU and the GPU. In addition, we also observed that a much cheaper GPU (GeForce GTX 760) report a higher performance than traditional multi-core architectures, on a set of well-known constraint satisfaction problems, for a lower price.

The current GPU version of AS allows the execution of only one local search process per block, and limits the number of threads for the parallel evaluation of the neighborhood to 512 (max. number of threads per block); however, different problems might require more computational power. For instance, more threads to evaluate the neighborhood (*e.g.,* Partition and Magic Square), and other problems can exploit better the parallelization by executing multiple local search processes per block (*e.g.,* Costas array). We also plan to study very large neighborhoods using the GPU, we recall that the current version is an adapted version of CPU implementation, however, the parallel nature of the GPU allows more flexibility. Furthermore, we plan to extend the current implementation to support execution on multiple GPUs and clusters of GPUs. Lastly, we believe that there is significant room for improving the performance of the parallel algorithm by means of including cooperation between the solvers. Local search solvers can exploit parallelization by exchanging information such as the best assignment for the variables found so far, and using this information to push the trajectory of the algorithm to promising areas of the search space [27].

IX. Acknowledgements

This work was supported by the Japan Society for the Promotion of Science (JSPS) under the JSPS Postdoctoral Program and the *kakenhi* Grant-in-aid for Scientific Research. The first author is currently supported by DISCUS, and FP7 Grant Agreement 318137.

References

[1] T. Gonzalez, Ed., *Handbook of Approximation Algorithms and Metaheuristics*. Chapman and Hall / CRC, 2007.

[2] P. Van Hentenryck and L. Michel, *Constraint-Based Local Search*. The MIT Press, 2005.

[3] M. Verhoeven and E. Aarts, "Parallel local search," *Journal of Heuristics*, vol. 1, no. 1, pp. 43–65, 1995.

[4] T. G. Crainic, M. Gendreau, P. Hansen, and N. Mladenovic, "Cooperative parallel variable neighborhood search for the -median," *Journal of Heuristics*, vol. 10, no. 3, pp. 293–314, 2004.

[5] P. Codognet and D. Diaz, "Yet another local search method for constraint solving," in *proceedings of SAGA'01*. Springer Verlag, 2001, pp. 73–90.

[6] P. Galinier and J.-K. Hao, "A general approach for constraint solving by local search," in *2nd workshop CP-AI-OR'00*, Paderborn, Germany, 2000.

[7] P. Codognet and D. Diaz, "An efficient library for solving CSP with local search," in *MIC'03, 5th International Conference on Metaheuristics*, T. Ibaraki, Ed., 2003.

[8] M. A. H. Newton, D. N. Pham, A. Sattar, and M. J. Maher, "Kangaroo: An efficient constraint-based local search system using lazy propagation," in *CP*, ser. Lecture Notes in Computer Science, J. H.-M. Lee, Ed., vol. 6876. Springer, 2011, pp. 645–659.

[9] S. Minton, M. D. Johnston, A. B. Philips, and P. Laird, "Minimizing conflicts: A heuristic repair method for constraint satisfaction and scheduling problems," *Artificial Intelligence*, vol. 58, no. 1-3, pp. 161–205, 1992.

[10] Y. Caniou, P. Codognet, D. Diaz, and S. Abreu, "Experiments in parallel constraint-based local search," in *EvoCOP'11*, ser. LNCS, P. Merz and J.-K. Hao, Eds., vol. 6622. Torino, Italy: Springer, April 2011, pp. 96–107.

[11] D. Diaz, F. Richoux, P. Codognet, Y. Caniou, and S. Abreu, "Constraint-based local search for the costas array problem," in *LION 6, Learning and Intelligent OptimizatioN Conference*. Paris, France: Springer LNCS, 2012.

[12] S. Kadioglu and M. Sellmann, "Dialectic search," in *CP'09, 15th Int. Conf. on Principles and Practice of Constraint Programming*, ser. Lecture Notes in Computer Science. Springer Verlag, 2009.

[13] A. Arbelaez, C. Truchet, and P. Codognet, "Using sequential runtime distributions for the parallel speedup prediction of sat local search," *Journal Theory and Practice of Logic Programming (TPLP)*, special issue, proceedings of ICLP13, to Appear.

[14] C. Truchet, F. Richoux, and P. Codognet, "Prediction of Parallel Speed-ups for Las Vegas Algorithms," in *Proceedings of ICPP-2013, 42nd International Conference on Parallel Processing*, J. Dongarra and Y. Robert, Eds. IEEE Press, October 2013.

[15] D. Diaz, S. Abreu, and P. Codognet, "Parallel constraint-based local search on the cell/be multicore architecture," in *proceedings of IDC2010, Intelligent Distributed Computing IV*. Springer Verlag, 2010.

[16] R. Machado, S. Abreu, and D. Diaz, "Parallel local search: Experiments with a pgas-based programming model," in *International Colloquium on Implementation of Constraint and Logic Programming Systems (CICLOPS)*, Budapest, Hungary, 2012.

[17] D. Diaz, F. Richoux, Y. Caniou, P. Codognet, and S. Abreu, "Parallel local search for the costas array problem," in *IEEE Workshop on new trends in Parallel Computing and Optimization (PC012), in conjunction with IPDPS 2012*. Shanghai, China: IEEE Press, May 2012.

[18] T. V. Luong, N. Melab, and E.-G. Talbi, "Gpu computing for parallel local search metaheuristic algorithms," *IEEE Trans. Computers*, vol. 62, no. 1, pp. 173–185, 2013.

[19] E.-G. Talbi and G. Hasle, "Metaheuristics on gpus," *J. Parallel Distrib. Comput.*, vol. 73, no. 1, pp. 1–3, 2013.

[20] I. M. Coelho, M. N. Haddad, L. S. Ochi, L. S. Ochi, and R. Farias, "A hybrid cpu-gpu local search heuristic for the unrelated parallel machine scheduling problem," in *3rd workshop on Applications for Multi-Core Architectures (WAMCA)*, 2012, pp. 19–23.

[21] C. Schulz, "Efficient local search on the gpu - investigations on the vehicle routing problem," *J. Parallel Distrib. Comput.*, vol. 73, no. 1, pp. 14–31, 2013.

[22] M. Burtscher and H. Rabeti, "A scalable heterogeneous parallelization framework for iterative local searches," in *IPDPS'13*, 2013.

[23] A. D. Palu, A. Dovier, A. Formisano, and E. Pontelli, "Exploiting unexploited computing resources for computational logics," in *9th Italian Convention on Computational Logic*, 2012.

[24] P. Manolios and Y. Zhang, "Implementing survey propagation on graphics processing units," in *SAT*, 2006, pp. 311–324.

[25] G. E. Blelloch, "Scans as primitive parallel operations," *IEEE Trans. Computers*, vol. 38, no. 11, pp. 1526–1538, 1989.

[26] K. Drakakis, "A review of costas arrays," *Journal of Applied Mathematics*, vol. 2006, pp. 1–32, 2006.

[27] A. Arbelaez and P. Codognet, "Massivelly Parallel Local Search for SAT," in *ICTAI'12*. Athens, Greece: IEEE Computer Society, November 2012, pp. 57–64.