A Dynamic Load-Balancing Algorithm for Heterogeneous GPU Clusters

Luis Sant'Ana, *CMCC-UFABC* Daniel Cordeiro, *DCC-USP* and Raphael Camargo, *CMCC-UFABC*

***Abstract*—The use of GPU clusters for scientific applications is becoming more widespread, with applications in areas such as physics, chemistry and bioinformatics. These clusters can be highly heterogenous, composed of several models of CPUs and GPUs, that could be used together to execute these applications. To use these heterogeneous devices in an efficient manner, it is necessary to load-balance the application task sizes among the GPUs and CPUs, minimizing the execution time of the application.**

We propose an algorithm for dynamic load balancing in heterogeneous GPU clusters that determines the execution profile of tasks on each device in runtime. This profile is used to find the best distribution of task sizes among devices. We implemented the algorithm in the StarPU framework and compared its performance with existing load-balancing algorithms, using applications from linear algebra, stock markets and bioinformatics.

***Keywords*—parallel computing, distribuited systems, GPU cluster, GPGPU.**

I. Introduction

The use of GPUs (Graphics Processing Units) is becoming increasingly popular among developers and users of applications that have high computational demands and can benefit from a large degree of parallelism [1]. Among these applications we can include fluid mechanics [2], visualization science [3], machine learning [4], bioinformatics [5] and neural networks [6].

Modern GPUs can have thousands of simple floating point units (FPUs) that, combined, can have a computational power several times superior to traditional CPUs. The main drawback is the difficult in optimizing GPU code, specially considering the architectural differences among different GPUs, with changes in the distribution of cores, shared memory, presence of cache, among others. But libraries for several classes of applications and algorithms are now available, facilitating the deevlopment process [7], [8].

Computationally demanding applications may also benefit from the usage of multiple GPUs, normally distributed on several machines, which are called GPU clusters [9], [10]. The development of applications for these clusters is more complex, since it requires the management of the multiple memory spaces, one for each GPU in the cluster, in addition to the main memory of the each node. This management includes transferring data between these memory spaces and ensuring the consistency of data.

There are several efforts by the scientific community for the creation of new programming models [11], [12] and frameworks [13], [14], [15], for facilitating the developement of GPU cluster applications. However, the combination of CUDA (Compute Unified Device Architecture) and MPI (Message Passing Interface) is still the standard choice to develop applications for GPU clusters.

High-performance GPU clusters are typically *homogeneous*, containing nodes and GPUs with the same configuration. Homogeneity facilitates the development of applications, since they can be optimized just for a single architecture. Moreover, if the application has exclusive access to the nodes, load-balancing is simpler, since the same task size, or number of tasks, can be allocateds to each node. But homogeneity can be difficult to maintain in a context where a new generation of hardware is launched every couple of years. In this case, joining heterogeneous machines can increase the available computational power to cluster users.

Another scenario where heterogeneity is more common is when connecting machines from several researchers groups using a standard Ethernet network. This enables a large amount of computing power from commodity hardware that could be efficiently used by application with low communication demands. These machines could be used by these applications outside the work time and the aquisition cost and space demand for this commodity-hardware GPU cluster would be almost zero.

Developing a load-balancing mechanism that works efficiently for all kinds of applications is difficult. With two or more GPUs (or CPUs), this problem is strictly equivalent to the classic problem of minimizing the maximum completion time of all tasks (makespan), which is known to be NP-hard [16]. An efficient load-balancing scheme must be considered on case by case basis. For example, there are several types of data-parallel applications [17] that could be divided using domain decomposition. Several scientific applications fit into this group, including applications in bioinformatics [5], neural networks [6], chemistry, physics and materials science.

When using homogeneous clusters, data can be distributed among the available GPUs using tasks of the same size. However, with heterogeneous GPUs, this distribution is more difficult. There are several existing GPU architectures, such as *Tesla*, *Fermi*, *Kepler* and *Maxwell*. These architectures have different organizations of FPUs (Floating-point Unit), caches, shared memory and memory speeds. Even for GPUs with the same architectures, their characteristics can vary considerably.

Luis Sant'Ana is with the Center Mathematics, Computation and Cognition, Federal Uniersity of ABC, Santo André, SP, Brazil e-mail: luis.ana@ufabc.edu.br

Daniel Cordeiro is with the Department of Computer Science, University of São Paulo, São Paulo, SP, Brazil e-mail: danielc@ime.usp.br

Raphael Carmargo is with the Center Mathematics, Computation and Cognition, Federal Uniersity of ABC, Santo André, SP, Brazil e-mail: raphael.camargo@ufabc.edu.br

A division of the load based on simple heuristics, such as a the number of cores in the GPU, is not effective and can be worse than a simple homogeneous division [9]. Also, different architectures require different low-level optimizations and a code could have been better optimized for one architecture. Finally, GPUs clusters normally have high-end CPUs in addition to the GPUs, and these CPUs can used by the applications.

The main task of the load-balancing mechanism is to devise the best data division amont the GPUs. A possible approach is to determine the performance profiles for each GPU type and application task and use it to determine the amount of work given to each GPU. This profiling can be done statically, before application execution [9], or dynamically, during application execution [18], [19]. Another solution is to use simple algorithms for task dispatching, such as the greedy algorithm from StarPU [15], where tasks are dispatched to the devices as soon as they become available.

In this work we propose a novel adaptive algorithm for dynamic load balancing of data-parallel applications in heterogeneous GPU clusters. The algorithm uses runtime task execution time measurements to create a performance model for each device (CPU or GPU). The algorithm dynamically adjust the size of data blocks allocated to each device based on this model. We compared the proposed algorithm with static [9] and dynamic (HDSS) [19] algorithms and with the standard StarPU greedy algorithm.

II. Related Work

A proposed solution for load balancing in distributed systems is to divide the tasks among CPUs according to a weight factor representing the processing speed of each processor. Early approaches used fixed weight factors determined at compile time with limited success [20]. But these weight factors are difficult to determine in heterogeneous GPUs, due to the architectural differences that affect the performance.

The problem of load balancing in heterogeneous GPUs, began to be studied recently. One proposal was the usage of a static algorithm that determines the distribution before the execution of the application, using profiles from previous executions [9]. The algorithm uses these profiles to find the distribution of data that minimizes the execution time of the application, ensuring that all GPUs to spend same amount of time performing the processing of kernels. The algorithm was evaluated using a large-scale neural network simulation. Its main drawback is that since it is static, an initial unbalanced distribution cannot be djusted in runtime. Another problem is that it require preivous executions of the appliction in the target deveices to determine its exexuction profiles. Finally, it does not consider the case where application behavior changes with the parameters. A dynamic algorithm, like the one that we propose, do not have these limitations.

Acosta *et al.* [18] proposed a dynamic load balancing algorithm that interactively finds a good distribution of work between GPUs during application execution. It uses a decentralized scheme in which synchronizations are done at each iteration to determine whether there is a need to rebalance the load. Each process, representing a single GPU, performs a redistribution of workload according to the size of the assigned task and the capabilities of the GPU. A vector of time is shared among all processors. Each processor processes the load itself and records the time this vector. The first step of the algorithm is to check whether the threshold defined by user is not being exceeded. If not the processors need not recalculate the load, otherwise the balancing is performed. The algorithm computes a vector called RP (Relative Power), which corresponds to the relationship between time spent to process a certain load versus versus the time taken for a computational unit as a function of the problem size. The processors calculate the SRP (Sum Relative Power) which is the sum of all RP vector calculated for each processor. Finally, each processor calculates the load itself, taking into account the size of the problem, the RP and the SRP. The disadvantage of this algorithm is that the function that calculates the load of each processor does not provide the precision, that generates unnecessary balancing. Our algorithm for generating a curve of execution is able to determine with certain accuracy the optimal size of the load assigned to each processor, preventing balancing unnecessary.

The Heterogeneous Dynamic Self-Scheduler (HDSS) [19] is a dynamic load-balancing algorithm for heterogenous GPU clusters. It is divided in two phases. The first is the adaptive phase, where it determines weights that reflect the speed of each GPU, similarly to Hummel *et al.* [20], but performed dynamically and using profiling data. A performance curve with the FLOPs per second for each task size is created for each GPU. The weights is determined based on logarithmic fits on the curves. These weights are used during the second phase, called completion phase, where it divides the remaining iterations among the GPUs based on their relative weights. It starts allocating larger block sizes to the GPUs, decreasing their size as the execution progresses. An important drawback of HDSS is that the weight model based on logarithmic curves is valid only for GPUs, where the number of FLOPs per second stabilizes with larger tasks. Also, after the weights are determined they cannot be changed. We use an execution model that fits for both GPUs, CPUs and other device types. Moreover, we use the complete performance models, instead of a single number, resulting in a better load-distribution. Finally, we use the task execution time data from all the execution, permiting later adjustments in the performance models of the devices.

III. Proposed Algorithm

In a typical data-parallel application, the application data is divided among the threads in a process called domain decomposition [17]. The threads then simultaneously process their part of the data and this processing can be on single or multiple steps, dependending on the task sizes. After finishing, the threads merge the processed results, and the application terminates or continues to the next phase of computing. The task of our load-balancing algorithm is determining the size of the data block assigned to each GPU and CPU in the system. We will use the term processor to mean a single CPU or GPU.

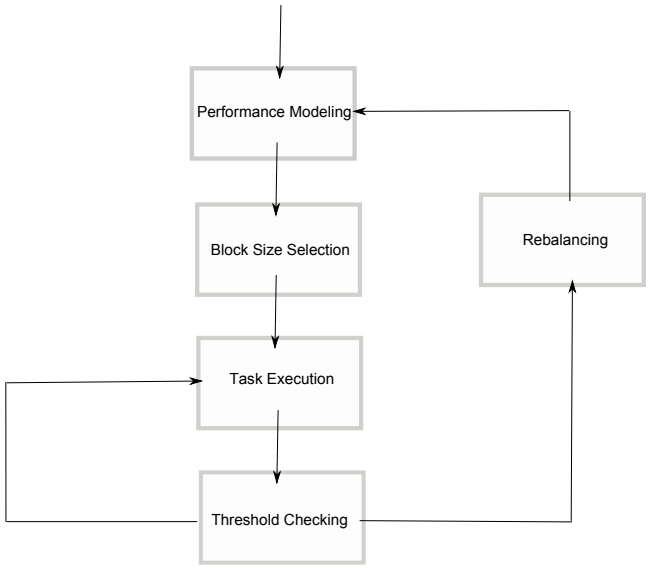

Fig. 1. Flowchart of the algorithm

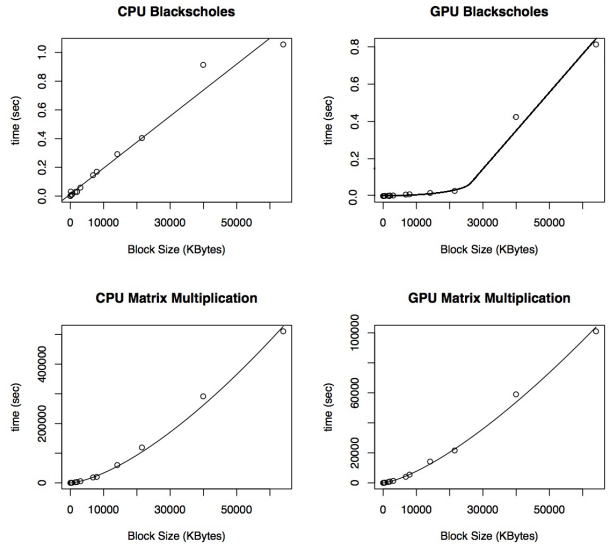

Fig. 2. GPU and CPU processing times for different block sizes.

Overview: The algorithm has three parts, which are: (1) processor performance modeling, where a performance model for each processor is determined during application execution based on task execution times; (2) optimal block size selection, where, based on the performance model, the algorithm determines the best distribution of block size among the processors; and (3) re-balancing, where the algorithm recalculates the optimal block sizes when the difference in execution time by different processors is larger than a threshold.

Processor performance modelling.: In this part, the algorithm constructs a function $P_p[x]$, which represents the execution time of a block of data of size x on processor p.

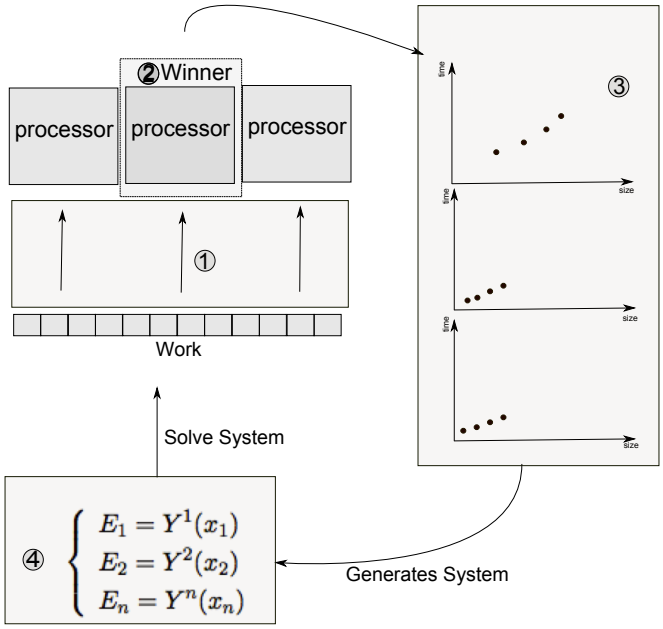

Fig. 3. Summary of the algorithm

For this, it executes blocks of different sizes on each processor and interpolate the curve that better fits these points.

A block of size $initialBlockSize$, defined by the user, is initially allocated and executed on each processor. After executing the first block, the algorithm selects the processor with the earliest finish time t_f and double the block size for this processor. For other processors, with finish times t_k, we select blocks of size $initialBlockSize * t_k/t_f$. This procedure is repeated until generating four points per curve. This process corresponds to (1) in Figure 2, which shows a schematic view of processor performance modelling part.

The algorithm then follows the step (2) of Figure 2, where it tries to fit a curve in the existing points using the least squares fitting method. If the coefficient of determination is greater than 0.7 for all processors, the algorithm considers the fit as good enough and finishes this phase. Otherwise, it generates more points in the curve following the previous procedure.

We find best fit curves by the method of *least squares*, using a function of the form $f(x) = ax - b$.

Figure 1 shows sample processing time measurements for a GPU and a CPU for different block sizes and two applications. We can see that the curves can be approximated by different types of functions.

The steps of this phase are shown in Algorithm 1, which is executed in a single node, called master node. Variable $blockListSize$ contains the size of the blocks assigned to each processor and is initialized with $initialBlockSize$, which is defined by the user. Variable $fitValues$ contains results from the last least square fitting, including the $error$ in the fitting, which is initialized as $+\infty$. In the main loop, while the error is larger than a predefined value, the function sends a chunk of data for each processor and obtains the finish time for each processor. After all processors finish their executions,

Algorithm 1 Processor performance model

```
function determineModel()
blockSizeList ← initialBlockSize;
while fitValues.error ≥ 0.5 do
    finishTimes = executeTasks(blockSizeList);
    blockSizeList = evaluateNextBlockSizes(finishTimes);
    fitValues = determineCurveProcessor();
end while
return fitValues;
```

it determine the block sizes for the next iteration, based on their finish times. Finally, it tries to fit model curves to each processor and receives the fitting results.

Optimal block size selection: In this part our algorithm determines the optimal block size for each processor with the objective of providing blocks with the same execution time on each processor.

Consider we have n processors and a input data size Z. The algorithm assigns a data chunk of size x^g for each processor $g = 1, ..., n$, corresponding a fraction of input data, such that $\sum_{g=1}^{n} x^g = Z$. We denote as $E^g(x^g)$ the execution time of task E in the processor g, for input of size x^g. To distribute the work among the processors, we find a set of values

$$X = \{x^g \in \mathbb{R} : [0, 1] / \sum_{g=1}^{n} x^g = Z\} \tag{1}$$

that minimizes $E_1(x_1)$ while satisfying the constraint

$$E_1 = E_2 = ... = E_n \tag{2}$$

which represents that all processors should spend the same amount of time performing the processing. To determine the set of values x, we solve the system of fitted curves for all processors, determined in the performance modelling part, and given by:

$$\begin{cases} E_1 = f(x_1) \\ E_2 = f(x_2) \\ E_n = f(x_n) \end{cases} \tag{3}$$

The equations system is solved applying an interior point line search filter method [21], which finds the minimum solution of a convex equation set, subject to a set of constraints, by transversing the interior points of its feasible region.

Execution and Rebalancing: After determining the task sizes, the scheduler sends a block of the selected size x_i for each processor i. When a processor finishes executing a task, it requests another task of the same size. The processors then continue the execution asynchronously, until completing all the tasks.

The scheduler also monitors the finish time of each task. If the difference in finishing times t_i and t_j between any two tasks of processors i and j goes above a threshold, the rebalancing process is executed. Small thresholds may cause excessive rebalancing while large thresholds may tolerate larger imbalances that will cause processor idlenesses.

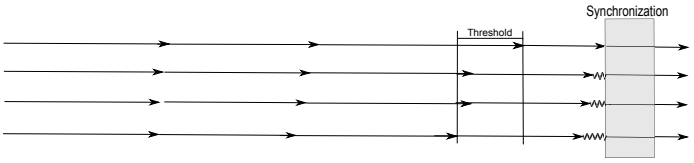

Fig. 4. Execution of tasks by four processors. The finish time difference of two tasks was above a threshold, which caused a rebalancing of the task sizes.

The threshold is determined empirically, but values of about XXX% of the execution time of a single value resulted in a good tradeoff.

During the rebalancing, the scheduler synchronizes the processors and apply the processor performance modelling algorithm to fit the best curves for the functions $P_p[x]$ updated with the execution times. The optimal block size selection routine is then applied to determine new block sizes x_i for each processor i.

Figure 3 shows a diagram of execution of four processors, where each arrow indicates the start of a task execution. The tasks start synchronized, but their finishes times start to diverge until the differences reaches a threshold. In this case, all threads are synchronized, and the new task sizes are recalculated for each processors. The task executions then restart.

Complete algorithm: Algorithm 2 shows the pseudo-code of the scheduling algorithm. The $determineModel$ function, shown in Algorithm 1 returns the performance model for each processor into variable $fitValues$. The algorithm then solves the system of equations contained in $fitValues$ to determine the best distribution X of task sizes for each processor.

Algorithm 2 Complete dynamic algorithm

```
function dynamic()
fitValues = determineModel()
X = solveEquationSystem(fitValues);
while there is data do
    finishTimes = executeTasks(X);
    if maxDifference(finishTimes)≥ threshold then
        fitValues = determineCurveProcessor();
        X = solveEquationSystem(fitValues);
        synchronize();
    end if
end while
```

The main loop repeats while there is still data for processing. It distributes tasks of different sizes for each processor in the system, obtaining the finish times for each task execution. It then checks if the maximum difference between the finish tasks is above a defined threshold. If the threshold is reached, the algorithm fits new curves for each processor performance model and solve the system of equations to determine a new distribution of tasks sizes for each processor.

IV. Implementation

The implementation was done in the C language with the framework StarPU. StarPU [15] is a tool for parallel

programming that supports hybrid architectures like multicore CPUs and accelerators. The StarPU proposes an approach of independent tasks based architecture. Codelets are defined as an abstraction of a task that can be performed on one core of a multicore CPU or subjected to an accelerator. Each codelet may have multiple implementations, one for each architecture in which codelet can be performed using specific languages and libraries for the target architecture. A StarPU application is described as a set of codelets with data dependencies.

The tool has a set of scheduling policies implemented that the programmer can choose according to the characteristics of the application. The main one is the use of static scheduling algorithm HEFT (Heterogeneous Earliest Finish Time) to schedule tasks based on cost models of task execution.

For each device one codelet has been programmed with the characteristics of the devices. A codelet is a structure that represents a computational kernel. Such a codelet may contain an implementation of the same kernel on different architectures (e.g. CUDA and x86). The applications were implemented by dividing the data set into tasks, implemented as codelets. The tasks are independent, with each task receiving a part of the input set proportional to the processor weight. Two codelets were implemented one for GPU/CUDA and one for CPU architecture.

To evaluate our load-balancing algorithm, we implemented it by modify the default StarPU balancing algorithm. The modification of the load balancing algorithm is realized by changing the STARPU_SCHED variable. The STARPU framework has an API that allows modifying the scheduling policies. There are data structures and functions that speed up the process of development. For example the function "double starpu_timing_now (void)" return the current date in micro seconds, which makes it easier for the determination of measures runtime. In StarPU there is a data structure called "starpu_sched_policy" This structure contains all the methods que Implement a scheduling policy.

Three other algorithms were implemented for comparison: the greedy, static and HDSS. The greedy consisted in dividing the input set in pieces and assigning each piece of input to any idle processor, without any priority assignment. The static [9], measures processing speeds before the execution and set static block sizes per processor at the beginning of the execution, with the block size proportional to the processor speed. Finally, The HDSS [19] was implemented using minimum square estimation to estimate the weights and divided into two phases: adaptation phase and completion phase.

The library used for solve the equation system like 3 was the IPOPT [21]. IPOPT (Interior Point Optimize) is an open source software package for large-scale nonlinear optimization. It can be used to solve general nonlinear programming problems.

A. Applications

For evaluating our algorithm, we adapted two applications from the CUDA SDK [22] to execute using the StarPU framework, the blackscholes and matrix multiplication applications. And we adapted a application to gene regulatory networks (GRN) inference [23].

A copy of the matrix A was distributed to all processing units and matrix B was divided according to the load-balancing scheme, the matrix multiplication version use the shared memory. Matrix multiplication has complexity $O(n^{3/2})$.

Blackscholes is a popular financial analysis algorithm for calculating prices for European style options. The Black-Scholes equation is a differential equation that describes how, under a certain set of assumptions, the value of an option changes as the price of the underlying asset changes. More precisely, it is a stochastic differential equation that includes a random walk term, which models the random fluctuation of the price of the underlying asset over time. The Black-Scholes equation implies that the value of a European call option.

The cumulative normal distribution function, gives us the probability that a normally distributed random variable will have a value less than x. There is no closed-form expression for this function, and as such it must be evaluated numerically. It is typically approximated using a polynomial function. The idea is to calculate the Black-Scholes the greatest amount of possible options. Thus, the input is a vector of data that the options should be calculated by applying a differential equation. The division of the task is given a position of providing input vector to each thread. The complexity of the algorithm is $O(n)$.

Gene regulatory networks (GRN) inference is an important bioinformatics problem in which the gene interactions need to be deduced from gene expression data, such as $microarray$ data. Feature selection methods can be applied to this problem. A feature selection technique is composed by two parts: a search algorithm and a criterion function. Among the search algorithms already proposed, there is the exhaustive search where the best feature subset is returned, although its computational complexity is unfeasible in almost all situations. The objective of work is the development of a low cost parallel solution based on GPU architectures for exhaustive search with a viable cost-benefit. The complexity of the algorithm is $O(n^3)$. The division of labor consisted in distributing genes between processors, and certain steps to synchronize the information search in the solution space, more details of the algorithm in [23].

V. Results

A. System Configuration

We used three different machines to evaluate our algorithm, presented in table **??**. We performed the experiments with four settings, with one machine (A), with two machines (A, B), with three machines (A, B, C) and four machines (A, B, C and D). The machine A has a CPU with 4 cores and two GPUs with 280 cores. The machine B has 1 CPU with 6 cores and two GPUs with 1536 cores. The machine C has 1 CPU with 6 cores and 1 GPU with 2688 cores and finally the machine D has 1 CPU with 14 cores and 2 GPUs with X cores.

To use all the n multiprocessors from a GPU, it is necessary to create at least n blocks. Moreover, each multiprocessor simultaneously executes groups (called warps) of m threads from a single block, and several warps should be present on each GPU for efficient usage of its cores.

TABLE I. CONFIGURATION OF THE MACHINES

Machine	Description						
A	**CPU Info**	intel i7 a20	4 cores	2.67 GHz	8192 MB cache	8 GB RAM	1 processor
	GPU Info	GTX 295	280 cores	223.8 GB/s Memory Bandwidth	896 MB	999 MHz Memory Clock	2 processors
B	**CPU Info**	intel i7 4930K	6 cores	3.4 GHz	12,288 KB	32 GB RAM	1 processor
	GPU Info	GTX 680	1536 cores	192.2 GB/s Memory Bandwidth	2 GB	6 GHz Memory Clock	2 processors
C	**CPU Info**	intel i7 3939K	6 cores	3.2 GHz	12,288 KB cache	32 GB RAM	1 processor
	GPU Info	GTX Titan	2688 cores	223.8 GB/s Memory Bandwidth	6 GB	6 GHz Memory Clock	1 processor
D	**CPU Info**	intel Xeon E5-2695V3	14 cores	2.3 GHz	35 MB	32 GB RAM	1 processor
	GPU Info	?	?	?	?	?	

In all tests, we used all the SM(Stream Multiprocessor) of the GPUs, launching kernels with k blocks per with 1024 threads for each block, where k is the number of processor in the GPU. For the used GPUs, k is **14, 8 and 30** in GTX Titan, GTX 680 and GTX 295 respectively. For the CPUs, we used all the CPU cores, launching one task per core.

The parameters used for each algorithm were obtained via experiments. A lot of values were tested. The best parameters were obtained specifically for each application.

B. Matrix Multiplication

We evaluated the execution time of the matrix multiplication application using one, two, three and four machines and four different scheduling algorithms: (1) our algorithm, (2) static, (3) HDSS, and (4) Greedy.

Figure 4 shows the results for matrices with sizes from 4096 x 4096 to 65536 x 65536 elements. In all scenarios, our scheduling algorithms obtained the best results, with the HDSS algorithm in second. The static and greedy algorithm were clearly inferior.

With one machine the difference was smaller because there are few types of devices for the scheduler to select. With two machines our algorithm starts to perform better, especially for larger matrices, which is explained by the fact that the execution time of the matrix multiplication increases quickly as we increase the matrix size. With the increasing size of the matrix there is a trend of increasing performance gap among algorithms. As the environment is heterogeneous the behavior is not well defined.

For matrices 4096 elements and four machines, the proposed algorithm spent 23.11 seconds while the HDSS spent 29.91 seconds, which results in 26.09% faster. And for matrices 65536, the algorithm proposed spent 428.01 seconds while the HDSS 488.22, which results in 14.02% faster, the summary is table II. The total cost of calculating the distribution of the blocks is 664.87 milliseconds and standard deviation of 98.3 mili seconds.

The results obtained shows that in more heterogeneous environment the advantage using our proposed algorithm is largest than the other. This advantage is due to a better distribution of the data along the execution. Figure 5 shows the time difference among the earliest and latest finishing threads, in the scenario with three machines and the size matrix fized with 65535 elements.

The greedy algorithm performed well considering that it do not directly use information on the processing speeds of the devices. But it uses this information indirectly, since faster devices finish their tasks earlier and, consequently, receive more tasks. The static algorithm had the worst performance, because it leaves the thread idle for a long time see the figure 5. There are big differences among end of threads.

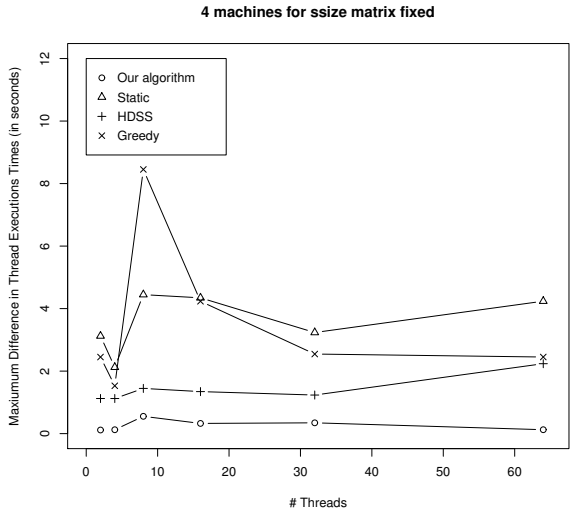

Fig. 6. Time differences between the earliest and latest finishing threads

HDSS uses the beginning of the execution to estimate the best block size distribution and uses this distribution through the complete execution. Moreover, larger blocks are used in the beginning of the execution, causing a larger delay due to disbalance when use different types of devices, such as GPUs and CPUs. Finally, HDSS uses a crude approximation to the device capabilities.

Our algorithm estimate fairly accurately the amount of data that should be provided to each processing unit, solving an optimization problem with the restriction that all units should finish the execution of the tasks at the same time. When the difference between finishing the execution of threads for certain data partition exceeds a certain threshold, the algorithm re-balances the data distribution.

C. Blackscholes

We evaluate the execution time of the Blackscholes application using different machine configurations. Figure 6 shows the execution times, where we varied the number of options on each execution and obtained the runtimes. Similarly to the matrix multiplication experiments, the performance gains

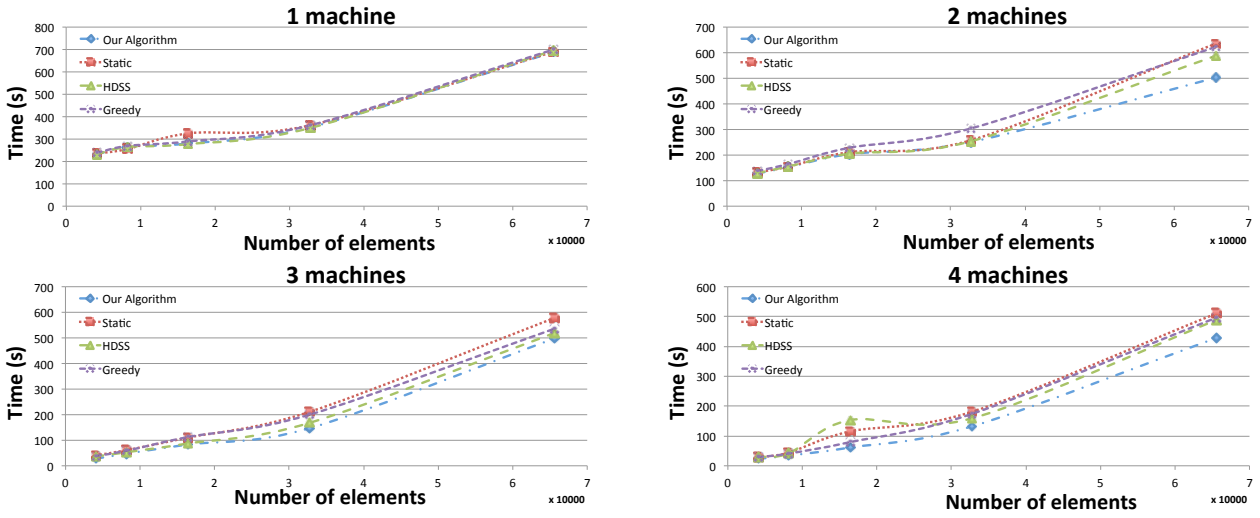

Fig. 5. Difference in runtime with different sizes of matrices for matrices multiplication

TABLE II. COMPARATIVE: HDSS X OUR ALGORITHM

	4096 elements			65536 elements		
Num. Machines	HDSS (s)	Our algorithm (s)	Diff. (%)	HDSS (s)	Our Algorithm (s)	Diff. (%)
1	233	231	0.86	696	689	1.01
2	129	128	0.78	587	502	16.93
3	36	28	28.57	543	497	9.25
4	29.91	23.11	26.09	488.22	428.01	14.02

were the largest with bigger problem sizes (number of options) and more heterogeneous environments, and the results can be explained using the same arguments. Interestingly, the Blackscholes application has linear complexity with the number of options, showing that our scheduling algorithm is also useful with this class of application. The table III shows the extreme values of 10,000 and 500,000 options, the best result was with 4 machines, and with the greatest number of options.

Also, considering that the application finishes in less than 4 seconds, for the scenario with 4 machines, we also see that the cost of solving the optimization problem to determine the best data distribution is small and the obtained gains outweighs the cost of the calculations. For applications tested in a few iterations around 6, it was possible to obtain the solution of the system in a few milliseconds. The total cost of calculating the distribution of the blocks is 201.32 milliseconds and standard deviation of 0.34 miliseconds

Figure 7 confirms this result, showing that the time difference among the earliest and latest finishing threads, in the scenario with four machines is always smaller for our proposed algorithm.

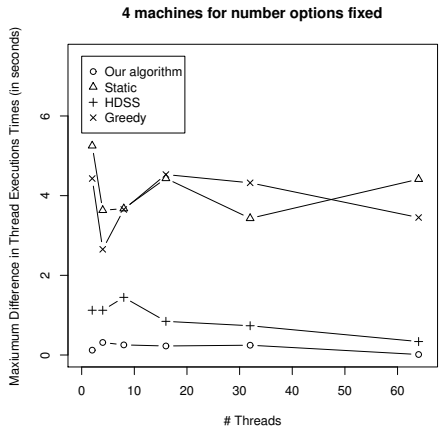

Fig. 8. Time differences between the earliest and latest finishing threads

D. Gene regulatory networks inference

As in previous applications, we compare the execution time in four different environments: with one machine, two machines, three machines and four machines. The results are shown in 8. We can notice that our algorithm outperformed in four cases. Especially in the case that three and four machines were used.

The table IV shows the extreme values, for 100,000 and 2,000,000 genes. To the environment with 3 and 4 machines the performance were similar due to limitation in data transfer, in the case of four machines the communication cost limited the gain. The total cost of calculating the distribution of the

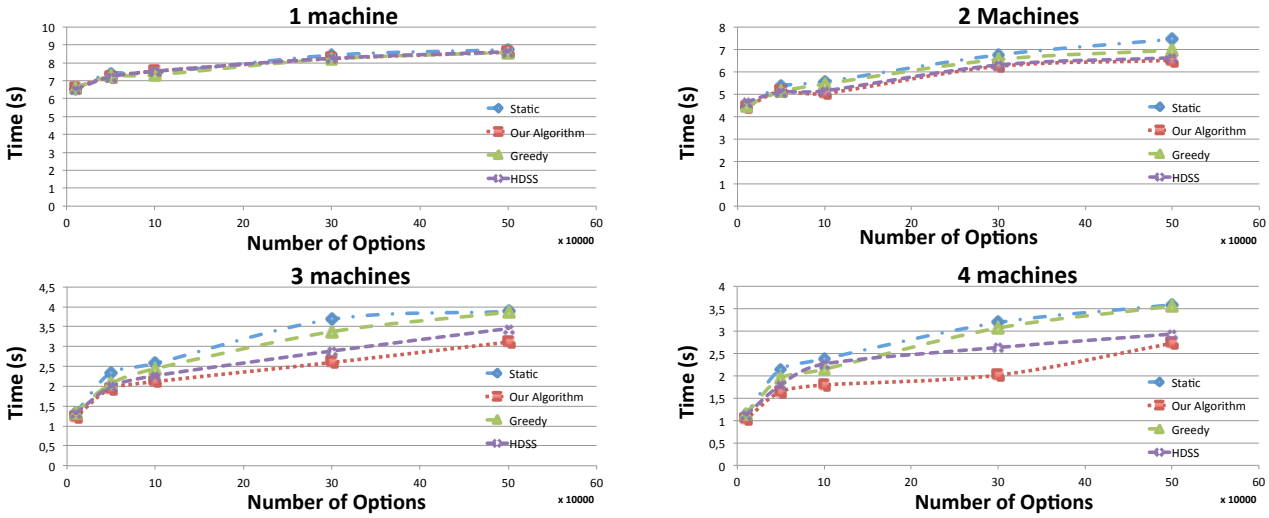

Fig. 7. Difference in runtime with different number options

TABLE III. COMPARATIVE: HDSS X OUR ALGORITHM

	10,000 Options			500,000 Options		
Num. Machines	HDSS (s)	Our algorithm (s)	Diff. (s)	HDSS (%)	Our algorithm (s)	Diff. (%)
1	6.52	6.56	-0.61	8.61	8.59	0.23
2	4.65	4.41	5.44	6.62	6.51	1.68
3	1.31	1.24	5.56	3.45	3.11	10.93
4	1.11	1.02	8.82	2.93	2.72	7.72

blocks is 821.54 milliseconds and standard deviation of 142.23 miliseconds.

Like the other experiments, the maximum difference among the threads followed the same principle. Our algorithm has the smallest difference among the threads, figure 9.

E. Block size evolution

In the figure 10 we compare the behavior in relation to submission of blocks among machines. It analyzed the block size for the HDSS and our algorithm. In this case three machines were used, it is possible to notice the difference between runs. Our algorithm in step 400 suffered a re-balancing, causing it to change the size of the block. In HDSS, a decrease of the block size remains over the iterations.

VI. CONCLUSIONS AND FUTURE WORK

In this paper we propose an algorithm for scheduling tasks in domain decomposition problems executing on clusters of heterogeneous CPUs and GPUs. It outperforms other similar existing algorithms, due to the online estimation of the performance curve for each processing unit and the selection of the best data distribution for the devices. We showed for two applications that our algorithm provides the highest gains for more heterogeneous and larger problems sizes.

Although we used dedicated clusters, we can also consider the usage of public clouds, where the user can request a number of resources allocated in virtual machines from shared machines. In this case, the quality of service may change during execution, and the addition of the execution time difference threshold permits readjustments in data distributions.

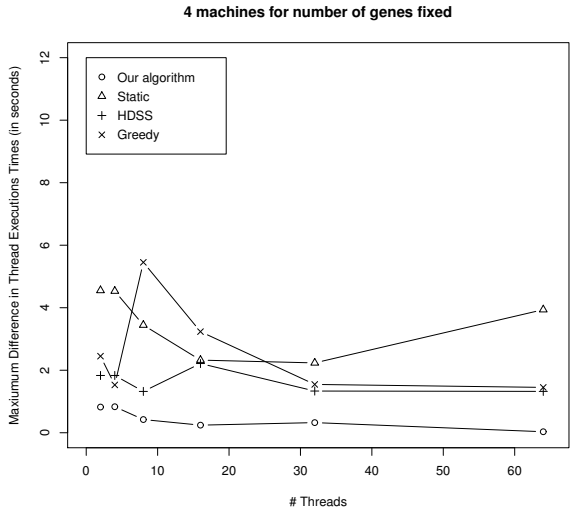

Fig. 10. Difference in runtime with different numbers of genes for gene regulatory network inference

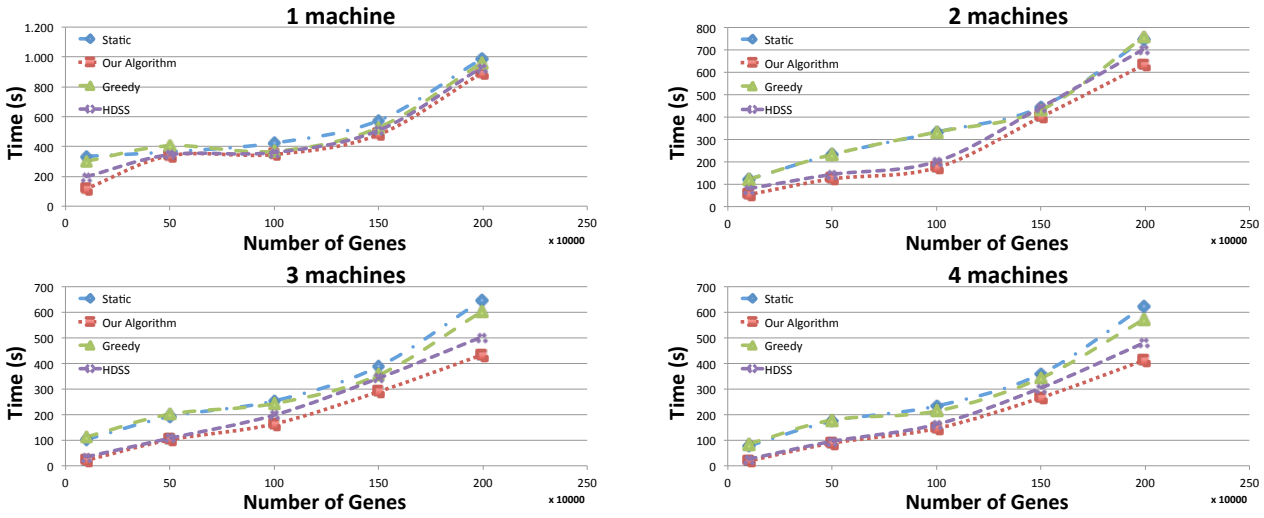

Fig. 9. Difference in runtime with different numbers of genes for gene regulatory network inference

TABLE IV. COMPARATIVE: HDSS X OUR ALGORITHM

	100,000 Genes			2,000,000 Genes		
Num. Machines	HDSS (s)	Algorithm (s)	Diff. (%)	HDSS (s)	Algorithm (s)	Diff. (%)
1	197.94	116.43	70.00	927.43	889.13	4.31
2	79.76	53.43	49.28	702.65	632.73	11.05
3	34.76	21.39	62.50	502.65	432.43	16.24
4	25.91	18.54	39.75	479.54	411.46	16.54

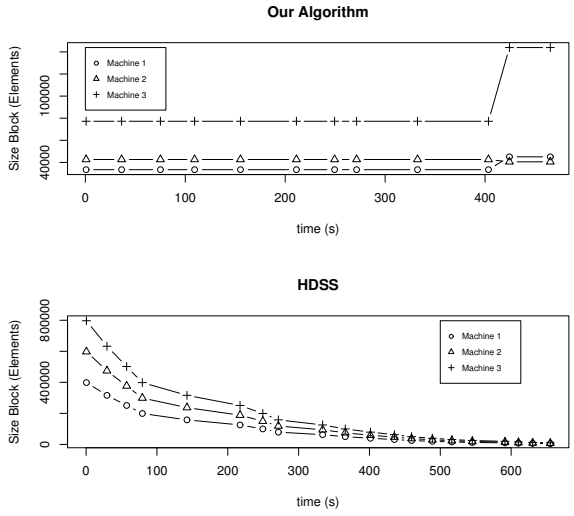

Fig. 11. Difference in runtime with different numbers of genes for gene regulatory network inference

We can also consider the scenario with added fault-tolerance, where machines may become unavailable during execution. In this scenario, a simple redistribution of the data among the remaining devices would permit the application to readapt to this scenario.

As ongoing work, we are including the cost of communication in the scheduling algorithm, which is essential for applications where the time spent with information exchange among the tasks cannot be ignored.

ACKNOWLEDGMENT

The authors would like to thank FAPESP (Proc. n. 2012/03778-0 and Proc. n. 2013/14603-9) for the financial support.

REFERENCES

[1] A. Jog, E. Bolotin, Z. Guz, M. Parker, S. W. Keckler, M. T. Kandemir, and C. R. Das, "Application-aware memory system for fair and efficient execution of concurrent gpgpu applications," in *Proceedings of Workshop on General Purpose Processing Using GPUs*, ser. GPGPU-7. New York, NY, USA: ACM, 2014, pp. 1:1–1:8.

[2] G. Oyarzun, R. Borrell, A. Gorobets, O. Lehmkuhl, and A. Oliva, "Direct numerical simulation of incompressible flows on unstructured meshes using hybrid cpu/gpu supercomputers," *Procedia Engineering*, vol. 61, no. 0, pp. 87 – 93, 2013, 25th International Conference on Parallel Computational Fluid Dynamics.

[3] H. K. Raghavan and S. S. Vadhiyar, "Efficient asynchronous executions of amr computations and visualization on a gpu system," *Journal of Parallel and Distributed Computing*, vol. 73, no. 6, pp. 866 – 875, 2013.

[4] Q. Li, R. Salman, E. Test, R. Strack, and V. Kecman, "Parallel multitask cross validation for support vector machine using gpu," *Journal of Parallel and Distributed Computing*, vol. 73, no. 3, pp. 293 – 302, 2013, models and Algorithms for High-Performance Distributed Data Mining.

[5] Y. Jeon, E. Jung, H. Min, E.-Y. Chung, and S. Yoon, "Gpu-based acceleration of an rna tertiary structure prediction algorithm," *Computers in Biology and Medicine*, vol. 43, no. 8, pp. 1011 – 1022, 2013.

[6] J. Liu and L. Guo, "Implementation of neural network backpropagation in cuda," in *Intelligence Computation and Evolutionary Computation*, ser. Advances in Intelligent Systems and Computing, Z. Du, Ed. Springer Berlin Heidelberg, 2013, vol. 180, pp. 1021–1027.

[7] "Nvidia® cudnn – gpu accelerated machine learning."

[8] "Cuda math library."

[9] R. de Camargo, "A load distribution algorithm based on profiling for heterogeneous gpu clusters," in *Applications for Multi-Core Architectures (WAMCA), 2012 Third Workshop on*, 2012, pp. 1–6.

[10] Z. Fan, F. Qiu, A. Kaufman, and S. Yoakum-Stover, "Gpu cluster for high performance computing," in *Proceedings of the 2004 ACM/IEEE conference on Supercomputing*, ser. SC '04. Washington, DC, USA: IEEE Computer Society, 2004, pp. 47–.

[11] L. Wang, M. Huang, V. K. Narayana, and T. El-Ghazawi, "Scaling scientific applications on clusters of hybrid multicore/gpu nodes," in *Proceedings of the 8th ACM International Conference on Computing Frontiers*, ser. CF '11. New York, NY, USA: ACM, 2011, pp. 6:1–6:10.

[12] L. Wang, W. Jia, X. Chi, Y. Wu, W. Gao, and L.-W. Wang, "Large scale plane wave pseudopotential density functional theory calculations on gpu clusters," in *High Performance Computing, Networking, Storage and Analysis (SC), 2011 International Conference for*, 2011, pp. 1–10.

[13] J. Kim, S. Seo, J. Lee, J. Nah, G. Jo, and J. Lee, "Snucl: an opencl framework for heterogeneous cpu/gpu clusters," in *Proceedings of the 26th ACM international conference on Supercomputing*, ser. ICS '12. New York, NY, USA: ACM, 2012, pp. 341–352.

[14] T. Miyoshi, H. Irie, K. Shima, H. Honda, M. Kondo, and T. Yoshinaga, "Flat: a gpu programming framework to provide embedded mpi," in *Proceedings of the 5th Annual Workshop on General Purpose Processing with Graphics Processing Units*, ser. GPGPU-5. New York, NY, USA: ACM, 2012, pp. 20–29.

[15] C. Augonnet, S. Thibault, and R. Namyst, "Starpu: a runtime system for scheduling tasks over accelerator-based multicore machines," Laboratoire Bordelais de Recherche en Informatique - LaBRI, RUNTIME - INRIA Bordeaux - Sud-Ouest, Rapport de recherche RR-7240, Mar 2010.

[16] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman and Company, 1979.

[17] W. D. Gropp, "Parallel computing and domain decomposition," in *Fifth International Symposium on Domain Decomposition Methods for Partial Differential Equations, Philadelphia, PA*, 1992.

[18] A. Acosta, V. Blanco, and F. Almeida, "Towards the dynamic load balancing on heterogeneous multi-gpu systems." in *Parallel and Distributed Processing with Applications (ISPA), 2012 IEEE 10th International Symposium on*, 2012, pp. 646–653.

[19] M. E. Belviranli, L. N. Bhuyan, and R. Gupta, "A dynamic self-scheduling scheme for heterogeneous multiprocessor architectures," *ACM Trans. Archit. Code Optim.*, vol. 9, no. 4, pp. 57:1–57:20, 2013.

[20] S. F. Hummel, J. Schmidt, R. N. Uma, and J. Wein, "Load-sharing in heterogeneous systems via weighted factoring," in *in Proceedings of the 8th Annual ACM Symposium on Parallel Algorithms and Architectures*, 1997, pp. 318–328.

[21] J. Nocedal, A. Wächter, and R. Waltz, "Adaptive barrier update strategies for nonlinear interior methods," *SIAM Journal on Optimization*, vol. 19, no. 4, pp. 1674–1693, 2009. [Online]. Available: http://dx.doi.org/10.1137/060649513

[22] CUDA Development Core Team, "CUDA 4.1 Programming Guide," Santa Clara CA, EUA, 2012. [Online]. Available: http://www.nvidia.com/cuda

[23] F. F. Borelli, R. Y. de Camargo, D. C. Martins Jr, and L. C. Rozante, "Gene regulatory networks inference using a multi-gpu exhaustive search algorithm," *BMC bioinformatics*, vol. 14, no. 18, pp. 1–12, 2013.