

Holistic Slowdown Driven Scheduling and Resource Management for Malleable Jobs

Marco D'Amico

marco.damico@bsc.es

Barcelona Supercomputing Center
Barcelona, Spain

Ana Jokanovic

ana.jokanovic@bsc.es

Barcelona Supercomputing Center
Barcelona, Spain

Julita Corbalan

juli@ac.upc.edu

Universitat Politècnica de Catalunya
Barcelona, Spain

ABSTRACT

In job scheduling, the concept of malleability has been explored since many years ago. Research shows that malleability improves system performance, but its utilization in HPC never became widespread. The causes are the difficulty in developing malleable applications, and the lack of support and integration of the different layers of the HPC software stack. However, in the last years, malleability in job scheduling is becoming more critical because of the increasing complexity of hardware and workloads. In this context, using nodes in an exclusive mode is not always the most efficient solution as in traditional HPC jobs, where applications were highly tuned for static allocations, but offering zero flexibility to dynamic executions. This paper proposes a new holistic, dynamic job scheduling policy, Slowdown Driven (SD-Policy), which exploits the malleability of applications as the key technology to reduce the average slowdown and response time of jobs. SD-Policy is based on backfill and node sharing. It applies malleability to running jobs to make room for jobs that will run with a reduced set of resources, only when the estimated slowdown improves over the static approach. We implemented SD-Policy in SLURM and evaluated it in a real production environment, and with a simulator using workloads of up to 198K jobs. Results show better resource utilization with the reduction of makespan, response time, slowdown, and energy consumption, up to respectively 7%, 50%, 70%, and 6%, for the evaluated workloads.

CCS CONCEPTS

• **Software and its engineering** → **Scheduling**; • **Computer systems organization** → *Multicore architectures*.

KEYWORDS

job scheduling, co-scheduling, malleability, resource management, high performance computing

1 INTRODUCTION AND MOTIVATION

In static HPC environments, it is a widely extended practice to allocate full nodes for exclusive utilization. This solution minimizes jobs interferences and applications can be tuned to exploit all the resources available in the node. However, execution environments, architectures, applications and workflows are more and more increasing their complexity. On one side, traditional HPC applications became just a piece of complex puzzles, including, data analytics applications, visualizations tools, creating elaborated workflows. To support those workloads, computing nodes increased their complexity, with the multicore and manycore technology delivering high computing power, in combination with different memory layers for better I/O management. It is difficult for every single piece

of the workload to exploit all the available resources, so dynamic resource management approaches need to be evaluated.

In a context where jobs can share nodes to take advantage of all the available resources, the execution of those workflows could highly improve the hardware efficiency.

Resource sharing as a strategy to improve resource utilization has been considered before with time-sharing approaches such as Gang-scheduling [17] or space-sharing, such as oversubscription for co-scheduling [31], both being not a specific policy but a family of policies in the same way backfill includes many variants. Resource sharing often poorly perform, because of the contention created by multiple applications that, tuned to run in exclusive job allocations, share the same resources, and by the overhead of context switching.

Our proposal overcome this problem by exploiting malleability [14] as a way for efficiently managing resources inside computing nodes. Programming models such as OpenMP [30], OpenMPs [6], and MPI [28] give basic support to changing the number of threads or tasks and their binding to resources, but the rest of HPC layers lack communication between them and this feature remain isolated. We started to address this problem by developing DROM, an API [5] that enables the scheduler to communicate with applications that can adapt at run time to changes in the computing resources. We took advantage of shared memory programming models to hide programming complexity and to allow dynamic cores allocation efficiently. We also showed that malleability outperform static resource sharing solutions and reduce response time.

In this paper we present SD-Policy, a Slowdown Driven, dynamic job scheduling policy, which exploits malleability offered by DROM as the key technology to reduce the average slowdown and response time of jobs. SD-Policy, based on backfill, applies malleability to running jobs to make room for jobs that will run with a reduced set of resources, only when the estimated slowdown improves over the static scheduling. SD-Policy supports mixed workloads with malleable, moldable and static applications, ideal for being used in transition to a malleable environment. Our approach is holistic, as we put effort connecting all the HPC software stack: applications, programming models, node resource management, and system level scheduling need to work in coordination to achieve the best performance. We implemented SD-Policy into the SLURM [1] Workload Manager by extending its plug-ins, and evaluated it with two methodologies: in a real environment, the Marenostrum4 (MN4) [2] system, using up to 49 computing nodes, and in the BSC SLURM simulator [19], using standard workloads [12] [13], up to 198K jobs on 5040 nodes. With the two methodologies, we demonstrate on one side the feasibility of the SD-Policy in a real production system, and, on the other hand, that significant performance benefits can

be obtained with dynamic scheduling policies when executing long, meaningful workloads. Our main contributions are:

- A dynamic job scheduling, Slowdown-Driven, policy for malleable jobs based on backfill and co-scheduling.
- An efficient resource selection algorithm based on the feedback of system metrics, i.e., average slowdown.
- The integration of our proposal in the full software stack: SLURM workflow manager, standard programming models like OpenMP, OpenMPs, and MPI, with a negligible effort from developers.

The next sections are organized as following: Section 2 introduces concepts, software, libraries used in our work, Section 3 describes the developed job scheduling policy and its implementation in a well know workload manager. Section 4 presents both simulations and real runs evaluations on the presented policy and its parameters, together with an in-depth analysis of their effects on system performance. Finally, Section 5 present the state of the art of the subject of scheduling in the context of malleable jobs, and Section 6 resumes and concludes our research, with insight on future work.

2 BACKGROUND

2.1 Malleability and DROM

A malleable job is able to adapt to changes in the allocated resources dynamically at runtime[12]. For modern HPC hardware architectures, it is important to divide malleability into two parts: malleability among different resources available in the nodes, and malleability in the number of computing nodes.

In previous work, we developed DROM (Dynamic Resource Ownership Management) [5], an interface that puts in communication applications and node managers, enabling the first level of malleability. DROM avoids the high overhead of re-configuring and transferring data needed by the second approach, as explained in the related work, in Section 5.

The interface enables malleability for jobs by integrating with shared memory programming models, like OpenMP [30] and OpenMPs [6]. It is designed to be integrated with any programming model, or directly with applications, enabling shrink and expand operations in malleability points, e.g., between tasks execution, where it is safe to change used resources. DROM eliminates overhead for developers that only need to dynamically link the library, without modifying or recompiling the code. The library also implements API for registering processes in the DROM environment, getting the list of recorded processes, and getting/setting their CPU masks. Those API can be used by the resource manager, in particular by the node manager, that is in charge to communicate with applications when the scheduler asks to change the job's allocation.

We evaluated the overhead of DROM in different use cases, and we found that the cost of shrinking and expanding operations are negligible, while response time and resource utilization improve. The overhead is given by higher level cache invalidation and additional code injected by the library every time a job shrinks or expands. Motivated by positive results, we decided to investigate the benefits of widely using DROM in an HPC system, e.g., by malleable-backfilling relatively short jobs, that complete fast, to improve the system slowdown. Low overhead allows SD-Policy to

apply malleability at high frequency, avoiding at the same time to penalize only some specific jobs. Using the second level of malleability at this granularity would require more significant amounts of time just to perform the shrinking and expanding operations, voiding the potential benefits on the system. On the other hand, our policy is compatible with malleability in the number of computing nodes, that applied at a different granularity, could bring of further improvements in system performance.

2.2 SLURM and SLURM simulator

SLURM workflow manager [1] is an open source, efficient job manager, popular in research and widely used in HPC systems, installed on 6 of the top 10 supercomputers of the top500 list.

SLURM is based on a master-slave architecture, with a master controller called *slurmctld* and a slave daemon for each compute node, called *slurmd*. *Slurmctld* is the central coordinator, it receives requests from users that submit jobs, it manages jobs in a priority queue, and it includes scheduling and placement algorithms for launching jobs in compute nodes. *Slurmd* is the node manager, it communicates with *slurmctld* to receive job launch requests, and it manages job's launch and supervises their execution.

BSC SLURM simulator [24] is based on the SLURM code, plus additional modifications that permit to simulate jobs instead of running them. In a previous work [19] we extended and improved the simulator to a stable and more accurate version.

SLURM simulator allowed us to reuse the implemented code into real SLURM, and at the same time, it gave us the possibility to keep using SLURM parameters to have a more precise evaluation, similar to real runs experiments. We developed and integrated into the simulator a runtime model for malleable jobs, as we describe in Section 3.4.

3 THE SLOWDOWN DRIVEN POLICY

This section presents Slowdown Driven policy and its implementation [33] in SLURM Workload Manager.

SD-Policy is a variant of backfill, co-scheduling malleable jobs in non exclusively allocated nodes, where they can efficiently partition the available resources using malleability. It is based on the simple idea that if an arriving job is malleable and not enough resources are available to run it by static backfill algorithm, SD-Policy selects some of the already running jobs, called mates, shrinking them to run the new job, only if predicted system slowdown is improved. Mates, selected by minimizing their increase in slowdown, will be expanded back when the new scheduled job terminates.

We divide the SD-Policy in three parts: *scheduling level*, *resource selection level*, and *node level*. We will start describing the scheduling policy in the controller, from a system level point of view. Then we will present the implementation of the resource selection and placement algorithm, based on the impact on job mates, and feedback from system metrics, used by the scheduler. Finally, we will explain how malleable applications interact with DROM and the node manager, and the implemented logic for shrinking and expanding operations, cores selection and distribution. At the end of the chapter, we present the used runtime models for the scheduling algorithm and the simulator.

3.1 The scheduling algorithm

The scheduling algorithm is a variant of backfill, that considers malleability. It first tries static placement of each job, and if not enough free resources are available, it attempts the flexible scheduling approach for the same job. Malleable backfill runs for each job right after the static trial, and not after the static backfill completed for all jobs. This strategy favors the scheduling of jobs in order of priority. The scheduling algorithm is detailed in Listing 1.

Listing 1: SD-Policy scheduling algorithm

```

schedule (new_job)
  j = new_job
  if (!(nodes = select_nodes(j, free_nodes, null))
      if (!malleable(j))
          return
  else run_job(j, nodes)
  static_end = get_wait_time(j) + j.req_time
  mall_end = j.req_time + runtime_increase(j)
  if (static_end > mall_end)
      s_mates = select_nodes(j, free_nodes, nodes)
      if (s_mates)
          update_stats(j, s_mates)
          s_nodes = get_nodelist(s_mates)
          run_job(j, s_nodes)

```

The scheduler uses an end time prediction, *static_end* and *mall_end*, to estimate if applying malleability would improve the new job slowdown, and, in the affirmative case, it calls the malleable resource selection algorithm. End times are calculated using an estimation of the wait time in the static case, by creating a map of jobs reservations in time to find out when the new job will start. In the malleable case, the job will begin immediately, while runtime will be calculated by adding Equation 6 to the requested time, from runtime models described in Section 3.4.

The scheduler collects the list of selected mates by calling *select_nodes*, using the algorithm described in Section 3.2, and for each mate and the new job, it updates the requested time and the predicted end time. Then it communicates with the node managers, starting procedure described in Section 3.3.

We modified SLURM backfill scheduler, by editing *sched/backfill* plug-in, to take advantage of malleability by implementing Listing 1.

3.2 The resource selection algorithm

The resource selection problem, when co-scheduling multiple jobs with malleability, is reduced to selecting best job mates that will shrink their allocation to make room for the new jobs. Selecting mates is a NP-complete problem, the objective function tries to find the best set of mates with minimum penalty p , i.e. the jobs that receive minimum performance impact when malleability is applied. We describe it with Equations 1-3. Given:

- $x_i \in \{0, 1\}$: mate i is selected
- n number of mates
- p_i penalty estimated for mate
- P maximum penalty for a single mate
- w_i weight of the mate i
- W weight of the scheduled job

we define the *Performance Impact (PI)* with the objective function 1:

$$PI = \min \sum_{i=1}^n x_i * p_i \quad (1)$$

With the constraints 2 and 3:

$$p_i < P, \forall i \in \mathbb{R} \quad (2)$$

$$\sum_{i=1}^n x_i * w_i = W \quad (3)$$

Following on, we define the Performance Impact PI , penalties p and P , weights w and W , and the heuristic used to solves this problem.

3.2.1 Performance Impact. PI is defined as the sum of slowdown increase for each mate when malleability is applied. As Feitelson [15] reports, there is no best metric when we talk about job scheduling evaluation, but it seems slowdown metric helps faster convergence in the preemptive scheduling, similar to our policy. Slowdown, normalizing response time by the runtime, does not give precedence to long jobs like response time, increasing fairness for users submitting short to medium jobs.

3.2.2 Penalties. We calculate the penalty p assigned to each mate based on estimation of the slowdown increase as:

$$p_i = (wait_time + increase + req_time)/req_time \quad (4)$$

Equation 4 considers *wait time*, increase in total run time based on Equation 6, *increase*, and the user *requested time* for the job. Equation 4 is an estimation, as the job duration is usually not equal to the requested time.

The penalty will give precedence to jobs that waited less in the queue and jobs that request a larger amount of time, so the impact in slowdown will be minimum. We define P as *MAX_SLOWDOWN*, a cut-off for p . A cut-off is needed for two main reasons: reducing the eligible mates to reduce the computation, and avoid penalizing jobs that have a high slowdown. We implemented this parameter in two ways:

- (1) *A static value* chosen by system administrators. The value can be chosen empirically or by analyzing the history of a system. In this case, the slowdown must be calculated by using user time estimation, not the real runtime, because this metric is the only one the system can use to predict the slowdown of running and waiting jobs.
- (2) *A dynamic value*: the cut-off is automatically set by the scheduler based on system average slowdown of running jobs, and it is updated every time the controller is not busy in scheduling jobs. The basic idea is to spread the slowdown in a similar way among running jobs, so jobs exceeding the average slowdown are not considered for malleability. Other metrics, like median and 70 percentile were analyzed, but they did not report improvement overall.

3.2.3 Weights. Following constraint 3, we define weight w as the number of allocated computing nodes for the mate, and W as the number of computing nodes requested by the scheduled job. Constraint 3 helps to keep balanced the system in the number of cores

per node jobs use in the allocated nodes, assuming jobs are statically load balanced, which is the common case in HPC environment.

3.2.4 Heuristic. The proposed heuristic tries different combinations of mates iterating recursively on the list of mates for the power of m times, where m is a configurable parameter representing the maximum number of mates. From our evaluation with standard workloads we did not see improvements in system metrics increasing m over two, so we kept it as an optimal value. For each combination that satisfies constraints 2-3, we calculate the Performance Impact and update the best solution, if improving it. Mates list is sort based on the mates' penalty p . If the list is too big, it can be reduced by considering only the first nm mates, with lower p . The algorithm supports contiguous allocations, node filtering by name, architecture, memory and network constraints. Options such as including free nodes to reduce fragmentation and more than two mates per node are supported.

We set a further constraint: new jobs must finish inside mates' allocation to avoid, in the case of multiple mates, that one of them finishes earlier. We avoid this case because the new job would expand to occupy the full nodes of just a part of its allocated space, creating unbalance in the case the application cannot balance its load dynamically. The constraint also avoids creating a further delay for jobs scheduled to run afterward. Listing 2 synthesizes the presented algorithm.

Listing 2: SD-Policy node selection heuristic

```
select_nodes(new_job, nodes, free_nodes)
    if (free_nodes.count > new_job.requested_nodes)
        return s_nodes = pick_nodes(new_job,
                                     free_nodes)
    if (nodes && malleable(newjob))
        mates = get_list_of_mates()
        mates = filter_and_sort(mates, MAX_SLOWDOWN)
        s_mates = pick_mates(newjob, mates, nodes)
    return s_mates
```

We implemented the malleable resource selection algorithm in SLURM *select/linear* plug-in. This plug-in is in charge to select entire nodes for jobs to be scheduled, respecting all job's constraints and optimizing the placement of the nodes using different criteria. We opted for the linear plug-in because malleable jobs can expand and shrink in the node, so there is no need to select individual cores at this point of the scheduling flow. The plug-in allocates entire nodes to give the node manager more freedom in binding specific cores to jobs, as it has a better view of the node usage, and it can get information directly from applications.

3.3 Node management algorithm

Node management is the bottom layer of resource management, and it directly interacts with jobs. In the SD-Policy, node managers are able to select the right amount of computing cores to give to each job, assigning them at launch time, and also controlling their number at runtime.

To achieve described malleability we integrated DROM API into the node manager, to automate the placement of jobs' tasks inside

computing nodes whenever one or more malleable jobs are co-scheduled in overlapping job's allocations. Node managers calculate tasks to cores distribution among jobs and automatically, keeping jobs balanced and isolated. Exceptional cases with nodes running different configurations or jobs asking a different distribution are supported, e.g., using a manual organization to optimize usage of resources for master-slave application architectures, or memory intensive jobs that do not need a high number of computing cores, but instead more memory bandwidth. In the second case, particularly common in the described HPC workloads, a distribution with few computing cores per socket will leave more resources to mates jobs, while fully taking advantage of the memory bandwidth.

We defined the *SharingFactor*, a limit on computational resources that can be taken from a running job in a computational node when shrunk, to implement fairness for mates and to study performance when changing the number of assigned resources. We evaluated static values for this parameter in Marenostum IV [2] and different cores distributions algorithms for the automatic distribution case. Results show that best overall performance is obtained when the applications run isolated in separate sockets. For Marenostum, the number of sockets is two, so the sharing factor is set to 0.5. In a dynamic approach, online performance analysis of running jobs would feed a tuning algorithm for selecting optimal values of *SharingFactor*, further increasing nodes efficiency.

The number of cores assigned to the malleable scheduled job depends on the *SharingFactor* and the minimum amount of cores to which the running jobs can shrink. This last value is equal to the static number of MPI ranks the application is running, to which we assign a minimum of one computing resource per rank.

Listing 3: SD-Policy node management algorithm

```
while (1)
    if (new_job = get_next_job())
        if (malleable(newjob))
            runningjobs.add(new_job)
            distribute_cpu(runningjobs)
            for job in runningjobs
                shrink_job(job)
            DROM_run(newjob)
            run(newjob)
    if (end_job = get_finished_job())
        if (malleable(end_job))
            if (end_job == mate)
                distribute_cpu(runningjobs)
            else
                owner = get_owner(end_job,
                                   runningjobs)
                expand_job(owner)
                DROM_clean(end_job)
```

The implementation is enclosed in the SLURM's *task/affinity* plug-in, in charge of controlling the resources assigned by *slurmctld* to the job's tasks. Task/affinity is dynamically loaded by *slurmd* and *slurmstepd*, dividing the code flow into two parts. In the first part, *Slurmd*, the node manager, is in charge of managing computing resources of a specific node, and thanks to the plug-in, calculating the affinity of tasks, i.e., the distribution onto the computing cores.

Slurmstepd, the job step manager, in the second part, controls the correct task’s launch and execution.

The description of the algorithm presented in Listing 3, and its implementation in SLURM follows:

- (1) At job start or end, the node manager (*slurmd*) interacts with DROM to get information about running tasks. In the case of a starting job, our algorithm recalculates affinities for all the tasks in the node. Cores distribution intelligently maintains running and new processes balanced in the number of cores per task, assuming that, without any other information, the imbalance degrades performance. Cores distribution keeps jobs in separate sockets to improve data locality and reduce interference between jobs. At job’s end, the algorithm returns cores of the ended job to the owner. If any of the already running owner terminates its execution before the new job, its cores will be distributed to remaining running tasks, to increase node utilization.
- (2) Once calculated cores distribution, the job manager (*slurmstepd*) checks if the dependencies created by redistribution of cores are satisfied. In other words, it checks if the jobs where the new job will take cores from are already running or about to start, to assure that DROM assigns cores correctly. Afterward, the job manager launches tasks using the DROM API.
- (3) At job’s launch time, DROM launches and sets the affinity for each new task and attaches them to the DROM space. At end time, it cleans information of tasks from the DROM space. DROM also sets the new affinities for running tasks, so when the tasks reach the next malleability point they can adapt to the change. At end time the API can be set to return cores to the original owners.

3.4 Runtime model for malleable jobs

SD-Policy is based on time estimations, so we implemented a model shaping how applications’ duration is affected when malleability is applied. We already showed that performance impact of DROM applications when changing the number of resources is negligible, so we assume the duration of jobs is proportional to the number of assigned computing resource, and the ability of jobs to adapt to load unbalance.

To estimate the increase in the runtime of a shrunk job, we partitioned the job duration while sharing nodes in T time slots t , each time slot is a different job’s resource configuration. We proportioned the runtime to the number of used resource for each t by putting in relation the static duration with the number of resources. We developed two models: the *ideal* and the *worst case*.

In the *ideal model*, applications do not suffer from the imbalance in the number of resources used by each process. If one job’s task uses one computing node and the other task just half, with this model, the performance will be linear with the number of resources. In this case we compute the increase in the runtime with Equation 5.

$$increase = \sum_{t=1}^T (req_cpus/used_cpus_t * time_t) \quad (5)$$

This model represent applications that can dynamically adapt their load to the resources at run time. tot_cpus is the original number of

cores used in the log, $used_cpus$ the assigned number at each time slot.

In the *worst case model* the same scenario would bring to lower utilization since performance is limited by the less used node n over all nodes N . We compute the increase in the runtime with Equation 6.

$$increase = \sum_{t=1}^T (req_cpus/\min(cpus_per_node_{nt}) * time_t) \quad (6)$$

$\forall n \in \{1, N\}$

The second model represents statically balanced applications. In this case, unbalanced changes in the used resources among nodes generate imbalance. Characterization of applications would led to a more detailed model that would improve simulated experiments precision, but in its absence considering the two models, we are able to give an upper and a lower bound.

We used the described models for time estimations in the SD-Policy and to calculate jobs’ runtime in the SLURM simulator. In the SD-Policy case, we use the *worst case model*, to be able to grant correct jobs execution and completion. In the simulator we try both models, and compare results, in Section 4.3. Running real runs experiments, in Section 4.4 will be the third source of information on how typical HPC applications perform when adapted to be malleable.

4 EVALUATION

Our evaluation is divided into two parts, depending on the used methodology: workloads run in a real environment and workloads simulations. We use the first approach to give a proof of the effective implementation of the policy running in a production system, and its integration with well-known schedulers and programming models, as well as a performance evaluation with benchmarks, neural network spiking simulators coming from Human Brain Project [32], and a computational multi-physics solver. We evaluate this workload on Marenstrum4 (MN4) [2] supercomputer, on up of 49 computing nodes with two sockets equipped with Intel Xeon Platinum 8160 processors, 48 cores, 96GB of main memory per node, for a total 2353 cores. Each workload runs for about two days.

The second methodology, based on whole systems simulations, allowed scaling the evaluation to workloads up to 80640 cores for eight months, allowing the analysis of the SD-Policy performance running on entire HPC systems.

Table 1 presents information about workloads used for the evaluation. All workloads and models come from Feitelson database web page [12]. We generated workloads 1, 2 and 5 with the model developed by Cirne [3], based on the characterization of four different logs. We configured it to use ANL arrival pattern, and we scaled the model to the considered system size. Since we were interested in the performance of our algorithm when the time-based predictions are precise, we generated *Workload 2*, equal in distribution to *workload 1* but with the job’s requested time same to the real duration. *Workload 3* is part of the RICC installation trace from 2010, a realistic workload characterized by a high number of small jobs requesting few nodes, ranging from short to long runtime, up to four days. *Workload 4* is the cleaned version of CEA-Curie log from 2011, only considering the primary partition.

Workload 5 was created from Cirne model, then converted to real applications submissions. We modeled applications behavior and scalability presented in Table 2, then we calculated parameters to adapt the applications to each entry of the workload, regarding the number of requested nodes and duration of the job. We chose a set of applications with different behaviors in the utilization of CPUs and main memory. In the lack of statistics about HPC workloads characterization in the literature, we organized the workload in three main types of applications, equally distributed: compute-bounded jobs with low memory utilization (PILS), memory-bounded jobs with lower CPU utilization (STREAM), big simulations, memory and compute intensive (CoreNeuron, NEST, Alya).

The following section presents different evaluations:

- Evaluation of MAX_SLOWDOWN
- Analysis of simulated workload 4
- Evaluation of runtime models
- Analysis of Workload 5 in a real environment

We used the following metrics for the evaluation:

- Makespan: the difference between the last job end time and the first job arrival time.
- Average response time: the average of jobs' response time, calculated as the difference between jobs' end time and submission time.
- Average slowdown: the average of jobs' slowdown, calculated as the response time divided by the static execution time of the job.
- Energy consumption: The energy consumed to run entire workloads, reported by system software.

4.1 Evaluation of MAX_SLOWDOWN

Different values of MAX_SLOWDOWN can have a high impact on the performance of SD-Policy. Low values will limit the number of times malleability can be applied, and high values could degrade jobs and system performance. We simulated workload 1, 2, 3 and 4 using SharingFactor of 0.5 and the ideal runtime model. We tried different values for MAX_SLOWDOWN represented by the following labels: MAXSD 5, MAXSD 10, MAXSD 50, MAXSD infinite, and the dynamic cut-off based on feedback from running jobs slowdown, DynAVGSD.

Figures 1, 2, and 3 show the makespan, average response time and average slowdown for the workloads, normalized to static backfill. In the first three workloads, increasing slowdown limit improves the system's slowdown, showing that not filtering mates still improves the system performance overall. We observe a reduction in slowdown up to 49.5%, 31%, 25.7%, and 70.4% respectively for Workload 1, 2, 3, and 4.

On one side, even with a high limit for MAX_SLOWDOWN, our algorithm tries to avoid applying malleability when it would not bring improvements, avoiding performance loss also when increasing the parameter to an infinite value. On the other side, a system administrator could still consider using a relatively low limit, to avoid some jobs to be penalized too much, or implement different queues with different QoS policies using different MAXSD configurations.

In Workload 2 slowdown does not monotonically decrease with the increase of the limit, showing an increase at value 50, but still

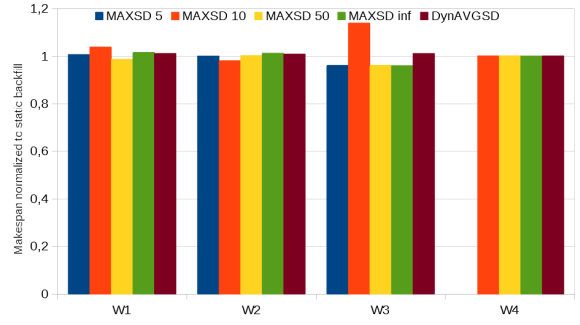


Figure 1: Makespan for workload 1 to 4 changing the MAX_SLOWDOWN parameter, normalized to static backfill simulation.

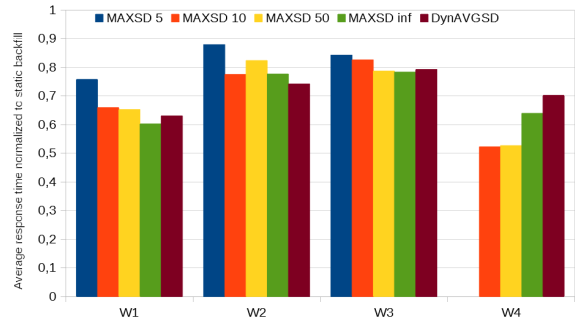


Figure 2: Average response time for workload 1 to 4 changing the MAX_SLOWDOWN parameter, normalized to static backfill simulation.

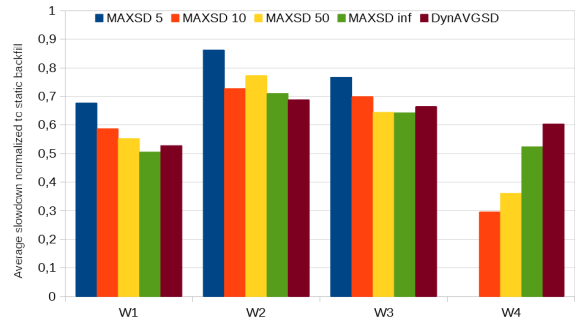


Figure 3: Average slowdown for workload 1 to 4 changing the MAX_SLOWDOWN parameter, normalized to static backfill simulation.

outperforming static backfill. SD-Policy DynAVGSD brings further improvements in the same workload, where the SD-Policy works with real jobs durations and not the user requested time. The explanation of the observed behaviors resides in the fact that variance in the real average slowdown is much higher than in the predicted average slowdown when using user requested time, so a dynamic value of MAX_SLOWDOWN benefits this evaluation. Also, using real predicted metrics allows the SD-Policy to be more precise. This observation suggests that using a predictive method for job's runtime, i.e. based on machine learning, rather than asking the user,

Table 1: Description of workloads

ID	Log/model	# jobs	System (nodes/cores)	max job (nodes/cores)	Avg resp. time (s)	Avg slowdown	Makespan (s)
1	Cirne	5000	1024/49152	128/6144	122152	3339,5	899888
2	Cirne_ideal	5000	1024/49152	128/6144	126486	3501	896024
3	RICC-sept	10000	1024/8192	72/576	43537	1341	407043
4	CEA-Curie	198509	5040/80640	4988/79808	29858,5	3666,5	21615111
5	Cirne_real_run	2000	49/2352	16/768	56482	4783,1	159313

Table 2: Workload characterization for real runs evaluation

Application	% workload	ReqNodes	ReqTime	CPU utilization	Memory utilization
PILS [16]	30.5%	small to high	small/med	high	low
STREAM [27]	30.8%	small to high	small/med	low	high
CoreNeuron [21]	35,5%	small to high	small to high	high	med
NEST [22]	2.6%	small to high	small to high	high	med
Alya [10]	0.6%	small	high	high	med

will improve the performance of our policy. Comparing workload 1 and 2, the static backfill for the first workload outperforms the static backfill for the second by 8.6%, showing the interesting result that precision of job’s duration does not always produce better average system slowdown. On the other side, having an exact job duration allows jobs not to be delayed in their start time, making backfill correct. Backfill behavior influences our approach, also base algorithm of SD-Policy, with Workload 1 performing 29.6% better than Workload 2.

In Workload 4, the best value is obtained in the MAXSD 10 case, while higher values and the dynamic version do not improve results. For this significant workload, we have the maximum observed reduction of the response time and the slowdown up to respectively 50% and 70%, showing the high benefits a system can have over a static backfill approach.

4.2 Analysis of a big workload

We simulated workload 4, big in length and number of nodes, and analyzed details by partitioning the jobs in categories depending on the requested resources and runtime. We compared the static backfill version with the MAXSD 10 version. The SD-Policy improved slowdown by 70.4% while keeping makespan constant. We compared average slowdown, average runtime and average wait time for static backfill and the SD-Policy, presenting the ratios between the two versions in Figures 4, 5 and 6. The malleable version highly improves the slowdown of jobs consuming up to 4 hours and asking up to 512 nodes, up to 569%. Relatively small and short jobs have very high slowdown, compared to larger and longer jobs. Those are the primary jobs that benefit from the SD-Policy.

The rest of the heatmap of the slowdown, in Figure 4 keeps having better values even for bigger and larger jobs, except for three categories. Two categories contain few jobs to take some conclusions, but the 121 jobs asking 512 to 1024 nodes with duration in 12 hours to 1 day range, show an increase of 15% in the average slowdown. Even if it seems this job category is penalized, the average slowdown for those jobs was lower than other neighbors in the heatmap, so result shows that the SD-Policy generates a more fair

distribution of the slowdown with respect to the static backfill. The rest of the jobs, even if affected in their runtime (Figure 5), because of the malleability, improve their wait time (Figure 6), showing fairness is not an issue for some particular category of jobs.

Figure 7 shows the trend of the average slowdown per day, comparing the static backfill and the SD-Policy per day, together with the number of jobs scheduled with malleability. It is visible that the peaks in slowdown are highly reduced all over the simulated time, and, apart from an initial spike of jobs that are selected as mates, the average slowdown for our policy never increases over the static. reduction of peaks in slowdown are usually associated with an high number of jobs scheduled with malleability. The total number of malleable scheduled jobs is 20476, the number of mates is 17102, corresponding to 10.3% and 8.6% of the workload.

4.3 Evaluation of runtime models

When one of the mates completes before the user requested time, it could lead the other job running in the same node to take the freed cores, but only for part of its job allocation, creating unbalance in its load. We run simulations with the two runtime models presented in Section 3.4 for workloads 1 to 4, using SD-Policy DynAVGSD, to estimate a lower and upper bounds for this overhead.

Figure 8 shows the impact in makespan, average response time and average slowdown for the two models. The worst case model increases response time, up to 11% for workload 1 with respect to the ideal model, negligible for workload 3 and 4, less than 1.5%. Average slowdown, similarly to average response time, increase by 16% in Workload 1, only 3.5% and 1% in Workloads 3 and 4, while still outperforming static backfill. Makespan increases by 9% in Workload 3, while less than 1% in the other cases. Workload 2 is not affected by using the worst case model, as the jobs requested times are exact and it allows the SD-Policy to avoid creating unbalances.

Using a predictive model that lower the gap between requested time and real runtime would reduce the evaluated overhead to 0, as the time predictions would be more precise.

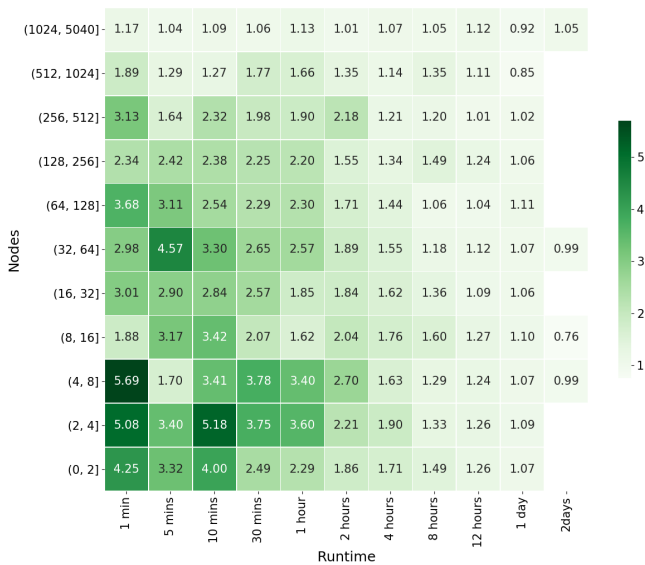


Figure 4: Heatmap showing the ratio between slowdown of static backfill and SD-Policy MAXSD 10 using workload 4, for different job categories.



Figure 5: Heatmap showing the ratio between run time of static backfill and SD-Policy MAXSD 10 using workload 4, for different job categories.

4.4 Analysis of a real run

Having a proposed policy implemented with evaluations in a real system is a complex task, many publications skip this part, only basing the evaluation on simulation results. This is particularly true in the case of malleability, where all the layers of HPC software stack need to be aware of the directly connected layers.

On one hand, simulations are essentials, as they allow evaluating a scheduling algorithm in a big system with large workloads, but they do not have the same precision and complexity of a real system.

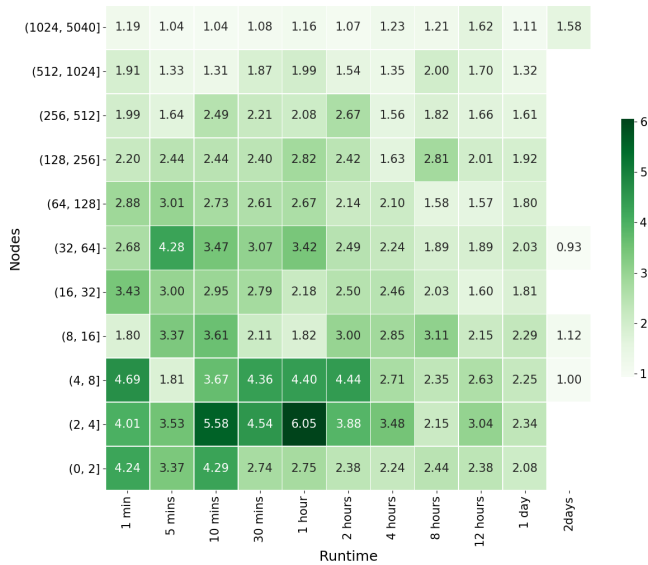


Figure 6: Heatmap showing the ratio between wait time of static backfill and SD-Policy MAXSD 10 using workload 4, for different job categories.

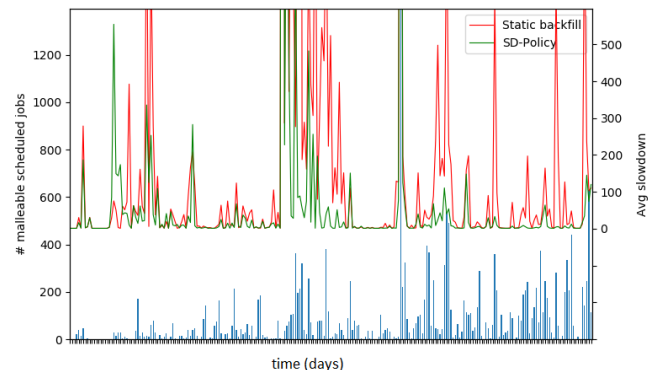


Figure 7: Columns represent the number of jobs scheduled with malleability per day, their y axis is on the left. The two lines represent slowdown per day of static backfill and SD-Policy MAXSD 10.

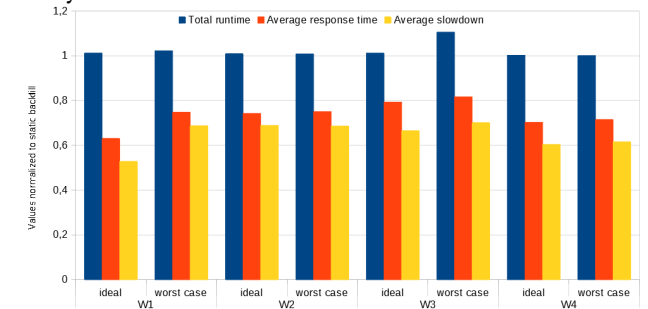


Figure 8: Makespan, average response time and slowdown for workloads 1-4, running SD-Policy with feedback, using ideal and worst case model. x axis represent the workloads, metrics are normalized to static backfill simulations.

It is difficult, for instance, evaluating the energy consumption of the SD-Policy and compare it with a real system, as well as modeling a real runtime model for different malleable jobs. We put effort in developing a working SLURM version, and we made its code available [33], and we used it to evaluate Workload 5, based on the Cirne model using ANL arrival pattern, with 2000 jobs. Jobs ask a maximum of 16 nodes, 768 cores per job, on a system of 49 nodes, 2352 cores. We converted Cirne log to a real job submission list for SLURM by using a set of malleable applications. We selected parameters to respect runtime and requested resources, and we generated, using a script, a list of submissions, respecting arrival time. Table 2 shows the list of used applications, the type and the percentage of jobs running them.

We run the workload in MN4 supercomputer by spawning a SLURM instance inside a job of the SLURM installation in MN4, to have exclusive access to a subset of 50 nodes of the machine to run the workload. The number of nodes and the makespan of this evaluation are constrained by system queue limits, 50 nodes for 48 hours, so we adapted Cirne model to it. The controller uses the first node, while the other nodes are the computing nodes.

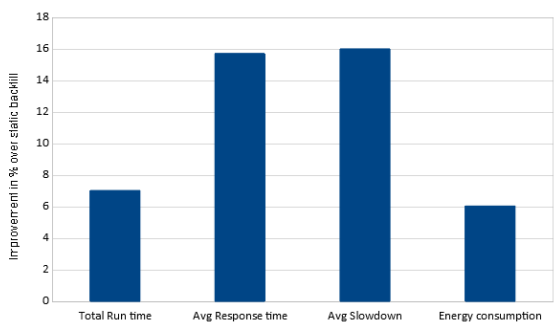


Figure 9: Improvement in percentage of the SD-Policy over the static backfill for makespan, average response time, average slowdown and energy consumption.

Results in Figure 9 show that makespan decreases by 7%, average response time and average slowdown by about 16%, compared to the static backfill run. Results show less gain over the simulations because in this evaluation the considered number of nodes is lower. Shorter makespan is obtained by better utilization of nodes’ resources, as 449 jobs out of 539 scheduled with malleability have a better runtime compared to the static execution, if we proportionate it to the number of used resources. This behavior is due two main reasons, already found in a previous research [5]. The first reason is the better use of resources by memory bounded jobs, or jobs with memory bounded phases, i.e. during application’s initialisation and finalization: in this situation, running computing-bounded jobs takes advantage of cores underutilized. The second reason is related to scalability problems of applications, that cannot perfectly scale to the high number of cores in the nodes: in this case, using malleability to partition available cores improves performance. As a consequence of better resources utilization, we save 6% of energy over the static approach, very important result considering the increasing energy consumption of HPC systems.

5 RELATED WORK

Malleability has been studied since many years ago in the context of scheduling, with the definition and study of Malleable Parallel Task Scheduling (MPTS) problem. The theoretical research shows its potential benefits [25] [8] [29]. These works are related to off-line schedulers, that select the number of resources that best improves the performance of the parallel task based on a model of its performance given at scheduling time, during the allotment. Once the task starts its execution, the number of its resources does not vary.

Today, the problem has been furtherer studied, Feitelson [14] classifies a malleable job as a job that can adapt to changes in the number of processors at run time. Following this definition, job scheduling simulations [18] showed the potential benefits of malleability on job’s response time, by using fair process distribution and shrink and expand operation on jobs that expand the number of used processors. Several studies have been done to reduce programming complexity when developing malleable applications. Utrera [35] uses folding techniques and a FCFS-malleable policy that uses co-scheduling and oversubscription to start MPI jobs when not enough resources are available. Kale [20] implement malleable and evolving jobs on the top of Charm++, and defined a scheduling policy based on equipartition. Prabhakaran [11], using Charm++ malleability, implemented shrink/expand operations in a production scheduler, together with a scheduling strategy based on equipartition and combining rigid, evolving and malleable jobs. Martin [26] introduces FLEX-MPI library for dynamic reconfiguration of MPI applications based on checkpoint and restart, while Comprés [4] implement MPI malleability with on-line data redistribution, plus shrink and expand operations in SLURM. In general, data redistribution in malleable MPI implies overhead of data movement and effort for developers, not in line with our research, being effortless and efficient. Cera [9] implements malleability based on dynamic CPUSets using MPI and a production resource manager. This approach is similar to how we use malleability, but in our case, we do not oversubscribe MPI processes because we demonstrated it could degrade application’s performance [23] and we integrate with shared memory programming models for better performance. While supporting MPI for multi-node applications, our approach uses DROM interface [5], that allows malleability in computational nodes by changing the number of threads OpenMP [30] or OmpSs [6] applications are using. This approach enables effortless, dynamic and zero overhead moldability and malleability in the number of cores application uses per node, and allows the scheduler to take decisions in real time with almost instantaneous adaptation from applications.

Many malleability approaches were studied in the context of the grid, taking advantage of application’s feedback, like [34] [36] [7]. Those approaches use application’s performance models and run time performance measurements. Our algorithm differs, because it uses feedback from the scheduler itself, scheduling at a higher level of abstraction, based on system metrics, e.g., average system slowdown rather than application’s feedback.

6 CONCLUSIONS AND FUTURE WORK

This paper presented a novel scheduling policy based on malleability and slowdown minimization, the SD-Policy. We described the

algorithm and its implementation into a resource manager, enabling an integrating of the whole HPC software stack. We presented different parametric evaluations, together with two more complete cases of study, a big simulation and a real run into a supercomputer. We showed that SD-Policy can reduce the response time and the slowdown up to 50% and 70% for the CEA Curie log, while we saw an improvement of nodes utilization that brought 7% makespan and 6% energy reduction for a real run.

Future work will focus on three main points:

- (1) Integration of online performance metrics collection: by having information about applications we will be able to perform a better placement of malleable jobs, and dynamic adapting SharingFactor.
- (2) Improved scheduler: feedback mechanism, heuristic and runtime estimation, based on machine learning approaches.
- (3) Malleability at MPI level, unlocking new possibilities for the scheduler, that can shrink or expand jobs allocations in the number of nodes.

ACKNOWLEDGMENTS

This work is partially supported by the Spanish Government through Programa Severo Ochoa (SEV-2015-0493), by the Spanish Ministry of Science and Technology through TIN2015-65316-P project, by the Generalitat de Catalunya (2017-SGR-1414) and from the European Union's Horizon 2020 under grant agreement No. 720270 (Human Brain Project SGA1) and No. 785907 (Human Brain Project SGA2).

REFERENCES

- [1] Yoo A. B., Jette M. A., and Grondon M. 2003. SLURM: Simple Linux Utility for Resource Management. In *Job Scheduling Strategies for Parallel Processing*. 44–60.
- [2] Barcelona Supercomputing Center. 2017. Marenostrom 4. <https://www.bsc.es/marenostrom/marenostrom>
- [3] W. Cirne and F. Berman. 2001. A comprehensive model of the supercomputer workload. In *Proceedings of the Fourth Annual IEEE International Workshop on Workload Characterization*. <https://doi.org/10.1109/WWC.2001.990753>
- [4] I. Comprés, A. Mo-Hellenbrand, M. Gerndt, and H. Bungartz. 2016. Infrastructure and API Extensions for Elastic Execution of MPI Applications. In *Proceedings of the 23rd European MPI Users' Group Meeting (EuroMPI 2016)*. 82–97.
- [5] Marco D'Amico, Marta Garcia-Gasulla, Víctor López, Ana Jokanovic, Raül Sirvent, and Julita Corbalan. 2018. DRoM: Enabling Efficient and Effortless Malleability for Resource Managers. In *Proceedings of the 47th International Conference on Parallel Processing Companion (ICPP '18)*. ACM, New York, NY, USA, Article 41, 10 pages. <https://doi.org/10.1145/3229710.3229752>
- [6] A. Duran, E. Ayguadé, R. M. Badia, J. Labarta, L. Martinell, X. Martorell, and J. Planas. 2011. Omppss: a Proposal for Programming Heterogeneous Multi-Core Architectures. *Parallel Processing Letters* 21 (2011).
- [7] F. Berman et al. 2003. Adaptive computing on the Grid using AppLeS. *IEEE Transactions on Parallel and Distributed Systems* 14, 4 (April 2003), 369–382. <https://doi.org/10.1109/TPDS.2003.1195409>
- [8] J. Turek et al. 1994. Scheduling Parallelizable Tasks to Minimize Average Response Time. In *Proceedings of the Sixth Annual ACM Symposium on Parallel Algorithms and Architectures*. 200–209.
- [9] M. Cera et al. 2010. Supporting Malleability in Parallel Architectures with Dynamic CPUSETs Mapping and Dynamic MPI. In *Proceedings of the 11th International Conference on Distributed Computing and Networking*. 242–257.
- [10] Mariano Vázquez et al. 2016. Alya: Multiphysics engineering simulation toward exascale. *Journal of Computational Science* 14 (2016), 15 – 27. <https://doi.org/10.1016/j.jocs.2015.12.007>
- [11] S. Prabhakaran et al. 2015. A Batch System with Efficient Adaptive Scheduling for Malleable and Evolving Applications. In *2015 IEEE International Parallel and Distributed Processing Symposium*. 429–438.
- [12] Dror G. Feitelson. Dec 8, 2005. *Logs of Real Parallel Workloads from Production Systems*. <http://www.cs.huji.ac.il/labs/parallel/workload/>
- [13] Dror G. Feitelson. Feb 21, 2006. *The Standard Workload Format*. <http://www.cs.huji.ac.il/labs/parallel/workload/swf.html>
- [14] Dror G. Feitelson and Larry Rudolph. 1996. Toward convergence in job schedulers for parallel supercomputers. In *Job Scheduling Strategies for Parallel Processing*. Berlin, Heidelberg, 1–26.
- [15] Dror G. Feitelson and Larry Rudolph. 1998. Metrics and benchmarking for parallel job scheduling. In *Job Scheduling Strategies for Parallel Processing*, Dror G. Feitelson and Larry Rudolph (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1–24.
- [16] M. Garcia, J. Labarta, and J. Corbalan. 2014. Hints to improve automatic load balancing with LeWI for hybrid applications. *J. Parallel and Distrib. Comput.* (2014).
- [17] B.C. Gorda and E.D. III. Brooks. 1991. Gang scheduling a parallel machine. (1991).
- [18] J. Hungershofer. 2004. On the combined scheduling of malleable and rigid jobs. In *16th Symposium on Computer Architecture and High Performance Computing*.
- [19] Ana Jokanovic, Marco D'Amico, and Julita Corbalan. 2018. Evaluating SLURM simulator with real-machine SLURM and vice versa.
- [20] L. V. Kale, S. Kumar, and J. DeSouza. 2002. A Malleable-Job System for Timeshared Parallel Machines. In *2nd IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGRID'02)*. <https://doi.org/10.1109/CCGRID.2002.1017131>
- [21] P. Kumbhar and M. Hines. 2016. CoreNeuron Neuronal Network Simulator Optimization Opportunities and Early Experience. In *GPU Technology Conference*.
- [22] Kunkel, S. et al. 2017. NEST 2.12.0. <https://doi.org/10.5281/zenodo.259534>
- [23] V. Lopez, A. Jokanovic, M. D'Amico, M. Garcia, R. Sirvent, and J. Corbalan. 2017. DJSB: Dynamic Job Scheduling Benchmark. In *Job Scheduling Strategies for Parallel Processing: 21st International Workshop, JSSPP 2017, Revised Selected Papers*. Springer.
- [24] A. Lucero. 2011. Simulation of batch scheduling using real production-ready software tools. In *Proceedings of the 5th IBERGRID*.
- [25] Walter Ludwig and Prason Tiwari. 1994. Scheduling Malleable and Nonmalleable Parallel Tasks.. In *SODA*, Vol. 94. 167–176.
- [26] Gonzalo Martín, Maria-Cristina Marinescu, David E. Singh, and Jesús Carretero. 2013. FLEX-MPI: An MPI Extension for Supporting Dynamic Load Balancing on Heterogeneous Non-dedicated Systems. In *Euro-Par 2013 Parallel Processing*. Springer Berlin Heidelberg, Berlin, Heidelberg, 138–149.
- [27] John D. McCalpin. 1995. Memory Bandwidth and Machine Balance in Current High Performance Computers. *Technical Committee on Computer Architecture* (1995).
- [28] Message Passing Interface Forum. 2015. MPI Specifications 3.1, <http://mpi-forum.org/docs/mpi-3.1/mpi31-report.pdf>.
- [29] G. Mounie, C. Rapine, and D. Trystram. 1999. Efficient Approximation Algorithms for Scheduling Malleable Tasks. In *Proceedings of the Eleventh Annual ACM Symposium on Parallel Algorithms and Architectures (SPAA '99)*. 23–32.
- [30] OpenMP. 27/4/2018. OpenMP 4.5 Specifications, <http://www.openmp.org/wp-content/uploads/openmp-4.5.pdf>.
- [31] J. Ousterhout. 1982. Scheduling Techniques for Concurrent Systems. (1982).
- [32] Human Brain Project. 2017. <https://www.humanbrainproject.eu/en/>.
- [33] "SD-Policy source code". 2018. https://github.com/marcodamico/slurm_sd-policy.
- [34] H. Sun, Y. Cao, and W. Hsu. 2011. Fair and Efficient Online Adaptive Scheduling for Multiple Sets of Parallel Applications. In *2011 IEEE 17th International Conference on Parallel and Distributed Systems*. <https://doi.org/10.1109/ICPADS.2011.62>
- [35] Gladys Utrera, Siham Tabik, Julita Corbalan, and Jesús Labarta. 2012. A Job Scheduling Approach for Multi-core Clusters Based on Virtual Malleability. In *Euro-Par 2012 Parallel Processing*. 191–203.
- [36] Sathish S. Vadhiyar and Jack J. Dongarra. 2005. Self Adaptivity in Grid Computing: Research Articles. *Concurr. Comput. : Pract. Exper.* 17, 2–4 (Feb. 2005), 235–257. <https://doi.org/10.1002/cpe.v17:2/4>