Obtaining Dynamic Scheduling Policies with Simulation and Machine Learning

Danilo Carastan-Santos
Univ. Grenoble Alpes, CNRS, Inria, LIG
Grenoble, France
Universidade Federal do ABC
Santo André, Brazil
danilo.santos@ufabc.edu.br

Raphael Y. de Camargo
Universidade Federal do ABC
Santo André, Brazil
raphael.camargo@ufabc.edu.br

ABSTRACT

Dynamic scheduling of tasks in large-scale HPC platforms is normally accomplished using ad-hoc heuristics, based on task characteristics, combined with some backfilling strategy. Defining heuristics that work efficiently in different scenarios is a difficult task, specially when considering the large variety of task types and platform architectures. In this work, we present a methodology based on simulation and machine learning to obtain dynamic scheduling policies. Using simulations and a workload generation model, we can determine the characteristics of tasks that lead to a reduction in the mean slowdown of tasks in an execution queue. Modeling these characteristics using a nonlinear function and applying this function to select the next task to execute in a queue improved the mean task slowdown in synthetic workloads. When applied to real workload traces from highly different machines, these functions still resulted in performance improvements, attesting the generalization capability of the obtained heuristics.

CCS CONCEPTS

• **Computing methodologies** → **Parallel computing methodologies**; **Machine learning approaches**;

KEYWORDS

Scheduling, High Performance Computing, Simulation, Machine Learning

ACM Reference Format:
Danilo Carastan-Santos and Raphael Y. de Camargo. 2017. Obtaining Dynamic Scheduling Policies with Simulation and Machine Learning. In *Proceedings of SC17*. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3126908.3126955

1 INTRODUCTION

The on-line scheduling of tasks in large-scale HPC platforms is a complicate subject to be tackled by scheduling systems. The need to address numerous scheduling factors leads to the development of sophisticated scheduling algorithms that are often difficult to reason about or hard to deploy in real systems. In this light, an appealing alternative is to use scheduling policies, which are functions that take as input the characteristics of the tasks (*e.g.* processing time, requested number of processors, waiting time, *etc.*) and output a value that denotes the priority of the task. These scheduling policies are often designed in an *ad-hoc* manner, generally based on intuitions regarding which rules the scheduling policies must impose to achieve good scheduling performance. One common practice of the scheduler systems is to add a backfilling mechanism in conjunction to the scheduling policy. The backfilling increases the utilization of the HPC platform and consistently improves scheduling performance.

Another common practice of HPC platform maintainers is to register information about the tasks that have been executed in the platform. These workload logs contain information regarding characteristics of tasks, such as the ones mentioned above and processing time estimates provided by users. In light of the ever increasing amount of information generated by HPC platforms and the need for simple and efficient scheduling solutions, the main question raised by this work is: *Is it possible to design a simple procedure, that employs simulation and machine learning techniques, to extract general and simple scheduling policies from existing workload logs, which perform better than the existing ad-hoc scheduling policies?*

In this work, we present a technique based on simulation and machine learning algorithms to generate simple scheduling policies represented by nonlinear functions. These nonlinear functions capture the characteristics of tasks that should be prioritized under several distinct situations and, when used as scheduling policies, improve the global scheduling of tasks. More specifically, this work presents the following contributions:

(1) We show that it is possible to generate efficient scheduling policies in the form of nonlinear functions, obtained from general workload characteristics, that improve global scheduling, when compared to classical and state-of-the-art *ad-hoc* scheduling policies;
(2) We propose a simple simulation procedure and a machine learning strategy, based on nonlinear regression, to observe the effects of scheduling decisions over tasks obtained from a workload model over distinct conditions, and to model these effects into nonlinear functions that can be used as on-line scheduling policies;
(3) We show that these obtained scheduling policies perform well when scheduling tasks from the same workload model

SC17, November 12–17, 2017, Denver, CO, USA

ACM ISBN 978-1-4503-5114-0/17/11…$15.00
https://doi.org/10.1145/3126908.3126955

used to observe the scheduling effects, with performances up to 11.8 times better than the best performing *ad-hoc* scheduling performance in the most realistic setting evaluated (*i.e.* in conjunction with a backfilling algorithm and considering the user estimates of task processing times to perform scheduling decisions);

(4) We show that the scheduling policies obtained by the procedure of the item 2 can generalize well, bringing good scheduling performances in all evaluated real world scenarios, using real workload traces from highly different HPC platform configurations.

The remainder of this paper is organized as follows: in Section 2 we present related work and in Section 3 we present the proposed strategy to obtain on-line scheduling policies. We present the main results (*i.e.* the obtained scheduling policies and its scheduling performances) in Section 4, and the conclusions in Section 5.

2 RELATED WORK

Several researchers attempted to tackle the problem of scheduling with a wide range of approaches, covering from integer linear programming [3, 12] to genetic algorithms [15, 25] and neural networks [1, 2]. Xhafa and Abraham [30] present a review of recent computational models and heuristics for scheduling in HPC platforms. While these works show improvements in scheduling performance, one aspect that these works share in common is that they are either computationally complex or hard to be deployed in real scenarios.

With regard to simpler scheduling policies, most of the policies are defined by *ad-hoc* intuitions about how the scheduling should behave in order to achieve good performance, from the classical policies with simple reasoning such as First-Come, First-Served (FCFS), Shortest Processing Time First (SPT) [24] and Longest Processing Time (LPT) [24], to the smarter and more complex reasoning policies such as WFP3 [28] and UNICEF [28].

One aspect of these *ad-hoc* scheduling policies is that they perform well only for some workload characteristics and HPC platform configurations. The main difference from our work is that we use observations over scheduling decision results, using simulations with several distinct workloads, to determine what are the best combinations of task characteristics that result in improvements over the average bounded slowdown of all tasks in the queue. The objective is to obtain policies that generalize better over different workloads and HPC platforms.

A noticeable phenomenon is the divergence between theory and practice in the problem of scheduling tasks in HPC platforms [10]. With the introduction of backfilling algorithms [19] – whose reasoning is to allow a task with low priority to be executed before a higher priority task if this low priority task does not delay the higher priority one – in practice most recent systems responsible for scheduling tasks in HPC platforms implement either FCFS with aggressive backfilling (called EASY [19] algorithm) or the same aggressive backfilling with some scheduling policy to sort the waiting tasks. For example, SLURM job management system, a well known scheduling system for HPC platforms [31], performs a multi-factor sorting of the waiting queue, using a linear combination of priority factors (tasks waiting time, tasks size, share factors, *etc.*), with coefficients – whose values establish the relative importance of these priority factors – defined by the HPC platform maintainer, and then performs backfilling over the sorted queue. Other job management systems such as TORQUE [26], PBSpro [20] or MOAB [6] implement similar approaches with their own specificities. Georgiou [14] presents a detailed review of these job management systems. The common point of these schedulers is that they use simple heuristics for scheduling, that work reasonably well in a variety of situations. Our work targets this kind of system, with the differential that, instead of *ad-hoc* heuristics, we define a procedural way to find simple and efficient scheduling policies for the type of tasks normally submitted on these systems.

A considerable amount of research was devoted to improve the performance of backfilling algorithms by performing tuning of its queues [17, 23, 27]. With regard to machine learning, most of the works are focused in predicting either performance [7] or communication cost [22]. More recently, Gaussier *et al.* [13] explored the usage of machine learning in scheduling and proposed a variant of the EASY algorithm that relies in a task runtime prediction model to perform the scheduling decisions. In our work we use machine learning to infer the scheduling policies to use, instead of predicting task behavior.

3 FINDING SCHEDULING POLICIES WITH SIMULATION AND MACHINE LEARNING

We tackled the on-line scheduling problem of executing a set of concurrent parallel tasks – whose resource requirements are known in advance (also known as rigid tasks) – on a HPC platform. The general idea proposed by this work is to design a simulation scheme to observe the scheduling behavior over sets of tasks under several distinct conditions and to use machine learning techniques to model the effects of task characteristics on scheduling performance into nonlinear functions. These functions can then be used by production on-line schedulers to determine – based on tasks characteristics, such as estimated processing time, resource requirements and submit time – the next task to choose from the queue for execution.

3.1 Scheduling Background

We consider an HPC platform as constituted by a set of n_{max} homogeneous resources connected by any interconnection topology and the tasks arrive over time (*i.e.* in an on-line manner) in a centralized waiting queue. A *task* t is some workload which has the following data:

- The estimated processing time $\tilde{p_t}$ of the task informed by the user;
- The actual processing time p_t of the task (only known after the task has been executed);
- The resource requirement of the task, measured as the number of processors q_t;
- The submit time r_t of the task (also called release date).

Although some data sets have additional information, the selected variables are available in most real workload traces, shared using the Standard Workload Format (SWF) [11].

There are many objective functions that can be considered in the on-line scheduling problem. In this regard, one reasonable expectation is that the waiting time of a task should be proportional to its processing time [9]. Hence, one can use the *bounded slowdown* objective function which is defined as follows for a task t:

$$bsld = \max\left(\frac{w_t + p_t}{\max(p_t, \tau)}, 1\right) \quad (1)$$

where w_t is the time that task t waited for execution (*i.e.* the time that t starts its execution minus the submit time r_t) and τ is a constant, with a typical value of 10s, that prevents small tasks from having excessively large slowdown values. Similarly, we can define the *average bounded slowdown* which is the slowdown average over a sequence of tasks T:

$$AVEbsld(T) = \frac{1}{|T|}\sum_{t\in T}\max\left(\frac{w_t + p_t}{\max(p_t, \tau)}, 1\right) \quad (2)$$

In this work, all scheduling evaluations are performed using this objective function. However, our proposed scheduling methodology can be applied with other objective functions, thus giving freedom for the HPC platform maintainer to choose the objective function that best suits the QoS requirements of his users.

3.2 Simulation Scheme

We considered for simulation an HPC platform represented by an homogeneous cluster compounded by n_{max} resources (processors). We defined two sets of tasks to be executed, S and Q. Set S contains $|S|$ tasks that are executed in order of arrival at the beginning of the simulation and it acts as a warm-up workload. Tasks from Q, which are used to extract information about scheduling performance, start to arrive after all tasks from S arrived. This scheme provides a way to represent an initial resource state of the cluster before the arrival of tasks from Q. From both sets, the tasks characteristics are obtained from a trace generated with the Lublin and Feitelson [18] workload model. This model is based on tasks characteristics of several real HPC platforms and it also models the release date of tasks, including peak periods. Another advantage of the workload model is that we can generate arbitrarily large traces. The task sets S and Q are generated in the following way: from a large enough trace N generated by the workload model, we randomly select a subtrace $M \subset N$ with size $|S| + |Q|$. The first $|S|$ tasks from M will belong to set S and the remaining $|Q|$ tasks from M will belong to set Q.

The first step is to determine, using simulations, how the scheduling performance is affected when a task t is selected for execution, under different sets of tasks and resource states. To obtain this information, several tuples of task sets (S, Q) were generated, where each tuple (S, Q) is a randomly generated trace as aforementioned. The key idea is to list as many scheduling situations – represented by tuples (S, Q) – as possible. For each tuple (S, Q), we define $\mathcal{P}$ as a collection of random permutations of Q, and $\mathcal{P}(t_0 = t)$ as the subset from all permutations where t is the first task in the permutation. We then simulate the scheduling execution for each pair (S, per) for all $per \in \mathcal{P}$. We call these pairs (S, per) as *trials* of the tuple (S, Q). On each trial, each task $t \in Q$ is submitted for execution in the order as they appear in per. We then assign a *score* for each task $t \in Q$:

$$score(t) = \frac{\sum_{per_j\in\mathcal{P}(t_0=t)} AVEbsld(per_j)}{\sum_{per_k\in\mathcal{P}} AVEbsld(per_k)} \quad (3)$$

The score denotes the impact of assigning a task $t \in Q$ to execute first, in the average bounded slowdown of all tasks in Q. The set of scores for all tasks $t \in Q$ constitute a *trial score distribution* of the tasks of Q under initial resource state S. Typical distributions, shown in Figure 1, contains most scores slightly above or below the mean $\frac{1}{|Q|} = \frac{1}{32} = 0.031$. Tasks with lower scores have a more positive impact in the average bounded slowdown when they are chosen to be executed first. It is important to note that, as the total number of possible trials grows quickly in function of $|Q|$, the size of the set Q must be chosen in a way that it is possible to perform an accurate estimation of the tasks scores, while maintaining a feasible computational cost. More information about this accuracy and computational cost can be found in Section 4.1.

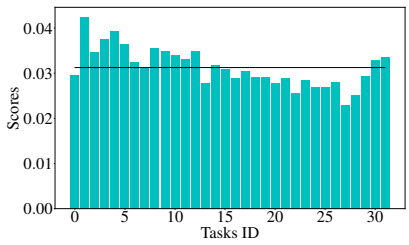
By joining the generated samples from multiple distinct tuples (S, Q), we generate a distribution $score(p, q, r)$, containing the sample means for tasks with processing time p, number of used processors q and arrival time r. This is the central result obtained from the simulations. The idea is that a scheduler from a HPC system will select the task from the queue with (p, q, r) values that has the smallest $score(p, q, r)$ value.

3.3 Machine Learning Scheme

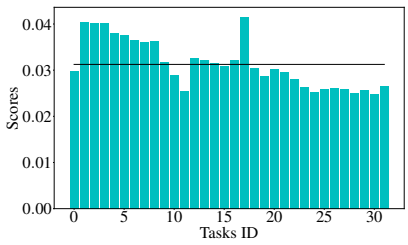
Using simulations we generate irregular $score(p, q, r)$ distributions, with values that can change each time a simulation is performed and that provide good estimates only for certain tasks characteristics. To obtain smother and more general representations of the score distributions, we can use a machine learning technique, called nonlinear regression, to determine a nonlinear function $f(p, q, r)$ that provides a good fitting to the distribution. This function can then be later assigned as a scheduling policy. In other words, the tasks arriving into a centralized queue of an HPC system can be prioritized according to these functions.

Let T be the set of all tasks from all sets Q generated in the simulation phase (see Section 3.2). For a task $t \in T$, we have a 4-tuple $(p_t, q_t, r_t, score(p_t, q_t, r_t))$ obtained from the previously computed trial score distributions. This 4-tuple denotes the *observation* of the scheduling performance behavior of the task t. Given a collection $\mathcal{F}$ of nonlinear functions, the problem consists in finding the function $f(p, q, r) \in \mathcal{F}$ that better fits the distribution $score(p, q, r)$ generated from all tasks $t \in T$.

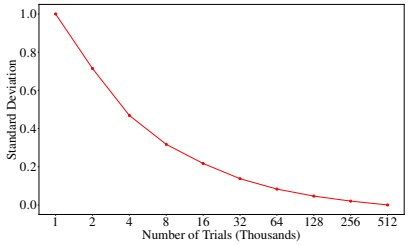
We defined $\mathcal{F}$ as all functions of form $f = (c_1\alpha(p))\ op_1\ (c_2\beta(q))\ op_2\ (c_3\gamma(r))$. We call α, β and γ as *base functions* and they can be any of the functions: identity, square root, logarithm and inverse. Operators op_1 and op_2 are any of the operators sum (+), multiplication ($\cdot$) or division ($\div$). Coefficients c_1, c_2 and c_3 denotes the relative importance of the base functions and are obtained by a nonlinear regression fitting. We chose the aforementioned base functions because they, in conjunction, can express a wide variety of nonlinear relationships, while maintaining a reasonable number of combinations to perform the fitting. Also, by using few parameters



(a) One example of a trial score distribution



(b) Another example of a trial score distribution



(c) Normalized standard deviations for the trial score distributions obtained with different numbers of the trials.

Figure 1: Examples of trial score distributions generated by the simulation procedure for a tuple of task sets (S, Q), with $|S| = 16$ and $|Q| = 32$, in a cluster with 256 nodes, and their estimation accuracy in function of the number of trials. The black horizontal in (a) and (b) line represents the mean $\frac{1}{|Q|} = \frac{1}{32} = 0.031$.

we reduce the possibility of overfitting. We employ a weighted nonlinear regression [4] procedure, which minimizes the error:

$$error = \sum_{t \in T} ((p_t q_t) \cdot (f(p_t, q_t, r_t) - score(p_t, q_t, r_t)))^2 \quad (4)$$

We used the weight $(p_t q_t)$ to emphasize that the fit must perform a good estimation of the score of large area tasks (*i.e.* tasks with large p and q values). This is based on the argument that tasks that consume a large amount of resources for a long period of time have a potential of blocking the execution of many smaller tasks, degrading the overall scheduling performance.

Once we perform the fitting for all functions $f \in \mathcal{F}$, we use a *AVGerror* function (Equation 5) to evaluate the average error of each nonlinear function f, when used to estimate the $score(p_t, q_t, r_t)$ for all tasks $t \in T$. The hypothesis is that nonlinear functions with the lower *AVGerror* values would perform well as scheduling policies.

$$AVGerror(f) = \frac{1}{|T|} \sum_{t \in T} \|f(p_t, q_t, r_t) - score(p_t, q_t, r_t)\| \quad (5)$$

4 RESULTS

In this section we present the main results obtained in our work. We first describe the simulation procedure and the nonlinear functions obtained with machine learning. Next, we evaluate the performance of the obtained functions when scheduling synthetic workloads under different conditions: (i) using actual task processing times, (ii) using user estimated task processing times, and (iii) using user estimated task times with backfilling. Finally, we evaluated the obtained functions using real workload traces, for the same three conditions used with the synthetic workloads.

In the simulations, to generate the distribution $score(p, q, r)$, we considered an HPC platform compounded by $n_{max} = 256$ homogeneous processors. We used sets of tasks S and Q with $|S| = 16$ and $|Q| = 32$ tasks. All simulations were performed using SimGrid [5].

We compared the performance of our scheduling policies with a selection of classical scheduling and smart *ad-hoc* policies, used in real HPC platforms (Table 1). Two classical and well known policies are First Come First Served (FCFS), where tasks are scheduled by the arrival order, and Shortest Processing Time First (SPT), where tasks with smaller processing times are scheduled first. We also used the WFP3 (WFP, for short) and UNICEF (UNI, for short) policies [28], which are based on the processing time (p_t), requested number of processors (q_t), and waiting time (w_t) of the task. The reasoning of WFP is that shorter and/or older tasks should be largely favored, while preventing the starvation of large tasks. UNI in its turn attempts to provide a fast turnaround for small tasks by favoring them.

Table 1: Scheduling policies used for comparison.

Name	Function
FCFS	$score(t) = r_t$
SPT	$score(t) = p_t$
WFP3	$score(t) = -(w_t/p_t)^3 \cdot q_t$
UNICEF	$score(t) = -w_t/(\log_2(q_t) \cdot p_t)$

4.1 Machine Learning: Obtained Nonlinear Functions

The first step towards obtaining the nonlinear functions for usage as scheduling policies is to produce the distribution $score(p, q, r)$. We start by generating permutations of the set of tasks Q, which are used to construct the trial score distributions. Enumerating and simulating the execution of all permutations of a set of tasks of size $|Q| = 32$ is unfeasible and, therefore, we need to define a suitable number of permutations that generate accurate trial score distributions. For that, we selected one tuple (S, Q) and generated the trial distributions with increasing amount of trials, repeating the simulation procedure ten times per number of trials, and measuring the standard deviation of the estimated scores. Figure 1c shows that the standard deviation drops quickly with increasing amount of trials. With 256 thousand trials, the resulting normalized standard deviation was 0.02. We decided to use 256 thousand trials, since its simulation takes less than 11 minutes using SimGrid [5] on an Intel Xeon E5-2620v2 six-core CPU.

After obtaining the trial score distributions, we generated the distribution $score(p, q, r)$ and performed the nonlinear regression

Table 2: The four best nonlinear functions obtained using nonlinear regression.

ID	Nonlinear Function	*AVGerror(f)*
F1	$\log_{10}(p) \cdot q + 8.70 \cdot 10^2 \cdot \log_{10}(r)$	$5.278 \cdot 10^{-3}$
F2	$\sqrt{p} \cdot q + 2.56 \cdot 10^4 \cdot \log_{10}(r)$	$5.317 \cdot 10^{-3}$
F3	$p \cdot q + 6.86 \cdot 10^6 \cdot \log_{10}(r)$	$5.408 \cdot 10^{-3}$
F4	$p \cdot \sqrt{q} + 5.30 \cdot 10^5 \cdot \log_{10}(r)$	$5.482 \cdot 10^{-3}$

using the function `leastsq()` from the SciPy [16] Python library for all functions $f \in \mathcal{F}$. Table 2 shows the four best functions obtained with regard to the Equation 5. We mathematically simplified the obtained functions, merging the coefficients c_1, c_2 and c_3 into a single coefficient $c_3/(c_1c_2)$, in front of the $log(r)$ term.

A noticeable phenomenon is their similarity, with all functions constituted by a sum of two factors, one containing parameters p and q and the other with the dependence on $log(r)$. Considering the large values of the constant before the $log(r)$ term, the functions emphasize that tasks that arrived earlier (*i.e.* with lower r values) must be prioritized in order to maintain lower slowdowns (recall that tasks with lower score value have a high priority). Moreover, these large constants effectively prevent starvation without any manual customization of the policies. Figures 2b and 2c illustrate the strong dependency on the submission time for all policies F1 to F4, with tasks that arrive earlier receiving a large priority (darker colors) over more recent tasks.

The second important factor is the size of the task, a product of two functions $f(p)$ and $g(q)$ of the processing time and number of processors used by the tasks. The policies F1 to F4 differ in the relative importance given to each of these values (p and q), with F1 and F2 imposing a heavier penalization for increasing numbers of requested processors q, F4 penalizing larger processing times p, and F3 penalizing higher values of p and q equally. Figure 2a shows that, for a fixed value of r, higher priorities are given for tasks that have either required smaller processing times or number of processors.

These results comply with the general intuition – in which tasks with small processing time, small requested amount of processors and that were submitted earlier should be prioritized – that is adopted by most of the *ad-hoc* scheduling policies. The main differences are the adoption of two separate terms, one considering only task size and the other submission time, and the large coefficient before the submission time term.

4.2 Scheduling Performance: Workload Model

In this subsection we aim to answer the following question: *Can the nonlinear functions, obtained using the procedure from Section 4.1, perform well as scheduling policies for tasks generated by the Lublin and Feitelson [18] workload model in the following scenarios?*

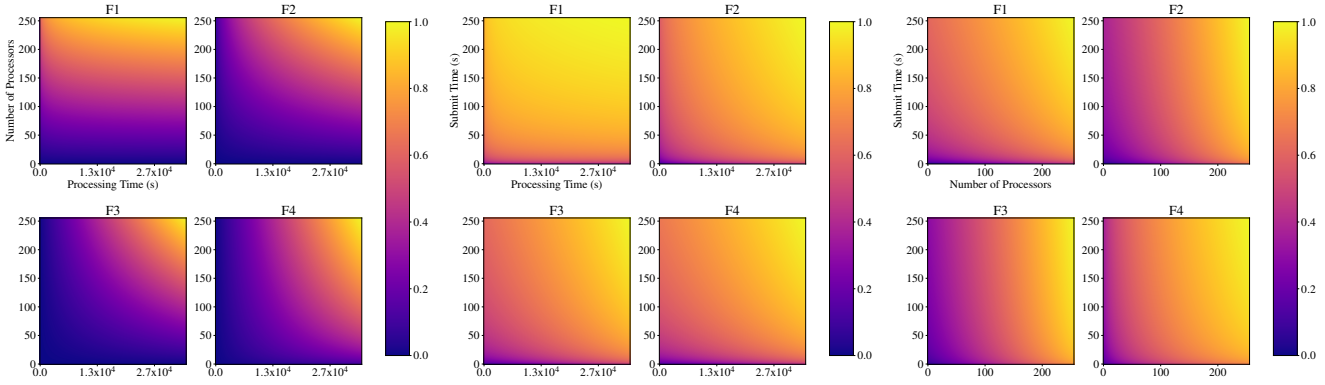
- Using the actual processing time p in the scheduling decisions and the same number of processors $n_{max} = 256$ from the simulation scheme;
- Using the actual processing time p in the scheduling decisions, but increasing the number of processors to $n_{max} = 1024$;
- Using the processing time estimate $\tilde{p}$ provided by the user, instead of the actual processing time p, to perform the scheduling decisions;
- Using the processing time estimate $\tilde{p}$ provided by the user, but performing the scheduling using the aggressive backfilling algorithm.

We should emphasize that in all scenarios we used the same set of nonlinear functions, obtained using the actual processing time p of tasks and $n_{max} = 256$ processors. Since the functions are parametrized by the number of processors (q), submission time (r) and processing time p (which can be substituted by its estimate $\tilde{p}$), they can be used in different scenarios. The objective was to evaluate the generalization capabilities of the obtained nonlinear functions.

For all experiments, we define a *dynamic scheduling experiment* as being multiple simulations of the execution of a sequence of tasks obtained from a certain workload trace or model. For each simulation we choose one scheduling policy from Tables 1 and 2 to schedule all tasks in the sequence. The output of the dynamic scheduling experiment is the average bounded slowdown from each simulation performed, using all policies present in Tables 1 and 2. The sequence of tasks contains all tasks submissions over a period of fifteen days. When using workload traces, we assured that there was no overlap between sequences used on distinct dynamic scheduling experiments.

The on-line scheduling algorithm works as follows: tasks arrive in a centralized waiting queue and the scheduler performs a reschedule – using a scheduling policy – of the tasks present in this queue in two distinct events: (i) when a task arrives in the queue or (ii) when a resource (set of processors) is released and becomes available. When a task t is selected for execution and if the requested number of processors q_t is lower than the total number of processors available, then q_t processors are reserved for this task and they become unavailable. These processors will become available again only when p_t units of time have passed since the start of the execution of t. If there is not enough processors to process t, then the scheduler waits for one of the two rescheduling events mentioned above.

4.2.1 Scheduling using actual task runtimes p. In this experiment we generated a task queue with characteristics from the workload model of Lublin and Feitelson [18] and considering a HPC platform constituted by $n_{max} = 256$ processors, which are the same settings used in the simulations to generate the nonlinear functions. Figure 3a shows the average bounded slowdowns for fifty dynamic scheduling experiments, with the orange line representing the median of the average bounded slowdown (whose values are shown in Table 3) over these experiments, the box limits representing the upper and lower quartiles, and the whiskers representing the lowest and highest values outside the the box limits but still inside the range of 1.5 times the difference between the upper and lower quartiles. The figure shows that all obtained nonlinear functions performed substantially better than the evaluated existing scheduling policies, from Table 1. The nonlinear function F2 provided the best results, followed by functions F1, F3 and F4. Although we



(a) Processing time *versus* number of processors **(b) Processing time *versus* submit time. Older tasks have small submit time values** **(c) Number of processors *versus* submit time. Older tasks have small submit time values**

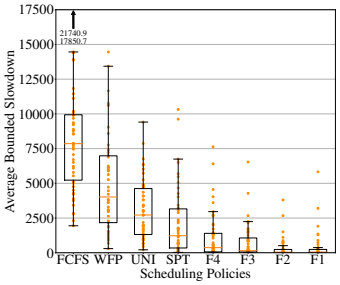
Figure 2: Dependency on the parameters p, q and r for the four best nonlinear functions obtained.

Table 3: Median of the average bounded slowdowns from Subsections 4.2 and 4.3.

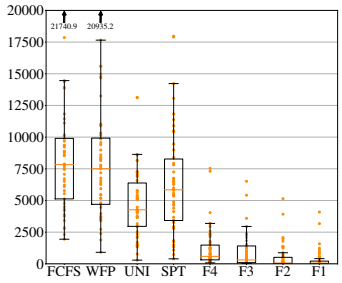
Experiment	FCFS	WFP	UNI	SPT	F4	F3	F2	F1
Workload model, n_{max} = 256, actual runtimes p	7830.58	4018.25	2717.84	1222.67	368.37	151.53	**37.47**	42.88
Workload model, n_{max} = 256, runtime estimates $\tilde{p}$	7830.58	7505.74	4265.58	5841.89	590.40	306.68	60.76	**50.67**
Workload model, n_{max} = 256, aggressive backfilling	1306.38	1220.52	508.95	863.76	122.72	54.49	**42.91**	50.77
Workload model, n_{max} = 1024, actual runtimes p	6910.82	5229.28	3176.86	2404.65	514.43	73.43	**21.41**	24.78
Workload model, n_{max} = 1024, runtime estimates $\tilde{p}$	6910.82	6564.50	3790.40	5220.45	1146.34	559.86	123.10	**32.33**
Workload model, n_{max} = 1024, aggressive backfilling	1262.63	1651.83	1079.08	1296.65	377.99	180.75	44.70	**30.24**
Curie workload trace, actual runtimes p	290.27	206.57	110.14	139.58	18.77	8.21	**4.27**	6.57
Anl Interpid workload trace, actual runtimes p	40.93	11.06	5.83	3.91	2.12	**1.76**	2.07	2.15
HPC2N workload trace, actual runtimes p	90.61	22.20	11.37	12.16	11.65	15.40	**10.47**	12.12
SDSC Blue workload trace, actual runtimes p	370.22	67.83	29.53	24.65	20.03	12.63	**4.41**	8.22
SDSC SP2 workload trace, actual runtimes p	629.60	59.60	25.84	21.78	29.14	45.33	**15.58**	19.53
CTC SP2 workload trace, actual runtimes p	438.37	309.72	33.86	90.24	18.77	15.21	**5.32**	9.13
Curie workload trace, runtime estimates $\tilde{p}$	290.27	258.17	145.19	218.77	44.63	12.15	**8.77**	9.86
Anl Interpid workload trace, runtime estimates $\tilde{p}$	40.93	18.33	11.94	9.17	4.18	3.29	**2.63**	2.74
HPC2N workload trace, runtime estimates $\tilde{p}$	90.61	62.24	28.34	46.67	27.81	20.08	**16.90**	19.83
SDSC Blue workload trace, runtime estimates $\tilde{p}$	370.22	136.11	57.00	46.82	18.69	9.66	**8.20**	11.52
SDSC SP2 workload trace, runtime estimates $\tilde{p}$	629.60	243.86	91.57	48.43	40.21	41.38	**30.04**	34.13
CTC SP2 workload trace, runtime estimates $\tilde{p}$	438.37	369.93	104.98	306.78	21.76	18.23	13.46	**12.33**
Curie workload trace, aggressive backfilling	54.94	49.23	32.27	35.72	14.90	9.57	**7.80**	9.81
Anl Interpid workload trace, aggressive backfilling	9.66	7.19	4.39	3.79	3.59	2.87	**2.65**	2.74
HPC2N workload trace, aggressive backfilling	36.43	28.08	16.22	16.26	15.31	**15.26**	16.83	20.46
SDSC Blue workload trace, aggressive backfilling	36.40	14.07	11.08	9.87	8.11	**8.02**	8.52	11.59
SDSC SP2 workload trace, aggressive backfilling	66.38	41.61	32.66	28.38	26.68	25.31	**23.26**	32.06
CTC SP2 workload trace, aggressive backfilling	83.38	54.32	24.96	14.40	**9.90**	11.07	10.23	12.88

would expect F1 to perform better, the *AVGerror* values of the generated functions are similar (Table 2) and, consequently, we cannot assume that F1 will always provide the best results.

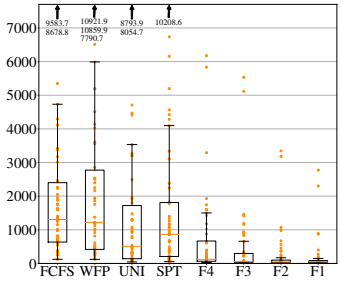
Figure 3d shows the results of fifty dynamic scheduling experiments considering an HPC platform with n_{max} = 1024 processors. The tasks were generated using the Lublin and Feitelson workload model configured for a cluster with 1024 processors, so that tasks sent to the waiting queue would use between 1 and 1024 processors. We can see that, compared to the other scheduling policies, the obtained nonlinear functions continued to perform well, indicating that the obtained scheduling policies have some generalization capability regarding the number of processors in the HPC platform.



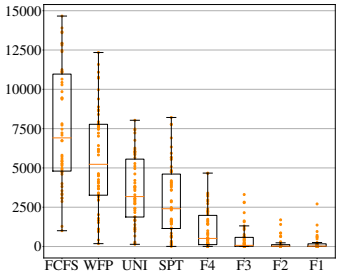
(a) Tasks characteristics generated from the Lublin and Feitelson [18] model and n_{max} = 256



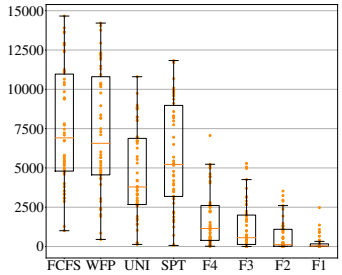
(b) Tasks characteristics generated from the Lublin and Feitelson [18] model, with processing time estimates generated from the Tsafrir *et al.* [29] model and n_{max} = 256



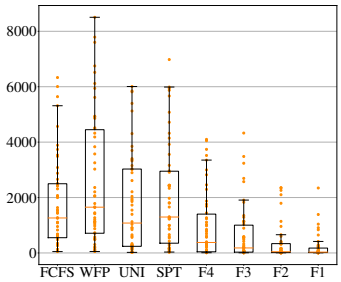
(c) Tasks characteristics generated from the Lublin and Feitelson [18] model, with processing time estimates generated from the Tsafrir *et al.* [29] model and n_{max} = 256 and with aggressive backfilling



(d) Same settings as (a), however with n_{max} = 1024



(e) Same settings as (b), however with n_{max} = 1024



(f) Same settings as (c), however with n_{max} = 1024

Figure 3: Scheduling performance results with tasks characteristics generated from a workload model.

4.2.2 Scheduling using user estimated task runtimes. In this experiment, instead of using the processing time p_t in the scheduling decisions, we utilize the *estimated* processing time $\tilde{p_t}$ of the task that is provided by the user. The actual processing time p_t in this case is used only to simulate the execution of the task. For the workload model used in this work, we used the user runtime estimate model of Tsafrir *et al.* [29] to generate $\tilde{p_t}$.

Traditionally the processing time estimates provided by the user are highly inaccurate. In this light, it is expected a reduction in the scheduling performance of all the scheduling policies, since none of these scheduling policies are designed to handle inaccuracies in the execution time of the tasks, and therefore the only aspect that we can evaluate is how tolerant the scheduling policies are when inaccurate processing time estimates are introduced.

Figures 3b and 3e show the results of fifty dynamic scheduling experiments, using estimated processing times $\tilde{p}$, for HPC platforms constituted by n_{max} = 256 and n_{max} = 1024 processors. As expected, all scheduling policies had a considerable performance degradation, except the FCFS, which does not use task processing times. Nevertheless, the median of the average bounded slowdown generated by policies F1, F2, F3 and F4 was between 7.22 and 84.18 times better for the scenario with n_{max} = 256 and between 3.30 and 117.24 times better for the scenario with n_{max} = 1024, when compared to the best performing *ad-hoc* scheduling policy.

4.2.3 Scheduling using estimated runtimes and agressive backfilling. In this experiment, we used the aggressive backfilling algorithm in conjunction with the scheduling policies. In such setting, when a rescheduling event occurs, the tasks are reordered in the queue using a scheduling policy and then we apply the aggressive backfilling algorithm to check if there is one or more tasks further back in the queue that, if selected for execution, will not delay the first task in the queue. In this case, all the scheduling decisions (the scheduling policy and the backfilling) are made over the estimated processing time $\tilde{p_t}$, with the actual processing time p_t used only to simulate task execution. This setting is the most realistic setting we can elaborate using tasks generated from a workload model.

Figures 3c and 3f show that the introduction of the aggressive backfilling algorithm resulted in a overall increase in performance, with the FCFS policy with backfilling (previously mentioned as the EASY algorithm) taking the most advantage of the backfilling strategy. Our obtained nonlinear functions had the least benefits from the backfilling, since the better initial schedules lowered the possibilities for task backfilling. Nevertheless, the performance of the obtained nonlinear functions is still superior to the other *ad-hoc* scheduling policies. For instance, the median average slowdown for the F2 strategy was more than 11 times smaller than the best *ad-hoc* policy for both 256 and 1024 core machines.

Table 4: Real workload traces used for evaluation of the scheduling policies.

Name	Year	# CPUs	# Jobs	Util %	Duration
Curie	2011	93,312	312,826	62.0	20 Months
ANL Interpid	2009	163,840	68,936	59.6	8 Months
HPC2N	2002	240	202,871	60.1	42 Months
SDSC Blue	2000	1,152	243,306	76.7	32 Months
SDSC SP2	1998	128	59,715	83.4	24 Months
CTC SP2	1997	338	77,222	85.2	11 Months

4.3 Scheduling Performance: Real Workload Traces

We also evaluated whether the obtained scheduling policies generalize well to highly different workloads types, obtained from real workload traces, and HPC platform configurations. In this light, in this subsection we attempt to answer the following questions:

- Can the obtained scheduling policies perform well when scheduling a set of tasks extracted from real workload traces and executed in a simulated HPC platform similar to the one where the traces were obtained?
- With the same setting from the previous question, but using the user estimated processing times $\tilde{p}$ to perform the scheduling decisions, can the obtained nonlinear functions perform well as scheduling policies?
- Can the obtained nonlinear functions benefit from the aggressive backfilling algorithm and perform well in the scenario from the previous question?

We used the traces described in Table 4, which are publicly available at the Parallel Workloads Archive [11]. To better evaluate the generality of the obtained nonlinear functions, we chose a set of traces from computer HPC configurations ranging from 128 to 163,840 processors, mean utilization values from 59.6% to to 85.2% and measurements dates from year 1997 to 2011.

On each trace we collected as many non overlapping half-month sequences of tasks to perform the dynamic scheduling experiments as possible. Each sequence contains all tasks submissions equivalent to a period of fifteen days. We made sure that there was no overlap between the sequences and the on-line scheduling algorithm works similarly to the scheduling algorithm used in the experiments of the previous subsection.

4.3.1 Scheduling using the actual task runtimes p. Figure 4 shows the dynamic scheduling experiments results using the actual processing times to perform the scheduling decisions. All the obtained nonlinear functions resulted in lower average slowdowns for all traces, with varying levels of improvements depending on the HPC workload characteristics. More importantly, the difference between the upper and lower quartiles (box limits) was lower when using the obtained nonlinear functions, meaning that the average slowdown was more predictable and stable, a desirable property for HPC systems. For all policies there were some cases where the average bounded slowdown was notoriously high, which occurred due to uncommon tasks characteristics in the traces. For example, in the HPC2N trace, some workloads contains bursts of tasks with large processing times, which overloaded the platform for all policies.

The nonlinear function F2 achieved overall best results for most of the traces, with the exception of the the Anl Interpid trace, in which F3 performed better. This result shows that, with real workload traces, the best scheduling policy can change from trace to trace. This behavior is not unexpected, since the workload traces evaluated are different from each other and from the workload generation model used in the simulation scheme. But choosing any of the obtained policies F1 to F4 resulted in improvements in the median average slowdowns in all scenarios and smaller differences between the upper and lower quartiles in most scenarios.

4.3.2 Scheduling using user estimated task runtimes $\tilde{p}$. We evaluated the scheduling policies using the processing time estimate $\tilde{p}$ obtained from the respective workload log when performing the scheduling decisions. Since the user estimates of the processing times are often rough and inaccurate, we expect a degradation in the performance of all scheduling policies. Figure 5 shows that the obtained functions F1 to F4 continued to generate lower median average slowdowns and differences between the upper and lower quartiles for all evaluated HPC platforms. Although the best function from F1 to F4 varied depending on the platform, any of them would result in performance improvements over existing *ad-hoc* policies.

The results from this section are noteworthy, considering the nonlinear functions F1 to F4 were trained using data from a single workload model in a simulated machine with 256 processors. These functions worked well as scheduling policies for HPC machines with highly different architectures, with up to 163,840 processors, very different workload types, and using inaccurate estimated task runtimes from real machine users.

4.3.3 Scheduling using estimated processing times and aggressive backfilling. In this experiment we considered the most realist scenario, where scheduling decisions are based on the user estimates of processing times $\tilde{p}$ and with the addition of aggressive backfilling to reduce resource idleness. Figure 6 shows the corresponding dynamic scheduling experiments. Once again, the FCFS policy with backfilling (the EASY scheduling algorithm) was the scheduling policy which benefited the most with the introduction of backfilling. The performance results obtained for the WFP and UNI policies also reinforces the favorable results obtained by Tang *et al.* [28], since we obtained similar comparative results for these policies, with the exception that we evaluated them with different workload logs.

The obtained nonlinear functions had smaller benefits from using backfilling. Similarly to the experiments with synthetic workloads, the better schedules provided by the proposed nonlinear functions resulted in less opportunities for backfilling tasks. Nevertheless, functions F1 to F4 still resulted in lower median average slowdowns and/or lower differences between the extreme quartiles for most scenarios, and continued to be a better general choice than the *ah-hoc* scheduling policies.

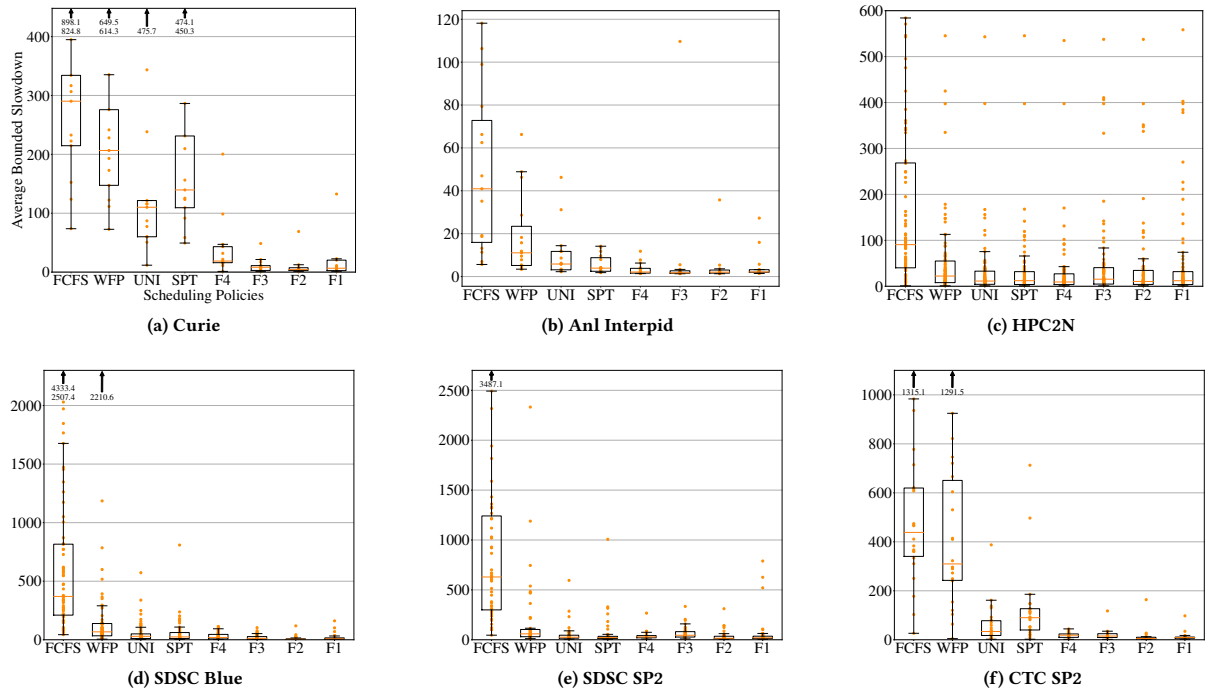


Figure 4: Scheduling performance results with tasks characteristics obtained from real HPC platform workload logs and using actual processing time in the scheduling decisions.

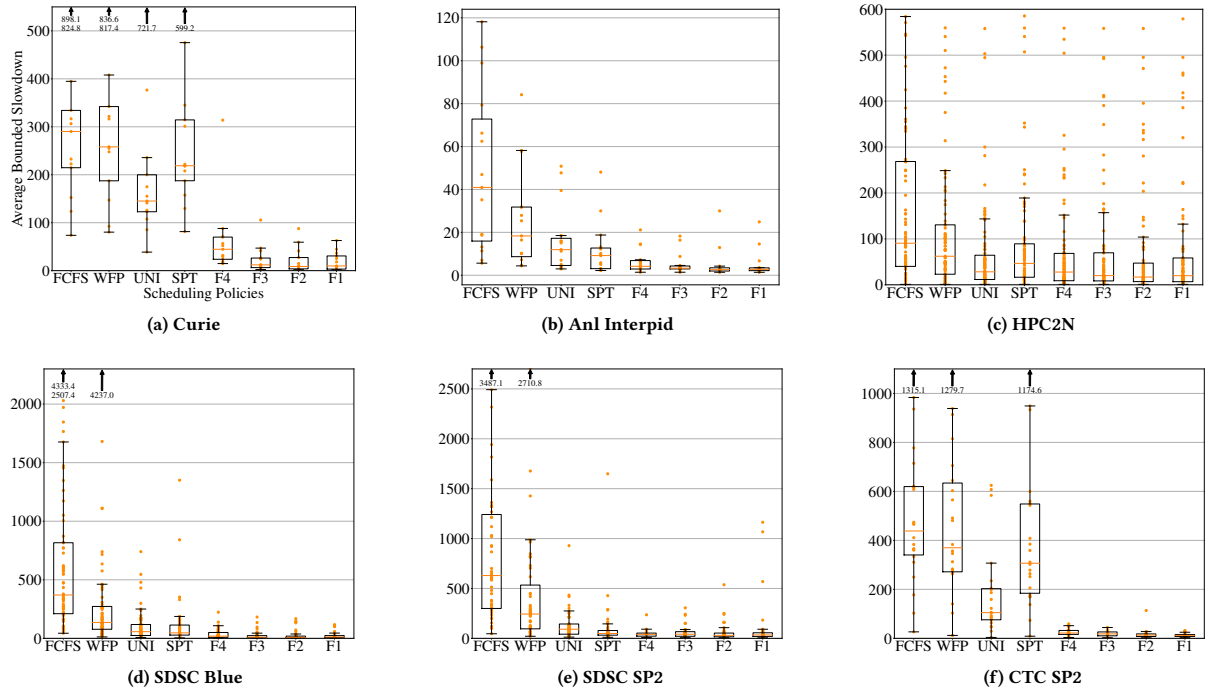


Figure 5: Scheduling performance results with tasks characteristics obtained from real HPC platform workload logs and using user estimated processing times obtained from the same logs in the scheduling decisions.

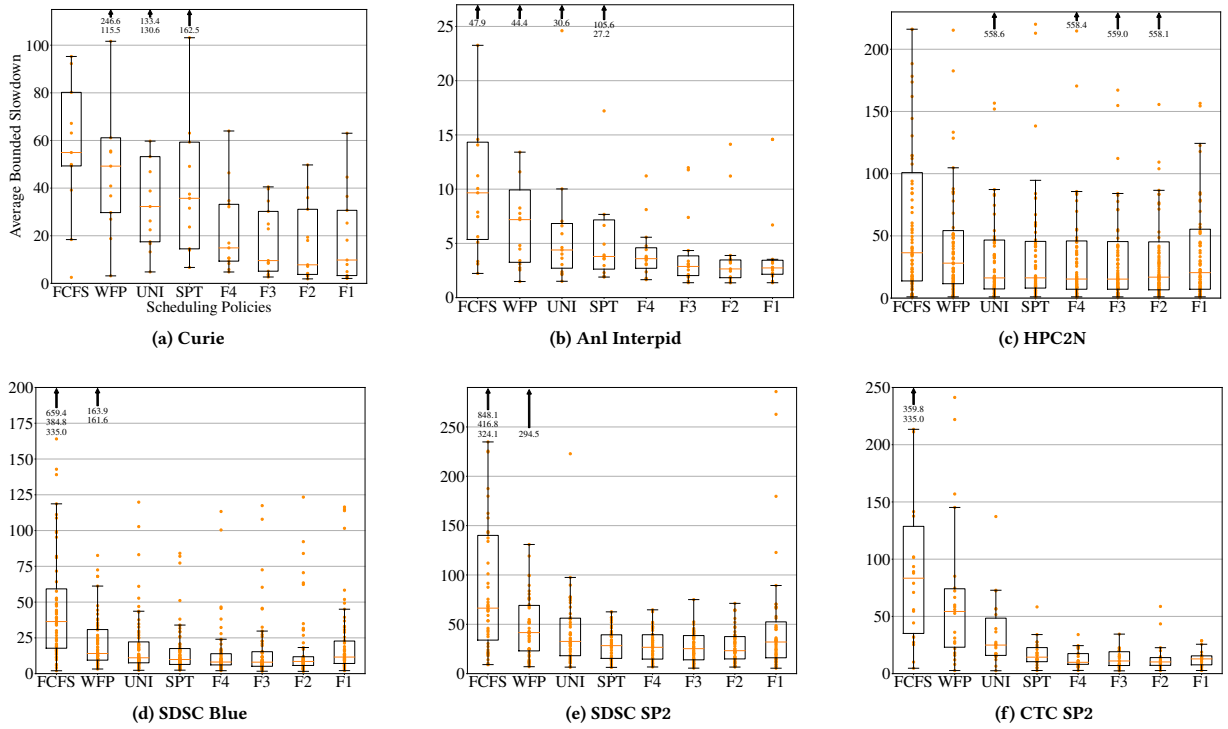


Figure 6: Scheduling performance results with tasks characteristics obtained from real HPC platforms, with the addition of the aggressive backfilling, and using user estimated processing times obtained from the same logs in the scheduling decisions.

5 CONCLUSIONS

Due to its simplicity, scheduling policies in the form of functions that take task characteristics into consideration, combined with the computationally inexpensive aggressive backfilling algorithm, are an appealing alternative for the problem of on-line scheduling of tasks in HPC platforms. In an equally simple manner, in this work we show that – by introducing a simulation procedure that captures the observations of task scheduling behavior under several distinct conditions and a simple machine learning strategy to model these observations into nonlinear functions – we can obtain scheduling policies that effectively capture the effects of task characteristics on scheduling behavior. We showed that these functions outperform other classical and smart *ad-hoc* scheduling policies in a variety of scenarios. Moreover, the large weights assigned to the submission time of the tasks (Table 2) result in a fast increase in priority as the tasks wait in the queue (Figures 2b and 2c), thus preventing starvation without any manual customization of the policies.

Using tasks characteristics obtained from the workload model of Lublin and Feitelson [18] and considering similar scenarios from the one used to capture the scheduling observations, the obtained scheduling policies achieved very good results. In the most realistic of these scenarios (*i.e.* using the aggressive backfilling in addition to the scheduling policies and using the processing time estimates to perform the scheduling decisions), the obtained scheduling policies achieved medians of the average bounded slowdowns up to 11.8 times better when compared to the best performing *ad-hoc* scheduling policy.

Although we used a workload model to generate the nonlinear functions, we could envision the same procedure being applied to obtain custom scheduling policies for a specific HPC platform, using its specific workload traces and architecture configurations. Our results using the workload model indicate that these custom policies could, in principle, bring important improvements in the achieved average bounded slowdowns in these platforms. We also have confidence that the procedures presented in this paper can be applied with other objective functions such as average waiting time or flow time. Therefore, our approach gives freedom for the HPC platform maintainer to choose the objective function that best suits the QoS requirements of his users. Also, one can applies the proposed procedure periodically, using data from newly executed tasks. This dynamic update will adapt the scheduling policies to changes in the pattern of tasks submitted in the platform.

Scheduling performance could also be improved by combining our machine learning scheme for obtaining efficient scheduling policies with a machine learning scheme for predicting task execution times, such as the one proposed by Gaussier *et al.* [13]. We showed that inaccurate execution time estimates from users greatly reduces the efficiency of our scheduling policies. We believe that by using execution time predictions we could improve the average bounded slowdown in the more realistic scenario where we do not know the actual task execution times.

Finally, we would like to note that using a workload model to capture the observations brought some important advantages. The generalized task properties present in the Lublin and Feitelson [18] workload model, when used to observe the scheduling behavior of the tasks in several distinct configurations, resulted in scheduling policies that are able to express efficient general patterns regarding to which task should be selected for execution first. In this light, the obtained scheduling policies showed consistent lower median values for the average bounded slowdown in simulation experiments considering very different task types and HPC platform configurations. Therefore, although it would be possible to define scheduling policies specifically for each HPC platform, as aforementioned stated, it seems that general policies, as the ones we found in this work, could be sufficient to efficiently schedule tasks for a range of specific workload types and HPC platform configurations.

As future work, we could improve the current work in two directions. The first would be to evaluate the scheduling policies by deploying them on real HPC platforms and check how they perform under real conditions. We used simplified simulations to evaluate the obtained scheduling policies using real traces which ignores, for instance, network and memory bottlenecks that could appear from interactions among task execution. Nevertheless, we believe the simulations are a reasonable approximation for these platforms.

The second direction is covering a wider range of HPC platforms. We plan to improve the strategy proposed by this work to obtain scheduling policies that also address the on-line scheduling of tasks in HPC platforms containing processing units with distinct architectures such as GPUs [21] and MICs [8], where multiple implementations, aiming a specific architecture, are available for the same task and the scheduler needs to select one of these implementations to be executed.

ACKNOWLEDGMENT

The authors would like to thank the CAPES Foundation and the Ministry of Education of Brazil for the scholarship grant, FAPESP (process 13/26644-1) for the financial support, and professor Denis Trystram for suggestions on workload generation models. We also would like to thank Ake Sandgren, Dan Dwyer, Joseph Emeras, Susan Coghlan, Travis Earheart and Victor Hazlewood for making their workload logs publicly available.

REFERENCES

[1] Anurag Agarwal, Selcuk Colak, Varghese S Jacob, and Hasan Pirkul. 2006. Heuristics and augmented neural networks for task scheduling with non-identical machines. *European Journal of Operational Research* 175, 1 (2006), 296–317.

[2] Derya Eren Akyol and G Mirac Bayhan. 2007. A review on evolution of production scheduling with neural networks. *Computers & Industrial Engineering* 53, 1 (2007), 95–122.

[3] Hadil Al-Daoud, Issam Al-Azzoni, and Douglas G Down. 2012. Power-aware linear programming based scheduling for heterogeneous computer clusters. *Future Generation Computer Systems* 28, 5 (2012), 745–754.

[4] Raymond J Carroll and David Ruppert. 1988. *Transformation and weighting in regression*. Vol. 30. CRC Press.

[5] Henri Casanova, Arnaud Giersch, Arnaud Legrand, Martin Quinson, and Frédéric Suter. 2014. Versatile, Scalable, and Accurate Simulation of Distributed Applications and Platforms. *J. Parallel and Distrib. Comput.* 74, 10 (June 2014), 2899–2917. http://hal.inria.fr/hal-01017319

[6] Adaptative Computing. 2017. Moab Workload Manager Documentation. http://www.adaptivecomputing.com/support/documentation-index/. (2017). Accessed: 2017-03-26.

[7] Rubing Duan, Farrukh Nadeem, Jie Wang, Yun Zhang, Radu Prodan, and Thomas Fahringer. 2009. A hybrid intelligent method for performance modeling and prediction of workflow activities in grids. In *Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid*. IEEE Computer Society, 339–347.

[8] Alejandro Duran and Michael Klemm. 2012. The Intel® many integrated core architecture. In *High Performance Computing and Simulation (HPCS), 2012 International Conference on*. IEEE, 365–366.

[9] Dror G Feitelson. 2001. Metrics for parallel job scheduling and their convergence. In *Workshop on Job Scheduling Strategies for Parallel Processing*. Springer, 188–205.

[10] Dror G Feitelson, Larry Rudolph, Uwe Schwiegelshohn, Kenneth C Sevcik, and Parkson Wong. 1997. Theory and practice in parallel job scheduling. In *Workshop on Job Scheduling Strategies for Parallel Processing*. Springer, 1–34.

[11] Dror G Feitelson, Dan Tsafrir, and David Krakov. 2014. Experience with using the parallel workloads archive. *J. Parallel and Distrib. Comput.* 74, 10 (2014), 2967–2982.

[12] Christodoulos A Floudas and Xiaoxia Lin. 2005. Mixed integer linear programming in process scheduling: Modeling, algorithms, and applications. *Annals of Operations Research* 139, 1 (2005), 131–162.

[13] Eric Gaussier, David Glesser, Valentin Reis, and Denis Trystram. 2015. Improving Backfilling by Using Machine Learning to Predict Running Times. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC '15)*. ACM, New York, NY, USA, Article 64, 10 pages. DOI: https://doi.org/10.1145/2807591.2807646

[14] Y Georgiou. 2010. *Resource and Job Management in High Performance Computing*. Ph.D. Dissertation. PhD Thesis, Joseph Fourier University, France.

[15] Edwin SH Hou, Nirwan Ansari, and Hong Ren. 1994. A genetic algorithm for multiprocessor scheduling. *IEEE Transactions on parallel and distributed systems* 5, 2 (1994), 113–120.

[16] Eric Jones, Travis Oliphant, Pearu Peterson, and others. 2001–. SciPy: Open source scientific tools for Python. (2001–). http://www.scipy.org/

[17] Jérôme Lelong, Valentin Reis, and Denis Trystram. 2017. Tuning EASY-Backfilling Queues. In *21st Workshop on Job Scheduling Strategies for Parallel Processing*.

[18] Uri Lublin and Dror G Feitelson. 2003. The workload on parallel supercomputers: modeling the characteristics of rigid jobs. *J. Parallel and Distrib. Comput.* 63, 11 (2003), 1105–1122.

[19] Ahuva W. Mu'alem and Dror G. Feitelson. 2001. Utilization, predictability, workloads, and user runtime estimates in scheduling the IBM SP2 with backfilling. *IEEE Transactions on Parallel and Distributed Systems* 12, 6 (2001), 529–543.

[20] Bill Nitzberg, Jennifer M Schopf, and James Patton Jones. 2004. PBS Pro: Grid computing and scheduling attributes. In *Grid resource management*. Springer, 183–190.

[21] J. D. Owens, M. Houston, D. Luebke, S. Green, J. E. Stone, and J. C. Phillips. 2008. GPU computing. *Proc. IEEE* 96, 5 (2008), 879–899.

[22] Nikela Papadopoulou, Georgios Goumas, and Nectarios Koziris. 2015. A machine-learning approach for communication prediction of large-scale applications. In *Cluster Computing (CLUSTER), 2015 IEEE International Conference on*. IEEE, 120–123.

[23] Dejan Perkovic and Peter J Keleher. 2000. Randomization, speculation, and adaptation in batch schedulers. In *Proceedings of the 2000 ACM/IEEE conference on Supercomputing*. IEEE Computer Society, 7.

[24] Michael L. Pinedo. 2008. *Scheduling: Theory, Algorithms, and Systems* (3rd ed.). Springer Publishing Company, Incorporated.

[25] Harmel Karam Singh and Abdou Youssef. 1995. *Mapping and scheduling heterogeneous task graphs using genetic algorithms*. Master's thesis. George Washington University.

[26] Garrick Staples. 2006. TORQUE resource manager. In *Proceedings of the 2006 ACM/IEEE conference on Supercomputing*. ACM, 8.

[27] Achim Streit. 2001. The self-tuning dynP job-scheduler. In *Parallel and Distributed Processing Symposium., Proceedings International, IPDPS 2002, Abstracts and CD-ROM*. IEEE, 8–pp.

[28] Wei Tang, Zhiling Lan, Narayan Desai, and Daniel Buettner. 2009. Fault-aware, utility-based job scheduling on BlueGene/P systems. In *Cluster Computing and Workshops, 2009. CLUSTER'09. IEEE International Conference on*. IEEE, 1–10.

[29] Dan Tsafrir, Yoav Etsion, and Dror G Feitelson. 2005. Modeling user runtime estimates. In *Workshop on Job Scheduling Strategies for Parallel Processing*. Springer, 1–35.

[30] Fatos Xhafa and Ajith Abraham. 2010. Computational models and heuristic methods for Grid scheduling problems. *Future generation computer systems* 26, 4 (2010), 608–621.

[31] Andy B Yoo, Morris A Jette, and Mark Grondona. 2003. Slurm: Simple linux utility for resource management. In *Workshop on Job Scheduling Strategies for Parallel Processing*. Springer, 44–60.

A ARTIFACT DESCRIPTION: OBTAINING DYNAMIC SCHEDULING POLICIES WITH SIMULATION AND MACHINE LEARNING

A.1 Abstract

We provide the source code implementation of all of the strategies proposed in the paper as well as the source code of the simulation experiments used to evaluate our approach so the reader can (i) generate their own distribution $score(p, q, r)$, (ii) reproduce the nonlinear regression to obtain the scheduling policies presented in the paper or generate their own scheduling policies with their own generated distribution, and (iii) reproduce the dynamic scheduling experiments results presented in the paper.

A.2 Description

A.2.1 Check-list (artifact meta information).

- **Program: (i) Prototypes to generate the distribution $score(p, q, r)$, (ii) prototypes for the nonlinear function enumeration and nonlinear regression and (iii) the dynamic scheduling experiments prototypes.**
- **Compilation: GCC**
- **Data set: Workload models and real workload traces obtained from the Parallel Workload Archive: http://www.cs.huji.ac.il/labs/parallel/workload/. See more details below**.
- **Run-time environment: Python 2.7**
- **Hardware: Various x86 or x64 CPUs.**
- **Output: (i) Generated distribution $score(p, q, r)$, (ii) enumerated functions, their coefficients c_1, c_2 and c_3, and their respective fitness score, and (iii) dynamic scheduling experiments results with the four best obtained nonlinear functions presented in the paper.**
- **Experiment workflow: See below.**
- **Experiment customization: See below.**
- **Publicly available?: Yes.**

A.2.2 How software can be obtained. All source material can be downloaded at GitHub http://github.com/hpcsched/gen-sched-policies. The Git repository is structured in three prototypes represented by the following directories:

- `training-data-generator`: Contains all the source material to generate the distribution $score(p, q, r)$;
- `nonlinear-regression`: Contains all the source material to perform the nonlinear function enumeration and nonlinear regression;
- `sched-performance-tester`: Contains all source material to reproduce the dynamic scheduling experiments results present in the paper.

A.2.3 Hardware dependencies. Any modern x86 or x64 CPU is appropriate to execute the prototypes. It is advised, however, that the system has at least 6GB of RAM since some experiments consume high amounts of memory.

A.2.4 Software dependencies. The main software requirements are a Linux distribution (preferably Ubuntu or CentOS), the GCC C compiler and Python 2.7. Below is presented a list with the specific software requirements:

- **Simgrid 3.13 C libraries**: Publicly available at http://gforge.inria.fr/frs/download.php/file/35817/SimGrid-3.13.tar.gz. Installation instructions are available at http://simgrid.gforge.inria.fr/simgrid/3.13/doc/install.html;
- **ScyPy Python libraries**: Download and installation instructions available at https://www.scipy.org/install.html;
- **Matplotlib Python libraries**: Download and installation instructions available at http://matplotlib.org/users/installing.html. Required to reproduce the plots of the paper.

A.2.5 Datasets.

- Lublin and Feitelson workload model publicly available in: http://www.cs.huji.ac.il/labs/parallel/workload/models.html#lublin99
- Tsafrir *et al.* processing time estimate model publicly avilable at: http://www.cs.huji.ac.il/labs/parallel/workload/models.html#tsafrir05
- CEA Curie workload log publicly avilable at: http://www.cs.huji.ac.il/labs/parallel/workload/l_cea_curie/index.html
- ANL Iterpid workload log publicly avilable at: http://www.cs.huji.ac.il/labs/parallel/workload/l_anl_int/index.html
- SDSC Blue workload log publicly avilable at: http://www.cs.huji.ac.il/labs/parallel/workload/l_sdsc_blue/index.html
- CTC SP2 workload log publicly avilable at: http://www.cs.huji.ac.il/labs/parallel/workload/l_ctc_sp2/index.html
- HPC2N workload log publicly avilable at: http://www.cs.huji.ac.il/labs/parallel/workload/l_hpc2n/index.html
- SDSC SP2 workload log publicly avilable at: http://www.cs.huji.ac.il/labs/parallel/workload/l_sdsc_sp2/index.html

A.3 Installation

Most of the source programs of this work are coded in Python and therefore no installation or compilation is necessary. However, the scheduling simulation programs are coded in C and they need to be compiled. Therefore, after the prototypes are downloaded from the repository (see Section A.2.2) and the software dependencies are installed (see Section A.2.4), it is necessary to perform the following commands from the root repository of the prototypes:

```
$ cd training-data-generator
$ make
$ cd ../sched-performance-tester
$ make
```

A.4 Experiment workflow

A.4.1 Workflow 1: Generating the distribution $score(p, q, r)$. To generate your own distribution $score(p, q, r)$ in the same way that it was performed in the paper, once the source codes is properly compiled (see Section A.3), perform the following commands from the root repository of the prototypes:

```
$ cd training-data-generator
$ nohup python generate_simulation_data.py &
```

The `nohup` command starts a background Python process that continuously generates tuples of tasks (S, Q) and simulate the trials of these tuples to obtain the distribution $score(p, q, r)$. It is recommended to leave this process running at least for a couple of days so enough scheduling observations are performed

in the simulations. From the `training-data-generator` directory there are two important directories: The `task-sets` directory contains all the task tuples (S, Q) generated. Each file present in this directory is a CSV file in which each line contain characteristics (`runtime,#processors,submit_time`) of a task. In its turn, the `training-data` directory contains all of the trial score distributions generated. Each file in this directory is a CSV file in which each line represents the observed scheduling behavior (`runtime,#processors,submit_time,score`) of a task. To join all of the trial score distributions to make the distribution $score(p, q, r)$, perform the following command in the same directory of the previous command:

```
$ python gather_data.py
```

This command creates a file called `score-distribution.csv` which contains the distribution $score(p, q, r)$ obtained from the simulations.

A.4.2 Workflow 2: Enumerating nonlinear functions and performing the fit with nonlinear regression. To reproduce the nonlinear regression and obtain the nonlinear functions presented in the paper, from the root directory of the prototypes perfom the following commands:

```
$ cd nonlinear-regression
$ python nlr_scipy_enumerate_functions.py \
        score-distribution.csv
```

The output of this command is the enumerated functions, their coefficients c_1, c_2 and c_3 and their respective fitness value, in decreasing order of $AVGerror$. More details about the output can be found in Section A.5. The `score-distribution.csv` file is the file that we used to obtain the nonlinear functions presented in the paper. This file can be changed with another distribuition $score(p, q, r)$ generated (see Section A.4.1).

A.4.3 Workflow 3: Reproducing the dynamic scheduling experiments results. To reproduce the dynamic scheduling experiments results presented in the paper, perform the following commands from the root repository of the prototypes:

```
$ cd sched-performance-tester
$ python test_all.py
```

The `test_all.py` is a wrapper script that calls all of the Python scrips present in the `sched-performance-tester` directory, performing all the dynamic scheduling experiments presented in the paper and outputs the statistics of the experiments (medians, means and standard deviations) plus the resulting plot similar to the ones presented in the paper. Alternatively, it's possible to execute each Python script present in the `sched-performance-tester` directory individually to perform only a specific experiment.

A.5 Evaluation and expected result

A.5.1 Generating the distribution $score(p, q, r)$ output. The output of this part is a CSV file named `score-distribution.csv` containing the observations of the scheduling behavior captured in the simulations. Each line of this file contains the characteristics and the resulting scheduling behavior (`runtime,#processors,submit_time,score`) of a task. Below is an example of the expected output:

```
50.0,8.0,88224.0,0.0347251055192
3.0,4.0,88302.0,0.0292281817457
7298.0,58.0,88334.0,0.0350921606481
98.0,1.0,88350.0,0.0333329252836
27.0,9.0,88356.0,0.0307780390597
11.0,32.0,88414.0,0.0278089667369
8758.0,8.0,88421.0,0.031821251468
```

A.5.2 Enumerating nonlinear functions and performing the fit with nonlinear regression. The output of this part is the enumerated functions, their coefficients c_1, c_2 and c_3 and their respective $AVGerror$ value, in decreasing order of $AVGerror$. One example of a expected output is presented below:

```
(-0.0155183403 x log10(runtime)) *
(-0.0005149209 x id(#cores)) +
(0.0069596182 x log10(submit)),
AVGerror=0.0052776
```

Note that algebraic equivalent functions can be enumerated and, in this case, their fitness value will be equal.

A.5.3 Reproducing the dynamic scheduling experiments results. The output of this part is the result statistics of the dynamic scheduling experiments performed. For one experiment, an expected output is the following:

```
Performing scheduling performance test for
the workload trace lublin_256.
Configuration:
Using actual runtimes, backfilling disabled
Experiment Statistics:
Medians:
FCFS=5846.87 WFP=3630.67 UNI=1799.74 SPT=943.59
F4=583.89 F3=89.94 F2=29.66 F1=29.58

Means:
FCFS=6194.92 WFP=3716.46 UNI=2336.92 SPT=1415.13
F4=721.77 F3=593.65 F2=143.40 F1=134.93

Standard Deviations:
FCFS=3321.45 WFP=2908.26 UNI=2050.01 SPT=1539.56
F4=700.10 F3=853.46 F2=227.07 F1=264.51
Boxplot saved in file plots/model_256_r.pdf
```

As mentioned in the output, A boxplot similar with the ones presented in the paper with the respective results is generated and stored in the `plots` subdirectory.