High Performance Computing Scheduler Parameter Optimization using Simulation and Regression

SCOTT HUTCHISON, Kansas State University, USA

DANIEL ANDRESEN (ADVISOR), Kansas State University, USA

WILLIAM HSU (ADVISOR), Kansas State University, USA

When a job is submitted to a High Performance Computing cluster, a scheduling application such as SLURM, PBS, or LoadLeveler will allocate resources for the execution of that job. HPC administrators can customize or optimize the behavior of these scheduling applications by configuring the scheduler's parameters. Although there are default values for these scheduling parameters provided by their maintainers, it is unclear which values would be optimal for a particular HPC system running the types of jobs its users typically submit. Using over 37,000 jobs from historic job log data from Kansas State University's High Performance Computing cluster, this research uses a SLURM simulator to execute over 90,000 scheduler simulations requiring over 840,000 compute hours along with gradient boosted tree regression to predict an optimal set of scheduler configuration parameters which results in a 79% decrease in the average job queue time when compared with the default scheduler parameters.

1 INTRODUCTION

When a job is submitted to a High Performance Computing (HPC) cluster, a scheduling application such as SLURM handles the allocation of HPC resources to the job's requirements as specified by the submitter. The behavior of these scheduling applications is highly customizable by HPC administrators through the use of the scheduler configuration parameters. Although these schedulers typically provide default values for the parameters controlling scheduler behavior, it is unclear if these default parameter values are optimized for a particular HPC system and/or the types of jobs usually submitted to the HPC by its users. This research uses Gradient Boosted Tree Regression (GBTR) [2] and a scheduler simulator in order to find optimal parameter settings for the SLURM scheduler.

2 METHODOLOGY

SLURM has hundreds of configuration parameters which modify the behavior of the scheduling algorithm. From the SLURM configuration documentation [1], eleven different scheduler parameters were chosen for investigation subjectively based on their perceived impact on scheduling behavior as depicted in Table 1. Values were chosen for each of these parameters around the default value, and a fully exhaustive set of nearly 80,000 combinations of parameters were generated. Detailed descriptions of the parameters can be found in the SLURM documentation at https://slurm.schedmd.com/slurm.conf.html. The training data set consisted of 18,523 jobs from one week in September 2018 from HPC log data of actual jobs submitted to the HPC. The training jobs were simulated using the SLURM simulator tool [4], and each simulation used a different set of SLURM configuration parameters. The queue time for each job was calculated as the number of seconds from when a job was submitted until it began execution. The average job queue time for all jobs was calculated, and this data was used to train a GBTR model in PySpark [3] using five fold cross validation which predicted the average job queue time for a unique set of scheduler parameters.

Next, ranges of values were chosen for each scheduler parameters as laid out in Table 1. The GBTR model was used to predict the average job queue time of over 12.4 billion different possible combinations of scheduler parameters. The validation data set consisted of 18,867 jobs from one week in April 2018. The best 10,000 parameters predicted by the

2022. Manuscript submitted to ACM

GBTR model to have the lowest average queue times were then simulated on the validation data using the SLURM simulator. These simulations were evaluated using the simulated average job queue time, and compared to a simulation of the validation data using the default values for these SLURM parameters.

Table 1. SLURM Parameters Investigated with Recommendations.

Scheduler Parameter	Test Values	Prediction Ranges	Experimental Recommendation
bf_continue	True, **False**	True, False	True. **Opposite of the default**
bf_interval	10, **30**, 3000	1-10800 by 1000's	Lower is better. **default too low.**
bf_max_time	10, **30**, 300	1-3600 by 500's	Lower is better. Default is ok.
bf_resolution	20, **60**, 360	1-3600 by 500's	Lower is better. Default is ok.
bf_running_job_reserve	True, **False**	True, False	False, Default is good.
bf_yield_interval	0.5, **2**, 9	1-10 by 3's	Lower is better. Default is good.
default_queue_depth	30, **100**, 3000	1-17000 by 3000's	Lower is better. Default is good.
sched_interval	20, **60**, 180	20-2000 by 400's	Inconclusive.
sched_min_interval	0, **2**, 6	2-200 by 20's	Inconclusive.
bf_max_job_test	100, **500**, 5000	1-1,000,000 by 100,000's	Inconclusive.
bf_yield_sleep	2, **5**, 45	1-10 by 3's	Inconclusive.

3 RESULTS AND DISCUSSION

Using five fold cross validation, GBTR model achieved a root mean squared error of 248.42 seconds when predicting the average job queue time on the training data. On the validation data, the default parameter setting produced an average queue time of 12,970 seconds. The lowest average queue time for the predicted best parameter settings was 2,719 seconds, **a decrease of 79%**. The average of all 10,000 predicted best parameter settings average queue time was 6,249 seconds, **a decrease of 51.8%**. **In all simulations**, the predicted best 10,000 jobs average queue time was lower than the average queue time using the default SLURM parameter settings. Figure 1 depicts the HPC CPU allocation for both the default parameter values and the best performing prediction. The scheduler simulation using recommended settings better utilizes HPC resources and finished scheduling all jobs earlier. The roughly 90,000 simulations required were executed in parallel and required over 840,000 compute hours.

Generalizing the recommendations for the SLURM configuration parameters yielded recommendations and insights for the eleven SLURM scheduler parameters under investigation for our HPC system and the types of jobs it typically runs. The summary of the recommendations for scheduler parameters can be found in Table 1.

4 CONCLUSION

Using a SLURM simulator and GBTR provided the ability to quickly experiment and make predictions in search of optimal scheduler configuration parameters without interrupting access to live HPC resources. This technique can allow HPC administrators to better optimize their HPC scheduler parameters for their particular hardware and the types jobs which are often run on their HPC systems.

REFERENCES

[1] [n. d.]. Slurm Documentation. https://slurm.schedmd.com/slurm.conf.html, accessed 2022-08-04.
[2] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

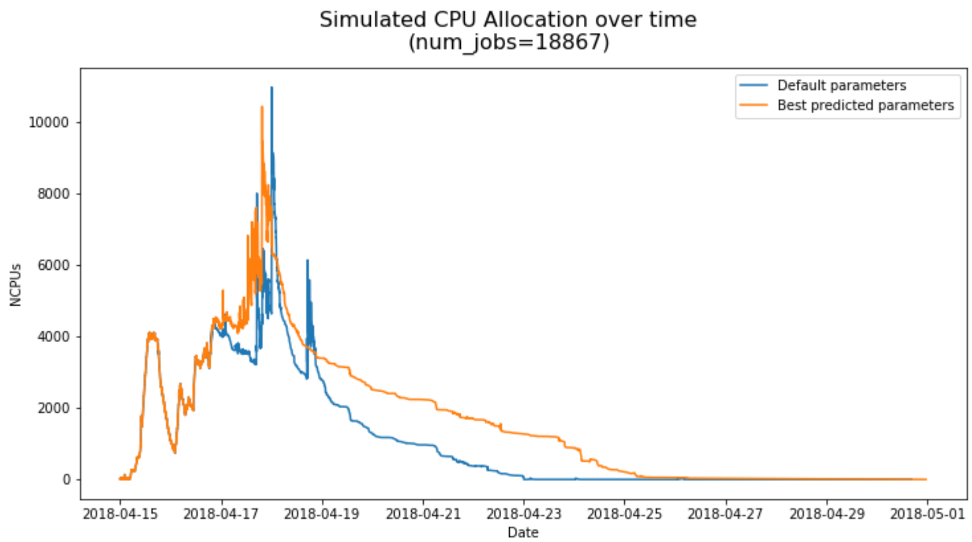

Fig. 1. The best recommended scheduler parameters settings better utilize HPC resources than the default parameter settings resulting in jobs being scheduled earlier.

[3] Xiangrui Meng, Joseph Bradley, Burak Yavuz, Evan Sparks, Shivaram Venkataraman, Davies Liu, Jeremy Freeman, DB Tsai, Manish Amde, Sean Owen, et al. 2016. Mllib: Machine learning in apache spark. *The Journal of Machine Learning Research* 17, 1 (2016), 1235–1241.

[4] Nikolay A Simakov, Martins D Innus, Matthew D Jones, Robert L DeLeon, Joseph P White, Steven M Gallo, Abani K Patra, and Thomas R Furlani. 2017. A slurm simulator: Implementation and parametric analysis. In *International Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems*. Springer, 197–217.