Usage Report for Fiscal Year 2019

Project Title:

Research and Development of system software for high performance big data applications

Name: Tonmoy Dey (1), ○Kento Sato (2), Jian Guo (2), Bogdan Nicolae (3), Jens Domke(2), Weikuan Yu (1), Franck Cappello (3), Kathryn Mohror (4)

Laboratory at RIKEN:
(1) Florida State University
(2) RIKEN Center for Computational Science, High Performance Big Data Research Team
(3) Argonne National Laboratory
(4) Lawrence Livermore National Laboratory

1. Background and purpose of the project, relationship of the project with other projects

With the emergence of fast local storage, multi-level checkpointing (MLC) has become a common approach for efficient checkpointing. However, since multi-level checkpoint/restart is one of major I/O workloads in HPC systems, to use multi-level checkpointing efficiently, the system needs to use optimized checkpoint/restart configuration. Current approaches, namely modeling and simulation, are either inaccurate or slow in determining the optimized configuration for a large-scale system.

In this project, our objective was to determine the optimized interval faster than the simulation approach without losing much of its accuracy. We present an idea to achieve this by combining the simulation approach with artificial intelligence (AI) techniques, such as machine learning and neural networks, so that we can leverage the accuracy of the simulation approach without spending a lot of time on simulating different CR parameters to determine the optimized configurations.

2. Specific usage status of the system and calculation method

For our approach, we generated a limited amount of data using the simulator which provides us with optimized interval and count for the provided CR configurations and we use that data to determine the optimized count and interval for other CR configurations using the AI techniques. We run the simulation with a specific configuration in HOKUSAI system. The simulator we have developed not only provides the elapsed time and efficiency for a system but also performs simulation across different settings of the checkpoint system to determine the optimal checkpoint count for each level of the multi-level checkpoint system. The optimized configurations are obtained by modulating the checkpoint configuration values starting with checkpoint interval. The simulator initially performs optimization of the checkpoint interval by comparing the current efficiency with the efficiency of the previous configuration until the maxima are obtained. The same procedure is also followed for both checkpoint count of level 1 and level 2. After the peaks for all the configuration parameters are captured by the simulator, the user is provided with the optimized configuration for the given system based on the overhead, restart time, and failure rates.

Once a significant amount of data is collected from the simulator, the information is passed on to different machine learning and neural network models to predict the optimized checkpoint configuration for other systems with different over-head, restart time, and failure rates. The machine learning models were training and predicting the optimized checkpoint configuration in HOKUSAI system.

3. Result

Two data sets were generated for evaluation using the CR simulator. The training data was simulated by providing overhead, latency, and restart time that closely matched some of the Top500 systems. We used the base failure rate of 2.1^{6} and 4.1^{7} failures/sec for Level-1 (L1) and Level-2 (L2) respectively and explored the impact of these failure rates by increasing them up to 50X the base value. The test data set was simulated using data mutually exclusive to the training data set but within the parameter space of the training data set.

Our random forest model could predict optimized checkpoint count with 99.365% accuracy when compared to the simulated result. For predicting the checkpoint interval with its values ranging up to 9960 seconds, both machine learning models could determine the optimized interval with a mean absolute error of 49.5 seconds, our baseline neural network predicted optimized interval with a mean absolute error of 116.63 seconds. However, daisy chaining shows a 40% performance improvement over the baseline design followed by a further 40% improvement with parameter reduction. The optimized neural network performs 18% better than the machine learning models with a mean absolute error 40.7 seconds. With majority of our dataset in the region 1000 - 8000 seconds, this converts to 0.5% - 4% mean error from the simulated values of the optimized interval.

4. Conclusion

In this project, we present an idea to combine simulation approach with AI techniques such as machine learning and neural network so that we can reduce the time taken to determine the optimized parameter values of checkpoint interval and checkpoint count. With our approach and design optimizations, we demonstrate that our models can predict the optimized parameter values with minimal error when compared to the simulation approach. For our future work, we will optimize the checkpoint levels and encoding for different CR configurations.

5. Schedule and prospect for the future

With our approach and design optimizations, we show that our models can predict the optimized parameter values when trained with the simulation approach. We also show that using more advanced deep neural network techniques can be used to improve the performance of neural network over the machine learning models up to 50%. which can further be used in future research works to optimize the checkpoint restart configuration of large-scale systems.

6. If no job was executed, specify the reason.

Fiscal Year 2019 List of Publications Resulting from the Use of the supercomputer

[Paper accepted by a journal]

[Conference Proceedings]

[Oral presentation]

[Poster presentation]

Dey T, Sato K, Guo J, et al. Optimizing Asynchronous Multi-level Checkpoint/Restart Configurations with Machine Learning[J], SC, 2019

[Others (Book, Press release, etc.)]