Accepted Manuscript

Improving the effectiveness of burst buffers for big data processing in HPC systems with Eley

Orcun Yildiz, Amelie Chi Zhou, Shadi Ibrahim

PII: S0167-739X(17)31730-2
DOI: https://doi.org/10.1016/j.future.2018.03.029
Reference: FUTURE 4041

To appear in: *Future Generation Computer Systems*

Received date : 31 July 2017
Revised date : 20 February 2018
Accepted date : 14 March 2018

Please cite this article as: O. Yildiz, A.C. Zhou, S. Ibrahim, Improving the effectiveness of burst buffers for big data processing in HPC systems with Eley, *Future Generation Computer Systems* (2018), https://doi.org/10.1016/j.future.2018.03.029

Available online at www.sciencedirect.com

ScienceDirect

Future Generation Computer Systems 00 (2018) 1–17

Future Generation Computer Systems

Improving the Effectiveness of Burst Buffers for Big Data Processing in HPC Systems with Eley

Orcun Yildiz[1], Amelie Chi Zhou[2], Shadi Ibrahim[3*]

[1] *Inria, Univ Rennes, CNRS, IRISA, Rennes, France*

[2] *Shenzhen University, China*

[3] *Inria, IMT Atlantique, LS2N, Nantes, France*

Abstract

Burst Buffer is an effective solution for reducing the data transfer time and the I/O interference in HPC systems. Extending Burst Buffers (BBs) to handle Big Data applications is challenging because BBs must account for the large data inputs of Big Data applications and the Quality-of-Service (QoS) of HPC applications – which are considered as first-class citizens in HPC systems. Existing BBs focus on only intermediate data of Big Data applications and incur a high performance degradation of both Big Data and HPC applications. We present *Eley*, a burst buffer solution that helps to accelerate the performance of Big Data applications while guaranteeing the QoS of HPC applications. To achieve this goal, *Eley* embraces interference-aware prefetching technique that makes reading data input faster while introducing low interference for HPC applications. Evaluations using a wide range of Big Data and HPC applications demonstrate that *Eley* improves the performance of Big Data applications by up to 30% compared to existing BBs while maintaining the QoS of HPC applications.

Keywords: HPC; MapReduce; Big Data; Parallel File Systems; Burst Buffers; Interference; Prefetch.

1. Introduction

With the arrival of the era of Big Data, we have witnessed the emergence of a new scalable data management and processing paradigm through the MapReduce model [1, 2] for data-intensive processing. MapReduce is adopted in both industry and academia due to its simplicity, transparent fault tolerance and scalability. For instance, Carnegie Mellon University is using MapReduce clusters for data analysis on several scientific domains including computational astrophysics, computational neurolinguistics, natural language processing and social networking analysis [3]. The success of MapReduce led to the emergence of new types of applications (e.g., stream data processing, graph processing, analysis of large scale simulation data) where obtaining timely and accurate responses is a must. However, commodity machines-based Big Data environments reach their limit due to being failure-prone and providing moderate performances.

Why Big Data processing on HPC Systems? High performance computing (HPC) systems are known for providing the maximum computing capability. They are equipped with high-speed networks, thousands of nodes with many

*Corresponding author

Email address: shadi.ibrahim@inria.fr (Shadi Ibrahim[3])