PABS 2015: 1st Workshop on Performance Analysis of Big Data Systems

Rekha Singhal
Tata Innovation Lab
Mumbai, India
rekha.singhal@tcs.com

Dheeraj Chahal
Tata Innovation Lab
Mumbai, India
d.chahal@tcs.com

ABSTRACT

The first ACM international workshop on performance analysis of big data system is held in Austin, Texas, USA on February 1, 2015 and co-located with the ACM fifth International Conference on Performance Engineering (ICPE). The main objective of the workshop is to discuss the performance challenges imposed by big data systems and the different state-of-the-art solutions proposed to overcome these challenges. The workshop aims at providing a platform for scientific researchers, academicians and practitioners to discuss techniques, models, benchmarks, tools and experiences while dealing with performance issues in big data systems. We have constructed an exciting program of one big data expert keynote talk, one invited talk and two refereed papers that will give participants a full dose of emerging research.

Categories and Subject Descriptors

A. General Literature

General Terms

Algorithms, Performance, Architectures

Keywords

Big Data; Performance; Analysis; Prediction; Case study

1. INTRODUCTION

Big data systems deal with velocity, variety, volume and veracity of the application data. We witness an explosive growth in the complexity, diversity, number of deployments and capabilities of big data processing systems such as Map-Reduce, Cassandra, Big Table, HPCC, Hyracks, Dryad, Pregel and Mongo DB. The big data system may use new operating system designs, advanced data processing algorithms, parallelization of application, high performance architectures and clusters to improve the performance. Looking at the volume of data to mine, and complex architectures, one may need to analyze, identify or predict bottlenecks to optimize the system and improve its performance.

The workshop on performance analysis of big data systems (PABS) aims at providing a platform for scientific researchers, academicians and practitioners to discuss techniques, models, benchmarks, tools and experiences while dealing with performance issues in big data systems. The primary objective is to discuss performance bottlenecks and improvements during big data analysis using different paradigms, architectures and technologies such as Map-Reduce, MPP, Big Table, NOSQL, graph based models (e.g. Pregel, giraph) and any other new upcoming paradigms. We propose to use this platform as an opportunity to discuss systems, architectures, tools, and optimization algorithms that are parallel in nature and hence make use of advancements to improve the system performance. This workshop shall focus on the performance challenges imposed by big data systems and on the different state-of-the-art solutions proposed to overcome these challenges.

2. TOPICS OF INTEREST

All novel performance analysis or prediction techniques, benchmarks, architectures, models and tools for data-intensive computing system for optimizing application performance on cutting-edge high performance solutions are of interest to the workshop. Examples of topics include but not limited to:

- Performance analysis and optimization of Big data systems and technologies.
- Case studies/ Benchmarks to optimize/evaluate performance of Big data applications/systems and Big data workload characterizations.
- Tools or models to identify performance bottlenecks and /or predict performance metrics in Big data
- Performance analysis while querying, visualization and processing of large network datasets on clusters of multicore, many core processors, and accelerators.
- Performance issues in heterogeneous computing for Big data architectures.
- Analysis of Big data applications in science, engineering, finance, business, healthcare and telecommunication etc.
- Data structure and algorithms for performance optimizations in big data systems.

3. WORKSHOP FORMAT

The workshop is scheduled for half a day. We have domain expert key note speaker Prof. D.K. Panda from Ohio State University, USA. He is well known for his contribution in the field of “Big data Performance Accelerators”. Key note lecture shall be of 45 minute duration with 15 min for Q/A. The key note will be followed by an invited talk by Prof. Amy W. Apon from Clemson University, USA. Two refereed papers will be presented for 40 minutes duration each including 30 minutes for presentation and 10 minutes for Q/A. Finally, the workshop will be concluded by the chairs.

ICPE'15, Jan. 31 – Feb. 4, 2015, Austin, Texas, USA.
ACM 978-1-4503-3248-4/15/01.
http://dx.doi.org/10.1145/2668930.2688199

4. KEYNOTE TALK

The keynote talk will be given by Prof. Dhabaleswar (DK) Panda from Ohio State University, USA. The title of the talk is “Accelerating Big Data Processing on Modern Clusters”.

Abstract:

Modern clusters are having multi-/many-core architectures, high-performance rdma-enabled interconnects and SSD-based storage devices. Hadoop framework is extensively being used these days for Big Data processing. Spark framework is emerging for real-time analytics. Similarly, Memcached is being used in data centers with Web 2.0 environment. This talk will provide an overview of challenges in accelerating Hadoop, Spark and Memcached on modern clusters. An overview of RDMA-based designs for multiple components of Hadoop (HDFS, MapReduce, RPC and HBase), Spark and Memcached will be presented. Performance benefits of these designs on various cluster configurations will be shown. The talk will also address the need for designing benchmarks using a multi-layered and systematic approach, which can be used to evaluate the performance of these middleware.

Bio:

Dhabaleswar K. (DK) Panda is a Professor of Computer Science and Engineering at the Ohio State University. He has published over 350 papers in major journals and international conferences. Prof. Panda and his research group members have been doing extensive research on modern networking technologies including InfiniBand, High-Speed Ethernet and RDMA over Converged Enhanced Ethernet (RoCE). The MVAPICH2 (High Performance MPI over InfiniBand, iWARP and RoCE) and MVAPICH2-X software libraries, developed by his research group (http://mvapich.cse.ohio-state.edu), are currently being used by more than 2,250 organizations worldwide (in 74 countries). This software has enabled several InfiniBand clusters to get into the latest TOP500 ranking during the last decade. More than 226,000 downloads of this software have taken place from the project's website alone. The new RDMA-enabled Apache Hadoop package, RDMA-enabled Memcached package, and OSU HiBD benchmarks (OHB) are publicly available from the High-Performance Big Data project site (http://hibd.cse.ohio-state.edu). Prof. Panda's research has been supported by funding from US National Science Foundation, US Department of Energy, and several industry including Intel, Cisco, Cray, SUN, Mellanox, QLogic, NVIDIA and NetApp. He is an IEEE Fellow and a member of ACM. More details about Prof. Panda are available at http://www.cse.ohio-state.edu/~panda

5. WORKSHOP ORGANIZERS

Dr. Rekha Singhal has 20 years of research and teaching experience. Currently she is working as Senior Scientist with TCS Innovation Lab and leading Big Data Performance Modelling and Analysis initiatives. She has numerous international publications and patents to her credit. One of the products, Revival 2000, developed under her guidance had received NASSCOM Technology award. Her research interests are Big Data System Performance, Query Performance Prediction, Database Performance Modelling, IP Storage Area Networks, Distributed Systems and Health IT. She is Ph.D and M.tech from IIT Delhi

Dr. Dheeraj Chahal is a Consultant and research team lead with Performance Engineering group at TCS innovations lab, Mumbai. Prior to joining TCS, he worked as Staff Software Engineer with HPC team at IBM, Bangalore and successfully conducted workshop on performance engineering at HiPC 2012 and 2013 (http://www.hipc.org/hipc2013/workshops.php). Dheeraj holds a PhD degree in Computer Science from Clemson University, SC, USA.

6. PROGRAM COMMITTEE

- Amitabha Bagchi , IITD, India
- Amy. W. Apon, Clemson University, USA
- Arno Jacobsen, University of Toronto, Canada
- Bojan Cukic, UNC, USA
- Dhableshwar Panda, Ohio State University, USA
- Gautam Shroff, TCS Innovation Lab, India
- Henrique Madeira, University of Coimbra, Portugal
- Kishor Trivedi, Duke University, USA
- Jeff Ullman, Stanford University and Gradiance, USA
- Narendra Bhandari, Intel, India
- Rajesh Mansharamani, CMG India
- Saumil Merchant, Shell, India
- Sebastien Goasguen, Citrix, Switzerland
- Steven J Stuart, Clemson University, USA
- Veena Mendiratta, Alcatel-Lucent, USA
- Vikram Narayana, George Washington University, USA
- Zia Saquib, CDAC, India

7. ACKNOWLEDGEMENTS

We thank Tata Consultancy Services for supporting our initiatives.