Components and Rationale of a Big Data Toolkit Spanning HPC, Grid, Edge and Cloud Computing

Geoffrey Fox
Indiana University, Department of Intelligent Systems Engineering
Bloomington, Indiana, USA
gcf@iu.edu

ABSTRACT

We look again at Big Data Programming environments such as Hadoop, Spark, Flink, Heron, Pregel; HPC concepts such as MPI and Asynchronous Many-Task runtimes and Cloud/Grid/Edge ideas such as event-driven computing, serverless computing, workflow, and Services. These cross many research communities including distributed systems, databases, cyberphysical systems and parallel computing which sometimes have inconsistent worldviews. There are many common capabilities across these systems which are often implemented differently in each packaged environment. For example, communication can be bulk synchronous processing or data flow; scheduling can be dynamic or static; state and fault-tolerance can have different models; execution and data can be streaming or batch, distributed or local. We suggest that one can usefully build a toolkit (called Twister2 by us) that supports these different choices and allows fruitful customization for each application area. We illustrate the design of Twister2 by several point studies. We stress the many open questions in very traditional areas including scheduling, messaging and checkpointing.

CCS Concepts/ACM Classifiers

• Computer systems organization~Cloud computing *• Computer systems organization~Grid computing* *• Computer systems organization~Sensor networks* **• Software and its engineering~Data flow architectures** **• Software and its engineering~Publish-subscribe / event-based architectures** • Software and its engineering~Message oriented middleware

Author Keywords

Cloud Computing; MapReduce; MPI; HPC; Dataflow; Edge Computing; Global Machine Learning

BIOGRAPHY

Fox received a Ph.D. in Theoretical Physics from Cambridge University where he was Senior Wrangler. He is now a distinguished professor of Engineering, Computing, and Physics at Indiana University where he is director of the Digital Science Center, and both Department Chair and Interim Associate Dean for Intelligent Systems Engineering at the School of Informatics, Computing, and Engineering. He previously held positions at Caltech, Syracuse University, and Florida State University after being a postdoc at the Institute for Advanced Study at Princeton, Lawrence Berkeley Laboratory, and Peterhouse College Cambridge. He has supervised the Ph.D. of 70 students and published around 1300 papers (over 450 with at least 10 citations) in physics and computing with an hindex of 76 and over 31000 citations. He is a Fellow of APS (Physics) and ACM (Computing) and works on the interdisciplinary interface between computing and applications.

REFERENCES

1.Supun Kamburugamuve, Kannan Govindarajan, Pulasthi Wickramasinghe, Vibhatha Abeykoon, Geoffrey Fox, "Twister2: Design of a Big Data Toolkit" Technical Report, October 7 2017 http://dsc.soic.indiana.edu/publications/twister2_design_big_data_toolkit.pdf

2. Geoffrey Fox, Judy Qiu, Shantenu Jha, Saliya Ekanayake, and Supun Kamburugamuve, "Big Data, Simulations and HPC Convergence" Workshop on Big Data Benchmarking, New Delhi India, December 14, 2015. Springer Lecture Notes in Computer Science LNCS 10044. http://dx.doi.org/10.1007/978-3-319-49748-8_1

3. Geoffrey C. Fox, Vatche Ishakian, Vinod Muthusamy, Aleksander Slominski, "Status of Serverless Computing and Function-as-a-Service(FaaS) in Industry and Research", Workshop on Serverless Computing (WoSC) Atlanta, June 5 2017 https://arxiv.org/abs/1708.08028

4.Supun Kamburugamuve, Pulasthi Wickramasinghe, Saliya Ekanayake, Geoffrey C Fox, "Anatomy of machine learning algorithm implementations in MPI, Spark, and Flink", The International Journal of High Performance Computing Applications, July 2, 2017 https://doi.org/10.1177/1094342017712976

5.http://www.spidal.org/ Middleware and High-Performance Analytics Libraries for Scalable Data Science SPIDAL project

6.High Performance Computing Enhanced Apache Big Data Stack http://hpc-abds.org/kaleidoscope/

UCC'17, December 5-8, 2017, Austin, Texas, USA.

ACM ISBN 978-1-4503-5149-2/17/12.
DOI: https://doi.org/10.1145/3126858.3133312