Integration of HPC, Big Data Analytics and the Software Ecosystem

Tutorial at 4th International Winter School on Big Data, Timişoara, Romania, January 22-26, 2018

Geoffrey Fox gcfexchange@gmail.com

Location of material:

https://drive.google.com/drive/folders/126NLJPTYnjzmzm_iHmv0HrE9v2NKVm--?usp=sharing

This link

https://docs.google.com/document/d/1jMfzWAvhlnynBPfQoEZSUR45x_L5zfyvY1V5A6g9kzs/edit?usp=sharing

Short Panel Talk on Requirements and Jobs

https://drive.google.com/open?id=16K961AY6SRMRvCbaNFtwAUT2aFMslRG6

Overview of Tutorial

Overview of Entire Tutorial

https://drive.google.com/open?id=1Gf2PFE52RX9NE-0y36YemV2V5O6lTnqm

General principles

Presentation on HPC-ABDS, Cloud Status, and Ogres Application Analysis; HPC-Cloud and Data-Simulation convergence

https://drive.google.com/open?id=1YOXWtLc2xHJOoidDEGH2Cgaifm7uJ__8

Twister2 Tutorial -- Initial Version

- The overview talk is https://drive.google.com/open?id=1FCOKLH3dBhCutNoTK_gUTvlJ-XjSEckd
- This includes a first hand experience on Twister2
- Download and install Twister2
- Run few examples (Working on the documentation)
  - Streaming word count

    - Batch word count
    - Install - https://github.com/DSC-SPIDAL/twister2/blob/master/INSTALL.md
    - Examples - https://github.com/DSC-SPIDAL/twister2/blob/master/docs/examples.md

Harp-DAAL Tutorial (using Docker)

- The Harp-DAAL overview technical talk is https://drive.google.com/open?id=1r9gzmT__3Xs11uo7n1XW_Zyd3wsPTnjz
- The Tutorial material can be found at
    - Video: https://www.youtube.com/watch?v=prfPewgMrRQ
    - This is built around a standalone Docker image (available) and covers Kmeans in detail
    - https://github.com/DSC-SPIDAL/harp/blob/master/Hands-on-kmeans.md
    - https://github.com/DSC-SPIDAL/harp/blob/master/Hands-on-NaiveBayes.md
    - https://github.com/DSC-SPIDAL/harp/blob/master/Hands-on-MFSGD.md
- The Harp-DAAL tutorial talk is https://drive.google.com/open?id=1F3CqdQS-nDbR4dfOpKlF0Hqh5sKdg5fB
- See SC17 tutorial https://dexterrules.github.io/SC-Demo-17/SC-Demo.html
    - There is video on use of Google Cloud https://drive.google.com/open?id=1wl_4kLXDqGXJFYf4qtam4Cc45wRhY1om
    - Compared to this site, there are few additional instructions in the direct instructions, that can help a user when they are running on a resource constrained environment like a laptop. These instructions are not present in the Tutorial website.
    - We can still use the interactive web site as it contains lots of explanations, examples etc. The instructions in the website are valid but users may encounter some problems because of the resource limitations of a laptop.
    - The Harp-DAAL video only covers the K-Means example. If one can follow that example other two examples are straight forward.
- The SC17 tutorial consists of
    - Setting up docker image that contains the Harp DAAL and Hadoop
    - K-Means algorithm with an exercise on filling the blanks
    - NB algorithm
    - MFSGD algorithm

SPIDAL Tutorial (on Linux -- tested on Ubuntu)

- The overview talk is https://drive.google.com/open?id=1mSZ-vPrnit_wNILMXlUsMqx5tqk3mciK

    - This goes through use of SPIDAL Clustering and Dimension Reduction as well as WebPlotViz online visualization system
    -
- The tutorial consist of Installation of SPIDAL software including openmpi
    - Fungi sequence clustering
    - Pathology data
- The tutorial material can be found at (all the materials are ready, working on the video)
    - Video: https://youtu.be/ZpYFKGYQ1Uk
    - https://dsc-spidal.github.io/tutorials/

Original Abstract

Level: Intermediate

Abstract:

We discuss high performance big data computing that supports hardware, algorithms and software allowing the use of rich functionality of big data systems, such as Apache Hadoop, Spark, HBase, Flink, Heron, and HDFS, on compute architectures ranging from commodity cloud, hybrid HPC cloud, and supercomputer, with possibly customized accelerator (e.g., FPGA, GPU, TPU), having performance and security that scales and fully exploits the specialized features (communication, memory, energy, I/O, accelerator) of each different architecture, for applications ranging over pleasingly parallelizable and mapreduce jobs, to classical machine learning (e.g., random forest, SVM clustering and dimension reduction), deep learning, LDA, and large Graph analysis tasks. We expect this area to be of growing importance and this tutorial covers three aspects of this.

General principles

- We introduce HPC-ABDS, the High-Performance Computing (HPC) enhanced Apache Big Data Stack (ABDS), which uses the major open source Big Data software environment but develops the principles allowing the use of HPC software and hardware to achieve good performance. We present several big data performance studies.
- We introduce the Ogres as an approach to classifying big-data applications and use this to explain problem classes that need particular hardware and software support.
- We present our analysis of the convergence between simulations and big-data applications as well as selected research about managing the convergence between HPC, Cloud, and Edge platforms.

Harp-Daal and SPIDAL (Scalable Parallel Interoperable Data Analytics Library)

- We introduce a novel HPC-Cloud convergence framework, Harp-DAAL and demonstrate that the combination of Big Data (Hadoop) and HPC (a Harp plugin for collective communication and DAAL for computation kernels) can simultaneously achieve productivity and performance on large scale data analytics. Harp is a distributed Java-based framework that orchestrates efficient node synchronization. Harp uses DAAL, Intel's Data Analytics Accelerator Library, for its highly optimized kernels on Intel Haswell and KNL architectures. This way, the high-level interfaces of big data tools can be combined with intra-node fine-grained parallelism that is properly optimized for different HPC nodes.
- Harp-DaaL supports the high performance SPIDAL machine learning library with currently 20 members which are being packaged for wide distribution.
- The tutorial covers both SPIDAL and Harp-DaaL with several examples

Twister2 Big Data Programming environment

- 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 describe status of Twister2 which is an open source project with an Apache 2.0 license. We go through the different parts of Twister2 and how we integrate the ideas present in existing HPC and Big Data systems,
- Twister2 is positioned as an appropriate software environment to support high performance big data computing and includes Harp-DaaL as a critical component to support scalable machine learning.