Tutorial Proposal

MapReduce for Scientist using FutureGrid

Gregor von Laszewski

Assist. Director of Cloud Computing, CGL, PTI, Indiana University

1 Abstract (200 words)

Todays scientific applications deal with the analysis of large amount of data. Scientists often want to be able to reuse existing analysis frameworks in order to be able to focus on the science aspect rather than developing an information technology. MapReduce is one framework that allows analyzing huge datasets using large numbers of compute elements. We will in this tutorial outline the concept of MapReduce, introduce criteria on which applications can successfully use MapReduce, show how to use it on FutureGrid. We will outline limitations and also introduce how to benchmark and improve performance, as well as showing how to setup your own MapReduce environment on FutureGrid.

2 Description

2.1 Overview and Goals of the tutorial (takeaways for the audience)

MapReduce programming model has simplified the implementations of many data parallel applications. The simplicity of the programming model and the quality of services provided by many implementations of MapReduce attract a lot of enthusiasm among parallel computing communities.

The goal of the tutorial is to introduce the participants to the concepts of map reduce. Furthermore we identify what scientific applications can benefit from it. After this introduction in concepts, that we will outline how to use MapReduce on FutureGrid allowing participants to set up their own environment on FG and use MapReduce for their applications. We will demonstrate various different setups and demonstrate Map reduce environments in FG utilizing traditional high-performance compute services, Infrastructure as a Service platforms such as OpenStack, Nimbus, and Eucalyptus. Participants will be able to apply for regular FutureGrid accounts to practically try out our tutorial in additional production services offered after the tutorial is completed.

In contrast to other tutorials, the tutorial will not be limited to the material offered at the conference, but we hope to engage tutorial participants in prolonged activities on the FutureGrid resources.

2.2 Will this tutorial be given somewhere else?

1. We have submitted this tutorial also to ISC in Germany.
2. The tutorial will be available online as part of the FG educational material.

2.3 Length of the tutorial (half-day (3.5 hours) or full-day (7 hours)

The tutorial will be a half day tutorial

2.4 Content level (split into: beginner, intermediate, advanced)

The content level of this tutorial is targeted towards beginners

2.5 Targeted audience (industry, academics, researchers, developers, system administrators)

Our target audience are from academy, industry, researchers that phase scientific problems and want to know more about Map Reduce. In addition it is attractive for those with intermediate knowledge that are looking for MapReduce services offered as part of a testbed.

2.6 Audience prerequisites

Some Java knowledge is of advantage

2.7 Outline of the tutorial

1. Introduction to MapReduce
   a. The Principal behind Map Reduce
   b. Simple examples of MapReduce

2. Software for MapReduce
   a. Hadoop
   b. Twister
   c. Others

3. FutureGrid Services Offering MapReduce
   a. Dynamic Provisioning of MapReduce on FutureGrid
   b. MapReduce Appliance
   c. MapReduce with HPC Queues

    d. MapReduce with OpenStack, Nimbus, and Eucalyptus
    e. Performance considerations and Measurements

4. Scientific Applications
    a. Example application usecases for MapReduce
    b. Example applications that run successful on Map Reduce

2.8 Description of the sections

5. Introduction to MapReduce

   In the first section we will introduce the tutorial participants to the concepts of MapReduce. We will be discussing general principals, and show usecases for which MapReduce is suited.

    a. The Principal behind Map Reduce
    b. Simple examples of MapReduce

6. Software for MapReduce

   In this section we will introduce environments that use MapReduce a basis. To each of them we will showcase an example on how to use the environment. Hadoop

    a. Hadoop
    b. Twister
    c. Others

7. FutureGrid Services Offering MapReduce

   While the previous section introduced general concepts and first examples, this section will introduce the participants to concrete environments available on FutureGrid supporting MapReduce. As each of the environments are available on FutureGrid and the users will be ale to try out the example in a working environment as part of a FutureGrid account that users fro this tutorial will obtain.

    a. Hadoop – Hadoop is one of the de-facto implementations of MapReduce
    b. Twister - Twister provides a set of extensions to the programming model and improvements to its architecture that will expand the applicability of MapReduce to more classes of applications.
        i.

c. Dynamic Provisioning of MapReduce on FutureGrid – users, especially technology developers, may want to deploy their own versions of MapReduce toolkits that provide enhancements or modifications. Our Dynamic provisioned MapReduce environments allow this. This include running MapReduce also in various clouds offered in FutureGrid such as Eucalyptus, Nimbus, and OpenStack

d. MapReduce Appliance – We will discuss how MapReduce can be used as part of a FutureGrid appliance

e. Performance considerations and Measurements – we will discuss how to measure performance of your application and compare them between the different approaches

8. Scientific Applications

In this section we will introduce a selected number of applications for which MapReduce has been successfully applied

a. Which applications are suitable for MapReduce

3 CV

Gregor von Laszewski
Assistant Director of Cloud Computing
Pervasive Technology Institute
Indiana University
2729 E 10th St.
Bloomington IN 47408

3.1 EDUCATION

Sep. 1991 - Nov. 1996: Ph.D., Computer Science, Syracuse University, Syracuse, NY.

Sep. 1990 – Sep. 1991: Graduate Fellowship The Ohio State University as exchange student, Columbus, OH

Sep. 1987 - Nov. 1990: Diploma (M.S., Grade A), Computer Science, University of Bonn, Germany.

Sep. 1984 - Apr. 1987: Pre-Diploma (B.S.), Computer Science, University, of Bonn, Germany.

APPOINTMENTS

Jul. 2009 – present: Indiana University, Bloomington, IN, Pervasive Technology Institute, Assistant Director of Cloud Computing

Aug. 2007 – Jul. 2009: As part of a two year leave of absence from Argonne National Laboratory, Argonne, IL. Scientist, Mathematics and Computer Science Division:

- Rochester Institute of Technology, NY, Director, Service Oriented Cyberinfrastructure Laboratory.

- Rochester Institute of Technology, NY, Associate Professor, Computer Science Department.
- Rochester Institute of Technology, NY, Associate Professor, PhD Program, GCCIS.

Apr. 2002 – Jul. 2007: Argonne National Laboratory, Argonne, IL. Scientist, Mathematics and Computer Science Division.

Jan. 2000 – Jul. 2007: Computation Institute. Fellow, Computation Institute, Chicago, IL, University of Chicago and Argonne National Laboratory.

Nov. 2004 - Jan. 2005: Department of Computer Science and Engineering, Denton, TX, Adjunct Professor, University of North Texas.

Nov. 1998 - Apr. 2002: Argonne National Laboratory, Argonne, IL. Assistant computer scientist, Mathematics and Computer Science Division.

Jan. 2002 - Dec. 2002: Illinois Institute of Technology. Visiting professor/Guest Lecturer, Computer Science Department of the Illinois Institute of Technology.

Nov. 1996 - Nov. 1998: Argonne National Laboratory, Argonne, IL. Postdoctoral researcher, Mathematics and Computer Science Division.

Jun. 1994 - Jan. 1995: NASA Goddard Space Flight Center, Greenbelt, MD, under contract with the University Research Space Agency (USRA). Research assistant.

Feb. 1987 - Sept. 1990: German National Research Center for Information Technology (GMD) (now Frauenhofer Gesellschaft), Bonn. Research assistant.

AWARDS

Mar. 2005: Sandia National Laboratories Recognition Award (as member of the CMCS Team), Livermore, CA, U.S.A.

Nov. 2004: Overall best research poster at Supercomputing 2004, Pittsburgh, PA, U.S.A.

Oct. 2003: Chicago Innovation Award (as member of the Globus Project), Chicago, IL, U.S.A.

Apr. 2003: Department of Energy Outstanding Mentor Award for Undergraduate Education, U.S.A.

Oct. 2001: R&D100 Award (as member of the Globus Project), Chicago, IL, U.S.A.

Nov. 1998: Best of show award in the High Performance Computing Challenge, Supercomputing, 98, Orlando, FL.

1995: University Space Research Agency (USRA) fellowship at Goddard Space Flight Center.

Oct. 1992: Overall best student paper at Supercomputing,Ä'92.

Sep. 1989: Financial assistance by the Department for Education and Research, Germany (due to outstanding grades upon graduation).

PUBLICATIONS

Dr. von Laszewski has published **over 110 academic papers** mostly in the areas of Grid, Cloud, and scientific computing.

EDUCATIONAL ACTIVITIES

Dr. von Laszewski taught classes on in the area of Grid and Distributed Computing at Illinois Institute of technology, Rochester Institute of Technology. Furthermore, he is most proud about his outstanding tutor award from DOE that recognizes his achievements to educate many students. He has given numerous tutorials at conferences and participated. He enjoys and accelerates working with talented students.

COMMUNITY ACTIVITIES

Dr. von Laszewski has in the past served on many program committees in the areas of Cloud Computing, Grids, and Distributed Computing. He has served on several international review committees.

4 Agreement

The presenters agree to release the tutorial notes and other material to the CCGrid attendees. All material will also be available online at the futuregrid.org Web site.