CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE
Concurrency Computat.: Pract. Exper. 2012; **00**:1–2
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cpe

Extensible, Scalable, Interoperable Pilot-Abstractions for MapReduce-based Applications on Clouds and Grids

Melissa Romanus, Andre Luckow, Pradeep Mantha, Shantenu Jha*

Radical Research Group, Rutgers University

SUMMARY

The data generated by scientific applications is experiencing an exponential growth. The efficient use of distributed resources will be the key to making meaningful sense of all of the data produced. In our recent work we showed how MapReduce can be used to efficiently process distributed data across on a distributed set of resources. PilotMapReduce (PMR) [1] is a flexible, infrastructure-independent runtime environment for MapReduce. PMR is based on Pilot-abstractions for compute (Pilot-Jobs) and data (Pilot-Data). Pilot-Jobs are used to couple the map phase computation to the nearby source data, and Pilot-Data is used to move intermediate data using parallel data transfers to the reduce computation phase.
In this work, we show how Pilot abstractions and PMR enable the processing of distributed data across multiple heterogeneous distributed infrastructure, including concurrent usage of clouds and traditional grids/clusters. We further show how PMR can efficiently support different infrastructure and application characteristics, e. g. applications that require iterative MapReduce. We further analyze different resource topologies and MapReduce configuration, such as both hierarchical and iterative MapReduce. In particular, we investigate typical infrastructure trade-offs (e.g. the overhead times in spawning virtual machines, the geographic distribution, etc).

Received . . .

KEY WORDS: MapReduce; Grid Computing; Cloud Computing; K-Means; Data-Intensive; Compute-Intensive

1. INTRODUCTION

Motivation of problem in scaling data intensive applications- Interoperability, Scalability, Extensibility/Flexibility/usability.
- Why is this a problem? Any real application requires this problem to be solved?
- CMS, Atlas generates PBs of data/ day.
Why Iterative MapReduce?
What Application? (k-means?)
Why k-means?
- twister mapreduce used k-means?
- k-means implemented using windows azure
What Infrastructure?
- FutureGrid/ XSEDE (Sierra, Kraken)
- OSG (Need Bliss Condor adaptor - Not yet developed)

*Correspondence to: Journals Production Department, John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK.
†Please ensure that you use the most up to date class file, available from the CPE Home Page at
http://www3.interscience.wiley.com/journal/117946197/grouphome/home.html

- Eucalyptus Cloud (Ashley's contributions) - Get widely distributed instances.
- OpenStack Cloud (Melissa contributions) - Get widely distributed instances.

Experiments -

- Scale data 1GB, 10GB, 100GB, 1000 GB - Scale Resources 1000 cores, 10000 cores, 50,000 cores

Some Research Questions?

Can we say something when to use cloud or When Grid ? Does minimizing queue wait time, distributed nature of Pilot-abstractions motivate Domain Scientists to use freely available Grid resources? Waiting time and cost increases as Number of instances required increase? HOw is it beneficial than Grids?

Why Domain scientists are moving to cloud? Hype? Due to non-availability of necessary simple abstractions to scale applications on Grid?

2. RELATED WORK

3. HADOOP IN THE CLOUD

- Elastic MapReduce
- Hadoop on Azure

REFERENCES

1. Mantha PK, Luckow A, Jha S. Pilot-MapReduce: an extensible and flexible MapReduce implementation for distributed data. *Proceedings of third international workshop on MapReduce and its Applications*, MapReduce '12, ACM: New York, NY, USA, 2012; 17–24, doi:http://doi.acm.org/10.1145/2287016.2287020.

DOI: 10.1002/cpe