Modeling and Control of MapReduce Systems
Mihaly Berekmeri, Sara Bouchenak, Bogdan Robu, Nicolas Marchand

To cite this version:
Mihaly Berekmeri, Sara Bouchenak, Bogdan Robu, Nicolas Marchand. Modeling and Control of MapReduce Systems. Conférence en Parallélisme Architecture et Système, Jan 2013, Grenoble, France. pp.n/c. hal-00940292

HAL Id: hal-00940292
https://hal.archives-ouvertes.fr/hal-00940292
Submitted on 31 Jan 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
I. Context and Problem Statement.

As our world is becoming ever more digitalized, companies are adopting more data-intensive processes. This trend is increasing the need for efficient distributed computing platforms capable of handling bulk data. The traditional database approach to computing is becoming less efficient for large-scale data processing. This is particularly true for unstructured data, which often requires specialized processing methods.

Cloud computing, using the pay-as-you-go model, offers an enticing new approach to deploying such applications. The dominant new perspective emerging in large-scale data processing is the MapReduce programming model.

II. MapReduce

MapReduce is a programming model especially designed for large scale unstructured data processing. Its success lies in its simplicity, scalability, and fault-tolerance. The runtime environments, which build upon the MapReduce framework, automatically take care of:
- data partitioning
- replication, data consistency
- load balancing, and fault tolerance
- task distribution and scheduling

Only two functions have to be implemented: the Map function and the Reduce function. The Map function takes an input set of (key, value) pairs and outputs an intermediate (key, value) pair. The Reduce function groups all the values associated with the same key and combines them to a Reduce function output. The Reduce functions process these values as needed.

IV. Current problems and development perspectives

Scalability and elasticity are a few of the key aspects of Cloud computing. Seemingly, resources scale up infinitely on demand. In the background, cloud computing takes care of such non trivial problems as data security and availability, scalability, virtualization, hardware, and software maintenance.

III. MapReduce

Cloud computing is the next milestone of IT evolution. It incorporates the long standing dream of providing computing as a service over the internet. Cloud computing can offer many computing forms as a service such as hardware, platform, software, storage. Infrastructure as a service is one of the most widely spread service models at the moment and it offers services as a service. Platform as a service models offer computing frameworks such as programming models and languages, operating systems. Software as a service provides an environment for running end user applications in the cloud.

V. Our goals

I. MapReduce Dynamic Model
- Transfer the problems into the control theory framework
- The model has to ease the definition of service level objectives
- Account for uncertainties
- Include faults

II. MapReduce Control Laws
- Provide guarantees in terms of performance, dependability and cost
- Optimize the system configuration
- Minimize deployment costs
- Handle multitenancy
- Test it on real life systems, such as AmazonEc2, Grid5000