Dynamic MapReduce Clusters on Demand

Bogdan Ghiţ, Nezih Yigitbasi, and Dick Epema
{b.i.ghit, m.n.yigitbasi, d.h.j.epema}@tudelft.nl

Multiple MapReduce Clusters

Why multiple MapReduce clusters?
- Performance isolation
- Data isolation
- Failure isolation
- Version isolation

Two Types of Isolation
- Driven by the infrastructure
  - intra-cluster isolation: within the same physical cluster
  - inter-cluster isolation: across multiple physical clusters

Figure 1. Intra-cluster isolation

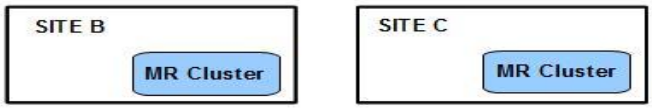
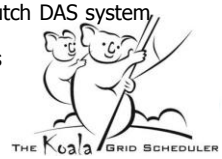


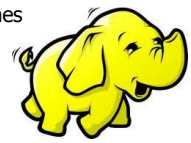
Figure 2. Inter-cluster isolation

Koala and Hadoop Technologies

Koala Grid Scheduler
- Developed at TU Delft and deployed on the Dutch DAS system
- Enables processor and data co-allocation
- Implements placement and scheduling policies
- Modules for different application types
  - e.g. cycle-scavenging, workflows



Hadoop Framework
- Open source implementation of MapReduce
- Scales to clusters of thousands of machines
- Stores data within the HDFS
- Relies on a master-worker paradigm
- Executes tasks close to their data



Koala Resource Management System

MR-Cluster Manager
- Maintains the meta-data of each MR cluster
  - HDFS location, node IP addresses
- Monitors the running jobs within each MR cluster
  - number of tasks per slot
- Dynamically changes the size of a given MR cluster
  - policies for growing or shrinking the cluster

MR-Runner
- The Koala module for scheduling MapReduce jobs
- Relies on SGE to reserve the nodes
- Deploys an MR cluster on the allocated nodes
- Registers the MR cluster with the manager
- Executes a given MapReduce job within an MR cluster

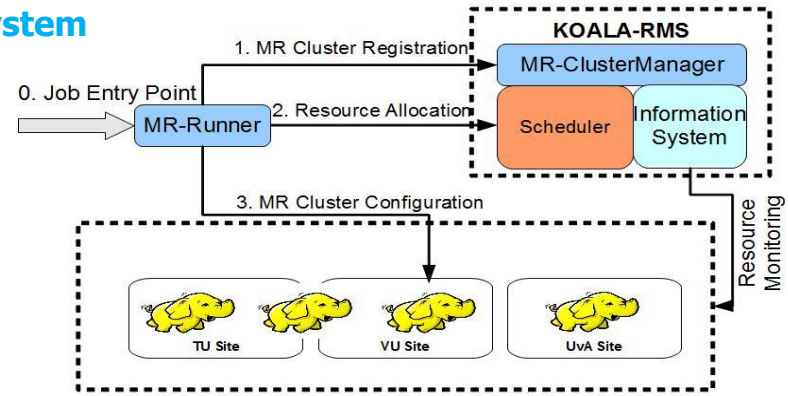


Figure 3. Koala and the MR-Runner

Experiments

Setup
- 40-node Hadoop deployments on DAS-4
- Two types of nodes
  - Core nodes – TaskTracker and DataNode
  - Transient nodes – only TaskTracker
- Two types of applications
  - CPU-intensive – WORDCOUNT
  - IO-intensive – SORT

Results
- WORDCOUNT **scales** well on MR clusters with a large number of transient nodes
- SORT **does not scale** when the number of transient nodes exceeds 20% of the MR cluster size

TODO
- Add support for co-allocation of MR clusters
- Provisioning policies to dynamically re-size the MR clusters

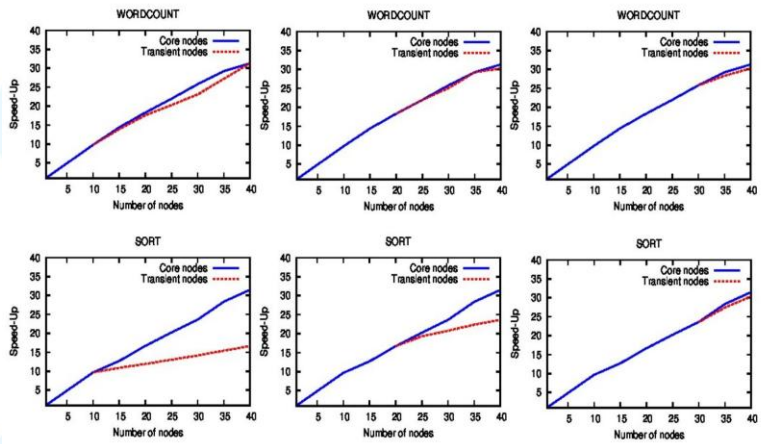


Figure 4. Speed-up on a mix of core and transient nodes. The MR clusters are configured with 10 (left), 20 (middle), and 30 (right) core nodes.

Supported by COMMIT/

http://pds.ewi.tudelft.nl