Hadoop-in-a-Hybrid-Cloud

Luis Russi[1], Carlos R. Senna[1], Edmundo R. M. Madeira[1], Xuan Liu[2], Shuai Zhao[2], Sheyda Mehr[2], and Deep Medhi[2]

[1]Institute of Computing – State University of Campinas – Brazil
[2]University of Missouri–Kansas City – USA

luisrussi@lrc.ic.unicamp.br[1], {csenna, edmundo}@ic.unicamp.br[1]
{xuan.liu, shuai.zhao, skkv6}@mail.umkc.edu[2], dmedhi@umkc.edu[2]

MOTIVATION

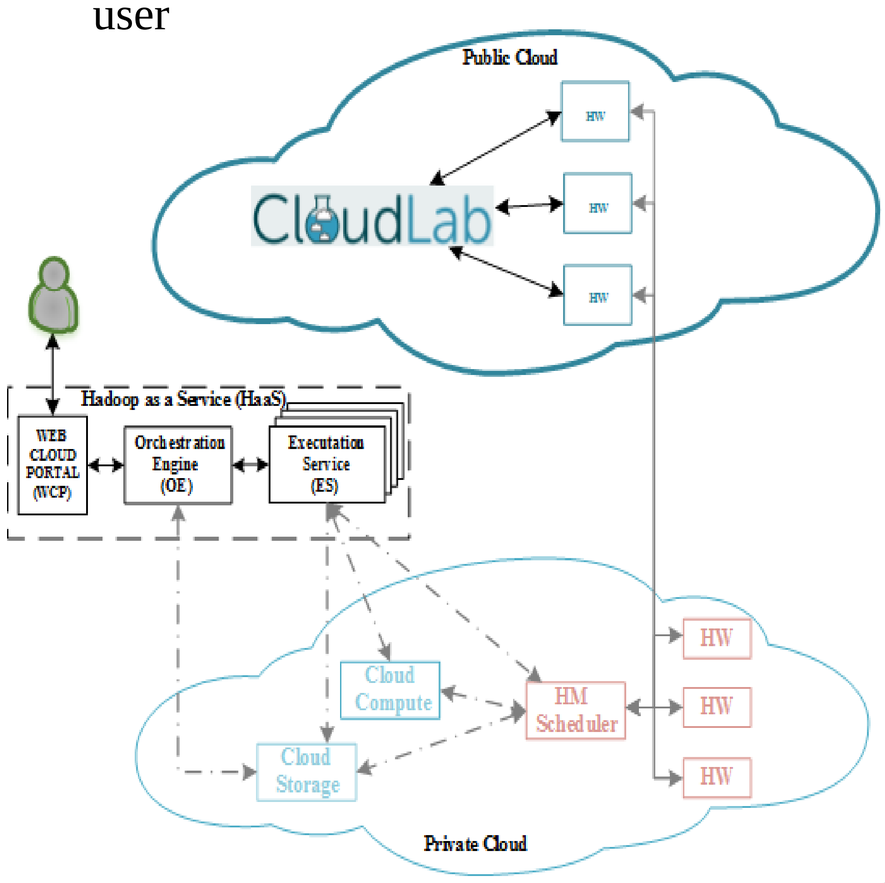
Execution of Hadoop applications in hybrid cloud is not easy!

- Spends time
- Needs technical knowledge
- Continuous evaluation of cloud resources
- On-demand preparation of public cloud resources
- Hybrid cloud requires an appropriate model that combines performance with minimal cost

PROPOSED ARCHITECTURE

An architecture to orchestrate Hadoop applications in hybrid clouds

- Automatic preparation of a cross-domain cluster
- Provisioning files
- Making the results available to the user



WEB CLOUD PORTAL

- User interface
- Management of files (application, data and submission)
- Simple XML-Based submission file
- Organizing the application workspace

Financial Support by the Brazilian Agencies CAPES, FAPESP, and CNPq.

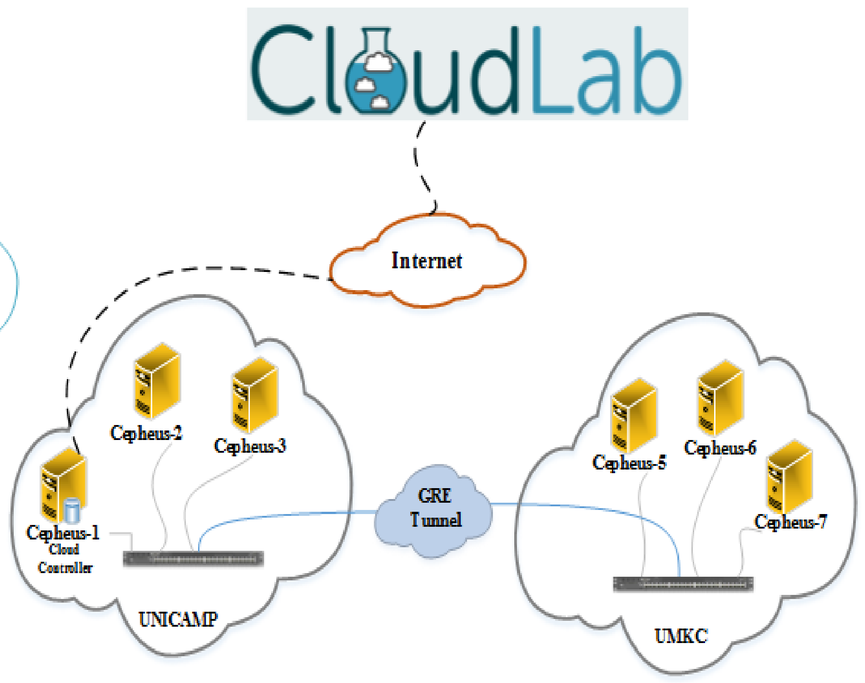
ORCHESTRATION ENGINE

•Prepares working place in the private cloud's storage
•Creates an Execution Service Instance (ESI) that is already associated with this cloud storage area
•Releases the ESI to manage the application execution (asynchronously)
•Copies the resulting files from the cloud storage to the user's working place
•Eliminates ESI and notifies WCP

EXECUTION SERVICE

•ES Instance interacts with the private cloud monitoring system to evaluate the computational resources conditions
•Checks for extra resources from the public cloud
•Automatic Hadoop Cluster preparation
•Makes an accessible copy of the resulting files
•Eliminates all involved VMs
•Notifies the OE about the end of processes
•Monitors all stages of processing

UNICAMP-UMKC HYBRID TESTBED



INITIAL RESULTS

✓Deploy CloudLab OpenStack cluster

✓Deploy and configure virtual machines with Hadoop version 2.6.0

✓Adapt orchestration scripts for running in CloudLab

✓Execute the Wordcount Hadoop application at the cluster

File size (Gb)	Adding to HDFS	Execution time
1	8 s	3m 24s
5	1m 16s	16m 37s
10	3m 40s	33m 4s
20	7m 32s	65m 41s
50	19m 13s	165m 30s
100	43m 52s	334m 13s

CONCLUSION

This work presents an architecture to orchestrate Hadoop applications in hybrid clouds.

- Full Management of Hadoop applications
- Releases the user of repetitive tasks
- Helps in a definition about the best resources
- Supports Hadoop-as-a-Service for complex environments like hybrid clouds.

REFERENCES

[1] The Apache Software Foundation., The Apache Hadoop Project, http://hadoop.apache.org/.
[2] F. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", Commun. ACM, v. 51 n. 1 pp. 107–113, January 2008.
[3] L. F. Bittencourt, E. R. M. Madeira, N. L. S. Fonseca, Scheduling in hybrid clouds, IEEE Communications Magazine, v. 50, pp. 42–47, 2012.
[4] Word Count Example, The Apache Hadoop Project, http://wiki.apache.org/hadoop/WordCount
[5] SENNA, CARLOS R.; RUSSI, LUIS G.C.; MADEIRA, EDMUNDO R.M. . An Architecture for Orchestrating Hadoop Applications in Hybrid Cloud, 14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), Chicago, USA, 2014