

Hadoop-in-a-Hybrid-Cloud

Luis Russi¹, Carlos R. Senna¹, Edmundo R. M. Madeira¹,
Xuan Liu², Shuai Zhao², and Deep Medhi²



¹Institute of Computing – State University of Campinas – Brazil

²University of Missouri–Kansas City – USA

luisrussi@lrc.ic.unicamp.br¹, {csenna, edmundo}@ic.unicamp.br¹
{xuan.liu, shuai.zhao}@mail.umkc.edu², dmedhi@umkc.edu²

MOTIVATION

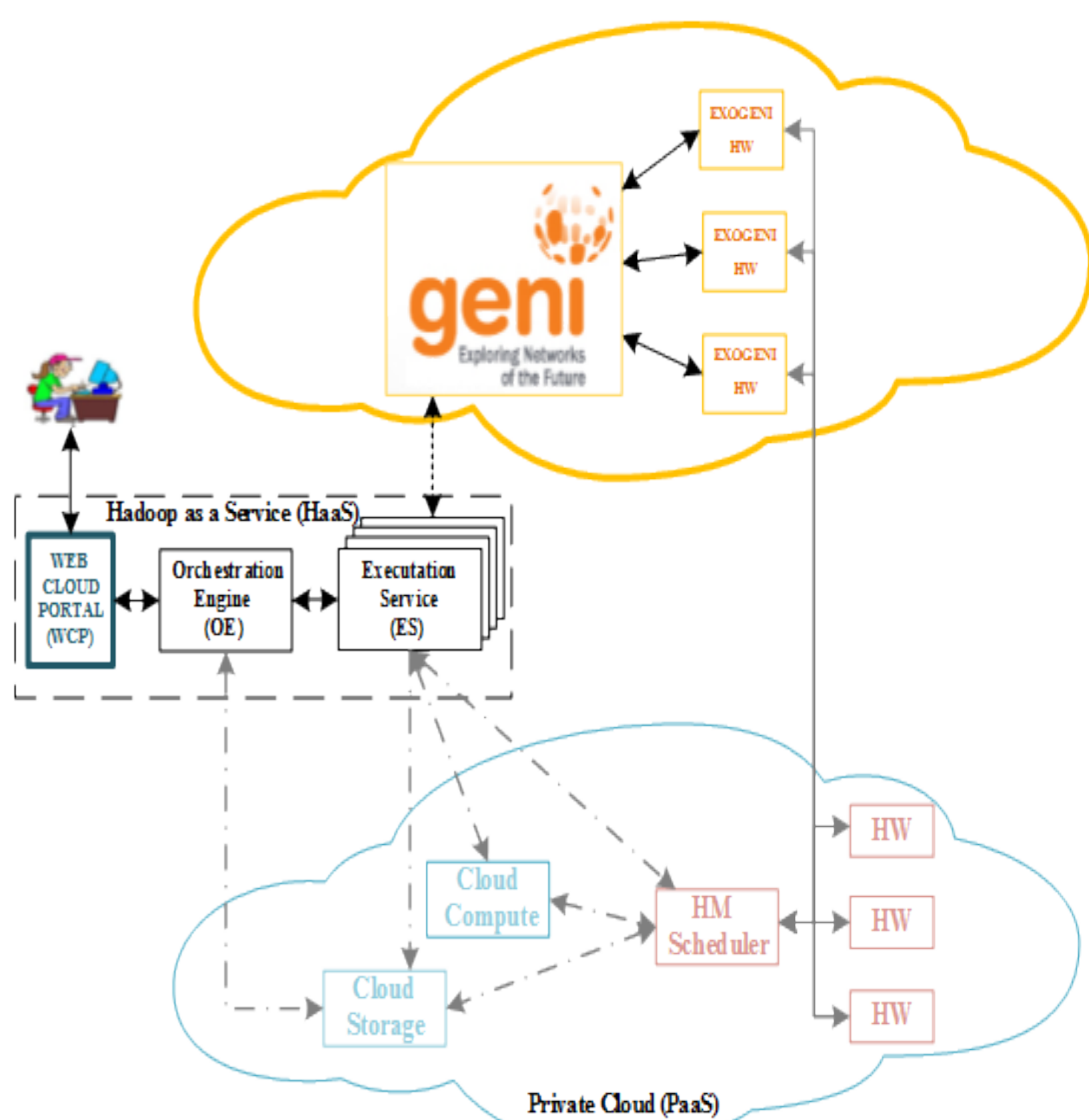
Execution of Hadoop applications in hybrid cloud is not easy!

- 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

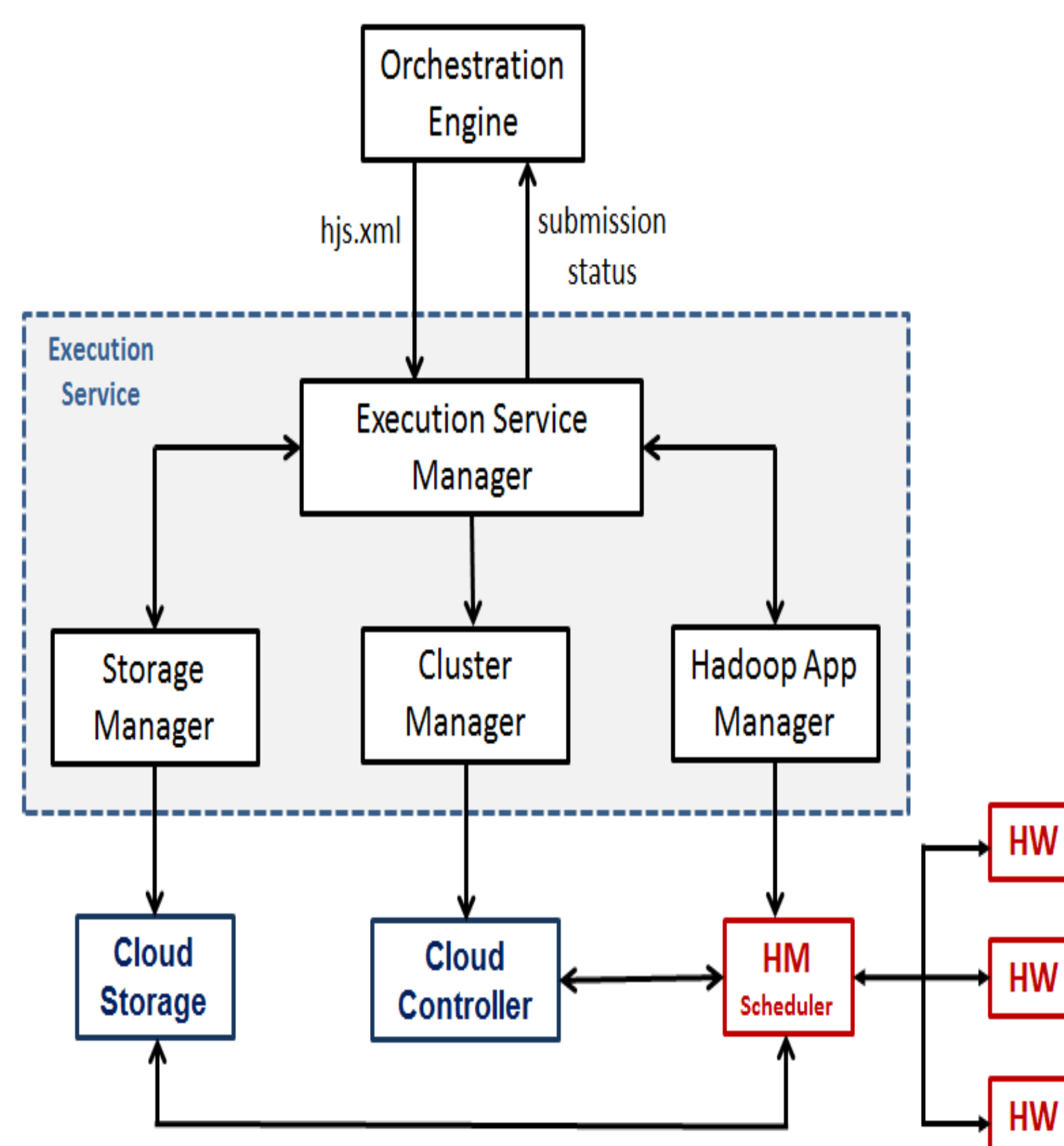
- User interface
- Management of files (application, data and submission)
- Simple XML-Based submission file
- Organizing the application workspace

ORCHESTRATION ENGINE

- Prepares working place in the private cloud's storage
- Creates an Execution Service Instance (ESI) 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

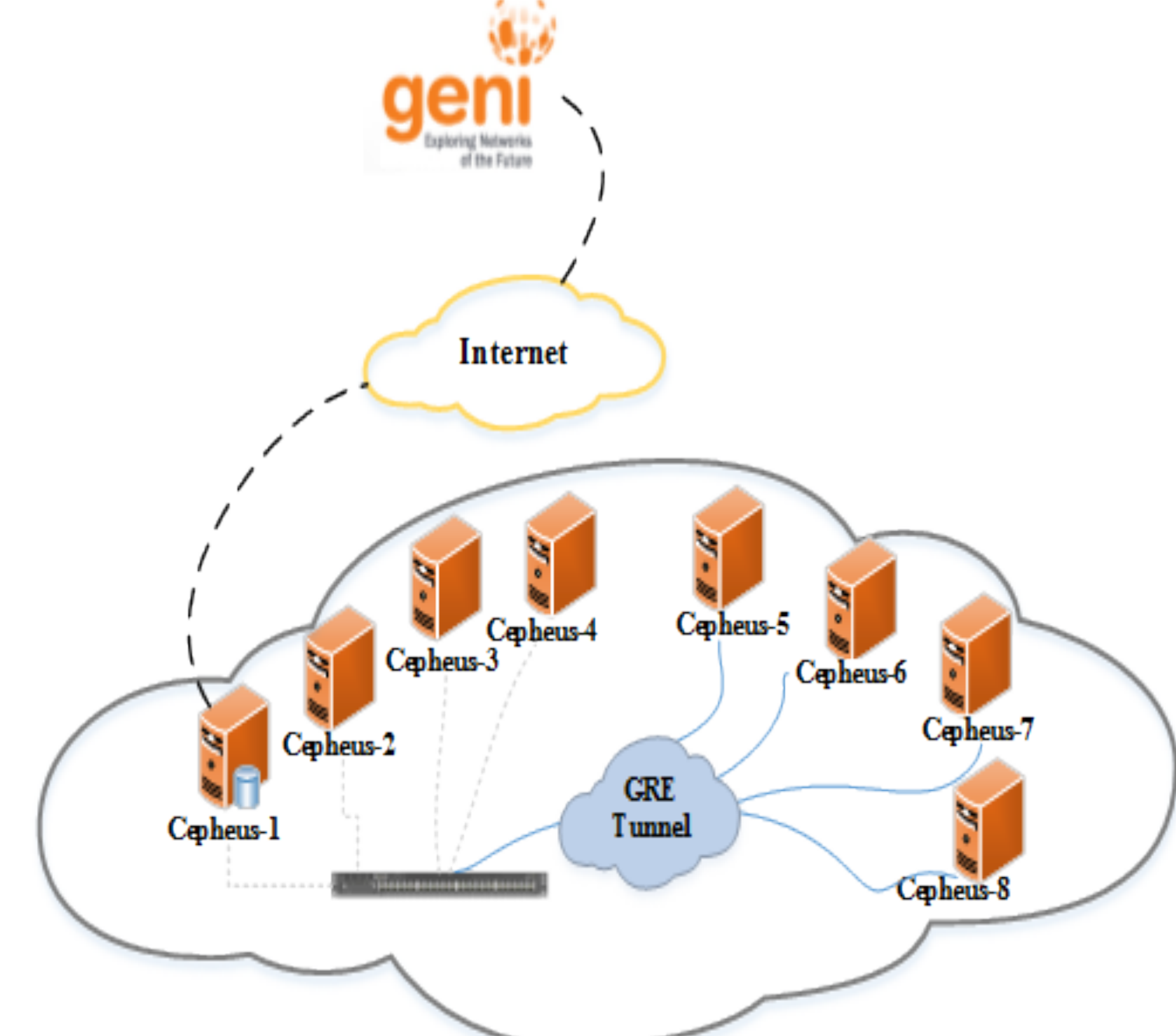
- ES Instance interacts with the private cloud monitoring system to evaluate the computational resources conditions
- Checks for extra resources from the public cloud
- Automatic Hadoop Cluster preparation
- Makes an accessible copy of the resulting files
- Eliminates all involved VMs
- Notifies the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



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

UNICAMP-UMKC HYBRID TESTBED



INITIAL RESULTS

- ✓ Include the UMKC compute nodes at the UNICAMP cloud controller
- ✓ Deploy exoGENI virtual machines with Hadoop 2.5.1
- exoGENI virtual machines and cloud Hadoop cluster joint
- Execute the Wordcount Hadoop application at the cluster
- Integrate GENI API to the private cloud framework

CONCLUSION

This paper 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.