Greek Research and Technology Network S.A.
GRNET
e-Science Activities & Plans

Platform of
National e-Science/Data Research Centres in Europe (PLAN-E)
Constituting Meeting
Amsterdam, 29 – 30 September, 2014

Panos Louridas
louridas@grnet.gr

The Big Picture

- Data-Intensive Science (up to the Petabytes level)
- Collaboration Platform (portal, wiki, instant messaging)
- Virtual Research Environment
- Scientific Instrument Data Sharing

Current Status

GRNET already operates an extensive IaaS platform:

- ~okeanos (http://okeanos.grnet.gr) offers computational resources
- Pithos+, integrated with ~okeanos, offers storage resources
- Built on open source software synnefo (http://www.synnefo.org), API-compatible with OpenStack

Some Numbers

History:

- Design started late 2010
- Production since July 2011

Numbers (9/2014):

- Users: > 8000
- VMs: > 6800 currently active
- ~400K VMs spawned so far (started/destroyed)
- More than 110K vlans spawned so far (user owned VLANs)

Typical VM flavor (more than 340 flavors available!)

- 4 cores (vCPUs), 80GB Hard Disk, 4 or 8GB RAM

Next Step

- Move from IaaS to PaaS
- PaaS for e-Science
- Leveraging computational and storage IaaS offered by ~okeanos and Pithos+

The Hadoop Ecosystem

- The most popular implementation of the MapReduce programming model
- Open source, commodity hardware
- Hadoop core (MapReduce, Hadoop distributed file system)
- Wide application ecosystem
  - (Pig, Hive, HBase)

From Hadoop to YARN

- YARN is the next generation Hadoop
- Remedies Hadoop shortcomings (like SPoF for namenode, performance for large clusters)
- Generalizes from a MapReduce implementation to a distributed job scheduling and execution system + MapReduce implementation on top
- Resource management: Many jobs can run on the same cluster

e-Science for Researchers

- Create and start YARN cluster, submit jobs, close cluster
  - via GUI and CLI (GUI and CLI communicate with same back-end API)
- Workflow capabilities with Apache Oozie and Apache Pig
- Other components of Hadoop Ecosystem (etc Hive) also available

Collaborative Research

- Virtual Research Environment
- Integrated group / project system
- Research/Project home page (portal, wiki)
- Project Management
- Teleconferences
- Digital repositories

Reproducible Research

- Save / Restore Research Environment
- Domain Specific Language (DSL) that will describe experiment and data
- Text editor => XML, JSON, YAML
- Re-execute experiment, possibly with different parameters

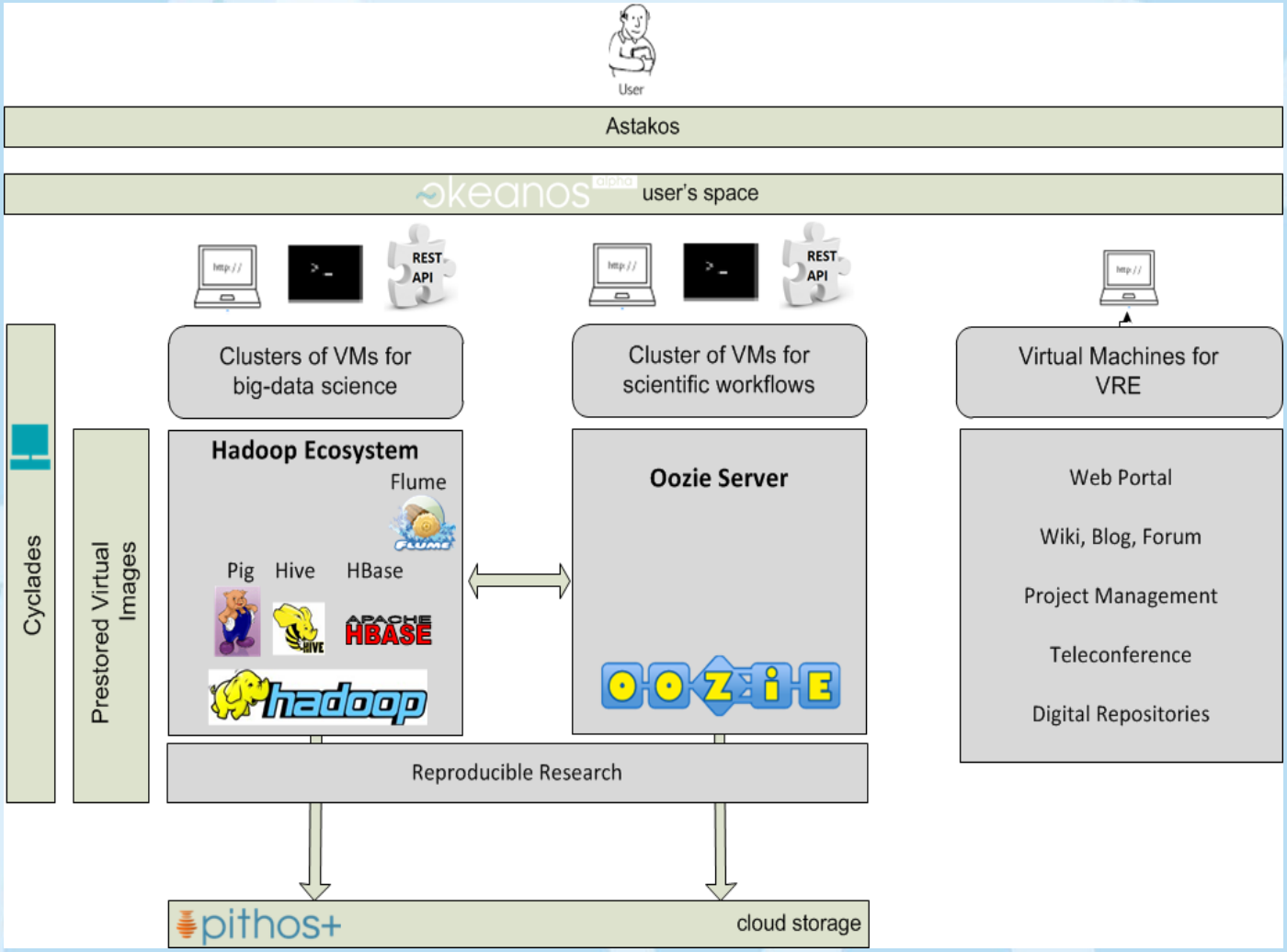
Access to Scientific Instruments

- Apache Flume
- Part of the Hadoop ecosystem
- Focus on streaming data

Requirements

- Interoperability with existing infrastructure (~okeanos, Pithos+)
- Web 2.0 Graphical User Interface
- Command-Line Interface
- Both GUI and CLI on top of single API, available to other 3rd party clients
- Deployment scripts

Architecture



Implementation

- Project already started (summer 2014)
- Completion time: summer 2015
- Development in rapid increments using Scrum
- All code open source, basic language is Python
- Adopting REST-based infrastructure with GUI on responsive framework (Ember.js) and asynchronous backend functionality

Q&A