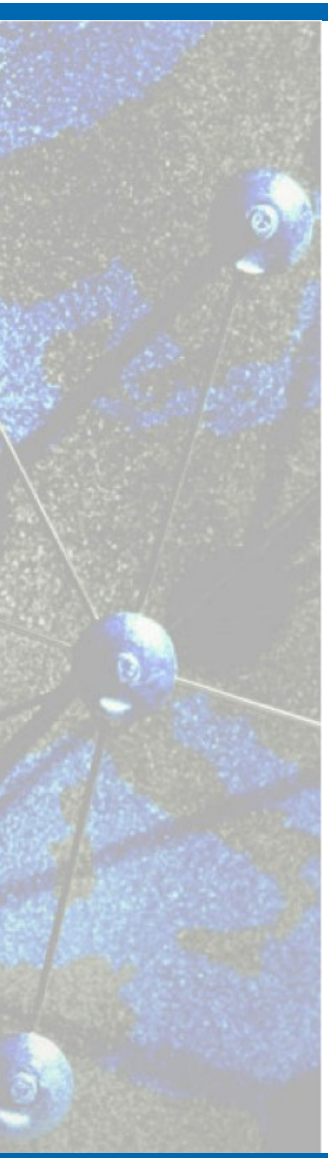

Exploiting Cloud Capacity in Scientific Research Platforms

Björn Hagemeier, Mirek Ruda, Boris Parak

Motivation

- Many use case images are not making use of cloud capacities
- Found number of use cases providing naive VM image layout for cloud usage
  - Does not scale, as some images were really large
  - Does not make use of cloud specific features, e. g.
    - Scalable storage (block/volume)
    - Scaling compute instances file systems at startup
- No common agreement on contextualization

FCTF current status

- Common ...
  - ... VM management
  - ... authentication
  - ... accounting
  - ... monitoring
  - ... image management/sharing
  - *... data management*
  - ... information discovery
- Mostly infrastructure related services
- Use cases hardly make any use of these
- Most promising for use cases: data management

Goals

- Define cloud capacities
  - What are they?
  - Who can make use of them and how?
- Make use of cloud capacity
  - Most obviously object/block storage
  - Auto-scaling both horizontally and vertically
- Minimal base images
  - Simplify initial setup and upgrade
  - Better integration with FedClouds infrastructure
  - General guidelines for contextualization

Milestones

- Identification of cloud capacities and eligible use cases
  - User surveys to match up capacities with use cases
- Exploit capacities in select use cases
- Evaluation of cloud platform services
- Document best practices and lessons learnt

Learn from the Best

AWS are providing a host of architectural guidelines to help users build their cloud based frameworks.

- `http://aws.amazon.com/products/`
- Our goal is to provide a similar level of documentation and similar services

Cloud Capacities (by area)

AWS

- Compute
- Content Delivery
- Database
- Software
- Networking
- Storage
- Payments and Billing
- Deployment and Management
- Web Traffic
- Application Services
- Workforce

Envisioned Use Cases

- BioVel/OpenModeller
  - Candidate for cloud object storage and potentially block storage
  - Needs subset of a big dataset for each simulation run
  - Chances are that only a subset of the full dataset will ever be used during the lifetime of an instance
- DCH-RP Community (Digital Cultural Heritage - Preservation)
  - Candidate for object storage, with guarantee terms for preservation
- LHCb DIRAC as a framework may be a good starting point, too
- ...

Outstanding Example

Peachnote use case

- Already using cloud capacities
  - Amazon SQS
  - Apache HBase (access to big data)

From the Applications session

Yesterday, we saw a number of inspiring works:

- Bioinformatics services on cloud
  - emphasis on small images
  - number of appliances
  - initial use of object storage, however not from within instances
  - persistent disks
- Phenomenology Tools on OpenStack Infrastructure
  - application contextualization
- Provision of Grid worker nodes
  - automatic scaling based on defined rules
  - can not only be used for Grid worker nodes

Project overview

- Timeline
  - Begin in April 2013, i. e. now
  - End in March 2014
  - 8 PM spread over 12 Months
- Partners
  - FZJ (lead)
    - Björn Hagemeier
  - CESNET
    - Mirek Ruda
    - Boris Parak