Datacloud, a Proposal for EINFRA-1 Topics 4,5

Davide Salomoni, INFN CNAF

Catania, 28/5/2014

Overall Goal

- The goal of the proposal is to develop a comprehensive computing platform, including and expanding a PaaS framework, which will allow public and private e-infrastructure service providers (such as EGI, EUDAT, PRACE, HelixNebula), to integrate their existing services and make them available to a wider user base in both the public and private sectors, for the benefit of scientific communities.
- The Consortium is currently being defined and will be a mix of technology providers, resource/ infrastructure providers and user communities.

EINFRA-1-2014 Items: Summary

1. Establishing a federated pan-European data e-infrastructure.
2. Services to ensure the quality and reliability of the e-infrastructure.
3. Federating institutional and, if possible, private data management and curation tools and services.
4. **Large scale virtualization of data/compute centre resources.**
5. **Development and adoption of a standards-based computing platform (with open software stack).**
6. Support to the evolution of EGI.
7. Proof of concept and prototypes of data infrastructure-enabling software (e.g. for databases and data mining) for extremely large or highly heterogeneous data sets.
8. Enable the creation of a platform and infrastructure for mining text aggregated from different sources/publishers.

EINFRA-1-2014 Item 4-5

- **Item 4:**
  - Large scale virtualization of data/compute centre resources to achieve on-demand compute capacities, improve flexibility for data analysis and avoid unnecessary costly large data transfers.
- **Item 5:**
  - Development and adoption of a standards-based computing platform (with open software stack) that can be deployed on different hardware and e-infrastructures (such as clouds providing infrastructure-as-a-service (IaaS), HPC, grid infrastructures…) to abstract application development and execution from available (possibly remote) computing systems. This platform should be capable of federating multiple commercial and/or public cloud resources or services and deliver Platform-as-a-Service (PaaS) adapted to the scientific community with a short learning curve. Adequate coordination and interoperability with existing e-infrastructures (including GÉANT, EGI, PRACE and others) is recommended.

Datacloud Proposal: Overall Vision

- Key Objectives:
  - Integration of Grid and Cloud computing models into existing data centers.
  - Support of hybrid (private/public) Clouds.
  - Dynamic, fair-share based scheduling of local resources for both Clouds and Grids.
  - Cloud federations and data center virtual extensions.
  - Support for new models (e.g. distributed objects, Cloud based) for data access and management.
  - Simplification of resource instantiation and use.
  - PaaS-level reusable component selection.

Main Technical Areas for Innovation

- These points are subject to modification and/or expansion, depending on the interest of the technology providers or communities.

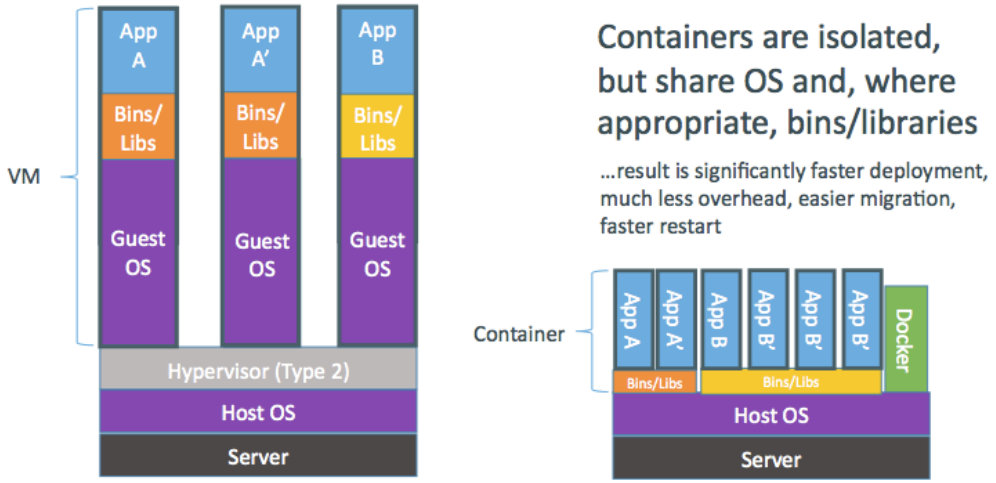
1. Scheduling within IaaS
2. Exploitation of Virtualization Techniques
3. Dynamically Distributed Virtual Data Centers
4. Service Orchestration
5. Exploitation of Data Locality
6. Security/AAI
7. Big Data: Management, Access, Transfer
8. Pilot Test Bed

Scheduling

- Introduction of flexible local, batch-like scheduling and prioritization capabilities into the most popular IaaS solutions.
- Scheduling across hybrid Clouds.
- On-demand resource pooling (“Cluster-as-a-Service”-like).
- Scheduling of virtual distributed clusters.
- Scheduling associated to network bandwidth allocation / availability.
- Scheduling based on resource location.
- Services for resource discovery and monitoring.

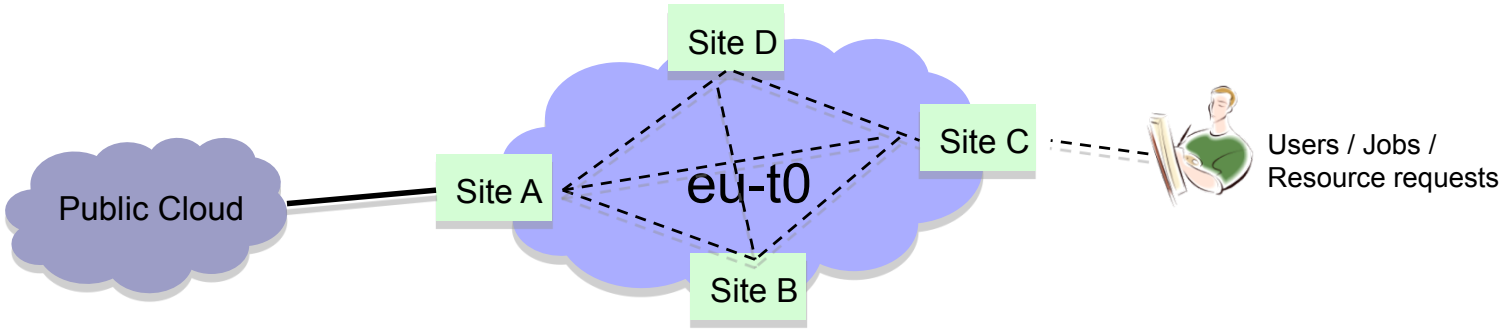
Exploitation of Virtualization Techniques

- Support of new patterns for application development and shipment through *containers* associated to Cloud provisioning.
- Use of special-purpose hardware (e.g.: GPUs, accelerators, low-latency networks) coupled with virtualization.



Dynamically Distributed Virtual Data Centers

- Given the constant increase and commoditizing of network bandwidth, the project will make it possible to dynamically:
  - Exploit hybrid Cloud deployments.
  - Transparently extend data centers to remote locations.



Service Orchestration

- Extend / connect different IaaS layers into a PaaS layer.
- Selection of components needed for a particular task and instantiation of the appropriate resources in multiple, federated, possibly hybrid Clouds. Components might for example be related to:
  - Highly-available solutions (mission-critical apps, disaster recovery).
  - Specific capabilities (low-latency connections, GPU devices e.g. for HPC-type resources).
  - Composition of data analysis workflows, potentially serving multiple scientific communities.
  - Interactive use-cases, providing services covering small / medium fast execution environments, with possible specific focus on QoS.
- Component/service selections will be made easier through the use and extension of Scientific Portals.

Exploitation of Data Locality

- The focus of this area is on the dynamic exploitation of data locality according to user requirements.
  - E.g. automatic data distribution to/across multiple zones of a given data center.
- This can be useful at both local (single data center) and distributed (multiple, federated data centers) levels.
- This is linked to the Scheduling area and requires consideration of network and storage topologies.
  - Users should not be aware of the details of data distribution.

Security/AAI

- Secure data access for all types of data.
- Federated user authentication and SSO based on the user credentials provided by home hosting laboratory or institutions.
- Support for VO attributes and policies in existing open-source Cloud platforms and in PaaS/SaaS solutions.
- “Catch-all” Identity Federation and Identity Providers for “homeless” users.
- Policy management and distribution service enabling Cloud federation-wide access control services for attribute and credential-based access requirements.
- Integration of SAML-based federated authentication and authorization mechanisms in order to simplify credential management and delegation.

“Big Data”: Management, Access, Transfer (1)

- Creation of a distributed archive and catalogue for the federation of storage systems with automatic failover for data access; implementation of mechanisms to enforce consistency checks and self-recovery procedures of data for both object and block data types.
- Development/support of platforms to provide high performance data analysis.
- Consolidation of technologies for data access and archiving through standard protocols (e.g. HTTP/ WebDAV); definition and development of a *Federated Data Storage as a Service* solution to satisfy different types of requests and needs, leveraging multiple storage solutions in different infrastructures, associated to standard protocols such as S3, CDMI, etc.

“Big Data”: Management, Access, Transfer (2)

- Definition of an abstraction layer aiming to provide a high-level and advanced data managements features.
- Definition of high level services to move very large data sets, able to deal with failures of software and hardware components in an heterogeneous Cloud infrastructure.
- Address topics such as integrity, security and privacy.
- Provide “personal-storage” services for small to medium scientific collaborations.
- Development of a Data Life Cycle management system (policy-based data management).

Pilot Test Bed

- Some of the project partners will make available a distributed pilot test bed to enable full testing of the new services and to provide a smooth migration path from current batch-based Grid Data Center to a new heterogeneous, hybrid Cloud-based federated distributed infrastructure.