Storage Management in INDIGO

Paul Millar

paul.millar@desy.de

with contributions from Marcus Hardt, Patrick Fuhrmann, Łukasz Dutka, Giacinto Donvito.
**INDIGO-DataCloud: cheat sheet**

- A Horizon-2020 project
  
  **Approved:** January 2015; **Started:** April 2015; **Ends:** September 2017.

- 26 partners from 11 European countries.

- Over €11 million

- **Objective:** develop an Open-Source platform for computing and data, deployable on public and private cloud infrastructures.

- Requirements from 11 INDIGO communities.

**More details:** [http://indigo-datacloud.eu/](http://indigo-datacloud.eu/)
The “golden era”
Collaborations & new equipment
More resources, but “cloud”!
Who is involved

- **Biological and medical science**
  Biological, molecular and medical imaging, life science research applied to medicine, agriculture, bio-industries and society, structural biology.

- **Social science, arts and humanities**
  Georeferencing (e.g., of current and historical maps), cultural heritage, smart sensors.

- **Environment and earth science**
  Biodiversity and ecosystem research, interactions between geosphere, biosphere and hydrosphere, earth system modelling.

- **Physical sciences**
  Astrophysics, theoretical and experimental research in physics.
How INDIGO-DataCloud helps

**WP4:**
Providing common interfaces for site-local resources

**WP5:**
Providing a useful, high-level service that combines multiple resources.
## IaaS: Quality of Service

<table>
<thead>
<tr>
<th>Media Quality</th>
<th>Access Latency</th>
<th>Durability</th>
<th>Data rate</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HIGH</td>
<td>OK</td>
<td>OK</td>
<td>Very low</td>
</tr>
<tr>
<td></td>
<td>MEDIUM</td>
<td>MEDIUM</td>
<td>OK</td>
<td>Reasonable</td>
</tr>
<tr>
<td></td>
<td>LOW</td>
<td>Not so clear</td>
<td>MEDIUM</td>
<td>Very high</td>
</tr>
<tr>
<td></td>
<td>MEDIUM</td>
<td>Quite OK</td>
<td>OK</td>
<td>MEDIUM</td>
</tr>
<tr>
<td></td>
<td>MEDIUM</td>
<td>OK</td>
<td>OK</td>
<td>MEDIUM</td>
</tr>
</tbody>
</table>

- **IaaS:** Infrastructure as a Service
- **High:** Excellent performance
- **Medium:** Satisfactory performance
- **Low:** Below satisfactory performance
- **Not so clear:** Unclear performance
- **Almost OK:** Almost satisfactory performance
- **Quite OK:** Quite satisfactory performance
- **Very high:** Very poor performance
- **Very low:** Very poor performance
- **Reasonable:** Reasonable performance
Making the choice meaningful

- **Access Latency / ms**
  - Low latency & lowest price → Class #1
  - High throughput & super durable → Class #2
  - Large volume & cheap & archive → Class #3

**Discover & Match**

- GUI
- REST API

**Canonical classes**
Federating QoS Choice

Property Information System

Discover & Match

GUI

REST API

Discover & Match

GUI

REST API

Discover & Match

GUI

REST API

Discover & Match

GUI

REST API

PaaS

IaaS

IaaS
IaaS: Data Lifecycle

Data Lifecycle is just time dependent changes of

- Storage Quality of Service
- Ownership and Access Control: PI Owned, limited access → Site Owned, Public access
- Payment model: pay-as-you-go → pay-in-advance for rest of lifetime
- Maybe other things
IaaS: Metadata-driven storage
IaaS: laying hierarchical storage
Ease of deployment

Grid computing

INDIGO-DataCloud
Identity and group-membership

- Allow **different** authentication mechanisms
  
  SAML, OpenID-Connect, X.509, ...

- **Harmonise** user identities:
  
  User is the same person, irrespective of how they authenticate

- **Support** group-membership:
  
  Membership can be used for authorisation decisions.

- **Support** third-party group membership:
  
  VOMS-style: where membership *not* asserted by authentication service.

For more details, see Andrea's Talk: “The Indigo AAI” tomorrow 10:15 in Scuderia.
Availability

- **First official release**: end of July next year
- We will start making available some services as soon as they are ready enough to be tested
- All the changes on the existing projects will be pushed back to the official releases.

  OpenStack, OpenNebula, dCache, OneData, Mesos, Accounting, QoS/SLA, etc...
The result: more time researching
PaaS: Unified data access

- **Data set registrar:**
  Unified vision of geographically distributed data set.

- **Data affinity:**
  Computation jobs started on resources close to data.

- **Automatic Staging:**
  Replicating data when not close to specialist hardware.

- **Optimised streaming access of remote data:**
  When data is not staged.

- **API for data and metadata management:**
  registration, migration, replication, sharing; federated ACL management

- **Optimised data movement**
- **Aggregate QoS through replication**
- **Gateway to external data repositories**
PaaS: Unified storage interfaces

- Data access methods and protocols:
  
  CDMI, Web GUI, WebDAV, S3, POSIX (mounted virtual volume)

- Data locations:
  
  via CDMI or WebDAV

- Data migration and replication:
  
  REST API or CDMI extension allowing replication based on metadata.
PaaS: Data Affinity

- Knowledge of where data is located
- Identify which IaaS computing resource is closest
- Allow deployment of computation activity close to where the data is located
- Minimise data transfers to improve efficiency.