Blueprint for an Open Source On-Premise Cloud to Serve as a Research Data Infrastructure for Universities

Raimund Vogl (rvogl@wwu.de), J. Hölters, M. Blank-Burian, M. Ketteler-Eising, D. Rudolph, H. Angenent, C. Schild, S. Ost
IT Center, University of Münster
Digitalization = Data Deluge
digital storage space at Münster University

<table>
<thead>
<tr>
<th>Jahr</th>
<th>TSM-Backup datavolume</th>
<th>Disk Space Raw (w/o sciebo)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>21.4</td>
<td>9</td>
</tr>
<tr>
<td>2003</td>
<td>31.4</td>
<td>9</td>
</tr>
<tr>
<td>2004</td>
<td>45.4</td>
<td>119</td>
</tr>
<tr>
<td>2005</td>
<td>63.6</td>
<td>148</td>
</tr>
<tr>
<td>2006</td>
<td>141.6</td>
<td>977</td>
</tr>
<tr>
<td>2007</td>
<td>261</td>
<td>1006</td>
</tr>
<tr>
<td>2008</td>
<td>320</td>
<td>1417</td>
</tr>
<tr>
<td>2009</td>
<td>394</td>
<td>1792</td>
</tr>
<tr>
<td>2010</td>
<td>550</td>
<td>2990</td>
</tr>
<tr>
<td>2011</td>
<td>730</td>
<td>4902</td>
</tr>
<tr>
<td>2012</td>
<td>1077</td>
<td>5258</td>
</tr>
<tr>
<td>2013</td>
<td>1097</td>
<td>7516</td>
</tr>
<tr>
<td>2014</td>
<td>1461</td>
<td>9412</td>
</tr>
<tr>
<td>2015</td>
<td>1830</td>
<td>10420</td>
</tr>
<tr>
<td>2016</td>
<td>2144</td>
<td></td>
</tr>
<tr>
<td>2017</td>
<td>2350</td>
<td></td>
</tr>
<tr>
<td>2018</td>
<td>3025</td>
<td></td>
</tr>
<tr>
<td>2019</td>
<td>3930</td>
<td></td>
</tr>
</tbody>
</table>
The challenges of Digitalization for HE & Research IT – demands on infrastructure

- Deluge of digital (research primary) data – Petabytes - hard to move (networks!)
- Long term availability and preservation
- Implement FAIR principles – findable, accessible, interoperable, reusable
- Collaboration on data
- Accommodate various (eScience) environments
- (Very) scalable Storage – with fast connect to (flexible – VM to HPC batch) Compute
- Sustainable on the long term:
  - Financially and technologically - hardware and software
  - Migration to future hardware generations (for decades)
- Unify operations environment of eScience/Research and general IT (Virtualization! Storage!)
The elements for the solution

- Open Source - Freeware: cost, openness, sustainability, no vendor lock-in
- Software Defined Storage – Object Storage – Ceph
- OpenStack (VMware Future?)
- Kubernetes - Serverless computing – cloud (micro) service orchestration
- Deep know how for technologies used in university IT staff
- Multi university cooperation for joint service design/implementation/operation
Data Center Net

Campus Backbone

2x 40GE

Data Center Net

NFS/SMB

2x 100GE

2x 100GE

Nx 40GE

data path
HPC – Cloud
> 10GE

Cloud
OpenStack/kubernetes/Ceph
2x 100GE Net

50+ hyperconverged hosts for:
- OpenStack VMs
- 15 PB+ Ceph Storage

CephFS

GPFS

HPC
10/40GE Net

2PB+ parallel GPFS Storage
Nx 40GE

500 compute nodes
10GE

OmniPath HPC Intercon.
The cloud stack

User Applications (Webserver, Databases, Kubernetes, ...)

Virtual Machines

OpenStack

Kubernetes, Docker, Ceph

Operating System

Node 1
Node 2
Node 3
Node (N-1)
Node N
Live Metrics, User Self Service
Setting up eScience applications on the cloud

Applications

Orchestration Layer

Resource Layer

Kubernetes

Pods (Containers)

OpenStack

Compute Instances (VMs)

Ceph

hyperconvergent, triply replicated Storage

deployment, management & scaling of applications and related services

dynamically provides computing, storage and network resources

(classic, monolithic Applications and Services)

App A

App B

BinderHub

JupyterHub
Research Data Services from the sciebo cloud

Research Data Project Manager (RDPM)

Research Data Services (RDS)
- Core Modules
- Modules for Internal Services
- Modules for External Services

Workflow Management System (WfMS)

Extract Transform Load (ETL) Tool

External Archives & Repositories

Research Data Infrastructure (RDI)

sciebo

owncloud service for 120.000 users in NRW
2300 research project boxes
containers/kubernetes
Supporting Researches throughout the Research Data Lifecycle

Private Domain/Group Domain
- Collection
- Processing
- Analysis
- Meta Data Enrichment

Research Data Lifecycle

Permanent Domain/Access Domain
- Reuse
- Access
- Archiving
- Publication
The eScience Support Center – eScience WWU
Jointly operated by IT center and Library

<table>
<thead>
<tr>
<th>DH</th>
<th>Digital Humanities</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS</td>
<td>Data Science</td>
</tr>
<tr>
<td>HPC</td>
<td>High Performance Computing</td>
</tr>
<tr>
<td>RDI</td>
<td>Research Data Infrastructure</td>
</tr>
<tr>
<td>RDM</td>
<td>Research Data Management</td>
</tr>
<tr>
<td>RSE</td>
<td>Research Software Engineering</td>
</tr>
</tbody>
</table>

Current Research Information System
Thank you for your attention!

Questions?

rvogl@wwu.de