URIKA-GX PLATFORM'S MULTI-TENANCY: LESSONS LEARNED

Oleksandr Shcherbakov, D. Hoppe, Dr. T. Bönisch, M. Gienger (HLRS)

S. Andersson, J. Kuebler, N. Mujkanovic (Cray)

OUTLINE

- Introduction to HLRS
- Data Analytics @ HLRS
- Urika-GX. Towards multi-tenancy
- Summer school
- Current state
- Summary

High-Performance Computing Center | Stuttgart

HIGH PERFORMANCE COMPUTING CENTER STUTTGART (HLRS)

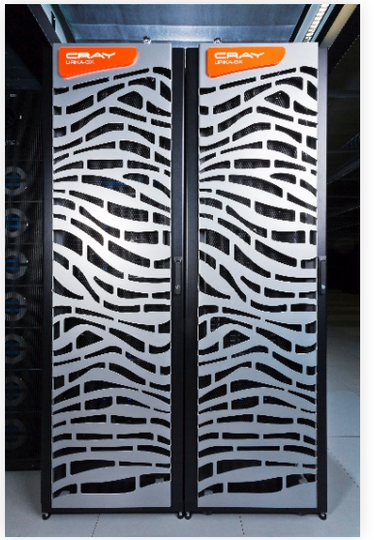
- Member of the Gauss Centre for Supercomputing
- Basic and applied research
  - Publicly funded national and European projects
  - Focused industrial collaborations
- Consultancy and training activities
- Providing High Performance Computing services
  - Academia
  - Industry

IMPORTANT HLRS SYSTEMS

HAZEL HEN CRAY XC40

- 7.712 nodes
- 185.088 cores Intel Haswell
- 7,40 PFLOPS Peak performance
- 1 PB main memory
- 15 PB disk storage

GILGAMESCH & ENKIDU CRAY URIKA-GX



- 64 nodes (48 + 16)
- 2.400 cores
- 33 TB main memory
- 100 TB HDFS Storage

DATA ANALYTICS @ HLRS

CATALYST

- Project established in 2016 to **evaluate and push the incorporation of data analytics for HPC**
- Cooperation with Cray and Daimler
  - Real-world case studies with partners from academia and industry
- Focus on the **engineering domain** in comparison to the general application of data analytics for natural sciences
- Integration and evaluation of **2 Cray Urika-GX systems** into the production environment of HLRS
  - Additional requirements concerning **multi-user support** and **security** arise

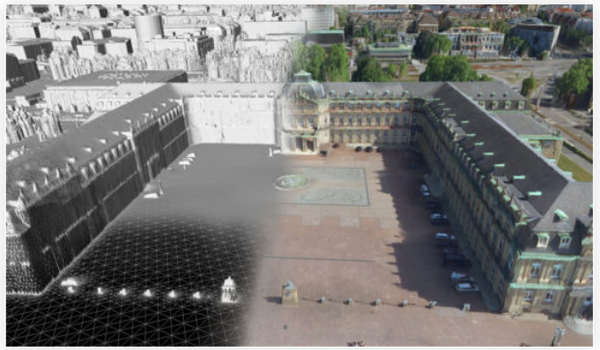
SELECTED CASE STUDIES

OVERVIEW

- “3D City over Night” (nFrames)
- EOPEN, an Europrean H2020 project
- SmartSHARK (University of Goettingen)
- Performance variations in HPC jobs (HLRS)

“3D CITY OVER NIGHT” (NFRAMES)

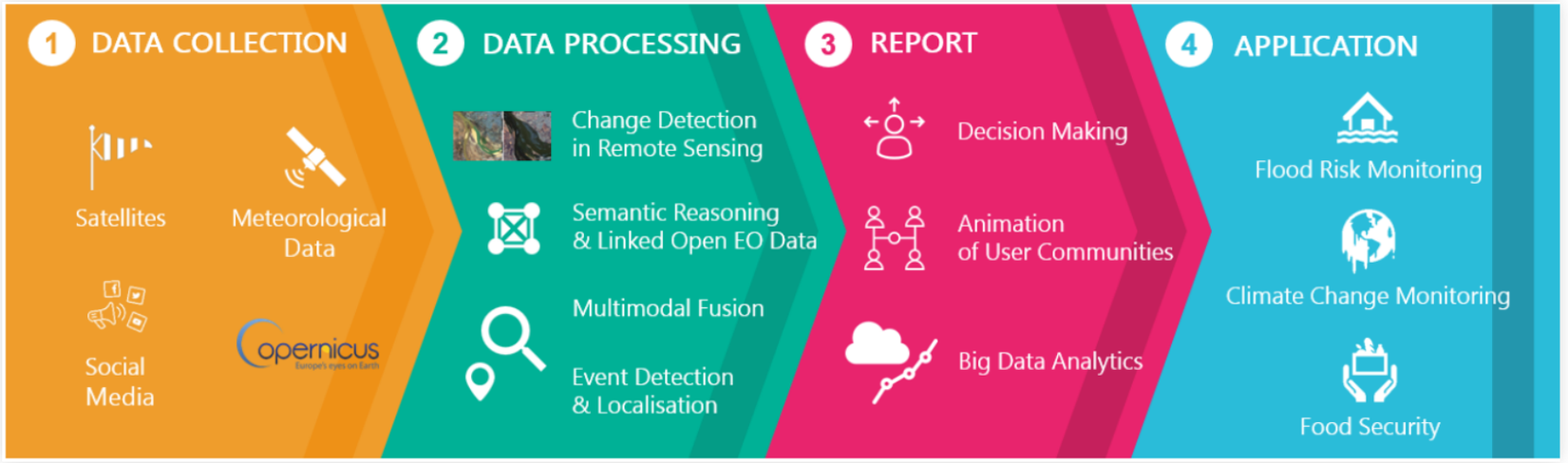
- Company provides software for **photogrammetry**
- **Aerial images** are huge, and thus processing large areas with high resolution is very compute intensive
- nFrames deployed their application on our system in order to process an entire city “over night” compared to 100 days with a single workstation



https://www.nframes.com/gallery/

EOPEN—H2020 EUROPEAN PROJECT (2017–2020)

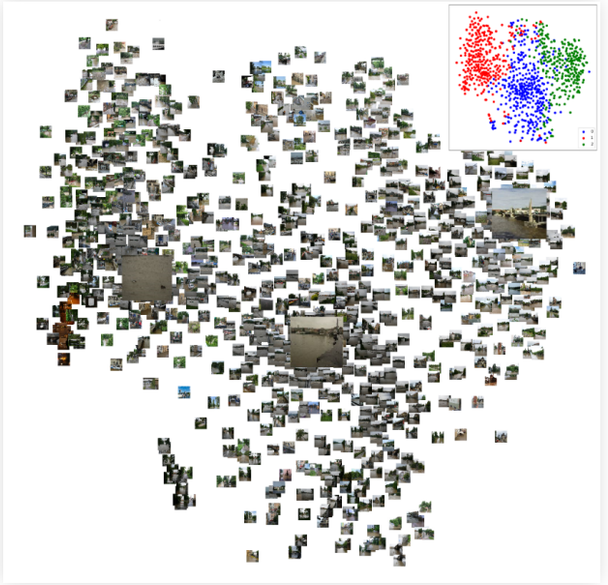
- **Platform** targeting non-expert Earth Observation (EO) data users, experts, and the SME community
- Make **Copernicus data** and services easy to use for Big Data
  - e.g. by providing infrastructure to support data analytics

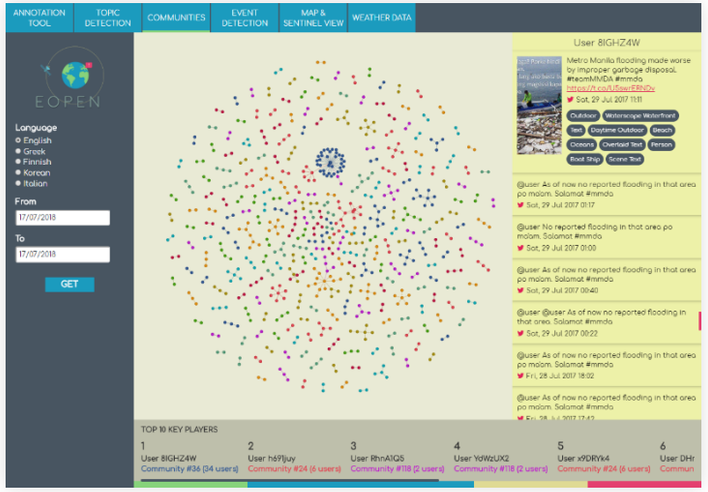


http://eopen-project.eu/

EOPEN—H2020 (CONT'D)

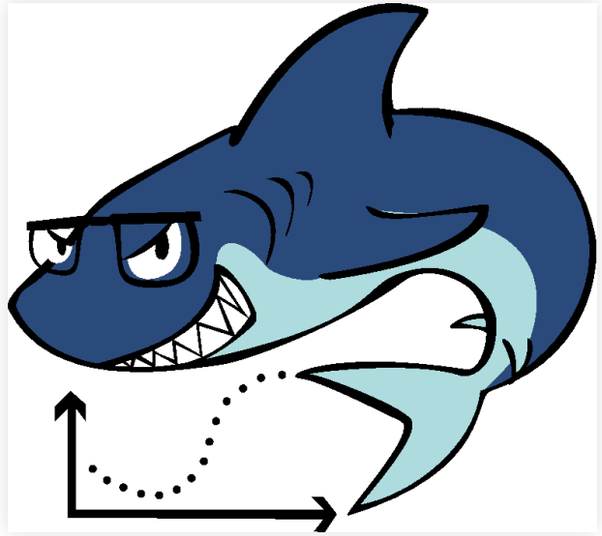
- Image clustering of flooding images
  - Using TensorFlow with Keras
- Community detection in Twitter
  - Exploit the Cray Graph Engine
  - TBD in 2019





SMARTSHARK

- Project by the University of Goettingen
  - Steffen Herbold, Software Engineering for Distributed Systems
- Analysis of software projects
  - defect prediction
  - sentiment mining
  - detection of social networks
- SmartSHARK collects data from version control systems, and allows to seamlessly perform data analytics on the data using Apache Spark



https://www.swe.informatik.uni-goettingen.de/research/smartshark

SMARTSHARK (CONT'D)

- University of Goettingen has no access to required hardware to analyze large software repositories with several TB of data
  - data consists not just from all revisions of available files, but also additional software metrics (≈ 200 features)
- HLRS case study
  - build a logistic regression model over software code entity states to predict which commits likely fix defects
  - University of Goettingen was able to analyze a GitHub repository with more than 2 TB of data for the 1st time

PERFORMANCE VARIATION IN HPC JOBS

- **Performance variability** on HPC platforms is a critical issue with serious implications for the users
  - **Irregular runtimes** prevent users from correctly assessing performance and from efficiently planning allocated machine time
  - Hundreds of **applications concurrently sharing thousands of resources** escalate the complexity of identifying the causes of runtime variations
- On production systems, implementing trial-and-error approaches is practically impossible !

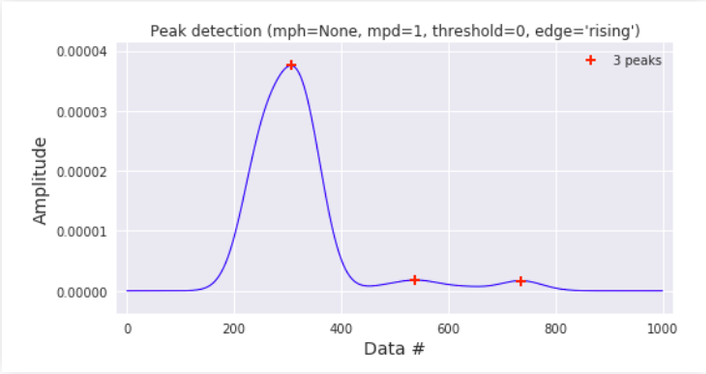
PERFORMANCE VARIATION IN HPC JOBS (CONT'D)

- What type of applications can impact the performance of other applications?
  - Victims
    - Applications that show high variability
  - Aggressors
    - Applications potentially causing the variability
- Understanding the nature of both types of applications is crucial for developing a meaningful detection mechanism

PERFORMANCE VARIATION IN HPC JOBS (CONT'D)

Victim Detection Approachs

1. Simple statistics
   - inter quartile distance
   - kernel density estimation
2. Machine learning and data mining techniques
   - neuronal networks
   - clustering (e.g. k-means)



URIKA-GX

TOWARDS MULTI-TENANCY

“Secure multi-tenancy is the key to utility computing, and now we can scale more securely”

Aisling MacRunnels CMO, JIVE Software

WHAT WAS WRONG?..

...apart from `admin:admin`

URIKA-GX V 1.0

- HDFS - "Simple" authentication by default
- No authentication in Mesos:
  - Any user could start jobs as any other user
  - Also as `root` (default configuration)
- `munge` credentials as environment variables for Marathon jobs
  - cron job with credentials for `root`
- No authentication in YARN
  - `http://.../?user.name=root`

URIKA-GX V 2.0

- Tenant VMs
  - Jobs are executed bare-metal
  - Commands are wrapper-scripts like (simplified):

```
ssh host $0
```

URIKA-GX V 2.1+

- Multi-tenancy
- Kubernetes (k8s)
  - Multi-tenancy is work-in-progress
- A user can mount host file systems, including /
- Default user in containers is `root`
- No authentication for Docker Registry
- *"It's not a bug, it's a feature"*

URIKA-GX 2.2UP02

released

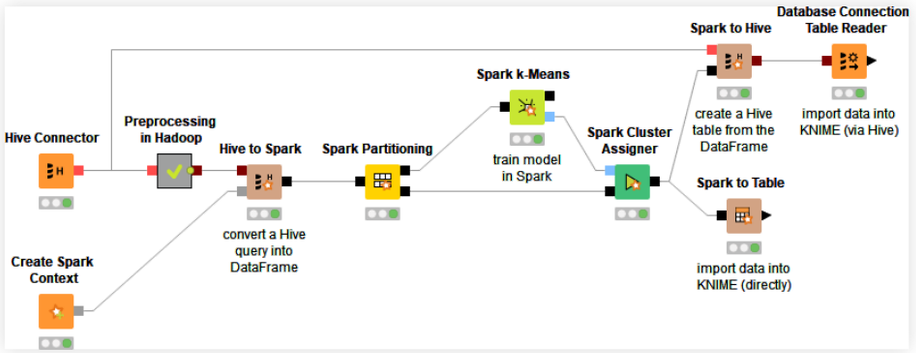
CUSTOMER'S REQUIREMENTS

SUMMER SCHOOL

Proof of concept

REQUIREMENTS

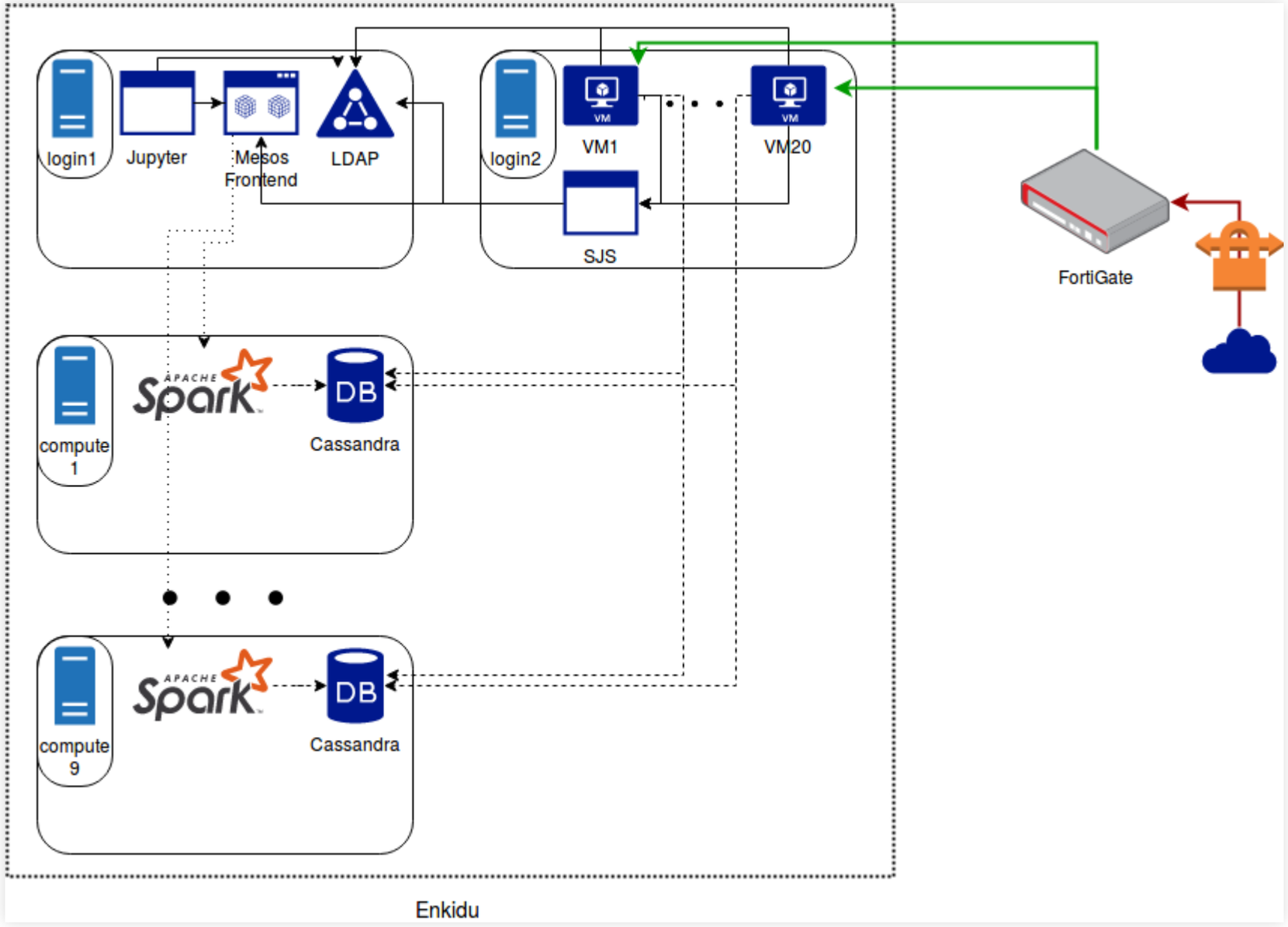
- Users must not be able do download or copy the data
- Cassandra
- Knime



https://www.knime.com/

IMPLEMENTATIONS

- (Default mode)
- xorgxrdp
- Desktop VMs for each user
- Spark Job Server (SJS) as Spark-Knime connector
- Mesos and Hadoop ACLs for SJS
- Cassandra
- Fortigate VPN
- LDAP connections for all services


login1
Jupyter
Mesos Frontend
LDAP
login2
VM1
VM20
SJS
FortiGate
compute 1
Cassandra
compute 9
Cassandra
Enkidu

PROBLEMS ENCOUNTERED

- SJS met `ulimit -u -S`
- `spark.port.maxRetries = 16`
- Summer school pilot was running in "Default" mode

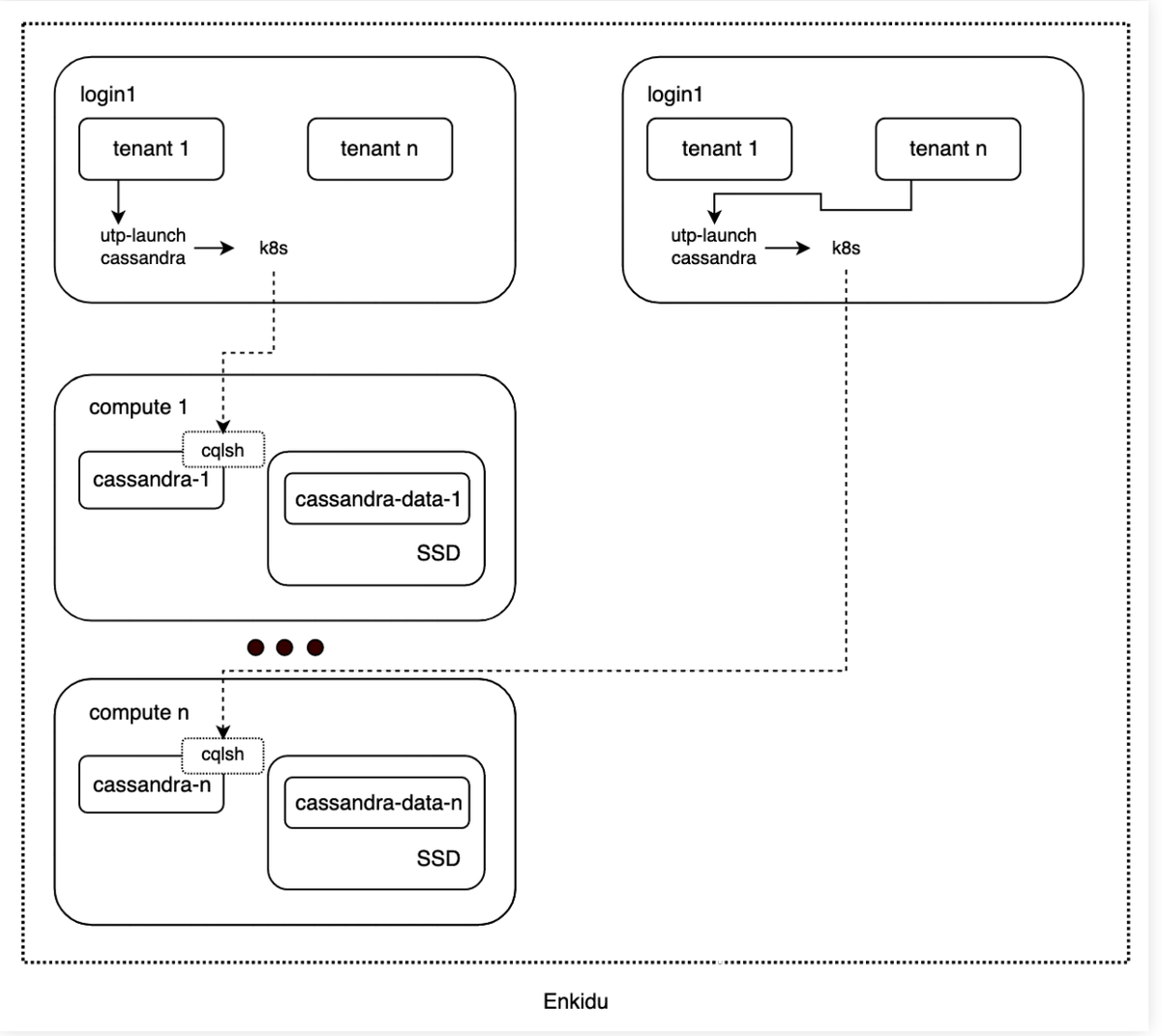
CURRENT STATE

- Urika-GX 2.1+, multi-tenancy on k8s with Cray wrappers
- Cassandra on k8s
- “`spark-submit` only”

CASSANDRA ON K8S

- Statefulset
  - Provides guarantee regarding ordering and uniqueness of pods
  - Pods created from same spec, but not interchangeable
- LocalPersistentVolume
  - Mounted local storage device
  - Ensures pod-to-node affinity
  - Must be manually created
  - Enabled alpha feature in Kubernetes 1.9
- PersistentVolumeClaim
  - Consumes persistent volume and binds to pod

CASSANDRA INSTALLATION (ABSTRACT VIEW)



SUMMARY

- Data analytics @ HLRS
  - Evaluation of Urika-GX in a real production environment
  - Focus on solutions for the engineering domain
  - Close collaboration with academia and industry
  - **Collaboration partners are always welcome**
- Security
  - Data analytics apps are not secure with default settings
  - Secure multi-tenancy support is still an ongoing challenge

THE END

Oleksandr Shcherbakov
High Performance Computing Center Stuttgart
Nobelstrasse 19
70569 Stuttgart
Phone: +49-711-685-87201
Email: shcherbakov@hlrs.de