IRIS Database Service at Edinburgh

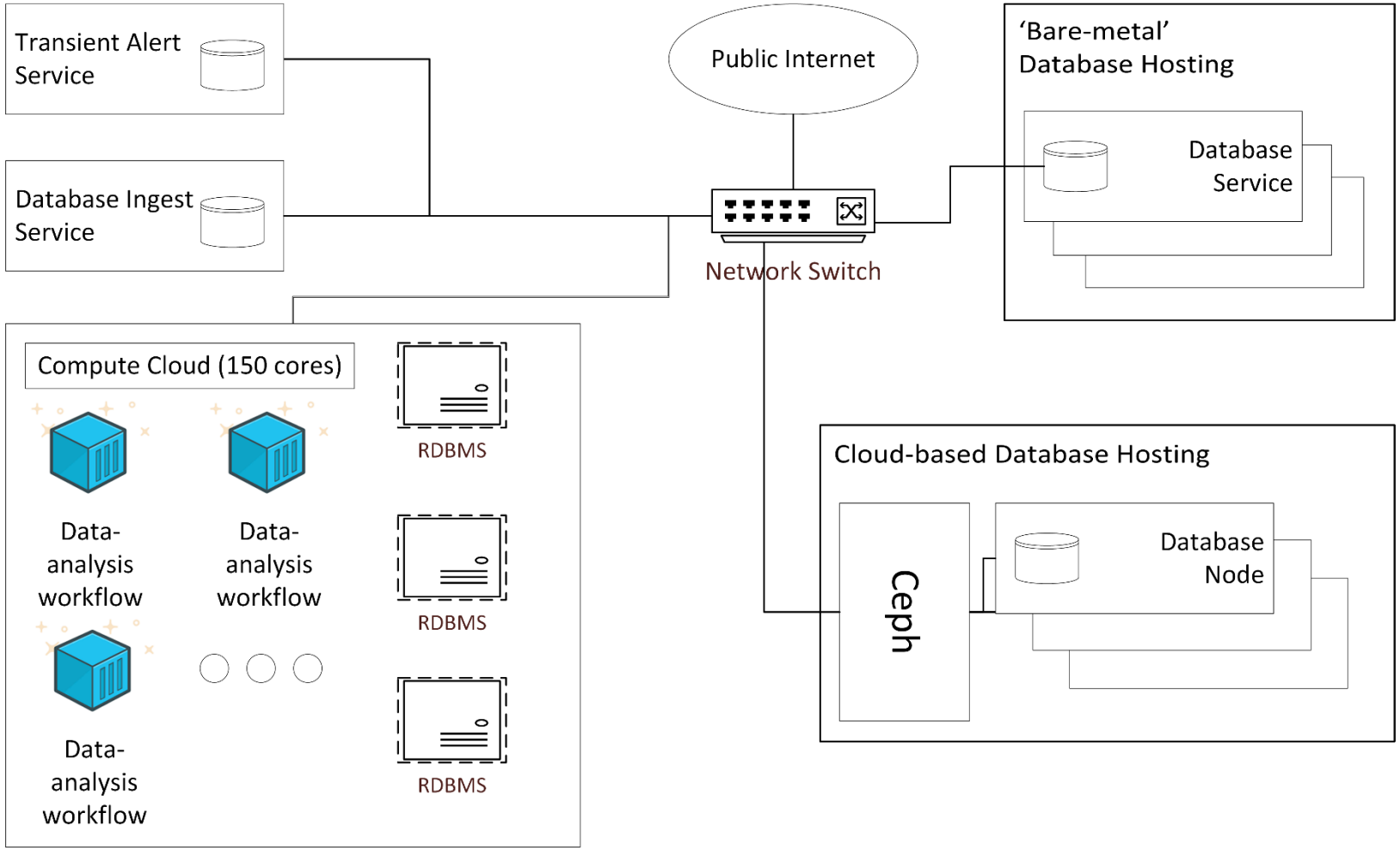
George Beckett (Teng Li and Mark Holliman), Edinburgh,

IRIS Face-to-face Meeting, Cambridge 3rd—4th April 2019

Motivation

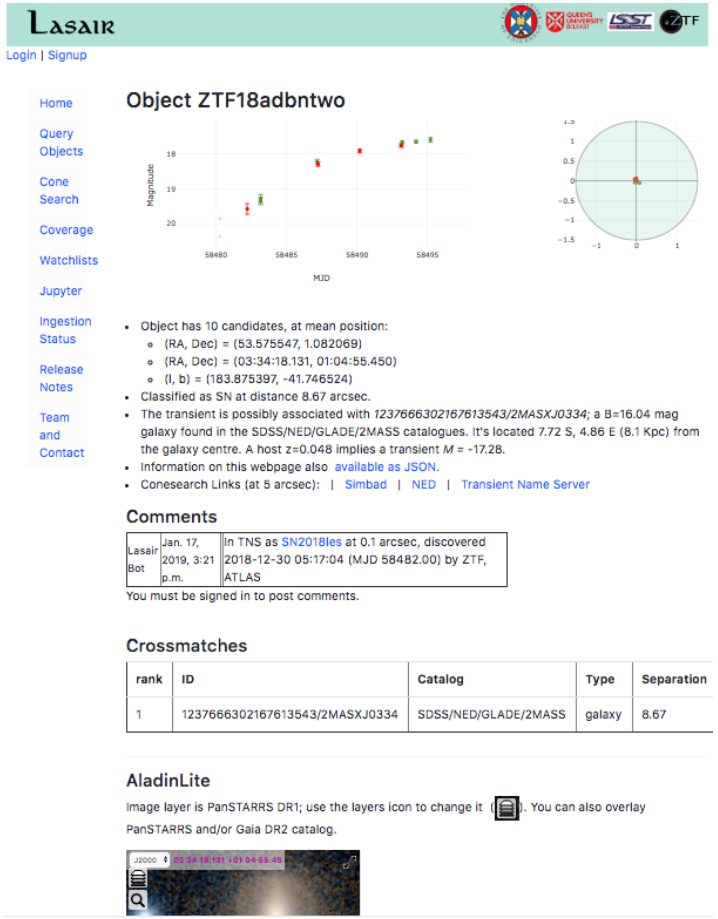
- Typically, in astronomy, sky surveys published in (RDBMS) databases
  - Growth in storage capacity has typically kept up with size of surveys
  - Workstation-based analysis of data has been practical
- Next generation of surveys (Euclid, LSST, SKA) will produce multi-PB databases
  - E.g. LSST will catalogue ~35 Billion objects
  - Not possible to (affordably) host on standard database servers
  - Nor practical to analyse data on local workstation
- Motivated inclusion of Database Testbed service in STFC e-Infra proposal (2017—2018)
  - Engage scalable (cloud) technologies to provide flexible platform for large-scale data repositories
  - Aim to address problems in astronomy
  - Identify/ integrate data-intensive applications from elsewhere in STFC community

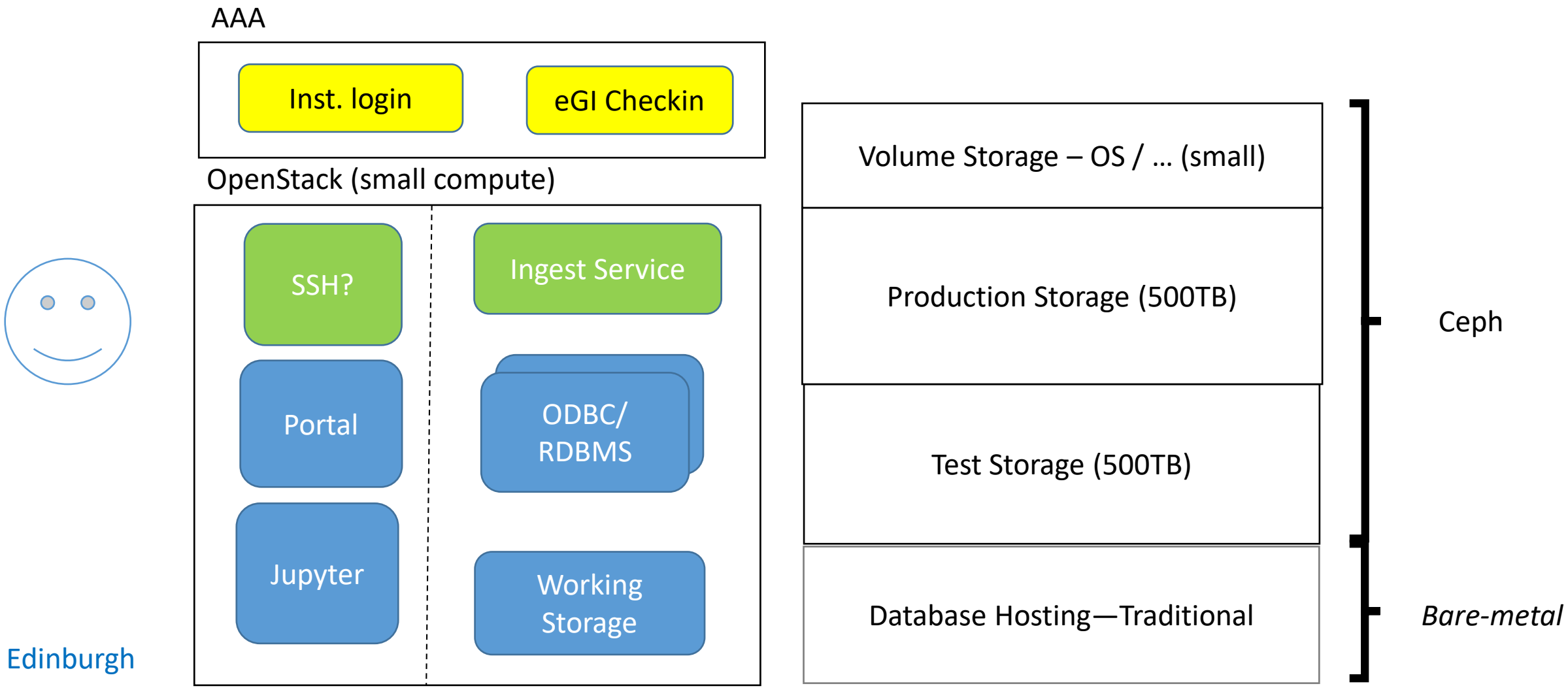
STFC Enabling e-Infrastructure Proposal



e-Infrastructure Pilot (2018—2019)

- Funding to set up 1 PB (usable) database storage
  - Guidance from other operators (RAL, Pawsey, UoE)
  - Fast network (40 GbE)
  - SSDs for Journals
  - Augmented with compute, funded by LSST:UK
- Deployed as two Ceph clusters
  - To allow tuning of configurations
  - Experiments based on OpenStack and Ceph (block storage vs. CephFS, block sizes, erasure code vs replication, …)
  - Best way to set up user-facing services (IP, interface, etc.)
- Initial tests using in-house relational databases suggest same order (!) of performance as bare-metal
- Early-adopter phase
  - Zwicky Transient Facility
  - Notes of interest from UK-DES, PanSTARSS, VVV, VIRGO




AAA
Inst. login
eGI Checkin
OpenStack (small compute)
SSH?
Ingest Service
Portal
ODBC/
RDBMS
Jupyter
Working
Storage
Volume Storage – OS / … (small)
Production Storage (500TB)
Test Storage (500TB)
Database Hosting—Traditional
Ceph
Bare-metal
Edinburgh
Man/ RAL/ Cam
…
OpenStack (big compute)
Compute
Transient/ Persistent Storage ???

IRIS High-performance Database Service

- Infrastructure available in 2020 RSAP call
  - Anticipate capacity for ~10 databases, in first instance
  - Ingestion to be tested
- Deploying Scientific OpenStack as 'interface‘
  - Working closely with Scientific OpenStack (Cambridge, Manchester, StackHPC)
- Planning to deploy Data Repository Dashboard
  - Bid to 2019 Other Digital Assets call
  - Augment Scientific OpenStack interface with database-orientated functionality (based on Trove project)
  - Streamlined management of databases through Horizon / command line
  - Support for range of relational/ NoSQL technologies
- Future work
  - Prototype infrastructure for low-latency event stream processing
    - Trialling stream processing technologies (SMACK)
  - Federation of cloud resources
    - Limited compute at Edinburgh node, so engage cloud resources elsewhere on IRIS / spill over into Commercial cloud

Thank-you

Database Service System Details

- Ceph Cluster – Storage for both the VM images and the database storage:
  - 8 x OSD Servers
    - Dual Intel Xeon E5-2650v4 2.20GHz CPUs (48 cores with hyperthreading enabled); 192GB RAM; 20 x 10TB 12Gbps SAS disks for data; 2 x 1TB SAS disks for OS (RAID1); LSI 9300-8i SAS HBA; Intel P4510 2TB NVMe SSD – Used to provide Ceph Journals for OSDs; Mellanox Dual Port 40Gb Ethernet Adapter
  - Ceph Mimic (13.2.2) installed on all nodes, nodes are running Ubuntu 18
  - Ceph Layout – machines effectively divided into two independent clusters of 4 servers (OSD servers also run MON, MDS, MGR services)
    - Cluster 1 – RADOS Gateway and block storage. Configured with replication to prevent data loss. OpenStack VMs/containers with OS image devices used for large database files on block devices
    - Cluster 2 – CephFS – “test” cluster, used to verify the optimal configuration for storing/running databases files. Configured to serve CephFS, using erasure coding
- Scientific OpenStack Deployment
  - 1 x controller server, and 4 x OS hosts
    - Dual Intel Xeon E5-2650v4 2.20GHz CPUs (48 cores with hyperthreading enabled); 256GB RAM; 2 x 1TB Enterprise SATA disks for OS (RAID1); Mellanox Dual Port 40Gb Ethernet Adapter
  - OpenStack has been deployed according to Scientific OpenStack recommendations
    - Machines provisioned manually (plans to use Kayobe in future)
    - Services deployed using Ansible Kolla
    - Cinder with connection to the Ceph cluster #1. To do, deploying Manila for access to the CephFS cluster

OpenStack Data Repository Dashboard

- Data Repository Dashboard for IRIS – aim to IRIS activities (and users) web-based and command-line interfaces to easily deploy and manage data repositories, making it straightforward and streamlined to publish substantial data via topical relational and "no-SQL" database platforms. We will:
  - Implement Trove configuration compatible with Scientific OpenStack and IRIS AAAI service(s)
  - Augment IRIS VM Image Repository with common database-configurations
  - Define and implement ingest mechanism by which substantial datasets may be ingested
  - Define potential mechanisms for fine-grain resource provisioning to be delegated by Experiments to Users
  - Acceptable User Policy between Experiments and Users, for database hosting, access, resource allocation/ recovery, data security and backups, and data migration