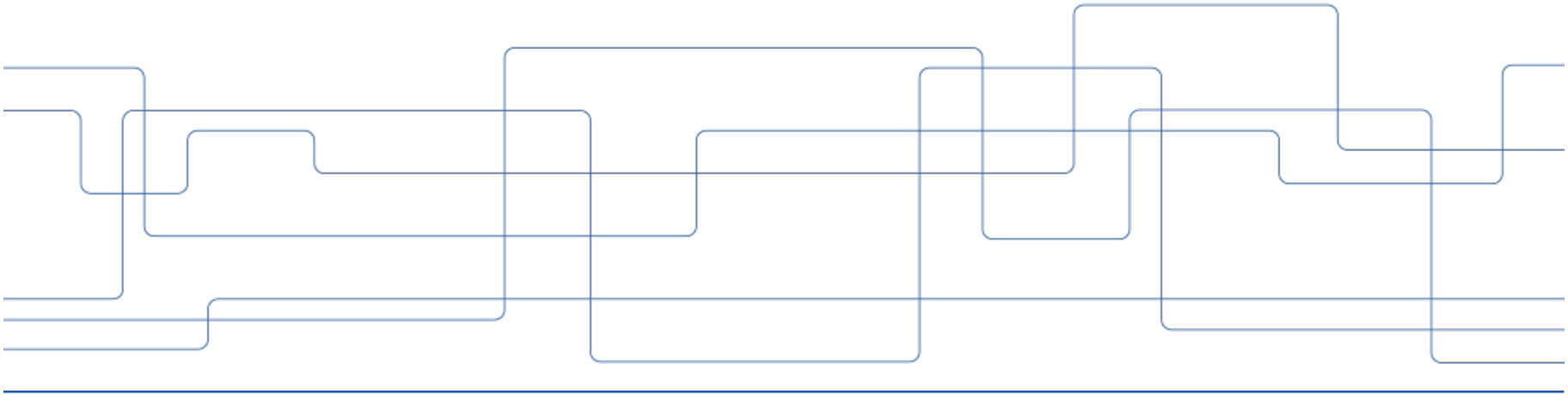
KTH ROYAL INSTITUTE
OF TECHNOLOGY

iRODS at KTH and SNIC - Status and Prospects

Ilari Korhonen

iRODS Users Group Meeting 2019, June 24th, Utrecht, Netherlands



Distributed National Infrastructure

- KTH Royal Institute of Technology in Stockholm is one of Europe's leading technical and engineering universities, founded in the year 1827
- For SNIC (Swedish National Infrastructure for Computing), we at the KTH supercomputing center PDC are jointly operating a national research data storage platform alongside with the supercomputing center NSC from Linköping University
- Development work for this distributed platform has been going on in several stages in the last few years, and finally was approved into production in 2017
- During the year 2017 we were able to procure the necessary hardware for somewhat large scale (although still entry-level in this context) use to start proper operations
- The entry level: dual replicated ~ 1 PiB of storage backing the iRODS-enabled distributed storage environment in a geo-replicated configuration: between PDC and NSC

Asymmetric and Asynchronous Geo-Replication

- Due to the differences in the infrastructures and procurements between the centers we were essentially forced to work in a very much asymmetric configuration - different hardware vendors, platforms, file systems, etc.

- Procurement at KTH PDC (in Stockholm) spec'd at approx. 1 PiB of user-visible storage with emphasis for throughput up to and exceeding 24 GiB/s

- Procurement at Linköping NSC (in Linköping) spec'd as well approx. of 1 PiB user-visible storage

- KTH PDC procurement resulted in a purchase of an IBM ESS system with 100 Gbps EDR InfiniBand connectivity towards the iRODS servers

- Linköping NSC procurement resulted in a purchase of HPE servers + SAS JBOD configuration running on top of ZFS file systems hosted by several independent iRODS servers

Asymmetric and Asynchronous Geo-Replication

- In practise, we ended up having a performance tier (or a landing zone) and a replication tier, with the other center
- For performance reasons, even without this asymmetric configuration of resources, we were essentially forced to operate in an asynchronously replicated configuration between the resources
- In practice this means: when a user object has been successfully written into the primary resource, a checksumming job will be invoked asynchronously from the iRODS rule engine, and afterwards the landed objects will be replicated to the secondary resource tree for high availability (with checksums)
- This is of course accomplished via the use of the (asynchronous) execution of the iRODS rule engine with delayed rules for new objects and bulk replication jobs between resources

Asymmetric and Asynchronous Geo-Replication

```
$ ilsresc pdc-gpfs
pdc-gpfs:passthru
└── pdc-gpfs-random0:random
    ├── pdc-gpfs-fs0-random0:random
    │   ├── fs0resc0
    │   ├── fs0resc1
    │   ├── fs0resc2
    │   └── fs0resc3
    └── pdc-gpfs-fs1-random0:random
        ├── fs1resc0
        ├── fs1resc1
        ├── fs1resc2
        └── fs1resc3

$ ilsresc nsc-zfs01
nsc-zfs01:passthru
└── nsc-b:random
    ├── nsc-b02p0
    └── nsc-b03p0
```

Asymmetric and Asynchronous Geo-Replication

- Naturally, this setup of ours will welcome other Swedish HPC centers to join in our venture and possibly even leverage features such as data locality
- Even if we are currently “only” replicating between two centers, if more partners were to emerge we of course would we able to deploy iRODS rules to easily accommodate a few more replication partners within the same zone
- We are naturally inviting more of our partnering Swedish supercomputing centers to participate in this nation-wide data management initiative of ours

Local Access from HPC Resources at KTH PDC

- We have (of course) provisioned iRODS clients to our local HPC compute environments to enable our users to login to the iRODS environments from the compute clusters available at our center - via our local Kerberos KDC

- There are (of course) two separate administrative domains: SNIC Swestore (`SWESTORE.SE`) and likewise KTH PDC (`NADA.KTH.SE`)

- In practise: the two Kerberos V5 realms are currently in a trust between: `NADA.KTH.SE -> SWESTORE.SE`

- In practise: the users from KTH PDC (which get imported to the SNIC iRODS after project approval at SNAC - Swedish National Applications Committee) - get their local KTH PDC Kerberos credentials approved as an authentication method

Local Access from HPC Resources at KTH PDC

- The Challenge: to enable our HPC users to offload their data from our high performance (compute) file system to the nationally accessible storage
- For this we have been collaborating with the iRODS Consortium (with whom we are members with) to test and further develop their Lustre / iRODS interface
- This would be a game-changer for us to be able to finally offload our so called heavy-weight (several hundred terabytes per project) users out of our primary high performance filesystems into a more suitable mid-term storage
- For example: at KTH PDC we have approx. 5 PiB of Lustre FS capacity - from which over 90% is full currently

Lustre and iRODS

- Progress has been made by the iRODS Consortium to further integrate a Lustre file system into an iRODS data grid - especially to cooperate with tiering
- Current approach: is to enable the change log in the Lustre file system in question and to activate a change log listener for the said file system to register the events which have taken place in the file system during the time period
- This is of course the same approach previously used by tools like `robinhood` developed at CEA in France

Lustre and iRODS

- Challenges remain:
  - the enabling of the Lustre change log slows down the metadata performance of the file system, since more writes are subject to the MDT(s)
  - also the “draining” of the Lustre change log becomes vital to the health of the file system itself, since the entries need to be acknowledged to be removed from the MDT(s) - and to prevent the file system metadata to flood
  - so what we need is a receiving end of the change log which would be basically guaranteed to remain in sync with the file system itself

Future Developments

- We would hope to be able to publish data from iRODS
  - For this there seems to be new code available :)
- We are also involved in the European PID Consortium EPIC at PDC
- Hopefully in the future we would be able to publish data sets out of our iRODS and reference those with European PID(s) such as EPIC HANDLE(s)

Thank you! Questions?