Optimizing Bioinformatics Workflow and Organizing Big Data with iRODs and Amazon S3

William Dixon, Victoria Cao, Andy Keen

Michigan State University

Objective

Due to the large amount of data involved in bioinformatics research, it has become increasingly necessary to find an effective way to move, organize, and analyze this data. To automate **metadata collection**, **efficiently query data**, and **customize data storage hierarchies**, we are experimenting with integrating OSIRIS's **Amazon S3-compatible services** with an **iRODS** (Integrated Rule-Orientated Data System) middleware server, and making these resources available to MSU's supercomputing center.

Methods

1. **Automated Ingest**
   After creating a **cacheless S3 resource** using **iRODS,** the **iRODS Celery Client** automatically ingests data from specific S3 buckets under one **Ceph** user (figure 1).

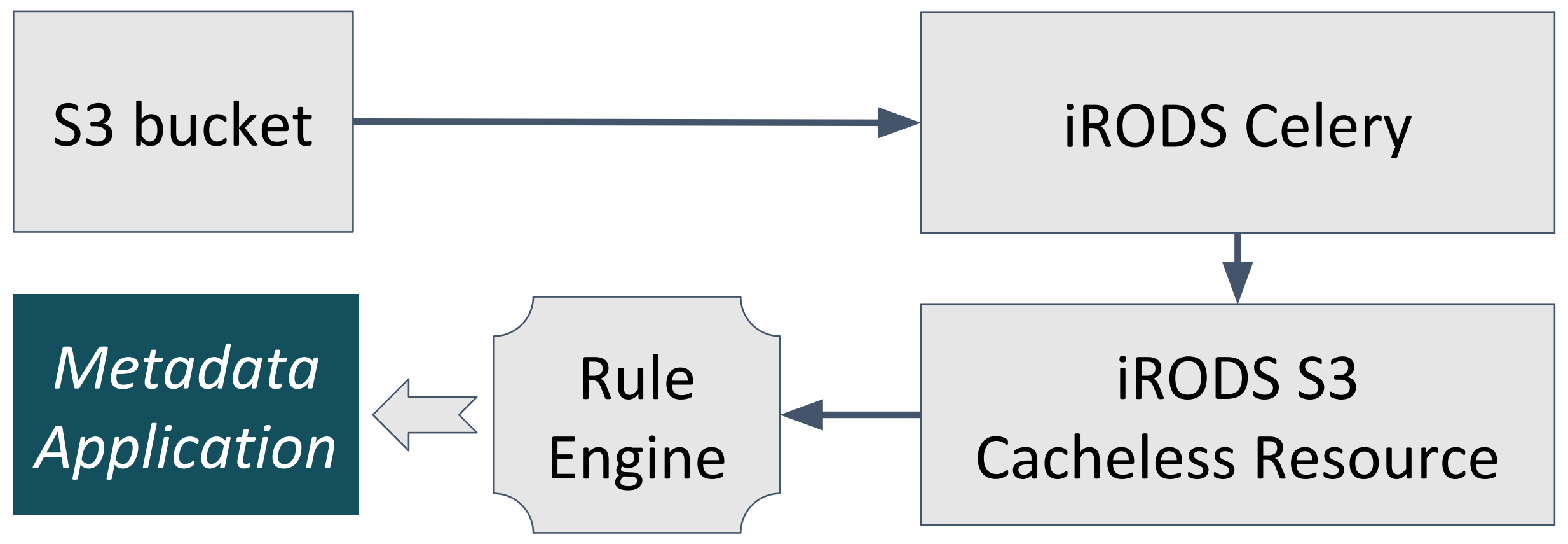

Figure 1: Workflow for automated virtual object data path ingest from Amazon S3 to iRODs resource

2. **Data Processing and Query**
   The **iRODS rule engine**, as seen in figure 1, applies metadata specified by a configuration file, and can be done independent of automatic ingest. To manage, view, and query virtual resource hierarchies and their data objects, we use **iRODS icommands,** a linux-like interface.

3. **Automated Output**
4. Specified data objects are replicated from the original resource to the S3 resource tethered to a new S3 bucket, using **icommands** (figure 2). Then, the Python Module **Boto3** manages the access control list (**ACL**) policies of a newly created output bucket, adding permissions to alias users under a specified **Ceph** user or other group users.

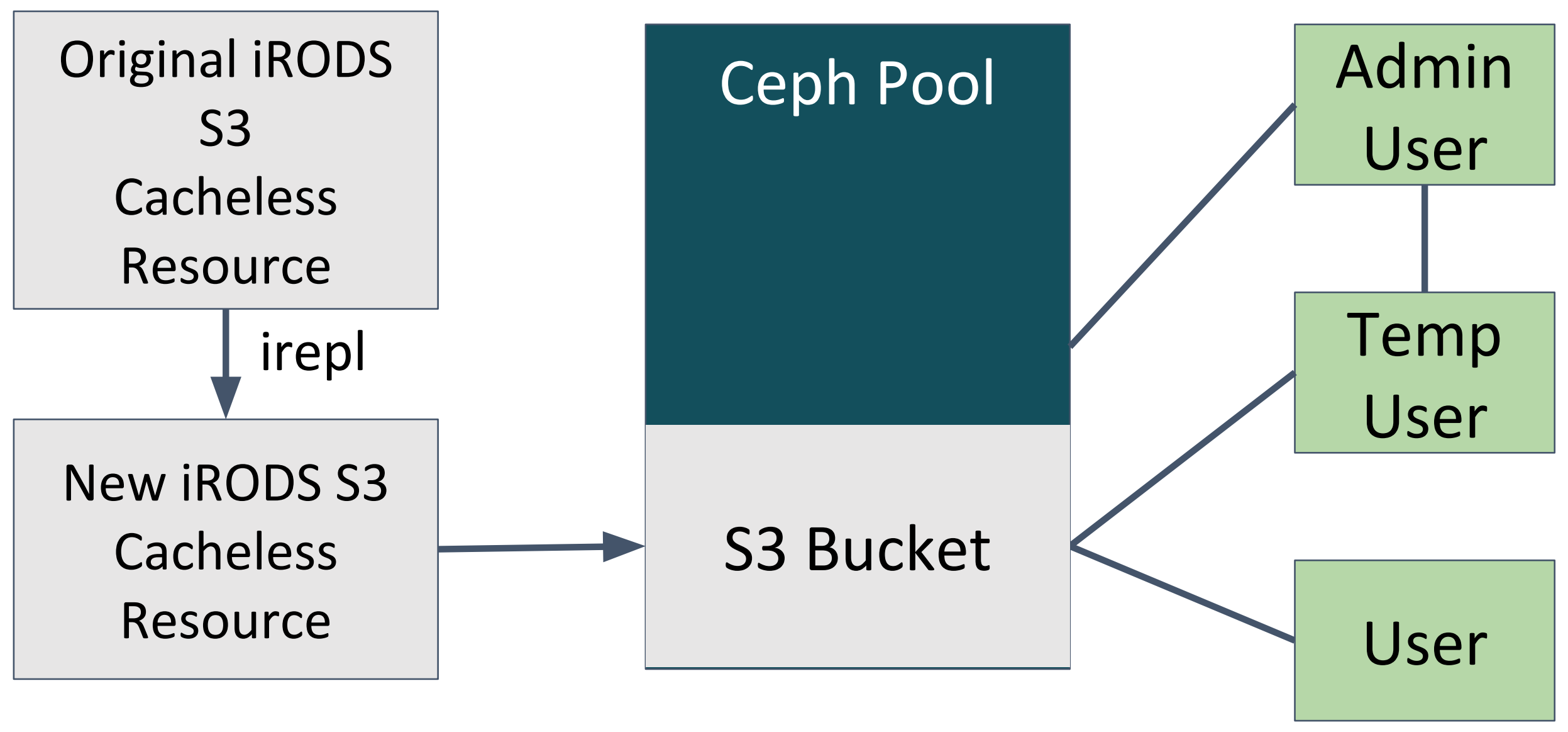
Original iRODS S3 Cacheless Resource
irepl
New iRODS S3 Cacheless Resource
Ceph Pool
S3 Bucket
Admin User
Temp User
User

Figure 2: Workflow for automated output from iRODS resource to new Amazon S3 bucket with multi-user access

Results

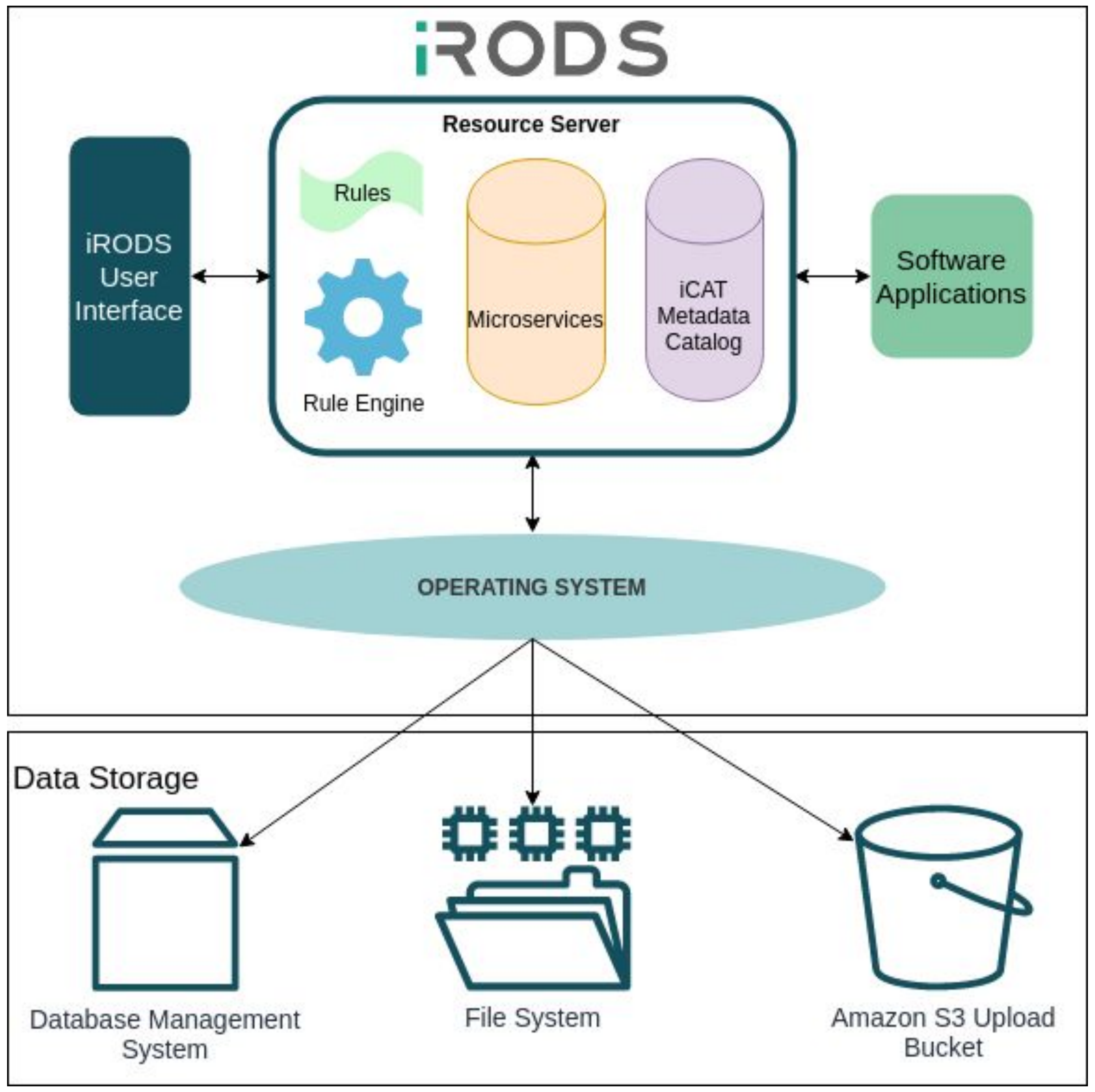

Figure 3: Outline of iRODS architecture with S3 storage application

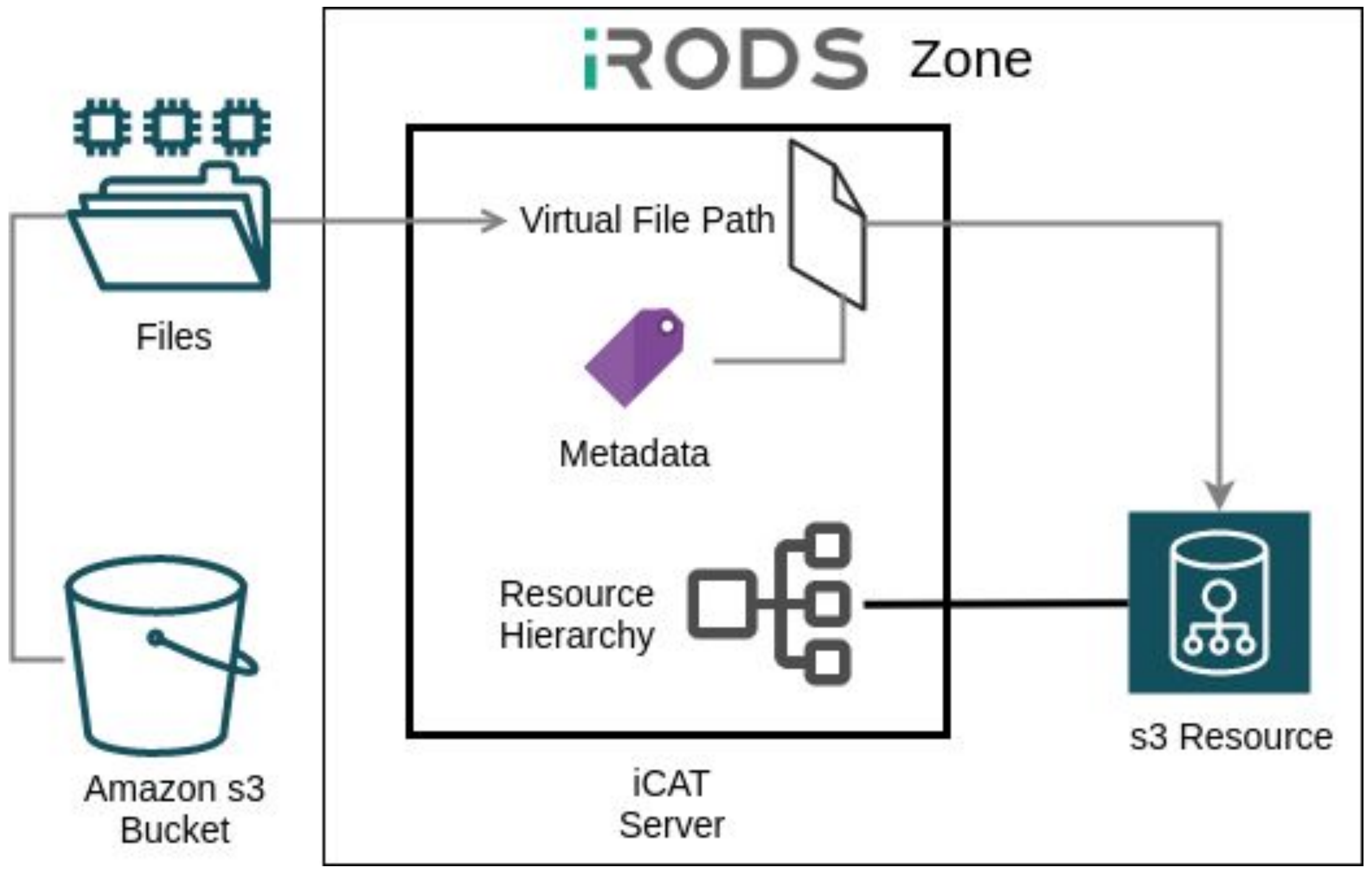

Figure 4: Outline of iRODs zone configuration with S3 storage

Conclusion

We were led to multiple deductions:

1. The S3 Cacheless Resource in iRODS, though relatively new, proves innovative in its **removal of extraneous copying processes** in data transfer and metadata application
2. For file sizes stretching into hundreds of gigabytes, replicating metadata and data between cacheless resources provides an **accessible way to customize the application of metadata**.
3. As the size of data in bioinformatics and other research fields increases, so will the **necessity of efficient data transfer and storage**. Thus, in the foreseeable future, similar optimization platforms will likely surface.

Discussion

To prepare our iRODS implementation for deployment in bioinformatics workflow between **MSU** and the **Van Andel Institute**, there are a number of tasks that must be completed:

1. Opening iRODS server(s) to be accessible **remotely** (in development, all was done locally)
2. Replicating an S3 cacheless resource to another S3 cacheless resource **without errors** for distribution into output buckets

References

ceph

iRODS

amazon S3

Acknowledgements

NSF CC*DNI DIBBs: Multi-Institutional Open Storage Research InfraStructure (MI-OSiRIS) - OAC1541335.