PROPOSAL FOR NEXT GENERATION ESGF SEARCH SERVICES

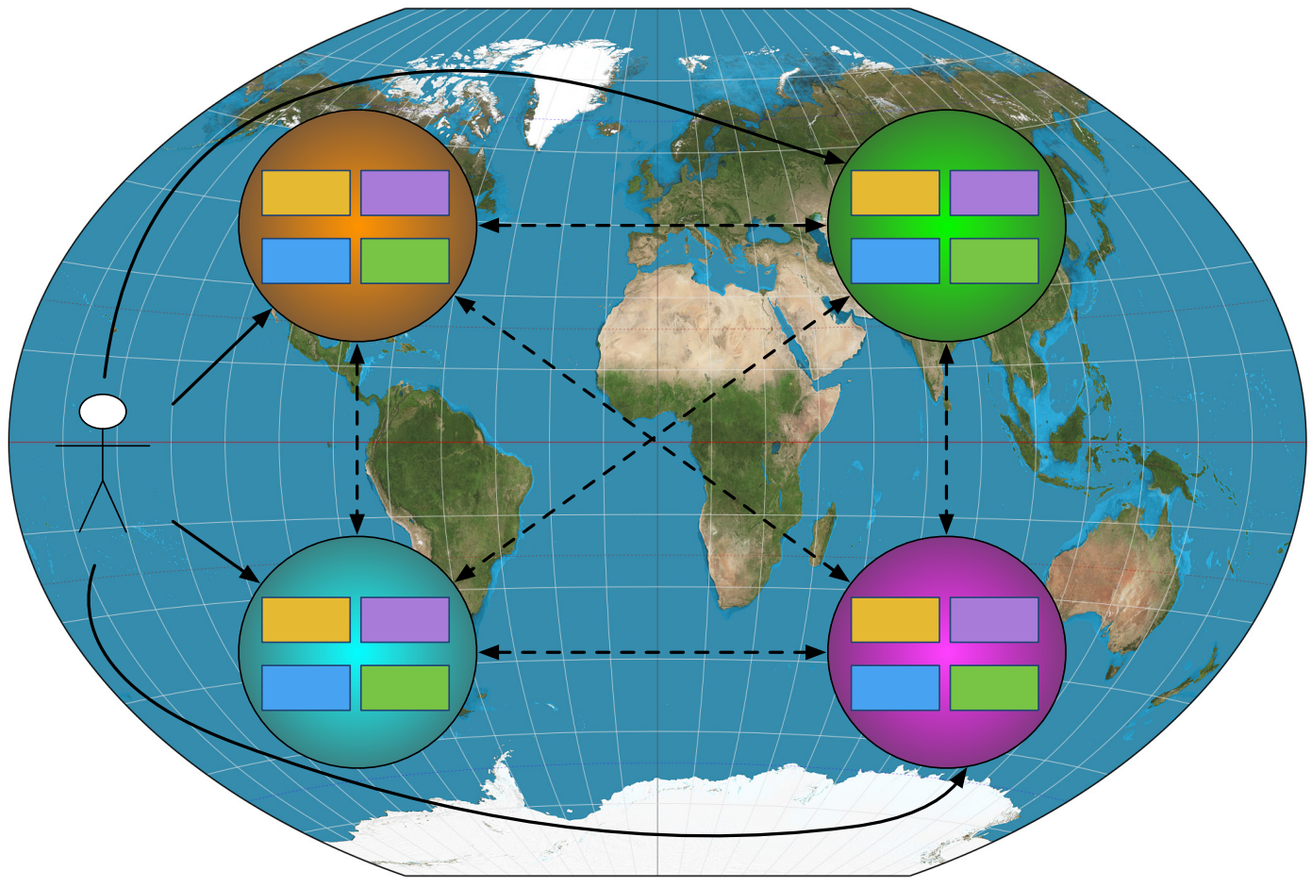
LUCA CINQUINI

NASA JET PROPULSION LABORATORY AND CALIFORNIA INSTITUTE OF TECHNOLOGY

JPL UNLIMITED RELEASE SYSTEM CLEARANCE NUMBER: # URS278802

Current ESGF Search Architecture

- Enables local administration of metadata catalogs, yet federation wide searches
- Based on Apache Solr, leverages functionality for distributed searches and replication
- Each node replicates the catalogs of all the other nodes to resolve searches locally
- A client can query any of the nodes in the federation and obtain the same results

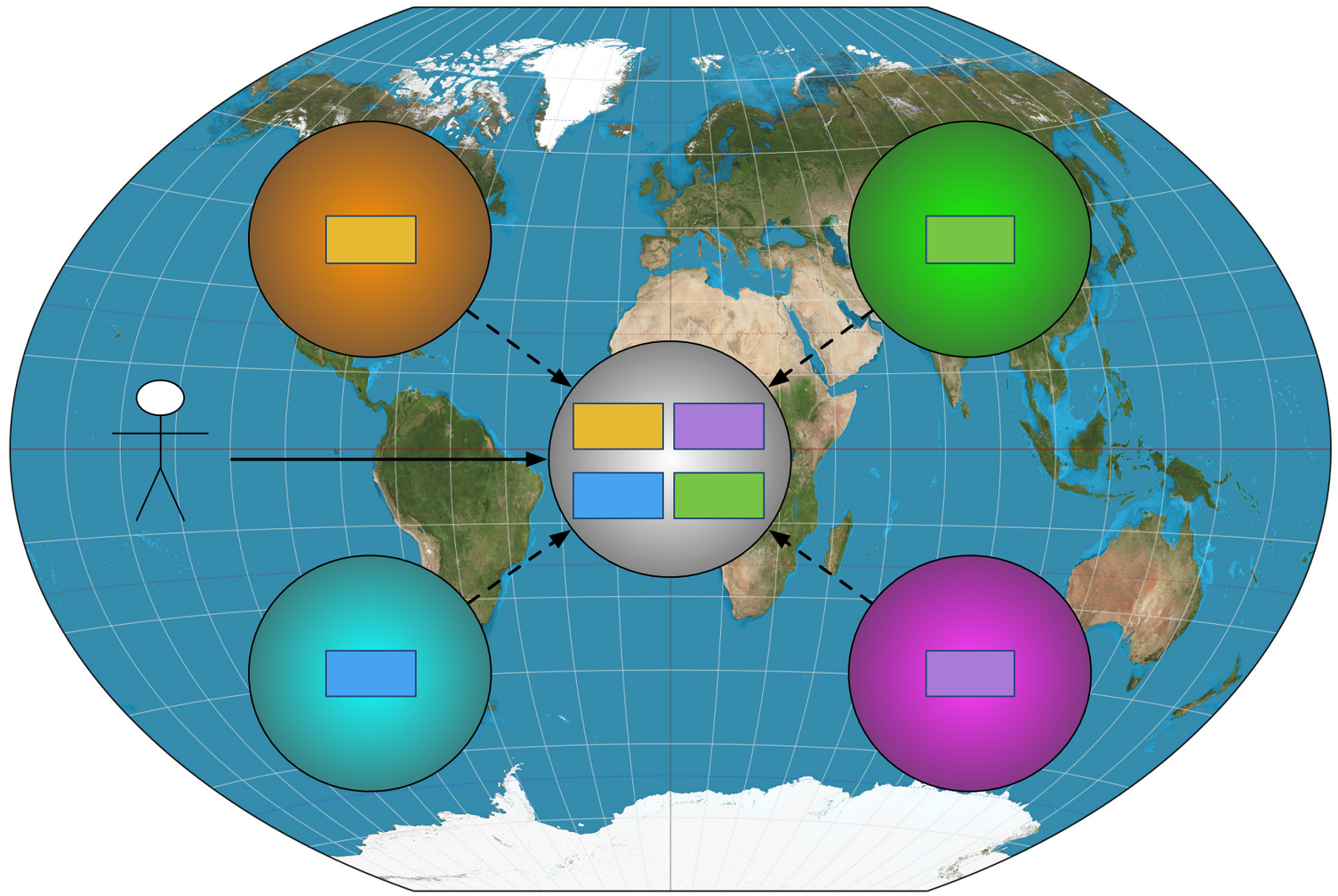


Current Shortcomings

- Each node administrator must manually configure a replica shard for all other nodes in the federation
- High potential for inconsistencies across nodes (for example, if one replica breaks at one node)
- All nodes must scale concurrently when the federation grows (number of index nodes or metadata holdings at each node)
- The current Solr installation is becoming obsolete and insecure, yet it is difficult to upgrade:
  - All sites must upgrade simultaneously for replication to keep working in both directions
  - Data must be re-indexed to upgrade the underlying Lucene version

Proposal for Next Generation ESGF Search Architecture

- Let each institution maintain only one index node where they publish their data (i.e. no replica shards)
- Establish a few “super-indexes” that aggregate metadata from all institutions
- Point all client applications to the super-indexes

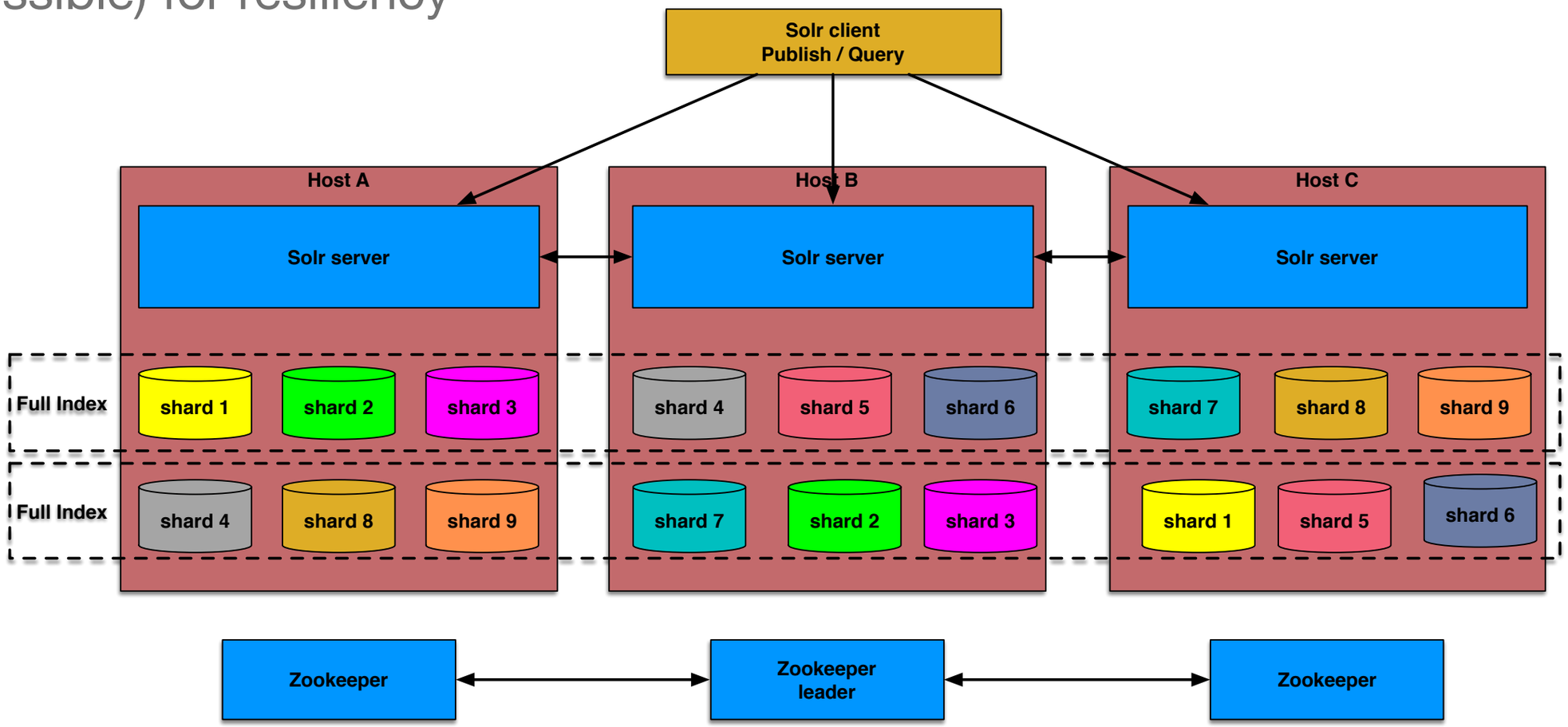


Technical Implementation

- Adopt Solr Cloud
- Deploy as Docker and Kubernetes
  - On the cloud, or in-premise
- Harvest and sync

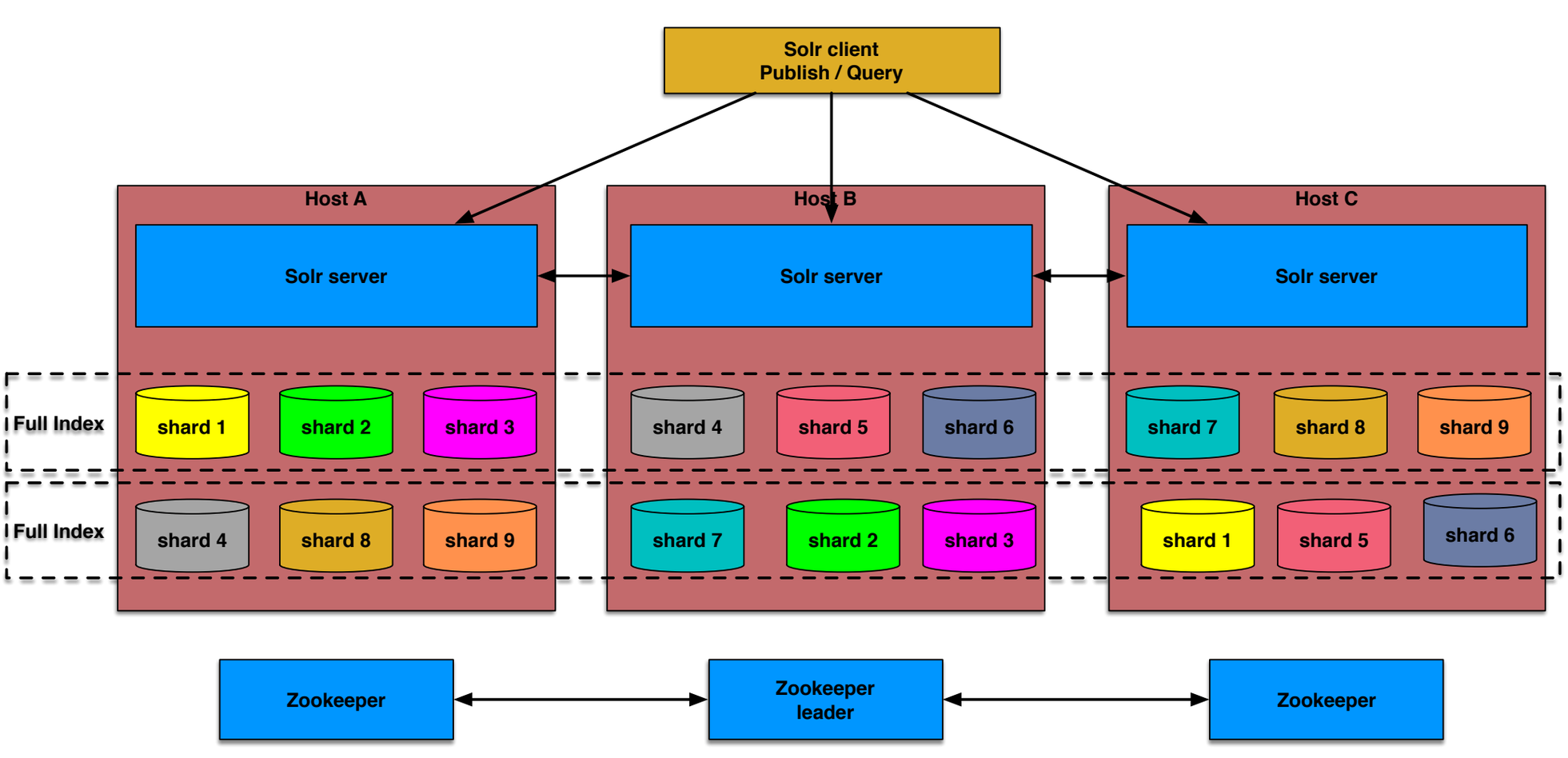
Solr Cloud

- Solr Cloud is a more advanced and scalable Solr architecture, designed to be deployed on a multiple hosts
- The full metadata index is partitioned into logical shards
- Each shard is physically instantiated as one or more replicas
- Replicas are automatically deployed onto Solr instances on separate hosts (if possible) for resiliency



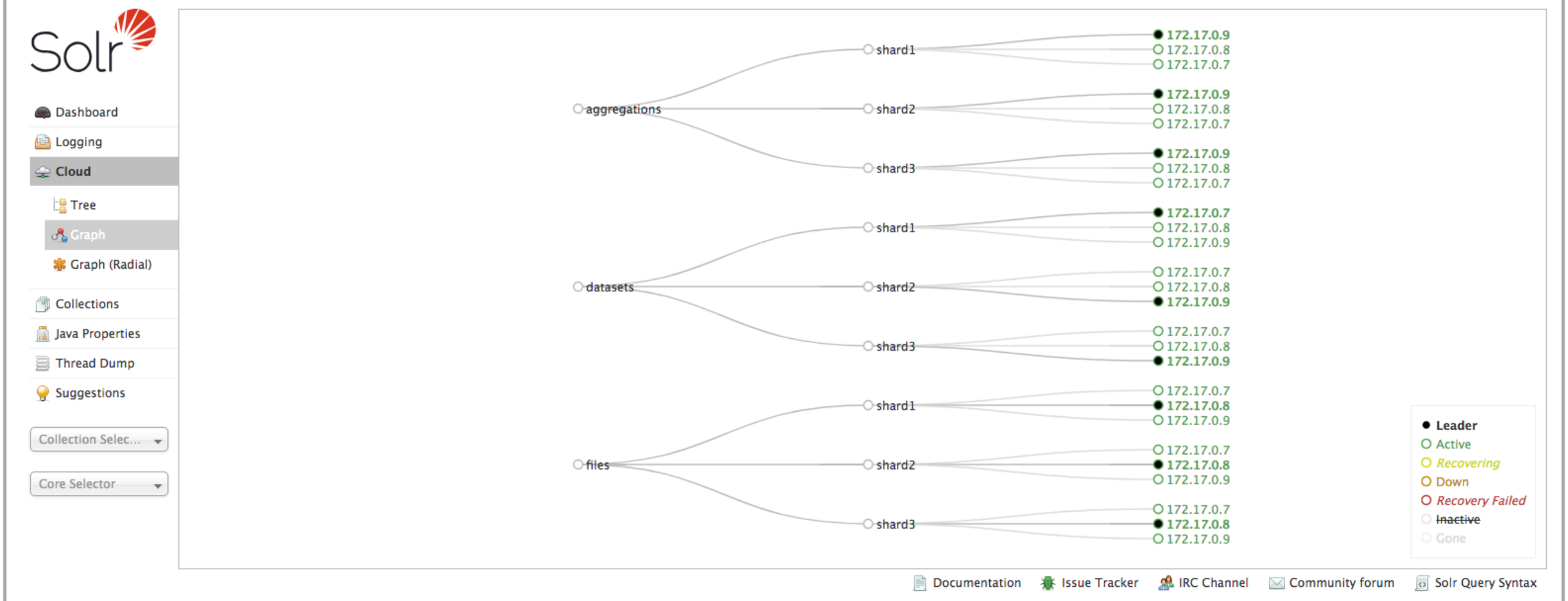
Solr Cloud

- Metadata can be published to any Solr instance and it will directed to the proper shard leader, then replicated (“distributed indexing”)
- Clients can query any Solr instance, and the query will be load balanced and resolved versus a complete set of shard replicas (“distributed querying”)
- A set of Zookeeper servers provides centralized configuration management



Prototype Deployment on AWS

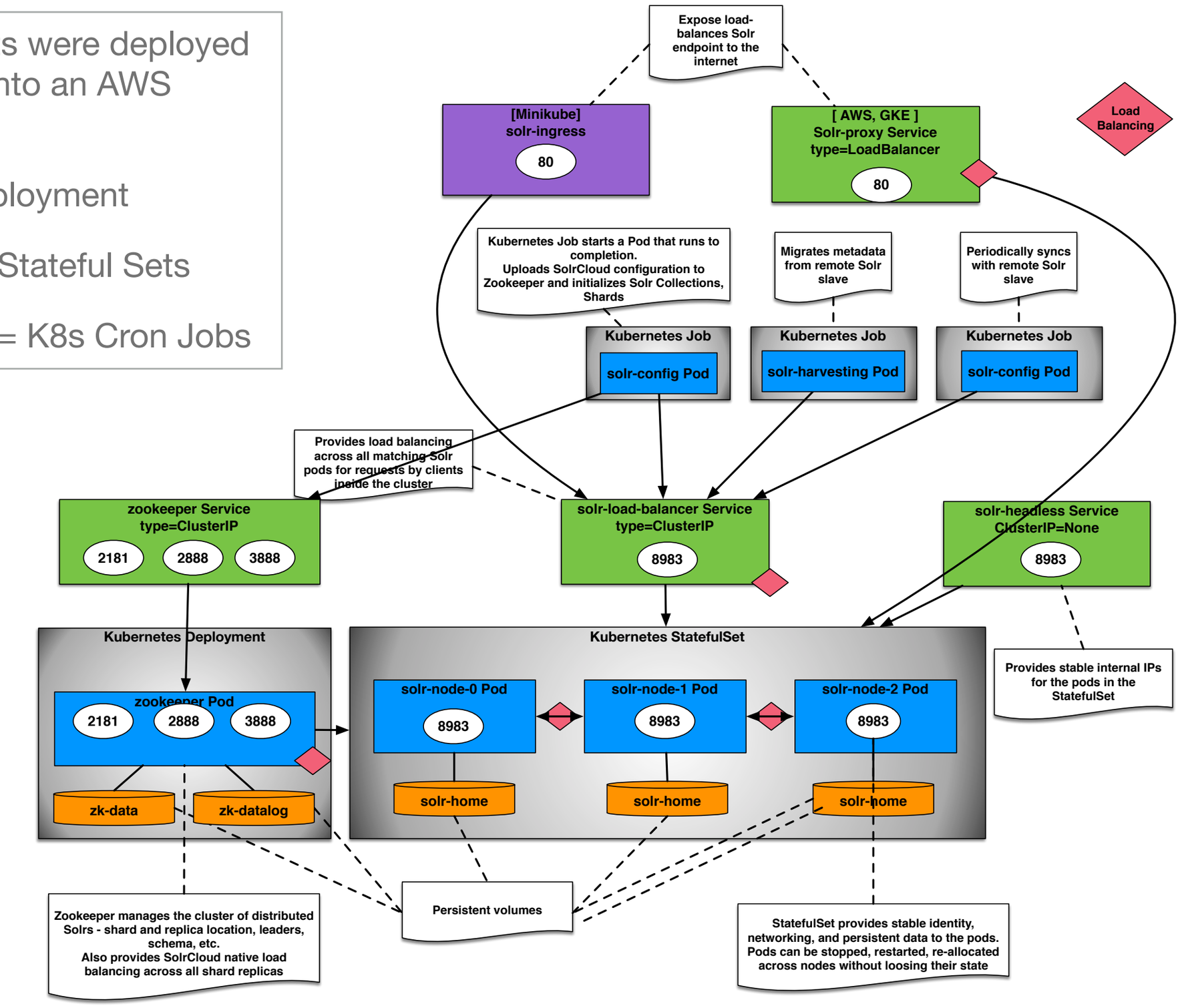
- Small cluster of 3 EC2 instances of type t2.medium (2 CPUs, 4GB memory)
- Solr configuration: 3 shards per collection, 3 replicas per shard
- Tracking ESGF global archive for over 2 months



Docker and Kubernetes

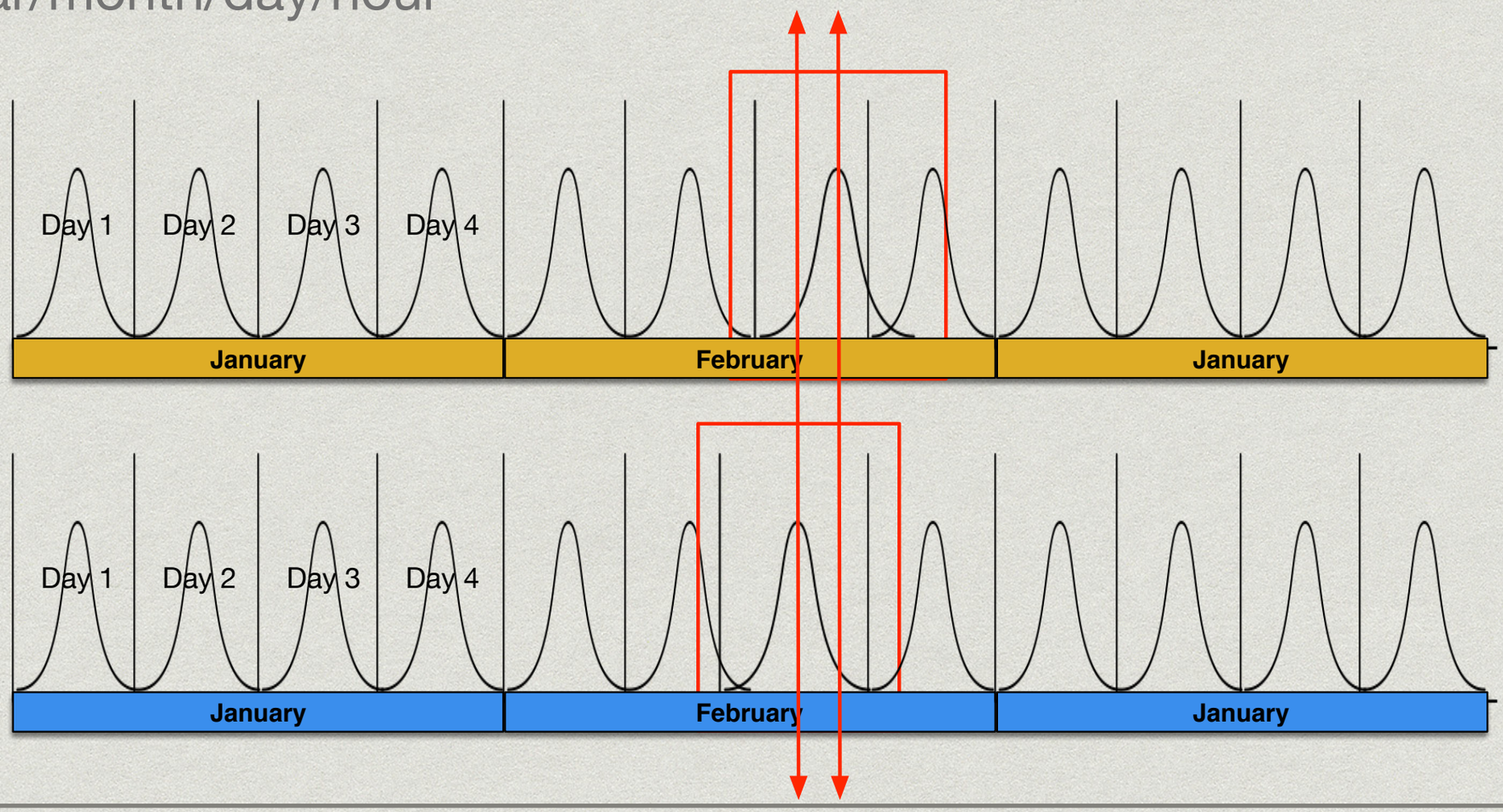
All software components were deployed as Docker containers onto an AWS Kubernetes cluster

- Zookeeper = K8s Deployment
- Solr instances = K8s Stateful Sets
- Harvest/Sync clients = K8s Cron Jobs



Harvesting and Syncing

- Harvesting clients are run initially to read all records from each existing master node into the super-index
  - May take up to several days for large indexes
- Syncing clients are run every hour to sync a remote index to the super-index
  - Algorithm uses timestamp stats to compare the indexes by time interval - year/month/day/hour



The ESGF super-index on AWS

esgfpy-solr — -bash — 85×46

```
(aws) kubectl get pods
NAME                                READY   STATUS      RESTARTS   AGE
solr-node-0                         1/1     Running     0          4d
solr-node-1                         1/1     Running     0          21d
solr-node-2                         1/1     Running     0          33d
solr-sync-ceda1-1544082300-2kpck    0/1     Completed   0          2h
solr-sync-ceda1-1544085900-b4vwg    0/1     Completed   0          1h
solr-sync-ceda1-1544089500-pp677    0/1     Completed   0          26m
solr-sync-ceda3-1544082000-hnvpq    0/1     Completed   0          2h
solr-sync-ceda3-1544085600-95p9p    0/1     Completed   0          1h
solr-sync-ceda3-1544089200-fwxns    0/1     Completed   0          31m
solr-sync-dkrz-1544081100-2qlpv     0/1     Completed   0          2h
solr-sync-dkrz-1544084700-5zdxv     0/1     Completed   0          1h
solr-sync-dkrz-1544088300-k64rd     0/1     Completed   0          46m
solr-sync-gfdl-1544083200-x5spc     0/1     Completed   0          2h
solr-sync-gfdl-1544086800-jstnl     0/1     Completed   0          1h
solr-sync-gfdl-1544090400-spwkg     0/1     Completed   0          11m
solr-sync-ipsl-1544080500-f2rb6     0/1     Completed   0          2h
solr-sync-ipsl-1544084100-jwfbk     0/1     Completed   0          1h
solr-sync-ipsl-1544087700-457b8     0/1     Completed   0          56m
solr-sync-jpl-1544081700-n74l8      0/1     Completed   0          2h
solr-sync-jpl-1544085300-jxnf7      0/1     Completed   0          1h
solr-sync-jpl-1544088900-6hsv7      0/1     Completed   0          36m
solr-sync-liu-1544081400-hlrbz      0/1     Completed   0          2h
solr-sync-liu-1544085000-pbqsg      0/1     Completed   0          1h
solr-sync-liu-1544088600-x2xrg      0/1     Completed   0          41m
solr-sync-llnl-1544080800-xd874     0/1     Completed   0          2h
solr-sync-llnl-1544084400-dkdj7     0/1     Completed   0          1h
solr-sync-llnl-1544088000-mq94n     0/1     Completed   0          51m
solr-sync-nci-1544082600-jg4pc      0/1     Completed   0          2h
solr-sync-nci-1544086200-7qzcp      0/1     Completed   0          1h
solr-sync-nci-1544089800-rtrgw      0/1     Completed   0          21m
solr-sync-pik-1544081700-dc22b      0/1     Completed   0          2h
solr-sync-pik-1544085300-mgcz2      0/1     Completed   0          1h
solr-sync-pik-1544088900-8pwtt      0/1     Completed   0          36m
zookeeper-5bbb44b7f9-d9578          1/1     Running     0          29d
(aws)
```

https://esgf-node.llnl.gov/esg-search/search/?offset=0&limit=10&type=Dataset

Apps | Saddle Ridge | My | Google Compute | AWS | CO2 | COG | JPL | EDRN | ESGF | tmp | DREAM | NOAA | Other Bookmarks

```
      <str name="facet">true</str>
      <str name="facet.sort">lex</str>
    </lst>
  </lst>
▼<result name="response" numFound="970199" start="0" maxScore="1.0">
  ▼<doc>
    ▼<str name="id">
        cmip3.CCCma.cccma_cgcm3_1_t63.piControl.mon.atmos.run1.ts.v1|aims3.llnl.gov
      </str>
      <str name="version">1</str>
    ▼<arr name="access">
      <str>HTTPServer</str>
```

Solr Admin

Not Secure | a1f27a255dea211e8b60502bd31928e2-1418248443.us-west-2.elb.amazonaws.com/solr/#/datasets/query

Apps | Saddle Ridge | My | Google Compute | AWS | CO2 | COG | JPL | EDRN | ESGF | tmp | DREAM | NOAA | Other Bookmarks

Analysis
Dataimport
Documents
Files
Query
Stream
Schema

df

Raw Query Parameters

key1=val1&key2=val

wt

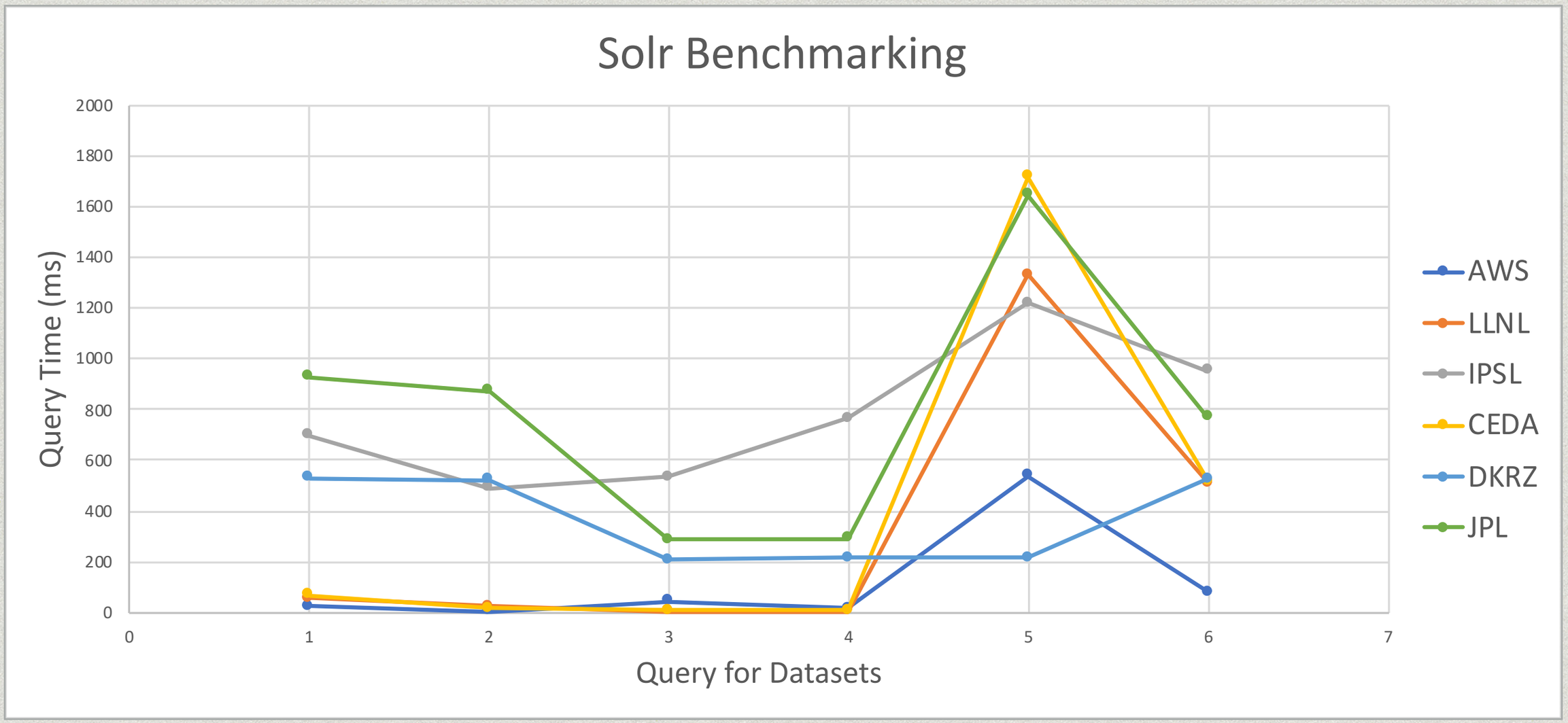
```
    "facet.sort":"lex",
    "_":"1544091627434"}},
"response":{"numFound":970199,"start":0,"maxScore":1.0,"docs":[
    {
      "id":"cmip5.output1.NASA-GMAO.GEOS-5.decadal1960.mon.atmos
      "version":"20160317",
      "access":["HTTPServer",
        "OPENDAP",
        "LAS"],
```

Advantages

- Automatic distributed indexing, querying and load balancing
- Resiliency and automatic failover
- Horizontal and vertical scalability
  - Add more servers and/or increase the memory of each server
  - The system can be scaled by increasing the resources at one location, not at all sites through the federation
- Upgrades can be executed by bootstrapping a new system in the background, and switching over the proxy when ready

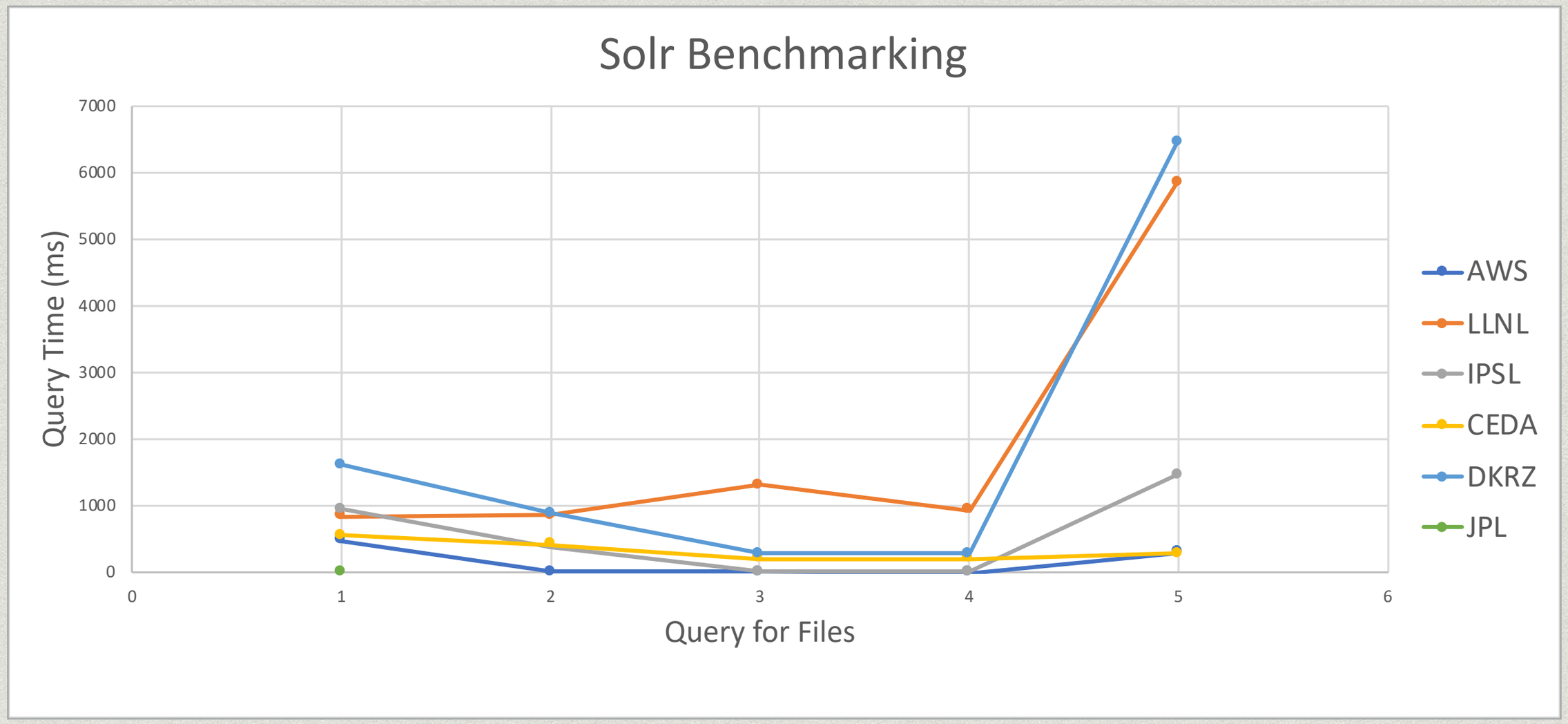
Benchmarking: Datasets

- Using “super-index” deployed on small AWS K8s cluster
- 3 EC2 instances of type “t2.medium” -2 CPUs, 4GiB memory



~1M Datasets

Benchmarking: Files



~18M Files

Conclusions

- Proof of concept successfully executed
- Software stack is ready for operational deployment as beta service
- Need to find resources - on the cloud or in-premise
- A timely deployment is recommended to enable smoother upgrades during CMIP6 operations