ESGF SEARCH: EVOLUTION

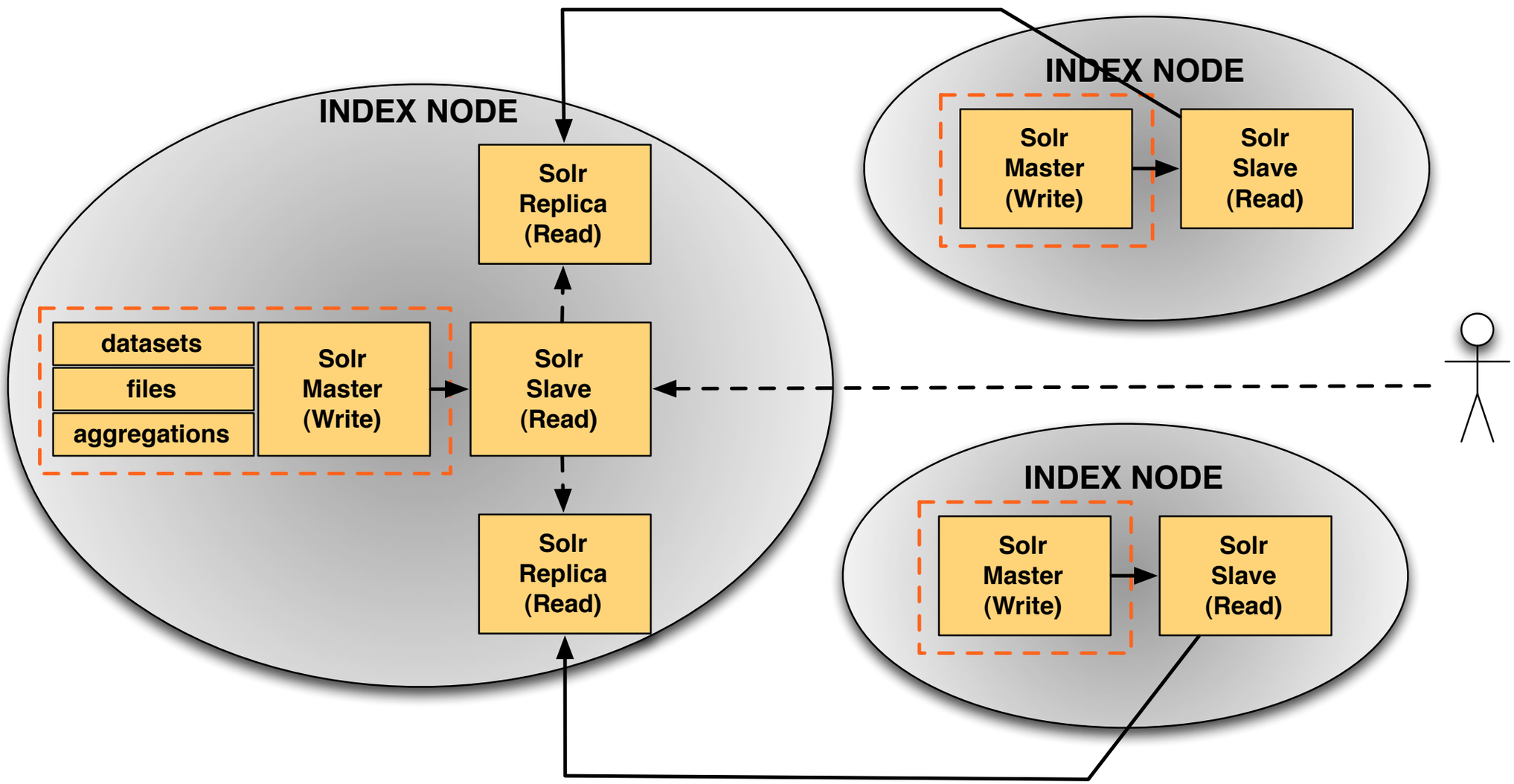
ESGF F2F Workshop,
Livermore, CA, December 2014

Luca Cinquini
California Institute of Technology & Jet Propulsion Laboratory (NASA)

JPL Unlimited Release Clearance Number: CL#14-5118

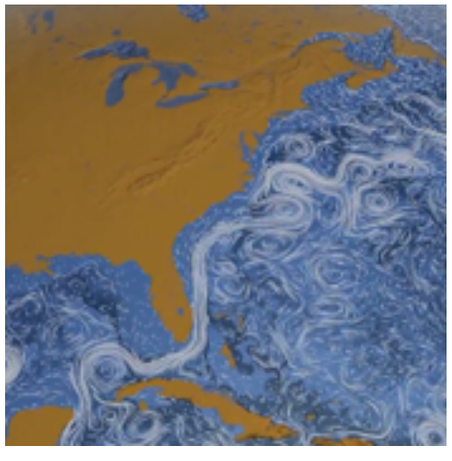
Introduction

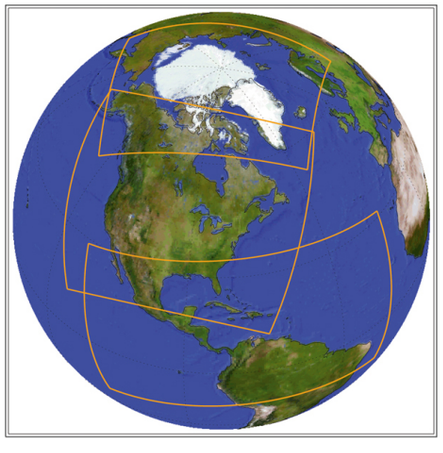
- ESGF offers a state of the art capability for searching across a federation of distributed and independent archives - no other public infrastructure has any equivalent!

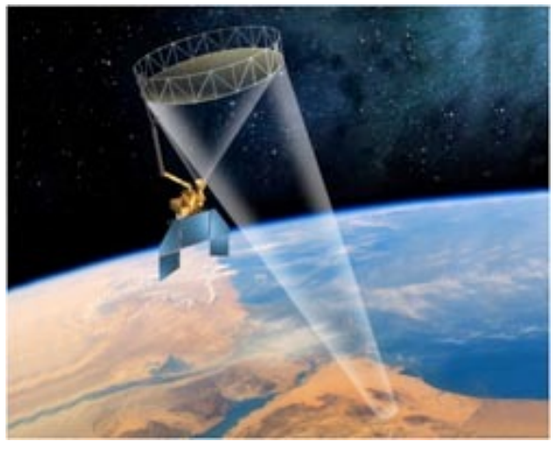


Introduction

- Nonetheless, the ESGF search infrastructure can be improved to fix some of the current problems, and <u>must</u> evolve to address the challenges of the next generation of climate data projects (CMIP6, NASA decadal surveys, etc.)
- Questions:
  - How do we improve the functionality, accuracy and reliability of the current search services ?
  - How do we scale to 10x more Index Nodes, 10x more data per Index Node, and to new projects and disciplines ?
- This talk will list the most critical areas of concern and suggest possible solutions



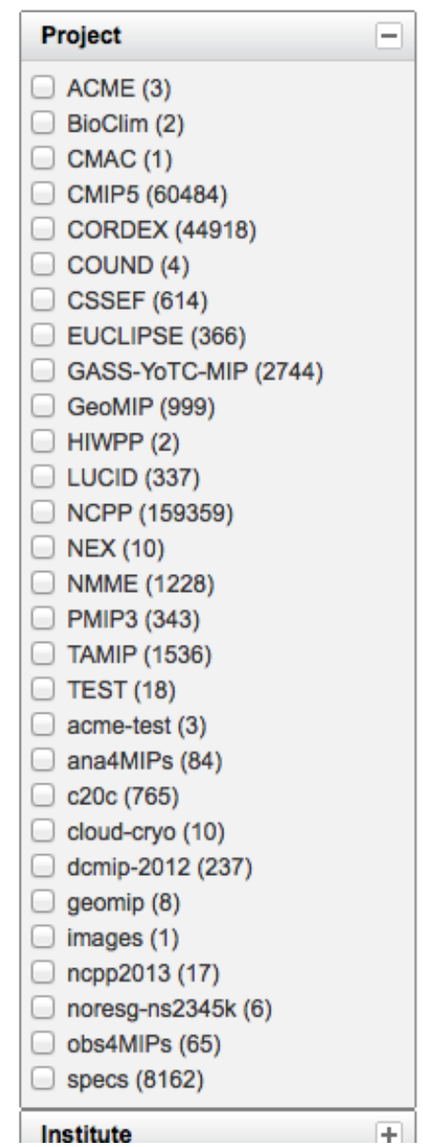


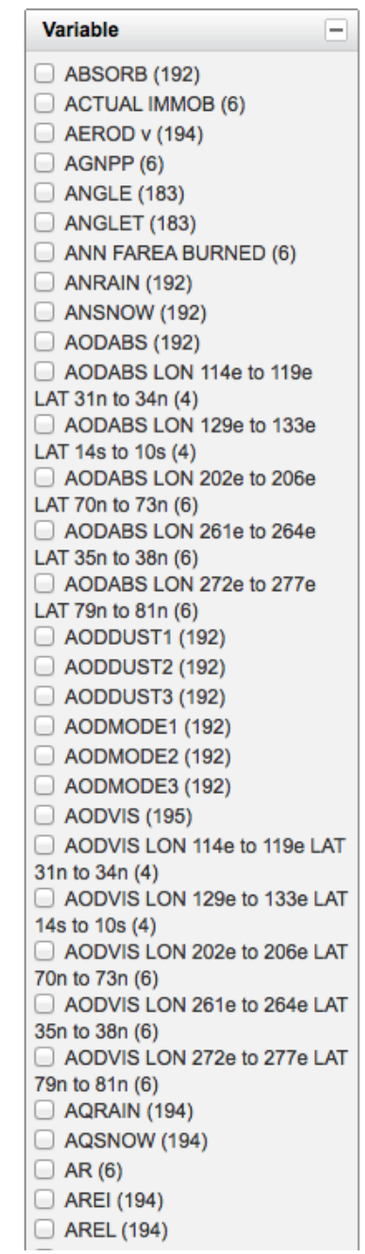


- Note: past year development has focused on improving the search UI (see CoG), but the underlying search service infrastructure has remained the same (last major upgrade: RESTful push/pull publishing services)

Proliferation of Nodes, Projects

- Problem: ESGF is being adopted by many new institutions, each publishing data from many projects, which often don't have any relation to each other
  - A global search must cover more and bigger indexes, it becomes slower
  - Each search facet presents too many options, degrading the user experience
- Solution:
  - CoG UI will help because it allows each project to define its specific search configuration: the target Index Node, one or more project constraints, specific facets
  - But to scale into the future, ESGF must partition the global search space into Virtual Organizations (VOs), a.k.a. "circles" of Nodes:
    - Each VO includes only Index Nodes with related data
    - Each VO is administered by a committee that decide which Node can join and which projects can be published
    - Each VO defines and maintains their metadata schemas and CVs





Metadata Validation

- Problem: metadata content of published datasets is often erroneous or incomplete
  - Metadata fields might be entirely missing (example: no time/space coverage)
  - Multiple spelling/case for the same facet value

- Solution: ESGF must enforce server-side validation of published metadata via project-specific schemas and Controlled Vocabularies (CVs)
  - Each project must define the list of valid facets and their multiplicity
  - Each facet can be assigned only values from its CV
  - CVs must be easy to develop and maintain by scientific (not technical) experts
  - Example: metadata schemas for CMIP5/6, Obs4MIPs, Ana4MIPs…

Upgrade to Solr4

- Problem: ESGF servers are running Solr3.6 and not taking advantage of new Solr features and performance improvements:
  - Geo-spatial searches
  - Atomic updates
    - Add QC information after dataset is already published
    - Add/update access control information
  - Solr Cloud: automatic configuration, sharding and replication

- Solution: must upgrade ESGF servers to Solr 4.10. But because Solr 3/4 indexes are incompatibles:
  - Each site must temporarily run 2 Solr indexes
  - Must find/write a tool to migrate metadata from one index to the other

Non-Standard Solr Port: 8983

- Problem: ESGF runs Solr query server ("slave" server) on non-standard port 8983 which is almost always blocked by firewalls (for both incoming and outgoing connections)

- Solution: must run Solr query server on standard web port 80
  - Run Solr slave within Tomcat server on port 80
  - Still run Solr master and optional replicas on separate Jetty servers, ports

Single-Variable Datasets

- Problem: often users complain that they cannot download only those files that match a given variable

- Solution:
  - Use filename matching expression when searching, generating wget scripts
    - CoG improves usability of filename searches
  - Republish all multiple variable datasets as single variable datasets ?
  - Push variable information to file level and use exact variable constraint when generating wget scripts ?
  - At least, mandate one-to-one dataset-variable correspondence for CMIP6

Inconsistent Search Results

- Problem: sometimes, searches initiated at different Index Nodes yield different number of results

- Solution: must first understand the cause of the problem…
  - Increase memory of Solr servers ?
  - Monitor the state of the federation by executing standard searches at all Index Nodes
  - Report any inconsistencies to the administrators

ESGF Search Roadmap

To be accomplished in roughly chronological order:

- Switch ESGF Search interface to CoG
- Upgrade all Index Nodes to Solr4, running on port 80
- Define ESGF Virtual Organizations, establish governance bodies
- Define, maintain and enforce metadata schemas and CVs

- Republish all data ?
  - Split multiple variables datasets into single variable
  - Generate time/space coverage metadata
  - Publish additional endpoints (OpenDAP, GridFTP, LAS)
  - Validate metadata
  - Most helpful exercise to prepare for CMIP6

- Establish monitoring/notification capabilities
- Develop widget for space/time search
- Revise ESGF Search documentation
- Cast ESGF Search as WPS service ?