CMIP6 Amazon-hosted Data Informational Session

Aparna Radhakrishnan
Kristopher Rand
Charles Stern
Julius Busecke

CMIP6 Data Informational session, Oct 13th, 2021

Community driven effort

Special thanks to..

Ryan Abernathey, V. Balaji, Philip Kershaw, Ana Privette , Ag Stephens, Naomi Naik, Serguei Nikonov, Hans Vahlenkamp, Mackenzie Blanusa, Anderson Banihirwe, Chris Blanton, Nkeh Perry Boh, Ben Evans, Richard Smith, Rhys Evans, Zac Flamig, Diana Gergel, Thomas Jackson, Rebecca Monge, Natalie O'Leary, Zouberou Sayibou.

ESGF Earth System Grid Federation

PANGEO

CMIP The World Climate Research Programme's Coupled Model Intercomparison Project

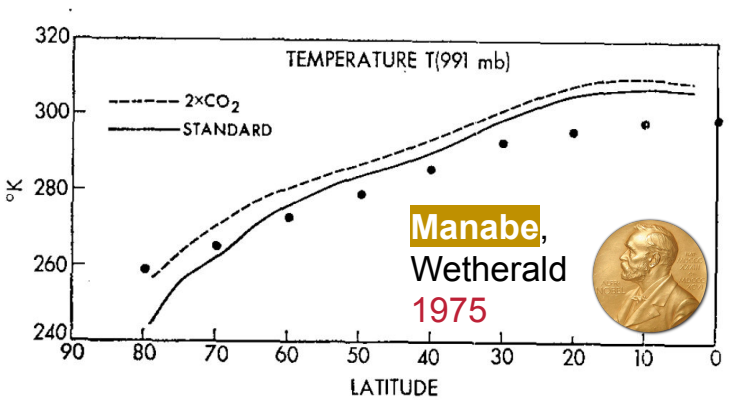
ipcc INTERGOVERNMENTAL PANEL ON climate change

WMO UNEP

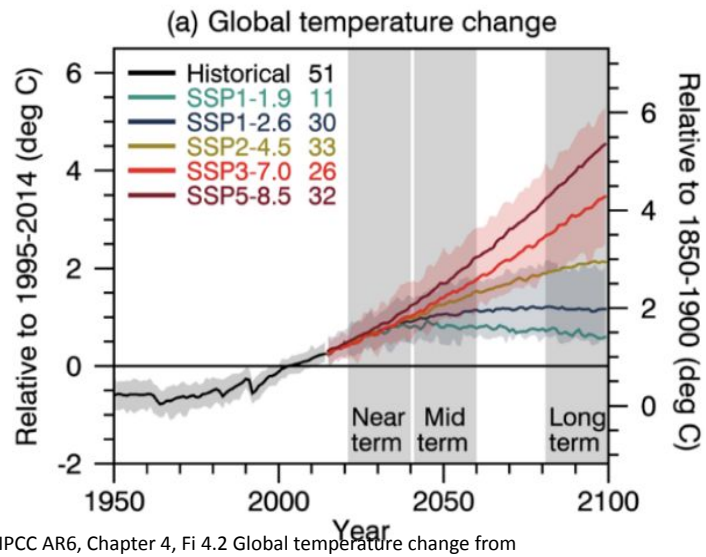
The Climate Pledge

Amazon is committed to building a sustainable business for our customers and the planet. In 2019, Amazon co-founded The Climate Pledge—a commitment to be net-zero carbon across our business by 2040, 10 years ahead of the Paris Agreement.

Motivation: Nature



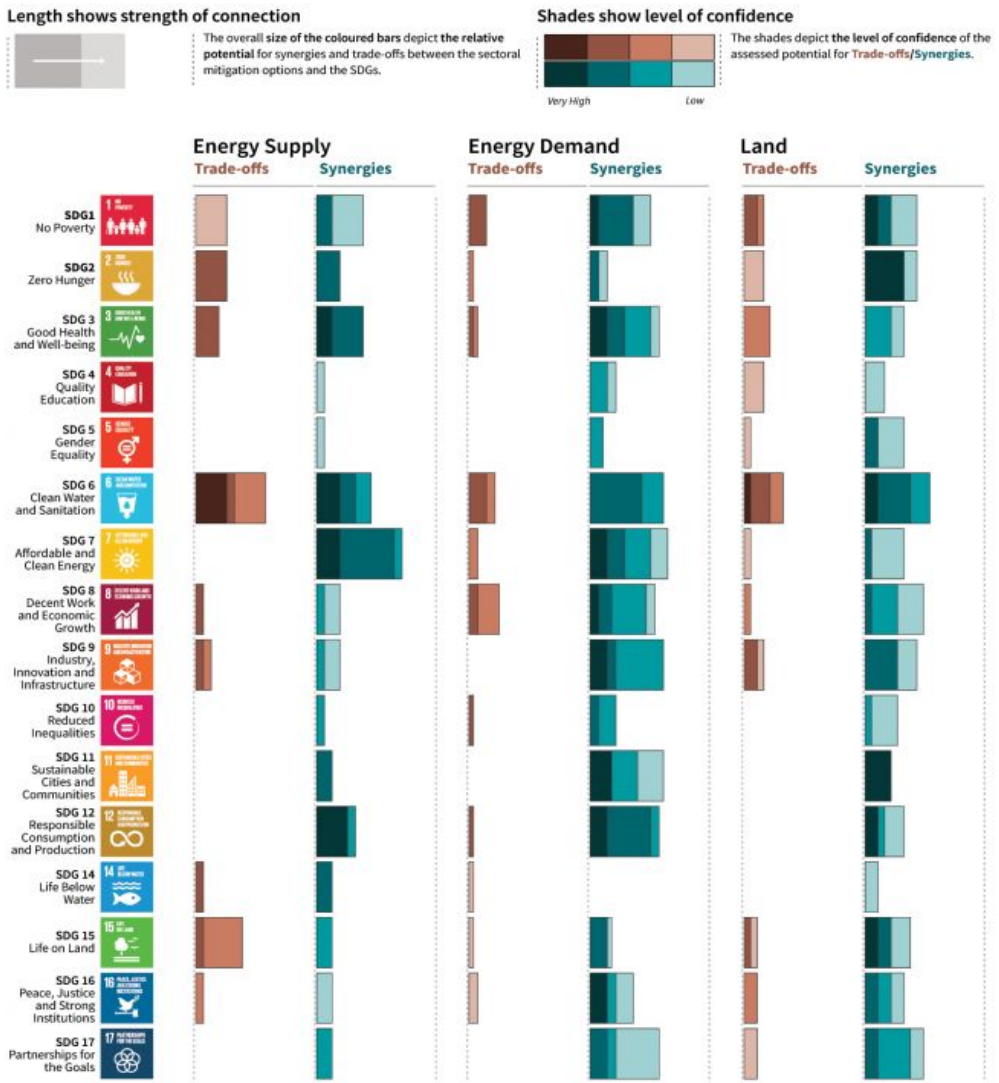
3 (NOAA GFDL)+ 2 (NASA GISS) -- In Charney's report, 1979



IPCC AR6, Chapter 4, Fi 4.2 Global temperature change from CMIP6 historical and scenario simulations.

Photo: Saikiran Kesari/Unsplash

https://science.thewire.in/environment/what-the-ipcc-report-forecasts-for-india-development-future/

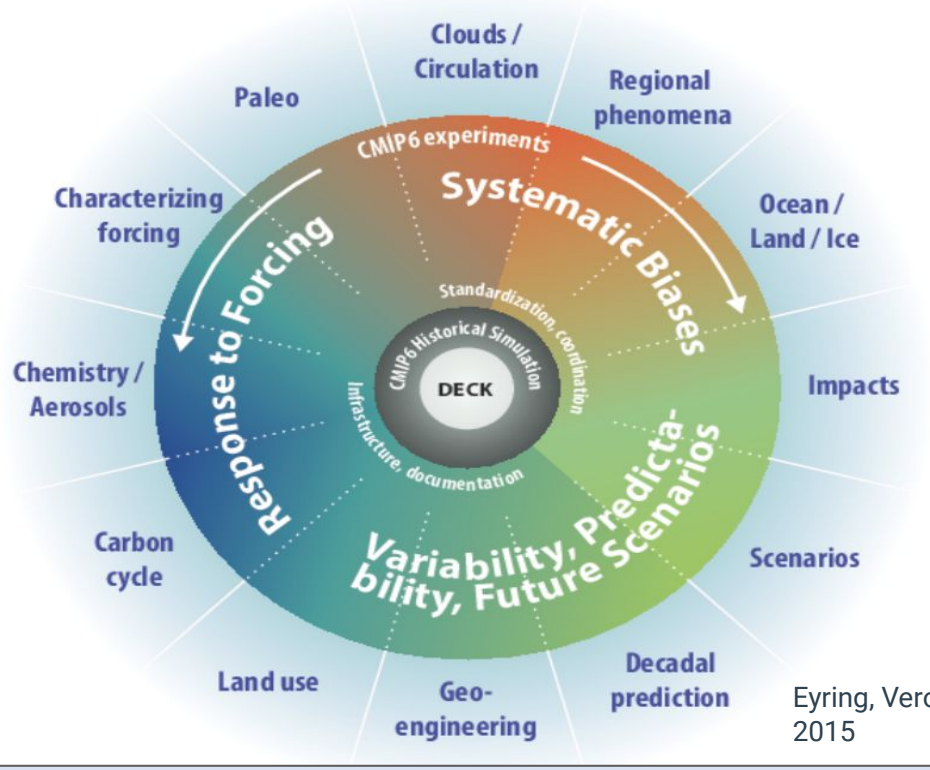


SUSTAINABLE DEVELOPMENT GOALS

IPCC, 2021: Summary for Policymakers. In: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [MassonDelmotte, V., P. Zhai, et al.

Overview: ESGF and CMIP

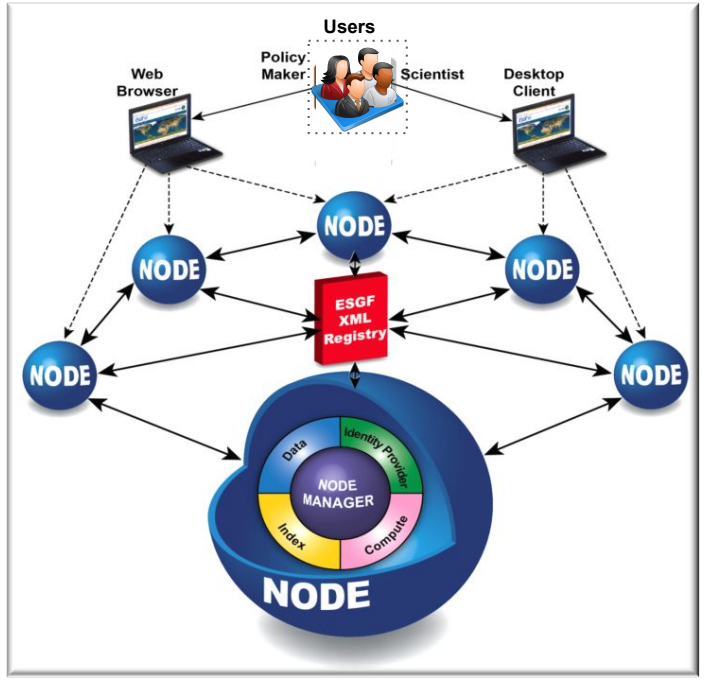
CMIP
The World Climate Research Programme's Coupled Model Intercomparison Project



Eyring, Veronika et al. 2015

The objective of CMIP is to better understand past, present, and future climate change arising from natural, unforced variability or in response to changes in radiative forcings in a multi-model context.

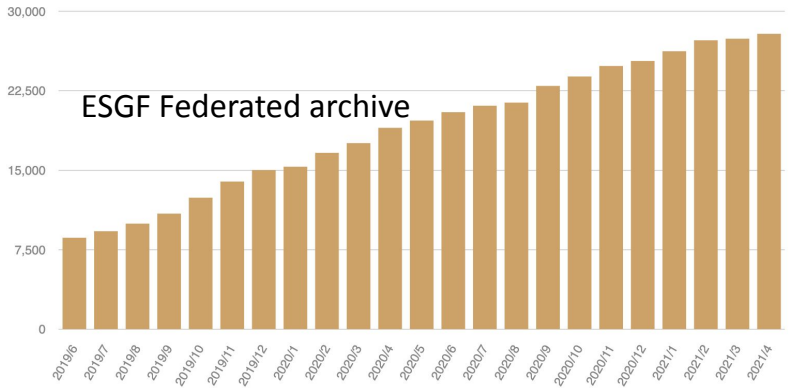
ESGF
Earth System Grid Federation



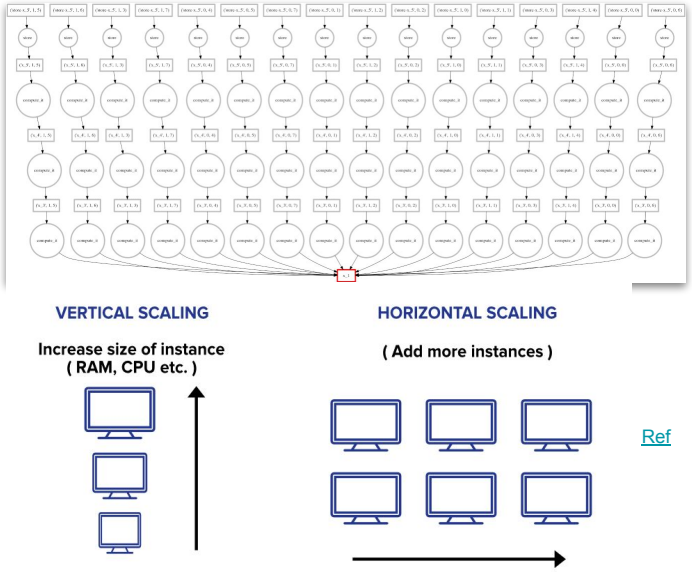
Adapted from https://esgf.github.io/

The Earth System Grid Federation (ESGF) is an international collaboration for the software that powers most global climate change research, notably assessments by the Intergovernmental Panel on Climate Change (IPCC).

Challenges and path forward

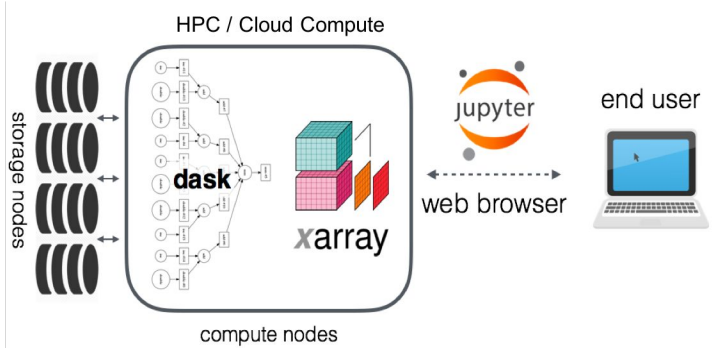


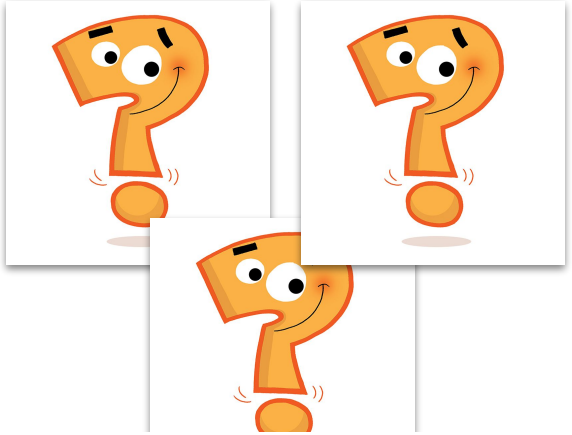
Expanding data archive..



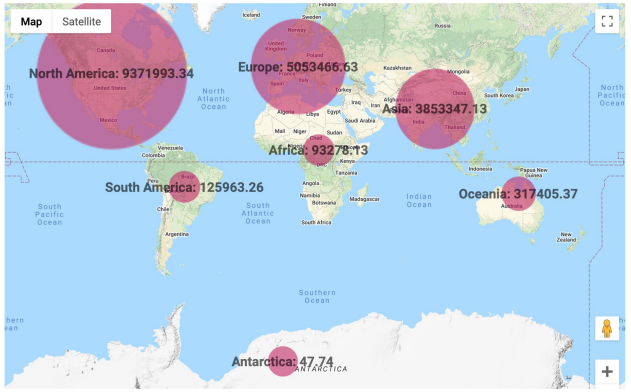
Ref

Different experimental design, model output formats, variable names, units, resolution,….





Numerous science questions..



Diverse groups of CMIP6 users all over the world. Cr. CMCC

Accelerate research

Democratize access to climate data

Community-driven research efforts: CMIP6 data in the cloud

CMIP6 S3 bucket

netCDF

arn:aws:s3:::esgf-world

AWS CLI Access aws s3 ls s3://esgf-world/ --no-sign-request

https://esgf-world.s3.amazonaws.com/index.html

Intake-esm catalog
https://cmip6-nc.s3.amazonaws.com/esgf-world.csv.gz
THREDDS catalog
https://aws-cloudnode.esgf.io/thredds/catalog/catalog.html
SpatioTemporal Asset Catalogs (STAC) underway

Zarr

arn:aws:s3:::cmip6-pds

AWS CLI Access aws s3 ls s3://cmip6-pds --no-sign-request

https://cmip6-pds.s3.amazonaws.com/index.html#CMIP6/

Intake-esm catalog
https://cmip6-pds.s3.amazonaws.com/pangeo-cmip6.csv
STAC catalogs underway

Checkout the CMIP6 registry in AWS to read more information, including CMIP6 data citations.

Community-driven best practices for improved data exploration

E.g. Directory Reference Syntax (DRS) established by the ESGF community makes cataloguing possible.

Taylor et al,2017, CMIP6-CV

E.g.
s3://esgf-world/CMIP6/AerChemMIP/NOAA-GFDL/GFDL-ESM4/hist-piNTCF/r1i1p1f1/Amon/tas/gr1/v20180701/tas_Amon_GFDL-ESM4_hist-piNTCF_r1i1p1f1_gr1_185001-194912.nc

Intake-esm → xarray

Search for all atmos monthly surface temperature fields for historical simulations.

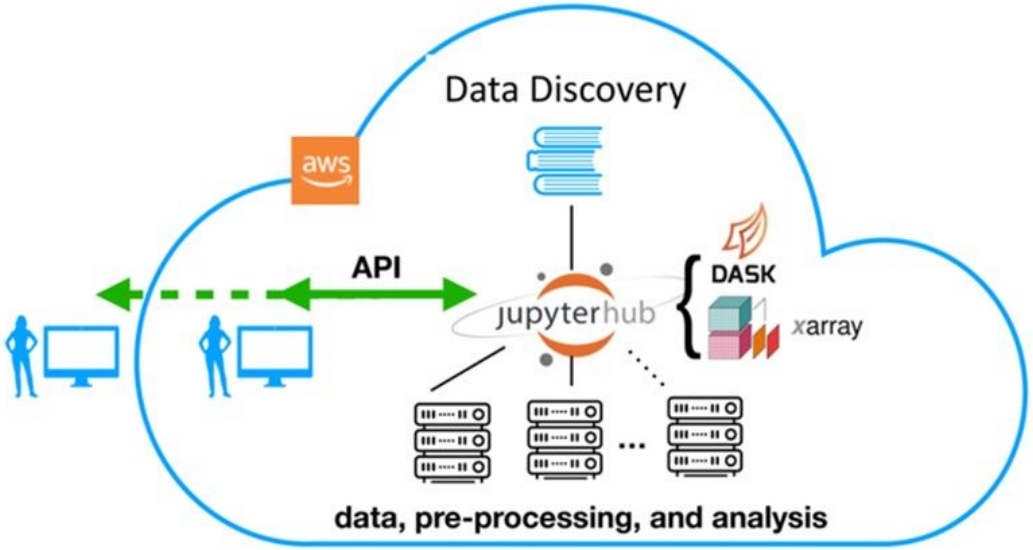
```
exp_filter = ['historical']
table_id_filter = 'Amon'
variable_id_filter = "tas"
cat = col.search(experiment_id=exp_filter,
                 table_id=table_id_filter,
                 variable_id=variable_id_filter)
```

catalog with 55 dataset(s) from 1872 asset(s):

intake-esm https://intake-esm.readthedocs.io/en/latest/
CMIP6 Controlled Vocabulary: https://github.com/WCRP-CMIP/CMIP6_CVs
intake-esm https://intake-esm.readthedocs.io/en/latest/
Example notebooks: https://github.com/pangeo-data/pangeo-example-notebooks
https://github.com/aradhakrishnanGFDL/gfdl-aws-analysis/blob/master/examples/intake-esm-s3-nc-simple-access.ipynb
Pangeo documentation: https://pangeo-data.github.io/pangeo-cmip6-cloud/

Community-driven research efforts: Bring analysis to data

ESGF federated nodes across the world



https://earthdata.nasa.gov/

Towards a more accessible, discoverable and performant CMIP6 data holding in the cloud.

Earth System Grid Federation (ESGF) in the cloud. -By Kristopher Rand

Towards a generic, flexible API to create analysis-ready cloud-optimized dataset. - By Charles Stern

Speed up your multi-model analysis, using CMIP6 pre-processor and Dask - By Julius Busecke

References

Towards a generic, flexible API to create analysis-ready cloud-optimized dataset. - By Charles Stern

Slides:

https://github.com/cisaacstern/pangeo-forge-slides

Pangeo Forge documentation:
https://pangeo-forge.readthedocs.io/

Speed up your multi-model analysis, using CMIP6 pre-processor and Dask - By Julius Busecke

Slides:

https://speakerdeck.com/jbusecke/aws-webinar-cmip6-preprocessing

cmip6_preprocessing:
https://github.com/jbusecke/cmip6_preprocessing

https://earthdata.nasa.gov/

Earth System Grid Federation (ESGF) in the cloud

ESGF federated nodes across the world

NEXT: Towards a more accessible, discoverable and performant CMIP6 data holding in the cloud.

Earth System Grid Federation (ESGF) in the cloud. -By Kristopher Rand

CMIP6 Data in the Cloud: 3 Main Objectives

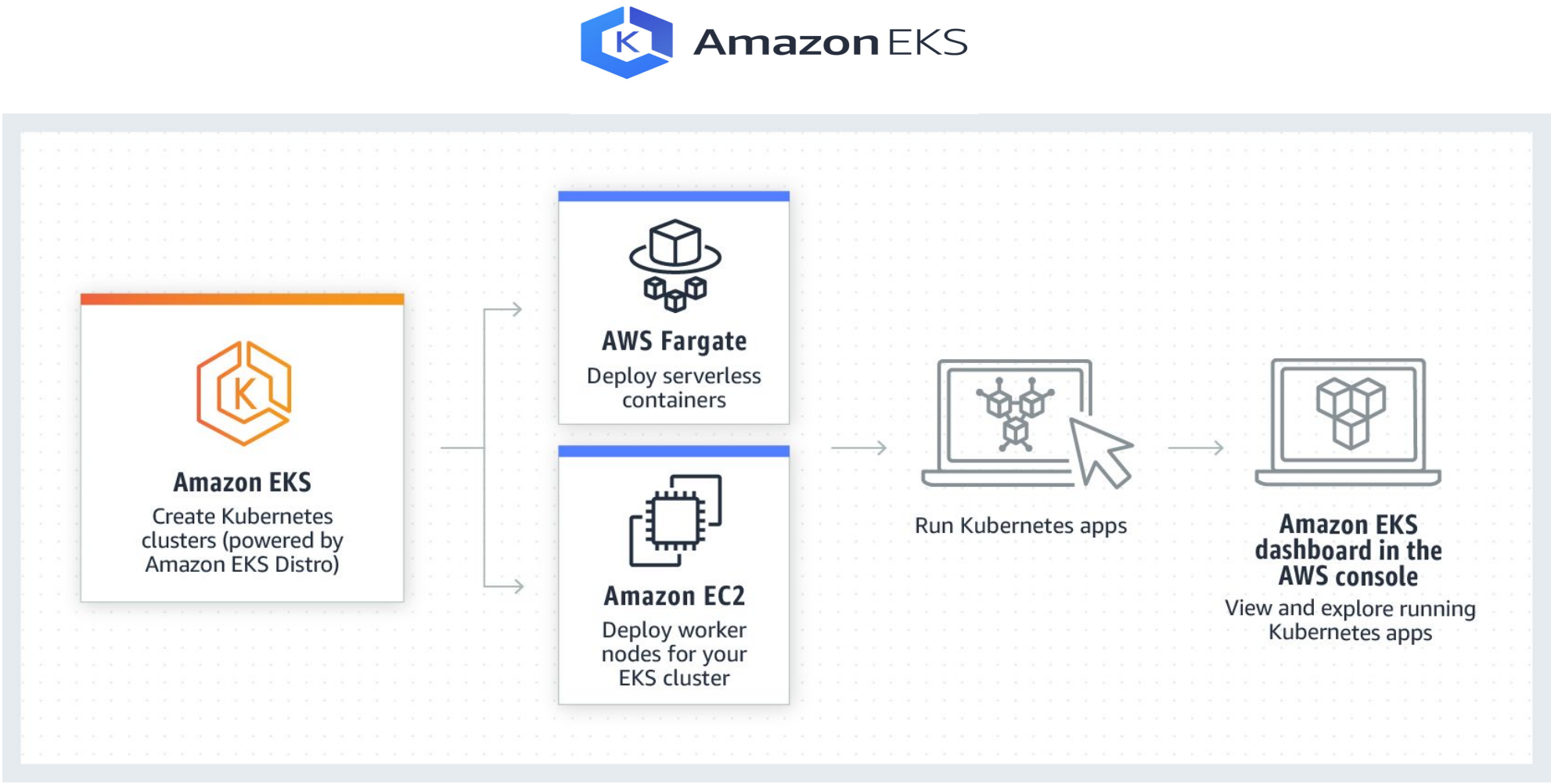
Accessibility

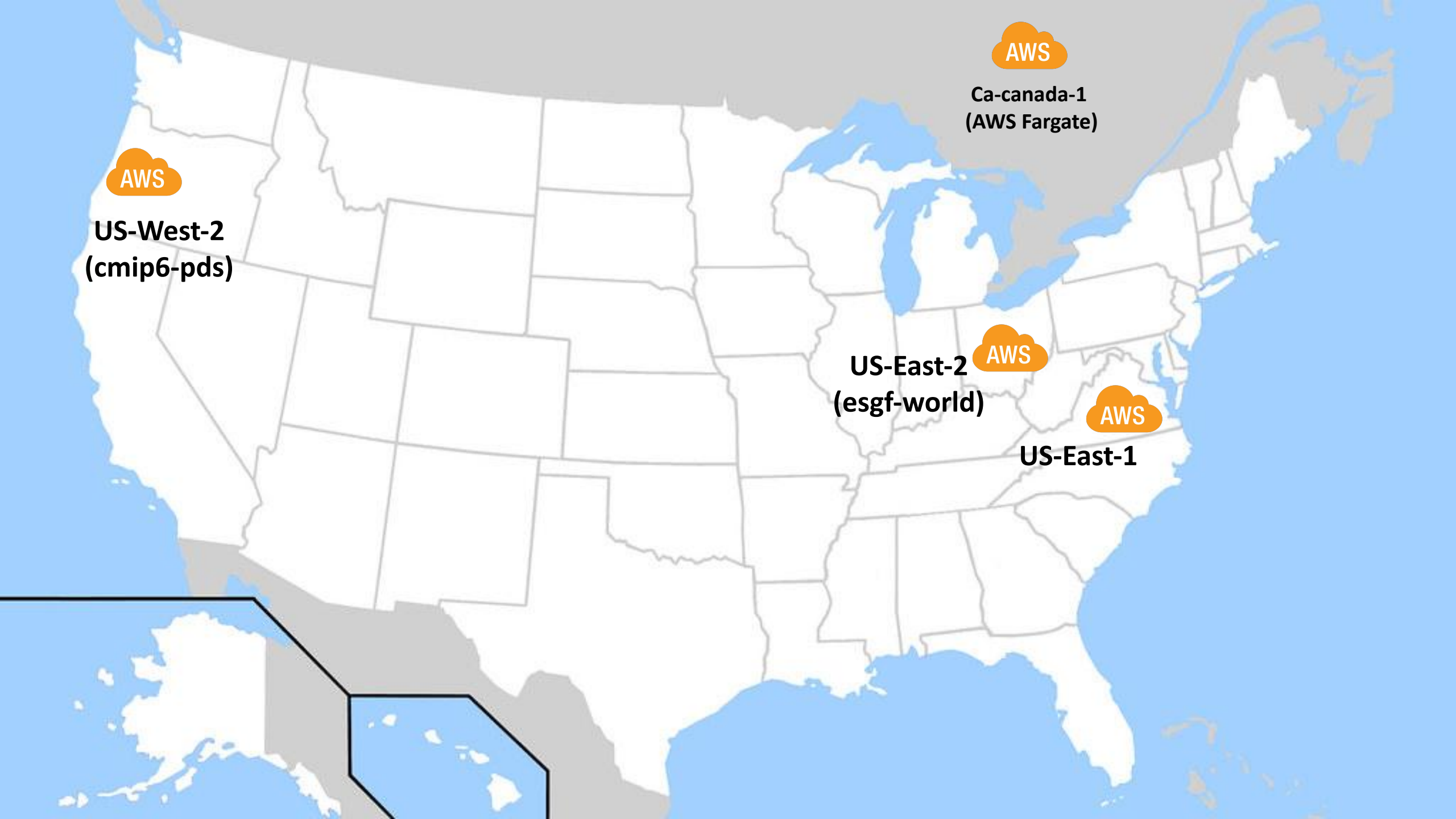
Discoverability

Performance

Others: 1. Scalable framework, 2. Data Management, 3. Data Metrics Analytics

Building a CMIP6 data node in the cloud utilizing Amazon's Elastic Kubernetes Service (EKS)




AWS
Ca-canada-1
(AWS Fargate)
AWS
US-West-2
(cmip6-pds)
AWS
US-East-2
(esgf-world)
AWS
US-East-1

EKS Clusters by Region

Ohio Region - US-East-2

- ESGF cluster
- 3 EC2 instances
- ESGF Data publication software
- Mounts “esgf-world” S3 bucket (1 PB) as file system

Virginia Region - US-East-1

- JupyterHub Dev/Test cluster
- 3-20 EC2 instances for autoscaling tests

Oregon Region - US-West-2

- ESGF Dev/Test cluster
- 1 EC2 instance
- The region itself also houses “cmip6-pds” Zarr S3 bucket (1 PB)

Canada Region - ca-canada-1

- AWS Fargate (serverless containers)

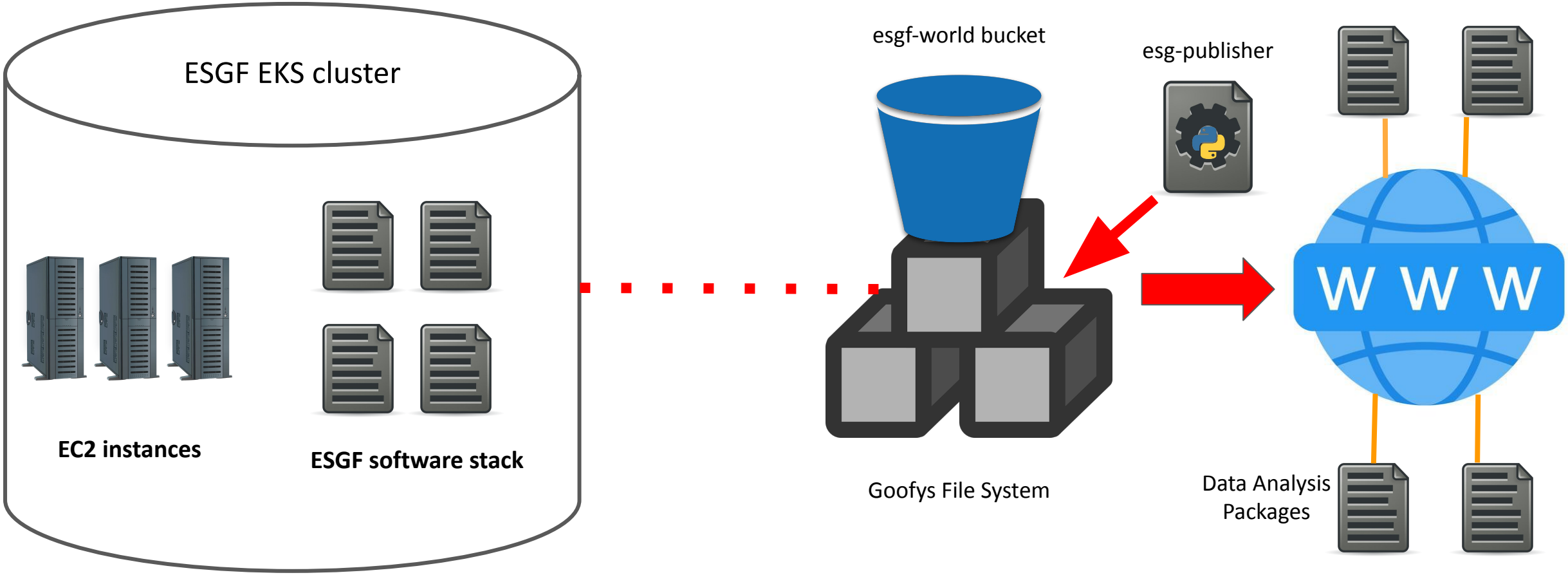
ESGF in the Cloud: Cloud Storage

With our EKS clusters established, we wanted to anchor data holdings to a storage infrastructure that provided sufficient security, organization, and efficiency

- Amazon Simple Storage Service (S3) bucket
  - “esgf-world” bucket (NetCDF data) mounted to EC2 instances on ESGF cluster (1 PB storage)
  - “cmip6-pds” bucket (Zarr data) utilized on EC2 instance for ESGF dev/test cluster (1 PB storage)
- Dynamically optimized for performance and cost effectiveness depending upon user access
- Security policies: S3 bucket policies, object access and bucket access control lists (ACLs), IAM (Identify and Access Management) roles.
- Objects organized compliant to CMIP6 Data Reference Syntax (DRS)
- 2 PB of storage total between the 2 buckets

ESGF Cluster and the “esgf-world” S3 bucket

- EC2 instances/nodes mount S3 bucket using the POSIX-like file system, Goofys
  - **ESGF software stack** for eventual data publication in AWS requires interaction with the Amazon S3 bucket to perform R/W operations; cannot natively communicate with the S3 bucket
  - Goofys is high-performance and easy to set up



ESGF Cluster Architecture

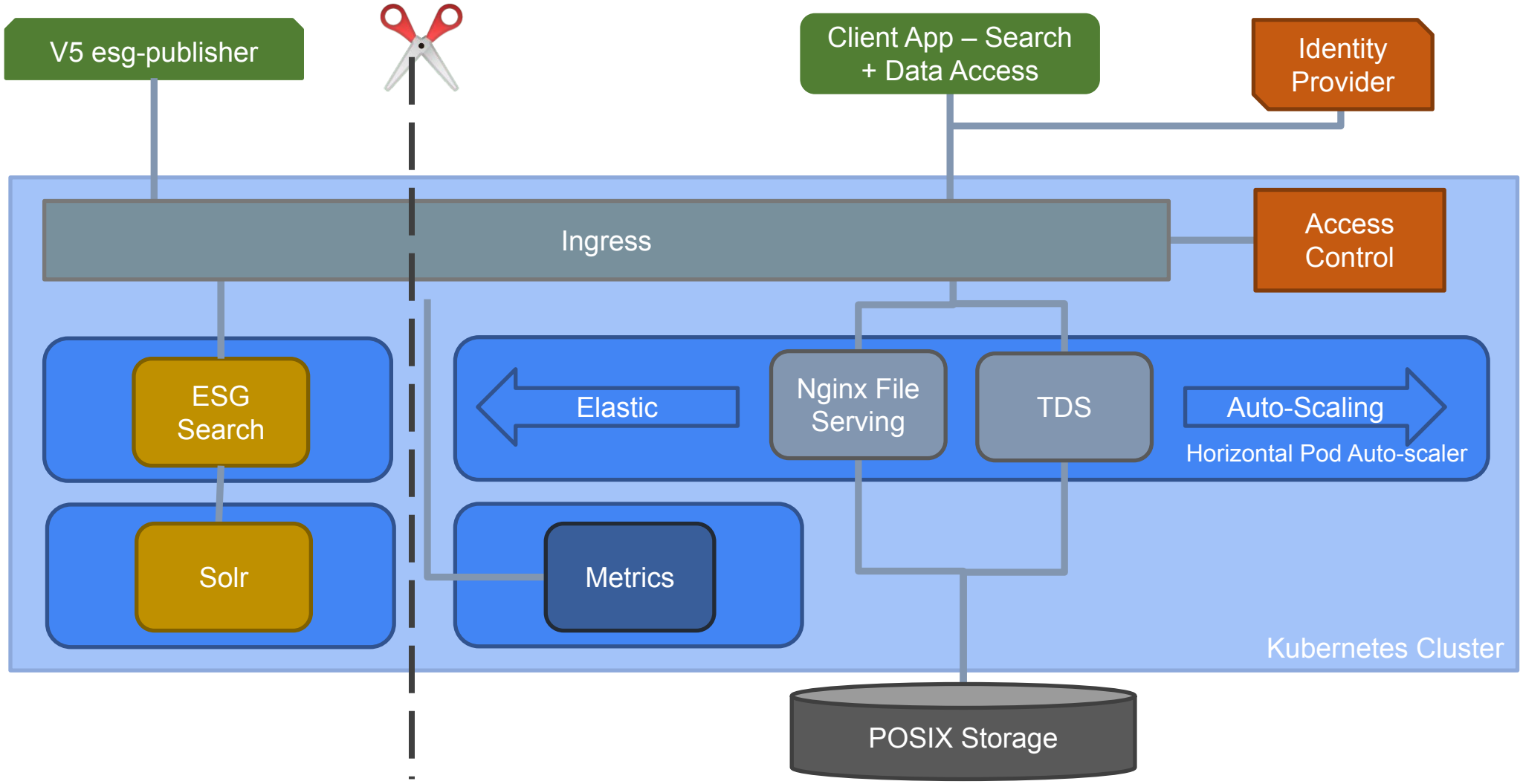


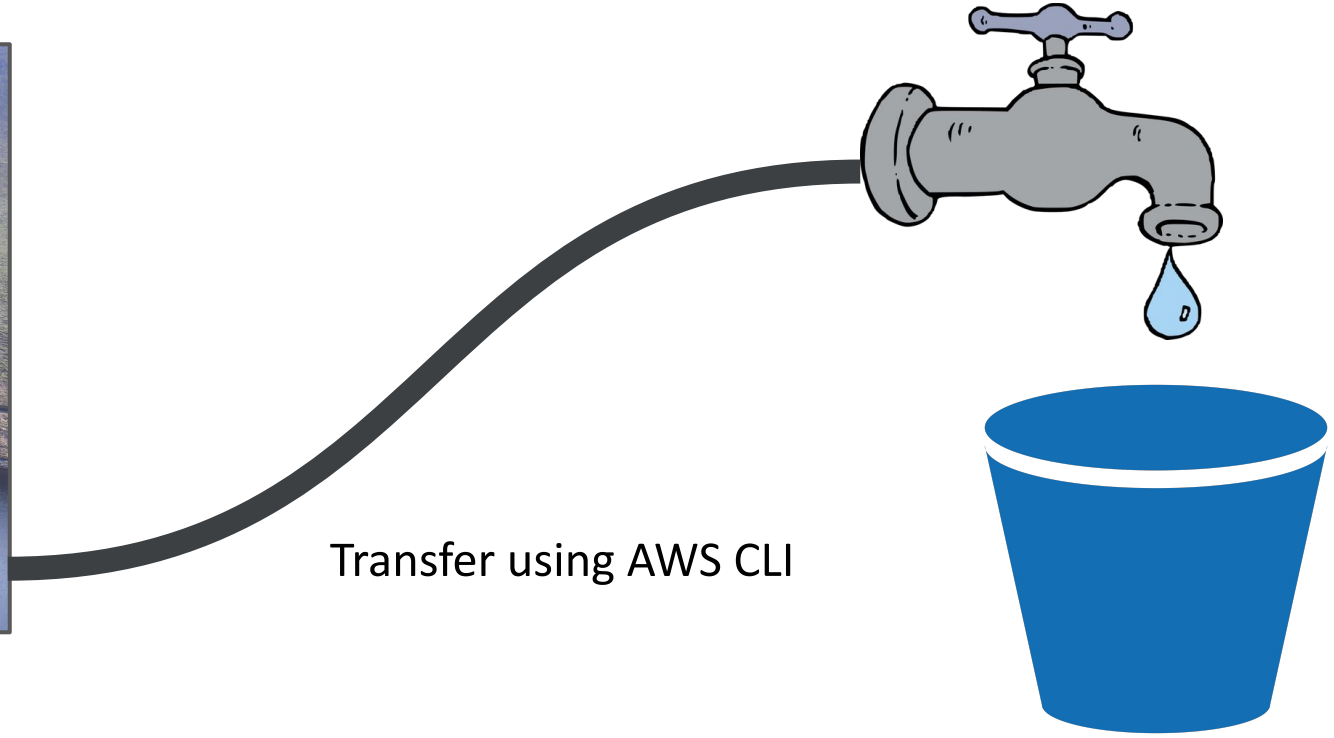
Figure Courtesy of: Phil Kershaw

Filling the bucket

Initial Data Transfer (esgf-world bucket)

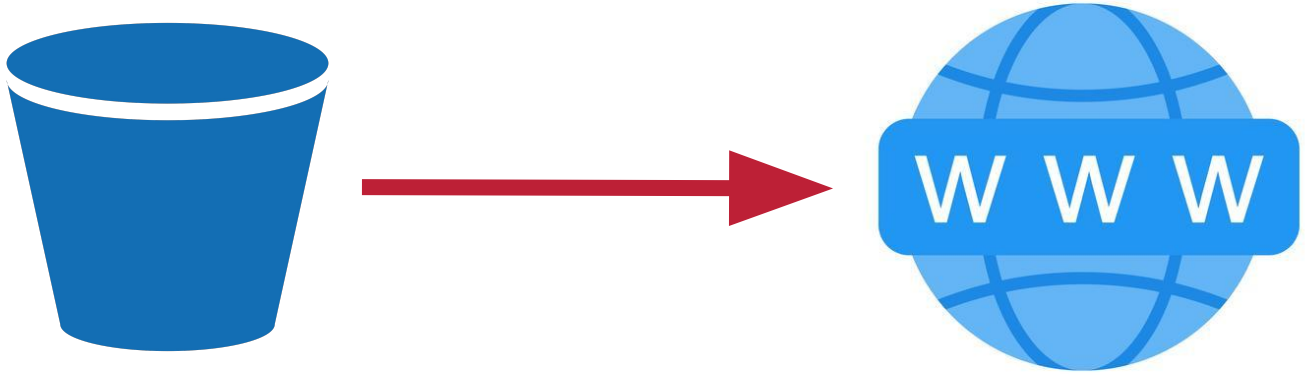
- Ingress: Transfer of ≅ 650 TB of high-value CMIP6 datasets primarily based on IPCC chapter variables and the consensus of the community
- Time period of transfer: 6 months

650 TB CMIP6 data reservoir



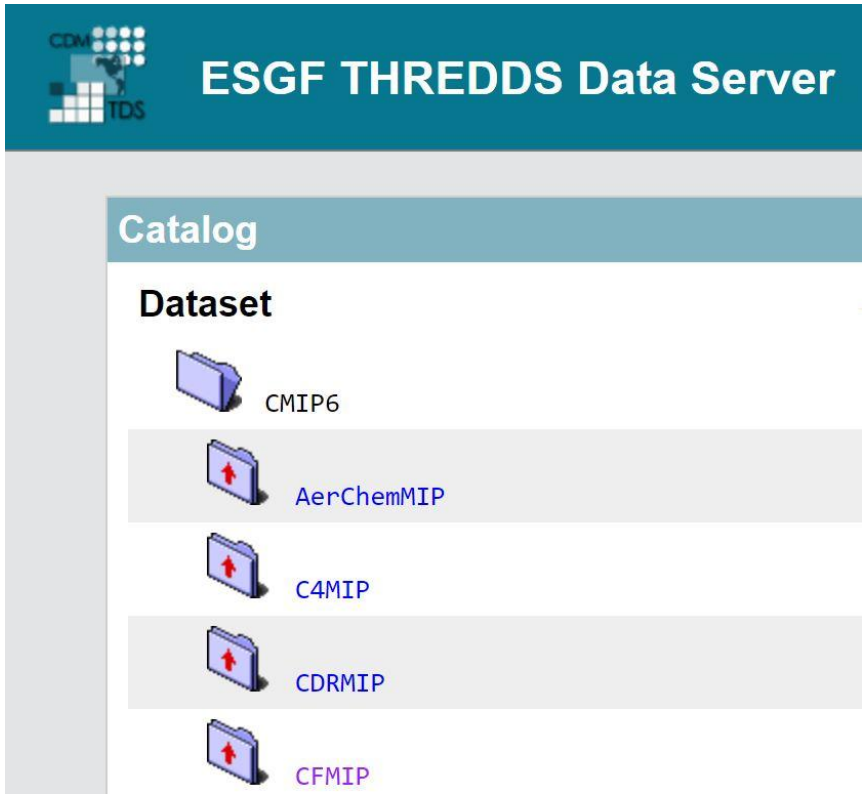
esgf-world bucket

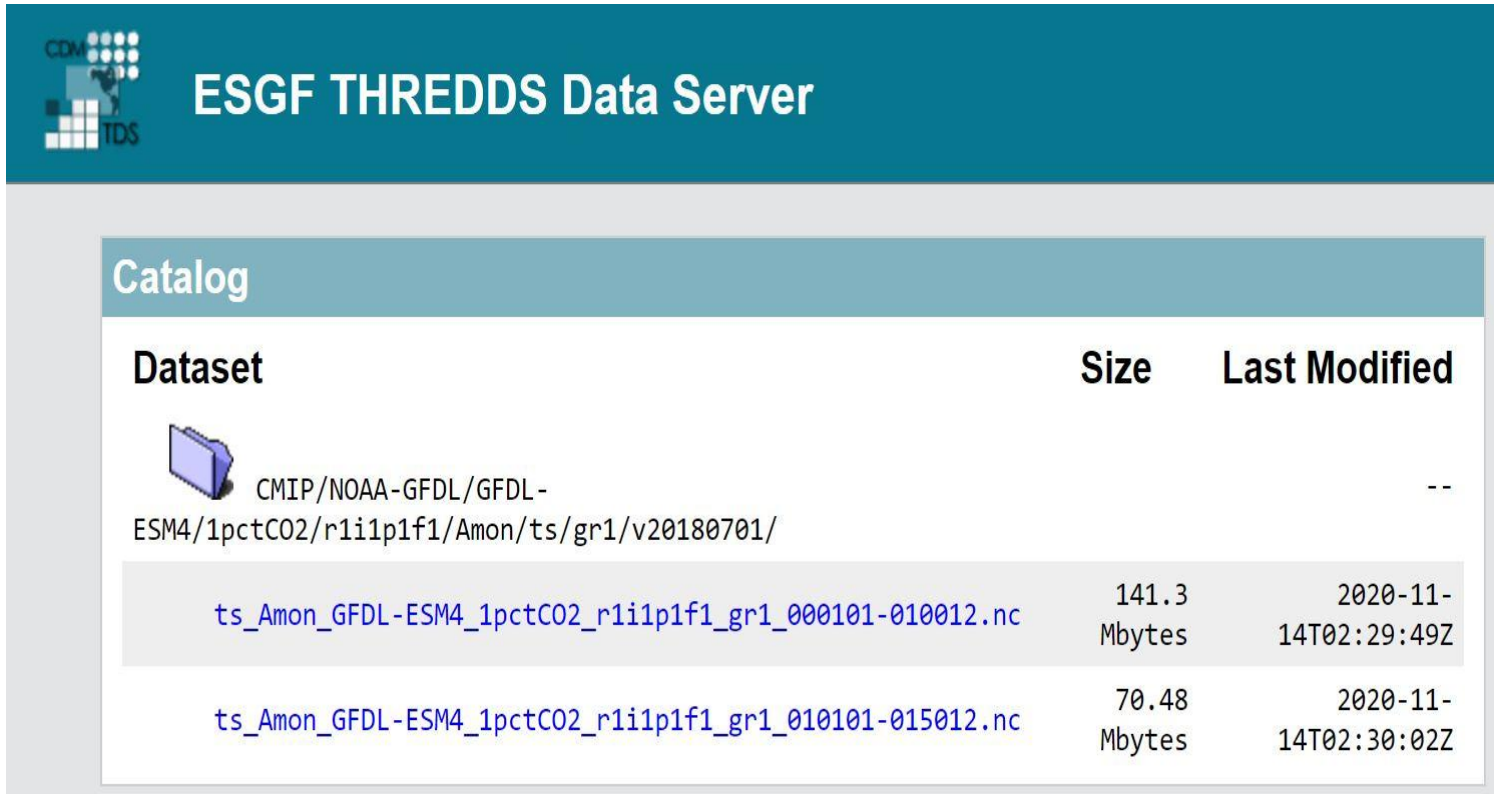
Access from the esgf-world S3 bucket



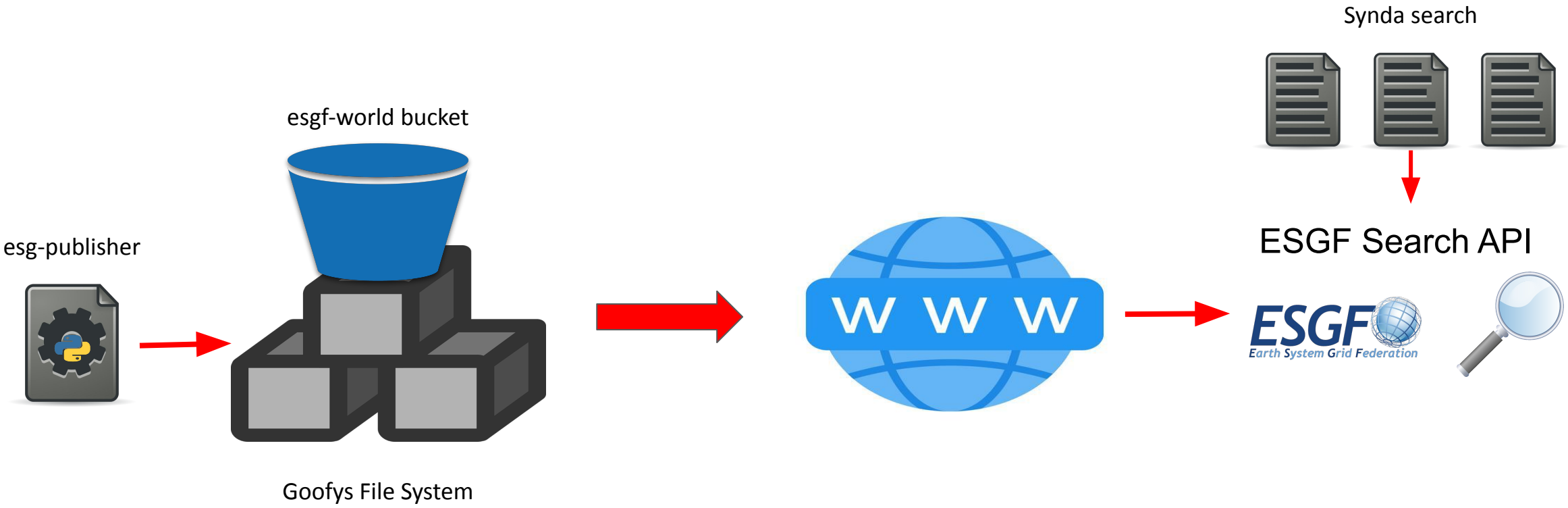
THREDDS Data Server (TDS)

- Web server that provides metadata and data access for scientific datasets, using OPeNDAP, OGC WMS and WCS, HTTP, and other remote data access protocols.
- ESGF IO domain (https://aws-cloudnode.esgf.io/thredds/catalog/catalog.html)





ESGF Data Publisher



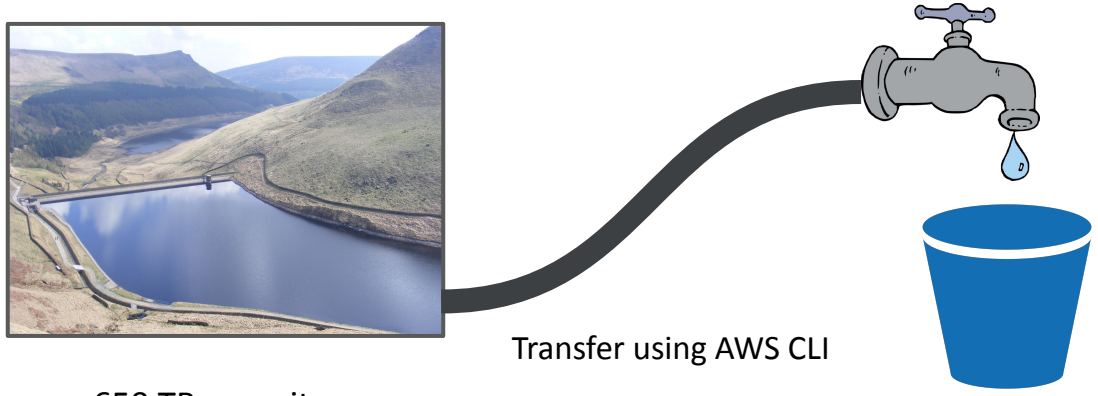
- The ESGF data publisher is run on the datasets and is processed into a record that becomes visible to applications that utilize the ESGF search API (i.e. Synda search)
- Further enhances discovery capability by enabling bulk search parameters using dozens of available dataset metadata fields
- As the publisher is decoupled from the main ESGF software package, there is greater flexibility for data egress

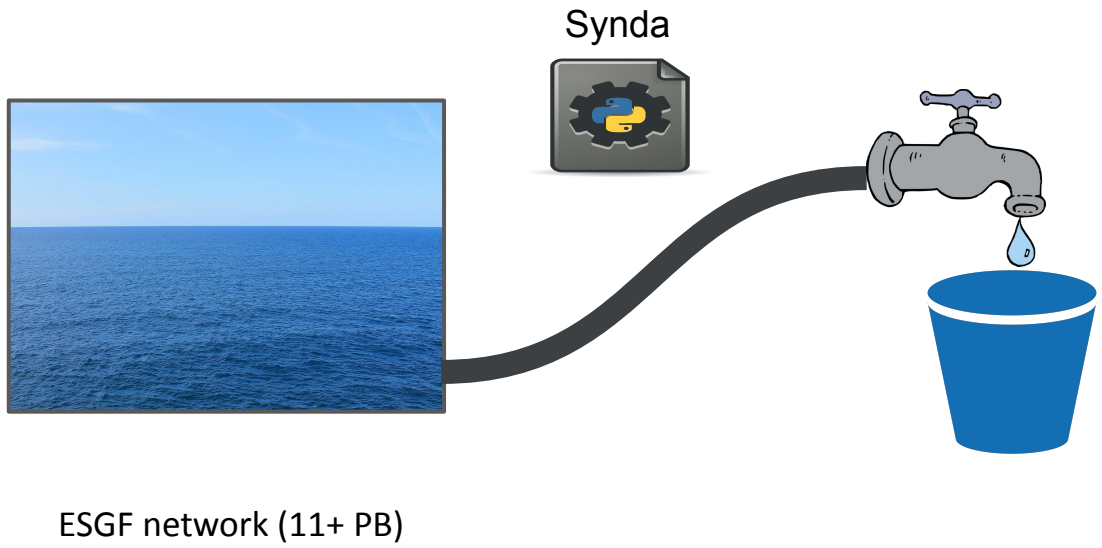
Next Steps: Egress - Federation of ESGF Cloud node

aws-cloudnode.esgf.io data node

esgf-world S3 bucket

Next Steps: Ingress - Filling our Bucket Faster

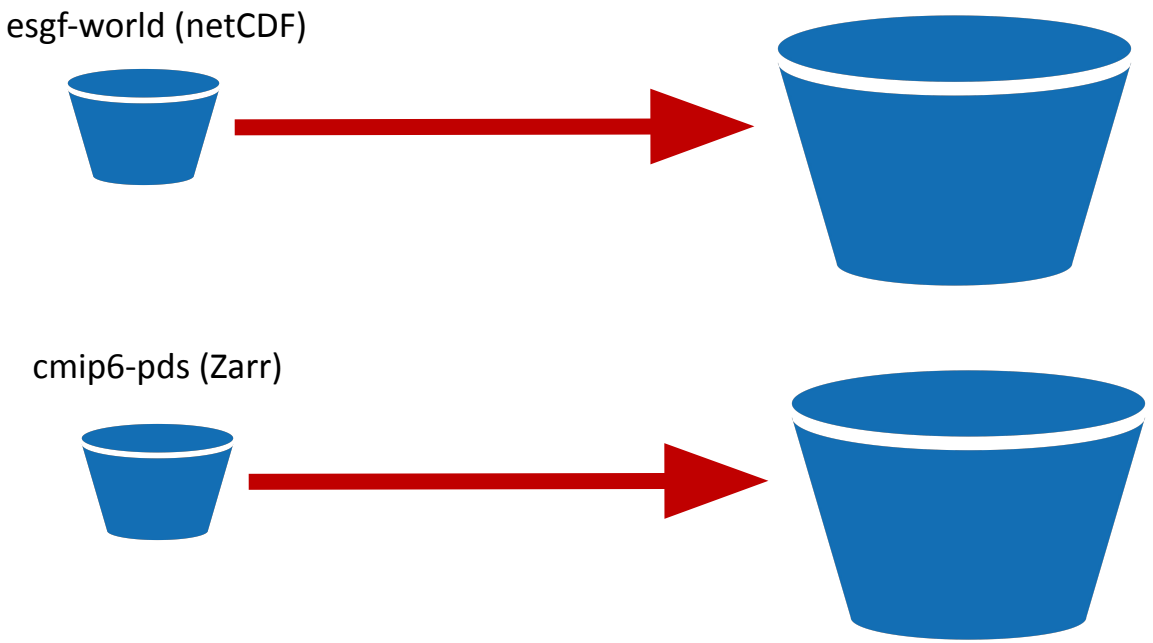




- Extending the ingress repository to the ESGF network itself
  - 11+ PB of unique datasets
  - Plentiful metadata
  - ESGF search API for bulk ingress
- Synda tool for bulk and parallel ingress (documentation can be found HERE)

We’re gonna need a bigger bucket!

Next Steps: Ingress - Expanding our Bucket



- Proposal submitted and pending for adding another 3 PB for esgf-world S3 bucket and Zarr S3 bucket for a total of 5 PB

- Comparison tracker of the two buckets can be found HERE

ESGF node: Future Architecture

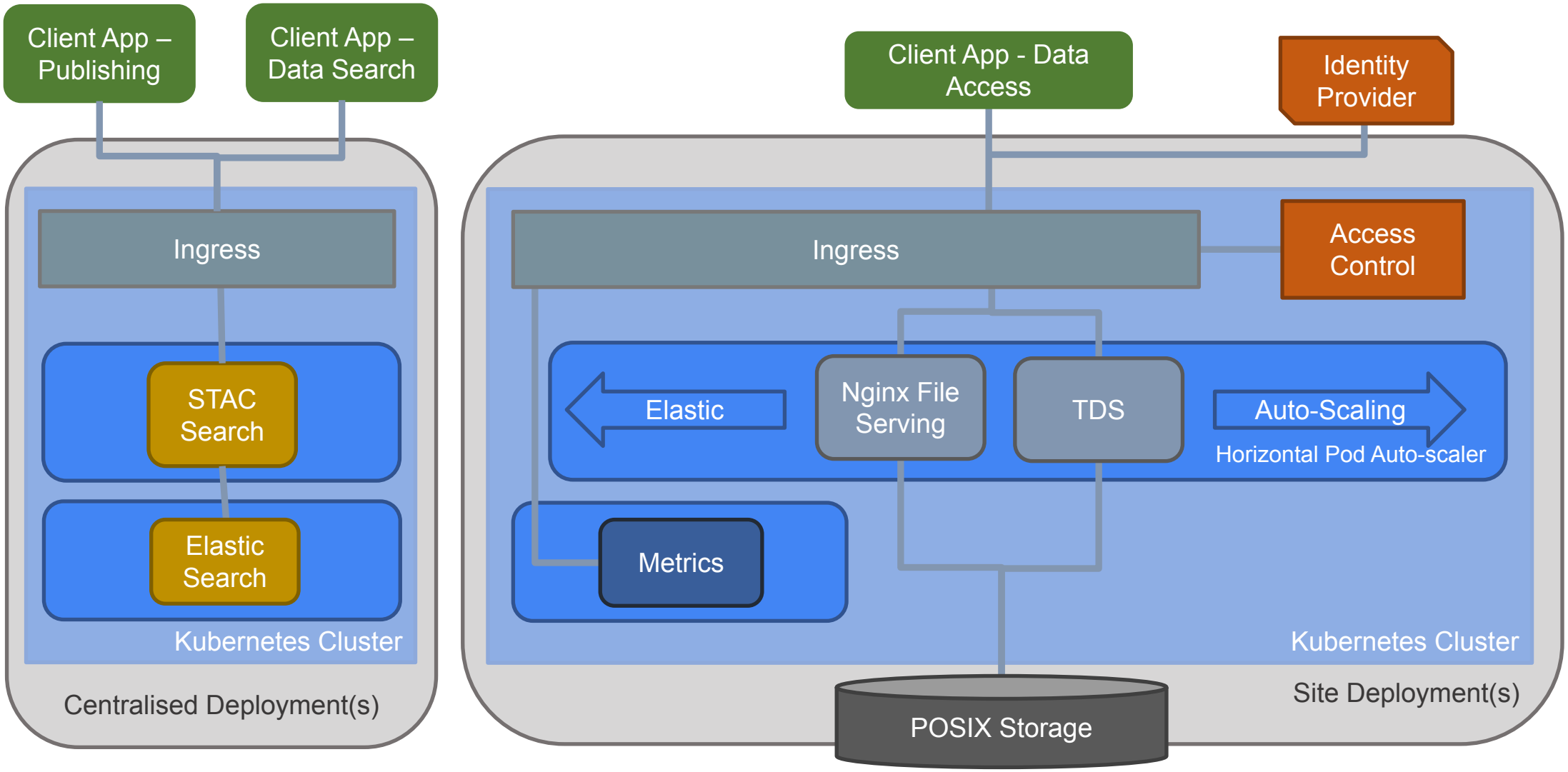


Figure Courtesy of: Phil Kershaw