JOINT GENOME INSTITUTE

A DOE OFFICE OF SCIENCE USER FACILITY

Distributed Computing at the JGI: A Grid-like approach for the life sciences

Georg Rath, Seung-Jin Sul, Ed Kirton,
Jeff Froula, Hugh Salamon

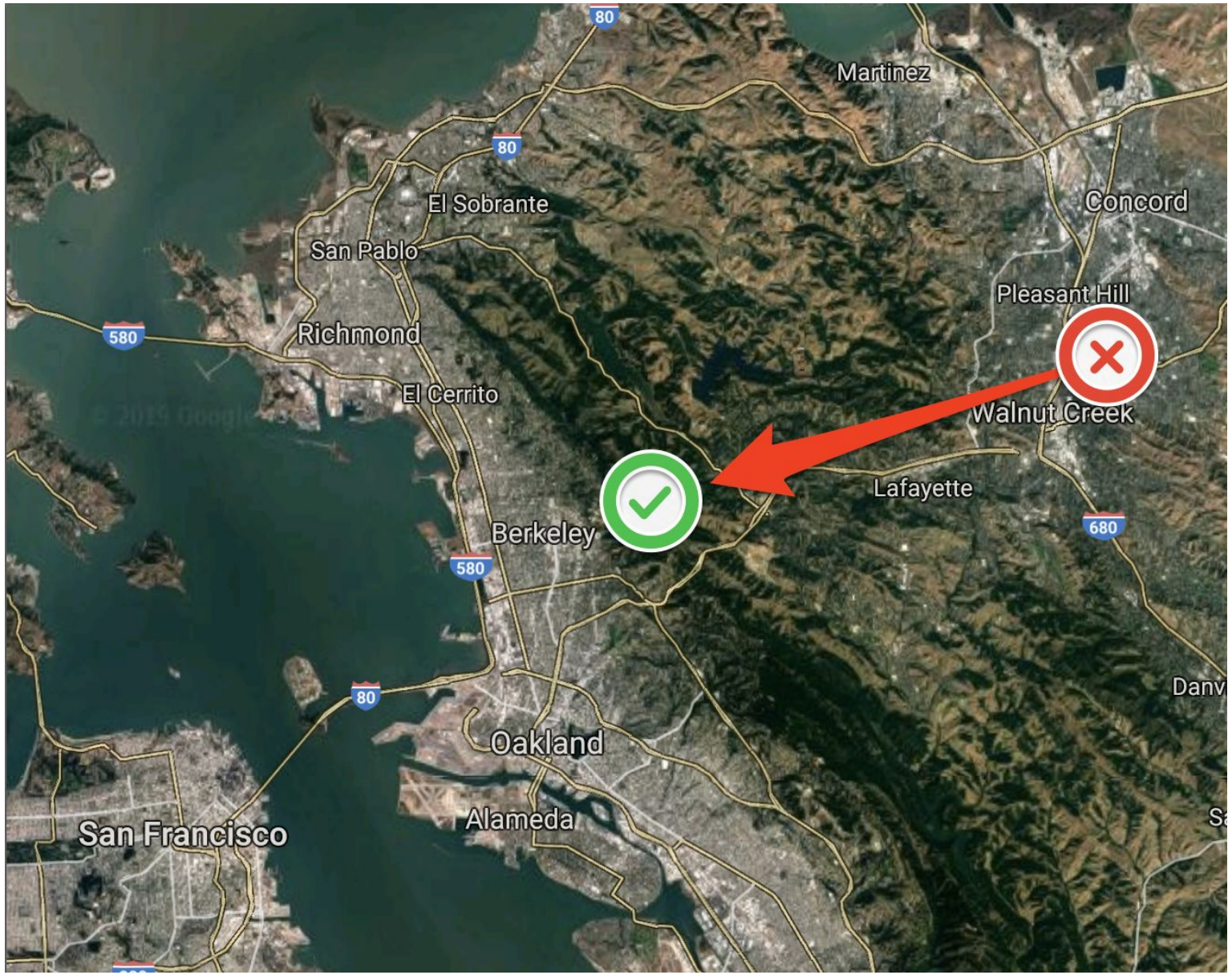
2019-10-18 - HEPIX

The Joint Genome Institute

- Mission

  "To provide the global research community with access to the most advanced integrative genome science capabilities in support of the US Department of Energy (DOE) research mission"
- National User Facility with ~ 1200 users worldwide
- ~280 staff
- ~$70M annual funding
- Services used by 1,598 DOE affiliated researchers in 2017
- DNA sequencing and other advanced genomic technologies
- Production & R&D
- Computational Analysis
- 75 Million core hours in 2018

Move to “The Hill” (LBNL proper)

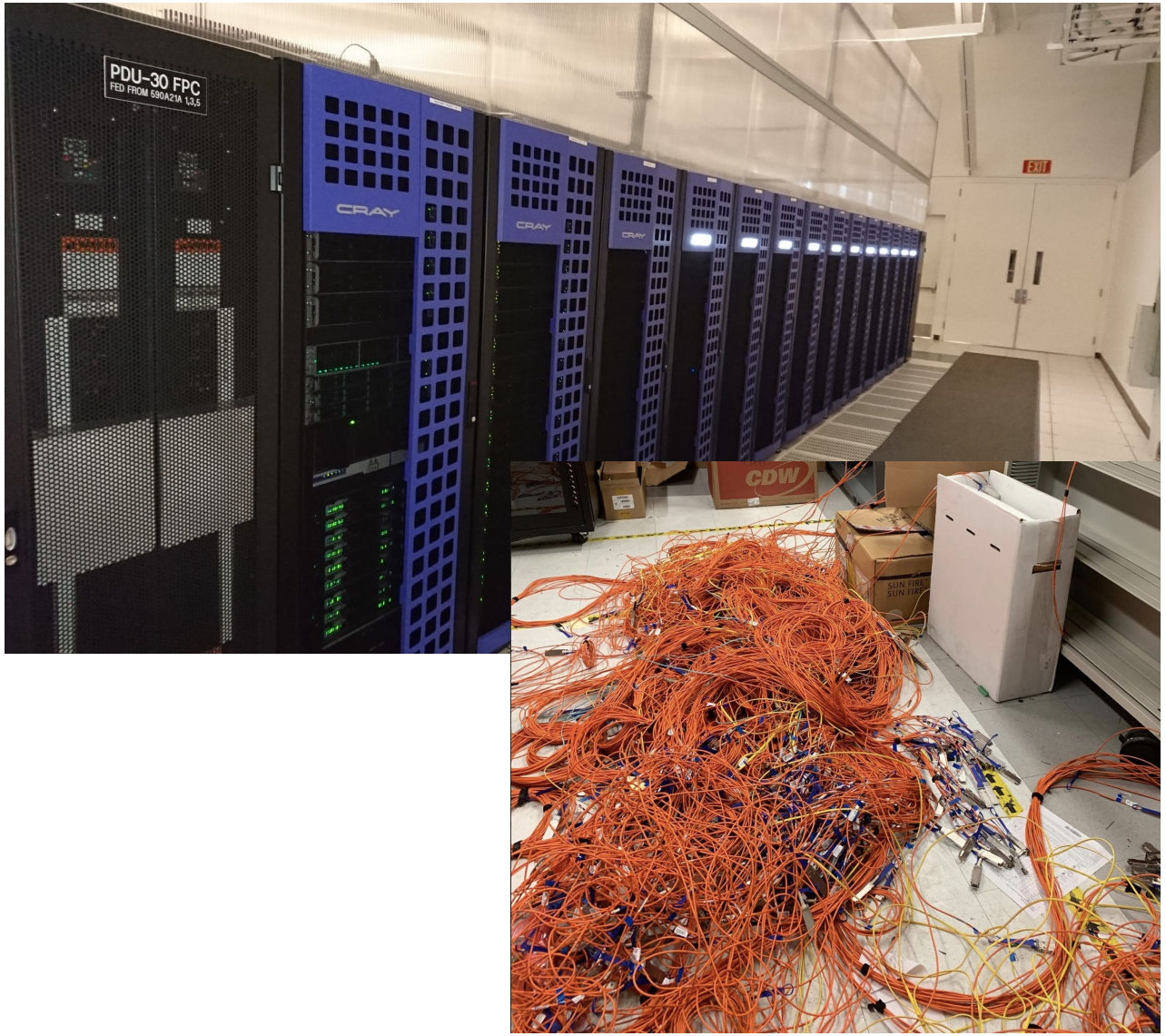


Integrative Genomics Building

Mendel Move

19-FA-5723 Visitor Map

Mendel Move


PDU-30 FPC
FED FROM 690A21A 1,3,5
CRAY
CRAY
CRAY
EXIT
CDW
SUN FIRE
SUN FIRE


CHU ROAD
SMOOT ROAD
91
90A
90E
G1
H1
90
90B
88
70
H2
K1
70A
54
59
G3
Blackberry
Canyon
Entrance
CYCLOTRON ROAD
F

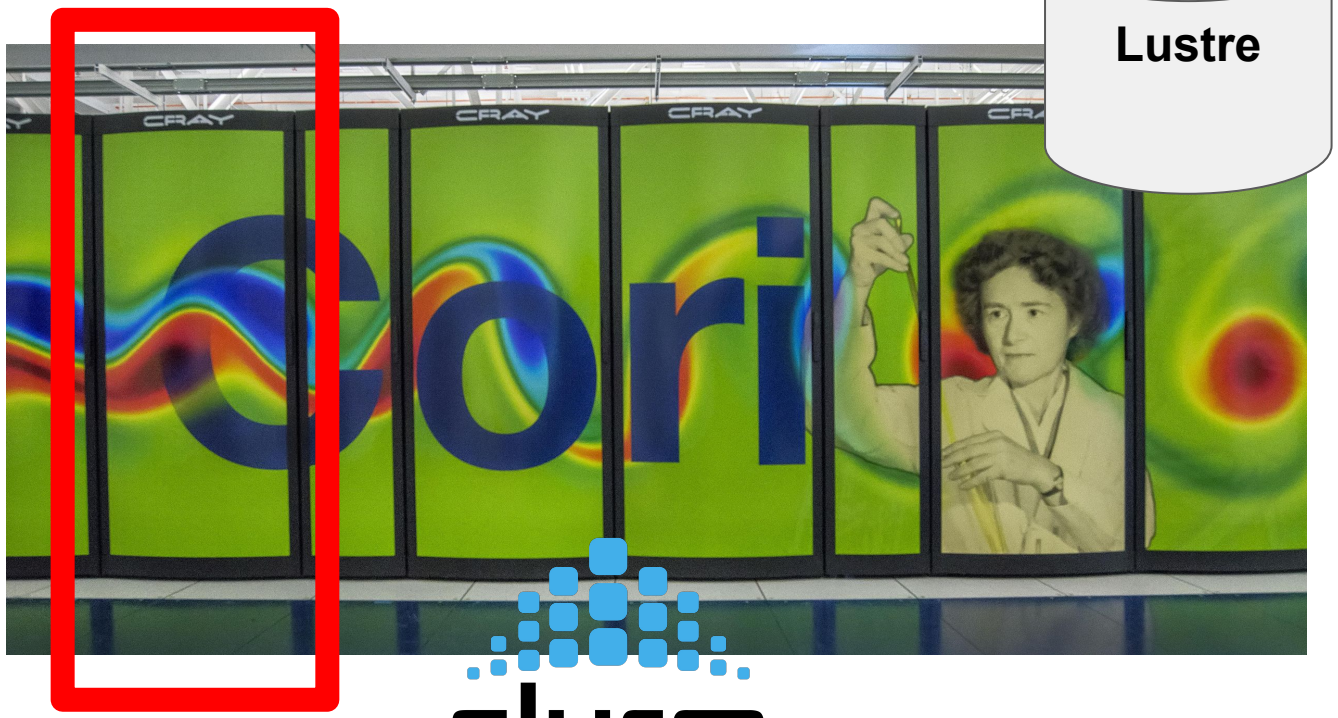
JGI Computational Resources

NERSC

GPFS
(multiple)
~ 5 PB

HPSS

~ 7 PB



Lustre

1 Rack of Cray XC40
(192 Nodes)

slurm
workload manager

Compute Allocations in
Cori proper
(m342 + science programs)

Rancher
Container
Platform

Spin

VM Platform
User Facing Services

vmware

JGI Computational Resources II

VM Platform
Business Systems
vmware®

coming
maybe

NetApp Seq Buffer ~ 1 PB

Coming Soon!

Instrument Support Servers

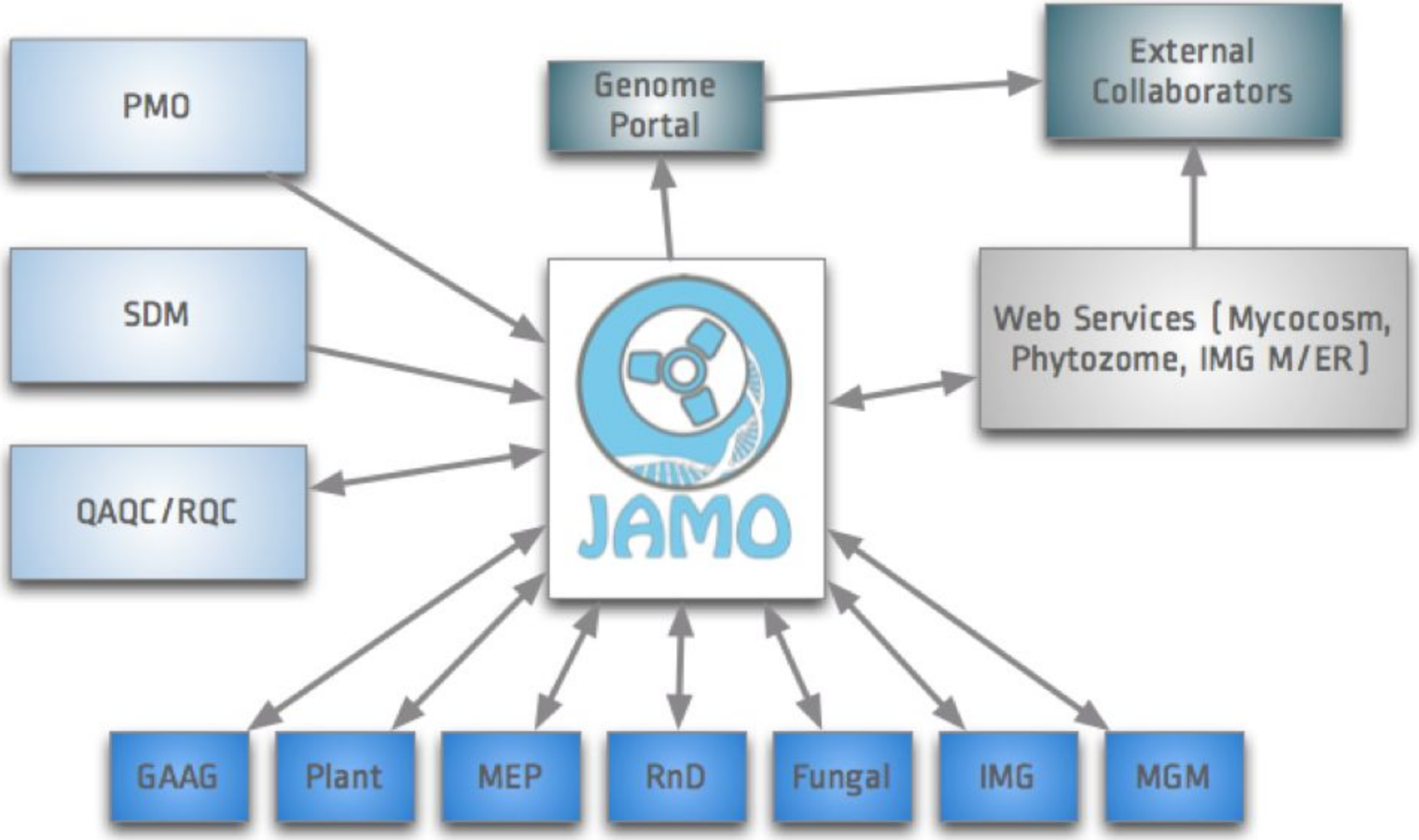
Coming Soon!

“JGI Cloud”

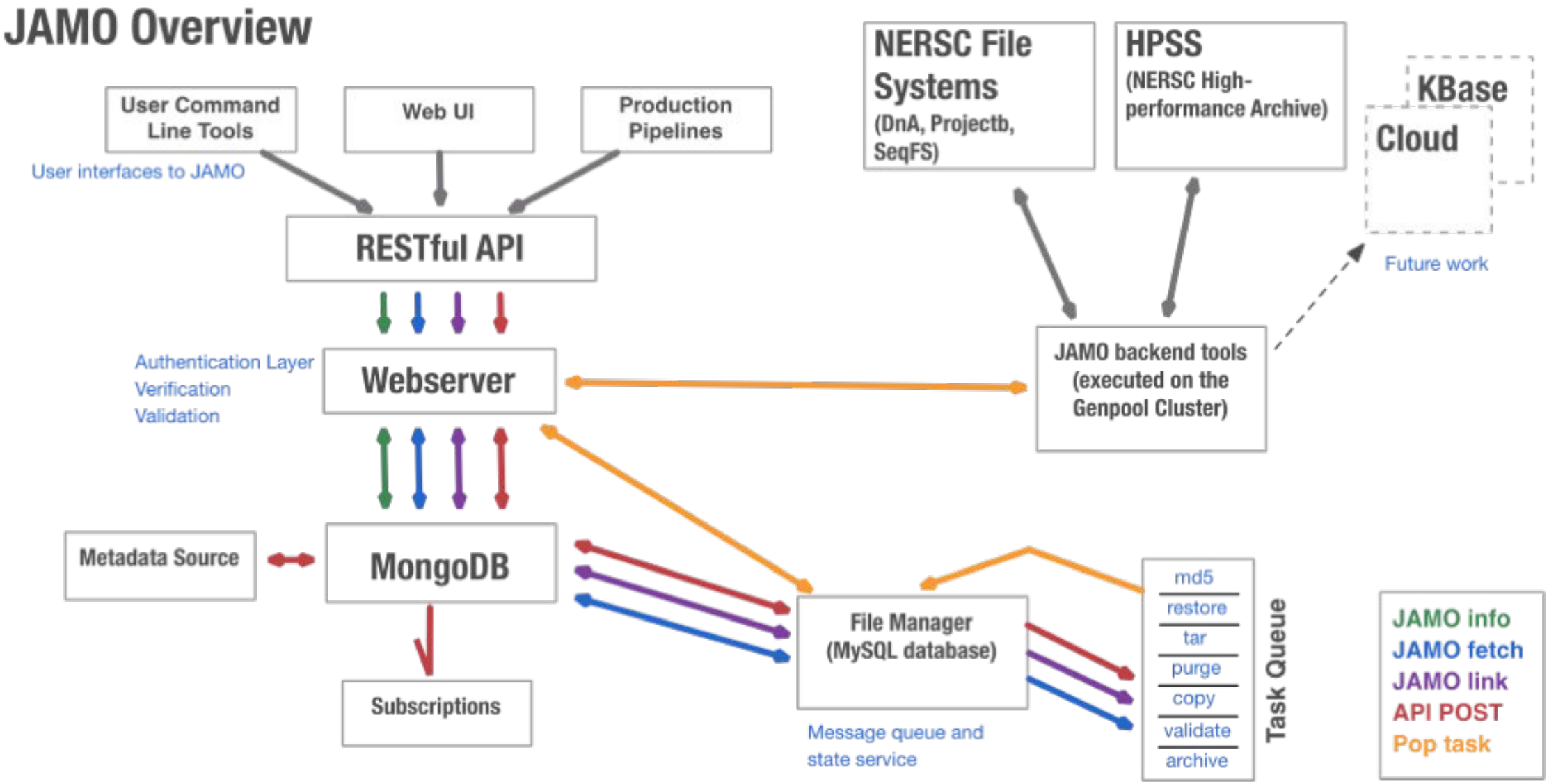
Data Movement - JAMO

- In-House Tool: JAMO
- JGI Archive and Metadata Organizer
- Online since 2006, manages 700 mio files, ~ 7 PB on tape
- Stores metadata of data produced by instruments/analysis
- Ties back to projects - “provenance”
- Moves and caches between HPSS and filesystems (not remote)
- Publish/Subscribe service

JAMO - “it really ties the room together”



JAMO - Architecture



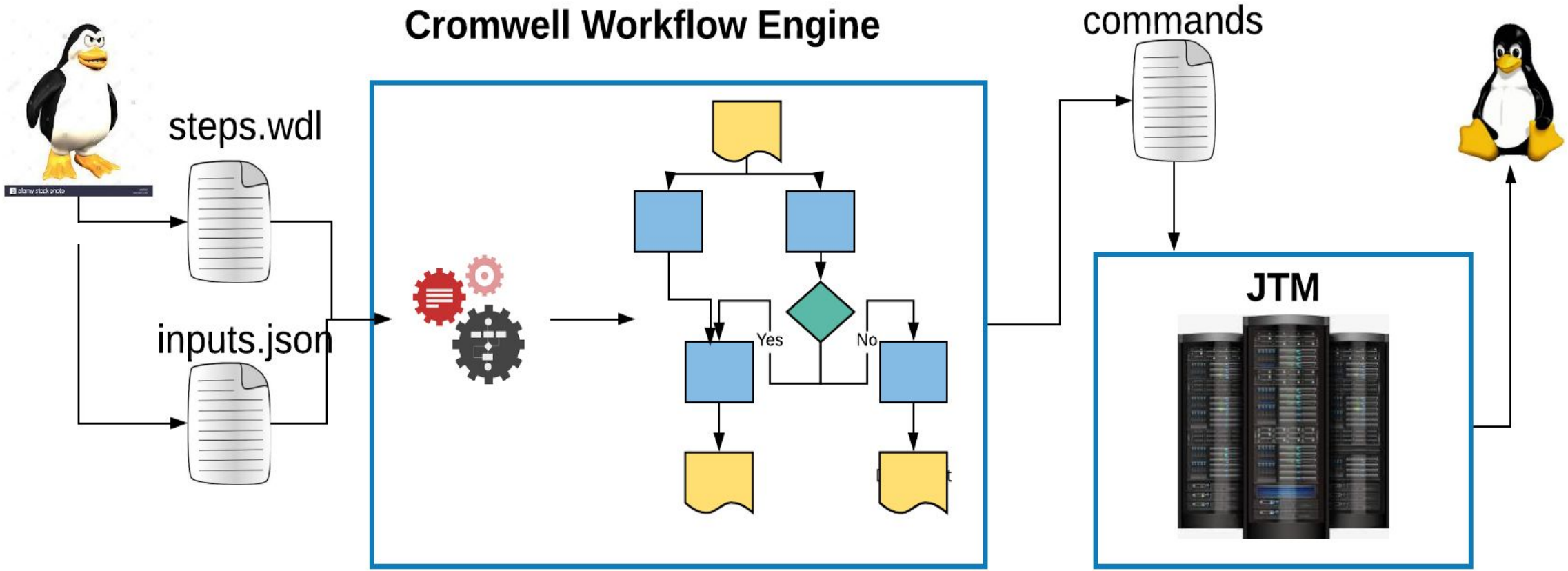
Problems

- **Users do not run jobs: Workflows**
  - Complex dependencies
- **Scheduler learning curve, idiosyncrasies**
  - HPC does not like a lot of jobs, necessitates “job packing”
  - Provokes an infinite number of ad-hoc workflow engine implementations
- **Workload tightly coupled to infrastructure**
  - can not be moved easily
- **Workflow reuse very hard**
- **Reproducible research?**

"All problems in computer science can be solved by another level of indirection"

- David Wheeler

JAWS - JGI Analysis and Workflow System



14

Workflow Example

```
{
    "ScatterGather.test_input": [
        "1.potato",
        "2.potato",
        "3.potato"
    ]
}
```

input.json

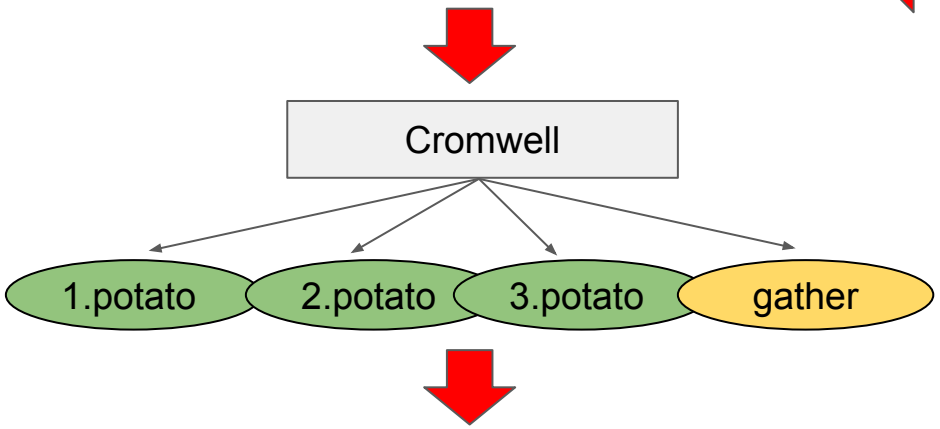
```
workflow ScatterGather {
  Array[String] test_input

  scatter (one in test_input) {
      call my_scatter { input: in=one }
  }
  call my_gather { input: files=stepA.out }
}
```

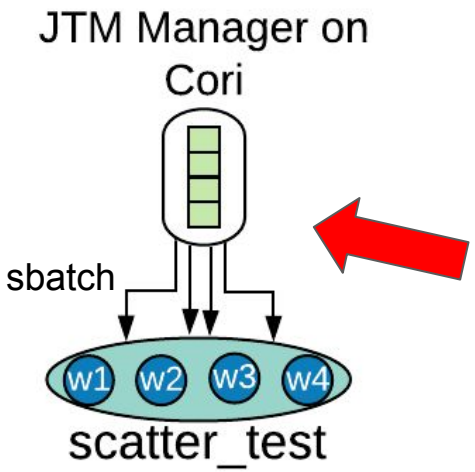
scatter.wdl

```
task my_scatter {

  String in

  runtime {
    cluster: "cori"
    poolname: "scatter_test"
    time: "00:30:00"
    mem: "115GB"
    cpu: 32
    node: 1
    nwpn: 4
  }

  command {
      echo ${in} is sleeping for 5 seconds > file;
      sleep 5
  }
  output { File out = "file" }
}
```

```
task my_gather {
  Array[File] files

  runtime {
    cluster: "cori"
    poolname: "scatter_test"
  }

  command { cat ${sep=' ' files} > final }
  output { String out = "final" }
}
```

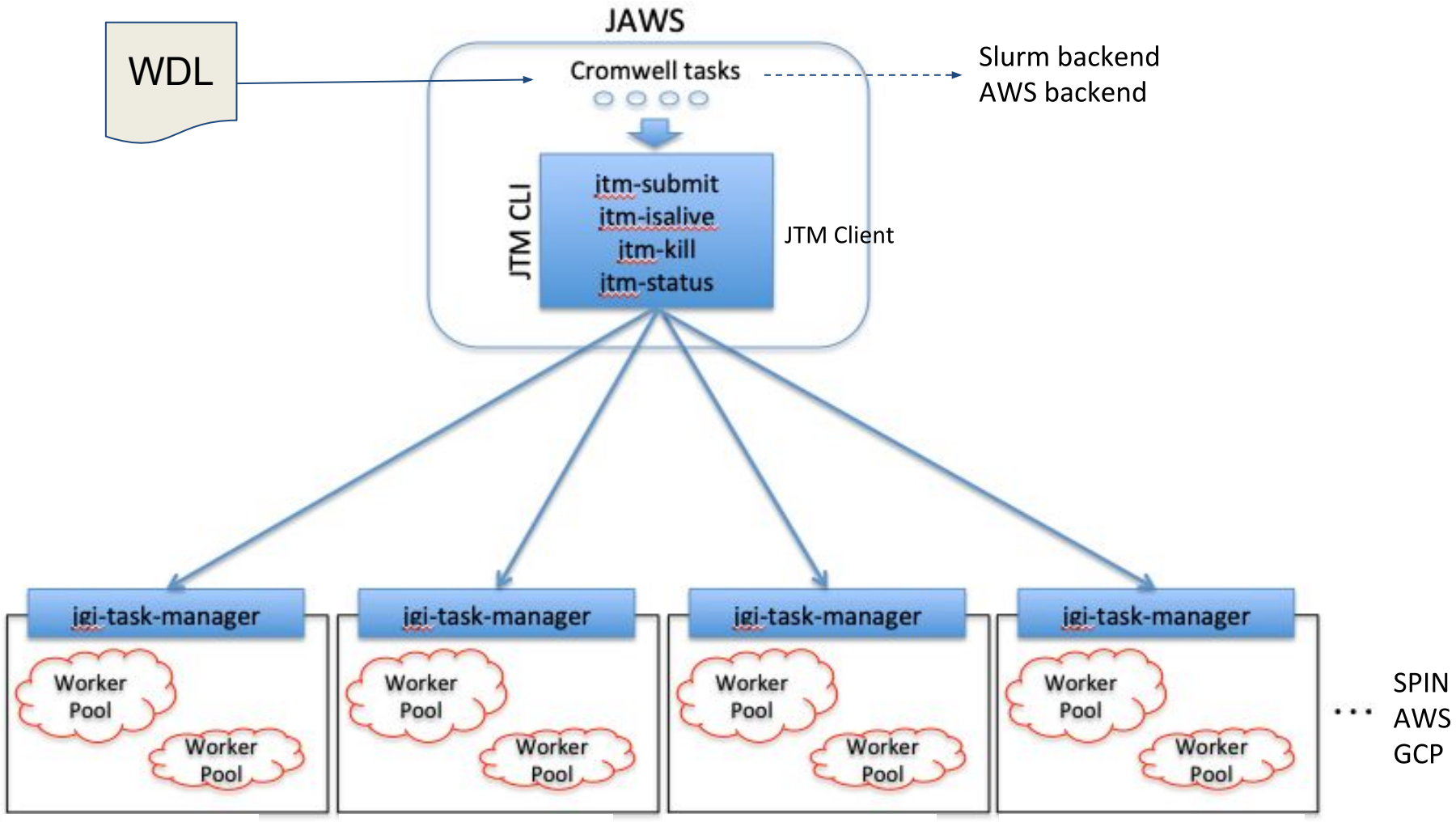
$ jaws submit scatter.wdl input.json





$ jtm-submit -cr <1.potato> -cl cori -tp scatter_test …
$ jtm-submit -cr <2.potato> -cl cori -tp scatter_test …
$ jtm-submit -cr <3.potato> -cl cori -tp scatter_test …
……
$ jtm-submit -cr <gather> -cl cori -tp scatter_test ...

JTM - JGI Task Manager. Pilots.



NERSC Genepool
- Slurm
- 22 (350) nodes
- Shifter

NERSC Cori
- Slurm
- 192 (12000) nodes
- Shifter

Other Site
- PBS
- 512 + 8 GPU nodes
- Singularity

Lab IT Lawrencium
- Slurm
- 600 + 4 GPU nodes
- Singularity

Data Movement - Quick n Dirty

- All of our sites are connected by Globus
- Data specified in input file and moved by JAWS
- Keep cache on scratch filesystem
- Transfer data back to NERSC
- Globus supports “science gateways” - JAWS can transfer as user

The Plan and Conclusion

- **Distributed Computing**
  - Use different sites (NERSC, JGI Cloud, etc)
  - JAWS as middleware
  - Containers to enable mobility of compute
- **JGI Cloud**
  - pilot project
  - access through JAWS exclusively
  - no direct access to NERSC filesystems
- **HEP & JGI - lessons to learn**
  - JAMO vs Rucio, xrootd, etc - replace parts of functionality of JAWS/JAMO/Globus?
  - JAWS vs HTCondor - eg pid namespaces

JGI JOINT GENOME INSTITUTE
A DOE OFFICE OF SCIENCE USER FACILITY

Questions?