JGI JOINT GENOME INSTITUTE
A DOE OFFICE OF SCIENCE USER FACILITY

NERSC User Group SIG: Experimental Facilities

Bryce Foster
2020-07-15

Agenda

JGI Data Factory

JGI's Pipelines

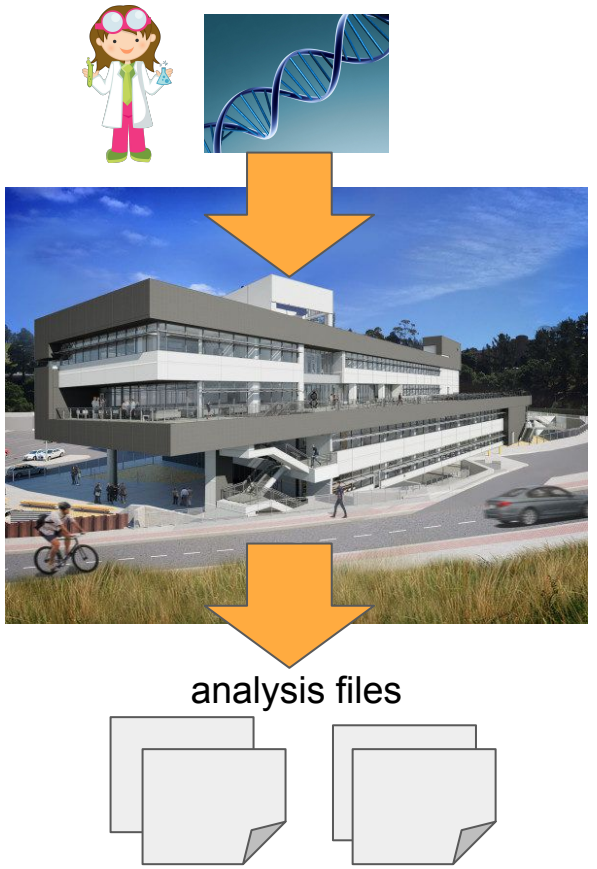
NERSC Usage

Challenges

JGI Overview

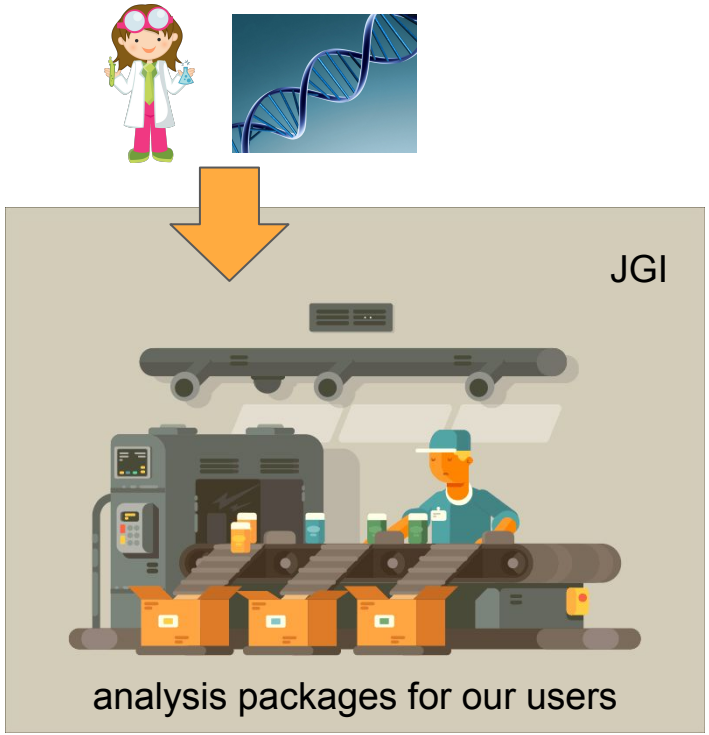
- **JGI's mission**
  - provide the scientific community at large with access to high-throughput, high-quality sequencing, DNA synthesis, metabolomics and analysis capabilities
  - projects involve many important multicellular organisms, microbes and communities of microbes called metagenomes related to the DOE mission areas of bioenergy, understanding global cycles such as the carbon cycle, and biogeochemistry

- **JGI is a data factory responsible for delivering project data to users**
  - Each product has a cycle time requirement to keep data flowing
  - Many projects involve multiple samples and take multiple years to process the data and provide all of the analyses



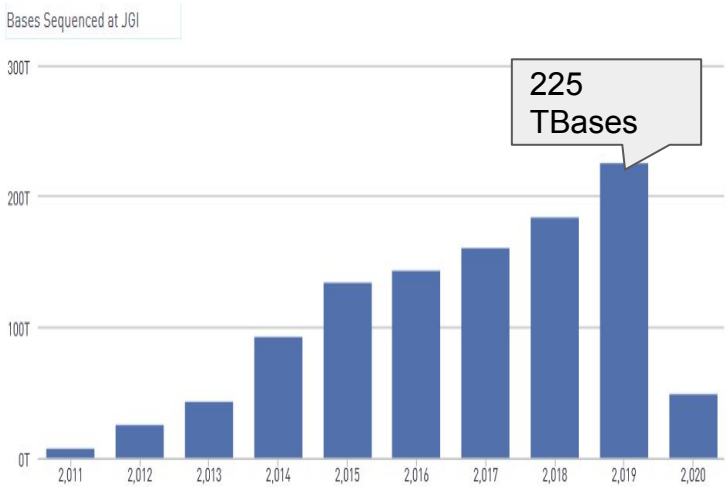
JGI - Data Factory

- **Employees: ~280**
- **FY2020 Budget: $77m**
- **For FY2019 ...**
  - Users: 1940
  - Active User Proposals: 600
  - Active projects: 16,000
  - Samples: 24,000
  - Pipeline runs: 95,000 (RQC only)
    - 30 active pipelines
  - Sequencers: 3 PacBio, 8 Illumina
- **Partnerships: Livermore Lab, Oakridge Lab**



JGI Growth & Cycle Time

- **Projects and sequencing have constantly been growing due to expanded capabilities and new sequencing technologies**
- **Products have expected Cycle Times**
  - some users have priority projects that require quick turn around from sequencing to analysis
  - the synthetic biology group needs to know quickly if the DNA sequence created is what the customer ordered
- **JGI analysts run analyses daily to keep up with the demand**
  - avg 200-300 daily pipelines runs
  - sequencers can produce 2 to 3TB per run



Big Picture - Data Factory

JGI JOINT GENOME INSTITUTE

Metadata

PPS

Converts sequence output to files of sequencing reads

Filtering, assemble reads, QC reports

Microbial, plant and fungal annotations

Physical

Digital

SDM

RQC

IMG

PMOS

GOLD

Genome Portal

Plant

Proposal

Data Warehouse

Researcher can download files from the website

NERSC

Fungal

Pre-Sequencing

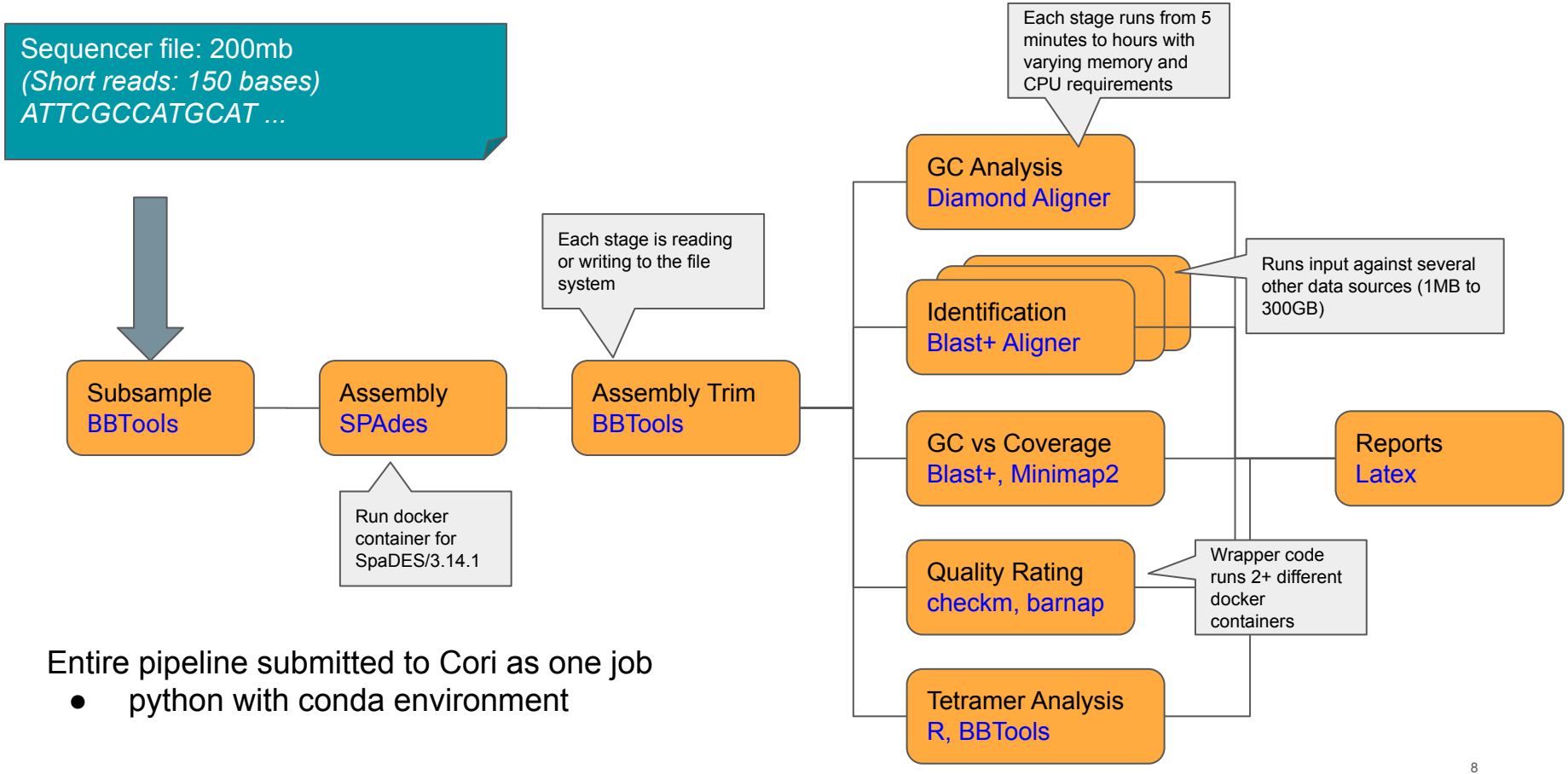
Post-Sequencing

Pipeline Characteristics

- **Sequence data is strings (ATGCGC…)**
  - Input sequence files are large (10MB to 100GB+)
  - Output analysis folders can be 10GB+
  - Input and output data archived to the tape system using HPSS
- **Pipelines run from 5 minutes to more than 7 days**
  - Dependent on pipelines, sample sizes and product types
  - Sequence data runs 2 to 5 different pipelines
- **Heavy disk I/O and high memory**
  - use Project B and CScratch
    - both are used to work around software bugs
  - loading input files or large databases into memory
  - difficult to predict memory requirements ahead of time
- **Wide variety of node usage**
  - Many pipelines run on one node for analysis
  - Some pipelines runs on several nodes for one analysis
  - Run on a workflow node and manage parallel analyses on cluster

Novaseq DNA sequencers runs twice a week and produce 2TB of sequence data for each run and can create 1000s of sequencing files as inputs to pipelines

Example Pipeline



Entire pipeline submitted to Cori as one job

- python with conda environment

NERSC Usage

- **Since 2011, JGI has depended on NERSC resources to run pipelines**
- **Almost all of JGI's analyses runs on the Cori cluster**
  - JGI has its own partition on Cori because JGI requires short queue wait times
  - JGI bought a high memory partition (19 nodes, 1.5 TB) because jobs need more than 128G of RAM
  - heavy usage of Shifter to run Docker images and conda environments
  - use Cori's general partition for overflow capacity, some KNL usage
    - KNL is 3x to 5x slower than Haswell nodes
  - 80% of JGI's usage of Cori is annotation of DNA sequences (what genes are in the DNA)
- **Disk usage**
  - 5 PB of spinning disk (project B, DNA, sandbox)
  - 20 PB of analysis files on tape (NERSC tape system - HSI)
- **Consultants**
  - JGI has 2 "full time" consultant positions split among 3 NERSC staff

NERSC & JGI Cluster Migration History

"Clusters"
Each JGI group had a small cluster

Genepool (UGE/SGE - 360 nodes 128Gb to 1Tb memory)
Racks dedicated to JGI, fairshare for each JGI group

Genepool nodes repurposed for Denovo. Wasn't ready for several months

Denovo
Genepool replacement

Cori (Slurm, Shifter)
New HPC for all of LBL

Need custom scheduler rules giving production users priority

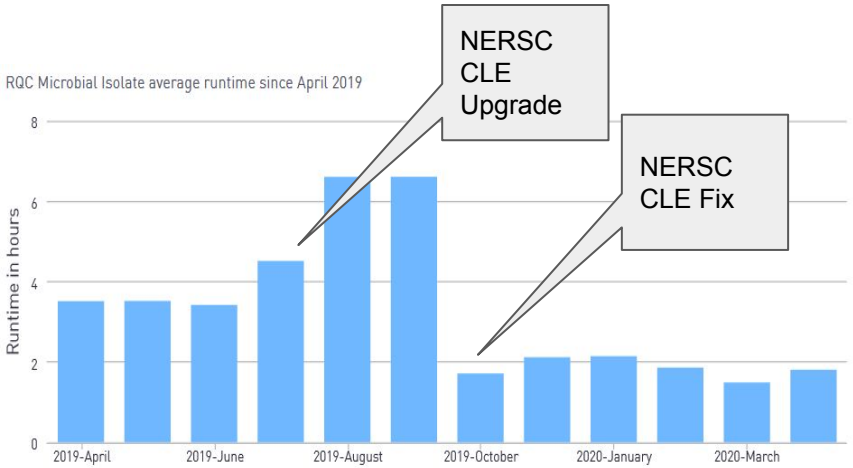
Cori Genepool
192 nodes only for JGI

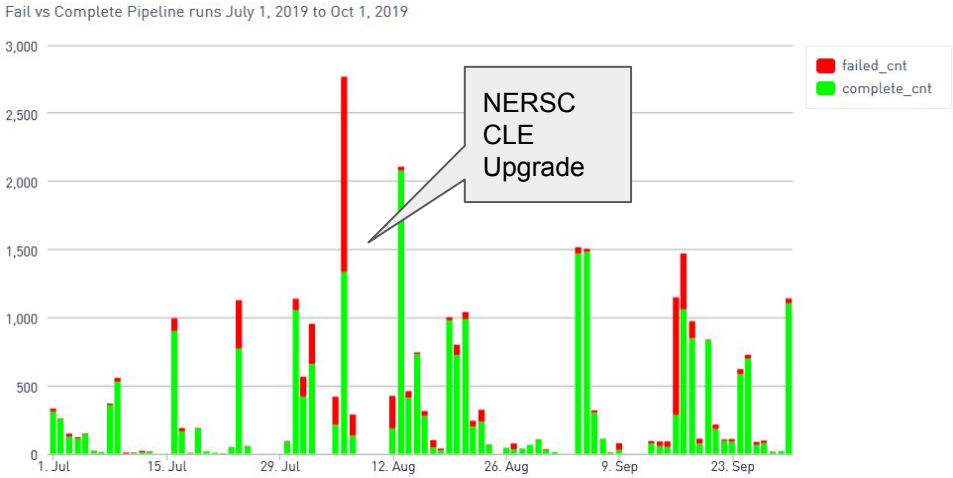
1.5 TB nodes with local disk for JGI's computing needs

Cori Ex Vivo
19 high memory nodes

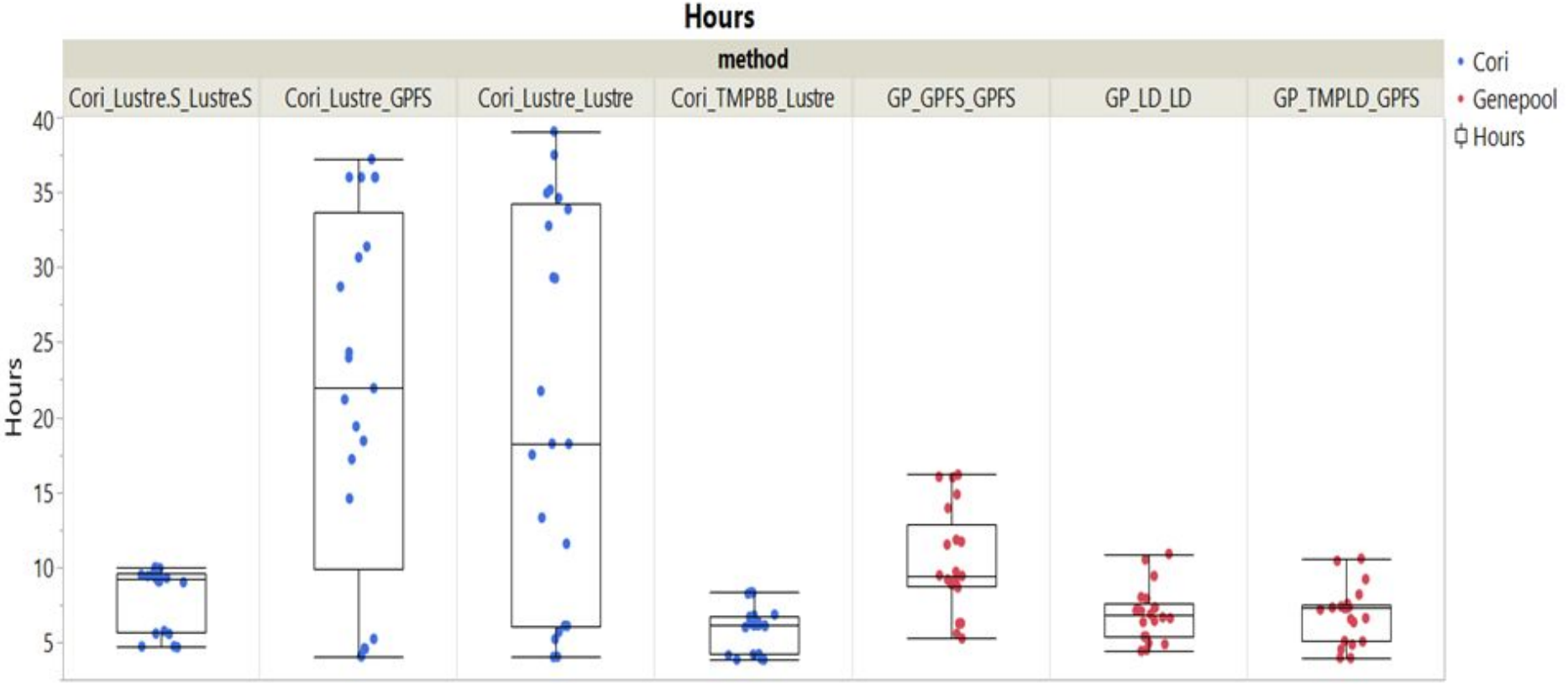
Challenges working on NERSC Clusters

- **"Weather" on Cori**
  - Regular problems with reading or writing files on the network file system (DVS)
  - Slower pipeline throughput because no local disk
  - Monthly maintenance can hold up analysis
    - e.g. need 5 days to run, maintenance in 4 days - jobs won't run until next week
  - Wasted resources spent debugging and rerunning failures





Run Time Experiment - Different Disk Systems

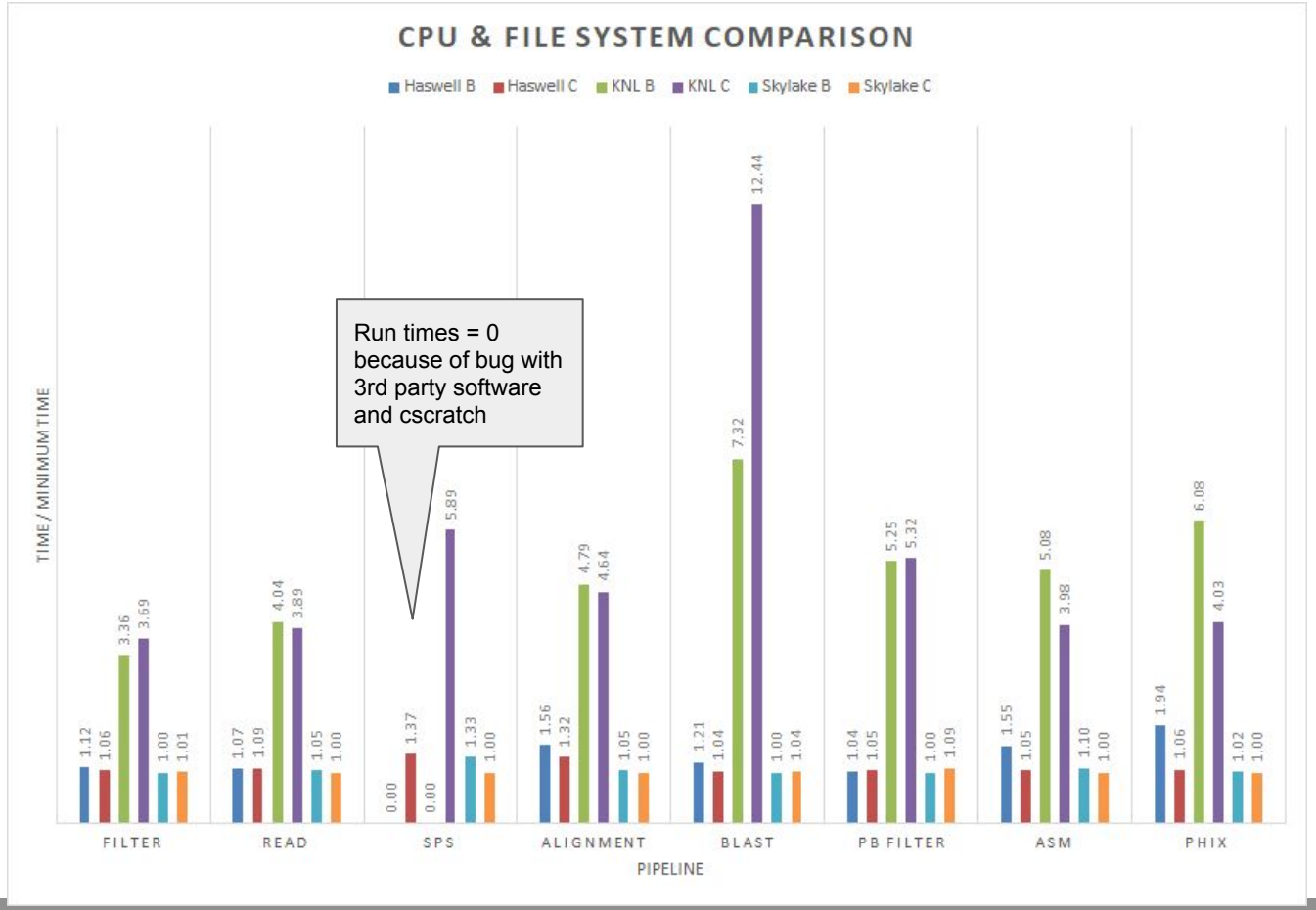


Haswell, KNL and Skylake CPU Comparison

Run times running commonly used pipelines on Haswell, KNL and Skylake using ProjectB (B) and CScratch (c)

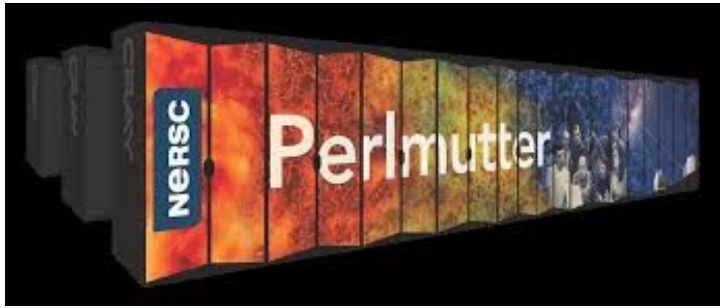
Conclusions

- Haswell and Skylake perform much better for JGI's pipelines
- Using ProjectB or CScratch have little affect on run time



Challenges with NERSC

- **Retooling every few years for a new cluster**
  - interruptions for installation (power work last weekend)
  - changed from using modules to using conda and docker for software packages
  - changed cluster scheduler from SGE/UGE to slurm
    - Our team uses more than 30 different 3rd party software packages
  - trying to use cluster not ready for production
    - file system not mounted or mounted read only
    - scheduler not configured correctly adding addtional cycle time to analysis
  - NERSC chasing the high performance systems isn't beneficial for us
    - what changes to the file system and nodes will need us to retool our pipelines again?
    - bioinformatic pipelines are not GPU-friendly or require a lot of retooling

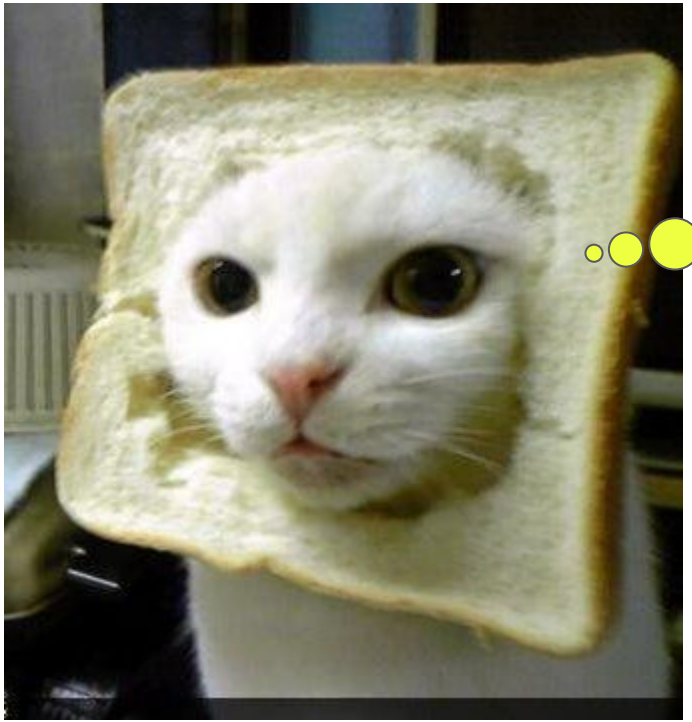


JGI's Wishlist for NERSC

- stable mid-range compute environment dedicated for JGI's computing needs
- local disk on nodes because I/O is much faster (25% faster run time)
- quarterly maintenance (or less) because interruptions affect product cycle time
- longer windows for computing (10 days+) because it is difficult to break up long running 3rd party software
- nodes that can be used to create docker containers at NERSC

- **Benefits**
  - Spend less time retooling code for new clusters
  - Spend more time doing analysis and creating new products for our customers

Discussion



What do we need to do to make our code work at NERSC this time?

Acknowledgements:

- Alicia Clum
- Alex Copeland
- Christa Pennacchio

Hidden Slide