Amazon-Free Big Data Analysis

Michael R. Crusoe
the GED Lab @ MSU
@JKhedron #NGS2013 2013-06-18

Overview

- Dedicated vs Shared computing
- Evaluating Computing Resources
  - XSEDE
    - Mason
    - Lonestar
    - Stampede
    - Blacklight
    - Extended Collaboration Support Service
  - Discovery Environment
  - DIY
- Data Transfer
  - Globus Online

Dedicated vs Shared computing

- Need for dedicated & interactive analysis
  - Amazon AWS or a local box is good for this
- Shared computing systems can offer a cost, performance, and support advantage
  - Free or near free for academics
  - Quality and availability of the support will vary
- There are always tradeoffs

Shared computing

- requires limits to play nicely with others
- a **Job** becomes the unit of computing
- Jobs have preset limits
    - how long they run
    - how many resources they use
- Limits are specified up front
- Requires experimentation to get right
    - Start small

Evaluating Computing Resources

- How do I load data?
  - SCP/SFTP, Globus Online, iRods, FTP
  - At what speeds?
- How long can my data stay there?
- What software is already installed?
  - Can I install on my own without assistance?
- How long can my jobs run?
- Can I share my data from the system with my collaborators?
  - Do they need their own account?

Evaluating Computing Resources (continued)

- Does this site have experience with NGS/bioinformatics?
- Can I run interactive jobs or just batch jobs?
- Do they have consulting services?
- What other support resources are available?
  - Wikis, discussion lists for users, phone support
- Can exceptions be made to any of their policies?
  - Ex: allowing a job to run for weeks or keeping data loaded for over three months.

XSEDE - Overview

- Single administrative interface to many systems
  - Only one application & one password needed
  - Unified directory of installed software
- Free for US academics (see next slide)
- For all XSEDE resources:
  - Data transfer via GridFTP, SSH, GlobusOnline
  - Extended collaboration support available
  - Free online training
  - Super easy to get startup allocation
  - Exceptions to policies can be made if needed

XSEDE - Eligibility

"the principal investigator (PI) must be a researcher or educator at a U.S.-based institution, including federal research labs or commercial organizations"

"[...] investigators with support from any funding source, not just NSF, are encouraged to apply. If your institution is not a university or a two- or four-year college, special rules may apply."

XSEDE - Mason

[glamour shot not available]

XSEDE - Mason

Part of the National Center for Genome Analysis Support (NCGAS)

- nodes
  - 16 largemem: ½TB memory, 32 1.86Ghz cores
- storage
  - 0.5PB shared scratch, 60 day purge, no quota
- Jobs
  - PBS / QSub interface
  - Max job time: 14 days (!)
- Lots of software pre-installed; more available upon request

XSEDE - Mason continued

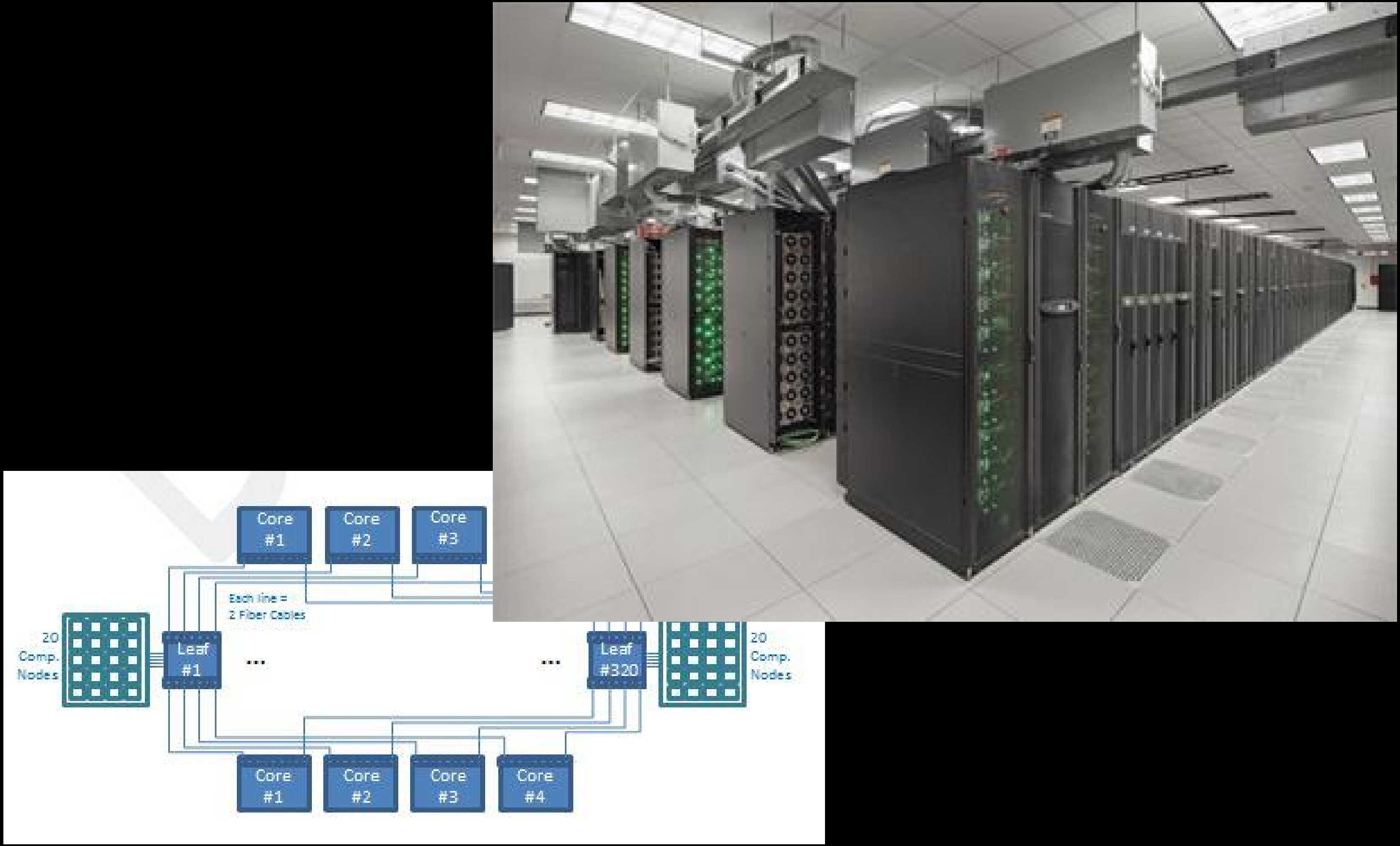
- Experience w/ NGS: yes!
- They provide: "[c]onsulting services for biologists [who are] undertaking genome analysis. [They also provide a]ssistance with genome analysis software on [their] systems."

XSEDE - Lonestar

XSEDE - Lonestar

- nodes
  - 1,888 compute nodes: 12 cores, 24 GB memory
  - 5 largemem nodes: 24 cores, 1TB memory
- storage
  - scratch: 10 day purge, no quota
  - work: no purge, 250GB quota
- jobs
  - PBS / QSUB interface
  - max job time: one day

XSEDE - Stampede



XSEDE - Stampede

- nodes
  - 6,400 compute nodes: 16 cores, 32GB memory
  - 16 large memory: 32 cores, 1TB memory
- storage
  - local ephemeral: 250GB, purged at end of job
  - scratch: 8.5PB total , 10 day purge, no quota
  - work: 450TB total, no purge, 400GB quota
- jobs
  - SLURM interface
  - max job time: one day

XSEDE - Blacklight

XSEDE - Blacklight

large shared memory system; single process can access up to 16TB w/o MPI

- blades
  - 256 blades: 16 cores @ 2.27 Ghz, 128GB memory
- storage
  - no local ephemeral storage**
  - scratch: 291TB total , 21 day purge, no quota
- jobs
  - PBS / QSub interface
  - max job time: 2 or 4 days**

**Exceptions: yes! fast local storage and extended job time available upon request

XSEDE - Extended Collaboration Support Service

Collaboration of weeks to a year

"Expertise is available in a wide range of areas, from performance analysis and petascale optimization to the development of community gateways and work and data flow systems."

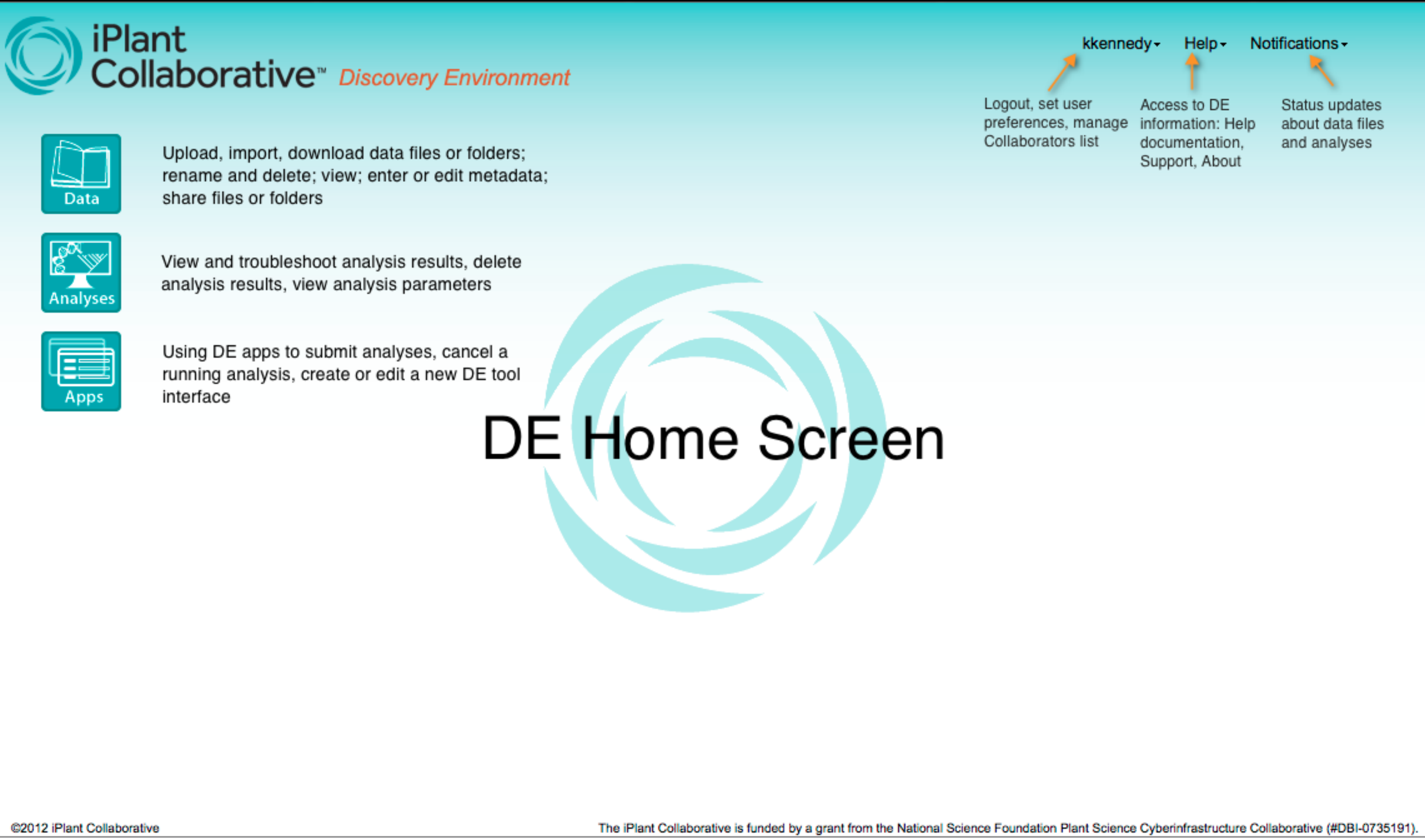
"[S]taff will [also] support extensive training, education, and outreach activities to foster integration of research and education."

iPlantCollaborative Discovery Environment

https://www.iplantcollaborative.org/discover/discovery-environment

- 90+ command line tools in a webapp
- Users can customize and integrate new tools through the graphical interface
- Storage
  - lots; no quota & no purge
  - can share with others via special links

DE Home Screen



Logout, set user preferences, manage Collaborators list

Access to DE information: Help documentation, Support, About

Status updates about data files and analyses

Upload, import, download data files or folders; rename and delete; view; enter or edit metadata; share files or folders

View and troubleshoot analysis results, delete analysis results, view analysis parameters

Using DE apps to submit analyses, cancel a running analysis, create or edit a new DE tool interface

The iPlant Collaborative is funded by a grant from the National Science Foundation Plant Science Cyberinfrastructure Collaborative (#DBI-0735191).

Sharing is hard: the DIY approach

Sharing is hard: the DIY approach

- $5,000-10,000 for a 24-32 core machine with a half-terabyte of memory and 15-20 terabytes of persistent storage
  - Yes, this will be cheaper next year
- Where will it go?
  - Very loud; you will also want to have good local network and internet connectivity
- What about data archiving?
- What about backups?
- People costs are likely to be larger than hardware costs

Sharing is hard: the DIY approach

TL;DR:

Don't DIY without hiring a fulltime person

(students don't count)

Data Transfer with Globus Online

Data Transfer with Globus Online

https://www.globusonline.org/

- fast, secure, & easy file transfer for big data
- web based & CLI
- Free!
- integrates with XSEDE and many other systems
- Optional data sharing ($)

Acknowledgments

My 2012-2013 employer: The Kusumi Lab @ Arizona State University

Especially Walter Eckalbar, Elizabeth Hutchins, and Prof. Kenro Kusumi

Chris Welcher for the HPC glamour shots