Intergalactic Travel

sending usegalaxy.org through the wormhole

Nate Coraor¹
Dannon Baker²
John Chilton¹

The Galaxy Team
¹Penn State University
²Johns Hopkins University
Major Challenges

• Aging infrastructure
  • No grant support for hardware
  • Grew organically over time
• Growth unsustainable
  • Not enough compute
  • Not enough storage
• Backups prohibitively expensive
Major Solutions

• Direct resources
  • from iPlant to move usegalaxy.org from Penn State to new, dedicated hardware at the Texas Advanced Computing Center
  • from the Pittsburgh Supercomputing Center to back up usegalaxy.org on the Data Supercell

• XSEDE Allocation (~600,000 SUs)
  • TACC Stampede
  • PSC Blacklight (16 TB shared memory!)
Down the Wormhole

• usegalaxy.org housed over 600 TB of user data
• How do we move the data from Penn State to TACC?
• How do we do move as quickly and with as little impact to users as possible?
Networking

- 10 Gb/s connection to XSEDE via PSC
- Galaxy: The first entity on the XSEDE network not an XSEDE member institution
Data Transfer Mechanisms

**Globus Online**
- Hands-off big data transfer
- Easy, fault-tolerant
- Performance auto-tuning
- Limited to 3 simultaneous transfers (we had 6 1Gb/s fileservers)

**Globus GridFTP**
- The transfer technology underlying Globus Online
- Not as resilient as rsync or Globus Online

**rsync over HPN-SSH**
- Massive improvement over standard SSH
- rsync capable of wire speed(!)
## Data Transfer Mechanisms

<table>
<thead>
<tr>
<th>Globus Online</th>
<th>Globus GridFTP</th>
<th>rsync over HPN-SSH</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Hands-off big data transfer</td>
<td>• The transfer technology underlying Globus Online</td>
<td>• Massive improvement over standard SSH</td>
</tr>
<tr>
<td>• Easy, fault-tolerant</td>
<td>• Not as resilient as rsync or Globus Online</td>
<td>• rsync capable of wire speed(!)</td>
</tr>
<tr>
<td>• Performance autotuning</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Limited to 3 simultaneous transfers (we had 6 1Gb/s fileservers)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
~6 days to copy 600 TB
 días to copy 600 TB
30 days to copy 600 TB
Galaxy's Hierarchical Object Store

Galaxy Server Processes

**Read Data**

- In Corral? (Yes → Corral)
  - No → In Staging?
    - Yes → Corral Staging
    - No → In PSU?
      - Yes → Penn State
      - No → Object Not Found

**Write Data**
LWR Pulsar

- Run jobs on remote resources **without a shared filesystem** or a scheduler/DRM
- Like Galaxy, interfaces with PBS, SGE, Condor, Slurm, etc.
- Runs jobs on Stampede, Blacklight, ???
- Communicates with Galaxy via AMQP
Walltime Resubmission

- Mean runtime of jobs over 120 seconds
  - bowtie: 20 minutes
  - bwa: 51 minutes
  - bowtie2: 28 minutes
  - cufflinks: 45 minutes
  - tophat: 153 minutes
  - tophat2: 165 minutes
- Walltime for jobs in this queue: 2 days
Walltime Resubmission

"Big" NGS/Multicore Job

- dynamic walltime

Galaxy dedicated cluster

- hit walltime?
  - Yes
  - Stampede
  - No

Sorry 🙁😔🙁

- hit walltime?
  - Yes
  - No

Done! /\(\cdot\,\cdot\,\cdot/)\/
State of Affairs

- **usegalaxy.org** running at TACC since October 8, 2013
- Data transfer did not complete until November
- Jobs running on dedicated resources
- Galaxy Test running jobs on Stampede

**Up next**
- Galaxy Main jobs on Stampede
- Trinity on Blacklight
- Charge jobs to users’ XSEDE Allocations
- Cloud Bursting
Credits

- Texas Advanced Computing Center
  - Dan Stanzione
  - Matt Vaughn
  - Chris Jordan
  - Mike Packard
  - Nathaniel Mendoza

- iPlant Collaborative
  - Stephen Goff

- Pittsburgh Supercomputing Center
  - Philip Blood
  - Kathy Benninger
  - Robert Budden
  - Jared Yanovich
  - Josephine Palencia

- and the Galaxy Team and community

Galaxy is supported in part by NSF, NHGRI, Pennsylvania Department of Public Health, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Johns Hopkins University
The Galaxy Team

Enis Afgan  Dannon Baker  Dan Blankenberg  Dave Bouvier  Marten Čech  John Chilton

Dave Clements  Nate Coraor  Carl Eberhard  Jeremy Goecks  Sam Guerler  Jen Jackson

Greg Von Kuster  Ross Lazarus  Nick Stoler  Anton Nekrutenko  James Taylor

http://wiki.galaxyproject.org/GalaxyTeam