Open Science Grid

Galaxy based BLAST submission to distributed high throughput computing resources

Rob Quick

Slides Prepared by Soichi Hayashi

Open Science Grid Operations

Indiana University / Research Technologies

Topics

- What is BLAST / Galaxy?
- Why BLAST on OSG?
- How to run BLAST on HTC?
- Conclusion and future TODO...

NCBI-BLAST

NCBI (National Center for Biotechnology Information)

BLAST (Basic Local Alignment Search Tool)

Popular application for Bioinformaticists

Compares biological sequences

- Identify unknown sequences
- Discover related organism

Database Source fasta

```
>gi|6226515|ref|NC_001224.1| Saccharomyces cerevisiae mitochondrion
TTCATAATTAATTTTTTATATATATATTATATTATAATATTAATTTATATTATAAAAATAATATTTATTATTAAAATAT
T
TATTCTCCTTTCGGGGTTCCGGCTCCCGTGGCCGGGCCCCGGAATTATTAATTAATAATAAATTATTATTAATAATTAT
T
TATTATTTTATCATTAAAATATATAAATAAAAAATATTAAAAAGATAAAAAAAATAATGTTTATTCTTTATATAAATTA
T
ATATATATATATAATTAATTAATTAATTAATTAATTAATAATAAAAATATAATTATAAATAATATAAATATTATTCTTT
A
TTAATAAATATATATTTATATATTATAAAAGTATCTTAATTAATAAAAATAAACATTTAATAATATGAATTATATATTA
T
TATTATTATTAATAAAATTATTAATAATATATCAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
T
... (150,000 lines)
```

```
$ makeblastdb -in
yeast.fasta -dbtype nucl -
out yeast
```

Blast DB

Input Query (Unknown Organism)

```
>CHR1.19971009 Chromosome I Sequence
CCACACCACACCCACACACCCACACACCACACCACACACCACACCCACACACACA
CATCCTAACACTACCCTAACACAGCCCTAATCTAACCCTGGCCAACCTGTCTCTCAACTT
ACCCTCCATTACCCTGCCTCCACTCGTTACCCTGTCCCATTCAACCATACCACTCCGAAC
CACCATCCATCCCTCTACTTACTACCACTCACCCACCGTTACCCTCCAATTACCCATATC
CAACCCACTGCCACTTACCCTACCATTACCCTACCATCCACCATGACCTACTCACCATAC
TGTTCTTCTACCCACCATATTGAAACGCTAACAAATGATCGTAAATAACACACACGTGCT
TACCCTACCACTTTATACCACCACCACATGCCATACTCACCCTCACTTGTATACTGATTT
TACGTACGCACACGGATGCTACAGTATATACCATCTCAAACTTACCCTACTCTCAGATTC
CACTTCACTCCATGGCCCATCTCTCACTGAATCAGTACCAAATGCACTCACATCATTATG
CACGGCACTTGCCTCAGCGGTCTATACCCTGTGCCATTTACCCATAACGCCCATCATTAT
CCACATTTTGATATCTATATCTCATTCGGCGGTCCCAAATATTGTATAACTGCCCTTAAT
ACATACGTTATACCACTTTTGCACCATATACTTACCACTCCATTTATATACACTTATGTC
AATATTACAGAAAAATCCCCACAAAAATCACCTAAACATAAAAATATTCTACTTTTCAAC
```

```
$ blastn -db mydb -query input_query.fasta -out output.txt -outfmt 1
```

```
comp10597_c0_seq1    Uextra    100.00    28 0 0 168    195    3953904    3953931    4e-06    52.8
comp10597_c0_seq1    Uextra    100.00    28 0 0 168    195    28550642   28550615   4e-06    52.8
comp12438_c0_seq1    2L        100.00    29 0 0 116    144    8509466    8509494    2e-06    54.7
comp12438_c0_seq2    2L        100.00    29 0 0 134    162    8509466    8509494    2e-06    54.7
```

Common Blast Databases

NCBI RefSeq Databases

NT/NR (10-20 parts 400-800M each compressed)

Collection of taxonomically diverse, non-redundant and richly annotated sequences.

* plasmids, organelles, viruses, archaea, bacteria, and eukaryotes.

patnt/pataa (1-4 parts 1G each)

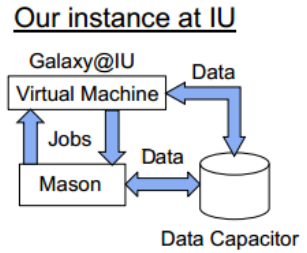
Patent database from USPTO or from EU/Japan Patent Agencies via EMBL/DDBJ

Flybase Databases

dmel-all-chromosome

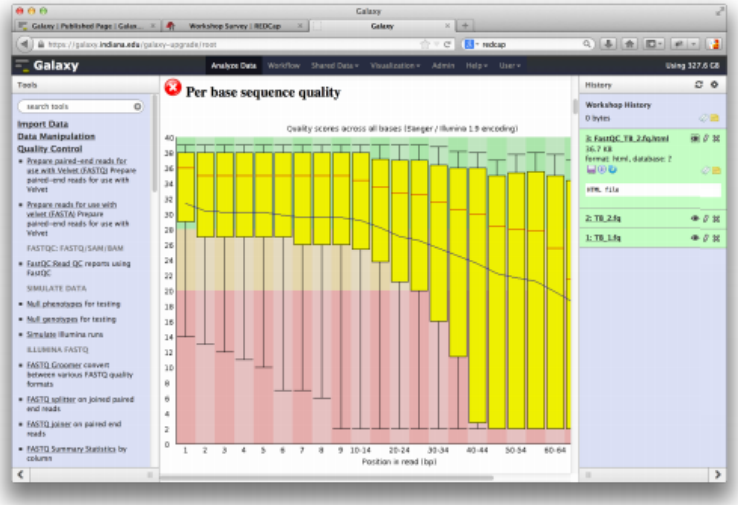
Galaxy

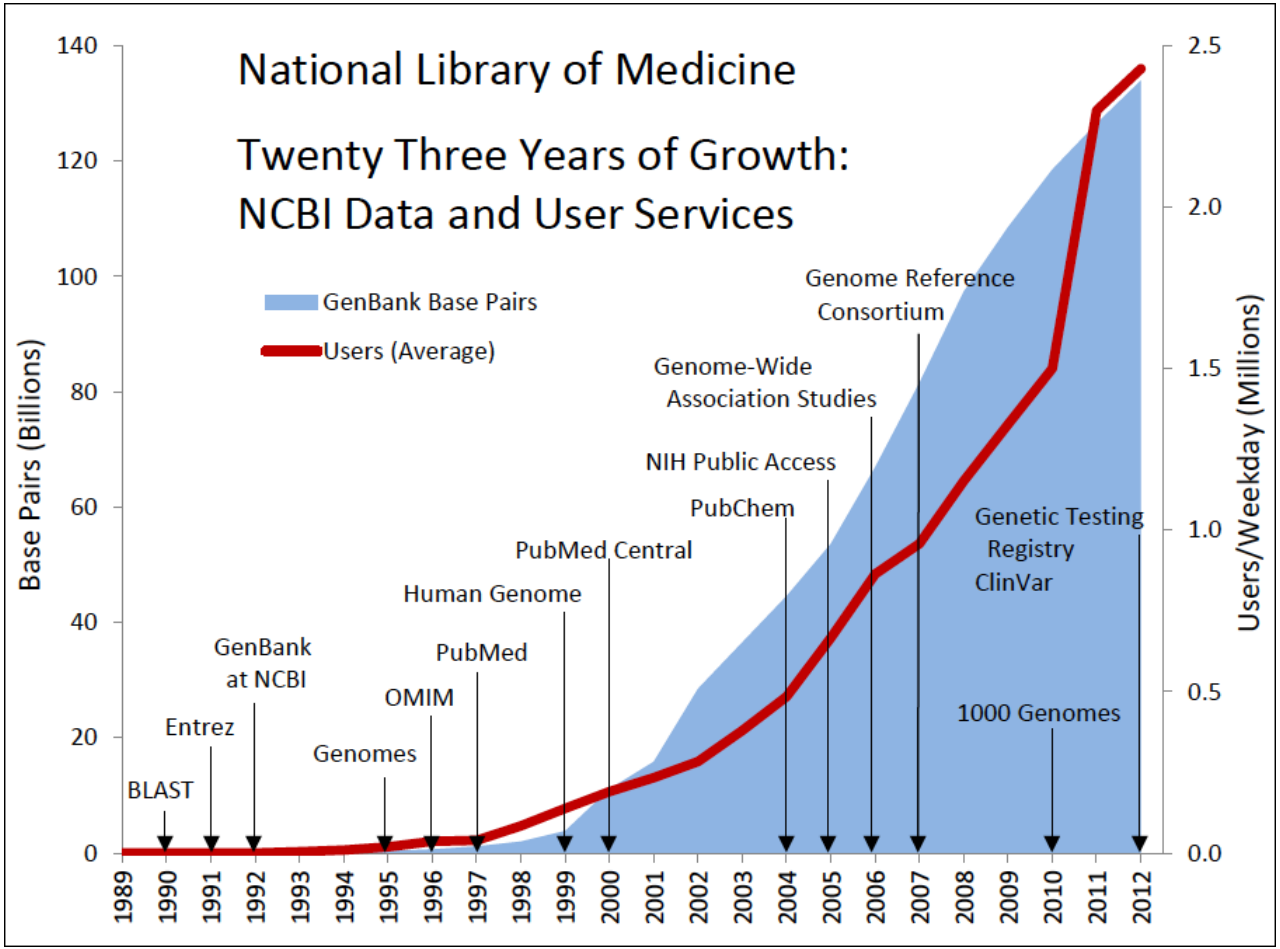
A popular Web-based platform for data intensive biomedical research



NCGAS (National Center for Genome Analysis Support) hosts an instance of Galaxy portal

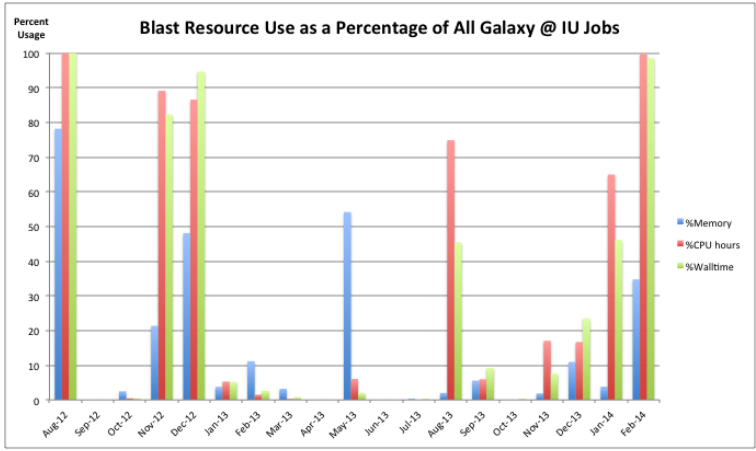
- IU Mason Cluster (8TB-memory)
- Access to IU DC2 (3.5PB)
- Genome assembly
- Large-scale phylogenetic software
- Blast

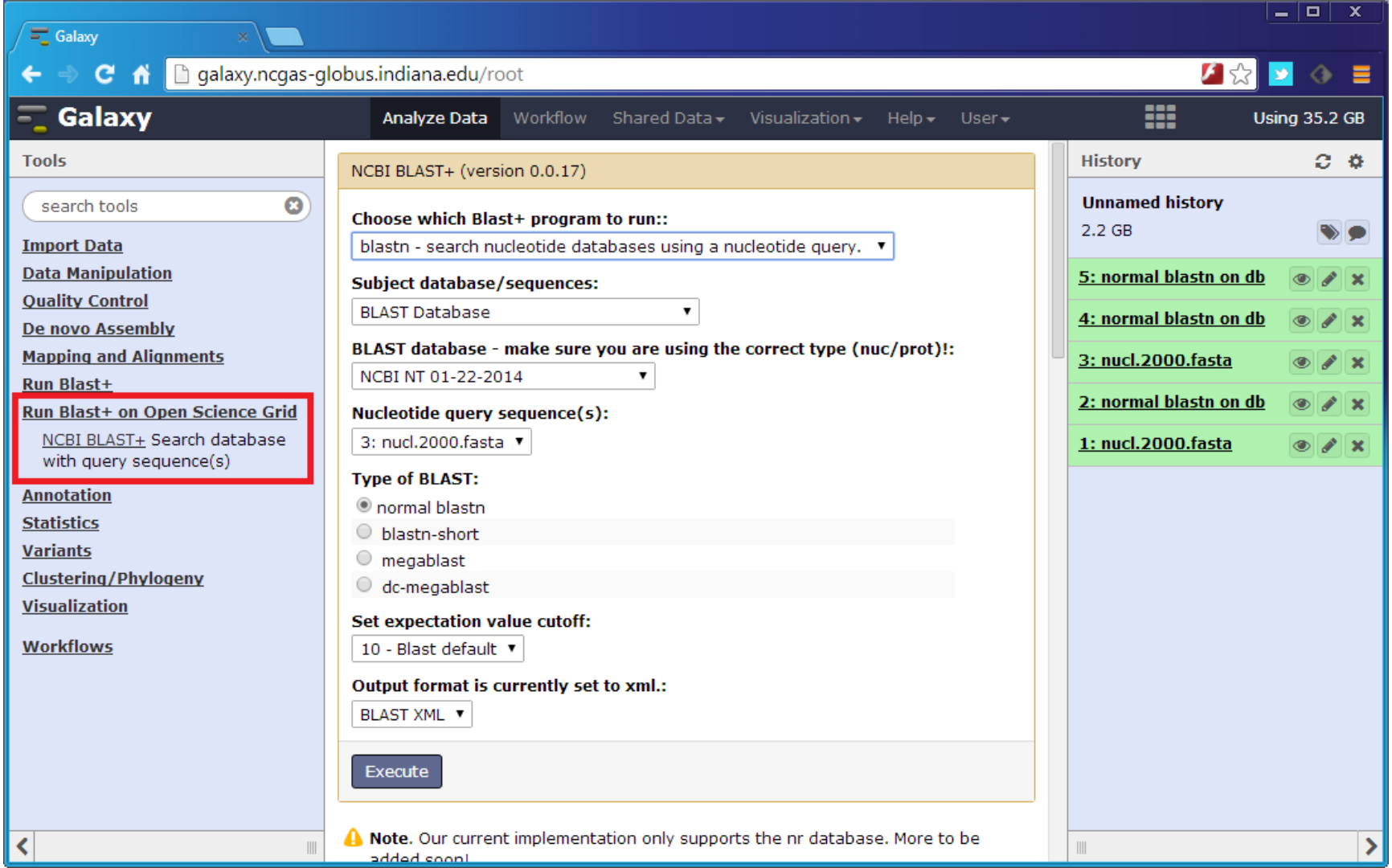



National Library of Medicine
Twenty Three Years of Growth:
NCBI Data and User Services
GenBank Base Pairs
Users (Average)
Base Pairs (Billions)
Users/Weekday (Millions)
BLAST
Entrez
GenBank at NCBI
Genomes
OMIM
PubMed
Human Genome
PubMed Central
PubChem
NIH Public Access
Genome-Wide Association Studies
Genome Reference Consortium
Genetic Testing Registry
ClinVar
1000 Genomes
0
20
40
60
80
100
120
140
0.0
0.5
1.0
1.5
2.0
2.5
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012

Why BLAST on OSG?

- BLAST is CPU intensive (not memory)
- IU/Mason is not an optimal resource to run BLAST
- Growth in data volume will squeeze available resource capacity at NCGAS in coming years.
- OSG's opportunistic resource could be used as an alternative for Mason and can provide necessary resource capacity.



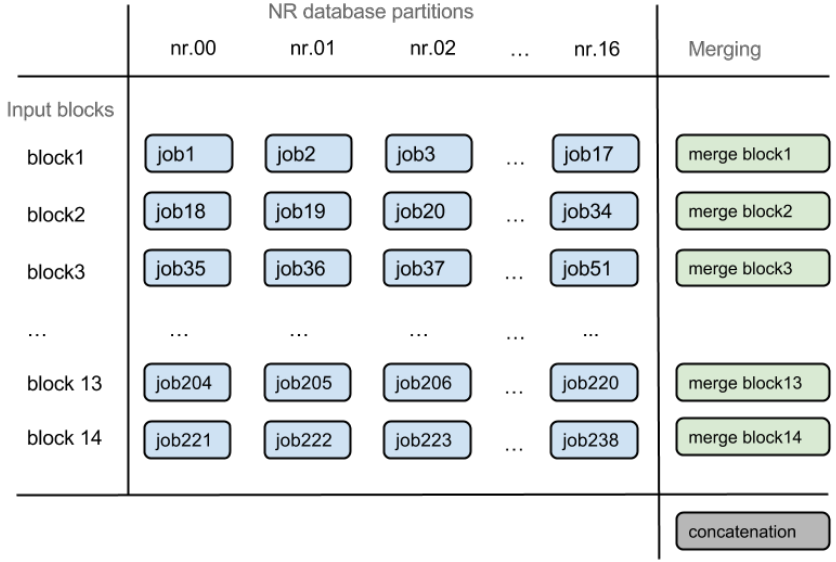

Galaxy
galaxy.ncgas-globus.indiana.edu/root
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 35.2 GB
Tools
search tools
Import Data
Data Manipulation
Quality Control
De novo Assembly
Mapping and Alignments
Run Blast+
Run Blast+ on Open Science Grid
NCBI BLAST+ Search database with query sequence(s)
Annotation
Statistics
Variants
Clustering/Phylogeny
Visualization
Workflows
NCBI BLAST+ (version 0.0.17)
Choose which Blast+ program to run::
blastn - search nucleotide databases using a nucleotide query.
Subject database/sequences:
BLAST Database
BLAST database - make sure you are using the correct type (nuc/prot)!:
NCBI NT 01-22-2014
Nucleotide query sequence(s):
3: nucl.2000.fasta
Type of BLAST:
normal blastn
blastn-short
megablast
dc-megablast
Set expectation value cutoff:
10 - Blast default
Output format is currently set to xml.:
BLAST XML
Execute
Note. Our current implementation only supports the nr database. More to be added soon!
History
Unnamed history
2.2 GB
5: normal blastn on db
4: normal blastn on db
3: nucl.2000.fasta
2: normal blastn on db
1: nucl.2000.fasta

osg-blast (v2)

- Written in nodejs / node-osg & node-htcondor modules
- Can be installed on any OSG submit hosts via “npm install osg-blast”
- Hosted databases (NT/NR) distributed via OASIS (CVMFS)
- Needs to be highly reliable and autonomous
  - Handle unexpected issues well
  - Needs to figure out the best configuration by itself.
  - Report site specific issues to GOC (and recover)
  - Cleanup after itself (removing temp files, canceling jobs)

osg-blast (v2)

- Splits both input queries / databases and run all jobs in parallel.
- Results are merged to create a single output sorted by e-value.

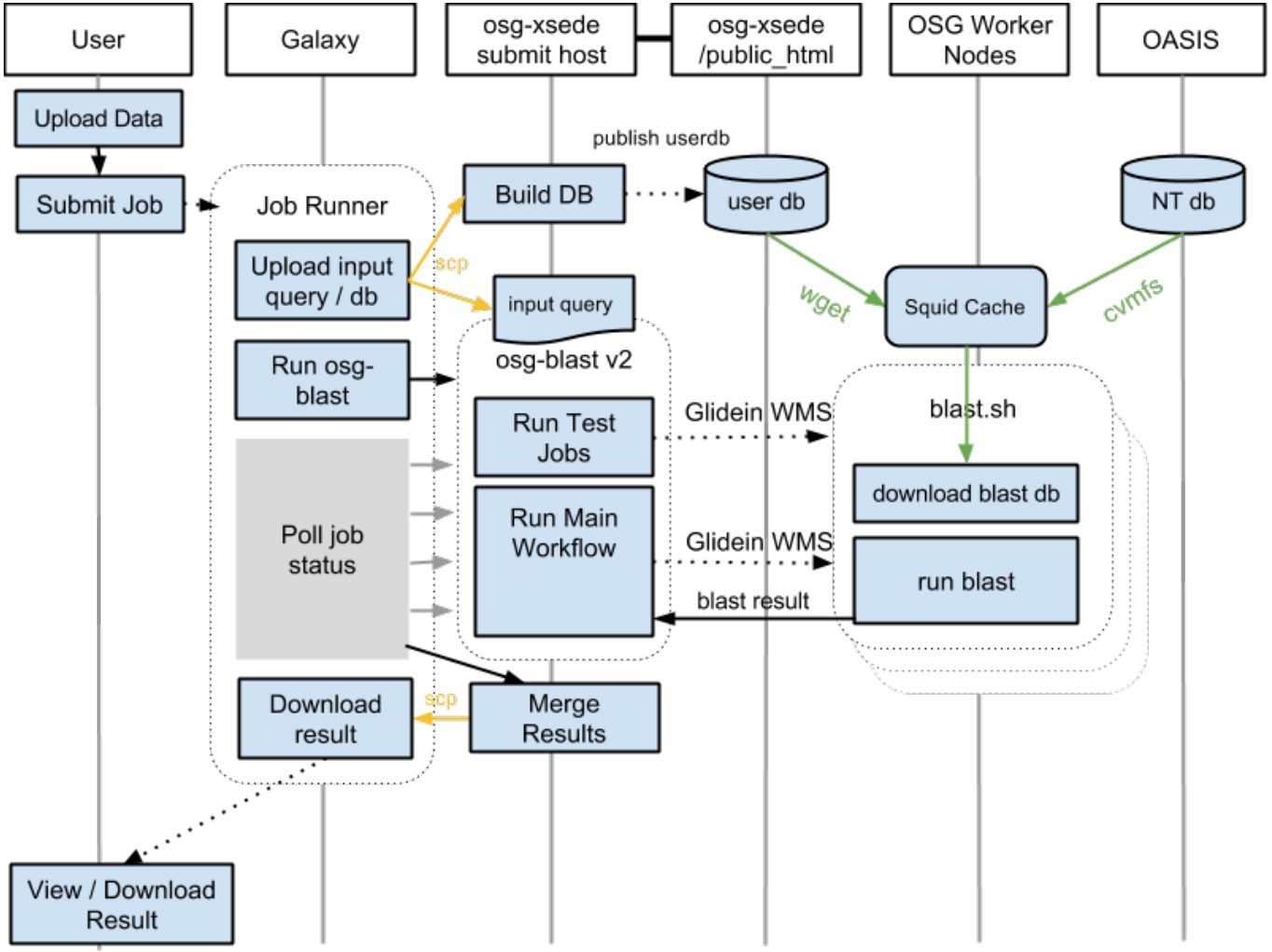


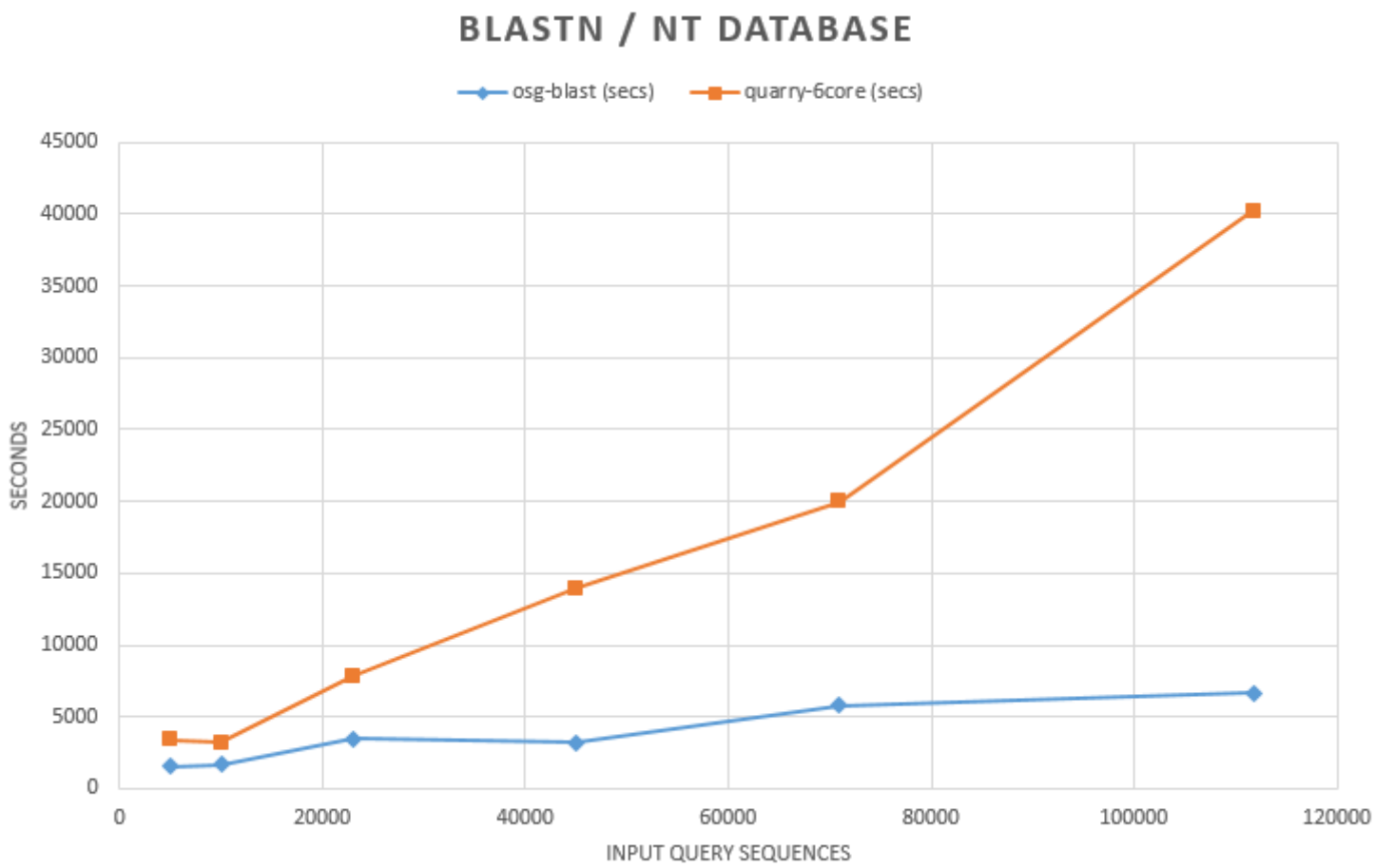
Test Stage

- Determine best input block size
- Detects issue with user input / OSG environment.

Main Stage

- Submit all jobs using information gathered during the test stage.
- Use -dbsize to correct e-value


User
Galaxy
osg-xsede submit host
osg-xsede /public_html
OSG Worker Nodes
OASIS
Upload Data
Submit Job
Job Runner
publish userdb
Build DB
user db
NT db
Upload input query / db
scp
input query
wget
Squid Cache
cvmfs
Run osg-blast
osg-blast v2
Run Test Jobs
Glidein WMS
blast.sh
download blast db
Poll job status
Run Main Workflow
Glidein WMS
run blast
blast result
Download result
scp
Merge Results
View / Download Result


BLASTN / NT DATABASE
osg-blast (secs)
quarry-6core (secs)
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
SECONDS
0
20000
40000
60000
80000
100000
120000
INPUT QUERY SEQUENCES

Conclusions

- Clearly, we will need more computing resources to run BLAST in coming years, and OSG's opportunistic environment can provide that need.
- Galaxy allows bioinformatics community to use existing UI to submit BLAST jobs.
- BLAST works well in HTC environment, and it seems to scale as expected using OSG's opportunistic resources.

Challenges / Future Goal

- osg-blast workflow needs to be highly robust (error-tolerant), reliable, and self-diagnosing to be practical (can't rely on users to fix problems)
- osg-blast output merger needs to be implemented for other output formats.
- Might need to explore alternative to CVMFS for hosting BLAST DBs.

Acknowledgements

Bill Barnett, Tom Doak, Rich LeDuc (SCT @ IU)

Ruth Pordes, Chander Seghal (Fermilab)

Derek Weitzel (UNL)

Mats Rynge (Information Science Institute @ USC)

Alain Deximo, Kyle Gross, Tom Lee, Vince Neal, Chris Pipes, Elizabeth Prout, Michel Tavares, Scott Teige(OSG Operations Center @ IU)

Contacts

Soichi Hayashi hayashis@iu.edu @soichih | soichi.us

Rob Quick rquick@iu.edu