PURDUE UNIVERSITY

# BLAST and Bioinformatics Applications on Purdue's DiaGrid

May 3, 2012

Brian Raub
Purdue University
braub@purdue.edu



ITaP INFORMATION TECHNOLOGY AT PURDUE

PURDUE UNIVERSITY

# Where were we?

- Over 37 kilocores across campus
  - Three community clusters (Steele, Coates, Rossmann)
  - Two "ownerless" clusters (Radon, Miner)
  - CMS Tier-2 cluster
  - Other small clusters
  - Instructional labs and academic departments



ITaP INFORMATION TECHNOLOGY AT PURDUE

PURDUE UNIVERSITY

# … and what about now?

- Nearly 50 kilocores across campus!
  - Two new community clusters
    - Hansen – Dell nodes w/ four 12-core AMD Opteron 6176 processors
    - Carter – HP nodes w/ 2 8-core Intel Xeon-E5 processors (Sandy Bridge)
  - Carter ranks 54th in the latest Top500.org list for fastest supercomputers
  - Carter is the nation's fastest campus supercomputer



ITaP INFORMATION TECHNOLOGY AT PURDUE

PURDUE UNIVERSITY

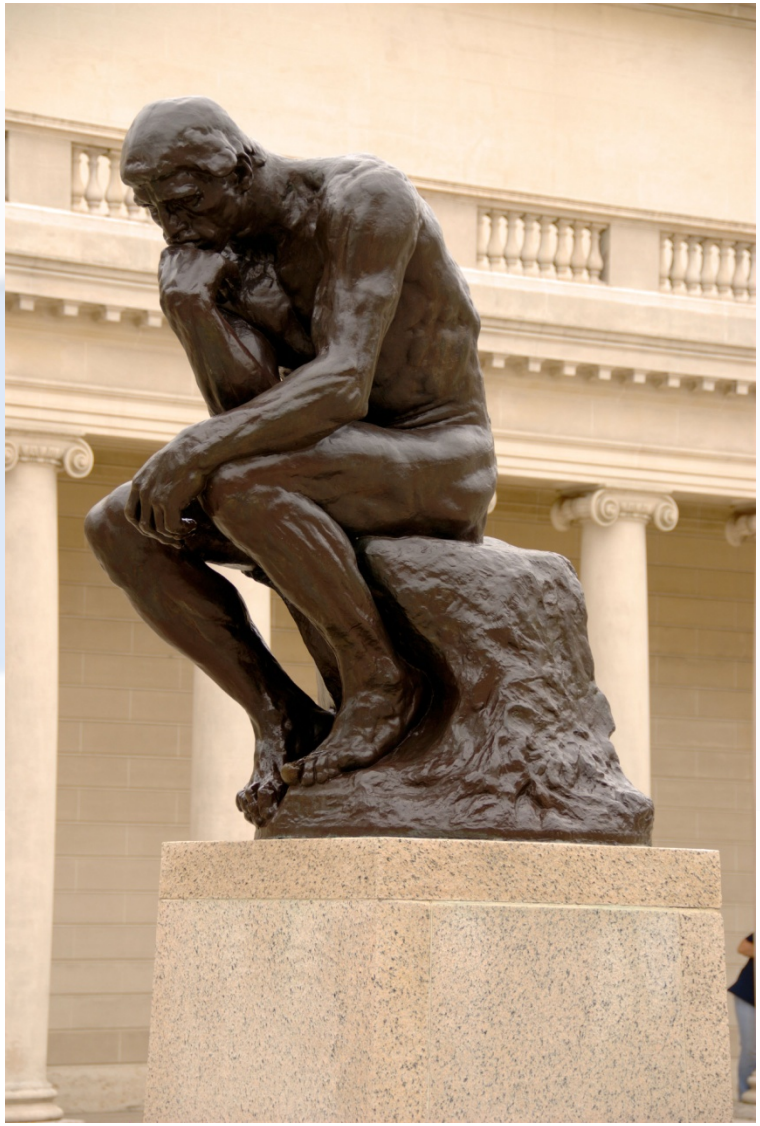
# DiaGrid?

- A large, high-throughput, distributed computing system
- Using Condor to manage jobs and resources
- Purdue leading a partnership of 10 campuses and institutions
  - University of Wisconsin, Notre Dame and Indiana University to name a few
- Including all Purdue (and other campus) clusters, lab computers, department computers, desktop, totaling 60,000+ cores



ITaP INFORMATION TECHNOLOGY AT PURDUE

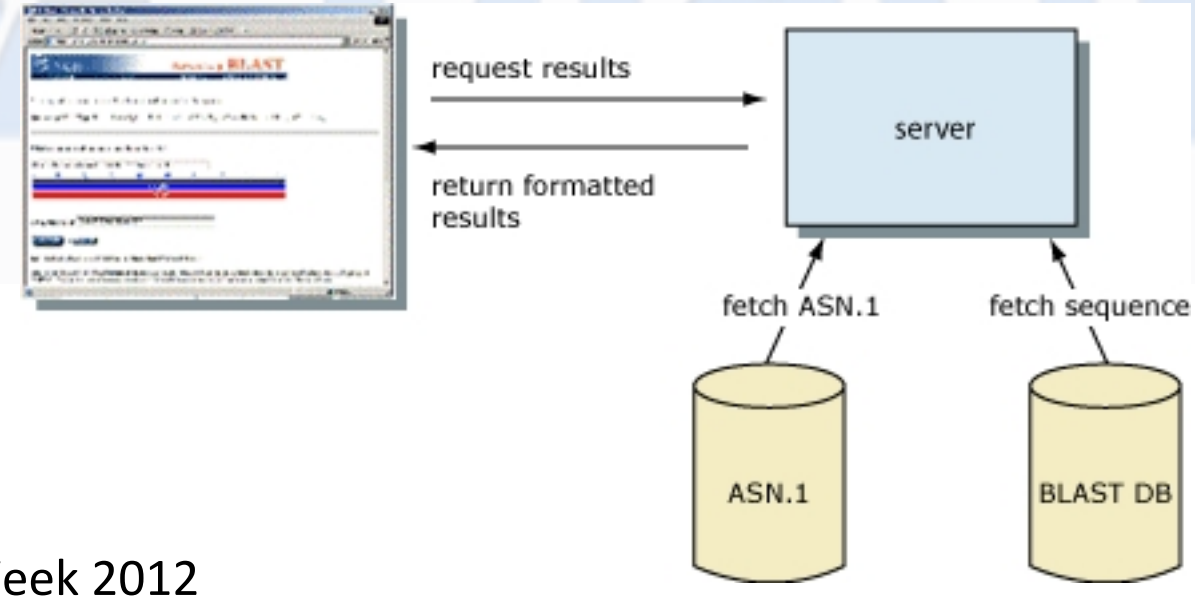
# Ok, cool… Now what?



# Basic Local Alignment Search Tool

- Comparing nucleotide or protein sequences
  - String and Substring pattern matching

- National Center for Biotechnology Information (NCBI)



PURDUE UNIVERSITY

# Why remake something?

- Input file size limitations (5MB, 10MB, etc.)
- # of sequences for comparison
- Timeliness
- Ease of use



ITaP INFORMATION TECHNOLOGY AT PURDUE

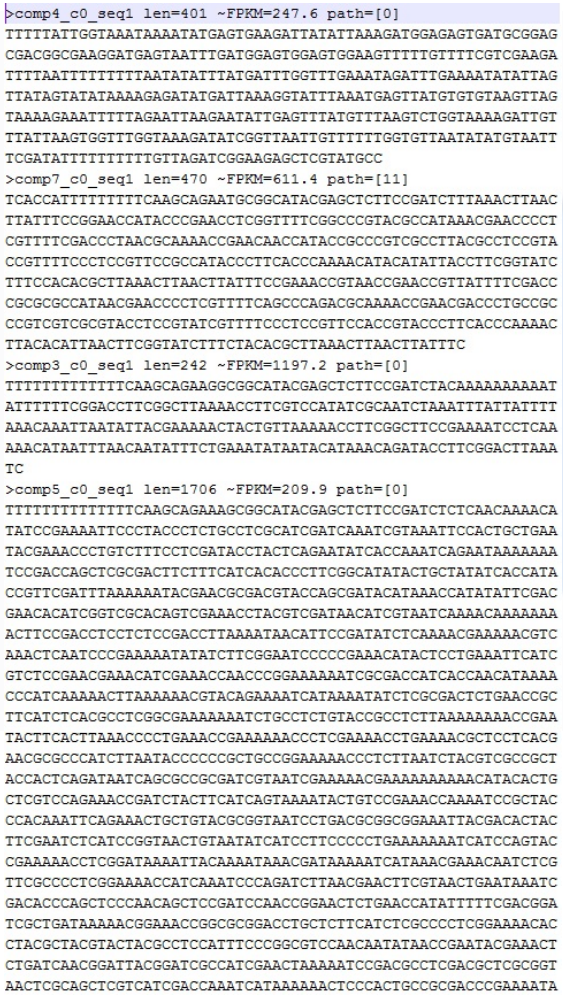
# BLAST and DiaGrid

- BLAST is highly parallelizable
  - No one sequence result depends on another (GREAT!!!)
  - Split input file with trusty friend AWK
  - Build a Condor DAG to maintain all jobs
    - Never more than 1500 individual jobs

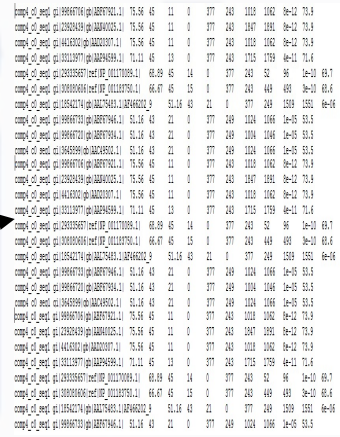
PURDUE UNIVERSITY

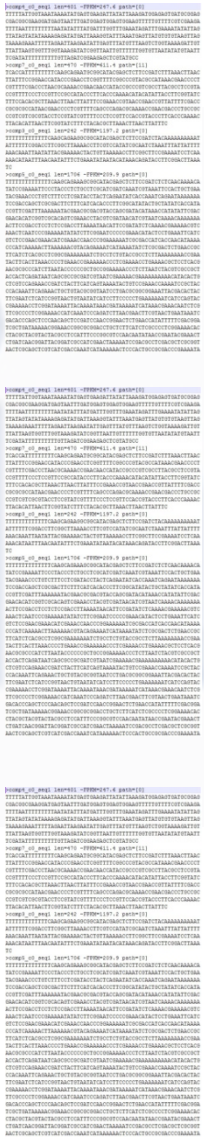
# BLAST and DiaGrid

## Input File




## Results





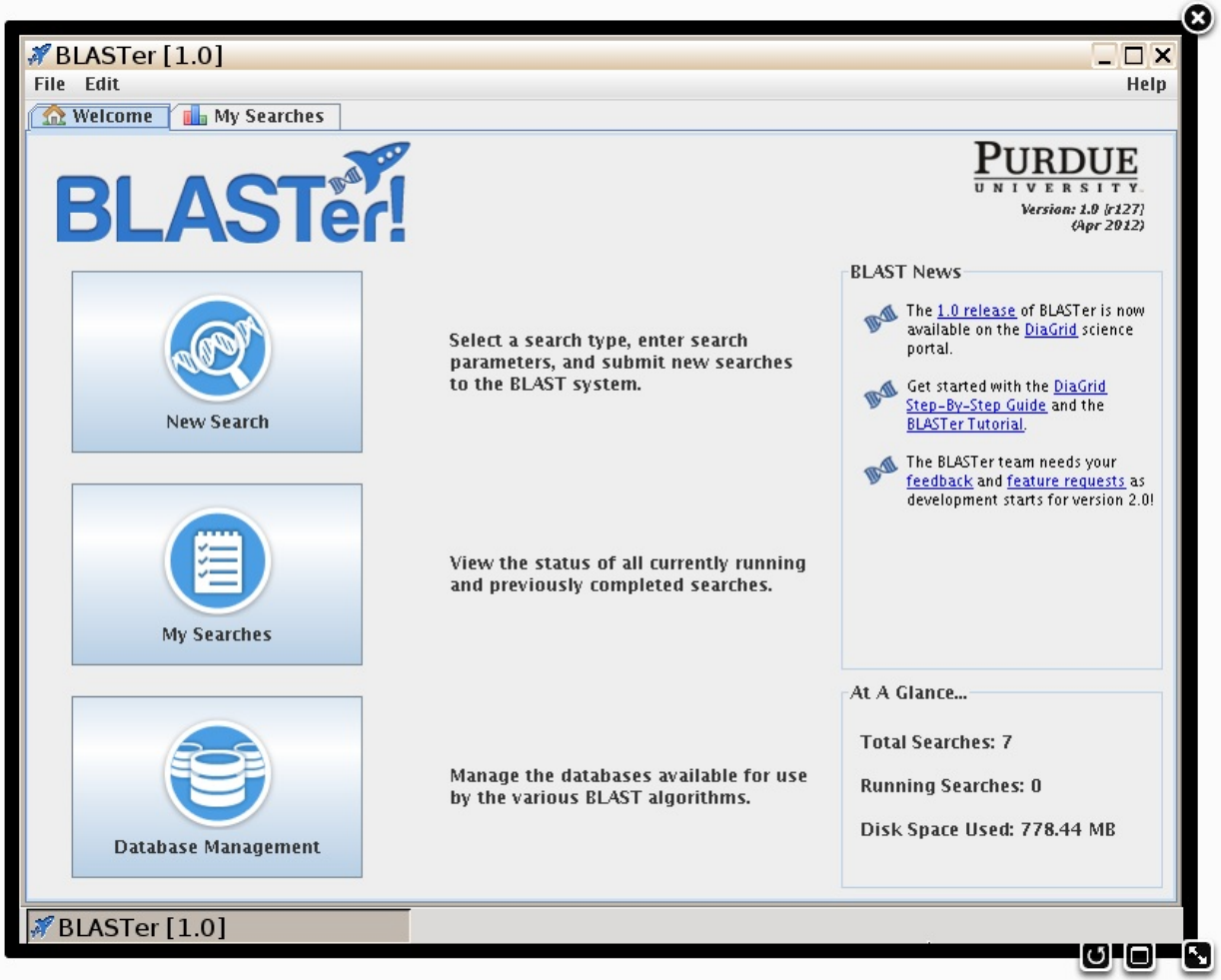
BLAST DB

Condor Week 2012

ITaP
INFORMATION TECHNOLOGY AT PURDUE

PURDUE UNIVERSITY

# BLASTer

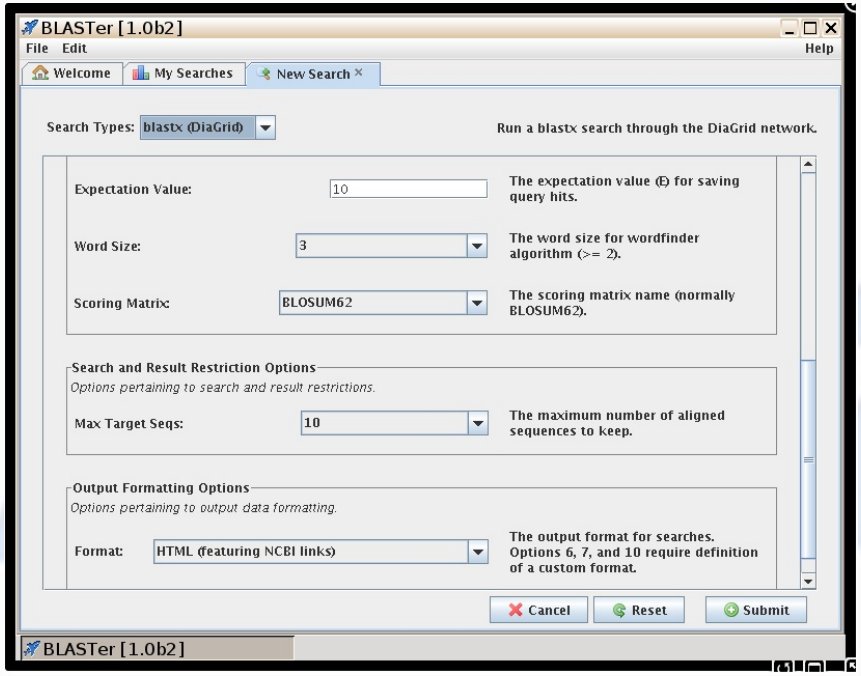


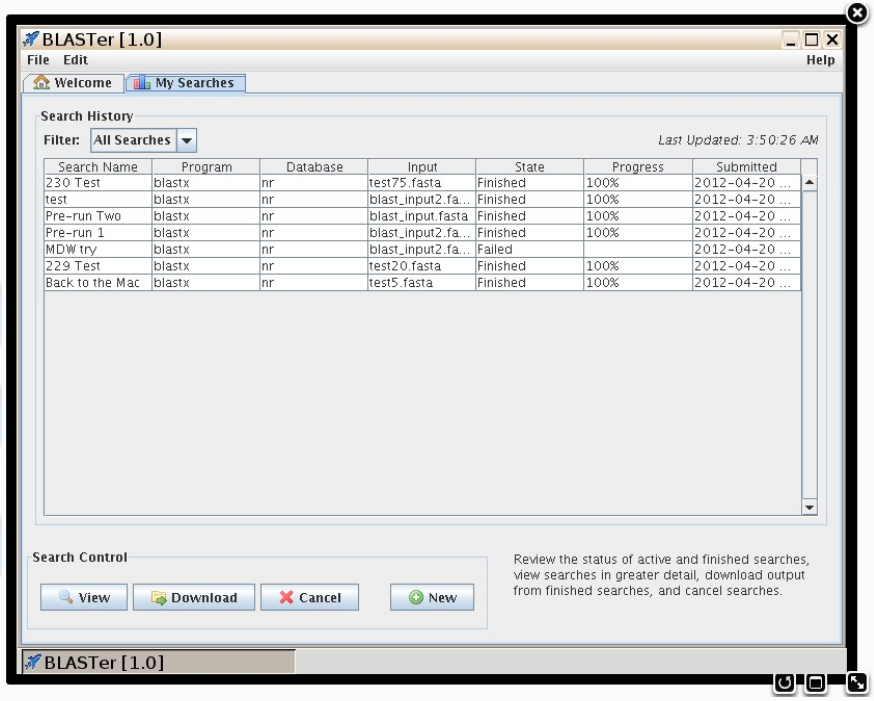



ITaP INFORMATION TECHNOLOGY AT PURDUE

PURDUE UNIVERSITY

# BLASTer



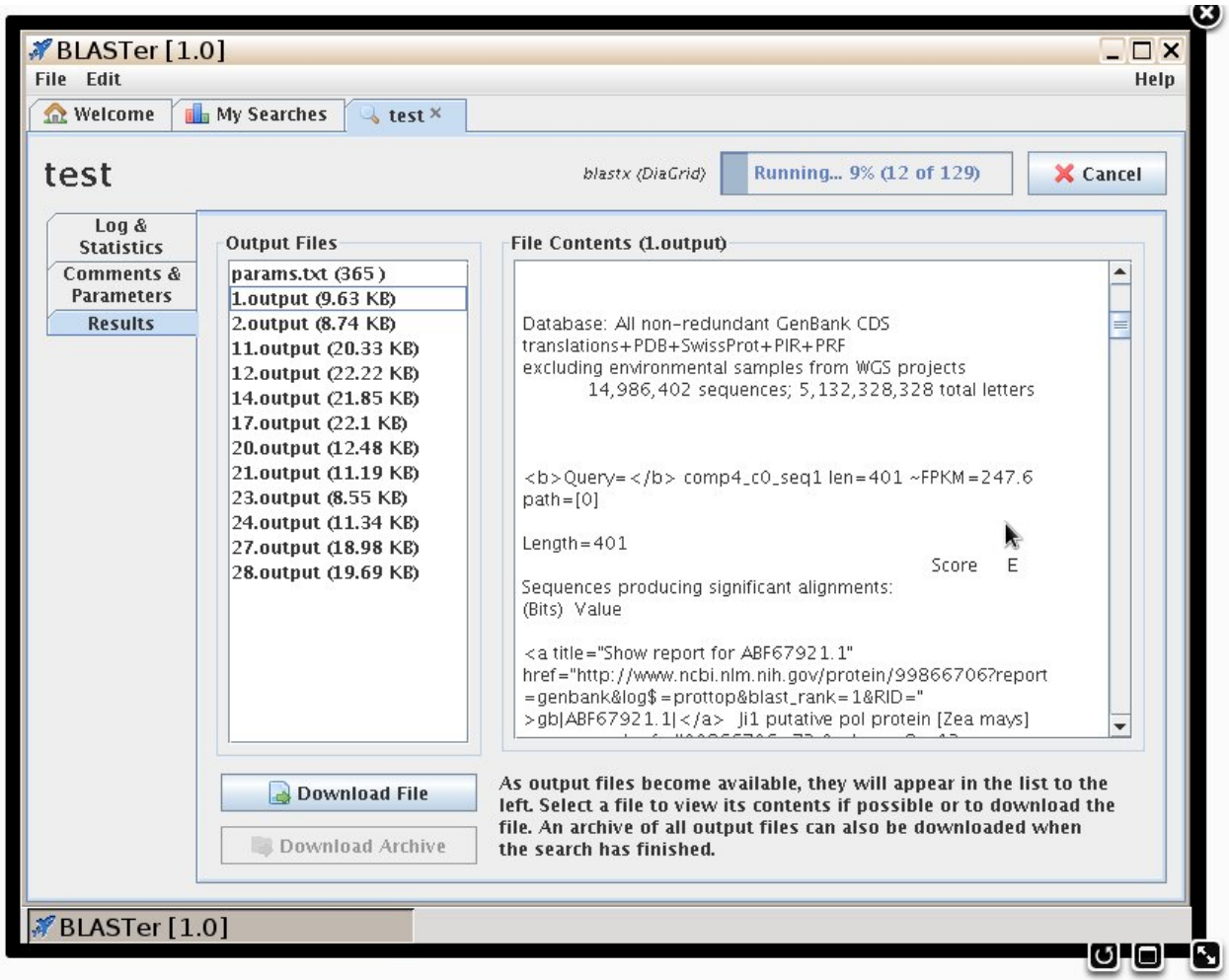






ITaP INFORMATION TECHNOLOGY AT PURDUE

PURDUE UNIVERSITY

# BLASTer




ITaP INFORMATION TECHNOLOGY AT PURDUE

# Big Benefits? We think so!

- Rick Westerman
  - Bioinformatics Specialist at the Purdue University Genomics Facility

| Sample | Sequences | Bases | N50 length | Cluster wall time | Condor wall time |
|---|---|---|---|---|---|
| #1 | 43 K | 16 M | 439 | 5:20 | 2:40 |
| #2 | 40 K | 15 M | 412 | 5:30 | 2:30 |
| #3 | 105 K | 49 M | 479 | 7:15 | 3:30 |
| #4 | 145 K | 72 M | 509 | 8:20 | 3:20 |

# Development Hurdles

- DiaGrid disk quota per user
  - Default 1GB -> NOT ENOUGH SPACE!!!
- Condor job failure
  - Set retry flag (We use 20 to be safe)
- Need more features!

PURDUE UNIVERSITY

# To the Future!




ITaP
INFORMATION TECHNOLOGY AT PURDUE

# BLASTer Plans

- Custom Databases
  - Nearly all researchers want this feature
  - Concern: Database permissions
- More output viewing options
  - Integrated HTML viewer
  - Blast2Go
- Better file management

# DiaGrid Plans

- R (programming language) statistical computing
  - Landscape Ecology & Biodiversity Department
- Cryo-Electron Microscopy Tools (Cryo-EM)
  - Single particle reconstruction (EMAN2 and similar tools)
  - Department of Biological Sciences

PURDUE UNIVERSITY

# Questions?



ITaP INFORMATION TECHNOLOGY AT PURDUE