PURDUE UNIVERSITY

ITaP INFORMATION TECHNOLOGY AT PURDUE

DiaGrid + hubzero

Power to the Masses

Carol Song

carolxsong@purdue.edu

Hubbub 2013

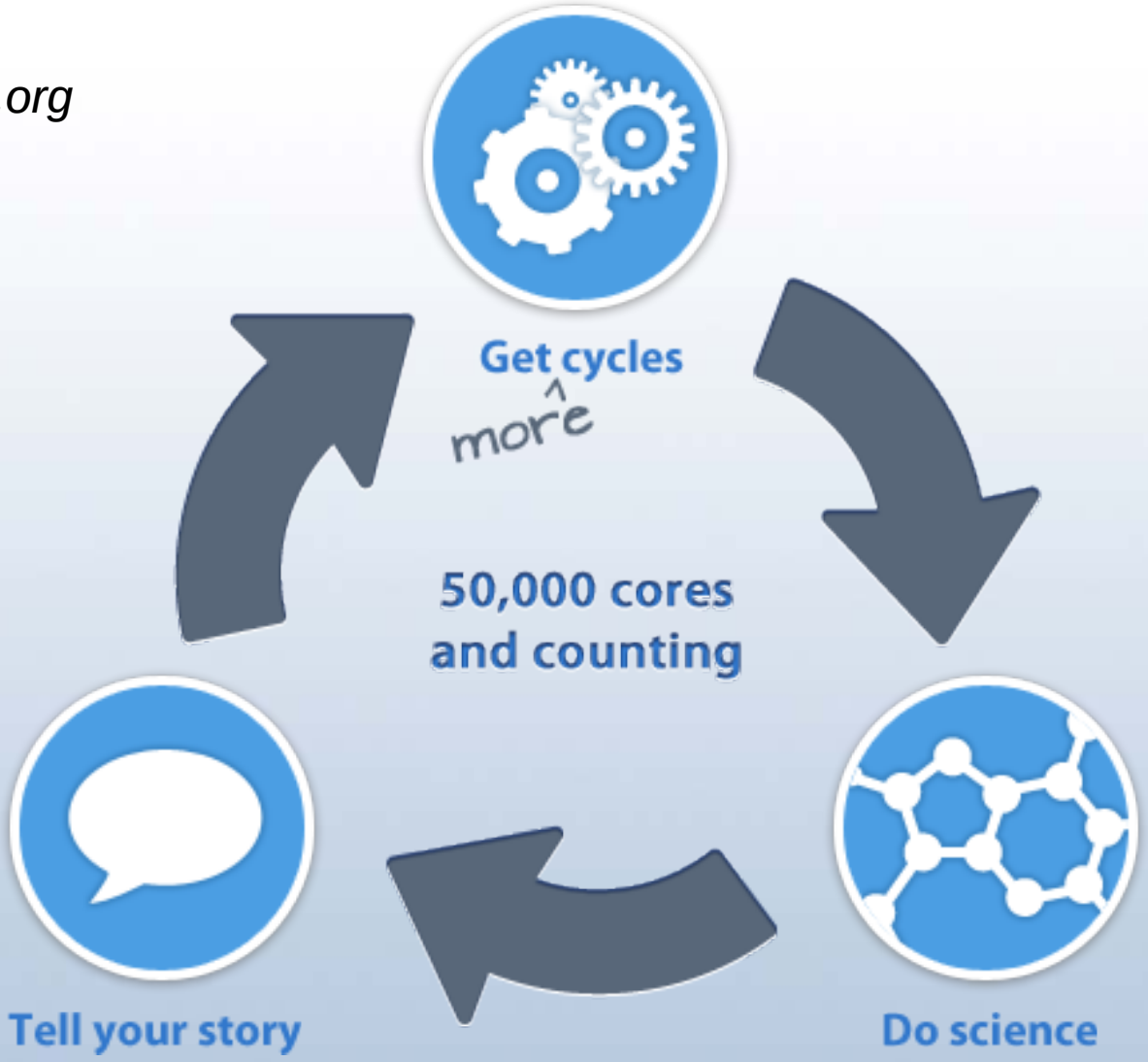
September 5, 2013

Contributors

- Rob Campbell, developer
- Kevin (Feng) Chen, developer
- Brian Raub, developer
- Chris Thompson, developer
- Steve Clark, HUBzero application dev
- Ben Cotton, project coordination, docs
- HUBzero team

What is DiaGrid?

Diagrid.org


Get cycles
more
50,000 cores
and counting
Do science
Tell your story

To users, DiaGrid is…..

Tools for science, easy to use, instant access, technical support, opportunity to help improve tools, ….

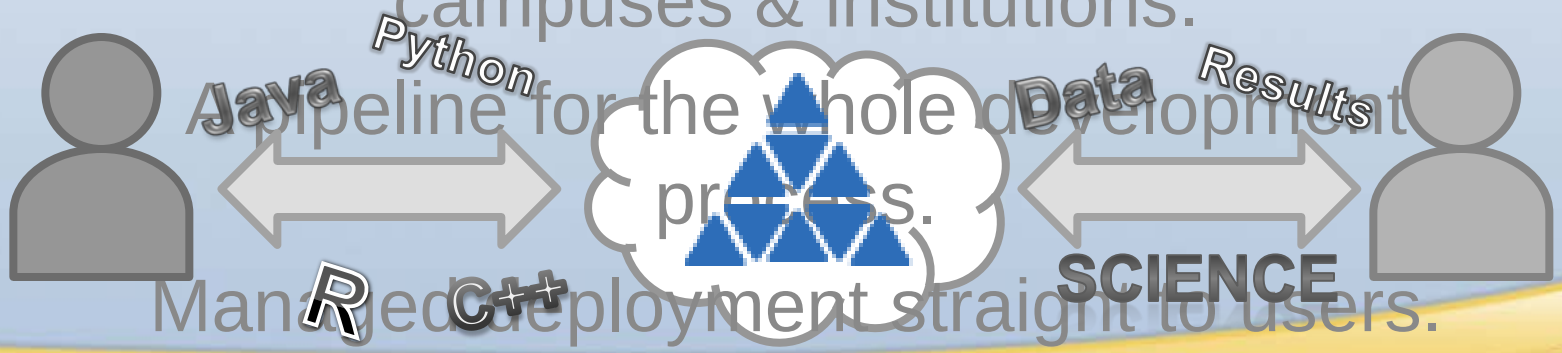
A hub for collaboration and community building

To app developers, DiaGrid is …

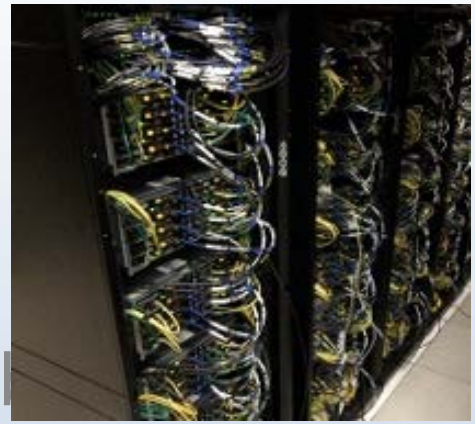
A federation of 50,000+ cores from computing resources across multiple campuses & institutions.

A pipeline for the whole development process.

Managed deployment straight to users.



Hardware



Large high-throughput and distri
network of *50,000*+ cores, available
through **HT Condor**.

Utilizes spare cycles from:

- Community clusters at Purdue
  - Steele, Coates, Rossmann, Hansen, & Carter
- Campus lab workstations

PURDUE UNIVERSITY

ITaP INFORMATION TECHNOLOGY AT PURDUE

The web site: *diagrid.org*

DiaGrid

101,860,020 jobs run to date

No Forms. No waiting. Just instant access to high-throughput computing

Announcements

CryoEM Updated

The cryogenic electron microscopy tool (CryoEM) has been updated. The latest version of CryoEM adds support for parallel processing, improving performance of some workflow steps. Additional performance improvements are planned for the next release.

Tools

We support BLAST, R scripts, CESM, SWATShare and other programs used by thousands of researchers. Find a tool, click the launch button, and start computing. Visit our DIY area to use your own tools on DiaGrid.

Incentives

Earn your way to VIP Status and receive more cycles and higher priority. Tell us about what you're doing and earn more cycles. Get involved by asking/answering questions in the community or suggest improvements.

Researcher Stories

Michael Delgado

Assistant Professor of Agricultural Economics
Purdue University

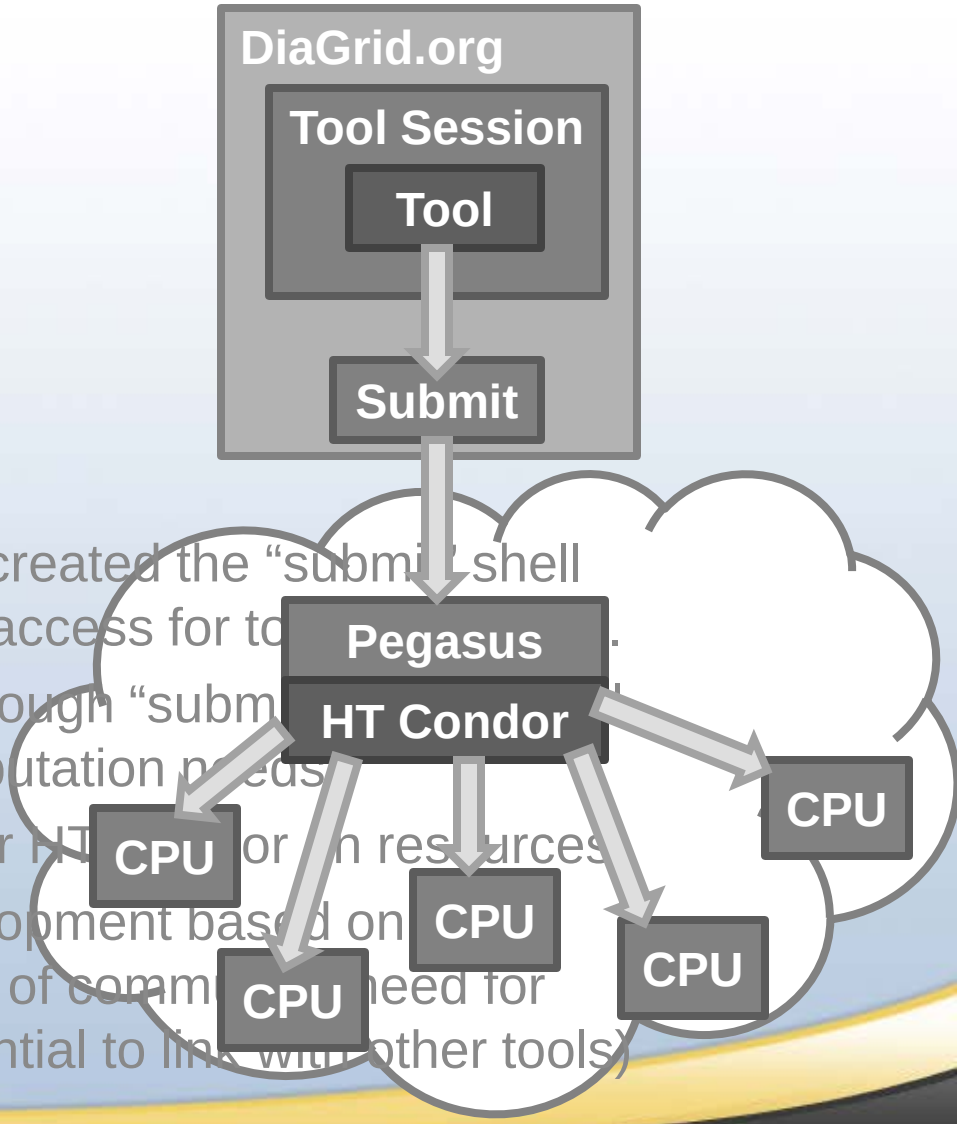
Michael Delgado uses DiaGrid and its SubmitR tool to help look at questions like how good voluntary pollution abatement programs actually are at reducing pollution or what the real impact is of

Bruce Hamaker & Osvaldo Campanella

Professor of Food Science & Professor of Agricultural and Biological Engineering
Purdue University

Bruce Hamaker and Osvaldo Campanella are exploring ideas for using natural starch and protein molecules as delivery vehicles with potential

Supporting Science



The HUBzero team has created the "submit" shell command to abstract grid access for tools.

Tools run a subprocess through "submit" to their grid computation needs.

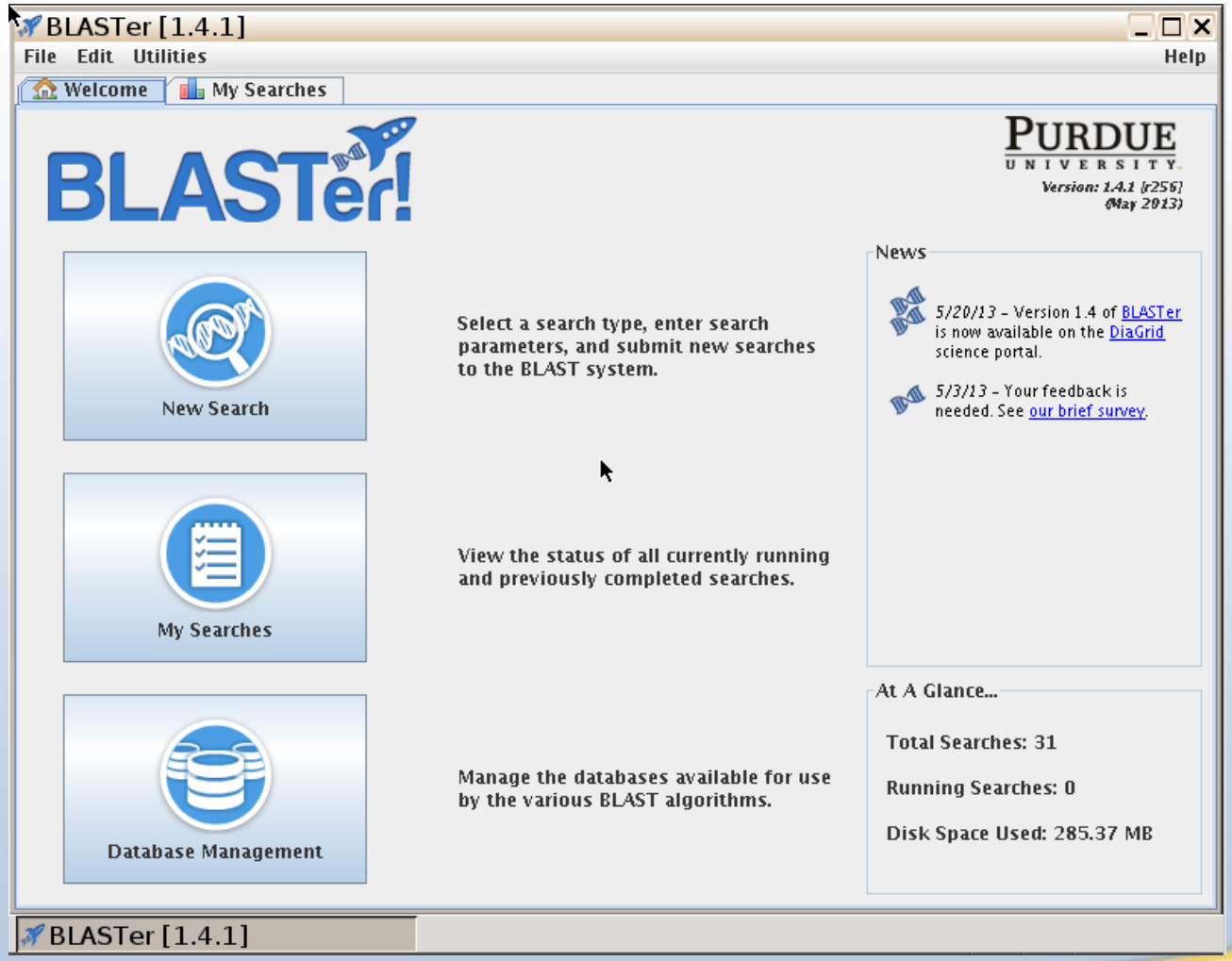
Utilizes Pegasus engine for HTC or campus resources.

Selects apps for development based on community needs (size of community, need for computing resources, potential to link with other tools)

PURDUE UNIVERSITY

ITaP INFORMATION TECHNOLOGY AT PURDUE

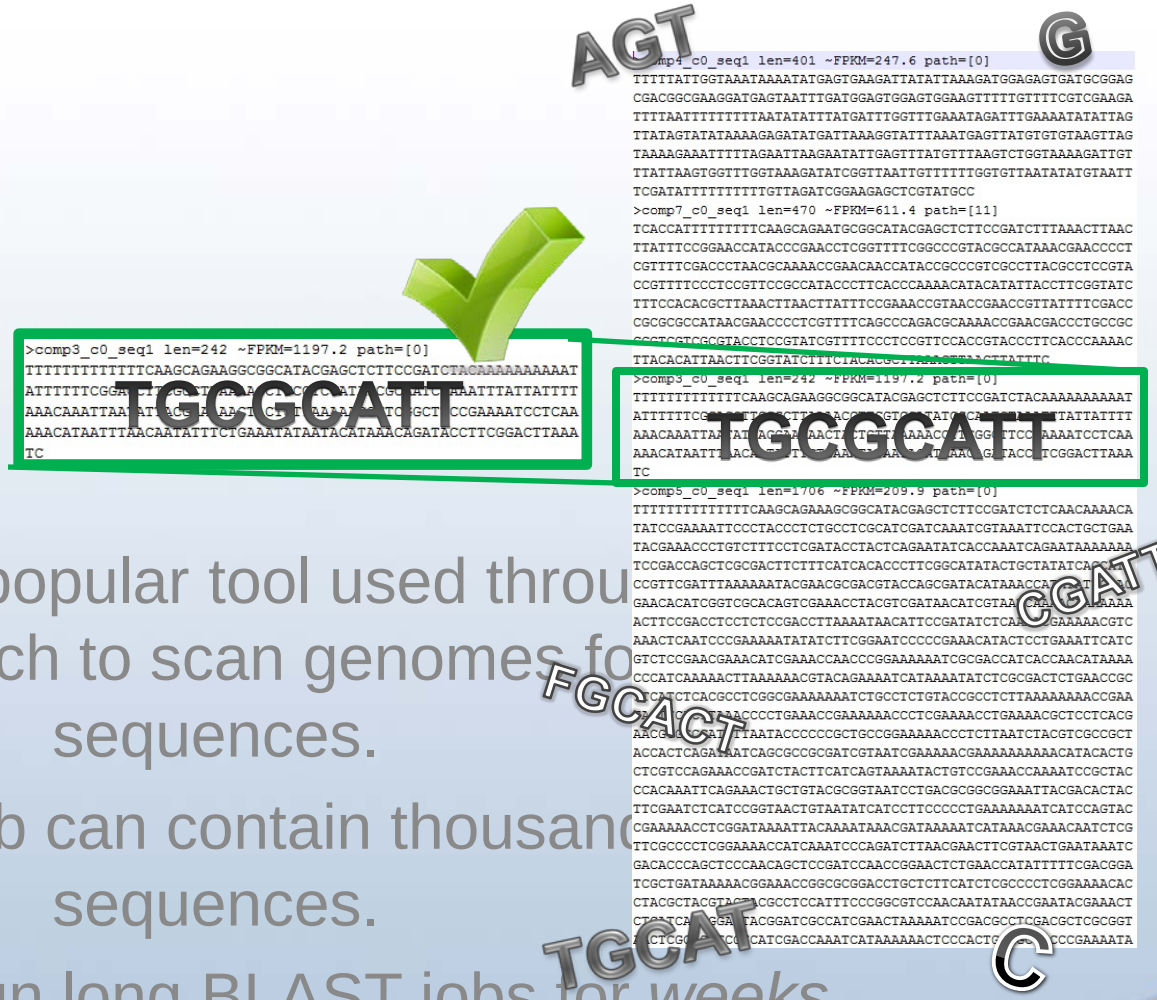
BLASTer



PURDUE UNIVERSITY

ITaP INFORMATION TECHNOLOGY AT PURDUE

BLASTer



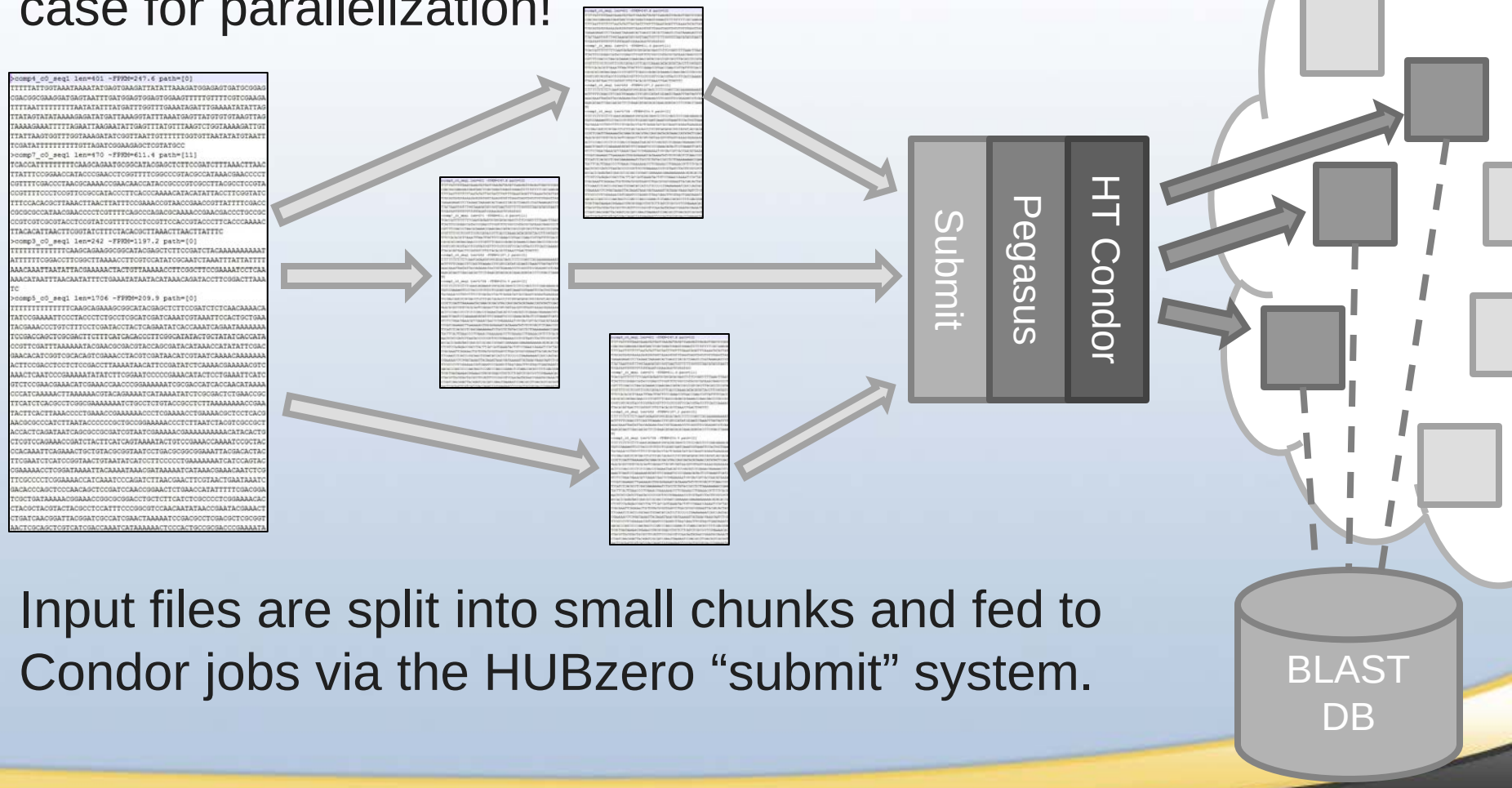
BLAST is a popular tool used throu
biology research to scan genomes fo
sequences.

A search job can contain thousand
sequences.

Many users run long BLAST jobs for *weeks*
on desktop workstations in their labs…

BLASTer

- Each sequence is independent, making a great case for parallelization!



- Input files are split into small chunks and fed to Condor jobs via the HUBzero “submit” system.

PURDUE UNIVERSITY

ITaP INFORMATION TECHNOLOGY AT PURDUE

Solving problems for users

- Speed up the searches
- Use custom databases for searches
- Manage data transfer
- Track search history
- Regular BLAST database update
- BLAST code update
- Post processing, link to other tools (BLAST2GO)
- Manage storage
- Share databases

In the past 12 months, BLASTer

- Completed 1.4 million search jobs (equivalent to searches of tens of millions of sequences against public and custom databases)
- Consumed 800K CPU hours (HT Condor)
- 111 researchers used Blaster
- Most of them are from domains that traditionally use desktops for computation.

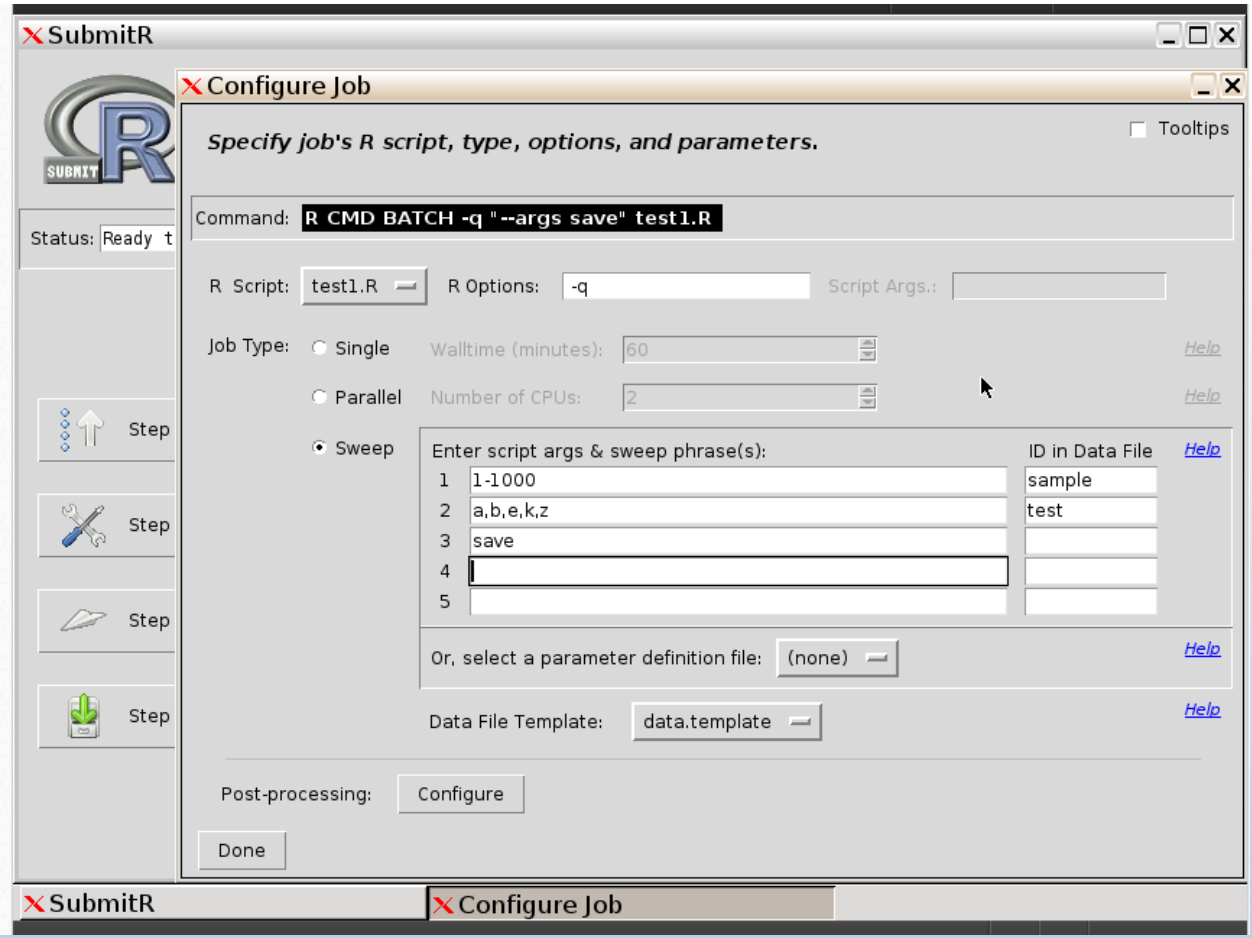
J. Andrew DeWoody, Nick Marra, Forestry & Natural Resources

- Using Blaster to annotate assembly of gene sequences (50,534 contigs) from E51K Illumina in study of gene evolution
- 8 days in the lab → less than 3 hours on DiaGrid

PURDUE UNIVERSITY

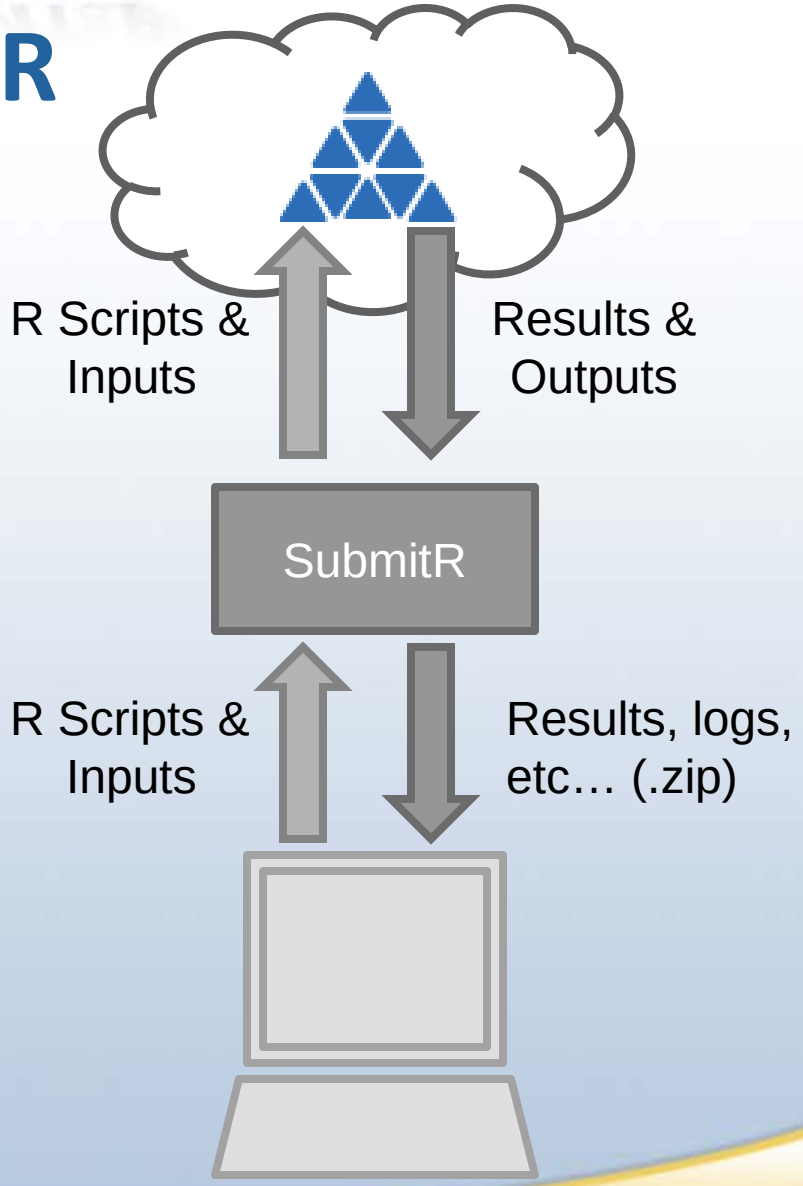
ITaP INFORMATION TECHNOLOGY AT PURDUE

SubmitR



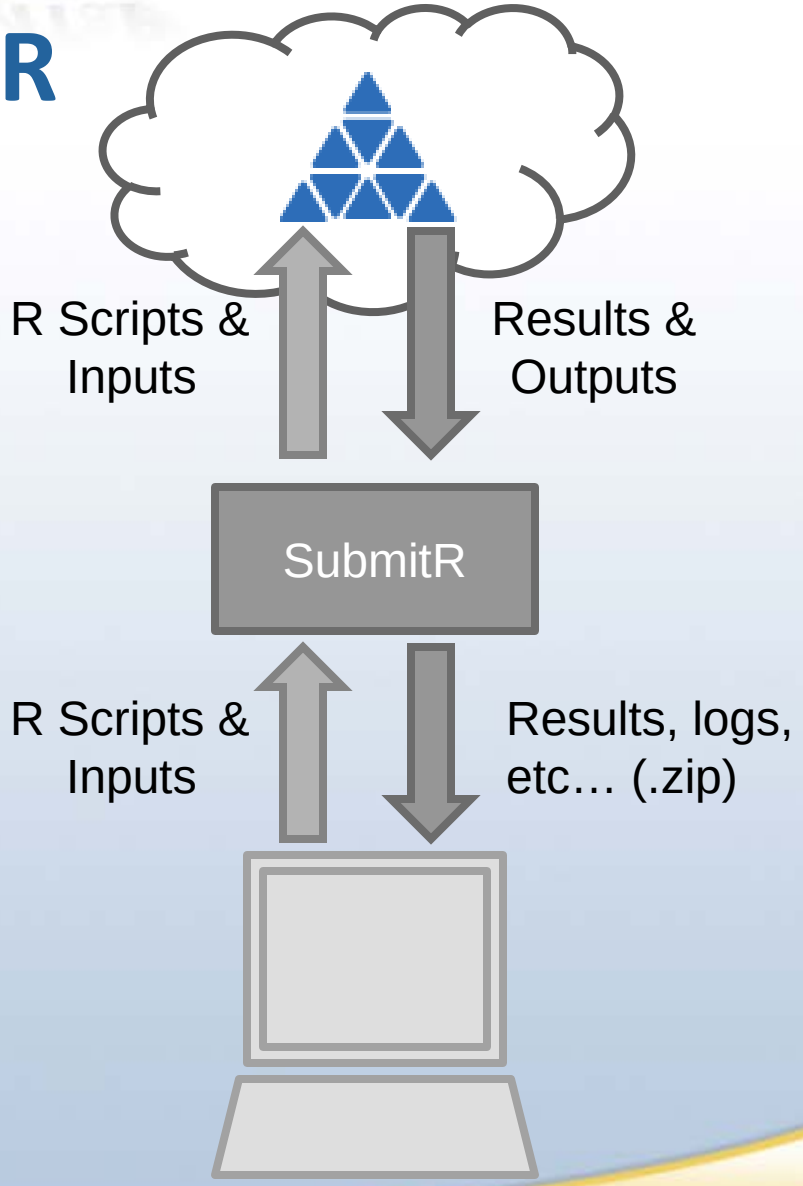
SubmitR

- Users create scripts to run their simulations all the time.
- A demand exists to run these jobs on the grid.
- SubmitR solves this issue for the R language on DiaGrid.



SubmitR

- SubmitR supports a wide range of R scripts:
  - **Single**: one process
  - **Parallel**: multiple processes communicating with each other
  - **Sweep**: many isolated processes with different parameters, inputs, or both



SubmitR

- SubmitR already supports a wide range of R libraries:

ElectroGraph	cubature	mvtnorm	rpart	survival
GWASExactHW	datasets	ncf	snow	tcltk
KernSmooth	deldir	nlme	snowfall	tools
MASS	foreign	nnet	sp	utils
Matrix	grDevices	np	spatial	…
PBSmapping	graphics	parallel	spatstat	
base	grid	plotrix	splancs	
boot	igraph	plyr	splines	
class	lattice	qtl	stats	
cluster	maptools	raster	stats4	
codetools	methods	rgdal	stpp	
compiler	mgcv	rgeos	stringr	

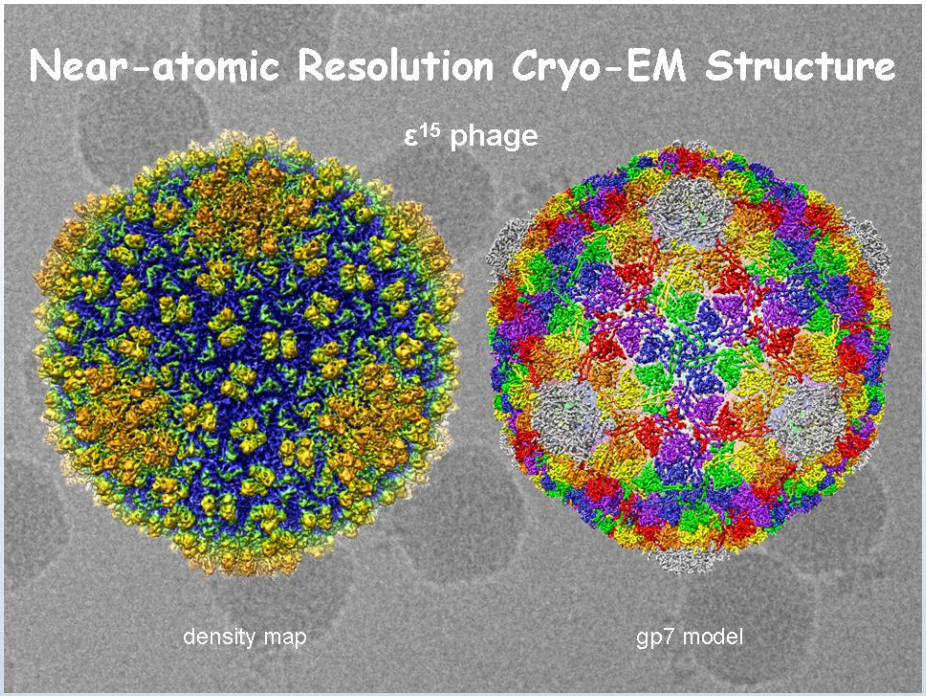
- And through the DiaGrid community features users can request more!

SubmitR usage examples

- Nutrition: (single, long running jobs)
  - Ingestive behavior research
- Bioinformatics: (single, long running jobs)
  - Genome association and prediction
- Agricultural Economics: (single and parallel jobs)
  - Distributed hydrological modeling
  - Effects of education on growth rates in developing countries
  - Consumer demand for hybrid cars
- In past 12 months, ~7550 simulation runs, 45 users. Together with workspace, nearly 3M hours consumed by R codes.

CryoEM

- The analysis of images taken at cryogenic temperatures within an electron microscope can reveal much about the structure of microscopic objects.
- Image processing is a good candidate for parallelization.

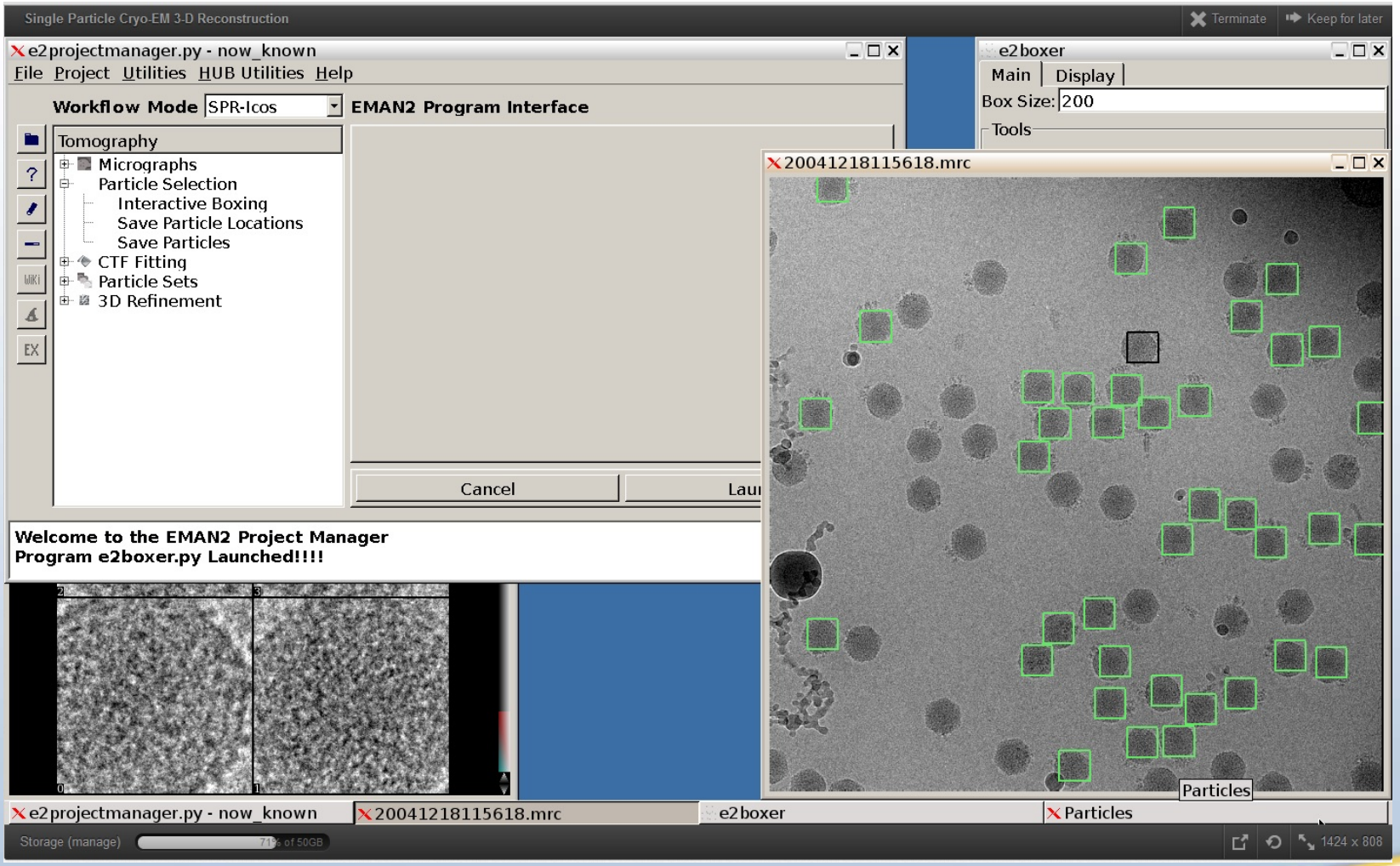


- The first user developed tool for the DiaGrid portal.
- DiaGrid staff utilized helping CryoEM authors split tasks for HT Condor then recombine with MPI for 3D visualization.

PURDUE UNIVERSITY

ITaP INFORMATION TECHNOLOGY AT PURDUE

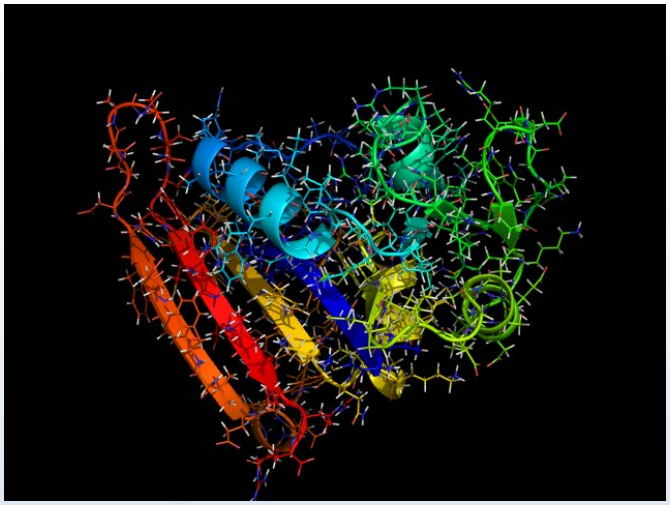
CryoEM



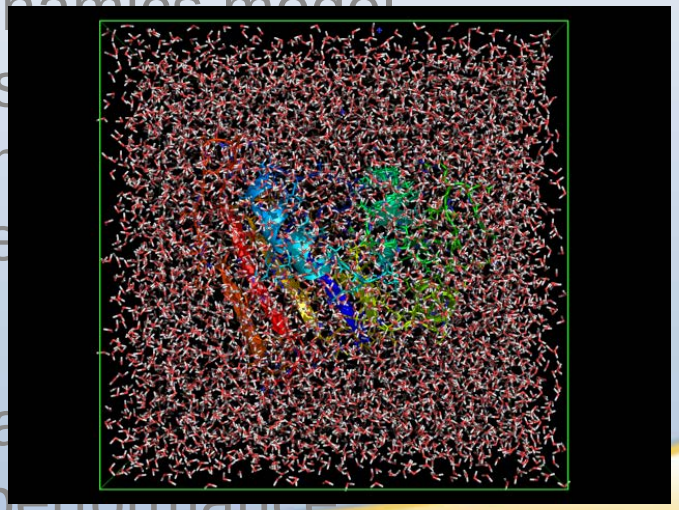
PURDUE UNIVERSITY

ITaP INFORMATION TECHNOLOGY AT PURDUE

GROMACSIMUM



GROMACS is a molecular dy
with a large community of us
scientific disciplines from cher
medicine, physics, e

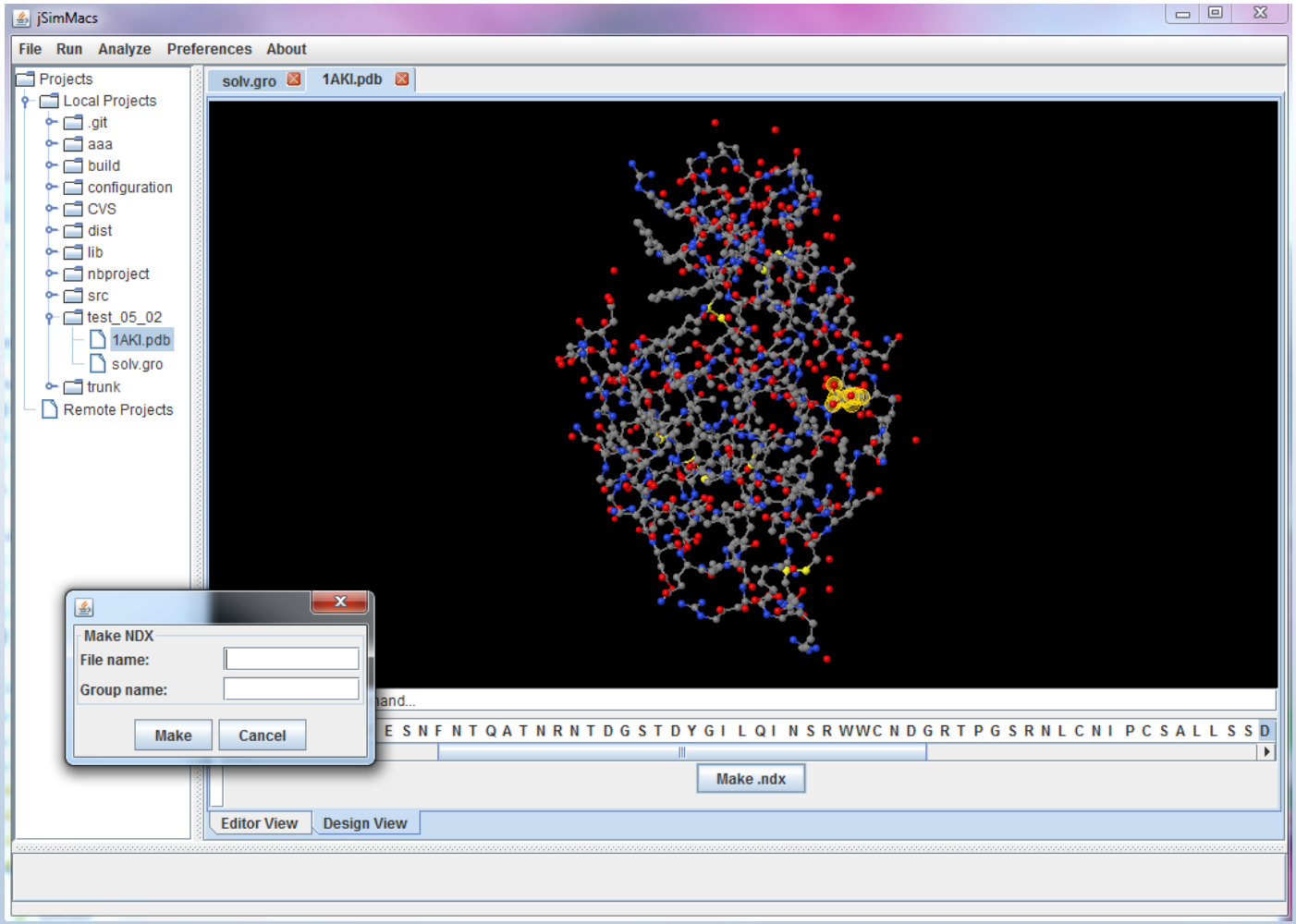


This project takes a popular
GROMACS GUI, jSimMacs, a
with new features for high-p
computing

PURDUE UNIVERSITY

ITaP INFORMATION TECHNOLOGY AT PURDUE

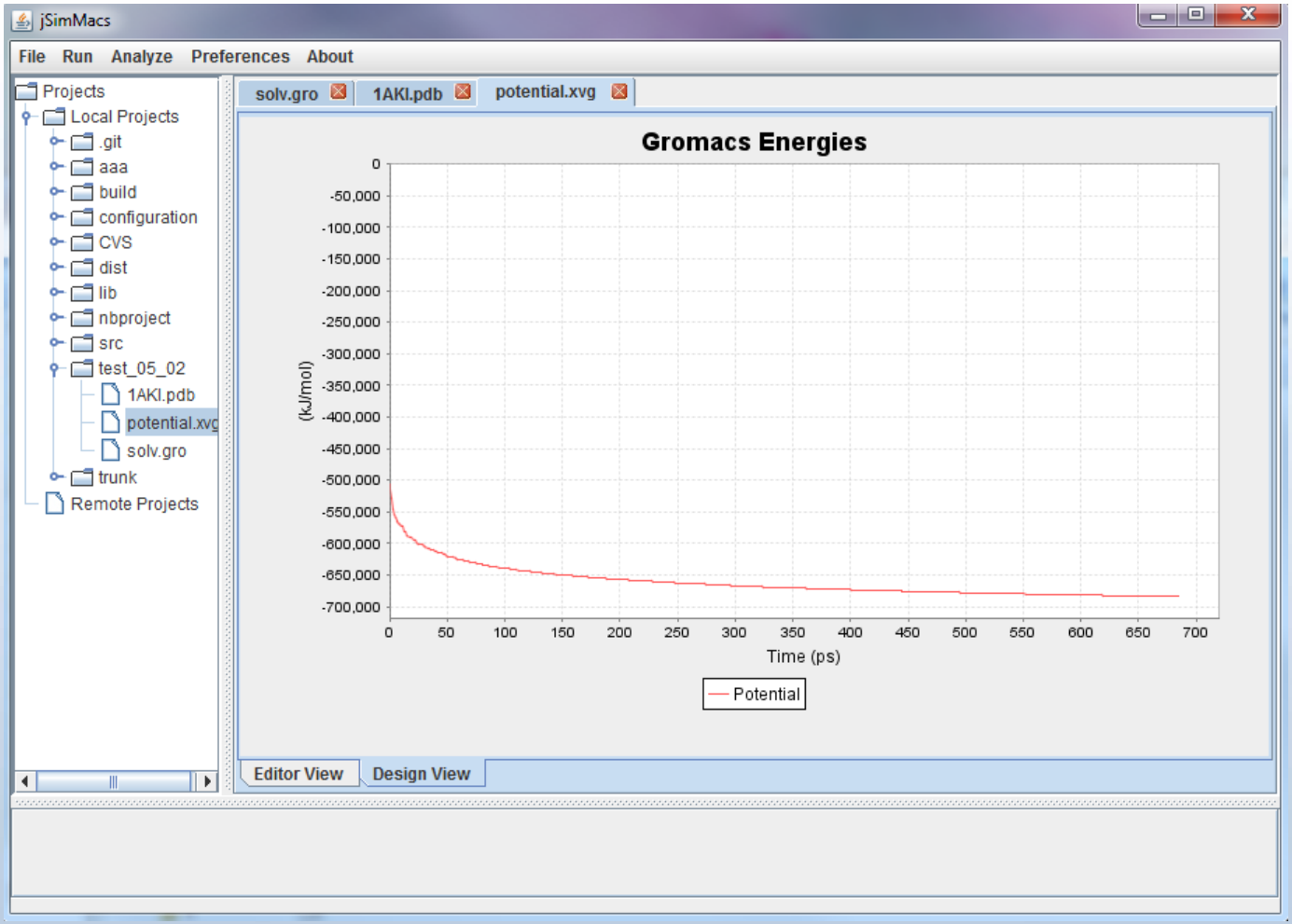
GROMACSIMUM



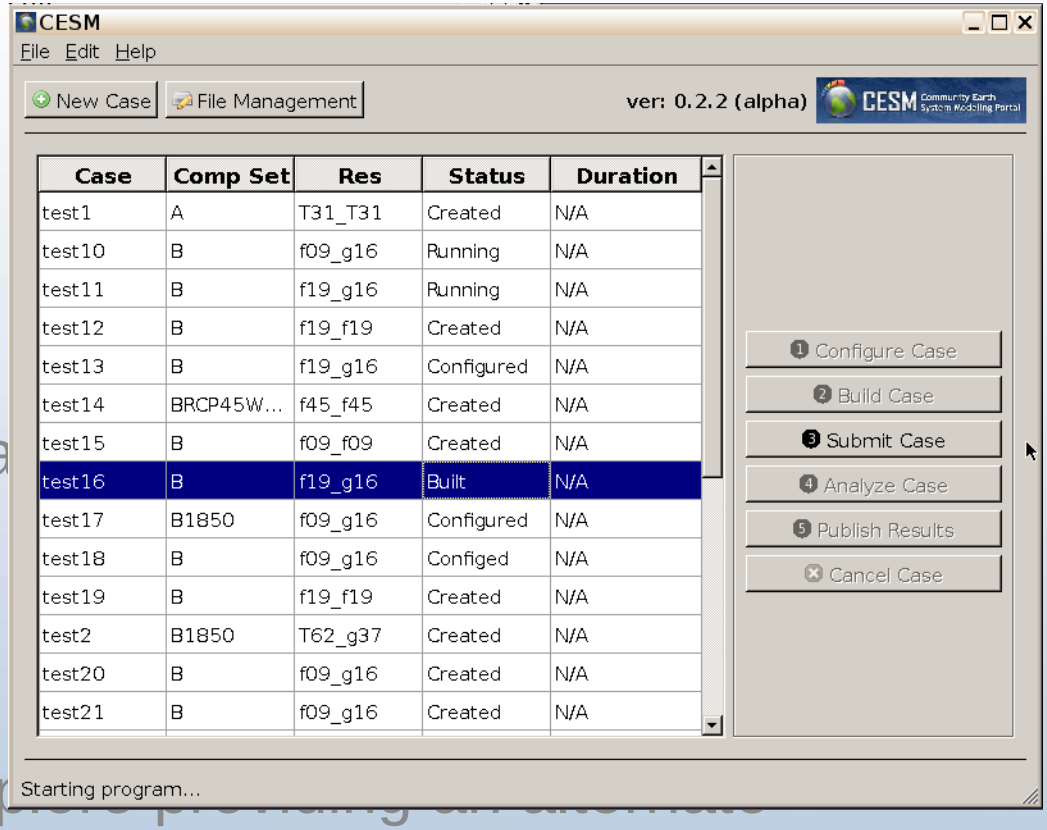
PURDUE UNIVERSITY

ITaP INFORMATION TECHNOLOGY AT PURDUE

GROMACSIMUM



CESM



CESM is a globa
many aspects of

Purdue developed
designed it to su

This project will exp
interface to the CESM gateway services from inside DiaGrid.

More apps!

More Applications that are:

for research or instruction

Requires high performance and/or high throughput computing

Solves workflow or ease-of-use problems

Tied to a computational resource or sufficiently portable as to be resource-agnostic

Not encumbered by license or patent restrictions