



Power to the Masses

Carol Song

carolxsong@purdue.edu

Hubbub 2013

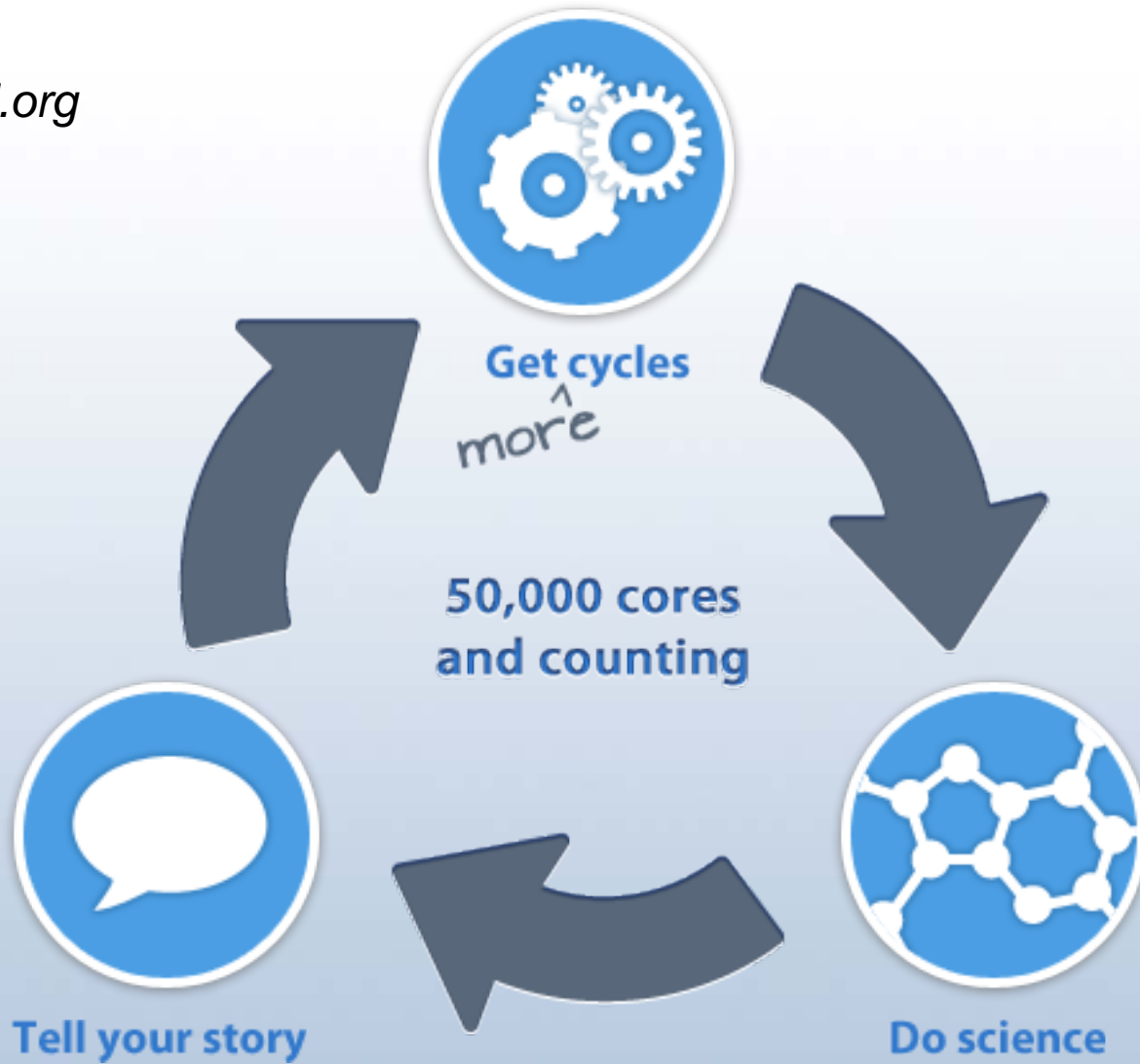
September 5, 2013

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



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.



Hardware



Large high-throughput and distributed 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

The web site: *diagrid.org*



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, CESH, 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*

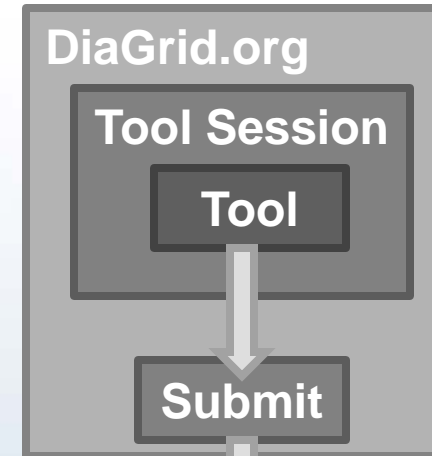
Michael Delgado uses DiaGrid and its SubmitR tool to help look at questions like how good voluntary pollution abatement programs actually are at reducing pollution or what the real impact is of



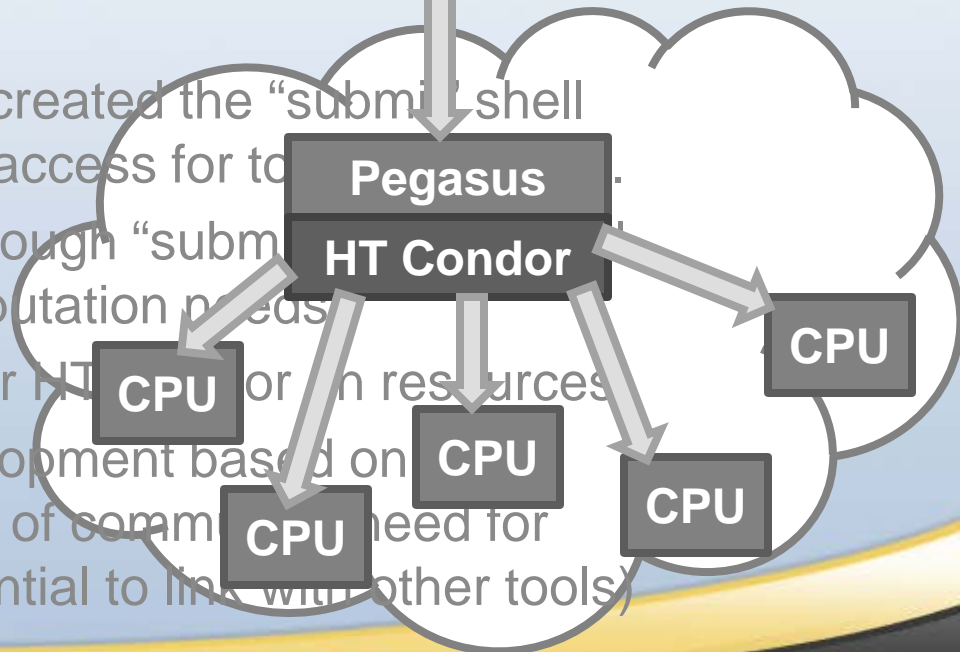
Bruce Hamaker & Osvaldo Campanella
*Professor of Food Science & Professor of Agricultural and Biological Engineering
Purdue University*

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

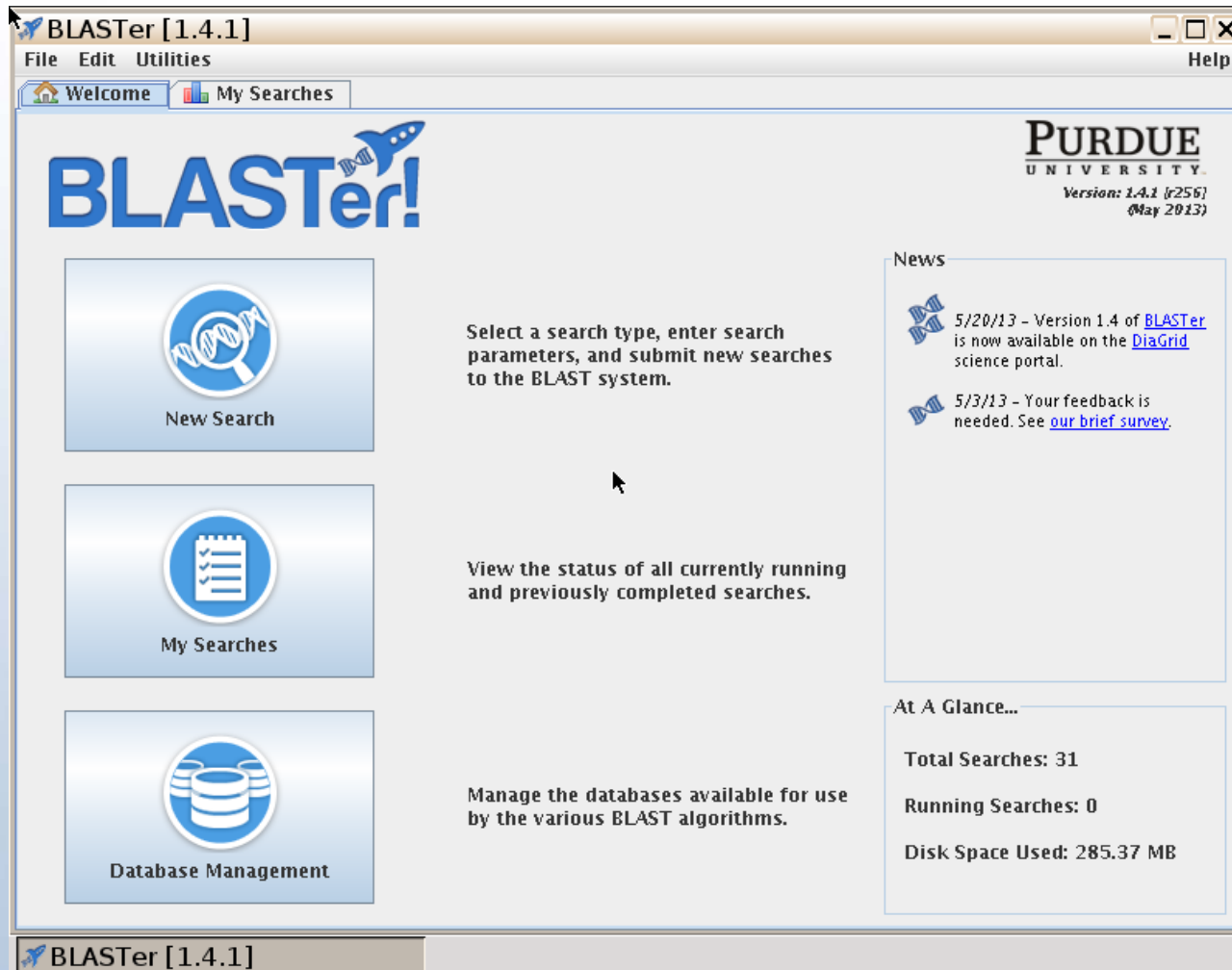
Supporting Science



The HUBzero team has created the "submit" shell command to abstract grid access for tools. Tools run a subprocess through "submit" to request their grid computation needs. Utilizes Pegasus engine for HT Condor or in resources. Selects apps for development based on community needs (size of community, need for computing resources, potential to link with other tools).



BLASTer



The screenshot shows the BLASTer web application interface within a browser window titled "BLASTer [1.4.1]". The interface includes a menu bar with "File", "Edit", "Utilities", and "Help". Below the menu bar are two tabs: "Welcome" and "My Searches". The main content area features the "BLASTer!" logo with a rocket icon. On the left side, there are three large blue buttons: "New Search" (with a magnifying glass icon), "My Searches" (with a checklist icon), and "Database Management" (with a database cylinder icon). The central area contains three instructions: "Select a search type, enter search parameters, and submit new searches to the BLAST system.", "View the status of all currently running and previously completed searches.", and "Manage the databases available for use by the various BLAST algorithms." On the right side, there is a "News" section with two entries: "5/20/13 - Version 1.4 of BLASTer is now available on the [DiaGrid](#) science portal." and "5/3/13 - Your feedback is needed. See [our brief survey](#)." Below the news is an "At A Glance..." section showing: "Total Searches: 31", "Running Searches: 0", and "Disk Space Used: 285.37 MB". The bottom status bar shows "BLASTer [1.4.1]".

BLASTer

```
>comp3_c0_seq1 len=242 ~FPKM=1197.2 path=[0]
TTTTTTTTTTTTTCAAGCAGAAGCGGCATACGAGCTCTCCGATCTCTGAAAAAAT
ATTTTTTCGGATTTGGATTTAGATTTTCTGATTTTCTGATTTTCTGATTTTCTG
AAACAATTAATTAATTAATTAATTAATTAATTAATTAATTAATTAATTAATTAAT
AAACATAATTTAACAATTTTCTGAAATATAATTAATTAATTAATTAATTAATTAAT
TC
```

TGGGCATT



```
comp4_c0_seq1 len=401 ~FPKM=247.6 path=[0]
TTTTTATGGTAATAAATATGAGTGAAGATTATTAAGATGGAGAGTATGGGGAG
CGACGGCGAAGGATGAGTAATTTGATGGAGTGGAGTGGAGTTTTTTGTTTCGTGAAGA
TTTTAATTTTTTTTTAATATAITATGATTTGGTTTGAATAGAITTGAAATATATTAG
TTATAGTATATAAAGAGATATGATTAAGGTAATTAATGAGTTATGTGTGAAGTTAG
TAAAGAAATTTTGAATTAAGATATTGAGTTTATGTTTAAAGTCTGGTAAAGATTGT
TTAATTAAGTGGTTGGTAAAGATATCGGTTAATGTTTTTTGGTGTAAATATATGTAAT
TCGATATTTTTTTTTGTTAGATCGGAAGAGCTCGTATGCC
>comp7_c0_seq1 len=470 ~FPKM=611.4 path=[11]
TCACCAATTTTTTTTCAAGCAGAATCGGCATACGAGCTCTTCGGATCTTTAAACTTAAC
TTAATTCGGGAACCATACCCGAACCTCGGTTTTTCGGCCGTACGCCATAAACGAAACCCCT
CGTTTTCGACCCATAACGCAAAACGCAACCATACGCCCGCTCGCCTTACGCTCCGTA
CGTTTTCCCTCCGTTCCGCATACCCCTTACCCAAACATACATATTACCTTCGATC
TTCCACACGCTTAACTTAACTTATTCGAAACCGTAACCCGAAACCGGTTATTTCCGAC
CGCGCGCATAACGAACCCCTCGTTTTTCAGCCAGACGCAAAACCGAAGCAGCCCTCGCGC
CGCTCGTCCGCTACCTCGTATCGTTTTCCCTCCGTTCCACCGTACCCCTTACCCAAAAC
TTACAGATTAACCTCGGTTATTTTCACACCGTTTTCGAAATTTTTCGAAATTTTTC
>comp5_c0_seq1 len=242 ~FPKM=1197.2 path=[0]
TTTTTTTTTTTTTCAAGCAGAAGCGGCATACGAGCTCTCCGATCTCTCAAAAAAAT
ATTTTTTCGGATTTGGATTTAGATTTTCTGATTTTCTGATTTTCTGATTTTCTG
AAACAATTAATTAATTAATTAATTAATTAATTAATTAATTAATTAATTAATTAAT
AAACATAATTTAACAATTTTCTGAAATATAATTAATTAATTAATTAATTAATTAAT
TC
```

TGGGCATT

```
>comp5_c0_seq1 len=1706 ~FPKM=209.9 path=[0]
TTTTTTTTTTTTTCAAGCAGAAGCGGCATACGAGCTCTTCGGATCTCTCAACAAAACA
TATCCGAAAATCCCTACCTCTGCTCGCATCGATCAAATCGTAATTCACCTGCTGAA
TAGGAACCCCTGTCTTCTCGATACCTACTCAGAATAACCAAAATCAGATAAAAAA
TCCGACCAGCTCGGACTTCTTTCATCACACCCCTTCGGCATATACTGCTATATCAATTA
CCGTTTCGATTTAAAAAATACGAACGGCAGTACGAGCATACATAAACCAATTAATTA
GAACACATCGGTCGACAGTCCGAACCTACGTCGATAACATCGTAAATTAATTAATTA
ACTTCGACCTCCTTCGACCTTAAATAACATTCGATATCTCAATTAATTAATTAATTA
AACTCAATCCGAAAAATATATCTTCGGAATCCCGGAAACATACCTCTGAAATTCATC
GTCTCCGAACCAACATCGAAACCAACCCCGAAAAAATCGGCACCATACCAACATAAAA
CCCATCAAAAATTAAAAAAGTACAGAAAATCATAAAATATCTCGGACTCTGAACCGC
CGATCTCAGCCTCGCGGAAAAAATCTGCTCTGTACCCGCTCTTAAAAAAAACCGAA
TTCGATCTCATCCGCTAATGTAATATCATCTTCCGCTGAAAAAATCATCCAGTAC
CGAAAAAATCCGGATAAATTAACAAATAAACGATAAAATCATAAACGAACAAATCTCG
TTCCGCTCGGAAACCATCAATCCGAGATCTTAACGAACCTCGTACTGAATAAATC
GACACCAGCTCCCAACAGCTCCGATCCAACCGGAACTCTGAACATATTTTTGACGGA
TCGCTGATAAAAAAGGAAACCGCGCGGAACTGCTCTTCACTCGCCCTCGGAAAAAC
CTACGCTACGCTACCGCTCCATTTCCCGCGCTCCAAACATATAACGAATACGAACAA
CTACTCAATTAACGATCCGATCGAACTAAAAATCCGAGCCTCGACGCTCGCGGT
CTCGGAAACCATCGACCAATCATAAAAAACTCCACTTTCGCGGAAAAATA
```

TGGCATT

BLAST is a popular tool used through biology research to scan genomes for sequences.

A search job can contain thousands of sequences.

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

AGT **G**

FGCAGT **CGATT**

TGGCATT **C**

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

SubmitR

SubmitR

Configure Job

Specify job's R script, type, options, and parameters. Tooltips

Command: `R CMD BATCH -q "--args save" test1.R`

R Script: R Options: Script Args.:

Job Type: Single Parallel Sweep

Walltime (minutes): [Help](#)

Number of CPUs: [Help](#)

Enter script args & sweep phrase(s):

	Enter script args & sweep phrase(s):	ID in Data File
1	1-1000	sample
2	a,b,e,k,z	test
3	save	
4		
5		

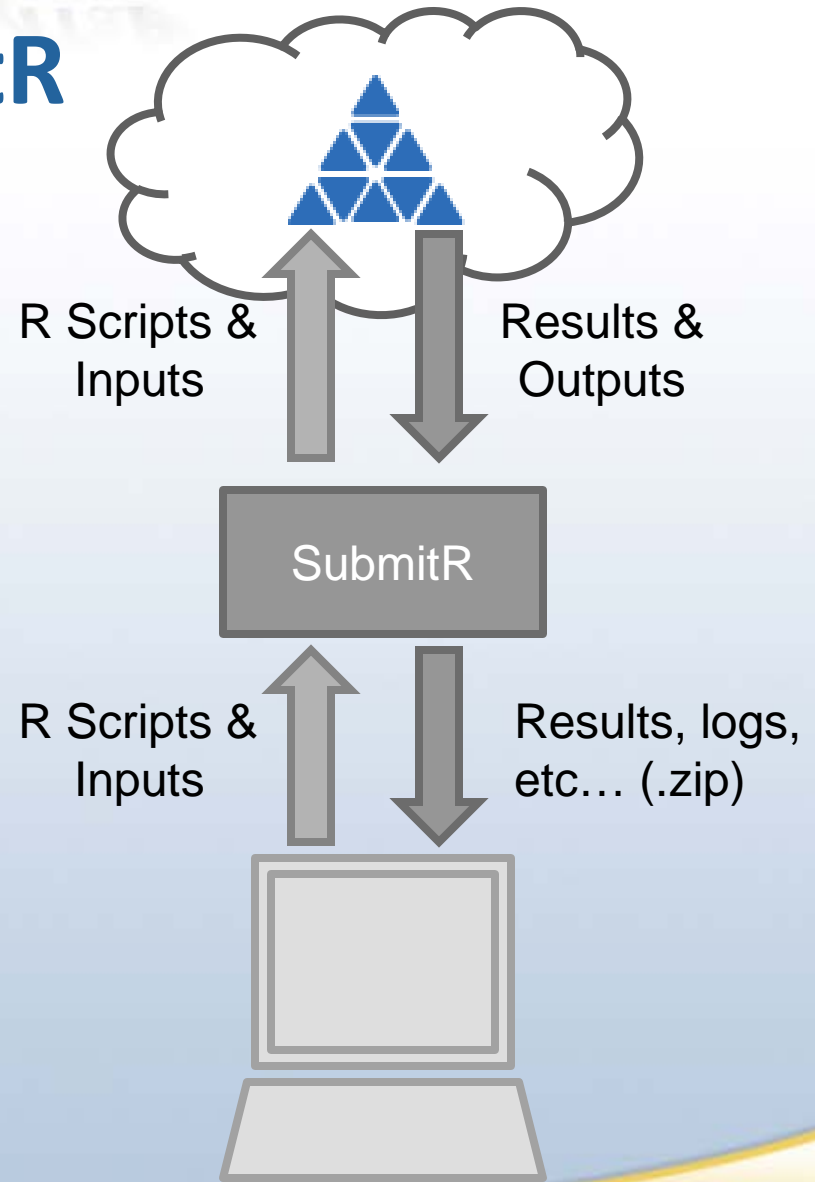
Or, select a parameter definition file: [Help](#)

Data File Template: [Help](#)

Post-processing:

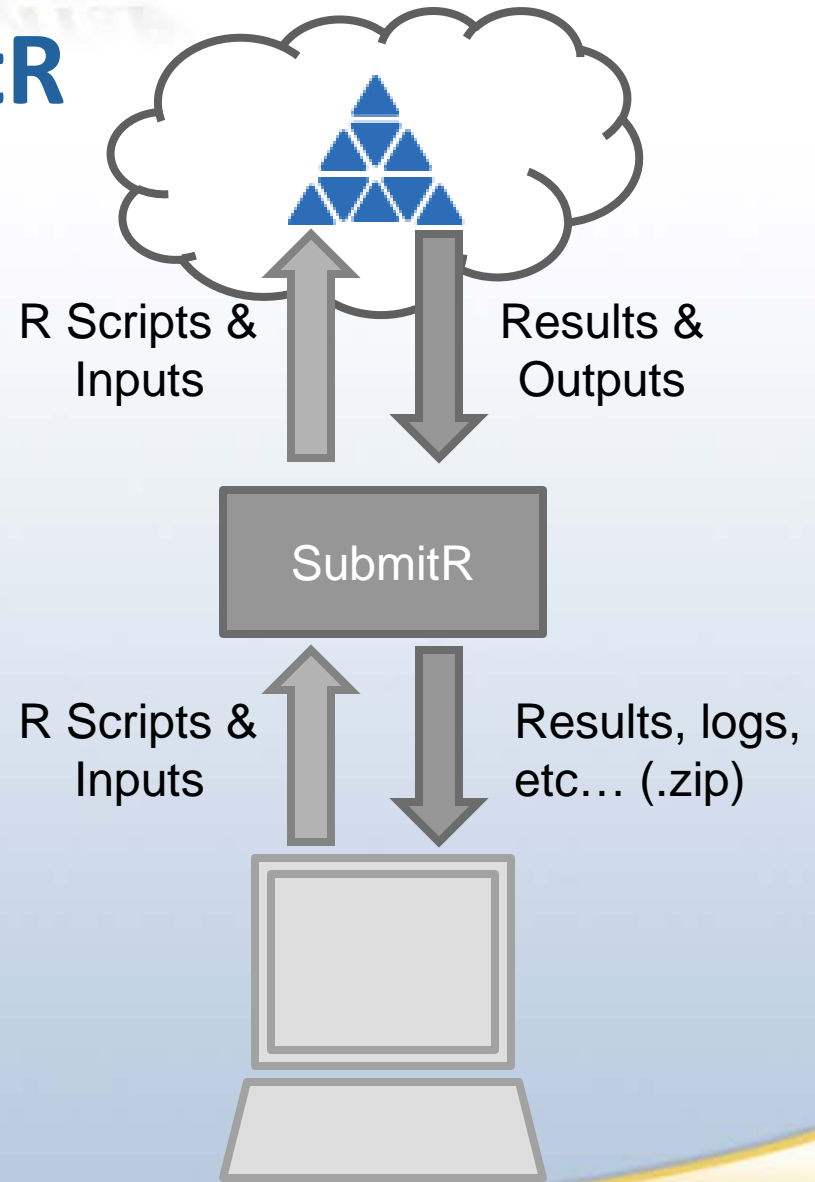
SubmitR

- 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	

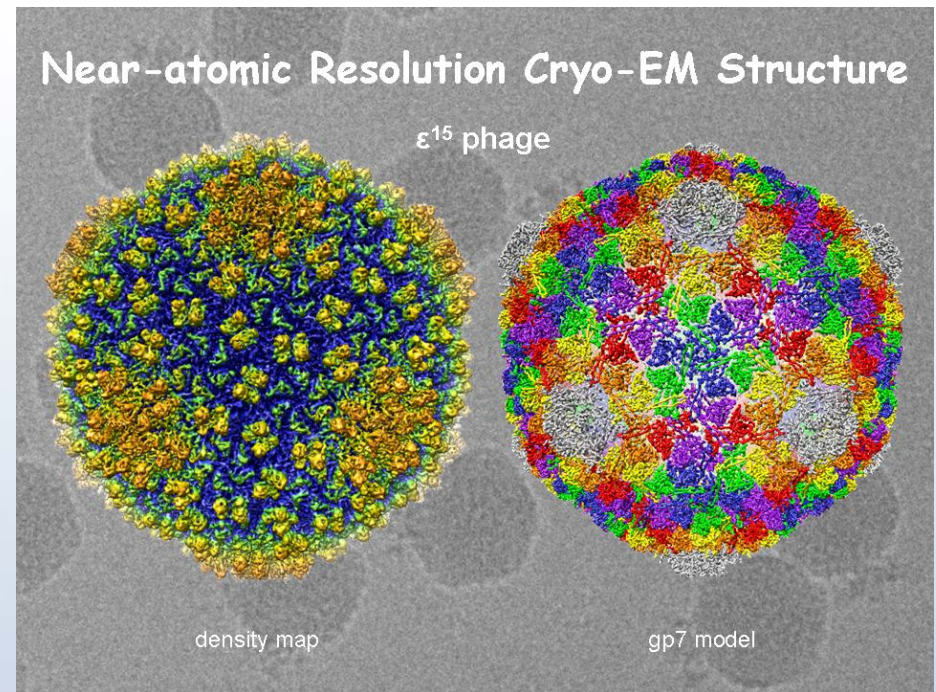
- And through the DiaGrid community features users can request more!

SubmitR usage examples

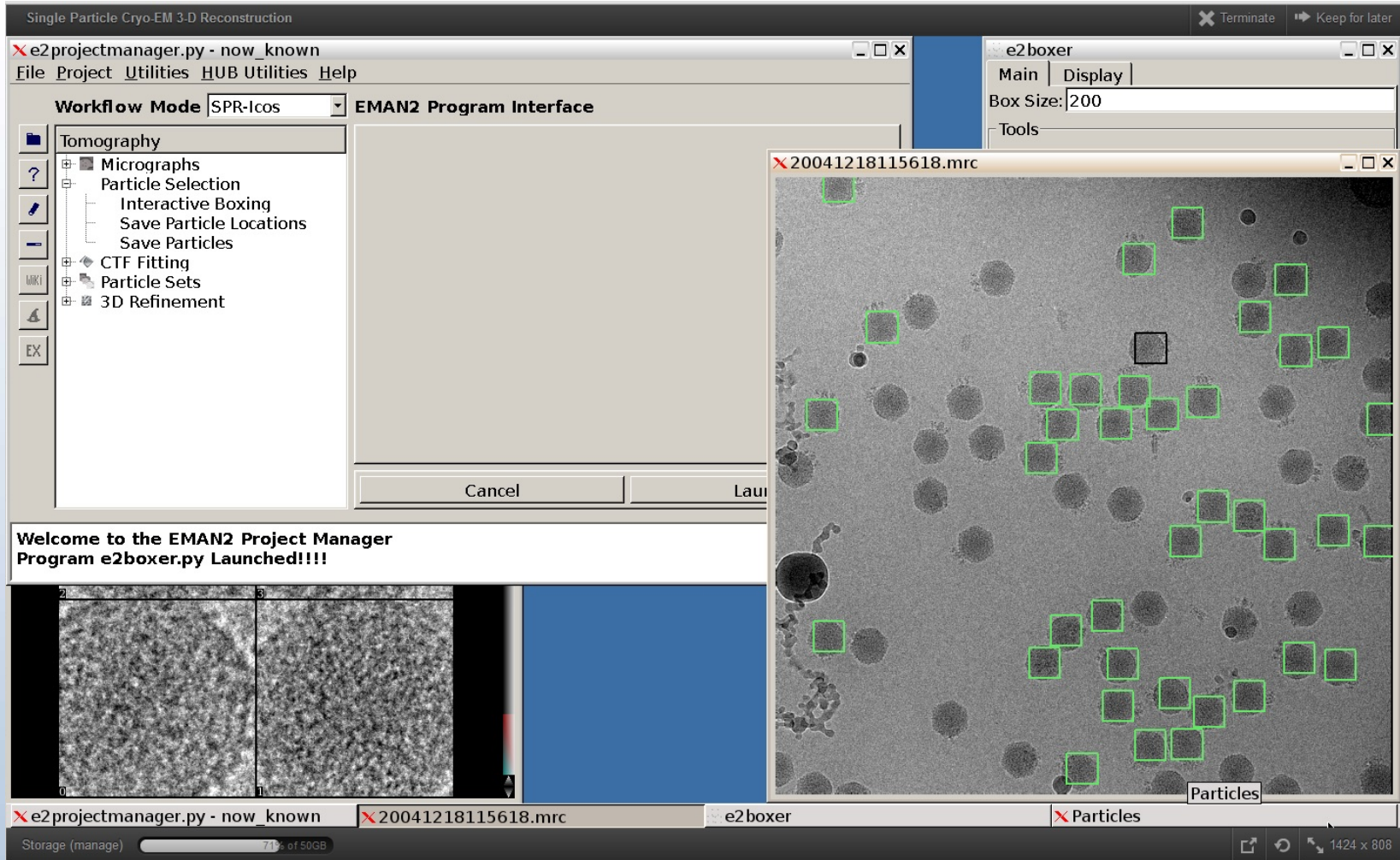
- Nutrition: (single, long running jobs)
 - Ingestive behavior research
- Bioinformatics: (single, long running jobs)
 - Genome association and prediction
- Agricultural Economics: (single and parallel jobs)
 - Distributed hydrological modeling
 - Effects of education on growth rates in developing countries
 - 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

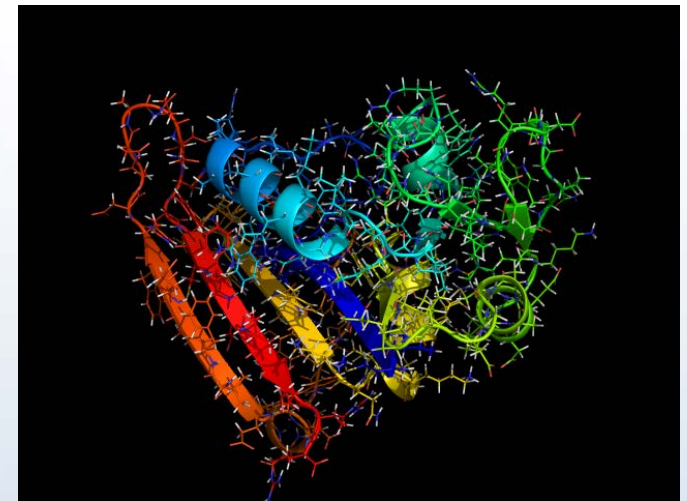
- The analysis of images taken at cryogenic temperatures within an electron microscope can reveal much about the structure of microscopic objects.
- Image processing is a good candidate for parallelization.
- The first user developed tool for the DiaGrid portal.
- DiaGrid staff utilized helping CryoEM authors split tasks for HT Condor then recombine with MPI for 3D visualization.



CryoEM

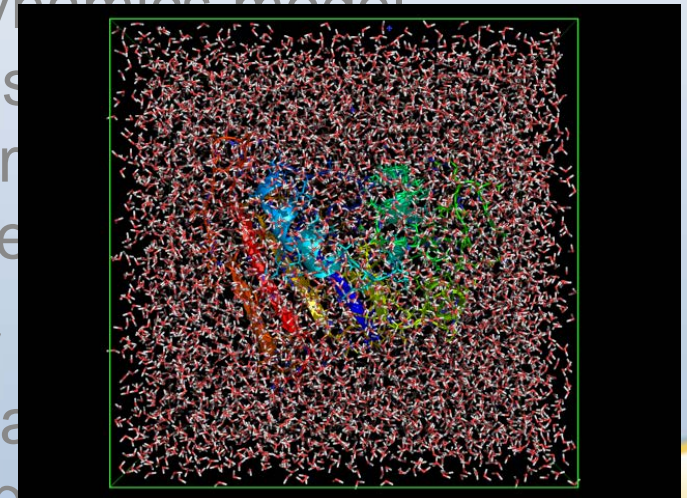


GROMACSIMUM



GROMACS is a molecular dynamics model
with a large community of users in
scientific disciplines from chemistry
medicine, physics, and biology.

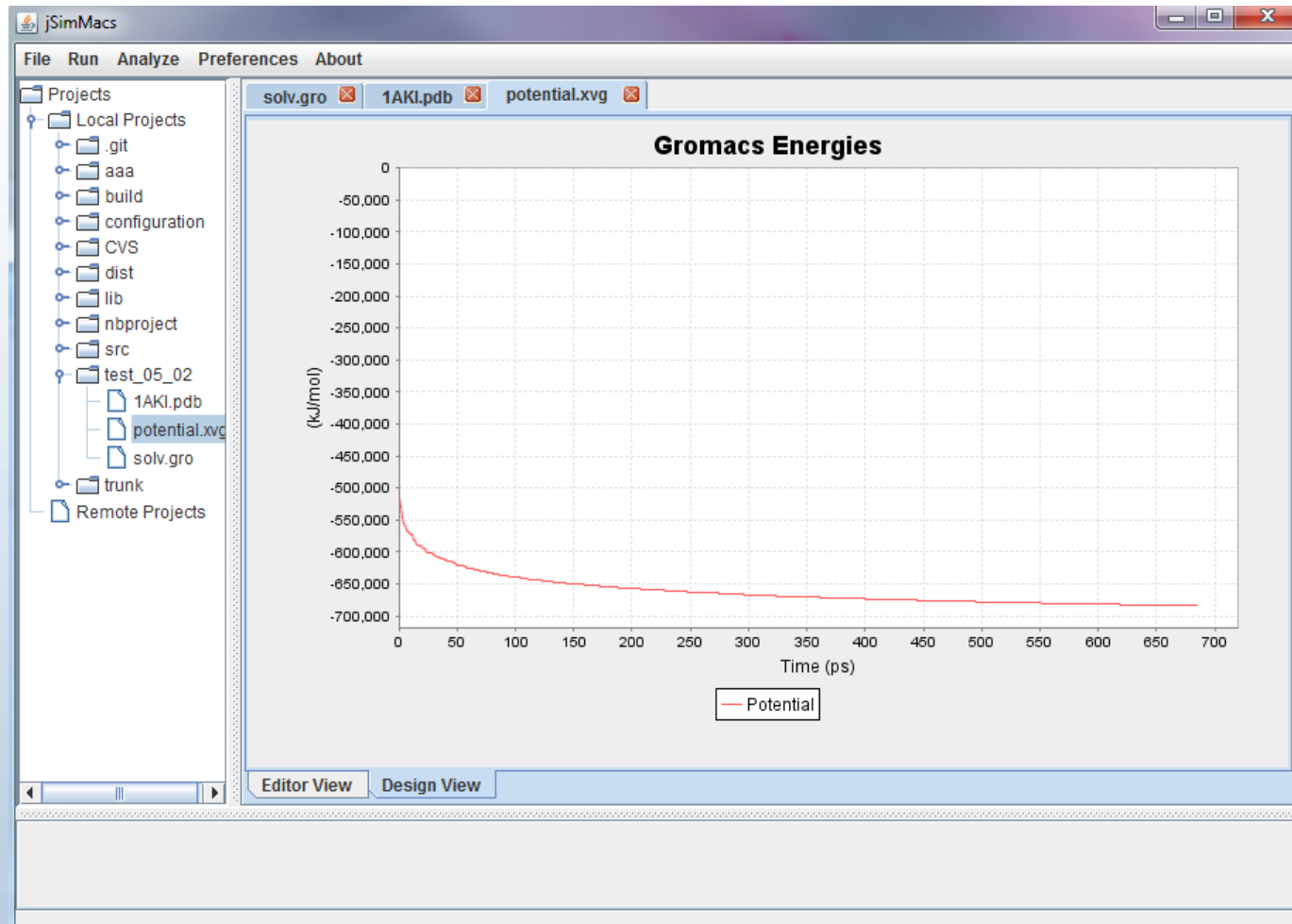
This project takes a popular
GROMACS GUI, jSimMacs, and
with new features for high-performance



GROMACSIMUM

The screenshot displays the jSimMacs application window. The main window has a menu bar with 'File', 'Run', 'Analyze', 'Preferences', and 'About'. On the left is a 'Projects' tree view showing a hierarchy of folders under 'Local Projects', including 'test_05_02' which contains '1AKI.pdb' and 'solv.gro'. The main area shows a 3D ball-and-stick model of a protein-ligand complex. A yellow stick model is highlighted within the protein structure. Below the 3D view is a sequence viewer showing a protein sequence: 'E S N F N T Q A T N R N T D G S T D Y G I L Q I N S R W W C N D G R T P G S R N L C N I P C S A L L S S D'. A 'Make .ndx' button is located below the sequence. In the foreground, a 'Make NDX' dialog box is open, with fields for 'File name:' and 'Group name:', and 'Make' and 'Cancel' buttons.

GROMACSIMUM



CESM

CESM is a global
many aspects of
Purdue developed
designed it to su

This project will explore providing an alternate
interface to the CESM gateway services from
inside DiaGrid.

The screenshot shows the CESM software interface. At the top, there is a menu bar with 'File', 'Edit', and 'Help'. Below the menu bar, there are two tabs: 'New Case' and 'File Management'. The version number 'ver: 0.2.2 (alpha)' is displayed on the right. A table with the following columns is visible: Case, Comp Set, Res, Status, and Duration. The table contains 17 rows of test cases. The row for 'test16' is highlighted in blue. To the right of the table is a vertical list of buttons: '1 Configure Case', '2 Build Case', '3 Submit Case', '4 Analyze Case', '5 Publish Results', and 'X Cancel Case'. At the bottom of the window, a status bar indicates 'Starting program...'

Case	Comp Set	Res	Status	Duration
test1	A	T31_T31	Created	N/A
test10	B	f09_g16	Running	N/A
test11	B	f19_g16	Running	N/A
test12	B	f19_f19	Created	N/A
test13	B	f19_g16	Configured	N/A
test14	BRCP45W...	f45_f45	Created	N/A
test15	B	f09_f09	Created	N/A
test16	B	f19_g16	Built	N/A
test17	B1850	f09_g16	Configured	N/A
test18	B	f09_g16	Configed	N/A
test19	B	f19_f19	Created	N/A
test2	B1850	T62_g37	Created	N/A
test20	B	f09_g16	Created	N/A
test21	B	f09_g16	Created	N/A

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