How to use Science Gateways
(and how to build them)

Amit Majumdar, Division Director
Data Enabled Scientific Computing Division

Wayne Pfeiffer, Distinguished Scientist

San Diego Supercomputer Center
University of California San Diego

SDSC Summer Institute 2017

Outline

- **Science Gateways**
  - Examples
- **Hands-on using a Science Gateway**

Science Gateways

- An online community space for science and engineering research and education

- A web-based resource for accessing data, software, computing services and equipment specific to the needs of a science or engineering discipline

HPC Resources

- **Available via XSEDE – Extreme Science and Engineering Discovery Environment**
  - NSF funded supercomputers, advanced support, services, allocation, EOT
- **You learned yesterday – how to**
  - Compile codes
  - Launch and manage jobs
  - Manage data on filesystems
  - Use HPC resources effectively
  - etc

Administrative and Technical tasks (barriers?)

- Write allocation proposals (peer-reviewed) for supercomputer time every year
- Understand HPC machines, policies, complex OS/software
- Install and benchmark complex applications on HPC resources
- Different machines have different schedulers
- Understand and manage remote authentication
- Figure out data transfer, file systems, storage

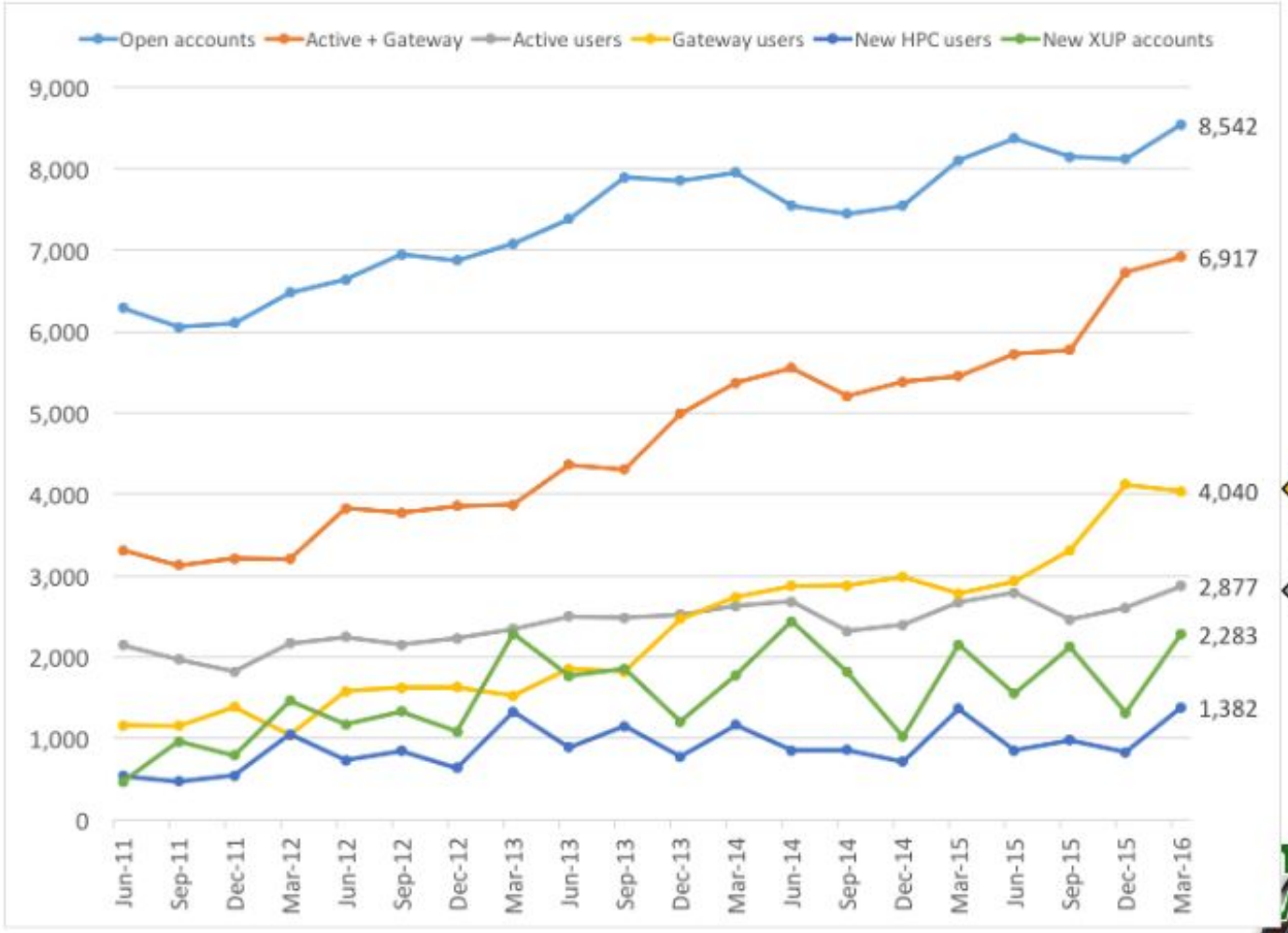
Science gateways

- **Easy web based user interface GUI**
  - Upload input files, models
  - Set application and HPC related parameters
  - Run jobs by the click of a button
- **Scientific applications already installed optimally on HPC resources at the backend**
- **Easily access, download output results**
  - Some provide post processing, viz
- **Some provide RESTful services**
- **Gateway team writes annual allocation proposal**

Catalyzes and democratizes computational science research for researchers and students from all universities, colleges and institutions

Gateway users surpass login users in 2013

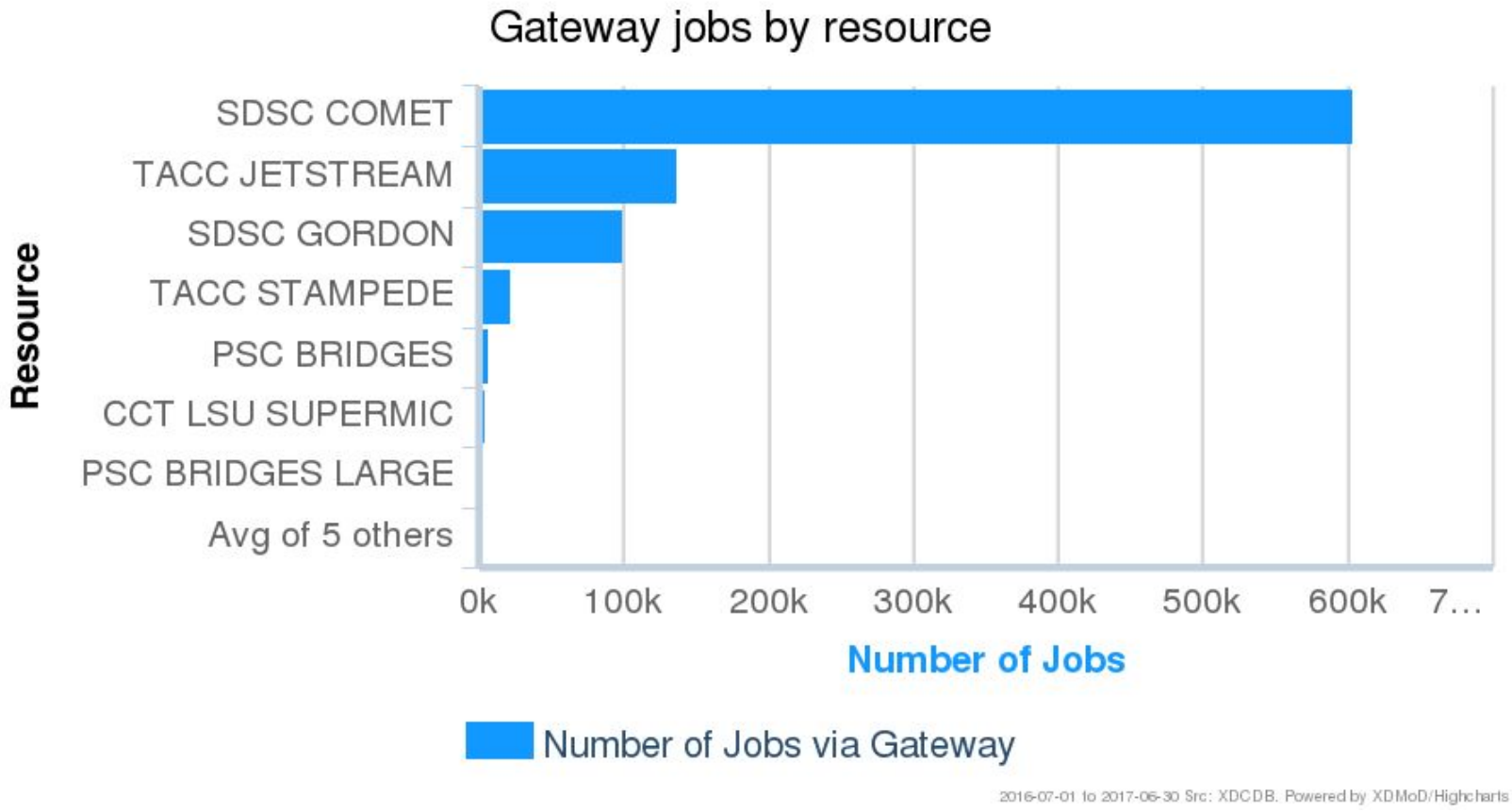
Automated user-counting in 2015



Source: David Hart

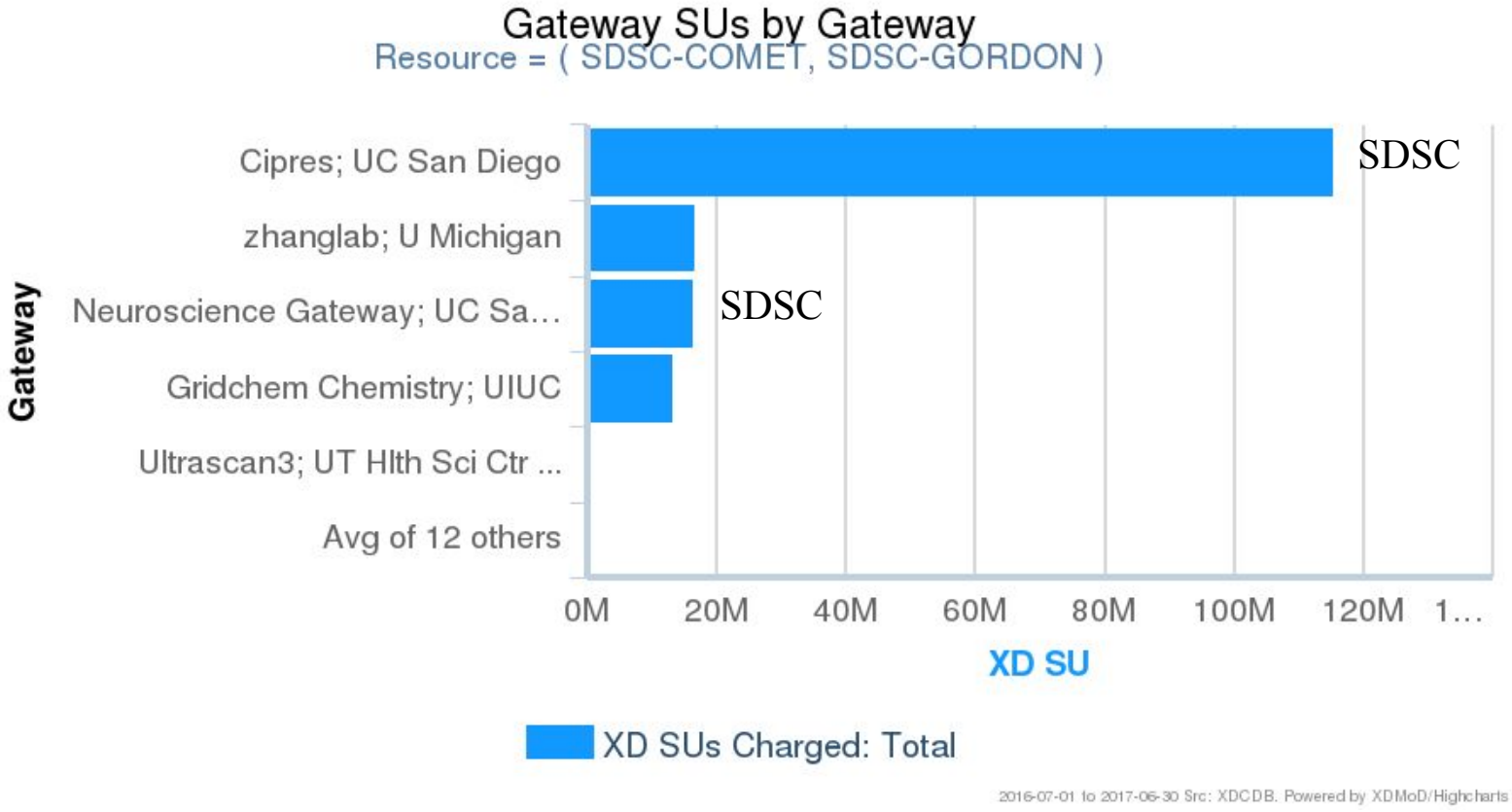
SDSC leads in hosting gateways

- **Comet and Gordon accounted for 80% of gateway jobs on XSEDE resources over the past year (7/16 thru 6/17)**



SDSC leads in developing & maintaining gateways

- 2 of top 4 gateways by usage over the past year are from SDSC



Tactics for Gateway Success:

Step 1: identify a user population in need

Step 2: commit to responding to user's needs

Step 3: let user behavior/needs drive improvements

Step 4: manage challenges that threaten productivity of high end users

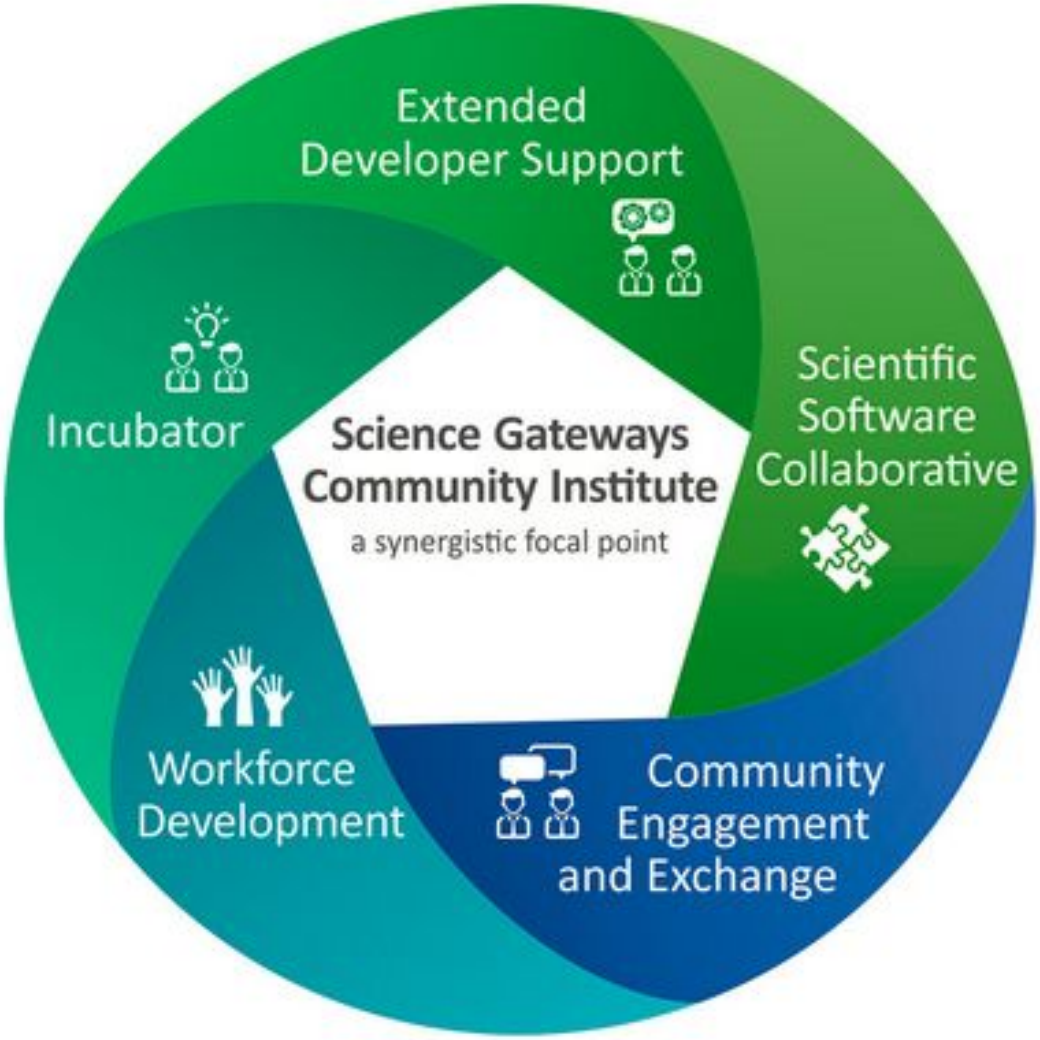
Step 5: with limited resources, prioritization is key

Step 6: stay in touch with your community

Step 7: embrace customer service

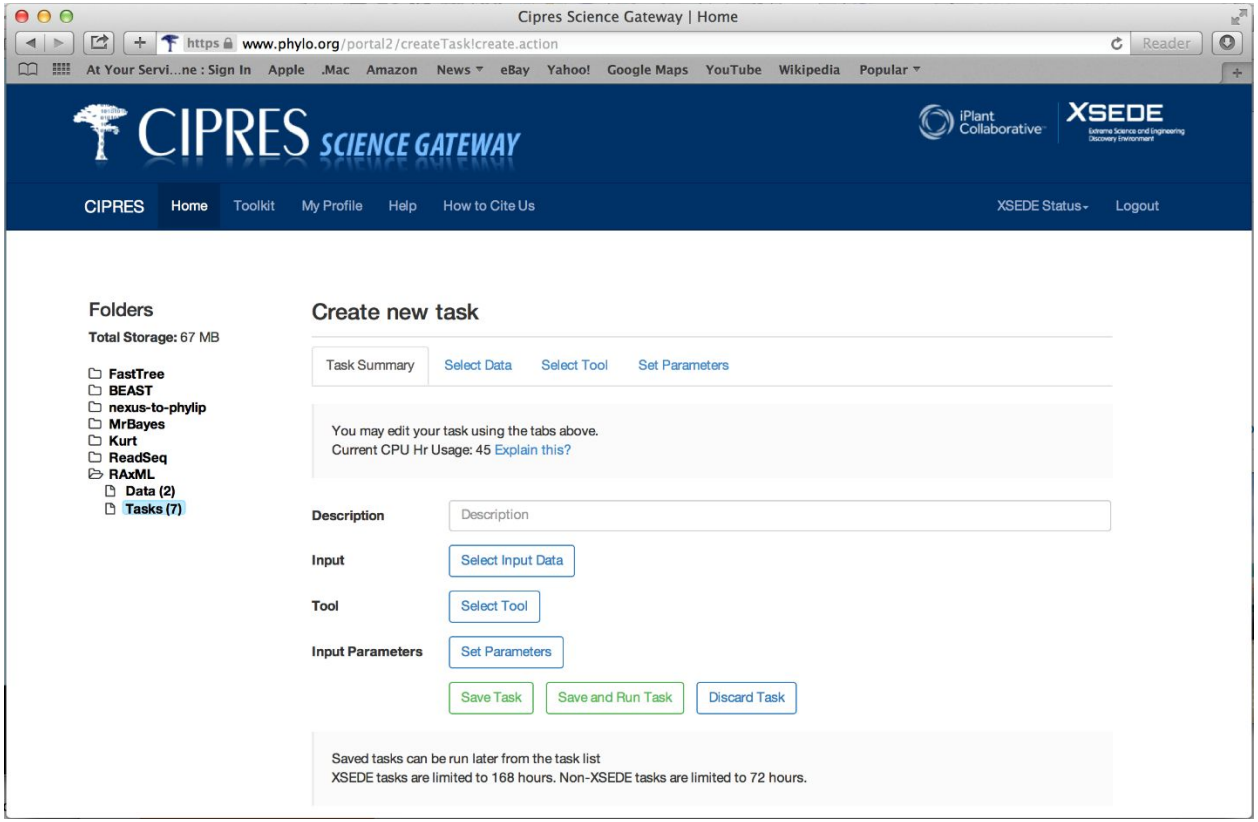
Nancy Wilkins-Diehr, SDSC – PI - http://sciencegateways.org/

Other institutions: Elizabeth City State in North Caronila, Indiana University, University of Notre Dame, Purdue University, the Texas Advanced Computing Center at the University of Texas, Austin, and the University of Michigan at Ann Arbor



The five key areas for the Science Gateways Community Institute to increase the number, ease of use, and effective application of gateways to serve the greater research and engineering community. *Source: SDSC*

The NSF-funded CIPRES gateway runs phylogenetics codes via a browser interface on supercomputers at SDSC



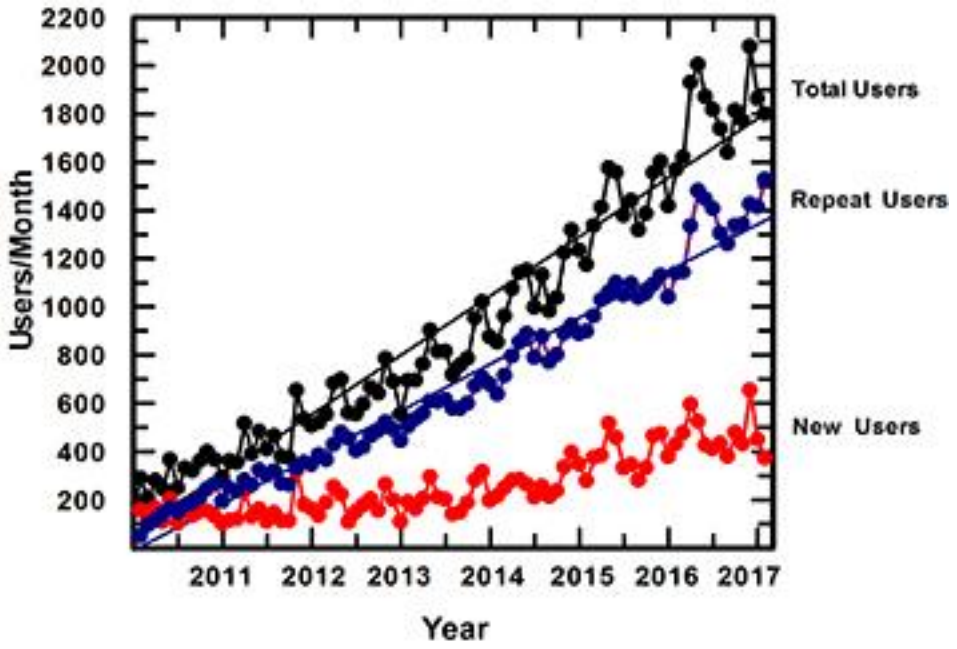
www.phylo.org

CIPRES has been developed & maintained by SDSC staff

Core team:
Mark Miller (PI)
Wayne Pfeiffer
Terri Schwartz

The CIPRES gateway has been extremely popular and supports thousands of researchers around the world

- ***>22,000 CIPRES users have run on NSF-funded supercomputers, including ~8,000 in the past year!***
- ***>3,000 publications have been enabled by CIPRES use!***
- ***US statistics from 2015***
  - *49 states + 2 territories + DC*
  - *252 universities & colleges*
  - *18 institutes*
  - *22 museums, gardens, & zoos*
  - *21 government agencies*
  - *4 high schools*
- ***Non-US statistics from 2015***
  - *85 countries*
  - *603 universities & colleges*
  - *161 institutes*
  - *80 museums, gardens, & zoos*
  - *134 government agencies*



The advent of DNA sequencing lets scientists infer phylogenetic trees from multiple sequence alignments

DNA, RNA, or AA sequences for multiple taxa

Multiple sequence alignment is matrix of taxa vs characters

```
                                  .   .   .    ......   .   .
Human
AAGCTTCACCGGCGCAGTCATTCTCATAAT...
Chimpanzee
AAGCTTCACCGGCGCAATTATCCTCATAAT...
Gorilla
AAGCTTCACCGGCGCAGTTGTTCTTATAAT...
Orangutan
AAGCTTCACCGGCGCAACCACCCTCATGAT...
Gibbon
AAGCTTTACAGGTGCAACCGTCCTCATAAT...
```

Final output is phylogeny or tree with taxa at its tips

```
 |---------- Chimpanzee
 +
 |   /---------- Gorilla
 |   |
 \---+             /------------------------------- Orangutan
     \-------------+
                   \------------------------------------------------ Gibbon
```

Multiple sequence alignment: ClustalW, MAFFT, …

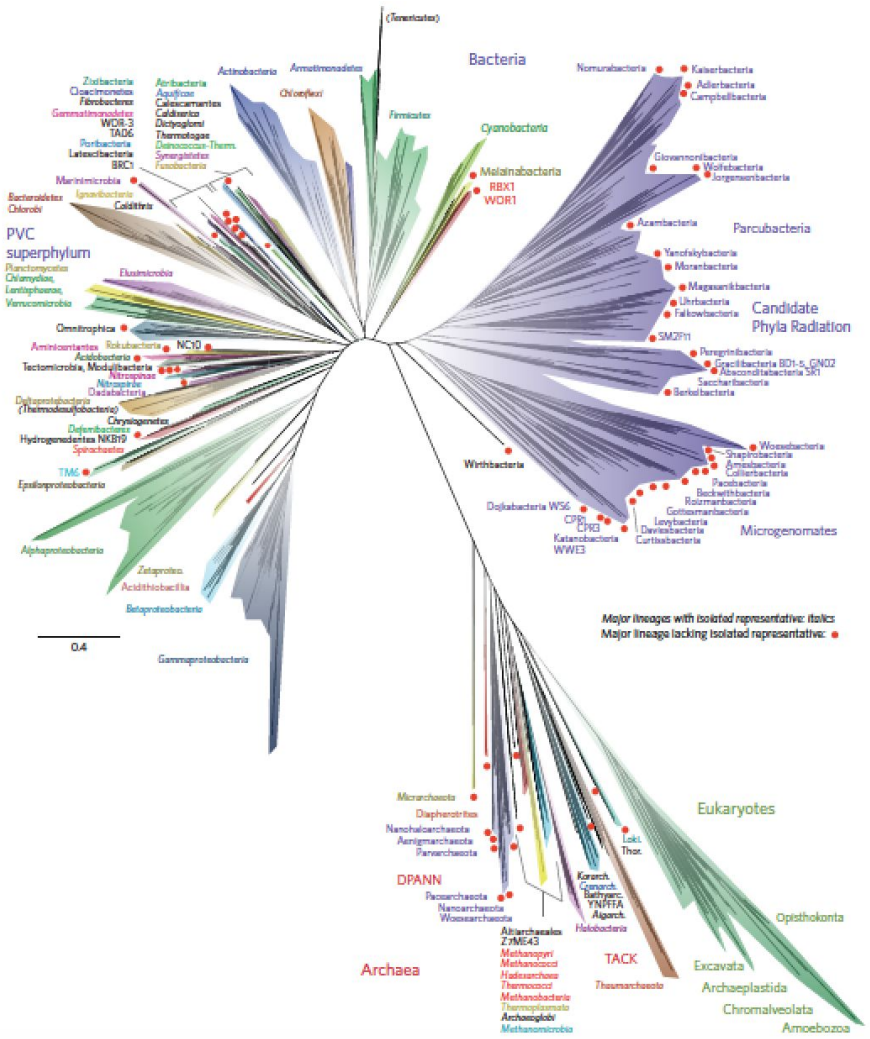
Aligned sequences

Phylogenetic tree inference: BEAST, MrBayes, RAxML, …

10 most popular codes that run via CIPRES on Comet; most have modest scalability; some run for days

Code	Latest version	Language	Computer	Cores charged
BEAST *	1.8.4	Java + C++	Comet	2 to 48
BEAST2	2.4.6	Java + C++	Comet	1 to 6
FastTree	2.1.9	C	Comet	3
GARLI	2.01	C++	Comet	1 to 24
jModelTest2	2.1.10	Java + C	Comet	24
MAFFT	7.305	C	Comet	12
MrBayes	3.2.6	C + C++	Comet	2 to 24
Migrate	3.6.11	C	Comet	1 to 72
Phylobayes	1.7b	C++	Comet	48
RAxML	8.2.10	C	Comet	12, 24, or 48

** Runs on GPUs as well as Intel x86 cores*



nature microbiology

LETTERS

PUBLISHED: 11 APRIL 2016 | ARTICLE NUMBER: 16048 | DOI: 10.1038/NMICROBIOL.2016.48

OPEN

A new view of the tree of life

Laura A. Hug[1†], Brett J. Baker[2], Karthik Anantharaman[1], Christopher T. Brown[3], Alexander J. Probst[1], Cindy J. Castelle[1], Cristina N. Butterfield[1], Alex W. Hernsdorf[3], Yuki Amano[4], Kotaro Ise[4], Yohey Suzuki[5], Natasha Dudek[6], David A. Relman[7,8], Kari M. Finstad[9], Ronald Amundson[9], Brian C. Thomas[1] and Jillian F. Banfield[1,9*]

Tree was generated with RAxML on 48 cores of Comet in a 3-day run via CIPRES

Vast, new superphylum of bacteria at upper right consists of phyla without isolated representatives identified only through metagenomic analyses

<- You are here in Opisthokonta, which includes animals & fungi

NSF funded Neuroscience Gateway at SDSC

Amit Majumdar, Subhashini Sivagnanam, Kenneth Yoshimoto, SDSC, UCSD
Ted Carnevale, Yale
Angus Silver, Padraig Gleeson, University College London
Maryann Martone, Anita Bandrowski, UCSD

- **NSG – in operation since early 2013 – nsgportal.org**
- **Built using the CIPRES gateway software**
- **NSG benefits the broader neuroscience research community in several ways, e.g.:**
- **Researchers can run larger complex neuronal networks, parameter sweep simulations, brain image processing tools**
- **Fully integrated The Virtual Brain (TVB) connectome pipeline workloads can be processed in parallel**
- **Easy access to widely used simulation tools, software such as:**
  Brian, NEST, NEURON, GENESIS, PyNN, MOOSE, FreeSurfer, Matlab, R, Tensorflow etc.
- **Researchers from EU Human Brain Project provided BluePyOpt optimization pipeline**
- **Access to new HPC resources – GPUs, Intel MICs**
- **Can be used by researchers with limited local (university-level) resources to address questions that require access to large scale, advanced systems**
- **Can be used by simulator developers to test, benchmark, and scale codes on large scale resources**
- **Can be used for classes, workshops, and tutorials**

Neuroscience Tools on NSG

BluePyOpt on Comet (1.1.27) - Running BluePyOpt analyses

Brian on Stampede (2.0b2) - Brian is a simulator for spiking neural networks

Brian on Comet (2.0b2) - Brian is a simulator for spiking neural networks

The Virtual Brain Personalized Multimodal Connectome Pipeline on Comet () - Connectome Pipeline on Comet

FREESURFER on Comet (5.3.0) - Freesurfer tool on Comet

PyMOOSE (3.0.1 Gulab Jamun) - Running Moose models on Comet

NEST on Stampede (2.6.0) - Neural Simulation Technology using Python

NEST using Python on Comet (2.2.1) - Neural Simulation Technology using Python

NEST on Stampede (2.6.0) - Neural Simulation Technology

NEST on Comet (2.2.1) - Neural Simulation Technology

NEURON7.3 Python on Stampede (7.3) - Using Python to run NEURON 7.3

NEURON7.3 Python on Comet (7.3) - Using Python to run NEURON 7.3

NEURON7.3 on Stampede (7.3) - Latest NEURON simulation software package on Stampede

NEURON7.3 on Comet (7.3) - Latest NEURON simulation software package on Comet

NEURON7.4 Python on Comet (7.4) - Using Python to run NEURON 7.4

NEURON7.4 on Comet (7.4) - Latest NEURON simulation software package on Comet

PGENESIS on Stampede (2.3) - Parallel Genesis software

PGENESIS on Comet (2.3) - Parallel Genesis software

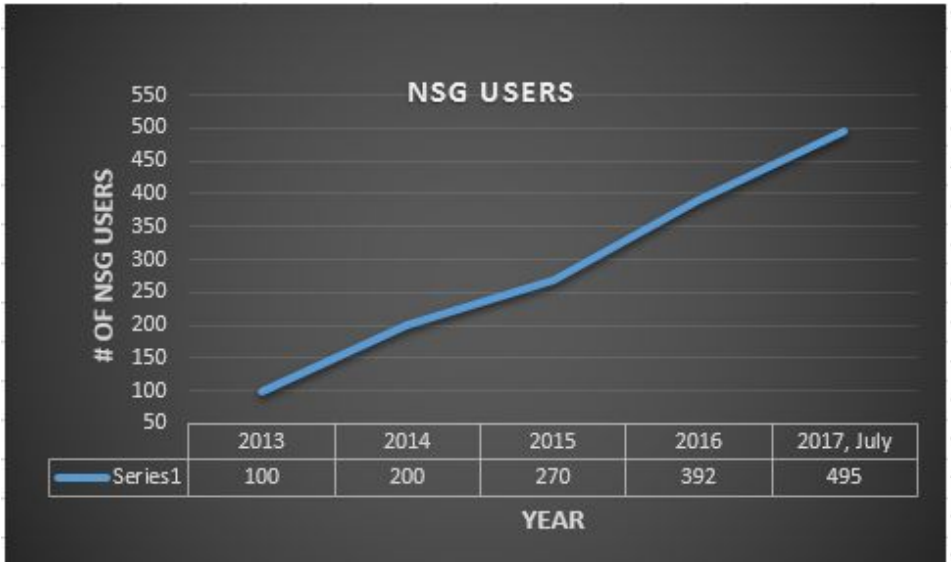
PyNN on Stampede (0.7.5) - Python package for simulator-independent specification of neuronal network models

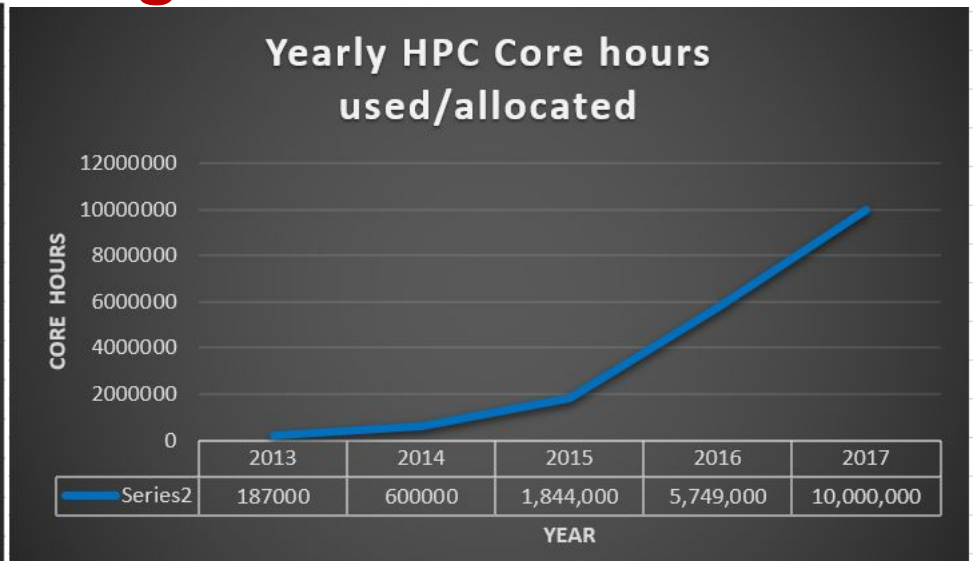
PyNN on Comet (0.7.5) - Python package for simulator-independent specification of neuronal network models

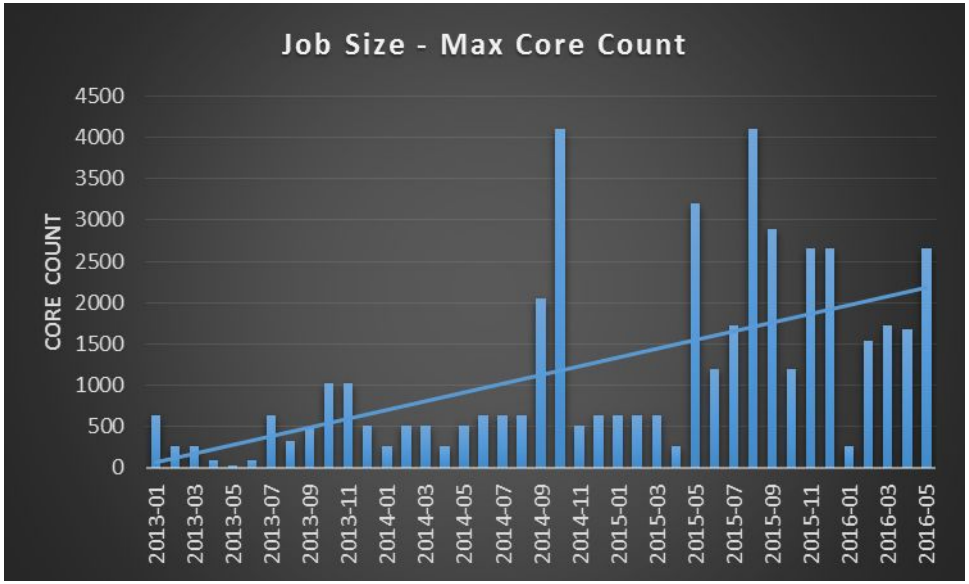
Python on Stampede (2.7.9) - Running Python models

Python on Comet (2.7.9) - Running Python models

NSG Usage


NSG USERS
OF NSG USERS
550
500
450
400
350
300
250
200
150
100
50
2013
2014
2015
2016
2017, July
Series1
100
200
270
392
495
YEAR


Yearly HPC Core hours used/allocated
CORE HOURS
12000000
10000000
8000000
6000000
4000000
2000000
0
2013
2014
2015
2016
2017
Series2
187000
600000
1,844,000
5,749,000
10,000,000
YEAR


Job Size - Max Core Count
CORE COUNT
4500
4000
3500
3000
2500
2000
1500
1000
500
0
2013-01
2013-03
2013-05
2013-07
2013-09
2013-11
2014-01
2014-03
2014-05
2014-07
2014-09
2014-11
2015-01
2015-03
2015-05
2015-07
2015-09
2015-11
2016-01
2016-03
2016-05

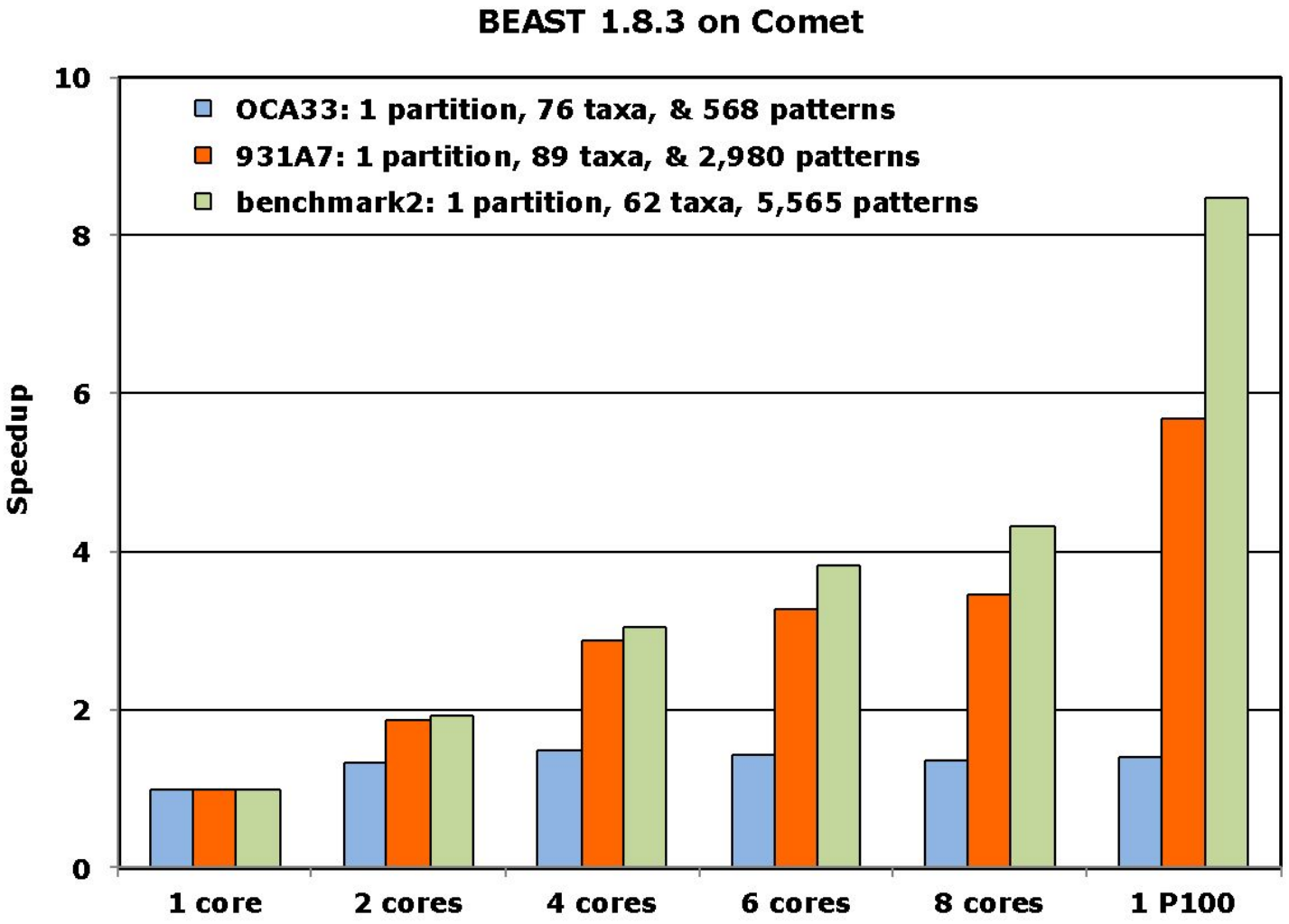
Hands-on – to show how easy it is!

0. Make sure you have the benchmark2.xml data set.
1. Go to https://www.phylo.org and click on Use the CIPRES Science Gateway.
2. Login as a guest without registering.
3. Click on Guest Folder, Data, and Upload/Enter Data. Then upload the benchmark2.xml data set to Data in Guest Folder.
4. Click on Tasks and Create new Task.
5. Select Input Data: i.e., the data set just uploaded.
6. Select Tool: BEAST on XSEDE.
7. Select Input Parameters. Use the default parameters with the following exceptions.

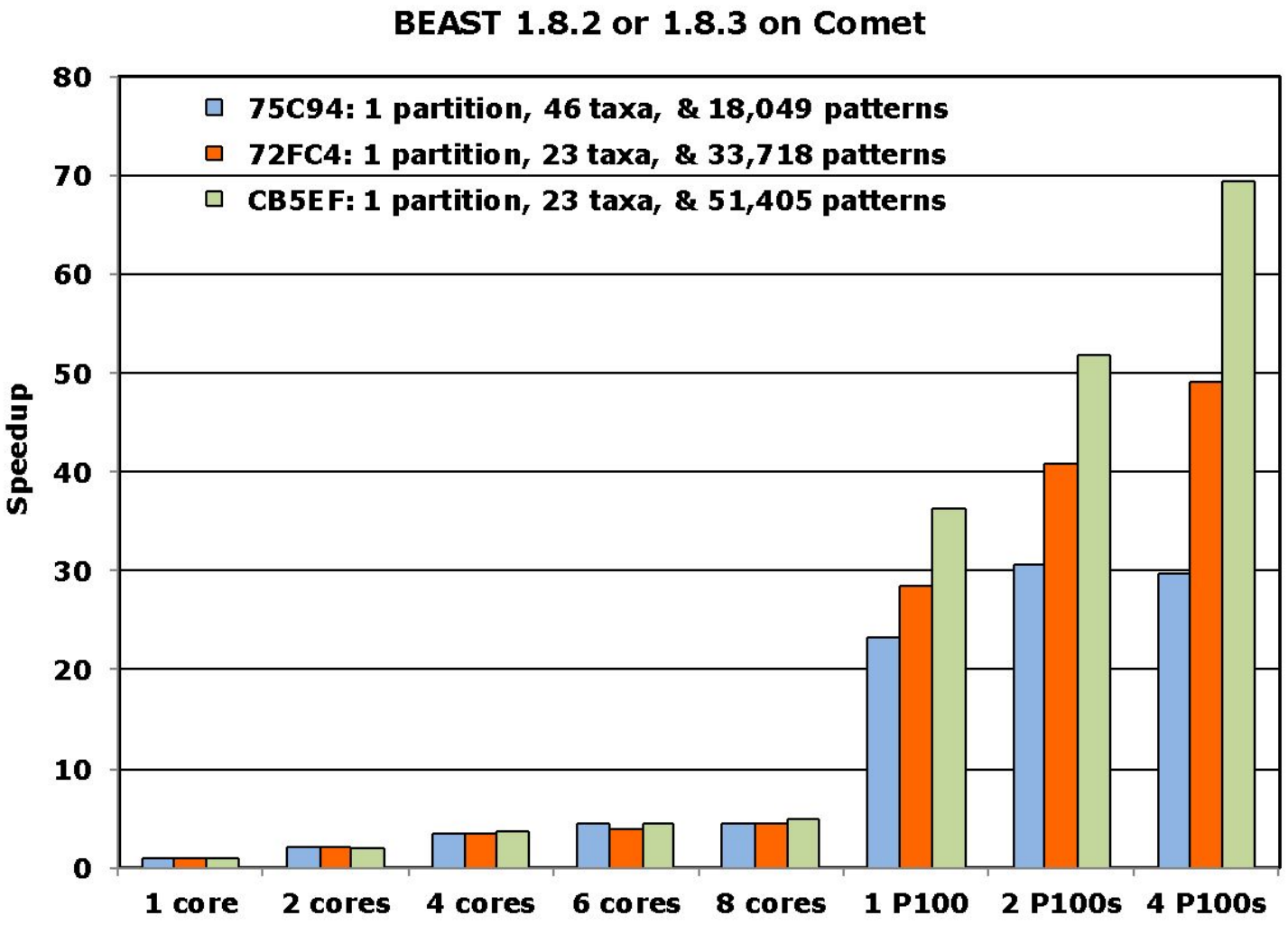
. First row folks specify 1,000 patterns. That will have the job run on 2 cores.
. Second row folks specify 2,000 patterns. Then the job will run on 4 cores.
. Third row folks specify 5,000 patterns. Then the job will run on 8 cores.
. Fourth row folks specify 5,000 patterns and two partitions. Then the job will run on 6 cores.

8. Save parameters.
9. Enter an appropriate Description for task, e.g., benchmark2.1000patterns
10. Click on Save and Run Task.
11. Click on View Status and then Intermediate Results while job is running or Output when job is done.
12. Then look at stdout.txt. Near the bottom, the time is output in seconds or minutes.

Speedup of BEAST improves with the number of patterns; CIPRES jobs use 2, 4, & 8 cores for these 3 data sets



Speedup of BEAST is outstanding on P100s for >10,000 patterns; CIPRES jobs use 1, 2, & 4 GPUs for these 3 data sets



Science Gateways - Summary

- **Allows anyone from anywhere to *easily* access and use HPC (and data, instrument etc.) for computational science**
  - All users start out with some amount of core hours (depends on the gateway and the science)
  - If you graduate out of a SGW, you can write your own allocation proposal (gateway/XSEDE staff can help)
    - In many cases still use the gateway to charge to your allocation
- **It creates a cyberinfrastructure environment for the science community to enable**
  - Research
  - Education
  - Sharing of information and data