XSEDE14

July 13-18 Atlanta

Genome Analysis

Birds of a feather

Craig A. Stewart
Philip Blood
Rich Knepper

National infrastructure serving genome science

Creators of new software

NCGAS – small, serving large community largely reactively

- Trinity
- Galaxy
- ABySS
- Velvet

iPlant – large collaborative serving plant science

- DNA Subway
- iPlant Discovery Environment
- Many bioinformatics software applications planned as part of group strategy

XSEDE – designed to serve all research communities

- Stampede
- Gordon
- Blacklight
- Comet
- Mason
- Wrangler
- FutureGrid

Network – essentially independent of any particular research community

- Internet2
- Regional providers

XSEDE resources

System	Type of resource	Type of Service Provider?
Stampede	Large scale distributed memory parallel	NSF funded, Level 1
Gordon	Large scale distributed memory parallel, pseudo large memory	“
Blacklight	Large memory	“
Comet	New - VMs	“
Wrangler	Storage	“
FutureGrid	Experimental computer science / cloud system	“
Mason	*Large memory (low cores)*	*IU-funded, Level 2*
Rockhopper	*Commercial “cluster as a service” owned by Penguin Computing and housed at / supported by IU*	*Commercially owned, Level 3*

Pittsburgh Supercomputing Center Blacklight (SGI Altix® UV 1000) - Massive Coherent Shared Memory Computer

- **2×16 TB of cache-coherent shared memory, 4096 cores**
  - *ideal for genome sequence assembly*
  - High bandwidth, low latency interprocessor communication
- **SUSE Linux operating system**
  - *excellent for portability:* supports OpenMP, C, C++, Java, Perl, Python, p-threads, MPI, UPC
  - *rapid algorithm development*

This slide courtesy Philip Blood, Pittsburgh Supercomputing Center, © PSC

Mason

- Supports data-intensive high performance computing tasks for IU researchers, faculty, staff, and students on all campuses.

Specs:

- Peak performance of 3.83 teraFLOPS
- 8 TB total RAM - 512 GB RAM per node – really a system of memory with a few processors attached
- Uses Lustre/Data Capacitor II as high performance file system
- Connects to IU's high speed research network via 10 Gbps connection

iPlant – Plant Cyberinfrastructure

Goals:

- *“to create a new type of organization a cyberinfrastructure collaborative for plant science”*
- *“to enable new conceptual advances through integrative, computational thinking”*
- *“to address an evolving array of grand challenge questions in plant science: the driving force and organizing principles for the collaborative”*
- ~ $10M / year ($50M NSF Funded Project – 5 years, renewed in 2013)
- iPlant is a cyberinfrastructure *platform*
- The *platform* is developed by iPlant and extensible by users
- NSF recommended scope beyond plants

XSEDE Novel and Innovative Projects program

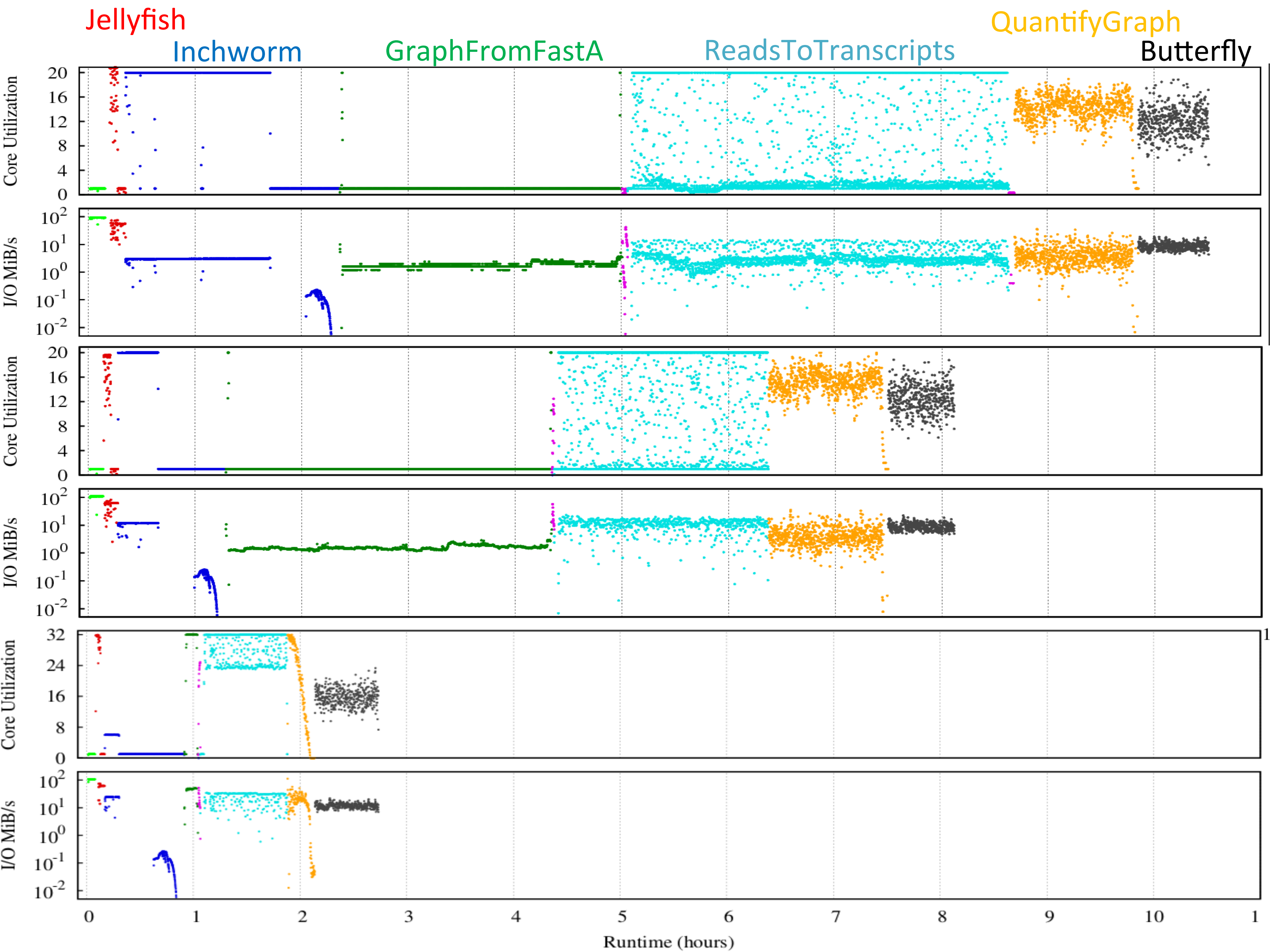
Researchers
XSEDE/PSC
Developers

- Novel and Innovative Projects within XSEDE is intended to be reactive to new user needs, with current focus on life sciences
- Work with developers to port key de novo assembly applications to large shared memory system, Blacklight
- Availability of Blacklight highlighted on Broad Institute developer web pages (ALLPATHS-LG and Trinity) and genomeweb.com
- Enthusiastic response from research community – dozens of new groups using Blacklight for de novo assembly every year
- Example projects:
  - **Cold Spring Harbor:** Assembled **5 and 10 gigabase wheat species** using **3 and 6 TB RAM** respectively. Targeting assembly of **16 gigabase** wheat genome (**ALLPATHS-LG**).
  - **Cornell** and **Broad Institute:** Assembled **20 primate transcriptomes at ~1 TB RAM each** (**Trinity**). Understanding evolutionary processes and gaining insight into human disease.

This slide courtesy Philip Blood, Pittsburgh Supercomputing Center, © PSC

National Center for Genome Analysis Support
“Mind the Gap”

Gap	How we fill it
System configurations offered by XSEDE and what people doing genome assembly need	Mason (IU contribution to facilities)
Software on XSEDE is not what people need	NCGAS installs and maintains
Software works slowly	NCGAS tunes / re-engineers
People just need help	NCGAS provides consulting NCGAS goes to conferences and informs people about our services
People need storage	***NCGAS provides tape storage (IU facilities)***
People need to publish data sets	***IU provides resources via IUScholarWorks***


Jellyfish
Inchworm
GraphFromFastA
ReadsToTranscripts
QuantifyGraph
Butterfly
Core Utilization
0
4
8
12
16
20
I/O MiB/s
10^{2}
10^{1}
10^{0}
10^{-1}
10^{-2}
Core Utilization
0
4
8
12
16
20
I/O MiB/s
10^{2}
10^{1}
10^{0}
10^{-1}
10^{-2}
Core Utilization
0
8
16
24
32
1
I/O MiB/s
10^{2}
10^{1}
10^{0}
10^{-1}
10^{-2}
0
1
2
3
4
5
6
7
8
9
10
1
Runtime (hours)

REUs

- NCGAS Virtual Interns leverage XSEDE experience to gain industry employment.
- NCGAS ran virtual REU program with Clark State University in Springfield OH, learning how to install and configure bioinformatics software.
- Two associate degree students participated.
- Both ended up working as professional HPC admins at Wright Patterson Air Force Base.

So… what are your pain points?

- And what can XSEDE and NCGAS do to help?

License Terms

- Please cite as: Stewart, C.A., Blood, P., and Knepper, R. “Genome Analysis” Birds of a Feather. 2014. Presented at XSEDE 14, July 13-18, Atlanta, GA. Available at: http://hdl.handle.net/2022/18710

Acknowledgements

- NCGAS is a collaboration led by the Indiana University Pervasive Technology Institute and includes partners the Texas Advanced Computing Center of the University of Texas, Austin; the San Diego Supercomputer Center of the University of California San Diego; and the Pittsburgh Supercomputing Center. NCGAS is affiliated with the Indiana University Pervasive Technology Institute as a Cyberinfrastructure and Service Center.
- This research was supported by NSF Award 1062432 – ABI Development: National Center for Genome Analysis Support (NCGAS).
- This research was also supported by a generous grant from the Lilly Endowment, Inc. to the Indiana University Pervasive Technology Institute.
- Any opinions expressed here are those of the presenter and do not necessarily reflect positions of the National Science Foundation or the Lilly Endowment.