

# Welcome to the XSEDE Big Data Workshop

John Urbanic  
Parallel Computing Scientist  
Pittsburgh Supercomputing Center

# Who are we?

Your hosts:  
Pittsburgh Supercomputing Center

## Our satellite sites:

Jefferson Lab  
Yale University  
Tufts University  
University of Utah  
Purdue University  
Cornell University  
New York University  
University at Albany  
University of Houston  
University of Delaware  
Texas Tech University  
Northwestern University  
Old Dominion University  
Arizona State University  
George Mason University  
University of Notre Dame  
North Dakota State University  
Pennsylvania State University  
University of Colorado Boulder  
Georgia Institute of Technology  
University of Houston-Clear Lake  
Lawrence Berkeley National Laboratory  
National Center for Supercomputing Applications  
University of Tennessee, Knoxville - National Institute for Computational Sciences

# XSEDE

Extreme Science and Engineering  
Discovery Environment

# Who am I?

John Urbanic

Parallel Computing Scientist

Pittsburgh Supercomputing Center

What I mostly do:

Parallelize codes with

- MPI, OpenMP, OpenACC, Hybrid
- Big Data, Machine Learning

**XSEDE**

Extreme Science and Engineering  
Discovery Environment



# XSEDE HPC Monthly Workshop Schedule

- October 2-3 *HPC Monthly Workshop: MPI*
- November 6 *HPC Monthly Workshop: OpenACC*
- December 4-5 *HPC Monthly Workshop: Big Data*
- January 16 *HPC Monthly Workshop: OpenMP*
- February 12-13 *HPC Monthly Workshop: Big Data*
- March 5 *HPC Monthly Workshop: OpenACC*
- April 2-3 *HPC Monthly Workshop: Big Data*
- May 6-7 *HPC Monthly Workshop: MPI*
- June 3-6 *Summer Boot Camp*
- **August 6-7** ***HPC Monthly Workshop: Big Data***
- September 3-4 *HPC Monthly Workshop: MPI*
- October 1-2 *HPC Monthly Workshop: Big Data*
- November 5 *HPC Monthly Workshop: OpenMP*
- December 3-4 *HPC Monthly Workshop: Big Data*

# HPC Monthly Workshop Philosophy

- Workshops as long as they should be.
- You have real lives...
  - in different time zones...
  - that don't come to a halt.
- Learning is a social process
  - This is not a MOOC
  - This is the **Wide Area Classroom**  
so raise your expectations

# Agenda

Tuesday, August 6

- 11:00 Welcome
- 11:25 A Brief History of Big Data
- 12:00 Hadoop
- 12:30 Intro to Spark
- 1:00 Lunch Break
- 2:00 Spark
- 3:30 Spark Exercises
- 4:30 Spark
- 4:45 A Big Big Data Platform
- 5:00 Adjourn

Wednesday, August 7

- 11:00 Machine Learning: A Recommender System
- 1:00 Lunch break
- 2:00 Deep Learning with TensorFlow
- 5:00 Adjourn

# *We do this all the time, but...*

- This is a very ambitious agenda.
- We are going to cover the guts of a semester course.
- We may get a little casual with the agenda.
- Three reasons we can attempt this now:
  - Tools have reached the point (Spark and TF) where you can do some powerful things at a high level.
  - We are going to assume you will use your extended access to do exercises. Usually this is just a bonus.
  - Worked last time.

# Resources

Your local TAs

Questions from the audience

On-line talks (no "pop ups")

[bit.ly/XSEDEWorkshop](http://bit.ly/XSEDEWorkshop)

Copying code from PDFs is very error prone. Subtle things like substituting “-” for “-” are maddening. I have provided online copies of the codes in a directory that we shall shortly visit. I strongly suggest you copy from there if you are in a cut/paste mood.

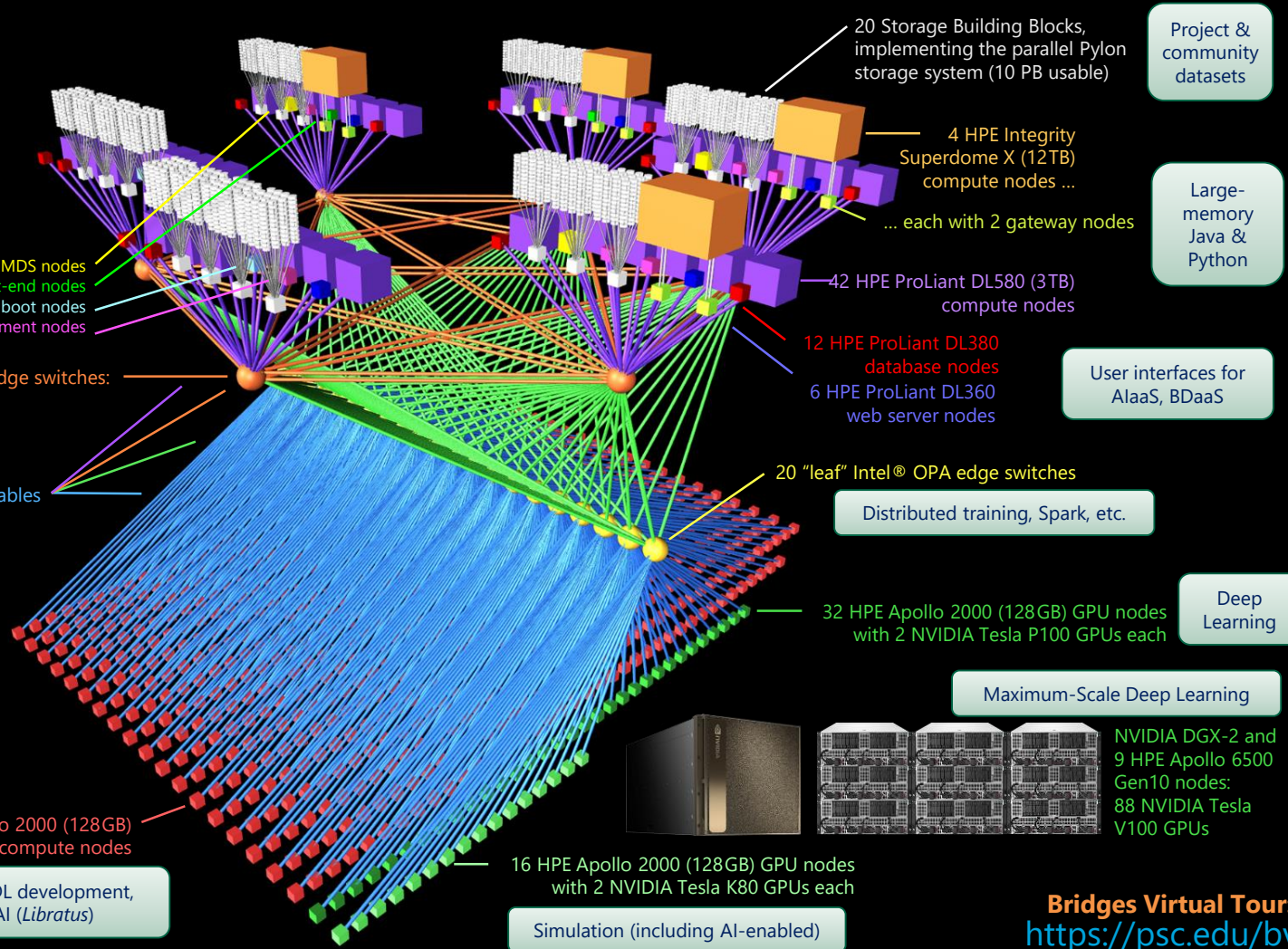
## The YouTube Channel Has Arrived!

Due to overwhelming demand, and a lot of editing, we have begun to post workshop videos on the XSEDE Monthly Workshop Training Channel:

***XSEDETraining***

They will be incrementally appearing in the coming months. Subscribe and give us feedback.





Purpose-built Intel® Omni-Path Architecture topology for data-intensive HPC

Maximum-Scale Deep Learning

ML, inferencing, DL development, Spark, HPC AI (*Libratus*)

Simulation (including AI-enabled)

**Bridges Virtual Tour:**  
<https://psc.edu/bvt>

This tool depicts the current plan for PSC's Bridges HPC system. Details are subject to change.

# Bridges Hardware

Type	RAM	#	CPU / GPU / SSD	Server
ESM	12 TB <sup>b</sup>	2	16 × Intel Xeon E7-8880 v3 (18c, 2.3/3.1 GHz, 45MB LLC)	HPE Integrity Superdome X
	12 TB <sup>c</sup>	2	16 × Intel Xeon E7-8880 v4 (22c, 2.2/3.3 GHz, 55MB LLC)	
LSM	3 TB <sup>b</sup>	8	4 × Intel Xeon E7-8860 v3 (16c, 2.2/3.2 GHz, 40 MB LLC)	HPE ProLiant DL580
	3 TB <sup>c</sup>	34	4 × Intel Xeon E7-8870 v4 (20c, 2.1/3.0 GHz, 50 MB LLC)	
RSM	128 GB <sup>b</sup>	752	2 × Intel Xeon E5-2695 v3 (14c, 2.3/3.3 GHz, 35MB LLC)	HPE Apollo 2000
RSM-GPU	128 GB <sup>b</sup>	16	2 × Intel Xeon E5-2695 v3 + 2 × NVIDIA Tesla K80	
	128 GB <sup>c</sup>	32	2 × Intel Xeon E5-2683 v4 (16c, 2.1/3.0 GHz, 40MB LLC) + 2 × NVIDIA Tesla P100	
GPU-AI16	1.5 TB <sup>d</sup>	1	16 × NVIDIA V100 32GB SXM2 + 2 × Intel Xeon Platinum 8168 + 8 × 3.84 TB NVMe SSDs	NVIDIA DGX-2 delivered by HPE
GPU-A8	192 GB <sup>d</sup>	9	2 × Intel Xeon Gold 6148 + 2 × 3.84 TB NVMe SSDs	HPE Apollo 6500 Gen10
DB-s	128 GB <sup>b</sup>	6	2 × Intel Xeon E5-2695 v3 + SSD	HPE ProLiant DL360
DB-h	128 GB <sup>b</sup>	6	2 × Intel Xeon E5-2695 v3 + HDDs	HPE ProLiant DL380
Web	128 GB <sup>b</sup>	6	2 × Intel Xeon E5-2695 v3	HPE ProLiant DL360
Other <sup>a</sup>	128 GB <sup>b</sup>	16	2 × Intel Xeon E5-2695 v3	HPE ProLiant DL360, DL380
Gateway	64 GB <sup>b</sup>	4	2 × Intel Xeon E5-2683 v3 (14c, 2.0/3.0 GHz, 35MB LLC)	HPE ProLiant DL380
	64 GB <sup>c</sup>	4	2 × Intel Xeon E5-2683 v3	
	96 GB <sup>d</sup>	2	2 × Intel Xeon	
Storage	128 GB <sup>b</sup>	5	2 × Intel Xeon E5-2680 v3 (12c, 2.5/3.3 GHz, 30 MB LLC)	Supermicro X10DRi
	256 GB <sup>c</sup>	15	2 × Intel Xeon E5-2680 v4 (14c, 2.4/3.3 GHz, 35 MB LLC)	
<b>Total</b>	<b>286.5 TB</b>	<b>920</b>		

Bridges-DL

- a. Other nodes = front end (2) + management/log (8) + boot (4) + MDS (4)
- b. DDR4-2133
- c. DDR4-2400
- d. DDR4-2666

# *Getting Time on XSEDE*

# XSEDE

Extreme Science and Engineering  
Discovery Environment

<https://portal.xsede.org/web/guest/allocations>

even simpler:

<https://psc.edu/apply-for-an-xsede-startup-grant>

# Getting Connected

The first time you use your account sheet, you must go to [apr.psc.edu](http://apr.psc.edu) to set a password. You may already have done so, if not, we will take a minute to do this shortly.

We will be working on [bridges.psc.edu](http://bridges.psc.edu). Use an ssh client (a Putty terminal, for example), to ssh to the machine.

At this point will be on a login node. It will have a name like “login001” or “login006”. This is a fine place to edit and compile codes. However we must be on compute nodes to do actual computing. We have designed Bridges to be the world’s most interactive supercomputer. We generally only require you to use the batch system when you want to. Otherwise, you get your own personal piece of the machine. For this workshop we will use

```
interact
```

to get a regular node of the type we will be using with Spark. You will then see name like “r251” on the command line to let you know you are on a regular node. Likewise, to get a GPU node, use

```
interact -gpu
```

This will be for our TensorFlow work tomorrow. You will then see a prompt like “gpu32”.

Some of you may follow along in real time as I explain things; some of you may wait until exercise time, and some of you may really not get into the exercises until after we wrap up tomorrow. It is all good.

# Modules

We have hundreds of packages on Bridges. They each have many paths and variables that need to be set for their own proper environment, and they are often conflicting. We shield you from this with the wonderful modules command.

You can load the two packages we will be using as

## *Spark*

```
module load spark
```

## *Tensorflow*

```
module load tensorflow/1.5_gpu
```

If you find either of these tedious to repeat, feel free to put them in your `.bashrc`.

# Editors

For editors, we have several options:

- emacs
- vi
- nano: use this if you aren't familiar with the others

For this workshop, you can actually get by just working from the various command lines.

# Programming Language

- We have to pick something
- Pick best domain language
- Python
- But not “Pythonic”
- I try to write generic pseudo-code
  - If you know Java or C or R, etc. you should be fine.



Warning! Warning!

Several of the packages we are using are very prone to throw warnings about the JVM or some python dependency.

We've stamped most of them out, but don't panic if a warning pops up here or there.

In our other workshops we would not tolerate so much as a compiler warning, but this is the nature of these software stacks, so consider it good experience.

# Our Setup For This Workshop

After you copy the files from the training directory, you will have:

/BigData

/Clustering

/MNIST

/Recommender

/Shakespeare

Datasets, and  
also **cut and  
paste code  
samples** are in  
here.



# Code of Conduct

XSEDE has an external code of conduct for XSEDE sponsored events which represents XSEDE's commitment to providing an inclusive and harassment-free environment in all interactions regardless of gender, sexual orientation, disability, physical appearance, race, or religion. The code of conduct extends to all XSEDE-sponsored events, services, and interactions.

Code of Conduct: <https://www.xsede.org/codeofconduct>

## Contact:

- Event organizer: Tom Maiden, PSC ([tmaiden@psc.edu](mailto:tmaiden@psc.edu))
- XSEDE ombudspersons:
  - Linda Akli, SURA ([akli@sura.org](mailto:akli@sura.org))
  - Lizanne Destefano, Georgia Tech ([lizanne.destefano@ceismc.gatech.edu](mailto:lizanne.destefano@ceismc.gatech.edu))

# Preliminary Exercise

Let's get the boring stuff out of the way now.

- Log on to apr.psc.edu and set an initial password if you have not.
- Log on to Bridges.

```
ssh username@bridges.psc.edu
```

- Copy the Big Data exercise directory from the training directory to your home directory.

```
cp -r ~training/BigData .
```

- Edit a file to make sure you can do so. Use emacs, vi or nano (if the first two don't sound familiar).
- Start an interactive session.

```
interact
```