PITTSBURGH SUPERCOMPUTING CENTER

# Pittsburgh Supercomputing Center Overview June 2017

James A. Marsteller
Chief Information Officer
Pittsburgh Supercomputing Center

# The Pittsburgh Supercomputing Center

- A University-based computing and research organization. A joint effort of Carnegie Mellon University and the University of Pittsburgh
- Established in 1986 with federal, state and corporate support.
- Currently staffed by 75 scientists, communications and computing professionals.
- Supported by NSF, NIH, the Commonwealth of Pennsylvania, foundations, DOE, DoD, and industry

*PSC offices are located at 300 S. Craig St., on the campus of Carnegie Mellon University in Pittsburgh, Pennsylvania.*

# PSC: 30+ years of national leadership in:

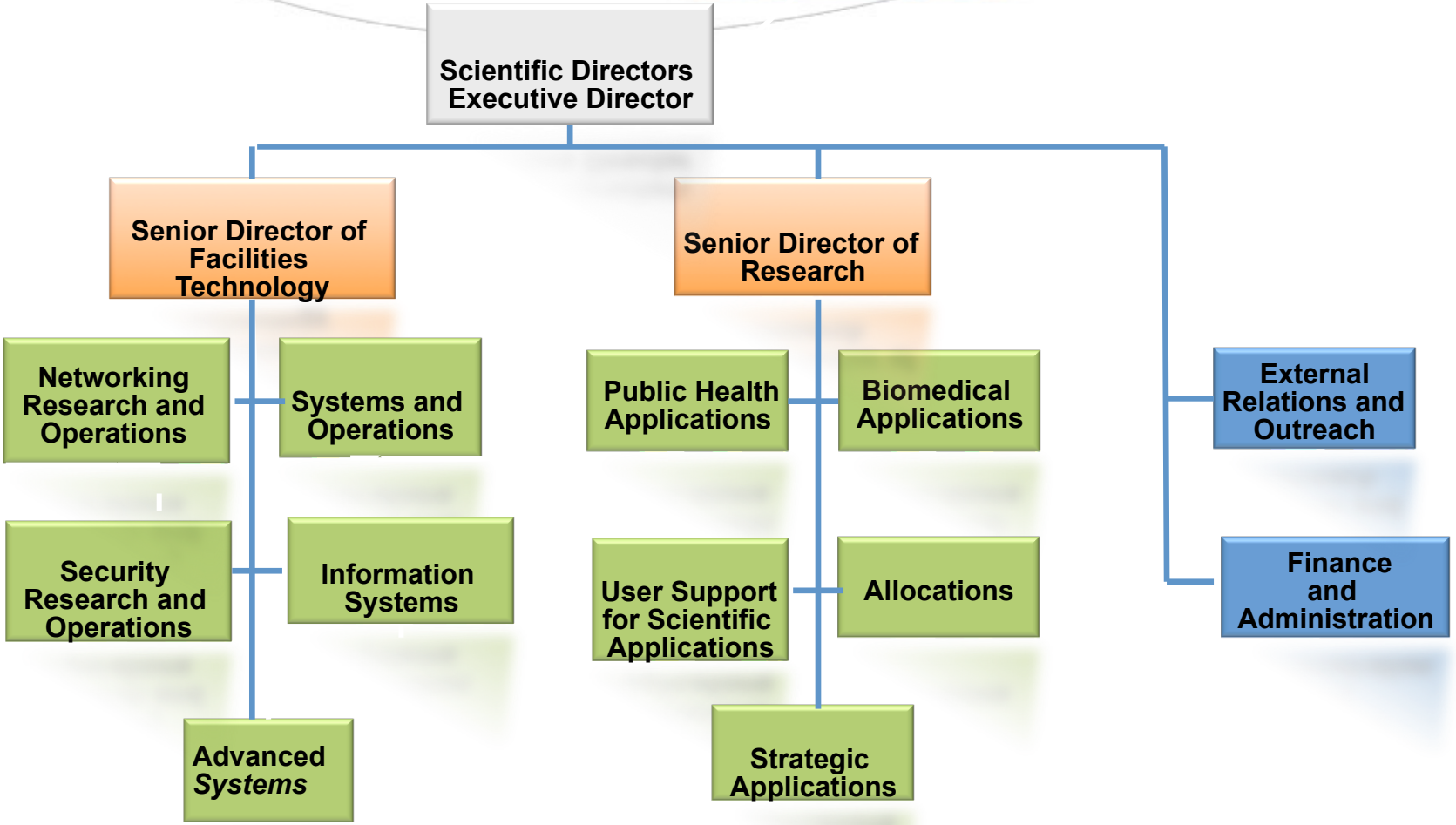
- High-performance and data-intensive computing
- 19 supercomputers, including 8 that were/are “serial #1”
- Data management technologies
- Software architecture, implementation, and optimization
- Enabling researchers nationwide
- Networking and network optimization
- Enabling ground-breaking science, computer science, and engineering
- Leading research in biology/biomedicine, epidemiology, neuroscience, filesystems, networking

# Mission

- Provide **integrated, flexible environment** for solving large-scale computational and data problems
- **Advance science** and technology through collaborative and internal research
- **Educate researchers** about benefits of supercomputing and train them in High Performance Computing (HPC)
- **Improve competitiveness** of industry through use of computational science
- **Fortify** local and national **education improvement efforts** with innovative outreach programs

PITTSBURGH SUPERCOMPUTING CENTER

# PSC Organization



# What Is a Supercomputer?

- Lots of definitions possible.
- One is it is a computer offering the performance that will be possible in a commodity computer in 10 years
- Another: a computer that costs $20 million
- More seriously, it's a computer that operates at the limit of current technology with regard to speed and/or memory, *and can solve a single computational problem too large to otherwise be accomplished*
- The latter rules out things like Amazon's and Google's huge clusters, though they're supercomputers by many definitions

# It Is, by Design, a Scale that Slides with Time

- A new iPhone has as much power as PSC's first supercomputer did in 1986!
- Cost differential: $18 million for the supercomputer, $200 for the iPhone (unadjusted dollars!)

# Supercomputing In Your Pocket

- We all have daily access to phenomenal computing power — a kind of public supercomputing
- Still, there's a series of machines currently operating that do something even more amazing

# PSC Major Computing Resources


**Bridges** is a uniquely capable resource for **empowering new research communities** and **bringing together HPC and Big Data.**

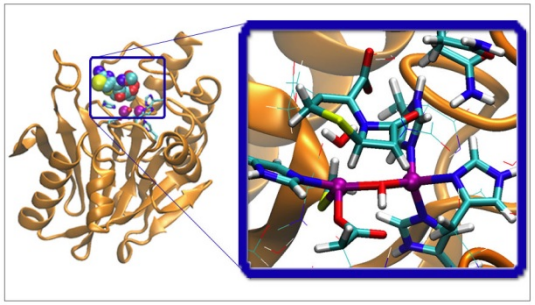
**Bridges'** richly-connected set of interacting systems offers exceptional flexibility for **data analytics, simulation, workflows and gateways**, leveraging interactivity, parallel computing, Spark and Hadoop.

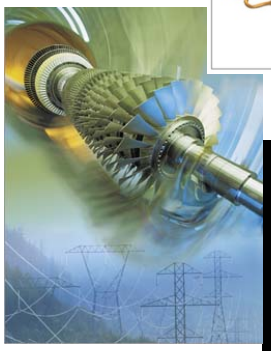
**Anton** is a special purpose resource for biomolecular simulation designed and constructed by D. E. Shaw Research.

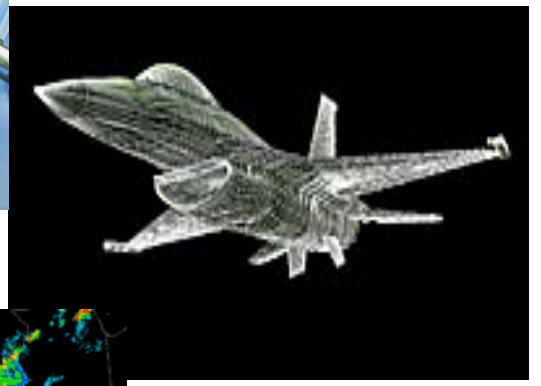
**Anton** was designed to dramatically **increase the speed of molecular dynamics** (MD) simulations, allowing biomedical researchers to **understand the motions and interactions of proteins** and other biologically important molecules over much longer time periods than was previously possible.

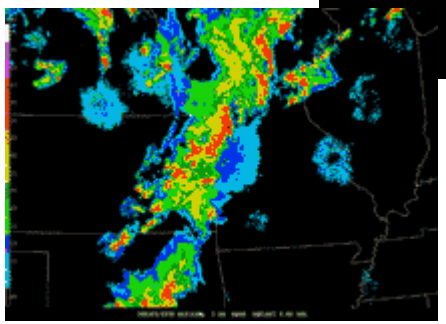
# Applications of High Performance Computing

- Molecular Biology
  - how proteins unfold when stretched
- Public Health
  - Models of disease spread
- Materials Science
  - analyze properties of new materials
- Storm modeling
  - more advance notice of major storm paths
- Earthquake simulation
  - Wave motion/input for building codes
- Origins of the universe
  - role of black holes







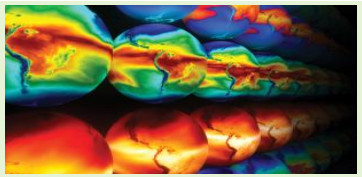


# The Shift to Big Data

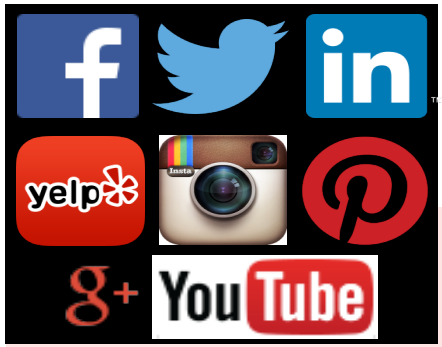
## *New Emphases*

**Pan-STARRS telescope**
http://pan-starrs.ifa.hawaii.edu/

**Genome sequencers**
**(Wikipedia Commons)**



**NOAA climate modeling**
http://www.ornl.gov/info/ornlreview/v42_3_09/article02.shtml



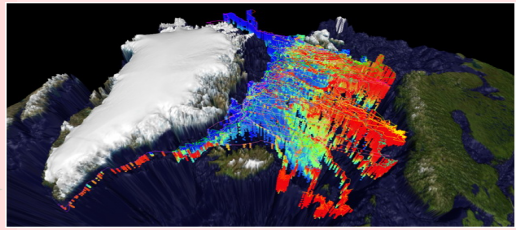

**Social networks and the Internet**

**Video**
Wikipedia Commons

**Library of Congress stacks**
https://www.flickr.com/photos/danlem2001/6922113091/

**Collections**
Horniman museum: http://www.horniman.ac.uk/get_involved/blog/bioblitz-insects-reviewed

**Legacy documents**
Wikipedia Commons



**Environmental sensors: Water temperature profiles from tagged hooded seals**
http://www.arctic.noaa.gov/report11/biodiv_whales_walrus.html

*Structured, regular, homogeneous*

***Unstructured, irregular, heterogeneous***

# Big Data

**Big Data** refers to collections of data sets so large and complex that it becomes difficult to process them in traditional ways.

**Examples** of Big Data include

- Big Science: Genomics, astronomy, Large Hadron Collider, …
- social networks,
- military surveillance
- photography and video archives
- large-scale e-commerce: Amazon, ebay, …

**Characteristics:**

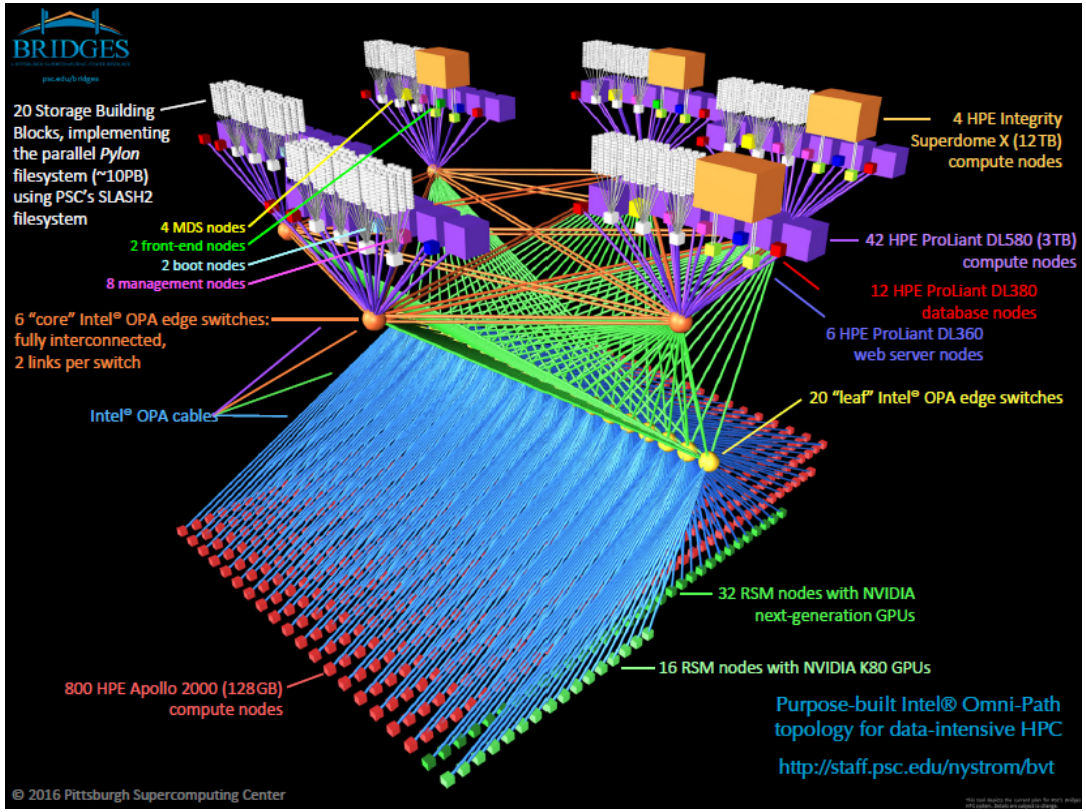
- Too big to handle the way we have been used to doing it.
- Too big to move
- Needs repeated reexamination

# Bridges: A "Custom Supercomputer" on the Fly

A heterogeneous system with nodes optimal for different tasks

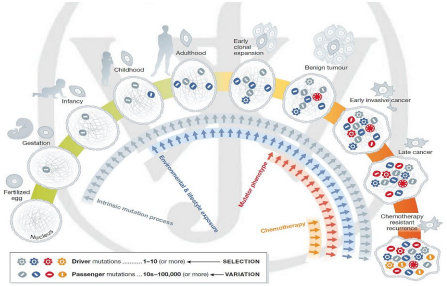
- Hundreds of CPU "compute nodes" for parallel crunching
- Dozens of large-memory nodes (3 TB)
- Several extreme-memory nodes (12 TB)
- GPU nodes for complex graphics-like computations
- Database nodes to allow use of databases

- Intel OmniPath architecture (first use in world) allows dexterous connections so users can tailor the system



# Bridges: Objectives and Approach

- Bring HPC to nontraditional users and research communities.
- Allow high-performance computing to be applied effectively to big data.
- Bridge to campuses to streamline access and provide cloud-like burst capability.
- Leveraging PSC's expertise with shared memory, *Bridges* features 3 tiers of large, coherent shared-memory nodes: 12TB, 3TB, and 128GB.



EMBO Mol Med (2013) DOI: 10.1002/emmm.201202388: ***Proliferation of cancer-causing mutations throughout life***

Alex Hauptmann et. al.: ***Efficient large-scale content-based multimedia event detection***

- *Bridges* implements a uniquely flexible environment featuring interactivity, gateways, databases, distributed (web) services, high-productivity programming languages and frameworks, and virtualization, and campus bridging.

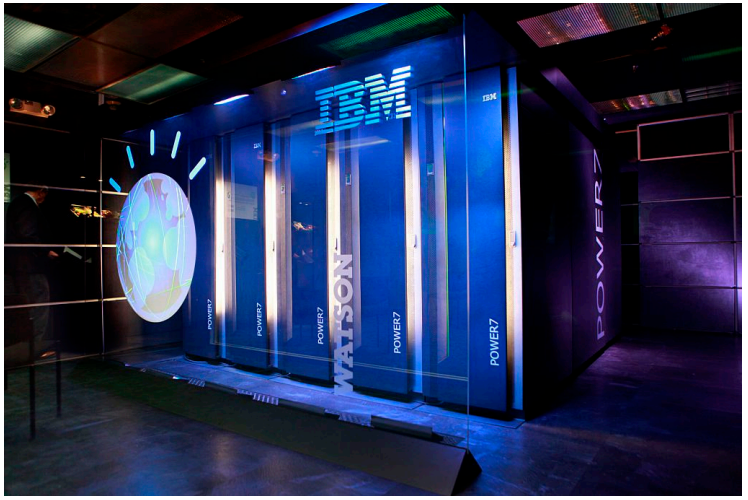
# Bridges: Applications *(Examples)*

- Finding causal relationships underlying various diseases
- Assembling large genomes and metagenomes
- Analysis of financial markets and policies
- Improving the effectiveness of organ donation networks
- Decision making with incomplete information
- Recognizing events and enabling search for videos
- Understanding how the brain is connected from EM data
- Addressing societal issues from social media data
- Analyzing large bodies of work in the digital humanities
- Agent-based modeling for epidemiology
- Cross-observational analyses in astronomy & other sciences
- Data integration for history, political science & cultural studies

# Life is not Chess

Tuomas Sandholm, Carnegie Mellon University

- Often need to make decisions with unknown or even deceptive information
- This makes real-world problems far more complicated than chess or Go
  - Trade negotiations
  - Treating diseases like cancer that react to treatment
  - Anticipating terror attacks
  - Military surveillance




- Other players have “hidden pieces” we need to anticipate
- In that, it’s far more like poker, e.g., No-limit Texas Hold-em

➢ Develop an Artificial Intelligence App in this type of decision-making

# No-Limit Texas Hold'em — Impossible to Calculate

- Poses a particular problem
  - Many bidding possibilities lead to $10^{161}$ situations (information sets)
  - More than atoms in the Universe
  - Way beyond any ability to calculate moves exactly
- An AI approach has to simplify choices
  - For example, pair of Jacks = pair of Queens
  - But those errors will multiply with each hand
  - Best to be as fine-grained as possible

# Taking on the Card Sharps

- In 2015, CMU group developed an AI-based No-limit Texas Hold'em poker player, Claudico based on their research
  - Claudico played on a commodity computer, using the PSC supercomputer Blacklight to formulate strategy and continue learning
  - Claudico beat all AI comers in 2015
  - It battled four top-10 human players in a tournament hosted by the Rivers Casiono to a statistical draw
- In 2017, the group developed Libratus with new strategies for bluffing and more optimal ways to simplify the game
  - Libratus used PSC Bridges supercomputer to plot each move in real time, and reformulate its strategy each night
  - In a rematch with four top human players, Libratus won decisively
- Research can be applied to any imperfect information game

# External Relations and Outreach

Government relations, at state and local level

Community relations (economic development, foundation, education)

K-12 STEM programs, including

- HS bioinformatics curriculum
- Pittsburgh Data Jam
- GCode – a new program to introduce coding to middle/high school girls of color
- SAFENet – resources to raise internet safety awareness

PSC Corporate Affiliates Program

All external communications, including

- Print, website and social media
- PSC Science Highlights magazine an electronic bi-annual review of recent PSC projects

# Education Outreach Programs

- **BEST** (Bioinformatics Education for Students - introduces teachers and their students to modern molecular biology concepts by incorporating computational biology and bioinformatics into their high school curriculum. BEST prepares teachers to introduce their students to emerging and exciting biomedical careers

- **Project GCODE** - introduces girls in grades 6 through 12 to basic programming concepts to build awareness of 21st century careers in technology and cyberscience. With particular emphasis on girls of color, participants learn basic programming structures as they create simple mobile apps, working closely with women instructors and mentors.

# Education Outreach Programs

- **Pittsburgh Data Jam** - is a high school competition to introduce Pittsburgh-area students to big data and data analytics. A program of Pittsburgh Dataworks, the Data Jam raises awareness about he impact of big data on students' lives and introduces them to career options created by the advent of big data. The Data Jam involves teams of students who select a city-centric problem to work on, identify data sources, conduct the analysis and present their results in the final poster presentation.

# PSC STEM Resources (available on-line)

**CAST** (Computation and Science for Teachers) - introduces an overall computational approach to teaching HS science; train teachers in using computational models/simulations to extend student understanding of the processes and theories of science

**CMIST** (Computational Modules In Science Teaching)- provides specific HS science teaching modules from PSC biomedical modeling and simulations research

**SAFE-Net** (Safety Awareness for Everyone on the Net) - focuses on raising the awareness of students, parents, and educators about cyber threats, measures of protection, and cyber ethics. The SAFE-Net website (https://safenet.3rox.net/ provides materials to teach parents, teachers and students about cyber security issues.

# Corporate Affiliates Program

## Collaborative Affiliations

- Alcoa
- PPG
- Phillips Respironics
- Bechtel Bettis
- ANSYS
- Medrad
- Westinghouse


The **Three Rivers Optical Exchange** (3ROX.net), a high-speed network hub, operates and manages network infrastructure that connects many universities and schools in Pennsylvania and West Virginia to research and education networks.

- Advanced Networking Services (Internet2, ESNet, Netflix/ Google Peering)
- Network Research Resource Pool
- Security as a Service (SECaaS)

# Questions

Find us on

and stay in touch!