Research computing at CSU

59 respondents to survey

Heterogeneous community

There is HPC Activities in all 8 colleges

Computing being done by Professional staff as well as grad students & post-docs

Users need fast processing, or large memory, or large disk (or 2 out of 3)

Application programs being used are homegrown, or open source, or commercial
Research computing at CSU

Majority of computation at CSU (by # of people or grant funding) is in support of experimental programs
Computation a critical part of a project, not the project

There is a significant developmental presence across the University

Active 53 funding (from VPR database)
Computing, modeling, simulation, bioinformatics... $78M
For reference:
Solar, biofuels, methane, wind, powerhouse... $41M
Tuberculosis $46M
Total CSU $1.1B
Application areas include:

- Statistics
  - large datasets
  - large number of datasets
- Simulation
  - greenhouse gas emissions
  - financial modeling
  - large-scale models in ECE
  - large scale network
  - molecular (bio, nano, polymer)
  - MCNPX Monte Carlo simulations
  - resource management for HPC systems
  - data center energy and thermal properties
- Atmospheric, fire, and ecological sciences
- Ecosystem modeling
- Geosciences
- Robotics
- Machine Learning
  - Exercise science
  - Atmospheric science
  - Biomedical engineering
- Distributed Computing and Big Data
  - Atmospheric Science
  - Epidemiology
  - Healthcare
- Smart Grid research
- Power systems engineering
- Computational Electromagnetics
- Civil Engineering (ordinary & partial differential equations)
- Bioinformatics
  - Genomics
    (short-reads analysis, sequence analysis, etc.)
  - Transcriptomics
  - Proteomics
  - Metabolomics
  - Population genetic evaluation
  - Structural RNA, RNA-RNA interaction prediction
  - Protein structure determination
- Math
- Physics
Needs/usage

Disk (TB up to PBs)

Processors
fast serial, single node-multicore,
Fast interconnect up to 1000s of cores, GPUs, accelerators

Ram (16 GB up to TB)

most applications don’t need to be/aren’t here

Cray
2,016 CPU cores
2.5 TB RAM (32 GB/node)
32 TB disk

Based on survey (hardware that people use)
applications tend to need/use
large memory (TB)
or large disk (PB)
or fast parallel (1000s of cores/GPUs)

Perhaps due to non-linearity of pricing with speed
Applications programs

open source > commercial > roll you own (based on # of apps)

Hardware needs generally dictated by software “vendor”
a number of applications take advantage of parallel processing but many can not, yet
Computers

Individual/group
20%/60%

Department/college
30%/30%

Cray/National resource
50%/10%

Based on number of cores
Based on number of active users

Work is getting done, grants are being funded, but more resources are needed to remain competitive.
Support needs

Heterogeneous User base

Professional staff & Faculty
busy doing their day job
need help in small bites

Grad students and post-docs
large learning curve
users & developers

consulting
scripting (python, matlab)
courses (discipline-specific & computing)
consulting
scripting (python, matlab)

Personal aside:
If programming assistance were available, perhaps:
scripts->applications
serial->parallel
Courses

ANEQ 575 Computational Biology in Animal Breeding.
BC 441 3D Molecular Models for Biochemistry.
CIVE542 Water Quality Modeling.
CIVE556 Seepage and Earth Dams.
CIVE607 Computational Fluid Dynamics.
CIVE631 Computational Methods in Subsurface Systems.
CS475 Parallel Programming
CS570 Advanced Computer Architecture.
CS575 Parallel Processing.
GRAD511 High Performance Computing and Visualization.
MECH650 Computational Materials from First Principles.
NB650 Computer Analysis of Neuronal Proteins.
SOCR731 Plant Breeding Data Management.
Bioinformatics Courses

BSPM 576/MIP 576 Bioinformatics.
Technical computing across platforms using bioinformatics tools in molecular analyses.

CS 425 Introduction to Bioinformatics Algorithms.
Algorithms for analysis of large scale biological data.

CS 548/STAT 548 Bioinformatics Algorithms.
Computational methods for analysis of DNA/protein sequences and other biological data.

CS 646 04(3-2-0). Machine Learning in Bioinformatics.
Recent research on the supplications of machine learning in bioinformatics.
Summary

Work is getting done, grants are being funded, but more resources are needed to remain competitive.

Majority of computation at CSU is in support of experimental programs. A critical part of a project, not the project.

There is a significant developmental presence across the University (that can be tapped through GSAs).

Hardware needs generally dictated by software “vendor”

- a number of applications can take advantage of parallel processing but many can not, yet

Applications tend to need/use

- large memory (TB)
- or large disk (PB)
- or fast parallel (1000s of cores/GPUs)

Support needed

- Courses (discipline-specific & computing)
- Consulting
- Scripting (python, matlab)
My personal opinion:
We need a heterogeneous environment modeled after campus usage/need with
a) large ram subsystem (several nodes)
b) gpu-rich subsystem
c) subsystem with large (cheap) scratch
d) subsystem with large number of nodes with fast interconnect
e) distributed nodes & fast communication to (pre)process where data is being generated

Over time users could/would add notes to the subsystem type(s) that they need

Need sophisticated queuing system and larger support staff

HPC Support staff should include GSAs in parallel programming and discipline-specific computing (eg. bioinformatics, statistics)

A lot of work going on a CSU, very grassroots
Central leverage could help funding success.
Grants tend to be "modular".
For example, in chemistry one can ask for $450K (not going to get much more), you can either buy hardware or pay people. It's better for CSU if you pay people (hardware has no overhead)
What do you think we need/what will help your research?

What should be in central infrastructure?

Software site licenses?

Support infrastructure?