Yale Center for High Performance Computation in Biology and Biomedicine

Robert Bjornson, Ph.D.
Nicholas Carriero, Ph.D.

NIDA Neuroproteomics Center
Advisory Committee Meeting
Dec 3, 2008

Providing HPC support for Proteomics/Bioinformatics

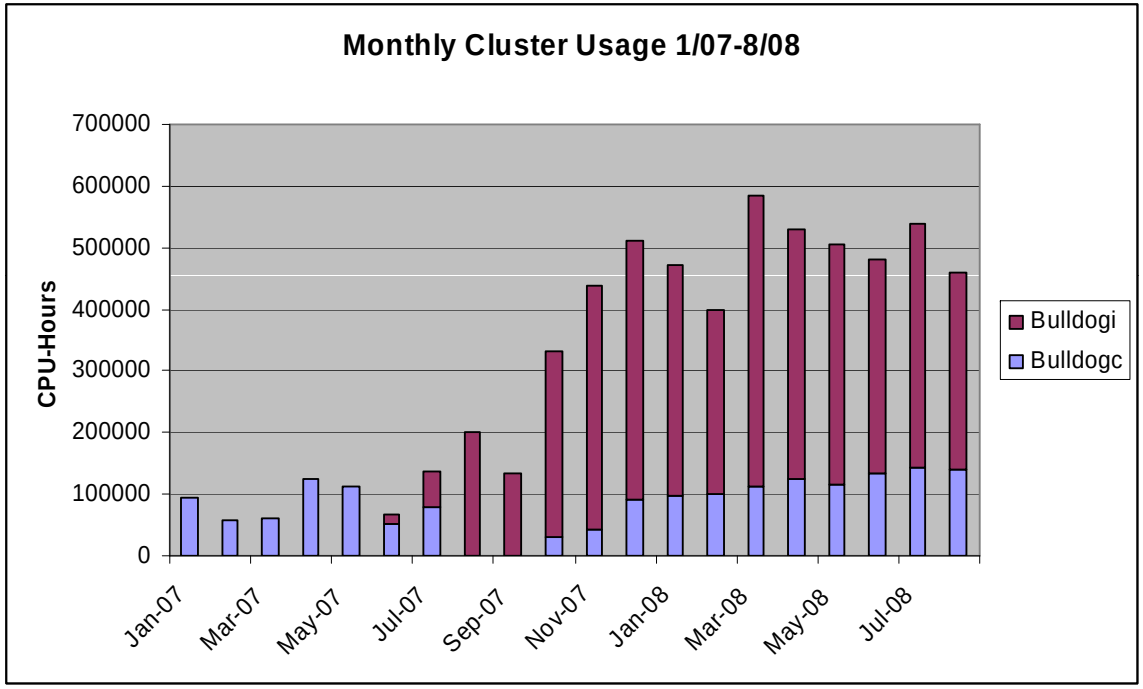
Goal: *Eliminate computing as the rate limiting step while preserving researchers' comfort factor with their codes.*

- From a computer scientist's perspective, the goal is "modest."
- From the researcher's perspective, the goal, by construction, meets computing needs without disrupting R&D effort.
- Not always, but surprisingly often, possible to reconcile these two.

Bulldogi Cluster

Center Statistics

- Large linux cluster, assorted other computers
- 300 compute nodes, 1000 cores (cpus)
- 100 TB File Space
- 500,000 cpu-hours/month
- 130 users


Monthly Cluster Usage 1/07-8/08
CPU-Hours
700000
600000
500000
400000
300000
200000
100000
0
Jan-07
Mar-07
May-07
Jul-07
Sep-07
Nov-07
Jan-08
Mar-08
May-08
Jul-08
Bulldogi
Bulldogc

Drug Fingerprinting Project Challenges

- Atypical HPC project?? *Not Really; becoming typical*
  - Data intensive
  - Substantial data “wrangling” pre- and post-computation
  - Computation via scripting language (Matlab)
  - Open Source software

- High-throughput Scaling:

  Difficult transition from manual processing on Windows PC

  to automated processing on Linux cluster

Drug Fingerprinting Project Challenges (2)

- Huge amount of raw data

  Typical screen: 192 plates x 384 wells x 4 sites x 3 wavelengths

  → ~1M image files, @ 2.9 MB → ~3 TB

  Due to network limitations, used portable USB hard drives

- Raw data massaging
  - Image files scattered and split across many directories by imaging instrument
  - Numerous missing, incomplete, truncated or duplicated files

    Used python scripts to consolidate and error-check files

Drug Fingerprinting Project Challenges (3)

- Image processing via CellProfiler (Broad Institute)
  - Matlab based, GUI and Batch versions

- Substantial Computation

  ~300,000 image sets. Each requires ~100sec

  → ~1 cpu-year

  In theory, each set is independent.

  Can be parallelized on ~100 CPUs of bulldogi cluster : ~4 days.

Drug Fingerprinting Project Challenges (4)

- Substantial Post-processing
  - Multiple result files must be merged properly
  - Raw results: ~40 objects (cells) per image set @ ~800 measurements
    → 12M x 800 matrix (70-80 GB file)

- Too large for direct manipulation by statistical packages.

- Use python and “out of core” techniques to postprocess file to manageable proportions.

Postprocessing steps (Python again)

- Sanity check output
- Detect and patch artifacts in output file
- Remove unwanted columns
- Reduce file to more reasonable size via averaging
- Convert to binary file for loading in R
- Parallel machine learning (Random Forest)

Summary

- Data handling more challenging than computation
- Multi-step process
- Sanity checking critical, difficult
- Python invaluable:
  - Scanning and organizing large directory trees
  - Manipulating huge files
  - Automating parallel processing, combining results
- Driven by needs of researcher and statistician