Clusters, High-Performance Computing, and Distributed Systems at Earlham

Presented by Craig and the Admins
Earlham College Computer Science

What are HPC systems?

- Large computers capable of running complex workflows on large datasets
- Common software
  - Operating system: Linux (or other UNIX-like system), in our case CentOS 7 or 8
  - Shell: HPC is mostly (but not exclusively) navigated using the Linux command line and a shell
  - Scheduler: in our case Torque (qsub), but others exist e.g. Slurm
  - Modules: load specific versions of particular software packages
  - File system mounting: access to user directories and other (relatively) small files across multiple machines
  - Languages: C (FORTRAN also works but at ECCS we tend to work in C), Bash
- Domain-specific software
  - Bioinformatics: mothur, samtools, bowtie2, blast, boost, ea-utils
  - Chemistry: Gaussian + WebMO
  - Computational biophysics: AMBER

Why use an HPC system?

- We want to solve big problems.
- Solving big problems in the sciences requires large datasets and complex workflows.
- That work takes time.
- Time is expensive.
- The primary benefit of HPC systems is to save **time**.

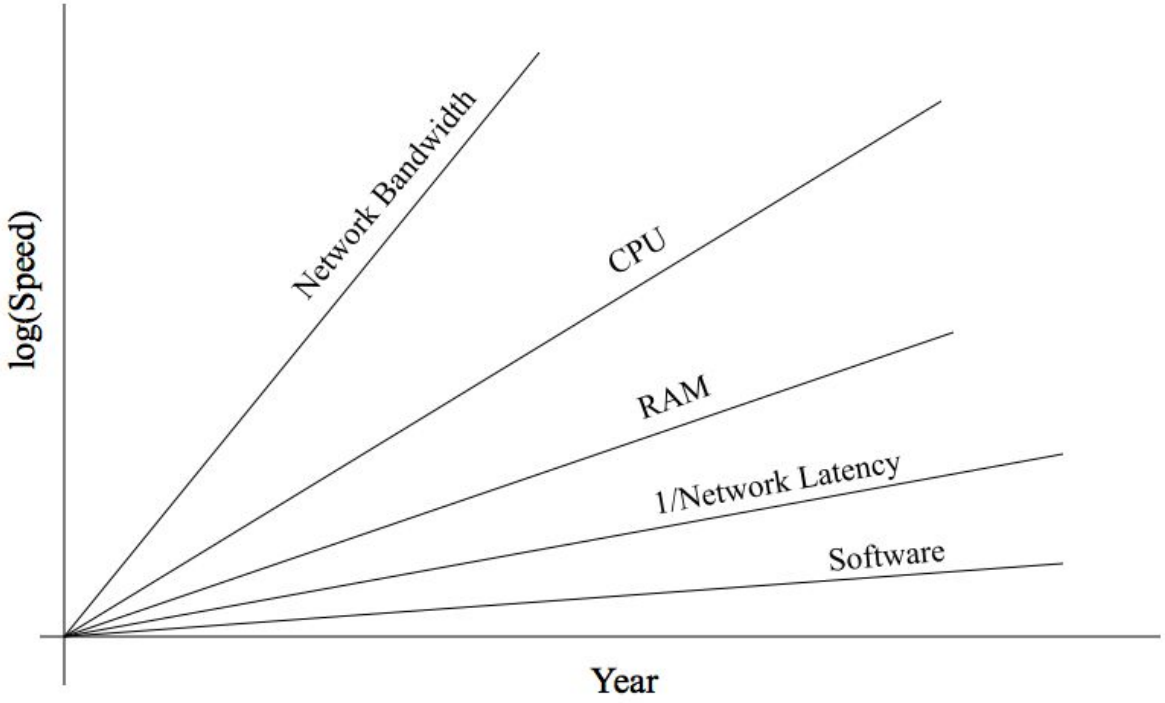
Types of HPC's

- Cluster: series of machines linked through a head node and a network switch
  - Common applications: Chemistry, Computational Biophysics
  - Textbook example of a **distributed system** - specifically, distributed-memory parallelism
  - Layout and Whedon are our current HPC clusters
- Jumbo server: a computer with a high ratio of (Storage+RAM)/Processing
  - Common applications: Bioinformatics
  - Shared-memory parallelism
  - Pollock and Lovelace are our current jumbo servers
- What about this "cloud" thing?

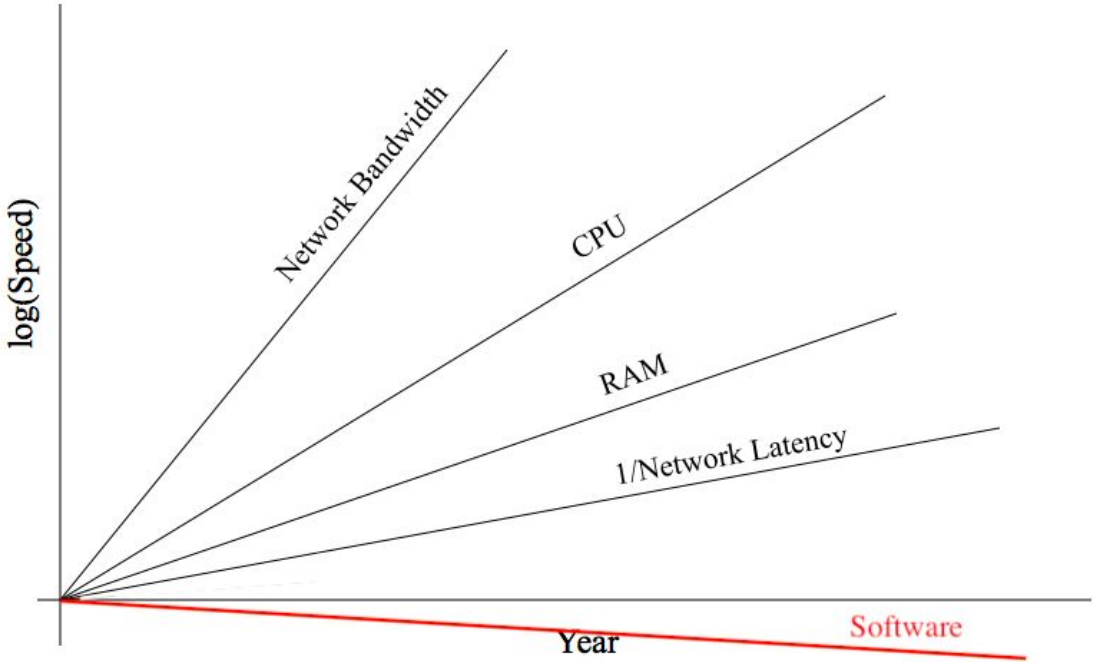
Opportunities to learn HPC at Earlham

- **Get a cluster account**
  - Email: admin @ cs . earlham . edu
  - wiki.cs.earlham.edu has resources for interfacing with our technology
- CS 360: Parallel and Distributed Computing
- Icelandic Field Studies
  - Collect field samples, which are then analyzed on-site
- System Administration Applied Group
  - Installs, configures, and retires computing resources including HPC clusters and phat nodes
- Senior projects
  - Machine learning lends itself to large resources
  - Projects in any of the above categories can become the basis for a senior project

Moore's Law in Practice

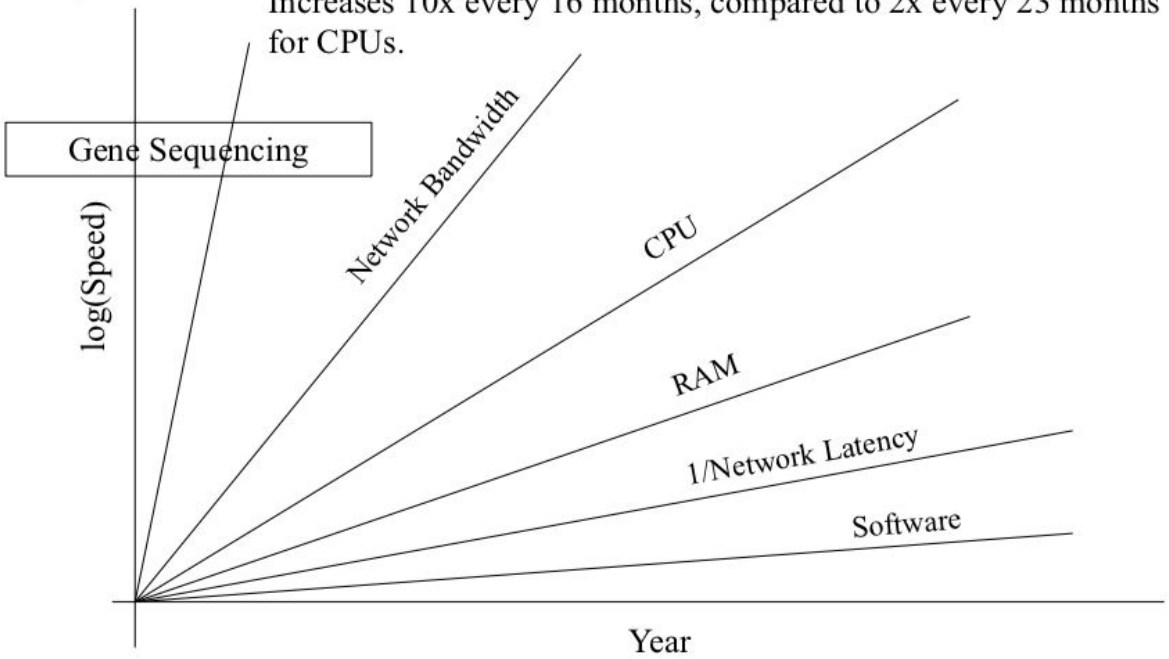


Moore's Law in Practice



Moore's Law on Gene Sequencers

Increases 10x every 16 months, compared to 2x every 23 months for CPUs.



Example

References and further reading

- Charlie Peck, Earlham CS
- Henry Neeman’s (OSU) “SupercomputingInPlainEnglish” series
- redhat.com, centos.org, debian.org

Thanks!
Questions/comments?