Case Study – Bioinformatics Research and Education at Arcadia University

Vitaly Ford
Assistant Professor
Computer Science and Math Department
Arcadia University, Glenside, PA
## Arcadia University & CSMA

<table>
<thead>
<tr>
<th>Arcadia</th>
<th>CSMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>• 4,000 students</td>
<td>• Math majors: 100</td>
</tr>
<tr>
<td>• Study abroad program</td>
<td>• Computer science majors: 100</td>
</tr>
<tr>
<td>• A few buildings in Europe</td>
<td>• Full-time faculty: 10</td>
</tr>
<tr>
<td>• Colocation in Valley Forge</td>
<td>• Research-oriented faculty: 8</td>
</tr>
<tr>
<td>• Limited SAN</td>
<td>• Clusters/HPC: none</td>
</tr>
<tr>
<td>• 5G with Cogent (burstable to 10G)</td>
<td>• Labs: 3 rooms, ~20 PC/Mac each</td>
</tr>
<tr>
<td>• It was 1G just a few weeks ago</td>
<td></td>
</tr>
<tr>
<td>• 10G with KINBER incoming</td>
<td></td>
</tr>
<tr>
<td>• Data Science programs in progress</td>
<td></td>
</tr>
</tbody>
</table>
Bioinformatics

• Interdisciplinary minor
  • Co-taught by three faculty (2 CS and 1 Biology)
• Students are mostly from CS, Math, or Biology
• Some of the core courses
  • Bioinformatics, Intro to Data Mining, Computer & Scientific Ethics, Probability, Elementary Statistics, Biology, Programming I
• Some of the electives
  • Biochemistry
  • Artificial Intelligence
  • Advanced Data Mining
  • Intermediate Statistics
Bioinformatics: Education

• Genomics Education Partnership (GEP, ~60 schools)
• Data from National Center for Biotechnology Information (NCBI)
• Public Galaxy server
• Tools: Basic Local Alignment Search Tool, Python, Anaconda, BloPython, Jupyter, GenScan, Tuxedo Suite of Tools, etc.
Bioinformatics: Research

• Illumina Miniseq + just purchased another sequencer
• Undergraduate projects
  • Capstone
  • Research
  • Bioinformatics course
Bioinformatics: Project Examples

• Projects using NCBI and GEP resources
  • DNA of species
  • SNPS in human genome
  • Hormonal response elements in the genome
  • Palindrome occurrences in virus’ RNA
    • RNA viral info
Bioinformatics: Challenges

• Public Galaxy server is slow
• Lab machines cannot do it either
  • 3M RNA to analyze :: 2 days :: 15 GB data => resource exhaustion
• Arcadia
  • Little-to-no storage capacity for genomics data
    • Existing storage does not segregate/limit users, no security
  • No computing capacity to process on premise
  • No ScienceDMZ to transfer big data
  • Multiple single points of failure (authentication, network)
  • Other researchers transfer data using their personal hard drives
Bioinformatics: Solutions

• NSF CC* Network Design Grant (July 9, 2018)
  • ScienceDMZ (DTN + SAN + perfSONAR + Globus)
  • KINBER -> Regional Networks + Internet2 + Cloud

• In progress: administration buy-in into the IT development
NSF CC* Network Design: Arcadia Team

• PI: Leslie Margolis Interim CIO
• Co-PI: Vitaly Ford Assistant Professor
• Science Advisory Committee
  • Associate Provost
  • College of Arts and Sciences Dean
  • College of Health Sciences Dean
• John Zottola, Director of Infrastructure Management
• Office of Sponsored Research and Programs
NSF CC* Network Design Grant

• Science Drivers
  • Bioinformatics
  • Computer Science
  • Chemistry/Physics

• Broader Impacts
  • Communications & Media
  • Physical Therapy (video recordings)
  • Health Science (Fox Chase/Temple Health, UMD, several more)
Back to the Future: 2-5+ years

- Cogent (5-10G) + KINBER (10G) redundant connectivity
- No single point of failure
- AWS (CloudLab, and other) for short-term computing
  - Individual Galaxy servers on AWS per student
- XSEDE/ERN?/VDC? resources
- HPC on premise for long-term computing
Back to the Future: Research Opportunities

• Sequencing and storage on premise
• Secure storage
• Bioinformatics projects that work
• Frictionless connection / data sharing with collaborators
• Publications based on the data generated by the sequencers
• Cloud computing and local storage after computation is done
Future Challenges (2+ years)

• Faculty awareness that ScienceDMZ exists
• Faculty training on Globus, Cloud, etc.
• Faculty one-on-one meetings
Suggestions? Solutions?

Questions?