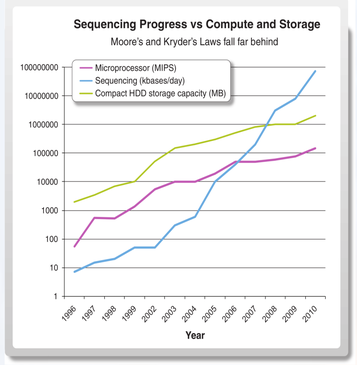
Petascale Informatics Applications Development on XSEDE Supercomputers

Bhanu Rekepalli[1], Paul Giblock[1], Christopher Reardon[1], Subhra Sarkar[2], Mark Fahey[1]
[1]National Institute for Computational Sciences, University of Tennessee, Oak Ridge, TN. [2] Clemson University, Clemson, SC.

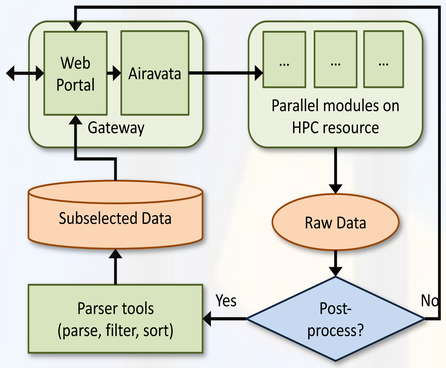
Problem: implementations of bioinformatics tools do not sustain the exponential growth in the genomic data, and HPC resources are difficult to utilize



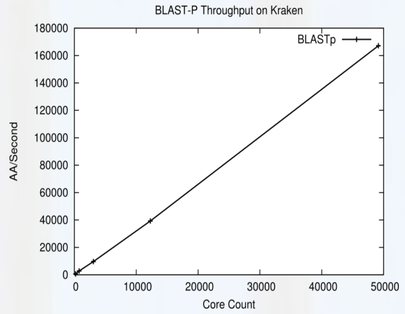
Solution

- **Highly-scalable Parallel Tool Design**
  - BLAST, HMMER, DOCK6, AutoDock, MUSCLE, LINUS
- **Software Modules**
  - Tools available as standard software environment modules on HPC architectures
- **Web-based Science Gateway**
  - Grant researchers simplified access to XSEDE HPC resources. (eg. Kraken)
- **Automated Workflows**
  - Allow for large-scale knowledge discovery
- **Database**
  - Accessible storage for long-term data analysis

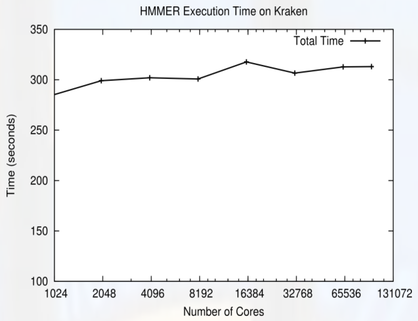
Component Architecture



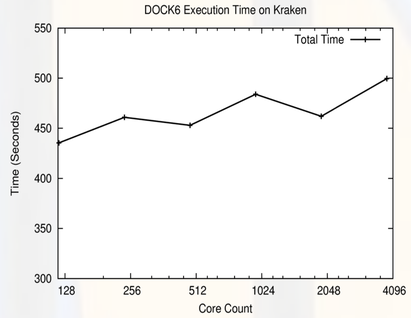
BLAST Scaling Results



HMMER Scaling Results

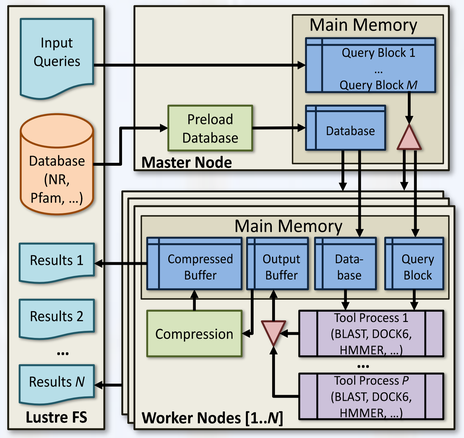


DOCK6 Scaling Results



Highly-Scalable Parallel Wrapper

- Reusable solution scales bioinformatics and cheminformatics tools for HPC environments.
- Critical performance optimizations:
  - Efficient database distribution
  - Dynamic hierarchical load balancing
  - High-throughput buffered parallel output
- Robust fault-tolerance
- Check-pointing and recovery



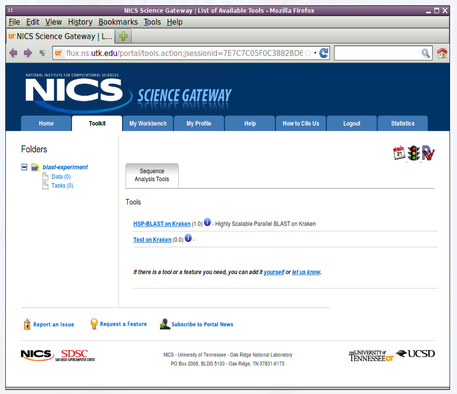
Abstract

We are focusing on two key informatics fields: bioinformatics and cheminformatics, specifically on fast data analysis and retrieval which will have a direct impact on quality of human health. The exponential growth of data generated in biology research, from small atoms to big ecosystems, necessitates an increasingly large computational component for analysis. Data-intensive bioinformatics applications are driven by novel DNA sequencing technologies and the high-throughput approaches that complement them, such as proteomics, genomics, and meta-genomics. At the same time, the number of small carbon-based compounds with molecular masses in the same range as those of living systems is estimated to be over 10^60, which is known as small molecule chemical space. Millions of compounds are available in various chemical libraries that are used for drug discovery and many hypothetical compounds that can be generated or synthesized if the simulation shows good results. Currently there are more than one hundred million compounds in different vendor and academic databases, and an even greater number in private labs. The chemical space is growing very rapidly, thus increasing the demand to speedup cheminformatics applications at large scales. Where individual research centers or universities could previously provide for these applications, this rapid data growth and resulting computational demand is changing the landscape to one where only specialized national centers can deliver the level of computing resources that are now required. We are addressing this demand through the development of massively parallel applications to analyze the growing flood of biological data and chemical space, and to rapidly contribute novel knowledge to the bioinformatics and cheminformatics fields. These analyses and the resulting discoveries will have real, tangible impacts on human health and the environment, for example by personalizing medicine and by creating new sources of renewable energy. Previous National Science Foundation (NSF) projects' efforts provided for the generation of parallel modules for widely-used bioinformatics and cheminformatics applications on the Kraken supercomputer. We have taken desktop and small cluster based applications which are most widely used by scientific community such as NCBI BLAST, HMMER, DOCK6, AutoDock, and MUSCLE; profiled the code; optimized the code and scaled applications to tens of thousands of cores on HPC architectures; and made them robust and portable to next generation architectures along with incorporating these parallel applications in science gateways with a web-based portal. This poster will discuss the various developmental stages, challenges and solutions involved in taking an informatics application from the desktop to petascale with a front-end portal for very-large-scale data analysis in life sciences. This research will help to bridge the gap between the rate of data generation and the speed at which scientists can study this data. The ability to rapidly analyze data at such a large scale is having a significant, direct impact on science achieved by collaborators who are currently using these tools on the Kraken supercomputer.

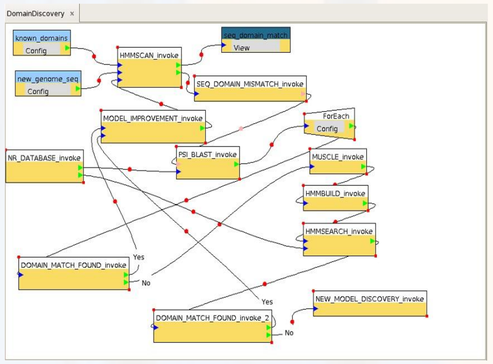
NICS Informatics Science Gateway

- Collaborative effort with SDSC to adapt CIPRES for the NICS Informatics Science Gateway
  - Add handling of data movement for large data sets
  - Incorporate complex computation workflows using Airavata
  - Provide database-supported output analysis tools
- Features:
  - User registry with fine-grained access and usage control
  - Logging of users and usage
  - Job monitoring and statistics tracking
  - Task-based tool parameter configuration
  - Remote job submission via community account
- Production URL: epscor.nics.tennessee.edu
- XSEDE project: Science Gateway for Systems Biology Workbench Portal

Science Gateway Web Interface



Workflow for Novel Domain Discovery



Acknowledgments

This research used resources at the National Institute for Computational Sciences (NICS) funded by The National Science Foundation (NSF) and also supported in part by the NSF grant EPS-091943