UNIVERSITY OF TROMSØ UiT

GeStore – Incremental Computations for Metagenomic Pipelines

Edvard Pedersen[1], Nils Peder Willassen[2], Lars Ailo Bongo[1]

[1]Department of Computer Science, University of Tromsø
[2]Department of Chemistry, University of Tromsø

Email: epe005@uit.no

Life science data processing

- Focus on analysis of metagenomic data
- Meta-data collections are used to compare sequence data to known organisms
  - Updated regularly
  - Important to integrate experimental results with update meta-data
- Must currently recompute entire experiment using complete data
- Computation is expensive
- Sequencing data production grows faster than processing and storage

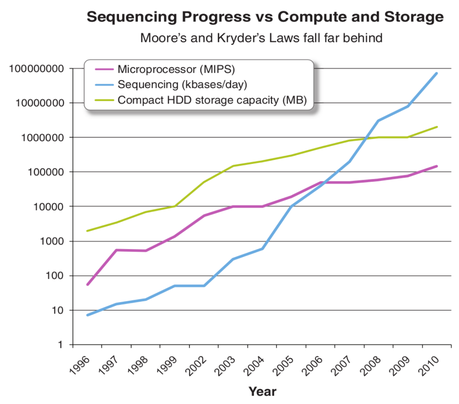

Figure1. DNA sequence data production growth, reprinted from [1]

GeStore

- System for on demand transparent creation of incremental meta-data collections
- Simple integration with metagenomic pipelines
- Extensible through plugin system to support new meta-data collections and file formats
- Distributed data-intensive processing

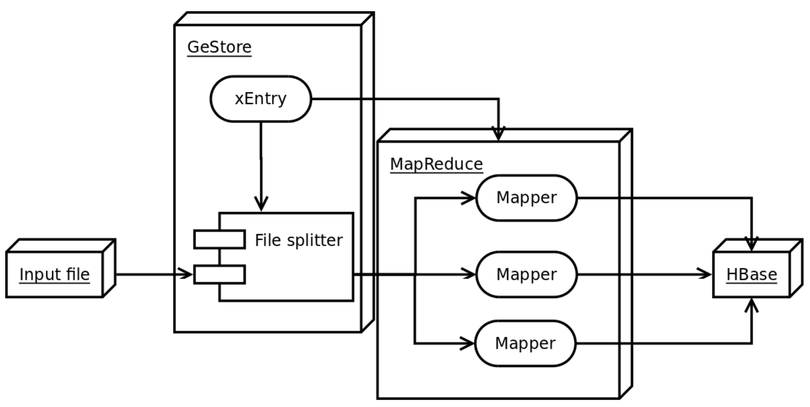

Figure3. Adding a meta-data collection to GeStore

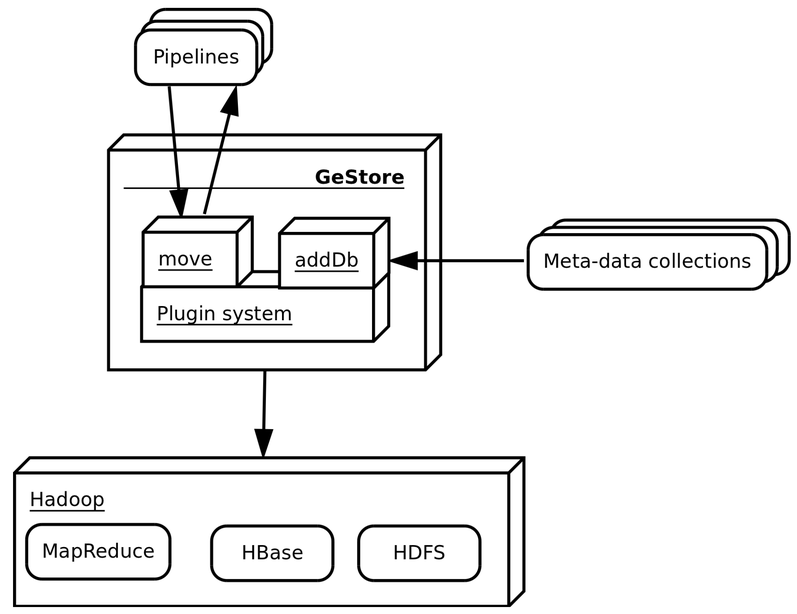

Figure2. GeStore architecture

GePan – Gene Prediction and Annotation

- Metagenomic pipeline system
- Uses several different tools and meta-data collections
- Locally developed by Tim Kahlke
- In use by local biotech company
- No native incremental update support
- Uses Sun Grid Engine for job scheduling

Scientific contributions

- Design and implementation of the GeStore system for incremental computations for metagenomic pipelines
- Experimental evaluation which shows that incremental techniques can reduce the processing requirements for metagenomic pipelines significantly
- Demonstrated the extensibility and simplicity of the GeStore system by integration with the GePan pipeline system and providing support for the UniprotKB meta-data collection

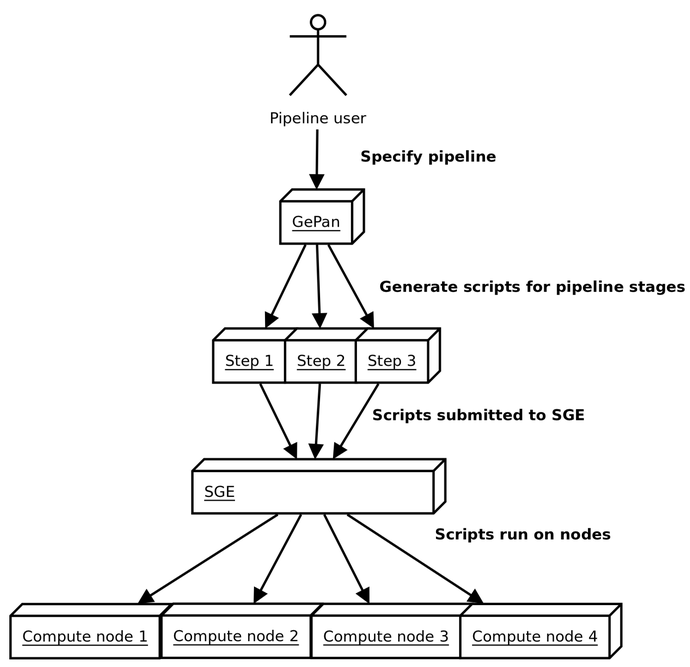

Figure4. GePan pipeline execution

Experimental evaluation

- GeStore used in the analysis of metagenomic data to reduce the size of meta-data collections when doing incremental updates
- Two years of monthly updates of the UniprotKB meta-data collection
- Metagenomic data collected in the Yellowstone national park

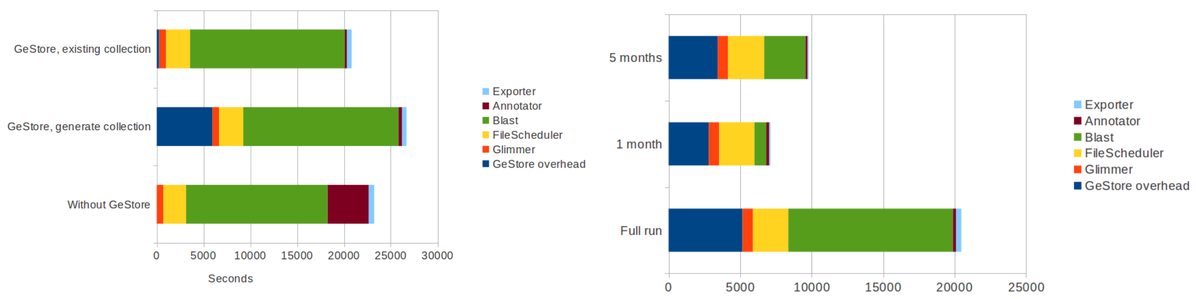

Figure5. GeStore overhead

Figure6. GeStore incremental update savings

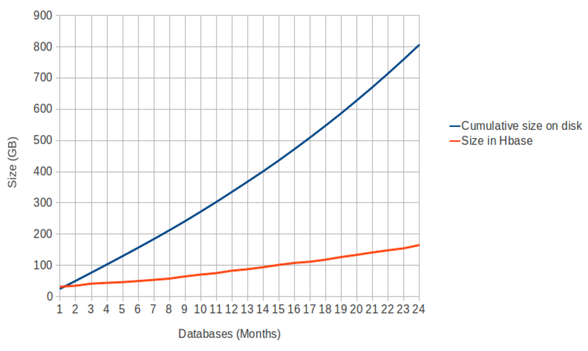

Figure7. Unreplicated storage measurements

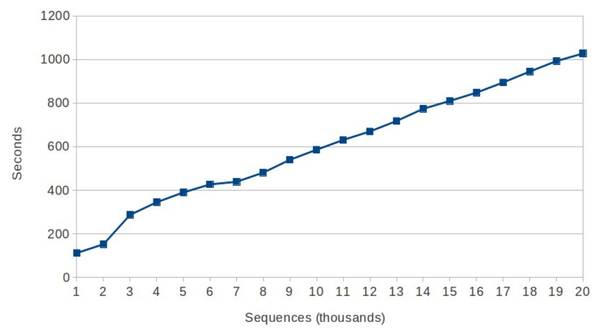

Figure8. BLAST scaling on collection size

How the system works

- Handles file operations for the pipeline system through a simple interface
- Transparently generates incremental meta-data collections when possible
- Replaces the current flat-file system for storage of meta-data collections
- Enables data processing previously considered impractical

Conclusions

- GeStore automatically generates incremental meta-data collections
- Up to 65% reduction in processing resource requirements
- Storage requirements reduced by up to 80%
- Easy to extend the system with new meta-data collections and file formats
- Easy to integrate GeStore with existing pipeline systems to enable incremental computations
- For further information, see [2]

Hardware

5 node cluster
2 quad-core Intel Xeon processors per node
24 GB RAM per node
11 TB network file storage
1.5 TB local storage
16 TB HDFS storage

Big data processing

Cloudera's Hadoop Distribution
Hadoop MapReduce
HDFS
Hbase
Java, Perl and Python

Research groups

This project is a collaboration between the High Performance Distributed Systems (HPDS) group at the Department of Computer Science and the Molecular Biosystems (MB) group at the Department of Chemistry. The HPDS group's research activities includes building and evaluating infrastructure systems for bioinformatics and genomics applications, high performance computing, and display wall systems. The research activities of the MB group is within omics technologies and system biology and includes gene regulation for marine bacteria and marine bioprospecting.

MB, HPDS, and the STAR group at the IT-department will be the partners of the Tromsø ELIXIR node.

Special thanks to Espen Mikal Robertsen, Tim Kahlke, Peik Haugen and Jon Ivar Kristiansen

[1] Kahn, S. D. *On the Future of Genomic Data*. Science, 331(6018):728–729, February 2011.
[2] Pedersen, E. *GeStore – Incremental Computations for Metagenomic Pipelines*. Master Thesis, University of Tromsø, Tromsø, 2012.