Improving the Performance of Bioinformatics Work Flows: Gap Statistic Acceleration

Dr. Ben Greene, CTO, Analytics Engines
Dr. Timothy Davison, VP of Bioinformatics and Biostatistics, Almac

"Using the accelerated gap statistic we have eliminated a substantial bottleneck in our pipeline for developing prognostic and predictive tests for stratified medicine."

Dr. Timothy Davison, VP of Bioinformatics and Biostatistics, Almac

"Accelerating gap statistic enables Almac Diagnostics to guarantee quality and statistical coverage whilst delivering faster results to customers."

Dr. Ben Greene, Chief Technology Officer, Analytics Engines

Customer Profile

The Almac Group is a leading contract research organization which provides comprehensive services over the complete pharmaceutical and clinical lifecycle from research to commercialisation.

Based in the UK and USA they work with over 600 companies worldwide and employ over 3,300 people.

Benefits

- Accelerates decision process from weeks to days or days to hours.
- 4 servers run at the same speed as 44 standard servers, meaning a lower foot print and power consumption for equivalent results.
- Less hands-on time required for setup and file management allowing for more analysis time.

The Challenge

Almac's bioinformatics pipeline is an essential component of Almac's diagnostics development and consultancy services. Improvements in time and quality within the bioinformatics pipeline translate directly to improvements in diagnostics R&D and product development.

Almac had identified a bottleneck in their bioinformatics workflow relating to gap statistic. Gap statistic is used in the discovery and validation of molecular subtypes from high-throughput data, an essential component in their development of prognostic and predictive tests for stratified medicine. The typical run time for the gap statistic was 33 hours which limited the number of concurrent projects and gap iterations that could be achieved in running the pipeline as part of diagnostic product development. An improved solution was needed that coudl provide a scalable way to increase performance, maintain the integrity of the algorithms, and adhere to data security protocols.

The Solution

Analytics Engines worked with Almac to optimize the gap statistic using a drop-in replacement gap statistic in R. This sits alongside the existing Almac workflows and runs on an Analytics Engines half rack-mount server appliance (pictured on right).

The accelerated R function makes parallelized calls to an optimized C++ software layer to perform multi-threaded bootstrapping. Implementation of the bootstrapping core in C++ enables efficient scaling across multiple cores and a more optimal mapping of the algorithm to the underlying hardware. By writing the code so that the algorithm runs in parallel across all available CPU cores it was possible to accelerate performance and reduce the bottleneck.

In addition, the solution also included a fully integrated, user friendly, Graphical User Interface (GUI) which replaces the existing Perl scripts.

The Results

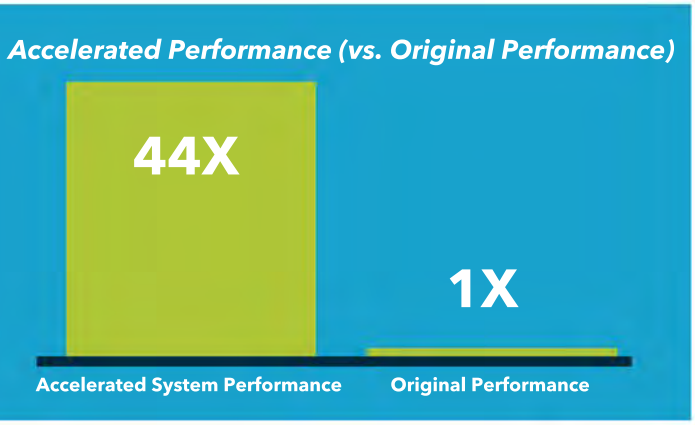
Optimization resulted in a 98% reduction in runtime for the typical full gap pipeline, reducing it from 33 hours down to 45 minutes. This performance improvement enables 44 times as many iterations over the same time period as before optimization, allowing for acceleration of timelines and increased throughput of diagnostic discovery projects.

Achieving the same degree of acceleration by simply scaling out the algorithm across multiple servers would require 44 servers. This Analytics Engine implementation adhered to Almac's data security practices so all information continues to be hosted on site without the associated costs of a large number of servers. For other organizations cloud hosting can also be utilized.

An additional benefit of the implementation was the introduction of a user friendly GUI. This allows for immediate visualisation of results on completion, removing time consuming manual steps and reducing the possibility of errors. Users no longer have to execute each step of the pipeline on the command line, removing a significant barrier to usage and making the pipeline accessible to a wider user base.

Results Overview

- 44X performance acceleration
- Runtime decreased from 33 hours to 45 minutes
- Accelerated decision-making from a weekly to a daily process
- Scalable for higher performance, as needed



For more information contact l.houston@analyticsengines.com

Analytics Engines