DISSERTATION

submitted

to the

Combined Faculty for the Natural Sciences and Mathematics

of

Heidelberg University, Germany

for the degree of

Doctor of Natural Sciences

Put forward by:

Sergei Yakneen (Hon BSc)

Born in Krasnoyarsk, Russia

Oral examination: .....

Modern Systems for Large-scale Genomics Data Analysis in the Cloud

Advisor: Prof. Dr. Michael Gertz

Abstract

Genomics researchers increasingly turn to cloud computing as a means of accomplishing large-scale analyses efficiently and cost-effectively. Successful operation in the cloud requires careful instrumentation and management to avoid common pitfalls, such as resource bottlenecks and low utilization that can both drive up costs and extend the timeline of a scientific project.

We developed the Butler framework for large-scale scientific workflow management in the cloud to meet these challenges. The cornerstones of Butler design are: ability to support multiple clouds, declarative infrastructure configuration management, scalable, fault-tolerant operation, comprehensive resource monitoring, and automated error detection and recovery. Butler relies on industry-strength open-source components in order to deliver a framework that is robust and scalable to thousands of compute cores and millions of workflow executions. Butler's error detection and self-healing capabilities are unique among scientific workflow frameworks and ensure that analyses are carried out with minimal human intervention.

Butler has been used to analyse over 725TB of DNA sequencing data on the cloud, using 1500 CPU cores, and 6TB of RAM, delivering results with 43% increased efficiency compared to other tools. The flexible design of this framework allows easy adoption within other fields of Life Sciences and ensures that it will scale together with the demand for scientific analysis in the cloud for years to come.

Because many bioinformatics tools have been developed in the context of small sample sizes they often struggle to keep up with the demands for large-scale data processing required for modern research and clinical sequencing projects due to the limitations in their design. The Rheos software system is designed specifically with these large data sets in mind. Utilizing the elastic compute capacity of modern academic and commercial clouds, Rheos takes a service-oriented containerized approach to the implementation of modern bioinformatics algorithms, which allows the software to achieve the scalability and ease-of-use required to succeed under increased operational load of massive data sets generated by projects like International Cancer Genomics Consortium (ICGC) Argo and the All of Us initiative.

Rheos algorithms are based on an innovative stream-based approach for processing genomic data, which enables Rheos to make faster decisions about the presence of genomic mutations that drive diseases such as cancer, thereby improving the tools' efficacy and relevance to clinical sequencing applications. Our testing of the novel germline Single Nucleotide Polymorphism (SNP) and deletion variant calling algorithms developed within Rheos indicates that Rheos achieves 98% accuracy in SNP calling and 85% accuracy in deletion calling, which is comparable with other leading tools such as the Genome Analysis Toolkit (GATK), freebayes, and Delly.

The two frameworks that we developed provide important contributions to solve the ever-growing need for large scale genomic data analysis on the cloud, by enabling more effective use of existing tools, in the case of Butler, and providing a new, more dynamic and real-time approach to genomic analysis, in the case of Rheos.

Zusammenfassung

Forscher verwenden zunehmend Cloud-Computing, um umfangreiche Genomik-Analysen effizient und kostengünstig durchzuführen. Erfolgreiche Anwendungen in der Cloud erfordern sorgfältige Instrumentierung und ein Management, das häufige Probleme wie Ressourcenengpässe und geringe Auslastung vermeidet, die sowohl die Kosten erhöhen als auch den Zeitplan eines wissenschaftlichen Projekts verlängern können.

Um diesen Herausforderungen zu begegnen, haben wir das Framework Butler für ein umfangreiches Management von wissenschaftlichen Workflows in der Cloud entwickelt. Die Eckpfeiler des Butler-Designs sind folgende: Unterstützung mehrerer Clouds, Infrastruktur-Konfigurationsmanagement, skalierbarer, fehlertoleranter Betrieb, umfassende Ressourcenüberwachung sowie automatisierte Fehlererkennung und -wiederherstellung. Butler setzt auf robuste Open-Source-Komponenten, um ein Framework bereitzustellen, das über Tausende von Rechnerkernen und Millionen von Workflow-Ausführungen skalierbar ist. Die Fehlererkennungs- und Selbstheilungsfunktionen von Butler sind einzigartig unter den Frameworks für wissenschaftliche Arbeitsabläufe und gewährleisten, dass die Analysen mit minimalem Eingriff des Menschen durchgeführt werden.

Butler wurde für die Analyse von über 725 TB DNA-Sequenzierdaten in der Cloud verwendet, unter Nutzung von 1500 CPU-Kernen und 6 TB RAM. Damit wurden im Vergleich zu anderen Tools Ergebnisse mit einer um 43% gesteigerten Effizienz erzielt. Das flexible Design des Butler Frameworks ermöglicht eine einfache Übernahme in andere Bereiche der Biowissenschaften und stellt sicher, dass es hinsichtlich nachgefragter wissenschaftlicher Analysen in der Cloud während der nächsten Jahre skaliert werden kann.

Da viele Bioinformatik-Werkzeuge mit kleinen Stichprobengrößen entwickelt wurden, ist es oft schwierig, mit Anforderungen an die Datenverarbeitung im dem Maßstab Schritt zu halten, der für moderne Forschungs- und klinische Sequenzierungsprojekte erforderlich ist. Das Rheos Softwaresystem wurde speziell für derart große Datenmengen entwickelt. Rheos nutzt die elastischen Rechenkapazitäten moderner akademischer und kommerzieller Clouds und setzt einen serviceorientierten, containerisierten Ansatz für die Implementierung moderner Bioinformatik-Algorithmen ein. Dies ermöglicht es der Software, Skalierbarkeit und Benutzerfreundlichkeit zu erreichen, um so bei hoher Betriebslast erfolgreich große Datensätze, die von Projekten wie dem Internationalem Krebsgenomkonsortium (ICGC) Argo und der Initiative All of Us generiert wurden, zu prozessieren.

Rheos basiert auf einem innovativen, Stream-basierten Ansatz für die Verarbeitung genomischer Daten. Mit Hilfe von Rheos können schnellere Entscheidungen über das Vorhandensein genomischer Mutationen getroffen werden, die Krankheiten wie Krebs auslösen, und so die Wirksamkeit für klinische Sequenzierungsanwendungen verbessert werden. Unsere Tests der innerhalb von Rheos entwickelten Tools zum Auffinden neuartiger Keimbahn-Einzelnukleotid-Polymorphismen (SNP) und

Deletionen deuten an, dass Rheos eine Genauigkeit von 98% beim SNP-Aufruf und 85% Genauigkeit beim Aufruf von Deletionen erreicht. Dies ist vergleichbar mit anderen führenden Tools wie dem Genome Analysis Toolkit (GATK), Freebayes und Delly.

Die beiden von uns entwickelten Frameworks liefern wichtige Beiträge zur Bewältigung des ständig wachsenden Bedarfs der Analyse großvolumiger genomischer Datensätzen in der Cloud, indem im Fall von Butler die vorhandenen Tools effektiver eingesetzt werden und im Fall von Rheos ein neuer, dynamischer und realer Ansatz zur Echtzeit-Genomanalyse geschaffen wird.

Acknowledgements

This work is dedicated to the memory of my mother Elena Gitelson.

I'd like to thank and acknowledge, first and foremost, my immediate family, wife Polina Litvak, and children Anastacia and Alexandra Yakneen for sticking with me through the thick and thin, and putting up with my perennial preoccupation with my research and this document.

I'd like to thank and acknowledge my mentors Jan Korbel and Michael Gertz, for helping me start, continue, and finish this project; for giving me the space to develop my ideas, yet providing the guidance necessary to stay on the right course.

I'd like to thank and acknowledge the Faculty of Mathematics and Computer Science at Ruprecht-Karls-Universität Heidelberg for giving me the privilege and opportunity to carry out my doctoral research here.

I have been inspired all my life by my parents' Elena Gitelson and Vladimir Iakhnin's dedication to research in medicine and physics, and especially by the example of my grandfather Joseph Gitelson, who today, at 90 years of age, still actively contributes to scientific research. May this work be a contribution worthy of their legacy.