Towards Real-time Proteomics Data Analysis using PEAKS Online 8.5

Lei Xin, Lin He, Zia Rahman, Tom Andersen, Brenton Morse, Ruilun Yang, Shaofeng Jiang, Sahar Rabinoviz

Bioinformatics Solutions Inc., Waterloo, Ontario, Canada

Bioinformatics Solutions Inc.

Introduction

As mass spectrometry technology advances, the prevalence of these tools in university research and clinical diagnosis labs are growing. The speed and accuracy advancements of this technology are resulting in large amount of data being generated. This poses a technological obstacle as this high-volume data cannot be interpreted for meaningful usage without complex computational processes. To address this fundamental obstacle and take advantage of now publicly or privately available cloud computing resources, we have developed a new distributed computing platform as the foundation for the next generation of our software. Our proposed technology is designed to perform complex algorithm in real-time on high-volume of mass spectrometry data with the support of hundreds to thousands of computing nodes running simultaneously, all while providing an easy-to-use and easy-to-access interface without the need to install local client.

Platform Architecture

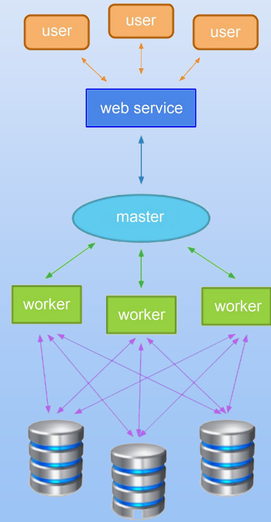

Figure 1: Distributed System Architecture with single master node and multiple workers.

Figure 1 shows the structure of our distributed system, where different users can access the web service simultaneously. The master node will be responsible for scheduling the tasks and sending them to the workers without blocking the master itself. The workers will work on individual tasks concurrently and communicate with master via messages.

To take advantage of all the possible computing resources and do the data analysis in real-time, we designed our architecture to be cross-platform, with support for several different hardware topology architectures. The architecture is also intrinsically distributed which ensures the computing nodes can start up and shut down at any time. To help facilitate the high bandwidth I/O requirements of our distributed systems for data access, we also designed a distributed database implemented with Cassandra system. Effective and efficient data structuring is also implemented to support high volume mass spectrometry data and results access. Further, an asynchronized programming model built on Akka actor system were applied in the platform to ensure deadlock is avoided as thousands of jobs are scheduled.

Experiment

A testing distributed system was setup using Google Cloud platform and Cassandra cluster. The Google Cloud was configured with multiple CPUs and threads allowing multiple user task submission. We setup a Cassandra cluster with three nodes for data access and with different testing configurations for number of workers, threads per worker, worker memory and total threads.

The testing data set is a public one from “Clinical Proteomic Tumor Analysis Consortium” study. The samples were analyzed by four Thermo LTQ and four Orbitrap instruments. Only four Orbitrap data (OrbiO@65, OrbiP@65, OrbiW@56, Orbi@86) were used in our study. The samples were prepared like this: a yeast lysate was spiked with a mixture of 48 human proteins (Sigma-Aldrich UPS1) at five levels: 0.25, 0.74, 2.2, 6.7, and 20 fmol/μL, each 3-fold higher than the last. Each sample was then analyzed three times by each instrument with each run lasting for 3 hours. From low level to high level, all 15 runs for each instrument were grouped into 5 groups: A, B, C, D, E, each contained triplicate runs for one sample giving a total of 60 samples. A customized FASTA database was constructed by combining 48 UPS proteins and yeast proteins from UNIPROT database together with the contaminant database.

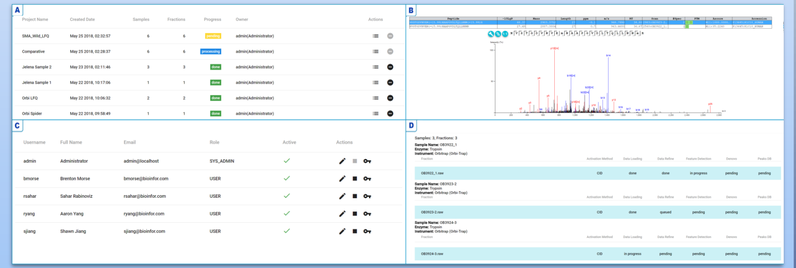

Figure 2: PEAKS Online Web Interface
A) User management page with multiple users
B) Peptide table and spectrum view on results page
C) Project listing page with multiple projects processing in parallel
D) Project processing page with multiple samples processing in parallel

Results

The data set was run using the PEAKS DB search algorithm on PEAKS Online platform and PEAKS DB search algorithm on PEAKS Studio for comparison. The PEAKS Studio run time results are shown in Table 1 and the PEAKS Online run time results are given in Table 2. The analysis time using 2, 4, 6, 12, 18, and 24 workers with 6 threads each are 9:33h, 4:42h, 3:07h, 1:40h, 1:12h, and 0:56 respectively. We noticed for the same dataset our distributed architecture runs more quickly than PEAKS Studio in all cases. Specifically the distributed system continues to show linear run time improvements with increased workers whereas PEAKS Studio stops achieving linear run time improvement by increasing thread count.

PEAKS Online is accessible through straight forward web interface detailed in Figure 2. By using a web interface client, users are not required to install additional software, instead they are able to connect to the hosted server as shown in Figure 1. Using their web browsers, multiple users can schedule and run multiple workflow submissions. The submitted workflows run in parallel and the processing begins once a worker becomes available.

Thread Count	Run Time	Run Time Change	Thread Count Change	Scalability
8	10:28:00	-	-	-
16	8:52:00	15.29%	100%	**0.15**
32	7:38:00	13.91%	100%	**0.14**
64	7:14:00	5.24%	100%	**0.05**

Table 1: PEAKS Studio 8.5 scalability performance.

Workers	Total Threads	Run Time	Run Time Change	Thread Count Increase	Scalability
2	12	9:33:00	-	-	-
4	24	4:42:00	50.79%	100%	**0.5079**
6	36	3:07:00	33.69%	50%	**0.6738**
12	72	1:40:00	46.52%	100%	**0.4652**
18	108	1:12:00	28.00%	50%	**0.5600**
24	144	0:56:00	22.22%	33%	**0.6667**

Table 2: PEAKS Online 8.5 scalability performance.

Conclusion

The benchmarking shows that PEAKS Online is able to significantly cut down on computing time with comparison to current PEAKS Studio architecture, offering approximately linear scaling of run time with respect to an increase in number of threads/workers.

Contact

Bioinformatics Solutions Inc. www.bioinfor.com support@bioinfor.com