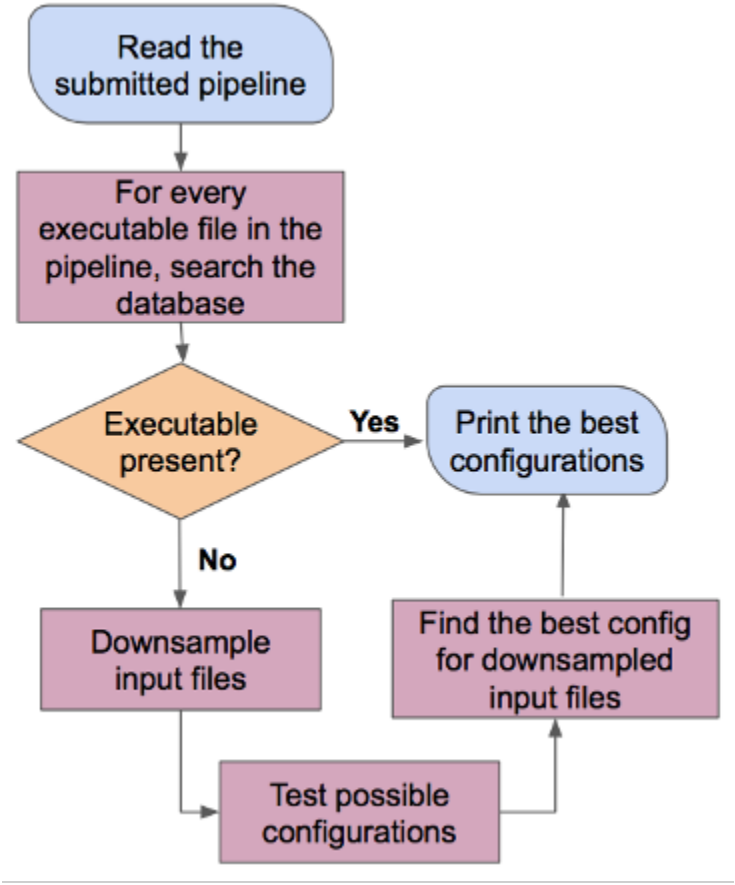
Docket #: S19-470

Hummingbird: Predicting Best Configurations for Genomics Cloud Computing

Stanford researchers developed a framework called 'Hummingbird' that predicts the cheapest, fastest and most efficient configurations to execute genomics pipelines on the cloud. Genomics researchers frequently do not know which cloud configuration is best to execute a pipeline, and thereby select a more expensive cloud tier than needed. Hummingbird provides three recommended pipeline configurations prior to running to assist the user in saving time and money:

> **Cheapest**: the least expensive configuration over the three categories and then selects the instance type (number of Virtual Central Processing Uintss) with the lowest cost.
> **Fastest**: the fastest configuration over the three categories and then selects the instance type with least execution time.
> **Fast and Cheap**: the highest normalized speedup, i.e., the instance which scaled best, and then selects the instance type with the lowest cost.

Hummingbird formulates the configuration prediction via Downsampling, the Memory Profiler, and the Prediction Model. **Downsampling** reduces the time required for the training phase prior to prediction. By reducing the size of the input files, Hummingbird can execute a genomics pipeline more quickly, thus decreasing the cost of training within the prediction framework. The **Memory Profiler** eliminates failures of the entire pipeline or any of its individual stages due to selection of inadequate main memory on the cloud. Once Downsampling is complete, Hummingbird uses the files to run the entire pipeline on different types of Google cloud instances. Once the execution is complete, the Hummingbird **Prediction Model** compares the execution time of all the instances and identifies the three different configurations: "fastest", "cheapest", and "most efficient" (fast and cheap).



Hummingbird Cloud Genomics Pipeline Configuration Determination
(image courtesy The Snyder Lab)

Stage of Development -Prototype
During testing, Stanford researchers compared Hummingbird's results with those obtained by executing the application on the whole input file and found that Hummingbird predicts the best configuration in many cases. Future work will improve heuristics and therefore accuracy, and tune the framework to provide more optimized solutions.

Applications

- Cloud genomics

Advantages

- Faster and lower cost – potentially an order of magnitude cost savings

Publications

- Amir Bahmani, Ziye Xing, Vandhana Krishnan, Utsab Ray, Frank Mueller, Amir Alavi, Philip S. Tsao, Michael P. Snyder, Cuiping Pan, Hummingbird: efficient performance prediction for executing genomic applications in the cloud, *Bioinformatics,* Volume 37, Issue 17, 1 September 2021, Pages 2537–2543,
- Ray, Utsab, Vandhana Krishnan, Amir Bahmani, Cuiping Pan, Keith Bettinger, Philip Tsao, Frank Mueller, and Michael Snyder. "Hummingbird: Efficient performance prediction for executing genomics applications in the cloud." In *Computational Approaches for Cancer Workshop*. 2018.

Innovators

- Amir Bahmani
- Vandhana Krishnan
- Cuiping Pan
- Ziye Xing
- Utsab Ray
- Michael Snyder
- Philip Tsao

Licensing Contact

Imelda Oropeza

Senior Licensing Manager, Physical Sciences

Email