Geophysical Research Abstracts
Vol. 20, EGU2018-11376, 2018
EGU General Assembly 2018

Exploiting Transparent Multimodal Parallelization for High-Performance Big Data Analytics and Machine Learning

Kwo-Sen Kuo (1,2), Hongfeng Yu (3), Michael Rilee (1,4), Yu Pan (3), and Feiyu Zhu (3)

(1) NASA Goddard Space Flight Center, Greenbelt, Maryland, United States (kwo-sen.kuo@nasa.gov), (2) Bayesics LLC, Bowie, Maryland, United States (kuo@bayesics.com), (3) University of Nebraska, Lincoln, Nebraska, United States (hfyu@unl.edu), (4) Rilee Systems Technologies LLC, Derwood, Maryland, United States (mike@rilee.net)

The challenge of Big Data may be succinctly summarized as: Achieving optimal scalability on data volume and variety to obtain analysis results with desirable speed (velocity). "Interactivity" is perhaps the analysis speed most desired by science researchers. Such high performance is obviously unattainable without employing parallel processing. It is even unlikely to be attained with only a single mode of parallelization, e.g. multithreaded parallelism (i.e. the pleasingly parallel kind). That is, multimodal parallelization, with both multithreaded and distributed modes and preferably on CPU-GPU hybrid compute architecture, must thus be judiciously exploited to optimize throughput. Few scientists, however, have the expertise in such sophisticated parallel programming. Thus, a platform that allows users to transparently utilize multimodal parallelization is an ideal solution, similar to how average smartphone users are utilizing multicore capabilities without being consciously aware of it.

A platform supports the analysis interactivity alluded to above will no doubt attract intense utilization for interactive analysis during the work hours. Such interactive usage, however, is likely to plummet in the off-work hours, during which time batch-suitable machine learning operations can be executed without leaving the system idle. With our innovation homogenizing data variety, data preparation (traditionally taking disproportionate effort and time) will be drastically simplified and accelerated, leading to better systemized machine learning.

In this presentation, we describe and demonstrate a prototype of such a platform and the technical innovations that have made it possible. We will also introduce additional unique advantages such a platform affords.