Geophysical Research Abstracts
Vol. 20, EGU2018-16171, 2018
EGU General Assembly 2018

Scaling DBSCAN towards exascale computing for clustering of big data sets

Ernir Erlingsson (1), Helmut Neukirchen (1), Gabriele Cavallaro (2), and Morris Riedel (2)

(1) University of Iceland, School of Engineering and Natural Sciences, Mechanical, Industrial Engineering and Computer Science, Iceland (ernire@gmail.com), (2) Jülich Supercomputing Centre, Forschungszentrum Jülich, Germany

Progress in sensor technology allows us to collect environmental data in more detail and with better resolution than ever before. One example are 3D laser scanners that generate 3D point-cloud datasets for land survey. Clustering can then be performed on these datasets to identify objects such as buildings, trees, or rocks in the unstructured point-clouds. Segmenting huge point-clouds (of whole cities or even whole countries) into objects is a computationally expensive operation and therefore requires parallel processing. Density-based spatial clustering of applications with noise (DBSCAN) is a popular clustering algorithm and HPDBSCAN is an efficient parallel implementation of it running on supercomputing clusters. Tomorrow's supercomputers will be able to provide exascale computing performance by exploiting specialised hardware accelerators, however, existing software needs to be adapted to make use of the best fitting accelerators. To address this problem, we present a mapping of HPDBSCAN to a pre-exascale platform currently being developed by the European DEEP-EST project. It is based on the Modular Supercomputer Architecture (MSA) that provides a set of accelerator modules which we exploit in novel ways to enhance HPDBSCAN to reach exascale performance. These MSA modules include: a Cluster Module (CM) with powerful multicore CPUs; the Extreme Scale Booster (ESB) module with manycore CPUs; the Network Attached Memory (NAM) module which stores datasets and provides extremely fast access to them; a fast interconnect fabric speeds up inter-process message passing together with the Global Collective Engine (GCE), which includes a multi-purpose Field Programmable Gate Array (FPGA) for, e.g., summing up values transmitted in messages collected. HPDBSCAN exploits the above accelerator modules as follows: the data that is to be clustered can be stored in the NAM, it gets subsequently distributed and load balanced, which is accelerated by the GCE, to the CPU nodes of the CM; the parallel clustering itself is performed by the powerful CPUs of the CM which also merges the obtained cluster IDs; the merged cluster IDs are stored in the NAM for further level of detail (LoD) studies, i.e. zooming in and out based on continuous, instead of fixed, levels of importance for each point, which can be regarded as an added dimension. The ESB module (supported by GCE) is most suitable to calculate these continuous level of importance (cLoI) values and add them to the dataset in the NAM. Based on the added cLoI data, the LoD studies can then be performed by re-clustering as described previously, i.e. distribution and load balancing of the cLoI value-enriched dataset followed by parallel clustering. The described approach will allow to scale HPDBSCAN-based clusterings on tomorrow's hardware towards exascale performance.

The DEEP-EST project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No. 754304 DEEP-EST.