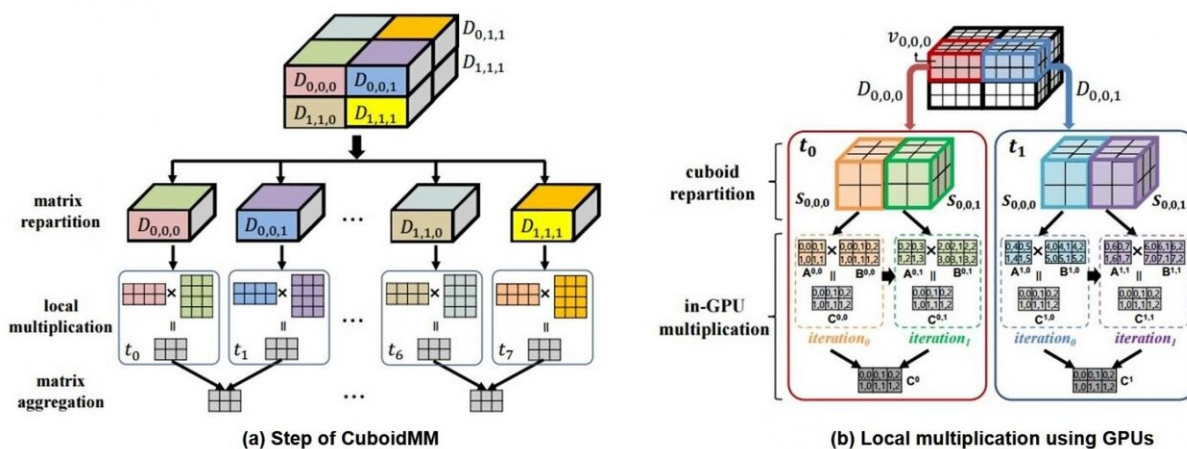# DistME: A fast and elastic distributed matrix computation engine using GPUs

July 17 2019

This is a mimetic diagram (a) of 3D data multiplication through CuboidMM and mimetic diagram (b) of data processing computation using GPU. Credit: DGIST

DGIST announced on July 4 that Professor Min-Soo Kim's team in the Department of Information and Communication Engineering developed the DistME (Distributed Matrix Engine) technology that can analyze 100 times more data 14 times faster than the existing technologies. This new technology is expected to be used in machine learning that needs big data processing or various industry fields to analyze large-scale data in the future.

'Matrix' data, which expresses numbers in row and column, is the most

widely used form of data in various fields such as machine learning and science technology. While 'SystemML' and 'ScaLAPACK' are evaluated as the most popular technologies to analyze matrix data, but the processing capability of existing technology has recently reached its limits with the growing size of data. It is especially difficult to conduct multiplications, which are required for data processing, for big data analysis with the existing methods because they cannot perform elastic analysis and processing and require a huge amount of network data transfer for processing.

In response, Professor Kim's team developed a distributed matrix multiplication method that is different from the existing one. Also called CuboidMM, this method forms matrix multiplication in a 3-D hexahedron and then partitions and processes to multiple pieces called cuboids. The optimal size of the cuboid is flexibly determined depending on the characteristics of the matrices, i.e., the size, the dimension, and sparsity of matrix, so as to minimize the communication cost. CuboidMM not only includes all the existing methods but also can perform matrix multiplication with minimum communication cost. In addition, Professor Kim's team devised an information processing technology by combining with GPU (Graphics Processing Unit) which dramatically enhanced the performance of matrix multiplication.

The DistME technology developed by Professor Kim's team has increased processing speed by combining CuboidMM with GPU, which is 6.5 and 14 times faster than ScaLAPACK and SystemML respectively and can analyze 100 times larger matrix data than SystemML. It is expected to open new applicability of machine learning in various areas that need large-scale data processing including online shopping malls and SNS.

Professor Kim in the Department of Information and Communication Engineering said 'Machine Learning Technology, which has been

drawing worldwide attention, has limitations in the speed for matrix-form [big data](#) analysis and the size of analysis processing. The [information](#) processing technology developed this time can overcome such limitations and will be useful in not only [machine learning](#) but also applications in wider ranges of science technology data analysis application."

This research was participated by Donghyoung Han, a Ph.D. student in the Department of Information and Communication Engineering as the first author and was presented on July 3 in ACM SIGMOD 2019, the top-renowned academic conference in the database field held in Amsterdam, Netherlands.

Provided by DGIST (Daegu Gyeongbuk Institute of Science and Technology)

Citation: DistME: A fast and elastic distributed matrix computation engine using GPUs (2019, July 17) retrieved 24 April 2024 from https://techxplore.com/news/2019-07-distme-fast-elastic-matrix-gpus.html