Sujet stage de master 2/fin d'études 2019-2020

Titre :

Sparse Matrix Linear Algebra and TensorFlow for Machine Learning on new Processors

Directeurs : Serge Petiton and Jérôme Gurhem, Maison de la simulation/CNRS, Saclay

Contexte :

Exascale supercomputers, and beyond, are expected to have highly hierarchical architectures, generating multi-level programming. The different programming associated to those new architectures and new hardware will generate new difficult challenges. Several programming paradigms and languages based on workflows or graphs of tasks are candidates. Machine learning applications on such supercomputers would generate important breakthroughs with crucial societal impacts. Nevertheless, a lot of researches on several topics still have to be addressed before any such results. An important topic is how we have to adapt linear algebra methods for future supercomputers manipulating huge data for machine learning applications, depending of programming paradigms in particular. As explain recently by Y. LeCun, future neural networks would have just a small amount of neurones actives for a given step and the main problem would be distributed and parallel spares matrix linera algebra problems, and in particular very large and sparse matrix vector multiplication.

TensorFlow is an "open source software library for high performance numerical computation", introduced by Google in 2015, well-adapted for machine learning. A TensorFlow program is based on a Graph where each node is an Operation that process Tensors (containing data). TensorFlow allows to manipulate tensors and dense and sparse matrices and propose several linear algebra methods.

Sujet

The main goal of this internship is to experiment existing linear algebra proposed by TensorFlow for sparse matrix computation and to develop new methods using different sparse matrix patterns. In particular, the student will analyse how it is possible to use new sparse matrix compression formats using TensorFlow in order to minimize communications and optimize computing time, on processors without any shared memories connecting all the cores. Some experiments on new processors or simulators will be proposed and analysed. We'll probably experiment with the new parallel processor proposed by the Silicon Valley star-up Cerebras and/or the new Fujitsu multi-core chip or simulator with a network on chip, developed for the future Japanese exascale machine..

Contributions attendues :

- Evaluation of TensorFlow for sparse matrix computation, with the existing COO format
- Implementation and evaluation of several other sparse formats (ELLPACk, SGP, CRS, …) into TensorFlow, with experiments on multi-core processors.
- Implementation of some linear algebra methods using those sparse formats and TensoFlow
- Depending of the obtained results, implementation and experiments using several multi-cores processors in parallel.
- Co-author of a paper to be submitted to an international conference with peer reviews
- Internship report in English

Lieu du stage : Maison de la Simulation, USR CNRS, Saclay

Durée: 6 months, “avec gratifications”.

Poursuite en thèse possible

Contact : serge.petiton@univ-lille.fr