FPGA-based Accelerators for "Learning to Rank" in Web Search Engines

Ning-Yi XU, Jing YAN, Rui GAO, Xiongfei CAI, Zenglin XIA, Feng-Hsiung Hsu
Platforms and Devices Center, Microsoft Research Asia
{Ningyixu, v-jiy, ruigao, xfcai, fhh}@microsoft.com

Introduction

Search relevance is a key measurement to judge the usefulness of search engines, which may change a search company's market cap by tens of billions of dollars. With the ever-increasing scale of the World Wide Web, machine learning technologies have become important tools to improve search relevance rankings. However, these machine learning algorithms are slow on large datasets using general purpose computing methods. Thus, at Microsoft Research Asia (MSRA), we have built two generations of FPGA-based accelerators to reduce the computation time for "learning to rank" algorithms (including but not limited to RankBoost [1], RankNet [2] and Lambda Rank [3][4]). This presentation will show our experience in designing efficient accelerators and cooperating with domain experts (MSRA researchers) to release the computing power of accelerators to support real world research activities. We will also discuss the most challenging part of the work, which is how to efficiently implement the parallelism in the algorithm, to meet both performance and quality goals with limited hardware resources, and providing good scalability at the same time. Experimental results show that the accelerators could accelerate RankBoost up to 1800x, LambdaRank about 10x ~15x, with equal or similar ranking quality compared to a pure software version.

The Accelerator Boards

Based on our understanding of the target applications, we have customized two generations of FPGA-based accelerators boards (Figure 1.), which are PCI/PCIe boards with FPGA and large on-board memory. The first board is a (32bit/33MHz) PCI card with FPGA and memories (SDRAMs, SRAMs). On this board, an Altera Stratix-II FPGA is used as the core computation component, and the on-board memories include DDR SDRAM (up to 2 GB) and SRAM (32MB). The second board continues to use the Altera Sratix-II family FPGAs as the main computation engine, and it also has two DDR2 modules (up to 16GB). A Xilinx Virtex-5 LXT FPGA is used to provide PCI Express (x8) interface to the host computer. Several Chinese universities (including Tsinghua, SJTU and BIT) are also developing various applications on these boards.

FPGA-based Accelerator for RankBoost (FAR)

We started our work with RankBoost in 2006. RankBoost is a promising algorithm in the learning to rank area, but it is not widely used because of its heavy computational load. At that time, a typical run of RankBoost on a Live Search dataset takes about two days, and to obtain optimal results, many runs are required. Therefore, we chose RankBoost as our first target algorithm of acceleration [5].

First, we redesigned the algorithm for better parallelism and similar quality to the original one, and reduced its computational complexity from O(N^2) to O(N). We then mapped the optimized algorithm to our first generation of FPGA-based PCI accelerator. In the FPGA, we designed a single instruction multiple data (SIMD) architecture with multiple processing engines (PEs). Each PE can operate at 180MHz in a full pipeline manner. Training data is compressed and well organized in order to enable high-bandwidth streaming memory access, which dramatically improves processing times.

The original software takes about 45 hours in a single complete run. The hardware accelerated implementation of the new algorithm takes just over 15 minutes, representing a speed-up of around 170 times. This enables intensive parameter exploration and relative studies with RankBoost. Several FAR accelerator systems have been successfully used in research work by the Web Search and Mining Group at Microsoft Research Asia. In certain applications for commercial search engines, our results suggest that the accelerator could be far more efficient than software that may take weeks instead of two days.

As the training data becoming larger and larger, the on-board memory (up to 2GB) on the first generation accelerator cannot load all required data. Thus we built the second generation accelerator which has large on-board memory (up to 16GB) in 2007. The performance is shown in Fig 2.

To have a better scalability, in 2008, we developed a distributed version of the system which can utilize multiple accelerators with a near linear performance increasing with the number of accelerator nodes. RankBoost accelerator system now provides a comprehensive high performance solution for research and engineering activities in learning to rank research.

Acceleration of RankNet and Lambda Rank

Lambda Rank ([2][3][4]) is another important ranking algorithm for which we are building an FPGA-based accelerator. Lambda Rank is proposed to learn with nonsmooth cost functions and we select its implementation for RankNet in [2] and [4], which uses a back-propagate neural network to learn the ranking function from human labeled datasets. The challenges in the acceleration exist in dense floating point operations, complex control flow, and implementation of special math functions (such as hyperbolic tangent, sigmoid etc.). This architecture requires more FPGA resources compared to FAR (FPGA-based Accelerated RankBoost), because the algorithm needs many floating point adders and multipliers. In

addition, larger on-board/on-chip memories are also required by a large number of intermediate data structures. This is why we plan to implement it on our second generation accelerator board, which is equipped with more powerful computing components. The hardware algorithm is also organized to SIMD architecture, and can operate at 100MHz with a deeper pipeline (the number of the pipeline stage is 9). Recently, we have just made the system run, and the speed-up is about 10 ~ 15 times. The preliminary ranking quality could be shown in Fig 3 in the measure of NDCG. Much better results have been observed with experiments on our virtual hardware model written in software. We are now in the process of tuning the reconfigurable hardware to further improve the ranking quality.

Acknowledgements

The authors would like to thank Lei ZHANG, Wei LAI, Jun-Yan CHEN, Tie-Yan LIU and Hang LI for their support and valuable feedback in this work.

References

[1] Yoav Freund, et al. An Efficient Boosting Algorithm for Combining Preferences. JMLR, 2003.

[2] Christopher J.C. Burges, et al. Learning to Rank using Gradient Descent. ICML, Pages: 89-96, 2005.

[3] Christopher J.C. Burges, et al. Ranking as Learning Structured Outputs. Proceedings of the NIPS 2005 Workshop on Learning to Rank, Dec 2005.

[4] Christopher J.C. Burges, Robert Ragno and Quoc Viet Le. Learning to Rank with Nonsmooth Cost Functions. Proceedings of the NIPS 2006.

[5] Ning-Yi XU, Xiong-Fei CAI, Rui GAO, Lei ZHANG and Feng-Hsiung HSU. “FPGA-based accelerator design for RankBoost in web search engines”. Proceedings of International Conference on Field-Programmable Technology (ICFPT), 2007.

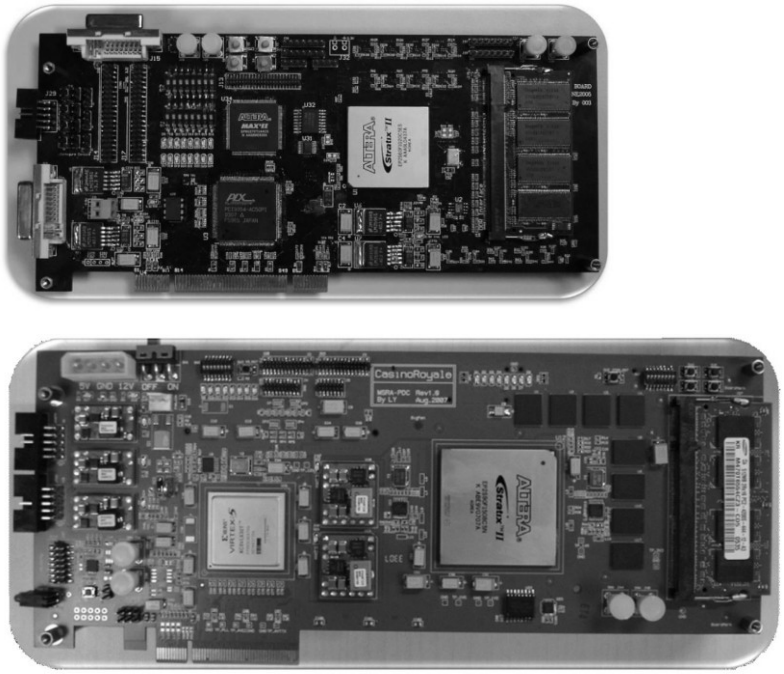

Figure 1. Two generations of accelerator boards.

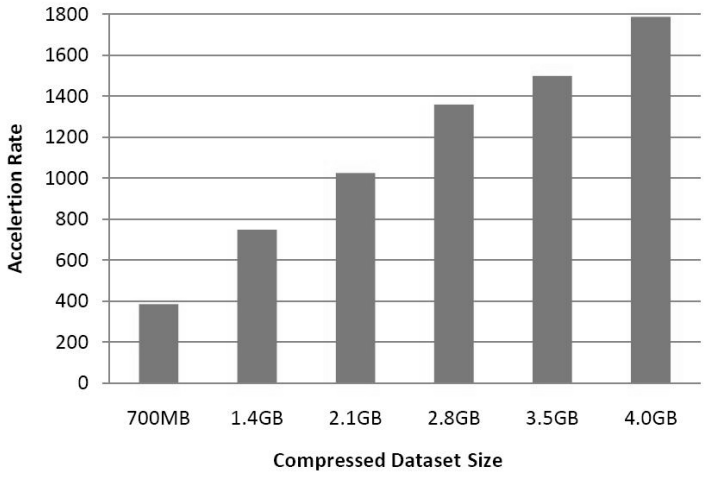

Figure 2. Speedup of FAR over the software implementation. Datasets are selected from real applications.

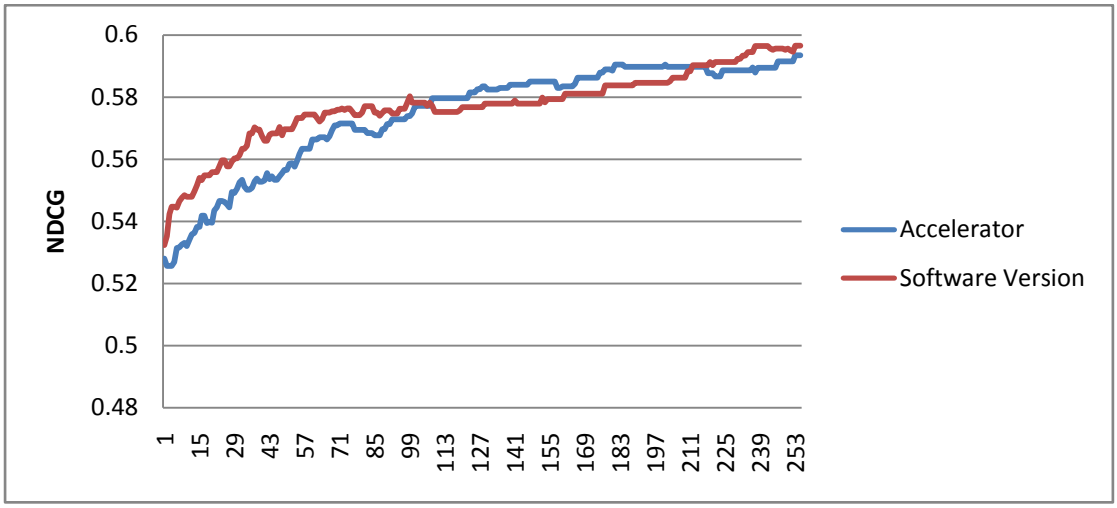

Figure 3. NDCG of ranking models over different training rounds. These models are trained by the accelerator and the original software version. The data set is from a commercial search engine.