Yi SHAN

http://nicsefc.ee.tsinghua.edu.cn/people/yi-shan/
Address：Baidu Building, Beijing, 10085, China
Phone：(+86)138-1098-6576
Email：shany04@gmail.com

Summary

Senior R&D Engineer in Baidu Research on large-scale heterogeneous computing for deep learning applications, such as image recognition. PhD in heterogeneous computing at Tsinghua University, and also at Imperial College London for a half-year visiting scholarship, with extensive experience in FPGA and GPU acceleration of computation-intensive algorithms in search engine, computer vision, and brain network analysis. Academic awards include IBM PhD Fellowships and Microsoft Research Innovation Internship Award. Industry experience includes search engine acceleration using FPGA and GPU (at Microsoft Research Asia), FPGA virtualization in cloud computing (at IBM Research China), and digital design for baseband communication (at Ericsson China).

Specialties: large-scale heterogeneous computing with FPGA and GPU for machine learning and computer vision

Education Background

Electronic Engineering, Tsinghua University **2008.09-2014.07**

- Doctor of Philosophy (Ph.D.), Electronic Science and Technology

Electronic Engineering, Tsinghua University **2004.09-2008.07**

- Bachelor of Engineering (B.Eng.)

Honors

- IBM PhD Fellowship Award, 2012 (among 84 around the world)
- Microsoft Research Innovation Internship Award, Microsoft Research Asia, 2011
- Second Prize in AMD GPU Acceleration Contest, China, 2010
- Second Prize of GuangHua Scholarship, Tsinghua, 2010
- Mitsubishi Heavy Industries Scholarship , 2010
- Outstanding graduate Communist, Tsinghua, 2010
- Academic excellence Award, Tsinghua, 2007

Working Experience

Institute of Deep Learning, Baidu Research 2014.07-now

- Large-scale heterogeneous computing for deep learning applications: I have built a large-scale GPU cluster, and designed some parallel strategies for efficient training of convolutional neuron network (CNN) model. I have also proposed data augmentation and multi-scale training to get the state-of-art result, 5.98% top-5 error rate, for ImageNet classification benchmark. Meanwhile, collaborated with Altera, I have designed FPGA based CNN system. This system is more suitable for low-latency and low-power scenario.
- paper：Deep Image: Scaling up Image Recognition, http://arxiv.org/abs/1501.02876
- FPGA demo：http://newsroom.altera.com/press-releases/altera-baidu-fpga-cloud-data-centers.htm

Visiting & Internship Experience

Next-Generation Systems, IBM Research China 2013.7-2013.10

- Low-latency heterogeneous cloud computing based on FPGA virtualization

Custom Computing Research Group, Imperial College London (supervisor Wayne Luk) 2012.10-2013.3

- Runtime adaptive computer vision system on Maxeler (multi-FPGA platform)
- Monte-Carlo parallel computing for pricing Asian options on Maxeler

Hardware Computing Group, Microsoft Research Asia 2009.11-2011.3

- FPGA and GPU acceleration of search engine applications, including online matching and ranking, offline Rankboost and PageRank algorithms.

2009.06-2009.11

Ericsson China R&D

- Baseband Digital Design using Verilog language on FPGA

China Mobile 2007.06-2007.8

- Communication Network Optimization using Matlab

Other Research Experience

Real-Time Detection, Tracking and Stereo Matching on FPGA

I have designed an FPGA system for Mitsubishi electronic road pricing including real-time vehicle detection, tracking and stereo measurement. For the stereo matching part, some memory optimizations are used to get a 47.6 fps for video size of 1920 × 1080 with a large disparity range of 256.

Publications (citations according to Google scholar: 124)

[1] **Yi SHAN**, *et al.*, "Hardware Acceleration for Accurate Stereo Vision System using Mini-Census Adaptive Support Region", **ACM Transactions on Embedded Computing Systems, 2014**.

[2] **Yi SHAN**, *et al.*, "FPGA based memory efficient high resolution stereo vision system", in International Conference on Field-Programmable Technology, **FPT 2012**.

[3] **Yi SHAN**, *et al.*, "FPMR: MapReduce Framework on FPGA", in ACM/SIGDA International Symposium on Field Programmable Gate Arrays, **FPGA 2010**.

[4] **Yi SHAN**, *et al.*, "FPGA and GPU Implementation of Large Scale SpMV", in IEEE Symposium on Application Specific Processors, **SASP 2010**.

[5] Fei Chen,**Yi Shan**, *et al.*, Enabling FPGAs in the Cloud, Proceedings of the 11th ACM Conference on Computing Frontiers, **CF 2012**.

[6] Guohao Dai, **Yi Shan**, *et al.*, "Online Scheduling for FPGA Computation in the Cloud", in International Conference on Field-Programmable Technology, **FPT 2014**.

[7] Boxun Li, **Yi SHAN**, *et al.*, "Memristor-based Approximated Computation", in International Symposium on Low Power Electronics and Design, **ISLPED 2013**.

[8] Wei WU, **Yi SHAN**, *et al.*, "FPGA Accelerated Parallel Sparse Matrix Factorization for Circuit Simulations", in Lecture Notes in Computer Science, **LNCS 2011**.

[9] Tianji WU, Bo WANG, **Yi SHAN**, *et al.*, "Efficient PageRank and SpMV Computation on AMD GPUs", in International Conference on Parallel Processing, **ICPP 2010**.

[10] Di WU, Tianji WU, **Yi SHAN**, *et al.*, "GPU Acceleration of Brain Network Analysis", in International Conference on Parallel and Distributed Systems, **ICPADS 2010**.

[11] Yu WANG, Yong HE, **Yi SHAN**, *et al.*, "Hardware Computing for Brain Network Analysis", in Asia Quality Electronic Design, **ASQED 2010**.