Heterogeneous Chip Architectures for Big Data Analytics

Friday, October 20th, 2017 at 10:00am
Johnson Center, Room F

Dr. Houman Homayoun

Department of Electrical and Computer Engineering, George Mason University

There is a fundamental shift underway now in various industry sectors to transport their big data to the clouds, thus increasing their efficiency and improving cash flow. In addition to traditional web service sector, this also includes sectors such as telecommunication (e.g., software defined network applications), and healthcare (e.g., GE Cloud Health and Cloud PACS). Cloud servers to process big data are now aggregating multiple disparate workloads with heterogeneously structured and unstructured data (image, text, video, graph) with varying performance goals (batch, streaming, real-time, and near-real-time). Emerging analytics applications in these domains rely heavily on specific deep machine learning and mining algorithms, and are running complex database software stack with significant interaction with I/O and OS and sharing many inherent characteristics that are fundamentally different from traditional CPU and parallel applications. Emerging big data applications require computing resources that can efficiently scale to manage massive amounts of diverse data. However, the rapid growth in the data yields challenges to process them efficiently using current cloud server architectures such as high performance Xeon. Furthermore, physical design constraints, such as power and density, have become the dominant limiting factor for scaling out servers. To respond to these challenges, heterogeneous architectures which integrates big and little cores with FPGA accelerators has emerged as a promising solution. In this talk, through methodical investigation of power and performance measurements, and comprehensive system level and micro-architectural analysis, I first show the characterization results of emerging big data applications on big Xeon and little Atom-based servers. The characterization results across a wide range of real-world big data applications and various software stacks demonstrate how the choice of big vs little core-based server for energy-efficiency is significantly influenced by the size of data, performance constraints, and presence of accelerator. Furthermore, the microarchitecture-level analysis highlights where improvement is needed in big and little core servers.

Second, I will show how in a heterogeneous architecture effective mapping of Hadoop and Spark MapReduce based applications to FPGA accelerator can significantly increase the energy-efficiency and performance. The real-system results show promising kernel speedup of more than 100X and significant energy-efficiency gains.

Biography: Houman Homayoun is an Assistant Professor in the Department of Electrical and Computer Engineering at George Mason University. He also holds a courtesy appointment with the Department of Computer Science. Houman joined GMU as a tenure-track Assistant Professor in August 2012. He is the director of GMU's Green Computing and Heterogeneous Architectures Laboratory (GOAL).

Prior to joining GMU, Houman spent two years at the University of California, San Diego, as NSF Computing Innovation (CI) Fellow awarded by the CRA-CCC working with Professor Dean Tullsen. Houman graduated in 2010 from University of California, Irvine with a Ph.D. in Computer Science. He was a recipient of the four-year University of California, Irvine Computer Science Department chair fellowship. His dissertation, entitled “Beyond Memory Cells for Leakage and Temperature Control in SRAM-based Units, the Peripheral Circuits Story”, was supervised by Professor Alex Veidenbaum from CS Department, and Professor Jean-Luc Gaudiot, and Professor Fadi Kurdahi from ECE Department. Out of thirty-one doctoral dissertations his work was nominated for 2010 ACM Doctoral Dissertation Award. Houman received the MS degree in computer engineering in 2005 from University of Victoria and BS degree in electrical engineering in 2003 from Sharif University of Technology.

Houman conduct research in big data computing, heterogeneous computing and hardware security and trust, which spans the areas of computer design and embedded systems, where he has published more than 80 technical papers in the prestigious conferences and journals on the subject. He is currently leading six research projects funded by DARPA, AFRL and NSF on the topics of hardware security and trust, big data computing, heterogeneous architectures, and biomedical computing. He successfully completed four projects on “Hybrid Spin Transfer Torque-CMOS Technology to Prevent Design Reverse Engineering”, “Persistence and Extraction of Digital Artifacts from Embedded Systems”, “Inter-core Selective Resource Pooling in a 3D Chip Multiprocessor”, and “Enhancing the Security on Embedded Automotive Systems” funded by DARPA, NIST, NSF and General Motors. Houman received the 2016 GLSVLSI conference best paper award for developing a manycore accelerator for wearable biomedical computing. He is currently serving as technical program chair of 2017 GLSVLSI conference, and will serve as the general chair of the conference for 2018. Since 2017 he has been serving as an Associate Editor of IEEE Transactions on VLSI.