Kartik Hegde

+1 217 721 2220 • kvhegde2@illinois.edu
www.kartikhegde.net

I am a computer architect passionate about building domain specific massively heterogeneous systems consisting of multiple hardware accelerators. I have a strong background in designing energy efficient accelerators for deep learning applications.

Education

- **University of Illinois at Urbana-Champaign** — **Urbana, IL**
  Ph.D. in Computer Science, GPA: 4.0 — *2017–2022*
- **National Institute of Technology (NIT-K), Surathkal** — **Mangalore, India**
  Bachelor of Technology Electronics & Communication Engineering — *2011–2015*

Fellowships

Facebook PhD Fellowship
Hardware & Software for Machine Learning Research — *May 2019–May 2021*

Experience

- **University of Illinois at Urbana-Champaign** — **Urbana,IL**
  Graduate Research Assistant — *Aug 2017–Present*
  - Building energy efficient accelerators for deep learning applications such as video and image recognition.
  - Developing programmable accelerators to achieve the efficiency of ASICs and the flexibility of general-purpose processors.
- **Facebook Research** — **Menlo Park**
  Research Intern — *May 2019–August 2019*
  - Developed a gradient based method for efficient accelerator-algorithm mapping space search.
  - Developed method not only outperformed other search methods, but also achives lower cost within fewer iterations.
- **ARM/ARM Research** — **Bangalore/San Jose**
  Graduate Engineer — *July 2015–July 2017*
  - Designed a sub-10mw edge accelerator for CNN inference.
  - Worked closely with architects to enable CPU-GPU coherency on mobile SoCs; a crucial step in enabling heterogeneity in modern SoCs.
  - Design & verification of state-of-the-art SoCs with accelerators such as Graphics, Video and Display.
- **Indian Institute of Science (IISc)** — **Bangalore**
  Research Intern — *May 2015–June 2015*
  - Worked in Super-computer Centre (SERC) that hosts India's fastest Supercomputer.
  - Explored the suitability of ARM architecture for micro-servers.
- **ARM** — **Bangalore**
  Graduate Intern — *May 2014–July 2014*

- Worked on Error Control Coding for L3 caches for server class ARM CPUs.
- Worked on verification of many-core ARM server class SoCs.

- **India Innovation Labs** — **Bangalore**
  Graduate Intern — *May 2013–July 2013*
  - Worked on accelerating image processing workloads on GPUs using OpenCL.

Publications

[1] *ExTensor: An Accelerator for Sparse Tensor Algebra*, **Kartik Hegde**, Michael Pellauer, Hadi Asghari-Moghaddam, Michael Pellauer, Neal Crago, Aamer Jaleel, Edgar Solomonik, Jo el Emer, and Christopher W. Fletcher, 52nd International Symposium on Microarchitecture, MICRO'19

[2] *Buffets: An Efficient and Composable Storage Idiom for Explicit Decoupled Data Orchestration*, Michael Pellauer, Yakun Sophia Shao, Jason Clemons, Neal Crago, **Kartik Hegde**, Rangarajan Venkatesan, Stephen W. Keckler, Christopher W Fletcher, Joel Emer, 24th International Conference on Architectural Support for Programming Languages and Operating Systems, ASPLOS'19

[3] *Morph: Flexible Acceleration for 3D CNN-based Video Understanding*, **Kartik Hegde**, Rohit Agrawal, Yulun Yao, Christopher Fletcher, 51st International Symposium on Microarchitecture, MICRO'18

[4] *UCNN: Exploiting Computational Reuse in Deep Neural Networks via Weight Repetition*, **Kartik Hegde**, Jiyong Yu, Rohit Agrawal, Mengjia Yan, Micheal Pelleaur, Christopher Fletcher, 45th International Symposium on Computer Architecture, ISCA'18

[5] *Adaptive Reconfigurable Architecture for Image Denoising.*, **Kartik Hegde**, Vadiraj Kulkarni, R. Harshavardhan and Sumam David, In Parallel and Distributed Processing Symposium Workshop, IPDPS'15

[6] *High Speed FFT for GPGPUs*, **Kartik Hegde**, Student Research Symposium, 21st International Conference on High Performance Computing, HiPC'14. Best Student Research Paper Award

Selected Projects

Academic Projects

- **Dense-Sparse Hardware Accelerators**
  - More and more applications, such as CNN inference, rely on dense-sparse tensor computations.
  - Project aims to design a fully programmable hardware accelerator for dense-sparse computations.[1]
- **Flexible Accelerator Design for Video Understanding**
  - Video understanding is a compute intensive application of deep learning, that requires hardware acceleration.
  - We developed novel hardware-software codesigned accelerator for 3D-CNNs to achieve significant improvement in efficiency [3].
- **Exploiting Weight Repetition in Modern DNNs**
  - We observed that a large amount of computation in DNNs is redundant due to repeated weights.
  - Our acceleator, UCNN, exploits such repetition to achieve significant gains in efficiency on modern DNN inference [4].
- **Dynamic Run-time Reconfiguration in FPGAs**
  - We used the feature of Dynamic reconfiguration of hardware on FPGAs to flexibly change the hardware based on the needs of the software in run-time.
  - We demonstrated significant improvements in performance in real-life image denoising on FPGAs [5].
- **Optimizing FFT Kernel for low-end GPUs**
  - We developed new kernels that are optimized to low-end GPUs, often found in edge/mobile devices.
  - We implemented a highly optimized FFT Kernel that performed better than several industry standard

implementations on low-end GPGPUs [6].

Industry Projects

- **Sub-10mw hardware accelerator for CNN Inference**
  - Used novel *dataflow* techniques to boost the efficiency and data reuse in the accelerator.
  - Accelerator could support CNN/RNN based inference, and scalable in terms of compute and memory.
- **Smart Scheduler for Accumulators**
  - Supported arbitrary sized data-sets in arbitrary order with any number of pipeline stages.
  - Advanced power saving methods and throughput optimization helped achieve up to 50% improvement over previous works.
- **Enabling CPU-GPU Coherency in modern SoCs**
  - Worked closely with architects to analyze the performance of CPU-GPU coherency implementations.
  - Several bugs found during the project helped seamless integration of system coherernсy in the SoC.
- **L3 Cache in modern CPUs with Error Control Coding**
  - Enabling RAS (Reliability, Availability and Serviceability) in modern CPUs is of paramount importance.
  - This project was aimed to test the ECC in L3 caches with techniques such as random error injections.

Relevant Courses

- **Mathematics:** Soft Computing, Discrete Mathematical Structures, Cryptography, Pattern Recognition
- **Computer Architecture:** Digital Electronics & Computer Architecture, Computer Organization and Design, Digital System Design, Embedded Systems, Microprocessors, DSP Architectures, Parallel Computer Architectures, VLSI design, Low power VLSI design, Advanced Operating Systems
- **DSP:** Digital Signal Processing, Digital Signal Compression, Linear Systems & Signals
- **Independent Coursework:** Electronics(6.002x,MIT), Python(6.00x,MIT), Machine Learning, Parallel Programming, Neural Networks for Machine Learning

Technical skills

- **Programming:** Verilog (Advanced), Python (Advanced), C (Advanced), VHDL (intermediate), OpenCL (intermediate), CUDA (intermediate), Perl (Beginner)
- **EDA/Simulation Tools:** Mentor Graphics Questasim, Modelsim, Synopsys VCS, Cadence Incisive
- **Environments/Frameworks:** Xilinx Vivado, Altera Quartus, ARM Keil μVision, MATLAB, TensorFlow, Caffe

Awards and Honors

- Best Student Research Paper award at HiPC-2014.
- Received HR&D Ministry's Fellowship, for top 1% scorers of Pre-University exams in India.
- Ranked 139th in Engineering Entrance Exam out of 100,000+ applicants.
- Outstanding Student Award, 2010 (Pre-University College)
- Academic Excellence award, 2011 (Pre-University College)
- Stood third in National Level Science Exhibition, 2009.