Kartik Hegde

+1 217 819 7789 • kvhegde2@illinois.edu
www.kartikhegde.net

I am a computer architect passionate about building domain specific, highly programmable hardware accelerators. My core research agenda is designing architectures that are at the sweet-spot in the trade-off spectrum between specialization and programmability.
I have been fortunate to have had the opportunity to build real-world products as a design engineer in industry, develop successful prototypes to drive new products as a researcher in industrial R&D, and bring crazy ideas into reality as a PhD student in academia.

Education

- **University of Illinois at Urbana-Champaign** — **Urbana, IL**
  Ph.D. in Computer Science — *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 Infrastructure for Machine Learning — *May 2019–May 2021*

Professional Experience

- **University of Illinois at Urbana-Champaign** — **Urbana,IL**
  Graduate Research Assistant — *Aug 2017–Present*
  - Building energy efficient, high-performance accelerators for areas such as deep learning [4,5], sparse tensor algebra[2], etc.
  - Auto-tuning and hyper-parameter search using Machine Learning[1].
  - Developing highly programmable accelerators to achieve the efficiency of ASICs and the flexibility of general-purpose processors (under work).
- **NVIDIA Research** — **Remote**
  Research Intern — *May 2021–August 2021*
  - Building Format Agnostic Tensor Expressions (FATE): A novel programming paradigm to program sparse tensor kernels in a compression format agnostic manner.
  - Building a specialized hardware that uses FATE as the ISA.
- **Facebook AI Research (FAIR)** — **Boston**
  Research Intern — *Jan 2020–May 2020*
  - Developed a novel programmable control paradigm for efficient hardware accelerators targeting irregular workloads.
  - Used the above to design a hardware accelerator for recommendation systems.
- **Facebook Reality Labs (Oculus Research)** — **Menlo Park**
  Research Intern — *May 2019–August 2019*

- Developed a gradient based method for efficient accelerator-algorithm mapping space search, which is a notoriously difficult combinatorial search problem.
- Proposed method approximates the search space as a smooth surrogate and uses gradients to incrementally update any valid mapping to optimal.

ARM/ARM Research — **Bangalore/San Jose**
Graduate Engineer — *July 2015–July 2017*

- Designed a sub-10mw edge accelerator for CNN inference, first project in the company targeting ML accelerator design, which further inspired multiple research projects and products.
- 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.

Peer Reviewed Publications

1. *DDGE: A Novel Control Paradigm for Irregular Algorithms*, **Kartik Hegde**, Vikram Sharma Mailthody, Christin David Bose, Shivam Potdar, Brandon Reagen, Wen-mei Hwu, Hsien-Hsin Sean Lee, and Christopher W. Fletcher, Under Review
   - Proposes a novel control paradigm that extracts irregular parallelism from irregular workloads. (**Next Generation Programmable Accelerators**)
   - Key insight is to combine the advantages of dataflow architectures and Von-Neumann machines, where an explicit dataflow graph is sequenced through with a program counter to discover parallelism in run-time; leading to an efficient frontier based computing.
2. *Mind Mappings: Enabling Efficient Algorithm-Accelerator Mapping Space Search*, **Kartik Hegde**, Po-An Tsai, Sitao Huang, Angshuman Parashar, Vikas Chandra, and Christopher W. Fletcher, ASPLOS 2021
   - Proposes a gradient-based method for the combinatorial search problem of algorithm-accelerator mapping space search. (**Machine Learning for Systems**)
   - Key insight is to approximate the search space as a smooth-function using a DNN based surrogate, hence making it differentiable, allowing gradients based search.
3. *ExTensor: An Accelerator for Sparse Tensor Algebra*, **Kartik Hegde**, Michael Pellauer, Hadi Asghari-Moghaddam, Michael Pellauer, Neal Crago, Aamer Jaleel, Edgar Solomonik, Joel Emer, and Christopher W. Fletcher, 52nd International Symposium on Microarchitecture, MICRO'19 MICRO Top-picks in Computer Architecture: Honorable Mention
   - Proposes a highly programmable accelerator for sparse tensor algebra. (**Domain-Specific Accelerators**.)
   - Key insight is to perform intersections at different granularity to aggressively skip work in sparse tensor operations.

4. *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 MICRO Top-picks in Computer Architecture: Honorable Mention
   - Proposes a reusable and composable storage idiom for programmable hardware accelerators. (**Agile Hardware Design**)
   - Modern hardware accelerators use explicit decoupled data orchestration for higher efficiency, which results in complicated buffer hierarchies, increasing the design time. Buffets are reusable storage idioms designed for hardware accelerator with superior efficiency over caches, scratch pads and FIFOs, with support for fine-grained synchronization.
5. *Morph: Flexible Acceleration for 3D CNN-based Video Understanding*, **Kartik Hegde**, Rohit Agrawal, Yulun Yao, Christopher Fletcher, 51st International Symposium on Microarchitecture, MICRO'18
   - Proposes a flexible hardware accelerator for 3D-CNNs for video understanding. (**Flexible Hardware Accelerators**)
   - Key insight is that different problem shapes require different dataflows for maximum efficiency, which can be enabled in hardware via simple programmable FSMs.
6. *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
   - Proposes a hardware accelerator to exploit repetition in DNN parameters. (**Hardware Accelerators for Machine Learning**)
   - Key insight is that the recent trend on DNN model compression via quantization and pruning is leading to fewer unique values in the network, which opens up opportunities to save computations with simple algebraic re-associations.
7. *Adaptive Reconfigurable Architecture for Image Denoising.*, **Kartik Hegde**, Vadiraj Kulkarni, R. Harshavardhan and Sumam David, In Parallel and Distributed Processing Symposium Workshop, IPDPS'15
   - Proposes an adaptive hardware based on partial reconfiguration for image denoising. (**Partial Reconfiguration in FPGAs**)
   - Different type of noises found in images require different denoising filters and the key insight is to use partial reconfiguration to instantiate the right type of filter based on noise detection, thereby saving area.
8. *High Speed FFT for GPGPUs*, **Kartik Hegde**, Student Research Symposium, 21st International Conference on High Performance Computing, HiPC'14. Best Student Research Paper Award
   - Proposes a modified Cooley-Tuckey FFT algorithm for low-end GPGPUs. (**GPUs for general purpose computing**)

Review Experience

- IEEE Computer Architecture Letters
- IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems
- IEEE Internet of Things Journal
- Journal of Systems Architecture

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
- **Machine Learning:** Introduction to optimization, Machine Learning, Reinforcement Learning
- **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)
- **Machine Learning:** TensorFlow, PyTorch, Caffe2
- **EDA/Simulation Tools:** Mentor Graphics Questasim, Modelsim, Synopsys VCS, Cadence Incisive
- **Environments/Frameworks:** Xilinx Vivado, Altera Quartus, ARM Keil μVision, MATLAB