Rangeen Basu Roy Chowdhury

North Carolina State University
Raleigh, NC, 27695
rangeen.basu.roy.chowdhury@intel.com
www.rangeen.org

Area of Expertise

Energy-efficient Microprocessor Architecture, Multicore CPU Architecture, CPU Performance and Power Modelling, 3D Die Stacked Architectures, RTL Design, Synthesis & Physical Design

Education

- **Doctor of Philosophy, Computer Engineering** **Jan 2014 – Oct 2016**
  North Carolina State University (NCSU)
  Address: North Carolina State University, Raleigh, NC-27695
  Dissertation: AnyCore: Design, Fabrication, and Evaluation of Comprehensively Adaptive Superscalar Processors.
  Advisor: Prof. Eric Rotenberg
  GPA: 4.0/4.0

- **Master of Science, Computer Engineering** **Aug 2011 – Dec 2013**
  North Carolina State University (NCSU)
  Address: North Carolina State University, Raleigh, NC-27695
  Project: Integrating OpenSparcT2 Memory Hierarchy into FabScalar Chips.
  Advisor: Prof. Eric Rotenberg
  GPA: 4.0/4.0

- **Bachelor of Technology, Electronics & Communication Engineering** **Aug 2005 – May 2009**
  National Institute of Technology (NIT) Durgapur
  Address: Mahatma Gandhi Avenue, Durgapur-713209, West Bengal, India
  Senior Project: Design and Evaluation of a High Speed Power Efficient OpAmp.
  GPA: 8.01/10.00

Professional Experience

- **Intel Corporation - Silicon Architecture Engineer (CPU Architect)** (40 Hours/week) **Oct 2016 – Present**
  Address: 3600 Juliette Lane, Santa Clara, CA-95054
  Manager: Sebastian Winkel
  - I work in the CPU architecture team researching ideas to improve the performance and power efficiency of Intel's big CPU cores. These cores power majority of the world's personal computers and servers and are sold under the brand names Xeon, Core-i3, Core-i5, and Core-i7.

- **North Carolina State University - Research Assistant** (20 Hours/week) **Jan 2012 – Aug 2016**
  Address: North Carolina State University, Raleigh, NC-27695
  Dissertation: AnyCore: Design, Fabrication, and Evaluation of Comprehensively Adaptive Superscalar Processors.
  Advisor: Prof. Eric Rotenberg
  - At North Carolina State University, I was part of the **Center for Efficient, Scalable, and Reliable Computing (CESR)** group, which performs research in various aspects of computer architecture. I was involved in two different projects in varying capacities. Both projects involved researching adaptivity (dynamic reconfigurability) in CPU architectures but they followed different paradigms of adaptivity.

AnyCore: The goal of my dissertation research was to design and evaluate a comprehensively adaptive CPU core, called *Anycore*, which can reconfigure itself as per the needs of the program running on the CPU, to improve performance while optimizing power consumption. A well designed adaptive CPU core can potentially outperform a single ISA heterogeneous multi-core processor, which is comprised of multiple differently tuned CPU cores and adapts by migrating the program between these cores. Hence, one of the primary goals of this project was to compare an adaptive CPU core with a heterogeneous multi-core processor. We also investigated hardware and software techniques that effectively utilize the hundreds of different configurations made available by an adaptive CPU core. Our findings were published in various papers and posters at conferences.

Anycore Chip: We fabricated an *Anycore* prototype using a ***130nm IBM process*** and used the chip for various experiments. The chip, named AnyCore-1 is the first academic prototype of an adaptive CPU core. Using this prototype, we researched various scheduling techniques to effectively use such an adaptive CPU.

Heterogeneity in 3D Stacked Architecture (H3): The goal of this project was to investigate the potential of single ISA heterogeneous multicore processor, wherein each CPU core is tuned for a different program behavior. A heterogeneous multicore processor improves performance and power efficiency by migrating the running program to the CPU core most optimal for each phase of the program. Although quite beneficial, the migration process is subject to overheads due to context save-restore and migration induced cache misses. The project aimed to reduce these inefficiencies in a heterogeneous multicore architecture by stacking CPU cores on top of each other using 3D stacked fabrication technology.

H3 Chip (Heterogeneity in 3D Stacked Architecture): We fabricated the H3 chip that consists of two 3D-stacked superscalar cores, a 1-wide superscalar and a 2-wide superscalar with respect to peak instruction fetch rate. This chip features high bandwidth connections between the two tiers to enable extremely fast context save-restore. The cache hierarchies of the two cores are also connected by low latency cross-tier buses to reduce migration induced cache misses. The chip was fabricated using ***GlobalFoundries 130nm*** process and the two tiers were bonded using ***Ziptronics*** face-to-face bonding process.

- **NVIDIA Corporation – CPU Architecture Intern** (40 Hours/week) **Aug 2014 –Dec 2014**
  Address: San Thomas Expressway, Santa Clara, CA-95050
  Project Title: Projecting next generation CPU power using *Top-N* inspired micro-benchmarks
  Manager: Aravindh Baktha
  - Worked on crafting micro-benchmarks inspired by a few of the most important mobile CPU benchmarks.
  - Developed a methodology for projecting power consumption of the next generation CPUs using these micro benchmarks.

- **Qualcomm Inc. – Interim Hardware Engineer** (40 Hours/week) **May 2014 –Aug 2014**
  Address: 8041 Arco Corporate Drive, Raleigh, NC-27617
  Project Title: Power estimation flow for the on-chip-network of a server grade CMP
  Manager: Jaya Ganasan
  - Worked as part of the team responsible for designing the on-chip network for Qualcomm's ARM server processor. I worked on the power analysis flow using Synopsys Prime Time and was responsible for creating the architectural power model for the various network components. The work involved making a test plan to obtain sufficient coverage for populating the power model database, automating power analysis flow to gather power data for each case, and creating an analytical power model for the design.

- **NVIDIA Corporation – CPU Design Intern** (40 Hours/week) **May 2012 – Aug 2012**
  Address: 20400 NW Amberwood Dr # 100, Beaverton, OR 97006
  Project Title: Improving Design for Debug (DFD) in Nvidia Denver processor
  Manager: Karol Menezes

- I worked with the Denver RTL team to design and implement a few key components of the "Design for Debug" features of the Denver CPU.
- I also worked closely with the designers of various units in the CPU to ensure sufficient debug coverage for those units.
- I was responsible for ensuring timing closure for all logic paths related to post silicon debug features.

- **Tejas Networks Ltd. - Research & Development Engineer** (40 Hours/week) **Aug 2009- Jun 2011**
  Address: Plot No. 25, JP Software Park, Electronics City, Phase-1, Hosur Road
  Bengaluru, Karnataka 560100, India
  Manager: Mahesh Prabhu
  - I was responsible for architecting and designing multi-gigabit serial communication protocols for chip-to-chip communication used in high capacity Ethernet switches. I was also involved in designing DDR2 and DDR3 memory controllers that were used as packet data stores before being forwarded to their destination. These designs were targeted for FPGAs the development process involved testing the designs on the actual FPGA boards in a test network.

- **Central Mechanical Engineering Research Institute - Research Assistant** **May 2008- May 2009**
  Address: CMERI, Mahatma Gandhi Road, City Center, Durgapur, West Bengal 713209, India
  Project Title: A CPLD based sub-state observer for fault detection in a mechanical system
  Advisor: Dr. Joydeb Roy Chowdhury
  - We developed a CPLD based observer that continuously monitors the condition of different components of an industrial ball stitching machine and alerts the operator when a component is about to fail. This "*about to fail state*", called a *sub state*, is not a directly observable state and can only be estimated by continuously monitoring other parameters of the component. We used training data gathered previously and linear regression model to estimate the sub state of the components.

Awards/Scholarships

- **ACM SIGARCH Student Travel Grant** for travel to ISCA 2015.
- **Certificate of Merit** for being among the top 0.1% of successful candidates in All India Secondary School Examination (10th Grade) 2003 in Science.

Publications

Ongoing Projects

- **R. Basu Roy Chowdhury**, E. Rotenberg. *Compiler assisted phase detection and scheduling on a system with an adaptive CPU.*

- E. Forbes, V. Srinivasan, **R. Basu Roy Chowdhury**, E. Rotenberg. *Heterogeneity in 3D – Design and evaluation of a 3D stacked heterogeneous architecture.*

Peer-reviewed Conferences

- V. Srinivasan, **R. Basu Roy Chowdhury**, E. Forbes, R. Widialaksono, Z. Zhang, J. Schabel, S. Ku, S. Lipa, E. Rotenberg, W. Rhett Davis and P. Franzon. *H3 (Heterogeneity in 3D): A Logic-on-logic 3D-stacked Heterogeneous Multi-core Processor.* ICCD, November 5-8, 2017.

- S. Ku, E. Forbes, **R. Basu Roy Chowdhury**, E. Rotenberg. *A Case for Standard-Cell Based RAMs in Highly-Ported Superscalar Processor Structures.* ISQED, March 13-15, 2017.

- R. Widialaksono, **R. Basu Roy Chowdhury**, Z. Zhang, J. Schabel, S. Lipa, E. Rotenberg, W. Rhett Davis, P. Franzon. *Physical Design of a 3D-Stacked Heterogeneous Multi-Core Processor.* 3DIC Conference, November 9-11 2016.

- **R. Basu Roy Chowdhury**, A. Kannepalli, S. Ku, E. Rotenberg. *AnyCore: A Synthesizable RTL Model for Exploring and Fabricating Adaptive Superscalar Cores.* ISPASS 2016*,* April 17-19 2016.

- E. Rotenberg, B. Dwiel, E. Forbes, Z. Zhang, R. Widialaksono, **R. Basu Roy Chowdhury**, N. Tshibangu, S. Lipa, W. Rhett Davis and P. Franzon. *Rationale for a 3D Heterogeneous Multi-core Processor.* Proceedings of the 31st IEEE International Conference on Computer Design (ICCD-31), pp. 154-168, October 2013.

Posters

- **R. Basu Roy Chowdhury**, A. Kannepalli, and E. Rotenberg. *AnyCore-1: A Comprehensively Adaptive 4-way Superscalar Core.* Poster session of Hot Chips 2016, August 21-23, 2016.

- S. Ku, E. Forbes, **R. Basu Roy Chowdhury**, E. Rotenberg. *A Case for Standard-Cell Based RAMs in Highly-Ported Superscalar Processor Structures.* Work-In-Progress Poster Session at DAC-53, June 5-9, 2016.

- E. Forbes, Z. Zhang, R. Widialaksono, B. Dwiel, **R. Basu Roy Chowdhury**, V. Srinivasan, S. Lipa, E. Rotenberg, W. R. Davis, and P. D. Franzon. *Under 100-cycle Thread Migration Latency in a Single-ISA Heterogeneous Multi-core Processor.* Poster session of Hot Chips 2015, August 23-25, 2015.

Talks and Workshops

- E. Forbes, **R. Basu Roy Chowdhury**, B. Dwiel, A. Kannepalli, V. Srinivasan, Z. Zhang, R. Widialaksono, T. Belanger, S. Lipa, E. Rotenberg, W. R. Davis, and P. D. Franzon. *Experiences with Two FabScalar-based Chips.* 6th Workshop on Architectural Research Prototyping (WARP'15), in conjunction with ISCA-42, June 14, 2015.

- **R. Basu Roy Chowdhury**, A. Kannepalli and E. Rotenberg. *FabScalar-RISCV*. 2nd RISC-V Workshop. Berkeley, CA, June 2015.

- **R. Basu Roy Chowdhury**. *Anycore-Towards a Universal Superscalar Core.* Computer Architecture Seminar at NC State University.

Professional Services

- Reviewer for Transactions on Architecture and Code Optimizations (TACO)
- Reviewer for The Computer Journal
- Member of Shadow Program Committee for Architectural Support for Programming Language and Operating System (ASPLOS) 2018
- External Reviewer for International Symposium on Workload Characterization (IISWC) 2017
- External Reviewer for International Parallel and Distributed Processing Symposium (IPDPS) 2017
- External Reviewer for International Conference on Distributed Computing Systems (ICDCS) 2017

Term Papers/Course Projects

- **R. Basu Roy Chowdhury** and Gregory Byrd. *A Case for Cache-Core Decoupling in 3D stacked chip multiprocessors*. The study was performed using the GEM5 full system simulator.

- **R. Basu Roy Chowdhury**. *Analysis of different FinFET design techniques used for Low-Power Robust SRAMs.*

- Implemented *demand paging* and *backing store* in Xinu for the x86 platform. A couple of different page replacement policies were also implemented and evaluated. Also wrote a *Linux like process scheduler* for Xinu and evaluated its performance and fairness.

- Developed a virtualization mechanism for ***CUDA*** applications to allow sharing ***GPU*** resources between Virtual Machine Instances. A remote procedure call (RPC) based client-server architecture was used to implement the virtualization mechanism.

- Implemented and evaluated a *Gshare branch predictor* in FabScalar RTL to predict multiple branches in the same cycle. This type of prediction is required to sustain the high fetch throughput required for a high ILP processor.

- Implemented a LLVM compiler pass for *profile driven Global Code Motion*. Also implemented a *majority voting based software fault tolerance* pass for the same compiler.

- Built a trace-driven configurable simulator for multilevel caches with inclusion/exclusion policies and performed design exploration for high cache-hit rates within a given budget of cache-size.

- Implemented a branch-predictor and quantified the performance of various types of branch-predictors while optimizing the prediction rate for a given memory budget.

- Implemented a superscalar pipeline simulator with out-of-order execution and Tomasulo's Algorithm based instruction scheduling.

Mentoring Experience

- Anil Kumar Kannepalli (Masters Student, graduated summer 2015): *Chip Bringup of the AnyCore Adaptive Superscalar Core.*

Teaching Experience

- Teaching Assistant at NCSU – ECE521 (Computer Design and Technology), Spring 2012

Relevant Coursework

- **Graduate Courses at NCSU:**

Computer Design and Technology (A+)
Advanced Microarchitecture (A+)
Architecture of Parallel Computers (A)
Advanced Parallel Architecture (A)
Data Center Design and Architecture (A)
Embedded System Design (A+)
Advanced Computer Design (A+)
VLSI System Design (A)
Digital ASIC Design (A)
Electronic System Level & Physical Design (A+)
Advanced Digital Electronics (B+)
Operating System Principles (A+)
Code Generation & Optimization (Compiler) (A)

- **Undergraduate Courses at NIT Durgapur:**

  Computer Organization and Architecture
  Embedded System Design
  Microprocessors and Microcontrollers
  Digital Electronics
  Integrated Electronics
  Computer Networks
  Introduction to System Programming

Abstracts

- V. Srinivasan, **R. Basu Roy Chowdhury**, E. Forbes, R. Widialaksono, Z. Zhang, J. Schabel, S. Ku, S. Lipa, E. Rotenberg, W. Rhett Davis and P. Franzon. "*H3 (Heterogeneity in 3D): A Logic-on-logic 3D-stacked Heterogeneous Multi-core Processor.*"
  Abstract: A single-ISA heterogeneous multi-core processor (HMP) is comprised of multiple core types that all implement the same instruction-set architecture (ISA) but have different microarchitectures. Performance and energy is optimized by migrating a thread's execution among core types as its characteristics change. Simulation-based studies with two core types, one simple (low power) and the other complex (high performance), has shown that being able to switch cores as frequently as once every 1,000 instructions increases energy savings by 50% compared to switching cores once every 10,000 instructions, for the same target performance. These promising results rely on extremely low latencies for thread migration. Here we present the H3 chip that uses 3D die stacking and novel microarchitecture to implement a heterogeneous multi-core processor (HMP) with ultra-fast thread migration capabilities. We discuss details of the H3 design and present power and performance results from running various benchmarks on the chip. The H3 prototype can improve performance of benchmarks by upto 25% while reducing power consumption by upto 20%.

- S. Ku, E. Forbes, **R. Basu Roy Chowdhury**, E. Rotenberg. "*A Case for Standard-Cell Based RAMs in Highly-Ported Superscalar Processor Structures.*"
  Abstract: Highly-ported memories are pervasive within superscalar processors. Accordingly, they have been targets for full-custom design using multi-ported versions of the 6T SRAM bitcell. Unfortunately, full-custom design of highly-ported memories is becoming exceedingly difficult in deep sub-micron technologies. This paper makes the case for implementing highly-ported memories with standard cells (flip-flops, muxes, and clock buffers). In lieu of exotic peripheral circuits for each port, standard-cell SRAMs use muxes. Consequently, area differences between full-custom and standard-cell designs are greatly reduced at a high number of ports. To also compete with full-custom memories in terms of timing and power, we introduce a standard-cell memory compiler with three key features: (i) per-row clock gating, (ii) a new tri-state based mux standard cell, and (iii) a modular layout strategy, which is the centerpiece of the memory compiler. For a 16-read/8-write 128-entry register file, our modular standard-cell memory consumes 13% more area and 4% more power, and is 35% faster, than the custom memory produced by FabMem. The automatic (built-in) robustness of standard cell designs further weigh in their favor, contrasted with exquisite transistor sizing/tuning of intertwined sub-circuits in a full-custom design.

- R. Widialaksono, **R. Basu Roy Chowdhury**, Z. Zhang, J. Schabel, S. Lipa, E. Rotenberg, W. Rhett Davis, P. Franzon. "*Physical Design of a 3D-Stacked Heterogeneous Multi-Core Processor.*"
  Abstract: With the end of Dennard scaling, three dimensional stacking has emerged as a promising integration technique to improve microprocessor performance. In this paper we present a 3D-SIC physical design methodology for a multi-core processor using commercial off-the-shelf tools. We explain the various flows involved and present the lessons learned during the design process. The logic dies were fabricated with GlobalFoundries 130 nm process and were stacked using the Ziptronix face-toface (F2F) bonding technology. We also present a comparative analysis which highlights the benefits of 3D integration. Results indicate an order of magnitude decrease in wirelengths for critical inter-core components in the 3D implementation compared to 2D implementations.

- **R. Basu Roy Chowdhury**, A. Kannepalli, S. Ku, E. Rotenberg. "*AnyCore: A Synthesizable RTL Model for Exploring and Fabricating Adaptive Superscalar Cores.*"
  Abstract: Adaptive superscalar cores have the ability to dynamically adjust their execution resources to match the instruction-level parallelism (ILP) of different program phases. The goal of adaptivity is to maximize performance in as energy-efficient a manner as possible. This is achieved by disabling execution resources that contribute only marginally to performance for the code at hand. Researchers have proposed many adaptive features, including structures, superscalar width, and pipeline depth. The benefits of adaptivity are eroded by its circuit-level overheads. Unfortunately, circuit-level overheads cannot be effectively estimated or appreciated without a hardware design. To this end, we developed a register-transfer-level (RTL) design of a highly adaptive superscalar core, called AnyCore. AnyCore can be used to quantify logic overheads of an adaptive core with respect to fixed cores, synthesize and compare different adaptive cores, synthesize and compare an adaptive core to a multi-core comprised of multiple fixed core types, and fabricate adaptive superscalar cores. We provide examples of these use-cases.

- E. Forbes, V. Srinivasan, **R. Basu Roy Chowdhury**, E. Rotenberg. *"Heterogeneity in 3D – Design and evaluation of a 3D stacked heterogeneous architecture.*"
  Abstract: In the H3 project, we are building a 3D Heterogeneous Multi-core Processor ("H3" stands for "heterogeneity in 3D"). The 3D IC will consist of two 3D-stacked superscalar cores. The two cores are a 1-wide superscalar and a 2-wide superscalar, with respect to peak instruction fetch rate. The 1-wide core has smaller ILP-extracting structures (active list, issue queue, LQ/SQ, and physical register file) than the 2-wide core. Both cores have three execution lanes (3-wide issue): simple/complex integer ALU lane, load/store lane, and branch lane. H3 features Fast Thread Migration (FTM) and Cache-Core Decoupling (CCD). FTM is a bulk swap of the architectural register state of the two cores, enabling fine-grain switching of threads between the two cores. CCD is the ability for a core to access either its caches or the other core's caches. This can be used to avoid migration-induced misses at the expense of higher hit latency due to crossing clock domains. Migrations can be initiated either from an off-chip interrupt signal (referred to as a global migration) or from the program itself via a new migrate instruction (a local migration). FTM and CCD require low-latency high-bandwidth interconnect. 3D face-to-face bonding is the ideal technology to meet this requirement.

- **R. Basu Roy Chowdhury** and Gregory Byrd. "*A Case for Cache-Core Decoupling in 3D stacked chip multiprocessors*."
  Abstract: Multi-threaded applications running on a chip multi-processor (CMP) exhibit different types of communication patterns. In some applications, each core most frequently communicates with its neighboring cores leading to a lot of coherence traffic being exchanged between these two entities. Examples of such applications are *ocean_cp, lu_cb, blackscholes* and *barnes* from the Parsec and Splash2x benchmark suites. With stock MESI directory based protocol, the directory acts as one level of indirection between the two neighboring cores. Eliminating this extra indirection might reduce the runtime for these applications. Providing direct connectivity between neighboring L1 caches is a way to eliminate this indirection.

  In a chip multiprocessor, cores can be designed and floorplanned in a way to allow fast connectivity between L1 caches of neighboring cores, essentially doubling the cache capacity while using only a modest amount of extra logic. The caches still remain relatively simple and only a few additional transient states are added to the coherence protocol. The microarchitecture is similar to Cache-Core Decoupling (CCD) proposed by Rotenberg et. al.. While the original CCD mechanism tries to avoid compulsory cache misses when a process migrates to a neighboring core and primarily focusses on multiple single threaded workloads, the approach can be generalized to increase cache capacity and reduce miss latency. The original CCD also does not deal with coherence issues. Although CCD can be implemented in a 2D layout, 3D die stacking can result in a much more efficient floorplan and lower latencies for cross-core cache accesses.

  The goal of this project is to design the cache-core decoupling mechanism in the context of a CMP that deals with multi-threaded applications and handles coherence issues correctly. If implemented correctly, it is also possible to accelerate single threaded programs by effectively increasing the L1 cache capacity of a core. If a core's neighbor is turned off or is idle, the neighbor's cache can be used as an extension to the core's own cache.