U.S CITIZEN • PORTLAND, OREGON, USA, 97229 •
CHEEMA.UMER@GMAIL.COM • 408-858-5362

# UMER CHEEMA

*Deep Learning Hardware Engineer*

## PROFESSIONAL SUMMARY

- Extensive experience in Accelerator Architectures, Neural Processing Unit (NPU), FPGAs, Micro-architecture/RTL Development, PPA Optimization, EDA Software Development, VLSI Physical Design Validation and Embedded Systems.
- Skilled in Computer Architecture, Digital Circuit Design, FPGA Based Accelerators and Microprocessor-based system Design
- PhD dissertation focused on Accelerator Architectures for memory and compute-intensive algorithms.
- Taught graduate and undergraduate level courses: Digital System Design, Microprocessor based System Design and Computer Architecture.

## EMPLOYMENT HISTORY

**DEEP LEARNING HARDWARE ARCHITECT** **2021 - Present**
***Intel Corporation*** ***Hillsboro, OR***

- Architect next-gen NPUs for Intel's client products, ensuring adaptability to evolving neural network needs.
- Developed HW-SW Codesign framework (sprLUT) for approximating Activations functions within NPU, boosting Transformer network performance by up to 62%.
- Identify Power, Performance, and Area (PPA) improvements to meet the customer defined goals.
- Enhanced power efficiency of the NPU's central Data Processing Unit by 20% through optimal data type selection and configuration related optimizations.
- Refined sprLUT framework over multiple NPU generations, enhancing numeric accuracy, approximation range, and data-type support.
- Lead cross-functional initiatives with RTL, Modeling, Validation, and Compiler teams to integrate cutting-edge features into NPU IP.
- Pioneer functional models, evaluate performance metrics, prototype RTL designs, and secure patents, pushing boundaries in Deep Learning Hardware.
- Lead cross-functional initiatives, integrating cutting-edge features into NPU IP and authoring High-Level Architecture Specifications for future generations.

**EDA TOOLS SOFTWARE ENGINEER** **2016 - 2021**
***Intel Corporation*** ***Hillsboro, OR***

- Enhanced run-time and memory efficiency of vital software for Design Rule Checking by 60% - Recognized by Intel Divisional Award.
- Led development and integration of software suite for physical design validation of top process nodes.
- Led exceptional and timely debug for a lead product that led to timely tape-in of critical Intel product - Recognized by Intel Divisional Award.
- Conducted data analysis to improve quality and reliability of latest process technologies.
- Provided debugging and tape-in support to product teams across the company.
- Conducted in-depth analysis of process limitations (e.g. Reliability, Optical Proximity Correction, yield, Lithography), contributing to the development of robust design rule checks for advanced process technologies.
- Led meticulous code reviews and performance optimizations, resulting in significant runtime and memory improvements for critical company-wide software.
- Conducted precise performance audits, identifying bottlenecks, and implementing solutions that enhanced software reliability.

**ASIC / FPGA DESIGN & VERIFICATION ENGINEER** **2008 - 2011**
***Whizz Systems Inc***

- Design, RTL development and validation of AMBA Bus-interface for the USB 3.0 Host controller
- Designed Micro-blaze systems on Xilinx FPGAs, boosting efficiency and reliability.
- Developed specifications, ensuring clear project guidelines and smooth execution.
- Implemented RTL logic, enhancing processing speed and system performance.
- Created UVM environments for RTL, enabling comprehensive verification.
- Conducted static timing analysis, performance optimizations and RTL implementation on FPGAs.

**ADJUNCT LECTURER** **2014 - 2022**
***University of Illinois at Chicago*** ***Chicago, IL/Virtual***

- Deliver lectures and develop course materials for "ECE465: Digital System Design" and "ECE 367: Microprocessor Based System Design".
- Design engaging projects and assessments, enhancing student understanding.
- Hold office hours, providing personalized support and guidance.
- Facilitate a collaborative learning environment, fostering academic growth.
- Implement innovative teaching methods, leading to improved student performance.

**TEACHING ASSISTANT** **2011 - 2013**
***University of Illinois at Chicago*** ***Chicago, IL***

- Assisted the following courses: ECE367 Lab, ECE465, "ECE267: Computer Organization & Design", "ECE466: Advanced Computer Architecture", "ECE101: Introduction to ECE"
- Conducted in-depth analysis of student performance, identifying trends and providing actionable insights to enhance educational outcomes.
- Provided personalized support to students, improving understanding and academic performance through tailored guidance and resources.

## SKILLS

Accelerator Design *(Expert)*,
Micro-architecture *(Expert)*,
Digital Logic Design *(Expert)*,
NPU Architecture *(Expert)*,
Hardware-Software Co-design *(Expert)*,
FPGA Design *(Expert)*, Verilog *(Expert)*,
System Verilog *(Expert)*,
Custom Silicon Architecture *(Expert)*,
Hardware Programming *(Expert)*,
Numerics *(Expert)*, IEEE754 *(Expert)*,
Power Performance Area (PPA) optimization *(Expert)*,
Embedded Systems *(Expert)*,
Hardware Prototyping *(Expert)*,
Software Development *(Expert)*,
High-Performance Computing *(Expert)*,
Computer Architecture *(Expert)*,
Python *(Experienced)*, Java *(Skillful)*,
Pytorch *(Beginner)*, OpenCL *(Skillful)*,
GPU Programming *(Skillful)*,
Assembly Language *(Expert)*, C *(Experienced)*,
VHDL *(Experienced)*, GIT *(Experienced)*,
High Level Synthesis *(Experienced)*,
Embedded Software *(Experienced)*,
MATLAB *(Expert)*, ModelSim *(Expert)*,
Algorithms *(Expert)*, Data-structures *(Expert)*,
Simulations *(Expert)*, Emulation *(Expert)*,
Deep Learning *(Experienced)*,
Neural Networks *(Experienced)*,
VLSI *(Experienced)*, GitLab *(Experienced)*,
Machine-Learning *(Experienced)*,
Object Oriented Programming *(Experienced)*,
Code-reviews *(Experienced)*,
AMBA Bus Protocol *(Experienced)*,
Design Rule Checking *(Expert)*,
Patent Generation *(Expert)*,
Data Structures *(Experienced)*,
GitHub *(Experienced)*, Jira *(Experienced)*,
Signal Processing *(Experienced)*,
SQL *(Experienced)*, C++ *(Skillful)*,
CPU Design *(Skillful)*,
Hardware Verification *(Skillful)*,
Scripting (Perl, TCL) *(Skillful)*,
Embedded Design *(Experienced)*,
Customer Debug *(Experienced)*.

## LINKS

LinkedIn: www.linkedin.com,
Publons: publons.com,
Google Scholar: scholar.google.com.

**RESEARCH ASSISTANT** **2012 - 2013**
***University of Illinois at Chicago*** ***Chicago***

- Collaborated with nursing and engineering teams on interdisciplinary health projects.
- Analyzed health records to identify key parameters for patient care improvement.
- Investigated nursing practices impacting death anxiety and anticipatory grieving.
- Utilized data analysis skills to enhance understanding of patient care outcomes.

## EDUCATION

**PH.D. ELECTRICAL & COMPUTER ENGINEERING** **Jan 2011 - May 2016**
***University of Illinois at Chicago*** ***Chicago***

Dissertation focused on High-performance Custom Silicon and programmable hardware Solutions for Memory and Compute-intense Applications:

- Memory-optimized and Power efficient Re-gridding architecture for accelerating Non-uniform FFT (NuFFT) on FPGAs - 9.6x improvement in power efficiency (FLOPS/Watt)
- High-performance architecture for accelerating Burrows Wheeler Transform (BWT) on FPGAs - 4x throughput improvement
- A hardware-efficient Architecture for Matrix Inversion - 80% reduction in floating-point resources on FPGAs - 4x throughput improvement

**MASTER OF SCIENCE (MS) IN ELECTRICAL & COMPUTER ENGINEERING** **Jan 2011 - May 2014**
***University of Illinois at Chicago*** ***Chicago***

**BACHELOR OF SCIENCE (B.SC.), ELECTRICAL ENGINEERING** **Sep 2004 - Aug 2008**
***University of Engineering & Technology*** ***Lahore***

## ADDENDUM

### SELECTED PUBLICATIONS & PATENTS

- Cheema U.I., et al. “Approximating Activation Functions In Neural Networks with Programmable Look-Up Table”, US. Patent Application 18/500,229.
- Cheema U.I., et al., “Accuracy-based approximation of activation functions with programmable look-up tables having area budget”. U.S. Patent Application 18/534,035.
- Cheema U.I. et al., “Approximating activation functions with taylor series”. U.S. Patent Application 18/346,992.
- Cheema U.I. et al., “Activation Function Approximation Based On Input Range Partition” PCT/US2024/047674
- Cheema U.I. et al., "PrepNN: Activation function-based preprocessor for compiler to enhance the Accuracy and Performance of Neural Networks on an AI accelerator" PCT/CN2024/121525
- Cheema, Umer I., et al., "Activation Function Approximation Based on Input Range Partition Activation Function Approximation Based on Input Range Partition" PCT/US2024/047674
- Cheema, Umer I.; et al., “Memory optimized Re-gridding Architecture for Non-Uniform Fast Fourier Transform” in IEEE Transactions on Circuits and Systems I: Regular Papers
- Cheema, Umer I. et al., "InvArch: A hardware efficient architecture for Matrix Inversion," in Computer Design (ICCD), 2015 33rd IEEE International Conference on , vol., no., pp.180-187, 18-21 Oct. 2015.
- Cheema Umer.I et al., "Power Efficient Re-gridding Architecture for accelerating Non-uniform Fast Fourier Transform Power Efficient Re-gridding Architecture for accelerating Non-uniform Fast Fourier Transform" 24th International Conference on Field Programmable Logic and Applications (FPL 2014)
- Cheema, Umer I., et al. "MedianPipes: An FPGA based Highly Pipelined and Scalable Technique for Median Filtering" Proceedings of 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA2015)
- Cheema, Umer I.; et al. "A high-performance architecture for computing burrows-wheeler transform on FPGAs," Reconfigurable Computing and FPGAs (ReConFig), 2013 International Conference on, Dec. 2013
- Raha, A.; Cheema, U.I “Dynamic Sparsity-Based Acceleration Of Neural Networks”, 18/543,356
- Pillai, Kamlesh; Cheema, U.I., “Configurable Multi-Precision Converter In Neural Network Accelerator”, PCT/US2024/044642
- Raha, A.; Cheema, U.I., et al., 2024. “Switchable one-sided sparsity acceleration”. U.S. Patent Application 18/476,594.
- Usman Tariq Cheema, Umer I. et al. “Power-Efficient and Highly Scalable Parallel Graph Sampling using FPGAs” Reconfigurable Computing and FPGAs (ReConFig), 2017 International Conference on, Dec. 2017
- Nash, Gregory T.; Cheema, Umer I., "Power-efficient RMA SAR imaging using pipelined Non-uniform Fast Fourier Transform," 2015 IEEE Radar Conference (RadarCon)
- "VPU-EM: An event-based modeling framework to evaluate NPU performance and power efficiency at scale", arXiv:2303.10271
- "Efficient hardware acceleration of emerging neural networks for embedded machine learning: An industry perspective", Book chapter Embedded Machine Learning for Cyber-Physical, IoT, and Edge Computing: Hardware Architectures

### SELECTED AWARDS

- Got multiple Intel awards demonstrating Intel values: Customer Obsessed, One-Intel and Quality
- Merit Scholarship for undergraduate studies for 4-years at University of Engineering & Technology, Lahore
- Awarded Dr. M.I.D Chughtai Medal by Forman Christian College for scoring overall first position out of over 1200 students

### SELECTED TALKS & PRESENTATIONS

- “Power-efficient Re-gridding architecture for NuFFT” at Halmstad University, Sweden; Bell Labs, USA; Intel, USA
- “High-Performance solutions for Memory & Compute intense applications” at WMU, CSULB, UCDenver
- Training workshop at SEECS, NUST on “Analyzing Design and Floor-planning using Xilinx PlanAhead tool” (Fall 2010)

### PROFESSIONAL SERVICES

- Peer Reviewer: IEEE Transactions on Circuits & Systems, IEEE Journal on Emerging and Selected Topics in Circuits & Systems, Healthcare Technology Letters
- Member Review Committee: IEEE International Symposium on Circuits & Systems (ISCAS2018,2019, 2025) and International Conference on Field-Programmable Technology (FPT'16, FPT'17, FPT'18)

### PROFESSIONAL MEMBERSHIPS

Senior Member at Institute of Electrical & Electronics Engineers (IEEE)