Fellow	PALLAB KUMAR / **NATH**

Host Organisation	Fraunhofer Institute for Integrated Circuits (IIS) – Erlangen, Germany

Scientific coordinator	Marco/ **Breiling**

I – SCIENTIFIC ACTIVITY DURING YOUR FELLOWSHIP

The overall objective of the ERCIM fellowship was to develop an efficient hardware architecture for the digital section of a single APU (Analog Processing Unit), which will be the part of a mixed-signal Deep Neural Network (DNN) inference accelerator (ADELIA) developed by Fraunhofer IIS.

In the existing design (ADELIA v2.0), Input Feature Maps (IFM) are stored in the input SRAM memory (e.g., 8 KB), which are accessed by the Address Generation Unit (AGU) and stored in the Input Shadow Buffer (SB) for further processing by the Crossbar Array (CA). The AGU generates the necessary addresses to access the SRAM. The existing design repeatedly accesses the same IFM, which is already fetched and available in the SB, hence incurs unnecessary memory access. The memory read/write consumes a significant amount of energy in any digital design. The proposed study targeted reducing unnecessary memory access and reusing the IFMs accessed from the memory.

During my fellowship, I have proposed two designs to reduce memory access in the APU:

Design 1:

In this design, each APU consists of two levels of memory (e.g., 8KB and 2KB memory blocks) hierarchy to reduce larger memory access and reuse of the already fetched data. When the contents of the cache (2KB) ring buffer crosses a given layer threshold value, the AGU generates the addresses and starts filling the SB. For computing the next Output Feature Map (OFM), the design utilized the overlapped data that is already available in SB, hence reducing memory access.

Design 2:

The basic difference between design 1 and design 2 is that there is no SB in design 2.
The memory hierarchy is the same as design 1. For design 2, a 256-bit barrel shifter (right shift) is required. In design 2, the SB has been removed, and Shift Register (SR) is responsible for both data storage and right shifting. The common data between two overlapped receptive fields would be rearranged in the SR during processing time. After this shifting operation, the common data between the two overlapped receptive fields will remain in the SR; hence, no need to access the memory again. The remaining new data (not available in SR) is fetched from the cache memory and stored in the SR. This technique can reduce a significant number of memory accesses. Design 2 increased over all latency of the system compared to that of design 1.

Analysis:

The behavioural modelling for both designs is implemented using SystemC simulation software. The functionality of both designs is verified with the help of available configuration data for a Voice Activity Detection (VAD) neural network. The proposed two designs reduce a significant amount of memory access (31%) compared to that of the existing design.

II – PUBLICATION(S) DURING YOUR FELLOWSHIP

A conference paper is currently being prepared based on the research conducted at Fraunhofer IIS, Erlangen, Germany.

III – ATTENDED SEMINARS, WORKHOPS, CONFERENCES

During the fellowship period, I did not participate in any seminars, workshops, or conferences. However, I conducted a research visit to the Chip Design for Embedded Computing Laboratory at Technische Universität Braunschweig from November 4 to 5, 2024. This visit was undertaken under the mentorship of Professor Guillermo Paya Vaya, Chair of Chip Design for Embedded Computing, with the objective of initiating a research collaboration in the field of embedded computing systems.

IV – RESEARCH EXCHANGE PROGRAMME (REP)

As part of the Research Exchange Programme (REP), I visited the Norwegian University of Science and Technology (NTNU) in Trondheim from October 7–11, 2024. During this time, I was mentored by Prof. Per Gunnar Kjeldsberg, Head of the Department of Electronic Systems. The experience was immensely enriching, offering valuable insights and discussions that significantly contributed to my research. It also laid the groundwork for future collaborative opportunities.

V – SUM UP OF THE FINAL SCIENTIFIC REPORT FOR THE ERCIM NEWSLETTER

Dr. Pallab Kumar Nath is currently an Assistant Professor at Pandit Deendayal Energy University, Gandhinagar, India. His research focuses on digital VLSI architecture design, FPGA-based system design, and hardware accelerators for Machine Learning algorithms. As part of his ERCIM Fellowship, he was hosted by Fraunhofer IIS, where he worked on the project *"Architecture Design Space Exploration for Mixed-Signal DNN Accelerators."* In this project, he proposed an energy-efficient data transfer technique that reduces input memory access and maximizes data reuse. During the fellowship, he also engaged in collaborative research with NTNU Norway and TU Braunschweig.

For further details, he can be contacted at pkniitkgp@gmail.com or via LinkedIn: https://www.linkedin.com/in/pallabkumarnath/