4-2017

Artificial Neural Network Based Prediction Mechanism for Wireless Network on Chips Medium Access Control

Ranjith Murugesan
rm2575@rit.edu

Artificial Neural Network Based Prediction Mechanism for Wireless Network on Chips Medium Access Control

By
Ranjith Murugesan

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Computer Engineering

Supervised by
Dr. Amlan Ganguly

Department of Computer Engineering
Kate Gleason College of Engineering
Rochester Institute of Technology
Rochester, NY
April 2017

Approved By:

Dr. Amlan Ganguly
Thesis Advisor – R.I.T. Dept. of Computer Engineering

Dr. Andreas Savakis
Secondary Advisor – R.I.T. Dept. of Computer Engineering

Dr. Sonia Lopez Alarcon
Secondary Advisor – R.I.T. Dept. of Computer Engineering

To my beloved Parents, Mentors and Friends

Acknowledgements

I would like to express my immense gratitude to Dr. Amlan Ganguly and Naseef Mansoor for their invariable kindness and support throughout the completion of my Masters.

I would like to thank Md. Shahriar Shamim, Sandeep Aswath Narayana, Pratheep Joe Siluvai Iruthayaraj, my lab friends, for their extreme guidance and encouragement.

In addition, I would like to thank Dr. Andreas Savakis and Dr. Sonia Lopez Alarcon to be a part of my thesis committee.

Finally, I would like to thank my family for their good will and blessings and my friends for their moral support, without which this would not have been possible.

Abstract

As per Moore's law, continuous improvement over silicon process technologies has made the integration of hundreds of cores on to a single chip possible. This has resulted in the paradigm shift towards multicore and many-core chips where, hundreds of cores can be integrated on the same die and interconnected using an on-chip packet-switched network called a Network-on-Chip (NoC). Various tasks running on different cores generate different rates of communication between pairs of cores. This lead to the increase in spatial and temporal variation in the workloads, which impact the long distance data communication over multi-hop wire line paths in conventional NoCs. Among different alternatives, due to the CMOS compatibility and energy-efficiency, low-latency wireless interconnects operating in the millimeter wave (mm-wave) band is nearer term solution to this multi-hop communication problem in traditional NoCs. This has led to the recent exploration of millimeter-wave (mm-wave) wireless technologies in wireless NoC architectures (WiNoC). In a WiNoC, the mm-wave wireless interconnect is realized by equipping some NoC switches with an wireless interface (WI) that contains an antenna and transceiver circuit tuned to operate in the mm-wave frequency. To enable collision free and energy-efficient communication among the WIs, the WIs is also equipped with a medium access control mechanism (MAC) unit. Due to the simplicity and low-overhead implementation, a token passing based MAC mechanism to enable Time Division Multiple Access (TDMA) has been adopted in many WiNoC architectures. However, such simple MAC mechanism is agnostic of the demand of the WIs. Based on the tasks mapped on a multicore system the demand through the WIs can vary both spatially and temporally. Hence, if the MAC is agnostic of such demand variation, energy is wasted when no flit is transferred through the wireless channel. To efficiently utilize the wireless channel, MAC mechanisms that can dynamically allocate token

possession period of the WIs have been explored in recent time for WiNoCs. In the dynamic MAC mechanism, a history-based prediction is used to predict the bandwidth demand of the WIs to adjust the token possession period with respect to the traffic variation. However, such simple history based predictors are not accurate and limits the performance gain due to the dynamic MACs in a WiNoC. In this work, we investigate the design of an artificial neural network (ANN) based prediction methodology to accurately predict the bandwidth demand of each WI. Through system level simulation, we show that the dynamic MAC mechanisms enabled with the ANN based prediction mechanism can significantly improve the performance of a WiNoC in terms of peak bandwidth, packet energy and latency compared to the state-of-the-art dynamic MAC mechanisms.

Table of Contents

List of Figures

Tables

Chapter 1. Introduction

Integration of multiple processor cores on a single chip has emerged as a de-facto design choice in computer industry due to the power, heat and reliability constraints of a single processor core. Such integration of large number of processing cores on a single die is possible due to the continuing progress and integration capabilities in silicon technologies following Moore's law. However, in order to cater the ever-increasing computational needs, these multicore systems require to integrate tens to hundreds of cores on the same die. Hence, the design of an interconnection fabric between these cores has become a significant problem in multicore systems. Traditional bus based interconnection fabric are not compatible for such large multicore systems as they are not scalable beyond a few number of cores. This has led to the recent exploration of the network-on-chip (NoC) fabric that provides a scalable communication infrastructure between the cores in a multicore chip.

1.1 Network-on-Chip and challenges

The NoC is basically a communication subsystem on an integrated circuit; typically between intellectual property (IP) cores in a system-on-chip (SoC). The NoC technology applies networking theory and methods to on-chip communication and brings notable improvements over conventional bus and crossbar interconnections. The NoC separates the computational cores from the interconnection and communication needs, and provides a scalable plug-and-play network for packet based data communication. The common characteristic of the NoC architectures is that the IP cores communicate with each other through switches and links as shown in Fig. 1.

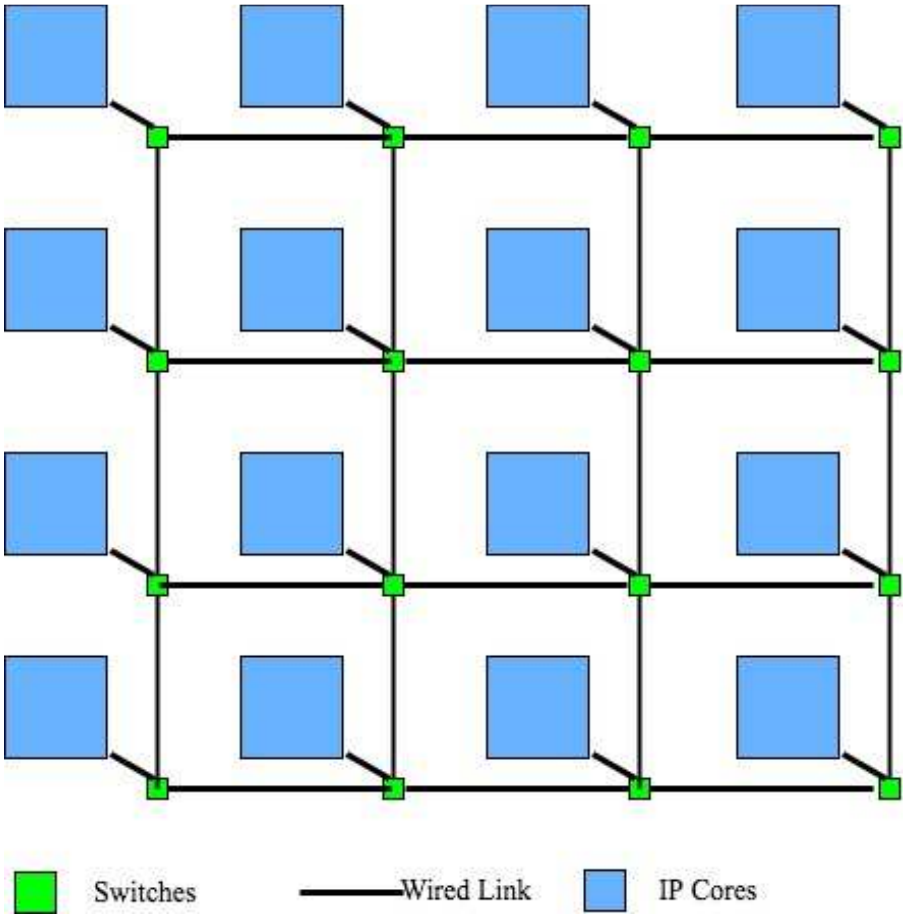

Fig 1: Mesh Network on Chip Architecture

In the NoC environment, wormhole switching is adopted to reduce the buffer requirement at the virtual channel (VC) based NoC switches. In the wormhole switching the packet are broken into fixed length flow control units or flits. The sizes of the flits are determined so that they can be transferred between adjacent switches in one clock cycle. The first flit or the header contains routing information that helps to establish a path from the source to destination. The subsequent body flits of the same packet follow this path to get routed to the destination. In order to connect the NoC switches many topologies has been explored in recent time. Some common NoC topologies are the Mesh, the

Folded-Torus and the Butterfly Fat-Tree. However, the performance benefits of such NoC topologies are limited due to the multi-hop communication over the metallic interconnect. Long-range metal wires in a mesh based NoC [38] and ultra-low-latency and low-power express channels between communicating cores [3] have been proposed to solve these multi-hop communication issues. However, the performance gains of these approaches are limited due to metal/dielectric-based interconnection. This has led to the exploration of novel interconnect technologies like on-chip photonic interconnect [5], multi-band RF transmission line interconnects (RFI) [6], and 3D integration [7].

The on-chip photonic interconnects are realized using on-chip optical waveguides, micro-ring resonators, and laser sources. With the data transmissions occurring at the speed of light, the optical interconnect have been predicted to enhance the bandwidth and reduce latency considerably [6]. However, the technological intricacy involved in manufacturing the photonic devices and integrating them with silicon-compatible circuits under area, power and delay constraints is a non-trivial challenge. On the other hand, multiband RF interconnects use wires as transmission lines to transfer data in the form of electromagnetic (EM) waves [7]. In this technique, the bandwidth can be increased by multiple access technique resulting in low latency data transfer at the speed of light using electromagnetic waves. However they are limited by the design of high frequency oscillators and filters on the chip for the transceivers. In 3D integration, multiple active layers are interconnected onto a single chip, which consequently reduces the hop-count and the average wire length of a single hop, thereby increases the performance. However, there is a temperature increase due to smaller footprint and higher resultant power densities, which cause high, heat dissipation [13] and require sophisticated cooling mechanism [14]. Also there is a difficulty in fabrication due to the issues with inter-layer alignments, bonding, inter-layer contact patterning [15] and increased risks of

manufacturing defects. Another alternative is NoCs equipped with wireless interconnects which form links between distance cores on the die enabling single hop communication. Such interconnection technology can be realized with CMOS compatible antenna and transceivers, making it a nearer term solution to the multi-hop issue of traditional wireline NoCs.

1.2 Wireless NoC

On-chip wireless interconnect is a promising alternative to the performance limitations in long-distance wired links. Advancement in on-chip antenna [22] and transceiver design in standard bulk CMOS technologies has made on-chip wireless interconnect possible. By utilizing such CMOS compatible transceiver and antennas operating in millimeter wave (mm-wave) band, research in recent time has shown that mm-wave wireless links can significantly improve the performance and energy-efficiency in on chip data transfer [8][38][39]. Generally, the mm-wave wireless NoC (WiNoC) architectures are hybrid architectures with both wireline and wireless interconnects. Some NoC switches are equipped with the wireless interfaces (WIs) that contains the antenna and transceiver to enable long range communication in a multicore environment. Furthermore, the wireless channel is shared among the WIs. In order to access the wireless channel in an interference free fashion, the WIs are also equipped with a medium access control (MAC) unit. The MAC unit enables the WIs to transmit through the energy efficient wireless medium. However, to utilize the full benefit of the novel mm-wave interconnect technology in a WiNoC, two critical challenges are required to be addressed. Firstly, to enable the energy-efficient communication using the shared wireless medium, design of efficient, simple and fair MAC mechanism is required. Secondly, in order to ensure efficient utilization of the wireless channel, the wireless bandwidth should be shared among the wireless interfaces (WIs) to maximize the performance and energy benefits of the wireless

interconnect. Many different MAC mechanisms to address these issues in the context of WiNoCs has been studied in literature. In the next section, we present different MAC mechanisms proposed for WiNoCs.

1.3 Issues in MAC for WiNoCs

In WiNoCs, the MAC mechanism enables the WIs to access the energy-efficient wireless medium. However, due to power and area constraints imposed by the on-chip environment, complex MAC mechanisms used in macro scale networks cannot be adopted directly for WiNoCs. Due to these constraints, simple and low overhead MAC mechanisms such as Frequency Division Multiple Access (FDMA), Code Division Multiple Access (CDMA) and Time Division Multiple Access (TDMA) has been proposed for WiNoC architectures. In a Frequency Division Multiple Access (FDMA) based MAC, design of transceivers operating in non-overlapping frequency range is required. However, design of such transceivers is a non-trivial challenge from the perspective of transceiver design and not easily scalable beyond a few number (e.g. 2-3) channels. Hence, multiple WIs share a single channel. The Code Division Multiple Access (CDMA) based MAC mechanism enables multiple concurrent wireless links at the penalty of reduced bandwidth for each link due to the orthogonal code words. Moreover, such MAC requires high power transceivers to maintain the orthogonality among the code channels [9]. On the other hand, simple and fair token-based TDMA MAC mechanism does not require sophisticated transceiver design and has low implementation overhead. Furthermore, no centralized control is required for the arbitration among the WIs. Due to these benefits, a token based MAC mechanism is adopted for many WiNoCs [5],[11]. In the token passing MAC mechanism, the WIs located at various parts on the chip are organized in a virtual token ring. The permission to transmit via the wireless channel is granted by the possession of a token

at a WIs. Each MAC unit is equipped with a token register that denotes the presence of token at the WIs [39]. Interference in transmission is avoided, as one WI possess the token at any point in time. The WI possessing the token transmits a predefined number of packets using the wireless medium before passing the token to the next WI in the ring. The transfer of token to the next WI is realized in the form of a token flit that contains information about the WI releasing the token, the WI that will be possessing the token and other control information [39]. The token circulates among the WIs in a round robin fashion to ensure fairness in channel access.

Although the token based MAC mechanism enables simple, fair and distributed medium access, in such mechanism the period for which each WIs possesses the token is independent of the volume of traffic passing through the WIs [5]. However, in a NoC, the traffic demand of the components varies both temporally and spatially due to the self-similar and bursty nature of the on-chip traffic [40] [41]. Such spatial and temporal variation is also observed among the WIs in a WiNoC [44]. To address such variation issue and enable efficient wireless channel access, dynamic MAC mechanisms are proposed in [44][42]. In these proposed MAC mechanisms, the token possession period is adjusted dynamically depending on the volume of traffic passing through the WIs [37][11]. However, in order to adjust the token possession period, accurate prediction mechanism is required to ensure maximum channel utilization. Existing dynamic MAC mechanisms uses simple history based prediction to predict the bandwidth demand of the WIs. Although such history based prediction mechanisms are simple, they limit the performance gain of the WiNoC architectures due to the lack in accuracy. Hence, in order to improve the efficiency in wireless channel access in a mm-wave WiNoC, accurate prediction mechanism is required. In this work, we investigate the possibilities with artificial neural network (ANN) based prediction mechanism for dynamic MAC mechanisms in WiNoCs.

1.4 Thesis Contribution

The ANN is a machine learning prediction algorithm that is inspired by animal nervous system. In recent literature, ANN is used to accurately predict the network traffic pattern in a NoC environment [19]. Hence, in this work we adopted such prediction mechanism to accurately predict the bandwidth demand of the WIs in WiNoCs. By integrating such ANN based prediction mechanism with a dynamic MAC mechanism, we evaluate the performance and energy efficiency of a WiNoC architecture. Our study shows that the proposed ANN based prediction can significantly improve the performance of a WiNoC over a wireline NoC due the increased accuracy in bandwidth prediction.

Specific contributions of this thesis are:

1. Design and training of the ANN to be able to predict the traffic demands of WIs in a WiNoC.

2. Integrate this prediction mechanism using ANNs with the token-time allocation mechanism

3. Perform system-level performance evaluation for various traffic patterns on a WiNoC architecture.

Chapter 2: Related Work

The design of the MAC mechanism enabling collision free and efficient utilization of the wireless channel in an on-chip environment is constrained by area, power and buffer overheads. Due to these restrictions, complex MAC mechanisms used in conventional networks are not suitable for WiNoCs [32]. Hence, design of efficient, low overhead, and fair MAC mechanisms are considered as one of the critical challenges for WiNoCs [45]. A synchronous and distributed MAC mechanism (SD-MAC) is proposed in [23] for the Ultra-Wide-Band (UWB) WiNoC architectures. Such WiNoC uses impulse based transceivers where the communication range is a few millimeters. Furthermore, in order to access the wireless channel, the WIs share the request control packets over wired links. Thus, such MAC mechanism cannot be adopted for WiNoCs where the WIs is more than a millimeter apart. A hybrid MAC mechanism combining both TDMA and FDMA based access is proposed for CNT based WiNoC architectures in [7]. Although, the CNT based antennas enables communication among the WIs distributed over the chip, their implementation in current CMOS process are challenging. Authors in [46], [47] has proposed an mm-wave WiNoC architecture with multiple non-overlapping channels to enable FDMA based medium access. However, such FDMA based approach is non-trivial from the perspective of transceiver design and the numbers of concurrent channels are not easily scalable. WiNoC architecture with CDMA based MAC mechanism proposed in [8] also enables concurrent communication among the WIs and efficiently utilizes the wireless bandwidth. However, CDMA requires coherent Binary Phase Shift Keying (BPSK) receiver along with Analog-to-Digital Converters (ADC) in the transceivers making the design significantly challenging. Similar to the CDMA, a distributed MAC with strictly synchronized orthogonal request packets is proposed in [32]. In [48], authors proposed a TDMA based CSMA MAC mechanism for WiNoC architectures. However, the CSMA MAC mechanism suffers from performance degradation at high traffic loads

due to back-off based collision recovery as demonstrated in [2]. Therefore, a simple, distributed and low-overhead token passing based MAC mechanism is adopted for many WiNoC architectures [10], [11]. In the TDMA-MAC mechanism, the access to wireless medium is granted to a WI by the possession of a token, circulating among the WIs, organized in a virtual ring. However, the token passing MAC mechanism is unaware of the varying bandwidth demand of the WIs, resulting in inefficient utilization of the energy efficient wireless channel. A dynamic radio access control mechanism (RACM) proposed in [49] shows that dynamic allocation of slots to the WIs improves the performance of a WiNoC. In the RACM MAC mechanism the unused slot by the WIs in an epoch is redistributed among the WIs in the next epoch based on their current slot usage. A similar MAC mechanism reported in [27] allocates slot to the wireless transceivers based on the link utilization. Although these dynamic MAC mechanisms are novel and provide promising solution for traffic additivity in WiNoC architectures, they are reactive with slow system level response as they depend only on the utilization of the WIs based on current traffic behavior.

In an orthogonal direction, artificial neural networks (ANNs), due to their ability to learn and adapt in various domains, has also been used for prediction with high accuracy [21]. In recent times, in order to avoid inter-core traffic congestion in a multicore chip, ANN based predictor was designed [19]. ANN with Back propagation (BP) learning algorithm is widely used in solving various classification and forecasting problems. They have the advantage to be easily implemented in parallel architectures (i.e. in multicore processors or systems with GPUs). ANN has the ability to learn about small network spatiotemporal variations which results in online congestion and can predict the next hotspot occurrence in advance. Also it can handle large amount of data sets and can implicitly detect complex non-linear relationships between dependent and independent variable. These characteristics of ANN motivated to adapt in this work to predict the bandwidth demand of the wireless nodes and

based on that, time slots can be reserved for each node for the transmission of data. In this work, we propose the design of an ANN based prediction mechanism for dynamic MAC mechanism [44] to proactively allocate token possession period to the WIs based on their predicted traffic demands.

Chapter 3: Token Based Dynamic MAC Mechanism for WiNoCs

Wireless NoCs can improve the energy-efficiency and performance of data communication in large multicore chips. However, the MAC mechanism enabling the on-chip wireless communication should ensure the interference free communication in the wireless medium. Complex MAC mechanism has high implementation overhead and hence it is not suitable for on chip environment [32]. For this reason, a low-overhead token passing MAC mechanism is adopted for WiNoCs in [10], [11] which can access the energy efficient wireless medium in a distributed fashion. However, such simple MAC mechanism is agnostic of the spatial and temporal traffic variation of the WIs. This limits the performance gain in a WiNoC due to inefficient utilization of the available wireless bandwidth. Dynamic MAC mechanism where the token possession period is adjusted dynamically based on the bandwidth demand of the WIs, has been proposed as a solution to this performance issue [42] and slip]. The dynamic MAC predicts the bandwidth demand of the WIs for the next token period and allocates the token possession period to each WI based on this predicted traffic demand. In this chapter, we discuss such a token based dynamic MAC mechanism [44].

3.1 History Based Prediction

In order to predict the traffic demand of the WIs, a simple history based prediction mechanism is used [44]. In the history based prediction, the bandwidth demand of a WI in a token period is predicted based on the moving average of bandwidth demand of that WI over previous token periods. Hence, the predicted bandwidth demand $\hat{B}^{TP^{j+1}}$ for token period j+1 is calculated by,

$$\hat{B}^{TP^{j+1}} = \frac{BD^{TP^{j}} + \overline{BD^{TP}}}{2} \quad (1)$$

Where, $BD^{TP^{j}}$ is the actual bandwidth demand of a WI over the token period j and $\overline{BD^{TP}}$ is the average bandwidth demand for that WI from token period 0 to j-1. The steady state demands of the WIs are captured by the moving average of the past token periods and the most recent variation in the demand is captured by the demand of the last period. Hence this prediction mechanism captures both long term and instantaneous bandwidth demands of a WI [28]. This predicted demand value is then used to adjust the token possession period for the next token period. However, the bandwidth demand prediction mechanism using the history-based prediction is unable to capture the trend when there are sudden variations in the traffic which is common in on-chip environments. ANNs on the other hand, can be trained to predict and track variations in traffic with higher accuracy. Hence in this work, ANN based prediction mechanism as mentioned in the next chapter has been explored to predict the bandwidth demand of the WIs. The integration of the ANN based prediction with the dynamic MAC mechanism to generate the dynamic token duration allocation is also discussed in the next chapter.

3.2 Proportionate Slot Allocation Mechanism (P-SAM)

In a baseline token based TDMA MAC for WiNoC each WI possesses the token for a fixed maximum duration before releasing it to the next WI for medium access. This neglects the unequal traffic demands at the WIs and their temporal variations by permitting WIs to transmit depending only on their current traffic demand without knowledge of the demand of others which maybe more than its own. In the P-SAM, the token possession period of a WI is dynamically adapted based on the proportion of the predicted traffic demand of the WI relative to other WIs. To make this proportional

slot allocation, a fixed token period similar to the token-based MAC is assumed which is distributed among the WIs depending on their relative predicted traffic demands. The allocated time slots of a WI, i at token period, j+1, $s_i^{TP^{j+1}}$ is given by,

$$s_i^{TP^{j+1}} = \frac{\widehat{B_i}^{TP^j}}{\sum_{i=1}^{N} \hat{B}_i^{TP^j}} \; X \; S_{TP} \tag{2}$$

Where, $\widehat{B_i}^{TP^j}$ is the predicted bandwidth demand for WI, i at token period j calculated using (1) and N is the number of WIs; S_{TP} *is* the number of time slots for data flits in the token period. With this scheme, WIs with greater predictive bandwidth demands will be assigned more time slots in the token possession period.

In dynamic MAC mechanism proposed in [44], this predicted demand value is shared with other WIs using the token flit. In order to achieve this, a *demand* field is added with the token flit. After all these demand values are shared by the WIs in a token period, the token possession period for the next token period is allocated based on these predicted demand values. However, the efficiency of the dynamic MAC mechanism depends on the accuracy of the prediction mechanism. In order to improve the efficiency in medium access, we propose an ANN based prediction mechanism and integrate such prediction with the dynamic MAC mechanisms as discussed in the next chapter.

Chapter 4: Integrating ANN based prediction with the MAC for WiNoC

Artificial Neural Network (ANN) is a machine learning prediction algorithm that is inspired by animal nervous system. It can be used to predict the network traffic pattern in NoC environment.

4.1 Artificial Neural Networks

Accurate traffic demand prediction mechanism is required for the higher transmission slot allocation efficiency. Hence in order to choose the prediction mechanism, both implementation complexity (i.e. area overhead and computational complexity) and prediction accuracy are considered as important factors. The history based bandwidth prediction mechanism suffers from high implementation complexity because of the nature of its algorithm. On the other hand, an ANN based prediction mechanism is used in control systems for its efficiency, simplicity, and robustness. Artificial Neural Networks (ANN) can be used to perform tasks such as pattern recognition, classification and function approximations in many situations where traditional approaches are not suitable. ANNs have been successfully applied to solve problems over the conventional methods. For example, ANN can be used successfully in synthesis and speech recognition, image processing, pattern recognition, coding, classification, power load forecasting, interpretation and prediction of financial trends of stock-market, composite structure manufacturing, processing, monitoring, modeling and controlling systems; and others [34]. In order to take advantage of the inherent parallelism in ANNs, most of the Hardware implementations have been presented especially in academicals works [34].

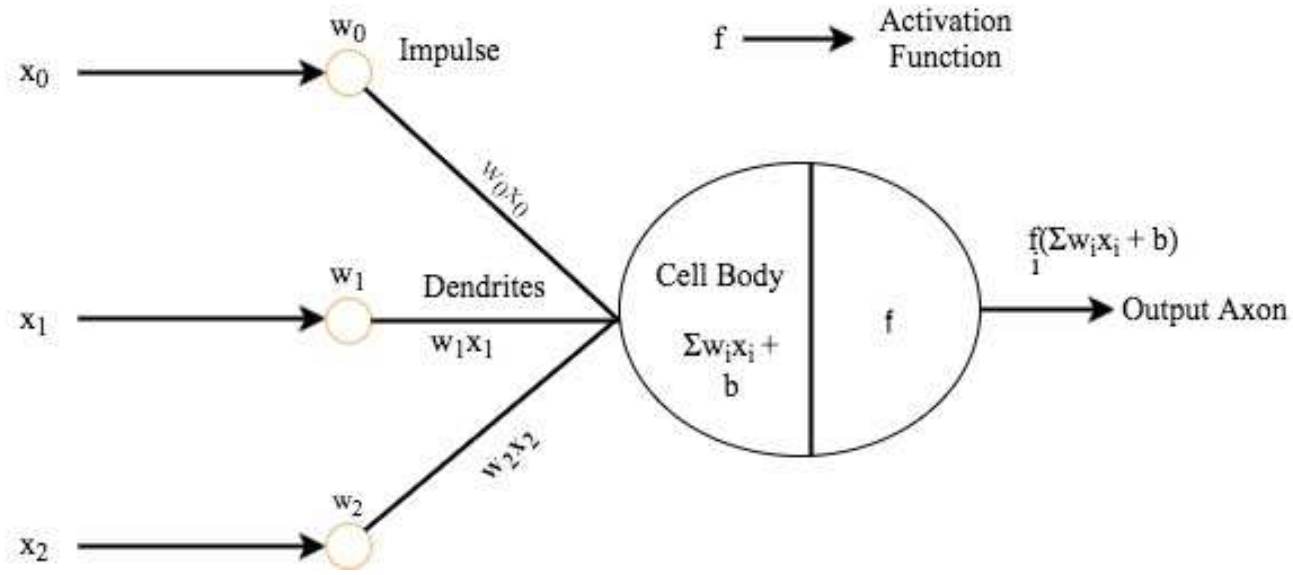

Fig. 2: Artificial Neuron

Artificial neurons are the building block of Artificial Neural Network (ANNs). Artificial neural network is the collection of nodes in a computational graph in an acyclic graph, which is inspired by the neurons in the human brain. Impulses are communicated between the neurons by dendrites. Artificial Neural Network has inputs (dendrites), final outputs (axon) to other neurons and the activation function (cell body). Neurons basically functions by emitting the signal through its axon, when activated. Activating the other neurons further processes the signal. The cell body processes the impulses further. Some of the activation functions are sigmoid, hyperbolic tangent, rectified linear units, leaky rectified linear units, and exponential linear units. Artificial neuron with some activation function f is shown in the Fig. 2,

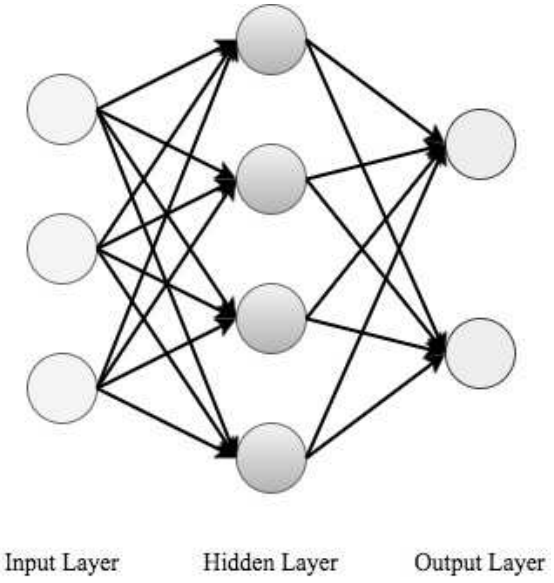

Fig. 3: ANN Example

ANN basic architecture consists of 3 layers: an input layer, hidden layer and output layer as in Fig. 3. Neural network process flow can be explained in three steps i.e. forward propagation, backward propagation and update weights.

4.2 Design of the ANN

The input to the ANN is the total number of incoming flits for each wireless node for the

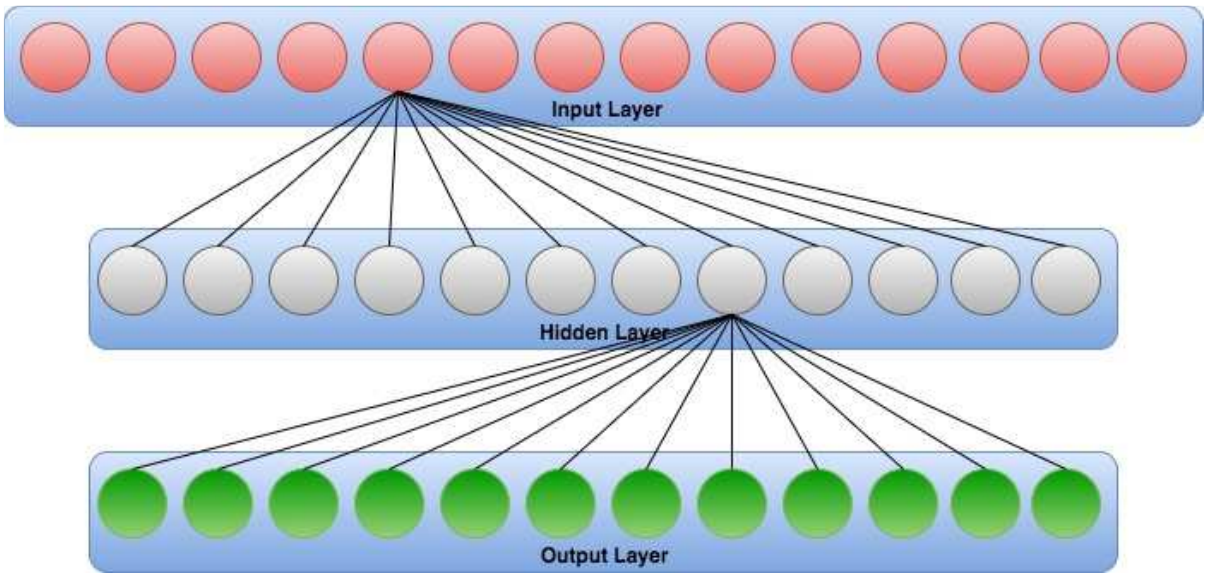

Fig 4: Trained ANN structure

given token period j. The input from the wireless node is passed to each individual nodes of the ANN input layer. The output is the predicted bandwidth demand for the token period j.

As shown in the Fig. 4, the three layers of the ANN are input layer, hidden layer and the output layer. Sigmoid activation function was found to produce best accuracy out of all activation functions for the hidden layer and the linear activation function for the output layer[50]. It was also observed from the simulation that there is the negligible loss in accuracy for the change in the number of neurons. So, the number of neurons should be selected carefully to have increased accuracy and to have lower hardware cost with increase in implementing the number of neurons. The selection of neurons are explained in section 4.4.

4.2.1 Forward Propagation

Forward propagation is one of the major processes done in all type of neural networks. With this, the output scores are obtained from the neural network. Neural networks are acyclic computational graphs consists of nodes as neurons and edges as weights. In forward propagation, the function (5) is performed at each node from inputs towards the outputs. The important note is the current layer input is the output of the previous hidden layer.

$$\alpha^{l}_{j} = f(\sum_{k} w^{l}_{jk} \alpha^{l-1}_{k} + b^{l}_{j}) \quad (5)$$

4.2.2 Backward Propagation

The error or loss generated by the ANN over all the network nodes is distributed using back propagation [30]. It helps to visualize the changes in the output cost when minute changes are done in weights or biases. Finding gradients for each and every weights and bias does it. To propagate the gradients chain rule is used. Gradients are found using below equations.

$$\delta^l_j = \delta C/\delta z^l_j \quad (6)$$

Error on other hidden neurons is given by $\delta^L = \upsilon_a C \Theta f'(z^L)$ where $\upsilon_a C = \delta C/\delta a^l_j$. These equations enable finding gradient of cost in any neuron network.

4.2.3 Parameter Update

In order to reduce the cost, parameters are updated using analytic gradients across all training samples. Some of the most common methods of updating parameters are Stochastic Gradient Decent (SGD), Momentum, Adagrad, Weight Initialization, and Regularization.

4.3 Training the ANN

In order to predict the bandwidth, ANN is trained using train data set. As in Fig. 5,the train data set consists of inputs i.e. utilization of the chip components and their corresponding outputs i.e. bandwidth demand of each wireless node, for the given token period. The relationship between the input and output is modeled while training the ANN using train data set. With the new mechanism, the volume of the traffic was effectively managed.

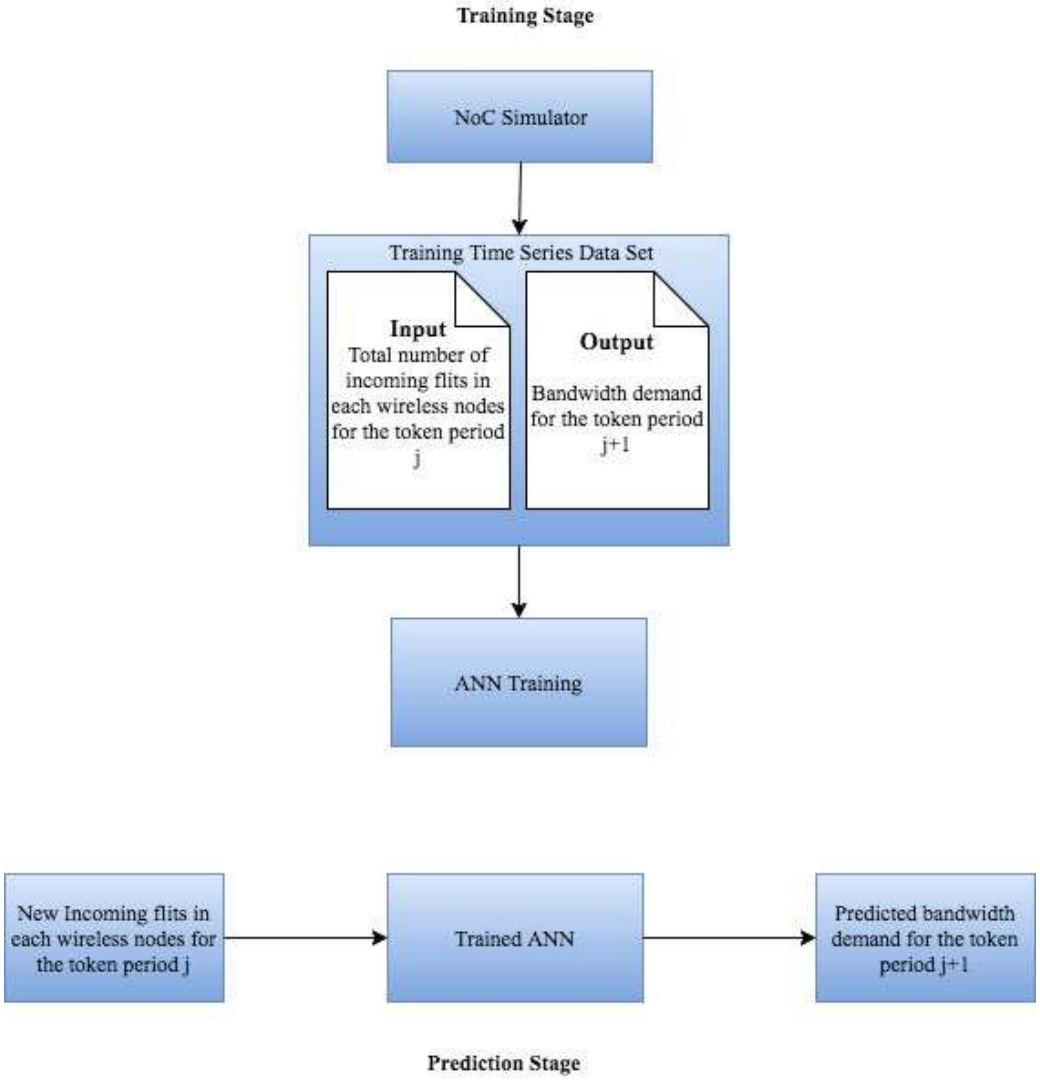

Fig. 5: ANN-NOC Methodology overview

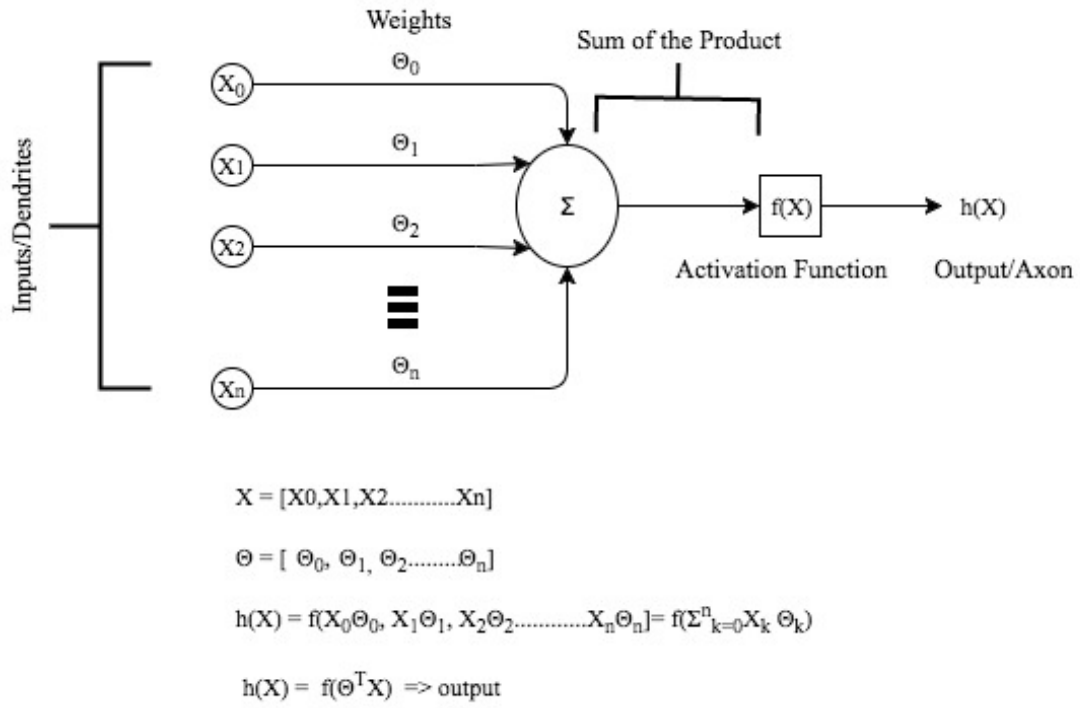

Fig. 6: Mathematical model of a single neuron

The basic neuron model of the ANN consists of two elements as mentioned in Fig. 6: First element adds product of inputs and weights coefficient. The second element is the neuron activation function, which is a nonlinear function. The weights 'Θ' are initialized to small ± values centered on 0. When the n dimensional input, including the bias input is sent through the ANN, it is initially multiplied by its initial weights. When the weighted sum of the n dimension exceed the threshold value, the neuron fires a floating point value, which is then passed through the activation function f(ΘTX) for mapping the value to give an output. In the next cycle, next input is fed and the weights are updated again. The above process is recursively followed, until the low cost function is obtained. Multiple such neurons connected together forms the Neural Network. This trained ANN understands the trend of any new incoming flits and produces the corresponding bandwidth demand as output.

4.4 Creating the training data set

The desire to have an ANN to model the dependency between the incoming flits in the core and the corresponding bandwidth of each wireless node creates the necessity to generate a training dataset. To gather the training data, random incoming flits monitor and the corresponding bandwidth is recorded for the particular token period, in the in-house multi-core NoC simulator. The simulator is allowed to run for 5000 cycles with the initialized conditions. At the end of 5000 cycles simulator outputs the bandwidth demand for each token period. Similarly, several random utilization conditions are initialized to the simulator, and then the corresponding output bandwidth demand is recorded as the training data. The total training data that was gathered is 5000 x 250 samples. This would help for the prediction of bandwidth in that particular token period in the operating range of the system.

The train function train (net,x,t) in nntool box of Matlab is used to train the neural network. While giving input to train (net,x,t) in nntool box, the incoming flits monitor input of each WI nodes is given to individual net, token period value is given to x and the next set of input is indicated in target t. Hence, the number of nodes in the input layer is 14, which get its input from 12 wireless nodes, x and t. The output is the bandwidth prediction for the same 12 WIs. So the total number of nodes in the output layer is 12. The number of hidden layer neurons for the core streams are 12, which are selected by means of trial and error. Initially, the number of neurons in the hidden layer is taken as 4 and the training is done with 250 samples. Where each sample was the real traffic utilization or bandwidth demand for 5000 cycles. The actual bandwidth demand is measured from a simulator, by measuring the data rate in each WI node for the given token period. The root mean square error is the measure of the difference between the predicted bandwidth value and the actual bandwidth value. The square root of the mean of the squares of the deviation between actual bandwidth and the predicted bandwidth gives the rms error. For the trained ANN with 250 samples,

the testing was done by running the simulation for another 5000 cycles, separate from the training samples and the rms error was calculated which is 2.9 cycles. Again the number of neurons in the hidden layer was increased to 12 and the training was done again with 250 samples. Now the testing was done again and the rms error is found to be 2.1 cycles. The number of hidden layer and the samples can be increased further, however it will increase the hardware cost due to the usage of lot of memory and connections.

4.5 Integrating the ANN prediction with dynamic MAC

In order to integrate the ANN based prediction mechanism with a dynamic MAC in a WiNoC, we equip one of the WIs with the ANN based prediction capabilities. The architecture of the WI with the ANN based prediction capabilities is shown in Fig. 7

The WI contains four registers, IDself, IDnext, HasToken and Demand*self*. The WIs shares their bandwidth demand value using the *demand* field in the token flit. These demand values are stored in the REGdemand of the WI with ANN based prediction capabilities. The registers IDself, HasToken and IDnext, and the counters Token Period Counter and Token Possession Period Counter are used for maintaining the token passing based medium access. The Demand Counter is used to determine the incoming bandwidth demand of a WI in a token period. At the expiration of the token period counter, the value of the Demand Counter is stored in the Demand*self* register. Hence, the Demand*self* stores the most recent value to the traffic demand for the WI. To predict the bandwidth demand of the WIs, the Demand*self* and the contents of the REGdemand is used by the ANN based prediction unit. Once, the ANN based prediction generates the predicted bandwidth demands, it is

broadcast to all the WIs as a control flit. Hence, all the WIs, uses the predicted demand information from the control flit to calculate the duration of the token possession period in the next token period.

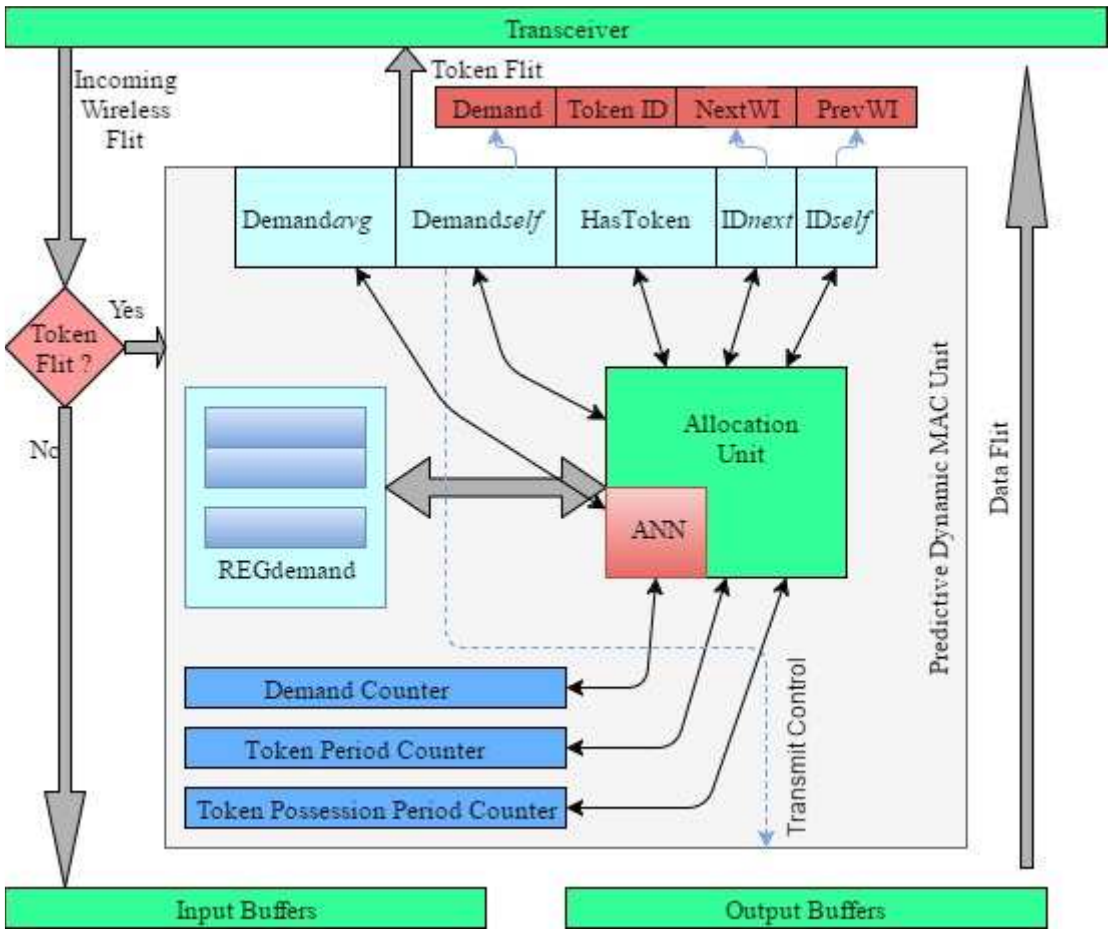

Fig. 7: Architecture of Predictive Dynamic MAC unit

4.6 ANN Hardware

The neurons are implemented as multiplier and accumulator (MAC) units in parallel for the efficient realization of ANN with the use of available resources as shown in Fig. 8. As discussed in section 4.4, the total numbers of input layer neurons are chosen as 14, the numbers of hidden layer neurons are chosen as 12 and the numbers of output layer neurons are chosen as 12. As soon as the information packets reach the ANN core in the pipelined fashion, the computation starts to predict the

bandwidth demand. The bandwidth is represented as 8 bit values. Wireless interconnection is proposed for the data transfer to and from the ANN to the different components of the chip. The incoming flits values are sent to the ANN through the nearest NoC switch. The wireless switch, which possesses the token, can only transfer the data to other wireless nodes.

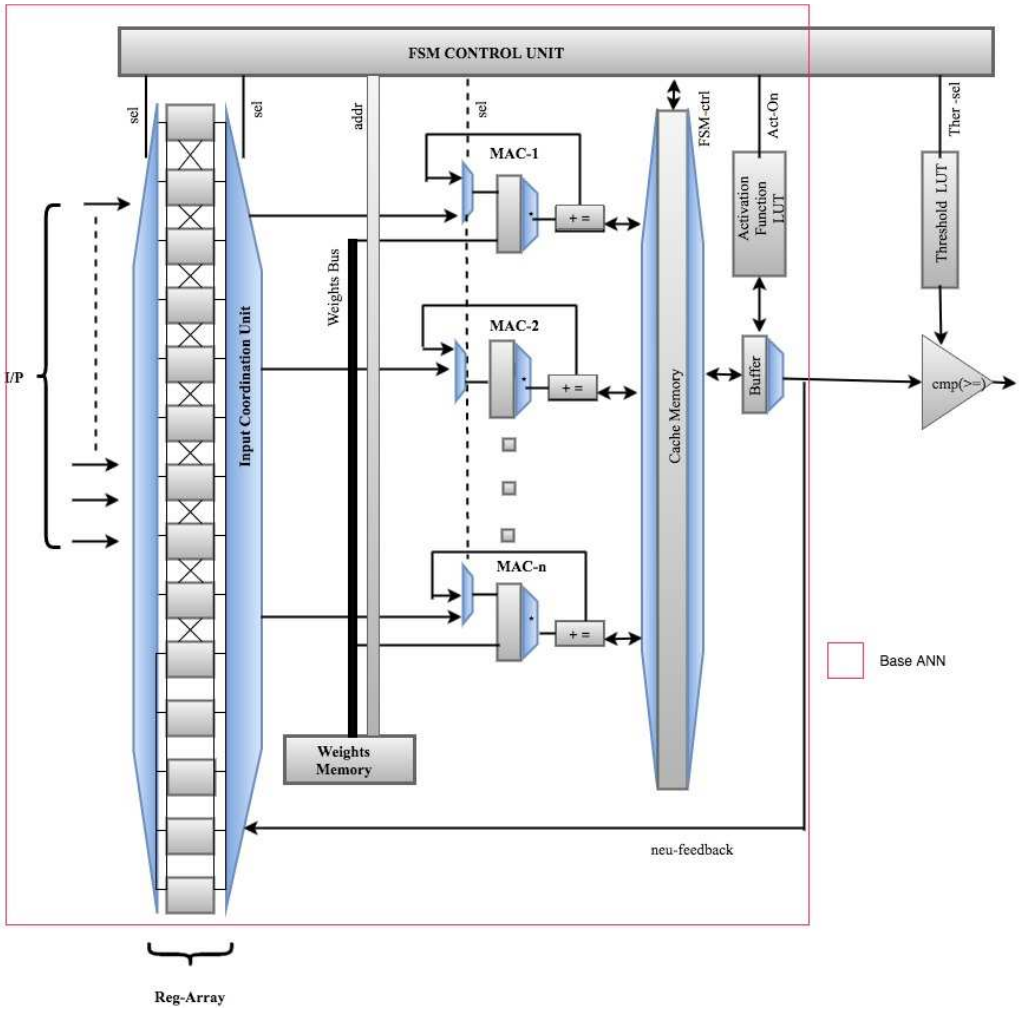

Fig. 8: ANN Hardware

The neurons are implemented as MAC units, which multiply and accumulate the input values and the weights. The register arrays are set to store the inputs, if the number of MAC unit available is less than the number of inputs. Weight values from the weight memory are multiplied with the input

and accumulated in the MAC unit. The outputs from the MAC are stored in the cache. When all the computation to a particular hidden layer is complete, the output is passed through the buffer. At the buffer, corresponding sigmoid activation function value is obtained from the look-up table. Once the computation is done between the input layer and the hidden layer, the LUT output is feedback to the input co-ordination unit. The same calculation between the input layer and the hidden layer is repeated for the hidden layer to output layer. During the first iteration, the predicted bandwidth is added to the initial bandwidth of the chip and then accumulated. For the second iteration, predicted bandwidth is added to the accumulated result. The result is compared with the threshold at the comparator. If the result is less than one, the threshold output is zero; otherwise the output is same as the result. The output from the comparator is the predicted bandwidth demand of each WI. This predicted bandwidth demand is used to dynamically adjust the token possession period of each WI in the dynamic MAC mechanisms.

Chapter 5: Experimental Results

In this chapter, we evaluate the performance and the energy efficiency of a WiNoC architecture with the ANN prediction based dynamic MAC mechanism. Although, many different WiNoC architectures was proposed in literature, Mesh based WiNoC architecture is considered in this work with both wired and wireless links as a test case. The performance of the WiNoC architecture is measured in terms of bandwidth and packet latency. The bandwidth of a WiNoC is measured as the number of bits successfully routed from the source to the destination core per second. The packet latency is the number of clock cycles required to transmit a packet successfully from the source core to the destination core. The energy efficiency is measured in terms of packet energy, which is the average energy required to route a whole packet from the source to the destination through the NoC components.

5.1 Mesh based WiNoC Architecture for the Test-bed

In this work, we adopt a mesh based wireless network on chip architecture for the evaluating the performance of the proposed ANN prediction based dynamic MAC mechanism. The performance and energy efficiency of the proposed ANN prediction based dynamic MAC mechanism is evaluated in a Wireless Mesh (WiMesh) architecture as a test case. The WiMesh is a hybrid topology with both wireline and wireless interconnects. The wireline topology is a regular mesh where each switch is connected to a core and the switches are interconnected with other switches in its cardinal direction. To deploy the WIs in the planar Mesh, we further divide the whole NoC into smaller logically subnets. Then, a WI is placed in one of the center switches in each subnet to form the WiMesh topology. As shown in [39], the performance of the WiMesh is optimized when the total number of WI is 12. However, due to subnet-based configuration in our test-bed WiMesh architecture, the total

number of WIs in the WiMesh is 12. This topology is shown in Fig. 10.

In order to realize the wireless interconnects, each WI contains on-chip antenna, and transceiver circuit. The adopted antenna should provide the best power gain with the smallest area overhead. Also, the antenna should have the capability of radiating in all directions with almost similar gain, as the WIs are located at different parts on the chip. A CMOS compatible metal mm-wave zigzag antenna operating at 60 GHz mm-wave bands and the bandwidth of 16 GHz has been demonstrated to possess such characteristics [8]. Hence, in this work, we adopt the design of such on-chip miniature zig-zag antenna to enable long range shortcuts among the WIs. On the other hand, the adopted transceiver circuit should provide a very wide bandwidth and consume low power, to ensure high performance and energy-efficiency. The transceiver design is adopted from [10] where low power design consideration is taken into account. The wireless transceiver is shown to dissipate 36mW or 2.3pJ/bit supporting the data rate of 16 Gbps with a bit error rate of less than 10^{-15} and with the area occupation of 0.3 mm^2 in the post-layout design using TSMC 65nm CMOS process.

The WiMesh contains both short and long links. The shortest path based routing method is used to optimize such network performance and it was adopted in this work. The shortest path between any two pairs of nodes in the network is determined using the minimum-spanning tree formed by Dijkstra's algorithm. The minimum spanning depends on the chosen start node but the length of paths between any particular pair is independent of the start node. Hence the minimum spanning tree is selected in random. Furthermore, the deadlock is avoided by transferring flits along the shortest path routing tree extracted by Dijkstra's algorithm, Hence, each switch results in a scalable routing mechanism, as it has only local forwarding information eliminating the need for non-scalable global routing information. We adopt wormhole switching for both the wireline and the wireless interconnects in the NoC architectures.

5.2 Simulation Environment

The NoC architectures (i.e. Mesh and WiMesh) are characterized using a cycle accurate NoC simulator that accurately models the progression of the flits over the switches and links per cycle as shown in Fig. 9. The simulator considers both flits that reach the destination as well as those that are stalled. The post-synthesis delay and the energy dissipation of the NoC components considering both dynamic and static power consumption are annotated into the simulator for evaluating the performance and energy efficiency of the NoC architectures. For our experiments, we consider a system size of 64 cores. For traffic pattern we have used uniform random synthetic traffic pattern. In the uniform random synthetic traffic pattern, a packet is addressed to any other core with equal probability. Hence, the uniform random traffic pattern captures both the long and short distance communication. In the system level simulation, ten thousand iterations were performed eliminating transients in the first thousand iterations. The NoC switch is adopted from a three-stage pipelined design [43]. Each switch is considered to have 4 VCs with a buffer depth of 2. However, as the WIs handle a large volume of traffic, an increased buffer depth of 16 is used with 4 VCs. A moderate packet size of 64 flits is considered for all our experiments. The width of all wired links is considered to be same as the flit size, which is considered to be 32 bits. Token possession period is considered to be 96 time slots for the baseline token passing mechanism.

For estimating the packet energy, the energy consumption in both links and switch are considered. The energy dissipation and delay of the wired link is obtained through Cadence simulations taking into account the specific lengths of each link based on the established topology in the 20mmx20mm die. For the wireless interconnect, we adopted the antenna from [8] and the transceiver design from [10] as mentioned in the previous section. The NoC switches and proposed ANN prediction based dynamic MAC are synthesized from a RTL level design using 65nm standard

cell libraries from TSMC using Synopsys. A 2.5 GHz clock and 1V V_{dd} frequency and voltage for the 65nm technology node is used for synthesis.

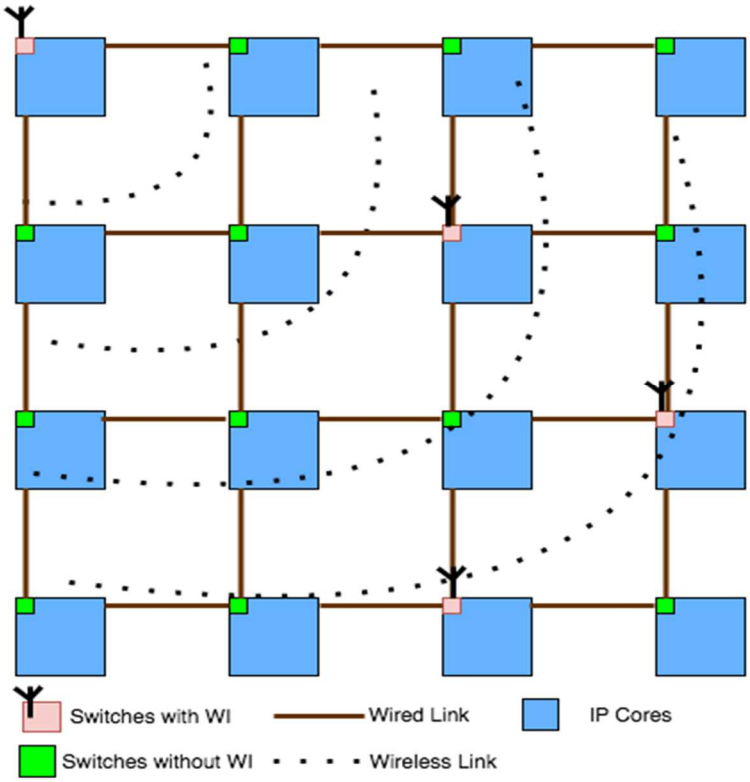

Fig. 9: WiMesh Topology

5.3 Comparative Performance Study

In this section we evaluate the performance of the ANN based prediction mechanism and compare it with the History based prediction. We also present the accuracy in prediction achieved by both these prediction mechanisms. Then, we present a comparative performance study between the proposed ANN based prediction enabled dynamic MAC unit with state-of-the-art dynamic MAC mechanism with history based prediction in a WiMesh architecture.

5.3.1 Overhead Analysis

In this section, we present the overhead analysis for the ANN prediction mechanism and compare it with the overhead of the history based prediction mechanism. The overhead is evaluated in term of power, delay and area as shown in Table 1. The experiment presented here consists of training an ANN model in which mathematical function approximation was done. The ANN was implemented using sigmoid activation function. The basic ANN architecture is equivalent to 14: 12: 12 configuration (i.e.) 14 neurons in input layer, 12 neurons in the hidden layer and 12 neurons in the output layer. This topology was developed, by doing the trials in which this topology offers the best relation between the network complexity and the output accuracy. The ANN is then compared with the History based predictor and the results are presented in table 1.

	History Based Prediction	**Artificial Neural Network**
Power (mW)	0.373	0.25
Area (um^2)	2629.02	912860
Delay (ns)	0.14	1122.6

Table 1: Artificial Neural Network comparison with History Based Predictor

The History based and ANN based predictors are both implemented in RTL and post-synthesis models using 65nm technology node is used for the comparison. The area of the ANN is very high because the activation function is implemented in the form of Look up table (LUT) [50]. The wireless transceiver is shown to dissipate 36mW and in this work total of 12 transceivers are used. The power consumption of ANN at each wireless node is 0.25 mW. Hence the total power

consumption of the ANN is only 0.06% of the power of wireless communication infrastructure. Though the delay is more for ANN, it works in parallel to the data path and hence it won't impact the overall delay of the dataflow. However, the prediction of the bandwidth demands improves the dataflow performance as can be seen in the next sections.

It can be seen from table 1 that the area overhead for the ANN based prediction is higher compared to the history-based prediction due to the LUTs used in the ANN. The trained ANN was simulated for 5000 cycles and as discussed in section 4.4, the rms error was calculated between the predicted bandwidth and the actual bandwidth, which is 2.1 cycles. The rms error for the history based prediction mechanism is 2.5 cycles. The ANN based prediction improves the prediction accuracy by 16% compared to the history based prediction. This improvement in accuracy results in increasing the efficiency in utilizing the wireless medium. This further results in an improvement in system performance and energy efficiency for the dynamic MACs as discussed in the next section.

5.3.2 Performance Analysis

In this section, we evaluate the performance of different MAC mechanisms for uniform random synthetic traffic pattern. The peak bandwidth at network saturation for different dynamic MAC mechanisms as well as for the wired Mesh is shown in Fig. 10. The wireline Mesh architecture has the lowest bandwidth compared to that of all the WiMesh architectures. This is because, in the Mesh architecture, inter-core communication requires multi-hop communication over wireline paths resulting in lower bandwidth. On the other hand, in the WiMesh architectures with dynamic MAC mechanisms, the wireless links help in reducing the average hop count of the network. Due to the long range wireless links and demand aware allocation of the time slots to the WIs, in the WiMesh architecture with history based prediction mechanism, the bandwidth improves by 7.4% compared to

the wireline Mesh architecture. However, the performance benefit due to the wireless interconnects can be further improved by increasing the accuracy of the prediction mechanism. However, compared to the history based prediction, the ANN based prediction improves the prediction accuracy by 16%. Due to the increase accuracy in ANN based prediction mechanism, the performance improvement for the WiMesh architecture is 12.05% compared to the wireline Mesh at the saturating point. Hence, the proposed ANN prediction based dynamic MAC has higher bandwidth compared to the state-of-the-art history based dynamic MAC.

The packet energy of the Mesh and WiMesh architecture with different MAC mechanisms for uniform random traffic pattern at network saturation is shown in Fig. 11. Due to the multi-hop inter-core communication, the packet energy for the wired Mesh architecture is higher than the WiMesh architecture with baseline token-based MAC mechanism that uses single hop wireline links. However, due to the dynamic allocation of the token possession period and use of the single hop wireless links, the packet energy is reduced for the WiMesh architecture with dynamic MAC mechanisms. The packet energy is lowest for the WiMesh architecture with the ANN prediction based dynamic MAC mechanism. This is because, the ANN based prediction can accurately predict the traffic demand of the WIs and allows efficient utilization of the wireless channel.

The impact of implementing demand-aware predictive dynamic MAC is more evident in Fig. 12 where the latency characteristics is shown for the architectures considered in this paper. From the figure, it can be seen that the WiMesh with ANN prediction based dynamic MAC mechanism yields lowest average packet latency among all the architectures. As the flit injection rates increase, the spatial variation in bandwidth demands among WIs also escalates. In such cases, the ANN based dynamic MAC results in 22.9% and 10.9% lower average packet latency compared to the wireline and dynamic WiMesh architecture with History based prediction at the saturation point.

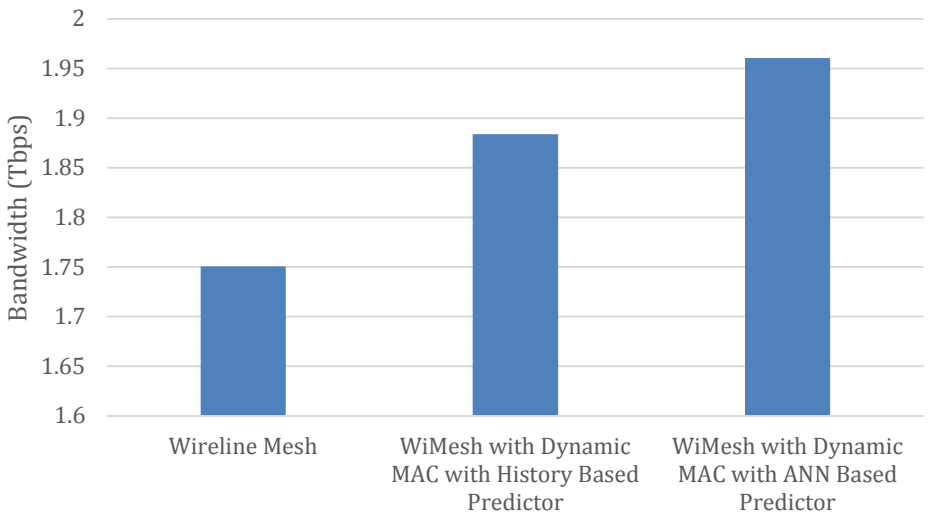

Fig. 10: Bandwidth for WiMesh architecture with Dynamic MAC mechanism

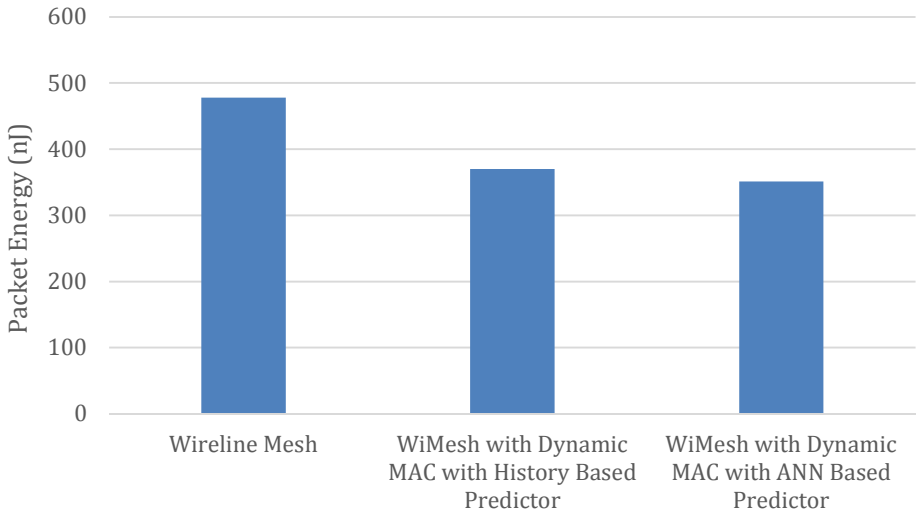

Fig. 11: Packet Energy for WiMesh architecture with Dynamic MAC

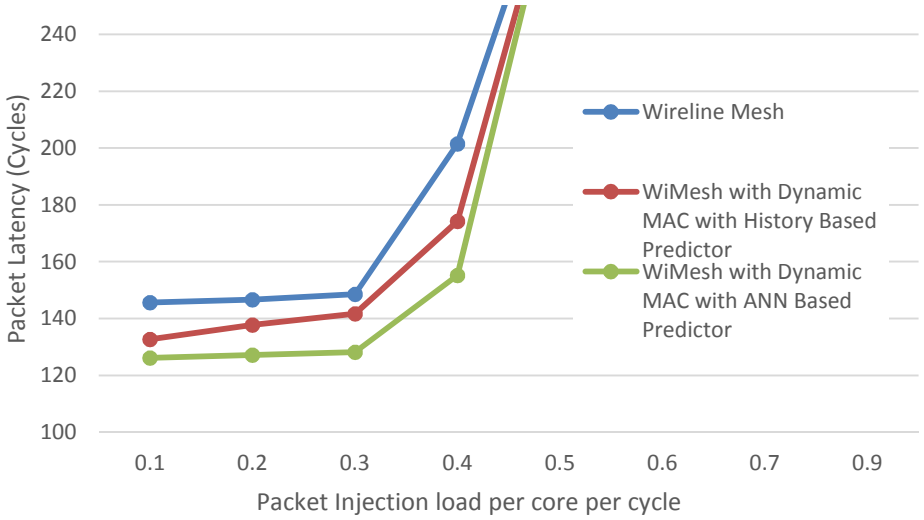

Fig. 12: Packet Latency for WiMesh architecture with Dynamic MAC

6 Conclusions and Future Work

Wireless interconnection is envisioned as an energy efficient communication backbone for future multicore systems. One of the key aspects for the adoption of such novel interconnect paradigm is the MAC mechanism that ensures the efficient utilization of the wireless channel based on the varying demand of the applications. In this work, we propose the design of ANN based prediction mechanism that is able to accurately predict the bandwidth demand of the WIs. The ANN based prediction mechanism improves the prediction accuracy by 16% compared to the state-of-the-art history based prediction mechanism. Furthermore, by equipping our proposed prediction mechanism with a dynamic MAC mechanism to dynamically adjust the token possession period of the WIs, we show that the system level performance of the WiNoC architectures can be improved significantly. The overhead of the ANN based prediction is also low. The power overhead for the ANN based prediction mechanism is only 0.06% of the wireless infrastructure in a WiMesh architecture with 12 WIs.

For the future computational demand, 100x more cores than the current state of art multicore design was expected. In order to predict the bandwidth demand of such complex system, the complex brain inspired algorithms such as the convolutional neural network can be used. The proposed ANN structure shows rms error of 2.1 to 2.5 with respect to the history based predictor model. Accuracy of the ANN can be further improved by increasing the number of hidden layer neurons. Due to the ANN system learning ability, future generations of self-learning and self-adjusting computer chips design can be possible. With the use of state-of-the-art low powered circuit design like neuromorphic engineering techniques, the efficiency of the ANN hardware can be improved. However, this requires analog structure. Further, Proportional-Integral-Derivative (PID) technique can be used to predict the

traffic demand of the WI, which has an advantage of simplicity, robustness and efficiency over Artificial Neural Network (ANN).

REFERENCES

[1] Chung, E. S., Milder, P. A., Hoe, J. C., and Mai, K., *"Single-Chip Heterogeneous Computing: Does the Future Include Custom Logic, FPGAs, and GPGPUs?,"* Microarchitecture (MICRO), 2010 43rd Annual IEEE/ACM International Symposium on*, Atlanta, GA, 2010, pp. 225-236.*

[2] Mansoor, N., and Ganguly, A., *"Reconfigurable Wireless Network-on-Chip with a Dynamic Medium Access Mechanism"*. In Proc of the *International Symp. on Networks-on-Chip*. NOCS '15. ACM, NY, USA.

[3] Kumar, A., Peh, L.S., Kundu, P., and Jha, N., "*Express virtual channels: towards the ideal interconnection fabric*". In Proc. of the International Symp. on Computer architecture. ISCA '07. ACM, 150-161, 2007.

[4] Ogras, U. Y. and Marculescu, R., ""*It's a small world after all": NoC performance optimization via long-range link insertion*," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 14, no. 7, pp. 693-706, July 2006.

[5] Shacham, A., Bergman, K. and Carloni, L. P., "*Photonic Networks-on-Chip for Future Generations of Chip Multiprocessors*," in IEEE Transactions on Computers, vol. 57, no. 9, pp. 1246-1260, Sept. 2008.

[6] Chang, M. *et al*., "CMP network-on-chip overlaid with multi-band RF-interconnect," *High Performance Computer Architecture, 2008. HPCA 2008. IEEE 14th International Symp. on*, Salt Lake City, UT, 2008.

[7] Ganguly, A. Chang, K., Deb, S., Pande, P. P., Belzer, B., and Teuscher, C., "*Scalable Hybrid Wireless Network-on-Chip Architectures for Multicore Systems*," in *IEEE Trans. on Comp.*, vol. 60, no. 10, pp. 1485-1502, 2011.

[8] Deb, S. et al., *"Design of an Energy-Efficient CMOS-Compatible NoC Architecture with Millimeter-Wave Wireless Interconnects," in* IEEE Transactions on Computers*, vol. 62, no. 12, pp. 2382-2396, Dec. 2013.*

[9] Vijayakumaran, V. et. al. "*CDMA Enabled Wireless Network-on-Chip*". ACM Journal on Emerging Tech and Comp Sys., 10(4). Article 28, 2014.

[10] Chang, K., et. al. "*Performance evaluation and design trade-offs for wireless network-on-chip architectures"*. ACM Journal of Emerg. Tech. in Comp. System, 8 (3), Article 23, 2012.

[11] DiTomaso, D., Kodi, A., Kaya, S., and Matolak, D., *"iWISE: Inter-router Wireless Scalable Express Channels for Network-on-Chips (NoCs) Architecture,"* High Performance Interconnects (HOTI), 2011 IEEE 19th Annual Symposium on*, Santa Clara, CA, 2011, pp. 11-18.*

[12] Duato, S.; Yalamanchili, S.; and NI, L.; "Interconnection Networks-An Engineering Approach", *Morgan Kaufmann,* 2002.

[13] L. Shang, L. Peh, A. Kumar, and N.K. Jha, "Temperature-Aware On-Chip Networks," *IEEE Micro*, 2006

[14] M. S. Shamim, A. Ganguly, C. Munuswamy, J. Venkatarman, J. Hernandez and S. Kandlikar, "Co-design of 3D wireless network-on-chip architectures with microchannel-based cooling," *Green Computing Conference and Sustainable Computing Conference (IGSC)*, 2015

[15] Guangshuo, L.; Jinpyo, P.; Marculescu, D., "Procrustes1: Power Constrained Performance Improvement Using Extended Maximize-Then-Swap Algorithm," *in Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on* , vol.34, no.10, pp.1664-1676, Oct. 2015

[16] M F. Chang, el. al., "CMP network-on-chip overlaid with multi-band RF- interconnect," *High Performance Computer Architecture, IEEE 14th International Symposium* on , vol., no., pp.191,202, 16-20 Feb. 2008

[17] Chang, Kevin, Sujay Deb, Amlan Ganguly, Xinmin Yu, Suman Prasad Sah, Partha Pratim Pande, Benjamin Belzer, and Deukhyoun Heo. "Performance evaluation and design trade-offs for wireless network-on-chip architectures." ACM Journal on Emerging Technologies in Computing Systems (JETC) 8, no. 3 (2012): 23.

[18] *Chip MultiProjects*. Retrived November, 2015, from: http://cmp.imag.fr

[19] Kakoulli, E.; Soteriou, V.; Theocharides, T., "Intelligent Hotspot Prediction for Network-on-Chip-Based Multicore Systems," *in Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on* , vol.31, no.3, pp.418-431, March 2012.

[20] Ferro and Pande 2009; Pavlidis and Friedman 2007; Park et al. 2008; Shacham et al. 2008; Joshi et al. 2009; Kurian et al. 2010.

[21] Haykin, S. and Network, N., 2004. "A comprehensive foundation. Neural Networks, 2(2004)".

[22] Lin, J. et al., "Communication Using Antennas Fabricated in Silicon Integrated Circuits," *IEEE Journal of Solid-State Circuits,* vol. 42, no. 8, August 2007, pp. 1678- 1687.

[23] Zhao, D; and Wang, Y; "SD-MAC: Design and Synthesis of A Hardware-Efficient Collision-Free QoS-Aware MAC Protocol for Wireless Network-on-Chip," *IEEE Transactions on Computers,* vol. 57, no. 9, September 2008, pp. 1230-1245.

[24] Deb, S.; Ganguly, A.; Pande, P.; Belzer, B.; and Heo, D.; "Wireless NoC as Interconnection Backbone for Multicore Chips: Promises and Challenges," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems,* vol. 2, no. 2, pp. 228–239, 2012.

[25] Murray, J; Klingner, J.; Pande, P.; and Shirazi, B.; "Sustainable Multi-Core Architecture with on-chip Wireless Links"*, Proceedings of ACM Great Lake Symposium on VLSI,* GLSVLSI 2012.

[26] Murray, J.; Wettin, P.; Pande, P.; Shirazi, B.; Nerurkar, N; and Ganguly, A.; “Evaluating Effects of Thermal Management in Wireless NoC-Enabled Multicore Architectures”, *Proceedings of IEEE International Green Computing Conference (IGCC),* 2013.

[27] Shamim, M.S.; Mhatre, A.; Mansoor, N.; Ganguly, A.; Tsouri, G., "Temperature- aware wireless network-on-chip architecture," *inGreen Computing Conference (IGCC), 2014 International* , vol., no., pp.1-10, 3-5 Nov. 2014.

[28] Benini, L. and Micheli, G.D. 2002. Networks on Chips: A New SoC Paradigm. *IEEE Computer Society*, 35(1),70-78. □

[29] Binkert, N. et. al. 2011. The gem5 simulator. *ACM SIGARCH Comput. Archit. News*, 39 (2), 1-7. □

[30] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by backpropagating errors," Cognitive modeling, vol. 5, p. 1, 1988.

[31] Chang, K. et. al. 2012. Performance evaluation and design trade-offs for wireless network-on-chip architectures. *ACM Journal of Emerg. Tech. in Comp. System*, 8 (3), Article 23.

[32] Duraisamy, K., Kim, R. G. and Pande,P.P., "Enhancing performance of wireless NoCs with distributed MAC protocols," Quality Electronic Design (ISQED), 16th International Symp. on, Santa Clara, CA, 2015. □

[33] Ogras, Umit Y., and Radu Marculescu. "" It's a small world after all": NoC performance optimization via long-range link insertion." IEEE Transactions on very large scale integration (VLSI) systems 14, no. 7 (2006): 693-706.

[34] A. L. S. Braga, C. H. Llanos, D. Göhringer, J. Obie, J. Becker and M. Hübner, "Performance, accuracy, power consumption and resource utilization analysis for hardware / software realized

Artificial Neural Networks," *2010 IEEE Fifth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA)*, Changsha, pp. 1629-1636, 2010.

[35] Vijayakumaran, Vineeth, Manoj Prashanth Yuvaraj, Naseef Mansoor, Nishad Nerurkar, Amlan Ganguly, and Andres Kwasinski. "CDMA enabled wireless network-on-chip." ACM Journal on Emerging Technologies in Computing Systems (JETC) 10, no. 4 (2014): 28.

[36] Mansoor, N., Ganguly, A., & Yuvaraj, M. P. (2013, October). An energy-efficient and robust millimeter-wave Wireless Network-on-Chip architecture. In Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT), 2013 IEEE International Symposium on (pp. 19-24). IEEE.

[37] Shah, A., Mansoor, N., Johnstone, B., Ganguly, A., & Alarcon, S. L. (2014, September). Heterogeneous photonic Network-on-Chip with dynamic bandwidth allocation. In System-on-Chip Conference (SOCC), 2014 27th IEEE International (pp. 249-254). IEEE.

[38] DiTomaso, Dominic, Avinash Kodi, Savas Kaya, and David Matolak. "iWISE: Inter-router wireless scalable express channels for network-on-chips (NoCs) architecture." In High Performance Interconnects (HOTI), 2011 IEEE 19th Annual Symposium on, pp. 11-18. IEEE, 2011.

[39] N. Mansoor, P. J. S. Iruthayaraj and A. Ganguly, "Design Methodology for a Robust and Energy-Efficient Millimeter-Wave Wireless Network-on-Chip," in IEEE Transactions on Multi-Scale Computing Systems, vol. 1, no. 1, pp. 33-45, Jan.-March 1 2015.

[40] A. K. Mishra, N. Vijaykrishnan and C. R. Das, "A case for heterogeneous on-chip interconnects for CMPs," 2011 38th Annual International Symposium on Computer Architecture (ISCA '11), San Jose, CA, pp. 389-399, 2011.

[41] M. Badr and N. E. Jerger, "SynFull: Synthetic traffic models capturing cache coherent behaviour," in ACM/IEEE 41st International Symposium on Computer Architecture (ISCA '14), Minneapolis, MN, pp. 109-120, 2014.

[42] M. Palesi, M. Collotta, A. Mineo, and V. Catania, "An Efficient Radio Access Control Mechanism for Wireless Network-On-Chip Architectures," in J. Low Power Electron. Appl., vol. 5, no. 2, pp. 38 – 56, 2015.

[43] P. P. Pande, C. Grecu, M. Jones, A. Ivanov, and R. Saleh, "Performance evaluation and design trade-offs for network-onchip interconnect architectures," in IEEE Trans. Comput., vol. 54, no. 8, pp. 1025–1040, Aug. 2005.

[44] Mansoor, N., Shamim, S., & Ganguly, A. (2016, June). A demand-aware predictive dynamic bandwidth allocation mechanism for wireless network-on-chip. In System Level Interconnect Prediction (SLIP), 2016 ACM/IEEE International Workshop on (pp. 1-8). IEEE.

[45] S. Abadal, M. Nemirovsky, E. Alarcón, and A. Cabellos-Aparicio. "Networking Challenges and Prospective Impact of Broadcast-Oriented Wireless Networks-on-Chip," In Proc. of the 9th ACM International Symp. on Networks-on-Chip (NOCS '15), Vancouver, Canada, Article 12, 2015.

[46] X. Yu, J. Baylon, P. Wettin, D. Heo, P. P. Pande and S. Mirabbasi, "Architecture and Design of Multichannel Millimeter-Wave Wireless NoC," in IEEE Design & Test, vol. 31, no. 6, pp. 19-28, Dec. 2014.

[47] C. Wang, W. H. Hu and N. Bagherzadeh, "A Wireless Network-on-Chip Design for Multicore Platforms," in Proc. of the International Euromicro Conference on Parallel, Distributed and Network-Based Processing, Ayia Napa, pp. 409-416, 2011.

[48] G. Piro, et al., “Initial MAC Exploration for Graphene-enabled Wireless Networks-on-Chip,” in Proc. of the International Conference on Nanoscale Computing and Communication (NANOCOM' 14). ACM, New York, USA, Article 7, 2014.

[49] M. Palesi, M. Collotta, A. Mineo, and V. Catania, “An Efficient Radio Access Control Mechanism for Wireless Network-On-Chip Architectures,” in J. Low Power Electron. Appl., vol. 5, no. 2, pp. 38 – 56, 2015.

[50] Aswath Narayana, Sandeep, "An Artificial Neural Networks based Temperature Prediction Framework for Network-on-Chip based Multicore Platform" (2016). Thesis. Rochester Institute of Technology. Accessed from http://scholarworks.rit.edu/theses/8994