HREN: A Hybrid Reliable and Energy-Efficient Network-on-Chip Architecture

Padmaja Bhamidipati, *Student Member, IEEE,* and Avinash Karanth, *Senior Member, IEEE,*

Abstract—As transistor scales down to sub-nanometer and processing cores with billions of transistors are integrated, reliable and energy-efficient Network-on-Chip (NoC) architectures are critical for improving performance of multicores. Near Threshold Voltage (NTV) scaling and approximate communication are popular techniques to reduce the energy consumption of NoC. Applications, which are insensitive to lower precision, can tolerate some loss in quality and take advantage of approximate communication. While approximate communication can improve energy-efficiency, these techniques are vulnerable to faults which in turn compromises reliability. In this paper, we propose **HREN: A Hybrid Reliable and Energy-efficient Network-on-Chip** architecture that improves the reliability of NoC while utilizing both approximate communication and NTV scaling techniques in a multi-layered reliability model. HREN architecture facilitates two-levels of data approximation by identifying and compressing frequently repetitive patterns in the application data, thereby reducing the number of packet transmissions in NoC. As applications exhibit different traffic patterns in NoC, HREN switches the voltage mode of the network globally at runtime, thereby reducing the dynamic energy consumption while performing data approximation. HREN carefully monitors and handles the faults occurred due to NTV scaling and approximation while maintaining the fine balance between energy consumption and error rate. From our simulation results, HREN demonstrates up to 2.8x dynamic energy savings while reducing latency up to 2x. HREN shows an improvement of 4x to 5.5x in Energy-Delay Product over the baseline model for AxBench approximation benchmark suit on a 4 $\times$ 4 concentrated mesh (CMESH) architecture.

Index Terms—Network-on-Chips, Reliability, Energy-Efficiency, Near-Threshold Voltage, Approximation.

1 Introduction

Recent advances in silicon technology has enabled integrating billions of transistors on a single chip to improve the execution speed of multicore applications. Advanced multicore architectures such as Tegra Xavier SoC [1], Ryzen-based Epyc by AMD [2], GV100 Volta [3], and Everest by Xilinx [4] contain billions of transistors. As more cores are integrated, dynamic and static power consumption of Network-on-Chips (NoCs), which ensures efficient inter-core communication, increases considerably. Prior research has proposed several energy-efficient techniques including Dynamic Voltage and Frequency Scaling (DVFS) [5], [6], [7], Near Threshold Voltage (NTV) Scaling [8] [9] [10] and power gating [11] [12] to improve the energy-efficiency of NoCs. As the supply voltage can only be scaled to 70% of the nominal voltage under standard DVFS technique, the region closest to the transistor threshold voltage (V_{th}) has not been extensively explored. In NTV scaling, the transistor operates at supply voltages closer to the threshold voltage and the energy-efficiency increases to more than 5X as the operating voltage is scaled by more than 25% of the nominal supply voltage [13], [9]. However, as energy efficiency techniques focus on improving the power consumed, lower supply voltage may lead to lowering the reliability of NoC.

To further improve energy-efficiency of multicores, approximate computing has been proposed which takes advantage of applications' ability to tolerate errors while balancing the power-latency trade-offs. Recently, researches have implemented approximate computing in various fields such as machine learning, fluid dynamics, video processing, image recognition, and many more [14] [15] [16]. Most recently, the approximate *communication*, which is an extension of approximate computing, is implemented to reduce the communication overhead of the NoC [16] [17]. In approximate communication, the data transmitted between two processing cores is approximated using compression, synchronization, and value prediction. Recent work [18] [19] [20] have showed the benefits of data compression and encoding techniques applied to the data packet to reduce the energy consumption and improve the performance of NoC.

With both NTV scaling and approximate communication, permanent and transient faults due to transistor aging and elevated device temperature could potentially cause irreversible damage to the NoC components and impact reliability. Prior work has shown that aging affects the shift in threshold voltage (V_{th}) due to the interface traps and the fluctuations in the charge density caused by Hot Carrier Injection (HCI) [21] and Negative Bias Temperature Instability (NBTI) [22], thereby reducing the lifetime of the transistor. As the variation in ΔV_{th} exceeds 10%, permanent faults are observed in the circuit [23]. Scaling down the supply voltage of the device slows down its aging process not only due to a decrease in temperature but also due to a decrease in the electric field. However, as the supply voltage is decreased, transient faults such as Single Event Upsets (SEUs) are observed in transistors, flipping binary bits (0 or 1) which results in logical errors in the data transmitted. Prior work has proposed various error handling techniques such as retransmission [24], error-correcting codes (ECC) [25], data

- *P. Bhamidipati is with the Electrical and Computer Engineering Department at University of Cincinnati and A. Karanth is with the School of Electrical Engineering and Computer Science, Ohio University, Athens, OH, 45701. This work was completed when P. Bhamidipati was working towards her MS at Ohio University.*
  E-mail: {pb261616, karanth}@ohio.edu

Manuscript received January 13, 2022

encoding and decoding [18] [19] and other techniques to improve the reliability of NoC.

In this paper, we propose **HREN: A Hybrid Reliable and Energy-efficient Network-on-Chip** that combines energy-efficient techniques such as voltage scaling and data approximation along with a hybrid reliability model to balance energy-efficiency, reliability and performance of NoC. HREN implements voltage scaling that includes five voltage modes which are carefully chosen such that the supply voltage is scaled according to the incoming traffic to reduce congestion and improve energy-efficiency. Along with voltage scaling, HREN proposes to reduce both the dynamic energy consumption and network packet latency using approximate communication by reducing the number of packet transmissions in NoC. The two-level data approximation framework of HREN identifies duplicate and similar data packet in the NoC for compression. It consists of a default approximation algorithm and an approximation engine (HREN-approx) that compresses data which is closest to the reference value. The default approximation algorithm of HREN is independent of traffic conditions and error-rate and compresses similar data patterns that are communicated from a given source to its destination. HREN-approx identifies and compresses the closest data patterns at different error rates.

The proposed reliability model of HREN is a hybrid error correction and detection scheme with two-layered architecture to mitigate soft errors caused due to lower voltage modes and data approximation. When NoC is operated at higher DVFS modes and at lower data compression rate, error rates are typically low, and therefore, the low-level encoding scheme which is also known as End-to-End (E2E) error correction, is applied to NoC. Similarly, when operating under low voltage modes (NTV) and at higher data compression rate, error rates are higher, and therefore, high-level encoding scheme such as switch-to-switch (S2S) error correction, is applied to NoC. This multi-layered hybrid reliability scheme handles Single Event Upsets (SEUs) and permanent faults that are encountered due to lower supply voltages and data compression, thus achieving a fine balance between power consumption and reliability. From our simulation results, HREN demonstrates up to 2.8X dynamic energy savings while reducing latency up to 2X. HREN shows an improvement of 4X to 5.5X in Energy-Delay Product over the baseline model for AxBench approximation benchmark suit on a 4 $\times$ 4 concentrated mesh (CMESH) architecture. The following are the major contributions of this article:

(1) Voltage scaling-aware NoC: We design link utilization aware voltage scaling technique (including NTV and DVFS) to improve energy-efficiency and manage congestion in NoC while maintaining upper bound on the energy-delay-product (EDP).

(2) Approximate communication: We explore repetitive patterns in the data that is communicated across NoC while performing data compression to reduce the number of packet transmissions and thereby improving the energy-efficiency of NoC.

(3) Reliability model for NoCs: We implemented a hybrid error correction and detection scheme with two-layered architecture to handle link wear-out and SEUs that are encountered at different supply voltage and approximation range, thus achieving a fine balance between power consumption and network performance.

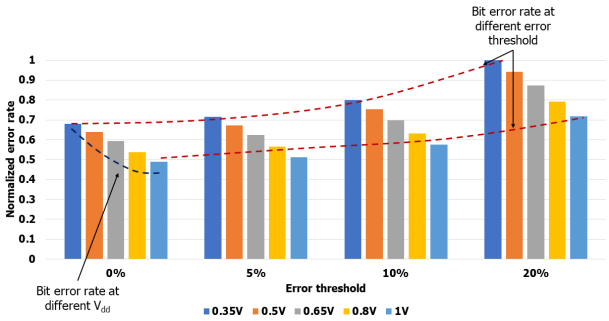


Fig. 1: Normalized error rate at various voltage modes and error threshold percentage to analyze the power-reliability trade-off.

2 HREN Architecture

In this section, we will discuss HREN architecture, voltage scaling mechanism, and data approximation algorithms used in the architecture. Later in this section, we will elaborate the proposed hybrid reliability model which handles the errors that are introduced due to approximation and voltage scaling.

HREN architecture evaluates the power-performance trade-offs in NoC by considering several critical parameters such as supply voltage, error rate, energy consumed and the number of packets routed across NoC. Figure 1 shows the plot of normalized error rate at various voltage modes and percentages of error threshold. The *error threshold* is the number of bits in error for every 100 bits transmitted. For example, at 5% error threshold, we induce errors in the data or approximate the data, such that there are 5 bits in error for every 100 bits transmitted from the source router. From Figure 1, as the supply voltage of NoC decreases, error rate increases but dynamic power decreases. On the other hand, as the error threshold (compression rate) increases, the error rate increases but the number of packet transferred decreases. In order to balance the power and reliability of NoC, we propose HREN protocol with four stages as described below:

STAGE 1: In stage 1, HREN decides on the appropriate voltage mode (among the five voltage modes) to improve energy-efficiency of NoC.

STAGE 2: In stage 2, HREN monitors the error rate of NoC according to the network parameters and the input parameters such as supply voltage/operating frequency set by the user. In this stage, HREN captures the bits in error (both single and multiple bits) caused due to voltage scaling from stage 1, and errors caused due to change in threshold voltage of the transistor (ΔV_{th}) which depends on aging, temperature, and other operating conditions.

STAGE 3: In stage 3, depending on the decisions made in stage 1 and stage 2, approximation algorithm is selected to improve the throughput and energy-efficiency of NoC.

STAGE 4: In stage 4, HREN captures the error rate due to data approximation and the error rate from stage 2 in order

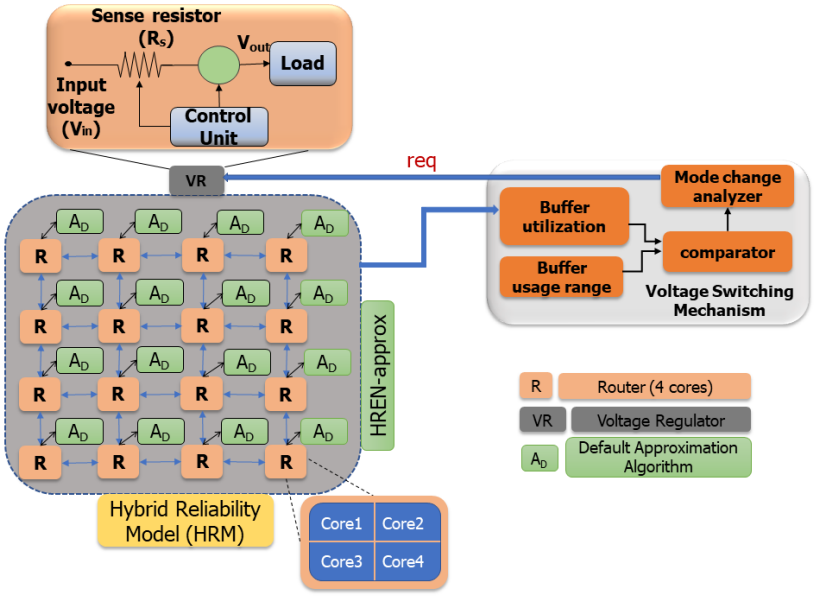
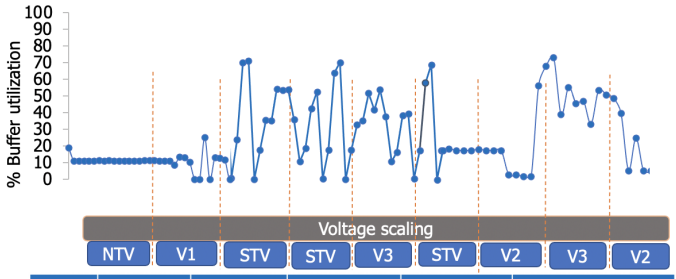


Fig. 2: Proposed HREN architecture, implemented on a 4x4 concentrated mesh (CMESH) topology. HREN consists of an automated voltage switching mechanism that depends on the buffer utilization of the NoC, the default approximation blocks at every router (AD), the HREN-approx approximation block, and the hybrid reliability model.

to tune the error detection strength using hybrid reliability model of HREN.

2.1 HREN Layout

Figure 2 shows the proposed HREN architecture that is implemented on a 4×4 concentrated mesh (CMESH) topology with a 64-core NoC. The voltage switching mechanism of HREN implements the voltage scaling technique which switches the supply voltage of NoC depending on the buffer utilization of the network. The mode change analyzer processes the input from the comparator which compares the buffer utilization of NoC with the buffer usage range set by the user. The mode change analyzer then sends out a request signal to the voltage regulator in order to change the voltage mode of NoC. This voltage switching mechanism tunes the supply voltage and the frequency of NoC globally between five-voltage modes as shown in Figure 3. The frequency, load, temperature and delay calculations are designed similar to RETUNES design [26]. HREN scales the supply voltage of NoC to the appropriate voltage mode and frequency globally at run-time for every epoch. HREN optimum epoch size is chosen to be 100 cycles by carefully monitoring the power and performance trade-off for several applications. The process of scaling voltage up/down of NoC is implemented in three stages. If the network senses a voltage scale-down request, each router reduces its operating frequency to the new appropriate frequency in the first stage, and then holds for a fixed number of cycles to account for the network to wake-up and adjust to the new environment in the second stage. In the third stage, supply voltage of NoC is stepped down concluding the step-down process. On the other hand, if the network senses a voltage scale-up request, each router steps up to its appropriate supply voltage in the first stage and then then holds for a fixed number of cycles to account for the network to wake-up and adjust for the new environment in the second stage. In the third stage, operating frequency of NoC is adjusted according to the current voltage mode.



Mode	Voltage (V)	Freq (Hz)	Load (flits/cycle)	Temp (Celsius)	Overhead Delay (Cycles)
NTV	0.35	200 M	0.01	75-77	8
v1	0.5	800M	0.1	76-82	5
v2	0.65	1.5G	0.2	80-93	4
v3	0.8	2G	0.3	90-101	2
STV	1.0	2.3G	0.4	97-104	1

Fig. 3: Represents the voltage switching mechanism of the five voltage modes. The calculated frequency, load, temperature and overhead delay values for all the proposed voltage modes are shown [26].

The two-layered data approximation framework of HREN is a combination of default algorithm (AD) and a HREN- approx, as shown in the Figure 2. The data approximation algorithms of HREN identify and compress the duplicate data patterns that are observed in the application data. The default algorithm of HREN is operated locally at every router while the HREN-approx is operated globally for the entire network. HREN is associated with error handling design to handle the errors observed due to

voltage scaling and data as shown in Figure 2. The Hybrid Reliability Model (HRM) of HREN monitors the Single Event Upsets (SEUs) and error threshold of the approximation algorithm globally for each epoch. Depending on the computed bit error rate of NoC at that epoch, HRM adjusts the strength of error correction code, thereby improving the reliability of NoC.

2.2 HREN Two-Layered Data Approximation Framework

We propose a two-layered data approximation design to identify duplicate patterns of data in an application to reduce the number of packet transmissions in NoC, thereby improving energy-efficiency and performance. The two layered hybrid reliability model consists of approximation algorithm (AD) in its first layer, and HREN-approx (a group of three algorithms) in its second layer. The default approximation algorithm (AD) is the primary layer of approximation that is applied throughout the execution process of an application. Figure 4 shows the default approximation algorithm where AD represents the default algorithm at every router operating individually. Before entering the router, a packet of 256 bits is split into 32-bit flits and each flit is sent to the default approximation algorithm. The default approximation algorithm at the source router then compares the payload of every flit (for example flit 1) with its adjacent flit (for example flit 2). If the comparator output is not equal, then the adjacent flit (flit 2) is transmitted to the destination router in the path directed by the routing algorithm. On the other hand, if the comparator output is equal, then the payload of the flit 1 is encoded to compress flit 2 at the approximation logic block. In this case, only the encoded flit 1 is transmitted as both the flits have the same destination router. The encoded bit is then sent back to source router from the AD block through the approximation logic block, and then transmitted to the destination router. At the destination router, the packet is then decoded to present flit 1 and flit 2 to the destination router. The default approximation algorithm of HREN improves the energy efficiency of NoC by reducing the number of bit transmissions.

HREN-approx is the second layer of approximation algorithm with three levels of data compression which is applied to the application data globally in NoC. Figure 5 shows the HREN-approx algorithm where the data at the source router is compressed before communication. Initially, in HREN-approx, the algorithm detection block chooses the appropriate algorithm among algo1, algo2, and algo3, depending on the error type of NoC: No Errors (NE), Few Errors (FE) and Many Errors (ME). The types of errors and the algorithm selection methodology is explained next in subsection 2.3. Figure 6 shows the logic tables of algo1, algo2, and algo3 respectively. If algo1 is selected, then the last two bits of every packet is approximated and compressed into a single bit. If algo2 is selected, then the last four bits of every packet is approximated and compressed as shown in Figure 6. Similarly, if algo3 is selected, then the last eight bits of every packet is approximated and compressed as shown in Figure 6. HREN-approx algorithm reduces number of bits that are transmitted in NoC, thereby reducing the dynamic power that is consumed in communicating packets from the source router to the destination router.

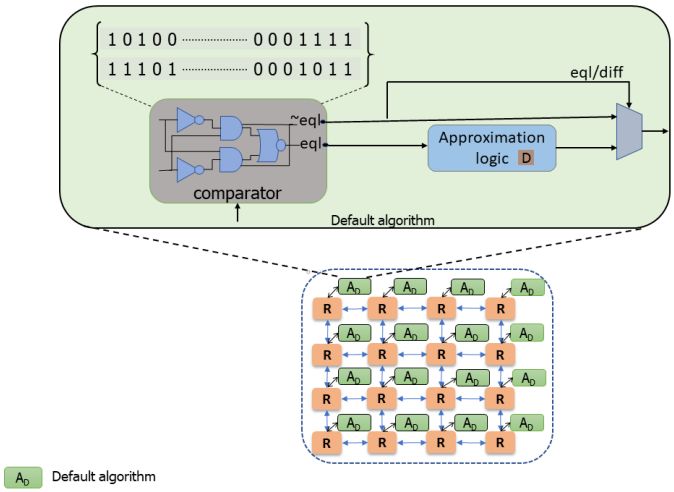


Fig. 4: Architecture of HREN: Micro-architecture of the first layer of the approximation algorithm also known as default approximation algorithm (AD) connected to a 4x4 NoC.

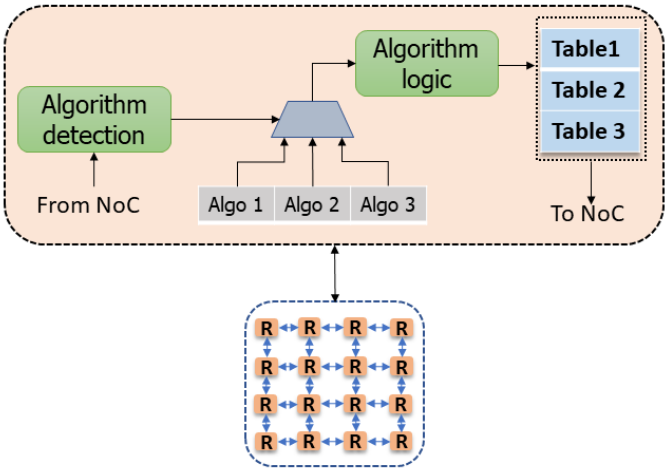


Fig. 5: Architecture of HREN: Micro-architecture of the second layer of the approximation algorithm also known as HREN-approx and its communication path to a 4x4 NoC.

2.3 Hybrid Reliability Model

The hybrid reliability model of HREN monitors and captures the error rate of the network before adjusting the error correction strength of NoC. Figure 7 shows the fault monitoring system of HREN, where the table produces various error types (NE, FE, ME) observed by our design due to voltage scaling [26]. Error type NE (No Errors) represents no bit-errors, FE (Few Errors) represents 1-bit or 2-bit errors, and ME (Many Errors) represents more than 2-bit errors in a flit. Initially, the fault monitoring system decides the error type depending on the active voltage mode and percentage of probability of error occurrence in NoC. The fault monitoring system receives the voltage mode of NoC from the voltage switching mechanism of HREN, whereas, the probability of error is calculated from the router fault model [27] using the temperature from the table in Figure 3. Hence, the probability of error (Pe) depends

Algo 1: Table1

bit	dup
00	0
01	0
10	1
11	1

Algo 2: Table2

bit	dup
0000	0
0001	0
0010	0
0011	0
0100	01
.	
.	
.	
1100	1
1101	1
1110	1
1111	1

Algo 3: Table3

bit	dup
00000000	0
00000001	0
00000010	0
00000011	0
00000100	0
00000101	0
00000110	0
00000111	0
00001000	01
.	
.	
.	
11111100	1
11111101	1
11111110	1
11111111	1

Fig. 6: Figure shows the logic tables for algo1, algo2, and algo3 of HREN-approx algorithm.

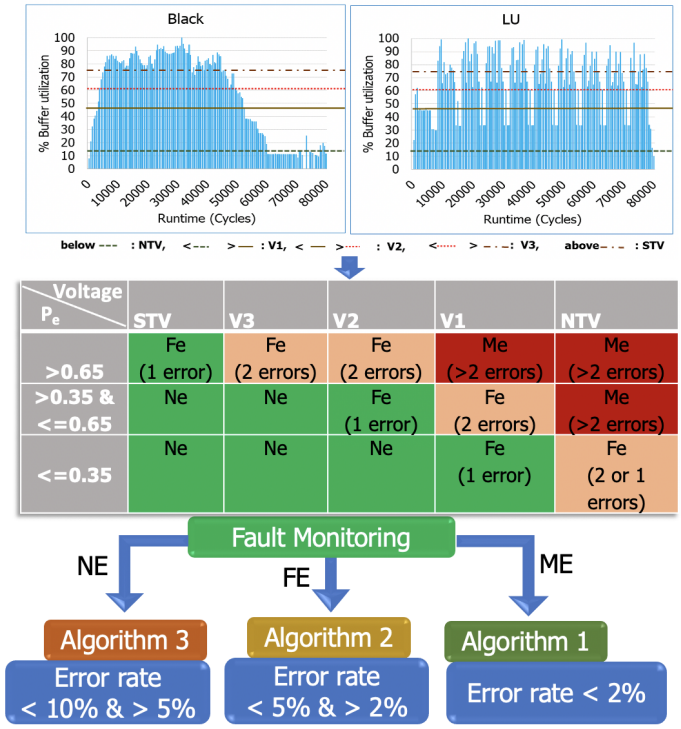


Fig. 7: Fault monitoring system (a) Monitors error type of the network for a supply voltage and link utilization (shown on the right for black and LU test cases) and (b) Selects algorithms for HREN-approx depending on the error type.

on the network parameters such as, temperature, supply voltage, router parameters and operating conditions. We correlated the error types (NE, FE, ME) with Pe to generate the table from Figure 7. Finally, the decision regarding the HREN-approx algorithm is made depending on the error type (NE, FE, ME) of NoC. If the error type of NoC is NE, algorithm 3 (algo3) is selected where the error rate is 5-10%. This mapping of error type with algorithm is performed in order to balance the total number of errors observed during a communication, as selecting algo3 for a higher error rate (ME) might result in permanent data loss. Similarly, if the error type is FE, algo2 is selected where the error rate is 2-5% , and if the error type is ME, algo1 is selected where the error rate is below 2%. The bit error rate for the algorithms is calculated using the BER formula (number of bit errors/total number of bits). As HREN does not change the HREN-approx algorithm dynamically, there is a probability that the error rate might increase while transmitting a packet from source to destination. In order to handle such errors, we designed a hybrid reliability model that tunes its error correcting strength depending on the error rate observed in NoC.

The hybrid reliability model of NoC consists of a two-layered encoding scheme with Switch-to-Switch (S2S), the strong Error Correcting Code (ECC) layer and End-to-End (E2E), the weak ECC layer. In S2S layer, every packet is encoded and decoded at each router in its path to ensure maximum reliability, whereas, in E2E layer, every packet is encoded at the source router and the destination router. Depending on error range of NoC that is calculated previously, the global control unit switches between E2E and S2S encoding layer dynamically. If NoC is at low error-rate, E2E encoding layer is activated. On the other hand, if NoC is at high error rate, S2S encoding layer is activated. Before initiating the transmission process, hybrid reliability model calculates the error range due to voltage scaling and approximation to decide on the encoding layer that needs to be activated initially. After switching the encoding layer, a flit is transmitted from one router to its next router in order to safeguard the communication process.

Figure 8 shows our proposed reliability model which includes error correcting and detecting codes (ECC) such as Cyclic Redundancy Check (CRC) and Single Error Correction and Double Error Detection (SECDED) hamming code. Initially, all the packets are encoded with CRC-32 irrespective of the encoding layer. In CRC-32, a 32-bit checksum is added to 224-bit data to form a 256-bit packet which is then split into eight 32-bit flits before entering source router. After entering the source router, each flit is encoded using H(39,32) SECDED hamming encoder. If E2E encoding layer is active, SECDED is applied at the source and at the destination routers only. However, if S2S encoding layer is active, SECDED is also applied at the intermediate router along with the source and the destination routers, thereby providing higher fault coverage.

If the encoding layer SECDED is not active and error cannot be corrected, a request for full-retransmission is sent to the source router from the current router. The source router then retransmits the packet to the current router and updates with a flag to indicate the retransmission of

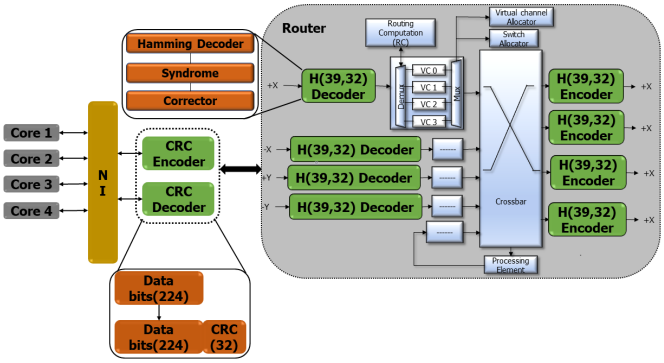


Fig. 8: Proposed hybrid reliability model which includes CRC and hamming codes.

the packet. If the encoding layer SECDED is active all the single-bit errors are corrected. However, if the error cannot be corrected, a single-hop retransmission is requested from the up-stream router to the current router. HREN avoids repetitive retransmission of packets using the status of the retransmission flag. If the flag is already set for a packet, HREN blocks the retransmission request of the same packet, thus avoiding number of retransmissions and reducing the energy consumption of NoC. All the operations of HREN are performed at each epoch cycle where the control unit for voltage scaling, HREN-approx algorithm, fault monitoring system, and, hybrid reliability model are operated globally.

Adaptive Routing: HREN uses west-first adaptive routing algorithm to distribute its packets uniformly throughout NoC to improve the lifetime of the transistor. For every epoch, the routing algorithm collects the average link utilization for the current router at runtime. The router then chooses the least utilized link from the available 4 links (directions) to route a packet along that direction. West-first routing algorithms prohibits north-west and south-west turns and therefore, prevents two of the possible eight turns of the Turn-model to avoid deadlocks. The routing algorithm determines the path of the packet based on the following conditions included in routing computation:

Criteria 1: If the average link utilization along x-axis is greater than y-axis, provided x/y-coordinate of current and destination routers are not equal, the packet is routed along the y- coordinate.

Criteria 2: Similarly, if the average link utilization along x-axis is less than y-axis, provided x/y-coordinate of current and destination router are not equal, the packet is routed along the x- coordinate

Criteria 3: If the x-coordinate of the current router is equal to that of the destination router, the routing algorithm routes the packet along y-direction and if the y-coordinate of the current router is equal to that of the destination router, the routing algorithm routes the packet along x-direction ignoring link utilization values.

3 HREN Evaluation Approach

In this section, we evaluate our proposed HREN architecture and compare its variations against four other schemes such as V^5 (five-voltage mode scheme), V^2 (two-voltage mode design), Always-NTV, and Always-STV, using AxBench workloads. The five evaluation schemes and their design approaches are explained below:

Always-NTV: Always-NTV scheme allows NoC to operate at the Near-Threshold Voltage (NTV) region. In this work, NTV is assumed to be 0.35V and is supplied to NoC globally, implementing coarse-grain approach. While operating in Always-NTV scheme, NoC achieves higher energy-efficiency at the cost of packet latency and degraded reliability.

Always-STV: Always-STV scheme allows NoC to operate at the Super-Threshold Voltage (STV) region also known as nominal voltage region. In this work, STV is assumed to be 1.0V and is supplied to NoC globally, implementing coarse-grain approach. While operating in Always-STV scheme, NoC experiences low latency, high reliability and low energy-efficiency.

V^2: V^2 scheme is a two-voltage mode design which switches the operating voltage of NoC between NTV and STV while globally supplying a single operating voltage to NoC for a given epoch. In this work, NTV is assumed to be 0.35V and STV is assumed to be 1.0V. The two voltage modes of V^2 are switched based on the buffer utilization of NoC. In V^2 scheme, we operated NoC in NTV mode when the buffer utilization of NoC is 25-30%, and operated NoC in STV mode when the buffer utilization of NoC is above 30%. This two-voltage design improved energy efficiency of NoC when compared to Always-STV scheme and reduced packet latency when compared to always-NTV scheme.

V^5: V^5 scheme is a five-voltage mode design that switches the supply voltage among five voltage modes based on the buffer utilization of NoC. The five voltage modes of this scheme include NTV at 0.35V, V1 at 0.55V, V2 at 0.6V, V3 at 0.8V, and STV at 1.0V. In this scheme, NoC is globally supplied with a single operating voltage for a given epoch. The buffer utilization range for each voltage mode is fixed and is carefully assigned to ensure maximum energy-efficiency. NTV voltage mode is activated when the buffer utilization range of NoC is below 15%. Similarly, V1, V2, V3, and STV voltage modes are activated when the buffer utilization range of NoC is between 15-45%, 45-60%, 60-75%, and above 75% respectively. Compared to V^2 scheme, V^5 scheme is more energy efficient as it can capture a higher number of variations in buffer utilization levels due to multiple voltage modes (5 voltage levels), thus reducing the dynamic energy consumption.

HREN: HREN scheme is our proposed architecture which consists of five-voltage mode design, a dual-layered approximation algorithm to improve energy-efficiency, and a hybrid-reliability model to improve reliability of NoC. The main difference between V^5 and HREN is that the HREN is equipped with an additional layer of data approximation when compared to V^5. However, both HREN and V^5 schemes include hybrid reliability model in order to improve reliability of NoC.

4 Simulation Results

In this section, we described the performance of our proposed HREN architecture by dividing the results into four main categories:

- In category 1, we compared the dynamic energy, latency, normalized error rate, and energy-Delay Product (EDP) of the approximation algorithms used in the design. In this category, we applied approximation algorithms on Always-STV, Always-NTV, V^2, and V^5 schemes to understand the energy-performance tradeoff in NoC.
- In category 2, we analyzed the threshold voltage change (that effects reliability) due to various factors such as supply voltage scaling, change in temperature, and aging. We also discussed the power-performance trade-offs of the proposed HREN hybrid reliability model.
- In category 3, we compared dynamic energy, latency, and energy-Delay Product (EDP) of all the five evaluation schemes which includes Always-STV, Always-NTV, V^2, V^5, and our proposed scheme, HREN.
- In category 4, we analyzed the area cost of approximation algorithms, control unit, and reliability design of HREN which includes CRC-32 and Hamming (H(39,32)) codes.

HREN is evaluated on a 4×4 CMESH with 64 cores and a 256-bit packet that is split into eight 32-bit flits before entering the network. In this work, we choose five applications with different domains from AxBench [28] CPU benchmark suit as shown in the Table 1. The variation in data patterns of AxBench benchmarks were evaluated by the proposed data approximation techniques to analyze the energy-efficiency and packet latency. The dynamic energy, latency, and the area cost of the routers and the links were determined by DSENT NoC modeling tool [29] using TSMC 45nm library. The area, dynamic energy, and latency cost of the data approximation design and hybrid-reliability model are obtained from Synopsis Design Compiler using 45nm technology library.

TABLE 1: Benchmarks Used in HREN

Applications	Domains
JPEG	Image Processing
BLACKSCHOLES	Financial Analysis
FFT	Signal Processing
JMEINT	3D Gaming
SOBEL	Image Processing

4.1 Category 1: Data Approximation Algorithms

Dynamic energy and Latency analysis: Figure 9 shows the normalized average dynamic energy savings of Always-STV, Always-NTV, V^2, and V^5 schemes under default approximation algorithm (AD). Even though we apply AD to all the packets at the source router, the algorithm under V^5 scheme (HREN's AD) shows approximately 15% of average dynamic energy savings when compared to the V^5 scheme with no data approximation applied. As the AD algorithm promises maximum energy savings with the V^5 scheme, we applied the algorithm for every packet at the source router to increase the overall energy savings and balance the energy consumption at every voltage mode (in V^5 scheme).

Figure 10 shows the normalized average dynamic energy of NoC for Always-NTV, Always-STV, V^2, and V^5 evaluation schemes with HREN-approx design. The plot analyzes the normalized average dynamic energy consumed by each packet to transmit from source to destination and to retransmit the packets if necessary. On average, algorithm 1 (algo 1) in V^5 scheme shows 45-50% savings in dynamic energy consumption when compared to Always-STV scheme and shows almost 25% savings in dynamic energy consumption when compared to V^2 scheme. Always-NTV consumes minimum dynamic energy when compared to all the other schemes as NoC is operated in NTV or lowest operating voltage throughout application execution process. Similarly, on an average, algorithm 2 (algo 2) and algorithm 3 (algo 3) in V^5 scheme showed dynamic energy savings of nearly 65% and 82% when compared to those in Always-STV scheme respectively. In the V^5 scheme, algo 2 and algo 3 showed up to 37% and 69% of dynamic energy savings when compared to those in V^2 scheme respectively. From Figure 10, the dynamic energy savings in V^5 scheme due to the algo 2, and algo 3 is approximately 47% and 72% on an average when compared to algo 1.

Figure 11 shows the normalized average packet latency of NoC for Always-NTV, Always-STV, V^2, and V^5 evaluation schemes. On an average, V^5 shows approximately 67-84% decrease in packet latency of NoC when compared to the packet latency observed in Always-NTV scheme. Similarly, V^5 scheme showed approximately 43-65% decrease in average packet latency when compared to the average packet latency in V^2 scheme. We also observed variation in average packet latency among the data approximation algorithms where, algo 2 and algo 3 showed decrease in average packet latency of up to 27% and 53% when compared to the average packet latency in algo 1. Hence, algo 3 shows minimum packet latency in V^5 scheme when compared to all the other schemes with any of the three data approximation algorithms.

Energy-Delay Product Analysis: Figure 12 shows the normalized Energy-Delay Product (EDP) which is the product of the time spent by a packet to reach to its destination router from the source router and the average energy consumption of each packet. The results showed a reduction in EDP on an average of 64.3%, 58.4%, and 34.1% for V^5 scheme when compared to Always-NTV, Always-STV, and V^2 schemes respectively. Figure 12 also shows that the EDP of algo 2 and algo 3 in V^5 scheme is 54% and 84% lower when compared to the EDP of algo 1. Hence, from this result we demonstrated that the EDP of algo 3 in V^5 scheme is minimum when compared to the other two approximation algorithms in the 4 evaluation schemes (Always-NTV, Always-STV, V^2, and V^5).

4.2 Category 2: Reliability Analysis

Our model considers the change in threshold voltage due to voltage scaling ($\Delta V_{th\ voltagescaling}$), aging ($\Delta V_{th\ aging}$), and temperature ($\Delta V_{th\ temp}$) change as the reliability metric. Traffic in NoC is correlated with temperature variations, and non-uniform distribution of this traffic causes hotspots affecting reliability of the network. We used Hotspot thermal model [30] and router fault model to monitor such variations in the network temperature.

Transistor aging caused due to HCI and NBTI are modeled as the threshold voltage variation using Synopsys

This article has been accepted for publication in IEEE Transactions on Emerging Topics in Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TETC.2022.3147407

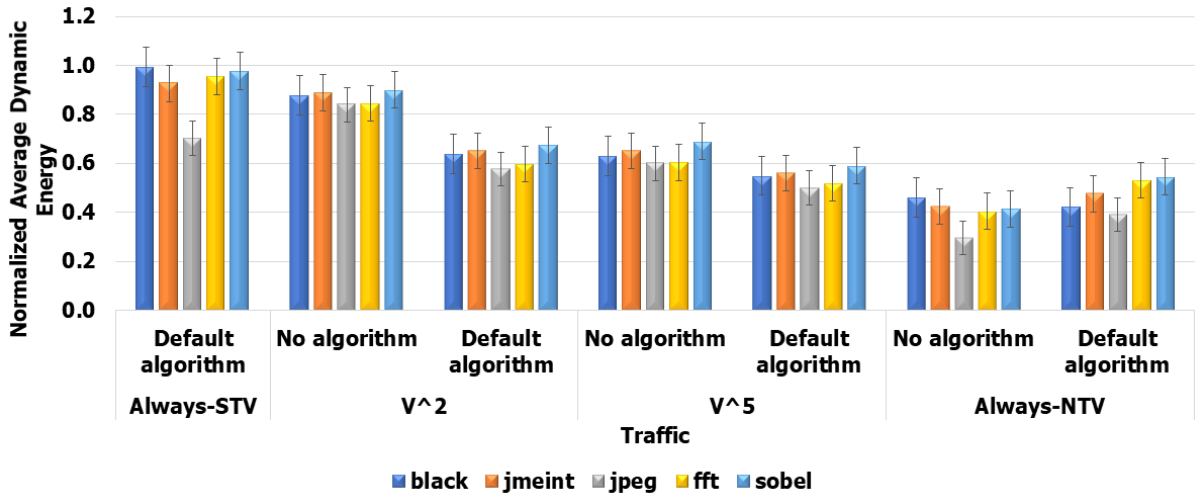


Fig. 9: Normalized average dynamic energy consumption of Always-STV, Always-NTV, V^2, and V^5 schemes for Default approximation algorithm and no approximation applied to AxBench benchmarks.The graph is normalized to Always-STV scheme when no algorithm is applied.

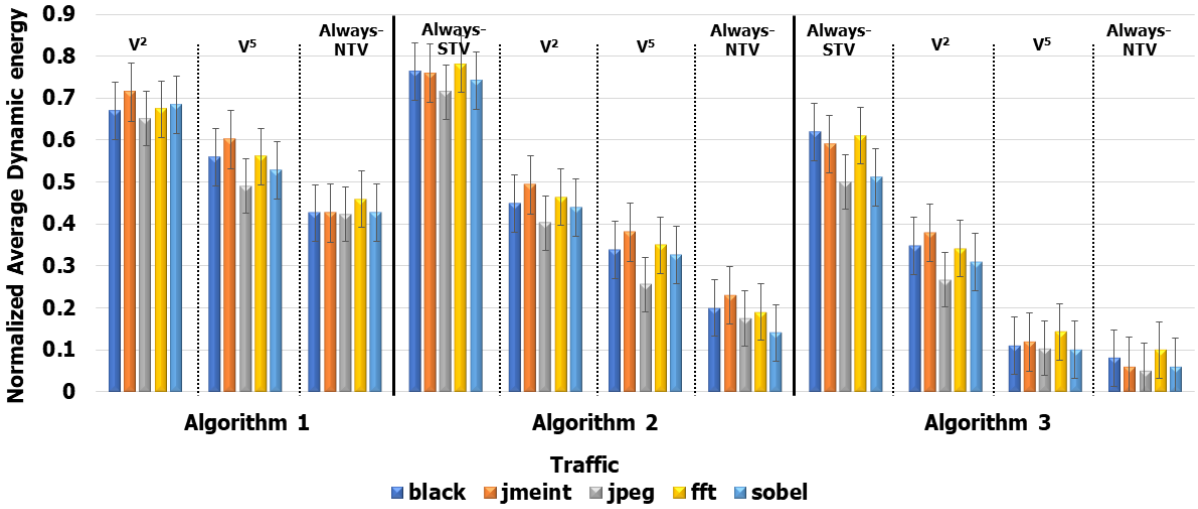


Fig. 10: Normalized average dynamic energy consumption of Always-STV, Always-NTV, V^2, and V^5 schemes for HREN-approx algorithms (algo 1, algo 2, and algo 3) applied to AxBench benchmarks.The graph is normalized to Always-STV scheme in algo 1.

HSPICE tool and Predictive Technology Model (PTM) for a 45nm transistor technology node using the temperature data from above. According to the results shown in Figure 13 for five voltage modes (Always-NTV, V1, V2, V3, and Always-STV) for a degradation period of 10 years, it is evident that the threshold voltage variations (ΔV_{th}) decrease with the operating voltage, mitigating the aging process. Figure 14 shows that HotSpot thermal map for Always-STV scheme under XY-routing (Base model on the left) and HREN (right). Our model showed better temperature distribution and slower aging process when compared to the base model which in turn supports reliability aspect of HREN. Please note that scales are different for both the models with HREN running 10 Kelvin cooler.

HREN hybrid reliability (S2S and E2E) models corrected all single-bit errors and requested re-transmission during multiple-bit errors at the cost of 6% energy consumption. The mean error rate of the base model under Always-NTV scheme is approximately 55% more when compared to our HREN model. where as, the mean error rate of HREN model is 2.5× when compared to Always-STV scheme. HREN offers an error rate in between Always-NTV and Always-STV schemes, thereby balancing reliability and power consumption. In addition to that, we reduced approximately 29% of full-retransmissions (source to destination) when compared to the model with V^5 scheme with full-retransmission of the erroneous packet and no error correction.

4.3 Category 3: Simulation Results of HREN

Dynamic Energy Analysis: Figure 15 shows the normalized average dynamic energy of the five evaluation schemes for

This article has been accepted for publication in IEEE Transactions on Emerging Topics in Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TETC.2022.3147407

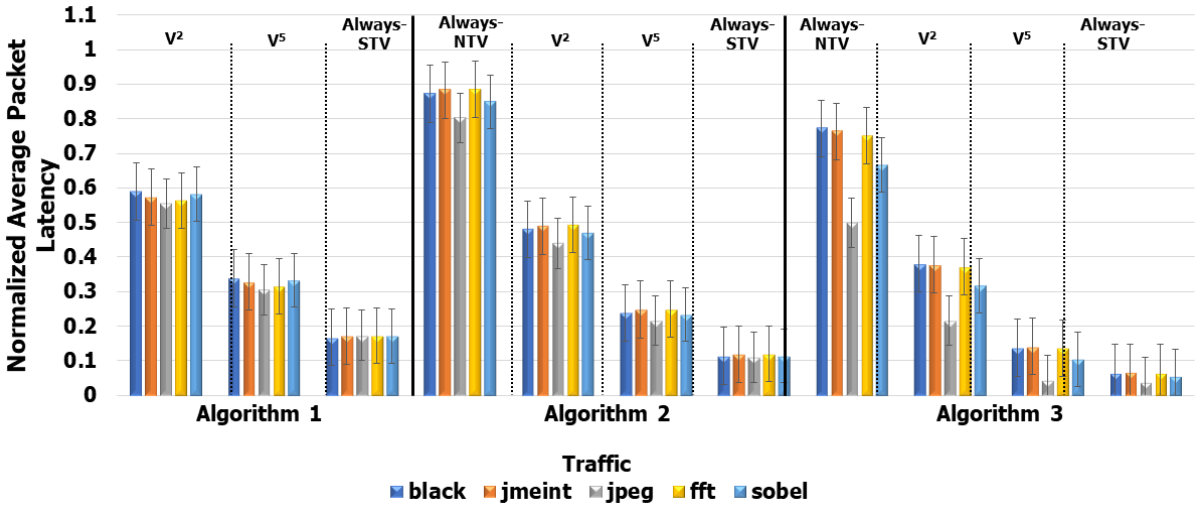


Fig. 11: Normalized average packet latency of Always-STV, Always-NTV, V^2, and V^5 schemes for HREN-approx algorithms (algo 1, algo 2, and algo 3) applied to AxBench benchmarks.The graph is normalized to Always-NTV scheme in algo 1.

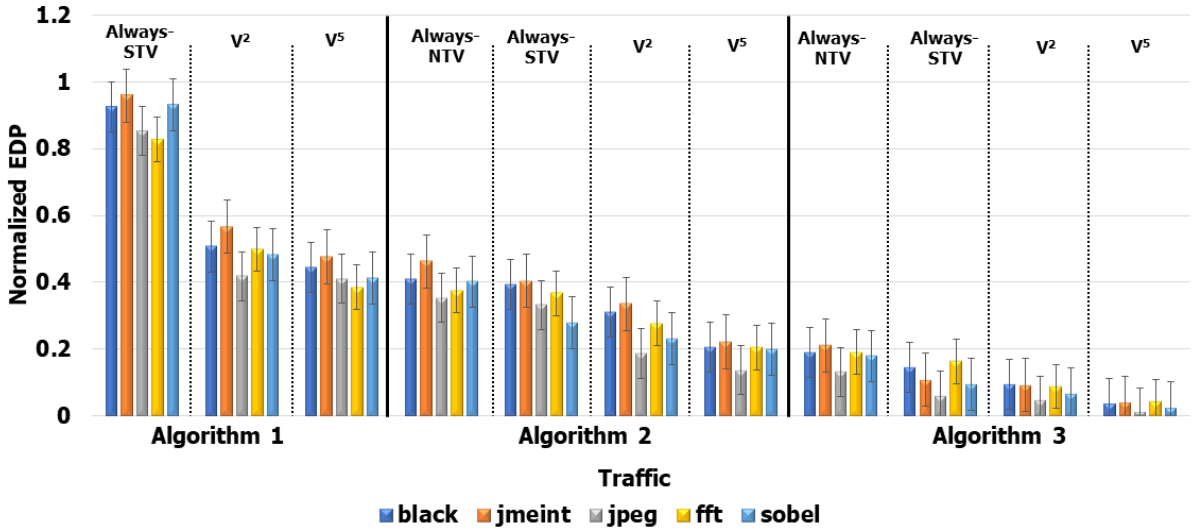


Fig. 12: Normalized Energy-Delay Product of Always-STV, Always-NTV, V^2, and V^5 schemes for HREN-approx algorithms (algo 1, algo 2, and algo 3) applied to AxBench benchmarks.The graph is normalized to Always-NTV scheme in algo 1.

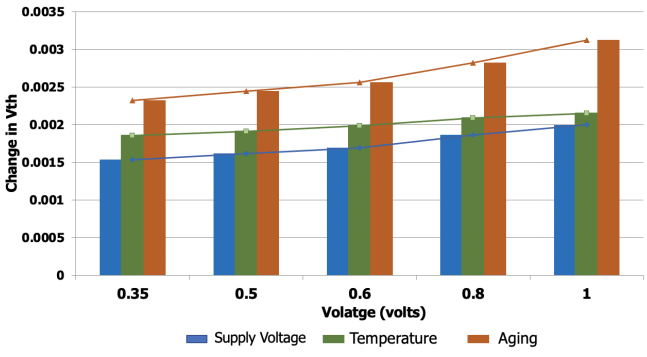


Fig. 13: Threshold voltage variation at five different supply voltages due to aging and temperature change.

five different applications of AxBench benchmark suit [28]. Our proposed evaluation scheme HREN showed approximately 59%, 47% and 27% decrease in dynamic energy consumption when compared to Always-STV, V^2 and V^5 scheme. Always-NTV scheme showed maximum energy savings among the five evaluation schemes as the supply voltage of NoC is NTV throughout the execution process whereas, the tradeoff is the increased error rate and packet latency of NoC.

Latency Analysis: Figure 16 shows the normalized average packet latency of NoC for the five evaluation schemes. HREN demonstrates approximately 34%, 53%, and 76% decrease in average packet latency of NoC when compared to V^5, V^2, and Always-NTV schemes. Among all the five evaluation schemes, Always-STV is expected to show minimum packet latency as NoC is operated in nominal voltage (highest voltage mode) throughout the execution process.

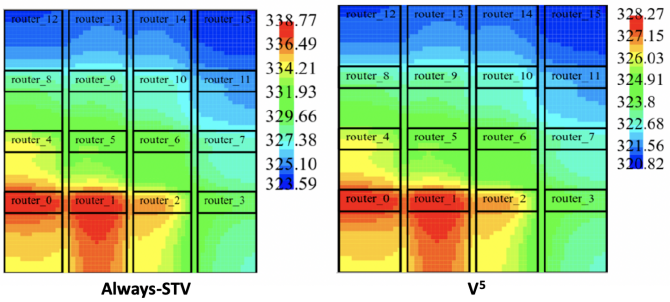


Fig. 14: Comparison of Hotspot thermal map of Always-STV under xy-routing (left) and V^5 scheme of HREN under adaptive routing (right). HREN showed uniform and lower device temperatures (slower aging) when compared to Always-STV under XY-routing.

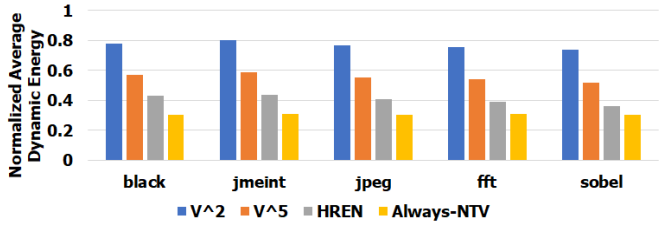


Fig. 15: Normalized average dynamic energy consumption of AxBench benchmark with the five evaluation schemes which is normalized to Always-STV scheme.

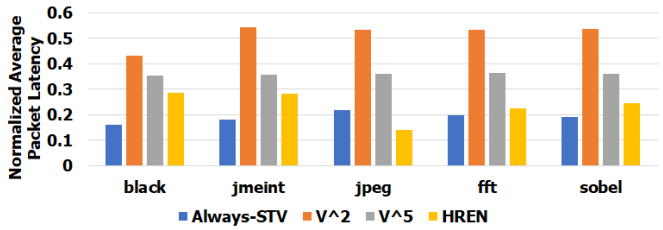


Fig. 16: Normalized average packet latency for AxBench benchmark with the five evaluation schemes which is normalized to Always-NTV scheme.

Energy-Delay Product Analysis: Figure 17 shows the normalized Energy-Delay Product (EDP) for the five evaluation schemes. HREN showed approximately 79% decrease in EDP when compared to Always-NTV scheme and 67% decrease in EDP when compared to Always-STV scheme. While comparing multiple voltage mode NoC architectures, HREN demonstrated nearly 50% decrease in EDP when compared to V^2 scheme and 41% decrease in EDP when compared to V^5 scheme. Our proposed design improved energy savings and decreased packet delay without compromising on the reliability of the network.

4.4 Category 4: Area Analysis

Table 2 represents the area overhead of the approximation algorithm, control unit and fault handling hardware of our proposed HREN architecture. The baseline model consists of buffer, crossbar, switch and other NoC components excluding the CRC, SECDED, and algorithms. The overall area

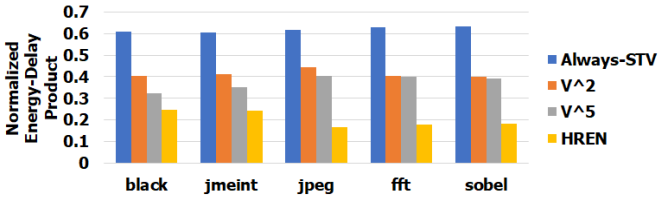


Fig. 17: Normalized Energy-Delay Product of AxBench benchmark with the five evaluation schemes which is normalized to Always-NTV scheme.

overhead of our proposed architecture, HREN, is approximately 8% of the overall chip area. We compared the area occupied by the default approximation algorithm and the HREN-approx which includes algo 1, algo 2, and algo 3 with the total area cost including NoC components. We observed that our data approximation (default approximation algorithm + HREN-approx) model occupied 3.5% of the total area, where as HREN-approx itself occupied 83% of the total area occupied by the data approximation model. The area overhead of the hybrid reliability model (CRC + SECDED) accounted for approximately 4.3% of the total area cost and the control unit used in our design occupied 0.037% of the total area cost whereas, the baseline NoC design occupied approximately 91.95% of the total HREN design.

TABLE 2: The percentage of area overhead of the approximation algorithm, control unit, fault handling hardware of our proposed HREN architecture, along with NoC components

Component	% of area occupied
Buffer (NoC)	63.85
Xbar (NoC)	26.17
Switch and Others (NoC)	1.93
Control unit	0.037
CRC	2.27
SECDED	2.12
Algorithm default	0.58
Algorithm 1	0.793
Algorithm 2	0.95
Algorithm 3	1.26

5 Prior Work

Voltage Scaling: As energy consumption and fault tolerance are the two critical factors influencing the communication cost, researchers have focused on improving the energy-efficiency and the reliability of NoC architecture. Prior work on improving energy-efficiency of NoC has focused on implementing techniques such as, approximate communication [16], Near Threshold Voltage (NTV) scaling [13], Dynamic Voltage and Frequency Scaling (DVFS) [5] [31], routing algorithms [32] [33], data encoding and decoding [18] [19] and power gating [11] [34] [12]. In [35], the authors proposed a dual voltage mode NoC architecture, where the supply voltage of NoC is switched between the nominal voltage and NTV depending on the traffic load. In their work, they showed that operating NoC in multiple voltage modes improves energy-efficiency of NoC. However, at

low supply voltage/operating frequency, faults and disturbances like Single Bit Upsets and Multiple Bit Upsets are observed. In [26], a five voltage-mode DVFS scheme with a dual-layered reliability model to improve energy-efficiency and reliability of NoC has been explored. The five-voltage scheme in our design improved energy efficiency of NoC by switching the supply voltage depending on the buffer utilization of NoC.

Data Approximation: Research on data approximation has introduced several techniques such as software modifications [16], computation related approximation [36], memory-based approximation [37] [38], which are inherently resilient to output errors. In [39], the authors proposed a reconfigurable NoC architecture to reduce the amount of data that is transferred between the source and destination by approximating the adder circuit. In a survey of data approximation [16], the authors demonstrated the influence of data approximation techniques on the energy efficiency and performance of an interconnection architecture. A recent work on approximate communication, such as [14], proposed data compression technique by identifying and encoding an approximately similar data in an application. In their work, they demonstrated that the communication cost of NoC depends on error threshold and data compression rate.

Reliability: Prior work on improving the reliability of complex hardware structures proposed error detecting and error correcting codes such as parity bits [40] , Cyclic Redundancy Checks (CRCs) [41], and Forward Error Correction (FEC) [42]. In [43], authors showed the impact of the variable error correcting methods of NoC. In their work, they demonstrated that End-to-End encoding scheme improves energy efficiency of NoC at low error rates by reducing the error correcting strength. On the other hand, switch-to-switch encoding scheme improves reliability of NoC by increasing the error correcting strength when the error rates are high. In [44], the authors proposed a configurable dual-layer error correction scheme that switches between two Error-correcting Codes (ECCs) depending on the error strength. Similar work on error correction [45] proposed an error correcting framework that dynamically detects and corrects the transient and permanent faults. As errors manifest by scaling down the supply voltage, in our previous publication [26], we proposed a dual-layer encoding scheme to handle Single Event Upsets (SEUs) due to voltage scaling and aging. However, there has been no prior work on evaluating the reliability-power trade-off of NTV scaling and data approximation of NoC.

In our work, we proposed a two-level approximation framework and variable voltage mode design to improve energy-efficiency and performance of NoC. Based on the buffer utilization, the supply voltage of NoC is varied globally. We proposed a hybrid reliability model that detects and corrects the errors that are introduced due to voltage scaling and data approximation.

6 Conclusions

In this paper, we propose a reliable and energy efficient NoC architecture while implementing approximate communication and NTV scaling techniques supported by a multi-layered reliability model. The V^5 scheme of HREN showed promising results while implementing voltage scaling (including NTV), adaptive routing, and approximate communication techniques in the design. Symmetrical distribution of traffic using dynamic adaptive routing algorithm showed balanced wear-out of links thus increasing the lifetime of NoC making it more reliable. We demonstrated that ΔV_{th} due to voltage scaling is less when compared to ΔV_{th} due to elevated temperature and aging effect in NoC. We then evaluated the combined effects of five voltage mode design with adaptive routing, which decreased NoC latency by 10-12 $\times$, and improved EDP by 1.3-7.5 $\times$ (including reliability) when compared to traditional NTV designs. The hybrid encoding scheme of HREN handles all the bit errors due to low supply voltage and aging, with a minimum area overhead of 2.79% in chip area (reliability design and control unit) and power cost of 6%. Results showed that the unified reliability model and the encoding scheme of HREN work together to improve NoC resiliency by tuning fault coverage. Approximate communication technique implemented in the design showed an additional power savings of 13%, while further reducing latency and EDP by 10% and 19% respectively. Overall, HREN demonstrated up to 2.8$\times$ dynamic energy savings while reducing latency up to 2$\times$. HREN simulation results showed an improvement of4x to 5.5$\times$ in Energy-Delay Product over the baseline model for AxBench approximation benchmark suit on a 4$\times$4 CMESH architecture.

References

[1] "Nvidia Xavier," Retrieved from: https://blogs.nvidia.com/blog/2016/09/28/xavier/.

[2] "Amd EPYC," Retrieved from: https://www.amd.com/en/products/epyc.

[3] "Quadro GV100," Retrieved from: https://blogs.nvidia.com/blog/2018/03/27/quadro-gv100-deep-learning-simulation/.

[4] "xilinx Everest," Retrieved from: https://www.anandtech.com/show/12509/xilinx-announces-project-everest-fpga-soc-hybrid.

[5] A. K. Mishra, R. Das, *et al.*, "A case for dynamic frequency tuning in on-chip networks," in *2009 42nd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2009.

[6] A. Bianco, P. Giaccone, and N. Li, "Exploiting dynamic voltage and frequency scaling in networks on chip," in *High Performance Switching and Routing (HPSR), 2012 IEEE 13th International Conference on*, pp. 229–234, IEEE, 2012.

[7] B. Lee, E. Nurvitadhi, R. Dixit, C. Yu, and M. Kim, "Dynamic voltage scaling techniques for power efficient video decoding," *Journal of Systems Architecture*, vol. 51, no. 10-11, pp. 633–652, 2005.

[8] J. Myers, A. Savanth, *et al.*, "A subthreshold arm cortex-m0+ subsystem in 65 nm cmos for wsn applications with 14 power domains, 10t sram, and integrated voltage regulator," in *IEEE Journal of Solid-State Circuits*, 2016.

[9] S. Khare and S. Jain, "Prospects of near-threshold voltage design for green computing," in *2013 26th International Conference on VLSI Design and 2013 12th International Conference on Embedded Systems*, 2013.

[10] S. Mittal, "A survey of architectural techniques for near-threshold computing," 2015.

[11] H. Bokhari, H. Javaid, M. Shafique, J. Henkel, and S. Parameswaran, "darknoc: Designing energy-efficient network-on-chip with multi-vt cells for dark silicon," in *Proceedings of the 51st Annual Design Automation Conference*, pp. 1–6, ACM, 2014.

[12] M. Casu, M. Yadav, and M. Zamboni, "Power-gating technique for network-on-chip buffers," *Electronics Letters*, vol. 49, no. 23, pp. 1438–1440, 2013.

[13] R. G. Dreslinski, M. Wieckowski, D. Blaauw, D. Sylvester, and T. Mudge, "Near-threshold computing: Reclaiming moore's law through energy efficient integrated circuits," *Proceedings of the IEEE*, vol. 98, no. 2, pp. 253–266, 2010.

[14] R. Boyapati, J. Huang, P. Majumder, K. H. Yum, and E. J. Kim, "Approx-noc: A data approximation framework for network-on-chip architectures," in *ACM SIGARCH Computer Architecture News*, vol. 45, pp. 666–677, ACM, 2017.

[15] J. S. Miguel, M. Badr, and N. E. Jerger, "Load value approximation," in *Proceedings of the 47th Annual IEEE/ACM International Symposium on Microarchitecture*, pp. 127–139, IEEE Computer Society, 2014.

[16] S. Mittal, "A survey of techniques for approximate computing," *ACM Computing Surveys (CSUR)*, vol. 48, no. 4, p. 62, 2016.

[17] F. Betzel, K. Khatamifard, H. Suresh, D. J. Lilja, J. Sartori, and U. Karpuzcu, "Approximate communication: Techniques for reducing communication bottlenecks in large-scale parallel systems," *ACM Computing Surveys (CSUR)*, vol. 51, no. 1, p. 1, 2018.

[18] M. Palesi, F. Fazzino, G. Ascia, and V. Catania, "Data encoding for low-power in wormhole-switched networks-on-chip," in *Digital System Design, Architectures, Methods and Tools, 2009. DSD'09. 12th Euromicro Conference on*, pp. 119–126, IEEE, 2009.

[19] N. Jafarzadeh, M. Palesi, A. Khademzadeh, and A. Afzali-Kusha, "Data encoding techniques for reducing energy consumption in network-on-chip," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 22, no. 3, pp. 675–685, 2014.

[20] R. Das, A. K. Mishra, C. Nicopoulos, D. Park, V. Narayanan, R. Iyer, M. S. Yousif, and C. R. Das, "Performance and power optimization through data compression in network-on-chip architectures," in *2008 IEEE 14th International Symposium on High Performance Computer Architecture*, pp. 215–225, IEEE, 2008.

[21] H. Kim, A. Vitkovskiy, P. V. Gratz, and V. Soteriou, "Use it or lose it: Wear-out and lifetime in future chip multiprocessors," in *Microarchitecture (MICRO), 2013 46th Annual IEEE/ACM International Symposium on*, pp. 136–147, IEEE, 2013.

[22] S. Ogawa and N. Shiono, "Generalized diffusion-reaction model for the low-field charge-buildup instability at the si-sio 2 interface," *Physical Review B*, vol. 51, no. 7, p. 4218, 1995.

[23] Y. Wang, S. Cotofana, and L. Fang, "A unified aging model of nbti and hci degradation towards lifetime reliability management for nanoscale mosfet circuits," in *2011 IEEE/ACM International Symposium on Nanoscale Architectures*, pp. 175–180, IEEE, 2011.

[24] D. Park, C. Nicopoulos, J. Kim, N. Vijaykrishnan, and C. R. Das, "Exploring fault-tolerant network-on-chip architectures," in *Dependable Systems and Networks, 2006. DSN 2006. International Conference on*, pp. 93–104, IEEE, 2006.

[25] S. R. Sridhara and N. R. Shanbhag, "Coding for system-on-chip networks: a unified framework," *IEEE transactions on very large scale integration (VLSI) systems*, vol. 13, no. 6, pp. 655–667, 2005.

[26] P. Bhamidipati and A. Karanth, "Retunes: Reliable and energy-efficient network-on-chip architecture," in *2018 IEEE 36th International Conference on Computer Design (ICCD)*, pp. 488–495, Oct 2018.

[27] K. Aisopos, C.-H. O. Chen, and L.-S. Peh, "Enabling system-level modeling of variation-induced faults in networks-on-chip," in *Proceedings of the 48th Design Automation Conference*, DAC '11, 2011.

[28] A. Yazdanbakhsh, D. Mahajan, H. Esmaeilzadeh, and P. Lotfi-Kamran, "Axbench: A multiplatform benchmark suite for approximate computing," *IEEE Design & Test*, vol. 34, no. 2, pp. 60–68, 2016.

[29] C. Sun, C.-H. O. Chen, G. Kurian, L. Wei, J. Miller, A. Agarwal, L.-S. Peh, and V. Stojanovic, "Dsent-a tool connecting emerging photonics with electronics for opto-electronic networks-on-chip modeling," in *2012 IEEE/ACM Sixth International Symposium on Networks-on-Chip*, pp. 201–210, IEEE, 2012.

[30] R. Zhang, M. R. Stan, and K. Skadron, "Hotspot 6.0: Validation, acceleration and extension," *University of Virginia, Tech. Rep*, 2015.

[31] S. Eyerman and L. Eeckhout, "Fine-grained dvfs using on-chip regulators," *ACM Transactions on Architecture and Code Optimization (TACO)*, vol. 8, no. 1, p. 1, 2011.

[32] K. Bhardwaj, K. Chakraborty, and S. Roy, "Towards graceful aging degradation in nocs through an adaptive routing algorithm," in *Proceedings of the 49th Annual Design Automation Conference*, pp. 382–391, ACM, 2012.

[33] K. Bhardwaj, K. Chakraborty, and S. Roy, "An MILP-based aging-aware routing algorithm for nocs," in *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2012*, pp. 326–331, IEEE, 2012.

[34] N. Nasirian, R. Soosahabi, and M. Bayoumi, "Traffic-aware power-gating scheme for network-on-chip routers," in *Circuits and Systems Conference (DCAS), 2016 IEEE Dallas*, pp. 1–4, IEEE, 2016.

[35] C. Rajamanikkam, R. JS, K. Chakraborty, and S. Roy, "Boostnoc: Power efficient network-on-chip architecture for near threshold computing," in *Proceedings of the 35th International Conference on Computer-Aided Design*, pp. 1–8, 2016.

[36] J. Huang, J. Lach, and G. Robins, "A methodology for energy-quality tradeoff using imprecise hardware," in *DAC Design Automation Conference 2012*, pp. 504–509, IEEE, 2012.

[37] Y. Fang, H. Li, and X. Li, "Softpcm: Enhancing energy efficiency and lifetime of phase change memory in video applications via approximate write," in *2012 IEEE 21st Asian Test Symposium*, pp. 131–136, IEEE, 2012.

[38] A. Shafiee, M. Taassori, R. Balasubramonian, and A. Davis, "Memzip: Exploring unconventional benefits from memory compression," in *2014 IEEE 20th International Symposium on High Performance Computer Architecture (HPCA)*, pp. 638–649, IEEE, 2014.

[39] C.-H. O. Chen, S. Park, T. Krishna, S. Subramanian, A. P. Chandrakasan, and L.-S. Peh, "Smart: A single-cycle reconfigurable noc for soc applications," in *2013 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 338–343, IEEE, 2013.

[40] K. Mohanram and N. A. Touba, "Cost-effective approach for reducing soft error failure rate in logic circuits," in *ITC*, vol. 1, pp. 893–901, 2003.

[41] É. Cota, A. de Morais Amory, and M. S. Lubaszewski, *Reliability, Availability and Serviceability of Networks-on-chip*. Springer Science & Business Media, 2011.

[42] P. P. Pande, A. Ganguly, B. Feero, B. Belzer, and C. Grecu, "Design of low power & reliable networks on chip through joint crosstalk avoidance and forward error correction coding," in *2006 21st IEEE International Symposium on Defect and Fault Tolerance in VLSI Systems*, pp. 466–476, IEEE, 2006.

[43] S. Murali, T. Theocharides, N. Vijaykrishnan, M. J. Irwin, L. Benini, and G. De Micheli, "Analysis of error recovery schemes for networks on chips," *IEEE Design & Test of Computers*, vol. 22, no. 5, pp. 434–442, 2005.

[44] Q. Yu and P. Ampadu, "A dual-layer method for transient and permanent error co-management in noc links," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 58, no. 1, pp. 36–40, 2010.

[45] T. Boraten and A. K. Kodi, "Runtime techniques to mitigate soft errors in network-on-chip (noc) architectures," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 37, no. 3, pp. 682–695, 2017.

Padmaja Bhamidipati Padmaja received her MS from Ohio University in 2018 and is currently working towards her PhD at University of Cincinnati. Her research interests include NoCs, SoCs, reliability, and security.

Avinash Karanth Avinash Karanth received the Ph.D. and M.S. degrees in electrical and computer engineering from the University of Arizona, Tucson, AZ in 2006 and 2003, respectively. He is currently the chair of the School of Electrical Engineering and Computer Science, and is the Joseph K. Jachinowski Professor in EECS at Ohio University, Athens, OH, USA. His current research interests include computer architecture, optical interconnects, machine learning, chip multiprocessors (CMPs), and networks-on-chip (NoCs). He was a recipient of the National Science Foundation CAREER Award in 2011, the Best Paper Award at the ICCD 2013 conference and his papers have been nominated for best paper at DATE 2019, NoCs 2010 and ASP-DAC in 2009. He is a senior member of IEEE.