RETUNES: Reliable and Energy-Efficient Network-on-Chip Architecture

Padmaja Bhamidipati  
School of Electrical Engineering and Computer Science  
Ohio university, Athens, Ohio 45701  
Email: pb261616@ohio.edu

Avinash Karanth  
School of Electrical Engineering and Computer Science  
Ohio University, Athens, Ohio 45701  
Email: karanth@ohio.edu

Abstract—As the number of cores integrated on the chip increases, the design of reliable and energy-efficient Network-on-Chip (NoC) to support the data movement needed by the multicores is becoming a major challenge. Reliability of NoC is affected by several aging affects such as Hot carrier injection (HCI) and Negative Bias Temperature Instability (NBTI) which vary the threshold voltage of the transistor, causing timing errors. Dynamic Frequency and Voltage Scaling (DVFS) along with Near Threshold Voltage (NTV) scaling allows the transistor to operate close to the threshold voltage, thereby aggressively minimizing dynamic power consumption by reducing voltage/frequency and minimizing the threshold voltage variation mitigating aging process. However, the trade-off is increased latency and reduced reliability due to lower voltage margins. In this paper, we propose a unified approach called RETUNES: Reliable and Energy-Efficient NoC where NTV scaling and reliability is both achieved while improving performance. RETUNES is a five-level voltage/frequency scaling scheme, which decreases power consumption and threshold voltage variation (ΔVth) during low network load with higher reliability and increases the network performance during high network load with reduced reliability. In order to even out the wear out and minimize the impact of aging in NoC, we propose adaptive routing algorithm in our design. RETUNES improves, total power savings by nearly 2.5 × and the energy-delay product (EDP) of the NoC by 3 × for Splash-2 and PARSEC benchmarks on a 4 × 4 concentrated mesh architecture.

I. INTRODUCTION

Technology scaling down combined with an exponential increase in the number of transistors that can be integrated at the on-chip level has resulted in the power density becoming one of the major challenges for multicore architectures. Network-on-Chips (NoCs) is the communication backbone that interconnects all the cores and the caches. Dynamic Voltage and Frequency Scaling (DVFS) has become one of the effective techniques to reduce power consumption and prior research has shown that 20-26% of dynamic power can be reduced [1]. Recently, NTV (Near Threshold Voltage) [2] scaling has been proposed which is an advanced power management technique that operates devices with a supply voltage (Vdd) near the transistor threshold voltage. Proper selection of Vdd close to transistor threshold voltage provides energy efficiency greater than 40% [3] [4]. Semiconductor industries such as ARM and Intel have focused on many applications of NTV scaling by varying device supply voltages.

Lowering operating voltage increases the susceptibility of devices to faults and compromises reliability. According to International Technology Roadmap for Semiconductors, reliability has become a major challenge as the lifetime of the transistor decreases with increase in temperature [5]. Fabrication methods, properties of the materials, and parameters such as supply voltage (Vdd) and temperature are the most common factors affecting the lifetime of the transistor degrading reliability [6]. As the transistor ages, variation in threshold voltage (Vth) increases due to interface traps and fluctuations in charge density. As the variation in ΔVth exceeds 10%, permanent faults are observed in the circuit. Scaling down the supply voltage of the device slows its aging process not only due to a decrease in temperature but also due to a decrease in the electric field. However, lowering operating voltage increases the susceptibility of the device to faults and in turn compromises reliability [7]. As NoCs are the only means for data transfer between the cores, faults, and disturbances like single bit upsets (SEUs), multiple bit upsets (MBUs), process variations and so on, have become a major challenge with NTV. To prevent serious reliability degradation, strong error correction techniques are critical to ensure that NoC operates reliably. As shown in [7], decreasing the supply voltage increases the bit error rate, and therefore, provides motivation to incorporate dynamic error handling techniques in NoC.

In this paper, we propose RETUNES: Reliable and Energy-efficient NoC, a unified model that includes, low-power design and fault tolerant techniques, to explore and analyze critical factors such as reliability, energy efficiency, and latency affecting NoC. In order to improve energy efficiency, we implement a voltage scaling methodology which includes NTV scaling. RETUNES uses five voltage modes which are carefully chosen such that the supply voltage regulation adapts to the incoming traffic to reduce congestion. We analyze the power-performance tradeoff of the network based on the buffer utilization and application traffic load. At low network load, voltage is scaled down ensuring maximum power savings and minimum ΔVth. At high network load, voltage is scaled up ensuring lower bit-error rate and minimum latency.

RETUNES enhances reliability by mitigating wearout at the inter-router level. We introduce an adaptive routing algorithm to even out the link wear-out by distributing the incoming traffic. The reliability model is a hybrid error correction and detection scheme with two-layered architecture to mitigate soft errors caused due to lower voltage modes and aging.
When operating the NoC in high voltage/frequency modes and at lower $\Delta V_{th}$, error rates are typically low, and therefore, we enable end-to-end (e2e) error correction scheme so that performance is not impacted. Similarly, when operating under low voltage modes (NTV) and at higher $\Delta V_{th}$, error rates are higher, and therefore, we enable switch-to-switch (s2s) error correction scheme so that reliability is improved. This multi-layered hybrid scheme handles SBUs and MBUs that are encountered at different supply voltages, thus achieving a fine balance between power consumption and network performance. The following are the major contributions of this paper:

- We introduce five voltage scaling design that includes NTV scaling to improve energy efficiency and effectively manage congestion in NoC while maintaining an upper bound on the energy-delay product (EDP).
- We propose an adaptive routing algorithm to proactively minimize the aging effect ($\Delta V_{th}$ range) due to unsymmetrical distribution of traffic. This improves the lifetime reliability of NoC architecture.
- RETUNES addresses both single and multi-bit errors caused due to voltage scaling (lower voltage/ frequency) and timing errors caused due to change in threshold voltage (due to aging, temperature and voltage scaling).

II. Prior Work

Dynamic Voltage and Frequency Scaling (DVFS) is an effective technique to reduce the dynamic power consumption of NoC. Typically DVFS designers have two important decisions to make at what circuit level (granularity) to apply DVFS and what is the appropriate voltage mode to apply. In fine-grain voltage/frequency domains, NoC routers and links operate independent of each other using multiple supply voltages to improve performance. However, since the current is drawn from different domains, the supply voltage guard-bands increases which results in a decrease in power-efficiency in NTV environment. In [3] author discussed this problem for IBM processor due to fine-grain domain approach. Instead of fine-grain voltage modes, coarse-grain on-chip multiple voltage mode approach in which NoC is controlled globally increases the power efficiency of the network, albeit at a cost of performance. The other crucial aspect of DVFS is the voltage mode selection. Operating voltage of the network can be scaled down to lowest mode if the buffer utilization/traffic within the NoC is low and similarly, supply voltage can be increased if the buffer utilization is high. Prior work has proposed several different metrics to measure traffic, such as buffer usage [9], predicted link usage algorithm [10], threshold-controlled algorithm [11] and DVFS algorithm [12].

Aging process in transistors cannot be stopped, however, controlled aging (voltage scaling) and symmetrical distribution on workload would increase the lifetime of the device. In [13] authors used aging-aware oblivious routing algorithm that dynamically chooses the routing scheme to distribute traffic symmetrically along NoC. However, voltage scaling technique to reduce the threshold voltage variation is not considered. In [14] authors proposed a design that interprets aging degradation by comparing runtime-delay increase (threshold voltage change) with the offline delay analysis. In the design, authors showed that lower supply voltage slows threshold voltage variation due to aging. Unfortunately, the need for offline calculation is the drawback of the design which loses the ability to dynamically tune design according to the wear-out levels.

NTV scaling has been proposed for processors, cores and memory and more recently applied to NoCs. In [15] the authors proposed a multi-layered NoC architecture that uses near threshold voltage technique which improved the energy efficiency of the NoC. In their work it was shown that switching the operating voltage based on application demand between NTV and normal voltage improves the performance of the NoC. However, operating NoC at lower voltage will increase the susceptibility of devices to faults. Prior work on error recovery schemes have showed the impact of encoding techniques on reliability of the network at low voltages [7]. In those experiments it was demonstrated that SBUs, MBUs and hard errors have a higher probability to occur at low supply voltage. Fortunately, error correction and detection techniques can be applied to handle soft errors. A 2-layered error management method for NoC to manage both permanent and transient error was proposed in [16]. ECC (Error correcting code) techniques like s2s and e2e error control mechanism is integrated in the data link layer, physical layer and network layer depending on the noise conditions of the NoC. However, none of the prior work have combined voltage scaling, reliability and aging for NoCs architecture together.

In what follows, we will describe the RETUNES architecture, voltage scaling (different modes), adaptive routing (aging) and reliability (different error correcting codes) model.

III. RETUNES Architecture

RETUNES is a multi-layered architecture which uniquely combines, hybrid reliability-layer and energy-efficiency layer evolving into an integrated NoC architecture. The energy efficient design of RETUNES uses voltage scaling which includes five voltage modes which are carefully chosen to ensure maximum energy savings. The most crucial design aspect is voltage mode selection and voltage mode switching. We explain our proposed voltage mode design, mode selection, and mode switching methods in subsection A. The hybrid reliability model of RETUNES analyses and handles faults observed due to transistor aging and voltage scaling which are explained in subsections B and C. In order to reduce aging effect and increase life-time of NoC, we effectively implement adaptive routing algorithm which is explained in subsection B.

A. Proposed Voltage Mode Selection

RETUNES is a reliable and energy efficient NoC design which is evaluated on a $4 \times 4$ concentrated mesh-topology.
with 64 cores. Figure 1 shows the centralized control unit and on-chip linear voltage regulator for our concentrated 4 × 4 2D concentrated mesh topology. The centralized control unit decides the voltage mode to operate NoC which is called the Mode Control Unit (MCU). The error correcting code to be used for the NoC is called the Layer Control Unit (LCU). While it is possible for switching decisions to be made locally on a per-router basis, we perform mode control and layer control to the entire NoC globally to reduce the cost and complexity of the controller. Recently voltage regulators are integrated on the same chip, to minimize the power and area cost of off-chip voltage regulators [17]. The linear on-chip voltage regulator used in our design is shown in Figure 1 which changes the voltage at a rate of 30 mV/ns with a minimum of 5% power loss. All the routers are instructed to wait an additional cycle to settle in the new voltage mode to prevent communication loss. Based on the buffer utilization, RETUNES selects the appropriate voltage mode for operation. The lowest voltage mode is the most power efficient mode while the highest voltage mode boosts the performance of the network with minimum latency. The voltage-frequency pairs used in our design as shown in Table I is similar to the previous works [15] [18].

**Buffer occupancy:** In RETUNES, we closely monitor buffer utilization, which serves as a metric for switching different voltage modes as shown in Figure 2. At the maximum buffer utilization level, operating NoC at nominal voltage (Super Threshold Voltage (STV)) is most effective. At the minimum buffer utilization level, operating NoC at lower voltages (NTV) is more power efficient. Since buffer utilization varies significantly for various applications during simulation, we carefully chose five voltage modes (NTV, V1, V2, V3, STV) to maximize power savings while providing sufficient opportunities to improve performance. We designed buffer utilization range for all the voltage modes to avoid congestion at lower voltage modes. When buffer utilization is below 5%, NTV mode is activated. Similarly, between 15% to 40% V1 mode is activated, 40% to 58-60% V2 mode is activated, 58-60% to 70% V3 mode is activated and above 70% STV mode is activated. These empirical values were determined by running several of the Splash-2 and PARSEC suite benchmarks. The buffer utilization from all of the router is collected for each epoch. The total buffer utilization is used to decide the voltage mode of the NoC. Table I shows the voltage modes in RETUNES and corresponding frequencies. NoC in its lowest voltage mode which is NTV (0.35V) is more energy efficient.

**Switching cost:** Switching cost is the overhead delay to switch between the voltage modes. Table I shows the overhead delays, temperature, and respective voltage modes in our proposed design. In order to calculate temperature for our five voltage modes, we modeled average link utilization to vary from 0.01 to 0.4 flits/cycle as shown in Table I. We assumed that the network saturates at 0.4 flits/cycle. The temperature values obtained at every voltage mode or network load is used to calculate the overhead delay of the network using a bias generator which is used to calculate wake-up latency [19]. The first input to find the overhead delay is the buffer utilization. The range of buffer utilization determines the supply voltage. Our next input parameter is link-utilization, where an increase in link utilization indicates that the traffic intensity has increased which in turn leads to increased power consumption thus raising the temperature. The link utilization, temperature, overhead delay, and corresponding voltage modes are shown in Table I. Routers operating in a voltage mode during step down, quickly ramp down the frequency and then waits for an overhead delay before ramping down the supply voltage. For a voltage step-up, we first ramp up the voltage then wait for the overhead delay and then ramp up the frequency. On receiving a signal to change the voltage mode of NoC, all the routers are instructed to complete the buffer transfers before ramping up/down the voltage, to prevent any loss in communication. In RETUNES, all voltage/frequency changes affect all links and routers simultaneously and the entire NoC operates at the same voltage/frequency (coarse-grain DVFS).

### Table I: RETUNES multiple voltage domains.

<table>
<thead>
<tr>
<th>Mode</th>
<th>Unit Volt</th>
<th>Freq Hz</th>
<th>Net Load</th>
<th>Temp Range</th>
<th>Delay cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTV</td>
<td>0.35</td>
<td>0.7</td>
<td>0.01</td>
<td>76-82</td>
<td>8</td>
</tr>
<tr>
<td>V1</td>
<td>0.55</td>
<td>0.8</td>
<td>0.1</td>
<td>70-77</td>
<td>5</td>
</tr>
<tr>
<td>V2</td>
<td>0.6</td>
<td>1.5</td>
<td>0.2</td>
<td>80-93</td>
<td>4</td>
</tr>
<tr>
<td>V3</td>
<td>0.8</td>
<td>2</td>
<td>0.3</td>
<td>90-101</td>
<td>2</td>
</tr>
<tr>
<td>STV</td>
<td>1</td>
<td>2.3</td>
<td>0.4</td>
<td>97-104</td>
<td>1</td>
</tr>
</tbody>
</table>

**B. Improving NoC Lifetime Reliability**

Lifetime of the device is a measure of wear-out (aging) experienced by that device over a period of time. In order to improve the lifetime of the transistor, we employ symmetrical distribution of stress for all the links in NoC. The symmetrical distribution of incoming traffic minimizes the aging process without sacrificing on performance (latency). In this section, we determine the aging process, which is the measure of...
threshold voltage change over different voltage modes used in the design.

Adaptive Packet Routing: In our design, we constantly monitor runtime link utilization to understand the stress levels of the link. As the stress level of the link increases, its temperature elevates, aging increases and eventually lifetime decreases. The unsymmetrical distribution of traffic leads to uneven aging which in turn degrades the reliability of the network. To minimize the unbalanced utilization, we use adaptive routing algorithm and reroute packets. At every router, the routing algorithm compares the link utilization for x, -x, y and -y-directions. The packet is adaptively routed using the least utilized link from the available links (directions) in the router. Our design effectively adapts to the runtime changes and makes in-flight routing decisions eliminating offline calculations and lookup tables to improve performance (reduce latency) of the NoC.

Determining Aging: We showed that lowering the supply voltage slows down the aging process, due to reduced temperatures and slower threshold voltage variations. Thus, our design is proved to be reliable even at lower voltage modes, where the soft errors discovered due to voltage scaling [7](lower voltages in this case) are handled by the fault model (discussed in subsection C). We calculated aging process as the threshold voltage variations over 10 years using Synopsys HSPICE tool. We modeled aging effect caused due to HCI and NBTI using Predictive Technology Models (PTM) [20] for a 45nm transistor.

C. Proposed Reliability Model

Reliability model: Using unified fault model as shown in [3] we determine the fault rate at every voltage mode and \( \Delta V_{th} \) range. We gathered raw data from our network simulator to calculate \( \Delta V_{th} \) due to voltage scaling. HotSpot thermal model [21] and router fault model [22] are used to calculate \( \Delta V_{th} \) due to temperature variations in NoC and Synopsys HSPICE tool is used to calculate \( \Delta V_{th} \) due to aging. The link utilization is correlated to temperature and transistor delay calculated (shown in subsection A) to determine the wear-out. The average dynamic power from our network simulator is provided as an input file to HotSpot thermal model and router fault model to calculate temperature. The temperature from HotSpot thermal model and router fault model is passed to Synopsys HSPICE to calculate wear-out. Reliability degradation is measured as the sum of all threshold voltage variations observed in the design.

Reliability degradation (Rd) : \( \Delta V_{th\ temp} + \Delta V_{th\ aging} + \Delta V_{th\ voltage} \)

The threshold voltage variation range (\( \Delta V_{th} \) range) is divided into three levels along with the error types (More errors (Me), Few errors (Fe) and No errors (Ne)) as shown in [3] We assume that, any variation of threshold voltage greater than 10% variation is considered to be a permanent fault. As the \( \Delta V_{th} \) range increases, the error type shifts to its higher error type (Ne is the lowest and Me is the highest) depending on the change in threshold voltage. Similarly, the fault model shown in [3] keeps track of the error type for all the voltage modes used in the design. Our reliability model checks the \( \Delta V_{th} \) range for every supply voltage change in fault model. The error rate at every \( \Delta V_{th} \) range is integrated with the fault model to generate unified fault model for the voltage mode active in NoC as shown in [3]. The unified reliability model determines error type, under each \( \Delta V_{th} \) range and the voltage mode (NTV, V1, V2, V3, STV) active at that instant of time. Our ECC scheme detects all two-bit errors and corrects all single bit errors obtained from the unified reliability model at each epoch. If a fault is detected and cannot be corrected, the flit is dropped and a request for retransmission is sent. When the number of faults increases in a packet, the entire packet is dropped and will be retransmitted to prevent the communication loss.

Encoding framework: In order to improve network performance towards varying faults, an adaptable error handling design is found to be more effective. When the probability of bit error is high, fault coverage should be increased to ensure reliability and fault coverage is reduced when the probability of error is low. To improve resiliency, we propose a two-layer encoding framework based on the voltage mode of the NoC. Initially, all the routers use CRC-32 at the lowest \( \Delta V_{th} \) range independent of the network voltage mode. When the NoC is in NTV or V1, s2s encoding layer is activated. In this layer,
every router uses strong ECCs to increase fault coverage for the input traffic. Upstream routers use counters to access the information of the number of bit errors that are detected in the downstream routers. Routers in STV, V3 and V2 modes employ weak ECC as the probability of bit error is low. Triple modular redundancy (TMR) control lines (layer control unit) are used to signal routers to switch between encoding schemes. 

**Fault model architecture:** Figures 4 and 5 show the proposed micro-architecture for e2e and s2s encoding layers. The network interface of the core connects the processing element to the network and each core is connected to the grid in a mesh topology. In our proposed design as shown in Figure 5, packets use CRC-32 for encoding before it enters the network. Our e2e encoding layer has 256-bit CRC-32 encoded packet with 224 data bits and 32-bit long check value. The routers with 3 pipeline stages would add stall cycles and increase flit delay at lower voltage/frequency modes (NTV, V1, V2). In order to decrease overall packet latency, the routers in s2s encoding layer consist of five pipeline stages as shown in the Figure 4. The proposed encoder and decoder design are implemented using hamming H(72,64) which is a single error correction and double error detection (SECDED) code. With SECDED hamming codes, all 1-bit errors are recovered, and 2-bit errors are detected. The codeword vector and the transpose of the generator matrix are multiplied to detect errors in the syndrome generator. The decoder design forwards data bits and parity bits together to the encoder design in the upstream router. The syndrome compares the forwarded parity bits with the new parity bits in the encoder design to detect faults. However, faults cannot be corrected in the encoder as the correcting hardware is not present in the encoder design.

**D. Voltage Mode and Error Correction**

We contemplate four stages of operation for our reliable and energy efficient RETUNES architecture as follows:

**Step 1:** Initially all the packets are encoded with CRC-32. The buffer utilization of the active layer is constantly monitored, which serves as the information to the mode change control unit. For every epoch the mode control unit updates the network utilization level. We observed traffic information from the buffer utilization trend for Splash-2 [23] and PARSEC [24] benchmarks and carefully choose the epoch size to be 100 cycles to avoid power loss due to frequent voltage mode switching.

**Step 2:** Once the voltage mode is switched, information regarding the current and previous voltage modes are passed to the encoding layer control unit and the voltage mode control unit. The mode control unit senses the local changes in the network and adjusts output voltage of the voltage regulator, which is used as supply voltage to NoC. The corresponding overhead delay and frequency is applied to NoC according to the mode control algorithm as shown in the figure. At the same time all the routers are instructed to complete the in-flight flit transmission to avoid data loss before switching the voltage modes.

**Step 3:** The encoding layer control unit plays a crucial role in decision making to switch between the encoding layers e2e or s2s. Current voltage mode of the network and the probability of error obtained from unified fault model are passed to layer control unit, where decision is made to upgrade or downgrade...
ECC mode. Our strong ECC uses SECDED hamming code at every switch to correct all 1-bit errors. If the error cannot be corrected, a request for retransmission is sent to the source router. The counters at each router constantly keeps track of the fault rate.

**Step 4:** On receiving the retransmission signal, the source router resends the requested flit to the destination router. Data is stored in the retransmission buffers until an acknowledgement (ACK) is received. Data and ACK lines are assumed to be separate, where fault coverage is not considered for the retransmitted data. Once the retransmission is successful, acknowledgement is sent to the source router to terminate the process.

### IV. PERFORMANCE EVALUATION

In this section, we evaluate our proposed reliable and power efficient scheme and discuss the results obtained from simulation.

#### A. Simulation setup:

In this section, we describe the performance of RETUNES architecture. RETUNES is evaluated on a 4 × 4 concentrated mesh-topology with 64 cores and unidirectional links. Each router has 4 VCs for every input port and 4 buffer slots per VC. Each packet has 256 bits and it is split into 4 equal 64-bit flits before injecting into the network. We used real traffic traces on our network using Splash-2 and PARSEC workloads. The power and area cost of the network is obtained from the Synopsys design compiler tool using the TSMC 45 nm technology libraries [25] and DSENT NoC modeling tool [26] using a 45 nm technology library and the power and area estimates for the reliability design was obtained from the Synopsys design compiler tool using the TSMC 45 nm technology libraries. To compare the performance of our design, we used 4 evaluation schemes, V2, Always STV (XY), Always STV (ADP), and Always NTV (XY) as follows:

**STV:** In STV/Always STV scheme the NoC is operated in nominal voltage mode, where, power consumption is maximum. However, Noc shows best performance with high application speedup in always STV scheme. We also compared always STV scheme under Dimensional Routing Algorithm (DOR)-XY routing (Always STV (XY)) and Adaptive Routing Algorithm (Always STV (ADP)).

**NTV:** In NTV/Always NTV scheme NoC and its cores are operated under lower voltages (near the threshold voltage). This scheme shows high energy efficiency at the cost of latency. The NoC and its cores suffer from performance loss due to increase in the error rate in always NTV scheme. However, this low-voltage scheme (Always NTV) have less impact on $\Delta V_{th}$ providing, lower error-probabilities when compared to other schemes. Always NTV scheme under DOR-XY routing (Always NTV(XY)) is considered as the baseline model for our design.

**V2 (2 voltage scheme):** In V2 scheme we switched the operating voltage applied to NoC and its core between 2 voltages (STV and NTV). In this scheme we operate NoC under NTV mode for 25-30% buffer utilization depending on the traffic to avoid congestion. STV mode is operated for buffer utilization higher than 25-30% of buffer utilization.

**V5 (5 voltage scheme):** V5 scheme is our proposed 5-level voltage scaling design. We provide energy efficient design that includes NTV, V1, V2, V3 and STV schemes, which are switched depending on the communication demand. The V5 scheme provides better performance by sacrificing chip area. The power consumption of the network is low when it uses our baseline model, whereas the bit error rate is low when it uses STV scheme. As discussed earlier, the bit error rate is also lower in always NTV scheme due to lower $\Delta V_{th}$ when compared to STV scheme. So, we determined the overall error rate with the unified reliability model depending on the $\Delta V_{th}$ calculation and voltage mode (fault model) as explained in section III(c). The NoC optimizes energy consumption when it uses NTV scheme and has lower latency when operated in STV mode.

#### B. Reliability analysis:

SECDED can prevent retransmissions by correcting all single bit errors. However, with increase in number of errors, full retransmission is needed. Our hybrid encoding scheme consumes 6% power to improve NoC resiliency by tuning fault coverage. We used the fault model to generate bit errors for different PARSEC and Splash-2 applications. On an average, we determined that the mean bit error rate for the V5 is 0.45 when compared to Always NTV scheme and 2.5 when compared to Always STV scheme. As expected, error rate is low at STV, whereas the error rate is high at NTV. The mean error rate due to voltage scaling occupies nearly 29% of the overall error rate (voltage scaling and aging) and mean error rate due to aging occupies approximately 71% of the overall error rate observed in NoC. RETUNES offers an error rate in-between STV and NTV, thereby balancing reliability and power consumption. This error rate was used to generate soft errors in the NoC. If s2s was enabled, and single but error is observed, the error was corrected, otherwise it retransmission was requested. If e2e was enabled, then the request was sent from the destination back to the source for retransmission. In this paper we deal with only soft errors (SBUs, MBUs). Permanent faults and faults during retransmission are not considered in this work.

#### C. Results:

**NoC lifetime analysis:** In this section we analyzed lifetime of a transistor over 10 years using Synopsys HSPICE under Predictive Technology Model (PTM). Figure 6 shows the threshold voltage change for 5 voltage modes (NTV, V1, V2, V3, STV) over 10 years. According to the results shown in 6, we can claim that the threshold voltage variations ($\Delta V_{th}$) decreases with the operating voltage, mitigating the aging process. Our design experience, lower stress levels due to HCI/NBTI effect at lower voltage modes and less vulnerable to bit errors due to voltage scaling (higher supply voltage in this
Fig. 6: Threshold voltage change ($\Delta V_{th}$) due to voltage scaling, elevated temperature, and aging at 5 different supply voltages. Lower is better.

Fig. 7: Total dynamic power cost for Splash-2 and PARSEC benchmarks of 64 core NoC when operated in 5 proposed schemes. Lower is better.

NoC power analysis: Figure 7 shows the dynamic power consumed (mW) by NoC in all the 5 simulation schemes (section IV - evaluation schemes) to transmit packets from source to destination. Therefore, the power to retransmit when multiple bits are in error is also taken into account. With unified reliability model, on an average across multiple applications, V5 scheme show 60-61% savings in power when compared to always-STV. And 23% savings when compared to V2 scheme. Since, always-NTV operates in the lowest mode irrespective of network load, always-NTV consumes the least power among the three.

NoC average delay analysis: In this section we analyzed the average latency of the NoC for different applications when operated in 5 proposed schemes of RETUNES. Figure 8 explicitly shows the breakdown when no reliability model is considered (blue) and when reliability costs are included (orange). With no reliability model the average packet latency is, approximately 1.2× more in V2 scheme and 10.8× more in baseline model, when compared to V5 scheme. The operating frequency of the network for always-NTV scheme is low throughout the application runtime which increases the overall packet latency. Similarly, the average packet latency is 1.6× more in V5 when compared to always-STV scheme. Overall when reliability delays are included as well, the average packet latency is 1.35× more in V2 and approximately 12× more in baseline model, when compared to V5. Most often, the retransmission delays vary at the speed of the operating mode and therefore, the results reflect the delay cost accordingly.

Energy-delay analysis: To provide meaningful insight, we combine the energy and delay into a single plot to analyze the advantage among the 5 schemes. When analyzing the energy-delay product (EDP), lower is considered to be better. Figure 9 shows our EDP plot. NoC in always-STV scheme shows decrease in packet latency and increase in overall power consumption, whereas always-NTV scheme shows decrease in overall power consumption and increase in packet latency. Our V5 scheme is expected to balance power consumption and packet latency. We see 7.5×, 2×, 1.6×, and 1.3× increase in the performance of the NoC (under V5) when compared to baseline model, Always-STV(XY), Always-STV (ADP), and V2 schemes respectively along with the reliability cost.

Area overhead: The area overhead of the fault handling hardware is very crucial to design an efficient model. The area efficient fault models would decrease the overall area of NoC.
In [10] we compare the area cost for our e2e and s2s encoding layers. We carefully model the decoder, encoder, control unit, switch, router buffers and the crossbar. We included the CRC-32 encoder area cost in the NoC area overhead even though e2e layer is implemented only at the NoC interface. Our reliability design occupies 2.4% chip area, whereas control unit consumes 0.39% of overall chip area.

V. CONCLUSIONS

In this paper we showed power savings of nearly 2.5 × by carefully choosing the appropriate voltage mode (including NTV scaling) for the varying traffic in NoC. 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 showed that the $\Delta V_{th}$ due to voltage scaling (down) is less when compared to elevated temperature and aging effect in NoC. We then evaluated the combined effects of five voltage design with adaptive routing, which decreased the NoC latency by 10-12× when compared to traditional NTV designs and improved EDP by 1.3-7.5× (including reliability). We also observed that the error rate of the NoC increases as the operating voltage of NoC decreases. Our hybrid encoding scheme handles all the bit errors due to lower supply voltage and aging, with minimum area overhead of 2.79% of chip area (reliability design and control unit) and power cost of 6% to improve NoC resiliency by tuning fault coverage.

ACKNOWLEDGMENT

This work was partially supported by National Science Foundation (NSF) grants CCF-1420718, CCF-1513606 and CCF-1703013.

REFERENCES

[17] J. GJANCZ, “On-chip voltage regulation for power management in system-on-chip,” in THESIS, Submitted as partial fulfillment of the requirements For the degree of Master of Science in Electrical and Computer Engineering In the Graduate College of the University of Illinois at Chicago, 2008.