Survey of Power Optimization Techniques for Network on Chip

Mohammed Waseem Khanooni[1], S. D. Chede[2]

Principal[2]

Department of ECE

Priyadarshini College of Engineering, Nagpur, India

Suryodaya College of Engineering, Nagpur, India

Corresponding Authors' Emails: *mwkhanooni@gmail.com[1], santoshchede@rediffmail.com[2]*

Abstract

Previously, research and design of Network-on-Chip paradigms where mainly focused on improving the performance of the interconnection networks. With emerging wide range of low-power applications and energy constrained high-performance applications, it is highly desirable to have NoCs that are highly energy efficient without incurring performance penalty. In the design of high-performance massive multi-core chips, power and heat have become dominant constraints. Increased power consumption can raise chip temperature, which in turn can decrease chip reliability and performance and increase cooling costs. Dynamic Voltage Scaling is an efficient technique for significant power savings in microprocessors. It has been proposed and deployed in modern microprocessors by exploiting the variance in processor utilization. On a Network-on-Chip paradigm, it is more likely that the wire line links and buffers are not always fully utilized even for different applications. Hence, by exploiting these characteristics of the links and buffers over different traffic, DVFS technique can be incorporated on these switches and wire line links for huge power savings

Keywords: *Embedded Multicore/Many-core Systems-on-Chip; Network-on-chip; Low-power design*

I. INTRODUCTION

The Network-on-Chip (NoC) paradigm has evolved to replace ad-hoc global wiring interconnects [7, 13, 14]. With this approach, system modules communicate by sending packets to one another over a network. The structured NoC wiring allows for the use of high-performance circuits to reduce latency and increase bandwidth [7, 13, 14].A conventional NoC consists of a packet- switched network with a two-dimensional mesh topology [7, 36]. NoCs typically employ wormhole routing, i.e., each packet is divided into smaller units called flits, which are forwarded individually on links [19, 36]. NoC routers typically employ multiple buffers, called virtual channels (VCs) [26, 36, 44], which allows them to transmit several flows in parallel by inter- leaving their flits on a single outgoing link. Currently proposed NoCs employ between two and four VCs [19, 26], but studies argue that this number should increase in future NoCs in order to supply higher throughput demands [36].

Power consumption is becoming a crucial factor in the design of high-speed digital systems, [1, 4, 5, 8, 12, 20, 21, 38, 44]. Whereas static power consumption is due to leakage and short-circuit currents, dynamic power consumption stems from switching activity, i.e., bit transitions. Interconnects consume the lion's share of dynamic power in modern chips. For example, studies show that interconnect links consume up to 60% of the dynamic power in NoCs [1, 42], more than 60% of the dynamic power in a modern microprocessor [21], and more than 90% in FPGA [15]. This portion is apparently growing [1, 9, 12, 33, 38, 44].

Modern device scaling results in deep sub micron noises, which cause inter- connect errors to be more dominant and harder to predict [10, 25, 27, 28, 29, 30, 41, 45], and also gives rise to new error sources [25, 27].

Traditional designs enhance interconnect reliability at the physical layer, using worst-case design margins such as aggressive inter-wire spacing, insertion of repeaters, and shielding of link wires [29, 32, 41]. Unfortunately, all these techniques incur high area and power costs [27, 40]. Moreover, they require knowledge of the circuit layout, thus inflicting design complexity [27, 30].

Furthermore, in novel technologies, the efficiency of these techniques decreases because transient errors are becoming

harder to predict [41]. A promising alternative to the traditional physical layer solutions is to add reliability at the data-link layer of the NoC, using error detection codes, as suggested in [30, 45]. Whereas error protection at the physical layer involves circuit design techniques that rely on specific device parameters, data link solutions are technology-independent [30].

In this thesis, we present a technique to save redundant bit transmissions resulting from error detection codes, along with the associated power penalty. System-level power design approaches include synthesis algorithms to in- crease the power efficiency in interconnection networks via better module placement [15] or improved application design [23]. In such methods, the traffic patterns among the cores need to be known a-priori. In contrast, the approach we present in this thesis does not require a-priori knowledge of the interconnect usage.

Embedded power design approaches include techniques for energy efficient micro architecture. For example, in [42] a power-driven design of router for NoC is presented. The technique we present in this thesis can complement this approach, and combining both schemes can help to further reduce power.

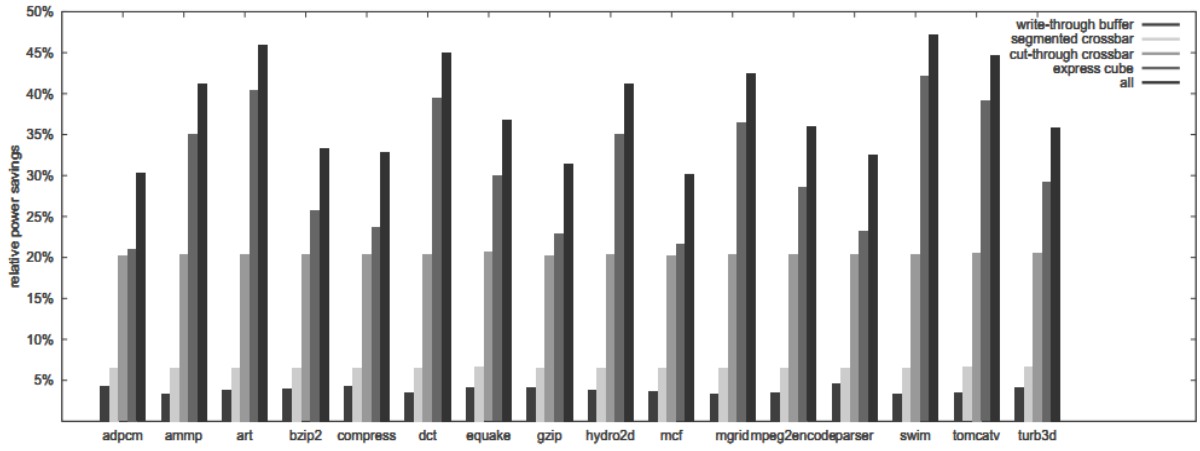

Figure 1

Data encoding is often employed to decrease the number of bit transitions over interconnects. Popular methods include Bus Invert (BI) [38], adaptive coding [16], gray coding [24] and transition method [35]. Of these, we elab- orate only on BI, which was shown to be the most effective in NoCs [31]. BI compares the data to be transmitted with the current data on link. If the Hamming distance (the number of bits in which the data patterns differ) between the new information and the link state is larger than half the number of bits (wires) on the link, then the data pattern is inverted before trans- mission.

To enable restoring the original data pattern, an extra control wire is added to the link, in which a transmission of 1 indicates data inversion. Analysis [37] shows that on link widths of more than 8 bits, the savings are insufficient to justify the overhead of encoding circuits, and therefore wider links are segmented.

Previous work [31] has investigated the reduction of NoC power consumption achieved using the four mentioned data encoding schemes. Experiments in 0:35u technology showed that BI achieves the best results. We therefore compare our technique to and combine our technique with BI in this thesis. Nevertheless, the same paper found that the achieved power gain is offset by the overhead required to implement the BI encoding scheme. In contrast, the power savings achieved by our technique are higher than the power consumed by the required overhead. The reasons for that are that our technique does not add any redundant control wires.

Power may also be reduced using low-power device and circuit design tech- niques, such as dynamic voltage and frequency scaling (DVFS) [17], which adjust the supply voltage and clock rate dynamically according to circuit parameters. The energy efficiency of DVFS is highly dependent on the slack of the circuit. Another approach uses low-swing signaling techniques [13, 18, 43], the efficiency of which depends on circuit layout and manufacturing parameters. In contrast, the approach we present in this thesis does not require knowledge about the circuit layout or manufacturing parameters.

II. LITERATURE REVIEW

A. Power-driven Design of Router

As demand for bandwidth increases in systems-on-a-chip and chip multiprocessors,

networks are fast replacing buses and dedicated wires as the pervasive interconnect fabric for on-chip communication. The tight delay requirements faced by on-chip networks have resulted in prior micro architectures being largely performance-driven. While performance is a critical metric, on-chip networks are also extremely power-constrained. In this paper , we investigate on-chip network micro architectures from a power-driven perspective. We first analyze the power dissipation of existing net- work micro architectures, highlighting insights that prompt us to devise several power-efficient network micro architectures: segmented crossbar , cut-through crossbar and write- through buffer . We also study and uncover the power saving potential of an existing network architecture: express cube. These techniques are evaluated with synthetic as well as real chip multiprocessor traces, showing a reduction in network power of up to 44.9%, along with no degradation in network performance, and even improved latency- throughput in some cases.

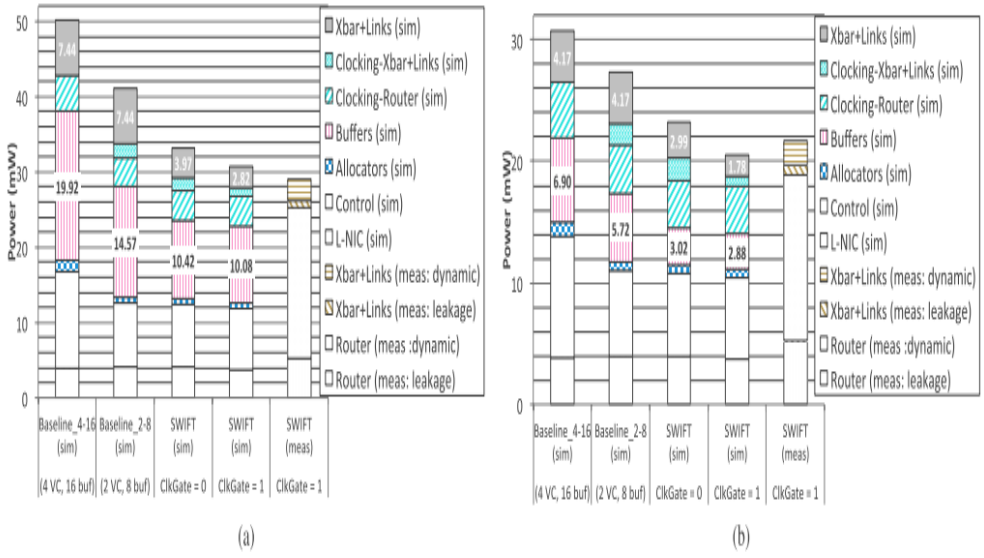

Figure: 2

Unlike prior approaches that are primarily performance- driven, in this paper we adopt a power-driven perspective towards the design of on-chip network micro architectures. As systems interconnected with on-chip networks become increasingly power-constrained, it is critical that we explore power-efficient network micro architectures. We first characterize the power profile of the on-chip network designs of two CMPs – the MIT Raw [20] and the UT Austin TRIPS [14], demonstrating that on-chip net- works take up a significant percentage of total system power (36% in Raw). This motivated us to propose three power- efficient router micro architectures and investigate the power efficiency of existing network architecture, evaluating their power-performance-area impact with detailed power modeling and probabilistic analysis. We then evaluated the proposed network micro architectures with synthetic as well as real CMP benchmark traffic traces, realizing 44.9% power savings with uniform random traffic, and 37.9%with TRIPS CMP traces as compared to a baseline network micro architecture based on current on-chip network designs. This substantial power saving is obtained with no degradation in network performance, and even improved performance in some cases.

Our study highlights the importance of a power-driven approach to on-chip network design. We will continue to investigate the interactions between traffic patterns and on-chip network architectures, and seek to reach a systematic design methodology for on-chip networks.

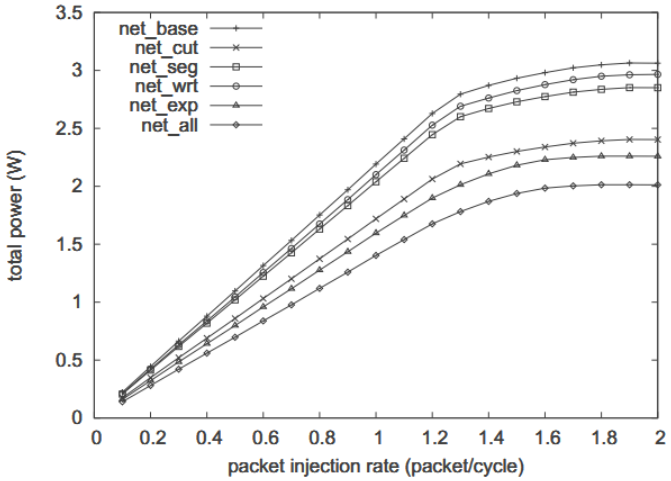

Figure: 3

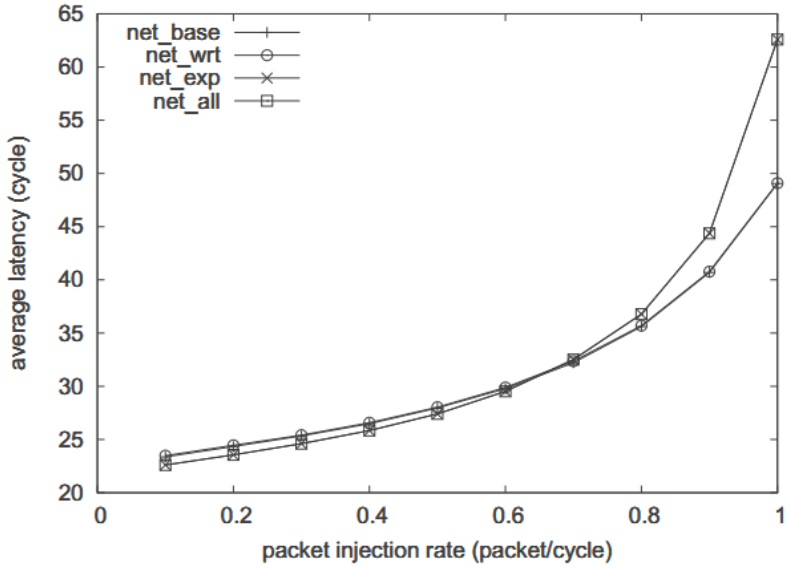

Figure: 4

B. Low-Power Network-on-Chip

An energy-efficient network-on-chip (NoC) is presented for possible application to high-performance system-on- chip (SoC) design. It incorporates heterogeneous intellectual properties (IPs) such as multiple RISCs and SRAMs, a reconfigurable logic array, an off-chip gateway, and a 1.6-GHz phase-locked loop (PLL). Its hierarchically-star-connected on-chip network provides the integrated IPs, which operate at different clock frequencies, with packet-switched serial-communication infrastructure.

Various low-power techniques such as low-swing signaling, partially activated crossbar, serial link coding, and clock frequency scaling are devised, and applied to achieve the power-efficient on-chip communications. The 5 5mm2 chip containing all the above features is fabricated by 0.18- m CMOS process and successfully measured and demonstrated on a system evaluation board where multimedia applications run.

The fabricated chip can deliver 11.2-GB/s aggregated bandwidth at 1.6-GHz signaling frequency. The chip consumes 160 mW and the on-chip network dissipates less than 51 mW.

Table: 1

Works	Power Reduction	Leakage Energy Saving	Variation Aware	Performance Up-gradation	Multiple V/F lines
[7]	65%	✓	×	×	×
[9]	50%	✓	×	×	×
[1]	N/A	✓	✓	✓	×
[5]	80%	×	✓	✓	✓
[12]	19%	✓	✓	-1%	×
[4]	10%	✓	✓	✓	✓
[3]	20-60%	✓	✓	×	✓
[13]	Info. N/A	×	✓	✓	✓
[2]	22-86%	✓	✓	×	✓
[6]	61.69%	×	✓	×	✓
[14]	63%	×	✓	✓	✓

A low-power NoC is designed and implemented for high-performance SoC applications. Heterogeneous IPs such as multiprocessors, memories, FPGA, and off-chip gateway with different timing references are interconnected in a hierarchical star topology. Various power-efficient techniques were suggested and implemented in each open system interconnection layer. Low-swing serial link and source-synchronous schemes in physical layer and low-energy serial link coding in data-link layer were proposed and realized on the NoC. Hierarchical circuit/packet switching, crossbar partial activation technique, and Mux-Tree based round-robin scheduler were also presented to reduce the power consumption in network layer. The on-chip network provides 11.2-GB/s bandwidth and consumes 51 mW at 1.6-GHz frequency. By using the proposed low-power techniques, the network power dissipation is reduced by 38%. The chip is implemented by 0.18- m CMOS process and success- fully operating with multimedia applications.

C. Power Reduction Techniques for Networks-on-Chip

Modern Network-on-Chip (NoC) links consume a significant fraction of the total NoC power, e.g., one study has shown that they consume up to 60% of total power and that this fraction is apparently growing. We present two algorithms for power reduction in NoC links.

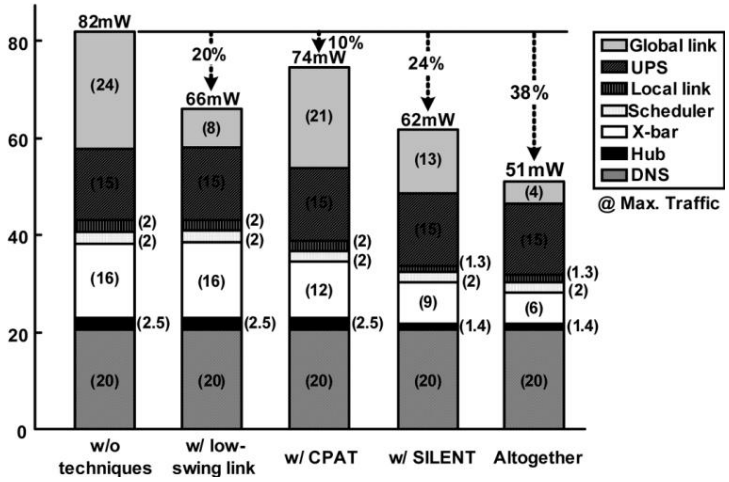

Figure: 5

We first present Parity Routing (PaR), a novel method to add redundant parity information to packets, while minimizing the number of redundant bits transmitted. PaR exploits NoC path diversity in order to avoid transmitting some of the redundant parity bits. Our analysis shows that, for example, on a 4x4 NoC with a demand of one parity bit, PaR reduces the redundant information transmitted by 75%, and the savings increase asymptotically to 100% with the size of the NoC. In addition, we show that PaR can yield power savings due to the reduced number of bit transmissions. Furthermore, PaR utilizes low complexity, small-area circuits.

Second, we present Selective Packet Interleaving (SPI), a flit transmission scheme that reduces power consumption in NoC links. SPI decreases the number of bit transitions in the links by exploiting the multiplicity of virtual channels in a NoC router. SPI multiplexes flits to the router's output link so as to minimize the number of bit transitions from the previously transmitted flit. Analysis and simulations demonstrate a reduction of up to 55% in the number of bit transitions and up to 40% savings in power consumed on the link. SPI benefits grow with the number of virtual channels. SPI works better for links with a small number of bits in parallel. While SPI compares favorably against bus inversion, combining both schemes helps to further reduce bit transitions.

Modern integrated circuits introduce low power design challenges. The lion's share of power consumption lies with the

interconnect switching activity, and this share is expected to grow in years to come [1, 12, 33, 34, 38, 44]. In this thesis we presented two algorithms for power reduction on NoC links.

First, we presented PaR parity routing, a low-overhead error detection solution for networks on chip. PaR can be used to provide any predefined error protection requirement. It exploits NoC path diversity, and selects routing paths based on parity bits. It thus saves the actual transmissions of these bits, along with the associated power penalty. PaR uses simple, low-complexity encoding and decoding circuits. We analyzed the savings achieved by PaR, and showed that it yields significant savings even on small NoCs, (for example, saving 75% of redundant bit transmissions on a 4x4 NoC mesh), and its savings asymptotically converge to 100% with the size of the NoC. We further showed that PaR can yield power savings (for example, saving 35% of redundant power consumption on a 3x3 NoC mesh NoC).

Second, we presented SPI - selective packet interleaving, a flit transmission scheme for energy efficient NoCs. SPI exploits the multiplicity of virtual channels to transmit a dynamically chosen flit so as to minimize bit transitions between consecutive flits. SPI uses simple, low-complexity circuits. We analyzed the savings achieved by SPI, and showed that SPI yields a significant improvement in power consumption, which outweighs the cost of implementing SPI. For example, with 8b width links and 4 VCs, SPI reduces the average number of bit transitions over the link by more than 35%, and reduces the power consumption by 25%. Analysis and simulations demonstrate a reduction of up to 55% in the number of bit transitions and up to 40% savings in power consumed on the link. Furthermore, SPI's benefits grow with the number of virtual channels.

Table: 2

Methods	Constraints			
	Power consumption (mW)	Area (number of slices)	Improved power	Improved area
IntelliBuffer [2]	410.42	1551	37.31%	49.4%
Adaptive data compression [3]	474.53	1054	45.79%	25.5%
Buffer clock-gating [10]	318.63	917	19.26%	14.4%
Newly proposed	257.05	785		

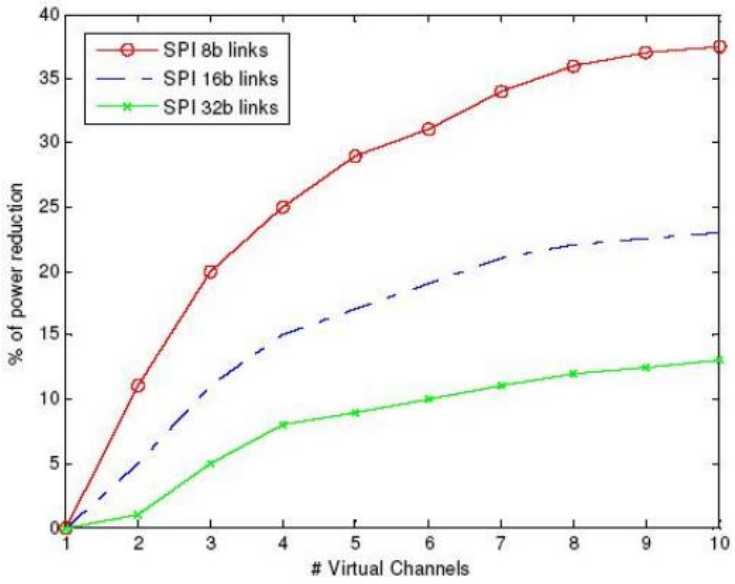

Figure: 6

D. SWIFT: A Low-Power Network-On-Chip

A 64-bit, 8 × 8 mesh network-on-chip (NoC) is presented that uses both new architectural and circuit design techniques to improve on-chip network energy-efficiency, latency, and throughput. First, we propose token flow control, which enables bypassing of flit buffering in routers, thereby reducing buffer size and their power consumption. We also incorporate reduced-swing signaling in on-chip links and crossbars to minimize datapath interconnect energy.

The 64-node NoC is experimentally validated with a 2 × 2testchipin90nm, 1.2 V CMOS that incorporates traffic generators to emulate the traffic of the full network. Compared with a fully synthesized baseline 8×8 NoC architecture designed to meet the same peak throughput, the fabricated prototype reduces network latency by 20% under uniform random traffic, when both networks are run at their maximum operating frequencies. When operated at the same frequencies, the SWIFT NoC reduces network power by 38% and 25% at saturation and low loads, respectively.

In this paper, we presented a NoC that utilizes low-power architecture and circuit co-design to improve the power, latency, and throughput of a NoC. In particular, a token- based smart pipeline bypassing scheme, and a reduced-swing crossbar and interconnect together contribute to latency and power improvements in an 8×8 network running uniform random traffic, while requiring half as many buffers as extracted simulations of a baseline NoC using virtual-channel routers. Under uniform random traffic, a reduction of 38% in peak network power was reported when networks were operated at identical frequency conditions, while a 20% reduction in low-load latency was reported when both networks are run at their maximum operating frequencies.

Reduced swing circuits achieve 62% power savings in the data path versus a full-swing, synthesized implementation. Differential mode shielding was also presented as a means to enable protected, reduced-swing signaling over digital logic with less capacitive loading than full ground plane shielding.

Many of the architectural and circuit novelties in SWIFT would enhance any NoC router/link design, as SWIFT performs more efficient allocation of net- work links and buffers, enabling low-power traversal. We hope this paper paves the way for more such prototype designs. Demonstrating a SWIFT-like NoC design on a multicore chip with real application traffic is part of our future work.

E. Area and Power Efficient Router Design

As network on chip (NoC) systems become more prevalent in today's industry. Routers and interconnection networks are the main components of NoC. Therefore, there is a need to obtain low area and power models for these components so that we can better understand the area and power tradeoffs. In this paper a low- area and power efficient NoC architecture is proposed by eliminating the virtual channels. Buffers are replaced by elastic buffer.

In order to get the advantage of both buffered and buffer less the cross bar is split in to two parts. Implementation is done in Micro wind 3.5 the proposed router area is reduced by 47.89% and power is reduced by 11.2% compared to base line router accordingly.

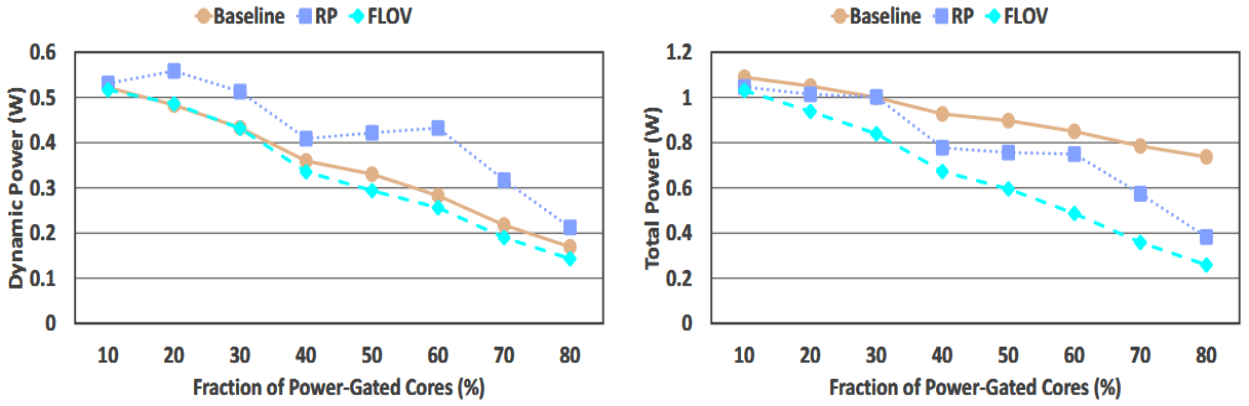

Figure: 7

In this paper we evaluate the performance of dual cross bar router design using elastic buffer with an objective of reducing power and area. The proposed design shows power reduction of 11.2% and area reduction 47.89% compare to baseline router result in increase in performance. With the proposed design we conclude that the advantage of both buffered and buffer less is achieved and single elastic buffer is enough to store a number of flit on each port with saving power when compared to the same number of VC router buffer based NoC architecture.

Design	Total average power(mW)	Area (mm2)
Baseline router	15.689	545496.0
Proposed router	13.93	284247.6

F. Voltage/Frequency Scaling in NoC

Voltage and frequency is dynamically scaled to produce energy efficient multi-core network- on-chip (NoC). A detailed analysis of the techniques employed for this purpose are studied and based on the optimized performance the most effective ones are reported. We also highlight the most promising high performance and energy minimizing techniques.

Techniques using DVFS have been studied in detail to find the most energy efficient technique of implementation of DVFS. Implementation of the traditional DVFS operations can be disintegrated into two parts: (a) clock transition and (b) voltage transition depending on the workload of the processor. Optimal core mapping also contributes to boost the performance. The

DVFS has become an integral part of the NoC for saving power as well as for reducing congestion by increasing the frequency of the congested router. Therefore, DVFS is essential for increasing the throughput of the system.

G. Smart Power-Saving Architecture for Network on Chip

In network-on-chip (NoC), the data transferring by virtual channels can avoid the issue of data loss and deadlock. Many virtual channels on one input or output port in router are included. However, the router includes five I/O ports, and then the power issue is very important in virtual channels. In this paper, a novel architecture, namely, Smart Power-Saving (SPS), for low power consumption and low area in virtual channels of NoC is proposed. The SPS architecture can accord different environmental factors to dynamically save power and optimization area in NoC. Comparison with related works, the new proposed method reduces 37.31%, 45.79%, and 19.26% on power consumption and reduces 49.4%, 25.5% and 14.4% on area, respectively.

The Smart Power-Saving (SPS) architecture for network-on- chip was presented. A clock control circuit and SPS algorithm are demonstrated to reduce the power consumption on the NoC architecture. From experimental results, the proposed SPS architecture is more efficient to reduce the power consumption than IntelliBufer [1], adaptive data compression [3], and buffer clock-gating [10] in the NoC architecture.

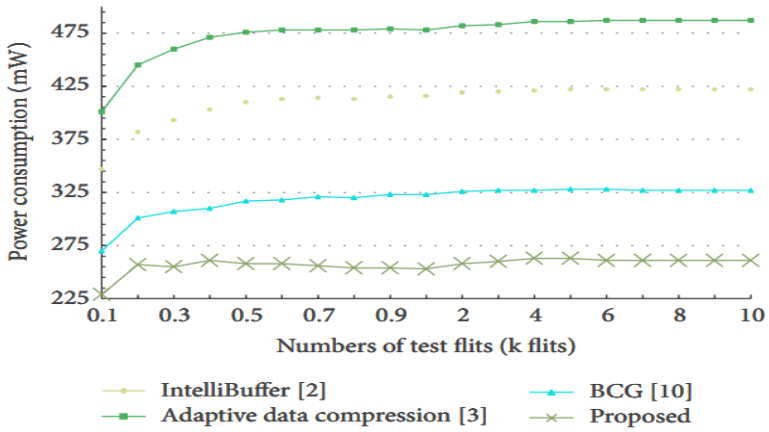

Figure: 8

H. Cognitive NoC Design

The number of cores in a multicore chip design has been increasing in the past two decades. The rate of increase will continue for the foreseeable future. With a large number of cores, the on-chip communication has become a very important design consideration. The increasing number of cores will push the communication complexity level to a point where managing such highly complex systems requires much more than what designers can anticipate for.

We propose a new design methodology for implementing a cognitive network-on-chip that has the ability to recognize changes in the environment and to learn new ways to adapt to the changes. This learning capability provides a way for the network to manage itself. Individual network nodes work autonomously to achieve global system goals, e.g., low network latency, higher reliability, power efficiency, adaptability, etc. We use fault-tolerant routing as a case study. Simulation results show that the cognitive design has the potential to outperform the conventional design for large applications. With the great inherent flexibility to adopt different algorithms, the cognitive design can be applied to many applications.

We have devised a methodology for designing and implementing cognitive NoCs. Cognitive NoCs have ability to learn and to recognize changes in the environment. Net- work constituents work autonomously but collectively achieve global goals. It is a holistic approach that crosses many architectural layers.

A case study on fault-tolerant routing shows that with a simple and straight-forward learning algorithm, Cognitive NoC outperforms slightly the conventional design over an extended period of time. It is able to quickly overcome the initial learning's large overhead. The combination of cognitive capabilities and architectural design makes Cognitive NoC a promising approach to address challenging issues effected in ever increasingly complex systems. We believe the Cognitive NoC approach is a good way to target many large applications that allow time for the system to overcome the initial learning over- head. Cognitive NoC provides many benefits: 1) better performance over time; 2) simpler and smaller router design; 3) great flexibility to accommodate different types of algorithms,

which broadens the design space and increases the applicability to many application domains; 4) better network scalability; 5) field upgradability and repair; and 6) reduction in costs in silicon validation and testing.

I. Configurable Power Efficient 3-Dimensional Crossbar Switch

In this paper, a 3-D crossbar switch is designed for 3-D mesh Network-on-Chip to meet the current requirements of smaller area, power efficiency and high speed. To increase the overall data transfer rate of the network, the number of ports are added in 3-D crossbar switch and found that it has been working at the same power as the 2-D crossbar switch (with lesser ports). The functional verification of the crossbar switch design has been done on ModelSim6.4a. Xilinx13.3 has been used for synthesis. Also, FPGA verification of the proposed design has been done.

The configurable power efficient 3-D crossbar switch has been designed for low power network. Because of the efficient use of transistors, the power for 3-D design remains same as the 2-D even when the numbers of ports and LUT count increases in 3-D crossbar. The propagation delay increase which is negligible in comparison to the increase in simultaneous data transfer rate.

In 3-D crossbar, data can be transferred from 7 input ports to 7 output ports simultaneously whereas in 2-D data from only 5 input ports can be transmitted to 5 output ports simultaneously. The design has been verified using FPGA.

DESIGN	AREA (LUT)	DELAY (ns)	POWER (mW)
2-D Crossbar Switch [9]	80	7.694 ns	14mW
Proposed 3-D Crossbar Switch	168	7.761ns	14mW

J. Power-Gating Mechanism for Energy-Efficient Networks-on-Chip

Reducing static NoC power consumption is becoming critical for energy-efficient computing as technology scales down since NoCs are devouring a large fraction of the on-chip power budget. We propose Fly-Over (FLOV), a light-weight distributed mechanism for power-gating routers.

With simple modifications to the baseline router architecture, FLOV links are facilitated over power-gated routers. A Handshake protocol that allows seamless

router power- gating in addition to a dynamic routing algorithm, that provides best-effort minimal path without the necessity for global network information, maintain normal NoC functionality. We evaluate our schemes using synthetic workloads as well as real workloads from PARSEC 2.1 benchmark suite. The results show that FLOV can achieve on average 19.2% latency reduction and 15.9% total energy savings.

We proposed Fly-Over (FLOV), a light-weight distributed router power-gating mechanism for NoCs. FLOV power-gates routers attached to powered-down cores without global network information, but still ensures network connectivity. Performance evaluations using synthetic and real workloads show that FLOV not only achieves better NoC power savings due to power-gating more routers but avoids aggregated traffic rerouting in the network unlike Router Parking.

K. Folded Torus-Like Network-on-Chip

Dark silicon refers to the phenomenon that a fraction of a many-core chip has to become "dark" or "dim" in order to guarantee the system to be kept in a safe temperature range and allowable power budget. Techniques have been developed to selectively activate non-adjacent cores on many-core chip to avoid temperature hotspot, while resulting unexpected increase of communication overhead due to the longer average distance between active cores, and in turn affecting application performance and energy efficiency, when Network-on-Chip (NoC) is used as a scalable communication subsystem.

To address the brand-new challenges brought by dark silicon, in this paper, we present FoToNoC, a Folded Torus-like NoC, coupled with a hierarchical management strategy for heterogeneous many-core systems. On top of it, objectives of maximizing application performance, energy efficiency and chip reliability are isolated and well achieved by hardware-software co-design in several different phases, including application mapping and scheduling, cluster management and DVFS control.

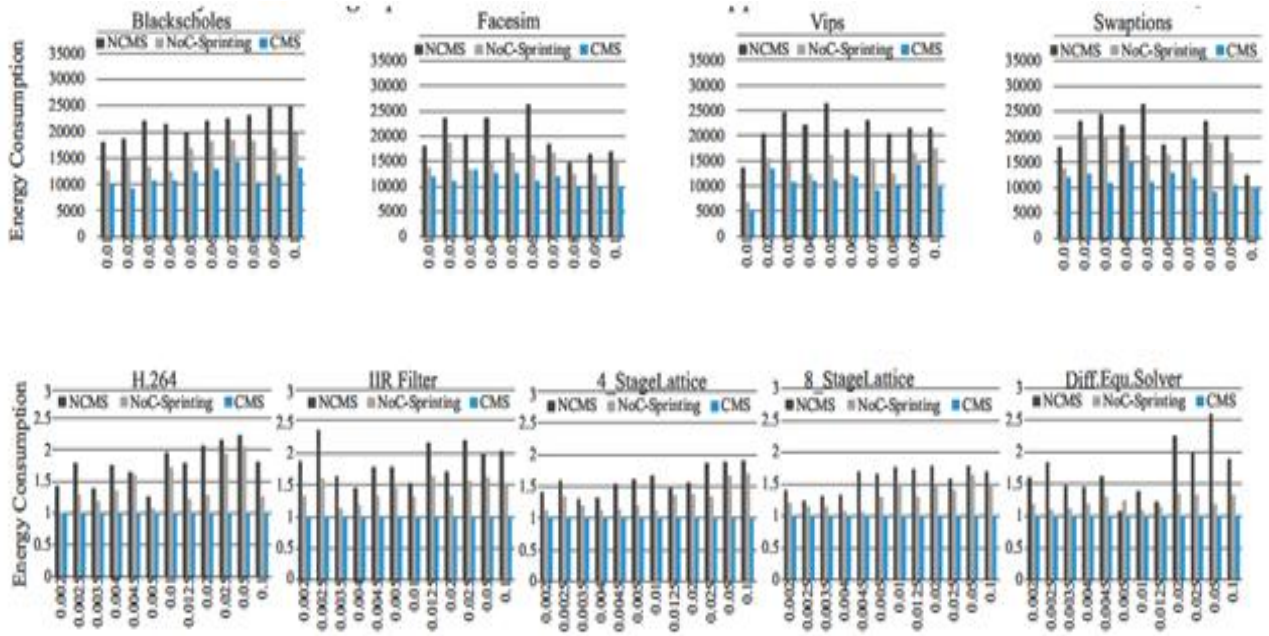

Figure: 9

Evaluations on PARSEC benchmark applications demonstrate the significance of the entire strategy. Compared with state-of-the-art approaches, the proposed FoToNoC organization can achieve on average 35:4% and 35:2% on communication efficiency and application performance improvement, respectively, when maintaining the safe chip temperature. The hierarchical cluster-based management strategy can further reduce an average 34:6% of the total energy consumption with a notable reduction on the chip peak temperature. The significant achievements on system energy efficiency and the reduction on chip temperature of H.264 decoder and DSP-stone benchmarks additionally verify the effectiveness of the proposed methods.

Dark silicon brings new challenges to the management of many-core systems due to the conflicting requirements on the selectively activated cores for safe temperature and short-distance communication consideration. In this paper, we present a hierarchical management policy that organizes heterogeneous cores in virtual clusters such that cores within a cluster are physically distant on the chip but logically adjacent to each other for the lowest chip temperature and the best inter-communication performance. On top of it, the proposed dark silicon- aware application

mapping and scheduling method has made trade-offs on optimizing application performance, chip temperature and system energy efficiency with the reduced complexity in several phases and aspects. Evaluation results on benchmark applications show the significant advantages of the proposed approaches in dark silicon many-core systems, compared to state-of-the-art management strategies.

L. Low-Power and Low-Latency Network-on-Chip

Packet-switched Network-on-Chip (NoC) is the shared global communication infrastructure for future large-scale chip multi-processors (CMPs). Recently, Single-cycle Multi-hop Asynchronous Repeated Traversal (SMART) on repeater-inserted wires to reduce packet delay was proposed. But current NoC with SMART support adds complexity to conventional routers and incurs high power consumption. In this paper, we propose a low- power and low-latency NoC design with SMART support, called Dimension Ordered Asynchronous Repeated Traversal (DOART). First we design a low-power interconnect called Single-cycle Intra-dimension Bridge (SIB) with SMART support, and then we propose an efficient construction framework to connect SIBs generating a large-scale low-power and low-latency NoC. In addition, the proposed DOART supports virtual channel and is protocol and routing-level deadlock-free. Experimental results show that DOART can reduce both the application execution time and network power consumption compared with state-of-the-art NoCs with SMART support.

The configuration of Sniper simulation is in Table I. Fig- ure 6(a) shows the execution time of SPLASH-2 workloads normalized to that of the baseline NoC. The three low-latency NoCs, LC, SMART 2D and DOART, all reduce applications' execution time, demonstrating NoC's impact on application performance. DOART NoC reduces applications' execution time by 14.4% in average, and 25.5% maximally, compared with the baseline. Figure 6(b) shows the four networks' power consumption normalized to that of the baseline NoC. DOART reduces network power consumption by 61.4%, 30.1% and 63.9% compared with baseline, LC, and SMART 2D, respectively. In contrast, SMART 2D causes 6.8% higher power consumption than the baseline NoC in average.

Version	Area (μm^2)	Power (Min.) Leakage power (mW)	Power (Min.) Dynamic power (mW)	Power (Min.) Total power (mW)	Power (Max.) Leakage power (mW)	Power (Max.) Dynamic power (mW)	Power (Max.) Total power (mW)
Base Router	229944	2.30	10.81	13.1028	2.06	449.2474	451.3119
Router with PM	232565 (+1.14%)	2.13	7.29	9.4197 (-28.11%)	2.08	487.71	489.7825 (+8.52%)

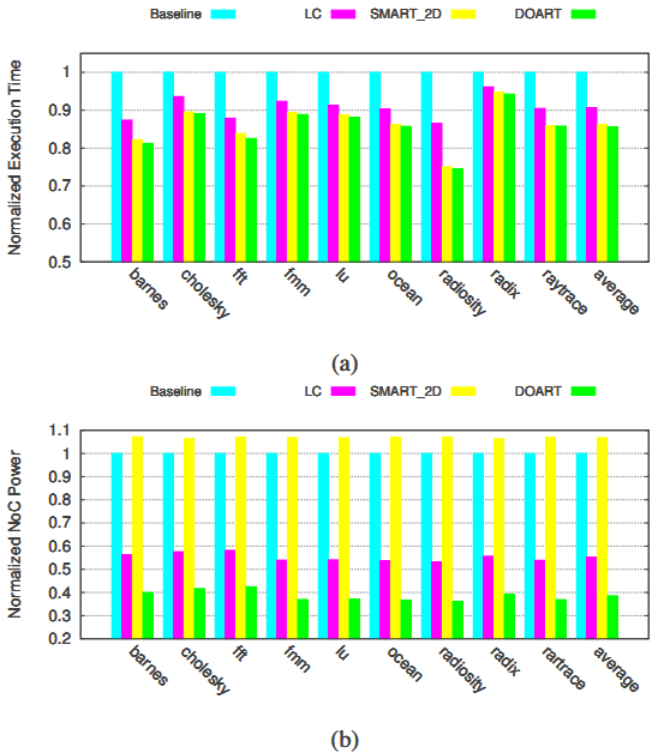

Figure: 10

M. Energy Aware Routing of Multi-level Network-on-Chip

The emergence of Network-on-Chip (NoC) as a communication paradigm for Multi-Processor System-on-Chips (MPSoCs) significantly exacerbates the need to provide a method- ology that optimizes the energy consumption of the overall system. This is especially important when factoring in current Network- on-Chip advances which have multiple communication media such as on-chip wireless or nano-photonics links, hybrid with traditional wired links. All of these media have different energy profiles, and if not taken into consideration the system will incur higher power consumption throughout the runtime of the application. In this work, the case for EDP (energy-delay

product) optimization between different levels of a multi-level Network- on-Chip is presented.

Using a dynamic, energy aware algorithm, the EDP improvement is compared to a multi-level Network-on- Chip using a statically optimized routing. The proposed routing algorithm handles the different types of energy-delay profiles of multiple links. The end product is a methodology that lowers the overall energy consumption by optimizing the energy profile of the Network-on-Chip while also minimizing the network delay.

A shift in focus has to be made between homogeneous and heterogeneous networks. This shift from load balancing to energy balancing is a necessity when dealing with multiple new communication media. This work adapted a dynamic routing algorithm to emphasize the energy characterization of the net- work and therefore make improved routing decisions by taking into consideration the importance of energy consumption of a packet transfer through different types of network links. The algorithm uses the concept of local router range [14] detailed in Section IV-A to dynamically modify the direction of the packet traversal through the multi-level Network-on- Chip.

An optimally selected static local router range shows good improvement to the overall performance of the network (measured with the reciprocal of EDP). Thus, validating the importance of a good routing algorithm, but the modifications that dynamically change the local router range by accounting for the energy expenditure of the various links offers an additional performance increase.

Simulating the network with multiple traffic patterns shows a performance improvement by up to 1.42× and a 21% savings in energy consumption compared to a optimally selected static local router range.

Future work should include improvements to the overall framework of the project including but not limited to using real application traffic to conduct the various simulations, testing additional multi-level topologies and adding additional runtime topological modifications (not limited to local router range).

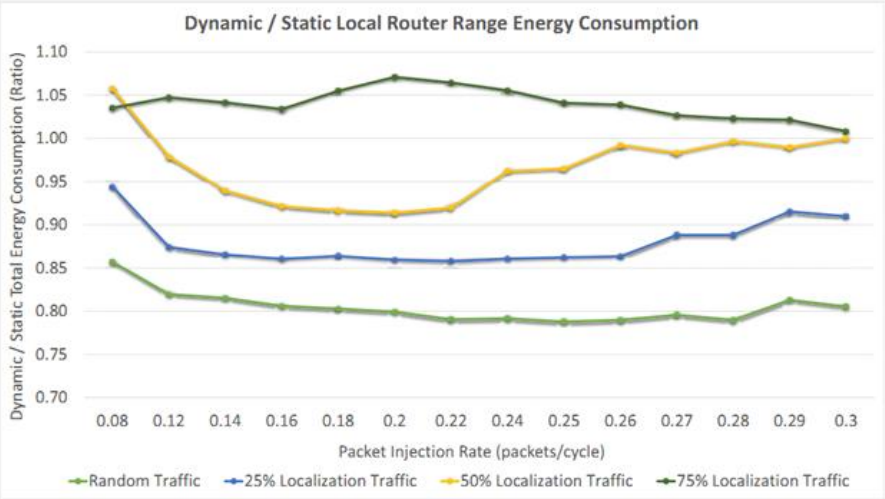

Figure: 11

N. Opportunistic Circuit-Switching

Modern on-chip networks (NoCs) rely on virtual channel (VC) flow control to allow effective utilization of link bandwidth at the cost of more power and longer per-hop latency. Despite many existing optimization techniques for NoCs under VC flow control, we take a further step on questioning its necessity. Our finding is, when the network is not busy, circuit- switching (CS) may already satisfy the performance requirements with much smaller power consumption and shorter per-hop latency. In this paper, we propose to opportunistically enable CS in NoCs under VC flow control. This allows us to effectively reduce the power consumption of NoCs through having less buffering and longer sleep intervals for power gating while retaining CS-like per-hop latency. Our evaluations reveal that this proposal leads to a reduction of network power by up to 70% while cutting the system energy footprint by up to 35%.

Flow control is an important topic for NoCs since it determines how traffic is treated and is highly related to both performance and power consumption of the network. In this paper, we proposed a novel flow control method which allows CS opportunistically in order to lower the per-hop latency and power consumption while still being able to maintain the network throughput when necessary since our proposal is built on networks under VC flow control. We found that OCS is able to reduce the energy consumption of the system by up to 35% while improving the system throughput by

up to 50%. In more details, power of the network is cut by up to 70% due to having less accesses to the buffers and longer sleep intervals to gate the routers. Such effectiveness proves that our proposal is very suitable for future NoC designs as energy efficiency is of more and more importance.

O. Power Management Controller for Online Power Saving

Growing chip integration density and increasing frequencies lead to tremendous leakage power and henceforth to chip heat problems. Power management is one possibility to reduce the power consumption and get the temperature problem under control. Current technology mainly focuses on power- gating techniques on basis of multi-core systems but leaving the network perspective out of scope. We provide a holistic concept, bringing together power-gating and frequency scaling techniques for network-on-chips. Following this, network static power consumption could be minimized without affecting the system performance. We present a light-weight power management controller for network-on-chips with online monitoring to optimize the power consumption of network resources. Our work comprises a hardware simulation model for design space exploration of varying technology specific parameters and an FPGS based prototype for verification. The power saving potential heavily depends on the network communication load. Our power controller adds only 2.1% of resources while 28.11% of the total power could be saved with clock gating.

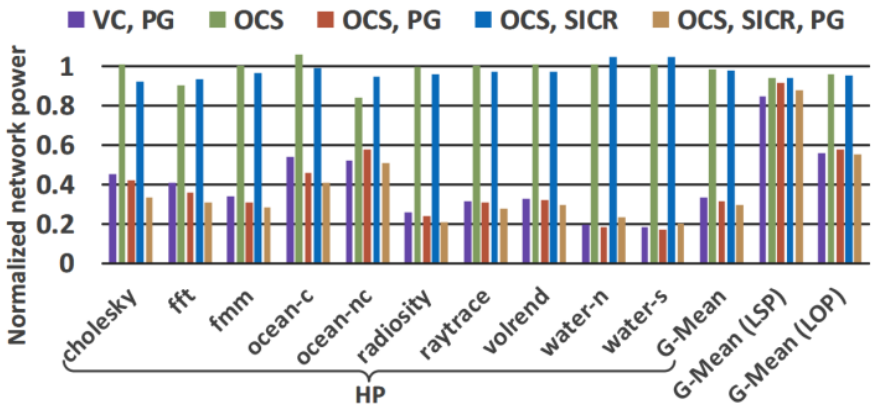

Figure: 12

In this paper, we presented a power management controller for on chip networks. Each router comprises its own control- ling unit to realize a decentralized power optimization while minimizing the performance loss. Depending on the degree of router utilization, the controller either suggest a frequency scaling factor or power gates the complete router in case of no utilization. The resource overhead of the PMC is kept to a minimum, while saving an average of 14.2% of the network power. A FPGA based implementation of the approach is available for evaluation as well as an ASIC design flow for power simulation.

CONCLUSION

Previously, research and design of Network-on-Chip paradigms where mainly focused on improving the performance of the interconnection networks. With emerging wide range of low-power applications and energy constrained high-performance applications, it is highly desirable to have NoCs that are highly energy efficient without incurring performance penalty. In the design of high-performance massive multi-core chips, power and heat have become dominant constraints. Increased power consumption can raise chip temperature, which in turn can decrease chip reliability and performance and increase cooling costs. Dynamic Voltage Scaling is an efficient technique for significant power savings in microprocessors. It has been proposed and deployed in modern microprocessors by exploiting the variance in processor utilization. On a Network-on-Chip paradigm, it is more likely that the wire line links and buffers are not always fully utilized even for different applications. Hence, by exploiting these characteristics of the links and buffers over different traffic, DVFS technique can be incorporated on these switches and wire line links for huge power savings.

REFERENCES

1) Mishra, A.K.; Das, R.; Eachempati, S.; Iyer, R.; Vijaykrishnan, N.; Das, C.R., "A case for dynamic frequency tuning in on-chip networks," Microarchitecture, 2009. MICRO-42. 42nd Annual IEEE/ACM International Symposium on , vol., no., pp.292,303, 12-16 Dec. 2009

2) Murray, J.; Pande, P.P.; Shirazi, B., "DVFS-enabled sustainable wireless NoC architecture," SOC

Conference (SOCC), 2012 IEEE International , vol., no., pp.301,306, 12-14 Sept. 2012

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

4) Wonyoung Kim; Gupta, M.S.; Gu-Yeon Wei; Brooks, D., "System level analysis of fast, per-core DVFS using on-chip switching regulators," High Performance Computer Architecture, 2008. HPCA 2008. IEEE 14th International Symposium on , vol., no., pp.123,134, 16-20 Feb. 2008

5) Wooyoung Jang; Pan, D.Z., "A Voltage-Frequency Island Aware Energy Optimization Framework for Networks-on-Chip," Emerging and Selected Topics in Circuits and Systems, IEEE Journal on, vol.1, no.3, pp.420,432, Sept. 2011

6) Ogras, U.Y.; Marculescu, R.; Marculescu, D., "Variation-adaptive feedback control for networks-on-chip with multiple clock domains," Design Automation Conference, 2008. DAC 2008. 45th ACM/IEEE , vol., no., pp.614,619, 8-13 June 2008

7) Niyogi, K.; Marculescu, D., "Speed and voltage selection for GALS systems based on voltage/frequency islands," Design Automation Conference, 2005. Proceedings of the ASP-DAC 2005. Asia and South Pacific , vol.1, no., pp.292,297 Vol. 1, 18-21 Jan. 2005

8) Donald, J.; Martonosi, M., "Techniques for Multicore Thermal Management: Classification and New Exploration," Computer Architecture, 2006. ISCA '06. 33rd International Symposium on , vol., no., pp.78,88, 0-0 0

9) Shang, L.; Li-Shiuan Peh; Kumar, A; Jha, N.K., "Temperature-Aware On-Chip Networks," Micro, IEEE , vol.26, no.1, pp.130,139, Jan.-Feb. 2006 doi: 10.1109/MM.2006.23

10) Jiong Luo; Li-Shiuan Peh; Niraj Jha, "Simultaneous dynamic voltage scaling of processors and communication links in real-time distributed embedded systems," Design, Automation and Test in Europe Conference and Exhibition, 2003 , vol., no., pp.1150,1151, 2003

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

12) Garg, Siddharth; Marculescu, D.; Marculescu, R.; Ogras, U., "Technology-driven limits on DVFS controllability of multiple voltage-frequency island designs: A system- level perspective," Design Automation Conference, 2009. DAC '09. 46th ACM/IEEE , vol., no., pp.818,821, 26-31 July 2009

13) Wettin, Paul; Murray, Jacob; Pande, Partha; Shirazi, Behrooz; Ganguly, Amlan, "Energy-efficient multicore chip design through cross-layer approach," Design, Automation & Test in Europe Conference & Exhibition (DATE), 2013 , vol., no., pp.725,730, 18-22 March 2013

14) J. Duato, S. Yalamanchili, and L. Ni, Interconnection Networks: An Engineering

15) Approach. Morgan Kaufmann, 2003, p. 600.

16) O. Lysne, T. Skeie, S. -a. Reinemo, and I. Theiss, “Layered routing in irregular networks,” IEEE Trans. Parallel Distrib. Syst., vol. 17, no. 1, pp. 51–65, Jan. 2006.

17) J. Lin, H. Wu, and Y. Su, “Communication using antennas fabricated in silicon integrated circuits,” Solid-State Circuits, vol. 42, no. 8, pp. 1678–1687, 2007.

18) A. Ganguly and Vineeth vijayakumaran, Manoj Prashanth Yuvaraj, Naseef Mansoor, “CDMA Enabled Wireless Network-on-Chip,”.

19) R. Jotwani and S. Sundaram, "An x86-64 core implemented in 32nm SOI CMOS," Proc. IEEE Int. Solid-State Circuits Conf., pp. 106–107, 2010.

20) P.P. Pande, C. Grecu, M. Jones, A. Ivanov, and R. Saleh, "Performance Evaluation and Design Trade-Offs for Network-on-Chip Interconnect Architectures," IEEE Tran. Computers, vol. 54, no. 8, pp. 1025-1040, Aug. 2005.

21) R. Ho, K.W. Mai, and M.A. Horowitz, "The Future of Wires," Proc. IEEE, vol. 89, no. 4, pp. 490-504, Apr. 2001.

22) P. Kapur, J.P. Mc Vittie, and K.C. Saraswat, "Technology and Reliability Constrained Future Copper Interconnects—Part II: Performance Implications," IEEE Tran. Electron Devices, vol. 49, no. 4, pp. 598-604, Apr. 2002.

23) D. Sylvester and K. Keutzer, "Impact of Small Process Geometries on Microarchitectures in Systems on a Chip," Proc. IEEE, vol. 89, no. 4, pp. 467-489, Apr. 2001.

24) Semiconductor Industry Association (SIA). (2003). International Roadmap for Semiconductors, 2003 edition, Austin, TX. International SEMATECH, 2003. [Online]. Available: http://www.itrs.net/links/2003itrs/home2003.htm

25) C. Grecu, P.P. Pande, A. Ivanov, and R Saleh, "Structured Interconnect Architecture: A Solution for the Non-Scalability of Bus-Based SoCs," Proc. Great Lakes Symp. VLSI, pp. 192-195, Apr. 2004.

26) C. Hsieh and M. Pedram, "Architectural Energy Optimization by Bus Splitting," IEEE Tran. Computer-Aided Design, vol. 21, no. 4, pp. 408-414, Apr. 2002. 41

27) M. Horowitz and B. Dally, "How Scaling Will Change Processor Architecture," Proc. Int. Solid-State Circuits Conf., pp. 132-133, Feb. 2004.

28) J. Nurmi, Interconnect-Centric Design for Advanced SoC and NoC. Springer Science + Business Media Inc., Germany, 2005.

29) M. Sgroi, M. Sheets, A. Mihal, K. Keutzer, S. Malik, J. Rabaey, and A. Sangiovanni-Vincentelli, "Addressing the System-on-a-Chip Interconnect Woes through Communication based Design," Proc. 38th Design Automation Conf., Las Vegas, pp. 667-72, Jun. 2001.

30) P. Macken, M. Degrauwe, M. V. Paemel, and H. Oguey, "A Voltage Reduction Technique for Digital Systems," IEEE Int. Solid-State Circuits Conf., pp. 238–239, Feb. 1990.

31) C. Lai, J. H. Lin, and Y. F. Wang, "DVFS SoC Architecture and Implementation," SoC Technology Journal, vol. 3, pp. 84–91, Nov. 2005.

32) C. Isci, A. Buyuktosunoglu, C.-Y. Cher, P. Bose, and M. Martonosi, "An Analysis of Efficient Multi-core Global Power Management Policies: Maximizing Performance for a Given Power Budget," Proc. 39th Annu. IEEE/ACM Int. Symp. Microarchitecture, vol. 26, no. 1, pp. 119-129, Feb. 2006.

33) G. Semeraro, G. Magklis, R. Balasubramonian, D. H. Albonesi, S. Dwarkadas, and M. L. Scott, "Energy-efficient Processor Design Using Multiple Clock Domains with Dynamic Voltage and Frequency Scaling," Int. Symp. High-Performance Computer Architecture, pp. 29-40, Feb. 2002. 42

34) T. Simunic, L. Benini, A. Acquaviva, P. Glynn, and G. D. Micheli, "Dynamic Voltage Scaling and Power Management for Portable Systems," Design Automation Conf., pp. 524-529, Jun. 2001.

35) Q. Wu, P. Juang, M. Martonosi, and D. W. Clark, "Voltage and Frequency Control with Adaptive Reaction Time in Multiple-Clock-Domain Processors," 11th Int. Symp. High-Performance Computer

Architecture, pp. 178-189, Feb. 2005.

36) W. Kim, M. Gupta, G. Y. Wei, and D. Brooks, "System Level Analysis of Fast, Per-core DVFS Using On-chip Switching Regulators," Int. Symp. High Performance Computer Architecture, pp. 123–134, Feb. 2008.

37) U.Y. Ogras, R. Marculescu, P. Choudhary, and D. Marculescu, "VoltageFrequency Island Oartitioning for GALS-Based Networks-on-Chip," Proc. 44th Annu. Design Automation Conf., pp. 110-115, Jun. 2007.

38) D. Bertozzi, "NoC Synthesis Flow for Customized Domain Specific Multiprocessor Systems-on-Chip," IEEE Tran. Parallel and Distributed Systems, vol. 16, no. 2, pp. 113-129, Feb. 2005.

39) J. Dielissen, A. Radulescu, K. Goossens, and E. Rijpkema, "Concepts and Implementation of the Philips Network-on-Chip," IP-based SoC Design, Nov. 2003.

40) M. Millberg, E. Nilsson, R. Thid, and A. Jantsch, "Guaranteed Bandwidth Using Looped Containers in Temporally Disjoint Networks within the Nostrum Network on Chip," Proc. Design Automation and Test in Europe (DATE), pp. 890-895, Feb. 2004. 43

41) Y. S. Dhillon, A. U. Diril, A. Chatterjee, and H. S. Lee, "Algorithm for Achieving Minimum Energy Consumption in CMOS Circuits Using Multiple Supply and Threshold Voltages at the Module Level," Proceedings of ICCAD, pp. 693-700, Nov. 2003.

42) K. Niyogi and D. Marculescu, "Speed and Voltage Selection for GALS Systems Based on Voltage/Frequency Islands," Proceedings of ASP-DAC, pp. 292-297, Jan. 2005.

43) M. Powell, S.-H Yang, B. Falsafi, K. Roy, and T.N. Vijaykumar, "Reducing Leakage in a High-Performance Deep-Submicron

Instruction Cache," IEEE Tran. VLSI Systems, vol. 9, no. 1, pp. 77-89, Feb. 2001.

44) S. Shigematsu, S. Mutoh, Y. Matsuya, and J. Yamada, "A 1-V High-Speed MTCMOS Circuit Scheme for Power-Down Application Circuits," IEEE Journal on Solid-State Circuits, vol. 32, no. 6, pp. 861-869, Jun. 1997.

45) B.H. Calhoun, F.A. Honore, and A.P Chandrakasan, "A Leakage Reduction Methodology for Distributed MTCMOS," IEEE Journal on Solid-State Circuits, vol. 39, no. 5, pp. 818-826, May 2004.

46) C. Long and L. He, "Distributed Sleep Transistor Network for Power Reduction," Proc. IEEE/ACM Design Automation Conf., pp. 181-186, Jun. 2003.

47) A. Ramalingam, B. Zhang, A. Davgan, and D. Pan, "Sleep Transistor Sizing Using Timing Criticality and Temporal Currents," Proc. ASP-DAC, pp. 1094-1097, Jan. 2005. 44

48) K. Shi and D. Howard, "Challenges in Sleep Transistor Design and Implementation in Low-Power Designs," Proc. 43rd Annu. Design Automation Conf., pp. 113-116, Jul. 2006.

49) D.E. Lackey, P.S. Zuchowski, T.R. Bednar, D.W. Stout, S.W. Gould, and J.M. Cohn, "Managing Power and Performance for System-on-Chip Designs Using Voltage Islands," IEEE/ACM Int. Conf. Computer-Aided Design, pp. 195–202, Nov. 2002.

50) J. Tschanz, S. Narendra, Y. Yibin, B. Bloechel, S. Borkar, and V. De, "DynamicSleep Transistor and Body Bias for Active Leakage Power Control of Microprocessors," IEEE Int. Solid-State Circuits Conf., vol. 1, pp. 102–481, Feb. 2003.

51) Q. Wu, M. Pedram, and X. Wu, "Clock-gating and Its Application to Low Power Design of Sequential Circuits," IEEE Custom Integrated Circuits Conf., pp. 479– 482, May 1997.

52) M. Pedram, “Power Minimization in IC Design: Principles and Applications,” ACM Tran. Design Automation, vol. 1, no. 1, pp. 3–56, Jan. 1996.

53) G. Friedman, “Clock Distribution Design in VLSI Dircuits: An Overview,” Proc.IEEE ISCAS, San Jose, CA, pp. 1475–1478, May 1994.

54) Q. Wu, M. Pedram, and X. Wu, “Clock-Gating and Its Application to Low Power Design of Sequential Circuits,” Proc. IEEE Custom Integrated Circuits Conf., vol 47, pp. 415-420, May 2000.

55) “International technology roadmap for semiconductors”, in http://www.itrs.net/reports.html, 2006.

56) Yongpan Liu; Huazhong Yang; Dick, R.P.; Hui Wang; Li Shang, "Thermal vs Energy Optimization for DVFS-Enabled Processors in Embedded Systems," Quality Electronic Design, 2007. ISQED '07. 8th International Symposium on , vol., no., pp.204,209, 26-28 March 2007.

57) Cheng, W.H.; Baas, B.M., "Dynamic voltage and frequency scaling circuits with two supply voltages," Circuits and Systems, 2008. ISCAS 2008. IEEE International Symposium on , vol., no., pp.1236,1239, 18-21 May 2008

58) Murray, J.; Tang, N.; Pande, P.P.; Deukhyoun Heo; Shirazi, B.A., "DVFS Pruning for Wireless NoC Architectures," Design & Test, IEEE , vol.32, no.2, pp.29,38, April 2015

59) Li Shang; Li-Shiuan Peh; Jha, N.K., "Dynamic voltage scaling with links for power optimization of interconnection networks," High-Performance Computer Architecture, 2003. HPCA-9 2003. Proceedings. The Ninth International Symposium on , vol., no., pp.91,102, 8-12 Feb. 2003