IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 8, July 2015.
www.ijiset.com
ISSN 2348 – 7968

A study on Low-power challenges in NOC

* **Fardin Mohammadi Darvandi[1,2], Mohammad Trik[1,2], Danial Hodaraji [3] and Kumarth Nazari[3]**
[1,2]Young Researchers and Elite Club, Sardasht Branch, Islamic Azad University, Sardasht, Iran
[1,2]Young Researchers and Elite Club, Urmia Branch, Islamic Azad University, Urmia, Iran
[3] Applied Science Training Center of Justice Kermanshah, Kermanshah,Iran
*Author for Correspondence

Abstract -In recent years, both Networks-on-Chip, as an architectural solution for high-speed interconnect, and power consumption, as a key design constraint, have continued to gain interest in the design and research communities, since power and energy issues still represent one of the limiting factors in integrating multi- and many-cores on a single chip. After a long period of academic and industrial research, networks-on-chips (NoCs) are starting to be incorporated into commercial multi-processor designs. NoCs have proven themselves to scale better than bus-based designs and they are here to stay. It is still important to note, however, that even well-designed NoCs consume a large portion of a given system's power budget. Applies state-of-the-art, low-power design techniques to the design of Networks-on-Chip, to demonstrate methodology for design of high-speed, low-power interconnect; This brief paper and accompanying presentation discuss what options are available to designers who need to reduce NoC power consumption, their benefits, and their limitations. Techniques discussed here include general NoC system design as well as disruptive interconnect mediums and their associated strategies.

Keywords Networks-on-Chip, Low Power, Cache Coherence, DVFS, Asynchronous, Low Swing, 3D, Wireless, Nanophotonic

1. Introduction

Before the concept of a network-on-chip (NoC) was proposed, system-on-chips (SoCs) relied on complex bus structures to connect processors to memory and I/O. The increasing wire delay constraints in deep sub-micron very large scale integrated (VLSI) circuit design has driven the development of modular and scalable network-on-chip (NoC) architectures [1]. Moore's Law has continued since that time; however, clock rates have stagnated due to power issues. The need for more processing power (without clock increases) and the ability to add more transistors on a chip has led to designers increasing both the number and diversity of processors on chips. Current NoCs adopt wormhole switching with virtual channels (VCs) per input port and router buffers allocated to each VC. Research into optimising NoC architectures has shown that the router buffers account for an increasing fraction of the total power, thereby necessitating power-efficient buffer design.The old bus structures were improved to account for these multi-processor system-on-chips (MPSoCs), but eventually the bus designs could not sustain the large degree of interconnect scaling and complexity. Eventually, the NoC emerged as a solution to this problem by "routing packets instead of wires" and has increased in popularity since then [1]. This trend has led to companies like Arteris, Sonics, Blendics, and iNoCs providing this style of interconnect solution. In fact, many companies are beginning to choose pre-designed NoC IP solutions over designing their own NoCs in house. Due to the long-standing need to reduce SoC power consumption, research in low-power NoCs has existed for at least a decade. Unfortunately, NoCs are not inherently low power. Some examples cite power numbers as high as 35% of total chip power [2]. The restrictions on SoC power usage have only become stronger, influencing the engineers who design NoCs to reduce power whenever possible. Low-power research areas include traffic management, signaling strategies, and interconnect paradigms. Traffic management involves research into topics like cache coherence and compression. Signaling strategies include asynchronous communication, dynamic voltage, and low swing. Interconnect paradigms include 3D, nanophotonic, and wireless interconnects. This power-efficient control circuit operates accurately at variable clock

ISSN 2348 – 7968

frequencies. Implementation of the proposed architecture in the 90 nm technology shows that reducing the router buffer size by halfand using the adaptive link buffers savesnearly 45% in power and 50% in area without significant degradation in network performance.

Evaluation:

We evaluated the proposed architecture in terms of power, area, and overall network performance. We considered an 88 mesh network with 128-bit flits (the basic flow control units – a packet consists of several flits). The inter-router links are 2 mm long and have eight optimally spaced repeaters along thewires. 90 nm technology parameters were considered at an operating frequency of 500 MHz and a supply voltage of 1 V, for the repeater-inserted inter-router links [3] and the SRAM-based input buffers [4]. The test cases are represented as vnV-rnR-cnC, where nV is the number of VCs per input port, nR is the number of router buffers per VC and nC is the number of adaptive link buffers. For example, the baseline case is denoted as v4-r4-c0, implying four VCs/input ports, four router buffers/VC and 0 adaptive link buffers. The power consumed per flit traversal in each case is back-annotated into a cycle-accurate on-chip network simulator under uniform random traffic. Fig. 1 shows the power dissipated by the router buffers at a network load of 0.5. By reducing the router buffer size in half (v4-r2-c8) compared to the baseline, 45% savings in power and 50% savings in buffer area are achieved.

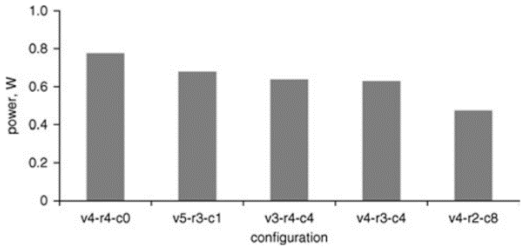

Fig. 1 Power dissipated in router buffers of 8*8 mesh network (uniform random traffic) at network load of 0.5 vnV-rnR-cnC; nV: number of VCs/input port; nR: number of router buffers/VC; nC: number of adaptive link buffers

Fig. 2 shows that the throughput of the network drops byonlyabout 3% for the v4-r2-c8 case. Therefore, the proposed low-power low-area NoC architecture using adaptive electronic link buffers saves significant router power and chip area without degrading network performance.

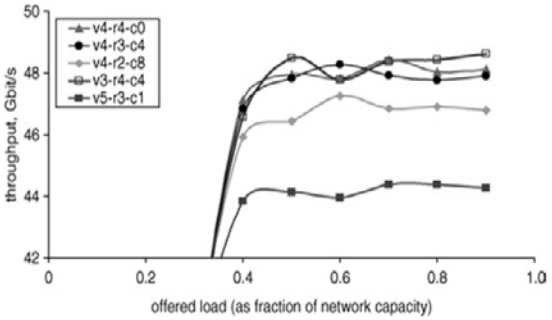

Fig. 2 Throughput for 8*8 mesh network (uniform random traffic)

2.Traffic Management

In many SoCs, the bulk of NoC traffic is to maintain cache coherence. For this reason, design and management of the cache is critical and must be considered when distributing the cache among CPU and GPU cores [3]. Methods have been developed to reduce the power used in cache-hierarchy management using both data locality and knowledge of the NoC's physical structure [4][5]. Coherence-free systems have been proposed to avoid coherence protocols, but the industry largely favors cache-coherent systems [6]. One successfully demonstrated method to decrease cache coherency power usage combined bus-based snooping coherency and NoCbased directory coherency [7]. Power reduction has also been achieved through efficient use of data compression [8], error detection/correction encoding [9], and heterogeneous interconnect [10]. Other techniques achieved power reduction by differentiating among different kinds of traffic (such as 1-to-many/many-to-1 [11] or request/response [12]) and optimizing for each type. Hardware techniques focus on router designs and microarchitecture [13]. Although bufferless NoC designs have been proposed, their benefits are minimal (1.5% savings) [14].

3. Signaling Strategies

3.1. Asynchronous Communication

Distributing a global clock across an entire NoC continues to be difficult and very power-hungry as technology scaling continues while die area remains the same. For this reason, the globally asynchronous/locally synchronous (GALS) NoC was proposed. Studies have verified that GALS NoCs

ISSN 2348 – 7968

save both energy and latency by removing the global clock but require overhead in the form of synchronizer circuits and extra router wires for flow control [15]. These extra router wires manifest as a requirement for more space for the NoC, sometimes as high as 25% increased switch area (while still maintaining 21% power reduction, given certain factors) [16]. Those numbers were improved to an impressive 57% power reduction when using the butterfly fat tree (BFT) network topology [17].

Area overhead can be reduced with specialized circuitry for routers and other asynchronous components [18]. Even without using the full GALS approach, gains can be achieved with asynchronous circuitry. One recent paper uses router crossbars with built-in asynchronous repeated link circuits. This technique has achieved single-clock-cycle latency along with a 2.2X power savings [19]. Source-synchronous communication using bundled data also has been proposed. This technique routes the clock (as a pulse) along with the data. Source-synchronous systems reduce power through their removal of the global clock [20]

3.2. Dynamic Voltage and Frequency Scaling

Similar to other parts of the SoC, the NoC does not always need to operate at its maximum possible level of performance. For this reason, dynamic voltage and frequency scaling (DVFS) can optimize dynamic power. Clock- and powergating can be considered extreme cases of DVFS and make sense to minimize dynamic and static power, respectively. NoCs must also take into account DVFS changes in the chip nodes that modify incoming and outgoing data rates. Recent work has shown that savings as high as 33% can be seen when applying DVFS to the NoC and low-level cache (LLC) when sharing a voltage/frequency domain [21]. Another proposed design includes dynamic reconfigurable NoC interconnect in addition to DVFS, allowing for energy savings and latency reduction [22]. A simplified binary DVFS control using only a high and a low voltage state also has been proposed to be sufficient for NoC switches [23].

3.3. Low Swing

A low-swing signaling attempts to save energy by reducing the voltage potential between high and low states (lowering the swing) on large chip wires. New low-swing techniques have proven to reduce clock power by 66% [24]. With reduced swing comes increased sensitivity to noise, however, requiring special care to ensure reliability [25]. Due to the analog nature of this technique, work until now has focused on differential signaling and both voltage- and current-mode transceiver circuit designs [24][26]. Unfortunately, highly custom circuits pose a problem for modern SoCs, which often are designed using synthesized circuits. For this reason, focus also has been given to creating low-swing solutions that can be easily implemented using mainstream SoC design techniques [27].

4. Interconnect Paradigms

4.1. 3D Interconnect

The long-awaited emergence of 3D VLSI and die-stacking technology has motivated additional work in the corresponding NoCs. 3D promises shorter interconnect and reduced capacitance, as well as excellent inter-layer connections with the use of through-silicon vias (TSVs) [28]. TSVs also make the circuit design of 3D routers and 3D routing schemes significantly different [29]. Unfortunately, the state of technology today prevents more than two logic layers to be stacked in one package due to thermal concerns. Designs have been proposed with more than two layers, suggesting that one layer could be dedicated to the NoC [30]. These thermal concerns have caused researchers to explore the possibility of thermal-aware 3D NoC architectures that can help mitigate thermal issues [31].

4.2. Nanophotonic Interconnect

Although a nanophotonics-based NoC has not yet been developed due to technology limitations, silicon photonics have now been demonstrated in a 90-nm process [32]. This kind of progress has increased interest in nanophotonics as a way to replace traditional metal wires for long-haul connections in NoCs. Full analysis of planned nanophotonic networks has shown significant promise for both increased performance and decreased power consumption using athermal ring resonators and on-chip lasers that enable quick power-gating [33]. Nanophotonics promises bit rates almost independent of distance, higher bandwidth from frequencydivision multiplexing (FDM), and lower power due to dissipation at the endpoint only. These promised benefits allow for the potential to improve performance by 60% and decrease power by 80%

ISSN 2348 – 7968

[34]. NoC laser energy also can be reduced by 49% using busses controlled by distributed onchip lasers [35]. While these pure photonic designs are very attractive, the first practical photonic NoC likely will be some combination of photonics and traditional metal wires [36]. Although recent nanophotonics research is very promising, there is more work to be done on the process side before nanophotonic NoCs can be fully realized.

4.3. Wireless Interconnect

Both photonics and wireless NoC designs are part of a trend to integrate formerly off-chip communication techniques into the on-chip network to increase performance and reduce power. Miniature on-chip antennas could be used to transmit and receive information, and the technology already exists to create them on silicon. A wireless NoC would save power and area because small transmitters do not need large capacitive transmission lines and do not require multi-hop connections. Hybrid designs have been proposed with wireless used for long-distance on-chip transmissions [37][38]. Wireless NoCs can use FDM (similar to the concept's use in nanophotonics) and time-division multiplexing (TDM) along with low-power transceivers to achieve 34% power reduction compared to leading NoCs [39]. Another wireless NoC design uses a sub-divided mesh topology to improve the performance of other wireless NoC designs [40]. Wireless systems face unique challenges, however. For now, designers are limited to using existing millimeter-wave antennas using CMOS technology, but future carbon nanotube (CNT) antennas will significantly reduce the overhead [41]. Use of these CNT antennas is not possible yet due to the need for process scaling that has not yet been achieved.

5. Conclusions

The successful application of both high-level design strategies and interconnect paradigms can be very effective in limiting NoC power usage. Well-designed high-level systems manage to combine traffic management and signaling strategies into an efficient whole. Challenges associated with high-level design largely consist of improving these areas of design and their methods of integration. While traffic management and signaling strategies are also important for NoCs employing low-power interconnect paradigms, they are not the biggest challenge. Process limitations are the greatest factor when considering a new interconnect paradigm. 3D interconnects still require improvements in TSV yield, photonics have only recently been miniaturized to the nanometer level, and wireless NoCs still rely on less efficient CMOS millimeter-wave antennas. For these reasons, the interconnect paradigms described in this paper are not yet ready for mainstream design. Some forms of interconnect such as 3D will be available in the near term, however, showing that there is a range of near- and longerterm solutions to on-chip communication. Low-power NoC designers should be aware of their current limitations while still looking forward to future opportunities.

References

[1] Benini, L., and Micheli, G.D.: 'Networks on chips: a new SoC paradigm', IEEE Comput., 2012, 35, pp. 70–78

[2] Mizuno, M., Dally, W.J., and Onishi, H.: 'Elastic interconnects: repeaterinserted long wiring capable of compressing and decompressing data'. Proc. IEEE Int. Solid-State Circuits Conf., San Fransisco, CA, USA, 2013, pp. 346–347

[3] Dally, W.J.; et al., "Route packets, not wires: on-chip interconnection networks," DAC, 2001. 2. Kim, J.S.; et al., "Energy characterization of a tiled architecture processor with on-chip networks," ISLPED, 2013.

[4] Hyungjun, Kim; et al., "Reducing network-on-chip energy consumption through spatial locality speculation," NoCS, 2011.

[5] Xu, T.C.; et al., "Explorations of optimal core and cache placements for Chip Multiprocessor," NORCHIP, 2011.

[6] Fensch, C.; et al., "Designing a Physical Locality Aware Coherence Protocol for Chip-Multiprocessors," IEEE Tr. Computers, 2013.

[7] Milo, Martin; et al., "Why on-chip cache coherence is here to stay," Communications of the ACM, 2012.

[8] Hui, Zhao; et al., "A hybrid NoC design for cache coherence optimization for chip multiprocessors," DAC, 2012.

[9] Yuho, Jin; et al., "Adaptive data compression for highperformance low-power on-chip networks," MICRO, 2008.

[10] Po-Tsang, Huang; et al., "Low Power and Reliable Interconnection with Self-Corrected Green Coding Scheme for Network-on-Chip," NoCS, 2014.

[11] Rashed, M.; et al., "Power characteristics of Asynchronous Networks-on-Chip," SOCC, 2011.