ELectronics EXpress

LETTER

IEICE Electronics Express, Vol.9, No.24, 1887–1892

Leakage-aware adaptive routing for pipelined on-chip networks in ultra-deep sub-micron technologies

Seongmin Jo[a)] **and Yong Ho Song**

Department of Electronics and Communication Engineering, Hanyang University, 17 Haengdang-dong, Seongdong-gu, Seoul, 133–791, Republic of Korea

a) *j.seongmin@gmail.com*

Abstract: As semiconductor process technology continues to scale down to the ultra-deep sub-micron level, leakage power becomes a critical design constraint for on-chip networks (OCNs). Power gating is widely used to reduce the OCN leakage power; however, it does not work well with adaptive routing owing to its aggressive use of free links and router buffers to achieve high performance. In this paper, a novel leakage-aware adaptive routing algorithm to increase the power-gating effect by routing packets with minimal link activation is proposed. Experimental results show that the proposed algorithm effectively achieves a reduction in the overall network leakage power of up to 11.6% greater than the conventional adaptive routing algorithm, with a little sacrificing network bandwidth.

Keywords: leakage power, power-gating, on-chip network, adaptive routing

Classification: Integrated circuits

References

[1] J. Howard, et al., "A 48-core IA-32 message-passing processor with DVFS in 45 nm CMOS," *Proc. ISSCC*, pp. 108–109, Feb. 2010.

[2] B. Li, et al., "Orion 2.0: a power-aware simulator for interconnection networks," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 20, no. 1, pp. 191–196, Jan. 2012.

[3] H. Matsutani, et al., "Ultra fine-grained run-time power gating of on-chip routers for CMPs," *Proc. Int. Symp. NOCS*, pp. 61–68, May 2010.

[4] W. J. Dally and B. Towles, *Principles and practices of interconnection networks*, Morgan Kaufmann Publishers Inc., 2003.

[5] A. Kumar, et al., "A 4.6 Tbits/s 3.6 GHz single-cycle NoC router with a novel switch allocator in 65 nm CMOS," *Proc. Int. Conf. Computer Design*, pp. 63–70, Oct. 2007.

DOI: 10.1587/elex.9.1887
Received November 06, 2012
Accepted November 29, 2012
Published December 28, 2012

IEICE Electronics Express, Vol.9, No.24, 1887–1892

1 Introduction

Pipelined on-chip networks (OCNs) are widely used in multi-processor systems-on-chip (MP SoCs) to extensively connect processing cores and hardware intellectual properties (IPs) [1]. With continuously increasing demands for communication bandwidth [2], achieving high OCN performance has become a critical avenue of research. Adaptive routing, which distributes multiple packets destined for the same node over different pathways to avoid network congestion, is one method for providing higher network bandwidth.

Because power consumption has become the most important design constraint affecting heat dissipation, packaging, and operating frequency, another area of research has been the development of low-power OCNs. Power consumption can be divided into two categories: dynamic power and static leakage power. Although dynamic power is still the most significant portion of the overall OCN power consumption, static leakage power continues to increase as process technologies are scaled down to the ultra-deep sub-micron (UDSM) level [2].

Power-gating techniques are widely used to cope with increases in SoC power leakage [3]. Power-gating cuts off supply power to unused components so that they cannot leak power while inactive. In OCNs, this technique can be effectively adopted by turning off unused links and buffers, which can consume a significant power load.

Although a number of studies [3] have proposed power-gating techniques for OCNs, these have only considered deterministic routing. Power gating is also significant to OCNs using adaptive routing because this requires the activation of more links and buffers to increase the network bandwidth. In this paper, a novel, leakage-aware adaptive routing (LAAR) algorithm to increase the effects of power gating by selecting a path that minimally activates links is proposed.

2 Motivation

There are two challenges in applying power gating to OCNs with adaptive routing. First, the effects of power gating in adaptive routing tend to be less than those in deterministic routing because in adaptive routing, there is a tendency to use as many links as possible to forward packets in order to avoid temporal network congestion. Second, a link's wake-up delay will increase router delay, thus degrading the overall network bandwidth. In deterministic routing, a look-ahead method [3] is used to hide wake-up delay. This method, however, is difficult to use in adaptive routing, where routers require prior knowledge of the congestion states of two-hop-away routers.

In order to overcome these challenges, the proposed LAAR algorithm favors the use of active links over power-gated links in routing decisions. Fig. 1, which shows two adaptive routing examples on a power-gated OCN, illustrates the effects of reusing active links on leakage power. The left-hand side of Fig. 1 shows network states, with active links shown in bold, whereas the right-hand side shows timing diagrams for two packet transfers. These

DOI: 10.1587/elex.9.1887
Received November 06, 2012
Accepted November 29, 2012
Published December 28, 2012

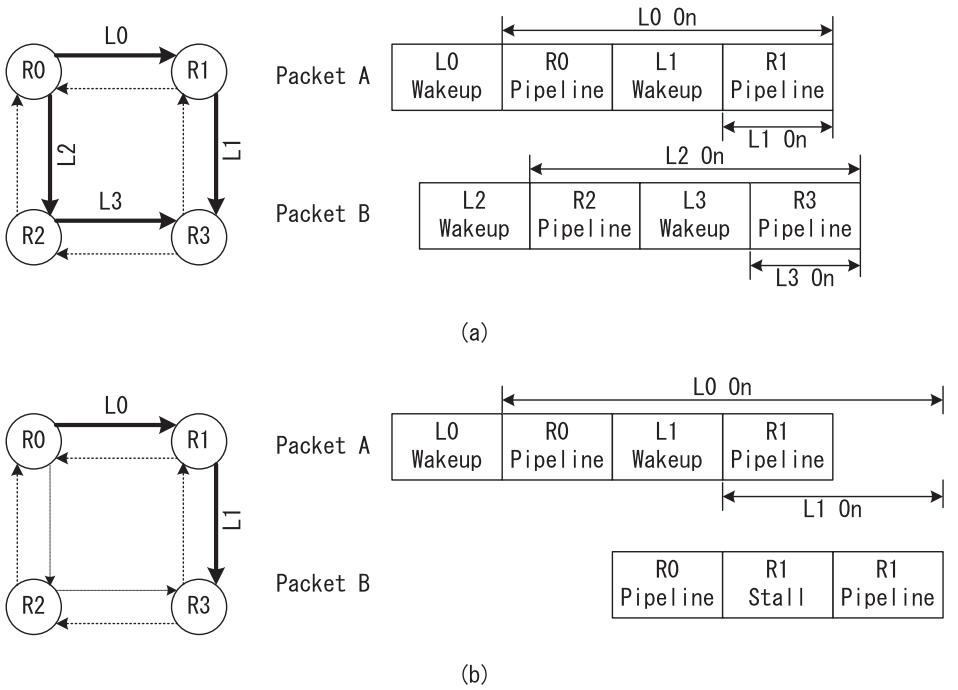

Fig. 1. Examples of routing scenarios using adaptive routing algorithms on a network with power gating: (a) Congestion-Aware Adaptive Routing (CAAR) and (b) Leakage-Aware Adaptive Routing (LAAR)

examples assume (1) that packets A and B, destined for router 3 (R3), are buffered in different virtual channels (VCs) of R0; (2) that the wake-up and pipeline latencies are the same; and (3) that there is no congestion in the network.

Fig. 1 (a) shows a typical example of congestion-aware adaptive routing (CAAR) [4], in which two packets are routed along different paths. First, R0 activates L0 and L2 and then transfers two packets along the router pipeline. Next, R1 and R2 activate L1 and L3 in order to deliver the packets. L0 and L2 remain active until R0 receives credits back for the transfer of the packets. The LAAR behavior is shown in Fig. 1 (b). Here, to transmit packet B, R0 reuses L0, which has already been activated by packet A. Since two packets are transferred over the same path, the later packet is delayed by one pipeline cycle owing to link contention. However, because the routing protocol uses state-of-the-art router micro-architectures [5] with single-cycle pipeline latency, the delay in packet transfer contributes minimally to network bandwidth degradation.

LAAR has an additional advantage in terms of link leakage power consumption. As shown in Fig. 1 (a), the overall leakage power of two packets by CAAR, P_{CAAR}, can be calculated using Eq. (1):

$$P_{CAAR} = P_L(6T_P + 2T_W) \tag{1}$$

where P_L is the leakage power consumption per clock cycle, T_P is the router pipeline latency, and T_W is the wake-up latency of the power-gated link. The

DOI: 10.1587/elex.9.1887
Received November 06, 2012
Accepted November 29, 2012
Published December 28, 2012

overall leakage power of two packets routed using LAAR, P_{LAAR}, is less than that consumed by CAAR, as the Eq. (2):

$$P_{LAAR} = P_L(5T_P + T_W) \tag{2}$$

The reuse of active links is effective in reducing the leakage power consumption for two reasons. First, this allows for the elimination of wake-up latency in later packets. Second, as links must be active until the upstream router receives a credit back, the pipeline cycles of the following packet will remain hidden (Fig. 1 (b)).

This method, however, encourages packets to concentrate on active links even during times of network congestion, which degrades network bandwidth. Since this is inconsistent with the original objective of adaptive routing, LAAR must avoid network congestion by reusing active links under specific conditions, as described in the next section.

3 Leakage-aware adaptive routing

Fig. 2 shows a flow chart of the proposed LAAR algorithm. LAAR is based on minimal adaptive routing [4], which returns at most two candidate output ports on the basis of the current and destination coordinates. To select an output port among these candidates, LAAR refers to the link activation status of the current router. If one of the candidates is connected to an active link, then the output port that uses that link is given higher priority than the other candidate.

If all of the candidates are connected to inactive links, new link activation is unavoidable, and LAAR must decide how to find an output port that

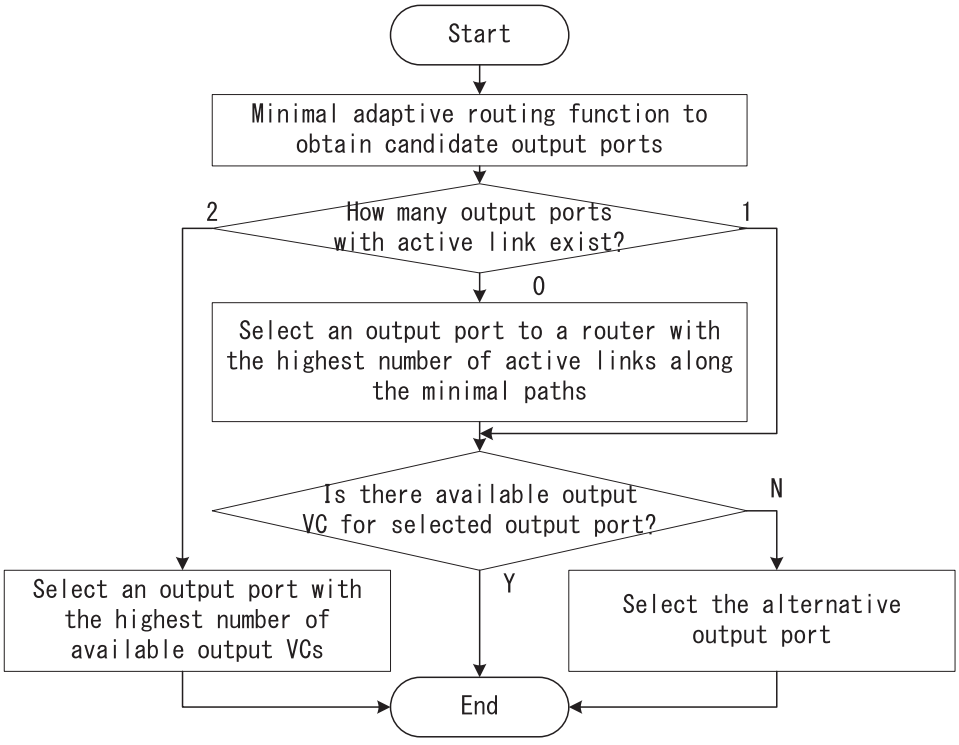

Fig. 2. Flow chart of proposed LAAR algorithm

DOI: 10.1587/elex.9.1887
Received November 06, 2012
Accepted November 29, 2012
Published December 28, 2012

requires activating the minimum possible number of links on the network. In order to achieve this, the routers must know the global link activation status, which may involve considerable overhead. Therefore, LAAR will select output ports using only the link status of neighboring routers. Because each router distributes its link activation status to its neighboring routers through sideband signals, LAAR can use these signals to determine the number of active links in the neighboring routers along the minimal paths in order to prioritize among output ports connecting to the router with the highest number of active links.

As mentioned above, the use of congestion avoidance to achieve high network bandwidth is an important role for adaptive routing. To facilitate this, LAAR examines the congestion status of the neighboring routers, checking to see if there is an output VC available for the port selected in the previous phase. If a VC is available, LAAR will use the selected output port; otherwise, an alternative output port will be selected.

When all the candidates have become active through this process, LAAR stops checking the activation statuses of the neighboring routers and instead begins to check their congestion statuses, as the use of neighboring routers' link activation statuses would at this point degrade network bandwidth by concentrating traffic onto the router with the highest number of active links.

4 Experimental results

In order to estimate and comparatively evaluate leakage power consumption caused by the LAAR algorithm, a cycle-accurate OCN simulator [5] incorporating an Orion 2.0 model [2] with a power-gating extension was used. CAAR and LAAR algorithms were evaluated on an 8×8 2-D mesh network with four 8-depth VCs per port. A power-gating technique was used to test both routing algorithms.

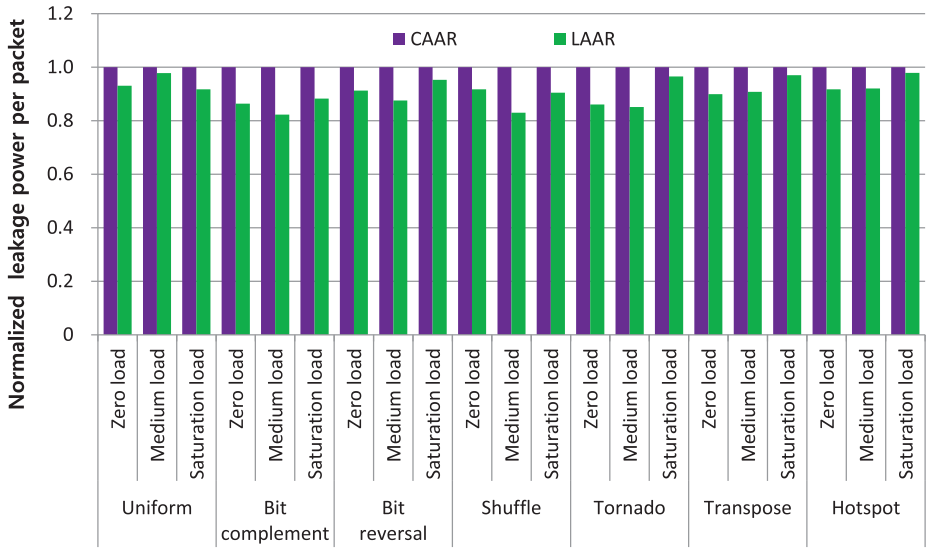

Fig. 3. Normalized leakage power per packet of CAAR and LAAR under seven representative traffic patterns

DOI: 10.1587/elex.9.1887
Received November 06, 2012
Accepted November 29, 2012
Published December 28, 2012

Fig. 3 compares the two algorithms in terms of their overall network leakage power under seven traffic patterns. In each case, the results are normalized to the leakage power of the CAAR algorithm. Leakage power was measured at three traffic levels: zero load, medium load (the halfway point between zero load and saturation load), and saturation load. From Fig. 3, it can be seen that using LAAR effectively reduced the network leakage power on an average by 10.0%, 11.6%, and 6.1% at zero load, medium load, and saturation load, respectively. At medium load, sufficient network traffic was provided to allow reuse of active links, and the network was not heavily congested, allowing parts of links to be continuously turned off by reusing other parts of the active links; as a result of these factors, LAAR performed best at this load. The LAAR algorithm only incurred about 2% penalty on the achievable network bandwidth with CAAR, demonstrating that LAAR was still capable of using multiple paths even under network congestion.

5 Conclusion

In this paper, a novel leakage-aware adaptive routing algorithm that enhances the effectiveness of power-gating techniques in pipelined OCNs was proposed. The experimental results described here show that the proposed LAAR algorithm effectively reduces leakage power by routing packets onto paths with minimal link activation.

Acknowledgments

This work was supported by Industrial Strategic Technology Development Program funded by the Ministry of Knowledge Economy (MKE, Korea) (10039188, Development of multimedia convergence programmable platform for smart vehicles).

DOI: 10.1587/elex.9.1887
Received November 06, 2012
Accepted November 29, 2012
Published December 28, 2012