The handle http://hdl.handle.net/1887/85165 holds various files of this Leiden University dissertation.

Author: Wang, P.
Title: On the power efficiency, low latency, and quality of service in network-on-chip
Issue Date: 2020-02-12
Chapter 7

Summary and Conclusion

With low network latency, high bandwidth, good scalability, and reusability, a Network-on-Chip is a promising communication fabric for the future many-core systems. However, NoCs consume too much power in real chips, which constraints the utilization of NoCs in future large-scale many-core systems. Meanwhile, with more advanced semiconductor technologies, applied in chip manufacturing, the static power consumption takes a larger proportion of the total power consumption. Thus, in this thesis, we have focused our attention on reducing the static power consumption of NoCs in two directions: applying efficient power gating on NoCs to reduce the static power consumption and realizing a confined-interference communication on a simplified NoC infrastructure to achieve energy-efficient packet transmission.

By powering off the idle components/routers in a NoC, power gating is an effective way to reduce the power consumption of a NoC. However, when the power gating is applied on a NoC, the powered-off components/routers block the packet transmission and cause significant packet latency increase. This is because the powered-off components/routers need some clock cycles to be fully charged (i.e., to be powered-on). During the time period of charging powered-off routers, some packets cannot be transferred and have to be blocked until the powered-off routers are fully charged. As a consequence, applying power gating on a NoC causes significant packet latency increase. Furthermore, the power gating process (i.e., switching off/on the power of components/routers) itself consumes extra power. This implies that frequent power gating or power gating in a short time may cause more power consumption or inefficient power consumption reduction. Thus, to reduce the packet latency increase caused by power gating and achieve significant reduction of the power consumption in NoCs, we have proposed three novel power gating approaches: duty buffer based (DB-based) power gating, dynamic bypass (D-bypass) power gating, and express virtual channel based (EVC-based) power gating. These power gating approaches are...
effective in reducing the power consumption of NoCs, but with different properties, they have different advantages. We summarize the properties of the DB-based power gating approach (DB_PG), the D-bypass power gating (D-bypass), and the EVC-based power gating approach (EVC_PG) in Figure 7.1. In Figure 7.1, the axes PL_l, PL_m, and PL_h represent the packet latency (PL) in a NoC under low traffic workloads (l), medium traffic workloads (m), and high traffic workloads (h), respectively. The axes PC_l, PC_m, and PC_h represent the power consumption (PC) of a NoC under low traffic workloads (l), medium traffic workloads (h), and high traffic workloads (h), respectively. For example, the PL_m axis crosses the block edges of DB_PG, D-bypass, and EVC_PG at three points, respectively. These points represent the packet latency (normalized to the same baseline) of DB_PG, D-bypass, and EVC_PG under medium traffic workloads. Thus, according to Figure 7.1, under medium traffic workloads, DB_PG has the highest packet latency among our three approaches, whereas EVC_PG has the lowest packet latency. Based on the different properties of our power gating approaches, shown in Figure 7.1, we draw the following conclusions:

- **Our DB-based power gating approach is effective in reducing the power consumption of a NoC in a wide range of traffic workloads, but at medium traffic workloads, it has the highest packet latency among our three power gating approaches.** This is because, our DB-based power gating is a fine-grained power gating approach, in which each input port of a router can be

![Figure 7.1: Packet latency (PL) and power consumption (PC) at low traffic workloads (l), medium traffic workloads (m), and high traffic workloads (h).](image-url)
separately powered-off. In this way, our DB-based power gating approach can fully utilize the idle time of each input port in a router to reduce the static power consumption. Thus, at different traffic workloads, our DB-based power gating approach achieves significant reduction of the power consumption in a NoC. Furthermore, taking advantage of our novel duty buffer (BD) structure to replace the powered-off input port to transfer packets, our DB-based power gating approach achieves lower packet latency than D-bypass at low traffic workloads as shown Figure 7.1. However, being a fine-grained power gating approach, our DB-based power gating approach needs to separately switch the power of each input port in a router. At medium traffic workloads, packets experience many power gating processes. As a consequence, our DB-based power gating approach has the highest packet latency among our three approaches at medium traffic workloads.

- At low traffic workloads, our D-bypass power gating is the most power-efficient approach among our three approaches, and it is effective in reducing the power consumption of a NoC only at low traffic workloads. However, at low traffic workloads, our D-bypass power gating has the highest packet latency among our approaches. This is because, in our D-bypass power gating approach, we add one special hardware bypass structure in each router. When a router is powered-off, only this special hardware bypass structure is kept powered-on. Compared with the DB-based power gating approach and the EVC-based power gating approach, our D-bypass power gating approach can power off more components in a router to reduce the static power consumption. Thus, at low traffic workloads, in which most of the routers are idle and can be powered-off, our D-bypass power gating approach consumes the least power among our three approaches. Furthermore, the special hardware bypass structure in each router makes it possible for packets to bypass powered-off routers. In this way, our D-bypass power gating approach can efficiently reduce the extra power consumption caused by power gating. However, being a course-grained power gating approach, our D-bypass power gating approach cannot fully utilize the idle time of each component in a router. When the traffic workload increases, most of the routers in a NoC become busy and cannot be powered off to reduce the static power consumption. As a consequence, our D-bypass power gating approach is effective only at low traffic workloads. In terms of the packet latency, as packets can bypass powered-off routers in our D-bypass power gating approach, the packet latency increase caused by power gating is reduced. However, limited by the low transmission capacity of the special hardware bypass structure in powered-off routers, our D-bypass power gating approach still causes significant increase of the packet latency. As a con-
sequence, our D-bypass power gating approach has the highest packet latency among our three approaches at low traffic workloads.

- **Our EVC-based power gating approach achieves the lowest packet latency among our three approaches at different traffic workloads. Furthermore, it is also the most effective approach in reducing the power consumption at high traffic workloads.** This is because, in the EVC-based power gating approach, we pre-define multiple virtual bypass paths between different routers. Packets can take these virtual bypass paths to bypass intermediate routers that can be powered-on or powered-off. Furthermore, compared with the D-bypass power gating approach, the pre-defined virtual bypass paths in our EVC-based power gating approach are much more efficient to allow packets to bypass the powered-on/powered-off routers. Therefore, our EVC-based power gating approach achieves the lowest packet latency among our three power gating approaches. In addition, packets can bypass not only powered-off routers but also they can bypass powered-on routers as well. Thus, even at high traffic workloads, our EVC-based power gating approach still can reduce the power consumption by allowing packets to bypass the powered-on routers.

A confined-interference communication in a NoC-based System-on-Chip is a useful quality-of-service. In confined-interference communication, the packets of different applications are grouped into different domains and packet interference can occur only in the same domain, whereas there is no packet interference between domains. By supporting a confined-interference communication, NoCs can support composability to facilitate the temporal verification of (hard) real-time applications. However, realizing a confined-interference communication on a conventional (virtual channel/buffer based) NoC requires a large number of virtual channels, which causes high power consumption. Therefore, there is an urgent need for realizing a confined-interference communication on a more power-efficient NoC architecture. Bufferless NoCs have simplified NoC architectures. By eliminating virtual channels/buffers in routers, bufferless NoCs consume much less power than conventional NoCs. However, as there are no buffers in bufferless NoCs to temporarily store packets, packets have to keep moving, which makes it more difficult to control the interference between packets. As a consequence, current bufferless NoCs do not support a confined-interference communication.

To overcome this issue, we have proposed a novel routing approach, called Surfing on a Bufferless NoC (Surf-Bless). **Based on our Surf-Bless routing approach, it becomes possible for bufferless NoC to support a confined-interference communication. Furthermore, our Surf-Bless routing approach is much more power/energy-efficient than related approaches.** This is because, our Surf-Bless approach is based
on a specific assignment and scheduling of the resources in a bufferless NoC. This specific assignment and scheduling can be visualized as multiple “waves” which move in space and time over the NoC in a specially designed repetitive pattern. The specially designed repetitive pattern for the waves guarantees that packets “surfing” on a wave can keep moving, which is essential to correctly use a bufferless NoC to transfer packets. This is because, in a bufferless NoC, there are no buffers and packets have to keep moving. Furthermore, the specially designed repetitive pattern also guarantees that there is no interference between different waves. Thus, by assigning different domains on different waves, there is no interference between domains and a confined-interference communication is achieved. In this way, we realize confined-interference communication on a bufferless NoC infrastructure. Furthermore, as the routers in our Surf-Bless approach do not have virtual channels/buffers, our Surf-Bless routing consumes much less power/energy than related approaches.
Bibliography


BIBLIOGRAPHY


List of Publications

First Author Publications


Co-author Publications

