On the power efficiency, low latency, and quality of service in network-on-chip

Wang, P.

Chapter 1

Introduction

MOORE's Law[1] has a profound impact on the semiconductor industry and the processor development. The downscaling of the manufacturing size of transistors has resulted in faster, smaller and more efficient transistors, which has created continuing increase of transistor budgets for designers to realize higher performance processors. In the past, taking the advantage of the new generation of advanced transistors, the designers had been respecting the Dennard's scaling [DGR+74][2] to improve the performance of uni-processors (single-thread) by increasing the clock frequency to speed up uni-processors' computations and by taking more transistors to realize more efficient micro-architectures, such as superscalar, superpipeline, branch prediction and so on. However, this situation came to an end around 2005-2007 [Boh07] because the further increase of the clock frequency caused exponential power consumption increase, while at the same time, the cooling technology did not scale exponentially to handle the temperature rise on the chip. As a consequence, the further increase of the clock frequency causes severe power dissipation problems and becomes impractical on a chip. Furthermore, the more advanced, but more complex micro-architectures will consume more power, which makes it difficult to further improve the performance of the uni-processors. As a result, system designers transfer their attention to build multi/many-core System-on-Chips (SoCs). Instead of using a huge and complex uni-processor, designers use a multi/many-core processor containing multiple, relatively simple processing elements (processing cores) to increase the performance with reasonable power consumption. In such a system, the performance almost linearly scales with the number of processing elements [Bor07]. Thus, without the need for scaling

[1] Moore's law is the observation that the number of transistors in a dense integrated circuit doubles about every two years (or 18 months).

[2] Dennard Scaling postulated that as transistors get smaller their power density stays constant, so that the power use stays in proportion with area. This allowed CPU manufacturers to raise clock frequencies from one generation to the next without significantly increasing the overall circuit power consumption.

up the frequency, performance improvements in many-core processors can be gained by simply increasing the number of processing elements.

At the beginning of the multi/many-core systems' development, bus-based communication infrastructures [HHS+00] ware used to support the communication between a few processing elements. However, as many-core systems constantly increase the number of processing elements to gain higher performance, the traditional bus-based communication infrastructure cannot meet the communication demands of large many-core systems. First, as a single centralized and shared infrastructure, a bus does not support parallel communication between multiple processing elements. The processing elements have to access the bus one at a time, which results in a very limited communication bandwidth. Furthermore, with the downscaling of the manufacturing size of transistors, the parasitic capacitance and signal propagation delay on the wires become significant, which results in that the wires delay becomes higher than the gate delay. Thus, the global wires on a bus have to operate at a much lower frequency than the processing elements. As a consequence, the communication over a bus has a significant delay. Finally, there is a significant amount of capacitance on the global wires, which causes high power consumption. The aforementioned drawbacks of low bandwidth, high wire delay, and high power consumption result in poor scalability of bus-based communication infrastructures. As a consequence, the bus-based communication infrastructures cannot efficiently support the communication between the processing elements of large-scale many-core systems.

In order to overcome the aforementioned drawbacks, researchers [DT01, HJK+00, BDM02] inspired by the concepts of off-chip networks have proposed the idea of Network-on-Chip (NoCs). In contrast to the conventional bus, NoCs utilize a distributed communication infrastructure. Routers as intermediate nodes are arranged on a particular topology and connected by short wires. The processing elements access the NoC though their own network interfaces without the need of any global controller. Thus, the processing elements can independently access the NoC. In other words, the NoCs support parallel communication, which not only enormously increases the communication bandwidth, but also significantly reduces the communication latency. In addition, without long global wires, the shorter wires in a NoC cause much less power consumption and much lower delay. Thus, it is possible to achieve a high frequency communication operation in a NoC. Moreover, a processing element uses the network interface to access the NoC, which completely separates the communication on the NoC and the computation on the processing elements. Thus, when designers develop a new chip, they could simply replace the processing elements and do not need to redesign the NoC. Such high reusability of NoCs is an attractive property for the SoC industry, which experiences a high Time-to-Market pressure. Based on the aforementioned advantages, NoCs become a promising communication fabric for future

Chip	NoC Topology	Technology	Frequency	VC setting	Power Consumption
Intel 80-tile [HVS+07]	8×10 Mesh	65 nm	$1.7 \sim 5$ GHz	2 16-flit VCs	28% (4 GHz)
SCCC [HDH+10, SJJ+11]	4×6 Mesh	45 nm	2 GHz	8 4-flit VCs	10% (1 GHz), 5% (250MHz)
TRIPS [GKS+07]	4×10, 5×5 Mesh	130nm	336 MHz	4 2-flit VCs, 1 4-flit VC	-
SCORPIO [DCS+14]	6×6 Mesh	45 nm	833 MHz	4 1-flit VCs, 2 3-flit VCs	19%
Title-64 [BEA+08]	8×8	90 nm	750 MHz	-	-
RAW [TKM+02, KTMW03]	4×4 Mesh	0.15um	420 MHz	4 4-flit VCs	36%

Table 1.1: NoCs on real chips.

large-scale many-core systems.

1.1 Design Trends in Network-on-Chip

Even though the idea of NoCs comes from the off-chip networks, the design constraints on the chip are different from the off-chip. For example, NoCs are power and hardware limited while networks off chip are often pin-bandwidth limited. Such different design constraints result in the NoCs forming their own design trends.

1.1.1 Low Power Consumption

The power consumption is a critical design constraint on a chip. Even though NoCs consume relatively lower power than the conventional bus-based communication infrastructures in large-scale many-core systems, the NoCs power consumption is still the major factor limiting the design and utilization of NoCs. Let us consider some NoCs implemented on real chips and shown in Table 1.1. Most of the NoCs consume significant power on the chip. For example, the NoC in the Intel's 80-tile chip [HVS+07] consumes 28% of the total power of the whole chip under 65 nm technology. Even, with the more power-efficient technology of 45 nm, the NoCs in SCCC [HDH+10, SJJ+11] and SCORPIO [DCS+14] consume 10% and 19% of the total power of the chip, respectively. Furthermore, with the NoC size increasing to connect more processing elements, the NoCs will consume much more power. According to the prediction in [Bor07], the power consumption of a NoC will reach 100 watts when the NoC size further increases to connect 1000 cores. Such high power consumption is unacceptable for the future many-core systems and will become the major factor limiting the utilization of a NoC. Therefore, it is critical and necessary to reduce the power consumption of NoCs.

The power/energy consumption of NoCs can be reduced in different ways. A straightforward way is to simplify the NoCs structure. For example, the complex crossbar in a router can be simplified into a few gates [BKA10, ZML+16] or the buffers in routers can be eliminated [MM09, LSMJ16]. By simplifying the NoCs structure, the NoCs need less hardware and consume less power. However, simpli-

fying the NoCs structure may cause other problems, such as only very limited power reduction [BKA10], poor throughput [MM09], losing packets [LSMJ16], etc.

Applying low power technologies on a NoC is another, more effective way to reduce the NoCs' power consumption. In a NoC with dynamic voltage and frequency scaling (DVFS), when the workload is low, the voltage-frequency (V/F) regulators let the routers work at a low voltage to reduce the power consumption. As the V/F regulators consume extra power, it is necessary to trade off between the power reduction and the extra power consumption caused by the V/F regulators. In practice, to use fewer V/F regulators to avoid the extra power consumption of the V/F regulators, a group of routers [YL16, LY18] or all the routers [BJS+14] share one V/F regulators. Compared with the aforementioned DVFS technique, the implementation of power gating on a NoC is much simpler and low cost power reduction technique. Power gating can be more flexibly applied on a NoC at different granularities. For example, in fine-grained power gating approaches, each component of a router [MKI+10] can be separately powered off. In course-grained power gating approaches, an entire router [WDLW17] or a group of routers [DNSD13] can be powered off by the same power gating controller. Furthermore, as NoCs have the characteristics of a distributed structure, a naturally unbalanced traffic workload, and a low average injection traffic rate, the static power consumption takes a high proportion of the total power consumption. Thus, power gating is an applicable and effective way to reduce the static power consumption of a NoC.

1.1.2 Low Latency

The network latency of a NoC has a significant impact on the system performance. For example, NoCs on general purpose processors should have a low network latency to realize the cache coherence protocol, which needs a low latency communication to synchronize the data located at different places in a very short time.

Many approaches are widely used to reduce the network latency. One approach is to reduce the blocking/stall or congestion, by using multiple virtual channels to reduce the Head-of-Line blocking [DT04] or by using efficient flow control to reduce the traffic congestion in a NoC [OM06, vdBCGB07]. As the packets need to go through multiple routers in order to reach their destinations, the pipeline stages of routers have a significant impact on the network latency. Thus, reducing the pipeline stages of routers in a NoC is necessary and is also widely used to reduce the network latency, such as in the look-ahead routing [Gal97] and in the speculative routing [PD01]. In some state-of-the-art router designs [KKS+07, MKAY09], the router pipeline can be aggressively reduced to only one stage.

Bypass-based approaches are also widely used to achieve a low network latency. The bypass paths in these approaches can be physical or virtual. The implementation

of physical bypass paths [MPK07, HY13] can be more efficient to reduce the network latency, but it needs more wires and complex routers, which causes more area and power overhead. In contrast, virtual bypass paths in NoCs, such as the express virtual channel [KPKJ07, KKC+08], do not need extra physical wires, so they do not cause such high area and power overhead, but the efficiency is lower. Recently, some researches show that, by adding asynchronous repeaters to reduce the wire delay, the wires on the chip can transfer packets to go through a long distance in a short time (under 45 nm manufacturing size, the packet can go up to 16mm in 1ns). Taking this advantage of such high speed wires, some bypass paths [CPK+13, KCKP13, CJ16] can be dynamically built, without the need of extra physical or virtual bypass paths, by reserving the existing wires between routers for a short time to dynamically build bypass paths.

1.1.3 Advanced Quality of Service

Quality-of-Service (QoS) is a strategy for allocating resources according to some service policies or purposes. Future NoCs will be expected to provide advanced QoS due to the growing popularity of service consolidation and real-time demands of SoCs.

At the system level, SoCs grow in complexity with an increasing variety of applications integrated on a single chip. One important branch of these applications is the hard real-time applications, in which the tasks must be finished before their deadline. Before executing a hard real-time application in a system, the functional and temporal behavior of such application should be verified. This verification process cannot be done on the application in isolation, because when multiple applications are simultaneously executed on a NoC-based many-core system, the communication interference on the NoC may influence the temporal behavior of the applications. In order to accurately verify the temporal behavior, the communication interference on the NoC must be considered. However, there are many possible combinations of simultaneously running applications and the communication interference is quit complex, which makes the verification process challenging. In order to facilitate this verification process, the NoCs should support composability, in which the communication (packet transmission) of different applications is completely isolated and cannot affect each other. Thus, the temporal behavior of applications will not be influenced by the communication interference on NoCs.

The circuit switching[3] is an effective way to achieve the aforementioned composability in a NoC. By reserving a routing path between the source router and the destination router, the packet transmission is completely isolated and there is no in-

[3]Circuit switching is a connection-oriented network switching technique. Here, a dedicated route is established between the source and the destination and the entire message/packet is transferred through it.

terference or contention between packets. Thus, there is no communication interference between different applications. In NoCs, the circuit switching can be implemented in different ways, such as time-division-multiplexed (TDM) circuit switching [GDR05, SMG14, GH10] or virtual TDM circuit switching [MNTJ04] (there are multiple virtual channels in the routers). The circuit switching avoids the packet interference but results in low bandwidth utilization and poor scalability. Compared with the circuit switching, a confined-interference communication [WGO+13, P+16] can more efficiently utilize the bandwidth. In the confined-interference communication, the packets of different applications are grouped into different domains and interference can occur only in the same domain, while there is no interference between domains. Thus, the packets in one domain have no influence on the transmission time of the packets in other domains. By supporting confined-interference communication, some composability support in a NoC can be realized in relatively simpler and easier way.

1.2 Contradictions between Design Trends

We have briefly introduced some NoC design trends in Section 1.1. However, there are a few notable contradictions between the design trend for low power consumption and the design trends for the low network latency and advanced QoS.

- **Low power consumption VS low network latency**

  To achieve low communication latency, NoCs need a certain amount of hardware to implement highly efficient communication infrastructure, such as multiple virtual channels to reduce the Head-of-Line blocking, a deep virtual channel to hide the credit-round trip delay, or the physical/virtual bypass paths to reduce the network latency. However, this hardware consumes significant power. If aggressively using a complex hardware structure to reduce the network latency, the power consumption of NoCs may sharply increase. On the other hand, if aggressively simplifying the NoC hardware structure to reduce the power consumption, the network latency may significantly increase. Moreover, in order to reduce the power consumption, low power techniques are applied on NoCs, such as DVFS and power gating. These low power techniques are effective in reducing the power consumption of NoCs, but this reduction is at the cost of increasing the network latency. Thus, to achieve low power consumption, the efficiency of the NoC may be undermined, which results in the network latency increase.

- **Low power consumption VS advanced QoS**

Typically, in order to achieve a certain QoS on NoCs, more hardware is required, which further increases the power consumption of a NoC. For example, in order to achieve a confined-interference communication [WGO+13], a NoC needs a large number of virtual channles in routers. This high hardware overhead in a NoC further increases the power consumption of the NoC. Furthermore, recent research [ZCPP15] shows that the routers in a NoC are potential hotspots on the chip. Thus, it becomes impractical to further increase the hardware overhead of routers to support QoS.

The aforementioned contradictions severely prevent the utilization of NoCs in future large-scale many-core systems. Thus, it becomes crucial to resolve these contradictions in a way which enables efficient untilization of NoCs in such future many-core systems.

1.3 Problem Statement

As discussed in Section 1.2, the high power consumption of a NoC constraints the utilization of NoCs in future large-scale many-core systems. Meanwhile, with more advanced semiconductor technologies, chips can work at lower voltages, which is helpful to achieve more energy-efficient many-core systems, but this also results in the static power consumption taking larger proportion of the total power consumption [BS00]. Thus, in this thesis, we focus 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.

1.3.1 Problem 1

Power gating is a promising low power technique to reduce the high static power consumption of the NoCs by powering off idle components that are not used. In a NoC, power gating can be flexibly applied at different granularities. In fine-grained power gating schemes, power gating can be separately applied on the components of a router, such as input ports, the crossbar, and the allocation unit. In course-grained power gating schemes, power gating can be applied on the whole router or a group of routers. When the components/routers are not used, the power of these components/routers is switched off to reduce the static power consumption. When the powered-off components/routers will be used to transfer packets, the power of these powered-off components/routers is switched on, but these components/routers cannot be immediately used because they have to be fully charged. This charging process is called the wakeup process. The extra clock cycles needed to charge the powered-off components is called

wakeup delay, which is about 6-12 clock cycles [CZPP15] in practice. The wakeup process blocks a routing path for a while. The packets have to be blocked until the powered-off components/routers are fully charged, which causes some packet latency increase. In addition, under a low traffic workload, a packet has high probability to experience multiple wakeup processes and accumulate large wakeup delay. As a consequence, the power gating causes significant packet latency increase. Furthermore, the power gating process itself consumes extra power. As a consequence, frequent power gating or power gating in a short time may cause more power consumption or inefficient power consumption reduction. Therefore, the challenge for applying power gating on NoCs is:

Problem 1: How to reduce the packet latency increase caused by power gating and achieve significant reduction of the power consumption?

1.3.2 Problem 2

As discussed in Section 1.1.3, a confined-interference communication is a promising QoS to meet real-time demands of SoCs. In a confined-interference communication, the communication interference is limited to the same domain, and there is no communication interference between different domains. By supporting a confined-interference communication, the NoCs support composability to facilitate the temporal verification of hard real-time applications. Furthermore, compared with the circuit switching, the confined-interference communication can better utilize the bandwidth of a NoC. However, realizing a confined-interference communication on a conventional (virtual channel/buffer based) NoC [WGO+13, P+16] requires a large number of virtual channels, which causes sharp power consumption increase. Therefore, there is an urgent need for realizing the confined-interference communication on a more power-efficient NoC architecture.

Simplified NoC architectures can be effective in reducing the power consumption. Bufferless NoCs [MM09, FCM11, LSMJ16] are one of the most power-efficient NoCs. By eliminating virtual channels/buffers in routers, the bufferless NoCs consume much less power than the conventional NoCs. However, as there are no buffers in bufferless NoCs to temporarily hold 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. Therefore, it is attractive and promising, to exploit the advantage of the low power consumption of bufferless NoCs and to try to realize a confined-interference communication. So, the challenge is:

Problem 2: How to realize a confined-interference communication on a bufferless NoC?

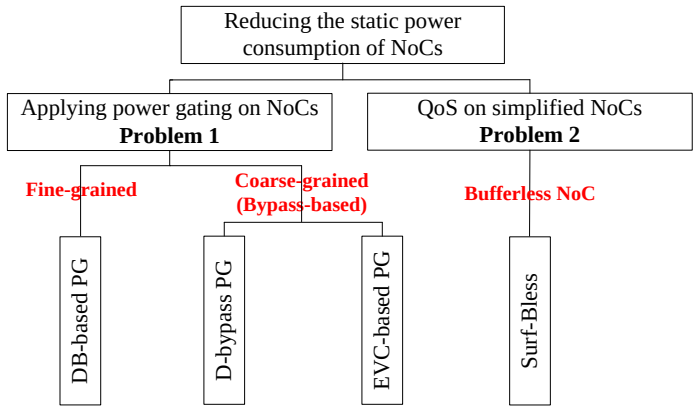

Figure 1.1: Contributions outline.

1.4 Contributions of The Thesis

By addressing the research problems formulated in Section 1.3, in this section, we summarize the research contributions as shown in Figure 1.1. To address Problem 1, described in Section 1.3.1, we propose 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 three power gating approaches can significantly reduce the power consumption of NoCs and are effective in reducing the packet latency increase caused by power gating. In addition, with different properties, they have their own advantages. As a fine-grained power gating approach, 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, our DB-based power gating approach is effective in reducing the power consumption of a NoC in a wider range of traffic workloads. The D-bypass power gating approach and the EVC-based power gating approach allow packets to bypass the powered-off routers without the need of experiencing the wakeup processes introduced in Section 1.3.1. Thus, they are more effective in reducing the packet latency increase caused by power gating and have less performance penalty. In addition, the EVC-based power gating approach allows packets to bypass powered-on routers as well to reduce the dynamic power consumption. Thus, at high workloads, the EVC-based power gating approach consumes less power than the D-bypass power gating approach. However, the EVC-based power gating approach may cause unfair allocation of the network resources, which may result in more performance penalty.

To address Problem 2, described in Section 1.3.2, we propose a novel routing

approach called Surfing on a Bufferless NoC (Surf-Bless), which achieves an energy-efficient confined-interference communication by taking advantage of the low power consumption of a bufferless NoC.

The specific novel contributions of the thesis are the following:

Contribution 1: Duty buffer based (DB-based) power gating approach

In this contribution, presented in Chapter 3, first, we propose a novel hardware structure, called duty buffer (DB), which can be used to replace any virtual channel in an input port. Then, taking advantage of our novel duty buffer, we propose a novel fine-grained power gating approach, called DB-based power gating approach. In our DB-based power gating approach, each input port of a router can be independently powered off to reduce the static power consumption. As the virtual channels in input ports are the main consumer of the static power in a NoC, our DB-based power gating approach is very effective in reducing the power consumption. Moreover, by using the duty buffer to replace the powered-off virtual channels, even when the input ports are powered-off, a router still keeps certain transmission ability to transfer packets. Thus, the packet can avoid as much as possible to be blocked by the powered-off input ports, and the packet latency increase caused by the power gating is reduced. Finally, by experiments in comparison with the related works [MKWA08, ZOG+15], our DB-based power gating approach achieves comparable power reduction. Moreover, our approach is more effective in reducing the packet latency increase caused by the power gating.

However, being a fine-grained power gating approach, our DB-based power gating needs to separately switch the power of each input port in a router. Thus, some times packets may experience more power gating processes, which may result in a lot extra packet latency increase.

Contribution 2: Dynamic bypass (D-bypass) power gating approach

In order to overcome the aforementioned drawback of the DB-based power gating, we propose the Dynamic bypass (D-bypass) power gating approach, presented in Chapter 4, which is a course-grained power gating approach. First, we apply power gating on each router and add one special hardware bypass structure in each router, which allows a packet to bypass a powered-off router from any input port to any output port at a time. Then, we propose a reservation mechanism to dynamically allocate the use of the special hardware bypass structure in a router. Thus, by dynamically reserving this hardware bypass structure in the powered-off router, packets can bypass the powered-off router more flexibly than in the related approaches [CP12, FTKH16, BHW+17, ZL18]. In our D-bypass power gating approach, as the packet can go through the powered-off routers, some packets do not need to experience the wakeup process and wait for the powered-off routers to be fully charged. Thus, the packet latency increase caused by the power gating is reduced. Furthermore, without

frequent interruption of powering on the powered-off routers, routers have more idle time to stay powered-off to reduce the static power consumption and have less power consumption overhead caused by the power gating. Finally, by experiments, we show that our D-bypass approach can efficiently reduce the power consumption. Moreover, compared with other related approaches [MKWA08, CP12, WNWS17, ZL18], our D-bypass approach has less performance penalty than the related approaches and the DB-based power gating approach.

However, just like most of the course-grained power gating approaches, our D-bypass power gating approach cannot fully utilize the idle time of each component in a router. When the traffic workload is high, 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 in reducing the power consumption only at low traffic workloads.

Contribution 3: Express virtual channel based (EVC-based) power gating approach

To overcome the aforementioned drawback of the D-bypass power gating approach, we propose the express virtual channel based (EVC-based) power gating approach, presented in Chapter 5. First, we extend the router micro-architecture and apply power gating on each router. Then, 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. Thus, our EVC-based power gating approach not only reduces the packet latency increase and extra power consumption caused by power gating, but also reduces the dynamic power consumption by transferring packets to bypass the powered-on routers. Thus, even at high traffic workloads when all of the routers are powered-on and power gating is ineffective in reducing the power consumption, our EVC-based power gating approach consumes less power than other related approaches [MKWA08, ZL18]. Moreover, the pre-defined virtual bypass paths are more efficient in transferring packets. Thus, our EVC-based power gating approach can achieve lower packet latency than the related approaches [MKWA08, WNWS17, ZL18]. However, compared with the D-bypass power gating approach, the EVC-based power gating approach may cause unfair allocation of the NoC resources thereby causing, in some cases, higher performance penalty in spite of the fact that the EVC-based power gating approach has lower packet latency.

Contribution 4: Surfing on a Bufferless (Surf-Bless) NoC routing approach

To address Problem 2 in Section 1.3.2, we propose a novel routing approach, called Surfing on a Bufferless NoC (Surf-Bless), presented in Chapter 6, which 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 critical to correctly use a bufferless NoC to transfer packets. It 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. By experiments, we show that our approach is effective to support such confined-interference communication and consumes much lower energy/power than the related approach [WGO+13].

1.5 Thesis Outline

Below, we give an outline of this thesis, which summarizes the content of the following chapters.

Chapter 2 provides basic information about NoCs to make it easier to understand the contributions of this thesis.

Chapter 3 to Chapter 6 contain the main contributions of this thesis. Each chapter is organized in a self-contained way, meaning that each chapter contains more specific introduction to the problem addressed, background information, related works, the proposed solution approach, an experimental evaluation, and a concluding discussion. Chapter 3, Chapter 4, and Chapter 5 are about applying power gating on a NoC to reduce the static power consumption and address the problem of the packet latency increase caused by the power gating. Chapter 6 is about realizing confined-interference communication on a bufferless NoC.

Chapter 3 presents our novel duty buffer structure and our DB-based power gating approach. This chapter is based on our publication [WNWS17].

Chapter 4 introduces our D-bypass power gating approach. This chapter is based on our publication [WNM+19a].

Chapter 5 elaborates our EVC-based power gating approach. This chapter is based on our publication [WNM+19b].

Chapter 6 presents our novel Surf-Bless routing to realize a confined-interference communication on a bufferless NoC. This chapter is based on our publication [WNM+19c].

Chapter 7 concludes this thesis.