

Chapter 2

Literature Survey

One of the most important reasons for using NoC architectures is their promise for scalability [15, 41, 60, 64, 148]. Several books provide an introduction to the NoC concept and discuss various research issues [50, 76, 87, 107, 125], while an exhaustive list of references can be found in some NoC bibliographies available on-line [121, 127]. Likewise, a comprehensive introduction to NoCs and existing design practices is presented in [21]. In what follows, we provide a systematic literature review which is structured along the lines discussed in [105].

2.1 Application Modeling and Optimization for NoC Communication

2.1.1 Traffic Models

Traffic models refer to the mathematical characterization of workloads generated by various classes of applications. With network performance being highly dependent on the actual traffic, it is obvious that accurate traffic models are needed for a thorough understanding of the huge design space of network topologies, protocols, and implementations. Since implementing real applications is time-consuming and lacks flexibility, such analytical models can be used instead to evaluate the network performance early in the design process.

Traffic characteristics have been long recognized as playing a major part in multicore systems design. For instance, in [171] the authors introduce an analytical traffic model based on identifying self-similar effects in multimedia traffic. These effects have important consequences for the design of on-chip multimedia systems since self-similar processes have properties which are completely different from traditional short-range dependent or Markovian processes that have been traditionally used in system-level analyses. Later, the authors in [160] derive a comprehensive traffic model for NoCs which exposes both spatial and temporal dimensions of traffic via three statistical parameters: hop count, burstiness, and

packet injection distribution. Interestingly enough, it has been subsequently reported that even the traffic generated by programmable cores consists of multiple program phases which exhibit the same type of self-similar behavior [145].

It should be noted, however, that the research in this area is still behind due to the lack of a widely accepted set of NoC benchmarks. This situation has two primary reasons. First, the applications suitable for NoC platforms are typically very complex. For instance, it is common for applications to be partitioned among tens of processes (or more) in order to allow for evaluations of scheduling, mapping, etc. For general purpose chip multiprocessors (CMPs), benchmarks such as SPLASH, originally designed for shared-memory multiprocessors, may be used but it is, however, unclear how effectively such benchmarks can actually stress the NoCs. Second, compared to traditional research areas like physical design where the design constraints are static (e.g., the aspect ratio of the blocks, number of wires between different blocks, etc.), the NoC research requires detailed information about the dynamic behavior of the system; this is hard to obtain even using detailed simulation or prototyping. As a result, most researchers and designers still rely on synthetic traffic patterns such as uniform random, bit-permutation traffic, to stress-test a network design [40, 90]. A first step towards a unified approach for embedded platforms has been made recently via the OCP-IP benchmarking initiative [58]. Similarly, there have been initial steps towards releasing parallel benchmarks targeting the future CMPs [19]. Such initiatives can certainly boost the research progress in this area. However, there needs to be more research aimed at developing accurate traffic models, as well as in-depth studies that project the NoC traffic for emerging workloads.

2.1.2 Application Mapping

Applications are typically described as a set of concurrent tasks that have been already assigned and scheduled onto a set of selected IP cores. The mapping problem for NoCs is to decide how to topologically place the selected set of cores onto the PEs of the network, such that some metrics of interest are optimized. We note that PE simply means a placeholder connected to one of the network routers. In other words, “mapping” here means determining which IP core connects to which router in the network; this, obviously, greatly impacts both performance and energy consumption of the NoC.

The mapping problem for NoCs has been first addressed by the authors of [68], where a branch and bound algorithm is proposed to map a given set of IP cores onto a regular NoC architecture such that the total communication energy is minimized. At the same time, the performance of the resulting communication system is guaranteed to satisfy the specified design constraints through bandwidth reservation. Follow-up work considered the mapping problem with increased path diversity [113] as well as additional latency constraints [162]. Likewise, a multi-objective mapping algorithm that finds the Pareto mappings with optimum

performance and power consumption is presented in [7]. Improving upon these studies, the authors in [62] propose a more general, unified approach for application mapping and routing path selection which considers both best effort and guaranteed service traffic.

One key component needed to solve the application mapping problem is the analytical model used for solution evaluation. For instance, if the goal is communication energy minimization, an accurate energy model is crucial. We note that many mapping algorithms use (directly or indirectly) the average packet hop count as a cost function, by relating the average number of packet hops to the communication energy consumption [71] or communication cost [162]. Along the same lines, effective performance models (such as those discussed in Chap. 5) are needed. When PEs have different sizes, the communication latency and power consumption per unit of data exchanged between any two neighboring routers may differ significantly. Therefore, embedding floorplanning information within the mapping loop becomes necessary to get more accurate energy/latency estimates [112].

With increasing level of programmability, MPSoCs are used under multiple use case scenarios. Hence, it becomes necessary to allocate the NoC resources based on different communication requirements (i.e., bandwidth and latency) and traffic patterns that characterize various use cases [116]. By the same token, with ever increasing power density and cooling costs, it is important to reduce or eliminate the potential hotspots and have a thermally-balanced design [73]. Finally, efficient techniques for *run-time* mapping and management of applications are needed. Towards this end, software development and code placement for embedded multiprocessors are discussed in [54]. Similarly, for applications launched dynamically, run-time mechanisms for mapping [3, 37] and/or migrating [17] are needed. Since execution time and arrival order of applications are not known a priori, finding optimal solutions is difficult and remains a big challenge.

2.1.3 Application Scheduling

Another important problem in NoC design is communication and task scheduling. Although scheduling is a traditional topic in computer science, most previous work focuses on maximizing performance [139, 176]. More recently, energy-aware scheduling techniques for hard real-time [59, 146, 156] and distributed [99, 109] systems have also been introduced, but they address only the bus-based or P2P communication. Without taking into consideration the network congestion which may change dynamically during tasks execution, such techniques cannot be directly applied to NoC scheduling. We note that mapping and scheduling problems can be considered jointly. However, finding the optimum solution remains an open problem due to its complexity.

Communication and task scheduling for NoCs is addressed in [69] where the authors present a scheduling algorithm which minimizes the overall energy consumption of the system while guaranteeing the real-time deadlines imposed on

tasks. Likewise, scheduling and arbitration policy for NoCs that use code division multiple access protocol is presented in [81].

Capturing the dynamic system behavior, i.e., the change system behavior due to the incoming and completed applications, is also important. To this end, the work in [140] models the applications as a set of independent jobs and presents exact timing models that capture both computation and communication of a job. Similarly, the work in [165] extends the NoC scheduling to consider multiple use case scenarios, hence communication patterns. By allowing the bandwidth to be shared among multiple communication scenarios, a better resource utilization of the NoC is obtained.

It should be also noted that dynamic voltage frequency scaling (DVFS) can be used in conjunction with scheduling in order to minimize the overall energy consumption. Such techniques have been proposed in the past for both bus- [59, 146] and NoC-based communication [155]. In these approaches, voltage scaling is applied to tasks and/or communication to minimize the power consumption, while accounting for the DVS overhead and satisfying the application deadlines.

We note that although we discuss here the mapping and scheduling problems separately, they can also represent a joint optimization problem. However, since they are both very hard problems to solve, such an integrated approach remains an open problem. This is particularly challenging for NoCs since communication delay is difficult to estimate and so deriving accurate models that can be used to guarantee hard deadlines is a huge problem by itself.

2.2 Communication Paradigm

2.2.1 Packet Routing

Given an underlying topology, the routing protocol determines the actual route taken by a message. The routing protocol is important as it impacts all network metrics, namely, latency (as the hop count is directly affected by the actual route), throughput (as congestion depends on the ability of the routing protocol to load balance), power dissipation (as each hop incurs a router energy overhead), QoS (as routing can be used to channel different message flows along distinct paths to avoid interference) and finally reliability (as the routing protocol needs to choose routes that avoid faults).

Routing has been extensively studied in classical interconnection networks, many of which have been leveraged in on-chip networks. One example is the dimension-ordered routing which routes packets in one dimension, then moves on to the next dimension, until the final destination is reached. While such a technique is very popular due to its simplicity, adaptive routing techniques (e.g., turn model routing, planar adaptive routing [40, 47]) can provide better throughput and fault tolerance by allowing alternative paths depending on the network congestion and

run-time faults. Oblivious routing algorithms which generate routes without any knowledge of traffic have also been extensively studied in the context of classical interconnection networks [147] and can be relevant to on-chip networks due to their low overhead [169].

New routing protocols have also been investigated specifically for NoCs. For instance, *deflective routing* in [119] routes packets to one of the free output channels belonging to a minimal path; if this is not possible, then packets are misrouted. Techniques have been also proposed to dynamically switch between deterministic and adaptive routing to exploit the trade-off between them [70].

Application-specific customization of routing protocols [71, 113] and techniques to provide low overhead routing algorithms with high path diversity [1, 114] have been explored. With NoCs being increasingly concerned with power, thermal and reliability issues, there exists recent work proposing thermal- [151], and reliability-aware [103] routing algorithms.

While some existing research into routing algorithms for off-chip interconnection networks can be leveraged for NoCs, the significantly different constraints of on-chip implementations lead to new challenges. First, the ultra-low latencies [56] and very high frequencies [170] of some NoCs make it difficult to incorporate sophisticated routing algorithms such as adaptive routing. The tight power constraints and reliability issues also lead to challenges in power-aware and fault-tolerant routing. With static topology irregularity, it is difficult to find minimal routes that can avoid deadlock and livelock situations (e.g., [30]). This is a research direction that needs more attention in the future. Relying on dimension-ordered routing as the escape routing function in irregular topologies becomes difficult and implementations typically require tables that incur delay, area, and power overheads [26]. To date, the vast majority of NoC routing solutions have focused on unicast traffic, that is, sending from one PE to another. Support for on-chip multicast [49] needs to be considered too, with emphasis on lightweight solutions such that the tight on-chip constraints can be met.

2.2.2 Switching Techniques

Switching, also called flow control,¹ governs the way in which messages are forwarded through the network. Typically, the messages are broken down into *flow control units (flits)* which represent the smallest unit of flow control. The switching algorithm then determines *if* and *when* flits should be buffered, forwarded, or simply dropped [40, 47]. As a result, the switching algorithm has the most direct impact on router microarchitecture and pipeline.

¹ The two terms “switching” and “flow control” have been used interchangeably in leading NoC texts [40, 47, 107].

Among the commonly used switching techniques in interconnection networks, wormhole switching seems the most promising for NoCs due to the limited availability of buffering resources and tight latency requirements. Virtual channels [42] are widely used in off-chip interconnection networks and are naturally adopted for NoC design to improve network bandwidth and tackle deadlock. However, as the design requirements change dramatically, the underlying substrate presents new opportunities for designing flow control algorithms.

Early work on NoC flow control aggressively drives down the router delay to a single cycle, through static compiler scheduling of network switching operations [168], dedicated look-ahead signals for setting up the switch ahead of time [56], by speculatively allocating resources to move the latency associated with resource allocation and multiplexing off the critical path at low traffic loads [111, 135], or through advanced reservation of resources [134]. While most studies focus on packet switching, several papers investigate the potential of circuit switching and time division multiplexing, to reduce the arbitration and buffering overheads of packet-switched routers [48, 62, 174].

Several techniques tackle NoC throughput such as dynamically varying the number of virtual channels (VCs) assigned to each port, to better adapt to the traffic load [117]. Express virtual channels aggressively drive down the router latency to just link latency, while extending throughput by having VCs that are statically defined to cross multiple hops [88]. Tackling latency and throughput simultaneously, layered switching in [98] hybridizes wormhole and cut-through switching by allocating groups of data words which are larger than flits but smaller than packets.

There is still a significant latency/throughput gap between the state-of-the-art NoCs and the ideal interconnect fabric of dedicated wires [88]. This disparity largely lies in the complex routers necessary at each hop for delivering ultra-low-latency and/or high bandwidth. In order for NoCs to be an efficient and effective replacement of dedicated wires as the primary communication fabric, there is a need for new switching techniques that can obviate this router overhead and truly deliver the energy-delay-throughput of dedicated wires.

2.2.3 QoS and Congestion Control

Conventional packet-switched NoCs multiplex message flows on links and share resources among these flows. While this results in high throughput, it also leads to unpredictable delays per individual message flows. For many applications with real-time deadlines, this non-determinism can substantially degrade the overall application performance. Thus, there is a need for research into NoCs that can provide deterministic bounds for communication delay and throughput.

QoS in NoCs is typically handled through three types of approaches. First, resources such as VCs can be pre-reserved with a fair mechanism for allocating resources between different traffic flows [20, 55, 94, 95, 108]. Second, multiple

priority levels can be supported within the network such that the urgent traffic can have a higher priority over the regular traffic [13, 25, 63, 104]. Techniques to ensure global fairness to network hot-spots have been proposed in [93]. Finally, QoS-aware congestion control algorithms have been proposed to avoid the spikes in delay when the traffic load approaches saturation by having congestion control at the network interface regulate traffic and ensure fairness [27, 46, 119, 124].

Future CMPs and MPSoCs impose increasingly more QoS demands on NoCs; yet, support for QoS has to be extremely light weight. Cache-coherent CMPs would benefit from NoC being able to preserve the ordering semantics of a bus, thereby easing consistency support. Being able to provide guarantees on packet deliveries such as snoop responses will also ease protocol design and lower the protocol overhead. Both SoCs and CMPs will benefit from NoCs that can support dynamically defined QoS levels and needs, as well as dynamically defined partitions of the NoC that support different QoS.

Further, it will greatly strengthen the fundamental basis of NoCs to have research into analytical models that can estimate network latency and/or throughput for arbitrary traffic patterns, as these can be used to feed into QoS engines with low implementation overhead.

2.2.4 Power and Thermal Management

Due to concerns on battery lifetime, cooling and thermal budgets, power issues are at the forefront of NoC design. Seminal work on router power modeling [133] along with position papers that highlight the importance of NoC power consumption [15, 41] motivated research into low-power NoCs. There has been research into run-time NoC power management using dynamic voltage scaling on links [150], as well as shutting links down based on their actual utilization [80, 161]. Globally Asynchronous Locally Synchronous (GALS) approaches to dynamic voltage and frequency scaling further leverage the existing boundaries between various clocking domains [14, 126].

Besides average power, peak power control mechanisms for NoCs have also been explored due to their impact on thermal hotspots [18]. Power and thermal management are also very tightly related to reliability. An approach for joint power and reliability management is presented in [159], while error coding schemes for improved reliability and power consumption are presented in [16, 175].

Thermal dissipation is another metric of interest and mechanisms have been investigated to control peak power with respect to its impact on thermal dissipation [73, 151]. To this end, thermally-aware task scheduling for MPSoCs is presented in [39] and thermal optimization of 3D implementations via task scheduling and voltage scaling has been studied in [166].

With NoCs facing highly-constrained power envelopes, run-time power management techniques are needed to reduce peak power consumption so as to avoid thermal emergencies. Challenges remain in dynamically estimating the thermal

hotspots as well as dynamic power profile in the presence of high workload variations. Holistic approaches covering hardware (for estimation and low level control), firmware (for implementing system level power manager) and operating system (for application characterization) are needed to tackle this problem.

Another wide open area for research is related to *distributed* control strategies for power and thermal management in NoCs. Techniques like [126, 158] are based on a centralized power manager but perhaps relying only on localized information may have its own advantages for NoCs; whether or not this is indeed the case remains to be investigated. It is important to note, again, that the accuracy of the energy models is crucial for these optimization techniques. Ideally, such models should target both dynamic and static power dissipation. While there exist preliminary efforts in this direction (e.g., [106] where adaptive body biasing is used to minimize static energy consumption), more work is needed to achieve practical solutions.

2.2.5 Reliability and Fault Tolerance

As CMOS technology approaches the nanoscale domain, there is an increasing need for studying how NoC architectures can tolerate faults and underlying process variations. For instance, shrinking transistor sizes, smaller interconnect features and higher operating frequencies of current CMOS circuits lead to higher soft-error rates and an increasing number of timing violations [154]. Moreover, the combination of smaller devices and voltage scaling in future technologies will likely result in increased susceptibility to transient faults. Therefore, in order to reduce the cost of design and verification, the future SoC architectures need to rely on fault-tolerant approaches.

Fault-tolerant multi-chip interconnection networks have been investigated, mostly in the areas of fault-tolerant routing or microarchitecture [47]. For NoCs, one of the earliest fault-tolerant communication approaches is the stochastic communication described in [23]. This approach is based on probabilistic broadcast where packets are forwarded randomly to the neighboring nodes. A theoretical model explaining the stochastic communication and relating the node coverage to the underlying properties of a grid topology was also proposed. Similarly, the studies in [138, 141] explore how NoC routing algorithms can route around faults and sustain network functionality in the presence of faults.

Researchers have also modeled the interaction between various NoC metrics, like delay, throughput, power and reliability [52]. Several studies look specifically into router design and ways to improve the NoC reliability through microarchitectural innovations that go beyond the expensive alternative of having redundant hardware [6]. The fault tolerance overhead of various flow control techniques is analyzed in [142]. Similarly, power consumption of link level and end-to-end data protection in NoCs is analyzed in [75], while energy efficiency of error correction at the receiver end is studied in [16, 115].

With devices moving into deep submicron technologies, reliability becomes a very important issue. However, research into NoC reliability is still in its infancy and thus realistic fault models that are a good representation of physical realities for NoCs are needed. Research exploring the trends in soft error rates for combinational and sequential logic (e.g., [110, 157]) can potentially be relevant to router microarchitectures as well. Future work that will critically impact the NoC power-performance-reliability trade-off includes the cost effectiveness of providing fault-tolerance, while maintaining suitable levels of fault isolation and containment.

2.3 Communication Infrastructure

2.3.1 Topology Design

The ability of the network to efficiently disseminate information depends largely on the underlying topology. Indeed, besides having a paramount effect on the network bandwidth, latency, throughput, overall area, fault-tolerance and power consumption, topology plays an important role in designing the routing strategy and mapping the IP cores to the network.

The simplicity and regularity of mesh structures makes design approaches based on such a modular topologies very attractive. More precisely, regularity improves timing closure, reduces dependence on interconnect scalability, and enables the use of high performance circuits. Typically, one-dimensional topologies (e.g., ring [136]) and two-dimensional topologies (e.g., mesh and torus [56, 167]) are the default choices for NoC designers. Node clustering to obtain topologies like the concentrated mesh [9] and hierarchical star [92] is a viable alternative to amortize the router overhead and reduce latency. Higher-radix networks like the flattened butterfly [85] reduce power and latency by reducing the number of intermediate routers and the wiring complexity over conventional butterfly but they increase the number of long wires.

Despite the benefits of regular network topologies, customization is also desirable for several reasons. First, when the size or shape of the cores varies widely, using regular topologies may waste area. Moreover, for real applications, the communication requirements of the components can vary widely. Designing the network to meet the requirements of highly communicating cores results in under utilization of other components, while designing it for the average case results in performance bottlenecks. Finally, for application-specific NoCs, a detailed a priori understanding of the communication workload can be exploited to fully customize the network topology [65, 122]. For instance, the approach proposed in [137] enables the automatic design of the communication architecture of a complex system using a library of pre-defined IP components. Similarly, the work in [164] presents a mixed integer linear programming-based technique for

NoC topology synthesis with the objective of minimizing the power consumption subject to performance constraints.

Interestingly enough, the two extreme points in the design space (i.e., completely regular and fully customized topologies) are not the only possible solutions for on-chip communication. Indeed, by inducing small world effects, the performance of regular topologies can be significantly improved with minimal impact on area and energy consumption [123]; this idea will be discussed in detail later in Chap. 6. However, for now it suffices to say that inducing small world effects via long-range links has a wide applicability as it has been demonstrated by its extension to on-chip radio-frequency links [33], express virtual channels [88] and wireless links [132].

Generally speaking, the problem of optimal topology synthesis for a given application does not have a known theoretical solution. Although the synthesis of customized architectures is desirable, distorting the regular grid structure leads to various implementation issues such as complex floorplanning, uneven wire lengths, etc. Although we have discussed only planar substrates, 3D die stacking provides the opportunity for higher radix topologies through the use of inter-die connections [84, 166, 177] and warrants further exploration.

2.3.2 Router Design

The design of a router involves determining the flow control techniques, number of virtual channels, buffer organization, switch design, pipelining strategy while adhering to target clock frequency and power budgets. All these issues require careful design since they have significant impact in terms of performance, power consumption and area.

The main focus in designing a router is to minimize the latency through it, while meeting bandwidth requirements. Reservation [134] and speculation [111, 135] can be used to hide the routing and arbitration latencies and achieve a single-stage router design. Decoupled parallel arbiters and smaller crossbars for row and column connections can reduce contention probability and reduce latency [83]. Moreover, techniques such as segmented crossbars, cut-through crossbars and write-through buffers can be used to design low power routers [172].

The impact of the number of VCs on performance varies with the network load. A lightly loaded network does not need many VCs, whereas a heavily loaded network does. A virtual channel regulator which dynamically allocates VCs and buffers according to traffic conditions, thereby reducing total buffering requirements and saving area and power, is presented in [117]. Arbitration during VC allocation is another area of potential optimization. Free virtual channel queues at each output port can effectively remove the need for VC arbitration by predetermining the order of grants [111].

An efficient algorithm for the buffer size allocation problem is proposed in [72]. The authors derive the blocking rate of each individual channel and then add more

buffering resources only to the highly utilized channels. Similarly, the properties of on-chip buffers and gate-level area estimates are studied in [143]. Finally, advanced circuit-level techniques have been employed to achieve high-speed and low power operation. For instance, the router presented in [170] employs a double pumped pipeline stage to interleave alternate data bits using dual edge-triggered flip-flops; this optimization reduces the crossbar area by 50 %. Similarly, serial on-chip links, partial crossbar activation, and low energy transmission coding techniques are used in the router design presented in [92].

Tools that enable micro-architecture exploration to trade off latency and bandwidth of the router against power consumption can help NoC designers make the right design decisions for particular application requirements. Accurate performance analysis of on-chip routers under arbitrary input traffic and methodologies for choosing the correct design parameters such as optimal channel width, buffer depth, pipeline depth, number of VCs for high performance and low power remain open problems. Finally, energy-efficient routers that can interface a variety of IP cores designed for legacy communication protocols with minimal performance overhead is an important challenge.

2.3.3 Network Channel Design

The links interconnecting the network routers also need to be designed efficiently in order to consume low power and area. The ideal interconnect should be such that its performance and cost come close to that of just the network channels (or links), with the performance delivered and power consumed by the network channel largely determined by the signaling techniques.

Non-uniform channel capacity allocation is presented in [61] where the traffic is assumed to be heterogeneous and critical delay requirements vary significantly. In addition to the effects mentioned above, the choice of W has implications on the wire sizing and spacing, which affect the channel operating frequency. The bandwidth of a network channel is given by $BW = f_{ch} \times W$. Hence, bandwidth cannot be optimized by simply considering f_{ch} and W separately. Pileggi et al. [96] discuss maximizing channel throughput by controlling the size and spacing of wires, as well as their number. In [78], the authors discuss a framework for equalized interconnect design for on-chip networks. The proposed approach finds the best link design for target throughput, power and area constraints, and enables architectural optimization for energy-efficiency.

It is also of interest to explore different implementation styles for network links. For instance, delay-insensitive current mode signaling [118] as well as low-swing, differential signaling techniques [77] can be used to improve performance and reduce power.

Alternatives to wire channels present interesting opportunities for future research. For instance, analog-digital hybrid routing approaches [102], optical links [120, 149], RF interconnects [33] and wireless links [179] have been

considered as alternatives to traditional on-chip repeated interconnects, but more work is needed to make such approaches applicable to real designs.

2.3.4 Floorplanning and Layout Design

Standard tile sizes help controlling the link lengths and ensuring that link delays do not limit the operating frequency. However, if the size of the network tiles varies significantly, or irregular topologies are used, the floorplanning step becomes mandatory. In this case, emphasis needs to be put on the shape and placement of tiles so as to control the link lengths. Reducing the total interconnect length is also important for reducing the power dissipation across the links. Another problem is the placement of special tiles like those connected to peripheral devices (e.g., memory controllers, I/Os) so as to minimize the average latency to these devices. In addition to link length, the goal here is to minimize link area by routing links over logic or caches as much as possible.

Layout-aware area, power and performance analysis of mesh-based NoCs is discussed in [130]. Likewise, the authors in [5] present a comparison in terms of performance, area and power scalability between crossbar designs within a pre-existing communication fabric and the NoC approach at layout level. Considering physical layout is also important while solving mapping and topology synthesis problems. To this end, floorplan aware solutions to these problems are presented in [112, 163]. We note that the size of the IP cores are assumed to be fixed in the SoC context. On the other hand, this does not necessarily hold for the CMP context where one can make trade-off between cache size and interconnect area [89]. For example, larger private caches can filter traffic and reduce the requirement on the network, while performance drawback of smaller caches can be mitigated by higher performance NoCs.

2.3.5 Clocking and Power Distribution

The traditional approach of designing fully synchronous chips with a single global clock is not attractive anymore due to smaller process geometries, larger wire delays, higher levels of integration of multiple cores on large dies. The large effort required for skew control and the significant power consumption of the global clock call for alternative clocking strategies [12]. Indeed, in addition to multiple frequencies, different cores can have their own optimal supply voltage to allow for fine-grain power/performance management.

Strategies such as asynchronous or mesochronous clocking [8, 22, 152] are alternatives that hold the promise of simplifying timing closure and global clock distribution. For instance, an approach to minimize the strict skew requirements without going fully asynchronous using all-in-phase clocking is presented in [153].

In [22], the authors present a mesochronous clocking strategy that avoids timing related errors while maintaining globally synchronous system perspective. The 80-tile teraflop NoC presented in [170] employs phase tolerant mesochronous interfaces between the routers with FIFO-based synchronization. Similarly, latency insensitive or synchronous elastic systems are developed to exploit the inherent advantages of synchronous design while decoupling the functionality from the channel delays [29, 38].

The GALS approach has been used with several tile-based multiprocessor implementations [178]. However, there are extra costs in terms of synchronization latency and power that need to be considered. Chelcea et al. [35] discuss interfacing different clock domains which is essential for implementing globally asynchronous systems. A systematic comparison between asynchronous and GALS implementations of an NoC is presented in [152]. The authors conclude that while the two approaches result in similar silicon area, power consumption and bandwidth, the asynchronous implementation has a clear advantage in terms of average packet latency. In [28], the authors propose asynchronous delay insensitive links to support the GALS NoC paradigm. This approach removes the constraints on wire propagation delays and enables designing links of any width with low wire and logic overhead.

Open problems in this area include robust design of clock crossing synchronizers with minimal latency penalty and low power consumption, since locally generated clocks for GALS SoCs are prone to synchronization failures due to clock delays [44]. Recent research in resonant clocking shows promise for reducing power and delivering high performance [32]; however, the use of resonant clocking with NoCs has yet to be investigated. Also, while controlling NoCs with multiple voltage-frequency islands has been discussed in [126], techniques that consider other control objectives such as chip temperature, power consumption, and nonlinear effects are needed.

2.4 NoC Evaluation and Validation

2.4.1 Analysis and Simulation

Fast and accurate approaches for analyzing critical metrics such as performance, power consumption or system fault-tolerance are important to guide the design process. However, in order to be used within an optimization loop or make early design choices, the analysis techniques need to be tractable and provide meaningful feedback to designers. At later design phases, one can obtain more accurate estimates through simulation.

Communication latency and network bandwidth are common performance metrics of interest. While it is relatively easier to find the communication latency for guaranteed service traffic [43, 108], analyzing the average latency for best

effort traffic is a challenging task. Therefore, the average hop count or free packet delay are commonly used to approximate the average packet latency [113, 164]. Techniques for analyzing the average communication latency in networks are proposed in [45, 67]. While not directly applicable to NoC performance analysis, these approaches can be used as a starting point and then account for NoC-specific constraints, such as application specific traffic and on-chip router parameters.

Analytical power models for early-stage power estimation in NoCs have also been investigated, starting with [133] which models the power consumed of multi-chip interconnection networks; this generated follow-up work that specifically targets on-chip network power dissipation [10, 31, 36, 51, 82, 151].

Performance analysis largely depends on various simplifying assumptions on the network or traffic characteristics (e.g., uniform traffic vs. bursty traffic) and typically assumes deterministic routing due to the difficulty in handling the more general problem. Approaches that relax the Markovian assumption and analytical power consumption models that accurately account for the application and architecture characteristics are highly needed [24].

Simulation-based approaches are still popular for architectural exploration of on-chip networks due to their accuracy, flexibility and ability to run real workloads [86, 87, 100, 129, 131, 173]

The major issue with simulation-based approaches is the trade-off between the level of implementation detail and simulation time [74]. Detailed models can deliver very accurate results, but the simulation time can be prohibitive. Realistic synthetic trace simulation [101, 171] or hardware acceleration [53] can be used for improving the simulation speed therefore these are wide open directions for research.

2.5 Prototyping, Testing and Verification

While simulation offers flexibility for power-performance evaluations under various network parameters, it still relies on many approximations that may affect the accuracy of the results. Prototyping can be further used to improve the evaluation accuracy by bringing the design closer to reality, at the expense of increased implementation effort and reduced flexibility. Finally, it is also important that testing and verification must be considered to ensure correctness.

Several concrete NoC architectures have been presented in the literature. In [2], the authors present the SPIN interconnect architecture and implement a 32-port network architecture using a 0.13 μ m process. This architecture uses credit-based flow control to provide QoS.

A flexible FPGA-based NoC design that consists of processors and reconfigurable components is presented in [11]. The FPGA prototype presented in [123] illustrates the impact of application-specific long-range links on the performance and energy consumption of 2D mesh networks. The aSoC architecture presented in [95] supports compile-time scheduling for on-chip communication and provides

software-based dynamic routing. The RAW chip [167], attacks the wire-delay problem by proposing a direct software interface to the physical resources. The static network used in the RAW chip also enables new application domains.

The 80-core teraflops chip recently introduced by Intel [170] is a good example of a major NoC prototyping effort. The chip uses a 2D mesh with mesochronous clocking for a high bandwidth, scalable design. The authors in [92] present a highly optimized NoC implementation using hierarchical star topology. Finally, the work presented in [79] addresses both architectural aspects and circuit level techniques for practical NoC implementation.

We note though that most studies dealing with concrete NoC implementations lack performance evaluation under *real* driver applications. This is an important issue that needs to be addressed in order to bring the NoC prototypes closer to real applications. Towards this end, the authors in [91] compare and contrast the NoC, bus- and P2P-based implementations of an MPEG-2 encoder using an FPGA-based prototype. The advantages of the NoC approach are illustrated in terms of scalability, throughput, energy consumption and area, both analytically and using direct measurements on the prototype.

In NoCs, the routers and links have been utilized to test the PEs and the network itself based on built-in self-test mechanisms [4, 57, 66, 97]. In [4], the authors propose a scalable test strategy for the routers in an NoC, based on partial scan and an IEEE 1500-compliant test wrapper. Similarly, the strategy proposed in [66] exploits the regularity of the switches, and broadcasts the test vectors through the minimum spanning tree to test the switches concurrently. The authors in [57] propose testing the routing logic and FIFO buffers recursively by utilizing the NoC component that already passed the test. The work presented in [97] also considers the power consumption required for testing purposes.

NoC verification has received less attention compared to other design aspects or even testing. The NoC verification approach in [55] relies on monitoring the network traffic and checking special events such as connection opened/closed, data received by a connection, etc. Likewise, the MAIA framework aims at automated generation and verification of NoC architectures [128]. More recently, formal verification of asynchronous NoC architectures is considered in [144], while a framework for quick formal modeling and verification of communication fabrics is presented in [34].

References

1. Abad P, Puente V, Gregorio JA, Prieto P (2007) Rotary router: An efficient architecture for CMP interconnection networks. In: Proceedings of the international symposium on computer architecture, June 2007
2. Adriahtantenaina A, Greiner A (2003) Micro-network for SoC: implementation of a 32-Port SPIN network. In: Proceedings of design, automation and test in Europe conference, March 2003

3. Al Faruque MA, Ebi T, Henkel J (2007) Run-time adaptive on-chip communication scheme. In: Proceedings of IEEE/ACM international conference on computer-aided design (ICCAD'07), San Jose, California, USA, 26–31, 2007
4. Amory AM, Briao E, Cota E, Lubaszewski M, Moraes FG (2005) A scalable test strategy for network-on-chip routers. In: Proceedings of IEEE international test conference, Nov 2005
5. Angiolini F, Meloni P, Carta S, Benini L, Raffo L (2006) Contrasting a NoC and a traditional interconnect fabric with layout awareness. In: Proceedings of design, automation and test in Europe conference, March 2006
6. Angiolini F, Atienza D, Murali S, Benini L, De Micheli G (2006) Reliability support for on-chip memories using networks-on-chip. In: Proceedings of the international conference on computer design, Oct 2006
7. Ascia G, Catania V, Palesi M (2004) Multi-objective mapping for mesh-based NoC architectures. In: Proceedings of international conference on hardware-software codesign and system synthesis, Sept 2004
8. Bainbridge W, Furber S (2001) Delay insensitive system-on-chip interconnect using 1-of-4 data encoding. In: Proceedings of international symposium on asynchronous circuits and systems, March 2001
9. Balfour J, Dally WJ (2006) Design tradeoffs for tiled CMP on-chip networks. In: Proceedings of the international conference on supercomputing, June 2006
10. Banerjee N, Vellank P, Chatha KS (2004) A power and performance model for network-on-chip architectures. In: Proceedings of design, automation and test in Europe conference, Feb 2004
11. Bartic TA et al (2003) Highly scalable network on chip for reconfigurable systems. In: Proceedings of international symposium system-on-chip, Nov 2003
12. Beerel P, Roncken ME (Dec. 2007) Low power and energy efficient asynchronous design. *J Low Power Electron* 3(3):234–253
13. Beigne E, Clermidy F, Vivet P, Clouard A, Renaudin M (2005) An asynchronous NOC architecture providing low latency service and its multi-level design framework. In: Proceedings of international symposium on asynchronous circuits and systems, May 2005
14. Beigne E, Clermidy F, Miermont S, Vivet P (2008) Dynamic voltage and frequency scaling architecture for units integration with a GALS NoC. In: Proceedings of IEEE international symposium on network on chip, 2008
15. Benini L, De Micheli G (Jan. 2002) Networks on chips: a new SoC paradigm. *IEEE Comput* 35(1):70–78
16. Bertozzi D, Benini L, De Micheli G (2005) Error control schemes for on-chip communication links: the energy-reliability tradeoff. *IEEE Trans Comput Aided Des Integr Circuits Syst* 24(6):818–831
17. Bertozzi S, Acquaviva A, Bertozzi D, Poggiali A (2006) Supporting task migration in multi-processor systems-on-chip: a feasibility study. In: Proceedings of design, automation and test in Europe conference March 2006
18. Bhojwani P, Lee JD, Mahapatra R (2007) SAPP: scalable and adaptable peak power management in NoCs. In: Proceedings of international symposium on low power electronic devices, Aug 2007
19. Bienia C, Kumar S, Singh JP, Li K (2008) The PARSEC benchmark suite: characterization and architectural implications. Princeton University Technical Report TR-811-08, Jan 2008
20. Bjerregaard T, Sparso J (2005) A router architecture for connection-oriented service guarantees in the MANGO clockless network-on-chip. In: Proceedings of design, automation and test in Europe conference, March 2005
21. Bjerregaard T, Mahadevan S (2006) A survey of research and practices of Network-on-chip. *ACM Comput Surv* 38(1):1–51
22. Bjerregaard T, Stensgaard MB, Sparso J (2007) A scalable, timing-safe, network-on-chip architecture with an integrated clock distribution method. In: Proceedings of design, automation and test in Europe conference, April 2007

23. Bogdan P, Dumitras T, Marculescu R (2007) Stochastic communication: a new paradigm for fault-tolerant networks-on-chip. Hindawi VLSI design, special issue on networks-on-chip, vol 2007, Hindawi Publishing Corporation
24. Bogdan P, Marculescu R (2010) Workload characterization and its impact on multicore platform design. In: Proceedings of 8th IEEE/ACM/IFIP international conference on hardware/software codesign and system synthesis (CODES/ISSS), 2010
25. Bolotin E, Cidon I, Ginosar R, Kolodny A (Feb. 2004) QNoC: QoS architecture and design process for network on chip. *J Syst Architecture (EUROMICRO J)* 50(2–3):105–128
26. Bolotin E, Cidon I, Ginosar R, Kolodny A (2007) Routing table minimization for irregular mesh NoCs. In: Proceedings of design, automation and test in Europe conference, April 2007
27. van den Brand JW, Ciordas C, Goossens K, Basten T (2007) Congestion-controlled best-effort communication for networks-on-chip. In: Proceedings of design, automation and test in Europe conference, April 2007
28. Campobello G, Castano M, Ciofi C, Mangano D (2006) GALS networks on chip: a new solution for asynchronous delay-insensitive links. In: Proceedings of design, automation and test in Europe conference, March 2006
29. Carloni LP, McMillan KL, Sangiovanni-Vincentelli AL (Sep. 2001) Theory of latency-insensitive design. *IEEE Trans Comput Aided Des Integr Circuits Syst* 20(9):1059–1076
30. Catania V, Holsmark R, Kumar S, Palesi M (2006) A methodology for design of application specific deadlock-free routing algorithms for NoC systems. In: Proceedings of CODES-ISSS, Oct 2006
31. Chan J, Parameswaran S (2005) NoCEE: energy macro-model extraction methodology for network on chip routers. In: Proceedings the of international conference on computer aided design, Nov 2005
32. Chan SC, Shepard KL, Restle PJ (2003) Design of resonant global clock distributions. In: Proceedings of the international conference on computer design, Oct 2003
33. Chang MF et al (2008) CMP network-on-chip overlaid with multi-band RF-interconnect. In: Proceedings of the international symposium on high-performance computer architecture, Feb 2008
34. Chatterjee S, Kishinevsky M, Ogras UY (2010) Quick formal modeling of communication fabrics to enable verification. In: Proceedings of IEEE international high level design validation and test workshop, 42–49 June 2010
35. Chelcea T, Nowick SM (2000) A low latency fifo for mixed-clock systems. In: Proceedings of IEEE computer society workshop on VLSI, April 2000
36. Chen X, Peh L (2003) Leakage power modeling and optimization in interconnection networks. In: Proceedings of the international symposium on low power electronics and design, Aug 2003
37. Chou C-L, Ogras UY, Marculescu R (2008) Energy- and performance-aware incremental mapping for networks-on-chip with multiple voltage levels. *IEEE Trans Comput Aided Des Integr Circuits Syst (TCAD)* 27(10):1866–1879
38. Cortadella J, Kishinevsky M, Grundmann B (2006) Synthesis of synchronous elastic architectures. In: Proceedings of design, automation conference, July 2006
39. Coskun AK, Rosing TS, Whisnant K (2007) Temperature aware task scheduling in MPSoCs. In: Proceedings of design, automation and test in Europe conference, April 2007
40. Dally WJ, Towles B (2004) Principles and practices of interconnection networks. Morgan Kaufmann Press, San Francisco
41. Dally WJ, Towles B (2001) Route packets, not wires: on-chip interconnection networks. In: Proceedings of design automation conference, June 2001
42. Dally WJ (1992) Virtual-channel flow control. *IEEE Trans Parallel Distrib Syst* 3(2):194–205
43. Dielissen J, Radulescu A, Goossens K, Rijpkema E (2003) Concepts and implementation of the Philips network-on-chip. In: Proceedings of IP-based SoC design, 2003

44. Dobkin R, Ginosar R, Sotiriou C (2004) Data synchronization issues in GALS SoCs. In: Proceedings of international symposium on asynchronous circuits and systems, April 2004
45. Draper J, Ghosh J (1994) A comprehensive analytical model for wormhole routing in multicomputer systems. *J Parallel Distrib Comput* 23(2):202–214
46. Duato J et al (2005) A new scalable and cost-effective congestion management strategy for lossless multistage interconnection networks. In: Proceedings of the international symposium on high-performance computer architecture, Feb 2005
47. Duato J, Yalamanchili S, Ni L (2002) Interconnection networks: an engineering approach. Morgan Kaufmann, San Mateo, CA
48. Enright-Jerger N, Peh L, Lipasti M (2008) Circuit-switched coherence. In: Proceedings of the international symposium networks-on-chips, May 2008
49. Enright-Jerger N, Peh L-S, Lipasti M (2008) Virtual circuit tree multicasting: a case of on-chip hardware multicast support. In: Proceedings of ISCA, June 2008
50. Enright-Jerger N, Peh L (2009) On-chip networks. Synthesis lecture. Morgan-Claypool Publishers
51. Easley N, Peh L (2004) High-level power analysis for on-chip networks. International conference on compilers, architectures and synthesis for embedded systems, Sep 2004
52. Ejlali A, Al-Hashimi BM, Rosinger P, Miremadi SG (2007) Joint consideration of fault-tolerance, energy-efficiency and performance in on-chip networks. In: Proceedings of design, automation and test in Europe conference, April 2007
53. Genko N, De Micheli G, Atienza D, Mendias J, Hermida R, Catthoor F (2005) A complete network-on-chip emulation framework. In: Proceedings of design, automation and test in Europe conference, March 2005
54. Goldfeder CM (2005) Frequency-based code placement for embedded multiprocessors. In: Proceedings of design automation conference, July 2005
55. Goossens K et al (2005) A design flow for application-specific networks-on-chip with guaranteed performance to accelerate SoC design and verification. In: Proceedings of design, automation and test in Europe conference, March 2005
56. Gratz P, Kim C, McDonald R, Keckler SW, Burger DC (2006) Implementation and evaluation of on-chip network architectures. In: Proceedings of international conference on computer design, Oct 2006
57. Grecu C, Pande PP, Wang B, Ivanov A, Saleh R (2005) Methodologies and algorithms for testing switch-based NoC interconnects. In: Proceedings of international symposium on defect and fault tolerance in VLSI systems, Oct 2005
58. Grecu C, Ivanov A, Pande P, Jantsch A, Salminen E, Ogras UY, Marculescu R (2007) An initiative towards open network-on-chip benchmarks. NoC benchmarking white paper, 2007. <http://www.ocpip.org/uploads/documents/NoC-Benchmarks-WhitePaper-15.pdf>
59. Gruian F (2001) Hard real-time scheduling for low energy using stochastic data and DVS processors. In: Proceedings of international symposium on low-power electronics and design, Aug 2001
60. Guerrier P, Greiner A (2000) A generic architecture for on-chip packet switched interconnections. In: Proceedings of design, automation and test in Europe conference, March 2000
61. Guz Z, Walter I, Bolotin E, Cidon I, Ginosar R, Kolodny A (2006) Efficient link capacity and QoS design for wormhole network-on-chip. In: Proceedings of design, automation and test in Europe conference, March 2006
62. Hansson A, Goossens K, Radulescu A (2007) A unified approach to mapping and routing on a network-on-chip for both best-effort and guaranteed service traffic. Hindawi VLSI Design, Hindawi Publishing Corporation
63. Harmanci M, Escudero N, Leblebici Y, Ienne P (2005) Quantitative modeling and comparison of communication schemes to guarantee quality-of-service in networks-on-chip. In: Proceedings of the international symposium on circuits and systems, May 2005

64. Hemani A, Jantsch A, Kumar S, Postula A, Oberg J, Millberg M, Lindvist D (2000) Network on a chip: an architecture for billion transistor era. In: Proceedings of the IEEE NorChip conference, Nov 2000
65. Ho WH, Pinkston TM (2003) A methodology for designing efficient on-chip interconnects on well-behaved communication patterns. In: Proceedings of the international symposium on high-performance computer, architecture, Feb 2003
66. Hosseinabady M, Dalirsani A, Navabi Z (2007) Using the inter- and intra-switch regularity in NoC switch testing. In: Proceedings of design, automation and test in Europe conference, April 2007
67. Hu P, Kleinrock L (1997) An analytical model for wormhole routing with finite size input buffers. 15th International teletraffic congress, June 1997
68. Hu J, Marculescu R (2003) Energy-aware mapping for tile-based NoC architectures under performance constraints. In: Proceedings of ASP-DAC, Jan 2003
69. Hu J, Marculescu R (2005) Communication and task scheduling of application-specific networks-on-chip. *IEE Proc comput Digital Tech* 152(5):643–651
70. Hu J, Marculescu R (2004) DyAD—Smart routing for networks-on-chip. In: Proceedings of design automation conference, June 2004
71. Hu J, Marculescu R (2005) Energy- and performance-aware mapping for regular NoC architectures. *IEEE Trans Comput Aided Des Integr Circuits Syst* 24(4):551–562
72. Hu J, Ogras UY, Marculescu R (2006) System-level buffer allocation for application-specific networks-on-chip router design. *IEEE Trans Comput Aided Des Integr Circuits Syst* 25(12):2919–2933
73. Hung W et al (2004) Thermal-aware IP virtualization and placement for Networks-on-Chip architecture. In: Proceedings of ICCD, 2004
74. Ibrahim KZ (2005) Correlation between detailed and simplified simulations in studying multiprocessor architecture. In: Proceedings of international conference on computer design, Oct 2005
75. Jantsch A, Lauter R, Vitkowski A (2005) Power analysis of link level and end-to-end data protection in networks on chip. In: Proceedings of the international symposium on circuits and systems, May 2005
76. Jantsch A, Tenhunen H (eds) (2003) *Networks-on-Chip*. Norwell, MA, Kluwer
77. Jose AP, Patounakis G, Shepard KL (2005) Near speed-of-light on-chip interconnects using pulsed current-mode signaling. In: Proceedings of symposium on VLSI Circuits, June 2005
78. Kim B, Stojanovic V (2007) Equalized interconnects for on-chip networks: modeling and optimization framework. International conference on computer-aided design, Nov 2007
79. Kim D, Kim K, Kim J, Lee S, Yoo H (2007) Solutions for real chip implementation issues of NoC and their application to memory-centric NoC. In: Proceedings of international symposium on networks-on-chips, May 2007
80. Kim EJ et al (2003) Energy optimization techniques in cluster interconnects. In: Proceedings of the international symposium on low power electronics and design, Aug 2003
81. Kim M, Kim D, Sobelman GE (2005) Adaptive scheduling for CDMA-based networks-on-chip. In: Proceedings of the IEEE northeast workshop on circuits and systems, May 2005
82. Kim JS, Taylor MB, Miller J, Wentzlaff D (2003) Energy characterization of a tiled architecture processor with on-chip networks. In: Proceedings of the international symposium on low power electronics and design, Aug 2003
83. Kim J, Nicopoulos CA, Park D, Vijaykrishnan N, Yousif MS, Das CR (2006) A gracefully degrading and energy-efficient modular router. In: Proceedings of the international symposium on computer architecture, June 2006
84. Kim J et al (2007) A novel dimensionally-decomposed router for on-chip communication in 3D architectures. In: Proceedings of the international symposium on computer architecture, June 2007
85. Kim J, Dally WJ, Abts D (2007) Flattened butterfly: a cost-efficient topology for high-radix networks. In: Proceedings of ISCA, June 2007

86. Kogel T et al (2003) A modular simulation framework for architectural exploration of on-chip interconnection networks. In: Proceedings of international conference on hardware-software codesign and system, synthesis, Oct 2003
87. Kogel T, Leupers R, Meyr H (2006) Integrated system-level modeling of network-on-chip enabled multi-processor platforms. Springer, New York
88. Kumar A, Peh L, Kundu P, Jha NK (2007) Express virtual channels: Towards the ideal interconnection fabric. In: Proceedings of the international symposium on computer architecture, June 2007
89. Kumar R, Zyuban V, Tullsen DM (2005) Interconnections in multi-core architectures: understanding mechanisms, overheads and scaling. In: Proceedings of the international symposium on computer architecture, June 2005
90. Lahiri K et al (2000) Evaluation of the traffic-performance characteristics of system-on-chip communication architectures. In: Proceedings of the international conference on VLSI design, Oct 2000
91. Lee HG, Chang N, Ogras UY, Marculescu R (2007) On-chip communication architecture exploration: a quantitative evaluation of point-to-point, bus and network-on-chip approaches. *ACM Trans Des Autom Electron Syst* 12(3):1–20
92. Lee K et al (2004) A 51mW 1.6GHz on-chip network for low-power heterogeneous SoC platform. *International Solid-State Circuits Conference*, Feb 2004
93. Lee JW, Ng A, Asanovic K (2008) Globally-synchronized frames for guaranteed quality of service in on-chip networks. In: *International symposium on computer architecture*, 2008
94. Leung LF, Tsui CY (2006) Optimal link scheduling on improving best-effort and guaranteed services performance in network-on-chip system. In: *Proceedings of design automation conference*, July 2006
95. Liang J, Laffely A, Srinivasan S, Tessier R (2004) An architecture and compiler for scalable on-chip communication. *IEEE Trans Very Large Scale Integr Syst* 12(7):711–726
96. Lin T, Pileggi LT (2002) Throughput-driven IC communication fabric synthesis. In: *Proceedings of the international conference on computer aided design*, 2002
97. Liu C, Shi J, Cota E, Iyengar V (2005) Power-aware test scheduling in network-on-chip using variable-rate on-chip clocking. In: *Proceedings of VLSI test symposium*, May 2005
98. Lu Z, Liu M, Jantsch A (2007) Layered switching for networks on chip. In: *Proceedings of design automation conference*, June 2007
99. Luo J, Jha NK (2000) Power-conscious joint scheduling of periodic task graphs and aperiodic tasks in distributed real-time embedded systems. In: *Proceedings of international conference on computer-aided design*, Nov 2000
100. Madsen J, Mahadevan S, Virk K, Gonzales M (2003) Network-on-chip modeling for system-level multiprocessor simulation. In: *Proceedings of the IEEE international real-time systems symposium*, 82–92, Dec 2003
101. Mahadevan S et al (2005) A network traffic generator model for fast network-on-chip simulation. In *Proceedings of design, automation and test in Europe conference*, March 2005
102. Mak TS, Sedcole P, Cheung PY, Luk W, Lam KP (2007) A hybrid analog-digital routing network for NoC dynamic routing. In: *Proceedings of the international symposium on networks-on-chip*, May 2007
103. Manolache S, Eles P, Peng Z (2005) Fault and energy-aware communication mapping with guaranteed latency for applications implemented on NoC. In: *Proceedings design automation conference*, July 2005
104. Marescaux T, Corporaal H (2007) Introducing the superGT network-on-chip. In: *Proceedings of design automation conference*, June 2007
105. Marculescu R, Ogras UY, Peh L, Jerger NE, Hoskote Y (2009) Outstanding research problems in NoC design: system, microarchitecture, and circuit perspectives. *IEEE Trans Comput Aided Des Integr Circuits Syst* 28(1):3–21

106. Martin S, Flautner K, Mudge T, Blaauw D (2002) Combined dynamic voltage scaling and adaptive body biasing for lower power microprocessors under dynamic workloads. In: Proceedings of international conference on computer aided design, Nov 2002
107. De Micheli G, Benini L (eds) (2006) Networks on chips: technology and tools (systems on silicon). Morgan Kaufmann, San Francisco
108. Millberg M, Nilsson E, Thid R, Jantsch A (2004) Guaranteed bandwidth using looped containers in temporally disjoint networks within the Nostrum network on chip. In: Proceedings of design, automation and test in Europe conference, Feb 2004
109. Mishra R et al (2003) Energy aware scheduling for distributed real-time systems. International parallel and distributed processing symposium, April 2003
110. Miskov-Zivanov N, Marculescu D (2010) Multiple transient faults in combinational and sequential circuits: a systematic approach. IEEE Trans CAD Integr Cir Syst 29(10):1614–1627
111. Mullins R, West A, Moore S (2004) Low-latency virtual-channel routers for on-chip networks. In: Proceedings of international symposium on computer architecture, June 2004
112. Murali S et al (2006) Designing application-specific networks on chips with floorplan information. In: Proceedings of ICCAD, Nov 2006
113. Murali S, De Micheli G (2004) Bandwidth-constrained mapping of cores onto NoC architectures. In: Proceedings of design, automation and test in Europe conference, Feb 2004
114. Murali S, Atienza D, Benini L, De Micheli G (2007) A method for routing packets across multiple paths in NoCs with in-order delivery and fault-tolerance guarantees. Hindawi VLSI Des 2007:11
115. Murali S et al (2005) Analysis of error recovery schemes for networks on chip. IEEE design and test of computers, 2005
116. Murali S, Coenen M, Radulescu A, Goossens K, De Micheli G (2006) A methodology for mapping multiple use-cases onto networks on chips. In: Proceedings of design automation and test in Europe conference, March 2006
117. Nicopoulos CA et al (2006) ViChaR: a dynamic virtual channel regulator for network-on-chip routers. In: Proceedings of the international symposium on microarchitecture, Dec 2006
118. Nigussie E, Lehtonen T, Tuuna S, Plosila J, Isoaho J (2007) High-performance long NoC link using delay-insensitive current-mode signaling. Hindawi VLSI Des (special issue on networks-on-chip) 2007:1–13
119. Nilsson E, Millberg M, Oberg J, Jantsch A (2003) Load distribution with the proximity congestion awareness in a network on chip. In: Proceedings of design, automation and test in Europe conference, March 2003
120. Connor IO, Gaffiot F (2004) Advanced research in on-chip optical interconnects. In: Piguet C (ed) Lower Power electronics and design, CRC Press
121. OCP International Partnership, http://www.ocpip.org/university_research_bibliography.php
122. Ogras UY, Marculescu R (2005) Energy- and performance-driven NoC communication architecture synthesis using a decomposition approach. In: Proceedings of design, automation and test in Europe conference, March 2005
123. Ogras UY, Marculescu R (2006) It's a small world after all": NoC performance optimization via long-range link insertion. IEEE Trans Very Large Scale Integr Syst Spec Sect Hardw Softw Codesign Syst Synth 14(7):693–706
124. Ogras UY, Marculescu R (2006) Prediction-based flow control for network-on-chip traffic. In: Proceedings of design automation conference, July 2006
125. Ogras UY, Marculescu R (2006) Communication-based design for nanoscale SoCs. In: Chen W-K (ed) VLSI handbook, 2nd edn. CRC Book Press
126. Ogras UY, Marculescu R, Marculescu D, Jung EG (2009) Design and management of voltage-frequency island partitioned networks-on-chip. IEEE Trans Very Large Scale Integr Syst 17(3):330–341

127. On-Chip Networks Bibliography, <http://www.cl.cam.ac.uk/~rdm34/onChipNetBib/browser.htm>
128. Ost L, Mello A, Palma J, Moraes F, Calazans N (2005) MAIA: a framework for networks on chip generation and verification. In: Proceedings of Asia South Pacific design automation conference, Jan 2005
129. Palermo G, Silvano C (2004) PIRATE: a framework for power/performance exploration of network-on-chip architectures. In: Proceedings of international workshop on power and timing modeling, optimization and simulation, Sept 2004
130. Pamunuwa D, Öberg J, Zheng LR, Millberg M, Jantsch A, Tenhunen H (2003) Layout, performance and power trade-offs in mesh-based network-on-chip architectures. In: IFIP international conference on very large scale integration, Dec 2003
131. Pande PP, Grecu C, Jones M, Ivanov A, Saleh R (Aug. 2005) Performance evaluation and design trade-offs for network-on-chip interconnect architectures. *IEEE Trans Comput* 54(8):1025–1040
132. Ganguly A et al (2010) Scalable hybrid wireless network-on-chip architectures for multi-core systems. *IEEE Trans Comput* 60(10):1485–1502
133. Patel CS, Chai SM, Yalamanchili S, Schimmel DE (1997) Power constrained design of multiprocessor interconnection networks. In: Proceedings of the international conference on computer design, Oct 1997
134. Peh L, Dally WJ (2000) Flit-reservation flow control. In: Proceedings of the international symposium on high-performance computer architecture, Jan 2000
135. Peh L, Dally WJ (2001) A delay model for router micro-architectures. *IEEE Micro*
136. Pham D et al (2005) The design and implementation of a first-generation CELL processor. In: Proceedings of the solid-state circuits conference, Feb 2005
137. Pinto A, Carloni LP, Sangiovanni-Vincentelli AL (2003) Efficient synthesis of networks on chip. In: Proceedings of international conference on computer design, Oct 2003
138. Pirretti M, Link GM, Brooks RR, Vijaykrishnan N, Kandemir M, Irwin MJ, (2004) Fault tolerant algorithms for network-on-chip interconnect. In: Proceedings of IEEE symposium on VLSI, Feb 2004
139. Pop P et al (2001) An approach to incremental design of distributed embedded systems. In: Proceedings of design automation conference, June 2001
140. Poplavko P, Basten T, Bekooij M, van Meerbergen J, Mesman B (2003) Task-level timing models for guaranteed performance in multiprocessor networks-on-chip. In: Proceedings of the international conference on compilers, architecture and synthesis for embedded systems, 2003
141. Puente V, Gregorio JA, Vallejo F, Beivide R (2004) Immunet: a cheap and robust fault-tolerant packet routing mechanism. In: Proceedings of the international symposium on computer, architecture, June 2004
142. Pullini A, Angiolini F, Bertozzi D, Benini L (2005) Fault tolerance overhead in network-on-chip flow control schemes. In: Proceedings of symposium on integrated circuits and system design, Sep 2005
143. Saastamoinen I, Alho M, Nurmi J (2003) Buffer implementation for proteo network-on-chip. In: Proceedings of international symposium on circuits and systems, May 2003
144. Salaun G, Serwe W, Thonnart Y, Vivet P (2007) Formal verification of CHP specifications with CADP illustration on an asynchronous network-on-chip. In: Proceedings of the IEEE international symposium on asynchronous circuits and systems, 2007
145. Scherrer A, Fraboulet A, Risset T (2006) Automatic phase detection for stochastic on-chip traffic generation. In: Proceedings International Conference on Hardware-Software Codesign, Oct 2006, pp 88–93
146. Schmitz MT, Al-Hashimi BM, Eles P (2004) Iterative schedule optimization for voltage scalable distributed embedded systems. *ACM Trans Embedded Comput Syst* 3(1):182–217. doi:10.1145/972627.972636

147. Seo D, Ali A, Lim W, Rafique N, Thottethodi M (2005) Near-optimal worst-case throughput routing for two-dimensional mesh networks. In: Proceedings of the international symposium on computer, architecture, June 2005
148. Sgroi M et al (2001) Addressing the system-on-a-chip interconnect woes through communication-based design. In: Proceedings of design automation conference, June 2001
149. Shacham A, Bergman K, Carloni LP (2007) The case for low-power photonic networks-on-chip. In: Proceedings of design automation conference, June 2007
150. Shang L, Peh L, Jha NK (2003) Dynamic voltage scaling with links for power optimization of interconnection networks. In: Proceedings of the international symposium on high-performance computer, architecture, Jan 2003
151. Shang L, Peh L, Kumar A, Jha N K (2004) Thermal modeling, characterization and management of on-chip networks. In: Proceedings of international symposium on microarchitecture, Dec 2004
152. Sheibanyrad A, Panades IM, Greiner A (2007) Systematic comparison between the asynchronous and the multi-synchronous implementations of a network-on-chip architecture. In: Proceedings of design, automation and test in Europe conference, April 2007
153. Shibayama A, Nose K, Torii S, Mizuno M, Eda Hiro M (2007) Skew-tolerant global synchronization based on periodically all-in-phase clocking for multi-core soc platforms. In: Proceedings of symposium on VLSI circuits, June 2007
154. Shim B, Shanbhag NR (2006) Energy-efficient soft-error tolerant digital signal processing. *IEEE Trans VLSI* 14(4):336–348
155. Shin D, Kim J (2004) Power-aware communication optimization for networks-on-chips with voltage scalable links. In: Proceedings of international conference on hardware/software codesign and system synthesis, Sept 2004
156. Shin D, Kim J, Lee S (2001) Intra-task voltage scheduling for low-energy, hard real-time applications. *IEEE Des Test* 18(2):20–30
157. Shivakumar P, Kistler M, Keckler S, Burger D, Alvisi L (2002) Modeling the effect of technology trends on soft error rate of combinational logic. In: Proceedings of the international conference on dependable systems and networks, June 2002
158. Simunic T, Boyd S (2002) Managing power consumption in networks on chip. In: Proceedings of design, automation and test in Europe conference, March 2002
159. Simunic Rosing T, Mihic K, De Micheli G (2007) Power and reliability management of SOCs. *IEEE Trans on VLSI* 15:391–403
160. Soteriou V, Wang H-S, Peh L (2006) A statistical traffic model for on-chip interconnection networks. In: Proceedings of the international symposium on modeling, analysis and simulation of computer and telecommunication systems, Sept 2006
161. Soteriou V, Peh L (2004) Design space exploration of power-aware on/off interconnection networks. In: Proceedings of the ICCD, Oct 2004
162. Srinivasan K, Chatha KS (2005) A technique for low energy mapping and routing in network-on-chip architectures. In: Proceedings of the international symposium on low power electronics and design, Aug 2005
163. Srinivasan K, Chatha KS (2006) A low complexity heuristic for design of custom network-on-chip architectures. In: Proceedings of design, automation and test in Europe conference, March 2006
164. Srinivasan K, Chatha KS, Konjevod G (2006) Linear programming based techniques for synthesis of network-on-chip architectures. *IEEE Trans on Very Large Scale Integr Syst* 14(4):407–420
165. Stuijk S, Basten T, Geilen M, Ghamarian AH, Theelen B (2008) Resource-efficient routing and scheduling of time-constrained streaming communication on networks-on-chip. *J Syst Architect (the EUROMICRO Journal)* 54(3–4):411–426
166. Sun C, Shang L, Dick RP (2007) Three-dimensional multi-processor system-on-chip thermal optimization. In: Proceedings of international conference on hardware/software codesign and system synthesis, Oct 2007

167. Taylor MB et al (2002) The Raw microprocessor: A computational fabric for software circuits and general purpose programs. *IEEE Micro*
168. Taylor MB, Lee W, Amarasinghe S, Agarwal A (2005) Scalar operand networks. *IEEE Trans Parallel Distrib Syst* (special issue on on-chip networks) 16(2):145–162
169. Towles B, Dally WJ (2002) Worst-case traffic for oblivious routing functions. *ACM Symp Parallel Algori Architect*
170. Vangal S et al (2007) An 80-Tile 1.28TFLOPS Network-on-Chip in 65nm CMOS. In: *Proceedings of solid-state circuits conference*, Feb 2007
171. Varatkar G, Marculescu R (2004) On-chip traffic modeling and synthesis for MPEG-2 video applications. *IEEE Trans VLSI* 12(1):108–119
172. Wang H, Peh L, Malik S (2003) Power-driven design of router microarchitectures in on-chip networks. In: *Proceedings of the international symposium on microarchitecture*, Nov 2003
173. Wang H, Zhu X, Peh L, Malik S (2002) Orion: a power-performance simulator for interconnection networks. In: *Proceedings of annual international symposium on microarchitecture*, Nov 2002
174. Wolkotte PT, Smit GJM, Kavaldjiev N, Becker JE, Becker J (2005) Energy model of networks-on-chip and bus. In: *Proceedings of the international symposium on system-on-chip*, Nov 2005
175. F. Worm, P. Ienne, P. Thiran, G. D. Micheli, “A robust selfcalibrating transmission scheme for on-chip networks. *IEEE Trans on Very Large Scale Integr Syst* 12(12):1360–1373
176. Xie Y, Wolf W (2001) Allocation and scheduling of conditional task graph in hardware/software co-synthesis. In: *Proceedings of design, automation and test in Europe conference*, March 2001
177. Yan S, Lin B (2008) Design of application-specific 3D networks-on-chip architectures. In: *Proceedings of ICCD*, 2008
178. Yu Z, Baas B (2006) Implementing tile-based chip multiprocessors with GALS clocking styles. In: *Proceedings of the international conference on computer design*, Oct 2006
179. Zhao D, Wang Y (2008) SD-MAC: design and synthesis of a hardware-efficient collision-free QoS-aware MAC protocol for wireless Network-on-Chip. *IEEE Trans Comput (TC)* 8:1046–1057



<http://www.springer.com/978-94-007-3957-4>

Modeling, Analysis and Optimization of Network-on-Chip
Communication Architectures

Ogras, U.Y.; Marculescu, R.

2013, XIV, 174 p., Hardcover

ISBN: 978-94-007-3957-4