Area and power savings via buffer reorganization in asymmetric 3D-NoCs for heterogeneous 3D-SoCs

Jan Moritz Joseph, Christopher Blochwitz, Thilo Pionteck
Universität zu Lübeck
Institute of Computer Engineering
23562 Lübeck, Germany
Email: {joseph, blochwitz, pionteck}@iti.uni-luebeck.de

Alberto García-Ortiz
University of Bremen
Institute of Electrodynamics and Microelectronics
28359 Bremen, Germany
Email: agarcia@item.uni-bremen.de

***Abstract*—In this paper, optimizations for asymmetric Network-on-Chip (NoC) router architectures are proposed for heterogeneous 3D-System-on-Chips (SoCs). The optimizations cover buffer reorganization among dies and focus on power and area savings. The architectures are compared to conventional, symmetric routers on the bases of synthesizable RTL models. Area savings of 8.3% and power savings of 5.4% for link buffers are achieved while accepting a minor average system performance loss of 2.1% in simulations. We thereby demonstrate the potentials of asymmetric NoC designs for heterogeneous 3D-SoCs.**

***Index Terms*—Network-on-Chip, heterogeneous 3D-System-on-Chip, asymmetric 3D-NoC, buffer reorganization**

I. Introduction

New production methods permit the design of heterogeneous 3D-SoCs, which consist of stacked silicon dies manufactured with different technologies interconnected by Through-Silicon Vias (TSVs). 3D-SoCs provide numerous advantages to overcome the limitations of current 2D-architectures, e.g. reduced power consumption and costs, increased performance, and heterogeneous integration [1]. The significant promises of the 3D-technology have generated a plethora of different architectures and paradigms: 3D-DRAM subsystems [2], 3D-FPGAs [3], and Vision Systems-on-Chip (VSoC) with stacked sensors [4]. There are critical challenges as well, such as thermal issues, low yield, and design complexity [5].

3D-technology allows for heterogeneous integration: The technology of each individual die is optimized for its system components, which may require different electrical characteristics. For instance, in the 3D-VSoC [4], a conservative mixed-signal technology (e.g. 130 nm) is connected with a high-speed digital technology (e.g. 65 nm). As another example, dedicated and interleaved dies with either memory or processing units offer a better performance than mixed dies [6].

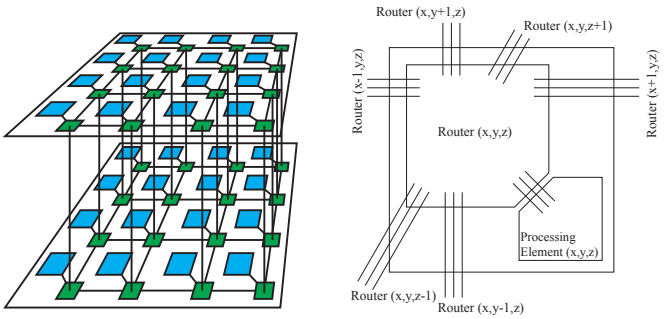
To exploit the potential of heterogeneous 3D-SoCs, 3D-NoCs offer a powerful, flexible, and scalable communication infrastructure. Most of the existing works on 3D-NoCs, as pointed out in [7], exploit incremental advantages of 3D-technology without addressing its unique features. These features, however, provide a larger optimization potential [8]. The existing works tacitly assume a multilayer homogeneous 3D-SoC that disregards the technological asymmetry intrinsically present in heterogeneous 3D-SoCs. Because of this heterogeneity, the actual cost and constraints of the communication infrastructure in each die are different. Different degrees of heterogeneity in the architecture, however, can be further exploited, e.g. the overall cost of the TSV-redundancy schemes required for yield improvement [9] are decreased. Thus, an important issue for 3D-NoC-design has not been considered yet: dedicated, asymmetric router architectures for 3D-NoCs that consider the heterogeneity of 3D-SoCs.

In this paper, we focus on buffer distribution in asymmetric 3D-NoCs (A-3D-NoCs) as one aspect to demonstrate their advantages in heterogeneous 3D-SoCs. Placing as many buffers as possible in a layer, which is optimized for memory, may be advantageous. Therefore, we introduce two novel optimizations for asymmetric router architectures with buffer reorganization among dies. We analyze their influence on architectural metrics such as area, power consumption, and performance. We put costs and performance loss into relation from a system-level point of view. The presented results show the potentials of asymmetries in communications infrastructures of heterogeneous SoCs, although we only consider a subset of asymmetric NoC design options. In addition, NoC design as presented here, which considers technological characteristics of heterogeneous SoCs, is not widely included in existing (asymmetric) 3D-NoCs.

II. Related work

A major concern of NoCs is their area footprint. An overview on different solutions is given in [10]. In general, routers are the main source of area requirements. Thus, many works try to reduce the number (e.g. [11]) or size (e.g. [8]) of routers. According to [12], the buffers within a router account for up to 75% of the overall area. Hence, numerous publications try to reduce the buffer area sizes within a router. This is done for application-specific scenarios by adapting the buffer depth [13], by reducing the number of virtual channels (VCs) [14], or by sharing buffers for several VCs and inputs/outputs [15]. The former approach additionally allows for asymmetric NoC designs by adapting the buffer depth of each router to traffic patterns [16].

When moving from 2D to 3D-system designs, the area footprint of 3D-NoC communication infrastructures becomes



(a) 3D-NoC with routers (green) and processing elements (blue)

(b) Architecture of a baseline symmetric 3D-router

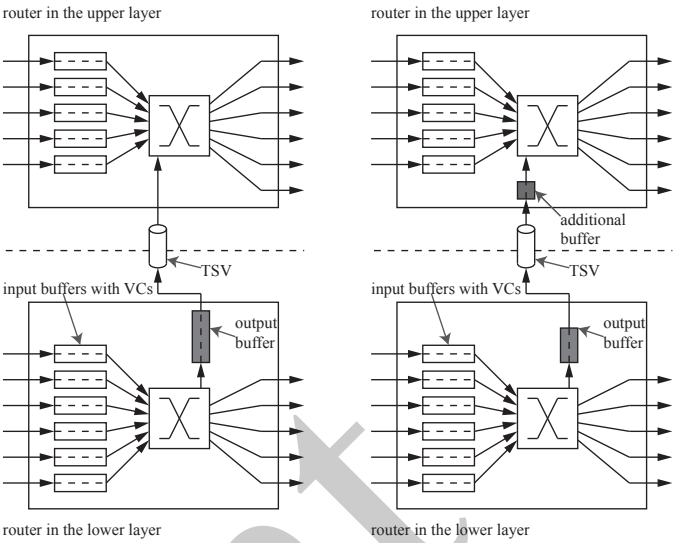
Fig. 1: 3D-NoC and symmetric router architecture

more severe. The additional dimension increases the number of ports per router and, thus, enlarges the size of the crossbar and the number of buffers. [17] showed that for standard NoC routers, the area increase per router when moving from 2D-mesh topology to 3D-mesh topology is about 50%. Apparent approaches to compensate for this are to reduce the number of ports per router, the overall number of routers [17], or the buffer depth per link [18].

In general, asymmetric router architectures can be implemented on homogeneous 3D-SoCs to decrease costs as well. For example, a standard router can be divided into multiple layers and be implemented on multiple dies [8], providing a multi-layered 3D-NoC router. Dynamically shutting down of unused layers achieves power savings of up to 67% for real workloads. In [7], dynamic link and crossbar sharing between adjacent routers decreases the latency by up to 21% compared with a standard 3D-NoC. As another example for asymmetry, different layers may incorporate different hybrid communication infrastructures, i.e. a mix of a symmetric NoC with a bus [19]. To the best of our knowledge, this the first work to evaluate the impact of asymmetric router designs, e.g. different buffer distributions per die for vertical links, while considering technology specific parameters. This extends the optimization design space and is orthogonal to all optimization approaches mentioned before.

III. Asymmetric 3D-NoC model

As a 3D-evaluation platform we, model a heterogeneous 3D-SoC consisting of two layers in a different silicon technology modeled on cycle-accurate abstraction level. Two layers are sufficient to evaluate the influence of buffer distributions. The 3D-NoC's topology is composed of stacked 4x4-grids (Fig. 1a). We assume the following simplifications although identical routers synthesized for different silicon nodes have varying clock frequencies and area footprints: The symmetric topology is not influenced since adjusted mapping and addresses for dimension order routing can be designed to match this property. Thus, it does not impact the traffic characteristics of the benchmarks and, thus, the results. Furthermore, we consider routers with synchronized clocks in both layers. We



(a) "Aggressive" router

(b) "delay oriented" router

Fig. 2: Architecture of the proposed asymmetric router pairs

assume that both layers run at the slower clock frequency, since this is the most pessimistic scenario.

A. Symmetric baseline router

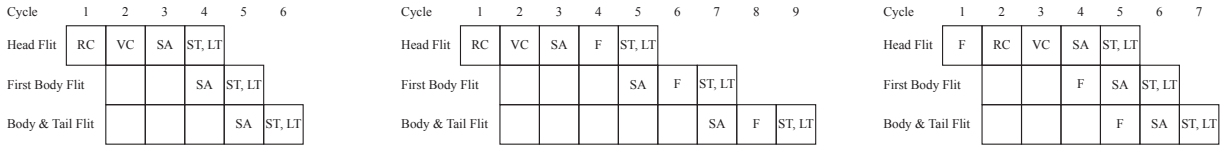
To establish a baseline for the comparison, a symmetric router model with input buffers is used. The routers are connected to each adjacent neighbor (Fig. 1b). Deterministic dimension order routing is used. The input buffers are eight flits deep. For Quality-of-Service (QoS), VCs with a lower number have a higher priority during arbitrations. The default router pipeline (Fig. 3a) is identical to conventional 3D-routers.

B. Asymmetric router architecture

In asymmetric router architecture, the (input) buffers of the upward, vertical links are placed as output buffer in the lower layer since this layer is manufactured in a smaller technology node. We do not consider single VCs individually; rather, we change the location of all VCs' buffers. The status registers stay in the upper layer. The locations of the router's remaining buffers are not modified. This results in an asymmetric architecture (Fig. 2a), in which routers on the upper layer have input buffers for all links except the incoming vertical links. In the lower layer, the routers have input buffers for all links and output buffers for the vertical link. This optimization reduces the area and power consumption of the routers and is, therefore, called "aggressive". The performance loss of the architecture is expected to be large. Thus, we introduce an additional "delay-oriented" design option, in which the buffers of the vertical link are divided and flits can be stored in an additional intermediate buffer in the upper layer (Fig. 2b).

C. Asymmetric router pipelines

The router pipeline of the routers in the lower layer (Fig. 3a) is not influenced by the buffer reorganization and is, thus,



(a) Baseline router pipeline (input buffers) (b) "aggressive" pipeline (c) "delay-oriented" pipeline

Fig. 3: Time behavior of vertical links *(RC – routing calculation, VC – VC calculation, SA – switch allocation, F – fetch, ST – switch traversal, and LT – link traversal)*

as usual. The length of the critical path for flits traveling to the upper layer is significantly shorter due to the directly connected output buffers. However, the routers in the lower layer do not have a performance advantage, since this would require a faster clock frequency for these links. This is not realistic: the faster clock area cannot be fed with incoming data. In the pipeline of the "aggressive" router architecture (Fig. 3b), an additional fetch cycle is needed in the vertical links to receive a flit, which is stored in the output buffers in the adjacent layer. Thus, flits in this link have a longer delay for the additional traversal via the TSV resulting in a critical path, which has an approximately doubled delay. Hence, these flits leave the router only at every other clock cycle. In contrast, in the "delay-oriented" router's pipeline (Fig. 3c) the fetch cycle is only required for head-of-line flits. Follow-up flits can be switched immediately as they reach the buffer in the upper layer. Since this takes place in parallel to the switch and link traversal of the prior flit, a flit can leave the router in every subsequent clock cycle after an initial delay.

IV. Results

As a case study, we estimate power and area savings using a standard router, which is synthesized from an RTL model for commercial 130 nm and 65 nm technologies. The baseline is set by the default router in a symmetric NoC with a buffer size of eight flits per VC and a flit size of 32 bit. The clock frequency of the routers depends mainly on the technology node. To evaluate the influence of the silicon technology and provide comparability to the results of our simulations, we set the buffer depth to 8 flits and use a single VC maximizing the influence of the asymmetry in the worst case. Although we assume synchronized routers in our model, we achieve different clock frequencies: For 65 nm-technology, the routers run at approximately 1 GHz and for 130 nm at 820 MHz.

A. Power savings

The total power values of the routers are given in Tab. II and were calculated using a single stacked pair of routers (with the same x- and y-coordinates). The power savings are compared with the symmetric baseline router. By moving buffers to the lower layer, the power consumption in the 65 nm-layer increases. However, single buffers consume less power in this layer reducing the overall power consumption. Using this linear dependency, the total power consumption declines by approx. 7.2% for the "aggressive", asymmetric architecture and by 5.4% for the "delay-oriented" design.

B. Area savings

The influence of area savings in the NoC router's buffers on the overall design costs depends on the RTL-design and the commercial technology. In our synthesized model of a standard router, the buffers occupied 84% of the routers area for 130 nm technology and 79% for 65 nm technology. The memory cell size, which is determined by the silicon node, is the only variable size on our router area model. Similar to the consideration for the power consumption, the area savings are a consequence of a reduced area footprint of units in the smaller technology. We measured an actual ratio of 3.7 for similar flip-flops in exemplary commercial 65 nm and 130 nm technologies. This yields 9.6% savings for the "aggressive" model and 8.3% savings for the "delay-oriented" router (see Tab. I).

C. Performance loss

We implemented a cycle-accurate simulator using SystemC, which supports abstract task graph modeling [22] with synthetic workloads and application traffic [20], [21]. The mapping prioritizes short communication distances to reduce the link load. In addition, the network load in the layer with the more conservative silicon technology is reduced. Depending on the number of tasks, not all cores of the SoC are utilized. We measured the performance of the A-3D-NoC with simulations (Tab. I). The system's performance is measured in delay of clock cycles for a single instance of inputs per application. Hence, the performance results are independent of the design's clock frequency. In general, the asymmetric architectures introduce additional latencies and the theoretical system performance decreases. The results support this. In addition, the "aggressive" router has a higher latency than the "delay-oriented" router. In consequence, the latter has a smaller performance loss with 2.1% in comparison to 14% of the former ones. Furthermore, for hot-spot traffic, there are relatively high losses of 33% and 4.9% due to the location of the global destination address, which was set at an edge of the upper layer of the NoC for a worst case approximation. Moreover, the mapping of the mp3 de- and encoder is disadvantageous since many packets travel between layers. The application's performance loss is the highest among the benchmarks with 40% and 6.4% respectively. Finally, due to the relatively small sample size, the performance for uniform traffic increases by 0.6% although the theoretical throughput of the NoC architecture decreases.

TABLE I: Benchmark results

Traffic/Application	# Tasks	# Packets	baseline router (sym.)	"aggressive" router		"delay-oriented" router	
			execution time	exec. time	perf. loss	exec. time	perf. loss
uniform (median)	32	64	1,318	1,310	-.6%	1,322	.3%
complement	32	64	170	210	19%	178	4.4%
hotspot	32	62	1,290	1,930	33%	1,656	4.9%
VOPD & Shape Dec. [20]	17	3,416	55,188	55,208	.7%	55,192	0.7%
VOPD [21]	16	4,044	63,766	69,766	8.6%	64,368	.9%
DVOPD [21]	32	8,762	65,202	77,316	16%	66,458	1.9%
MPEG-4 [21]	12	3,467	142,010	142,010	.0%	142,010	.0%
PIP [21]	8	576	9,984	11,364	11%	10,114	1.3%
MWD [21]	12	1,120	14,114	16,216	13%	14,338	1.6%
H.263 dec., mp3 dec. [21]	13	19,636	172,554	181,068	4.7%	173,232	.4%
mp3 enc., mp3 dec. [21]	12	1,652	18,500	30,752	40%	19,730	6.2%
H.263 enc., mp3 dec. [21]	14	24,291	373,778	430,884	13%	378,704	1.3%
			average performance loss		14%		**2.1%**
			area saving for router input buffers		9.6%		**8.3%**

TABLE II: Power consumption of a router pair

Router Architecture	Power Consumption			Savings
	65 nm	130 nm	Router Pair	
Symmetric	20.7 mW	58.3 mW	79 mW	0%
"Aggressive"	23.6 mW	49.7 mW	73.3 mW	7.2%
"Delay-oriented"	22.9 mW	51.9 mW	74.8 mW	**5.4%**

V. Conclusion

We propose two novel asymmetric router architectures with a focus on area savings for asymmetric 3D-NoCs. The "aggressive" router design does not compensate its area and power savings due to the large performance losses. The "delay-oriented" router model performs better and provides an advantageous trade-off between overall cost and performance. In comparison with conventional symmetric routers, we achieved area savings of 8.3% and power savings of 5.4% while accepting a minor average performance loss of 2.1%. Summing up, we demonstrated the potentials of asymmetric NoC designs for heterogeneous 3D-SoCs.

Acknowledgements

This work is funded in part by the German Research Foundation (DFG) project PI 447/5-1.

References

[1] X. Dong and Y. Xie, "System-level cost analysis and design exploration for three-dimensional integrated circuits (3d ics)," 2009.

[2] C. Weis, I. Loi, L. Benini, and N. Wehn, "Exploration and optimization of 3-d integrated dram subsystems," *TCAD*, vol. 32, no. 4, pp. 597–610, 2013.

[3] V. F. Pavlidis and E. G. Friedman, *Three-dimensional Integrated Circuit Design*. Elsevier Science, 2010.

[4] Á. Zarándy, *Focal-plane sensor-processor chips*. Springer, 2011.

[5] R. S. Patti, "Three-dimensional integrated circuits and the future of system-on-chip designs," *Proc. of the IEEE*, 2006.

[6] X. Yu, L. Li, Y. Zhang, H. Pan, and S. He, "Performance and power consumption analysis of memory efficient 3d network-on-chip architecture," *ICCA*, 2013.

[7] S. Seyyedaghaei Rezaei, A. Mazloumi, M. Modarressi, and P. Lotfi-Kamran, "Dynamic resource sharing for high-performance 3-d networks-on-chip," *LCA*, 2015.

[8] D. Park, S. Eachempati, R. Das, A. Mishra, Y. Xie, N. Vijaykrishnan, and C. Das, "Mira: A multi-layered on-chip interconnect router architecture," in *ISCA*, June 2008.

[9] C. Osewold, W. Buter, and A. Garcia-Ortiz, "A coding-based configurable and asymmetrical redundancy scheme for 3-d interconnects," *ReCoSoC*, 2014.

[10] A.-M. Rahmani, K. Latif, P. Liljeberg, J. Plosila, and H. Tenhunen, "Research and practices on 3d networks-on-chip architectures," *NORCHIP*, 2010.

[11] M. Coppola, R. Locatelli, G. Maruccia, L. Pieralisi, and A. Scandurra, "Spidergon: a novel on-chip communication network," *Int. Symp. on SoC*, 2004.

[12] P. Gratz, C. Kim, R. McDonald, S. W. Keckler, and D. Burger, "Implementation and evaluation of on-chip network architectures," *ICCD*, 2006.

[13] A. S. Kumar, M. P. Kumar, S. Murali, V. Kamakoti, L. Benini, and G. d. Micheli, "A simulation based buffer sizing algorithm for network on chips," *ISVLSI*, 2011.

[14] C. Nicopoulos, D. Park, J. Kim, N. Vijaykrishnan, M. Yousif, and C. Das, "Vichar: A dynamic virtual channel regulator for network-on-chip routers," *Microarchitecture, 2006. MICRO-39. 39th Annual IEEE/ACM Int. Symp. on*, pp. 333–346, 2006.

[15] K. Latif, A.-M. Rahmani, L. Guang, T. Seceleanu, and H. Tenhunen, "Pvs-noc: Partial virtual channel sharing noc architecture," *Parallel, Distributed and Network-Based Processing (PDP), 2011 19th Euromicro Int. Conf. on*, pp. 470–477, 2011.

[16] J. Hu, U. Y. Ogras, and R. Marculescu, "System-level buffer allocation for application-specific networks-on-chip router design," *TCAD*, 2006.

[17] B. S. Feero and P. P. Pande, "Networks-on-chip in a three-dimensional environment: A performance evaluation," *IEEE Trans. on Comp.*, 2009.

[18] Y. Ghidini, T. Webber, E. Moreno, F. Grando, R. Fagundes, and C. Marcon, "Buffer depth and traffic influence on 3d nocs performance," *Symp. RSP*, 2012.

[19] M. O. Agyeman, A. Ahmadinia, and A. Shahrabi, "Low power heterogeneous 3d networks-on-chip architectures," *HPCSim*, 2011.

[20] van der Tol, Erik B. and E. G. Jaspers, "Mapping of mpeg-4 decoding on a flexible architecture platform," *Media Processors*, 2002.

[21] P. K. Sahu and S. Chattopadhyay, "A survey on application mapping strategies for network-on-chip design," *Jour. of Sys. Arc.*, 2013.

[22] J. Joseph and T. Pionteck, "A cycle-accurate network-on-chip simulator with support for abstract task graph modeling," in *System-on-Chip (SoC), 2014 International Symposium on*, Oct 2014, pp. 1–6.