Bridging the Frequency Gap in Heterogeneous 3D SoCs through Technology-Specific NoC Router Architectures

Jan Moritz Joseph
RWTH Aachen University, Germany
joseph@ice.rwth-aachen.de

Lennart Bamberg
GrAI Matter Labs, Eindhoven
lbamberg@graimatterlabs.ai

Geonhwa Jeong
Georgia Institute of Technology, Atlanta, GA
geonhwa.jeong@gatech.edu

Ruei-Ting Chien
Georgia Institute of Technology, Atlanta, GA
rchien6@gatech.edu

Rainer Leupers
RWTH Aachen University, Germany
leupers@ice.rwth-aachen.de

Alberto Garía-Ortiz
University of Bremen, Germany
agarcia@item.uni-bremen.de

Tushar Krishna
Georgia Institute of Technology, Atlanta, GA
tushar@ece.gatech.edu

Thilo Pionteck
Otto-von-Guericke Universität Magdeburg, Germany
thilo.pionteck@ovgu.de

Abstract

In heterogeneous 3D System-on-Chips (SoCs), NoCs with uniform properties suffer one major limitation; the clock frequency of routers varies due to different manufacturing technologies. For example, digital nodes allow for a higher clock frequency of routers than mixed-signal nodes. This large frequency gap is commonly tackled by complex and expensive pseudo-mesochronous or asynchronous router architectures. Here, a more efficient approach is chosen to bridge the frequency gap. We propose to use a heterogeneous network architecture. We show that reducing the number of VCs allows to bridge a frequency gap of up to 2×. We achieve a system-level latency improvement of up to 47% for uniform random traffic and up to 59% for PARSEC benchmarks, a maximum throughput increase of 50%, up to 68% reduced area and 38% reduced power in an exemplary setting combining 15-nm digital and 30-nm mixed-signal nodes and comparing against a homogeneous synchronous network architecture. Versus asynchronous and pseudo-mesochronous router architectures, the proposed optimization consistently performs better in area, in power and the average flit latency improvement can be larger than 51%.

1 Introduction

3D integration allows to vertically integrate chips by stacking with Through-Silicon Via (TSV)-based links, by monolithic integration using monolithic inter-tier vias (MIVs) or by face-to-face bonding using a flipped chip. From a technology perspective, the optimization potential of 3D integration is large with improved performance, power and area (PPA) [6]. Ultimately, 3D integration also challenges fundamental limitations of computation by reducing compute time asymptotically from t to $t^{0.75}$ [16]. These encouraging promises make it worth tackling thermal limitation and limited yield [15].

Heterogeneous 3D integration [26] is game-changing. While the technology of all dies is equivalent in homogeneous 3D integrated circuits (3D ICs), e.g., [20], heterogeneous integration allows to align the technologies of dies with the requirements of the components integrated (e.g., processors in 15nm logic dies interleaved with memory in a 22nm die [25]). As even larger PPA gains are possible than in homogeneous 3D ICs, commercial products have been announced, e.g., Intel Lakefield with Foveros 3D technology.

3D integration allows manifold architectural innovations like new processor architectures [7] which are not possible with contemporary planar chips by using the third dimension. For interconnection architectures, there has been vast research effort, as well, especially for 3D NoCs. These works usually assume homogeneous 3D integration using identical manufacturing technologies for dies. Therefore, homogeneous and heterogeneous network architectures exist; homogeneous architectures extend the traditional 2D NoC to a third dimension primarily exploiting topology advantages [19]; heterogeneous architectures mainly target application-specific NoCs with non-uniform resource distribution, e.g., by providing more buffers and VCs in network areas with more application traffic [17]. Few works exist on architectures for heterogeneous 3D ICs.

In previous works on NoCs for heterogeneous 3D ICs [10, 12], two challenging integration issues have been identified:
Varying area: In homogeneous 3D ICs, the size of logic and memory is independent of their die. This is not the case for heterogeneous 3D ICs. In other words, *iso-architectures for routers do not yield iso-area among dies*. This can be exploited to improve PPA: In [10], buffer resources are split up among routers so that area and power are saved in mixed-signal dies. [10] improves power up to 15% and area up to 28% vs. a homogeneous network architecture.
Varying clock frequency: The achievable clock frequency differs with dies in heterogeneous 3D ICs. In other words *iso-architectures for routers do not yield iso-frequency among dies*. The resulting frequency gap can be larger than factor 16 [12]. Thus, it is suboptimal to use a homogeneous network architecture and clock the routers in the faster technology nodes below their capability as this will reduce system performance. Asynchronous or pseudo-mesochronous routers are not optimal either, because they do not fully levitate throughput bottlenecks and come at area costs [4, 12, 13, 24].

To summarize, in heterogeneous 3D ICs, NoCs encounter challenging integration issues; Especially, varying clock frequencies

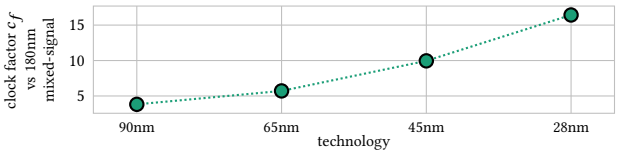


Figure 1: Different achievable clock frequency factor (c_f) for 180nm mixed-signal vs. digital nodes. Fig. based on [12]

limit system performance. In this work, we tackle this by using a heterogeneous network architectures. Specifically, we propose an optimization that uses VC-less routers in parts of the network; this allows to increase the clock frequency of these routers while being implemented in a less-scaled technology or mixed-signal node. Even though the VC-less routers provide less performance capabilities locally, the global system-performance will increase from matched clock frequencies. By that, *we bridge the frequency gap* found in NoCs for heterogeneous 3D ICs, improve PPA and reduce design complexity. To the best of the authors' knowledge, this is the first work to use a heterogeneous network architecture to tackle technology-imposed performance limitations.

2 Related Work

Many works adapt the network architecture of NoCs by optimizing the VC count and the buffer depth. STORM [21] optimizes the router architectures by leveraging traffic pattern biases, introducing destination-based VC partitioning. It offers lower latency and higher throughput at small area costs. [9] proposes a greedy algorithm to allocate buffer resources in different input channels on a 2D Mesh network to match the application traffic characteristics. ViChaR [17] utilizes a unified buffer structure, that could dynamically allocate the number of VC and the buffer depth, to fit current traffic load. Thus, ViChaR uses the application-traffic to optimize the VC count.

Due to 3D's promises, there are many works on 3D NoCs regarding network topology [2], routing algorithms [1], router architectures [21], vertical link placement [5], etc. However, the research focus of these previous works is dissimilar to this paper's approach.

3 Background

A heterogeneity in 3D SoC yields new challenges for NoC architectures, mainly varying area and clock frequency. To adopt to this, previous works advocate heterogeneous network architectures, i.e., routers that do vary among layers. [10] distributes router memory non-uniformly over layers to reduce area and power by trading a small performance loss. [12] focuses on routing algorithms to send packets along the fastest routes in the network accepting nonuniform router utilization.

One observation of [12] is that the clock frequency of uniform (homogeneous) routers synthesized for disparate technologies varies significantly. The model for the clock scaling factor c_f [12, Eq. 3] expresses this deviation. It is between 4× to 16× combining digital and mixed-signal nodes with a difference in feature size of 1:2 and 1:4 as shown in Fig. 1 (For example, this could be applied to 15 nm digital technology and 30 nm or 60 nm mixed-signal node, respectively.). This means, that the same architecture can be clocked up to 16× faster in a digital than a mixed-signal node.

This large clock-speed difference leaves two consequences. First, the latency for a packet transmission at the same distance varies

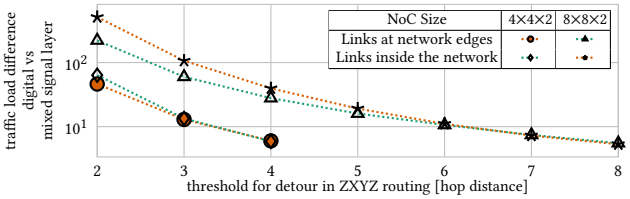


Figure 2: Load difference between the digital and the mixed-signal dies for uniform random traffic.

between dies. Second, the throughput of a packet transmission is limited by the slowest-clocked router along a path. For these systems, there are three network architectures so far:

1) Homogeneous synchronous routers are all clocked the same. Thus, the network is limited by the slowest frequency, commonly found in the layer in the mixed-signal node. This does not improve latency or throughput but is a straight-forward architecture.

2) Homogeneous asynchronous routers clock as fast as possible. This is a viable option but does not improve throughput. It requires additional logic to synchronize between layers.

3) Pseudo-mesochronous routers [12] improve throughput and latency by a technology-architecture co-optimization. Throughput limitations are tackled by the pseudo-mesochronous router architecture. It combines multiple smaller flits into multi-flit transmission using mesochronous shift registers between dies improving throughput by 2×. This is only possible at reasonable area costs because ZXYZ routing limits the number of throughput-critical paths. Latency limitations are tackled by the ZXYZ routing algorithm. In a nutshell, packets are sent along paths that offer the best latency under zero load, even if this means a detour through a neighbored faster clocked layer. This allows for an average latency improvement of 2.26× in a Vision SoC case study. However, the non-standard routing algorithm poses one important shortcoming. As packets take detours, the load on the digital layers is increased. While ZXYZ routing was proven beneficial for use cases such as camera chips, it is not applicable universally. Fig 2 shows the load difference between the digital and the mixed-signal dies in a 4×4×2 and a 8×8×2 NoC depending on the threshold for taking a detour, measured in hop counts. Links at the edges of the network and links inside the network are differentiated. The data are obtained from a mathematical model using the bisection width of the network and are cross-validated against simulations. (Later on, our simulations with uniform random traffic prove the point, cf. Fig. 5). The load difference is consistently above 5× for all evaluated scenarios, but can be up to 512×. An off-the-shelf solution for this would be resource redistribution on network level, such as ViChar [17]. It allows to increase the buffer depths and VC count in the digital layer. However, this is not applicable realistically, because the load difference is too large, effectively. Therefore, using standard XYZ routing algorithm is preferable, as done in this work.

More generally, both the asynchronous or pseudo-mesochronous architectures have two further shortcomings. First, frequency differences induce throughput bottlenecks between clock domains. Second, a synchronization logic between the different clock domains in each layer is required at additional area and performance costs. In [4], the transmission speed is reduced by one hop and the router logic is increased by 10% [4, Tab. II]. [24] has similar results. [13] requires 4.4% more total NoC area [13, Table 2]. To summarize,

Table 1: Maximum achievable clock frequency for input-buffered 3D routers in a 4×4×2 NoC.

	1 VC buffer depth ≤ 8	≥2 VCs buffer depth ≤ 8
15 nm digital node	4 GHz	2 Ghz
30 nm mixed-signal node	1 GHz	0.5 GHz

these three asynchronous 3D NoCs induces approx. ≤10% area costs. Furthermore, their custom physical design drives the design complexity. The same considerations hold for the pseudo-mesochronous architecture, as the router logic is modified. It accommodates one additional crossbar (cf. [12, Fig. 10]) and modified input-buffers [12, Figs. 13, 15, 16] increasing the router logic area by 10.6% for a design with comparable parameters to the previous ones [12, Sec. VIII-D]. The modified input-buffers drive physical design complexity: If the buffers are implemented in the digital layer, a large TSV array will be required to transmit data in parallel. This reduces yeild. If the buffers are located in the mixed-signal layer, a serial transmission is possible through a smaller TSV array but requiring the buffer to be driven by the digital layer's clock using a dedicated TSV. To summarize, the design complexity for asynchronous and pseudo-mesochronous routers are far from trivial and a network-architecture based solution is preferable.

4 Architectural Optimization

In summary of the background, we identify shortcomings of previous solutions that this work tackles with an optimized architecture:

(1) Homogeneous synchronous NoCs are globally clocked at a slow frequency.
(2) Homogeneous asynchronous NoCs yield area overheads of approx. 10% vs. synchronous routers.
(3) Heterogeneous pseudo-mesochronous NoCs yields a similar area overhead of 10% for synchronization and are not universally applicable for all traffic patterns.

The key idea of this paper is applying a *heterogeneous synchronous architecture* using VC-less routers in the mixed-signal layers. Being VC-less increases the maximum achievable clock speed so that we can use a synchronous architecture instead of paying the costs of asynchronicity. To elucidate that point further, we implemented an input-buffered 3D NoC on RTL and synthesized it using 15 nm Nangate Open Cell Library (OCL) targeting 0.5 GHz, 1 GHz, 2 GHz and 4 GHz and scale this for 30 nm mixed-signal (cross-validated data). The timing is violated for a 2 GHz target frequency when the number of VCs is ≥12 and buffer depth is ≥8 flits. With a target frequency of4 GHz, the timing constraint is violated for all except the VC-less router as a result of a less complex router architecture. The results are summarized in Table 1. We found that the buffer depth merely effects the achievable clock frequency. To conclude, we can double the clock frequency of the NoC by removing VCs in the mixed-signal layer.

Taking advantage of a globally increased clock frequency from partially VC-less routers, leads to the proposed architectural optimization shown in Table 2. In all three existing approaches, the routers have the equal number of VCs. In the homogeneous synchronous architecture, the clock frequency is limited by the slowest routers in the mixed-signal layer at 0.5 GHz. In the homogeneous asynchronous and the pseudo-mesochronous architectures, the routers in the digital layer are clocked to their full capability of 2 GHz. Our proposed optimized heterogeneous synchronous architecture removes the VCs in the mixed-signal layer to achieve an improved clock frequency of 1 GHz for the whole system.

The proposed heterogeneous architecture opt for a VC-less router. This enables a better area and power; E.g., it is well-known that VC-less routers can achieve over a magnitude of area reductions.

The next positive outcome of the proposed heterogeneous synchronous architecture is that it results in a rather straight-forward physical design. Although the performance of VC-less routers in the mixed-signal layers is reduced locally, we will show that the overall system performance increases from clock frequency gains.

To summarize, the analysis of various network architectures show that VC-less routers can achieve 2× clock frequency compared with routers with more than 2 VCs implemented in the same technology. Therefore, using VC-less routers instead of routers with VCs in mixed-signal, i.e, less-scaled, technologies is able to bridge the frequency gap present in a 3D heterogeneous integration.

4.1 Limitations

Multiprocessors often rely on request-reply communication protocols (e.g., for cache coherence), in which VCs are used to avoid request-reply protocol deadlocks. Similarly, system-level deadlock might occur despite the deadlock-free routing algorithm for dependencies in mapped task-graphs. In our VC-less architecture, this is avoided by using subactive methods, e.g., DRAIN [18]. Furthermore, VC-less routers are only used in mixed-signal layers that typically will not host cores of multiprocessors.

5 Evaluation

Area, power and network performance, i.e., maximum achievable throughput and flit latency are analyzed. Even though thermal performance plays an important role in 3D ICs in general, we do not analyze it here, because thermal performance is mainly driven by the processing elements and not the routers.

We use a two-layer 3D IC that allows to evaluate the effects of heterogeneous 3D integration that combines a 15 nm digital node with a 30 nm mixed-signal node. This is a representative example for a heterogeneous 3D IC as it has a similar difference in scaling as found in [14]. We compare the three most common 3D manufacturing techniques; stacked 3D integration using TSVs; monolithic 3D integration using MIVs; and face-to-face bonding for two layers, in which one is flipped.

We implement a 8×8×2 NoC using 3D mesh topology as an exemplary evaluation platform. The routers are input-buffered with 4 and 8 VCs. We use XYZ routing and set the packet length to 16 flit. We evaluate 4 and 8 flit deep input buffers as representative points in the design space with 25% and 50% of the packet length.

The Ratatoskr design framework [11] is used for evaluations. It provides a hardware implementation of the network. We simulate the NoC using its cycle-accurate simulator.

The optimized architecture (heterogeneous, synchronous) is compared against 1) the conventional architecture (homogeneous, synchronous), against 2) the asynchronous architecture (homogeneous,

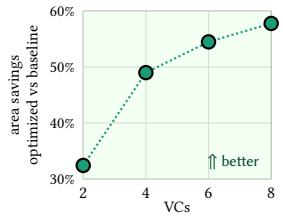


Figure 3: Area savings for 4 flit deep buffers (MIVs).

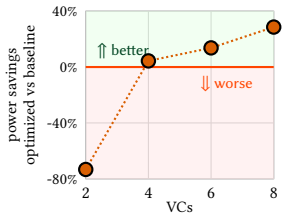


Figure 4: Power saving for 4 flit deep buffers (MIVs).

asynchronous), and against 3) the pseudo-mesochronous architecture (heterogeneous, asynchronous) [12]. The properties of the architectures are shown in Table 2, including the maximum clock frequency for the aforementioned setting based on synthesis results.

5.1 Area

We synthesize the routers for 15 nm digital node, scale the results for an exemplary 30 nm mixed-signal node (models from [12]) and cross-validate against a 28 nm node. Links are 32b wide. The area of TSVs and keep-out-zones (KOZ) is from [23]; the MIV area from [3]. The results for 4 and 8 VCs are shown in Tables 3 and 4, respectively.

5.1.1 Optimized vs. conventional Using the optimized architecture, we save 48.1% to 57.9% NoC area in comparison against the conventional one. We observe two trends in Tables 3 and 4. First, the area savings are higher for higher VCs count or buffer depths. This trend is shown in Fig. 3, in which the area savings comparing the optimized vs. the conventional architecture are plotted for a varying number of VCs. The trend is as expected, because more VCs from the conventional architecture are removed. Second, the optimized architecture saves more area using MIVs or face-to-face bonding than using TSVs; reason being the larger area for TSVs and KOZ. However, this difference between the 3D manufacturing methods is small when comparing the two architectures as the savings from buffer area by reducing the VC count outnumbers the savings from moving away from TSVs.

To summarize, 48.1% area savings (4 VCs, 4 flit deep buffers) with the optimized architecture meets expectations. As the resources on the mixed signal layer are reduced, the NoC area declines.

5.1.2 Optimized vs. Asynchronous We achieve between 48.1% and 57.9% area savings in vs. an asynchronous design. We did not synthesize a handshaking logic between the clock regions as this requires an physical design effort beyond an architectural optimization. Exemplary designs [4, 13, 24] show, the synchronization along adds a logic overhead of approx. 10%. Thus, the real area savings are expected to be larger than the ones reported. The trends in the data are identical to those seen in the previous comparison.

To summarize, the asynchronous architecture requires more area than the proposed optimized architecture.

5.1.3 Optimized vs. Pseudo-mesochronous Comparing the proposed architecture against the pseudo-mesochronous one, we observe larger area savings of 58.4% to 68.3% than in the previous evaluations. Since the pseudo-mesochronous router targets increased throughput, it requires additional logic for buffering of multi-flits and a second crossbar (cf. [12, Fig. 14]). The results here do not account for this, so that the actual area savings will be higher.

To summarize, the pseudo-mesochronous pays a hefty area-price shy of 70% for the same throughput as in the optimized architecture.

5.2 Power

We compare the static power of the optimized architecture with the conventional architecture. We do not conduct a power analysis for the other architectures as a considerable portion of the power consumption depends on the synchronization logic, which is a physical design challenge and not an architectural problem. E.g., the pseudo-mesochronous router requires a shift register in the mixed-signal layer in a separate clock region driven by the clock network of the digital layer using a vertical link to propagate the clock signal or a very large array of TSVs/MIVs, as explained above. A power analysis would require a simulation on the level of individual wires and transistors. As we do not optimize the synchronization logic here as we target an architectural optimization.

We compare the power of a 3D IC with TSVs, MIVs and face-to-face bonding against 2D IC. The RC-parasitics of wires, TSVs and MIVs vary: TSVs have a very high capacitance of about 10fF [23], which will increase energy consumption, while MIV only have about 0.2fF capacitance [22]. The power results are shown in Tables 3 and 4. We see that the power savings do not depend on the 3D integration technology at the given accuracy (rounding), because the power differences are a small constant offset to all architectures that evens out at the comparison.

5.2.1 Optimized vs. conventional The optimized network architecture saves between 4.4% to 37.7% power for the given configuration with 4/8 VCs and 4/8 flit deep buffers. The power savings are higher for more VCs and an increased buffer depth. This trend is generally similar to the one for area. However, there is an important difference; Fig. 4 plots the power savings comparing the optimized vs. the conventional architecture for a varying number of VCs. While we are able to save area for 2 VCs (cf. Fig. 3), this is not the case for power, in which the NoC draws nearly 80% more. Reason being that the higher clocked router with a single VC in the mixed signal layer draws a disproportional higher amount of power because it is clocked at the very limit of the circuit for that technology node.

To summarize, the optimized architecture allows to save 4% to 38% power vs. the conventional one despite of its higher clock frequency. However, power savings are not given for all parameters, as the higher clock may result in disproportionate power consumption in mixed-signal layers.

5.2.2 Optimized vs. Asynchronous and vs. Pseudo-mesochronous One can expect that the power consumption of both the asynchronous and the pseudo-mesochronous router are higher than the conventional router as of the synchronization logic. Therefore, the proposed optimization improves the power consumption in general.

5.3 Performance

For the performance evaluation, we use synthetic uniform random traffic pattern (urand) and PARSEC benchmarks (Netrace [8]). The results are shown in Tables 3 and 4 for urand traffic (averaging from 0.02% to 1.8% injection rates in packets/cycle for 16 flits/packet) and in Fig. 6 for the PARSEC.

5.3.1 Optimized vs. conventional The average latency improvement for urand traffic comparing the optimized vs. the conventional

Table 2: Architectural options of a 3D NoC using exemplary 3D IC with 15 nm digital and 30 nm mixed-signal nodes.

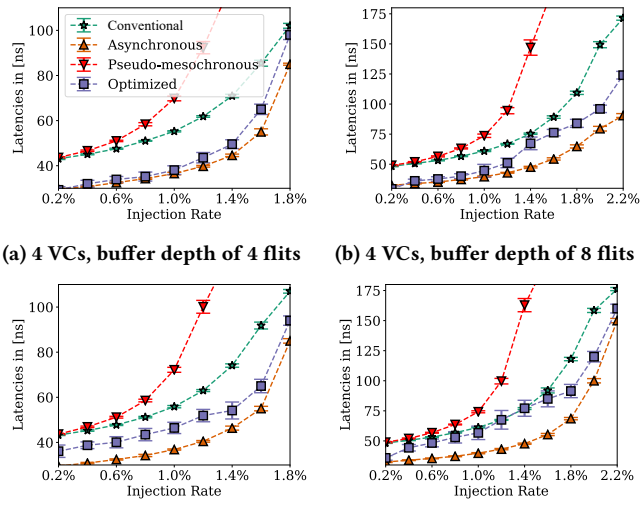
	Conventional	**Asynchronous**	**Pseudo-mesochronous [12]**	**Optimized**
Architecture	Homogeneous	Homogeneous	Heterogeneous	Heterogeneous
Timing	Synchronous	Asynchronous	Asynchronous	Synchronous
15 nm digital	2-8 VCs, ≤ 8 Buffer Depth	2-8 VCs, ≤ 8 Buffer Depth	2-8 VCs, ≤ 8 Buffer Depth	2-8 VCs, ≤ 8 Buffer Depth
	0.5 GHz	2 GHz	2 GHz	1 GHz
30 nm mixed signal	2-8 VCs, ≤ 8 Buffer Depth	2-8 VCs, ≤ 8 Buffer Depth	2-8 VCs, ≤ 8 Buffer Depth	1 VC, ≤ 8 Buffer Depth
	0.5 GHz	0.5 GHz	0.5 GHz	1 GHz

Table 3: PPA gains and losses for a 8×8×2 NoC with 4 VCs.

4 VCs		**Buffer Depth = 4 flits**				**Buffer Depth = 8 flits**			
		Δ *Area*	Δ *Power*	Δ *Flit lat. urand (avg)*	Δ *Max throughput*	Δ *Area*	Δ *Power*	Δ *Flit lat. urand (avg)*	Δ *Max throughput*
Optimized vs. conventional	*TSV*	-48.1%	-4.4%	+30.6%	+50%	-48.8%	-15.9%	+46.9%	+50%
	MIV	-49.0%	-4.4%			-49.3%	-15.9%		
	face-to-face	-49.0%	-4.4%			-49.3%	-15.9%		
Optimized vs. asynchronous	*TSV*	≥ -48.1%	n.a.	-6.1%	+50%	≥ -48.8%	n.a.	-19.2%	+50%
	MIV	≥ -49.0%				≥ -49.3%			
	face-to-face	≥ -49.0%				≥ -49.3%			
Optimized vs. pseudo-mesochronous [12]	*TSV*	≥ -58.4%	n.a.	+45.4%	+0%	≥ -59.2%	n.a.	‘+43.0%	+0%
	MIV	≥ -59.4%				≥ -59.7%			
	face-to-face	≥ -59.4%				≥ -59.7%			

Table 4: PPA gains and losses for a 8×8×2 NoC with 8 VCs.

8 VCs		**Buffer Depth = 4 flits**				**Buffer Depth = 8 flits**			
		Δ *Area*	Δ *Power*	Δ *Flit lat. urand(avg)*	Δ *Max throughput*	Δ *Area*	Δ *Power*	Δ *Flit lat. urand (avg)*	Δ *Max throughput*
Optimized vs. conventional	*TSV*	-57.3%	-28.6%	+20.9%	+50%	-57.6%	-37.7%	+6.1%	+50%
	MIV	-57.8%	-28.6%			-57.9%	-37.7%		
	face-to-face	-57.8%	-28.6%			-57.9%	-37.7%		
Optimized vs. asynchronous	*TSV*	≥ -57.3%	n.a.	-21.3%	+50%	≥ -57.6%	n.a.	-46.2%	+50%
	MIV	≥ -57.8%				≥ -57.9%			
	face-to-face	≥ -57.8%				≥ -57.9%			
Optimized vs. pseudo-mesochronous [12]	*TSV*	≥ -67.7%	n.a.	+37.6%	+0%	≥ -68.0%	n.a.	+30.4%	+0%
	MIV	≥ -68.2%				≥ -68.3%			
	face-to-face	≥ -68.2%				≥ -68.3%			



(a) 4 VCs, buffer depth of 4 flits **(b) 4 VCs, buffer depth of 8 flits**

(c) 8 VCs, buffer depth of 4 flits **(d) 8 VCs, buffer depth of 8 flits**

Figure 5: Flit latency, uniform random traffic (8x8x2 NoC).

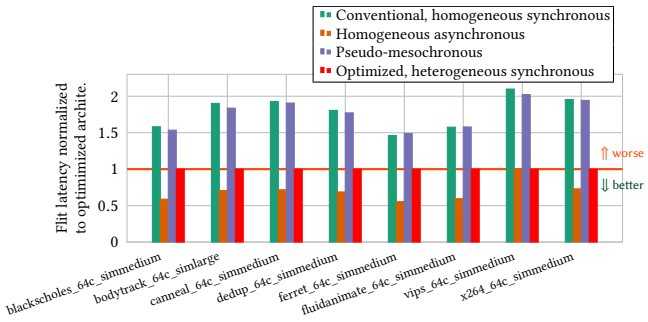
architecture is shown in Tables 3 and 4. The maximum achievable throughput for the optimized architecture is +50% better than that of the conventional one as easily calculated from the clock speeds.

For urand traffic, the optimized architecture achieves 20.9% to 46.9% better average latency. A more detailed view is given in Fig. 5 for different injection rates. The conventional (red) is consistently slower than the optimized architecture (blue). Only for 8 VCs and 8 flit deep buffers, the performance is similar for medium injection rates between 1% to 1.5% packets/cycle.

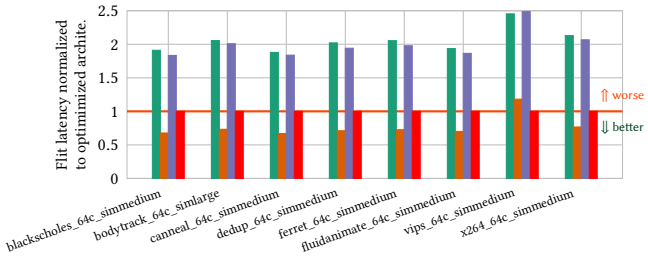
The results for PARSEC benchmarks are shown in Fig. 6. Comparing the optimized vs. the conventional architecture, the minimum improvement of flit latency is 31% for *ferret* with 4 VCs and 4 flits deep buffers. The maximum improvement of 59% is observed for *vips* with 4 VCs and 8 flits deep buffers. The average improvement in latency for individual benchmarks ranges from 42.5% to 51.5%; the improvement is larger for more VCs and a larger buffer depth.

To summarize, the optimized architecture outperforms the conventional architecture in terms of flit latency and maximum achievable throughput. There are some cases where the latency is on par, but the area and power savings are the highest in these cases.

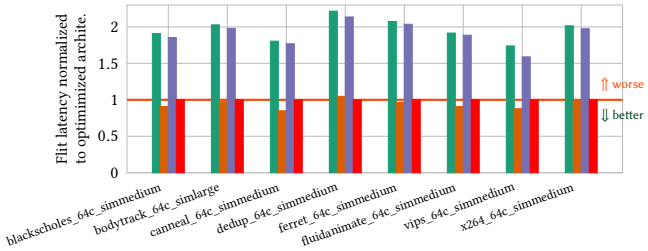
5.3.2 Optimized vs. Asynchronous For urand traffic, the optimized vs. the asynchronous architecture yields a between 6.1% to 46.2% worse



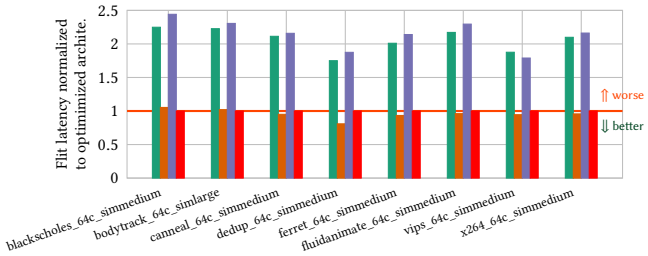
(a) 4 VCs, buffer depth of 4 flits



(b) 4 VCs, buffer depth of 8 flits



(c) 8 VCs, buffer depth of 4 flits



(d) 8 VCs, buffer depth of 8 flits

Figure 6: Flit latency for PARSEC benchmarks (8x4x2 NoC).

average latency, cf. Fig. 5. While this is clearly a large performance degradation, the optimized architectures still offers a 50% better maximum achievable throughput. To further analyze the impact of the traffic pattern, we use PARSEC benchmarks. Here, the flit latency of the optimized router offers in worst case 55% reduced (for *ferret* at 4 VCs and 4 flit deep buffers) and in best case 18% improved (for *vips* with 4 VCs and 8 flits deep buffers). Averaging over all benchmarks, the flits latency of the optimized network architecture is between 4.5% to 31.5% worse.

Although the performance results comparing the optimized and the asynchronous architecture are ambiguous, the area savings of approx. 48% - 58%, the power savings of approx. ≥ 4% - 38% and the improved throughput make the proposed architecture a compelling low-power and low-area alternative.

5.3.3 Optimized vs. Pseudo-mesochronous The optimized architecture offers better average latency between 30.4% to 45.4% and the same maximum achievable throughput as the pseudo-mesochronous architecture for urand traffic. For PARSEC, the optimized architecture has between 33% to 59.7% better flit latency (for *fluidanimate* at 4 VCs and 4 flit deep buffers and for *vips* at 4 VCs and 8 flit deep buffers, respectively). Averaging all benchmarks, the optimized architecture offers 41% to 53% better flit latency.

The pseudo-mesochronous router comes at costs. First, area and power are worse than the proposed architecture because the synchronization logic is complex; at least 58.4% additional area is required (see above). Second, the synchronization logic is complex in physical design and thus poses a severe challenge for system integration. Third, the pseudo-mesochronous router requires to use ZXYZ routing [12], which increases the load on the digital layer considerably. Thus, the network saturates faster as shown in Fig. 5. The results in Fig. 6 show that the performance benefits for the pseudo-mesochronous router demonstrated in [12] depend on the application. Since it targets Vision-SoCs, it may not be favorable in other scenarios than presented here.

To summarize, the optimized router offers consistently a better flit latency than the pseudo-mesochronous router for the benchmarks evaluated. Moreover, it saturates slower and offers smaller area and lower power. Therefore, the proposed solution is a more efficient architecture for a wider range of applications.

6 Conclusion

The maximum achievable clock speed of the identical router architecture in heterogeneous 3D ICs varies with manufacturing technology. Conventional synchronous network architectures are limited in performance while asynchronous network architectures yield large practical overheads. In this work, a simple yet striking architectural optimization for 3D NoCs in heterogeneous 3D ICs is proposed to tackle the problem. The clock frequency of the routers in the less-scaled or mixed signal layers can be increased by using VC-less routers. This allows to close down the frequency gap at reduced implementation costs. Specifically, in comparison to a homogeneous synchronous baseline, the proposed architecture shows 6.1%-46.9% better average flit latency, 50% better throughput at 48.1%-57.9% reduced area and 4.4%-37.7% reduced power (for exemplary 8×8×2 NoCs with 4 or 8 VCs and 4 or 8 flit deep buffers). In comparison to a homogeneous asynchronous architecture, the average flit latency difference depends on the application; an 31.5% loss to an improvement of up to 4.5% can be observed on average. Furthermore, the proposed architecture achieves 50% better maximum throughput, reduces area by at least 48%, reduces power by at least 4.4% and removes the need for a complex physical design of a synchronization logic, in which the cost are not included to evaluate conservatively. Finally, in comparison to a pseudo-mesochronous router architecture, the proposed optimized architecture achieves 30.4%-51.5% better average flit latency, and at least 58.4% better area. Furthermore, the simpler physical design complexity is reduced so that one can either reduce the number of vertical links to improve yield or implement a simpler clock network with less clock regions.

Acknowledgments

This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 328514428. Furthermore, it was supported by a fellowship within the IFI programme of the German Academic Exchange Service (DAAD).

References

[1] A. B. Ahmed and A. B. Abdallah. 2012. LA-XYZ: Low Latency, High Throughput Look-Ahead Routing Algorithm for 3D-NoC Architecture. In *MCSoC*.

[2] M. Bahmani, A. Sheibanyrad, F. Petrot, F. Dubois, and P. Durante. 2012. A 3D-NoC Router Implementation Exploiting Vertically-Partially-Connected Topologies. *2012 IEEE Comp. Society Annual* (2012).

[3] K. Chang, S. Sinha, B. Cline, G. Yeric, and S. K. Lim. 2016. Match-making for Monolithic 3D IC: Finding the right technology node. In *DAC*.

[4] F. Darve, A. Sheibanyrad, P. Vivet, and F. Petrot. 2011. Physical Implementation of an Asynchronous 3D-NoC Router Using Serial Vertical Links. In *2011 IEEE Computer Society Annual Symposium on VLSI*. 25–30.

[5] S. Das, J. R. Doppa, D. H. Kim, P. P. Pande, and K. Chakrabarty. 2015. Optimizing 3D NoC design for energy efficiency: A machine learning approach. In *ICCAD*.

[6] X. Dong and Y. Xie. 2009. System-level cost analysis and design exploration for three-dimensional integrated circuits (3D ICs). *ASPDAC* (2009).

[7] B Gopireddy and J Torrellas. 2019. Designing Vertical Processors in Monolithic 3D. In *ISCA*.

[8] J. Hestness, B. Grot, and S.W. Keckler. 2010. Netrace: Dependency-Driven Trace-Based Network-on-Chip Simulation. In *NOCS*.

[9] J. Hu, U.Y. Ogras, and R. Marculescu. 2006. System-level buffer allocation for application-specific networks-on-chip router design. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 25, 12 (2006).

[10] J.M. Joseph, C. Blochwitz, A. García-Ortiz, and T. Pionteck. 2017. Area and power savings via asymmetric organization of buffers in 3D-NoCs for heterogeneous 3D-SoCs. *Microprocessors and Microsystems* 48 (2017), 36–47.

[11] J.M. et al. Joseph. 2019. Ratatoskr: An open-source framework for in-depth power, performance and area analysis in 3D NoCs. arXiv:cs.AR/1912.05670

[12] J. M. Joseph, L. Bamberg, D. Ermel, B. R. Perjikolaei, A. Drewes, A. García-Ortiz, and T. Pionteck. 2019. NoCs in Heterogeneous 3D SoCs: Co-Design of Routing Strategies and Microarchitectures. *IEEE Access* 7 (2019), 135145–135163.

[13] A. Karthikeyan and P.S. Kumar. 2018. GALS implementation of randomly prioritized buffer-less routing architecture for 3D NoC. *Cluster Computing* (2018).

[14] S. Khushu and W. Gomes. 2019. Lakefield: Hybrid Cores in a Three Dimensional Package. In *HotChips*.

[15] P. Leduc, F. de Crecy, M. Fayolle, B. Charlet, T. Enot, M. Zussy, B. Jones, J. Barbe, N. Kernevez, N. Sillon, S. Maitrejean, D. Louis, and G. Passemard. 2007. Challenges for 3D IC integration: bonding quality and thermal management. In *IITC*. 210–212.

[16] I. L. Markov. 2014. Limits on fundamental limits to computation. *Nature* (2014).

[17] C. Nicopoulos, D. Park, J. Kim, N. Vijaykrishnan, M. Yousif, and C. Das. 2006. ViChaR: A Dynamic Virtual Channel Regulator for NoC Routers. *MICRO*.

[18] M. Parasar, H. Farrokhbakht, N. Enright Jerger, P.V. Gratz, T. Krishna, and J. San Miguel. 2020. DRAIN: Deadlock Removal for Arbitrary Irregular Networks. In *HPCA*.

[19] V. F. Pavlidis and E. G. Friedman. 2007. 3-D Topologies for Networks-on-Chip. *IEEE Trans. on Very Large Scale Integration (VLSI) Systems* 15, 10 (2007), 1081–1090.

[20] S. Pentapati, L. Zhu, L. Bamberg, D. E. Shim, A. García-Ortiz, and S. K. Lim. 2019. A Logic-on-Memory Processor-System Design With Monolithic 3-D Technology. *IEEE Micro* 39, 6 (2019), 38–45.

[21] S. Rasheed, P. V. Gratz, S. Shakkottai, and J. Hu. 2014. STORM: A Simple Traffic-Optimized Router Microarchitecture for Networks-on-Chip. In *NOCS*.

[22] S. K. Samal, D. Nayak, M. Ichihashi, S. Banna, and S. K. Lim. 2016. Monolithic 3D IC vs. TSV-based 3D IC in 14nm FinFET technology. In *S3S*.

[23] Taigon Song, Chang Liu, Yarui Peng, and Sung Kyu Lim. 2013. Full-chip Multiple TSV-to-TSV Coupling Extraction and Optimization in 3D ICs. In *DAC*.

[24] P. Vivet, D. Dutoit, Y. Thonnart, and F. Clermidy. 2011. 3D NoC using through silicon Via: An asynchronous implementation. In *VLSI-SoC*.

[25] X. Yu, L. Li, Y. Zhang, H. Pan, and S. He. 2013. Performance and power consumption analysis of memory efficient 3D Network-on-Chip architecture. *ICCA*.

[26] L. Zhu, L. Bamberg, A. Agnesina, F. Catthoor, D. Milojevic, M. Komalan, J. Ryckaert, A. Garcia-Ortiz, and S. K. Lim. 2020. Heterogeneous 3D Integration for a RISC-V System with STT-MRAM. *IEEE Computer Architecture Letters* (2020).