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

Wang, P.

Chapter 7

Summary and Conclusion

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

By powering off the idle components/routers in a NoC, power gating is an effective way to reduce the power consumption of a NoC. However, when the power gating is applied on a NoC, the powered-off components/routers block the packet transmission and cause significant packet latency increase. This is because the powered-off components/routers need some clock cycles to be fully charged (i.e., to be powered-on). During the time period of charging powered-off routers, some packets cannot be transferred and have to be blocked until the powered-off routers are fully charged. As a consequence, applying power gating on a NoC causes significant packet latency increase. Furthermore, the power gating process (i.e., switching off/on the power of components/routers) itself consumes extra power. This implies that frequent power gating or power gating in a short time may cause more power consumption or inefficient power consumption reduction. Thus, to reduce the packet latency increase caused by power gating and achieve significant reduction of the power consumption in NoCs, we have proposed three novel power gating approaches: duty buffer based (DB-based) power gating, dynamic bypass (D-bypass) power gating, and express virtual channel based (EVC-based) power gating. These power gating approaches are

effective in reducing the power consumption of NoCs, but with different properties, they have different advantages. We summarize the properties of the DB-based power gating approach (DB_PG), the D-bypass power gating (D-bypass), and the EVC-based power gating approach (EVC_PG) in Figure 7.1. In Figure 7.1, the axes PL_l, PL_m, and PL_h represent the packet latency (PL) in a NoC under low traffic workloads (l), medium traffic workloads (m), and high traffic workloads (h), respectively. The axes PC_l, PC_m, and PC_h represent the power consumption (PC) of a NoC under low traffic workloads (l), medium traffic workloads (h), and high traffic workloads (h), respectively. For example, the PL_m axis crosses the block edges of DB_PG, D-bypass, and EVC_PG at three points, respectively. These points represent the packet latency (normalized to the same baseline) of DB_PG, D-bypass, and EVC_PG under medium traffic workloads. Thus, according to Figure 7.1, under medium traffic workloads, DB_PG has the highest packet latency among our three approaches, whereas EVC_PG has the lowest packet latency. Based on the different properties of our power gating approaches, shown in Figure 7.1, we draw the following conclusions:

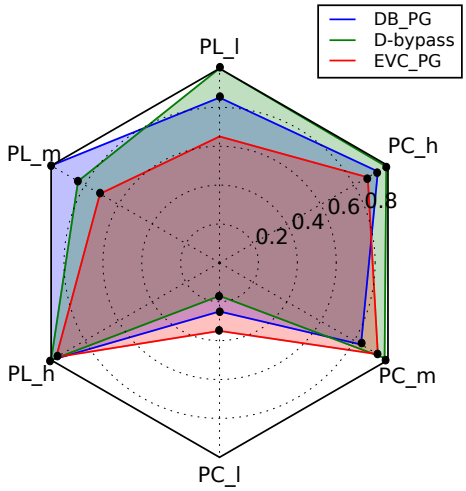

Figure 7.1: Packet latency (PL) and power consumption (PC) at low traffic workloads (l), medium traffic workloads (m), and high traffic workloads (h).

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

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

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

sequence, our D-bypass power gating approach has the highest packet latency among our three approaches at low traffic workloads.

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

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

To overcome this issue, we have proposed a novel routing approach, called Surfing on a Bufferless NoC (Surf-Bless). **Based on our Surf-Bless routing approach, it becomes possible for bufferless NoC to support a confined-interference communication. Furthermore, our Surf-Bless routing approach is much more power/energy-efficient than related approaches.** This is because, our Surf-Bless approach is based

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

Bibliography

[AABC18] Fawaz Alazemi, Arash Azizimazreah, Bella Bose, and Lizhong Chen. Routerless network-on-chip. In *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pages 492–503. IEEE, 2018.

[AKPJ09] Niket Agarwal, Tushar Krishna, Li-Shiuan Peh, and Niraj K Jha. Garnet: A detailed on-chip network model inside a full-system simulator. In *2009 IEEE international symposium on performance analysis of systems and software*, pages 33–42. IEEE, 2009.

[BBB+11] Nathan Binkert, Bradford Beckmann, Gabriel Black, Steven K Reinhardt, Ali Saidi, Arkaprava Basu, Joel Hestness, Derek R Hower, Tushar Krishna, Somayeh Sardashti, et al. The gem5 simulator. *ACM SIGARCH Computer Architecture News*, 39(2):1–7, 2011.

[BD14] James Balfour and William J Dally. Design tradeoffs for tiled cmp on-chip networks. In *ACM International Conference on Supercomputing 25th Anniversary Volume*, pages 390–401. ACM, 2014.

[BDM02] Luca Benini and Giovanni De Micheli. Networks on chips: A new soc paradigm. *computer*, 35(1):70–78, 2002.

[BEA+08] Shane Bell, Bruce Edwards, John Amann, Rich Conlin, Kevin Joyce, Vince Leung, John MacKay, Mike Reif, Liewei Bao, John Brown, et al. Tile64-processor: A 64-core soc with mesh interconnect. In *2008 IEEE International Solid-State Circuits Conference-Digest of Technical Papers*, pages 88–598. IEEE, 2008.

[BHW+17] Rahul Boyapati, Jiayi Huang, Ningyuan Wang, Kyung Hoon Kim, Ki Hwan Yum, and Eun Jung Kim. Fly-over: A light-weight distributed power-gating mechanism for energy-efficient networks-on-chip. In *2017 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 708–717. IEEE, 2017.

[BJ14] Mario Badr and Natalie Enright Jerger. Synfull: synthetic traffic models capturing cache coherent behaviour. In *2014 ACM/IEEE 41st International Symposium on Computer Architecture (ISCA)*, pages 109–120. IEEE, 2014.

[BJS+14] Haseeb Bokhari, Haris Javaid, Muhammad Shafique, Jörg Henkel, and Sri Parameswaran. darknoc: Designing energy-efficient network-on-chip with multi-vt cells for dark silicon. In *Proceedings of the 51st Annual Design Automation Conference*, pages 1–6. ACM, 2014.

[BKA10] Ali Bakhoda, John Kim, and Tor M Aamodt. Throughput-effective on-chip networks for manycore accelerators. In *Proceedings of the 2010 43rd annual IEEE/ACM international symposium on microarchitecture*, pages 421–432. IEEE Computer Society, 2010.

[BKSL08] Christian Bienia, Sanjeev Kumar, Jaswinder Pal Singh, and Kai Li. The parsec benchmark suite: Characterization and architectural implications. In *Proceedings of the 17th international conference on Parallel architectures and compilation techniques*, pages 72–81. ACM, 2008.

[Boh07] Mark Bohr. A 30 year retrospective on dennard's mosfet scaling paper. *IEEE Solid-State Circuits Society Newsletter*, 12(1):11–13, 2007.

[Bor07] Shekhar Borkar. Thousand core chipsa technology perspective. In *2007 44th ACM/IEEE Design Automation Conference*, pages 746–749. IEEE, 2007.

[BS00] J Adam Butts and Gurindar S Sohi. A static power model for architects. In *Proceedings 33rd Annual IEEE/ACM International Symposium on Microarchitecture. MICRO-33 2000*, pages 191–201. IEEE, 2000.

[CJ16] Xianmin Chen and Niraj K Jha. Reducing wire and energy overheads of the smart noc using a setup request network. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 24(10):3013–3026, 2016.

[CP12] Lizhong Chen and Timothy M Pinkston. Nord: Node-router decoupling for effective power-gating of on-chip routers. In *Proceedings of the 2012 45th Annual IEEE/ACM International Symposium on Microarchitecture*, pages 270–281. IEEE Computer Society, 2012.

[CPK+13] Chia-Hsin Owen Chen, Sunghyun Park, Tushar Krishna, Suvinay Subramanian, Anantha P Chandrakasan, and Li-Shiuan Peh. Smart: a single-cycle reconfigurable noc for soc applications. In *2013 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 338–343. IEEE, 2013.

[CZPP15] Lizhong Chen, Di Zhu, Massoud Pedram, and Timothy M Pinkston. Power punch: Towards non-blocking power-gating of noc routers. In *2015 IEEE 21st International Symposium on High Performance Computer Architecture (HPCA)*, pages 378–389. IEEE, 2015.

[CZPP16] Lizhong Chen, Di Zhu, Massoud Pedram, and Timothy M Pinkston. Simulation of noc power-gating: Requirements, optimizations, and the agate simulator. *Journal of Parallel and Distributed Computing*, 95:69–78, 2016.

[CZZ+15] Hsiang-Yun Cheng, Jia Zhan, Jishen Zhao, Yuan Xie, Jack Sampson, and Mary Jane Irwin. Core vs. uncore: the heart of darkness. In *2015 52nd ACM/EDAC/IEEE Design Automation Conference (DAC)*, pages 1–6. IEEE, 2015.

[DCS+14] Bhavya K Daya, Chia-Hsin Owen Chen, Suvinay Subramanian, Woo-Cheol Kwon, Sunghyun Park, Tushar Krishna, Jim Holt, Anantha P Chandrakasan, and Li-Shiuan Peh. Scorpio: a 36-core research chip demonstrating snoopy coherence on a scalable mesh noc with in-network ordering. In *2014 ACM/IEEE 41st International Symposium on Computer Architecture (ISCA)*, pages 25–36. IEEE, 2014.

[DGR+74] Robert H Dennard, Fritz H Gaensslen, V Leo Rideout, Ernest Bassous, and Andre R LeBlanc. Design of ion-implanted mosfet's with very small physical dimensions. *IEEE Journal of Solid-State Circuits*, 9(5):256–268, 1974.

[DMMD09] Reetuparna Das, Onur Mutlu, Thomas Moscibroda, and Chita R Das. Application-aware prioritization mechanisms for on-chip networks. In *Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture*, pages 280–291. ACM, 2009.

[DNSD13] Reetuparna Das, Satish Narayanasamy, Sudhir K Satpathy, and Ronald G Dreslinski. Catnap: energy proportional multiple network-on-chip. In *ACM SIGARCH Computer Architecture News*, pages 320–331. ACM, 2013.

[DT01] William J Dally and Brian Towles. Route packets, not wires: on-chip interconnection networks. In *Proceedings of the 38th annual Design Automation Conference*, pages 684–689. Acm, 2001.

[DT04] William James Dally and Brian Patrick Towles. *Principles and practices of interconnection networks*. Elsevier, 2004.

[EBA+11] Hadi Esmaeilzadeh, Emily Blem, Renee St Amant, Karthikeyan Sankaralingam, and Doug Burger. Dark silicon and the end of multicore scaling. In *2011 38th Annual international symposium on computer architecture (ISCA)*, pages 365–376. IEEE, 2011.

[FCM11] Chris Fallin, Chris Craik, and Onur Mutlu. Chipper: A low-complexity bufferless deflection router. In *2011 IEEE 17th International Symposium on High Performance Computer Architecture*, pages 144–155. IEEE, 2011.

[FDC+09] David Fick, Andrew DeOrio, Gregory Chen, Valeria Bertacco, Dennis Sylvester, and David Blaauw. A highly resilient routing algorithm for fault-tolerant nocs. In *Proceedings of the Conference on Design, Automation and Test in Europe*, pages 21–26. European Design and Automation Association, 2009.

[FKM+02] Krisztián Flautner, Nam Sung Kim, Steve Martin, David Blaauw, and Trevor Mudge. Drowsy caches: simple techniques for reducing leakage power. In *ACM SIGARCH Computer Architecture News*, pages 148–157. IEEE Computer Society, 2002.

[FLY+16] Haohuan Fu, Junfeng Liao, Jinzhe Yang, Lanning Wang, Zhenya Song, Xiaomeng Huang, Chao Yang, Wei Xue, Fangfang Liu, Fangli Qiao, et al. The sunway taihulight supercomputer: system and applications. *Science China Information Sciences*, 59(7):072001, 2016.

[FTKH16] Hossein Farrokhbakht, Mohammadkazem Taram, Behnam Khaleghi, and Shaahin Hessabi. Toot: an efficient and scalable power-gating method for noc routers. In *2016 Tenth IEEE/ACM International Symposium on Networks-on-Chip (NOCS)*, pages 1–8. IEEE, 2016.

[FYNM11] Chris Fallin, Xiangyao Yu, Gregory Nazario, and Onur Mutlu. A high-performance hierarchical ring on-chip interconnect with low-cost routers. *Computer Architecture Lab, Carnegie Mellon Univ, Tech. Rep*, 7:2011, 2011.

[Gal97] Mike Galles. Spider: A high-speed network interconnect. *IEEE Micro*, 17(1):34–39, 1997.

[GDR05] Kees Goossens, John Dielissen, and Andrei Radulescu. Æthereal network on chip: concepts, architectures, and implementations. *IEEE Design & Test of Computers*, 22(5):414–421, 2005.

[GH10] Kees Goossens and Andreas Hansson. The aethereal network on chip after ten years: Goals, evolution, lessons, and future. In *Design Automation Conference*, pages 306–311. IEEE, 2010.

[GHKM09] Boris Grot, Joel Hestness, Stephen W Keckler, and Onur Mutlu. Express cube topologies for on-chip interconnects. In *2009 IEEE 15th International Symposium on High Performance Computer Architecture*, pages 163–174. IEEE, 2009.

[GHKM11] Boris Grot, Joel Hestness, Stephen W Keckler, and Onur Mutlu. Kilo-noc: a heterogeneous network-on-chip architecture for scalability and service guarantees. In *ACM SIGARCH Computer Architecture News*, pages 401–412. ACM, 2011.

[GKS+07] Paul Gratz, Changkyu Kim, Karthikeyan Sankaralingam, Heather Hanson, Premkishore Shivakumar, Stephen W Keckler, and Doug Burger. On-chip interconnection networks of the trips chip. *IEEE Micro*, 27(5):41–50, 2007.

[GN92] Christopher J Glass and Lionel M Ni. The turn model for adaptive routing. *ACM SIGARCH Computer Architecture News*, 20(2):278–287, 1992.

[HDH+10] Jason Howard, Saurabh Dighe, Yatin Hoskote, Sriram Vangal, David Finan, Gregory Ruhl, David Jenkins, Howard Wilson, Nitin Borkar, Gerhard Schrom, et al. A 48-core ia-32 message-passing processor with dvfs in 45nm cmos. In *2010 IEEE International Solid-State Circuits Conference-(ISSCC)*, pages 108–109. IEEE, 2010.

[HGBH09] Andreas Hansson, Kees Goossens, Marco Bekooij, and Jos Huisken. Compsoc: A template for composable and predictable multi-processor system on chips. *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, 14(1):2, 2009.

[HGR07] Andreas Hansson, Kees Goossens, and Andrei Rădulescu. Avoiding message-dependent deadlock in network-based systems on chip. *VLSI design*, 2007, 2007.

[HHS+00] Lance Hammond, Benedict A Hubbert, Michael Siu, Manohar K Prabhu, Michael Chen, and K Olukolun. The stanford hydra cmp. *IEEE micro*, 20(2):71–84, 2000.

[HJK+00] Ahmed Hemani, Axel Jantsch, Shashi Kumar, Adam Postula, Johnny Oberg, Mikael Millberg, and Dan Lindqvist. Network on chip: An architecture for billion transistor era. In *Proceeding of the IEEE NorChip Conference*, volume 31, page 11, 2000.

[HSG09] Andreas Hansson, Mahesh Subburaman, and Kees Goossens. aelite: A flit-synchronous network on chip with composable and predictable services. In *Proceedings of the conference on design, automation and test in Europe*, pages 250–255. European Design and Automation Association, 2009.

[HVS+07] Yatin Hoskote, Sriram Vangal, Arvind Singh, Nitin Borkar, and Shekhar Borkar. A 5-ghz mesh interconnect for a teraflops processor. *IEEE Micro*, 27(5):51–61, 2007.

[HY13] Syed Minhaj Hassan and Sudhakar Yalamanchili. Centralized buffer router: A low latency, low power router for high radix nocs. In *2013 Seventh IEEE/ACM International Symposium on Networks-on-Chip (NoCS)*, pages 1–8. IEEE, 2013.

[JBB+13] Nan Jiang, James Balfour, Daniel U Becker, Brian Towles, William J Dally, George Michelogiannakis, and John Kim. A detailed and flexible cycle-accurate network-on-chip simulator. In *Performance Analysis of Systems and Software (ISPASS), 2013 IEEE International Symposium on*, pages 86–96. IEEE, 2013.

[KBD07] John Kim, James Balfour, and William Dally. Flattened butterfly topology for on-chip networks. In *Proceedings of the 40th Annual IEEE/ACM International Symposium on Microarchitecture*, pages 172–182. IEEE Computer Society, 2007.

[KCKP13] Tushar Krishna, Chia-Hsin Owen Chen, Woo Cheol Kwon, and Li-Shiuan Peh. Breaking the on-chip latency barrier using smart. In *2013 IEEE 19th International Symposium on High Performance Computer Architecture (HPCA)*, pages 378–389. IEEE, 2013.

[KKC+08] Tushar Krishna, Amit Kumar, Patrick Chiang, Mattan Erez, and Li-Shiuan Peh. Noc with near-ideal express virtual channels using global-line communication. In *2008 16th IEEE Symposium on High Performance Interconnects*, pages 11–20. IEEE, 2008.

[KKS+07] Amit Kumary, Partha Kunduz, Arvind P Singhx, Li-Shiuan Pehy, and Niraj K Jhay. A 4.6 tbits/s 3.6 ghz single-cycle noc router with a novel switch allocator in 65nm cmos. In *2007 25th International Conference on Computer Design*, pages 63–70. IEEE, 2007.

[KKY11] Gwangsun Kim, John Kim, and Sungjoo Yoo. Flexibuffer: Reducing leakage power in on-chip network routers. In *Design Automation Conference (DAC), 2011 48th ACM/EDAC/IEEE*, pages 936–941. IEEE, 2011.

[KPKJ07] Amit Kumar, Li-Shiuan Peh, Partha Kundu, and Niraj K Jha. Express virtual channels: towards the ideal interconnection fabric. In *ACM SIGARCH Computer Architecture News*, pages 150–161. ACM, 2007.

[KTMW03] Jason Sungtae Kim, Michael Bedford Taylor, Jason Miller, and David Wentzlaff. Energy characterization of a tiled architecture processor with on-chip networks. In *Proceedings of the 2003 International Symposium on Low Power Electronics and Design, 2003. ISLPED'03.*, pages 424–427. IEEE, 2003.

[LCL+16] Shaoli Liu, Tianshi Chen, Ling Li, Xiaoxue Feng, Zhiwei Xu, Haibo Chen, Fred Chong, and Yunji Chen. Imr: High-performance low-cost multi-ring nocs. *IEEE Transactions on Parallel and Distributed Systems*, 27(6):1700–1712, 2016.

[LSMJ16] Zimo Li, Joshua San Miguel, and Natalie Enright Jerger. The runahead network-on-chip. In *2016 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pages 333–344. IEEE, 2016.

[LY18] Zhonghai Lu and Yuan Yao. Thread voting dvfs for manycore nocs. *IEEE Transactions on Computers*, 67(10):1506–1524, 2018.

[MJW12] Sheng Ma, Natalie Enright Jerger, and Zhiying Wang. Supporting efficient collective communication in nocs. In *IEEE International Symposium on High-Performance Comp Architecture*, pages 1–12. IEEE, 2012.

[MKAY09] Hiroki Matsutani, Michihiro Koibuchi, Hideharu Amano, and Tsutomu Yoshinaga. Prediction router: Yet another low latency on-chip router architecture. In *2009 IEEE 15th International Symposium on High Performance Computer Architecture*, pages 367–378. IEEE, 2009.

[MKI+10] Hiroki Matsutani, Michihiro Koibuchi, Daisuke Ikebuchi, Kimiyoshi Usami, Hiroshi Nakamura, and Hideharu Amano. Ultra fine-grained run-time power gating of on-chip routers for cmps. In *2010 Fourth ACM/IEEE International Symposium on Networks-on-Chip*, pages 61–68. IEEE, 2010.

[MKWA08] Hiroki Matsutani, Michihiro Koibuchi, Daihan Wang, and Hideharu Amano. Run-time power gating of on-chip routers using look-ahead routing. In *Proceedings of the 2008 Asia and South Pacific Design Automation Conference*, pages 55–60. IEEE Computer Society Press, 2008.

[MM09] Thomas Moscibroda and Onur Mutlu. A case for bufferless routing in on-chip networks. *ACM SIGARCH Computer Architecture News*, 37(3):196–207, 2009.

[MNTJ04] Mikael Millberg, Erland Nilsson, Rikard Thid, and Axel Jantsch. Guaranteed bandwidth using looped containers in temporally disjoint networks within the nostrum network on chip. In *Proceedings Design, Automation and Test in Europe Conference and Exhibition*, volume 2, pages 890–895. IEEE, 2004.

[MPK07] George Michelogiannakis, Dionisios Pnevmatikatos, and Manolis Katevenis. Approaching ideal noc latency with pre-configured routes. In *First International Symposium on Networks-on-Chip (NOCS'07)*, pages 153–162. IEEE, 2007.

[OM06] Umit Y Ogras and Radu Marculescu. Prediction-based flow control for network-on-chip traffic. In *Proceedings of the 43rd annual design automation conference*, pages 839–844. ACM, 2006.

[P+16] Anastasios Psarras et al. Phasenoc: Versatile network traffic isolation through tdm-scheduled virtual channels. *IEEE TCAD*, 2016.

[PD01] L-S Peh and William J Dally. A delay model and speculative architecture for pipelined routers. In *Proceedings HPCA Seventh International Symposium on High-Performance Computer Architecture*, pages 255–266. IEEE, 2001.

[PP84] Mark S Papamarcos and Janak H Patel. A low-overhead coherence solution for multiprocessors with private cache memories. *ACM SIGARCH Computer Architecture News*, 12(3):348–354, 1984.

[SCK+12] Chen Sun, Chia-Hsin Owen Chen, George Kurian, Lan Wei, Jason Miller, Anant Agarwal, Li-Shiuan Peh, and Vladimir Stojanovic. Dsent-a tool connecting emerging photonics with electronics for opto-electronic networks-on-chip modeling. In *2012 IEEE/ACM Sixth International Symposium on Networks-on-Chip*, pages 201–210. IEEE, 2012.

[Sil] Silvaco. Nangate 45nm library. `http://www.nangate.com/`.

[SJJ+11] Praveen Salihundam, Shailendra Jain, Tiju Jacob, Shasi Kumar, Vasantha Erraguntla, Yatin Hoskote, Sriram Vangal, Gregory Ruhl, and Nitin Borkar. A 2 tb/s 6×4 mesh network for a single-chip cloud computer with dvfs in 45 nm cmos. *IEEE Journal of Solid-State Circuits*, 46(4):757–766, 2011.

[SMG14] Radu Andrei Stefan, Anca Molnos, and Kees Goossens. daelite: A tdm noc supporting qos, multicast, and fast connection set-up. *IEEE Transactions on Computers*, 63(3):583–594, 2014.

[TKM+02] Michael Bedford Taylor, Jason Kim, Jason Miller, David Wentzlaff, Fae Ghodrat, Ben Greenwald, Henry Hoffman, Paul Johnson, Jae-Wook Lee, Walter Lee, et al. The raw microprocessor: A computational fabric for software circuits and general-purpose programs. *IEEE micro*, 22(2):25–35, 2002.

[Val82] Leslie G. Valiant. A scheme for fast parallel communication. *SIAM journal on computing*, 11(2):350–361, 1982.

[VB81] Leslie G Valiant and Gordon J Brebner. Universal schemes for parallel communication. In *Proceedings of the thirteenth annual ACM symposium on Theory of computing*, pages 263–277. ACM, 1981.

[vdBCGB07] Jan Willem van den Brand, Calin Ciordas, Kees Goossens, and Twan Basten. Congestion-controlled best-effort communication for networks-on-chip. In *Proceedings of the conference on Design, automation and test in Europe*, pages 948–953. EDA Consortium, 2007.

[WDLW17] Ji Wu, Dezun Dong, Xiangke Liao, and Li Wang. Energy-efficient noc with multi-granularity power optimization. *The journal of Supercomputing*, 73(4):1654–1671, 2017.

[WGO+13] Hassan MG Wassel, Ying Gao, Jason K Oberg, Ted Huffmire, Ryan Kastner, Frederic T Chong, and Timothy Sherwood. Surfnoc: a low latency and provably non-interfering approach to secure networks-on-chip. In *ACM SIGARCH Computer Architecture News*, pages 583–594. ACM, 2013.

[WNM+19a] Peng Wang, Sobhan Niknam, Sheng Ma, Zhiying Wang, and Todor Stefanov. A dynamic bypass approach to realize power efficient network-on-chip. In *Proceedings of the 21st IEEE International Conference on High Performance Computing and Communications (HPCC-2019)*, 2019.

[WNM+19b] Peng Wang, Sobhan Niknam, Sheng Ma, Zhiying Wang, and Todor Stefanov. Evc-based power gating approach to achieve low-power and high performance noc. In *Proceedings of Euromicro Conference on Digital System Design*, 2019.

[WNM+19c] Peng Wang, Sobhan Niknam, Sheng Ma, Zhiying Wang, and Todor Stefanov. Surf-bless: A confined-interference routing for energy-efficient communication in nocs. In *Proceedings of the 56th Annual Design Automation Conference 2019*, page 50. ACM, 2019.

[WNWS17] Peng Wang, Sobhan Niknam, Zhiying Wang, and Todor Stefanov. A novel approach to reduce packet latency increase caused by power gating in network-on-chip. In *2017 Eleventh IEEE/ACM International Symposium on Networks-on-Chip (NOCS)*, pages 1–8. IEEE, 2017.

[YL16] Yuan Yao and Zhonghai Lu. Dvfs for nocs in cmps: A thread voting approach. In *2016 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pages 309–320. IEEE, 2016.

[ZCPP15] Di Zhu, Lizhong Chen, Timothy M Pinkston, and Massoud Pedram. Tapp: Temperature-aware application mapping for noc-based many-core processors. In *Proceedings of the 2015 Design, Automation & Test in Europe Conference & Exhibition*, pages 1241–1244. EDA Consortium, 2015.

[ZL18] Hao Zheng and Ahmed Louri. Ez-pass: An energy & performance-efficient power-gating router architecture for scalable nocs. *IEEE Computer Architecture Letters*, 17(1):88–91, 2018.

[ZML+16] Xia Zhao, Sheng Ma, Chen Li, Lieven Eeckhout, and Zhiying Wang. A heterogeneous low-cost and low-latency ring-chain network for gpgpus. In *2016 IEEE 34th International Conference on Computer Design (ICCD)*, pages 472–479. Ieee, 2016.

[ZOG+15] Jia Zhan, Jin Ouyang, Fen Ge, Jishen Zhao, and Yuan Xie. Dimnoc: A dim silicon approach towards power-efficient on-chip network. In *2015 52nd ACM/EDAC/IEEE Design Automation Conference (DAC)*, pages 1–6. IEEE, 2015.

List of Publications

First Author Publications

1. **Peng Wang**, Sobhan Niknam, Sheng Ma, Zhiying Wang, and Todor Stefanov, "Surf-Bless: A Confined-interference Routing for Energy-Efficient Communication in NoCs", In Proc. "56th ACM/IEEE Int. Design Automation Conference (DAC'19)", pp. 50, Las Vegas, NV, USA, June 2-6, 2019, **(Winner of 2019 HiPEAC Paper Award)**.

2. **Peng Wang**, Sobhan Niknam, Sheng Ma, Zhiying Wang, and Todor Stefanov, "A Dynamic Bypass Approach to Realize Power Efficient Network-on-Chip", In Proc. "21st IEEE International Conference on High Performance Computing and Communications (HPCC-2019)", Zhangjiajie, Hunan, China, August 10-12, 2019.

3. **Peng Wang**, Sobhan Niknam, Sheng Ma, Zhiying Wang, and Todor Stefanov, "EVC-based Power Gating Approach to Achieve Low-power and High Performance NoC", In Proc. "Euromicro Conference on Digital System Design (DSD)", Chalkidiki, Greece, August 28-30, 2019.

4. **Peng Wang**, Sobhan Niknam, Zhiying Wang, and Todor Stefanov, "A Novel Approach to Reduce Packet Latency Increase caused by Power Gating in Network-on-Chip", In Proc. "11th ACM/IEEE International Symposium on Networks-on-Chip (NOCS'17)", pp. 1-8 , Seoul, South Korea, Oct. 19-20, 2017.

Co-author Publications

1. Sobhan Niknam, **Peng Wang**, and Todor Stefanov, "Resource Optimization for Real-Time Streaming Applications using Task Replication", IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), vol. 37, No. 11, pp. 2755-2767, Nov 2018.

2. Sobhan Niknam, **Peng Wang**, and Todor Stefanov, "Hard Real-Time Scheduling of Streaming Applications Modeled as Cyclic CSDF Graphs", In Proc. "22nd Int. Conf. Design, Automation and Test in Europe (DATE'19)", pp. 1528-1533, Florence, Italy, Mar. 25-29, 2019.

3. Di Liu, Jelena Spasic, **Peng Wang**, and Todor Stefanov, "Energy-Efficient Scheduling of Real-Time Tasks on Heterogeneous Multicores Using Task Splitting", In Proc. "22nd IEEE International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA'16)", pp. 149-158, Daegu, South Korea, Aug. 17-19, 2016.

4. Fuchs, Christian M, Murillo, Nadia M, Plaat, Aske, van der Kouwe, Erik, **Peng Wang**, "Towards affordable fault-tolerant nanosatellite computing with commodity hardware", In Proc. "27th IEEE Asian Test Symposium (ATS'2018)", pp. 127-132, Heifei, China, Oct. 15-18, 2018.