electronics

Article

Edge-Based Heuristics for Optimizing Shortcut-Augmented Topologies for HPC Interconnects

Kazi Ahmed Asif Fuad †, Kai Zeng † and Lizhong Chen *

Kelley Engineering Center, School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR 97331, USA

* Correspondence: chenliz@oregonstate.edu; Tel.: +1-541-737-3317

† These authors contributed equally to this work.

Abstract: Interconnection network topology is critical for the overall performance of HPC systems. While many regular and irregular topologies have been proposed in the past, recent work has shown the promise of shortcut-augmented topologies that offer multi-fold reduction in network diameter and hop count over conventional topologies. However, the large number of possible shortcuts creates an enormous design space for this new type of topology, and existing approaches are extremely slow and do not find shortcuts that are globally optimal. In this paper, we propose an efficient heuristic approach, called *EdgeCut*, which generates high-quality shortcut-augmented topologies. EdgeCut can identify more globally useful shortcuts by making its considerations from the perspective of edges instead of vertices. An additional implementation is proposed that approximates the costly all-pair shortest paths calculation, thereby further speeding up the scheme. Quantitative comparisons over prior work show that the proposed approach achieves a 1982× reduction in search time while generating better or equivalent topologies in 94.9% of the evaluated cases.

Keywords: high-performance computing system; interconnection network; topology; shortcut; design space exploration; heuristic search; shortest path; hop count

Citation: Fuad, K.A.A.; Zeng, K.; Chen, L. Edge-Based Heuristics for Optimizing Shortcut-Augmented Topologies for HPC Interconnects. *Electronics* **2022**, *11*, 2778. https://doi.org/10.3390/electronics11172778

Academic Editor: José L. Abellán

Received: 7 July 2022
Accepted: 31 August 2022
Published: 3 September 2022

1. Introduction

High-performance computing (HPC) systems are essential in order to run a variety of large-scale applications in multiple domains such as bio-informatics, astronomical analysis, nuclear simulations, financial services, etc., as well as large machine learning models applied to numerous use cases. As the backbone of HPC systems, interconnection networks are responsible for connecting up to hundreds or thousands of compute nodes (a compute node may, in turn, consist of hundreds to thousands of processing cores) [1] by providing fast and low-cost communication in the order of microseconds [2]. A primary factor in dictating the performance of interconnection networks is topology, which specifies the structure that is used to connect compute nodes. To achieve high interconnected performance, it is critical to design network topologies that have small diameters and low hop counts.

Prior works have proposed a number of regular and irregular topologies including rings, meshes, tori, hypercubes, fat trees, Clos, Butterfly, and Dragonfly. Many of them have been deployed in practical HPC systems. Interestingly, despite the seemingly mature development of topologies, recent work has demonstrated the large potential of a very different class of topologies, which we refer to as *shortcut-augmented topologies*. Such a topology starts with a base topology (e.g., a ring) and adds a series of shortcuts on top of that. Surprisingly, with careful selection, shortcut-augmented topologies are able to reduce both network diameter and average shortest hop count by multiple folds, compared with existing widely used regular and irregular topologies [3].

While this is promising, a major roadblock of further improving this class of topologies is the enormous design space that is formed by the combinations of possible shortcuts.

As an example, for a 64-node network, there are around 2×10^3 possible edges. If 128 edges are selected, there are over 4×10^{205} combinations! To make things worse, this design space grows super-exponentially as the network size increases, thus greatly exceeding the capabilities of exhaustive methods. Conventional methods of design space searching, such as simulated annealing (SA), would be extremely slow. This calls for novel, efficient heuristic approaches that exploit unique characteristics of the shortcut selection problem.

A straightforward heuristic is to add shortcuts randomly (subject to available free ports in the switches). With a large number of repetitions, a good shortcut-augmented topology may be found for a small network size. However, given the rapid increase in design space, this method quickly becomes insufficient for larger networks. Several papers have pointed out that adding random shortcuts can enhance the performance of their proposed network topologies [4–7], but these works do not directly propose approaches that can generate shortcuts more effectively. The state-of-the-art heuristic [3] is to consider the usefulness of shortcuts when adding shortcuts for a given vertex. Specifically, for each vertex v, the method examines a set of random nodes that are connected to v, and selects the top y (say 3) nodes that have the longest shortest paths from v. It then adds y shortcuts from v—one to each of the y nodes. We refer to this method as *vertex-based random shortcut* (VRS) approach. While this improves the quality of the generated topologies, our analysis reveals that VRS often adds shortcuts that are locally useful at a vertex but are not much use globally. This is because the y longest shortest paths at a vertex may not necessarily be considered as long paths in the entire network. Consequently, VRS requires more shortcuts to be added, which leads to additional costs of links and higher-degree switches.

To address this issue, this paper proposes *EdgeCut*, an effective heuristic approach for generating high-quality shortcut-augmented topologies. The main novelty is to identify more globally useful shortcuts by thinking from the perspective of edges, rather than from the perspective of vertices in prior work. We propose two variations of EdgeCut. EdgeCut-Full considers the performance impact on all node pairs after a shortcut is added. It produces the best topologies but needs to update all pairs' shortest paths after each addition, thus resulting in longer search time. EdgeCut-Lite approximates this function, leading to 11× reduction over EdgeCut-Full in search time while still generating satisfying topologies. Evaluation results show that, compared with simulated annealing, the proposed EdgeCut reduces the search time by 1982× and generating better or equivalent diameter in 94.9% of the test network topologies and sizes. Compared with VRS, EdgeCut reduces the network diameter by 55.1% while being slightly faster. These results highlight the effectiveness of the proposed approach.

The rest of the paper is organized as follows. Section 2 provides more background on HPC interconnection networks and shortcut-augmented topologies. Section 3 describes details of the proposed EdgeCut approach for identifying high-quality shortcuts. Section 4 presents evaluation methodology and results, and Section 5 includes further discussions. Finally, Section 6 concludes the paper.

2. Related Work

2.1. HPC Topologies

The topology of HPC interconnection networks is a very active research direction. While many topologies have been proposed in the past, new topologies are continually being proposed due to new challenges in latency, costs, scalability, reliability, etc., that are associated with ever-growing HPC systems. Only limited research has been conducted on shortcut-augmented topologies, leaving many opportunities for further improvement.

In *direct* topologies such as tori, meshes, and hypercubes, every switch is connected to a compute node, whereas in *indirect* topologies such as fat trees, Clos, and Butterfly, only the input and output switches at the edge of an network are associated with computer nodes, and packets sent from computer nodes are forwarded indirectly through middle-stage switches before reaching their destination [8]. Both direct and indirect topologies have been used in practice, e.g., 3D and 5D torus networks are used in Cray Gemini [9] and IBM

BlueGene/Q, respectively [10], and Dragonfly networks [11] with virtual routers are used in Cray Cascades [12]. In a sense, various direct and indirect topologies differ in how they trade-off among degree, diameter, and hop count [3].

Variations of regular topologies that result in irregular or ad hoc designs have also been proposed. For example, the Jellyfish topology [6] utilizes random graphs to develop high-capacity networks that allow incremental expansion on a daily basis to large-scale data centers, at a higher of cost of cabling. Slim Fly [4,5] approximates the optimal diameter, which results in lower latency, cost, and energy consumption while sustaining high bisection bandwidth. However, it is not suitable for gradual size expansion due to limited flexibility in the small design space. Distributed Loop Networks (DLN) add chordal edges or shortcuts to a simple ring topology to reduce diameter while maintaining low degree distribution. By adding shortcuts in a less regular manner than being evenly spaced, DLN can achieve more efficient designs, e.g., the diameter of a 36-vertex ring can be reduced from 18 to 9 by adding only five shortcuts [13,14].

Another related line of topology research stems from the famous *small-world phenomenon*, first proposed by [15], that demonstrates that people living in a country constitute societies of a network with only short path lengths. Later, Wattz and Strogatz characterized the small-world phenomenon into the Wattz–Strogatz (WS) model [16], which allows networks to be generated with short average distance and large clustering coefficient [17,18]. The model uses a few additional long edges to reduce the diameter in random graphs for social networks and Internet topologies [3,18]. Since then, researchers have been exploring the small-world phenomenon in computer networks [19–21]. In particular, to exploit the small-world effect in HPC, [3] proposes several methods that add random shortcuts to a base topology, the best of which is the vertex-based random shortcut (VRS) method that is mentioned in Section 1. Although shortcuts are selected optimally at each local vertex, they are not necessarily the most useful shortcuts to add globally. This deficiency is addressed by our proposed approach.

2.2. Design Space Exploration

The design space of shortcut-augmented topologies is enormous. This is not only because of the large number of possible shortcut candidates (especially for networks with high-radix switches), but also because of the huge number of combinations of the shortcut candidates. There are three typical ways of exploring design space. The first one is exhaustive search, which is impractical in this problem. The second one is general-purpose search algorithms, such as ant colony algorithm and simulated annealing [22]. These algorithms may be able to find the optimal or near-optimal solutions but usually require extremely long search time for large design space. The third one is heuristic approaches that leverage problem-specific characteristics to enable approximate but fast searches. In this paper, we aim to demonstrate that, for the problem of shortcut-augmented topologies, it is possible to design heuristic approaches that can find comparable solutions of general-purpose search algorithms but take only a tiny fraction of their time.

3. Proposed Approach

In this section, we describe the details of the proposed EdgeCut approach (implementation and instructions on running the proposed approach are available at https://github.com/OSU-STARLAB/EdgeCut (accessed on 1 September 2022)). We start by introducing some notations and definitions, and then present two versions of EdgeCut, namely EdgeCut-Full and EdgeCut-Lite, that offer different trade-offs between efficiency and effectiveness.

3.1. Definitions, Notations, and Assumptions

An interconnection network topology for N compute nodes can be abstracted as a graph with N vertices. The hop count between two nodes is the length of the path between the corresponding two vertices in the graph. Without additional information, a typical way

5.5. Additional Considerations

The evaluation in this paper focuses on average hop count, diameter, and number of edges, all of which are well-established metrics to assess topologies. For a given system, however, the choice of topologies also depends on several other considerations, such as traffic patterns, queuing effects, cost of links (e.g., electrical vs. optical), cost of switches, etc. Additional evaluation is needed to take these factors into account, such as simulating on a cycle-accurate interconnection network simulator.

6. Conclusions

Shortcut-augmented topologies have the potential to surpass conventional regular and irregular topologies but have been hindered by the challenge in searching their enormous design space. In this paper, we address this important issue by proposing an efficient and effective heuristic approach that aims to generate more globally useful shortcuts. The proposed EdgeCut-Full considers the performance impact of shortcut candidates more comprehensively but also incurs higher computation, whereas EdgeCut-Lite simplifies the search process while retaining the ability to find good shortcuts. Evaluation results show that the proposed approach is able to generate comparable high-quality topologies as simulated annealing and achieve faster search time than vertex-based method, essentially reaping the benefits of both worlds and offering a better trade-off.

Author Contributions: Conceptualization, K.A.A.F., K.Z. and L.C.; formal analysis, K.A.A.F. and L.C.; methodology, K.A.A.F. and K.Z.; software, K.A.A.F. and K.Z.; supervision, L.C.; validation, L.C.; writing—original draft, K.A.A.F., K.Z. and L.C.; writing—review & editing, L.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research (including APC) was funded, in part, by the National Science Foundation grant number 1750047.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Top 500 Supercomputer Sites. Available online: https://www.top500.org/lists/top500/2022/06/ (accessed on 30 June 2022).
2. Tomkins, J. Interconnects: A buyers point of view. In Proceedings of the ACS Workshop, Baltimore, MD, USA, 13–14 June 2007.
3. Koibuchi, M.; Matsutani, H.; Amano, H.; Hsu, D.F.; Casanova, H. A case for random shortcut topologies for HPC interconnects. In Proceedings of the 2012 39th Annual International Symposium on Computer Architecture (ISCA), IEEE, Portland, OR, USA, 9–13 Jun 2012; pp. 177–188.
4. Besta, M.; Hoefler, T. Slim Fly: A Cost Effective Low-Diameter Network Topology. In Proceedings of the SC '14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, New Orleans, LA, USA, 16–21 November 2014; pp. 348–359. [CrossRef]
5. Besta, M.; Hoefler, T. Slim Fly: A Cost Effective Low-Diameter Network Topology. *arXiv* **2019**, arXiv:1912.08968v2. [CrossRef].
6. Singla, A.; Hong, C.Y.; Popa, L.; Godfrey, P.B. Jellyfish: Networking data centers randomly. In Proceedings of the 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12), San Jose, CA, USA, 25–27 April 2012; pp. 225–238.
7. Fujiwara, I.; Koibuchi, M.; Matsutani, H.; Casanova, H. Skywalk: A Topology for HPC Networks with Low-Delay Switches. In Proceedings of the 2014 IEEE 28th International Parallel and Distributed Processing Symposium, Phoenix, AZ, USA, 19–23 May 2014; pp. 263–272. [CrossRef]
8. Baboli, M.; Husin, N.S.; Marsono, M.N. A comprehensive evaluation of direct and indirect network-on-chip topologies. In Proceedings of the 2014 International Conference on Industrial Engineering and Operations Management, Bali, Indonesia, 7–9 January 2014; pp. 2081–2090.
9. Alverson, R.; Roweth, D.; Kaplan, L. The Gemini System Interconnect. In Proceedings of the Proceedings of the 2010 18th IEEE Symposium on High Performance Interconnects, IEEE Computer Society, Mountain View, CA, USA, 8–20 August 2010; pp. 83–87. [CrossRef]

10. Chen, D.; Eisley, N.; Heidelberger, P.; Kumar, S.; Mamidala, A.; Petrini, F.; Senger, R.; Sugawara, Y.; Walkup, R.; Steinmacher-Burow, B.; et al. Looking under the Hood of the IBM Blue Gene/Q Network. In Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis, Salt Lake City, UT, USA, 10–16 November 2012; IEEE Computer Society Press: Washington, DC, USA, 2012.
11. Kim, J.; Dally, W.J.; Scott, S.; Abts, D. Technology-Driven, Highly-Scalable Dragonfly Topology. *SIGARCH Comput. Archit. News* **2008**, *36*, 77–88. [CrossRef]
12. Faanes, G.; Bataineh, A.; Roweth, D.; Court, T.; Froese, E.; Alverson, R.; Johnson, T.; Kopnick, J.; Higgins, M.; Reinhard, J. Cray cascade: A scalable HPC system based on a Dragonfly network. In Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis, Salt Lake City, UT, USA, 10–16 November 2012; IEEE/ACM: New York, NY, USA, 2012; p. 103.
13. Bollobás, B.; Chung, F.R.K. The Diameter of a Cycle Plus a Random Matching. *SIAM J. Discret. Math.* **1988**, *1*, 328–333. [CrossRef]
14. Bermond, J.C.; Comellas, F.; Hsu, D.F. Distributed loop computer-networks: A survey. *J. Parallel Distrib. Comput.* **1995**, *24*, 2–10. [CrossRef]
15. Milgram, S. The Small-World Problem. *Psychol. Today* **1967**, *1*, 61–67.
16. Watts, D.J.; Strogatz, S.H. Collective dynamics of 'small-world'networks. *Nature* **1998**, *393*, 440–442. [CrossRef] [PubMed]
17. Yasudo, R.; Koibuchi, M.; Nakano, K.; Matsutani, H.; Amano, H. Designing High-Performance Interconnection Networks with Host-Switch Graphs. *IEEE Trans. Parallel Distrib. Syst.* **2019**, *30*, 315–330. [CrossRef]
18. Yasudo, R.; Nakano, K.; Koibuchi, M.; Matsutani, H.; Amano, H. Designing low-diameter interconnection networks with multi-ported host-switch graphs. In *Concurrency and Computation: Practice and Experience*; Wiley: Hoboken, NJ, USA, 2000; p. e6115. [CrossRef]
19. Kleinberg, J. The small-world phenomenon and distributed search. *SIAM News* **2004**, *37*, 1–2.
20. Bonnet, F.; Kermarrec, A.M.; Raynal, M. Small-World Networks: From Theoretical Bounds to Practical Systems. In *Principles of Distributed Systems, Proceedings of the 11th International Conference, OPODIS 2007, Guadeloupe, French West Indies, France, 17–20 December 2007*; Tovar, E., Tsigas, P., Fouchal, H., Eds.; Springer: Berlin/Heidelberg, Germany, 2007; pp. 372–385.
21. Nguyen, V.; Martel, C. Designing Low Cost Networks with Short Routes and Low Congestion. In Proceedings of the IEEE INFOCOM 2006. 25TH IEEE International Conference on Computer Communications, Barcelona, Spain, 23–29 April 2006; pp. 1–12. [CrossRef]
22. Kirkpatrick, S.; Gelatt, C.D.; Vecchi, M.P. Optimization by simulated annealing. *Science* **1983**, *220*, 671–680. [CrossRef]
23. Parhami, B.; Yeh, C.H.; Parameters, T. Why Network Diameter is Still Important. In Proceedings of the International Conference in Communications (CIC-2000), Las Vegas, Nevada, USA, 26–29 June 2000.
24. Silla, F.; Duato, J. High-performance routing in networks of workstations with irregular topology. *IEEE Trans. Parallel Distrib. Syst.* **2000**, *11*, 699–719. [CrossRef]
25. Ausavarungnirun, R.; Fallin, C.; Yu, X.; Chang, K.K.W.; Nazario, G.; Das, R.; Loh, G.H.; Mutlu, O. Design and Evaluation of Hierarchical Rings with Deflection Routing. In Proceedings of the 2014 IEEE 26th International Symposium on Computer Architecture and High Performance Computing, Paris, France, 22–24 October 2014; pp. 230–237. [CrossRef]
26. Kim, J.; Dally, W.J.; Abts, D. Flattened Butterfly: A Cost-Efficient Topology for High-Radix Networks. In Proceedings of the 34th Annual International Symposium on Computer Architecture, San Diego, CA, USA, 9–13 June 2007; Association for Computing Machinery: New York, NY, USA, 2007; pp. 126–137. [CrossRef]