Network Partitioning and Avoidable Contention

Yishai Oltchik
yishai.oltchik@inf.ethz.ch
ETH Zurich

Oded Schwartz
odedsc@cs.huji.ac.il
Hebrew University of Jerusalem

ABSTRACT

Network contention frequently dominates the run time of parallel algorithms and limits scaling performance. Most previous studies mitigate or eliminate contention by utilizing one of several approaches: communication-minimizing algorithms; hotspot-avoiding routing schemes; topology-aware task mapping; or improving global network properties, such as bisection bandwidth, edge-expansion, partitioning, and network diameter. In practice, parallel jobs often use only a fraction of a host system. How do processor allocation policies affect contention within a partition?

We utilize edge-isoperimetric analysis of network graphs to determine whether a network partition defined by a processor allocation has optimal internal bisection. Increasing the bisection allows a more efficient use of the network resources, decreasing or completely eliminating the link contention. We study torus networks and characterize partition geometries that maximize internal bisection bandwidth, and examine the allocation policies of Mira and JUQUEEN, the two largest publicly-accessible Blue Gene/Q torus-based supercomputers. Our analysis shows that the bisection bandwidth of their partitions can often be improved by changing the partitions’ geometries, yielding up to a $\times 2$ speedup for contention-bound workloads. Benchmark experiments validate the predictions. Our analysis applies to allocation policies of other networks.

ACKNOWLEDGMENTS

We thank Ivo Kabadshow and Dorian Krause of Jülich Supercomputing Centre for their help in arranging the JUQUEEN experiments. We thank Adam Scovel of Argonne National Laboratory for his help and support in setting up custom partitions on Mira. Our experiments could not have been done without their help.

The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this project by providing computing time through the John von Neumann Institute for Computing (NIC) on the GCS Supercomputer JUQUEEN at Jülich Supercomputing Centre (JSC). This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.

Research is also supported by the Einstein Foundation and the Minerva Center for their help in arranging the JUQUEEN experiments. We thank Adam Scovel of Argonne National Laboratory for his help and support in setting up custom partitions on Mira. Our experiments could not have been done without their help.

The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this project by providing computing time through the John von Neumann Institute for Computing (NIC) on the GCS Supercomputer JUQUEEN at Jülich Supercomputing Centre (JSC). This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.

Research is also supported by the Einstein Foundation and the Minerva Foundation. This work was supported by the PetachTikva industry-academia consortium. This research was supported by a grant from the United States-Israel Bi-national Science Foundation (BSF), Jerusalem, Israel. This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 818252). This work was supported by The Federmann Cyber Security Center in conjunction with the Israel national cyber directorate.

1 INTRODUCTION

Network contention frequently dominates the run time of parallel algorithms and limits scaling performance [6]. Optimizing the internal bisection bandwidth of allocated partitions can decrease or completely eliminate the link contention of a parallel computation, improving overall performance for contention-bound workloads.

Our contribution. Our analysis utilizes a novel generalization of Bollobás and Leader’s bounds on the edge-isoperimetric problem on torus graphs [11]. A solution was known for tori with dimensions of equal size, whereas our new bound applies to torus graphs with arbitrary dimension sizes. This is useful, as the vast majority of torus networks with 3 dimensions or more have unequal dimension sizes. We apply isoperimetric analysis to compute node partition allocations allowed by the allocation of Mira and JUQUEEN, the two largest publicly-accessible Blue Gene/Q torus-based supercomputers. Our analysis demonstrates that the bisection bandwidth of their current partitions can often be improved by changing the partitions’ geometries, yielding up to a $\times 2$ speedup for contention-bound workloads. Benchmarking experiments on both systems validate the predictions.

CCS CONCEPTS

• Computer systems organization → Distributed architectures; • Networks → Network performance evaluation; • Mathematics of computing → Extremal graph theory;

ACM Reference Format:
Yishai Oltchik and Oded Schwartz. 2020. Brief Announcement: Network Partitioning and Avoidable Contention. In Proceedings of the 32nd ACM Symposium on Parallelism in Algorithms and Architectures (SPAA ’20), July 15–17, 2020, Virtual Event, USA. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3350755.3400242

2 PRELIMINARIES

Torus graphs. Let $D$ and $a_1, \ldots, a_D$ be integers, and let $G = (V, E)$ be a graph. If $V = \{a_1\} \times \ldots \times \{a_D\}$, and every two vertices $u = (a_1, \ldots, a_D), v = (b_1, \ldots, b_D)$ are adjacent if and only if $\exists k$ such that $a_k = b_k \pm 1 \mod a_k$ and $\forall j \neq k, a_j = b_j$, then $G$ is said to be a $D$-torus (also, $D$-dimensional torus). If $a_1 = \ldots = a_D$ then $G$ is
3.1 The Edge-Isoperimetric Problem

We obtain a novel generalization of Theorem 1 to arbitrary torus graphs. We show that the bound is optimal for cuboid subcubes, and conjecture that it is optimal for arbitrary subcubes as well.

Theorem 2 (Edge-isoperimetric ineq. for tori). Let \( G = (V, E) \) be a \( D \)-dimensional torus with \( V = [a_1] \times [a_2] \times \ldots \times [a_D] \), and let \( t \leq \frac{\sqrt{V}}{\sqrt{D}} \). Suppose, without loss of generality, that \( a_1 \geq a_2 \geq \ldots \geq a_D \). Then, for any cuboid \( S \subset V \), \( |S| = t \):

\[
|E(S, \bar{S})| \geq \min_{r \in \{0, \ldots, D-1\}} 2(D-r) \cdot n^{D-r} \cdot \frac{n^{D-r}}{D-n^{D-r}}
\]

As is the case with Theorem 1, our bound can be attained in some cases. Let \( k = \prod_{i=0}^{D-1} a_{D-i} \). If \( \exists r \) such that \( (\frac{k}{n})^{\frac{1}{D-r}} \) is an integer, define the cuboid \( S_r = \left( (\frac{k}{n})^{\frac{1}{D-r}} \right)^{D-r} \times [a_{D-r+1}] \times \ldots \times [a_D] \).

Claim 3. Let \( G = (V, E) \) be a \( D \)-dimensional torus as in Theorem 2, and let \( t, r' \) be integers such that \( t \leq \frac{\sqrt{V}}{\sqrt{D}} \) and \( (\frac{k}{n})^{\frac{1}{D-r'}} \) is an integer. Define \( S_{r'} \) as in Theorem 2, with arg min \( r' \). Then:

\[
|E(S_{r'}, \bar{S}_{r'})| = 2(D-r') \cdot \left( \prod_{i=0}^{r'-1} a_{D-i} \right)^{\frac{1}{D-r'}} \cdot t^{\frac{D-r'}{D-n^{D-r'}}}
\]

Claim 4. Let \( G, t, D, r, S_r \) be defined as in Claim 3. Let \( A \subset V \) be some cuboid \([A_1] \times \ldots \times [A_D]\) with \( |A| = t \). Suppose there are exactly \( r \) indices \( i_1, \ldots, i_r \) such that \( A_{i_k} = a_{i_k} \). Then, \( |E(S_r, \bar{S_r})| \leq |E(A, \bar{A})| \).

The proof of Theorem 2 follows directly from combining Claims 3 and Claim 4. See full version of this paper [29] for complete proofs of those Claims. The small-set expansion of \( G \) for vertex subsets of size \( t \) is defined \( h_t = \frac{|E(A, \bar{A})|}{|E(A)| + |E(\bar{A})|} \), and it can be used to test whether a given network will be inevitably asymptotically contention-bound when executing a parallel algorithm with known per-processor communication costs [7]. Isoperimetric sets in torus graphs also attain the small-set expansion as a result of Equation 1, our analysis here is sufficient also for deriving those lower bounds.

3.2 Analysis of Blue Gene/Q Systems

**JUQUEEN.** Formerly at Jülich Supercomputing Centre, JUQUEEN was ranked 5th in the November 2012 Top 500 supercomputers [16]. It had 28,672 compute nodes, with network size \( 7 \times 2 \times 2 \times 2 \). JUQUEEN’S scheduler permits all partitions that are cuboids of midplanes, and geometries may be user-specified. Partitions optimal and sub-optimal in bisection bandwidth are both permitted.

**Mira.** Formerly at Argonne National Laboratory, Mira [5] was ranked 3rd in the June 2012 Top 500 supercomputers [16]. It had 49,152 compute nodes, with network size \( 4 \times 4 \times 3 \times 2 \). Mira’s scheduler is based on a pre-defined list of partitions; in Table 1 we list the partitions whose bisection bandwidth is sub-optimal.

4 EXPERIMENTS

Executions on proposed partitions on Mira were done via special support allowing the use of a temporarily modified processor allocation policy. Only the job placement rules were changed.
We compared the average wallclock time when using currently-used algorithms. Each round, every pair of nodes simultaneously exchanges random messages sized 0.1342 Gigabyte.

We compared the average wallclock time when using currently-used and proposed partitions (best-case and worst-case on JUQUEEN). Additional experiments appear in [29].

REFERENCES
[1] Rudolf Ahlswede and Sergei J. Bezrukov. Edge isoperimetric theorems for integer point arrays. *Applied Mathematics Letters*, 8(2):75–80, 1995.
[2] Jung Ho Ahn, Nathan Binkert, Al Davis, Motay McGracen, and Robert Schreiber. HyperX: topology, routing, and packaging of efficient large-scale networks. In *Proceedings of the Conference on High Performance Computing, Networking, Storage and Analysis*, pages 41. ACM, 2009.
[3] Yuuichirou Ajima, Tomohiro Inoue, Shinya Hiramoto, and Toshiyuki Shimizu. Tofu: Interconnected for the K computer. *Fujitsu Sci. Tech. J*, 23:1–155, 2014.
[4] Bob Alverson, Tim Johnson, Joe Kopnick, Mike Higgins, and James Reinhard. Cray Cascade: a Scalable HPC System based on a Dragonfly Network. In *Proceedings of the 23rd International Conference on High Performance Computing, Networking, Storage and Analysis*, pages 103. IEEE Computer Society Press, 2012.
[5] Argonne National Laboratory website. Official machine overview of Mira. https://www.alcf.anl.gov/user-guides/machine-overview-bbg.
[6] Grey Ballard, E Carson, J Demmel, M Hoemmen, Nicholas Knight, and Oded Schwartz. Communication lower bounds and optimal algorithms for numerical linear algebra. *Acta Numerica*, 23:1–155, 2014.
[7] Grey Ballard, James Demmel, Andrew Gearhart, Benjamin Lipshitz, Yishai Olichch, Oded Schwartz, and Sivan Toledo. Network topologies and inevitable contention. In *Communication Optimizations in HPC (COMHPC)*, *International Workshop on*, pages 39–52. IEEE, 2016.
[8] Grey Ballard, James Demmel, Olga Holtz, Benjamin Lipshitz, and Oded Schwartz. Communication-optimal parallel algorithm for Strassen’s matrix multiplication. In *Proceedings of the twenty-fourth annual ACM symposium on Parallelism in algorithms and architectures*, pages 193–204. ACM, 2012.
[9] Maciej Besta and Torsten Hoefler. Load-balanced routing in interconnection networks. PhD thesis, Stanford University, 2005.
[10] Abdinav Bhatele, Nikhil Jain, Katherine E. Isaaq, Ronak Buch, Todd Gamblin, Steven H Langer, and Laxmikant V Kale. Optimizing the performance of parallel applications on a 5d torus via task mapping. In *High Performance Computing (HiPC), 2014 21st International Conference on*, pages 1–10. IEEE, 2014.
[11] Bela Bollobas and Imre Leader. Edge-isoperimetric inequalities in the grid. *Combinatorica*, 11(4):299–314, 1991.
[12] Dong Chen, Noel Easley, Philip Heidelberger, Sameer Kumar, Anitha Mamidala, Fabrizio Petreni, Robert Senger, Yutaka Sugawara, Robert Walkup, Burkhard Steinmacher-Burrow, Anamitra Choudhury, Yoshig Sabharwal, Swati Singhal, and Jeffrey J. Parker. 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*, page 69. IEEE Computer Society Press, 2012.
[13] Dong Chen, Noel Easley, Philip Heidelberger, Robert Senger, Yutaka Sugawara, Sameer Kumar, Valentina Salapura, David Satterfield, Burkhard Steinmacher-Burrow, and Jeffrey Parker. The IBM Blue Gene/Q interconnection fabric. *IEEE Micro*, 32(1):52–63, 2012.
[14] James Demmel, Laura Grigori, Mark Hoemmen, and Julien Langou. Communication-optimal parallel and sequential QR and LU factorizations. *SIAM Journal on Scientific Computing*, 34(1):A246–A259, 2012.
[15] Jack Dongarra, Piotr Luszczek, and Antoine Petitet. The LINPACK benchmark: past, present and future. *Concurrency and Computation: practice and experience*, 15(9):803–820, 2003.
[16] Jack Dongarra, Hans W Meuer, and Erich Strohmaier. Top500 Supercomputer Sites, November 2017. 2017. www.top500.org.
[17] Michael R Garey, David S Johnson, and Larry Stockmeyer. Some simplified NP-complete problems. In *Proceedings of the sixth annual ACM symposium on Theory of computing*, pages 47–63. ACM, 1974.
[18] Lawrence Hueso Harper. Optimal assignments of numbers to vertices. *Journal of the Society for Industrial and Applied Mathematics*, 12(1):131–155, 1964.
[19] Nikhil Jain, Abhinav Bhatele, Xiang Ni, Todd Gamblin, and Laxmikant V Kale. Partitioning low-diameter networks to eliminate inter-job interference. In *Parallel and Distributed Processing Symposium (IPDPS), 2017 IEEE International*, pages 439–448. IEEE, 2017.
[20] Sangeetha Abdu Jyothi, Ankit Singla, P Brighten Godfrey, and Alexandra Kolla. Measuring and understanding throughput of network topologies. In *High Performance Computing, Networking, Storage and Analysis*, SC’16: *International Conference for*, pages 761–772. IEEE, 2016.
[21] John Kim, Wilham J Dally, Steve Scott, and Dennis Abts. Technology-driven, highly-scalable dragonfly topology. In *Computer Architecture*, 2008. DCA’08. 35th International Symposium on, pages 77–88. IEEE, 2008.
[22] Lawrence Livermore National Laboratory website. Sequoia supercomputer transitions to classified work. *https://www.llnl.gov/news/sequoia-supercomputer-transitions-classified-work*. Lawrence Livermore National Laboratory website. Sequoia supercomputer transitions to classified work. *https://www.llnl.gov/news/sequoia-supercomputer-transitions-classified-work*.
[23] Lawrence Livermore National Laboratory website. Using the Sequoia and Vulcan BG/Q Systems. *https://computing.llnl.gov/tutorials/bgq/.*
[24] James R Lee, Shayon Oveis Gharan, and Luca Trevisan. Multivariate spectral partitioning and higher-order cheeger inequalities. *Journal of the ACM (JACM)*, 61(6):37, 2014.
[25] John H Lindsey. Assignment of numbers to vertices. *The American Mathematical Monthly*, 71(5):508–516, 1964.
[26] Benjamin Lipshitz, Grey Ballard, James Demmel, and Oded Schwartz. Communication-avoiding parallel strassen: Implementation and performance. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, page 101. IEEE Computer Society Press, 2012.
[27] NASA Website. Official machine overview of the Pleiades supercomputer. *https://www.nas.nasa.gov/hecc/resources/pleiades.html*. Oak Ridge National Laboratory website. Official machine overview of Titan. *https://www.olcf.ornl.gov/olcf-resources/compute-systems/titan/*.
[28] Yishai Olichch and Oded Schwartz. Network partitioning and avoidable contention. *arXiv*, 2020.
[29] Arjun Singh. *Load-balanced routing in interconnection networks*. PhD thesis, Stanford University, 2005.
[30] Edgar Solomonik and James Demmel. Communication-optimal parallel 2.5D matrix multiplication and LU factorization algorithms. In *European Conference on Parallel Processing*, pages 90–109. Springer, 2011.
[31] Asad Valadarsky, Michael Dittrich, and Michael Schapira. Xpander: unveiling the secrets of high-performance datacenters. In *Proceedings of the 14th ACM Workshop on Hot Topics in Networks*, page 16. ACM, 2015.