NTNU
Norwegian University of
Science and Technology

TDT4900 Computer Science, Master's Thesis

Optimization of Seed Selection for Information Diffusion with High Level Synthesis

Julian Lam

Fall 2016

Department of Computer and Information Science
Norwegian University of Science and Technology

Supervisor 1: Donn Alexander Morrison
Supervisor 2: Yaman Umuroglu

0.1 Assignment

Information diffusion is a field of network research where a message, starting at a set of seed nodes, is propagated through the edges in a graph according to a simple model. Simulations are used to measure the coverage and speed of the diffusion and are useful in modelling a variety of phenomena such as the spread of disease, memes on the Internet, viral marketing and emergency messages in disaster scenarios.

The effectiveness of a given spreading model is dependent on the initially infected nodes, or seeds. Seed selection for an optimal spread is an NP hard problem and is normally approximated by selecting high-degree nodes or using heuristic methods such as discount-degree or choosing nodes at different levels of the k-core.

High-level synthesis (HLS) is becoming an important tool in the optimization/acceleration of algorithms in hardware. Starting with an algorithm written in a high-level language such as C or C++, HLS aids with hardware design by providing a methodology and tools that guide the developer through the design process.

This project should employ HLS as a design methodology for hardware accelerated seed selection in large graphs. The student will study seed selection for a given diffusion model, write a high-level model, and use HLS to implement a hardware design that exploits parallelism in the seed selection algorithm in order to improve performance over a GPCPU implementation. –

Abstract

Information Diffusion are often used for different simulations in network research because it simulates how information propagetes thorough a network, from memes on the Internett, spreading of disease in populations, to viral marketing. Measuring spread and speed, we can find influential targets in the network, such targets are optimal targets to pass message during disaster scenario, vaccinate to prevent spreading of a disease, or even targets for viral marketing.

High Level Synthesis have in recent years matured greatly. With HLS, designing custom architectures is no longer a

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Information diffusion is a field of network research where a message, or data, is propagated through a *network* or a *graph*. The message originates from a chosen set of nodes, known as *seed nodes*. These seed nodes passes the message to it's neighbour through the edges and thus propagate the message over theentire network. There are different models used in Information diffusion, *Independent Cascade Model*, and *Linear Threshold Model*. Information Diffusion can be used to model different phenomena such as the spread of disease, viral marketing, or even spread of viral videos and *"memes"*[2] [3]. The effectiveness of the simulation is measured in the spread and the speed of propagation. The effectiveness of the simulation is dependent on the chosen seed nodes. By finding the most optimal set of seed node, we can potentially stop an epidemic by vaccinating influential nodes, we can find important target for viral marketing by giving free sample, and use this information to quickly spread message during disaster scenarios.

There are multiple studies done regarding information diffusion, [4], [2], [5], [6]. There are few that focus on optimizing the seed selection, especially in hardware. Finding the most optimal set of seed node is useful in multiple fields. We prevent the spread of a disease by vaccinating influential nodes in the network, we can pass critical message through a population in disastrous scenario, or even find optimal target for viral marketing. The current seed selection algorithm is an greedy solution[7][DOUBLE CHECK THIS SOURCE], where every set of node is tested and the set with best coverage and time is chosen. This is a time consuming process and highly parallelizable. This makes it a good candidate for *Field-programmable gate arrays*(FPGAs).

High Level Syntesis(HLS) synthesis high level behaviour and constrains to lower level design.[8]. It allows users to implement an algorithm in high level language, C or C++, and generate an optimal design in *verilog* or *VHDL*. Verilog and VHDL are hardware descriptive language designed to describe digital

systems [9]. In recent years, High Level Synthesis have gotten more attention and more support, the xilinx forums are anwsered quickly by the developers and highly populated with seasoned hardware designers and novices.

Unlike traditional hardware design, HLS allows programmer with limited knowledge to desgn an optimal custom *Intellectual property core*(IP-core). In HLS, programmers can test out multiple different optimization schemes in short period of time. Thus allowing the programmer to quickly test out different optimization schemes.

For our implementation, we focused mainly on the ICM. The ICM is a special case of the common graph traversal algorithm *Breadth First Search*(bfs). For our implementation, we chose to implement the ICM as a custom *sparse matrix vector multiplication*(spmv). By performing ICM as spmv, we can utilize the parallelism options that spmv uncovers.

1.2 Assignment Interpretation

From the assignment text, these task were chosen as the main focus of this thesis:

Task 1 ***(mandatory)*** Implement Information Diffusion as Sparse matrix vector multiplication, with high level language C.

Task 2 ***(mandatory)*** Tailor the implementation of Information Diffusion for synthesise with Vivado HLS.

Task 3 ***(optional)*** Implement said design on a Zynq FPGA board.

Task 4 ***(optional)*** Extend the system to be able to handle graph in the size of toy graphs(containing 2^{26}) nodes)

1.3 Report Structure

We have here the basic outline for this report and a short overview of the remainder of this report:

Chapter 2: Background contains the information regarding network, Information diffusion, matrix vector multiplication and High level synthesis. Most of the background information regarding this report can be found in this chapter.

Chapter 3: Related Work shows what the related works and state of the art regarding information diffusion.

Chapter 4: Architecture

Chapter 5:.

Chapter 6: Future Work

Chapter 7: Conclusion Find something

Chapter 2

Related Work

Here, we will give you a short overview of the current state regarding Information Diffusion, High Level Synthesis and different optimization options.

- Yamans paper, where there are some works that shows the solution i use
- parallalization of the algorithm
- maybe some examples of HLS to show that HLS is used.
- showe that there are not many HLS implementation, recently matured.
- show that there are not many hardware implementation for information diffusion.
- need to look through Yamans paper and get some refrences from there.
- might be good to look at how this type of sparce matrix multiplication can be used
- show other implementation of SPmv
- Show some examples where image processing is done through vivado HLS.
- [21] A good paper showing the state of the art for HLS.

This chapter, we will look at the state of research regarding High Level Synthesis, network research regarding Information Diffusion, and Optimization of Independent cascade model and Breadth first search.

2.1 Information Diffusion

There are multiple studies done regarding Information Diffusion. One studies shows how information diffusion can be applied during an disease outbreak[2], viral marketing[20], coordinat during crisis situation[22].

Models of influence have been done on blogs[23][24], and twitter[25]. We can see that in an age of social media, the studies of information diffusion is more relevant then ever.

while other[6] have argued that the emerging of social network and media, have changed the traditional model. The activation is no longer only relying on neighbour nodes, but also an external influence. They found that large amount of information volume in Twitter is the result of network diffusion, while a small amount is due to external events and factors outside the network[6]. Another studies shows during the 2011 Egyptian Uprising, how larg amount of such a movement were "tweeted"[22].

As we mentioned in **??**, we mainly focus on 2 common information diffusion models, ICM and LTM. But there are different models too. One report[5] proposed several different problems with traditional models where each node is either *activated*(infected, influenced, '1') or *inactive*(healthy, not reached, '0'), and passes the *contagion*(information, data, infection, influence) to a neighbouring nodes through the edges. The report mentioned different assumptions that such models take. Among them is that a complete graph is provided, the spread of contagion is from a known source, and that the structure in the network is sufficient to explain the the behaviour[5]. The report propose an alternative model, *Linear Influence Model*(LIM), where the focus is on the global influence a infected node has on the rate of diffusion through the implicit network. This model takes the assumptions, that newly activated nodes is dependent on previous activated nodes. The LIM does not need explicit knowledge of the entire network, instead the model takes the newly activated nodes and model them as a *influence function*, which is used to find the global influence.

2.2 High Level Synthesis

High Level Synthesis as a concept have been around since the mid-1980s and early-1990s. Early tools, known as Carnegie-Mellon University design automation (CMU-DA)[26][27] was a pioneering early version of HLS tool. The tool gathered quickly considerable interest. A number of HLS tools were built in later year mostly for prototyping and research[28][29][30]. Some of these early tools was able to produce real chips, but the reason for lack of further development and adaptation, was that RTL synthesis was not a widely accepted and immature field. This often lead to suboptimal solutions.

In the 2000, new HLS tools was developed in academia and in the industry. These tools, used hihg level language, C and C++. Vivado HLS, designed by Xilinx [31], is one such HLS tool. The Vivado HLS became free during their 2015.4 update[32]. This resulted in an revived interest in HLS. The community around HLS is also evolving, on the Xilinx-forum, there are multiple anwsers and active members. We can see that the solution designed by HLS tools is close to traditional hand-crafted designs[33].

Different solutions that have

2.3 Different optimization scheme

Chapter 3

Result

as we can see, the algorithm was able to finish a

Chapter 4

Conclusion

Bibliography

[1] Deepayan Chakrabarti, Yiping Zhan, and Christos Faloutsos. R-mat: A recursive model for graph mining. 4:442–446, 2004.

[2] Daniel Gruhl, R. Guha, David Liben-Nowell, and Andrew Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pages 491–501, New York, NY, USA, 2004. ACM.

[3] Daniel M. Romero, Brendan Meeder, and Jon Kleinberg. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 695–704, New York, NY, USA, 2011. ACM.

[4] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and P Krishna Gummadi. Measuring user influence in twitter: The million follower fallacy. *ICWSM*, 10(10-17):30, 2010.

[5] J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. In *2010 IEEE International Conference on Data Mining*, pages 599–608, Dec 2010.

[6] Seth A. Myers, Chenguang Zhu, and Jure Leskovec. Information diffusion and external influence in networks. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 33–41, New York, NY, USA, 2012. ACM.

[7] David Kempe, Jon Kleinberg, and va Tardos. Influential nodes in a diffusion model for social networks. 3580:1127–1138, 2005.

[8] M. C. McFarland, A. C. Parker, and R. Camposano. The high-level synthesis of digital systems. *Proceedings of the IEEE*, 78(2):301–318, Feb 1990.

[9] Donald Thomas and Philip Moorby. *The Verilog® Hardware Description Language*. Springer Science & Business Media, 2008.

[10] Jeremy Kepner and John Gilbert. *Graph algorithms in the language of linear algebra*, volume 22. SIAM, 2011.

[11] Y. Umuroglu, D. Morrison, and M. Jahre. Hybrid breadth-first search on a single-chip fpga-cpu heterogeneous platform. pages 1–8, Sept 2015.

[12] M. E. J. Newman. *The Structure and Function of Complex Networks*, volume 45. 2003.

[13] Stanley Milgram Jeffrey Travers. An experimental study of the small world problem. *Sociometry*, 32(4):425–443, 1969.

[14] Stanley Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.

[15] Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. *Sociometry*, pages 425–443, 1969.

[16] Mark EJ Newman. Assortative mixing in networks. *Physical review letters*, 89(20):208701, 2002.

[17] James Moody. Race, school integration, and friendship segregation in america1. *American journal of Sociology*, 107(3):679–716, 2001.

[18] Éva Tardos David Kampe, Jon Klein. Maximizing the spread of influence through a social network. pages 137–146, 2003.

[19] MH McAndrew. On the product of directed graphs. *Proceedings of the American Mathematical Society*, 14(4):600–606, 1963.

[20] Pedro Domingos and Matt Richardson. Mining the network value of customers. pages 57–66, 2001.

[21] J. Cong, B. Liu, S. Neuendorffer, J. Noguera, K. Vissers, and Z. Zhang. High-level synthesis for fpgas: From prototyping to deployment. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 30(4):473–491, April 2011.

[22] Kate Starbird and Leysia Palen. (how) will the revolution be retweeted?: Information diffusion and the 2011 egyptian uprising. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 7–16, New York, NY, USA, 2012. ACM.

[23] Eytan Adar and Lada A. Adamic. Tracking information epidemics in blogspace. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '05, pages 207–214, Washington, DC, USA, 2005. IEEE Computer Society.

[24] Manuel Gomez Rodriguez, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 1019–1028, New York, NY, USA, 2010. ACM.

[25] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone's an influencer: Quantifying influence on twitter. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 65–74, New York, NY, USA, 2011. ACM.

[26] S. Director, A. Parker, D. Siewiorek, and D. Thomas. A design methodology and computer aids for digital vlsi systems. *IEEE Transactions on Circuits and Systems*, 28(7):634–645, Jul 1981.

[27] A. Parker, D. Thomas, D. Siewiorek, M. Barbacci, L. Hafer, G. Leive, and J. Kim. The cmu design automation system: An example of automated data path design. In *Proceedings of the 16th Design Automation Conference*, DAC '79, pages 73–80, Piscataway, NJ, USA, 1979. IEEE Press.

[28] John Granacki, David Knapp, and Alice Parker. The adam advanced design automation system: Overview, planner and natural language interface. In *Proceedings of the 22Nd ACM/IEEE Design Automation Conference*, DAC '85, pages 727–730, Piscataway, NJ, USA, 1985. IEEE Press.

[29] P. G. Paulin, J. P. Knight, and E. F. Girczyc. Hal: A multi-paradigm approach to automatic data path synthesis. In *Proceedings of the 23rd ACM/IEEE Design Automation Conference*, DAC '86, pages 263–270, Piscataway, NJ, USA, 1986. IEEE Press.

[30] H. D. Man, J. Rabaey, P. Six, and L. Claesen. Cathedral-ii: A silicon compiler for digital signal processing. *IEEE Design Test of Computers*, 3(6):13–25, Dec 1986.

[31] D. Navarro, Luca, L. A. Barragn, I. Urriza, and Jimnez. High-level synthesis for accelerating the fpga implementation of computationally demanding control algorithms for power converters. *IEEE Transactions on Industrial Informatics*, 9(3):1371–1379, Aug 2013.

[32] MS Windows NT kernel description. @miscWinNT, title = MS Windows NT Kernel Description, howpublished = http://web.archive.org/web/20080207010024/http://www.808multimedia.com/winnt/kernel.htm, note = Accessed: 2010-09-30 . Accessed: 2016-06-10.

[33] F. Winterstein, S. Bayliss, and G. A. Constantinides. High-level synthesis of dynamic data structures: A case study using vivado hls. In *Field-Programmable Technology (FPT), 2013 International Conference on*, pages 362–365, Dec 2013.