Hardware Accelerated Linear Programming

Parallelizing the Simplex Method with OpenCL

Bradley de Vlugt, Maysam Mirahmadi, Serguei L. Primak, Abdallah Shami

Western University · Canada

SOSCIP Southern Ontario Smart Computing Innovation Platform

Introduction

- Linear Programming (LP) models are frequently encountered optimization problems in medicine, engineering and business , where a linear objective function needs to be either minimized or maximized subject to multiple linear constraints.
- High performance implementations of the Simplex Algorithm facilitate solutions to large problems in these disciplines.
- This work proposes an energy-efficient hardware accelerated LP solver for a class of problems in radiation therapy. The proposed accelerator is implemented and tested on various devices using OpenCL.

Linear Programming

- The standard form of a linear programming problem is (1), where A represents the constraint system, b represents the constraint bounds, c is the cost function, and x is the decision variables.

$$\begin{aligned} &\min\ \mathbf{cx} \\ \text{s.t.}\ &\mathbf{Ax} \le \mathbf{b} \qquad (1) \\ &\mathbf{x} \ge 0 \end{aligned}$$

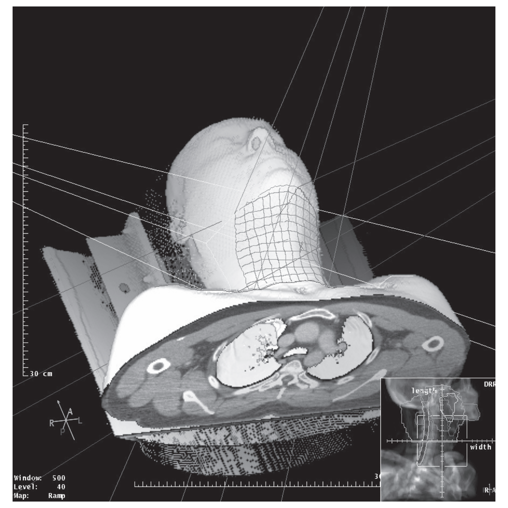
Application: Radiation Therapy Treatment Planning

- Radiation therapy uses several beams to deliver a radiation dose to tumour cells. Recent literature has formulated the problem of optimizing the beam weights as an LP problem [3-4].
- The affected area is divided into target t and normal n voxels. The radiation from each beam to each voxel is measured to form the dose matrix D.
- The model (2) minimizes the total dose delivered to non-target areas subject to constraints based on the properties of the tissue in each voxel.

$$\begin{aligned} \min\ &(\mathbf{D}_n^T \mathbf{c}_N)^T \mathbf{w} \\ &\mathbf{D}_t \mathbf{w} \le \mathbf{x}_t^U \\ &\mathbf{D}_t \mathbf{w} \ge \mathbf{x}_t^L \qquad (2) \\ &\mathbf{D}_n \mathbf{w} \le \mathbf{x}_n^U \\ &\mathbf{w} \ge 0 \end{aligned}$$

- Vectors x represent the upper U and lower L bounds for the target and normal areas. Beam weights w are given costs c_N.
- Since each beam will deliver radiation to most voxels, the dose matrix of the LP problem will be dense.

Figure 1: Radiation Therapy Beam Configuration [4]



OpenCL

- Language designed by Khronos Group for portability across hardware accelerators.
- A Hardware accelerator, or OpenCL Device, is managed by a Host processor.
- The OpenCL linear programming engine was tested on multiple devices to measure acceleration (see Table 1 for specifications).
- Effort was made to optimize OpenCL kernels for each device without losing design portability.

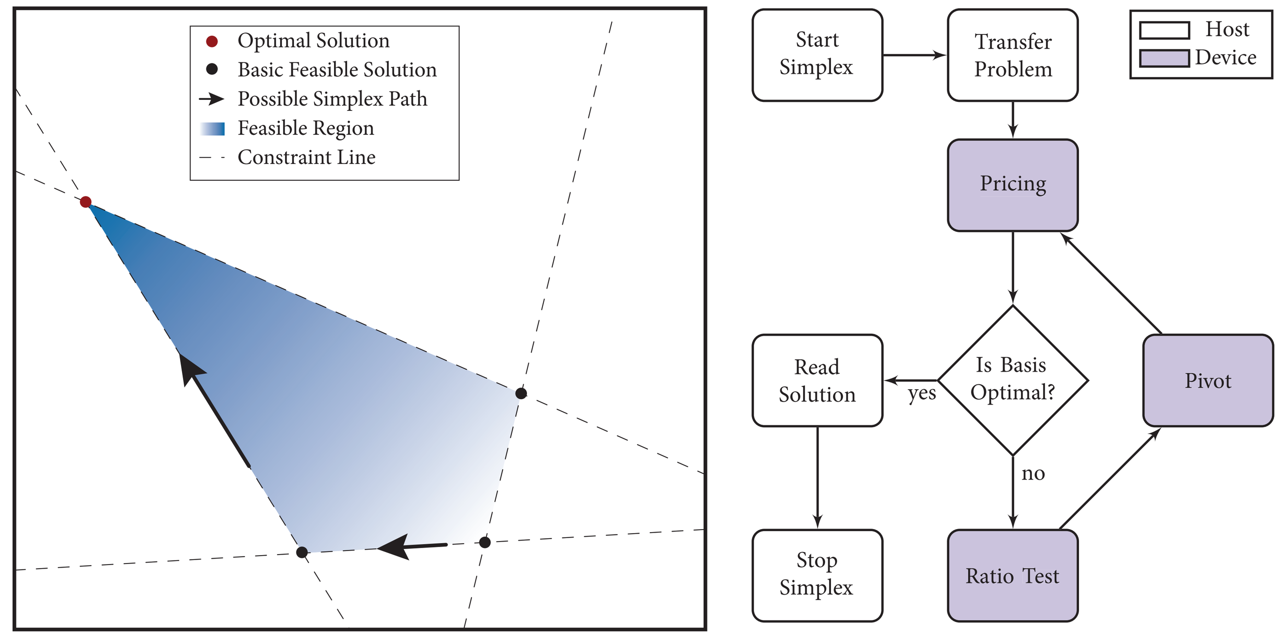
Table 1: OpenCL Test Device Specifications

Device	Power (W)	Memory Bandwidth (GB/s)
Intel Core i7 4930k (6 cores)	130	59.7
Nvidia GeForce GTX-780	250	288.4
Altera Nallatech PCIe-385N	25	24.0

The Simplex Algorithm

- The Simplex Algorithm solves LP problems by iterating over subsets of the problem variables to generate improving solutions.
- Each simplex iteration consists of three subroutines: pricing, ratio test, and pivot.
- Pivot finds the direction of the adjacent solution, ratio test calculates the magnitude of the reduction in the cost function, and pivot updates the problem data structures.
- The three subroutines are performed on the OpenCL Device in a loop managed by the Host processor.

Figure 2: OpenCL Simplex Algorithm Design



a) The Simplex Algorithm b) OpenCL Implementation

Performance Benchmarking

Figure 3: Speed Up for the OpenCL Solver Over the Sequential Solver

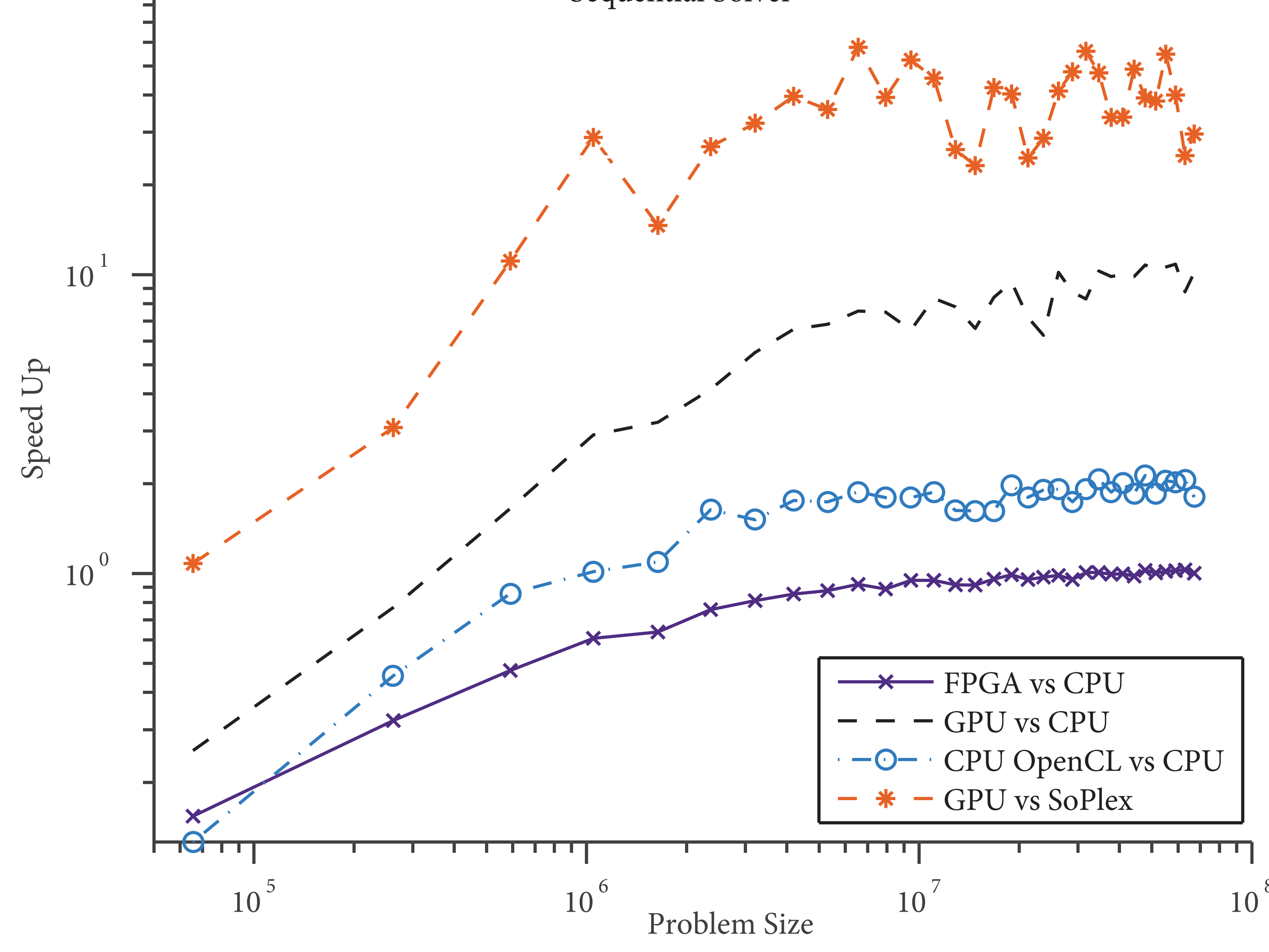
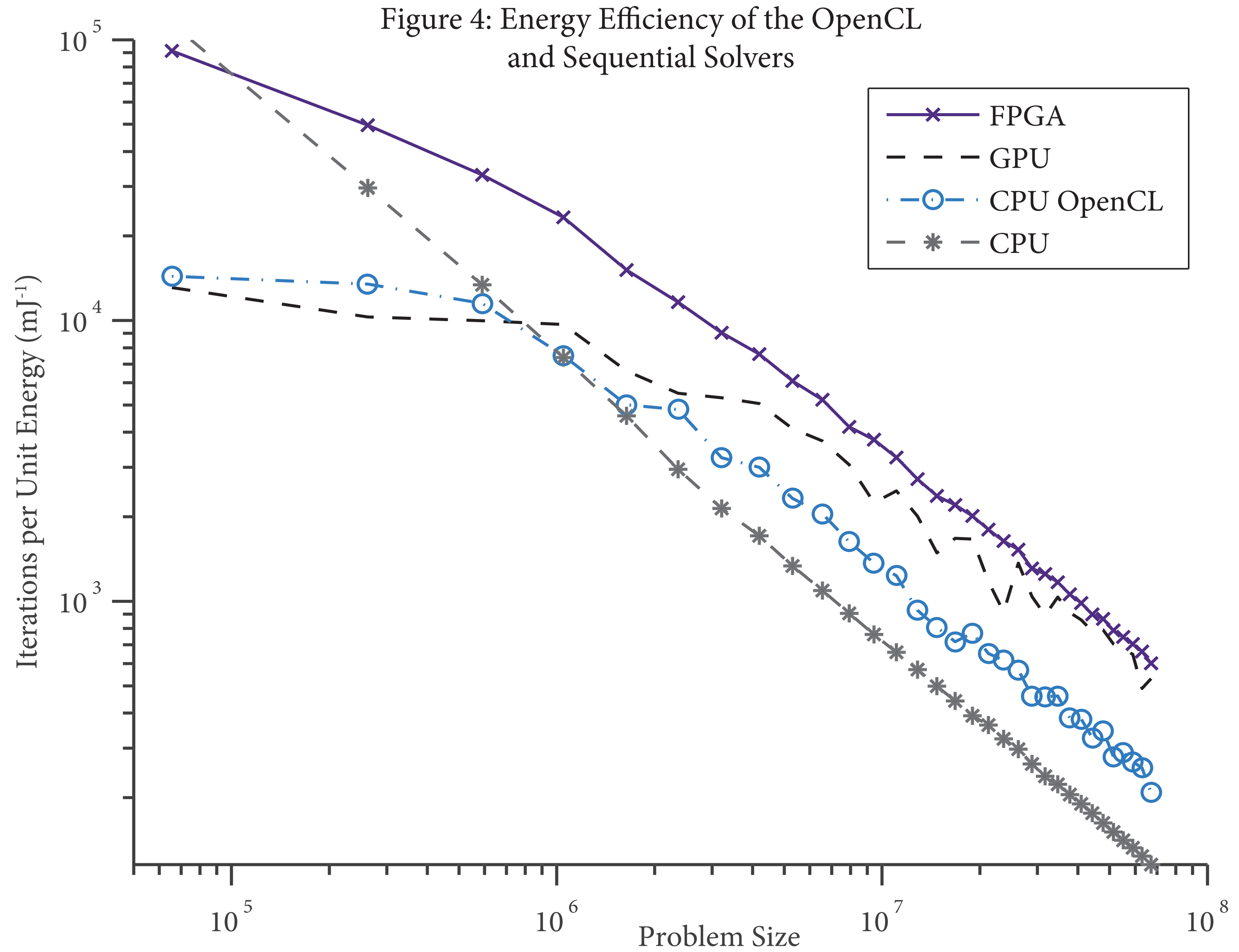


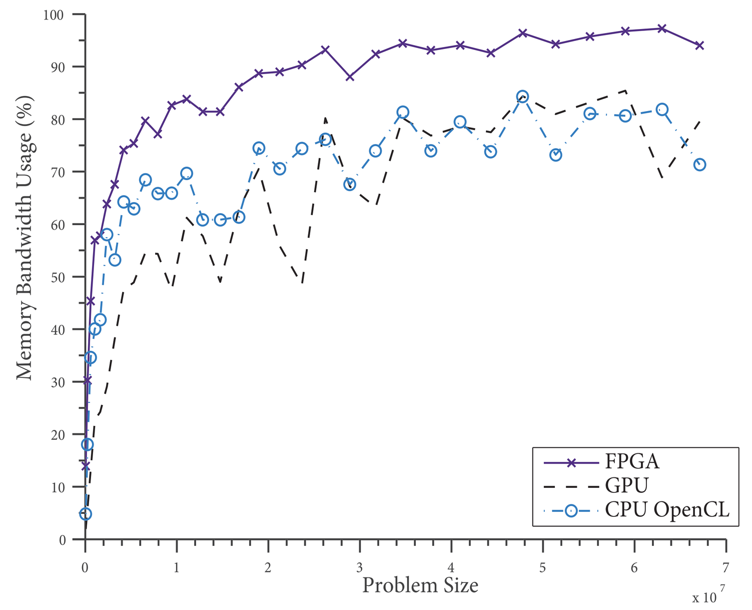
Figure 4: Energy Efficiency of the OpenCL and Sequential Solvers



Design Analysis

- Performance benchmarking reveals speed ups relative to sequential code that approach 2 and 10 on a CPU and GPU.
- The FPGA exhibits close to unity speed up but proved to be the most efficient in terms of Simplex iterations processed per unit energy with an efficiency 5 times greater than the CPU.
- The design effectively saturates the memory bandwidth of each test device as shown in Figure 5.

Figure 5: Memory Bandwidth Usage



Conclusions

- A heterogenous OpenCL system is proposed to solve dense LP problems and tested on multiple hardware accelerators.
- Test results indicate that the GPU provides the fastest solutions to LP problems as shown in Figure 3.
- The selected FPGA provides superior energy efficiency for the application as shown in Figure 4.

Acknowledgement

This work was supported in part by the Southern Ontario Smart Computing Innovation Platform Consortium (SOSCIP) and IBM Canada Ltd.

The authors would like to thank Dr. Mary Fenelon, Dr. Sean Wagner, Mr. Laszlo Ladanyi and Mr. Blair Adamache from IBM for their valuable comments and suggestions.

References

[1] A. Hamzic, A. Huseinovic, and N. Nosovic, "Implementation and performance analysis of the simplex algorithm adapted to run on commodity opencl enabled graphics processors," in *2011 XXIII International Symposium on Information, Communication and Automation Technologies (ICAT)*, 2011, pp. 1–7.

[2] S. Bayliss, C.-S. Bouganis, G. Constantinides, and W. Luk, "An fpga implementation of the simplex algorithm," in *IEEE International Conference on Field Programmable Technology*, 2006, pp. 49–56.

[3] A. Olafsson and S. Wright, "Linear programming formulations and algorithms for radiotherapy treatment planning," *Optimization Methods and Software*, vol. 21, no. 2, pp. 201–231, April 2006.

[4] H. Romeijn, R. Ahuja, J. Dempsey, and A. Kumar, "A new linear programming approach to radiation therapy treatment planning problems," *Operations Research*, vol. 54, no. 2, pp. 201–216, April 2006.

[5] R. Wunderling, "Paralleler und objektorientierter Simplex-Algorithmus," Ph.D. dissertation, Technische Universitat Berlin, Algorithmus," Ph.D. dissertation, Technische Universit¨at Berlin, 1996, http://www.zib.de/Publications/abstracts/TR-96-09/.