How GPUs can find your next hit: Accelerating virtual screening with OpenCL

Simon Krige
ACS 2013
Agenda

> Background
> About **blaze**V10
> What is a GPU?
> Heterogeneous computing
> OpenCL: a framework for parallel computing
> Porting **blaze**V10 to OpenCL: **blaze**V10 GPU
> **blaze**V10 GPU benchmark
> Science Advantages of GPUs
> Conclusions
About me

> Originally from Geneva, Switzerland.
> Graduated with a Masters degree in computer science from the University Of Bristol two years ago.

> Working on an 18 months project at Cresset in collaboration with the University of Bristol, funded by the Knowledge Transfer Partnership
> About 14 months into the project now!
About Cresset

> Founded in 2002 by Dr Andy Vinter
> Use shape and electrostatics of ligands to compare molecules in 3D

> Software
  > Ligand based virtual screening
  > Develop pharmacophores and understand structure activity relationships
  > Find novel bioisosteric replacements for parts of your molecule

> Services
  > Full range of computational chemistry services
Non-Classical Molecular Comparisons

Non-Classical Molecular Comparisons

Fields 0.66
Shape 0.98

> Ligand based virtual screening - search a database with a query structure, retrieve a hit list

> Runs on a Linux cluster

> Can screen ~5 million compounds in a few hours
  > 100→500 cpu cluster
  > i.e. a high number of CPUs working together

> We would like a less CPU hungry (cheaper) solution!
We want to do this as fast as possible!
GPUs explained

- GPU: Graphical Processing Unit
- Designed to build images and output to display: calculations related to 3D computer graphics
- Many-core architecture makes them ideal parallel processors
- Shader pipeline now used for general-purpose computing power, as opposed to being hard wired solely to do graphical operations
- We now talk about GPGPUs, HPC GPUs and accelerator devices
Nvidia Tesla K20 GPU

7.1 Billion transistors, 2496 cores!
Heterogeneous computing

> Definition (wikipedia) : electronic system that uses a variety of different types of computational units

> A modern platform includes:
  > One or more CPUs
  > One or more GPUs
  > DSP processors
  > … other?

Samsung Exynos 4 Quad

  > Quad-core ARM Cortex-A9 1.6 GHz
  > Quad-core Mali-400
It’s a heterogeneous world!

- Some familiar heterogeneous devices:

- The Heterogeneous many-core challenge: **How are we to build a software ecosystem for the Heterogeneous many core platform?**
OpenCL: Industry Standard for Programming Heterogeneous Platforms

- **CPUs**
  - Multiple cores driving performance increases
  - Multi-processor programming – e.g. OpenMP

- **GPUs**
  - Increasingly general purpose data-parallel computing
  - Graphics APIs and Shading Languages

**OpenCL**

- **Emerging Intersection**

**Heterogeneous Computing**

**OpenCL – Open Computing Language**

Open, royalty-free standard for portable, parallel programming of heterogeneous parallel computing CPUs, GPUs, and other processors
Creating **blaze**V10 GPU

- Full port from highly optimised FORTRAN77 to OpenCL.
- **Steps:**
  1. Download OpenCL drivers and SDK for your platform
  2. Profile serial code and identify computationally intense functions
  3. Line by line conversion of serial code to OpenCL kernels
  4. Optimize memory accesses, vector operations, work group sizes
  5. Test and benchmarks on different platforms
- **Challenges:**
  - Some device drivers are not fully mature and can be unstable
  - Optimizing parallel code is not easy, it requires excellent knowledge of the underlying architecture
  - Porting serial code to OpenCL can be a lengthy process overall
blazeV10 GPU benchmark

- ~1000 molecules – 80k conformations
- Standard instance: 12 conformations processed per second on a single core of Intel® Core i7-3770 CPU @ 3.40GHz
- CPUs:
  - Intel® Core i7-3770 CPU @ 3.40GHz (4 cores - 4 threads)
  - Intel® Xeon CPU X5570 @ 2.93GHz (4 cores – 8 threads)
  - Intel® Xeon CPU E5645 @ 2.40GHz (6 cores - 12 threads)
- GPGPUs:
  - NVIDIA GTX580
  - AMD HD7970
- HPC GPUs:
  - NVIDIA M2050
  - NVIDIA K20
- CPUs and GPUs will work together, it’s a heterogeneous world!
blazeV10 GPU benchmark

Speedup graph

Higher is better

Top speedup: 63x

Amazon EC2

GPU node

Nvidia K20

45x

M2050
GTX580
HD7970
K20
K20+Intel i7
M2050*2 + Intel X5...
GTX580+Intel E5645
GTX580*2
Saving you space and money

- **Less hardware**: blazeV10 GPU is ~45 times faster on a K20 than blazeV10 on a single core of Intel i7 CPU!

- **It’s cheaper**: for a $2.10/hour GPU instance on AmazonEC2 you can process 2m conformations, whereas you can only process 1.3m conformations with 14 dual-core High-CPU Medium instances.

- **It’s greener**: one GPU consumes ~400watts, one quad-core CPU workstation consumes ~200watts: we achieve 5x performance per watt = 5x less gCO2 consumed per answer.
Science Advantages

> Faster Virtual screening
  > Easier deployment
  > Cheaper
  > Desktop box with 4GPUs vs. 150 node cluster

> New science
  > Using multiple molecule 3D comparisons in new ways
  > Similarity matrices

> Easier
  > Manage fewer instances on AmazonEC2

> Accurate
  > Results accuracy is preserved i.e. we are not sacrificing accuracy for speed.
> GPUs are an excellent solution to accelerating your software

> OpenCL is the way to go if you want platform independent parallel code that will take advantage of all available resources

> Accelerating your code on a GPU can make your software much faster, greener and hardware resource efficient

> **But** it can be a lengthy process that requires a good programmer(s).
Acknowledgements

> Bristol
  > Prof. Simon McIntosh-Smith
  > Prof. Richard Sessions

> Cresset
  > Dr. Mark Mackey
  > Dr. Tim Cheeseright

> Nvidia
  > Mark Berger for donation of a K20 GPU
Thank you

simon@cresset-group.com

Cresset Group
Cressetgroup
Cresset

www.cresset-group.com