IEEE Software

Sunday, January 8, 2017

BarraCUDA in the Cloud

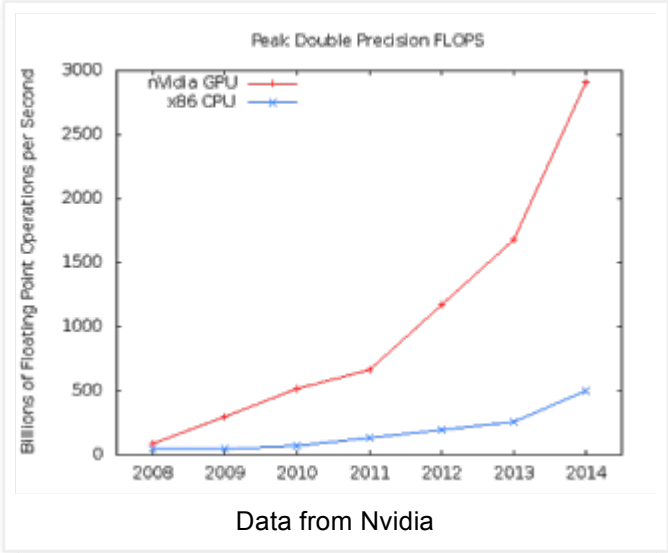
by W.B. Langdon, UCL and Bob Davidson, Microsoft
Associate Editor: Federica Sarro (@f_sarro), University College London, UK

What is this, flying fishes? Well no. BarraCUDA is the name of a Bioinformatics program and the cloud in question is Microsoft's Azure, which is in the process of being upgraded with copious nVidia K80 tesla GPUs which support CUDA in instances of virtual machines. BarraCUDA has been around for a few years [1]. It is a port of BWA [2] which takes advantage of the massive parallelism available on graphics hardware (GPUs) to greatly speed up approximate matching of millions of short DNA strings against a reference genome. For example, the human reference genome [3]. Approximate matching is necessary, because of noise but primarily because the medical purpose of many DNA scans is to reveal differences between them and "normal" (i.e. reference) DNA. A typical difference is to substitute one character for another, but tools like BarraCUDA also find matches where a character is inserted and where one is deleted. Although there are many sources of DNA data, BarraCUDA and similar programs are targeted at strings generated by "Next Generation Sequencing" (NGS) machines. These are amazing devices. A top end NGS machine is now capable of generating more than a billion DNA strings, sequences of A, C, G or T letters. Part of the trade-off for this speed is the strings are short (typically a hundred letters long) and noisy. The first step is to find where the short fragments of DNA came from by aligning the strings against a reference genome. To account for the various sources of noise, NGS is usually run with three fold redundancy and sometimes a particularly important part of a person's genome may be scanned ten or more times. Given multiple alignments to the same part of the reference genome, it becomes possible to look for consistent variations.

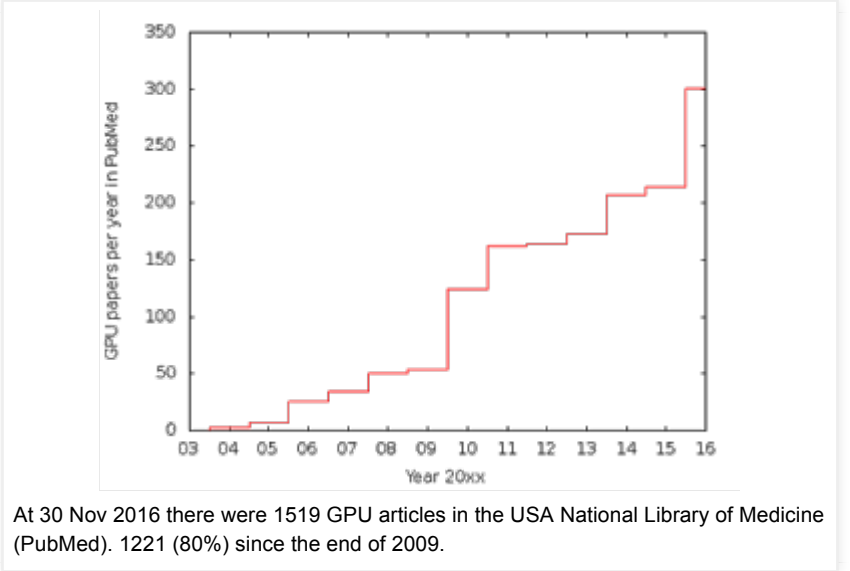
BWA, BarraCUDA and Bowtie are members of a family of Bioinformatics tools which have proved successful because they are able to compress the human reference genome into less than 4 gigabytes of RAM, making it possible to run an important part of the DNA analysis tool chain on widely available computers. Indeed in the case of BarraCUDA, GPUs with 4GB are also widely available. Recently BarraCUDA was optimised using genetic improvement[4,5] (see blog posting February 3, 2016). This updating prompted the question was it possible to use BarraCUDA with epigenetics data.

To grossly oversimplify, whilst (to a good approximation) all the cells in your body contain the same DNA, what makes your cells different from each other is how that DNA is used. It is

thought that to a large extent how DNA is enabled and disabled is controlled by epigenetic makers on that DNA itself. These epigenetic markers differ between cells. Indeed the markers change not only between cells but also with the person's age and factors outside the cell. Since this is not fully understood, the study of epigenetics, particularly how it relates to disease is a very active topic. Much of the Next Generation Sequencing technology can be reused by epigenetics. However when matching epigenetic sequences against a reference, the reference is twice the size of the DNA reference. Fortunately this need has coincided with the launch of GPUs with larger memory (e.g. the Tesla K40 has 12GB). Which in turn has coincided with the introduction of Azure cloud nodes with multiple K40s or K80s. Recently we have been benchmarking [6] BarraCUDA on epigenetics data supplied by Cambridge Epigenetics on Azure nodes.



Data from Nvidia



At 30 Nov 2016 there were 1519 GPU articles in the USA National Library of Medicine (PubMed). 1221 (80%) since the end of 2009.

References

[1] P. Klus et al., BarraCUDA. BMC Res Nts, 5(27), 2012.

[2] Heng Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009, 25(14):1754-1760.

[3] Initial sequencing and analysis of the human genome. Nature 409, 6822, (15 Feb 2001), 860–921.
[4] W.B. Langdon. Genetically improved software. In Amir H. Gandomi et al., editors, Handbook of Genetic Programming Applications, chapter 8, pages 181–220. Springer, 2015.
[5] W.B. Langdon, Brian Yee Hong Lam, M. Modat, J. Petke, M. Harman. Genetic Improvement of GPU Software. Genetic Programming and Evolvable Machines. Online first.
[6] W.B. Langdon, A. Vilella, Brian Yee Hong Lam, J. Petke and M. Harman, Benchmarking Genetically Improved BarraCUDA on Epigenetic Methylation NGS datasets and nVidia GPUs. In Genetic Improvement 2016, (GECCO 2016) workshop, pages 1131-1132, 20-24 July, Denver.

You might also enjoy reading
Genetic Improvement, IEEE Software blog, February 3, 2016
GPGPUs for bioinformatics, Oxford Protein Informatics Group, April 17, 2013
Genome Sequencing in a Nutshell, Deborah Siegel and Denny Lee, May 24, 2016

Posted by Federica Sarro at 2:36 PM

No comments:

Post a Comment

Home Older Post

Subscribe to: Post Comments (Atom)