Dealing With Big Data Outside Of The Cloud GPU Accelerated Sort

John Vidler[1] Paul Rayson[1] Laurence Anthony[2] Andrew Scott[1]
John Mariani[1]

[1] School of Computing and Communications, Lancaster University
{j.vidler, p.rayson, a.scott, j.mariani}@lancaster.ac.uk

[2] Faculty of Science and Engineering, Waseda University
anthony@waseda.jp

31 May 2014

Table of Contents

1 Motivation

2 Solution

3 Data

4 Results

5 Summary

Motivation

Corpus data is used in ...

- Digital Humanities
- Natural Language Processing
- (Historical) Text Mining
- Corpus Linguistics

Motivation

Big Data!

- Corpora are becoming un-processable due to their large size
  - Large digitisation initiatives (Digital Humanities)
  - Web as Corpus (Corpus Linguistics)
- Fitting them in memory is increasingly a challenge! (24G max in xeon)
- Processing the data held in memory is cumbersome (long processing times)

Motivation

Current solutions

- International infrastructure projects (CLARIN, DARIAH)

Motivation

Current solutions

- International infrastructure projects (CLARIN, DARIAH)
  - Do not allow for local access to support researchers during resource creation and iterative analysis

Motivation

Current solutions

- International infrastructure projects (CLARIN, DARIAH)
    - Do not allow for local access to support researchers during resource creation and iterative analysis
- Online tools (Sketch Engine, BYU Corpora)

Motivation

Current solutions

- International infrastructure projects (CLARIN, DARIAH)
  - Do not allow for local access to support researchers during resource creation and iterative analysis
- Online tools (Sketch Engine, BYU Corpora)
  - Remotely hosted, not easy to replicate locally

Motivation

Current solutions

- International infrastructure projects (CLARIN, DARIAH)
  - Do not allow for local access to support researchers during resource creation and iterative analysis
- Online tools (Sketch Engine, BYU Corpora)
  - Remotely hosted, not easy to replicate locally
- Semi-cloud based tools (GATE, Wmatrix, CQPweb)

Motivation

Current solutions

- International infrastructure projects (CLARIN, DARIAH)
    - Do not allow for local access to support researchers during resource creation and iterative analysis
- Online tools (Sketch Engine, BYU Corpora)
    - Remotely hosted, not easy to replicate locally
- Semi-cloud based tools (GATE, Wmatrix, CQPweb)
    - Installation and configuration not accessible to SSH researchers

Motivation

A remaining need

- Investigate processing efficiency improvements for locally controlled and installed corpus retrieval software
- Core tasks such as indexing, n-grams, collocations, sorting results in concordances cannot be carried out locally in reasonable time

Motivation

A Case Study

Can we leverage the power of GPUs to aid corpus processing?

Table of Contents

1 Motivation

2 Solution

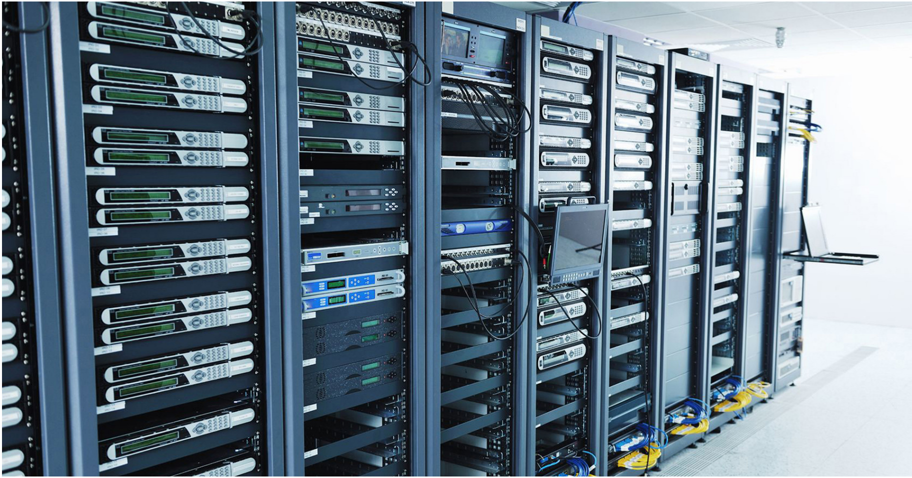
3 Data

4 Results

5 Summary

Hardware

The traditional way



Hardware

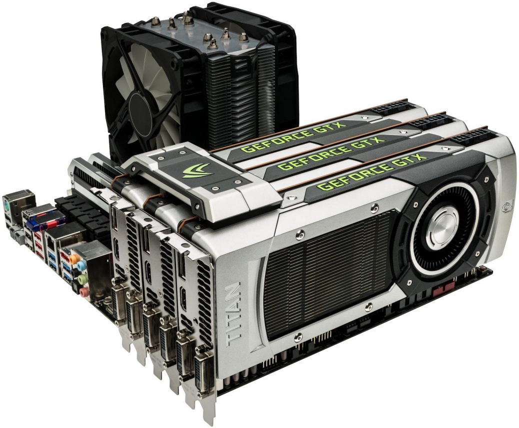
The not-so-traditional way

Card Comparison

	 GT 620	 **GTX Titan**	 **Tesla K40**
Cores	96	192	2880
Memory	128 MB	6 GB	12 GB
Address Width	64 bit	384 bit	384 bit
Copy Engines	1	1	2
Cost (GBP)	≈ £30	≈ £500 – 600	≈ £3200

Hardware

Scalability



It is possible to run several cards at once - our experiments only used one.

Table of Contents

1 Motivation

2 Solution

3 Data

4 Results

5 Summary

Data Sources

Corpus Source:

Data Sources

Corpus Source: Project Gutenberg's Library

1. Download the snapshot DVD
2. Extract the text-format books
3. Walk the files grabbing collocations lines for specific common words

Data Sources

Corpus Source: Project Gutenberg's Library

1. Download the snapshot DVD
2. Extract the text-format books
3. Walk the files grabbing collocations lines for specific common words
    - A quick Java tool was used for this ...
    - ... normally to be done by querying a database

Data Sources

Corpus Source: Project Gutenberg's Library

1. Download the snapshot DVD
2. Extract the text-format books
3. Walk the files grabbing collocations lines for specific common words
   - A quick Java tool was used for this ...
   - ... normally to be done by querying a database

Data Sources

Example Input

Preceeding 10 words	Pivot	Subsequent 10 words
… began to diminish and soon	there	were no more visitors …
… as though it had been	there	for months He even went the …
… that as yet	there	were no signs of decomposition …
… the stairs were distinctly heard	There	was silence for a few …
… ready to go downstairs when	there	appeared before her her son …
… terms of this agreement	There	are a few things that …
… agreement See paragraph C below	There	are a lot of things you …

A section of input data, similar to that which might
be generated by LWAC, or AntConc, for example.

Table of Contents

1 Motivation

2 Solution

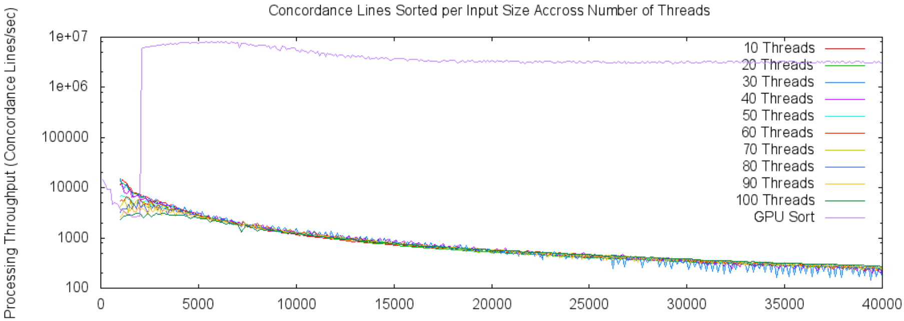
3 Data

4 Results

5 Summary

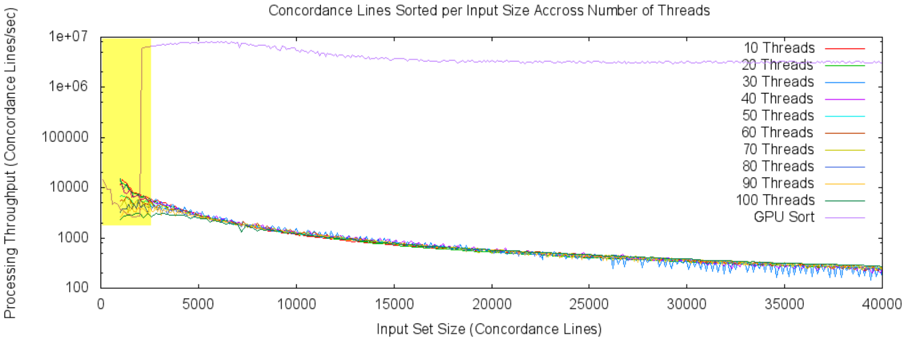
Results

Running on the GPU



Results

Running on the GPU



Results

Running on the GPU

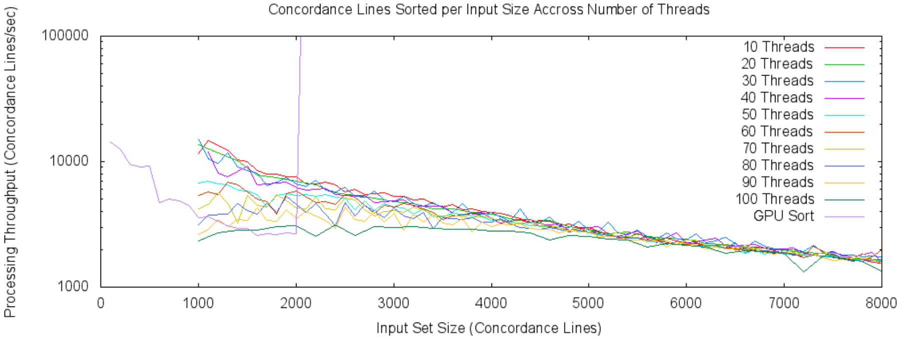


Table of Contents

1 Motivation

2 Solution

3 Data

4 Results

5 Summary

Summary

- GPU computing does offer time gains for linguistic processes

Summary

- GPU computing does offer time gains for linguistic processes

But...

- The program design has to be carefully considered
  - Not a ‘normal’ set of processors!
  - Current equipment is very batch-mode, dynamic pipelines are either difficult or impossible.
- Longer, more complex processes work better, earlier
  - Our experiments actually do too little on the GPU!

Questions

Thank You

Any comments, questions?

References

GT 260 specification (nvidia). 2014. URL http://www.geforce.co.uk/hardware/desktop-gpus/geforce-gt-620/specifications.

GTX titan specification (nvidia). 2014. URL http://www.nvidia.co.uk/gtx-700-graphics-cards/gtx-titan-black/.

Daniel Cederman and Philippas Tsigas. Gpu-quicksort: A practical quicksort algorithm for graphics processors. *J. Exp. Algorithmics*, 14:4:1.4–4:1.24, January 2010. ISSN 1084-6654. doi: 10.1145/1498698.1564500. URL http://doi.acm.org/10.1145/1498698.1564500.

Yangdong Steve Deng. IP routing processing with graphic processors. *2010 Design, Automation & Test in Europe Conference & Exhibition (DATE 2010)*, pages 93–98, March 2010. doi: 10.1109/DATE.2010.5457229. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5457229.

Carlos Aguilar Melchor, Benoit Crespin, Philippe Gaborit, Vincent Jolivet, and Pierre Rousseau. High-Speed Private Information Retrieval Computation on GPU. In *Proceedings of the 2008 Second International Conference on Emerging Security Information, Systems and Technologies*, pages 263–272, Washington, DC, USA, August 2008. IEEE Computer Society. ISBN 978-0-7695-3329-2. doi: 10.1109/SECURWARE.2008.55. URL http://portal.acm.org/citation.cfm?id=1447563.1447928.

Layali Rashid, WessamM. Hassanein, and MoustafaA. Hammad. Analyzing and enhancing the parallel sort operation on multithreaded architectures. *The Journal of Supercomputing*, 53(2):293–312, 2010. ISSN 0920-8542. doi: 10.1007/s11227-009-0294-5. URL http://dx.doi.org/10.1007/s11227-009-0294-5.

Weibin Sun, Robert Ricci, and Matthew L. Curry. GPUstore. In *Proceedings of the 5th Annual International Systems and Storage Conference on - SYSTOR '12*, pages 1–12, New York, New York, USA, 2012. ACM Press. ISBN 9781450314480. doi: 10.1145/2367589.2367595. URL http://dl.acm.org/citation.cfm?id=2367595.

Stephen Wattam, Paul Rayson, Marc Alexander, and Jean Anderson. Experiences with Parallelisation of an Existing NLP Pipeline : Tagging Hansard. In *Proceedings of The 9th edition of the Language Resources and Evaluation Conference*, 2014.