Using GPUs to Mine Large Scale Software Problem Repositories

Tim Dunn

Department of Electrical and Computer Engineering, Clarkson University
Department of Computer Science, Clarkson University

Large Software Projects

Mozilla Firefox contains over 18.6 million lines of code

People submit observed failures in what are known as “bug reports”

An Example Bug Report

Bug ID:	915
Summary:	(col-align-inherit) implement inheritance of alignment attributes from columns (align, valign, char, charoff, (lang, dir)?)
Description:	something about a missing colframe...
Submitter:	kipp
Submitted:	9 - 26 - 1998
Modified:	6 - 7 - 2015
Status:	NEW
Product:	Core
Component:	CSS Parsing and Computation
Version:	Trunk
Comments:	396 total

Organizing Reports

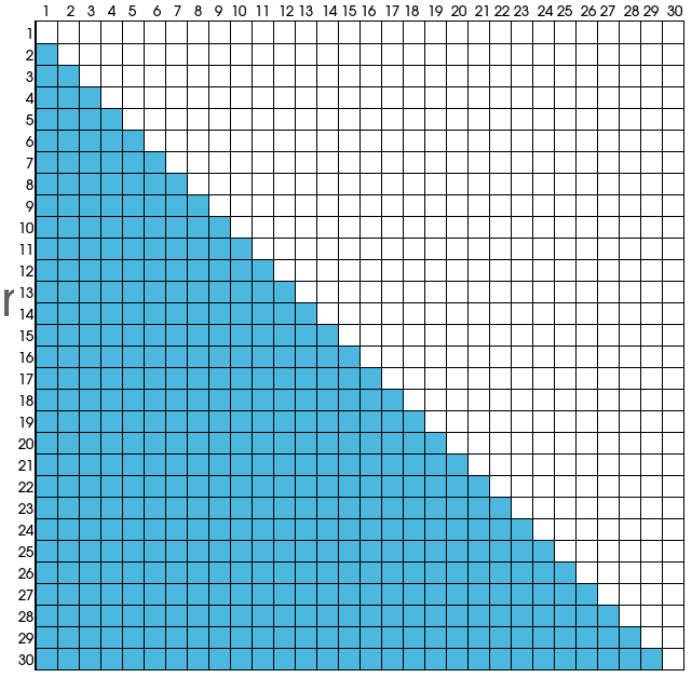
Repository	Total Reports	Average Reports per Day
RedHat	1,357,998	210
Mozilla	1,287,896	197
Novell	989,610	233
Eclipse	498,161	92

Table 1: Software Problem Repository Sizes (as of July 19, 2016)

How Long Will This Take?

0 + 1 + 2 + 3 + ... + (n-2) + (n-1) compar

$$\sum_{k=0}^{n-1} k = \frac{n(n-1)}{2} \approx \frac{n^2}{2}$$



Mozilla has about 1,290,000 reports, so a $8.2 \cdot 10^{11}$ comparisons are required

If you compare one report to another per second, this would take about 26,000 years

Automated Duplicate Report Detection

Generate a list of 20 most similar bug reports, and have a developer look through that list to search for duplicates

Current methods have **recall rates** of 60-80%

But, are evaluated on small subsets of the full dataset

NEW REPORT #183

114:	0.257325
29:	0.237171
121:	0.221359
1:	0.217584
168:	0.216930
53:	0.204124
14:	0.202444
18:	0.200000
23:	0.197642
132:	0.195180
.	
.	
.	

Why is this Challenging?

Authors can use different language to describe the same issue

Many reports are of poor quality and fairly short

Numerous spelling errors hinder effective report comparisons

Many non-native English speakers are submitting problem reports

Report Comparison Methods

Sequence Length (words)

Longest Common Substring

right clicking the mouse is not working for me

clicking the help icon does not do anything for several seconds

2

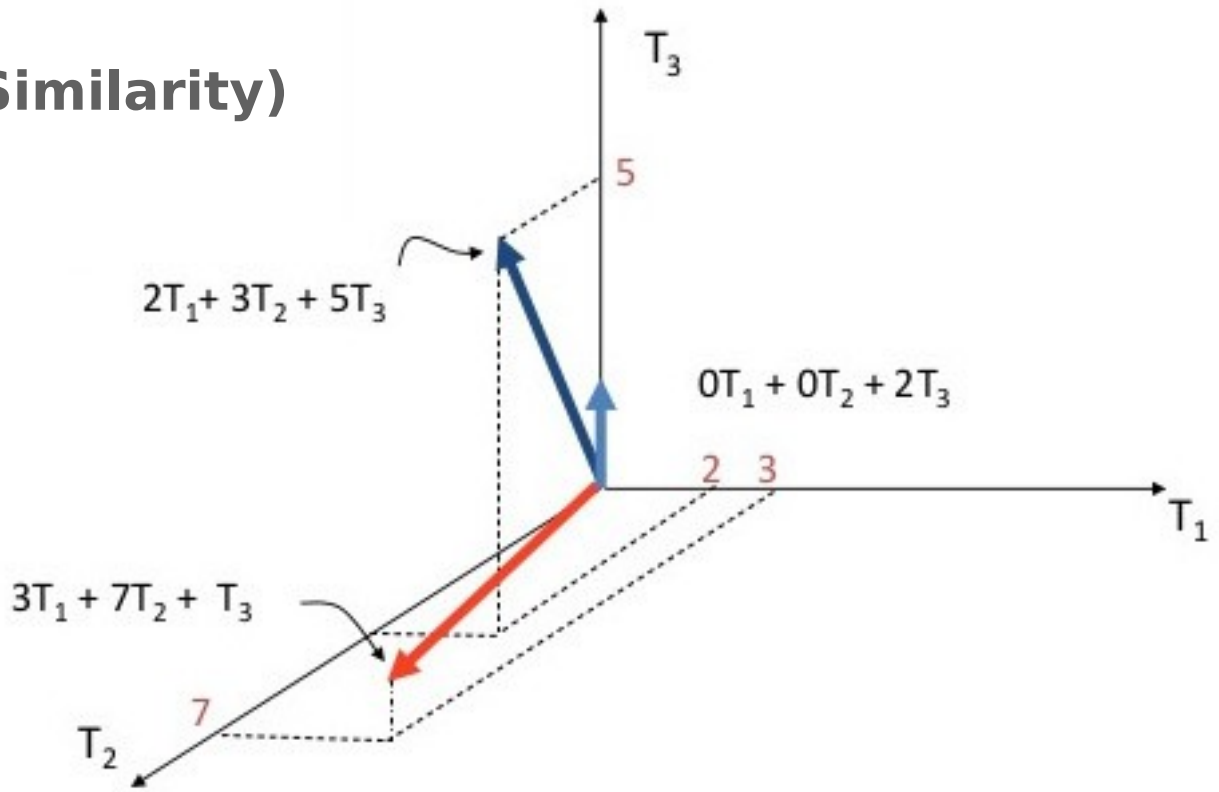
Longest Common Subsequence

right clicking the mouse is not working for me

clicking the help icon does not do anything for several seconds

4

The Vector Space Model
(Cosine Similarity)



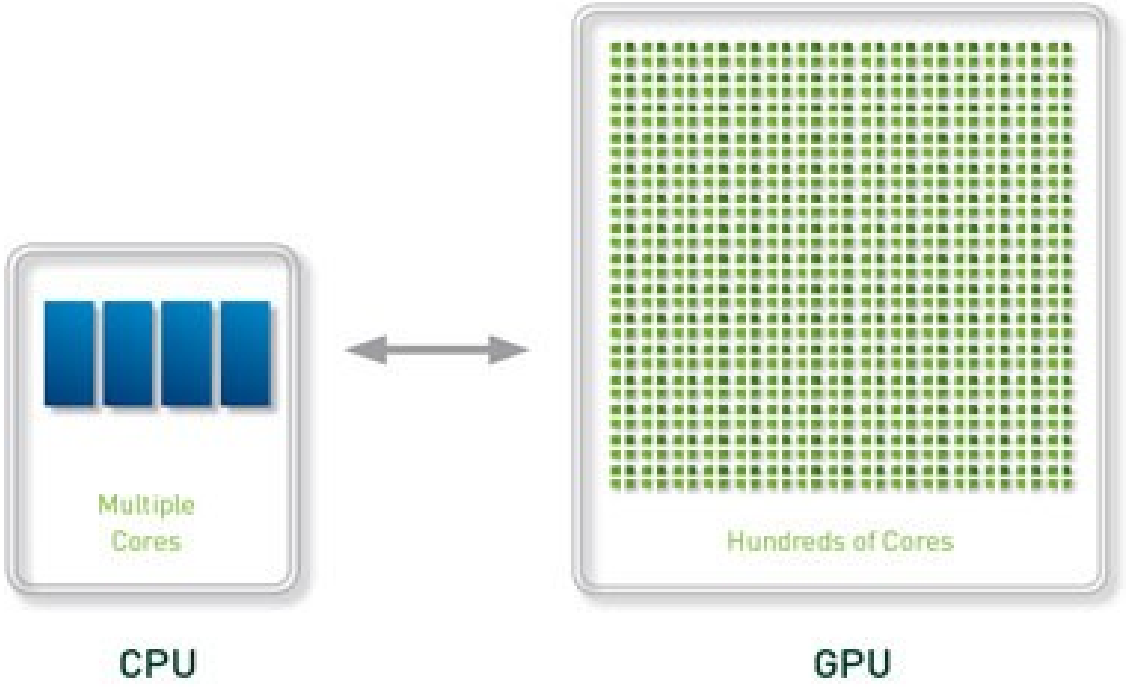
How Fast are These Methods?

Test setup:

Intel Xeon E5-2660-v3 10 core processor with 128GB of RAM
Maximizing computations to fully utilize 1 of the 10 cores

Repository	Longest Common Subsequence	Longest Common Substring	Cosine Similarity	Total Time
Testing Set (1,000 reports)	12.8 seconds	12.8 seconds	3.45 seconds	**29 seconds**
Eclipse (498,161 reports)	36.8 days	36.8 days	9.9 days	**83.5 days**
Mozilla (1,287,896 reports)	245.7 days	245.7 days	66.2 days	**557.6 days**

Graphics Processing Unit

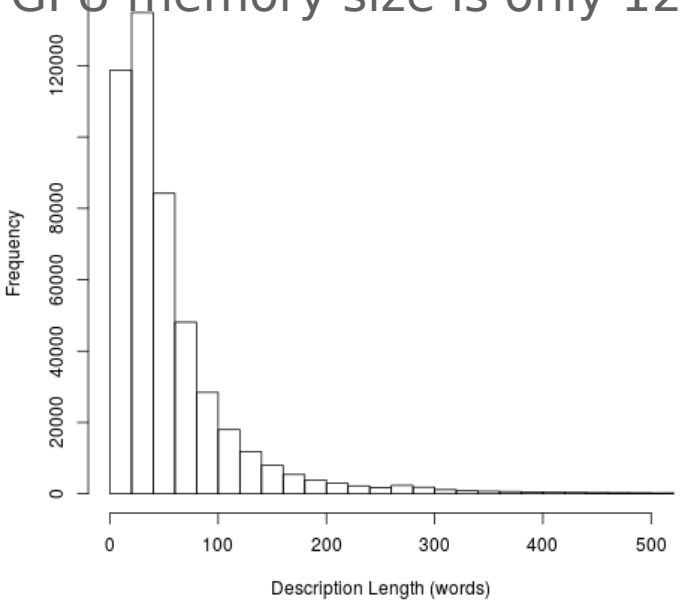


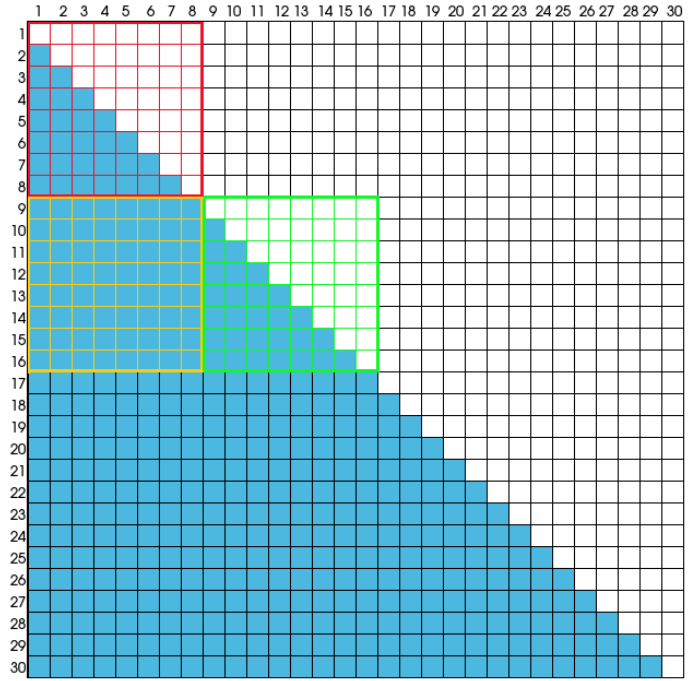
http://www.nvidia.com/object/what-is-gpu-computing.html

http://www.nvidia.com/content/tesla/images/tesla-k20-series.jpg

Challenges of Using GPUs

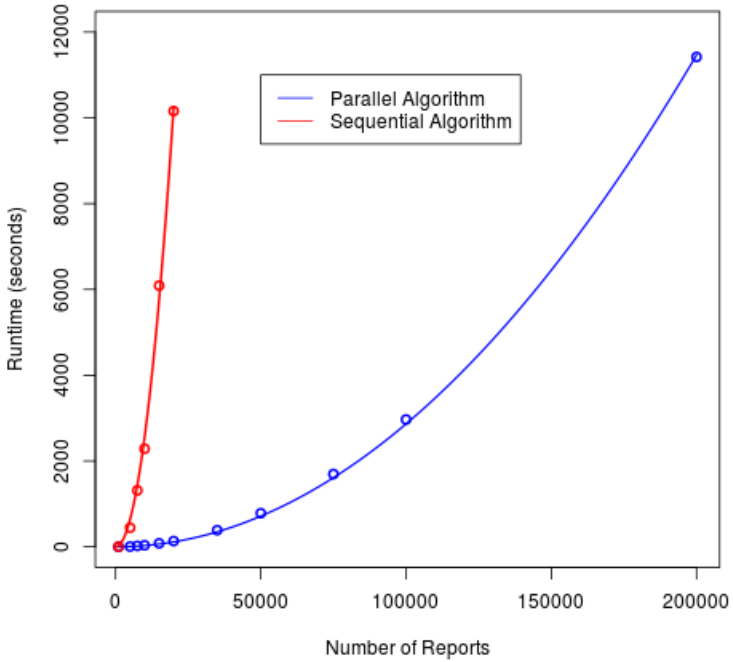
- Fast GPU memory uses constant sized arrays
- Total GPU memory size is only 12GB





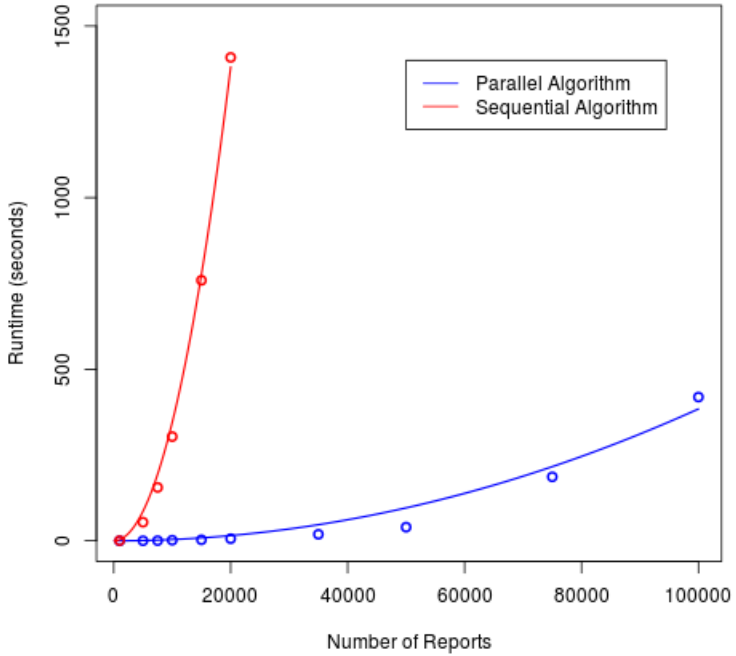
Results

Longest Common Subsequence and Substring



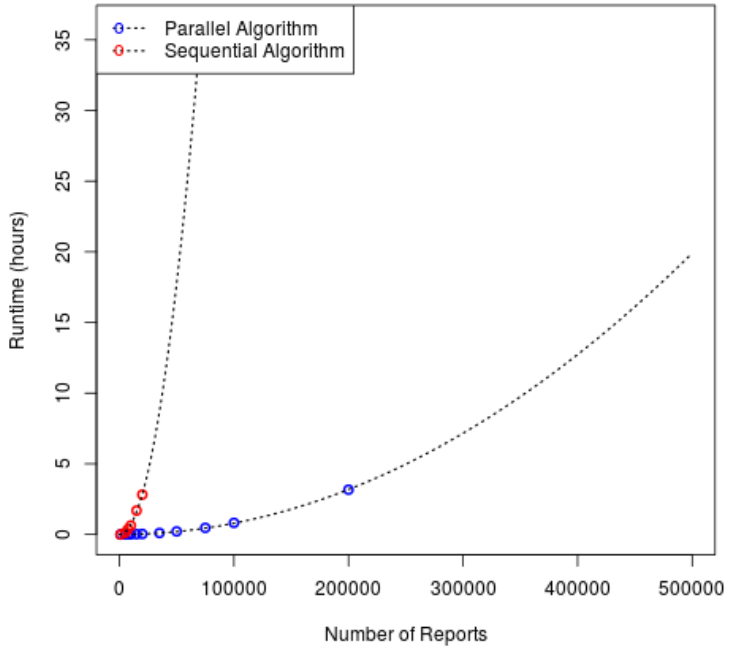
85.6x speedup

Cosine Similarity



89.8x speedup

Summary



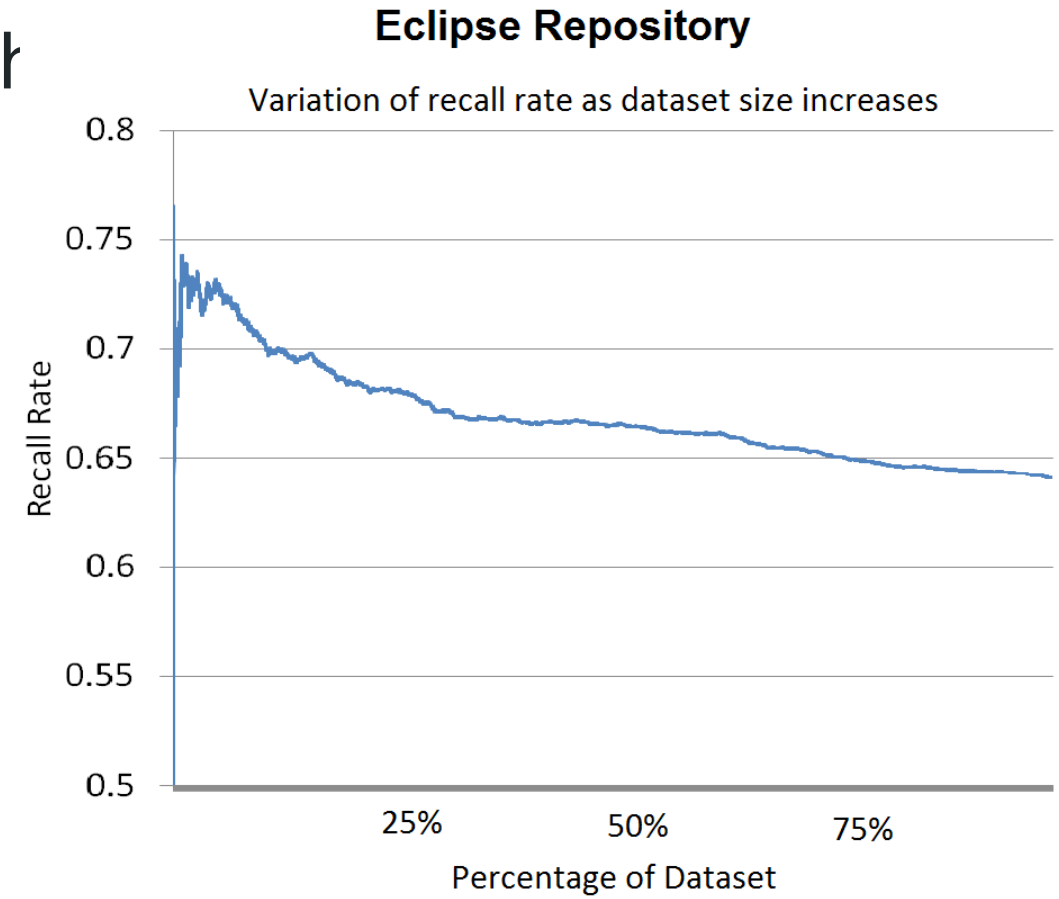
Repository	Total Reports	Human	Computer	GPU
Testing Set	1,000	5.8 days	29 seconds	0.337 seconds
Eclipse	498,161	3,882 years	83.4 days	23.3 hours
Mozilla	1,287,896	25,833 years	557.7 days	6.5 days

Continued Research

Faster runtimes will allow us to intelligently combine multiple similarity metrics

- e.g. topic modeling, cosine similarity, longest common substring, and longest common subsequence

- This addresses the challenge of declining recall rate as repository size increases



Acknowledgements

Thank you to:

- NVIDIA Corporation for donating a GPU
- Corning Inc. for providing a research stipend
- Clarkson University Honors Program for providing housing
- Professor Banerjee for his help and advice