CS Capstone Spring Midterm Progress Report

May 6, 2018

# Privacy Preserving Cloud, Email, and Password Systems

Prepared for

OSU

Attila Yavuz

Prepared by

Group 38

The Secret Bunny Team

Andrew Ekstedt
Scott Merrill
Scott Russell

**Abstract**

This document provides a report of progress our team has made on our project during the first half of the Spring 2018 term. We outline our current status, problems we've encountered, and future work. We also provide individual accounts of what each group member has accomplished.

## CONTENTS

## 1 Project Purpose and Goals

The "Privacy Preserving Cloud, Email and Password Manager" Capstone project is a research-oriented project that aims to find a way to implement the DSSE scheme proposed by David Cash. This implementation will be executed through command line prompts and hosted on OSUs engineering servers. A user can use this system with a client-server model to perform actions, such as search or update, on a "cloud-based" database. User interface is not considered a priority as this project is not intended to be used in any commercial capacity. There are four primary goals of this project. First the implementation of David Cash's Basic DSSE Algorithm. Next is to implement Cash's piPack. Finally to implement Pointer optimization. 4th we will be comparing these 3 version of David Cash's scheme against Attila Yavuz's IM-DSSE Bit Matrix Scheme. Finally we have benchmarking analysis between these two schemes on a small and medium size dataset.

## 2 Current Status

Implementation of the Core DSSE proposed by David Cash is the priority of this project. So that's what we worked on. Functionality is all complete except for that of the pointer optimization. We have the following methods implemented

- Setup - create an initial encrypted search index from a set of documents
- Search - search for a token in the encrypted database
- Add - update the encrypted index by adding tokens to a file
- Basic - These above methods are part of the Basic implementation

Also:

- piPack optimization
- Benchmarking analysis on Small and Medium Data Sets.

### 2.1 Andrew Ekstedt

I finished the bulk of my work last term, so this term has been about helping my teammates and getting ready for expo. This has included gathering benchmarking data for Scott Russell, and talking with Scott Merrill about the finer points of implementing the optimized variants of the DSSE algorithm. I also worked on improving the UI of our code a little.

In week 3, I started writing a simple web app that we can use to demonstrate our project at expo. This isn't strictly part of our project requirements, but we feel that it would help people understand our project, and it also meshes well with our project theme of showing how these encryption techniques can be applied in a way that people can actually use. However, this plan got pushed to the back burner while we focus on finishing the last bits of our project. Hopefully there will be time to work on it more during the next couple weeks before expo, but if not, that's fine.

Other things I've done this term include writing a WIRED-style article about another capstone project, and contributing to the group progress report and presentation.

### 2.2 Scott Russell - Benchmarking

For me this term has been entirely focused on the benchmarking and data testing. With expo coming up and with our implementation of Basic, piPack and compilation of IM-DSSE complete I now have all the materials I need to begin deep dive benchmarking. These benchmarks include various database sizes ranging from a single file with 100k keywords

up to 40k files. Benchmarking analysis concluded that our implementation of David Cash's algorithm was substantially faster in all test cases. Also that the optimization of piPack performed marginally better as well relative to the basic scheme.

Relative to the Capstone side of things we all worked together on creating this report, individual sections, and presentation where we demoed our results of benchmarking with our finalized implementation of piPack, basic and compilation of IM-DSSE capabilities. I also revised the final poster draft with our benchmarking results and submitted to expo for printing. I wanted our design to stand out. Therefor we have a mascot bunny.

### 2.3 Scott Merrill - Optimization

Optimization was the main focus for this last term and a half. Outlined in David Cash's paper were three levels of optimization that could be done to potentially improve the efficiency of both space and speed of the DSSE scheme that we have been implementing. We were able to complete the first level of optimization by the end of last term. This term I have been focused on trying to complete the pointer optimization. This provides a lay of obfuscation to the overall scheme which allows for increased misdirection and difficulty for an "adversary" to learn any useful information about data access. This optimization has proven to add quite a bit of complexity to the core code that this program uses and it has been frustratingly slow to find a solution.

## 3 Problems that impeded progress, with solutions

As we begin implementation of our capstone project it is important to be able to look back at progress throughout the term. If we can understand what went well and poorly we can use that information to be able to have a more cohesive plan for the following terms. We plan to meet as a team twice every week in these final weeks leading up to expo to come together as a team to finalize work.

### 3.1 Missing TA

Our capstone TA was missing in action for the first few weeks of the term. We were eventually able to work out a meeting time in Week 4.

### 3.2 Group Dynamic

In the fall term we had a clear cut goal for each week. With a new writing assignment assigned and due it kept us on a good pace with what we imagined the project process would be. Now that we are "cut lose" to do our own implementation we are finding it hard to stay on track compared to our Gantt Chart that we imagined progress would be this term.

Communication with our client and TA has been difficult throughout the capstone process. Not being able to meet with them until later in the term it has impaired our ability to gain feedback and understanding of project requirements. We still have our initial Gantt Chart to compare progress to, and used that until meeting with our client. Another problem that Ive experienced this term is a lack of hard deadlines. In the fall term weve had deadlines for every document. From requirements to specifications these have all been created for our capstone portfolio. However, in this term there has been a very hands-off approach from the capstone team. This is the first time I have worked on such a style of project. It is very realistic to how real-world companies divvying out tasks, so I am grateful that we can practice these self-motivational skills in a less stressful environment when our jobs are not on the line.

In additional, being a research project, it is harder to put into words the progress that weve made outside of project code. For those teams focused on a more implementation heavy projects it is easier to show progress week by week. For example, one week I spent hours looking over and comparing POP3 algorithms against one another to find one that works well without specific implementation. I have listed these in my OneNote as progress, but it is hard to put into progress without code pushes on GitHub. It is directly relevant to our project and vital to the overall success of meeting our clients specifications.

### 3.3 AppleDouble directories

Benchmarking was impeded due to the mysterious appearance of ".AppleDouble" directories in the email database we were attempting to perform benchmarks on. These directories were owned by root and not readable by normal users, which for some reason caused IM-DSSE to crash when we tried to get it to load the database. We eventually traced this to some interaction with the university SMB fileshare. The solution was to not browse the directory over the fileshare.

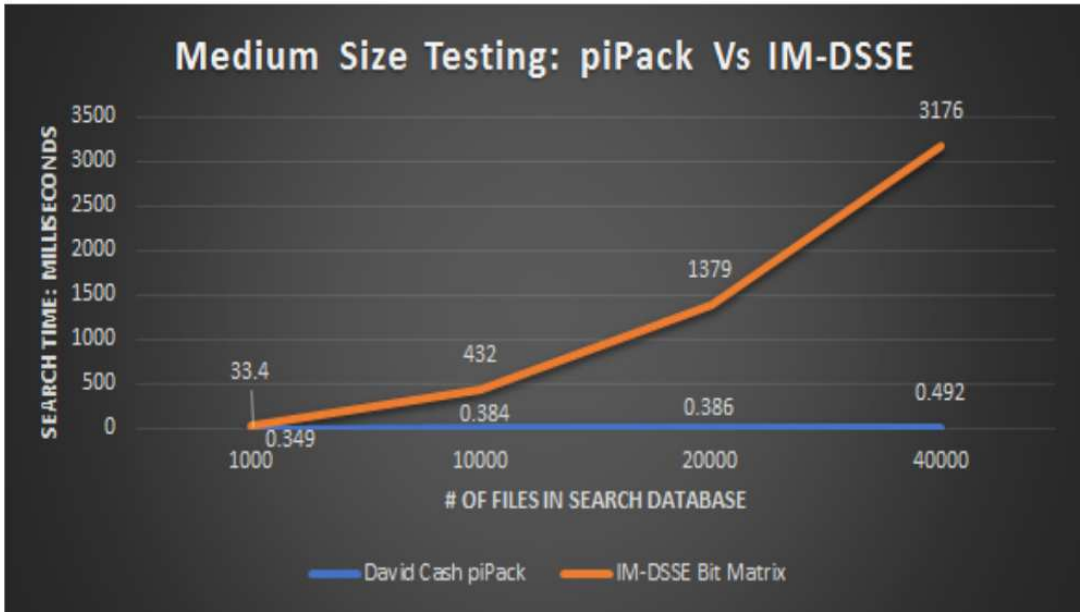
### 3.4 Disk space & Memory usage

Another problem that impeded benchmarking was a lack of disk space on the university servers. We wanted to try benchmarking very large data sets, but were limited to less than 6GB of file space. Additionally, trying to load a medium data set of around 1GB in size caused our implementation to run out of memory.

## 4 Results / Benchmarks

This section lists all the benchmarking graphs and discusses their important to the overal goal of the project.

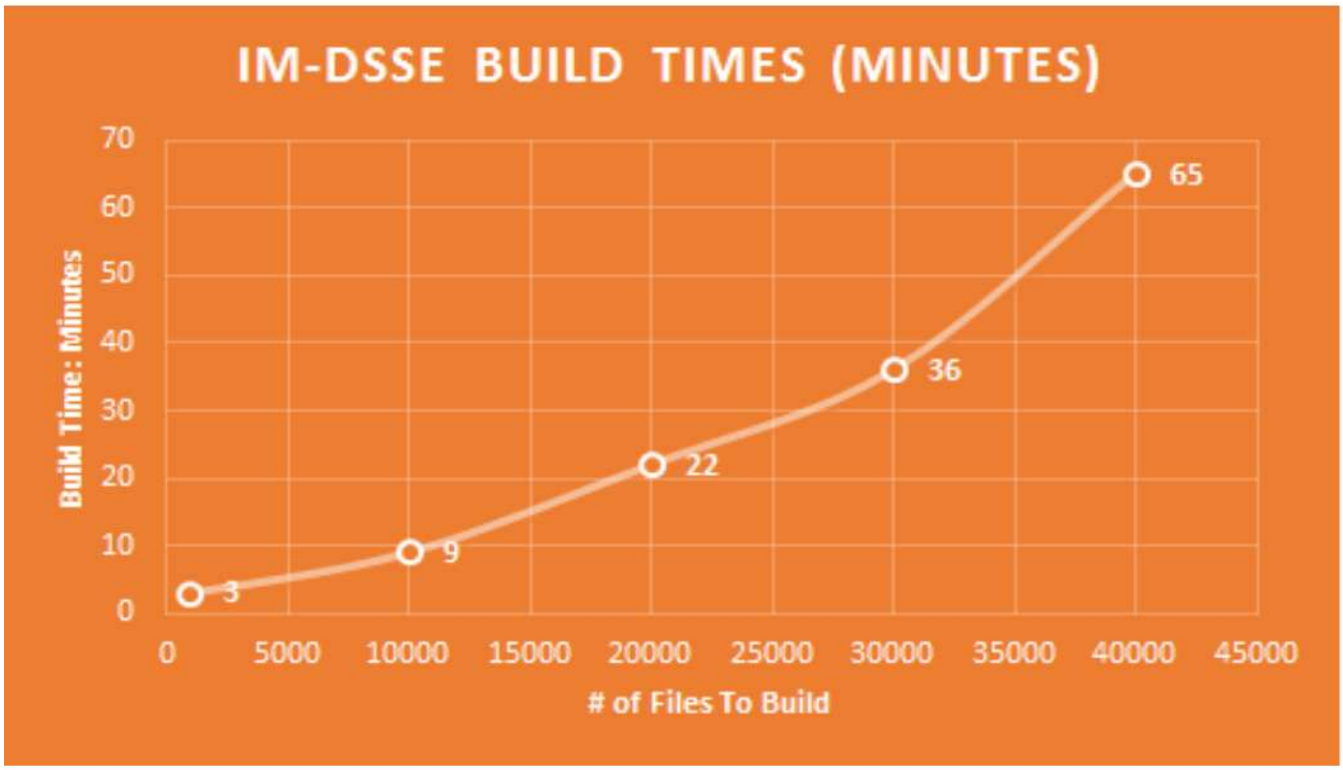
### 4.1 Medium Size File Benchmarking




The medium size file benchmarking compares the results of IM-DSSE, David Cash Basic and David Cash piPack optimizations to compare the runtime speed of a search call to a database of 1000-40,000 files. Total number of keywords range up to 250,000. As we can see it is very clear that both version of David Cash scheme vastly outperform that of the IM-DSSE Bit Matrix. This makes sense, as we observe that the time complexity for IM-DSSE is $\mathcal{O}(n^2)$ where the search complexity of David Cash

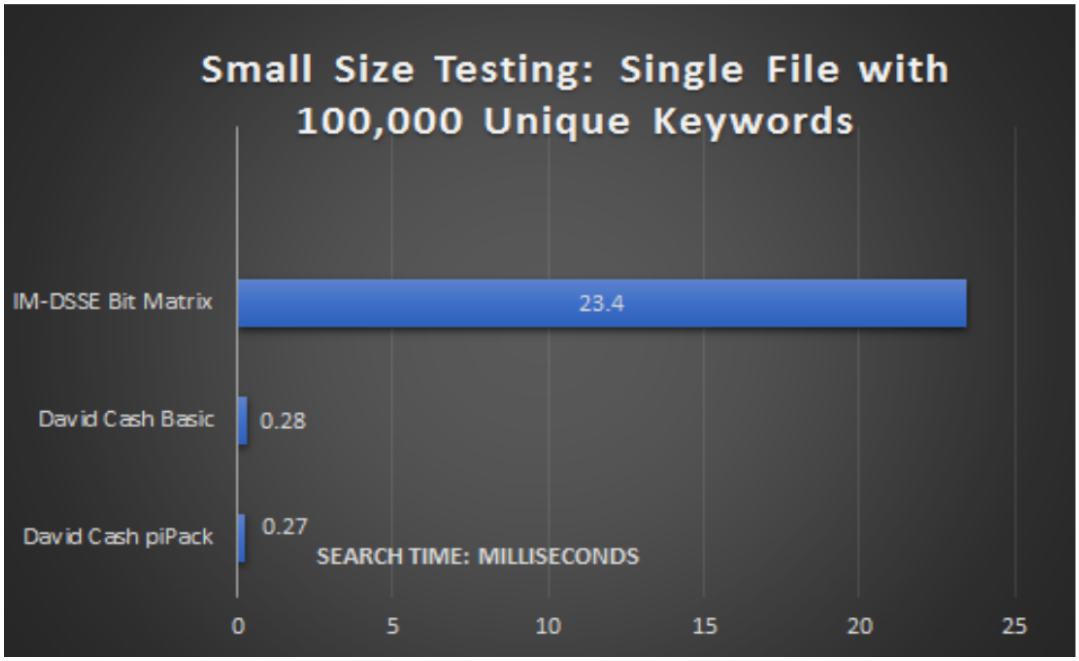
is closer to that of $\mathcal{O}(\log n)$. As file sizes increase the runspeed for IM-DSSE is substantially slower than David Cash. We will compare the basic to the optimization of piPack of David Cash in a later graph but we can clearly see here that David Cash has a vastly superior search speed, especially on larger data sets.

### 4.2 IM-DSSE Bit Matrix Build time



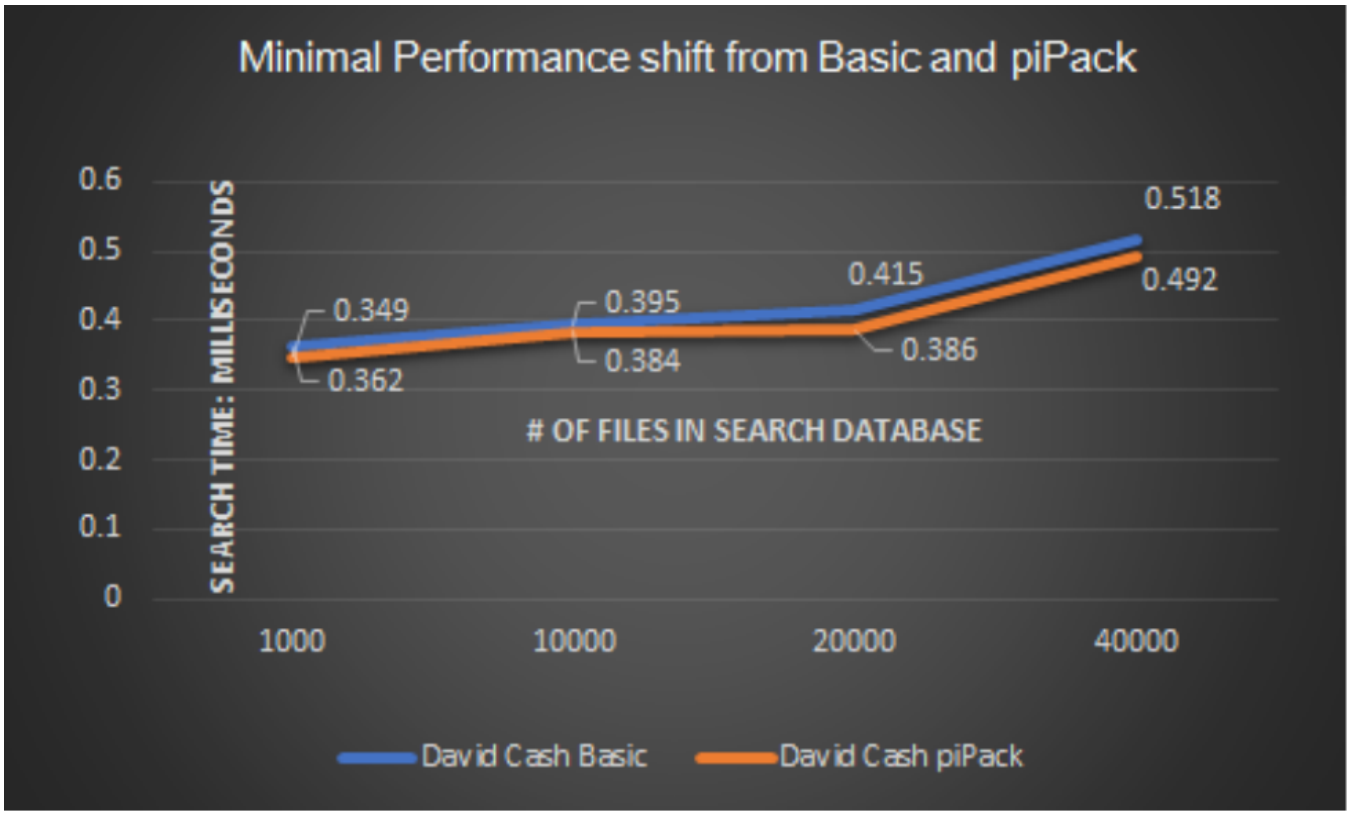

The build time of the IM-DSSE Bit matrix scheme is also $\mathcal{O}(n^2)$ as it needs to create a key file pair with every single file and keyword found in the dataset. This takes a long time, and as a result our testing showed that the quadratic graph confirms that build speed. We did not graph it here but we also averaged out a build time, on the largest dataset of 40,000 files, to be around 2-3 minutes; thus vastly outperforming IM-DSSE at over an hour in build time.

### 4.3 Small Data Single file benchmarking




Similar to the large data set we also see that IM-DSSE is outperformed by David Cash substantially. The overhead of creating this key-file 2d matrix pair is simply much more expensive than the algorithm used by David Cash. The optimization between piPack and Basic is unsubstantial here as we are only working on a single file.

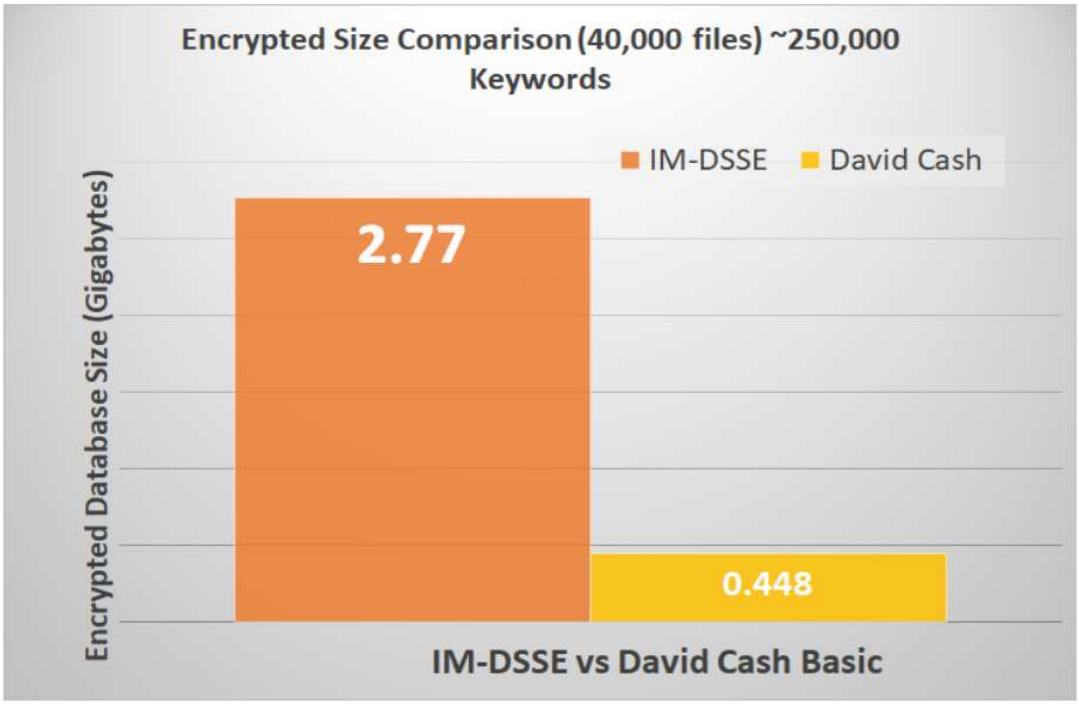
### 4.4 DSSE Basic vs DSSE piPack Optimization Graph




From our results we believe that this graph is the most interesting to talk about. It is clear that there is a slight optimization between the piPack scheme over the basic. However it was not as large as we were expecting. We believe

this is because the piPack scheme would better optimize disk access and data storage retrieval. However, for our testing purposes all the data is ran locally, therefore is on ram. Thus the piPack scheme does not provide the

### 4.5 Encrypted Size Comparison DSSE Basic VS IM-DSSE




The encrypted size comparison shows just how much larger the Encrypted Index is that is being searched through between cash and IM-DSSE. In addition to having a slower search time IM-DSSE indirectly composes its own larger dataset to search through because of its implementation of key-file pairs. Thus the size of the datasets are different. I expect this size to become even more drastic were we to test on a large database set such as the Wikipedia text data set.

## 5 Future work

All that is left to do this term as part of our project requirements in Pointer Optimization.

- Final steps towards pointer optimization implementation.
- Benchmark pointer against IM-DSSE and other Cash algorithms.

### 5.1 Looking back and looking ahead:

Overall, I feel that this term has flown bye. From week to week the time till expo has been rapidly approaching. Luckily for our team we have been on top of our requirements and have given ourselves enough time to continue working on final optimizations while obtaining benchmarking results. We had a shaky start to the term missing our direct contact with client and TA. Luckily we are still feeling very good about our position and are excited as a team to be ready for expo.

Practicing pitches to different audiences, revising the poster within regulation, exploring ways of demoing our research project to appease a general audience. This project has been successful this term and our group hopes to finish strong to be able to deliver the product that our client expects of us.

### 5.1.1 Future work

For the remainder of the term our group work time together will specifically work on finalizing pointer optimization and preparing for expo. This will include testing our pitch ideas, from a general audience to a technical one having the right verbiage for our audience is very important. Benchmarking Pointer against DSSE Basic and piPack are the second part of finalization with pointer optimization. As the term comes to a close and expo is finished we will be mostly working on compiling our final report, which includes our final term report at the end of the year with feedback from teammates, TA, and client.

## 6 RETROSPECTIVE / CONCLUSION

| What Went Well? | What Didn't Go Well? | What Can We Change? |
|---|---|---|
| Finished piPack Implementation | Was unable to load Large data set, Client agreed not possible with current server capabilities so we scrapped large testing | Finish the year strong with working towards expo presentation and final report. |
| Benchmarking on Small/Medium datasets against IM-DSSE, Basic and piPack completed | Delayed meeting with TA until week 4 | Keep communication up with TA to discuss specifics about Expo and Final Report requirements. |

Last term consisted mainly of implementation of DSSE. As we headed into this term the tasks for the project were separated into sections as an attempt to parallelize the work being done. Our primary goal for this term has been small and medium file size benchmarking between all our algorithm implementations. We ended up not having time to access our stretch goals of Email and Cloud integration and instead are focusing on expo preparation. This is fine as they were part of our stretch goals but we do wish we have another month or so to be able to get to the stretch goals. Aside from these changes the project is moving along and the group, as a whole seem to be in a good place. We are looking forward to expo and compile all our documentation into the final paper at the end of the term.