Research Assessment Exercise 2020
Impact Case Study

University: City University of Hong Kong
Unit of Assessment (UoA): 13 - computer studies/science (incl. information technology)
Title of case study: I/O Stack Optimization from Mobile Devices to Servers

1. Summary of the impact

I/O stack includes all layers between memory and physical storage, from page cache, virtual file system, to block I/O layer. I/O access latency is critical in computing systems, from mobile devices to servers. This impact case presents the work in the department on I/O stack optimization. Dr. Hong Xu designed a new graph processing system which minimized the overall disk I/O access for random walk algorithms. It is deployed in several systems of Tencent and supports hundreds of millions of users. Dr. Xue's cross-layer mobile storage optimization is used in all of Huawei's mobile production lines, contributing to their market-share success.

2. Underpinning research

Dr. Hong Xu's research focuses on computer networking and systems, particularly data analytics and machine learning systems as well as data center networking. Starting from 2017, Dr. Xu's group has been working on various resource scheduling problems in modern data analytics systems such as Apache Spark.

In one thread, Dr. Xu's team empirically find that in running an application's entire analytics pipeline, the Spark executors naturally exhibit salient time-varying resource usages: for example, an application may use much CPU initially and then start to bottleneck on network bandwidth in a later stage [1]. By exploiting this workload characteristic, they designed Elasecutor [1], a novel executor scheduler that dynamically allocates and explicitly sizes resources to executors over time according to the predicted time-varying resource demands. Elasecutor is open source on GitHub at https://github.com/NetX-lab/Elasecutor. In another thread, Dr. Xu's team also study geo-distributed analytics systems where an analytics query or job is executed across data centers in different parts of the world with their own datasets. Here a unique characteristic of the workload is the similarity between input data in different data centers [2]. They exploit this characteristic to design a new scheduler called Bohr that moves input data that are "similar" to those in the destination site, which can significantly reduce the amount of intermediate data that needs to be shuffled across the wide area Internet.

Both projects lead to publications at top international conferences: Elasecutor appears in ACM SoCC, the top conference in cloud computing organized by ACM SIGMOD and SIGOPS, and Bohr appears in ACM CoNEXT, a top conference organized by ACM SIGCOMM on computer networking. More importantly, their approach of exploiting the unique characteristics of workloads in systems design, together with the experience in data analytics systems, built a solid foundation for their collaboration with Tencent. They started to work closely with Tencent on graph processing systems, a topic of significant interest to their production teams.

Dr. Jason Xue's research focuses on computer storage and systems, particularly on cross layer optimisation and data reliability. In particular, Dr. Xue's team worked on revealing the root causes for user perceived latency for Android based mobile devices at that time. Contrary to the common belief, Dr. Xue finds that fragmentation is still a significant problem on mobile storages, which all use Flash memories [3][4][5]. Fragmentation was a known problem for hard-disks, and it was not

considered as a problem for Flash memory which has decent random access performance. Dr. Xue's team was the first to show that, while Flash memory has an advantage in random access performance, fragmentation which causes longer I/O queueing time and increases cache misses for mobile flash storage is still significant for mobile devices. Full analysis across multiple layers have been investigated with a couple of remedies presented as a result of Dr. Xue's work. The solutions in fixing the fragmentation problem have shown the effectiveness in reducing user perceived latency for Android based mobile devices after a period of usage.

3. References to the research

[1] Libin Liu, Hong Xu. "Elasecutor: Elastic Executor Scheduling in Data Analytics Systems." Proc. ACM Symposium on Cloud Computing (SoCC), 2018.

[2] Hangyu Li, Hong Xu, Sarana Nutanong. "Bohr: Similarity Aware Geo-Distributed Data Analytics." Proc. ACM International Conference on emerging Networking EXperiments and Technologies (CoNEXT), 2018.

[3] Cheng Ji, Li-Pin Chang, Liang Shi, Chao Wu, Qiao Li, Chun Jason Xue: An Empirical Study of File-System Fragmentation in Mobile Storage Systems. HotStorage 2016

[4] Sangwook Shane Hahn, Sungjin Lee, Cheng Ji, Li-Pin Chang, Inhyuk Yee, Liang Shi, Chun Jason Xue, Jihong Kim: Improving File System Performance of Mobile Storage Systems Using a Decoupled Defragmenter. USENIX Annual Technical Conference 2017: 759-771

[5] Cheng Ji, Li-Pin Chang, Sangwook Shane Hahn, Sungjin Lee, Riwei Pan, Liang Shi, Jihong Kim, Chun Jason Xue: File Fragmentation in Mobile Devices: Measurement, Evaluation, and Treatment. IEEE Trans. Mob. Comput. 18(9): 2062-2076 (2019)

4. Details of the impact

Context

Graph processing system is one of the fundamental infrastructures for processing data analytics workloads, simply because a lot of data are organized and analysed as graphs. Using Tencent, our collaborator as an example, data about users and user activities from their online social platforms (WeChat, QQ, etc.) are stored as large social graphs and are used for many high-value tasks such as anti-fraud and anti-money laundering operations. These analysis tasks are often based upon variants of random walk algorithms on the social graphs. Since the scale of the graph is extremely large, existing systems cannot handle the random walk algorithms well. One group in Tencent initially used a Spark-based GraphX cluster to execute random walk based algorithms of a graph with over millions of vertices and billions of edges, and the running time exceeds 3 days. Another group used a graph database cluster (Neo4j) which cannot even finish executing their random walk based algorithms. Based on their research on large-scale data analytics systems, Dr. Xu's team developed a new out-of-core graph processing system called RWGraph. RWGraph is optimized for running random walk algorithms over hyper-scale social graphs with billions of edges on just a single machine. RWGraph uses a novel graph representation and a separate value storage to store the graph, so that the overall I/O access is minimized for random walk algorithms. RWGraph also exposes new stream processing APIs for developing different random walk algorithms on its backend.

Smartphones and mobile devices have changed the computing paradigm in the past decade. Smartphones are no longer only used to make phone calls, they are now the de facto computing platform for daily usage. While this is happening in the world, the computing systems supporting the smartphones and mobile devices need to catch up. Early generations of Android phones suffer from user perceived latency after a certain period of usage. Overcoming user perceived latency after extended usage was one of the top challenges of Huawei's mobile division. The research by Dr. Xue's team conducts systematic cross-layer optimizations for Android based mobile systems. Specifically, for high-end product series of Huawei's mobile devices, Dr. Xue's team developed customized solution to maintain superior performance after years of usage time. For low-end product series of Huawei's mobile devices, Dr. Xue's team developed transparent file system level compression to improve storage utilization as well as storage life time.

Reach and Significance of Impact

Since 2018, RWGraph has been deployed in production at Tencent and serving a variety of workloads from both internal groups and external clients. Internally, it is used for the LingKun financial fraud detection system to analyze and detect fraud activities and fraudulent user clusters. RWGraph is able to run random walk based algorithms on a graph with over 100 million users within 1-2 hours. The estimated annual revenue of the LingKun system is over $10 million RMB. The second internal system RWGraph is deployed for TenPay, Tencent's online payment processing platform supporting WeChat Pay with over 800 million users worldwide. RWGraph runs their capital chain analysis queries over the user transactions. The third internal system RWGraph has impact on is Anquan Guanjia, Tencent's computer security and protection solution for mobile and desktop environments. RWGraph is used to run their social graph analytics in the backend with over 200 million users in the graph, which was not possible before with the previously deployed Neo4j graph databases.

Externally, RWGraph has also reached various clients of Tencent and helped support their graph processing needs. It has served China Construction Bank at Shenzhen for many graph analytics tasks such as financial transactions analysis, criminal gang detection, and anti-money laundering. Each task, depending on the specifics of the scenario and requirements, is estimated to generate $1-$3 million RMB revenue annually for Tencent. Some government agencies in China are also in talks with Tencent as of September 2019 to use RWGraph along with other solutions to detect criminal gangs using social graph analysis.

Since 2015, several research findings of Dr. Xue's team are used in Huawei's mobile production line and contribute to the recent success of Huawei's smartphones. The first is the study of fragmentation on mobile flash storage and how to remove fragmentation effectively and efficiently. The second is to optimize page cache mechanism for android base systems for improved performance. The third is re-designing memory allocation for Android on mobile devices. These findings are now applied in Huawei's full production line, while Huawei's smartphone now has the highest market share in China, as well as in Europe. Eliminating user perceived latencies and improving user experience are instrumental in achieving this status.

5. Sources to corroborate the impact

[A] Testimonial from Tencent on the details of RWGraph's impact to their business

[B] Testimonial from Huawei on the details of Dr. Xue's research impact to their business