An Intelligent Software-Defined Storage

(A very efficient Big Data processing)

By: Maryam Yammahi, Adi Alhudhaif, Simon Berkovich

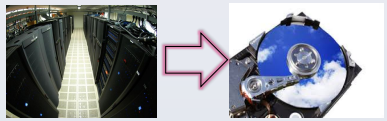
School of Engineering & Applied Science
THE GEORGE WASHINGTON UNIVERSITY

Imagine with the huge availability of the data in the internet, an intelligent Storage that can store, access, retrieve and processes amassment of data items (beyond terabytes)!!!

Imagine a new intelligent Software-defined Storage that can efficiently process Big Data from several order of magnitude including: high speed, accuracy, less resources!!!!

Abstract

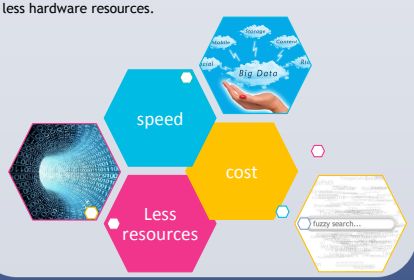
The invention presents a special repository for diversity of information items as a practical realization of Big Data systems. It is a very large Software-Defined Storage for the construction of intelligent system. The suggested system emulates the basic features of a suggested memory organization of the brain that based on a new type of computational model for processing Big Data [1]. This new Software-Defined Storage will enable the access to very large data of diversified files. Thus it enhances speed and efficiency to the storage for various data. It uses different and unique operational techniques to allow efficient access to the stored data items in the storage.



Objective

The objective of this research is to develop an innovative Software-Defined Storage, in which it can:

- process terabytes of files (Big Data),
- provides an efficient access speed.
- requires less resources since the algorithms used with it require less memory space.
- It will also provide a huge cost reduction, since it will require less hardware resources.



Methodology

The invention uses new operational techniques to implement Software Defined Storage (SDS). SDS refers to systems in the storage infrastructure (hardware) is managed and automated by intelligent software, as opposed to by the storage hardware itself. The overall scheme is shown in the diagram of Fig.1.

The invention combines two operational techniques for accessing the data in very large (more than terabytes) systems;

1) Approximate search using a novel technique called "Pigeonhole search"[2], which uses to ensure the feasibility of processing terabytes of data. "Pigeonhole access" provide Content-addressable access that is arranged by inverted files for each type of attribute.

2) Resolution of multiple responses performed by a Novel principle for the stream extraction of data [3][4], that achieves selection of the most appropriate (frequent) items from the output of the approximate search.

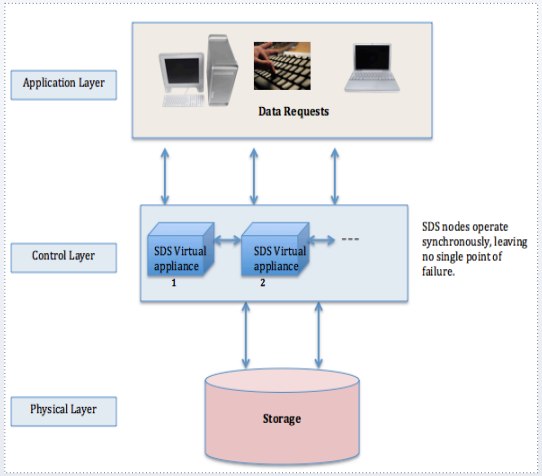


Fig. 1 : The general structure of the suggested SDS

Results

The combined techniques improve speed enormously while still achieving acceptable levels of accuracy. Specifically, the approximate search can achieve "speed-up" ratios of 500x - 6000x over baseline naïve search technique; and the stream extraction can be tailored to recover from 10>95% recovery of predominant element(s). By accessing multi-attribute items using any combination of attributes, it produces a vast amount of data to be processed in stream processing fashion to select the most appropriate items based on their occurrence frequencies.

#1: " Pigeonhole access " (The central part of the SDS).

TABLE I
TIME RESULTS AND SPEEDUP BETWEEN PIGEONHOLE SEARCH AND SEQUENTIAL SEARCH ALGORITHMS

N	Sequential Search	Pigeonhole search	Speedup
10^3	195 ns	112 ns	2
10^4	1,455 ns	96 ns	15
10^5	13,785 ns	96 ns	144
10^6	138,857 ns	460 ns	347
10^7	1,338,165 ns	2240 ns	597

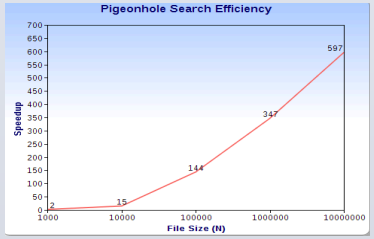
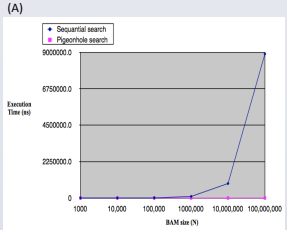


Fig.2: Pigeonhole search efficiency compared to the sequential search in a 92-bit vector



The time taken by both Pigeonhole search and Sequential search

The rate that Pigeonhole search is faster than Sequential search

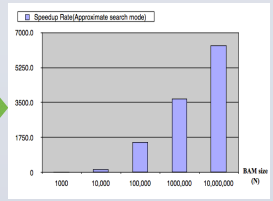


Fig.3: (A):shows the the time taken by Pigeonhole access is much less than sequential access. (B) shows the speed up rate, which is faster by more than 6000.

Results (cont.)

#2: Stream processing for resolution of the multiple responses.

Stream processing can be applied using one of the two approaches;

1- Cyber-physical stream (CPS) is used to extract the most frequent item (Single item) Fig. 4.

2- Multi-Buffer Based Algorithm to extract the most frequent k items and their frequencies in single process (Multiple items) Fig. 5.

(A) (B)

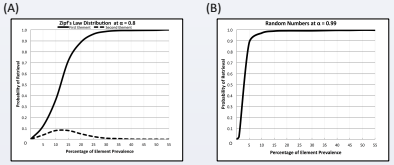


Fig.4. (A): Probabilities of retrieving the prevalent item. (B): Probabilities of retrieving first and second most frequent item.

(A) (B)

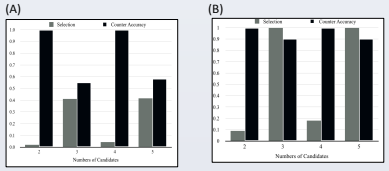


Fig.5. (A): Probabilities of retrieving the prevalent items and their frequencies accuracy for an item of a frequency of 1%. (B): Probabilities of retrieving the prevalent items and their frequencies accuracy for an item of a frequency of 5%

Conclusions & applications

This research contributes to the area of processing large data files (Big Data). The suggested Software-Defined Storage will enable the access to very large data of diversified files. Thus it enhances speed and efficiency to the storage for various data. This invention, could get numerous applications in Big Data intelligent systems.

Status and future plan

Patent Status: US Provisional Application filed.

Academic Recognition: Two journal publications.

Development phase: we are in implementation phase of combining the two algorithms into one storage.

Next Step: Building a prototype for the device.

References

[1] S. Berkovich, ""Organization of the Brain in Light of the Big Data Philosophy", COM. Big Data' 14 Proc. of the 1st International Summits on Big Data Computing, IEEE, Washington DC, 2014.

[2] M. Yammahi, K. Kowsari, C. Shen and S. Berkovich, "An efficient technique for searching very large files with fuzzy criteria using the Pigeonhole Principle", COM. BigData' 14 Proc. of the 1st International Summits on Big Data Computing, IEEE, Washington DC, 2014.

[3] A. Alhudhaif, M. Yammahi, T.Yan and S. Berkovich. Article: A Cyber-Physical Stream Algorithm for Intelligent Software Defined Storage. *International Journal of Computer Applications* 109(5):21-25, January 2015

[4] Alhudhaif, Adi, Tong Yan, and Simon Berkovich. "On the organization of cluster voting with massive distributed streams." In *Computing for Geospatial Research and Application (COM. Geo), 2014 Fifth International Conference on*, pp. 55-62. IEEE, 2014.