An Intelligent Software-Defined Storage
(A very efficient Big Data processing)
By: Maryam Yamnahi, Adi Alhudaif, Simon Berkovich

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 simulates 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.

Methodology
The invention uses new operational techniques to implement Software Defined Storage (SDS). SDS refers to systems in the storage infrastructure (hardware) that 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.

1) Approximate search using a novel technique called “Pigeonhole search”[2], which uses to ensure the feasibility of processing terabytes of data. “Pigeonhole access” provides 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.

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).

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.

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.

Results (cont.)
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 simple process (Multiple items) Fig. 5.

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

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

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

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%