

## Towards the Integration of MEMS-based Storage in Computing Systems

a report by

**Dr Hailing Yu, Professor Divyakant Agrawal and Professor Amr El Abbadi**

Computer Science Department, University of California, Santa Barbara

### Introduction

Although magnetic disks have persistently dominated storage technology, recently, memory hierarchy has suffered from problems of latency, bandwidth and cost gap. In particular, due to advances in processor technology and semiconductor manufacturing, the processor-to-disk performance gap has been growing consistently. Currently, this gap has widened to six orders of magnitude, and future trends indicate that unless a breakthrough occurs in disk technology, this gap will continue to widen by about 50% annually.

Due to advances in semiconductor manufacturing, microelectromechanical system (MEMS)-based storage systems<sup>1-3</sup> are being developed as an alternative to conventional disks. IBM's 'Millipede' Project<sup>4</sup> promises to deliver by 2005 a postage-stamp-sized memory card that can hold several gigabytes of fast, non-volatile memory. This project, according to the MEMS designers, only scratches the surface, as the digital bits of future-generation MEMS-based storage devices will continue to shrink until they are individual molecules or even atoms. As material scientists and mechanical engineers work feverishly to develop more efficient MEMS-based storage devices, the role of computer science research is to integrate such devices effectively into computer systems for different applications.

The challenges are major and intriguing due to the fact that MEMS-based storage devices have very different characteristics from disks: the ability to

move in both x and y dimensions and thousands of concurrent tips that can be activated simultaneously to access data, etc. Thus, from the operating system point of view, the algorithms for input/output (I/O) scheduling, data layout and failure management techniques designed for traditional disks need to be revisited. Some initial steps have been taken by various research groups, notably the Carnegie Mellon University's (CMU's) Center for Highly Integrated Information Processing and Storage Systems (CHIPS) Project, which has explored various operating systems issues such as request scheduling, data placement and others by mapping MEMS-based storage into disk-like devices.<sup>5</sup>

Even though the results showed that stand-alone MEMS-based storage devices improve the overall application run time by several folds as a result of this mapping, MEMS-based storage devices lose the two-dimensional property. Thus, further performance gains from this property are lost, especially for data-intensive applications such as database management systems (DBMS).

Building on the preliminary models of MEMS-based storage devices that have emerged,<sup>1,2</sup> several areas have been investigated where considerable potential exists for improving the limits of performance of such systems. Based on the characteristics of MEMS-based storage, the new I/O scheduling being proposed guarantees to perform within twice the optimal time for any workload. The data placement scheme being developed for relational DBMS can significantly improve the system performance.

Dr Hailing Yu is pursuing her PhD at the Computer Science Department of the University of California, Santa Barbara. Her research interest is integrating microelectromechanical system-based storage into computation systems.

Professor Divyakant Agrawal has been on the faculty of the Department of Computer Science at the University of California, Santa Barbara, since 1987. His research interests are in the areas of distributed systems and databases. Professor Agrawal received a BE (Hons) in Electrical Engineering from the Birla Institute of Technology and Science (India) in 1980, and an MS and PhD in Computer Science from the State University of New York at Stony Brook in 1984 and 1987, respectively.

Professor Amr El Abbadi joined the Department of Computer Science at the University of California, Santa Barbara in August 1987. His research interests are in the areas of distributed algorithms, multidimensional and spatial databases and large-scale information systems. Professor El Abbadi received his BEng from Alexandria University (Egypt) in 1980 and his PhD in Computer Science from Cornell University in 1987.

1. L Richard Carley, Gregory R Ganger and David F Nagle, "MEMS-based integrated-circuit mass-storage systems", Communication of the ACM, 43(11) (November 2000), <http://www.lcs.ece.cmu.edu/research/MEMS>
2. P Vettider, M Despont, U Durig, W Haberle, M I Lutwyche, H E Rothuizen, R Stuz, R Widmer and G K Binnig, "The 'millipede' – more than one thousand tips for future AFM storage", IBM Journal of Research and Development, 44(3) (May 2000), pp. 323–340.
3. Tara Madhyastha and Katherine Pu Yang, "Physical modeling of probe-based storage", Proceedings of the 18th IEEE Symposium on Mass Storage Systems and Technologies (April 2001), pp. 207–224.
4. P Vettiger and G Binnig, "The nanodrive project", Scientific American (January 2003), pp. 47–53.
5. J Griffin, S Schlosser, G Ganger and D Nagle, "Operating systems management of MEMS-based storage devices", Symposium on Operating Systems Design and Implementation (OSDI) (October 2000), <http://www.lcs.ece.cmu.edu/research/MEMS>

### An Architecture for MEMS-based Storage

MEMS are extremely small mechanical structures formed by the integration of mechanical elements, actuators, electronics and sensors. These are fabricated on silicon chips using photolithographic processes similar to those employed in manufacturing standard semiconductor devices. As a result, MEMS-based storage can be manufactured at a very low cost. MEMS represent a compromise between slow traditional disks and expensive storage based on electrically erasable programmable read-only memory (EEPROM) technologies. Unlike traditional disks, MEMS-based storage devices do not make use of rotating platters due to the difficulty of manufacturing efficient and reliable rotating parts in silicon. The emerging paradigm for such systems is that of a large-scale MEMS array that, like disk drives, has read/write heads and a recording media surface. The read/write heads are probe tips mounted on microcantilevers embedded in a semiconductor wafer and arranged in a rectangular fashion. The recording media is another rectangular silicon wafer (called the media sled) that can use conventional techniques for recording data. This structure gives MEMS-based storage two-dimensional characteristics.

MEMS-based storage devices are composed of tens to thousands of recording heads and a recording media surface. The recording heads (probe tips) form a two-dimensional array and are fabricated on silicon chips. The recording media (the media sled) is spring-mounted above the probe tips' array and can move in x and y dimensions. There are several different approaches for recording data. For example, IBM's Millipede uses pits in the polymers made by tip heating, while CMU CHIPS adopts the same technique as data recording on a magnetic surface.

Following is an overview of the CMU CHIPS Project, which is dedicated to next-generation MEMS-based storage devices. To access data under this model, the media sled is moved from its current position to a specific position defined by (x, y) co-ordinates. After the 'seek' is performed, the media sled moves in the y direction while the probe tips access the data. The media sled also moves in the z direction to actuate the distance between the probe tips and the media sled. The x and y actuators provide the force for moving the media sled in the x and y directions, while the spring provides the restoring motion. These two actuators work independently.

The media sled is divided into rectangular regions. Each of these rectangular regions contains an array of  $m \times n$  bits and is serviced by one probe tip. The relation between the regions and tips is a one-to-one mapping, i.e. the number of regions is the same as the number of probe tips. In theory, all the probe tips can be activated simultaneously. For the CMU CHIPS device, the system has 6,400 tips, arranged in an array of 80 x 80 tips per rectangular region with each region having 2,500 x 2,500 ( $m \times n$ ) bits. Due to power and heat constraints, only 1,280 tips can be activated simultaneously.

### The Minimum Spanning Tree-based I/O Schedule

MEMS-based storage devices are two-dimensional, while disks have traditionally always been modelled as one-dimensional storage devices by organising them in terms of cylinders, sectors and tracks. Even though existing disk scheduling algorithms such as first-come, first-service (FCFS), cyclical look (CLOOK), shortest seek time first (SSTF) and shortest position time first (SPTF) can be adapted to MEMS-based storage devices, some characteristics of MEMS-based storage devices have not been considered adequately. It has been established that optimal scheduling for MEMS-based storage is NP-complete.<sup>6</sup> Even though it is not practical to design an optimal algorithm, offline and online scheduling algorithms have been proposed with guaranteed upper bound for any workload.

MEMS-based storage devices are designed to move in both x and y dimensions independently. Based on this property, the distance between two points is measured using the infinity distance measure. In the proposed algorithm, several properties of a minimum spanning tree are first developed in the infinity distance space and used to reduce the construction cost of the minimum spanning tree. Then, a minimum spanning tree in the infinity distance space is built for all the requests, where requests' positions are treated as vertices and edge costs are seek time between two vertices. Finally, the requests are served in the double-walk order of the minimum spanning tree. Theoretical analysis shows that these algorithms are guaranteed to perform within twice the optimal performance time. ■

### Additional Information

*This article is continued, with a graphic and further discussion, in the Reference Section on the CD-ROM accompanying this business briefing.*

6. The complexity class of decision problems for which answers can be checked for correctness, given a certificate, by an algorithm, the run time of which is polynomial in the size of the input (i.e. it is NP) and no other NP problem is more than a polynomial factor harder. Informally, a problem is NP-complete if answers can be verified quickly and a quick algorithm to solve this problem can be used to solve all other NP problems quickly.