Solid-State Electronics 102 (2014) 2–11

Contents lists available at ScienceDirect

Solid-State Electronics

journal homepage: www.elsevier.com/locate/sse

Emerging memories

Livio Baldi [a,*], Roberto Bez [a], Gurtej Sandhu [b]

[a] Micron Semiconductor Italia, via C. Olivetti, 2, 20864 Agrate Brianza, Italy
[b] Micron Technology Inc., 8000 S. Federal Way PO Box 6, Boise, ID 83707-0006, USA

ARTICLE INFO

Article history:
Available online 22 July 2014

The review of this paper was arranged by Prof. S. Cristoloveanu

Keywords:
Memory
Flash
DRAM
Mass storage
Emerging memories

ABSTRACT

Memory is a key component of any data processing system. Following the classical Turing machine approach, memories hold both the data to be processed and the rules for processing them. In the history of microelectronics, the distinction has been rather between working memory, which is exemplified by DRAM, and storage memory, exemplified by NAND. These two types of memory devices now represent 90% of all memory market and 25% of the total semiconductor market, and have been the technology drivers in the last decades. Even if radically different in characteristics, they are however based on the same storage mechanism: charge storage, and this mechanism seems to be near to reaching its physical limits. The search for new alternative memory approaches, based on more scalable mechanisms, has therefore gained new momentum. The status of incumbent memory technologies and their scaling limitations will be discussed. Emerging memory technologies will be analyzed, starting from the ones that are already present for niche applications, and which are getting new attention, thanks to recent technology breakthroughs. Maturity level, physical limitations and potential for scaling will be compared to existing memories. At the end the possible future composition of memory systems will be discussed.

1. Introduction

Memory has been from the beginning one of the key components of any computing system, from the perforated tape and the hard-wired set of rules of the Turing machine, to the arrays of magnetic nuclei and magnetic hard disks of early computers. It is not exaggerated to say that it was the availability of the first solid state DRAM memories, together with compact magnetic hard disk drives, that made possible the introduction of the personal computer, while the development of the first solid state nonvolatile memories, first EPROM and afterward NOR Flash, made possible the widespread use of microcontrollers in a variety of applications. But it was with the introduction of the low cost NAND Flash that a drastic change took place in the utilization of memories: from data and program storage in computer to general information storage for all kind of consumer application, from music to pictures and movies. As a consequence, a radical change took place in the memory market: while the amount of memory required for computer programs is increasing at a limited rate, anyhow controlled by the complexity of software development, there seems to be no end to the demand of memory for information storage, with increasing demand in terms of resolution and quality. In a sort of chain reaction, the availability of low cost memory has driven the shift from analogue to digital data storage, and the growing market has fuelled a continuous growth of the technology. It is not the lack of demand that could break the circle, but some doubts start appearing about the continuation of the technology development at the same pace.

2. The role of memory

Memory is a quite generic term, and in general it can have many different meanings. In physiology we can distinguish between short term and long term memory, and the mechanism that regulates the transfer of information from one to the other is not fully clear, but other type of memories can be distinguished as well: procedural memory, semantic memory, episodic memory, and so on, depending on the approach used.

In computing systems, the original model of the Turing machine included one alterable memory, the infinite tape on which data were punched, and a set of rules hard wired in the machine, while the von Neumann architecture, used in computers, introduced a distinction between main memory, which stores instruction to be executed and data being processed, and mass storage unit, which stores both instruction/programs and data not needed in the short term. The first distinction is functional while the second is

* Corresponding author. Tel.: +39 0396375015.
E-mail addresses: lbaldi@micron.com (L. Baldi), rbez@micron.com (R. Bez), gsandhu@micron.com (G. Sandhu).

http://dx.doi.org/10.1016/j.sse.2014.06.009
0038-1101/

essentially technological and related to the large cost of fast access memory. This short introduction shows that there are several types of memory, and that a large variety of performance parameters can be used to characterize them. In general computer architectures include different levels of fast access cache memory (SRAM or DRAM) for the data being processed and the instructions of the program being run and a long term, slower access mass-storage memory, where programs and data not in use at the moment are stored. The trend towards mobile computing is shifting mass-storage memory technology from magnetic hard disks to solid-state solutions, while the convergence among consumer, communication and data processing applications has lead to an explosive growth in mass-storage requirements. In the end the choice of one specific memory technology depends on the right mix of several parameters, and this aspect must always be taken in consideration when assessing emerging memories.

2.1. The memory landscape

At present memories represent about 21% of total semiconductor market [1]. Although a large variety of memory types is available, the market is dominated by two memory types: DRAM and NAND Flash that together make up 89% of the memory market (Fig. 1).

It is interesting to note two facts: these memories have radically different performances, so as to be considered as exemplification of the concepts of Volatile and NonVolatile memory: DRAM content is byte-alterable and programming takes place in tens of nanoseconds, but data retention is in the order of hundred of milliseconds, even in the presence of a power supply, while NAND require milliseconds for programming, and the content can be changed only by very large blocks, but data can be retained for years, even in the absence of a power supply. It is also worth noticing that both do not show the best performances in their respective categories: SRAM is faster than DRAM in programming and can retain information without refreshing, while NOR has a better single byte programming speed and faster access time than NAND. However these two memory types have one thing in common: the very low cost, thanks to a small cell size, and the good scalability. Different as they are in performances, DRAM and NAND are based on the same information storage mechanism: charge storage. Both feature a capacitor in which a certain amount of charge can be stored and retained for a certain period of time; what is different is the mechanism that is used to store the charge. In DRAM the capacitor is a separate component and accessed through a MOS devices, which allows for fast programming, but does not guarantee a long data retention because of the leakage of the transistor. In Flash NAND the capacitor is an extra gate inserted in a MOS device and the charge is stored via a tunnelling mechanism. It requires high energy to pass the energy barrier formed by the gate oxide for programming (and therefore long programming times), but it guarantees a long data retention. Reading mechanism is also different: in DRAM it implies the detection of the residual charge stored in the capacitor, while in the Flash cell the stored charge directly controls the threshold voltage, and therefore the current of a transistor. These two reading mechanisms can be found also in all emerging memory devices.

2.2. Limits to scaling

The success of DRAM and NAND Flash has been largely connected to their small cell size and its scalability. NAND Flash, in which the capacitor is directly integrated in the transistor, has an advantage over DRAM, with a cell size of $4–4.5F^2$, where F is the minimum lithography size, while DRAM cell size has been reduced from over $10F^2$ to $8F^2$. Thanks to the smaller intrinsic cell size, in the last decade NAND Flash is now playing the role of technology driver, pushing the limits of lithography, as it can be seen in Fig. 2.

When speaking of density, NAND has a further advantage over DRAM since the threshold window between "1" and "0" levels and especially their stability over time allows storing more than one bit/cell. Devices with 2 bit/cell have been introduced already in the 90's [2], and more recently NAND memories with 3 bit/cell [3] have become usual especially for mass-storage applications. As a consequence NAND Flash memory has overshot Moore's law for density in the last decade with an impressive growth rate of $3\times$/year. The key question is: can this rate be maintained, or will the scaling of NAND and DRAM devices come to a sudden stop in the next years?

Two main factors can affect the further evolution of both memories: one technological and one physical. The technological factor is the availability of a cost effective lithography for sub-20 nm feature size, and it is common to all kind of devices and could affect also emerging memory technologies, the physical factor is more typical of the data storage mechanism of DRAM and NAND, and it is directly related to scaling. Since the data are stored as charge in capacitors, with the reduction of the physical size also the stored charge is bound to reduce, with a negative effect on data sensing and data retention capability. In DRAM the reduction of capacitor area has been compensated by going to 3-D capacitor integration first in trenches, and afterwards on top of the access transistor, and by introducing high-k dielectrics. However there are technological limitations to the aspect ratio of the cylindrical capacitors, and to the performance of high-k materials that can be easily integrated with CMOS process. For NAND devices, the main impact has been on the sensitivity to the electrostatic interference by the charge on neighboring cells, and on the enhanced sensitivity to dielectric leakage and read disturbs. These issues have been until now successfully solved with sophisticated design solutions, the introduction of Error Correction Codes and of Managed NAND (memory managed externally by a microcontroller) and by a

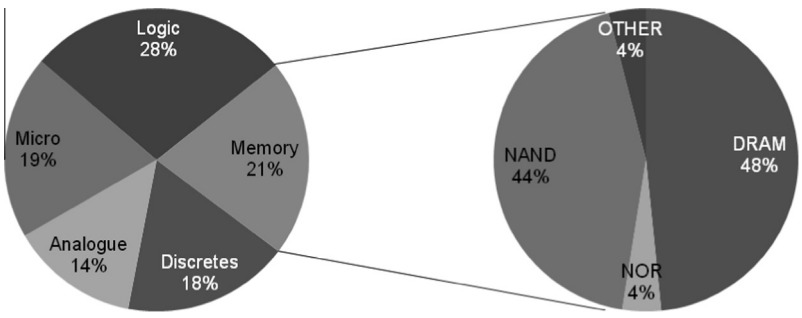

Fig. 1. Semiconductor and memory market (source: WSTS Q1 2013).