A stacked memory device on logic 3D technology for ultra-high-density data storage

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

2011 Nanotechnology 22 254006

(http://iopscience.iop.org/0957-4484/22/25/254006)

View the table of contents for this issue, or go to the journal homepage for more

Download details:
IP Address: 129.186.252.109
The article was downloaded on 18/10/2011 at 20:59

Please note that terms and conditions apply.
A stacked memory device on logic 3D technology for ultra-high-density data storage

Jiyoungh Kim\textsuperscript{1,2}, Augustin J Hong\textsuperscript{1}, Sung Min Kim\textsuperscript{1,2}, Kyeong-Sik Shin\textsuperscript{1}, Emil B Song\textsuperscript{1}, Yongha Hwang\textsuperscript{1}, Faxian Xiu\textsuperscript{1}, Kosmas Galatsis\textsuperscript{1}, Chi On Chui\textsuperscript{1}, Rob N Candler\textsuperscript{1}, Siyoung Choi\textsuperscript{2}, Joo-Tae Moon\textsuperscript{2} and Kang L Wang\textsuperscript{1}

\textsuperscript{1} Device Research Laboratory, Department of Electrical Engineering, University of California, Los Angeles, CA 90095, USA
\textsuperscript{2} Advanced Technology Development Team and Process Development Team, Memory R&D Center, Samsung Electronics Co. Ltd, Korea

E-mail: hbt100@ee.ucla.edu

Received 22 June 2010, in final form 29 September 2010
Published 16 May 2011
Online at stacks.iop.org/Nano/22/254006

Abstract

We have demonstrated, for the first time, a novel three-dimensional (3D) memory chip architecture of stacked-memory-devices-on-logic (SMOL) achieving up to 95% of cell-area efficiency by directly building up memory devices on top of front-end CMOS devices. In order to realize the SMOL, a unique 3D Flash memory device and vertical integration structure have been successfully developed. The SMOL architecture has great potential to achieve tera-bit level memory density by stacking memory devices vertically and maximizing cell-area efficiency. Furthermore, various emerging devices could replace the 3D memory device to develop new 3D chip architectures.

Some figures in this article are in colour only in the electronic version

1. Introduction

Over the last few decades, the market for mobile and wireless electronic gadgets has increased tremendously. The availability of low-cost and high-density memory chips has enabled a variety of memory intensive applications that run on those smart devices. Being the key driver behind these low-cost memories with gigabyte capacity, the non-volatile NAND Flash memory has the niche of simple cell structure and small cell size [1–10]. Owing to its noiseless, faster, and more mechanically stable operations, the utilization of NAND Flash memory has been rapidly expanded into the SSD (solid-state drive) market that has begun to threaten the conventional magnetic hard drive market shares [11, 3, 12–15]. As a result of geometric scaling, more Flash memory cells can be packed onto the same chip and at a reduced cost per technology generation. These scaling efforts are, however, fast approaching unsustainable limits such as lithographic patterning resolution, transistor short-channel effects, and ever increasing interference between adjacent cells [3, 5–8].

In order to achieve much higher-density Flash devices with alleviated scaling problems, 3D cell structures [16–24] and multi-level-cells (MLCs) [25, 3, 5, 26–28, 9] are feasible pathways forward and are currently being developed in nearly all semiconductor memory corporations. MLC chips containing two or three memory bits in a single cell have already been adopted as a standard in consumer electronics. Since the continual increase of MLC memory density has resulted in slower operation and instability during data reading [29–31], the emphasis in research and development is shifting toward the 3D cell approach. Although several 3D structures have been proposed to date, including Bit Cost Scalable (BiCS, figures 1(a)–(c)) from Toshiba [18, 19, 23], their acceptance is restrained by the complex and high-cost fabrication steps (supplementary figure S1 available at stacks.iop.org/Nano/22/254006/mmedia). Furthermore, those
structures turn out to be unsuitable for 3D chip architecture to maximize cell-area efficiency (ratio of memory area to chip area). In this paper, we report a breakthrough to increase memory density in a cost-effective manner by wafer level 3D memory chip architecture, called SMOL (stacked-memory-devices-on-logic). The SMOL has been realized through an innovative 3D memory device of a vertical-stacked-array-transistor (VSAT) [22] and interconnection structure of planarized-integration-on-the-same-plane (PIPE) [20].

2. VSAT and PIPE structure

A VSAT is a series of vertical NAND devices with a polysilicon channel that overcomes the above-mentioned scaling limits and process complexity, and achieves ultra-high-density memory devices. The polysilicon channel covers stacked control gates (CGs, figure 1(d)) and each CG has a double-gate-in-series structure embedding a SONOS layer on both sidewalls (figure 1(e)). The electric current flows horizontally in general from the source-select-gate (SSG) to the ground-select-gate (GSG), but locally up-and-down along the vertical strings. The VSAT memory density is proportional to the number of stacked layers, which can be freely increased by depositing extra CG layers without needing advanced and expensive lithography processes. In other words, the effective cell area decreases in reverse proportion to the number of stacked layers and hence the chip area could be reduced by 45% (=60% − 60% × 1/4, where 60% is the ratio of memory area to chip area) in the case of four multiple stacks or 52.5% (=60% − 60% × 1/8) in the case of eight multiple stacks. PIPE (figure 1(f)) is a unique interconnection structure providing a more convenient and cost-efficient method than a conventional stair-like structure (figure 1(c)). The PIPE will also impact various crossbar-typed devices [32–37], which have ultimate memory density, to develop much easier and cost-effective interconnections (supplementary figure S2 available at stacks.iop.org/Nano/22/254006/mmedia). We have successfully demonstrated the VSAT prototypes with PIPE structure having a 100 nm feature size (figures 2(b)–(d)). The process starts from creating an oxide mesa structure (figure 2(a)). Next, multiple layers of doped polysilicon for CGs and insulating film are alternately deposited. Afterward, the multi-layer stack is defined into a vertical fin array, and all the CG electrodes are exposed on the same plane through a CMP process to provide easy access for outside connections. The tunneling-oxide, charge-trapping-nitride, and control-oxide films conformally wrapped around the fin array are sequentially deposited, followed by depositing a conformal ultra-thin polysilicon film as the channel material. Finally, a process to pattern the polysilicon into line-shaped channels is carried out. The combined VSAT and PIPE structure offers numerous advantages over other 3D structures such as ultra-high memory density, a simple process sequence, easy integration with peripheral circuits, small overhead area for the vertical interconnection, and highly suppressed short-channel effects (supplementary figure S3 available at stacks.iop.org/Nano/22/254006/mmedia).
Figure 2. Features of the VSAT structure. (a) Process sequence of VSAT. Multiple layers of gate electrodes and isolating films are deposited over an oxide mesa. The control gates (CGs) are defined through a lithography and dry etching process, followed by a CMP process flattening the CGs. After depositing SONOS and polysilicon films on CGs and separating the channels, an interconnecting process is followed. (b) SEM image of the PIPE structure. A pair of PIPEs is formed on both sides of an oxide mesa, enabling the via-holes to leave out every second CG, thus leading to an easy via-hole process. (c) SEM image of the VSAT. A VSAT with four multiple layers (280 nm high) has been developed at the 100 nm technology node using KrF lithography. (d) TEM images of VSAT. Al$_2$O$_3$/nitride/tunneling-oxide are deposited in turn on CGs, finally followed by a 20 nm thick polysilicon film for a channel material.

3. SMOL, 3D memory chip architecture

In a typical 2D architecture, memory arrays and peripheral logic devices are generally located on the same plane above the Si substrate (figure 3(a)) since both devices use single crystalline Si as the channel material. These 2D chips have a cell-area efficiency of approximately 60% and in other words, peripheral logic devices use 40% of the chip area. In order to increase the cell-area efficiency, the 3D vertical-chip architecture is preferred to have the memory and logic cells stacked vertically (figure 3(b)). With this configuration, the VSAT structure can be readily implemented for the 3D chip architecture, SMOL (figures 3(c) and (d)), because the SGs are spatially separated from the memory arrays. The underlying Si substrate can thus be freed up in a VSAT (figure 1(e)) as compared to the BiCS structure in which the Si substrate is utilized as the SGs (figure 1(b)). Accordingly, the peripheral CMOS devices are first built up over the whole chip area together with interconnect lines and landing-pads for vertical interconnection. Stacked-memory-devices are then formed above it. In order to have vertical connections between VSAT and logic devices, the polysilicon channels are connected to the SGs through the landing-pads, while the CGs are connected to the CG-drivers through the landing-pads as well as the PIPE structures (figures 3(c) and (d)).

The SMOL structure with VSATs has many advantages over other stand-alone 3D memory device implementations. The first and main advantage is the increased cell-area efficiency originating from a vertical arrangement of memory and logic devices. Figure 3(d) shows the configuration of a SMOL structure composed of four multiple-stacked layers, and a pair of PIPE structures at both ends of each CG. An impressive cell-area efficiency of more than 95% is achieved, assuming that the unit block consists of 2 kB/page, 16 vertical stacks per string, and four multiple layers in a vertical stack (supplementary figure S4 available at stacks.iop.org/Nano/22/254006/mmedia). A 5% overhead area is sufficient to build up the PIPE structure with a 50 nm thick CG, 40 nm vertical spacing, and 70$^\circ$ angle on a 100 nm feature size (figure 3(e)). As a result, memory density increases by 35% for the same chip area. Second, the requirement of scaling down logic devices becomes less stringent because the whole Si substrate can be used to configure the peripheral logic devices, which results in lower subthreshold leakage current and thus lower power consumption. Third, the SMOL process sequence is simpler and more cost-effective than the 2D chip architecture, which usually fabricates the 3D memory devices separately from the 2D logic devices, and thus requires additional processing steps for protecting the devices. Furthermore, higher system speed might be possible through the elimination of I/O buffers and with, instead, a direct interface with internal host logic devices in the same chip, when the SMOL technology is realized above microprocessor circuits [38]. In this study, a prototype SMOL structure has been successfully developed with a feature size of 100 nm, fabricating VSAT above the pre-fabricated logic devices. Tungsten is used as the interconnection material for logic devices to endure the subsequent high processing temperature involved in fabricating VSAT devices. Multiple layers of n-type doped polysilicon and oxide thin films are deposited on an oxide buffer layer of 300 nm thick (figure 3(f)). Next, the SMOL prototype is constructed using the fabrication steps for the VSAT (figure 2(a)) and demonstrates highly integrated 3D memory on logic chip architecture (figure 3(g)).
20 nm thick and a hydrogen annealing process that reduces the total number of traps. With an ultra-thin polysilicon film (from 6.0 V), a reduced subthreshold slope of 300 mV dec−1 and an increased drive current of 3A m−2 of phosphorus (figures 5(b) and (c)), even disconnect the channels. MEDICIT™ simulations have been carried out for varying dopant types and doping levels in the polysilicon film, while fixing the vertical spacing to 50 nm in order to ensure a small cell-to-cell interference. The simulation results show that the electrical characteristics are uniform through a wide range of doping levels from 10^{16} cm^{−3} of boron to 10^{16} cm^{−3} of phosphorus (figures 5(b) and (c)), even though the parasitic resistance increases for heavily doped p-type cases and leakage current rapidly increases in heavily doped n-type cases (figure 5(c)). An ultra-thin polysilicon film makes this wide doping window possible because the total amount of charge (Q_{total} = N_{doping} \times t_{body}) is small even with high doping levels. This in turn yields a small

4. Electrical performance

Although polysilicon is widely used as a channel material in 3D devices, there is still a critical concern about its electrical performance as a FET channel due to the trapping sites at the polysilicon grain boundaries. In order to lower the device threshold voltage (V_{th}) and improve the subthreshold characteristics, an ultra-thin polysilicon film has been employed [18] as it reduces the polysilicon volume and total number of traps. With an ultra-thin polysilicon film of 20 nm thick and a hydrogen annealing process that reduces the total number of traps, the device electrical performance has been significantly improved with a lowered V_{th} of 1.5 V (from 6.0 V), a reduced subthreshold slope of 300 mV dec^{−1} (from 800 mV dec^{−1}), and an increased drive current of 1.3 A m^{−1} (from 0.3 A m^{−1}) (figures 4(a) and (b) and supplementary figure S5 available at stacks.iop.org/Nano/22/254006/mmedia). The channel doping level and vertical distance between WLs are therefore the crucial parameters for desired electrical performance. A larger vertical spacing yields less interference yet it results in larger parasitic resistance in the source–drain region, which might even disconnect the channels. MEDICIT™ simulations have been carried out for varying dopant types and doping levels in the polysilicon film, while fixing the vertical spacing to 50 nm in order to ensure a small cell-to-cell interference. The simulation results show that the electrical characteristics are uniform through a wide range of doping levels from 10^{16} cm^{−3} of boron to 10^{16} cm^{−3} of phosphorus (figures 5(b) and (c)), even though the parasitic resistance increases for heavily doped p-type cases and leakage current rapidly increases in heavily doped n-type cases (figure 5(c)). An ultra-thin polysilicon film makes this wide doping window possible because the total amount of charge (Q_{total} = N_{doping} \times t_{body}) is small even with high doping levels. This in turn yields a small

5. Results of device simulation

The source–drain region of a VSAT has an identical doping level to the channel region because no intentional doping has been introduced into the former, unlike in conventional FETs. The electrons in the source–drain region are induced by the fringing electric field (figure 5(a)) of CGs, which electrically connect the individual channels together as one conducting path (supplementary figure S6 available at stacks.iop.org/Nano/22/254006/mmedia). The channel doping level and vertical distance between WLs are therefore the crucial parameters for desired electrical performance. A larger vertical spacing yields less interference yet it results in larger parasitic resistance in the source–drain region, which might even disconnect the channels. MEDICIT™ simulations have been carried out for varying dopant types and doping levels in the polysilicon film, while fixing the vertical spacing to 50 nm in order to ensure a small cell-to-cell interference. The simulation results show that the electrical characteristics are uniform through a wide range of doping levels from 10^{16} cm^{−3} of boron to 10^{16} cm^{−3} of phosphorus (figures 5(b) and (c)), even though the parasitic resistance increases for heavily doped p-type cases and leakage current rapidly increases in heavily doped n-type cases (figure 5(c)). An ultra-thin polysilicon film makes this wide doping window possible because the total amount of charge (Q_{total} = N_{doping} \times t_{body}) is small even with high doping levels. This in turn yields a small
Figure 4. Electrical performance of VSAT. (a) $V_{GS}$–$I_{DS}$ plot. DC performance is improved by adopting a thin polysilicon channel and a hydrogen annealing process, showing 1.5 V of $V_{th}$, 300 mV dec$^{-1}$ of subthreshold slope, and 15 cm$^2$ V$^{-1}$ s$^{-1}$ of mobility extracted from the maximum transconductance. (b) $V_{GS}$–$I_{DS}$ plot. 1.3 µA µm$^{-1}$ of drive current is achieved at $V_{GS} = 4.0$ V. (c) Mechanism of programming or erasing. Electrons move in either the forward or reverse direction between a channel and a trapping layer of nitride via the polarity of the gate bias. Trapping or de-trapping electrons in a nitride film shifts the threshold voltage of the device and serves as the data storage mechanism. (d) Memory functions. Positive bias on gates pulls electrons from the channel to the trapping layer, and thus $V_{th}$ moves up. (e) Negative bias pushed electrons from the trapping layer to the channel, and $V_{th}$ moves down. (f) $V_{th}$ changes by as much as 4 V after a 100 ms programming and erasing cycle.

Figure 5. Device simulations. (a) Fringing electric field in the source–drain region. The electric field affects the source–drain region through the fringing capacitance, and induces electrons, electrically connecting the whole channel. (b) $V_{th}$ depending on the channel doping levels simulated by MEDICI. $V_{th}$ varies less than 50 mV under a wide window of doping levels. (c) $I_{ON}$–$I_{OFF}$ characteristics simulated by MEDICI. $I_{ON}$ and $I_{OFF}$ shows stable performance from B $10^{16}$ to Ph $10^{16}$ cm$^{-3}$, although it shows highly conducting or non-conducting characteristics in extremely highly doped cases. (d) Interference mechanism in VSATs. A $V_{th}$ shift occurs on the center cell because of the neighboring cells with an applied bias of $V_{READ}$. (e) and (f) $V_{th}$ shift by interference simulated by MEDICI. The lateral spacing is more sensitive than the vertical spacing, owing to the larger coupling capacitance. VSAT can be scaled down to 40 nm vertically and 50 nm laterally maintaining 0.25 V of $V_{th}$ shift.
potential change \( (V = Q_{\text{total}}/C_{\text{ox}}) \) and thus a stable operation. Finally, the interference between neighboring cells in VSAT is examined because it would limit the proximity between cells, and thus the ultimate memory density and scalability. Through MEDICT™ simulations, direct interference effects have been studied by monitoring the \( V_{th} \) shift induced by the neighboring CGs’ potential (figure 5(d)). The amount of interference is proportional to the coupling capacitance between the channel and neighboring CGs. The lateral coupling is stronger than the vertical coupling since the CGs and channels face each other, which constitutes a larger coupling area. The amount of \( V_{th} \) shift is estimated to be 0.25 V, an acceptable number by the industrial standard, with a 50 nm lateral spacing and 40 nm vertical spacing (figures 5(e) and (f)). This result strongly supports the VSAT strategy to increase the memory density by staking more layers, which is manufactured using only relatively low-cost lithography, deposition, and etching processes.

6. Conclusion

A novel SMOL structure has been successfully developed through a simple and cost-effective method. The cell-area efficiency of more than 95% can be achieved through the SMOL structure with a feature size of 100 nm. Electrical performance of the polysilicon channel has been dramatically improved by adopting an ultra-thin film and hydrogen annealing process. MEDICT™ simulations exploring interference and channel doping effects show that the SMOL structure shows many desirable memory characteristics. The advantages of the SMOL architecture make it a viable candidate for low-cost and high-density memory devices. This technology is compatible with the CMOS process and is also applicable to various emerging memory devices.

References

[27] Park J-H et al 2006 Highly manufacturable 32Gb multi-level NAND flash memory with 0.0098 \( \mu m² \) cell size using TANOS(Si-oxide-Al₂O₃-TaN) cell technology IEDM ’06: Int. Electron Devices Mtg pp 1–4


[34] Deng J et al 2007 Metallization for crossbar molecular devices Nanotechnology 18 155202

[35] Vontobel P O et al 2009 Writing to and reading from a nano-scale crossbar memory based on memristors Nanotechnology 20 425204

[36] Yun D K et al 2009 Mass fabrication of resistive random access crossbar arrays by step and flash imprint lithography Nanotechnology 20 445305

[37] Jeong H Y et al 2010 A low-temperature-grown TiO₂-based device for the flexible stacked RRAM application Nanotechnology 21 115203