Challenges of nickel silicidation in CMOS technologies ☆

Nicolas Breil [a],*, Christian Lavoie [b],*, Ahmet Ozcan [a], Frieder Baumann [a], Nancy Klymko [a], Karen Nummy [a], Bing Sun [a], Jean Jordan-Sweet [b], Jian Yu [a], Frank Zhu [a], Shreesh Narasimha [a], Michael Chudzik [a]

[a] IBM Semiconductor Research and Development Center (SRDC), East Fishkill, NY 12533, USA
[b] IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA

ABSTRACT

In this paper, we review some of the key challenges associated with the Ni silicidation process in the most recent CMOS technologies. The introduction of new materials (*e.g.* SiGe), and of non-planar architectures bring some important changes that require fundamental investigation from a material engineering perspective. Following a discussion of the device architecture and silicide evolution through the last CMOS generations, we focus our study on a very peculiar defect, termed *NiSi-Fangs*. We describe a mechanism for the defect formation, and present a detailed material analysis that supports this mechanism. We highlight some of the possible metal enrichment processes of the nickel monosilicide such as oxidation or various RIE (Reactive Ion Etching) plasma process, leading to a metal source available for defect formation. We also investigate the NiSi formation and re-formation silicidation differences between Si and SiGe materials, and between (1 0 0) and (1 1 1) orientations. Finally, we show that the thermal budgets post silicidation can lead to the formation of *NiSi-Fangs* if the structure and the processes are not optimized. Beyond the understanding of the defect and the discussion on the engineering solutions used to prevent its formation, the interest of this investigation also lies in the fundamental learning within the Ni–Pt–Si–Ge system and some additional perspective on Ni-based contacts to advanced microelectronic devices.

© 2015 Elsevier B.V. All rights reserved.

Introduction

For several decades, the microelectronics industry has been fueling the development of the Information Technology and Communications sectors, virtually enabling most of the recent innovations and revolutions observed across the research and industry communities. The exponential increase observed in the available computing power for a given cost – also known as *Moore's Law* [1] – has essentially been allowed by the ability to scale down all dimensions of a transistor, directly increasing performance by the same factor while maintaining a constant power dissipation per unit area. This scaling law, as described first by Dennard [2], is at the origin of the continuous performance increase and the phenomenal rise in transistor densities observed over half a century.

When the transistor gate length dimensions reached the sub-100 nm regime, the scaling started to deviate from ideality because of the limitation in gate oxide thickness since gate leakage could not be easily controlled. At first, these deviations in scaling were mainly compensated by a slower decrease in operating voltages in order to achieve the expected performance gains. The partly scaled voltages lead to drastic increases in power densities and heat dissipation problems that have limited CMOS performance. In the last few technology generations, because of material limitations of the traditional doped silicon source and drain, silicon channel and silicon oxide dielectric, the scaling alone, although surely increasing transistor density, has not nearly provided the expected performance improvements. Maintaining some of the scaling expectations down to the smallest devices ($\sim$10 nm) in these conditions has only been possible through the implementation of technology boosters which usually consisted in integrating modified substrates, new materials or stressors to improve the properties of silicon oxide and doped silicon.

Within IBM, the fabrication of devices, often referred to as Front End of Line (FEOL) was first modified by changing substrate from bulk Si(1 0 0) substrates to Si on insulator substrates (SOI). The generations afterwards saw the additions of silicon nitride stressors to increase mobility of carriers in the channel [3], high-K metal gates to decrease equivalent oxide thickness and reduce leakage [4,5], epitaxial S/D stressors [6,7] which are now followed by new device geometries (nanowires, UTSOI [8–13], triple gates, finfets

[14–17]).This reduction in device size obviously occurred in concert with a reduction in size of the interconnections in the Back End of Line (BEOL) as well as with multiple materials modifications. The past two decades have witnessed BEOL changes from Al to Cu as interconnection material, addition of low K then ultra low K dielectrics, test of air gaps as well as change in liners and capping layers for Cu lines.

As the device size continues to decrease, the connection between the semiconducting device and the multiple levels of Cu in the BEOL becomes critical. This first contact level to the device has historically been made using a self aligned process to form a conducting silicide only over exposed Si areas [18]. In earlier technologies, the source and drain areas were silicided concurrently with the gate structure. In more recent technologies, only source and drain are silicided since the gate is now fully replaced using metals with desired workfunctions on n or p FET transistors. The scaling of the contact size through generations has lead to the arrival of many new yield limiting defects. These defects have forced the evolution of metal silicide contacts from titanium to cobalt and to nickel silicides. Lately, the appearance of 3D geometries has lead to the formation of silicides in constrained volumes which have caused a new family of defects controlled by the stress driven diffusion of Si, causing extrusion-like defects with the unreacted metal. Within IBM such defects have been termed filaments, stringers or wings depending on their extent and location in a device.

In the last few technologies, the constraints expected on the contacts from regular scaling of the area were dramatically increased with the appearance of trench contacts and 3D devices. When performing a silicidation reaction at the bottom of a contact trench, the contact length is typically two times smaller compared to a direct silicidation of the source and drain regions. The latest change to triple gate (or finFET) has increased the channel width also by more than a factor of two for the same contact area. With increasing fin heights, these two factors can decrease the equivalent contact length by close to an extra order of magnitude. With these important process evolutions, an intrinsic contact resistivity of less than 5E-9 $\Omega-cm^2$ is a necessity as far as performance is concerned for upcoming technologies.

Nickel silicide based contacts have the ability to reach intrinsic contact resistivities of about 1E-9 $\Omega-cm^2$ [19,20] and are therefore adequate for current technologies. However, as the device size is reduced and the geometry is changing, stress in localized volume when the silicide forms and proximity of substrates materials that are different in composition lead both to unexpected diffusion and to the formation of defects that are considerably more cumbersome to understand and control. As a result, it remains unclear what the contacts of tomorrow will become. Much research has been done using Metal–Insulator–Semiconductor (MIS) contacts [20b] and attempts to form silicide contacts using higher temperature materials is also underway. In this paper, we direct our attention to maintaining a defect free Ni silicide through scaling and new geometries. We do so using our recent materials learning related to the elimination of an unexpected and startling yield limiting defect.

Evolution of device architectures and silicidation techniques

Traditional devices are usually built using a so-called *planar* architecture, as shown in Fig. 1a. The need for an increased scaling and a better control of the channel electrostatics has pushed the devices toward 3D architectures, such as the FinFET device shown in Fig. 1b. From a silicidation perspective, it is important to note that both these devices exhibit some significant *non-planarities* that have major implications in the engineering of the silicidation module.

As shown in Fig. 1a, the source and drain areas are typically formed by an epitaxial process, with SiGe being used for the p-type MOSFETs and, in the case of the latest IBM 22SOI technology, with Si:C being used for the n-type MOSFETs [21]. These epitaxially grown materials exhibit a (1 1 1) facet that grows from the intersection between the source/drain area and the surrounding oxide insulation zone. This intersection is obviously non-planar, as shown in the Fig. 1a, and brings some challenges in the silicidation process as will be discussed later in this paper.

In the case of the FinFET architecture shown in Fig. 1b, the gate surrounds the silicon fin channel, which gives a significant advantage for the control of the channel electrostatics. However, the narrow and tall fin is a major drawback from a silicide contact perspective, because it strongly limits the available surface for the contact, and increases the electrical resistance of the fin that connects the contact to the channel. Thus an *in situ* doped epitaxial film (i.e. SiGe, Si:C) is an important process step in the source/drain module, in order to increase the effective contact surface, and to bring more volume and dopants to reduce the source/drain resistance. Hence, the introduction of epitaxial films is an enabler
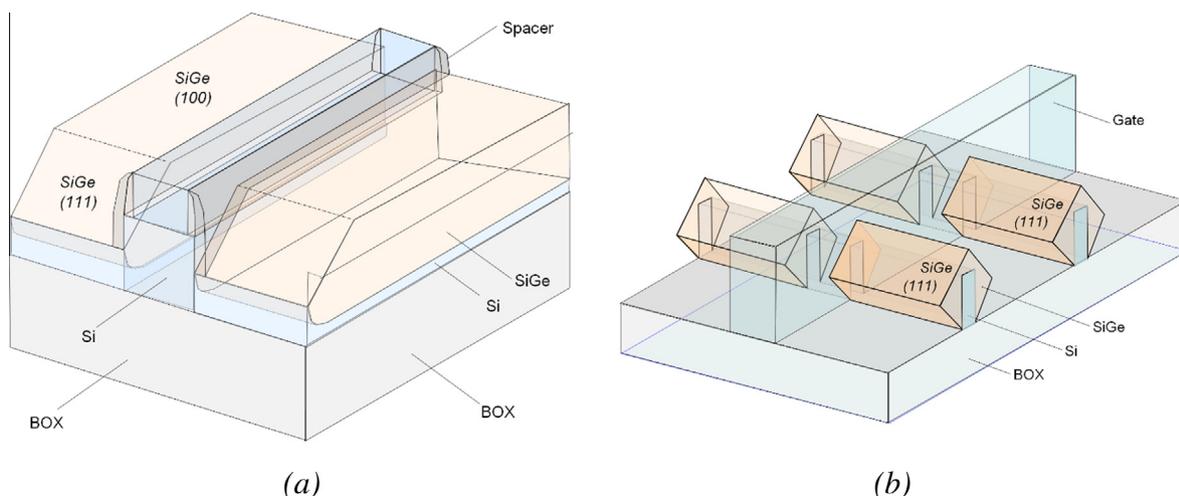

Fig. 1. (a) Schematic view of a gate first planar device on a SOI substrate. The (1 1 1) facet typically appears during epitaxial overgrowth of an embedded source or drain area. (b) Schematic view of a 3D FinFET device on a SOI substrate. Because of different epitaxial growth kinetics on (1 1 1) facets, the epitaxial process results in the formation of diamond shapes source/drain regions.