THERMAL ANNEAL PROCESS FOR STRAINED-SI DEVICES

BACKGROUND

[0001] The present invention relates generally to semiconductor devices, and more particularly to a method to improve carrier mobility, through the incorporation of strained silicon, in semiconductor devices.

[0002] Metal-Oxide-Semiconductor field effect transistors (MOSFETs) are a common component of integrated circuits (ICs). By continually reducing the gate oxide thickness, and the length of the gate electrode, the semiconductor industry has doubled the transistor's switching speed every eighteen to twenty-four months, a phenomenon popularly known as the Moore's Law. The strategy of shrinking both the gate oxide thickness, and the source drain length, has allowed the Moore's Law to uphold in the past, but the technology market's insatiable desire for increasing switching speed is causing the strategy to reach its limits. Newer strategies, such as switching to new materials for the transistor's substrate, and the incorporation of new process methods to effectively use these new materials, are, therefore, designed and applied.

[0003] One such strategy for increasing the performance of MOSFETs is to enhance the carrier mobility of silicon (Si), thereby reducing resistance and power consumption, and increasing drive current, frequency response and operating speed. A method of enhancing carrier mobility that has become a focus of recent attention is the use of silicon material to which a tensile strain is applied. "Strained" silicon, or strained-Si, may be formed by growing a layer of silicon on a silicon germanium (SiGe) substrate or base. The silicon germanium lattice is generally

more widely spaced than a pure silicon lattice as a result of the presence of the larger germanium atoms in the lattice. Because the atoms of the silicon lattice align with the more widely spread silicon germanium lattice, a tensile strain is created in the silicon layer. In other words, the silicon atoms are essentially pulled apart from one another. The amount of tensile strain applied to the silicon lattice increases with the proportion of germanium in the silicon germanium lattice.

[0004] Relaxed silicon has six equal valence bands. The application of tensile strain to the silicon lattice causes four of the valence bands to increase in energy and two of the valence bands to decrease in energy. As a result of quantum effects, electrons effectively weigh 30 percent less when passing through the lower energy bands. Thus the lower energy bands offer less resistance to electron flow. In addition, electrons encounter less vibrational energy from the nucleus of the silicon atom, thereby causing them to scatter at a rate of 500 to 1,000 times less than in relaxed silicon. As a result, carrier mobility is dramatically increased in strained silicon as compared to relaxed silicon, thereby offering a potential increase in mobility of 80 percent or more for electrons and 20 percent or more for holes. The increase in mobility has been found to persist for current fields of up to 1.5 megavolts/centimeter. These factors are believed to enable a device speed increase of 35 percent without further reduction of device size, or a 25 percent reduction in power consumption without a reduction in performance.

[0005] The abruptness of the source and drain extensions is critical in strained-Si technology. One difficulty is that dopants are observed to diffuse faster in strained-Si than in Si, thereby causing drain extensions to move further under the gate and increasing gate-to-drain capacitance. Also, the diffusion of germanium across the SiGe-to-silicon boundary will release stress and reduce mobility in the strained Si layer.

[0006] In further consideration, the junction depth for source and drains should be only 35-75 nm deep for the 100 nm generation of devices to go into production in the near future, while drain extensions should only be less than 20-35 nm deep.

[0007] Therefore, desirable in the art of semiconductor designs are additional designs that may reduce the implant energy and annealing time, thereby achieving higher carrier mobility as well as shallower junctions.

SUMMARY

[0008] In view of the foregoing, this invention provides an advanced thermal annealing process for forming strained-Si devices.

[0009] In one embodiment, a method is disclosed for forming a semiconductor device using strained silicon. After forming a first substrate material, with a first natural lattice constant on a device substrate, and a second substrate material with a second natural lattice constant, on the first substrate material, channel, source and drain regions of a field effective transistor are further defined using the first and second substrate materials. After implanting one or more impurity materials to the source and drain regions, and the transistor goes through an annealing process using a high speed heat source other than a Tungsten-Halogen lamp.

[0010] The construction and method of operation of the invention, however, together with additional objects and advantages thereof, will be best understood from the following description of specific embodiments when read in connection with the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0011] FIG. 1 illustrates a cross-sectional drawing of a conventional MOSFET.

[0012] FIG. 2 illustrates a cross-sectional drawing of a MOSFET fabricated with two layers of substrates, in accordance with a first embodiment of the present invention.

[0013] FIG. 3 tabulates the heat pulse times of various heat sources.

[0014] FIG. 4 illustrates a cross-sectional drawing of a MOSFET fabricated with a second semiconductor material in the source-drain regions in accordance with a second embodiment of the present invention.

[0015] FIG. 5 illustrates a cross-sectional drawing of a MOSFET fabricated with one layer of substrate in accordance with a third embodiment of the present invention.

DESCRIPTION

[0016] The present invention provides various design embodiments to demonstrate the effects of incorporating a strained channel and an advanced thermal annealing process to improve transistor performance.

[0017] FIG. 1 illustrates a cross-sectional drawing 100 of a conventional MOSFET. The MOSFET includes a gate electrode 102 that is separated from a channel region 104 in the substrate 106 by a thin first gate insulator 108, such as silicon oxide or oxide-nitride-oxide (ONO). The gate electrode 102 is typically formed from a doped semiconductor material such as polysilicon.

[0018] The source and drain of the MOSFET are formed on opposing sides of the gate electrode 102 as deep source and drain regions 110. Source and drain silicide contacts 112 are formed on the source and drain regions 110. These silicides typically include a compound comprising the substrate semiconductor material and a metal such as cobalt (Co), nickel (Ni) or titanium (Ti) used to reduce contact

resistance to the source and drain regions 110. The source and drain regions 110 are formed deep enough to extend beyond the depth to which the source and drain silicide contacts 112 are formed. The source and drain regions 110 are implanted only after the formation of spacers 114 around the gate electrode 102. The spacers 114 serve as an implantation mask to define the lateral position of the source and drain regions 110 relative to the channel region 104 beneath the gate.

[0019] The gate electrode 102, likewise, has a silicide gate contact 116 formed on its upper surface. The gate structure comprising a polysilicon material and an overlying silicide is sometimes referred to as a polycide gate.

[0020] The source and drain of the MOSFET further includes shallow source and drain extensions 118. As dimensions of the MOSFET are reduced, short channel effects resulting from the small distance between the source and drain cause a degradation of MOSFET performance. The use of shallow source and drain extensions 118, rather than the deep source and drain regions near the ends of the channel 104, help to reduce short channel effects and excessive device element capacitances. The shallow source and drain extensions are implanted prior to the formation of the spacers 114. Diffusion, during subsequent annealing causes the source and drain extensions 118 to extend slightly beneath the gate electrode 102.

[0021] FIG. 2 illustrates a cross-sectional view 200 of a MOSFET fabricated with two layers of substrates in accordance with a first embodiment of the present invention. The MOSFET structure is formed on a substrate 202, typically silicon. On top of the substrate 202 is a silicon germanium layer 204, further on which is formed an epitaxial layer of strained silicon 206. The natural lattice constants of the silicon germanium layer 204 and the silicon layer 206 are different. The MOSFET uses other conventional MOSFET structures, including the deep source and drain regions 110,

the shallow source and drain extensions 118, the gate oxide layer 108, the gate electrode 102 surrounded by spacers 114, silicide source and drain contacts 112, and the silicide gate contact 116. The channel region 104 of the MOSFET includes the strained silicon material which provides enhanced carrier mobility between the source and drain. It is understood that the channel lattice distance may be bigger than five angstroms due to the strained silicon arrangement.

[0022] After the introduction of dopant impurities, typically by implant, annealing is performed to activate the implants. The use of a fast heat source such as a noble gas lamp is herein proposed as an improved method for performing this annealing over the typically used Tungsten-Halogen lamp. Such a heat source shall be able to provide a heat ramp-up rate of 250 degrees Celsius per second. While in the annealing process, a peak temperature time should be controlled so that no overheating is caused. Typically, for the recommended heat sources described in more details below, no more than 10 ms of the peak temperature time is required.

[0023] In a first suggested embodiment, this type of heat lamp may be a DC water-wall argon lamp or an AC xenon lamp. These types of lamps provide flash times as low as 0.8 ms, and as high as 1.5 ms. In addition, the wavelength of the xenon lamp is typically between 1.4 and 5 um. The use of these types of lamps allows the annealing of the semiconductor without causing excessive germanium diffusion, which reduces channel mobility. In addition, the use of these sources may reduce the diffusion of impurities under the gate and the relaxation of the strained-Si layer caused by diffusion of germanium into the strained layer. In the first suggested embodiment, the use of any one of the proposed high speed heat sources may produce MOSFETs, as shown in FIG. 2, with junction abruptness of less than 10 nm per decade drop of doping concentration.

[0024] In another suggested embodiment, an XeCl excimer laser may be used as an annealing heat source. The heat pulse time of an excimer laser provides the shortest pulse time of 90 ns. A Table 300 of the heat pulse times of various heat sources is presented in FIG. 3.

[0025] FIG. 4 illustrates a cross-sectional drawing 400 of a MOSFET according to another embodiment of the present invention. In this embodiment, the source and drain regions 402 at the opposing sides of the gate may have different lattice constants due to different material combinations. Further, the MOSFET structure may have the same structure as the one illustrated in FIG. 2, but may also have a strained-Si layer 404 only in the channel regions. The formation of different materials in the source or drain region can be done through conventional photo masking and etching techniques. For embodiment, a portion of the source or drain region containing the compound semiconductor, typically SiGe, can be removed. Then, semiconductor material is deposited to form the source or drain region. When the lattice constants of the materials in the source and drain regions are different, the anneal process ad described can be used for a better resulting product. One of the improved anneal heat sources, as tabulated in the Table 300, may be used in the fabrication of this type of device, which may have a junction abruptness of less than 10 nm per decade drop of doping concentration. The same device improvements of minimal germanium and implanted impurity diffusion may then be realized.

[0026] FIG. 5 illustrates a cross-sectional drawing 500 of a MOSFET fabricated with one layer of substrate in accordance with a third embodiment of the present invention. A strained-Si layer 502 is formed on a substrate layer 504, typically SiGe. Alternatively, such a device may be fabricated in the substrate layer 504 with a channel lattice distance greater than 5.43 Angstroms. One of the improved anneal heat sources, as tabulated in the Table 300, may be used in the fabrication of this

type of device, which may have a junction abruptness of less than 10 nm per decade drop of doping concentration.

[0027] The above invention provides many different embodiments or embodiments for implementing different features of the invention. Specific embodiments of components and processes are described to help clarify the invention. These are, of course, merely embodiments and are not intended to limit the invention from that described in the claims.

[0028] Although the invention is illustrated and described herein as embodied in a design and method for forming strained silicon devices, it is nevertheless not intended to be limited to the details shown, since various modifications and structural changes may be made therein without departing from the spirit of the invention and within the scope and range of equivalents of the claims. Accordingly, it is appropriate that the appended claims be construed broadly and in a manner consistent with the scope of the invention, as set forth in the following claims.

WHAT IS CLAIMED IS:

1. A method for forming a semiconductor device using strained silicon, the method comprising:

forming a first substrate material with a first natural lattice constant on a device substrate;

forming a second substrate material with a second natural lattice constant on the first substrate material;

defining a channel, source and drain regions of a field effective transistor using the first and second substrate materials;

implanting one or more impurity materials to the source and drain regions; and

annealing the transistor using a high speed heat source other than a Tungsten-Halogen lamp,

wherein the channel has a channel lattice distance more than about five angstroms.

2. The method of claim 1 wherein the heat source is a noble gas lamp.

3. The method of claim 2 wherein the heat source is a DC water-wall argon lamp.

4. The method of claim 2 wherein the heat source is an AC xenon lamp.

5. The method of claim 4 wherein the xenon lamp has a wavelength between 1.4 um to 5 um.

6. The method of claim 1 wherein the heat source is a XeCl excimer laser.

7. The method of claim 1 wherein the annealing provides a junction abruptness of less than 10 nm per decade drop for a doping concentration of the impurity materials.

8. The method of claim 1 wherein a predetermined peak temperature time for the annealing is no longer than 10 ms.

9. The method of claim 1 wherein the annealing provides a heat ramp-up rate no slower than 250 degrees Celsius per second.

10. The method of claim 1 wherein the first substrate material is silicon germanium.

11. The method of claim 1 wherein the second substrate material is an epitaxial strained silicon.

12. The method of claim 1 wherein the channel region is formed in the second substrate material only.

13. The method of claim 1 wherein the source and drain regions have different lattice constants.

14. A method for forming a semiconductor device using strained silicon, the method comprising:

defining a channel, source and drain regions of a field effective transistor

using a strained silicon substrate having a first and second substrate materials having different lattice constants;

implanting one or more impurity materials to the source and drain regions; and

annealing the transistor using a high speed heat source other than a Tungsten-Halogen lamp,

wherein the annealing provides a junction abruptness of less than 10 nm per decade drop for a doping concentration of the impurity materials.

15. The method of claim 14 wherein the heat source is a noble gas lamp.

16. The method of claim 15 wherein the heat source is a DC water-wall argon lamp.

17. The method of claim 15 wherein the heat source is an AC xenon lamp having a wavelength between 1.4 um to 5 um.

18. The method of claim 14 wherein the heat source is a XeCl excimer laser.

19. The method of claim 14 wherein a predetermined peak temperature time for the annealing is no longer than 10 ms.

20. The method of claim 19 wherein the annealing provides a heat ramp-up rate no slower than 250 degrees Celsius per second.

21. The method of claim 14 wherein the first substrate material is silicon germanium.

22. The method of claim 14 wherein the second substrate material is an epitaxial strained silicon.

23. The method of claim 1 wherein the source and drain regions have different lattice constants.

24. The method of claim 1 wherein the channel has a channel lattice distance more than about five angstroms.

25. A semiconductor device comprising:

a strained silicon substrate having a first and second substrate materials with different lattice constants;

a channel region defined on the substrate having a channel lattice distance more than about five angstroms; and

a source and drain regions of a field effective transistor defined on the substrate doped with one or more impurity materials with a junction abruptness of less than 10 nm per decade drop for a doping concentration thereof.

26. The device ofclaim 25 wherein the junction abruptness is achieve by annealing the transistor using a high speed heat source having a heat ramp-up rate no slower than 250 degrees Celsius per second.

27. The device of claim 25 wherein the first substrate material is silicon germanium.

28. The device of claim 27 wherein the second substrate material is a strained silicon.

29. The device of claim 25 wherein the source and drain regions have different lattice constants.