Nanoelectronics: Devices and Materials

Prof. Navakanta Bhat
Centre for Nano Science and Engineering
Indian Institute of Science, Bangalore

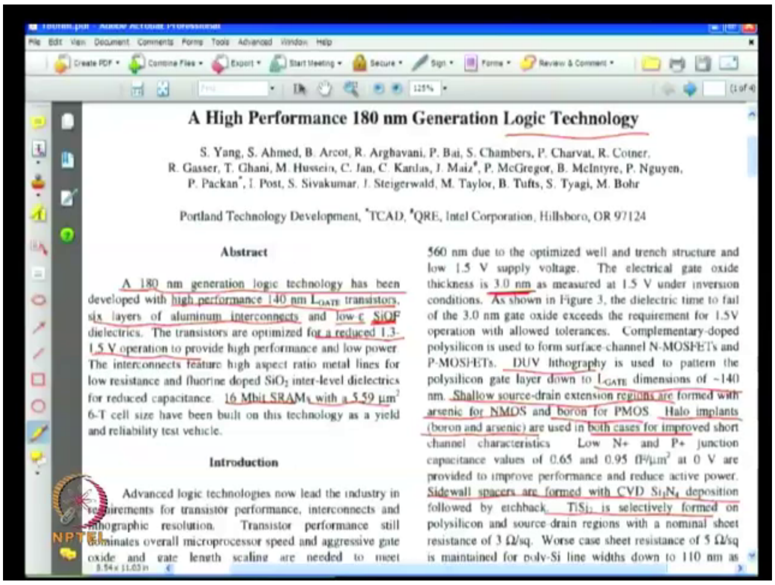
Lecture - 11
Industrial CMOS Technology

Today, what we will do is; go through some representative CMOS technology. And you will recognize that all that we have studied so far is very relevant and it is in fact, being implemented in state of our CMOS technology.

So, what I thought I will do is; walk you through some representative notes. We will start with 180 nanometer generation CMOS technology and then progressively look at scaling; look at 130 nanometer, 90, 65, 45, 32; each one being 0.7 X of the previous technology. So, let us then start with this; what I will do here is essentially I will make use of these papers, which are published in conference called IDM; which is a annual conference called International electron Devices Meeting; it is a premier conference in electron devices. And any new device technology, which gets developed, gets reported in these conferences for the first time.

So, what I have done is that; I have taken a series of papers published in past several years. We will go through these papers one by one and of course, we will not really read the entire paper, but I will highlight a few important things to draw your attention to some of the very important technological implementations.

(Refer Slide Time: 01:45)



A High Performance 180 nm Generation Logic Technology

S. Yang, S. Ahmed, B. Arcot, R. Arghavani, P. Bai, S. Chambers, P. Charvat, R. Cotner,
R. Gasser, T. Ghani, M. Hussein, C. Jan, C. Kardas, J. Maiz[#], P. McGregor, B. McIntyre, P. Nguyen,
P. Packan[*], I. Post, S. Sivakumar, J. Steigerwald, M. Taylor, B. Tufts, S. Tyagi, M. Bohr

Portland Technology Development, [*]TCAD, [#]QRE, Intel Corporation, Hillsboro, OR 97124

Abstract

A 180 nm generation logic technology has been developed with high performance 140 nm L_{GATE} transistors, six layers of aluminum interconnects and low-ε SiOF dielectrics. The transistors are optimized for a reduced 1.3-1.5 V operation to provide high performance and low power. The interconnects feature high aspect ratio metal lines for low resistance and fluorine doped SiO_2 inter-level dielectrics for reduced capacitance. 16 Mbit SRAMs with a 5.59 μm^2 6-T cell size have been built on this technology as a yield and reliability test vehicle.

Introduction

Advanced logic technologies now lead the industry in requirements for transistor performance, interconnects and lithographic resolution. Transistor performance still dominates overall microprocessor speed and aggressive gate oxide and gate length scaling are needed to meet

560 nm due to the optimized well and trench structure and low 1.5 V supply voltage. The electrical gate oxide thickness is 3.0 nm as measured at 1.5 V under inversion conditions. As shown in Figure 3, the dielectric time to fail of the 3.0 nm gate oxide exceeds the requirement for 1.5V operation with allowed tolerances. Complementary-doped polysilicon is used to form surface-channel N-MOSFET's and P-MOSFETs. DUV lithography is used to pattern the polysilicon gate layer down to L_{GATE} dimensions of ~140 nm. Shallow source-drain extension regions are formed with arsenic for NMOS and boron for PMOS. Halo implants (boron and arsenic) are used in both cases for improved short channel characteristics. Low N+ and P+ junction capacitance values of 0.65 and 0.95 fF/μm^2 at 0 V are provided to improve performance and reduce active power. Sidewall spacers are formed with CVD Si_3N_4 deposition followed by etchback. $TiSi_2$ is selectively formed on polysilicon and source-drain regions with a nominal sheet resistance of 3 Ω/sq. Worse case sheet resistance of 5 Ω/sq is maintained for poly-Si line widths down to 110 nm as

Let us look at this paper here now; which is essentially A High Performance 180 nanometer Generation Logic Technology. And just to be consistent; 180 nanometer technology and all subsequent technologies of course, will be implemented by various companies, various foundries. Just I have picked papers published from Intel; just without any bias to any company, but just to tell you a representative technology and also with respect to Dell technology, Intel technology; we will see how the technology has progressed.

So, this is essentially a 180 nanometer Generation Logic Technology, if you see here 180 nanometer Generation Logic Technology; has been developed with high performance 140 nanometer gate length transistor. You see, I had told you that there is so called technology node, but the gate lengths of the transistor could be less than the technology node. So, this technology node is 180 nanometer. However, the gate length of the transistor is actually 140 nanometer. It has 6 layers of aluminum interconnects, the interconnect in this technology is still aluminum. However, later on you will see that they have migrated from aluminum to copper; in terms of the interconnect.

And the dielectric between interconnects, which is metal interconnect; if you recall, we have discussed this in our CMOS process flow; between one metal level to another metal level, you have an insulator. And that insulator is silicon oxide; except that it is little bit variant of silicon oxide; it is fluorine doped silicon oxide. The idea of doping it with fluorine is to reduce the dielectric constant of silicon oxide; that is why they say it is a low Epsilon; Epsilon is a dielectric constant.

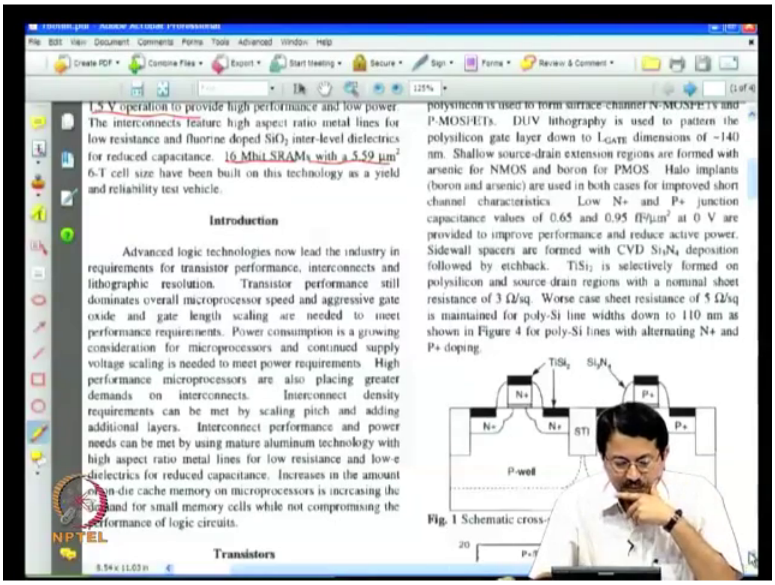
Because in interconnects you want to minimize the parasitic capacitances, parasitic resistances. And hence you need to minimize the capacitances, this is opposite of high k gate dielectric, these are low k gate dielectrics. So, you minimize the interconnect delay and interconnect capacitances. The transistors are optimized for a reduced 1.3 to 1.5 volt operation you see the transistors operate at low voltage; at 1.5 volt maximum, not beyond that.

The interconnects which are high aspect ratio, metal lines and all that and using this technology they have demonstrated a 16 bit SRAM; with certain memory cell size and this is sort a demonstration of the technology. What they are showing here is that; not only they have developed the transistors, but they have put together transistor to make a circuit. And in this case, the circuit is static random access memory; which is a representative circuit.

Eventually in this technology is a logic technology; you see here logic technology, not a memory technology. Logic technology will be used to build microprocessors, which are very high and logic chips, but microprocessors on chips will have cache memory. So, these SRAMs will be used for cache memories in the microprocessor, these are not going to be standalone SRAM chips to be sold.

So, we have been able to demonstrate that SRAM circuit can also be implemented using this technology.

(Refer Slide Time: 05:10)



1.5 V operation to provide high performance and low power. The interconnects feature high aspect ratio metal lines for low resistance and fluorine doped SiO$_2$ inter-level dielectrics for reduced capacitance. 16 Mbit SRAMs with a 5.59 μm^2 6-T cell size have been built on this technology as a yield and reliability test vehicle.

Introduction

Advanced logic technologies now lead the industry in requirements for transistor performance, interconnects and lithographic resolution. Transistor performance still dominates overall microprocessor speed and aggressive gate oxide and gate length scaling are needed to meet performance requirements. Power consumption is a growing consideration for microprocessors and continued supply voltage scaling is needed to meet power requirements. High performance microprocessors are also placing greater demands on interconnects. Interconnect density requirements can be met by scaling pitch and adding additional layers. Interconnect performance and power needs can be met by using mature aluminum technology with high aspect ratio metal lines for low resistance and low-e dielectrics for reduced capacitance. Increases in the amount of on-die cache memory on microprocessors is increasing the demand for small memory cells while not compromising the performance of logic circuits.

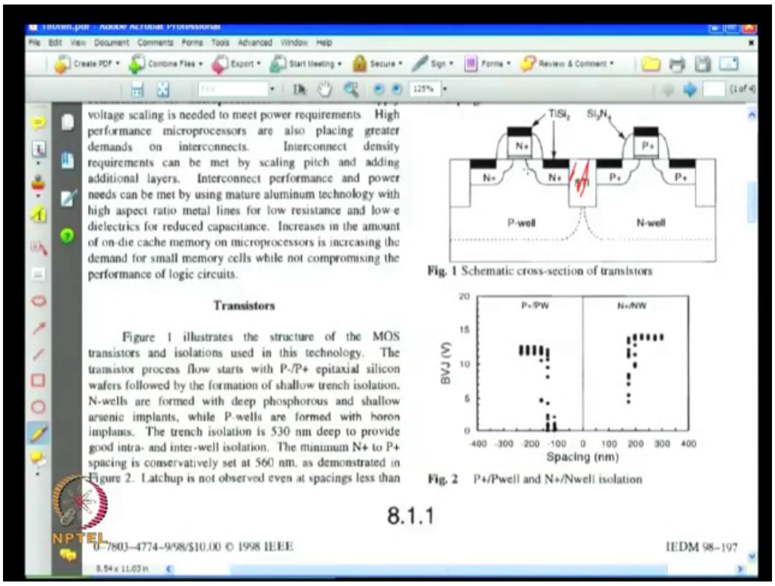
Transistors

polysilicon is used to form surface channel N-MOSFETs and P-MOSFETs. DUV lithography is used to pattern the polysilicon gate layer down to L$_{GATE}$ dimensions of ~140 nm. Shallow source-drain extension regions are formed with arsenic for NMOS and boron for PMOS. Halo implants (boron and arsenic) are used in both cases for improved short channel characteristics. Low N+ and P+ junction capacitance values of 0.65 and 0.95 fF/μm^2 at 0 V are provided to improve performance and reduce active power. Sidewall spacers are formed with CVD Si$_3$N$_4$ deposition followed by etchback. TiSi$_2$ is selectively formed on polysilicon and source drain regions with a nominal sheet resistance of 3 Ω/sq. Worse case sheet resistance of 5 Ω/sq is maintained for poly-Si line widths down to 110 nm as shown in Figure 4 for poly-Si lines with alternating N+ and P+ doping.

Fig. 1 Schematic cross-

So, let us then just go through some key highlights here and they give the introduction; why do you need scaling and all that; that is very typical.

(Refer Slide Time: 05:23)



voltage scaling is needed to meet power requirements. High performance microprocessors are also placing greater demands on interconnects. Interconnect density requirements can be met by scaling pitch and adding additional layers. Interconnect performance and power needs can be met by using mature aluminum technology with high aspect ratio metal lines for low resistance and low-e dielectrics for reduced capacitance. Increases in the amount of on-die cache memory on microprocessors is increasing the demand for small memory cells while not compromising the performance of logic circuits.

Transistors

Figure 1 illustrates the structure of the MOS transistors and isolations used in this technology. The transistor process flow starts with P-/P+ epitaxial silicon wafers followed by the formation of shallow trench isolation. N-wells are formed with deep phosphorous and shallow arsenic implants, while P-wells are formed with boron implants. The trench isolation is 530 nm deep to provide good intra- and inter-well isolation. The minimum N+ to P+ spacing is conservatively set at 560 nm, as demonstrated in Figure 2. Latchup is not observed even at spacings less than

Fig. 1 Schematic cross-section of transistors

Fig. 2 P+/Pwell and N+/Nwell isolation

8.1.1

0-7803-4774-9/98/$10.00 © 1998 IEEE

IEDM 98-197

Now, let us come to the transistor; so, the transistor is sort of illustrated in this figure; you have CMOS technology, you have N channel and P channel transistor. Figure one illustrates the structure of the MOS transistor and isolation used in this technology. The isolation as we know is a shallow trench isolation; correct now tello coat isolation. You have made a trench inside the silicon and that isolates the neighboring transistors. Start with a P minus P plus epitaxial silicon wafers, followed by the formation of shallow trench isolation, N wells are formed with deep phosphorous and shallow arsenic implants; they have not used in this particular case antimony and indium yet.

But none the less they use phosphorous and arsenic combination for N well; while P wells are formed with boron implants that are how they are forming these wells. And then they also have the electrical gate oxide thickness is; 3 nanometer. Gate oxide what they show here is a electrical thickness, when we talk of C V characterization subsequently. From next class, we will see what is the difference between physical thickness and electrical thickness?

Electrical thickness is always little more than physical thickness. So, if they are saying 3 nanometer is their electrical thickness, their physical thickness could be something like 2 nanometer; less than 3 nanometer for sure. And they have ensured that; this dielectric has sufficient reliability as they will demonstrate and their gate lengths have been patterned using deep UV lithography; it is an optical lithography essentially.

Shallow source drain extensions; remember all that we talked about in terms of source drain engineering, they have used shallow source drain extensions with arsenic for NMOS, boron for PMOS; these are source drain extension for N channel and P channel transistors. Halo implants, boron and arsenic are used in both cases; for improved short channel characteristics, low N plus P plus junction capacitance values; that is fine.

Sidewall spacers are formed using CVD nitride; remember the nitrides spacer that we talked about, that is in fact; used to isolate deep sourced extension from your shallow extensions, they make use of that. They use silicide; titanium silicide, remember the parasitic region that we talked about; after the gating of the transistor, there is a contact and in between there is a silicon and if you convert it into a silicide, you get lower resistivity and in this technology they are using titanium silicide.

So, that the typical on; and source drain regions with a normal sheet resistance of 3 ohms per square; that is the kind of sheet resistance that you can get, which is a very low sheet resistance by the way. So, hence this is how the cross section would look you see, there is a N channel transistor in a P well, there is a deep source drain, there is a shallow extension, there is a spacer; which separates the two and there is a silicide here. And similarly silicide will be on poly, as well as on this junction and this is a PMOS transistor; complimentary metal oxide silicon technology.

So, this is how you would define the transistor and then you will have to characterize a lot of things. You will have to make sure that the isolation is sufficient, for that you need to characterize the breakdown voltage of these; that is between this and this. If you apply 2 voltage; it should not break down, the breakdown voltage should be more like 5, 10 volt. So, for all practical purposes these two are perfectly isolated, so there is essentially demonstrating all these characteristics.

You can go back and go through these papers, these papers are certainly available if you are an I EEE member or if your institute is subscribing I EEE journals; you will certainly have this listed in the I EEE publication.

(Refer Slide Time: 09:31)

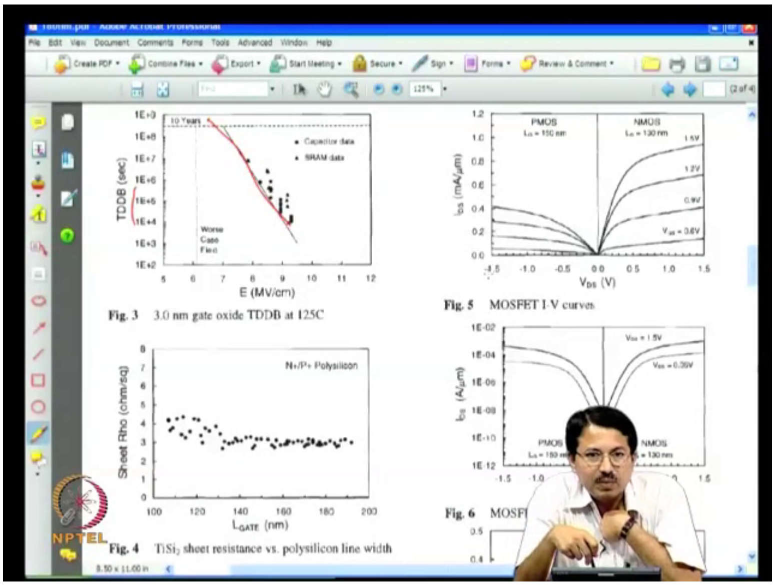


Fig. 3 3.0 nm gate oxide TDDB at 125C

Fig. 5 MOSFET I-V curves

Fig. 4 TiSi$_2$ sheet resistance vs. polysilicon line width

Fig. 6 MOSF

Another important aspect that we have discussed; all these reliability issues: remember I said when you do the gate oxide reliability; there is something called TDDB; which is time dependent dielectric breakdown. Even though you are going to operate your transistor at very low electric field, such as may be in this case 6 mega volt per centimeter; what you do is that, you apply large electric field and characterize the breakdown times and based on that you sort of extrapolate it, and say that in the real use condition certainly this will be reliable up to 10 years; they have done a similar thing, what we have discussed; these are typical reliability characterizations.

And here they show typical N channel and P channel characteristics, this is drain current versus drain voltage; the so called output characteristics, which are well behaved; going up to 1.5 volt for NMOS and minus 1.5 volt for P channel transistor. Of course, here N current is more than the P current because of the electron mobility being higher than the whole mobility.

(Refer Slide Time: 10:54)

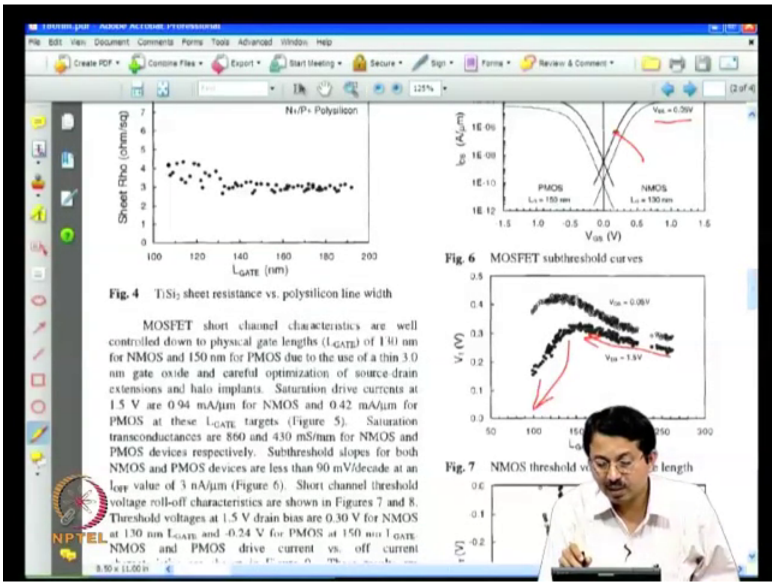


Fig. 4 TiSi$_2$ sheet resistance vs. polysilicon line width

Fig. 6 MOSFET subthreshold curves

Fig. 7 NMOS threshold v length

MOSFET short channel characteristics are well controlled down to physical gate lengths (L_{GATE}) of 130 nm for NMOS and 150 nm for PMOS due to the use of a thin 3.0 nm gate oxide and careful optimization of source-drain extensions and halo implants. Saturation drive currents at 1.5 V are 0.94 mA/μm for NMOS and 0.42 mA/μm for PMOS at these L_{GATE} targets (Figure 5). Saturation transconductances are 860 and 430 mS/mm for NMOS and PMOS devices respectively. Subthreshold slopes for both NMOS and PMOS devices are less than 90 mV/decade at an I_{OFF} value of 3 nA/μm (Figure 6). Short channel threshold voltage roll-off characteristics are shown in Figures 7 and 8. Threshold voltages at 1.5 V drain bias are 0.30 V for NMOS at 130 nm L_{GATE} and -0.24 V for PMOS at 150 nm L_{GATE}. NMOS and PMOS drive current vs. off current

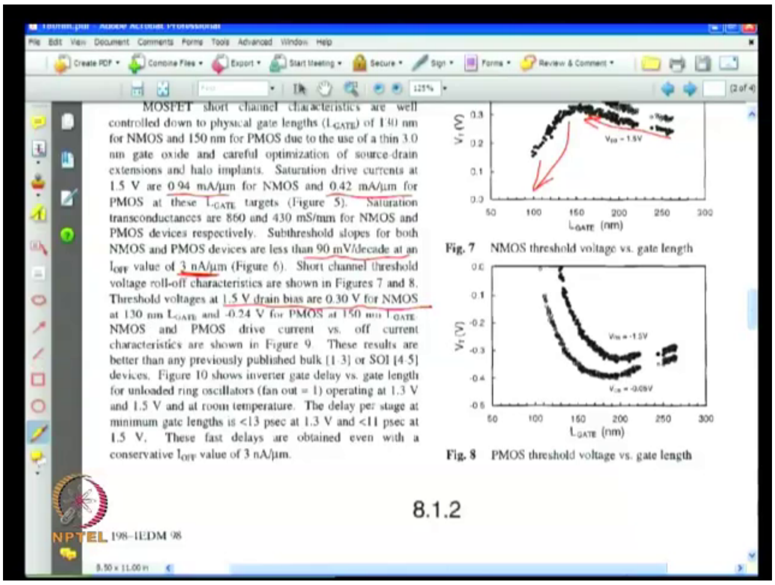
This is sub threshold characteristics your threshold voltage is about 0.5 volt or so, and below that this is IDS and log scale and VGS in linear scale; it is a semilog plot really and this really is a sub threshold slope, as we have already discussed. And again notice that; depending on whether you have 0.05 volt here or 1.5 volt here, you will have different characteristics; that are essentially due to double effect drain induced by real lowering.

If you have larger drain voltage for the same gate voltage, you get larger current because your drain electric field will influence the injection at the source side; especially in these nanometric dimension transistors. So, this is just to show that; your polysilicon gate because of the silicidation has reasonably low resistivity.

Sheet resistance is the order of 3 to 4 ohm per square, which is fair good. These are typical Vt versus length plots, remember what our discussion; there is a reverse short channel effect. I told you whenever you use a halo; invariably you will have reverse short channel effects. And then usual Vt roll off that you will have in this region here this; Vt is coming down. This is Vt measured at 1.5 volt on the drain and this is Vt measured at 0.05 volt on the drain and this difference is essentially DIBL.

Vt at low voltage on the drain is always higher than Vt at high voltage on the drain; your DIBL can be characterized based on that.

(Refer Slide Time: 12:32)



MOSFET short channel characteristics are well controlled down to physical gate lengths (L_{GATE}) of 130 nm for NMOS and 150 nm for PMOS due to the use of a thin 3.0 nm gate oxide and careful optimization of source-drain extensions and halo implants. Saturation drive currents at 1.5 V are 0.94 mA/μm for NMOS and 0.42 mA/μm for PMOS at these L_{GATE} targets (Figure 5). Saturation transconductances are 860 and 430 mS/mm for NMOS and PMOS devices respectively. Subthreshold slopes for both NMOS and PMOS devices are less than 90 mV/decade at an I_{OFF} value of 3 nA/μm (Figure 6). Short channel threshold voltage roll-off characteristics are shown in Figures 7 and 8. Threshold voltages at 1.5 V drain bias are 0.30 V for NMOS at 130 nm L_{GATE} and -0.24 V for PMOS at 150 nm L_{GATE}. NMOS and PMOS drive current vs. off current characteristics are shown in Figure 9. These results are better than any previously published bulk [1-3] or SOI [4-5] devices. Figure 10 shows inverter gate delay vs. gate length for unloaded ring oscillators (fan out = 1) operating at 1.3 V and 1.5 V and at room temperature. The delay per stage at minimum gate lengths is <13 psec at 1.3 V and <11 psec at 1.5 V. These fast delays are obtained even with a conservative I_{OFF} value of 3 nA/μm.

Fig. 7 NMOS threshold voltage vs. gate length

Fig. 8 PMOS threshold voltage vs. gate length

8.1.2

NPTEL 198–IEDM 98

So, they have NMOS and PMOS devices which have a sub threshold slope of 90 millivolt per decade. Remember 60 is the best you can get, but you will never get 60; because you have 1 plus CD over C ox term, which is always more than 1. So, because of that it will be more than 60; so, they have been able to achieve 90 millivolt per decade. And their off currents are of the order of 3 nano ampere per micrometer, we always specify off current in terms of ampere per unit width; depending on how many microns your transistor width is you multiply that, you get the off state current.

And your on state currents; saturation drive currents are 0.94 milliampere for NMOS and 0.42 milliampere for PMOS. So, in other words your on current to off current ratio is almost 10 power 6, if it were to be 10 power 6; then it would have been 3 milliampere here, but it is not 3 milliampere; it is close to 1 milliampere, which is reasonably good on current to off current ratio. This is NMOS, this is PMOS same story; reverse short channel effect and typical Vt roll off; the Vts are negative.

Again lower drain voltage, higher drain voltage; short channel, threshold, voltage, roll off are shown in these figures; threshold voltages are 1.5 volt; drain bias are 0.3 volt for NMOS and minus 0.24 volt for PMOS. And these results are of course, better than any previously published bulk or SOI devices; they are trying to say that, they have been able to come up with much better devices.

Subsequently, they have also made use of these transistors and dealt what is called ring oscillator and from that they extract inverter gate delay. If you were to build a simple inverter, a not gate; what is the gate propagation delay? You have a input transition from low to high, how long does it take for output to go from high to low? That is your propagation delay and they have been able to get the propagation delay and typically, it is very difficult to measure propagation delays of the order of these propagation delays that they have are of the order of few picoseconds.

You do not construct just one single inverter; because if you want to construct one single inverter, you will have to measure the time resolution of the order of picoseconds. What you do is that; you cascade the inverters, a large number of inverters are in series; very interestingly, if you have a odd number of inverters in series and connect the output of the last inverter back to the input of the first inverter.

Create a feedback circuit; that starts working as an oscillator that is why it is called a ring oscillator. Ring meaning, you have created a inverter ring; you need to have an odd number of inverter; if you have even number of inverter, it will go to a steady state; input will also be a study, output will also be steady, but if you have a odd number of inverter; you can do this exercise yourself; output will try to change the input and it will keep changing; it will become a oscillator circuit.

So, you do this oscillator; you measure the oscillators time period and you know how many stages of inverters you have and you back calculate; the delay per inverter and that is what is done typically. And they do it for unloaded ring oscillator meaning there is no extra load on the oscillator. The only load is the subsequent inverter stage; operating at 1.3 and 1.5 at room temperature; they have been able to get very low delays of the ring oscillators.

(Refer Slide Time: 16:39)

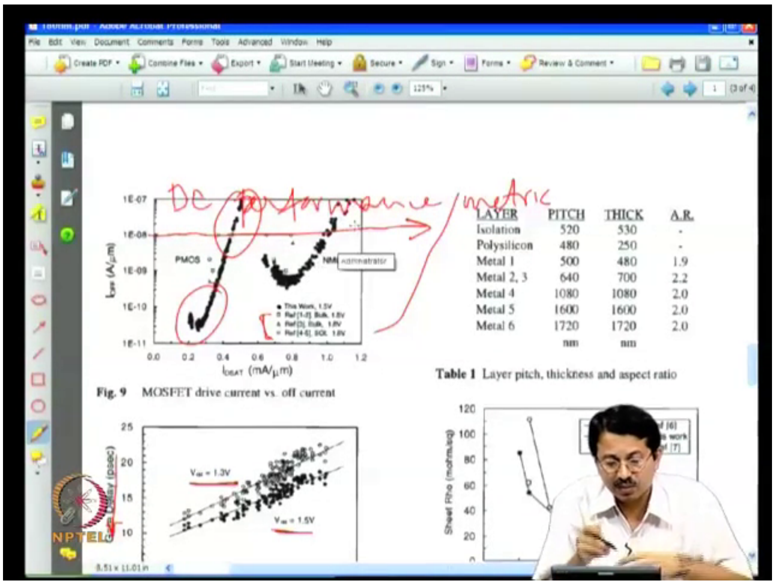


Fig. 9 MOSFET drive current vs. off current

LAYER	PITCH	THICK	A.R.
Isolation	520	530	-
Polysilicon	480	250	-
Metal 1	500	480	1.9
Metal 2, 3	640	700	2.2
Metal 4	1080	1080	2.0
Metal 5	1600	1600	2.0
Metal 6	1720	1720	2.0
	nm	nm	

Table 1 Layer pitch, thickness and aspect ratio

We will come to this in a minute, but this is typically what is plotted; gate delay as a function of gate length. If you have a long transistor, your delays are also more; that is why we are scaling; I mean the reason why we scale is that, the circuit start operating faster as we have discussed in the very beginning. As you start shrinking this gate length, the gate length is decreasing here, the delay is also decreasing.

And the nominal gate length for this technology is of the order of 150 nanometer and that 150 nanometer, your gate delays are of the order of ten picoseconds. Again, if you operate this at lower voltage; your delay is little longer, little more and if you apply higher voltage for this circuit; not very high, I mean if you apply very high voltage; it will break down; then you can decrease the delay a little further.

One good way of bench marking the transistors is to really plot this; so, called DC performance metric; what is DC performance metric? This is a XY plot with on current on the X axis and off current on the Y axis. And the requirement for a good transistor is that; for any given off current, whatever that off current is that; let us say 10 nanometer off current.

I want to be furthest right of this curve, if you have two technologies for example, if I have a new technology; which looks like this. This is much better transistor because for the same leakage current, it is giving you much much better on state current; on state current is much larger. Here on to off current ratio is huge, so that it will be less leaky; it will not consume large static power, at the same time it will be very fast. So, it is essentially a 2 D plot of off current versus on current and you want to really create your transistor to really go along this; towards the right.

And of course, NMOS tends to be for the same leakage current; to the right of PMOS, simply because you have much better electron hole mobility; compared to hole mobility; but what they are trying to do here is that, the filled simple they say this work and their some previous works; from references, these are empty symbols. So, what they have plotted here is that; these empty symbols are to the left of; all these are PMOS transistors from previously published papers.

And all these empty symbols here are NMOS transistors from previously published papers. And all they are trying to say is that; compared to previously published PMOS devices; these are much better. Because for the same off current, your curve is to the right; you get much higher on current and that is what you want to generate. The way you generate this curve by the way; if you want to generate such a curve, you cannot generate this curve with a one gate length transistor.

What you do is that; using the same technology, you print transistors of different gate lengths; gate lengths from 130 nanometer, 150 nanometer, 200 nanometer, 300 nanometer so on and so forth. As your gate length is increasing, your leakage current decreases, your on current also decreases. So, all these points here; you see correspond to shorter gate length; because these are large on current and large off current. All these points here correspond to longer gate length pitch channel transistor and that is how you generate series of these points by characterizing transistors of different dimensions.

And you plot a universal curve which is a off current versus on current curve. Similarly, for N channel transistors; these are the transistors which have very high on current and off current, which are coming from very short channel 130 nanometer kind of transistor. Whereas, along this these are longer channel transistors and that is how this is generated and they do interconnects. Interconnect is essentially aluminum in this case they have not yet used copper.

But the only thing that they have done different in interconnect is to use fluorinated silicon oxide as I mentioned earlier here.

(Refer Slide Time: 21:28)

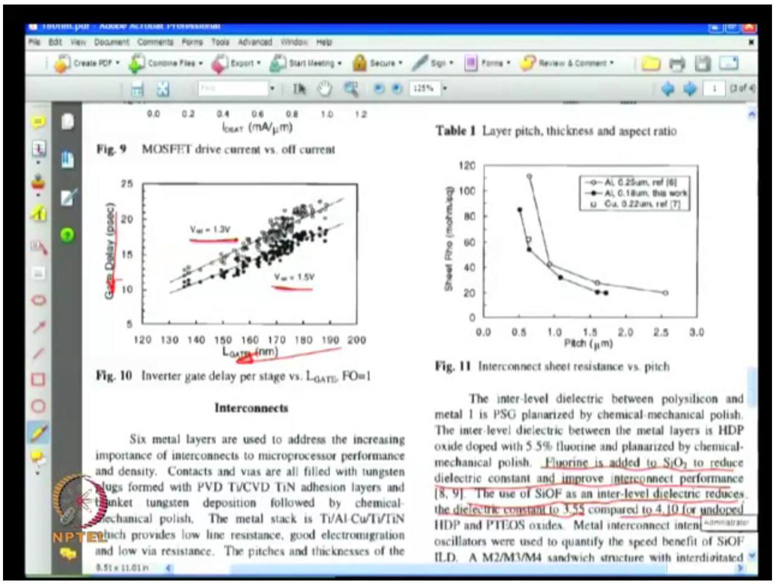


Fig. 9 MOSFET drive current vs. off current

Fig. 10 Inverter gate delay per stage vs. L_{GATE}, FO=1

Interconnects

Six metal layers are used to address the increasing importance of interconnects to microprocessor performance and density. Contacts and vias are all filled with tungsten plugs formed with PVD Ti/CVD TiN adhesion layers and blanket tungsten deposition followed by chemical-mechanical polish. The metal stack is Ti/Al-Cu/Ti/TiN which provides low line resistance, good electromigration and low via resistance. The pitches and thicknesses of the

Table 1 Layer pitch, thickness and aspect ratio

Fig. 11 Interconnect sheet resistance vs. pitch

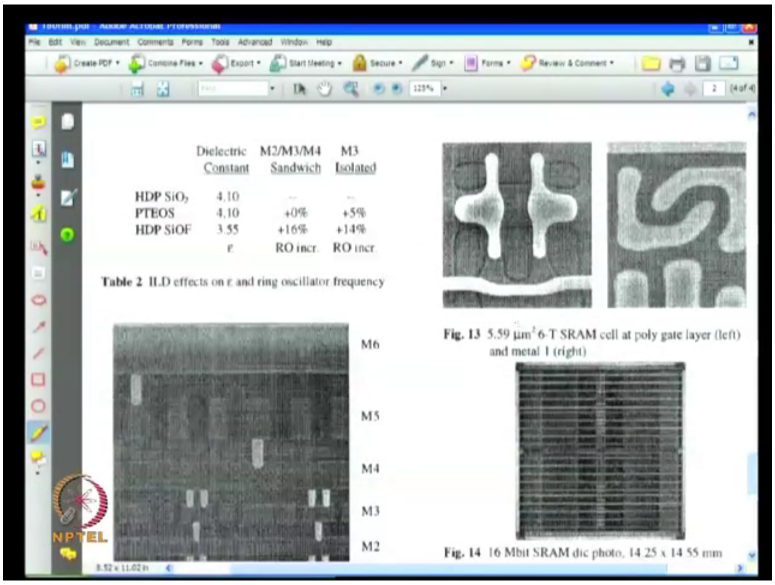
The inter-level dielectric between polysilicon and metal 1 is PSG planarized by chemical-mechanical polish. The inter-level dielectric between the metal layers is HDP oxide doped with 5.5% fluorine and planarized by chemical-mechanical polish. Fluorine is added to SiO_2 to reduce dielectric constant and improve interconnect performance [8, 9]. The use of SiOF as an inter-level dielectric reduces the dielectric constant to 3.55 compared to 4.10 for undoped HDP and PTEOS oxides. Metal interconnect inten oscillators were used to quantify the speed benefit of SiOF ILD. A M2/M3/M4 sandwich structure with interdigitated

You see fluorine is added to a silicon oxide; to reduce dielectric constant and improve interconnect performance. The use of SiOF as an inter level dielectric, reduces the dielectric constant to 3.55 compared to 4.1 for undoped silicon oxide; remember silicon oxide, the dielectric constant is around 44.1 and whereas, that comes down to 3.55.

You see thermal oxide, if oxide is grown at let us say high temperature; 1000 degree centigrade then its dielectric constant is more like 3.9; that is what we used earlier when we are talking of gate dielectric; 3.9, we round it off to 4. But here we say, 4.1 because this oxide is not thermally grown, it is a deposited oxide. Because you already have a metal line and top of metal line, you have to put oxide; so there is no silicon to grow oxide, you will have to do a chemical vapor deposition of the oxide.

And invariably these oxides which are done at low temperature, because you already have a metal; tend to have little higher dielectric constant and hence the dielectric constant is of the order of 4.1; after fluorine doping, it comes all the way down to 3.55.

(Refer Slide Time: 23:00)



	Dielectric Constant	M2/M3/M4 Sandwich	M3 Isolated
HDP SiO_2	4.10	--	--
PTEOS	4.10	+0%	+5%
HDP SiOF	3.55	+16%	+14%
	ε	RO incr.	RO incr.

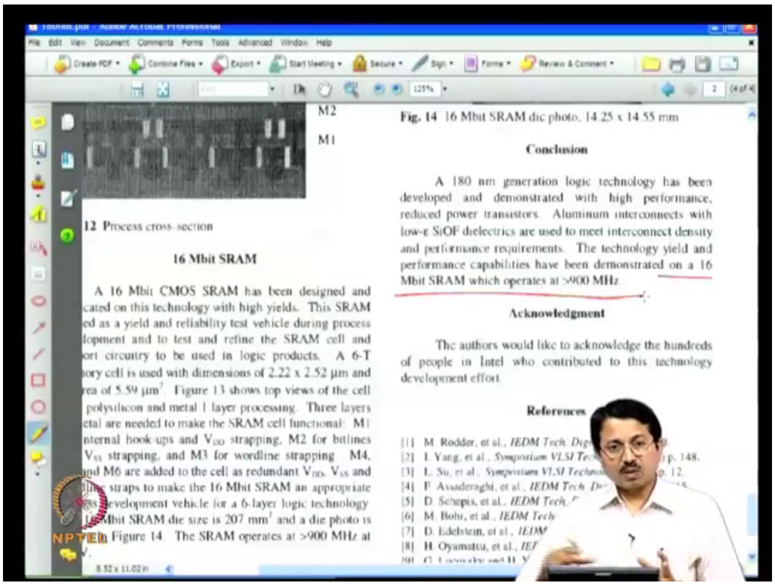
Table 2 ILD effects on ε and ring oscillator frequency

Fig. 13 5.59 μm^2 6-T SRAM cell at poly gate layer (left) and metal 1 (right)

Fig. 14 16 Mbit SRAM die photo, 14.25 x 14.55 mm

So, they actually build a SRAM cell; this is the top view of the SRAM cell where in you have 6 transistors, this is 6 transistors SRAM cell. This is their interconnect stack; starting from the transistor at the very bottom, these are the contacts; metal 1 via 1; metal 2 via 2; metal 3, via 4, metal 4 and so on and so forth; going up to 6 levels of metals; in this case.

(Refer Slide Time: 23:34)



12 Process cross-section

16 Mbit SRAM

A 16 Mbit CMOS SRAM has been designed and cated on this technology with high yields. This SRAM ed as a yield and reliability test vehicle during process lopment and to test and refine the SRAM cell and ort circuitry to be used in logic products. A 6-T ory cell is used with dimensions of 2.22 x 2.52 μm and rea of 5.59 μm^2. Figure 13 shows top views of the cell polysilicon and metal 1 layer processing. Three layers etal are needed to make the SRAM cell functional: M1 nternal hook-ups and V_{DD} strapping, M2 for bitlines V_{SS} strapping, and M3 for wordline strapping. M4, nd M6 are added to the cell as redundant V_{DD}, V_{SS} and line straps to make the 16 Mbit SRAM an appropriate ss development vehicle for a 6-layer logic technology 16 Mbit SRAM die size is 207 mm^2 and a die photo is n Figure 14. The SRAM operates at >900 MHz at V.

Fig. 14 16 Mbit SRAM die photo, 14.25 x 14.55 mm

Conclusion

A 180 nm generation logic technology has been developed and demonstrated with high performance, reduced power transistors. Aluminum interconnects with low-ε SiOF dielectrics are used to meet interconnect density and performance requirements. The technology yield and performance capabilities have been demonstrated on a 16 Mbit SRAM which operates at >900 MHz.

Acknowledgment

The authors would like to acknowledge the hundreds of people in Intel who contributed to this technology development effort.

References

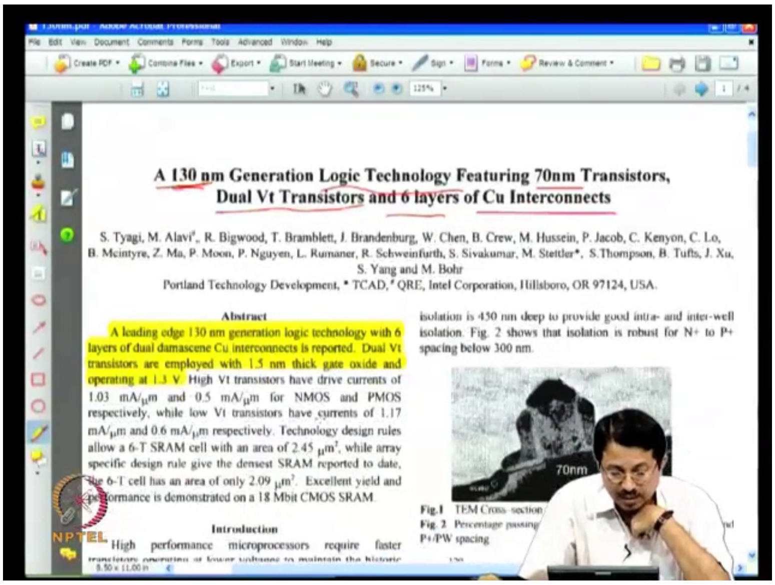
[1] M. Rodder, et al., *IEDM Tech. Dig*
[2] I. Yang, et al., *Symposium VLSI Tec* p. 148.
[3] L. Su, et al., *Symposium VLSI Techno* p. 12.
[4] F. Assaderaghi, et al., *IEDM Tech. D*
[5] D. Schepis, et al., *IEDM Tech.*
[6] M. Bohr, et al., *IEDM Tech*
[7] D. Edelstein, et al., *IEDM*
[8] H. Oyamatsu, et al., *IEE*

And using all these, they have been able to show a working SRAM and this is what they are sort of summarizing; 180 nanometer generation logic technology has been developed and demonstrated with high performance, reduce power transistors; aluminum interconnects with low Epsilon silicon; oxide with fluorine doping dielectrics are used to meet interconnect density and performance requirement.

The technology yield and performance capabilities have been demonstrated on a 16 megabit SRAM, which operates at greater than 900 mega hertz frequency. So, eventually if you have to build a microprocessor; if the cache has to respond at that frequency, it would respond at that frequency.

So, they have been able to start from the basic semiconductor processes; using all the process integration, now build the transistors, demonstrate a simple circuit like ring oscillator, characterize the gate delay and take it forward to make an SRAM cell and show that the SRAM is working. So, this is really a demonstration of a technology, but still this technology is still not; after this it may take another year for you to make a real product. Because your real product is going to be a microprocessor subsequently, then you will have to make a microprocessor, make sure that yield is good.

(Refer Slide Time: 24:59)



A 130 nm Generation Logic Technology Featuring 70nm Transistors, Dual Vt Transistors and 6 layers of Cu Interconnects

S. Tyagi, M. Alavi[#], R. Bigwood, T. Bramblett, J. Brandenburg, W. Chen, B. Crew, M. Hussein, P. Jacob, C. Kenyon, C. Lo, B. Mcintyre, Z. Ma, P. Moon, P. Nguyen, L. Rumaner, R. Schweinfurth, S. Sivakumar, M. Stettler*, S.Thompson, B. Tufts, J. Xu, S. Yang and M. Bohr
Portland Technology Development, * TCAD, # QRE, Intel Corporation, Hillsboro, OR 97124, USA.

Abstract

A leading edge 130 nm generation logic technology with 6 layers of dual damascene Cu interconnects is reported. Dual Vt transistors are employed with 1.5 nm thick gate oxide and operating at 1.3 V. High Vt transistors have drive currents of 1.03 mA/μm and 0.5 mA/μm for NMOS and PMOS respectively, while low Vt transistors have currents of 1.17 mA/μm and 0.6 mA/μm respectively. Technology design rules allow a 6-T SRAM cell with an area of 2.45 μm^2, while array specific design rule give the densest SRAM reported to date, the 6-T cell has an area of only 2.09 μm^2. Excellent yield and performance is demonstrated on a 18 Mbit CMOS SRAM.

Introduction

High performance microprocessors require faster transistors operating at lower voltages to maintain the historic

isolation is 450 nm deep to provide good intra- and inter-well isolation. Fig. 2 shows that isolation is robust for N+ to P+ spacing below 300 nm.

Fig.1 TEM Cross-section

Fig. 2 Percentage passing nd P+/PW spacing

And that requires lot of optimization; so, let us now look at the subsequent technology which is the 130 nanometer generation logic. Again logic technology because Intel does not make standalone memories, the only product they offer and that is; their essentially lifeline is microprocessors, very high end microprocessors. And hence this is the logic technology, not a memory technology again.

Now 70 nanometer transistors; although it is a 130 nanometer technology; you see the gate length is 70 nanometer. They have dual Vt transistor; remember something that we discussed, there is a leakage power problem, you need to really make sure that you minimize the leakage power; at the same time you need high performance. So, what you do? It turns out there are only is few critical parts, in any complicated chip.

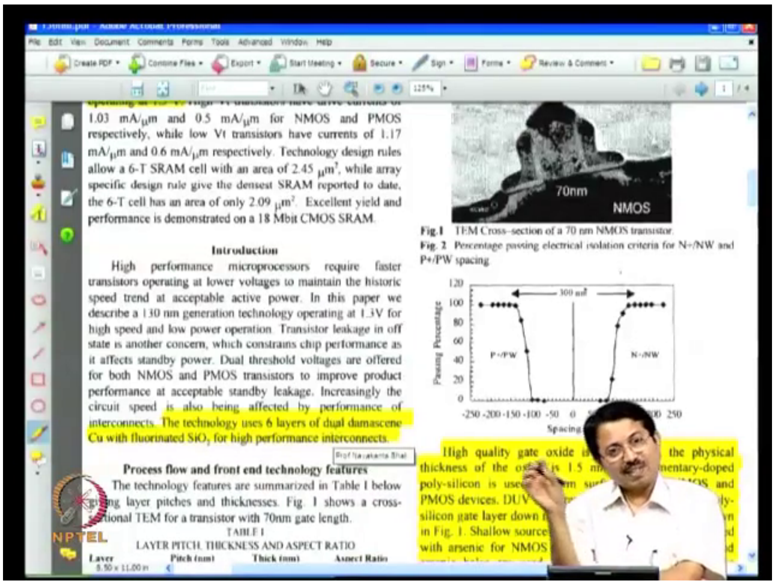
It is important to make the transistors in the critical part very fast transistors and hence make only those transistors a low Vt transistors and others could be high Vt transistors. As a result of that you get very fast chip as well as low power consuming chip. So, in other words here; you have two flavors of N channel transistors and two flavors of P channel transistors. One is called the low Vt transistor and the other one is called a high Vt transistor.

And hence it is a dual Vt transistor technology and what else? 6 layers copper interconnects, a big departure from aluminum to copper. Because this is the only way to scale the interconnects, very briefly may be we will look at the issues and interconnect scaling; now let us get started what they have here. A leading edge 130 nanometer generation logic technology with 6 layers of dual damascene copper interconnect. Dual damascene is a particular process sequence which is used to realize copper interconnects.

We will not really go over the details of that process sequence; dual Vt transistors are employed with 1.5 nanometer thick gate oxide, the gate oxide thickness as gone down compared to what it was in the previous generation; operating at 1.3 volt; previously it was 1.5 maximum; now it is 1.3; you see. High Vt transistors have drive currents of 1 milliampere and 0.5 milliampere per micron; for NMOS and PMOS.

While low Vt transistors have 1.17 for NMOS and 0.6 for PMOS respectively; so two flavors of transistors. Again they have been able to make an SRAM cell, demonstrate a SRAM cell; demonstrate a technology on a 18 bit SRAM.

(Refer Slide Time: 27:56)



1.03 mA/μm and 0.5 mA/μm for NMOS and PMOS respectively, while low Vt transistors have currents of 1.17 mA/μm and 0.6 mA/μm respectively. Technology design rules allow a 6-T SRAM cell with an area of 2.45 μm^2, while array specific design rule give the densest SRAM reported to date, the 6-T cell has an area of only 2.09 μm^2. Excellent yield and performance is demonstrated on a 18 Mbit CMOS SRAM.

Introduction

High performance microprocessors require faster transistors operating at lower voltages to maintain the historic speed trend at acceptable active power. In this paper we describe a 130 nm generation technology operating at 1.3V for high speed and low power operation. Transistor leakage in off state is another concern, which constrains chip performance as it affects standby power. Dual threshold voltages are offered for both NMOS and PMOS transistors to improve product performance at acceptable standby leakage. Increasingly the circuit speed is also being affected by performance of interconnects. The technology uses 6 layers of dual damascene Cu with fluorinated SiO_2 for high performance interconnects.

Process flow and front end technology features

The technology features are summarized in Table I below giving layer pitches and thicknesses. Fig. 1 shows a cross-sectional TEM for a transistor with 70nm gate length.

TABLE I

LAYER PITCH, THICKNESS AND ASPECT RATIO

Layer	Pitch (nm)	Thick (nm)	Aspect Ratio

Fig.1 TEM Cross-section of a 70 nm NMOS transistor.

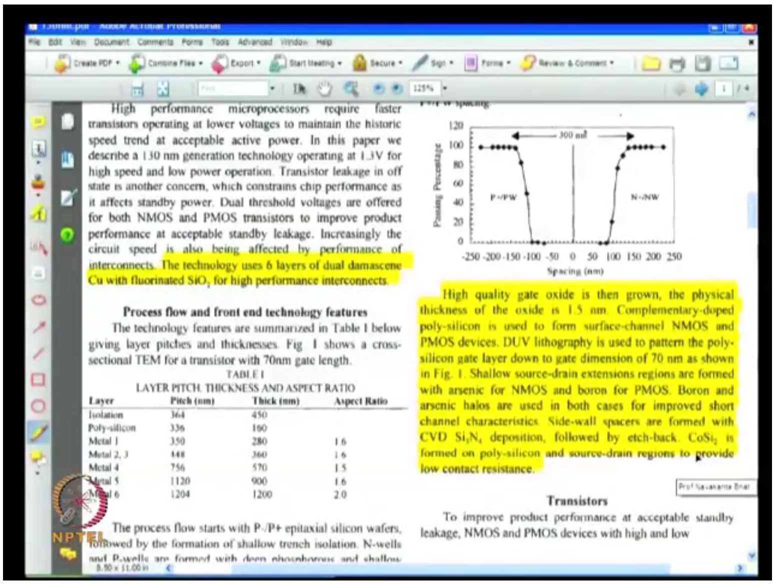
Fig. 2 Percentage passing electrical isolation criteria for N+/NW and P+/PW spacing

High quality gate oxide is … the physical thickness of the ox… is 1.5 n… …mentary-doped poly-silicon is use… …m sur… …MOS and PMOS devices. DUV … …ly-silicon gate layer down … …n in Fig. 1. Shallow source … …d with arsenic for NMOS … …d

The typical process flow is the same thing; except that you have a two dual Vt; meaning you will have to create two different N wells, two different P wells. So, that doping concentration is little different in two N wells and little different in two P wells and that is how you will get dual Vt technology; for N and P, which means more photolithography step, little more complex technology. But that is worth it; only if you do that, you would be able to get the best of it.

So various technological features; this is how a transistor looks; N channel transmission electron, micrograph of showing 70 nanometer gate length; source drain, ultra thin gate electrode, polysilicons, silicide at the top and so on and so forth.

(Refer Slide Time: 28:52)



High performance microprocessors require faster transistors operating at lower voltages to maintain the historic speed trend at acceptable active power. In this paper we describe a 130 nm generation technology operating at 1.3V for high speed and low power operation. Transistor leakage in off state is another concern, which constrains chip performance as it affects standby power. Dual threshold voltages are offered for both NMOS and PMOS transistors to improve product performance at acceptable standby leakage. Increasingly the circuit speed is also being affected by performance of interconnects. The technology uses 6 layers of dual damascene Cu with fluorinated SiO_2 for high performance interconnects.

Process flow and front end technology features

The technology features are summarized in Table I below giving layer pitches and thicknesses. Fig 1 shows a cross-sectional TEM for a transistor with 70nm gate length.

TABLE I
LAYER PITCH, THICKNESS AND ASPECT RATIO

Layer	Pitch (nm)	Thick (nm)	Aspect Ratio
Isolation	364	450	
Poly-silicon	336	160	
Metal 1	350	280	1.6
Metal 2, 3	448	360	1.6
Metal 4	756	570	1.5
Metal 5	1120	900	1.6
Metal 6	1204	1200	2.0

The process flow starts with P-/P+ epitaxial silicon wafers, followed by the formation of shallow trench isolation. N-wells and P-wells are formed with deep phosphorous and shallow

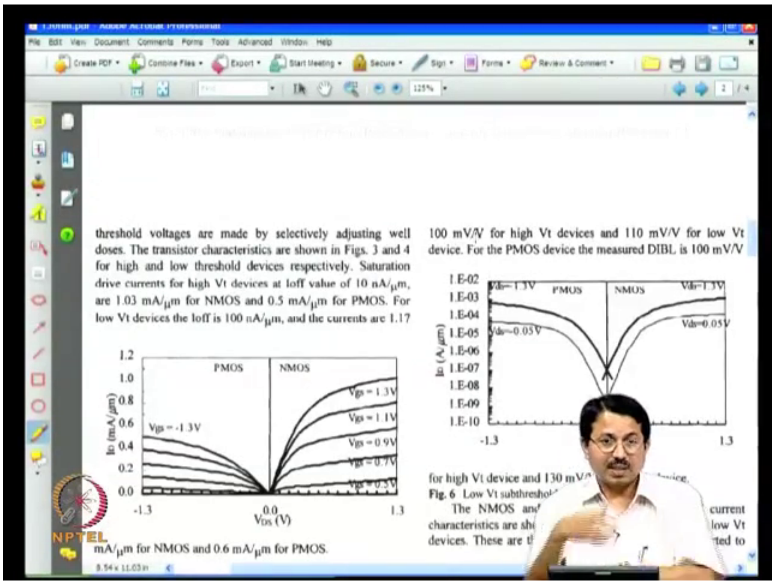
High quality gate oxide is then grown, the physical thickness of the oxide is 1.5 nm. Complementary-doped poly-silicon is used to form surface-channel NMOS and PMOS devices. DUV lithography is used to pattern the poly-silicon gate layer down to gate dimension of 70 nm as shown in Fig. 1. Shallow source-drain extensions regions are formed with arsenic for NMOS and boron for PMOS. Boron and arsenic halos are used in both cases for improved short channel characteristics. Side-wall spacers are formed with CVD Si_3N_4 deposition, followed by etch-back. $CoSi_2$ is formed on poly-silicon and source-drain regions to provide low contact resistance.

Transistors

To improve product performance at acceptable standby leakage, NMOS and PMOS devices with high and low

Or else will remain identical; you will use halo; you will use shallow extension except that you may change the dose; because the transistor gate length is going down especially halo dose, you may want to manipulate. Shallow source drain extension regions are formed with arsenic for NMOS, boron for PMOS. Boron and arsenic halos are used; this will remain same, silicon nitride is used for the spacer; correct? Followed by etch-back; the silicide here is cobalt silicide as a (Refer Time: 29:28) titanium silicide.

(Refer Slide Time: 29:36)



threshold voltages are made by selectively adjusting well doses. The transistor characteristics are shown in Figs. 3 and 4 for high and low threshold devices respectively. Saturation drive currents for high Vt devices at Ioff value of 10 nA/μm, are 1.03 mA/μm for NMOS and 0.5 mA/μm for PMOS. For low Vt devices the Ioff is 100 nA/μm, and the currents are 1.17 mA/μm for NMOS and 0.6 mA/μm for PMOS.

100 mV/V for high Vt devices and 110 mV/V for low Vt device. For the PMOS device the measured DIBL is 100 mV/V for high Vt device and 130 mV

Fig. 6 Low Vt subthreshol

The NMOS and current characteristics are sh low Vt devices. These are t ted to

Cobalt silicide gives you much better, lower resistance compared to titanium silicide. Again the demonstrator well behaved NMOS and PMOS; it is the story looks similar right, you will see similar graphs here except that now you have a high Vt and low Vt; two flavors.

And sub threshold plot, drain current in a log scale versus gate voltage; for low drain voltage and high drain voltage. And the DIBL for the 70 nanometer NMOS device is measured to be 100 millivolt per volt that is how we specify the DIBL. What is the difference in Vt per unit voltage; on the drain; unit voltage change on the drain? And that is how they are characterizing the DIBL.

(Refer Slide Time: 30:26)

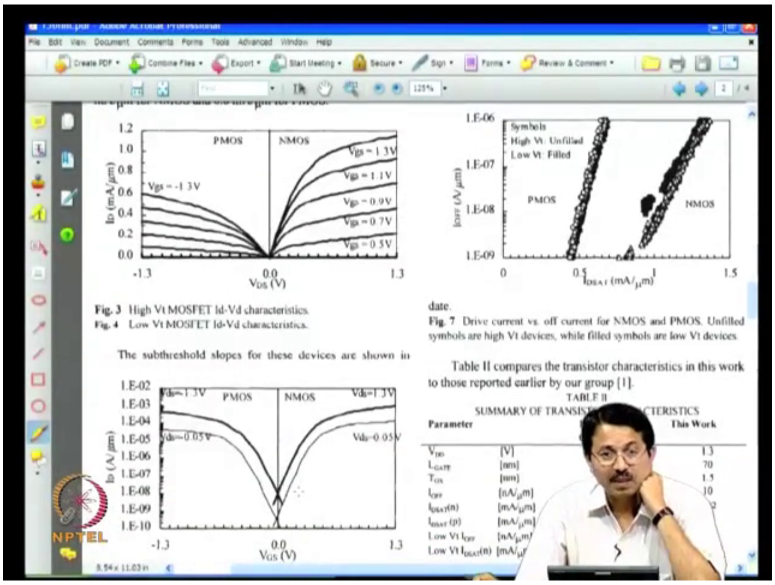


Fig. 3 High Vt MOSFET Id-Vd characteristics.
Fig. 4 Low Vt MOSFET Id-Vd characteristics.

The subthreshold slopes for these devices are shown in

date.

Fig. 7 Drive current vs. off current for NMOS and PMOS. Unfilled symbols are high Vt devices, while filled symbols are low Vt devices.

Table II compares the transistor characteristics in this work to those reported earlier by our group [1].

TABLE II
SUMMARY OF TRANSIS[illegible]CTERISTICS

Parameter		This Work
V_{DD}	[V]	1.3
L_{GATE}	[nm]	70
T_{OX}	[nm]	1.5
I_{OFF}	[nA/μm]	10
I_{DSAT}(n)	[mA/μm]	[illegible]
I_{DSAT} (p)	[mA/μm]	
Low Vt I_{OFF}	[nA/μm]	
Low Vt I_{DSAT}(n)	[mA/μ	

And this is again the typical off current versus on current and they are sort of saying that; this again two flavors unfilled and filled, high Vt and low Vt; there two flavors of PMOS and two flavors of N channel transistors; these are some important matrix.

(Refer Slide Time: 30:46)

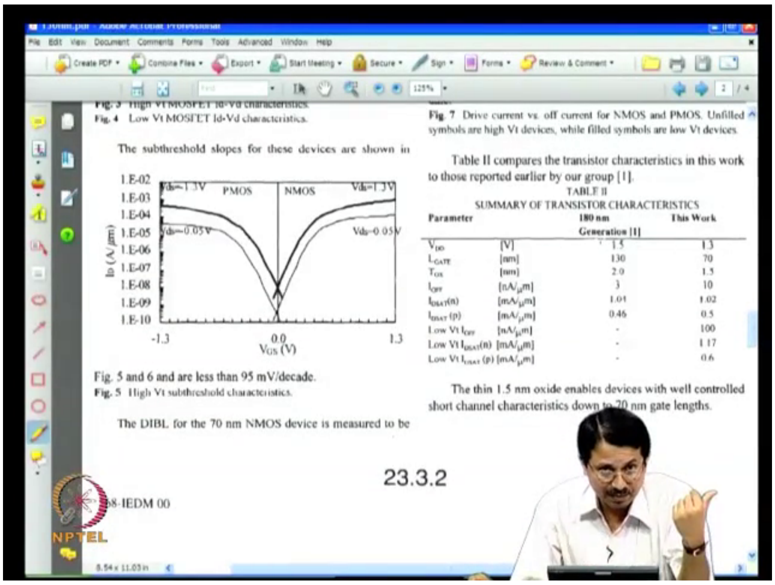


Fig. 4 Low Vt MOSFET Id-Vd characteristics.

The subthreshold slopes for these devices are shown in

Fig. 5 and 6 and are less than 95 mV/decade.

Fig. 5 High Vt subthreshold characteristics.

The DIBL for the 70 nm NMOS device is measured to be

23.3.2

8-IEDM 00

Fig. 7 Drive current vs. off current for NMOS and PMOS. Unfilled symbols are high Vt devices, while filled symbols are low Vt devices.

Table II compares the transistor characteristics in this work to those reported earlier by our group [1].

TABLE II

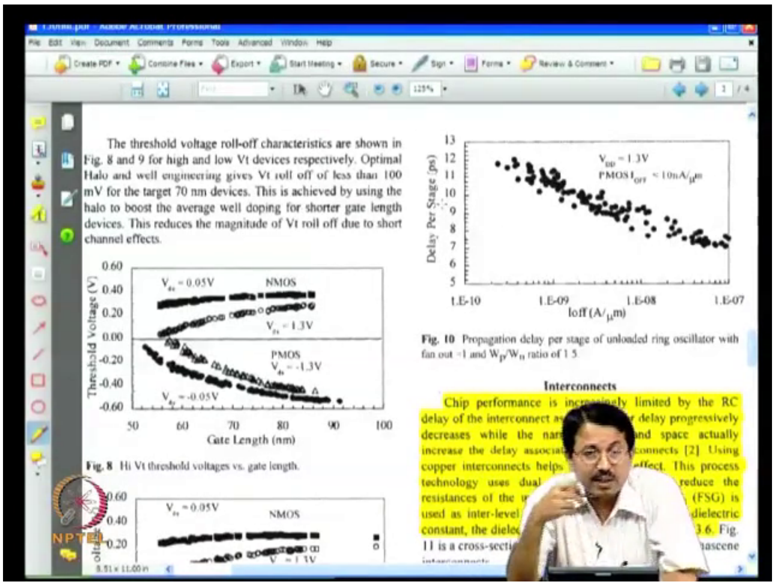
SUMMARY OF TRANSISTOR CHARACTERISTICS

Parameter		180 nm Generation [1]	This Work
V_{DD}	[V]	1.5	1.3
L_{GATE}	[nm]	130	70
T_{OX}	[nm]	2.0	1.5
I_{OFF}	[nA/μm]	3	10
I_{DSAT}(n)	[mA/μm]	1.01	1.02
I_{DSAT} (p)	[mA/μm]	0.46	0.5
Low Vt I_{OFF}	[nA/μm]	-	100
Low Vt I_{DSAT}(n)	[mA/μm]	-	1.17
Low Vt I_{DSAT} (p)	[mA/μm]	-	0.6

The thin 1.5 nm oxide enables devices with well controlled short channel characteristics down to 20 nm gate lengths.

Such as VDD, gate length, oxide thickness; they are comparing 180 nanometer generation, which is a previous paper that we discussed and this generation; supply voltage from 1.5 to 1.3, gate length from 130 to 70, oxide thickness from 2 nanometer to 1.5 nanometer and so and so forth.

(Refer Slide Time: 31:13)



The threshold voltage roll-off characteristics are shown in Fig. 8 and 9 for high and low Vt devices respectively. Optimal Halo and well engineering gives Vt roll off of less than 100 mV for the target 70 nm devices. This is achieved by using the halo to boost the average well doping for shorter gate length devices. This reduces the magnitude of Vt roll off due to short channel effects.

Fig. 8 Hi Vt threshold voltages vs. gate length.

Fig. 10 Propagation delay per stage of unloaded ring oscillator with fan out =1 and W_p/W_n ratio of 1.5.

Interconnects

Chip performance is inc[illegible]ly limited by the RC delay of the interconnect a[illegible] delay progressively decreases while the nar[illegible]nd space actually increase the delay associat[illegible]connects [2]. Using copper interconnects helps [illegible]effect. This process technology uses dual [illegible] reduce the resistances of the [illegible] (FSG) is used as inter-level [illegible] dielectric constant, the dielec[illegible] 3.6. Fig. 11 is a cross-secti[illegible]ascene

You see off current has increased a little bit; from 3 to 10 nano ampere per micron. Typical threshold voltage role of; threshold voltage versus gate length, N channel P channel; N channel positive and P channel negativity; low drain voltage, high drain voltage; this is because of DIBL. This is now high Vt and this is the low Vt transistor.

So, again they characterize delay; they build ring oscillator. Now they have depicted in slightly different fashion, in the previous paper; you saw delay plotted against gate length. Here, delay is plotted against off state current of a transistor; remember off state current is inversely related to gate length. In other words, if you have to plot gate length here; this would have been a lower gate length, may be 100 nanometer here, may be 200 nanometer here, may be 500 nanometer here.

500 nanometer will also have; I am sorry, I got it the other way around; this would be 100 nanometer, let me see if I can delete this; let me just take it away; that is ok.

(Refer Slide Time: 32:40)

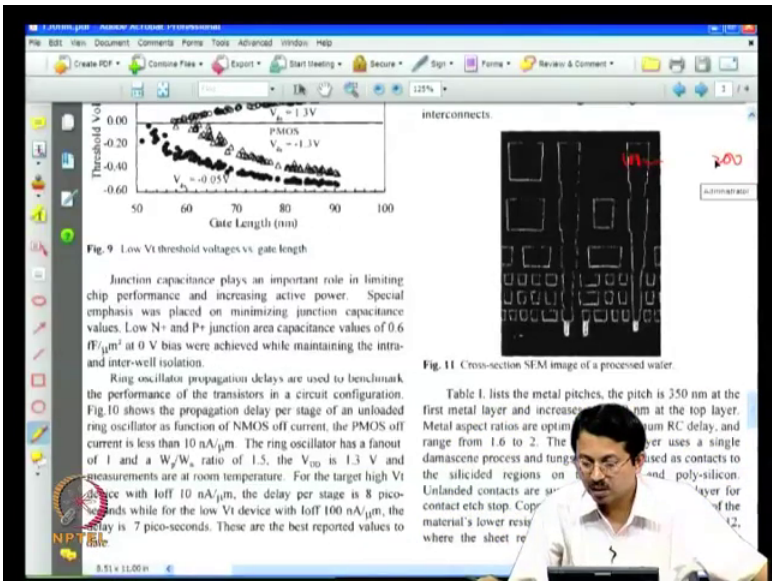


Fig. 9 Low Vt threshold voltages vs. gate length

Junction capacitance plays an important role in limiting chip performance and increasing active power. Special emphasis was placed on minimizing junction capacitance values. Low N+ and P+ junction area capacitance values of 0.6 fF/$_{\mu}$m^2 at 0 V bias were achieved while maintaining the intra- and inter-well isolation.

Ring oscillator propagation delays are used to benchmark the performance of the transistors in a circuit configuration. Fig.10 shows the propagation delay per stage of an unloaded ring oscillator as function of NMOS off current, the PMOS off current is less than 10 nA/$_{\mu}$m. The ring oscillator has a fanout of 1 and a W_p/W_n ratio of 1.5, the V_{DD} is 1.3 V and measurements are at room temperature. For the target high Vt device with Ioff 10 nA/$_{\mu}$m, the delay per stage is 8 pico-seconds while for the low Vt device with Ioff 100 nA/$_{\mu}$m, the delay is 7 pico-seconds. These are the best reported values to date.

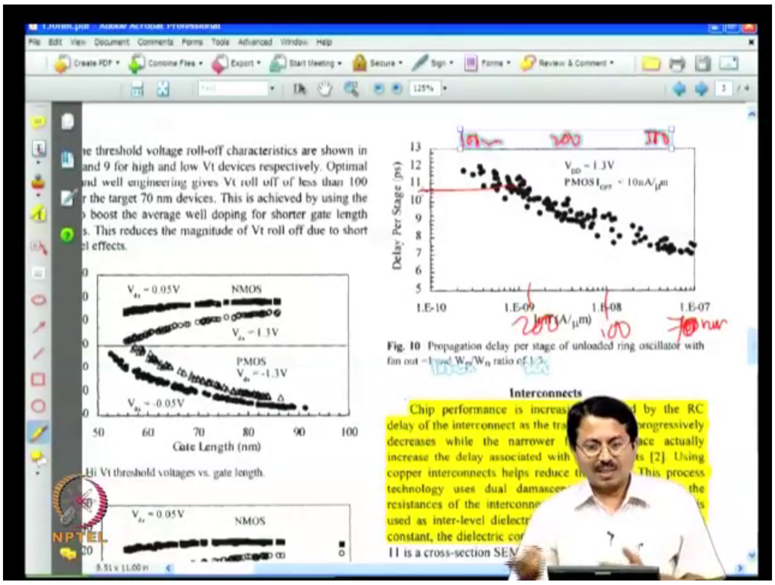
interconnects.

Fig. 11 Cross-section SEM image of a processed wafer.

Table I. lists the metal pitches, the pitch is 350 nm at the first metal layer and increases nm at the top layer. Metal aspect ratios are optim um RC delay, and range from 1.6 to 2. The yer uses a single damascene process and tungs used as contacts to the silicided regions on and poly-silicon. Unlanded contacts are s layer for contact etch stop. Cop f the material's lower resis 12, where the sheet re

So, this needs to be sort of reverse; 100 nanometer is out here, 100 nanometer also results in large leakage current; longer channel length also results in small leakage current correct, I mean let me write it here so, that becomes very clear.

(Refer Slide Time: 33:09)



he threshold voltage roll-off characteristics are shown in and 9 for high and low Vt devices respectively. Optimal nd well engineering gives Vt roll off of less than 100 r the target 70 nm devices. This is achieved by using the boost the average well doping for shorter gate length s. This reduces the magnitude of Vt roll off due to short el effects.

Hi Vt threshold voltages vs. gate length.

Fig. 10 Propagation delay per stage of unloaded ring oscillator with fan out =1 and W_p/W_n ratio of 1.5.

Interconnects

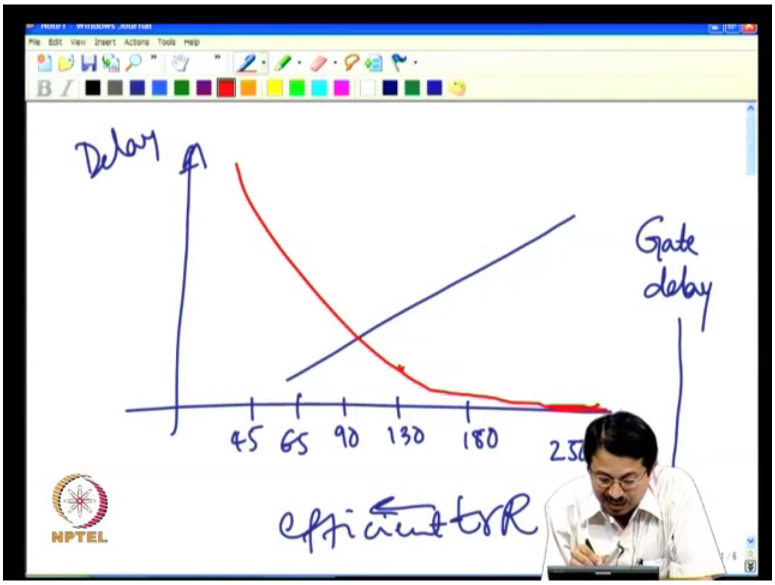
Chip performance is increasi… by the RC delay of the interconnect as the tra… progressively decreases while the narrower … ace actually increase the delay associated with … ts [2]. Using copper interconnects helps reduce t… This process technology uses dual damascen… the resistances of the interconne… is used as inter-level dielectr… constant, the dielectric co… 11 is a cross-section SEM

So, this is 100 nanometer transistor, let us say 100 or 100; 70 nanometer in this case, they have gone down to 70 nanometer. So, this would be something like 70 nanometer out here and this may be 100 nanometer and this may be 200 nanometer. So, the gate lengths is a lowest here and the lowest gate length gives the lowest delay.

And when the gate length is increasing, the leakage current also decreases as is evident here; lower leakage current. Because you have larger gate length, the on current will also be lower; just as off current is decreasing, on current is also decreasing and hence you will get larger delays. It is essentially looking at this delay versus gate length or delay versus leakage current; one and the same essentially.

So, chip performance is increasingly limited by the RC delay of interconnect; now why is that?

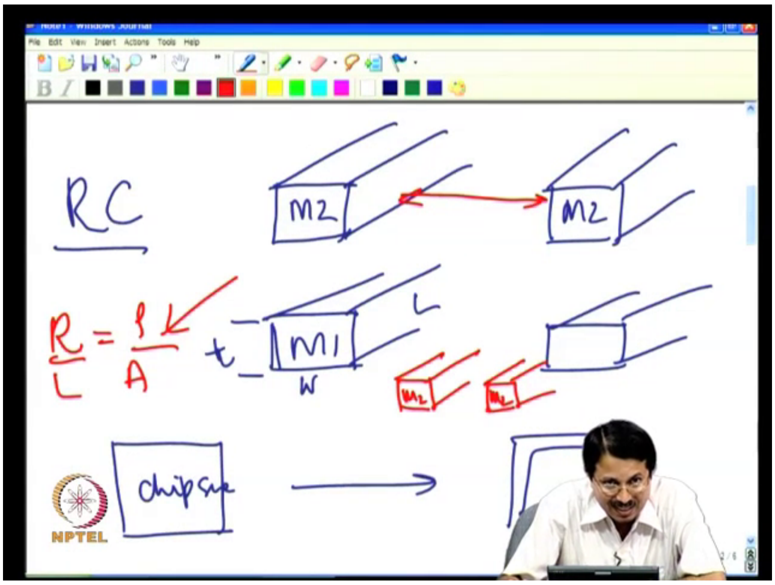
(Refer Slide Time: 34:10)



Let us look at what is happening with scaling; what have we done over scaling? We have started from let us say technology of the order of the 250 nanometer, quarter micron technology 180 nanometer, 130; 90, 65, 45 and so on and so forth. And let us look at what happens to delay; we know that as I scale transistor, transistor become very efficient. In other words, if we are talking of gate delay or inverter delay; over the technology generation, inverter delay or gate delay has been decreasing and that is why in the first place we want to scale.

You build any gate, it could be a not gate and gate or what have you right. So, called gate delay decreases; gate delay decreases because remember? Gate delay depends on transistors and transistors with scaling become efficient, smaller transistors are always better and the so, called CV or I metric become better; as I starts scaling down the transistor. What is the story with interconnects? Interconnects, it turns out the story is completely opposite unfortunately.

(Refer Slide Time: 35:37)



What is an interconnect? Interconnect is a metal line, let us say this is metal 2 and there may be another neighboring metal 2 line here; this is another line running at the same level. And let us say; there is another metal one underneath M 1; between these two, you have inter level dielectric, two metal levels between that you have a dielectric; that is how you have insulated; this is metal 1.

What is scaling of interconnects? Scaling of interconnects is scaling the dimensions of the interconnects; just as scaling of the transistor, dimension has to come down and transistor distance has to come down. And same here with the interconnects, the interconnect; their dimension has to come down, what is the dimension? You see; this is the so called line width; we call it line width.

We print the metal line; metal line is a long metal line; how long it is; depends on where is this metal going in a chip from one location to the other location in a chip. This is the thickness; let us say t and this is length, let us normalize the length. Unfortunately, the chips are not becoming small; the chips are really becoming bigger and bigger. You would have thought that with scaling because you are trying to bring down the number of transistor, the chip should become smaller. That is why we say at 0.7 x; so, that the chip area comes down, transistor area comes down and all that. But what has happened over the years is that, with scaling we also put in more transistors in a new technology; we are not just happy with the performers gained that come just because of the mere scaling.

We want to increase a number of transistors very significantly compared to what it was earlier. So, if previously you had million transistor, you want to put 4 million transistor in a new chip, in a new technology. And hence you continue to use the same area or even bigger area. So, in other words if you are looking at a chip; the chip size from one technology to the other technology really does not scale; if anything it is becoming bigger.

The chips are becoming bigger; what is the interconnect? Interconnect is essentially a metal line which takes the signal from one location to the other location in a chip you see. The message here is that, the lengthwise it is not really helping us; it is really becoming as big. So, let us look at normalized length what happens with interconnects scaling; what is interconnect delay? Interconnect delay depends on R and C of the interconnect; this is the R C delay; due to interconnects.

Scaling, I need to scale the width, I need to scale t; if I go from this technology to a new technology of metal interconnect, in that the thickness has come down, the width has come down. What is the implication? Resistance goes up; because your cross section area for conduction is going down. So, per unit length; you see, if you were to look at ask the question what is resistance per unit length? What is rho L by A? Let us say I am interested in R per L; it is rho by A; how is A scaling?

A has t and W; as a result your interconnect resistance per unit length goes up as k square; which is not a good thing, this is not the end of the story. So, the message is that interconnect lines are becoming more resistant with scaling; not the end of the story, there is even more serious problem and what is that serious problem? You would have thought that at least because I am making interconnect smaller, the capacitance should go down.

Because capacitance should be due to; let us say, this metal line and another metal line; there is a coupling capacitance and the area for that coupling is t; I mean this that is determine by this width, that is coming down and hence it should go down.

Yes this so called parallel plate capacitance or inter metal capacitance indeed goes down, but there is a more serious issue and what is that? I am also decreasing the distance between the two metal lines in the new generation of the technology. So, the new generation of the technology has M 2 lines; which were far apart, have been brought very close to each other.

So, there is a new capacitive component; which is intralayer capacitance. Earlier this distance was very large, you could have ignored that capacitance. Now this is increasing very dramatically, so net result is that with scaling your R is increasing, your C is also increasing; as per as interconnects are concerned.

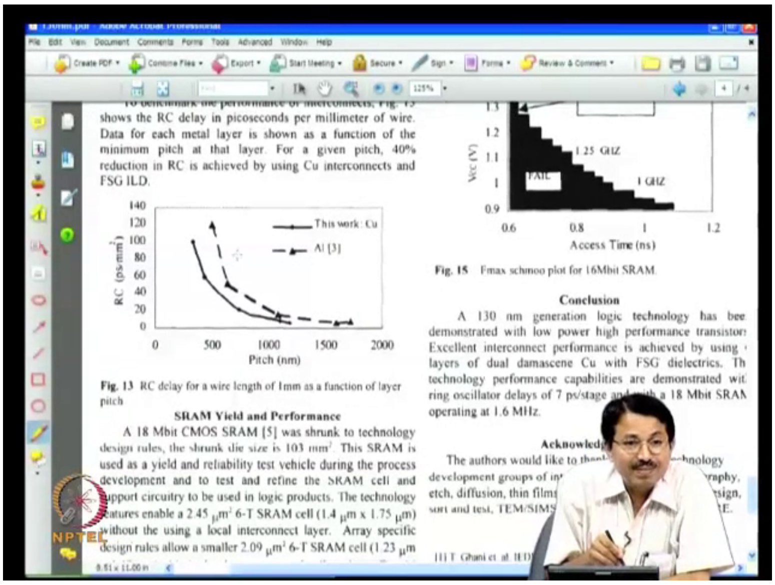
C is increasing very dramatically because these interconnects lines are coming close to each other. And hence the R C delay of interconnect is increasing. So, if you were to look at interconnects, in older generation technology; 250 nanometer and beyond, interconnects were afterthought; you never bothered about interconnects. Their ideal metal lines takes signal from one location to other location, they do not contribute any further delay.

So, it was nonexistent compared to gate delay; your interconnect delay was very very low. But what has happened in the recent past is something like this; your gate delays are increasing, your interconnect delays are increasing. And hence interconnects are becoming bottle X; it is like just imagine your road network in the city, you have the fastest driving car. And that is transistor for you; that is a gate for you, but you have a very bad road network.

Then you are limited; not by the speed of the car, but you are limited how fast the roads will allow you to travel. And these are the roads; interconnects are the road that allows you to take signals from one location to other location; and that has become a serious problem. And then you see that in this technology; there are lot things that are being done; to minimize R and to minimize C.

In previous paper; we saw fluorine was doped to minimize capacitance and today I mean in this technology, we see aluminum is replaced with copper because copper has lower resistance compared to aluminum. RC delay becomes much better and that is exactly what is done here. And they have optimized this fluorinated oxide and that gives a k of about 3.6 and this is again the hierarchy of interconnects.

(Refer Slide Time: 43:06)



shows the RC delay in picoseconds per millimeter of wire. Data for each metal layer is shown as a function of the minimum pitch at that layer. For a given pitch, 40% reduction in RC is achieved by using Cu interconnects and FSG ILD.

Fig. 13 RC delay for a wire length of 1mm as a function of layer pitch

SRAM Yield and Performance

A 18 Mbit CMOS SRAM [5] was shrunk to technology design rules, the shrunk die size is 103 mm². This SRAM is used as a yield and reliability test vehicle during the process development and to test and refine the SRAM cell and upport circuitry to be used in logic products. The technology eatures enable a 2.45 μm² 6-T SRAM cell (1.4 μm x 1.75 μm) without the using a local interconnect layer. Array specific design rules allow a smaller 2.09 μm² 6-T SRAM cell (1.23 μm

Fig. 15 Fmax schmoo plot for 16Mbit SRAM.

Conclusion

A 130 nm generation logic technology has bee demonstrated with low power high performance transistor: Excellent interconnect performance is achieved by using layers of dual damascene Cu with FSG dielectrics. Th technology performance capabilities are demonstrated wit ring oscillator delays of 7 ps/stage and with a 18 Mbit SRAN operating at 1.6 MHz.

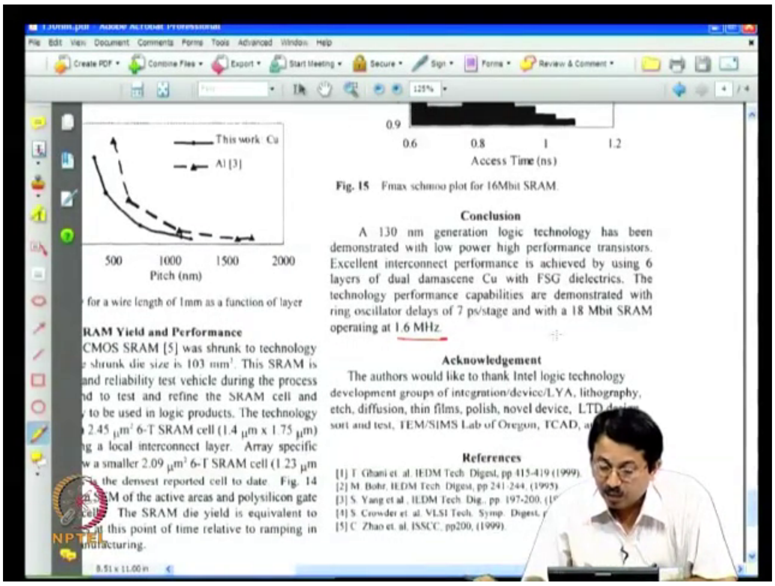
Acknowled

The authors would like to th nology development groups of in raphy, etch, diffusion, thin films sign, sort and test, TEM/SIMS E.

[1] T. Ghani et. al. IED

And they have used this to build; what they are showing here is that; if they use aluminum, you see with decreasing pitch, decreasing pitch is two metals are coming closer to each other. Your delays are increasing, if you are use aluminum; it would have gone up like this. Because you use copper, you could decrease it; at the same pitch, copper technology gives much lower delay compared to aluminum technology.

(Refer Slide Time: 43:42)



for a wire length of 1mm as a function of layer

RAM Yield and Performance

CMOS SRAM [5] was shrunk to technology shrunk die size is 103 mm². This SRAM is and reliability test vehicle during the process nd to test and refine the SRAM cell and to be used in logic products. The technology 2.45 μm² 6-T SRAM cell (1.4 μm x 1.75 μm) g a local interconnect layer. Array specific w a smaller 2.09 μm² 6-T SRAM cell (1.23 μm the densest reported cell to date. Fig. 14 M of the active areas and polysilicon gate cell. The SRAM die yield is equivalent to this point of time relative to ramping in nufacturing.

Fig. 15 Fmax schmoo plot for 16Mbit SRAM.

Conclusion

A 130 nm generation logic technology has been demonstrated with low power high performance transistors. Excellent interconnect performance is achieved by using 6 layers of dual damascene Cu with FSG dielectrics. The technology performance capabilities are demonstrated with ring oscillator delays of 7 ps/stage and with a 18 Mbit SRAM operating at 1.6 MHz.

Acknowledgement

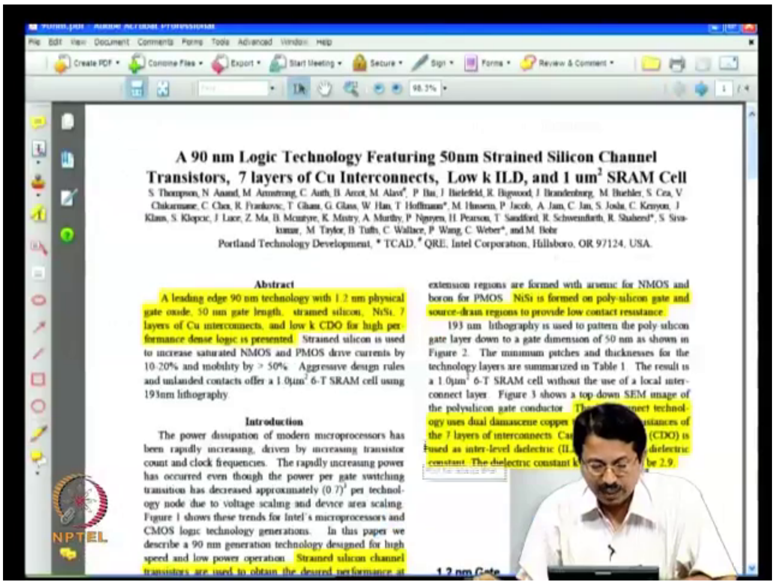
The authors would like to thank Intel logic technology development groups of integration/device/LYA, lithography, etch, diffusion, thin films, polish, novel device, LTD design, sort and test, TEM/SIMS Lab of Oregon, TCAD, a

References

[1] T. Ghani et. al. IEDM Tech. Digest, pp 415-419 (1999).
[2] M. Bohr, IEDM Tech. Digest, pp 241-244, (1995).
[3] S. Yang et al., IEDM Tech. Dig., pp. 197-200, (1
[4] S. Crowder et. al. VLSI Tech. Symp. Digest,
[5] C. Zhao et. al. ISSCC, pp200, (1999).

And that is what they have done and they have been able to make an SRAM and demonstrate that 18 bit SRAM is operating now at what; 1.6 mega hertz; much higher frequency. So, actually it is not very clear here, I take back what I said; I mean the previous one was a lower density SRAM was at 900 mega hertz, but this is a higher density SRAM; that is operating at little lower frequency. As you start increasing the density, there is also an impact on the speed of the SRAM that you can build.

(Refer Slide Time: 44:38)



A 90 nm Logic Technology Featuring 50nm Strained Silicon Channel Transistors, 7 layers of Cu Interconnects, Low k ILD, and 1 um^2 SRAM Cell

S. Thompson, N. Anand, M. Armstrong, C. Auth, B. Arcot, M. Alavi#, P. Bai, J. Bielefeld, R. Bigwood, J. Brandenburg, M. Buehler, S. Cea, V. Chikarmane, C. Chen, R. Frankovic, T. Ghani, G. Glass, W. Han, T. Hoffmann*, M. Hussein, P. Jacob, A. Jain, C. Jan, S. Joshi, C. Kenyon, J. Klaus, S. Klopcic, J. Luce, Z. Ma, B. Mcintyre, K. Mistry, A. Murthy, P. Nguyen, H. Pearson, T. Sandford, R. Schweinfurth, R. Shaheed*, S. Sivakumar, M. Taylor, B. Tufts, C. Wallace, P. Wang, C. Weber*, and M. Bohr
Portland Technology Development, * TCAD, # QRE, Intel Corporation, Hillsboro, OR 97124, USA

Abstract

A leading edge 90 nm technology with 1.2 nm physical gate oxide, 50 nm gate length, strained silicon, NiSi, 7 layers of Cu interconnects, and low k CDO for high performance dense logic is presented. Strained silicon is used to increase saturated NMOS and PMOS drive currents by 10-20% and mobility by > 50%. Aggressive design rules and unlanded contacts offer a 1.0μm^2 6-T SRAM cell using 193nm lithography.

Introduction

The power dissipation of modern microprocessors has been rapidly increasing, driven by increasing transistor count and clock frequencies. The rapidly increasing power has occurred even though the power per gate switching transition has decreased approximately $(0.7)^3$ per technology node due to voltage scaling and device area scaling. Figure 1 shows these trends for Intel's microprocessors and CMOS logic technology generations. In this paper we describe a 90 nm generation technology designed for high speed and low power operation. Strained silicon channel transistors are used to obtain the desired performance at

extension regions are formed with arsenic for NMOS and boron for PMOS. NiSi is formed on poly-silicon gate and source-drain regions to provide low contact resistance.

193 nm lithography is used to pattern the poly-silicon gate layer down to a gate dimension of 50 nm as shown in Figure 2. The minimum pitches and thicknesses for the technology layers are summarized in Table 1. The result is a 1.0μm^2 6-T SRAM cell without the use of a local interconnect layer. Figure 3 shows a top down SEM image of the polysilicon gate conductor. Th[illegible]nect technology uses dual damascene copper [illegible]stances of the 7 layers of interconnects. Ca[illegible] (CDO) is used as inter-level dielectric (IL[illegible] dielectric constant. The dielectric constant k[illegible] be 2.9.

1.2 nm Gate

Whereas the ring oscillator; obviously, are much more efficient compared to the basic technology is much better. Previous generation was 10 picosecond; 10 to 13 picosecond, now it is a 7 pico second. So, you can clearly see a better technology and if you go down to 90 nanometer; now what has happened now? 7 layers, 6 layers have become 7 layers of copper; 90 nanometer has a 50 nanometer gate length transistors.

From 70, it came down to 50 nanometer now, it is using a strained silicon channel; we have not discussed about strained silicon channel, but very soon; down the line, you will start discussing strained silicon also. So, we will postpone that discussion for the time being. This uses 1.2 nanometer oxide, this is even thinner than previous generation and nickel silicide; instead of cobalt silicide, it is nickel silicide.

(Refer Slide Time: 45:26)

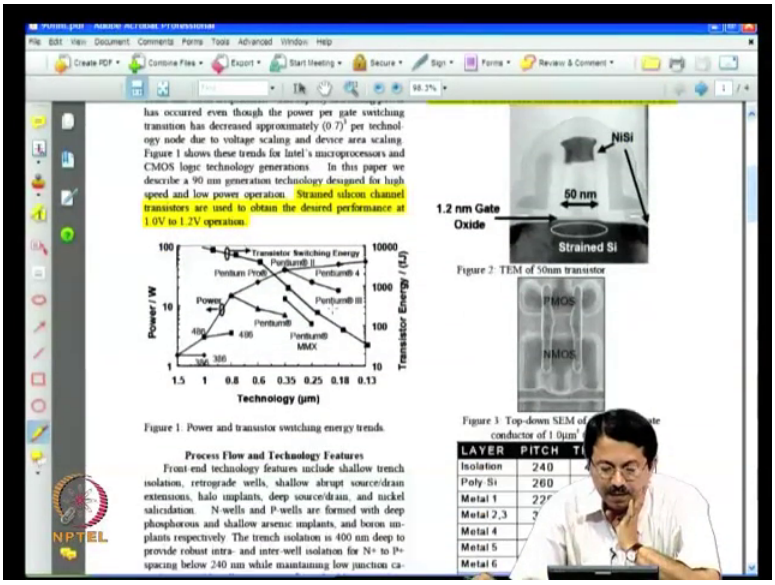

has occurred even though the power per gate switching transition has decreased approximately (0.7)³ per technology node due to voltage scaling and device area scaling. Figure 1 shows these trends for Intel's microprocessors and CMOS logic technology generations. In this paper we describe a 90 nm generation technology designed for high speed and low power operation. Strained silicon channel transistors are used to obtain the desired performance at 1.0V to 1.2V operation.

Figure 1: Power and transistor switching energy trends.

Figure 2: TEM of 50nm transistor

Figure 3: Top-down SEM of

conductor of 1.0µm²

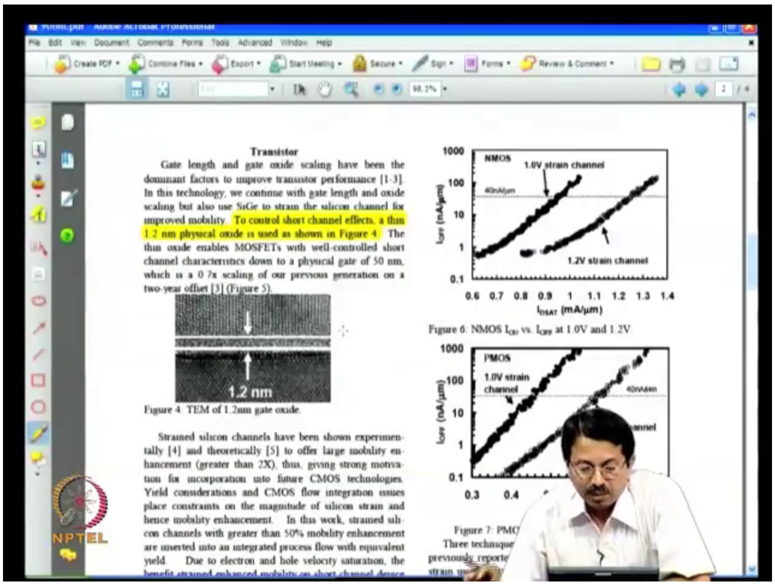
LAYER	PITCH	T
Isolation	240	4
Poly Si	260	
Metal 1	22	
Metal 2,3	3	
Metal 4		
Metal 5		
Metal 6		

Process Flow and Technology Features

Front-end technology features include shallow trench isolation, retrograde wells, shallow abrupt source/drain extensions, halo implants, deep source/drain, and nickel salicidation. N-wells and P-wells are formed with deep phosphorous and shallow arsenic implants, and boron implants respectively. The trench isolation is 400 nm deep to provide robust intra- and inter-well isolation for N+ to P+ spacing below 240 nm while maintaining low junction ca-

Now, there is a some more optimization and so, this is the cross section of a transistor; channel is strained silicon 1.2 nanometer gate oxide, 50 nanometer gate length, nickel silicide is shown here at the source drain and at the polysilicon. And they have been able to make a 60 SRAM cell using the technology.

(Refer Slide Time: 45:48)


Transistor

Gate length and gate oxide scaling have been the dominant factors to improve transistor performance [1-3]. In this technology, we continue with gate length and oxide scaling but also use SiGe to strain the silicon channel for improved mobility. To control short channel effects, a thin 1.2 nm physical oxide is used as shown in Figure 4. The thin oxide enables MOSFETs with well-controlled short channel characteristics down to a physical gate of 50 nm, which is a 0.7x scaling of our previous generation on a two-year offset [3] (Figure 5).

Figure 4: TEM of 1.2nm gate oxide.

Figure 6: NMOS I_{ON} vs. I_{OFF} at 1.0V and 1.2V.

Figure 7: PMO

Strained silicon channels have been shown experimentally [4] and theoretically [5] to offer large mobility enhancement (greater than 2X), thus, giving strong motivation for incorporation into future CMOS technologies. Yield considerations and CMOS flow integration issues place constraints on the magnitude of silicon strain and hence mobility enhancement. In this work, strained silicon channels with greater than 50% mobility enhancement are inserted into an integrated process flow with equivalent yield. Due to electron and hole velocity saturation, the benefit strained enhanced mobility on short channel device

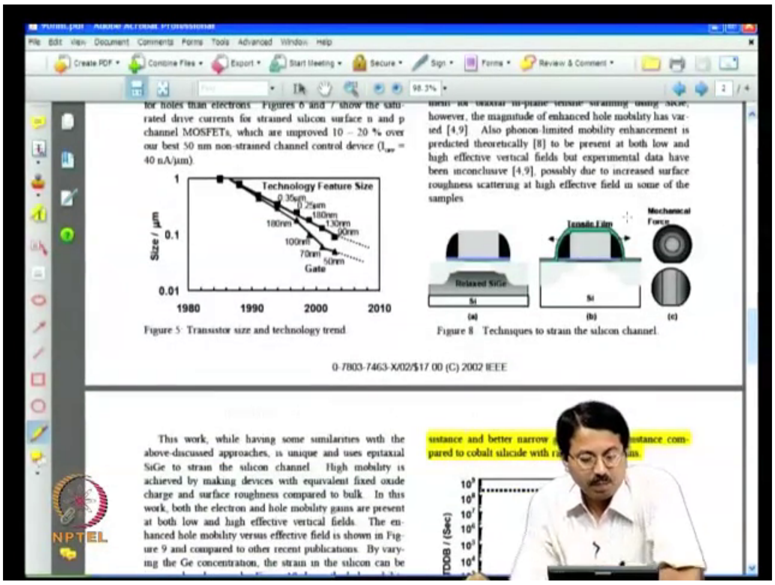
Three technique

previously reporte

strain u

This here just shows that you have silicon and 1.2 nanometer oxide and polysilicon gate and typical off current verses on current; for N and P channel transistors.

(Refer Slide Time: 46:03)



for holes than electrons. Figures 6 and 7 show the saturated drive currents for strained silicon surface n and p channel MOSFETs, which are improved 10 – 20 % over our best 50 nm non-strained channel control device (I_{off} = 40 nA/μm).

Figure 5: Transistor size and technology trend

however, the magnitude of enhanced hole mobility has varied [4,9]. Also phonon-limited mobility enhancement is predicted theoretically [8] to be present at both low and high effective vertical fields but experimental data have been inconclusive [4,9], possibly due to increased surface roughness scattering at high effective field in some of the samples.

Figure 8. Techniques to strain the silicon channel.

0-7803-7463-X/02/$17.00 (C) 2002 IEEE

This work, while having some similarities with the above-discussed approaches, is unique and uses epitaxial SiGe to strain the silicon channel. High mobility is achieved by making devices with equivalent fixed oxide charge and surface roughness compared to bulk. In this work, both the electron and hole mobility gains are present at both low and high effective vertical fields. The enhanced hole mobility versus effective field is shown in Figure 9 and compared to other recent publications. By varying the Ge concentration, the strain in the silicon can be

sistance and better narrow stance compared to cobalt silicide with ra ns.

And they are just showing here as a function of here; different technology nodes, we already looked at 180 nanometer technology, 130 nanometer in 180 nanometer; you had 100 nanometer, 130 nanometer, 70 nanometer gate length; 90 nanometer, 50 nanometer gate length; this is what I said, now we have a departure in terms of what gate length you print is actually smaller than the so called technology definition; 90 nanometer.

But in the past; they were exactly identical, we have had discussion on this; in the one of the earlier lectures. Why nickel silicide? Because with scaling; cobalt silicide will start increasing the resistance, whereas nickel silicide; gives very low resistance and that is why they have gone to nickel silicide.

(Refer Slide Time: 46:43)

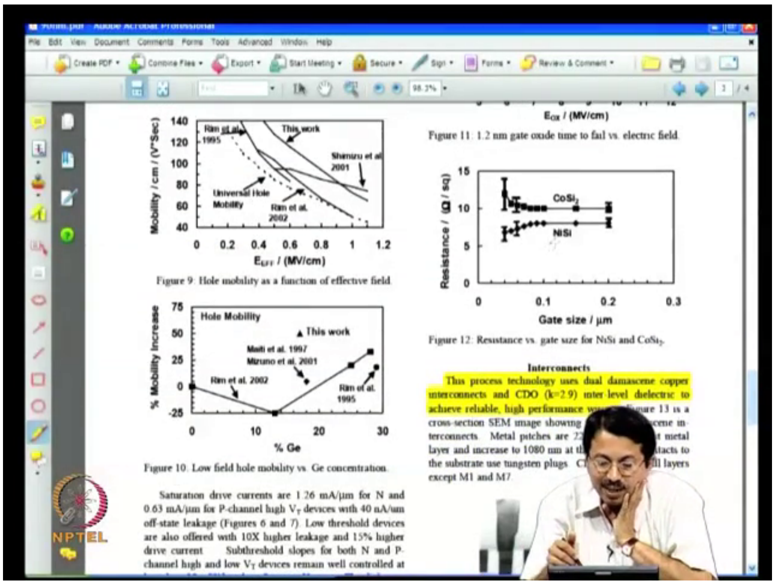


Figure 9: Hole mobility as a function of effective field.

Figure 10. Low field hole mobility vs. Ge concentration.

Saturation drive currents are 1.26 mA/µm for N and 0.63 mA/µm for P-channel high V_T devices with 40 nA/um off-state leakage (Figures 6 and 7). Low threshold devices are also offered with 10X higher leakage and 15% higher drive current. Subthreshold slopes for both N and P-channel high and low V_T devices remain well controlled at

Figure 11: 1.2 nm gate oxide time to fail vs. electric field.

Figure 12: Resistance vs. gate size for NiSi and $CoSi_2$.

Interconnects

This process technology uses dual damascene copper interconnects and CDO (k=2.9) inter-level dielectric to achieve reliable, high performance wi gure 13 is a cross-section SEM image showing cene interconnects. Metal pitches are 22 t metal layer and increase to 1080 nm at th tacts to the substrate use tungsten plugs. C ll layers except M1 and M7.

Again multi level metal interconnects; now instead of fluorine doped oxide, they have done carbon doped oxide.

(Refer Slide Time: 46:57)

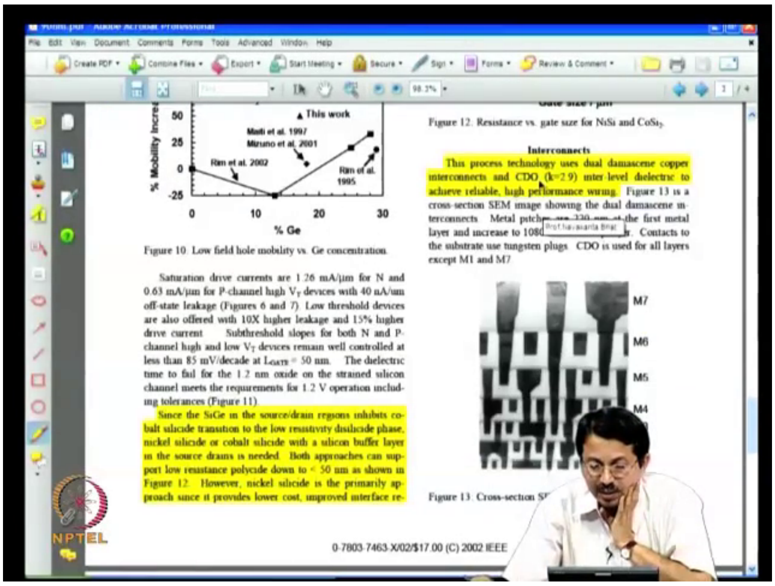


Figure 10. Low field hole mobility vs. Ge concentration.

Saturation drive currents are 1.26 mA/µm for N and 0.63 mA/µm for P-channel high V_T devices with 40 nA/um off-state leakage (Figures 6 and 7). Low threshold devices are also offered with 10X higher leakage and 15% higher drive current. Subthreshold slopes for both N and P-channel high and low V_T devices remain well controlled at less than 85 mV/decade at L_{GATE} = 50 nm. The dielectric time to fail for the 1.2 nm oxide on the strained silicon channel meets the requirements for 1.2 V operation including tolerances (Figure 11).

Since the SiGe in the source/drain regions inhibits cobalt silicide transition to the low resistivity disilicide phase, nickel silicide or cobalt silicide with a silicon buffer layer in the source drains is needed. Both approaches can support low resistance polycide down to < 50 nm as shown in Figure 12. However, nickel silicide is the primarily approach since it provides lower cost, improved interface re-

Figure 12. Resistance vs. gate size for NiSi and $CoSi_2$.

Interconnects

This process technology uses dual damascene copper interconnects and CDO (k=2.9) inter-level dielectric to achieve reliable, high performance wiring. Figure 13 is a cross-section SEM image showing the dual damascene interconnects. Metal pitch the first metal layer and increase to 1080 r. Contacts to the substrate use tungsten plugs. CDO is used for all layers except M1 and M7.

Figure 13: Cross-section S

0-7803-7463-X/02/$17.00 (C) 2002 IEEE

I know again a new improvement; that is how you need newer and newer materials; k has come down to 2.9, instead of 3.5. Again to decrease the R C delay of your metal interconnects.

(Refer Slide Time: 47:18)

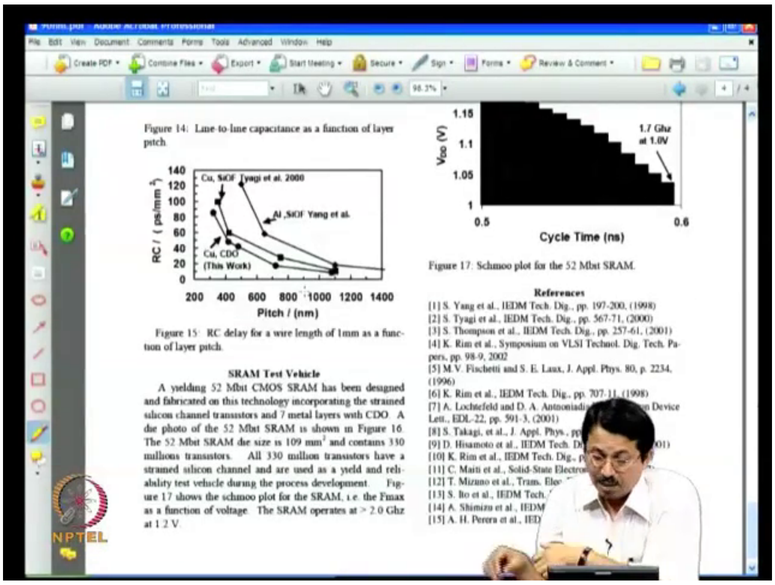


Figure 14: Line-to-line capacitance as a function of layer pitch

Figure 15: RC delay for a wire length of 1mm as a function of layer pitch.

SRAM Test Vehicle

A yielding 52 Mbit CMOS SRAM has been designed and fabricated on this technology incorporating the strained silicon channel transistors and 7 metal layers with CDO. A die photo of the 52 Mbit SRAM is shown in Figure 16. The 52 Mbit SRAM die size is 109 mm^2 and contains 330 millions transistors. All 330 million transistors have a strained silicon channel and are used as a yield and reliability test vehicle during the process development. Figure 17 shows the schmoo plot for the SRAM, i.e. the Fmax as a function of voltage. The SRAM operates at > 2.0 Ghz at 1.2 V.

Figure 17: Schmoo plot for the 52 Mbit SRAM.

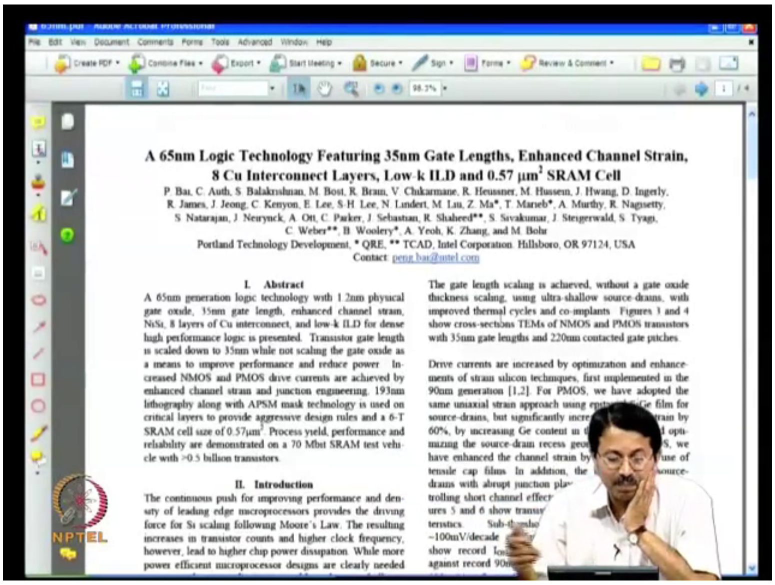
References

[1] S. Yang et al., IEDM Tech. Dig., pp. 197-200, (1998)
[2] S. Tyagi et al., IEDM Tech. Dig., pp. 567-71, (2000)
[3] S. Thompson et al., IEDM Tech. Dig., pp. 257-61, (2001)
[4] K. Rim et al., Symposium on VLSI Technol. Dig. Tech. Papers, pp. 98-9, 2002
[5] M.V. Fischetti and S. E. Laux, J. Appl. Phys. 80, p. 2234, (1996)
[6] K. Rim et al., IEDM Tech. Dig., pp. 707-11, (1998)
[7] A. Lochtefeld and D. A. Antnoniadis Device Lett., EDL-22, pp. 591-3, (2001)
[8] S. Takagi, et al., J. Appl. Phys., pp
[9] D. Hisamoto et al., IEDM Tech. D 01)
[10] K. Rim et al., IEDM Tech. Dig., p
[11] C. Maiti et al., Solid-State Electro
[12] T. Mizuno et al., Trans. Elec
[13] S. Ito et al., IEDM Tech.
[14] A. Shimizu et al., IEDM
[15] A. H. Perera et al., IED

And they again show that; with carbon doped oxide, you can get significantly lower RC delay as you start decreasing the pitch. As compared to aluminum with fluorine doped oxide, copper with fluorine doped oxide and this is copper with carbon doped oxide; which is much better. And they have been able to do an SRAM here, and this SRAM operates at 2 gigahertz; which is much higher frequency.

I suspect that the previous paper that probably should have been 1.6 giga hertz; that could be a typo there; rather than 1.6 mega hertz; now that is more like it. Because on 180 nanometer, you had a 900 mega hertz, very close to gigahertz and then you have 1.6 gigahertz and now you have about 2 gigahertz or little more than 2 gigahertz.

(Refer Slide Time: 48:51)



A 65nm Logic Technology Featuring 35nm Gate Lengths, Enhanced Channel Strain, 8 Cu Interconnect Layers, Low-k ILD and 0.57 µm^2 SRAM Cell

P. Bai, C. Auth, S. Balakrishnan, M. Bost, R. Brain, V. Chikarmane, R. Heussner, M. Hussein, J. Hwang, D. Ingerly, R. James, J. Jeong, C. Kenyon, E. Lee, S-H. Lee, N. Lindert, M. Liu, Z. Ma*, T. Marieb*, A. Murthy, R. Nagisetty, S. Natarajan, J. Neirynck, A. Ott, C. Parker, J. Sebastian, R. Shaheed**, S. Sivakumar, J. Steigerwald, S. Tyagi, C. Weber**, B. Woolery*, A. Yeoh, K. Zhang, and M. Bohr
Portland Technology Development, * QRE, ** TCAD, Intel Corporation. Hillsboro, OR 97124, USA
Contact: peng.bai@intel.com

I. Abstract

A 65nm generation logic technology with 1.2nm physical gate oxide, 35nm gate length, enhanced channel strain, NiSi, 8 layers of Cu interconnect, and low-k ILD for dense high performance logic is presented. Transistor gate length is scaled down to 35nm while not scaling the gate oxide as a means to improve performance and reduce power. Increased NMOS and PMOS drive currents are achieved by enhanced channel strain and junction engineering. 193nm lithography along with APSM mask technology is used on critical layers to provide aggressive design rules and a 6-T SRAM cell size of 0.57µm^2. Process yield, performance and reliability are demonstrated on a 70 Mbit SRAM test vehicle with >0.5 billion transistors.

II. Introduction

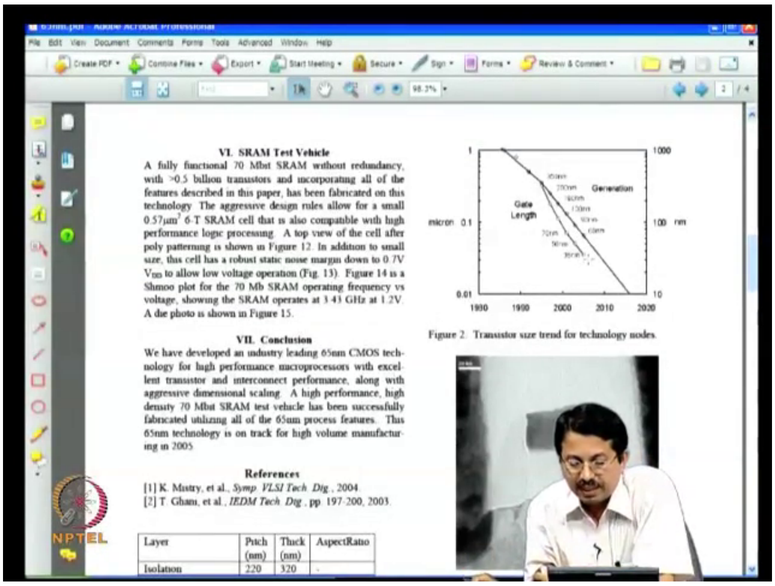
The continuous push for improving performance and density of leading edge microprocessors provides the driving force for Si scaling following Moore's Law. The resulting increases in transistor counts and higher clock frequency, however, lead to higher chip power dissipation. While more power efficient microprocessor designs are clearly needed

The gate length scaling is achieved, without a gate oxide thickness scaling, using ultra-shallow source-drains, with improved thermal cycles and co-implants. Figures 3 and 4 show cross-sections TEMs of NMOS and PMOS transistors with 35nm gate lengths and 220nm contacted gate pitches.

Drive currents are increased by optimization and enhancements of strain silicon techniques, first implemented in the 90nm generation [1,2]. For PMOS, we have adopted the same uniaxial strain approach using [illegible] Ge film for source-drains, but significantly incre[illegible] strain by 60%, by increasing Ge content in [illegible] opti-mizing the source-drain recess geo[illegible] S, we have enhanced the channel strain by [illegible] use of tensile cap films. In addition, the [illegible] source-drains with abrupt junction pla[illegible] trolling short channel effec[illegible] ures 5 and 6 show transi[illegible] teristics. Sub-th[illegible] ~100mV/decade [illegible] show record lo[illegible] against record 90[illegible]

So, this is about 90 nanometer and if you come to 65 nanometer; all that has happened really is gate length of scale down further; 35 nanometer, 8 levels of copper interconnects, interconnect hierarchy is also becoming more complex and SRAM cell is also shrinking. Continue to use nickel silicide. And let us see if there is anything; I want to highlight here.

(Refer Slide Time: 48:58)



VI. SRAM Test Vehicle

A fully functional 70 Mbit SRAM without redundancy, with >0.5 billion transistors and incorporating all of the features described in this paper, has been fabricated on this technology. The aggressive design rules allow for a small 0.57µm^2 6-T SRAM cell that is also compatible with high performance logic processing. A top view of the cell after poly patterning is shown in Figure 12. In addition to small size, this cell has a robust static noise margin down to 0.7V V_{DD} to allow low voltage operation (Fig. 13). Figure 14 is a Shmoo plot for the 70 Mb SRAM operating frequency vs voltage, showing the SRAM operates at 3.43 GHz at 1.2V. A die photo is shown in Figure 15.

VII. Conclusion

We have developed an industry leading 65nm CMOS technology for high performance microprocessors with excellent transistor and interconnect performance, along with aggressive dimensional scaling. A high performance, high density 70 Mbit SRAM test vehicle has been successfully fabricated utilizing all of the 65nm process features. This 65nm technology is on track for high volume manufacturing in 2005.

References

[1] K. Mistry, et al., *Symp. VLSI Tech. Dig.*, 2004.
[2] T. Ghani, et al., *IEDM Tech. Dig.*, pp. 197-200, 2003.

Layer	Pitch (nm)	Thick (nm)	AspectRatio
Isolation	220	320	-

Figure 2. Transistor size trend for technology nodes.

So, this is a typical trend; so, on a 65 nanometer technology, the gate lengths are 35 nanometer; this is a typical cross section of a transistor.

(Refer Slide Time: 49:14)

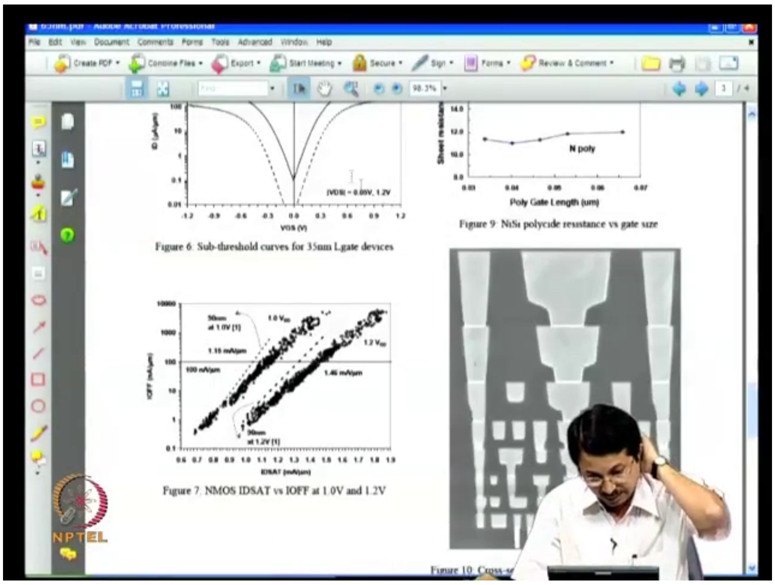


Figure 6: Sub-threshold curves for 35nm Lgate devices

Figure 9: NiSi polycide resistance vs gate size

Figure 7: NMOS IDSAT vs IOFF at 1.0V and 1.2V

Figure 10: Cross-s

N and P channel transistor ID, VDS characteristics and sub threshold characteristics, this is off current versus on current; typical DC performance metric. And this is; the same for PMOS transistor, this is a very nice cross section showing different levels of metal interconnects.

(Refer Slide Time: 49:38)

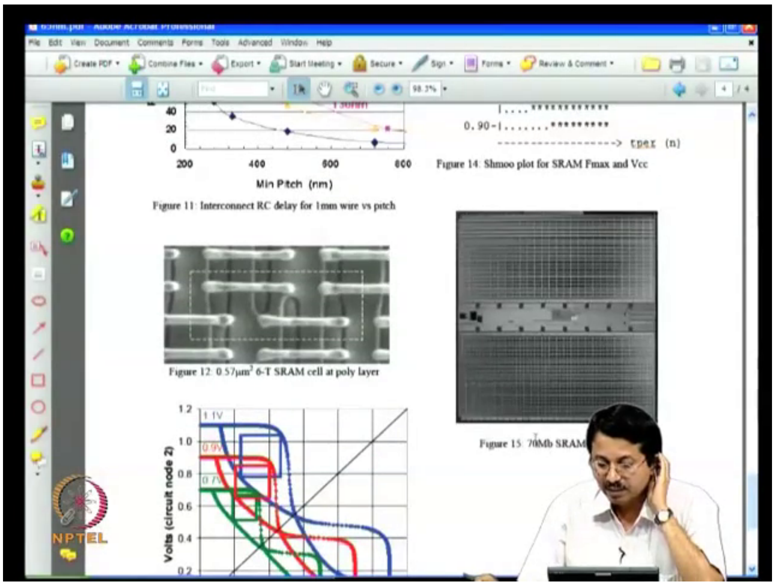


Figure 11: Interconnect RC delay for 1mm wire vs pitch

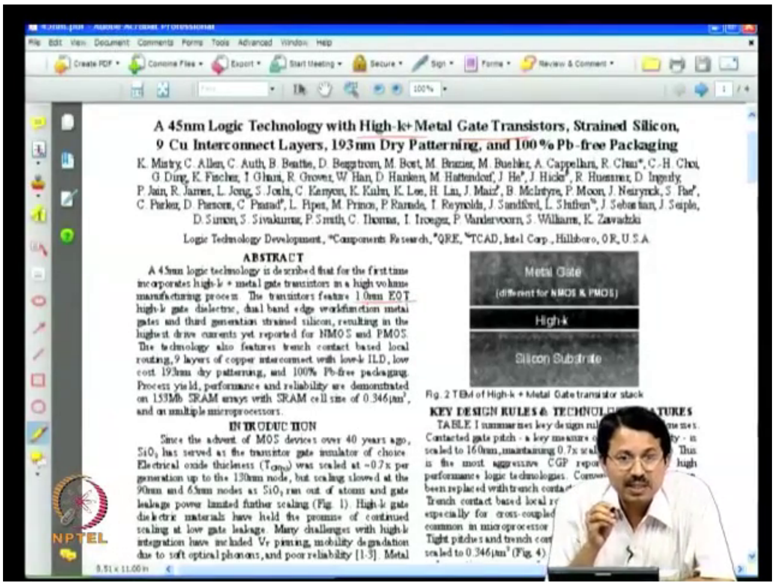
Figure 14: Shmoo plot for SRAM Fmax and Vcc

Figure 12: 0.57μm² 6-T SRAM cell at poly layer

Figure 15: 70Mb SRAM

And they have made an SRAM, this is 70 mega bit SRAM test vehicle. And they have been able to assure that this SRAM operates at frequency is close to 2 gigahertz and that they have been able to demonstrate; no major departures. Then at 45 nanometer; you see a very significant departure.

(Refer Slide Time: 50:01)



Adobe Acrobat Professional
Document Comments Forms Tools Advanced Window Help
Create PDF Combine Files Export Start Meeting Secure Sign Forms Review & Comment
100%

A 45nm Logic Technology with High-k+Metal Gate Transistors, Strained Silicon, 9 Cu Interconnect Layers, 193nm Dry Patterning, and 100% Pb-free Packaging

K. Mistry, C. Allen, C. Auth, B. Beattie, D. Bergstrom, M. Bost, M. Brazier, M. Buehler, A. Cappellani, R. Chau*, C.-H. Choi, G. Ding, K. Fischer, T. Ghani, R. Grover, W. Han, D. Hanken, M. Hattendorf, J. He#, J. Hicks#, R. Huessner, D. Ingerly, P. Jain, R. James, L. Jong, S. Joshi, C. Kenyon, K. Kuhn, K. Lee, H. Liu, J. Maiz#, B. McIntyre, P. Moon, J. Neirynck, S. Pae#, C. Parker, D. Parsons, C. Prasad#, L. Pipes, M. Prince, P. Ranade, T. Reynolds, J. Sandford, L. Shifren%, J. Sebastian, J. Seiple, D. Simon, S. Sivakumar, P. Smith, C. Thomas, T. Troeger, P. Vandervoorn, S. Williams, K. Zawadzki

Logic Technology Development, *Components Research, #QRE, %TCAD, Intel Corp., Hillsboro, OR, U.S.A

ABSTRACT

A 45nm logic technology is described that for the first time incorporates high-k + metal gate transistors in a high volume manufacturing process. The transistors feature 1.0nm EOT high-k gate dielectric, dual band edge workfunction metal gates and third generation strained silicon, resulting in the highest drive currents yet reported for NMOS and PMOS. The technology also features trench contact based local routing, 9 layers of copper interconnect with low-k ILD, low cost 193nm dry patterning, and 100% Pb-free packaging. Process yield, performance and reliability are demonstrated on 153Mb SRAM arrays with SRAM cell size of 0.346μm², and on multiple microprocessors.

INTRODUCTION

Since the advent of MOS devices over 40 years ago, SiO_2 has served as the transistor gate insulator of choice. Electrical oxide thickness ($T_{OX(e)}$) was scaled at ~0.7x per generation up to the 130nm node, but scaling slowed at the 90nm and 65nm nodes as SiO_2 ran out of atoms and gate leakage power limited further scaling (Fig. 1). High-k gate dielectric materials have held the promise of continued scaling at low gate leakage. Many challenges with high-k integration have included V_T pinning, mobility degradation due to soft optical phonons, and poor reliability [1-3]. Metal

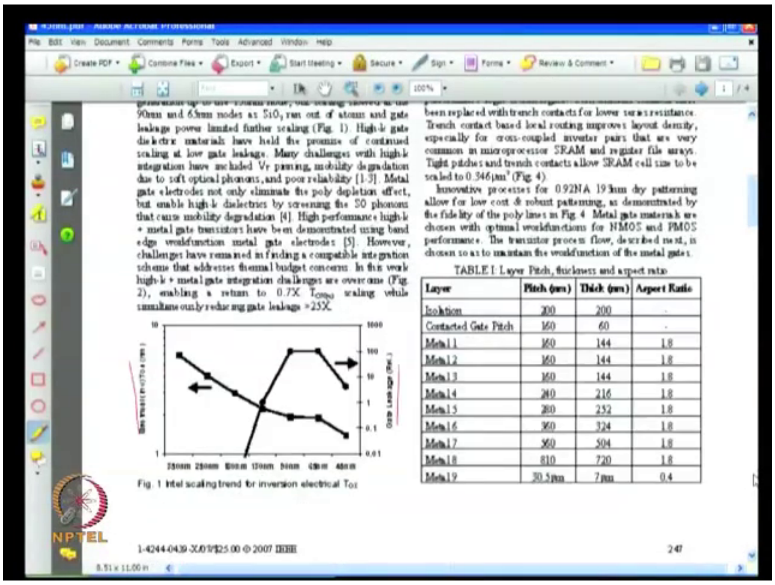
Fig. 2 TEM of High-k + Metal Gate transistor stack

KEY DESIGN RULES & TECHNOL TURES

TABLE I summarizes key design rul nemes. Contacted gate pitch - a key measure o ty - is scaled to 160nm, maintaining 0.7x scal This is the most aggressive CGP repor high performance logic technologies. Conv been replaced with trench conta Trench contact based local r especially for cross-coupled common in microprocessor Tight pitches and trench con scaled to 0.346μm² (Fig. 4)

And what is that? For the first time, you see high k gate dielectrics; along with metal gate. Here high k metal gate for the first time; 45 nanometer logic technology and transistors feature; 1 nanometer EOT, that is equivalent oxide thickness, but they actually use a hafnium oxide gate dielectric, which is thicker than 1 nanometer. They use a high k gate dielectric, dual band edge work function metal gates; two different metals; one for N channel, one for P channel transistors. So, what they have been able to show here is the following.

(Refer Slide Time: 50:50)



90nm and 65nm nodes as SiO$_2$ ran out of atoms and gate leakage power limited further scaling (Fig. 1). High-k gate dielectric materials have held the promise of continued scaling at low gate leakage. Many challenges with high-k integration have included V$_T$ pinning, mobility degradation due to soft optical phonons, and poor reliability [1-3]. Metal gate electrodes not only eliminate the poly depletion effect, but enable high-k dielectrics by screening the SO phonons that cause mobility degradation [4]. High performance high-k + metal gate transistors have been demonstrated using band edge workfunction metal gate electrodes [5]. However, challenges have remained in finding a compatible integration scheme that addresses thermal budget concerns. In this work high-k + metal gate integration challenges are overcome (Fig. 2), enabling a return to 0.7X T$_{OXinv}$ scaling while simultaneously reducing gate leakage >25X.

Fig. 1 Intel scaling trend for inversion electrical T$_{OX}$

been replaced with trench contacts for lower series resistance. Trench contact based local routing improves layout density, especially for cross-coupled inverter pairs that are very common in microprocessor SRAM and register file arrays. Tight pitches and trench contacts allow SRAM cell size to be scaled to 0.346μm^2 (Fig. 4).

Innovative processes for 0.92NA 193nm dry patterning allow for low cost & robust patterning, as demonstrated by the fidelity of the poly lines in Fig. 4. Metal gate materials are chosen with optimal workfunctions for NMOS and PMOS performance. The transistor process flow, described next, is chosen so as to maintain the workfunction of the metal gates.

TABLE I: Layer Pitch, thickness and aspect ratio

Layer	Pitch (nm)	Thick (nm)	Aspect Ratio
Isolation	200	200	-
Contacted Gate Pitch	160	60	-
Metal 1	160	144	1.8
Metal 2	160	144	1.8
Metal 3	160	144	1.8
Metal 4	240	216	1.8
Metal 5	280	252	1.8
Metal 6	360	324	1.8
Metal 7	560	504	1.8
Metal 8	810	720	1.8
Metal 9	30.5μm	7μm	0.4

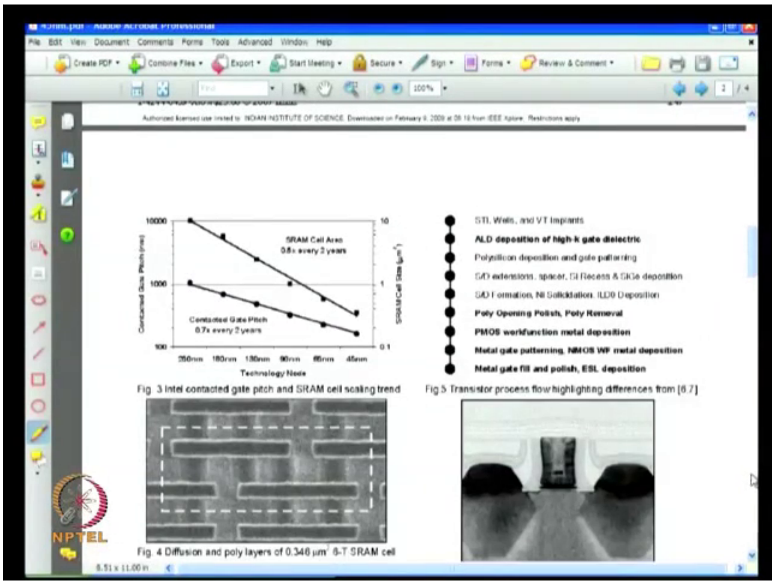
1-4244-0439-X/07/$25.00 © 2007 IEEE

247

You see this is gate leakage here and this is oxide thickness; EOT, when you use SiO 2; EOT and SiO 2 are one and the same. You see, from previous this is a 45 nanometer generation, at 65 nanometer generation; they were not able to scale the gate oxide thickness. Because the leakage current; as we start a decreasing the gate oxide thickness, leakage current suddenly started increasing because of direct tunneling current. And because from 90 nanometer to 65 nanometer; they did not have a mature high k gate dielectric technology, they could not scale the oxide thickness, they kept the oxide thickness as is and then the leakage current was kept at the same time, same level.

Now, with the high k technology; the EOT came down, but the leakage did not increase; the leakage also came down. Because it is a EOT of 1 nanometer, but indeed it is a thicker; physical oxide, direct tunneling is suppressed and hence you got the best of the world. You got lower EOT and lower leakage current; that is why you want to use high k dielectric. That is illustrated very nicely here, showing the trends in different technologies. So, this is the high k and metal gate is different for NMOS and PMOS.

(Refer Slide Time: 52:28)



Authorized licensed use limited to: INDIAN INSTITUTE OF SCIENCE. Downloaded on February 9, 2009 at 05:10 from IEEE Xplore. Restrictions apply.

Fig. 3 Intel contacted gate pitch and SRAM cell scaling trend

- STI, Wells, and VT Implants
- **ALD deposition of high-k gate dielectric**
- Polysilicon deposition and gate patterning
- S/D extensions, spacer, Si Recess & SiGe deposition
- S/D Formation, Ni Salicidation, ILD0 Deposition
- **Poly Opening Polish, Poly Removal**
- **PMOS workfunction metal deposition**
- **Metal gate patterning, NMOS WF metal deposition**
- **Metal gate fill and polish, ESL deposition**

Fig.5 Transistor process flow highlighting differences from [6,7]

Fig. 4 Diffusion and poly layers of 0.346 μm² 6-T SRAM cell

So, what is the typical process flow? They do a STI; Shallow Trench Isolation, wells and adjust the Vts, they deposit high k gate dielectric using Atomic Layer Deposition; ALD is Atomic Layer Deposition of the dielectric. They do a disposable gate; metal gate, remember the discussion that we had. So, polysilicon is done as usual, gate patterning, source drain extension, spacers; everything is done.

Source drain formation, nickel silicidation and then you deposit an insulator. Then poly opening you yetch-back the poly; poly is removed that is what we mean by disposable metal gate. Poly was sitting there as a place holder so that you get self aligned transistors. Then you do one metal for PMOS transistor; which has a work function appropriate with P channel transistor. Do patterning of the gate and then again do a NMOS work function metal deposition.

You need two metals for defining two different kinds of transistor and you complete the transistors and then of course, you do the multi level metal interconnects.

(Refer Slide Time: 53:40)

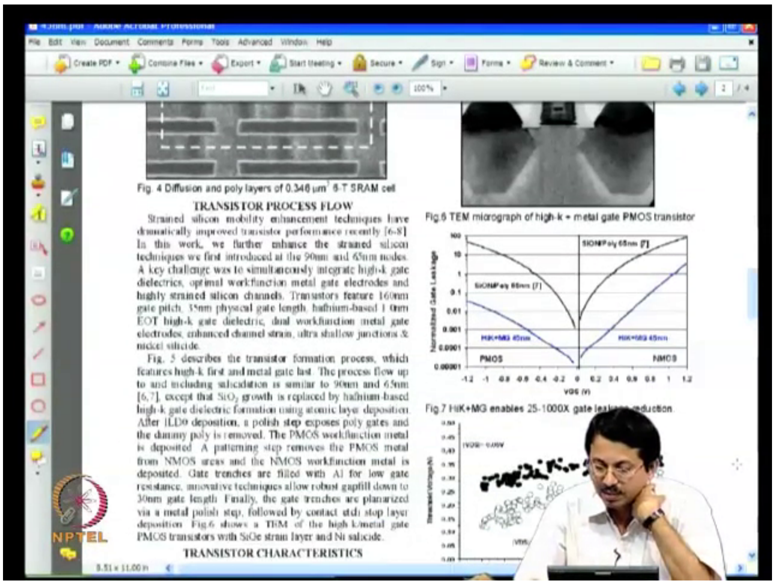


Fig. 4 Diffusion and poly layers of 0.346 μm² 6-T SRAM cell

TRANSISTOR PROCESS FLOW

Strained silicon mobility enhancement techniques have dramatically improved transistor performance recently [6-8]. In this work, we further enhance the strained silicon techniques we first introduced at the 90nm and 65nm nodes. A key challenge was to simultaneously integrate high-k gate dielectrics, optimal workfunction metal gate electrodes and highly strained silicon channels. Transistors feature 160nm gate pitch, 35nm physical gate length, hafnium-based 1.0nm EOT high-k gate dielectric, dual workfunction metal gate electrodes, enhanced channel strain, ultra shallow junctions & nickel silicide.

Fig. 5 describes the transistor formation process, which features high-k first and metal gate last. The process flow up to and including salicidation is similar to 90nm and 65nm [6,7], except that SiO_2 growth is replaced by hafnium-based high-k gate dielectric formation using atomic layer deposition. After ILD0 deposition, a polish step exposes poly gates and the dummy poly is removed. The PMOS workfunction metal is deposited. A patterning step removes the PMOS metal from NMOS areas and the NMOS workfunction metal is deposited. Gate trenches are filled with Al for low gate resistance, innovative techniques allow robust gapfill down to 30nm gate length. Finally, the gate trenches are planarized via a metal polish step, followed by contact etch stop layer deposition. Fig.6 shows a TEM of the high k/metal gate PMOS transistors with SiGe strain layer and Ni salicide.

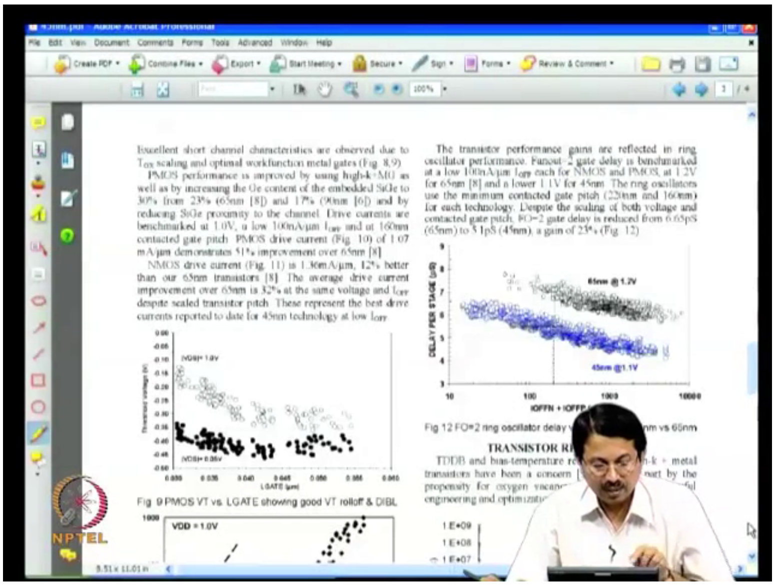
TRANSISTOR CHARACTERISTICS

Fig.6 TEM micrograph of high-k + metal gate PMOS transistor

Fig.7 HiK+MG enables 25-1000X gate leakage reduction

So, this is again to show that with high k; your gate leakage has come down very dramatically. If you had used the silicon oxide or even so called nitride silicon oxide, you would had a very leakage current, but now that is not the case. This is the typical Vt versus gate length; threshold voltage role off, that we have seen; measured that low drain voltage and high drain voltage; that is for N channel and this is for P channel transistors.

(Refer Slide Time: 54:16)



Excellent short channel characteristics are observed due to T_{OX} scaling and optimal workfunction metal gates (Fig. 8,9).

PMOS performance is improved by using high-k+MG as well as by increasing the Ge content of the embedded SiGe to 30% from 23% (65nm [8]) and 17% (90nm [6]) and by reducing SiGe proximity to the channel. Drive currents are benchmarked at 1.0V, a low 100nA/μm I_{OFF} and at 160nm contacted gate pitch. PMOS drive current (Fig. 10) of 1.07 mA/μm demonstrates 51% improvement over 65nm [8].

NMOS drive current (Fig. 11) is 1.36mA/μm, 12% better than our 65nm transistors [8]. The average drive current improvement over 65nm is 32% at the same voltage and I_{OFF} despite scaled transistor pitch. These represent the best drive currents reported to date for 45nm technology at low I_{OFF}.

Fig 9 PMOS VT vs. LGATE showing good VT rolloff & DIBL

The transistor performance gains are reflected in ring oscillator performance. Fanout=2 gate delay is benchmarked at a low 100nA/μm I_{OFF} each for NMOS and PMOS, at 1.2V for 65nm [8] and a lower 1.1V for 45nm. The ring oscillators use the minimum contacted gate pitch (220nm and 160nm) for each technology. Despite the scaling of both voltage and contacted gate pitch, FO=2 gate delay is reduced from 6.65pS (65nm) to 5.1pS (45nm), a gain of 23% (Fig. 12).

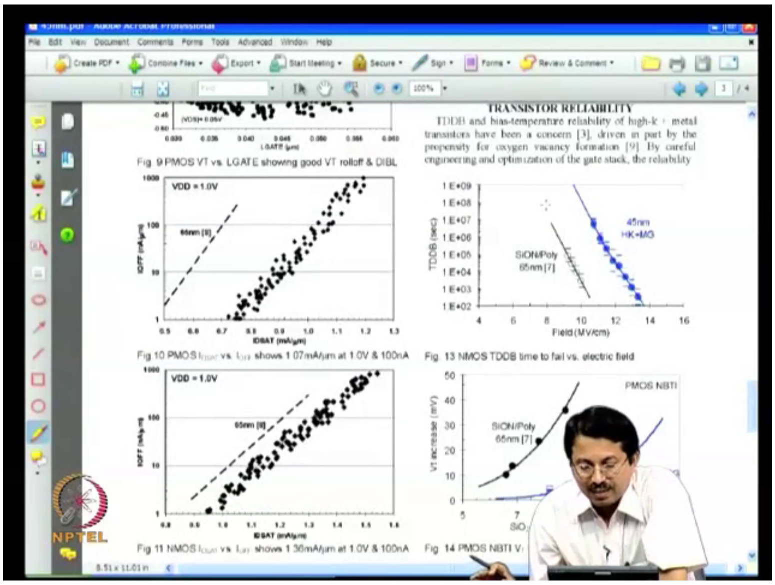
Fig 12 FO=2 ring oscillator delay v nm vs 65nm

TRANSISTOR R

TDDB and bias-temperature r -k + metal transistors have been a concern [art by the propensity for oxygen vacan ful engineering and optimizatio

Typical off current versus on current plots; looks similar, everything looks similar.

(Refer Slide Time: 54:21)



TRANSISTOR RELIABILITY

TDDB and bias-temperature reliability of high-k + metal transistors have been a concern [3], driven in part by the propensity for oxygen vacancy formation [9]. By careful engineering and optimization of the gate stack, the reliability

Fig. 9 PMOS VT vs. LGATE showing good VT rolloff & DIBL

Fig 10 PMOS I_{DSAT} vs. I_{OFF} shows 1.07mA/μm at 1.0V & 100nA

Fig 11 NMOS I_{DSAT} vs. I_{OFF} shows 1.36mA/μm at 1.0V & 100nA

Fig. 13 NMOS TDDB time to fail vs. electric field

Fig 14 PMOS NBTI V

And now typical delay versus off current; remember? This is how it looks, when you have very large off current; it means the smallest length transistor has a highest leakage current. And hence because you have build the inverter using the smallest length transistor you have the best delay that you can get. And the delays are now of the order of very low delays, 5 picoseconds or even less than 5 picoseconds.

Reliability of the transistors; again they have characterized TDDB by accelerated testing; extrapolation, just to show that these high k gate dielectrics are as reliable as silicon oxide.

(Refer Slide Time: 55:07)

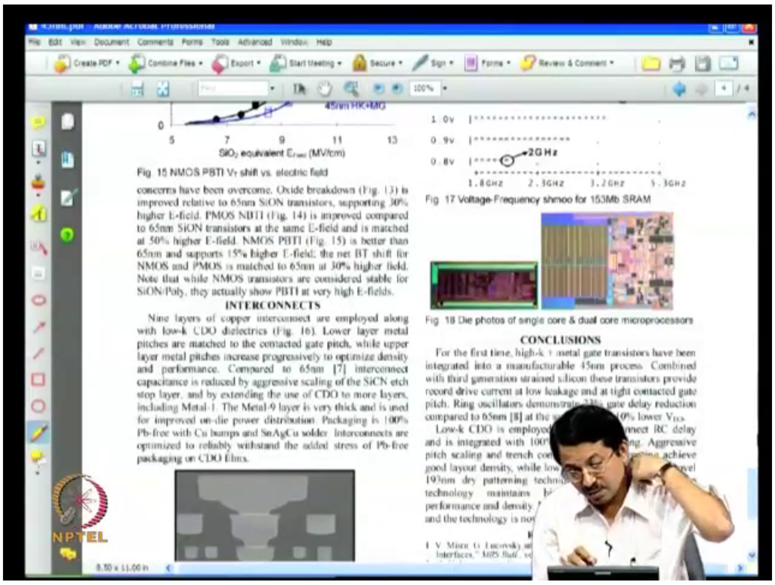


Fig. 15 NMOS PBTI V_T shift vs. electric field

concerns have been overcome. Oxide breakdown (Fig. 13) is improved relative to 65nm SiON transistors, supporting 30% higher E-field. PMOS NBTI (Fig. 14) is improved compared to 65nm SiON transistors at the same E-field and is matched at 50% higher E-field. NMOS PBTI (Fig. 15) is better than 65nm and supports 15% higher E-field; the net BT shift for NMOS and PMOS is matched to 65nm at 30% higher field. Note that while NMOS transistors are considered stable for SiON/Poly, they actually show PBTI at very high E-fields.

INTERCONNECTS

Nine layers of copper interconnect are employed along with low-k CDO dielectrics (Fig. 16). Lower layer metal pitches are matched to the contacted gate pitch, while upper layer metal pitches increase progressively to optimize density and performance. Compared to 65nm [7] interconnect capacitance is reduced by aggressive scaling of the SiCN etch stop layer, and by extending the use of CDO to more layers, including Metal-1. The Metal-9 layer is very thick and is used for improved on-die power distribution. Packaging is 100% Pb-free with Cu bumps and SnAgCu solder. Interconnects are optimized to reliably withstand the added stress of Pb-free packaging on CDO films.

Fig. 17 Voltage-Frequency shmoo for 153Mb SRAM

Fig. 18 Die photos of single core & dual core microprocessors

CONCLUSIONS

For the first time, high-k + metal gate transistors have been integrated into a manufacturable 45nm process. Combined with third generation strained silicon these transistors provide record drive current at low leakage and at tight contacted gate pitch. Ring oscillators demonstrate 23% gate delay reduction compared to 65nm [8] at the [illegible] 10% lower V_{CC}.

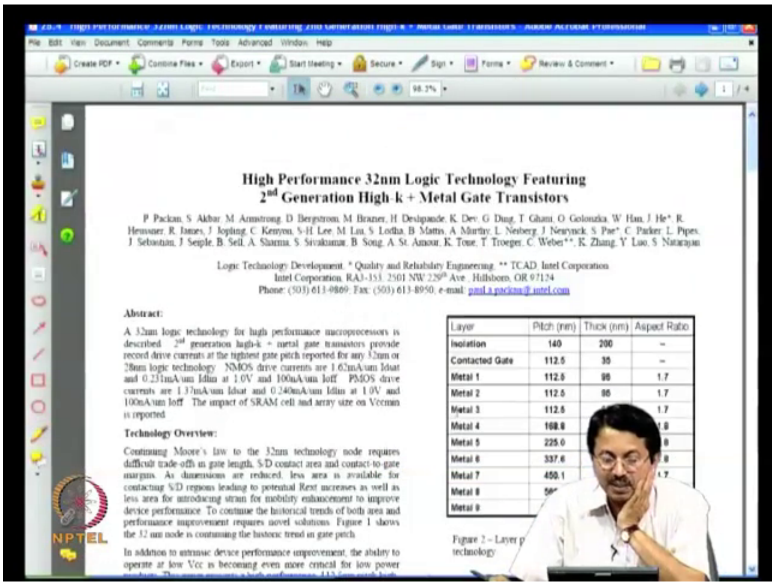
Low-k CDO is employed [illegible] nnect RC delay and is integrated with 100 [illegible] ng. Aggressive pitch scaling and trench co [illegible] achieve good layout density, while low [illegible] vel 193nm dry patterning techni [illegible] technology maintains hi [illegible] performance and density [illegible] and the technology is no [illegible]

1 V. Misra, G. Lucovsky ... Interfaces," MRS Bull., ve [illegible]

8.50 x 11.00 in

Then they have actually made, not only SRAM; in this case they have actually made a microprocessor. They have made a single core microprocessor and dual core microprocessor and been able to show that; you can actually get a fairly good circuit working with all these new technology; high k and metal gate so, on and so forth.

(Refer Slide Time: 55:37)



High Performance 32nm Logic Technology Featuring
2nd Generation High-k + Metal Gate Transistors

P. Packan, S. Akbar, M. Armstrong, D. Bergstrom, M. Brazier, H. Deshpande, K. Dev, G. Ding, T. Ghani, O. Golonzka, W. Han, J. He*, R. Heussner, R. James, J. Jopling, C. Kenyon, S-H. Lee, M. Liu, S. Lodha, B. Mattis, A. Murthy, L. Neiberg, J. Neirynck, S. Pae*, C. Parker, L. Pipes, J. Sebastian, J. Seiple, B. Sell, A. Sharma, S. Sivakumar, B. Song, A. St. Amour, K. Tone, T. Troeger, C. Weber**, K. Zhang, Y. Luo, S. Natarajan

Logic Technology Development, * Quality and Reliability Engineering, ** TCAD, Intel Corporation
Intel Corporation, RA3-353, 2501 NW 229th Ave., Hillsboro, OR 97124
Phone: (503) 613-9869; Fax: (503) 613-8950, e-mail: paul.a.packan@intel.com

Abstract:

A 32nm logic technology for high performance microprocessors is described. 2nd generation high-k + metal gate transistors provide record drive currents at the tightest gate pitch reported for any 32nm or 28nm logic technology. NMOS drive currents are 1.62mA/um Idsat and 0.231mA/um Idlin at 1.0V and 100nA/um Ioff. PMOS drive currents are 1.37mA/um Idsat and 0.240mA/um Idlin at 1.0V and 100nA/um Ioff. The impact of SRAM cell and array size on Vccmin is reported.

Technology Overview:

Continuing Moore's law to the 32nm technology node requires difficult trade-offs in gate length, S/D contact area and contact-to-gate margins. As dimensions are reduced, less area is available for contacting S/D regions leading to potential Rext increases as well as less area for introducing strain for mobility enhancement to improve device performance. To continue the historical trends of both area and performance improvement requires novel solutions. Figure 1 shows the 32 nm node is continuing the historic trend in gate pitch.

In addition to intrinsic device performance improvement, the ability to operate at low Vcc is becoming even more critical for low power

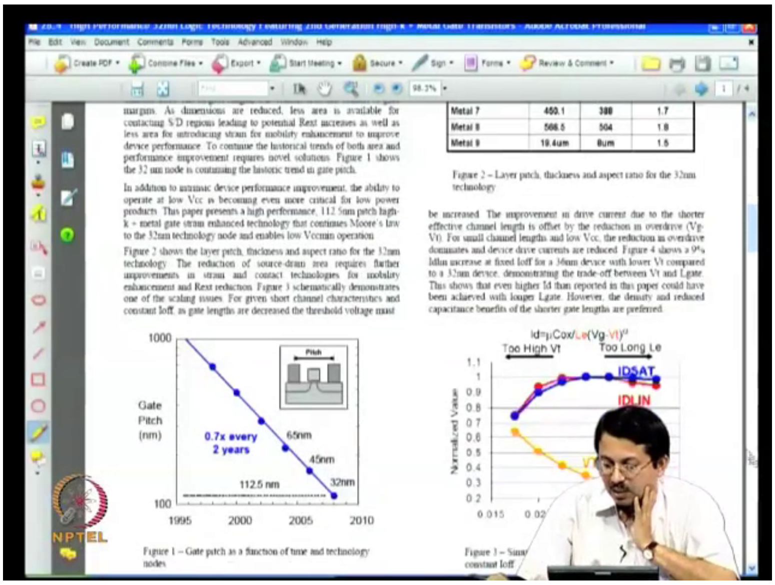
Layer	Pitch (nm)	Thick (nm)	Aspect Ratio
Isolation	140	200	–
Contacted Gate	112.5	35	–
Metal 1	112.5	95	1.7
Metal 2	112.5	95	1.7
Metal 3	112.5	[illegible]	1.7
Metal 4	168.8	[illegible]	[illegible]
Metal 5	225.0	[illegible]	[illegible]
Metal 6	337.6	[illegible]	[illegible]
Metal 7	450.1	[illegible]	[illegible]
Metal 8	[illegible]	[illegible]	[illegible]
Metal 9	[illegible]	[illegible]	[illegible]

Figure 2 – Layer p [illegible] technology

And that is what happened in 45 nanometer technology and in 32 nanometer technology; with high k and metal gate continued, it continues to be high k and metal gate transistors, we have stuck to that right now. Going forward, we will continue to use high k and metal gate, but this called second generation of high k and they have tried to do some optimization; improve the hafnium oxide performance, improve the metal gate performance and so on and so forth.

So, 32 nanometer; the actual gate length is little lower than 32 nanometer.

(Refer Slide Time: 56:12)


28.4 High Performance 32nm Logic Technology Featuring 2nd Generation High-k + Metal Gate Transistors - Adobe Acrobat Professional
File Edit View Document Comments Forms Tools Advanced Window Help
Create PDF
Combine Files
Export
Start Meeting
Secure
Sign
Forms
Review & Comment
98.3%
1 / 4
NPTEL

margins. As dimensions are reduced, less area is available for contacting S/D regions leading to potential Rext increases as well as less area for introducing strain for mobility enhancement to improve device performance. To continue the historical trends of both area and performance improvement requires novel solutions. Figure 1 shows the 32 nm node is continuing the historic trend in gate pitch.

In addition to intrinsic device performance improvement, the ability to operate at low Vcc is becoming even more critical for low power products. This paper presents a high performance, 112.5nm pitch high-k + metal gate strain enhanced technology that continues Moore's law to the 32nm technology node and enables low Vccmin operation.

Figure 2 shows the layer pitch, thickness and aspect ratio for the 32nm technology. The reduction of source-drain area requires further improvements in strain and contact technologies for mobility enhancement and Rext reduction. Figure 3 schematically demonstrates one of the scaling issues. For given short channel characteristics and constant Ioff, as gate lengths are decreased the threshold voltage must

Figure 1 – Gate pitch as a function of time and technology nodes

Metal 7	450.1	388	1.7
Metal 8	566.5	504	1.8
Metal 9	19.4um	8um	1.5

Figure 2 – Layer pitch, thickness and aspect ratio for the 32nm technology

be increased. The improvement in drive current due to the shorter effective channel length is offset by the reduction in overdrive (Vg-Vt). For small channel lengths and low Vcc, the reduction in overdrive dominates and device drive currents are reduced. Figure 4 shows a 9% Idlin increase at fixed Ioff for a 36nm device with lower Vt compared to a 32nm device, demonstrating the trade-off between Vt and Lgate. This shows that even higher Id than reported in this paper could have been achieved with longer Lgate. However, the density and reduced capacitance benefits of the shorter gate lengths are preferred.

Figure 3 – Simu
constant Ioff

This is just showing that previous generation technology and this is 32 nanometer technology.

(Refer Slide Time: 56:23)

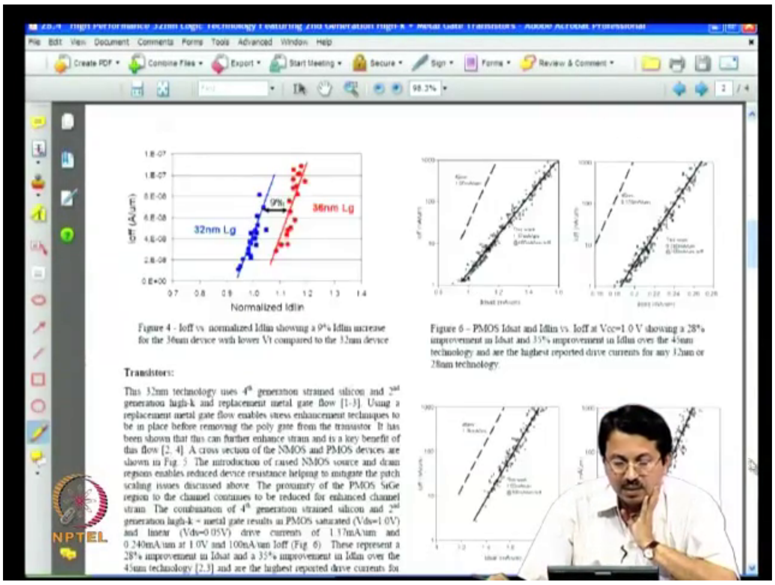


Figure 4 - Ioff vs. normalized Idlin showing a 9% Idlin increase for the 36nm device with lower Vt compared to the 32nm device

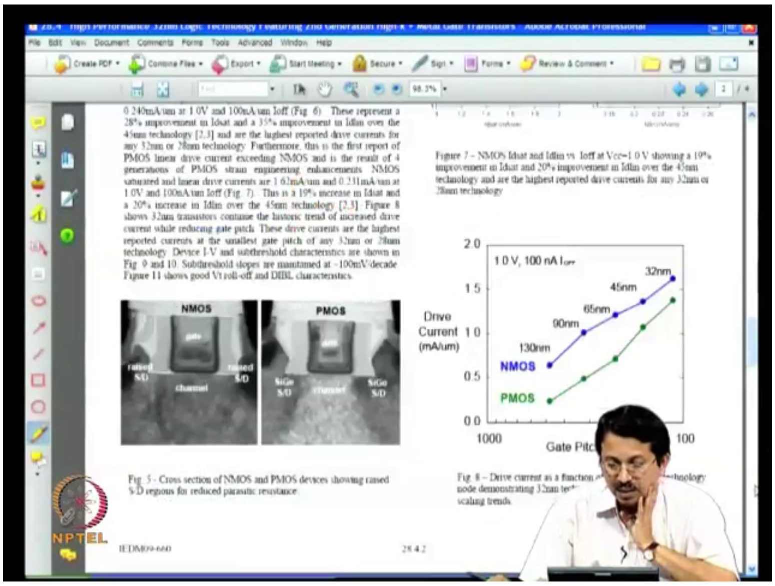
Transistors:

This 32nm technology uses 4th generation strained silicon and 2nd generation high-k and replacement metal gate flow [1-3]. Using a replacement metal gate flow enables stress enhancement techniques to be in place before removing the poly gate from the transistor. It has been shown that this can further enhance strain and is a key benefit of this flow [2, 4]. A cross section of the NMOS and PMOS devices are shown in Fig. 5. The introduction of raised NMOS source and drain regions enables reduced device resistance helping to mitigate the pitch scaling issues discussed above. The proximity of the PMOS SiGe region to the channel continues to be reduced for enhanced channel strain. The combination of 4th generation strained silicon and 2nd generation high-k + metal gate results in PMOS saturated (Vds=1.0V) and linear (Vds=0.05V) drive currents of 1.37mA/um and 0.240mA/um at 1.0V and 100nA/um Ioff (Fig. 6). These represent a 28% improvement in Idsat and a 35% improvement in Idlin over the 45nm technology [2,3] and are the highest reported drive currents for

Figure 6 - PMOS Idsat and Idlin vs. Ioff at Vcc=1.0 V showing a 28% improvement in Idsat and 35% improvement in Idlin over the 45nm technology and are the highest reported drive currents for any 32nm or 28nm technology.

And off current versus on current, the typical plots that we have seen. All these are off current versus on current plots, we trying to indicate that; they have been able to optimize these transistors.

(Refer Slide Time: 56:36)



0.240mA/um at 1.0V and 100nA/um Ioff (Fig. 6). These represent a 28% improvement in Idsat and a 35% improvement in Idlin over the 45nm technology [2,3] and are the highest reported drive currents for any 32nm or 28nm technology. Furthermore, this is the first report of PMOS linear drive current exceeding NMOS and is the result of 4 generations of PMOS strain engineering enhancements. NMOS saturated and linear drive currents are 1.62mA/um and 0.231mA/um at 1.0V and 100nA/um Ioff (Fig. 7). This is a 19% increase in Idsat and a 20% increase in Idlin over the 45nm technology [2,3]. Figure 8 shows 32nm transistors continue the historic trend of increased drive current while reducing gate pitch. These drive currents are the highest reported currents at the smallest gate pitch of any 32nm or 28nm technology. Device I-V and subthreshold characteristics are shown in Fig. 9 and 10. Subthreshold slopes are maintained at ~100mV/decade. Figure 11 shows good Vt roll-off and DIBL characteristics.

Figure 7 – NMOS Idsat and Idlin vs. Ioff at Vcc=1.0 V showing a 19% improvement in Idsat and 20% improvement in Idlin over the 45nm technology and are the highest reported drive currents for any 32nm or 28nm technology.

Fig. 5 - Cross section of NMOS and PMOS devices showing raised S/D regions for reduced parasitic resistance.

Fig. 8 – Drive current as a function of ... technology node demonstrating 32nm te... scaling trends.

IEDM09-660

28.4.2

For a given off current; they have been able to get as large on current as possible. Typical output characteristics; IDS versus VDS; notice that the difference between N and P has come down.

(Refer Slide Time: 56:44)

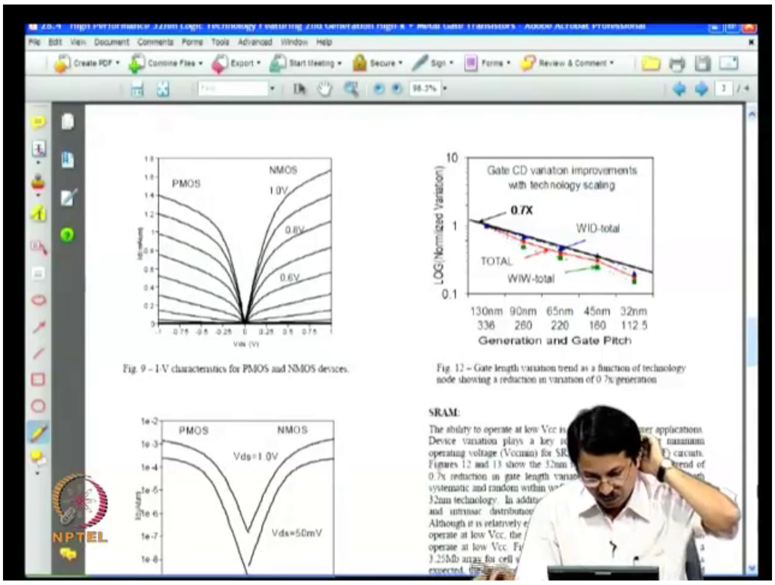


Fig. 9 – I-V characteristics for PMOS and NMOS devices.

Fig. 12 – Gate length variation trend as a function of technology node showing a reduction in variation of 0.7x/generation.

SRAM:

The ability to operate at low Vcc is ... er applications. Device variation plays a key ro... minimum operating voltage (Vccmin) for SR... circuits. Figures 12 and 13 show the 32nm ... trend of 0.7x reduction in gate length varia... systematic and random within w... 32nm technology. In additi... and intrinsic distribution... Although it is relatively e... operate at low Vcc, the ... operate at low Vcc. Fi... 3.25Mb arrays for cell ... expected, the ...

If you recall 180 nanometer technology, your P channel current was almost 50 percent of the N channel current. Now, using the so called strain engineering which; as I said we will be discussed in one of the future classes. We are able to enhance both electron and whole mobilities and so much so that the whole mobility can approach very close to electron mobility.

And that is why you do not see a huge difference between electron and hole mobility and electron and hole current; N channel and P channel current.

(Refer Slide Time: 57:32)

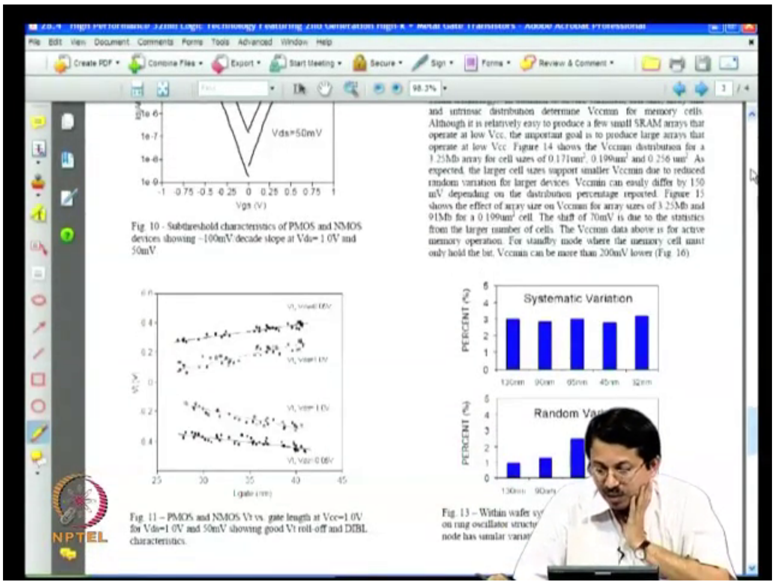


Fig. 10 - Subthreshold characteristics of PMOS and NMOS devices showing ~100mV/decade slope at Vds= 1.0V and 50mV

Fig. 11 – PMOS and NMOS Vt vs. gate length at Vcc=1.0V for Vds=1.0V and 50mV showing good Vt roll-off and DIBL characteristics.

and intrinsic distribution determine Vccmin for memory cells. Although it is relatively easy to produce a few small SRAM arrays that operate at low Vcc, the important goal is to produce large arrays that operate at low Vcc. Figure 14 shows the Vccmin distribution for a 3.25Mb array for cell sizes of 0.171um^2, 0.199um^2 and 0.256 um^2. As expected, the larger cell sizes support smaller Vccmin due to reduced random variation for larger devices. Vccmin can easily differ by 150 mV depending on the distribution percentage reported. Figure 15 shows the effect of array size on Vccmin for array sizes of 3.25Mb and 91Mb for a 0.199um^2 cell. The shift of 70mV is due to the statistics from the larger number of cells. The Vccmin data above is for active memory operation. For standby mode where the memory cell must only hold the bit, Vccmin can be more than 200mV lower (Fig. 16).

Fig. 13 – Within wafer sy
on ring oscillator structu
node has similar variat

That difference has come down; this is a typical sub threshold plot that you would see. And typical Vt versus length plot; Vt is decreasing with decreasing channel length. So, all these is consistent with what we have been discussing and they have again made an SRAM and they demonstrate that; SRAM works fairly well with this 32 nanometer technology.

(Refer Slide Time: 57:46)

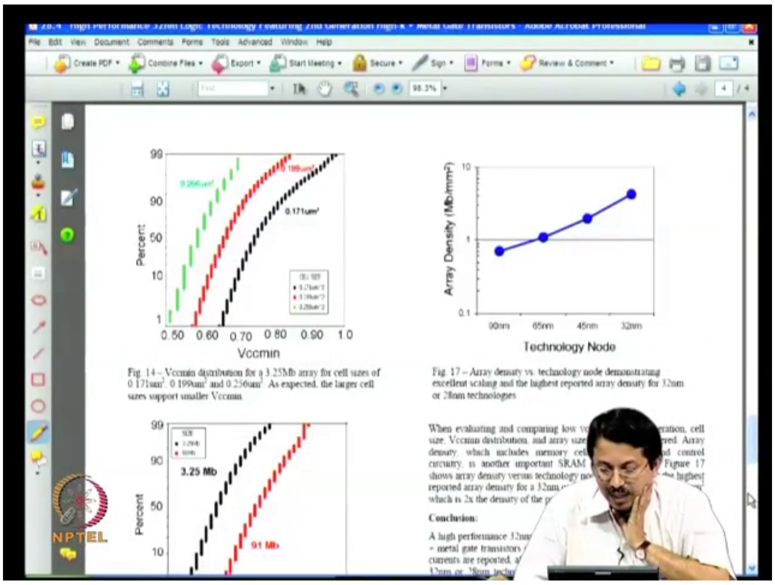

28.4 High Performance 32nm Logic Technology Featuring 2nd Generation High-k + Metal Gate Transistors - Adobe Acrobat Professional
File Edit View Document Comments Forms Tools Advanced Window Help
Create PDF Combine Files Export Start Meeting Secure Sign Forms Review & Comment
98.3%
99
90
50
10
1
Percent
0.266um²
0.199um²
0.171um²
0.50 0.60 0.70 0.80 0.90 1.0
Vccmin
10
1
0.1
Array Density (Mb/mm²)
90nm 65nm 45nm 32nm
Technology Node
3.25 Mb
91 Mb
NPTEL

Fig. 14 – Vccmin distribution for a 3.25Mb array for cell sizes of 0.171um², 0.199um² and 0.256um². As expected, the larger cell sizes support smaller Vccmin.

Fig. 17 – Array density vs. technology node demonstrating excellent scaling and the highest reported array density for 32nm or 28nm technologies.

When evaluating and comparing low v… …eration, cell size, Vccmin distribution, and array size… …ered. Array density, which includes memory cell… …nd control circuitry, is another important SRAM … Figure 17 shows array density versus technology no… …the highest reported array density for a 32nm… which is 2x the density of the p…

Conclusion:

A high performance 32nm… + metal gate transistors… currents are reported, a… 32nm or 28nm techn…

(Refer Slide Time: 57:48)

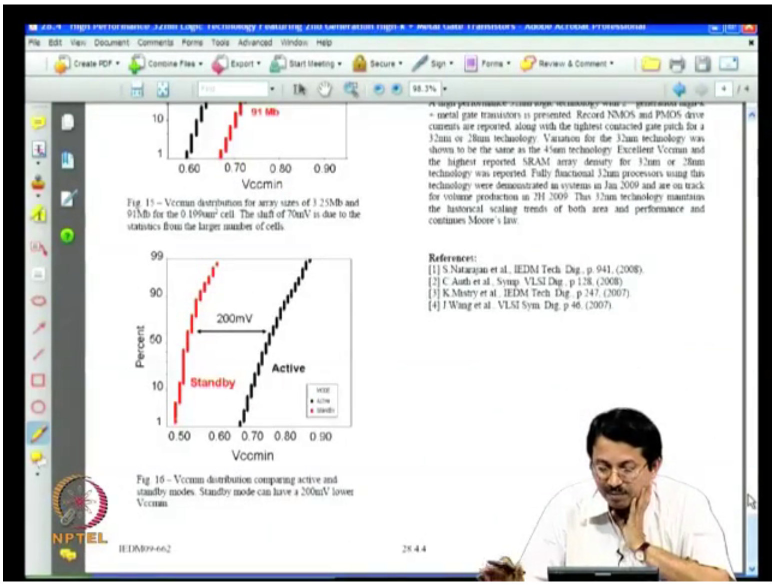

28.4 High Performance 32nm Logic Technology Featuring 2nd Generation High-k + Metal Gate Transistors - Adobe Acrobat Professional
File Edit View Document Comments Forms Tools Advanced Window Help
Create PDF Combine Files Export Start Meeting Secure Sign Forms Review & Comment
98.3%
10
1
91 Mb
0.60 0.70 0.80 0.90
Vccmin
99
90
50
10
1
Percent
200mV
Active
Standby
0.50 0.60 0.70 0.80 0.90
Vccmin
NPTEL

Fig. 15 – Vccmin distribution for array sizes of 3.25Mb and 91Mb for the 0.199um² cell. The shift of 70mV is due to the statistics from the larger number of cells.

Fig. 16 – Vccmin distribution comparing active and standby modes. Standby mode can have a 200mV lower Vccmin.

A high performance 32nm logic technology with 2nd generation high-k + metal gate transistors is presented. Record NMOS and PMOS drive currents are reported, along with the tightest contacted gate pitch for a 32nm or 28nm technology. Variation for the 32nm technology was shown to be the same as the 45nm technology. Excellent Vccmin and the highest reported SRAM array density for 32nm or 28nm technology was reported. Fully functional 32nm processors using this technology were demonstrated in systems in Jan 2009 and are on track for volume production in 2H 2009. Thus 32nm technology maintains the historical scaling trends of both area and performance and continues Moore's law.

References:
[1] S.Natarajan et al., IEDM Tech. Dig., p. 941, (2008).
[2] C.Auth et al., Symp. VLSI Dig., p.128, (2008).
[3] K.Mistry et al., IEDM Tech. Dig., p.247, (2007).
[4] J.Wang et al., VLSI Sym. Dig. p.46, (2007).

IEDM09-662 28.4.4

So, what we have been able to do in this lecture is really to look at representative CMOS logic technology, not memory technology; these technologies are not for DRAM fabrication. Not for flash memory fabrication, not for even SRAM fabrication, but to make a very high performance microprocessor. So, their performance is the key; you want to make the transistor, which gives a lowest delay.

And various techniques we have seen, all source drain engineering, channel engineering is consistently use in all these transistors. And as we have continued to scale the technologies, we have transition from silicon oxide to high k gate dielectrics; to metal gates and so on and so forth.

Similarly, in the interconnect arena; we have gone from aluminum and conventional silicon oxide to dope silicon oxide, copper and there by minimize the R C delay of interconnects very significantly. So, hopefully now you should be able to read up any CMOS technology paper and be able to understand that; based on all the discussion that we have had so far.

So, we will stop this lecture today and starting from next lecture, we will actually start looking at electrical characterization; that is C V characterization and I V characterization.