Expanded Application Space for Laser Spike Annealing of CMOS Devices

Jeff Hebb, Yun Wang, Shaoyin Chen, Michael Shen, Senquan Zhou, Xiaoru Wang, and David Owen

Ultratech Inc., 3050 Zanker Rd., San Jose, California, 95134, USA
E-mail: jhebb@ultratech.com, Phone: +1-408-321-8835

Abstract-**LSA was first introduced into mainstream semiconductor manufacturing for logic IC's at the 65nm node, continuing the natural evolution of semiconductor thermal processing to higher temperatures (>1200°C) and shorter times (100's of microseconds). The initial application was a simple one-step LSA to assist spike-RTA in dopant activation of the source/drain and polysilicon gate regions. Since then, LSA has proliferated to other junction engineering steps in the high temperature/low dwell time regime. As devices scale to sub-45nm nodes, there are opportunities for LSA to expand to alternative applications in other regions of temperature-time (T-t) process space. This paper will discuss the application of LSA to logic IC manufacturing within the framework of T-t process space, which is divided into three regimes: 1. High temperature, 2. Long dwell time, and 3. Low temperature. We describe how the design of the LSA system enables access to these three T-t regimes, and discuss current and future applications for each regime.**

I. INTRODUCTION

Laser spike annealing (LSA) was first introduced into mainstream semiconductor manufacturing at the 65nm logic node for the junction activation application. For the initial application, a single high temperature LSA step was added after the SD spike RTA in the standard CMOS process flow, to increase activation of the source drain extension (SDE), deep source drain (SD) and polysilicon gate, with negligible diffusion [1]. Beyond one-step junction activation, the use of LSA has become more pervasive as IC manufacturers find other insertion points and new applications for LSA which can boost device performance. Compared to other device performance boosters such as HK-MG or strain engineering, the integration of LSA is relatively simple and cost effective, as it does not require new materials to be introduced into the transistor structure.

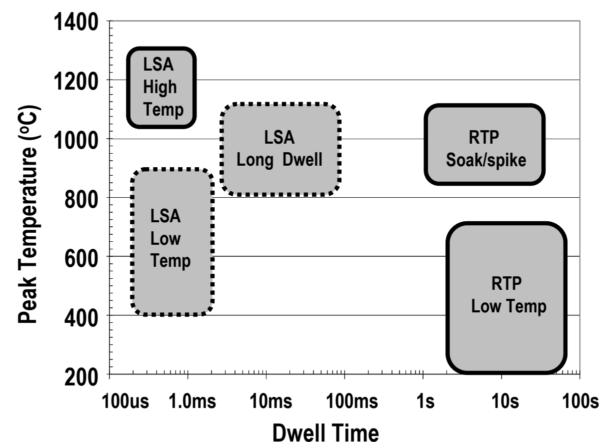

Figure 1: Time-temperature process regimes of LSA and RTP.

Three time temperature-time (T-t) regimes for LSA are shown in Fig.1. Two RTP T-t regimes are also shown for comparison. The "high temperature" regime covers most LSA applications being used in IC manufacturing today, although there are still new applications being developed in this T-t regime. Two new LSA T-t regimes are also shown – the "long dwell time" regime and the "low temperature" regime. In this paper, we will describe how the design of the LSA system enables access to all three T-t regimes, and discuss current and future applications for each regime.

II. LSA TOOL ARCHUTECTURE

A. Standard LSA system

The general architecture of a standard laser spike

978-1-4244-3815-0/09/$25.00 ©IEEE

annealing tool is shown in Figure 2 [2]. A high-power CO_2 laser at a wavelength of 10.6 μm is conditioned through a system of reflective optics to form a "line beam" at the wafer plane. The combination of long wavelength, p-polarization, and Brewster angle minimizes within-die reflectance variations, almost eliminating within-die temperature variations (pattern effects) [3]. The beam is approximately 100 μm in width (scan direction) and 5-10mm in length (step direction), depending on tool configuration. The wafer is sitting on a heated chuck which scans the wafer under the CO_2 laser beam. The dwell time is defined as the duration for which a point on the silicon wafer is exposed to the beam, and can be varied by changing the stage speed. The peak temperature is measured in real time by measuring the emitted radiation from the "hot spot" at multiple wavelengths, and converting the emitted intensity measurements into a wafer temperature. This temperature is compared to the temperature set point, and the error is used to make the appropriate correction to the laser power at an update rate of 10kHz [4]. Through a combination of uniform absorption and high frequency temperature control, the instantaneous peak temperature of the wafer surface is kept uniform on blanket and patterned wafers.

This standard configuration of the tool covers the "LSA high temperature" process space, with dwell times typically ranging from 200μsec to 1.5msec, and peak temperatures from 1100 to 1350°C. A simulation of a typical T-t profile is shown in Fig. 3.

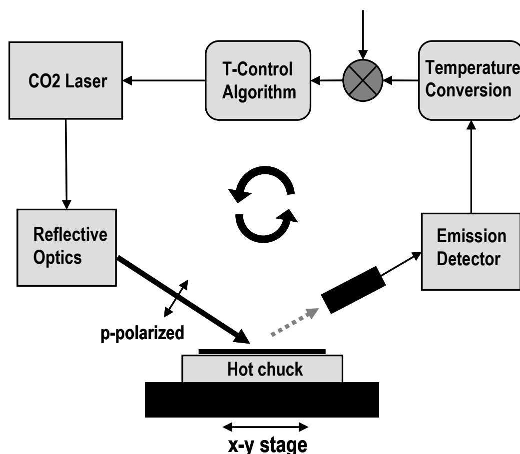

Figure 2: Block diagram of a laser spike annealing system (CO_2 beam only).

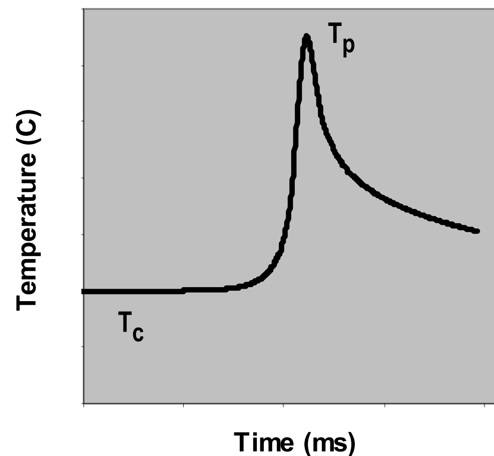

Figure 3: Simulated temperature profile of a point on the silicon wafer for an LSA system with a CO_2 beam. T_c is chuck temperature and T_p is the peak temperature

B. LSA with preheat beam

In order to expand the T-t space of LSA, a second laser beam is added as secondary heat source, referred to as the preheat beam [5, 6]. In general, the length of preheat beam length is the equal to or greater than the CO_2 beam, and the width is approximately an order of magnitude larger than that of the CO_2 beam. The dimensions, wavelength, angle, and polarization of the beam can be optimized depending on the application. The closed-loop temperature measurement and control approach used for the standard single-beam LSA described above is extendible to the dual beam approach

Figure 4 shows a general T-t profile for a point on the silicon wafer as it is scanned under the preheat and CO_2 beams. The points starts in thermal equilibrium with the chuck, then ramps to the intermediate temperature as it scans under the pre-heat beam. The point then encounters the CO_2 beam, which brings the wafer to the peak temperature, with a dwell time determined by the width of the CO_2 beam. The overall temperature profile can be thought of as a superposition of two separate anneals – a long dwell time anneal at intermediate temperature and a short dwell time anneal at higher peak temperature. The system can be configured to use the either beam alone, or with both preheat and CO_2 beam simultaneously.

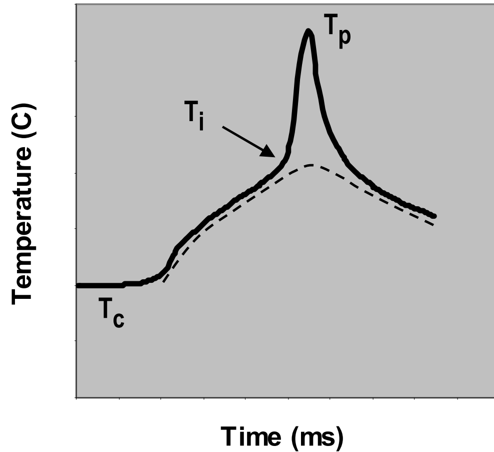

Figure 4: Simulated temperature profile a point on the silicon wafer for an LSA system with a preheat beam and a CO_2 beam. T_i is the intermediate temperature achieved by the pre-heat beam. The dashed line to illustrate the approximate temperature profile of pre-heat beam only.

This dual beam approach offers a wide range of process flexibility, where the adjustable parameters include the CO_2 dwell time, peak temperature, pre-heat dwell time, intermediate temperature, and chuck temperature. Depending on how these parameters are set, one can access the various T-t regimes described in Fig. 1. We will now discuss current and future applications covered by each of these LSA T-t regimes.

III. HIGH TEMPERATURE REGIME

A. LSA integration for junction engineering

When LSA was first introduced during the 65nm node, the most common integration scheme was to add a high-temperature LSA step after the source/drain spike RTA in a standard CMOS flow [1]. The device performance benefits from this scheme are due to reduction of poly depletion, and reduction of series resistance of the SDE and SD areas. By lowering the spike RTA temperature and adjusting the halo, SDE, and SD implants, the benefit of scaling is added to the benefits of higher activation [7]. This one-step LSA integration scheme is most effective for NMOS because higher temperatures are generally required to achieve the highest dopant activation. Typical improvements are 5-10% gain in NMOS drive current, and more modest improvements in PMOS drive current (0-4%).

Recently, alternative junction engineering schemes using different LSA insertion points have been shown to give additional device performance improvements. For example, Yamamoto et al [8] showed large performance improvements in NMOS (12.8%) and PMOS (8.2%) by using a two-step LSA integration scheme. LSA was inserted immediately after SDE and halo implants, which lowered the resistance and suppressed transient enhanced diffusion for the PMOS SDE. LSA was also inserted between SD implant and spike RTA, which enhanced NMOS activation and reduced junction leakage. Various multi-step LSA integration schemes are currently being used in IC manufacturing [9].

B. Compatibility with strain engineering:

For LSA applications in the high temperature regime described above, compatibility with strain engineering is essential for successful integration of LSA. The most common and challenging issue is compatibility of LSA with embedded SiGe (e-SiGe), which is commonly used to introduce uniaxial compressive stress on the PMOS channel to enhance mobility. LSA has been shown to be compatible with e-SiGe techniques [10], but the challenge becomes greater as devices are scaled and Ge concentrations are increased. One of the common manifestations of excessive stress during advanced annealing is wafer-level plastic deformation leading to gate-to-contact photolithographic overlay errors.

Silicon behaves as a viscoelastic material, so the yield stress increases with increasing strain rate. This behavior is advantageous for LSA integration, as the wafer becomes more resistant to plastic deformation as the dwell time is reduced. The increased resistance to plastic deformation allows higher peak temperatures to be maintained, which can enhance activation and device performance. Lower dwell times can be achieved readily on the LSA system by simply increasing the stage speed, while simultaneously increasing the power density to the wafer plane in order to maintain the desired peak temperature. The effect of LSA dwell time on wafer stress and overlay errors has been studied extensively [11, 12, 13, 14, 15]. In one case study [14], reducing the dwell time by a factor of two reduced the maximum gate-to-contact overlay error by a factor of two. In addition to the lower dwell time capability, LSA has another characteristic that is advantageous for suppressing plastic deformation [11]. During LSA, only a small portion of wafer is heated up at a given instant (~5%), so the mechanical energy can be effectively dissipated to the bulk wafer in three dimensions. This greatly reduces the thermal shock

that is experienced by the wafer compared to other advanced annealing techniques which deliver all of the energy to the wafer in a few milliseconds. The effective dissipation of mechanical energy also makes the system robust to wafer breakage, which is critical for implementing any advanced annealing tool in high volume production.

LSA has also been shown to be a promising technology for forming Si:C, which is a strain engineering approach being considered for NMOS at the 22nm node. Carbon incorporation into the NMOS SD can be achieved by selective epitaxy, or by solid phase epitaxial (SPE) of the SD region after ion implantation of carbon. For the SPE approach, LSA has demonstrated a higher concentration of substitutional carbon compared to RTA due to the high temperature of SPE regrowth [15]. The high LSA annealing temperature also improves dopant activation, which is necessary to fully take advantage of the strain enhanced channel mobility [16].

C. Compatibility with High k – Metal gate

The use of High k – Metal gate (HK-MG) has been introduced into IC manufacturing to drastically reduce gate leakage compared to the traditional PolySiON gate [17, 18]. Although HK-MG nullifies the LSA advantage of reduced poly depletion, the high activation and negligible diffusion of advanced annealing is still required for high performance devices to reduce series resistance while maintaining scaling. In fact, the Rs reduction becomes more important as devices scale because of the increasing series-to-channel resistance ratio. Therefore any advanced annealing technology must be compatible with HK-MG. One question for a gate-first process is whether the metal gate could affect surface emissivity and hence the within-die absorption uniformity. Wang et al. [19] showed that the LSA approach is still effective for HK-MG through reflectance measurements on blanket films as well as real device layouts. There are two reasons for the effective suppression of pattern effects by LSA in the presence of metal gate: 1) The metal electrode layers are thin enough that the are semi-transparent at CO_2 wavelength, and 2) on a real device wafer, only the gate area has HK-MG, which will generally comprise less than 50% of the die area.

IV. LSA LONG DWELL TIME REGIME

As devices are scaled to the 22nm node, scaling requirements may prohibit the use of spike RTA to form shallow junctions. The thermal budget of spike RTA even at low temperatures will lead to diffusion of the SDE and SD implants, which may result in unacceptable short channel effects. One of the advantages of spike RTA is that it anneals the end of range (EOR) defects caused by the SDE and SD implants. If one eliminates the RTA step, and uses high temperature LSA only, these defects may not be annealed sufficiently due to the short dwell time. In order to anneal defects without excessive diffusion, we may need to access the time scale range between LSA high temperature regime the RTA spike anneal regime, depicted in Fig. 1 and the LSA "long dwell" regime. There has been some previous work showing that the use of longer dwell time anneal may have advantages for defect recovery in HK-MG devices [20].

As described earlier, the introduction of the preheat beam with a larger beam width allows access to the "long dwell" process space, where dwell times from a few milliseconds to 10's of milliseconds can be achieved. The wider beam allows one to achieve longer dwell times while still maintaining a high enough stage speed for high throughput. Peak temperatures for such applications are expected to be in the range of 800-1200°C, depending on the particular device and integration scheme. To characterize the effect of dwell time on junction leakage for SDE, we have done some initial characterization work on blanket wafers with a Ge pre-amorphization implant (PAI) followed by B 200eV 1e15cm^{-2} implant. The junction leakage is measured using the RSL technique for dwell times from 275usec to 3.2msec, and for temperatures from 1150 to 1350°C. The data, shown in Fig.5, clearly shows that increasing the dwell time reduces the temperature for minimum RSL leakage, presumabely due to annealing of EOR defects. Further characterization is underway to explore longer dwell times and lower temperatures.

Another benefit of introducing the preheat beam is stress reduction. In one experiment, we annealed blanket implanted wafers using the preheat beam and CO_2 beam simultaneously. Implant conditions are B 5keV 2e15cm^{-2}, and the dwell time of the CO_2 beam is 800usec. Wafers were annealed at increasing peak temperatures, and the slip threshold was determined by optical inspection using a Nomarski microscope. The results are shown in Fig.6. The introduction of the preheat beam increase the slip threshold by 70°C. In this experiment, the strain rate near the peak temperature is independent of the presence of the preheat beam, from which we conclude that the increased slip threshold is due to the reduction in thermal gradients during anneal. This characteristic of the preheat beam may be advantageous for

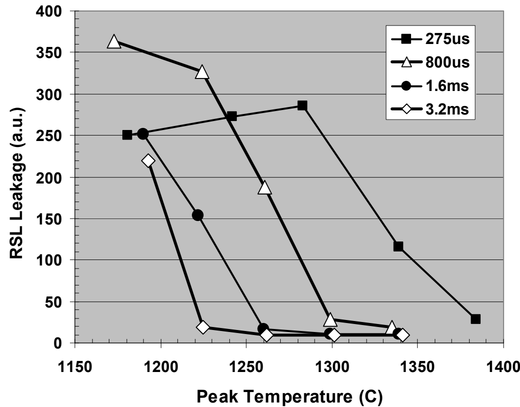

Figure 5: Effect of dwell time and peak temperature on RSL leakage (a.u.) for implanted blanket wafers (Ge PAI + B 200eV 2e15cm^{-2}).

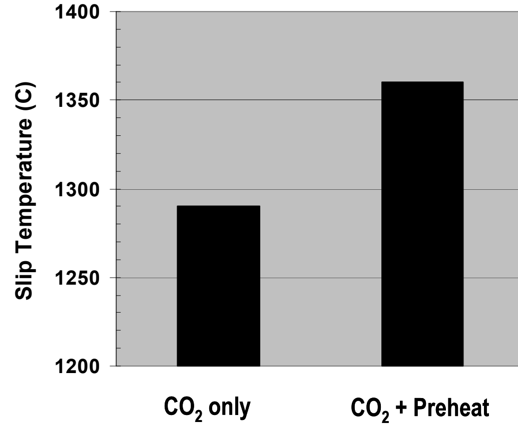

Figure 6: Slip temperature for blanket implanted wafers (B 5keV 2e15cm^{-2}), annealed with and without the preheat beam.

combined long dwell/high temperature processes, where the high temperature process provides high activation with low stress, and the preheat results provides annealing of EOR defects.

V. LOW TEMPERATURE LSA

Nickel has emerged as the material of choice for silicidation of contacts for sub-90nm device nodes. Compared to its predecessor, cobalt silicide, the benefits of nickel silicide include scalability to narrow line widths, compatibility with SiGe, and reduced silicon consumption [21]. Nickel silicide is widely used at the 32nm node, and it is likely that it will be extended to the 22nm node, at least for planar transistors. The integration of the nickel silicide process has its own challenges, including excessive nickel diffusion and poor thermal stability of the low resistivity phase. Optimization of thermal processes used to form nickel silicide has played a key role in addressing these integration issues. A two-step silicide process using rapid thermal annealing has been widely adopted to avoid excessive silicidation. Recent improvements in low temperature measurement and control in RTP have been critical in extending the nickel silicide technology to smaller device nodes [22]. The T-t regime for the RTP silicide steps are shown in Fig 1, where temperature as low as 250°C are used for the first step, and temperatures of 400 to 500°C are typically used for the second step which forms the low-resistivity NiSi phase.

As devices scale to 32nm and beyond, micro- and millisecond annealing techniques may play a key role in improving the nickel silicide process. Potential benefits of higher temperatures and/or shorter times of LSA are: 1) reducing the diffusion of nickel to undesirable areas of the transistor such as the gate dielectric or STI, and 2) maintaining high dopant activation close to the NiSi/Si interface. The expected process window for LSA is shown in Fig. 1, where the dwell time range is 100's of microseconds to a few milliseconds, and the temperature range is 400 to 900°C.

We have made improvements to the laser spike annealing system for low temperature processes appropriate for nickel silicide. First, the temperature measurement and control system has been modified to allow closed-loop temperature control for temperatures as low as 400°C. Also, we can lower the chuck temperature to less than 200°C by using the preheat beam at low power to heat the silicon wafer just enough to allow the absorption of CO_2 wavelength. The preheat beam heating is minimized so that almost all of the heating is done by the CO_2 laser.

We have used this LSA system for basic characterization of 100Å nickel films on blanket silicon wafers, from which we make the following observations. The temperature range for the transition from the nickel-rich phase to nickel monosilicide phase takes place between 500 and 600°C. This transition temperature range is similar to previously reported results using short wavelength laser annealing [23]. Although we do not have the RTP transition curves for these particular wafers, data from the literature indicates that transition range for similar wafers is 300 to 400°C [24]. The critical

temperature for agglomeration using LSA (as determined by increase in Rs and optical inspection) is approximately 850 to 900°C, depending on Pt concentration. The typical temperature for this transition using RTP is approximately 600°C [24]. For these two critical transition points in the silicidation process, the use of LSA increases the transition temperatures by approximately 200°C compared to the RTP processes used today. This drastic increase in process temperatures and reduction of annealing times may prove to be beneficial for nickel silicide formation for the reasons stated previously.

One concern when characterizing any new process for micro- or millisecond annealing is pattern effects, which can degrade within-die uniformity so that the process is not manufacturable. The largest concern for pattern effects in nickel silicide is the second annealing step, where the unreacted nickel is etched away, leaving a large variation in pattern density within a die. We expect that the nature of pattern effects for this step will be similar to the pattern effects seen for junction formation step, but could be worse due to the presence of a highly conductive silicide. To quantify the within-die temperature uniformity during the second silicidation step, we have mapped the reflectance of a nickel silicide device wafer during LSA, shown in Fig. 7. The reflectance mapping tool is a specially designed lab system with a small beam size, which allows the mapping of the die with a spatial resolution of approximately 100μm. Since the spatial resolution is on the order of the heat diffusion length, the measured reflectance variations will be a good representation of the expected temperature variations. The histogram of the absorptance variations within a single production die is shown in Fig. 7, where the absorptance is calculated as unity minus measured reflectance at each point. The mean absorptance for LSA is approximately 97%, and the absorptance variation around the mean is approximately ± 1%. Using a simple model that assumes temperature variations are linearly proportional to variations in absorbed power, the within-die temperature range is approximately ±5°C at a process temperature of 800°C. We have confirmed very similar behavior for several other device layouts. We believe that this tight temperature control using LSA will be critical achieving the high yields required for mass production.

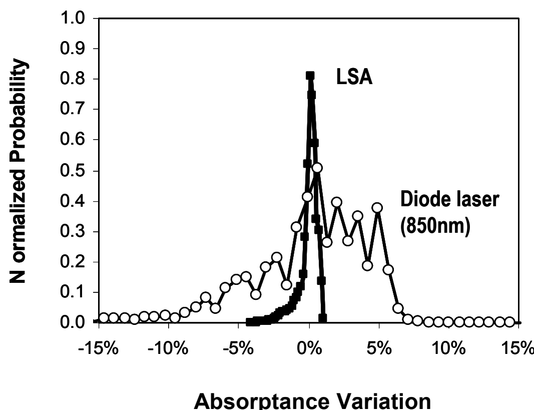

Figure 7: Histogram of within-die absorptance variation of a NiSi device wafer, for LSA and a diode laser (λ=850nm) at near normal incidence.

To further understand the effect of the wavelength of the annealing source, we also measured the reflectance of the same device wafer using a diode laser (λ = 850nm) at near normal incidence, also shown in Fig. 7. For this heating method, the mean absorptance is approximately 70%, and the range of absorptance across the die is approximately ±8%. This variation in absorptance will lead to within-die temperature range of approximately ±45°C for a mean temperature of 800°C.

VI. SUMMARY

The addition of a preheat beam to the standard CO_2 beam used in LSA allows access to all three T-t regimes discussed in this paper. For the high temperature regime, which is mainly used for dopant activation, the low dwell time capability of LSA is effective in maintaining high annealing temperatures for devices which use aggressive strain engineering techniques such as such as e-SiGe. Also, the LSA approach to pattern suppression is extendible to advanced devices using HK-MG. For the long dwell time regime, which may be used for defect annealing in an LSA-only approach to junction activation, initial data presented in this paper shows that increasing the dwell time beyond one millisecond reduces junction leakage in blanket implanted wafers. The introduction of the preheat beam also has advantages for reducing stress during high temperature LSA, as demonstrated by the significant increase in slip threshold observed on blanket implanted wafers. Access to the low temperature regime for middle-of-line applications such as nickel silicide formation is achieved by the use of a preheat beam and improvements in the LSA temperature measurement and control system. Within-die

reflectance measurements show that the LSA approach to pattern effect suppression is still very effective for nickel silicide device wafers (±1%), whereas the use of short-wavelength diode laser leads to pattern effects which are much larger than long-wavelength LSA.

REFERENCES

1. S.K.H. Fung et al., *Symp. VLSI Tech. Dig*., p. 92, 2004.
2. S. Chen et al., *Proc.RTP2007*, 2007.
3. L. Feng et al., *Proc. IWJT2006*, p.25, 2006.
4. J. McWhirter et al, *Proc. RTP2008*, 2008
5. S. Talwar et al., *US patent 7,148,159*, granted Dec. 12, 2006.
6. S, Talwar et al., *US patent 7,494, 492*, granted Feb. 24, 2009.
7. T. Yamashita and P. Fisher, ESSDERC, September 2006.
8. T. Yamamoto et al., *IEDM Tech. Dig*., p. 143, 2007.
9. D. Lammers, interview published in *Semi. Int.,*, Feb. 7, 2008.
10. Z. Luo et al., *IEDM Tech. Dig*., p. 495, 2005.
11. D.M. Owen et al., *RTP2008 Conf. Proc.*, p. 127, 2008.
12. D.M. Owen et al., *Proc. IWJT 2009*, p. 15, 2009.
13. S. Shetty et al., *Proc. IWJT2009,* p. 119, 2009.
14. Y. Wang, *Proc. RTP2008*, 2008
15. A. Li-Fatou et al., *ECS Trans*. **11**(6), p. 125 (2007).
16. H. Maynard et al., *Proc. RTP2008*, 2008.
17. X. Chen et al., *Symp. VLSI Tech. Dig*., p. 88, 2008.
18. C.H. Diaz, *Symp. VLSI Tech. Dig*., p. 629, 2008.
19. Y. Wang et al., Proc. *IWJT 2009*, p. 110, 2008.
20. A. Takayuki et al., Proc. *IWJT 2009*, p. 110, 2009.
21. J.P. Lu et al.*,Proc. IWJT2006*, p. 127, 2006.
22. B, Ramachandran et al., *Solid State Tech*, October 2006.
23. B. Adams et al., *Proc. RTP2007*, 2007.
24. V. Carron, *RTP2008 Workshop*, 2008.