Design for manufacturing (DFM) and reducing device variation are key to attaining high yields at the 65 nm node and beyond. Examples of changing equipment designs to meet these needs include a switch from batch to serial high-current implanters, and a switch from radiative to conductive rapid thermal processing. At 45 nm, a switch to iso-scan medium-current implanters and millisecond annealing will be necessary. At 32 nm, elevated source/drain (S/D) doping may be dictated by retained dose limitations.

In a keynote address in September, Thomas Skotnicki from STMicroelectronics (Crolles, France) projected that, if the NMOS channel doping level is kept constant at 10^{19}/cm³, the gate at the 32 nm node would have a gate length (L_g) of 16 nm and only 3.5 boron atoms under the gate; at an L_g of 12 nm, only 2.6 boron atoms would be needed; and at an L_g of 8 nm, 1.7 boron atoms would be under the gate. If a single boron dopant atom is missing, it can change the device threshold voltage (V_t) significantly. Asen Asenov, professor at the University of Glasgow (Scotland), recently reported that a change in the poly gate grain size can also significantly change the device V_t, especially for gate lengths <30 nm.

Also in September, Sunit Rikhi of Intel (Santa Clara, Calif.) was quoted in defining DFM as keeping variability in check. That same month, Tze-Chiang Chen of IBM (White Plains, N.Y.) identified one of the key challenges in extending silicon CMOS technology as reducing process variability in terms of chip/wafer/lot uniformity (mean deviation), regionally (systematic and random), and locally (systematic and random). He said the following process variabilities can cause V_t shifts of >0.1 V:

- Process proximity effects
- Layout loading effects
- Gate line edge roughness
- Implant dopant positioning
- Thermally induced variation by RTA

With scaling, implant and annealing process variability begins to cause significant device threshold voltage (V_t) variation. Device manufacturers have two choices: either design the process around it or change the process equipment. We explore the options at the 65 to 32 nm nodes for DRAM and logic devices.

The key lies in characterization, reduction and accommodation of these effects.

65 nm DRAM gate doping

With the introduction of p+ dual poly gate doping for DRAMs at the 65 nm node, new, very high-dose boron implantation techniques in the 1-4 keV energy range and up to 10^{17}/cm² doses are now being implemented. In situ phosphorus-doped n+ amorphous polysilicon deposition is used for DRAM poly gate electrode, but for dual poly gate (n+ and p+) the PMOS p+ electrode is formed by counter doping with boron to compensate the phosphorus-doped n+ regions. This boron implant dose is much higher than that used for logic device p+ poly doping (10^{16}/cm² to 10^{17}/cm² vs. mid-10^{16}/cm²). With traditional beamline ion implanters and boron implant energies in the 1-4 keV range, such high doses would require deceleration-mode implantation for productivity.

However, the gate-channel doping level is sensitive to any boron dopant penetration and therefore energy contamination. As DRAM companies developed p+ poly processes, they quickly discovered that very high deceleration ratios would be required to achieve production with hafnium-based oxide with poly for low standby power (LSTP) devices, while most other companies will stay with SiON with higher nitrogen content. Poly gate will possibly switch to fully silicided (FUSI) nickel or a metal/poly stack.

1. At the 45 nm node, one company is likely to use hafnium-based oxide with poly for low standby power (LSTP) devices, while most other companies will stay with SiON with higher nitrogen content. Poly gate will possibly switch to fully silicided (FUSI) nickel or a metal/poly stack.

John Ogawa
Borland,
J.O.B. Technologies,
Aiea, Hawaii
Process Variability Reduction for Gate Doping and USJs

boron beamline implantation throughputs, but device degradation from boron penetration made this impractical, leaving two alternative choices: plasma or beamline implantation using the molecular dopant species $B_{18}H_{22}$.

For this reason, DRAM companies like Micron (Boise, Idaho), Hynix (Seoul, Korea) and Qimonda (Dresden, Germany) have recently reported on higher productivity boron p+ poly doping using $B_{18}H_{22}$ for beamline and $B_{2}H_{6}$ for plasma implantation. Micron compared beamline boron at 4.5 keV/6 $\times$ 10$^{15}$/cm$^2$ dose with $B_{2}H_{6}$ plasma doping at 6 keV/1 $\times$ 10$^{15}$/cm$^2$ dose. The plasma dose was almost 2x higher than the beamline dose because only ~40% of the dose is retained after plasma implantation (4 $\times$ 10$^{15}$/cm$^2$). Plasma implantation has a unique wedge-like dopant depth profile signature with dopant pileup at the surface, while beamline implantation has a retrograde surface dopant profile. Because of these differences, after annealing it was shown that the beamline-diffused boron dopant profile level near the poly/oxide interface was >2.5 $\times$ 10$^{20}$/cm$^3$ by secondary ion mass spectroscopy (SIMS) analysis compared with plasma, which was <1 $\times$ 10$^{20}$/cm$^3$. The spreading resistance profile (SRP) showed the beamline electrical boron dopant level to be >2 $\times$ 10$^{20}$/cm$^3$ and dopant level from the plasma was <8 $\times$ 10$^{19}$/cm$^3$.

Hynix also reported that, with plasma doping, the boron dopant profile is piled up at the wafer surface, so that after photore sist strip and cleaning, which removes 4 nm of surface silicon, up to 70% of the boron dose is lost. After beamline implantation, the boron profile is deeper and retrograde near the surface, so after strip and cleaning only 10% of the boron dose is lost. They concluded that a plasma implantation process would require 3x higher boron dose (4.5 $\times$ 10$^{19}$/cm$^3$) compared with beamline implantation (1.5 $\times$ 10$^{16}$/cm$^3$) to achieve equivalent poly depletion rate.

In other papers, results on developing a drop-in $B_{18}H_{22}$ alternative beamline implantation process to monomer boron for the p+ poly DRAM dual poly gate doping were reported. Qimonda’s DRAM uses undoped poly, so no compensation of an n+ poly layer was needed and therefore the p+ poly boron dose was lower (only 6 $\times$ 10$^{19}$/cm$^3$).

Therefore, for DRAM dual poly gate, both approaches will be used for p+ poly boron implantation. In a July 2006 press release, Varian Semiconductor Equipment Associates (Gloucester, Mass.) claimed its plasma implantation system would be used for production DRAM p+ dual poly gate doping, and in October, Axcelis (Beverly, Mass.) announced a design win of the Optima HD Imax implanter, also for advanced memory dual poly gate p+ boron doping with $B_{2}H_{6}$.

Another problem for DRAM p+ poly doping not seen with logic devices is additional boron loss from the p+ poly caused by out-diffusion into the top tungsten layer. There was a 2.5x reduction in the poly boron level from 2.5 $\times$ 10$^{20}$/cm$^3$ to 1.0 $\times$ 10$^{20}$/cm$^3$ near the poly/oxide interface. A TiN/WN barrier layer was needed to prevent diffusion into the tungsten layer. This result was based on using beamline boron implantation; if plasma implantation was used instead, significantly more boron surface loss and out-diffusion would have occurred. However, at the 65 nm node, WSi$_2$ and not tungsten is still used by many DRAM companies, and no comparative data on boron out-diffusion into WSi$_2$ has yet been reported.

### 65 nm logic device, $T_{ox}$ inversion reduction

Starting at the 65 nm node, the SiON gate equivalent oxide thickness (EOT) stopped scaling at 1.1 nm, but the oxide thickness at inversion ($T_{inv}$) did not. By increasing the nitrogen content and poly dopant activation level, continual reduction in $T_{inv}$ is realized. The poly dopant activation level can be increased by using high-temperature, millisecond-laser or flash/rapid thermal anneal (RTA), reducing $T_{inv}$ by 0.1-0.2 nm, but a spike/RTA or soak/RTA is needed for dopant diffusion through the poly gate electrode material. The p+ poly sheet resistance ($R_s$) was reduced from 425 to 155 $\Omega$/sq, while the n+ poly resistance was reduced from 125 to 100 $\Omega$/sq. However, Chen showed no difference in the SIMS dopant depth profile level between spike/RTA and spike+ laser treatment, so SRP may have been a better technique to look for changes in profiles.

Therefore, the process integration flow is to add the flash or laser annealing step after the spike/RTA annealing process for the 65 nm node production. This offers <5% improvement to the p source/drain extension (SDE) activation if the spike/RTA temperature is kept above 1050°C from 6 $\times$ 10$^{19}$/cm$^3$ to 7 $\times$ 10$^{19}$/cm$^3$, yet without the spike/RTA, 1.2 $\times$ 10$^{20}$/cm$^3$ is achieved. If the spike/RTA temperature is reduced to 950-1000°C, the additional flash or laser annealing improves pSDE activation by -100% from 5 $\times$ 10$^{19}$/cm$^3$ to 9 $\times$ 10$^{19}$/cm$^3$. Intel reported similar results using phosphorus for nSDE, 6 $\times$ 10$^{20}$/cm$^3$ for spike/RTA, 1 $\times$ 10$^{21}$/cm$^3$ for spike or laser annealing.
for spike/RTA+laser and 1.8 × 10^{21}/\text{cm}^3 for laser alone.\textsuperscript{15}

**45 nm node logic metal gates**

At the 45 nm node, only NEC (Tokyo) will use hafnium-based oxide (medium-k) with poly for low standby power (LSTP) devices,\textsuperscript{16} while most other companies will use SiON with higher nitrogen content, but possibly switch from poly to either a fully silicided (FUSI) metal gate electrode\textsuperscript{16} or a metal/poly stack gate electrode (Fig. 1). Depending on the thickness of the poly in the metal/poly stack, some dopant diffusion will be needed, requiring a low-temperature spike or soak anneal in a gate poly pre-doping and annealing step. In all-metal gate electrodes, likely to be used in 32 nm logic devices, poly doping will not be needed.

**65 nm USJ & HALO engineering**

For the 65 nm node, there are no major issues with medium-current implanter design for ultrashallow junction (USJ) and HALO doping. However, there are changes being made with respect to high-current systems for junction depths in the 20-30 nm range.

**65 nm high-current implantation**

For the 65 nm node, DFM issues related to ion implantation can be found with the sudden change from batch implanters to serial (single-wafer) high-current implanters in manufacturing fabs around the world. The batch implanter spinning disk was a device yield killer because of the ballistic collision of particles with poly gate structures as gate length scaled to <90 nm.\textsuperscript{17} The cone-angle effects of the batch implanter was also seen as a source of \(V_t\) variation caused by asymmetrical SDE implantation across the wafer from left to right, especially for NMOS devices.\textsuperscript{18} The industry largely switched from batch to serial high-current implanters at 65 nm.

**RTA design**

At 65 nm, device \(V_t\) variation caused by lamp-based spike/RTA sensitivity to patterns can result in \(\Delta V_t\) variation to be 2x worse with radiative systems vs. conductive heating as gate length scales <40 nm.\textsuperscript{19,20} The within-die pattern density variation resulted in localized lamp heating and therefore deeper junctions. This effect could be reduced with much slower ramp-up rates or by changing the heating method to conductive heating using hot-wall annealers.\textsuperscript{19,21} Mattson (Fremont, Calif.) recently reported modifying its RTA lamp heating system by placing a “hot shield” in front of the wafer to block direct radiation from the lamps to eliminate this pattern sensitivity.\textsuperscript{22}

At 65 nm, a couple of high-performance logic companies are using spike/RTA-flash msec annealing sequence after S/D formation in production for improved poly activation and \(T_{\text{inv}}\) scaling.\textsuperscript{13} The spike/RTA diffuses the dopant in the poly layer, SDE and deep S/D, while the flash anneal improves poly activation. If the spike/RTA temperature is high (1050-1100°C) then <5% improvement in SDE activation is observed (7-8 × 10^{20}/\text{cm}^3). If the spike/RTA temperature is lower (950-1000°C), then >100% improvement in SDE activation is observed (5 × 10^{20}/\text{cm}^3 to 1 × 10^{21}/\text{cm}^3).

**45 nm node \(V_t\) variation**

Starting at the 45 nm node and definitely with the 32 nm node, where SDE is formed by diffusion-less activation, the precision of the HALO or pocket implant becomes critical in both angle and dose control. NEC reported the need for multiple HALO/pocket implants when using diffusion-less laser annealing for USJ formation and short-channel effect optimization.\textsuperscript{23} Therefore, the HALO implant dose and angle precision, reproducibility and variation across a 300 mm wafer can be minimized with an iso-centric scanning system design.\textsuperscript{24}

Another change is the switch to indium (In) dopant species for p-type HALO and antimony (Sb) for n-type HALO to achieve super-steep retrograde doping profile in the channel region and dopant-free channel, so that fluctuations in the channel will not be an issue, as mentioned earlier by Skotnicki.\textsuperscript{1} To enhance HALO dopant activation with USJ diffusion-less annealing, molecular dopant species (\(\text{B}_9\text{H}_{14}\), \(\text{B}_{18}\text{H}_{22}\)) and higher mass dopants (\(\text{P}_2\), \(\text{P}_4\), \(\text{As}_2\), \(\text{As}_4\), etc.) may also be introduced for HALO implantation.

**High-current implanter (\(x_j=12-20\) nm)**

With less dopant diffusion, localized variation in angle control with serial high-current implanters means <7° tilt quad-mode implantation is required for SDE to minimize gate length variation and therefore \(V_t\) variation. Varian reported on using quad-mode 4° tilt implantation for SDE to minimize local angle variation on the VIISta-HC ribbon beam implanter design.\textsuperscript{25} This solution is very similar to that reported by Axcelis, which used a >5° tilt quad-mode SDE implantation to eliminate the batch implanter cone-angle effect on local implant angle variation and asymmetrical device.\textsuperscript{26}

The other four serial high-current implanters also have their own unique implant non-uniformity signatures based on the wafer scanning mechanism and method:

- Applied Quantum-X with its 2-D mechanical scan
- Axcelis Optima-HD with its high-speed 2-D mechanical scan
- Sumitomo Eaton Nova SHX with its 1-D mechanical scan and 1-D beam scan
- Nissin Exceed Cluster with its 1-D mechanical scan and 1-D beam scan

To correct for any across-wafer and local variation in angle or dose, tilted quad-mode or multi-mode (8 to 16 mode) implantation may be necessary. But when high-angle 15-30° tilted SDE implant is needed to
achieve the desired gate overlap control with diffusion-less annealing, then any localized angle and dose implant variations will require octa-mode (45° rotation) or even 16-mode (22.5° rotation) tilted implantation. If gate line edge roughness dominates $V_t$ variation because of $L_g$ variation, then high-tilt multi-mode SDE implantation will also be needed.

With low-energy beam blow-up, energy contamination concern and productivity issues, new higher-mass dopant species will be introduced, such as $B_{10}H_{14}$ and $B_{18}H_{22}$ for p-type dopants and $As_7$, $As_9$, $P_9$, $P_4$, and $Sb$ for n-type dopants. Also, enhanced dopant activation using molecular dopant species and higher-mass dopant species when using flash, laser or SPE diffusion-less activation annealing techniques have been reported.27

**Millisecond flash/laser anneals**

At the 45 nm node, lower-temperature spike/RTA or soak/RTA annealing will be used only for poly dopant activation and diffusion, and not for SDE (Fig. 2). If metal gate electrodes are successfully introduced, replacing poly for gate electrode, then RTA will not be used at all. As described by NEC, millisecond flash or laser annealing will be integrated into the process flow after HALO, SDE and deep S/D implantation for 45 nm node LSTP devices to achieve diffusion-less activation with <3 nm of dopant movement for SDE.23 In its process flow, NEC reported first using a separate gate poly pre-doping and pattern followed by spike/RTA annealing to diffuse the dopant in the poly layer. Then off-set spacer, HALO and SDE implantation, sidewall spacer and S/D implantation are performed, followed by millisecond laser annealing (or flash annealing option) for USJ formation with diffusion-less activation, followed by nickel salicidation and metallization. If the millisecond flash or laser annealing techniques are sensitive to within-die pattern density variation and result in localized heating and $V_t$ variation, this effect can be reduced by using a capping layer for DFM uniformity improvement or changing anneal equipment to a longer-wavelength laser or longer-pulse-duration flash that is not sensitive to pattern density variation.

However, laser overlap stitching pattern (stripping) non-uniformity is another issue that can affect $V_t$ variation and DFM. Detection of localized heating variation on a micro level requires new metrology equipment not usually used by the industry today. Photo luminance (PLi) analysis and elastic material probe ($R_s$ and $N_{surf}$) microline scans with microresolution of <50 µm have been used for both flash lamp and laser scan annealing localized microvariation detection. Each individual flash lamp signature was detected by PLi and correlated to localized $R_s$ and $N_{surf}$ microvariations. This was also seen with the laser beam scanned stitching signature. New flash and laser equipment designs to eliminate these localized heating signature

---

**Retained dose at such shallow junctions both after implant and annealing could be the No. 1 issue at the 32 nm node.**

---

**Lydall Affinity Chillers**

**Accomplish one thing: Provide the customer peace of mind**

Lydall global customer support network provides comprehensive solutions to semiconductor manufacturing centers worldwide.

---

**Service includes:**

- Installation assistance
- Preventative/Predictive Maintenance
- Complete Refurbishment Services
- Training Programs
- Regional Consumables & Spares
- Regional Repair Facilities

---

**Worldwide Locations:**

- New Hampshire: 800 738-8581
- Texas: 512 442-2832
- Singapore: 65 6841-6283
- California: 650 508-2200
- Germany: 49 8709 926-10
- Taiwan: 886 3578-1585

---

www.lydallaffinity.com
variations must be developed to reduce $V_t$ variation and achieve the DFM requirements for 45 nm node production and beyond.

**High-current implant ($\chi$, of 9-12 nm)**

If 3 nm of dopant diffusion is assumed with flash, laser or SPE annealing techniques, then a 6 nm implanted junction is required. Figure 3 shows the implant energy required to realize these shallow junctions using beamline implantation with boron, and BF$_2$, dopant species or BF$_3$, plasma implantation defined at $5 \times 10^{19}$/cm$^2$. Retained dose at such shallow junctions both after implant and annealing could be the No. 1 issue at the 32 nm node to achieving the desired dopant activation level of $>1 \times 10^{19}$/cm$^2$ (Fig. 4). Using a BF$_2$ or BF$_3$, dopant source, the maximum retained dose for a sub-5 nm junction is $<2 \times 10^{19}$/cm$^2$; after annealing, this level can drop to $<5 \times 10^{19}$/cm$^2$, making a $1 \times 10^{20}$/cm$^2$ dopant activation target level impossible to realize. With boron, $1 \times 10^{19}$/cm$^2$ retained dose can be realized but data for BF$_3$, and BF$_2$, and BF$_3$, is still needed. With millisecond flash annealing, if the retained dose before annealing is $3 \times 10^{19}$/cm$^2$, a BF$_3$, of $5 \times 10^{19}$/cm$^2$ could only be realized if $6 \times 10^{19}$/cm$^2$ was the retained dose before annealing then a BF$_3$, of $9 \times 10^{19}$/cm$^2$. Therefore, plasma doping with BF$_3$, is out, but is usable with BF$_2$, while BF$_3$, is out but boron is usable for beamline doping. Enhanced activation with diffusion-less activation is achieved when using BF$_3$, and BF$_2$, so this effect may be desirable to meet the 32 nm node targets. Another option reported by Wakabayashi of NEC is to use undoped elevated source/drains, 4 nm thick. This allows higher-energy implantation for higher retained dose with controlled scaling of the SDE junction below the gate edge.

**Summary**

At 65 nm, DRAM dual poly gate formation will use either B$_2$H$_6$, beamline or plasma implantation at doses in the $10^{16}$/cm$^2$ to $10^{17}$/cm$^2$ range. Logic devices will use millisecond flash or laser annealing for improved gate poly dopant activation, with 0.1-0.2 nm reduction in $T_{inv}$. Also, serial high-current implanters will replace batch implanters, and lamp heating sensitivity to pattern density is likely to cause significant $V_t$ variation. At the 45 nm node, millisecond annealing for diffusion-less activation will be used for USJ formation so implant precision becomes critical, requiring iso-centric scan for median-current implanters and multi-mode tilt implantation on serial high-current implanters to minimize their unique non-uniformity signature effects. At the 32 nm node, implantation retained dose may limit what dopant species options are acceptable. Metrology tools to detect localized micro-process variations on the micron level caused by implantation and annealing equipment-based signatures are becoming more critical.

**Metrology tools to detect localized micro-process variations on the micron level caused by implantation and annealing equipment-based signatures are becoming more critical.**

**References**


**Acknowledgements**

The author would like to thank Paul Cheng of AIBT, Masayasu Tanjo of Nissin, and Hiroyuki Ito of UJT Labs for providing the retained dose information.

**John Ogawa Borland** is founder of J.O.B. Technologies. He received his B.S. and M.S. in material science and engineering from the Massachusetts Institute of Technology (MIT). Borland completed his thesis research on InP crystal growth at NTU Labs (Mitsubishi, Tokyo, Japan). During his 27 years in the semiconductor industry, he has worked for Varian Semiconductor, Genus, Applied Materials, National Semiconductor, Hughes Research Labs and NASA Marshall Space Flight Center. He is a senior member of IEEE and a member of the Electrochemical Society (ECS) and Materials Research Society.