Flat-Top Flash Annealing™ For Advanced CMOS Processing

Paul Timans, Gary Xing, Silke Hamm, Steve McCoy, Joseph Cibere, Greg Stuart, and David Camm
Outline

• Introduction
  – Key Annealing Requirements for Advanced Devices
  – The Flash-Assisted RTP™ Approach for Millisecond Annealing

• Advanced Doping Requirements
  – The Balancing Act:
    • Diffusion, Activation & Defect Annealing vs. Integration Issues
      o Opportunities through long-pulse millisecond annealing

• Limits on Millisecond Anneal Duration
  – Bulk Wafer Heating
  – Total Energy Requirement

• Process Benefits
  – Improving Kinetic Trade-off by Controlling the T-t profile
  – Dopant Activation Improvements - As & B Doping

• Conclusions
Key Annealing Processes for Advanced CMOS

- **Optimized dopant activation**
  - High Activation: Decrease $R_{\text{EXT}}$ - for both ion-implanted & deposited dopants
  - Control diffusion: Optimize gate overlap & placement of junction relative to interfaces/defects
  - Sufficient defect annealing

- **Optimize integration of strain and new materials**
  - Low thermal budget Ni(Pt)Si: Improve $R_C$, prevent “piping defects”
  - Low thermal budget anneals for high-K film
  - Maintaining channel strain (SiGe & Si:C)
  - Annealing for new channel materials Ge, InGaAs, ....

Many thermal budget constraints
Reduce $T$? and/or Shorten Time?
Millissecond Annealing with Flash-Assisted RTP™

- For anneals of < 0.5 s, surface-specific heating is essential:
  - Requires a pulsed energy source (>10 kW/cm²)
  - Flash-lamps or scanned lasers
- Flash-Assisted RTP™ (fRTP™):
  - Fast ramp (150 K/s) to $T_i$
  - Pulsed surface heating with proprietary water-wall arc lamps
  - Millios® tool provides real-time T measurement and control on front & back of wafer

**Diagram:**
- **Flash Lamps:** 4 Flash Lamps above wafer, exposing device side to an intense flash
- **Arc Lamps:** 2 Arc Lamps below wafer to heat to intermediate temperature

**Graph:**
- Fast ramp by flash heating
- Rapid cooling by radiation and thermal conduction to bulk

**Temperature (°C):**
- Top Temp
- Bot Temp

**Time (s):**
- 4.835 - 4.840 - 4.845 - 4.850 - 4.855
- 5.0 - 5.5 - 6.0
The Whole Wafer Can be Uniformly Processed in Just One Flash

- Process Uniformity: $3\sigma < 5K$; Range < 9K
- Fast ramp rates & whole-wafer flash exposure provide very cost-effective processing
The Ultra-Shallow Junction “Balancing Act”

- Millisecond annealing greatly improves activation compared to conventional RTA
- BUT....USJ needs a careful balance between diffusion, activation & defect annealing
  - Conventional MSA heating duration is too short for complete defect annealing: Longer pulses needed?
- Also: Many integration factors........

H.W. Kennel et al., RTP 2006 Conference

R.A. Camillo-Castillo et al. APL 88(1), 2006


---

- Time for Diffusion
- Time for Activation (50%)
- Temperature (ºC)
- Time (s)
- Furnace
- RTP
- Diffusion Length
- ms-anneal

---

- 1 nm
- 2 nm
- 5 nm
- 10 nm
- 20 nm

---

- 30kV Ge PAI Flash-assist RTP
- 8kV Ge PAI Flash-assist RTP
- 30kV Ge PAI 960ºC Spike RTA
- 30kV Ge PAI 760ºC RTP, 1100ºC RTA + 960ºC Spike RTA
- 8kV Ge PAI 706ºC RTP, 1100ºC RTA + 960ºC Spike RTA
- TEM Detection Limit
Integration Also Imposes (Dominant?) Requirements

- Activation, diffusion & defect annealing are just the starting point

- Many additional integration issues affect the choice of doping/annealing conditions
  - Some materials systems impose peak T limitations

- Sophisticated optimization of annealing conditions is required
  - Spike anneal / MSA integration
  - Process conditions: ramp rates, preheat T, peak temperatures, pulse durations……
  - Flexibility of anneal conditions is very valuable!
Extending the Length of Millisecond Anneal Duration Opens Up New Opportunities

• Motivation to explore the “long pulse” MSA regime
  – “Traditional MSA” < $t_{\text{anneal}}$ < Spike RTA
  – MSA without the need for a separate spike anneal
  – Longer pulse allows lower peak temperatures: More “integration friendly”

• Junction objectives:
  – Adequate defect annealing
  – Controlled nm-level diffusion (e.g. gate overlap control)
  – Maintain high activation

• Integration aspects:
  – Compatibility with strain scheme
  – Compatibility with gate dielectric approach

• Motivation for a “few ms” anneal at temperatures <1250°C
Millisecond Annealing Can Be Extended to ~10 ms

- T-t models holding the device side at a constant yet elevated temperature for longer times
- If time at temperature >10 ms the bulk of the wafer approaches temperatures where excessive diffusion and deactivation may occur

Beyond 10 ms, MSA looks more & more like spike annealing

Practical range of a long ms flash
- Peak temp up to 1300°C
- Duration up to 10 ms
Extending Pulse Length Increases Wafer Processing Energy Requirement

- Keeping the Si surface hot for longer increases in total energy requirement
- Calculations illustrate the effect for an increasing anneal duration (T-50K)
  - Assume semi-infinite wafer, so that heat-sinking remains effective
  - The flat-top profile minimizes the total energy requirement, less energy is wasted during preheat
  - Laser scanning at low scan velocity increases energy requirement
    - Heat loss parallel to the wafer surface as well as down into the bulk of the wafer
    - Very challenging for wafer throughput, especially given strong scan overlap requirement
Improving Kinetic Trade-Off with Flat-Top Profile

• MSA can exploit kinetics to optimize the trade-off between competing processes: e.g. One desired process & one undesired process
  - e.g. Activation of B implants:
    • $E_A$ (Diffusion) $\sim 3.5$ eV
    • $E_A$ (Activation) $\sim 5$ eV
  - Short anneals at high $T$ are very useful for activation/diffusion trade-off

• The effect of the anneal cycle can be summarized in $t_{eff}$

• Maximize:

\[
\frac{t_{eff - desired\_process}}{t_{eff - undesired\_process}}
\]

• The flat-top profile is most effective, because a larger fraction of the anneal is spent close to the peak temperature

$t_{eff} = \int_{0}^{t} \exp \left\{ -\frac{E_A}{k} \left( \frac{1}{T(t')} - \frac{1}{T_{REF}} \right) \right\} dt'$
Dopant Profile Tuning with the Flat-Top Flash: As

- Extended pulse duration at ~1200°C allows profile tuning
- Longer time reduces $R_S$ and increases profile abruptness
Dopant Profile Tuning with the Flat-Top Flash: B

- Once again, we see improved activation and profile abruptness as pulse length at 1200°C is extended
Improving Dopant Activation

- Increasing the pulse duration demonstrated improved activation with a small increase in junction depth
- **As**: Concentration-enhanced diffusion and increased abruptness improves $R_S$
- **B**: BIC dissolution combined with concentration-enhanced diffusion improves $R_S$

**Graph:**

- **Green line**: $2 \times 10^{15}$ As/cm$^2$, 2.5 keV
- **Blue line**: $2 \times 10^{15}$ B/cm$^2$, 1 keV

**Axes:**
- **Y-axis**: Sheet Resistance ($\Omega$/sq.)
- **X-axis**: Junction Depth ($5 \times 10^{18}$ cm$^{-3}$) (nm)

**Legend:**
- Increasing anneal duration (1 to 4 ms)
Millios® Flat-Top Flash Anneal Provides Device Benefits Compared to Spike RTA+Laser Annealing

- Long ms-flash was compared with a “combo process” of Spike-RTA + Laser Anneal (SOI process, poly/SiON gate)
- \( I_d - V_{gs} \) and \( V_{t(sat)} \) roll-off were very similar
- Long ms-flash NFETs needed only ~ \( \frac{1}{2} \) of the B halo dose and exhibit no anomalous corner leakage which is sometime found in spike RTA+laser NFETs. (Better B halo localization)

K.L. Lee et al. (IBM) and S. McCoy et al. (Mattson), IWJT 2010
Conclusions

• Extending the annealing time for millisecond anneals opens up a new regime for dopant activation & damage annealing combined with controlled diffusion

• Pulsed anneals at <1250°C for times up to ~ 10 ms can help overcome integration issues with new device structures
  – Beyond 10ms, the benefits of fast surface cooling are lost and the processing becomes very costly

• The Flat-Top Flash Annealing™ method brings additional advantages to millisecond annealing
  – Increasing the fraction of the anneal time spent at peak temperature can improve kinetic trade-offs
  – Most effective use of heat source energy, together with high wafer throughput
  – Significant improvements in doping profile shapes and activation