# Advanced Activation Using Various Thermal Budget Regimes Such As Flash, Multiple Flashes And Flash + Spike Annealing

J. Gelpey, W. Lerch, S. Paul, J. Niess, S. McCoy,
Sing-Pin Tay*
Mattson Technology Inc.

* Presenter

NCCAVS THIN FILM USER GROUP
August 20, 2008

mattson

A Partner You Trust. Performance You Value.

# Overview

- Introduction
- Experimental Procedures/Definitions
  - Time/Temperature Profiles
- Results and Discussions
  - Multi-Flash Annealing
  - Ambient Effects
  - Dopant Activation Strategies
- Summary and Conclusions
- Acknowledgements

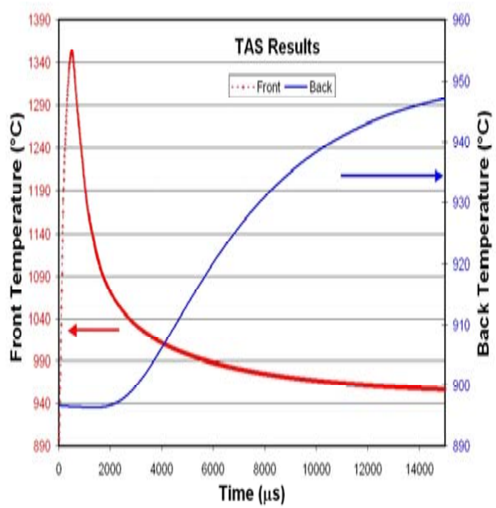
# Why ms-Annealing?

- As transistors continue to shrink, junctions must get shallower and more abrupt while electrical activation must get higher to improve or even maintain performance
- As power consumption becomes more of a concern, junction leakage becomes a more important issue so defect control is also critical
- Conventional spike RTP and beamline implantation can no longer meet the roadmap requirements

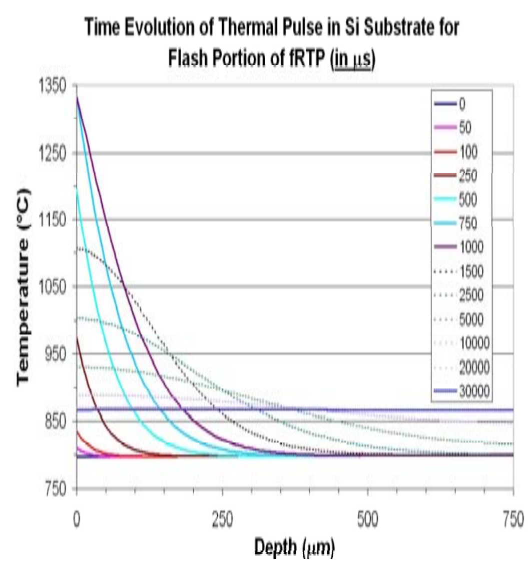
| Year of Production | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 |
|---|---|---|---|---|---|---|
| MPU/ASIC Metal 1 (M1) ½ Pitch (nm)(contacted) | 68 | 59 | 52 | 45 | 40 | 36 |
| Drain extension $X_j$ (nm) for bulk MPU/ASIC [F] | 12.5 | 11 | 10 | 9 | 8 | 7 |
| Maximum allowable parasitic series resistance for bulk NMOS MPU/ASIC × width ((Ω?µm) from PIDS [G] | 200 | 200 | 200 | 180 | 180 | 180 |
| Maximum drain extension sheet resistance for bulk MPU/ASIC (NMOS) (Ω/sq) [G] | 650 | 740 | 810 | 900 | 1015 | 1160 |
| Extension lateral abruptness for bulk MPU/ASIC (nm/decade) [H] | 2.5 | 2.3 | 2.0 | 1.8 | 1.6 | 1.4 |

# How ms-Annealing Works

- Spike anneals cannot get appreciably shorter since the entire volume of the wafer must be heated and cooled
- MSA works because the radiant energy pulse is shorter than the thermal time constant of the wafer so only the top surface of the wafers is heated by the flash or laser.  This allows much faster heating and cooling rates since the bulk of the wafer acts as a heat sink to the top of the wafer.




Front and back side temperature as a function of time in flash lamp annealing (modeled)



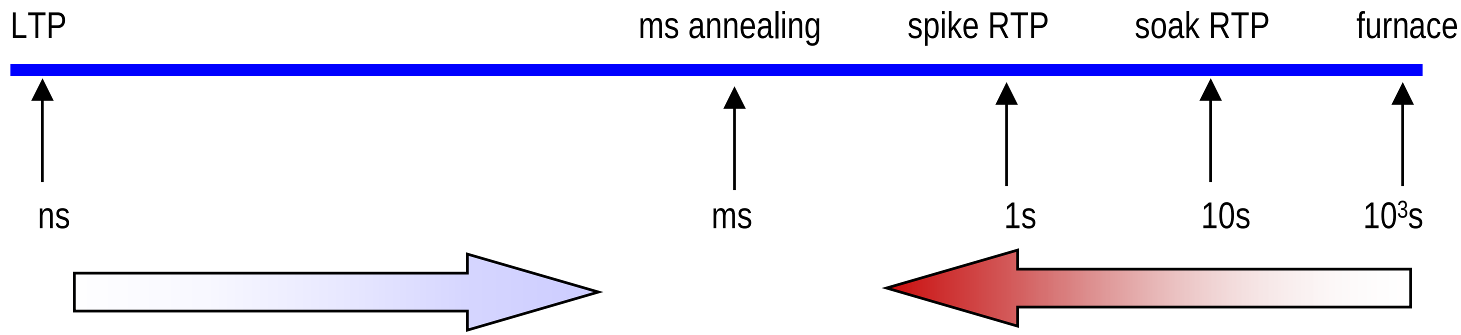

Evolution of temperature pulse in flash lamp annealing



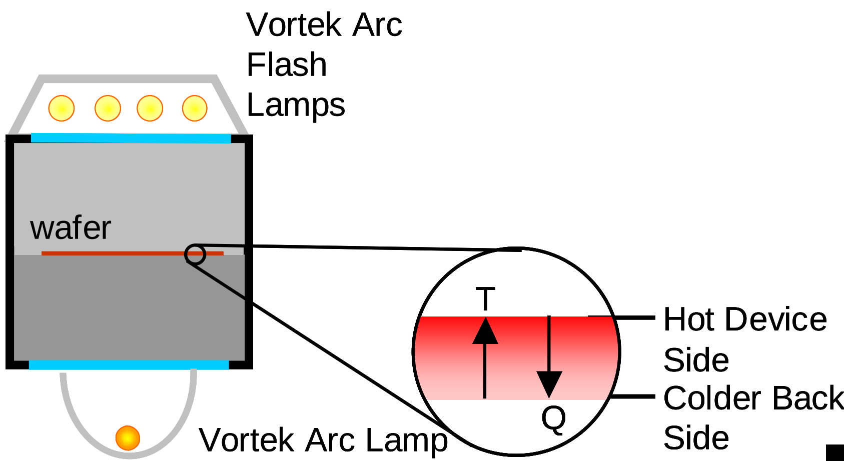
A Partner You Trust.
Performance You Value.

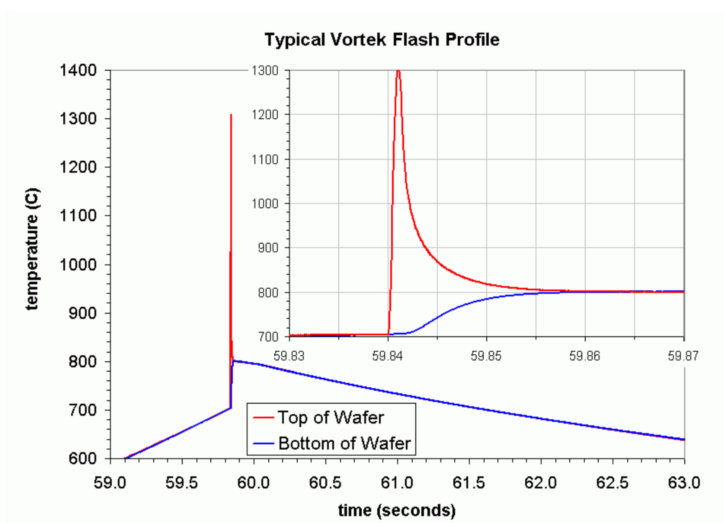
# History

- RTP times have been getting shorter (soak anneals ~10s, spike anneals ~1s), but these techniques heat the entire thickness of the wafer, so there is a practical limit on reduction of thermal budget.
- Laser annealing times have been getting longer (melt LTP in the ns to μs range), but these techniques have shown difficult integration issues.
- ms annealing by either flash lamps or lasers seems to hit the “sweet spot” (minimal diffusion, good activation, reasonable defect removal)



# Flash Lamp MSA Concept



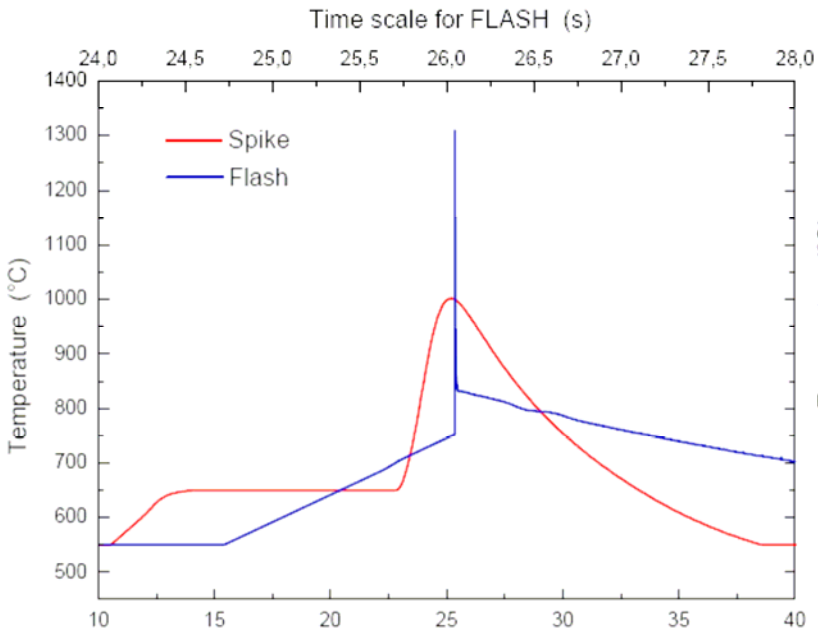



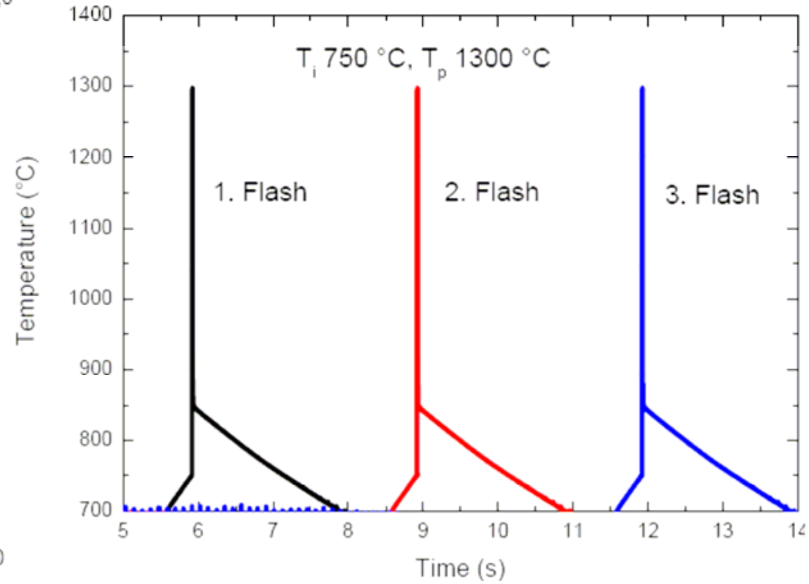

Real-time temperature measurement of both bottom and top of wafer

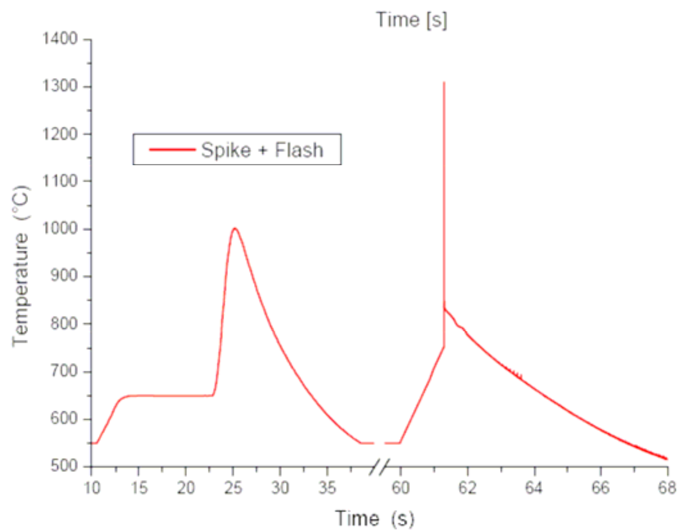
# Experimental Details

- Implant Conditions
  - $B^+$ 500eV ($\alpha$-Si, 30keV Ge [EOR≅50nm] and c-Si)
  - $As^+$ 1keV (c-Si) dose up to 1E15cm$^{-2}$
- Spike anneals in Mattson 3000 Plus
  - Prestabilization @650°C for 10s
  - Peak temperature 1000°C
  - 100ppm or 10% oxygen in nitrogen
- Flash anneals in Mattson Millios™ fRTP
  - Intermediate temperatures of 750° or 950°C
  - Peak temperatures of 1250° or 1300°C
  - 100ppm or 10% oxygen in nitrogen
- Analysis
  - Four-point probe sheet resistance KLA-Tencor RS100
  - Hall Effect measurement Accent HL5500
  - SIMS quadrupole CAMECA SIMS 4600
  - TEM JEOL 2100-HC with weak-beam dark-field technique

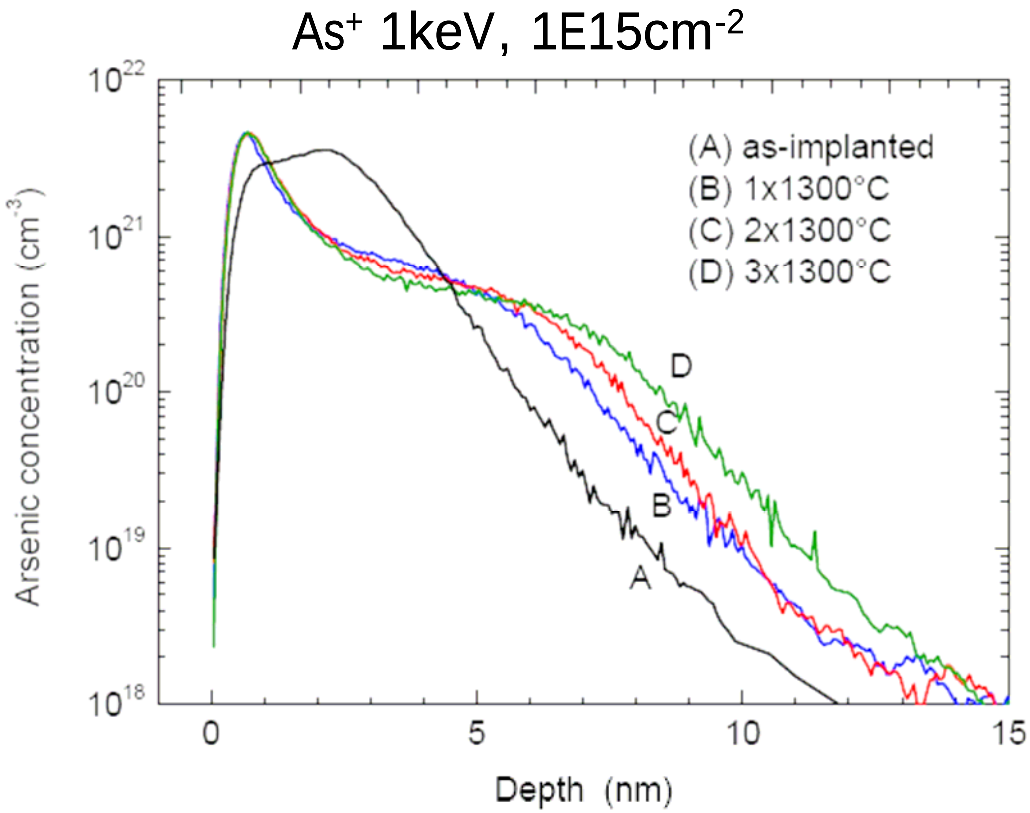
# Temperature/Time Profiles



# NMOS: Multiple Flash Results

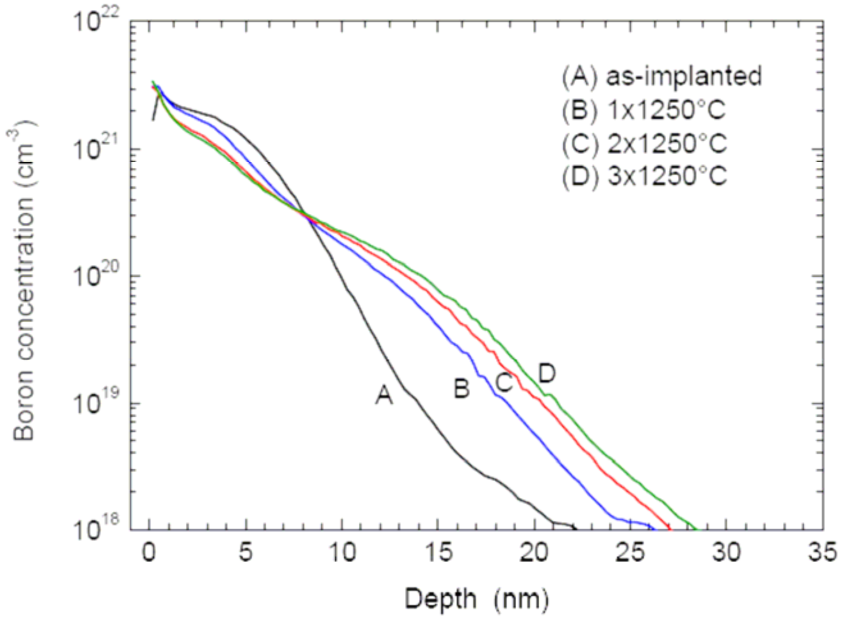



| | 4PP $R_s$ | Hall $R_s$ |
|---|---|---|
| B | 649Ω/sq | 660 Ω/sq |
| C | 571 Ω/sq | 576 Ω/sq |
| D | 532 Ω/sq | 539 Ω/sq |

Crystal Defects below TEM WBDF detection limit

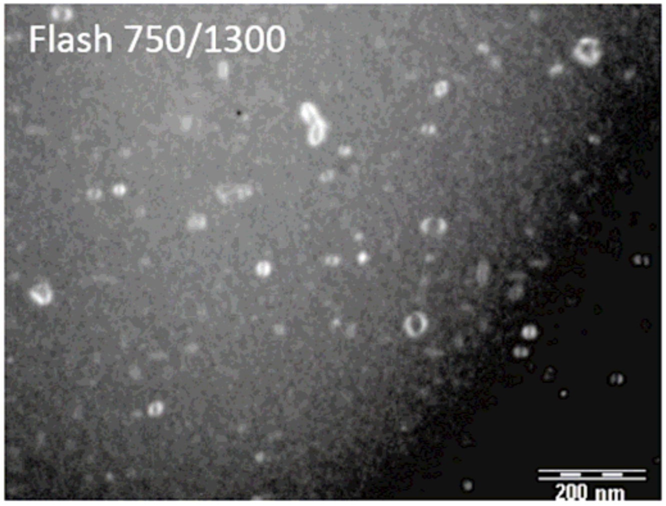
# PMOS: Multiple Flash Results

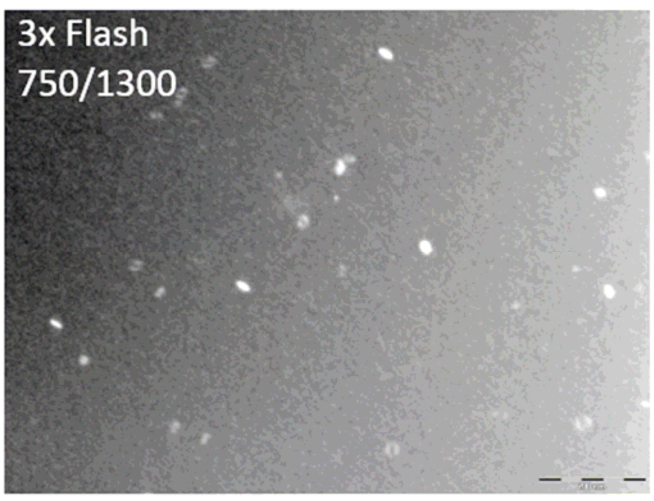
B$^{+}$ 500eV, 1E15cm$^{-2}$ into $\alpha$-Si




B 522Ω/sq., C 433Ω/sq., D 401Ω/sq.

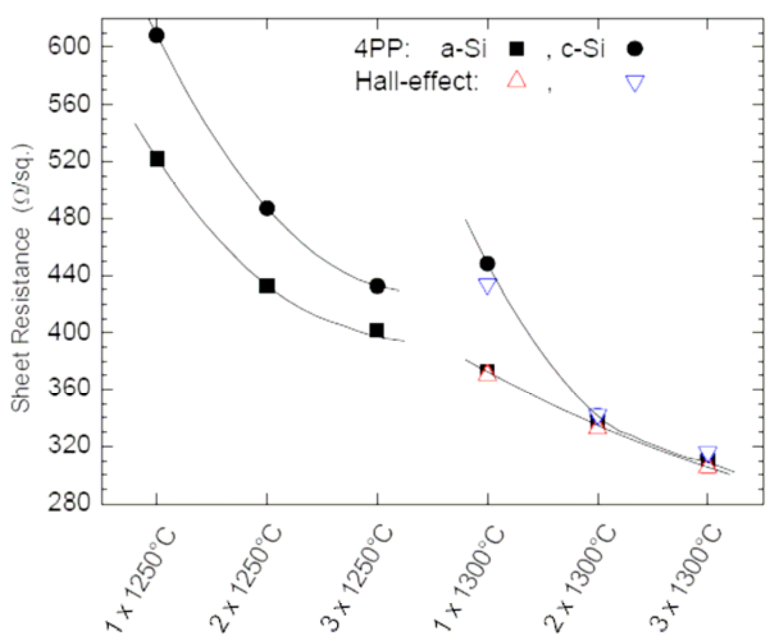
Extended defects remain after anneal.



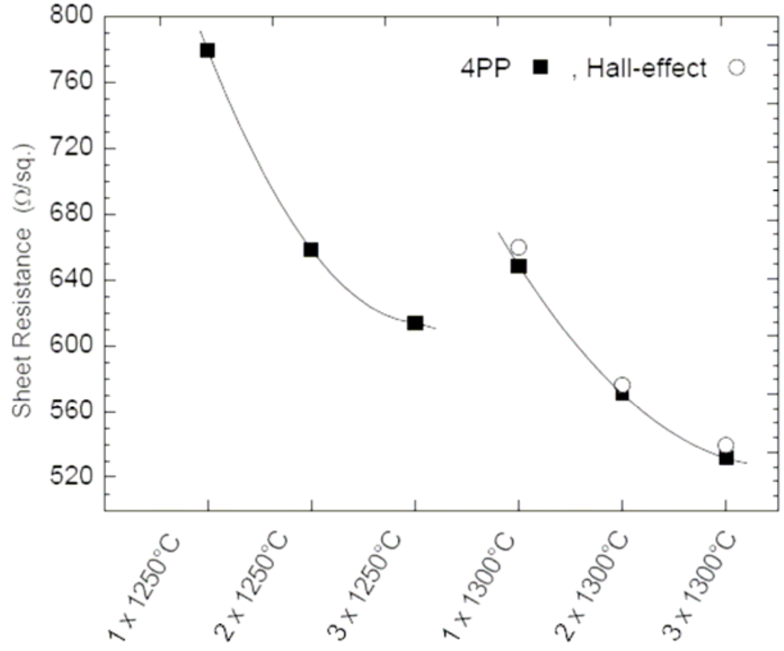



# Summary of Multiple Flash Processes

Boron



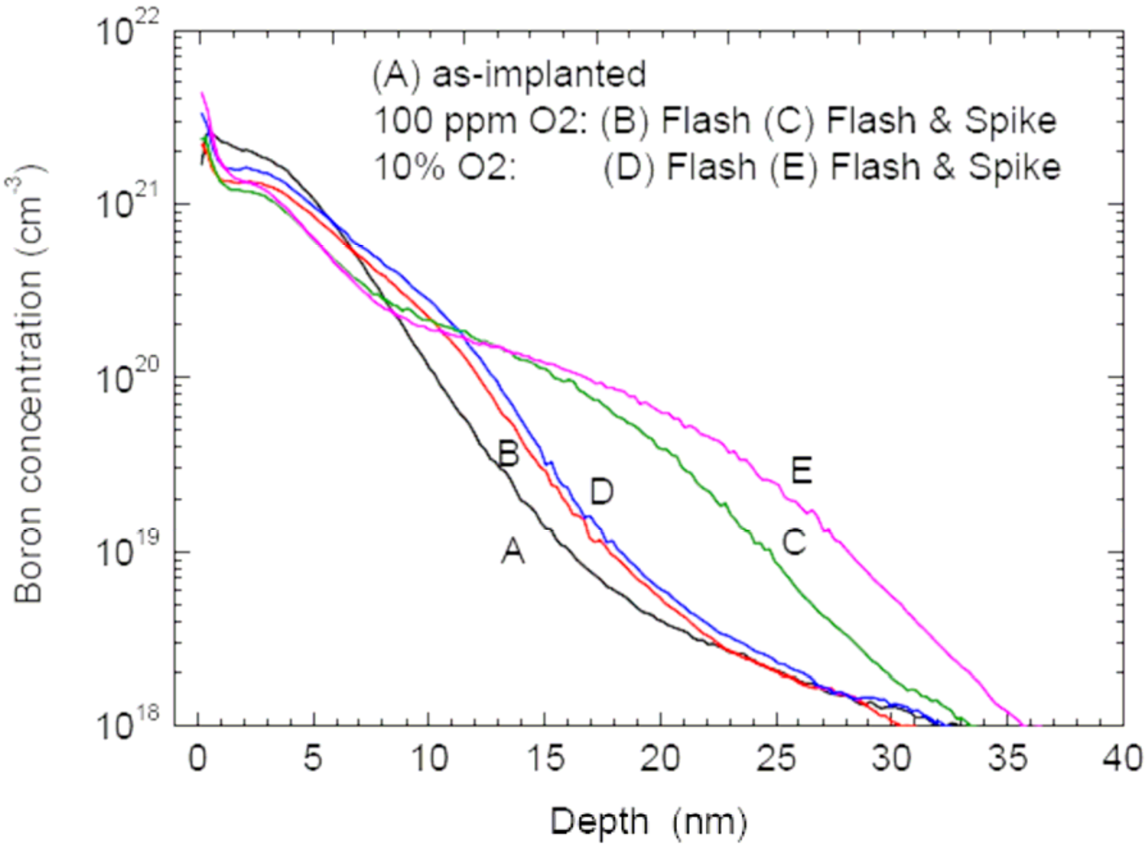

Arsenic




Multiple flash processing increases activation significantly (up to 25%) with little additional diffusion for both boron and arsenic.

# Oxidation Enhanced Diffusion during Flash Annealing: PMOS

$B^+$ 500eV, 1E15cm$^{-2}$ into crystalline Silicon




Spike: 950°C

Flash: 750°C/1300°C

Sheet Resistance Values

B=448Ω/sq.

C=395Ω/sq.

D=453Ω/sq.

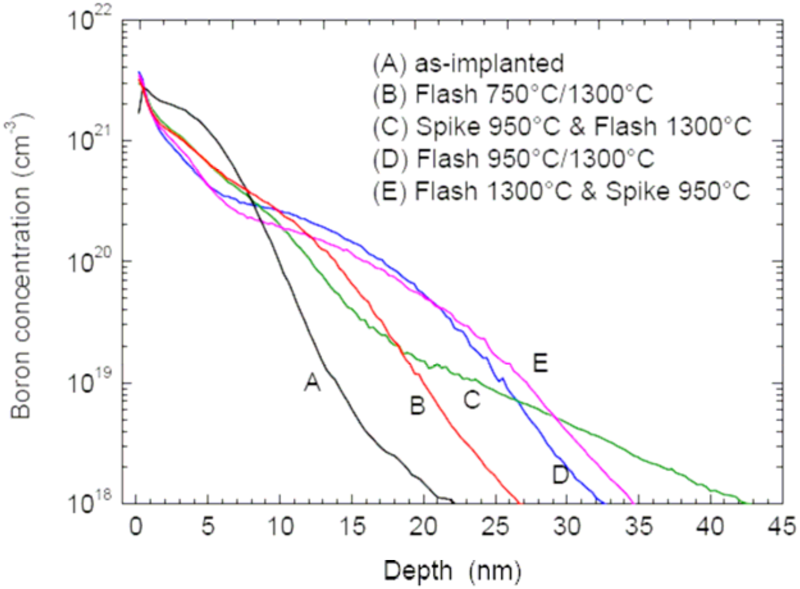
E=463Ω/sq.

For the 100ppm $O_2$ case (C), the boron retained dose is reduced by about 20% compared to 10% $O_2$ case (E) with Flash & Spike.

# Annealing Strategy for Dopant Activation: PMOS

B+ 500ev, 1E15cm-2

α-Si

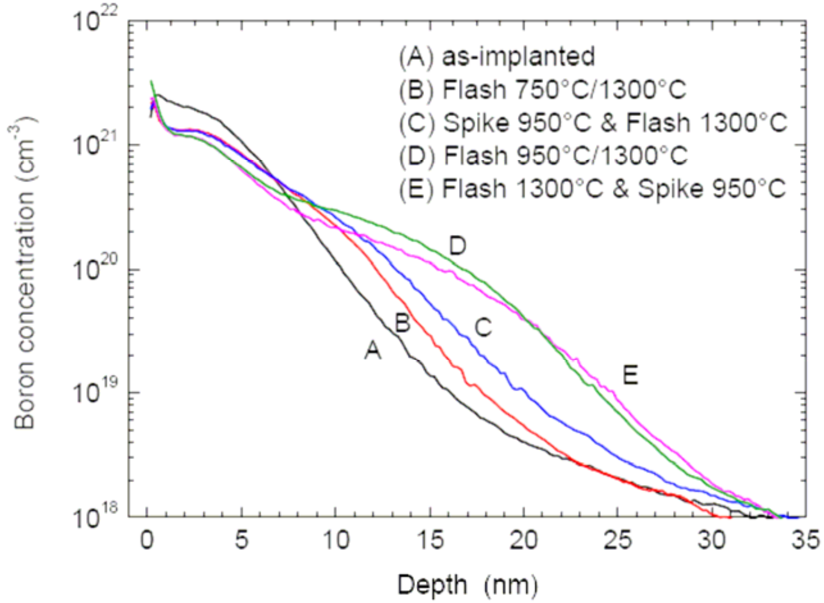



B: 373Ω/sq., C: 416Ω/sq.

D: 360Ω/sq., E: 443Ω/sq.

Single or Multiple Flash seems to be the optimum strategy

c-Si

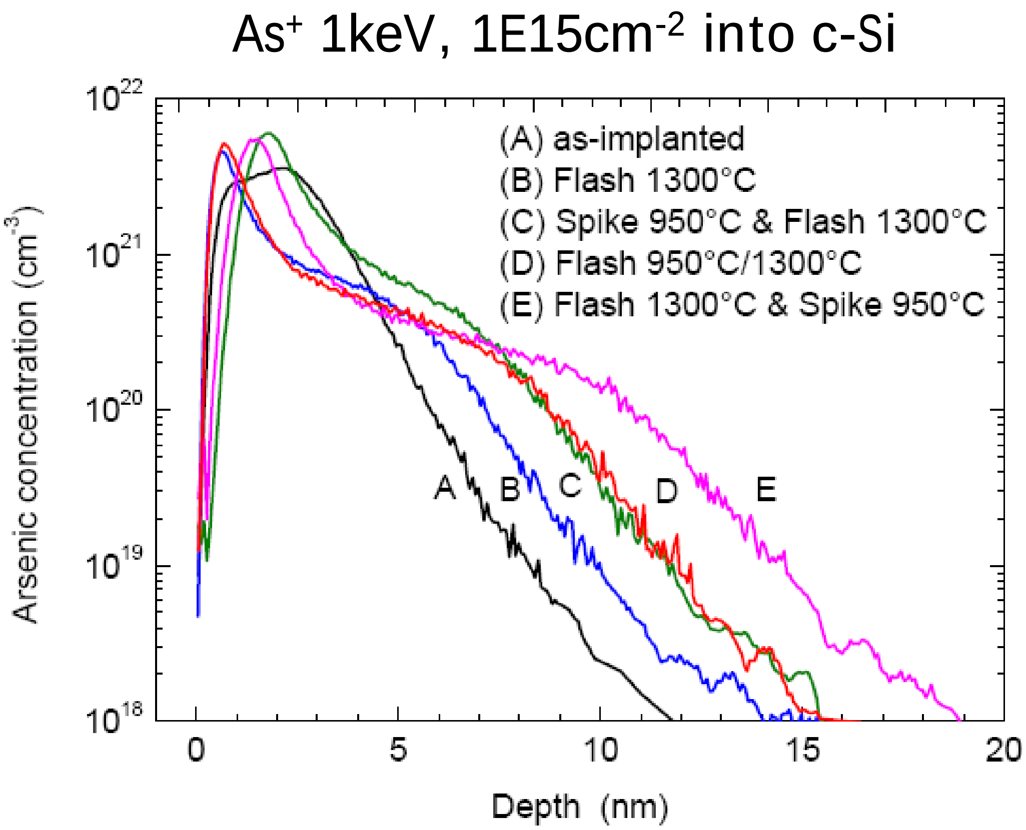



B: 448Ω/sq., C: 375Ω/sq.

D: 436Ω/sq., E: 453Ω/sq.

Spike+Flash seems to be the optimum strategy

# Annealing Strategy for Dopant Activation: NMOS



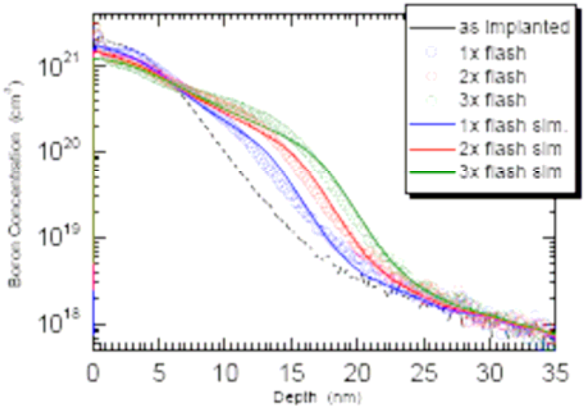

B: 648Ω/sq., C: 591Ω/sq.
D: 694Ω/sq., E: 693Ω/sq.

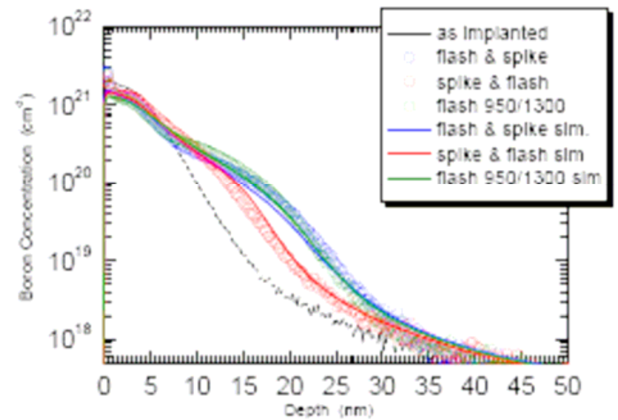
For As implants into c-Si, the defect density for spike+flash anneals is below the TEM detection limit.

Spike+Flash seems to be the optimum anneal strategy

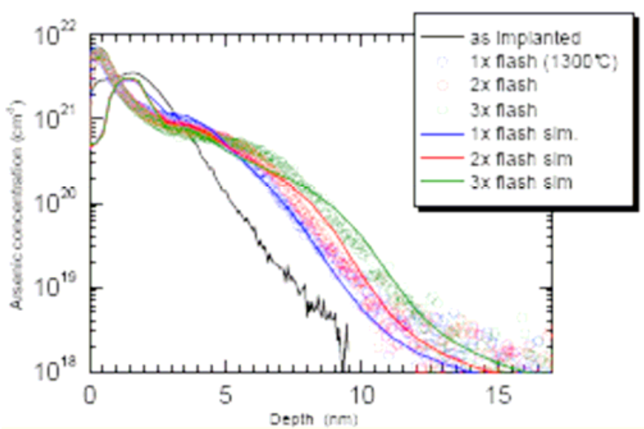
# Simulations Agree with Experiments

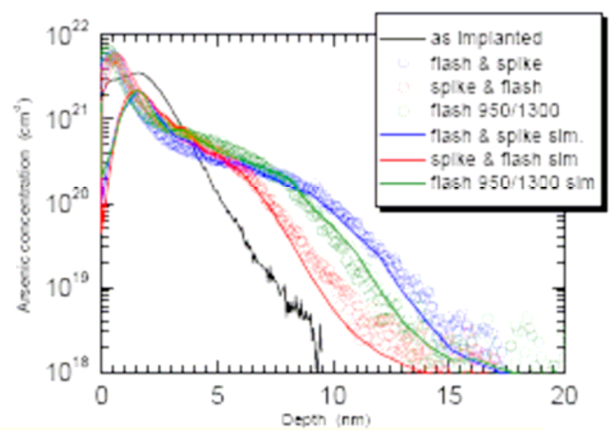
c-Si, 500 eV $B^+$ $1\cdot10^{15}$ $cm^{-2}$




c-Si, 1 keV $As^+$ $1\cdot10^{15}$ $cm^{-2}$




Diffusion and activation due to various thermal budgets and their combinations are well described by TCAD simulations!

# Summary and Conclusions

- Taken individually, each implant has an optimum anneal condition
  - Boron in c-Si: Spike + Flash
  - Boron in $\alpha$-Si: Multiple Flash
  - Arsenic in c-Si: Spike + Flash
- Oxygen control is important; even for very short MSA
- Dopant loss can be effectively controlled with ambient engineering
- Multiple flash processes can increase activation by up to 25% and improve defect removal, but some extended defects still remain
- For both B and As implants into crystalline silicon, the extended defects are below the TEM (WBDF) detection limit using spike + flash annealing
- Simulation can do an effective job of modeling activation and diffusion
- For integration of a CMOS process flow, this is essential to optimize the entire process

# Acknowledgements


- Partial support for this work was funded by the European Union under the ATOMICS project
- Much of this work was initially presented at the EMRS earlier this year
  - Thanks to the co-authors of that paper: W. Lerch, S. Paul, J. Niess: Mattson Thermal Products GmbH; S. McCoy, J. Gelpey: Mattson Technology Canada Inc.; F. Cristiano, F. Severac: LAAS/CNRS; P. Fazzini: CEMES/CNRS; A. Martinez-Lima, P. Pichler: Fraunhofer-IISB; H. Kheyrandish: CSMA–MATS and D. Bolze: IHP
- Thanks to Zsolt Nenyei and Paul Timans for helpful discussions