Variation in 45nm and Implications for 32nm and Beyond

Kelin J. Kuhn
Intel Fellow
Director of Advanced Device Technology
AGENDA

Technology scaling
I. Physical Variation Sources and Mitigation
II. Measurements, results and interpretation
II. Next generation challenges
Closing thoughts
Technology Scaling
Lithography Scaling Limitations
From Broers [1] IEDM Plenary Session 1980

In the limit, microscope objectives with 0.95 N.A. are available and, provided very small fields (200μ x 200μ) are adequate, linewidths < 0.4μ should be achievable under carefully controlled laboratory conditions, and in very thin resist layers.

Depth of field will be reduced to about ± 0.2μ. Deep U.V. (λ = 200nm - 260nm) lenses will be difficult to build because of the lack of materials that are transparent at these wavelengths and yet have relatively high refractive indices.

1980:
Optical Lithography Limit
~ 400nm
Transistor Scaling Limitations

For conservative design margins, typical results suggest that IGFET channel lengths can be reduced to approximately 0.40 microns in E/D NMOS logic gates; 0.30 microns in CMOS transmission gates; and 0.20 microns in E/E CMOS logic gates. Smaller channel lengths can be projected for more aggressive designs. The dominant mechanism imposing these limits is subthreshold drain current due to short channel charge sharing and drain induced barrier lowering.

1983: Transistor architecture limit
200-400nm (SCE)
Transistor Scaling Limitations
From Heilmeier [4] IEDM Plenary Session 1984

Other factors limiting the scaling of ICs come into play. Some of these factors are interconnect capacitance, channel capacitance, interconnect resistance, parasitic resistances, velocity saturation, ionizing radiation, drain breakdown, gate oxide breakdown, hot carrier injection, subthreshold current, punch through, and statistical control of oxide thickness and channel doping. It appears that minimum geometries for high-volume ICs will saturate in the range of 0.3 to 0.5 microns.

1984:
Transistor architecture limit
300-500nm (laundry list of reasons...)
How small is a 32nm memory cell?

Small enough that a 2008 32nm SRAM cell is dwarfed by a human red blood cell.

1983-84 limits on gate size, are commensurate with the dimensions of 2008’s entire 32nm SRAM cell!
How small is a 32nm memory cell?

1980 SRAM Cell: 1700 um²

32nm SRAM Cell: 0.171 um²

Small enough that a 2008 32nm SRAM cell is dwarfed by a 1980 SRAM cell CONTACT

M. Bohr 2007
How small is a 32nm memory cell?

Small enough that a 2008 32nm SRAM cell is dwarfed by a 1980 SRAM cell CONTACT

M. Bohr, ISCC, 2009
Atomic dimensions are now routine
Part I:
Physical Variation Sources
and Mitigation
Part I – Physical Variation Sources and Mitigation

Patterning

Polish

Strain
Part I – Physical Variation Sources and Mitigation

Patterning

Polish

Strain
How small is a 45nm transistor?

- 5.5X smaller than the 193nm light that prints it
- ~15X smaller than visible green light
Putting it all together for the gate layer of a 65nm MPU

(magnified 25,000X)
Putting it all together for the gate layer of a 65nm MPU

(magnified 25,000X)
Optical Proximity Correction (OPC)  
As a Resolution Enhancement Technique

Contour prediction – no OPC  Contour prediction – with OPC

SEM Image – no OPC  SEM Image – with OPC

K. Wells-Kilpatrick: 2007
45nm: OPC as a Variation Management Technique

Top-down resist CD meets spec, but poor contrast leads to poor resist profile which gets transferred to metal pattern after etch, resulting in shorting marginality

Computational lithography solution

K. Kuhn, IEDM 2007
Design

OPC/RET

Putting it all together for the gate layer of a 65nm MPU

(magnified 25,000X)

Etch

Reticle manufacturing

Phase mask

Trim mask

Phase mask data

Trim mask data

C. Kenyon
TOK conf.
Dec. 2008
MEEF
Mask Error Enhancement Factor

• MEEF is a scaling factor that causes certain layout geometries to exhibit a greater sensitivity to mask dimension tolerances.

• Any dimensional error in the mask is magnified on the wafer by the MEEF value.

\[ \Delta W_{\text{wafer}} = \text{MEEF} \times \Delta W_{\text{mask}} \]

• Depending on the value of the mask error and the lithography exposure/focus conditions the final printed pattern can be either larger or smaller.
MEEF Impact on Ze Error

Ze error can be either positive or negative.

Yellow: DCCD contour after OPC
Green: with -3.375 nm mask making error
Red: with 3.375 nm mask making error

65nm Simulation
Notch width = 120nm
Notch height = 250nm

MEEF = 8.4
Low MEEF requires targeting in the “flat” portion of CD vs. pitch

Process innovations continue this trend in the 32nm node
Putting it all together for the gate layer of a 65nm MPU

(magnified 25,000X)
FLARE

- Flare is unwanted scattered light arriving at the wafer.
- Flare is caused by interactions that force the light to travel in a "non-ray trace" direction.
- Flare is both a function of local environment around a feature (short range flare) and the total amount of energy going through the lens (long range flare).
Impact of flare on gate CDs

- During 65nm process development, large CD deviations were observed for structures having identical pitch and reticle CD due to flare
- Gates only 500μm away from one another could be >5nm different in CD

C. Kenyon
TOK conf.
Dec. 2008
Development effort produced an algorithm capable of scanning designs and binning regions by local chrome fraction. Binning algorithm is combined with flare-calibrated OPC model.
Putting it all together for the gate layer of a 65nm MPU (magnified 25,000X)
45nm highlights role of lithography/etch in resolving LER/LWR
Critical to management of variation is the ability to deliver a 0.7X gate CD variation improvement in each generation enabled by continuous process technology improvements.
Part I – Physical Variation Sources and Mitigation

Patterning

Polish

Strain
CMP Integration at 45 nm – HiK Metal Gate

First Generation HiK – Replacement Metal Gate
Three critical CMP operations in the FE

STI deposition and polish
Wells and VT implants
ALD deposition of high-k gate dielectric
Polysilicon deposition and gate patterning
S/D extensions, spacer, Si recess and SiGe deposition
S/D formation, Ni silicidation, 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

K.Mistry et al., IEDM (2007)
C.Auth et al. VLSI Symp, (2008)
J. Steigerwald, IEDM (2008)
CMP Integration at 45 nm – HiK Metal Gate

First Generation HiK – Replacement Metal Gate

Three critical CMP operations in the FE

- STI deposition and polish
- Wells and VT implants
- ALD deposition of high-k gate dielectric
- Polysilicon deposition and gate patterning
- S/D extensions, spacer, Si recess and SiGe deposition
- S/D formation, Ni silicidation, 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

References:
- K.Mistry et al., IEDM (2007)
- C.Auth et al. VLSI Symp, (2008)
- J. Steigerwald, IEDM (2008)
STR Pattern Density Variation Impact

High Pattern Density

Low Pattern Density

Post STI Deposition

Nitride

Silicon

Oxide

Slower Polish Rate

Faster Polish Rate

Post STI Polish
STI Step Height Variation

High Pattern Density Area

Low Pattern Density Area

STI topography impacts transistor $L_e$ and $Z_e$

Positive Step Height

Zero Step Height
STI Step Height Variation

High Pattern Density Area

Low Pattern Density Area

STI topography impacts transistor Le and Ze

Poly
STI Step Height Impact on Gate CD

Negative Step Height

- "Dogbone"
  - Lg is longer at the diffusion boundary

Positive Step Height

- "Icicle"
  - Gate CD is shorter at the diffusion boundary
SRAM Density Scaling

65nm to 32nm: Patterning and polish enhancements

- Improved CD uniformity across STI boundaries
- Square corners (eliminate “dogbone” and “icicle” corners)

90nm – TALL
1.0 \( \mu m^2 \)

65nm – WIDE - 0.57 \( \mu m^2 \)

45nm – WIDE
0.346 \( \mu m^2 \)

32nm – WIDE
0.171 \( \mu m^2 \)
CMP Integration at 45 nm – HiK Metal Gate

First Generation HiK – Replacement Metal Gate
Three critical CMP operations in the FE

STI deposition and polish
STI CMP

Wells and VT implants

ALD deposition of high-k gate dielectric
Polysilicon deposition and gate patterning
S/D extensions, spacer, Si recess and SiGe deposition
S/D formation, Ni silicidation, ILD0 deposition

Poly Opening Polish, Poly removal

PMOS workfunction metal deposition

S/D formation, Ni silicidation, ILD0 deposition
Poly Opening Polish, Poly removal

Metal gate patterning, NMOS WF metal deposition

Metal gate fill and polish, ESL deposition

K.Mistry et al., IEDM (2007)
C.Auth et al. VLSI Symp, (2008)
J. Steigerwald, IEDM (2008)
Variation Challenges of RMG CMP Steps

- Gate height control critical to reducing variation
- PMOS/NMOS differences complicate CMP

C. Auth et al. VLSI Symp, (2008)

J. Steigerwald, IEDM 2008
Variation Challenges of RMG CMP Steps

OVERPOLISH
Exposes raised S/D
Rext/mobility impact

UNDERPOLISH
Underetched contact
Rext impact

S/D region – attacked
during poly etch

Gate region

S/D region – marginal
contact

NMOS S/D region
contact

J. Steigerwald, IEDM 2008
Poly Opening Polish (POP) Thickness Control

45nm: within die (WID) and within wafer (WIW) improvement
High selectivity between films is required.
Key aspect is control of polish rate at edge of wafer.

J. Steigerwald, IEDM 2008
45 nm: POP CMP Improvement
Overscaling Topography Improvement

Improvements in polish enabled dramatic improvements in topography variation

J. Steigerwald, IEDM 2008
Part I – Physical Variation Sources and Mitigation

Patterning

Polish

Strain
Strain: Importance in scaling

Strain (first introduced at 90nm) is a critical ingredient in modern transistor scaling
Strain: Pitch dependence

NMOS
Pitch degradation increases with film pinchoff, requires higher stress, thinner films

PMOS
eSiGe S/D mobility strongly dependent on pitch

C. Auth, VLSI 2008
NMOS strain: Scaling with pitch

Tensile trench contacts

Compressive gate stress

C. Auth, VLSI 2008
PMOS strain: Scaling with pitch

- Increase %Ge
- Move SiGe
- Remove Gate

IDSAT (a.u.)

65nm 45nm

Removal of Gate

Increase to 30% Ge

Proximity Reduction

Technology node

C. Auth, VLSI 2008
Random $V_T$ variability and strain

Similar $V_T$ matching with CESL while 35% $I_{ON}$ enhancement is achieved

$650\mu A/\mu m$ – $30pA/\mu m$ at $V_{dd}=1V$ and $L_g=25nm$
Part II: Measurements, results and interpretation
Systematic and Random

• Statistician’s viewpoint:

• Process engineer’s viewpoint:

• Device engineer’s viewpoint:
Measurement “food pyramid”

- In-line or off-line physical measurements of test wafers (TEM, SIMs, Auger, etc.)
- Device parametric measurements on test material ($I_{on}/I_{off}$, $I_{G}/V_{G}$ etc.)
- In-line physical measurements of selected sites in product (CD, thickness, etc.)
- Device parametric measurements on product ($I_{dsat}/I_{lin}$, $V_{T}$)
- Device parametric measurements on simple circuits ($f_{max}$, $f_{min}$, etc)
- Device sort on completed product ($V_{ccmin}$ and performance)

Increasing data quantity
Decreasing ability to segment origin

Very limited data
Huge sample size

Highly detailed data
Tiny sample size
Measurement of Random and Systematic VT Variation at the Device Level

Traditional method:
1. Measure two identical adjacent devices and extract the difference $\sigma(VT_A - VT_B)$
2. Measure the entire population of all devices and extract $\sigma(VT_{pop})$

Random Variation for a matched pair

$$Random_{mp} = \text{StdDev}(VT_A - VT_B) = \sigma(DVT)$$

Random Variation for a single device

$$Random_{one-device} = \frac{\text{StdDev}(VT_A - VT_B)}{\sqrt{2}} = \frac{\sigma(DVT)}{\sqrt{2}}$$

Systematic Variation for a single device

$$Systematic = \sqrt{(\sigma VT_{pop})^2 - \left(\frac{\sigma(DVT)}{\sqrt{2}}\right)^2}$$
Pelgrom Plots: What is $A_{VT}$ anyway?
Two choices are widely used in the literature

IEDM 2008: Weber

**Choice A**
Slope of $\sigma VT$ vs $1/\sqrt{LW}$

**Choice B**
Slope of $\sigma \Delta VT$ vs $1/\sqrt{LW}$

$\sqrt{2}$
What did Pelgrom say?

Pelgrom “Matching properties of MOS transistors”

- Eq. 5 defines a generic $A_P$ for a parameter $\Delta P$; implying $AV_T$ would then be the parameter for $\Delta VT$

\[
\sigma^2(\Delta P) = \frac{A_P^2}{WL} + S_{x}^2 D_x^2.
\]

- However, one page further in the paper, he explicitly defines $AV_T$ in terms of $VT$ only in equation 8:

\[
\sigma^2(V_{T0}) = \frac{A_{VT0}^2}{WL} + S_{VT0}^2 D^2
\]

- So – which is did he mean? Well, I asked him.
What is $\Delta V_T$ anyway?

Two choices are widely used in the literature:

Choice A

Slope of $\sigma_{VT}$ vs $1/\sqrt{LW}$

Choice B

Slope of $\sigma_{\Delta VT}$ vs $1/\sqrt{LW}$

This is $\Delta V_T$
What is $\Delta V_T$ anyway?

Two choices are widely used in the literature:

**Choice A**

Slope of $\sigma V_T$ vs $1/\sqrt{LW}$

**Choice B**

Slope of $\sigma \Delta V_T$ vs $1/\sqrt{LW}$

I will call this $C_{VT}$ (or $C_2$)

$C_{VT} = \Delta V_T / \sqrt{2}$

This is $\Delta V_T$
RDF is frequently described by (Stolk):

\[
\sigma V_{\text{Tran}} = \left( \frac{\sqrt{4q^3 \varepsilon_{\text{Si}} \phi_B}}{2} \right) \cdot \frac{T_{\text{ox}}}{\varepsilon_{\text{ox}}} \cdot \left( \frac{\sqrt{N}}{\sqrt{L_{\text{eff}} \cdot Z_{\text{eff}}}} \right) = \frac{1}{\sqrt{2}} \left( \frac{c_2}{\sqrt{L_{\text{eff}} \cdot Z_{\text{eff}}}} \right)
\] (1)

RDF is frequently described by (Stolk):

$$\sigma V_{\text{Tran}} = \left( \frac{4q^3 \varepsilon_{\text{Si}} \phi_B}{2} \right) \frac{T_{\text{ox}}}{\varepsilon_{\text{ox}}} \left( \frac{4\sqrt{N}}{\sqrt{L_{\text{eff}} \cdot Z_{\text{eff}}}} \right) = \frac{1}{\sqrt{2}} \left( \frac{c_2}{\sqrt{L_{\text{eff}} \cdot Z_{\text{eff}}}} \right)$$  \hspace{1cm} (1)


K. Kuhn, IEDM 2007

Additional propagation of confusion (By me, it turns out …)

C2 is proportional to the slope of the $1 / \sqrt{L_{\text{eff}} \cdot Z_{\text{eff}}}$ line vs $\sigma VT$
What is $B_{VT}$ then?

$$B_{VT} = \text{Slope of } \sigma_{VT} \text{ vs } \sqrt{T_{inv}(V_T+0.1)/LW}$$

Fig. 3: Takeuchi, IEDM 2007
But what about simple circuits?

One powerful tool for assessment of variation is locating ring-oscillators (ROs) routinely in all product designs.
# Random and Systematic Variation for Matched Ring Oscillators

**Random:**
- **Calculate Delta**
  \[
  \Delta = \frac{FreqA - FreqB}{FreqA + FreqB} \times 200
  \]
  \[
  \text{Rand} = \text{StdDev}(\Delta)
  \]

**Systematic:**
- **Total Sigma**
  \[
  \sigma = \text{StdDev}(FreqA)
  \]
- **Grand Mean**
  \[
  \mu = \frac{\text{Mean}(FreqA) + \text{Mean}(FreqB)}{2}
  \]
- **Systematic Variation**
  \[
  \text{Syst} = \sqrt{\left(\frac{\sigma}{\mu}\right)^2 - \text{Rand}^2}
  \]

**Total Variation:**
\[
\text{Total} = \frac{\text{StdDev}(FreqA)}{\text{Mean}(FreqA)} \times 100
\]
45nm: Within Wafer Variation

For random variation: Uniform across wafer
For systematic variation: More variation at the wafer edge

15cm 14cm 13cm 12cm 11cm 10cm 9cm 8cm
45nm: Within Die (WID), Within Wafer (WIW) and Wafer to Wafer (WTW)

For random variation: Uniform with population choice
For systematic variation: Variation increases significantly going from within-die (WID) to within-wafer (WIW)
45nm Product wafer: Random variation

K. Kuhn, ITJ 2008
Random and Systematic Variation Trends

Systematic WIW variation is comparable from one generation to the next.

Random WIW variation in 32nm is comparable to 45nm and significantly improved over 65nm and 90nm due to HiK-MG.
What about more complex circuits?
RSM Methodology for Variation Model Parameters

- Identify the set of input parameters in variation modeling files that can be allowed to vary
- Create DOE to vary all parameters within selected limits
- Create a series of variation modeling files, using the matrix of parameters from the DOE
- Simulate an appropriate set of circuits and devices to obtain responses to the set of variation modeling files
- Enter simulation results back into DOE to determine sensitivity to model parameters
- Optimize variation modeling file parameters to get best match to measured data
**Example Matrix of Inputs and Associated Responses**

<table>
<thead>
<tr>
<th>Inputs</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
</tr>
<tr>
<td>D</td>
<td>4</td>
</tr>
<tr>
<td>E</td>
<td>2</td>
</tr>
<tr>
<td>F</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td></td>
</tr>
<tr>
<td>J</td>
<td></td>
</tr>
</tbody>
</table>

Not all responses are sensitive to all inputs—key is to determine which responses are appropriate for setting each input parameter.
SRAM test chip with advanced test features (PBIST, eFUSE, ECC, etc.) to support development of 32nm high-volume manufacturing process

K. Zhang, ISCC 2009
SRAM $V_{CCmin}$ – Silicon to Simulation

Transistor random $V_T$ variation ($\sigma_{VT,\text{random}}$)

Read: Static-Noise-Margin (SNM) &
Dynamic Stability
Write: Dynamic Writability

Die-level
Read / Write $V_{CCmin}$

K. Zhang, ISCC 2009

Wafer-level SRAM P/NMOS transistor systematic $V_T$ variation
32nm Voltage-Frequency Shmoo

3.25Mb SRAM Macro

- 32nm SRAM operates over a broad range of supply voltages, enabling dynamic voltage scaling for low-power application
- 32nm SRAM achieves operating frequency of 4GHz at 1.0V, 15% better than 45nm design

K. Zhang, ISCC 2009
Part III: Next generation challenges
Lithography Pipeline

Extend 193nm Optical Lithography as far as possible
Deploy EUV Lithography when available/affordable
Extreme Ultraviolet Lithography

Cymer beta source

Intel EUV Mask

ASML ADT printed wafer

Philips beta source

Photoresist Development

Nikon EUV1 printed wafer

Continued progress towards EUV implementation

M. Bohr, ISCC, 2009
Non-EUV Lithography Beyond 32 nm

Double Patterning
- Pitch doubling
- Improved 2-D features

Spacer Gate Patterning
- Pitch doubling
- Improved variation
Pitch doubling and gate CD control

Neither Resist Freeze nor Double Pattern Transfer achieve full benefit of patterning at \( \frac{1}{2} \) pitch

Both techniques still require resolution of a very small space (MEEF, LWR etc.)

C. Kenyon, TOK conf., Dec. 2008
Disadvantages of Double-patterning

Misalignment between the 2 exposures is a crucial liability for this technique and can limit its usability. Transistor parameters can be affected by asymmetry between the source and drain regions.
Pitch doubling eliminates the close correlation which currently exists between the CDs of adjacent gates. This has implications for memory cells and other circuits which depend upon this CD matching.

C. Kenyon, TOK conf., Dec. 2008
Pitch doubling and gate CD matching

Single patterning: the distribution of CD mismatches between adjacent gates is a very small fraction of total gate CD variation

Pitch doubling: the distribution of CD mismatches is GREATER than the total gate CD variation

C. Kenyon, TOK conf., Dec. 2008
Non-EUV Lithography Beyond 32 nm

Double Patterning
- Pitch doubling
- Improved 2-D features

Spacer Gate Patterning
- Pitch doubling
- Improved variation

M. Bohr, ISCC, 2009
Bencher et al, Proc. of SPIE Vol. 6924 69244E-7
Alternative: Spacer patterning

Spacer patterning retains correlation between doubled features
Alternative: Spacer patterning

Spacer inhomogenities not transferred to patterned features

Bencher et al, Patterning by CVD Spacer Self Alignment DoublePatterning (SADP), Proc. of SPIE Vol. 6924 69244E-7
Uniformity matters: Logic images vs. technology node
Layout Restrictions 65nm to 32nm

65 nm Layout Style

- Bi-directional features
- Varied gate dimensions
- Varied pitches

32 nm Layout Style

- Uni-directional features
- Uniform gate dimension
- Gridded layout

M. Bohr, ISCC, 2009
Transistor Architecture Enhancements

Fully depleted devices (such as UTB or FinFET) are examples of innovations which permit significant improvement in RDF due to the ability to maintain channel control at lower channel doping.

Weber et al. IEDM 2008 pp. 245-248

Vellianitis et al. IEDM 2008 pp. 681-683
Fully depleted devices (such as UTB or FinFET) are examples of innovations which permit significant improvement in RDF due to the ability to maintain channel control at lower channel doping.

Weber et al. IEDM 2008 pp. 245-248
Closing Thoughts
Random and Systematic Variation Trends

Systematic WIW variation is comparable from one generation to the next.

Random WIW variation in 32nm is comparable to 45nm and significantly improved over 65nm and 90nm due to HiK-MG.
45 nm: POP CMP Improvement
Overscaling Topography Improvement

Improvements in polish enabled dramatic improvements in topography variation

J. Steigerwald, IEDM 2008
Gate CD variation improvements with technology scaling

Critical to management of variation is the ability to deliver a 0.7X gate CD variation improvement in each generation enabled by continuous process technology improvements.
SRAM Density Scaling

Improved fidelity / uniformity on 32nm vs 90nm

K. Zhang, ISCC, 2009
Q&A

For further information on Intel's silicon technology, please visit our Technology & Research page at

www.intel.com/technology