The New Era of Scaling in an SoC World

Mark Bohr
Intel Senior Fellow
Logic Technology Development
The End of Scaling is Near?

“Optical lithography will reach its limits in the range of 0.75-0.50 microns”

“Minimum geometries will saturate in the range of 0.3 to 0.5 microns”

“X-ray lithography will be needed below 1 micron”

“Minimum gate oxide thickness is limited to ~2 nm”

“Copper interconnects will never work”

“Scaling will end in ~10 years”

Perceived barriers are meant to be surmounted, circumvented or tunneled through
Outline

• Transistor Scaling
• Microprocessor Evolution
• Vision of the Future
Scaling Trends

Transistor dimensions scale to improve performance, reduce power and reduce cost per transistor.
Scaling Trends

Transistor dimensions scale to improve performance, reduce power and reduce cost per transistor.
## MOSFET Scaling

<table>
<thead>
<tr>
<th>Device or Circuit Parameter</th>
<th>Scaling Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Device dimension ( tox, L, W )</td>
<td>( 1/\kappa )</td>
</tr>
<tr>
<td>Doping concentration ( Na )</td>
<td>( \kappa )</td>
</tr>
<tr>
<td>Voltage ( V )</td>
<td>( 1/\kappa )</td>
</tr>
<tr>
<td>Current ( I )</td>
<td>( 1/\kappa )</td>
</tr>
<tr>
<td>Capacitance ( \varepsilon A/t )</td>
<td>( 1/\kappa )</td>
</tr>
<tr>
<td>Delay time/circuit ( VC/I )</td>
<td>( 1/\kappa )</td>
</tr>
<tr>
<td>Power dissipation/circuit ( VI )</td>
<td>( 1/\kappa^2 )</td>
</tr>
<tr>
<td>Power density ( VI/A )</td>
<td>1</td>
</tr>
</tbody>
</table>

Classical MOSFET scaling was first described in 1974

---

R. Dennard, IEEE JSSC, 1974
30 Years of MOSFET Scaling

Dennard 1974

Gate Length: 1.0 µm
Gate Oxide Thickness: 35 nm
Operating Voltage: 4.0 V

Intel 2005

Gate Length: 35 nm
Gate Oxide Thickness: 1.2 nm
Operating Voltage: 1.2 V

Classical scaling ended in the early 2000s due to gate oxide leakage limits
90 nm Strained Silicon Transistors

Strained silicon provided increased drive currents, making up for lack of gate oxide scaling
High-k + Metal Gate Transistors

65 nm Transistor

45 nm HK+MG

SiO₂ dielectric
Polysilicon gate electrode

Hafnium-based dielectric
Metal gate electrode

High-k + metal gate transistors
break through gate oxide scaling barrier
Transistor Performance Increase

45 nm HK+MG provides average 30% drive current increase or >5x $I_{OFF}$ leakage reduction

Ref. K. Mistry, IEDM ’07
Gate Leakage Reduction

HK+MG significantly reduces gate leakage
SRAM bitcell leakage reduced ~10x
Minimal oxide scale  HiK+MG

Tox scaling

$C_2$

Normalized to 180nm

$\sigma_{V_{\text{Tran}}} = \left( \frac{4q^3\varepsilon_{si} \phi_B}{2} \right) \left( \frac{T_{ox}}{\varepsilon_{ox}} \right) \left( \frac{4\sqrt{N}}{\sqrt{Leff \cdot Zeff}} \right) = \frac{1}{\sqrt{2}} \left( \frac{C_2}{\sqrt{Leff \cdot Zeff}} \right)$

HK+MG provides oxide scaling needed for variability reduction

Ref. K. Kuhn, IEDM '07
Interconnect Trends

Added metal layers + material improvements enable interconnect scaling

Technology Generation (nm)

M2 Pitch (µm)

# Metal Layers

0.7x per generation
Interconnect Trends

Added metal layers + material improvements enable interconnect scaling
Interconnect Trends

Added metal layers + material improvements enable interconnect scaling
45 nm Interconnects

Loose pitch + thick metal on upper layers
- High speed global wires
- Low resistance power grid

Tight pitch on lower layers
- Maximum density for local interconnects

Hierarchical interconnect pitches
45 nm Interconnects

Thick M9 for very low resistance on-die power routing
45 nm Microprocessor Products

45 nm process serves microprocessor applications from low power to high performance
32 nm Generation
32 nm Logic Technology

• 2nd generation high-k + metal gate transistors
  - High-k EOT scaled from 1.0 nm to 0.9 nm
  - Replacement metal gate process flow
  - 4th generation strained silicon

• 9 copper + low-k interconnect layers
  - Hierarchical interconnect pitches
  - Thick M9 for power routing

• Immersion lithography on critical layers
  - 70% transistor and interconnect pitch scaling
  - 50% SRAM cell area scaling

• Pb-free and halogen-free packages

Higher performance, lower power, lower cost per transistor
Contacted Gate Pitch Trend

Transistor gate pitch continues to scale 0.7x every 2 years
Transistor Performance

Drive currents continue to increase while gate pitch scales.
32 nm Interconnects

Hierarchical interconnect pitches

M9
Pitch (nm)
M8 566.5
M7 450.1
M6 337.6
M5 225.0
M4 168.8
M3 112.5
M2 112.5
M1 112.5
SRAM Cell Size Scaling

Transistor density continues to double every 2 years

32 nm Generation 0.171 um² Cell
SRAM Cell Scaling

<table>
<thead>
<tr>
<th>Feature Size</th>
<th>Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>65 nm</td>
<td>0.570 µm²</td>
</tr>
<tr>
<td>45 nm</td>
<td>0.346 µm²</td>
</tr>
<tr>
<td>32 nm</td>
<td>0.171 µm²</td>
</tr>
</tbody>
</table>

Good pattern resolution while scaling feature size and continuing with 193 nm exposure wavelength
32 nm SRAM Test Chip

- 291 Mbit
- 0.171 \( \mu m^2 \) cell size
- >1.9 billion transistors
- >3.8 GHz operation
- Functional silicon in Aug ‘07

32 nm SRAM test vehicle included all transistor and interconnect features used on 32 nm microprocessors

Ref. Y. Wang, paper 27.1, ISSCC ’09
30 Years of Scaling

Contact 1978

Ten 32nm SRAM Cells 2008

Scale 1 µm
## The Old Era of Device Scaling

It has served us well for >30 years

<table>
<thead>
<tr>
<th>Device or Circuit Parameter</th>
<th>Scaling Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Device dimension $t_{ox}, L, W$</td>
<td>$1/k$</td>
</tr>
<tr>
<td>Doping concentration $N_a$</td>
<td>$k$</td>
</tr>
<tr>
<td>Voltage $V$</td>
<td>$1/k$</td>
</tr>
<tr>
<td>Current $I$</td>
<td>$1/k$</td>
</tr>
<tr>
<td>Capacitance $\varepsilon A/t$</td>
<td>$1/k$</td>
</tr>
<tr>
<td>Delay time/circuit $VC/I$</td>
<td>$1/k$</td>
</tr>
<tr>
<td>Power dissipation/circuit $VI$</td>
<td>$1/k^2$</td>
</tr>
<tr>
<td>Power density $VI/A$</td>
<td>$1$</td>
</tr>
</tbody>
</table>
The New Era of Device Scaling

Modern CMOS scaling is as much about material and structure innovation as dimensional scaling.
Outline

• Transistor Scaling

• Microprocessor Evolution

• Vision of the Future
Microprocessor Evolution

More transistors

Higher frequency

More data bits per cycle

Instruction parallelism

Out-of-order issue

Multi-threading

Many of these innovations have been for improved performance, now the challenge is to innovate for power efficiency
45 nm Nehalem CPU

Modern microprocessors are a complex system on a chip with multiple functional units and multiple interfaces
45 nm Nehalem CPU

- 23 master DLL circuits
- 11 PLL circuits
- 5 digital thermal sensors

Multiple clocking domains, local control
SRAM Dynamic Sleep Transistors

5-10x leakage reduction during “retention/standby”

Ref. K. Zhang, VLSI Circuits ‘04
Integrated Power Gates

- Shuts off both switching power and leakage power
- Enables idle cores to go to ~0 power, independent of state of other cores on die

Ref. R. Kumar, paper 3.2, ISSCC '09
Power Gates Enabled with Design+Process Co-optimization

Thick metal 9 layer for low resistance on-die power routing

Ultra-low leakage transistor for high off-resistance power gates
Dynamically delivering optimal performance and energy efficiency

Nehalem Turbo Mode

Many threaded workloads

- All cores operating

Lightly threaded workloads - Turbo Mode

- Power gates shut off some cores
- Zero power for inactive cores
- Higher frequency for active cores

Ref. R. Kumar, paper 3.2, ISSCC '09
Nehalem Power Control Unit

- Integrated proprietary microcontroller
- Shifts control from hardware to embedded firmware
- Real time sensors for voltage, temperature, current/power
- Flexibility enables sophisticated algorithms, tuned for current operating conditions

Ref. R. Kumar, paper 3.2, ISSCC ’09
Adaptive Frequency System

- Adaptive PLL frequency
  - Higher frequency during voltage peaks
  - Lower frequency during voltage droops
- Up to 5% frequency improvement at same voltage
- Lower power at same frequency

Ref. N. Kurd, VLSI Circuits ’08
Modern microprocessors integrate many of the separate system components from past platforms
## Microprocessor Evolution

<table>
<thead>
<tr>
<th>Feature</th>
<th>Intel386™</th>
<th>Nehalem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transistor Count</td>
<td>280 thousand</td>
<td>731 million</td>
</tr>
<tr>
<td>Frequency</td>
<td>16 MHz</td>
<td>&gt;3.6 GHz</td>
</tr>
<tr>
<td># Cores</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Cache Size</td>
<td>None</td>
<td>8 MB</td>
</tr>
<tr>
<td>I/O Peak Bandwidth</td>
<td>64 MB/sec</td>
<td>50 GB/sec</td>
</tr>
<tr>
<td>Adaptive Circuits</td>
<td>None</td>
<td>Sleep Mode</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Turbo Mode</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Power Gating</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Adaptive Frequency Clocking</td>
</tr>
</tbody>
</table>
45 nm SoC Transistors

Wider range of transistor types provided for SoC:
High performance and low power

Ref. C. Jan, IEDM ‘08
Wider range of transistor types provided for SoC:
High speed, high voltage I/O

Ref. C. Jan, IEDM '08
Devices for SoC Analog Circuits

Passive Elements
- Precision resistor
- High Q varactor
- High Q inductor

Active Elements
- RF CMOS

RF + Mixed Signal Circuits
- A to D, D to A converters
- RF transceiver
- LCPLL
- High speed I/O

Ref. C. Jan, IEDM ‘08
The Old Era of Microprocessor Scaling

Larger Cores
Higher Frequency
Higher Power

It has served us well for >30 years
The New Era of Microprocessor Scaling

Many-Core    Multi-Core    Multi-Function
System on a Chip

Avoiding the power wall requires a systemic approach from process technology through circuit design to micro-architecture to deliver products with power efficient performance
Outline

• Transistor Scaling

• Microprocessor Evolution

• Vision of the Future
Future Scaling Challenges

- Patterning ever-smaller features sizes
- Transistor and interconnect technologies that provide higher performance at lower power
- Continued voltage scaling for low power
- Integrating a wider range of device types for system-on-chip or system-in-package products
193 nm enhancements got us to the 32 nm generation
Layout Restrictions

65 nm Layout Style
- Bi-directional features
- Varied gate dimensions
- Varied pitches

32 nm Layout Style
- Uni-directional features
- Uniform gate dimension
- Gridded layout
Lithography Options for Beyond 32 nm

Double Patterning
- Pitch doubling
- Improved 2-D features

Computational Lithography
- Pixilated mask
- Existing 193 nm litho tools
Extreme Ultraviolet Lithography

Continued progress towards EUV implementation
Transistor Options

Substrate Engineering
+ Increased p-channel mobility
? Impact on n-channel mobility

Multi-Gate FETs
+ Improved electrostatics
+ Steeper sub-threshold slope
? Higher parasitic resistance
? Higher parasitic capacitance
III-V Transistor Options

- InSb PMOS QWFET
  - Peak $f_T > 140$ GHz at $V_{cc} = -0.5V$

- InGaAs NMOS QWFET
  - Peak $f_T > 400$ GHz at $V_{cc} = 0.5V$

III-V materials for improved performance at low voltage
3-D Chip Stacking

+ High density chip-chip connections
+ Small form factor
+ Combine dissimilar technologies

? Added cost
? Degraded power delivery, heat sinking
? Area impact on lower chip

3-D chip stacking using through-silicon vias
Optical Interconnects

Nearer term: High bandwidth chip-chip interconnects
Longer term: On-chip interconnects

Ref. I. Young, paper 28.1, ISSCC ‘09
High Density Memory

Dense memory increasingly important
Several novel directions being explored
System integration needed for performance, power, form factor
Challenge is to integrate wider range of heterogeneous elements
Higher Level System Integration

We’re trying to emulate nature’s capabilities

Organic
- Reptile
- Computing
- Sensors
- Power Supply
- Motion

Electronic
- Autonomous Vehicle
  Stanford entry
  2007 DARPA challenge
Evolutionary Comparison

Organic

- Complex Molecule
- Single-Cell Organism
- Multi-Cell Organism
- Reptile
- Human

Electronic

- Transistor
- Integrated Circuit
- Microprocessor
- Autonomous Vehicle
- Robot

What did nature have to "invent" to evolve to higher forms?
Brain Neuron

- Up to 1 meter in length
- ~50 um

<table>
<thead>
<tr>
<th></th>
<th>Neuron</th>
<th>Transistor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Charge carrier:</td>
<td>Ions</td>
<td>Electrons</td>
</tr>
<tr>
<td>Voltage swing:</td>
<td>100 mV</td>
<td>1.0 V</td>
</tr>
<tr>
<td>Threshold voltage:</td>
<td>10-20 mV</td>
<td>~300 mV</td>
</tr>
</tbody>
</table>

Nature is a master of low power operation

Neuron image from J. Nolte [36]
Organic vs. Electronic Circuits

Brain circuits are slow but massively parallel

Neuron image from J. Nolte [36]
Organic vs. Electronic Interconnects

Myelinated Axon

Myelin coating improves axon signal speed ~10x, but still slow

Cu + Low-k + Repeaters

Axon image from J. Nolte [36]
Organic vs. Electronic Systems

<table>
<thead>
<tr>
<th># Devices:</th>
<th>10^{11} Neurons</th>
<th>&gt;10^8 CPU Transistors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10^{14} Synapses</td>
<td>(✓) 10^{11} System Total</td>
</tr>
<tr>
<td>Input Devices:</td>
<td>Eyes, Ears, Taste, Touch, Smell</td>
<td>(✓) Keyboard, Radio, USB Port</td>
</tr>
<tr>
<td>Operating Freq:</td>
<td>100 Hz</td>
<td>&gt;2 GHz (✓)</td>
</tr>
<tr>
<td>Power:</td>
<td>20 Watts (✓)</td>
<td>40 Watts</td>
</tr>
</tbody>
</table>

We have a way to go and much to learn
Conclusion

• Moore’s Law continues, but the formula for success is changing
  – New materials and device structures are needed to continue scaling
  – Circuit design and micro-architecture innovations focus more on power efficiency

• System level integration is increasingly important
  – Success will be determined by ability to integrate a wider and more heterogeneous set of components

• Organic evolution has given us some clues for effective higher level system integration
  – Low power operation
  – Massive parallelism
  – Integrated sensors