Scaling of MOSFETs

Introduction to scaling

Scaling is a discipline of science applicable to many areas such as mechanics (*e. g*. MEMS), electronics (*e. g*. ICs), and optics (*e. g*. micro-optics). As the dimensions of an object are linearly scaled up or down, many properties scale do not scale linearly.

Example: As the linear dimensions of an object are scaled, how do the following properties scale?

Surface area	Volume
Mass and weight	Mechanical stability
Electrical characteristics	Optical properties

➔ Different properties can scale very differently
➔ It will be useful to know the scaling laws

© E. F. Schubert, Rensselaer Polytechnic Institute, 2003

Introduction to MOSFET scaling

The lateral geometric dimensions of devices and interconnects are reduced. This reduction in size is referred to as "scaling" of the geometric dimensions of the integrated circuit (IC).

Minimum feature size is smallest size of object (*e. g.* gate length or interconnect linewidth) on IC.

Minimum feature size of ICs has shrunk considerably over the time of several decades. As a consequence, the number of transistors has increased over time.

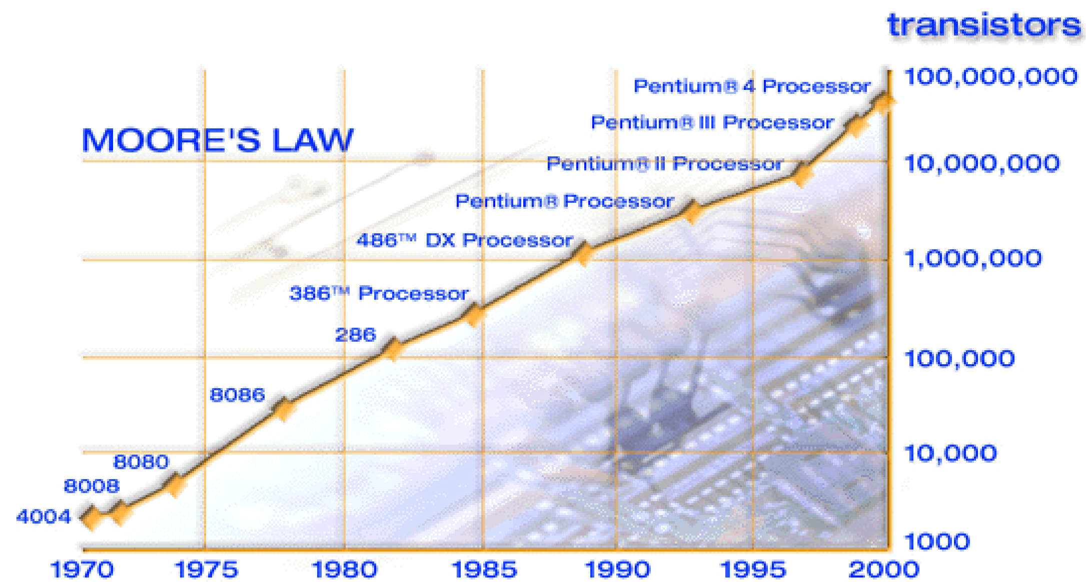
© E. F. Schubert, Rensselaer Polytechnic Institute, 2003

Processor name	Year of introduction	Transistors
4004	1971	2,250
8008	1972	2,500
8080	1974	5,000
8086	1978	29,000
286	1982	120,000
386	1985	275,000
486 DX	1989	1,180,000
Pentium	1993	3,100,000
Pentium II	1997	7,500,000
Pentium III	1999	24,000,000
Pentium 4	2000	42,000,000

This table shows the year of introduction and the number of transistors on the processors of Intel Corporation.

© E. F. Schubert, Rensselaer Polytechnic Institute, 2003

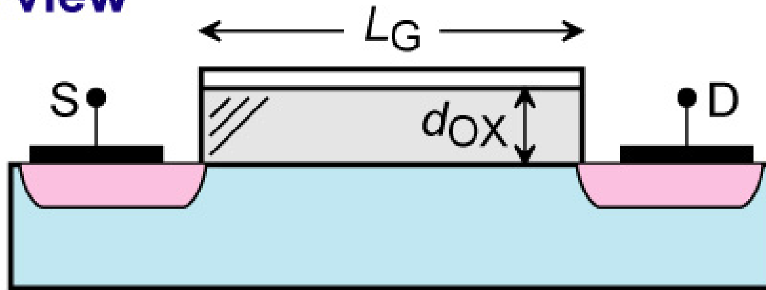
The name Gordon E. Moore (co-founder of Intel Corp.) is frequently associated with MOSFET scaling (see, for example Gordon E. Moore "Cramming more components onto integrated circuits" *Electronics*, April 19, 1965).

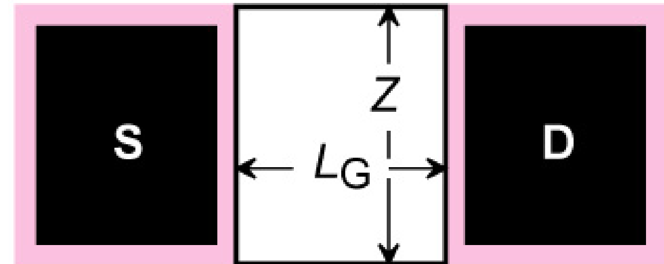


(after Intel Corporation)

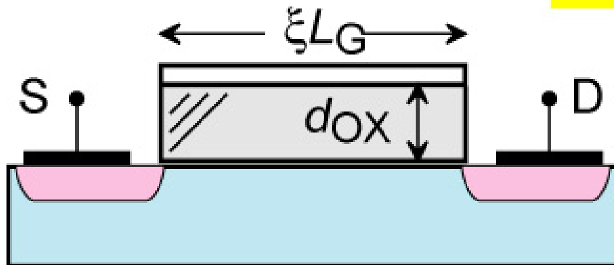
© E. F. Schubert, Rensselaer Polytechnic Institute, 2003

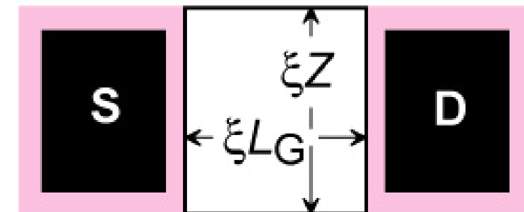
Scaling of a MOSFET:





Scaling factor ξ





Definition: The scaling parameter ξ is the pre-factor by which dimensions are reduced. It is $\xi < 1$.

Gate length	L_G	$\Rightarrow$	$\xi\ L_G$
Gate width	Z	$\Rightarrow$	$\xi\ Z$

© E. F. Schubert, Rensselaer Polytechnic Institute, 2003

Recall the basic result of the gradual channel approximation:

$$I_{\mathrm{D,\,sat}} = \frac{\varepsilon_{\mathrm{OX}}\,\mu Z}{d_{\mathrm{OX}}\,L_{\mathrm{G}}}\,\frac{1}{2}\,V_{\mathrm{DS,\,sat}}^{2} = \frac{\varepsilon_{\mathrm{OX}}\,\mu Z}{2\,d_{\mathrm{OX}}\,L_{\mathrm{G}}}\left(V_{\mathrm{GS}} - V_{\mathrm{th}}\right)^{2}$$

and

$$g_{\mathrm{m,\,sat}} = \frac{\mathrm{d}I_{\mathrm{D,\,sat}}}{\mathrm{d}V_{\mathrm{GS}}} = \frac{\varepsilon_{\mathrm{OX}}\,\mu Z}{d_{\mathrm{OX}}\,L_{\mathrm{G}}}\left(V_{\mathrm{GS}} - V_{\mathrm{th}}\right)$$

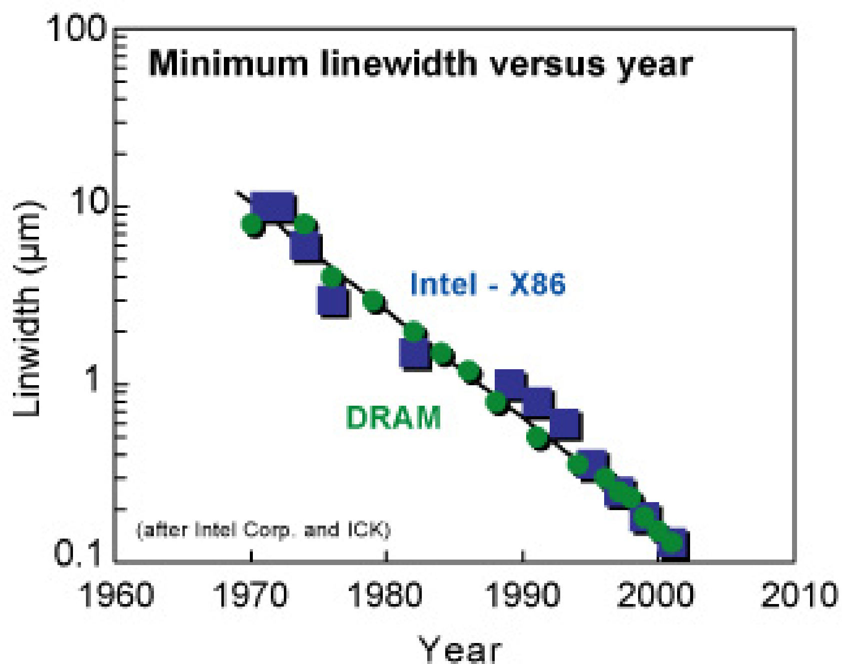
The gradual channel approximation shows:

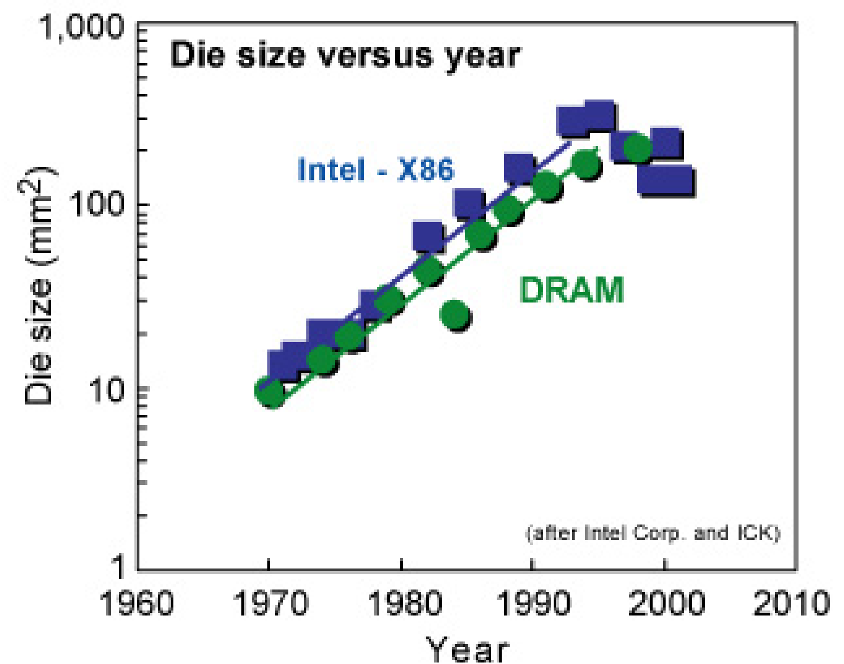
Shorter gate length results in higher transconductance.
Smaller gate thickness results in higher transconductance.

© E. F. Schubert, Rensselaer Polytechnic Institute, 2003

Gate capacitance ∝ L_G Z

Gate capacitance scales with ξ^2.
Smaller gate width results in smaller input capacitance.



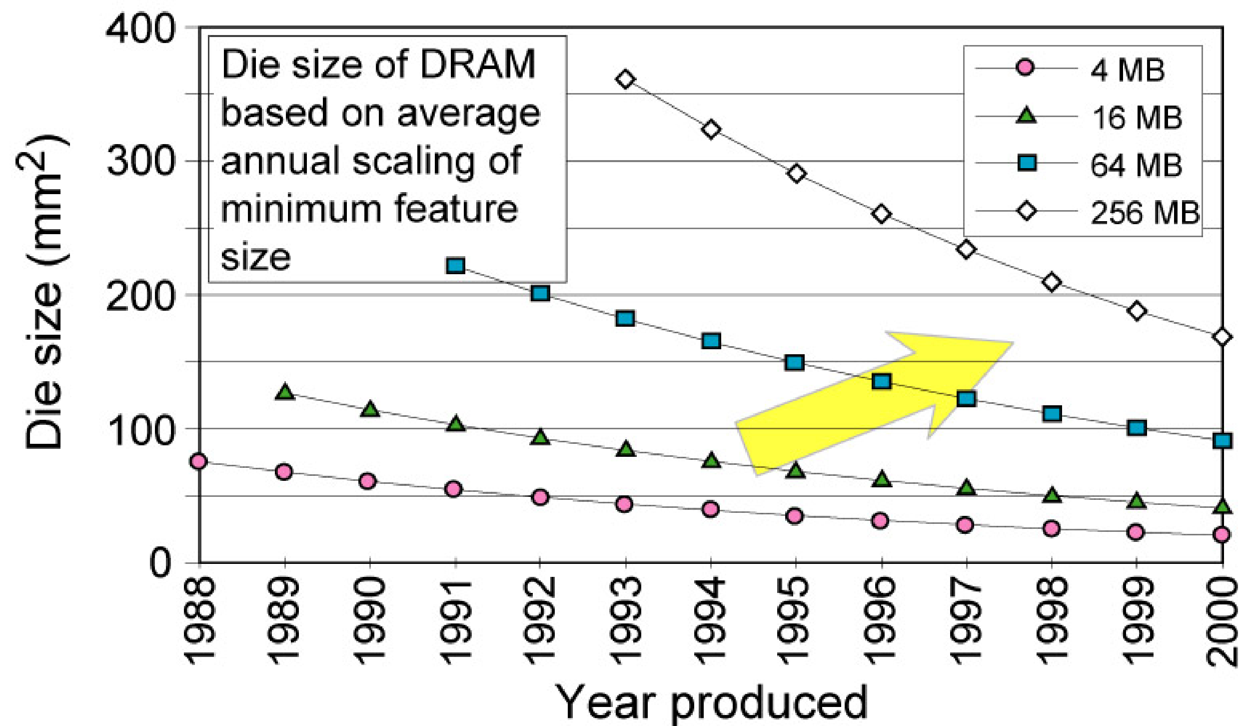


Linewidths are shrunk on a two to three year cycle. The linewidth in 2002 was 0.13 – 0.18 µm and a typical die size was 2 cm^2.

© E. F. Schubert, Rensselaer Polytechnic Institute, 2003

The number of transistors per IC grows due to higher computing demand, shrinking linewidth, and growth of die size. Even if die size were a constant, the number of transistors per IC would grow.

Illustration of die size of DRAMs versus year:



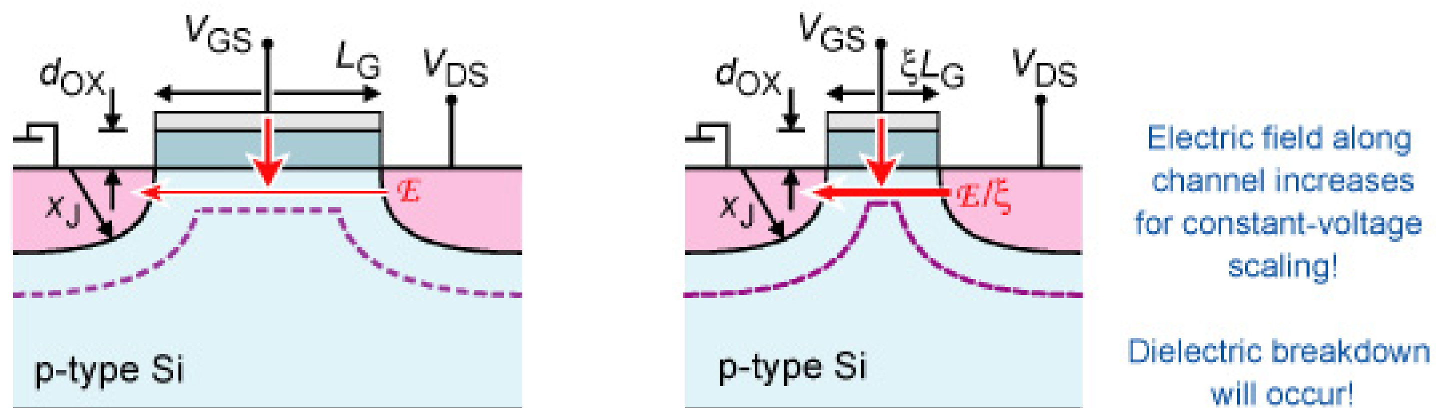
© E. F. Schubert, Rensselaer Polytechnic Institute, 2003

Constant-voltage scaling

There are two types of scaling, namely constant-voltage scaling and constant-field scaling. Next, we discuss constant-voltage scaling.

In constant-voltage scaling, only the lateral dimensions of the MOSFET are scaled. Thus, constant-voltage scaling is a purely geometrical process.

What happens as the dimensions shrink?



© E. F. Schubert, Rensselaer Polytechnic Institute, 2003

If constant-voltage scaling (*i. e*. pure geometrical scaling) is performed, L_G and Z are reduced by the scaling factor ξ. Because the drain-source voltage remains unchanged, the lateral field increases. It increases approximately by the factor ξ, since $\mathcal{E} \propto V_{DS} / L_G$, at least in the ohmic regime.

This will lead eventually to very high fields in the channel so that dielectric breakdown (avalanche breakdown) will occur.

Therefore, the voltages must be reduced as well. As V_{DS} is reduced, so must be V_{th}. As V_{th} is reduced, so must be d_{OX}.

Thus the electric field must stay constant. This brings us to constant-field-scaling.

Consequences of constant-voltage scaling

Consider an integrated circuit that consists of transistors. Assume that one transistor charges the input capacitance of a following transistor.

© E. F. Schubert, Rensselaer Polytechnic Institute, 2003

As L_G and Z of the transistor are scaled by the scaling factor, ξ, the drain current of the transistor does not change because

$$I_D \propto Z$$

and

$$I_D \propto 1 / L_G$$

so that

$$\text{Drain current} \qquad I_D \qquad \Rightarrow \qquad \xi \, \xi^{-1} \, I_D = I_D$$

Thus the drain current is not changed.

However, as L_G and Z of the transistor are scaled by the scaling factor, ξ, the gate capacitance changes because

$$C_G \propto Z \, L_G$$

so that

© E. F. Schubert, Rensselaer Polytechnic Institute, 2003

Gate capacitance $C_G \Rightarrow \xi^2 C_G$

Thus the gate capacitance strongly changes upon scaling.

The time required to charge the gate capacitance with a drain current so that a threshold voltage V_{th} is reached, can be derived form the equation $Q = C_G V_{th}$. The time is given by:

$$\tau = C_G / (\mathrm{d}Q / \mathrm{d}t) = C_G / I_D$$

Thus the charging time changes according to

Charging time $\tau \Rightarrow \xi^2 \tau$

As a consequence, scaled integrated circuits can operate at higher frequencies.

These factors allow one to attain higher clock speed in integrated circuits as they are scaled down.

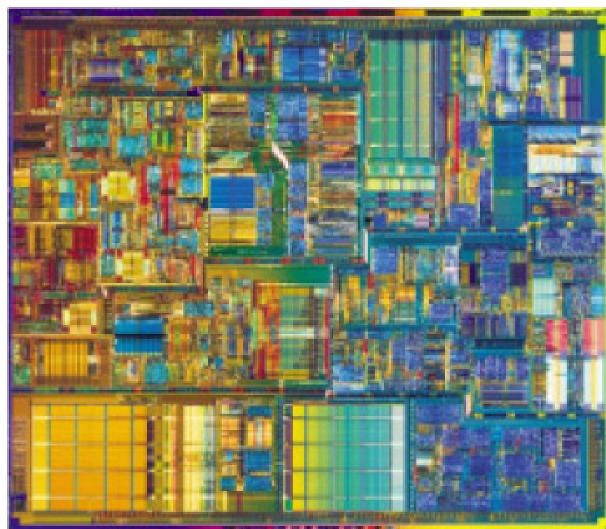
© E. F. Schubert, Rensselaer Polytechnic Institute, 2003

The following table illustrates performance increase as linewidths shrink.

Year	Linewidth	Clockspeed
Beginning of 1970s	10 μm	1 MHz
End of 1970s	3 μm	5 MHz
Beginning of 1980s	2 μm	20 MHz
End of 1980s	0.8 μm	50 MHz
Beginning of 1990s	0.5 μm	100 MHz
End of 1990s	0.25 μm	750 MHz
Beginning of 2000s	0.13 μm	2 GHz

Table shows typical linewidths and clock speeds versus year.

© E. F. Schubert, Rensselaer Polytechnic Institute, 2003



Intel's Pentium 4 processor

(after Intel Corp.)

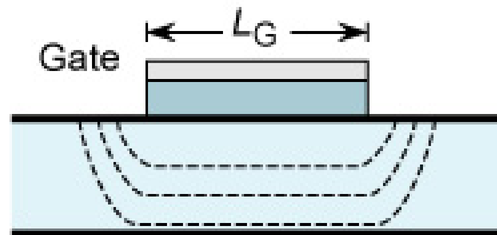
Picture shows that processor IC has different areas, namely "random logic" and "ordered arrays". Interconnect structure is very different in these areas. Scaling laws will be different for these different areas.

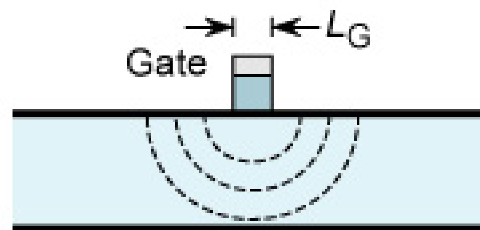
© E. F. Schubert, Rensselaer Polytechnic Institute, 2003

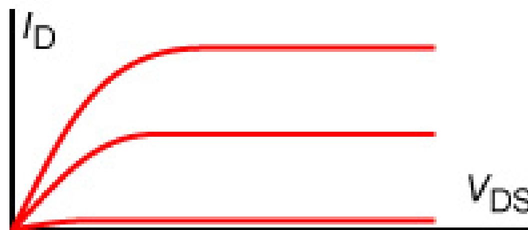
Short-channel effects

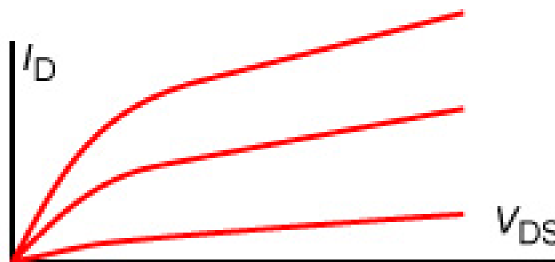
As the gate length is reduced, the characteristics of a MOSFET change due to short-channel effects, *i. e.* effects that arise at very short gate (and channel) lengths.

At short gate lengths, the electrostatic field no longer resembles that of a planar capacitor.









Short-channel effects:

Threshold-voltage shift

Lack of pinch-off

Increased leakage current

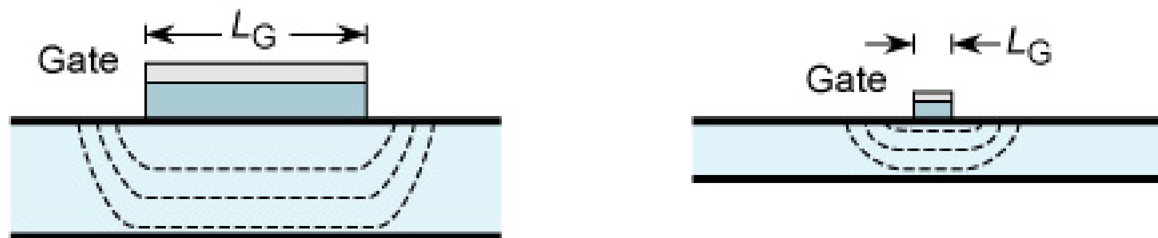
Increase of output conductance

© E. F. Schubert, Rensselaer Polytechnic Institute, 2003

At short gate lengths, FETs suffer from;

- Threshold voltage shift
- Clear pinch-off of channel
- Increased leakage current
- Increased output conductance
- Minimization of leakage currents is enormously important in VLSI. **Why?**

How can short-channel effects be mitigated? By "vertical" scaling, *i. e.* by scaling the lateral dimensions but also by the dimensions perpendicular to the wafer plane!



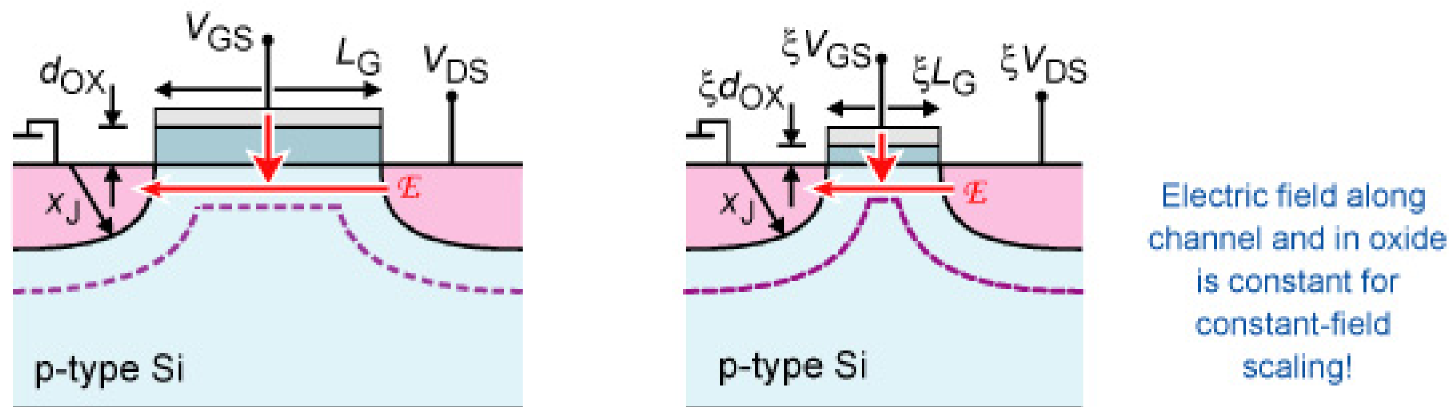
By scaling all dimensions, the original field distribution can be re-created.

© E. F. Schubert, Rensselaer Polytechnic Institute, 2003

Constant-field scaling

In constant-field scaling, the lateral dimensions (L_G, Z), the perpendicular dimensions (d_{OX}), and the voltages (V_{DS}, V_{GS}, V_{th}) of the MOSFET are scaled. Thus, constant-field scaling is ***not*** a purely geometrical process.

What happens as the dimensions shrink?



If constant-field scaling is performed, L_G, Z, d_{OX}, V_{DS}, V_{GS}, and V_{th} are reduced by the scaling factor ξ. Electric field in the channel and oxide are

© E. F. Schubert, Rensselaer Polytechnic Institute, 2003

independent of the factor ξ (with $\xi < 1$), since $\mathcal{E}_{\text{Channel}} \propto V_{DS} / L_G$ (at least in the ohmic regime) and $\mathcal{E}_{OX} \propto V_{th} / d_{OX}$.

Gate length	L_G	$\Rightarrow$	ξL_G
Gate width	Z	$\Rightarrow$	ξZ
Oxide thickness	d_{OX}	$\Rightarrow$	ξd_{OX}
DS voltage	V_{DS}	$\Rightarrow$	ξV_{DS}
Threshold voltage	V_{th}	$\Rightarrow$	ξV_{th}

Thus, by constant-field scaling, the original magnitudes of $\mathcal{E}$ fields are re-established (approximately), hence the term constant field scaling.

© E. F. Schubert, Rensselaer Polytechnic Institute, 2003

Punch-through effect

However, the scaled structure has another problem, namely the closer proximity of source and drain depletion regions.

As a consequence, the device is now more susceptible to punch-through breakdown.

To avoid this problem, the substrate doping is increased to reduce the depletion widths

Junction depths are also reduced to mitigate short-channel V_{th} effect.

Doping

The acceptor doping N_A is increased by a factor by ξ in order to reduce the depletion widths to prevent punch-through breakdown.

In addition, the n-type implants ("bath tubs") are scaled, *i. e*. the n-type doping concentration is increased by a factor ξ and the junction depth is decreased by a factor ξ. This lessens short-channel V_{th} effects.

© E. F. Schubert, Rensselaer Polytechnic Institute, 2003

Consequences of constant-field scaling within framework of Shockley's gradual-channel-approximation model

Recall the drain saturation current and transconductance:

$$I_{\mathrm{D,\,sat}} = \frac{\varepsilon_{\mathrm{OX}}\,\mu\,Z}{d_{\mathrm{OX}}\,L_{\mathrm{G}}}\,\frac{1}{2}\,V_{\mathrm{DS,\,sat}}^{2} = \frac{\varepsilon_{\mathrm{OX}}\,\mu\,Z}{2\,d_{\mathrm{OX}}\,L_{\mathrm{G}}}\left(V_{\mathrm{GS}} - V_{\mathrm{th}}\right)^{2}$$

$$g_{\mathrm{m,\,sat}} = \frac{\mathrm{d}I_{\mathrm{D,\,sat}}}{\mathrm{d}V_{\mathrm{GS}}} = \frac{\varepsilon_{\mathrm{OX}}\,\mu\,Z}{d_{\mathrm{OX}}\,L_{\mathrm{G}}}\left(V_{\mathrm{GS}} - V_{\mathrm{th}}\right)$$

Assuming V_{GS} = 0 V and the scaling dependences for L_G, Z, d_{OX} mentioned above, the following dependences are found:

Drain current I_D $\Rightarrow$ $\xi\,\xi^{-1}\,\xi^{-1}\,\xi^{2}\,I_D = \xi\,I_D$

Transconductance

Transconductance g_m $\Rightarrow$ $\xi \, \xi^{-1} \, \xi^{-1} \, \xi \, g_m = g_m$

The static power dissipation of one transistor is given by $P = I \, V$:

Static power per device P $\Rightarrow$ $\xi \, \xi \, P = \xi^2 \, P$

Power per transistor area:

Power / area $P / (Z \, L_G)$ $\Rightarrow$ $(\xi^2 / \xi^2) \, P = P$

The delay is given by $\tau = C_G / I_D$:

Delay $\tau = C_G / I_D$ $\Rightarrow$ $[(\xi^2 / \xi) / \xi] \, \tau = \tau$

It follows the power delay product:

Power × delay $P \, \tau$ $\Rightarrow$ $\xi^2 \, P \, \tau$

© E. F. Schubert, Rensselaer Polytechnic Institute, 2003

As the delay time does not change, the clock frequency would remain constant:

Clock frequency	$f = 1 / \tau$	$\Rightarrow$	f

The energy consumed per device per clock cycle was given above:

Power × delay	$P \tau$	$\Rightarrow$	$\xi^2 P \tau$

Thus the dynamic power consumption is given by:

Dynamic power	$P \tau f$	$\Rightarrow$	$\xi^2 P$

Dynamic power per transistor area:

Dynamic power / area	$P \tau f / (Z L_G)$	$\Rightarrow$	$(\xi^2 / \xi^2) P = P$

However, we would like the next generation of scaled-down integrated circuits to work faster than the last generation, i.e. ICs should work at an

© E. F. Schubert, Rensselaer Polytechnic Institute, 2003

increasingly higher clock frequency. This can be accomplished by scaling the supply voltage not as aggressively.

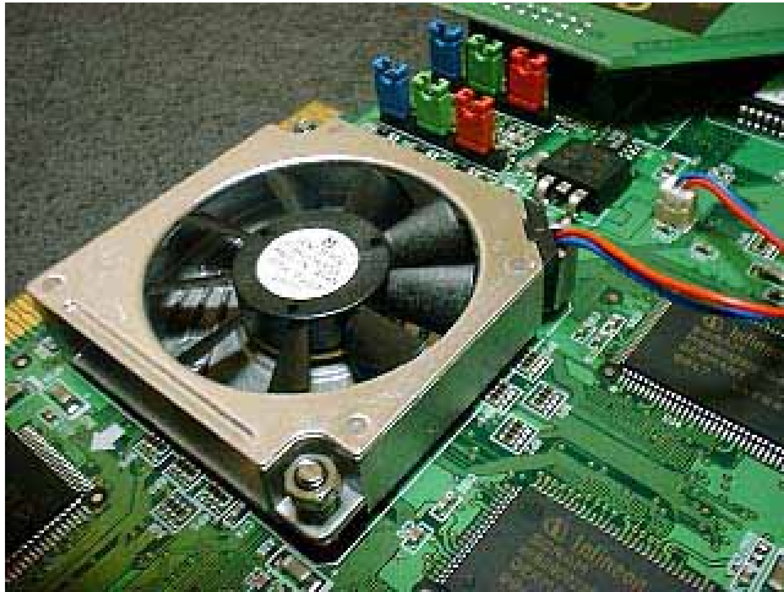
- Supply voltage does is not scaled by ξ but by a smaller factor, e. g. $(\xi)^{1/2}$.

Such scaling results in an increase in the channel electric field and in a higher drain current. As a result the delay time shrinks and transistors can be operated at a higher frequency. This in turn increases the dynamic power consumption per unit area. (Also note that velocity saturation effects become important at very small gate lengths leading to smaller-than-expected amplifications.)

There is an important consequence from this. The power per device reduction by ξ^2 is optimistic and the realistic power consumption scales with less than ξ^2. However, the area reduction scales exactly with ξ^2. Thus the dynamic power consumption per unit area scales with a factor ($> \xi^2/ \xi^2$). As a result, the power consumption per unit area increases as ICs are scaled down. That means, scaled ICs run increasingly hot. This is a severe performance limitation for future ICs.

© E. F. Schubert, Rensselaer Polytechnic Institute, 2003

Solution: Cooling fins and chip fans.



© E. F. Schubert, Rensselaer Polytechnic Institute, 2003

Consequences of constant-field scaling within framework of the saturated-velocity model

Recall the drain saturation current and transconductance:

$$I_{\mathrm{D,\,sat}} = -\frac{\varepsilon_{\mathrm{OX}}}{d_{\mathrm{OX}}} \left(V_{\mathrm{GS}} - V_{\mathrm{th}} \right) v_{\mathrm{sat}}\, Z$$

$$g_{\mathrm{m,\,sat}} = \frac{\mathrm{d}I_{\mathrm{D,\,sat}}}{\mathrm{d}V_{\mathrm{GS}}} = -\frac{\varepsilon_{\mathrm{OX}}}{d_{\mathrm{OX}}}\, v_{\mathrm{sat}}\, Z$$

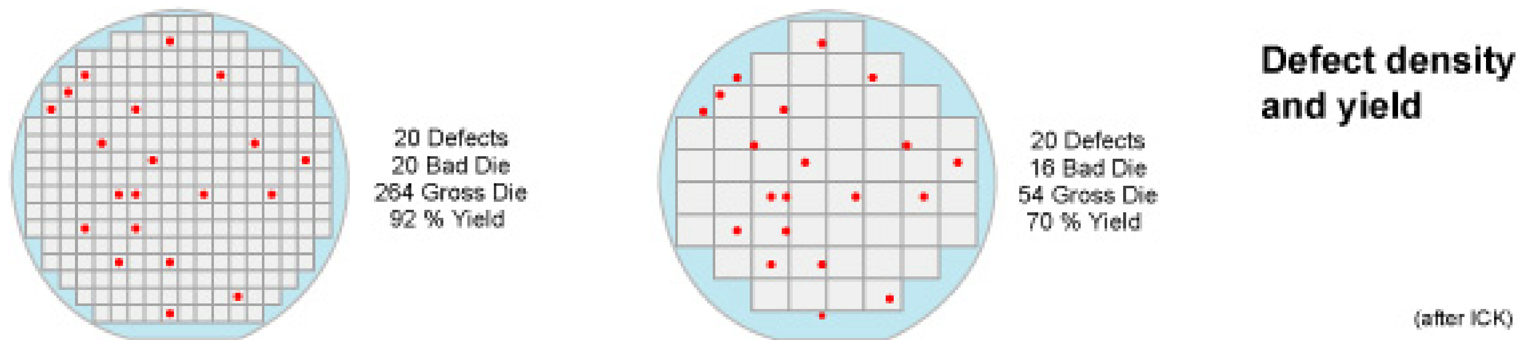
If the same analysis as done above is done for the saturated-velocity model, identical results are obtained.

© E. F. Schubert, Rensselaer Polytechnic Institute, 2003

Yield, defect density and die size

Yield, defect density, and die size are related.

Figure illustrates the relation of yield, die size and yield.



For large die sizes, yield can go to zero!

Reduction of defect density is especially important for large die size manufacturing.

© E. F. Schubert, Rensselaer Polytechnic Institute, 2003

Maintaining a clean environment is especially important for large die size manufacturing.

One small defect can ruin a 2 cm^2-area die!

The situation is very different for discrete device manufacturing, e. g. small-area transistor, LED, or laser manufacturing.

For a given defect density, the yield approaches 0 % and 100 % in the limit of large and small die sizes, respectively.

© E. F. Schubert, Rensselaer Polytechnic Institute, 2003

After scaling process reaches limits …

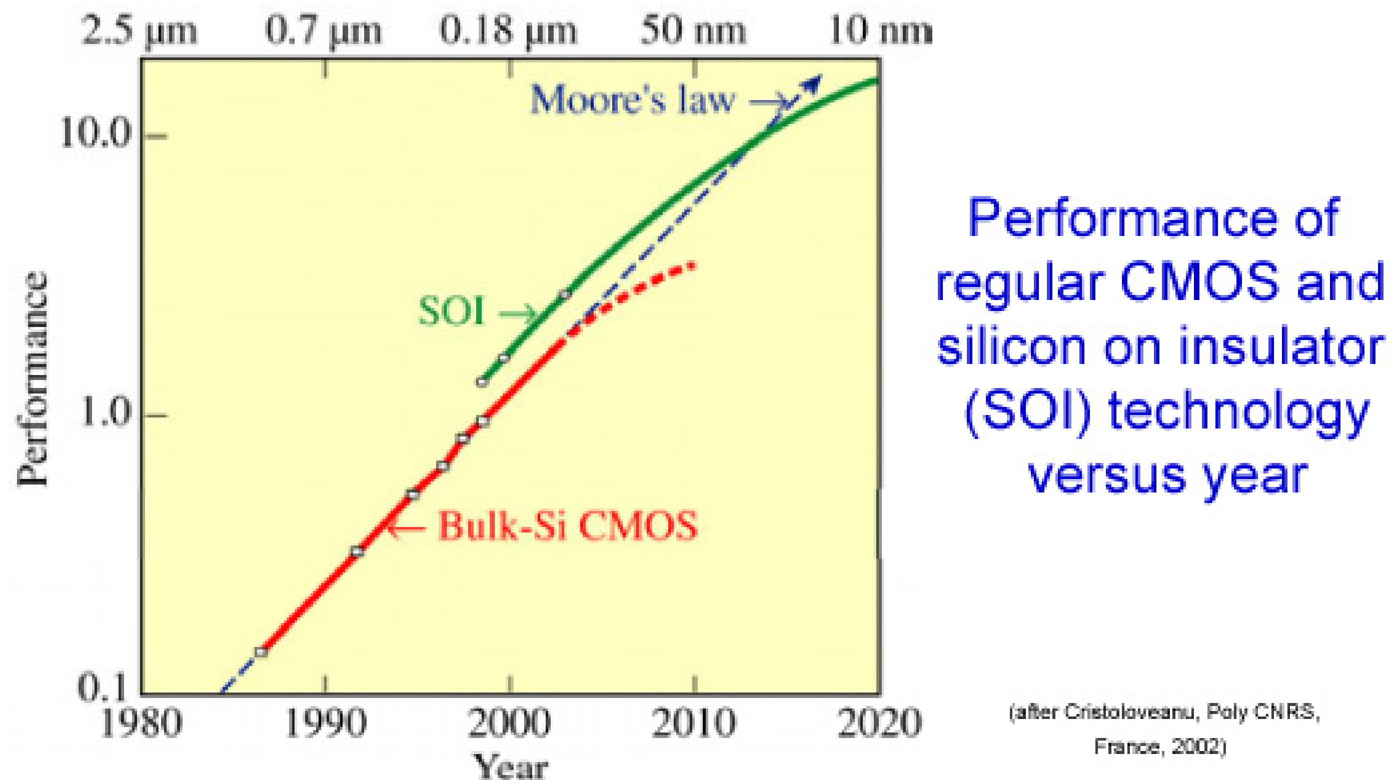
… new ideas, technologies, and concepts will be required for continued performance improvements.

… examples of new ideas, technologies, and concepts:

- Silicon on insulator (SOI) technology
- SiGe and SiC technology
- 3D integration
- Optical interconnects
- DRAM, S-DRAM, DDR memory, RAMBUS memory
- Replacement of hard drives by flash memory

© E. F. Schubert, Rensselaer Polytechnic Institute, 2003

Example: Expected advances based on SOI technology



Due to inherent performance advantages, SOI out-performs conventional CMOS technology.

© E. F. Schubert, Rensselaer Polytechnic Institute, 2003