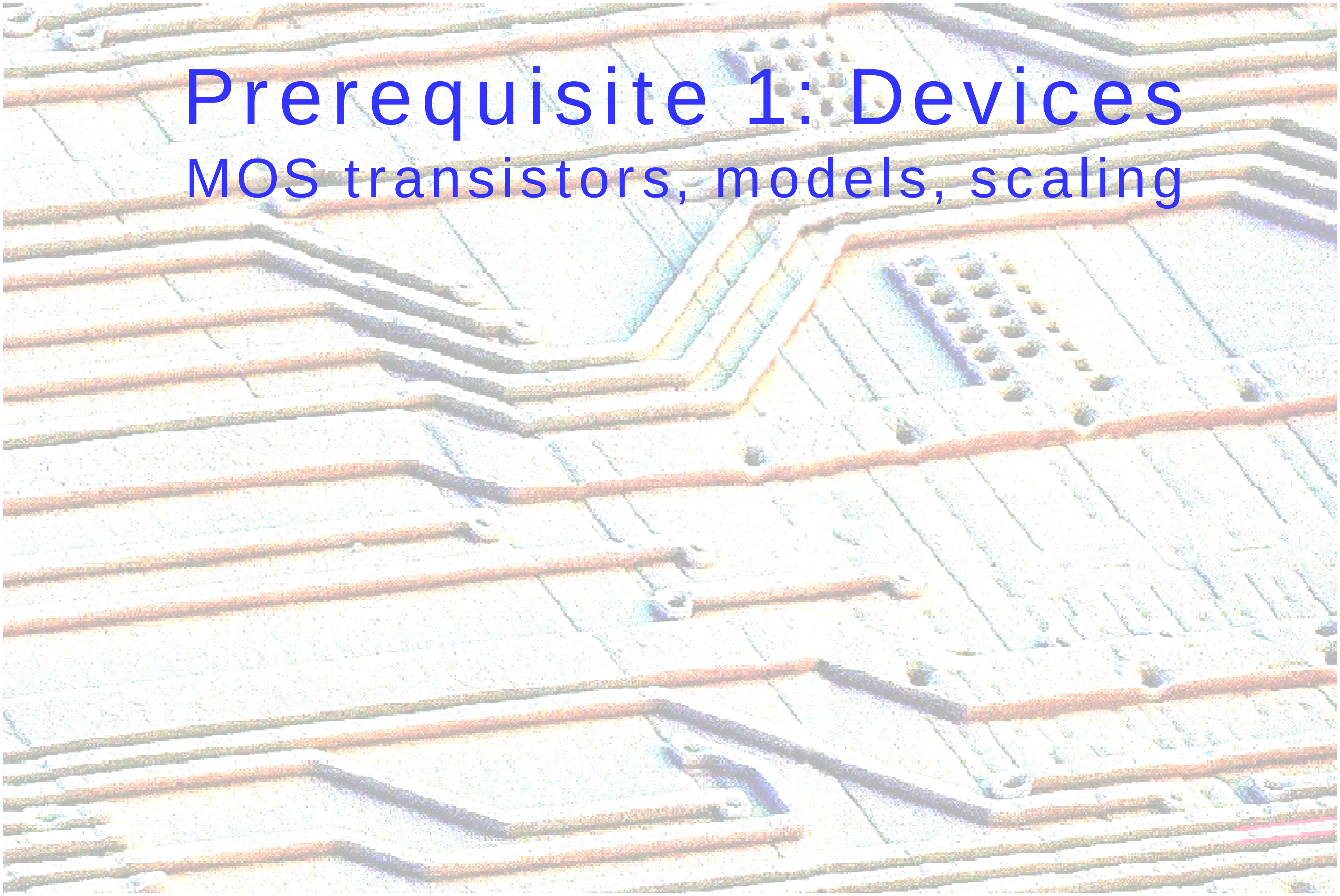
# Prerequisite 1: Devices

## MOS transistors, models, scaling

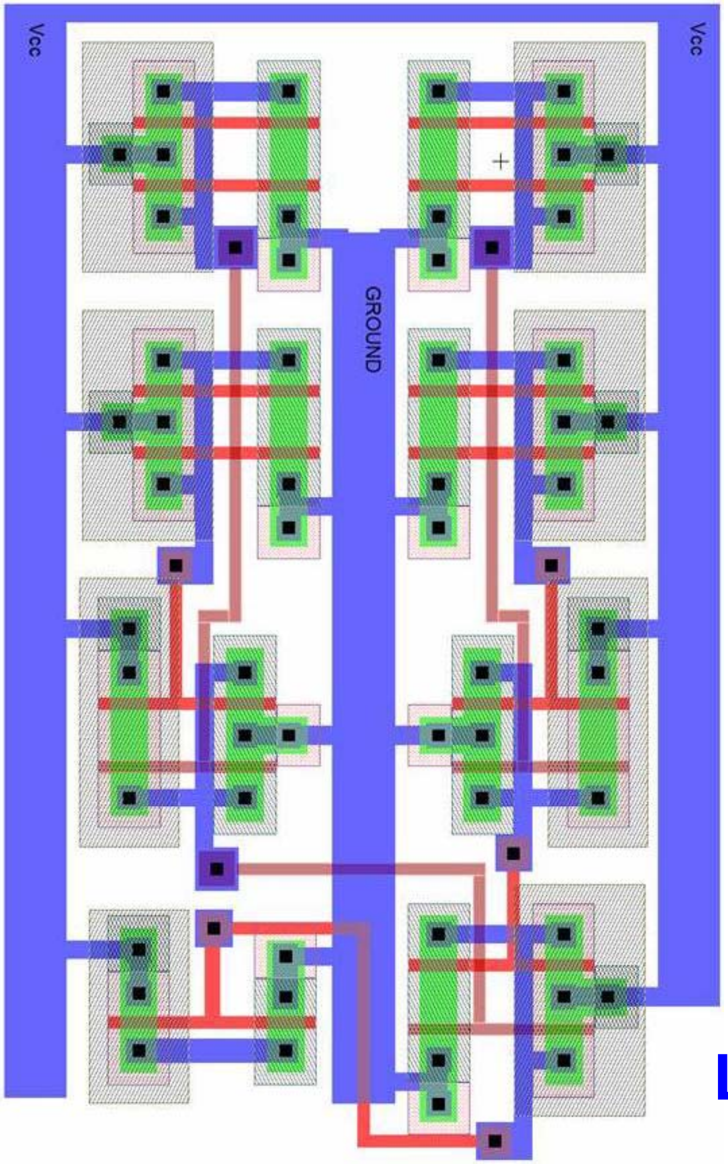


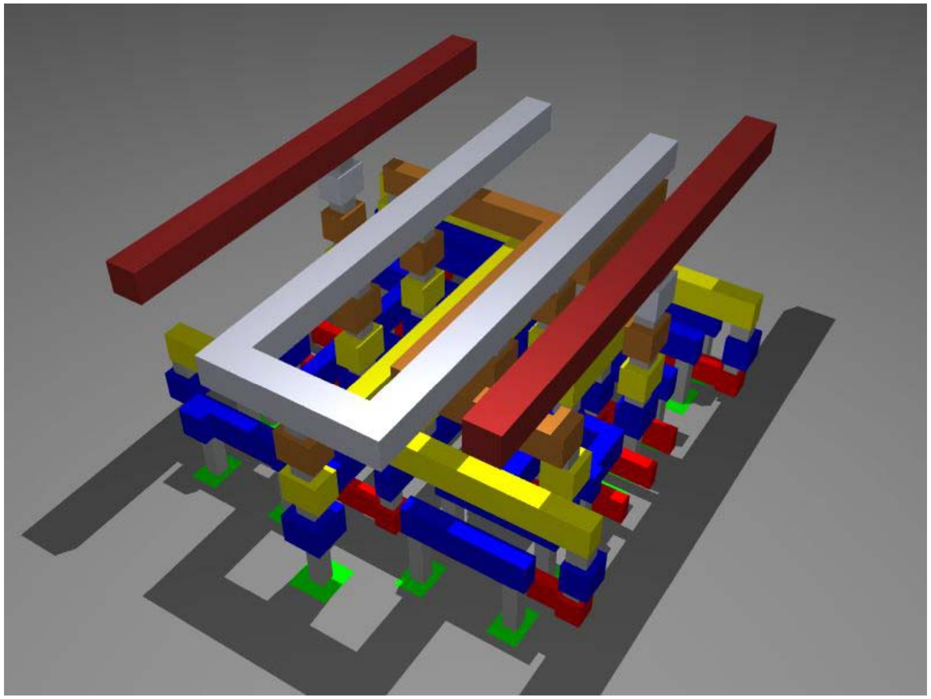
# Prerequisites

- **Qualitative (intuitive) understanding of device operation**
- **basic device equations**
- **models for manual analysis**
- **Understanding models for SPICE simulation**
- **understanding of secondary and deep-sub-micron effects**
- **Future trends**

- **In depth:**
  - **ET4392 - Physics of Semiconductor Devices**
    **0/0/2+2/0 (René van Swaaij), compulsory**
  - **Neamen: *Semiconductor Physics and Devices, basic principles*,**
    **2003, McGraw Hill**
  - **Taur and Ning: *Fundamentals of Modern VLSI Devices***
    **1998, Cambridge University Press**

# Chip Anatomy





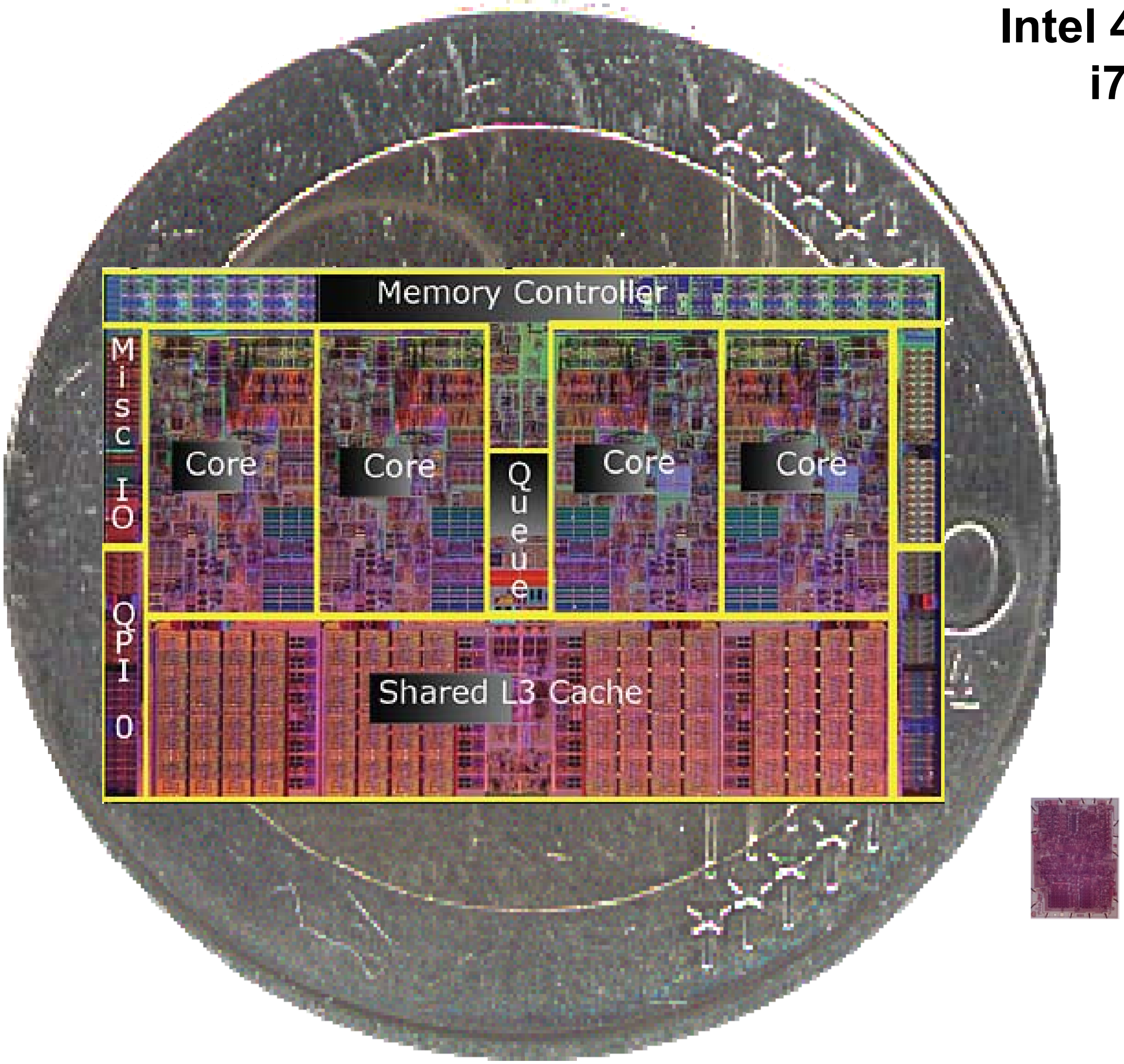
(Rendering of) result after fabrication

Not 1-to-1, sorry

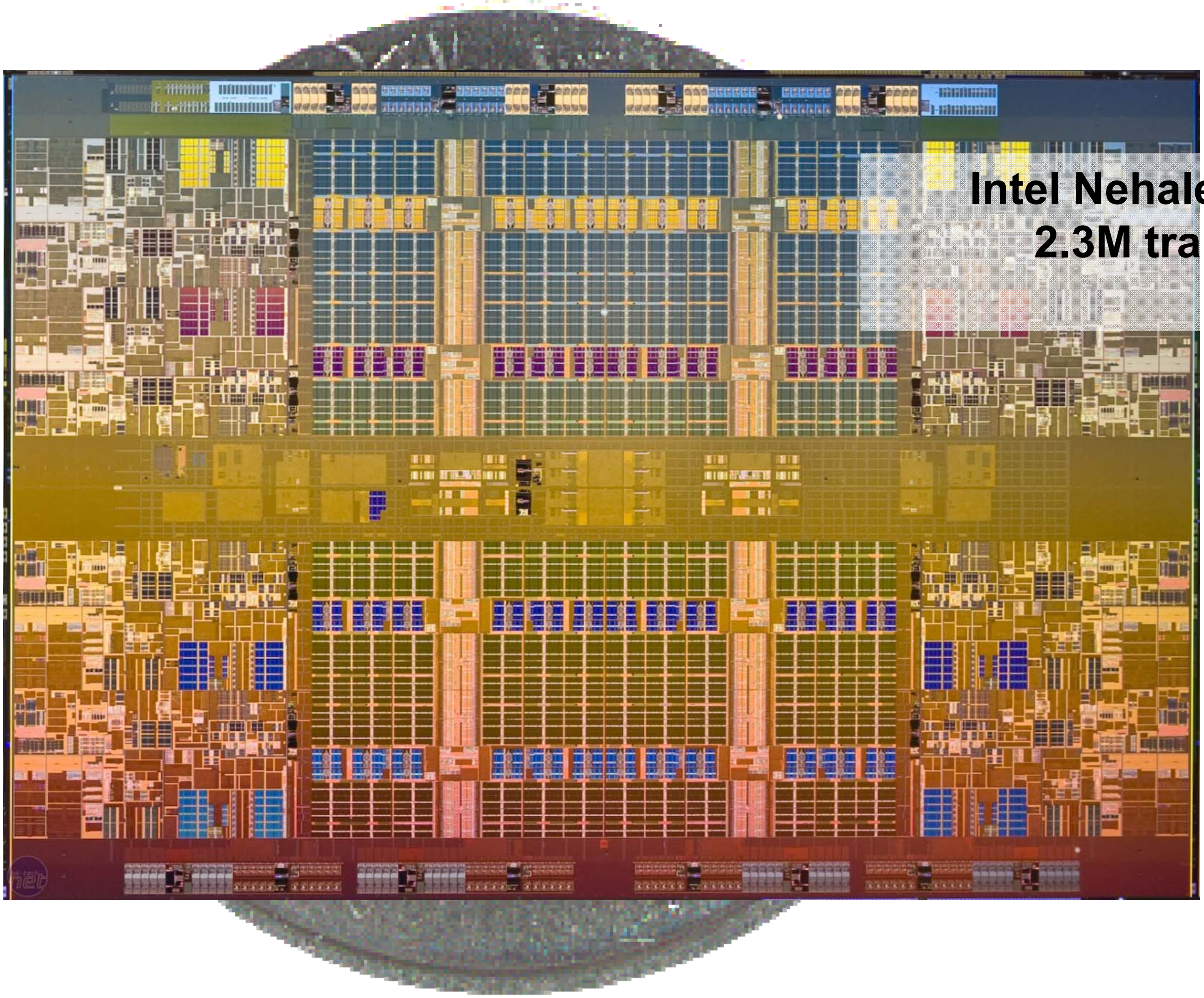
Layout of chip, final design result

# Evolution

**Intel 4004 (1971) and Core i7 (Nehalem, 2008) die compared to 2€ coin**

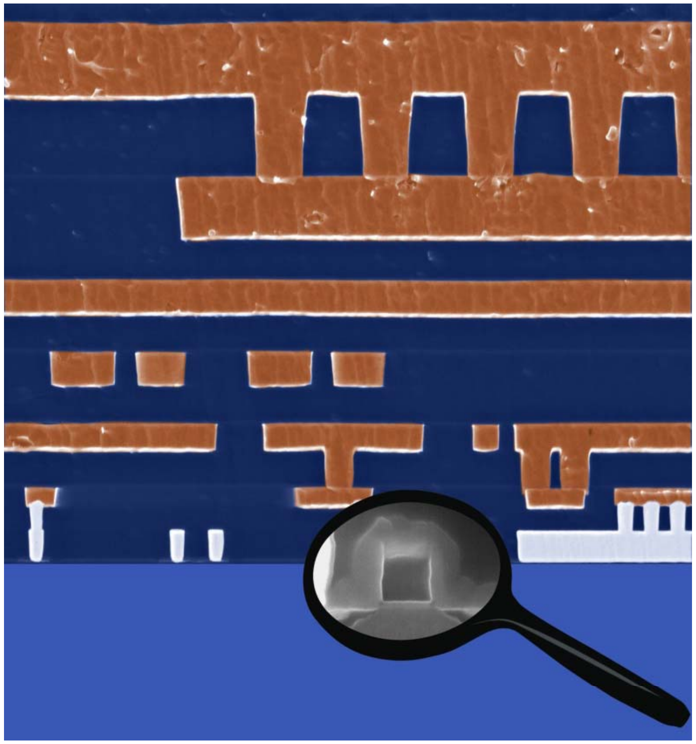


# Evolution

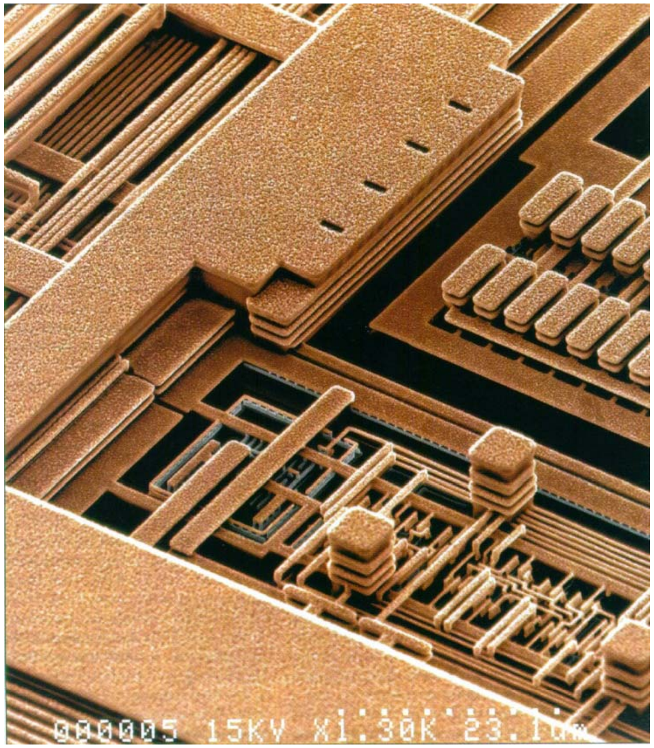


**Intel Nehalem EX, 8 core,**
**2.3M transistors, 2010**
**(684 mm$^2$)**

# SEM Images

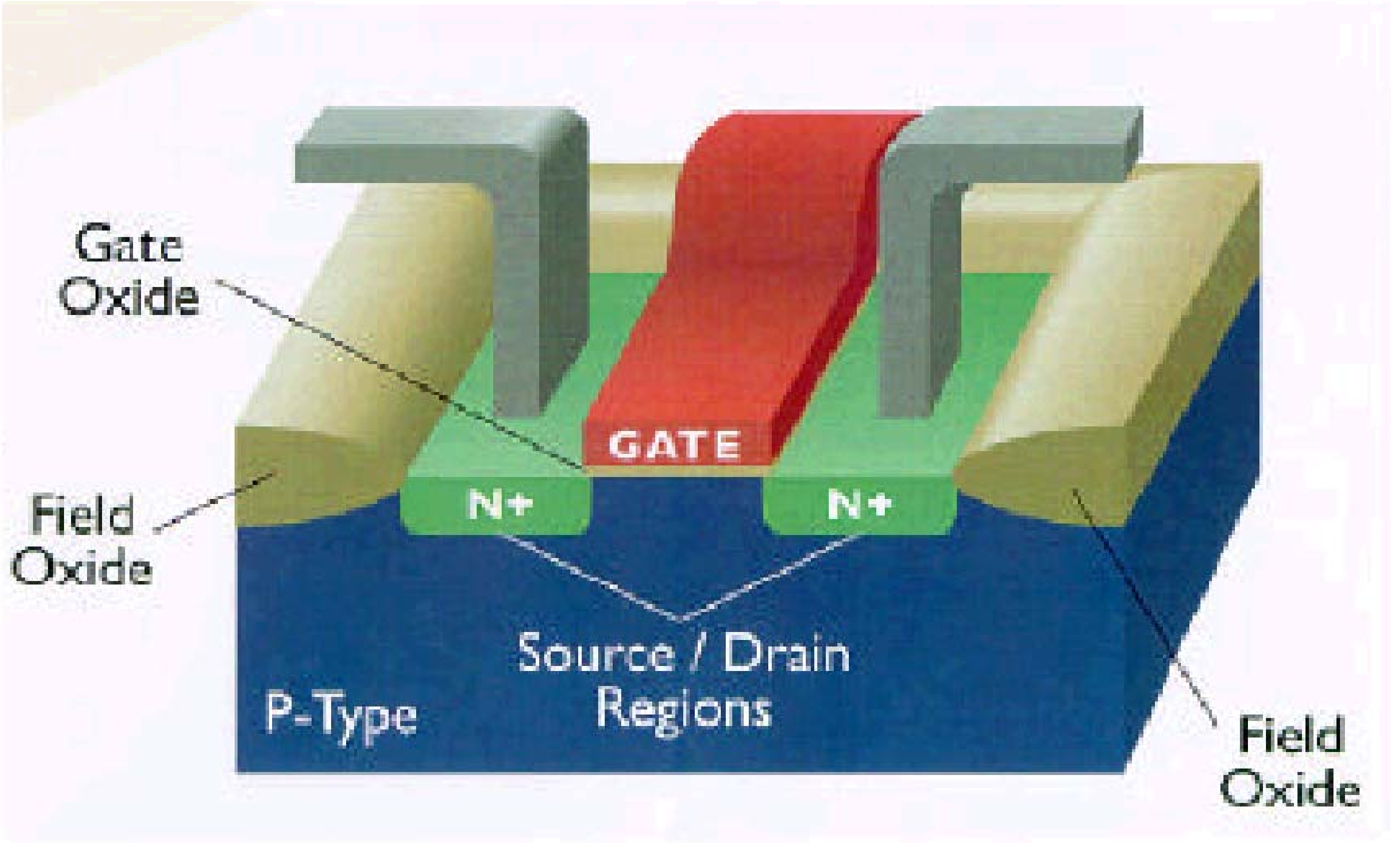


Cross-section

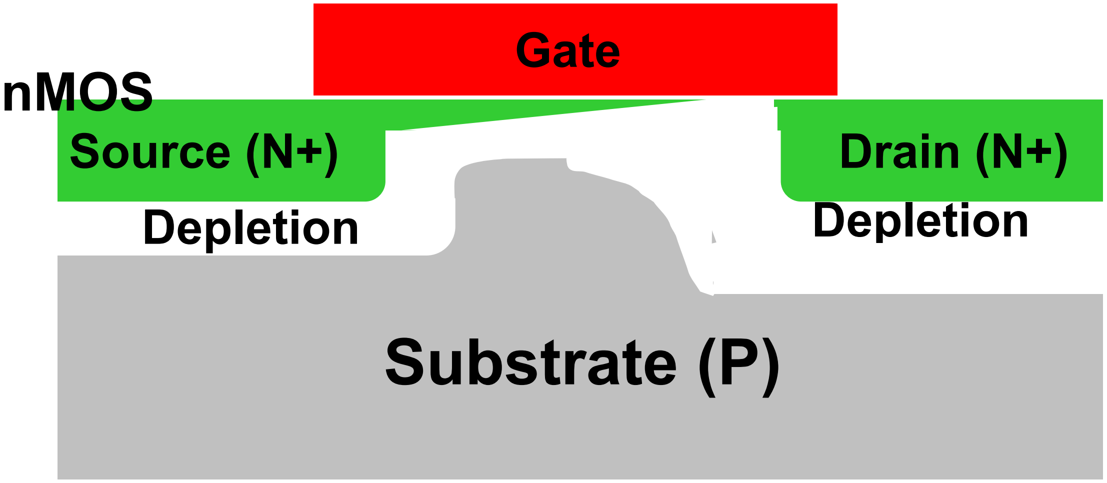


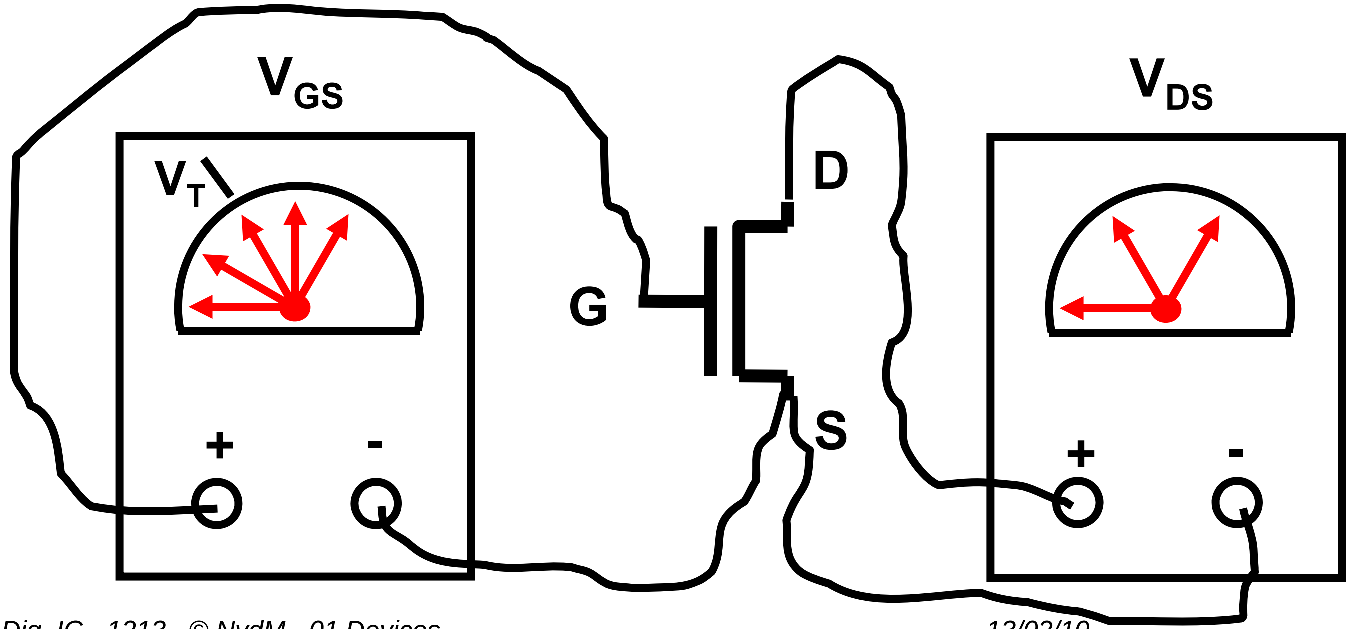
3d perspective

# MOS Transistor 3D Structure



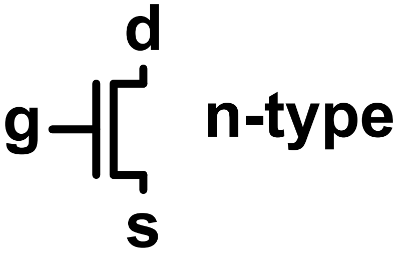
# nMOS Transistor Operation

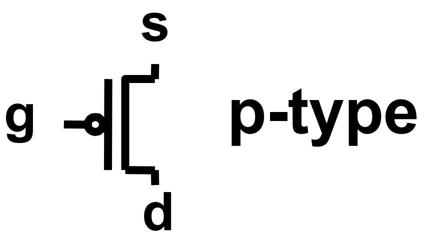




# CMOS – *Complementary* Metal Oxide Semiconductor Technology

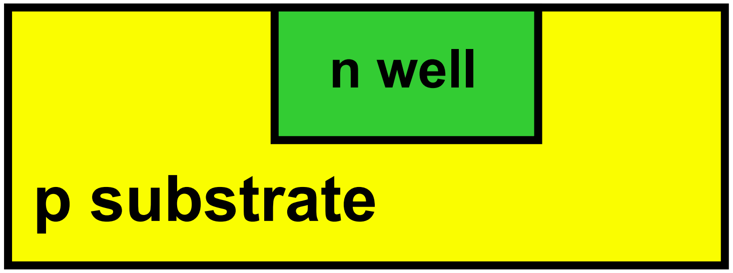
## 2 Distinct Transistor Types





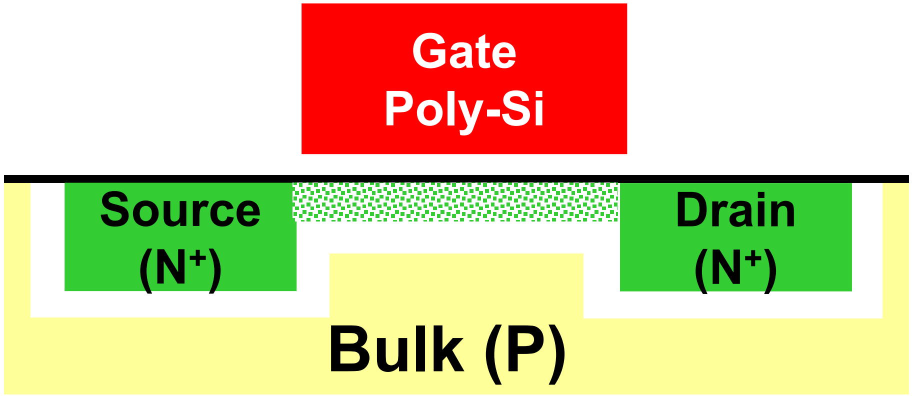
- "on" when $V_g$ is high
- With n-type s/d
- Electrons (n) as carrier
- Built in p-type Si

- "on" when $V_g$ is low
- With p-type s/d
- Holes (p) as carrier
- Built in n-type Si



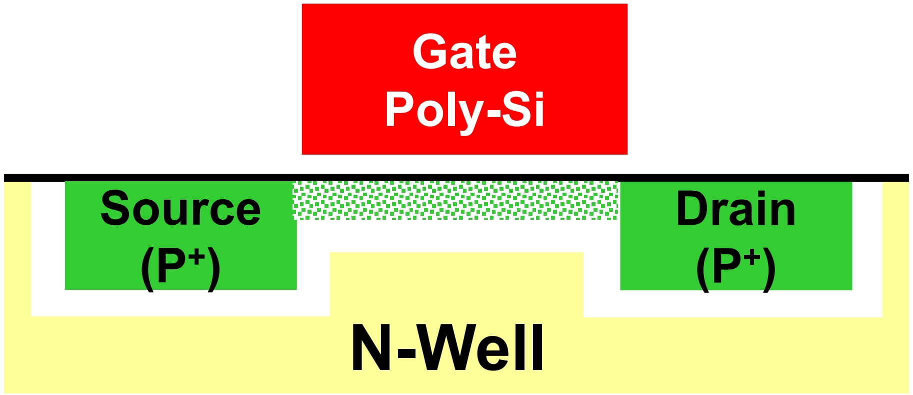
n-well (for PMOS) in p-type substrate (for NMOS)

# NMOS vs PMOS



**NMOS**

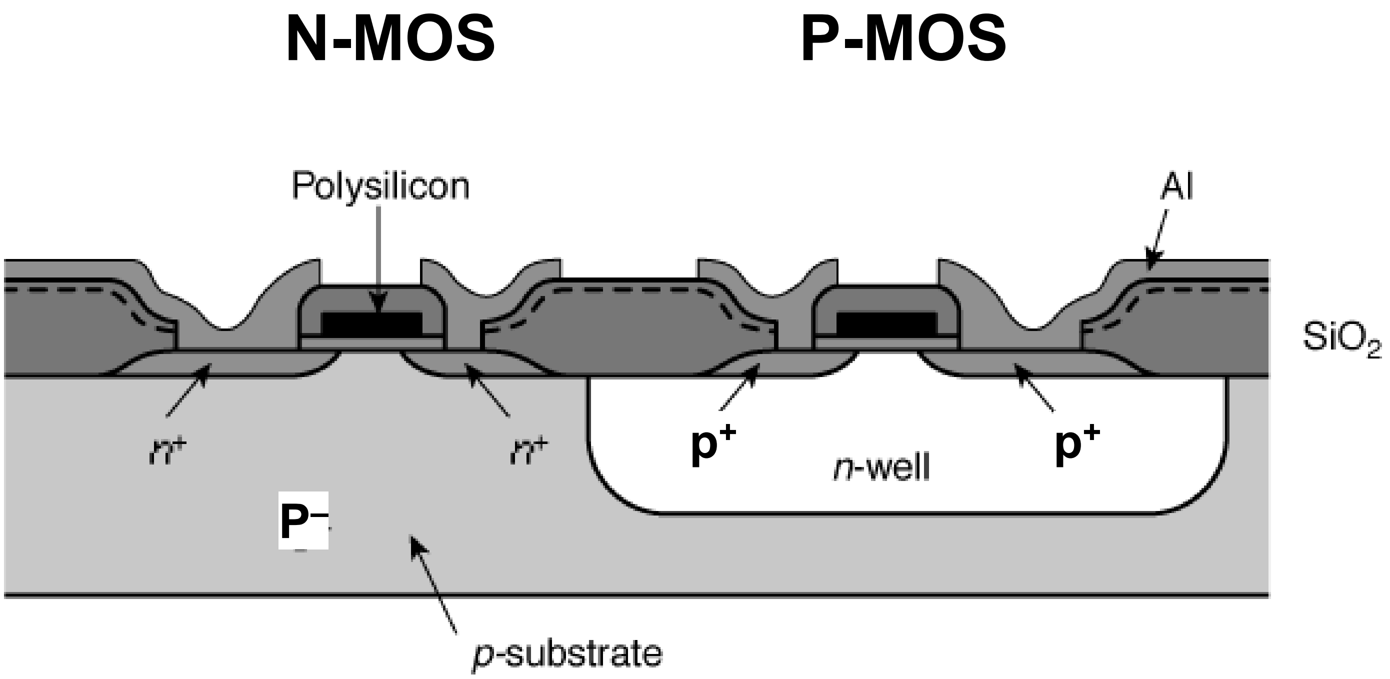
- On when gate voltage is high
- Off when gate voltage is low



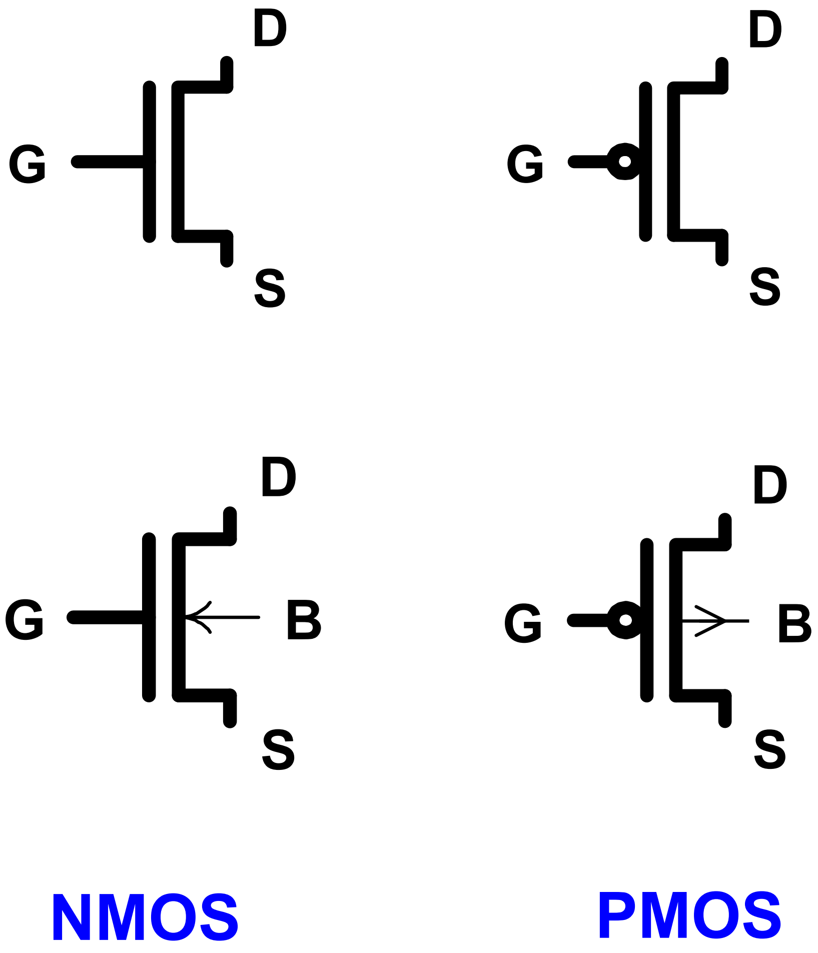
**PMOS**

- N and P regions opposite of NMOS
- On when gate voltage is low
- Off when gate voltage is high

# Cross-Section of CMOS Technology



# MOS Transistors



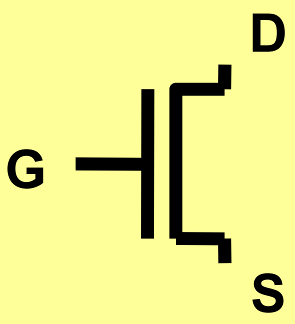
**3-terminal model**

**bulk assumed to be connected to appropriate supply**

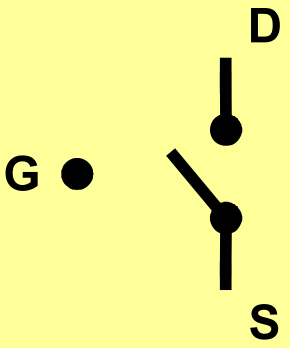
**4-terminal model**

**B = bulk (substrate)**

# MOS Transistor Switch Level Models



NMOS



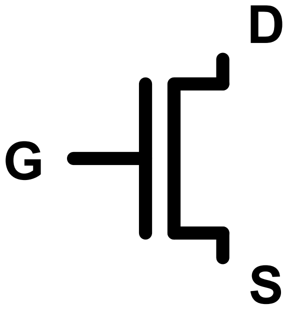
Simplest possible useful model

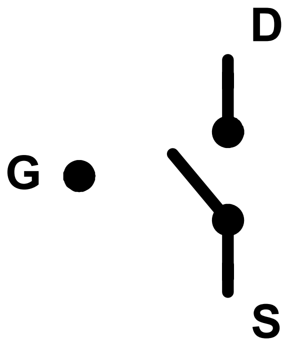
**Position of switch depends on gate voltage (relative to source)**

| $V_{GS}$ | NMOS | PMOS |
|---|---|---|
| $V_{GS}>V_T$ | closed | open |
| $V_{GS}<V_T$ | open | closed |

- **Connection between source and drain depends on gate voltage, current can flow from source to drain and vice versa if closed**
- **No static current flows into gate terminal**

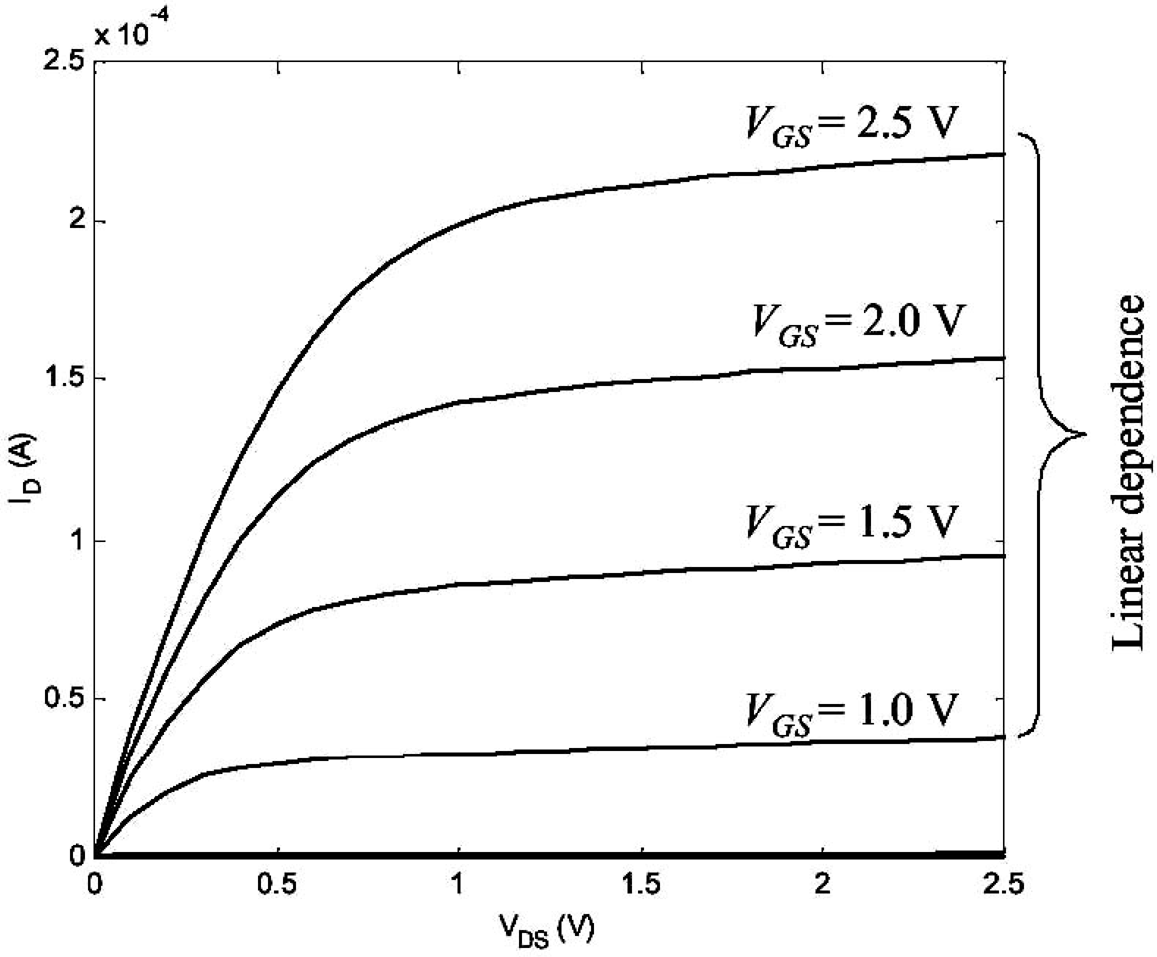
# Is this all there is?





- You don’t believe that (CMOS) life can be so **simple**, do you?
- {TPS} some of the things that you would expect to be **non-idealities** of CMOS as a switch
- Since we want to design CMOS circuits, we need a **deeper understanding** of CMOS circuits

# Drain Current $I_D$

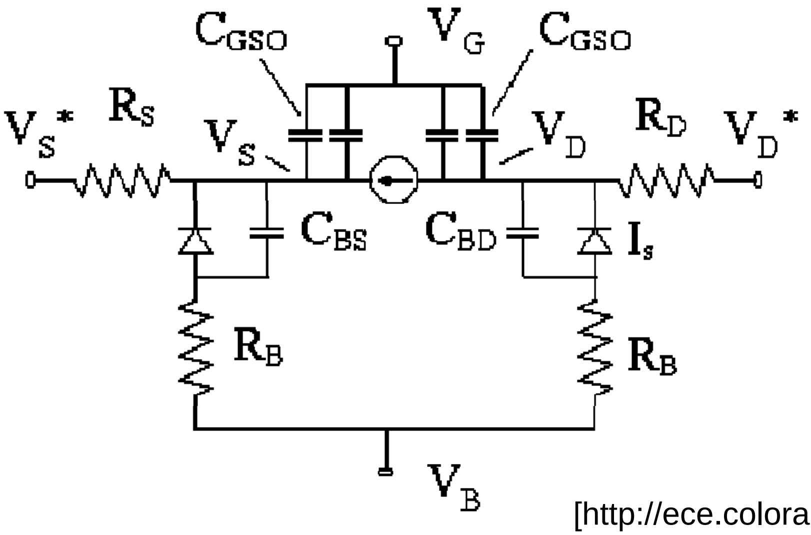


(b) Short-channel transistor ($L_d$ = 0.25 µm)

- You might remember quadratic dependence
- Not true anymore for short-channel devices (velocity saturation)

# SPICE Model

- Some very advanced models to describe MOS devices over all combinations of terminal voltage
  - Includes dynamic behavior, thermal, …
  - BSIM3, BSIM4, PSP, (MEXTRAM), …
  - Can be used for Circuit Simulation
  - >10k lines of C-code
  - See ET4292 – Ramses van der Toorn – Device Modeling



[http://ece.colorado.edu/~bart/book/book/index.html]

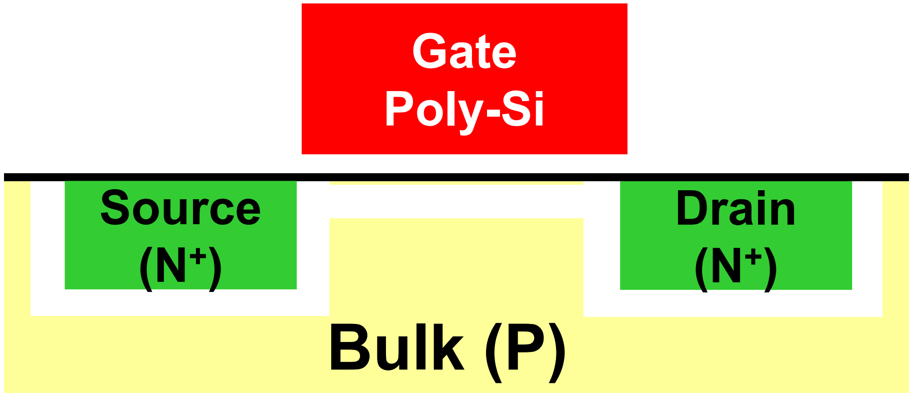
# MOS Operating Regimes

- Off
- Saturation
- Linear – Triode – Resistive
- Velocity Saturation
- (Sub-threshold)

- Different formulas for drain current $I_D$ in each region
- You need to understand these principles

p. 88-106

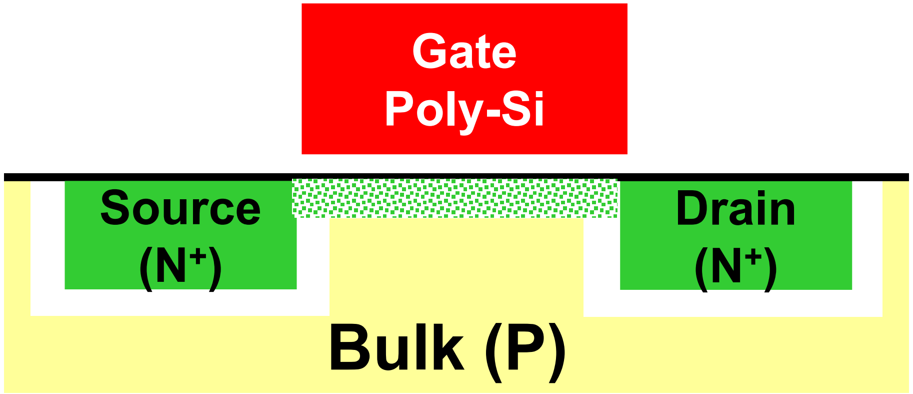
# Off-regime (NMOS)



$V_{GS} < V_T$

- Easiest – current $I_D$ is essentially zero
- When $V_{GS}$ approaches $V_T$, a small current starts to flow (sub-threshold current)
  - Important phenomenon for small, low-voltage devices
  - Important opportunity for ultra-low power circuits

# Triode-regime (NMOS)



$V_{GS} > V_T, V_{GD} > V_T$

$\Leftrightarrow$

$V_{GS} > V_T, V_{DS} < V_{GS}\text{-}V_T$

- **Inversion both at source-side and drain-side of channel**
- **$I_D$ depends on $V_{DS}$: triode behavior**
- **$I_D$ depends on $V_{GS} – V_T$**
- **Inversion not completely symmetric if $V_{DS} > 0$**

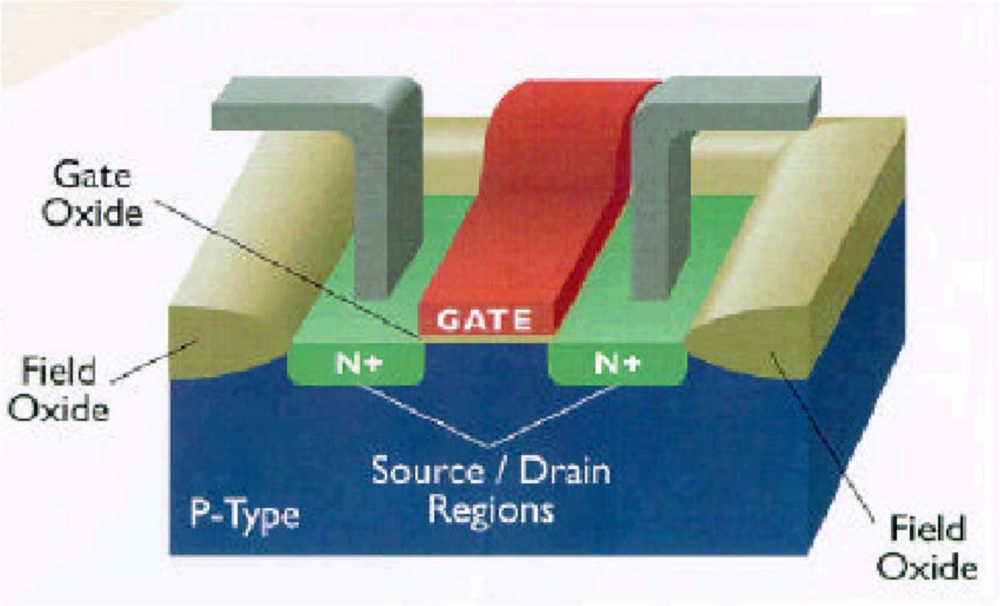
$$I_D = k\left[(V_{GS} - V_T)V_{DS} - \frac{1}{2}V_{DS}^2\right]$$

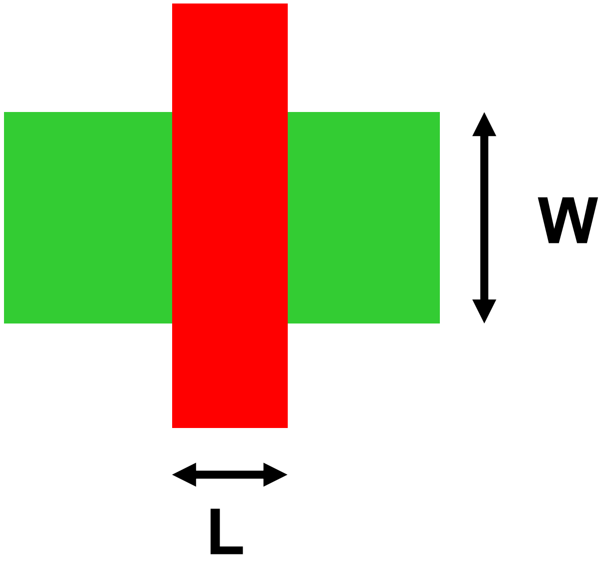
***k*, *V*$_T$: device parameters**

# Device Sizing

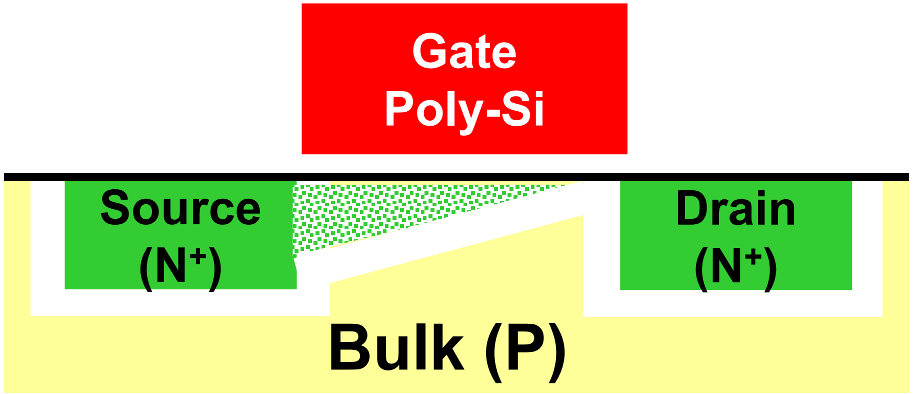
$$I_D = k\left[(V_{GS} - V_T)V_{DS} - \frac{1}{2}V_{DS}^2\right] \qquad k = k'\frac{W}{L}$$

| | |
|---|---|
| *k* | **Device transconductance** |
| *k'* | **Process transconductance** |
| *W* | **Device width** |
| *L* | **Device length** |





# Saturation (NMOS)



$V_{GS} > V_T, V_{GD} < V_T$

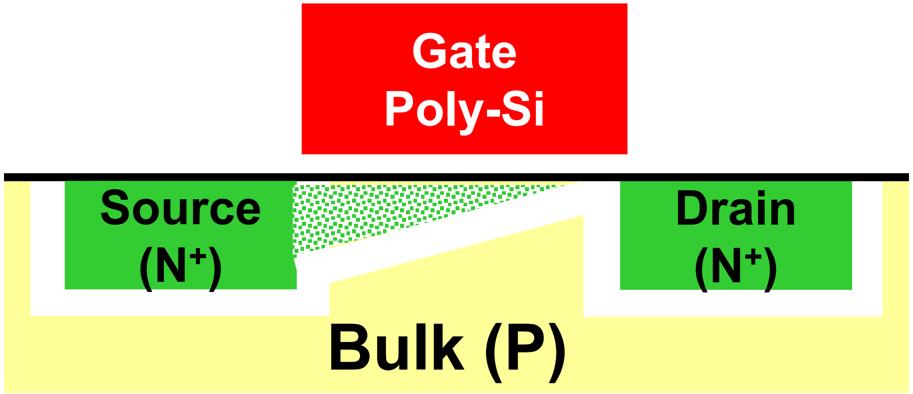
$\Leftrightarrow$

$V_{GS} > V_T, V_{DS} > V_{GS}\text{-}V_T$

- Inversion only at source-side of channel
- Still, current will be flowing between S and D
- Current does not (strongly) depend on $V_{DS}$: current source behavior

$$I_D = \frac{1}{2}k\left(V_{GS} - V_T\right)^2$$

# Velocity Saturation



$V_{DS} > V_{DSAT}$

- When $V_{DS}$ large enough – current doesn't increase anymore since carrier velocity is limited by scattering to lattice
- Visible at 250 nm and below
- Can happen in (otherwise) saturation conditions, but also in triode conditions

$$I_D = k\left[(V_{GS} - V_T)V_{DSAT} - \frac{1}{2}V_{DSAT}^2\right]$$

(*k*, $V_T$,$V_{DSAT}$: device data)

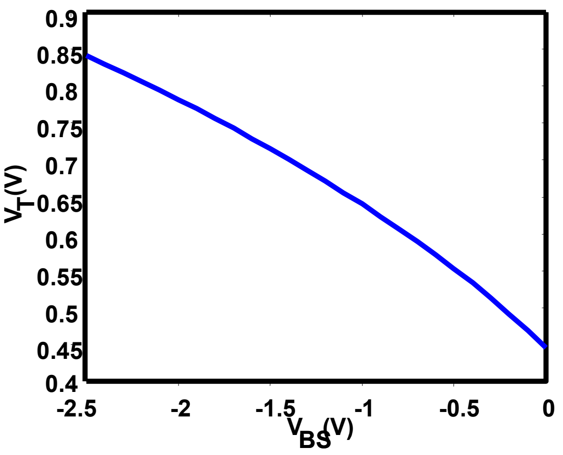
# More Effects

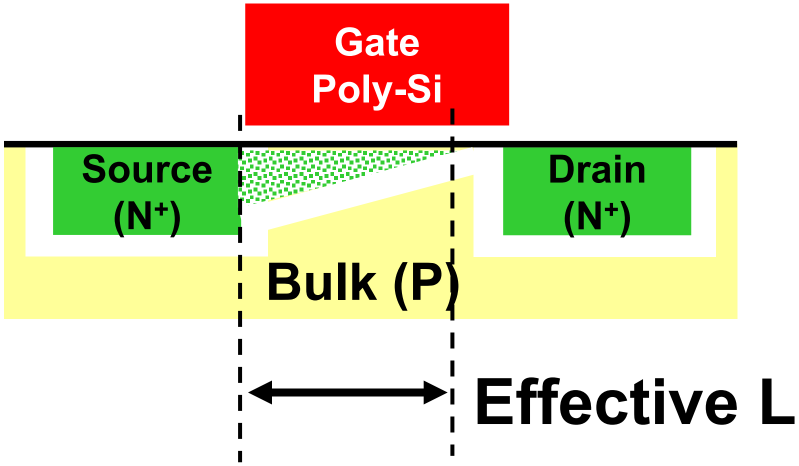
- **Body Effect: $V_T$ depends on $V_{SB}$**

$$V_T = V_{T0} + \gamma\left(\sqrt{\left|-2\phi_F + V_{SB}\right|} - \sqrt{\left|-2\phi_F\right|}\right)$$

- **Channel length Modulation:**
  **$I_D$ depends on $V_{DS}$ also in saturation**

$$I_D = I_D \times \left(1 + \lambda V_{DS}\right)$$





# Summary

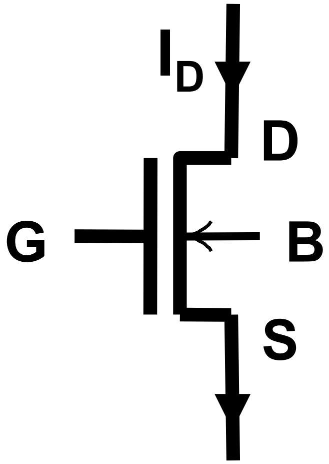
$$I_D = \begin{cases} k\left[(V_{GS} - V_T)V_{DS} - \frac{1}{2}V_{DS}^2\right] & V_{GS} > V_T, V_{DS} < V_{GS}\text{-}V_T \\ \frac{1}{2}k(V_{GS} - V_T)^2 & V_{GS} > V_T, V_{DS} > V_{GS}\text{-}V_T \\ k\left[(V_{GS} - V_T)V_{DSAT} - \frac{1}{2}V_{DSAT}^2\right] & V_{DS} > V_{DSAT} \end{cases}$$

$$I_D = I_D \times (1 + \lambda V_{DS})$$

$$V_T = V_{T0} + \gamma\left(\sqrt{|-2\phi_F + V_{SB}|} - \sqrt{|-2\phi_F|}\right)$$

**Next slide presents alternative formulation**

# MOS Models for Manual Analysis



**determined by circuit**

$V_{DS}, V_{GS}, V_{SB}$

**determined by designer**

$W, L$

**determined by technology**

$k', \lambda, V_{DSAT}, V_{T0}, \gamma, \phi_F$

## MOS model for manual analysis

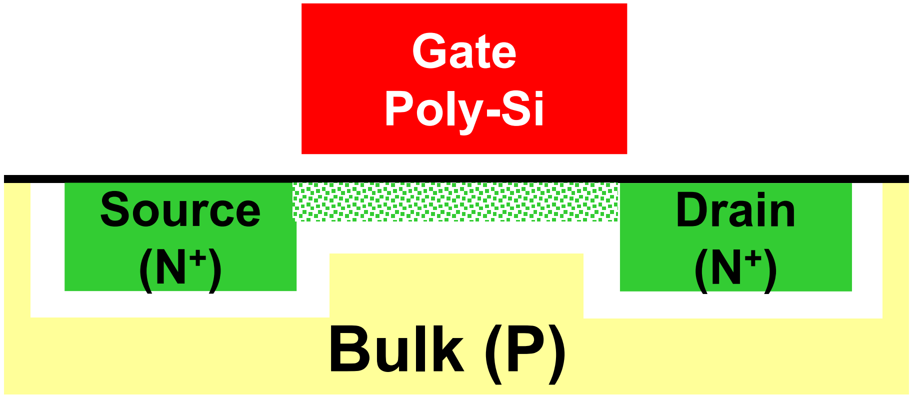
$$I_D = k\left(V_{GT}V_{MIN} - 0.5V_{MIN}^2\right)\left(1 + \lambda V_{DS}\right) \qquad \text{for } V_{GT} \geq 0$$

$$= 0 \qquad \text{for } V_{GT} \leq 0$$

$$V_{MIN} = MIN\left(V_{DS}, V_{GT}, V_{DSAT}\right) \qquad k = k'\frac{W}{L}$$

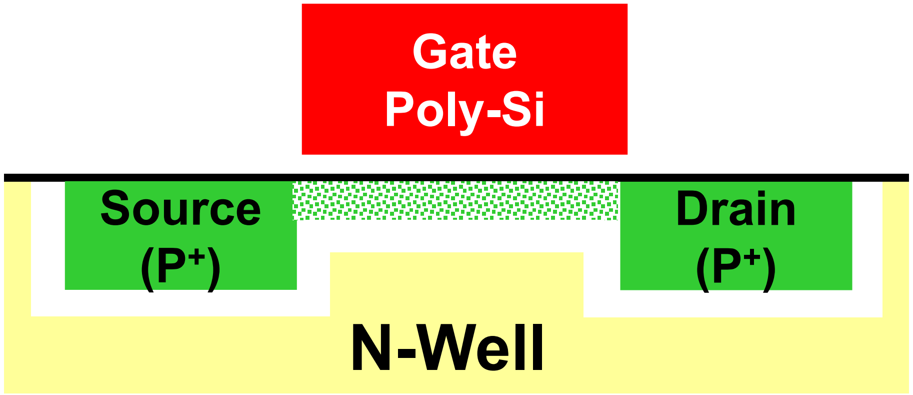
$$V_{GT} = V_{GS} - V_T, \qquad V_T = V_{T0} + \gamma\left(\sqrt{\left|-2\phi_F + V_{SB}\right|} - \sqrt{\left|-2\phi_F\right|}\right)$$

# NMOS vs PMOS



**NMOS**

- On when gate voltage is high
- Off when gate voltage is low

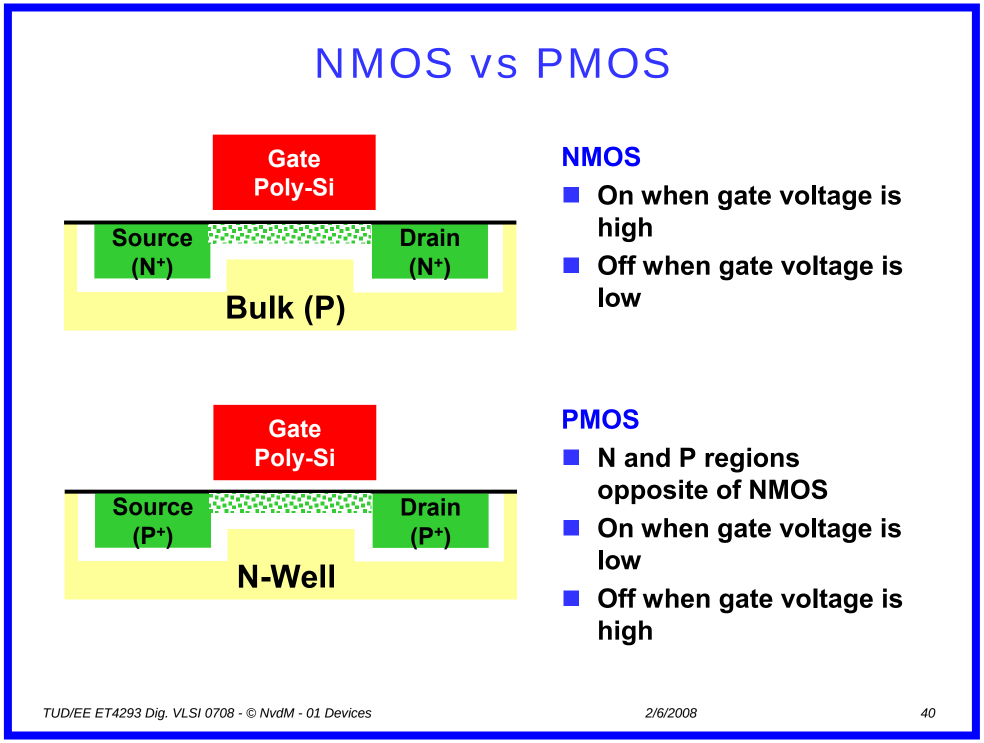


**PMOS**

- N and P regions opposite of NMOS
- On when gate voltage is low
- Off when gate voltage is high

# NMOS vs PMOS

- **PMOS: all polarities reversed for same operation region**
- **Hole-mobility < Electron-mobility**
- **$|k'_{PMOS}| < k'_{NMOS}$**



## NMOS vs PMOS

**NMOS**

- On when gate voltage is high
- Off when gate voltage is low

**PMOS**

- N and P regions opposite of NMOS
- On when gate voltage is low
- Off when gate voltage is high

*TUD/EE ET4293 Dig. VLSI 0708 - © NvdM - 01 Devices* *2/6/2008* *40*

## NMOS Regions of Operation

Triode ⇔ $V_{gs} > V_t$ and $V_{ds} < V_{gs} - V_t$ and $V_{ds} < V_{DSAT}$

Saturation ⇔ $V_{gs} > V_t$ and $V_{ds} > V_{gs} - V_t$ and $V_{ds} < V_{DSAT}$

Vel. sat ⇔ $V_{gs} > V_t$ and $V_{ds} > V_{DSAT}$

Cut-off ⇔ $V_{gs} \leq V_t$

## PMOS Regions of Operation

Triode ⇔ $V_{gs} < V_t$ and $V_{ds} > V_{gs} - V_t$ and $V_{ds} > V_{DSAT}$

Saturation ⇔ $V_{gs} < V_t$ and $V_{ds} < V_{gs} - V_t$ and $V_{ds} > V_{DSAT}$

Vel. sat ⇔ $V_{gs} < V_t$ and $V_{ds} < V_{DSAT}$

Cut-off ⇔ $V_{gs} \geq V_t$

## Universal

Triode ⇔ $|V_{gs}| > |V_t|$ and $|V_{ds}| < |V_{gs}| - |V_t|$ and $|V_{ds}| < |V_{DSAT}|$

Saturation ⇔ $|V_{gs}| > |V_t|$ and $|V_{ds}| > |V_{gs}| - |V_t|$ and $|V_{ds}| < |V_{DSAT}|$

Vel. sat ⇔ $|V_{gs}| > |V_t|$ and $|V_{ds}| > |V_{DSAT}|$

Cut-off ⇔ $|V_{gs}| \leq |V_{t|}$

# Unified Model Parameters

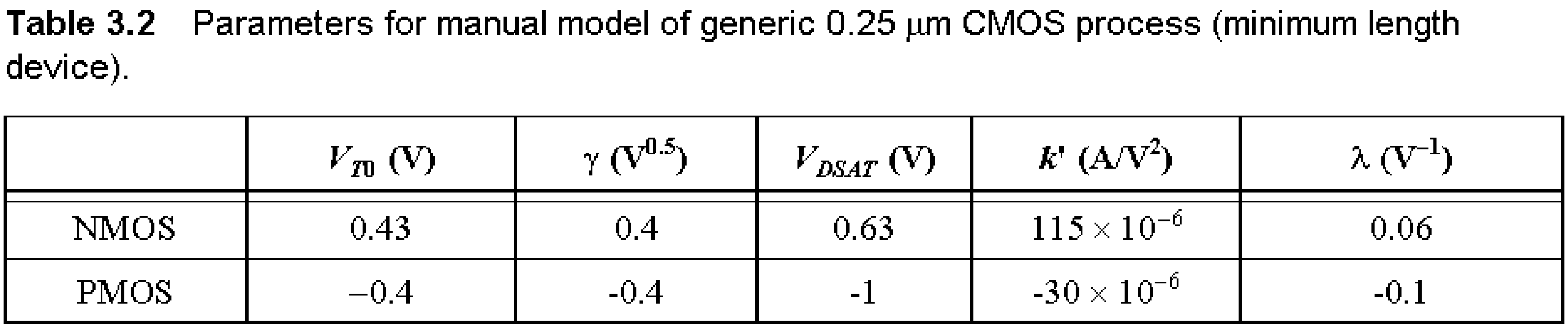


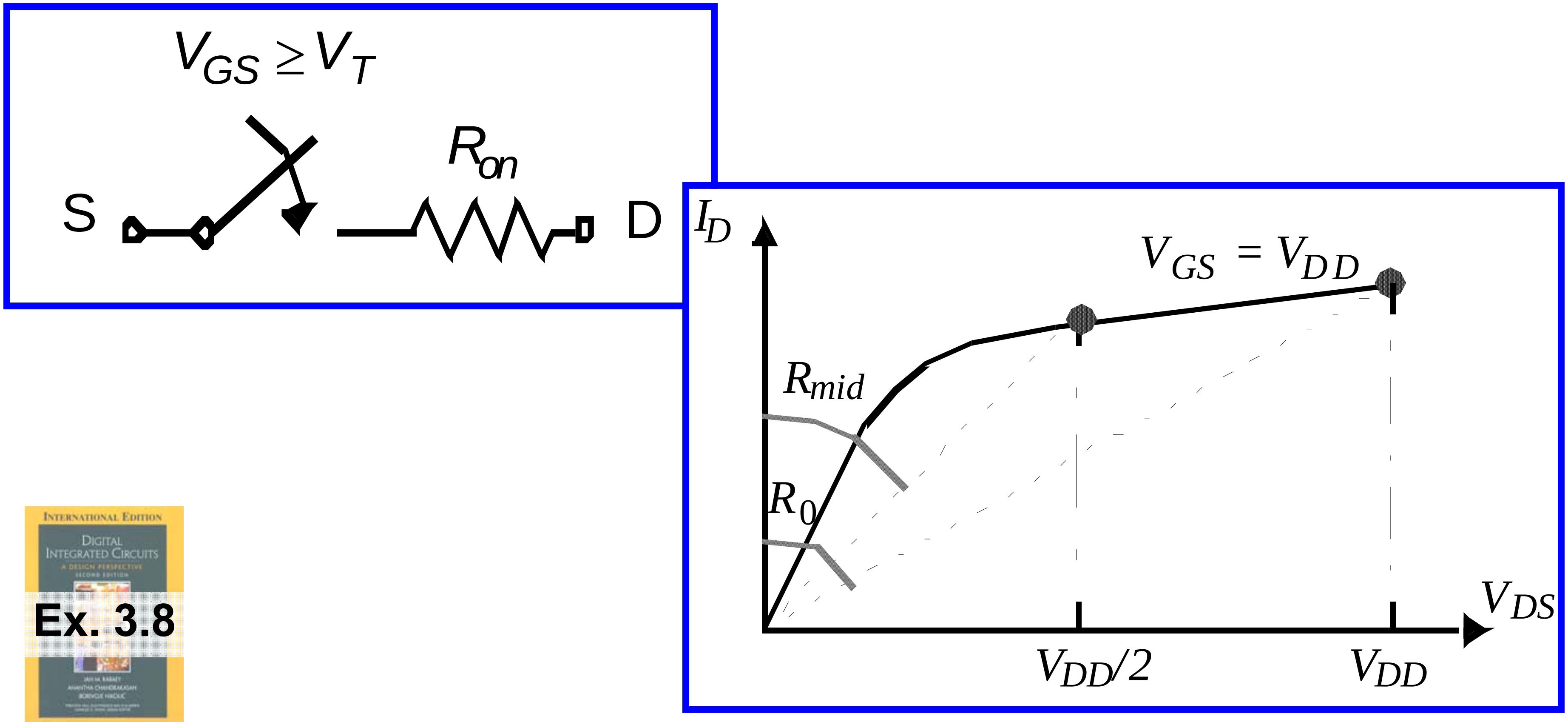
**Table 3.2** Parameters for manual model of generic 0.25 μm CMOS process (minimum length device).

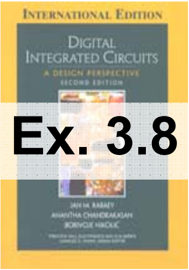
| | $V_{T0}$ (V) | $\gamma$ ($V^{0.5}$) | $V_{DSAT}$ (V) | $k'$ ($A/V^2$) | $\lambda$ ($V^{-1}$) |
|---|---|---|---|---|---|
| NMOS | 0.43 | 0.4 | 0.63 | $115 \times 10^{-6}$ | 0.06 |
| PMOS | –0.4 | -0.4 | -1 | $-30 \times 10^{-6}$ | -0.1 |

- **Parameters depend on technology**
- **See tables in front and back cover of book**

- **Modern processes offer various threshold voltages**
- **{TPS} Why?**

# The Transistor as a Switch

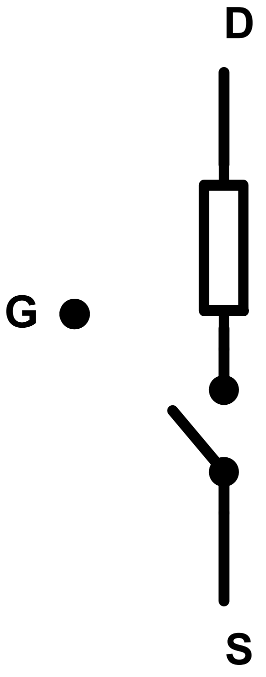




Ex. 3.8

$$R_{eq} = \frac{1}{2}\left(\frac{V_{DD}}{I_{DSAT}(1+\lambda V_{DD})} + \frac{V_{DD}/2}{I_{DSAT}(1+\lambda V_{DD}/2)}\right) \approx \frac{3}{4}\frac{V_{DD}}{I_{DSAT}}\left(1 - \frac{5}{6}\lambda V_{DD}\right)$$

# MOS Transistor
# Switch Level Model (Empirical).



$$R_{on} = \frac{L}{W} R_{eq}$$

**Position of switch depends on gate to source voltage**

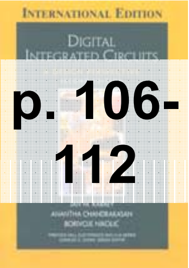
| $V_{GS}$ | NMOS | PMOS |
|---|---|---|
| hi | closed | open |
| lo | open | closed |

## $R_{eq}$: Practice!

| $R_{eq}$ \ $V_{dd}$ (V) | 1 | 1.5 | 2 | 2.5 |
|---|---|---|---|---|
| NMOS (kΩ) | 35 | 19 | 15 | 13 |
| PMOS (kΩ) | 115 | 55 | 38 | 31 |

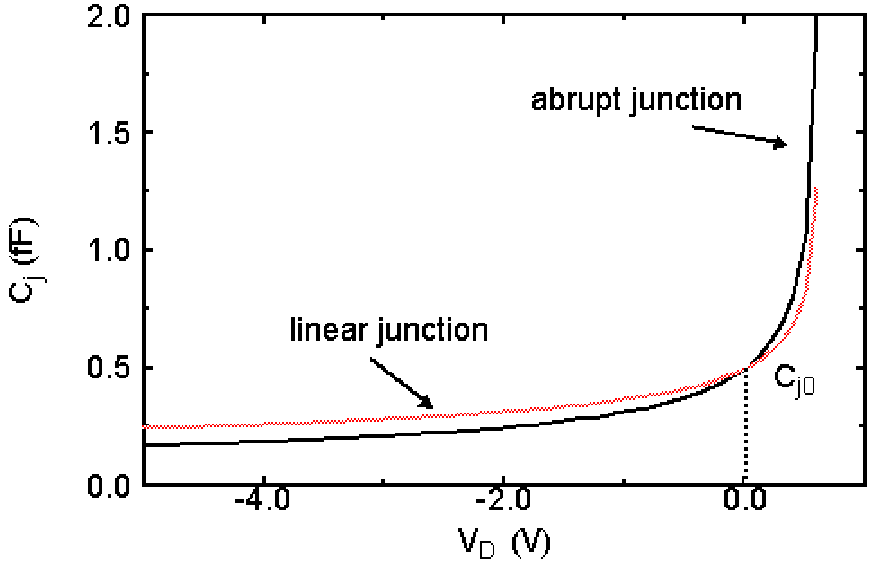
# Dynamic Behavior

- (Solely) governed by **time needed to (dis)charge** (intrinsic, parasitic) **capacitances** associated with device and interconnect
- **Essential knowledge** for designing high-quality ckts.
- Many caps are **non-linear**
- **Spice models** can accurately take C into account
- Need **simplification** and insight for *design*
  - {TPS} How?
  - Linearization
  - Lumping/merging



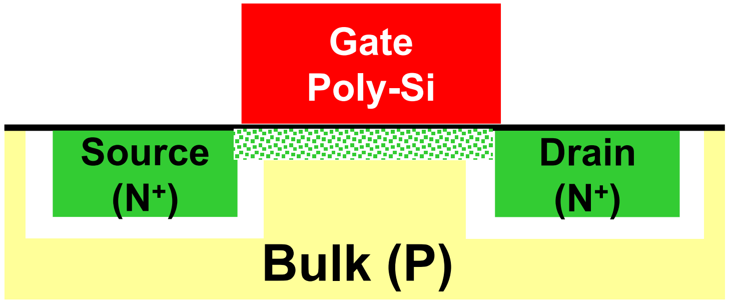
p. 106-112

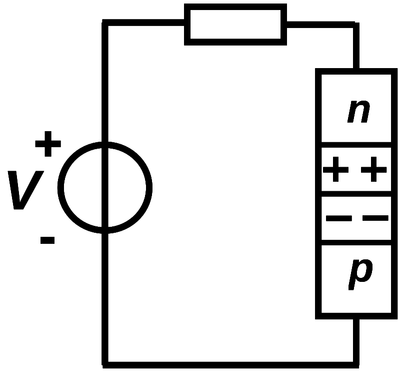
# Junction Capacitance



$$C_j = \frac{C_{j0}}{(1 - V_D / \phi_0)^m}$$

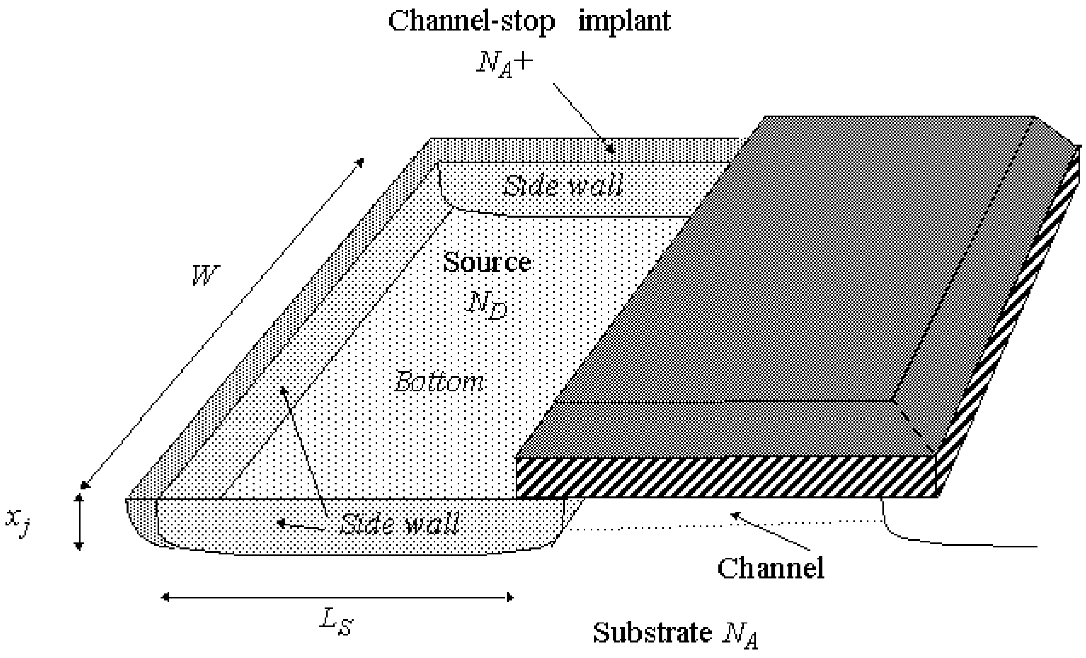
m = 0.5: abrupt junction
m = 0.33: linear junction





- Space-charge / depletion region creates electric field
- Electric field energy works like capacitor (energy is ½CV²)
- Width of depletion region depends on voltage: non-linear C

# MOS S/D Capacitance



$$C_{diff} = C_{bottom} + C_{sidewall}$$

$$= C_j \times AREA + C_{jsw} \times PERIMETER$$

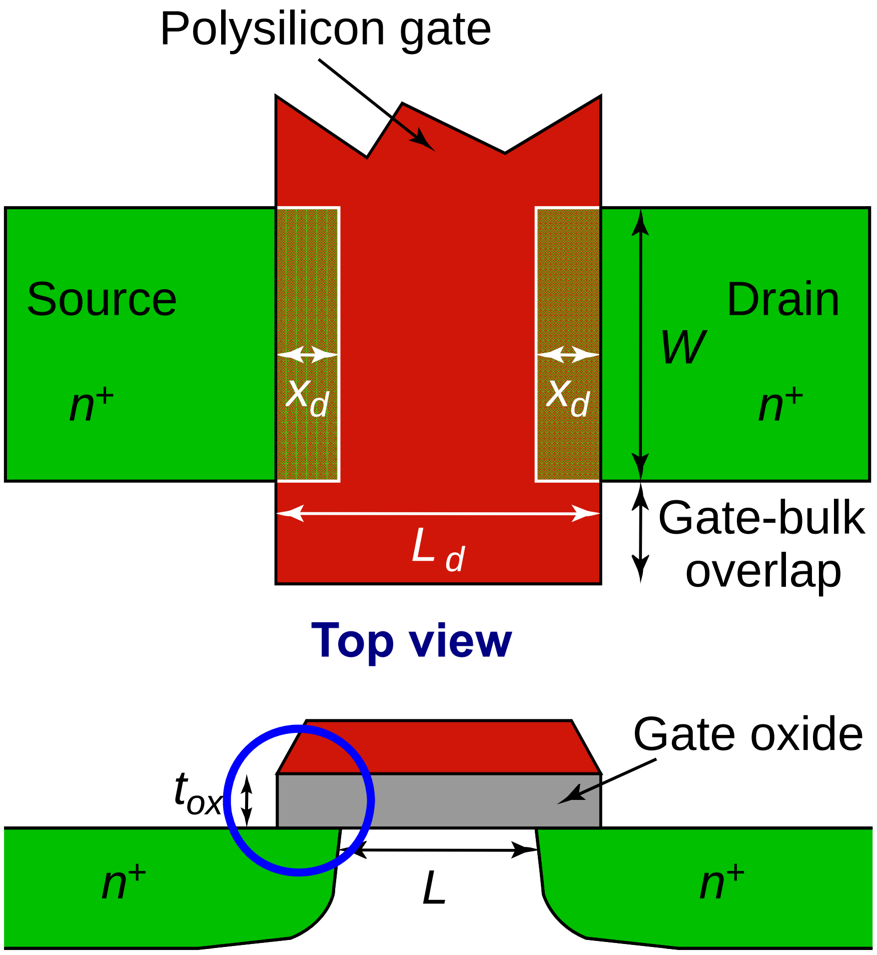
$$= C_j L_s W + C_{jsw} (2L_s + W)$$

**Diode capacitances ⇒ use linearized values**

# Gate Capacitance

- **Gate-to-channel**
  - **Complex function of operating voltages**
  - **Because channel charge depends on voltages**
- **Gate-source and gate-drain overlap capacitance**
  - **Just (almost) linear capacitances depending only on geometry**

# Gate-Overlap Capacitance

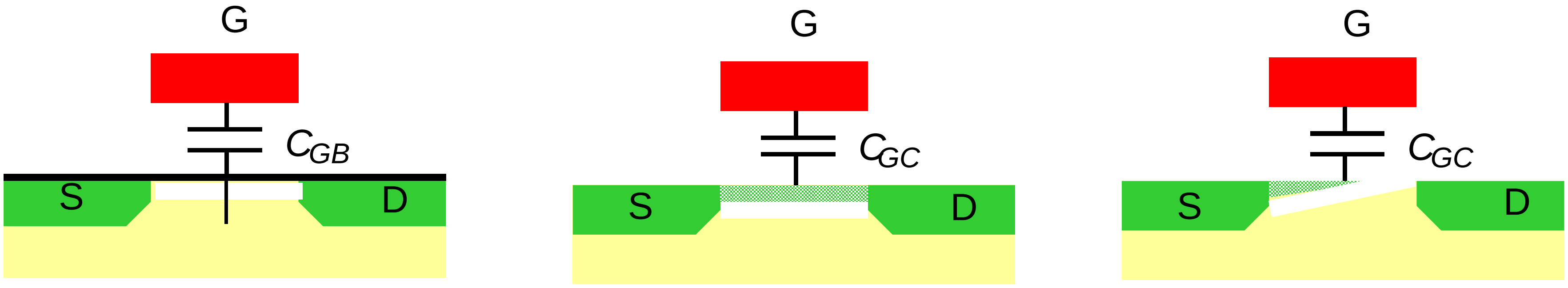


**Top view**

**Cross section**

$$C_{gate} = \frac{\varepsilon_{ox}}{t_{ox}} WL$$

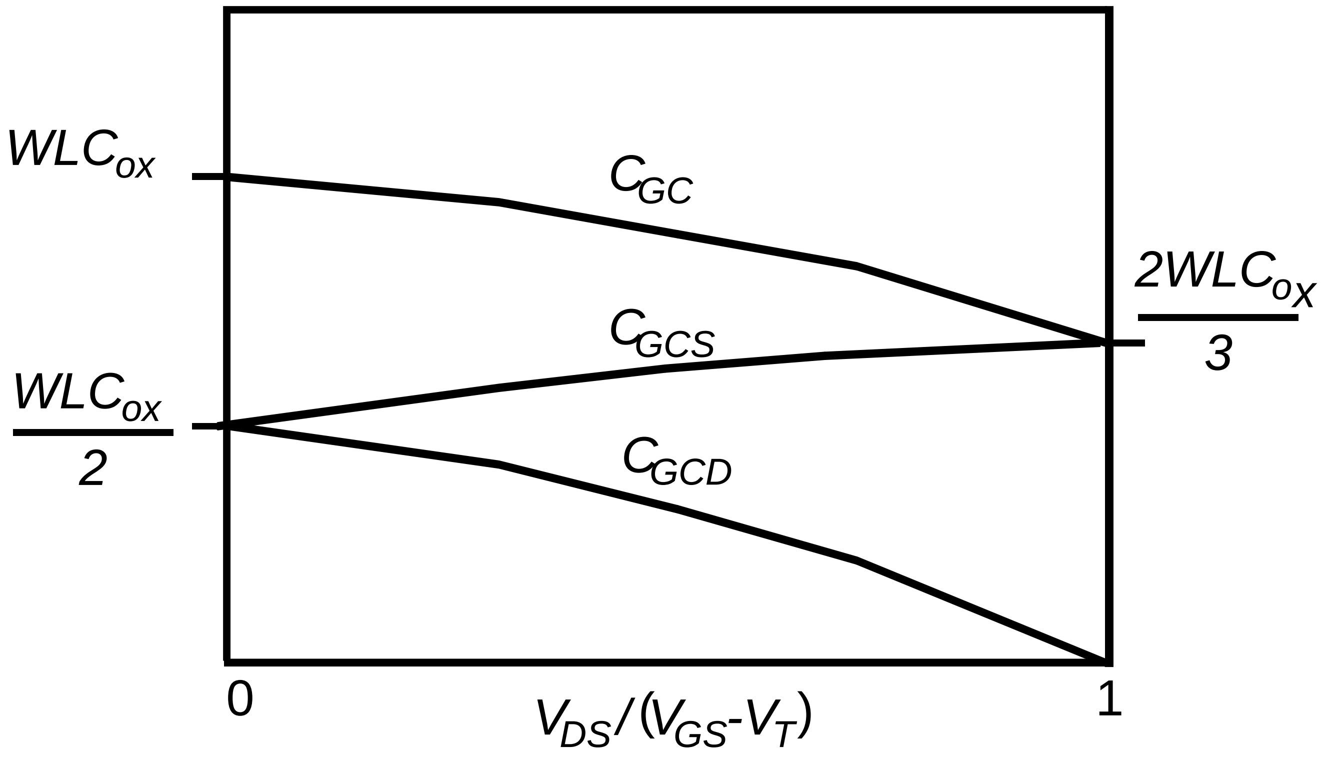
# Gate-Channel Capacitance



| Operation Region | $C_{gb}$ | $C_{gs}$ | $C_{gd}$ |
|---|---|---|---|
| Cutoff | $C_{ox}WL_{eff}$ | 0 | 0 |
| Triode | 0 | $C_{ox}WL_{eff}/2$ | $C_{ox}WL_{eff}/2$ |
| Saturation | 0 | $(2/3)C_{ox}WL_{eff}$ | 0 |

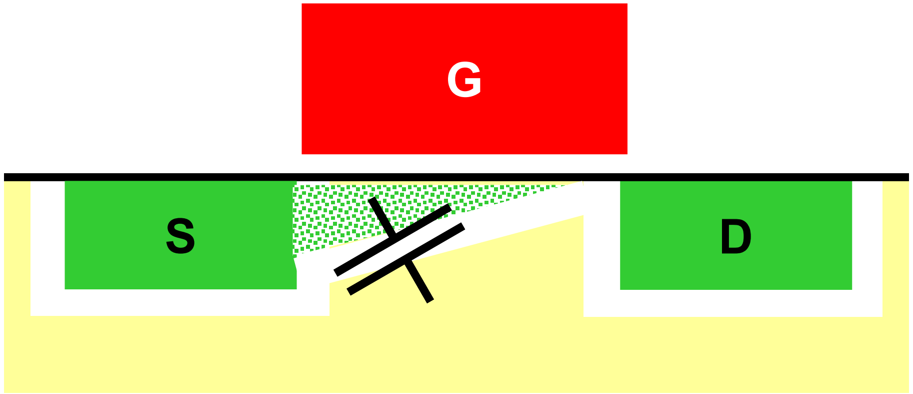
**Most important regions in digital design: saturation and cut-off**

# Gate-Channel Capacitance



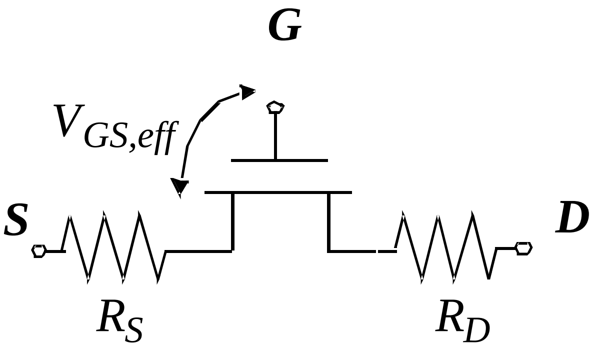
**Degree of saturation**

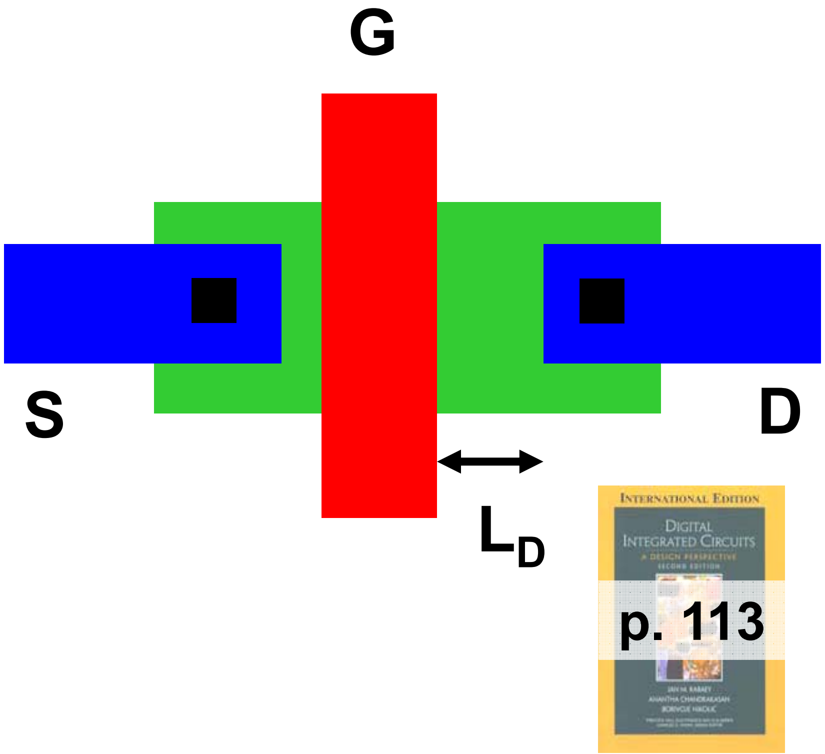
# Channel-Bulk Capacitance



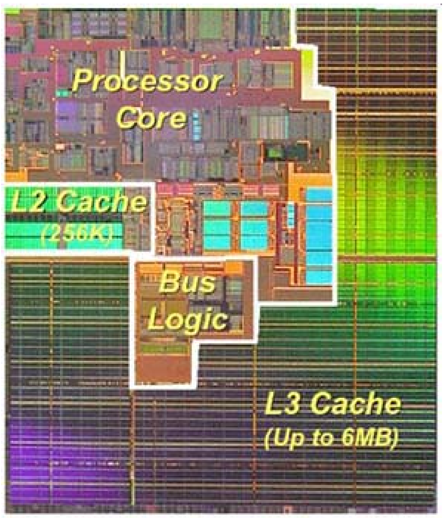
- Channel-bulk cap $C_{CB}$
- Only when transistor is on
- Parallel to $C_{SB}$

# Device Parasitic Resistance





# Technology Scaling



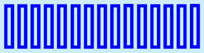
McKinley

4004

| Processor | 4004 | Montecito | Bloomfield |
|---|---|---|---|
| Year | 1971 | 2005 | 2008 |
| Feat. size | 10μm | 90nm | 45nm |
| Die size | 12mm$^2$ | 596mm$^2$ | 263mm$^2$ |
| Transistors | 2300 | 1.7x10$^9$<br>1550M for 24 MB L3 | 0.731x10$^9$ |
| Clock | 108 kHz | 1.8GHz | 3.3GHz |
| Perform.<br>(spec2000) | 0.01 | ~1600 | |

4004 →

Comparative interconnect dimensions

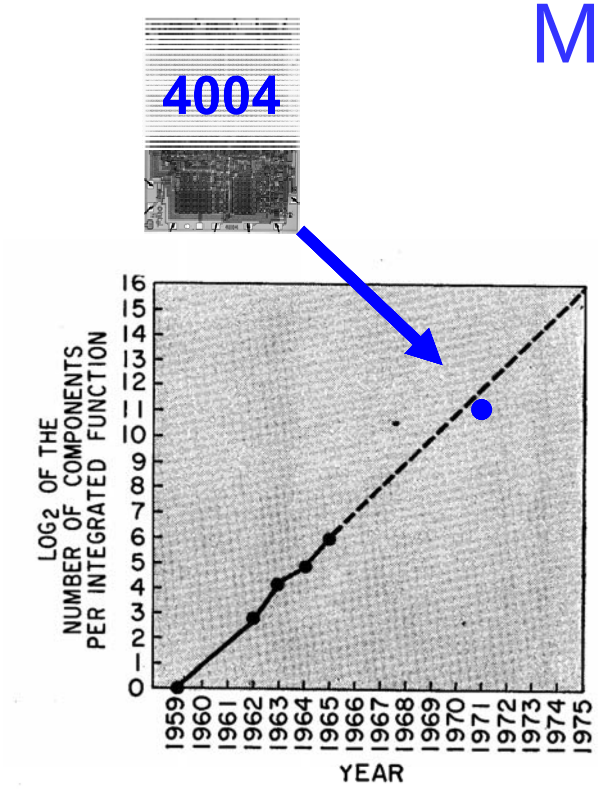


← 15 lines @ 45 nm

International Edition
Digital Integrated Circuits
A Design Perspective
Second Edition

§3.5

# Moore’s Law

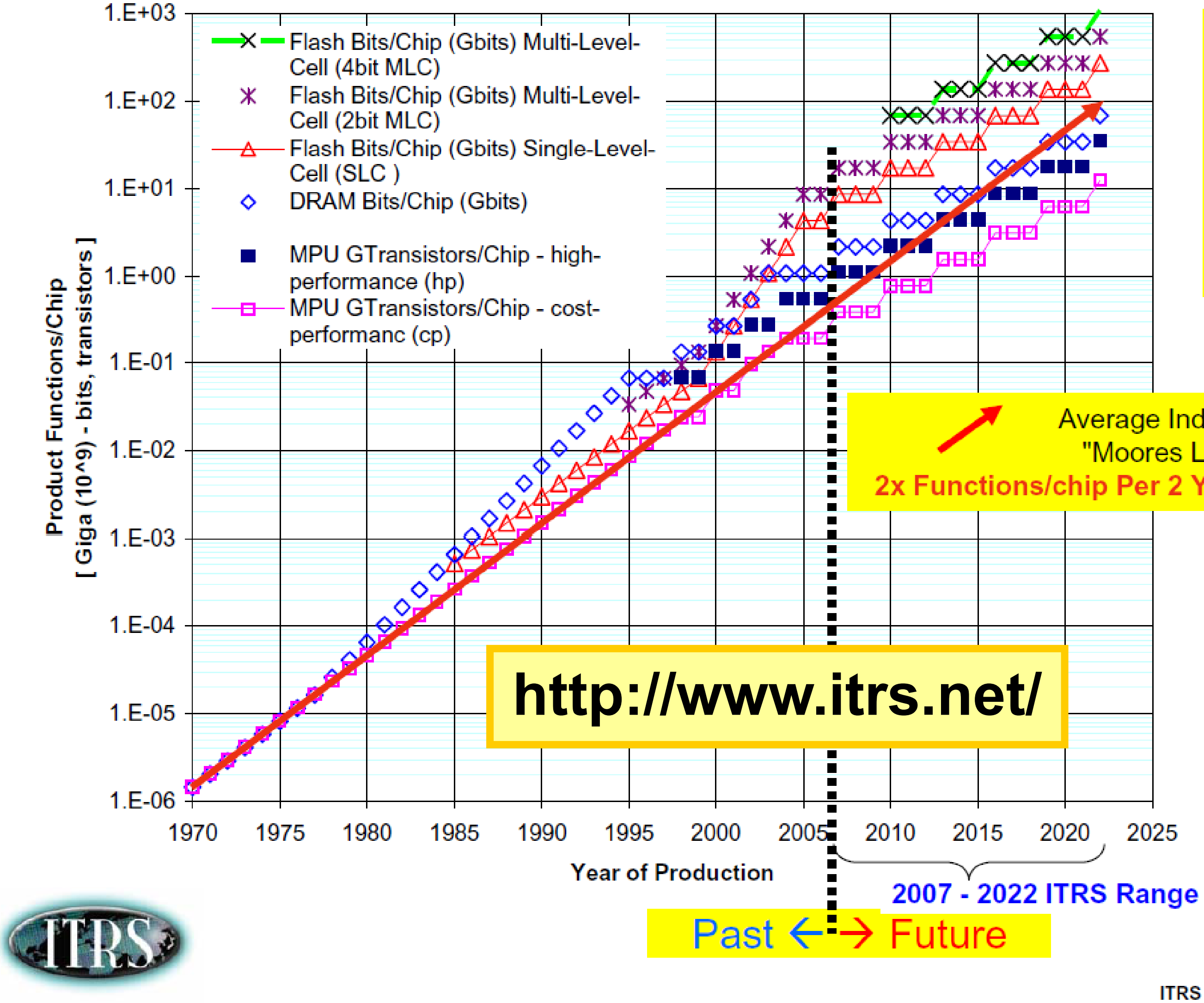


**The number of transistors that can be integrated on a single chip will double every 18 months**

**Gordon Moore, co-founder of Intel [Electronics, Vol 38, No. 8, 1965]**

# Chip Size Trends – 2007 ITRS Functions/Chip Model

## 2007 ITRS Product Technology Trends - Functions per Chip (@Volume Production, Affordable Chip Size**)



Average Industry "Moores Law" : 2x Functions/chip Per 2 Years

**http://www.itrs.net/**

Past ← → Future

** Affordable Production Chip Size Targets:
DRAM, Flash < 145mm$^2$
hp MPU < 310mm$^2$
cp MPU < 140mm$^2$

MPU ahead or = "Moore's Law" 2x Xstors/chip Per 2 years Thru 2010

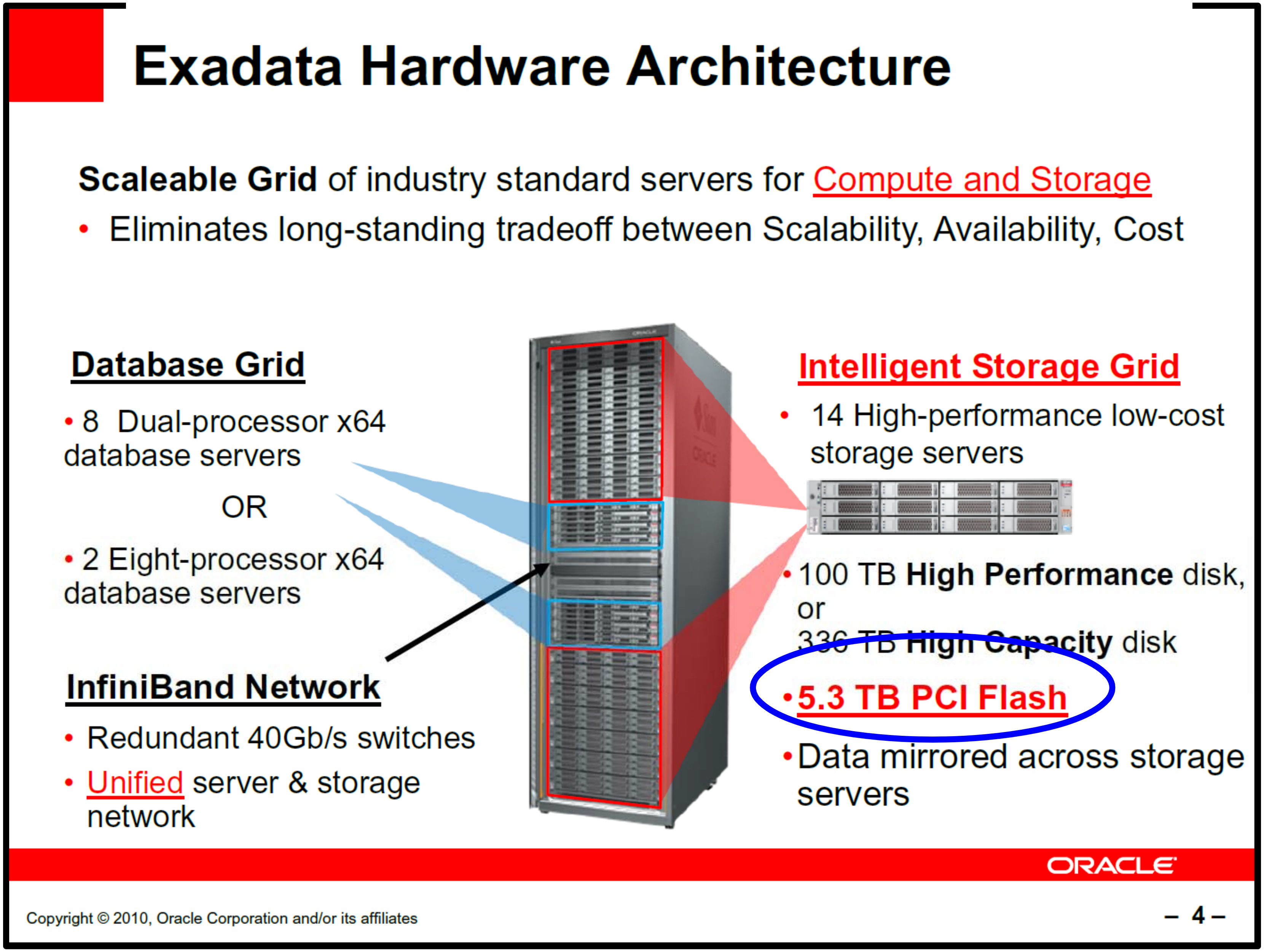
** Example Chip Size Targets:
1.1Gt P07h MPU
@ intro in 2004/620mm$^2$
@ prod in 2007/310mm$^2$

** Example Chip Size Targets:
0.39Gt P07c MPU
@ intro in 2004/280mm$^2$
@ prod in 2007/140mm$^2$

ITRS Winter Conference 2007 Makuhara, Japan

# Flash Memory

# Oracle Database Server (source: Oracle)



## Exadata Hardware Architecture

**Scaleable Grid** of industry standard servers for Compute and Storage

- Eliminates long-standing tradeoff between Scalability, Availability, Cost

**Database Grid**

- 8 Dual-processor x64 database servers

OR

- 2 Eight-processor x64 database servers

**InfiniBand Network**

- Redundant 40Gb/s switches
- Unified server & storage network

**Intelligent Storage Grid**

- 14 High-performance low-cost storage servers
- 100 TB **High Performance** disk, or 336 TB **High Capacity** disk
- **5.3 TB PCI Flash**
- Data mirrored across storage servers

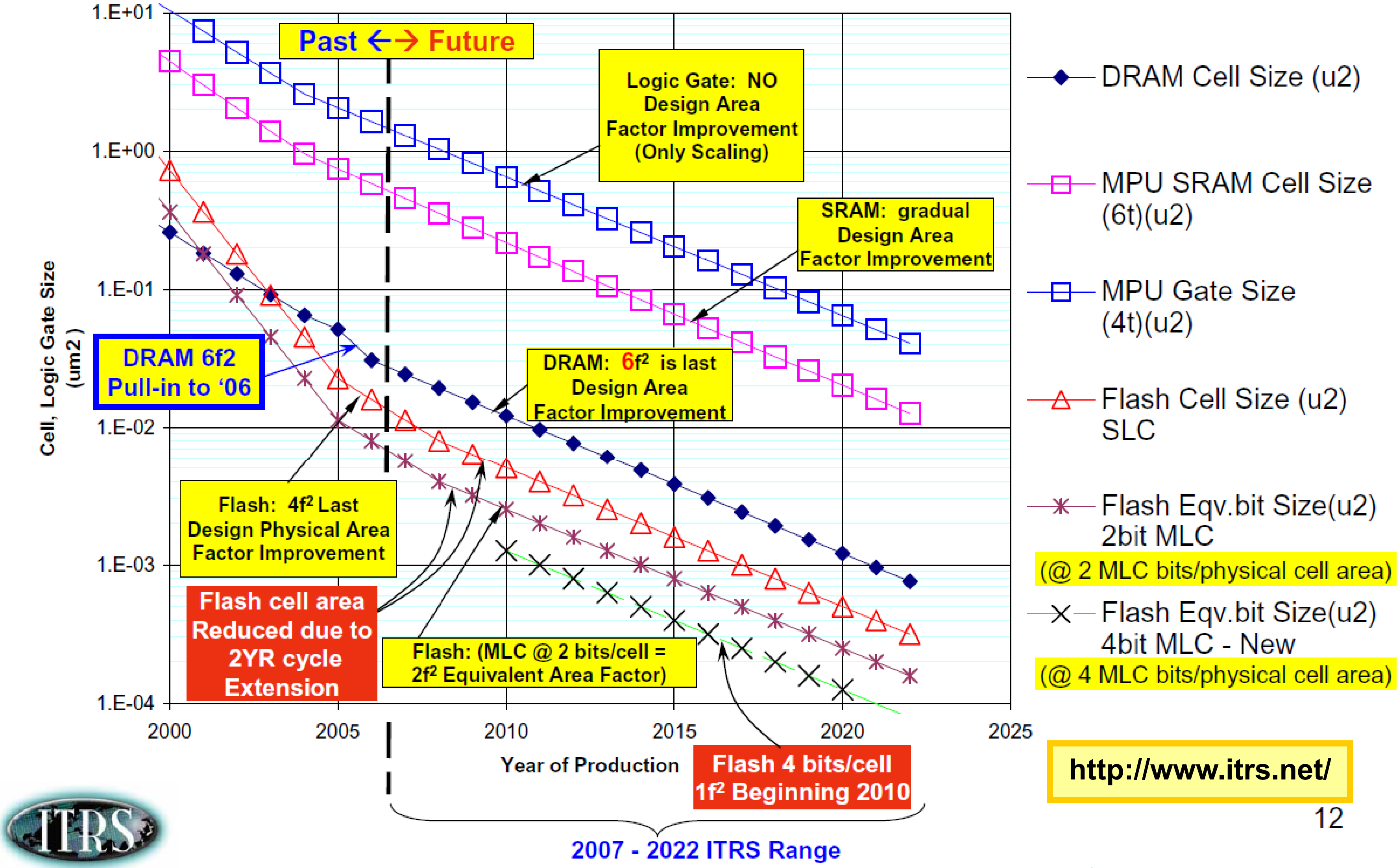
ORACLE

– 4 –

# Figure 9 ITRS Product Function Size

**[changes to DRAM and Flash; plus extend all to 2022]**

**2007 ITRS Product Function Size Trends - Cell Size, Logic Gate(4t) Size**



Past ←→ Future

Logic Gate: NO Design Area Factor Improvement (Only Scaling)

SRAM: gradual Design Area Factor Improvement

DRAM 6f2 Pull-in to '06

DRAM: 6f² is last Design Area Factor Improvement

Flash: 4f² Last Design Physical Area Factor Improvement

Flash cell area Reduced due to 2YR cycle Extension

Flash: (MLC @ 2 bits/cell = 2f² Equivalent Area Factor)

Flash 4 bits/cell 1f² Beginning 2010

(@ 2 MLC bits/physical cell area)

(@ 4 MLC bits/physical cell area)

2007 - 2022 ITRS Range

**http://www.itrs.net/**

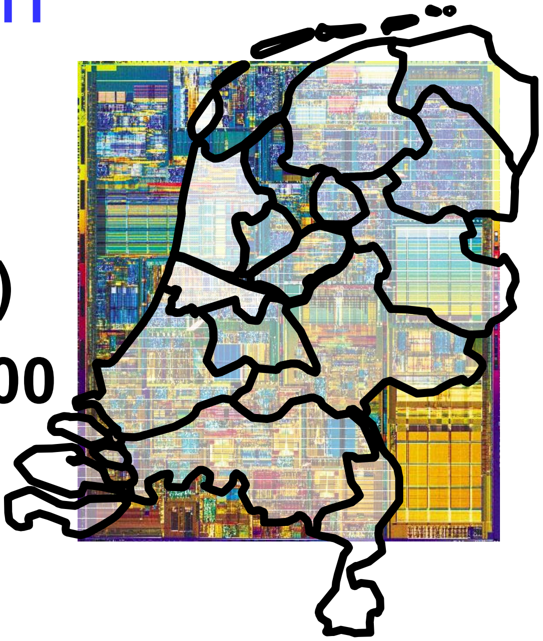
ITRS Winter Conference 2007 Makuhara, Japan

# IC Technology—Comparison

**Chip: 4 $cm^2$**

**Netherlands: 40,000 $km^2$ (approximately)**

**Scale: 2 cm / 200 km = 1:10,000,000**



**A 45 nm chip compares to Netherlands full of roads:**

**0.45 meters wide**

**0.45 meters apart**

**10 layers**

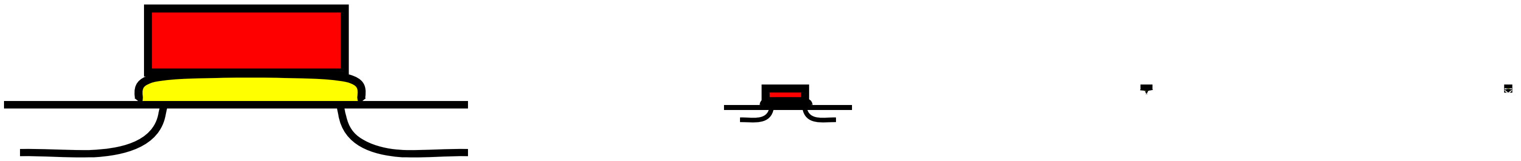
# IC Technology Scaling

**Scaling improves density and performance**

- First order scaling theory

| | | 2010 / 1971 |
|---|---|---|
| dimensions, voltage | 1/S | 0.005 |
| intrinsic delay | 1/S | 0.005 |
| power per transistor | $1/S^2$ | 0.000025 |
| power-delay product | $1/S^3$ | 0.000000125 |

- Scaling trend

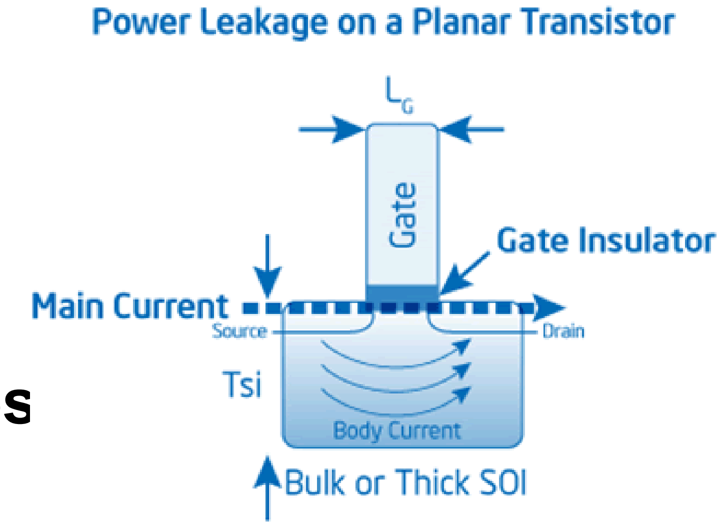


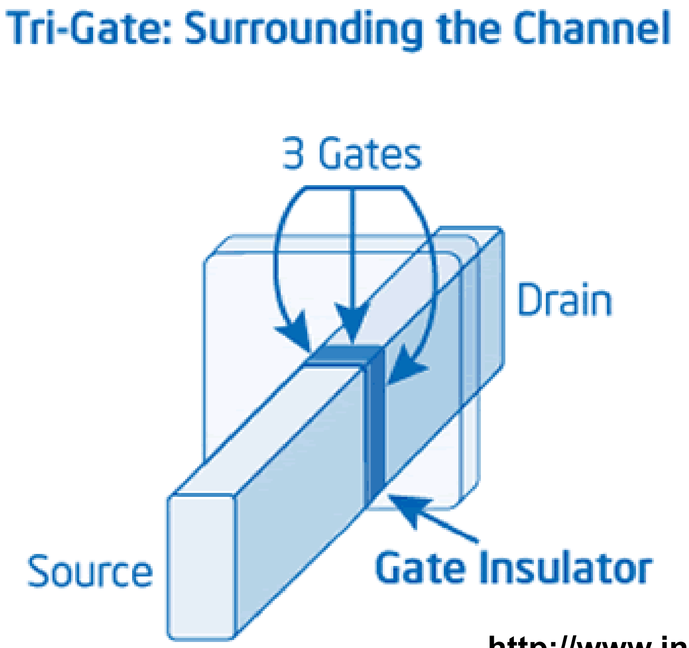
| 1971 | 1982 | 2000 | 2010 |
|---|---|---|---|
| S=1 | S≈5 | S≈50 | S≈200 |

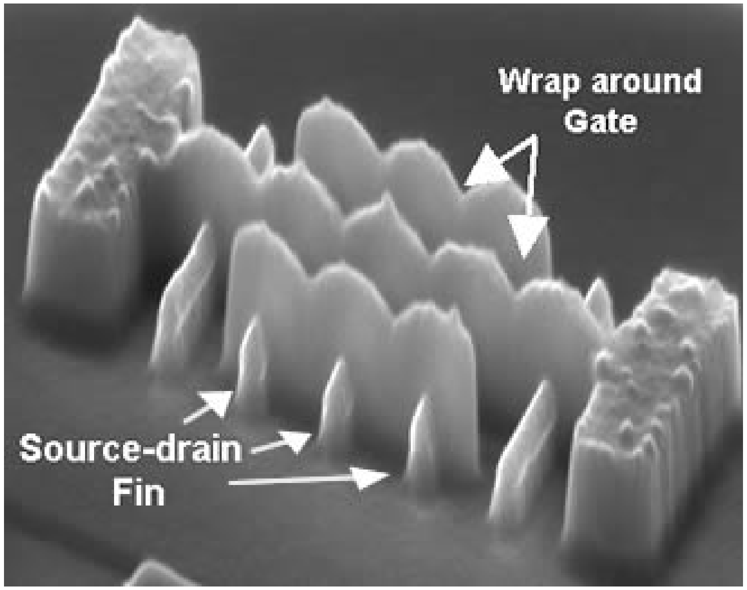
first µproc

# Advanced Issues

- Variability, manufacturing tolerances
- Scaling
- Reliability
- Advanced device architectures


Power Leakage on a Planar Transistor


Tri-Gate: Surrounding the Channel



**http://www.intel.com/technology/silicon/tri-gate-demonstrated.htm**

# Summary

- Overview of important concepts
  - MOS devices
  - Operating regions
  - Models
  - Scaling
- Outlook

- Study details yourself!