Devices and Circuits of the Nanoscale

Sandip Tiwari
st222@cornell.edu

A discussion of electronics and some of its devices and circuits with emphasis on nanoscale effects in the context of applications and systems

Background on electronics and CMOS devices
Nanoscale in Silicon
Nanoscale in Other Materials
Nanoscale Devices
Circuits in the Context of Systems

Electronics

![Graph showing Moore's Law](source: Intel)

Foundation of a trillion dollar information industry: smaller area, faster & cheaper year after year: Moore’s Law
Lemma: An industry that works hard and spends billions at putting itself out of business
Elements of an Electronic System

- Logic
  - Logic execution, Logic interfacing (drivers, bus, interface, …)
- Memory
  - Cache, Data, Code, Storage, … (dynamic and non-volatile, … fast and slow)
- Communication
  - On and off chip to other chips, boards, …
- Interfaces
  - Display, touch, sound, keyboard, sensors, other input/output
- Hierarchical system design
**Bulk Transistor**

We want:
- High on current $I_{on}$
- Low off current $I_{off}$
- Rapid control between the two states
  - ideal is 0 mV
  - practical is 60 mV for a decade change in current
- Reproducible
- Low sensitivity to variations
- Low energy

*Source: Skotnicki et al. (2005)*

---

**Electron Transport in FET**

*Credit*
**Transistor**

Charge confinement (doping)
Mobility and effective velocity

Resistances
Doping and conductivity

Mobility

Charge confinement
Gate control

Charge confinement (doping and body)
Gate control

**Nanoscale**

Scaling doesn’t quite work below 100 nm

*IBM*

*T. Kuroda*
**Power**

Passive power increasing at rapid rate due to gate and inter-junction leakage

Dielectrics and junctions with increasing tunneling

---

**Interconnects: 2D**

- Technology scaling occurs with increasing average interconnect length and routing density and increased interconnect aspect ratio
  - Interconnects grow linearly with cells in ordered arrays (memories, e.g.)
  - Interconnects grow as the square of the elements in random logic
- Local (intra-block) wires scale with block size, but global (inter-block) wires do not. Global wiring and increasing buffers become an increasingly problem
## Throughput & Power Dissipation in Buffers

- Use of repeaters means more power, and absence means increased delays with global delays more dominant.
- In 65 nm high speed designs, the # of buffers is ~850K (more area, power, and congestion).

![Diagram of driver and receiver with throughput and power dissipation](source: Deodhar et al.)

### Application Dependent Limits

<table>
<thead>
<tr>
<th>Device Type</th>
<th>Application</th>
<th>T (C)</th>
<th>Power (mW/cm)</th>
<th>VDD (V)</th>
<th>Lat (nA/um)</th>
<th>Vth (mV)</th>
<th>Ibias (nm)</th>
<th>Wmax (nm)</th>
<th>Lmax (nm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulk</td>
<td>High Performance</td>
<td>85</td>
<td>1000-30</td>
<td>0.9-1.2</td>
<td>1000-110</td>
<td>140-235</td>
<td>1.0-1.3</td>
<td>8.5-11</td>
<td>14.5-19</td>
</tr>
<tr>
<td></td>
<td>...burn-in limited</td>
<td>-40</td>
<td>1000-30</td>
<td>0.7-1.0</td>
<td>1000-115</td>
<td>95-150</td>
<td>0.9-1.2</td>
<td>8-10</td>
<td>13.5-18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-170</td>
<td>1000-30</td>
<td>0.5</td>
<td>1000-155</td>
<td>50-75</td>
<td>0.9</td>
<td>8</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>140</td>
<td>-</td>
<td>1.8</td>
<td>1000</td>
<td>180</td>
<td>1.3</td>
<td>13.5</td>
<td>23</td>
</tr>
<tr>
<td>Bulk</td>
<td>Medium-High Performance</td>
<td>85</td>
<td>30-5</td>
<td>0.8-1.2</td>
<td>120-20</td>
<td>235-300</td>
<td>1.2-1.5</td>
<td>10-14</td>
<td>17-24</td>
</tr>
<tr>
<td>Bulk</td>
<td>Moderate Performance</td>
<td>85</td>
<td>5-0.5</td>
<td>0.6-1.0</td>
<td>25-2</td>
<td>300-360</td>
<td>1.3-1.6</td>
<td>10-14</td>
<td>17-24</td>
</tr>
<tr>
<td>Bulk</td>
<td>Low-Power</td>
<td>65</td>
<td>0.5-0.001</td>
<td>0.7-0.9</td>
<td>1.0-0.01</td>
<td>410-550</td>
<td>1.7-2.0</td>
<td>13-17</td>
<td>22-29</td>
</tr>
<tr>
<td>Bulk</td>
<td>Ultra-Low-Power</td>
<td>40</td>
<td>&lt;0.001</td>
<td>0.7-1.0</td>
<td>&lt;0.008</td>
<td>550-710</td>
<td>2.1-2.6</td>
<td>16-22</td>
<td>27-36</td>
</tr>
<tr>
<td>Bulk</td>
<td>Low-Power SRAM</td>
<td>65</td>
<td>0.1-0.01</td>
<td>0.9-1.2</td>
<td>2.0-4</td>
<td>300-360</td>
<td>1.4-1.6</td>
<td>12-15</td>
<td>20-26</td>
</tr>
<tr>
<td>Bulk</td>
<td>Ultra-Low-Power SRAM</td>
<td>60</td>
<td>&lt;0.001</td>
<td>0.5-0.5</td>
<td>425-510</td>
<td>1.7-2.0</td>
<td>15-19</td>
<td>25-30</td>
<td></td>
</tr>
<tr>
<td>Bulk</td>
<td>DRAM - metal gate</td>
<td>85</td>
<td>-</td>
<td>1.0</td>
<td>0.0001</td>
<td>790</td>
<td>2.5</td>
<td>28</td>
<td>49</td>
</tr>
<tr>
<td>Bulk</td>
<td>DRAM - neg. wordline</td>
<td>85</td>
<td>-</td>
<td>1.0</td>
<td>0.0001</td>
<td>790</td>
<td>2.5</td>
<td>28</td>
<td>49</td>
</tr>
<tr>
<td>DG-FET</td>
<td>High Performance</td>
<td>85</td>
<td>30-5</td>
<td>0.8-1.2</td>
<td>90-15</td>
<td>245-305</td>
<td>1.3-1.6</td>
<td>5-7</td>
<td>13-17</td>
</tr>
<tr>
<td>DG-FET</td>
<td>Medium-High Performance</td>
<td>85</td>
<td>5-0.5</td>
<td>0.6-1.0</td>
<td>20-2</td>
<td>300-360</td>
<td>1.3-1.7</td>
<td>5-6</td>
<td>13-16</td>
</tr>
<tr>
<td>DG-FET</td>
<td>Low-Power</td>
<td>65</td>
<td>0.5-0.001</td>
<td>0.7-0.9</td>
<td>0.7-0.007</td>
<td>420-510</td>
<td>1.7-2.1</td>
<td>5.8</td>
<td>14-20</td>
</tr>
<tr>
<td>DG-FET</td>
<td>Ultra-Low-Power</td>
<td>40</td>
<td>&lt;0.001</td>
<td>0.7</td>
<td>&lt;0.005</td>
<td>530-660</td>
<td>2.1-2.5</td>
<td>5-6</td>
<td>16-22</td>
</tr>
<tr>
<td>DG-FET</td>
<td>Moderate Performance</td>
<td>85</td>
<td>0.1-0.01</td>
<td>0.9-1.2</td>
<td>1.3-3.04</td>
<td>315-355</td>
<td>1.5-1.7</td>
<td>5-9</td>
<td>15-27</td>
</tr>
<tr>
<td>DG-FET</td>
<td>Low-Power SRAM</td>
<td>65</td>
<td>5-1</td>
<td>0.9-1.2</td>
<td>1.2</td>
<td>425-475</td>
<td>2.5</td>
<td>17</td>
<td>30</td>
</tr>
<tr>
<td>DG-FET</td>
<td>Ultra-Low-Power SRAM</td>
<td>60</td>
<td>&lt;0.001</td>
<td>0.5-0.5</td>
<td>730</td>
<td>1.2-2.0</td>
<td>17-22</td>
<td>11-15</td>
<td>25-33</td>
</tr>
</tbody>
</table>

*Source: D. Frank et al. IEDM (1998)*

- **Red**: constrained by source-drain tunneling
- **Blue**: constrained by functionality
- **Green**: constrained by noise margin

---

*Short Course, South Korea, 2006* Sandip Tiwari; Cornell University
Power Dissipation in Small Dimensions & Temperature

- $10^5$ W/cm² => 100°C with package at 50°C at 0.18 μm dimension
- Area in which this dissipation occurs critical to temperature

Liu et al., IEEE EDL (2002)

Energy determines density for electronic nanosystems
Consequences of Improving Electrostatics

<table>
<thead>
<tr>
<th>Higher body doping</th>
<th>Lower carrier mobility</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Higher junction capacitance</td>
</tr>
<tr>
<td></td>
<td>Higher junction leakage</td>
</tr>
</tbody>
</table>

Thinner gate dielectric → Higher leakage

Shallower junctions → Higher resistance

So, there are always compromises to be made

Statics and Dynamics

**Electrostatics**
- **Gate Control**
  - Gate dielectrics, work-functions, ...
- **Substrates**
  - Sharp halo’s and improved junctions
  - Thin silicon bodies
- **Threshold**
  - Work functions, doping, new geometries

**Electrodynamics**
- **Transport**
  - Strained materials (Si, SiGe, ...), new orientations, new materials (Nanotubes, Ge, III-V, ...)
- **Parasitics**
  - New contact materials, raised source-drain structures, etc.
Non-Classical CMOS (Single Gate)

Transport enhancement

Strained Si, Ge, SiGe

Substrate Enhancement:
Thin Body

SiGe, SiCGe, …on
bulk Si & SOI

Source/Drain Enhancement

Silicide

Low barrier Schottky
source/drain or non-
overlapped extensions

Source: ITRS

Non-Classical CMOS (Multi-Gate)

Tied Gates

Channels on multiple surfaces

Vertical transistor

Channels on side-walls

Independent gates

Channels on planar surfaces

Source: ITRS
**Planar Transistors**

Bulk

- Scaling limited
- Enhanced by Si/SiGe channels

Silicon on Insulators: Various Forms

- Scaling limited
- But improved performance that can be enhanced by bulk-like approaches
- Enhanced scaling limits if thin silicon and low parasitics feasible
- Enhanced scaling limits if thin silicon, two gates and low parasitics feasible

**Strain by Orientation**

Electron Mobility (cm²/V·s) vs. \( N_{inv} (\text{cm}^{-2}) \)

- \( (100)/<110> \)
- \( (111)/<112> \)
- \( (110)/<110> \)

Hole Mobility (cm²/V·s) vs. \( N_{inv} (\text{cm}^{-2}) \)

- \( (110)/<110> \)
- \( (111)/<112> \)
- \( (100)/<110> \)

Yang et al. (2003)
Transport Improvement by Orientations

Bond (110) onto (100) Si – oxide based
Use epitaxy of 100 with oxide isolation for (110)

Yang et al. (IBM)

Strain

Reduced average conductivity mass
Reduced intervalley scattering
Higher effective mobility

Takagi (2003)
**Strained Si**

![Strained Si on SiGe](image)

Strain by Process

Uniaxial Strain

**Mobility Enhancements with Silicon(strained) on Insulator**

K. Rim (IBM)

Thompson et al. (2004)
**High k (Permittivity)**

HfO$_2$ has $10^4$ less leakage than equivalent SiO$_2$.

But, large interface state issues

---

**Combinations**

To date, mobility degradation with high permittivity materials is substantial.

---

Gousev et al. (IBM)

Rim (IBM)
**FinFet**

![FinFet Diagram](image)

**Majkusiak (1998)**

---

**Thin Si**

![Thin Si Diagram](image)

**Majkusiak (1998)**
**Transport in Thin Silicon**

- Front Gate: 5.8 nm
- Front Oxide: 5.3 nm
- Silicon: 5.8 nm
- Back Oxide: 7.0 nm
- Back Gate

Charge distributed across silicon: 
-3.0 V < VBG < 3.0 V

FG

+++

---

Si

Channel

A. Kumar, et al. (2005)

Good electron transport still maintained in thin silicon

**Power: Switching and Standby and Adaptive Control**

- L = 90 nm with 2 nm front oxide, 5 nm back oxide, 25 nm Si, and using 21 stage ring oscillator
- Devices provide tuning of standby power and switching performance with good noise margin

**Thin Si**

![Image of thin Si structure with gate oxide thickness](image)

**Flash Non-Volatile Memories**

**NOR Write**
- Programming: Channel Hot Electron
  - Vwl: 8-10 V, Vbl: 4-5 V
  - $T_{\text{pulse}}$: 1 us, $I$: 10-100 uA
  - 0.5 MB/s

**NOR Erase**
- Erasing: FN Tunneling
  - Vwl: -8 V, Vbody: 6-8 V
  - $T_{\text{pulse}}$: 100 ms, $I$: ~0 uA

**NAND Write**
- Programming: FN Tunneling
  - Vwl: 18-20 V, Vbody: 0 V
  - $T_{\text{pulse}}$: 300 us, $I$: ~0 uA
  - 7-10 MB/s

**NAND Erase**
- Erasing: FN Tunneling
  - Vwl: 0 V, Vbody: 18-20 V
  - $T_{\text{pulse}}$: 2 ms, $I$: ~0 uA

Source: R. Bez
Nanoscale Classical Picture

- Achieving quantum confinement
  - Quantum wells (2D)
  - Quantum wires (1D)
  - Quantum dots (0D)
**Confinement & Degeneracy**

Degeneracy capacitance is non-geometrical
- does not scale with layer thicknesses.
- Constant in 2-D (single subband occupancy)
- Large in n-Si because of valley degeneracy and large effective

\[ \eta = \frac{eN}{C_{2D}} \]

(Equivalent Circuit)

\[ C_{2D} = 4\pi n_v e^2 m/\hbar \]

Solomon & Laux, IEDM 2001

**Limits**

- Classical (semi-) physics has sufficed to date
- Behavior changes when electron-wavelength approaches device dimensions
  - Few-electrons per device questions
  - quantization perpendicular to transport with confinement-energy penalties in threshold voltage control and transport.
  - wave function penetration in transport direction introduces tunneling leakage in off-state
    
    \[ \exp(-2J\alpha dx) \]
    
    conflicting requirements between low mass for transport vs. tunneling.
- Statistics of small numbers of impurities limit reproducibility of small devices
- Timing fluctuations, even though above the threshold electron individuality is lost
- Related to timing, energy and power limits
Nanoscale: Power and Performance

- Suppose we could make devices at a 10 nm x 20 nm minimum dimension with a cell size of 50 nm x 60 nm (3.3 x 10^{10} cm^{-2})
- And, suppose we limit the power density to 100 W/cm^2 and 1 V supply

- If all elements were continuously switching the average power per device is 3.33 nW/device at 6 nA/device, or 1 electron transiting every 27 ps (TOO SLOW)
- Present digital design handles this by partitioning functions and allocating power according to speed desired: clocks high and cache low
  - Needs multiple threshold voltages and a variety of circuits
- Temperature of 100 C (50 C package) in an isolated small element implies current of <0.5 \mu A

An Electron in a Semiconductor

- Unhindered movement of a single electron is \mu A’s of current
  - However, to observe it, requires constraints (barriers, e.g.) and the current drops – typically nA

- A 10 nm x 10 nm x 10 nm cube of silicon has ~50 available states in ~1 eV of energy range

- Variance of an ensemble of \( n \) that follows Poisson distribution is \( 1/\sqrt{n} \)

- Mean free path of a hot electron is 5-40 nm
**Discreteness**

Frank et al. (2000)

**Quantum Dots – Single Electron Effects**

\[
\Theta_{\text{Bath}} = 15 \text{ mK} \\
\Theta_{\text{Electrons}} = 80 \text{ mK}
\]
Charging Effects

- Charging of a small particle with an extra electron requires an energy: \( E_C = \frac{e^2}{2C} \)
- A small particle (~10 nm) in a dielectric (SiO$_2$, e.g.) has \( C = 2 \) aF, \( E_C = 40 \text{ meV} \sim 150 \text{ C} \)

Coulomb Blockade and Staircase

- **Blockade**: no current flow until an electron can charge the particle

- **Staircase**: When particle charged by > 1 electron
**Impedance, Currents and Size Effects**

- For a clear observation of Coulomb blockade
  - System energy change much larger than eigenstate width (which is related to lifetime of state/tunnel escape rate)
  - requires $R_{T} \gg \hbar/2\pi e^2$ or 4.1 kΩ
  - In real structures, $R$ is typically $G\Omega$
  - Poor Gains (Power and Voltage) and Impedance mismatch
  - Time constants ($RC$) of ns and currents of nA
- Size has significant effect through charging energy
  - $E_c \sim 1/C$ and $\Delta E_c/E_c \sim \Delta L/L$
- Size has significant effect through subband energy
  - Sub-band Energies: $\Delta E_0 \sim 1/L^2$ and $\Delta E_0 \sim 2 \Delta L/L$

---

**Single Electron Transistor**

Requires atomic fab accuracy for room temperature

Low currents, but allows larger variety of materials

Yu. Pashkin et al. (2002)
Single Electron Latching Switch

![Image of single-electron transistor and latching switch]

- **Single-electron trap**
  - $C_0$
  - $C_c$
  - $V_S$
  - $V_D$

- **Single-electron transistor**

- **I-V curve within the "Orthodox" theory**

- **Low-temperature prototype (trapping time > 12 hrs)**
  - P. Dresselhaus et al. (1994)

Nanoscale: Classical Charge Effect

If $C = 1 \text{ aF}$, $e^2/2C = 160$ meV

- $C = 1 \text{ aF}$ is a 18 nm metal particle in free space or ~4 nm in oxide
- Single electron charging occurs with blockade regions (Coulomb Blockade)

![Image of nanoscale quantum dot and tunnel oxide]

- **Nano-Crystal Quantum Dot**
- **Control Oxide**
- **Tunnel Oxide**
- **Channel Surface**

- **Tiwari et al. APL (1996)**
- **Muralidhar et al., IEDM (2003)**

Makes low power memories possible
Nanoscale in Other Materials

Nanotubes

- High carrier mobility
  - Ballistic transport (<1-10μm)
  - >10,000 cm²/V.s (>10μm)
- High current carrying capabilities
  - J=10⁹ A/cm² (Most metal fails at <10⁶ A/cm²)
- Nearly-ideal surface (!)
  - Wider choice of dielectrics
- All atoms on surface
  - Potential for sensors
- Can be direct bandgap
  - Potential for optical devices
- Diameter determines semiconducting (2/3) vs metallic tubes (1/3), and placement
Carbon Nanotubes

- Good transport
- But, poor control of:
  - Placement
  - Thickness
  - Chirality

Nanotube Band Structure and Mobility

Graphite
Zero gap

semiconducting

Quantized $k_{\perp}$

$\pi D k_{\perp} = 2\pi n$

Reported max. mobilities:
1,000 - 100,000 cm²/V.s
Diameter Dependence

$G(4e^2/h)$

$P(10^3 \text{cm}^2/\text{V-s})$

$V_g (\text{V})$

$G$ increases with $d$

$G_{\text{max}}$ versus $d$

$\langle \rho \rangle = \left( R_{300K} - R_{50K} \right) / L$

$G_{\text{max}} \sim d$, $\mu_{\text{peak}} \sim d^2$

$\mu (10^3 \text{cm}^2/\text{V-s})$

$\mu_{\text{peak}}$ versus $d$

$\mu (10^3 \text{cm}^2/\text{V-s})$

$d^2$ Ref.
Temperature Scaling

\[ G_{\text{max}} \sim 1/T, \ \mu_{\text{peak}} \sim 1/T \]

Model – Acoustic Phonons

Effective mass:

\[ m^* = \frac{h^2}{\partial^2 E/\partial k^2} = \frac{2h}{3v_0} \left( \frac{1}{d} \right) \]

\[ \tau_0^{-1} = \alpha \frac{T}{d} \]

\[ G_{\text{max}} = \frac{4e^2 l_0}{hL} = \frac{4e^2 v_0 d}{h\alpha L T} \]

\[ \mu_{\text{peak}} = 0.32 \frac{e \tau_0}{m^*} = 0.48 \frac{e v_0 d^2}{\hbar \alpha T} \]

Molecules

- Small, and digitized size, shape and functionality
  - forgiving tolerance, and can perform specific electrical and mechanical functions, and can be self-assembled

- But,
  - Based on stochastic processes
  - Fragility of organic structures
    - Charge states depend on current flow
      - Stability dependent on charge/oxidation state and temperature
  - Molecules are difficult to access
    - Interfacing difficult
  - Proximity of contacts broaden levels and induces gap states
  - Line shapes do not have a sharp cut-off

Molecular Rectifier

- Analogy with semiconductor diode:

Chemical potential close to LUMO

Chemical potential close to HOMO

**Molecular Resonant Tunneling Diode**

- CH₂ groups act as tunnel barriers
- Negative differential resistance (NDR)
- Like a resonant tunnelling diode (RTD)


**Molecular Transistors**

- Behavior is very complex
- Conceptual models are needed

DFT Simulations: N. Lang
**Single Electron Molecular Transistors**

J. Park et al. (2002)

![Diagram of molecular transistors](image1.png)

**Magnetic RAM**

Stores information using the magnetic polarity of a thin, ferromagnetic layer.

Information read by measuring current or resistance across the MRAM stack.

Current determined by the rate of electron quantum tunneling, which is affected by magnetic polarity of the cell.

The “Free Layer” polarization is allowed to change, depending on if the cell is High or Low

The resistance across the stack is measured to determine the cell state
**What is MRAM? How it works**

What is MRAM? How it works

**Conductance in Magnetic Layers**

Conductance in Magnetic Layers

(a) Parallel
Low Resistance

(b) Anti-parallel
High Resistance

Source: Slaughter (2004)
Variance in Magnetic Structures


Nanoscale Devices
NanoCrystal Floating-Gate Memory

Charging and Erasure

Electrostatic energy change upon addition of an electron
\[ \Delta E_s = \frac{Ne^2}{C} + \frac{e^2}{2C} \]

Hamiltonian for the system:
\[ H = H_{2deg} + H_{qd} + H_T, \]

where
\[ H_{2deg} = \sum_n (\epsilon_n + eV)a_n^{\dagger}a_n \]
with \( n \) identifying the indices of the ladder in the inversion layer
and
\[ H_T = \sum_{n,m} T_{nm}a_n^{\dagger}b_m + c.c. \]
with \( m \) identifying the indices of the ladder in the quantum dot

Equation of motion for the density matrix:
\[ i\hbar \frac{\partial \hat{P}_H(t)}{\partial t} = \left[ H, \hat{P}_H(t) \right] \]
**RTS: Nano-Crystal Memory**

**Single-Electron Events**

- Fast and slow processes – surface states; and correlated processes.

![Graph showing RTS Amplitude](image)

RTS Amplitude ~ 14%  \( \Delta t = 1 \) msec

**Nanocrystal Memories**

4Mb Array in a 6V 90 nm process

*Muralidhar et al., 2003 IEDM*
Scaled Front-Side SONOS Memories

Use of higher defect density to counter statistical effects

\[ \Delta V = 2.7 \text{ V} \]

\[ L = 46 \text{ nm, } W = 33 \text{ nm} \]

ONO stack = 2 / 6 / 12 nm

Kim et al., IEEE SNW (2003)

SONOS Memories

A Constant Window

75 electrons

W/L 33nm/46nm

W/L 62nm/60nm

W/L 75nm/100nm

W/L 90nm/100nm

Kim et al., IEEE SNW (2003)
Memory Using Defects on Back

Silva et al. (2005)

---

Defects on the Back

ONO stack = 2 / 6 / 13 nm; L = 50 nm, W = 100 nm

S = 70 mV/decade

\[ V_D = 1 \text{ V} \]

\[ S = 70 \text{ mV/decade} \]

\[ t_{Si} = 20 \text{ nm} \]
**Single Electrons in Output Characteristics**

\[ V_D = 100, 200 \text{ mV} \]

\[ V_G - V_T = 0, 0.1, \ldots 0.5 \text{ V} \]

Oxide/Nitride/Oxide : 8 / 15 / 40 nm; \( t_{Si} = 50 \text{ nm} \)

*Silva et al., IEEE SOI Conf. (2003)*

**Carbon Nanotube Field-Effect Transistor**

Subthreshold Characteristics

Output Characteristics

\( V_{gs} = 0.4 \text{ to } -1.6 \text{ V} \)

Step -0.4V
**Nanotube FETs**

- 1D (ultra-thin body) channel
- Ballistic transport (at low Vds)
- Switching can be dominated by the contact Schottky barriers
  - Screening length
  - Barrier width ~ oxide thickness tox (on-state)
  - Ambipolar behavior

\[ \lambda \sim \sqrt{t_{ox}d_i} \]

Gate oxide: 10-nm SiO₂
Contact metal: Ti

**Performance**

Drain current \( I_d \) vs. Al gate voltage \( V_{g-Al} \)

Bulk Switching

Theoretical limit: \( S \sim 60 \text{ mV/dec} \)

Contact Switching

IBM


**Band to Band Tunneling in Nanotubes**

- The semiconductor is one-dimensional
- The body of the semiconductor is ultra-thin
- Transport in the semiconductor is ballistic
- The effective masses of electrons and holes are small
- The effective masses of electrons and holes are similar
- The semiconductor has a direct band gap

**Molecule as Filter**

\[ V_G = \eta V_G + \frac{(1 - \eta)}{2} V_B \]

- Molecular state spectrum shifts with gate potential
- Symmetric molecule with unity peak resonant transmission
- State spectrum fixed relative to central molecular potential.
- Electrochemical potentials represented by source & drain (n or p) Fermi levels
- State occupancy is 0 above VP, 1 below VN and ½ between VN and VP
- Charge in filled state = # of electrons in orbital represented by that state
- Central Potential given by electrostatic coupling to S, D and G potentials and increased by \( Q_{sc}/C \)
- Polarization by fixed dielectric permittivity.

Molecules and Self-Assembly

If molecule mimics MOSFET
For gate field to penetrate molecular channel
- Dielectric thickness to be comparable to the molecular length
- Intimacy between molecule gate dielectric
- Molecule sufficiently long and chemically functionalized and the gate dielectric is sufficiently thick to limit tunneling between source and drain electrodes and to ensure an “OFF” state of the device and between source-gate

- If self-assembly used as a technique for fabrication
  - Low energy scales of assembly process (~ eV)
  - Higher defect rate with consequences for larger scale
- Is current sufficient

Circuits and Systems
**Stability and Signal Recovery**

Analog: open & closed loop/feedback
Regenerative effects and signal stability issues

Digital
Signal restoration using gain

\[
\frac{V_{\text{out}}}{V_{\text{in}}} = \frac{Z_2}{Z_1} \left[ 1 + \frac{1}{K \left( 1 + \frac{Z_2}{Z_1} \right)} \right]
\]

Feedback between input and output \((Z_2, \text{e.g.})\) leads to a larger input load because the 180° phase shift during amplification
At the input \(C\) appears as \((K+1)C\) (Miller effect)

**Devices and Circuits**

Digital CMOS design:
- Only two circuit forms matter
  - (maybe three)
- Static CMOS, and Dynamic CMOS

These forms employed because:
- They are not highly demanding of devices because of power gain
  - So they work with transistors
  - Robust, especially static circuits
Sensing

- Voltage • stable
- Semi-latch • improved
- Full-latch • fast
  • precharge
- Current • fast
  • high power
### MRAM Designs

<table>
<thead>
<tr>
<th>Twin Cell</th>
<th>Averaging</th>
<th>Self-referencing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read time [ns]</td>
<td>6.96</td>
<td>6.16</td>
</tr>
<tr>
<td>Power [μW]</td>
<td>179.37</td>
<td>176.67</td>
</tr>
<tr>
<td>No. of Transistors</td>
<td>11</td>
<td>11+1 capacitor</td>
</tr>
<tr>
<td>Density Index</td>
<td>1</td>
<td>500</td>
</tr>
</tbody>
</table>

Source: Sudheeran and Chang (2006)

### NT Mixer/Transistor

FET transistor: \[ I = \frac{\partial G}{\partial V_g} \left( |\Delta V_g| + \frac{V_s}{2} \right) V_s \]

Mixing signal for ac applied to source: \[ I_{\text{mix}} = \frac{1}{4} \frac{\partial G}{\partial V_g} (V_s^{\text{ac}})^2 \]

![NT Mixer/Transistor schematic](attachment:image.png)

\( f = 10 \text{ MHz} \)
**NT Mixer/Transistor up to 50 GHz**

**Classical vs. Quantum Computing**

- Classical bit: 1 (On) and 0 (Off)
  - Stable pointer states of the computer hardware
- Quantum bit: Qubit (superposition of two states)
  \[
  |\psi\rangle = \alpha|0\rangle + \beta|1\rangle
  \]
  - Every two level system can serve as qubit
- For any digital computer, its set of computational states is some set of mutually distinguishable abstract states
  - The specific computational state that is in use at a given time represents the specific digital data currently being processed within the machine
  - In quantum computing the computational state is not always a pointer state
### Classical versus Quantum Bit

<table>
<thead>
<tr>
<th></th>
<th>Classical</th>
<th>Quantum</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Info unit</strong></td>
<td>Bit 0 or 1</td>
<td>Super position $</td>
</tr>
<tr>
<td><strong>Storage Capacitance</strong></td>
<td>Linear - N</td>
<td>Exponential O($2^n$)</td>
</tr>
<tr>
<td><strong>Processing</strong></td>
<td>Serial $x \rightarrow f(x)$</td>
<td>Parallel $\sum</td>
</tr>
<tr>
<td><strong>Universal Gates</strong></td>
<td>Nand</td>
<td>Single qubit Rotations + Cnot</td>
</tr>
<tr>
<td><strong>Measurement</strong></td>
<td>0 -&gt; 0</td>
<td>Problem : destroy coherence</td>
</tr>
<tr>
<td><strong>Algorithms</strong></td>
<td>Many</td>
<td>Factorization, Search Based on quantum interference</td>
</tr>
</tbody>
</table>

---

### The Square Root of NOT

- If input is either basis state (0 or 1) you get a state that appears *random* when measured…

\[
\begin{array}{c|c|c}
0 & N^{1/2} & 0 (50\%) \\
1 & N^{1/2} & 1 (50\%)
\end{array}
\]

- But if you feed the output back into another $N^{1/2}$ *without measuring it*, you get the *inverse* of the original value!

- “How is that possible?”

\[
\begin{array}{c|c|c}
0 & N^{1/2} & 0 (50\%) \\
1 & N^{1/2} & 1 (50\%)
\end{array}
\]

\[
\begin{array}{c|c|c}
0 & N^{1/2} & 0 (50\%) \\
1 & N^{1/2} & 1 (50\%)
\end{array}
\]
**NOT\(1/2\): Unitary Implementation**

\[
\sqrt{N} := \begin{bmatrix}
0 & 1 \\
1+i & 1-i \\
2 & 2 \\
1-i & 1+i \\
2 & 2
\end{bmatrix}
\]

\[
(\sqrt{N})^2 = \begin{bmatrix}
0 & 1 \\
1 & 0
\end{bmatrix}
\]

\[
\sqrt{N}\left|0\right\rangle = \sqrt{N}\left[\begin{array}{c}
1 \\
0
\end{array}\right] = \frac{1+i}{2}\left|0\right\rangle + \frac{1-i}{2}\left|1\right\rangle
\]

\[\text{Prob. } \frac{1}{2} \quad \text{Prob. } \frac{1}{2}\]

**Optical Implementation of N\(1/2\)**

- Beam splitters (semi-silvered mirrors) form superpositions of reflected and transmitted photon states

![Diagram of optical implementation]
**SET Probing Qbit**

Box: 700x50x15 nm; 10⁸ electrons

\[ e^2/2C = 117 \, \mu V, \quad T = 30 \, mK, \quad kT = 3 \, \mu V \]

Initialize by preparing pure state \[ |0> \] away from resonance

Apply fast voltage pulse \( \Delta t \) to gate (non-adiabatic) to create degenerate charge state at resonant condition (\( |0> \) and \( |1> \) are now superposed)

Sample measurement of \( |1> \) by tunneling probe

Repeat by sweeping \( \Delta t \) in pico-seconds  

\[ \text{Nakamura} \]

---

**Quantum Computing Requirements**

- DiVicenzo’s check list:
  - Identifiable qubits and their scalability
  - Initialization procedure
    - Preparation of ground state of the whole system
  - Low decoherence
    - Long decoherence time versus gate time
  - Quantum gating – ability to realize a universal gate through control of system Hamiltonian
  - Controlled evolution
  - Reliable readout

- Qubit Coupling: nearest neighbor versus common mode
- Engineering Correction Code (ECC): to address decoherence – redundant qubit register and majority voting
**Systems, Hierarchy, Complexity and Architecture**

A critical look

---

**Molecular Memories?**

- Memory by using switching behavior in a crosspoint configuration
- Logic by diode logic with open and diode cross-points and resistors (!)

<table>
<thead>
<tr>
<th>Current (mA)</th>
<th>Voltage (V)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-10</td>
<td>-2.0</td>
</tr>
<tr>
<td>-5</td>
<td>-1.0</td>
</tr>
<tr>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>5</td>
<td>1.0</td>
</tr>
<tr>
<td>10</td>
<td>2.0</td>
</tr>
</tbody>
</table>

molecular switch
**Molecular: 2 terminals or 3 terminals?**

Customized combinatorial logic scales polynomially with function size ($\sim N^2$).
Array logic scales exponentially with function size ($\sim 2^{2N}$)
Array peripherals scale in proportion to function size

---

**CMOL**

CMOL combines:
- (relatively sparse, but highly functional) CMOS subsystem
- very dense nanowire crossbar, and
- a molecular-scale device at each nanowire crosspoint

*Likharev (2003)*
**CMOL**

Tilt $\alpha = \sin^{-1}(F_{\text{nano}}/F_{\text{CMOS}})$

Nanodevice addressed via two CMOS cells

Each and every nanodevice may be addressed!

*Likharev (2003)*

---

**CMOL FPGA**

CMOS inverters + nanodevice latches for (re)configuration

CMOL FPGA fabric

*Likharev (2005)*
**NanoFabrics**

![Diagram](image-url)

*deHon (2004)*

---

**Speed Gaps in Processor Hierarchy**

![Graph](image-url)

*Source: T.C. Mowry*
3-D Microprocessors

- Goal of improving logic-memory interactions and to compensate logic and memory performance divergence
  - Current designs exceedingly complex (→ power^) focused on
    - Superscalar (> 1 inst/cycle), out-of-order execution, instruction-level parallelism, hiding memory latency, …
- 3-D in μP:
  - High density, low latency, large bandwidth

Latency and Bandwidth

2-D: Connections on the periphery
- Long global connections
- CPU to off-chip main memory with latency and misses

3-D: Connections across the area
- Connections short + vertical
- Suitable for high-bandwidth and vector operations
- No pin cost, large block access of data

Latency: Important for random access (servers, e.g.), single core
Bandwidth: Multiple cores, multi-threads, graphics

The following example uses a baseline 2-D processor core representative of current technology
3 GHz CPU, 750 MHz memory, 64 KB L1I, 64 KB L1D, 1 MB L2
Expanding L2 Cache

- Performance peaks at 8 MB for integer programs with standard DRAM
  - Example of trade-off between fitting the working data (4-16 MB for integer programs) into the cache (better performance) and increased access latency for larger caches (worse performance)
- Larger working data of floating-point programs continue to improve with cache size despite cache hit latency with large cache size

\[ C. C. Liu \text{ \& Tiwari}, \text{IEEE D&T Mag. (2005)} \]

L2/L3 Cache Sizing with Stream Prefetching

- Performance within 8-10% of perfect L2
  - Large speedups achievable with small L2/L3 cache because of significant reduction in main memory latency
  - Hierarchy critical to performance

\[ C. C. Liu \text{ \& Tiwari}, \text{IEEE D&T Mag. (2005)} \]
**Complexity**

- Rent’s Rule:
  - Terminal count is related to number of gates (at all hierarchical levels)
  \[ T = t N^p \]
  \((0 < p < 1; \ t \ is \ number \ of \ terminals \ per \ logic \ block)\)
  - \(p=1\) is un-optimized placement
  - Number of interconnections among a group of sub-components at any level is proportional to the total terminal count of all the sub-components
  - With placement optimization \((p<1)\), only a fraction of logic blocks accessible
  - This accessibility defines how much of the circuitry do iterative testing procedures access and test for usefulness
  - If logic blocks defective: \(N_{accessible} \sim ((1-d_{LOGIC})N)^p\)
  - If wiring defective, the number of testable logic blocks:
    \(N_{accessible} \sim (1-d_{LOGIC}) N^p\)
    - a considerably more serious problem

**Configurability: Defects and Testing**

- Suppose, we work with \(10^5\) logic blocks, each employing 1000 device elements
- What does it mean that one can work with chips that are 90% functional (or that 10% of logic blocks are faulty)
  - If the probability of failure is \(p\) for each element, probability of a logic block being functional is \((1-p)^{1000}\)
  - Rate of connectivity is non-linearly related to defects, and affects congestion, power, ....
  - Probability of 90% yield in logic blocks implies 1 in \(10^4\) device level faults
    - *We still need extremely high reproducibility and yield*
Observability in Presence of Defects

Testing of N modules in a defect-free system

- Number of modules left for testing
- Number of iterations

- Observation: Usable logic blocks of a chip reduce rapidly with interconnect defects since their correctness cannot be tested. Testability is cumbersome.

Kumar, DFT (2004)

Interconnect Defect Penalty

- Factor increase in power dissipation
- Factor increase in area

- Observation: Usable logic blocks of a chip reduce rapidly with interconnect defects since their correctness cannot be tested. Testability is cumbersome.

Kumar, DFT (2004)
Conclusion

- Nanoelectronics will certainly be evolutionary, and may be revolutionary
  - Complex applications (beyond sensing, …) require a systematic, robust and reproducible framework that requires a number of properties across scales
- Logic applications will require 3-D structures and non-Manhattan layouts
  - These usually do not work with “bottoms-up” approaches
- Multiplexing schemes to manage the interconnect pitch transformation from nano- to microscale require real estate
- Charge-based devices at nanoscale have inherent power dissipation problems
- Other approaches, spin-based or photon-based or others, need to demonstrate size scale, gain and ability to transform signal to charge and vice versa for connection to the external world