Devices and Circuits of the Nanoscale

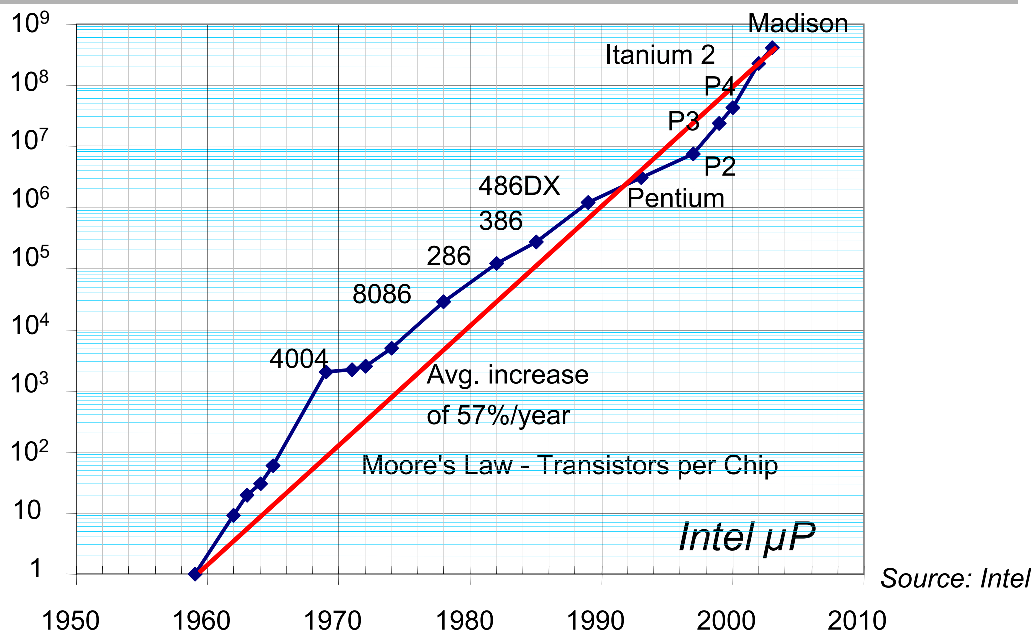
Sandip Tiwari
st222@cornell.edu

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

Cornell University

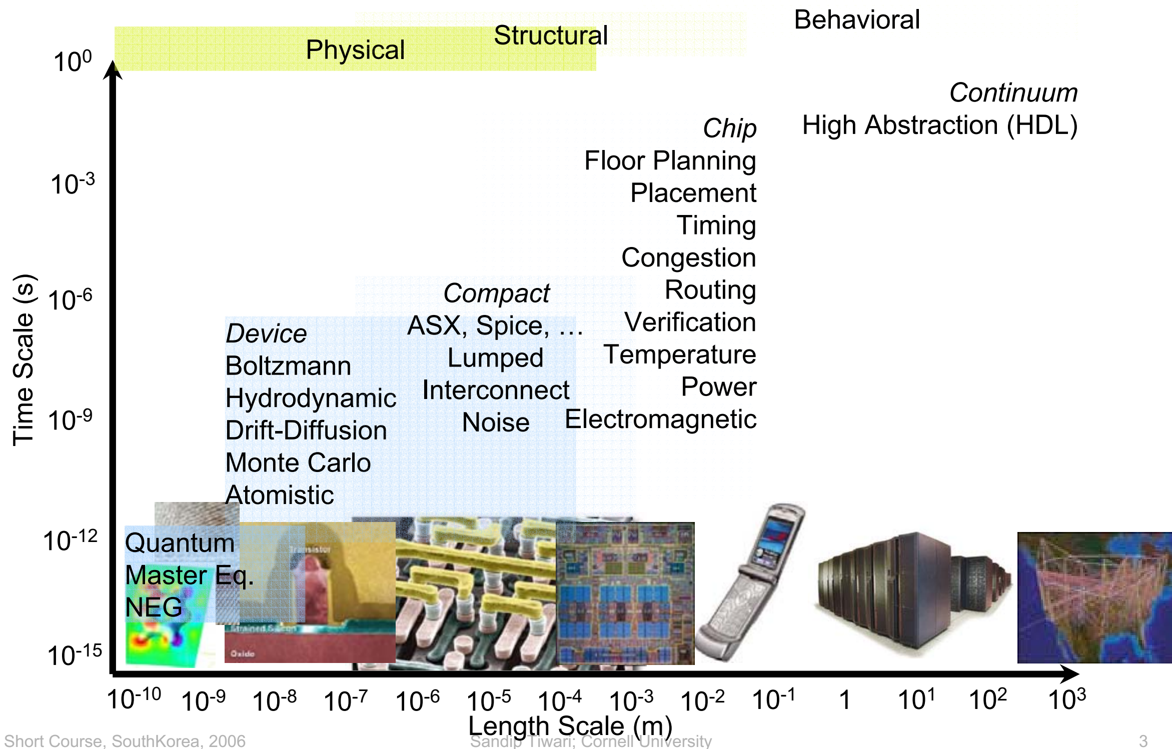
Electronics



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

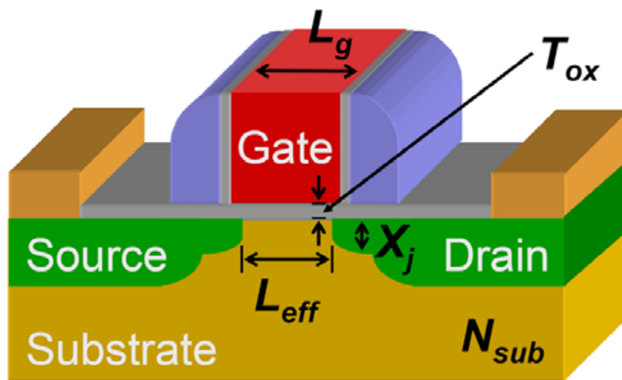
Electronics



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

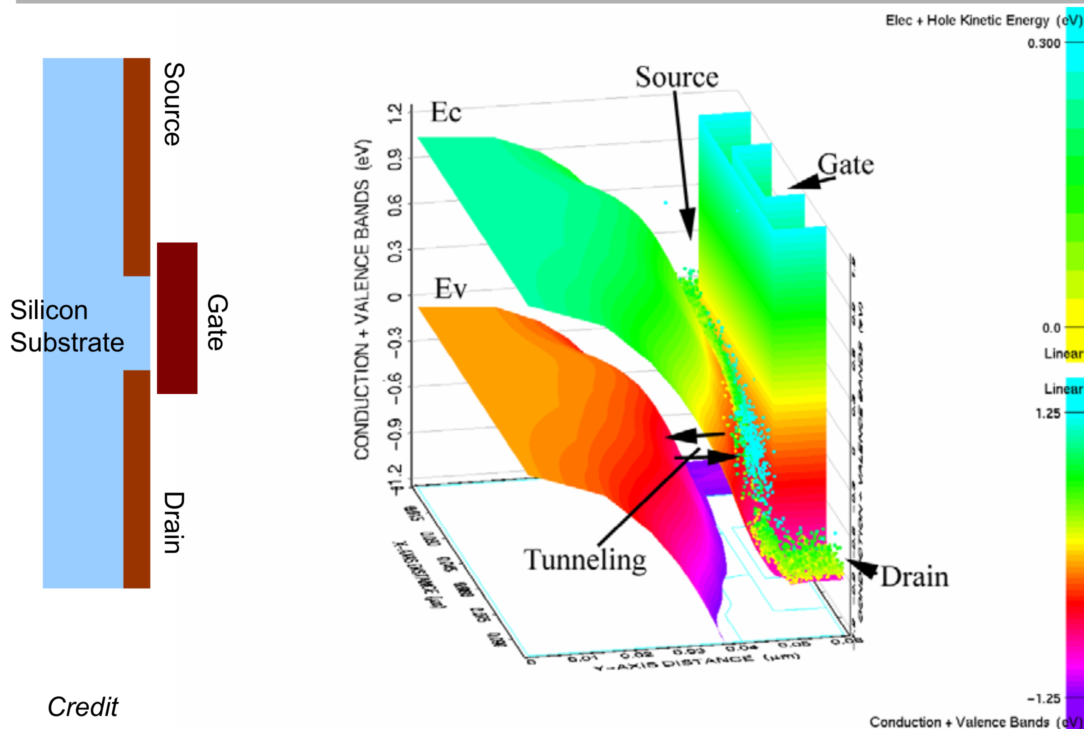


We want:

- 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

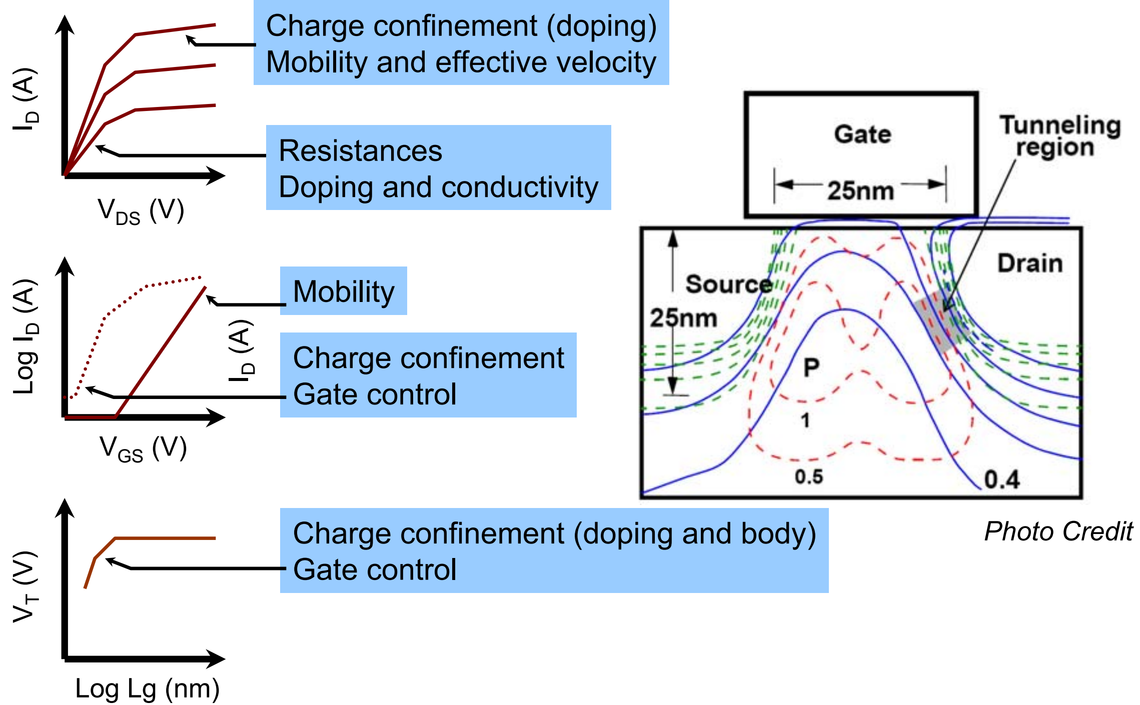
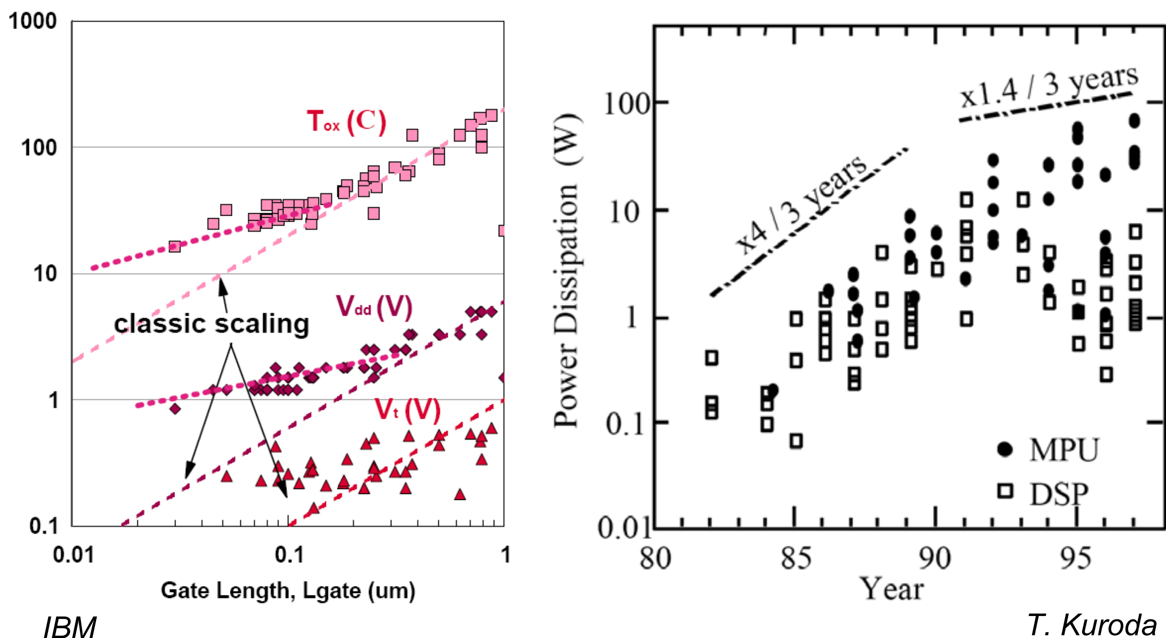


Photo Credit

Nanoscale

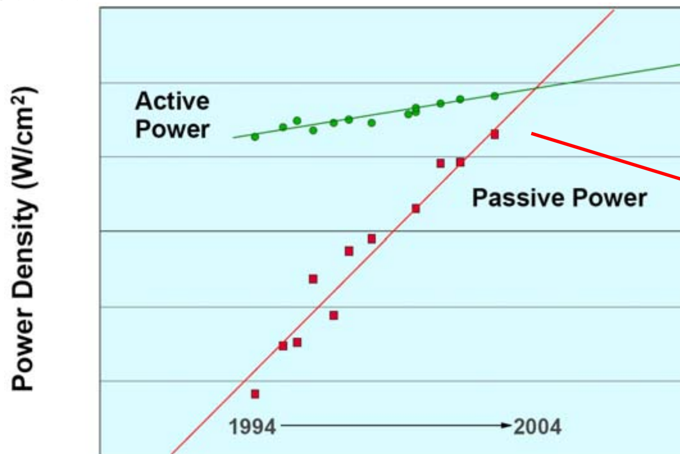


IBM *T. Kuroda*

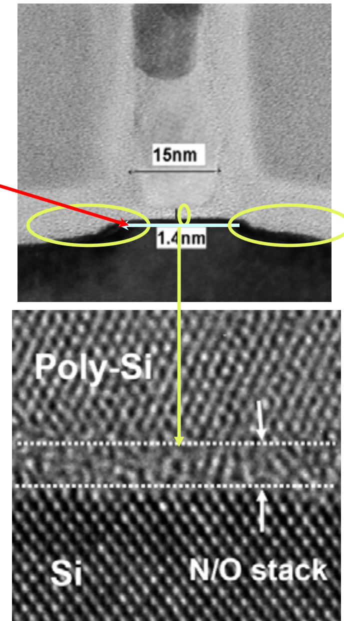
Scaling doesn't quite work below 100^{+} nm

Power

Source: IBM



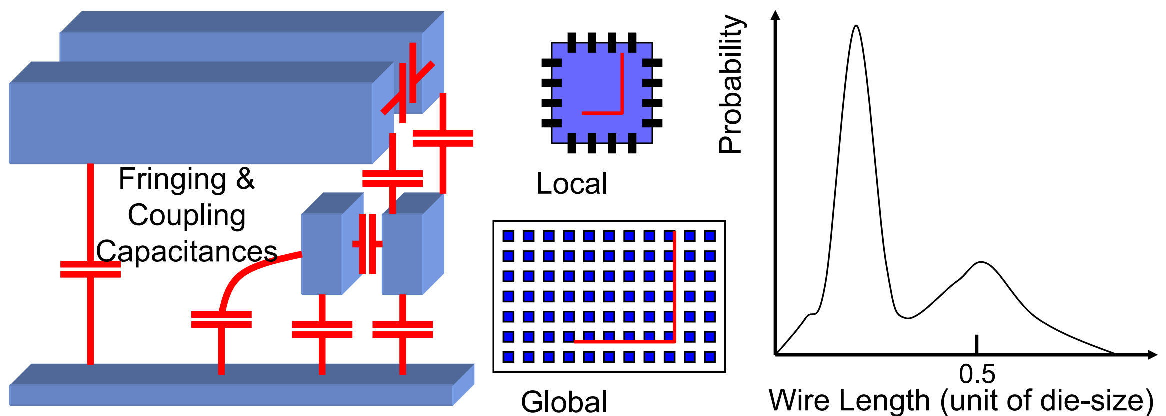
Source: B. Yu



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

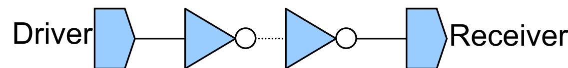
Dielectrics and junctions with increasing tunneling

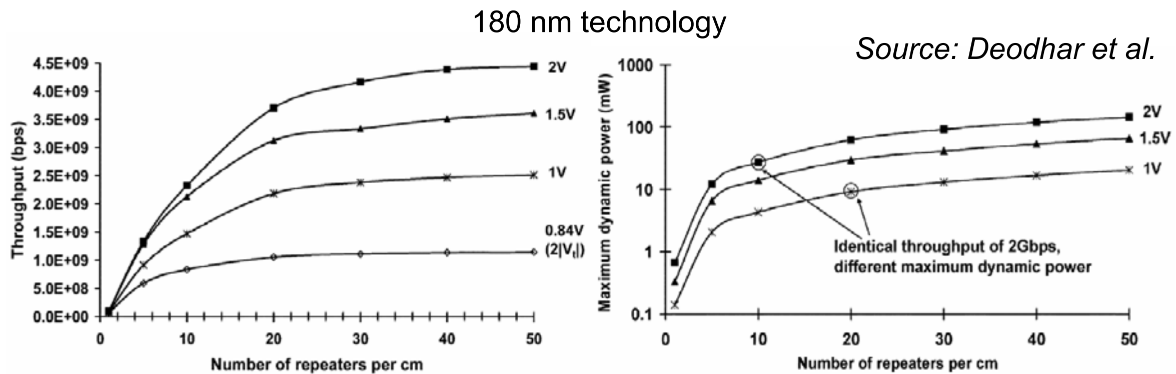
Interconnects: 2D



- 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

Application Dependent Limits

Device type	Application	T (C)	Power (W/cm2)	VDD (V)	Ioff (nA/um)	VTn (mV)	toxTeq (nm)	Wmin (nm)	Lnom (nm)
Bulk	High Performance	85	1000-30	0.8-1.2	1000-110	140-235	1.0-1.3	8.5-11	14.5-19
		-40	1000-30	0.7-1.0	1000-115	95-150	1.0-1.2	8-10	13.5-18
		-170	1000-30	0.5	1000-155	50-75	0.9-1.1	8	14
	...burn-in limited	140	-	1.8	1000	180	1.3	13.5	23
Bulk	Medium-High Performance	85	30-5	0.8-1.2	120-20	235-300	1.2-1.5	10-14	17-24
Bulk	Moderate Performance	85	5-0.5	0.6-1.0	25-2	300-390	1.3-1.6	10-14	17-24
Bulk	Low Power	65	0.5-0.001	0.7-0.9	1.0-0.01	410-550	1.7-2.0	13-17	22-29
Bulk	Ultra-Low Power	40	<0.001	0.7-1.0	<0.008	550-710	2.1-2.6	16-22	27-38
Bulk	Moderate Perf. SRAM	85	5-1	0.9-1.2	20-4	300-360	1.4-1.6	12-15	20-26
	Low Power SRAM	65	0.1-0.01	0.9-1.2	0.5-0.05	425-510	1.7-2.0	15-19	25-32
	Ultra-Low Power SRAM	40	0.0001	1.2	0.0006	635	2.4	23	39
Bulk	DRAM - metal gate	85	-	1.0	0.0001	790	2.5	28	49
	DRAM - neg. wordline	85	-	1.0	0.0001	250	2.5	28	49
DG-FET	High Performance	85	1000-30	0.8-1.2	1000-75	155-255	1.0-1.3	5-6	13
		-40	1000-30	0.7-1.0	1000-85	100-160	1.0-1.3	5-6	13
		-170	1000-30	0.5	1000-155	45-65	1.0-1.2	6-7	14
DG-FET	Medium-High Performance	85	30-5	0.8-1.2	90-15	245-305	1.3-1.6	5-7	13-17
DG-FET	Moderate Performance	85	5-0.5	0.6-1.0	20-2	300-390	1.3-1.7	5-6	13-16
DG-FET	Low Power	65	0.5-0.001	0.7-0.9	0.7-0.007	420-510	1.7-2.1	5-8	14-20
DG-FET	Ultra-Low Power	40	<0.001	0.7	<0.005	530-660	2.1-2.5	5-8	16-22
		40	<0.001	1.0	<0.005	515-645	2.2-2.6	11-15	25-33
DG-FET	Moderate Perf. SRAM	85	5-1	0.9-1.2	14-3	315-355	1.5-1.7	5-9	13-20
	Low Power SRAM	65	0.1-0.01	0.9-1.2	0.3-0.04	425-475	1.8-2.1	6-13	15-27
	Ultra-Low Power SRAM	40	0.0001	1.2	0.0006	570	2.5	17	36

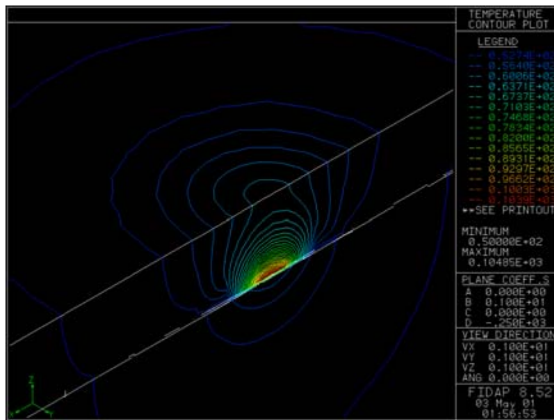
red = constrained by source-drain tunneling, blue = constrained by functionality, green = constrained by noise margin

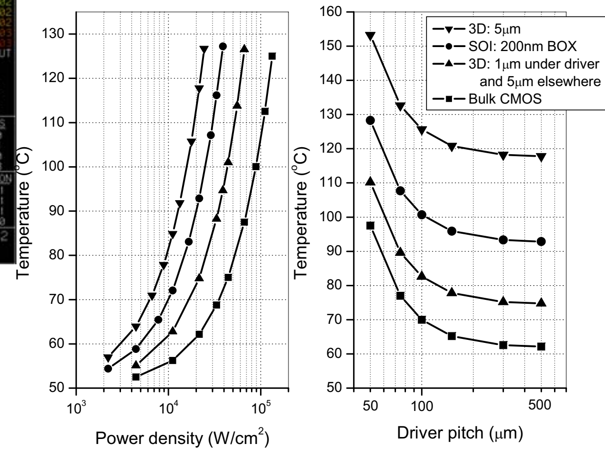
Source: D. Frank et al. IEDM (1998)

Power Dissipation in Small Dimensions & Temperature

- 10^5 W/cm^2 => 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

Higher body doping	Lower carrier mobility Higher junction capacitance Higher junction leakage
Thinner gate dielectric	Higher leakage
Shallower junctions	Higher resistance

So, there are always compromises to be made

Statics and Dynamics

Electrostatics

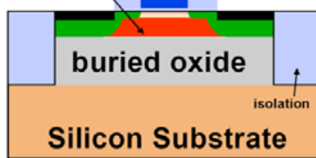
- Gate Control
  - Gate dielectrics, work-functions, …
- Substrates
  - Sharp halo's and improved junctions
  - Thin silicon bodies
- Threshold
  - Work functions, doping, new geometries

Electrodynamics

- Transport
  - Strained materials (Si, SiGe, …), new orientations, new materials (Nanotubes, Ge, III-V, …)
- Parasitics
  - New contact materials, raised source-drain structures, etc.

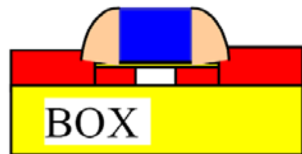
Non-Classical CMOS (Single Gate)

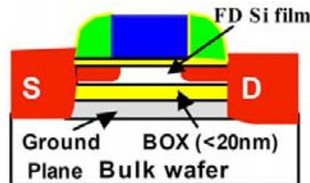
Transport enhancement



Strained Si, Ge, SiGe, SiCGe, …on bulk Si & SOI

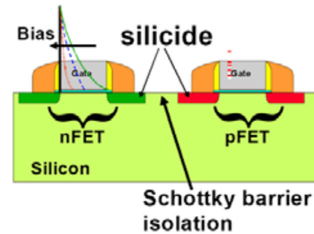
Substrate Enhancement: Thin Body

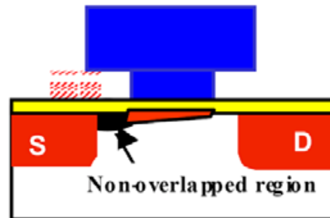




Fully depleted SOI with sub-10 nm body or ultra-thin channel and buried oxide

Source/Drain Enhancement



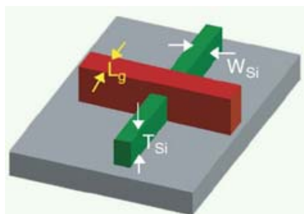


Low barrier Schottky source/drain or non-overlapped extensions

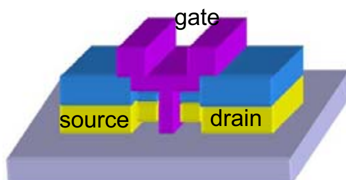
Source: ITRS

Non-Classical CMOS (Multi-Gate)

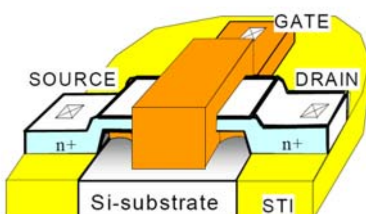
Tied Gates



Channels on multiple surfaces

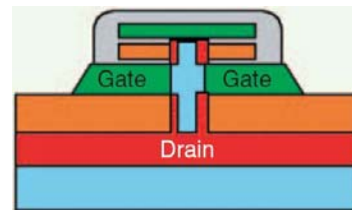


Channels on side-walls

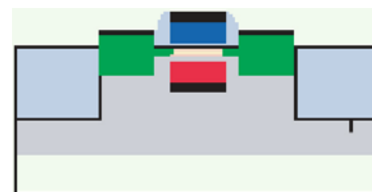


Channels on planar surfaces

Vertical transistor

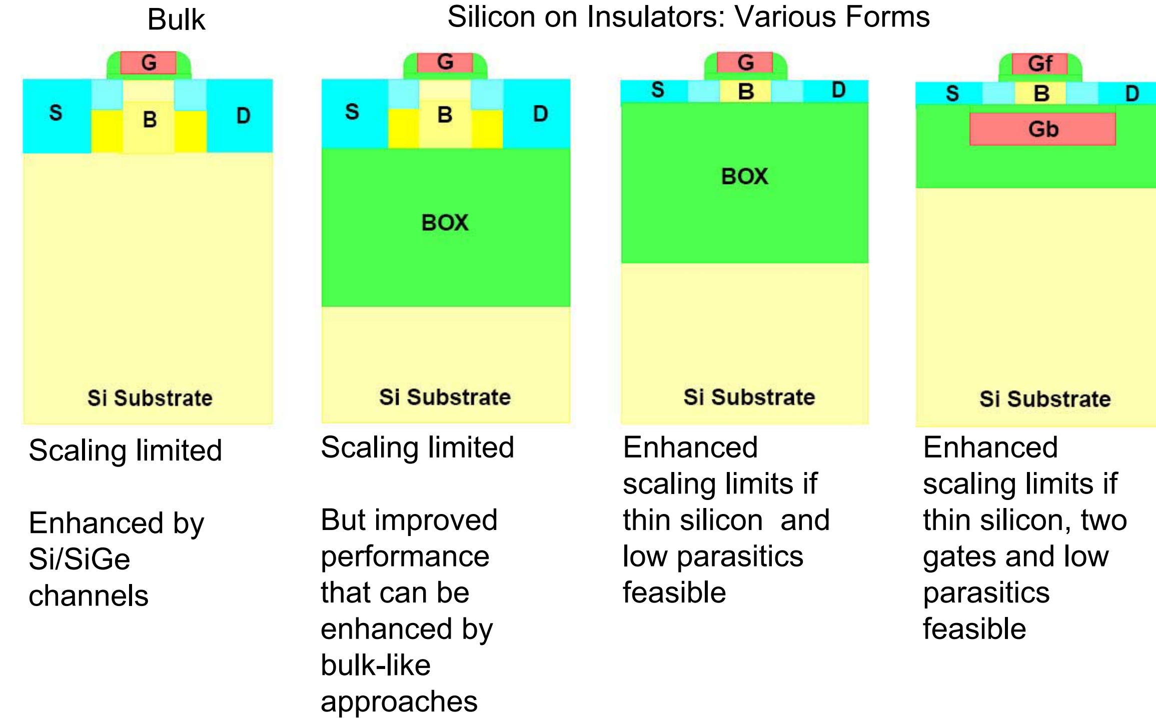


Independent gates



Source: ITRS

Planar Transistors



Bulk: Scaling limited

Enhanced by Si/SiGe channels

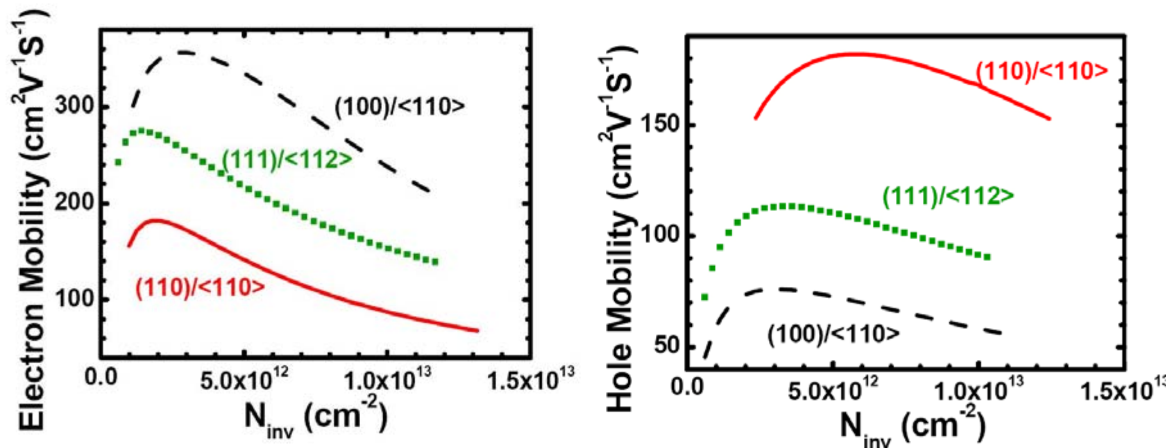
Scaling limited

But improved performance that can be enhanced by bulk-like approaches

Enhanced scaling limits if thin silicon and low parasitics feasible

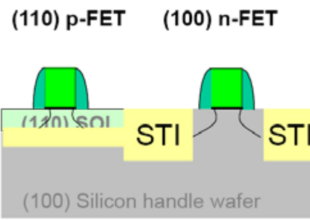
Enhanced scaling limits if thin silicon, two gates and low parasitics feasible

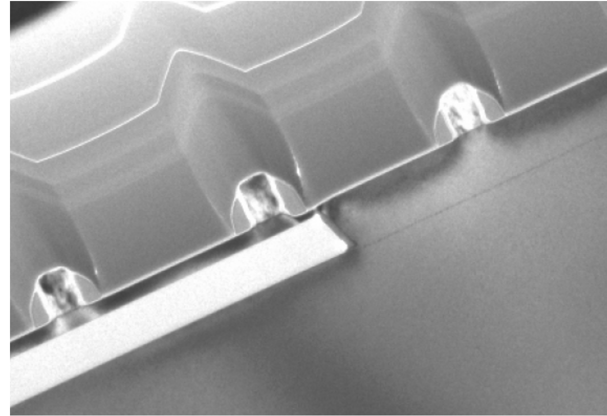
Strain by Orientation

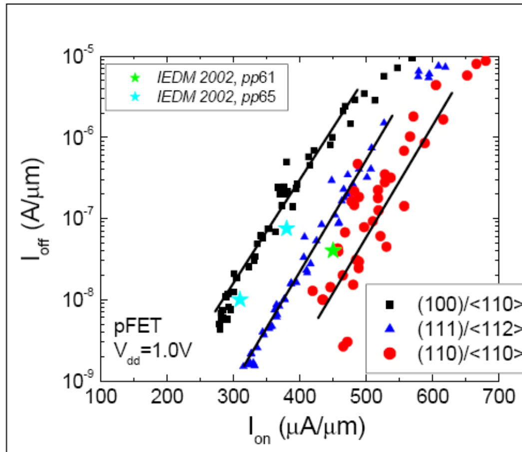


Yang et al. (2003)

Transport Improvement by Orientations



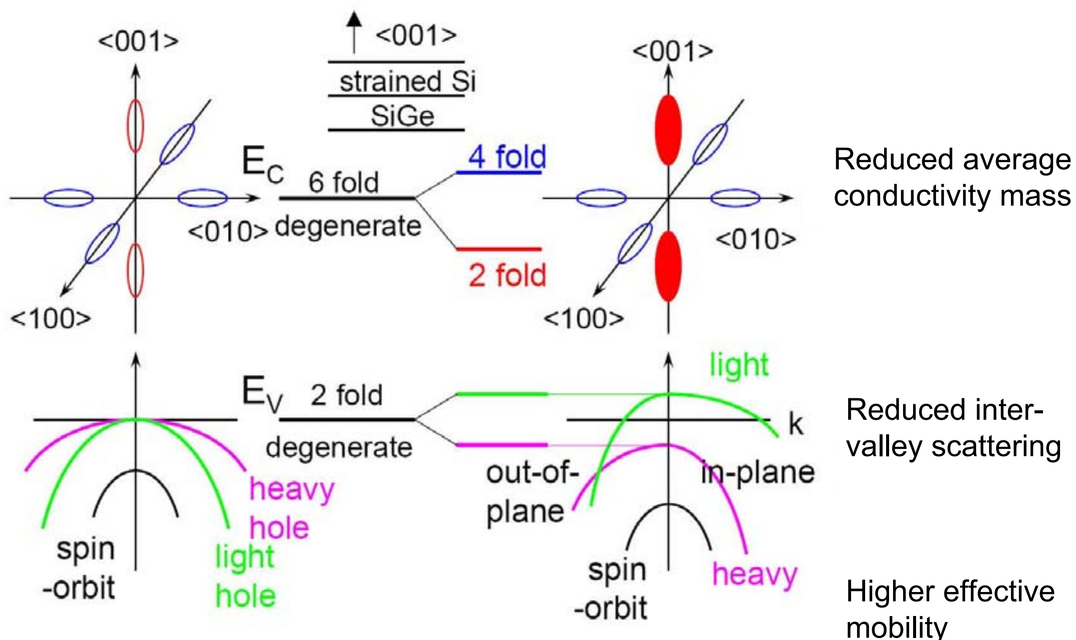




Bond (110) onto (100) Si – oxide based
Use epitaxy of 100 with oxide isolation for (110)

Yang et al. (IBM)

Strain



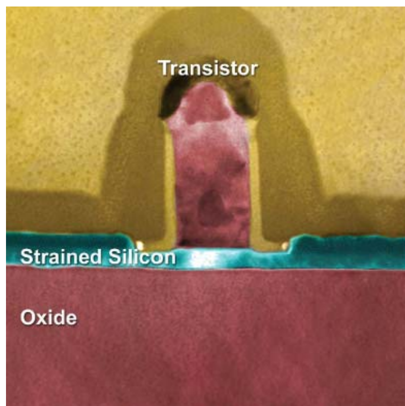
Reduced average conductivity mass

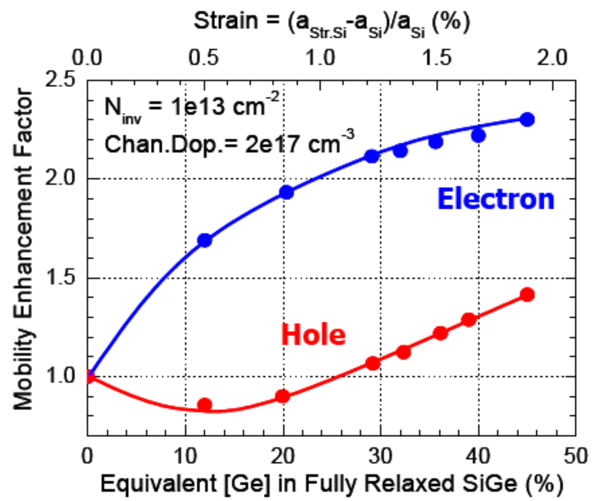
Reduced inter-valley scattering

Higher effective mobility

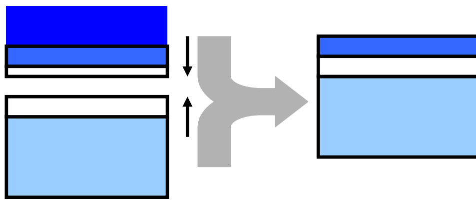
Takagi (2003)

Strained Si





Strained Si on SiGe



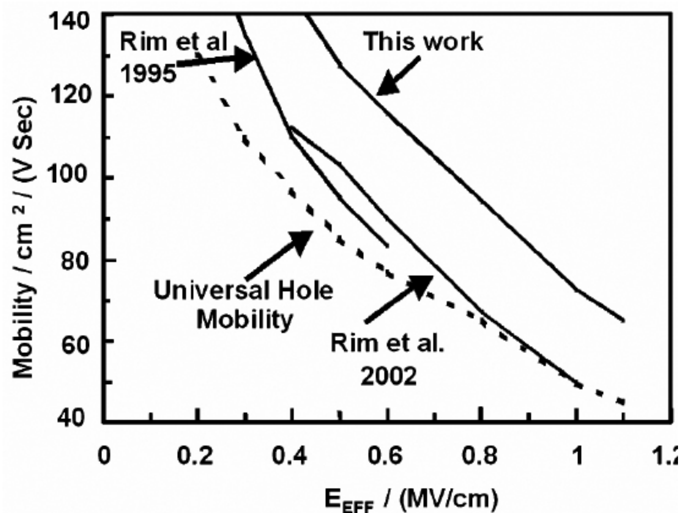
Oxidized Si

Mobility Enhancements with Silicon(strained) on Insulator

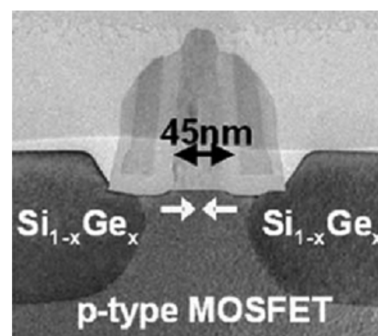
K. Rim (IBM)

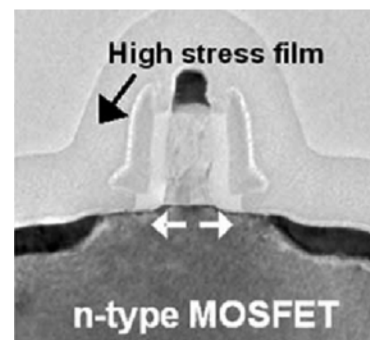
Strain by Process

Uniaxial Strain

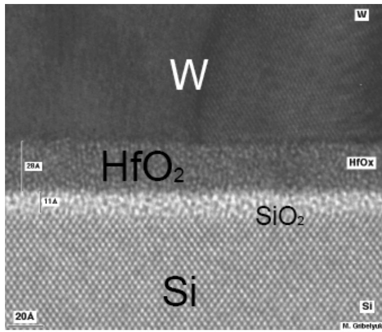


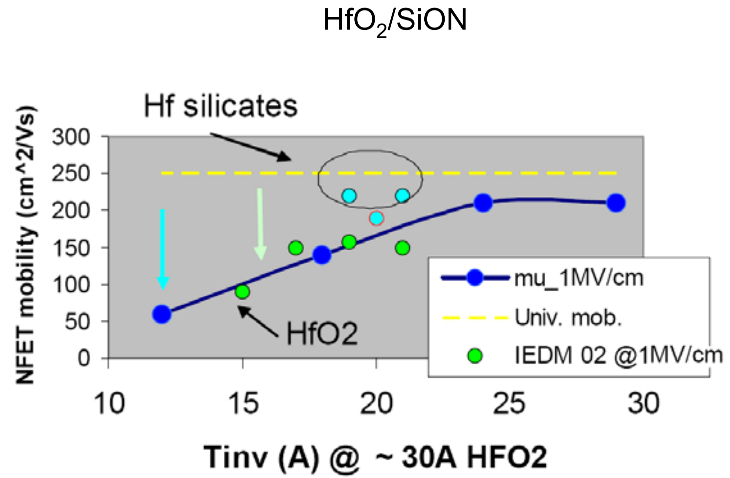
Thompson et al. (2004)





High k (Permittivity)

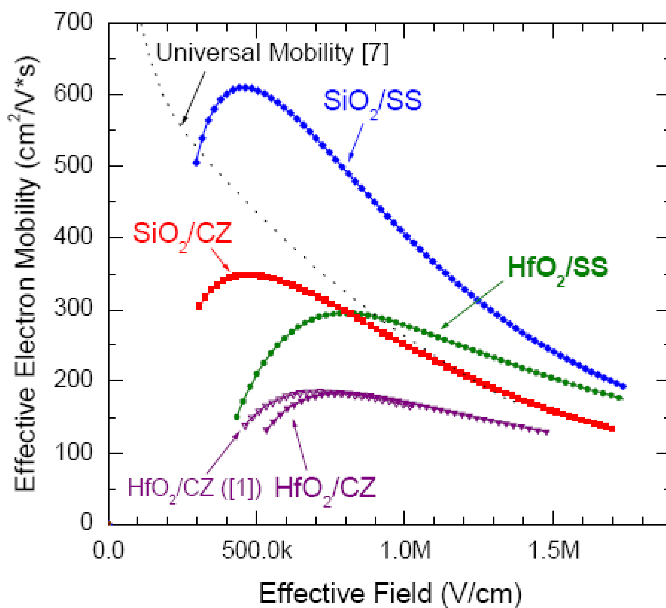


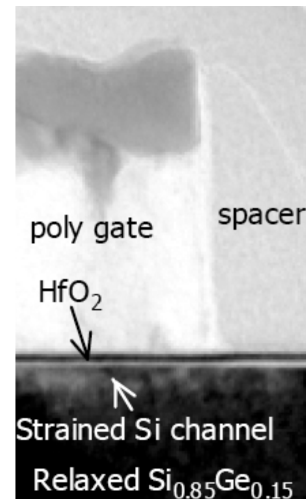


HfO_2 has x10^4 less leakage than equivalent SiO_2
But, large interface state issues

Gousev et al. (IBM)

Combinations

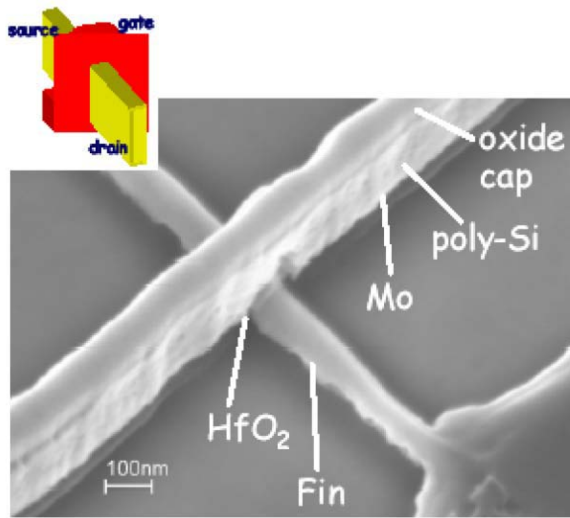


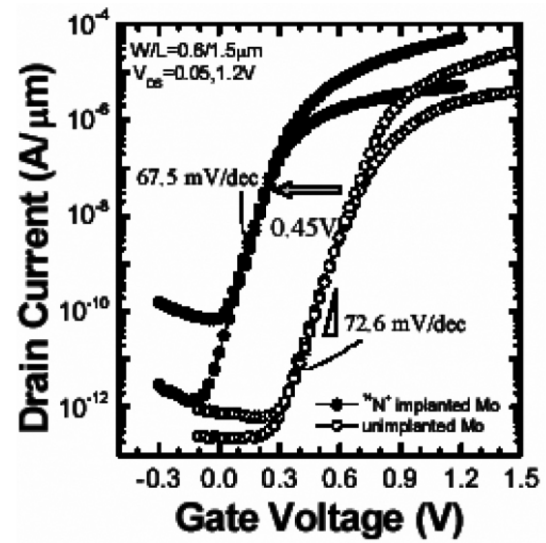


To date, mobility degradion with high permittivity materials is substantial

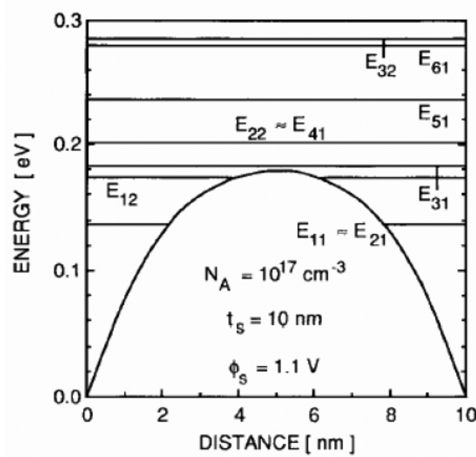
Rim (IBM)

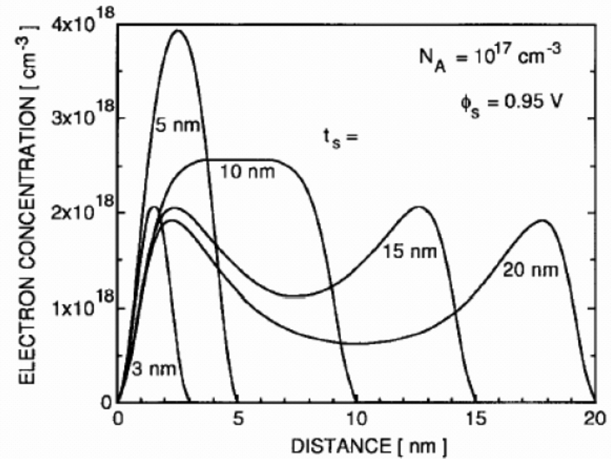
FinFet





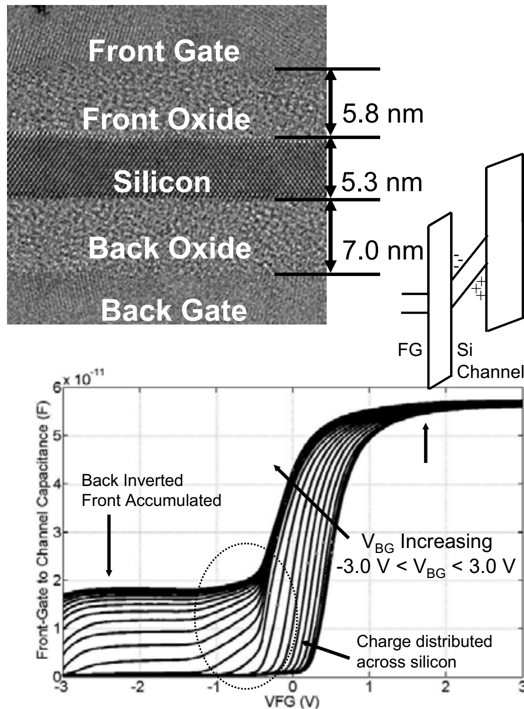
Thin Si

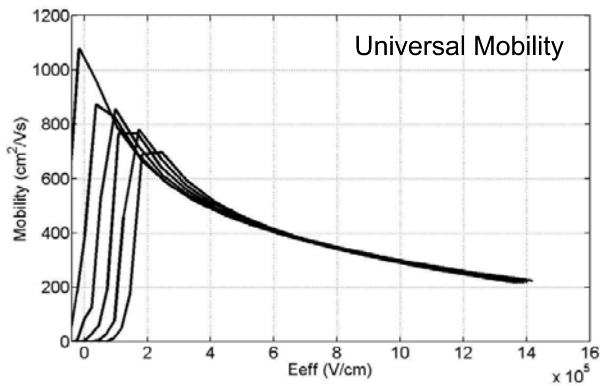




Majkusiak (1998)

Transport in Thin Silicon



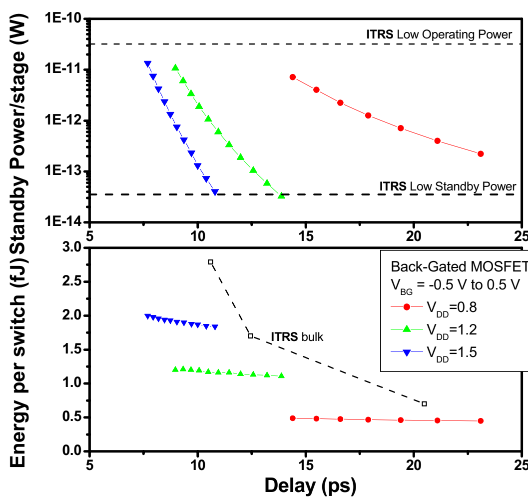


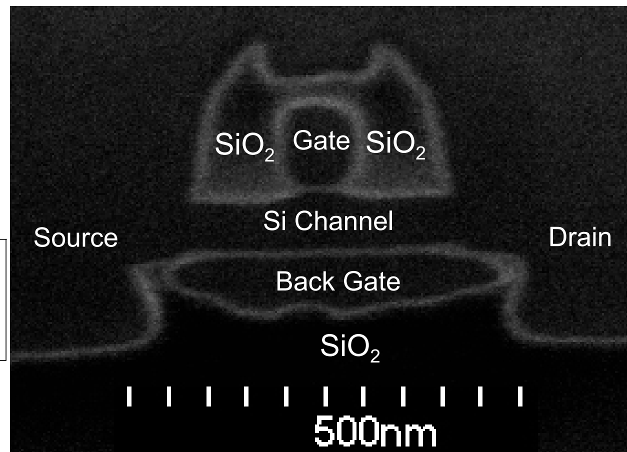
$$E_{eff-front} = \left(\frac{Q_F C_{oxB} + C_{oxF} C_{oxB} (V_{FG} - V_{FBF} - V_{BG} + V_{FBB})}{C_{oxF} C_{oxB} + C_{oxF} C_{Si} + C_{oxB} C_{Si}} \cdot \frac{1}{t_{Si}} \right) + \frac{Q_F}{2\varepsilon_{Si}}$$

Good electron transport still maintained in thin silicon

A. Kumar, et al. (2005)

Power: Switching and Standby and Adaptive Control

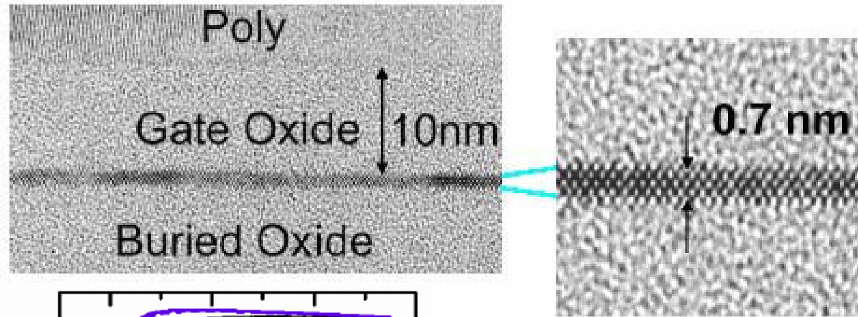


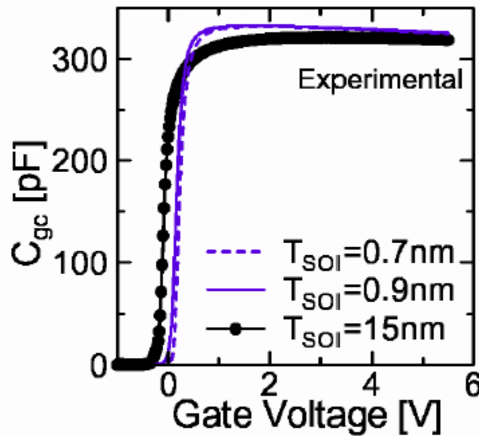


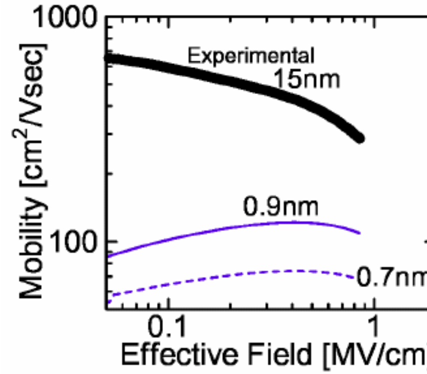
- L = 90 nm with 2 nm front oxide, 5 nm back oxide, 25 nm Si, and using 21 stage ring oscillator
- Devices provide tuning of standby power and switching performance with good noise margin

Source: Avci et al. (2005) & Lin (2006)

Thin Si



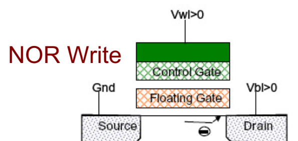




Uchida et al., IEDM (2004)

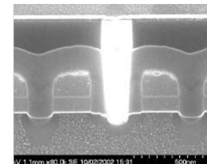
Flash Non-Volatile Memories

NOR Write

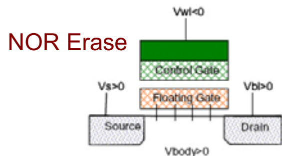


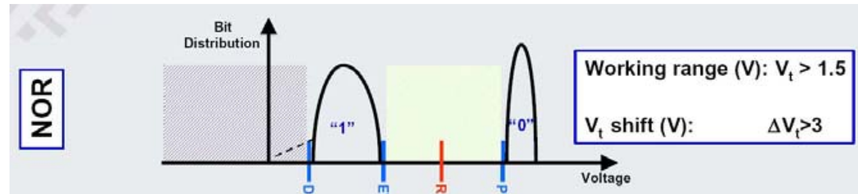
Programming: *Channel Hot Electron*
Vwl: 8-10 V, Vbl: 4-5 V
T_{pulse}: 1 us, I: 10-100 uA
0.5 MB/s

Erasing: *FN Tunneling*
Vwl: -8 V, Vbody: 6-8 V
T_{pulse}: 100 ms, I: ~0 uA

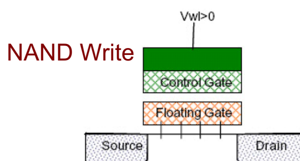


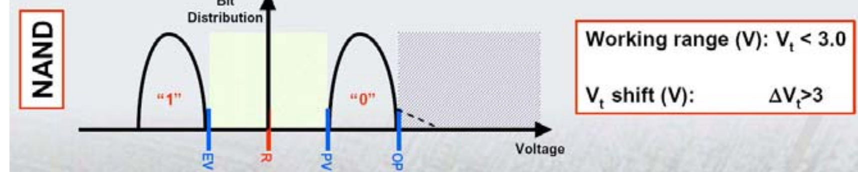
NOR Erase



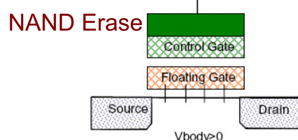


NAND Write



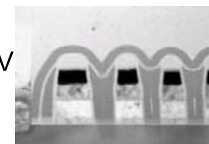


NAND Erase



Programming: FN tunneling
Vwl: 18-20 V, Vbody: 0 V
T_{pulse}: 300 us, I: ~0 uA
7-10 MB/s

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

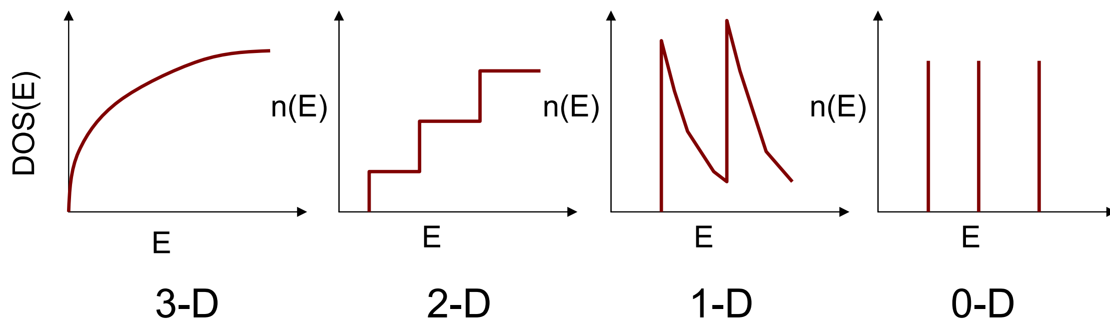


Source: R. Bez

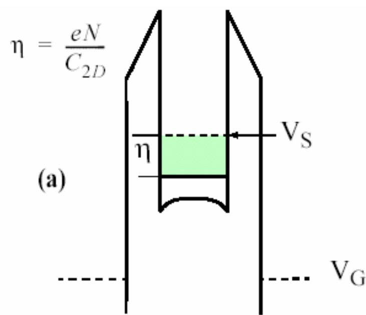
Nanoscale in Silicon

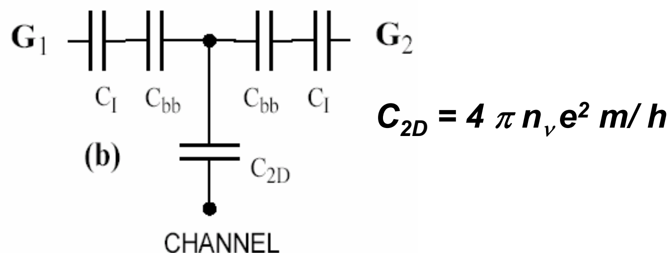
Nanoscale Classical Picture

- Achieving quantum confinement
  - Quantum wells (2D)
  - Quantum wires (1D)
  - Quantum dots (0D)



Confinement & Degeneracy





Solomon & Laux, IEDM 2001

- 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

$$C_{2D} = 4\,\pi\, n_\nu e^2\, m / h$$

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

$$\exp\left(-2\int \alpha dx\right)$$

  - 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

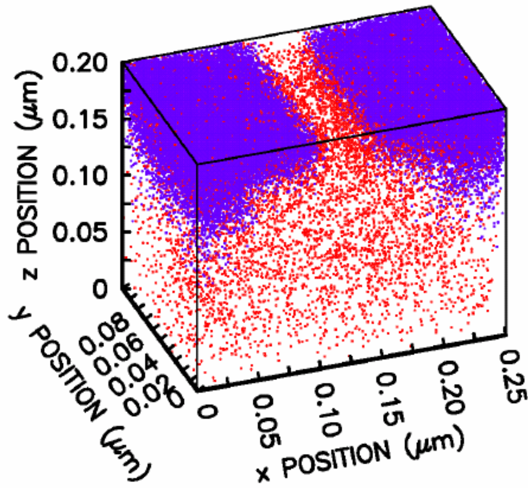
Nanoscale: Power and Performance

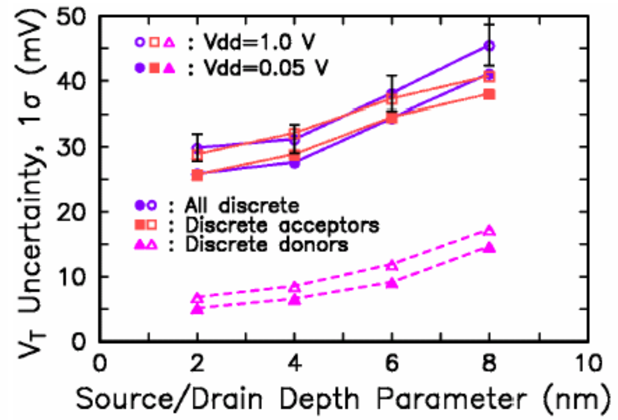
- Suppose we could make devices at a 10 nm x 20 nm minimum dimension with a cell size of 50 nm x 60 nm (3.3×10^{10} cm^{-2})
- And, suppose we limit the power density to 100 W/cm^2 and 1 V supply
- If all elements were continuously switching the average power per device is 3.33 nW/device at 6 nA/device, 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 μA

An Electron in a Semiconductor

- Unhindered movement of a single electron is μA'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

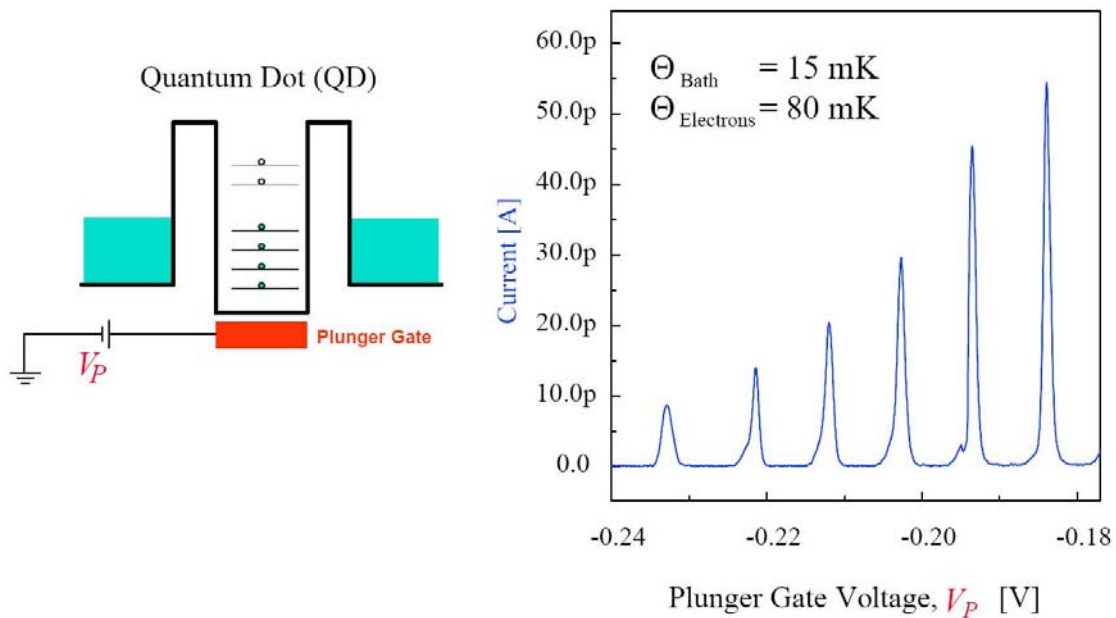
Discreteness





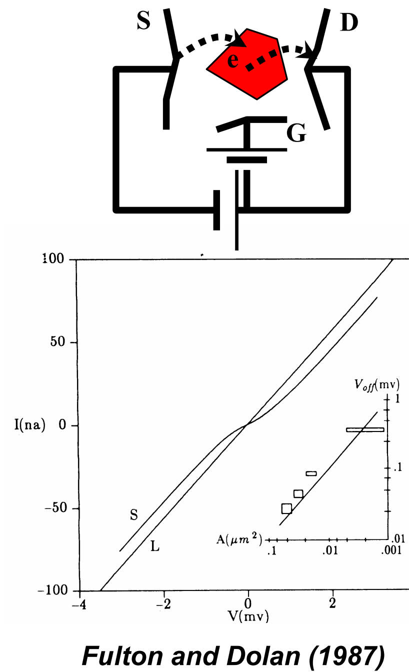
Frank et al. (2000)

Quantum Dots – Single Electron Effects



Charging Effects

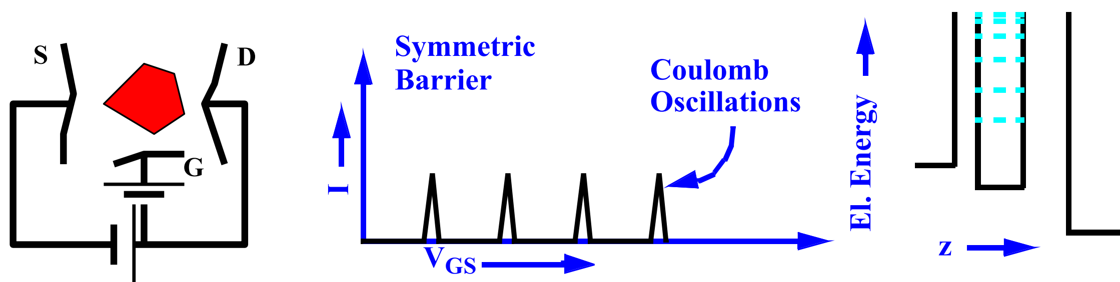
- Charging of a small particle with an extra electron requires an energy: $E_C = e^2/2C_\Sigma$
- A small particle (~10 nm) in a dielectric (SiO_2, e.g.) has C_Σ=2 aF, E_C = 40 meV ~150 C
- Observations by Neugebauer and Webb (1962), Zeller and Giaver (1969) and Lambe and Jaklevic (1969)



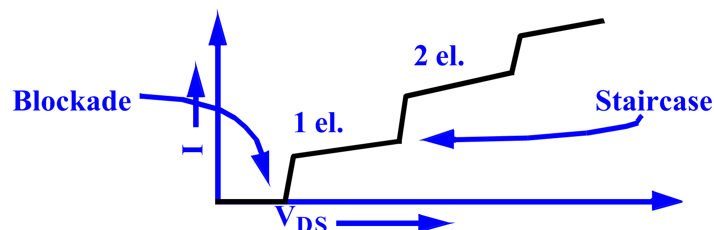
Fulton and Dolan (1987)

Coulomb Blockade and Staircase

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



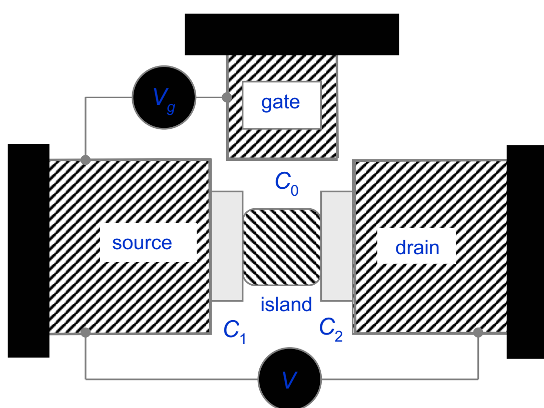
- Staircase: When particle charged by > 1 electron

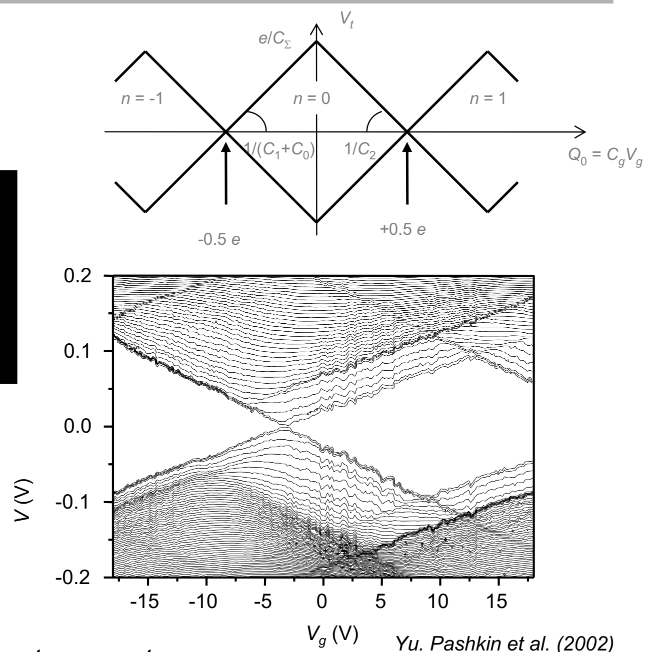


Impedance, Currents and Size Effects

- 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>> $h/2\pi e^2$ or 4.1 kΩ
  - In real structures, R is typically GΩ
  - ⇒ 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 ~1/C and $\Delta E_c/E_c$ ~$\Delta L/L$
- Size has significant effect through subband energy
  - Sub-band Energies: ΔE_0~$1/L^2$ and ΔE_0 ~2 $\Delta L/L$

Single Electron Transistor



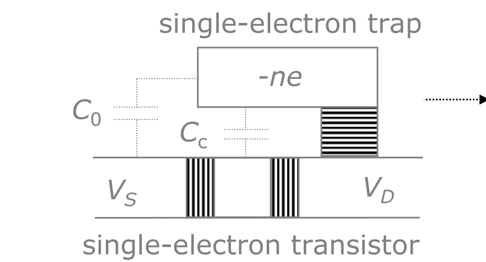


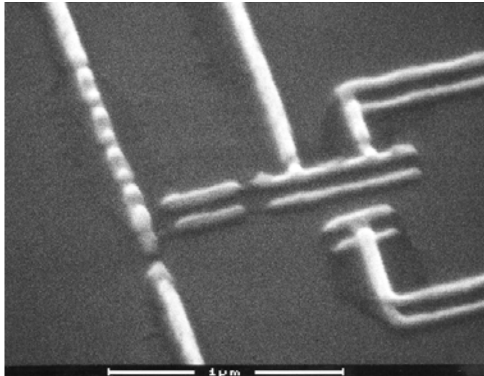
Yu. Pashkin et al. (2002)

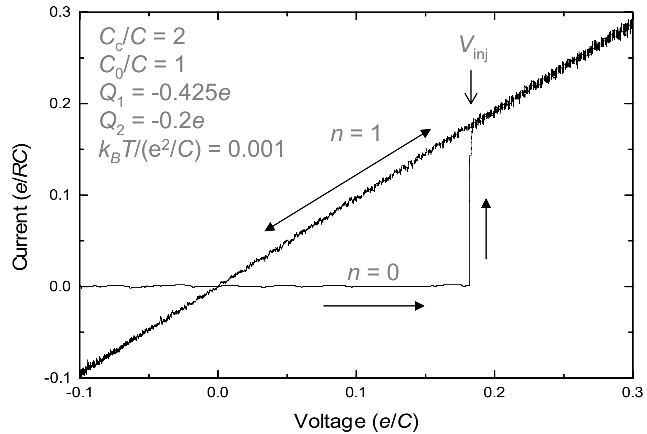
Requires atomic fab accuracy for room temperature

Low currents, but allows larger variety of materials

Single Electron Latching Switch





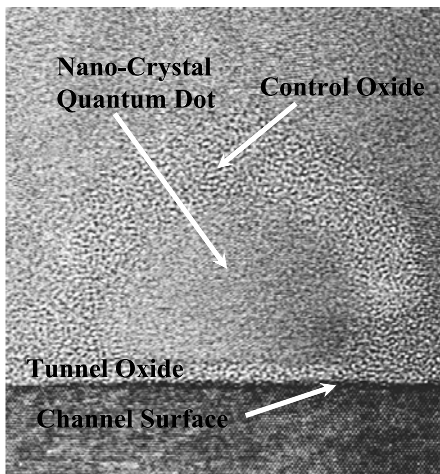


I-V curve within the "Orthodox" theory

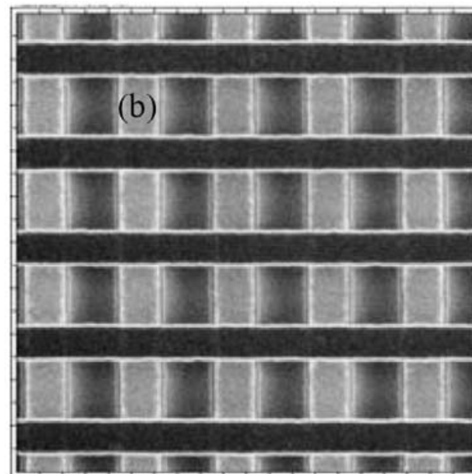
Low-temperature prototype (trapping time > 12 hrs)
P. Dresselhaus *et al*. (1994)

Nanoscale: Classical Charge Effect

If C = 1 aF, $e^2/2C$ = 160 meV
C = 1 aF is a 18 nm metal particle in free space or ~4 nm in oxide
Single electron charging occurs with blockade regions (Coulomb Blockade)



Tiwari et al. APL (1996)



Muralidhar et al., IEDM (2003)

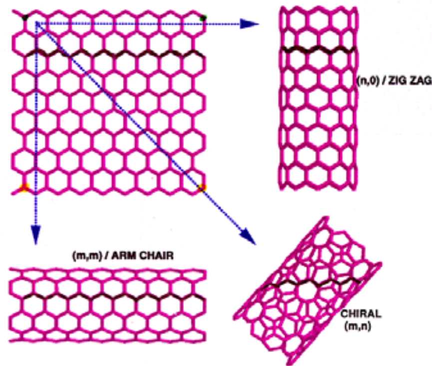
Makes low power memories possible

Nanoscale in Other Materials

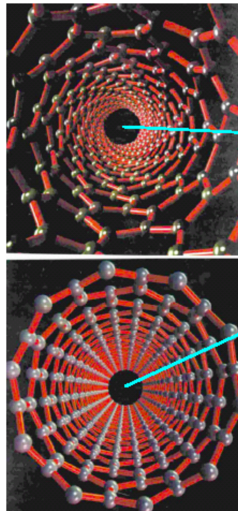
Nanotubes

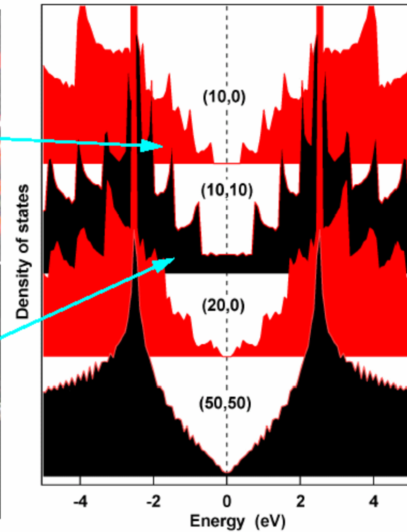
- High carrier mobility
  - Ballistic transport (<1-10μm)
  - >10,000 $cm^2/V.s$ (>10μm)
- High current carrying capabilities
  - $J=10^9$ A/cm^2 (Most metal fails at $<10^6$ A/cm^2)
- Nearly-ideal surface (!)
  - Wider choice of dielectrics
- All atoms on surface
  - Potential for sensors
- Can be direct bandgap
  - Potential for optical devices
- Diameter determines semiconducting (2/3) vs metallic tubes (1/3), and placement

Carbon Nanotubes

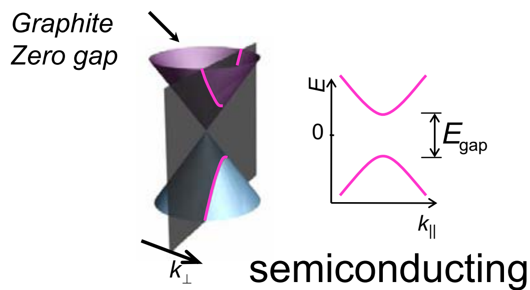


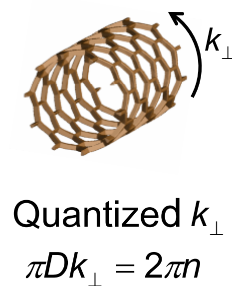
- Good transport
- But, poor control of
  - Placement
  - Thickness
  - Chirality





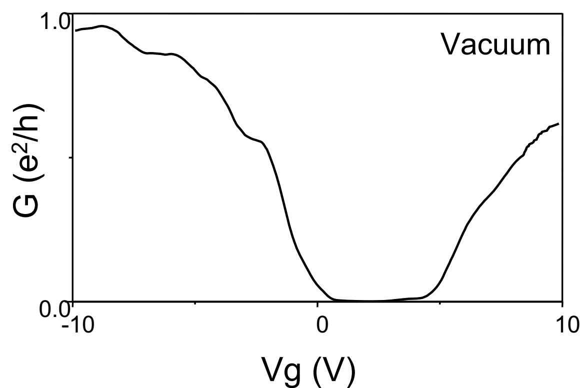
Nanotube Band Structure and Mobility

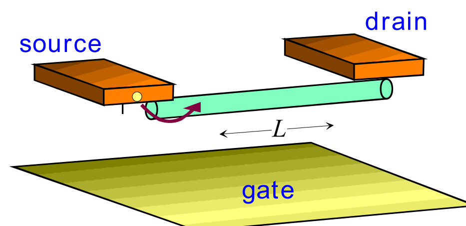




Quantized $k_{\perp}$

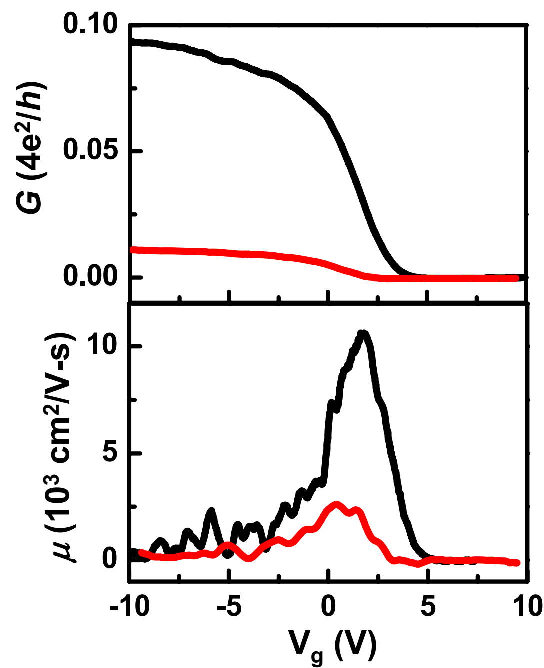
$\pi D k_{\perp} = 2\pi n$

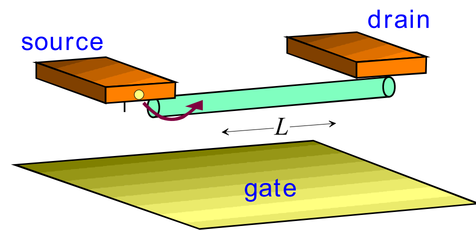




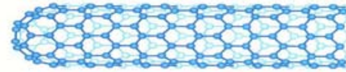
Reported max. mobilities: 1,000 - 100,000 $cm^2/V.s$

Diameter Dependence

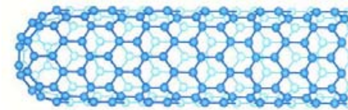




d = 1.5 nm



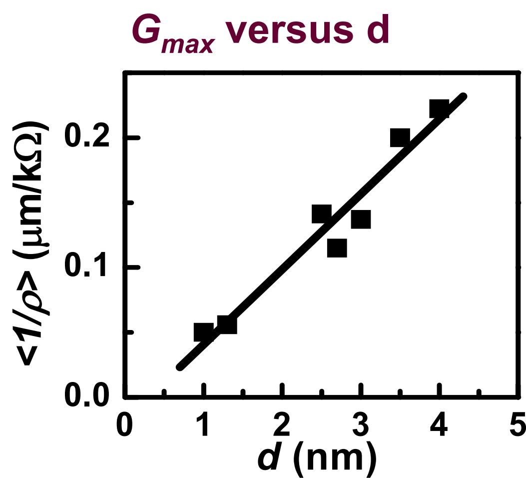
d = 3.4 nm



G increases with d

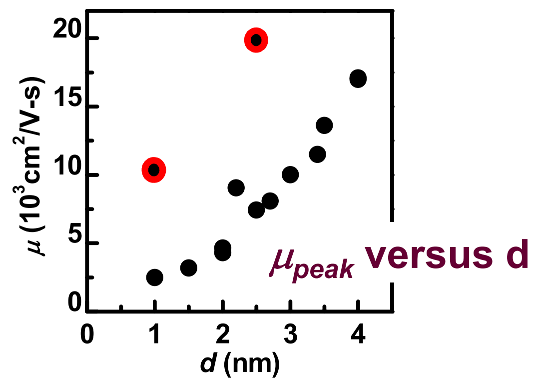
Diameter Dependence

G_{max} versus d

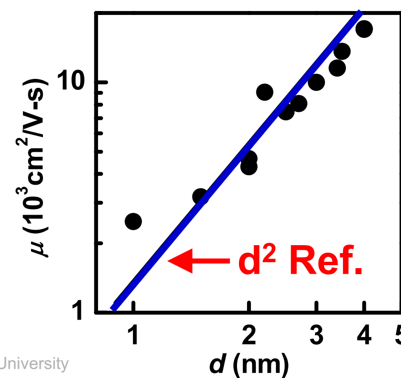


$$\langle \rho \rangle = \left(R_{300K} - R_{50K} \right) / L$$

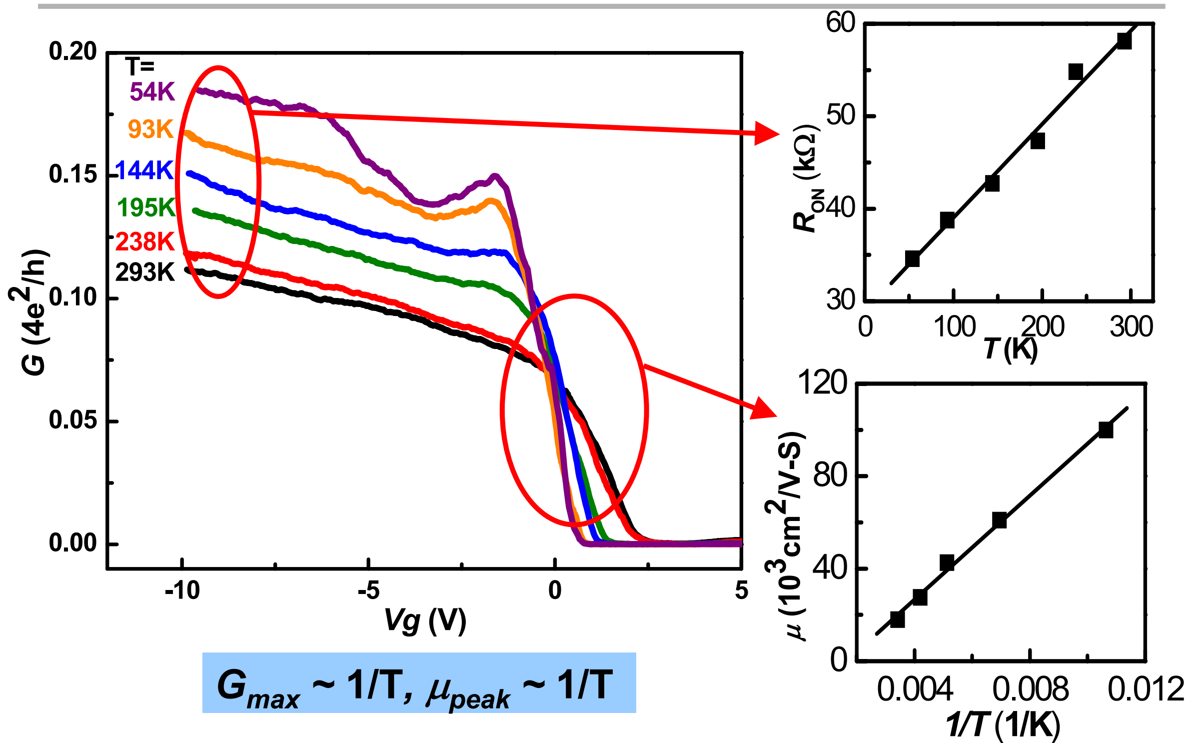
$G_{max} \sim d$, $\mu_{peak} \sim d^2$

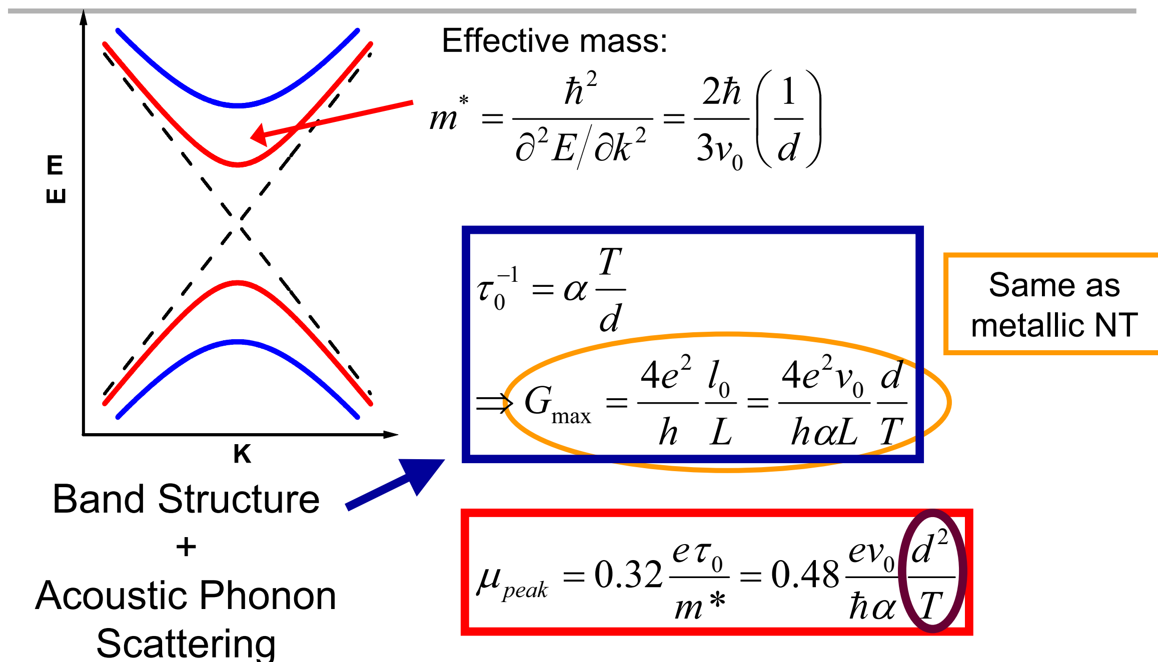


μ_{peak} versus d



Zhou *et al.*





G. Pennington *et al.*, *Phys. Rev. B* **68**, 045426 (2003)
V. Perebeinos *et al.*, *Phys. Rev. Lett.* **94**, 086802 (2005)

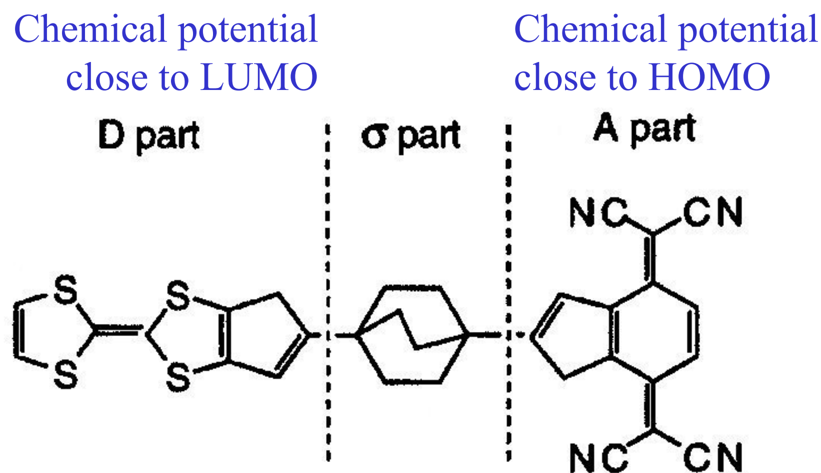
Molecules

- 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

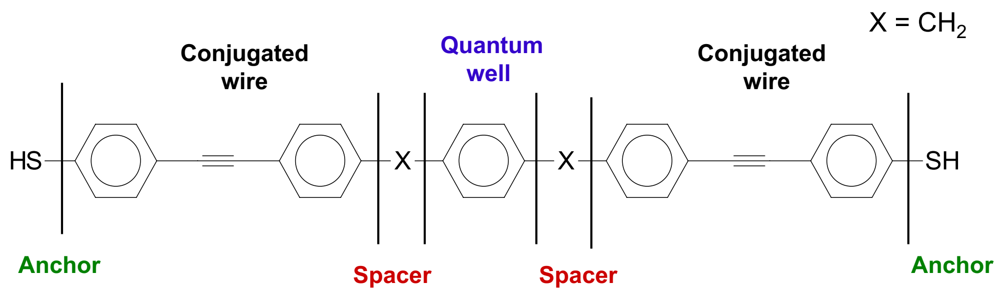
Molecular Rectifier

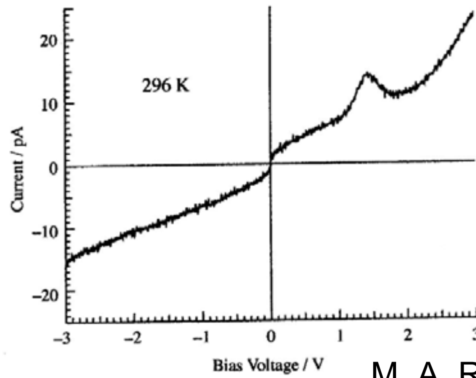
- Analogy with semiconductor diode:



Aviram and Ratner, *Chem. Phys. Lett.* **29** 277 (1974)

Molecular Resonant Tunneling Diode

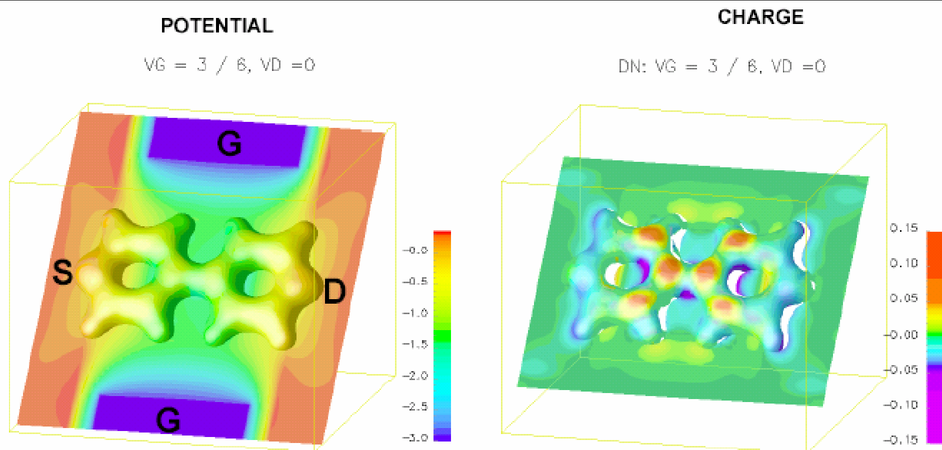




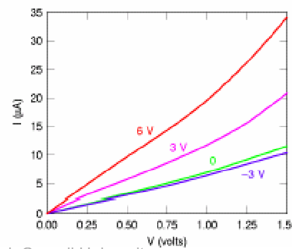
- CH_2 groups act as tunnel barriers
- Negative differential resistance (NDR)
- Like a resonant tunnelling diode (RTD)

M. A. Reed, *Proc. IEEE* **87** 652 (1999)

Molecular Transistors



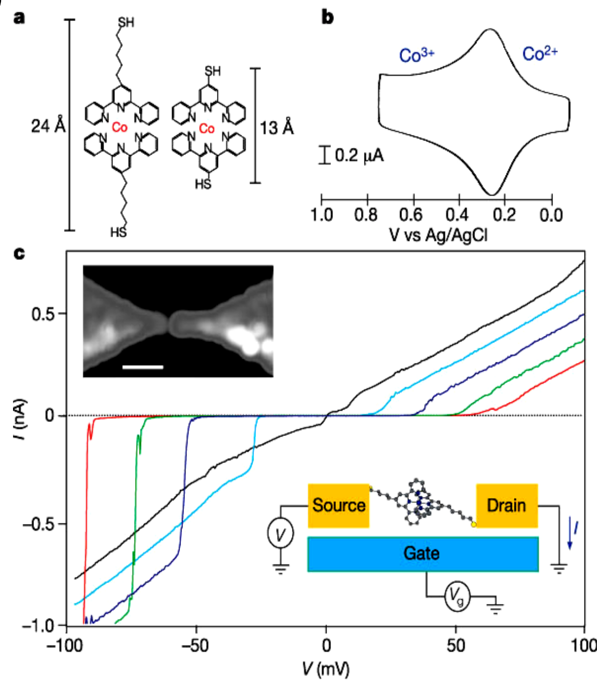
- Behavior is very complex
- Conceptual models are needed



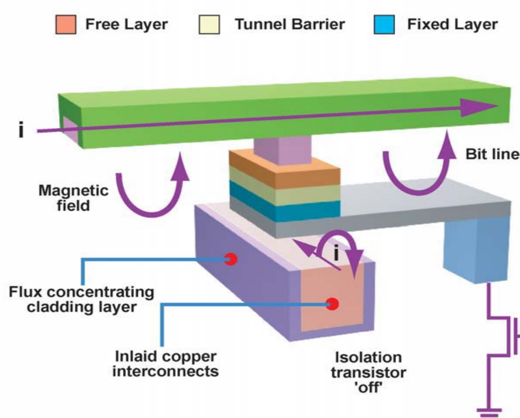
DFT Simulations: N. Lang

Single Electron Molecular Transistors

J. Park *et al*. (2002)



Magnetic RAM



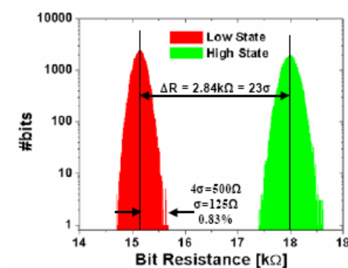
Stores information using the magnetic polarity of a thin, ferromagnetic layer.

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

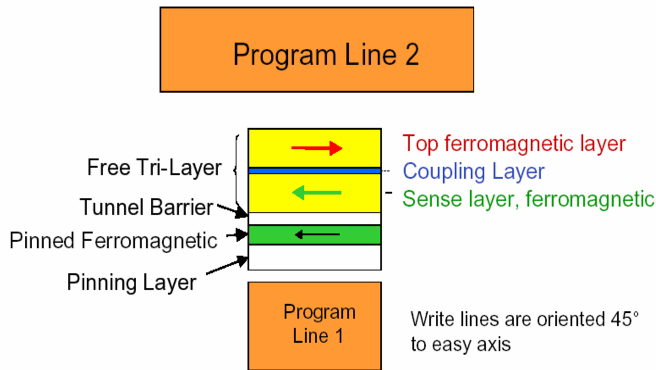
Current determined by the rate of electron quantum tunneling, which is affected by magnetic polarity of the cell.

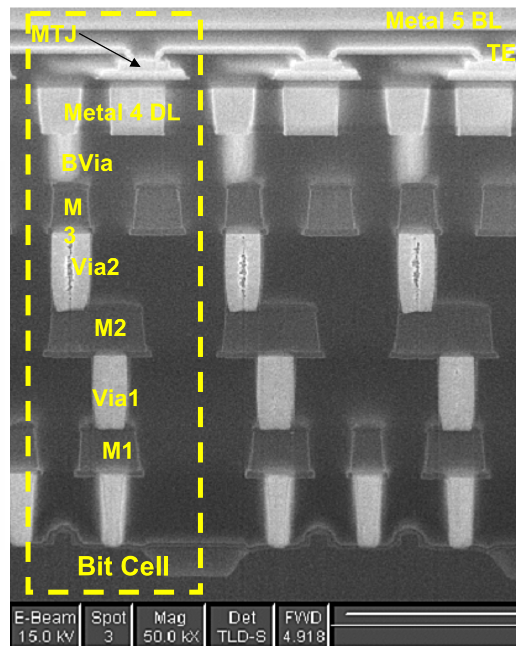
The “Free Layer” polarization is allowed to change, depending on if the cell is High or Low

The resistance across the stack is measured to determine the cell state



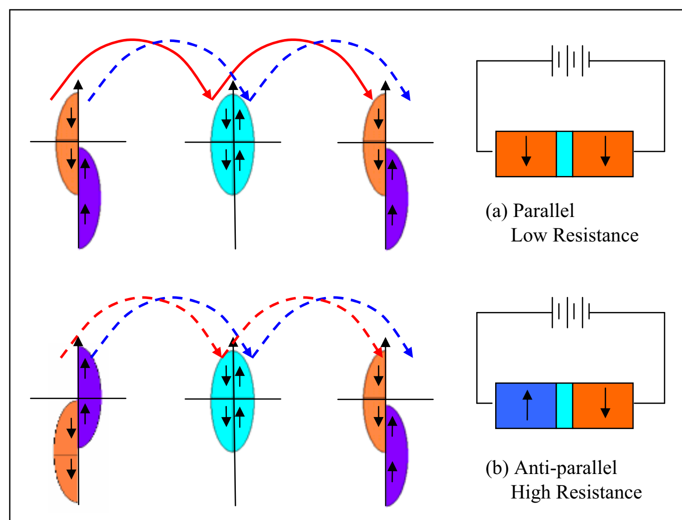
What is MRAM? How it works



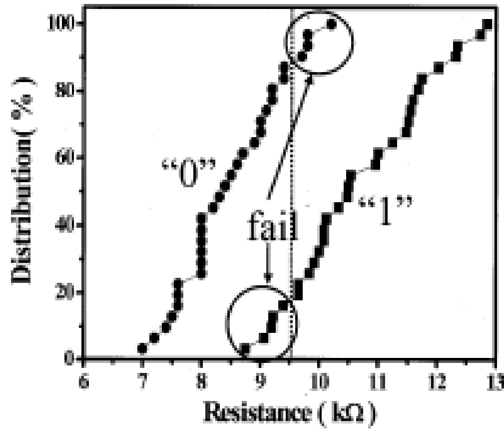


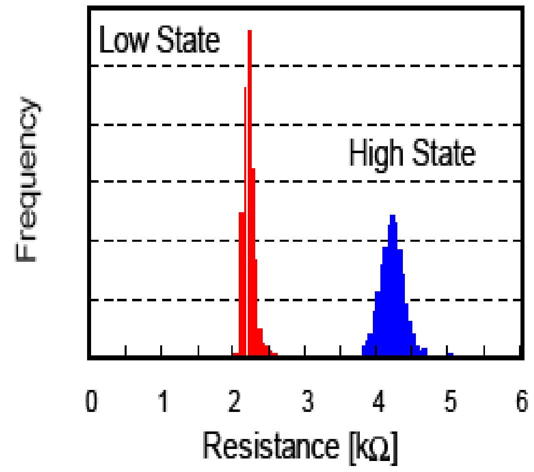
Source: Slaughter (2004)

Conductance in Magnetic Layers



Variance in Magnetic Structures



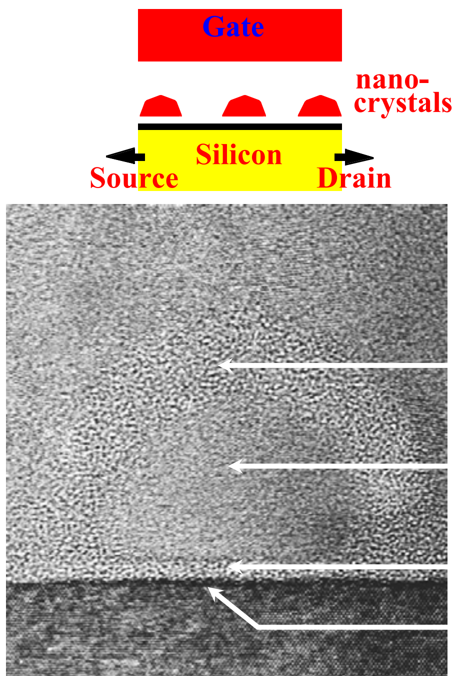


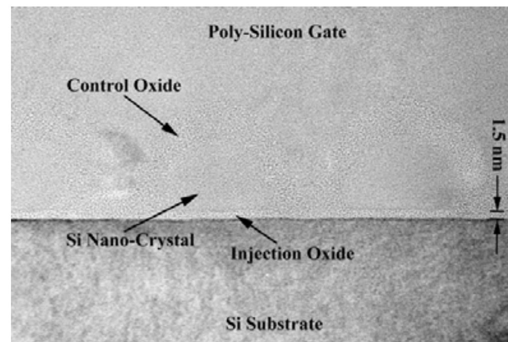
W.J Gallagher et al. IBM J. R&D (2006)

M .Hosomi et al. IEDM (2005)

Nanoscale Devices

NanoCrystal Floating-Gate Memory



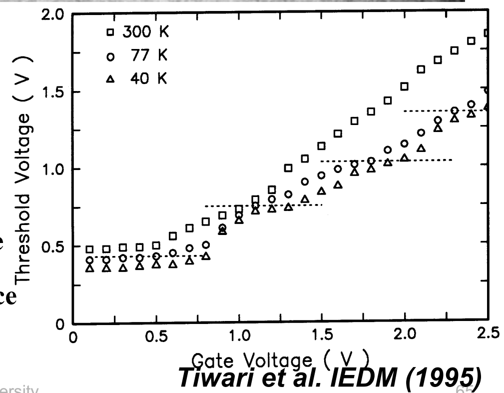


Control Oxide

Nano-Crystal
Floating Gate

Injection Oxide

Channel Surface



Tiwari et al. IEDM (1995)

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

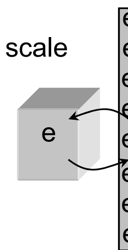
$$H_T = \sum_{n,m} T_{nm} a_n^\dagger b_m + c.c.$$

with m identifying the indices of the ladder in the quantum dot

Equation of motion for the density matrix:

$$i\hbar \frac{\partial \widehat{P}_H(t)}{\partial t} = \left[H, \widehat{P}_H(t) \right]$$

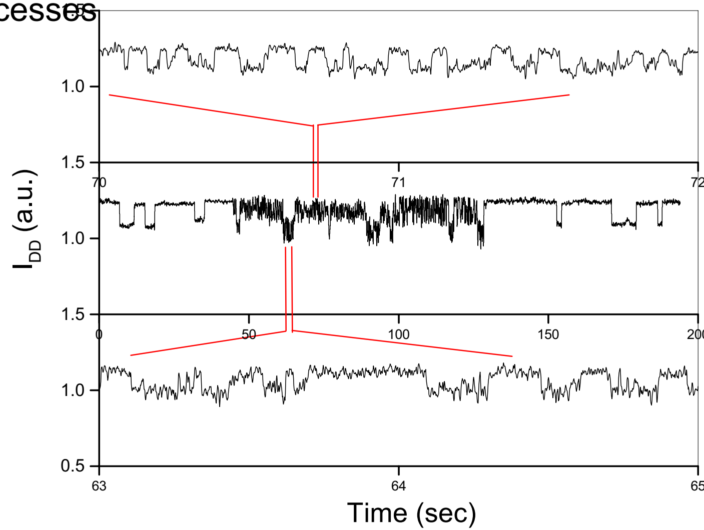
Quantum dot
5-10 nm length scale



Inversion layer
1-2 nm thick

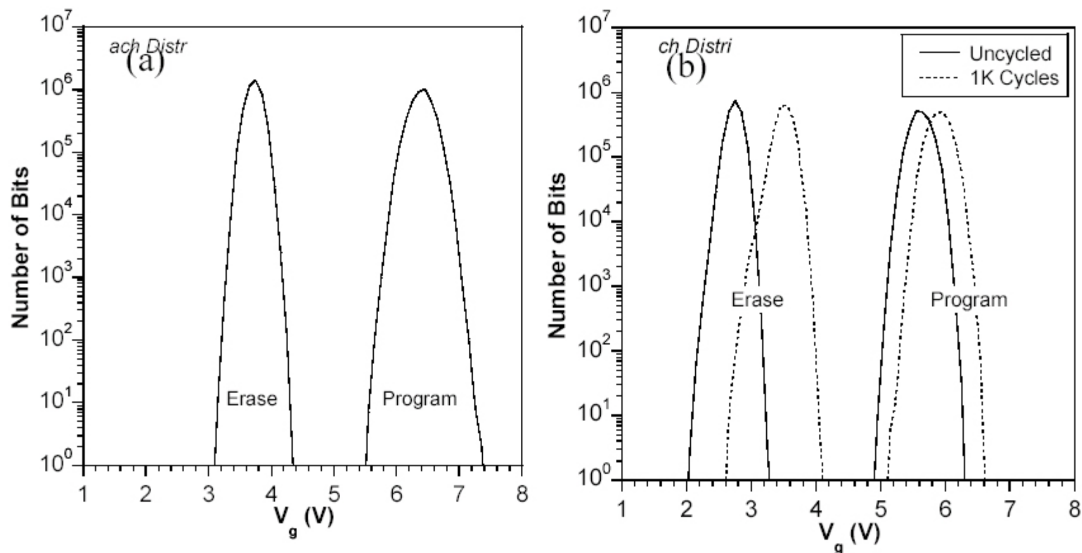
RTS: Nano-Crystal Memory Single-Electron Events

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



RTS Amplitude ~ 14%   Δt = 1 msec

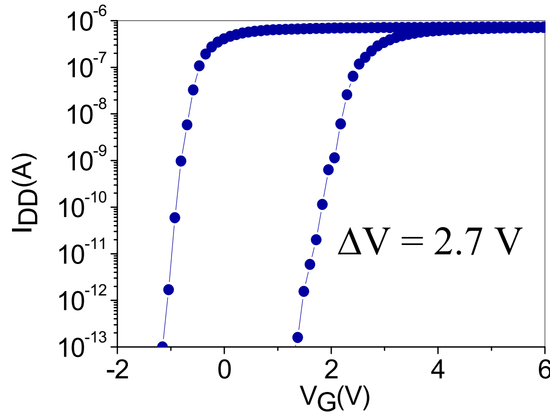
Nanocrystal Memories



4Mb Array in a 6V 90 nm process
Muralidhar et al., 2003 IEDM

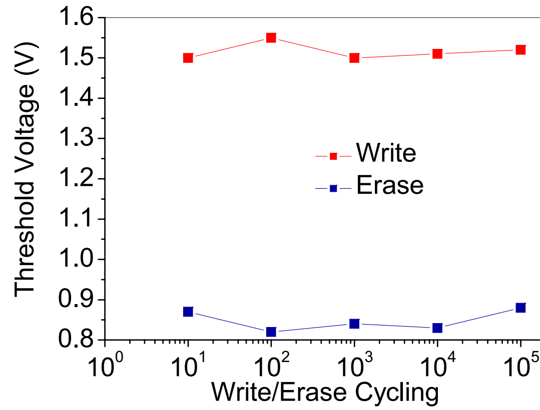
Scaled Front-Side SONOS Memories

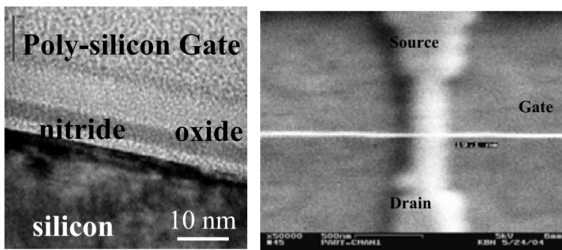
Use of higher defect density to counter statistical effects



L = 46 nm, W = 33 nm

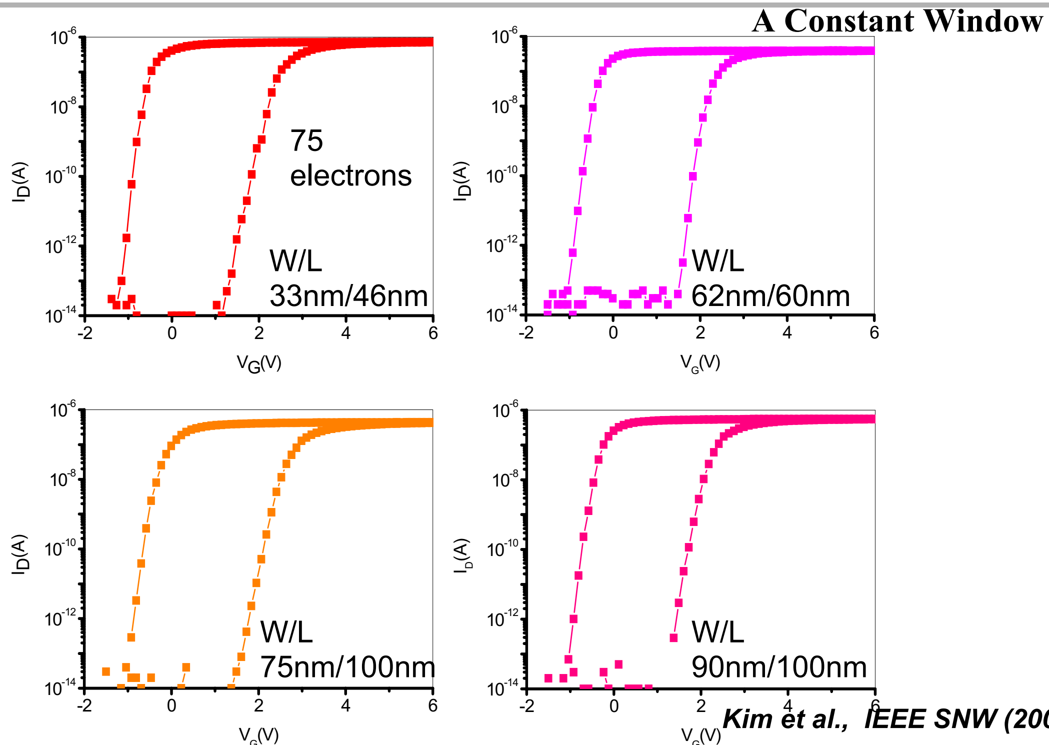
ONO stack = 2 / 6 / 12 nm





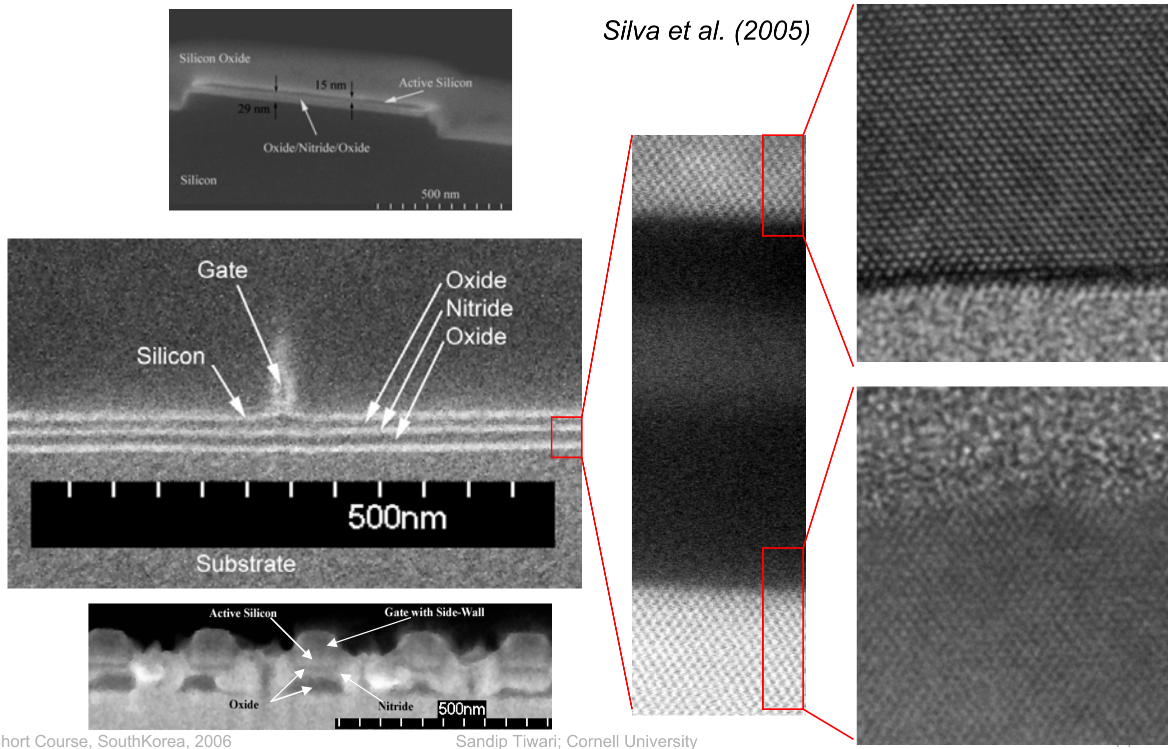
Kim et al., IEEE SNW (2003)

SONOS Memories

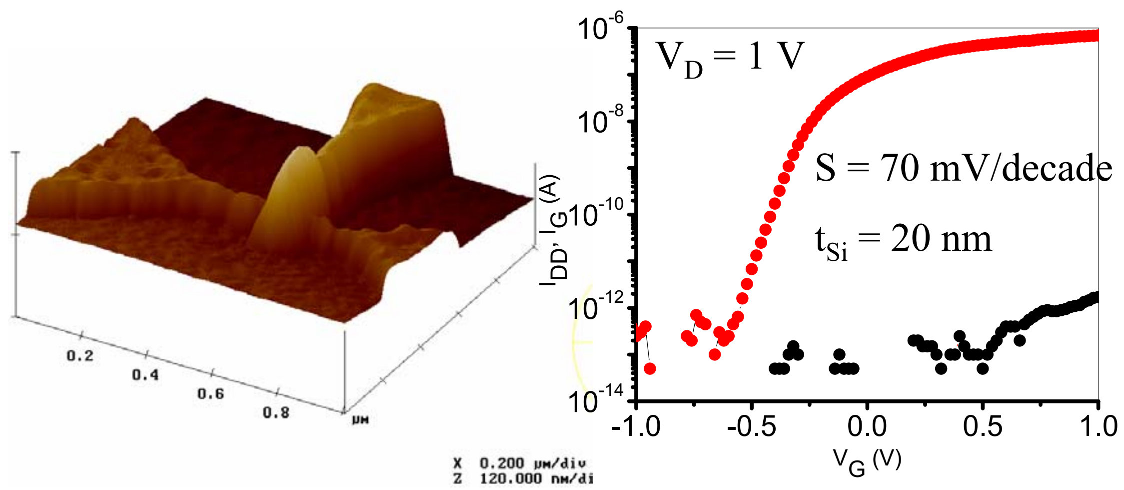


Kim et al., IEEE SNW (2003)

Memory Using Defects on Back

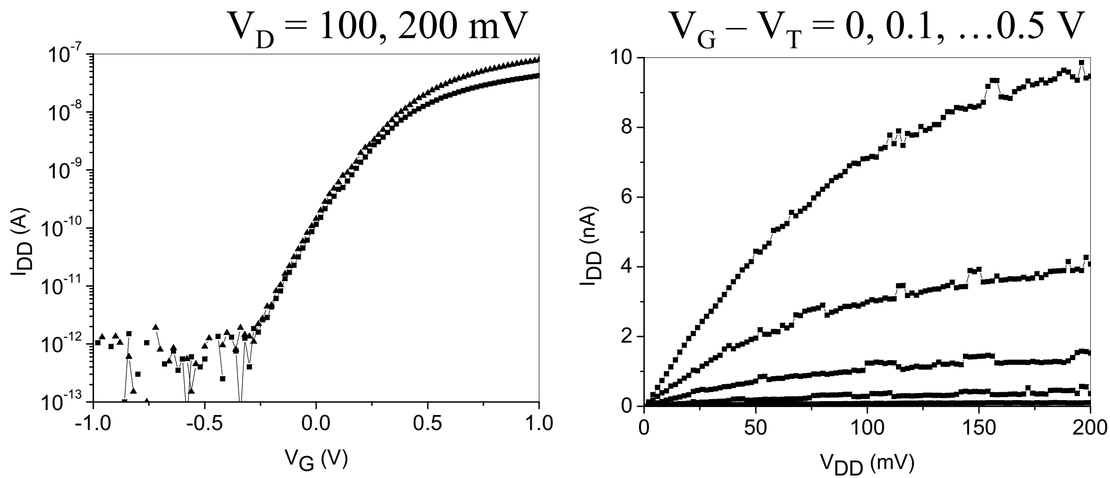


Defects on the Back



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

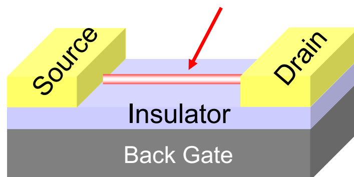
Single Electrons in Output Characteristics

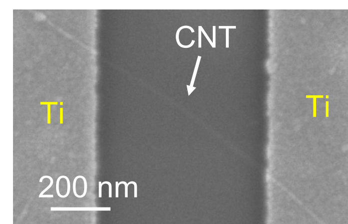


Oxide/Nitride/Oxide : 8 / 15 / 40 nm; $t_{Si} = 50$ nm

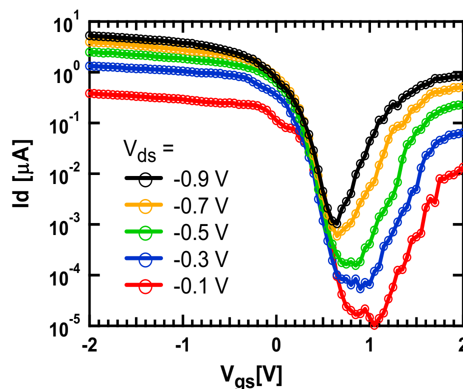
Silva et al., IEEE SOI Conf. (2003)

Carbon Nanotube Field-Effect Transistor

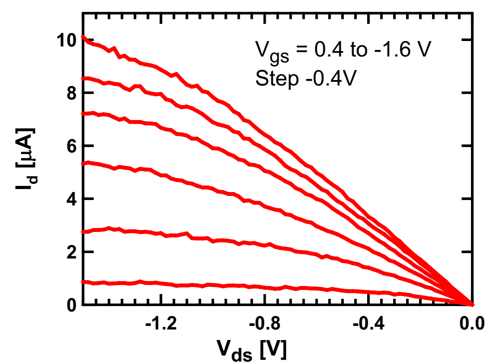




Subthreshold Characteristics



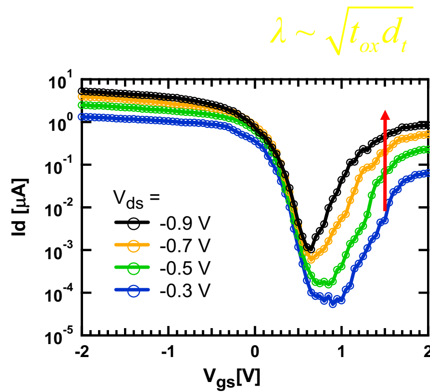
Output Characteristics



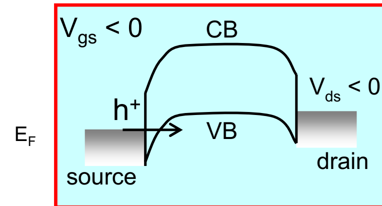
Nanotube FETs

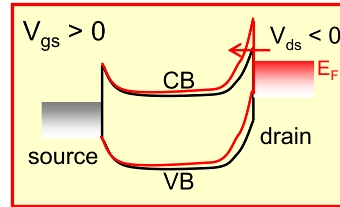
- 1D (ultra-thin body) channel
- Ballistic transport (at low Vds)
- Switching can be dominated by the contact Schottky barriers
  - Screening length
  - Barrier width ~ oxide thickness tox (on-state)
  - Ambipolar behavior

$$\lambda \sim \sqrt{t_{ox} d_t}$$



Gate oxide: 10-nm SiO_2
Contact metal: Ti

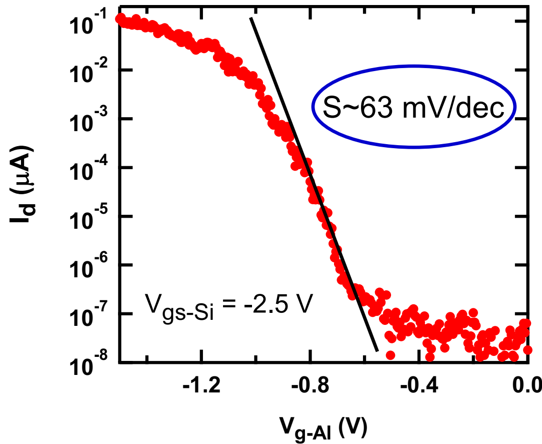




CB: conduction band
VB: valence band

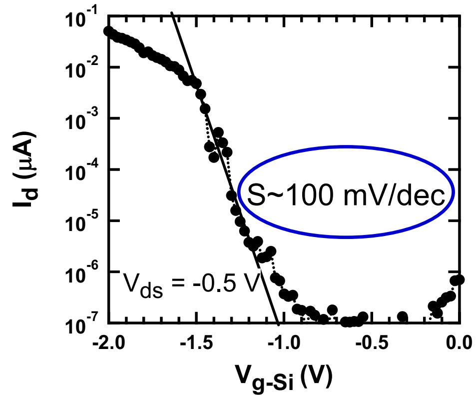
Performance

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



Bulk Switching

I_d vs. Si gate voltage V_{g-Si}



Contact Switching

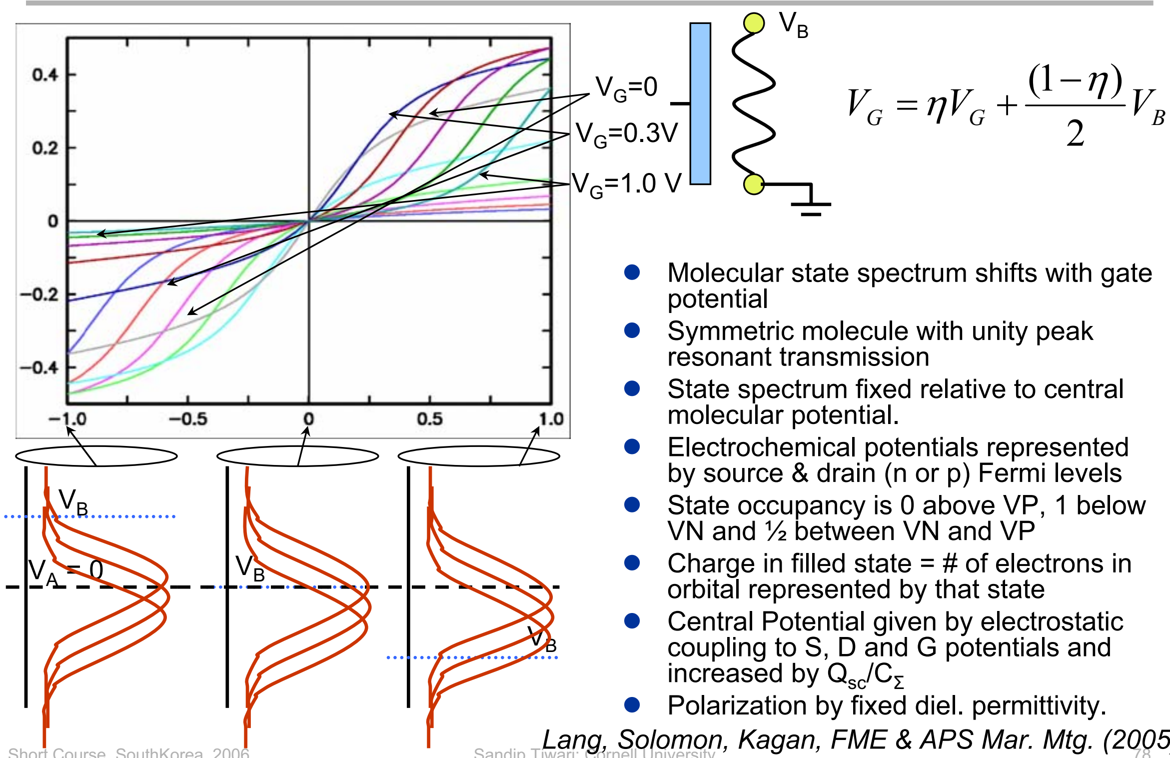
Theoretical limit: S ~ 60 mV/dec

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_Σ
- Polarization by fixed diel. permittivity.

Lang, Solomon, Kagan, FME & APS Mar. Mtg. (2005)

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

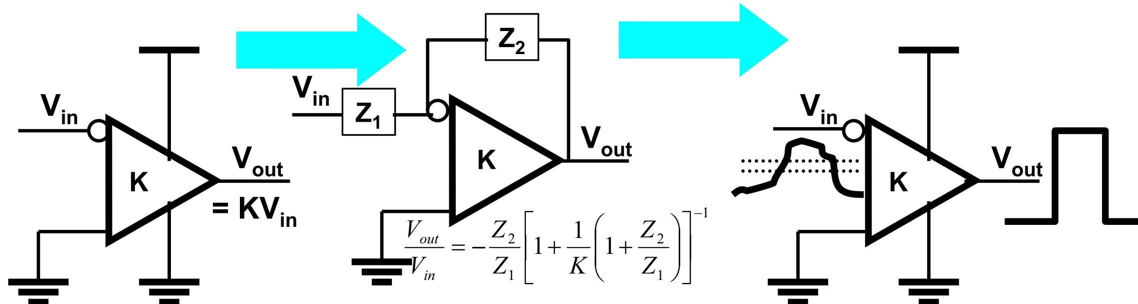
Stability and Signal Recovery

Analog: open & closed loop/feedback

Regenerative effects and signal stability issues

Digital

Signal restoration using gain



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

$$Z_{in} = \frac{Z_2}{K+1}$$

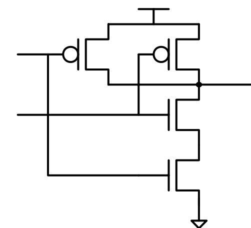
Feedback between input and output (Z_2, e.g.) leads to a larger input load because the 180° phase shift during amplification

At the input C appears as (K+1)C (Miller effect)

Devices and Circuits

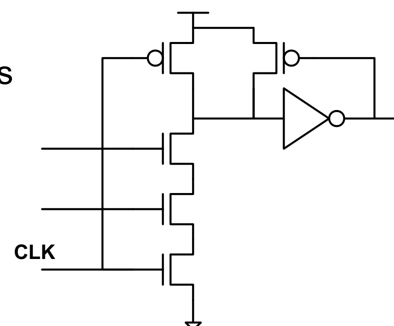
Digital CMOS design:

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

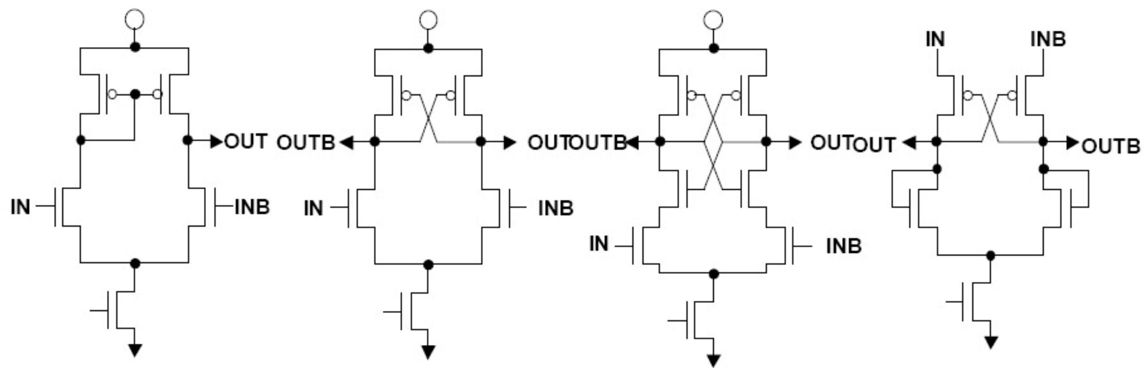


These forms employed because:

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

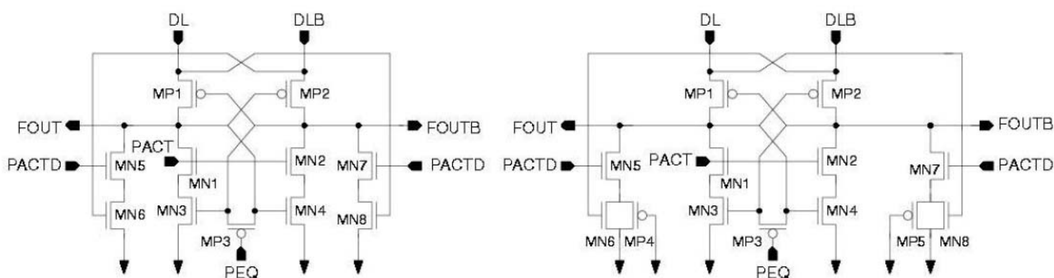


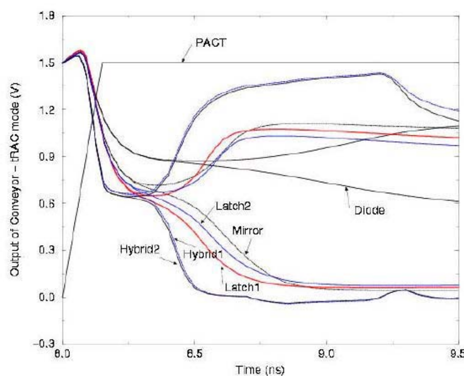
Sensing



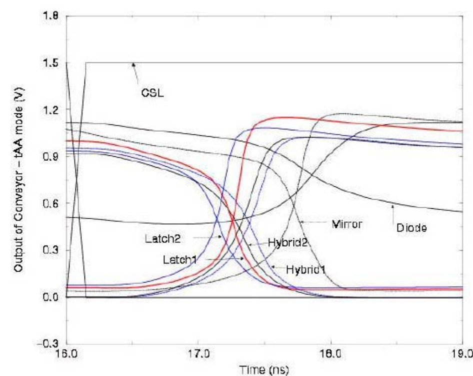
- **Voltage**
  - stable
- **Semi-latch**
  - improved
- **Full-latch**
  - fast
  - precharge
- **Current**
  - fast
  - high power

Sensing





tRAC



tAA

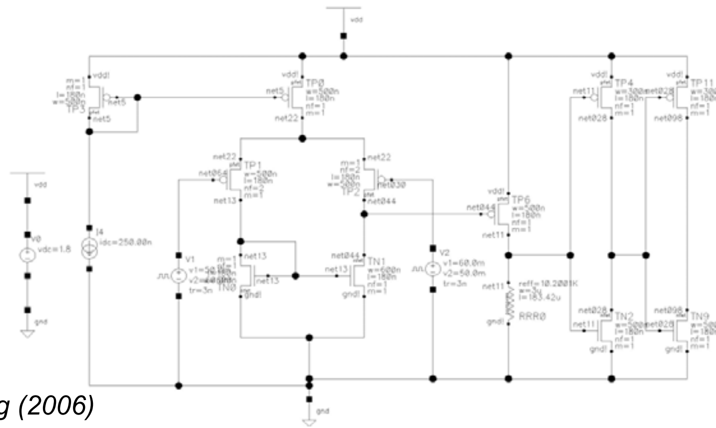
MRAM Designs

Twin Cell

Averaging

Self-Referencing

Differential Amplifier Design			
	Twin	Averaging	Self-referencing
Read time [ns]	6.96	6.16	70.02
Power [uW]	179.37	176.67	44.62
No. of Transistors []	11	11+1capacitor	16 +1 capacitor
Density Index	1	500	∞

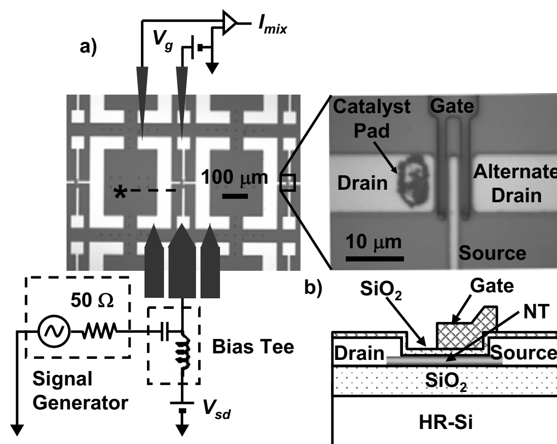


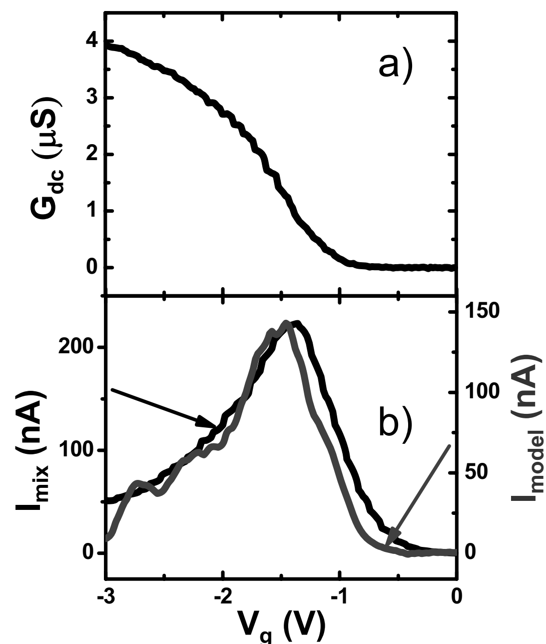
Source: Sudheeran and Chang (2006)

NT Mixer/Transistor

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

Mixing signal for ac applied to source: $$I_{mix} = \frac{1}{4}\left|\frac{\partial G}{\partial V_g}\right| (V_s^{ac})^2$$

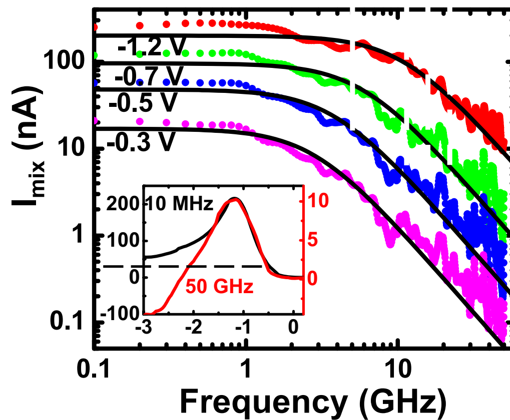


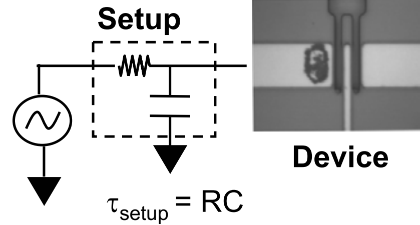


f = 10 MHz

NT Mixer/Transistor up to 50 GHz

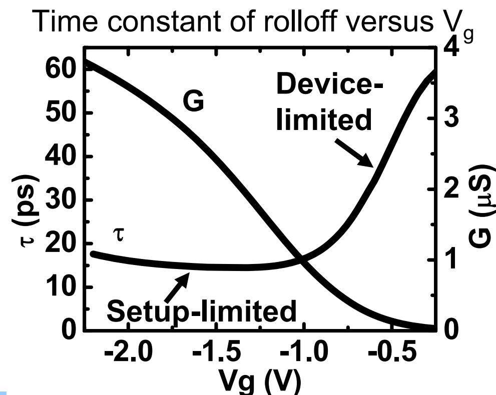
Mixing signal versus frequency





High freq. roll off caused by setup, not device (except near turnoff)

Time constant of rolloff versus V_g



NT FET mixer operates up to 50 GHz

Rosenblatt et al. , APL (2005)

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)

$$|\psi>=\alpha|0>+\beta|1>$$

  - 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

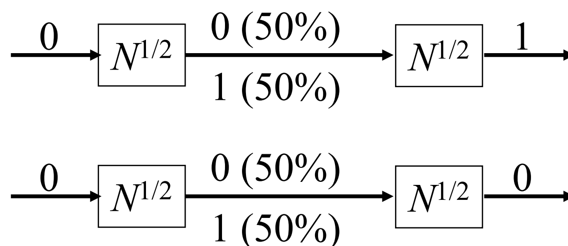
Classical versus Quantum Bit

	Classical	Quantum
Info unit	Bit 0 or 1	Super position $\lvert\psi\rangle = \alpha\lvert 0\rangle + \beta\lvert 1\rangle$
Storage Capacitance	Linear - N	Exponential O(2^n)
Processing	Serial x -> f(x)	Parallel $\sum\lvert\psi_n\rangle \rightarrow \sum\lvert f(\psi_n)\rangle$
Universal Gates	Nand	Single qubit Rotations + Cnot
Measurement	0 -> 0 1 -> 1	Problem : destroy coherence
Algorithms	Many	Factorization, Search Based on quantum interference

The Square Root of NOT

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

- 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?"



NOT$^{1/2}$: Unitary Implementation

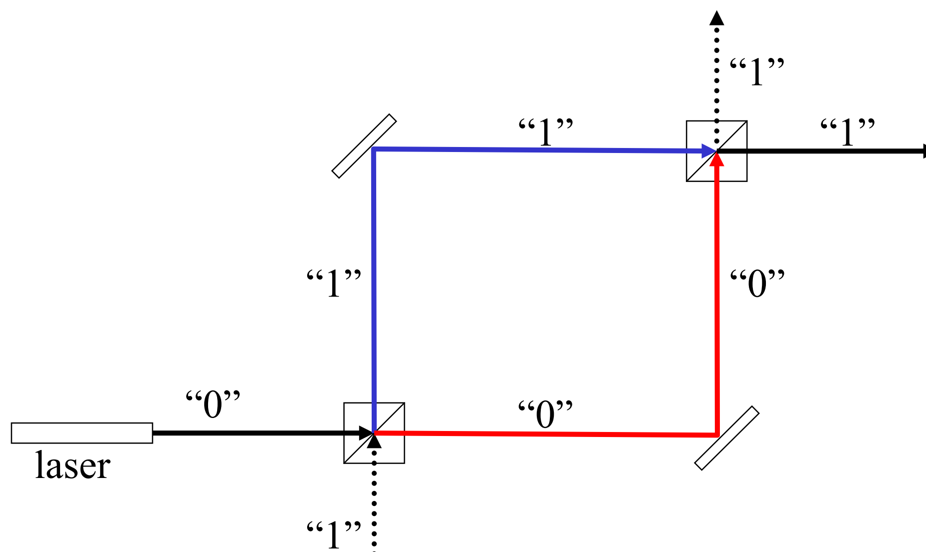
$$\sqrt{N} :\equiv \begin{array}{c} \begin{array}{cc} \mathbf{0} & \mathbf{1} \end{array} \\ \begin{bmatrix} \frac{1+i}{2} & \frac{1-i}{2} \\ \frac{1-i}{2} & \frac{1+i}{2} \end{bmatrix} \end{array} \begin{array}{c} \mathbf{0} \\ \mathbf{1} \end{array}$$

$$\left(\sqrt{N}\right)^2 = \begin{array}{c} \begin{array}{cc} \mathbf{0} & \mathbf{1} \end{array} \\ \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \end{array} \begin{array}{c} \mathbf{0} \\ \mathbf{1} \end{array} = N$$

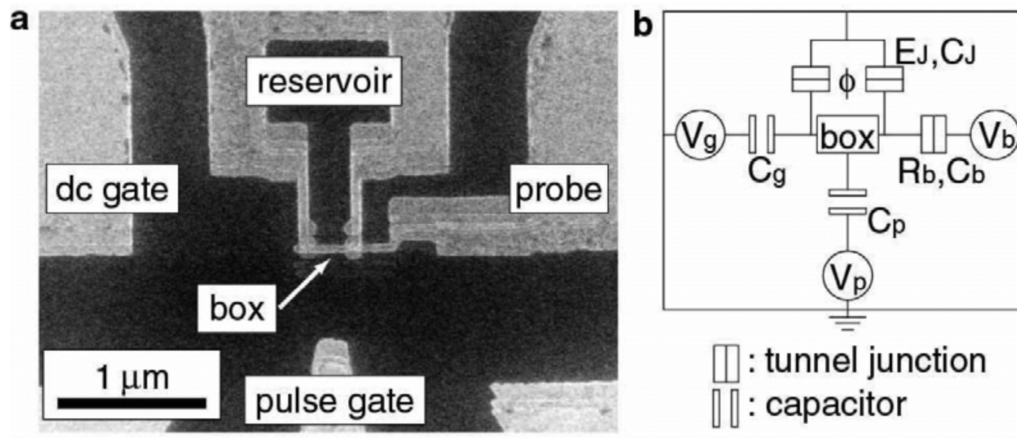
$$\sqrt{N}\left|\mathbf{0}\right\rangle = \sqrt{N}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{1+i}{2} \\ \frac{i-i}{2} \end{bmatrix} = \underbrace{\frac{1+i}{2}}_{\text{Prob. } 1/2}\left|\mathbf{0}\right\rangle + \underbrace{\frac{1-i}{2}}_{\text{Prob. } 1/2}\left|\mathbf{1}\right\rangle$$

Optical Implementation of N$^{1/2}$

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



SET Probing Qbit



Box: 700x50x15 nm; 10^8 electrons

$e^2/2C$ = 117 μV, T = 30 mK, kT = 3 μV

Initialize by preparing pure state |0> away from resonance

Apply fast voltage pulse Δt 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 Δt in pico-seconds

Nakamura

Quantum Computing Requirements

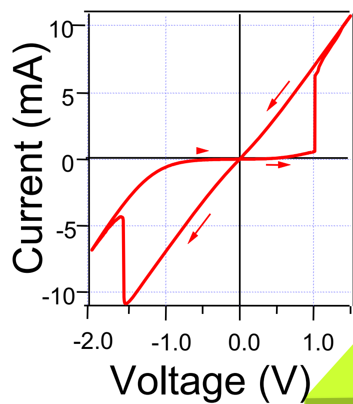
- DiVicenzo's check list:
  - Identifiable qubits and their scalability
  - Initialization procedure
    - Preparation of ground state of the whole system
  - Low decoherence
    - Long decoherence time versus gate time
  - Quantum gating – ability to realize a universal gate through control of system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

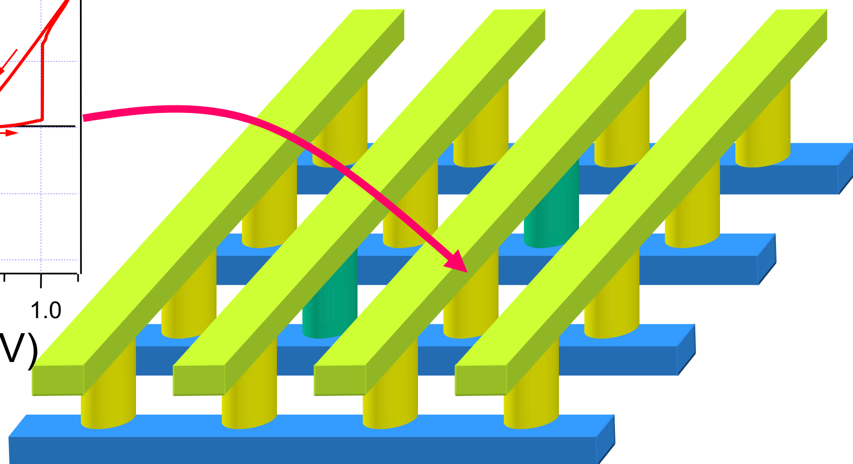
Systems, Hierarchy, Complexity and Architecture

A critical look

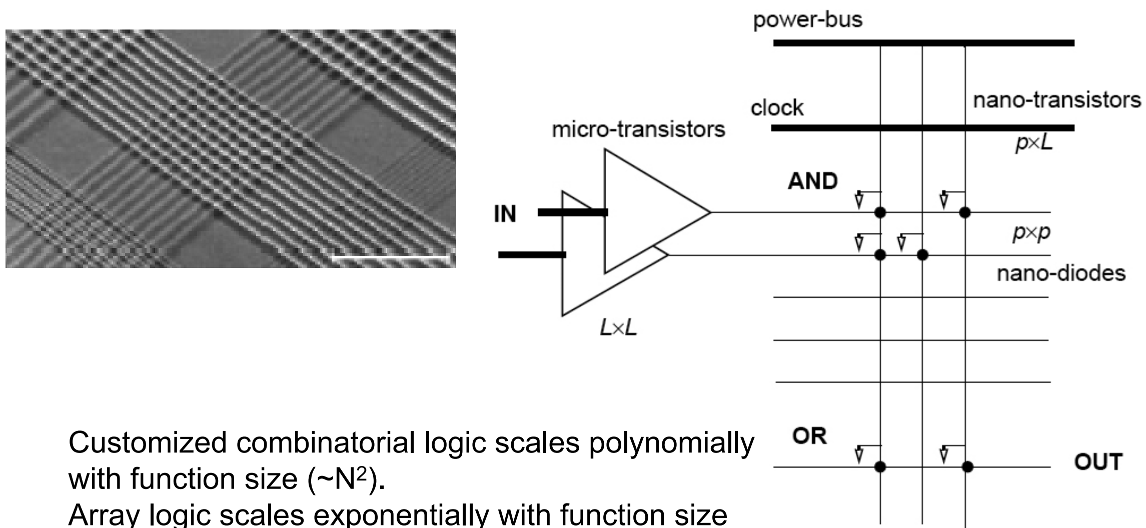
Molecular Memories?

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





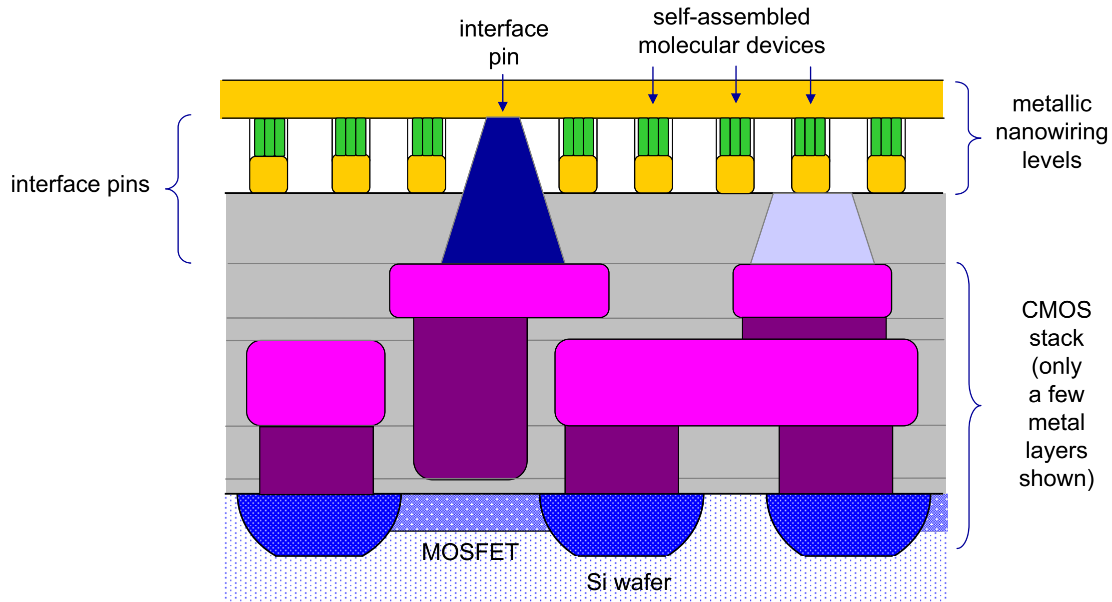
Molecular: 2 terminals or 3 terminals?



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

UCLA, CalTech, HP

CMOL

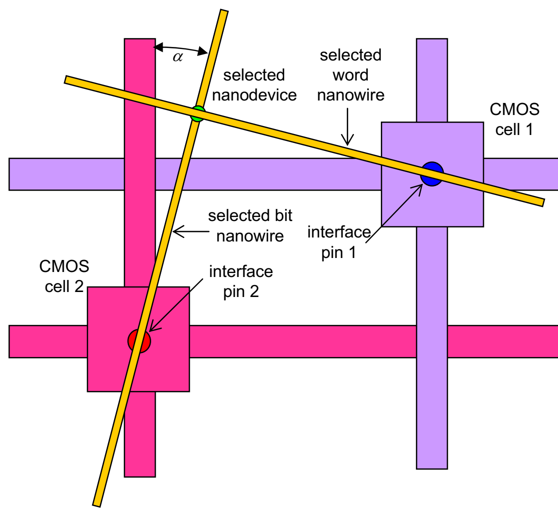


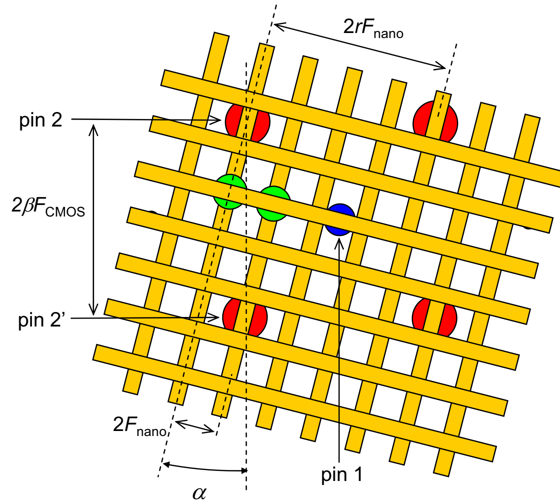
CMOL combines: - (relatively sparse, but highly functional) CMOS subsystem
- very dense nanowire crossbar, and
- a molecular-scale device at each nanowire crosspoint

Likharev (2003)

CMOL

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





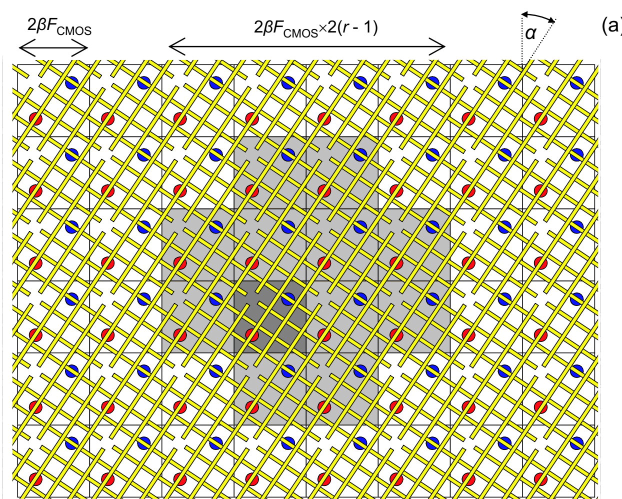
Nanodevice addressed via two CMOS cells

Each and every nanodevice may be addressed!

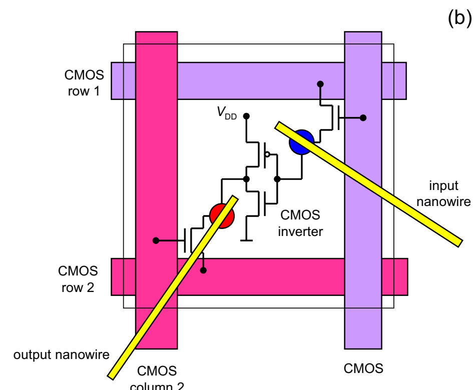
Likharev (2003)

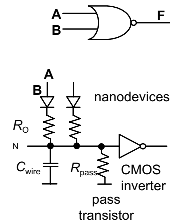
CMOL FPGA

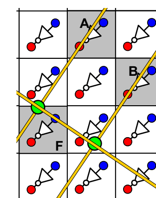
CMOS inverters + nanodevice latches for (re)configuration



CMOL FPGA fabric

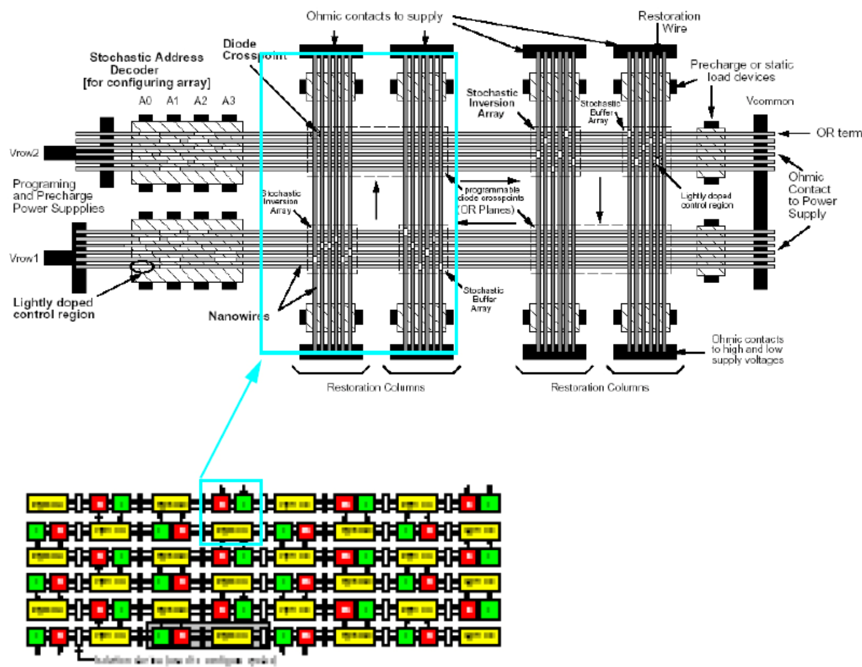






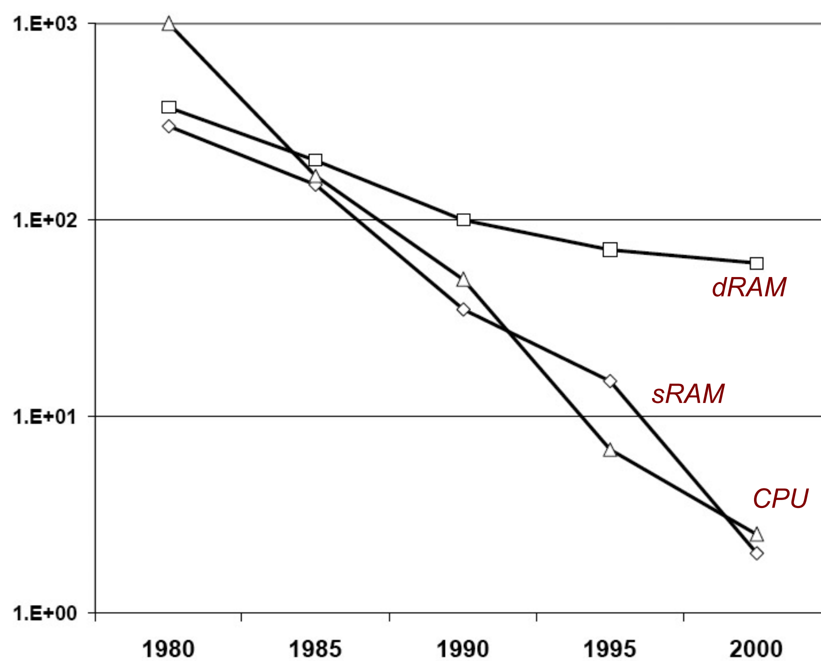
Likharev (2005)

NanoFabrics



deHon (2004)

Speed Gaps in Processor Hierarchy

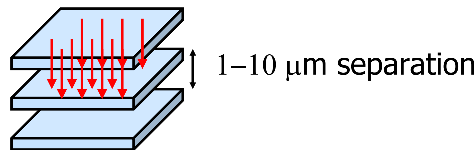


Source: T.C. Mowry

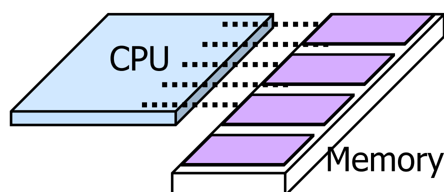
3-D Microprocessors

- 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

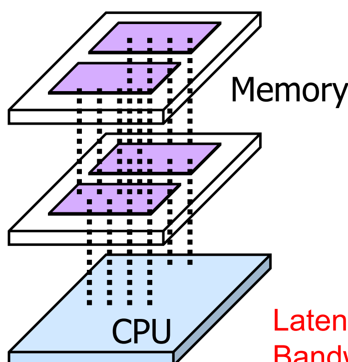
Vertical connections throughout the design area



Latency and Bandwidth



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



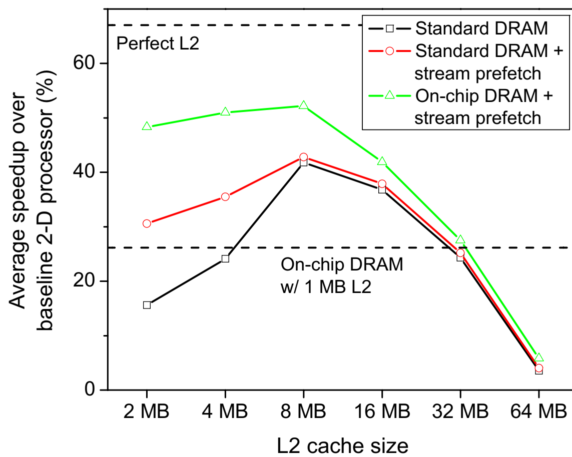
3-D: Connections across the area
- Connections short + vertical
- Suitable for high-bandwidth and vector operations
- No pin cost, large block access of data

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

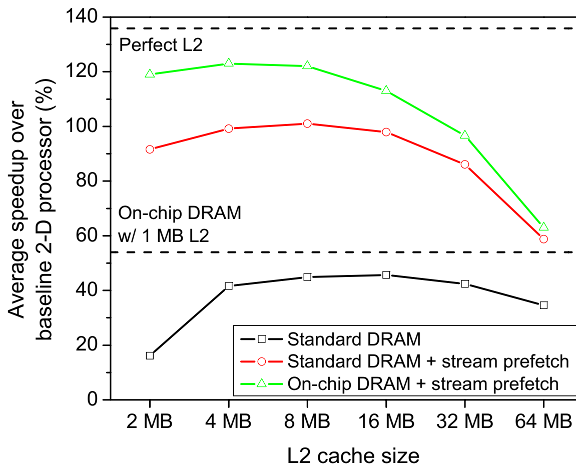
The following example uses a baseline 2-D processor core representative of current technology
3 GHz CPU, 750 MHz memory, 64 KB L1I, 64 KB L1D, 1 MB L2

Expanding L2 Cache

SPEC2000 Integer programs



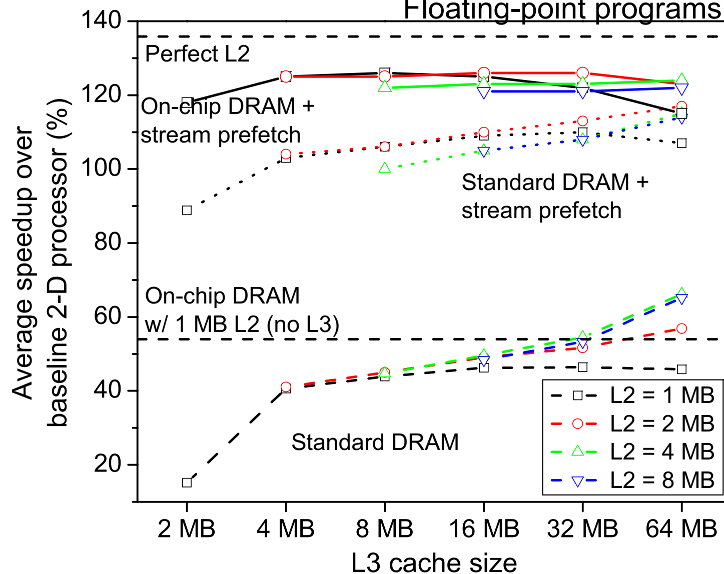
Floating-point programs



- 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 .. Tiwari., 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 .. Tiwari., IEEE D&T Mag. (2005)

Complexity

- Rent's Rule:
  - Terminal count is related to number of gates (at all hierarchical levels)

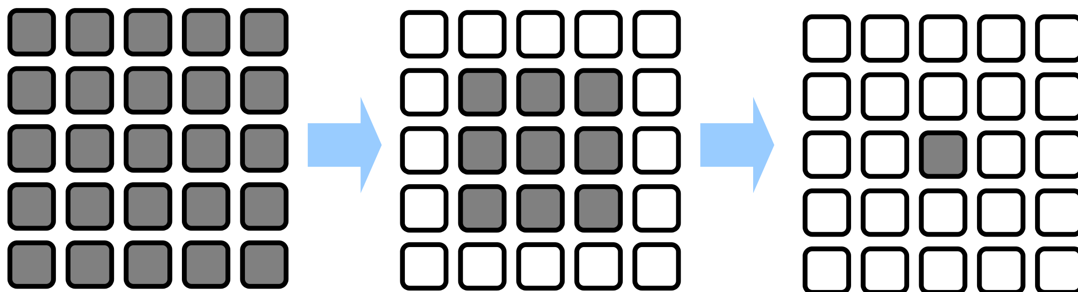
    $T = t N^p$

    (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

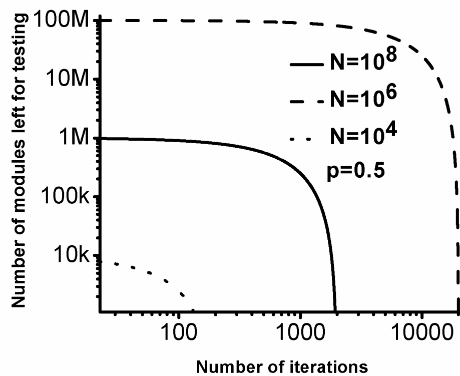


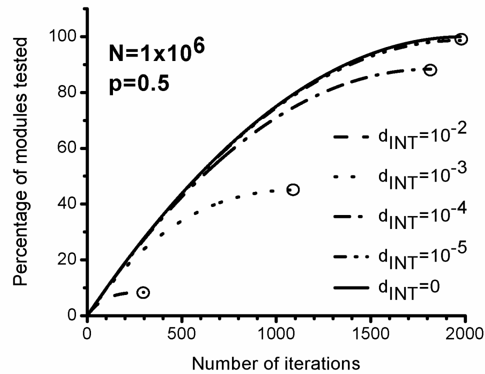
- Suppose, we work with 10^5 logic blocks, each employing 1000 device elements
- What does it mean that one can work with chips that are 90% functional (or that 10% of logic blocks are faulty)
  - If the probability of failure is p for each element, probability of a logic block being functional is $(1-p)^{1000}$
  - 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

Percentage of modules tested/testable in presence of interconnect defects



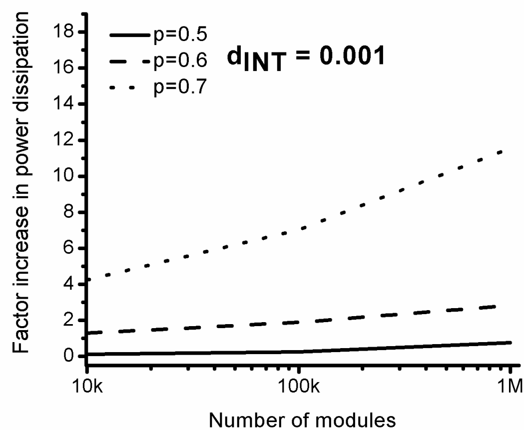


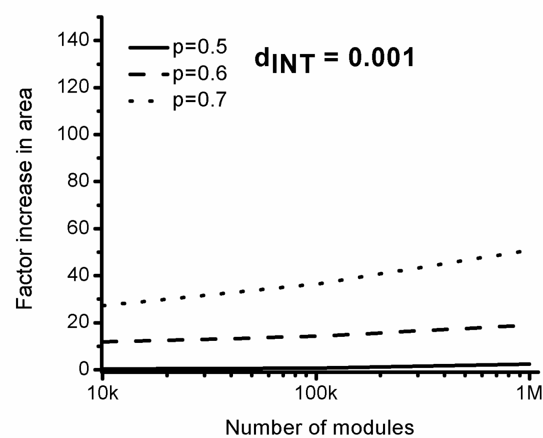
Usable logic blocks of a chip reduces rapidly with interconnect defects since their correctness can not be tested.

Testability is cumbersome.

Kumar, DFT (2004)

Interconnect Defect Penalty





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