# The future of scaling

George Bourianoff, Intel

Dublin, Ireland

August 28,2011

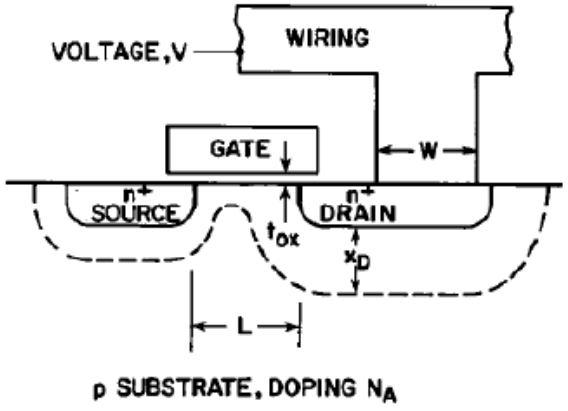
Acknowledgement to Kelin Kuhn, Intel Fellow

# Outline

- Scaling the FET
  - Classical Dennard scaling
  - 2003-2011 - performance enhancement
  - 2011-20?? -
- Beyond the FET
  - More powerful unit devices
  - Temperature, order and non-equilibrium operation
  - Bioinspired, Non Boolean operation
- Conclusions

# Classic MOSFET scaling

| Device or Circuit Parameter | Scaling Factor |
|---|---|
| Device dimension *tox, L, W* | 1/κ |
| Doping concentration *Na* | κ |
| Voltage *V* | 1/κ |
| Current *I* | 1/κ |
| Capacitance *εA/t* | 1/κ |
| Delay time/circuit *VC/I* | 1/κ |
| Power dissipation/circuit *VI* | $1/\kappa^2$ |
| Power density *VI/A* | 1 |




*R. Dennard, IEEE JSSC, 1974*

**Classical MOSFET scaling was first described by Dennard in 1974**

Classical MOSFET scaling ended at the 130 nm node

And no one noticed !

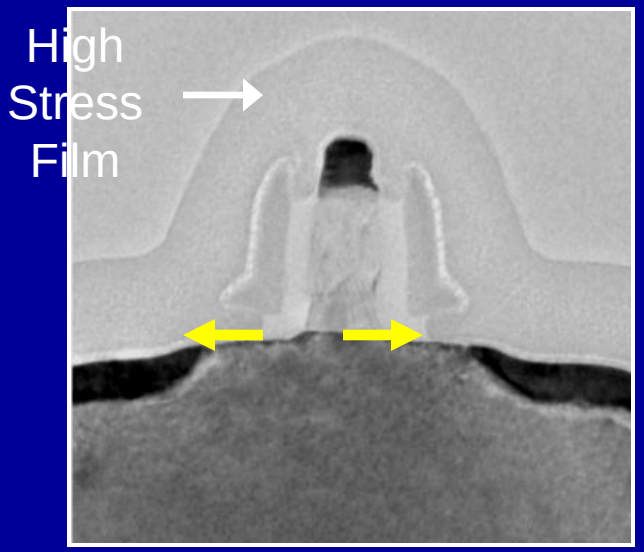
Why did no one notice?

What has been happening since 2003 (130 nm node)?

Performance Boosters

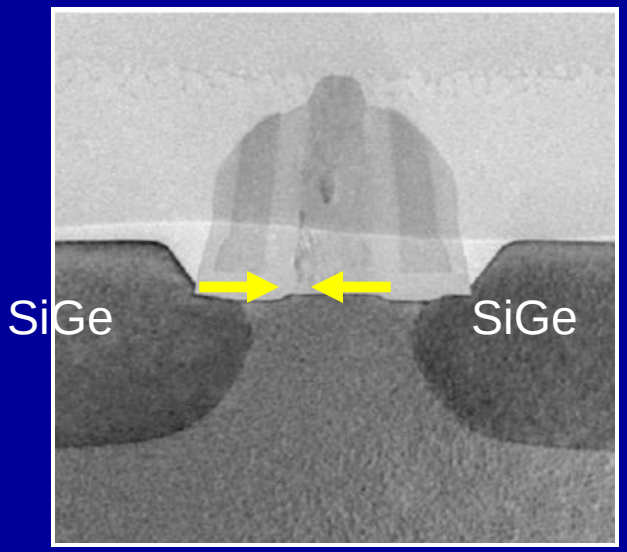
# 90 nm Strained Silicon Transistors

NMOS




SiN cap layer
Tensile channel strain

PMOS

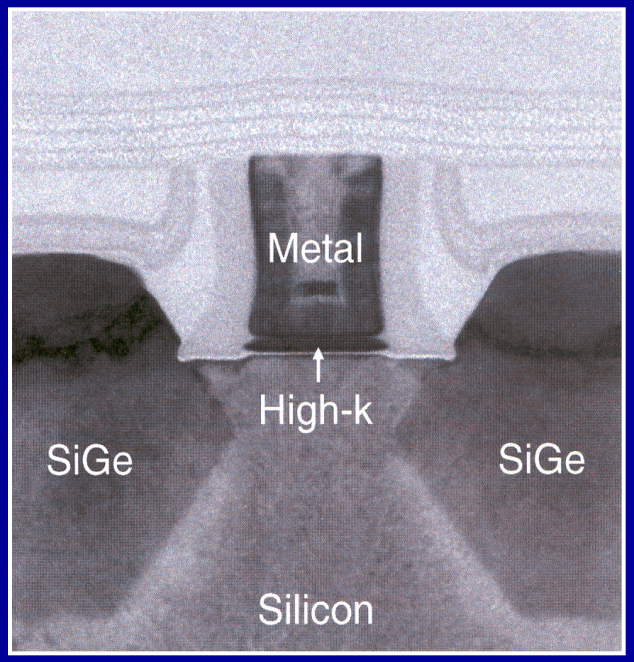



SiGe source-drain
Compressive channel strain

**Strained silicon provided increased drive currents, making up for the loss of classical Dennard scaling**

# 45nm High-k + Metal Gate Transistors

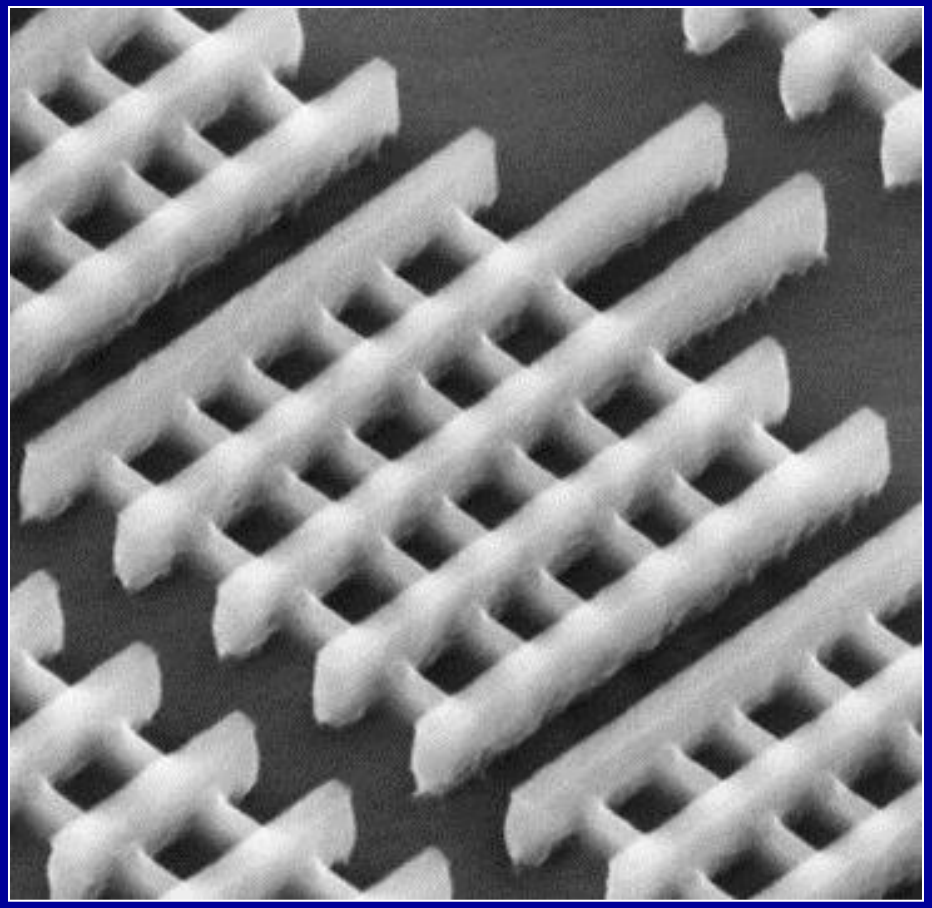
45 nm HK+MG




Hafnium-based dielectric
Metal gate electrode

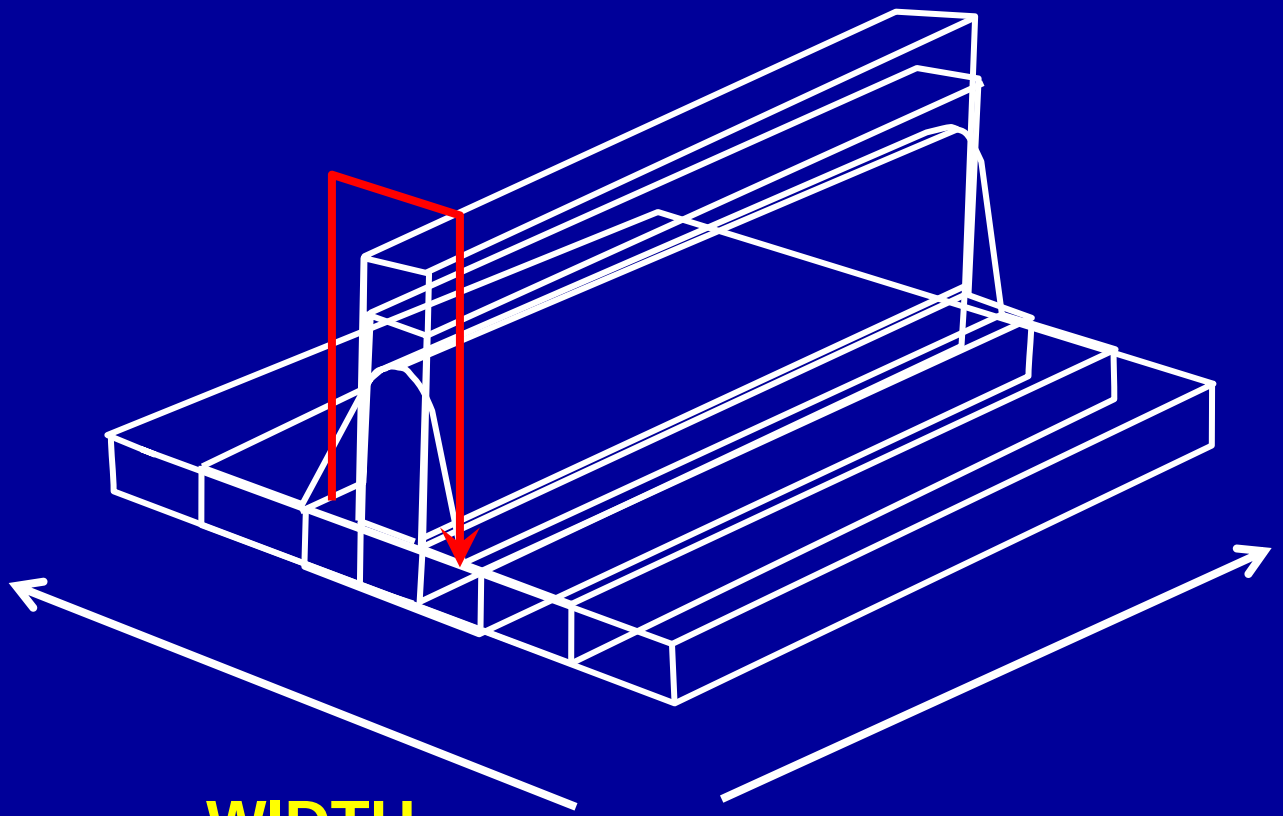
**High-k + metal gate transistors restored gate oxide scaling at the 45nm node**

# 22nm production TriGate process



**Mark Bohr, Kaizad Mistry: Intel, April 25th, 2011 press release**

# TriGate

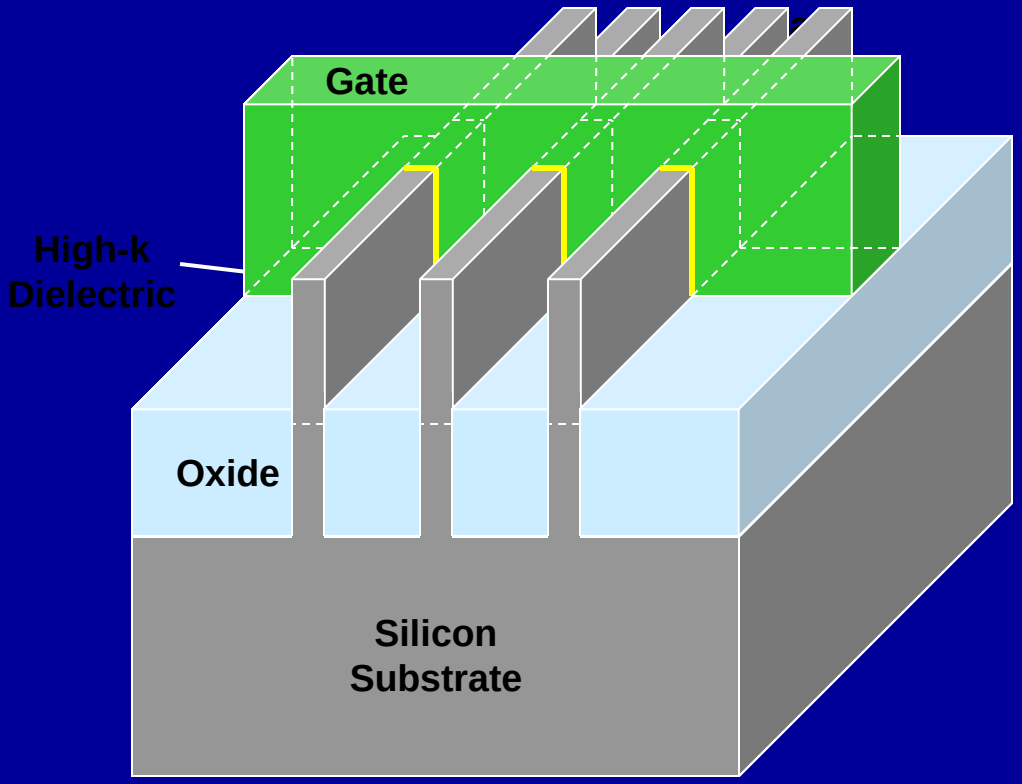


**WIDTH**

**WIDTH = 2H + W**

**LENGTH**

# TriGate

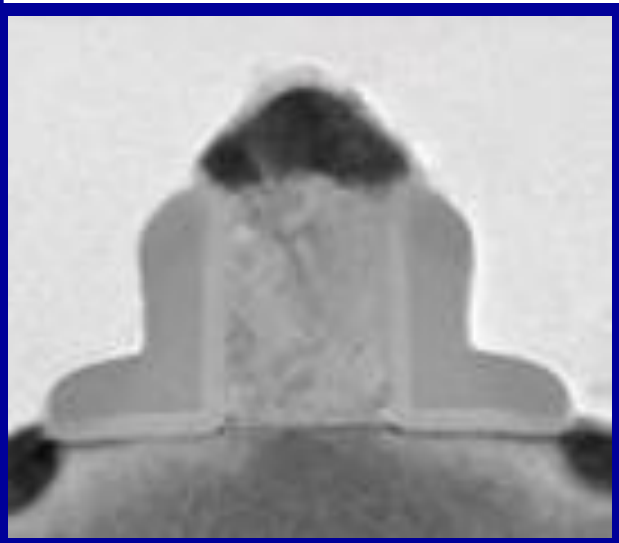



**Tri-Gate transistors can have multiple fins connected together to increase total drive strength for higher performance**

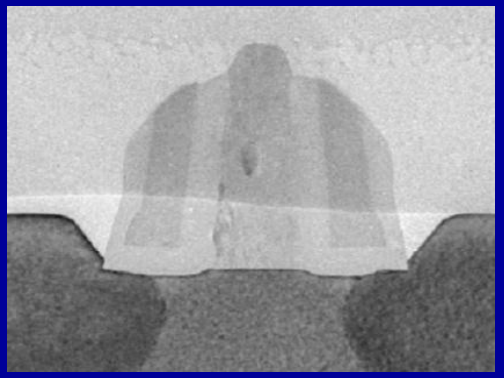
# Changes in Scaling

## THEN

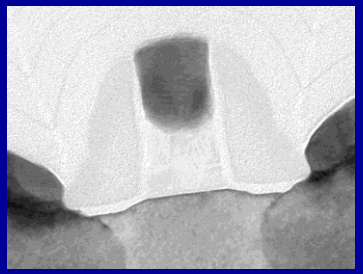
- Scaling drove down cost
- Scaling drove performance
- Performance constrained
- Active power dominates
- Independent design-process



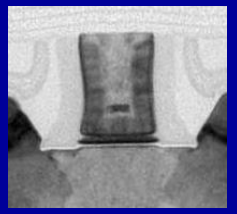
**130nm**



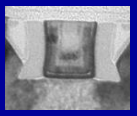
**90nm**



**65nm**



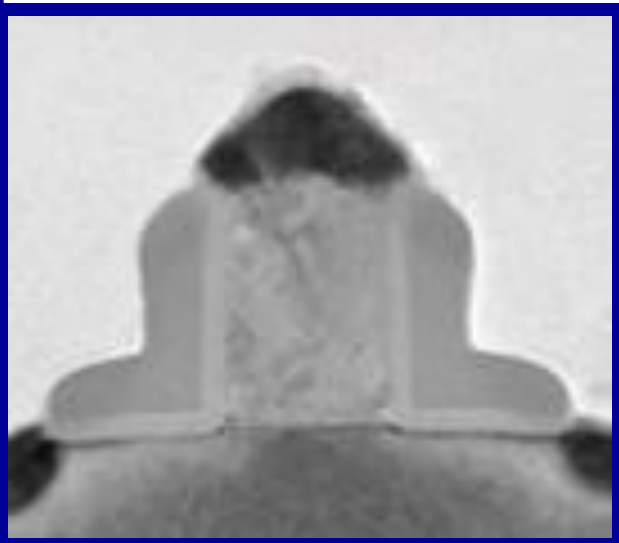
**45nm**



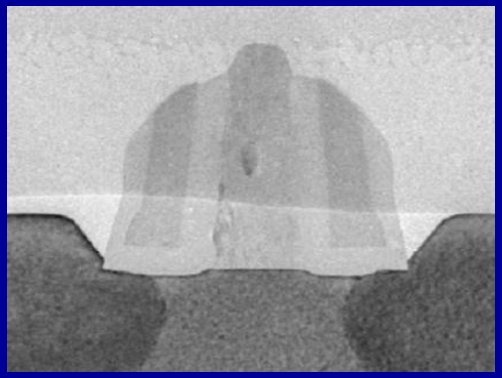
**32nm**

# Changes in Scaling

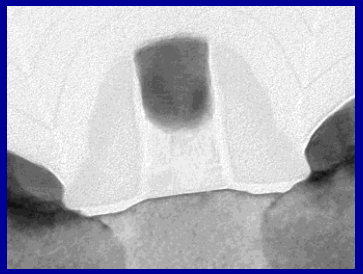
| THEN | NOW |
| --- | --- |
| • Scaling drove down cost<br>• Scaling drove performance<br>• Performance constrained<br>• Active power dominates<br>• Independent design-process | • Scaling drives down cost<br>• Materials drive performance<br>• Power constrained<br>• Standby power dominates<br>• Collaborative design-process |



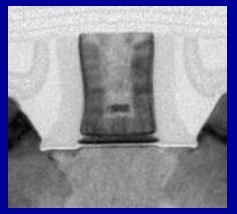
130nm



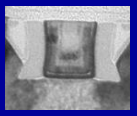
90nm



65nm

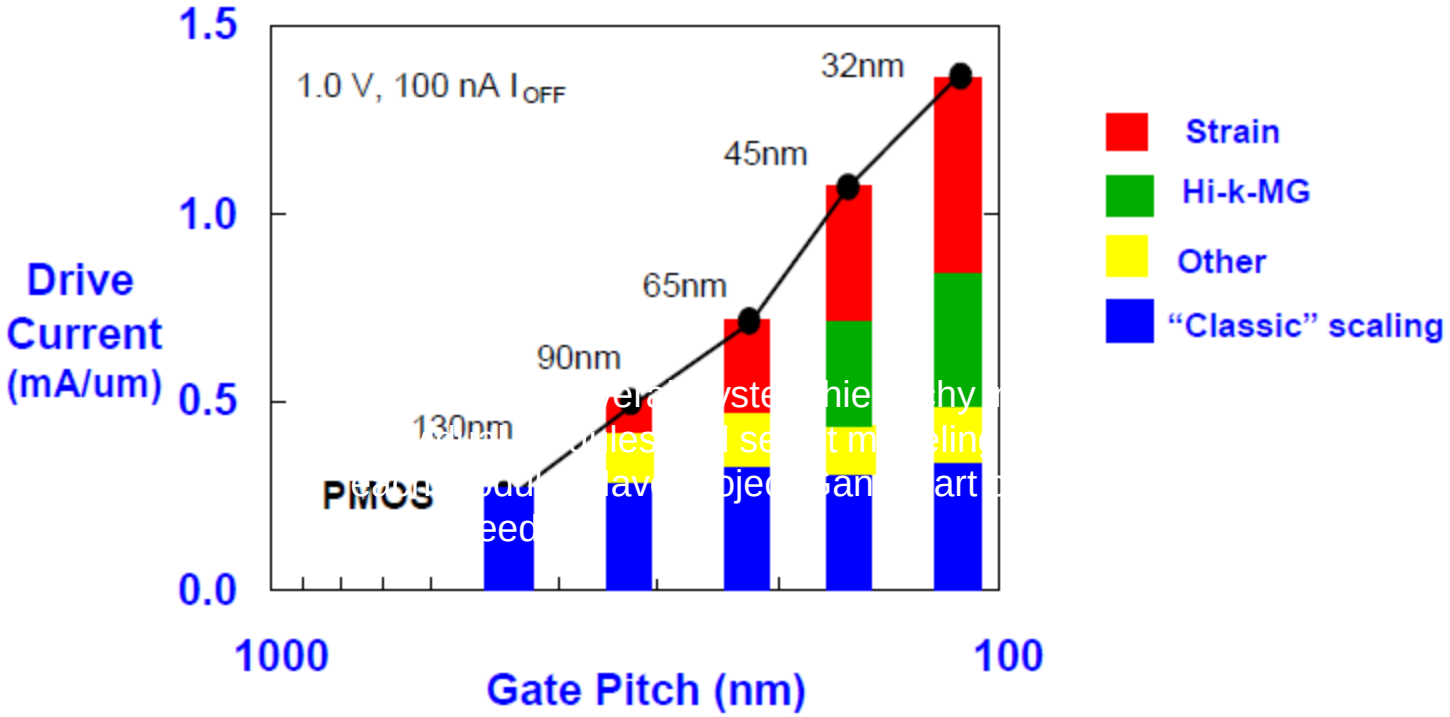


45nm



32nm

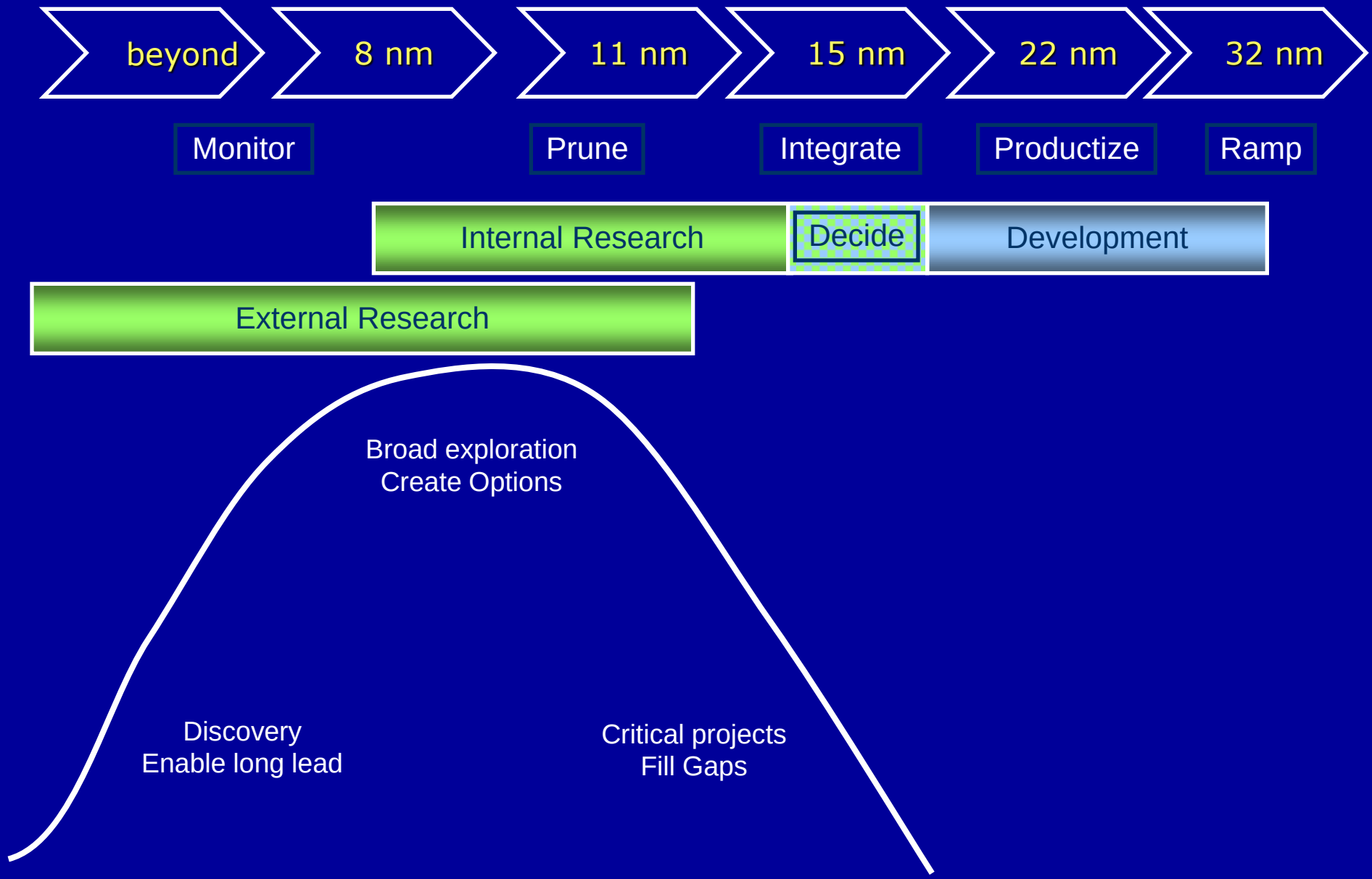
# Transistor Performance Trend




**Strain is a critical ingredient in modern transistor scaling**

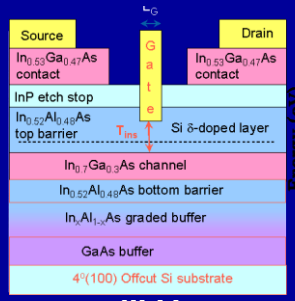
**Strain was first introduced at 90nm, and its contribution has increased in each subsequent generation**

# Ideal View of Research

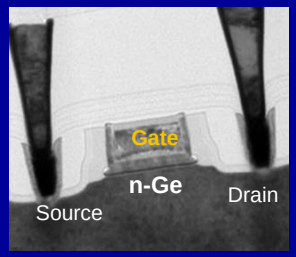


Are there other performance boosters?

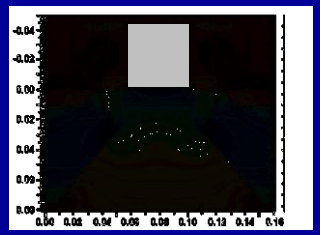
MOBILITY



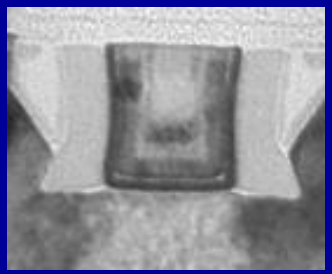

III-V




Ge



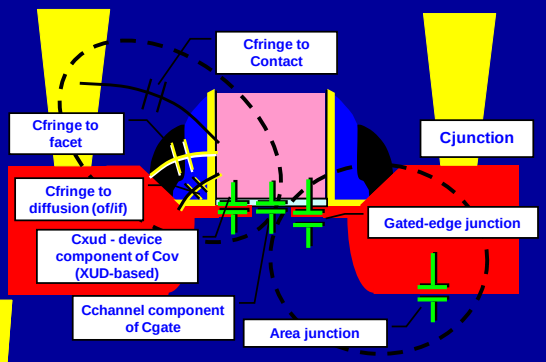
Strain



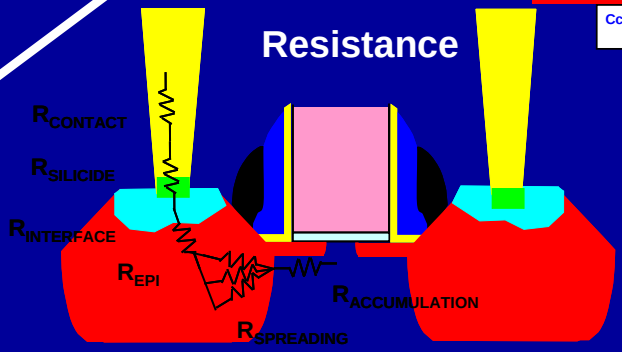
32nm

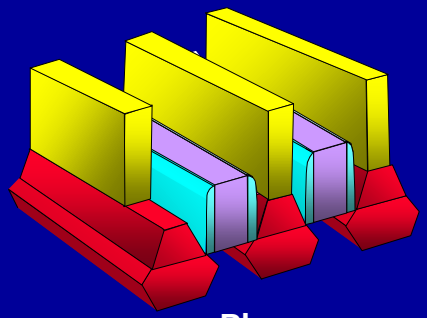
PARASITICS

Capacitance

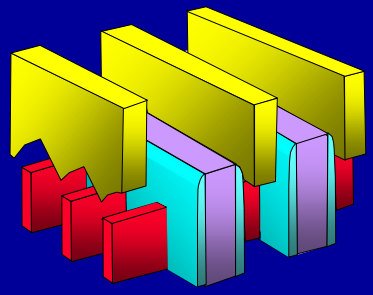



Resistance

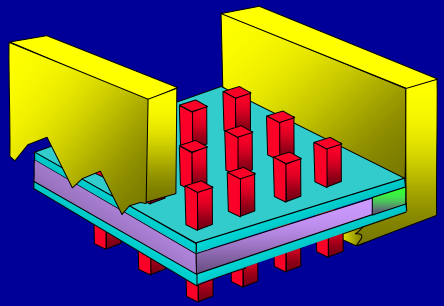


Planar



Fins

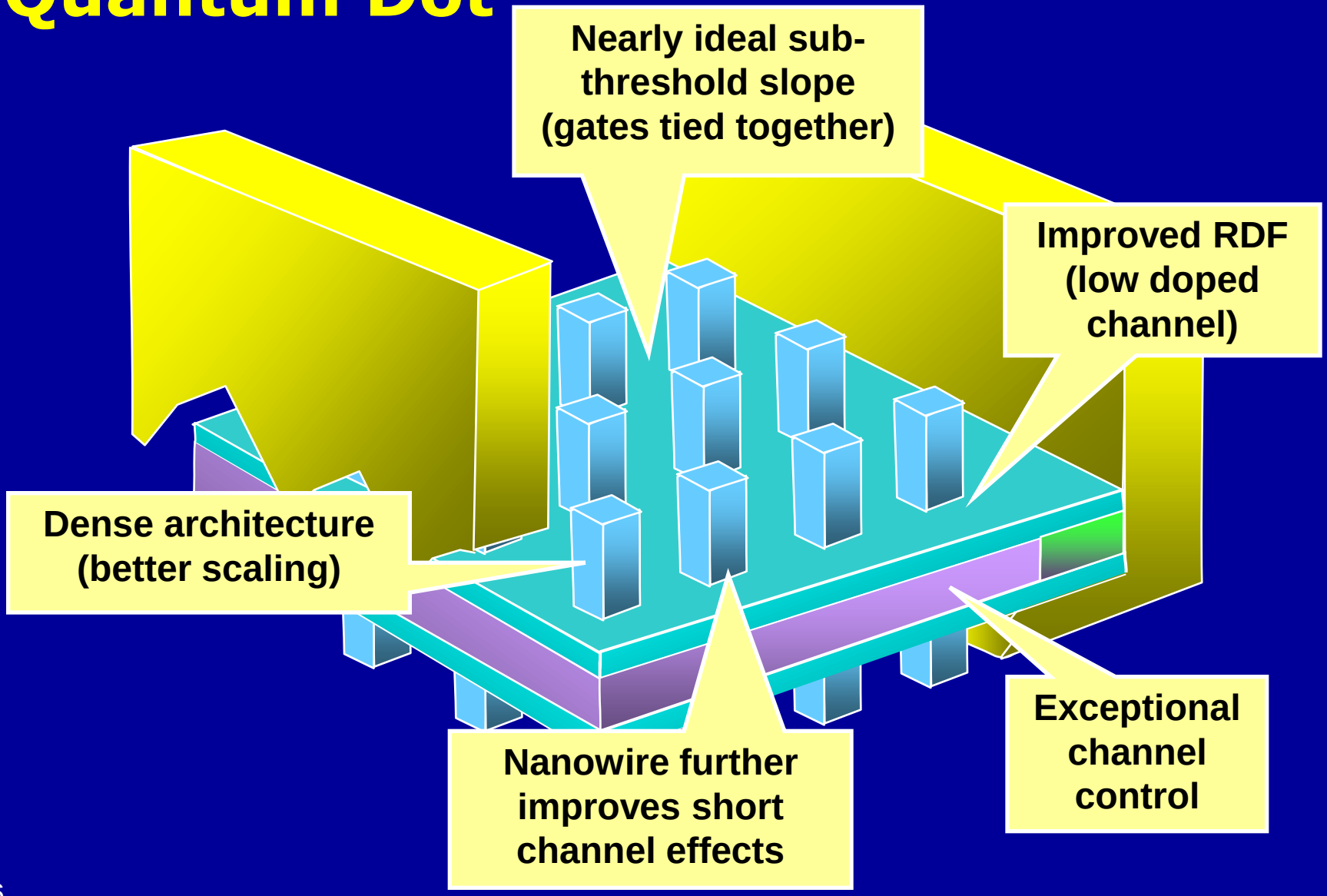


Wires/Dots

ELECTROSTATIC CONFINEMENT

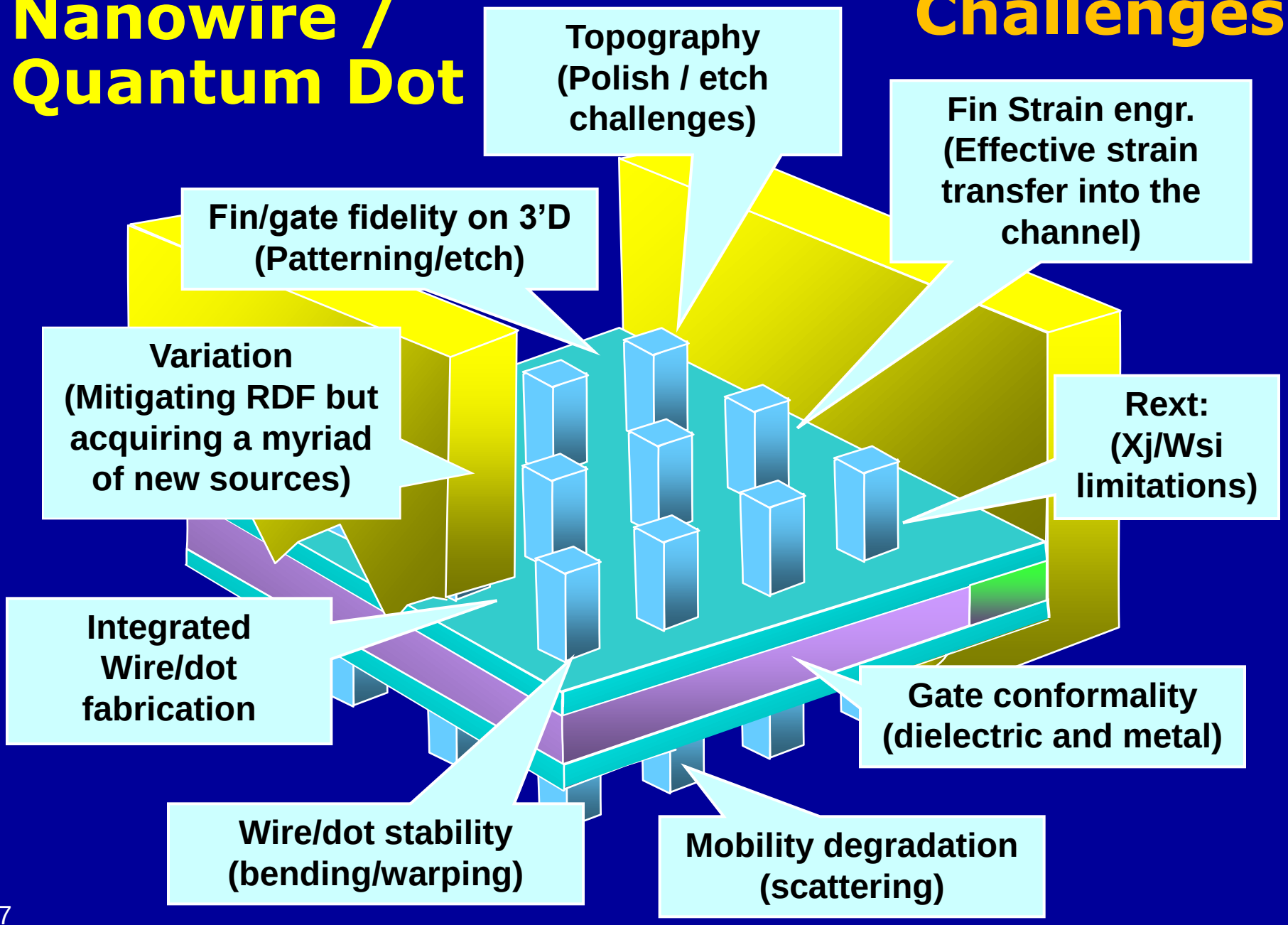
# Nanowire / Quantum Dot

# Benefits



# Nanowire / Quantum Dot

# Challenges

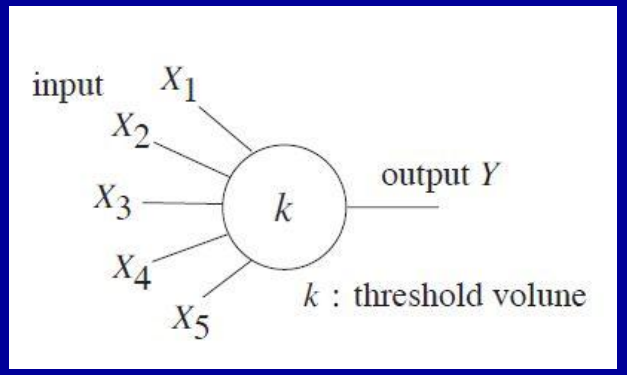


# Outline

- Definition
- Device “scaling” 2003 -2011
- Device “scaling” 2011 - ?
- Beyond the FET
  - More powerful unit devices
  - Temperature, order and non-equilibrium operation
  - Bioinspired, Non Boolean operation
- Conclusions

# More powerful “unit” devices

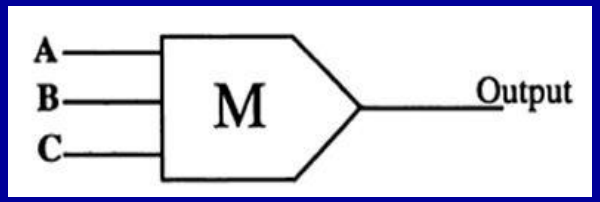
- Greater functionality per device - less interconnect per function
- Interconnects losses large fraction of total power*




•Use more powerful “unit logic devices” –multi input threshold gate

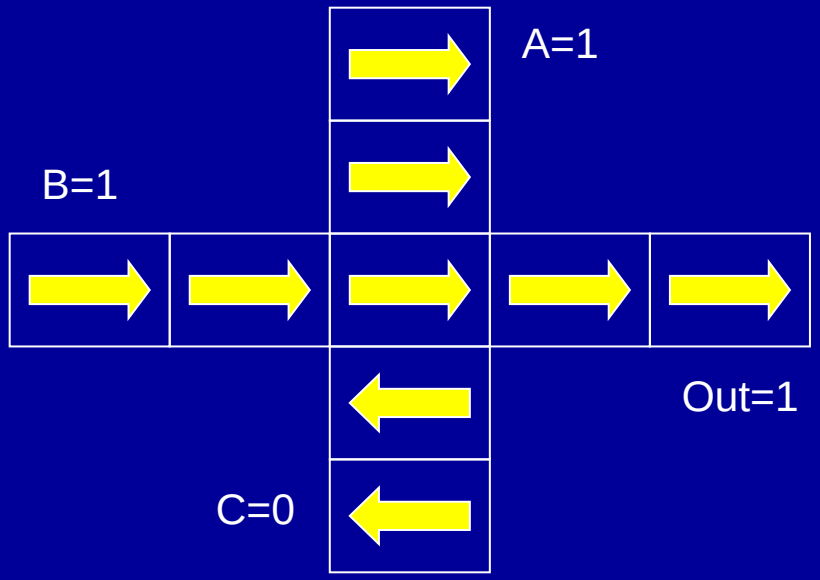
• 3 input majority gate is specific example

•Can be implemented using spin torque transfer technology




* Magen et. al. Interconnect-Power Dissipation in a Microprocessor

# Majority Gate functionality



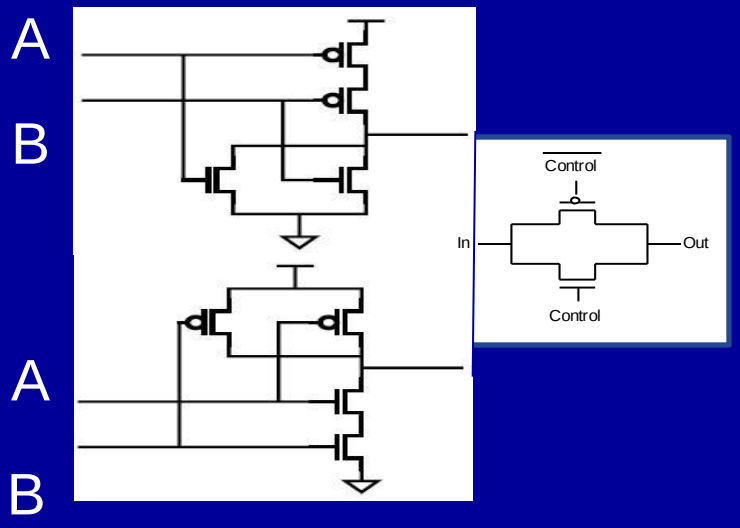

| | A | B | C | Out |
|---|---|---|---|---|
| AND | 0 | 0 | 0 | 0 |
| | 0 | 0 | 1 | 0 |
| | 0 | 1 | 0 | 0 |
| | 0 | 1 | 1 | 1 |
| OR | 1 | 0 | 0 | 0 |
| | 1 | 0 | 1 | 1 |
| | 1 | 1 | 0 | 1 |
| | 1 | 1 | 1 | 1 |

- A gate with 3 inputs and 1 output
- Output is a majority voting of inputs
- Binary output requires a read op with sense amp
- Logically equivalent to reconfigurable AND/OR gate

# Majority Gate Equivalent circuit functionality

CMOS

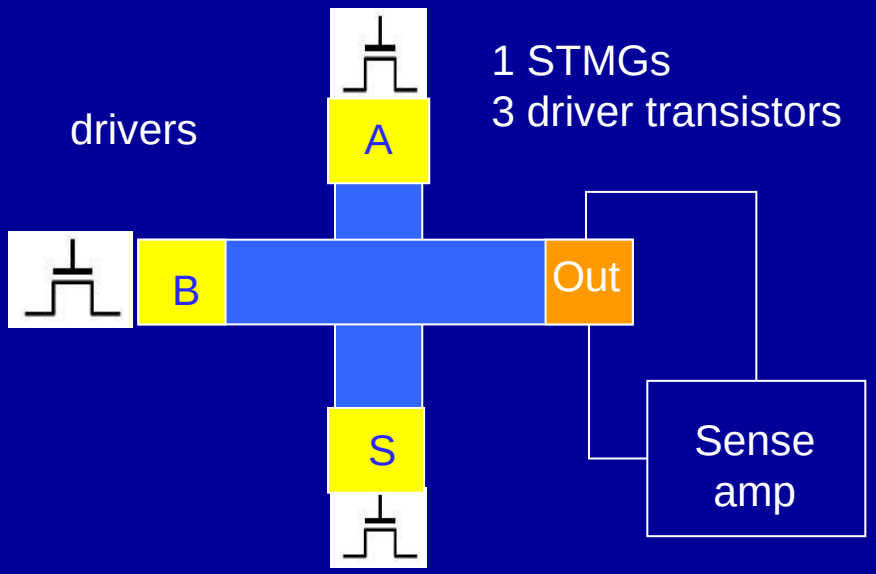
12 transistors




Average area per transistor in a MPU, according to ITRS
Requires 12 transistors

$A_{gate} \sim 12 * 72\ F^2 = 824\ F^2$

STMG

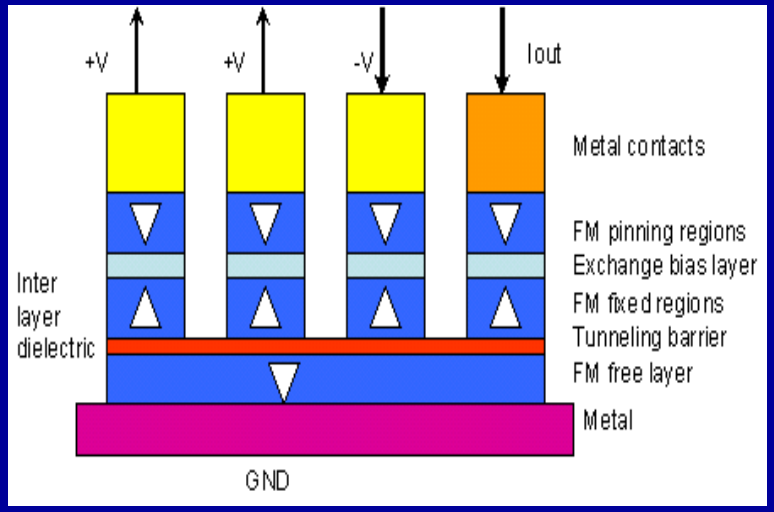



1 STMGs
3 driver transistors

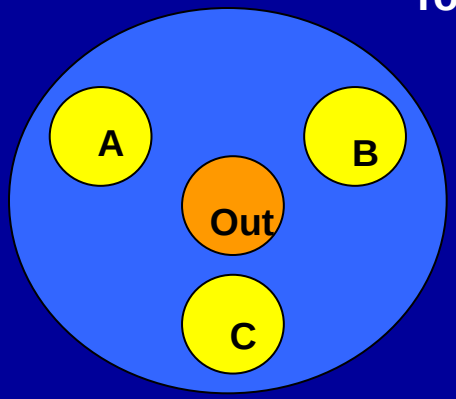
Magnetic circuits in the metal layers.
Drive transistors no additional area.
Scalable with process feature size.

$A_{gate} \sim 36\ F^2$

# Spin torque transfer majority gate (STTMG)




**Top view**




- Four stacks of ferromagnetic materials, similar to perpendicular MTJs
- Three stacks - inputs. One stack - output
- Free layer is common to all four stacks
- Polarity of free layer can be controlled by polarity of voltage applied to 2 of the 3 input stacks
- Polarity of free layer can be sensed by magneto-resistance of forth stack

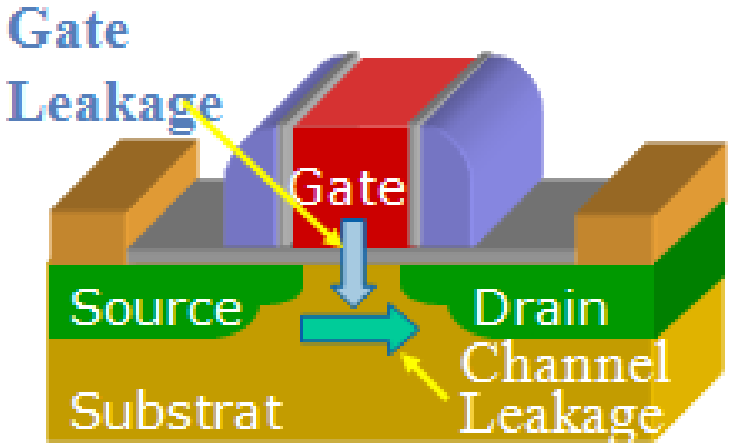
# Summary of STTMG operation

- Single non volatile magnetic device replaces 12 FETs plus their interconnect
- Magneto dynamics of the free layer replaces cascaded switching operations in CMOS gates
- Potential performance benefits in some application areas
- Shows the potential for “More Powerful” logic devices

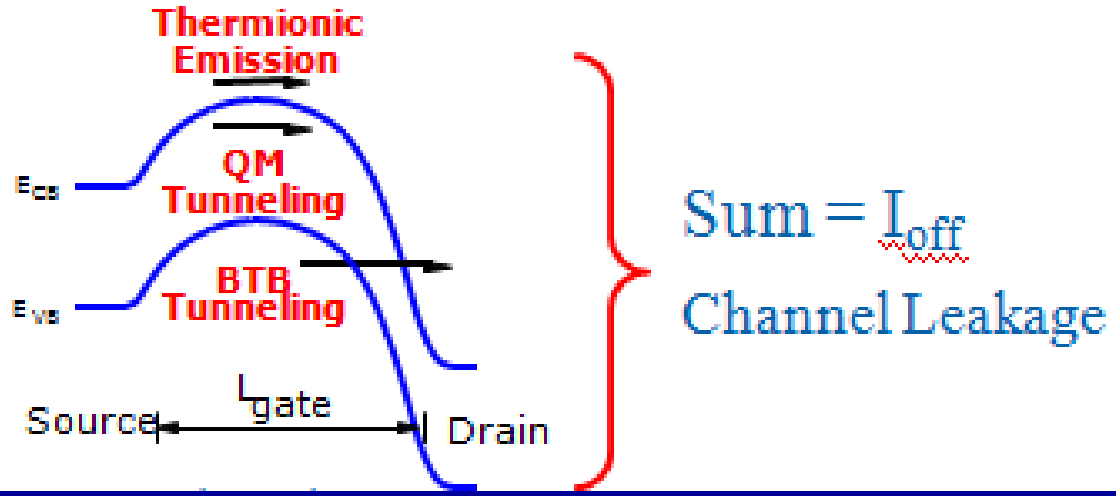
# Outline

- Definition
- Device “scaling” 2003 -2011
- Device “scaling” 2011 - ?
- Beyond the FET
  - More powerful unit devices
  - Temperature, order and non-equilibrium operation
  - Bioinspired, Non Boolean operation
- Conclusions

# Temperature, order and non equilibrium operation




- Charge transport across a barrier
- Assumes thermal distribution
- Assumes thermal equilibrium

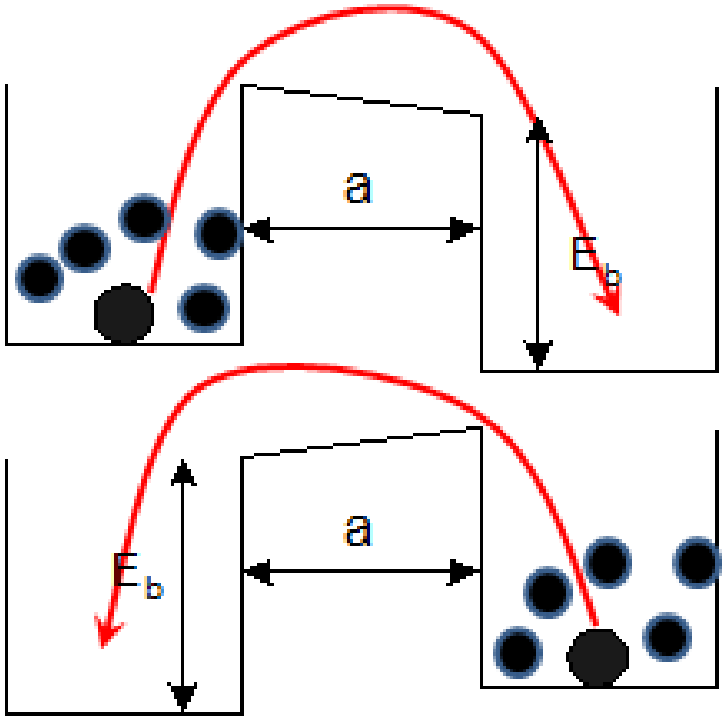



$$\text{Sum} = I_{off}$$

Channel Leakage

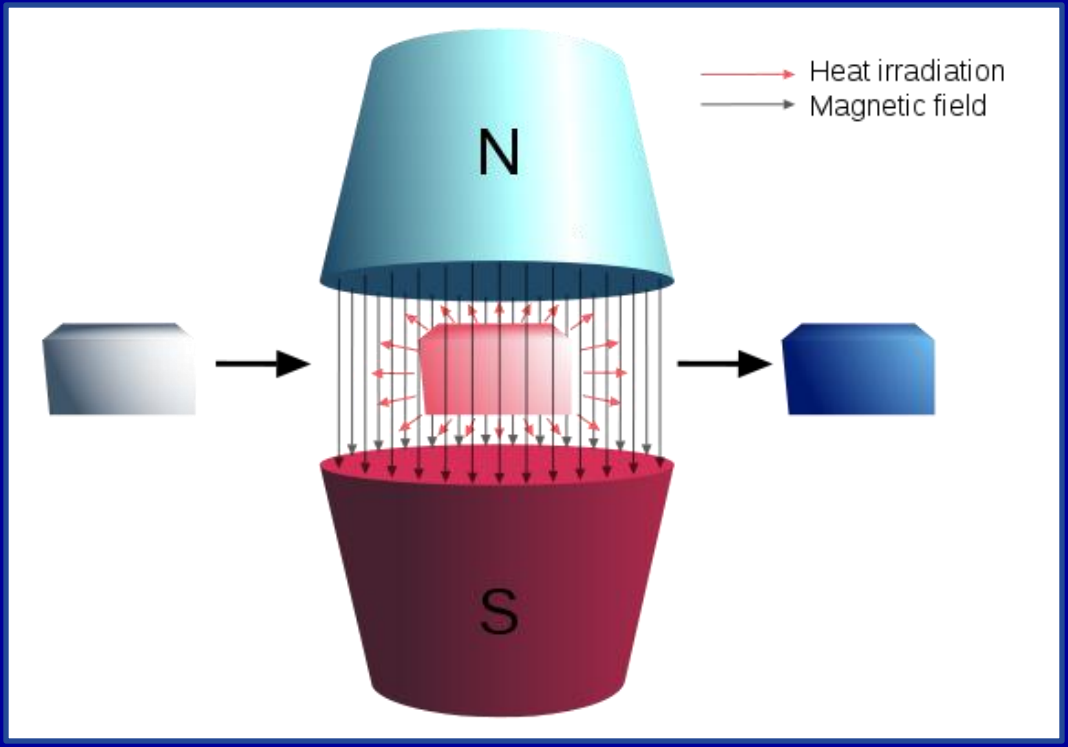
# Temperature, order and transition probability

- Higher temp = > dis order =>higher transition probability
- Lower temp = <dis order =<transition probability
- Alternate ways to introduce order:
  - Magnetic fields
  - Lattice forces




Classical

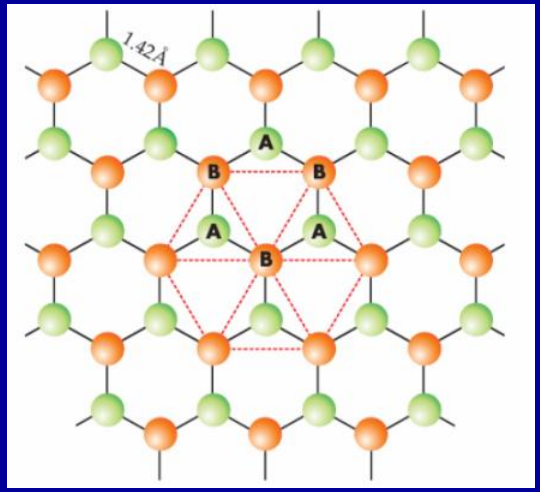
# Magnetic cooling

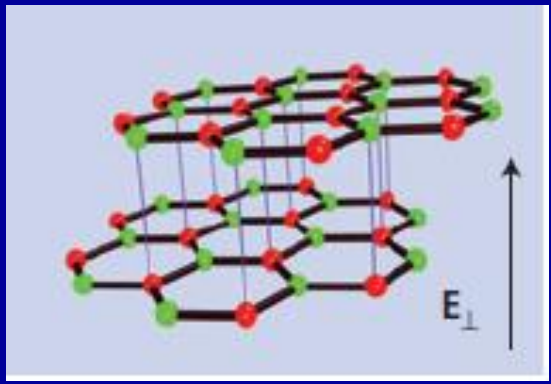



•Magnetoelectric alloy (e.g. Gadolinium) heats up inside the magnetic field and loses thermal energy to the environment, so it exits the field cooler than when it entered

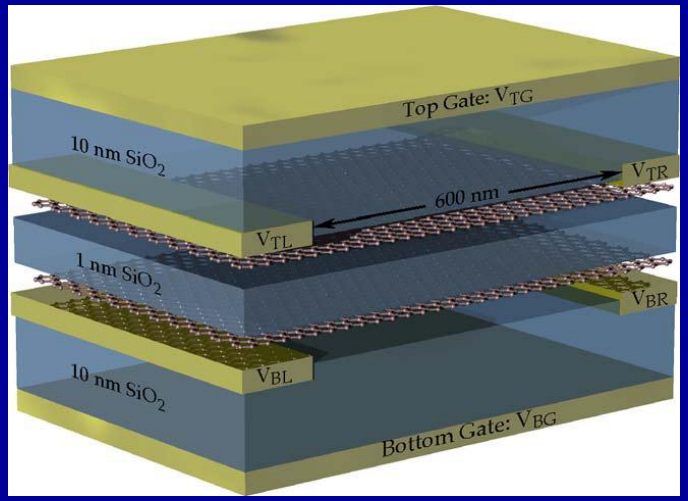
# Monolayer and bilayer graphene has extremely regular atomic structure

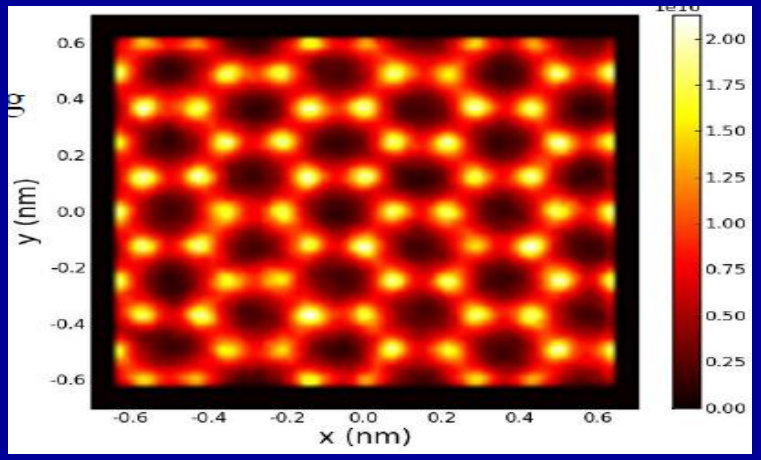
- 2D hexagonal lattice with 2 sublattices A,B
- Ultra high conductivity $\mu_e$~200000 $cm^2$/V-sec
- Regular atomic structure *theoretically* permits ordered electronic states Bose-Einstein condensates usually seen at much lower temperatures
- Sublattices support excitonic condensates
-

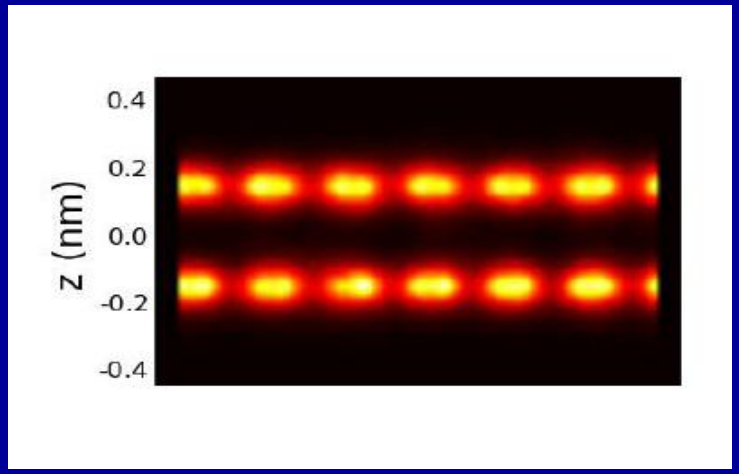


# Bilayer Graphene structure




- Electrons and holes behave collectively as condensates on A and B and form exciton pairs between the layers
- Recombination is allowed only when electron and hole densities exactly match
- Tc ~400 K

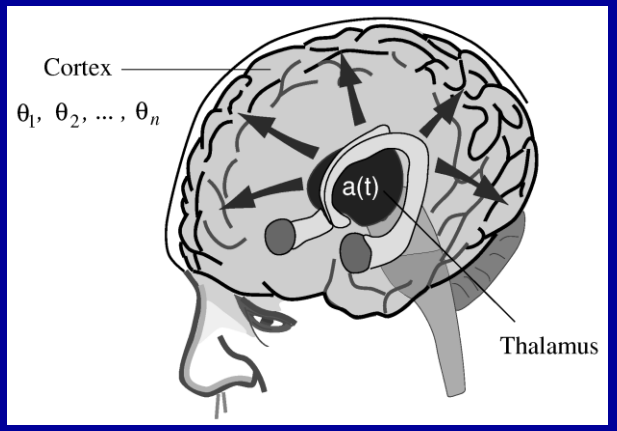
M. Gilbert et.al J Comput Electron (2009) 8: 51–59, DOI 10.1007/s10825-009-0286-y

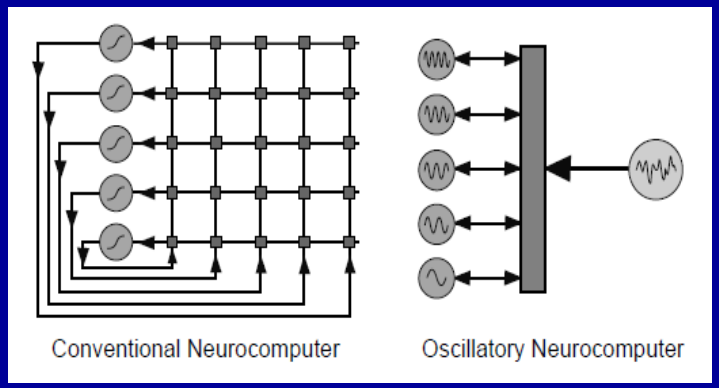
# Outline

- Definition
- Device “scaling” 2003 -2011
- Device “scaling” 2011 - ?
- Beyond the FET
  - More powerful unit devices
  - Temperature, order and non-equilibrium operation
  - Bioinspired, Non Boolean operation
- Conclusions

# Oscillatory model of neurocomputing

- Oscillations experimentally observed in visual cortex after stimulus
- Synchronized oscillations observed in parts of the brain not geometrically close
- Synchronized oscillations proposed as dynamic interconnect media




Hoppensteadt and Izhikevich, Phys Rev L, VOLUME 82, NUMBER 14, April 5, 1999

# Associative memory model of neurocomputing

- Physiologically inspired model of Brain: Human intelligence results from recalling, matching and synthesizing past memory fragments
- Associative system built with NMOS resonance elements

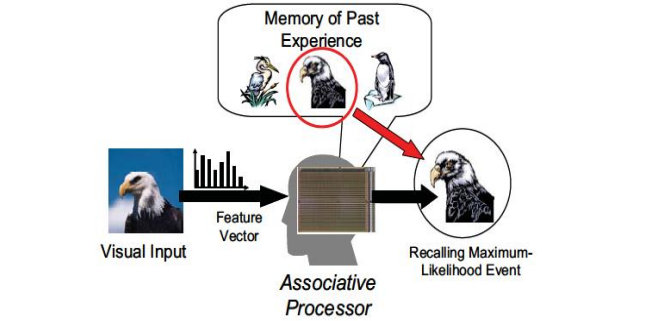
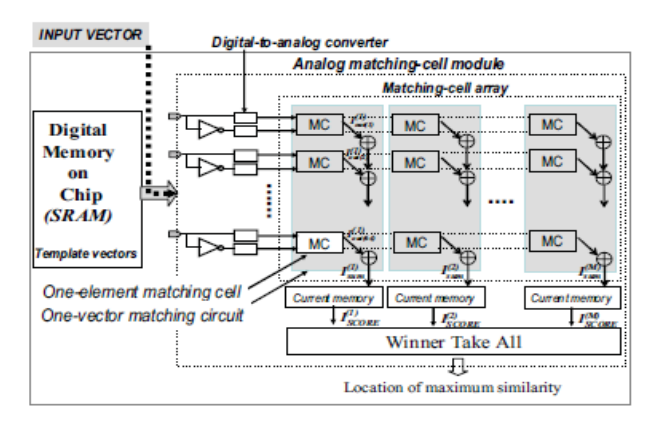



Fig. 1. Psychologically-inspired VLSI brain model based on the associative principle.




Tadashi Shibata, Proceedings of the ECS, Volume 25, Issue 42, 2009

# Conclusions

- Classical Dennard scaling ended at the 130 nm node in 2003
- Moore's law scaling has continued since then
- Advance research enabled seamless transition from Dennard to material scaling
- We have visibility for about 10 years into the future
- Potential solutions beyond the FET are being investigated