Emerging Technologies for Computing

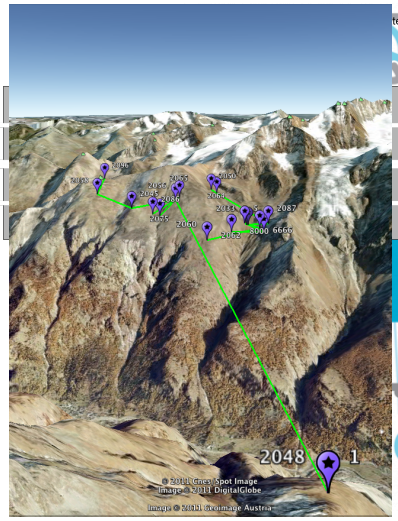
Giovanni De Micheli

Outline

- **Introduction and motivation**
- Technological innovations
  - Emerging nanotechnologies and devices
- Design with emerging technologies
  - Physical and logic synthesis
- New technologies for broader computing systems
  - Device fusion
- Conclusions

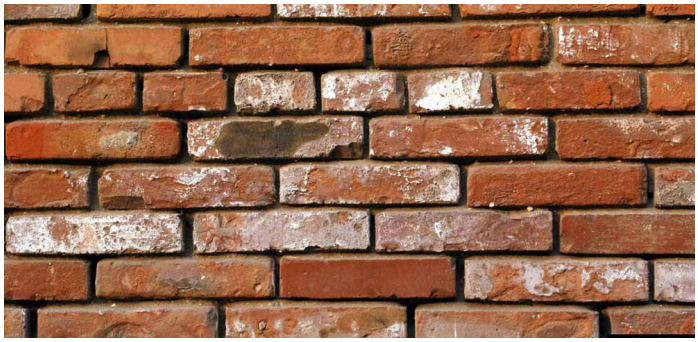
Computing today

Computing today



Walls

- High-performance, energy-proportional servers
  - High speed computation and data retrieval
- Ultra-low power computing and communication
  - Connect myriad of devices for *Internet of Things*



Walls and game changers

- New computing paradigms
  - Quantum computing (superposition, entanglement)
  - Analog computing (memristors, dynamical systems)
  - Neuromoprhic computing
  - In-memory computing
- New materials and devices
  - Enhanced CMOS devices
  - Exploit heterogeneous integration
- Parallelism in algorithms and software
  - Exploit new computational methods
- Use new design methods and tools
  - Revisit hardware synthesis and design techniques

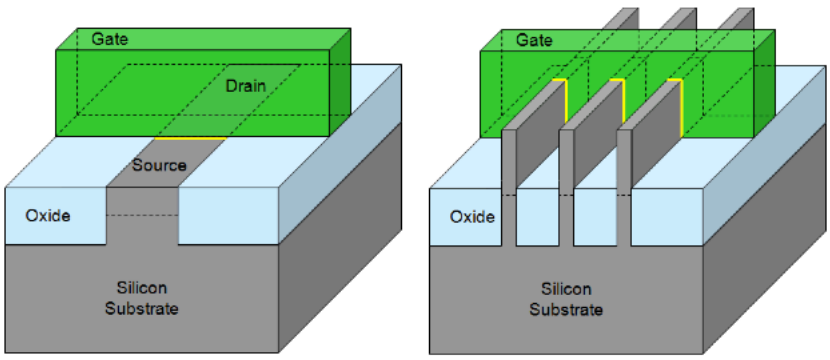
Outline

- Introduction and motivation
- **Technological innovations**
  - Emerging nanotechnologies and devices
- Design with emerging technologies
  - Physical and logic synthesis
- New technologies for broader computing systems
  - Device fusion
- Conclusions

Semiconductor technologies

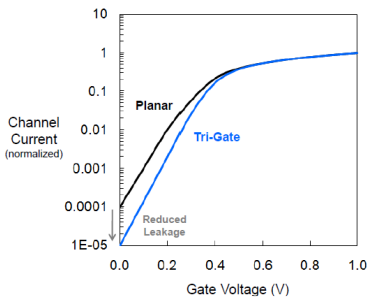
- Most manufacturing technologies have geometries in the nanometer range
- Recently-established nano-electronic technologies
  - Tri-Gate (FinFET) transistors
  - Fully-depleted Silicon on Insulator (FDSOI)
- Downscaling geometries is still effective
- Emerging nano-electronic technologies
  - New materials and devices for processing and memory

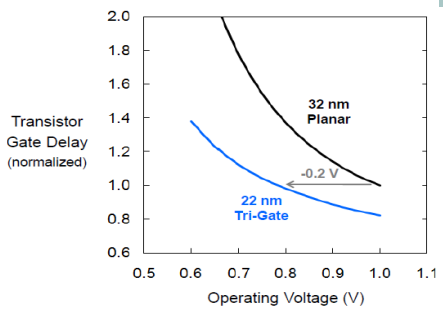
22 nm Tri-Gate Transistors



[Courtesy: M. Bohr]

Tri-Gate vs. planar transistors

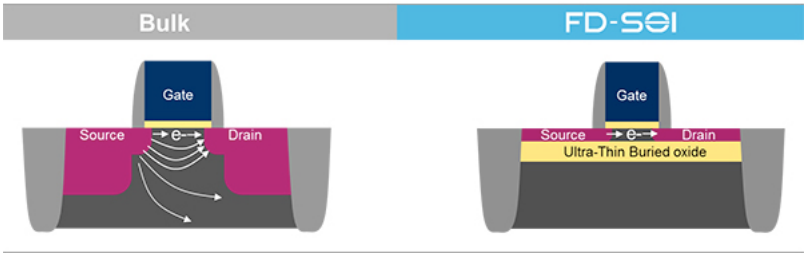




- Smaller current for same gate voltage (when offf)
- Same gate delay for smaller operational voltage

[Source: Intel]

Fully Depleted SoI Transistors

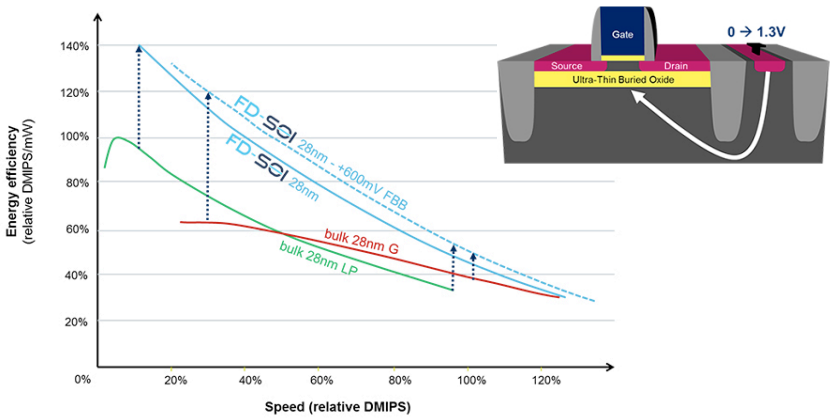


[Courtesy: STMicroelectronics]

- Transistor is built on top of buried oxide (BOX)
- Thin, undoped channel (fully depleted)
- Fine power-consumption control through body bias

11

Fully Depleted SoI Transistors



[Courtesy: STMicroelectronics]

- Energy efficiency
- Forward body biasing

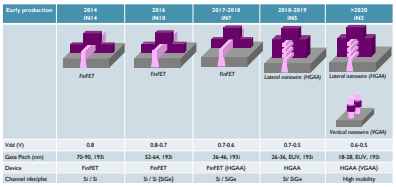
12

Outline

- Introduction and motivation
- **Technological innovations**
  - **Emerging nanotechnologies and devices**
- Design with emerging technologies
  - Physical and logic synthesis
- New technologies for broader computing systems
  - Device fusion
- Conclusions

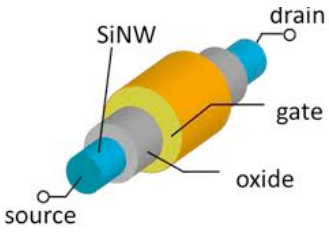
Emerging nano-technologies

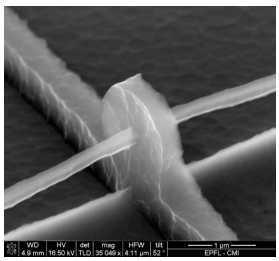
- *Enhanced* silicon CMOS is likely to remain the main manufacturing process in the medium term
  - The 7nm and 5nm technology nodes are on the way
- What are the candidate technologies beyond the 5nm node?
  - Silicon Nanowires (SiNW)
  - Carbon Nanotubes (CNT)
  - 2D devices (Flatronics)
  - … and many others



- What are the differentiators and common denominators from a design standpoint?

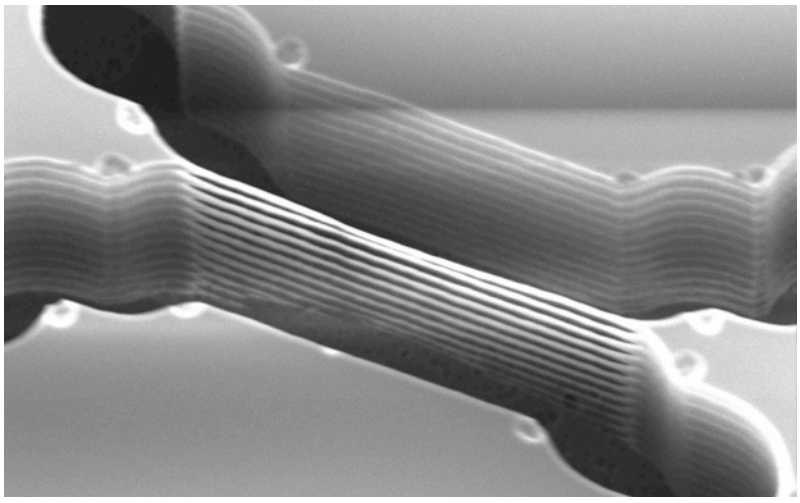
Silicon Nanowire Transistor



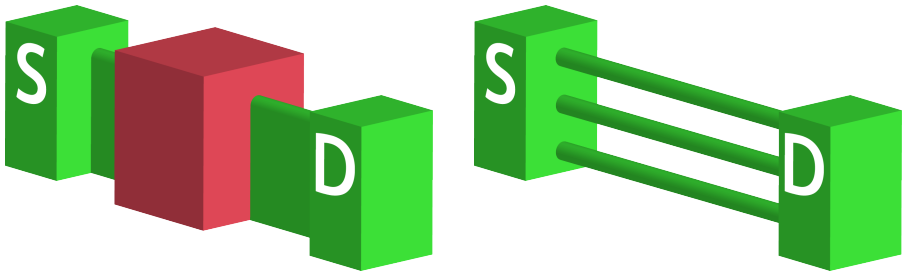


- Fully compatible with CMOS process
- Gate all around
- High I_{on} / I_{off} ratio

Vertically-aligned horizontal SiNW



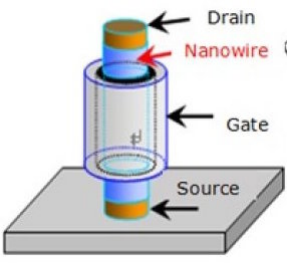
FinFET to Nanowire FET

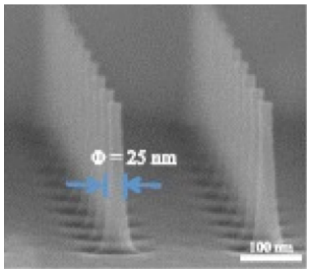


FinFET **NW FET**

18

Vertical silicon nanowire transistors



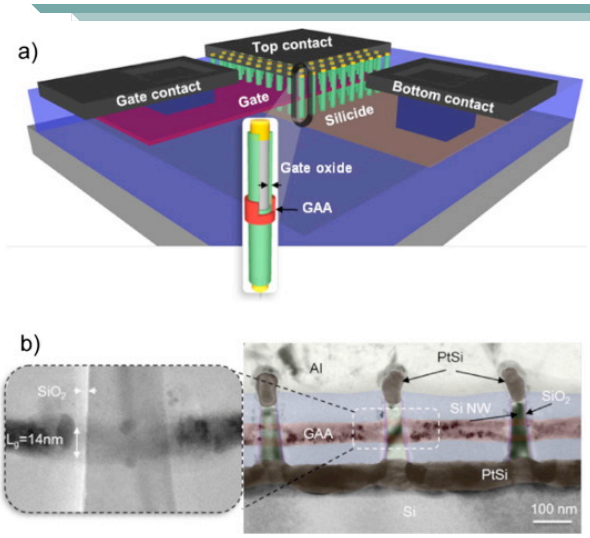


- Fully compatible with CMOS process
- Higher device density
- More complex fabrication process

[Guerfi, Nanoscale 16]

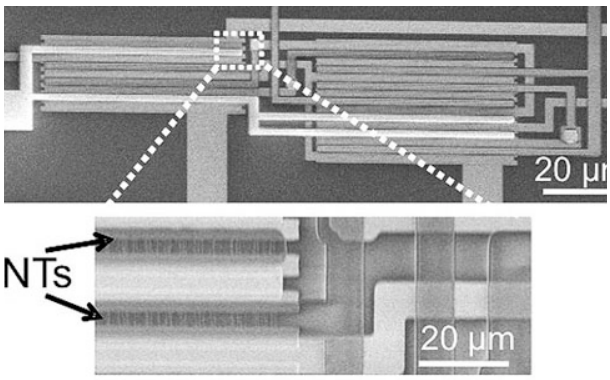
19

Vertical silicon nanowire arrays



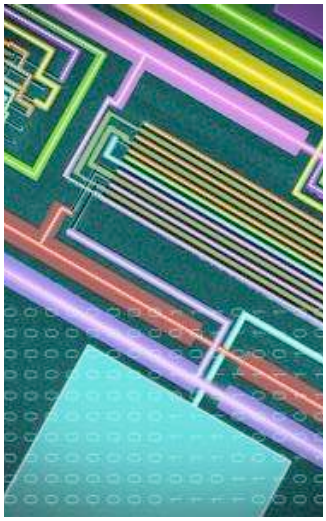
[Larrieu, SS Electronics17]

Carbon Nanotube Transistors

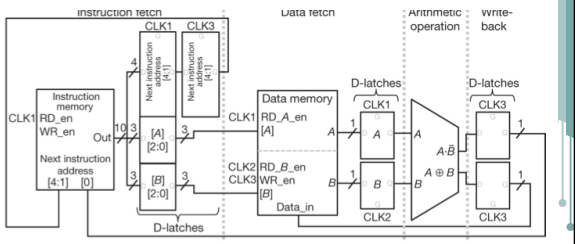


- CNTs benefit from higher mobility and thus higher currents
- CNTs grown separately but can be ported to Si wafers
- Handling CNT imperfection is major design and fabrication issue

CNT nanocomputer

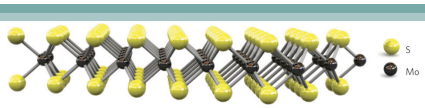


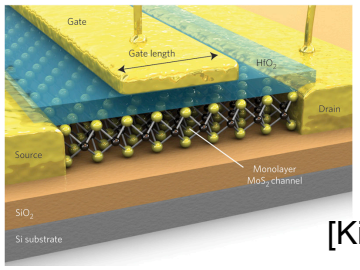
- First CNT computing engine
- Runs 20 MIPS instructions
- Multitasking



[Shulaker, Wong, Mitra et al, NatureNano 13]

2D electronic technologies

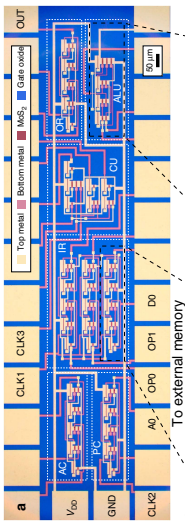




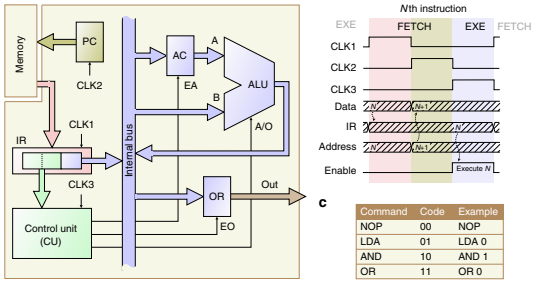
[Kis, Nature Nano 2011]

- Graphene, MoS_2 and other materials
- Single or few atomic layers
- High I_{on} / I_{off} ratio for MoS_2 (10^8) but n-type mainly

MoS_2 nanocomputer



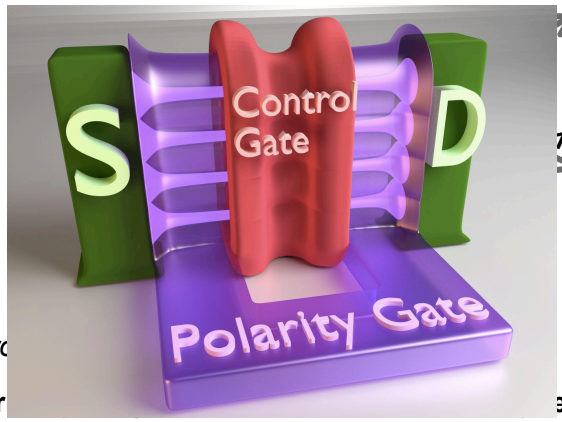
- First MoS_2 computing engine
- Runs 4instructions
- 115 N-xtors (enhancement load)
- 2 micron feature size

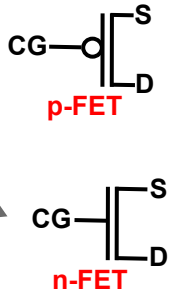


Command	Code	Example
NOP	00	NOP
LDA	01	LDA 0
AND	10	AND 1
OR	11	OR 0

[Wacter et al. Nature Comm, 2017]

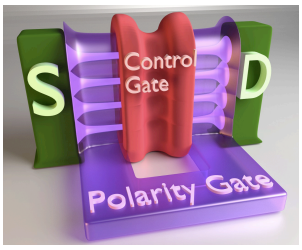
Double gate SiNW FET

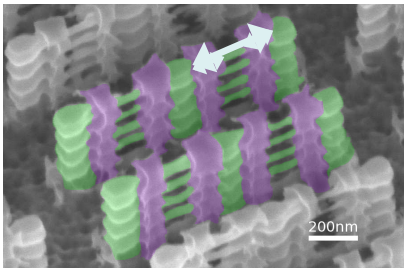


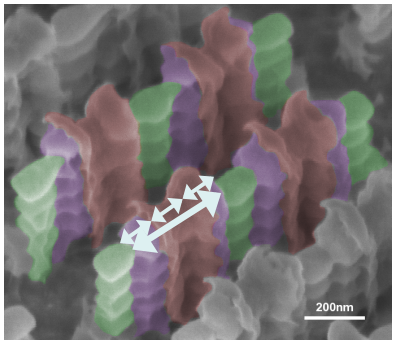


- *Electro*
- **Electr** **e** or **n-type**
- **Comparator**-activated switch

Fabricated device view



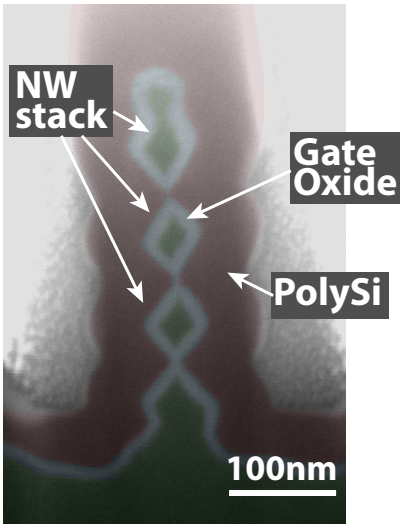


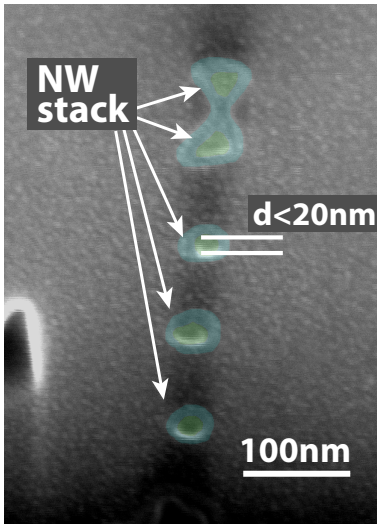


- 100 nm gate segments
- 350-nm long nanowires
- 20-40 nm wire diameter

26

Device cross-sections



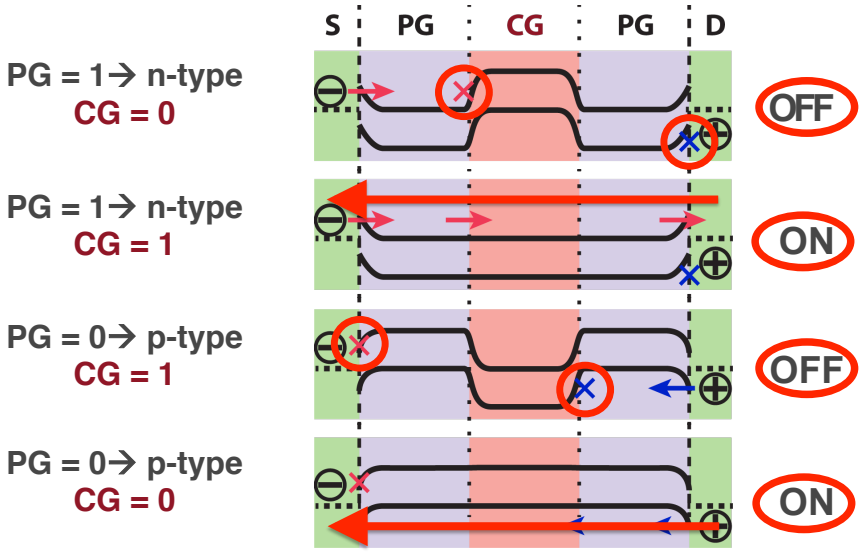


2

M. De Marchi et al., IEDM 2012, TNANO 2013.

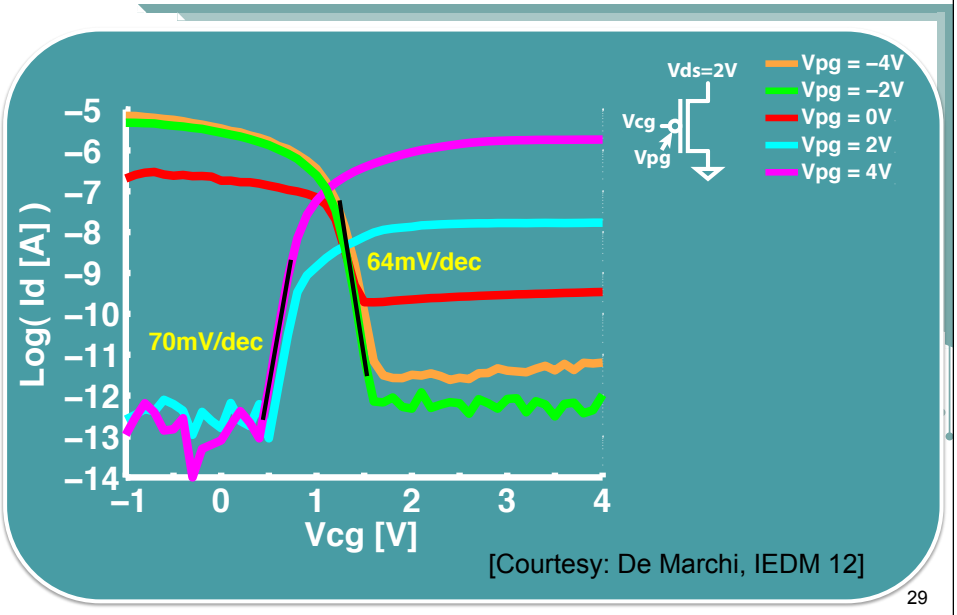
27

Device working principle



28

Device I_d/V_{cg}



[Courtesy: De Marchi, IEDM 12]

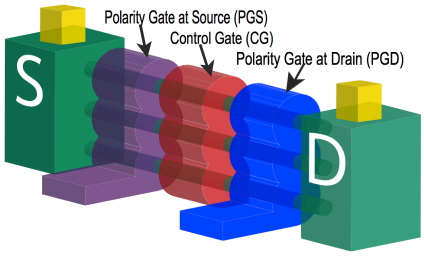
29

Similar devices

- Controlled devices can be realized with various materials and shapes (e.g., FINFET)
- SiNW controlled-polarity devices can be made with one polarity gate **on one side** [Heinzig]
- Polarity-gate bias can enable:
  - Steep Subthreshold
  - Multiple threshold voltages

Three-independent-gate SiNWFET

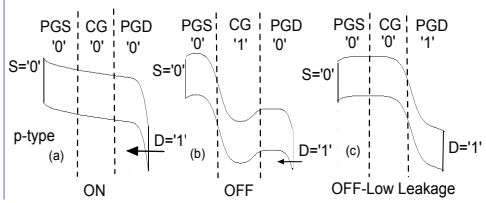
❖Structure



- **Vertically stacked nanowires**
- **3 independent gate regions**
- **Schottky barrier contacts at S/D**
- **Polarity and Vt controllability**

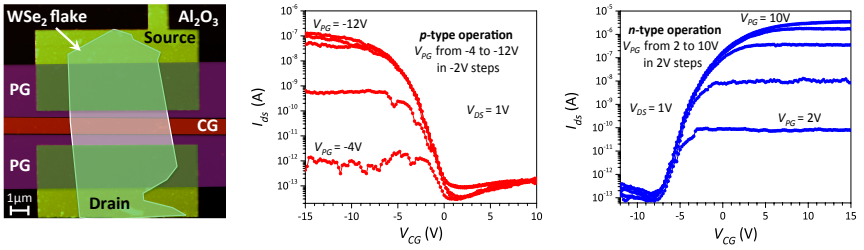
❖Electrostatic control

S	D	PGS	CG	PGD	State
0	1	0	0	0	ON (P-type)
		1	1	1	ON (N-type)
		0	1	0	OFF (LVT)
		1	0	1	OFF (LVT)
		0	0	1	OFF (HVT)
		0	1	1	OFF (HVT)



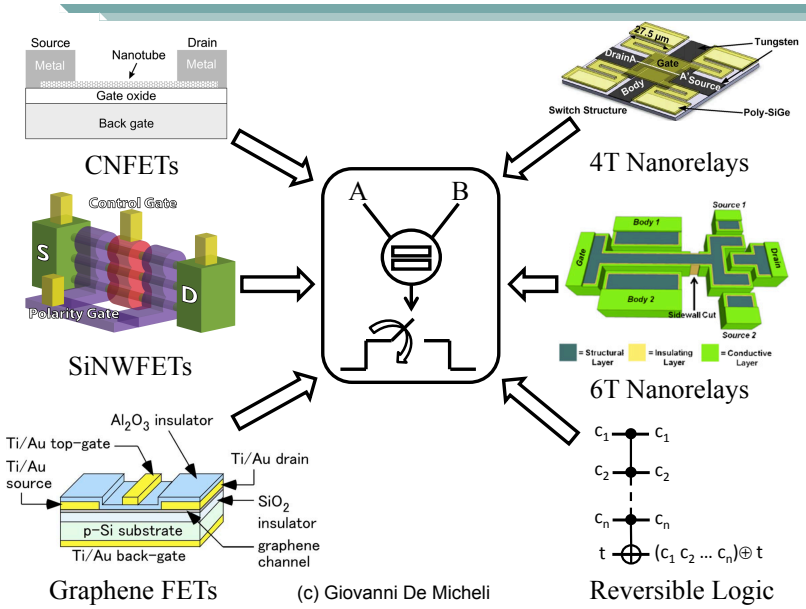
Controllable polarity in 2D

2D Controllable-polarity transistor (WSe_2)



[Resta, Scientific Reports 2016]

Modeling various emerging nanogates



Outline

- Introduction and motivation
- Technological innovations
  - Emerging nanotechnologies and devices
- **Design with emerging technologies**
  - **Physical and logic synthesis**
- New technologies for broader computing systems
  - Device fusion
- Conclusions

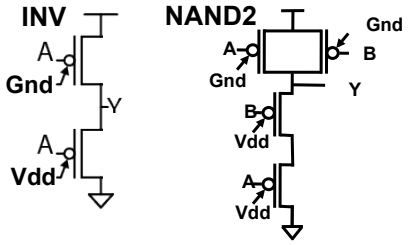
Logic level abstraction

- Three terminal transistors are switches
  - A loaded transistor is an *inverter*
- Controllable-polarity transistors compare two values
  - A loaded transistor is an *exclusive or* (EXOR)
- The intrinsic higher computational expressiveness leads to more efficient data-path design
- The larger number of terminals must be compensated by smart wiring
- Fine-grained programmability

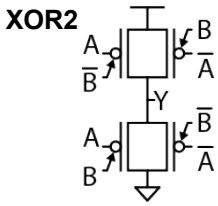
Logic cell design

- CMOS complementary logic is efficient only for negative-unate functions (INV, NAND, NOR…etc)
- Controllable-polarity logic is efficient for all functions
- Best for XOR-dominated circuits (binate functions)

Negative Unate functions	Binate functions
INV, NAND2	XOR2

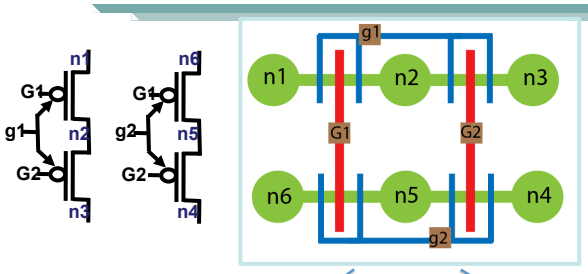


Similar to regular CMOS

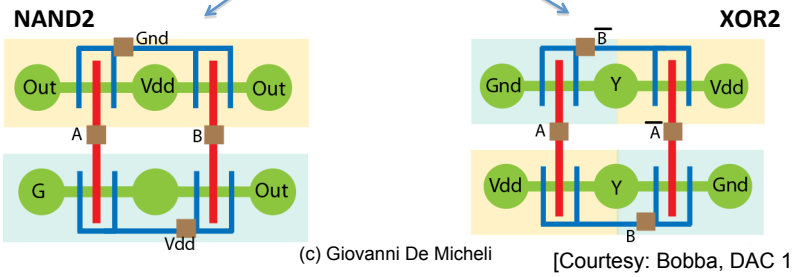


Only 4 transistors when compared to 8 transistors with a regular CMOS

Layout abstraction and regularity



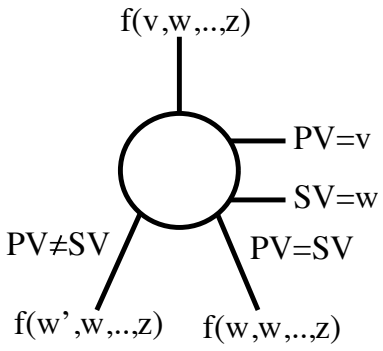
Two transistor pairs grouped together



Logic Design Abstraction:
Biconditional Binary Decision Diagrams

- Native **canonical** data structure for logic design
- *Biconditional* expansion:

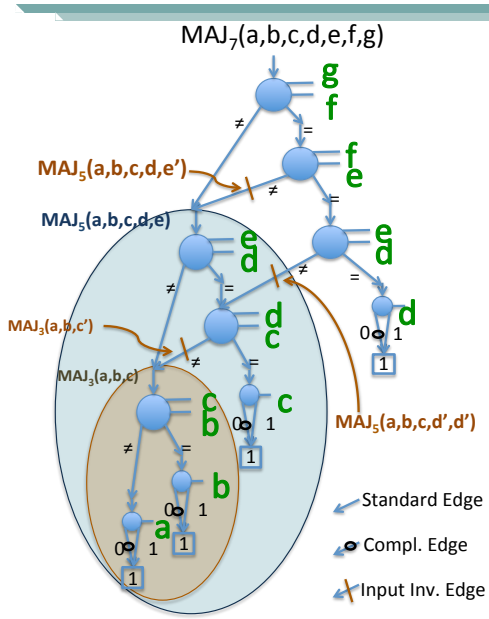
$$f(v,w,..,z) = (v \oplus w) f(w',w,..,z) + (v \overline{\oplus} w) f(w,w,..,z)$$



- Each BBDD node:
  - Has two branching variables
  - Implements the *biconditional* expansion
  - Reduces to Shannon's expansion for single-input functions

39

BBDDs are Compact (Majority Function)



Number of nodes of MAJ(n):

$$\frac{1}{8}n^2 + \frac{1}{2}n + \frac{11}{8}$$

MAJ(3): 4

MAJ(5): 7

MAJ(7): 11

....

40

New logic models and data structures

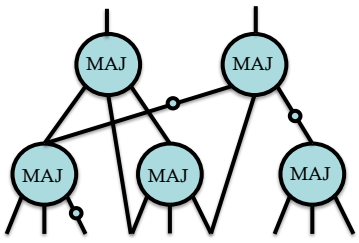
- Design with emerging devices requires exploring new logic models combining:
  - XOR primitives (programmable complementation)
  - MAJority functions (programmable AND/OR)
- The resulting models and algorithms have wide applicability to logic design (including CMOS)

Majority logic: a new/old paradigm?

In fact <x,y,z> is probably the most important ternary operation in the entire universe, because it has amazing properties that are continuously being discovered and rediscovered.

Donald Knuth, *The Art of Computer Programming, Vol. 4A*

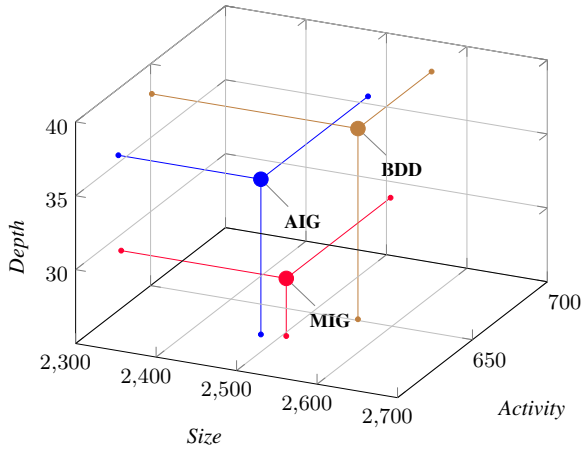
- Majority Inverter graphs as data structure for logic synthesis
- Reachable design space
- Surprising experimental results



[Courtesy: Amaru', DAC 14]

Experimental Results: MCNC circuits

Logic Synthesis: MIGthy



MIGs & AIGs better than BDDs

MIGs size & activity ~ AIGs

MIGs depth -20% w.r.t AIGs

CMOS Design Results

Advanced 22nm CMOS
MIG as front-end to LS & PD

Well-established 90nm CMOS
MIG as front-end to LS & PD

Behavioral

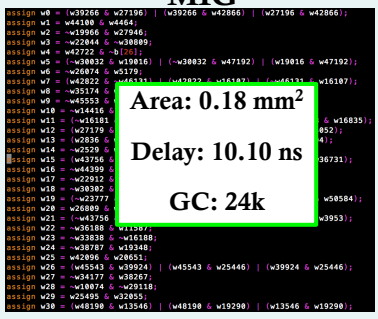
```
module div32 (a, b, quotient_uns, quotient_tc, remainder_uns,
 remainder_tc);
 parameter width = 32;
 input [width-1 : 
 output [width-1 :
 output signed [wi                    r_tc;
 // operators for                     ainder
 assign quotient_ur
 assign quotient_tc = $signed(a) / $signed(b);
 assign remainder_uns = a % b;
 assign remainder_tc = $signed(a) % $signed(b);
endmodule
```

Area: 0.21 mm²

Delay: 11.22 ns

GC: 37k

MIG



Area: 0.18 mm²

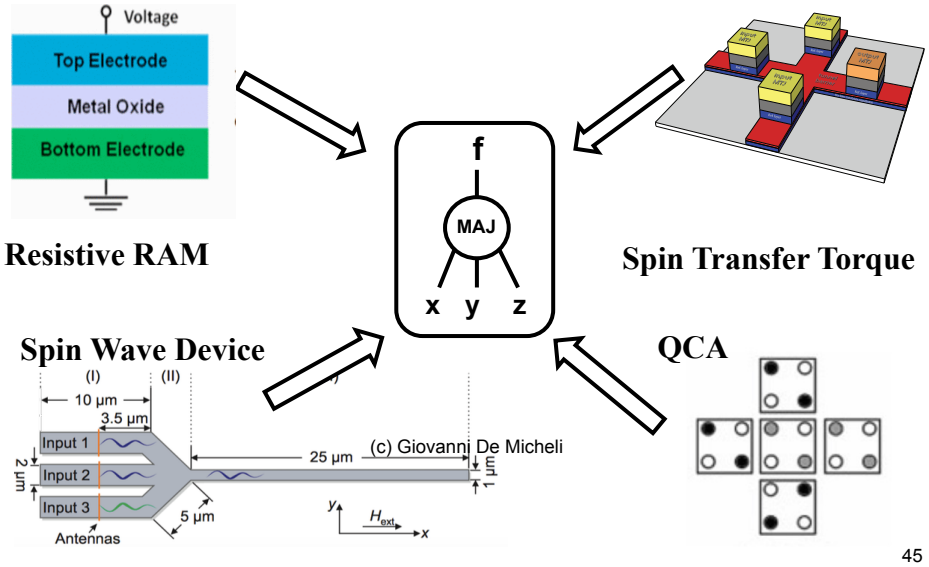
Delay: 10.10 ns

GC: 24k

All circuits underwent formal verification with success

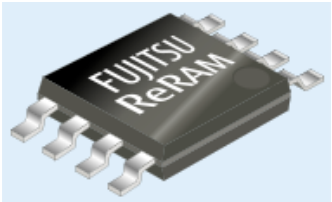
Both circuits underwent formal verification with success

Modeling various emerging nanogates



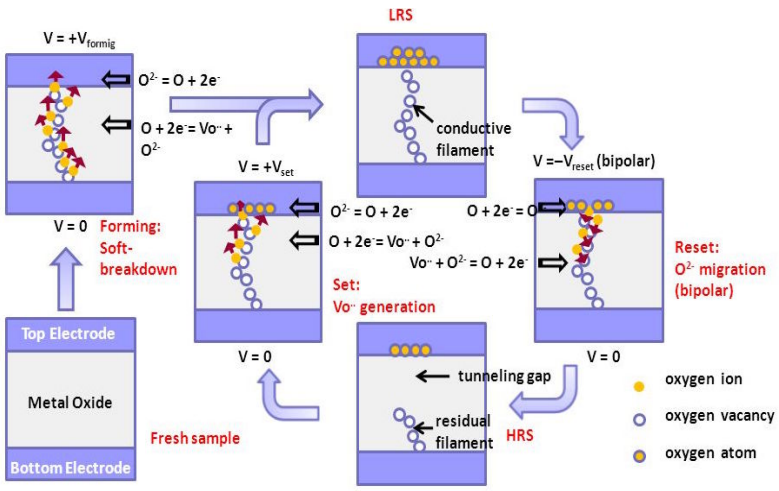
Resistive RAMs

- Non-volatile, low-power dense RAM arrays
- Based on resistive switching:
  - Various physical mechanisms
- Can be realized in the BEOL
  - 3D integration



[Fujitsu MB85AS4MT]
4M (512kx8)

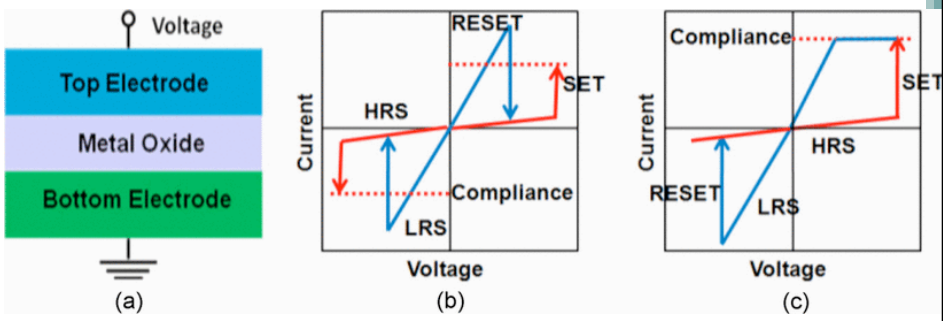
Oxide ReRAM switching



[Su et al.Functional Metal Oxide Nanostructures, 2011]

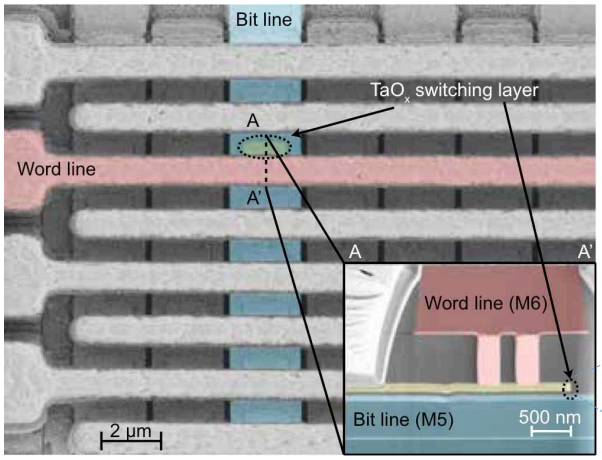
Oxide Resistive RAMs

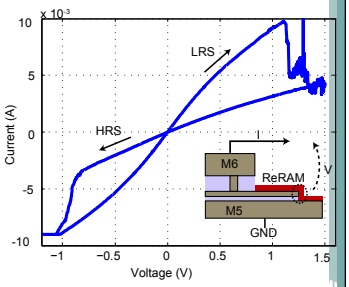
- Formation of an oxygen vacancy filament
  - Reversible write: set/reset
  - Low-current read
  - Unipolar and bipolar switching

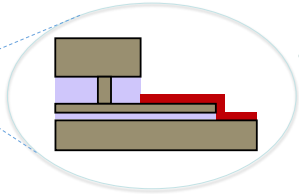


[P. Wong et al 2016]

ReRAM integrated devices



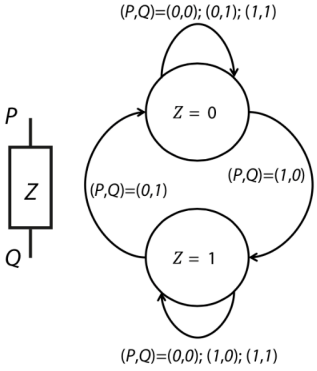




TiN/TaOx/TiN forming-free, V_{set} = -1V, V_{reset} = 1.3V

Sandrini, Microelectr. Eng. 2015 [Courtesy: Leblebici] 49

ReRAM as computational element



P	Q	Z	Z_n
0	0	0	0
0	1	0	0
1	0	0	1
1	1	0	0

$Z_n = P \cdot \overline{Q}$

P	Q	Z	Z_n
0	0	1	1
0	1	1	0
1	0	1	1
1	1	1	1

$Z_n = P + \overline{Q}$

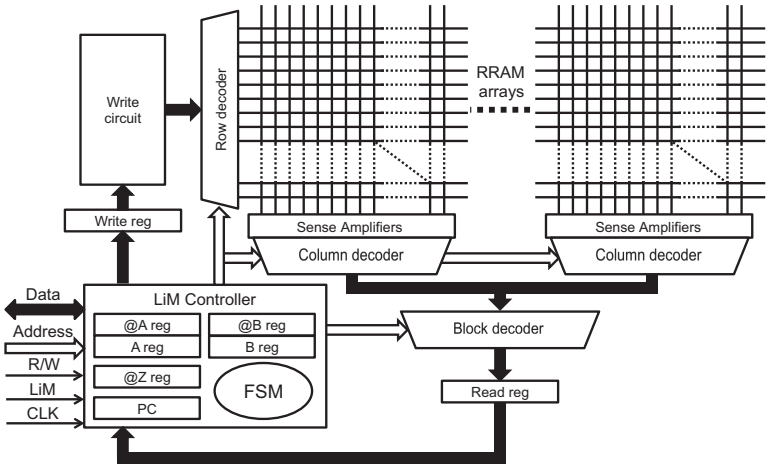
Z_n = PZ+Q'Z+PQ'
Z_n = Maj(P,Q',Z)

(c) Giovanni De Micheli 50

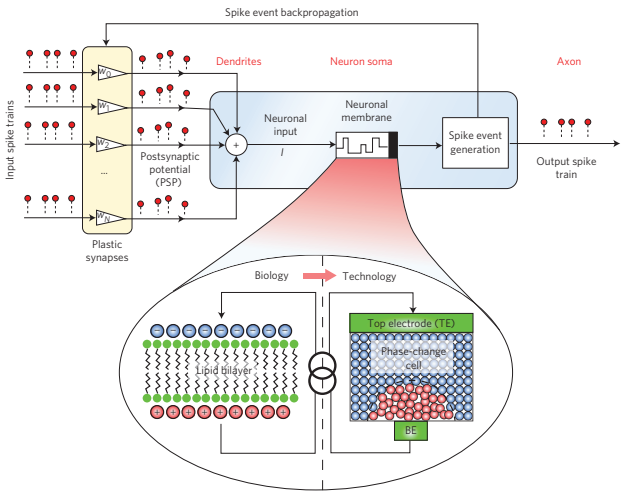
In memory processing/computing

- New paradigm for big data processing
- Data in main memory and operated upon locally
- Hardware:
  - Large memory arrays with embedded processors
  - Technology compatibility
  - What about ReRAM arrays?
- Convert small portion of memory array to perform computation
  - Map data flow computation to memory cell control
  - Create appropriate local controller
  - Use memory for storage

PLIM Architecture



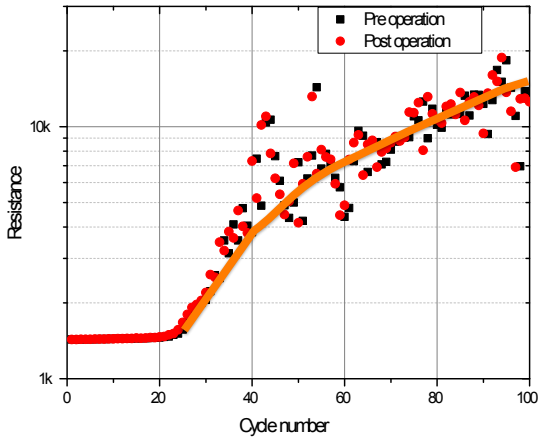
ReRAMs for artificial neurons



[Tuma et al, NatureNano 16]

Resistance modulation (# of pulses)

Pulses: -1V, 1us, 20%, 10ms between pulses



10x resistance modulation with less than 100 pulses

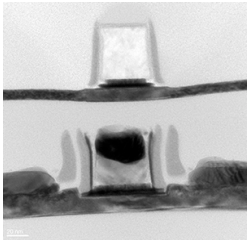
[Courtesy: Leblebici]

Outline

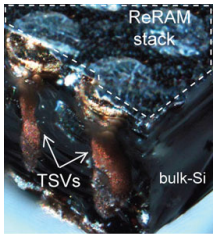
- Introduction and motivation
- Technological innovations
  - Emerging nanotechnologies and devices
- Design with emerging technologies
  - Physical and logic synthesis
- **New technologies for broader computing systems**
  - **Device fusion**
- Conclusions

What is next ?

- Technology hybridization
  - Fusion of sensing and computing
  - 3D integration with sensors
- Heterogeneous integration
  - Sequential integration



[Batude, IEDM 14]



[Sacchetto, Nanoscale12]

SiNWs: ideal biosensing support

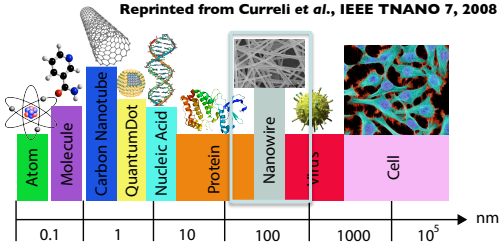
1. Nano size
   - Best interface to proteins

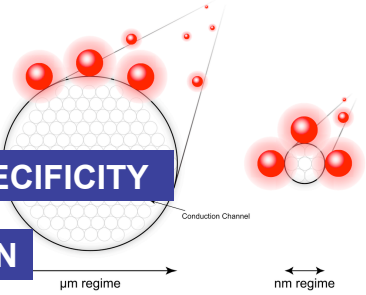
SENSITIVITY

2. Surface-to-volume ratio
   - Larger interaction area
   - Charge confinement

3. Silicon biomodification → **SPECIFICITY**

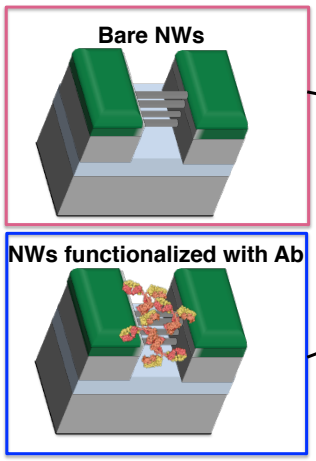
4. Compatibility → **INTEGRATION**

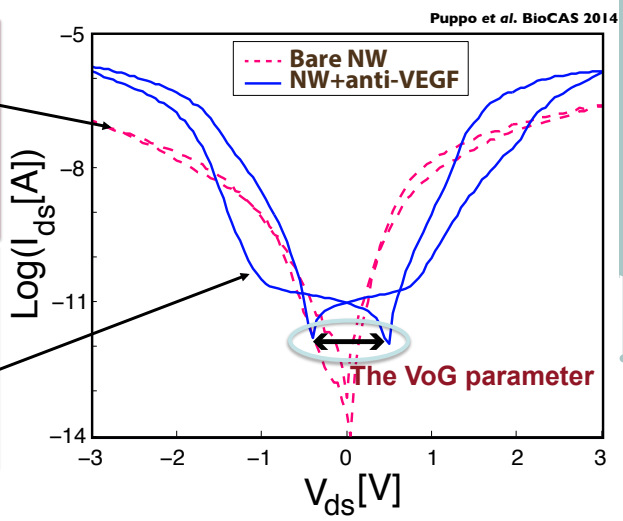




58

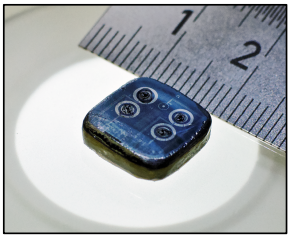
SiNW biosensors



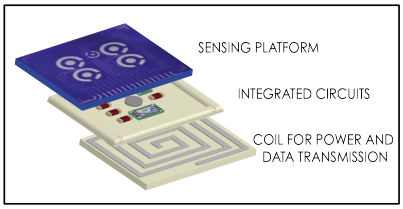


59

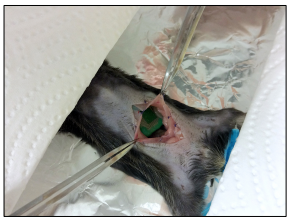
Example of sensor integration



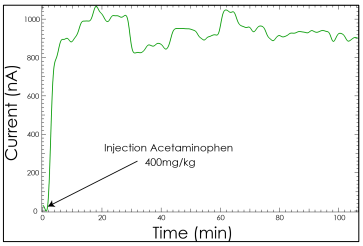
Muti-sensor for lab animals



Chip layers



Chip implant in mouse



Step injection response

[Baj-Rossi, 15]

Outline

- Introduction and motivation
- Technological innovations
  - Emerging nanotechnologies and devices
- Design with emerging technologies
  - Physical and logic synthesis
- New technologies for broader computing systems
  - Device fusion
- **Conclusions**

Conclusions

- Computing is evolving in various directions and permeates everyday life and activities
- Computing is still mainly based on *von Neuman* architectures, switching theory and silicon devices
- New materials and devices can change the physical substrate of computation, making it more efficient and broader in scope
- Progress will require a strong coordination of technology, architecture and software as well as design methods and tools

Special thanks

- Yusuf Leblebici (EPFL)
- Pierre Emmanuel Gaillardon (U. Utah)
- Davide Sacchetto (CSEM)
- Michele de Marchi (ESPROS Photonics)