# X-GDR BarCamp sur les Défis d'Implémentation de l'IA
# Sécurité, Fiabilité, Soutenabilité et Nouvelles Technologies

## Emerging technology-based 3D compute cubes for edge intelligence

FVLLMONTI

*Ian O'Connor*[1], David Atienza[2], Jens Trommer[3],

Oskar Baumgartner[4], Guilhem Larrieu[5] and Cristell Maneux[6]

[1]Lyon Institute of Nanotechnology, FR; [2]École Polytechnique Fédérale de Lausanne (EPFL), CH;
[3]Namlab gGmbH, DE; [4]Global TCAD Solutions, AT; [5]LAAS – CNRS, FR; [6]University of Bordeaux, FR

# Thanks to the team!

FVLLMONTI

we are

RUST

**R**eliable
**U**ltra-low power
**S**ecure
compu**T**ing

Alberto Bosio
Damien Deleruyelle
Bastien Deveautour
Cédric Marchand
Sara Mannaa
Ian O'Connor

Giovanni Ansaloni
Alireza Amirshahi
David Atienza

Jens Trommer
Cigdem Cakirlar
Thomas Mikolajick

Oskar Baumgartner
Zlatan Stanojevic
David Pirker

Guilhem Larrieu
Sylvain Pelloquin
Konstantinos Moussakas

Cristell Maneux
Chhandak Mukherjee
Marina Deng
François Marc
Jean-Luc Rouas
Axel Guihard

# How do Han Solo and Greedo communicate?


- In-ear translation device!

https://www.timekettle.co

3000mAh

BLE

0.5-3s latency

5G

100M parameters (32b)
100M MACs/inference
95% accuracy rate

- But it needs the cloud …
  - Could latency nullify Han's lightning reflexes?
  - Could Jabba hack the communication to find Han?
  - Is there even a signal on Tattoine?!

# Edge intelligence

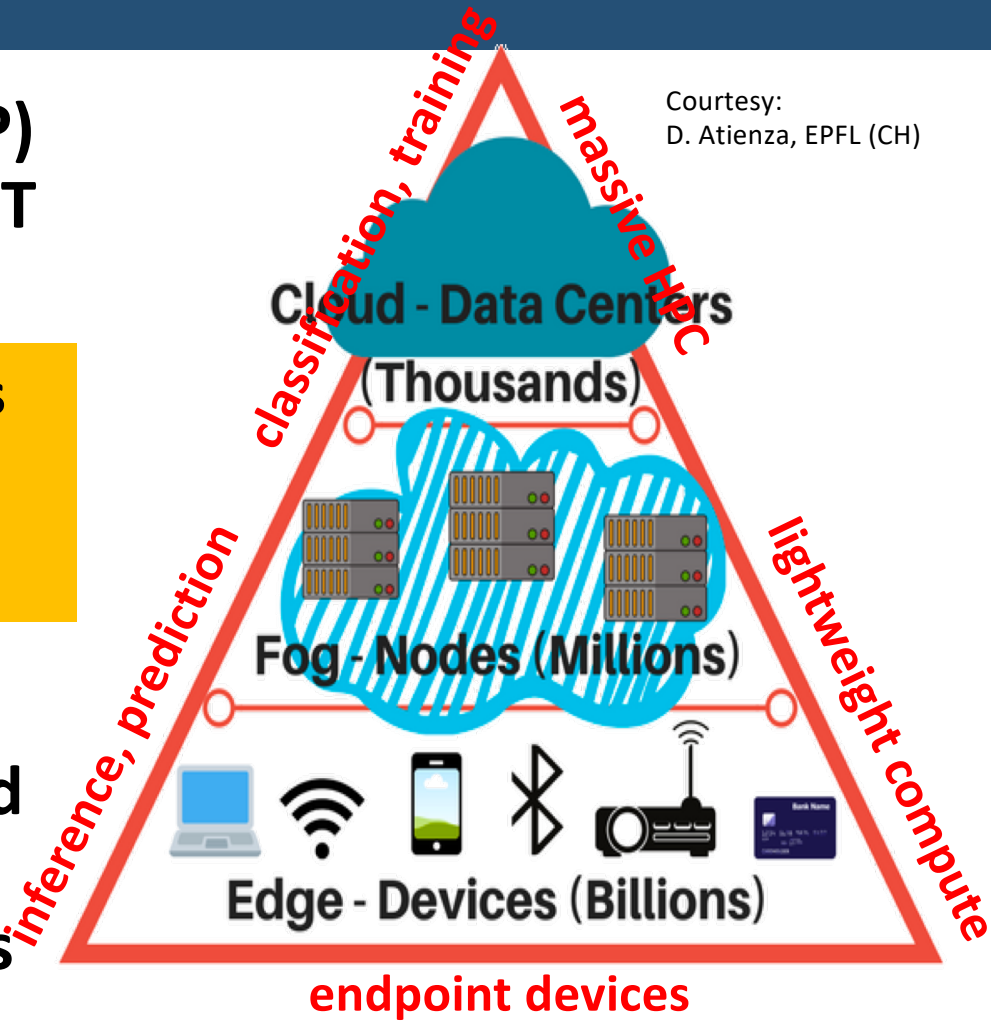
- **General purpose AI models (CV+NLP) are 10k larger! e.g. WuDao2.0 at > 1T parameters**

**Exploding carbon footprint of datacenters**

**Data transport energy is a large fraction**

**AI energy doubles every 3-4 months**

- **So: move intelligence to the edge!**
- **But: edge AI (<10nm) models limited to ~10M parameters (storage, computing) due to SWaP constraints**




Courtesy:
D. Atienza, EPFL (CH)

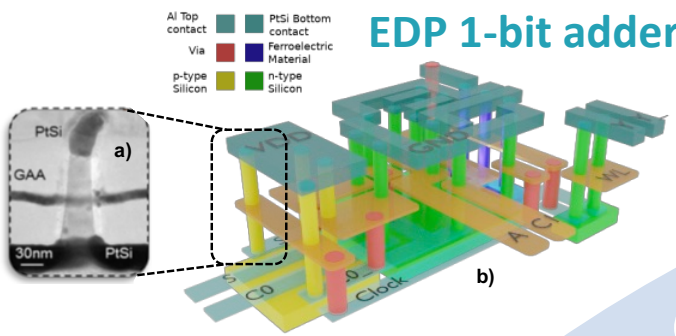
# FVLLMONTI

Ferroelectric Vertical Low energy Low latency low volume Modules fOr Neural network Transformers In 3D

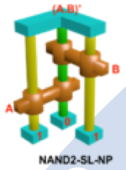
Horizon 2020
European Union Funding
for Research & Innovation

Grant no 101016776

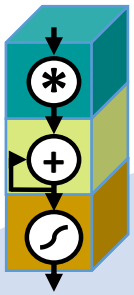
## Compact and low EDP 1-bit adder




## High-expressivity (non-volatile) logic cells



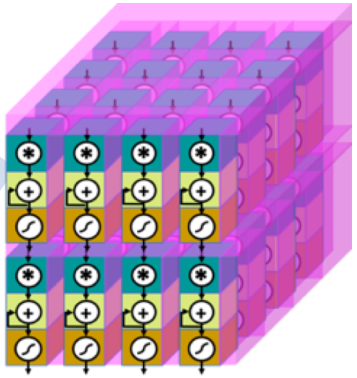

Dedicated library of 3D logic cells leveraging **VNWFET** devices



**Versatile** VNWFET-logic cell based 3D neural network compute cube ($N^2C^2$) for NN-based architecture design

## $N^2C^2$ concept

**Scalable** and versatile 3D architectural model leveraging reconfigurable 3D interconnect framework

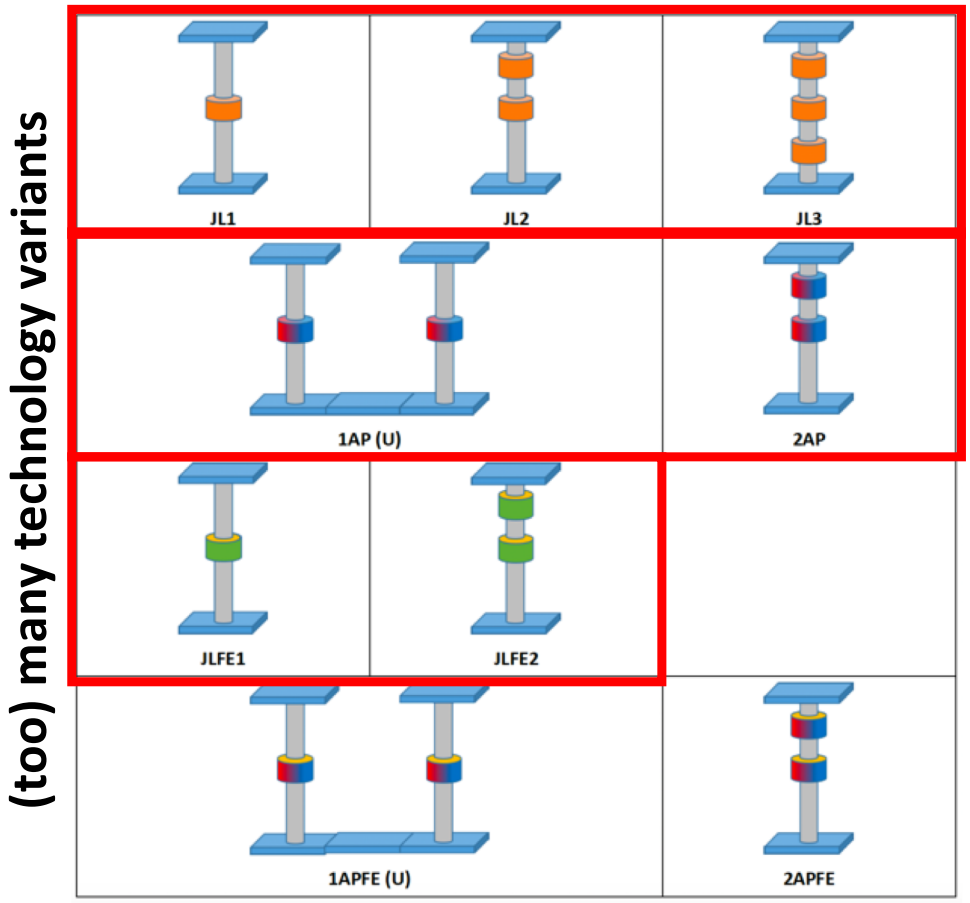


**Physical regularity**
**Functional versatility**
**In-memory vector processing**

**Suitable for aggressive exploration of hw/sw co-design + approximate computing for machine learning**

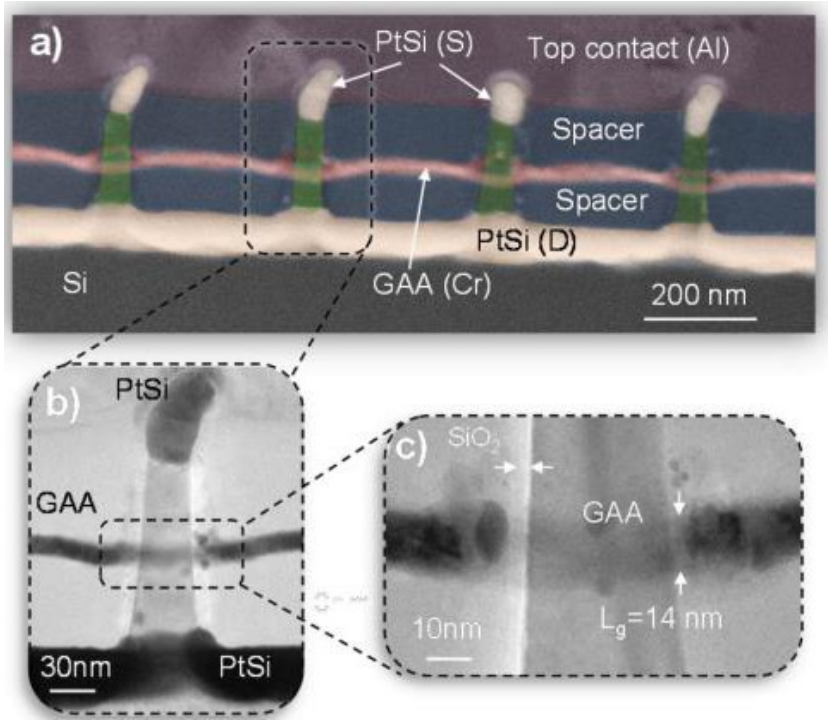
# Agenda

- Why vertical?
  - CStatic logic JL1, JL2, JL3
- Why ferroelectric?
  - Complementary non-volatile logic JLFE1, JLFE2
- Why ambipolar?
  - Ambipolar reconfigurable logic 1AP, 2AP
- The importance of DTCO
  - DTCO scope (device, model, technology variants, template)



# Why vertical?

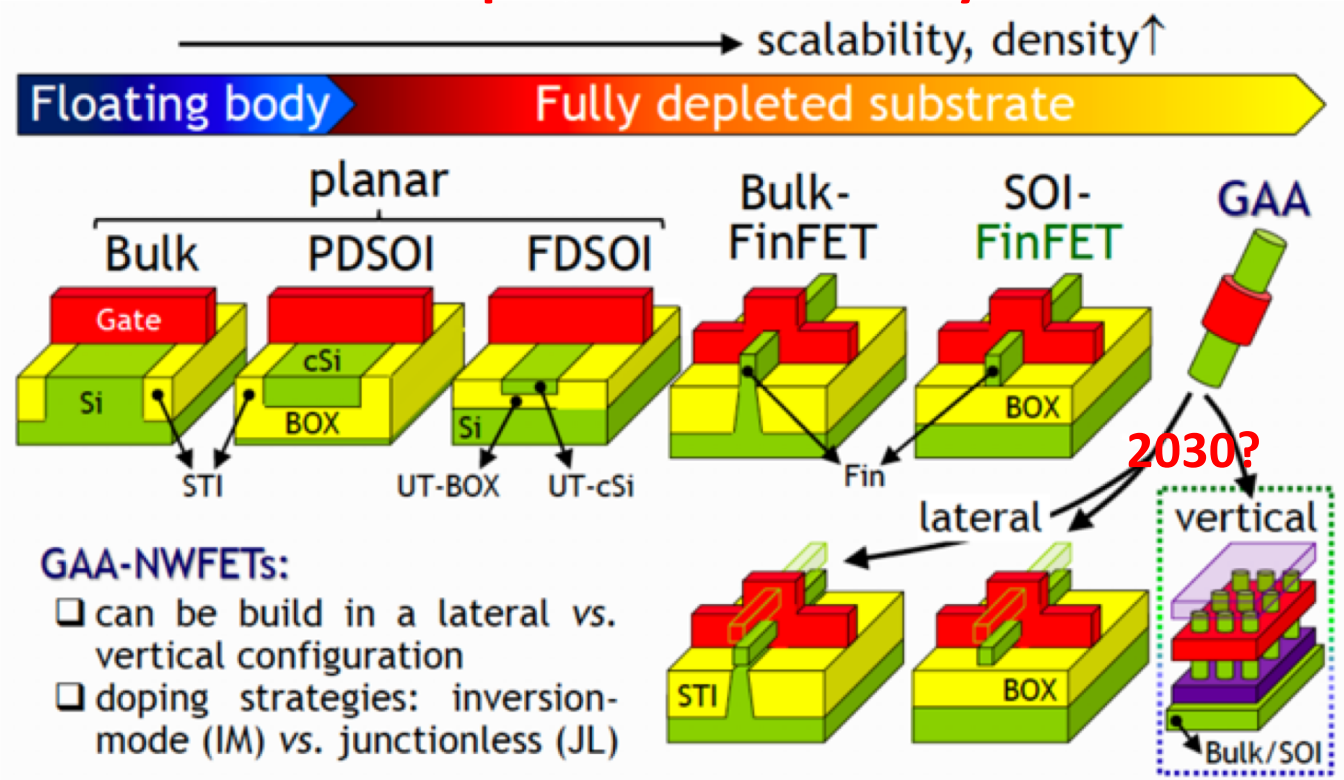
- **Vertical transistor technologies:**
  - **Minimal gate length is not limited by lithography resolution**
  - **Straightforward stacking of multiple gate-all-around devices in series**
- **Scaling continues! Gate capacitance, device-device wiring capacitance reduced – energy efficiency**
- **Not a new idea … but renewed interest**




G. Larrieu, (2013) Nanoscale 5 (6), 2437-2441.

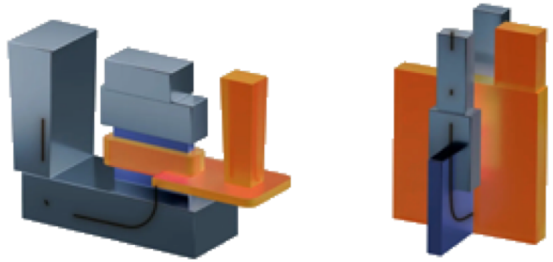
# In the news

## Int. Roadmap for Devices and Systems™ 2021 Update: More Moore




A. Veloso et al., "Challenges and opportunities of vertical FET devices using 3D circuit design layouts", IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S), 2016.

*"… next-generation transistors that will enable a trend of **smaller**, more **powerful** and **energy-efficient** devices …"*



https://research.ibm.com/blog/vtfet-semiconductor-architecture

**2022**

# Energy-efficiency analysis - EDP

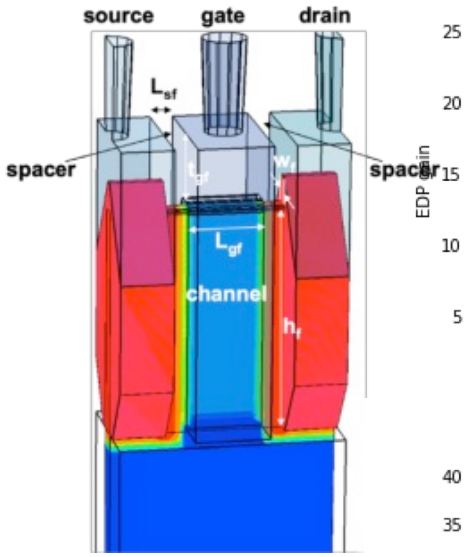
- **EDP (Energy×Delay Product) is a useful metric to compare energy-efficient circuits**
- **The ratio between EDP values of FinFET and VNWFET technologies, for identical supply voltage, can be expressed as**

$$G_{Evf} = \frac{C_{Lf}^2}{C_{Lv}^2}\frac{I_{satv}}{I_{satf}}$$

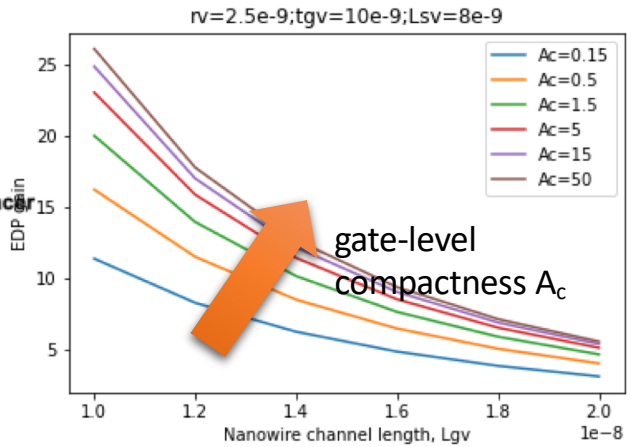
$I_{sat}$: saturation current based on device geometry
$C_L$: load capacitance based on contacts
$_V$: VNWFET device ; $_F$: FinFET device

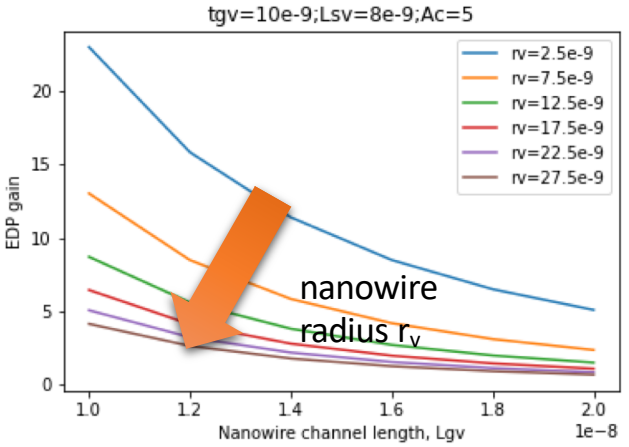
$$G_{Evf} = \frac{\left[\frac{\varepsilon_{ox} L_{gf} 2\left(h_f+w_f\right)}{EOT}+\frac{\varepsilon_s\, 2t_{gf}\left(2t_{gf}+h_f+w_f\right)}{L_{sf}}+C_{wf}\right]^2}{\left[\frac{\varepsilon_{ox} L_{gv} 2\pi r_v}{EOT}+\frac{\varepsilon_s\left(4\left(t_{gv}+r_v\right)^2-\pi r_V^2\right)}{L_{sv}}+\frac{C_{wf}}{\sqrt{A_c}}\right]^2}\frac{\pi r_v}{\left(h_f+w_f\right)}\frac{L_{gf}}{L_{gv}}$$

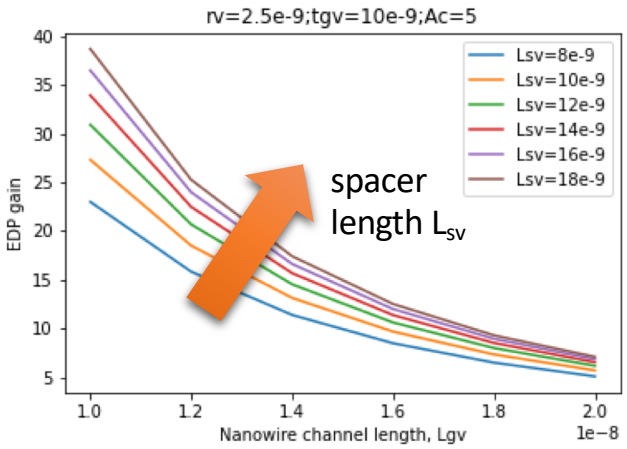
# EDP analysis and physical DSE

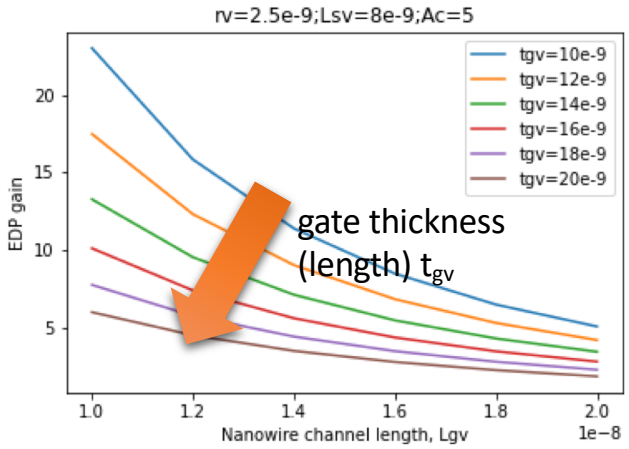



**FinFET**



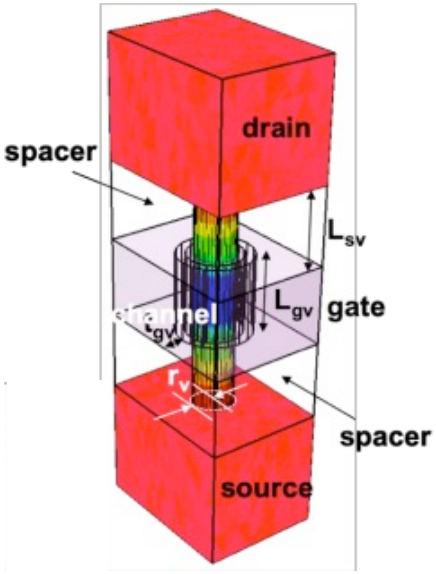




I. O'Connor et al., "Analysis of Energy-Delay Product of a 3D Vertical Nanowire FET Technology," EuroSOI-ULIS, 2021

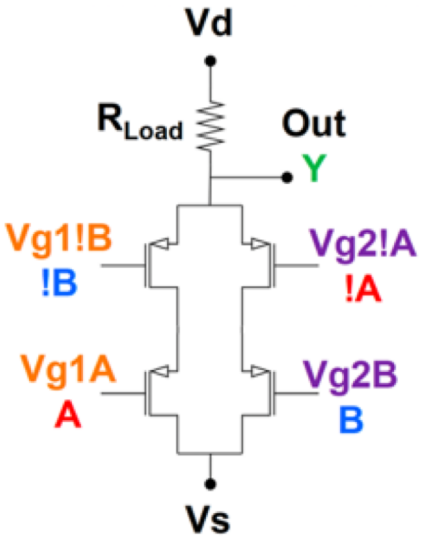
Improved performance gain for **smaller** and **fewer** nanowires per VNWFET (pitch overhead, $I_{on}$ varies sublinearly)

**VNWFET**



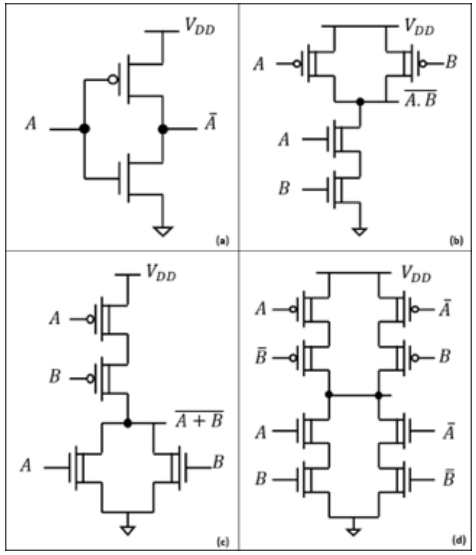
# JL1/JL2 – hardware and projections

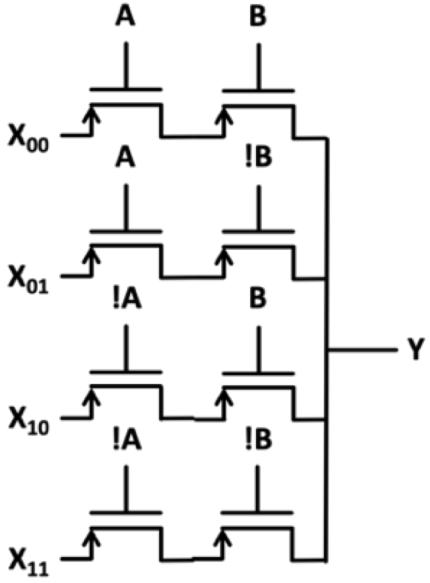
**DMP nomenclature :**
[LogicFunction]
_[Datawidth]
_[DesignStyle]
_[TechnologyVariant]




**PStatic cell hardware**
**XOR2_1_PStatic_JL2**

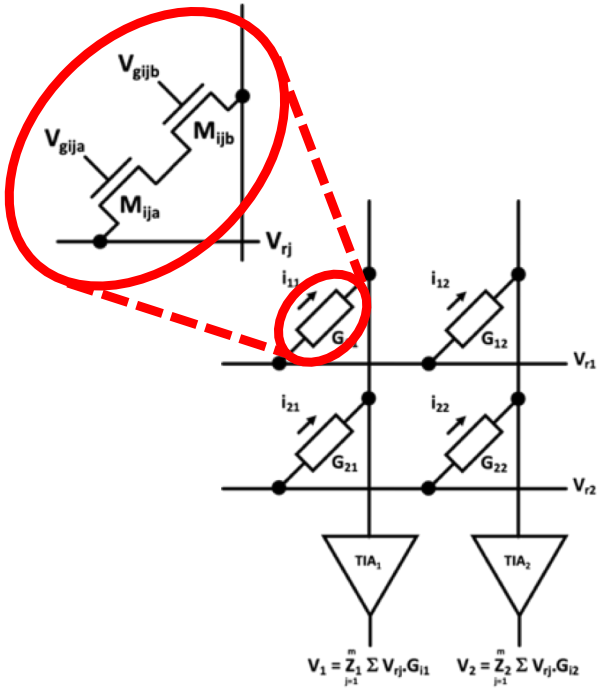
**CStatic cell library**
**Simulations**
**INV1_1_CStatic_JL1**
**NAND2_1_CStatic_JL1**
**NOR2_1_CStatic_JL1**
**XOR2_1_CStatic_JL1**





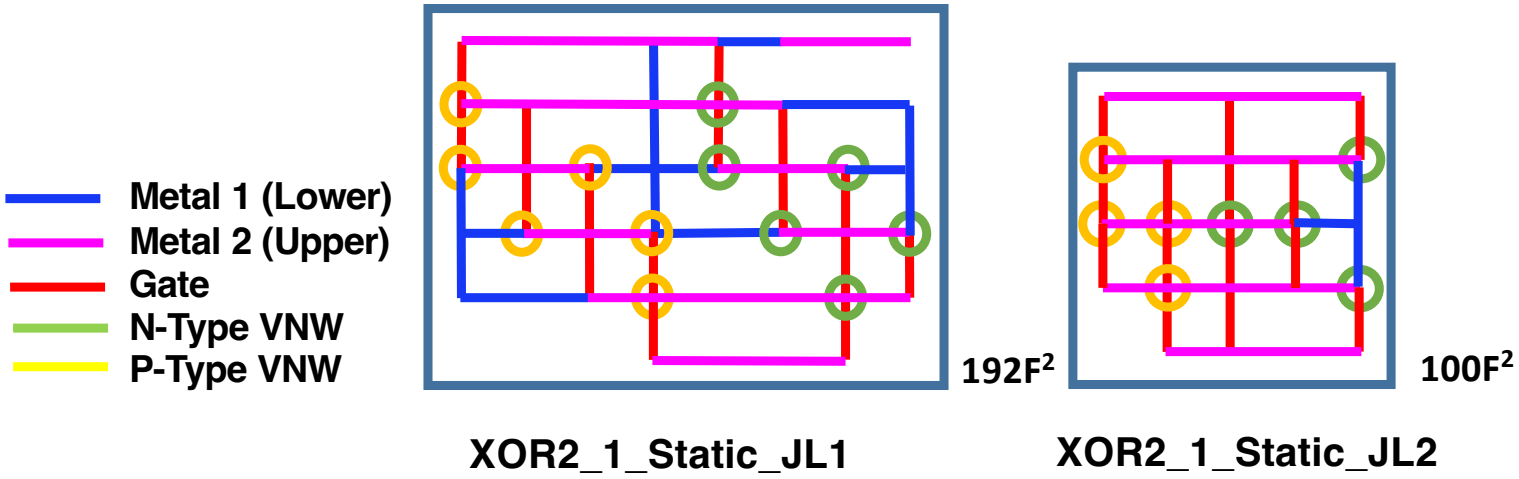

**PPTL hardware**
**MUX2_1_PTL_JL2**

**Xbar hardware**
**2X2_1_XBAR_JL2**



# JL1/JL2 – STX and area estimation

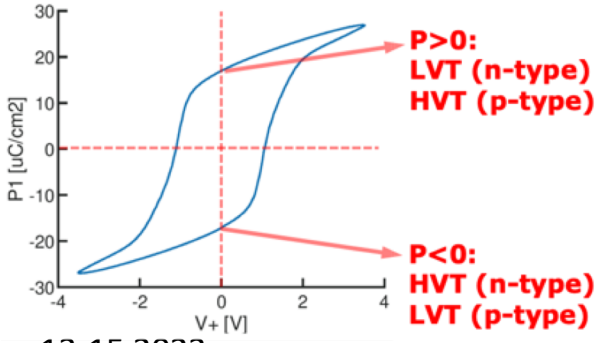
- Cell area is expressed in units of $F^2$
- F is the minimum lithographic feature of the densest process layer
- F = Half pitch dimension of M1 layer



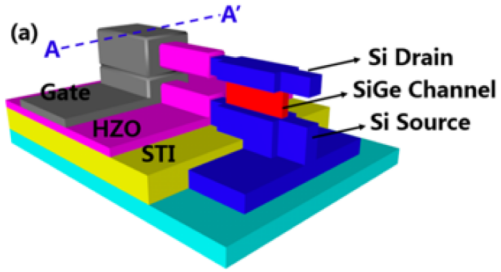

| Function | JL1 $F^2$ | JL2 $F^2$ | $F^2$ red | JL1 #NW | JL2 #NW | #NW red |
|---|---|---|---|---|---|---|
| INV1_1_Static | 8 | 8 | 0% | 2 | 2 | 0% |
| NOR2_1_Static | 60 | 36 | 40% | 4 | 3 | 33% |
| NAND2_1_Static | 60 | 36 | 40% | 4 | 3 | 33% |
| XOR2_1_Static | 192 | 100 | 48% | 12 | 8 | 50% |

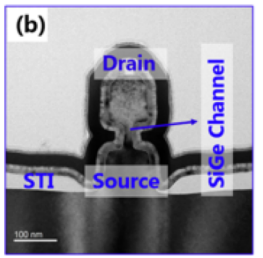
# Why ferroelectric?

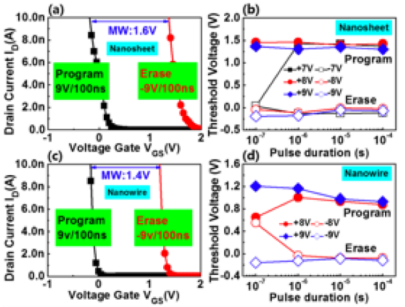
- Non-volatile behavior
- Best-in-class energy-efficiency
- $HfO_2$ based:
  - CMOS-compatible
  - VNWFET-compatible
- NV memory, NV logic
- In-memory computing




| | FLASH | MRAM | PCM | RRAM | FeRAM | FeFET |
|---|---|---|---|---|---|---|
| Programming power | ~200pJ/bit | ~20pJ/bit | ~300pJ/bit | ~100pJ/bit | ~10fJ/bit | ~10fJ/bit |
| Write speed | 20 µs | 20 ns | 10-100 ns | 10-100 ns | 14ns @ 2.5V | 50 ns |
| Endurance | $10^5$ - $10^6$ | $10^6$-$10^{15}$ | $10^8$ | $10^5$ - $10^6$ | > $10^{11}$ | $10^5$ - $10^6$ |
| Retention | > 125°C | 85°C - 215 °C | 165°C | > 125°C | 85°C | 125°C? |
| Extra masks | Very high (>10) | Limited (3-5) | Limited (3-5) | Low (2) | Low (2) | Low (2) |
| Process flow | Complex | Medium | Medium | Simple | Simple | Simple |
| Scalability | Bad | Medium | High | High | Poor (2D) High (3D) | Good |



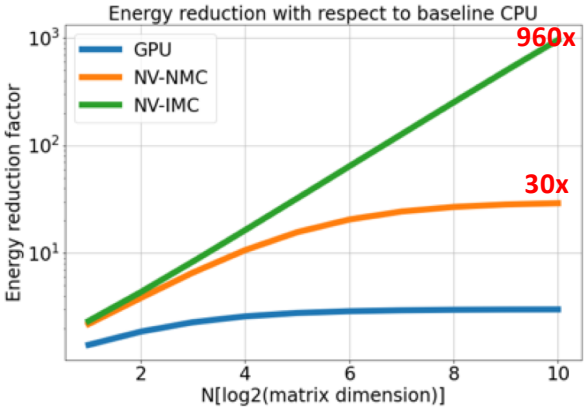



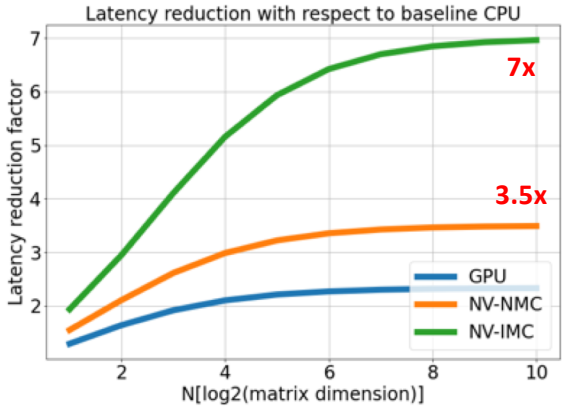




W. Huang et al., "Ferroelectric Vertical Gate-All-Around Field-Effect-Transistors With High Speed, High Density, and Large Memory Window," in IEEE Electron Device Letters, vol. 43, no. 1, pp. 25-28, Jan. 2022, doi: 10.1109/LED.2021.3126771

# NMC/IMC for matrix multiplication

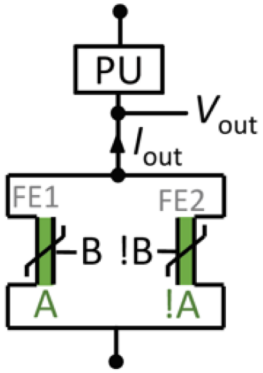
- Operation costs assessed for $2^{10}$ multiplications of $2^N x 2^N$ matrices with 32-bit data representation
- Transfer of matrix coefficients to non-volatile memories at varying degrees of proximity to CPU
- Assumptions:
  - 1pJ/bit/mm communication costs (inst., operands)
  - 10aJ/bit computation costs
  - 10fJ/bit programming costs
- Gain increases with data size for NV-NMC and even more for NV-IMC computing (almost 3 orders of magnitude between N=1 and N=10)

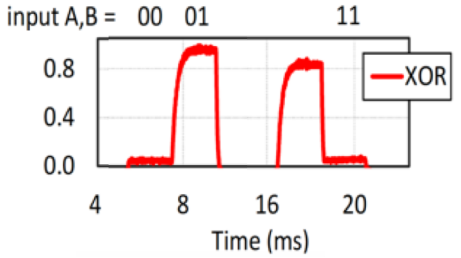


# NV-XOR

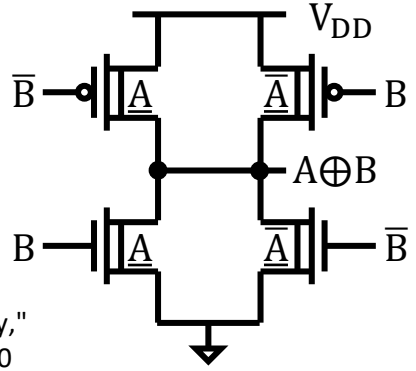
- **NV-programming of coefficients in data-intensive applications e.g. convolutional filters, neural networks**
- **XOR is a key operation for adders, multipliers, MAC …**
- **For XOR3/JLFE2, 8 devices instead of 20!**



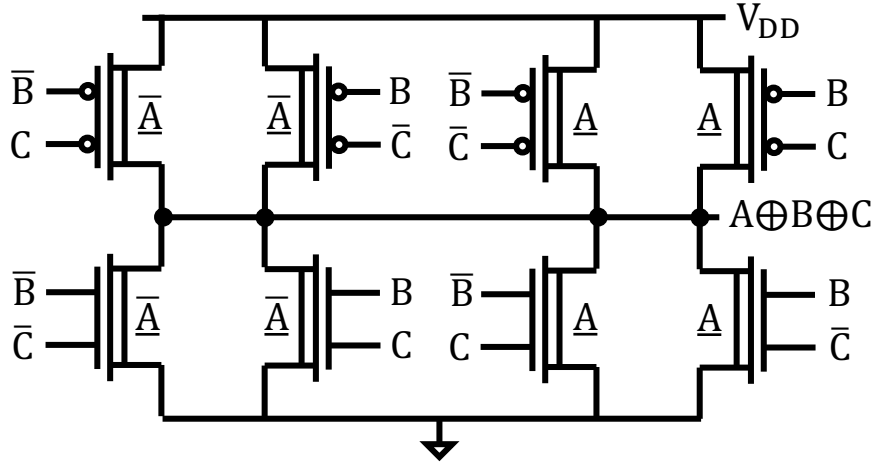




E.T. Breyer et al., "Compact FeFET circuit building blocks for fast and efficient nonvolatile logic-in-memory," IEEE J. Electron Devices Society, 2020

**XOR2_1_NMOS_JLFE1**



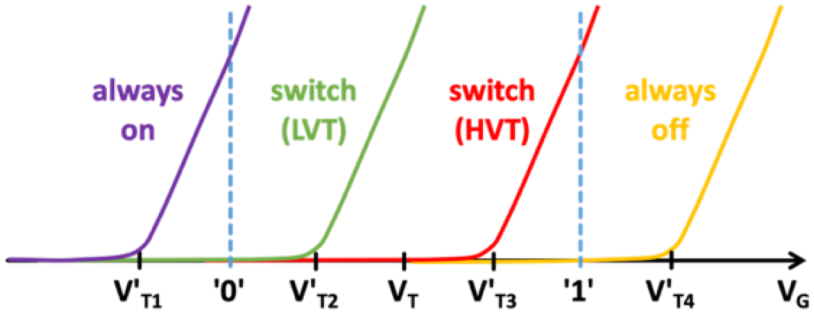

**XOR2_1_Static_JLFE1**

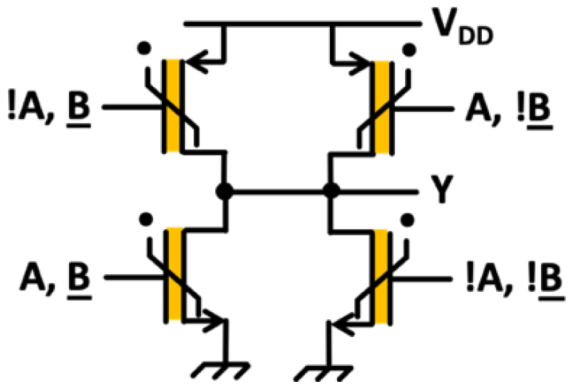



**XOR3_1_Static_JLFE2**

# CNVL JLFE1/JLFE2 cell library

**Principle: non-volatile programming of $V_t$**

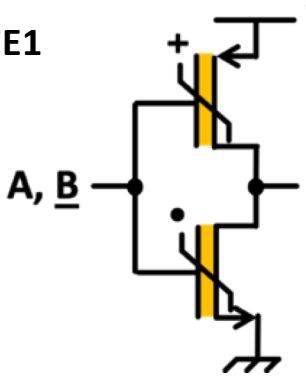



**XOR2_1_CNVL_JLFE1**

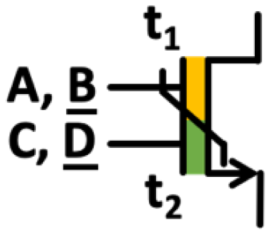
**Illustration: NAND2_1_CVNL_JLFE1**

| line | A | B | Y | |
|---|---|---|---|---|
| 1 | 0 | 0 | 1 | |
| 2 | 0 | 1 | 1 | PUN |
| 3 | 1 | 0 | 1 | |
| 4 | 1 | 1 | 0 | PDN |



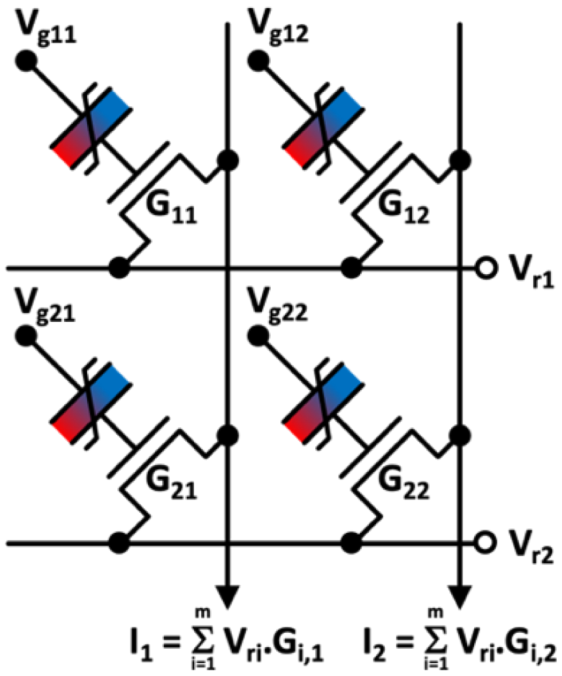

B=0 always on (L1,L3)
B=1 switch (L2)

B=0 always off
B=1 switch (L4)

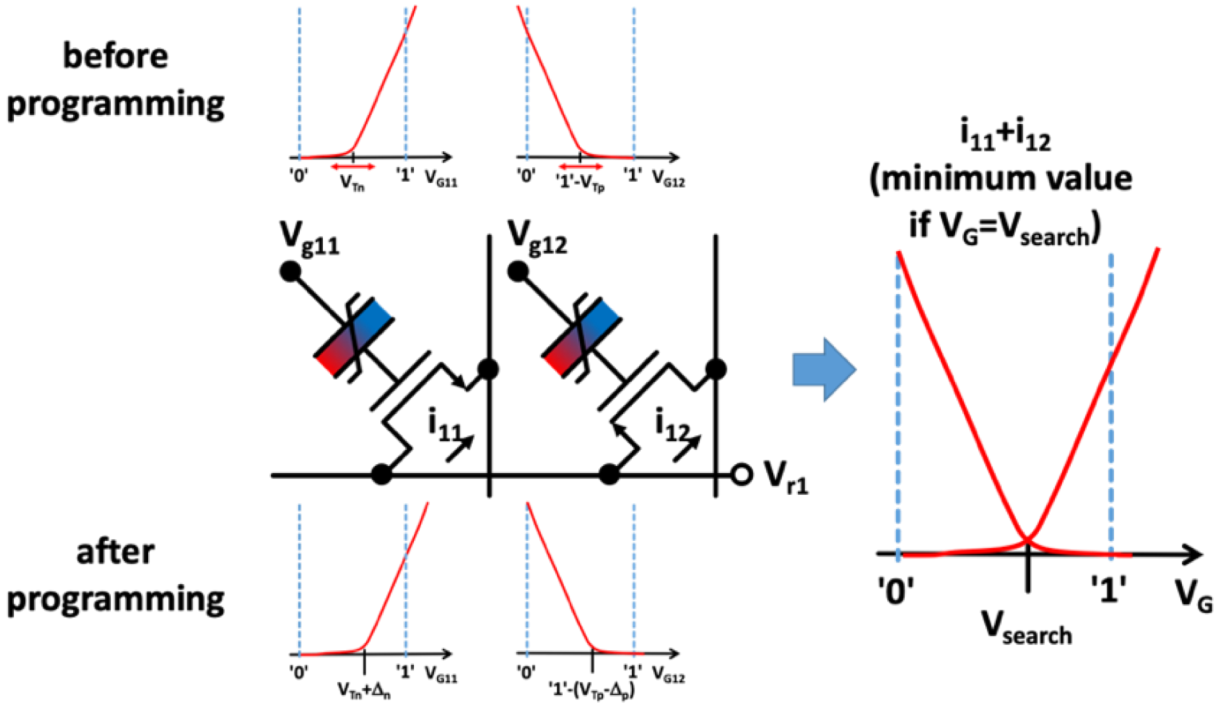



**CNVL_JLFE2 building block**

# JLFE1 NVXbar



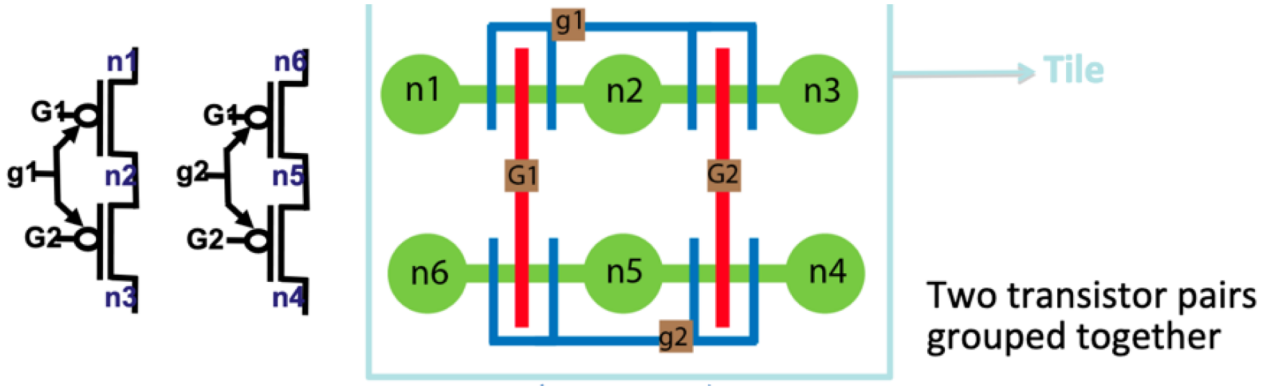

**2x2 analog matrix multiplication**
**2X2_1_NVXBAR_JLFE1**




**Analog TCAM tile**
**(analog distance computation)**
**TCAM_1_NVXBAR_JLFE1**

# Why ambipolar?

- **Reconfigurable transistors via electrostatic doping using $g_x$ (polarity gate)**
  - **$g_x$=1: n-type**
  - **$g_x$=0: p-type**
- **Ultimate logic flexibility**
- **Tuning hardware to application needs**

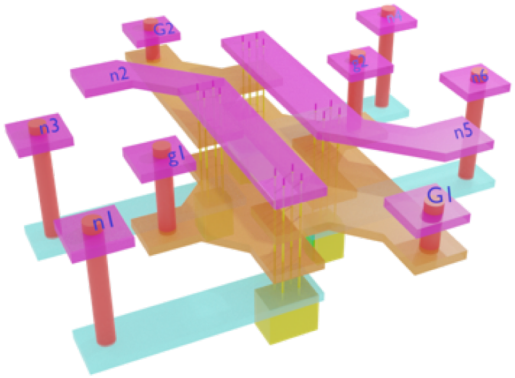


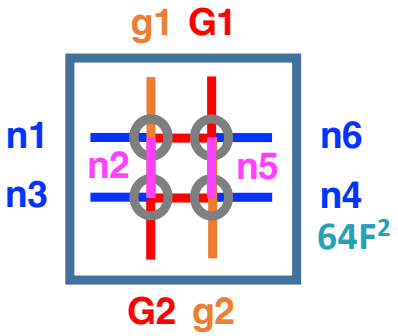

S. Bobba et al., "Physical synthesis onto a Sea-of-Tiles with double-gate silicon nanowire transistors", DAC 2012 (DOI: 10.1145/2228360.2228369)

| Logic | n1 | n2 | n3 | n4 | n5 | n6 | G1 | G2 | g1 | g2 |
|---|---|---|---|---|---|---|---|---|---|---|
| XOR2 | Gnd | Out | Vdd | Gnd | Out | Vdd | A | A' | B' | B |
| XNOR2 | Gnd | Out | Vdd | Gnd | Out | Vdd | A | A' | B | B' |
| NAND2 | Out | Vdd | Out | Out | - | Gnd | A | B | Gnd | Vdd |
| NOR2 | Vdd | - | Out | Out | Gnd | Out | A | B | Gnd | Vdd |
| INV | Vdd | Out | Vdd | Gnd | Out | Gnd | A | A | Gnd | Vdd |
| BUF | O1 | Vdd | Out | Out | Gnd | O1 | A | O1 | Gnd | Vdd |

# Vertical implementation of ambipolar tile

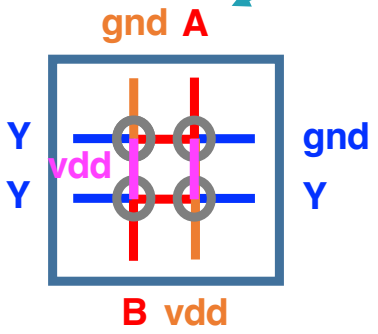
- **2AP implementation: (ferroelectric) polarity gate stacked with control gate**
- **Regular tile facilitates fabrication**
- **Routing strategy critical: separation between data plane and power plane**



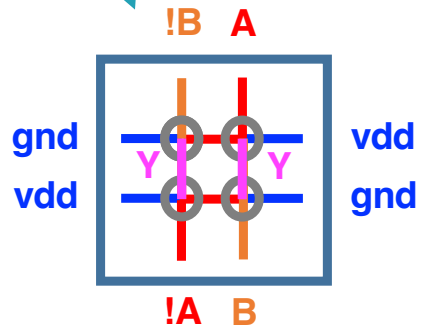



Tile2_1_Ambipolar_2AP

metal 1 (lower)
metal 2 (upper)
gate 1
gate 2
n-Type VNW
p-Type VNW
ambipolar VNW

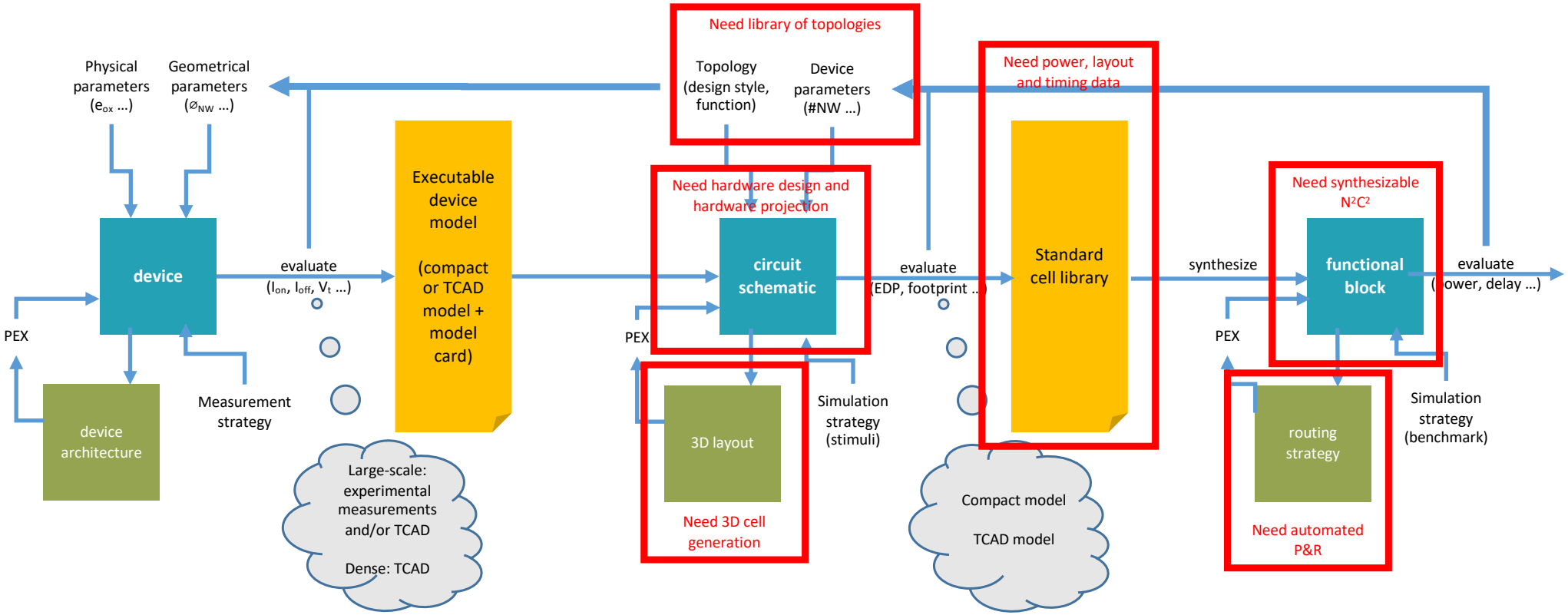



NAND2



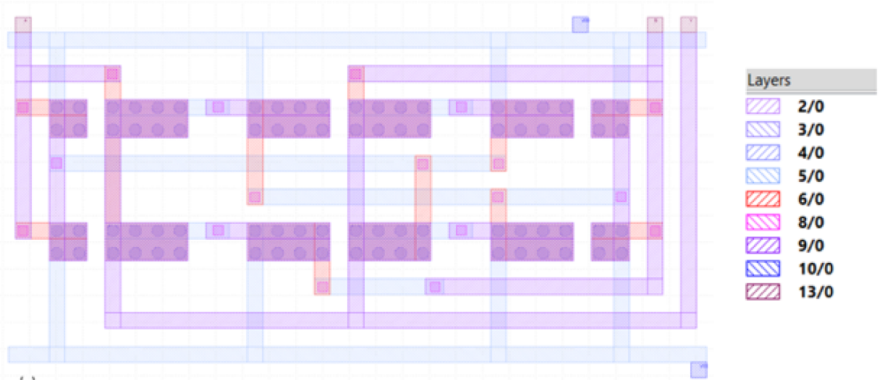

XOR2
(164F² in JL1, 100F² in JL2)
2.5X footprint reduction

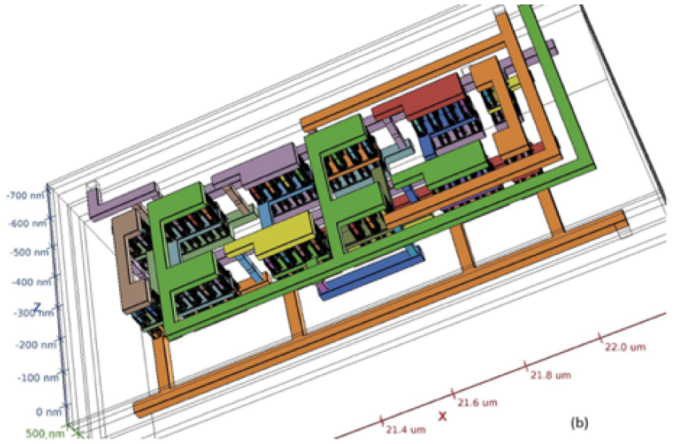
# Design-Technology Co-Optimization



# Physical design – tool integration

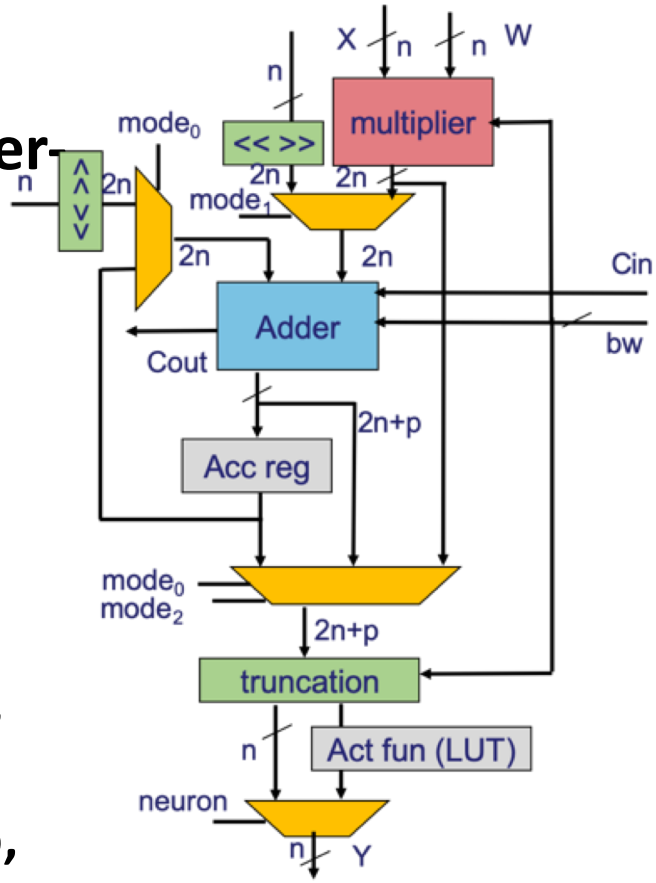
- Automated GDSII generation
- Import into GTS TCAD toolset
- PEX (parasitic extraction)
- and/or TCAD simulation
- Spectre simulation with compact model and PEX results
- Energy/delay metric quantification



# $N^2C^2$ – neural network compute cube

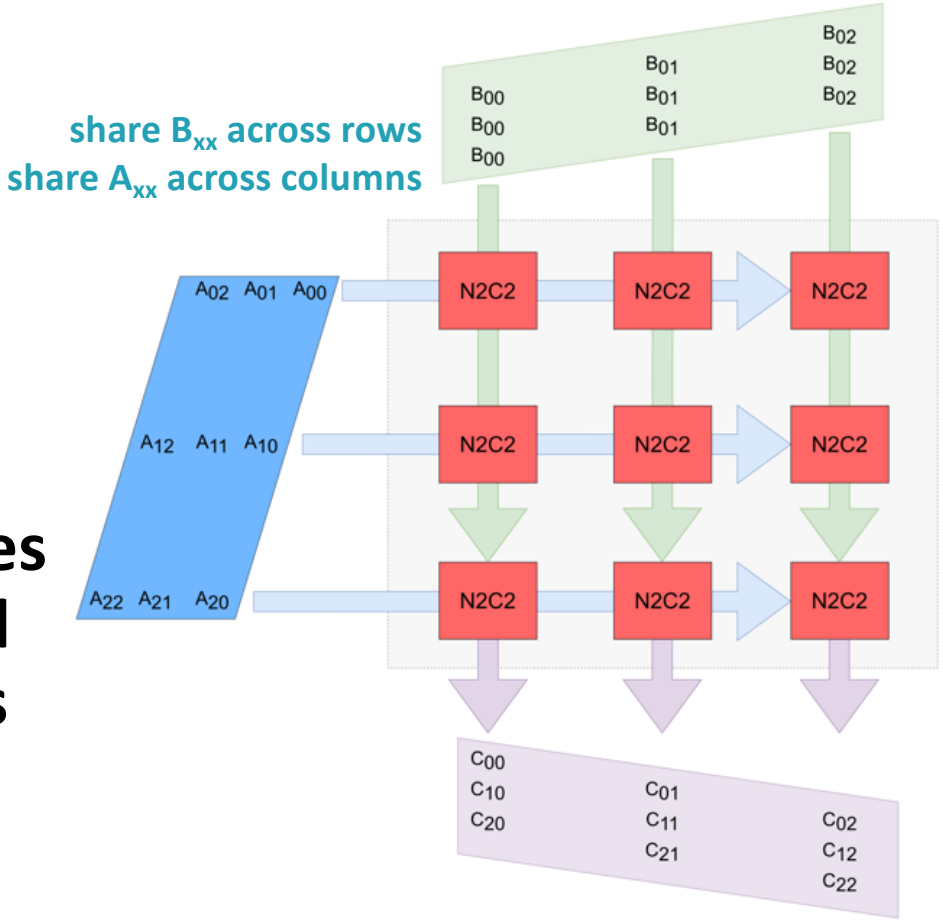
- **$N^2C^2$ = flexible computing hardware block for transformer-based neural networks**
  - **Function-configurable: 32-bit integer MAC / * / + w/wo activation**
  - **Connectivity-configurable: 2-8 operands**
  - **Datawidth-configurable**
  - **Intra-$N^2C^2$ scaledown to 2*16b, 4*8b, 8*4b …**
  - **Inter-$N^2C^2$ scaleup to 64b, 128b, 256b …**




| Output | Control Signals | Description |
|---|---|---|
| Y <= X * W<br>Acc <= Y + A | $mode_0$ = 1<br>$mode_1$ = 1<br>$mode_2$ = 0<br>neuron = 0 | multiplication |
| Y <= X * W + A<br>Acc <= Y | $mode_0$ = 0<br>$mode_1$ = 1<br>$mode_2$ = 0<br>neuron = 0 | multiplication addition |
| Y <= X * W + Acc<br>Acc <= Acc + Y | $mode_0$ = 1<br>$mode_1$ = 1<br>$mode_2$ = 1<br>neuron = 0 | multiplication accumulation (MAC) |
| Y <= A + B<br>Acc <= Y | $mode_0$ = 0<br>$mode_1$ = 0<br>$mode_2$ = 0<br>neuron = 0 | addition |
| Y <= B + Acc<br>Acc <= Acc + Y | $mode_0$ = 1<br>$mode_1$ = 0<br>$mode_2$ = 1<br>neuron = 0 | accumulation |
| Y <= Act(R) | neuron = 1 | neuron mode with activation function. R can be the result of any of the above arithmetic functions. |

# Systolic array approach

- homogeneous network of tightly coupled hard-wired data processing units (i.e. $N^2C^2$)
- adapted to dense linear algebra computation
- each $N^2C^2$ independently computes a partial result from data received from upstream neighbours, stores the result within itself and passes it downstream



# Conclusion

- **Vertical transistors: 10x EDP reduction (wrt 7nm FinFET) for 14nm VNWFET gate length**
- **Ferroelectric IMC: 10x energy + 10x latency reductions (wrt CPU-based approach) for $2^{10}$ multiplications of $2^5$x$2^5$ matrices with 32-bit data**
- **Ambipolar: access compact, flexible and disruptive logic cells**
- **$N^2C^2$ approach: 10x inference time reduction (wrt conventional approach) – cf Alberto's presentation**


**improving energy-efficiency by 10k!**

# Mandatory closing slide

- PhD and postdoc positions
  - Energy-efficient circuit and architecture design and test
  - AI/ML
  - Hardware security
  - Emerging technologies
    - 3D, ferroelectric, silicon photonics
  - Emerging paradigms
    - AxC
    - Stochastic
  - Design-Technology co-optimization


alberto.bosio
cedric.marchand
ian.oconnor @ec-lyon.fr
damien.deleruyelle @insa-lyon.fr
bastien.deveautour @cpe.fr