The Era of Hyper Scaling in Electronics

Suman Datta

Contributors: Salahuddin, Bokor, Ramesh (Berkeley), Schlom, Jena (Cornell), Kummel (UCSD), George (Colorado), Raychowdhury, Swaminathan, Khan, Naemi, Bakir, Yu (Georgia Tech), Wong, Pop, Goodson, Wang (Stanford), Ye, Datta (Purdue), JP Wang (Minnesota), Mishra (UCSB), Iyer (UCLA), Cho (UTDallas), Winter (Wayne State), Hock (Illinois Tech), Fay, Niemier (Notre Dame)

Former Colleagues at Intel: Doyle, Kavalieros, Doczy, Metz, Dewey, Chau

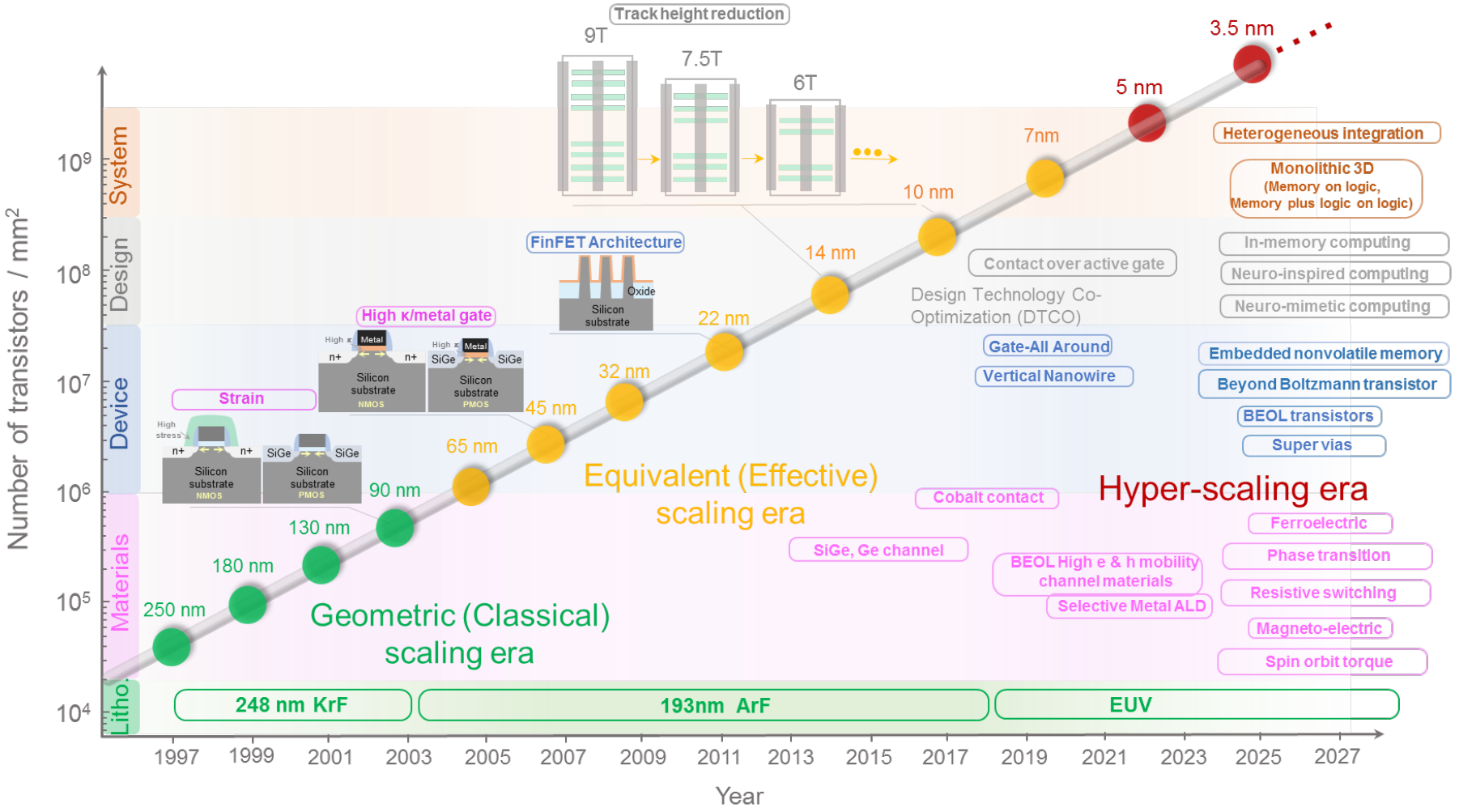
UNIVERSITY OF NOTRE DAME

Electronics

UNIVERSITY OF NOTRE DAME

- Past
- Present
- Future

Three Eras of Scaling



Three eras of scaling: geometric (past), equivalent (present) and hyper-scaling (future)

S. Salahuddin, K. Ni, and S. Datta, Nature Electronics, Aug 2018

Geometric (Classical) Scaling

UNIVERSITY OF NOTRE DAME

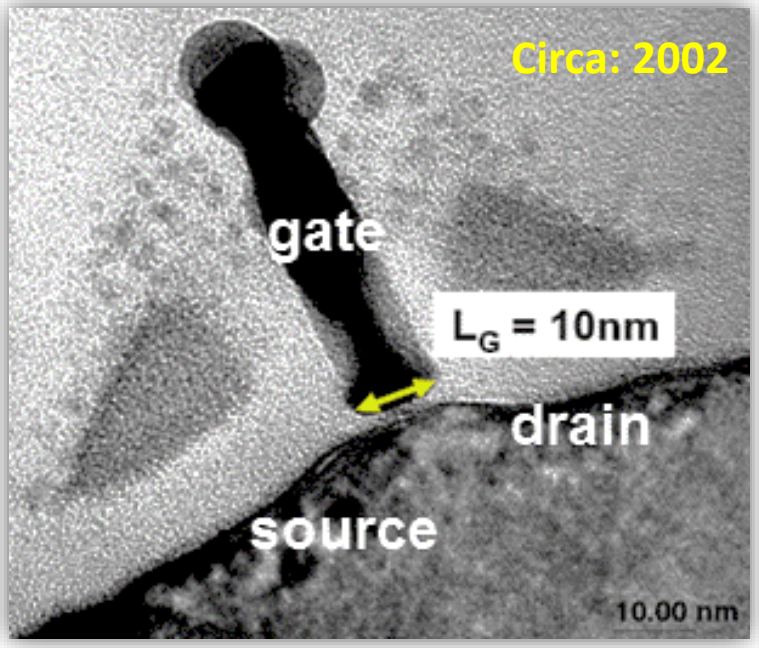
Table 1
Scaling Results for Circuit Performance

Device or Circuit Parameter	Scaling Factor
Device dimension t_{ox}, L, W	$1/\kappa$
Doping concentration N_a	κ
Voltage V	$1/\kappa$
Current I	$1/\kappa$
Capacitance $\epsilon A/t$	$1/\kappa$
Delay time/circuit VC/I	$1/\kappa$
Power dissipation/circuit VI	$1/\kappa^2$
Power density VI/A	1

R. Dennard, Proceedings of the IEEE, April 1999

10nm Transistor (circa 2002)

UNIVERSITY OF NOTRE DAME



B. Doyle, S. Datta, J. Kavalieros, R. Chau, *T*ransistor Elements for 30nm Physical Gate Lengths and Beyond" Intel Technology Journal, May 2002

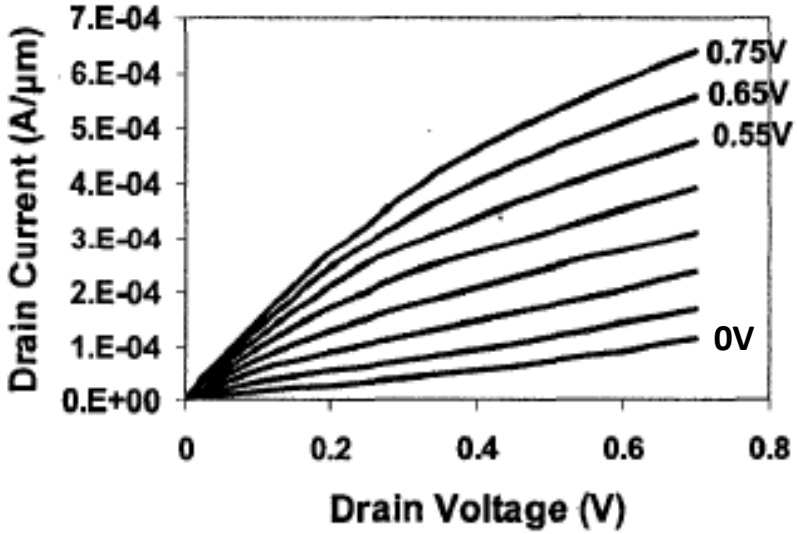


Figure 8: Family of Id-Vd curves for the 10nm research device.

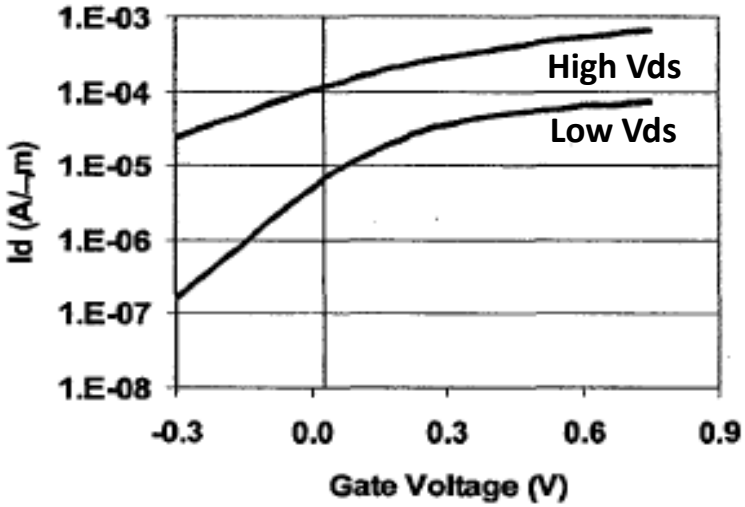
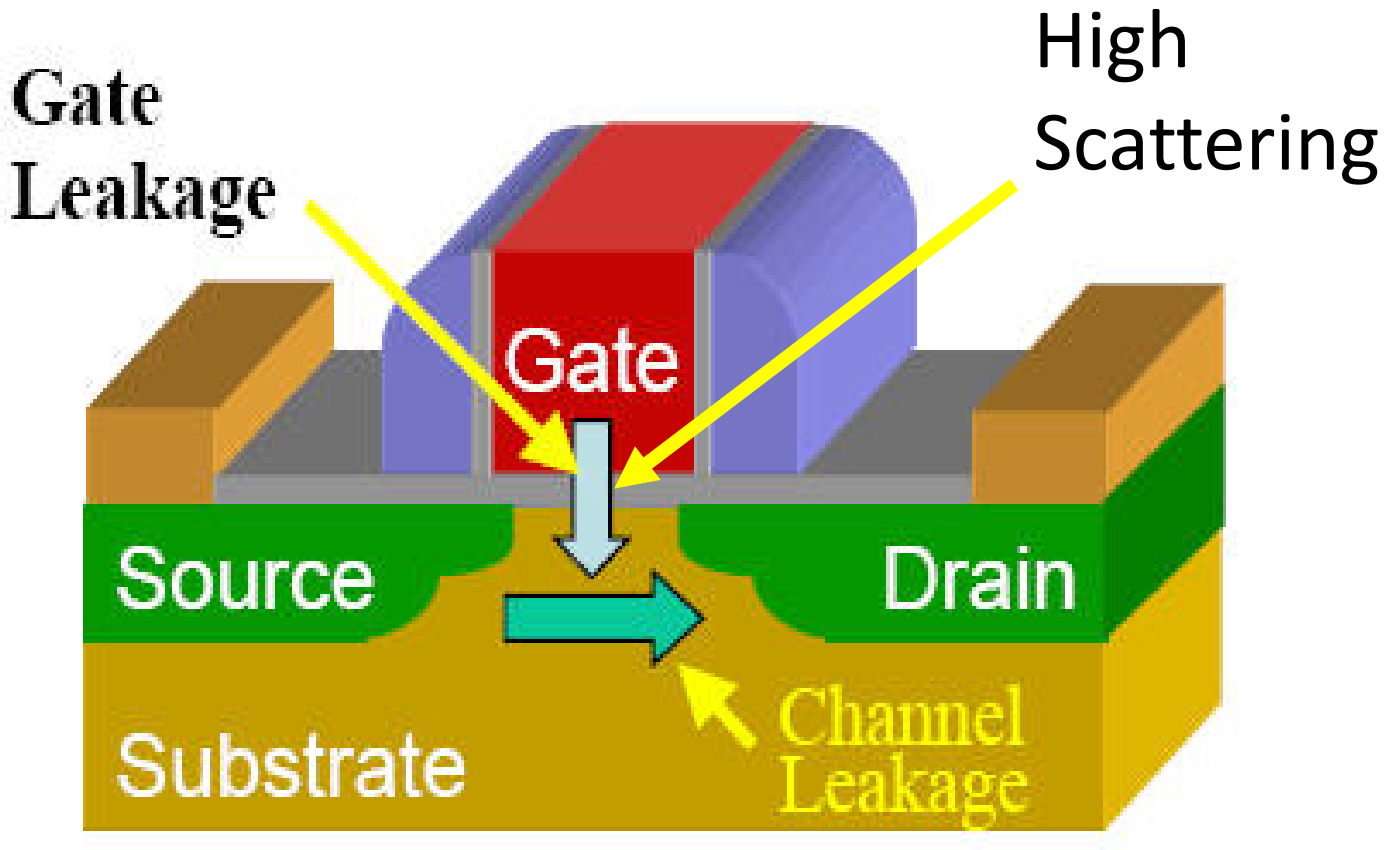


Figure 7: Id-Vg characteristics at Vd=50mV and 0.75V for the 10nm experimental device.

Classical Scaling Ends

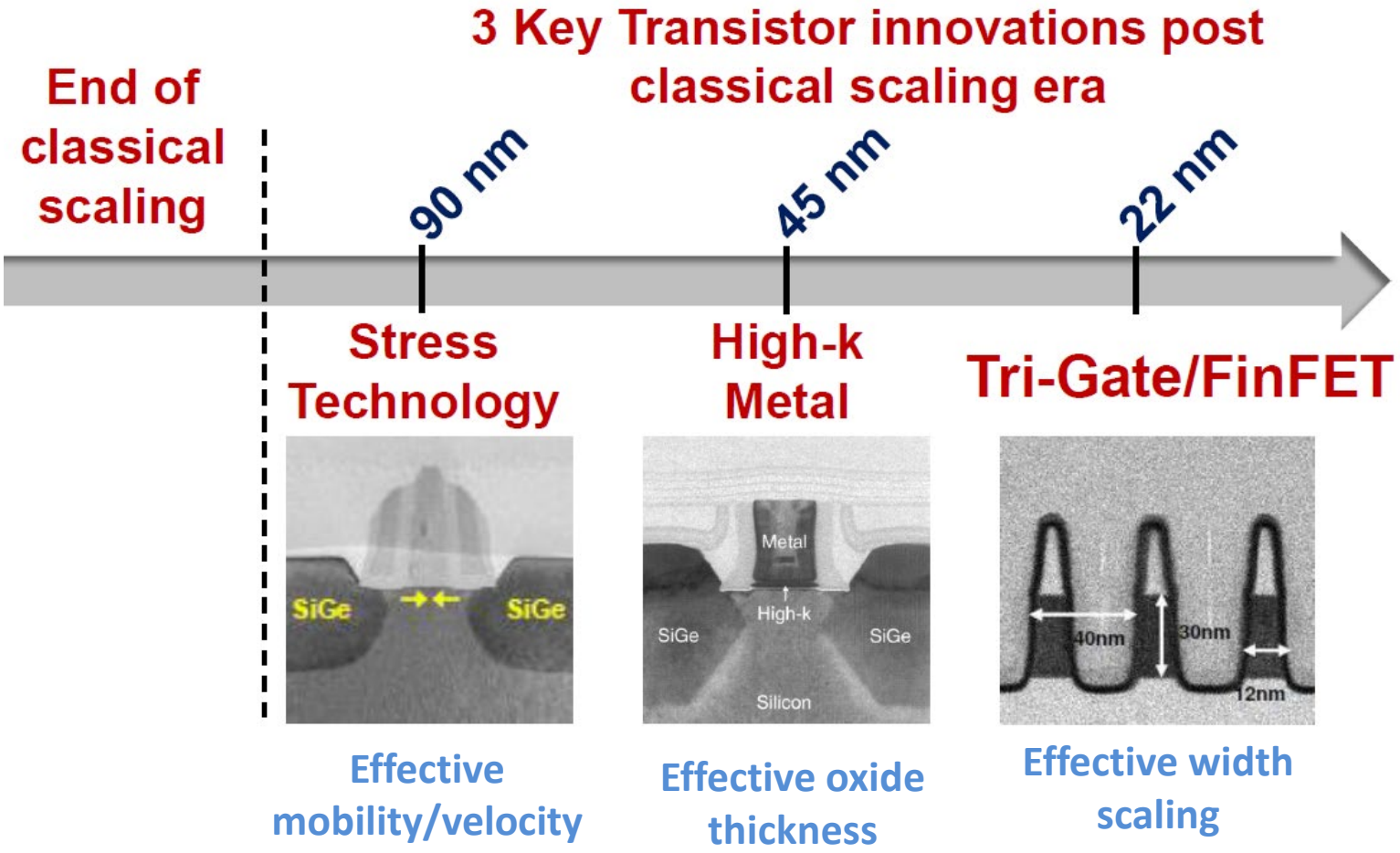


Classical scaling will not work.

Need new solutions involving new **materials**, new **device architecture** and new **switching mechanism**

Equivalent (Effective) Scaling

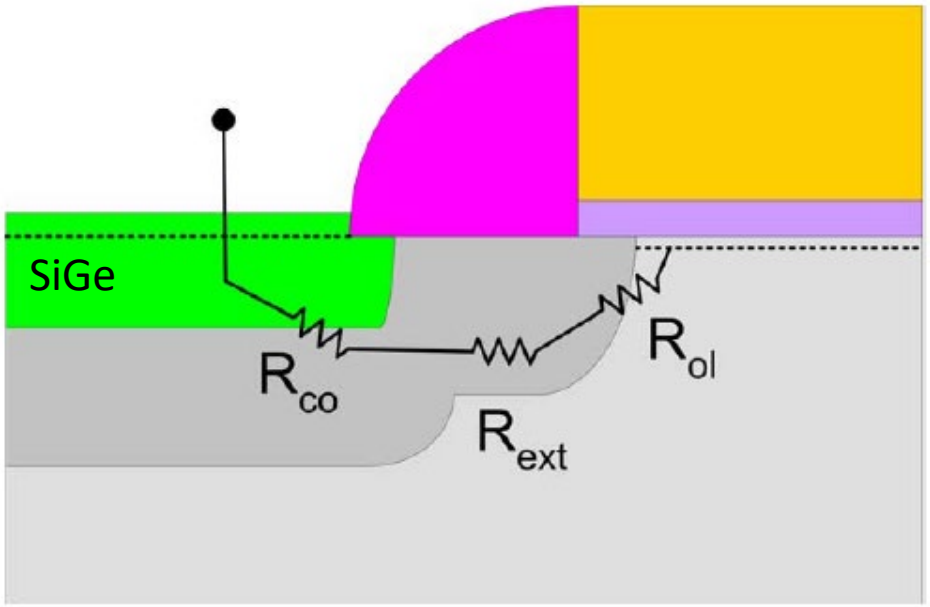
UNIVERSITY OF NOTRE DAME

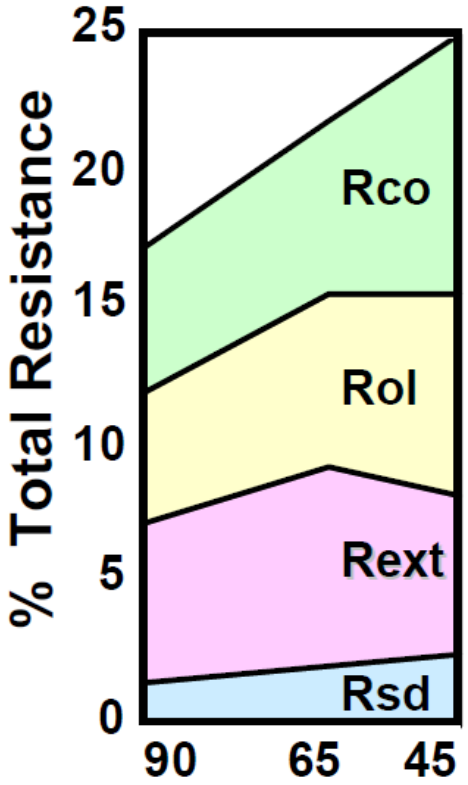


3 key transistor innovations that shaped this era of effective scaling

Lessons in serendipity embedded in each

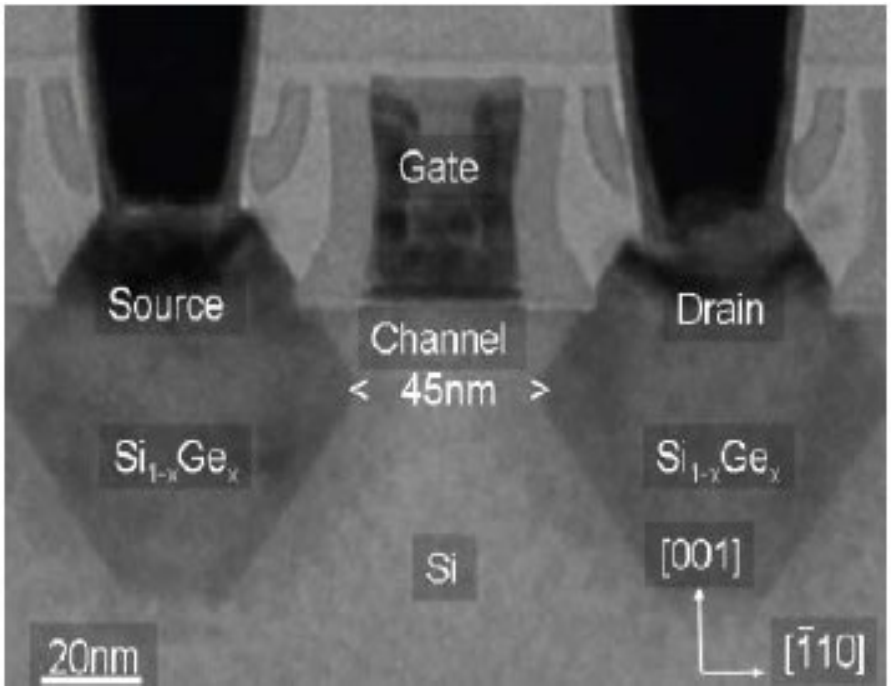
Elevated Source Drain

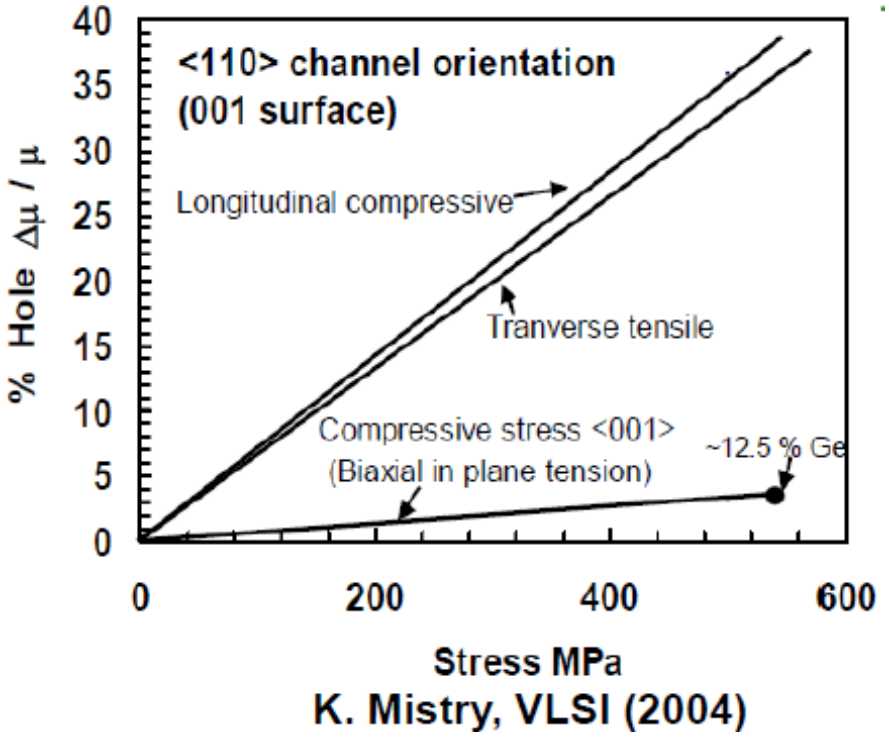




Series resistance becomes bigger fraction of on-resistance

Partially Embedded Source Drain

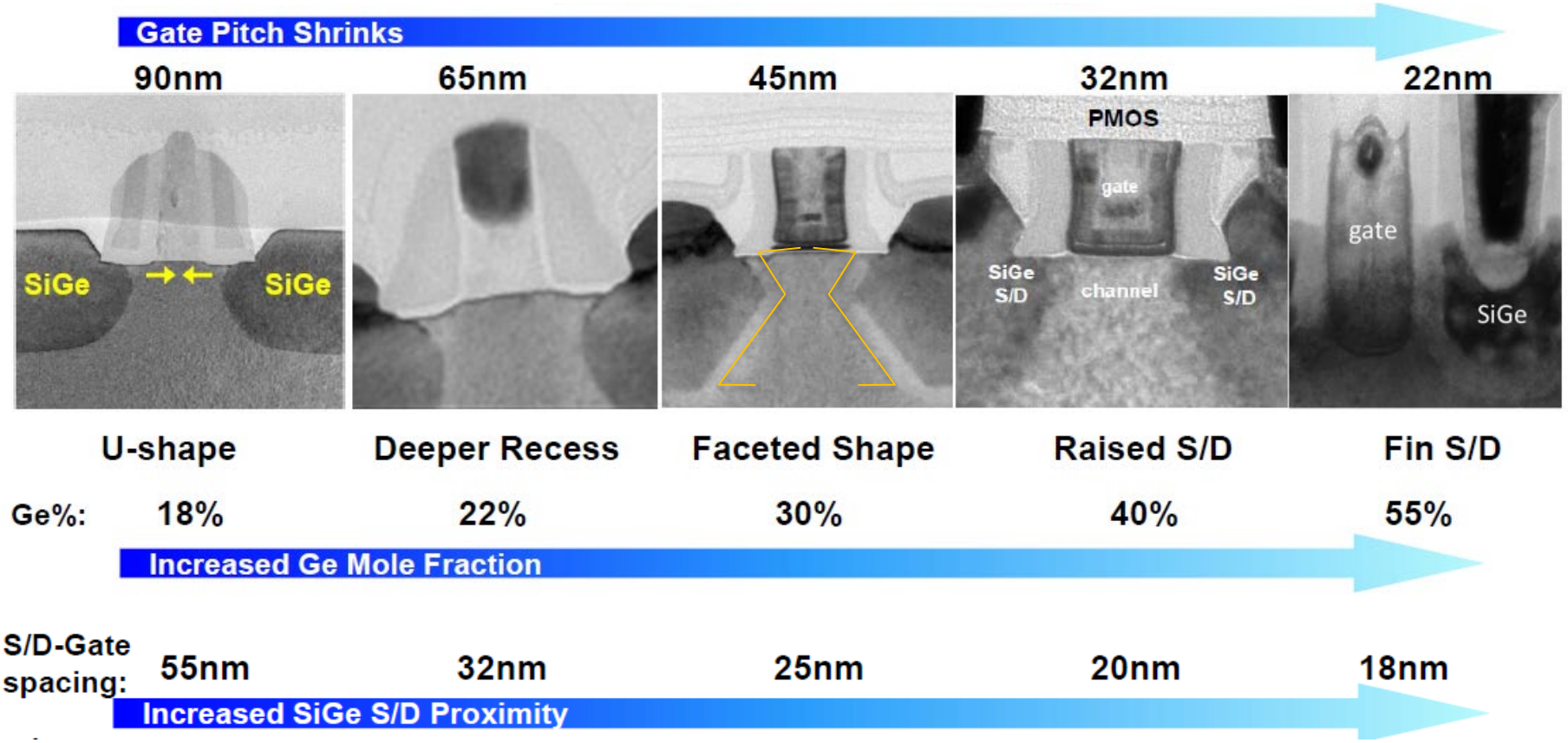




K. Mistry, VLSI (2004)

Hole mobility responds in a remarkable fashion
Local strain in PMOS and does not affect NMOS

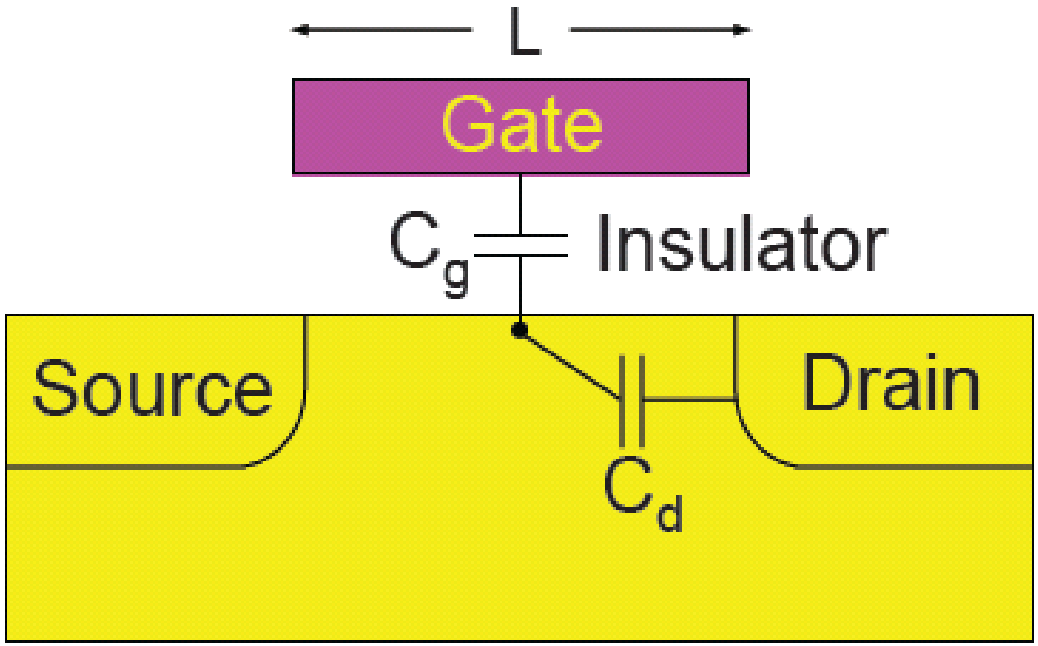
Local Stressor – an act of serendipity



SiGe S/D was intended for lower external resistance

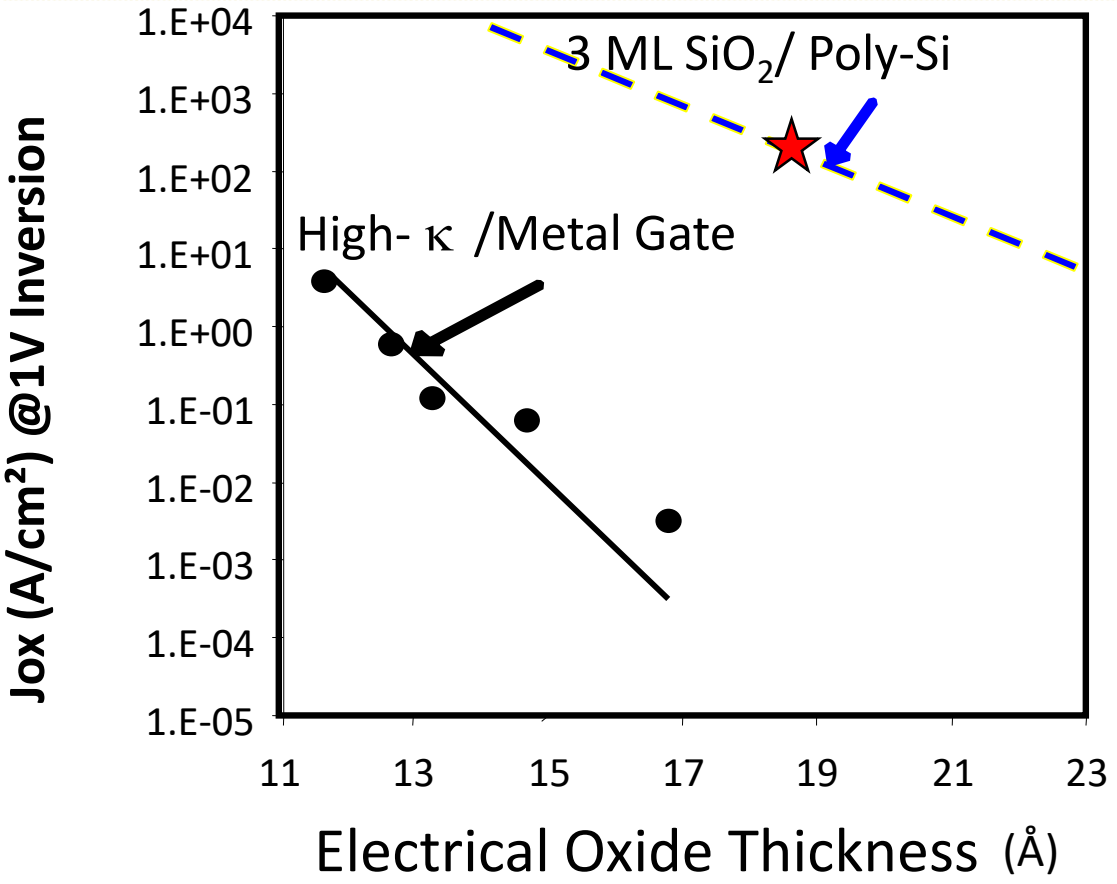
eSiGe S/D ends up as effective local stressor for the channel

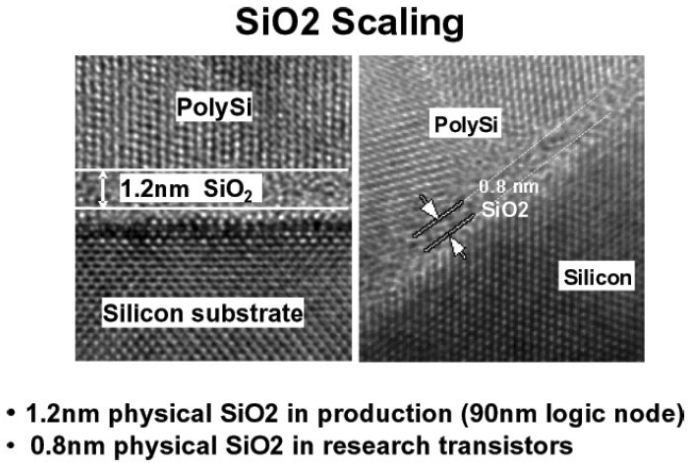
Gate Insulator



"Transistors behave like resistors at small L unless the insulator thickness is reduced with L"

High-k metal-gate



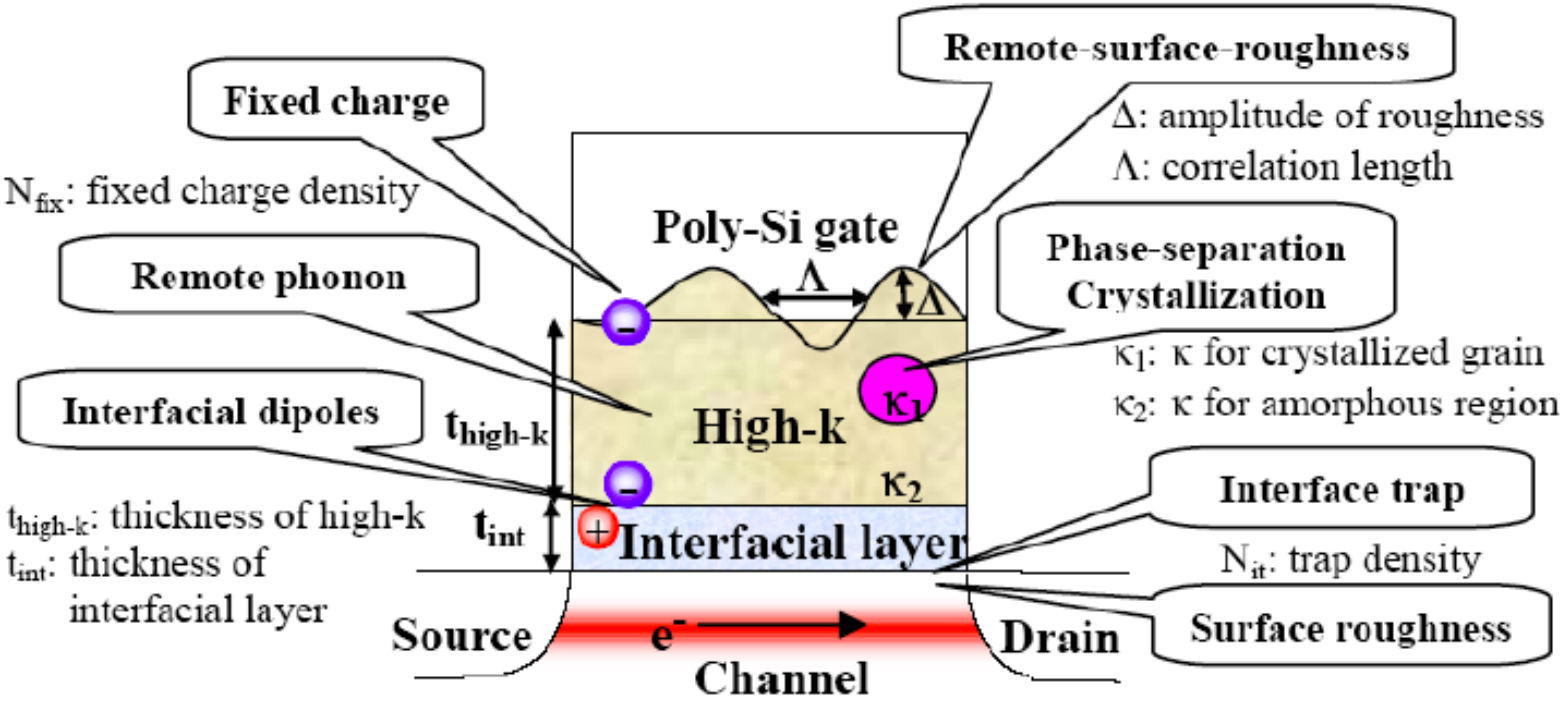


- 1.2nm physical SiO2 in production (90nm logic node)
- 0.8nm physical SiO2 in research transistors

Gate leakage increases with SiO_2 scaling -> running out of atoms

High-k (HfO_2) Gate Stack enables T_{elec} scaling with low gate leakage

Gate Stack Challenges



S. Saito, et al., IEEE IEDM, Washington, DC, Dec., 2003.

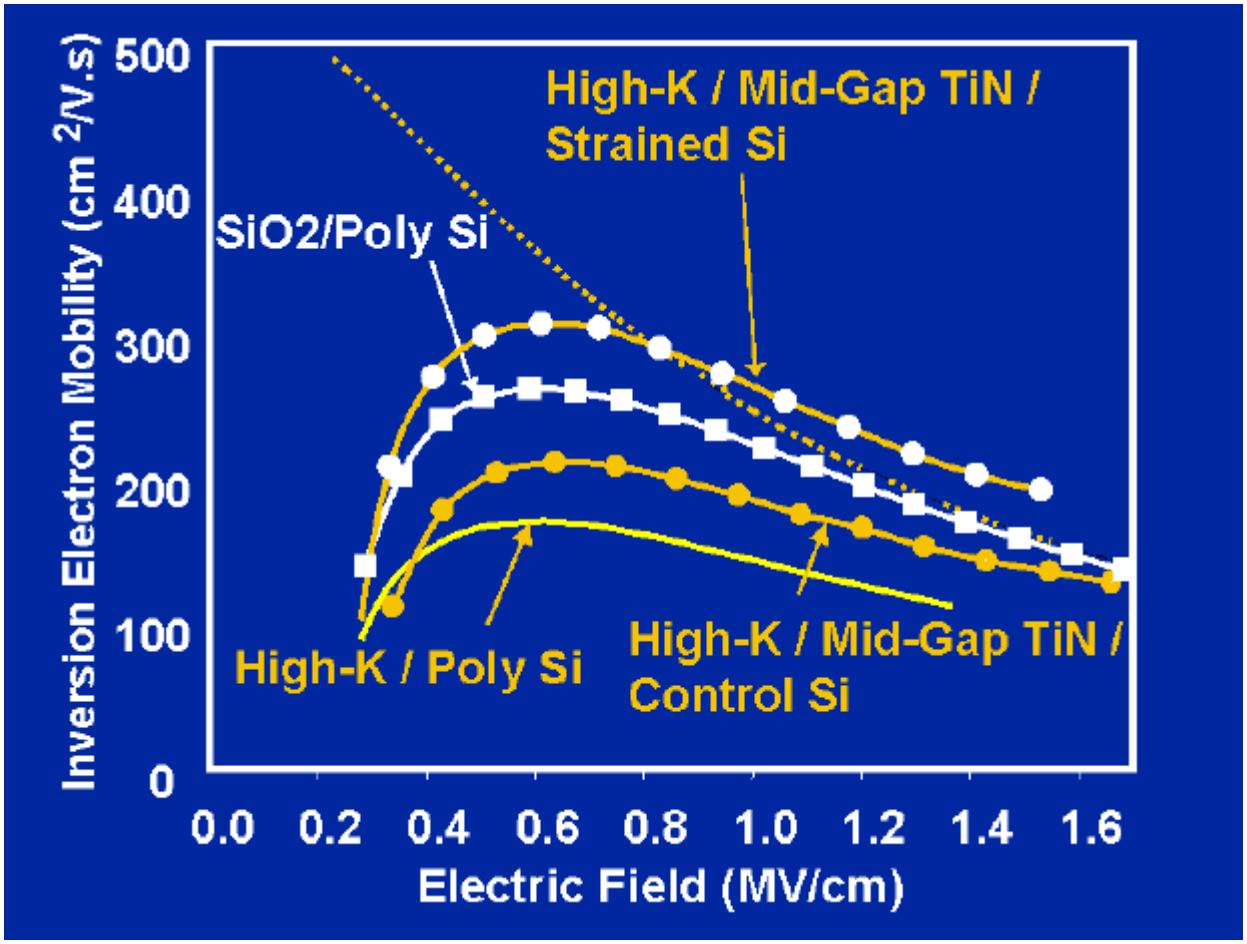
Bulk & interface traps: Poor reliability

New scattering modes: Poor mobility

Technology Integration: Complexity and cost

Strained Si + High-k / MG stack

UNIVERSITY OF NOTRE DAME

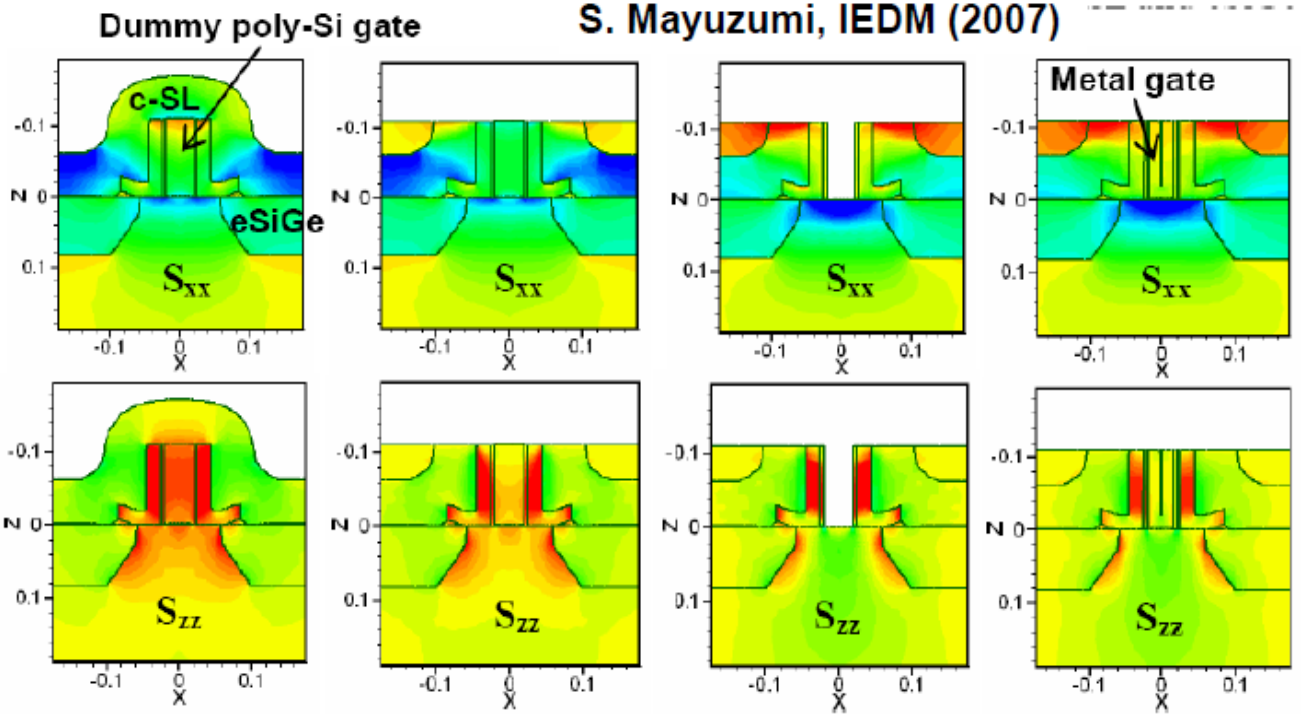


Need strain + high-k + metal gate to recover mobility

S. Datta, J. Kavalieros, R. Chau, *IEDM 2003*

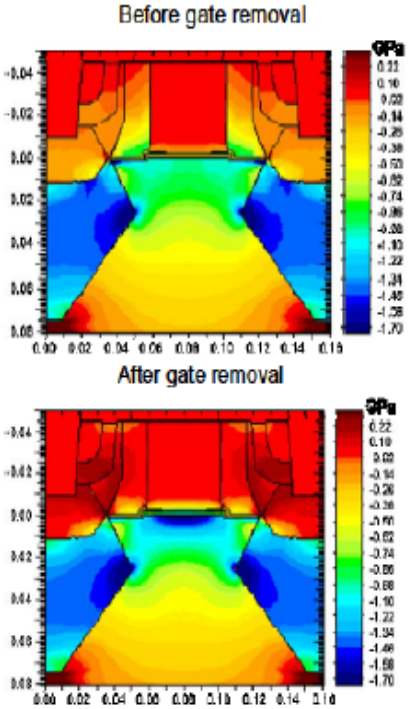
Gate-Last Enhances Channel Stress

UNIVERSITY OF NOTRE DAME



Intel's 45nm PMOS

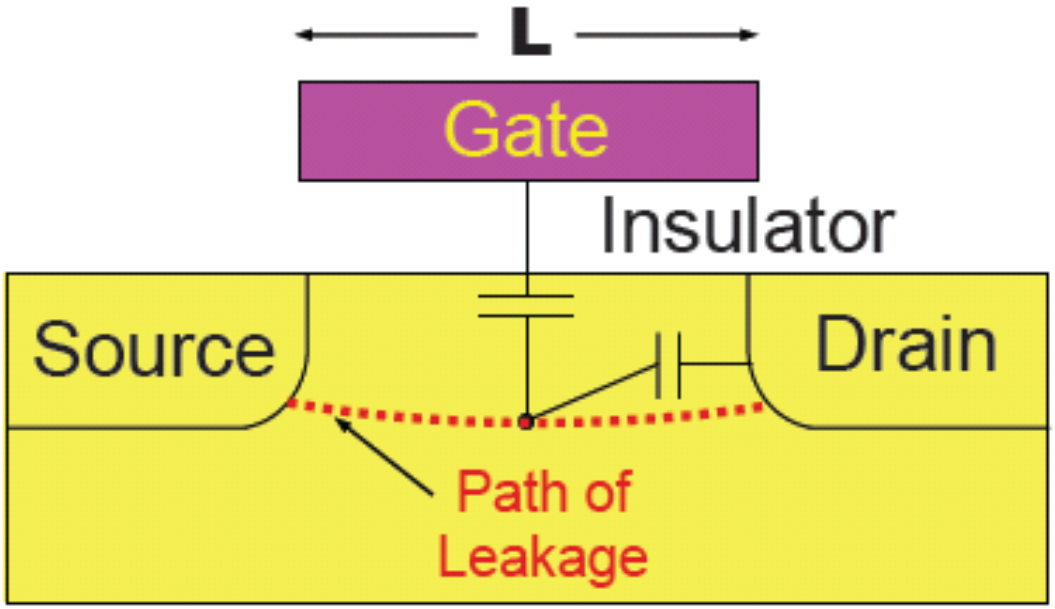
C. Auth, VLSI (2008)



Longitudinal compressive strain ε_{xx} is enhanced

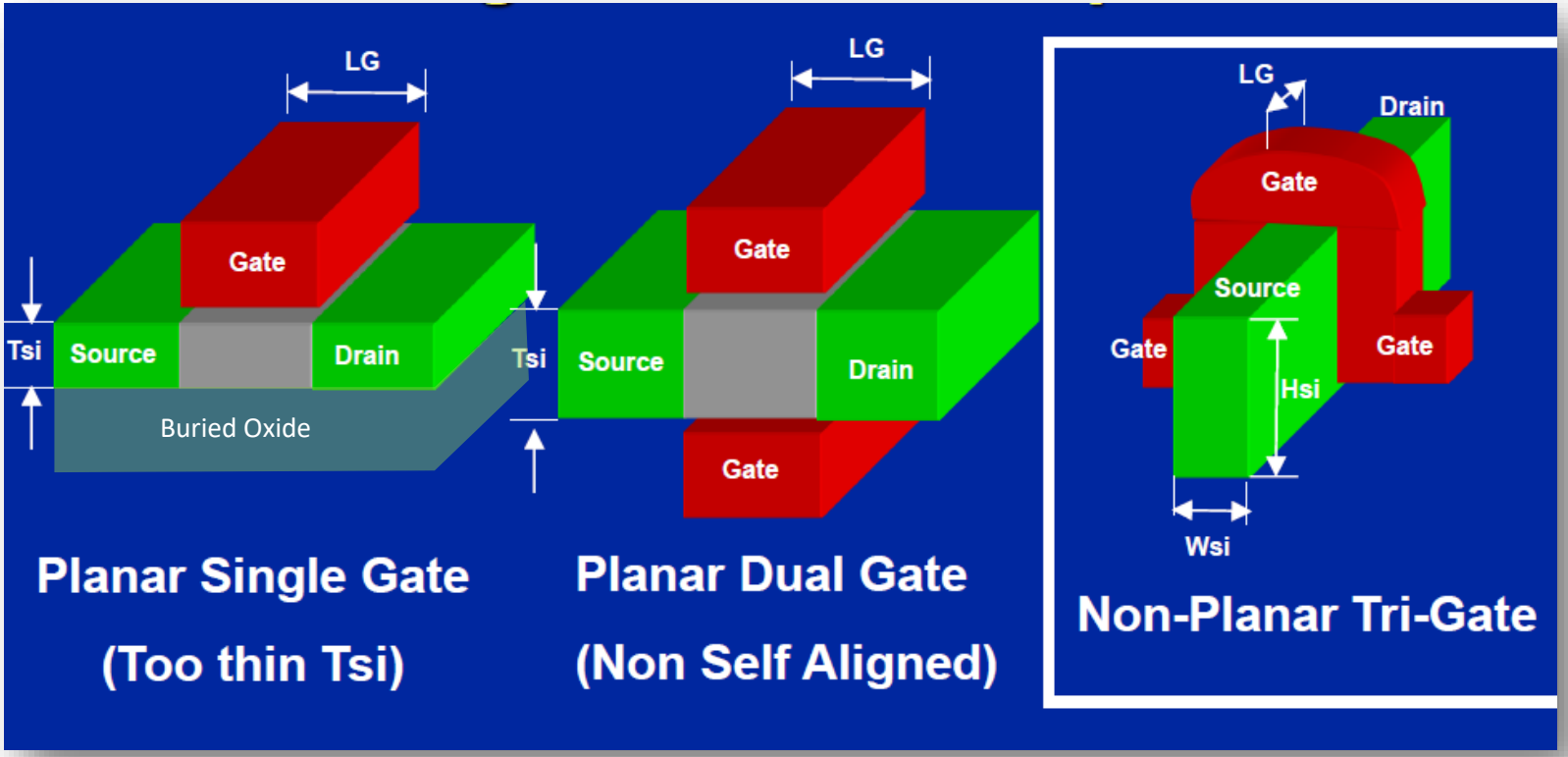
Replacement metal gate acts as effective stress enhancers – another act of serendipity

Multi-GateTransistor



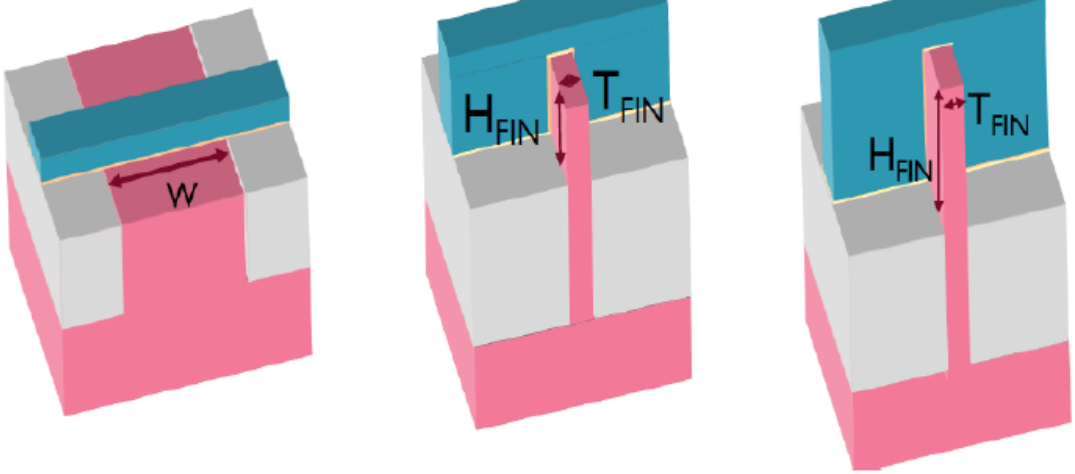
"Even thin insulators cannot control leakage paths that are away from the gate"

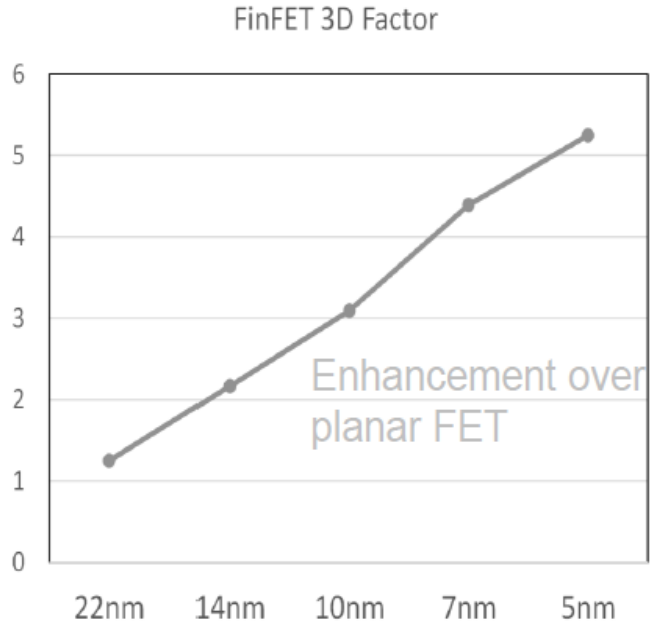
From Single-Gate to Tri-Gate



Ultra thin body, double-gate or Tri-Gate architecture allows gate length scaling

Tri-gate / FinFET



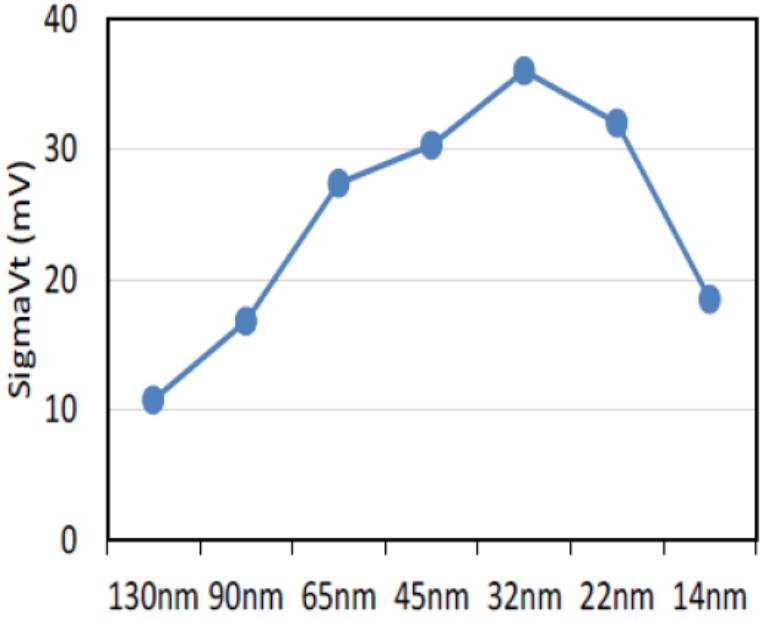


$$\text{3D Factor} = \frac{\text{Total width of FinFET}}{\text{Planar width used}} = \frac{2 \times H_{FIN} + T_{FIN}}{\text{Fin Pitch}}$$

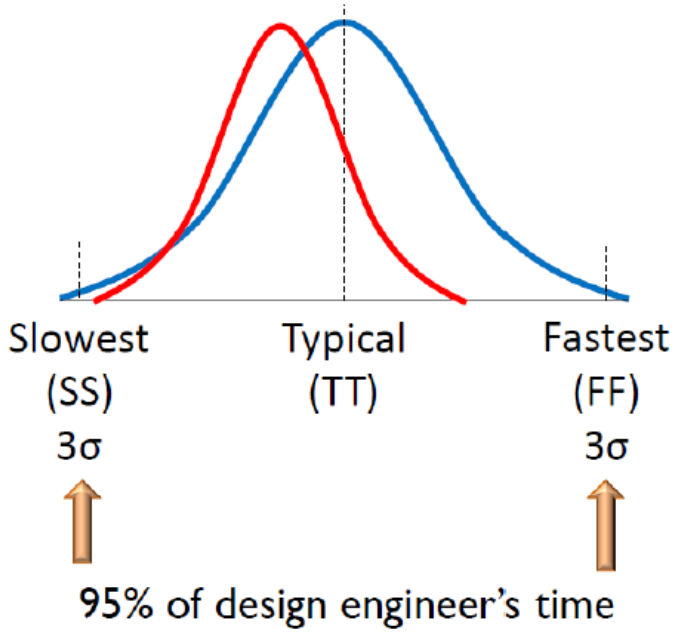
Part of Tri-gate appeal is improved electrostatics
Other part is **folded width**

Improved Variation – an act of serendipity

$$\text{Variation} \propto \frac{A_{VT}}{\sqrt{W \cdot L}}$$



S. Natarajan (Intel), IEDM 2014

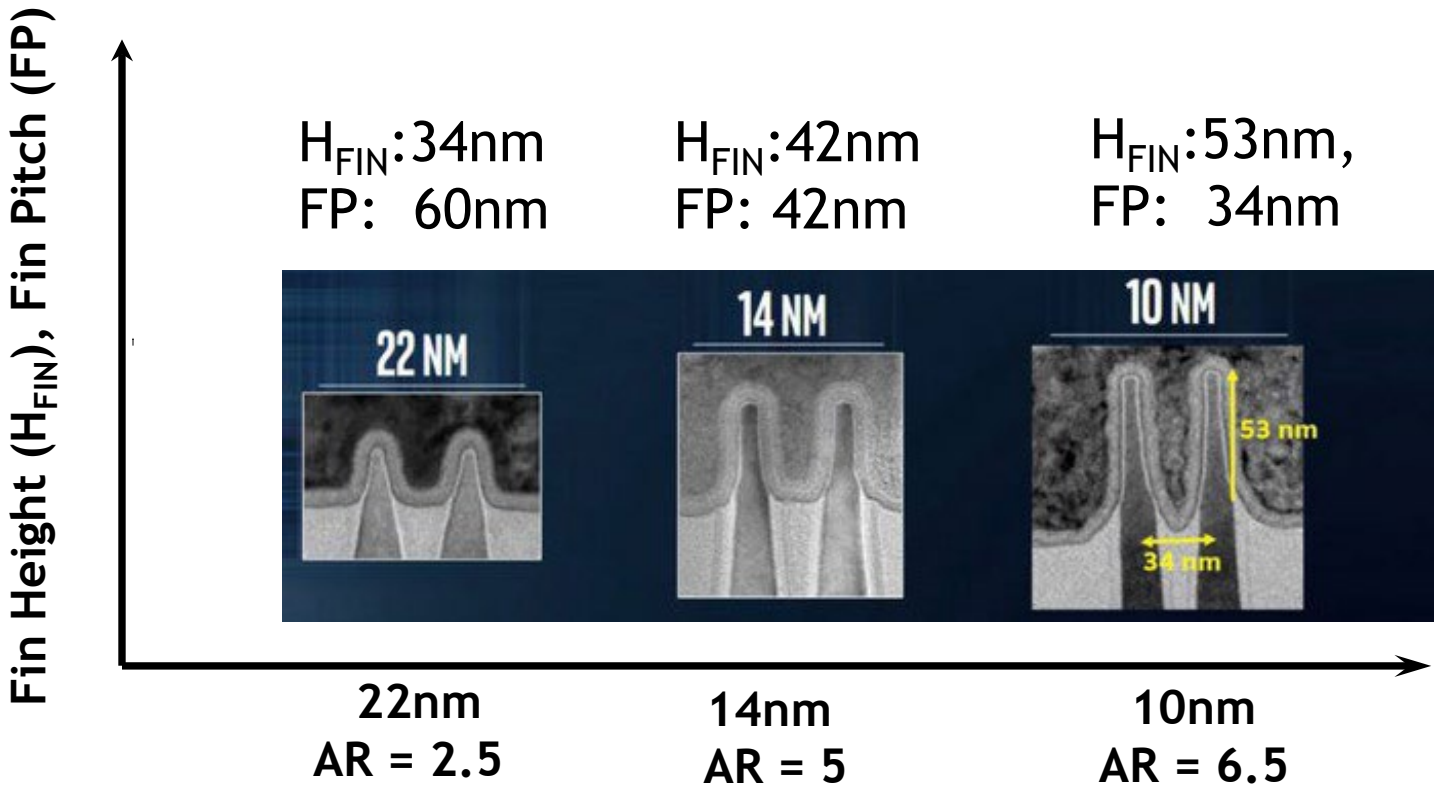


Greg Yeric (ARM), IEDM 2015

Tri-gate was intended primarily for electrostatics

3D factor enabled effective width scaling with improved variation

FinFET Evolution



Fin height and fin pitch evolution over last 3 generations

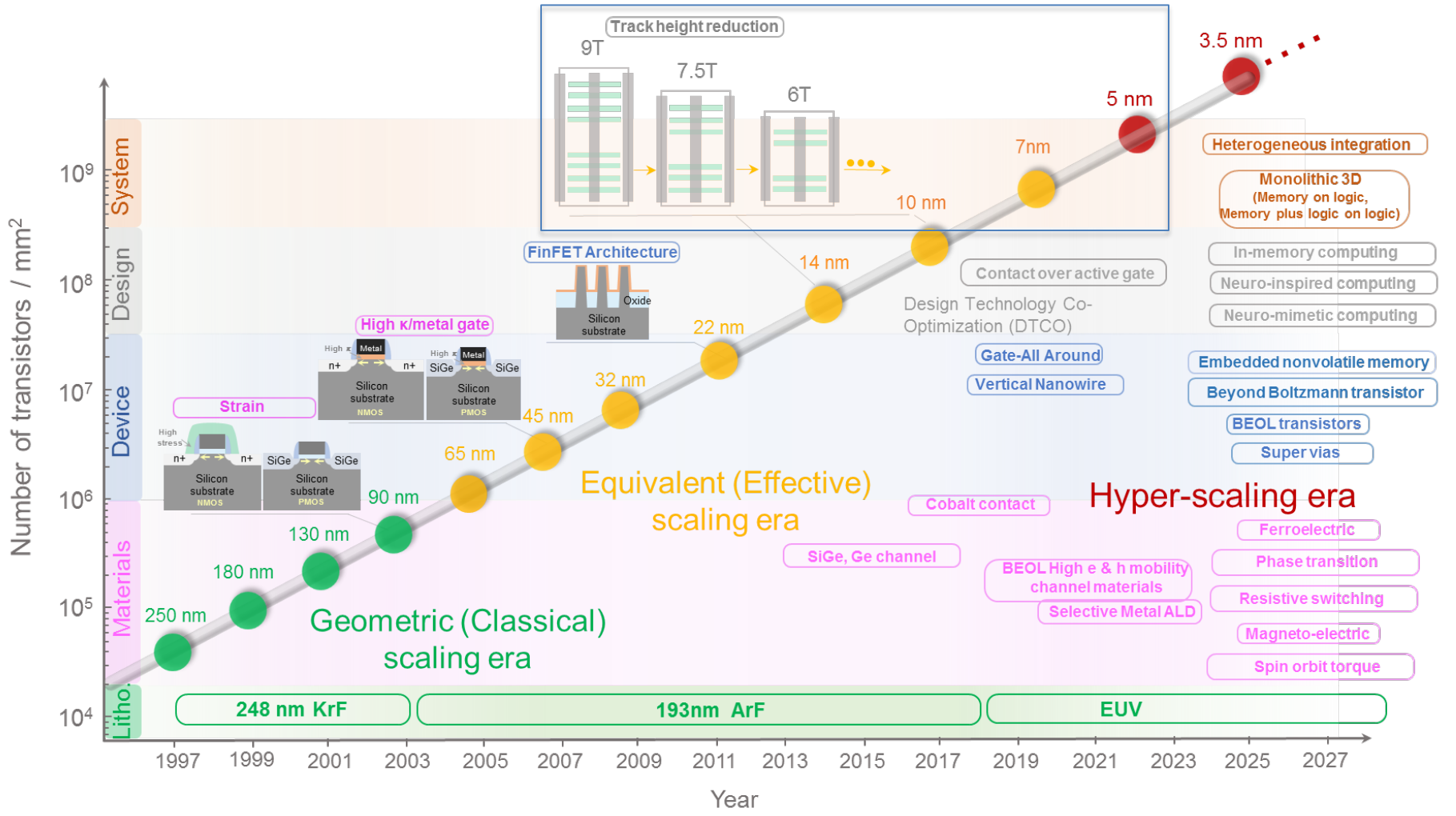
Conductance density is increasing monotonically

Outline

UNIVERSITY OF NOTRE DAME

- Past
- Present
- Future

Present

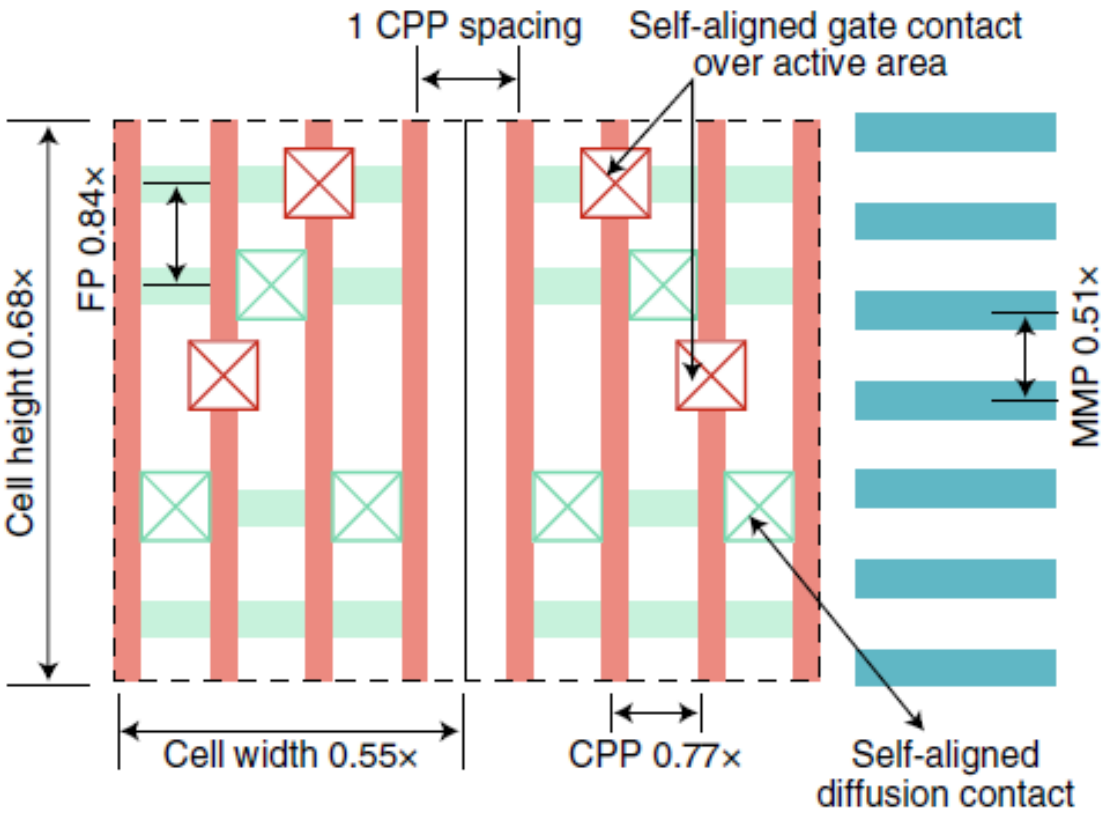


Present scaling relies on design technology co-design (DTCO)

S. Salahuddin and S. Datta, Nature Electronics, 2018

10nm (IDM) / 7nm (Foundry)

UNIVERSITY OF NOTRE DAME



Feature	14 nm	10 nm
CPP	1	0.77
FP	1	0.84
MMP	1	0.51
Cell height	1	0.68
Cell width	1	0.55
Cell area	1	0.37
Transistor density (10^6 mm^{-2})	~40	~100

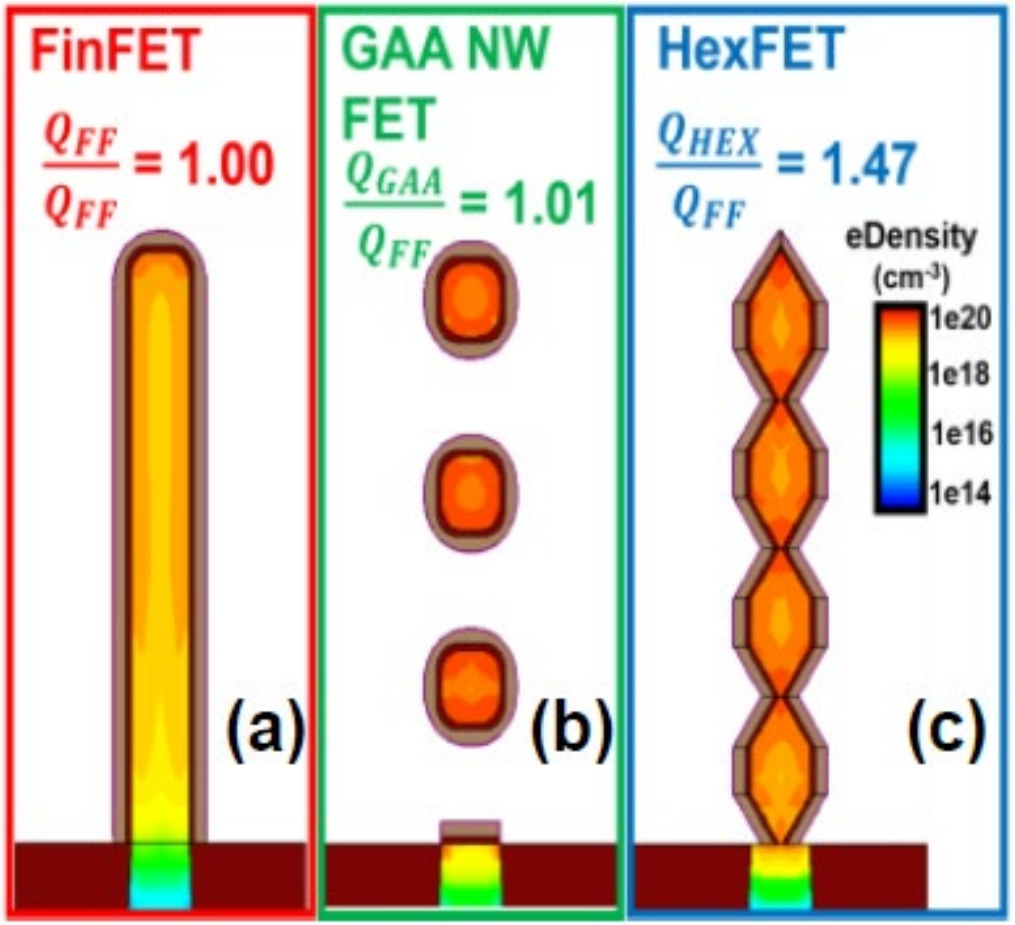
Ten nanometre CMOS logic technology

Through some unconventional approaches to improving transistor density and performance, the latest logic technology from Intel delivers 100 million transistors per square millimetre — and in the process, reaffirms Moore's law.

S. Datta Nature Electronics, Sept 2018

Extending FinFETs to GAA FETs

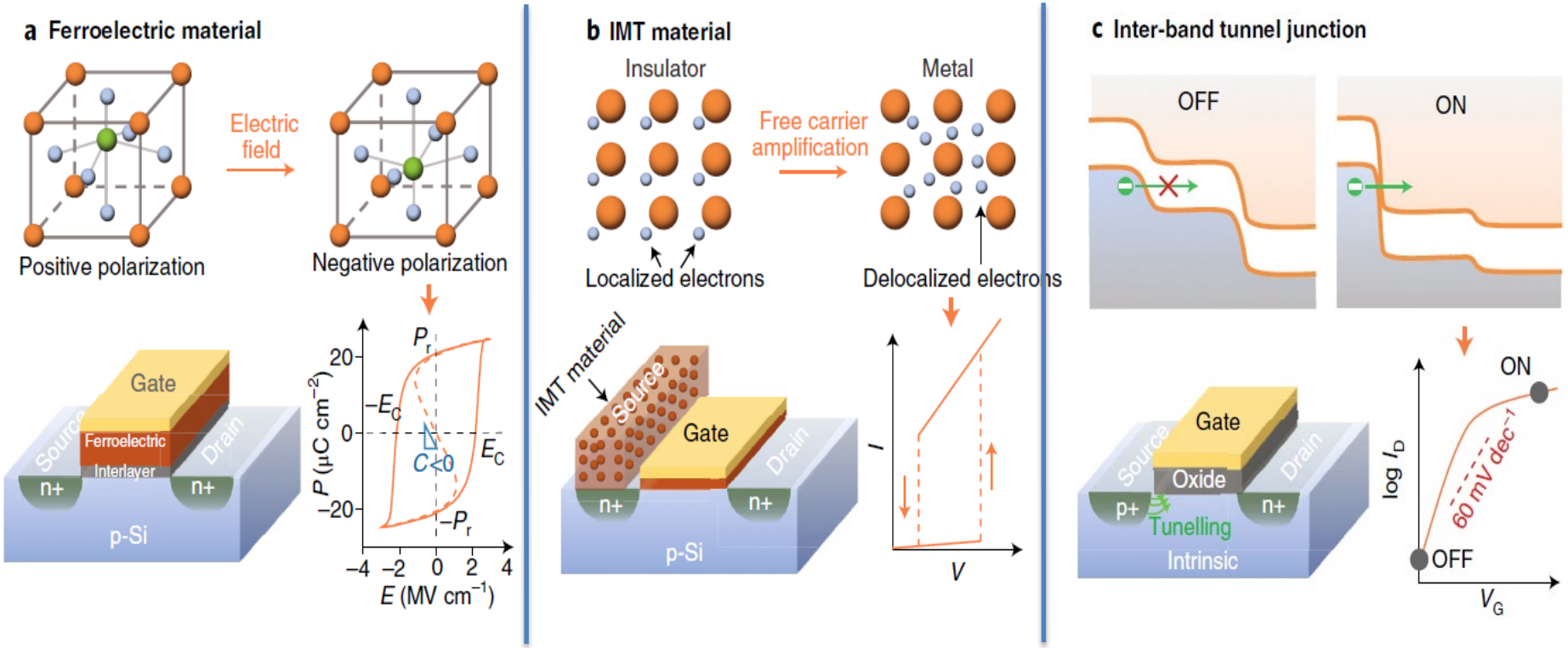
UNIVERSITY OF NOTRE DAME



GAA NW FET is being explored extensively

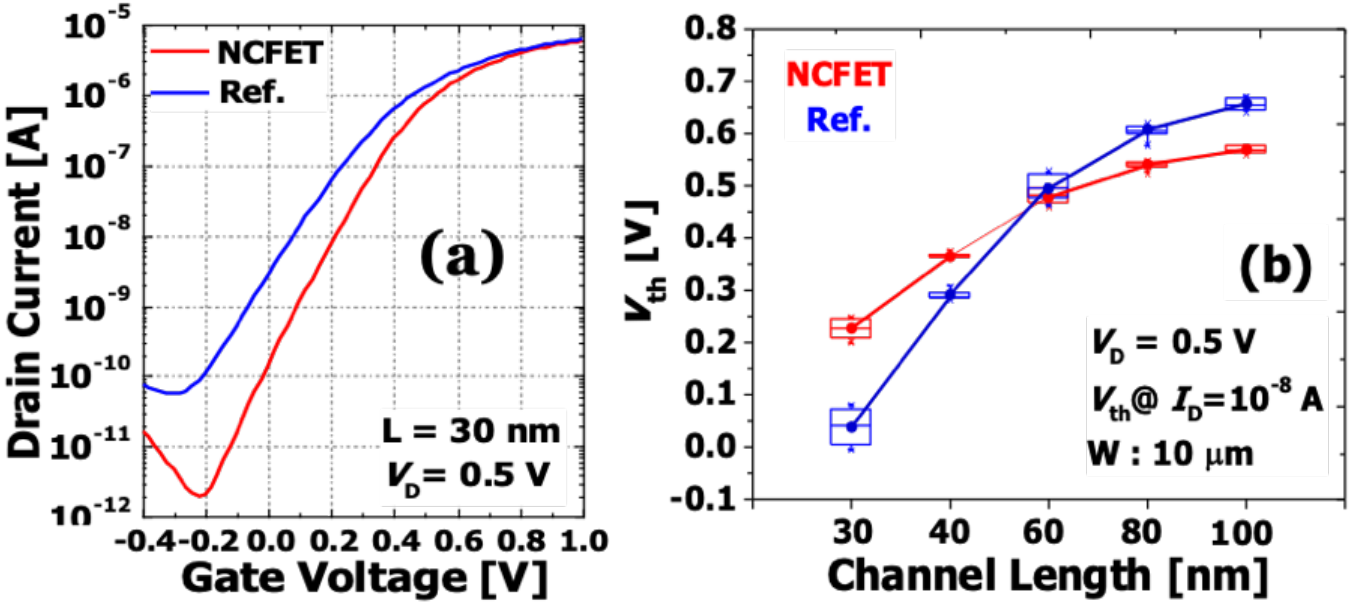
Hexagonal FinFETs as alternatives to FinFETs and GAA NWFETs

Beyond FinFETs – steep slope FETs



Ferroelectrics, insulator-metal phase transition materials, inter-band tunnel junctions are being explored for transistors with steep slope and low voltage operation

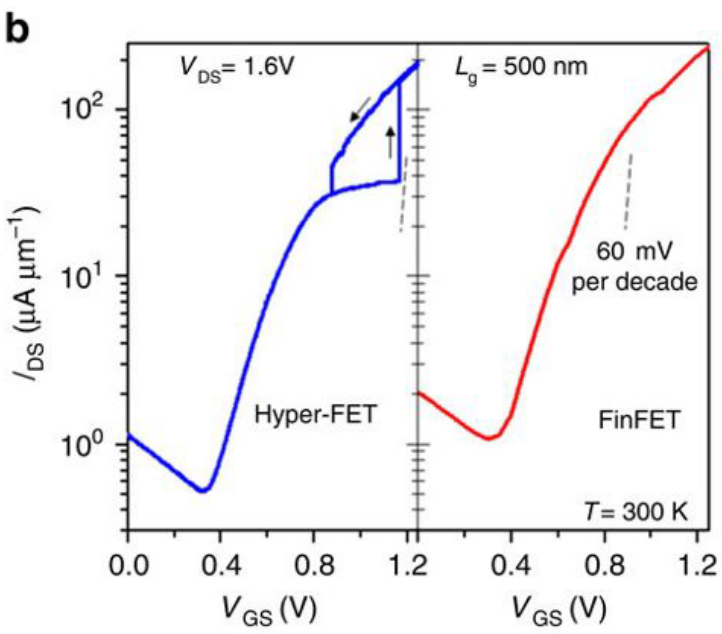
Negative Capacitance FETs

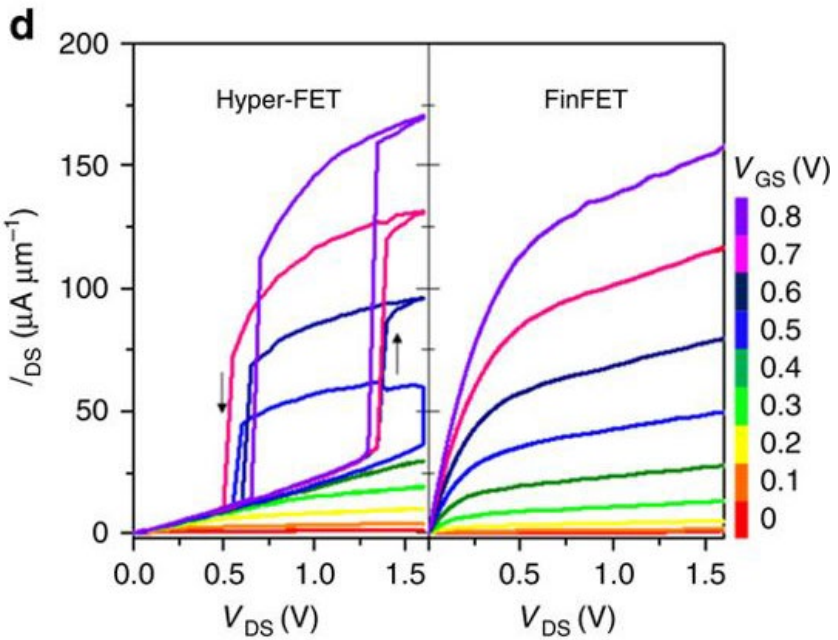


Negative capacitance (NC) FETs show short channel effect improvement over traditional FETs

S. Salahuddin, EDL, 2018

Phase Transition FETs



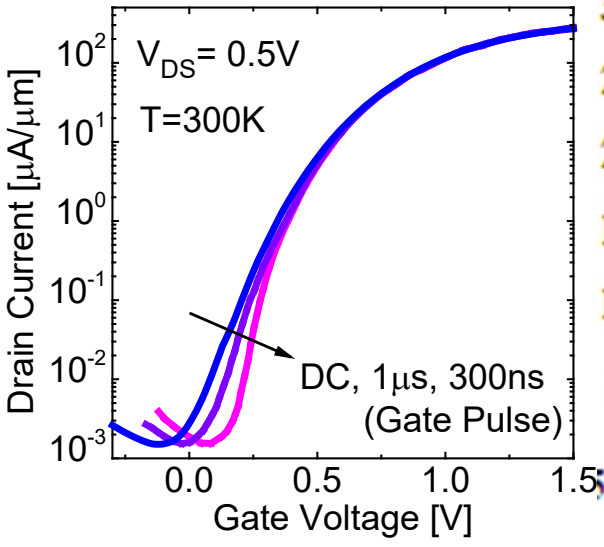


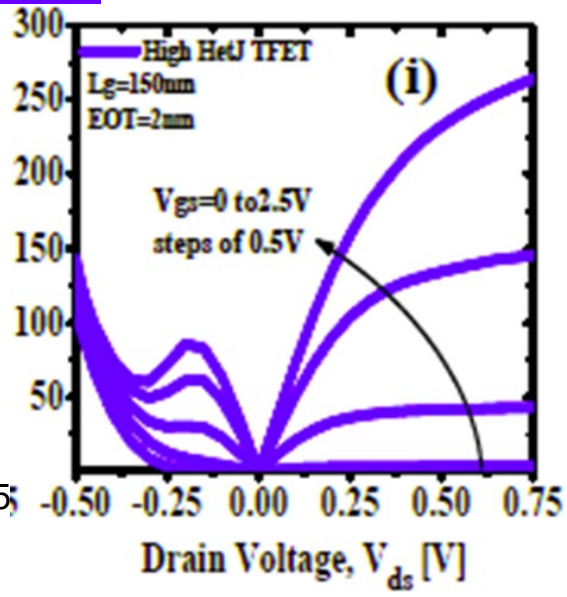
Phase FETs show on-off ratio benefit over traditional FETs (albeit with hysteresis)

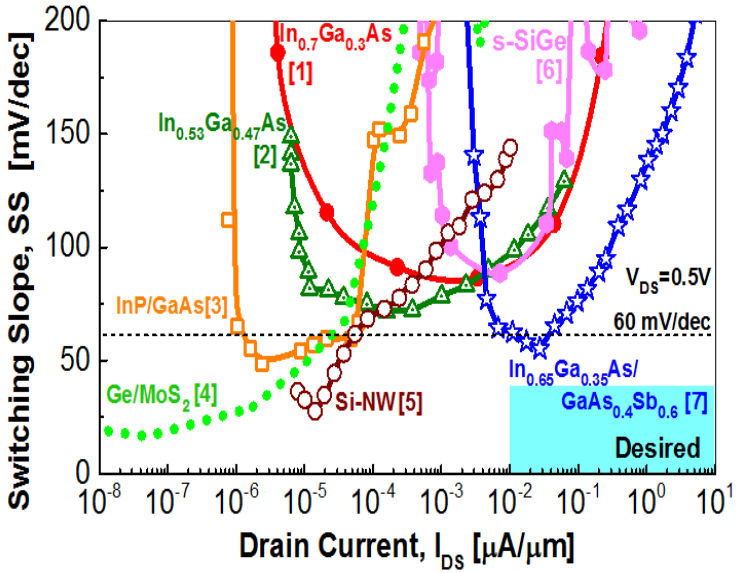
N. Shukla, S. Datta, Nature Communications, 2015

Tunnel FETs

UNIVERSITY OF NOTRE DAME







Tunnel FETs show benefit over sub-threshold CMOS

[1] H. Zhao et al, IEEE EDL, Dec. 2010

[2] M. Noguchi et al., IEDM 2013

[3] B. Ganjipour et al., ACS Nano, Apr. 2012

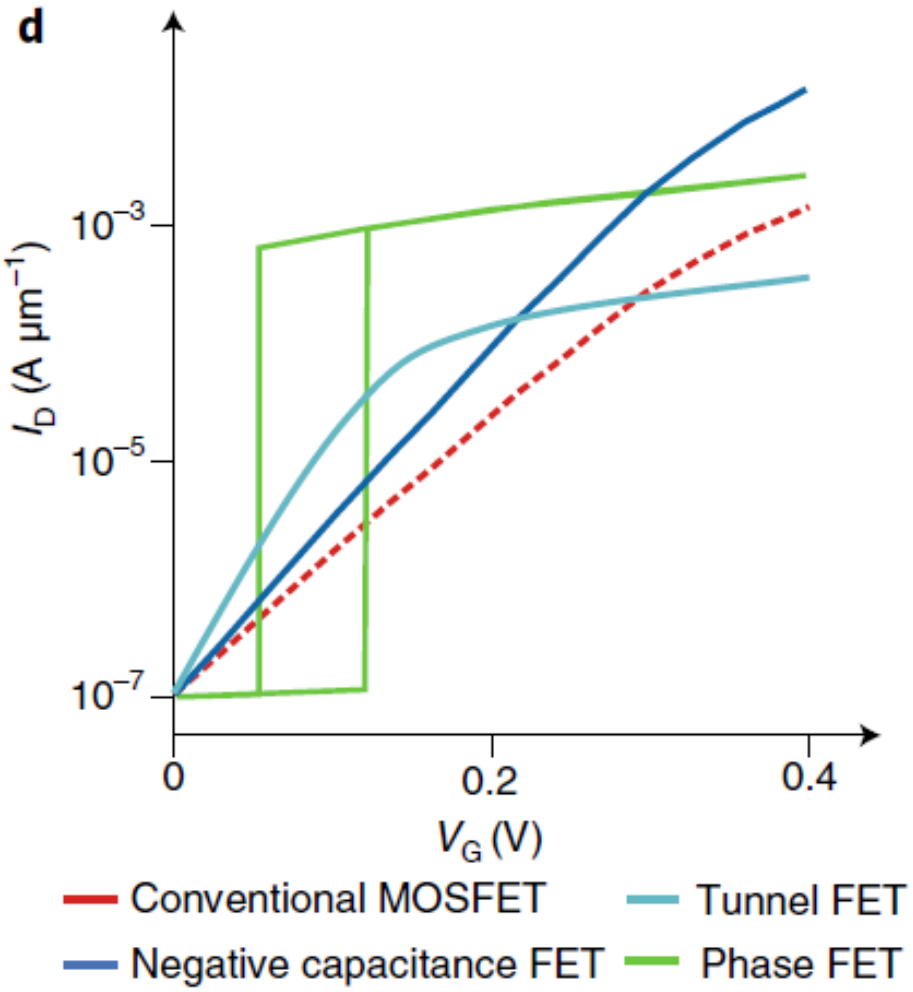
[4] D. Sarkar. et al., Nature, Oct. 2015

[5] L. Knoll et al., IEEE EDL, June 2013

[6] A. Villalon et al., VLSI 2012

[7] R. Pandey et al., VLSI 2015

Steep Slope FETs



Negative capacitance FET has emerged as the most promising steeper slope transistor option

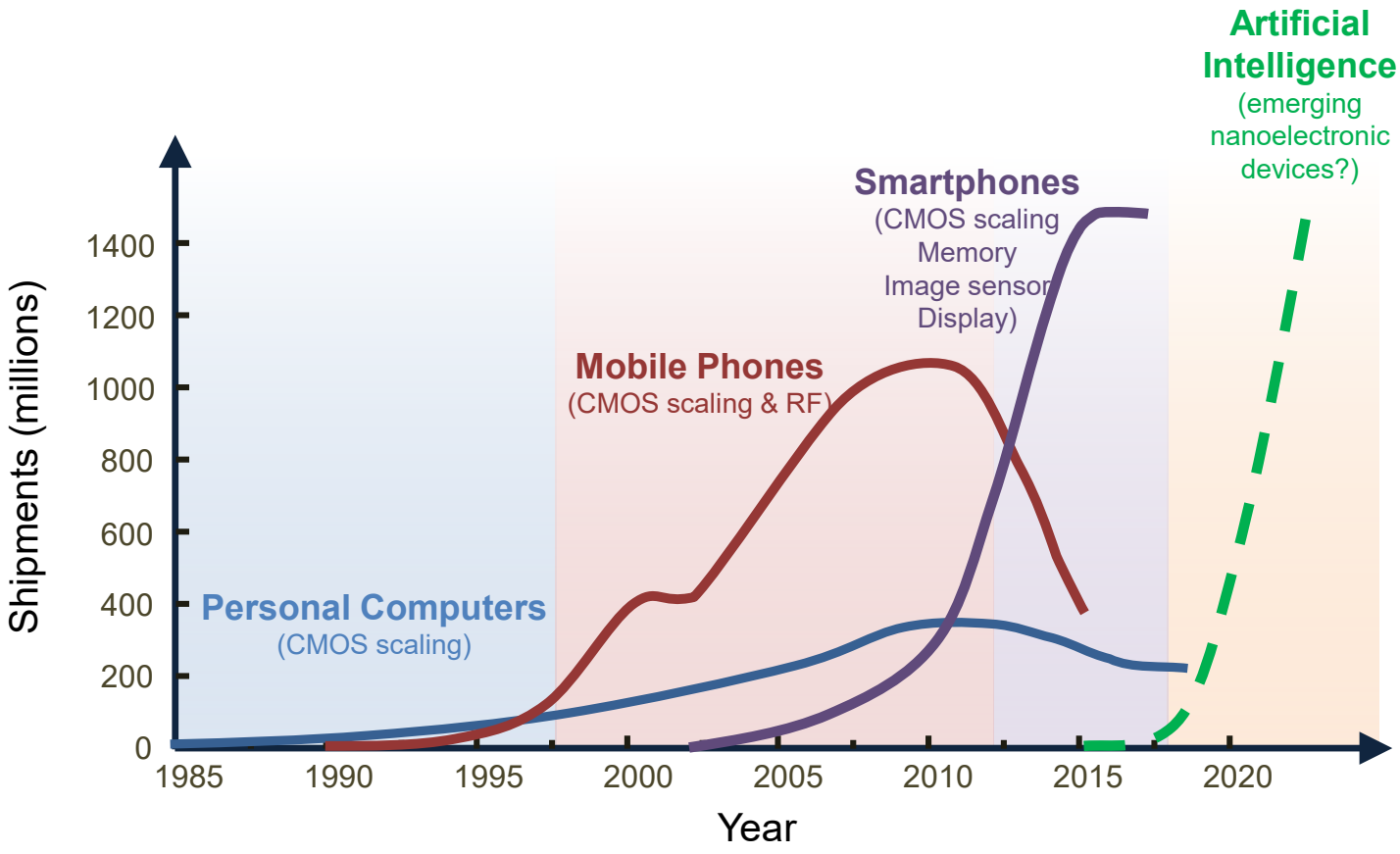
S. Salahuddin and S. Datta, Nature Electronics, 2018

Outline

UNIVERSITY OF NOTRE DAME

- Past
- Present
- Future

The Fourth Wave of Computing



Data intensive computing will drive the next era of hardware scaling

Source: https://www.economist.com/briefing/2015/02/26/the-truly-personal-computer
https://www.statista.com/chart/12798/global-smartphone-shipments/

Hyper-scaling

UNIVERSITY OF NOTRE DAME

- Architecture: Ability to scale appropriately to meet increased workload demand

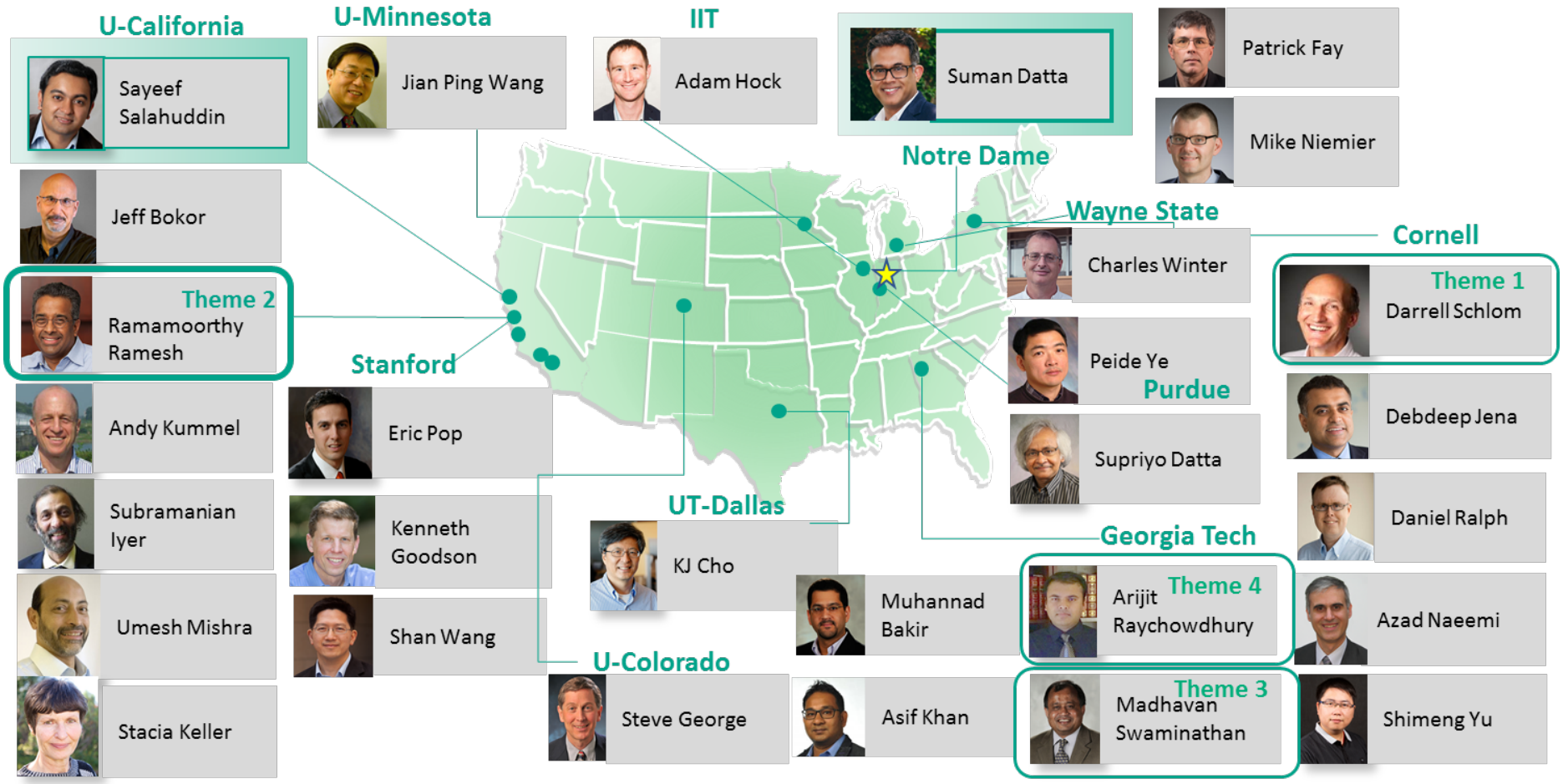
- Technology: Ability to scale beyond x-y dimension to meet system specifications of power, performance, area, function, form-factor and cost
  - Monolithic integration
  - Pseudolithic integration
  - Merged logic-memory fabrics
  - Neuro-inspired computing fabrics

ASCENT

Applications and Systems Driven Center for Energy-Efficient Integrated Nanotechnologies

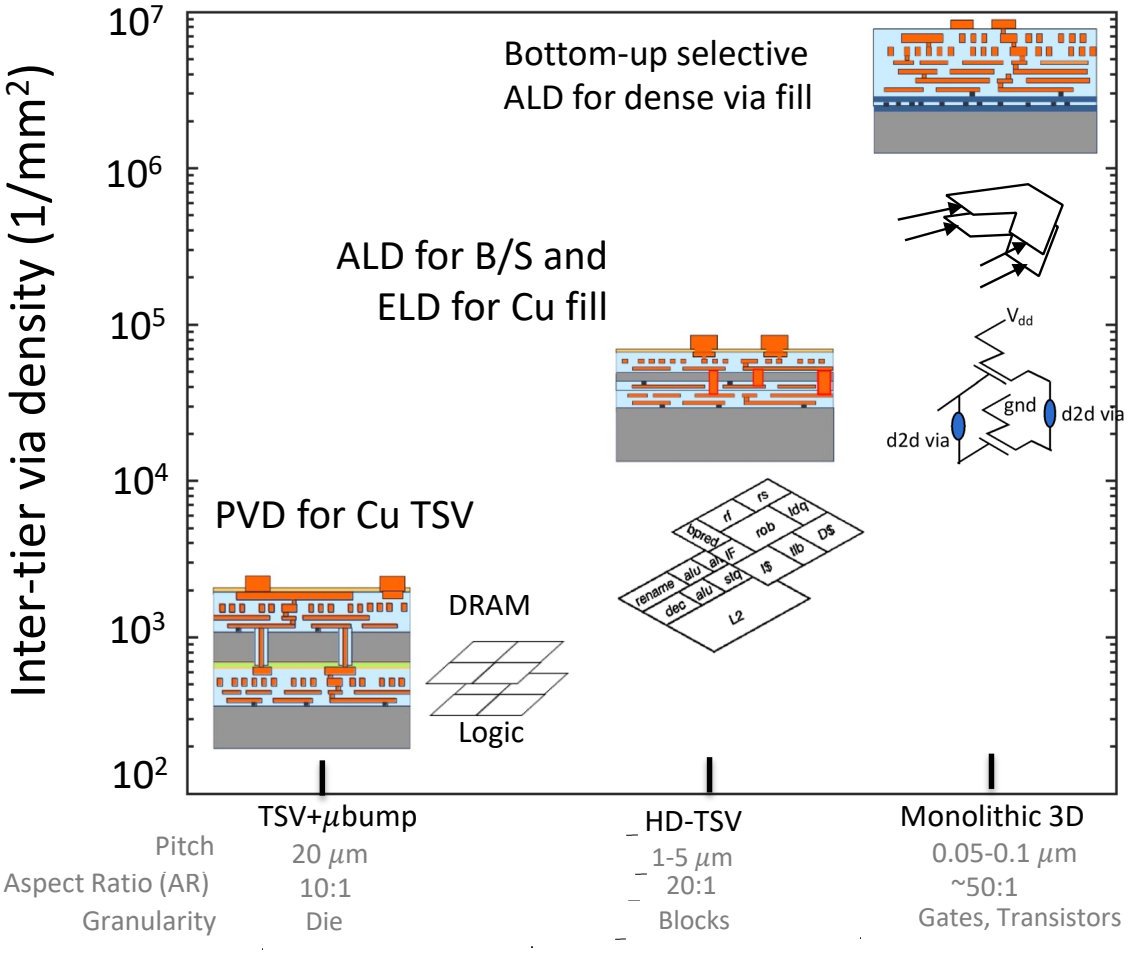
ascent.nd.edu

UNIVERSITY OF NOTRE DAME



ASCENT is sponsored jointly by the Semiconductor Research Corporation (SRC) and DARPA

Monolithic 3D



Levels of granularity:

- ❑ Die level
- ❑ Block level
- ❑ Gate level
- ❑ Transistor level

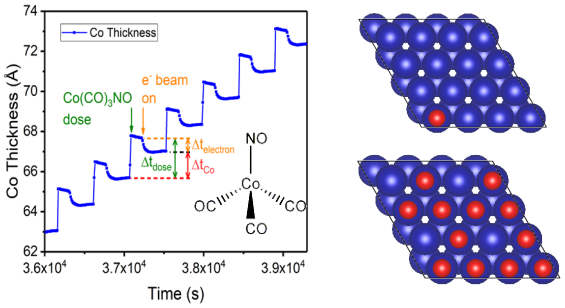
Grand Challenges:

- ❑ Protect bottom layer transistors
- ❑ Fabricate upper layer transistors
- ❑ Align top layer with bottom layer
- ❑ Fill high AR inter-layer vias with low resistivity metals
- ❑ Embed thermal management structures
- ❑ Assess system performance (ROI)

Monolithic 3D

UNIVERSITY OF NOTRE DAME

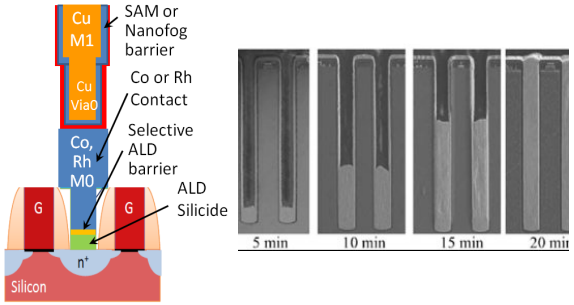
Electron Enhanced ALD



- EE-ALD of Cobalt
- Computational growth study

George, Hunter (Colorado, Wayne St)

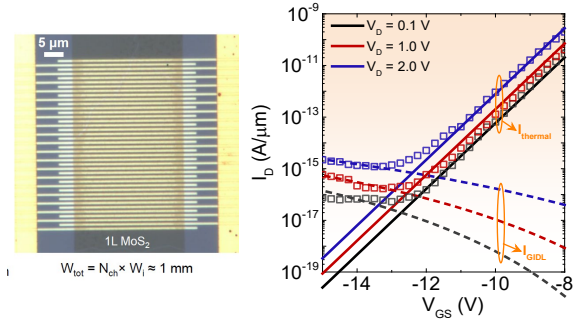
Self-aligned Contacts & Vias



- Hyper Selective ALD of silicide
- Hyper-elective Co (and Ru) ALD

Kummel (UCSD)

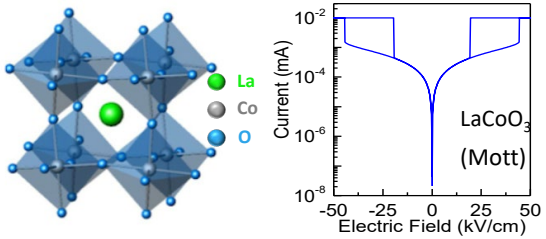
Atto-amp I_{off} FETs



- ML MoS_2 layer on silicon
- 5nm thick In_2O_3 FETs on silicon

Pop, Datta (Stanford, Notre Dame)

3D Memory & Selectors

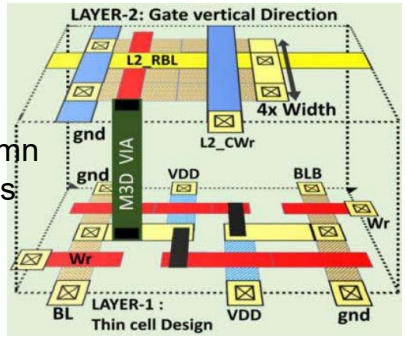


- Selectors for 3D Memory; Mott, Ovonic and Metal Ion Threshold selectors

Schlom, Wong, Datta (Cornell, Notre Dame, Stanford)

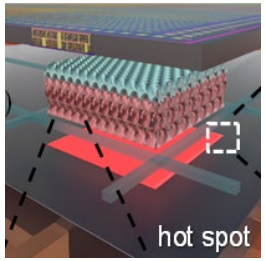
Sequential M3D Design

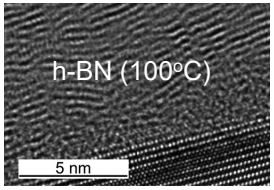
- 3D SRAM with concurrent row & column data access
- 3D DRAM



Datta, Narayanan (Penn St, Notre Dame)

M3D Thermal Management

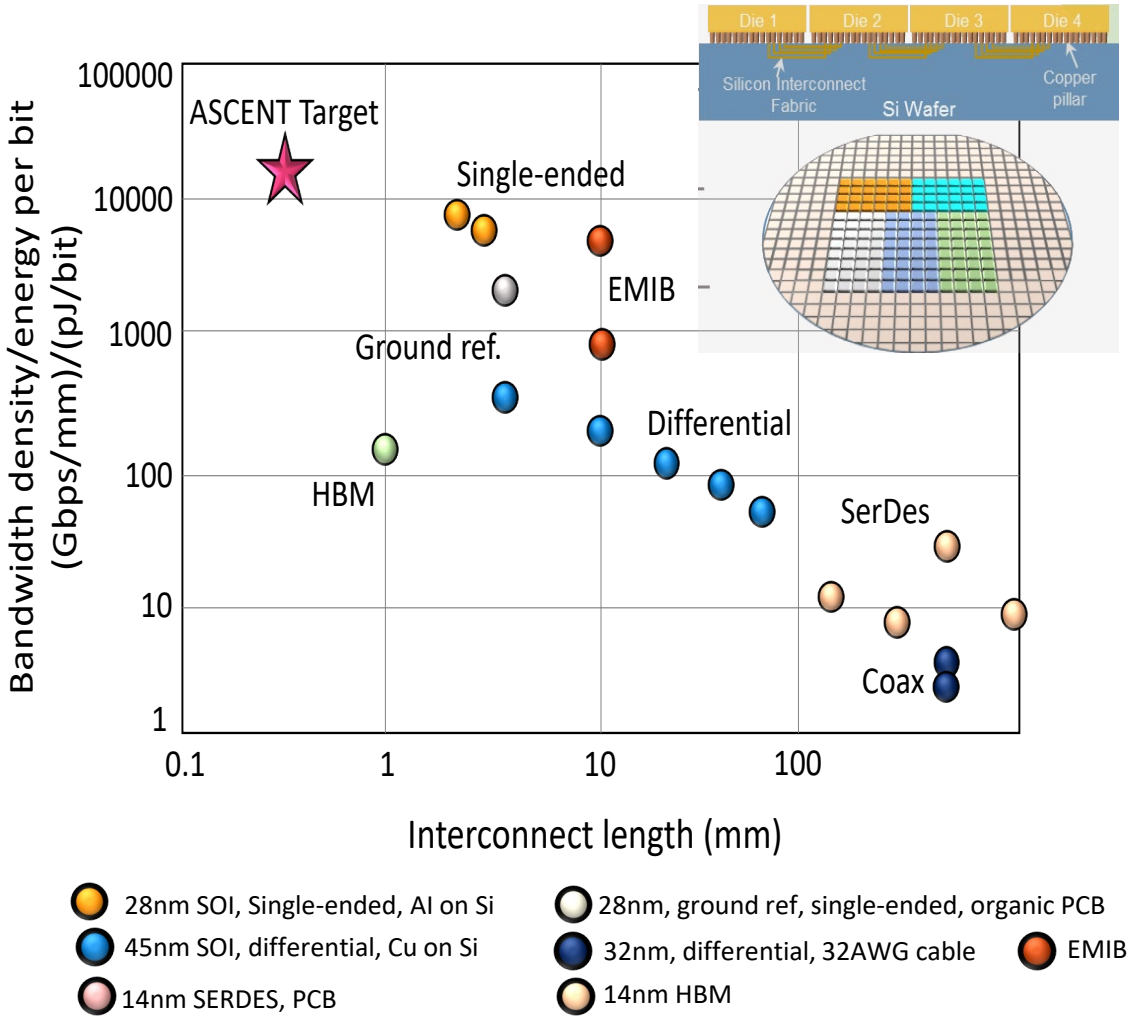




- Layered materials (hBN) with anisotropic thermal conductivity

Pop, Goodson (Stanford)

Heterogeneous Integration & Advanced Packaging



Why Heterogeneous Integration

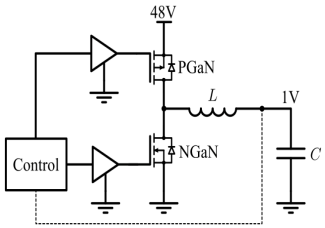
- PCB ultimately limits size, weight, area, performance (SWAP) of microsystems
- Enables silicon IP reuse, reduces design cost and improves system performance

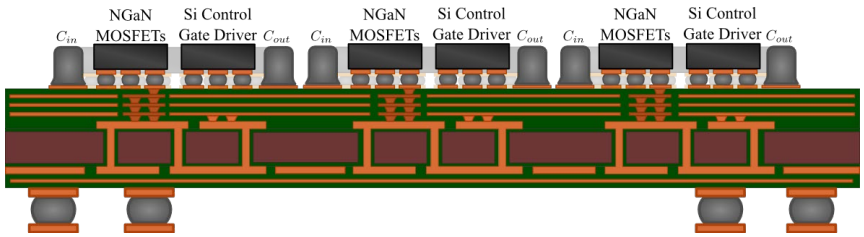
Grand Challenges

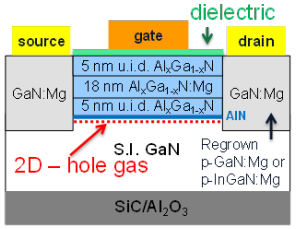
- Target 1um interconnect spacing and <20 um die-to-die spacing
- Target aggregate data transfer rate of 1,000 Gb/s/mm at < 0.1pJ/bit over 50-500 um distance

Heterogeneous Integration Fabric

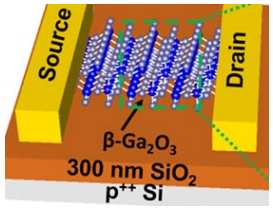
Power Delivery







GaN n/p FET



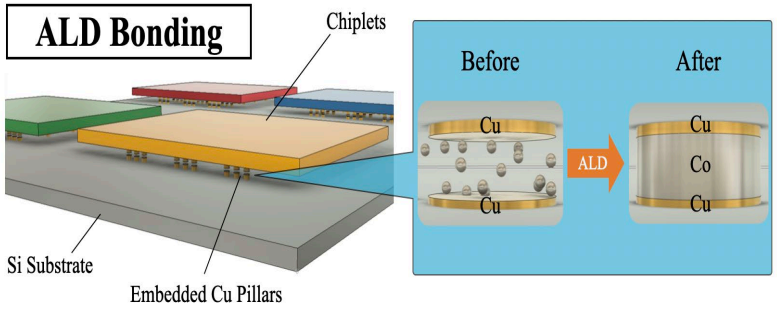
β-Ga_2O_3 nFET

Package integrated 48-to-1V voltage converter & regulator for data centers

Raychowdhury, Misra (Georgia Tech, UCSB)

Ye (Purdue)

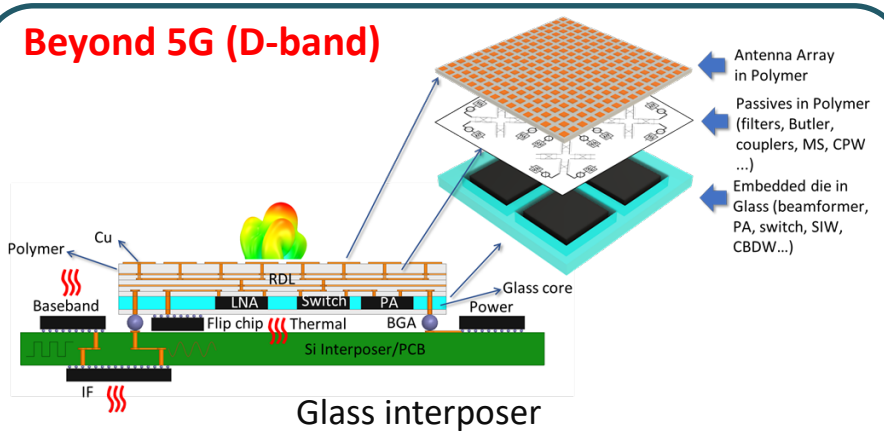
High-speed Digital Signaling (Logic + Memory)



Low temperature (200°C) ALD bonding shows the potential to form nanoscale I/O's

Kummel, Bakir (UCSD, Georgia Tech)

Beyond 5G (D-band)



Glass interposer

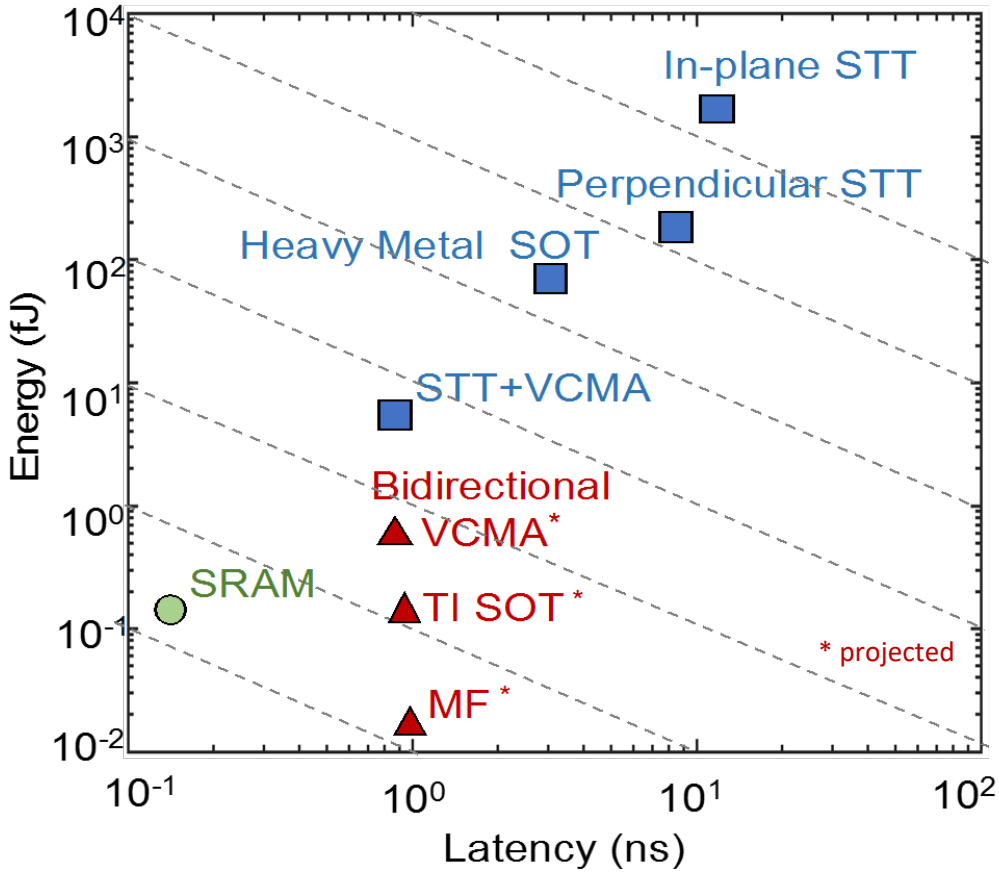
RF Switches

BAW FET Filters

High Gain 3D Antennas + SIW

Datta, Fay, Jena, Swaminathan (Notre Dame, Cornell, Georgia Tech)

Beyond CMOS (spin)



Grand Challenges

Efficient Write

- ❑ Spin-orbit torque (SOT)
- ❑ Magneto-electric (voltage controlled magnetism)
- ❑ Voltage controlled magnetic anisotropy (VCMA)

Efficient Read

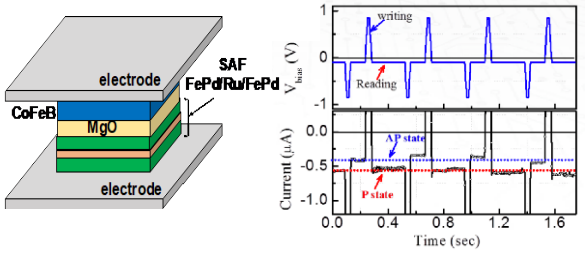
- ❑ MTJ free read-out of magnetic information

  (e.g. Inverse Rashba effect)

Beyond CMOS (Merged logic-memory)

UNIVERSITY OF NOTRE DAME

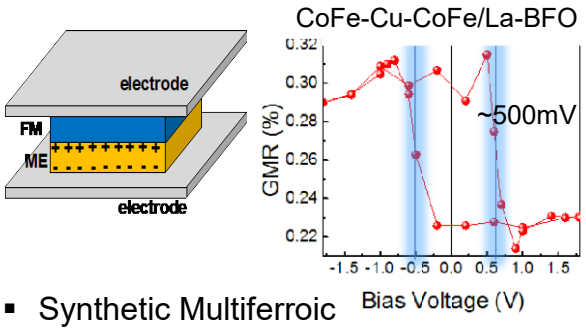
Bi-directional VCMA



- Field-free bidirectional VCMA with SAF
- p-MTJ with J~10^5 A/cm^2 demonstrated

JP Wang (Minnesota)

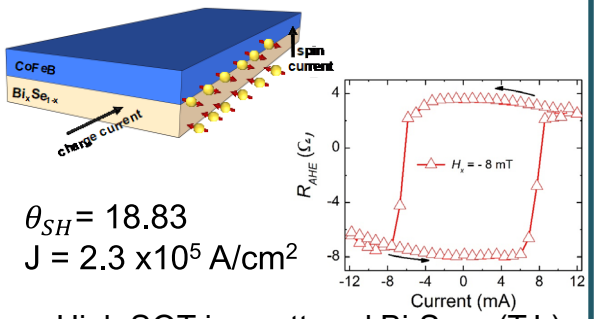
Magneto-electric (ME)



- Synthetic Multiferroic
- Voltage-driven bidirectional switching

Ramesh (Berkeley)

Topological Insulator (TI)

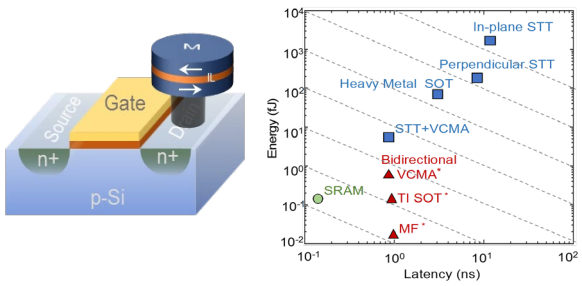


θ_{SH} = 18.83
J = 2.3 x10^5 A/cm^2

- High SOT in sputtered Bi_xSe_{1-x} (T.I.) and Iridate

Wang (Stanford, Minnesota)

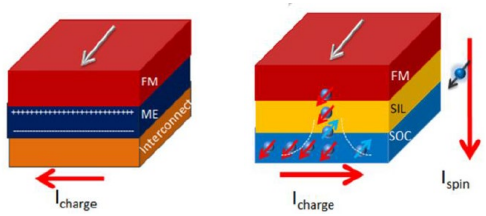
Embedded Memory



- Logic compatible, high endurance, embedded memory applications

JP Wang, Salahuddin (Berkeley)

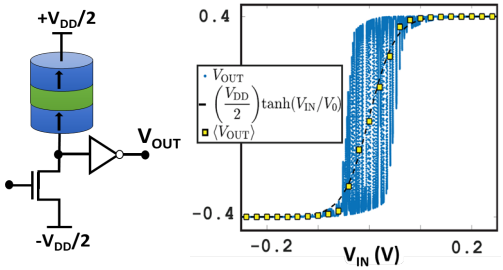
Spin Logic (Majority gate)



- 10aJ class logic operation

Ramesh (Berkeley)

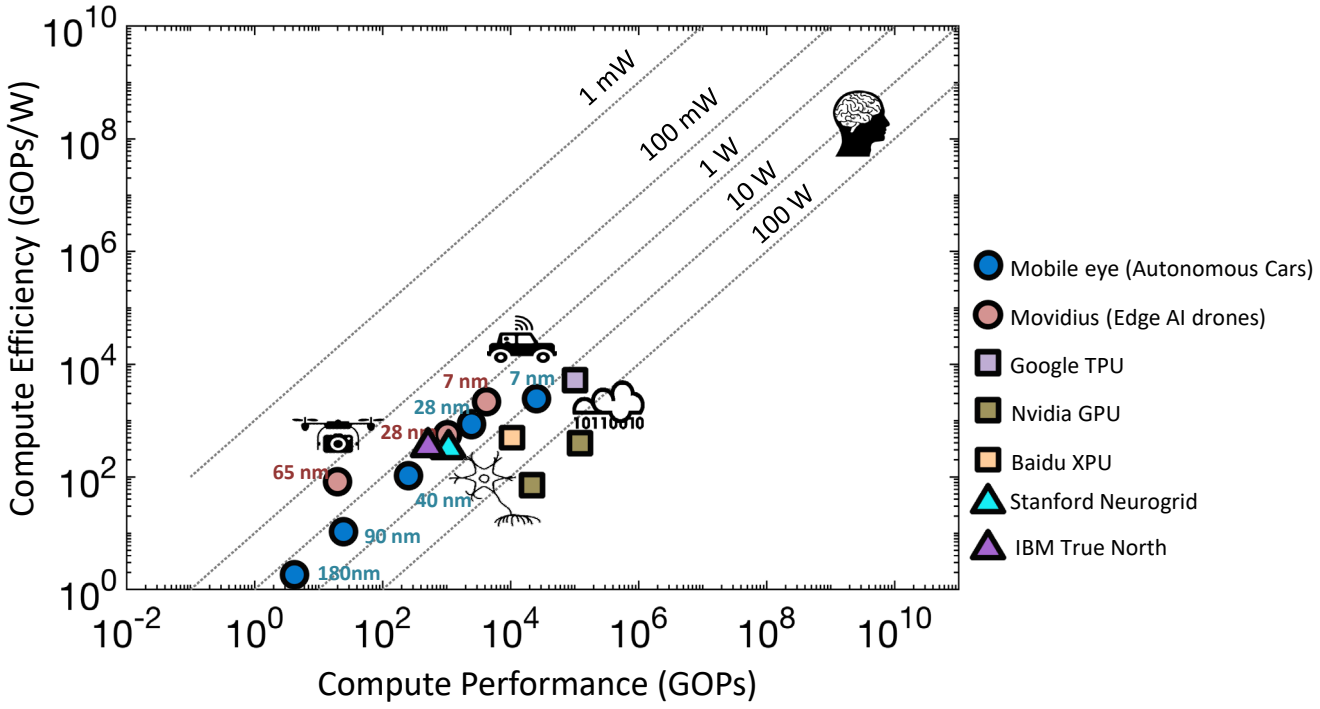
Unconventional Computing



- p-bits for Deep Belief Network, Ising machines

Salahuddin, Datta (Berkeley, Purdue)

Hardware for AI

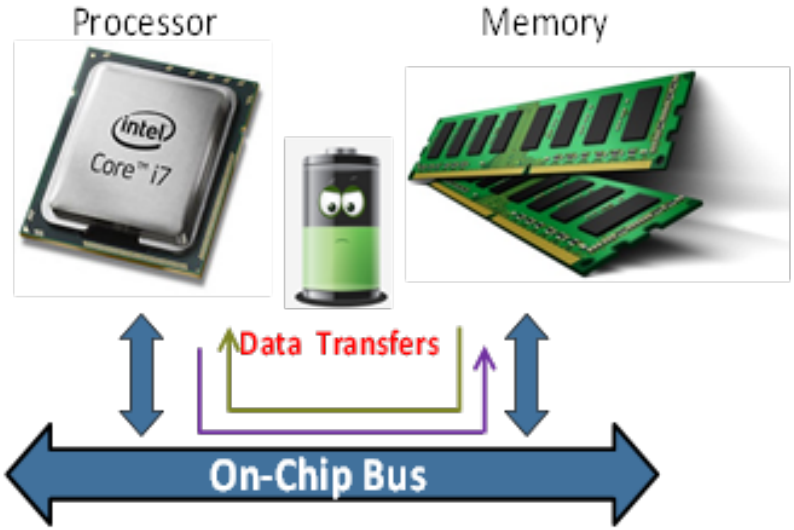


AI hardware addresses different applications with appropriate energy efficiency and performance, but far from that of brain

Neuro-inspired Computing Fabric

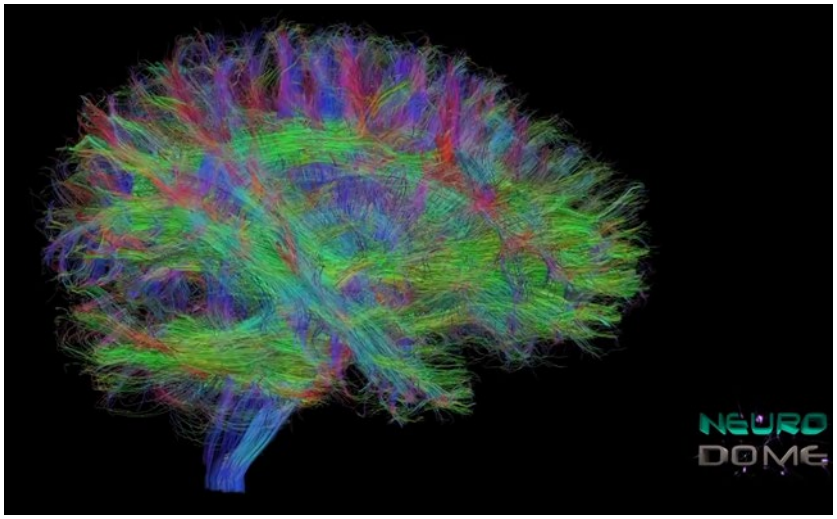
UNIVERSITY OF NOTRE DAME

Von Neumann



- Fully programmable
- Suitable for high precision computing
- Separate logic and memory
- Limited by memory bandwidth
  where read/writes are performed
  at course granularities

Brain

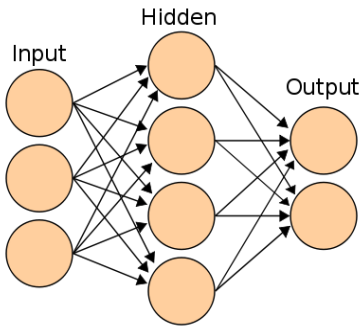


- Massively parallel architecture supported by dense connectivity
- Suitable for life long learning and decision making in changing environment
- Co-located logic and memory
- Based on Markov decision processes and resilient to instantaneous errors

Neuro-inspired Computing Fabric

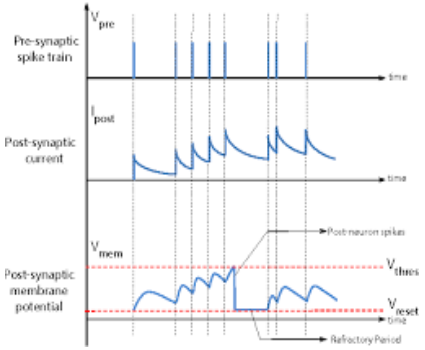
UNIVERSITY OF NOTRE DAME

Artificial Neural Networks (ANNs)



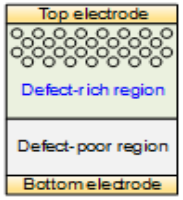
- Digital
- Memory-centric
- Synchronous

Spiking Neural Networks (SNNs)

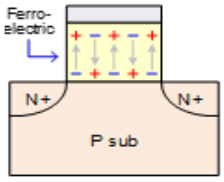


- Analog Primitives
- Time-encoding
- Asynchronous

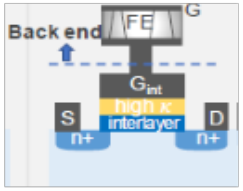
Non-filamentary RRAM Synapse



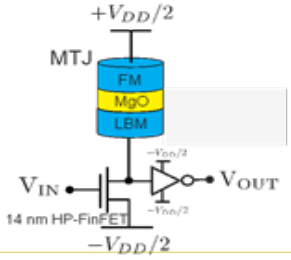
FerroFET Analog Synpase



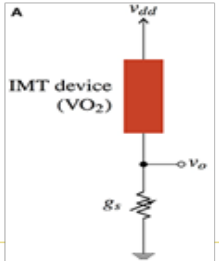
FeMFET Synapse



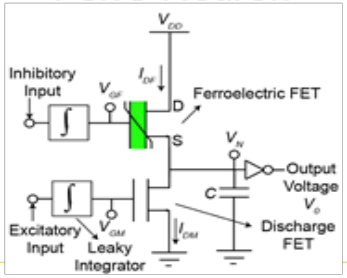
Spin Neuron



IMT Neuron



Ferro Neuron

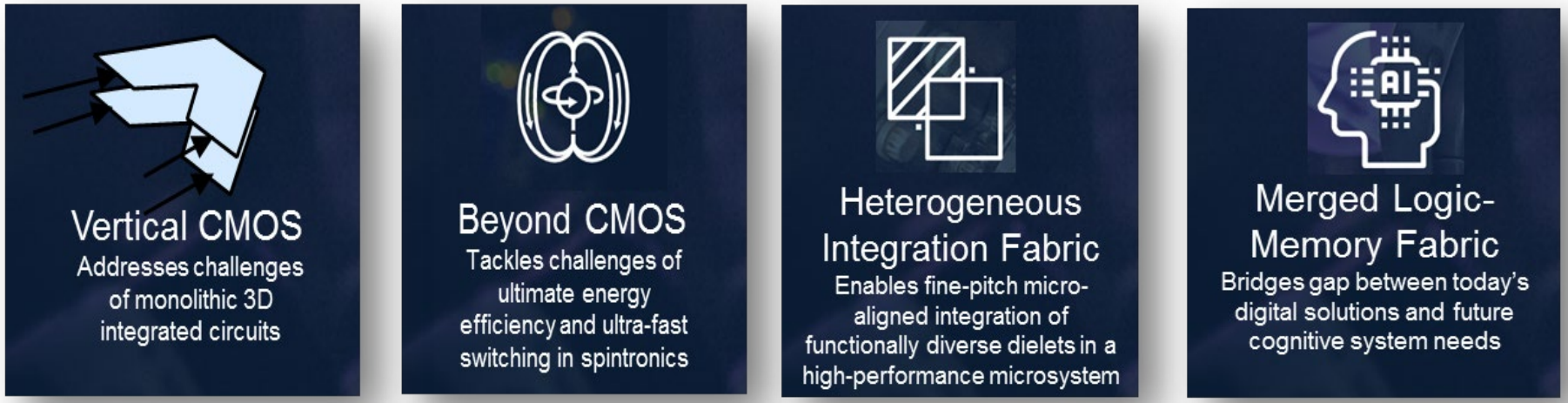


Raychowdhury, Yu, Khan (Georgia Tech)

Datta, Niemier (Notre Dame)

Salahuddin (Berkeley)

Key Takeaway



Lessons from the Past: Serendipity and not luck. It's being in the right place at the right time with the right training and mindset

Present state of Moore's Law: Not a Law of Physics. It's hard work by scientists and engineers, our ingenuity and audacity to compete and thrive, driven by free market economics

Future of Electronics: Materials scientists, device physicists, circuit designers, chip architects will collaborate to provide semiconductor innovations that provide system level benefit

"Such an exponential will continue for a very long time"

Thank You

UNIVERSITY OF NOTRE DAME