Heterogeneous Integration of Nanomaterials & Devices: 1D, 2D, 3D

Eric Pop

with M. Aly, C. English, G. Hills, S. Mitra, A. Salleo, P. Wong

Electrical Engineering (EE)
Materials Science & Engineering and SystemX Alliance
Stanford University

http://poplab.stanford.edu

Abstract

Abundant-data applications of the 21st century will arise from social networks, e-commerce transactions, genome sequences, and trillions of interconnected sensors which could create a data deluge. Future technologies must be able to process and classify such data in real-time, in an energy-efficient manner, a challenge which will expose gross inefficiencies in traditional computing architectures. This short course will introduce the **Nano-Engineered Computing Systems Technology (N3XT) approach** [1] within the context of the **Stanford SystemX Alliance** [2], with particular focus on heterogeneous integration of low-power nanoelectronics. Our groups of researchers within SystemX are pursuing the **3D monolithic integration of logic, memory and thermal management**, as well as energy-efficient and energy-harvesting design opportunities. This approach capitalizes on several recent nanotechnology breakthroughs, such as high-performance and energy-efficient **1D and 2D transistors**, **non-volatile memory** such as RRAM, 3D fine-grained interconnects, and **embedded cooling technologies** based on conduction and convection breakthroughs. The benefits of monolithically co-integrating logic and memory into 3D systems can be up to **1000-fold in terms of energy-delay-product (EDP)** benefits, especially for future abundant-data applications.

[1] M. Aly et al., *Computer* 48, 24-33 (2015)
[2] https://systemx.stanford.edu

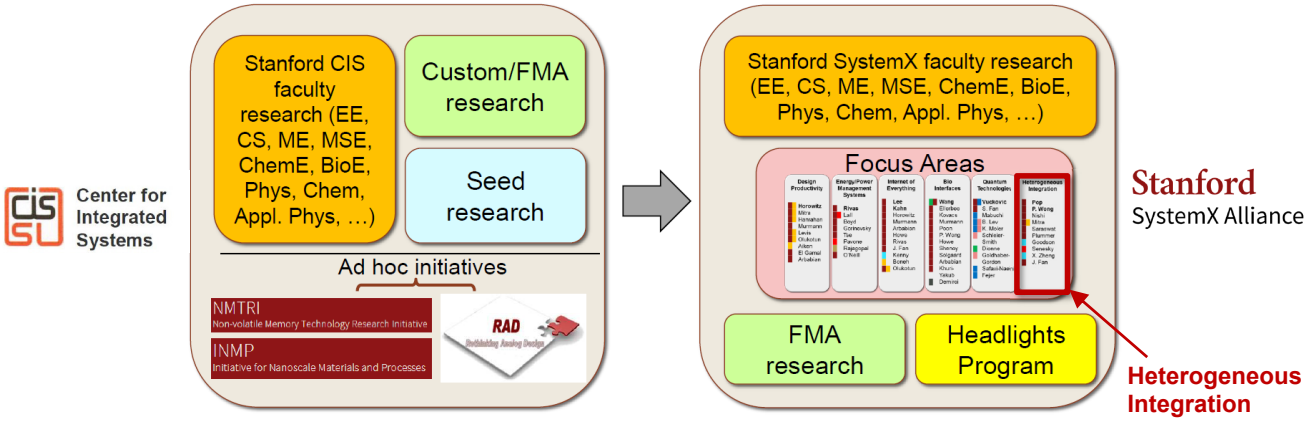
Most updated version of this short course at:
http://poplab.stanford.edu/pdfs/Pop_SSDM_N3XT_2018.pdf

Outline

- **Introduction and SystemX @ Stanford**
- Today's Computing Roadblocks
- The N3XT Solution
- Nanosystems Today
- Summary and Key Takeaways

What Is the SystemX Alliance at Stanford?

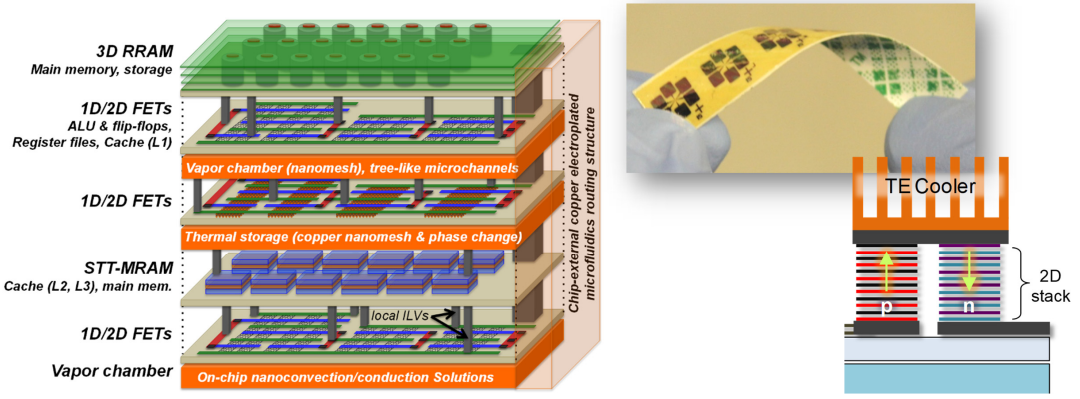
- Center for Integrated Systems (CIS) → SystemX Alliance
  - **Industry-academic Alliance to repositioned for 21st century research**
  - **"The hub" for electronics research at Stanford**



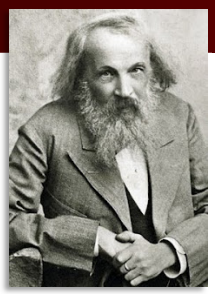
- What's New?
  - Stronger emphasis of top-down, **systems research**
  - Introduce **focus areas** to create coherent thrusts
  - Additional **sponsor benefits** including workshops, E-Seminars

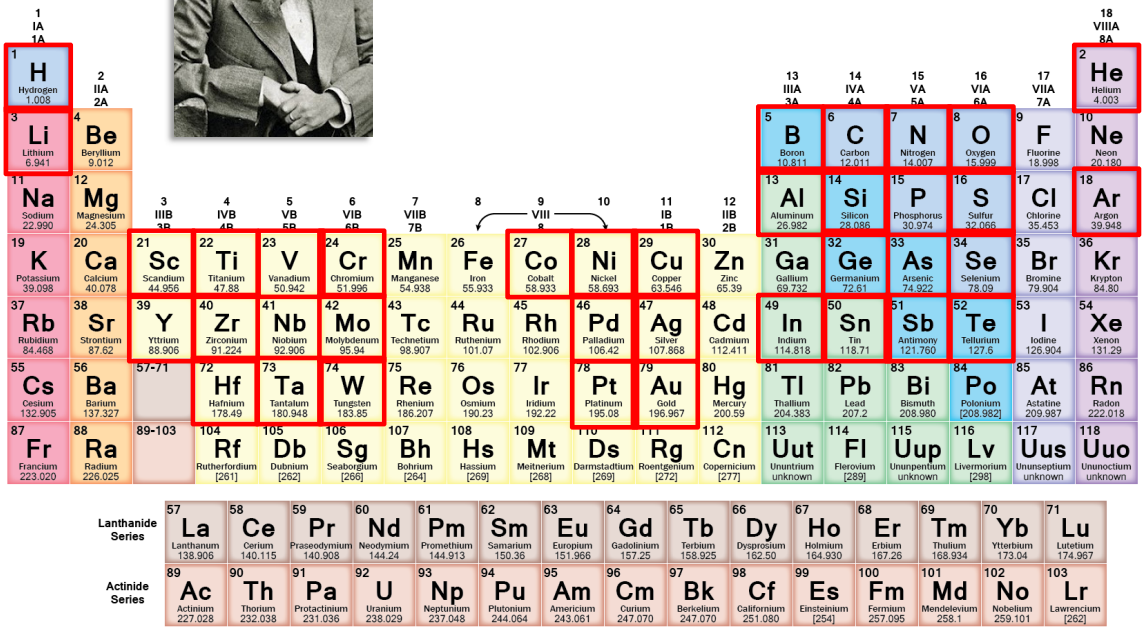
Heterogeneous Integration in SystemX

- **Heterogeneous Integration (HI) focus area in SystemX**
- Integration of "beyond-Si" platforms for "beyond Moore" applications
  - **3D integration** of logic, memory, sensors, thermal management, flexible substrates…
  - **Energy-efficient** design and energy-harvesting opportunities



- Core faculty: E. Pop (lead), H.-S.P. Wong, J. Fan, K. Goodson, S. Mitra, Y. Nishi, J. Plummer, K. Saraswat, D. Senesky, X. Zheng + Z. Bao

From…  Mendeleev



To today…

Outline

- Introduction and SystemX @ Stanford
- **Today's Computing Roadblocks**
- The N3XT Solution
- Nanosystems Today
- Summary and Key Takeaways

21st Century Computing Workload

Large amounts of loosely-structured data…

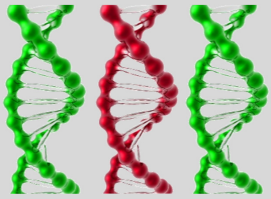
- Streaming video/audio
- Natural languages
- Real-time sensing
- Contextual environment

(IBM Watson, *Jeopardy!* champion)

200 kiloWatts

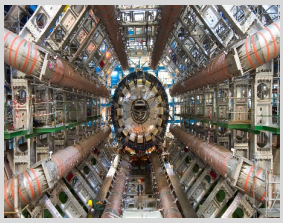
Abundant-Data Applications

Genomics



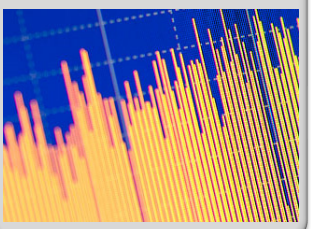
Smart Cities

Science



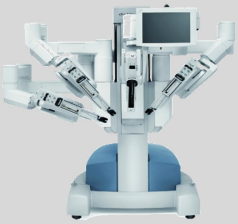
Retail

Finance



Security

Health Care

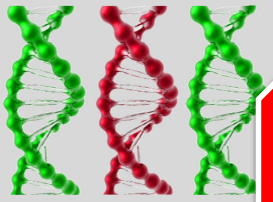


Government

source: H.-S.P. Wong

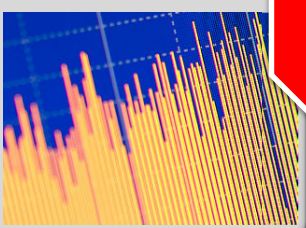
Abundant-Data Applications

Genomics



Retail

Finance



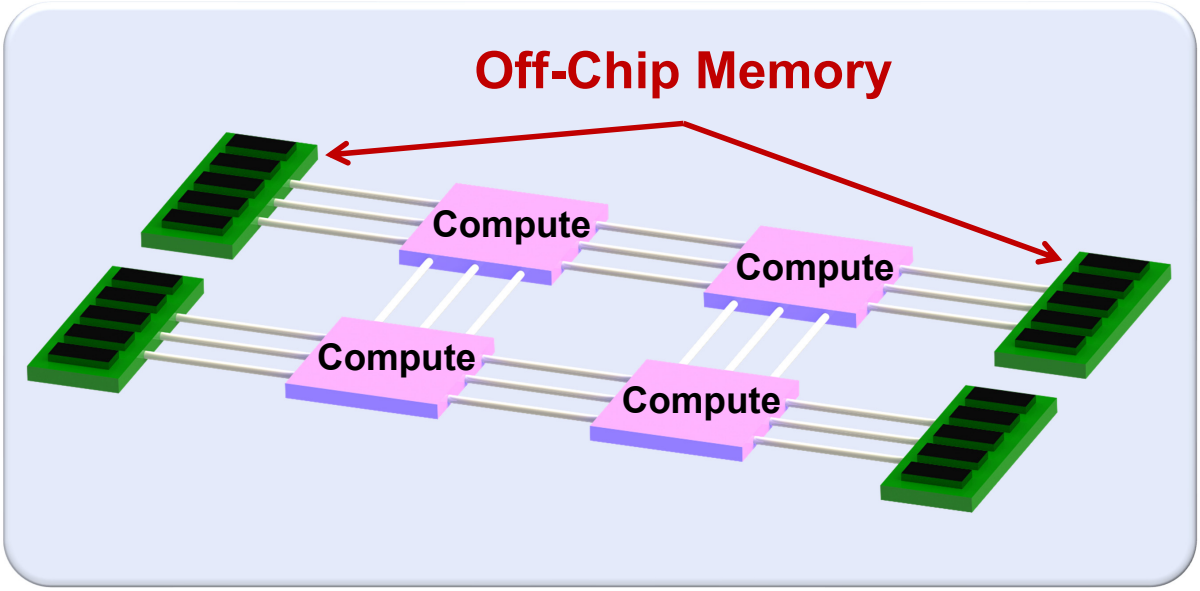
Government

Computational demands *exceed* Processing capability

cannot pursue 21st century applications with 20th century hardware

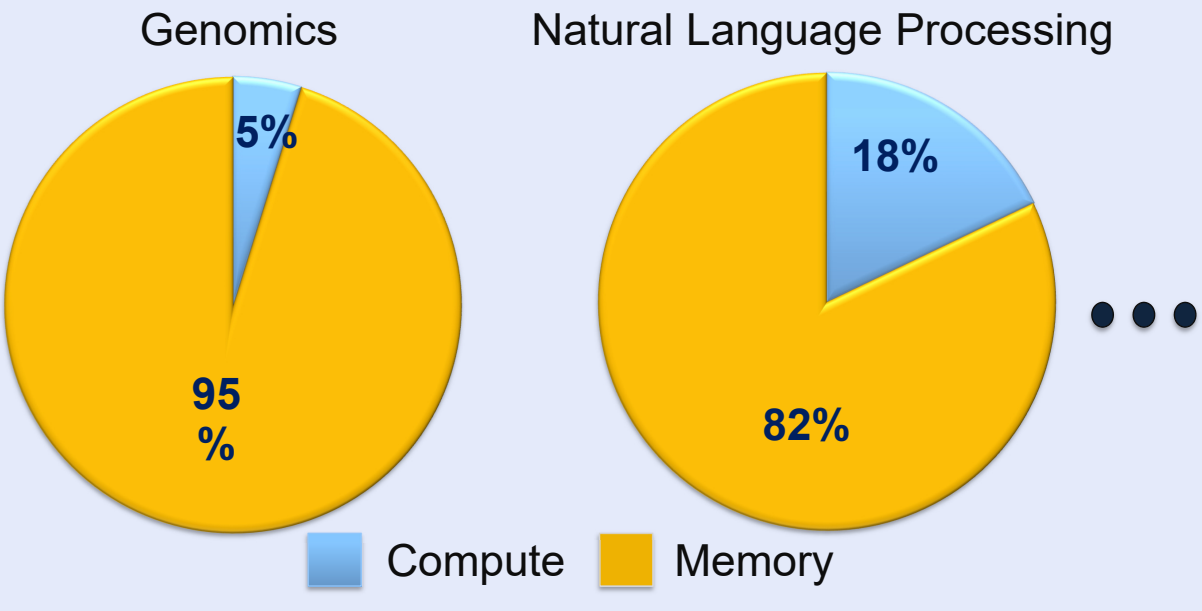
Identifying the Computing Bottleneck TODAY

- **Separate** compute and memory chips
- Large on-chip buffers: **expensive** and **inadequate**



Memory Wall

Abundant-data applications: Energy measurements



Intel performance counter monitors 2 CPUs, 8-cores/CPU + 128 GB DRAM

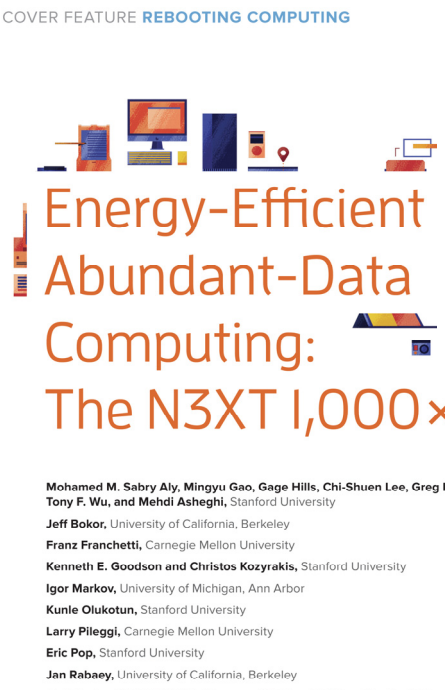
Source: M. Aly, S. Mitra

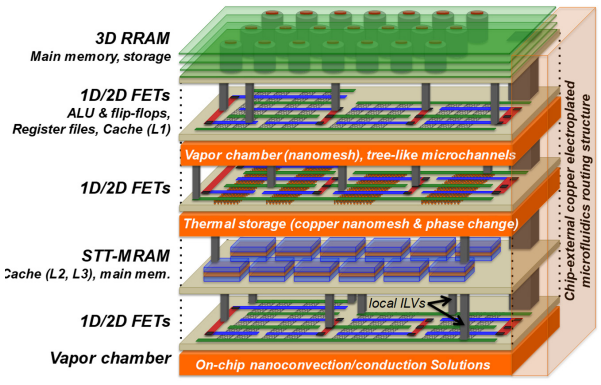
Outline

- Introduction and SystemX @ Stanford
- Today's Computing Roadblocks
- **The N3XT Solution**
- Nanosystems Today
- Summary and Key Takeaways

N3XT Solution: 3D Monolithic Integration

M. Aly, […], K. Goodson, C. Kozyrakis, E. Pop, J. Rabaey, C. Re, H.-S.P. Wong, S. Mitra, *IEEE Computer* (2015)



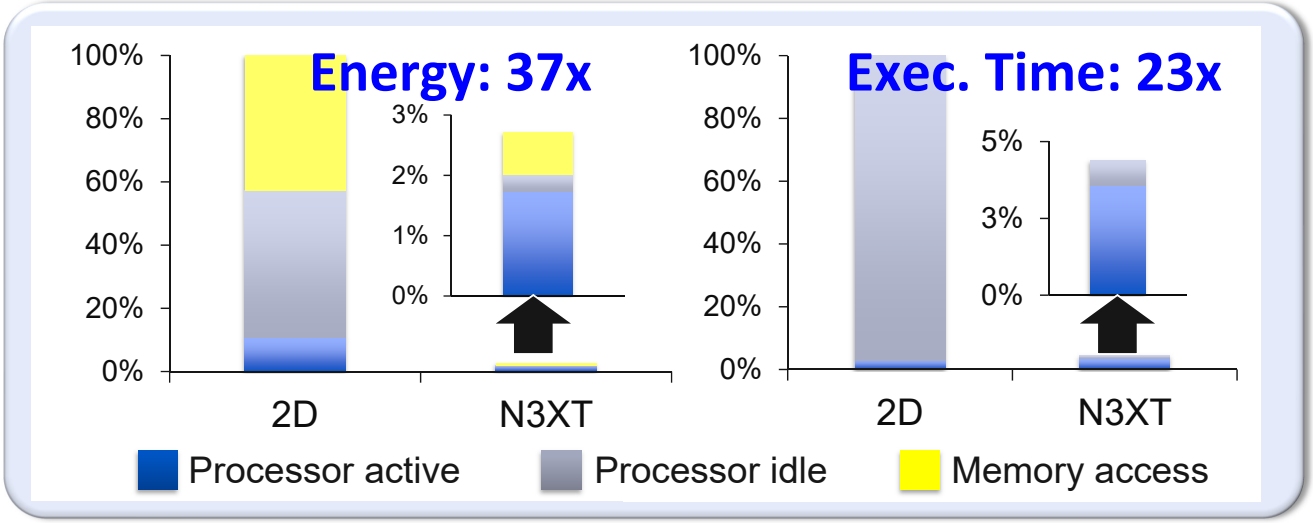


Next-generation information technologies will process unprecedented amounts of loosely structured data that overwhelm existing computing systems. N3XT improves the energy efficiency of abundant-data applications 1,000-fold by using new logic and memory technologies, 3D integration with fine-grained connectivity, and new architectures for computation immersed in memory.

What Are the Benefits?

M. Aly, […], K. Goodson, C. Kozyrakis, E. Pop, J. Rabaey, C. Re, H.-S.P. Wong, S. Mitra, *IEEE Computer* (2015)

PageRank Application



Energy x Exec. Time = 851x benefits

Source: M. Aly, S. Mitra (Stanford)

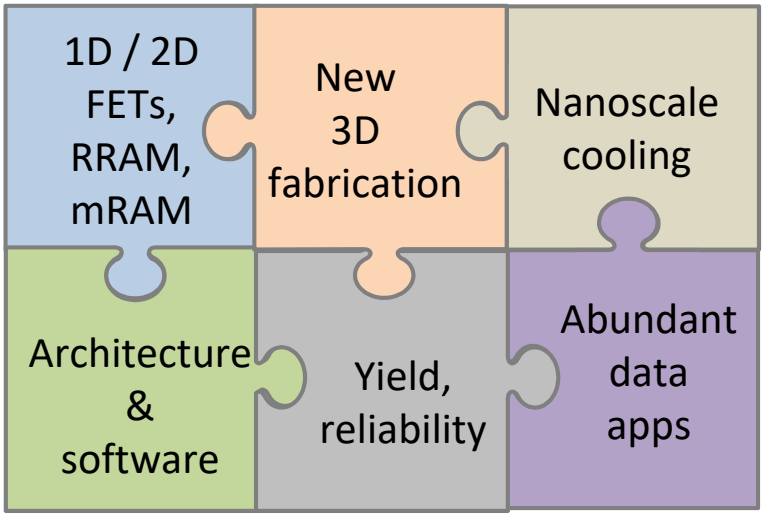
Note: N3XT has CNFETs, STT-RAM cache, RRAM memory

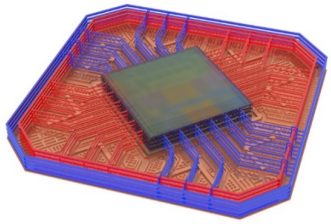
N3XT: End-to-End Approach

COVER FEATURE **REBOOTING COMPUTING**

Energy-Efficient Abundant-Data Computing: The N3XT 1,000×

Mohamed M. Sabry Aly, Mingyu Gao, Gage Hills, Chi-Shuen Lee, Greg Pitner, Max M. Shulaker, Tony F. Wu, and Mehdi Asheghi, Stanford University
Jeff Bokor, University of California, Berkeley
Franz Franchetti, Carnegie Mellon University
Kenneth E. Goodson and Christos Kozyrakis, Stanford University
Igor Markov, University of Michigan, Ann Arbor
Kunle Olukotun, Stanford University
Larry Pileggi, Carnegie Mellon University
Eric Pop, Stanford University
Jan Rabaey, University of California, Berkeley
Christopher Ré, H.-S. Philip Wong, and Subhasish Mitra, Stanford University





Monolithic Thermal Platform (Goodson)

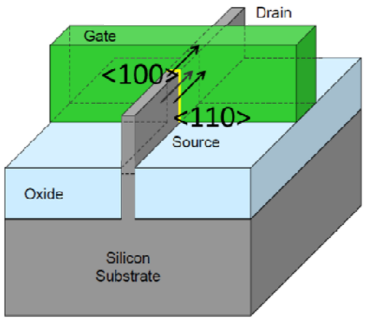
So What Is Needed?

Transistor Scaling Limits

$L_G \sim (t_{ch} t_{ox})^{1/2}$

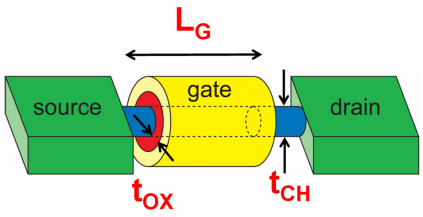
- Gate lengths L_G scale with channel thickness t_{ch}
- Can be made of 3D materials like Si, but…

2D



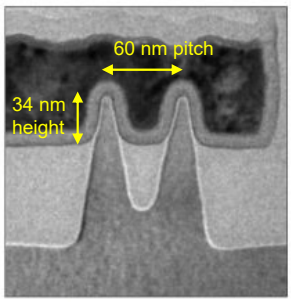
"Fin"FET or tri-gate

1D

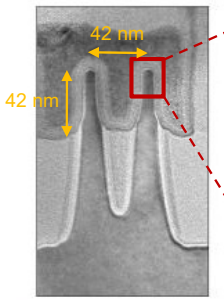


wrap-gate nanowire or nanotube

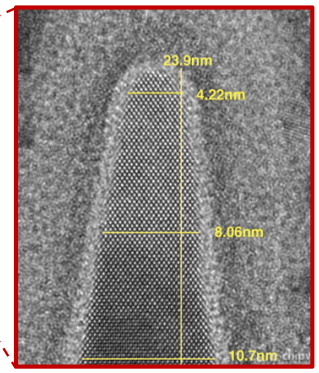
50 Years of Scaling: State-of-the-Art FinFETs



22 nm 1st Generation Tri-gate Transistor



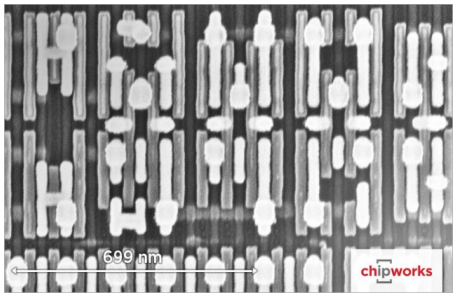
14 nm 2nd Generation Tri-gate Transistor



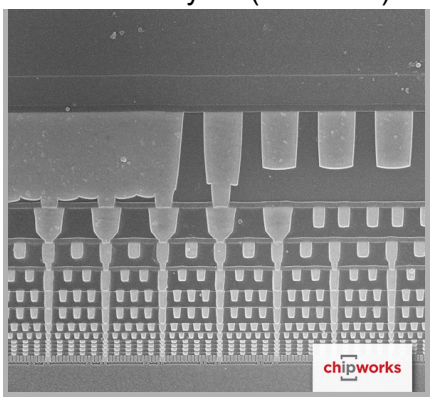
tapered fin
average width = 8 nm

	22 nm Node	14 nm Node
Fin Pitch	60	42
Gate Pitch	90	70

Intel 14 nm SRAM (top view)



13 metal layers (side view)

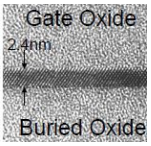


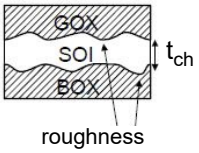
source: Intel and ChipWorks

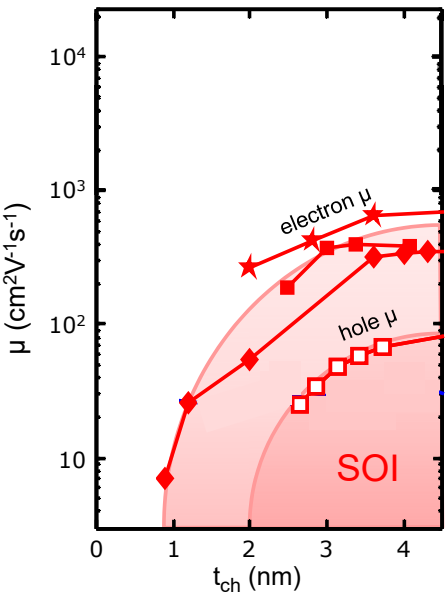
Transistor Scaling: Atomically Thin Materials

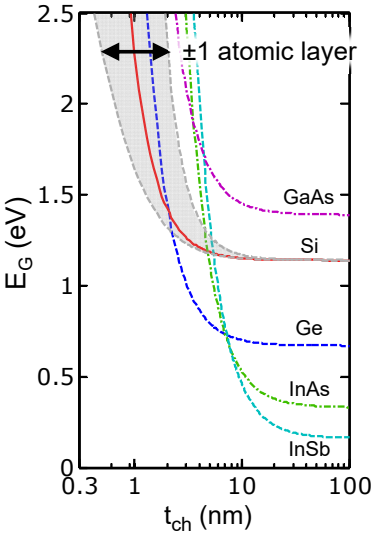
- **Problem:** 20th century transistors "carved" out of **3D** materials (Si) → surface roughness restricts mobility, band gap, variability

3D





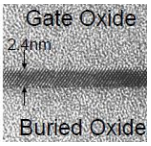


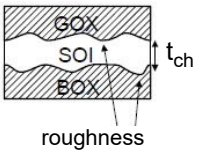


Transistor Scaling: Atomically Thin Materials

- **Problem:** 20th century transistors "carved" out of **3D** materials (Si) → surface roughness restricts mobility, band gap, variability

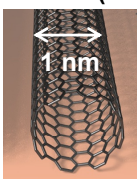
3D



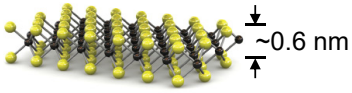


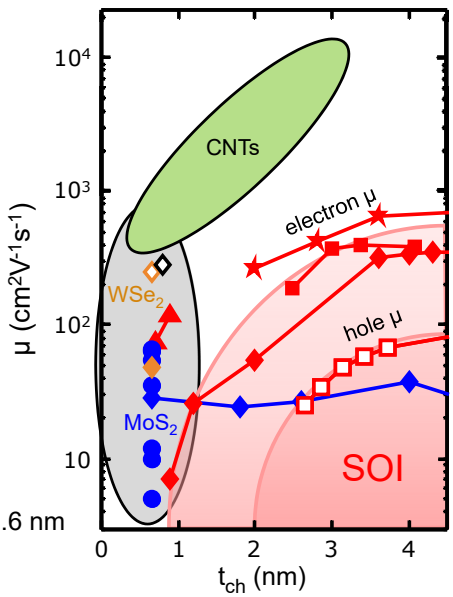
- **Solution:** 21st century atomically thin materials

1D carbon nanotube (CNT)

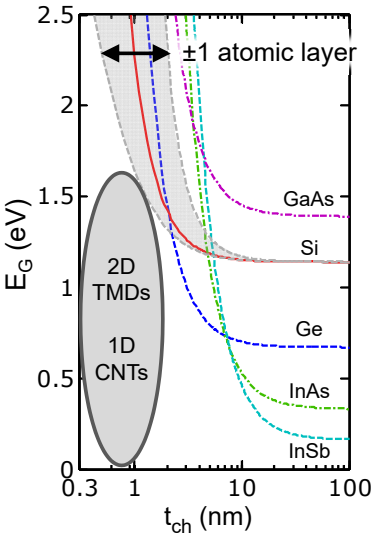


2D TMD (MoS_2, WTe_2, $ZrSe_2$)



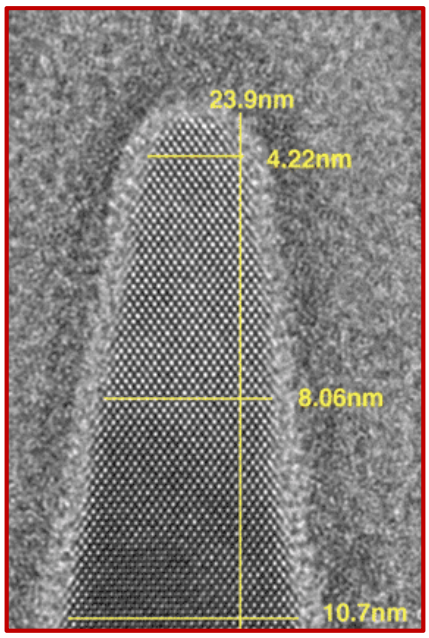


C. English, E. Pop, et al., *Nano Lett.* 16, 3824 (2016)



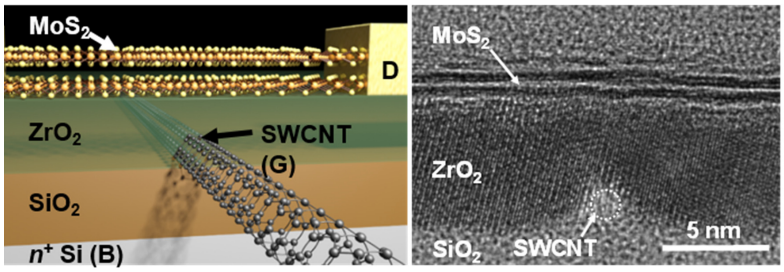
Modern FinFET vs. Atomically Thin FETs

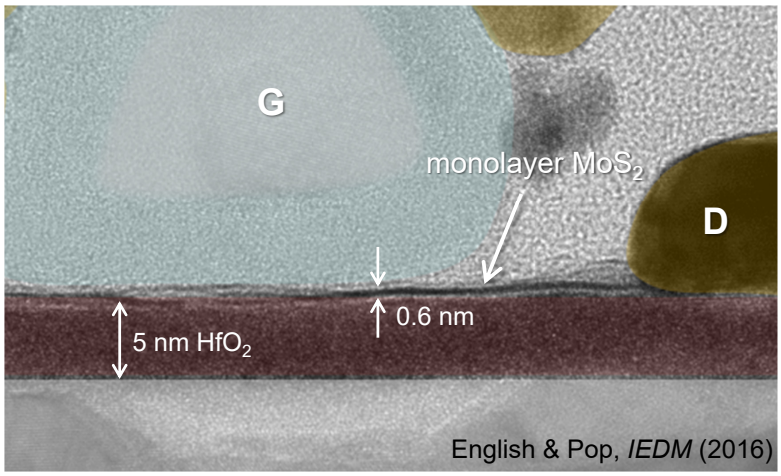
(note same scale)



Intel "14 nm" FinFET

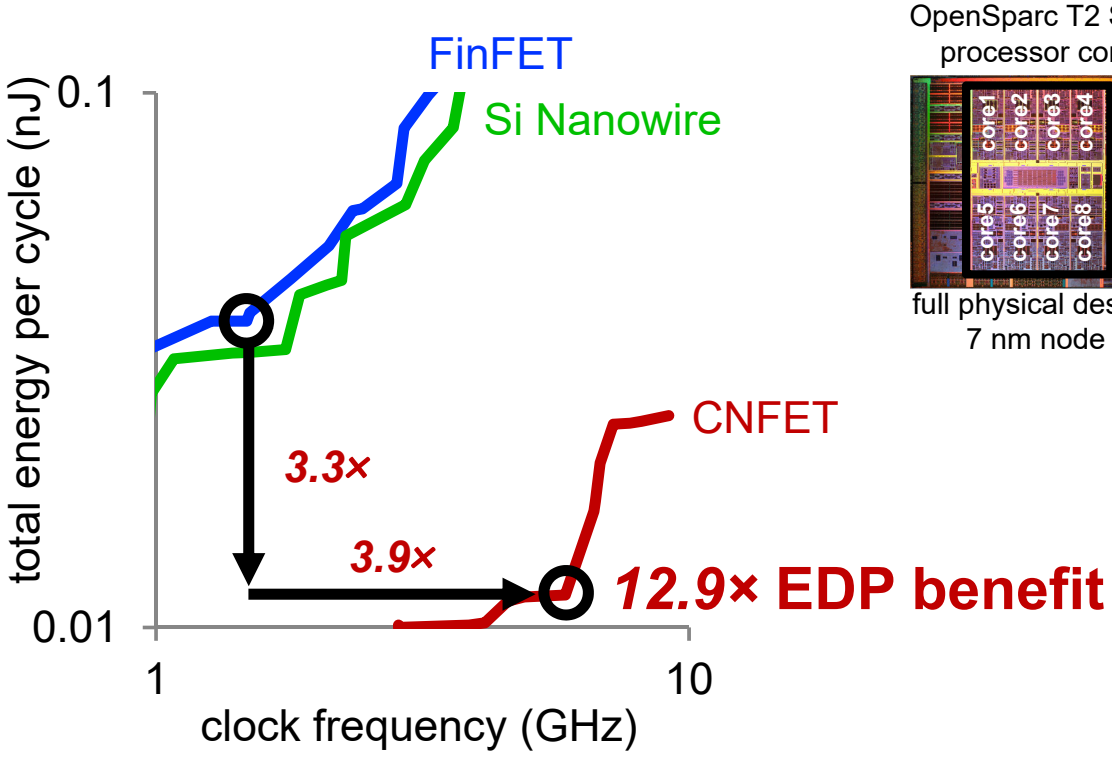
Desai, Wong & Javey, *Science* (2016)





English & Pop, *IEDM* (2016)

Benefits of Carbon Nanotube FETs



OpenSparc T2 SoC
processor core

full physical design
7 nm node

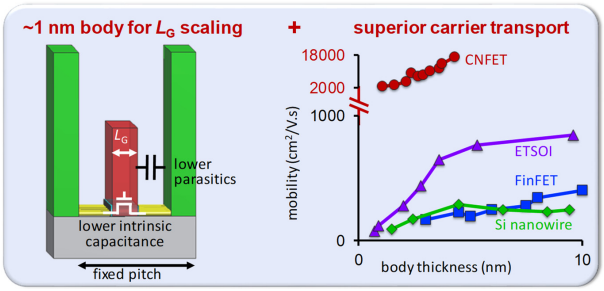
Source: G. Hills, S. Mitra (Stanford) + IMEC

Where Do Benefits Come From?

- Simultaneous 1D ultra-thin body (1 nm) + high drive current
- Challenging for bulk (3D) materials!

$$EDP \sim \underbrace{CV^2}_{\text{energy}} * \underbrace{CV/I}_{\text{delay}}$$

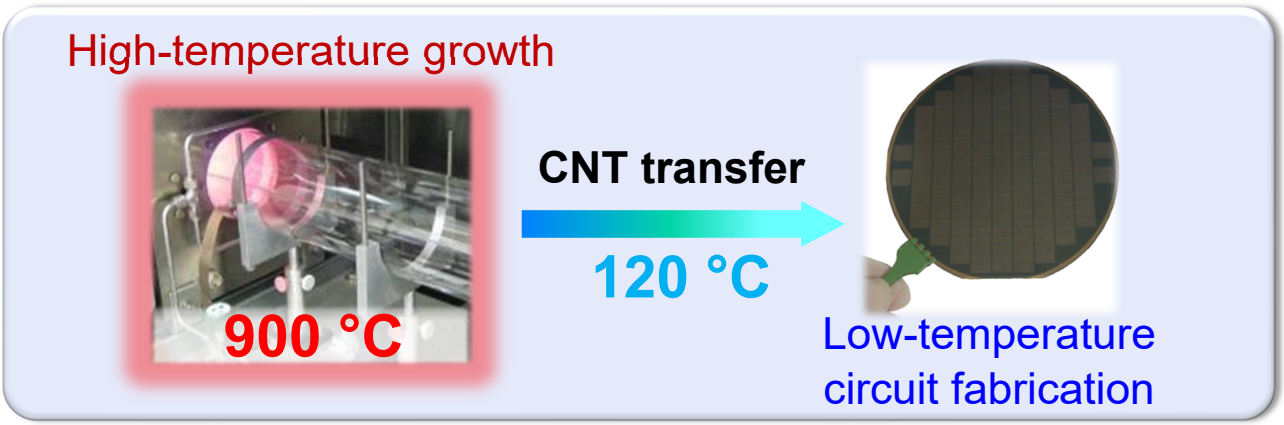
- 20% lower V_{DD} ~2x benefit
- 25% higher I_D ~1.25x benefit
- 2x lower capacitance ~4x benefit
  - Shorter L_G
  - Reduced parasitics
  - Smaller FET widths required to meet timing

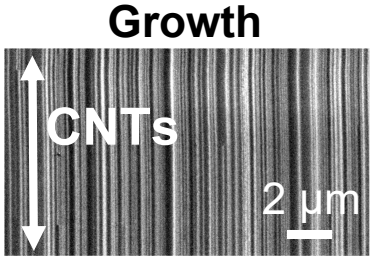


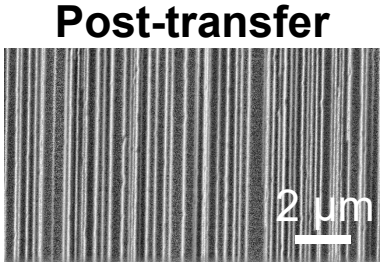
Source: G. Hills, S. Mitra

1D and 2D Materials Enable Monolithic 3D

- Material layer transfer decouples high temperature growth



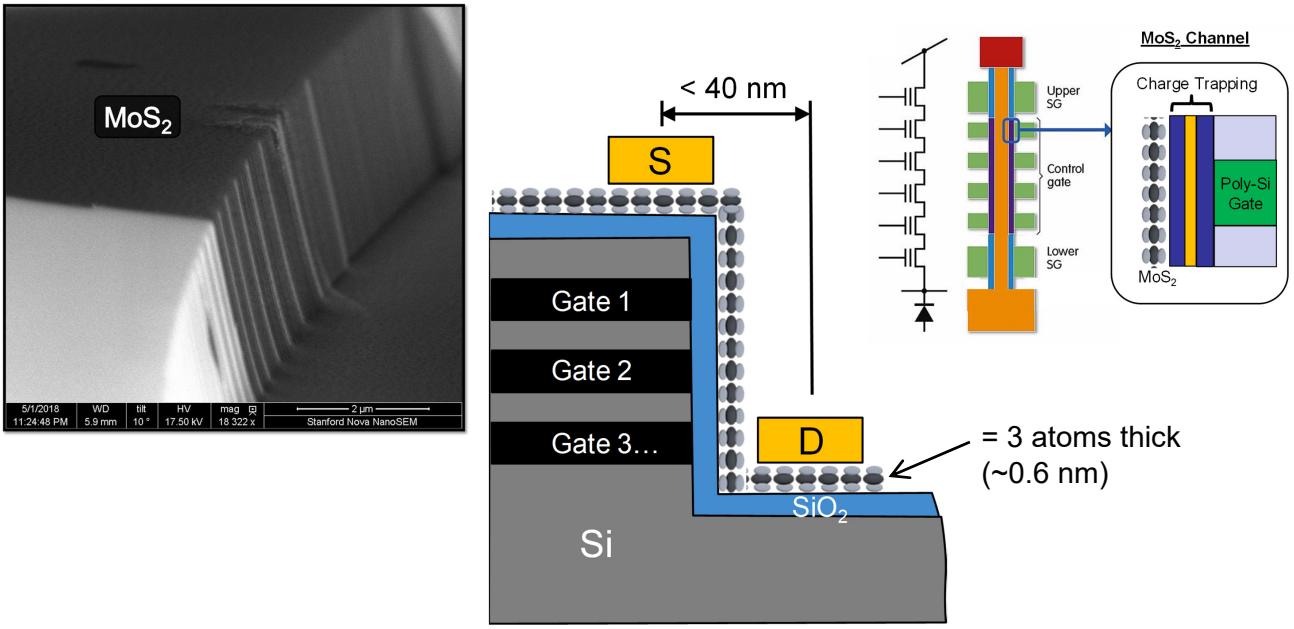




Vertical 2D Material Growth & Integration

Work of C. McClellan and A. Yu

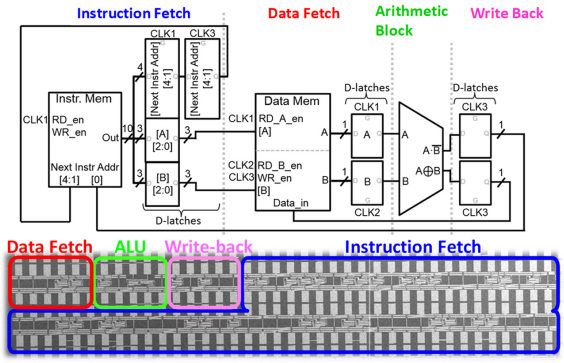
- CVD growth conformal over 3 to 6 μm steps, at 700°C (for now)
- Potential applications for vertical transistors, e.g. high density memory



Outline

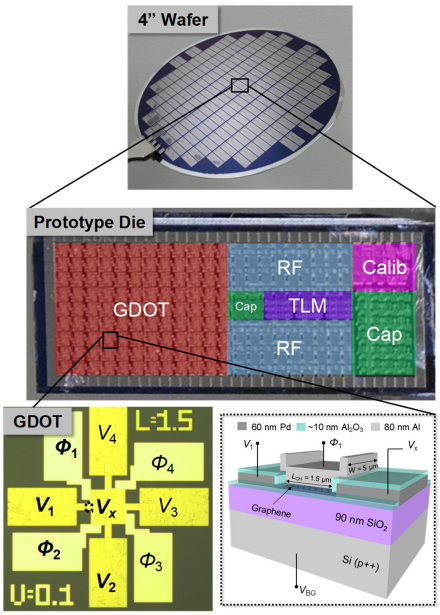
- Introduction and SystemX @ Stanford
- Today's Computing Roadblocks
- The N3XT Solution
- **Nanosystems Today**
- Summary and Key Takeaways

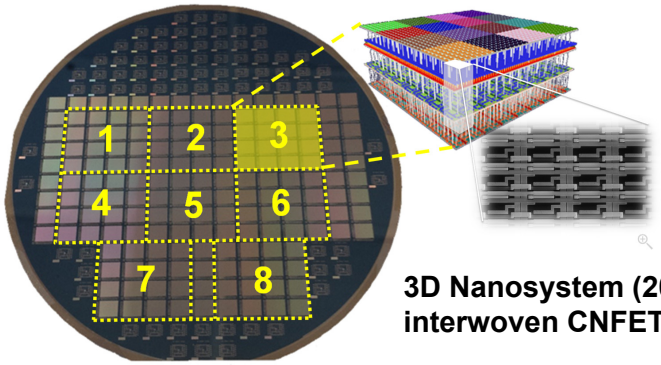
3D Nanosystems Can Be Built TODAY…



CNT Computer (2013)

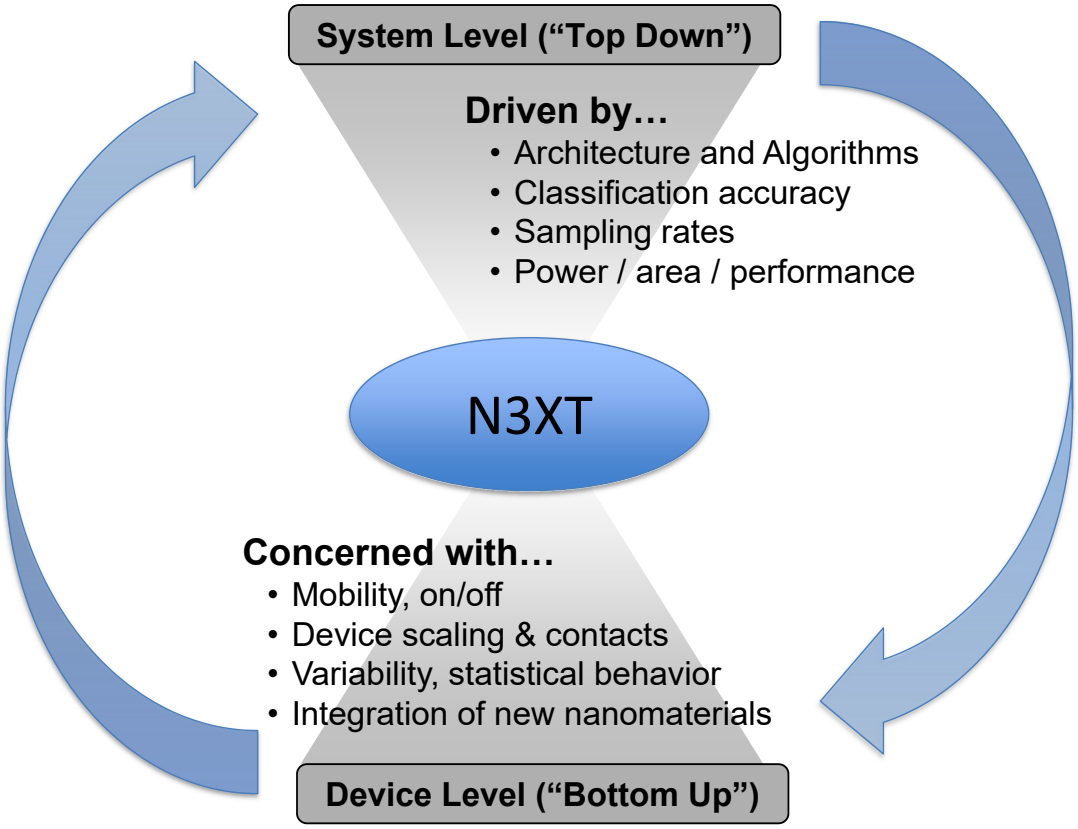
Graphene Dot Product (2016)





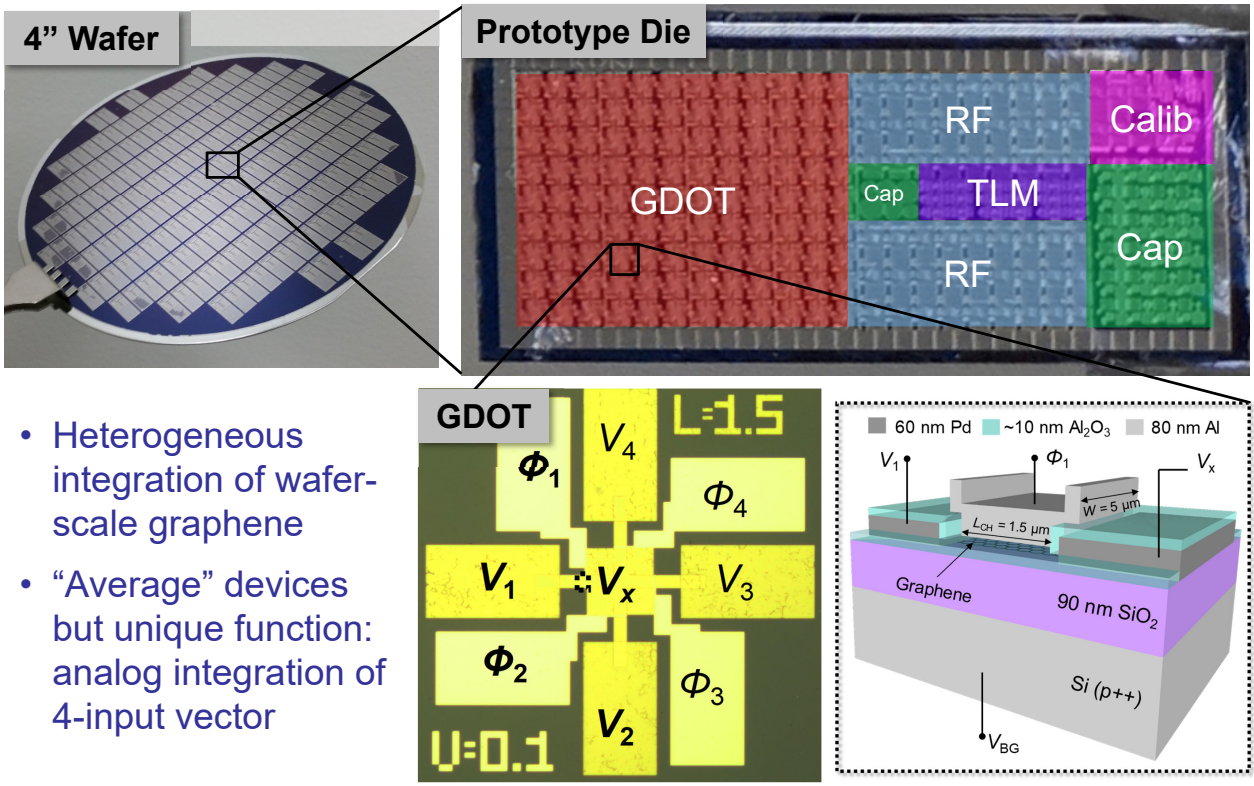
3D Nanosystem (2017)
interwoven CNFETs + memory + sensing

Top-Down and Bottom-Up Approach



Ex: Large-Scale Graphene Nanosystems

N. Wang, S. Gonugondla, I. Nahlus, N. Shanbhag, E. Pop, *VLSI Symp.* (2016)

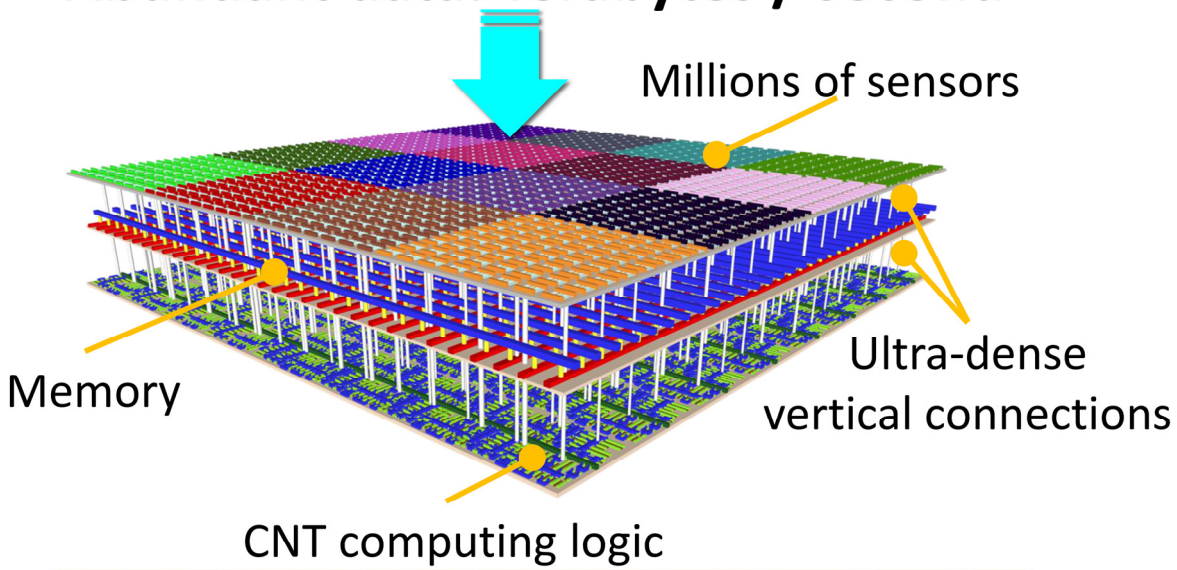


- Heterogeneous integration of wafer-scale graphene
- "Average" devices but unique function: analog integration of 4-input vector

Ex: 3D Monolithic Computing & Data Storage

M. Shulaker, G. Hills […], H.-S.P. Wong, S. Mitra, *Nature* **547**, 74 (2017)

Abundant data: Terabytes / second

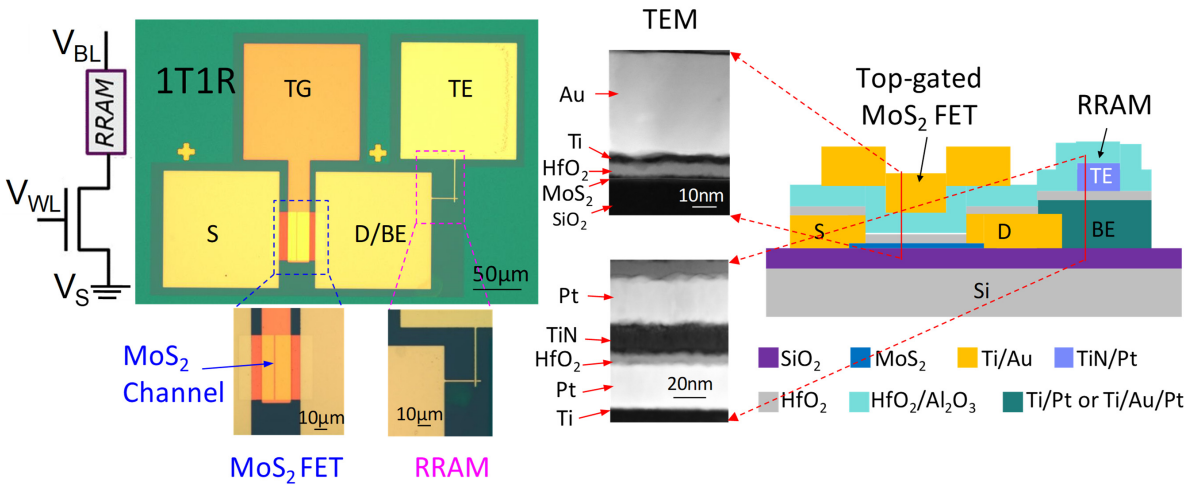


In-situ classification:
Extensive, accurate

Ex: 2D + RRAM Heterogeneous Integration

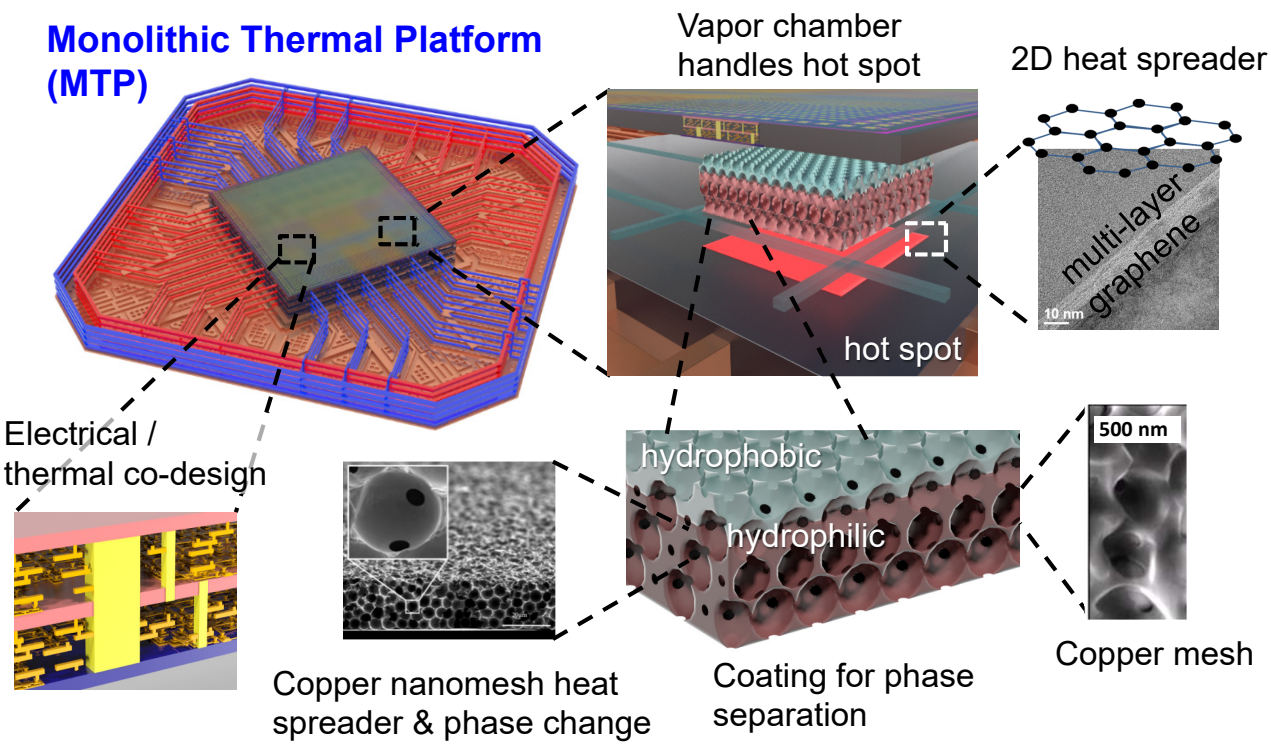
R. Yang, H. Li, K. Smithe, T. Kim, K. Okabe, E. Pop, J. Fan, H.-S.P. Wong, *IEDM* (2017)

- 1-transistor-1-resistor (1T1R) memory cells
- Monolayer CVD-MoS_2 FETs: low OFF-state current reduces leakage
- HfO_x oxygen-vacancy RRAM: low-T fabrication < 200 °C

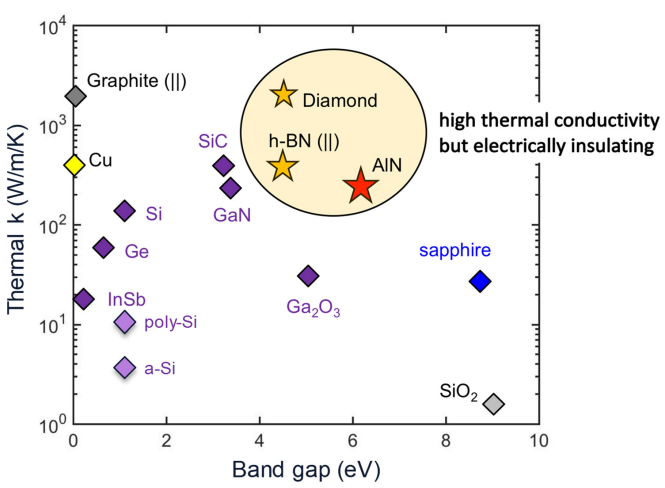


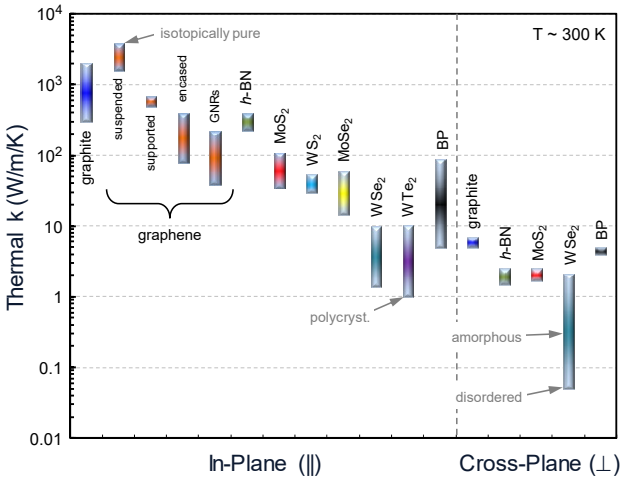
Ex: 3D Integration Thermal Management

K. Goodson group (Stanford)



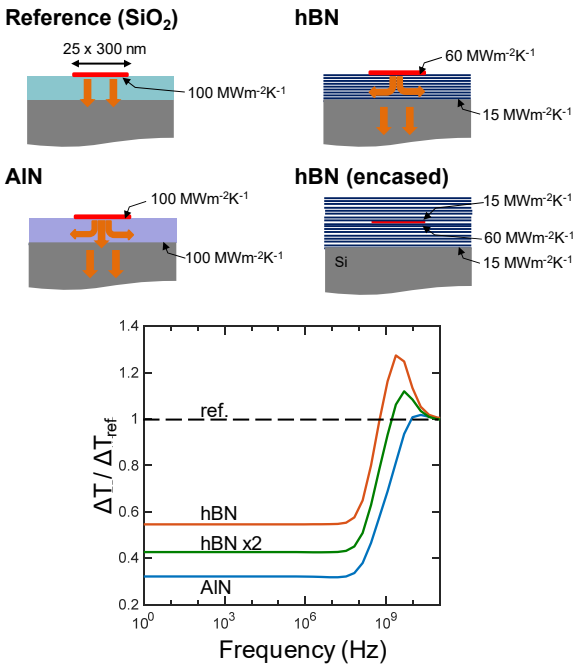
Ex: New Thermal (Nano)Materials

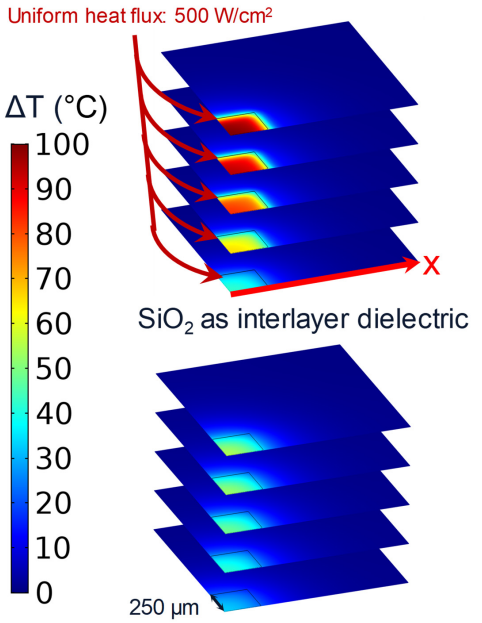




- AlN and h-BN are *insulators with large thermal conductivity*
- Could be grown at low temperature → what are quality & thermal properties?
- Can we leverage *layered material anisotropy*? (h-BN, graphene)

Ex: Built-In AlN and h-BN Heat Spreaders

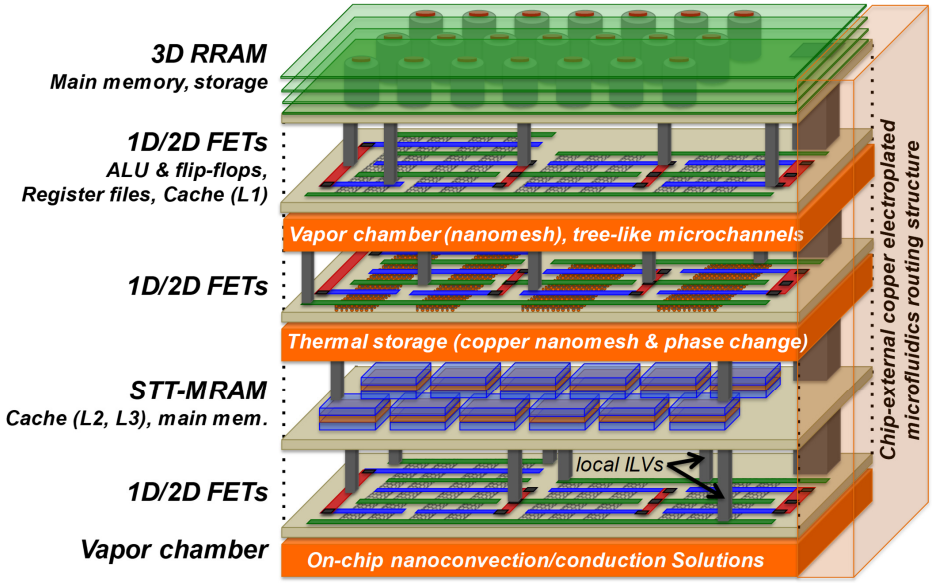




- Evaluating h-BN and AlN as heat spreading interlayer dielectrics
- Thermal resistance of interfaces becomes important at high frequencies

N3XT Solution: 3D Monolithic Integration

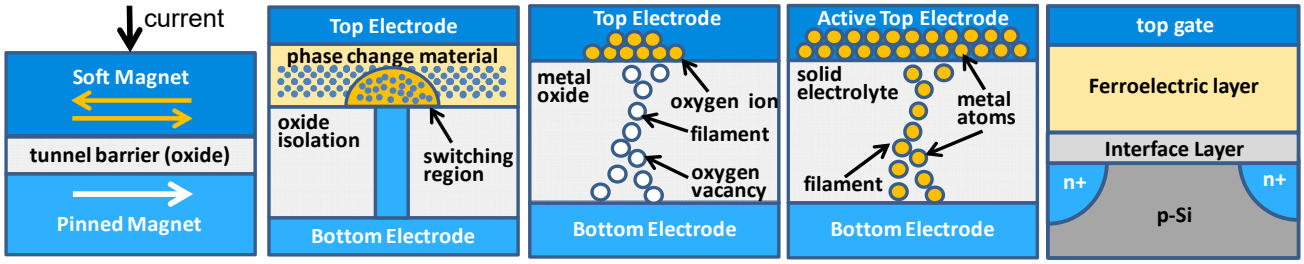
M. Aly, […], K. Goodson, C. Kozyrakis, E. Pop, J. Rabaey, C. Re, H.S.-P. Wong, S. Mitra, *IEEE Computer* (2015)



computation immersed in memory

"New" Memories

Random access, non-volatile, no erase before write, on-chip integration

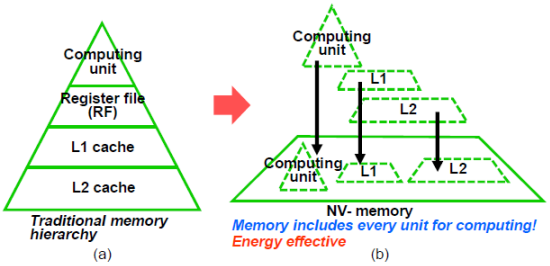


STT-MRAM	PCM	RRAM	CBRAM	FERAM
Spin torque transfer magnetic random access memory	Phase change memory	Resistive switching random access memory	Conductive bridge random access memory	Ferro-electric random access memory

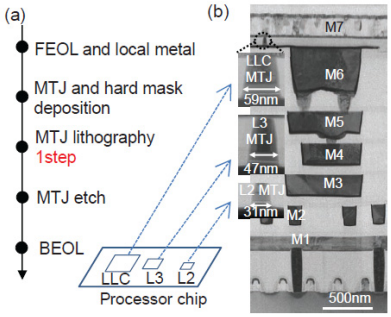
Source: H.-S.P. Wong (Stanford)

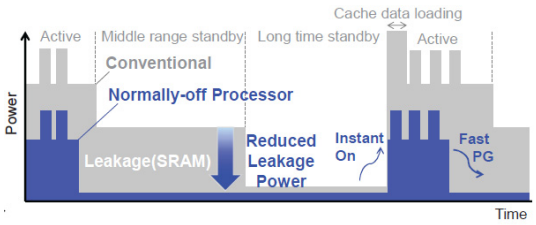
STT-MRAM Caching for Energy Reduction

- Much work at Toshiba, TSMC, SK Hynix, Everspin (GF)
- Computing immersed in (non-volatile) memory
- LLC based on STT-MRAM
- 75% energy reduction vs. SRAM-based L2 and L3 cache



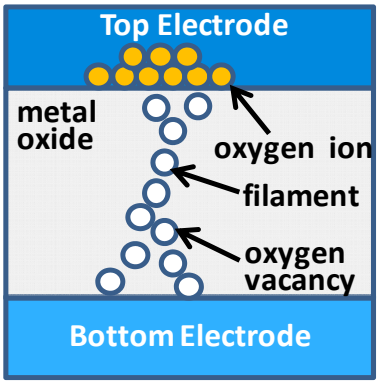
S. Fujita (Toshiba), IEDM 2015



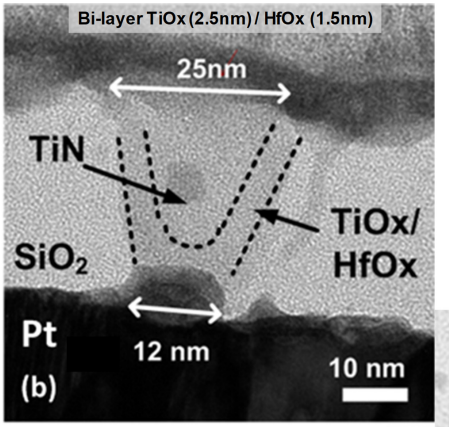


Conventional processor (SRAM, leakage) vs. normally-off processor (STT-MRAM, non-volatile).

RRAM: Resistive Random Access Memory

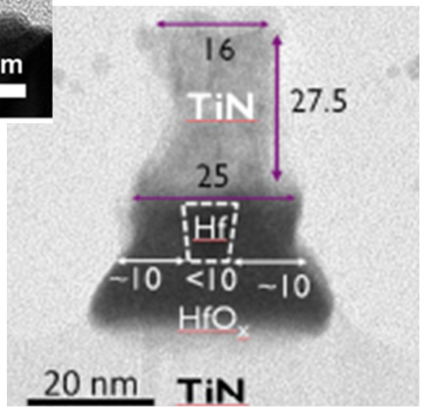


Scalable: 12 nm



Y. Wu, H. Yi, Z. Zhang, Z. Jiang, J. Sohn, S. Wong, H.-S. P. Wong, *IEDM* 2013. (Stanford)

Scalable: 10 nm



B.Govoreanu et al., *IEDM* 2011 (IMEC)

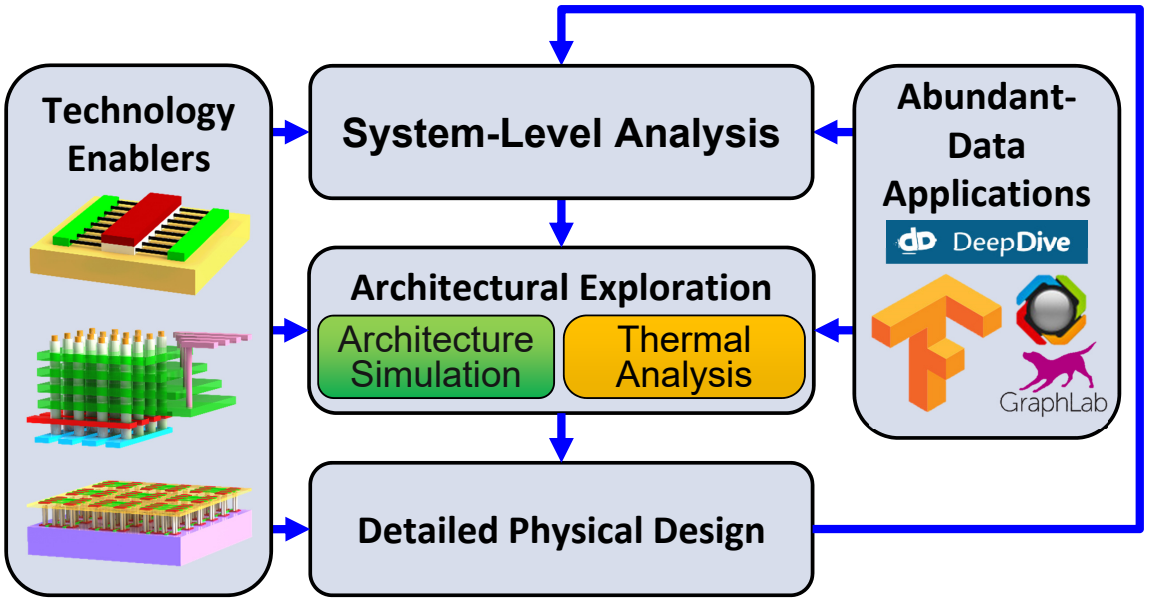
- Random access
- Non-volatile
- Scalable to sub-10 nm
- Low-T fabrication on CMOS

Outline

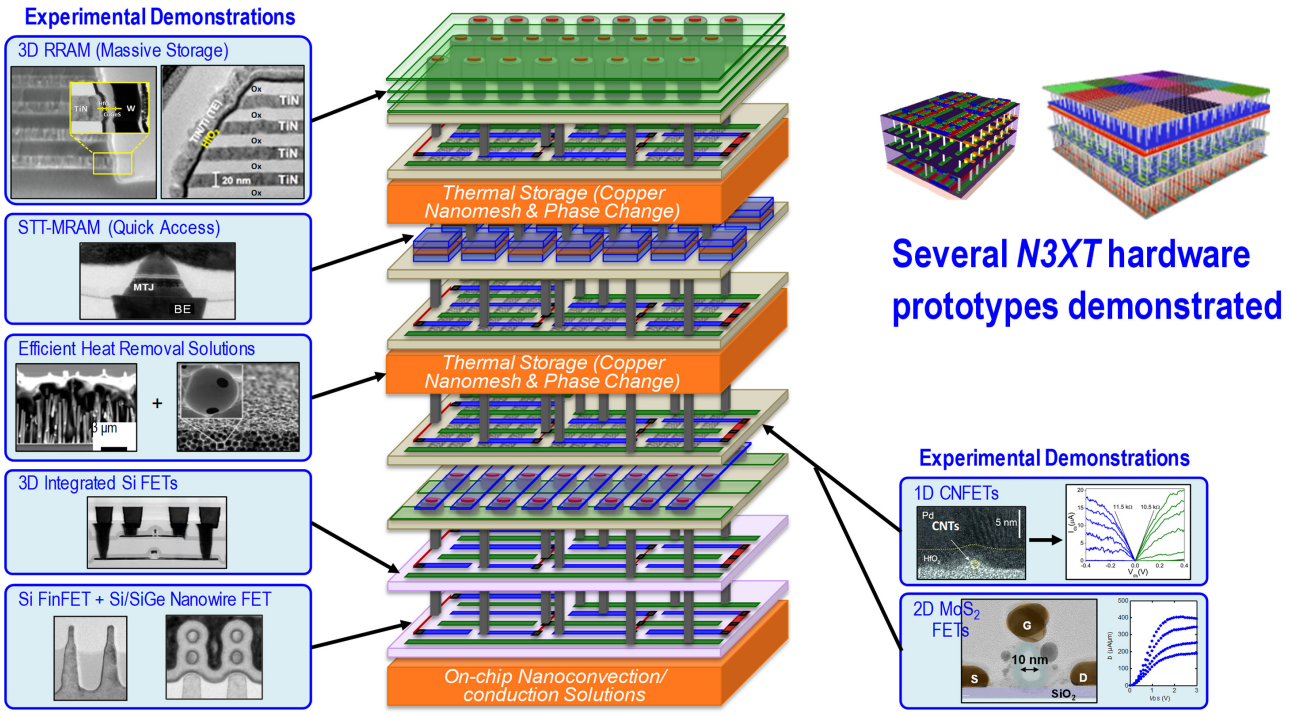
- Introduction and SystemX @ Stanford
- Today's Computing Roadblocks
- The N3XT Solution
- Nanosystems Today
- **Summary and Key Takeaways**

N3XT Exploration Framework

- Joint technology, design, and applications



N3XT: Computation Immersed in Memory

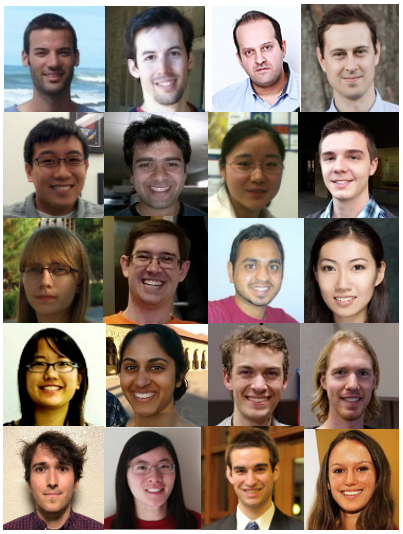


Key Takeaways

- N3XT: massive energy-efficiency benefits
  - Computation immersed in memory
  - 3D monolithic integration with fine-grained interconnects
- Isolated Improvements are NOT enough
  - Need improvements in logic, memory, thermal, …
  - Low-temperature (<450°C) integration
- N3XT 3D nanosystems demonstrated

Acknowledgements

http://poplab.stanford.edu



- Alumni:
  - Profs. D. Estrada, F. Xiong, S. Islam. Drs. A. Behnam, Z.-Y. Ong, A. Liao, Z. Li, V. Dorgan, E. Carrion, K. Grosse, M. Mleczko, C. English
- Post-docs:
  - Eilam Yalon, Miguel Munoz, Sam Vaziri, Kevin Brenner
- Grad students (40% women):
  - F. Lian, S. Deshmukh, S. Bohaichuk, K. Smithe, C. Neumann
  - I. Datye, M. Chen, A. Gabourie, V. Chen, K. Schauble, R. Grady
  - S. Suryavanshi, N. Wang, R.L. Xu, C. McClellan, C. Bailey, C. Koroglu
- Undergrads:
  - Andrew, Aria, Megan, Stephone, Tim, Job, Justin, Erin, Priyanka…
- Sponsors:
  - NSF, Army, Air Force, SRC, SystemX, Intel, HRL, NGC, SRI
- Collaborators:
  - Z. Bao, M. Brongersma, W. Cai, Y. Cui, R. Dauskardt, Y. Fisher, D. Goldhaber-Gordon, K. Goodson, T. Heinz, R. Howe, A. Lindenberg, N. Melosh, S. Mitra, Y. Nishi, K. Olukotun, P. Pianetta, E. Reed, A. Salleo, K. Saraswat, D. Senesky, Z.X. Shen, H.-S.P. Wong, X. Zheng (Stanford)
  - D. Cahill, W. King, J. Lyding, N. Shanbhag, M. Shim (UIUC), M. Rudan (Bologna), C. Jacoboni (Modena), D. Jena & G. Xing (Cornell), T. Grasser (TU Wien), D. Ielmini, R. Sordan (Milano), J. Shiomi (Tokyo), R. Wallace (UTD), S. Datta (Notre Dame), A. Newaz (SFSU), S. Koester (UMN), J. Moon (HRL), I. Karpov (Intel)

Acknowledgements: Funding

This work was supported in part by ASCENT, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA

STARnet

SONIC

SRC

ASCENT
Applications and Systems Driven Center for Energy-Efficient Integrated Nanotechnologies

NSF

NEEDS

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH · UNITED STATES AIR FORCE ·

Stanford | SystemX Alliance

SystemX ALLIANCE Stanford University

Stanford | Non-Volatile Memory Technology Research Initiative (NMTRI)