

# FUTURE OF SEMICONDUCTOR HARDWARE

## ASCENT

*Applications and Systems Driven Center on Energy Efficient  
Integrated Nanoelectronics*

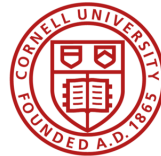
**Suman Datta**, Sayeef Salahuddin, Muhannad Bakir, Jeffrey Bokor, Kyeongjae Cho, Supriyo Datta, Patrick Fay, Steve George, Ken Goodson, Adam Hock, Sharon Hu, Subramanian Iyer, Debdeep Jena, Siddharth Joshi, Asif Khan, Andy Kummel, Umesh Mishra, Azad Naeemi, Michael Niemier, Eric Pop, Ramesh Ramamoorthy, Dan Ralph, Arijit Raychowdhury, Darrell Schlom, Madhavan Swaminathan, Jianping Wang, Shan Wang, Chuck Winter, Peide Ye, and Shimeng Yu

UC San Diego



UCLA

ILLINOIS INSTITUTE  
OF TECHNOLOGY



Georgia  
Tech



Stanford  
University



PURDUE  
UNIVERSITY.



WAYNE STATE  
UNIVERSITY

UNIVERSITY OF  
NOTRE DAME



UT DALLAS



UC SANTA BARBARA

Berkeley  
UNIVERSITY OF CALIFORNIA

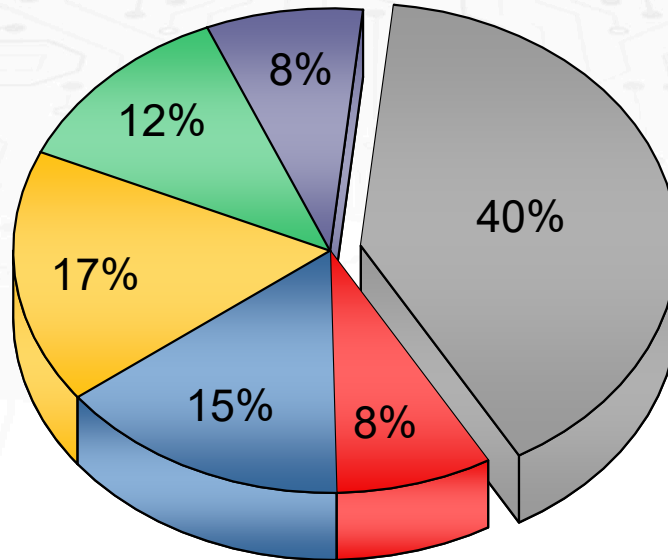
ASCENT

Applications and Systems Driven Center for  
Energy-Efficient Integrated Nanotechnologies

# Compute Performance Gain Comes from

- Compilers
- Power Management
- Microarchitecture

- ❖ Software Advancement
- ❖ Integration of system components
- ❖ Architectural Efficiency



Process Technology

- ❖ Higher performance, denser logic transistors
- ❖ Back-end-of-line RC reduction

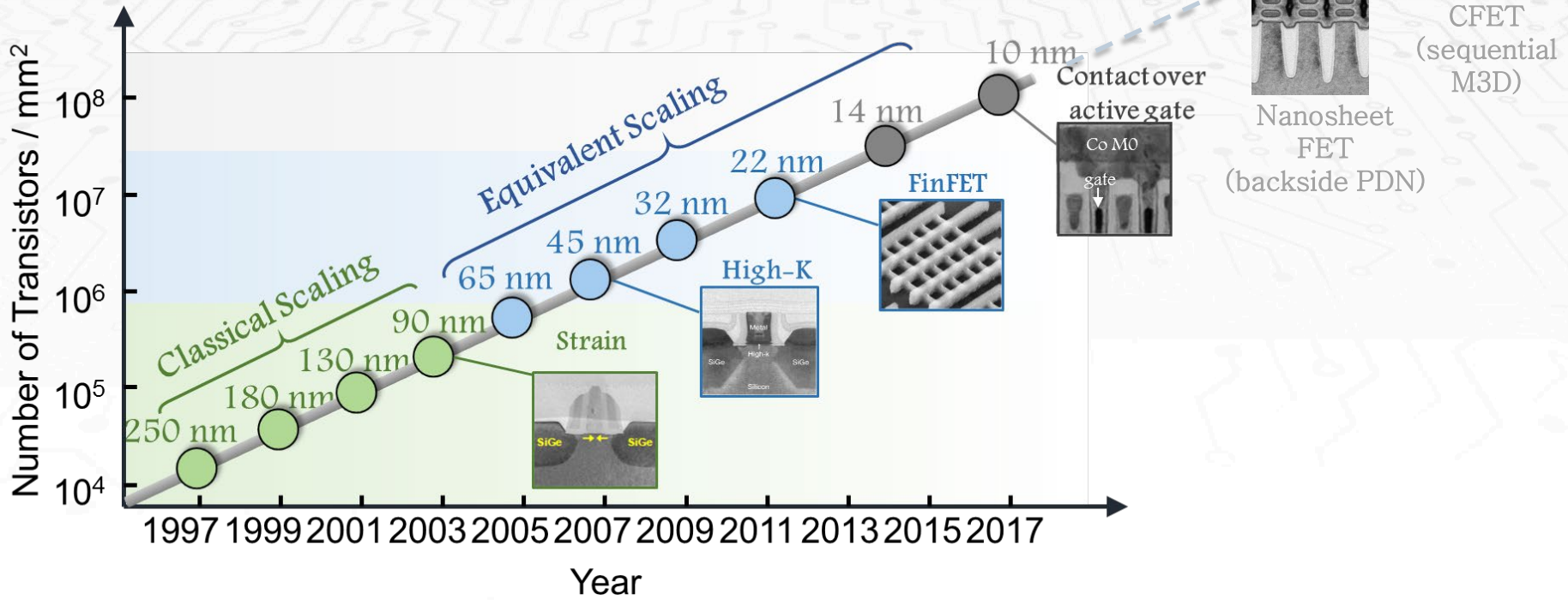
Additional Die Size  
Additional TDP

- ❖ Bigger Die
- ❖ Increased thermal design point

60% of compute performance gain over last decade came from process technology advancement

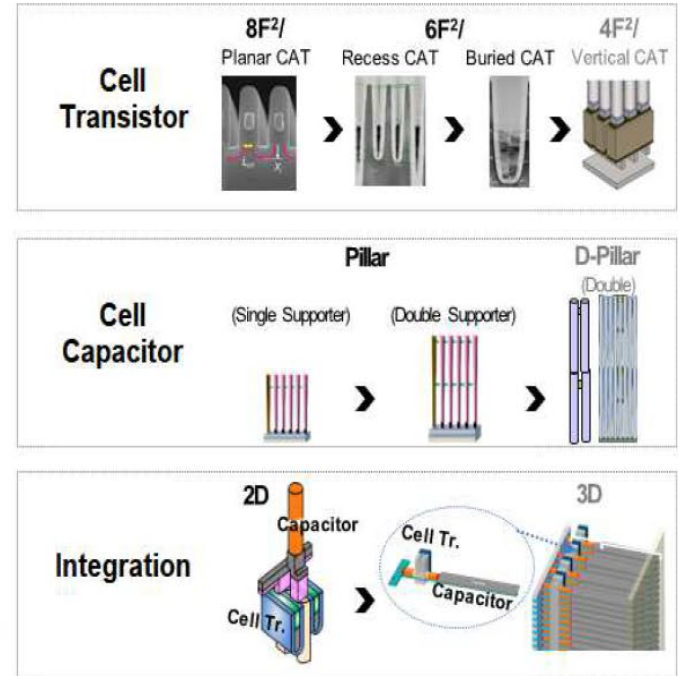
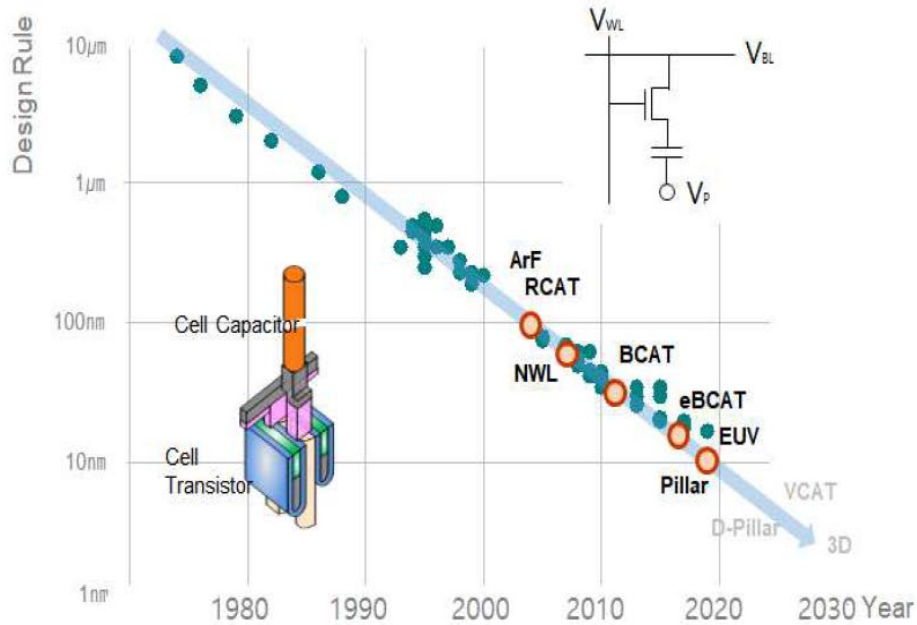
Source: Lisa Su (AMD), ERI DARPA Summit, 2019

# Logic Scaling



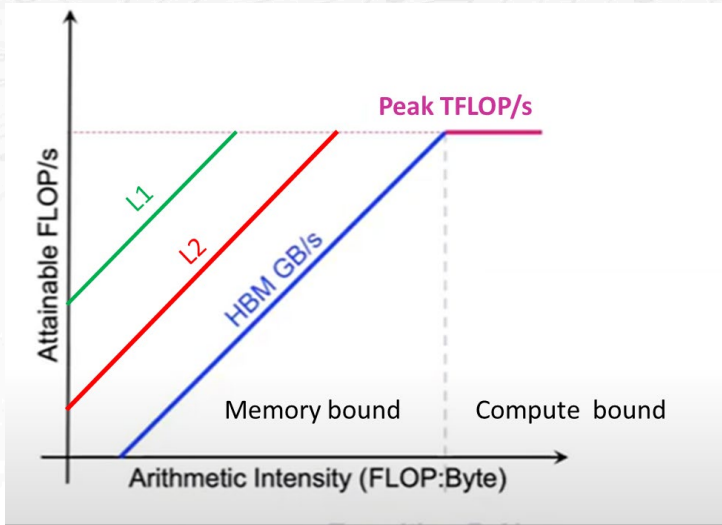
Scaling continues enabled by innovation in materials, devices, process integration & DTCO

# Memory Scaling

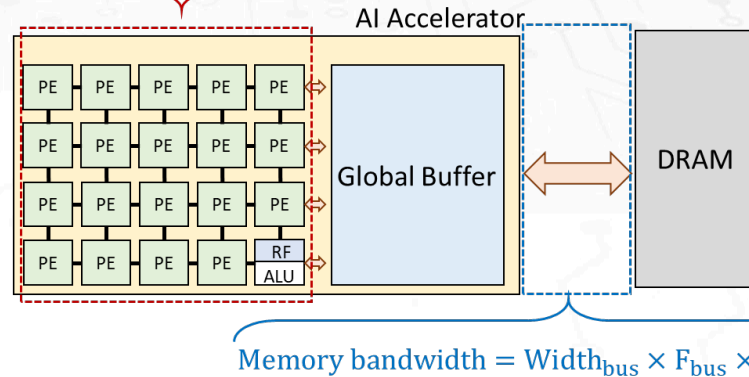


Scaling continues enabled by devices, structures and process integration innovation

# Roofline Performance



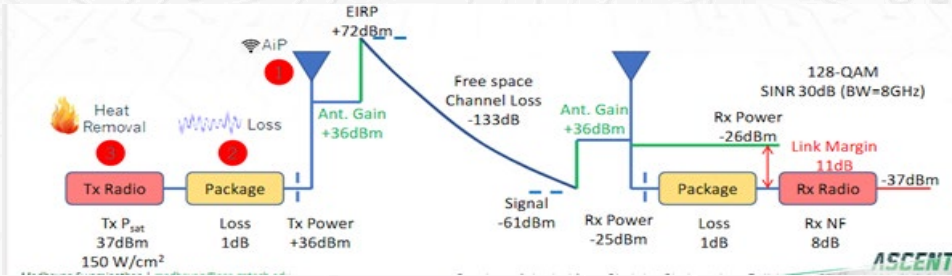
$$\text{Peak throughput} = N_{\text{cores}} \times F_{\text{cores}} \times \frac{\text{OP}}{\text{cycle}}$$



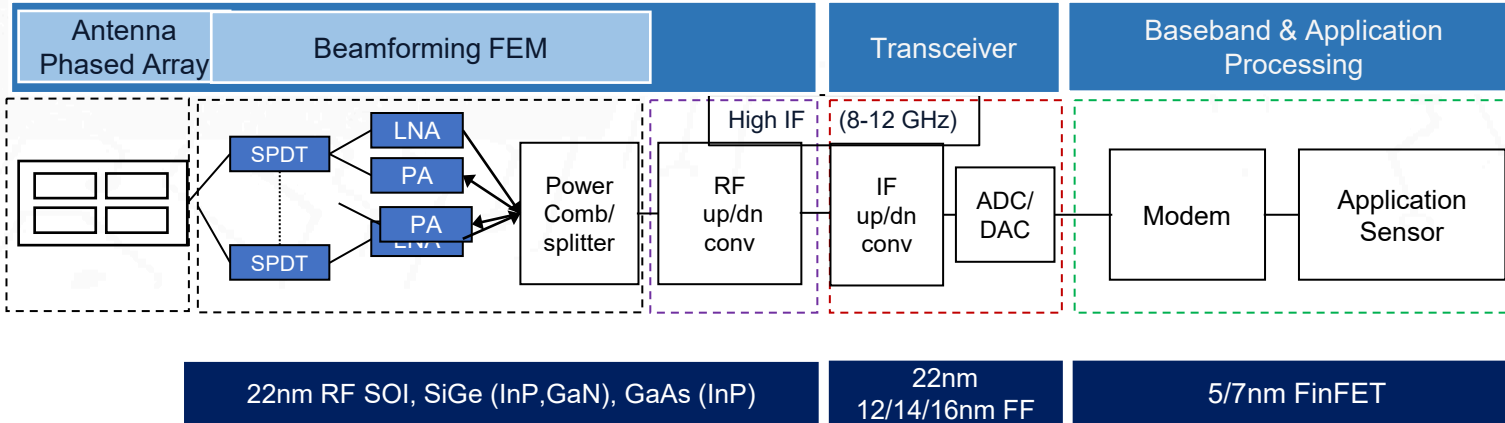
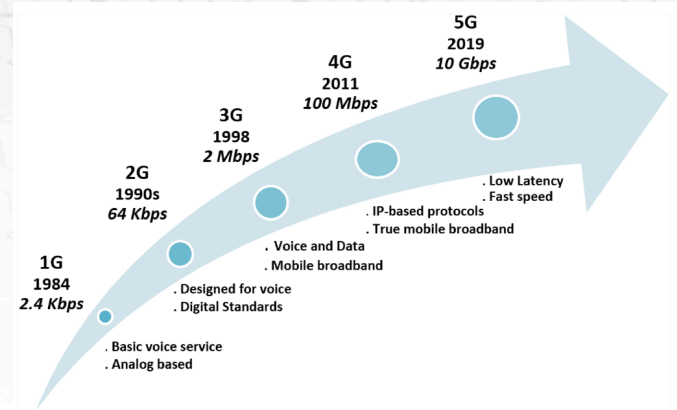
Arithmetic Intensity = FLOPS / Bytes (moved)

- **Compute Bound:** logic transistor performance, improve wire RC, stack more layers with high intertier via density
- **Memory bandwidth Bound:** Internal: Memory layer stacking with high TSV density, External: core to HBM interconnect using Si interposer

# Communications



Antenna Module



Chip partitioning and technology adoption depend on Tx power, power efficiency, cost and target form factor


Heterogeneous integration (HI) critical for mmwave communication

## Go vertical


## Augment charge with spin

## Embrace Heterogeneity


## Compute with memory



**Vertical CMOS**  
Addresses challenges of monolithic 3D integrated circuits



**Beyond CMOS**  
Tackles challenges of ultimate energy efficiency and ultra-fast switching in spintronics



**Heterogeneous Integration Fabric**  
Enables fine-pitch micro-aligned integration of functionally diverse dielets in a high-performance microsystem



**Merged Logic-Memory Fabric**  
Bridges gap between today's digital solutions and future cognitive system needs

- ❖ *BEOL logic, memory*
- ❖ *Barrier less HAR Vias*
- ❖ *Fine-grained thermal*
- ❖ *FEOL FETs (NC, Cryo)*

- ❖ *SOT MRAM*
- ❖ *VCMA MRAM*
- ❖ *Magneto-electric*
- ❖ *Probabilistic bit*

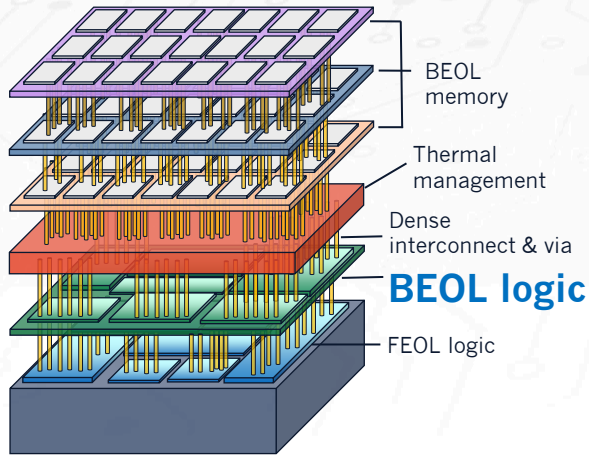
- ❖ *Millimeter-wave packaging*
- ❖ *Chip to chip signaling*
- ❖ *Power delivery*
- ❖ *Reconfigurable RF*
- ❖ *Package-level Thermal*

- ❖ *Analog weight cells*
- ❖ *Stochastic compute*
- ❖ *Collective compute*
- ❖ *Secure compute*

*Design, build and benchmark semiconductor prototypes to establish technology roadmap*

# BEOL Transistors (n-type)

## Monolithic 3D



Integration at

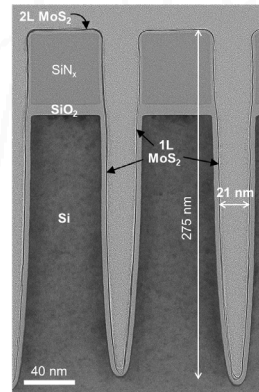
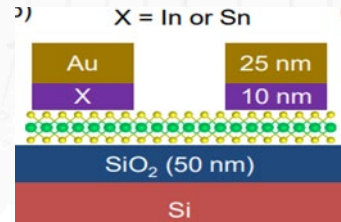
- Transistor level
- Gate level
- Block level

**Highlights:** ML MoS<sub>2</sub> growth on trench sidewall below 550°C; R<sub>c</sub> ~ 190 Ωμm (In/Au) and 220 Ωμm (Sn/Au)

**Challenges:** Defect control in ML TMD and dielectric interface

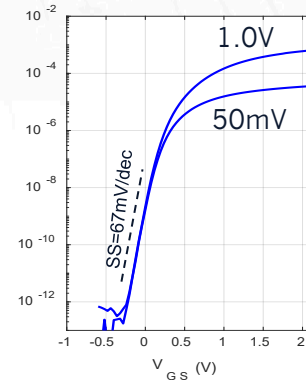
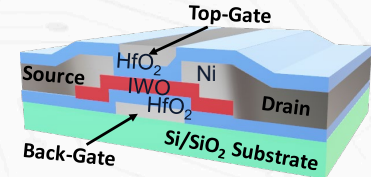
## 2D Materials

### Monolayer MoS<sub>2</sub> nFET



## Amorphous Oxide Semiconductors

### Dual-Gate W-doped In<sub>2</sub>O<sub>3</sub> nFET



**Highlights:** L<sub>G</sub> = 50 nm, EOT = 0.8 nm, SS 67 mV/dec, R<sub>C</sub> ≈ 500 Ω·μm, I<sub>D</sub> ≈ 720 μA/μm; improved V<sub>T</sub> stability

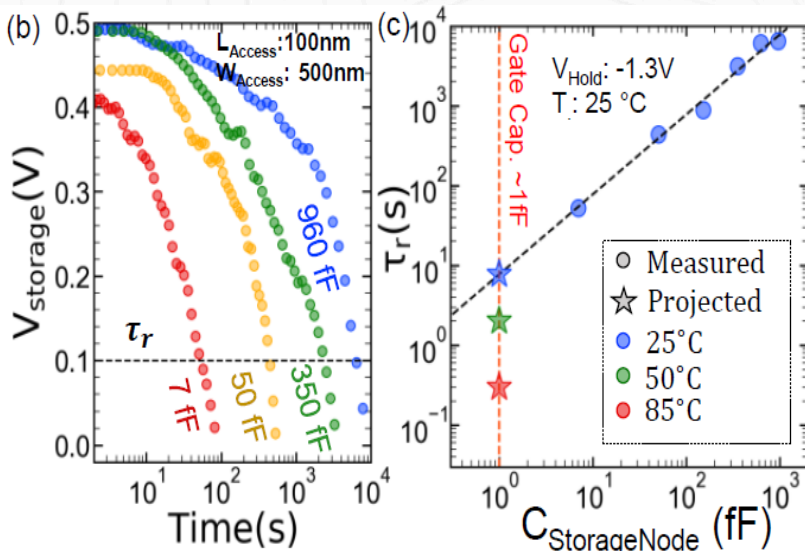
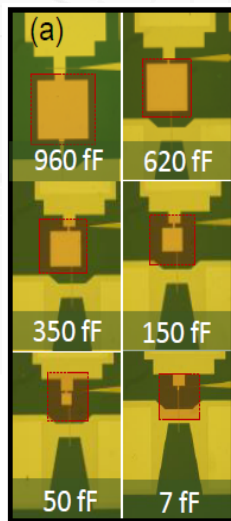
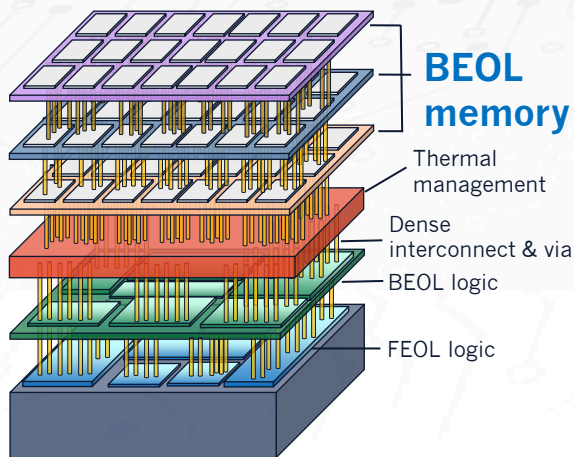
**Challenges:** Defect control in oxide thin films



# BEOL Memory (Monolithic 3D eDRAM)

2T (capacitorless) DRAM memory with sub femto ampere Ioff BEOL oxide FETs

## Monolithic 3D



Integration at

- Transistor level
- Gate level
- Block level

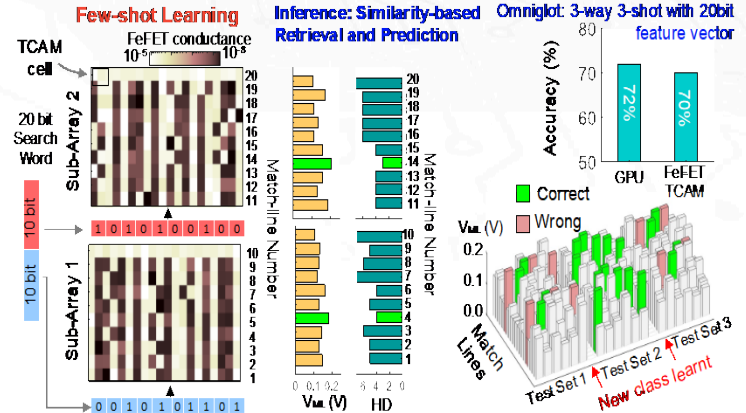
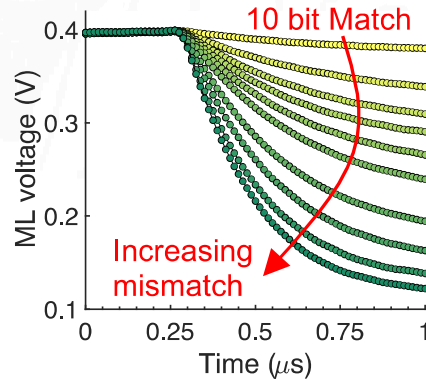
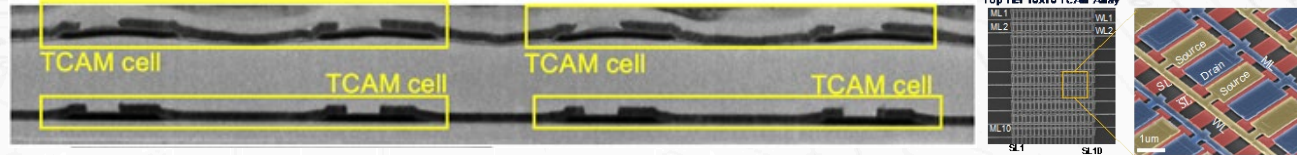
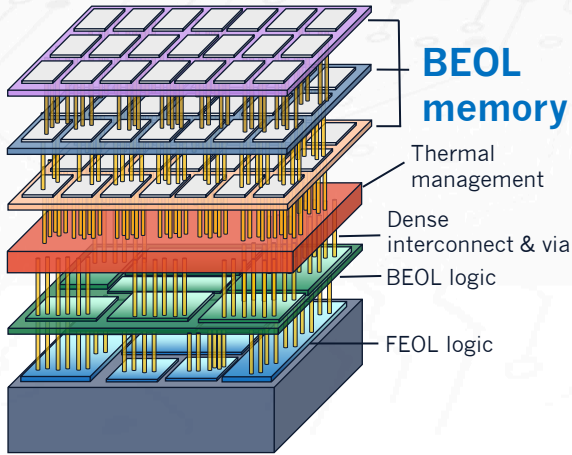
**Highlights:** Demonstration of 2T gain cell embedded DRAM (eDRAM) exhibiting (a) cell level leakage current of  $\sim 1 \times 10^{-15} \text{ A}/\mu\text{m}$  and  $\sim 1 \times 10^{-14} \text{ A}/\mu\text{m}$  at 25C and 85C

**Challenges:** Stability/Variability of IWO FETs without affecting mobility

# BEOL Memory (Monolithic 3D TCAM)

2 layers of Ferroelectric FETs as TCAMs for few shot learning

## Monolithic 3D



Integration at

- Transistor level
- Gate level
- Block level

**Highlights: Demonstration of M3D TCAM array demonstrating in situ compute of Hamming distance and 3-way 3-shot learning with 20-bit feature vectors**

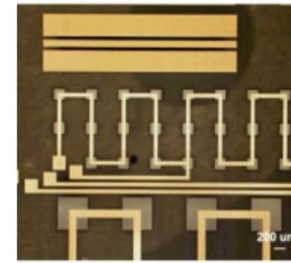
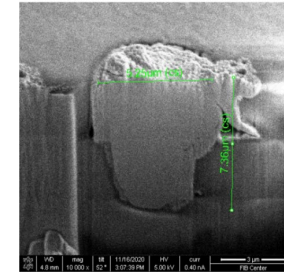
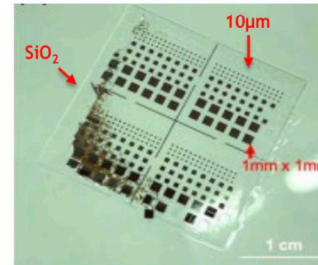
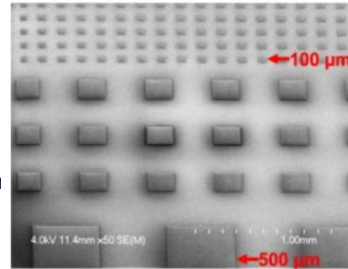
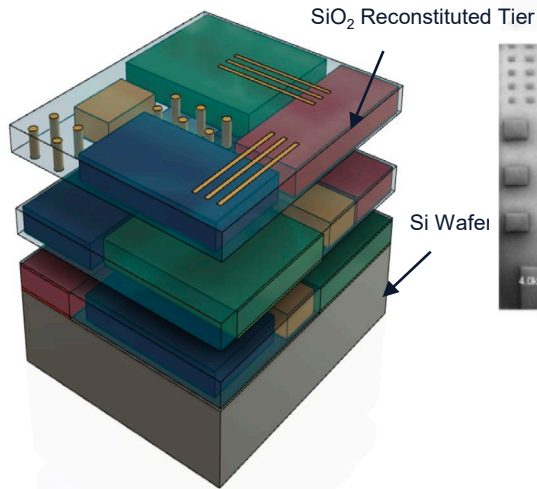
**Challenges: Fabricate arrays for larger statistics; SOC compatible voltage;  $V_t$  drift**

# BEOL Chiplets

## Polyolithic 3D

### 3D Integrated Chiplets Encapsulation (3D ICE)

#### Transfer of SiO<sub>2</sub>-reconstituted-tier on a glass wafer



‘Sea of Chiplets’ of varying dimensions after encapsulation

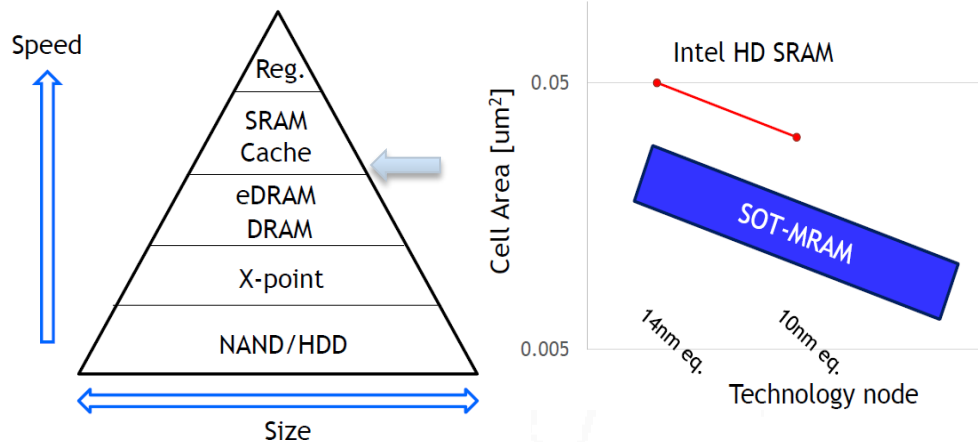
SiO<sub>2</sub>-reconstituted tier with through-oxide vias (TOV) and RDL

- Fills gap between M3D and heterogeneous packaging
- SiO<sub>2</sub>-reconstituted Tier for BEOL-level polyolithic integration

**Highlights:** Demonstrated SiO<sub>2</sub>-reconstituted Tier with via and RDL transferred onto glass substrate.

**Challenges:** Staying competitive with industrial 3DIC approaches; thermal management

# Augment Charge with Spin

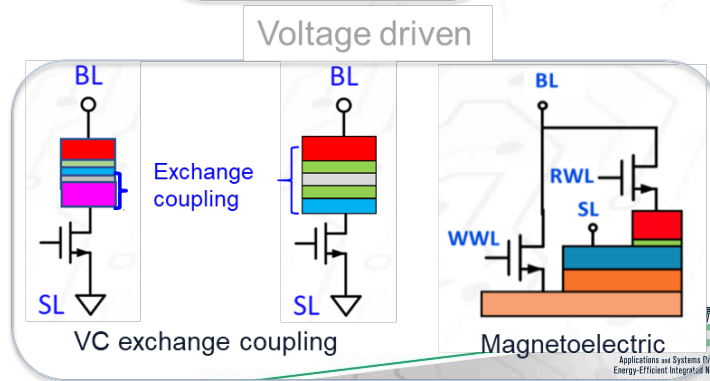
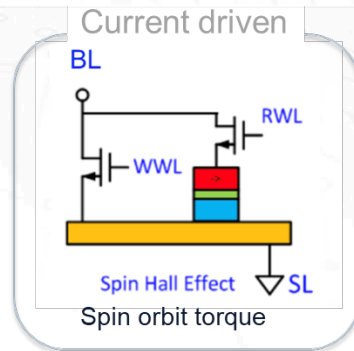


Noriyuki Sato, Ian Young (Intel) 2020 VLSI Symposium

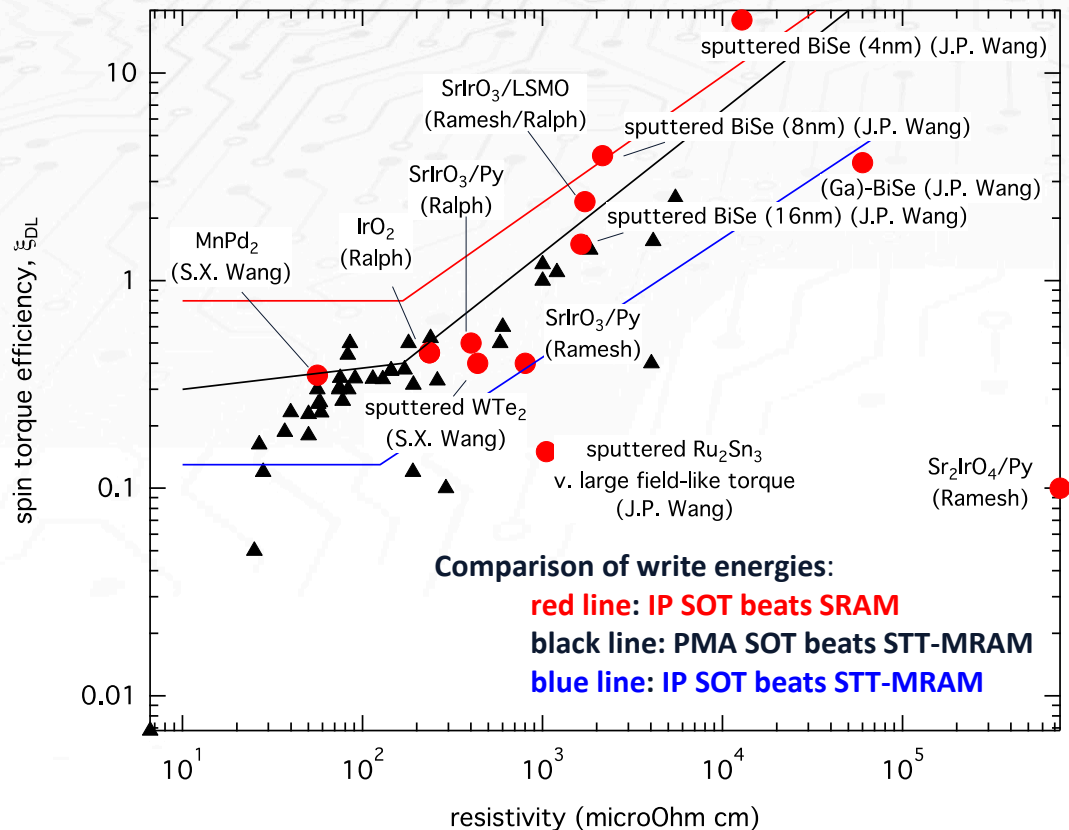
*Explore conventional and new symmetry materials to switch magnet (by current / voltage) more efficiently than STT*

## Beyond STT-MRAM

- Copper
- Fixed FM
- ME
- Oxide
- Free FM
- SOT material
- Spacer
- Fixed FM 2



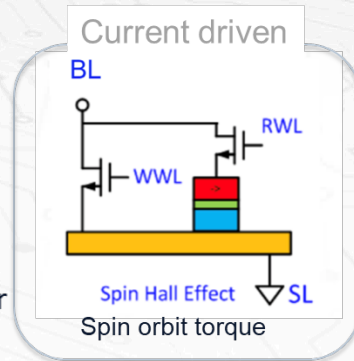
# Spin Orbit Torque (SOT) Memory



SOT efficiency

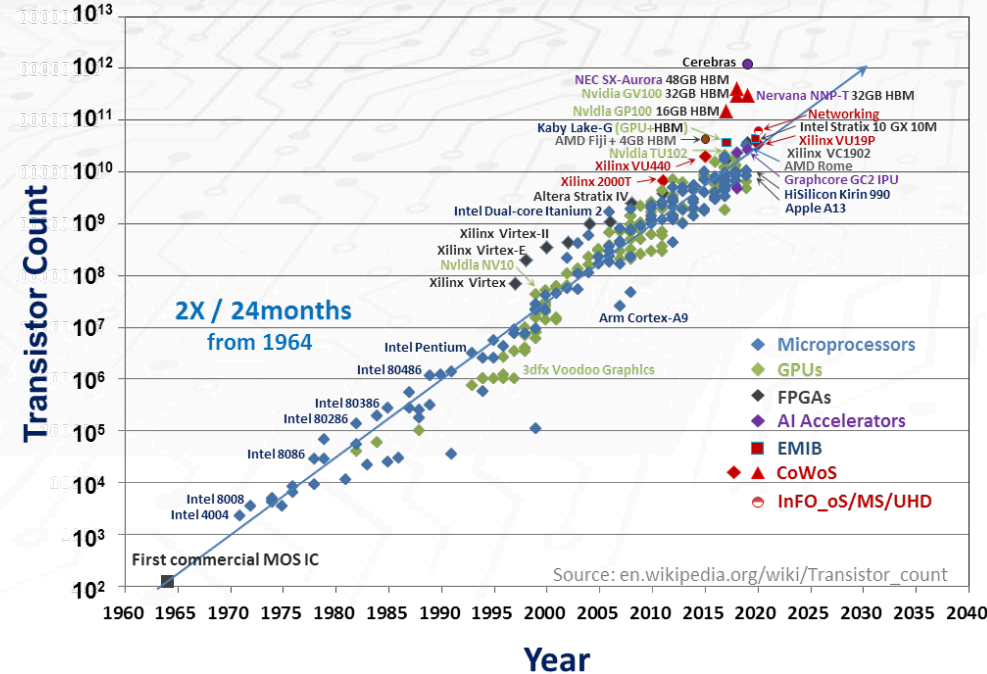
$$\xi_{DL} \frac{1}{\lambda_{sf}} \text{ (shunting factor)}$$

allows thin SO layer



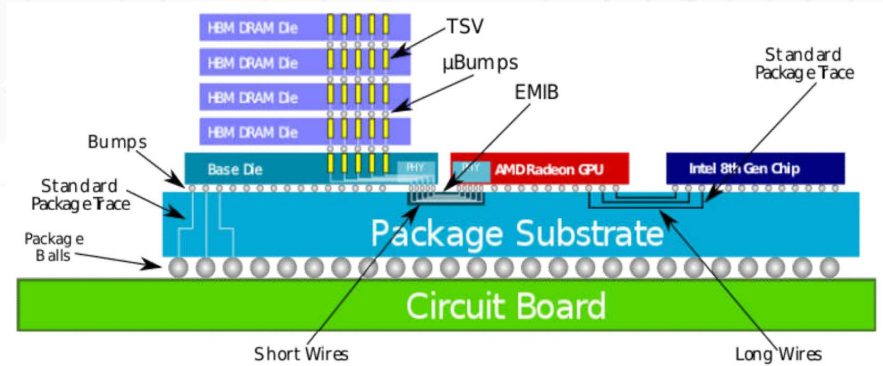
- Maximum spin torque efficiencies fall within resistivity range of  $10^3$ - $10^4 \mu\Omega\text{cm}$ .
  - Topological materials allow torque efficiencies  $> 1$
- Existing materials allow write energies less than STT-MRAM for in-plane SOT-MRAM

# Embrace Heterogeneity



Doug Yu (TSMC) 2019 IEDM Evening Panel

*It may prove to be more economical to build large systems out of smaller functions, which are separately packaged and interconnected - Gordon Moore, 1965*

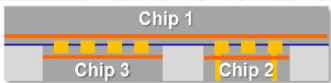


R. Viswanath (Intel) et al., 2018 IEEE EDAPS

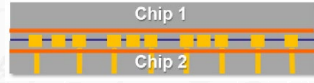
# Embrace Heterogeneity

## Today's HI trends

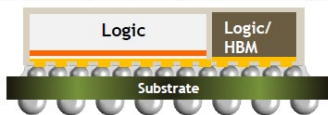
SoIC™



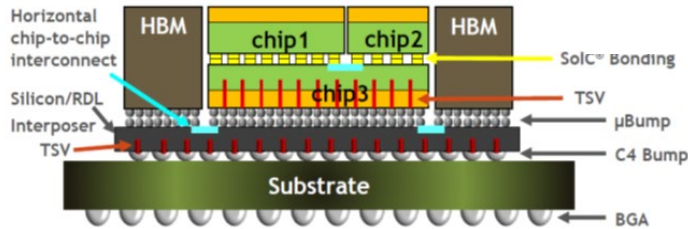
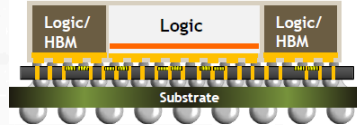
WoW



InFO



CoWoS

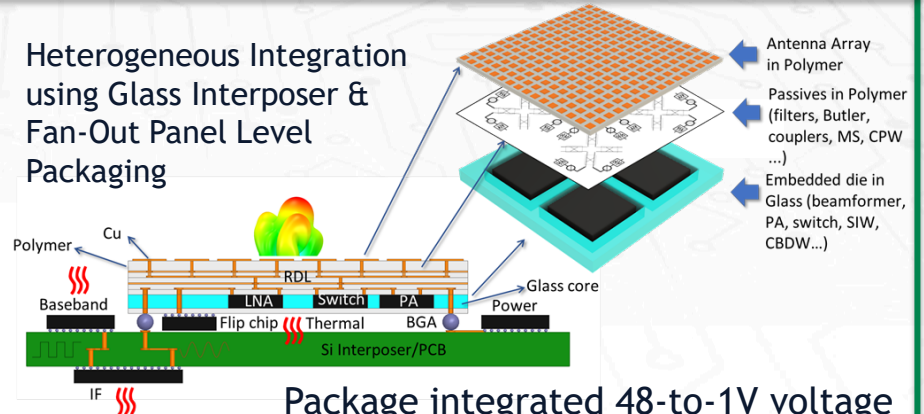


Smaller bump/bond pitch  
Larger package size

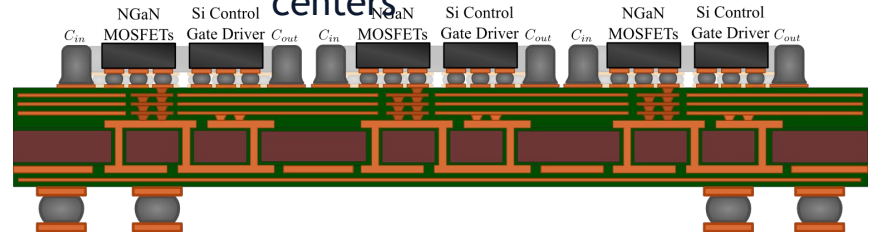
3D integration

## Heterogeneous fabric (ASCENT)

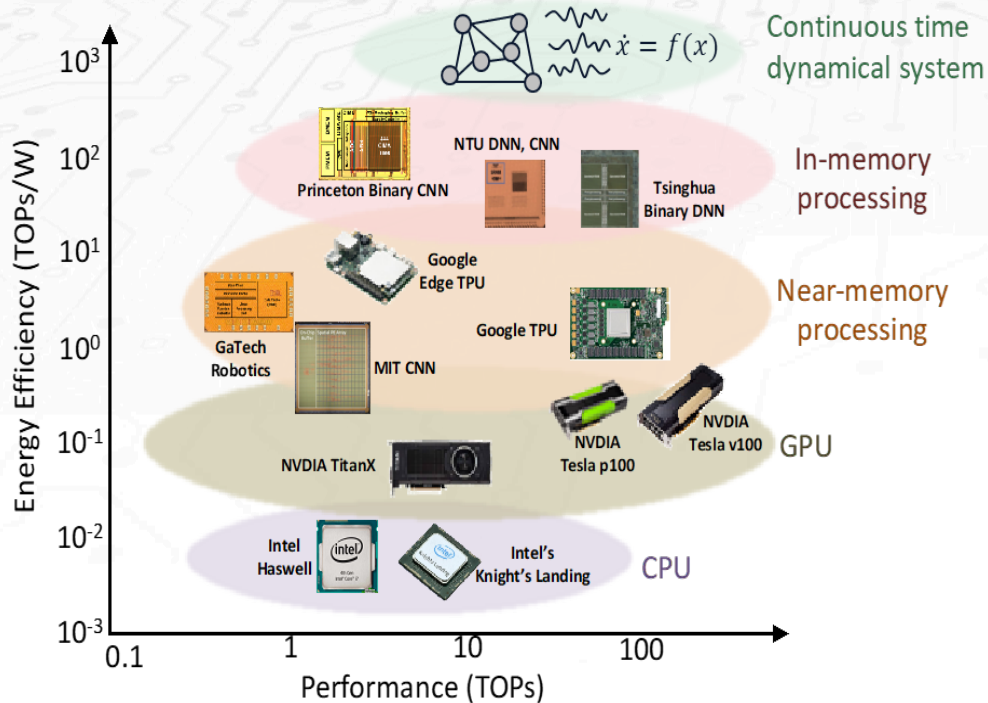
Heterogeneous Integration using Glass Interposer & Fan-Out Panel Level Packaging



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

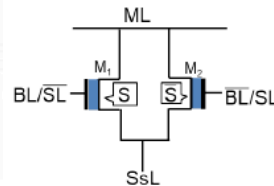


# Compute with Memory

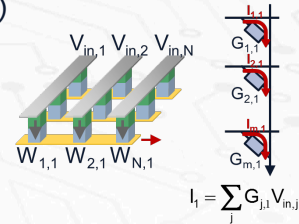


## Compute with memory (ASCENT)

### Compute-in-memory (CIM)

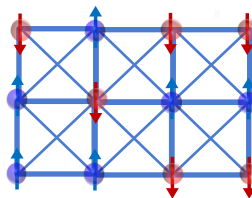


Fast search or Homomorphic encryption with TCAM

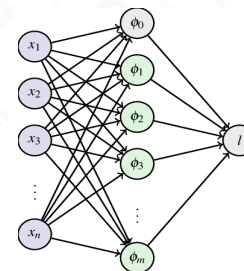


Analog MAC with cross-bar

### Dynamical systems



Ising machine

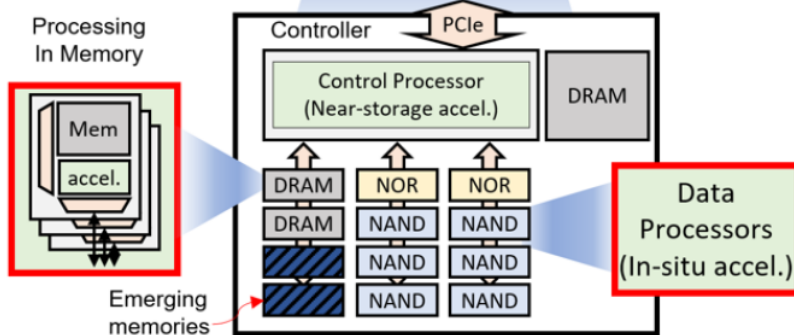
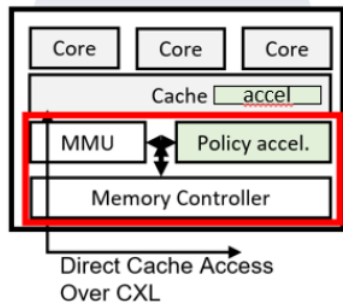
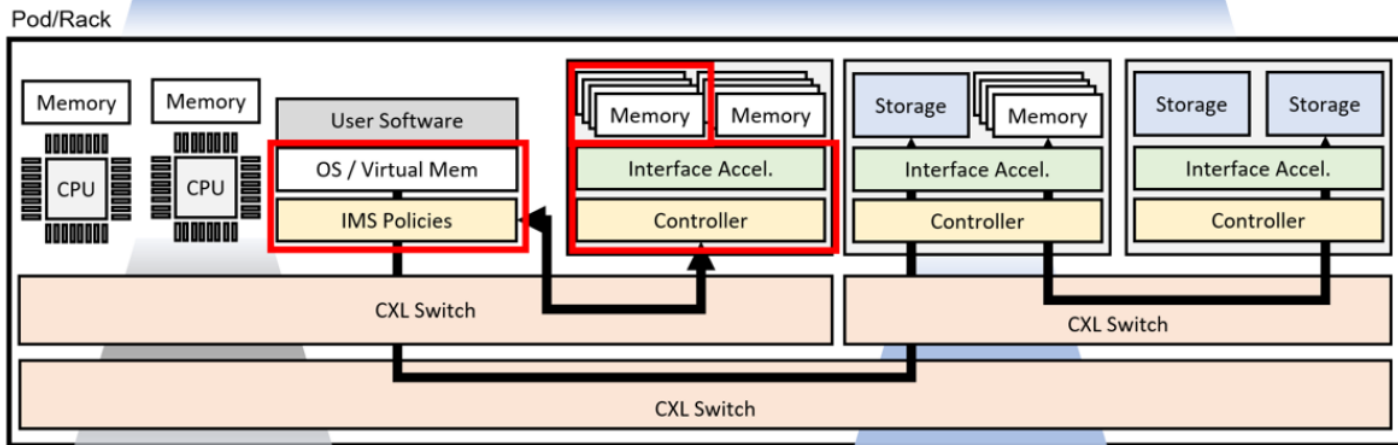


Stochastic Restricted Boltzmann machines

Explore specialized hardware for solving computationally hard problems



# The future of data-centric compute



.... is heterogenous and distributed



# JUMP

Joint University Microelectronics Program




[www.src.org/program/jump](http://www.src.org/program/jump)

Semiconductor Research Corporation

@srcJUMP





SIA-SRC Webinar on the  
Collaboration towards  
Decadal Plan Goals:  
Advances and Challenges  
in Semiconductor  
Hardware

**Successes and  
Learnings from JUMP  
ComSenTer**

Dr. Farhana Sheikh,  
Intel Corporation



intel®



# Legal Notices & Disclaimers

Intel technologies may require enabled hardware, software or service activation.

No product or component can be absolutely secure.

Your costs and results may vary.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

All product and service plans, and roadmaps are subject to change without notice. Any forecasts of goods and services needed for Intel's operations are provided for discussion purposes only. Intel will have no liability to make any purchase in connection with forecasts published in this document. Code names are often used by Intel to identify products, technologies, or services that are in development and usage may change over time. No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request. Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade. You may not use or facilitate the use of this document in connection with any infringement or other legal analysis concerning Intel products described herein. You agree to grant Intel a non-exclusive, royalty-free license to any patent claim thereafter drafted which includes subject matter disclosed herein.

Statements in this document that refer to future plans or expectations are forward-looking statements. These statements are based on current expectations and involve many risks and uncertainties that could cause actual results to differ materially from those expressed or implied in such statements. For more information on the factors that could cause actual results to differ materially, see our most recent earnings release and SEC filings at [www.intc.com](http://www.intc.com).

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others. This document contains information on products and/or processes in development.

# Tech has never been more important to humanity

Computing has become pervasive, and the entire world is becoming digital

*Revolutionary technologies to address high-frequency radar, sensing, and communications: systems-to-circuits, beyond-CMOS devices, CMOS optimizations for sub-THz circuits – convergence of communications, sensing, and compute*

### ComSenTer System Highlights:

- 140GHz Channel measurements up to 100m in urban environments, 1GHz BW
- Digital MU-MIMO: low-cost, energy-efficient digital beamforming → adaptability
- Systems-to-circuits modeling and optimization: 4X to 5X power reduction

### ComSenTer Circuits Highlights:

- CMOS-based PA with highest power & PAE at 140GHz – record-setting work
- CMOS-based D-band phased array with 13Gb/s & integration into COTS system
- 32-element DBF ASIC + SERDES + ADC-DAC array + baseband
- GaN PA at 140GHz with >24dBm output, 10dB gain; 220GHz InP PA: 30% PAE

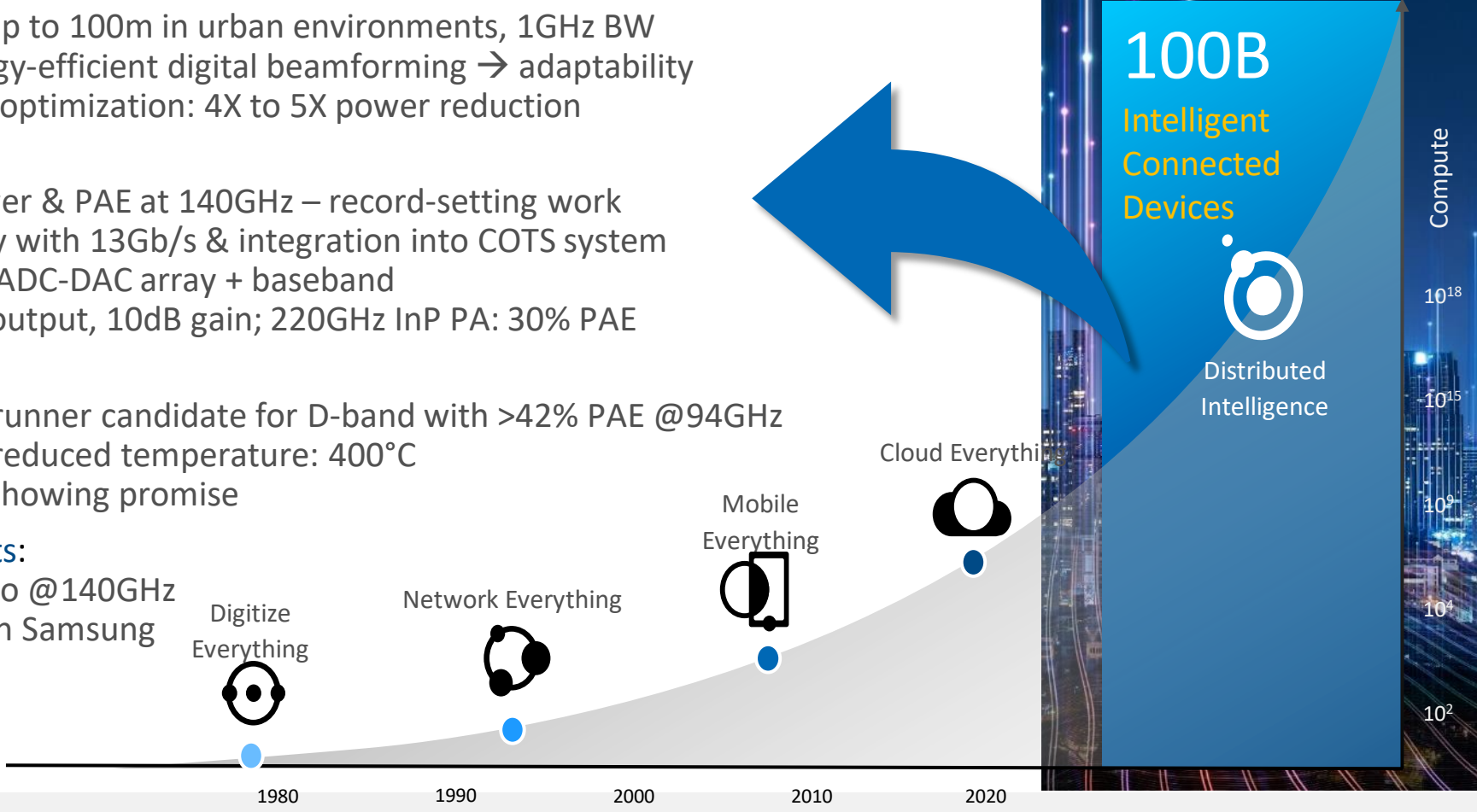
### ComSenTer Devices Highlights:

- N-Polar GaN emergence as front-runner candidate for D-band with >42% PAE @94GHz
- Successful growth of diamond at reduced temperature: 400°C
- AlN/GaN transistors healthy and showing promise

### ComSenTer Demonstration Highlights:

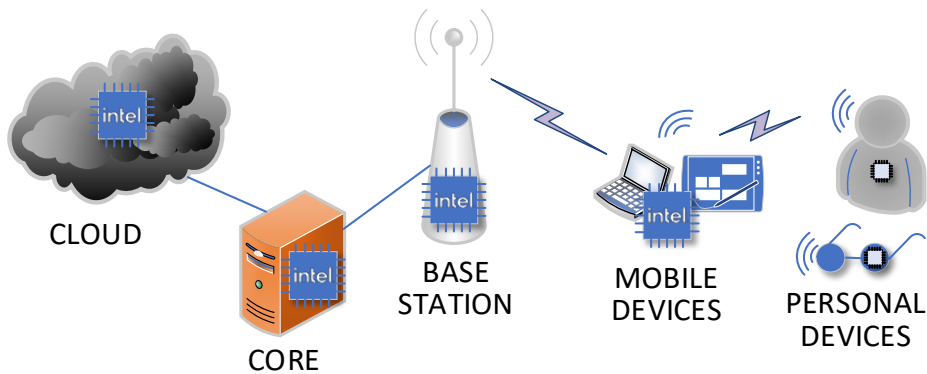
- Massive MIMO COTS + ASICs demo @140GHz
- 140GHz massive MIMO demo with Samsung

*Additional information posted on src.org under JUMP ComSenTer*

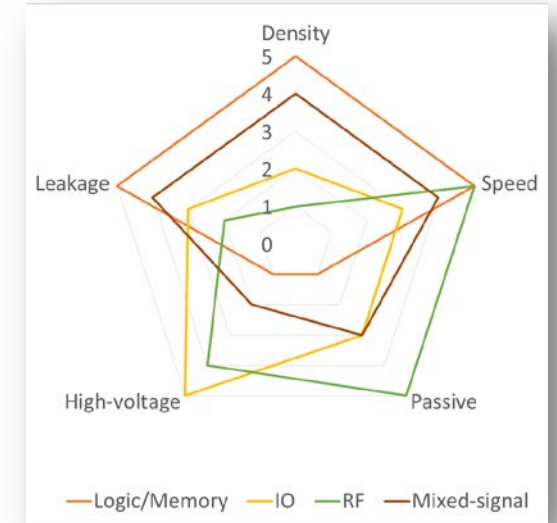


# Heterogeneity and Modularity: Systems Integration

Are we ready for a new wave in Semiconductor Technologies & Education?



*Convergence of communications,  
sensing, and compute:  
End-to-end integration of  
heterogeneous materials/devices,  
technologies, circuits, and platforms*



- 1. Optimized systems integration → Platform Centric Design focusing on Functional Density**  
*High-frequency RF (III-V, SiGe, CMOS) + Analog/Mixed-Signal + Digital + Configurable Arrays with Domain Specific Compute + Advanced memory architectures*
- 2. Security + Intelligence → Design and integration from Day 1:** *integral part of algorithms, architecture, ...*
- 3. Optimized packaging + 2.5D/3D Chiplet based devices, circuits, & architectures → Interface I/O power can be high**  
*Need for 2.5D / 3D and multi-die heterogeneous integration at nano, micro, macro levels for comms and sensing  
Optimal partitioning and automated design methodologies/flows*
- 4. Work force development in emerging sub-THz comms/sensing/imaging with focus on diversity**  
*Build technical leadership broadly across USA schools in devices, circuits, packaging, & systems: multi-disciplinary workforce → materials and devices + circuits and tapeout + software + systems/architecture classes*



“Don’t be encumbered by history.  
Go off and **do something wonderful.**”



**Robert Noyce**  
Co-Founder of Intel

The Intel logo is centered on a solid blue background. It features the word "intel" in a white, lowercase, sans-serif font. A small blue square is positioned above the letter "i". To the right of the word "intel" is a white registered trademark symbol (®).

intel®



# Advances and Challenges in Semiconductor Hardware

## Madhavan Swaminathan, Georgia Tech

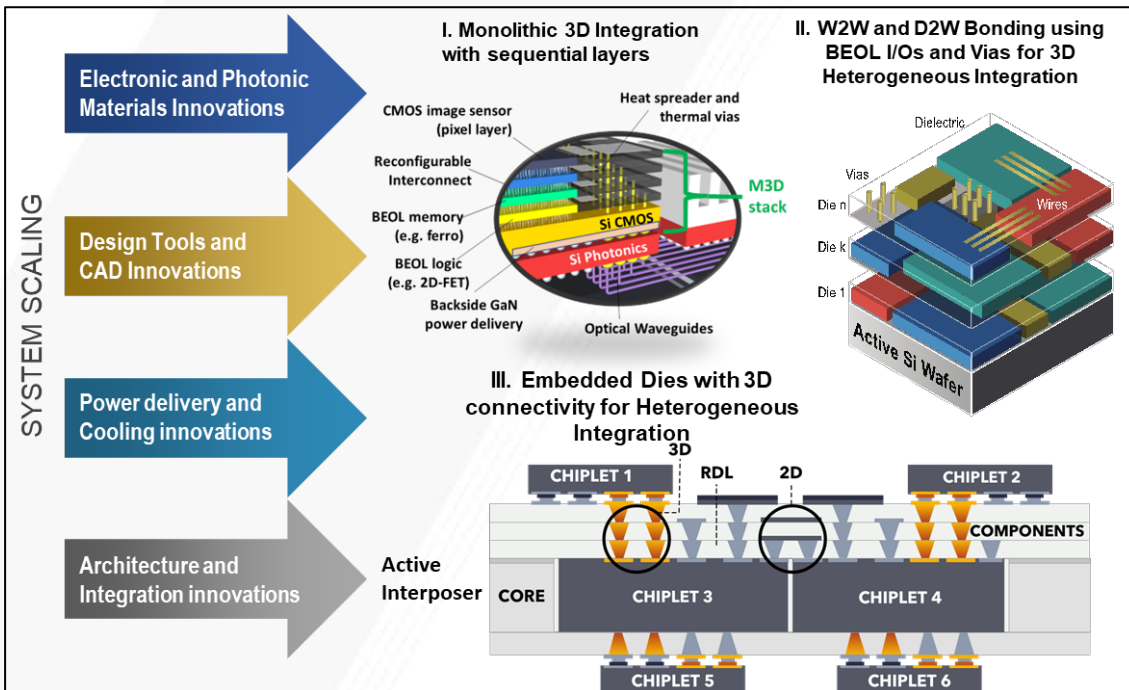


- ❑ Global semiconductor industry projected to become a trillion-dollar industry by 2030 (Source: McKinsey & Company)
  - 55 years to become a \$0.5 Trillion industry
  - 10 years to become a \$1.0 Trillion industry

### ❑ Drivers & Challenges

- Compute & Storage (Deep Learning: 300X “brute force” System Scale Out since 2012 leading to energy crisis)
- **Wireless** (Single autonomous car expected to produce 4000 Gigabytes of data per day)

- ❑ Need Energy Efficient solutions (femto-Joules/bit) with Large Bandwidth Density (500TBps/mm<sup>2</sup>) in the future



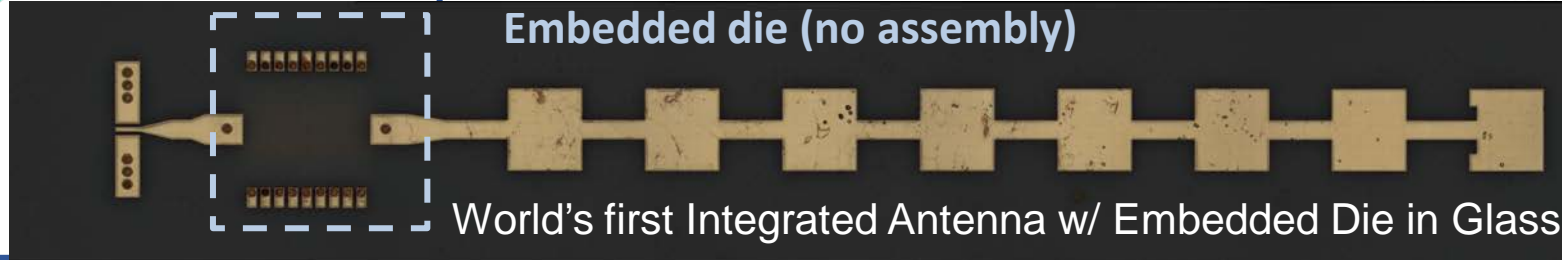
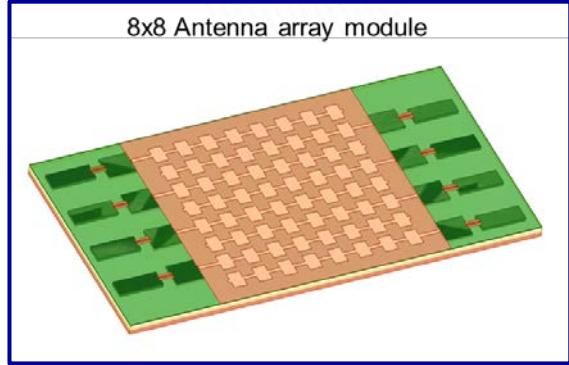
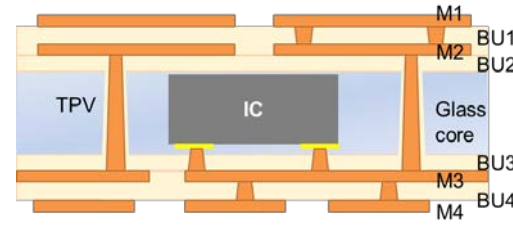
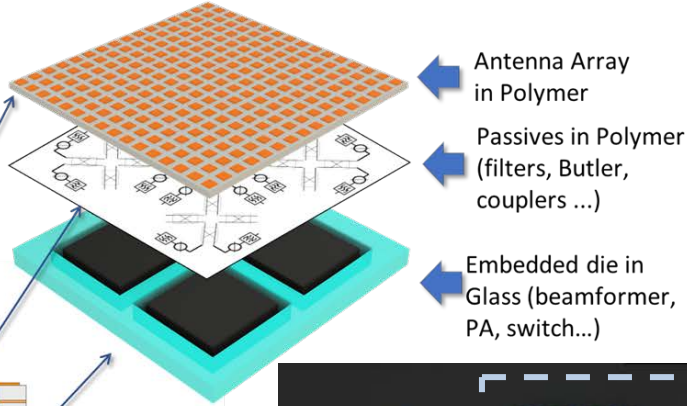
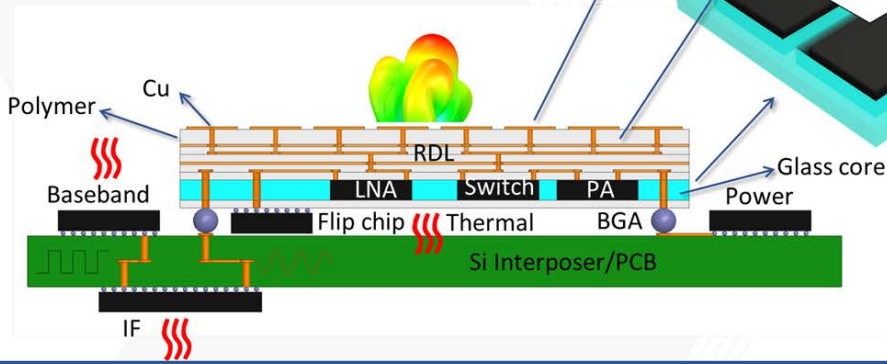
With Moore’s law slowing down, need **Heterogeneous Integration** platforms which combine both sequential Monolithic 3D Integration (M3D) and parallel polyolithic chiplet integration to achieve 100X improvement in transistor, IO, and Bandwidth Densities.

<https://www.inovex.de/de/blog/edge-computing-introduction/>

Andrew John and Micah Musser, “AI and Compute – How much longer can computing power drive artificial intelligence progress”, CSET, 2022.

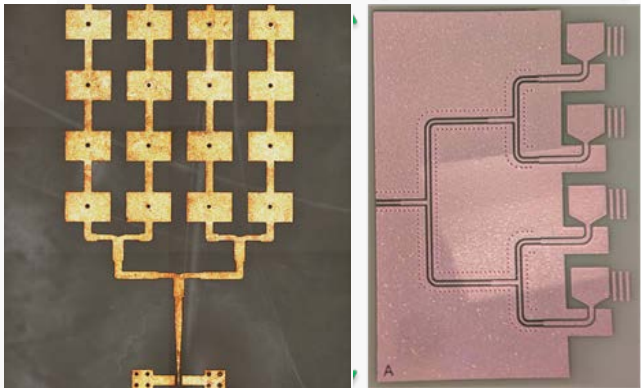
# Recent Key Accomplishments: Antenna in Package for 6G (D-Band)

## Glass Interposer for D-Band (Wireless)



Xiaofan Jia et al, ECTC 2022 (10.7dB Gain @ 138GHz)

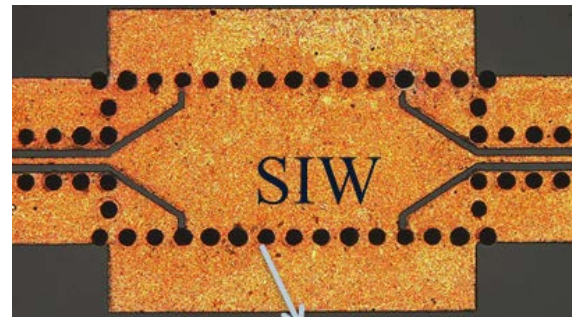
## ANTENNA ARRAY



14-16dBi Gain      11dB Gain

Kai-Qi Huang et al, ECTC 2021  
Serhat Erdogan et al, IMS 2022

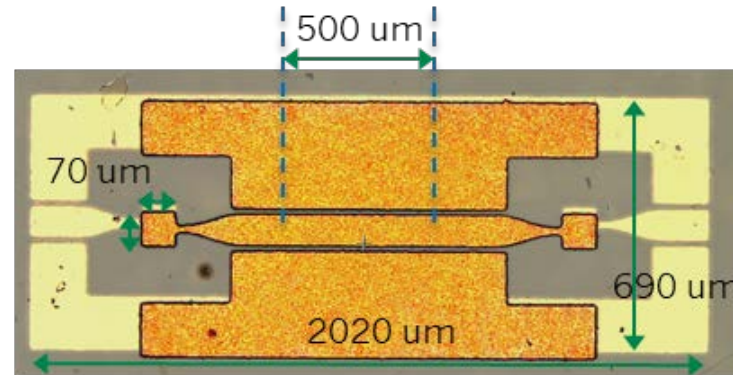
## SIW



100/200um TGV

SIW Loss: 0.7dB/mm

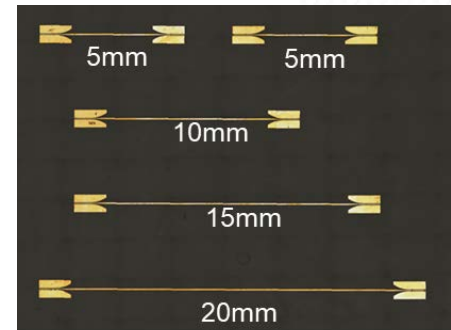
Mutee Rehman et al, IMS 2021



Via-less Loss: 1.8dB

L. Vijaykumar et al, ECTC 2022

## Planar Goubau Lines Loss: 0.34dB/mm

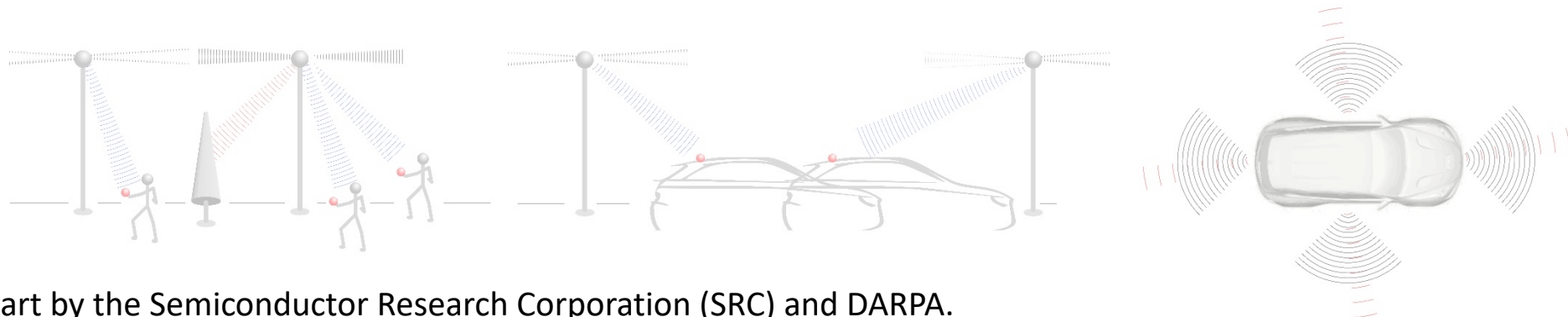


Xiaofan Jia et al, IMS 2022

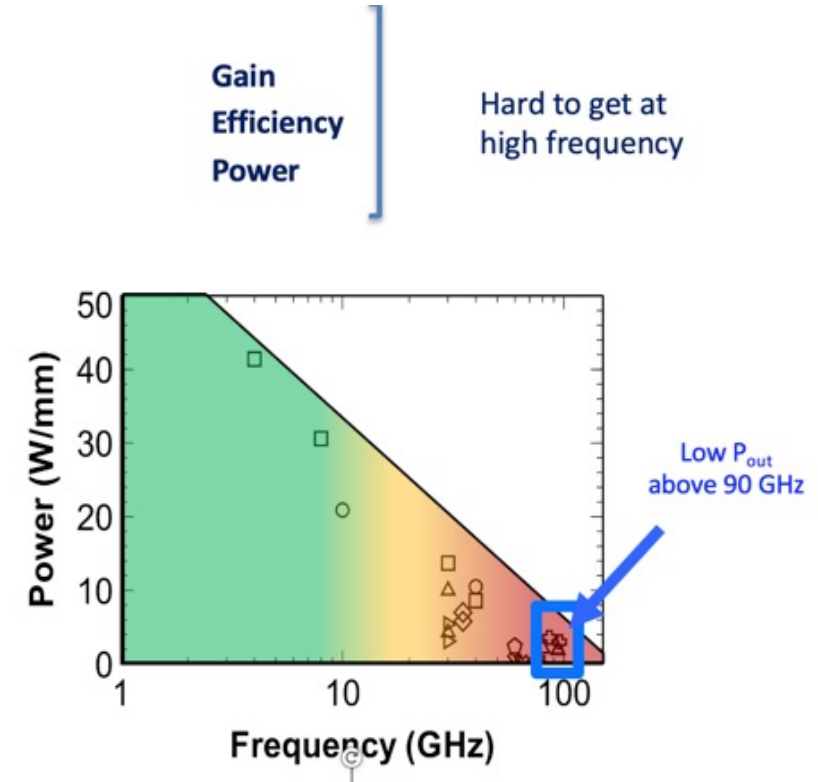
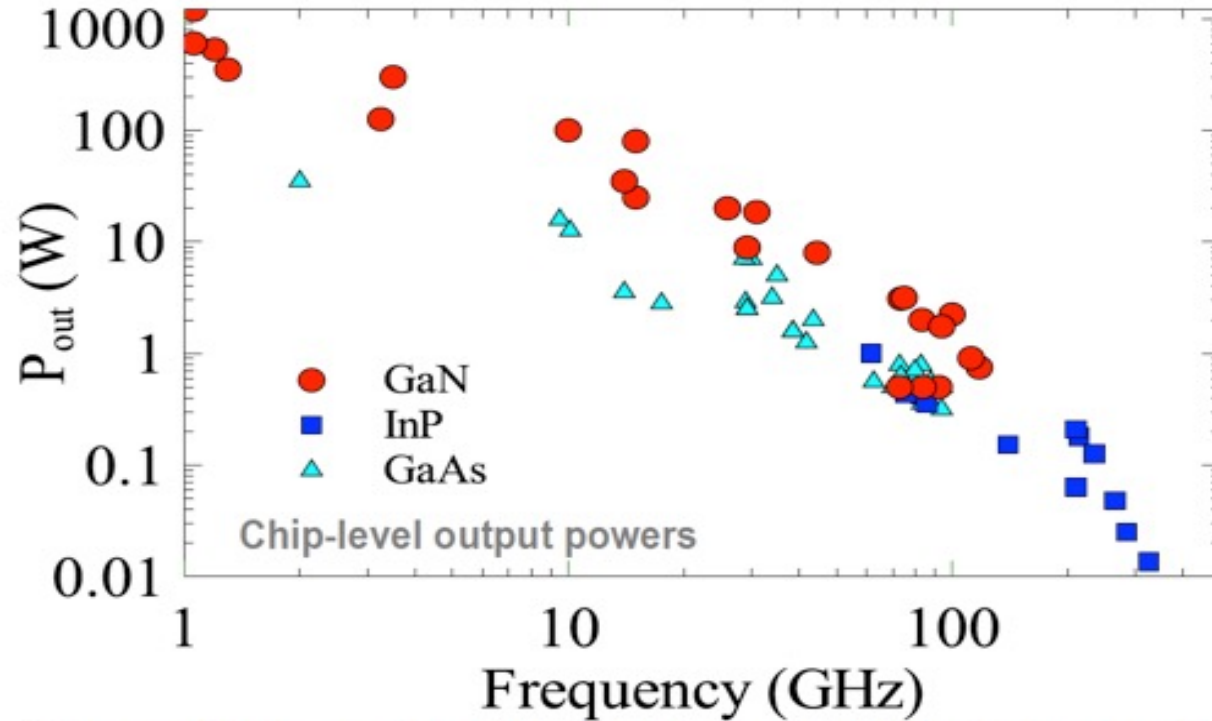
# Progress Towards High-Performance THz and mm-Wave Transistors for Wireless Systems

***Srabanti Chowdhury***

**[srabanti@stanford.edu](mailto:srabanti@stanford.edu)**



This work was supported in part by the Semiconductor Research Corporation (SRC) and DARPA.



The center's application goals require high transmit power and low receiver noise figure beyond the state of the art.

The improved device-level performance needed is addressed through application-specific THz transistors

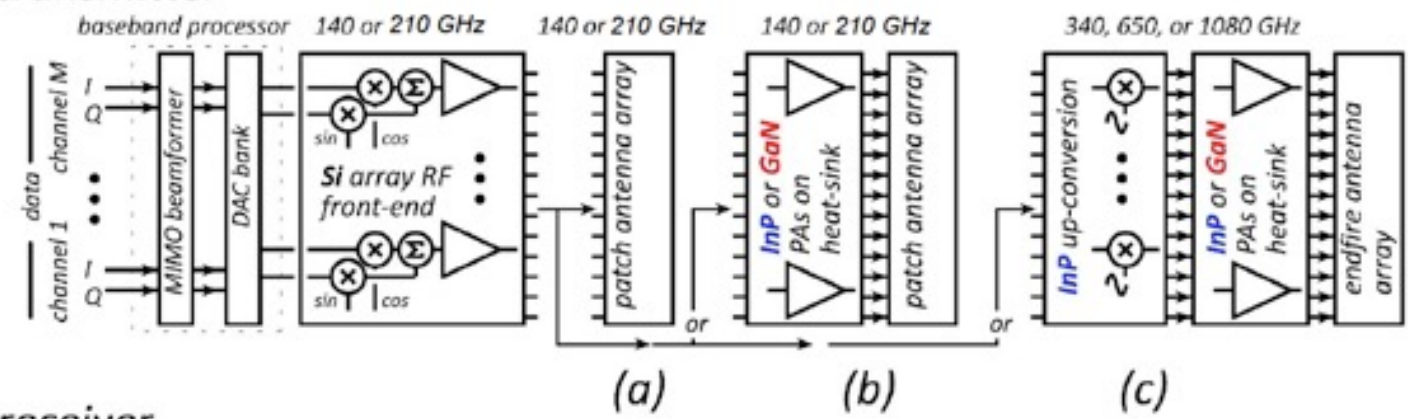
**Silicon**  
baseband processing at all frequencies  
RF sections @ 140, 200GHz  
PAs, LNAs in short-range 140, 210 GHz links

**GaN**  
high-power amplifiers in long-range 140,210GHz links  
(possibly 340GHz ?), with integrated Diamond cooling  
→ Mishra, Xing, Jena, Chowdhury

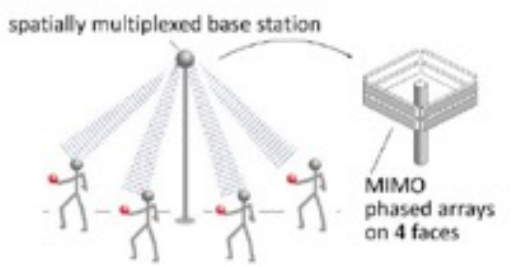
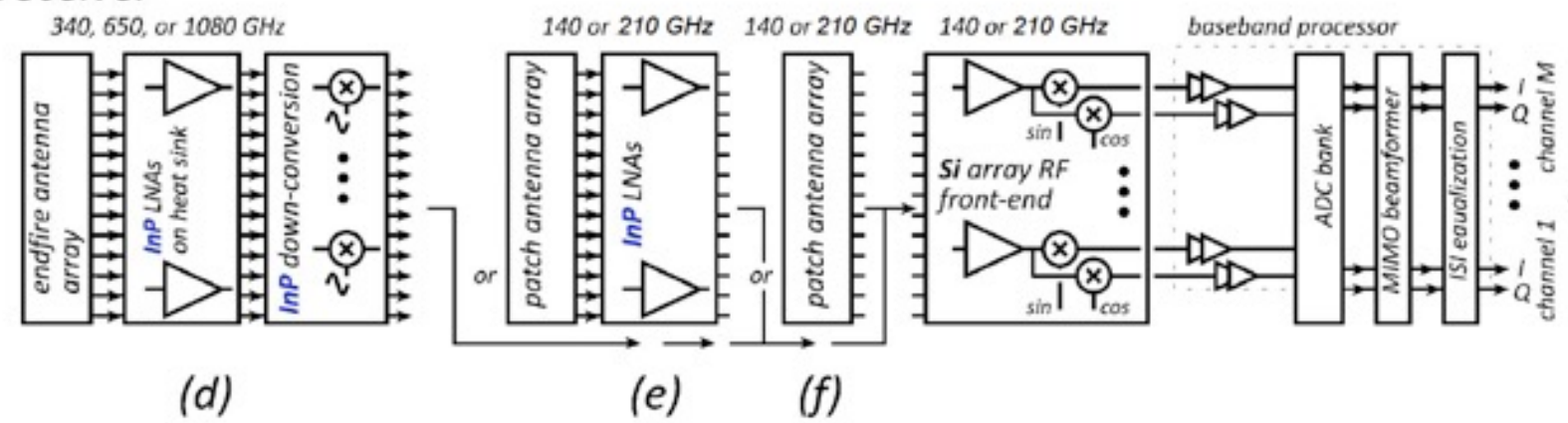
**InP MOS-HEMT**  
low-noise amplifiers in long-range 140,210GHz links  
low-noise amplifiers @ 290, 650GHz  
→ Rodwell

**InP HBT**  
medium-power amplifiers in long-range 140, 210GHz links  
power amplifiers @290, 650GHz  
RF sections @ 290, 650GHz  
→ Rodwell (devices), Buckwalter (IC in foundry process)

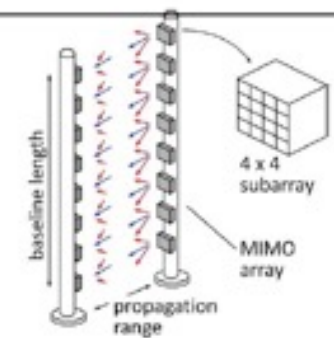
**transmitter**



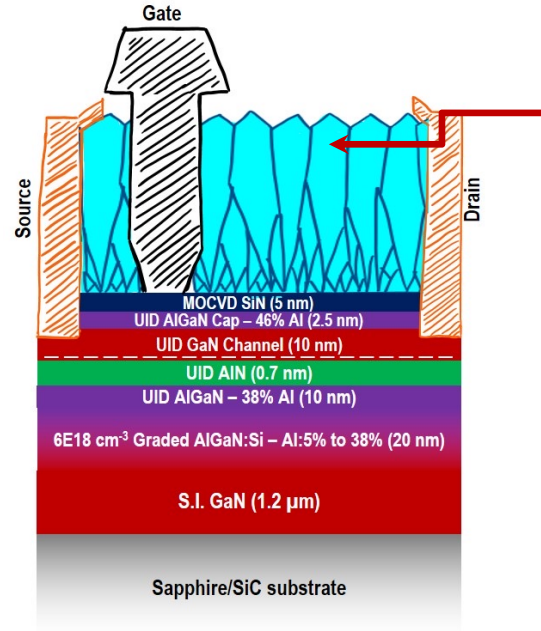
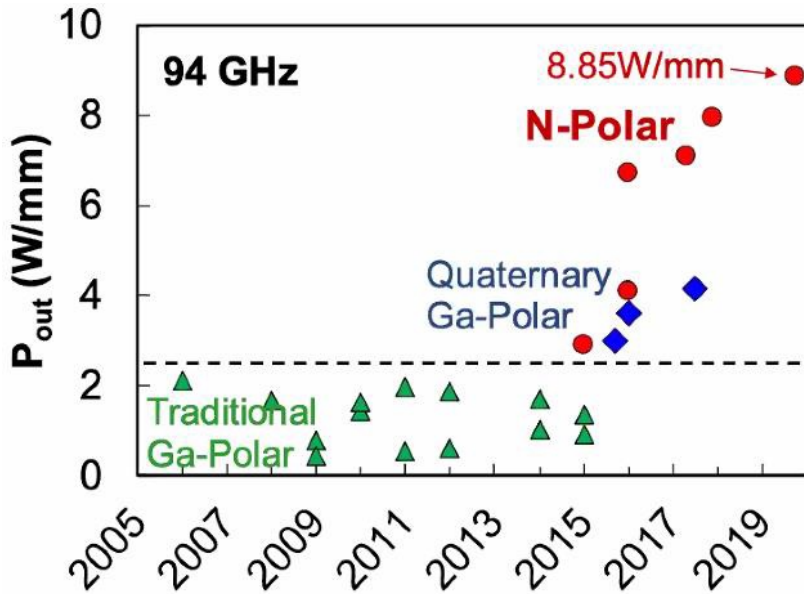
**receiver**



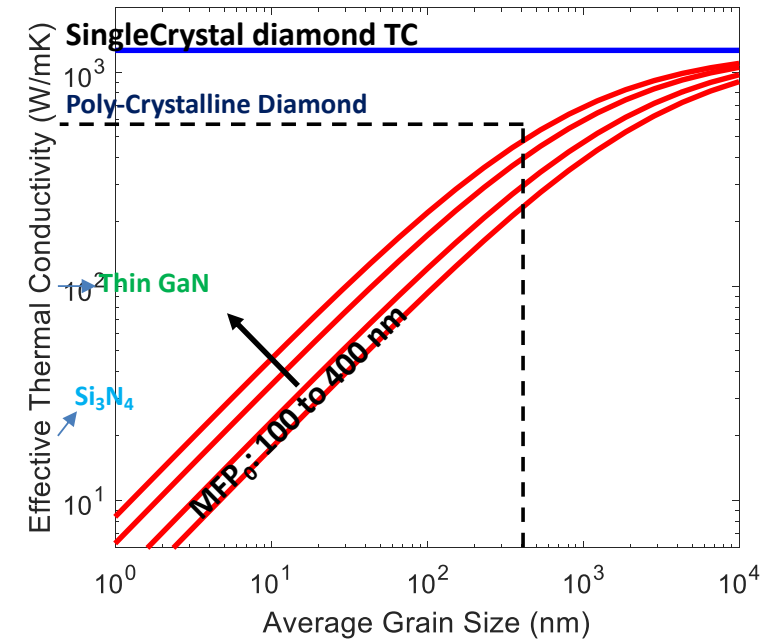
**MIMO hub:**  
140GHz: F= 8dB,  $P_{avg}$ =21dBm,  $P_{1dB}$  ≈ 25dBm



**Point-point MIMO:**  
210GHz: F= 6dB,  $P_{avg}$ =16dBm,  $P_{1dB}$  ≈ 21dBm



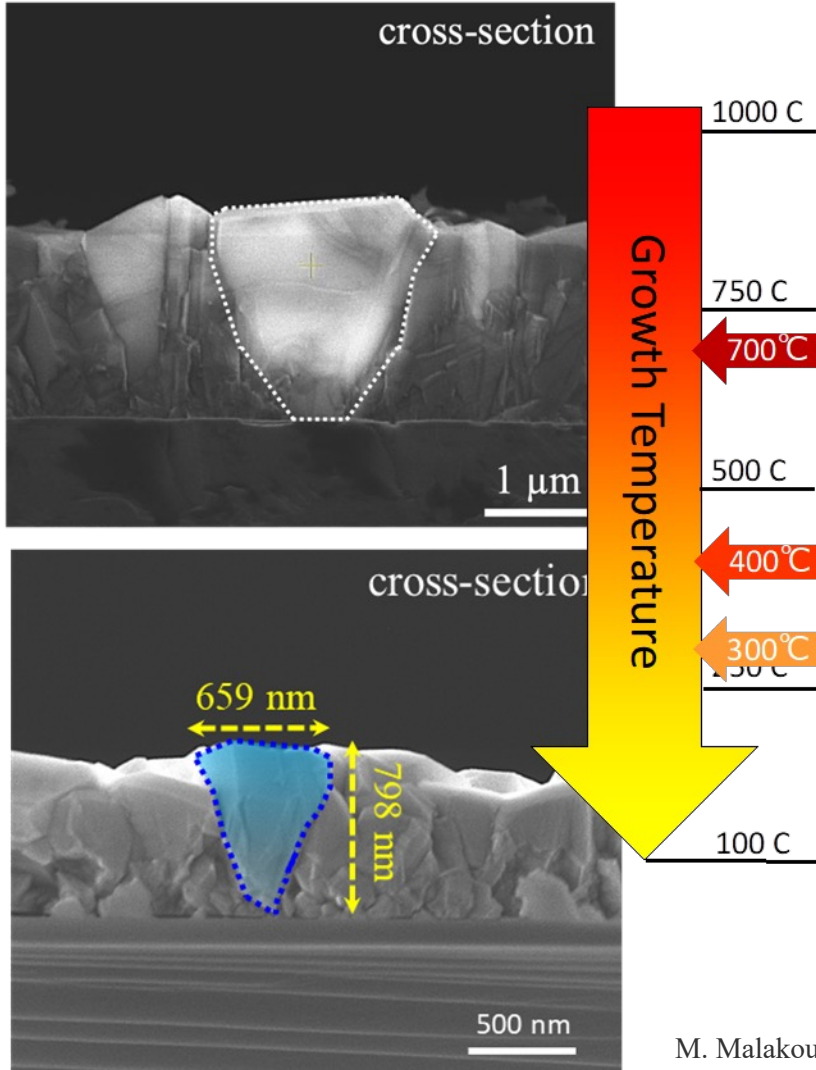
|         | Thermal Expansion ( $10^{-6}/K$ ) | Thermal Conductivity (W/mK) | BandGap (eV) | Dielectric Constant |
|---------|-----------------------------------|-----------------------------|--------------|---------------------|
| Diamond | 1.0                               | 100-2000                    | 5.45         | 5.7                 |



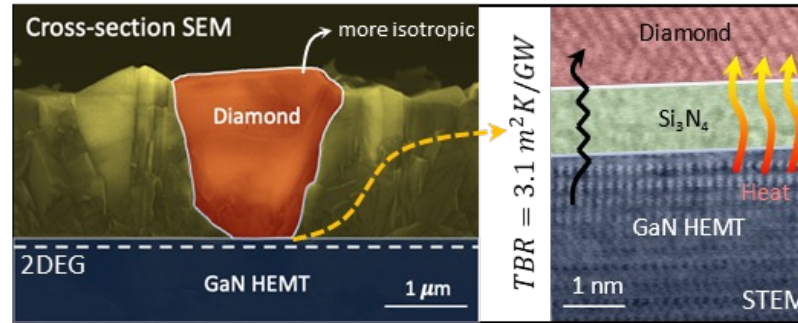
- ❖ To achieve over  $x$  W operation in mm-wave domain (94GHz, 240Ghz and 300Ghz) with a PAE of 20-25%.
- ❖ The device should be able to transfer heat over 3-4x W to deliver  $x$  W without losing the performance.

→ Leverage the thermal conductivity of diamond without hurting the GaN channel mobility.

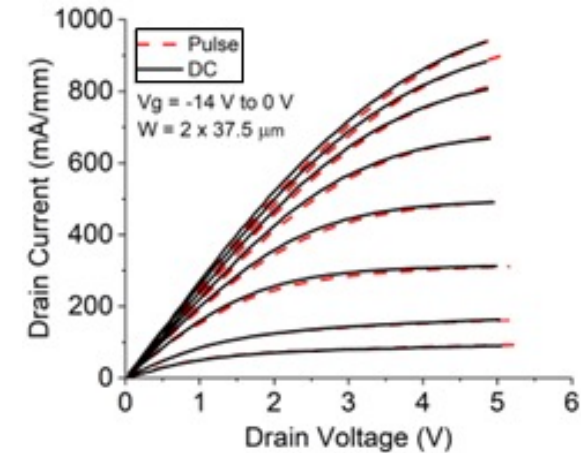
## Diamond growth temperature reduction



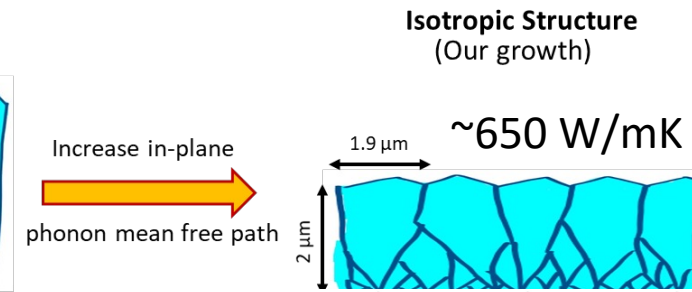
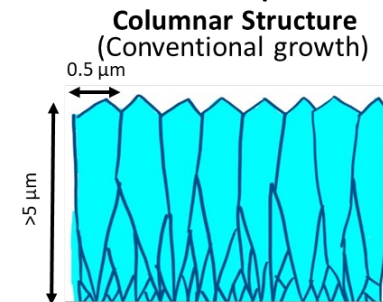
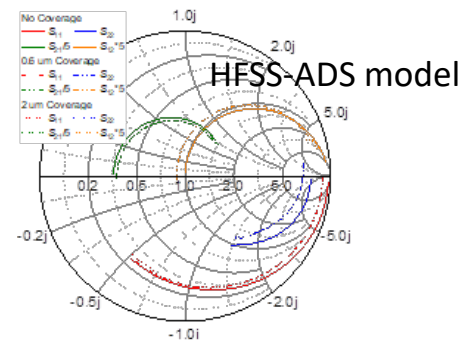
## Lowest reported diamond-GaN TBR



- ✓ Lowest diamond-GaN TBR ( $3.1 \text{ m}^2\text{K/GW}$ )
- ✓ High-quality seeding and nucleation
- ✓ Thin but high thermal conductivity diamond
- ✓ Near-isotropic grains at  $400^\circ\text{C}$
- ✓ Dispersion free HEMT w/ diamond on top
- ✓ HFSS-ADS high-frequency model development



Dispersion free HEMT w/ diamond



M. Malakoutian, ..., S. Chowdhury, Accepted for TECHCON 2022, Austin, TX, United States, 2021.

M. Malakoutian, ..., S. Chowdhury, Cryst. Growth Des., vol. 21, no. 5, pp. 2624–2632, 2021.

M. Malakoutian, ..., S. Chowdhury, Appl. Phys. Express, vol. 14, p. 055502, 2021

M. Malakoutian, ..., S. Chowdhury, ACS Appl. Mater. Interfaces, vol. 13, no. 50, pp. 60553–60560, 2021.

- SRC enabled a platform where industry and academic researchers can collaborate, observe, develop and nurture technologies
- The “energy” felt during our ComSenTer reviews was exceptional
- Collaboration without boundaries helped set up successful seed ideas/projects (DARPA)



The strength of these relationships will drive the future innovations, workforce development and supply of talents to achieve ambitious yet practical goals.

We need to remove any roadblocks that threaten such collaboration and progress.



# *Collaboration towards Decadal Plan Goals: Advances and Challenges in Semiconductor Hardware*

**Vijaykrishnan Narayanan**  
**Pennsylvania State University**

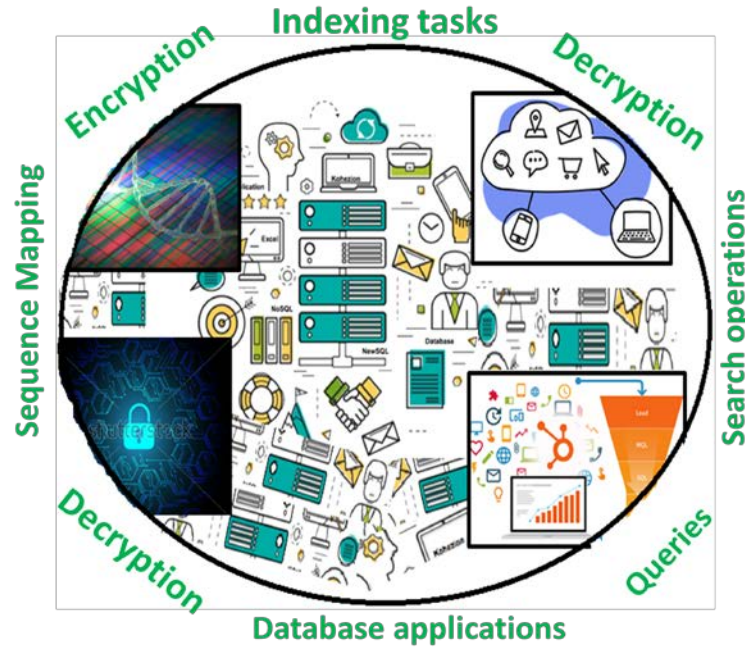


# Evolution from Compute-to-Memory Centric Systems

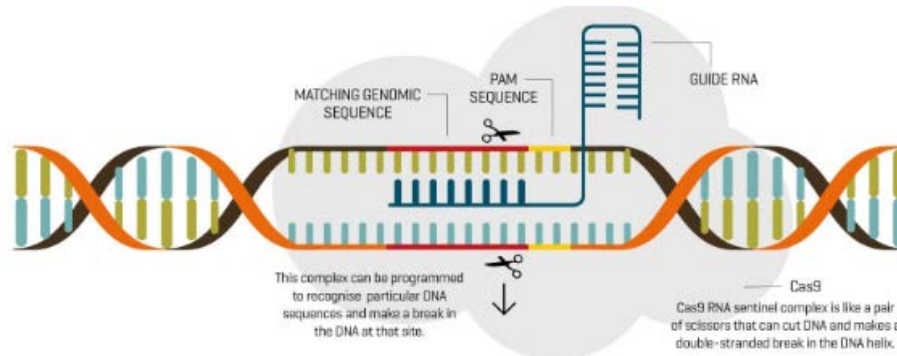
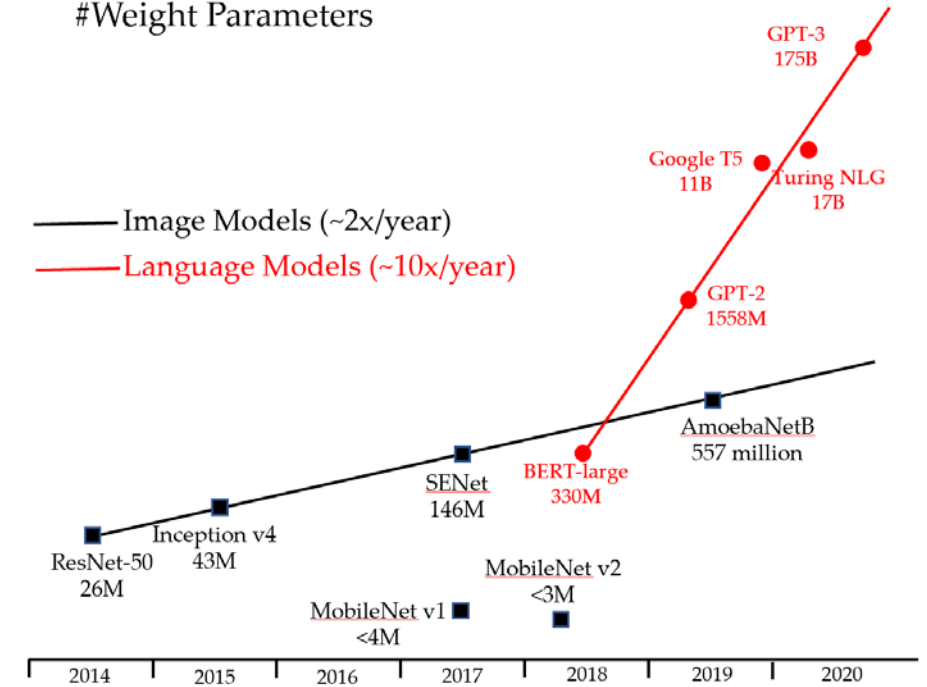
4 million billion bytes of data to Image a single black hole



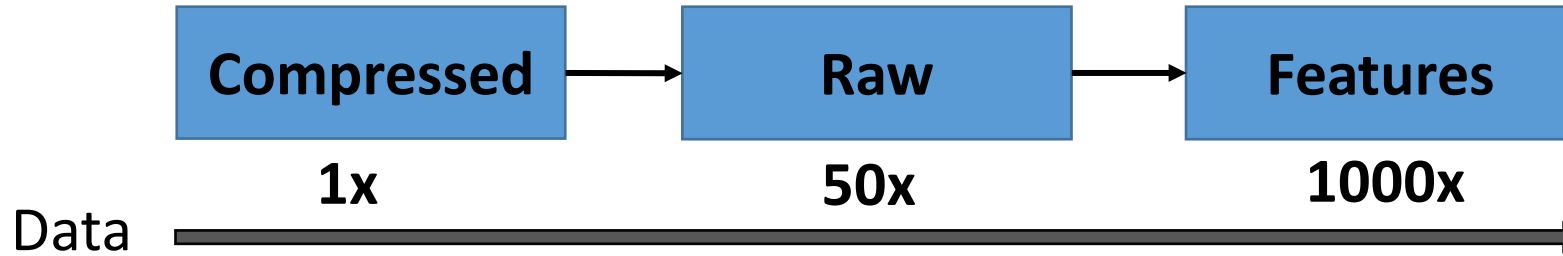
2019: The time is now for memory centric designs



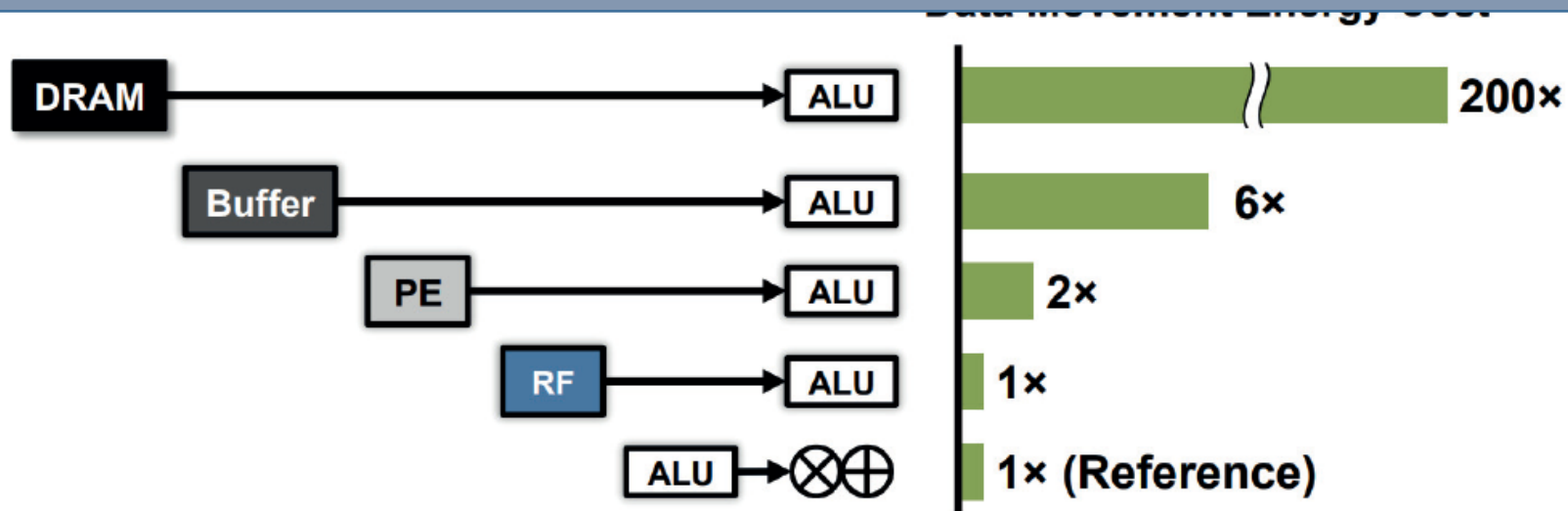
#Weight Parameters



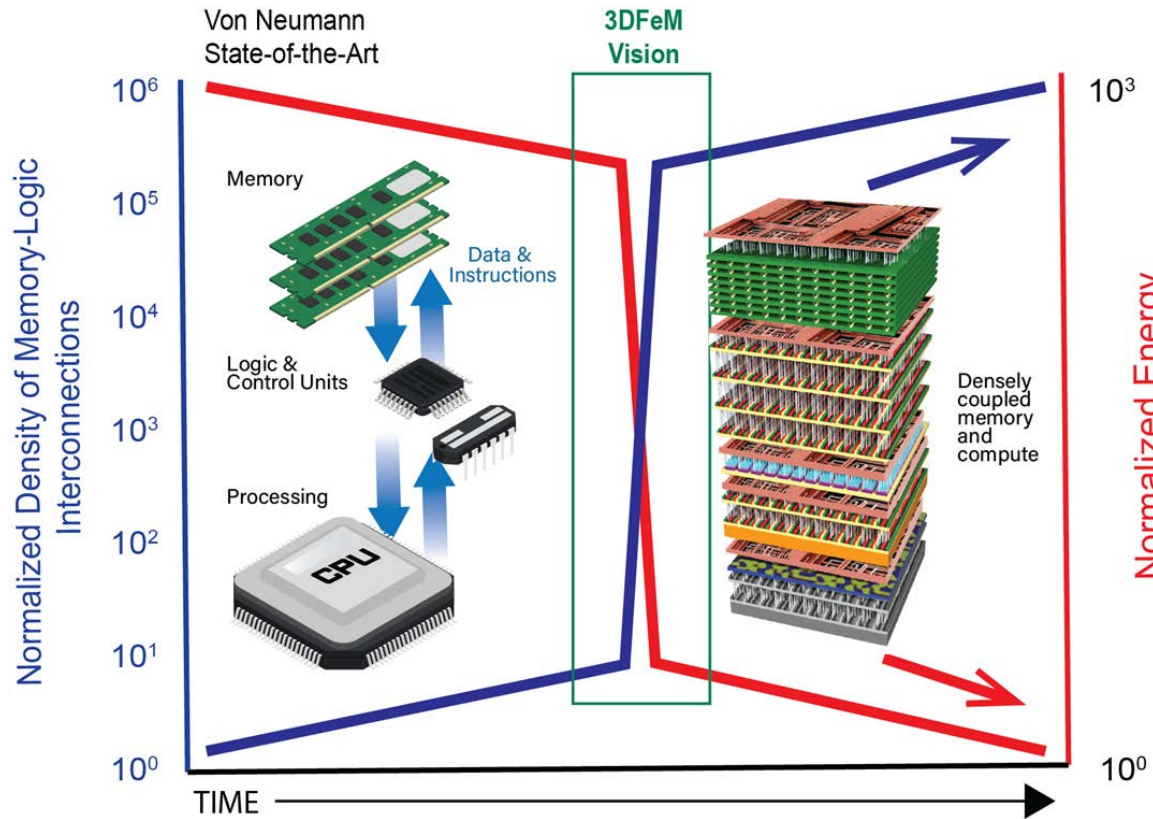
# Latency-Storage Tradeoff



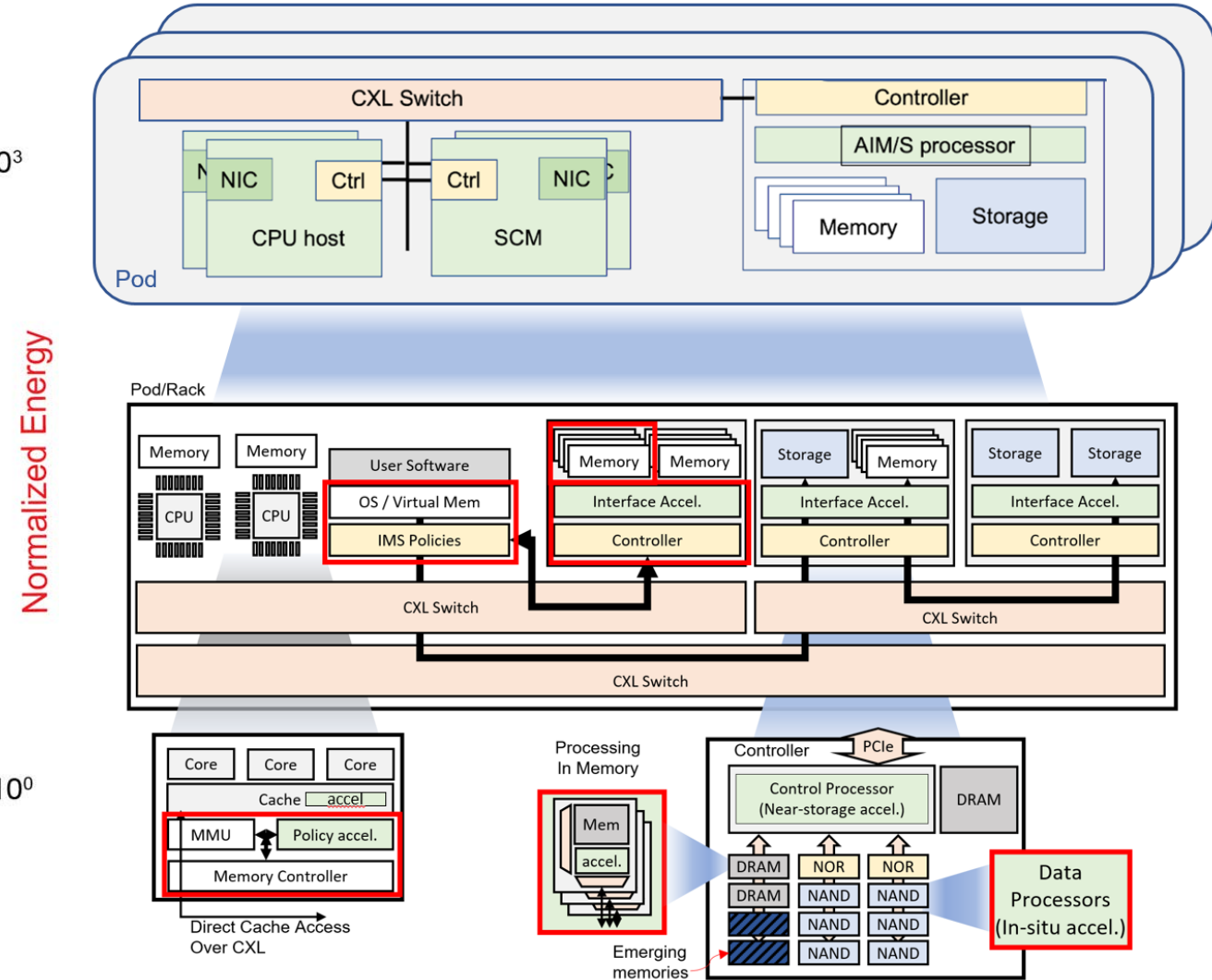
Efficient algorithms, Hardware Acceleration, High density memory/storage, Compute near memory/storage, 3D Integration



# Enabling Data-Centric Systems



Source: DOE 3DFeM

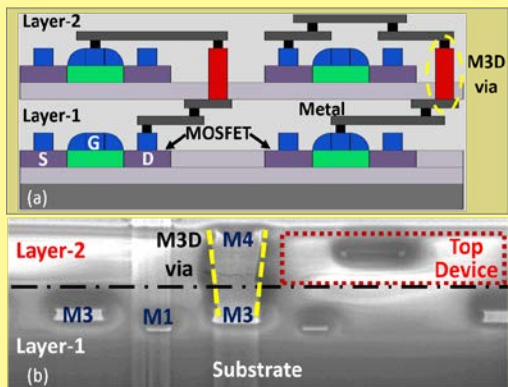


Source: Tajana Simunic

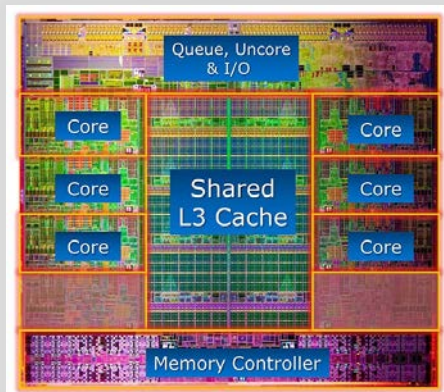
# Technology-System Interactions

Technology

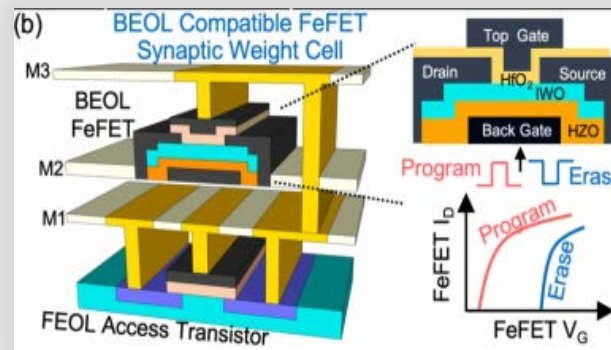
## Monolithic 3D (M3D) Integration Process



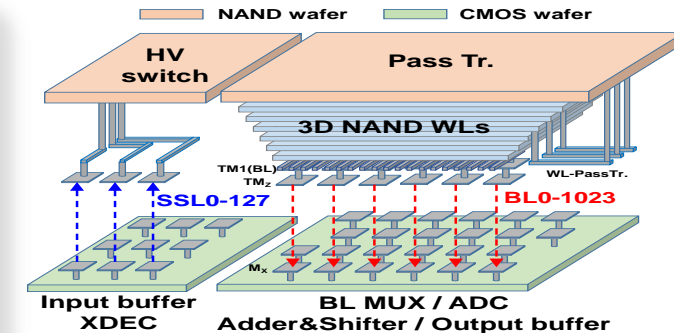
## Conventional SRAM



## M3D Integration Process + Ferro-electric FET (FeFET)

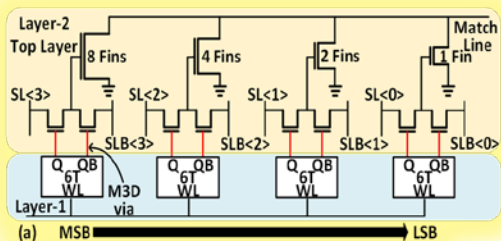


## In Storage Compute



PIM Approach

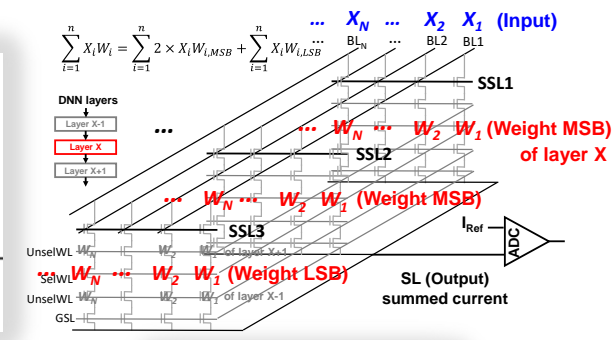
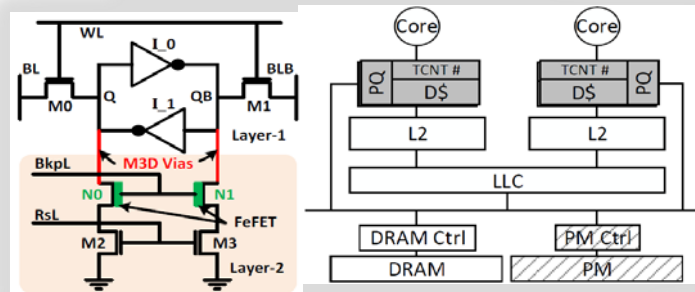
## In-Memory Comparator



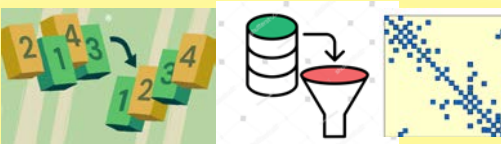
## LUT-based Compute

| Operand         | 3 <sub>d</sub>  | 5 <sub>d</sub>  | 7 <sub>d</sub>   | 9 <sub>d</sub>   | 11 <sub>d</sub>  | 13 <sub>d</sub>  | 15 <sub>d</sub>  |
|-----------------|-----------------|-----------------|------------------|------------------|------------------|------------------|------------------|
| 3 <sub>d</sub>  | 9 <sub>d</sub>  | 15 <sub>d</sub> | 21 <sub>d</sub>  | 27 <sub>d</sub>  | 33 <sub>d</sub>  | 39 <sub>d</sub>  | 45 <sub>d</sub>  |
| 5 <sub>d</sub>  | 15 <sub>d</sub> | 25 <sub>d</sub> | 35 <sub>d</sub>  | 45 <sub>d</sub>  | 55 <sub>d</sub>  | 65 <sub>d</sub>  | 75 <sub>d</sub>  |
| 7 <sub>d</sub>  | 21 <sub>d</sub> | 35 <sub>d</sub> | 49 <sub>d</sub>  | 63 <sub>d</sub>  | 77 <sub>d</sub>  | 91 <sub>d</sub>  | 105 <sub>d</sub> |
| 9 <sub>d</sub>  | 27 <sub>d</sub> | 45 <sub>d</sub> | 63 <sub>d</sub>  | 81 <sub>d</sub>  | 99 <sub>d</sub>  | 117 <sub>d</sub> | 135 <sub>d</sub> |
| 11 <sub>d</sub> | 33 <sub>d</sub> | 55 <sub>d</sub> | 77 <sub>d</sub>  | 99 <sub>d</sub>  | 121 <sub>d</sub> | 143 <sub>d</sub> | 165 <sub>d</sub> |
| 13 <sub>d</sub> | 39 <sub>d</sub> | 65 <sub>d</sub> | 91 <sub>d</sub>  | 117 <sub>d</sub> | 143 <sub>d</sub> | 169 <sub>d</sub> | 195 <sub>d</sub> |
| 15 <sub>d</sub> | 45 <sub>d</sub> | 75 <sub>d</sub> | 105 <sub>d</sub> | 135 <sub>d</sub> | 165 <sub>d</sub> | 195 <sub>d</sub> | 225 <sub>d</sub> |
| 7 <sub>d</sub>  | 21 <sub>d</sub> | 35 <sub>d</sub> | 49 <sub>d</sub>  | 63 <sub>d</sub>  | 77 <sub>d</sub>  | 91 <sub>d</sub>  | 105 <sub>d</sub> |

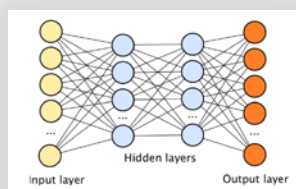
## Non-Volatile Cache



Apps



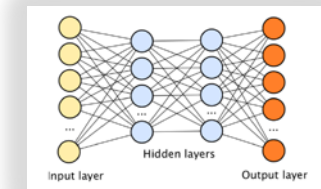
## Sort, Data Filtering, SpGEMM



## Machine Learning

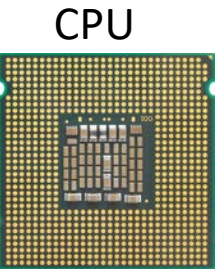


## Persistent Applications

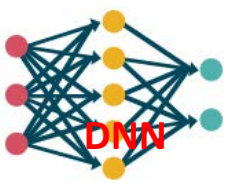


## Machine Learning

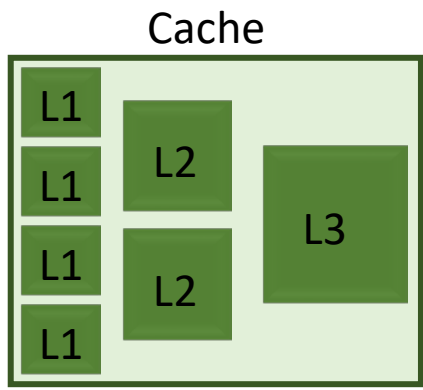
# Compute-Memory-Storage Hierarchy



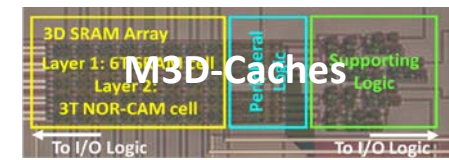
FPGA  
ASIC  
Sensors



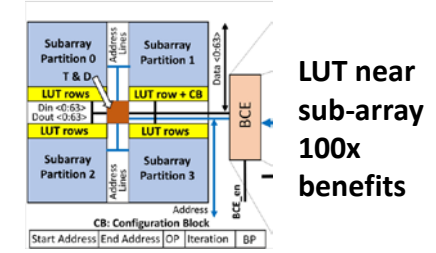
Computer  
Vision



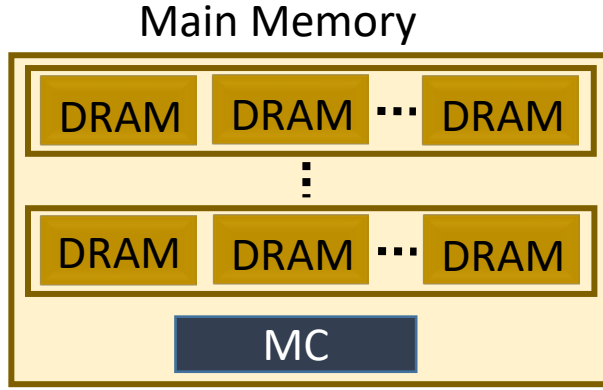
Multi-dimensional  
Caches  
In Mem Compute  
Content-based Retrieval  
In Mem Sorting



Compute Support, 50x benefits



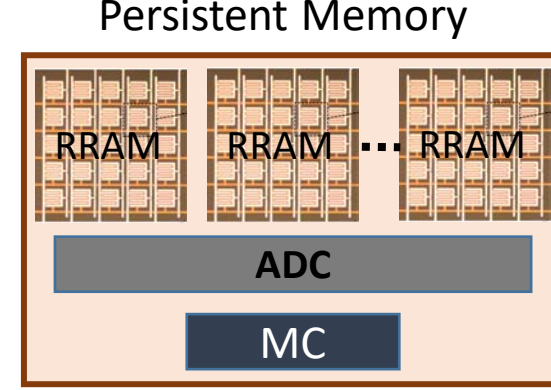
LUT near  
sub-array  
100x  
benefits



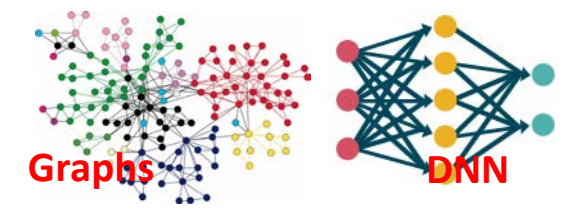
Near Mem Compute  
CPU cores – DRAM Bank



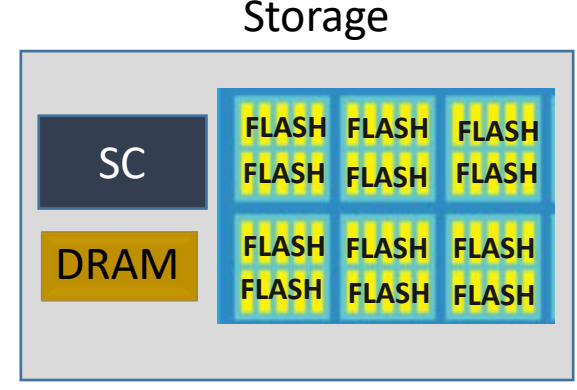
200x benefits



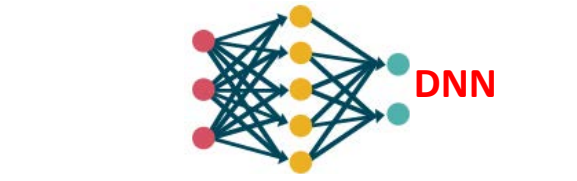
In Mem Compute  
Analog and Neuro Computing  
Energy Harvesting  
Content-based Retrieval



Visual Analytics  
500x to 2000x benefits



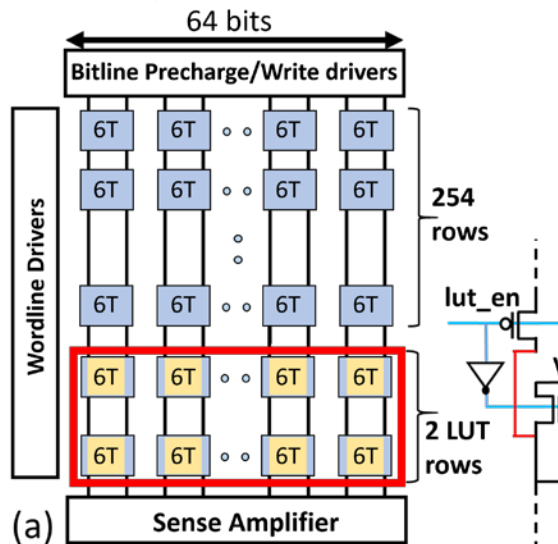
Near Storage Compute  
LUT near Flash Chips



Visual Analytics  
10000x benefits

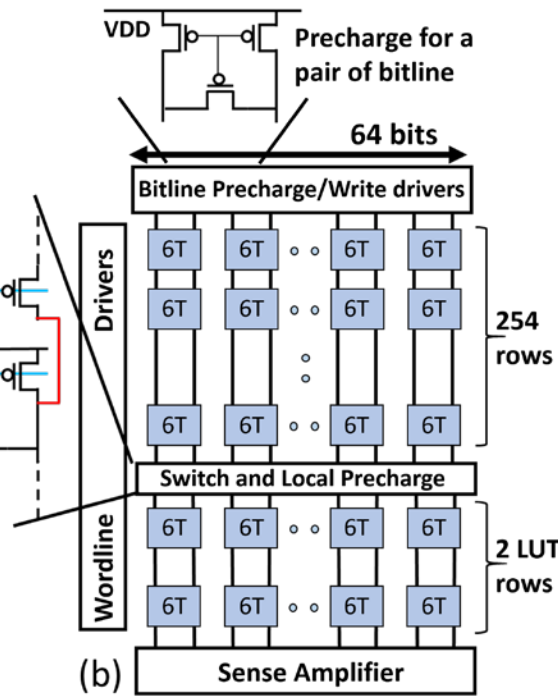
# BFree Architecture: Sub-array with Reduced Access LUT rows and Compute Engine (BCE)

Conventional subarray partition

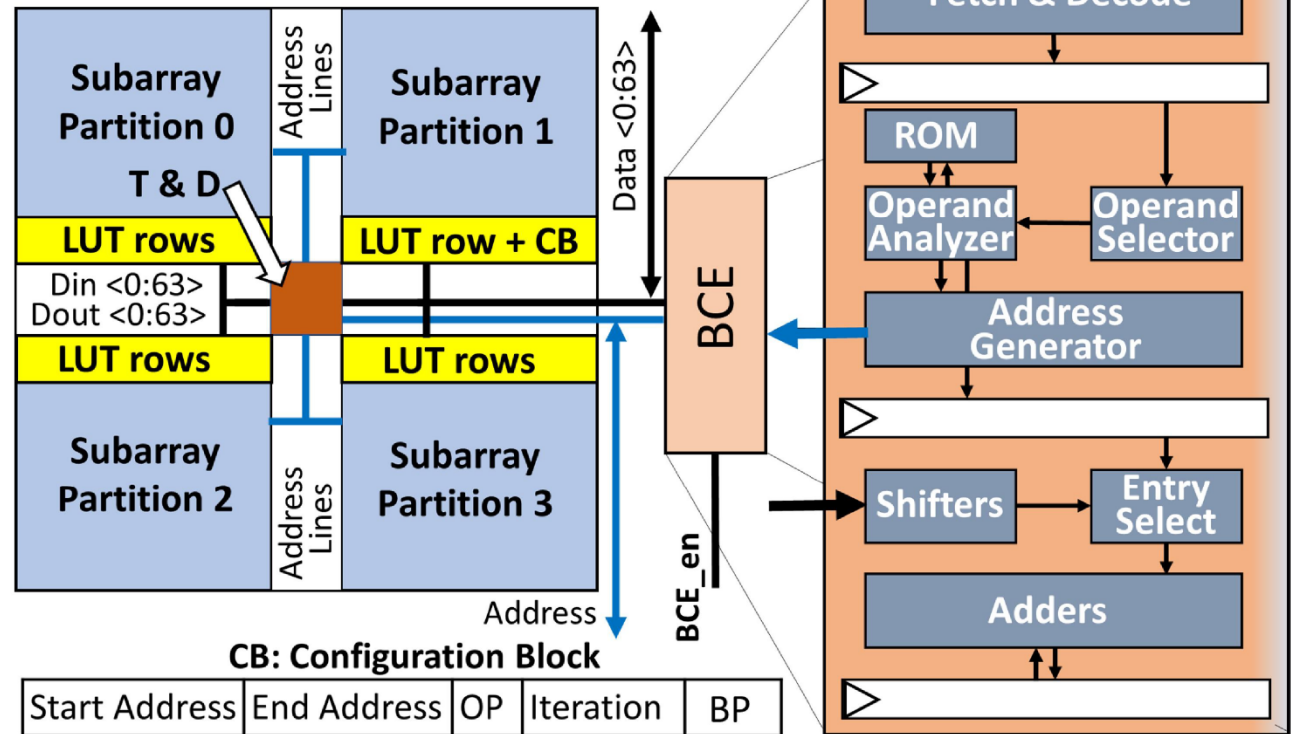


| Design | Latency (ps) | Energy (pJ) |
|--------|--------------|-------------|
| (a)    | 654          | 13.9        |
| (b)    | 192          | 0.060       |

Subarray partition with reduced access cost for LUTs

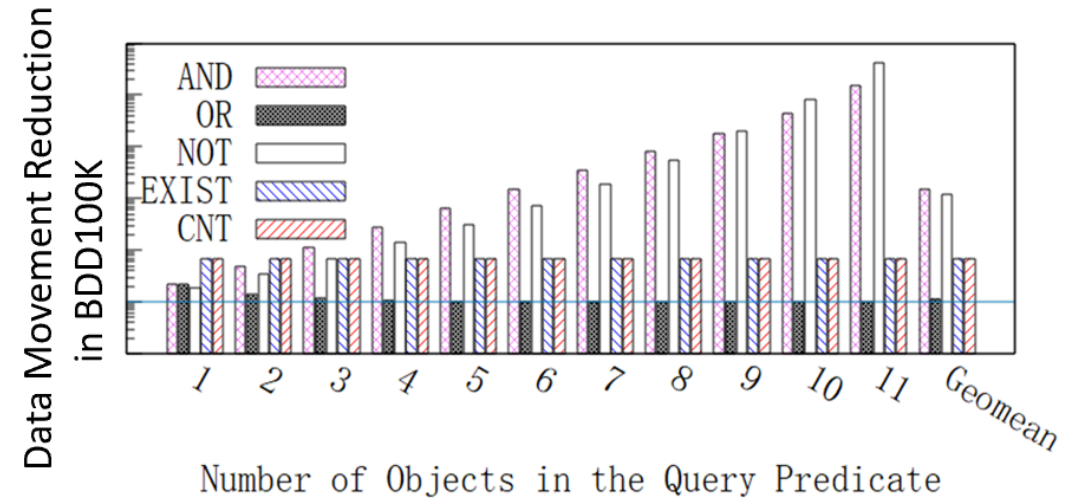
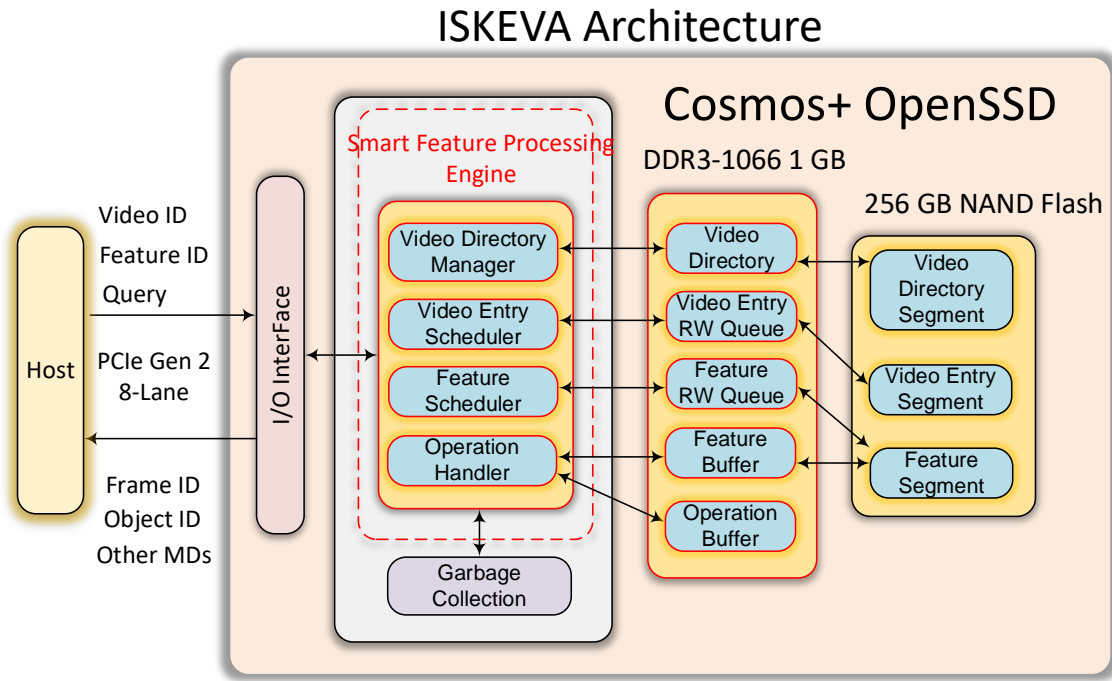


Subarray with the BFree Compute Engine (BCE)



BCE is a 3-stage pipelined in-order core, placed at the sub-array level. BCE is connected to the timer and decoder ports of the sub-array without perturbing the custom-built sub-arrays.

# Near Storage Processing



Great potential, but many more challenges

- Programming Ease
- Scalability
- Security
- Endurance
- Power/Thermal



# CROSS-LAYER DESIGN FROM DEVICES TO APPLICATIONS

**X. Sharon Hu, University of Notre Dame**

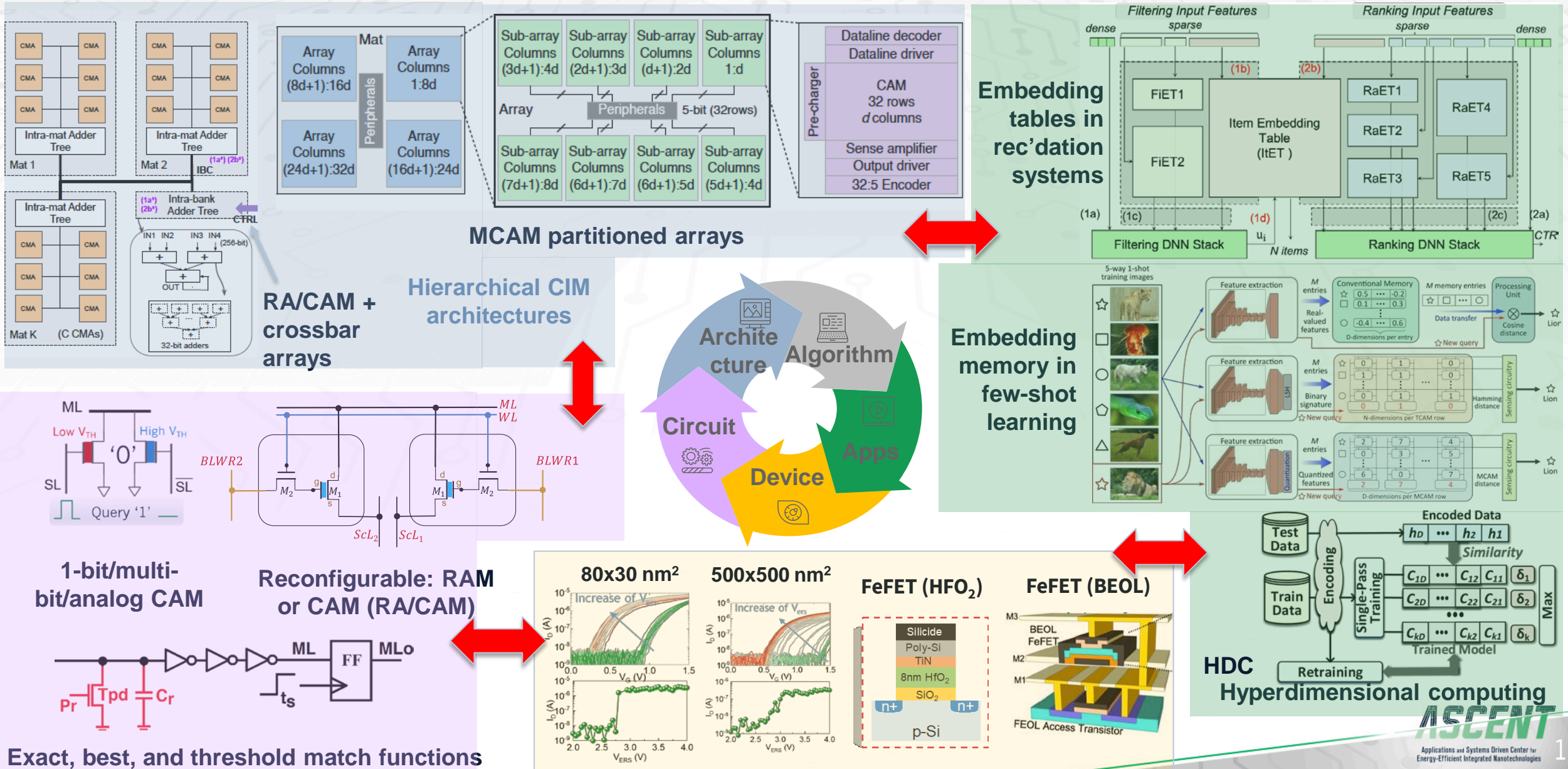
**Faculty collaborators:** Michael Niemier (notre Dame), Suman Datta (Notre Dame), Mohsen Imani (UCI), Kai Ni (RIT), Thomas Kampfe (Fraunhofer IPMS-CNT, Germany)

**Students at Notre Dame:** Arman Kazem, Ann Franchesca Laguna, Liu Liu, Mohammad Mehdi Sharifi

The logo for ASCENT, where the letters are stylized with circuit traces and nodes. The word "ASCENT" is written in a bold, green, sans-serif font. A green diagonal line cuts across the letters from the bottom left to the top right.

Applications and Systems Driven Center for  
Energy-Efficient Integrated Nanotechnologies

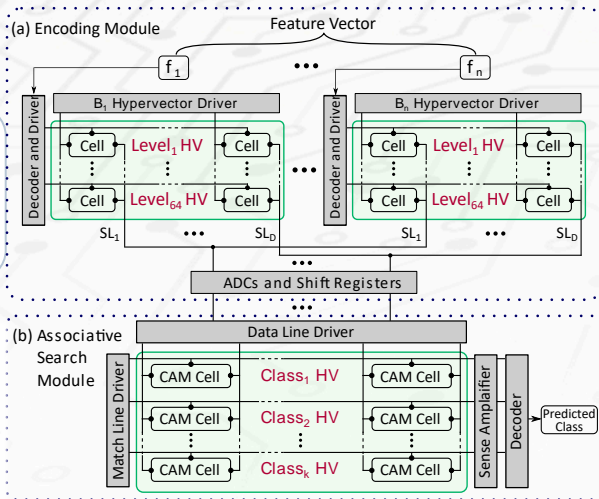
# CROSS-LAYER DESIGN: FEFET-BASED COMPUTE-IN-MEMORY FABRICS



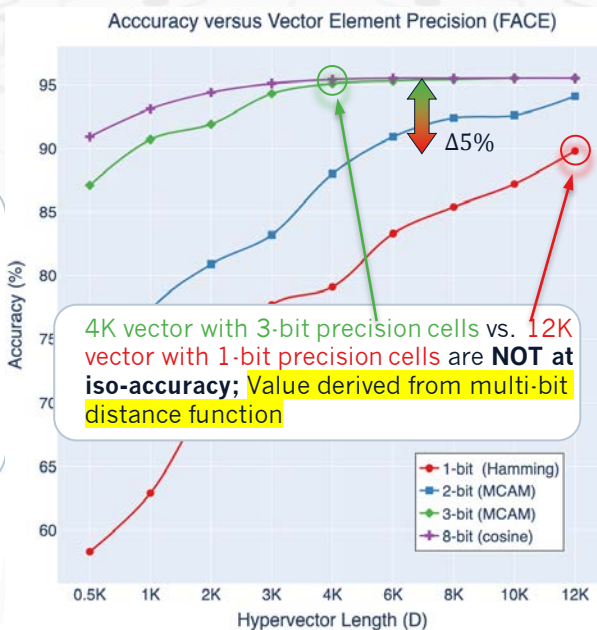
# VALUE PROPOSITION OF MULTI-BIT FEFETS?

NEED CROSS-LAYER ANALYSIS TO ANSWER!

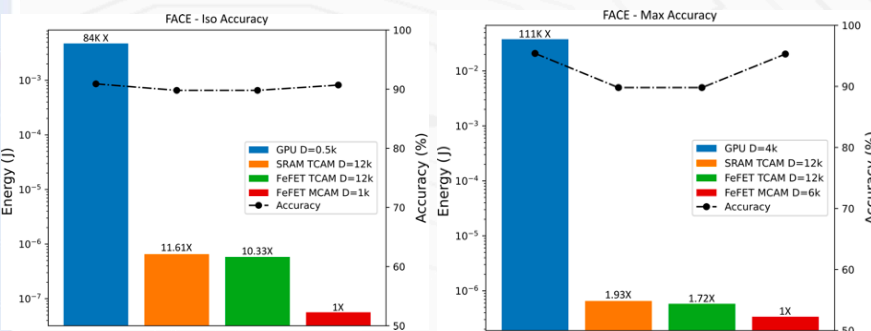
## 1 Crossbar, CAM HDC architecture



## 2 How is accuracy impacted by CAM-based distance function?



## 3 End-to-end energy comparison: value proposition

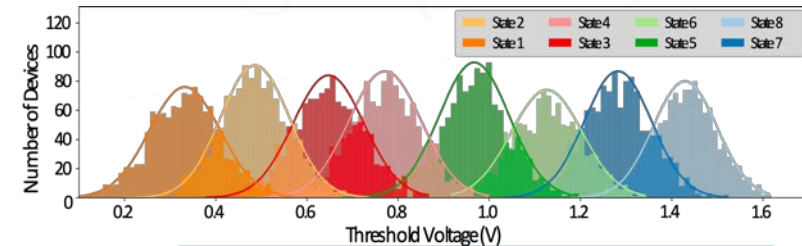
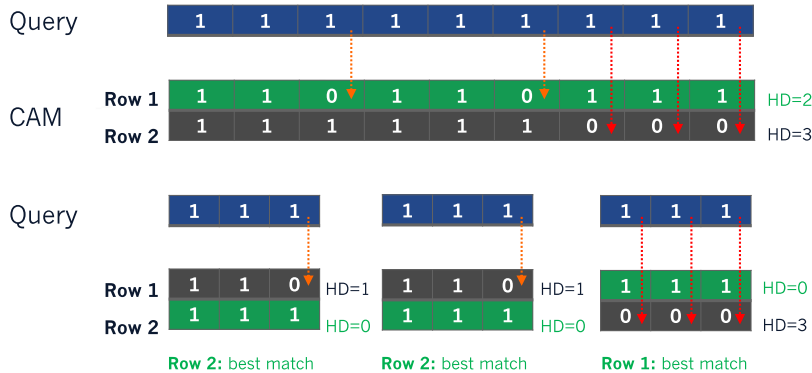


10X gain in energy efficiency by MCAM at iso accuracy

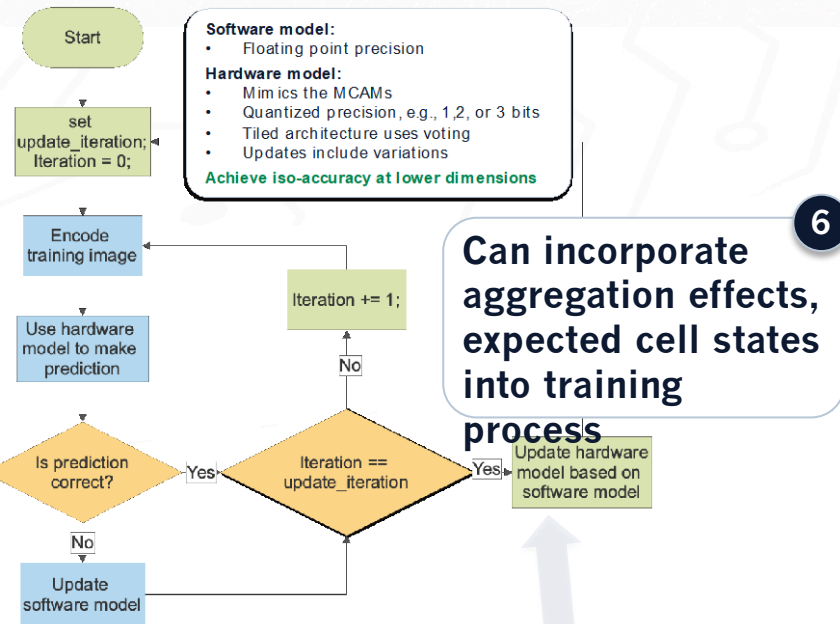
1.7X gain in energy efficiency by MCAM at max accuracy

Technology-based solutions needed for iso-accuracy

## 5 What is the impact of a realistic associative memory architecture?



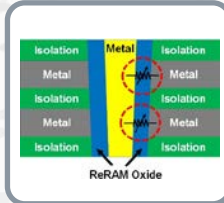
## 4 What is the impact of realistic, cell state distributions? What σ of variation tolerable?



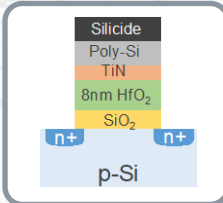
Can incorporate aggregation effects, expected cell states into training process

# DEVICE-ARCHITECTURE-ALGORITHM DSEs

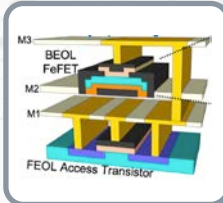
RRAM



Si-FeFET



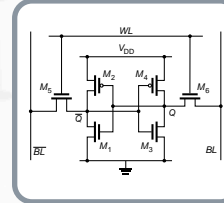
BEOL FeFET



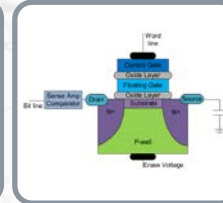
SOT-MRAM



CMOS



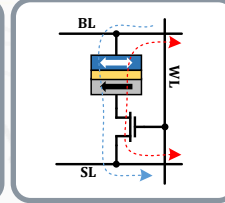
Flash



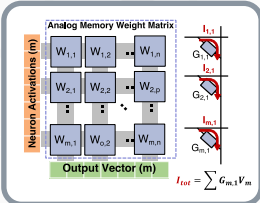
PCM



STT-MRAM



Crossbars



TBD

40X energy  
50X latency  
Iso-accuracy

TBD

TBD

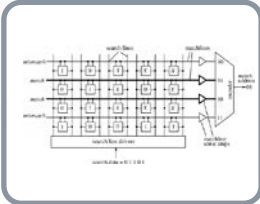
TBD

TBD

TBD

TBD

CAMs



Accuracy drop with Hamming distance

40X energy  
50X latency  
Iso-accuracy

TBD

TBD

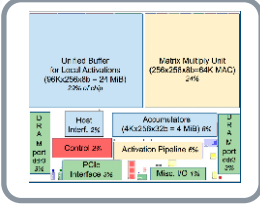
TBD

TBD

TBD

TBD

TPUs



TBD

TBD

TBD

TBD

TBD

TBD

TBD

TBD

GPUs



TBD

TBD

TBD

TBD

TBD

TBD

TBD

TBD

**Tools to support cross-layer modeling, design and optimization**

Non-HDC, other DNN, non-DNN algorithms?