Plenty of room at the top and the bottom

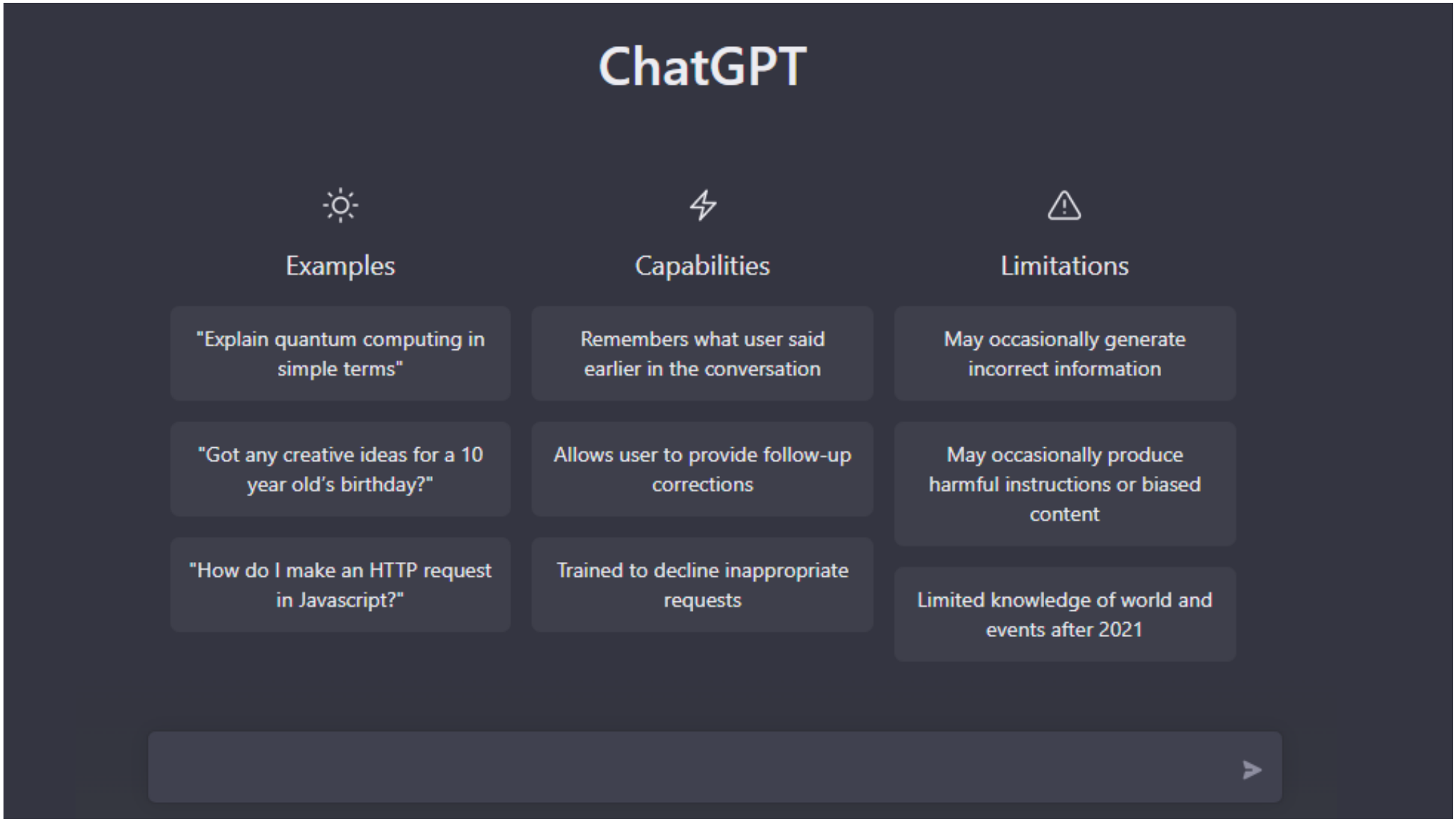
Suman Datta
Joseph M. Pettit Chair in Advanced Computing
Georgia Research Alliance Eminent Scholar

Asif Khan
Assistant Professor
onsemi Junior Professor

School of Electrical and Computer Engineering
and of Materials Science and Engineering
sdatta68@gatech.edu akhan40@gatech.edu

Georgia Tech Chips Day, 5/2/2023

Georgia Tech


ChatGPT
Examples
"Explain quantum computing in simple terms"
"Got any creative ideas for a 10 year old's birthday?"
"How do I make an HTTP request in Javascript?"
Capabilities
Remembers what user said earlier in the conversation
Allows user to provide follow-up corrections
Trained to decline inappropriate requests
Limitations
May occasionally generate incorrect information
May occasionally produce harmful instructions or biased content
Limited knowledge of world and events after 2021

Model: Default (GPT-3.5)

AK Tell me an interesting fact about UIUC.

Model: Default (GPT-3.5)

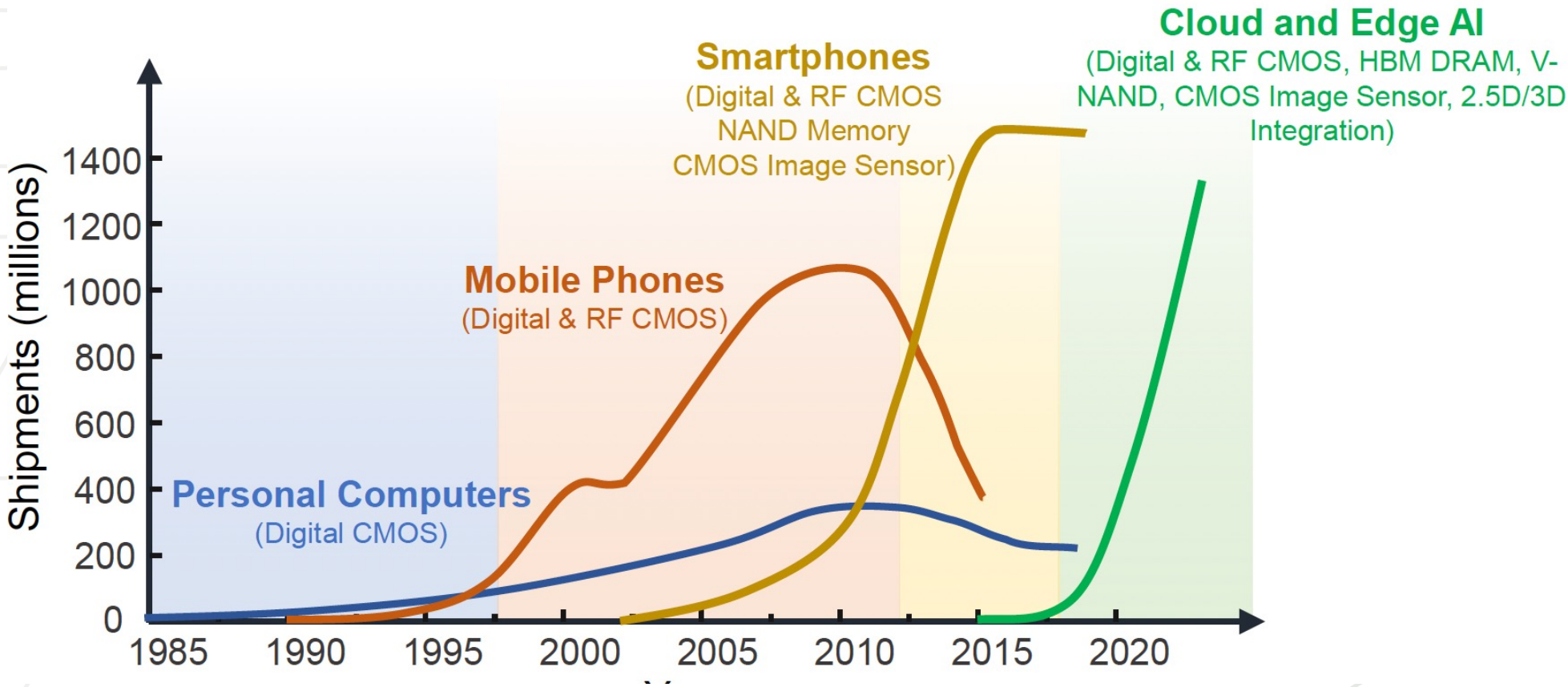
AK I am going to UIUC for a seminar. Which is the nearest airport within 50 miles? I cannot use CMI because there is no direct flight from ATL.

Model: Default (GPT-3.5)

AK I will be giving an seminar at UIUC on microelectronics and ferroelectrics tomorrow. Can you suggest me five interesting opening line for me?

Stop generating

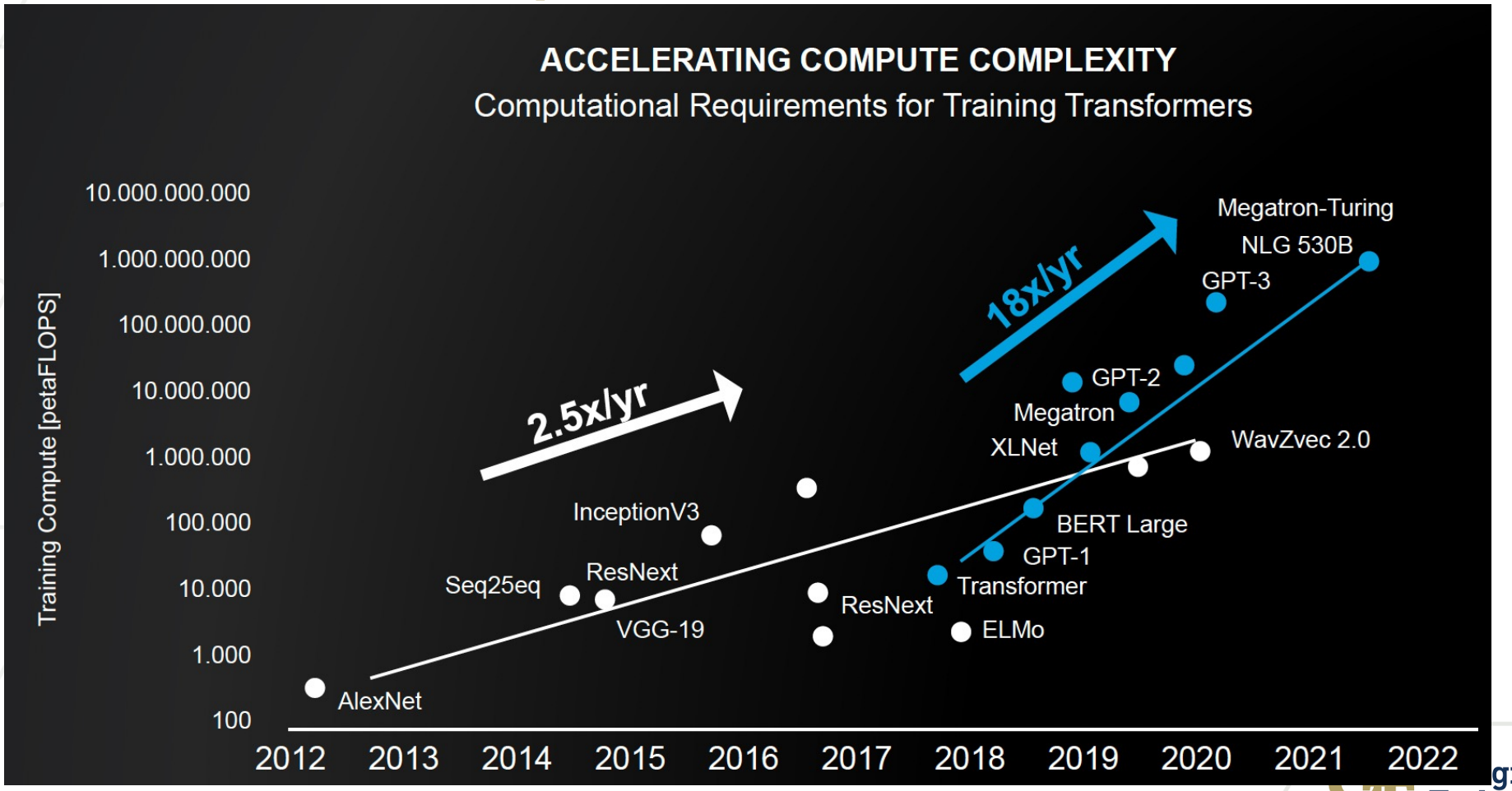
How the computing paradigms have evolved



Adapted from: Datta, EDTM 2023
https://www.economist.com/briefing/2015/02/26/the-truly-personal-computer
https://www.statista.com/chart/12798/global-smartphone-shipments/

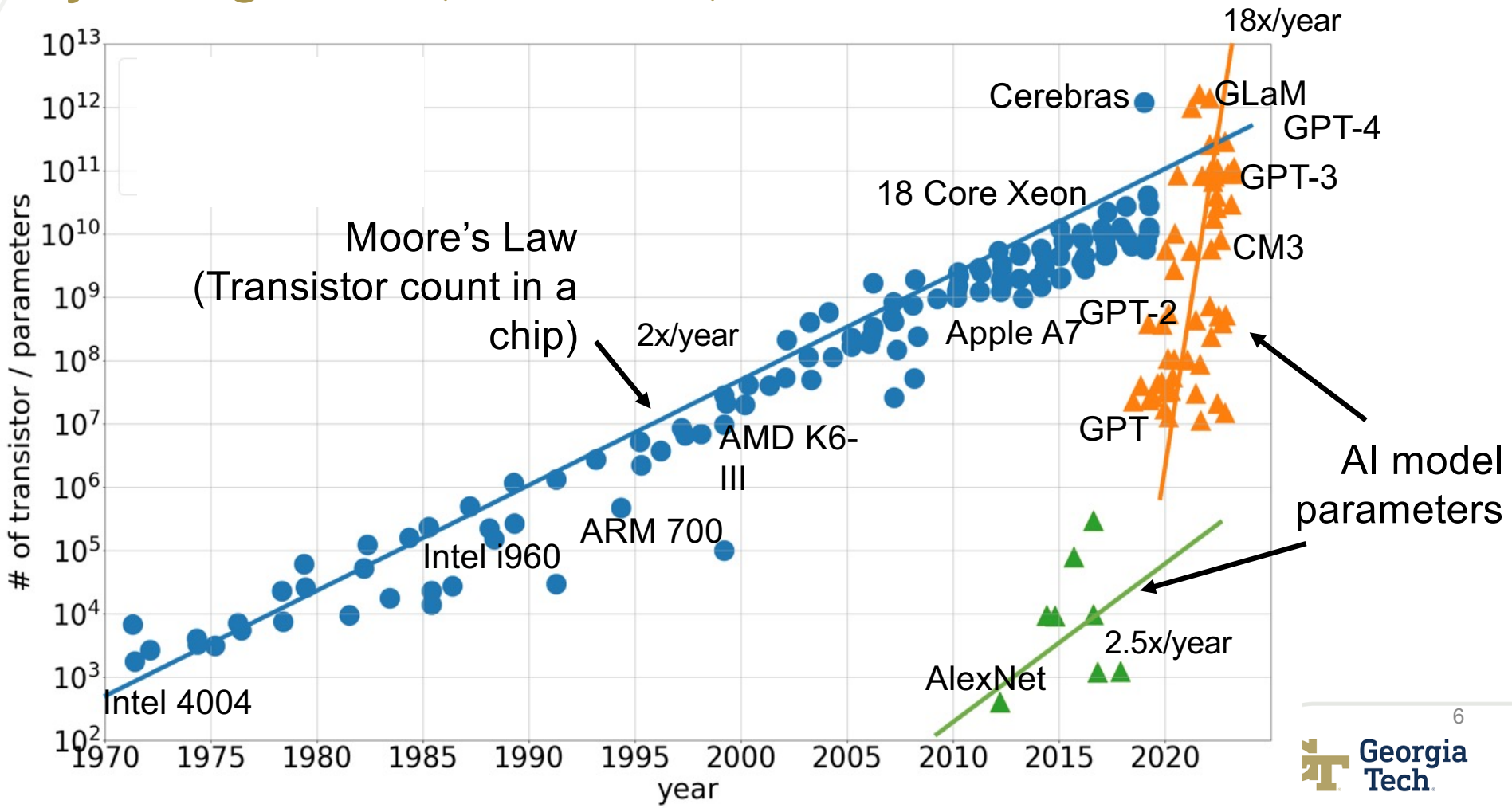
Georgia Tech

Evolution of AI model parameter count



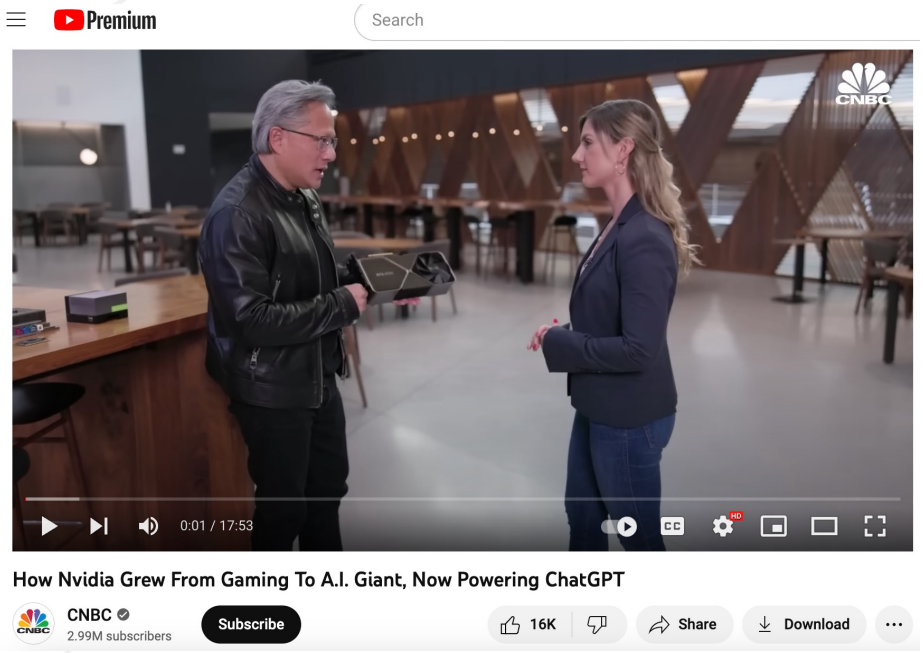
Source: NVIDIA

A story of algorithm, software, and hardware

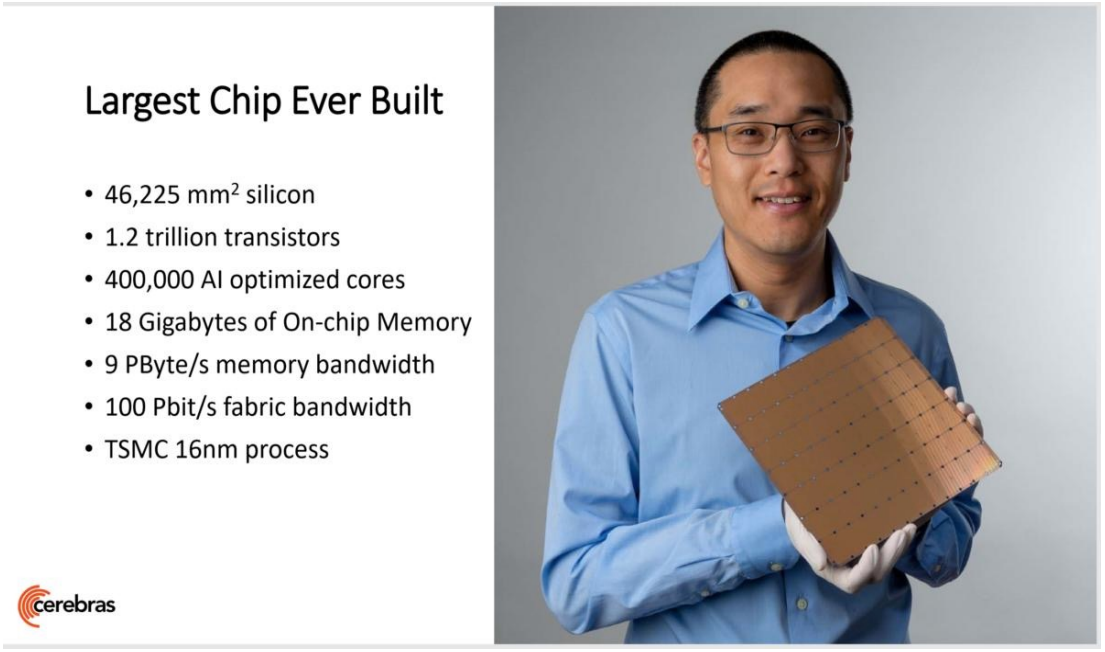


Georgia Tech

Hardware that powers the AI

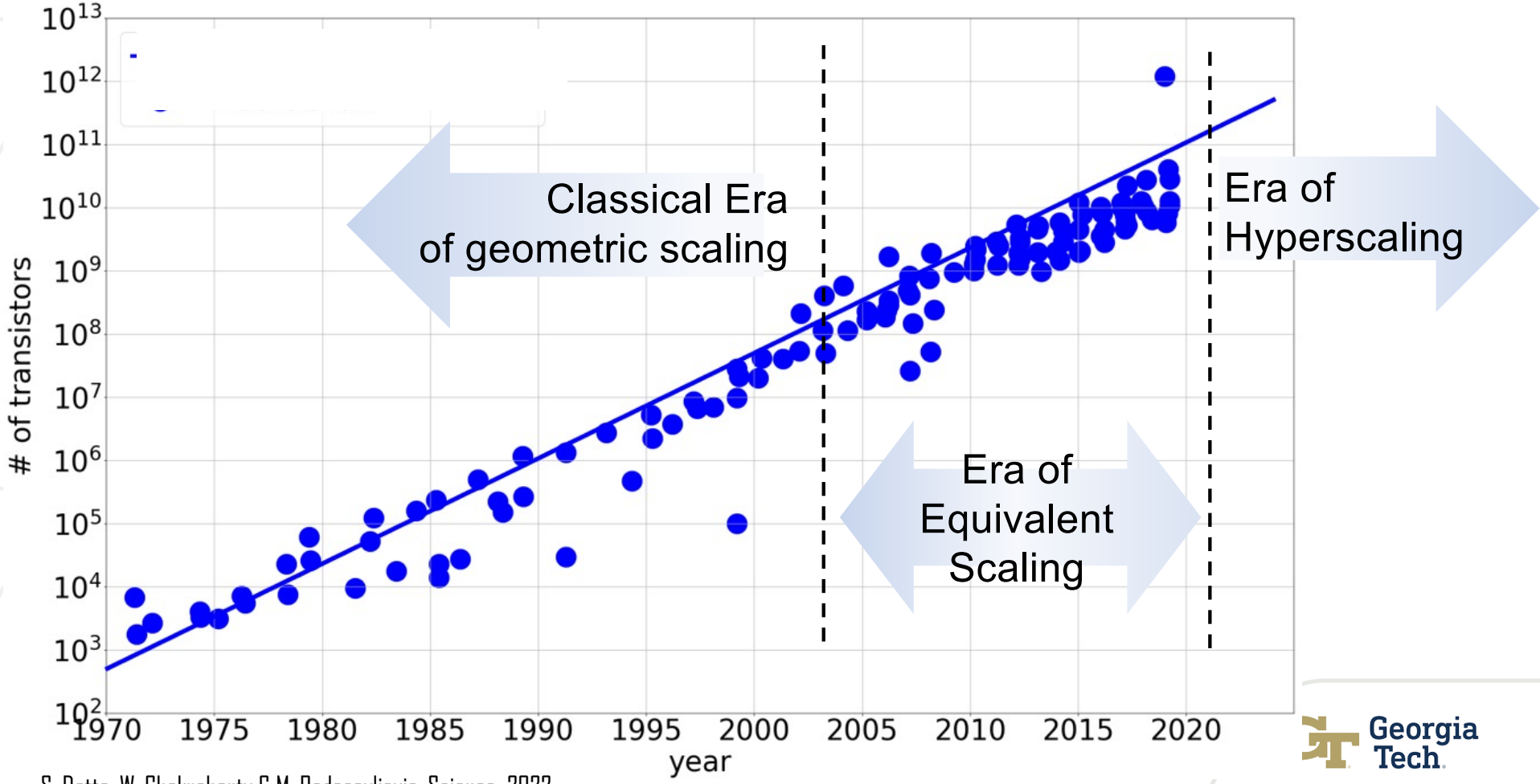


https://www.youtube.com/watch?v=d3L2uPuxOxU



https://www.pcworld.com/article/397925/cerebras-systems-new-deep-learning-chip-is-as-big-as-your-keyboard-and-the-largest-ever.html

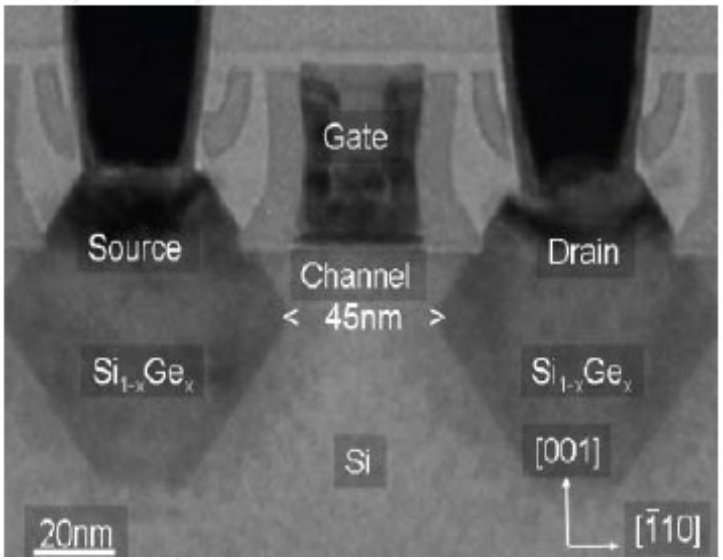
A brief history of scaling

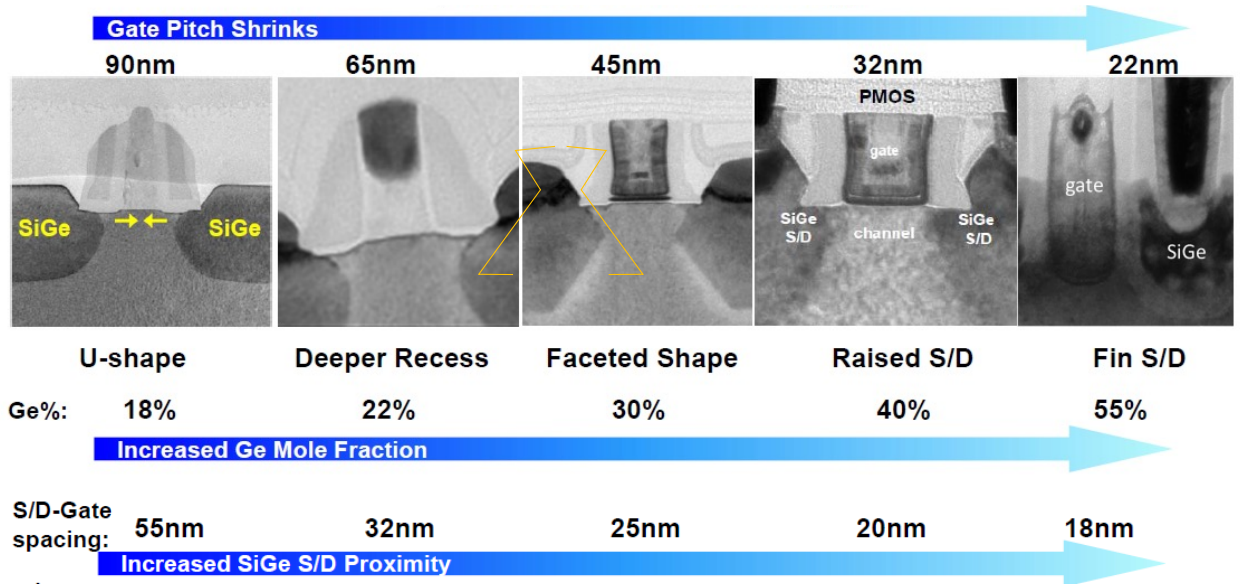


S. Datta, W. Chakraborty & M. Radosavljevic, Science 2022

Georgia Tech

Serendipity-1: Mobility enhancement using $Si_{1-x}Ge_x$ source-drain

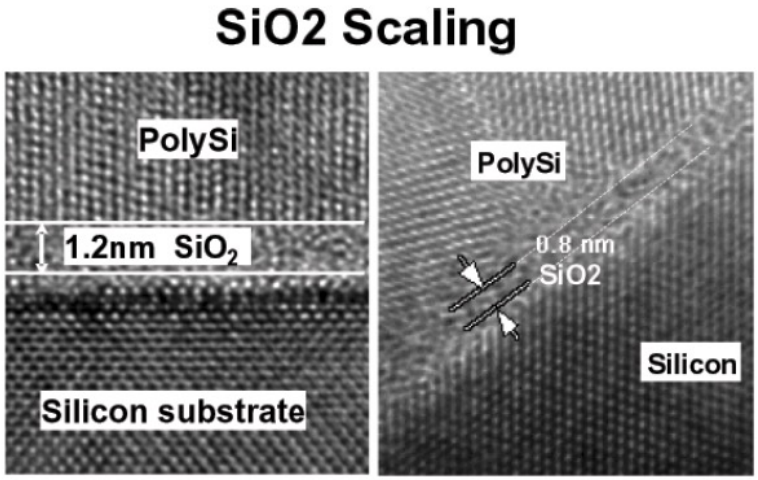




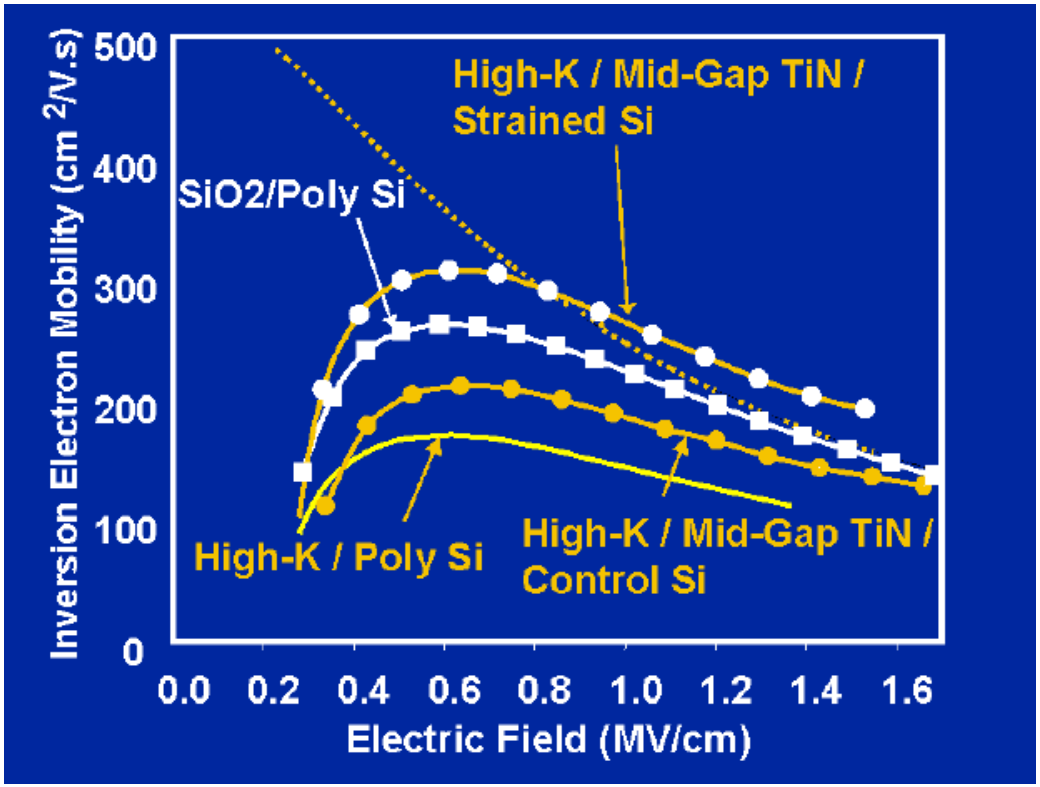
Embedded SiGe S/D reduces external resistance and imposes longitudinal uniaxial compressive strain leading to an 40% or higher enhancement effective hole mobility/velocity in the channel.

Georgia Tech

Serendipity-2: High-K metal gate and plasmon dynamics



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

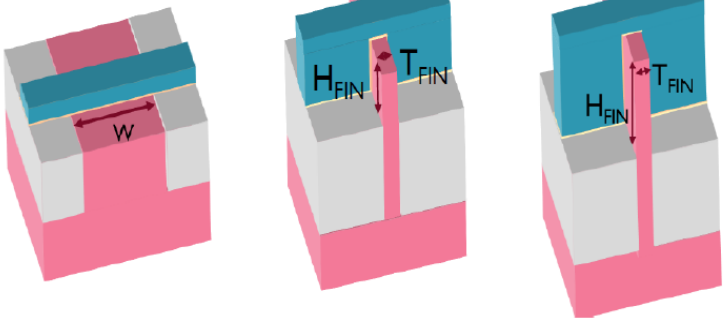


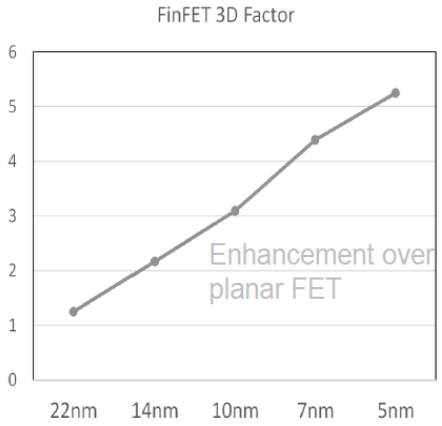
Surface phonon limited mobility recovers with high-k / metal gate

Combination of strained channel, high-k dielectric and metal gate recovers mobility

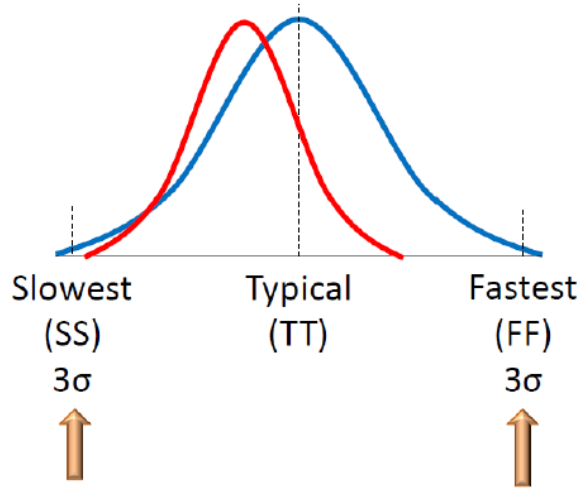
Georgia Tech

Serendipity-3: Reduced variation in 3D transistors (FinFETs)





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



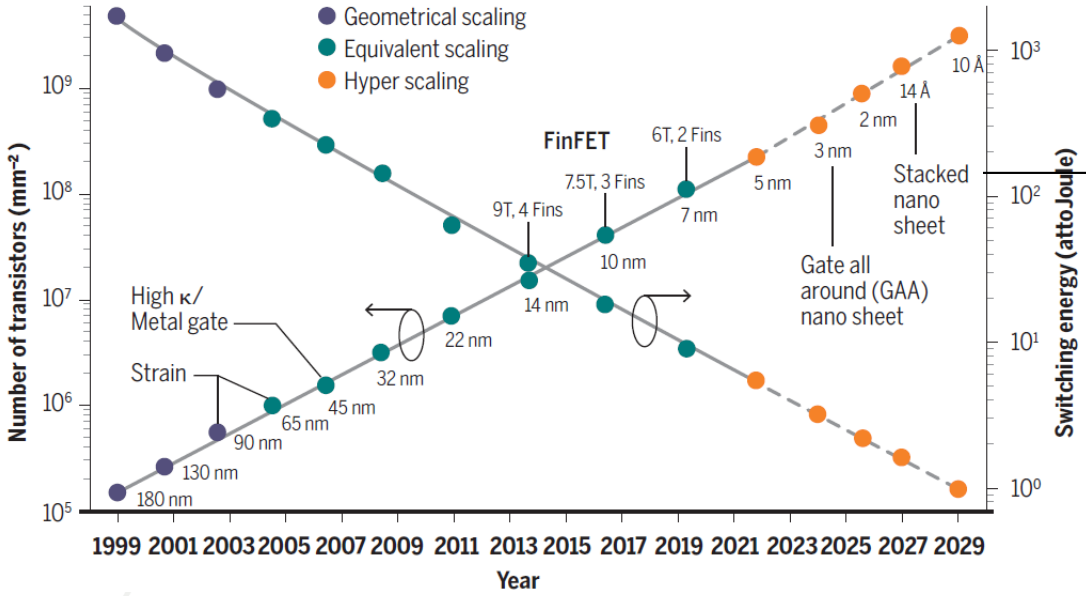
Greg Yeric (ARM), IEDM 2015

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

Tri-gate was intended primarily for electrostatics
FinFET 3D factor allows for effective width scaling with improved variation

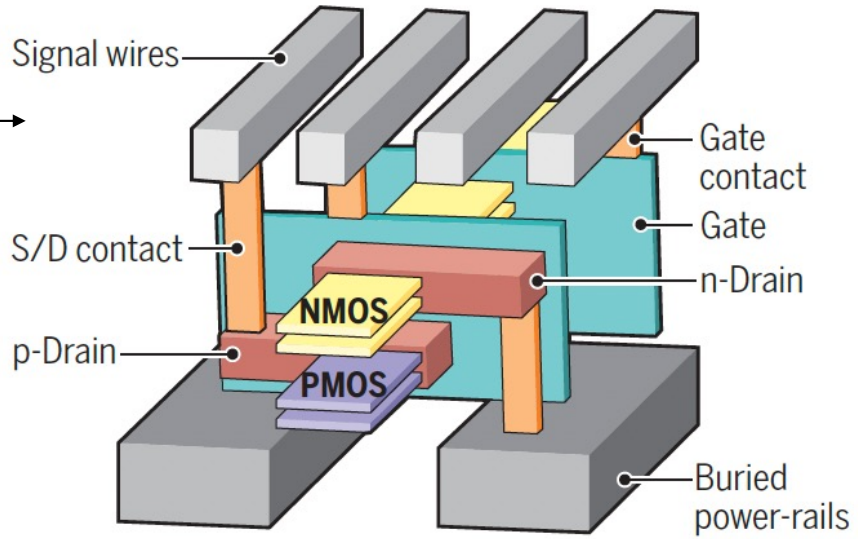
Georgia Tech

The era of hyper scaling: Plenty of room at the bottom



S. Datta, W. Chakraborty & M. Radosavljevic, Science 2022

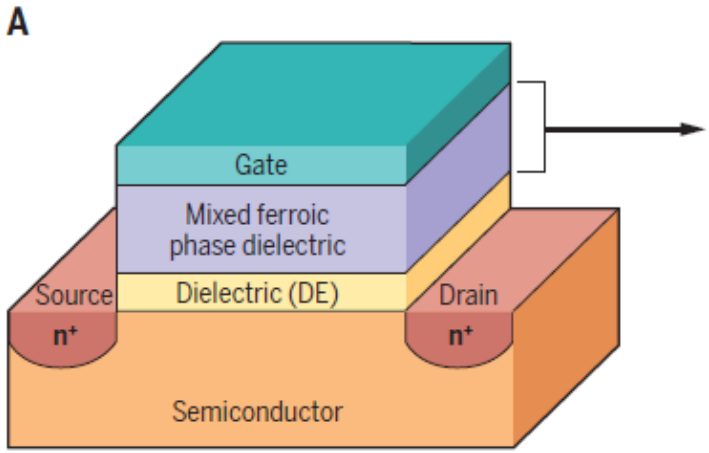
Stacked nano-sheet based inverter

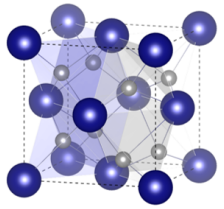


Buried Power Rails (BPR) reduce number of tracks (T).
Stacked transistors further reduce # of tracks (T).

Georgia Tech

Plenty of Room at the Bottom: Negative Capacitance FETs

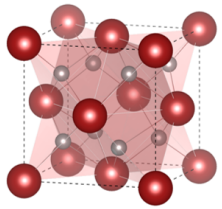




Ferroelectric (FE)
Orthorhombic
[Pca2$_1$]

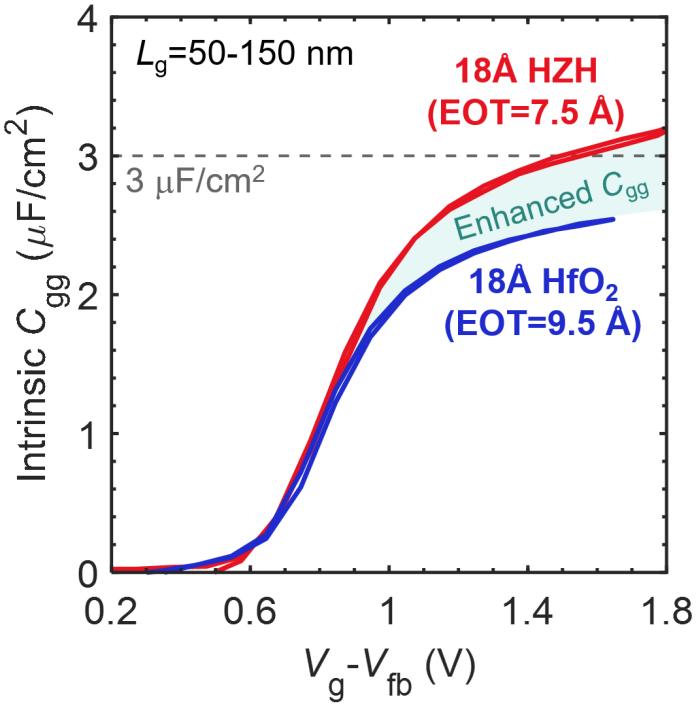
P

HfO_2



Antiferroelectric (AFE)
Tetragonal
[P4$_2$/nmc]
P=0

ZrO_2



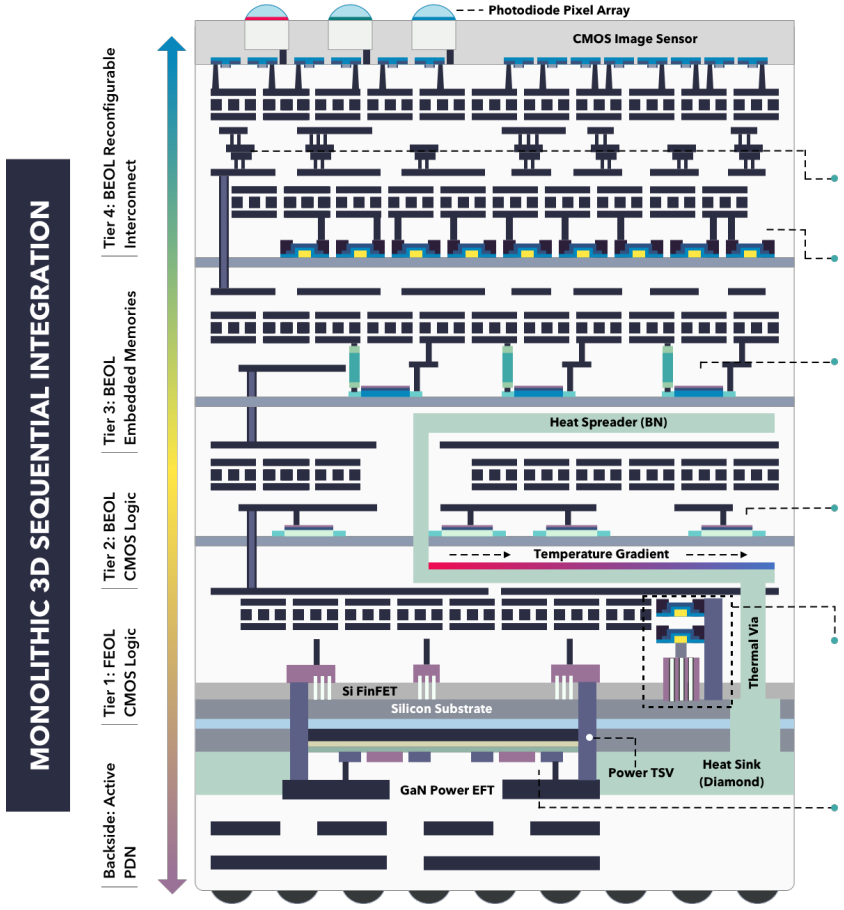
Negative capacitance FETs use a mixed phase ferroelectric (FE) and anti-ferroelectric (AFE) to enhance the permittivity for ideal kT/q switching transistors operating at low V_{DD}

S. Datta, W. Chakraborty & M. Radosavljevic, Science 2022

S. Cheema, S. Salahuddin et al, Nature 2022

Georgia Tech

Plenty of room at the bottom: BEOL and backside

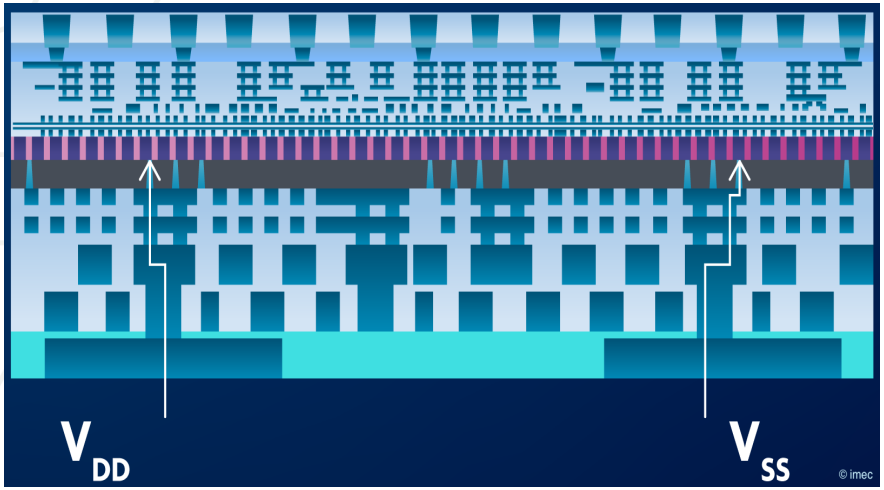


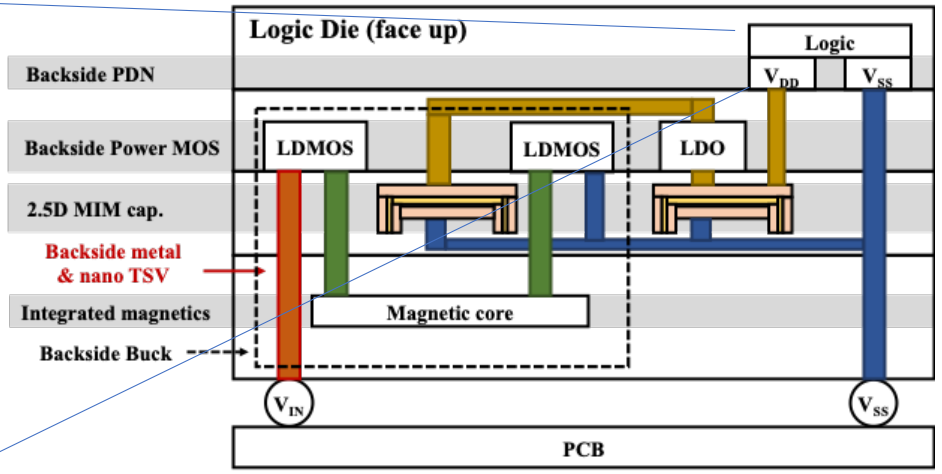
The back-end-of-the-line and the backside of the si wafer can be likened to the wild wild west.

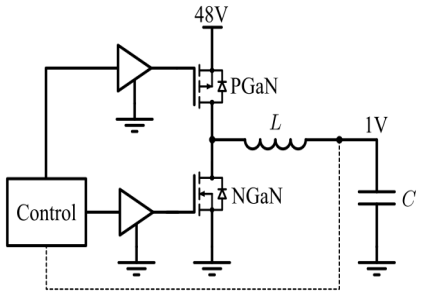
Adapted from a Center Proposal on Heterogenous Integration

Georgia Tech

Backside Power Delivery Network (BPDN)







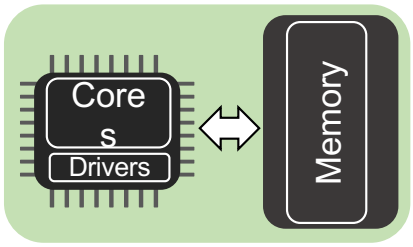
Deliver 500W to billions of transistors operating at 0.5V.

Mitigate I*R drop and electromigration.

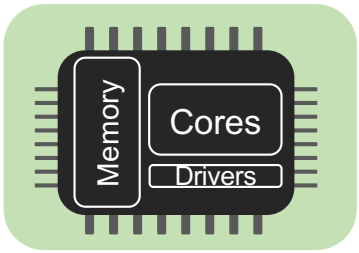
Backside PDN for power conversion will require **high-k MIM cap, high-μ inductor, high Ron/Roff ratio transistor and thermal management.**

Georgia Tech

High bandwidth, Low latency On-Chip Memory



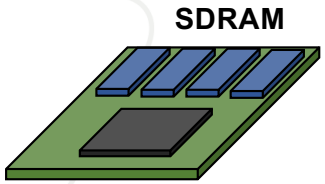
Off Chip Memory



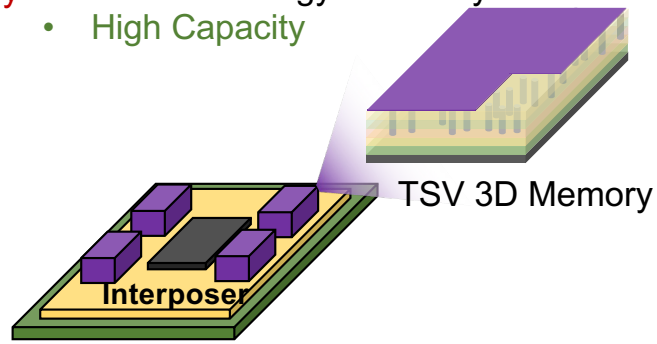
On Chip Memory

BW: ~20 Gb/s ~2 Tb/s ~10 Tb/s

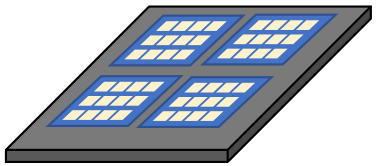
GDDR:
- Low Energy Efficiency
- High Capacity



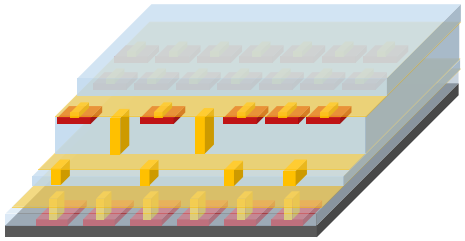
HBM 3D TSV Chip Stacking:
- Medium Energy Efficiency
- High Capacity



2D On-chip memory:
- High Energy Efficiency
- Low Capacity



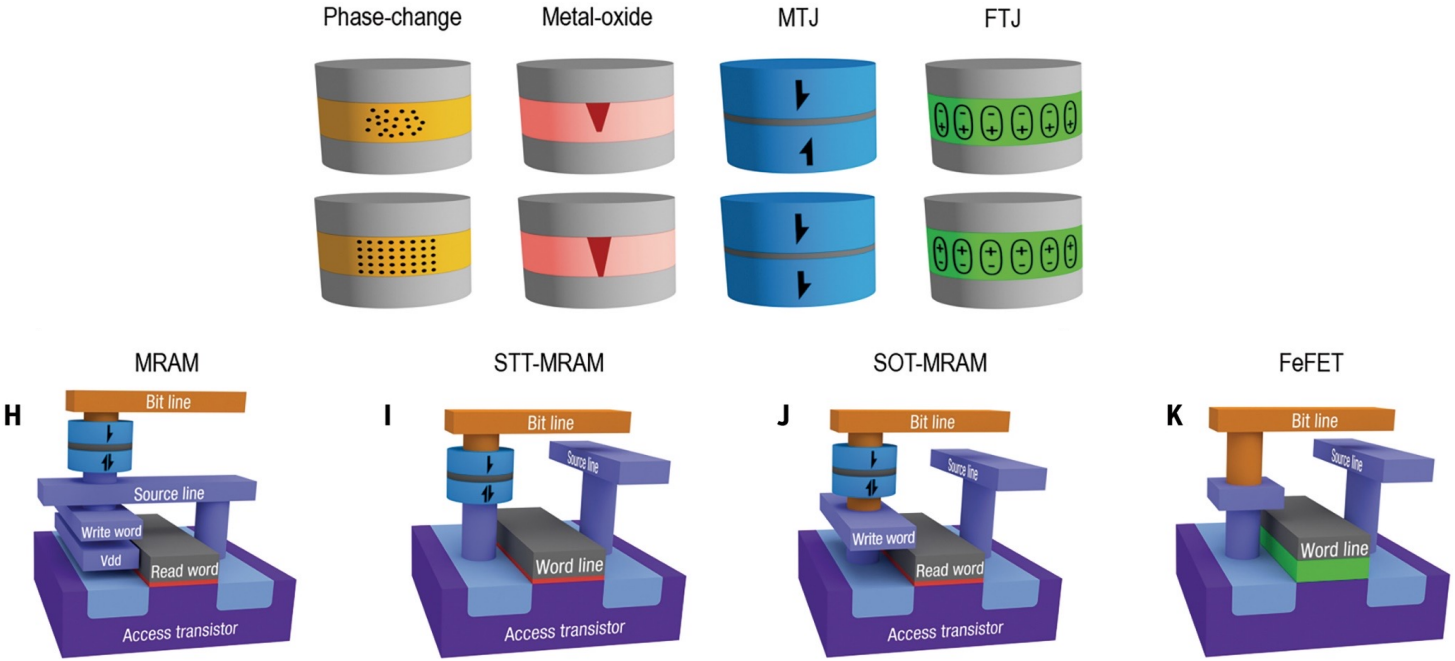
3D Monolithic On-chip memory:
- High Energy Efficiency
- High Capacity



FEOL CMOS under array

Georgia Tech

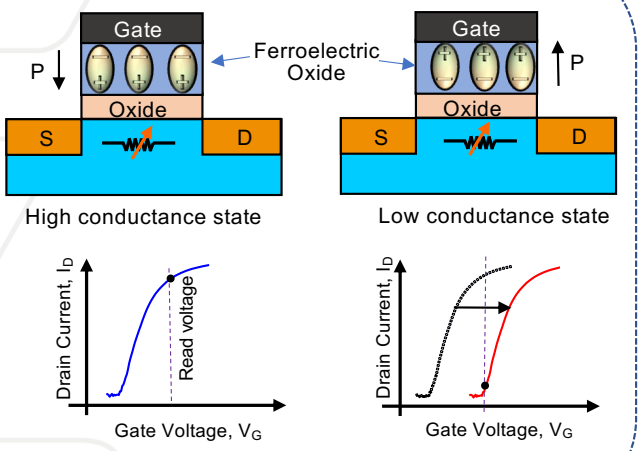
Many candidates but no clear winner



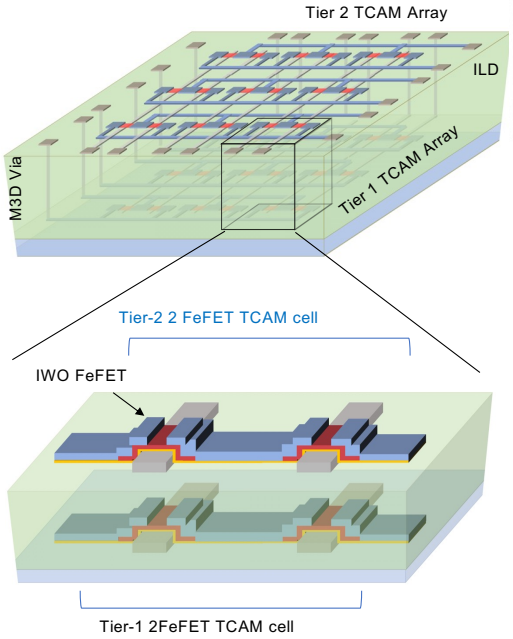
Each emerging memory candidate has its pros and cons.

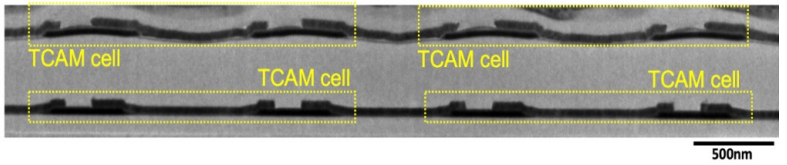
Georgia Tech

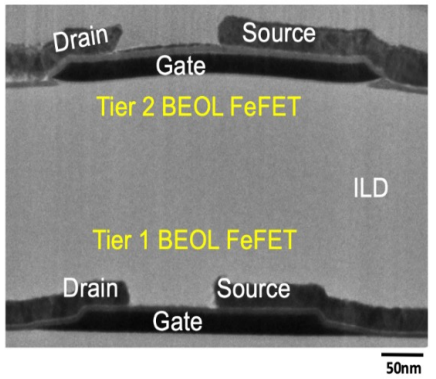
Monolithic 3D Single Transistor (1T) Ferroelectric FET memory



Data stored as the polarization state of the FE



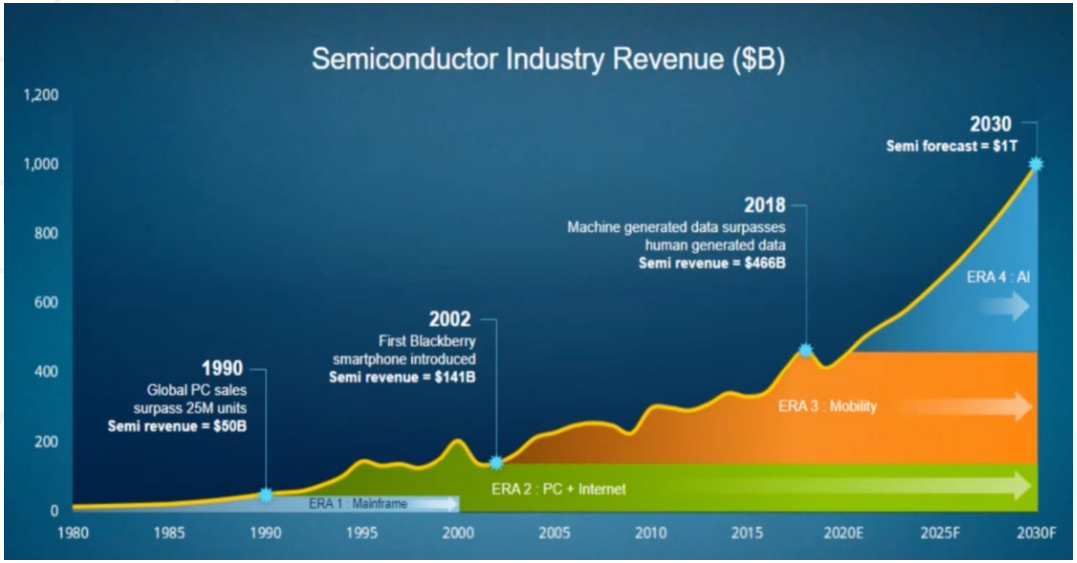




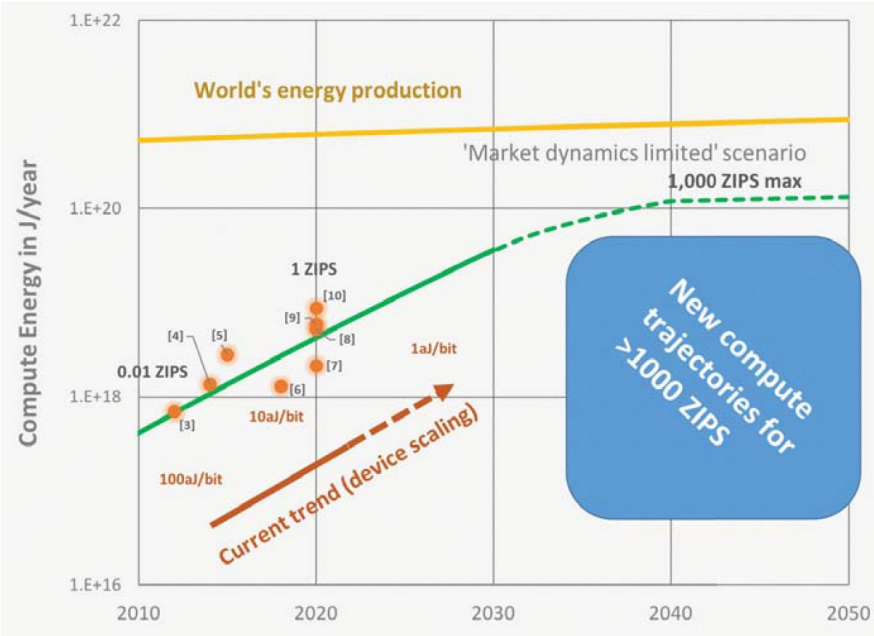
2-tier monolithic 3D array using Back-end-of-line (BEOL) compatible Ferroelectric FETs

Georgia Tech

The energy and sustainability cost of computation



Source: Applied Materials



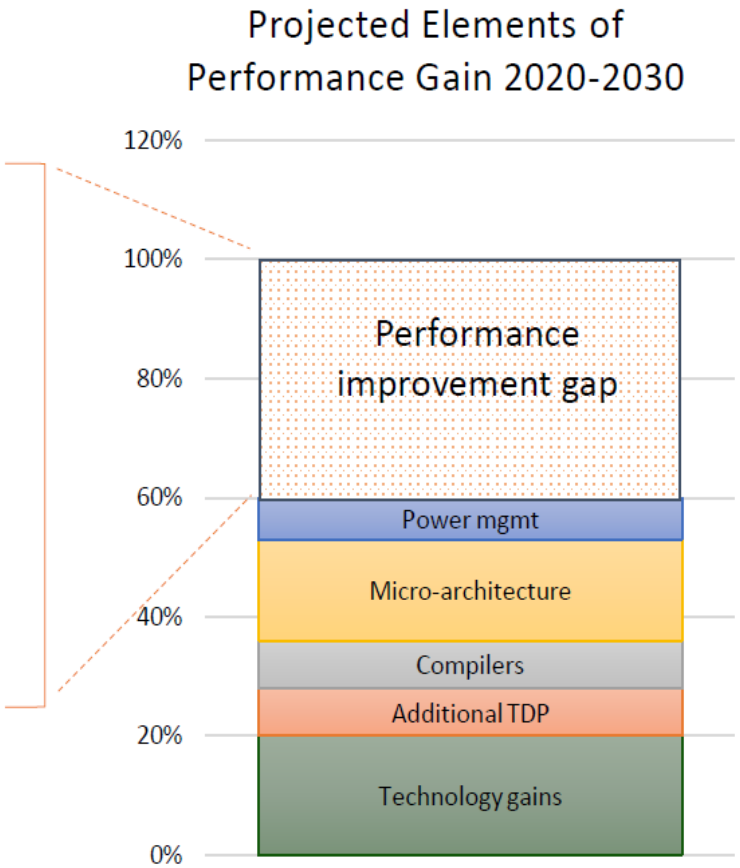
The semiconductor industry is projected to grow by ~100% in the 7 years!

The growth of compute energy can also be on an unsustainable path!

Georgia Tech

Compute Performance Gain for next decades

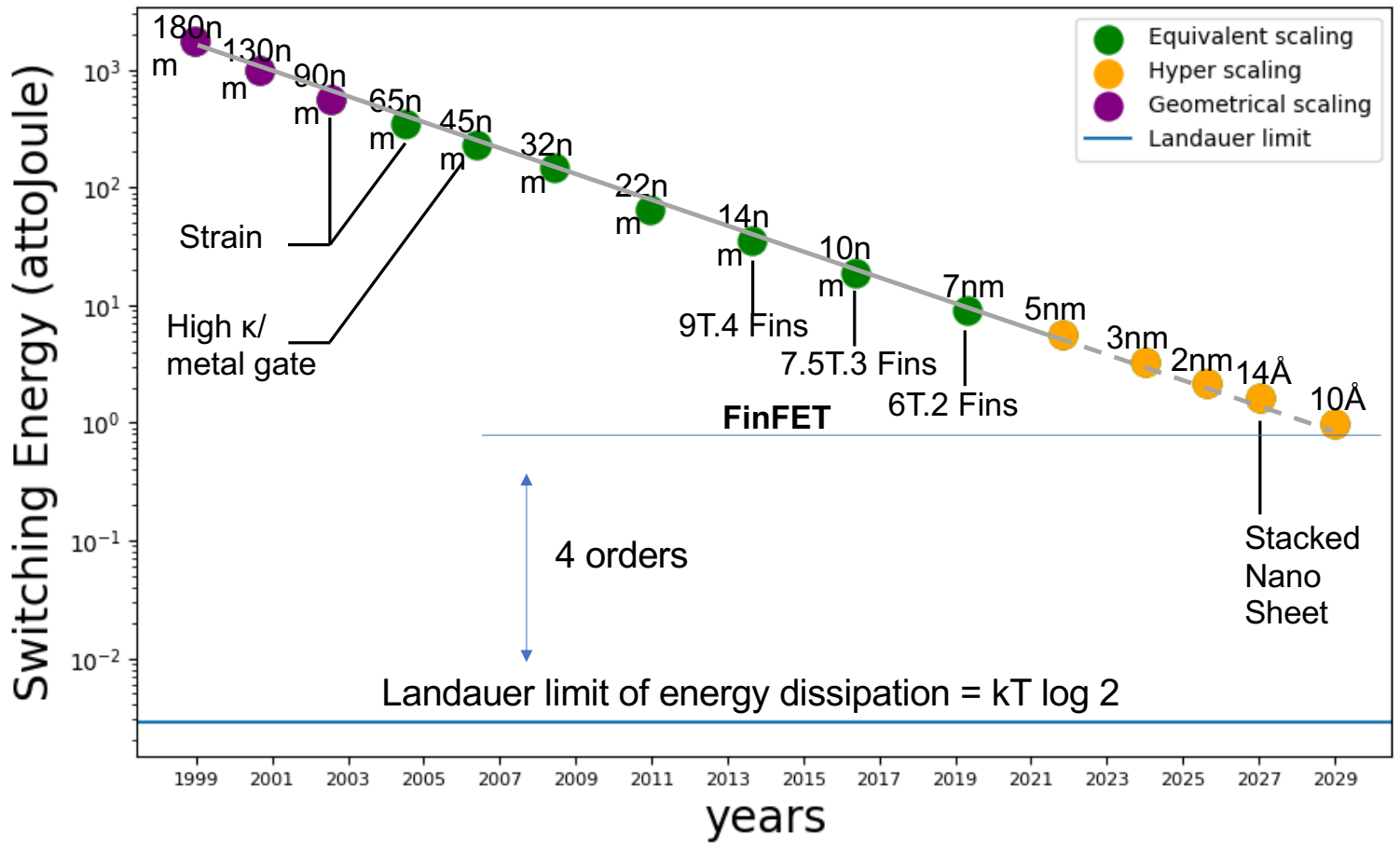
- ❑Materials with tailored properties aided by high throughput materials discovery
- ❑Monolithic 3D technologies for logic, memory, power delivery, thermal management
- ❑Design automation for modular chiplet design, placement and assembly for flexible and composable heterogenous compute system
- ❑Device-circuit-system-application co-optimization for maximal use of hardware resources



Lisa Su, AMD
2019 DARPA ERI Summit

Georgia Tech

Plenty of Room at the Top and Bottom



Georgia Tech


Georgia Tech
midtown
ATLANTA