Leveraging Monolithic 3D Technology for Data-Centric Applications

Lightening Talk at the EES2 Technical Workshop
September 14, 2022

Emre Salman and Ayse K. Coskun (Boston University)

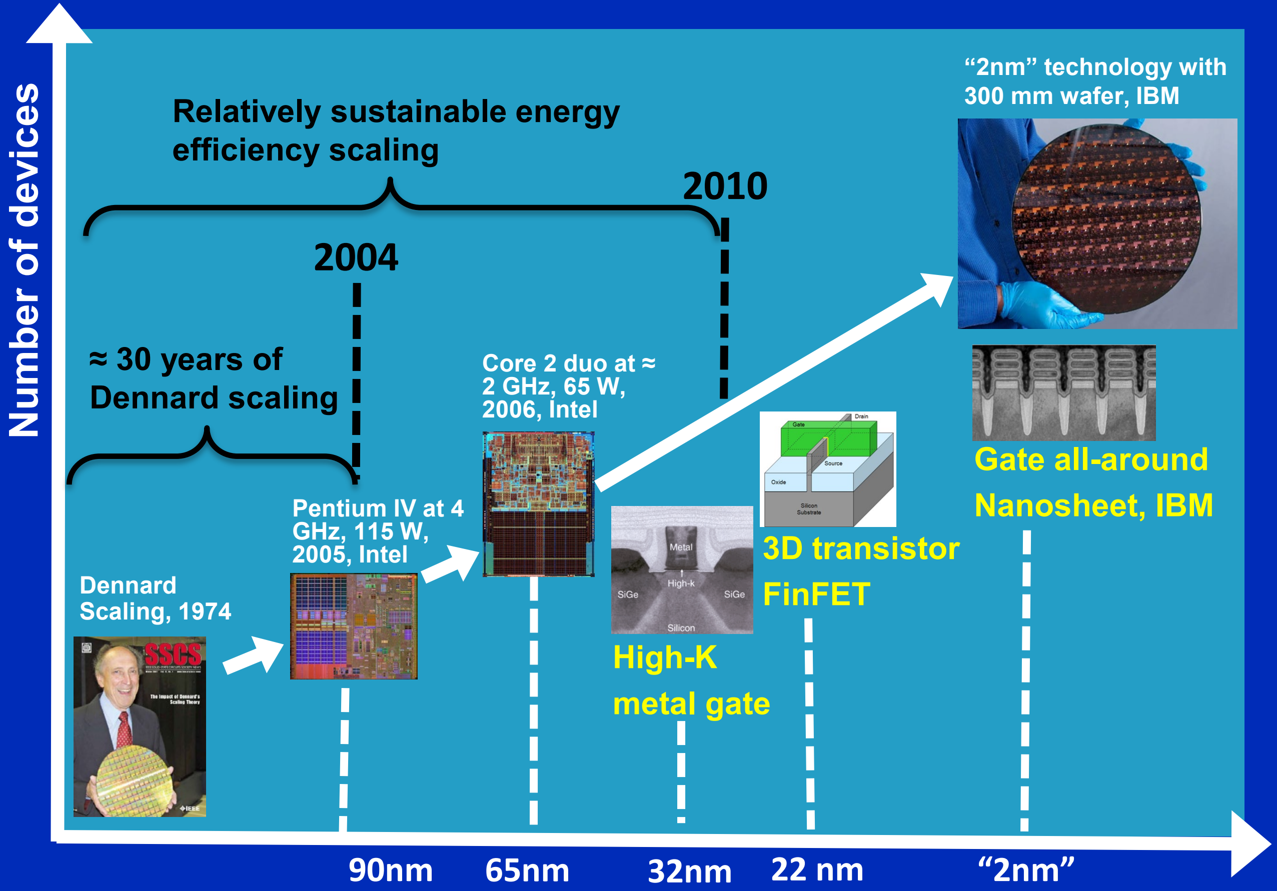
Department of Electrical and Computer Engineering
Stony Brook University (SUNY), Stony Brook, New York

Stony Brook University

https://nanocas.ece.stonybrook.edu

emre.salman@stonybrook.edu

Sustainable Energy Efficiency Scaling



Number of devices

Relatively sustainable energy efficiency scaling

2004

2010

≈ 30 years of Dennard scaling

Dennard Scaling, 1974

Pentium IV at 4 GHz, 115 W, 2005, Intel

Core 2 duo at ≈ 2 GHz, 65 W, 2006, Intel

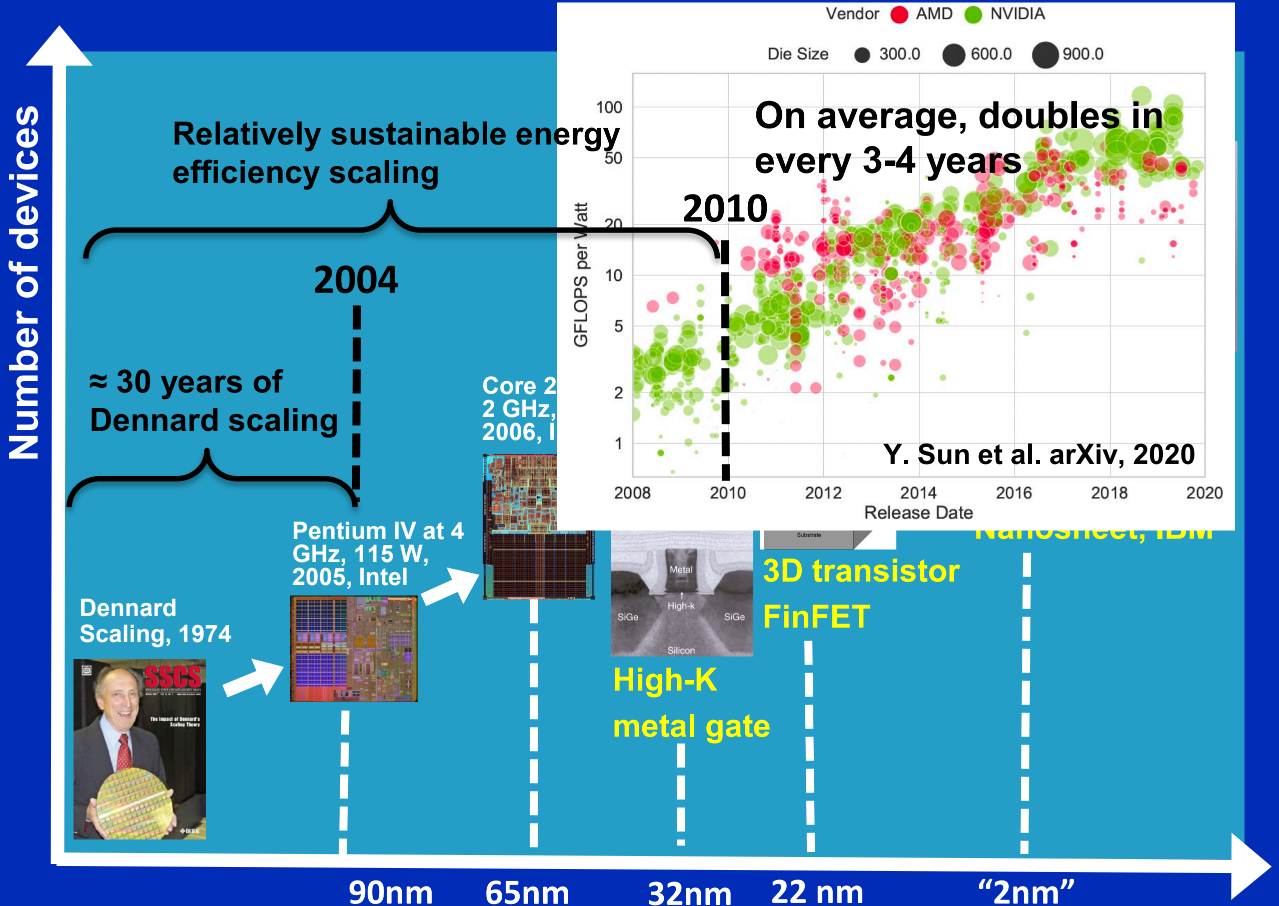
High-K metal gate

3D transistor FinFET

"2nm" technology with 300 mm wafer, IBM

Gate all-around Nanosheet, IBM

Sustainable Energy Efficiency Scaling



Relatively sustainable energy efficiency scaling

2004

2010

≈ 30 years of Dennard scaling

Dennard Scaling, 1974

Pentium IV at 4 GHz, 115 W, 2005, Intel

Core 2
2 GHz,
2006, I

High-K metal gate

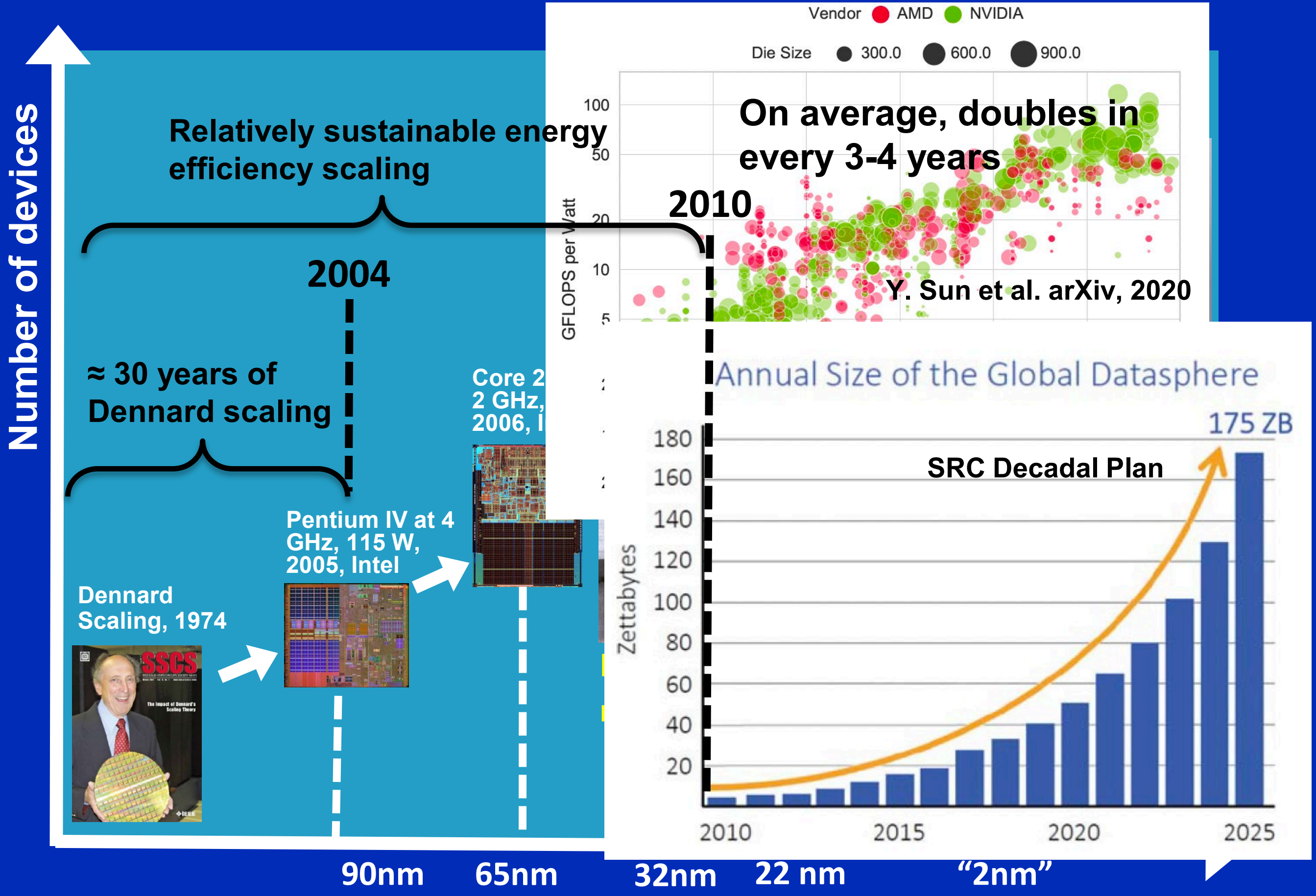
3D transistor FinFET

Nanosheet, IBM

On average, doubles in every 3-4 years

Y. Sun et al. arXiv, 2020

Sustainable Energy Efficiency Scaling



Relatively sustainable energy efficiency scaling

2004

2010

≈ 30 years of Dennard scaling

Dennard Scaling, 1974

Pentium IV at 4 GHz, 115 W, 2005, Intel

Core 2 2 GHz, 2006, I

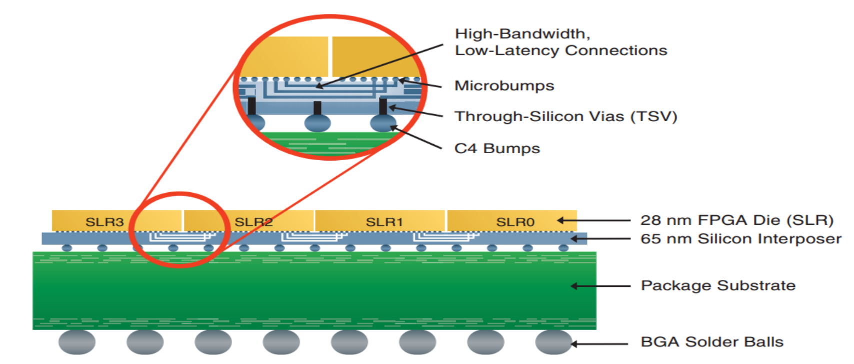
On average, doubles in every 3-4 years

Y. Sun et al. arXiv, 2020

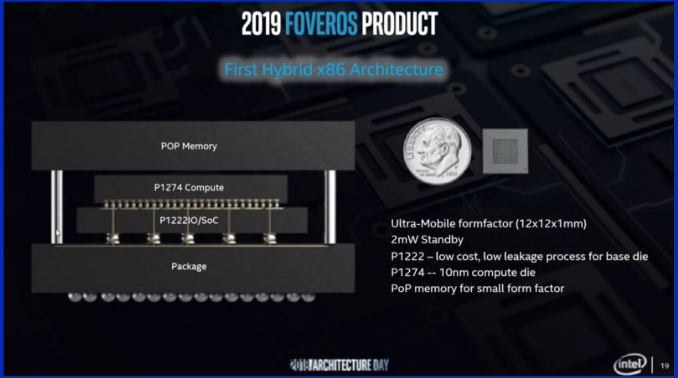
SRC Decadal Plan

- Dominance of data-centric applications
  - Integration platforms for mitigating dominance of data on performance/power
    - 2.5D (chiplet)
    - TSV based 3D
    - Monolithic 3D
- FLOPS / Watt
  - Upper bound in power
  - Lower bound in performance
  - Application specific
    - Cloud vs. edge

Emerging Integration/Packaging Technologies

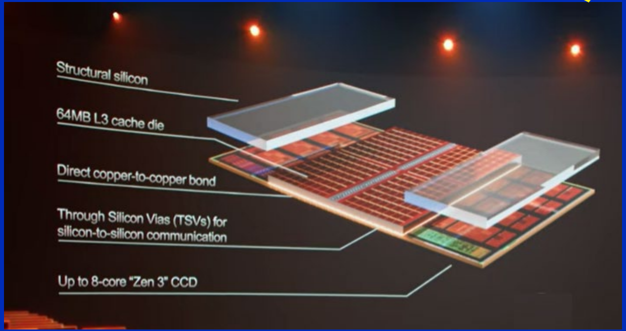


Interposer Based Integration (2.5D), Xilinx



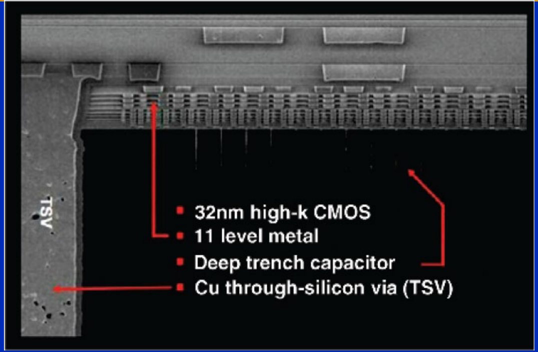
Ultra-Mobile formfactor (12x12x1mm)
2mW Standby
P1222 – low cost, low leakage process for base die
P1274 -- 10nm compute die
PoP memory for small form factor

Through Silicon Via (TSV), Intel FOVEROS

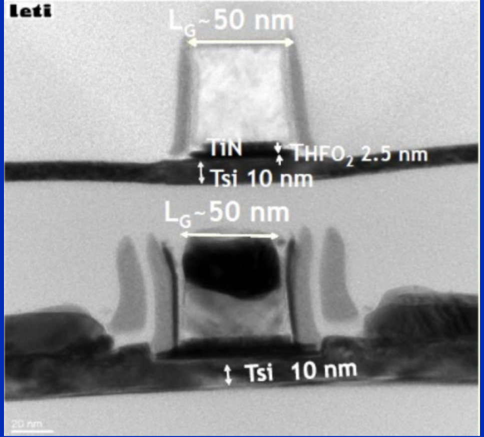

Structural silicon
64MB L3 cache die
Direct copper-to-copper bond
Through Silicon Vias (TSVs) for
silicon-to-silicon communication
Up to 8-core "Zen 3" CCD

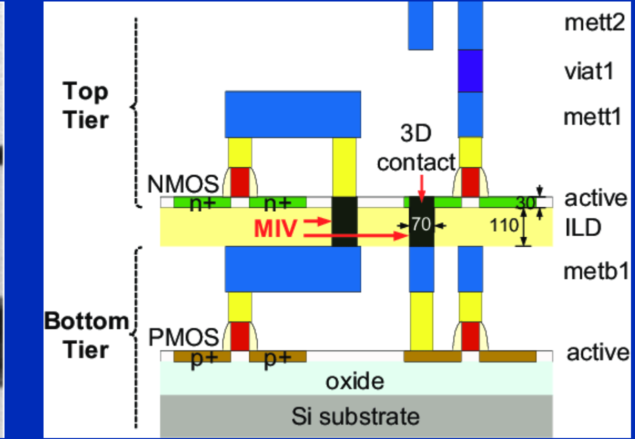
Cache-core stack, AMD Ryzen 5000 series

Monolithic 3D Integration Technology



Courtesy of IBM

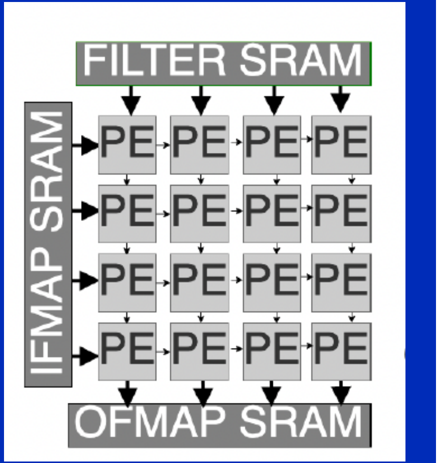


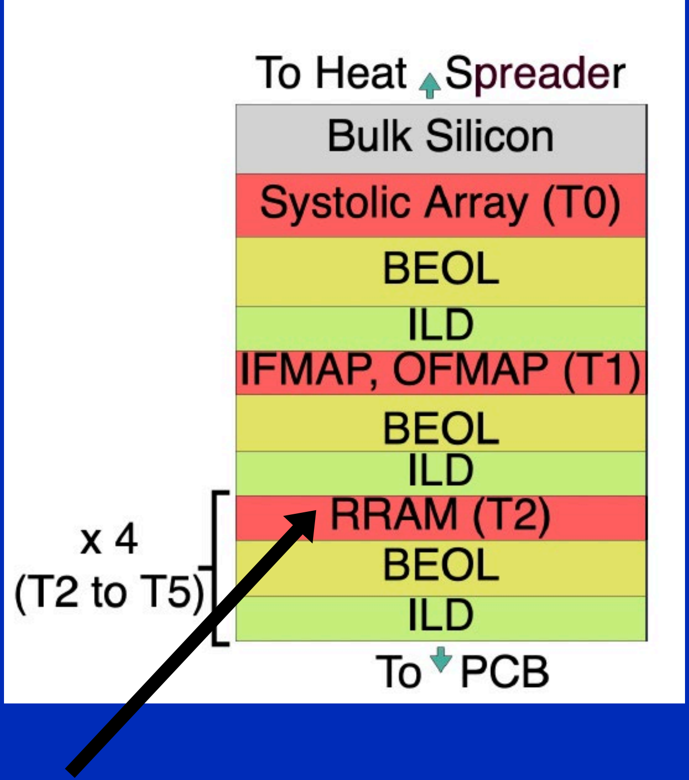


Monolithic 3D ICs with MIVs (LETI)

- TSVs are large
  - Diameter in the 2 to 10 μm range
  - Height in the 8 to 60 μm range
  - Large pitch (30 to 50 μm) and keep-out zone (KOZ) requirements
- MIVs provide unprecedented interconnect density
  - Diameter ≈ 50 nm (similar to a metal via)

Mono3D Deep Neural Network Accelerator

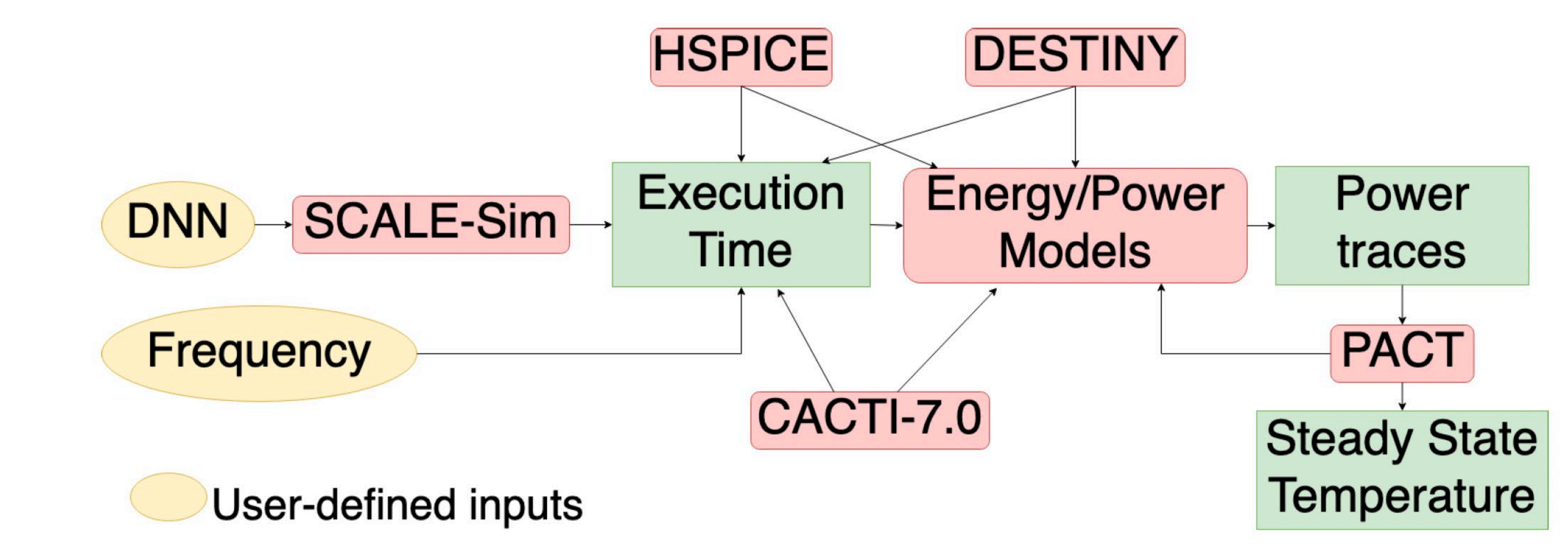




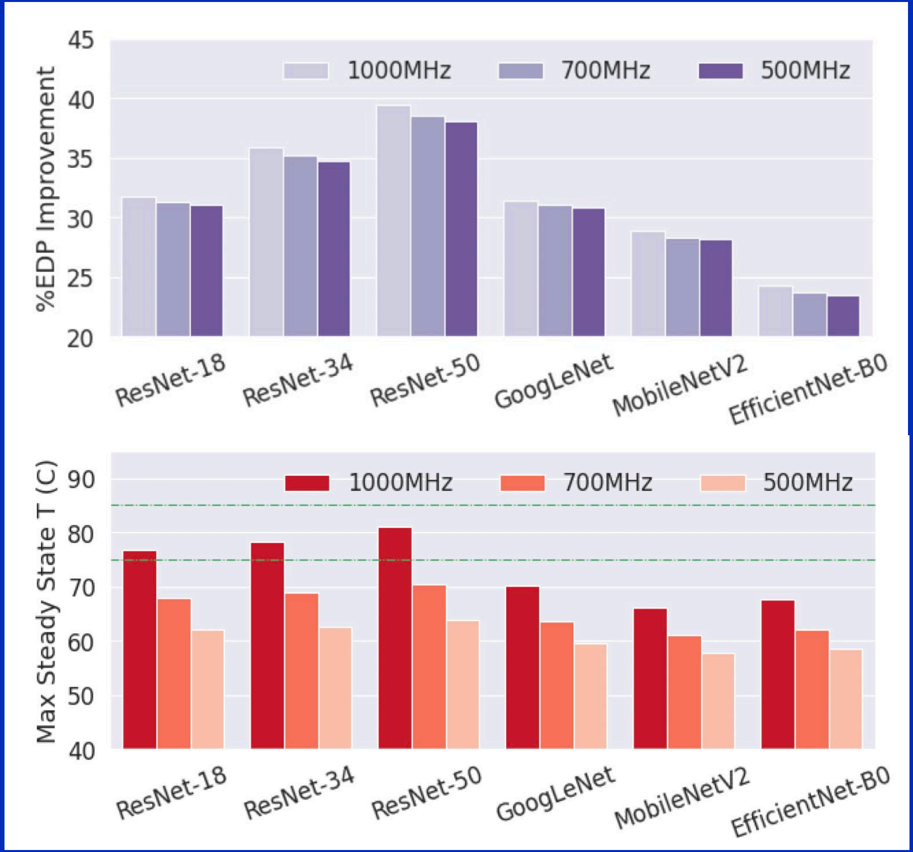
Resistive RAM to enable high bandwidth interface

- Accelerator hardware in 22nm CMOS
- Footprint of 8 mm^2
  - 256X256 array
  - 2 MB SRAM for IFMAP
  - 2 MB SRAM for OFMAP
  - 32 MB RRAM for weights
    - RRAM is read only
    - Endurance issue is mitigated
- Architectural optimizations
  - Read all weights into PE array in one cycle
  - Multicast all IFMAPs to all the PEs

Thermal-Aware Evaluation Framework



Improvements w.r.t. 2D System



- Up to 40% reduction in **energy-delay product**
- Up to 81% improvement in **inference per second per Watt**
- Up to 10X improvement in **inference per second per Watt per footprint**

Final Notes

- Impractical to achieve the same bandwidth with TSV-based 3D
  - Footprint would increase to more than 100 mm^2
- Optimizations to dataflow, architecture, and circuits are critical
  - Switching to Mono3D and RRAM alone does not produce desired improvements
    - EDP improvements less than 5%
  - Importance of co-design
- Thermal-awareness will play a key role in future advanced integration/packaging technologies
  - Higher power density
  - Temperature dependence of emerging nonvolatile memory
  - Edge applications with limited cooling capability

  - Questions: emre.salman@stonybrook.edu