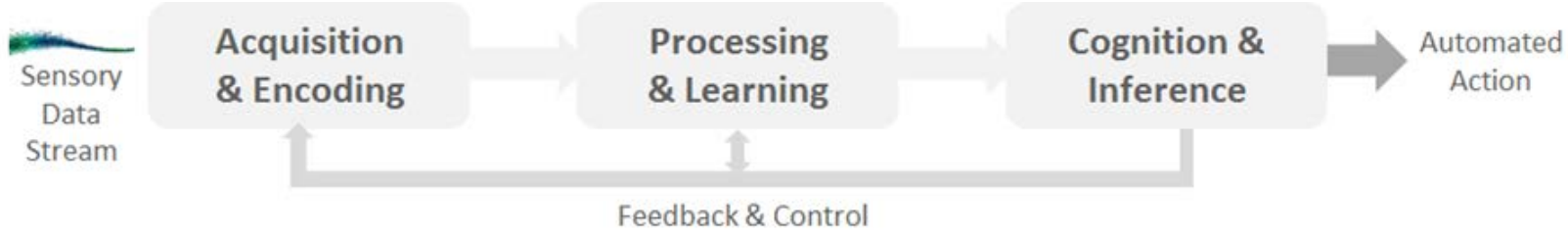
Energy Efficient Computing in Nanoscale CMOS

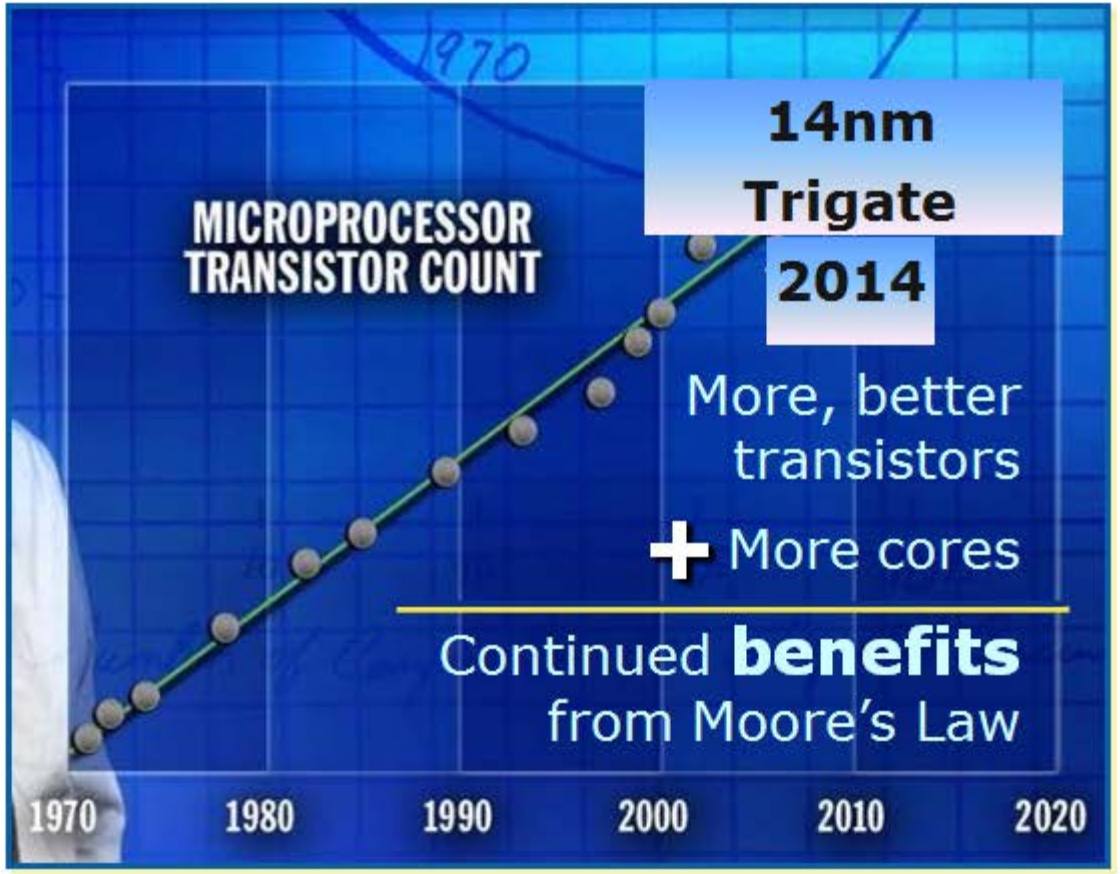
Vivek De
Intel Fellow
Director of Circuit Technology Research
Intel Labs

Internet of Everything (IoE)

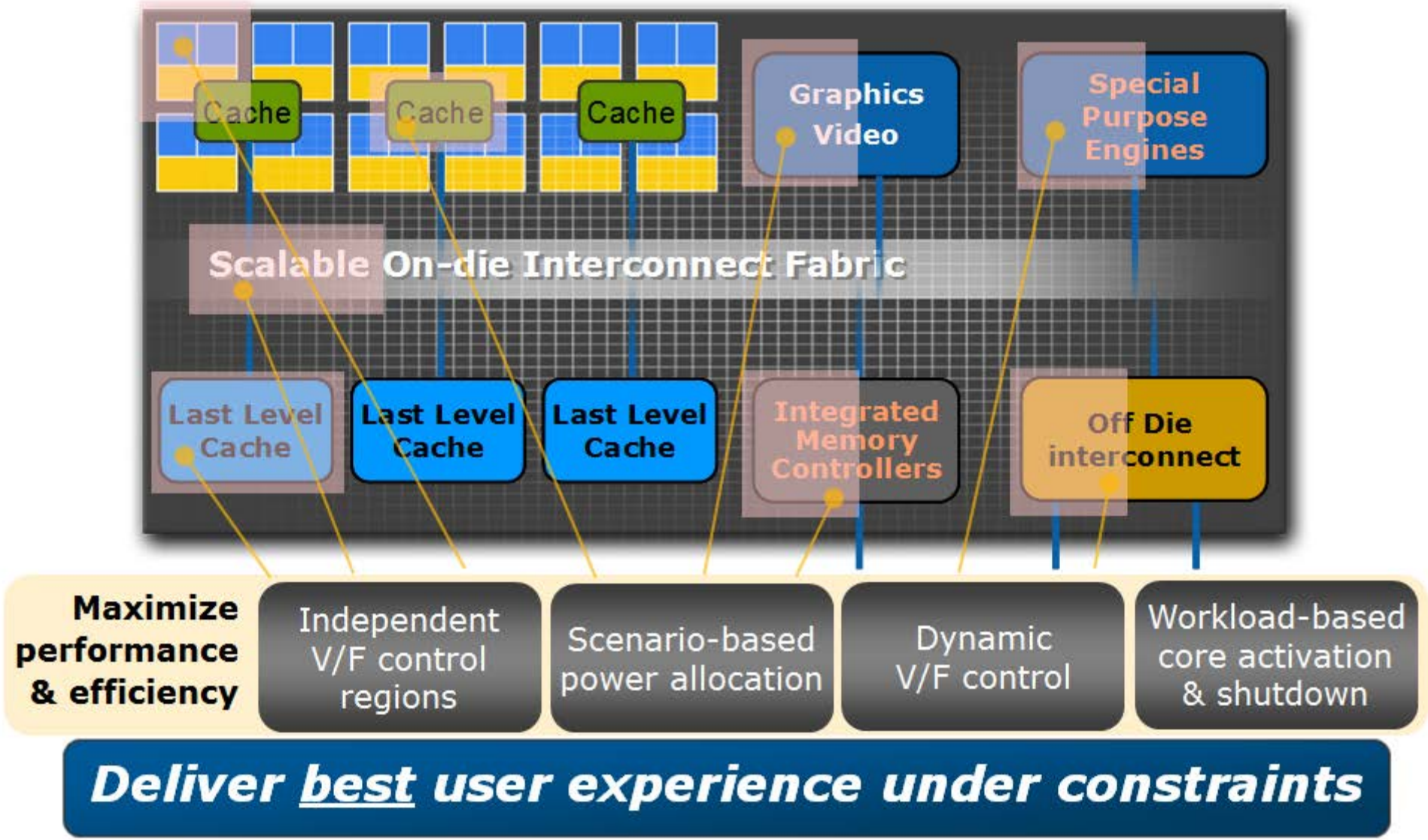


Paradigm shifts across hardware and software

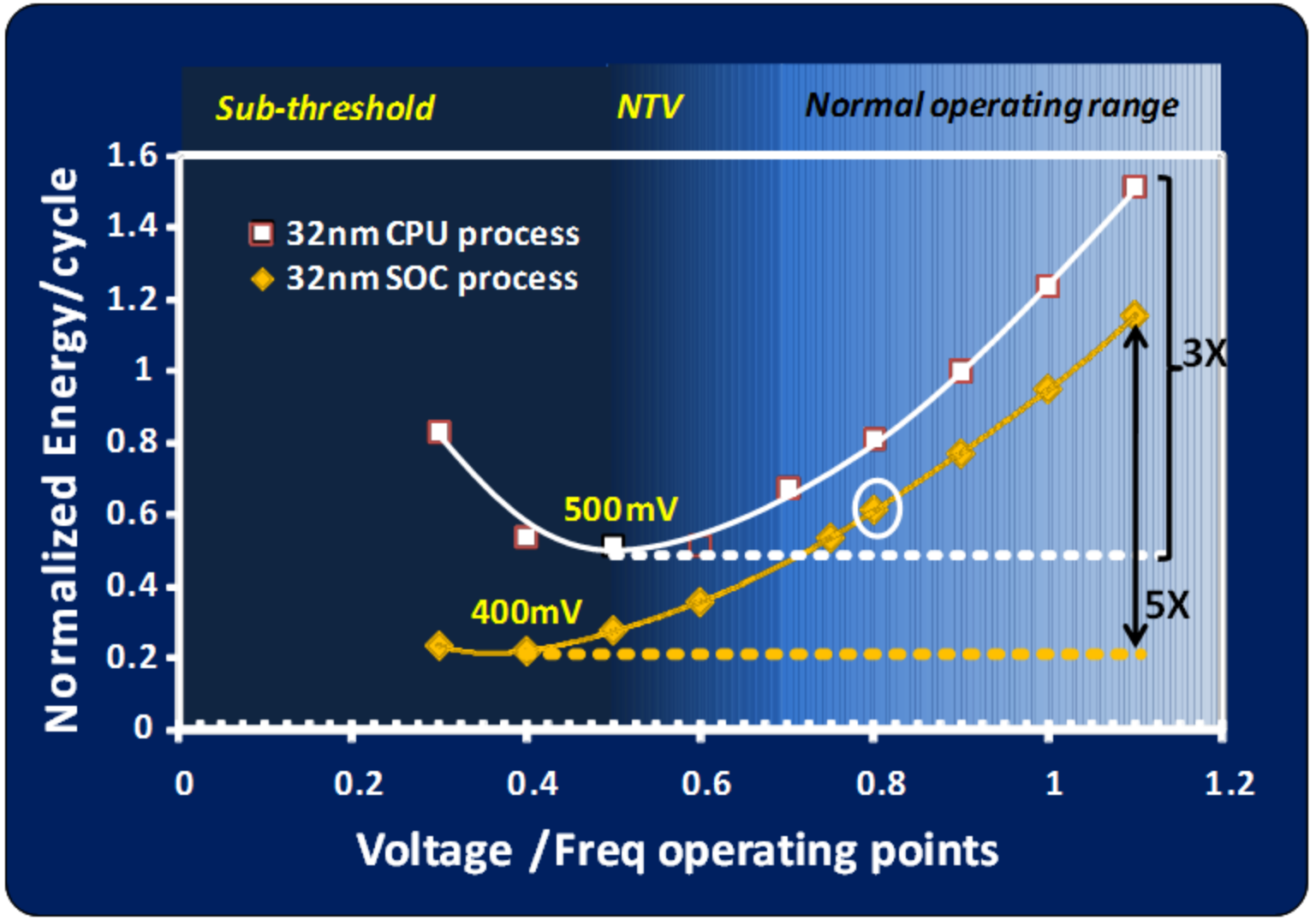
Moore’s Law scaling



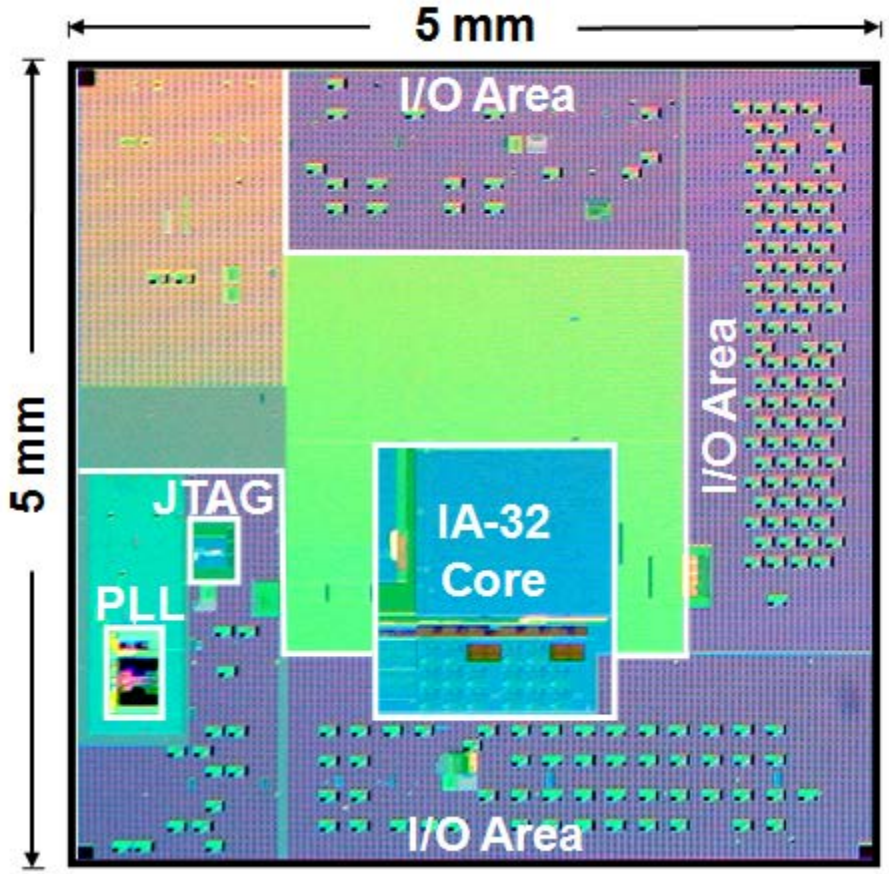
Dynamic platform control

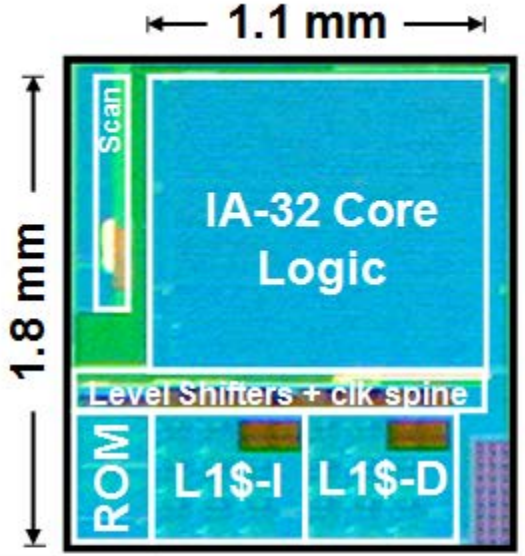


Near Threshold Voltage (NTV) computing



NTV IA processor



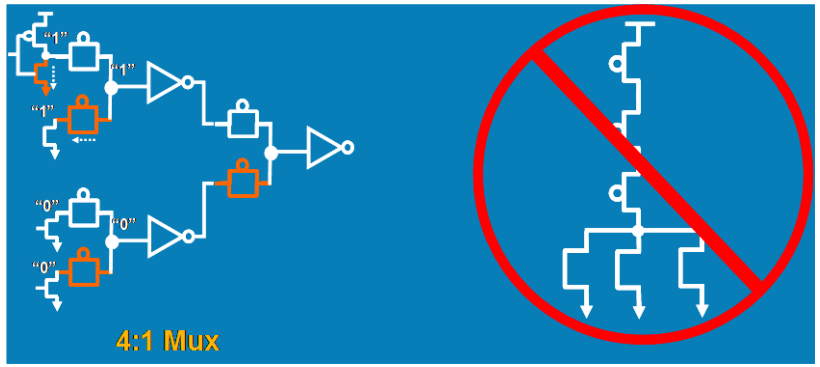


Technology	32nm High-K Metal Gate
Interconnect	1 Poly, 9 Metal (Cu)
Transistors	6 Million (Core)
Core Area	2mm^2

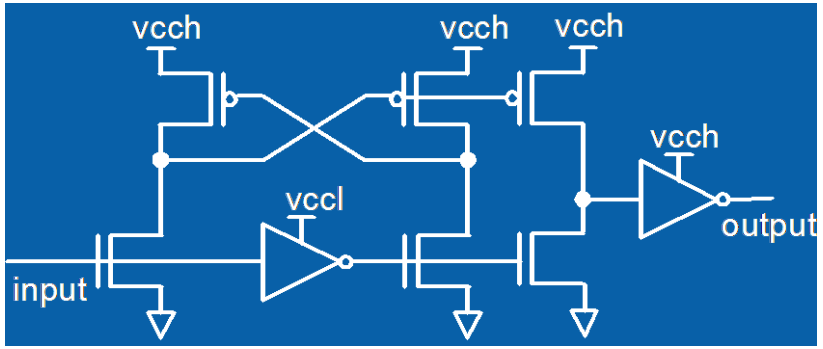
NTV design techniques

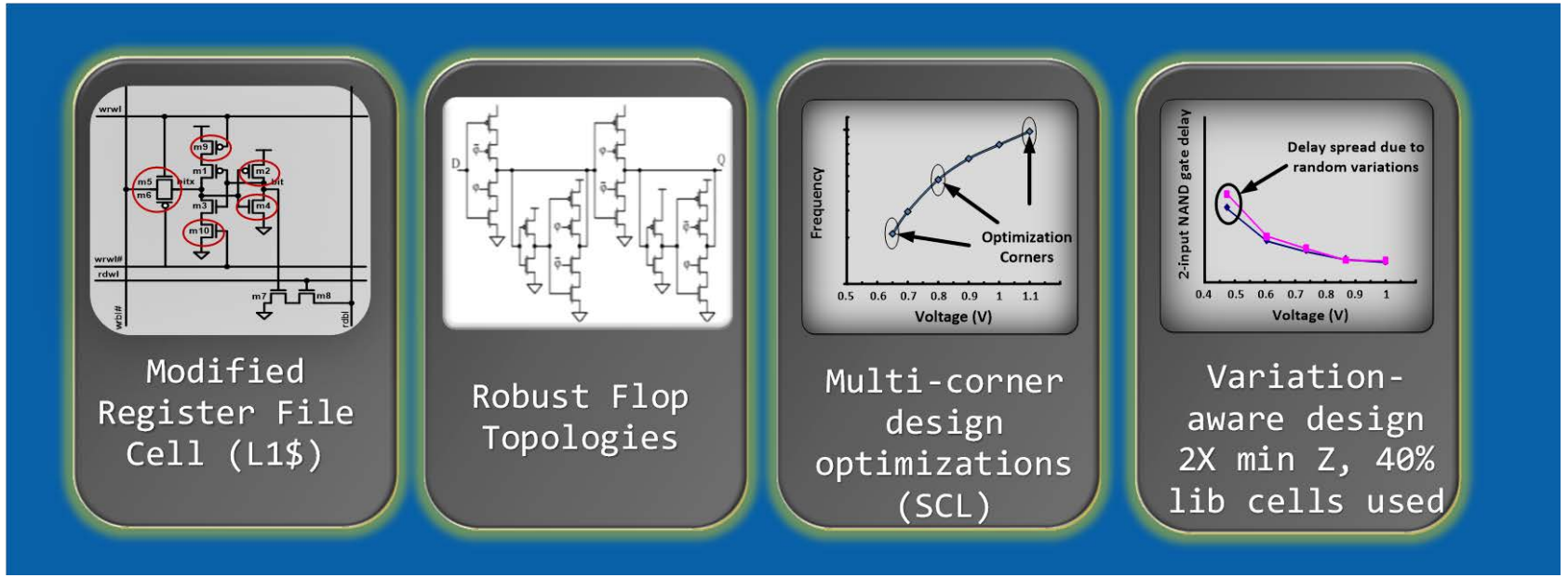
Narrow muxes

No stack height > 2

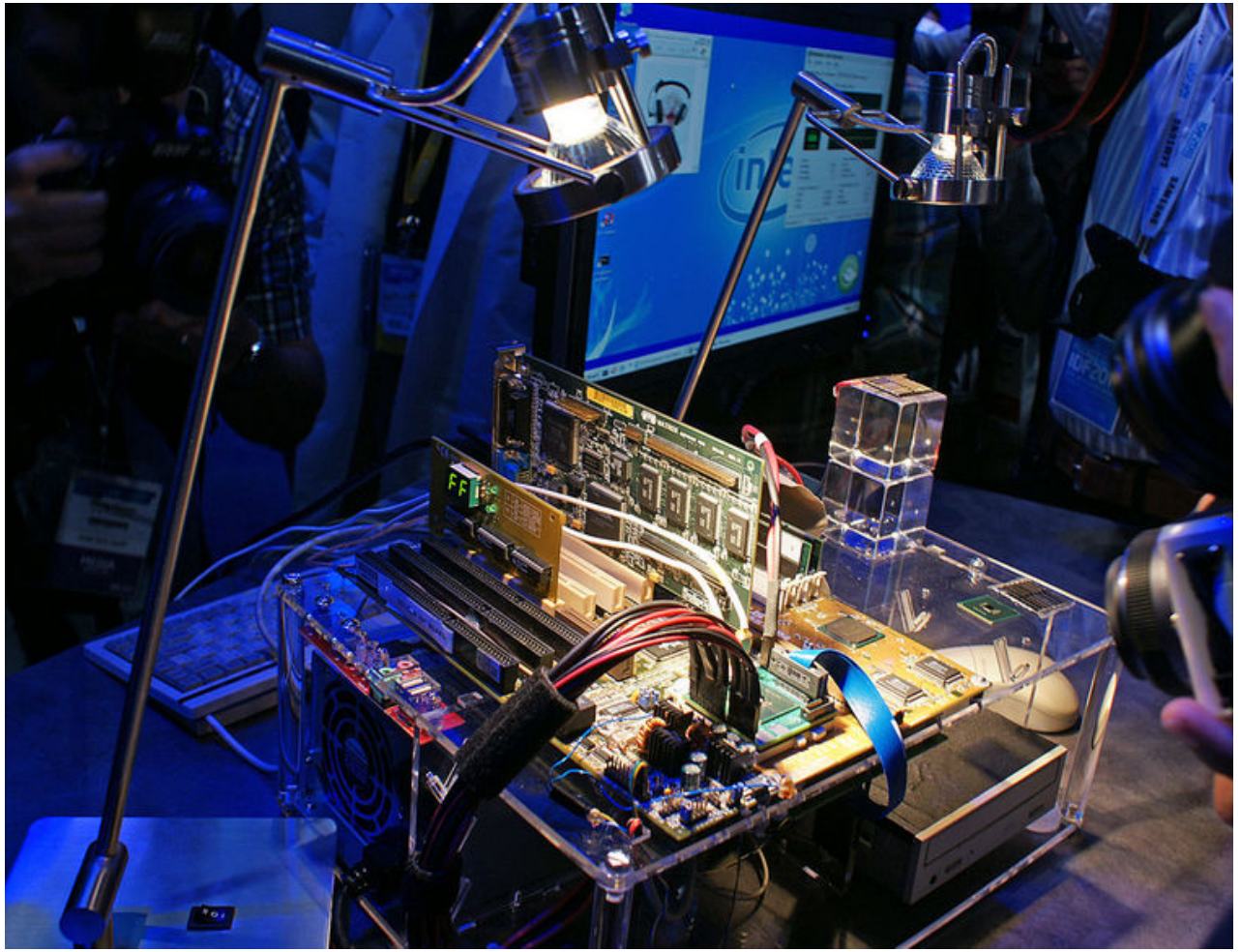


Robust level converters

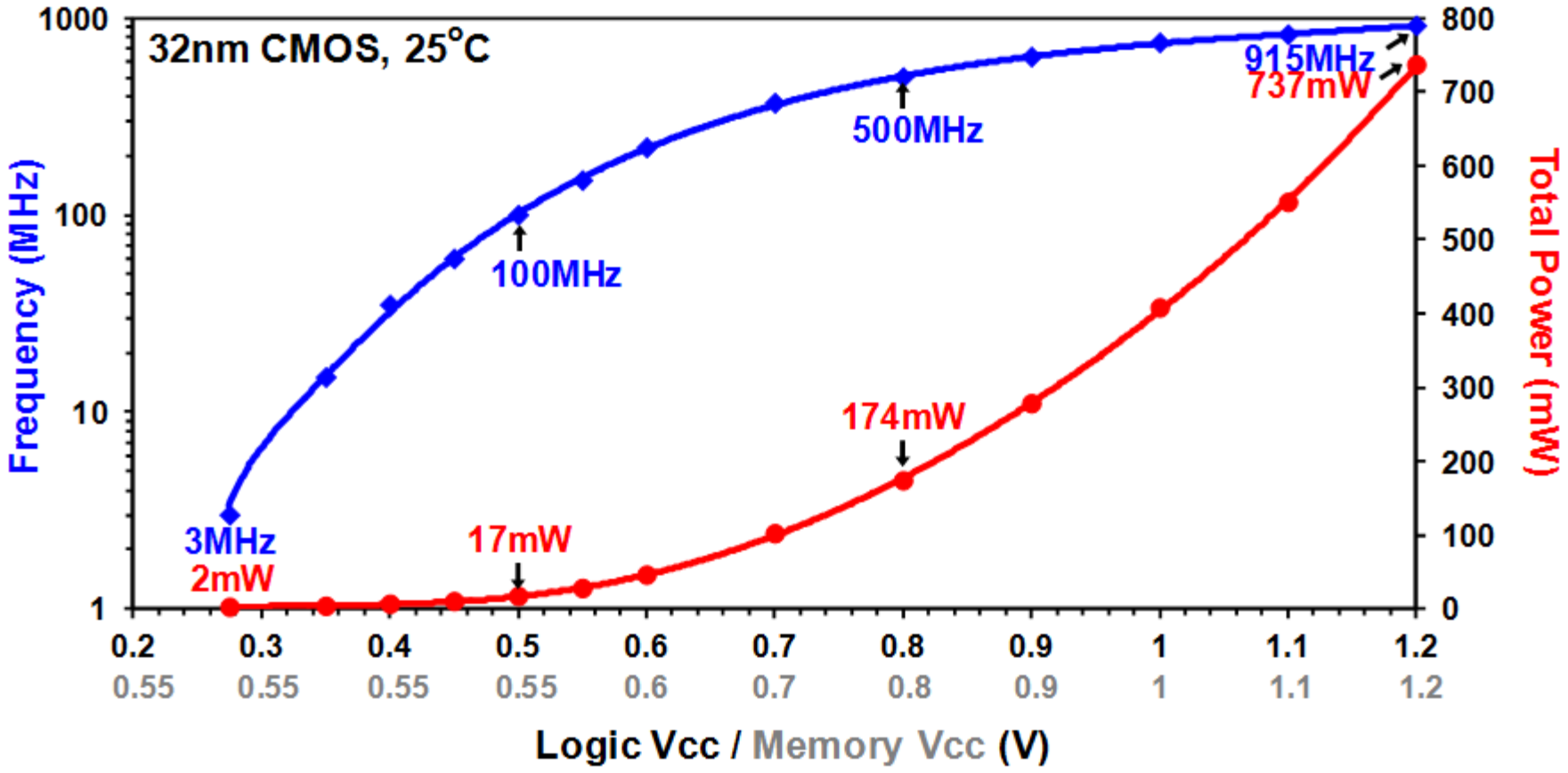




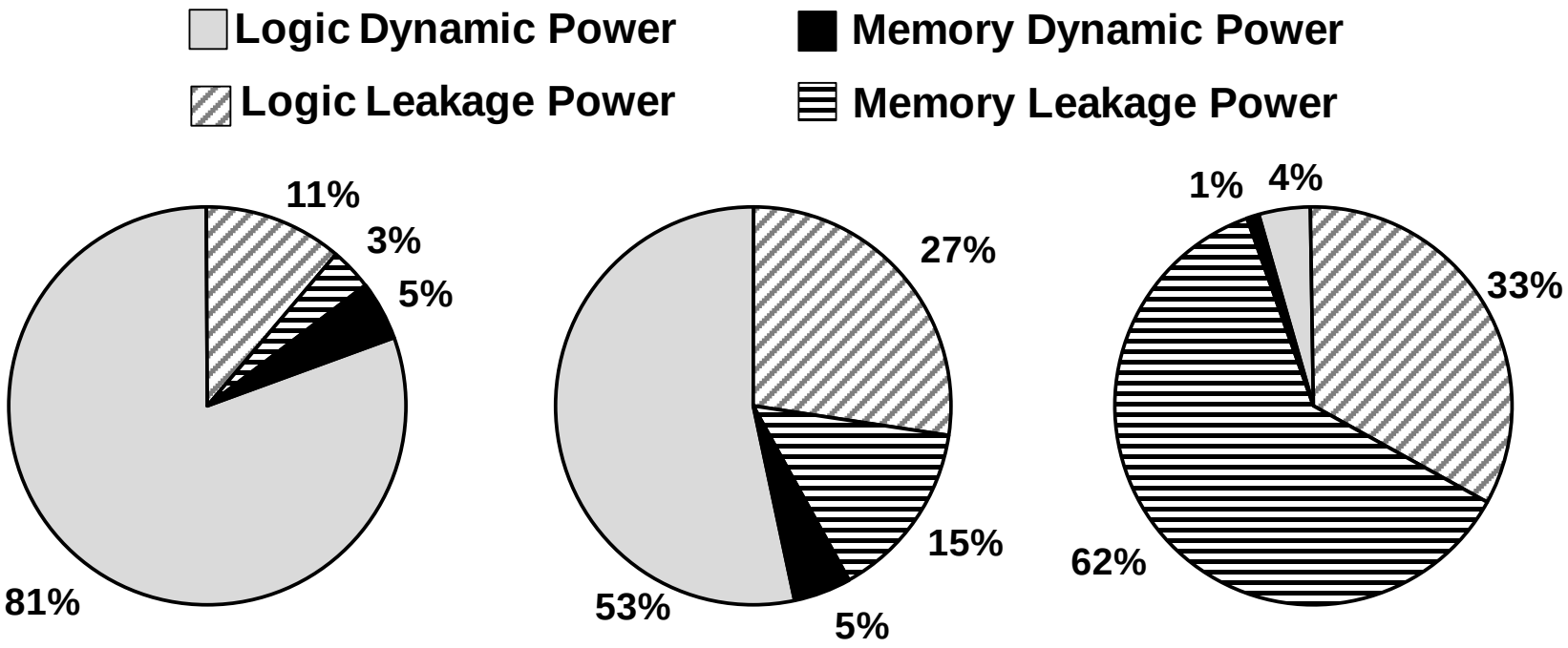
NTV IA – powered by solar cell!



Power performance measurements



Power components



Vcc-max (Super-Threshold)

Logic Vcc: 1.2V
Memory Vcc: 1.2V

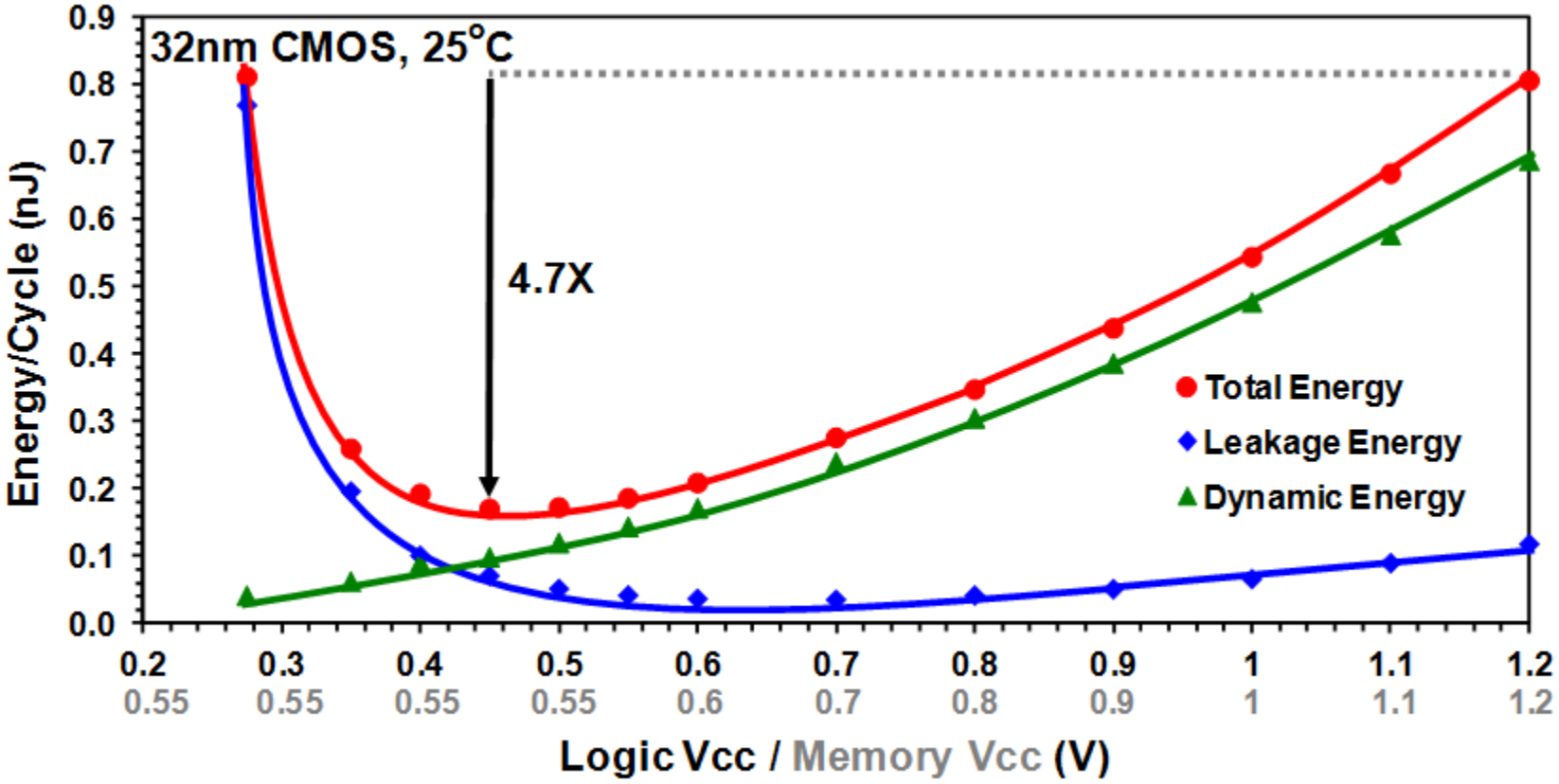
Vcc-opt (Near-Threshold)

Logic Vcc: 0.45V
Memory Vcc: 0.55V

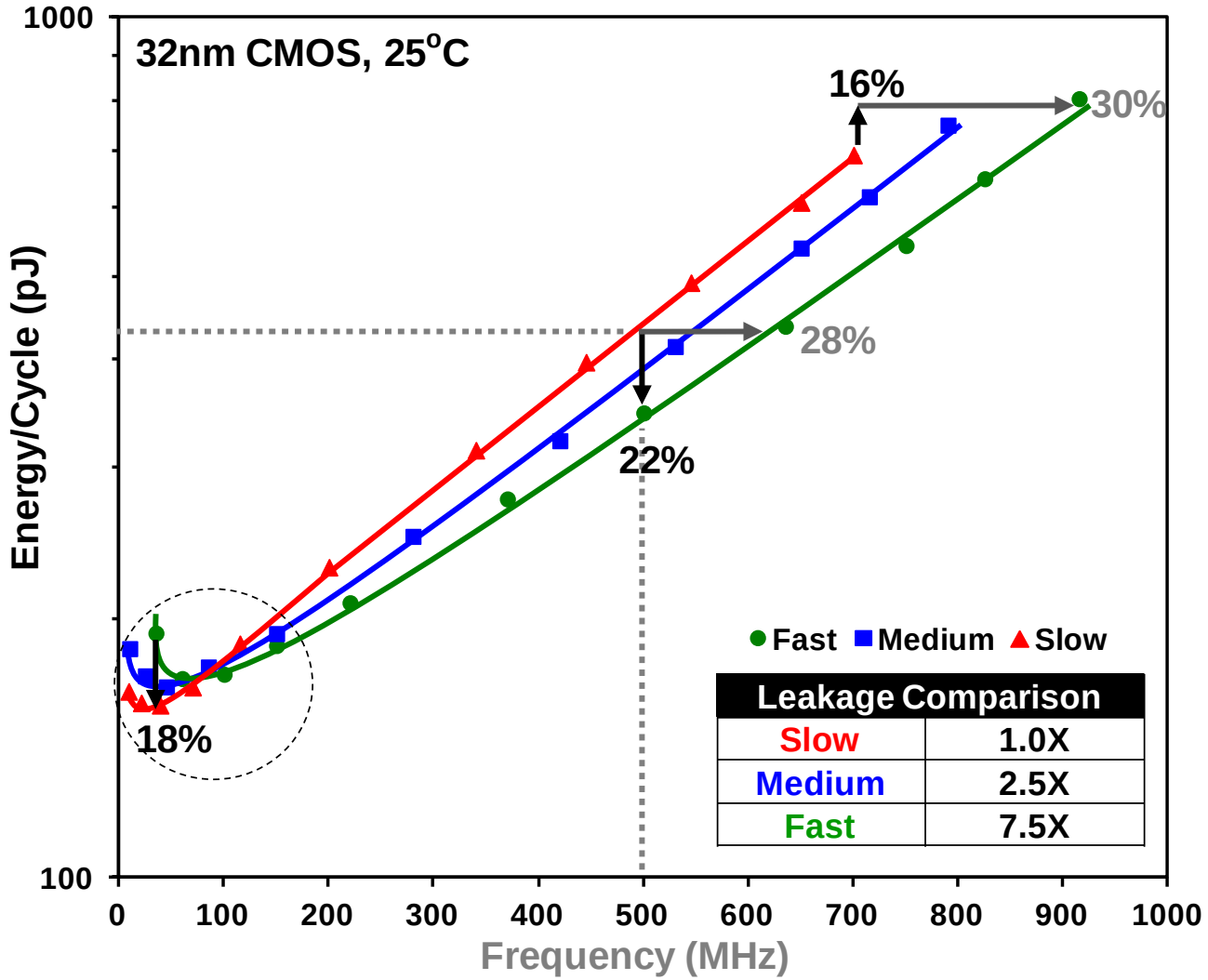
Vcc-min (Sub-Threshold)

Logic Vcc: 0.28V
Memory Vcc: 0.55V

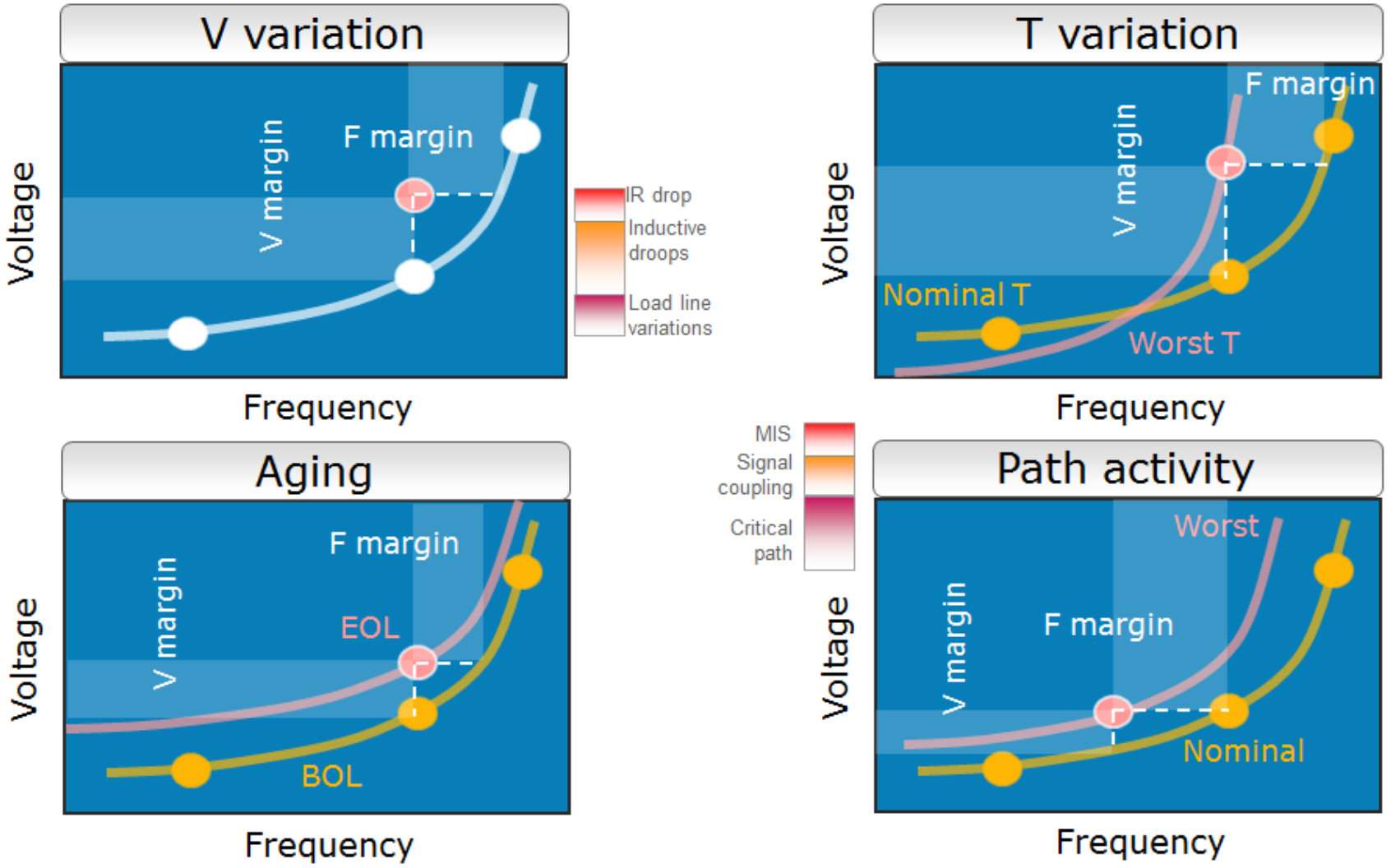
Minimum energy operation



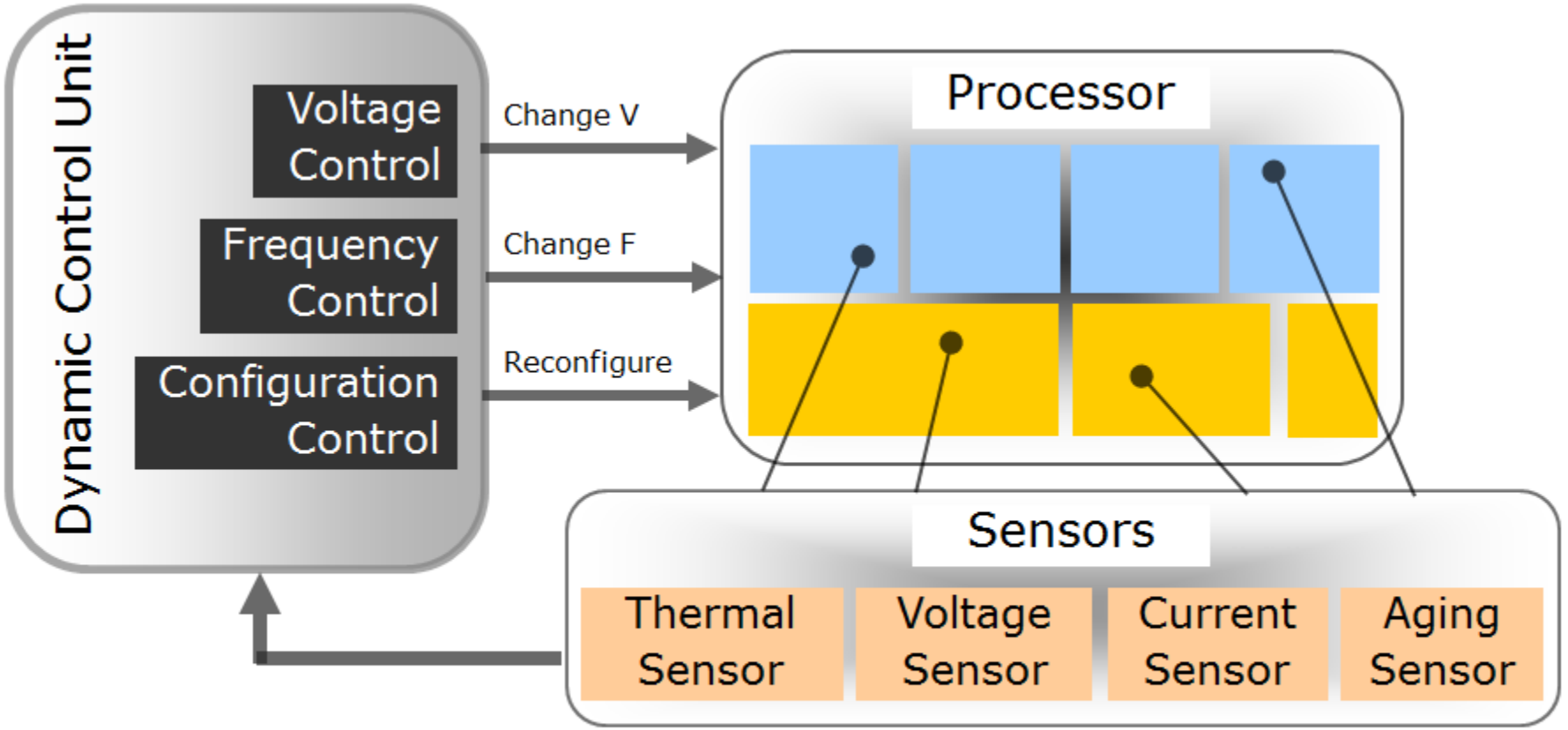
NTV and variability



Voltage-frequency margins

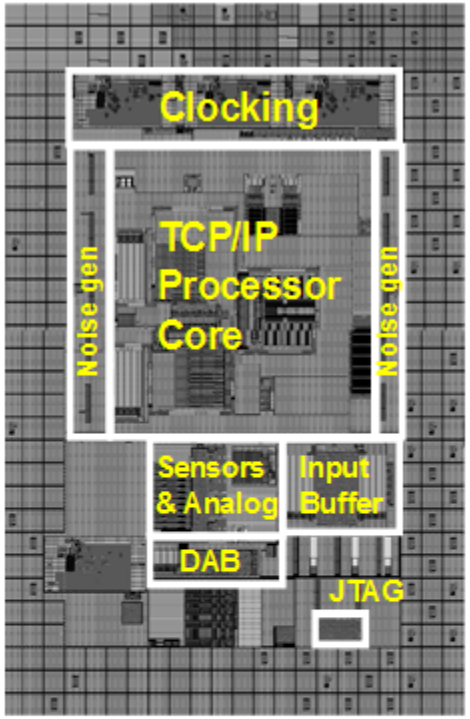


Dynamic adaptation & reconfiguration

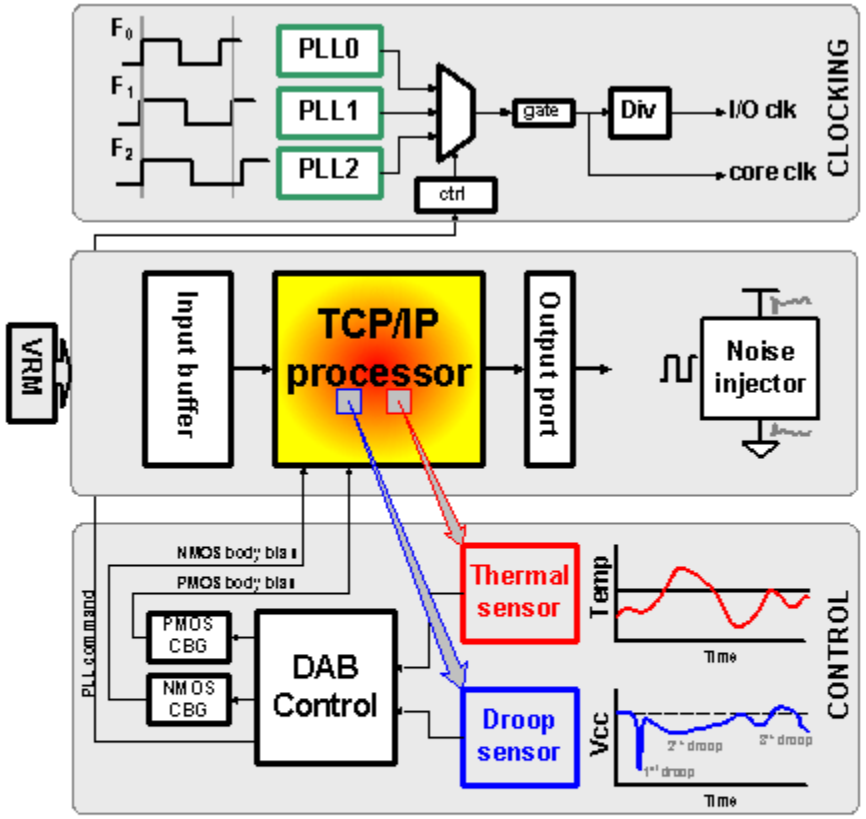


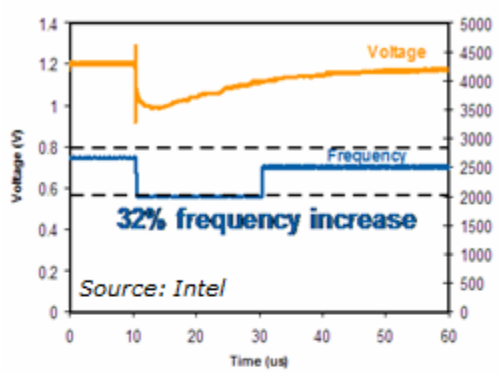
Adapt & reconfigure for best power-performance

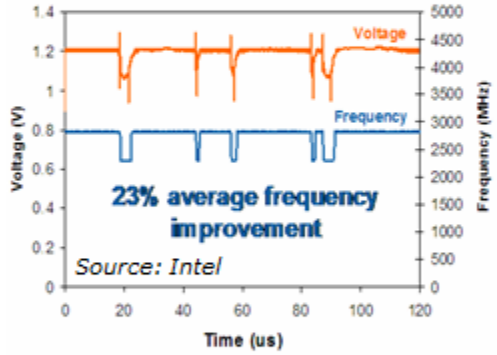
Dynamic V & F adaptation



Prototype chip in 90nm



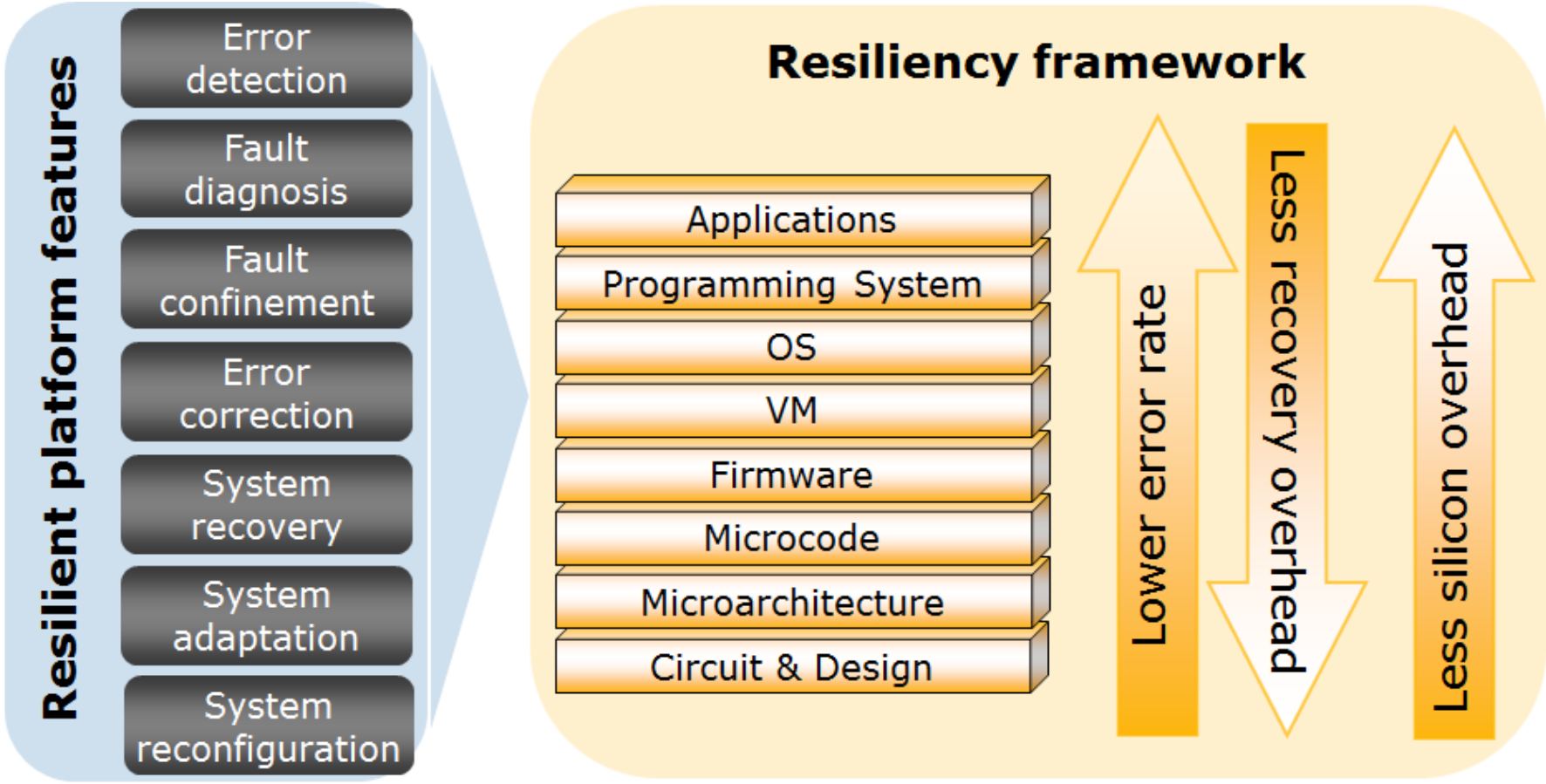




Environment-aware dynamic adaptation

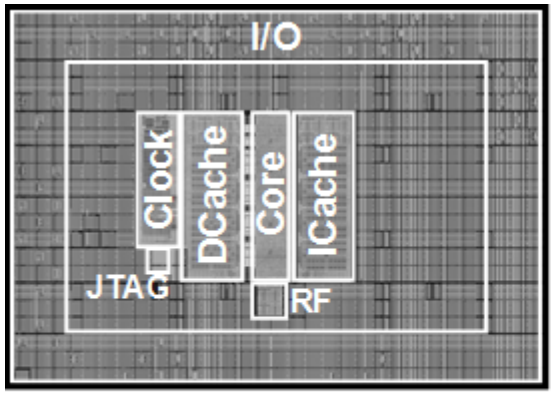
- Adapt F/V to V/T change → reduce V/T margin
- Adapt F/V to aging → reduce aging margin

Resilient platforms

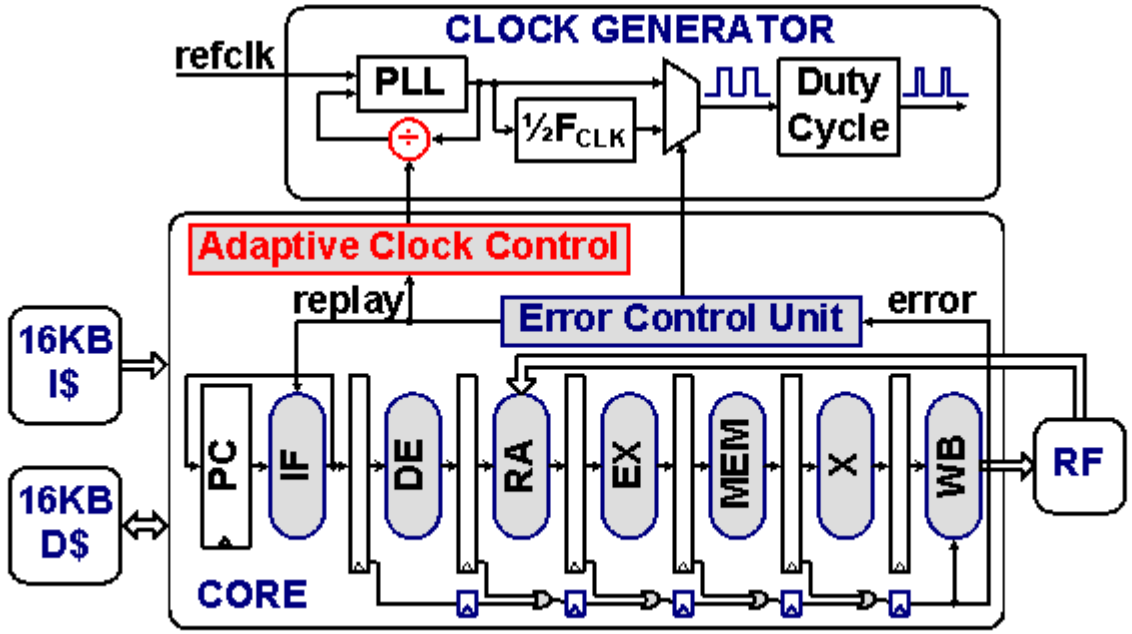


***Resiliency* for performance, efficiency & reliability**

Resilient & adaptive core



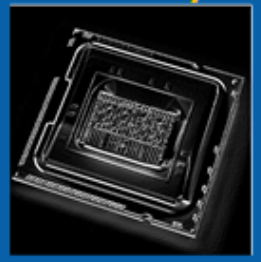
Technology	45nm CMOS
Die Area	13.64 mm^2
Core Area	0.39 mm^2
Core F_{MAX}	1.45GHz at 1.0V
Core Power	135mW at 1.0V



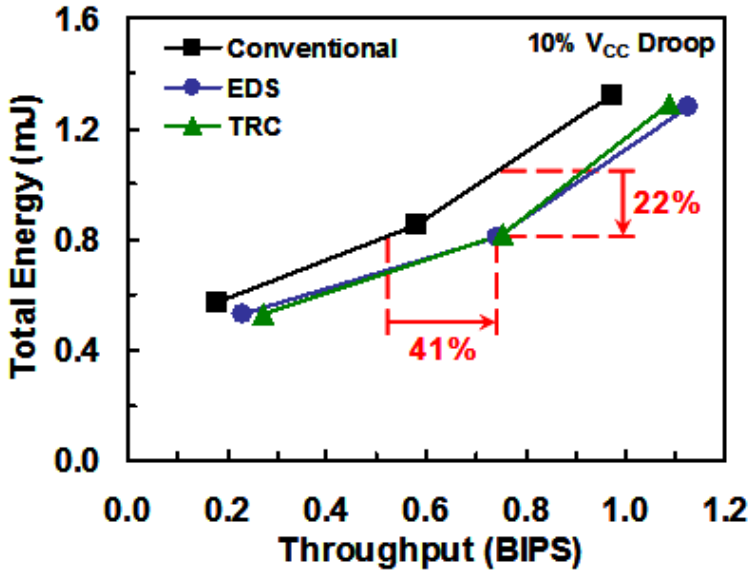
Performance & efficiency gains

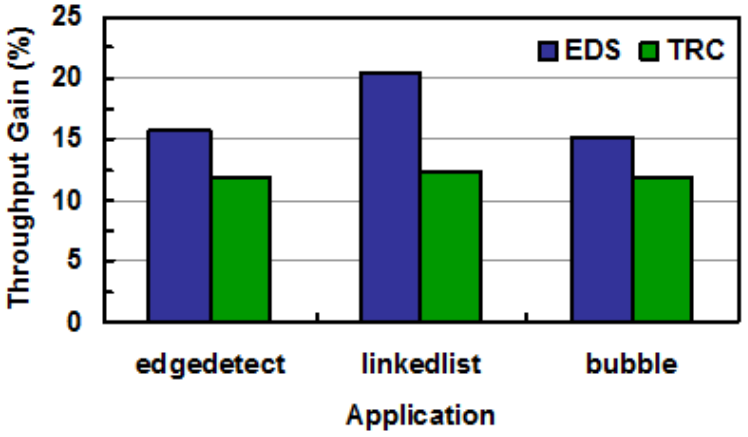
Input Image

Output Image: Resiliency ON



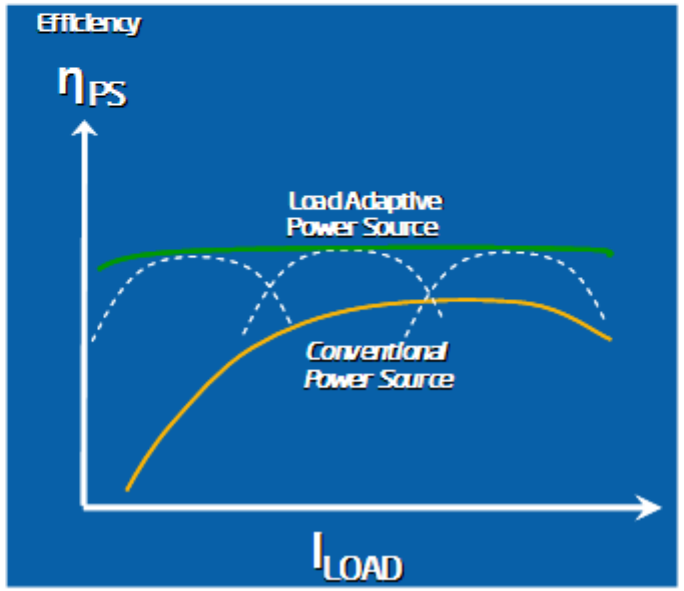
Output Image: Resiliency OFF



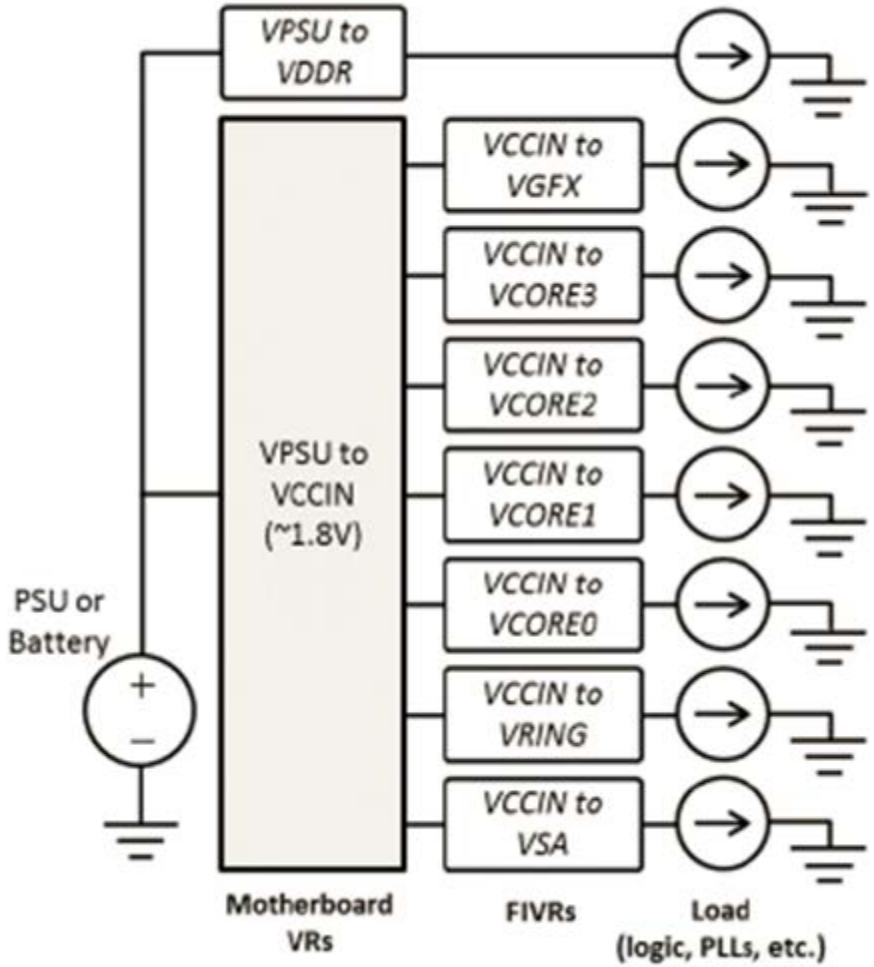


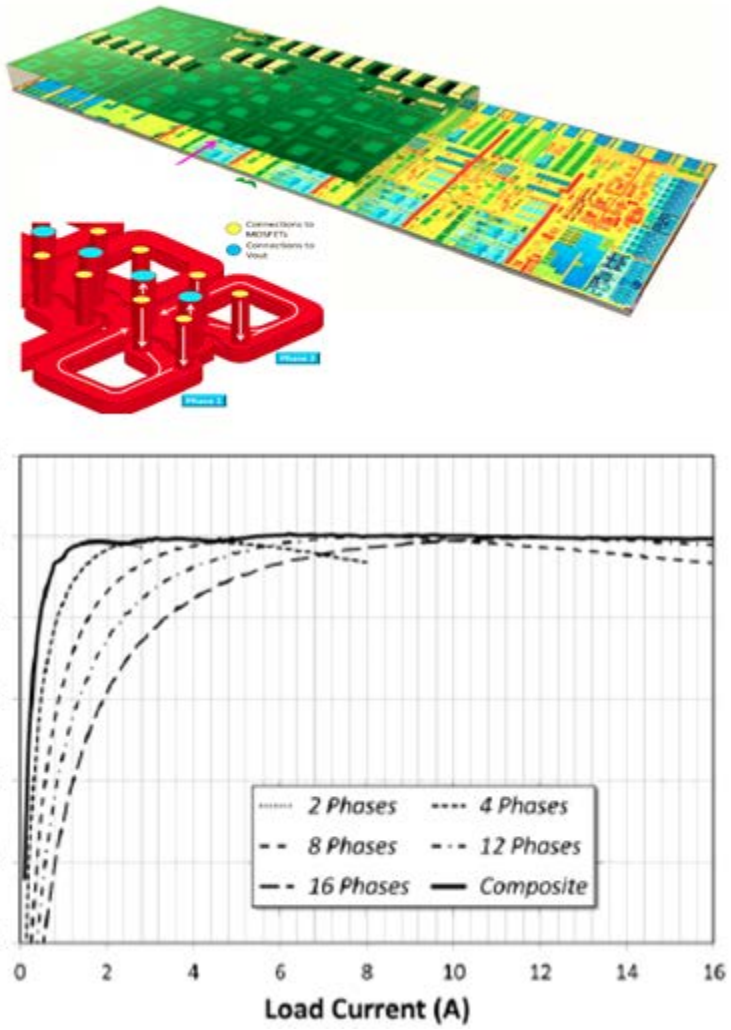
Integrated voltage regulators

Conversion	Distribution	Control
• Area efficient • Scalable • Persistent rail	• Lower loss • Higher fidelity • Simpler	• Fast & efficient • Load adaptive • Independent rails
Efficient	**Low loss**	**Fine-grain**

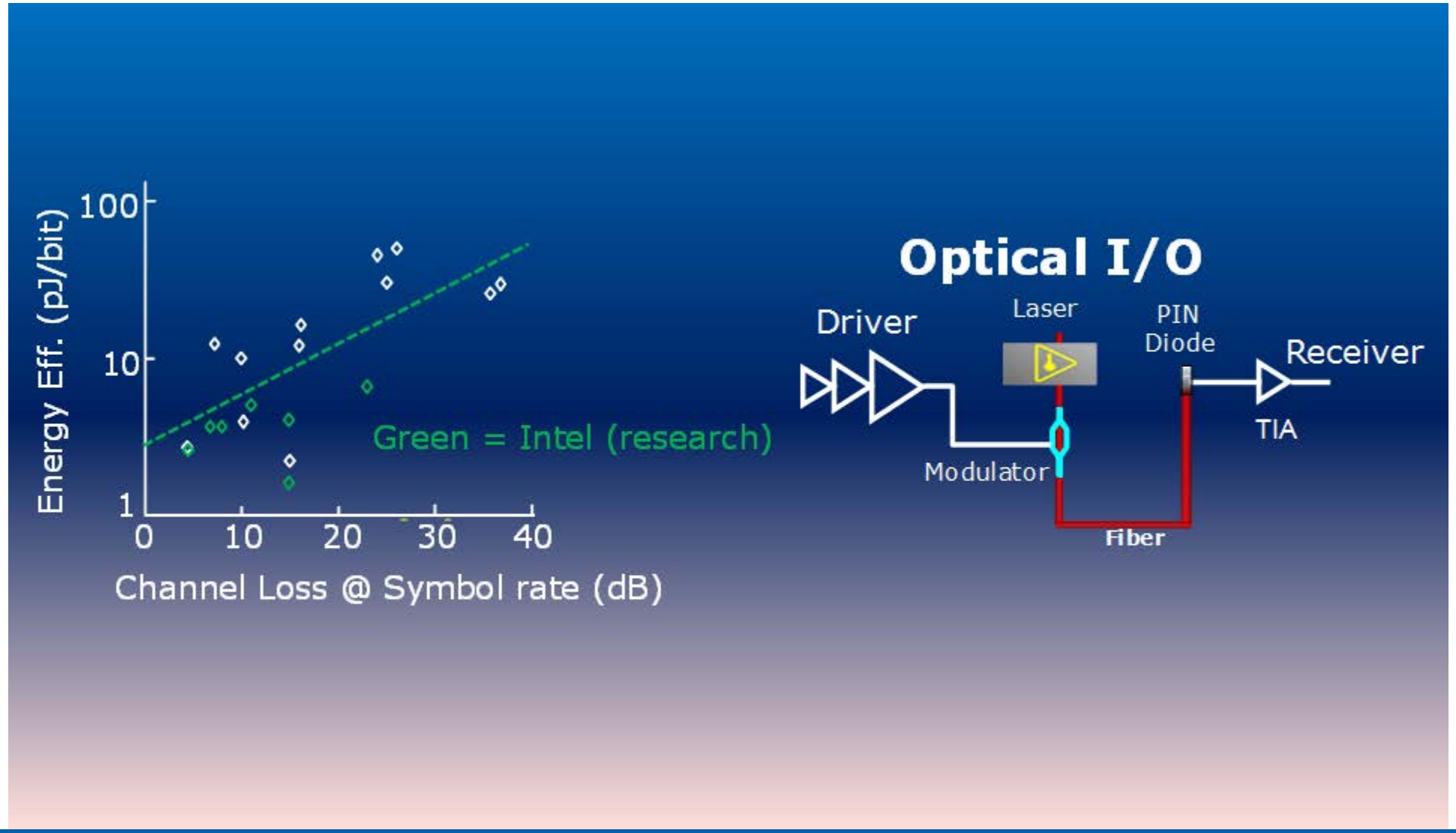


Fully integrated VR

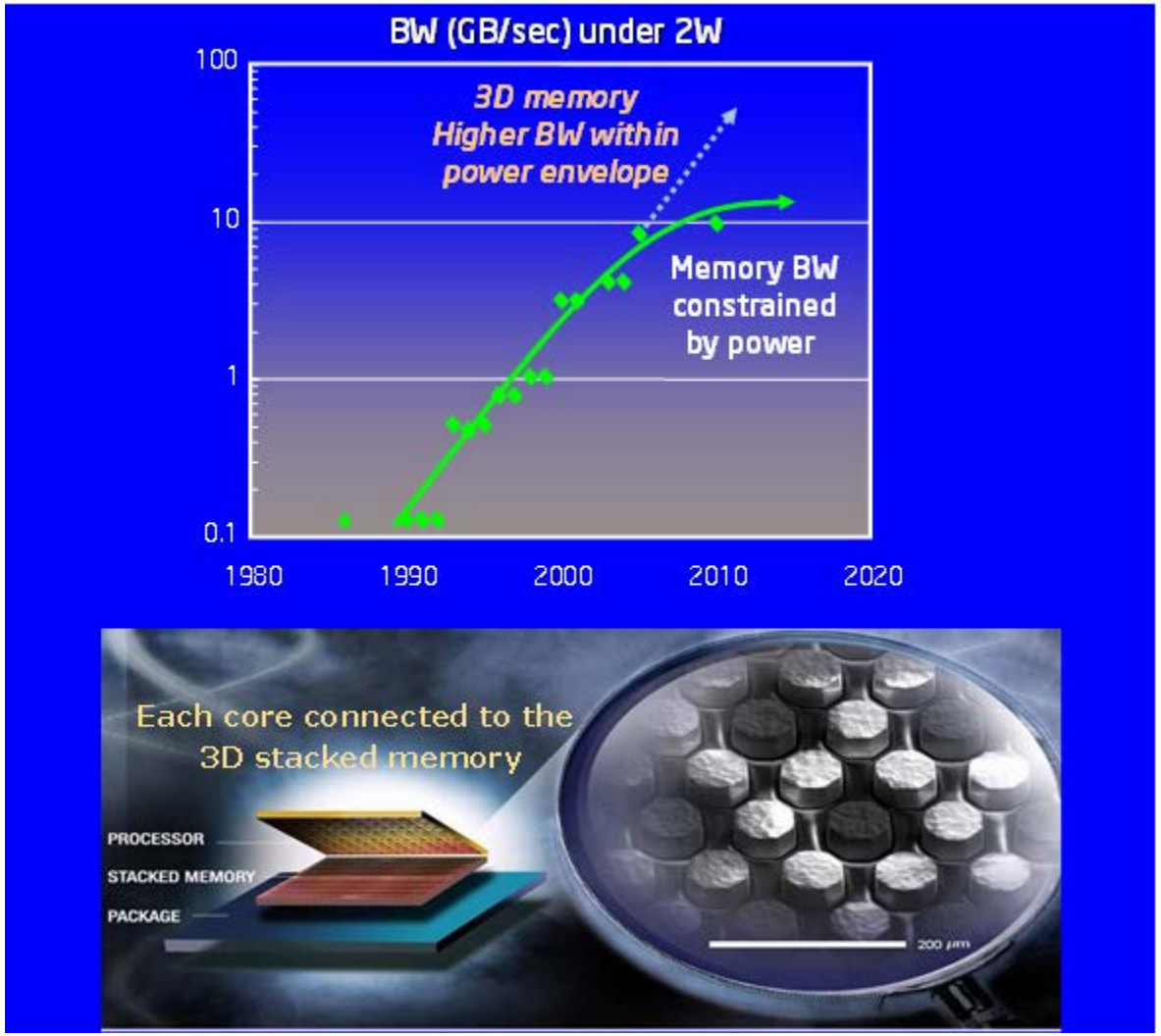




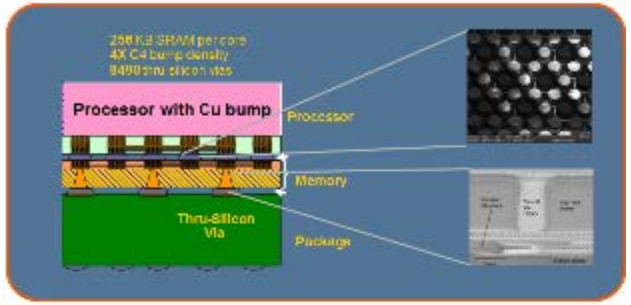
Energy efficient interconnects



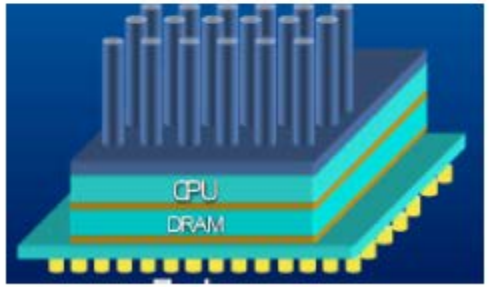
Memory capacity & bandwidth



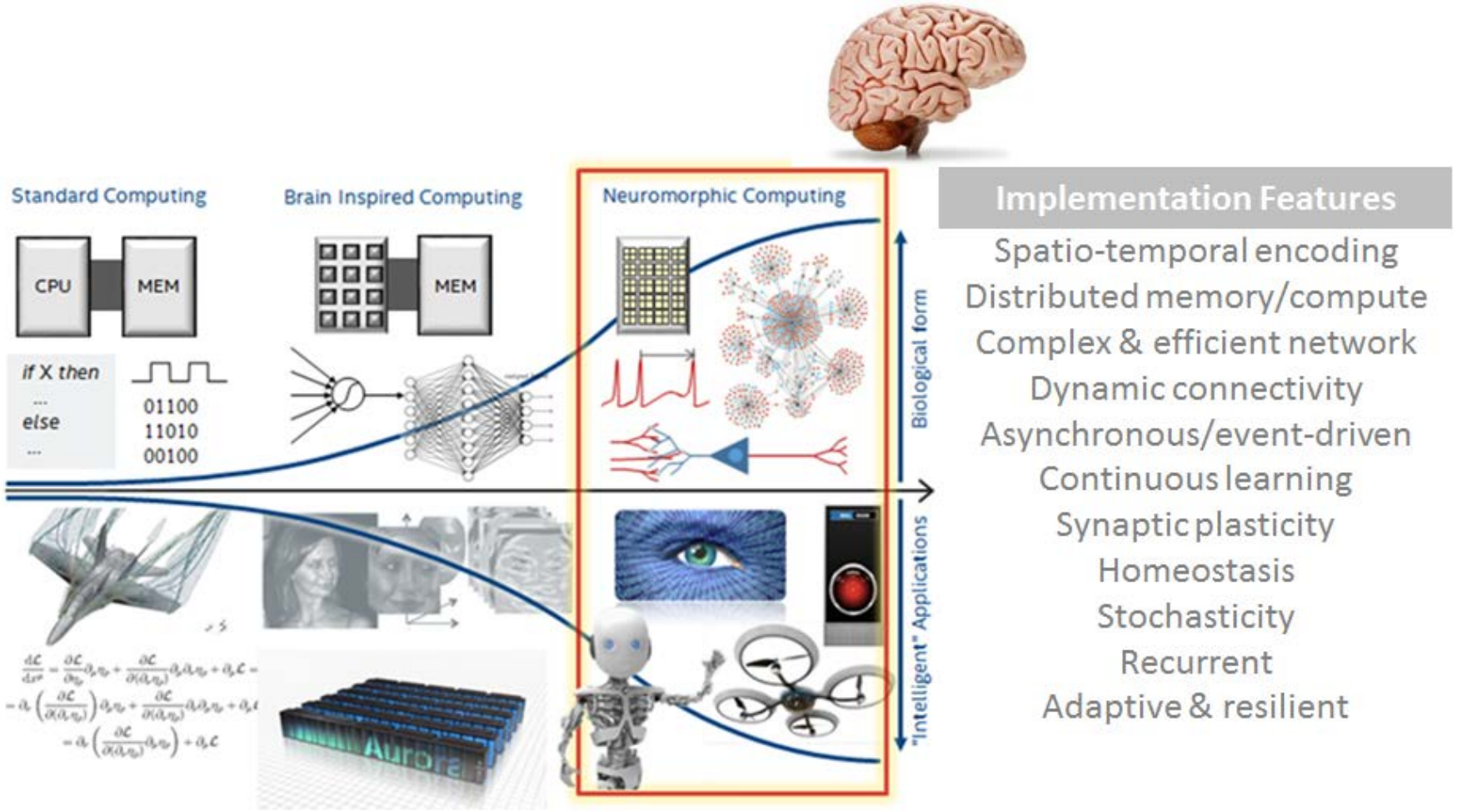
3D Integration: SRAM



3D Integration: DRAM



Efficient & scalable neuromorphic systems

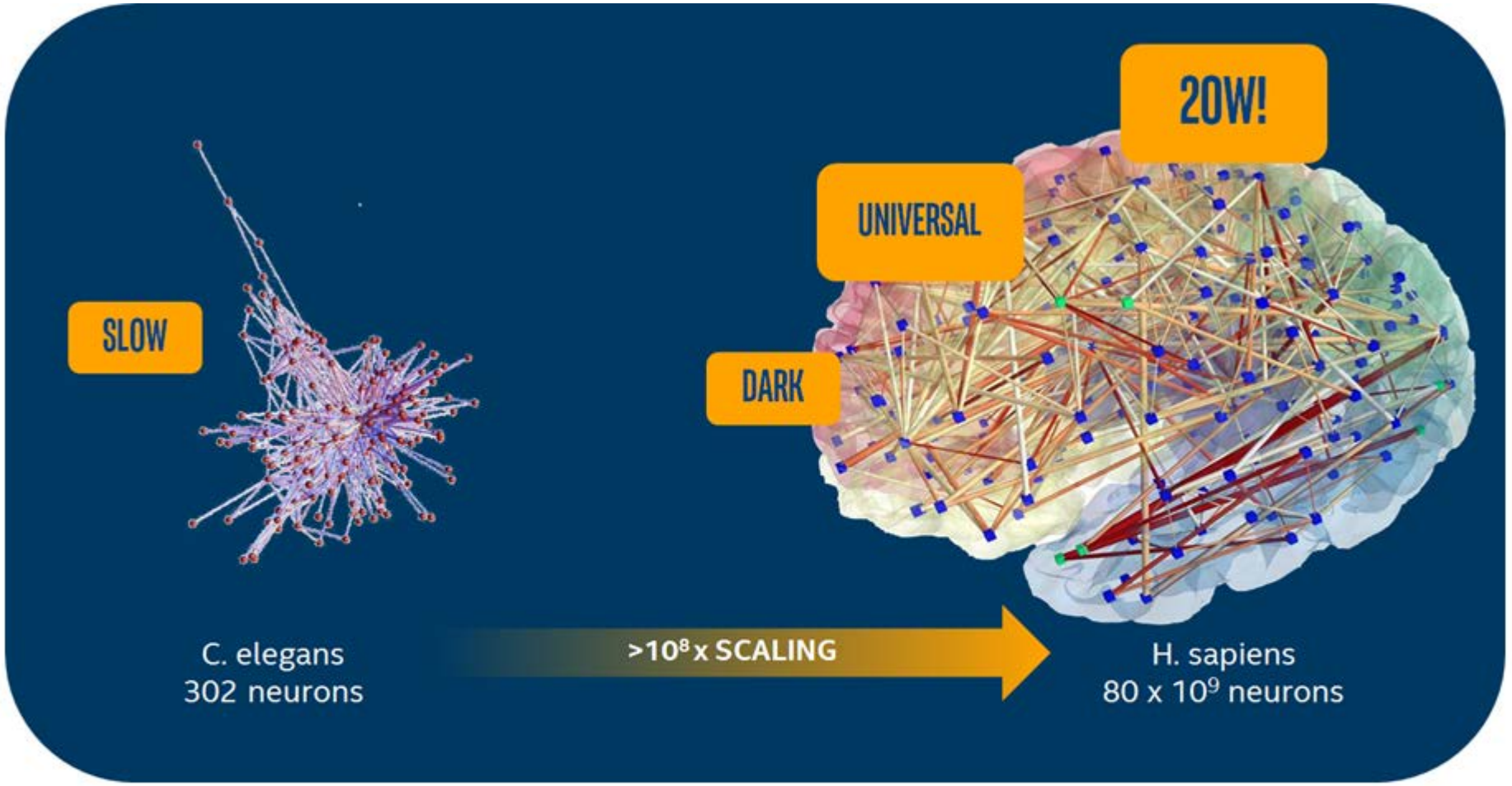


Implementation Features

Spatio-temporal encoding
Distributed memory/compute
Complex & efficient network
Dynamic connectivity
Asynchronous/event-driven
Continuous learning
Synaptic plasticity
Homeostasis
Stochasticity
Recurrent
Adaptive & resilient

Exploit integration, spike timing, sparsity, plasticity & resiliency

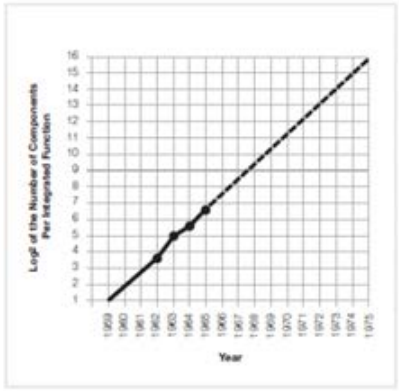
Efficient & scalable neuromorphic architecture



Phenomenal architecture efficiency & scalability in nature!

The next big leap…

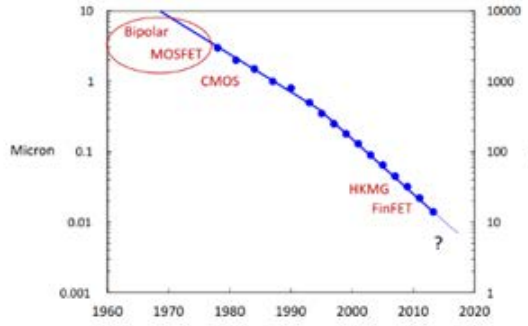
Moore's Law: Economics AND Power



Heat problem

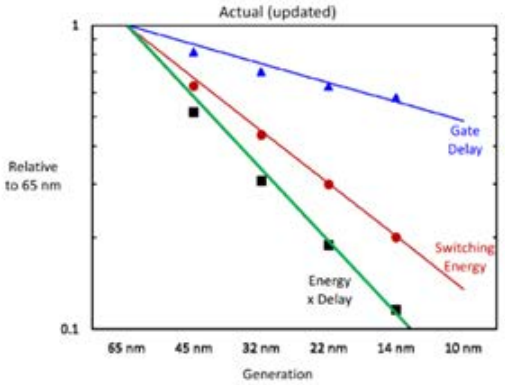
"Will it be possible to remove the heat generated by tens of thousands of components in a single silicon chip?"

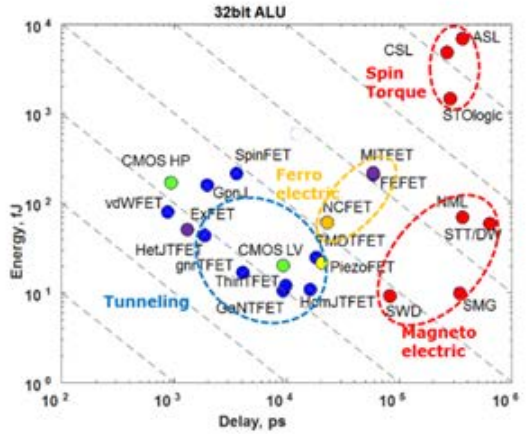
"Cramming more components onto integrated circuits", Electronics, Volume 38, Number 8, April 19, 1965



W. Holt, 2016 ISSCC

CMOS & Beyond





D. Nikonov & I. Young, 2015 JXCDC

Big leaps in energy efficiency trigger technology transitions!


LIMITLESS
POSSIBILITIES
WE ARE HERE