Energy Efficiency in the AI Computing Era

The Future of Power Delivery Inside the Processor

F. Carobolante, Intel Corp.

A Journey

- “What got us here won’t get us there”
- One size doesn’t fit all
- Looking into the crystal ball

A homage to Intel – 20 years of research in PwrSoC

PWR SOC 08

- *Ted DiBene "Power on Silicon with on-die magnetics"*
- *Dominik Schmidt "Challenges and Solutions: Power Delivery and Regulation in NanoCMOS SoCs"*
- *Donald S. Gardner "Integrated On-Chip Inductors Using Magnetic Material"*
- *Ted DiBene "Fine Grain On-die Integrated Magnetics; Breaking the Power/Performance Barriers"*
- *Rinkle Jain "Enabling Aggressive Dynamic Voltage and Frequency Scaling in Many Voltage Domains"*
- *Mondira (Mandy) Pant "The Era of Intelligent Power Delivery"*
- *Rinkle Jain "Distributed Power Conversion – An Answer to Power Delivery Challenges in SoCs?"*
- *Amit Jain "FIVR Control topology and design for distributed loads"*
- *Christopher Schaef "Potential of Hybrid Converters in Compute Platform Power Delivery"*
- *Rinkle Jain "Fine Grain Voltage Domains on Graphics"*
- *Ravi Mahajan "Advanced Packaging Architectures for Heterogeneous Integration"*
- *Vivek De "System-Level Power Management Strategies for Integrated Platforms"*
- *Ravi Mahajan "Advanced Packaging Architectures for Heterogeneous Integratio*

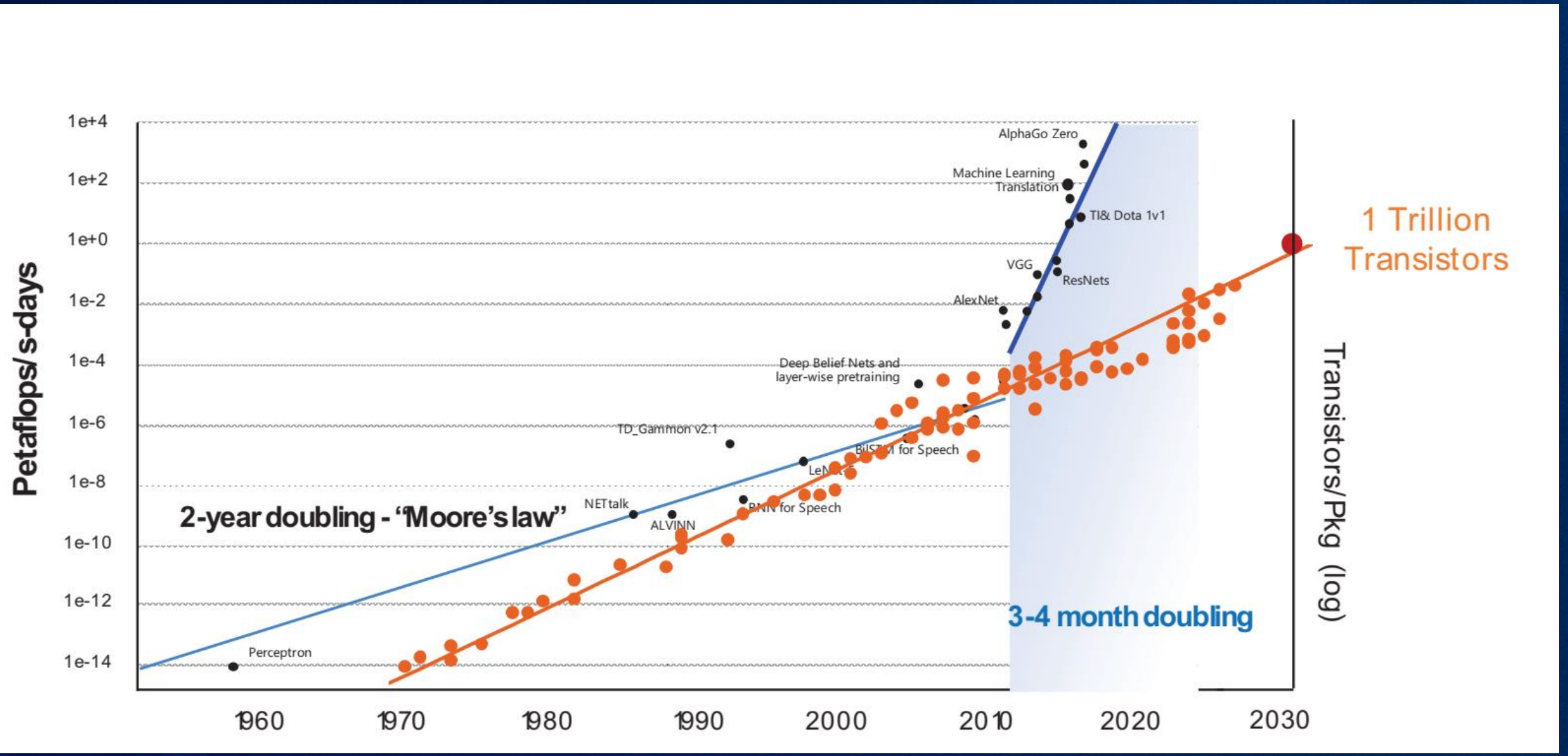
PWR SOC 23

- *Kaladhar Radhakrishnan "Magnetic Inductors for Next Generation IVR"*
- *Han Wui Then "GaN-on-Si Process Featuring GaN MOSHEMT Transistor Technology a Integrated Silicon CMOS on 300mm Wafers"*
- *Nicolas Butzen "Next-Generation Switched-Capacitor Converters using High-Densit MIM Capacitors"*

“What got us here won’t get us there”

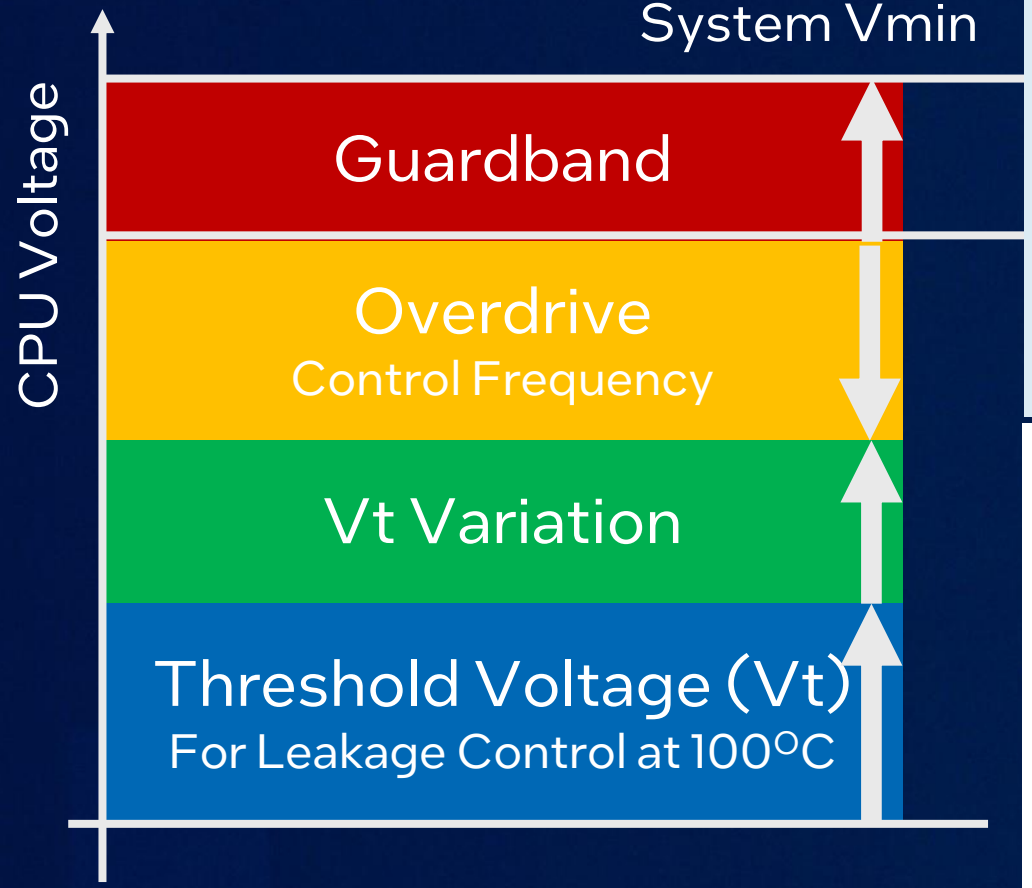
intel

The Challenges

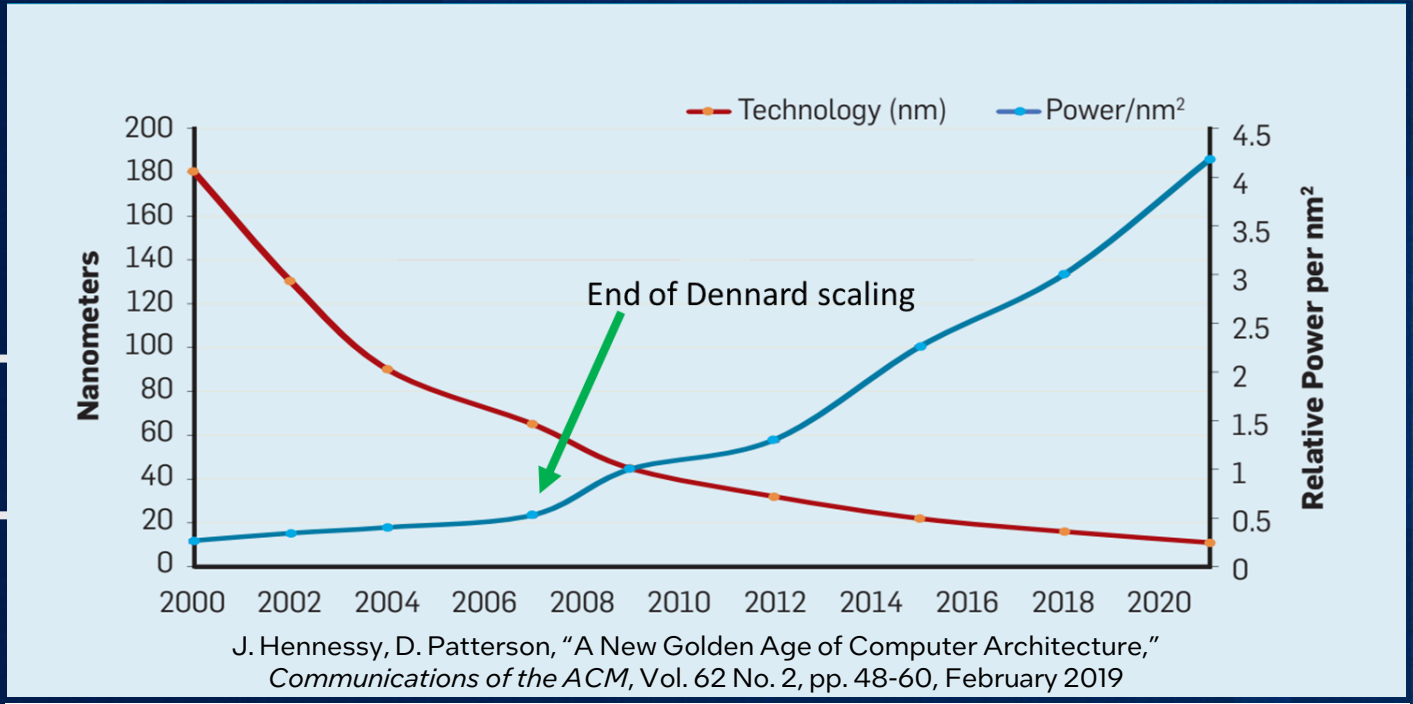


Wilfred Gomes "Beyond Exascale: A paradigm shift for AI and HPC" (Invited), IEDM 2023

The 1,000 A challenge



Power increases by 3% for every 10mV increase in supply voltage!



J. Hennessy, D. Patterson, "A New Golden Age of Computer Architecture," *Communications of the ACM*, Vol. 62 No. 2, pp. 48-60, February 2019

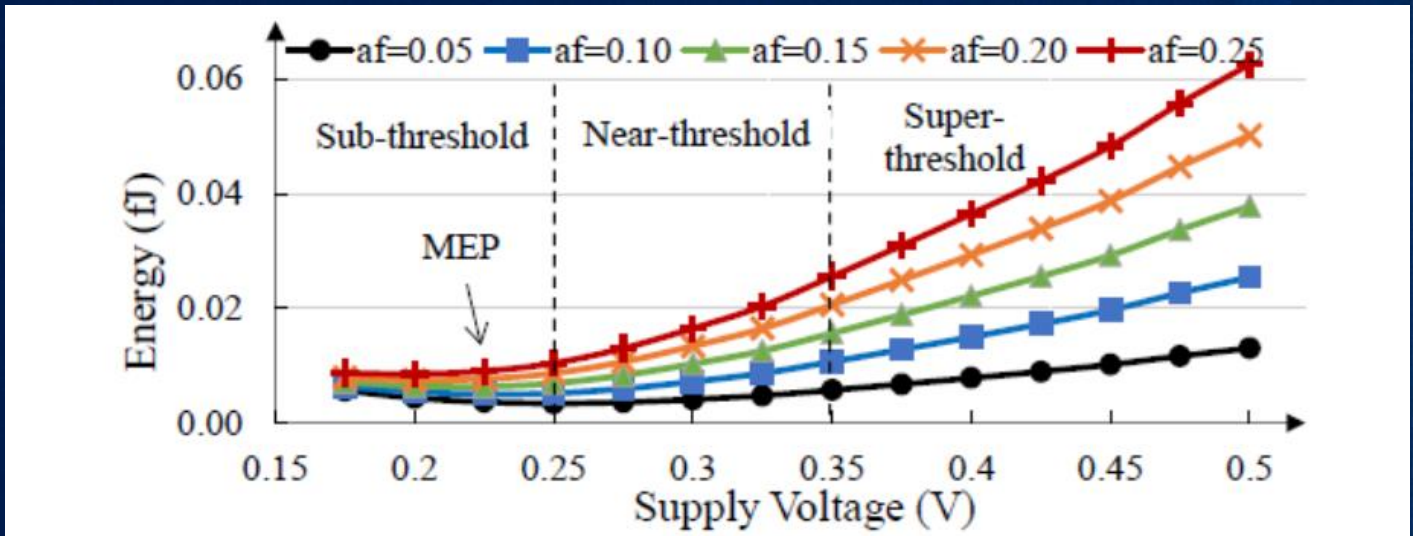
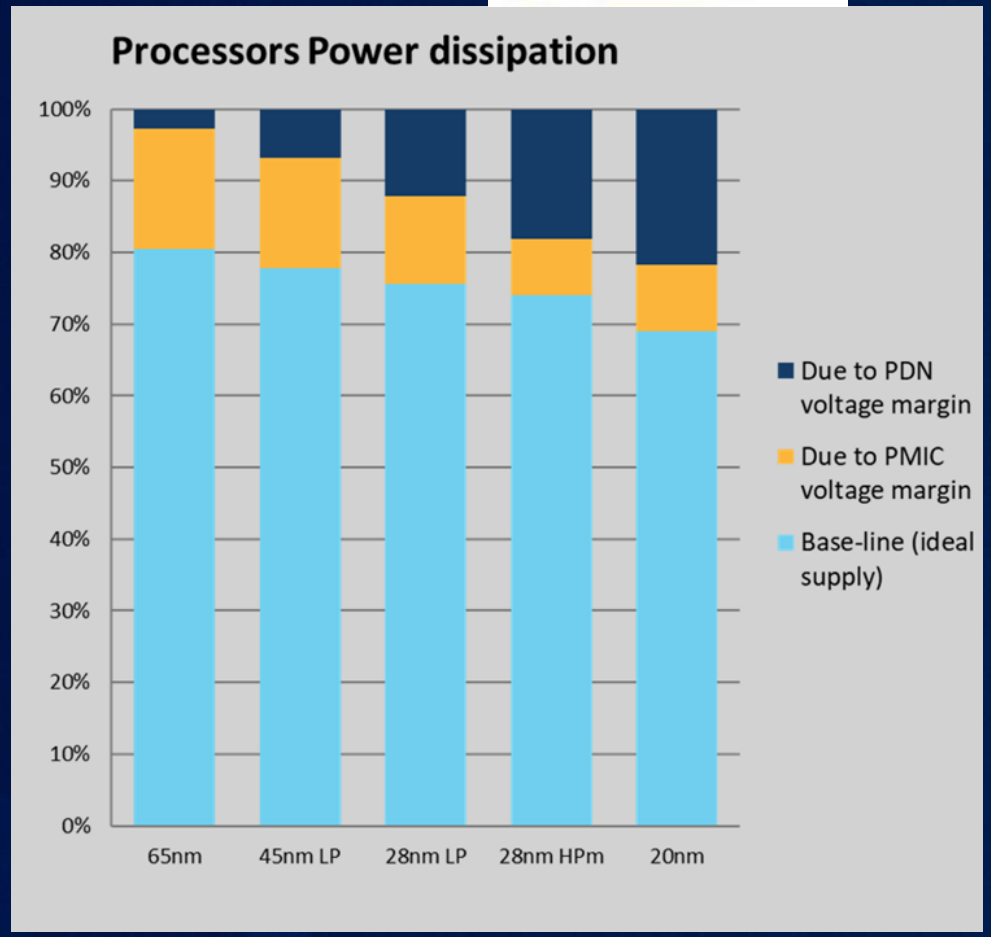


Figure 4. Energy consumption of a 20-stage inverter chain versus supply voltage at different activities factors (af) for FinFET 7nm normal V_{th} device.

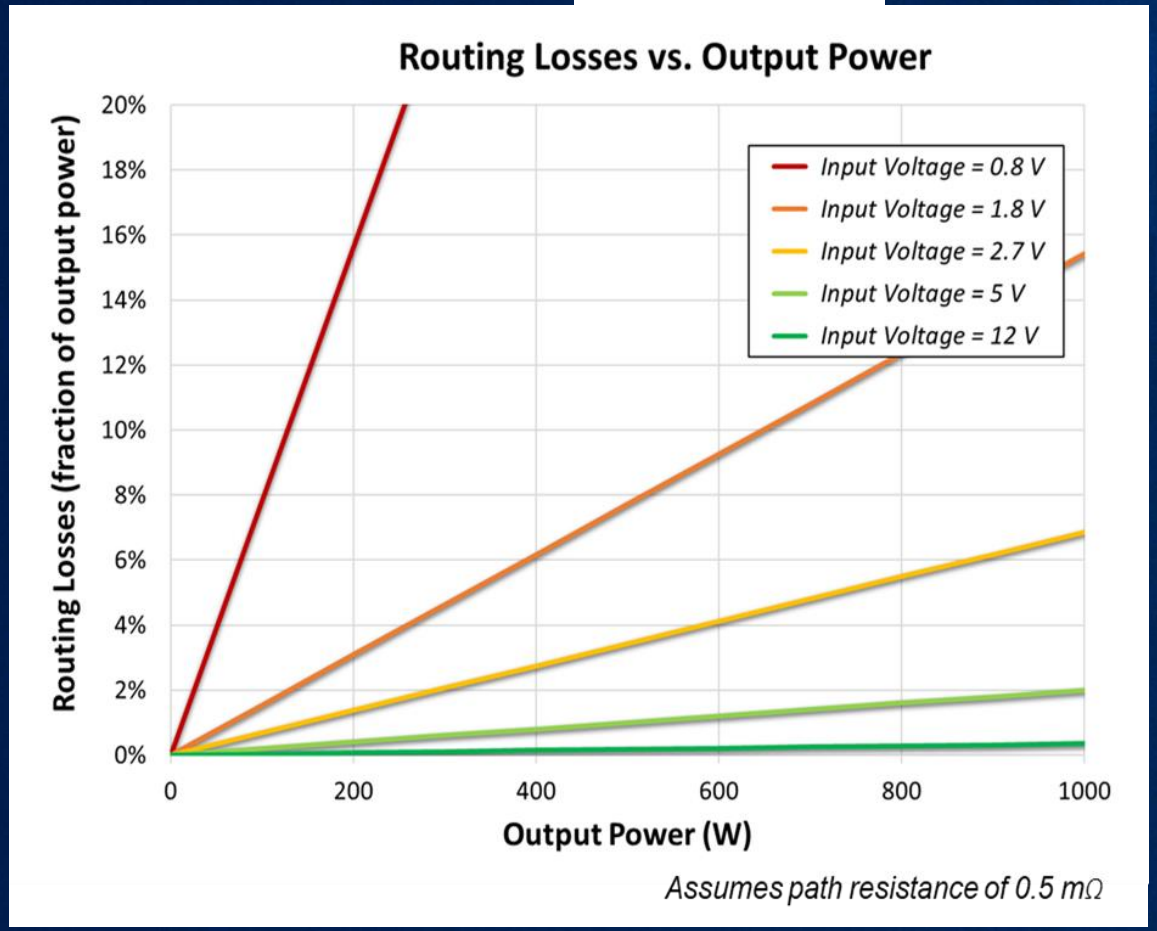
Qing Xie et al. "Performance Comparisons between 7nm FinFET and Conventional Bulk CMOS Standard Cell Libraries," IEEE Transaction on Magnetics, 2015

The snowballing effect of high current and power density

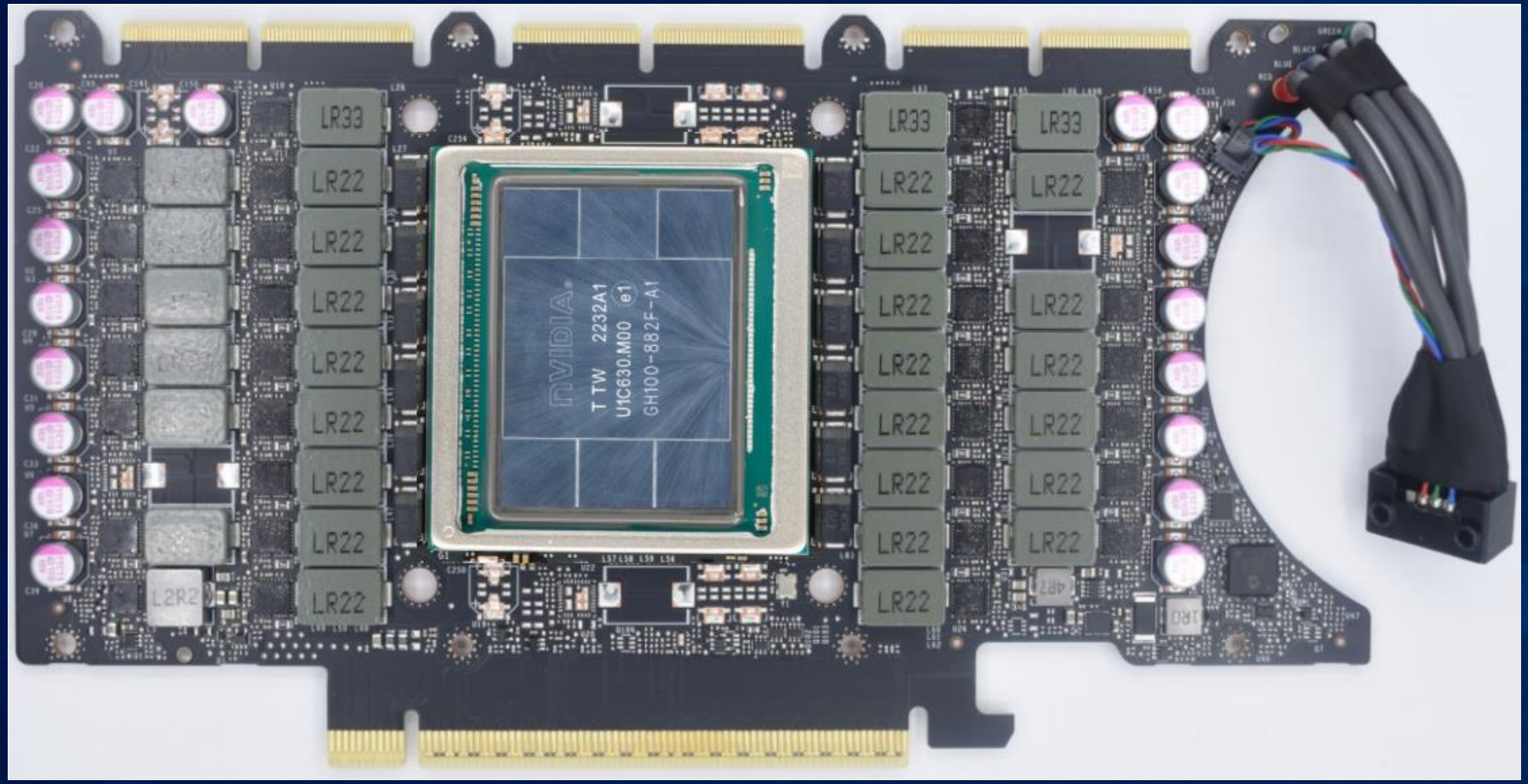
F. Carobolante



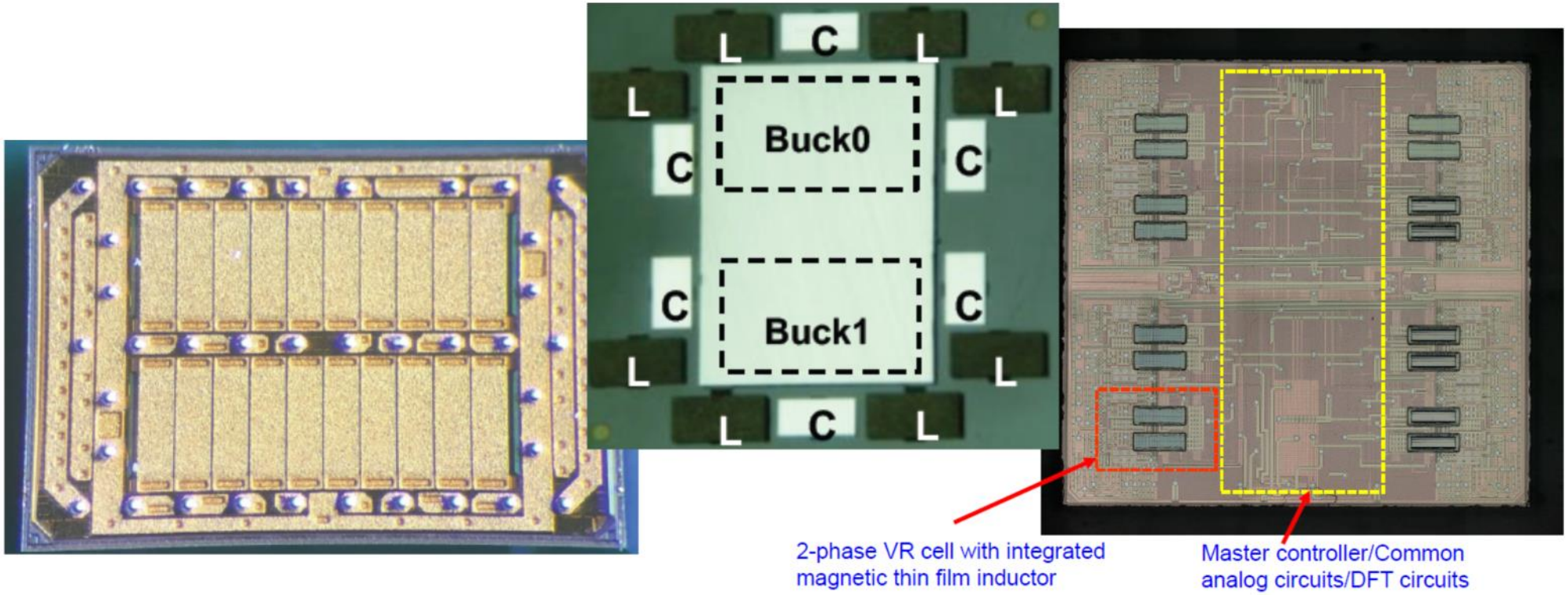
K. Radhakrishnan



When you have a single stage conversion…



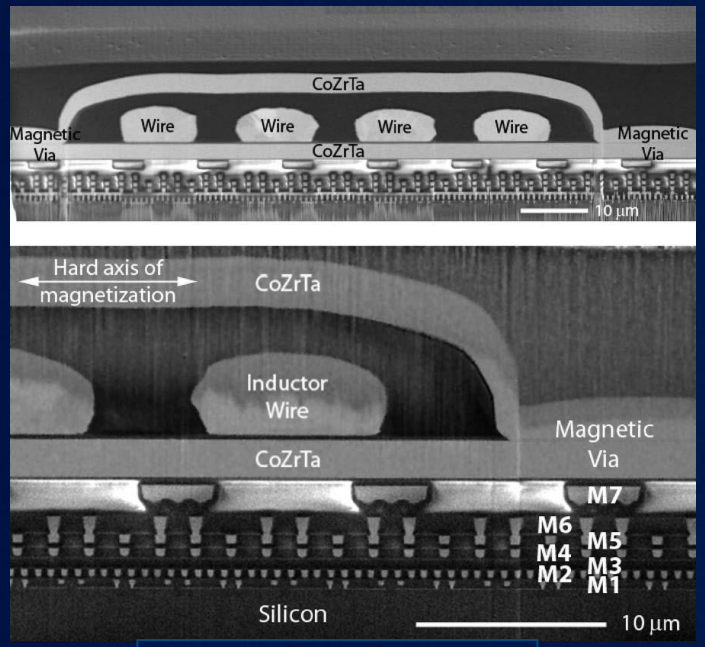
Ferric, Dialog and Huawei show fully integrated PwrSoC's



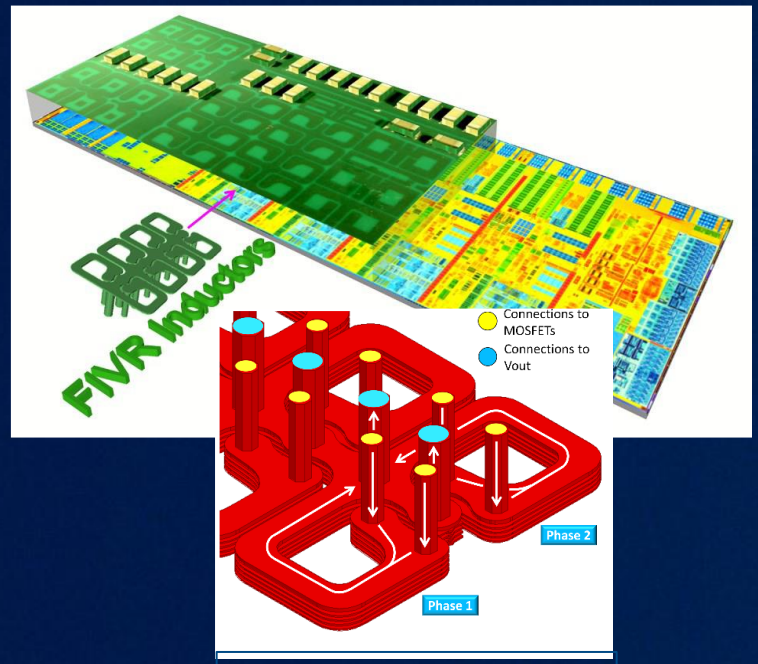
14

F. Carobolante "PwrSoC at an Inflection Point From R&D to Market Relevance" APEC 2021

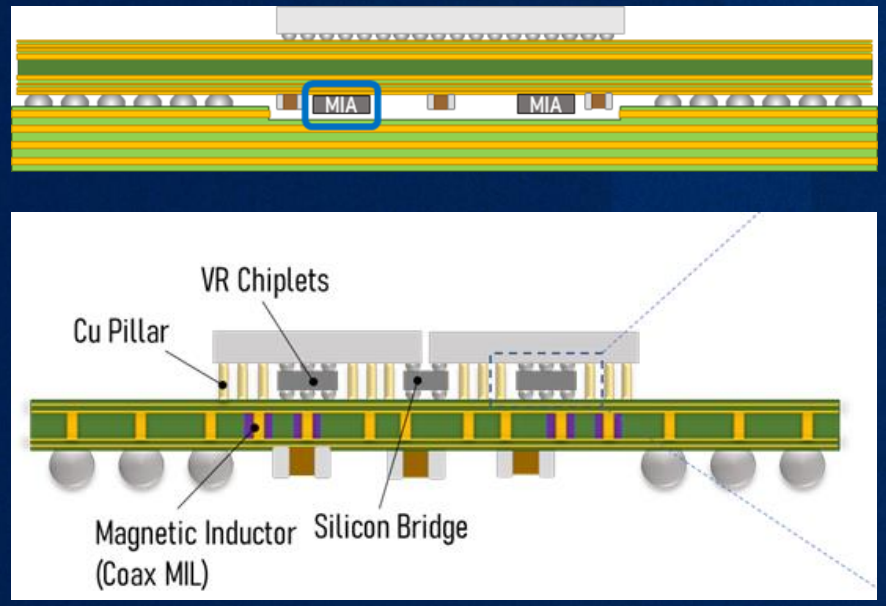
A Few Key IVR Milestones from Intel



Intel Magnetics on Silicon 2008



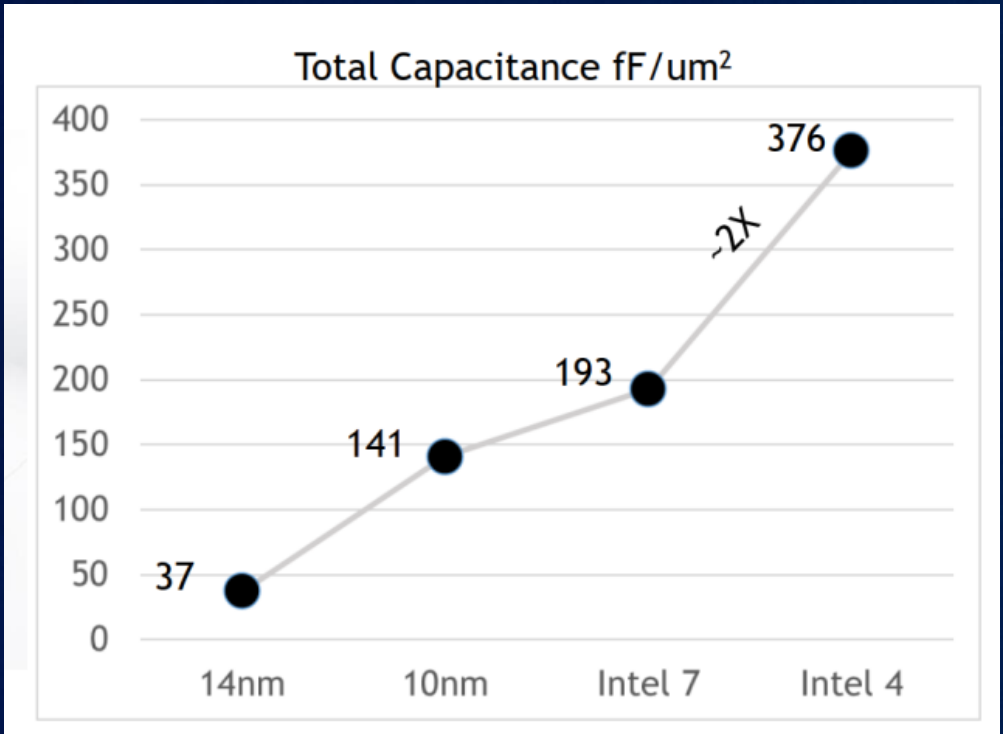
Intel Haswell 2014



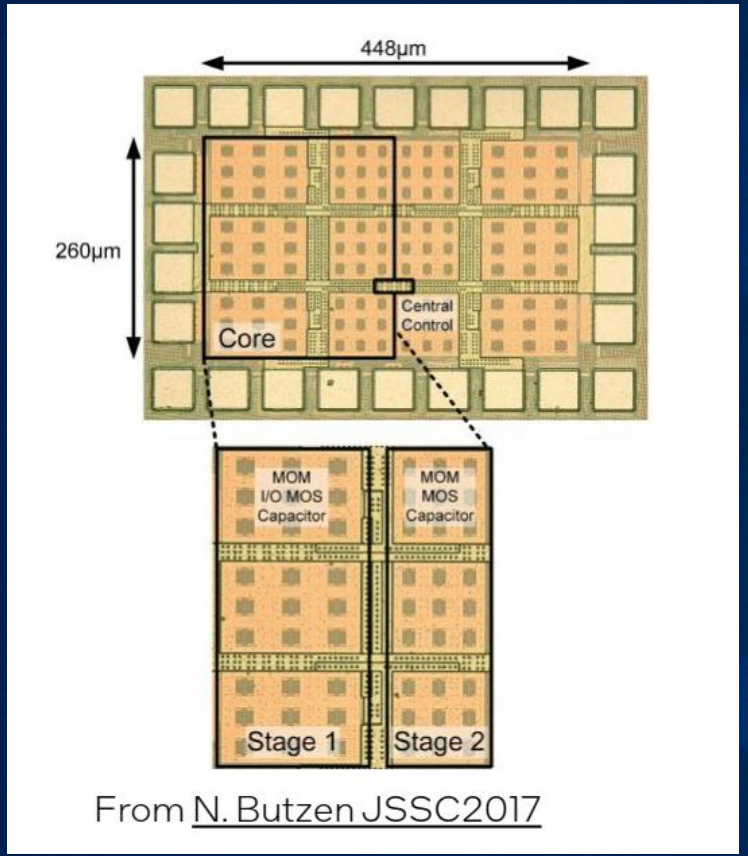
Intel @ PwrSoC 2021

Good capacitors are your best friends

- Deep Trench… not compatible with transistors
- MIM caps: achieving high density!



From https://www.semiconductor-digest.com/intel-4-process-drops-cobaltinterconnect-goes-with-tried-and-tested-copper-with-cobalt-liner-cap/

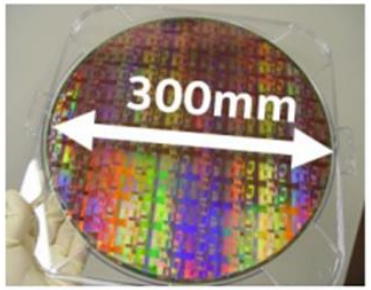


From N. Butzen JSSC2017

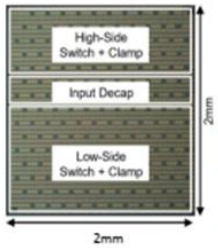
N. Butzen "Next-Generation Switched-Capacitor Converters using High-Density on-die MIM Capacitors" PwrSoC 2023

Need a Doctor? Dr. GaN [GaN Research at Intel]

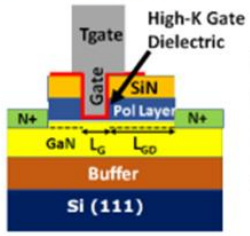
300mm GaN-on-Si(111) Process

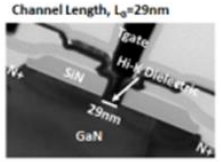


Power GaN Die (W=1000mm)



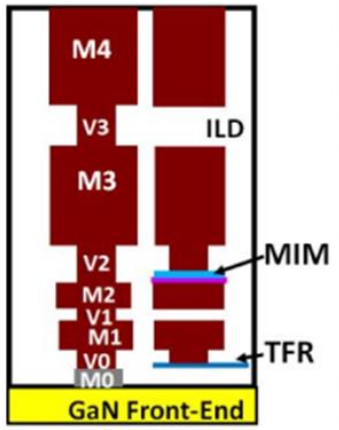
- 300mm Si (111) HR substrate
- High-k E-mode MOSHEMT
- Schottky GaN HEMT
- Min channel Lg 30nm
- Regrown N+ Source/drain

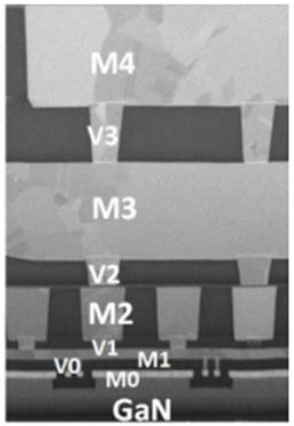




E-mode high-K GaN Transistor

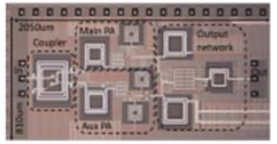
Backend Metal Interconnect

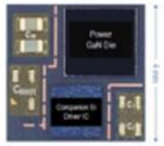


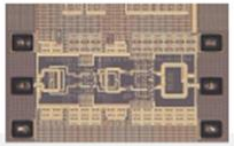


- 4 Cu metal layers
- Passives: inductor, MIM and TFR

Circuit Research



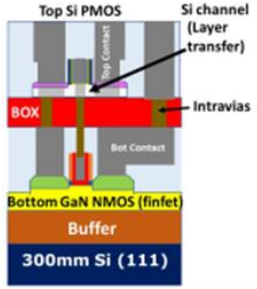


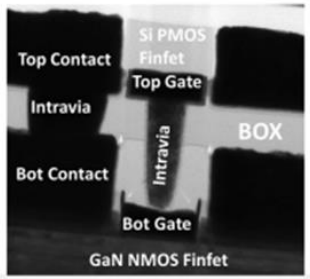


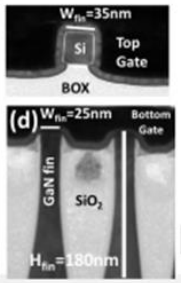
- Power electronics
- Sub-7Ghz Doherty PA
- 28 – 40 GHz PA, LNA

VLSI ‘21, ‘22

GaN and CMOS Integration



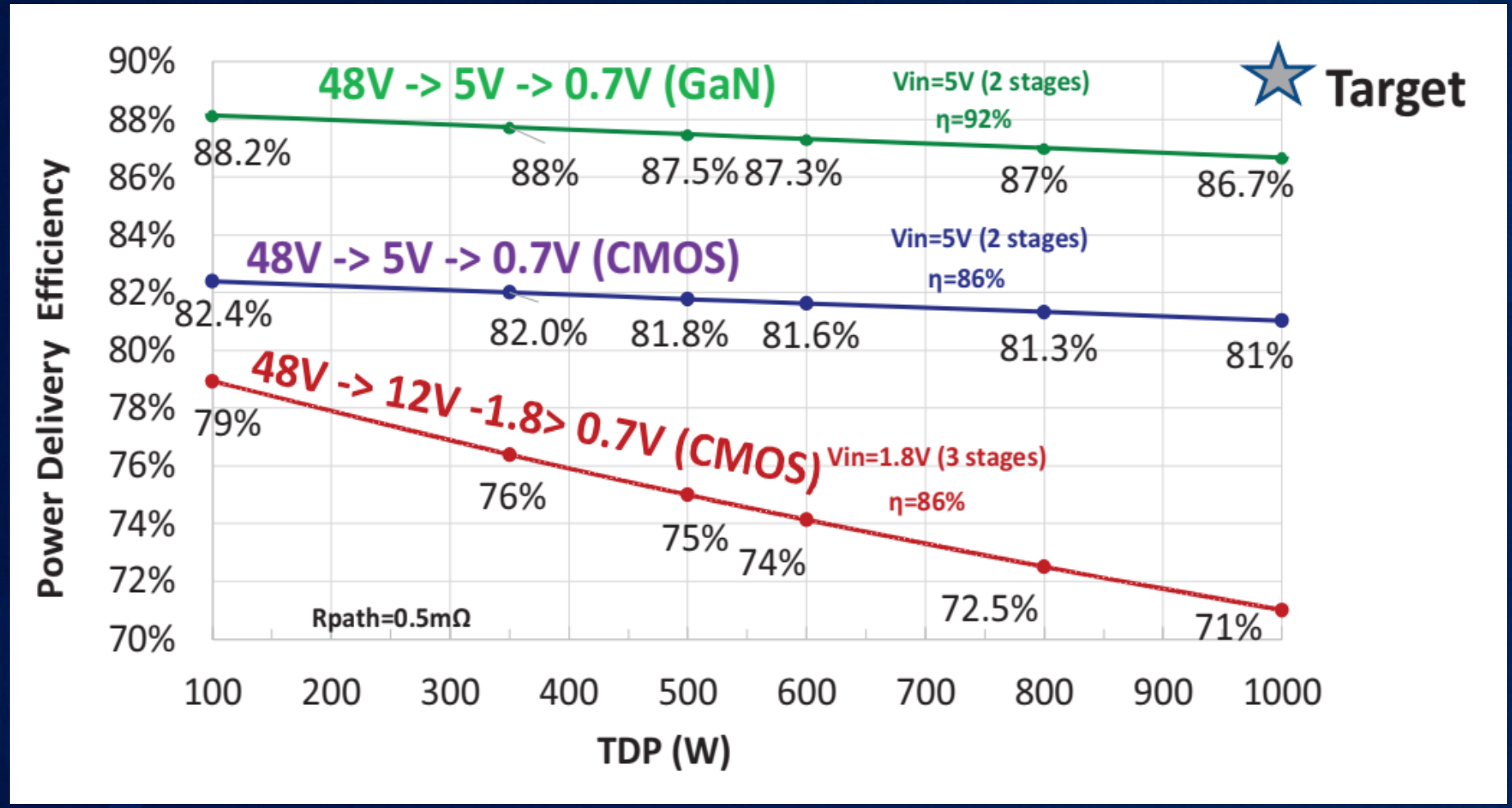




IEDM ‘19, ‘21

Han Wui Then “GaN-on-Silicon Process Technology” PwrSoC 2023

Efficiency projections

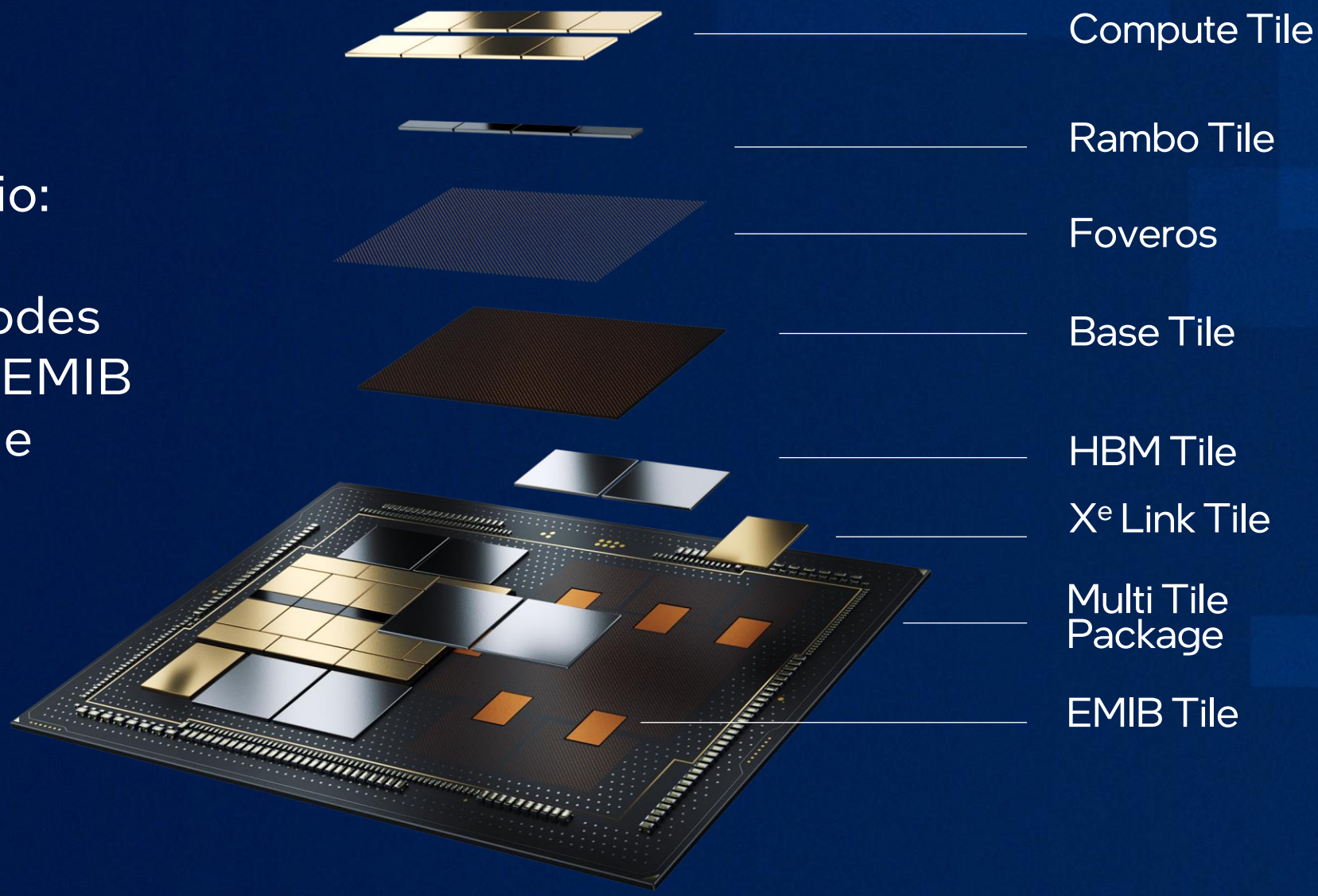


Wilfred Gomes "Beyond Exascale: A paradigm shift for AI and HPC" (Invited), IEDM 2023

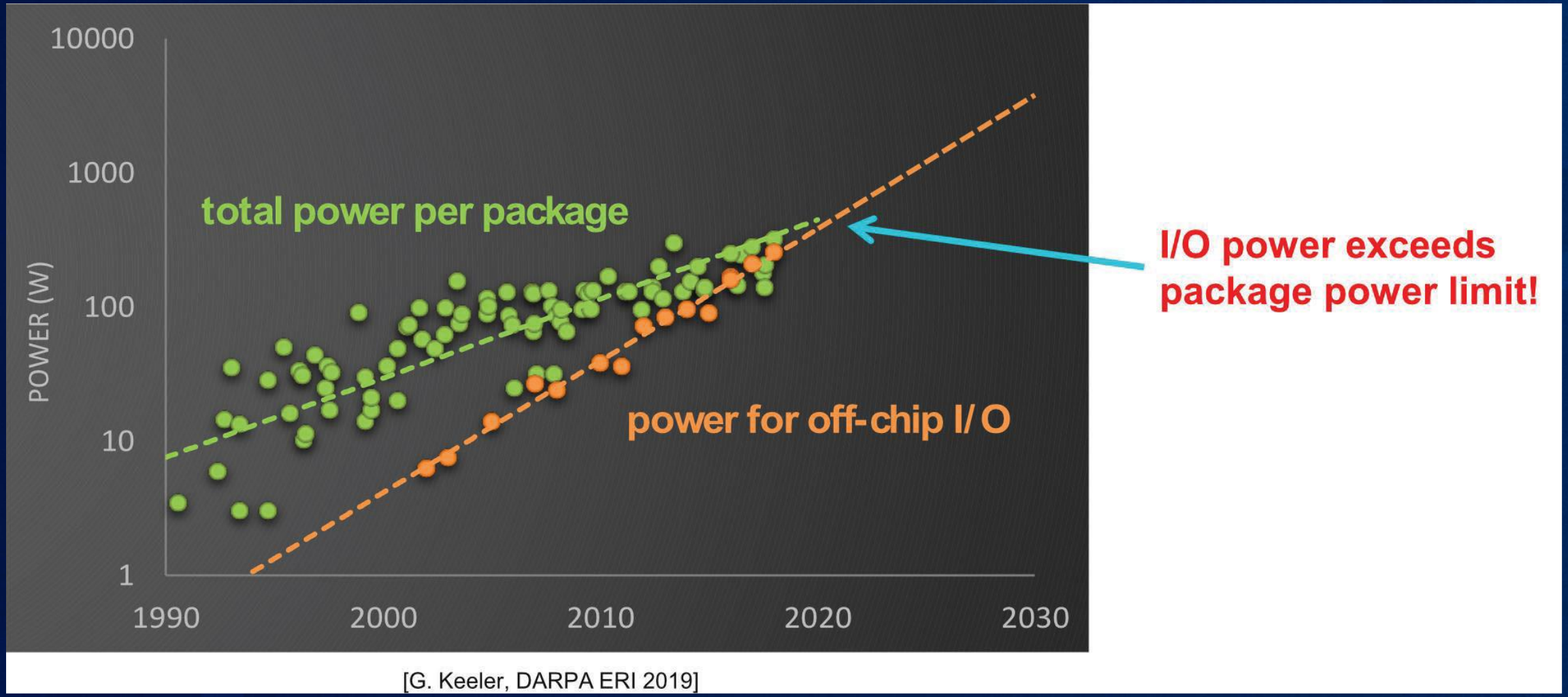
Now: Chiplets and Heterogeneous Integration

Intel Ponte Vecchio:

- 47 Tiles
- Five Process Nodes
- FOVEROS and EMIB Integration in the same Package



...and the need for integrating Optics



[G. Keeler, DARPA ERI 2019]

One size doesn’t fit all

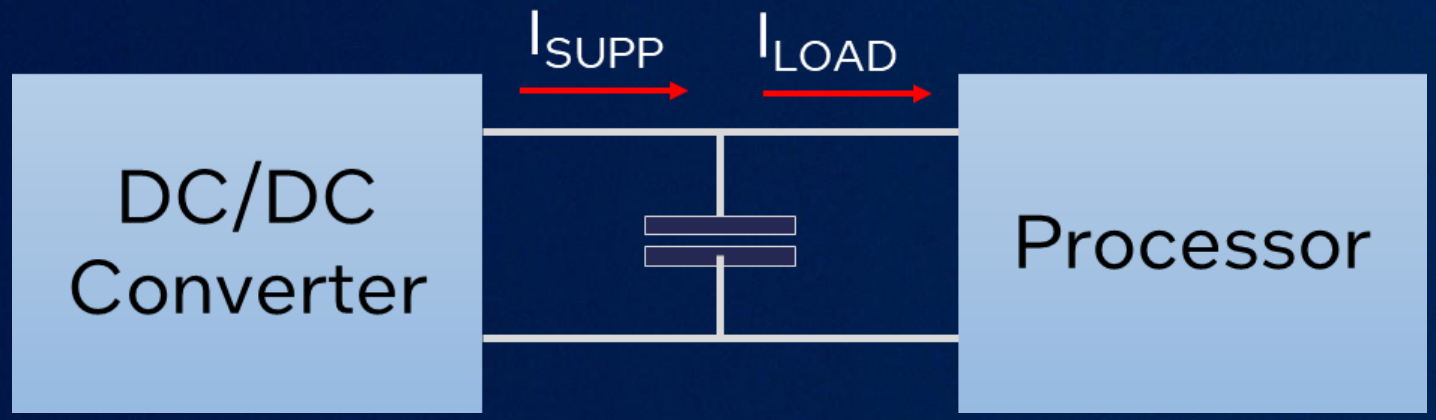
CPU Vs. GPU

CPU

- High Peak-to-Average ratio
  $I_{SUPP} = I_{LOAD\_AVERAGE} << I_{LOAD\_PEAK}$
- Unpredictable loads
- Hot spots limited

GPU

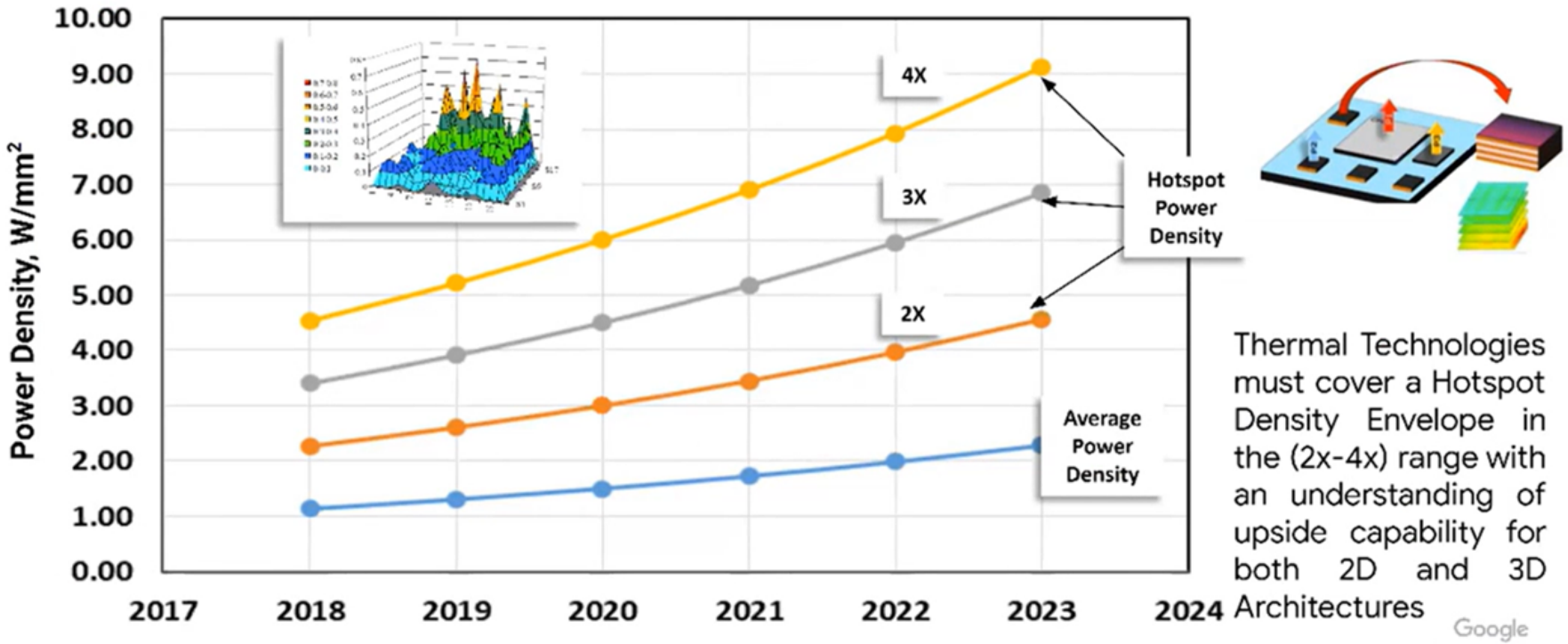
- Sustained high current
  I_{SUPP} approx. = I_{LOAD}
- Predictable loads, but…
- Thermally limited



Fine-grained power delivery & load balancing

Example of chip power density trends

Source: Ravi Mahajan, Weihua Tang, Intel, Materials to be published in 2020 HIR Thermal Chapter

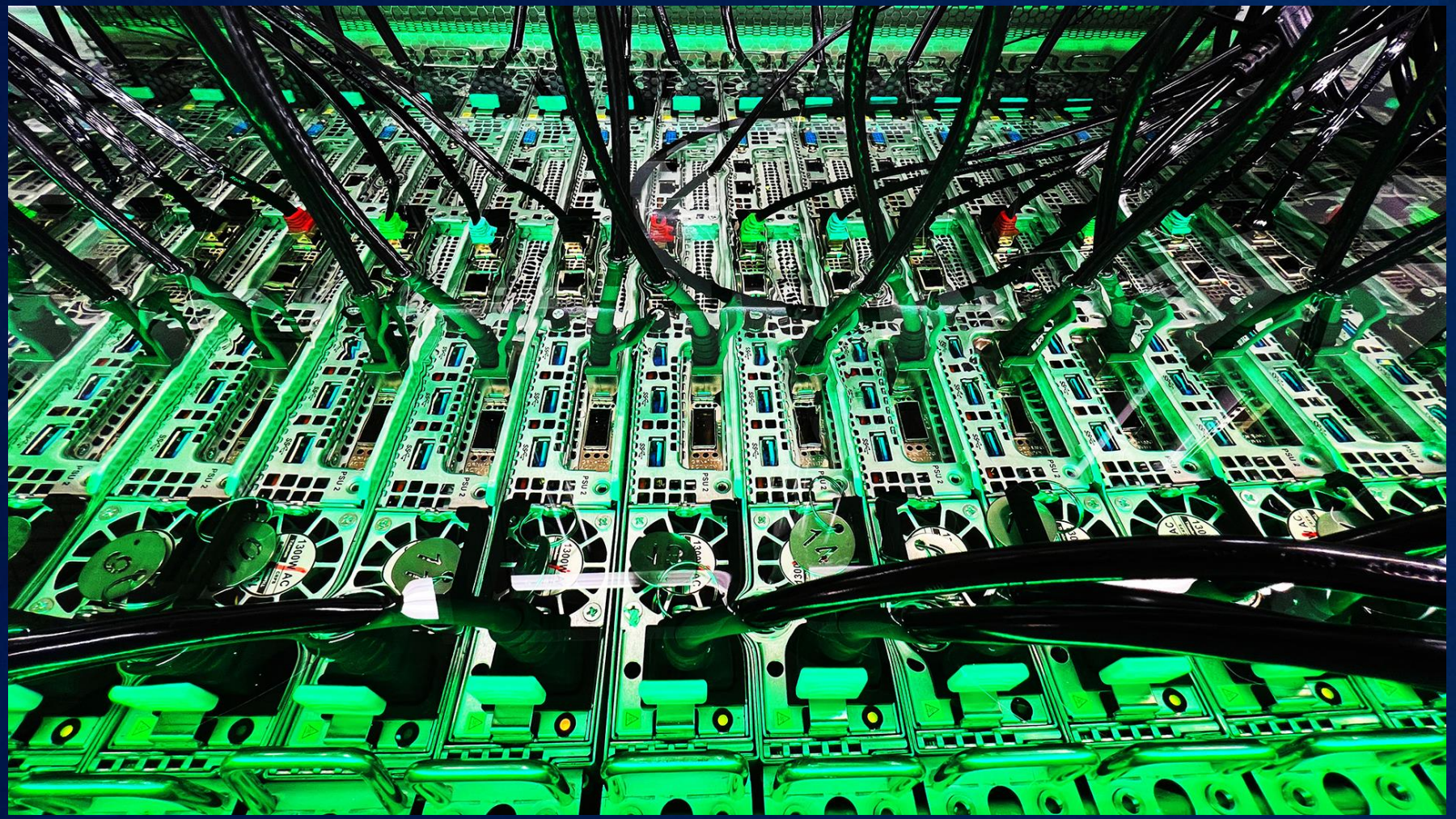


Thermal Technologies must cover a Hotspot Density Envelope in the (2x-4x) range with an understanding of upside capability for both 2D and 3D Architectures

Google

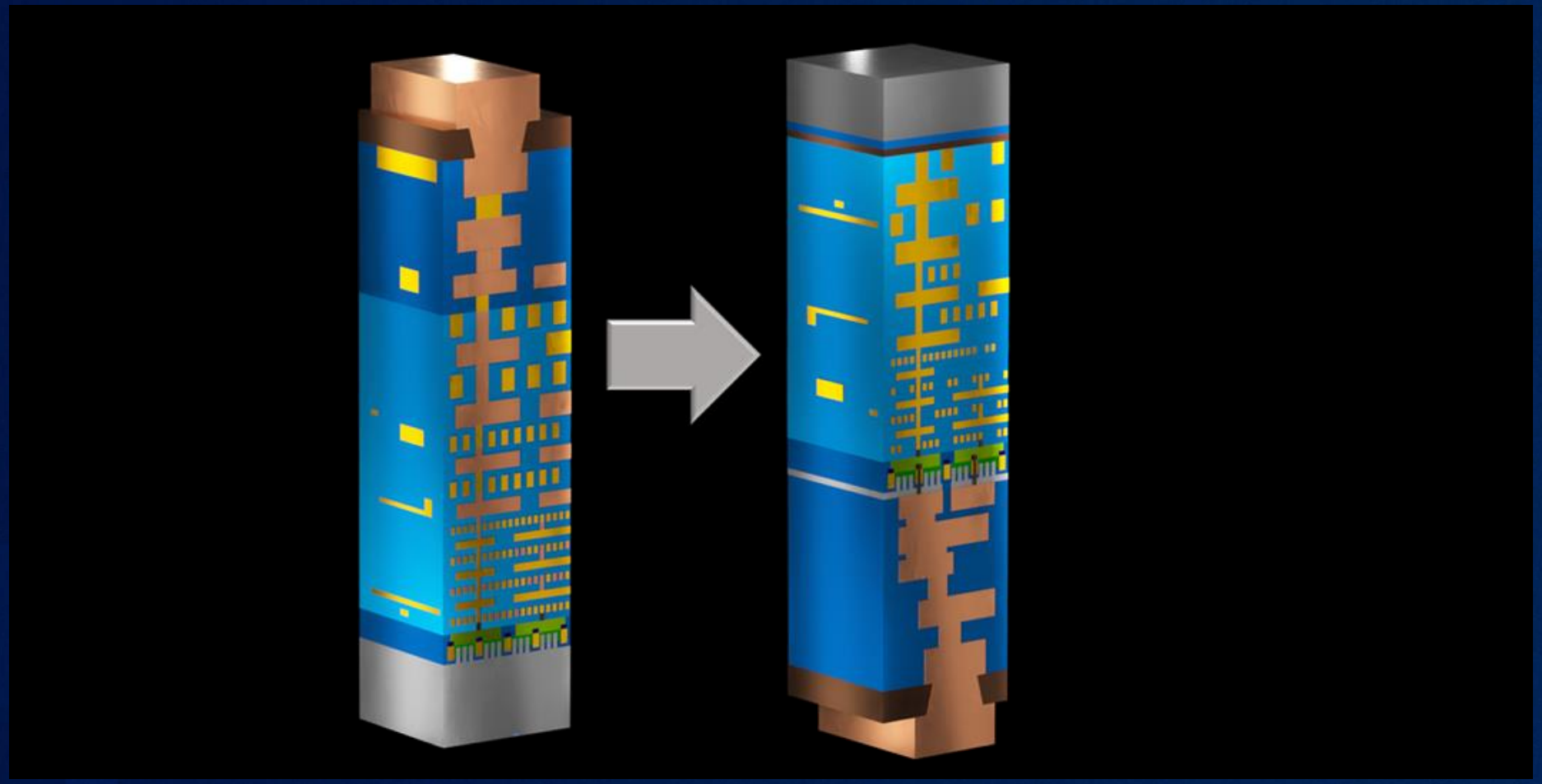
Ravi Mahajan & Weihua Tang, 2020 IEEE Heterogeneous Integration Roadmap (HIR)

Liquid immersion cooling



https://www.intel.com/content/www/us/en/newsroom/news/intel-dives-into-future-of-cooling.html

Optimizing both Signal and Power Integrity

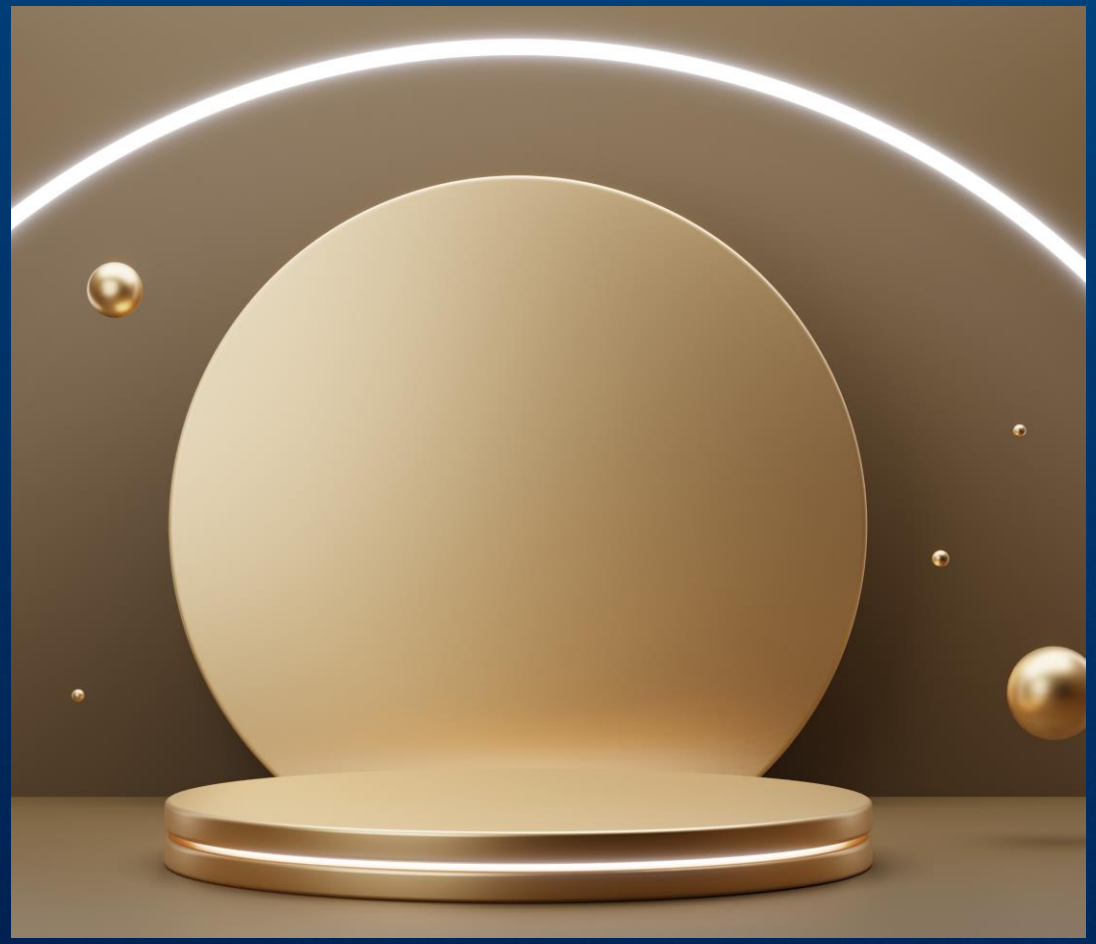


R. Mahajan, Advanced Packaging Architectures for Heterogeneous Integration, PwrSoC 2021

What about AI

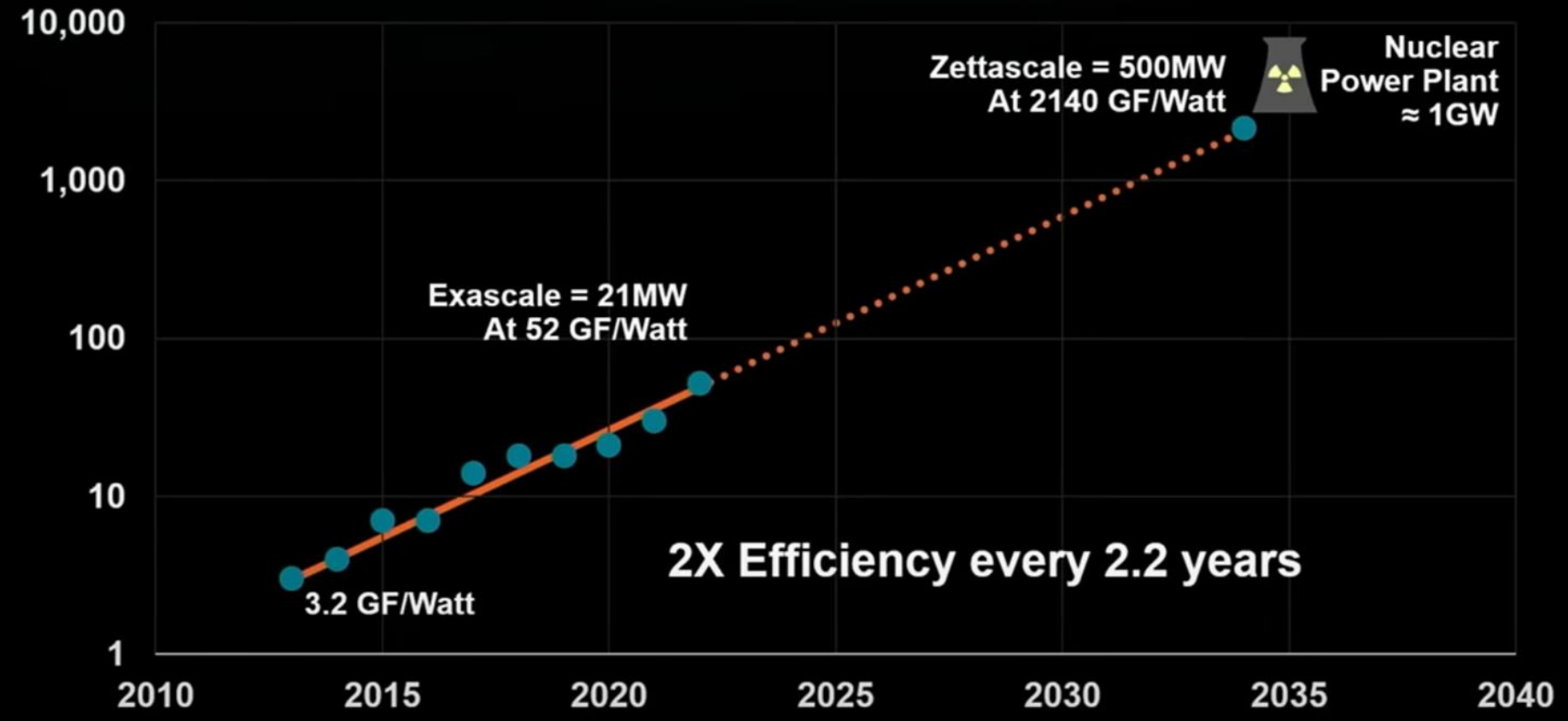
- Proactive Power Mgmt with ML
- Security

Looking into the crystal ball



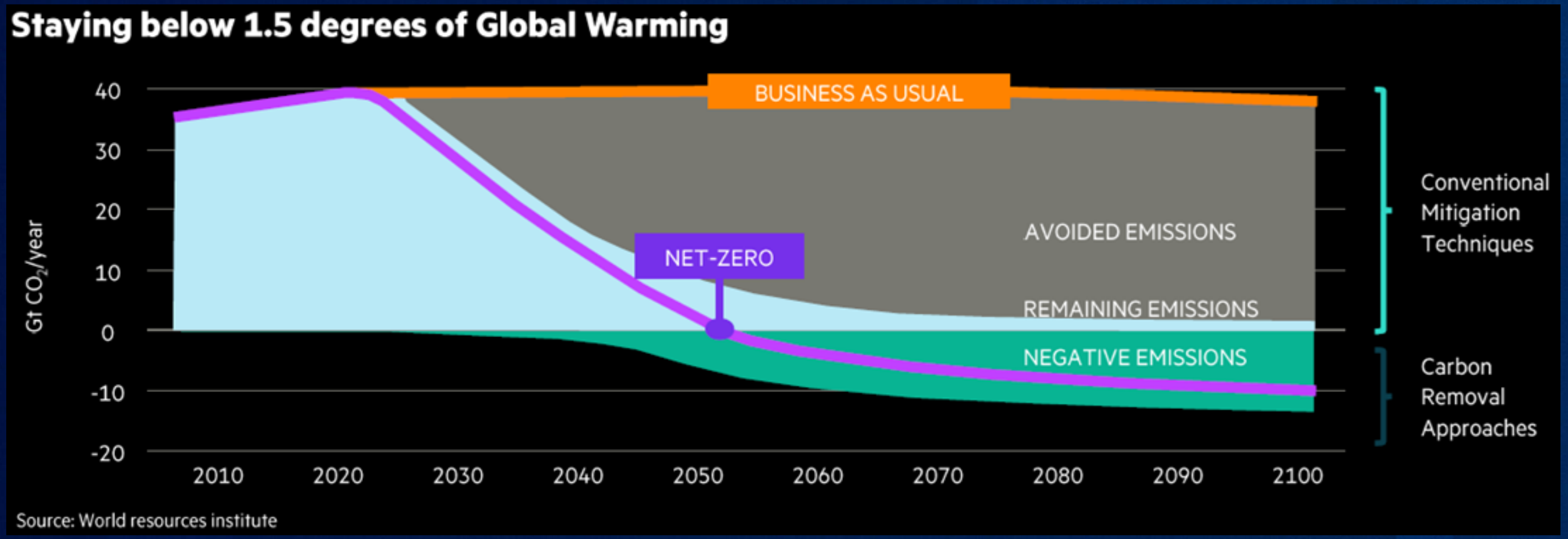
Supercomputer Energy Use Trajectory

Green500 Supercomputer GFLOPs/Watt and Projection



Lisa Su
ISSCC 2023

The Sustainability Challenge… Many are working on it!



IEEE HIR HETEROGENEOUS INTEGRATION ROADMAP

Semiconductor Research Corporation

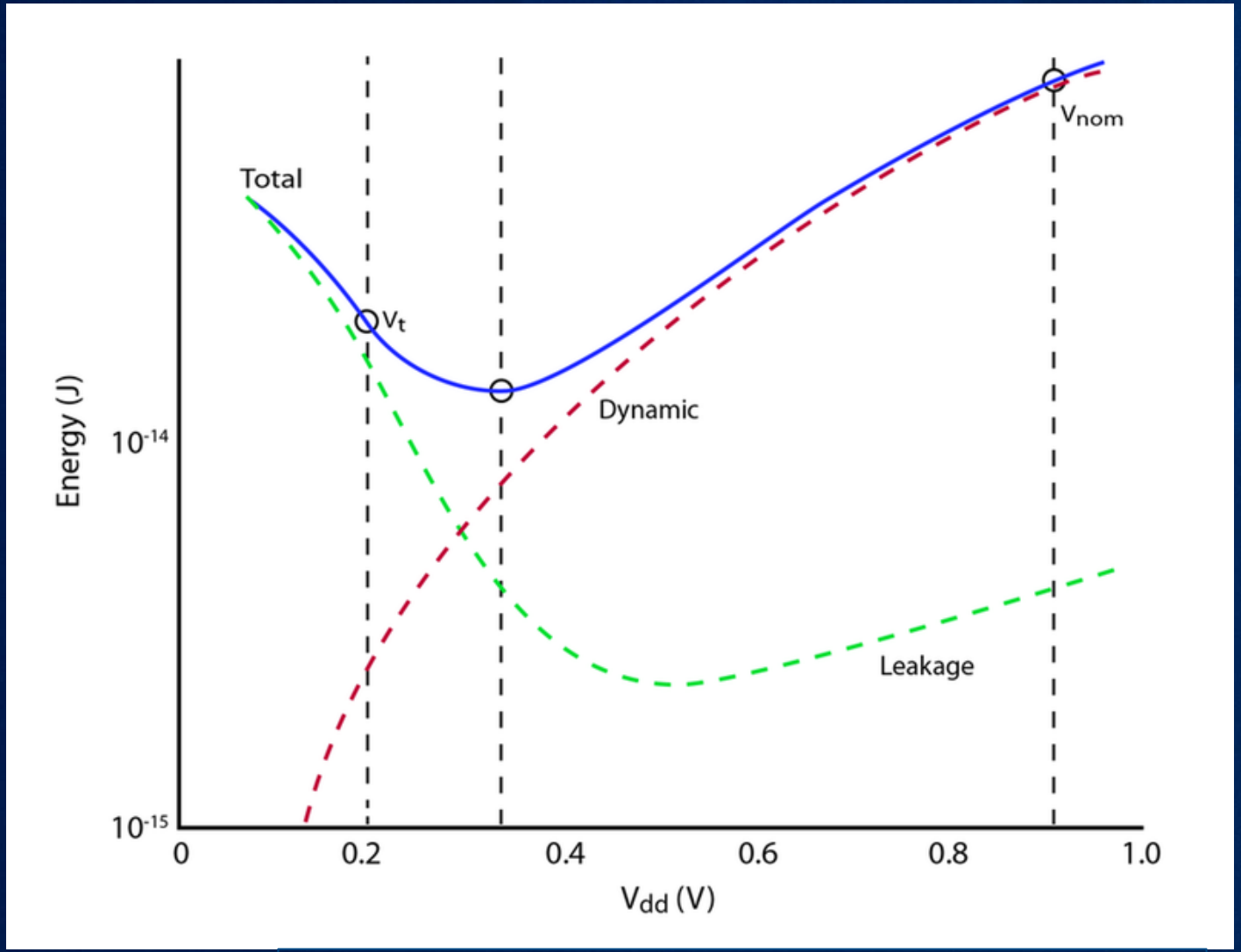
MAPT

DOE EES2

HPC 4 ENERGY INNOVATION

CMOS Operation Close to Threshold… not a solution!

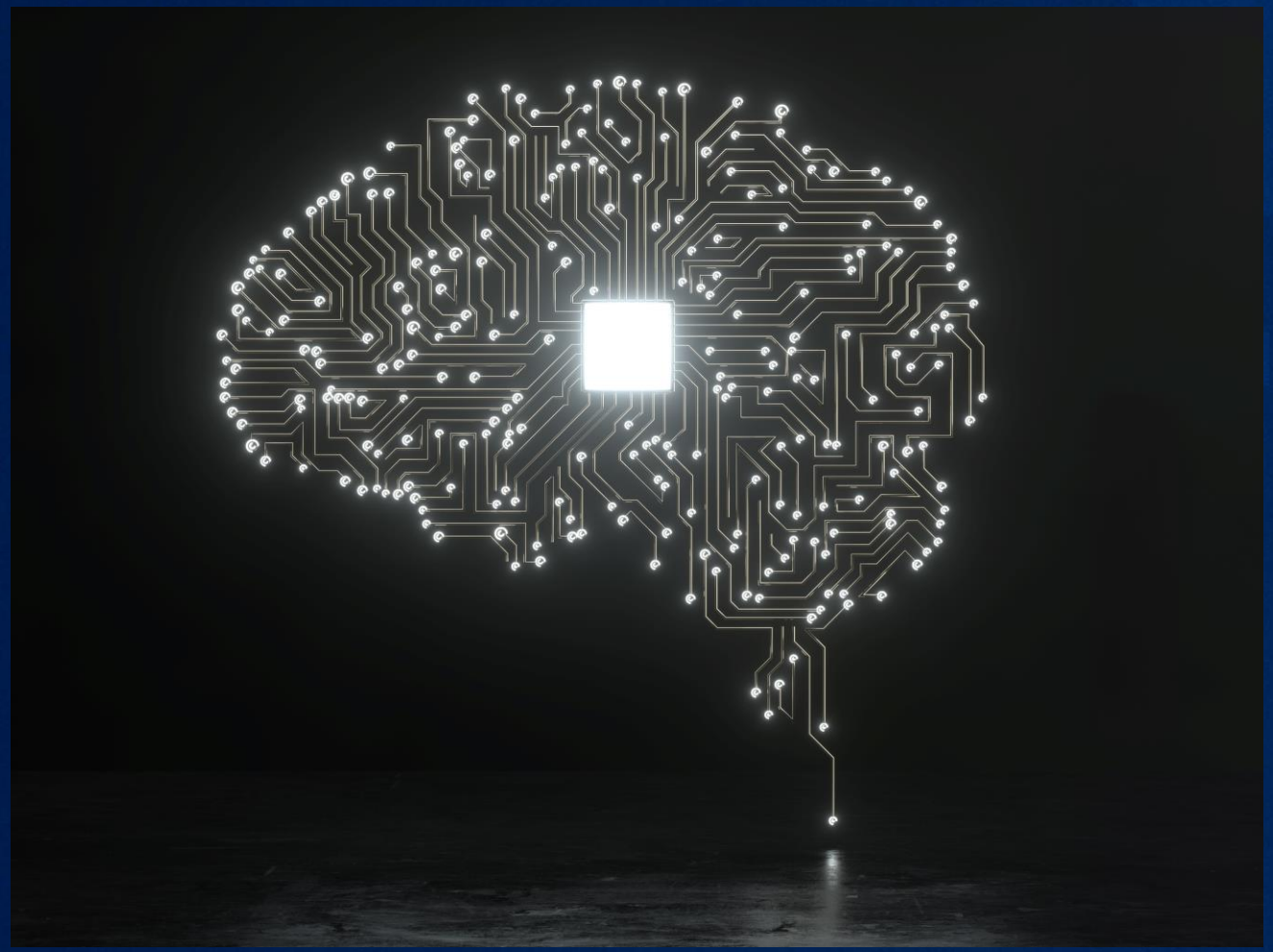
- Theoretically, you can optimize efficiency, but…
- Minimum core voltage limited by circuit functional and timing failures from process variation and noise



https://www.techdesignforums.com/practice/files/2014/05/variability.png

Our brain: massively parallel at low power

- Neural Network based architecture
- Beyond CMOS transistors



Questions?


Apollo 13 – Universal Pictures
Efficiency
"Power is Everything"*
John Aaron- Apollo 13 Flight Controller