A Roadmap for Heterogeneous Integration for the Next Decade and Beyond
An Academic Perspective

Madhavan Swaminathan
Dept. Head Electrical Engineering
William E. Leonhard Endowed Chair
Director, CHIMES (an SRC JUMP 2.0 Center)
The Pennsylvania State University
Emeritus Professor, ECE & MSE, Georgia Tech
Former Director, 3D Systems Packaging Research Center (NSF-ERC), Georgia Tech

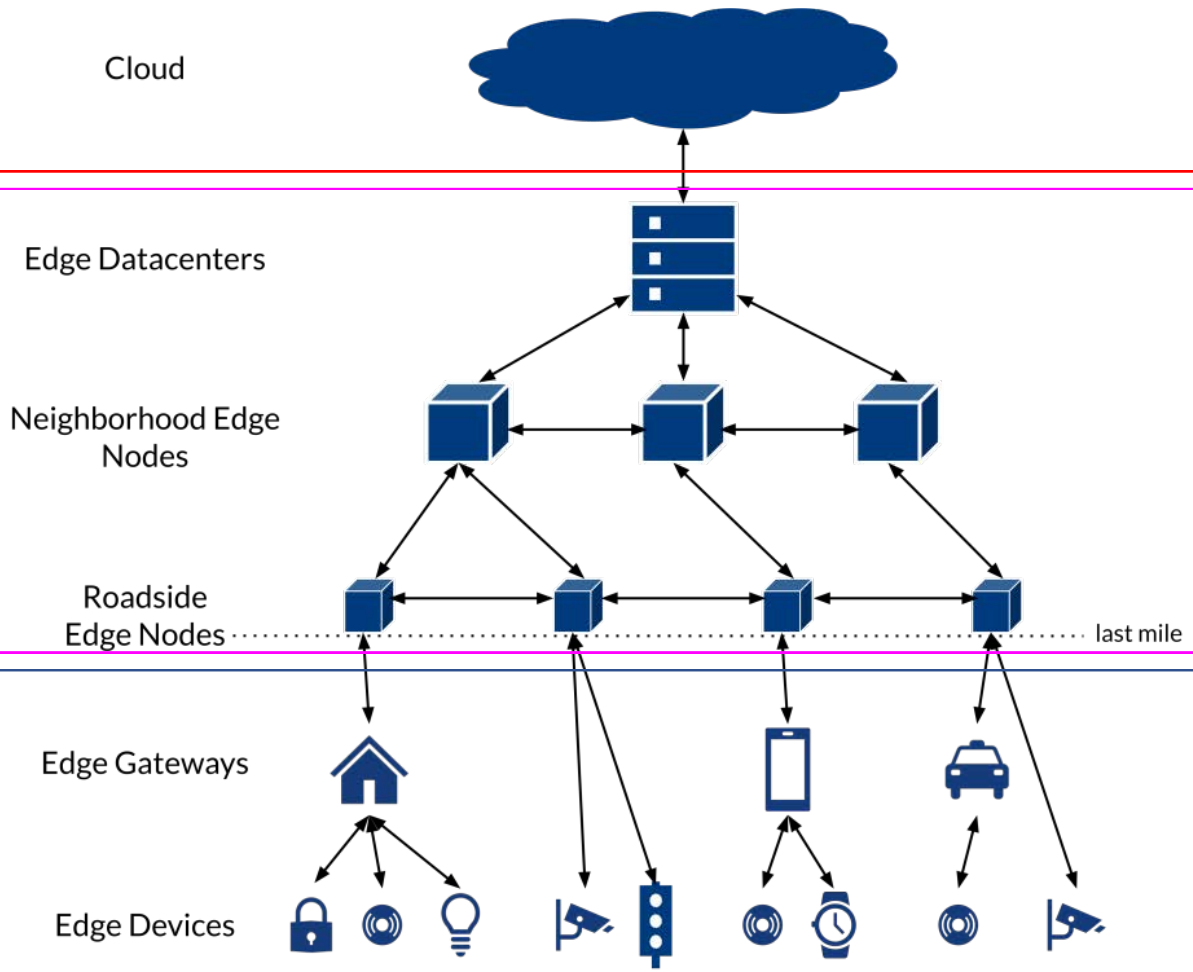
Outline

- ❑ Emerging Applications
- ❑ Need for Heterogeneous Integration
- ❑ A 10 Year and Beyond Roadmap
- ❑ CHIMES – A Snap Shot

Emerging Applications

- ❑ Cognition (AI)
- ❑ Communication (C)
- ❑ Sensing (S)
- ❑ Distributed Computing (DC)
- ❑ Intelligent Memory (M)
- ❑ Harsh Environments (H)

Emerging Distributed & Edge Computing



- Storage, Coordinator of lower levels
- Limitless Compute Power
- ~8000
- Power: 10s of KWs - 100s of KW

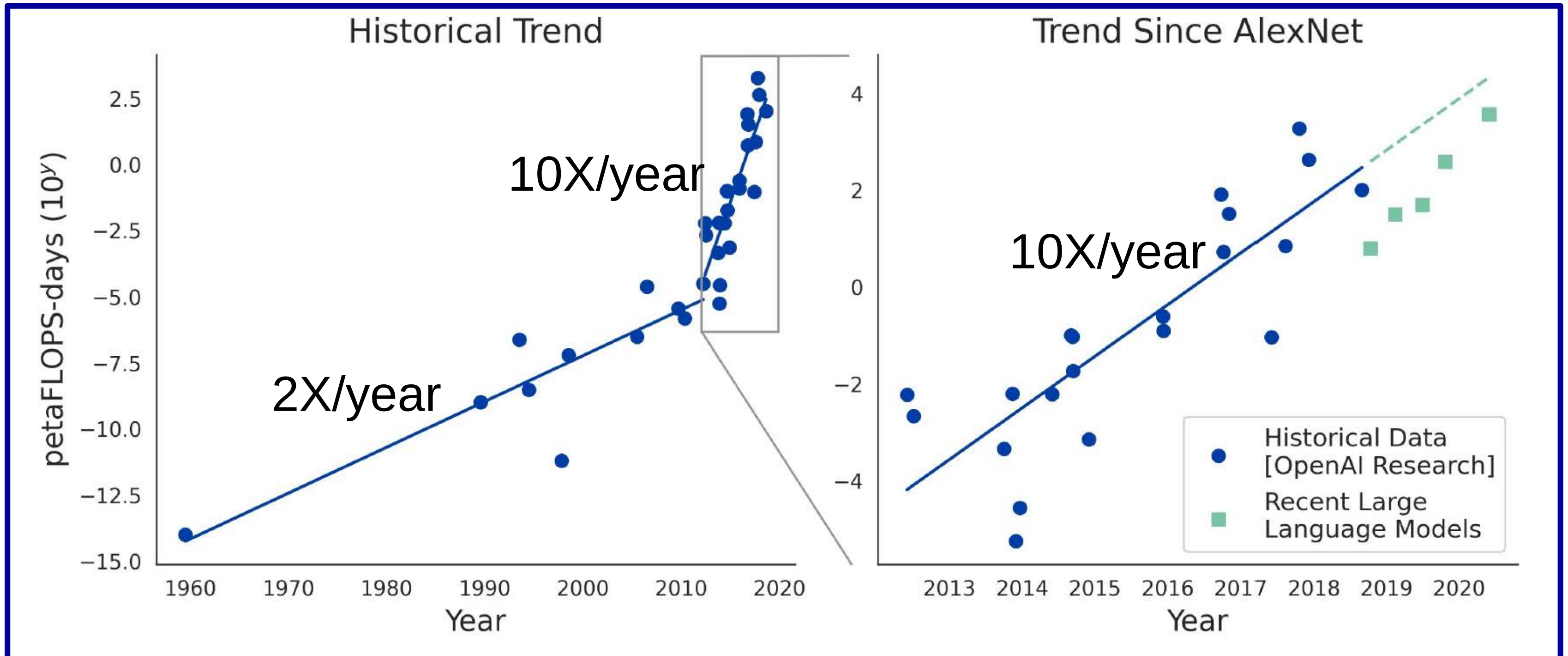
- Base stations, Routers, Switches
- High Compute Power
- 10-100 million devices
- Power: 10s of W – few KWs

- Sensors, Smart Phones, Automobiles
- Billions of devices
- Power: 10s of mW – 10s of W

Ref: Jeff Burns, "Systems and Architectures for Distributed Compute", SRC Workshop, 2022.

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

Deep Learning & Computational Power (Large Data Centers)

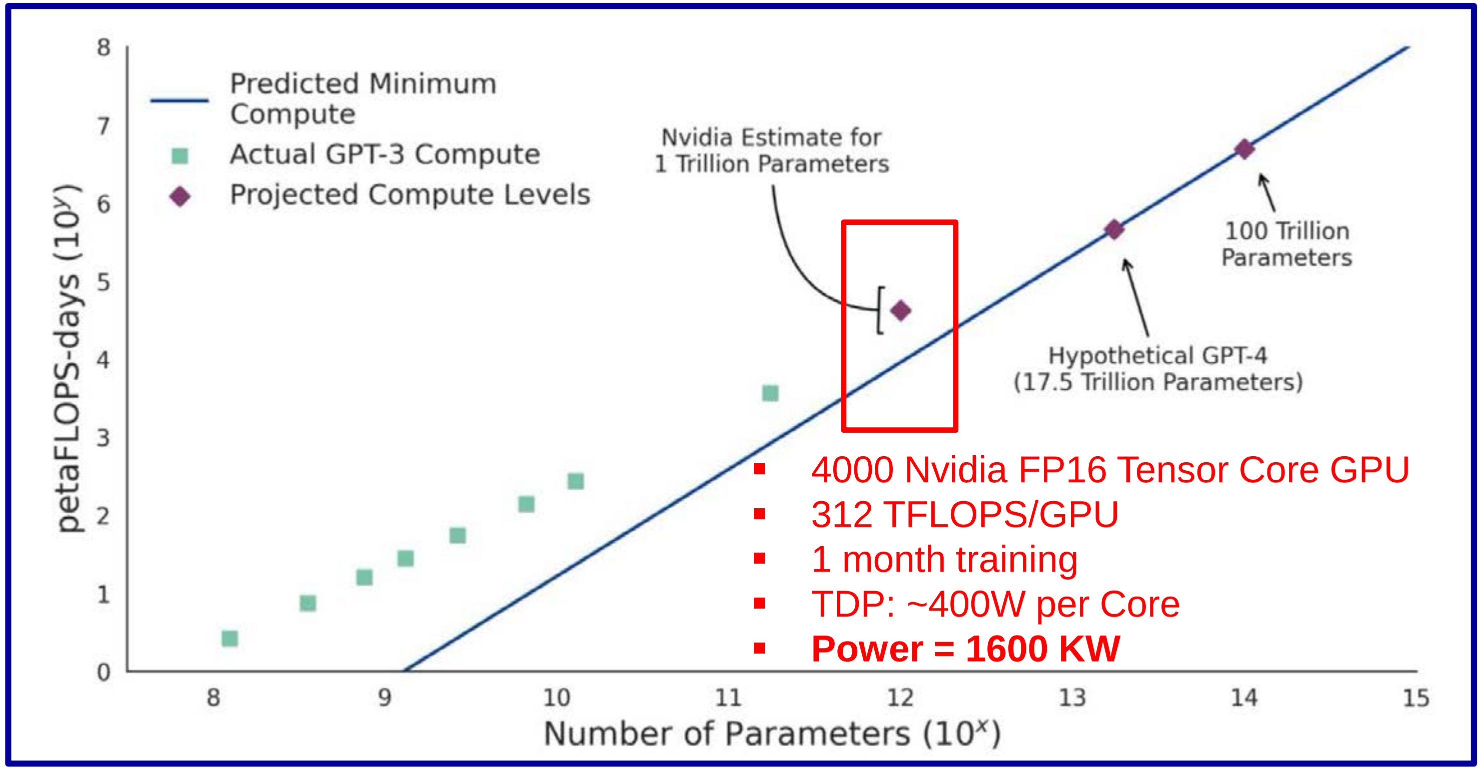


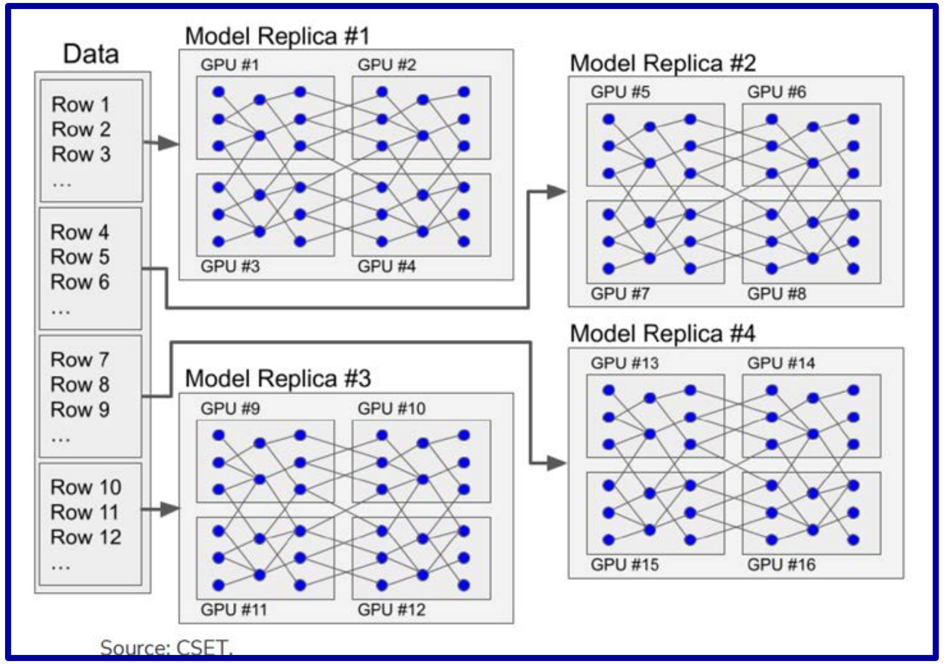
Source: OpenAI and CSET.

- Observations
  - Prior to 2012: Compute demand doubling every 2 years at the same rate as Moore
  - Since 2012: Compute demand growing 10X/year due to Deep Learning

Ref: Andrew John and Micah Musser, "AI and Compute – How much longer can computing power drive artificial intelligence progress", CSET, 2022

Compute Power & Scaling

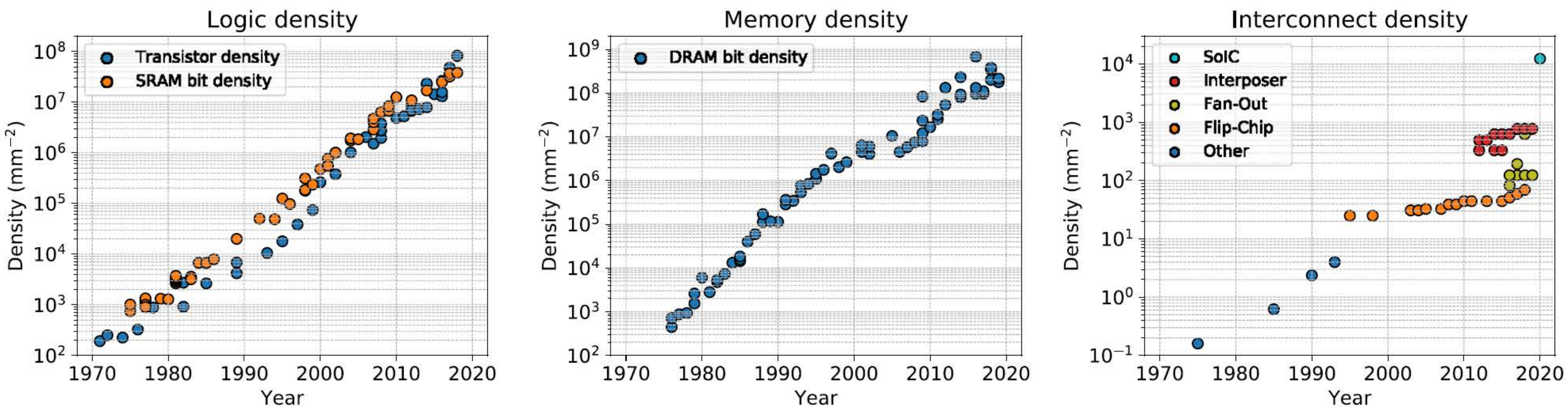




- 2X Process Gains (8nm)
- 3X Architecture Gains (Tensor Flow)
- 300X System Scale out (Data Centers) – OUCH!
- Energy crisis if we continue on this path …..

Ref: Andrew John and Micah Musser, "AI and Compute – How much longer can computing power drive artificial intelligence progress", CSET, 2022
Anna Herr and Quentin Herr, Superconducting AI & HPC, IMEC

Historical Trends & Future (Estimated) Needs



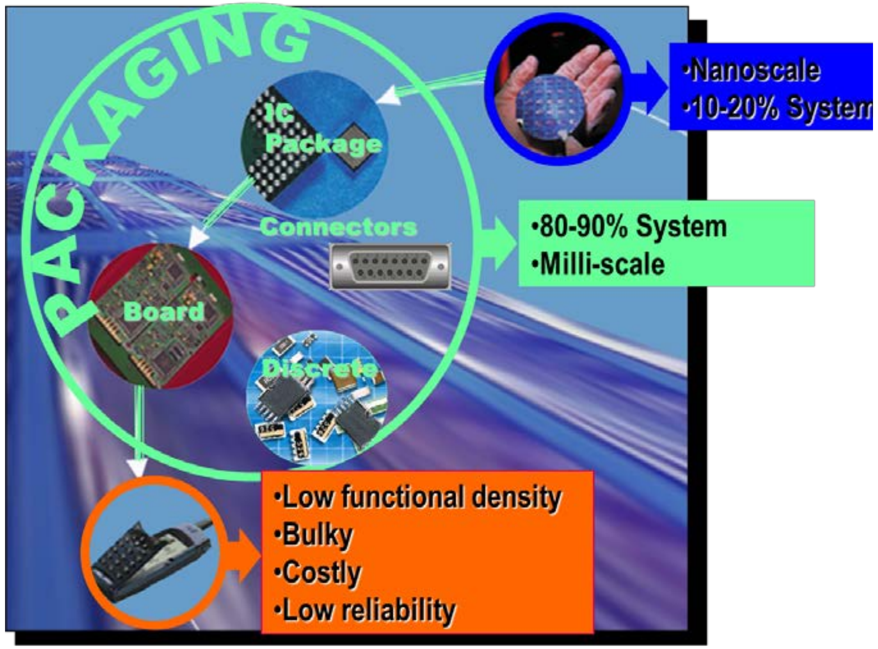
H.-S. Philip Wong, et al, “A Density Metric for Semiconductor Technology”, Proceedings of the IEEE, April 2020

Current State of the Art

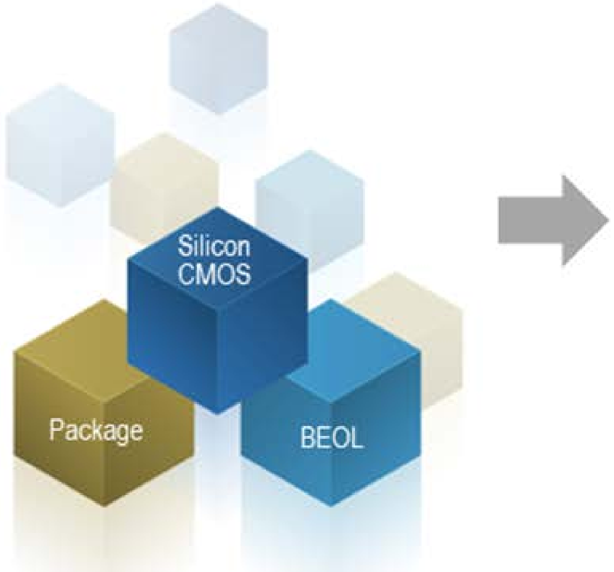
- Monolithic logic 10^8 transistors/mm^2
- DRAM 10^9 transistors/mm^2
- IO density 10^4 IO/mm^2
- SRAM Access 20-50 TBps

10-100X increase in transistor densities
Interconnect densities 10^6 and higher (100X)
Energy per bit (EPB) reduced to femto-joules/bit
500TBps/mm^2 of bandwidth (10X increase)
Wireless communication at 1Tbps

Is Heterogeneous Integration the Answer?



Packaging (Past)



Advanced Packaging (Present)

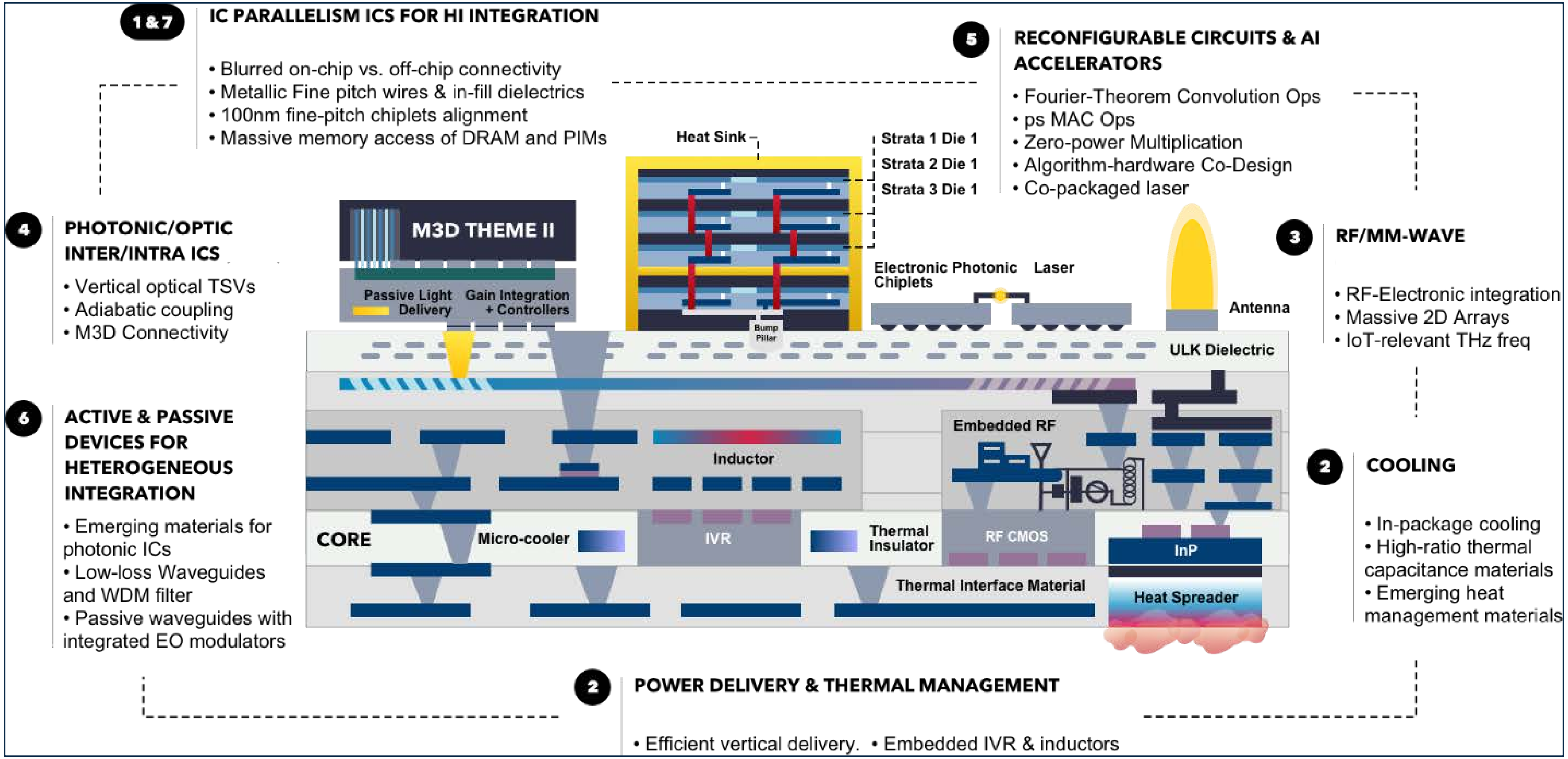
Heterogeneous Integration (Future)

- **Past/Present:** FEOL Transistor, BEOL Wiring & Package individually developed and combined
- **Future:** New and transformative logic, memory, and interconnect technologies that overcome the inevitable slowdown of traditional dimensional scaling of CMOS by interconnecting a **diversity of transistors and integrated circuit components**, **blurring the line between what is on-chip and what is off-chip**.

Grand Challenges

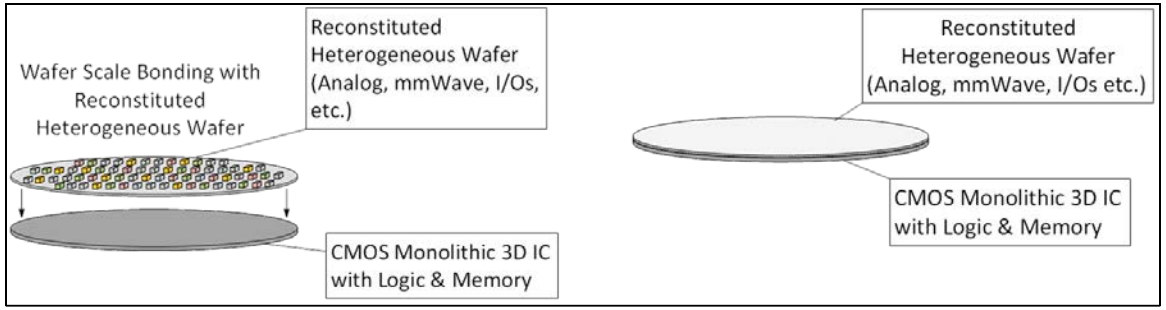
1. Heterogeneous integrated systems with volumetric transistor and interconnect densities approaching monolithic integration through 250-nm pitch self-aligned chip I/Os, and massive and seamless photonic connectivity with bandwidth density of 100Tbps/mm.
2. A generalized framework for materials synthesis and implementation with optimized properties.
3. Power management with segmented and deeply integrated power delivery solutions at 50kW with >80% efficiency.
4. Ultra-compact multi-scale thermal management solutions for densely integrated electronic, photonic, and sub-Terahertz components with micron-scale heat spreaders for >1kW/cm^2 heat flux, and superior energy efficiency and reliability.
5. Reconfigurable photonic interconnects enabling adaptive & optimized system configurations achieving > 100X enhancement in energy-efficiency and throughput compared to SOTA.
6. Fully integrated electronics, photonics and thermal design automation platform.

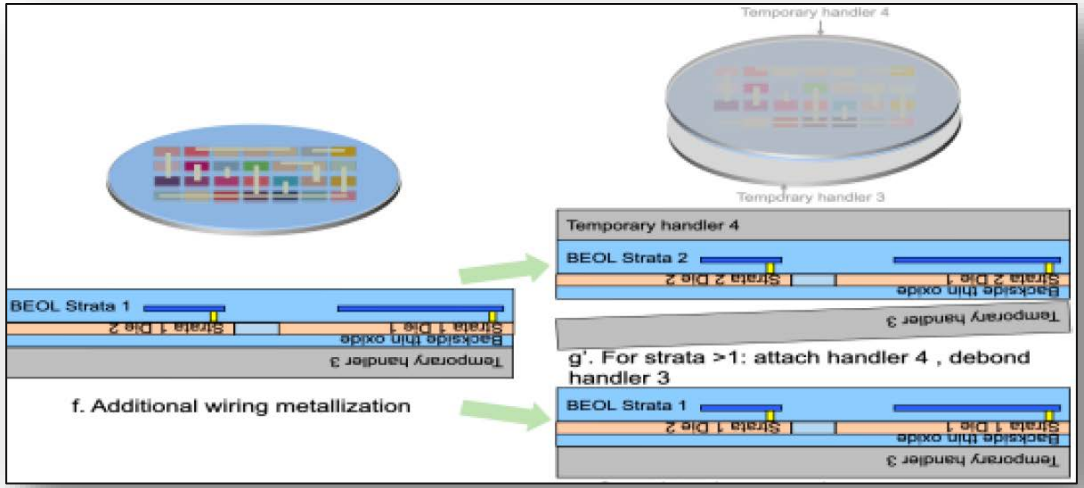
Ultra-dense Heterogeneous Platform – A Futuristic Outlook



The Roadmap - Volume Scaling

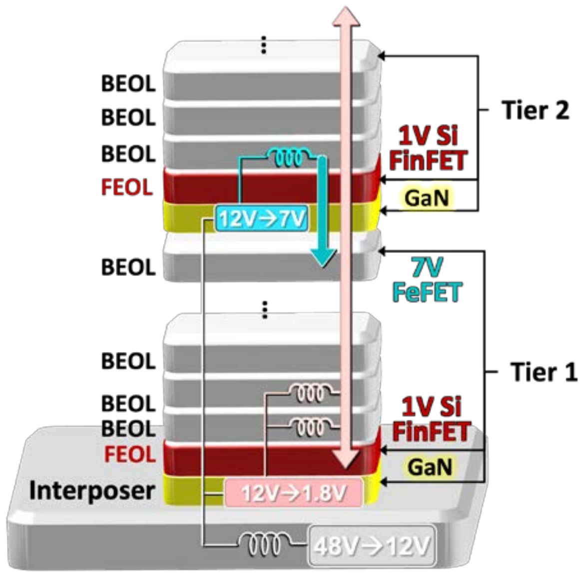
Metric & Drivers	Metric (10 Years)	State of the Art (SOTA)
Electrical IC Parallelism & Volume Scaling (AI, DC, M)	IO density 16M/mm^2	IO density 10K/mm^2
	BW Density > 500Tbps/mm^2; <0.01 pJ/bit	BW density 28Tbps/mm^2; 0.021 pJ/bit (ARM/GT IEDM 2020)
Cooling & Power Delivery (S, C, H, AI)	Power 1000W; Current density 2–5A/mm^2; Efficiency >80% (HIR 10 year)	Power 400W; Current Density 1A/mm2; Efficiency <70%
	Current 50 KAmps; Power 50kW; PDN Power <1KW (2%); Efficiency>80%	Current 18 KAmps; Power 10KW; PDN 5KW (33%); η <67% (Tesla Dojo)
Wireless IC & I/O (C, S, H)	Insertion Loss <1dB (PA output to Antenna input) @ 0.3 – 1THz; DARPA ELGAR 1dB	Insertion Loss >5dB in D-Band (110-170 GHz) (ComSenTer)
	Antenna efficiency >90% @ 0.3–1THz	Antenna efficiency >90% (< 100GHz)
	SNR 30dB (128 QAM)	SNR 20dB (128 QAM) D-Band
Photonic/Optic IC and Chiplet Communication (AI, C, DC)	Coupling loss < 0.9dB; Misalignment; +/−1mm	Coupling Loss ~ 1dB (O-Band)
	100 (Tbps/mm)/(pJ/bit) (link-level)	1 (Tbps/mm)/(pJ/bit) (DARPA PIPES)
	Bandwidth (BW) Reconfigurable ICs between chiplets; 0.5-5Tbps aggregated BW	NA
	Optical gain/laser integration via photonic wire bonds (PWB) and/or mono-int. (DARPA LUMOS)	Laser is off-chip → requiring packaging (costly & not reliable)
Reconfigurable Circuits & Modules (AI, S, DC)	100 TOP/J, 1ns latency, infinite bit-resolution MAC ops (DOD Labs e.g., AFRL, ARL)	<5 TOP/J, 1000+ns latency, 4-12 bit
	AI accelerator: 10–100 fJ/MAC, 1–50 TMACs/mm^2, 1ns–25 ps operation (Google X)	0.5–1 pJ/MAC, 0.5–1 TMAC/s/mm^2, 0.5–1 GMVM/s, and 1–2 us per MVM
Active & Passive Devices for Heterogeneous Integration (C, S, DC)	EO Modulator: ER > 12dB; BW > 15GHz per channel; (next Gen AIM Photonics components)	NA
	Interposer Waveguides; losses < 0.1dB/cm	1.5dB/cm
	Interposer Filters for WDM: filter sharpness > 15dB over < 0.2nm BW (AIM Photonics Foundry)	NA

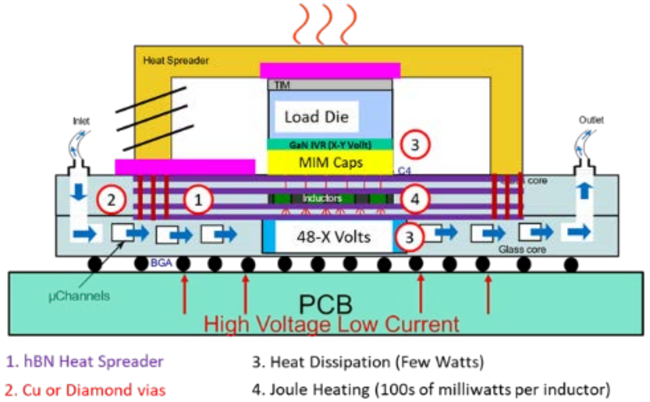


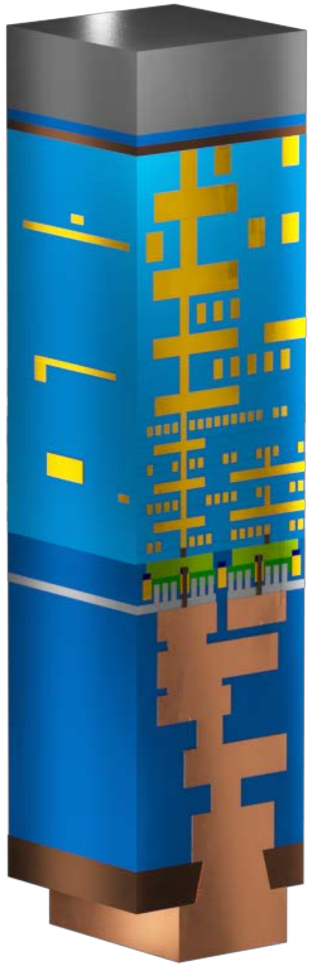


The Roadmap - Addressing the Current Problem

Metric & Drivers	Metric (10 Years)	State of the Art (SOTA)
Electrical IC Parallelism & Volume Scaling (AI, DC, M)	IO density 16M/mm^2	IO density 10K/mm^2
	BW Density > 500Tbps/mm^2; <0.01 pJ/bit	BW density 28Tbps/mm^2; 0.021 pJ/bit (ARM/GT IEDM 2020)
Cooling & Power Delivery (S, C, H, AI)	Power 1000W; Current density 2–5A/mm^2; Efficiency >80% (HIR 10 year)	Power 400W; Current Density 1A/mm2; Efficiency <70%
	Current 50 KAmps; Power 50kW; PDN Power <1KW (2%); Efficiency>80%	Current 18 KAmps; Power 10KW; PDN 5KW (33%); η <67% (Tesla Dojo)
Wireless IC & I/O (C, S, H)	Insertion Loss <1dB (PA output to Antenna input) @ 0.3 – 1THz; DARPA ELGAR 1dB	Insertion Loss >5dB in D-Band (110-170 GHz) (ComSenTer)
	Antenna efficiency >90% @ 0.3–1THz	Antenna efficiency >90% (< 100GHz)
	SNR 30dB (128 QAM)	SNR 20dB (128 QAM) D-Band
Photonic/Optic IC and Chiplet Communication (AI, C, DC)	Coupling loss < 0.9dB; Misalignment; +/−1mm	Coupling Loss ~ 1dB (O-Band)
	100 (Tbps/mm)/(pJ/bit) (link-level)	1 (Tbps/mm)/(pJ/bit) (DARPA PIPES)
	Bandwidth (BW) Reconfigurable ICs between chiplets; 0.5-5Tbps aggregated BW	NA
	Optical gain/laser integration via photonic wire bonds (PWB) and/or mono-int. (DARPA LUMOS)	Laser is off-chip → requiring packaging (costly & not reliable)
Reconfigurable Circuits & Modules (AI, S, DC)	100 TOP/J, 1ns latency, infinite bit-resolution MAC ops (DOD Labs e.g., AFRL, ARL)	<5 TOP/J, 1000+ns latency, 4-12 bit
	AI accelerator: 10–100 fJ/MAC, 1–50 TMACs/mm^2, 1ns–25 ps operation (Google X)	0.5–1 pJ/MAC, 0.5–1 TMAC/s/mm^2, 0.5–1 GMVM/s, and 1–2 us per MVM
Active & Passive Devices for Heterogeneous Integration (C, S, DC)	EO Modulator: ER > 12dB; BW > 15GHz per channel; (next Gen AIM Photonics components)	NA
	Interposer Waveguides; losses < 0.1dB/cm	1.5dB/cm
	Interposer Filters for WDM: filter sharpness > 15dB over < 0.2nm BW (AIM Photonics Foundry)	NA



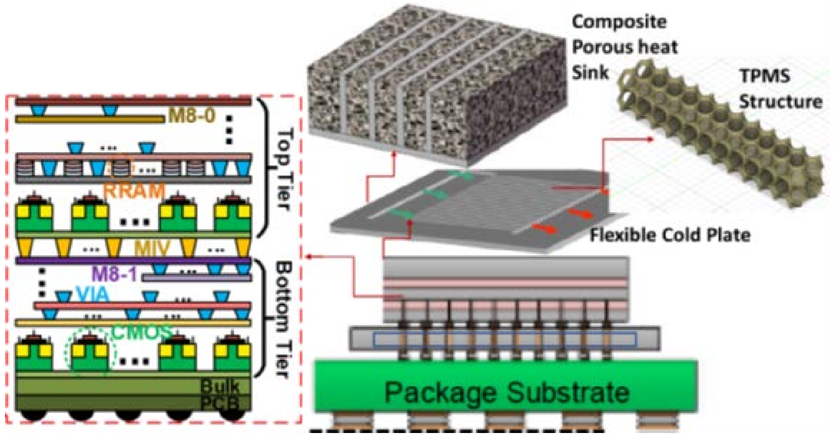


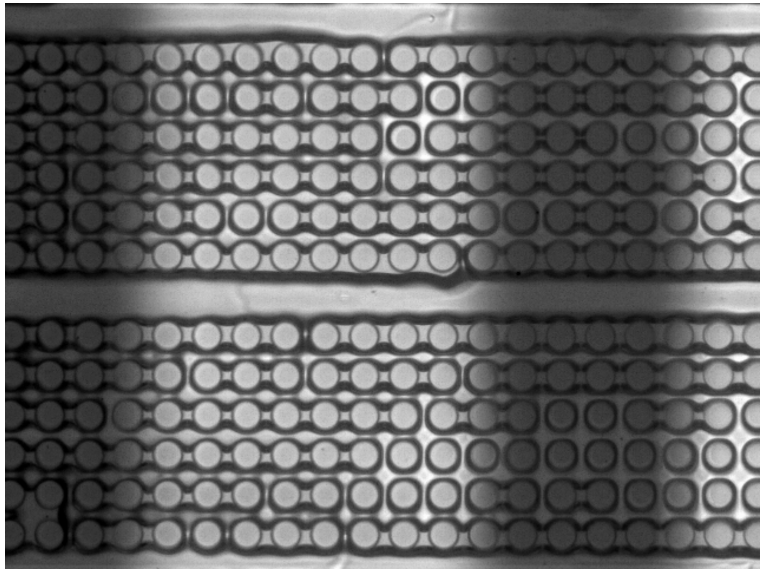


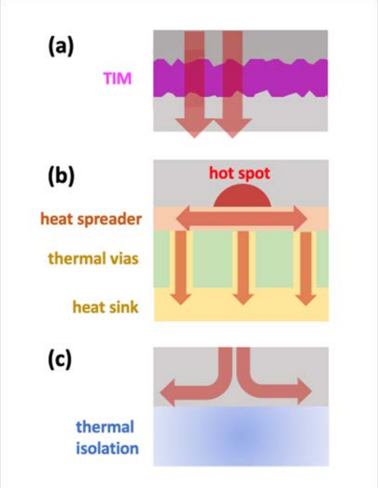
Courtesy: K. Radhakrishnan, Intel

The Roadmap - Staying Cool

Metrics & Drivers	Metrics (10 Years)	State of the Art (SOTA)
Benchmarking (AI, C, S, DC, M, H)	Automated cross-layer pathfinding framework with Power, Performance, Form-factor, Cost and Reliability (PPFCR) prediction for full systems + applications.	Pathfinding limited to technology or technology + circuit only; limited automation.
Co-Design (AI, C, S, DC, M, H)	Physical design tools for M3D chiplets integrated on systems spanning > 20,000mm^2	No tools for M3D; interposer tools limited to ~2000mm^2
Power Delivery (AI, DC, H)	Power 1000W; Current density 2-5A/mm^2; Efficiency >80% (HIR 10 year)	Power 400W; Current density 1A/mm^2; Efficiency <70%
	Current 50 kAmps; Power 50kW; PDN Power <1kW (2%); Efficiency>80%	Current 18 kAmps; Power 10kW; PDN 5kW (33%); Efficiency <67% (Tesla Dojo)
Thermal (AI, C, DC, H)	Chiplet: Background heat flux (entire die) 2kW/cm^2; hot spot heat flux (0.1mm x 0.1mm) 30kW/cm^2; Vol. removal 2kW/cm^2 mm; Thermal time constant 1μs	Background heat flux (entire die) 200W/cm^2; hot spot heat flux (1mm x 1mm) 1kW/cm^2; Vol. removal 200W/cm^2 mm; Thermal time constant 20μs
	Stack: Thermal isolation ratio 0.95; TIM specific resistance 0.3 mm^2 K/W for 30 μm; k 200 W/mK, elastic modulus 30 MPa	Thermal isolation ratio 0.5; TIM specific resistance 1 mm^2 K/W (DARPA NTI)
	Interposer: Heat spreader effective k 10,000 W/mK; Heat spreader thickness 0.2mm	Heat spr. effective k 4000 W/mK; Heat spreader thickness 2mm (DARPA TGP)
	Server (1U): Volumetric R_{th} 0.015 ^oCcm^3/W	Volumetric R_{th} 0.03 ^oCcm^3/W
Electrical/Optical Test (AI, DC)	Fault coverage >99% (Short/Open) & >95% (device/components)	<95% coverage (Short/Open), devices/components (no test or BIST)
	10X reduction in test cost (Package BIST)	(Relative) test cost growing
Hardware Security	Probability of trojan evasion >99%	No universal security metrics available
	Probability of reverse engineering <1%	No universal security metrics available
	Split manufacturing based on fine pitch HI	Not available



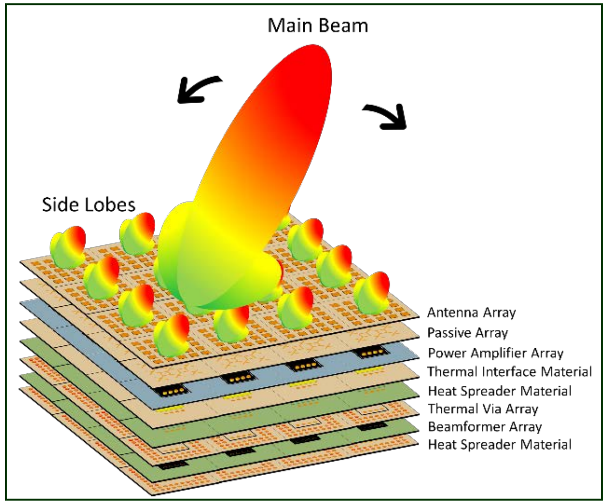


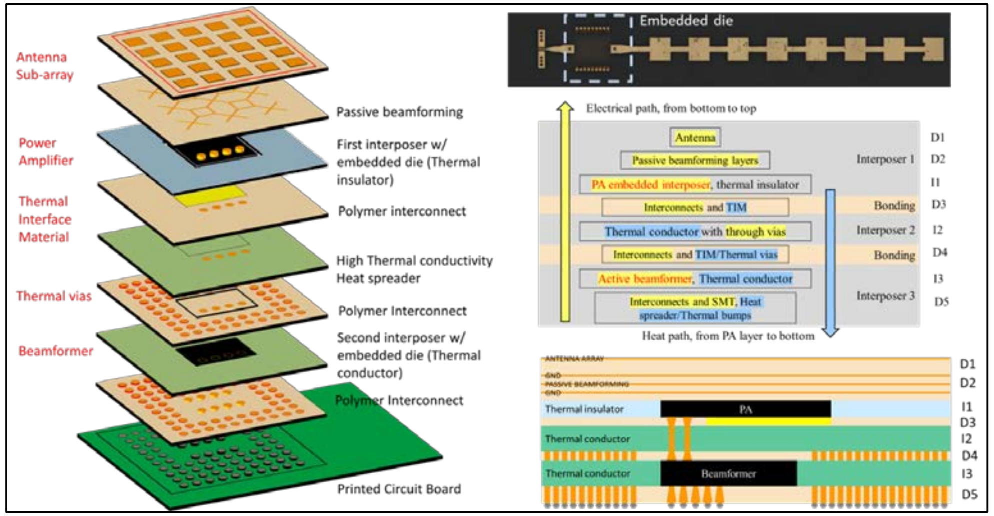


Desired heat flow through critical junctions enabled by proposed thermal materials

The Roadmap - Ensuring Connectivity the Wireless Way

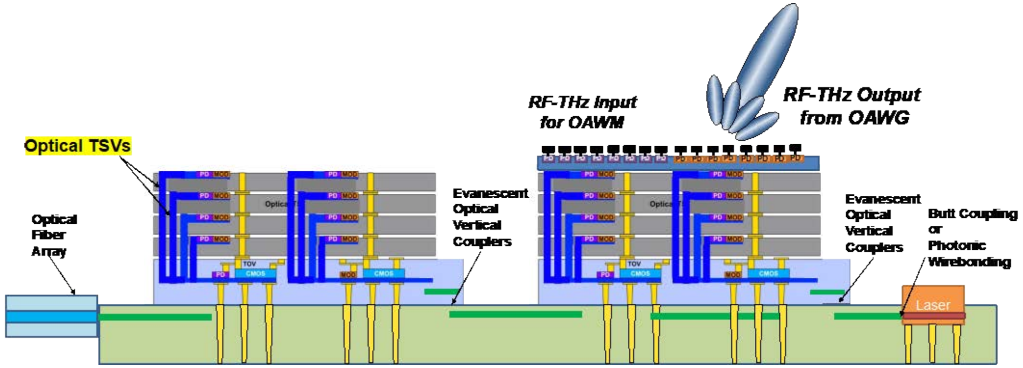
Metric & Drivers	Metric (10 Years)	State of the Art (SOTA)
Electrical IC Parallelism & Volume Scaling (AI, DC, M)	IO density 16M/mm^2	IO density 10K/mm^2
	BW Density > 500Tbps/mm^2; <0.01 pJ/bit	BW density 28Tbps/mm^2; 0.021 pJ/bit (ARM/GT IEDM 2020)
Cooling & Power Delivery (S, C, H, AI)	Power 1000W; Current density 2–5A/mm^2; Efficiency >80% (HIR 10 year)	Power 400W; Current Density 1A/mm2; Efficiency <70%
	Current 50 KAmps; Power 50kW; PDN Power <1KW (2%); Efficiency>80%	Current 18 KAmps; Power 10KW; PDN 5KW (33%); η <67% (Tesla Dojo)
Wireless IC & I/O (C, S, H)	Insertion Loss <1dB (PA output to Antenna input) @ 0.3 – 1THz; DARPA ELGAR 1dB	Insertion Loss >5dB in D-Band (110-170 GHz) (ComSenTer)
	Antenna efficiency >90% @ 0.3–1THz	Antenna efficiency >90% (< 100GHz)
	SNR 30dB (128 QAM)	SNR 20dB (128 QAM) D-Band
Photonic/Optic IC and Chiplet Communication (AI, C, DC)	Coupling loss < 0.9dB; Misalignment; +/−1mm	Coupling Loss ~ 1dB (O-Band)
	100 (Tbps/mm)/(pJ/bit) (link-level)	1 (Tbps/mm)/(pJ/bit) (DARPA PIPES)
	Bandwidth (BW) Reconfigurable ICs between chiplets; 0.5-5Tbps aggregated BW	NA
	Optical gain/laser integration via photonic wire bonds (PWB) and/or mono-int. (DARPA LUMOS)	Laser is off-chip → requiring packaging (costly & not reliable)
Reconfigurable Circuits & Modules (AI, S, DC)	100 TOP/J, 1ns latency, infinite bit-resolution MAC ops (DOD Labs e.g., AFRL, ARL)	<5 TOP/J, 1000+ns latency, 4-12 bit
	AI accelerator: 10–100 fJ/MAC, 1–50 TMACs/mm^2, 1ns–25 ps operation (Google X)	0.5–1 pJ/MAC, 0.5–1 TMAC/s/mm^2, 0.5–1 GMVM/s, and 1–2 us per MVM
Active & Passive Devices for Heterogeneous Integration (C, S, DC)	EO Modulator: ER > 12dB; BW > 15GHz per channel; (next Gen AIM Photonics components)	NA
	Interposer Waveguides; losses < 0.1dB/cm	1.5dB/cm
	Interposer Filters for WDM: filter sharpness > 15dB over < 0.2nm BW (AIM Photonics Foundry)	NA

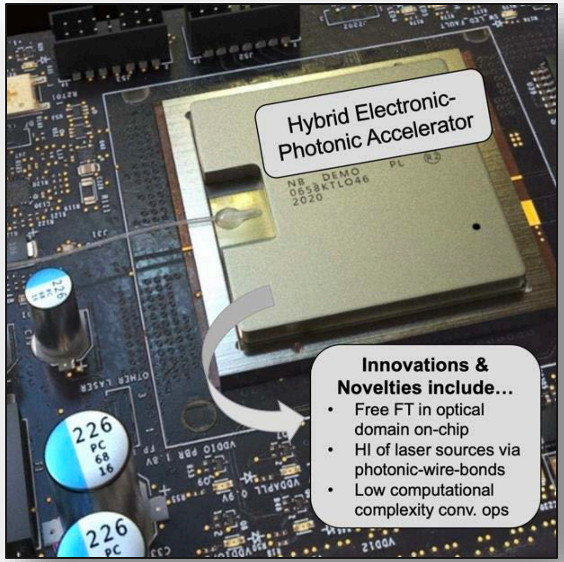


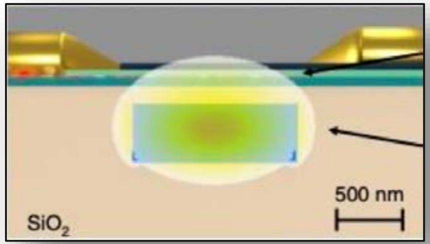


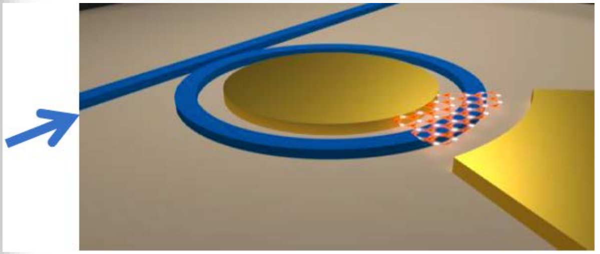
The Roadmap – Lighting it up

Metric & Drivers	Metric (10 Years)	State of the Art (SOTA)
Electrical IC Parallelism & Volume Scaling (AI, DC, M)	IO density 16M/mm^2	IO density 10K/mm^2
	BW Density > 500Tbps/mm^2; <0.01 pJ/bit	BW density 28Tbps/mm^2; 0.021 pJ/bit (ARM/GT IEDM 2020)
Cooling & Power Delivery (S, C, H, AI)	Power 1000W; Current density 2–5A/mm^2; Efficiency >80% (HIR 10 year)	Power 400W; Current Density 1A/mm2; Efficiency <70%
	Current 50 KAmps; Power 50kW; PDN Power <1KW (2%); Efficiency>80%	Current 18 KAmps; Power 10KW; PDN 5KW (33%); η <67% (Tesla Dojo)
Wireless IC & I/O (C, S, H)	Insertion Loss <1dB (PA output to Antenna input) @ 0.3 – 1THz; DARPA ELGAR 1dB	Insertion Loss >5dB in D-Band (110-170 GHz) (ComSenTer)
	Antenna efficiency >90% @ 0.3–1THz	Antenna efficiency >90% (< 100GHz)
	SNR 30dB (128 QAM)	SNR 20dB (128 QAM) D-Band
Photonic/Optic IC and Chiplet Communication (AI, C, DC)	Coupling loss < 0.9dB; Misalignment; +/−1mm	Coupling Loss ~ 1dB (O-Band)
	100 (Tbps/mm)/(pJ/bit) (link-level)	1 (Tbps/mm)/(pJ/bit) (DARPA PIPES)
	Bandwidth (BW) Reconfigurable ICs between chiplets; 0.5-5Tbps aggregated BW	NA
	Optical gain/laser integration via photonic wire bonds (PWB) and/or mono-int. (DARPA LUMOS)	Laser is off-chip → requiring packaging (costly & not reliable)
Reconfigurable Circuits & Modules (AI, S, DC)	100 TOP/J, 1ns latency, infinite bit-resolution MAC ops (DOD Labs e.g., AFRL, ARL)	<5 TOP/J, 1000+ns latency, 4-12 bit
	AI accelerator: 10–100 fJ/MAC, 1–50 TMACs/mm^2, 1ns–25 ps operation (Google X)	0.5–1 pJ/MAC, 0.5–1 TMAC/s/mm^2, 0.5–1 GMVM/s, and 1–2 us per MVM
Active & Passive Devices for Heterogeneous Integration (C, S, DC)	EO Modulator: ER > 12dB; BW > 15GHz per channel; (next Gen AIM Photonics components)	NA
	Interposer Waveguides; losses < 0.1dB/cm	1.5dB/cm
	Interposer Filters for WDM: filter sharpness > 15dB over < 0.2nm BW (AIM Photonics Foundry)	NA



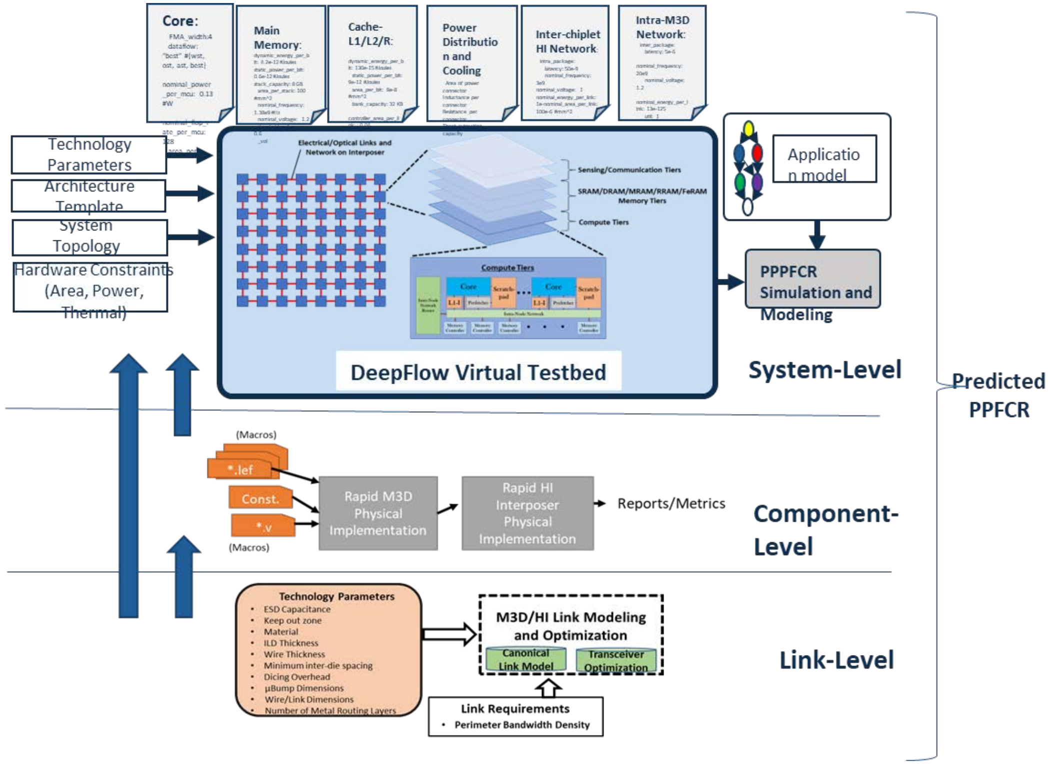


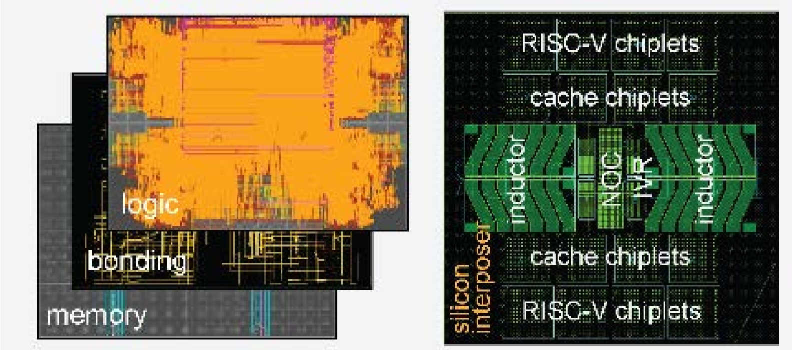




The Roadmap – Democratizing Heterogeneous Integration

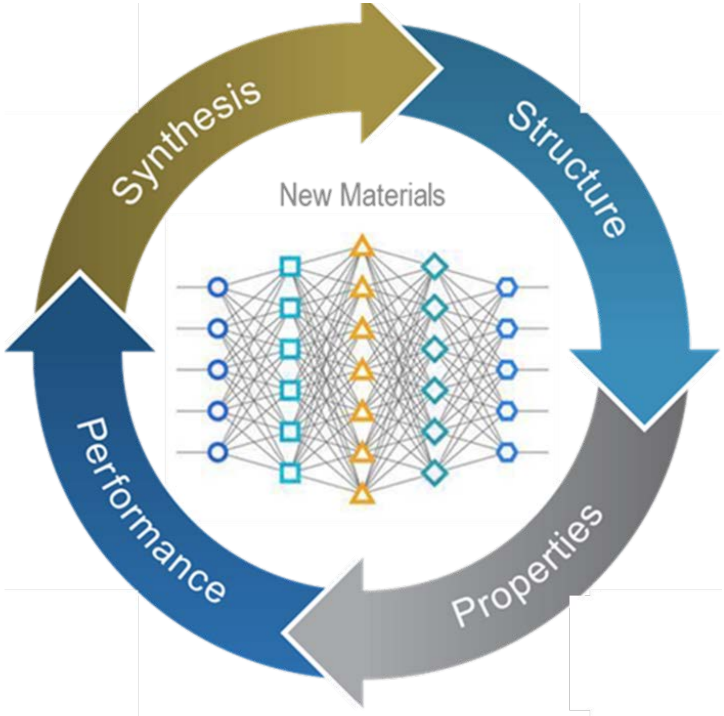
Metrics & Drivers	Metrics (10 Years)	State of the Art (SOTA)
Benchmarking (AI, C, S, DC, M, H)	Automated cross-layer pathfinding framework with Power, Performance, Form-factor, Cost and Reliability (PPFCR) prediction for full systems + applications.	Pathfinding limited to technology or technology + circuit only; limited automation.
Co-Design (AI, C, S, DC, M, H)	Physical design tools for M3D chiplets integrated on systems spanning > 20,000mm^2	No tools for M3D; interposer tools limited to ~2000mm^2
Power Delivery (AI, DC, H)	Power 1000W; Current density 2-5A/mm^2; Efficiency >80% (HIR 10 year)	Power 400W; Current density 1A/mm^2; Efficiency <70%
	Current 50 kAmps; Power 50kW; PDN Power <1kW (2%); Efficiency>80%	Current 18 kAmps; Power 10kW; PDN 5kW (33%); Efficiency <67% (Tesla Dojo)
Thermal (AI, C, DC, H)	Chiplet: Background heat flux (entire die) 2kW/cm^2; hot spot heat flux (0.1mm x 0.1mm) 30kW/cm^2; Vol. removal 2kW/cm^2 mm; Thermal time constant 1μs	Background heat flux (entire die) 200W/cm^2; hot spot heat flux (1mm x 1mm) 1kW/cm^2; Vol. removal 200W/cm^2 mm; Thermal time constant 20μs
	Stack: Thermal isolation ratio 0.95; TIM specific resistance 0.3 mm^2 K/W for 30 μm; k 200 W/mK, elastic modulus 30 MPa	Thermal isolation ratio 0.5; TIM specific resistance 1 mm^2 K/W (DARPA NTI)
	Interposer: Heat spreader effective k 10,000 W/mK; Heat spreader thickness 0.2mm	Heat spr. effective k 4000 W/mK; Heat spreader thickness 2mm (DARPA TGP)
	Server (1U): Volumetric R_{th} 0.015 $^oCcm^3/W$	Volumetric R_{th} 0.03 $^oCcm^3/W$
Electrical/Optical Test (AI, DC)	Fault coverage >99% (Short/Open) & >95% (device/components)	<95% coverage (Short/Open), devices/components (no test or BIST)
	10X reduction in test cost (Package BIST)	(Relative) test cost growing
Hardware Security	Probability of trojan evasion >99%	No universal security metrics available
	Probability of reverse engineering <1%	No universal security metrics available
	Split manufacturing based on fine pitch HI	Not available

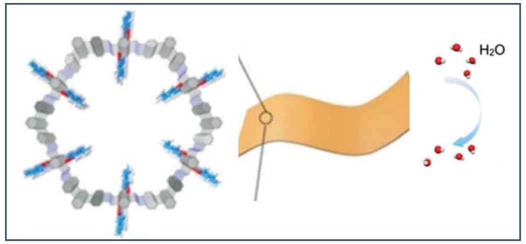


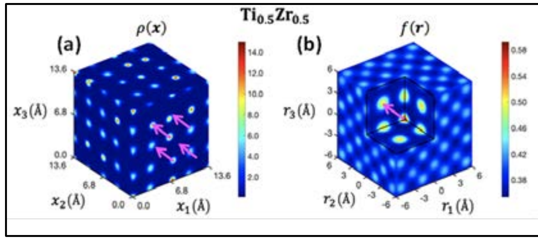


The Roadmap – We live in a Material World

Metrics & Drivers	Metrics (10 Years)	State of the Art (SOTA)
Ultra-low K Dielectrics AI, C, DC, M)	Permittivity < 2.5, Loss 0.01 @ 1THz, Thickness: 1μm – 100μm, CTE: <30 ppm/C	Permittivity 3.0, Loss 0.05 @ 150GHz, Thickness: 5mm, CTE: 30-40 ppm/C
Interconnects (AI, DC, M)	For BEOL Devices: 100nm pitch, 50nm dia., aspect ratio 2-10 for 2-5 tiers for via bottom-up fill.	400nm pitch, 200nm dia., aspect ratio up to 5 (selective ALD for Co from ASCENT)
	TSVs: 100 nm diameter, 250 nm pitch; aspect ratios >30:1	9 μm pitch and ~2μm diameter [AMD 3D V-cache]
	Misalignment-tolerant bonding process for inter-layer interconnects at 250 nm pitch	5.5 μm pitch using face-to-face hybrid bonding [ARM/GF]
Thermal (AI, C, S, DC, H)	TIM: 200W/m-K; Thickness <30μm; CTE 20 ppm/C	TIM: 25-100 W/m-K; CTE 55-200 ppm/C
	Heat Spreader: - cBN 200 W/mK at 100 nm thick, 500W/mK at 1 μm - Diamond 400 W/mK at 1 μm, 1500 W/mK at 10 μm, 2000 W/mK at 50μm.	Heat Spreader: Cu 400 W/mK bulk
	Ultra-low interface thermal resistance, e.g. GaN/diamond 3 m^2K/GW, Si/diamond 3 m^2K/GW	GaN/diamond 6.5 m^2K/GW Si/diamond 9.5 m^2K/GW
	Thermal isolation materials: 0.02 W/m-K, dielectric constant < 1.5, thermally stable above 300°C	Thermal isolation materials (SiCOH): 0.6 W/m-K, dielectric constant 1.8-2.5, and thermally stable above 300°C

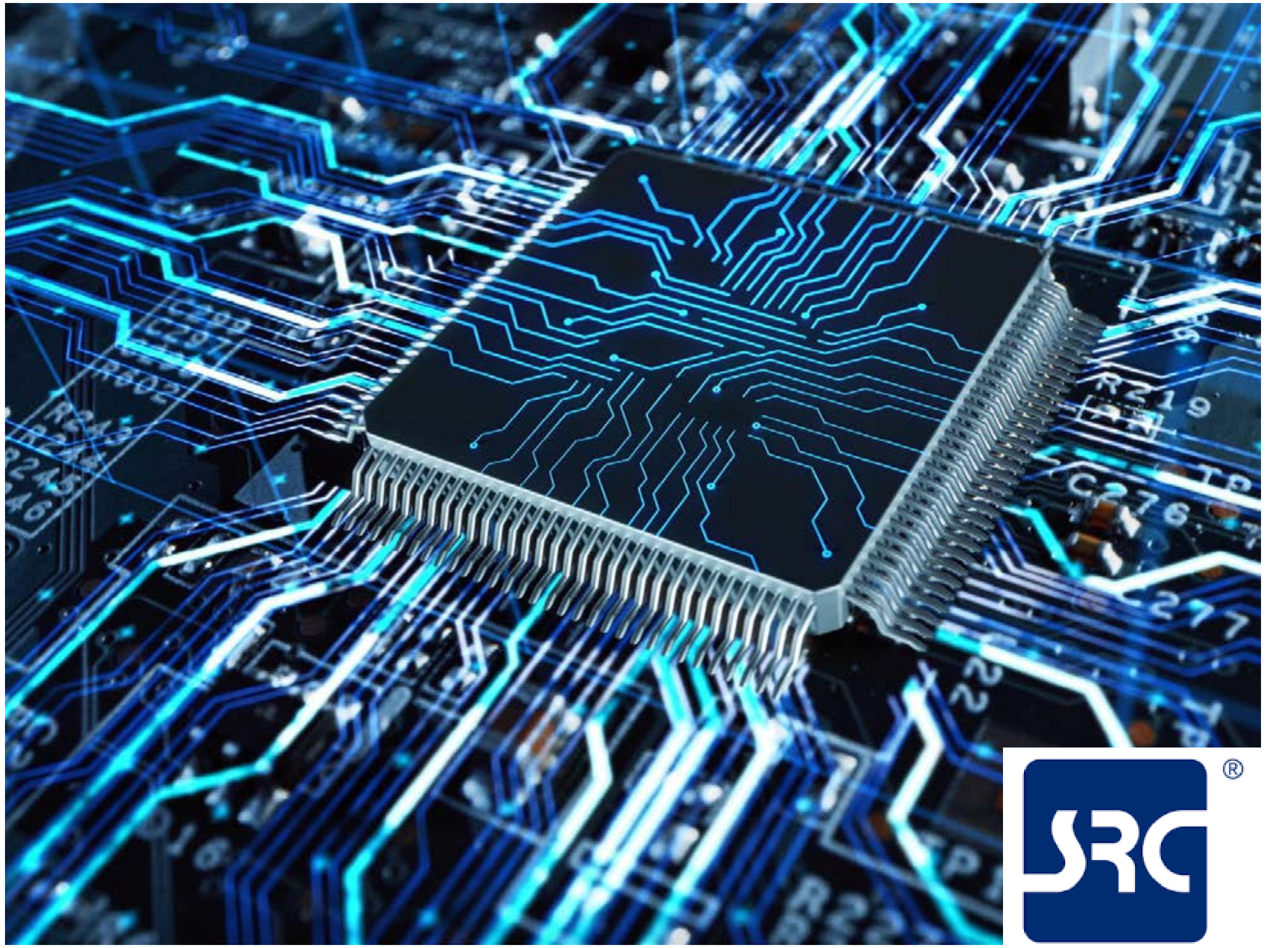






Introducing JUMP 2.0 Center @ Penn State (14 Univ. Partners)

Penn State leads semiconductor packaging, heterogeneous integration center



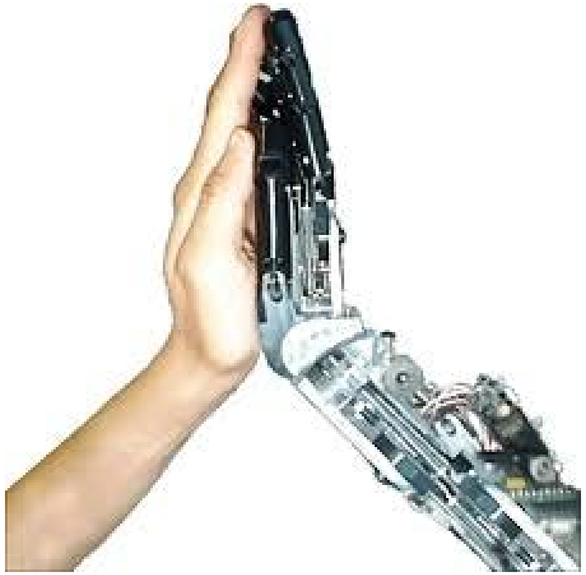
- Center for Heterogeneous Integration of Micro Electronic Systems (CHIMES)
- Supported by the Semiconductor Research Corporation (SRC)'s Joint University Microelectronics Program 2.0 (JUMP 2.0), a consortium of industrial partners in cooperation with the Defense Advanced Research Projects Agency (DARPA)

https://www.psu.edu/news/engineering/story/penn-state-leads-semiconductor-packaging-heterogeneous-integration-center/

The CHIMES Team

PennState
Georgia Tech
UCLA
Columbia University in the City of New York
Cornell University
The George Washington University, Washington, DC
Arizona State University
Massachusetts Institute of Technology
Stanford University
Rice University
UC Davis University of California
UC San Diego
University of Illinois Chicago
University of Colorado Boulder


Thank
You
Obrigado
Merci
Multumesc
Asante
Grazie
Spasibo
Maake
Kiitos
Dank Je
Arigato
Vinaka
Matondo
Mamana
Juspaxar
Welalin
Kia Ora
Nirringrazzjak
Mochchakkeram
Ua Tsaug Rau Koj
Cam on ban
Chokrane
Dankon




NAMASTÉ