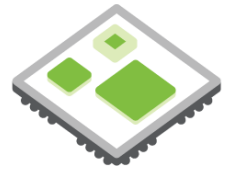
SERVER:
OPEN CHIPLET
ECONOMY



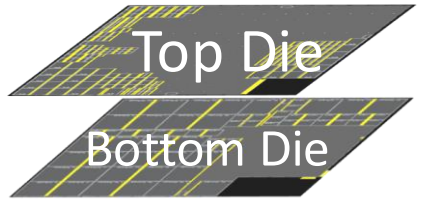
Scaling up AI Compute, Memory and Connectivity

Igor Elkanovich

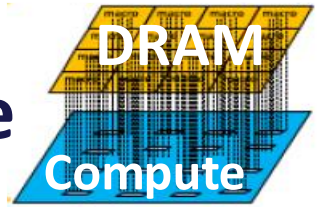
GUC CTO

Agenda

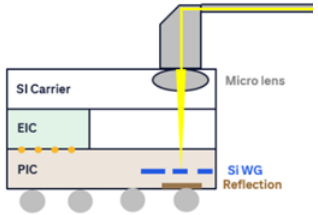
- **Scaling up compute: 2.5D CoWoS® and 3D SoIC®-X**



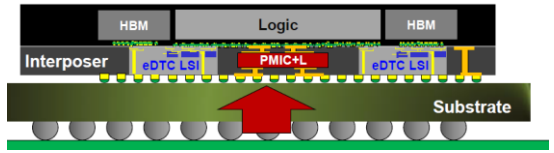
- **Memory: HBM5 and DRAM on Compute**



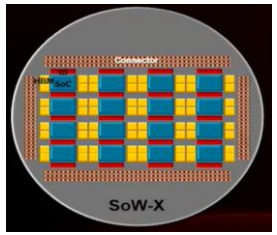
- **Co-Packaged Optics (CPO) and COUPE™**



- **Vertical power delivery and Integrated Voltage Regulator (IVR)**

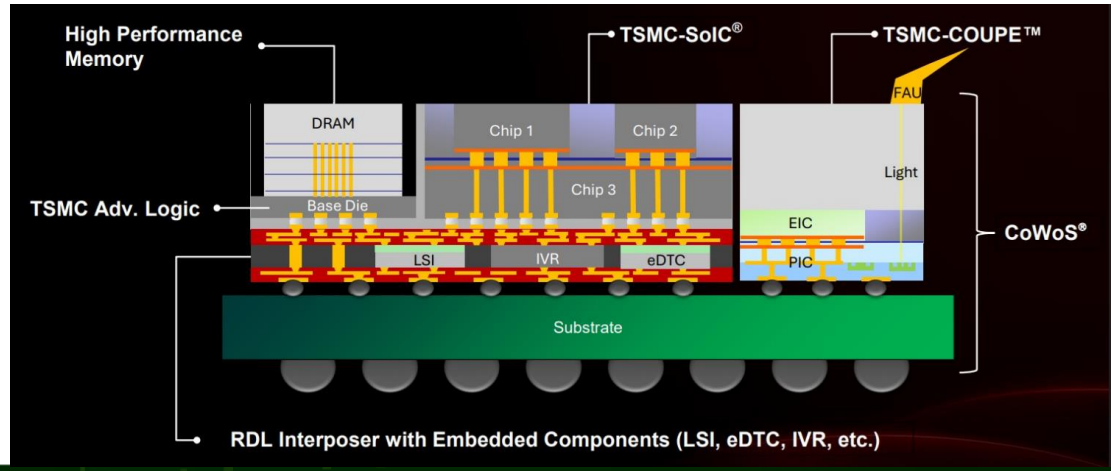


- **System on Wafer**



Scaling up AI Compute, Memory and Connectivity

- Compute
  - 3D integration (SoIC-X) is a key approach to increase compute density
  - High power density (5-10W/mm^2) is a key challenge
  - Vertical power delivery to minimize loss
    - Voltage Regulators integration into PCB cavity → package substrate → CoWoS-L interposer
  - Cooling compute dies, memories (max 95C), CPO (temp. stability) and VRs is a big challenge
- Memory
  - HBM5 doubles memory bandwidth vs. HBM4E
  - DRAM on Compute: architectural flexibility, high bandwidth, low latency and power
- Connectivity: CPO
  - COUPE: high bandwidth, high radix and power efficient Optical Engines
  - COUPE supports both Scale Out and Scale Up



Low Power & Latency 2.5D Die-to-Die Interface

- UCIe limitations
  - UCIe architecture is complicated – extra power and latency
  - In most of use cases standard compliance is not required
- GLink-LL: low power & latency D2D
  - Moderate bandwidth
  - Power <0.1 pJ/bit
  - Latency: max 3ns including CDC FIFO
  - BER <1E-30, no CRC and Retry required

	UCIe-S	UCIe-A	GLink-LL
Speed	64Gbps	64Gbps	4-8Gbps
Bump/Bond Pitch	Standard package 110um	CoWoS® 45um	CoWoS® 35um SoIC®-X-CH 6um
IP Depth	2mm	1.8mm	0.2-0.8mm
Trace Distance	25-40mm	2mm	0.4-1.0mm
Edge BW Efficiency	0.37 TBps/mm	2.6 TBps/mm	0.3-0.6 TBps/mm
Area BW Efficiency	0.19 TBps/mm^2	1.4 TBps/mm^2	0.7 TBps/mm^2
Power Efficiency	1.25 pJ/bit	0.75 pJ/bit	**0.05-0.10 pJ/bit**
BER	BER<1E-15	BER<1E-15	**BER<1E-30**
CRC+Retry	CRC+Retry	CRC+Retry	**Not required**
Latency with FIFOs	10ns + Retry	10ns + Retry	**1-3ns**

3D Top-Bottom Dies Interface

- 3D interface is customized for every project
  - Bond map, DDR/SDR architecture, signal/clock ratio, redundancy, power domains, dimensions
- GUC's GLink-3D has been silicon-proven and adopted in customer's AI ASICs
  - High bandwidth (80 Tbps/mm^2), low power (0.03 pJ/b) and low latency (0.25ns)
- Supports DVFS: voltage and frequency can change while keeping data transfer
- Flexible IP delivery types: soft, semi-soft or hardmacro

Silicon Proven

GLink-3D 1.0
Architecture: DDR
Process: N6/N5
Power: 0.2 pJ/bit
Bandwidth*: 9 Tbps/mm^2
Latency: 1-2ns

Silicon Proven

GLink-3D 2.0
Architecture: SDR, DDR
Process: N2P/N2P
Power: 0.03 pJ/bit
Bandwidth*: 50 Tbps/mm^2
Latency: 0.125-0.5ns

Silicon Proven

GLink-3D 2.1
Architecture: SDR
Process: N3P/N5
Power: 0.03 pJ/bit
Bandwidth*: 40 Tbps/mm^2
Latency: 0.25ns

Execution

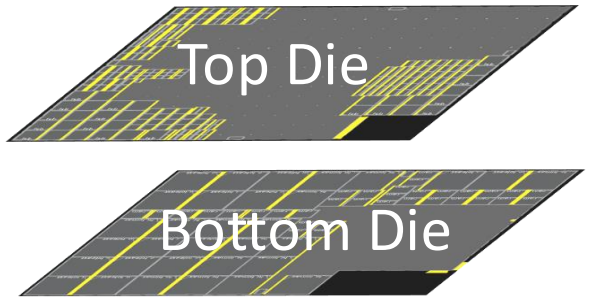
GLink-3D 3.0
Architecture: SDR, DDR
Process: N2P/N2P
Power: 0.03 pJ/bit
Bandwidth*: 80 Tbps/mm^2
Latency: 0.125-0.5ns

* Full Duplex Bandwidth

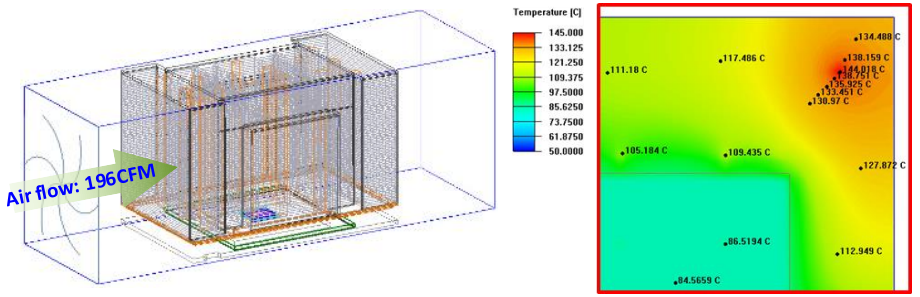
3D Key Challenges

- Power delivery for high power density (7.5 W/mm² average TDP) solutions
  - Optimal distribution of bottom die's TSVs to meet global TSV density
  - Optimizing TSVs channels between bottom die's large SRAMs
  - In N2P TSV overhead was reduced by ~40% vs N3P increasing routing resources
- Power and thermal analysis at early stage to avoid late stage iterations
  - Architecture and floorplan are adjusted for better power integrity and thermal
- Mechanical strength:
  - Reducing warpage, package structure and chiplets floorplan optimization, materials selection

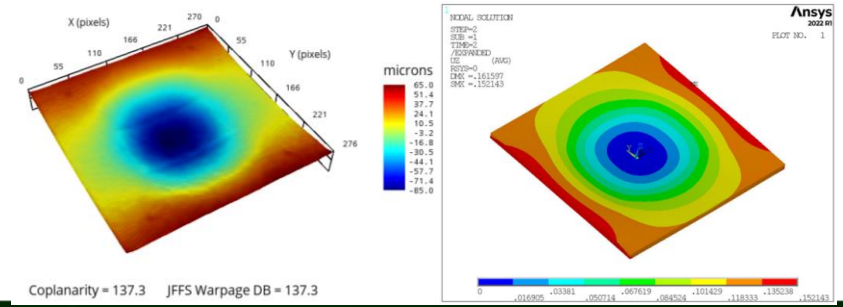
Power Delivery Optimization



SoIC®-X Thermal Analysis



Warpage Minimization



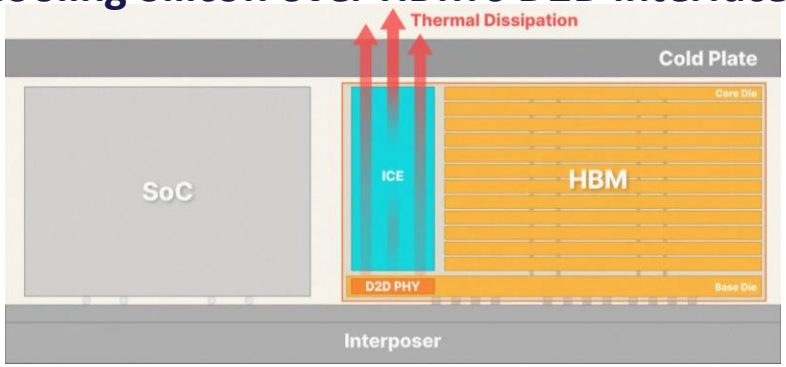
From HBM4E to HBM5

- GUC's HBM4E PHY/Controller at 16Gbps is ready
  - PHY supports face-up configuration for SoIC-X bottom die and top die's logic over the PHY
  - Controller is optimized for high bandwidth utilization at random access
    - It supports advanced RAS functionality to prevent silent data corruption
- HBM5 targets 24-32Gbps, will be ready for production in 2029
  - HBM5 PHY area will be reduced by 40% by using smaller u-bump pitch
  - Thermal is a key challenge, HBM4E at 16Gbps is ~60W while HBM5 at 32Gbps will be ~90W
    - HBM5 D2D interface is the main heat source, solution: cooling silicon over D2D interface

HBM5 vs. HBM4E

	HBM4E	HBM5-HBM5	Difference
Speed per pin	16Gbps	24-32Gbps	**1.5x-2x**
PHY size	14 mm^2	8 mm^2	**0.6x**
PHY area efficiency	0.3 TB/mm^2	0.75-1.0 TB/mm^2	**2.6x-3.5x**
Availability	2026	2029	

Cooling Silicon over HBM's D2D interface



Source: SK Hynix Newsroom May 2026

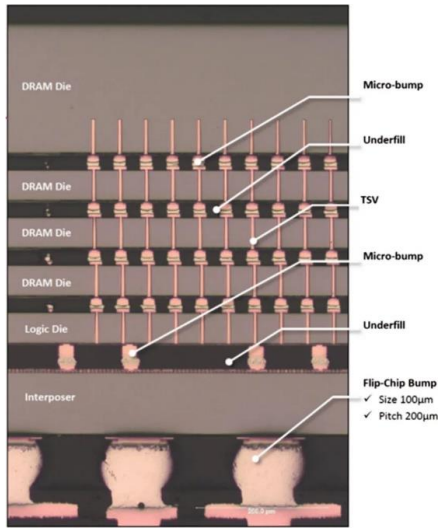
DRAM on Compute Logic

- DRAM on Logic advantages: higher bandwidth, lower power and latency
- HBM uses TCB 3D stacking with pitch ~20um (HBM4E)
  - Due to low TSV density HBM shares TSVs between banks
- DRAM on Logic uses Hybrid Bonding (HB) 3D stacking with pitch 5-8um
  - Interface is 10-15x denser, separate wide buses per every bank are used
- Maturity and cost
  - HBM is cost efficient for its long term manufacturing and improvements
  - HB stacking is not widely used yet but there is strong dynamics of HB adoption

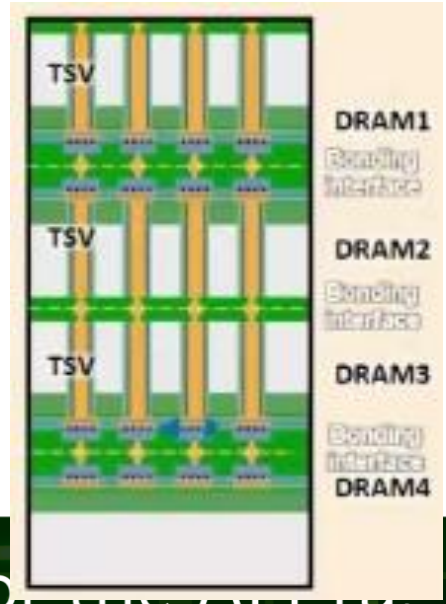
	DRAM on Compute	HBM4E
Total BW	32** TBps	24*** TBps
Memory density	4 layers: 110 MB/mm^2	12 layers: 310 MB/mm^2
Latency*	30 ns	200 ns
Power*	0.5 pJ/bit	2.5 pJ/bit
Max temperature	125C	95C

* Including PHY and Controller **Full reticle size 4 layer DRAM ***6xHBM4E at 16Gbps

HBM Stacking: TCB

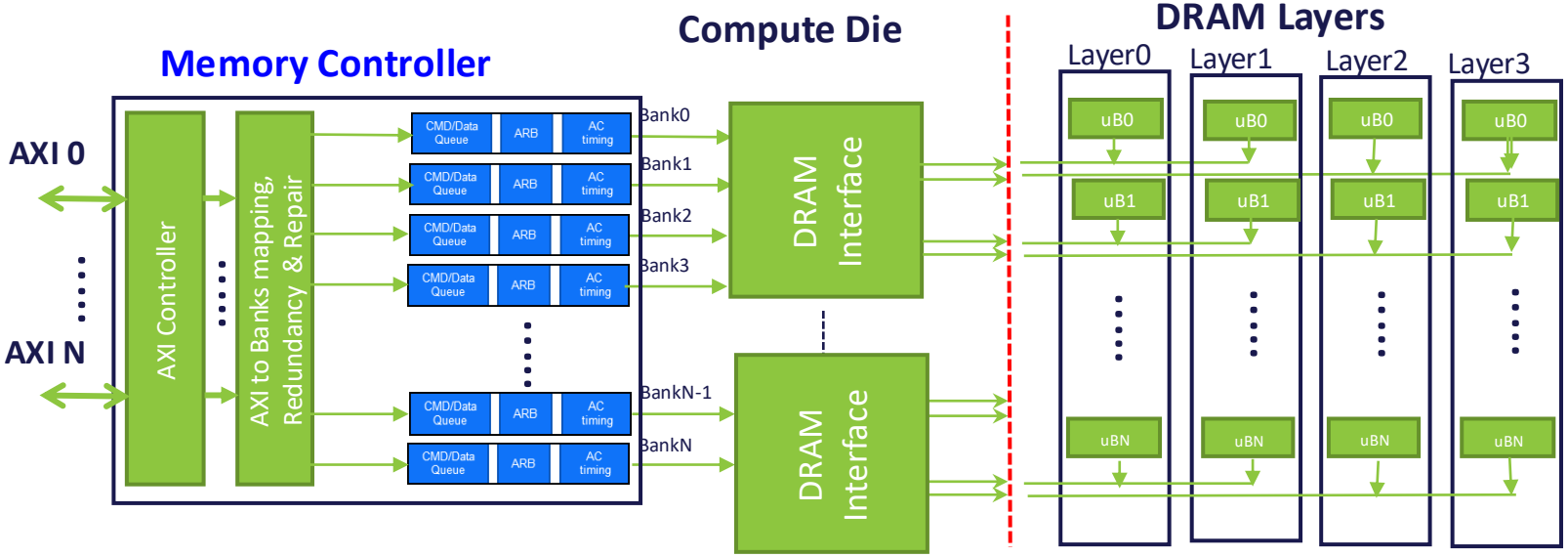


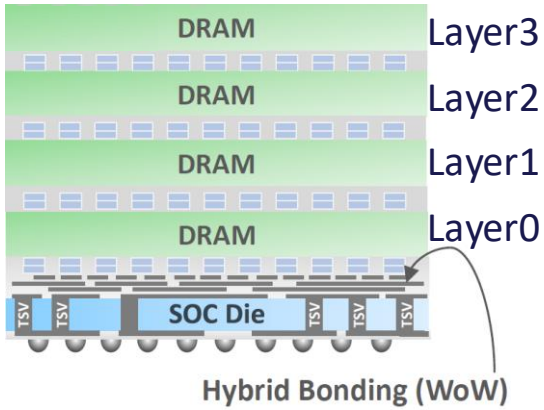
DRAM on Logic Stacking: Hybrid Bonding (HB)



Memory Controller for DRAM on Compute

- Memory Controller is a key enabler for DRAM on Compute
- GUC's Memory Controller
  - Based on our HBM4E Controller, optimized for area efficiency and direct access to all DRAM banks
  - Supports multi-bank interleaving and access reordering for high bandwidth utilization
  - Interoperable with any DRAM vendor
  - DRAM yield management by banks redundancy and repair
  - Bandwidth utilization: ~100% at sequential access, 80%-70% at random access
  - Latency: <30ns from AXI read command to AXI read data

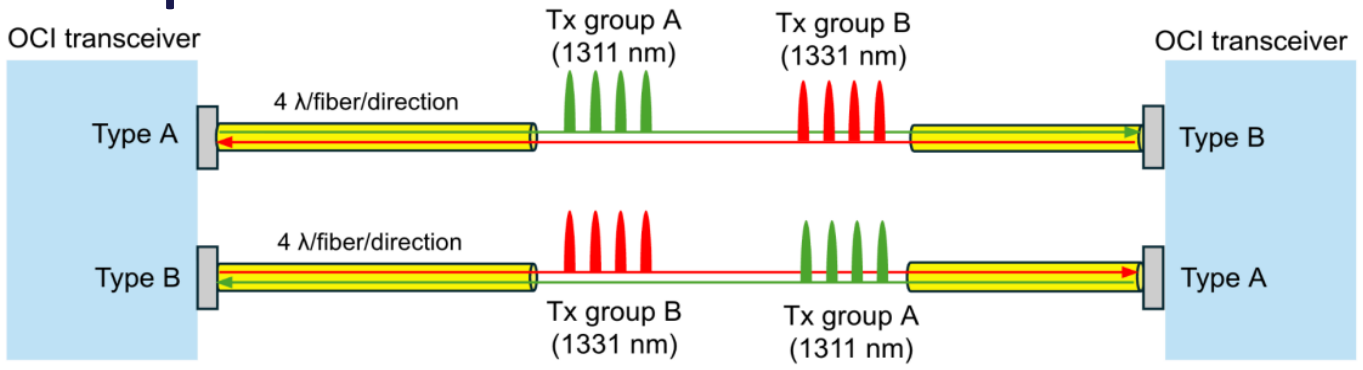


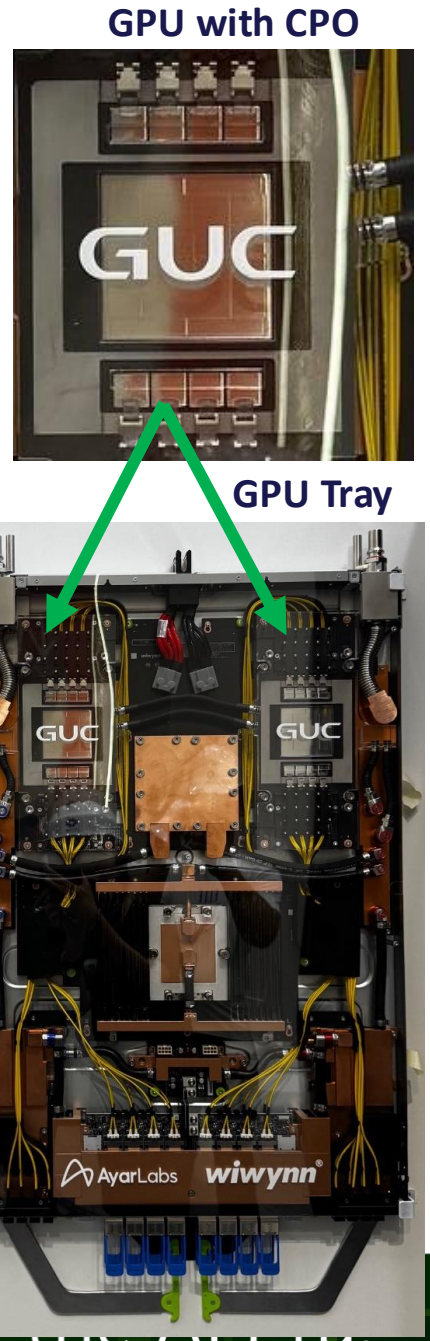


TSMC COUPE for Co-Packaged Optics (CPO)

- TSMC COUPE: high bandwidth, high radix, power efficient Optical Engines
  - TSMC COUPE's tiny MRM/MRR and dense Fiber Attach (FAU) fit Slow-Wide
- Optical Engines functionality:
  - Scale Out: IEEE-DR 200G/PAM4 per fiber, per direction
  - Scale Up, Slow-Wide OCI-gen2: 4xλ per direction 100G/NRZ
    - DWDM of 4xλ NRZ datastreams in a single bidirectional fiber
    - Area and power efficient NRZ modulation instead of PAM4
    - Light, low latency FEC is used due to much lower NRZ BER

OCI Spec: DWDM channels for a Bidirectional Link

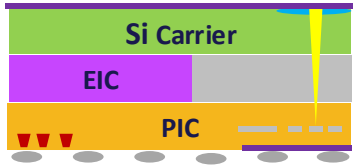




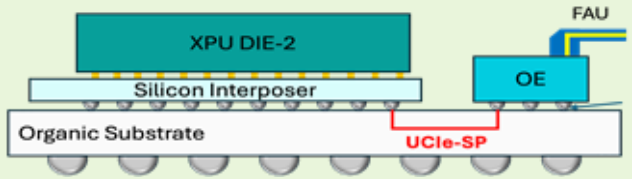
COUPE Manufacturing and Testing Flow

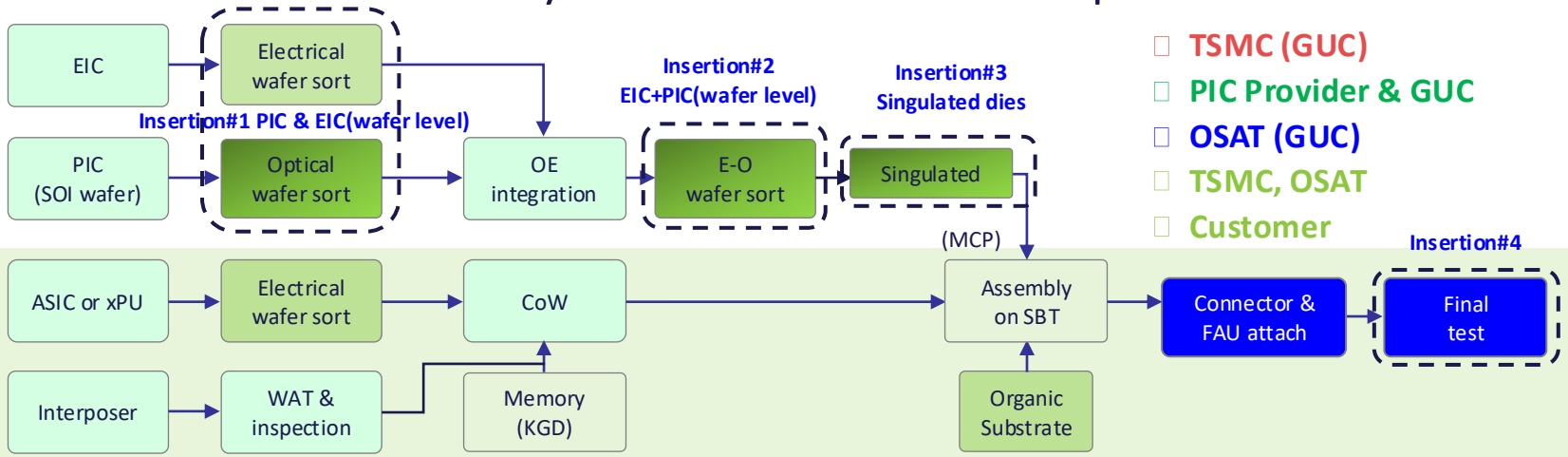
- GUC's COUPE manufacturing and testing flow uses robust and scalable ecosystem
  - Optical Engines are assembled on substrate aside CoWoS island
- Four test insertion steps to ensure Optical Engines (OE) quality before XPU package assembly
  - Insertion 1: separate PIC and EIC wafer level tests
  - Insertion 2: PIC+EIC wafer level test after 3D stacking
  - Insertion 3: OE test after dies singulation
  - Insertion 4: Final test after entire xPU with OEs assembly with attached fiber and optical connector

Optical Engine (OE) Assembly



xPU with OE Assembly



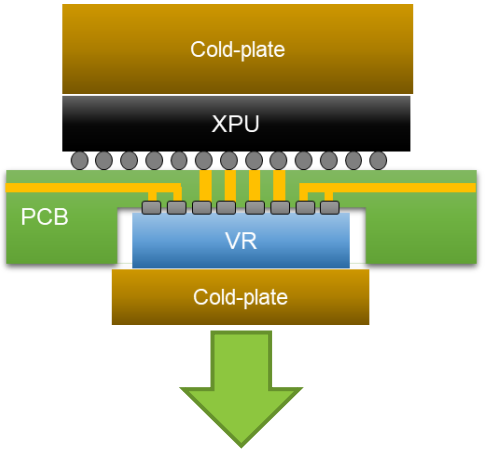




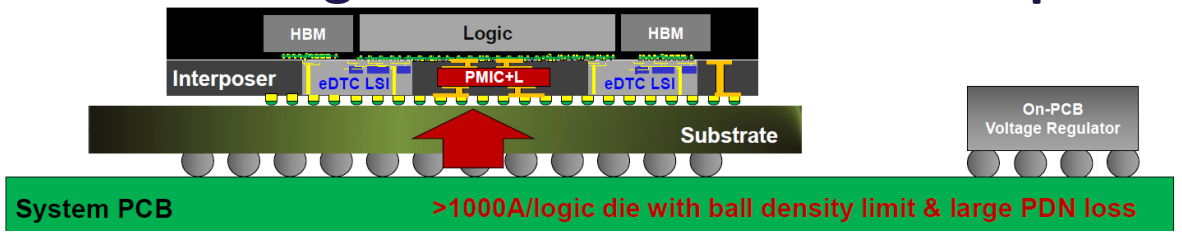
Vertical Power Delivery

- For few kW ASICs vertical power delivery is a must to minimize loss
- Voltage Regulators integration steps
  - PCB cavity → package substrate → CoWoS-L interposer
- GUC's IVR advantages
  - Low IR drop, Integrated inductor
  - Response time is 10x-100x faster than on-board VRM
  - DVFS per CPU/AI core
  - Current per ball and C4 bump (EM) is >2x lower
- GUC demonstrated IVR using TSMC 16nm
  - Dual-phase IVR per CPU/AI core
  - Input voltage: 1.8V
  - Programmable output voltage: 0.5V-1.0V
  - Efficiency: 85% at high current, 75% at low current
  - Power density: 3W/mm2

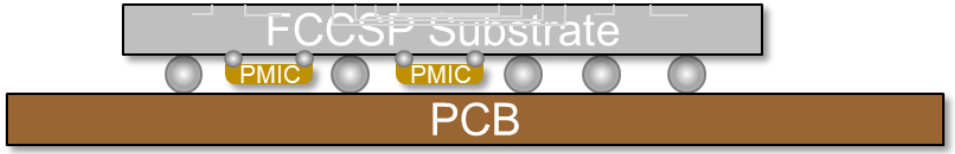
Vertical Power Delivery



IVR Integrated into CoWoS®-L Interposer

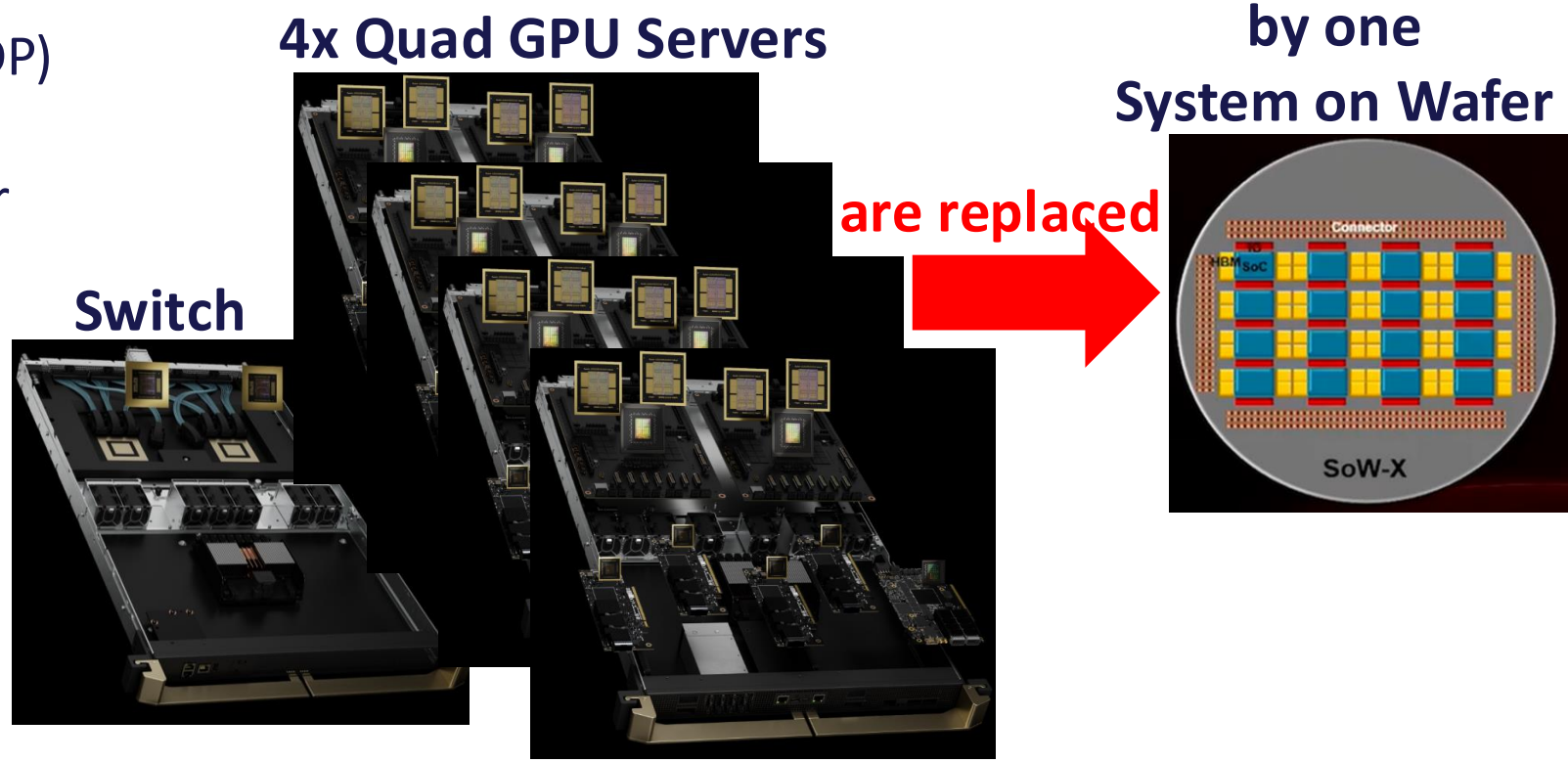


VR at Land Side (Handheld Application)



System on Wafer (SoW)

- Tens dies and HBMs in one wafer, for huge scale AI and networking applications
  - >40x reticle size of silicon area (>33,000mm^2), up to 64 HBM4E, high speed IOs and connectors
  - UCIe interfaces replace SerDes: 10x less power, 20x lower latency
- Key challenges:
  - Heat dissipation: 20kW power (TDP)
    - HBM max temperature: 95C
  - Power delivery: 30kW peak power
- SoW enabling technologies:
  - UCIe 64Gbps Die-to-Die interface
  - 224Gbps routing using RDL layers
  - Power efficient implementation
  - Thermal/mechanical
  - Redundancy and repair
  - 3DIC and many more…



Summary

- 3D significantly increases compute density, its key enablers:
  - High bandwidth, low power, low latency 3D interface
  - Power delivery for high power density and thermal optimization
  - Few generations of 3D products were validated in customer's systems
- Vertical power delivery supports high power density with minimal loss
  - IVR will be integrated in CoWoS-L interposer for fast response time and per-core DVFS
- Memory
  - HBM4E PHY/Controller at 16Gbps is ready
  - HBM5 doubles memory bandwidth vs. HBM4E and increases PHY area efficiency
  - DRAM on Compute for flexible architecture, high bandwidth, low latency and low power
- Connectivity: TSMC's COUPE for CPO Optical Engine
  - COUPE enables high-bandwidth, high-radix and power-efficient system scaling
  - It is mature and supports both Scale Out and Scale Up requirements
- System on Wafer: a game changer, 5x more silicon area and HBMs than the biggest CoWoS


Thank
You!
OCP
GLOBAL
SUMMIT
October 12-15, 2026
San Jose, California