GUC
The Advanced ASIC Leader

GLink-2.5D and GLink-3D

Customers Presentation

By Igor Elkanovich, GUC CTO

April, 2022

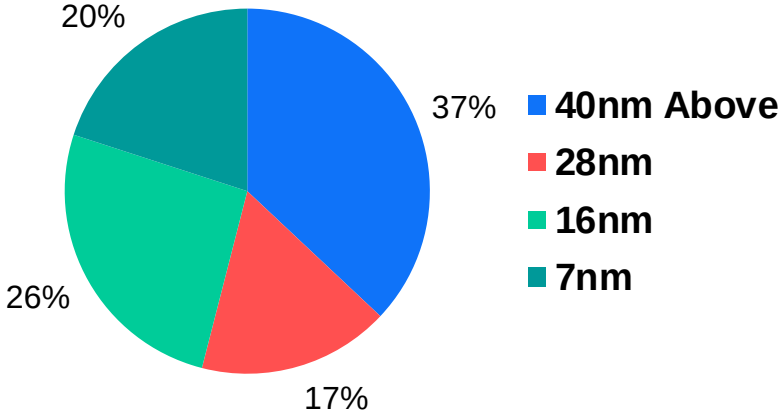
Agenda

- GUC Introduction
- GLink-2.5D
- GLink-2.5D in multi-die processors
- GLink-3D

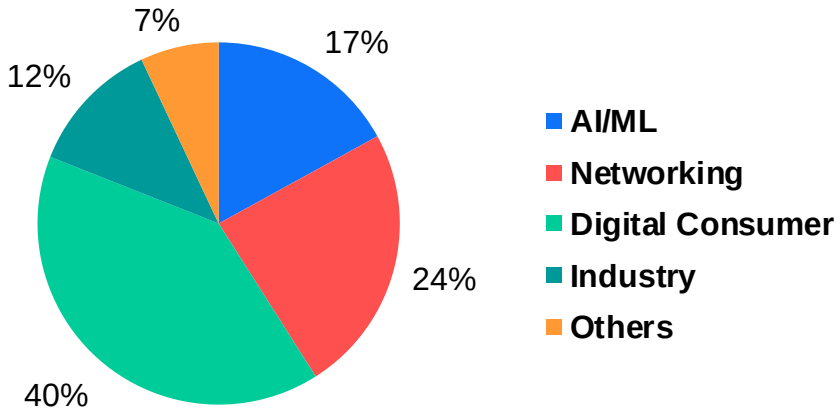
GUC Introduction

- Established in 1998, HQ in Hsinchu, next to TSMC HQ
- Traded in Taiwan Stock Exchange, TSMC holds ~35% shares
- Services : ASIC, IP, Production
- Offices : USA, China, Europe, Japan and Korea
- Employees: ~820, 2021 revenue: ~$540M

2021 Revenue Breakdown By Technology



2021 Revenue Breakdown By Application



GUC 2.5D and 3D Multi-die APT Platform

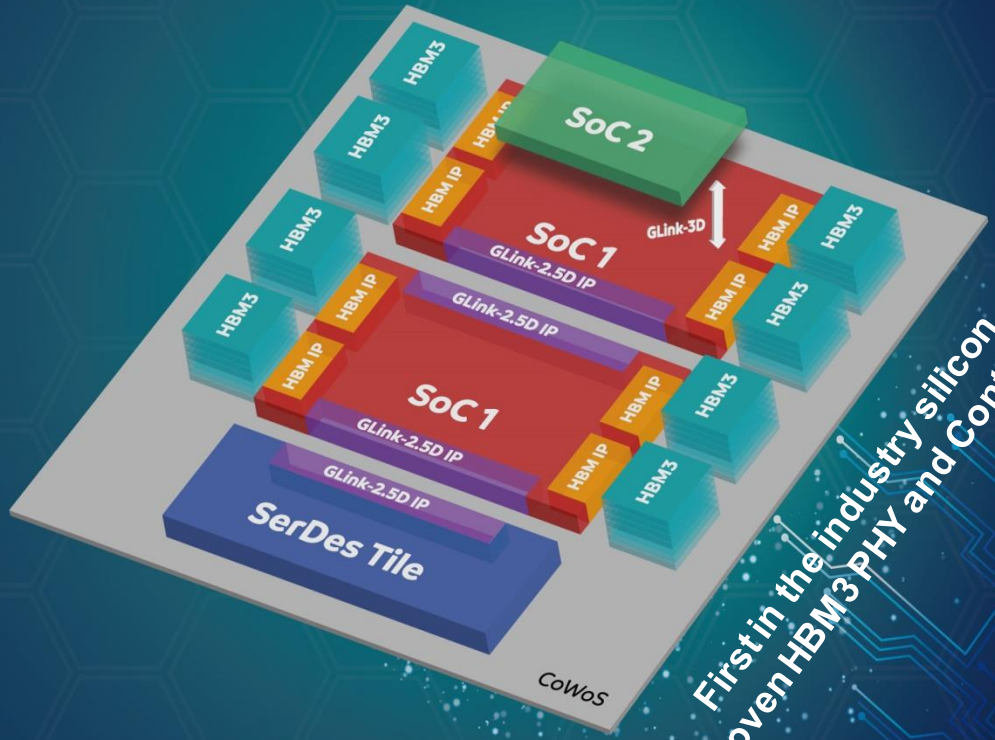
GUC
The Advanced ASIC Leader

Platform Highlights

- HBM3 PHY + Controller IP
- Die-to-die IP (GLink-2.5D)
- Die-on-die IP (GLink-3D)
- High speed IP integration (112G SerDes, PCIe-5, GDDR6)
- Advanced Packaging Technology (CoWoS / InFO / 3D-SoIC)

Service Scope

- SoC and ASIC design to production
- Interposer and RDL design
- SI/PI/IR/THM simulation
- Package and substrate design

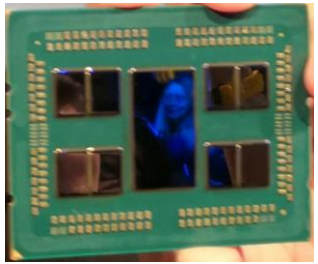


First in the industry silicon proven HBM3 PHY and Controller

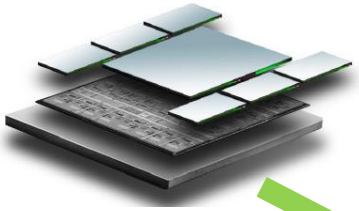
Chiplets: Substrate Integration vs. 2.5D/3D

- **Multi-dies integration:**
  - Traditional: on substrate (MCM)
  - Emerging (GUC): 2.5D on Silicon/RDL (CoWoS®/InFO), 3D stacked dies (TSMC-SoIC™)
- **Next (GUC): 3D stacked dies (TSMC-SoIC™) on top of 2.5D (CoWoS®/InFO)**

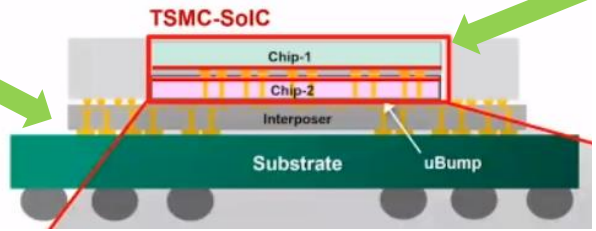
Dies on Substrate Integration (MCM)



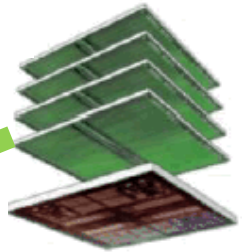
GLink-2.5D for CoWoS®/InFO_oS



3D (SoIC™) on 2.5D (CoWoS®/InFO_oS)



GLink-3D for SoIC™ CoW/WoW



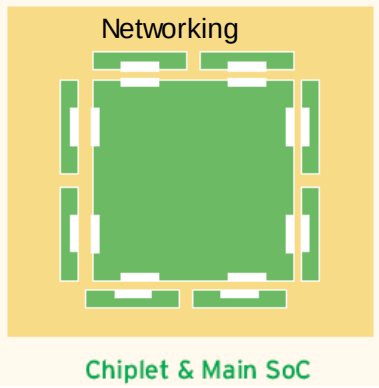
Agenda

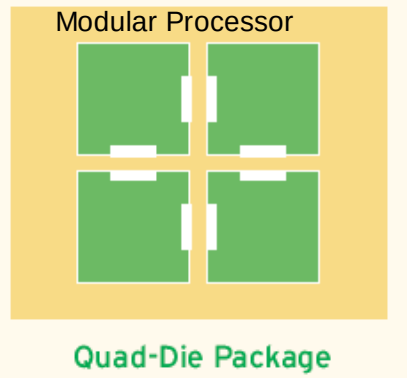
- GUC Introduction
- GLink-2.5D
- GLink-2.5D in multi-die processors
- GLink-3D

GLink-2.5D Functionality and Use Cases

- **The most power, area and speed optimized solution for multi-die integration using InFO_oS or CoWoS**
- **Reliable Solution**
  - No errors, error correction is not used
  - DFT functionality for separate dies testing and InFO_oS/CoWoS assembly testing
  - Redundant lanes embedded to achieve better yield
- **GUC provides Total Service Package:**
  - Sub-system design
  - InFO_oS RDL and CoWoS Interposer routing
  - SI/PI/Thermal simulation services

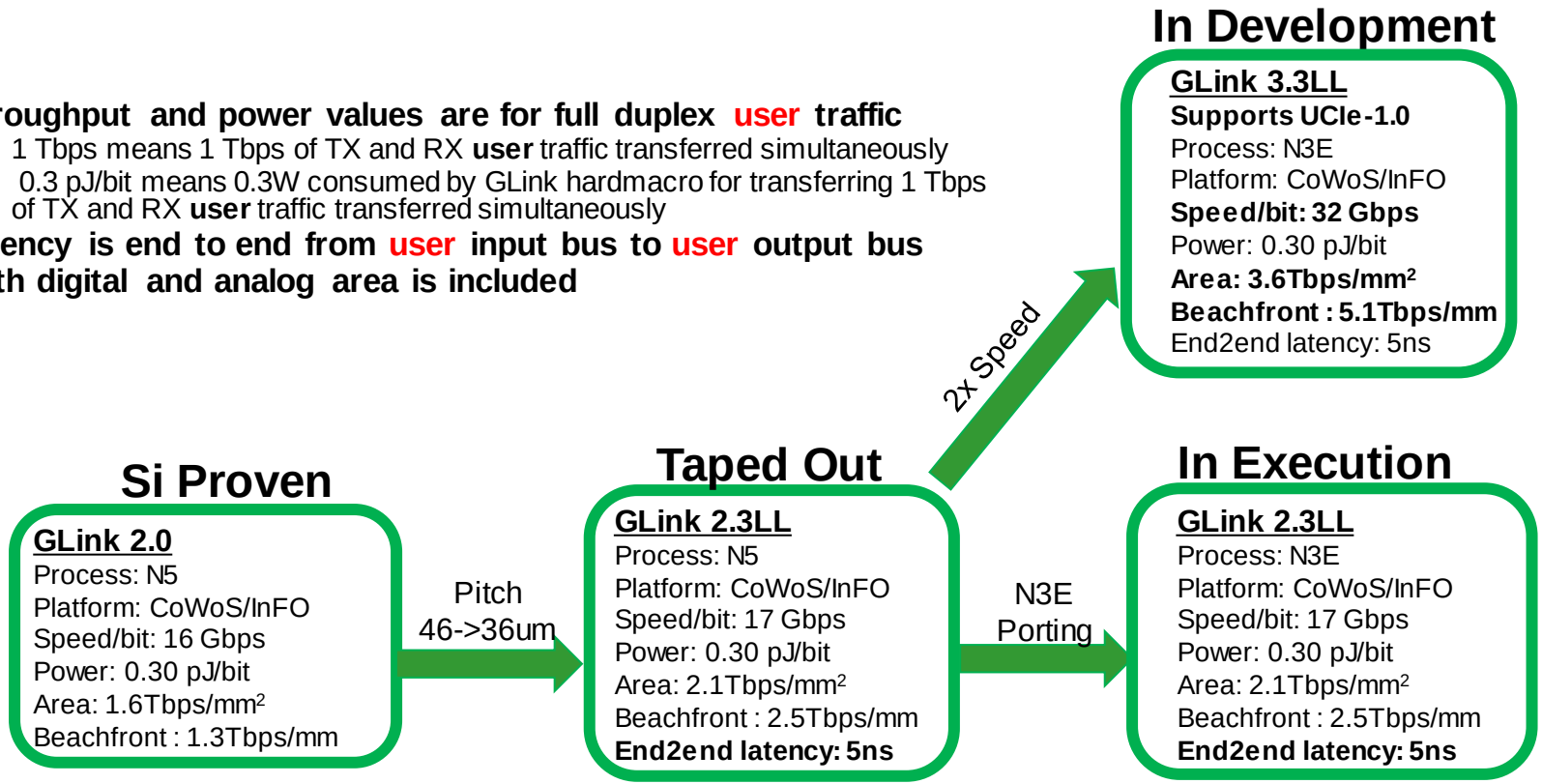
Use Cases





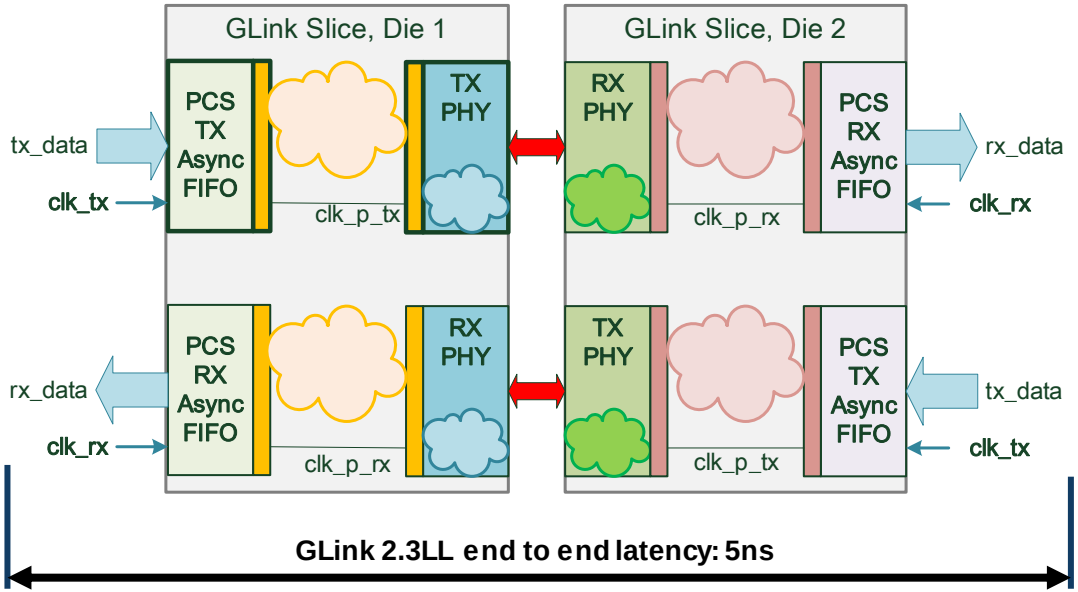
GLink 2.5D Roadmap: 2nd and 3rd Generations

- **Throughput and power values are for full duplex user traffic**
  - 1 Tbps means 1 Tbps of TX and RX **user** traffic transferred simultaneously
  - 0.3 pJ/bit means 0.3W consumed by GLink hardmacro for transferring 1 Tbps of TX and RX **user** traffic transferred simultaneously
- **Latency is end to end from user input bus to user output bus**
- **Both digital and analog area is included**



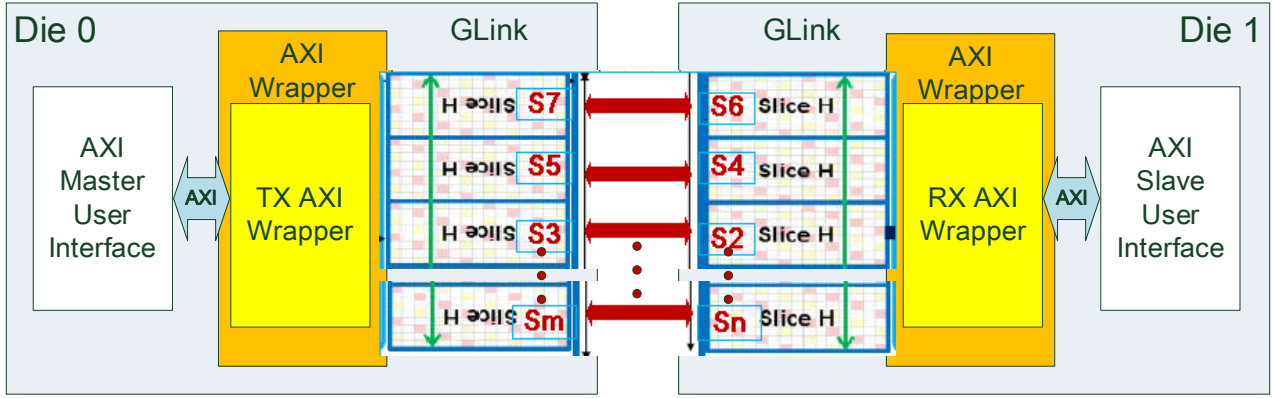
GLink Architecture

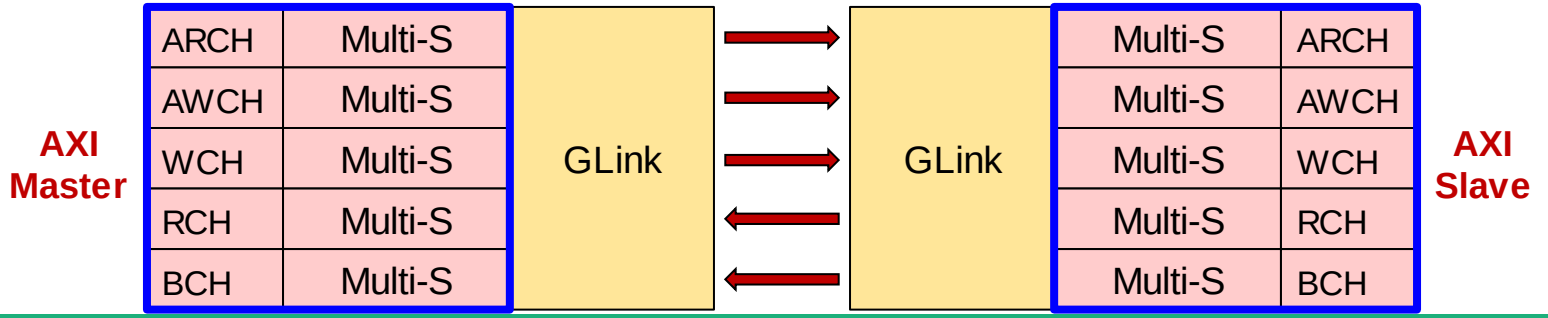
- Two fully asynchronous FIFOs are placed at RX and TX sides to allow asynchronous clocks



GLink as AXI Bridge Physical Layer

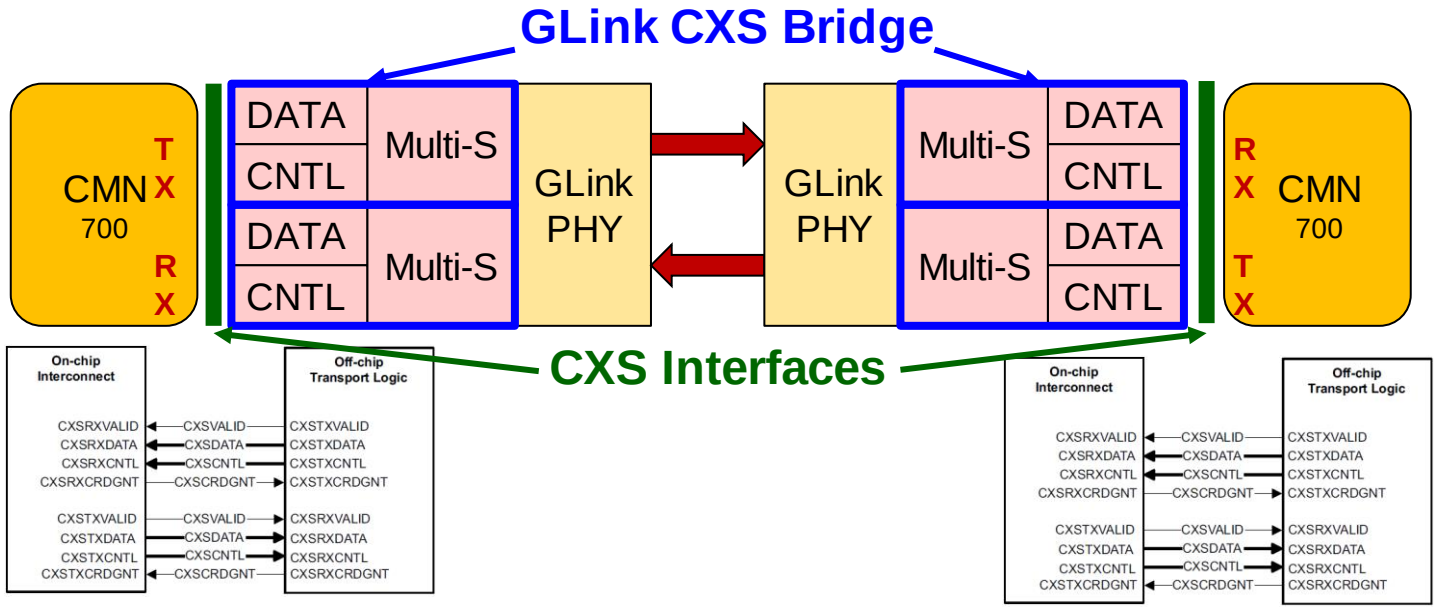
- AXI Bridge Master and Slave channels are mapped to GLink IP
- GLink adjusts data widths to Bridge Channels widths
- Fully reliable data transfer by GLink is a key for robust AXI end to end operation





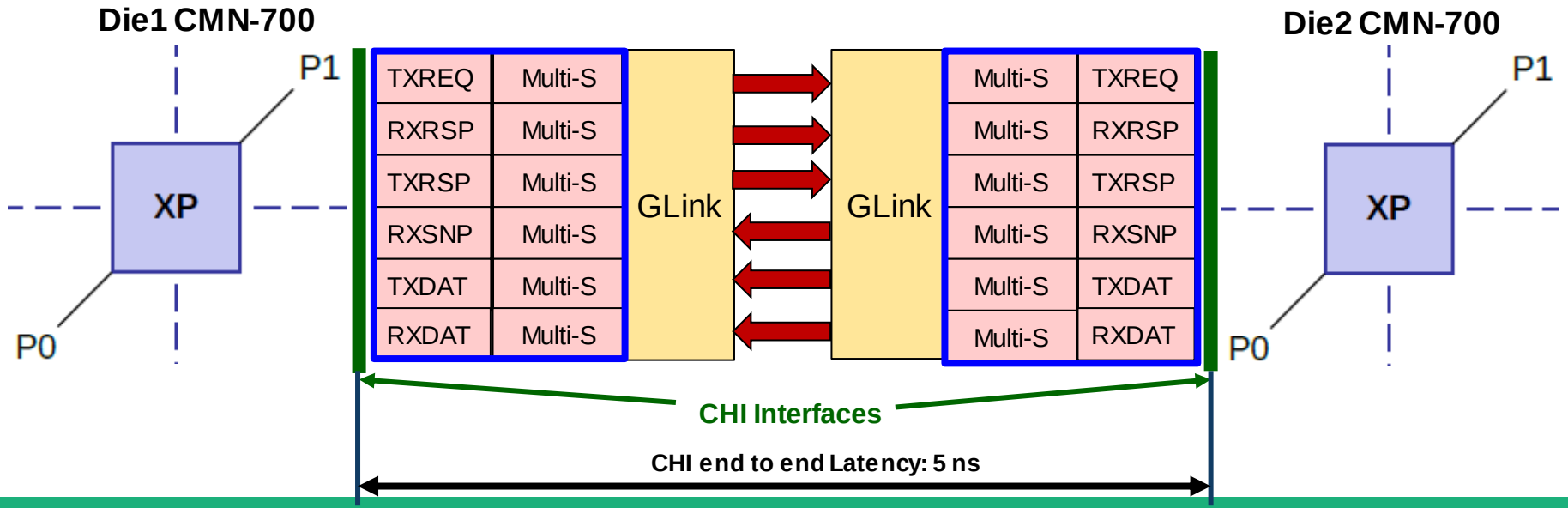
GLink as CXS Bridge Physical Layer

- CXS bus is used as CMN-700 die to die interconnect
- CXS Issue B is supported
- CXS bus uses credit-based Flow Control
- Fully reliable data transfer by GLink allows to skip Link and Transaction Layers



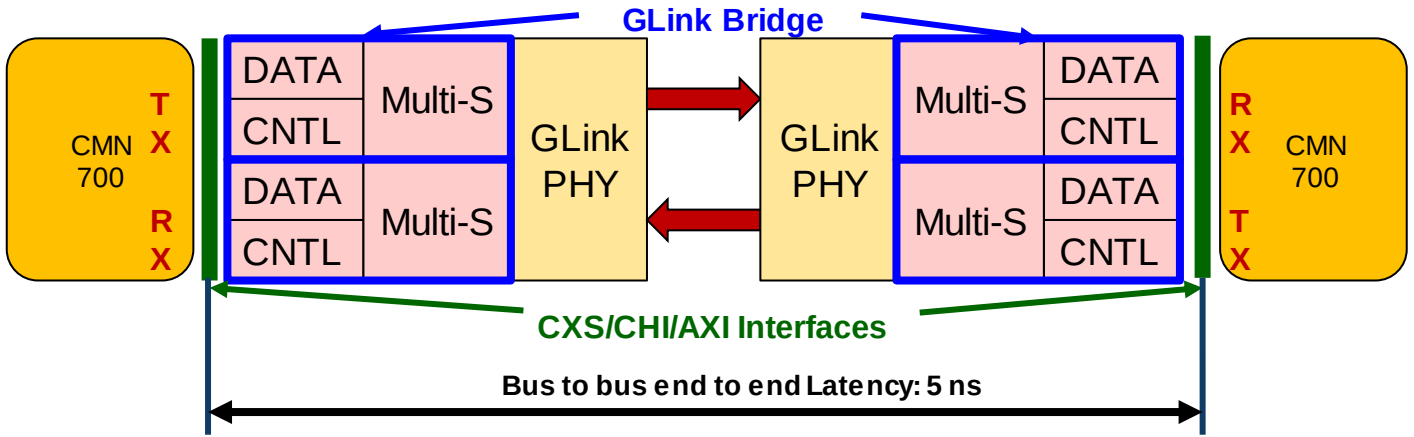
CHI Bridge

- CHI bus is used as direct CMN-700 XP to XP die to die interconnect
- CHI Issue E is supported
- CHI bus uses credit-based Flow Control similar to CXS
- Error-free data transfer by GLink allows to skip Link and Transaction Layers



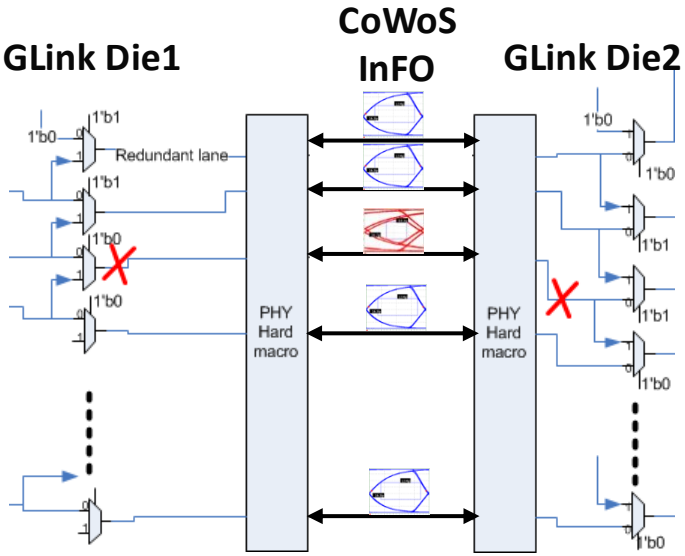
CXS/CHI/AXI Bridges Latency

- **Bridges end to end latency: 5ns**



GLink Signal Quality Monitoring by proteanTecs

- CoWoS/InFO connectivity quality is monitored in field, in mission mode
- Bumps and traces defects are identified before they cause interface fail
- Repair algorithm replaces marginal I/Os with redundant ones
- System operation failure is prevented, chip lifetime is extended
- Silicon proven in N7 and N5 HBM2E testchips
- Silicon proven in N5 GLink 2.0 testchip

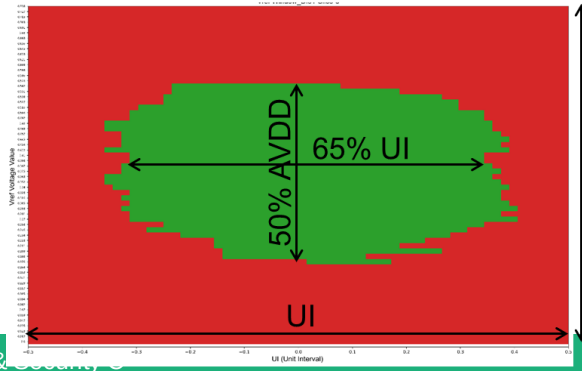


CoWoS u-bumps resistance sensitivity measured in GUC's N7 HBM2E Testchip

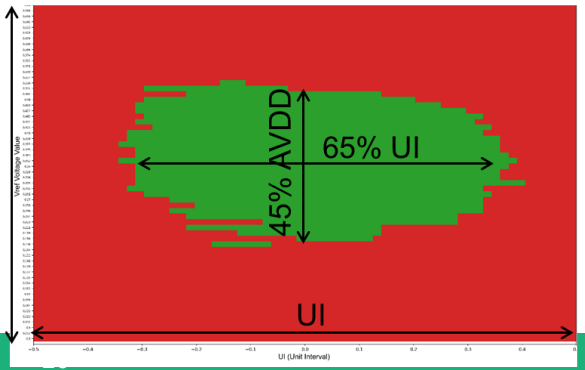
GLink 2.0 Testing Status Summary

- **Robust and stable chip operation at 16 Gbps**
  - Wide eye opening using HW training with default settings
  - Months of continues error-free operation in corner conditions: silicon proven BER<1E-20
  - >20% supply voltage margin in the worst corner
  - >15% data rate margin in the worst corner
- **Power is slightly lower than expected 0.30 pJ/bit**
  - Power variation vs. corners is low
- **High yield, consistent behavior vs. corners, voltage and temperature**
  - In spite that some of the chips are out of process window
- **Full PVT silicon testing report is ready**

Typical Corner:
TT, AVDD=0.75V, AVDD12=1.2V, 25C



Worst Corner:
SS, AVDD=0.675V, AVDD12=1.08V, 125C



GLink 2.3LL Supports all Types of CoWoS & InFO_oS

- **GLink 2.3LL allows to achieve 7.6Tbps/3mm density (full duplex)**
  - It uses more aggressive 36 um bump pitch
  - It uses new wider physical bus
- **GLink 2.3LL supports the following platforms:**
  - CoWoS-S (silicon interposer) with 5xMi metal stack, 0.8/0.8um width/space routing rules
  - CoWoS-R (organic interposer) with 5 metal layers, 2/2um width/space routing rules
  - CoWoS-L (organic interposer with silicon bridge) with 5xMi metal stack, 0.8/0.8um width/space routing rules
  - InFO_oS (RDL) with 5 metal layers, 2/2um width/space routing rules

UCIe Adoption

- **GLink already supports most of UCIe-1.0 digital and analog functionality**
- **GLink to support UCIe: minor, low risk changes:**
  - UCIe bump map, Link Training, Sideband Channel (digital), FDI Interface, in-band CRC

Analog Functionality

UCIe-1.0	GLink
64-bit single ended data, diff. clock	The same
Data rates 4,8,12,16,24, 32 Gbps	16Gbps Si proven, 32Gbps in design
Latency 2ns TX and 2ns RX	End to end latency 5ns Including TX & RX async FIFOs
BER<1E-15	BER<1E-20
Power: 0.3 pJ/bit	Power: 0.3 pJ/bit
Channel: up to 2mm	Channel: up to 2mm
Bump map	To be adopted

Digital Functionality

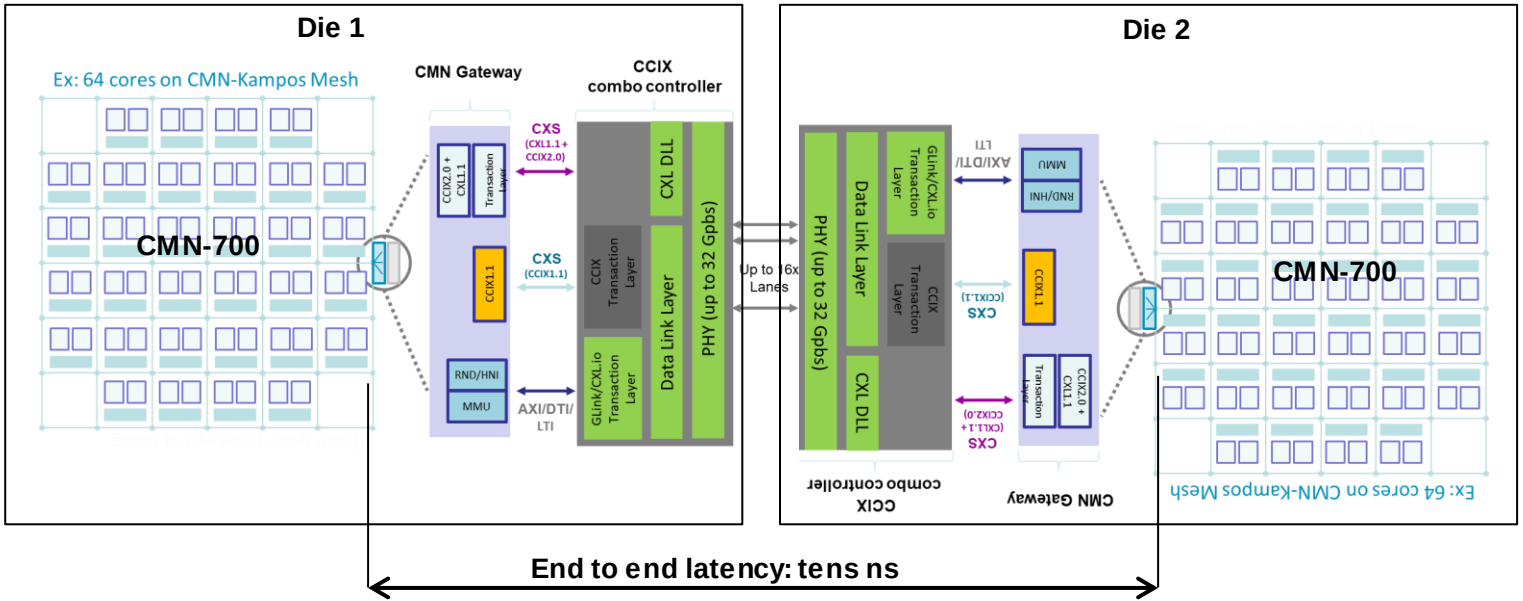
UCIe-1.0	GLink
CRC and Retry	Supported
Lane Repair	Supported
Scrambling/Descrambling	Supported
Async FIFOs	Supported
Link Training	To be adjusted
SideBand Channel	I/O – supported Logic – to be added
FDI Interface	Interface to be adjusted

Agenda

- GUC Introduction
- GLink-2.5D
- **GLink-2.5D in multi-die processors**
- GLink-3D

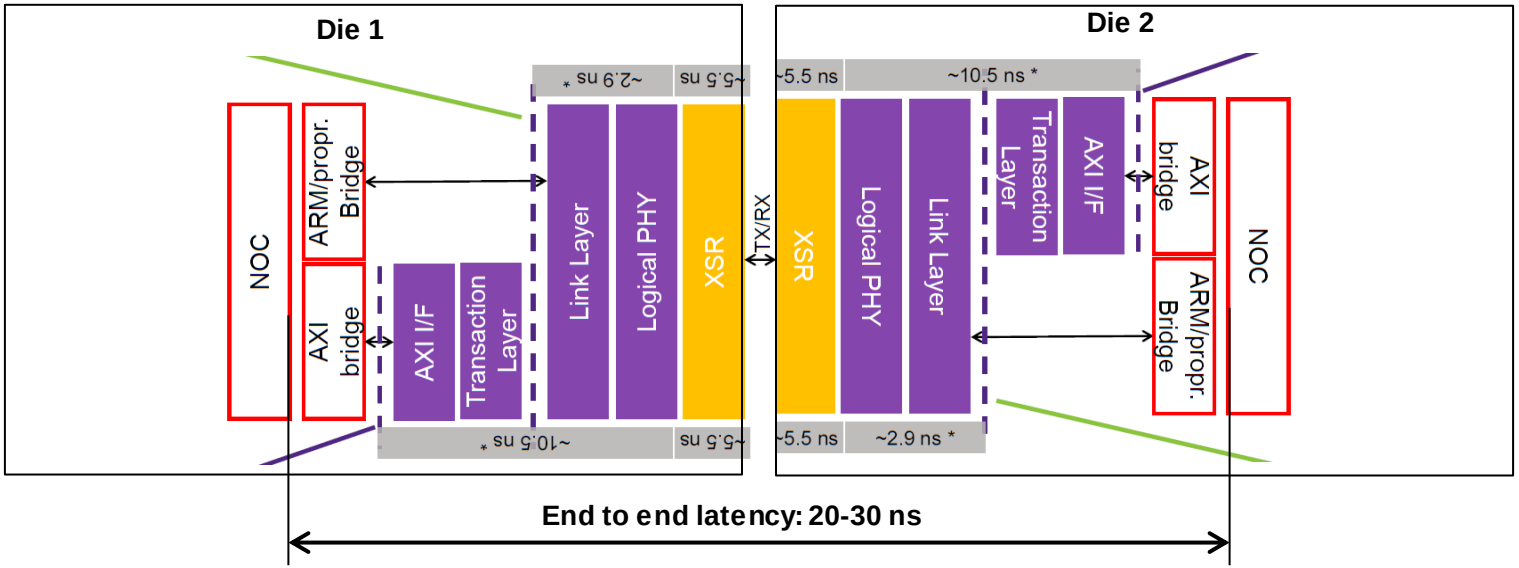
Traditional Solution: Multi-die Integration using CCIX/CXL

- Typical multi-die solution uses serial links with many layers between CMN meshes of two dies resulting in tens ns of end to end latency



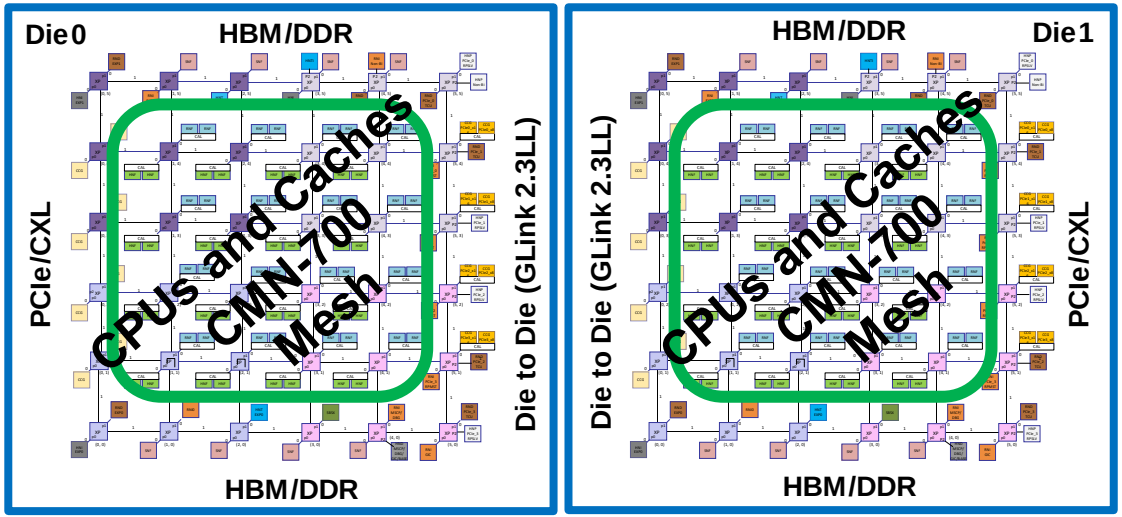
Traditional Protocol Stack

- See below specific example with 20-30 ns end to end latency



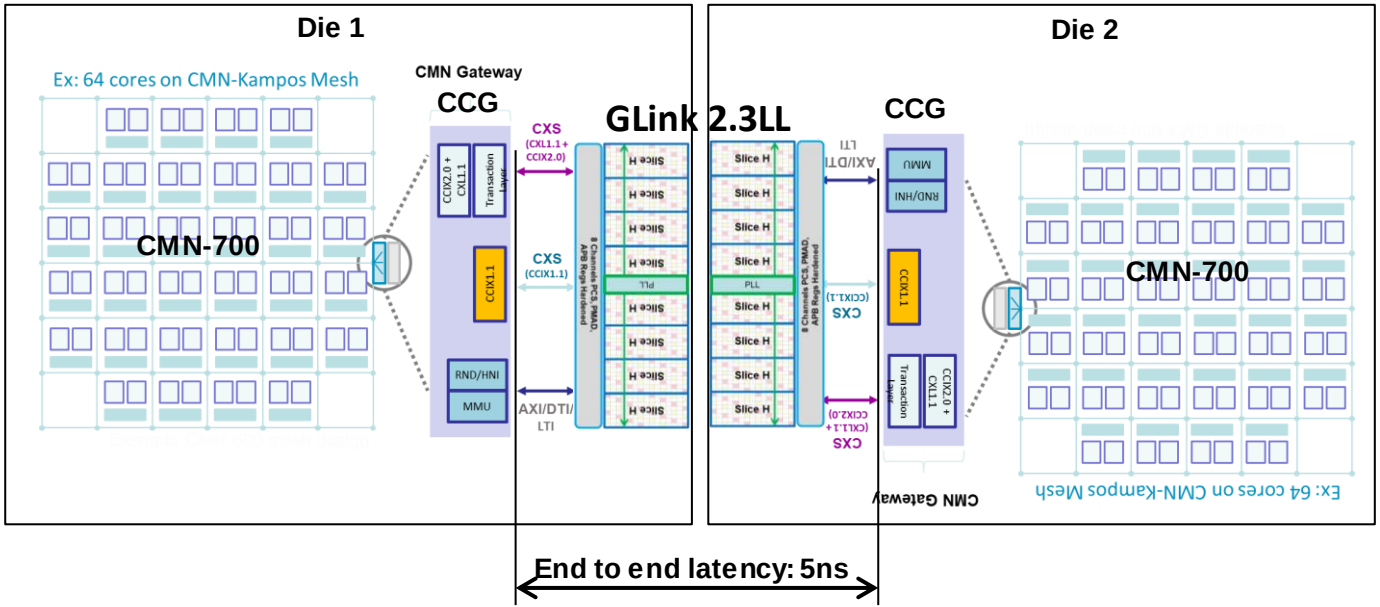
GLink 2.3LL for ARM Processors Mesh

- **High bandwidth, low latency GLink 2.3LL interface allows seamless integration of dies with ARM CMN-700 multi-CPU Mesh**
- **GLink 2.3LL:**
  - No errors, no data link and transaction layers
  - Look and feel as single die integration
  - Low latency connection
- **Very high GLink 2.3LL bandwidth allows to connect many mesh boundary Cross Points (XPs)**



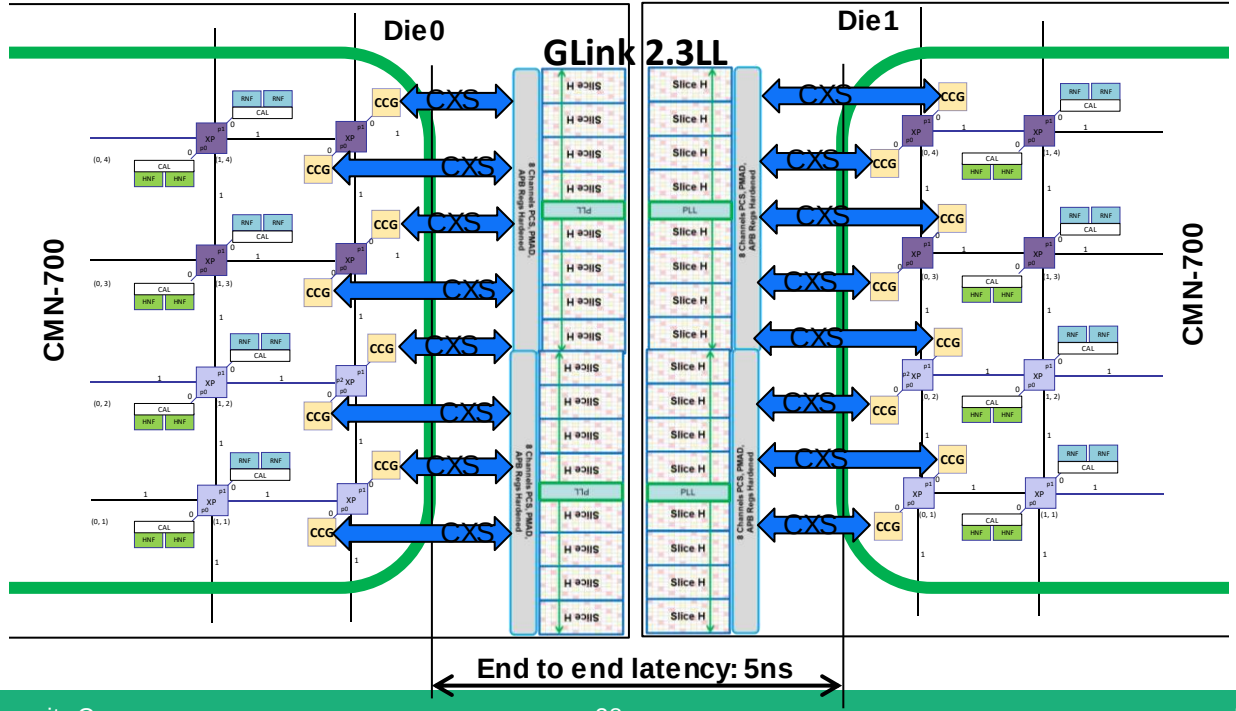
GLink 2.3LL Connection using CXS Bus

- GLink 2.3LL is used as CXS and AXI buses physical layer
- GLink 2.3LL allows 5 ns CXS to CXS end to end latency
- CMN Gateway (CCG) is used to connect GLink 2.3LL to Cross Point (XP)



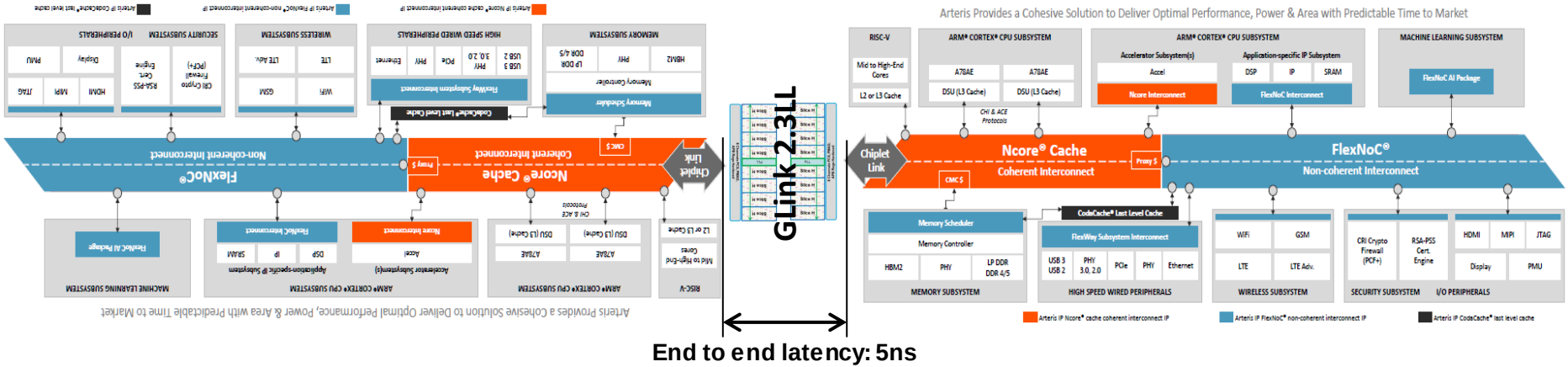
Multiple CXS Buses are connected using GLink 2.3LL

- High GLink 2.3LL bandwidth allows to connect multiple boundary XPs of two dies
- Boundary XPs use two CHI ports to connect to CCGs
- Every XP CHI port is connected through CCG to 512-bit@2GHz CXS bus
- GLink 2.3LL is used as CXS physical layer will 5ns CXS/Die0 to CXS/Die1 latency



GLink 2.3LL for Arteris Ncore Die to Die Interconnect

- **High bandwidth, low latency GLink 2.3LL interface allows seamless integration of dies Arteris Ncore**
- **GLink 2.3LL:**
  - No errors, no data link and transaction layers
  - Look and feel as single die integration
  - Low latency connection



Summary

- **GLink 2.3LL 17 Gbps achieves full duplex 2.5 Tbps/mm and 0.30 pJ/bit while using silicon-proven GLink 2.0 architecture**
- **GLink 3.3LL 32 Gbps achieves full duplex 5.1 Tbps/mm while keeping 0.30 pJ/bit**
- **GLinks support all types of 2.5D platforms InFO_oS, CoWoS-S/R/L**
- **Error-free die to die interconnect with 5ns end to end (AXI/CXS/CHI to AXI/CXS/CHI) latency**
- **GLink add-on IPs (GPIO, Bus Bridges, Multi-Slice PCS) enable seamless system integration**
  - Common AMBA bus bridges (AXI, CXS, CHI, etc...) are provided
  - Wide bandwidth connection of multiple dies
    - Connecting CMN-700 to CMN-700 in many XP points with 5ns XP to XP latency

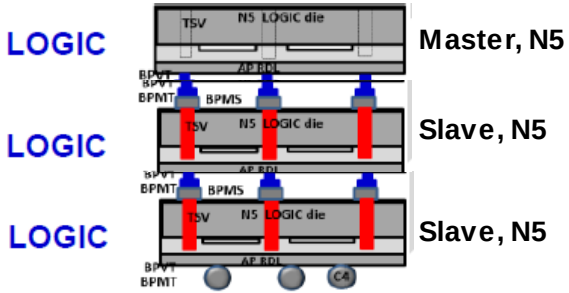
Agenda

- GUC Introduction
- GLink-2.5D
- GLink-2.5D in multi-die processors
- **GLink-3D**

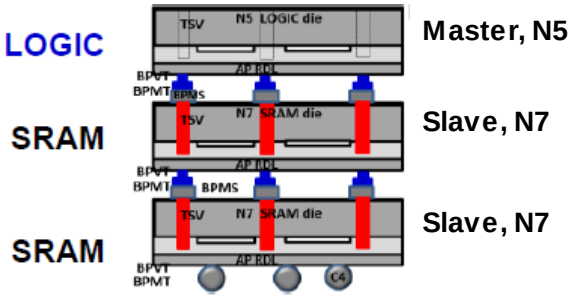
GLink-3D Applications

- **GLink-3D supports the following applications:**
  - Logic on SRAM: N5 Master Logic die on N7 Slave SRAM dies
    - Expandable SRAM application: Processor on top of SRAM dies
  - Logic on Logic: N5 Master Logic die on N5 Slave Logic dies
    - Expandable Compute application: computation units added to a Processor

Logic on Logic
Expandable Compute

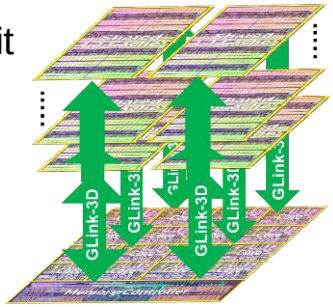


Logic on SRAM
Expandable SRAM



3D SRAM Comparison vs. HBM3

- **3D SRAM is a game changer for AI/HPC/Networking:**
  - Flexible amount of SRAM (GBytes!) are placed exactly where Processor needs it
- **3D SRAM comparison vs. HBM3:**
  - 1000x more bandwidth per 1 Mbyte than HBM3
  - 100x lower latency than HBM3
  - 20x less ASIC area than HBM3 Controller+PHY
  - 15x less power than HBM3
  - SRAM cost per bit is 10x higher than HBM DRAM
- **HBM3/DDR5 DRAM advantage: memory size and cost per bit**
- **3D SRAM advantage: bandwidth, latency, area and power**



	3D SRAM	HBM3
Stacked dies layers	Multiple	8 or 12 Hi
Bandwidth per 1 MByte	320 Gbps	0.27-0.4 Gbps
Read Latency	Constant 5ns	Variable 150ns-1000ns
Power efficiency	0.4 pJ/bit	5.8 pJ/bit
Processor area efficiency	9 Tbps/mm^2	0.4 Tbps/mm^2
Memory density	1.2 MBytes/mm^2	16.9 MBytes/mm^2
Max memory size (2x500mm^2)	4 GBytes	192 GBytes

Comparison of Die to Die Interfaces

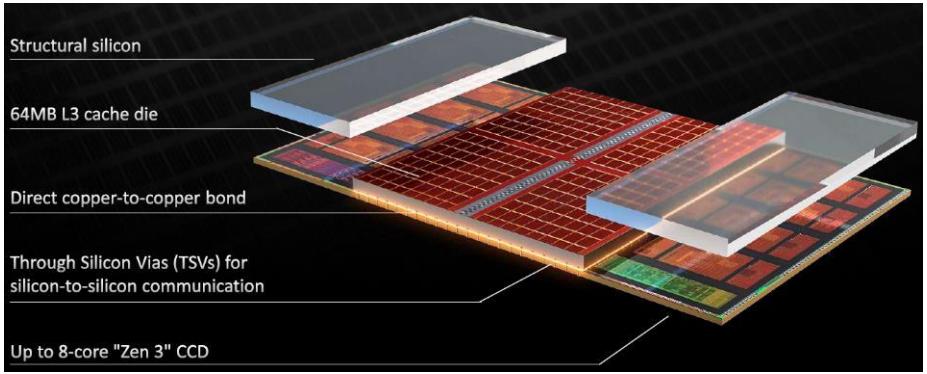
- **GLink interfaces data transmission is reliable: no errors, no error correction - low latency**
- **GLink-2.5D interface over CoWoS or InFO_oS allows ~5x lower power and ~3x higher bandwidth density**
- **GLink-3D interface over SoIC further reduces power, increases bandwidth density and reduces latency by order of magnitude**
- **XSR IPs require additional Link and Transaction Layers and FEC – additional area/power/latency**
- **GLink IP includes all analog and digital layers, it can be connected directly to AXI, CXS, CHI or any other bus**

	112G-XSR*	GLink-2.5D	GLink-3D 1.0	GLink-3D 2.0LL
Connectivity	Substrate	CoWoS/InFO_OS	SoIC	SoIC
Bit Error Rate	1E-7…1E-9	Error-free	Error-free	Error-free
Power efficiency	1.5 pJ/bit	0.30 pJ/bit	<0.2 pJ/bit	0.1-0.05 pJ/bit
Beachfront efficiency	0.8 Tbps/mm	2.5 Tbps/mm	NA	NA
Area efficiency	0.7 Tbps/mm^2	2.3 Tbps/mm^2	9 Tbps/mm^2	8-20 Tbps/mm^2
End to end latency	10ns	<10ns	1-2ns	<500ps

* Only Serdes, not including Link and Transaction Layers and FEC

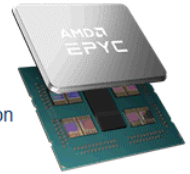
AMD 3DIC Products

AMD V-Cache Technology



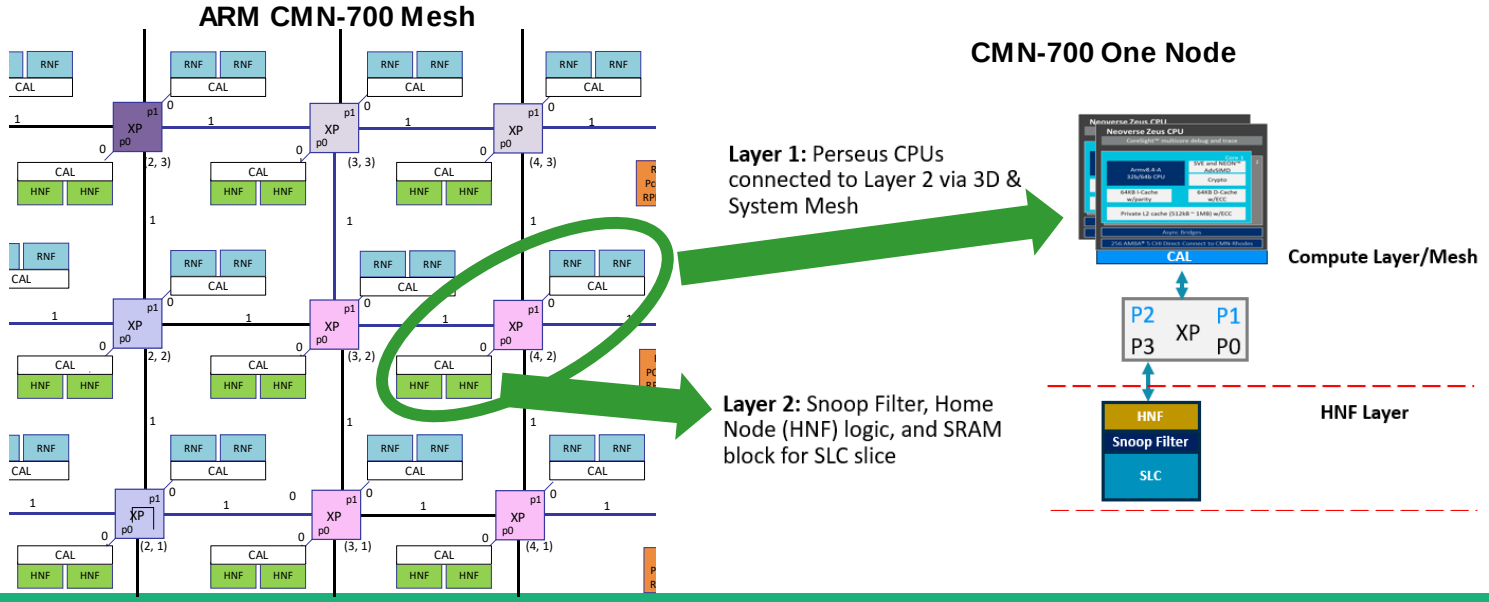
AMD V-Cache Based Processors

- **Milan-X(EPYC server CPU, 64 cores, 768MB L3)**
  - L3: 256MB → 804MB
  - Performance enhancement
    - ✓ 50% uplift across targeted technical computing workloads
    - ✓ 66% ↑ for Synopsys VCS functional verification



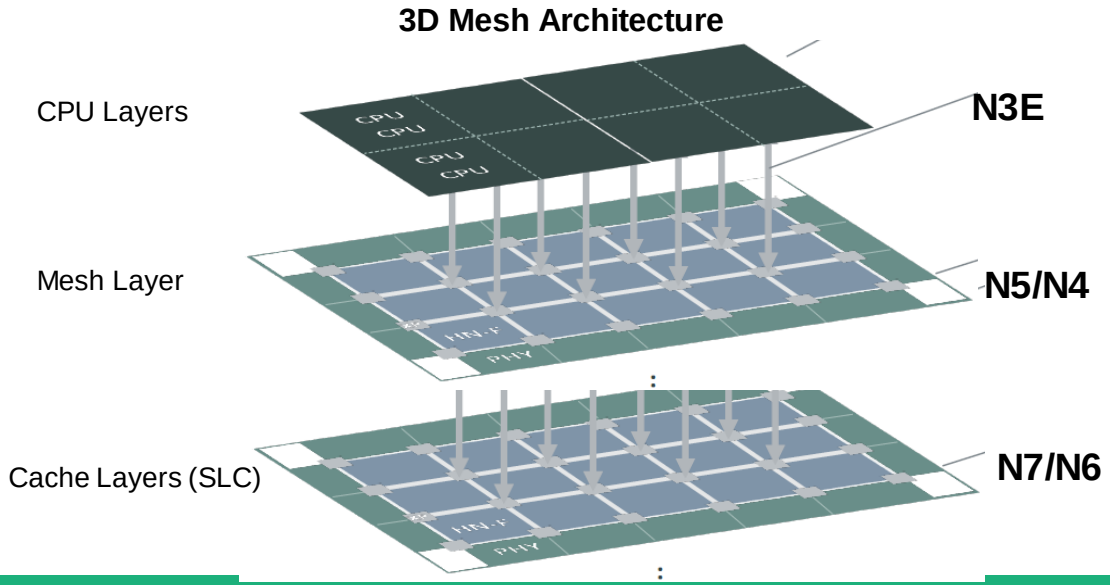
ARM CMN-700 Mesh Cross-Point (XP) Ports

- **ARM CMN-700 Mesh is built of identical Cross-Points (XP)**
- **Every XP has:**
  - 4 interfaces connected to adjustment XPs
  - One Port connected to a dual CPU CAL port (RN-H)
  - One Port connected to System Layer Cache and Snoop Filter (HN-F)



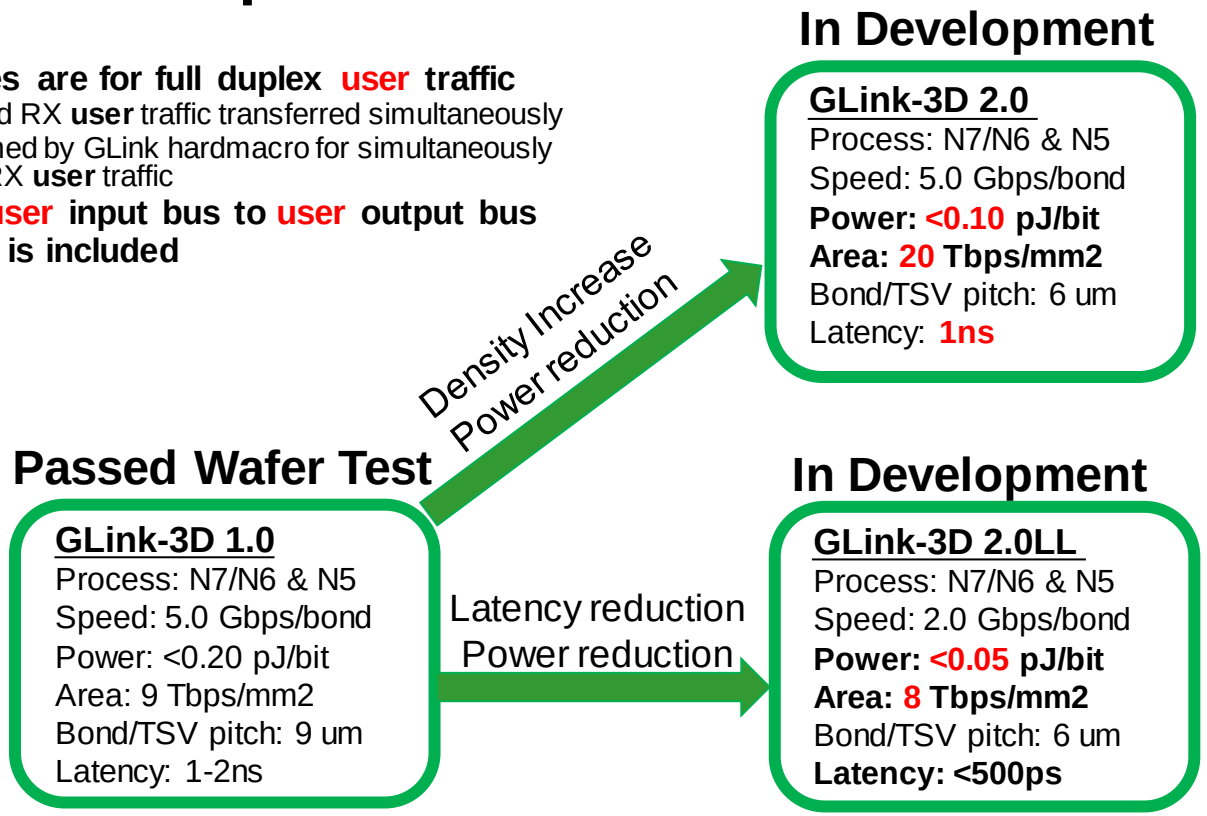
3D Architecture Summary

- **3DIC allows CPUs and Cache (SLC) disintegration from Mesh**
  - Implementation of CPUs, Mesh and Cache, each in the most optimal technology node
- **Increasing Cache and Cores without increasing mesh distance and latency**
- **Simulated performance gain: 25%-50%**



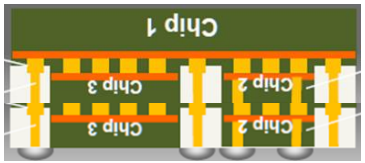
GLink-3D IP Roadmap

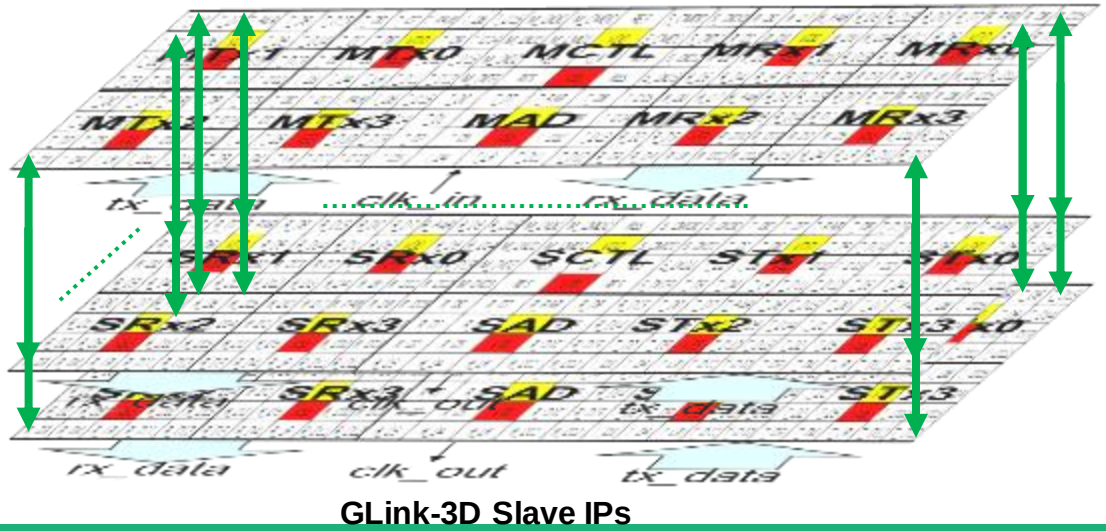
- **Throughput and power values are for full duplex user traffic**
  - 1 Tbps means 1 Tbps of TX and RX **user** traffic transferred simultaneously
  - 0.1 pJ/bit means 0.1W consumed by GLink hardmacro for simultaneously transferring 1 Tbps of TX and RX **user** traffic
- **Latency is end to end from user input bus to user output bus**
- **Both digital and analog area is included**



GLink-3D Physical Connection

- GLink-3D Master and Slave bond maps are built in the way so that T_Dn bits will perfectly match R_Dn
- Multiple Slaves can be attached over/under a Master
- TSVs and Bonds are used to connect Master's and all corresponding Slaves' pins to a single physical net





GLink-3D as AMBA CHI Physical Layer

- AMBA CHI Interconnect and Node TX and RX buses are mapped to GLink-3D Master and Slave IP
- GLink-3D adjusts data widths to AMBA CHI widths

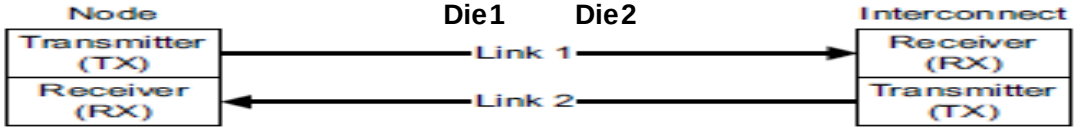
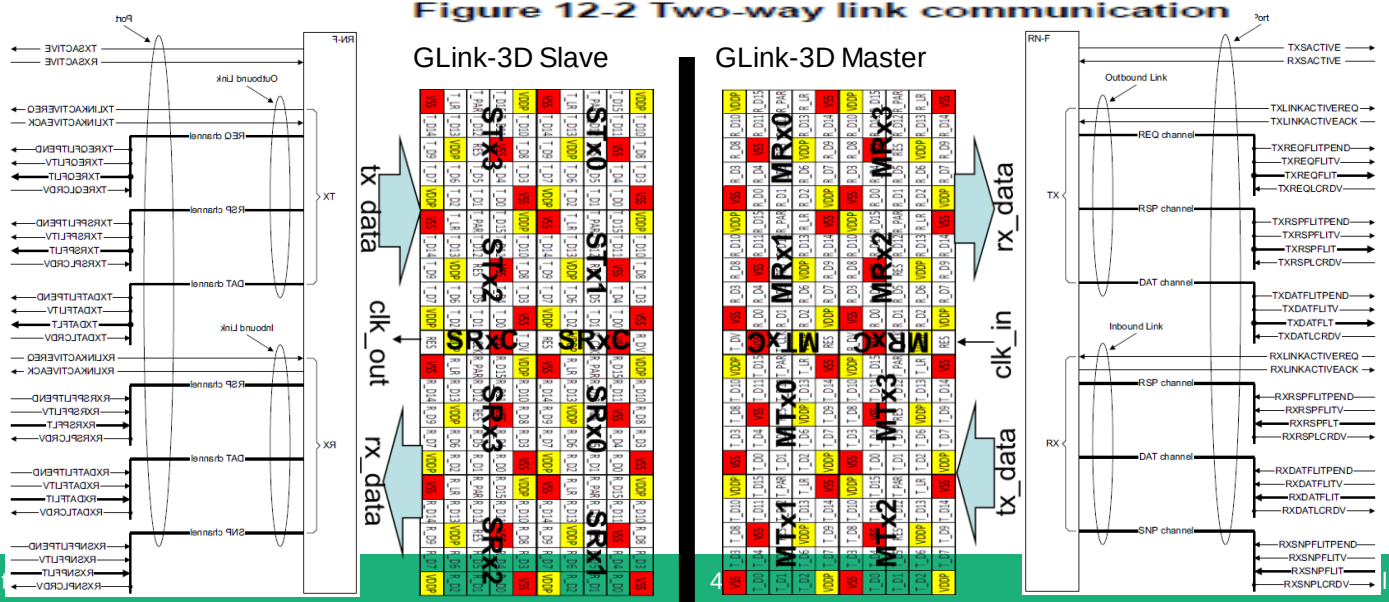


Figure 12-2 Two-way link communication

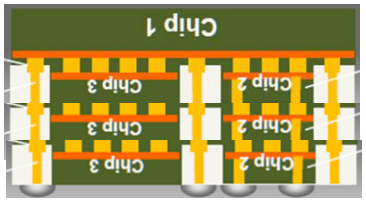


SoIC Assembly Methods

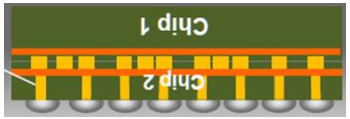
- GLink-3D supports both CoW and WoW assembly
- Amount of Slave dies attached to the Master die varies according to application
- SRAM address space can be increased by assembling more Slave dies

SoIC Assembly Methods

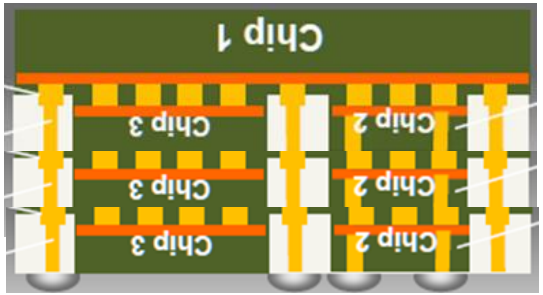
CoW



WoW



Variable amount of Slave dies attached to Master die



Master Die, assembled face down

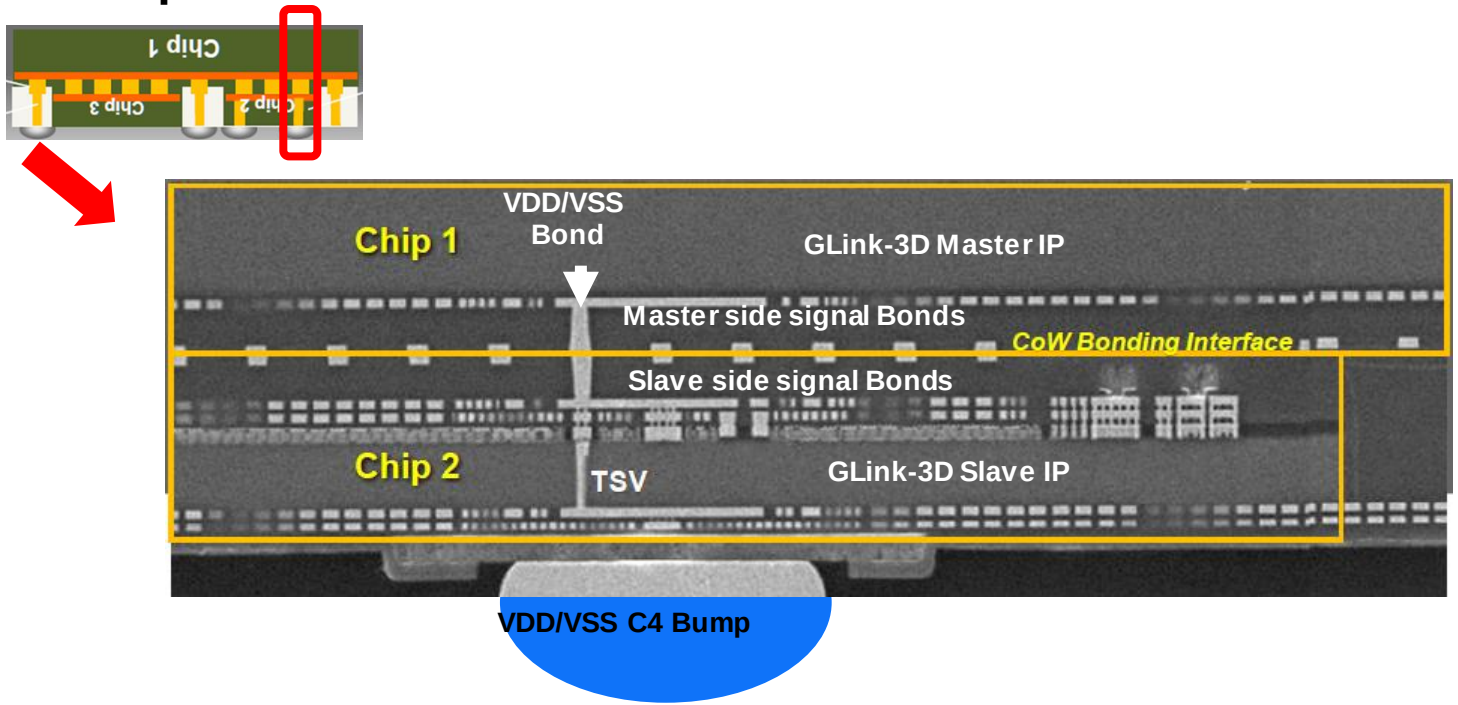
Slave Die 0, assembled face to face to Master

Slave Die 1, assembled face to back to Slave Die 0

Slave Die 2, assembled face to back to Slave Die 1

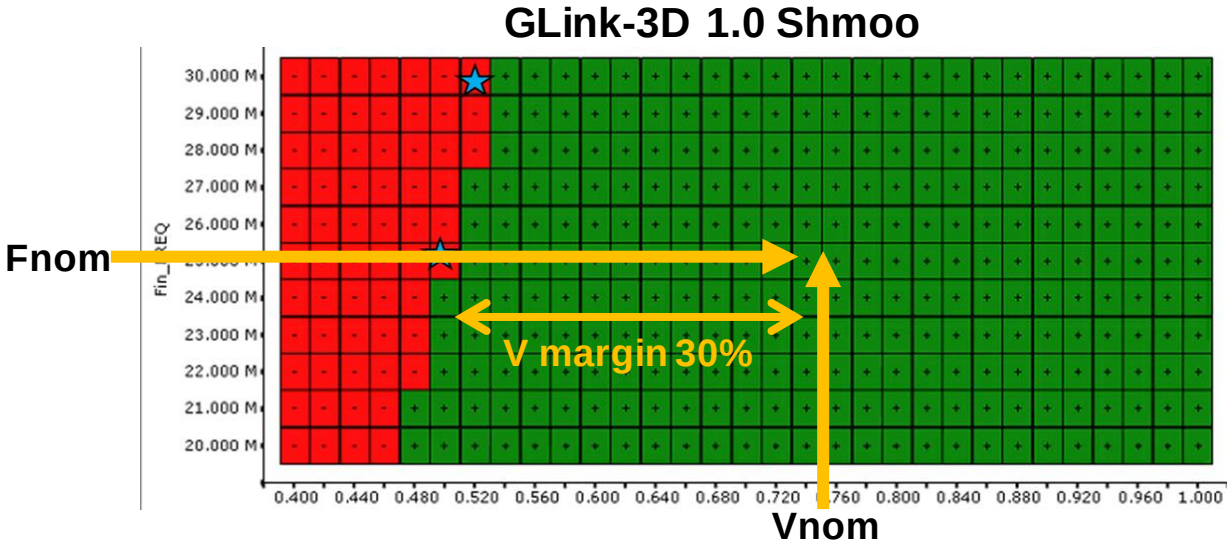
GLink-3D IP Includes TSV

- VDD/VSS TSVs connect VDD/VSS bonds of GLink-3D IP to VDD/VSS C4 bumps



GLink-3D 1.0 Testchip Silicon Testing Results

- **Wafer test results:**
  - Robust functional operation of all GLink-3D interfaces in all process corners
  - Yield: 99.3%
  - 30% supply voltage margin



GLink-3D Summary

- **GLink-3D IP family supports all types of TSMC 3D SoIC platforms: both CoW and WoW, both Face to Face and Face to Back**
  - Point to Multi-point architecture allows a Master die to interface several Slave dies using a shared interface
- **1st gen silicon-proven GLink-3D 1.0 achieves full duplex 9 Tbps/mm² and <0.20 pJ/bit**
- **2nd gen GLink-3D 2.0 targets to increase bandwidth to 20 Tbps/mm² and twice reduce power while using proven DDR architecture**
- **GLink-3D 2.0LL version targets to reduce end to end latency to <500ps and further to reduce power**

GUC
The Advanced ASIC Leader

Presentation Disclaimer

The content may only represent current status and are subject to change without prior notice.


GUC
The Advanced ASIC Leader
Thank You
for your attention