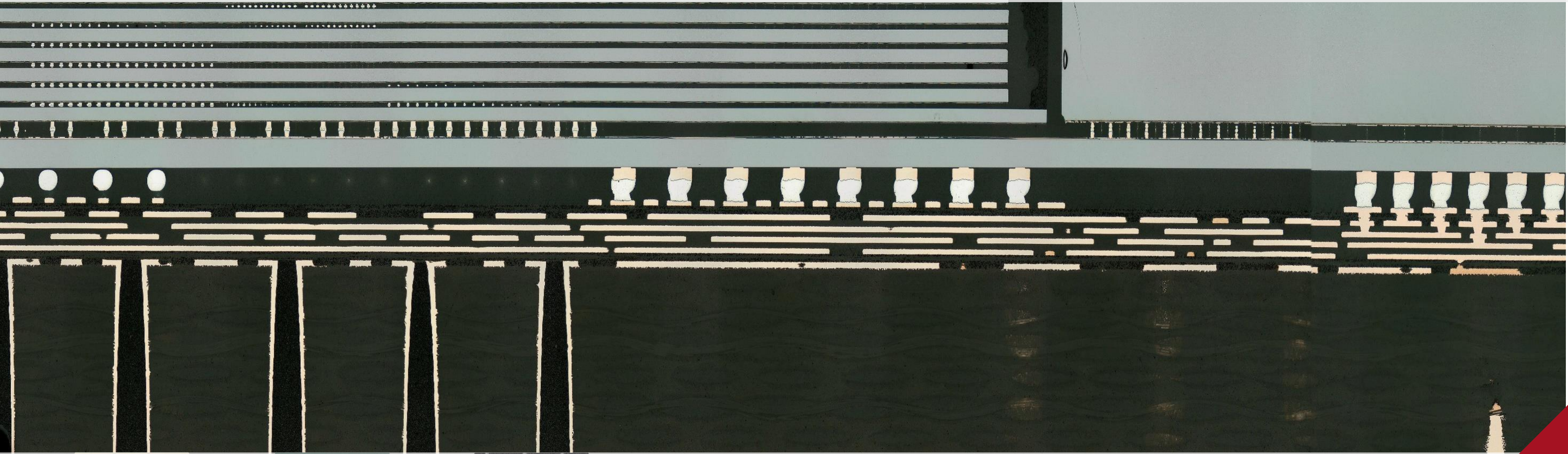
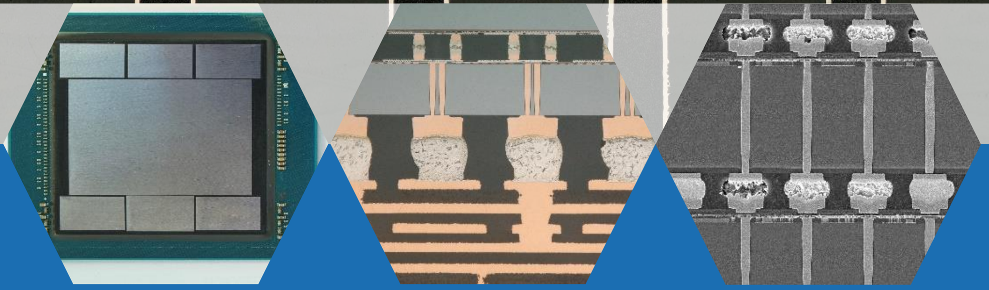




NVIDIA A100 Ampere GPU

NVIDIA's new generation Graphics Processing Unit (GPU) wit
CoWoS, 40GB Samsung HBM2, 2.5D and 3D packaging.

SP20579 - Advanced Packaging report by Belinda Dube
Laboratory Analysis by Véronique Le Troadec

February 2021 – Sample

SAMPLE

22 Bd Benoni Goullin
44200 NANTES - FRANCE

+33 2 40 18 09 16

info@systemplus.fr

systemplus.fr

Table of Contents

OVERVIEW

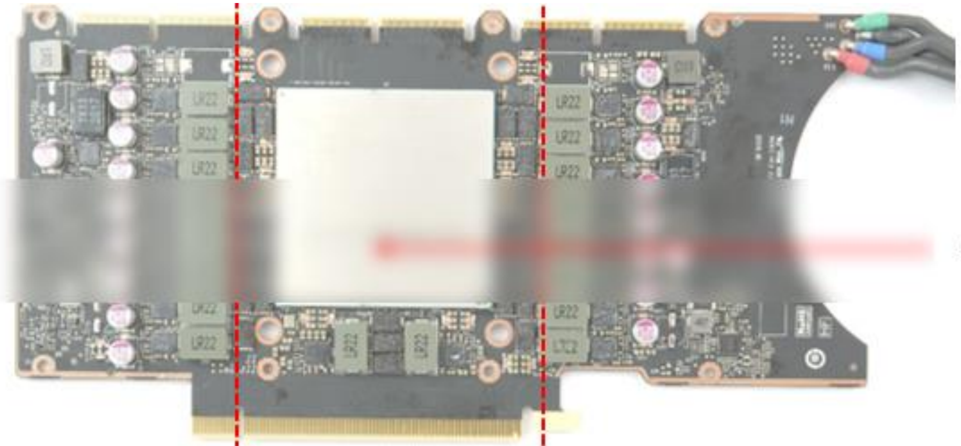
METHODOLOGY

Executive Summary

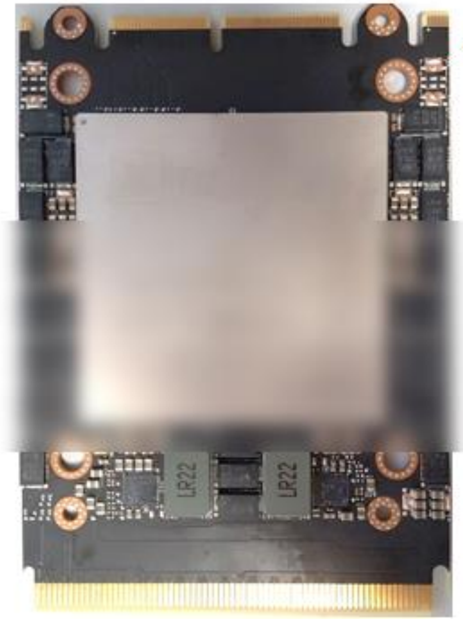
- The high end electronic packaging market was worth more than $880 million dollars in 2019. The biggest market for high end performance packaging comes from telecom and infrastructure. It has more than a 50% market share according to Yole Développement's report High-End Performance Packaging: 3D/2.5D Integration 2020.
- The NVIDIA Ampere A100 targets high performance data centers, artificial intelligence applications, data analytics, and High-Performance Computing (HPC). It uses advanced technologies, including TSMC's 7nm FinFET chip, 3D stacked memory with 2.5D integration on a silicon interposer in a Chip-on-Wafer-on-Substrate (CoWoS) process.
- The new generation GPU provides significantly higher performance compared to the previous generation. The NVIDIA Ampere A100 provides the consumer market with 40GB or 80GB of Samsung's high bandwidth memory (HBM2). Samsung's HBM2 is engineered to sustain and support high speed data transfer. The HBM2 DRAM solution satisfies the market need for high performance, energy efficiency, and compact integration. The NVIDIA Ampere A100 is characterized by GPU Memory Bandwidth of 1,555 GB/s. A 3D assembly process yields HBM2 stacks composed of eight 1GB memory dies and one logic die, connected with via-middle through-silicon vias and micro-bumps.
- More than 6,000mm^2 of silicon area is integrated in a single 55mm x 55mm 12-layer ball grid array (BGA) package of the NVIDIA Ampere A100. Two major semiconductor leaders, Samsung and TSMC, collaborate to deliver this much silicon in a single package. TSMC is the main provider for the NVIDIA Ampere A100. Using its 2.5D CoWoS platform, it manufactures the world's largest processor built on 7nm process technology. This GPU die features a 7nm FinFET transistor process and 54 billion transistors on a single chip. It also produces a large silicon interposer on which the GPU and HBM2 memory is assembled at the wafer-level.
- The report includes a complete physical analysis of the package, the GPU die, interposer die and the HBM2 DRAM. Along with the manufacturing process of the silicon dies, CoWoS process and final assembly, this report comes with a cost analysis and a price estimation of the NVIDIA Ampere A100. Finally, the report includes a comparison to highlight the similarities and differences between the NVIDIA Ampere A100 and NVIDIA's Tesla P100 and V100.

PHYSICAL ANALYSIS

Graphic Card Teardown



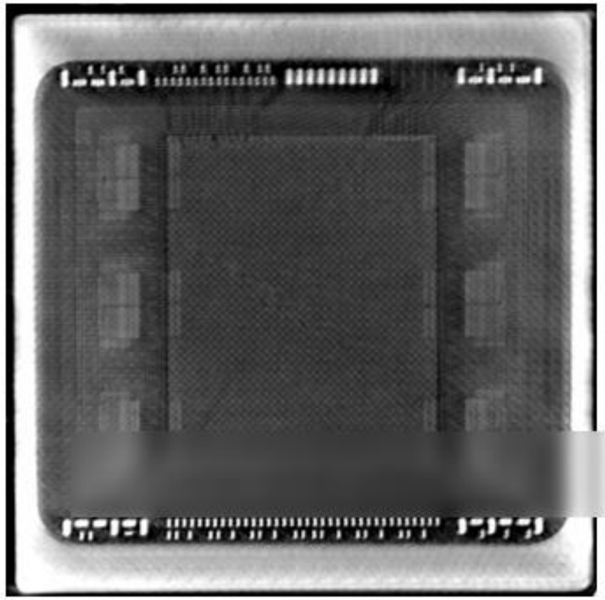
NVIDIA Ampere A100 Teardown
©2021 by System Plus Consulting



NVIDIA Ampere A100 Teardown – PCB Front Side
©2021 by System Plus Consulting

NVIDIA Ampere A100 Teardown – PCB Back Side
©2021 by System Plus Consulting

NVIDIA AMPERE 100 Package - XRAY Images-

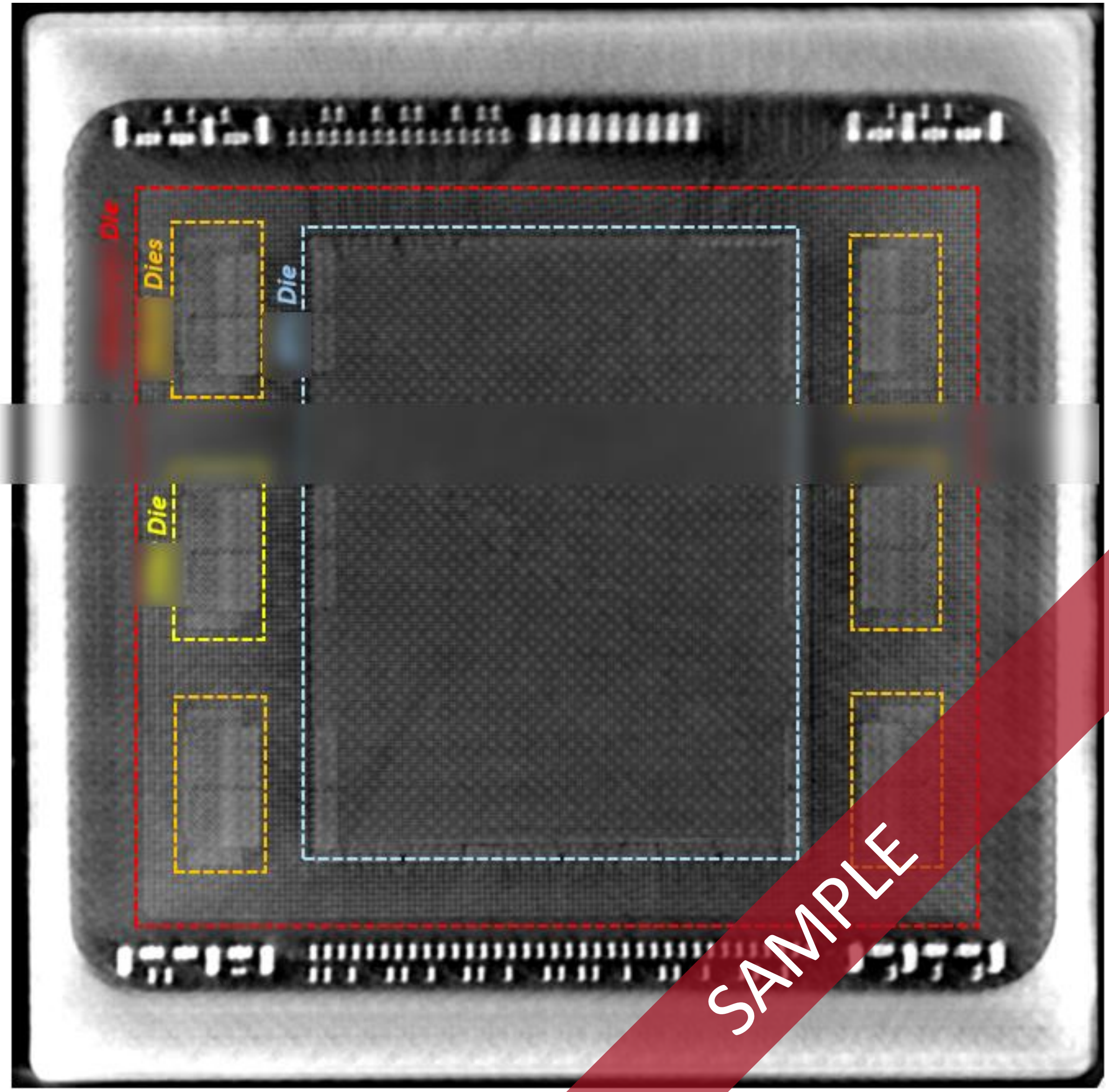


NVIDIA Ampere A100 X-Ray Image

©2021 by System Plus Consulting

The Nvidia Ampere 100 package integrates:

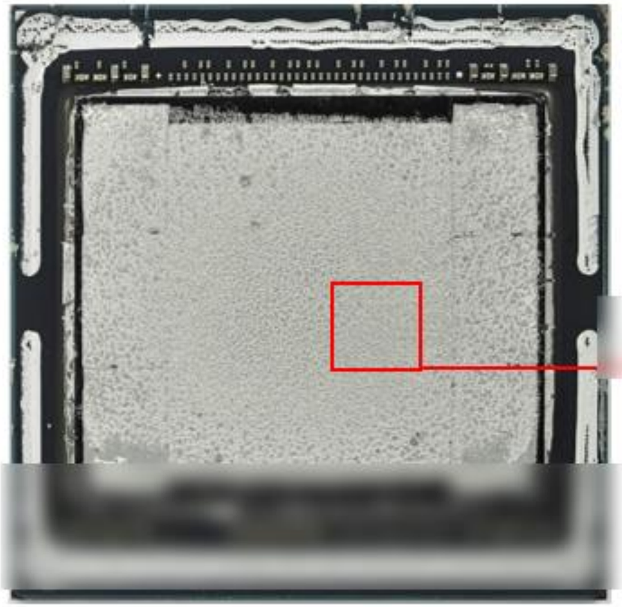
- Interposer die
- GPU Die on die
- HBM memory packages assembled on die
- Filler Die



NVIDIA Ampere A100 X-Ray Image

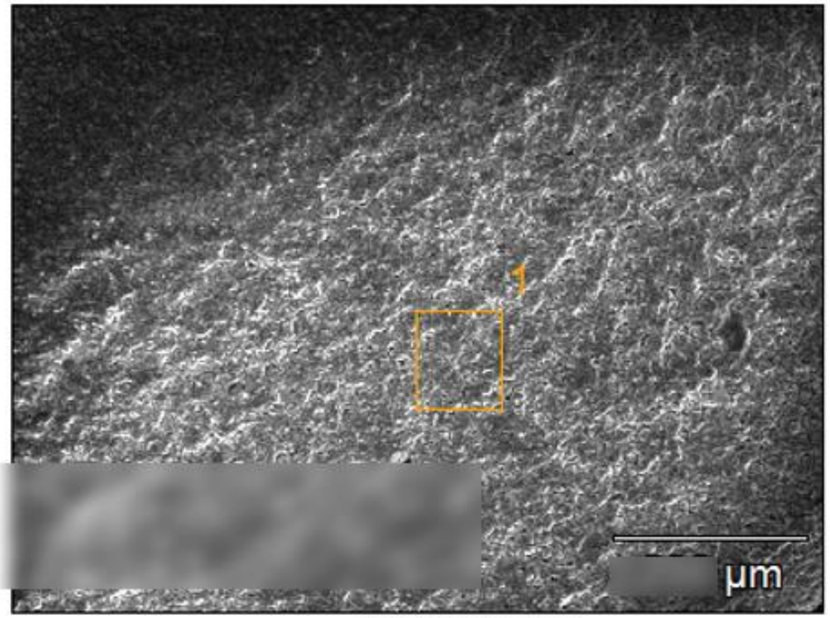
©2021 by System Plus Consulting

Package Opening



Package Opening

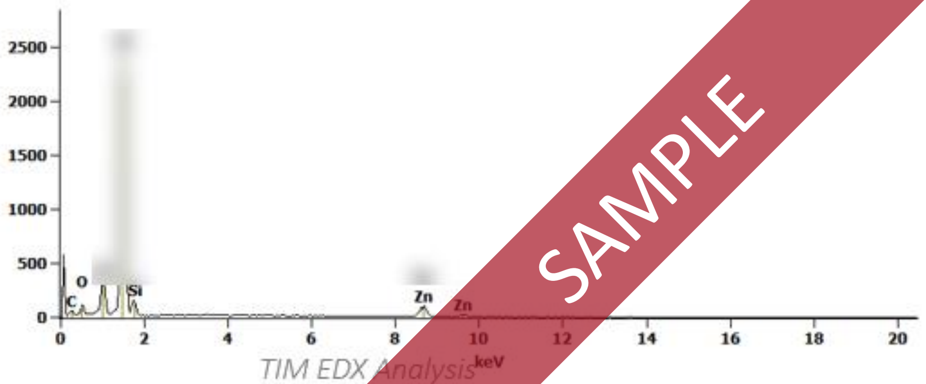
©2021 by System Plus Consulting



TIM EDX Analysis

©2021 by System Plus Consulting

- A [illegible] is applied between the [illegible] memory and the top metal frame to ensure that there are so [illegible]
- The [illegible] helps [illegible] to avoid [illegible] of the active components.
- The [illegible] is made of an [illegible] compound.

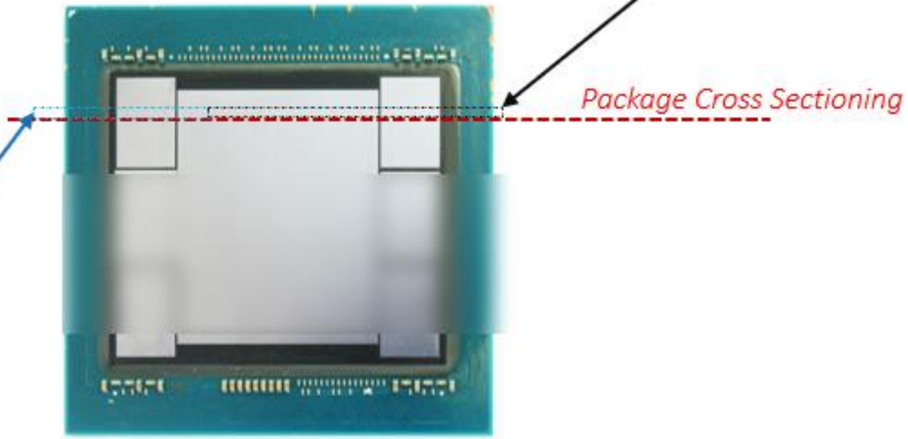


TIM EDX Analysis

©2021 by System Plus Consulting

Package Cross Section

Package Cross Section
©2021 by System Plus Consulting

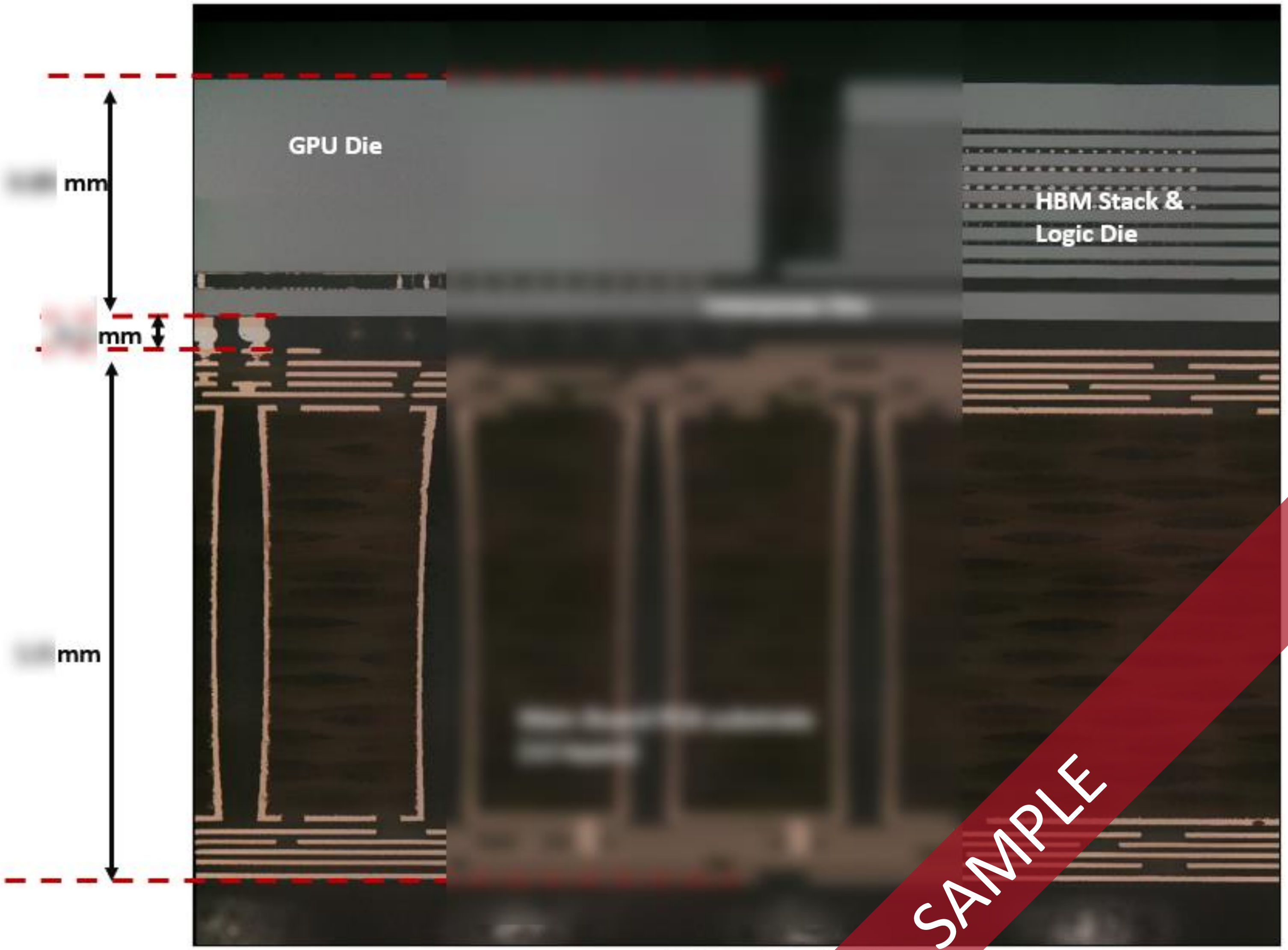


Package Opening Top View
©2021 by System Plus Consulting

Package Cross Section
©2021 by System Plus Consulting

Board Cross-Section

Package Opening Top View
©2021 by System Plus Consulting



Package total mm

Package Cross Section
©2021 by System Plus Consulting

[Overview / Introduction](#)

[Company Profile & Supply Chain](#)

[Physical Analysis](#)
- Summary
- Graphic Card View & Teardown
- Package X-Ray Images
- Package Overview
- Package Opening
- Package Cross Section
- HBM Stack-Logic/DRAM Die
- HBM Stack Cross Section
- GPU Die
- Cross-Section – GPU
- Filler Die
- Cross-Section – Filler
- Interposer Die
- Cross-Section – Interposer

[Manufacturing Process Flow](#)

[Cost Analysis](#)

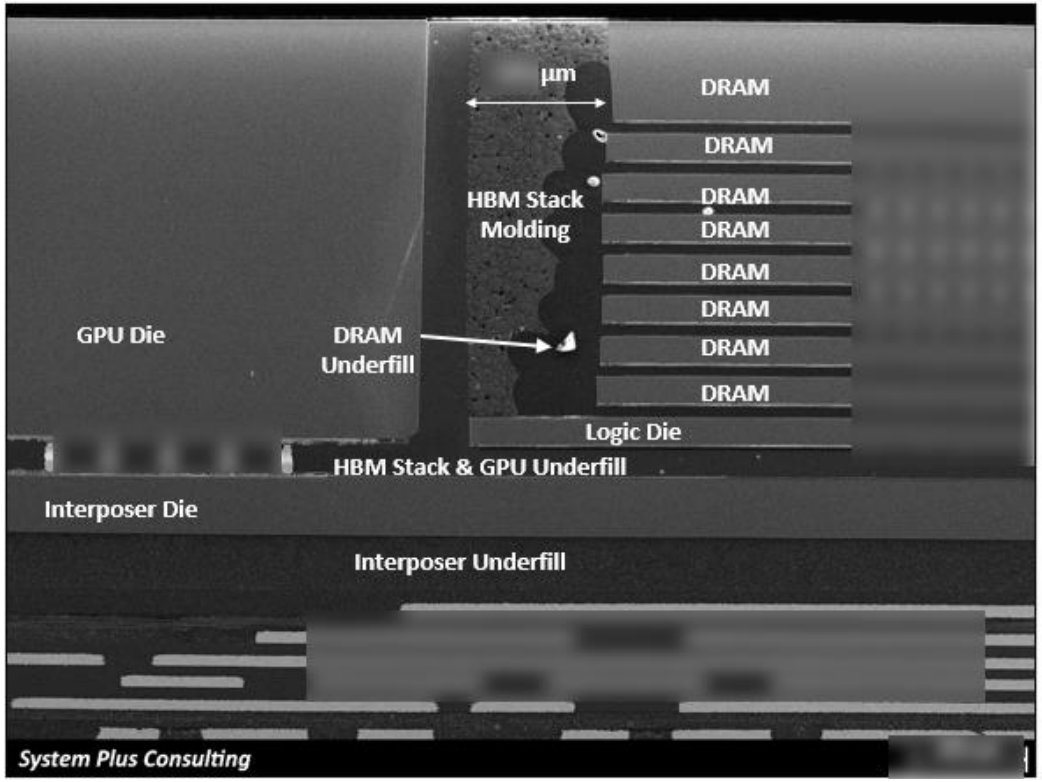
[Selling Price Analysis](#)

[Feedback](#)

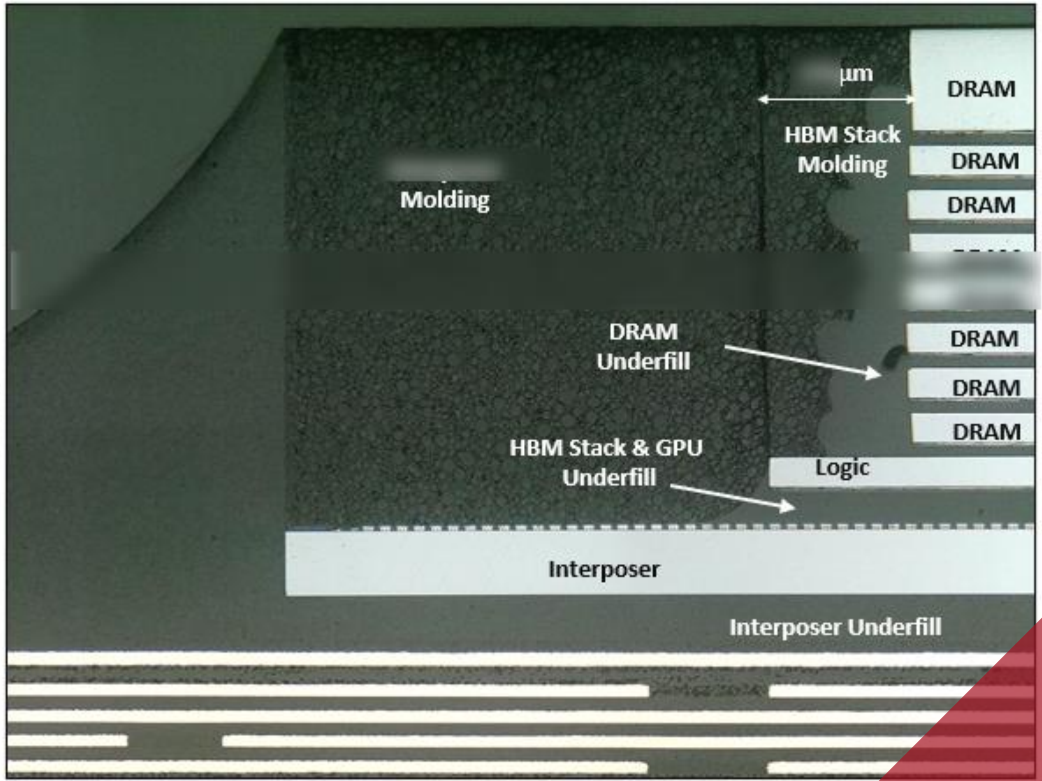
[Related Reports](#)

[About System Plus](#)

Package Cross-Section – HBM2 Stack



HBM Stack Cross-Section – SEM View
©2021 by System Plus Consulting

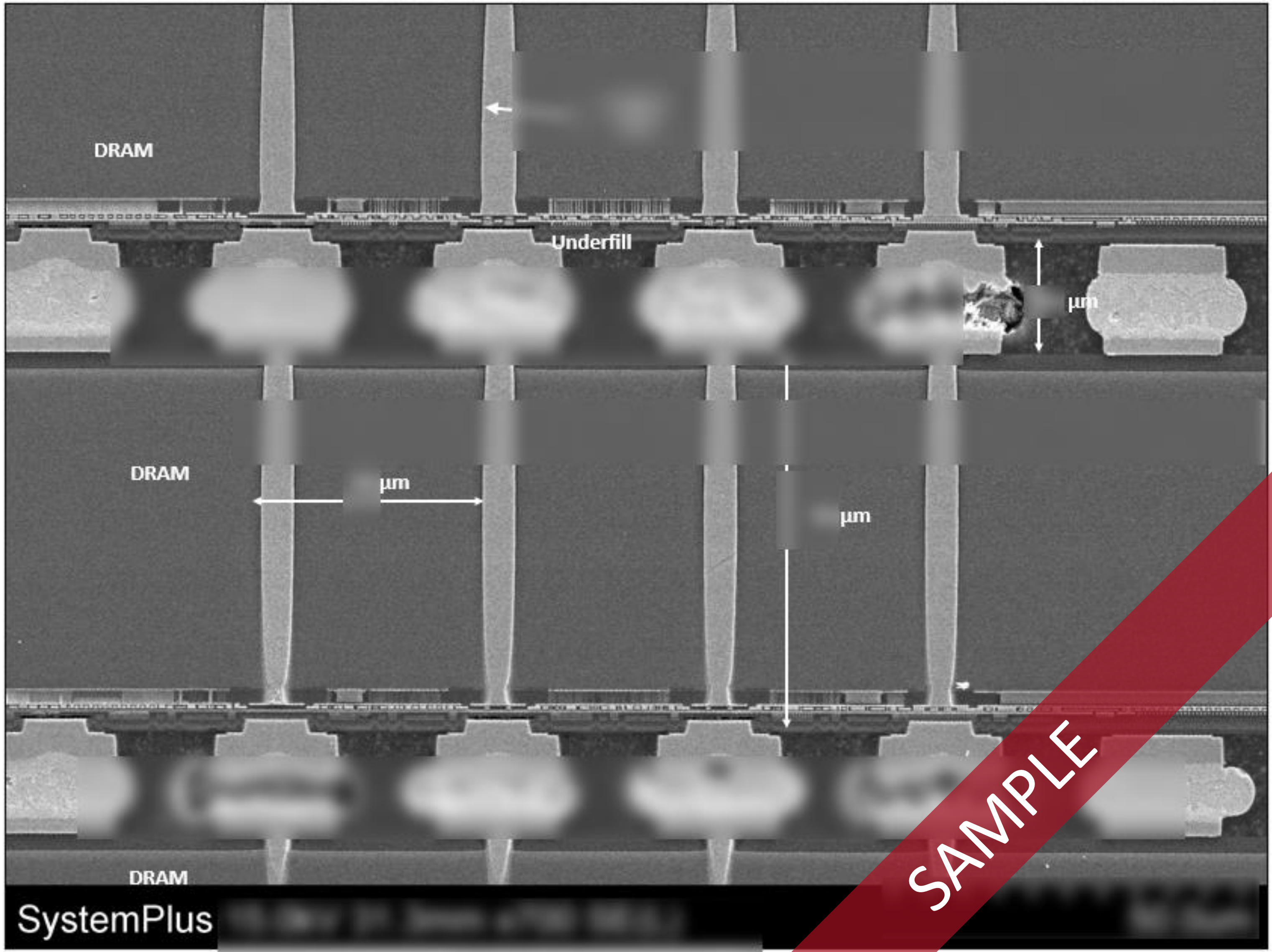


HBM Stack Cross-Section – Optical View
©2021 by System Plus Consulting

- The HBM stack is molded on the side.
- The side mold is 225-235µm wide.
- DRAM dies do not have the same exact size, they are diced before being bonded together.
- The top DRAM Die is thicker is size compared to the other DRAM dies. This could be a technic to provide mechanical stability and protect the DRAM dies from fracture.

Package Cross-Section – HBM2 Stack

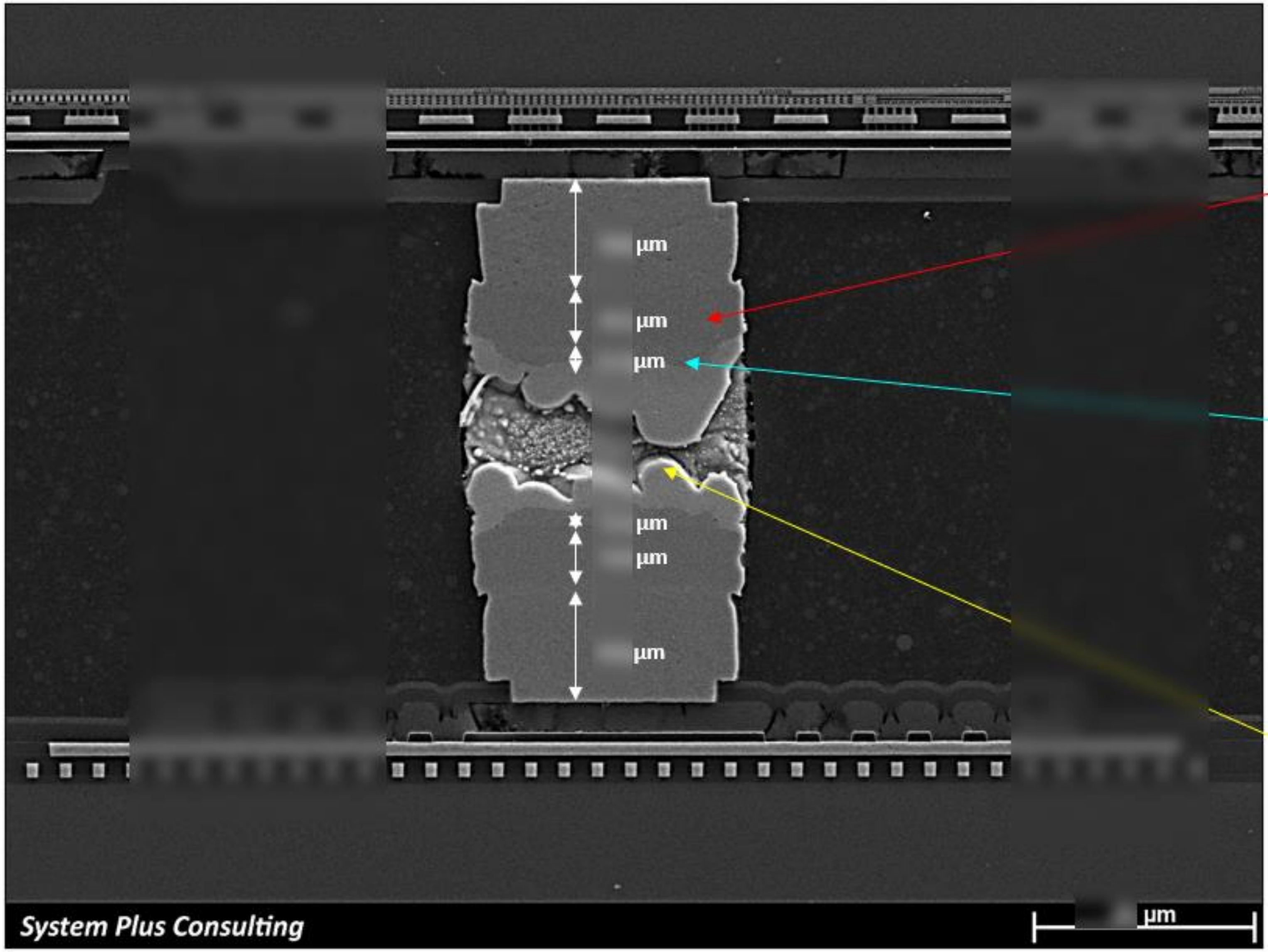
- HBM die thickness(Si substrate, capacitors & top metal layers) (except top die): µm
- HBM stack TSV & micro-bumps pitch: µm
- Underfill thickness: µm



HBM Stack Cross-Section – SEM View
©2021 by System Plus Consulting

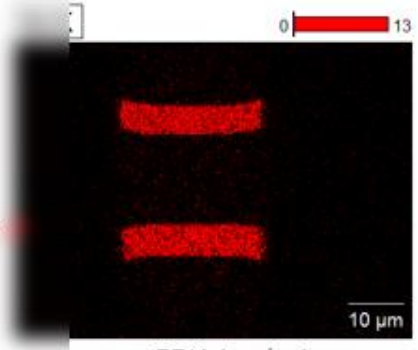
SYSTEMPlus CONSULTING

Package Cross-Section – GPU – Interposer Bumps



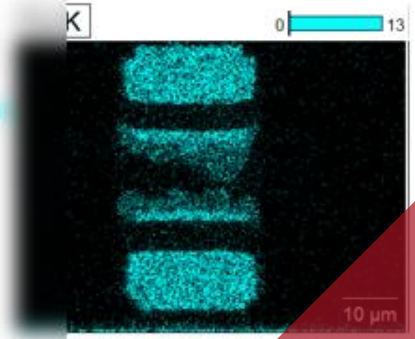
Interposer Cross-Section – SEM View

©2021 by System Plus Consulting



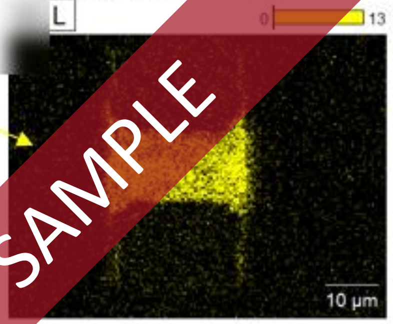
EDX Analysis

©2021 by System Plus Consulting



EDX Analysis

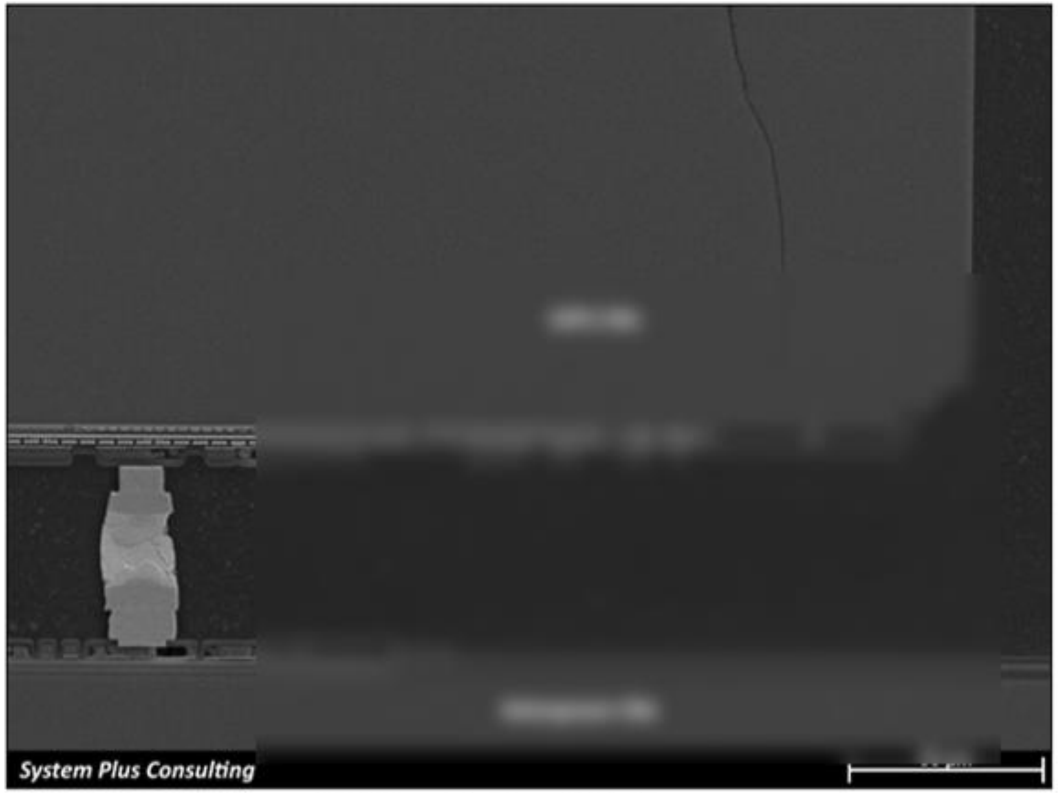
©2021 by System Plus Consulting



EDX Analysis

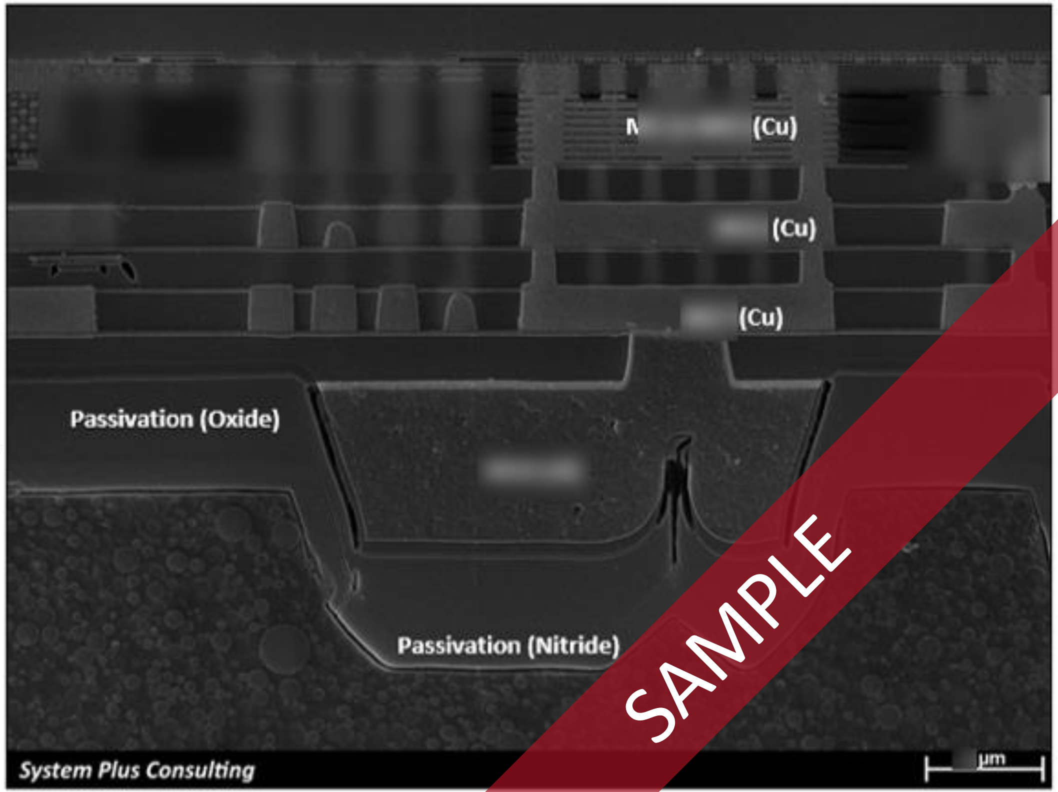
©2021 by System Plus Consulting

Die Cross-Section – GPU



GPU Cross-Section – SEM View

- The GPU Die process uses metal layers. (Cu + 1 Al)



GPU Cross-Section – SEM View

Package Cross-Section – Interposer (under GPU Die)

Interposer Cross-Section – SEM View
©2021 by System Plus Consulting

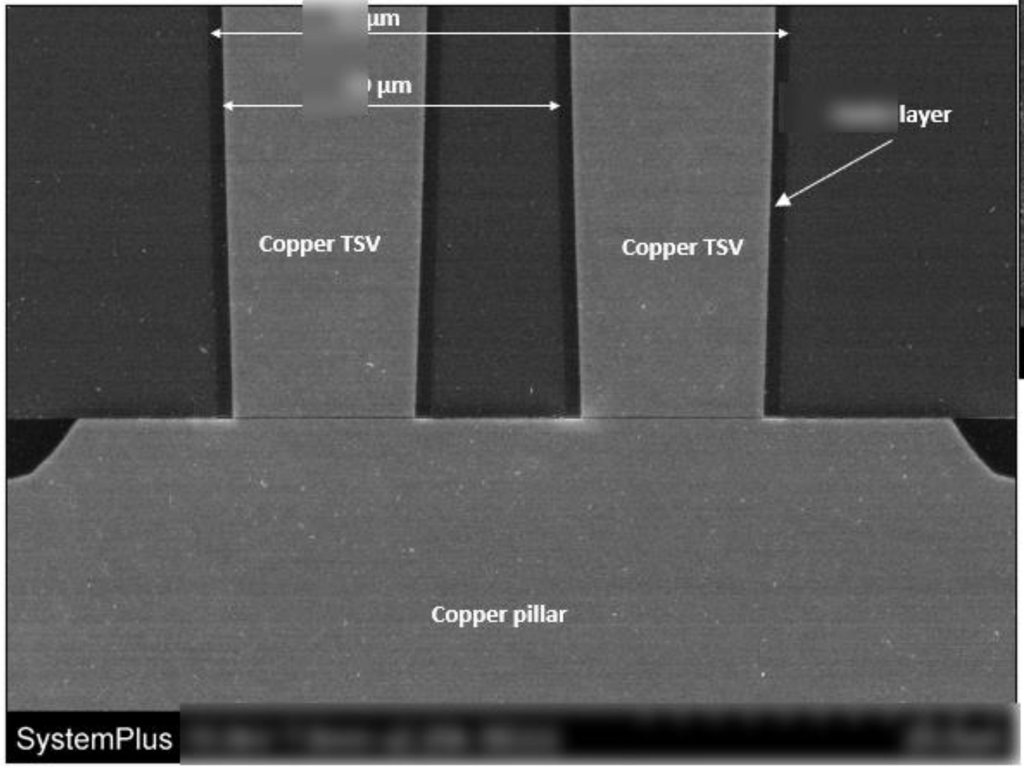
Interposer Cross-Section – SEM View
©2021 by System Plus Consulting

Interposer Cross-Section – SEM View
©2021 by System Plus Consulting

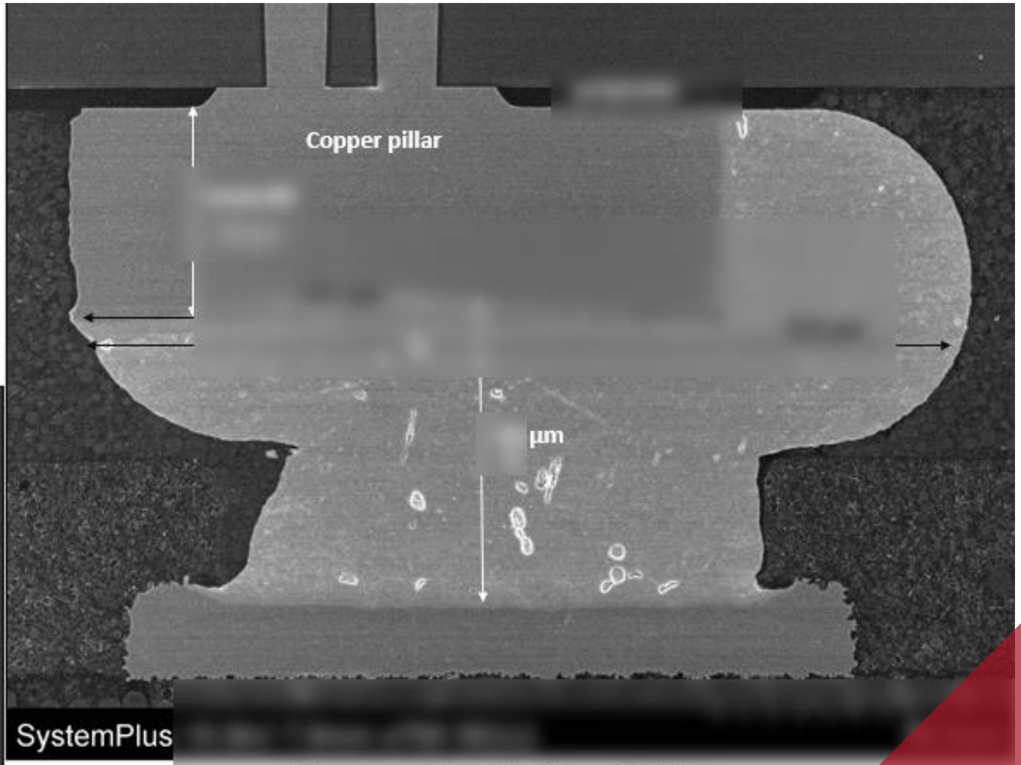
Interposer Cross-Section – SEM View
©2021 by System Plus Consulting

- The interposer die is connected to the package PCB using [illegible]. The [illegible] pitch could be regrouped into two groups: one in [illegible] region and one in [illegible] region.
- [illegible] we assume that the interposer-PCB micro bumps could [illegible] size.

Package Cross-Section – Interposer



Interposer Cross-Section – SEM View
©2021 by System Plus Consulting



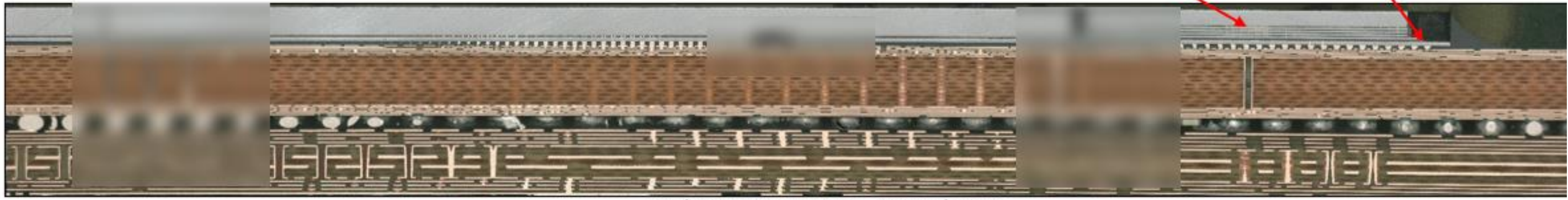
Interposer Cross-Section – SEM View
©2021 by System Plus Consulting

- Interposer copper pillar diameter: [illegible] µm

PHYSICAL
COMPARISON

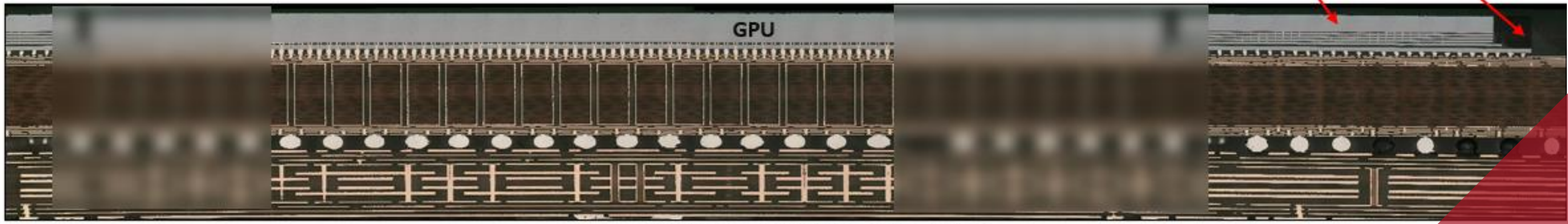
Comparison NVIDIA Tesla V100, P100 and Ampere A100- Package Cross Section

Tesla P100



Package Cross-Section – Optical View
©2021 by System Plus Consulting

Tesla V100



Package Cross-Section – Optical View
©2021 by System Plus Consulting

Ampere A100

Package Cross-Section – Optical View
©2021 by System Plus Consulting

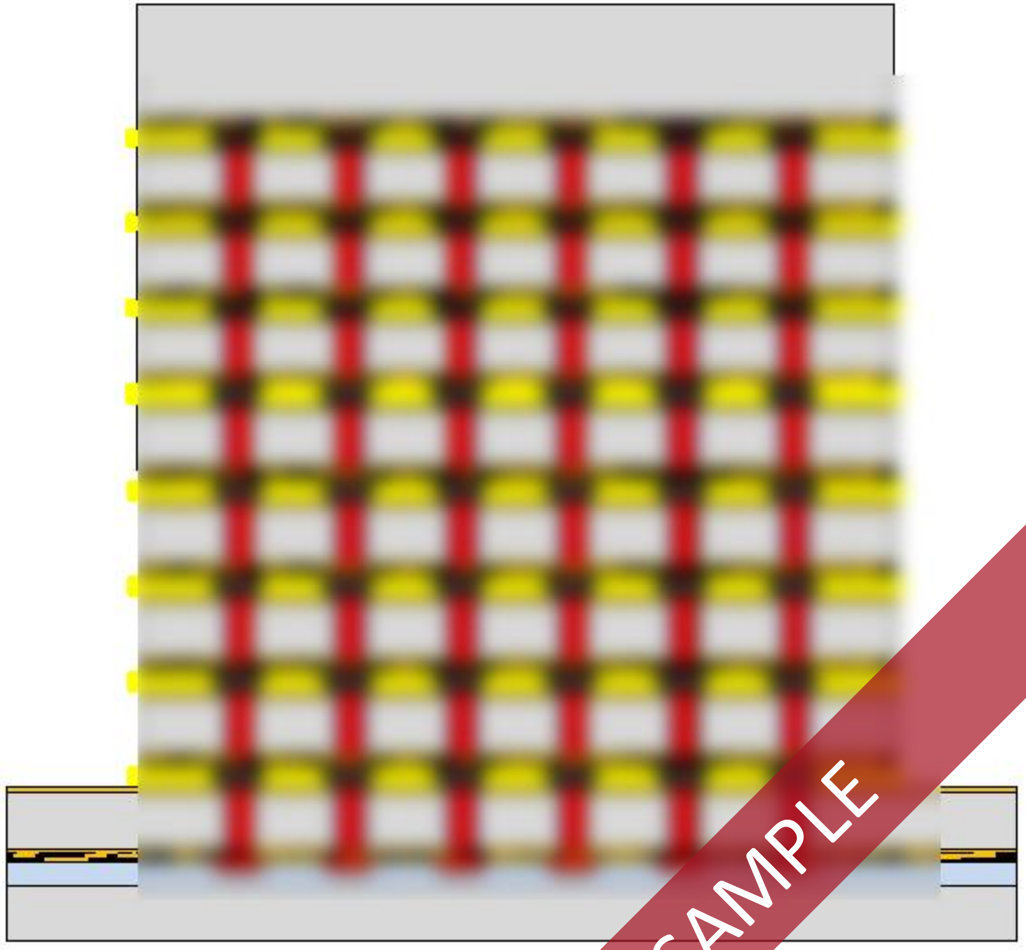
- The package and die lay-out is three packages.
- Ampere A100 HBM stack has memory dies yet the Tesla V100 and Tesla P100 had DRAM memory dies.

MANUFACTURING
PROCESS FLOW

HBM Stacking Process Flow (2/4)

HBM Stack

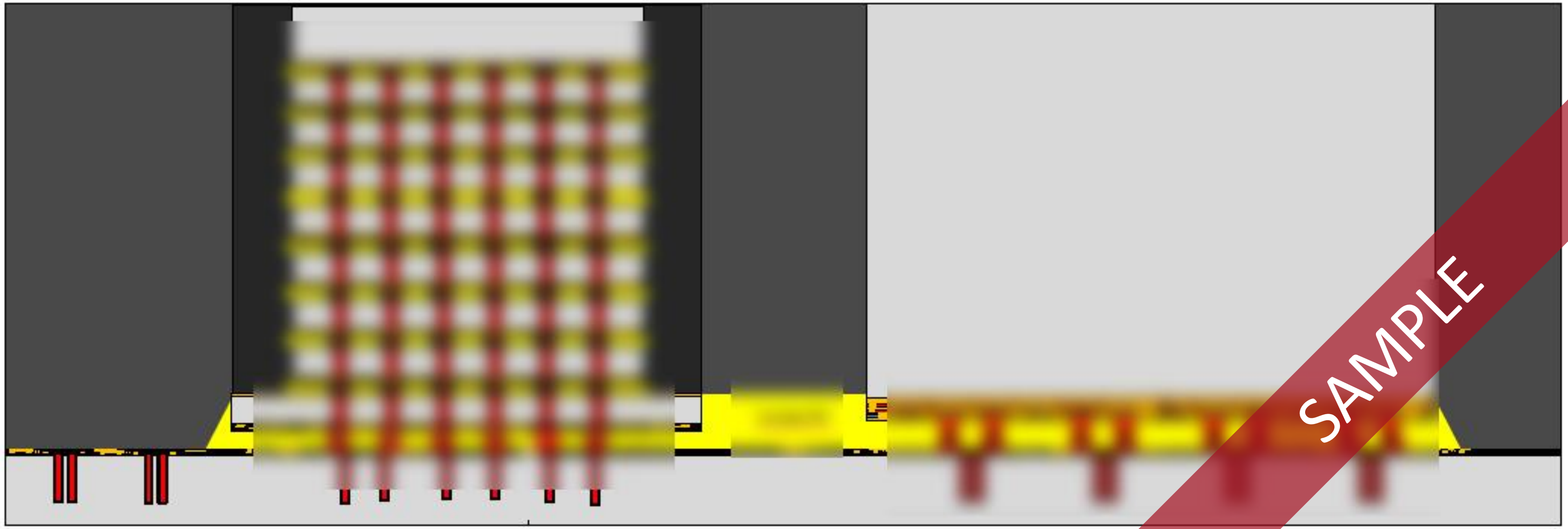
- TCB – DRAM dies Thermocompression Bonding on logic wafer



Interposer – CoW Process Flow (2/7)

Interposer

- Interposer Molding
- Mold



C O S T ANALYSIS

GPU Wafer & Die Cost

	Low Yield		Medium Yield		High Yield	
	Cost	Breakdown	Cost	Breakdown	Cost	Breakdown
Front-End Price	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Probe Test Cost	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
µBump Cost	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Backgrinding and Dicing Cost	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Total Wafer Cost (including foundry margin)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Nb of potential good dies per wafer	[illegible]		[illegible]		[illegible]	
Nb of good dies per wafer	[illegible]		[illegible]		[illegible]	
Front-End Cost	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Probe Test Cost	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
µBump Cost	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Backgrinding and Dicing Cost	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Yield Losses Cost	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Die Cost (including foundry margin)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

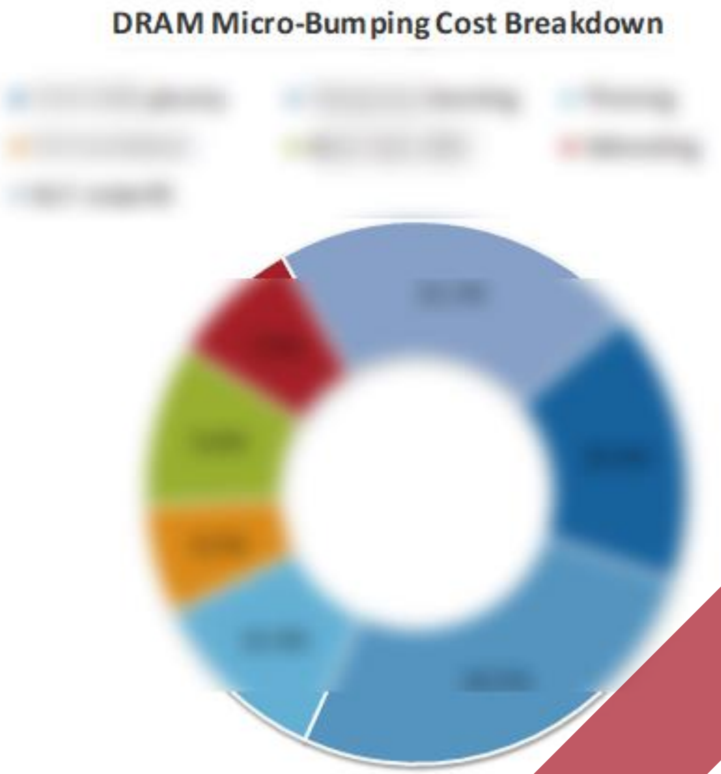
Die Cost Breakdown (Medium Yield)

The **wafer cost** for the GPU is estimated to [illegible] **in low yield and** [illegible] **in high yield. This cost includes TSMC foundry overheads.**

The number of **good dies per wafer** is estimated to range from [illegible] **to** [illegible] according to yield variations, which results in a **GPU die cost** ranging from [illegible] **to** [illegible]

DRAM Microbumping Cost

DRAM Micro-Bumping Cost	Cost	Breakdown
Front-Side µbump	[illegible]	[illegible]
Temporary bonding	[illegible]	[illegible]
Thinning	[illegible]	[illegible]
TSV Via Reveal	[illegible]	[illegible]
Back-Side UBM	[illegible]	[illegible]
Debonding	[illegible]	[illegible]
NCF Underfill	[illegible]	[illegible]
DRAM Micro-Bumping Cost	[illegible]	**100%**

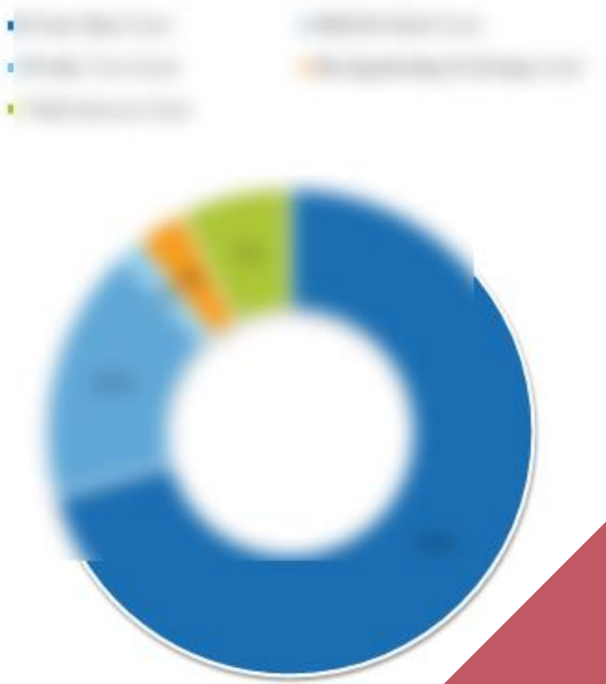


The DRAM micro-bumping manufacturing cost is estimated to $[illegible] per wafer.

DRAM Wafer & Die Cost

	Low Yield		Medium Yield		High Yield	
	Cost	Breakdown	Cost	Breakdown	Cost	Breakdown
Front-End Cost	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Middle-End Cost	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Probe Test	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Backgrinding & Dicing Cost	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Total Wafer Cost	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Nb of potential good dies per wafer	[illegible]		[illegible]		[illegible]	
Nb of good dies per wafer	[illegible]		[illegible]		[illegible]	
Front-End Cost	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Middle-End Cost	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Probe Test Cost	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Backgrinding & Dicing Cost	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Yield Losses Cost	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
DRAM Die Cost	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

DRAM Die Cost Breakdown (Medium Yield)

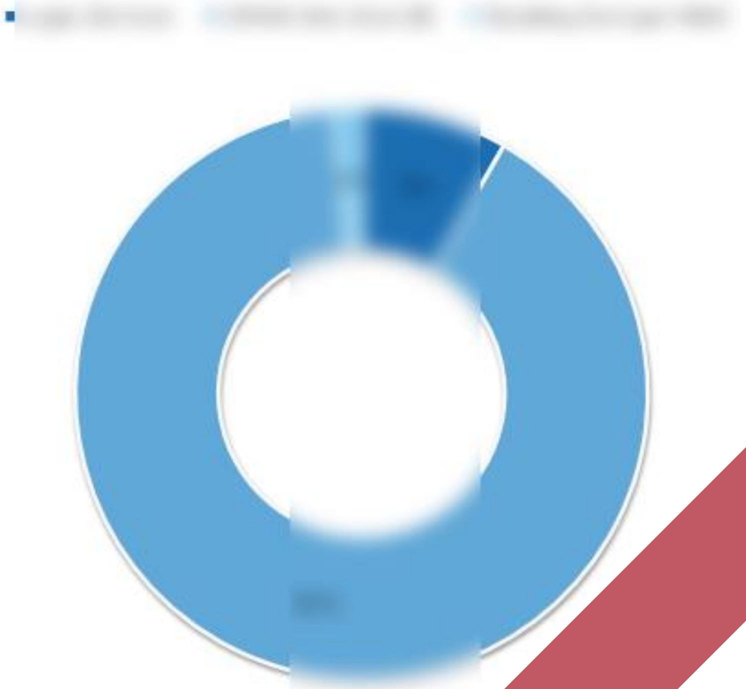


Total DRAM Wafer Cost is estimated to range between [illegible] and [illegible] between low and high yield.
The number of good dies per wafer is estimated to ranges from [illegible] to [illegible] according to yield variations, which results in a die cost ranging from [illegible] to [illegible].

HBM Stack Cost

	Low Yield		Medium Yield		High Yield	
	Cost	Breakdown	Cost	Breakdown	Cost	Breakdown
Logic Front-End Cost	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Logic Middle-End Cost	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Probe Test Cost	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Total Logic Wafer Cost	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
DRAM Dies Cost	[illegible]		[illegible]		[illegible]	
HBM Stacking Cost	[illegible]		[illegible]		[illegible]	
Total HBM Wafer Cost	[illegible]		[illegible]		[illegible]	
Nb of potential good stack per wafer	[illegible]		[illegible]		[illegible]	
Nb of good stack per wafer	[illegible]		[illegible]		[illegible]	
Logic Die Cost	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
DRAM Dies Cost (8)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Stacking Cost per HBM	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
HBM Stack Cost	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Samsung Gross Profit	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
HBM Stack Price	[illegible]		[illegible]		[illegible]	

HBM Stack Cost Breakdown (Medium Yield)



The number of good HBM stack per wafer is estimated to ranges from [illegible] to [illegible] according to yield variations, which results in a HBM stack cost ranging from $[illegible] to [illegible] according to yield.

We estimate a gross margin of 50% for Samsung, which results in a HBM stack price ranging from [illegible] to [illegible]. This corresponds to the selling price to NVIDIA.

SAMPLE

Chip-on-Wafer (CoW) Stack Wafer Cost

	Low Yield		Medium Yield		High Yield	
	Cost	Breakdown	Cost	Breakdown	Cost	Breakdown
Interposer Wafer Manufacturing Cost	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
CoW Cost	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Yield Losses Cost	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Total Interposer + CoW Wafer Cost	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Foundry Gross Profit	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Total Interposer + CoW Wafer Price	[illegible]		[illegible]		[illegible]	
HBM Stacks Cost	[illegible]		[illegible]		[illegible]	
GPU Dies Cost	[illegible]		[illegible]		[illegible]	
Filler Dies Cost	[illegible]		[illegible]		[illegible]	
Total CoW Stack Wafer Cost	[illegible]		[illegible]		[illegible]	

The manufacturing cost of the interposer including the Chip-on-Wafer assembly steps is estimated to range from [illegible] to [illegible] according to yield variations.

By considering a gross margin for TSMC (estimated to 50%), the interposer + CoW wafer price is estimated to range from [illegible] to [illegible] according to yield variations.

By adding the cost of the HBM Stacks (8) and the GPU die with fillers, the total Chip-on-Wafer stack wafer cost ranges from [illegible] to [illegible] according to yield variations.

Related Reports

REVERSE COSTING ANALYSES - SYSTEM PLUS CONSULTING

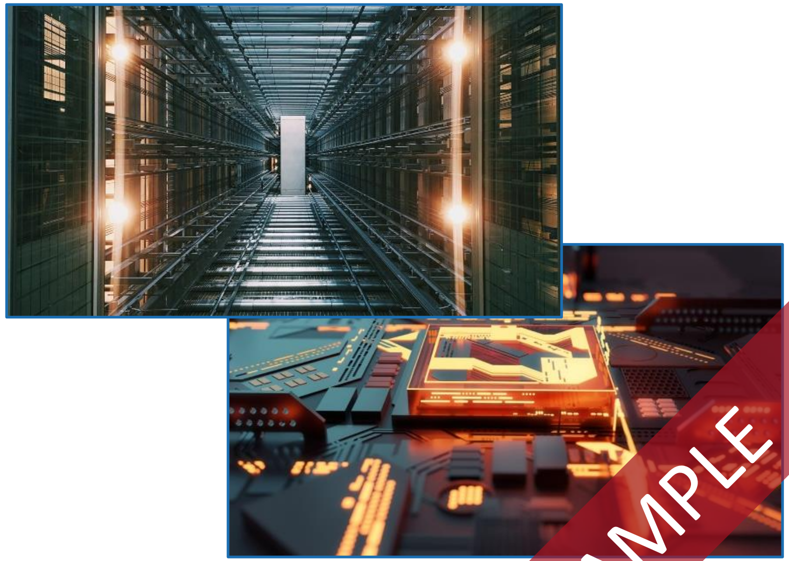
ADVANCED PACKAGING

- NVIDIA Tesla P100 GPU with HBM2
- Intel Foveros 3D Packaging Technology

MARKET AND TECHNOLOGY REPORTS - YOLE DÉVELOPPEMENT

ADVANCED PACKAGING

- Computing for Datacenter Servers 2021
- Processor Quarterly Market Monitor



COMPANY SERVICES

Our Core Activity : The Reverse Costing®

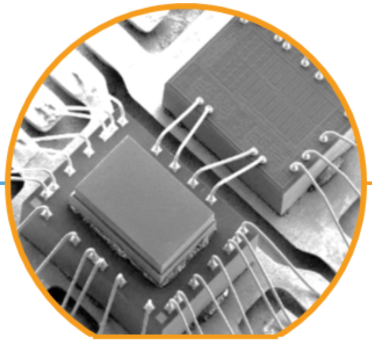
A Structure, Process and Cost Analysis

Reverse Costing® consists in disassembling a device or a system, in order to identify its technology and determine its manufacturing processes and cost, using in-house models and tools.

Fields Of Expertise

Electronic System

- Automotive
  - ADAS
  - Infotainment
  - Telematics
  - Electrification
  - Safety
- Consumer
  - Smartphone
  - Smart Home
  - Wearable
- Telecom
  - Router/Set-Top Box
  - Base Station
- Industrial
- Medical



Semiconductor Device

- Advanced Packaging
  - WLP (Fan in, Fan out)
  - SiP
  - Embedded
  - 3D Packaging
- Imaging
  - Infrared
  - Visible
- Integrated Circuit
  - ASIC
  - SOC
  - MPU/GPU/MCU/DSP
- MEMS & Sensors
  - Inertial Sensor
  - Environmental Sensor
  - Fingerprint Sensor
  - Oscillator
  - Microphone
  - Inkjet
  - RF MEMS
  - Light / Optics
- Memory
  - NAND
  - DRAM
  - Emerging
- Power Electronics
  - Discrete
  - Module
  - Compound (GaN, SiC)
  - Power RF
- RF
  - Radar
  - Filter
  - Module (FEM, Wifi/BT)
  - Power Amplifier
- Solid State Lighting
  - LED
  - Laser / VCSEL
- Photonics

Business Model

Report
60+ per year

Monitor
1 per year quarterly updated

Teardown Track
175+ teardowns per year

Custom Analysis
150 custom analyses per year

Costing Tools
5 process-based and 3 parametric costing tools

Cost Methods Training
On demand

Worldwide Presence

100+ collaborators in 8 different countries

Headquarters

- *Nantes – System Plus Consulting*
- *Lyon – Yole Développement*

Contacts

REPORTS, MONITORS & TRACKS

Western US & Canada
Steve Laferriere - steve.laferriere@yole.fr
+ 1 310 600 8267

Eastern US & Canada
Chris Youman - chris.youman@yole.fr
+1 919 607 9839

Europe and RoW
Lizzie Levenez - lizzie.levenez@yole.fr
+49 15 123 544 182

Benelux, UK & Spain
Marine Wybranietz - marine.wybranietz@yole.fr
+49 69 96 21 76 78

India and RoA
Takashi Onozawa - takashi.onozawa@yole.fr
+81 80 4371 4887

Greater China
Mavis Wang - mavis.wang@yole.fr
+886 979 336 809 +86 136 6156 6824

Korea
Peter Ok - peter.ok@yole.fr
+82 10 4089 0233

Japan
Miho Ohtake - miho.ohtake@yole.fr
+81 34 4059 204

Japan and Singapore
Itsuyo Oshiba - itsuyo.oshiba@yole.fr
+81 80 3577 3042

Japan
Toru Hosaka – toru.hosaka@yole.fr
+81 90 1775 3866

HEADQUARTER & CUSTOM PROJECT SERVICES

22 Bd Benoni Goullin
44200 Nantes, France
sales@systemplus.fr - +33 2 40 18 09 16

REPORTS & MONITORS WEBSITE

www.systemplus.fr

TRACKS WEBSITE

www.reverse-costing.com

MARKETING, COMMUNICATION & PR

Camille Veyrier, Marketing & Communication
camille.veyrier@yole.fr - +33 472 83 01 01

Sandrine Leroy, Public Relations
sandrine.leroy@yole.fr - +33 4 72 83 01 89

Follow us on