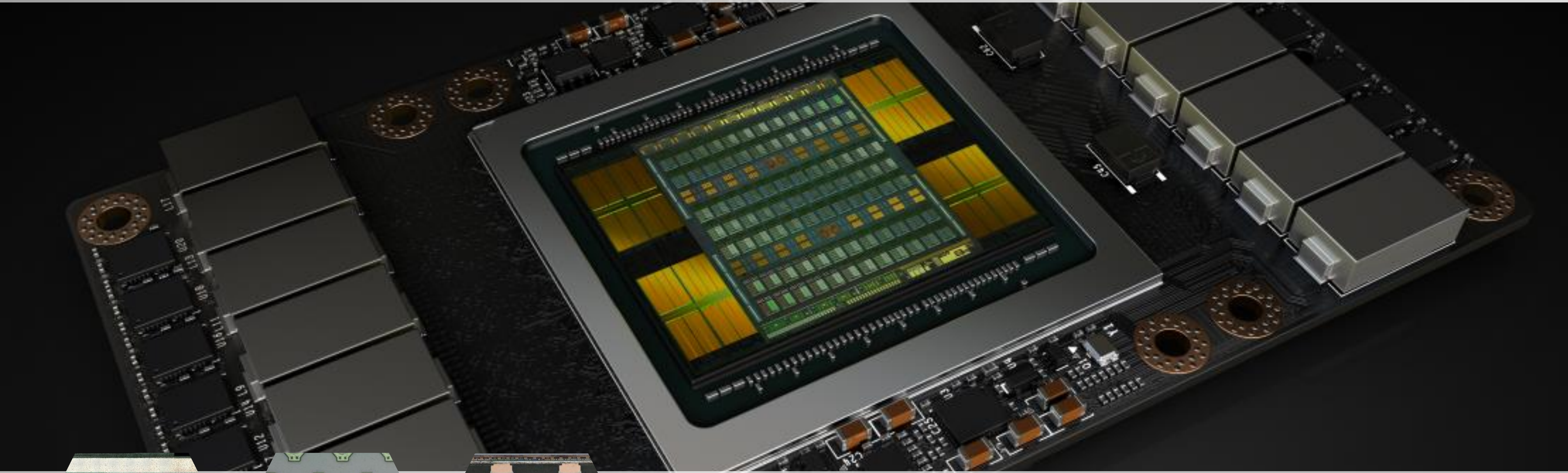
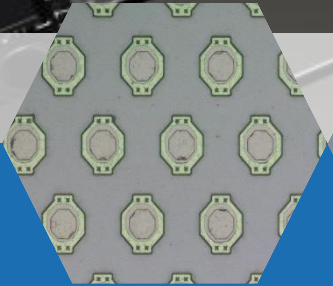
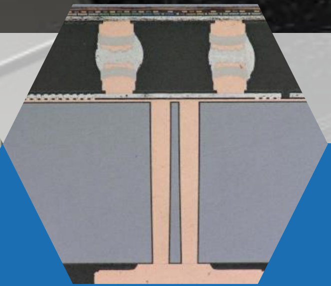
REVERSE COSTING® – STRUCTURAL, PROCESS & COST REPORT







Nvidia Tesla V100 16GB HBM2

2.5D & 3D Packaging – TSMC CoWoS – Samsung HBM2

Adv. Packaging report by Romain FRAUX

December 2018 – Version 2

22 Bd Benoni Goullin
44200 NANTES - FRANCE

+33 2 40 18 09 16

info@systemplus.fr

www.systemplus.fr

Versions of the Report

Version	Date	Updates
V1	15/12/2018	○ Initial release
V2	02/01/2019	○ Added DRAM Driver Die ✓ Interposer µBumps Tweak ○ Added GPU Die ✓ µBumps dimensions ✓ Interposer µBumps dimensions: added middle and side differences ○ Added Interposer Die ✓ Reticle Stitching size and location ✓ µBumps Dimensions ✓ L/S Width

Table of Contents

OVERVIEW
METHODOLOGY

Executive Summary

Overview / Introduction
- Executive Summary
- Reverse Costing Methodology
- Glossary

Company Profile & Supply Chain

Physical Analysis

Manufacturing Process Flow

Cost Analysis

Selling Price Analysis

Feedback

About System Plus

- Targeted for High Performance Computing (HPC) and deep learning, the NVIDIA Tesla V100 includes3D stacked memory with 2.5D integration on a silicon interposer in a Chip-on-Wafer-on-Substrate (CoWoS) process.
- The Tesla V100 accelerators are equipped with 16GB or 32GB of second generation high bandwidth memory (HBM2).
- HBM2 greatly increases memory capacity and bandwidth over first generation HBM1 technology. HBM1 was limited to 1GB of memory per stack of four dynamic random access memory (DRAM) die with maximum capacity of 256MB and 125GB/sec of bandwidth. That compares to 8GB of memory per stack of eight stacked DRAM dies with maximum capacity of 1GB and 180GB/sec bandwidth for HBM2.
- The single 55mm x 55mm 12-layer ball grid array (BGA) package of the NVIDIA Tesla V100 includes more than 4,000 mm^2 of silicon area. Two industry leaders, TSMC and Samsung, had to come together to deliver this much silicon area in a package.
- TSMC is the main provider for the Tesla V100. Using its 2.5D CoWoS platform, it manufactures the GV100 GPU die, featuring a 12nm FinFET process and 21.1 billion transistors. It also produces a large silicon interposer on top of which the GPU is assembled at the wafer-level with its four HBM2 stacks.
- Samsung provides the HBM2 stacks. A 3D assembly process yields HBM2 stacks composed of four 1GB DRAM memory dies and one buffer die, connected with via-middle through-silicon vias and micro-bumps.
- The report includes a complete physical analysis of the packaging process, with details on all technical choices regarding process, equipment and materials. Also, the complete manufacturing supply chain is described and manufacturing costs are calculated.

Reverse Costing Methodology

The reverse costing analysis is conducted in 3 phases:

Teardown analysis

- Package is analyzed and measured
- The dies are extracted in order to get overall data: dimensions, main blocks, pad number and pin out, die marking
- Setup of the manufacturing process.

Costing analysis

- Setup of the manufacturing environment
- Cost simulation of the process steps

Selling price analysis

- Supply chain analysis
- Analysis of the selling price

Glossary

Acronym	Definition
Al	Aluminum
ASIC	Application Specific Integrated Circuit
BGA	Ball Grid Arra
CMOS	Complementary Metal–Oxide–Semiconductor
CoWoS	Chip-on-Wafer-on-Substrate
DRAM	Dynamic Random Access Memory
DRIE	Deep Reactive Ion Etching
EDX	Energy Dispersive X-ray spectroscopy
G&A	General & Administrative
GPU	Graphics Processor Unit
HBM	High Bandwidth Memory
HMC	Hybrid Memory Cube
OEM	Original Equipment Manufacturer
PCB	Printed Circuit Board
PGDW	Potential Good Dies per Wafer
R&D	Research and Development
SEM	Scanning Electron Microscope
Si	Silicon
SiO2	Silicon Dioxide
TSV	Through-Silicon Via

COMPANY PROFILE

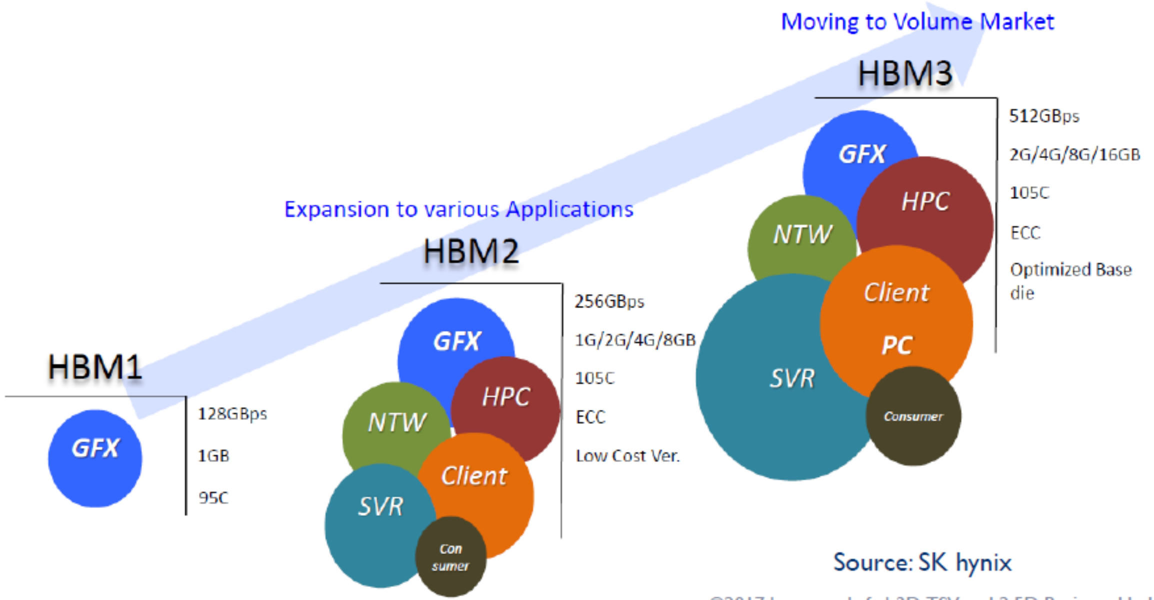
Market Forecast

HIGH-BANDWIDTH MEMORY ROADMAP

Because of further adoption of HBM2, a 3rd generation is under development

- High-Bandwidth Memory product is continuously adopted by players and is integrated in many advanced and essential products.
- AMD adopted first HBM 1st generation in 2015 in high-performance graphic card for gaming.
- Since AMD, others have followed such as Nvidia, Xilinx, Intel pushing HBM2 products in HPC, servers and other applications.
- A 3rd generation is in preparation by memory manufacturers (Samsung & SK hynix) enabling twice the data flowrate.

HBM will penetrate various market segments in the short future



Source: SK hynix

source: Yole 3D TSV and 2.5D Business Update - Market and Technology Trends 2017

Market Forecast

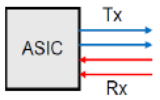
3D STACKED MEMORY CUBE

Technical comparison between HBM and HMC Memory Systems

In terms of packaging, one of the main differences between HMC and HBM is the final integration. HBM will require a **silicon interposer** mounted on a substrate whereas HMC can be placed directly on an **organic substrate.**

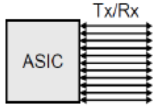
High Bandwidth Memory Interfaces

- HMC: Stand-Alone Memory Module on Board
  - Uses Serial Interface between ASIC and HMC – few high speed signals.
  - Wired across system board.
  - High speed signals need isolation; Drives number of BGA's up.



- *Single fast SERDES lane.*
- *15-30 Gbps bidirectional bandwidth.*
- *Also requires many isolation pins.*

- HBM: On-Module Memory Integrated w/ Si Interposer
  - Uses Parallel Interface – many, many, fairly slow signals.
  - Generally simpler IO and lower power.
  - Too many signals to get off package.
  - Lower cost memory, but higher complexity package integration.



- *Many slow parallel lanes.*
- *~2 Gbps/line unidirectional.*
- *30 Gbps bidirectional BW needs15 parallel signal lines for each direction.*

Source: GlobalFoundries

Variables	HBM1	HBM2	HMC1	HMC2
VDD (V)	1.5	1.5	1.35	1.2
Max. Data Rate (Gbps)	1	2	15	30
Bus Width (bits)	1024		4 Links (16 TX/RX lanes per link)	
Max Stack Bandwidth (GB/s)	128	256	120	320
Signaling	Single ended		Differential	
Interface	Wide parallel		Serial	
Channel overhead	Short		Long	
Format	In a Si Interposer		Stand-alone as a complete package	
Control distribution in logic	Simple DRAM		Advanced Transactional	

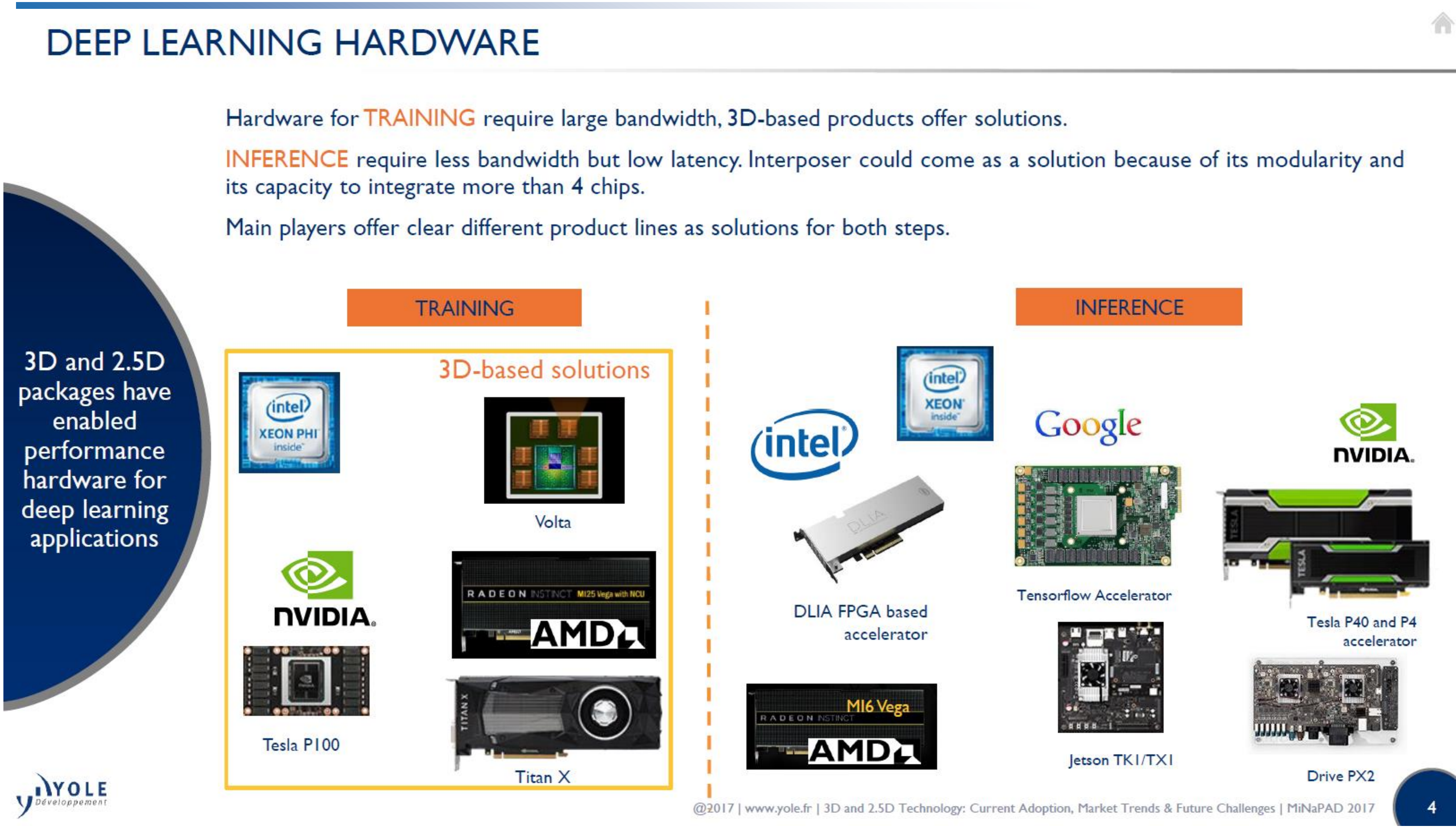
Source: Rambus

YOLE Développement

 3

source: Yole 3D TSV and 2.5D Business Update - Market and Technology Trends 2017

Market Forecast



source: Yole 3D TSV and 2.5D Business Update - Market and Technology Trends 2017

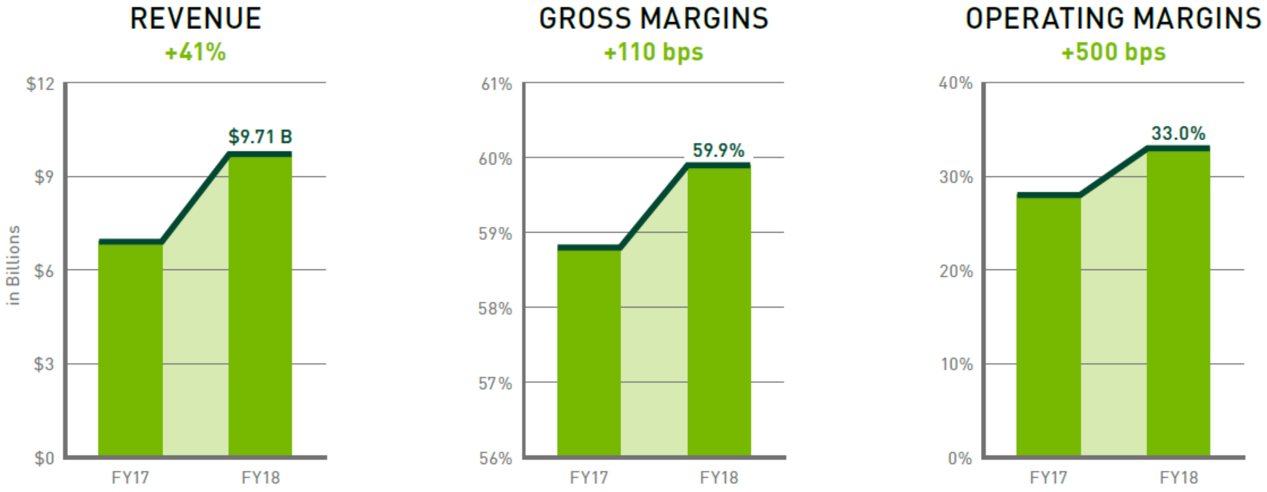
Company Profile – NVIDIA

NVIDIA Financial Highlights:

- Sales revenues 2018: $9.7 Billion
- Gross margin 2018: 59.9%
- Net income 2018: $3.0 Billion

NVIDIA Employees:

- Date of Establishment
  - ✓ April, 1993
- Headquarter:
  - ✓ Santa Clara, California, USA
  - ✓ 10,299 employees worldwide (as of 29 January, 2017).



NVIDIA Records (source: NVIDIA)

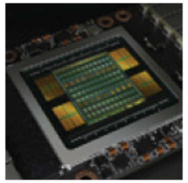
Company Profile – NVIDIA

Manufacturing Supply Chain (source: NVIDIA)

- Semiconductor wafers:
  - TSMC
  - Samsung
- Assembly, Testing and Packaging:
  - ASE
  - BYD Auto
  - Hon Hai Precision
  - JSI Logistics
  - King Yuan Electronics
  - Siliconware Precision Industries
- Substrates:
  - Ibiden
  - Nanya Technology
  - Unimicron Technology
- Memories:
  - Samsung
  - SK Hynix

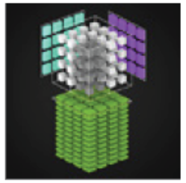
NVIDIA Telsa V100 Characteristics

NVIDIA Tesla V100 with Volta GV100 GPU



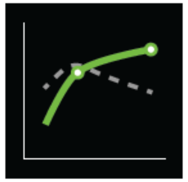
VOLTA ARCHITECTURE

By pairing CUDA Cores and Tensor Cores within a unified architecture, a single server with Tesla V100 GPUs can replace hundreds of commodity CPU servers for traditional HPC and Deep Learning.



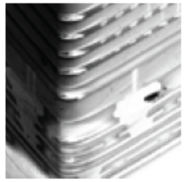
TENSOR CORE

Equipped with 640 Tensor Cores, Tesla V100 delivers 125 teraFLOPS of deep learning performance. That's 12X Tensor FLOPS for DL Training, and 6X Tensor FLOPS for DL Inference when compared to NVIDIA Pascal™ GPUs.



MAXIMUM EFFICIENCY MODE

The new maximum efficiency mode allows data centers to achieve up to 40% higher compute capacity per rack within the existing power budget. In this mode, Tesla V100 runs at peak processing efficiency, providing up to 80% of the performance at half the power consumption.



HBM2

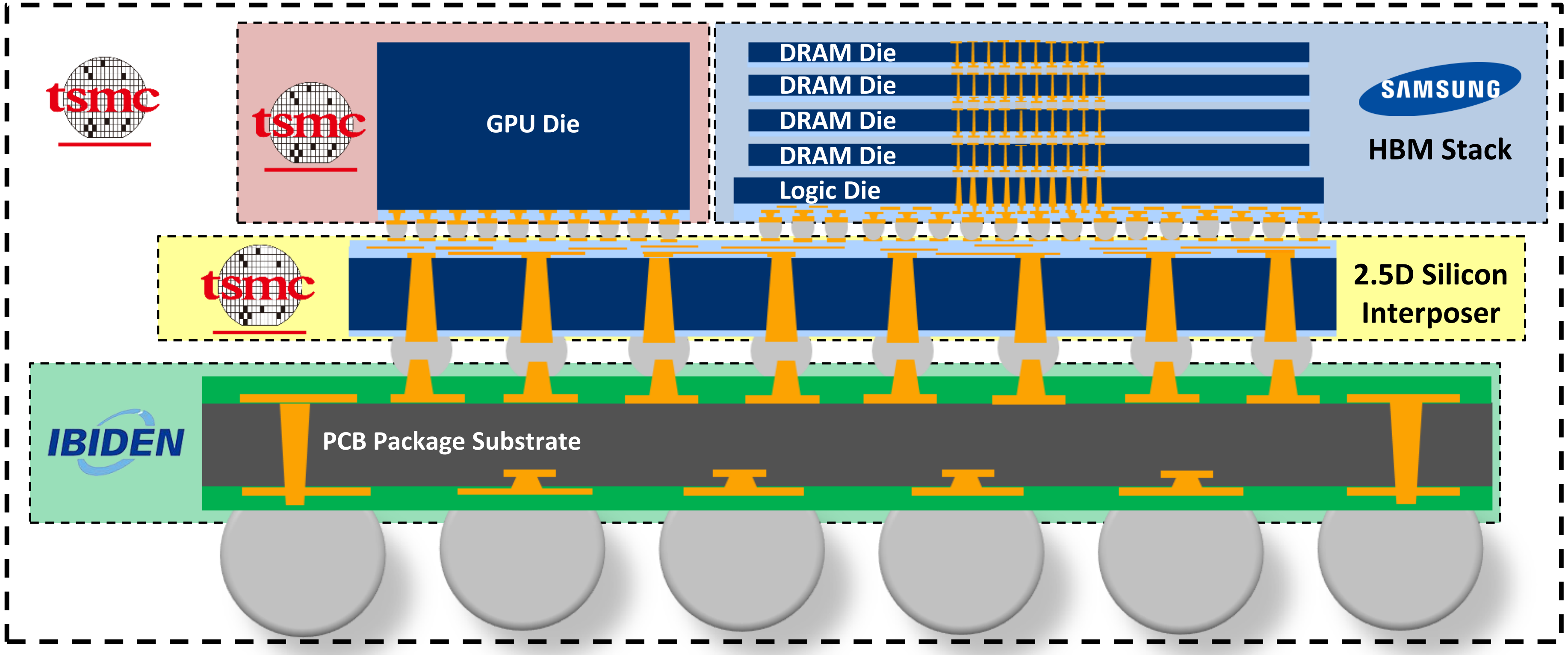
With a combination of improved raw bandwidth of 900GB/s and higher DRAM utilization efficiency at 95%, Tesla V100 delivers 1.5X higher memory bandwidth over Pascal GPUs as measured on STREAM. Tesla V100 is now available in a 32GB configuration that doubles the memory of the standard 16GB offering.

SPECIFICATIONS


	Tesla V100 PCle	**Tesla V100 SXM2**
GPU Architecture	**NVIDIA Volta**	
NVIDIA Tensor Cores	**640**	
NVIDIA CUDA® Cores	**5,120**	
Double-Precision Performance	**7 TFLOPS**	**7.8 TFLOPS**
Single-Precision Performance	**14 TFLOPS**	**15.7 TFLOPS**
Tensor Performance	**112 TFLOPS**	**125 TFLOPS**
GPU Memory	**32GB /16GB HBM2**	
Memory Bandwidth	**900GB/sec**	
ECC	**Yes**	
Interconnect Bandwidth	**32GB/sec**	**300GB/sec**
System Interface	**PCIe Gen3**	**NVIDIA NVLink**
Form Factor	**PCIe Full Height/Length**	**SXM2**
Max Power Comsumption	**250 W**	**300 W**
Thermal Solution	**Passive**	
Compute APIs	**CUDA, DirectCompute, OpenCL™, OpenACC**	

Volta GPU Supply Chain

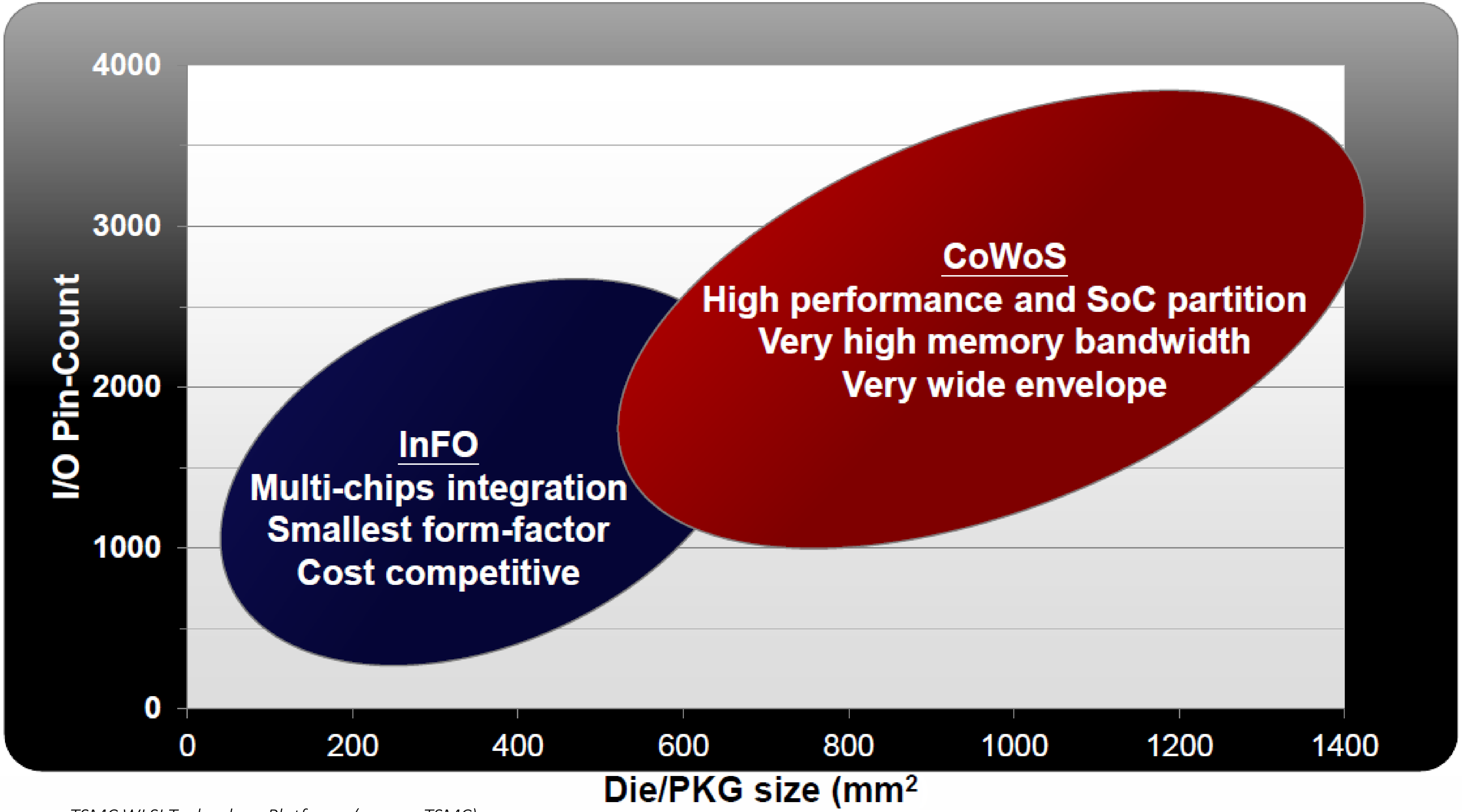
Component manufacturing supply chain:



Package Manufacturing Supply Chain:

- ✓ The HBM stack (memory dies, logic die and 3D interconnection) is made by Samsung in Korea.
- ✓ The GPU die is manufactured by TSMC in Taiwan.
- ✓ The Interposer is produced by **TSMC in Taiwan**.
- ✓ The PCB package substrate is made by Ibiden in Japan.
- ✓ The final assembly (HBM and GPU on interposer, interposer on PCB, passives assembly and BGA balls) is realized by TSMC in Taiwan.

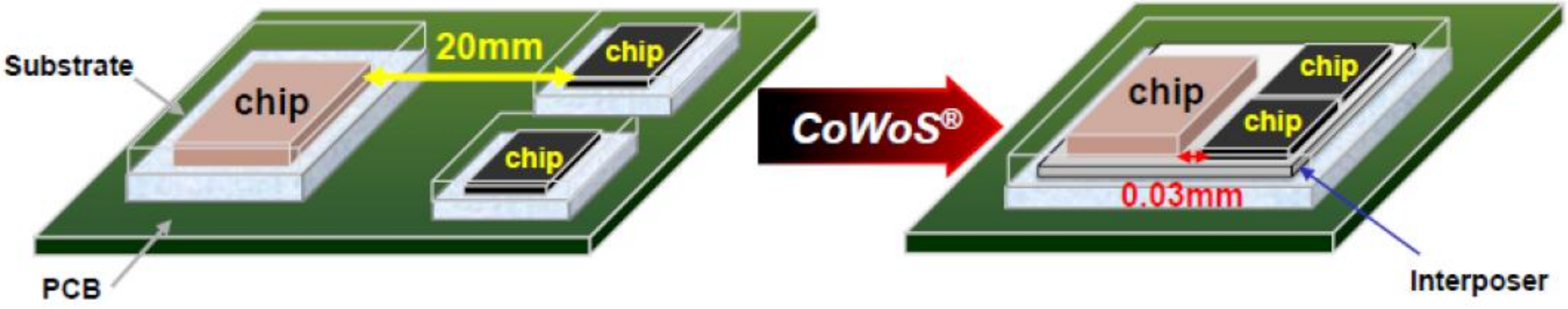
TSMC CoWoS



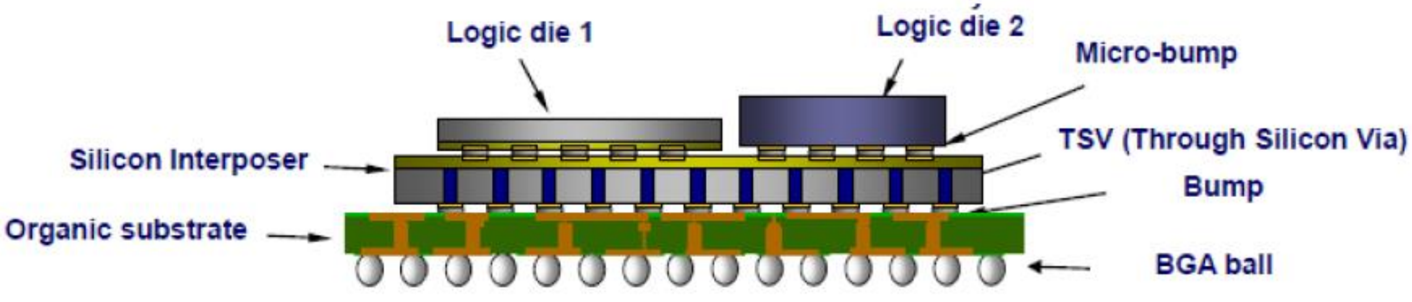
TSMC WLSI Technology Platforms (source: TSMC)

TSMC CoWoS

- Integrate multiple chips into one single package using a sub-micron scale silicon interface (interposer)



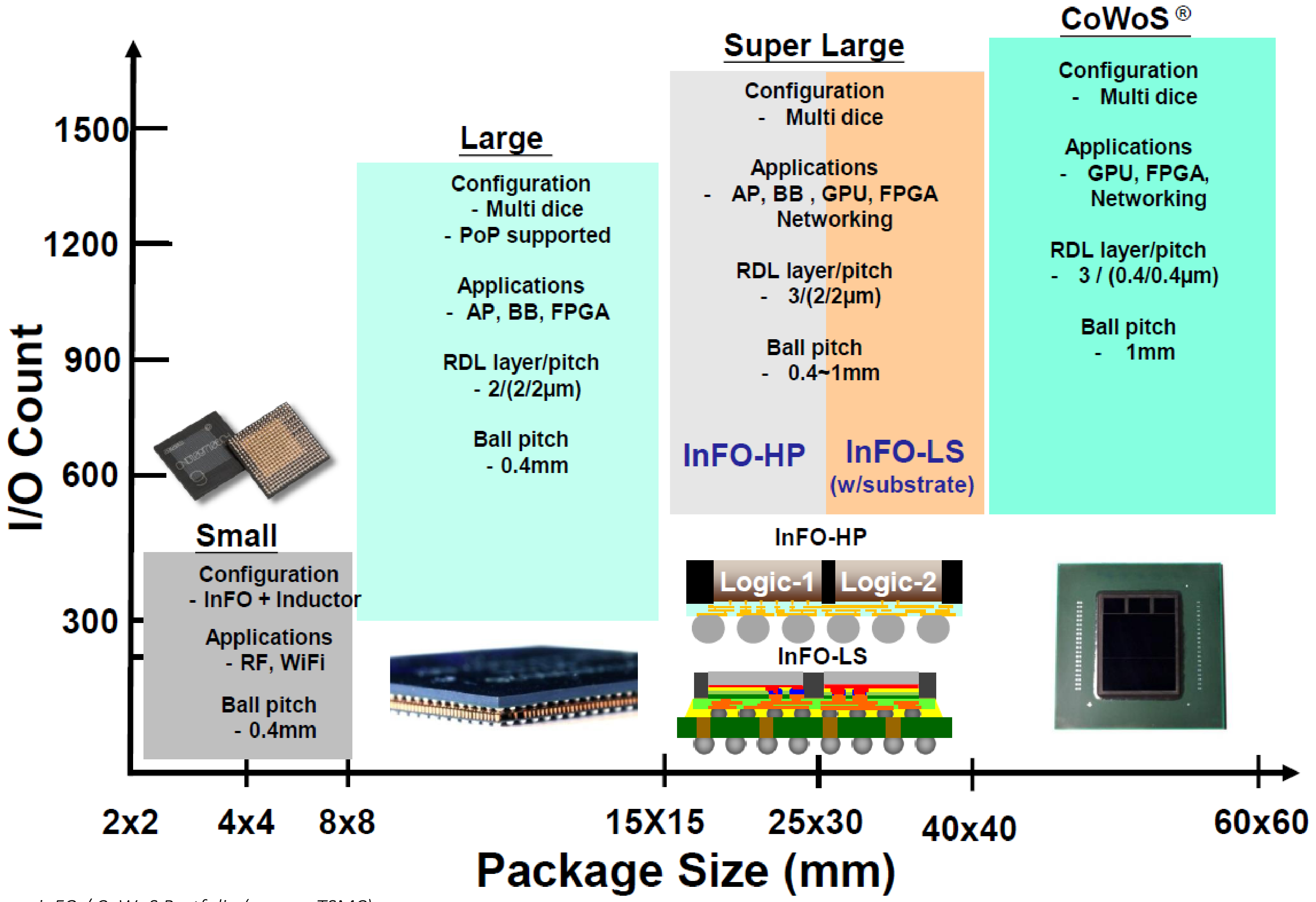
- Enable higher performance, lower power consumption, and smaller form factor
- Best integrated flow for high yield and reliability



Heterogeneous Integration

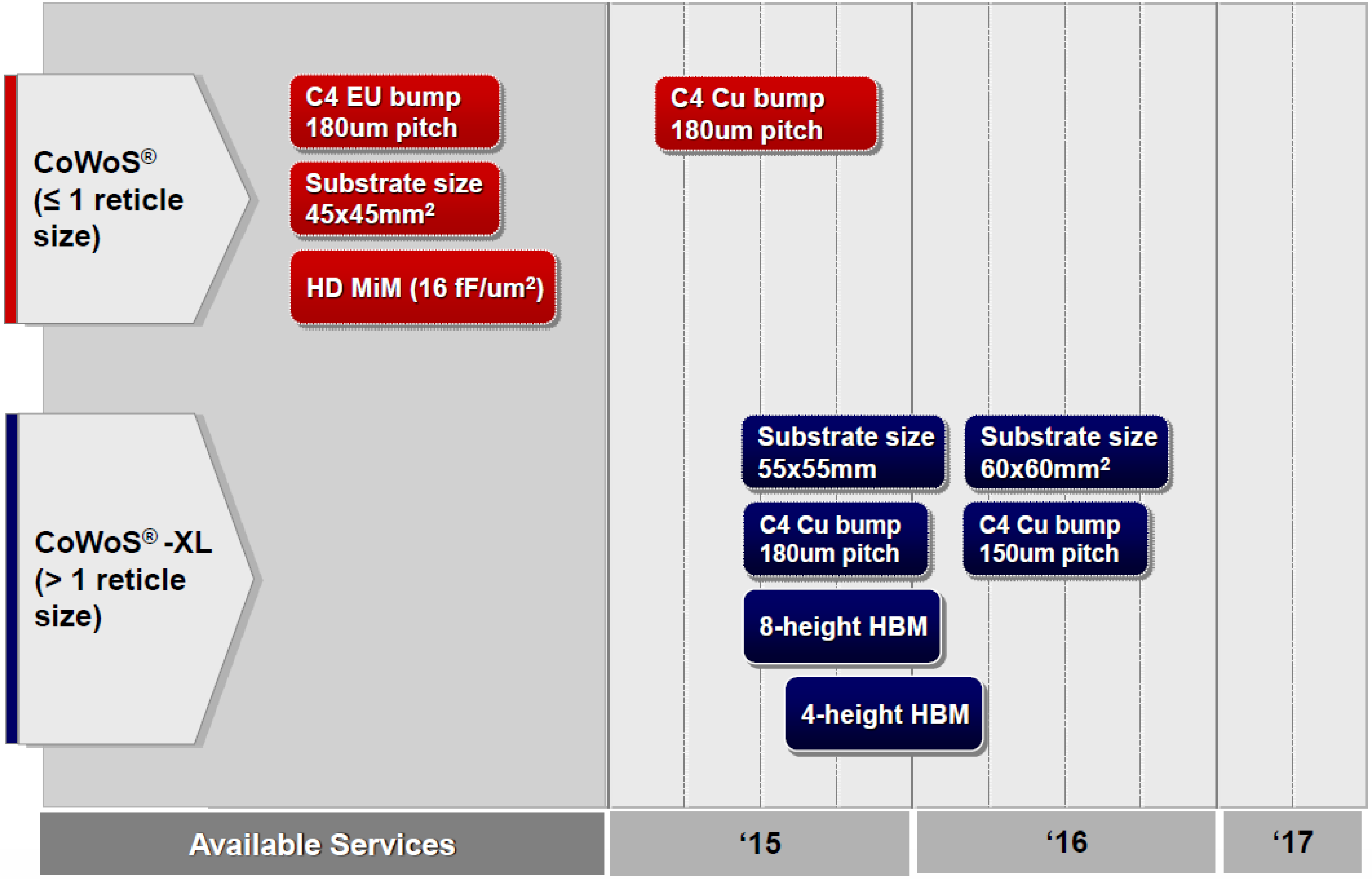
CoWoS Technology (source: TSMC)

TSMC CoWoS



inFO / CoWoS Portfolio (source: TSMC)

TSMC CoWoS

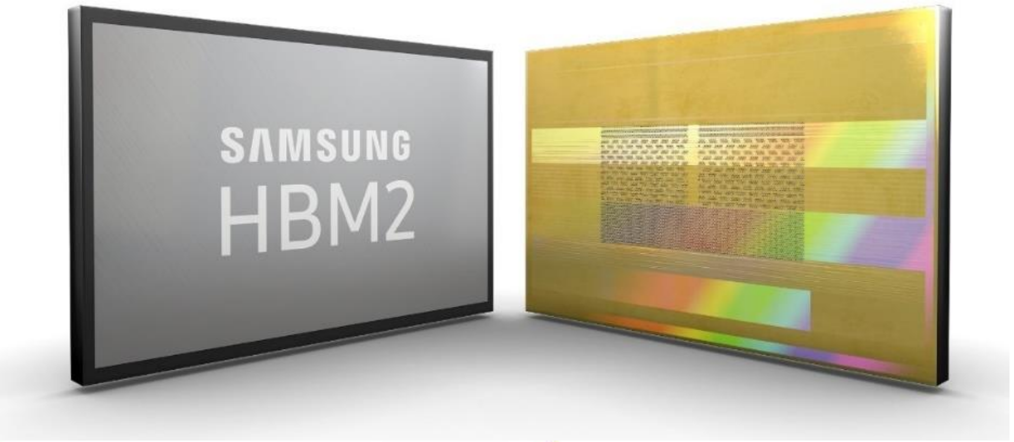


CoWoS Roadmap (source: TSMC)

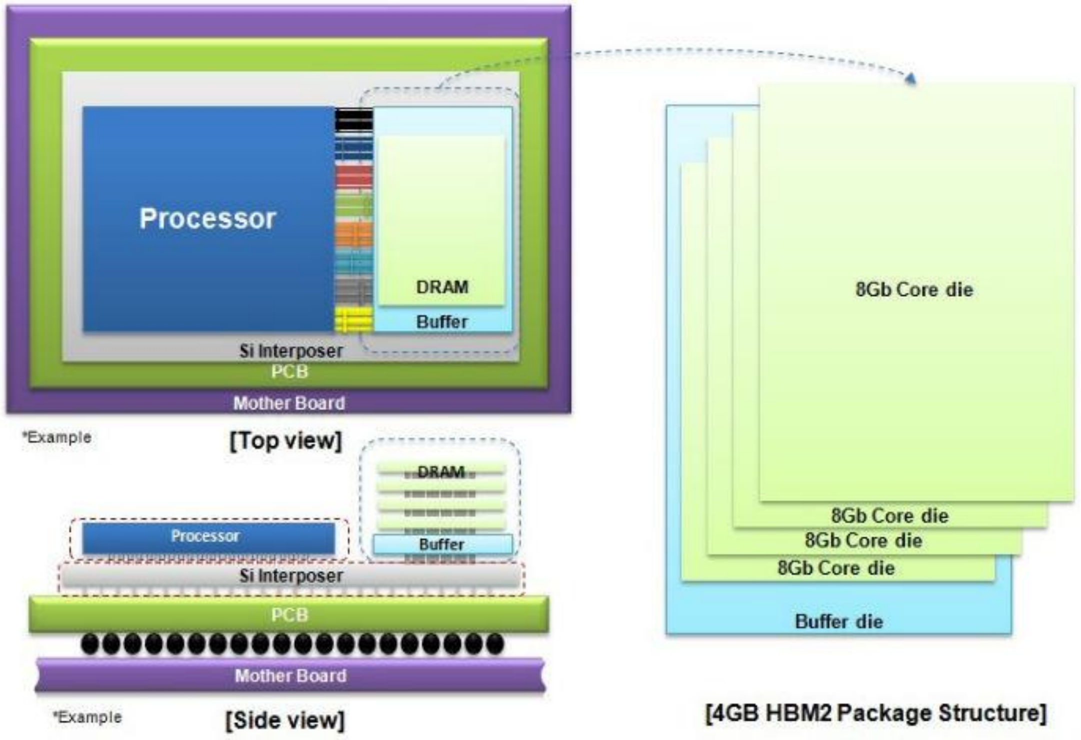
Samsung HBM2

Samsung HBM2 Stack

- Mass production: beginning of 2016
- Stack of 4GB
  - 4 stacked dies of 1GB



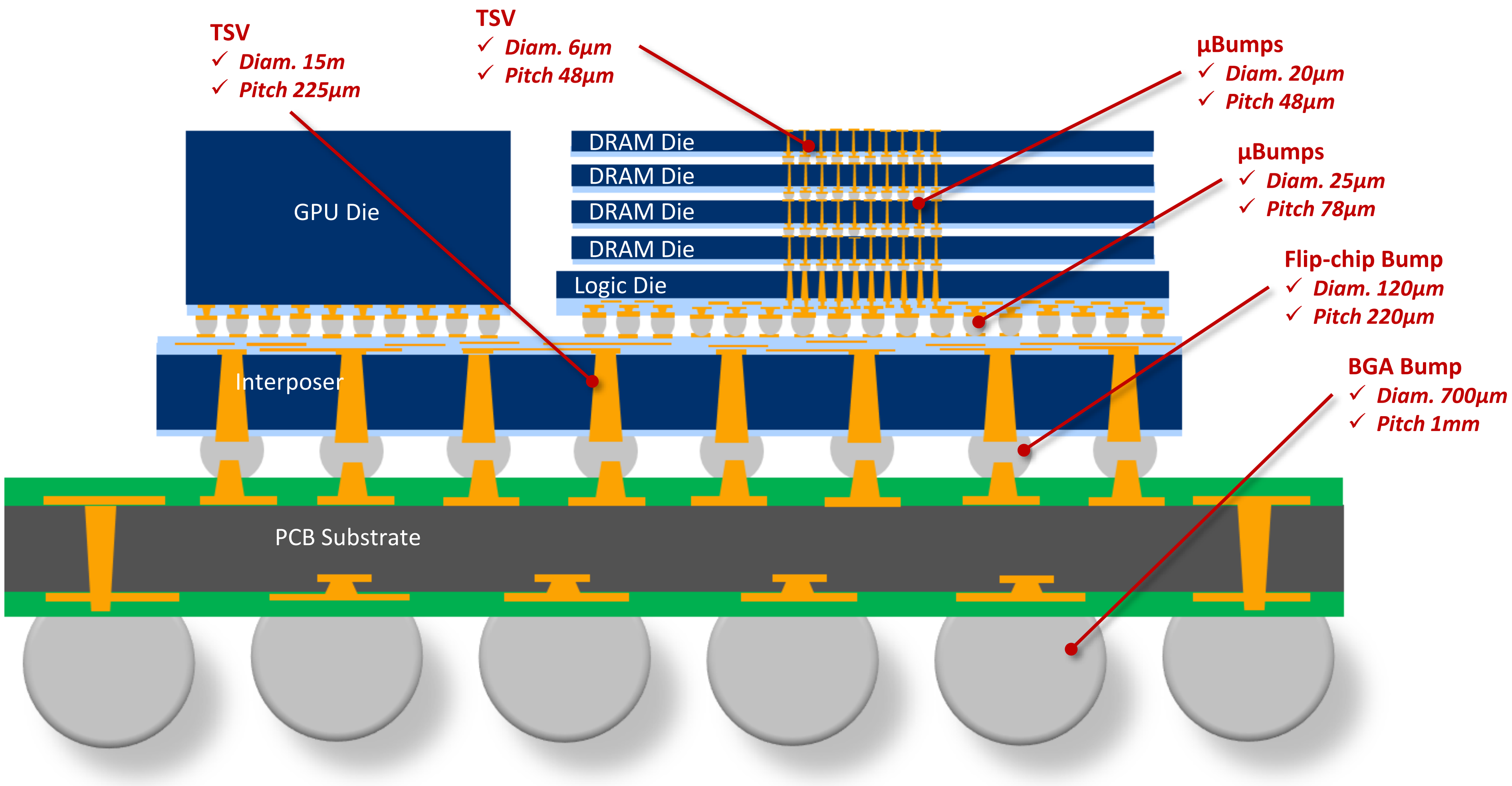
- The 4GB HBM2 package is created by stacking a buffer die at the bottom and four 8-gigabit (Gb) core dies on top. These are then vertically interconnected by TSV holes and microbumps. A single 8Gb HBM2 die contains over 5,000 TSV holes, which is more than 36 times that of a 8Gb TSV DDR4 die, offering a dramatic improvement in data transmission performance compared to typical wire-bonding based packages.
- Samsung's new DRAM package features 256GBps of bandwidth, which is double that of a HBM1 DRAM package. This is equivalent to a more than seven-fold increase over the 36GBps bandwidth of a 4Gb GDDR5 DRAM chip, which has the fastest data speed per pin (9Gbps) among currently manufactured DRAM chips. Samsung's 4GB HBM2 also enables enhanced power efficiency by doubling the bandwidth per watt over a 4Gb-GDDR5-based solution, and embeds ECC (error-correcting code) functionality to offer high reliability.



Source: Samsung

PHYSICAL ANALYSIS

Summary of the Physical Analysis



PACKAGE STRUCTURE:

- 3D Packaging: 5 stacked dies with TSV & µBumps (HBM stack).
- 2.5D Packaging: HBM stack and GPU stacked with µBumps on a silicon interposer holding TSV.
- Flip-chip BGA: silicon interposer flip-chipped to a 12-layers PCB substrate

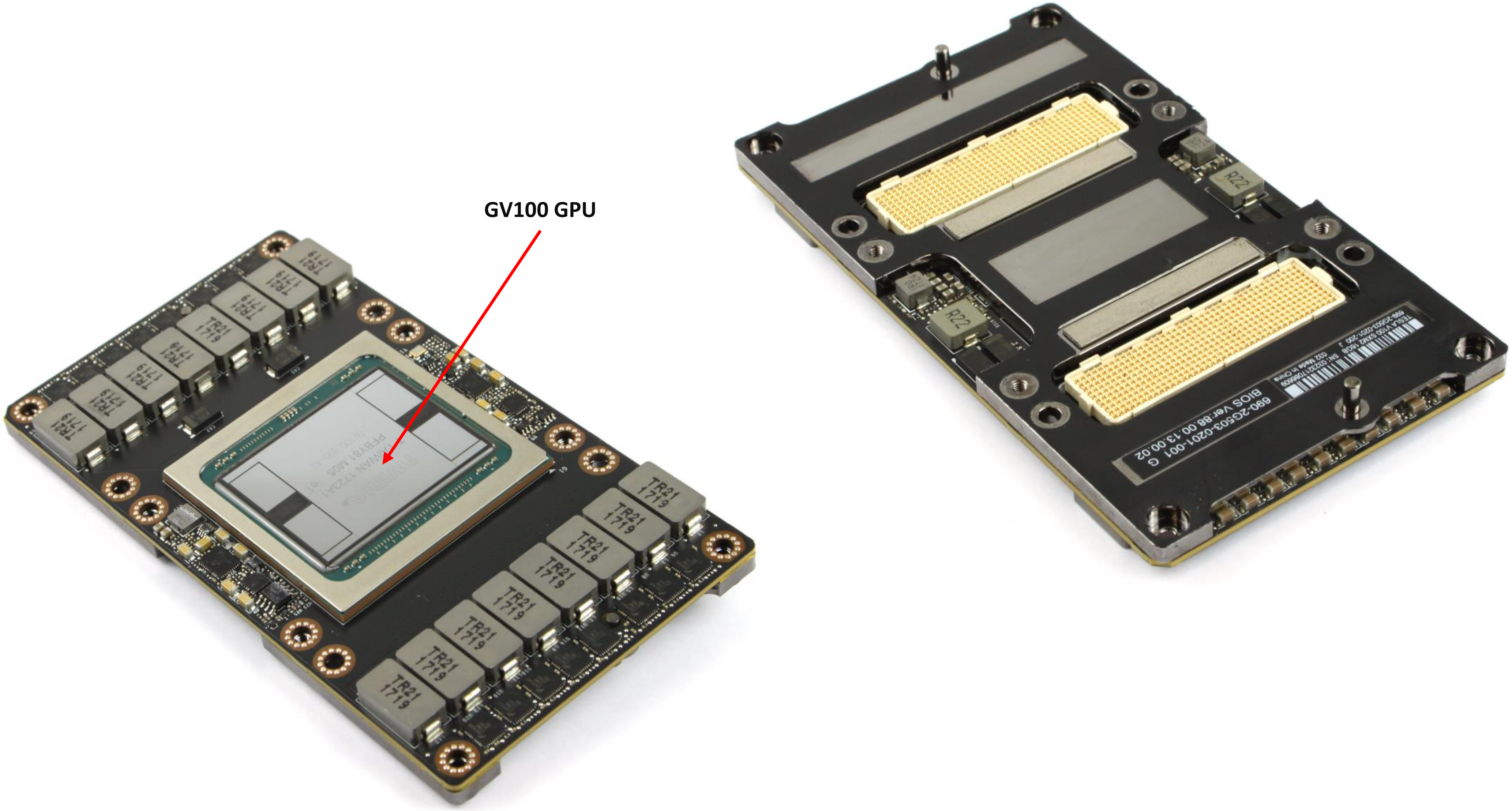
Physical Analysis Methodology

- Package is analyzed and measured.

- The package is opened to get overall die data: dimensions, main characteristics, device marking.
  - Pictures of selected area are made in order to understand the assembly.

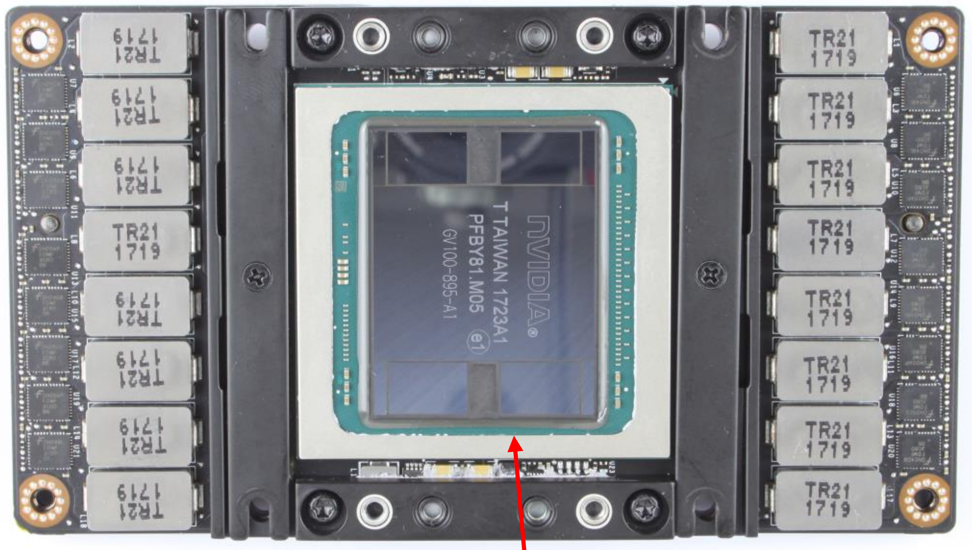
- The dies are separated to get overall die data: dimensions, main blocks, pad number and pin out, die marking.
  - Removal of metal layers.
  - Pictures (SEM & optical) of selected areas.
  - Cross section to measure thicknesses.

Graphic Card Teardown

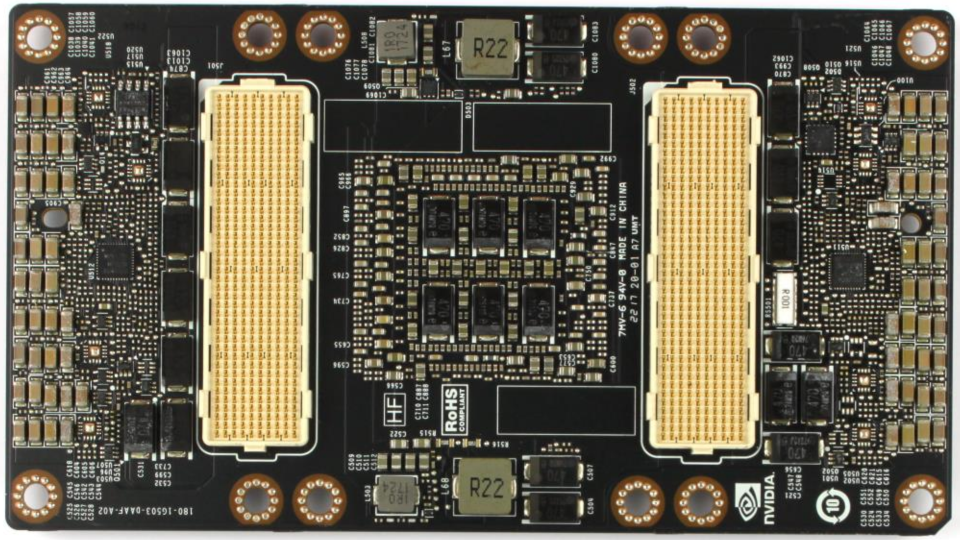


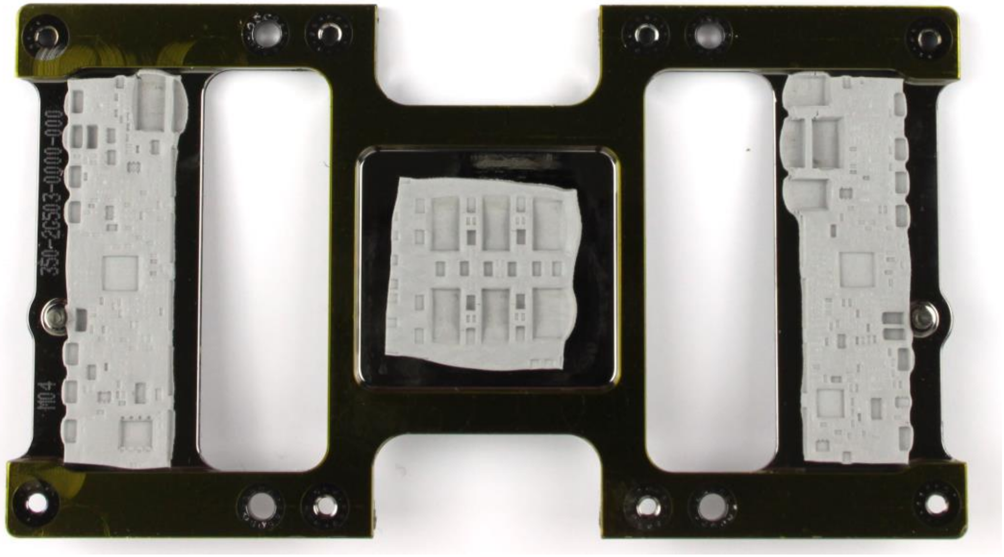
NVIDIA Tesla V100 Opening
©2018 by System Plus Consulting

Graphic Card Teardown



GV100 GPU



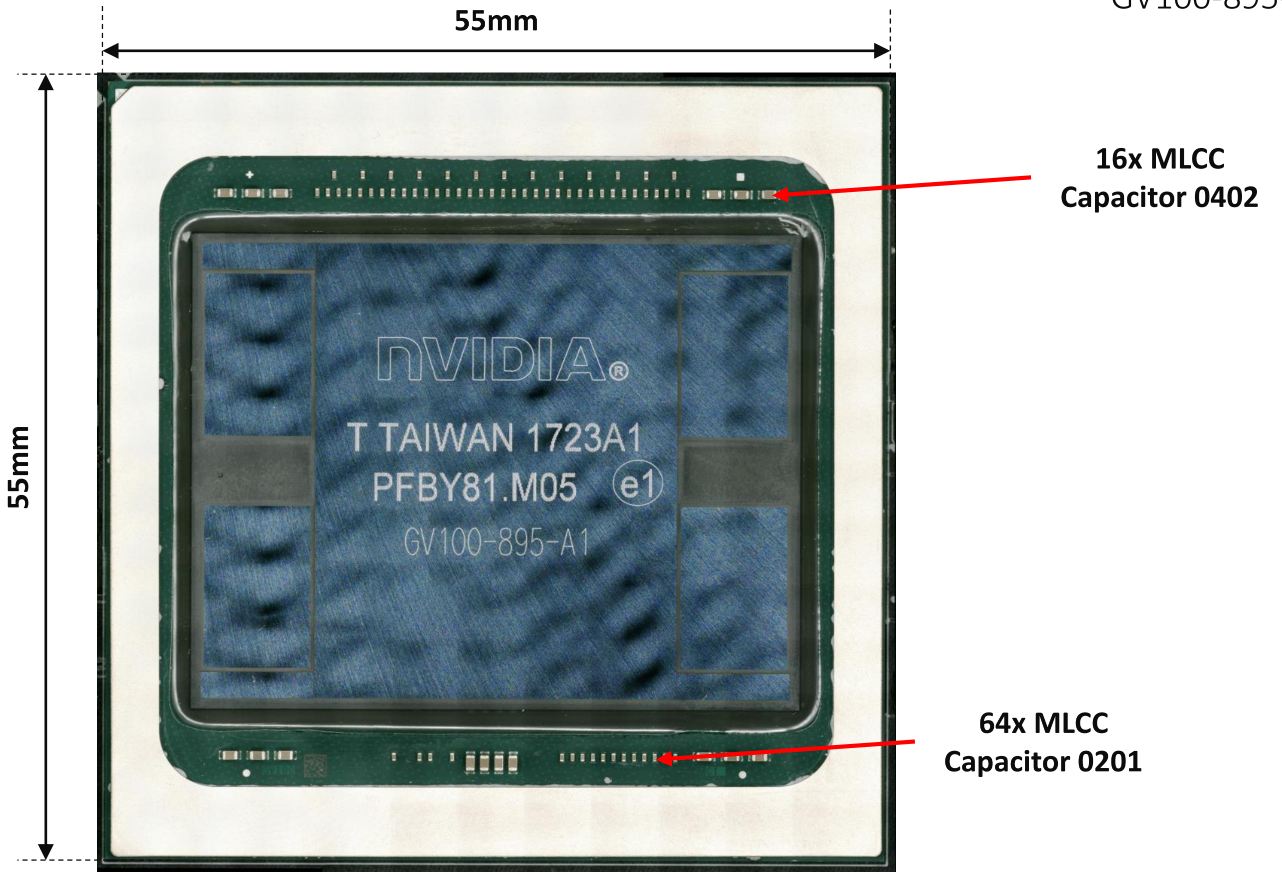


NVIDIA Tesla V100 Opening

Package Views & Dimensions

- Package: FCBGA 2540-ball
- Dimensions: 55 x 55 mm
- Pin Pitch: 1mm

- Marking:
  <logo Nvidia>
  T TAIWAN 1723A1
  PFBY81.M05
  GV100-895-A1

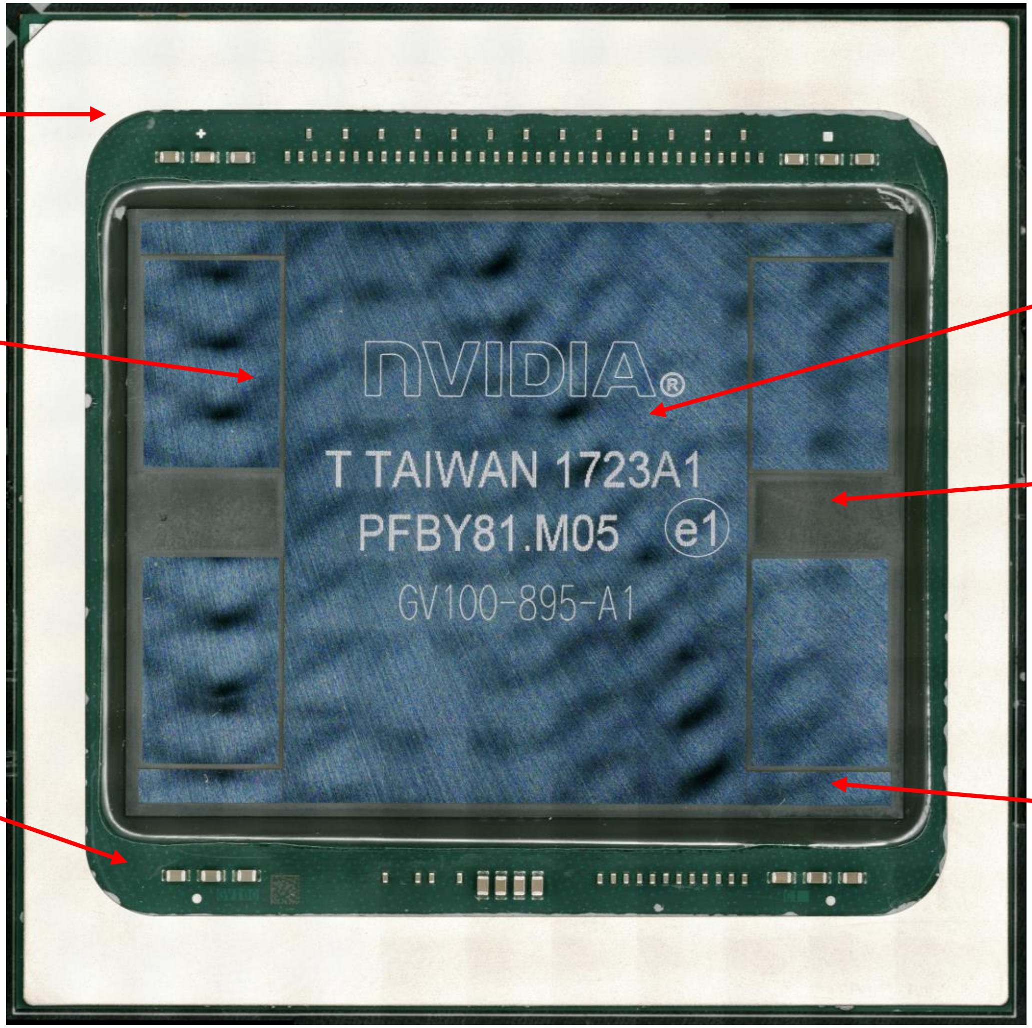


Package Top View

Package Bottom View

Package Views & Dimensions

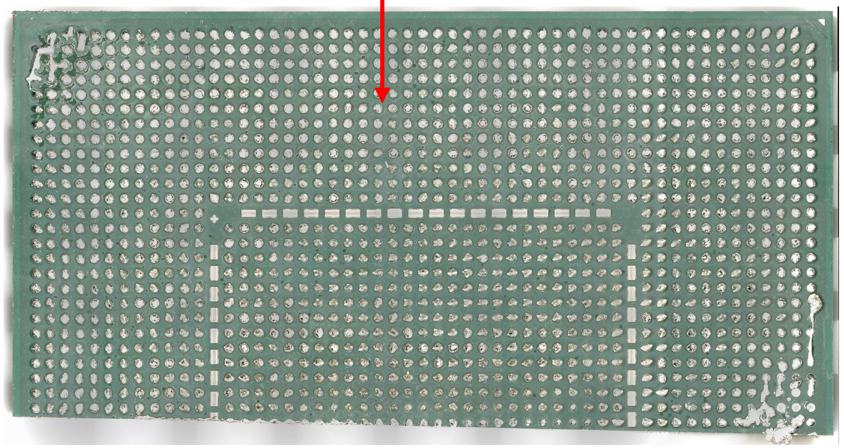
- Single package with GV100 GPU and 16GB HBM2 Memory on a silicon interposer.

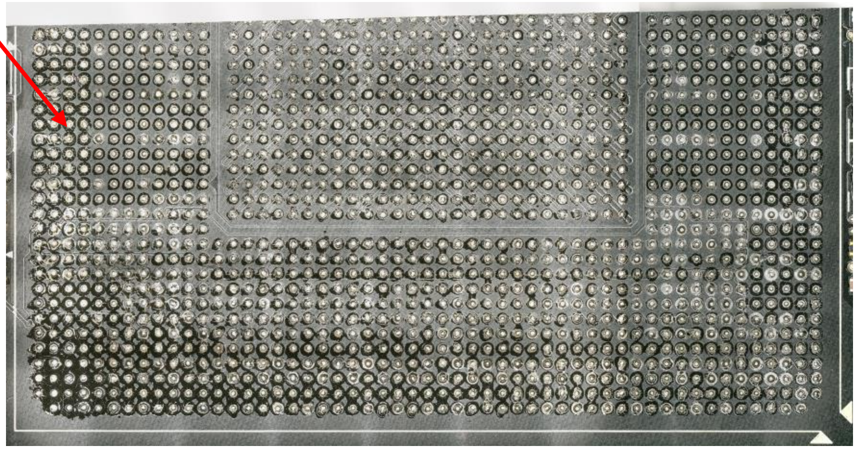


Package Top View

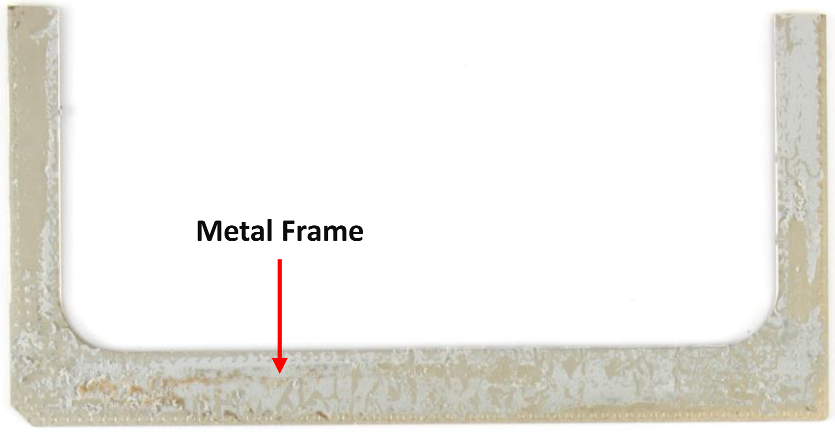
Package Views & Dimensions

- A cross-section of the package was made before unsoldering to have access to the complete assembly structure.





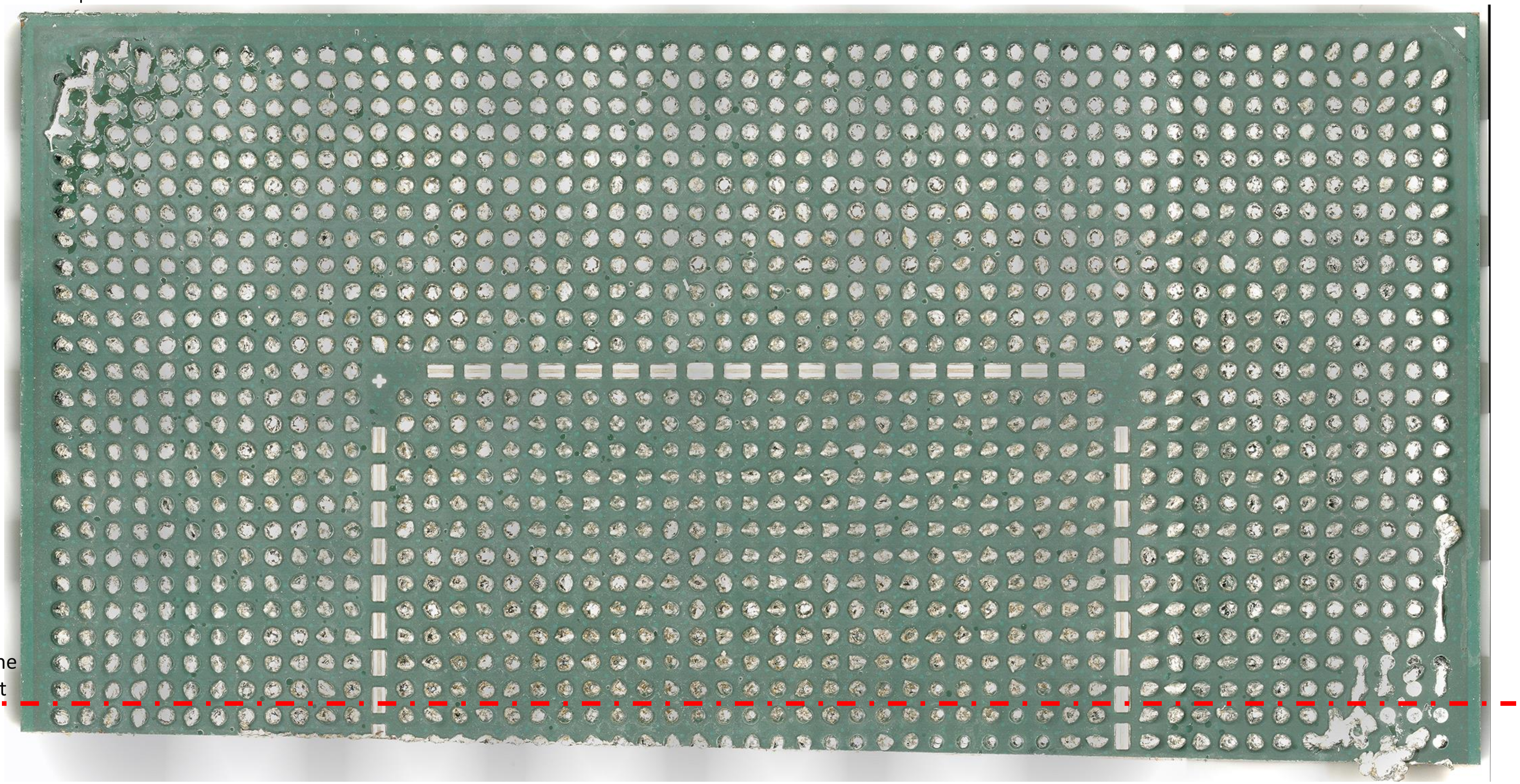
Package Unsoldered from PCB Board



Metal Frame Removed

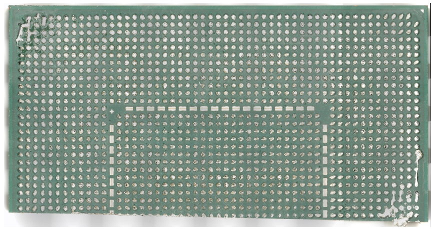
Package Views & Dimensions

- Estimated number of balls: 2,540
- Ball pitch: 1mm

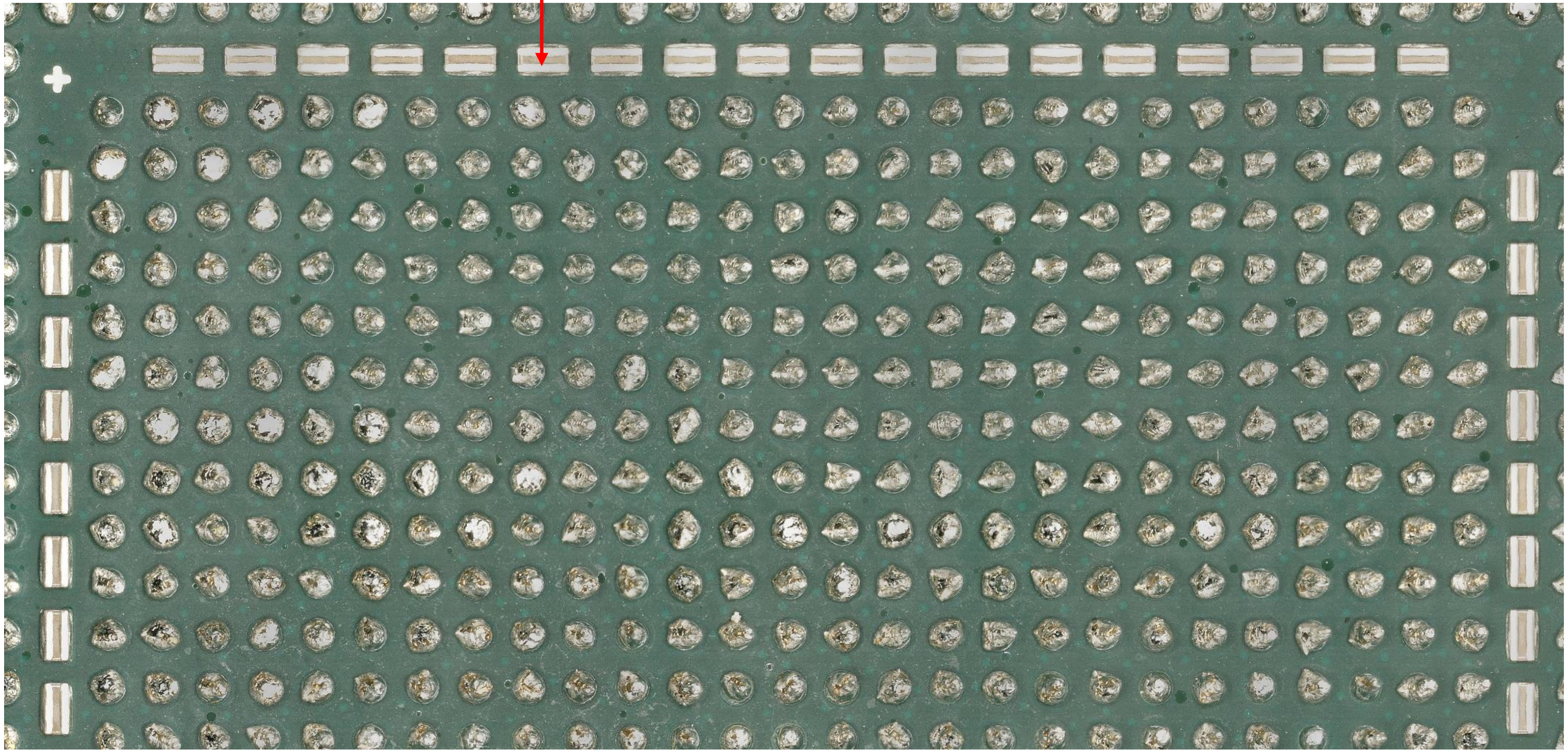


Package Bottom View after Cross-Section

Package Views & Dimensions

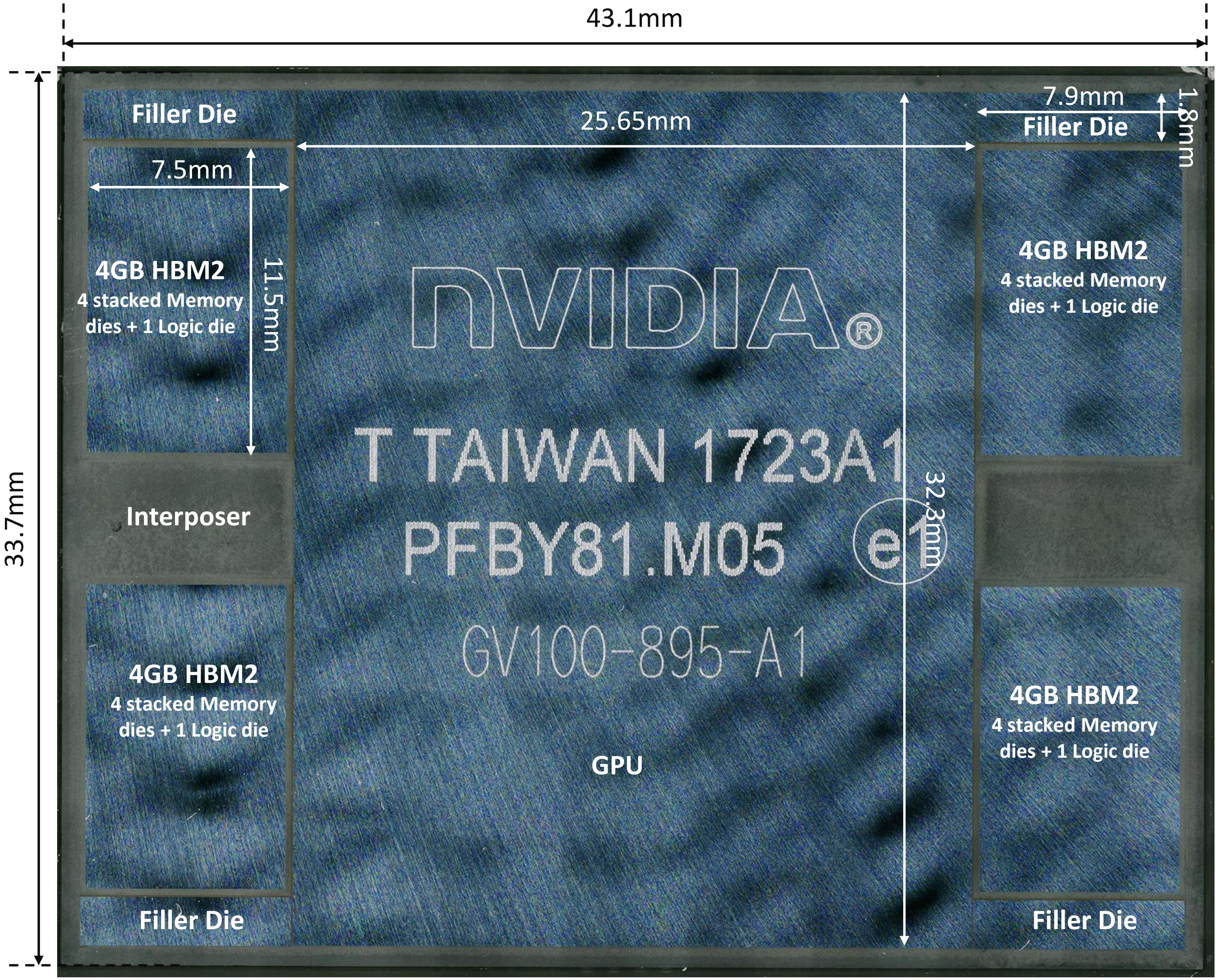


- MLCC Capacitors are soldered on the bottom side of the substrate.
  - MLCC size: 0402 (1.0x0.5mm)
  - Estimated number :66



MLCC View

Dies Size



Die Size (with scribe line):

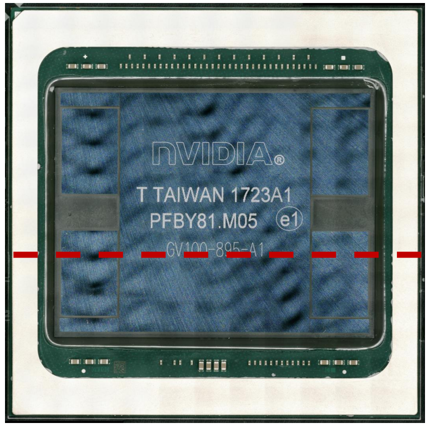
- Interposer Area: 1,452mm² (43.1x33.7mm)
- Nb of candidates per 12-inch wafer: 36
- GPU Die Area: 828.5mm² (25.65x32.3mm)
- Nb of candidates per 12-inch wafer: 64
- DRAM Dies Area: 86mm² (11.5x7.5mm)
- Nb of candidates per 12-inch wafer: 728
- Logic Dies Area: 96mm² (12.0x8.0mm)
- Nb of candidates per 12-inch wafer: 656
- Filler Die Area: 14.2mm² (7.9x1.8mm)
- Nb of candidates per 12-inch wafer: 4,396

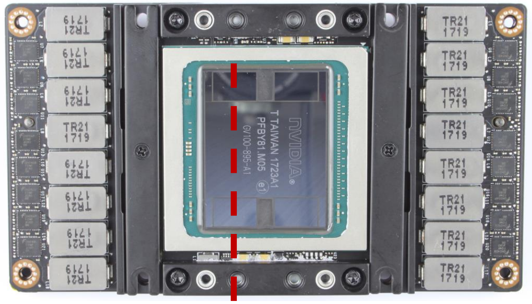
Package Opening

Package Top View – Metal Frame Removed

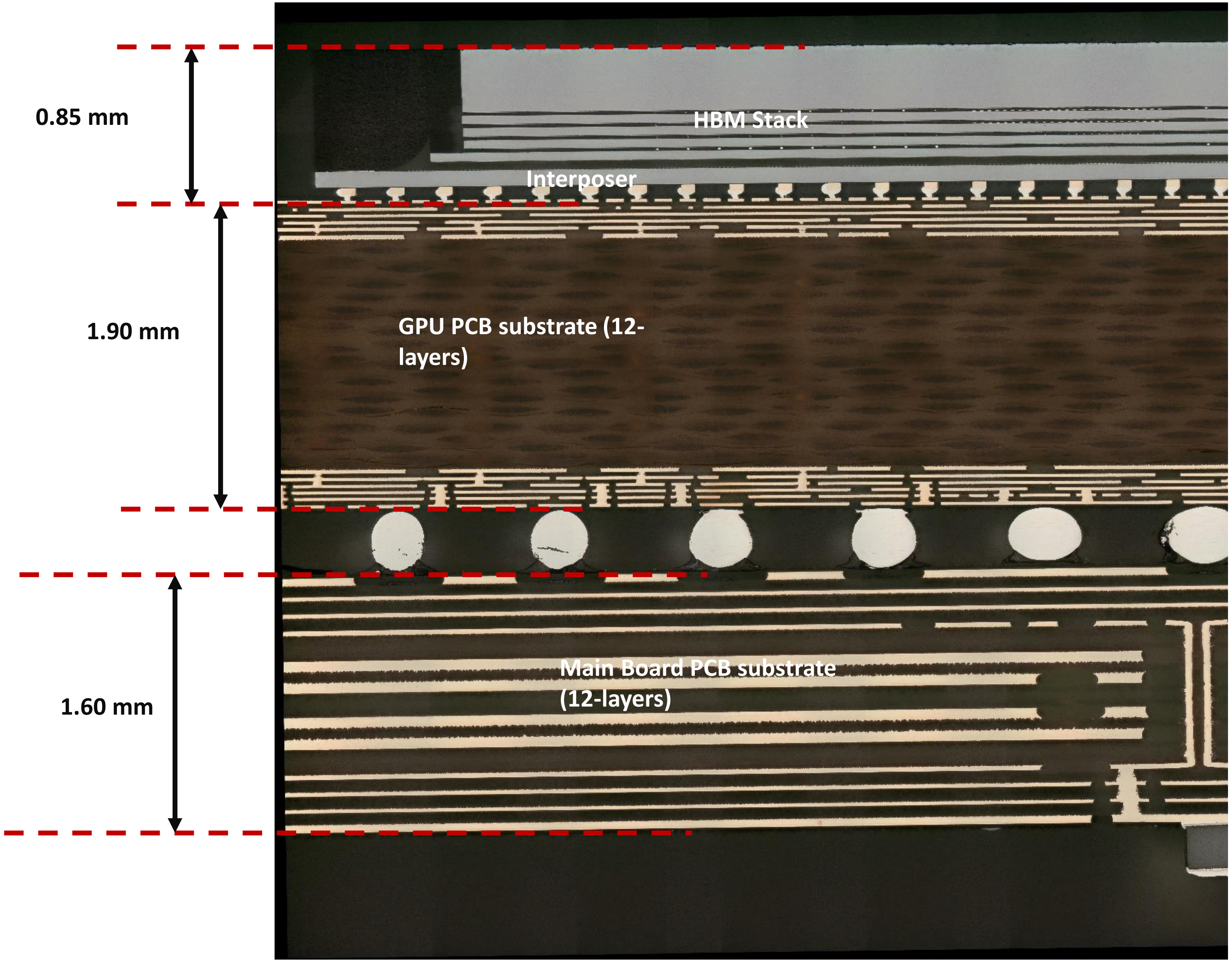
Package Top View – Underfill removed and Right HBM Memory Removed

Board Cross-Section

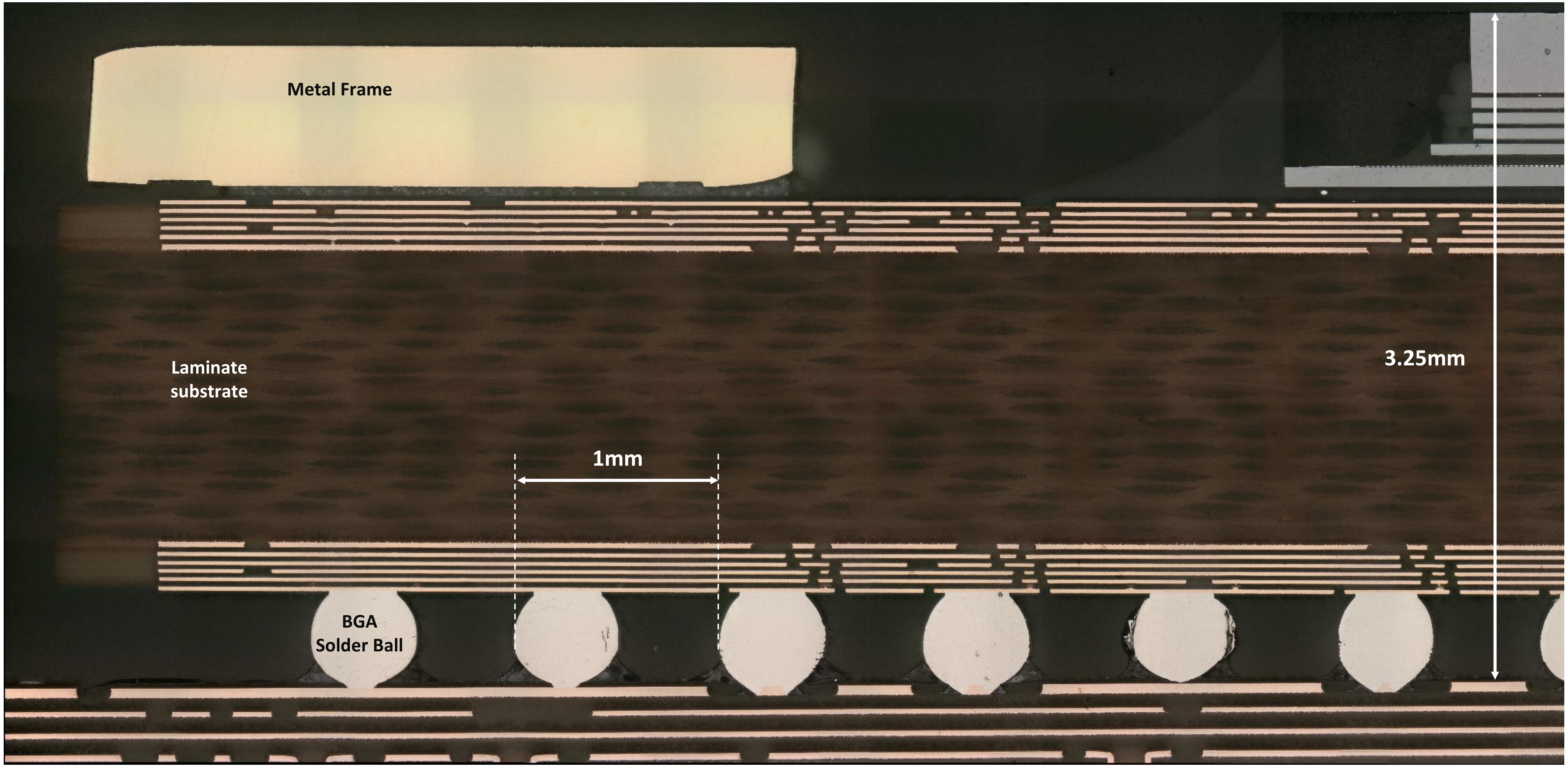




Board Cross-section plane



Board Cross-Section – Laminate Substrate



Substrate Cross-Section – Optical View
©2018 by System Plus Consulting

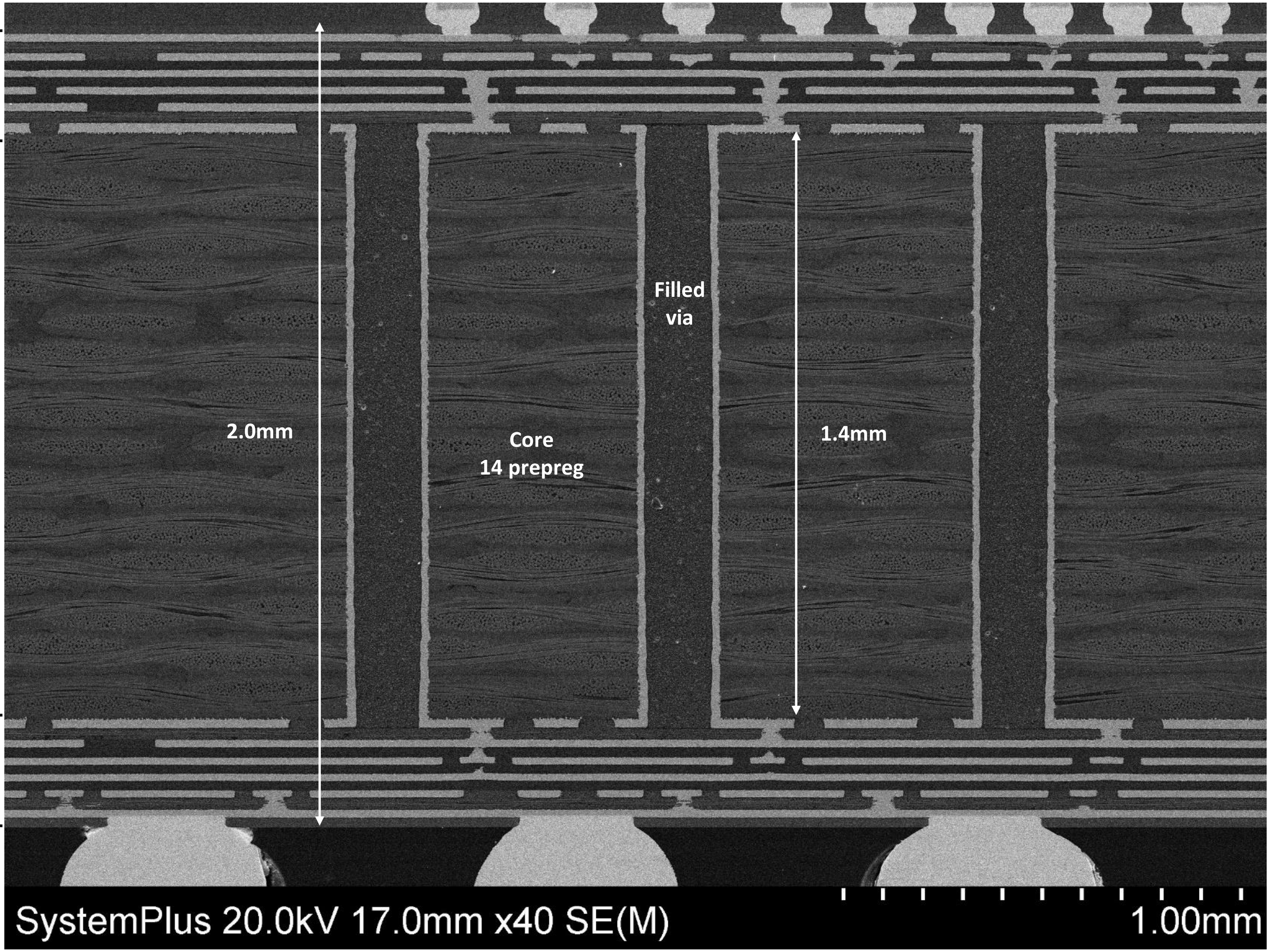
- Package total thickness: 3.25mm
- Metal frame thickness: 0.7mm

- Laminate substrate layers: 12
- Laminate substrate thickness: 2mm
- Laminate core thickness: 1.4mm

Board Cross-Section – Laminate Substrate

- The package laminate is a 12 layers HDI PCB.
  - ✓ PCB thickness: 2mm
  - ✓ Copper layers thickness: 18µm
  - ✓ Microvia diameter: 65µm

6 copper layers

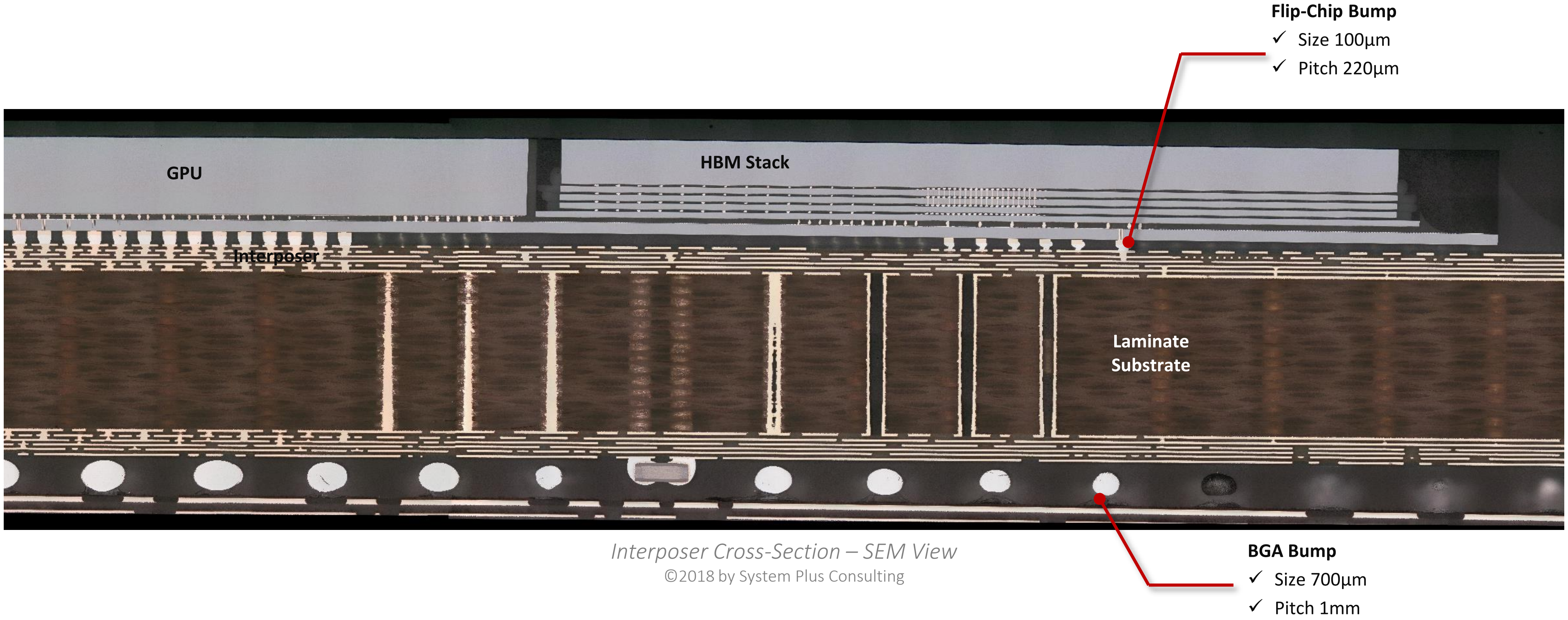


6 copper layers

Substrate Cross-Section – SEM View

Board Cross-Section – Interposer

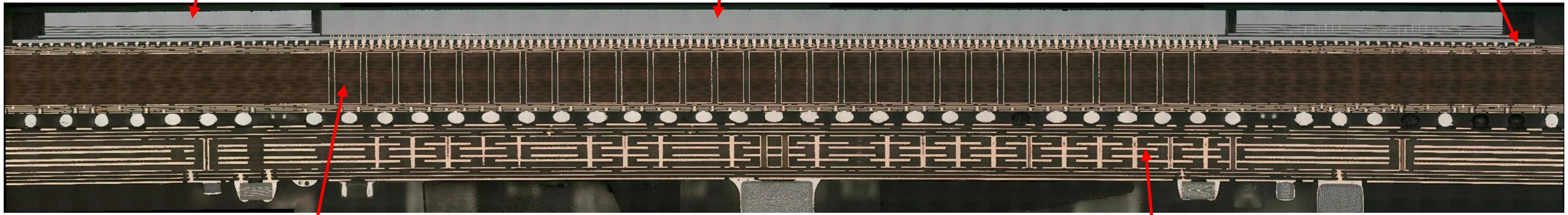
- HBM stack and GPU die are bonded on an interposer which is flip-chipped to the PCB substrate.



Interposer Cross-Section – SEM View

- BGA bump pitch: 1000µm
- BGA ball diameter: 700µm
- Flip-Chip bump pitch: 220µm
- Flip-Chip bump diameter: 120µm

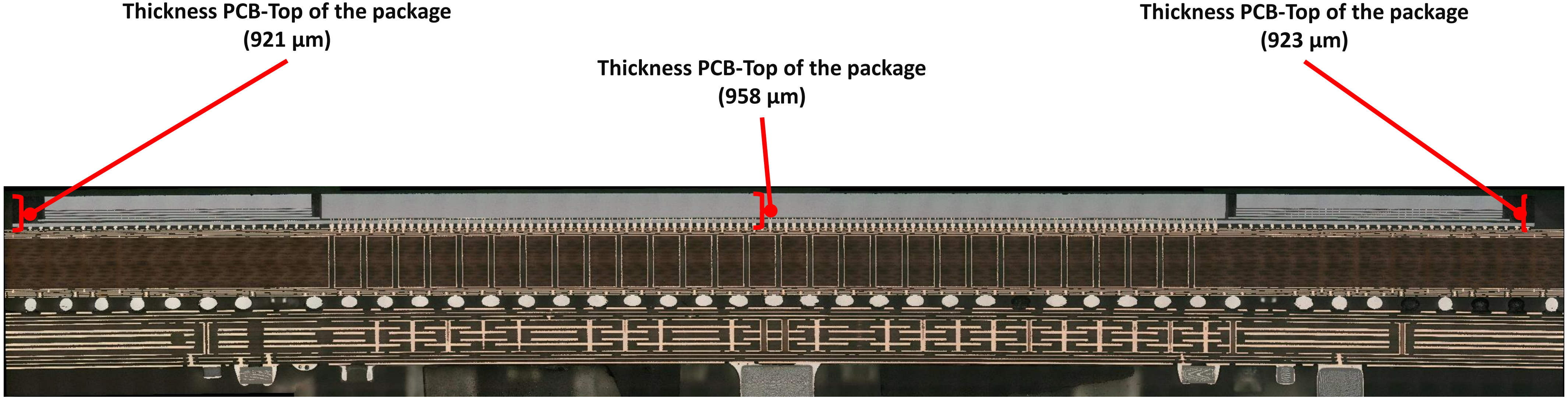
Board Cross-Section



Package Cross-Section – Optical View

©2018 by System Plus Consulting

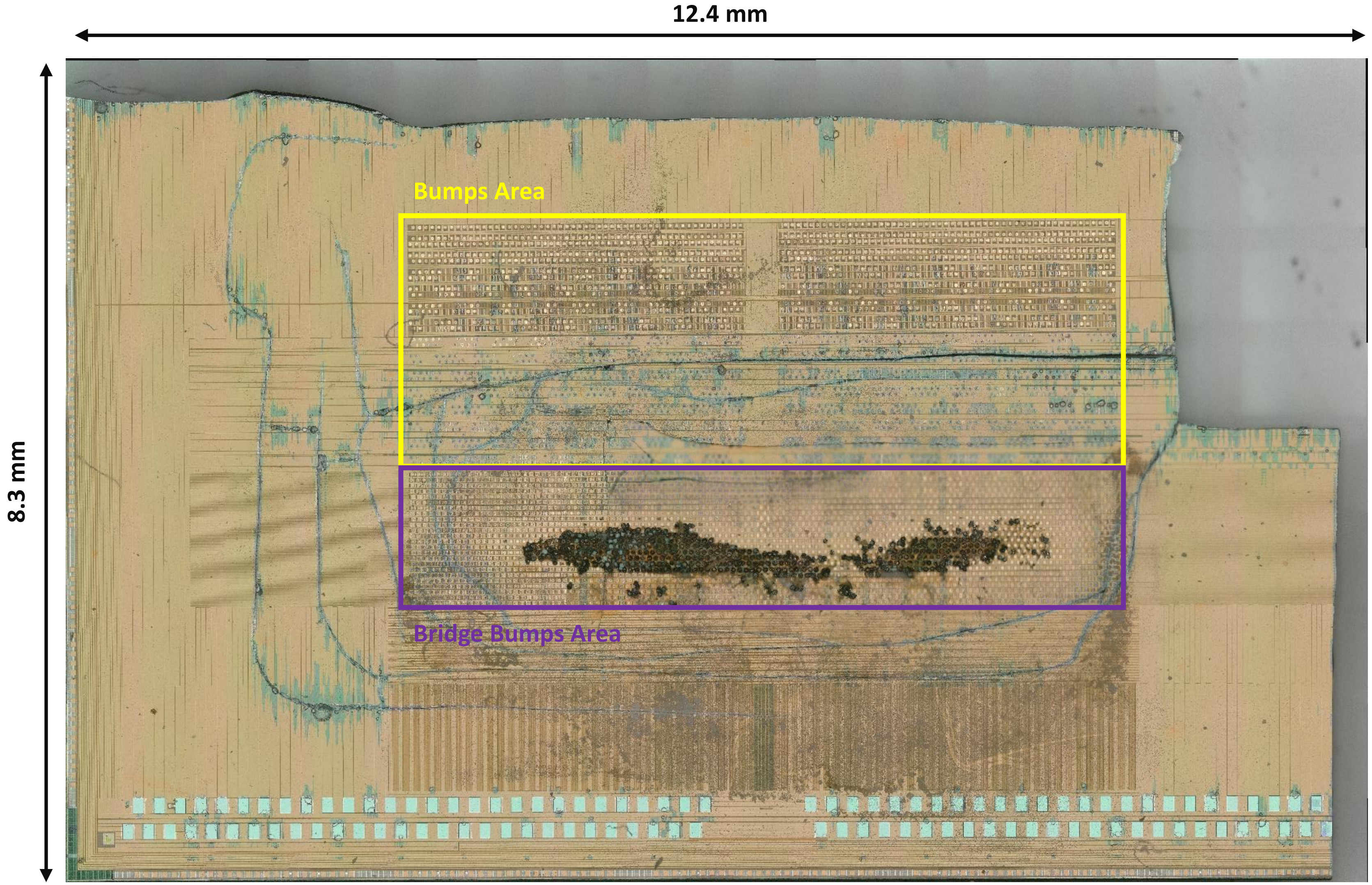
Board Cross-Section



Package Cross-Section – Optical View

- Warpage Estimation: 40 µm

Samsung 1GB HBM2 – Driver Die View & Dimensions



Die Overview

©2018 by System Plus Consulting

Die Area: 102.9 mm² (12.4x8.3 mm)

Nb of PGDW per 12-inch wafer: 604

Pad number: 116

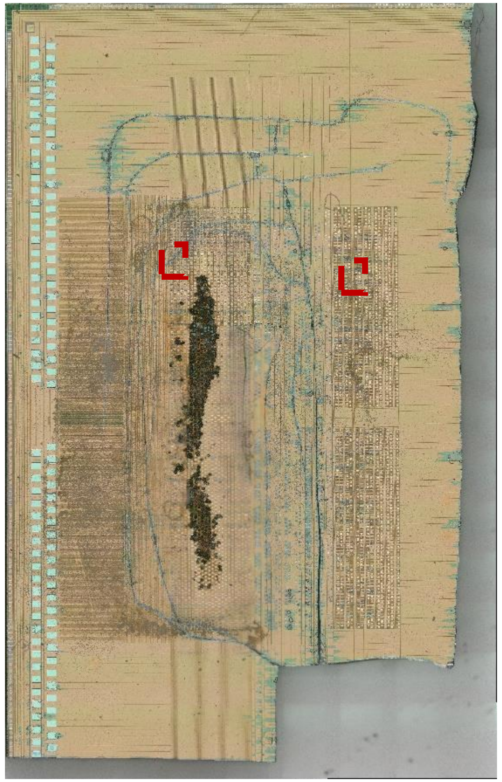
Bridge Bumps Number: 2,400

Bridge Bumps Area: 6.76 mm² (6.04 x 1.12 mm)

Bridge Bumps Fill Factor: 6.5 %

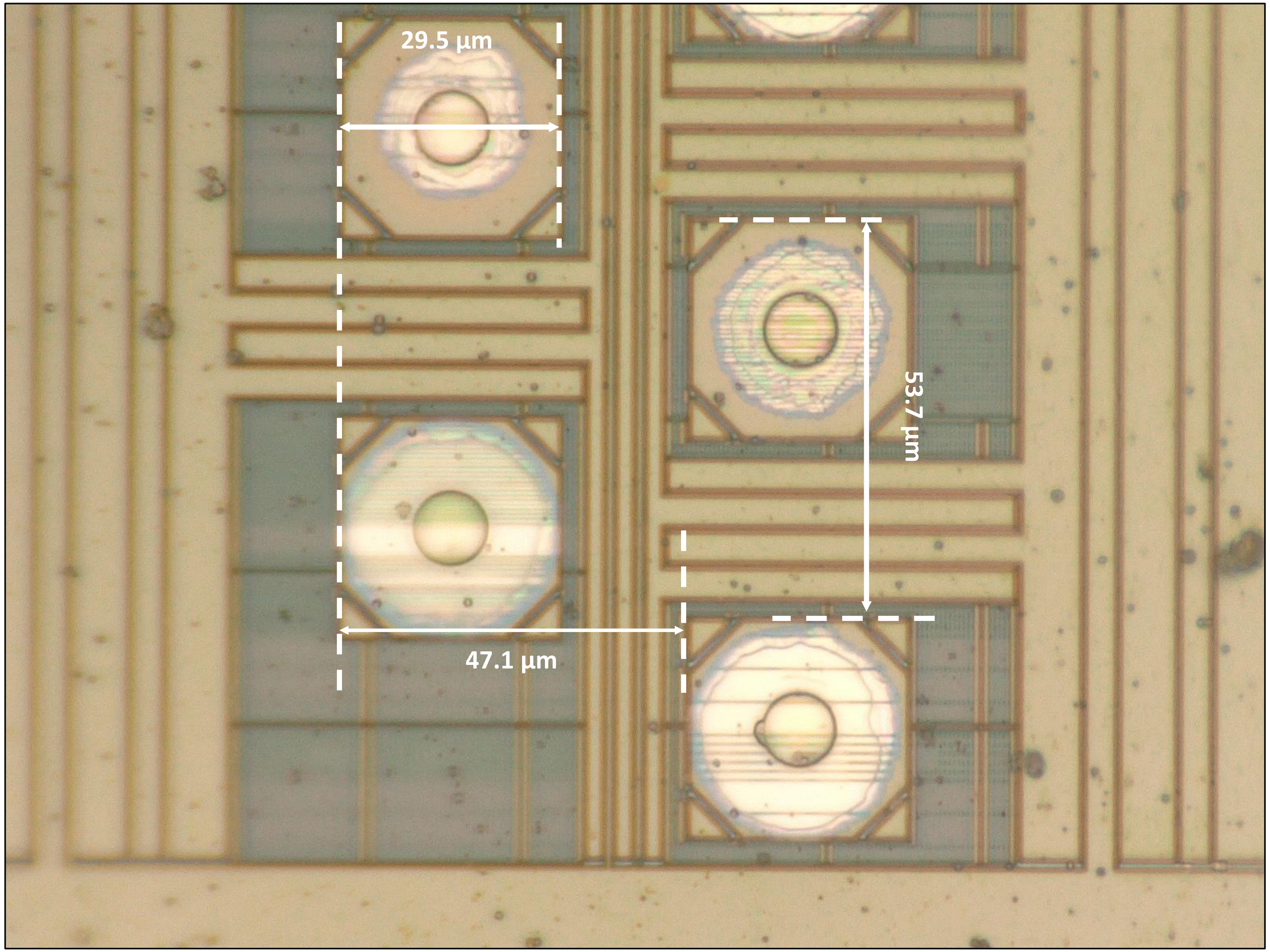
Bumps Number: 1,143

Samsung 1GB HBM2 – Driver Bumps



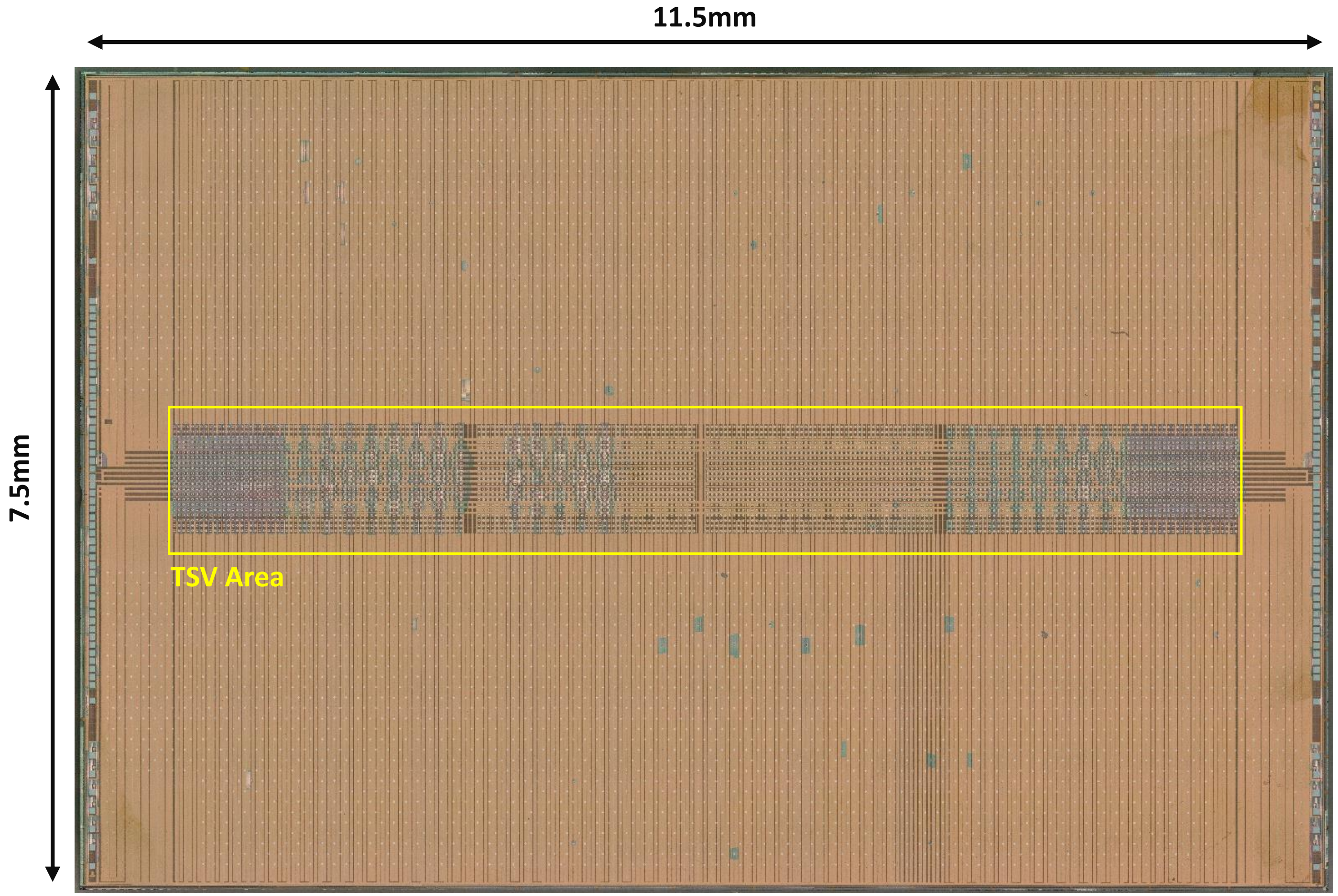
Die Overview
©2018 by System Plus Consulting

- µBump pitch: 47.1 µm
- µBump diameter: 29.5 µm



Bumps Areas
©2018 by System Plus Consulting

Samsung 1GB HBM2 – DRAM Die View and Dimensions



Die Overview

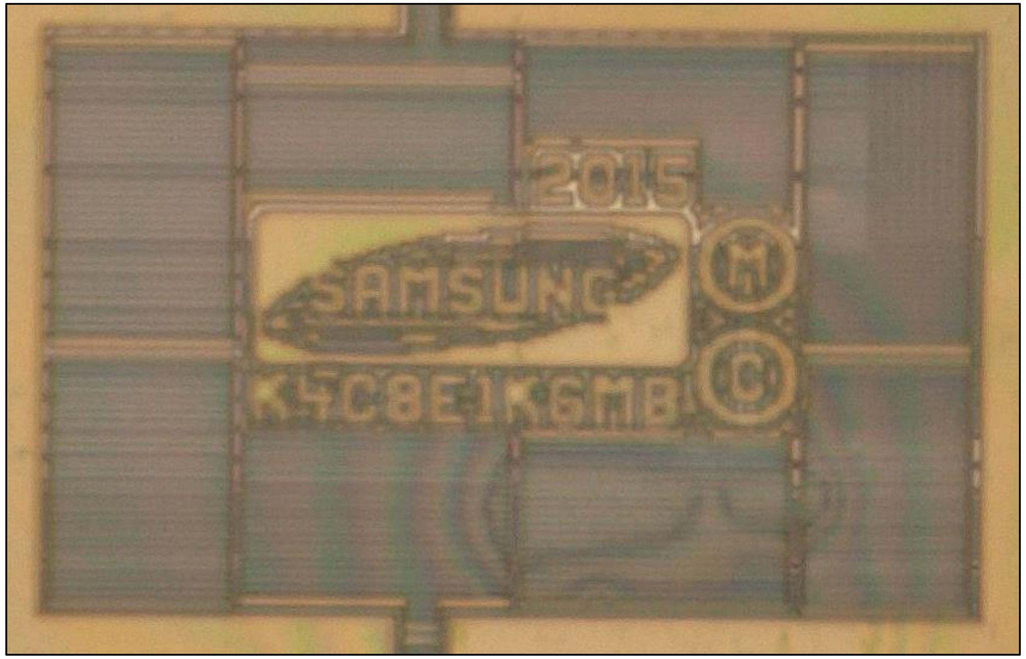
Die Area: 86mm² (11.5x7.5mm)

Nb of PGDW per 12-inch wafer: 728

Pad number: 126

TSV number: 4,830

Samsung 1GB HBM2 – DRAM Die Marking



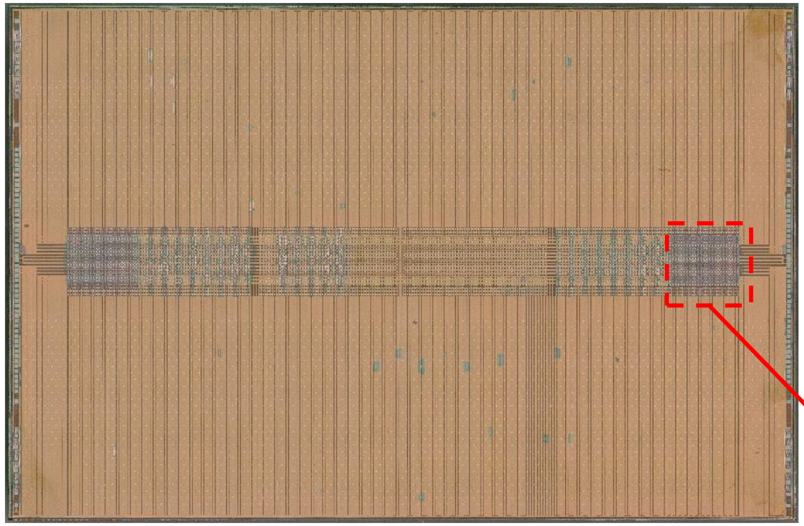
Die Marking

The die marking includes the logo of Samsung and :

K4C8E1K6MB

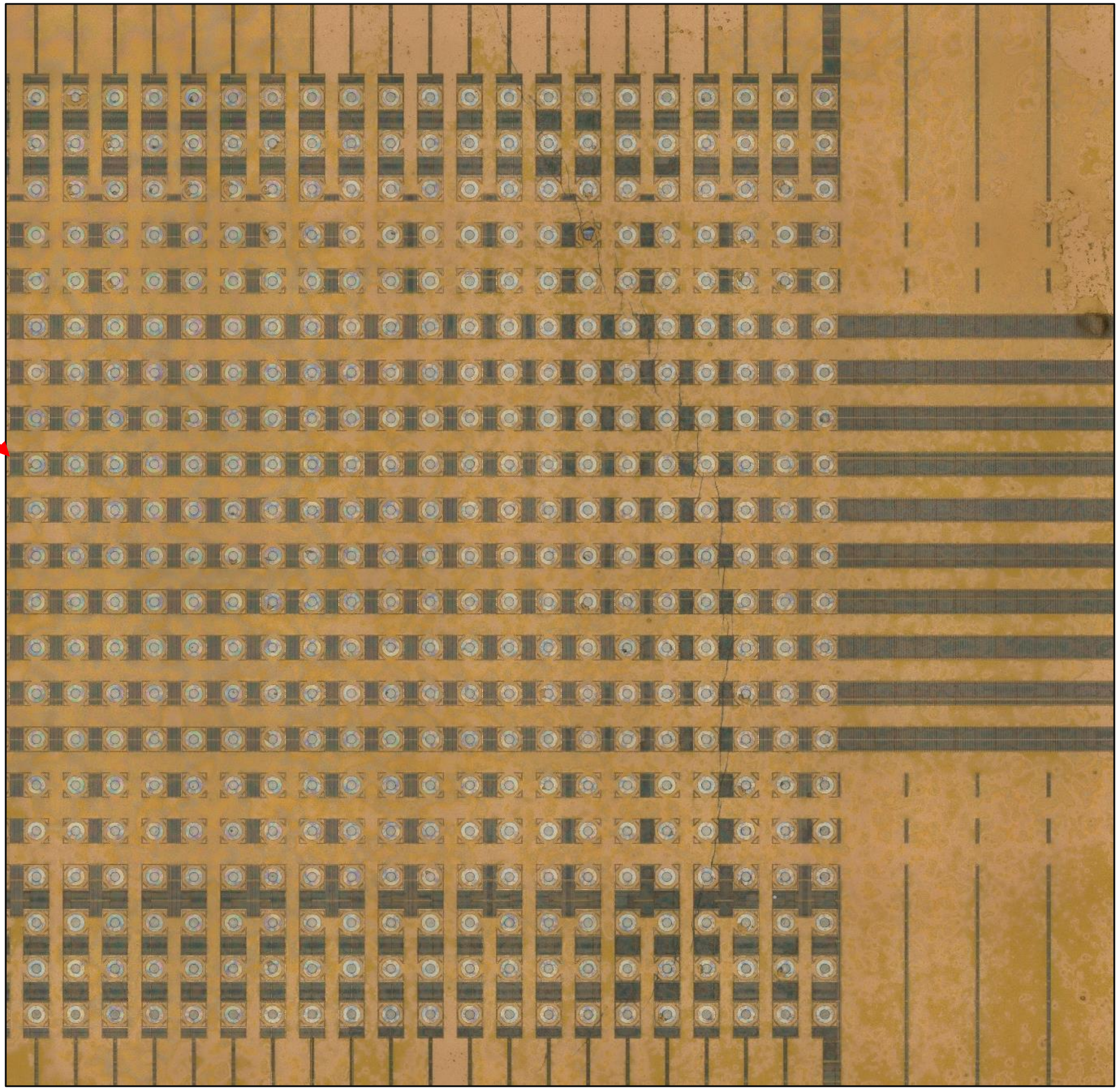
2015

Samsung 1GB HBM2 – DRAM Die – µBumps & TSVs



Die Overview

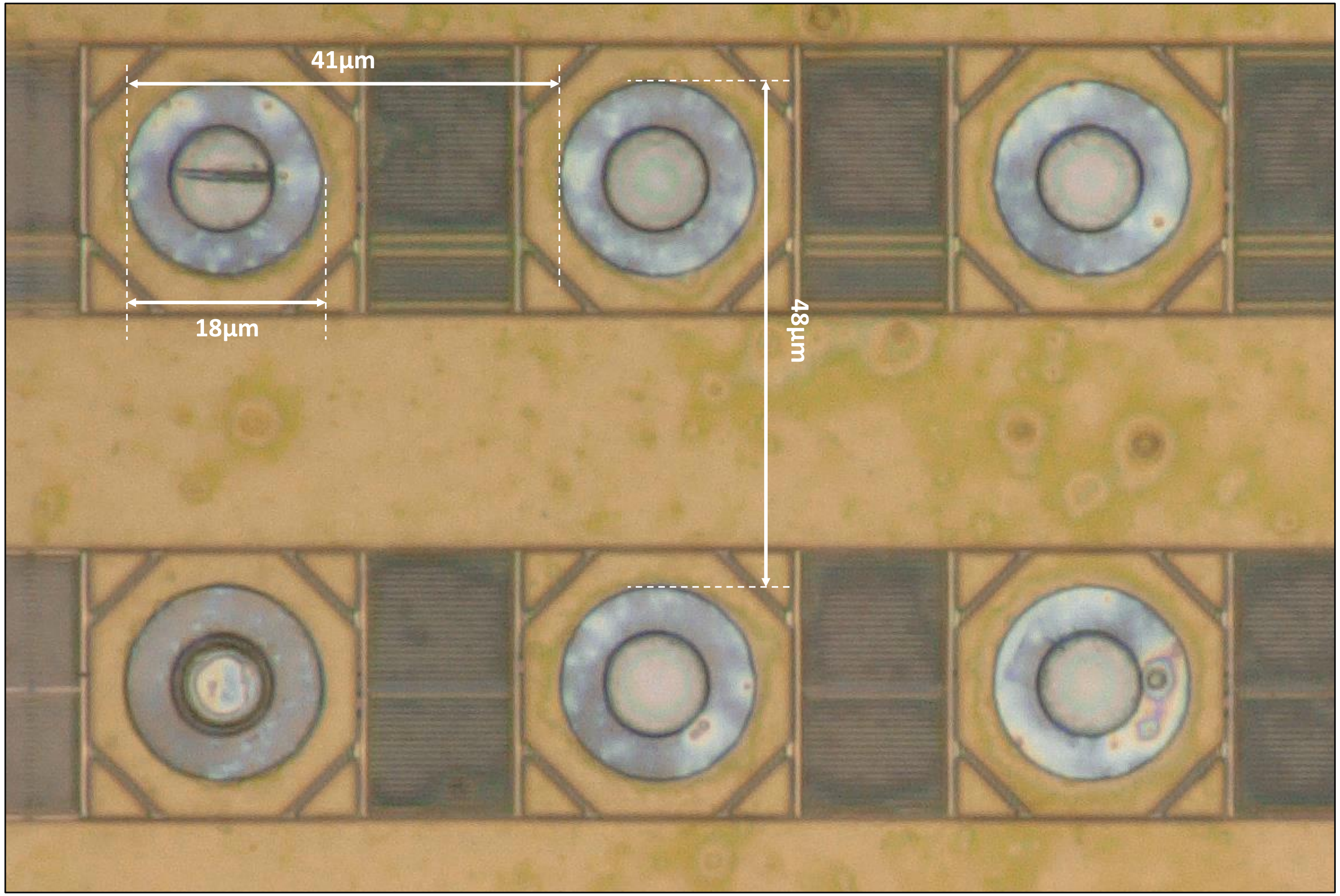
- TSVs are located at the center of the dies.



TSVs Areas

Samsung 1GB HBM2 – DRAM Die – µBumps & TSVs

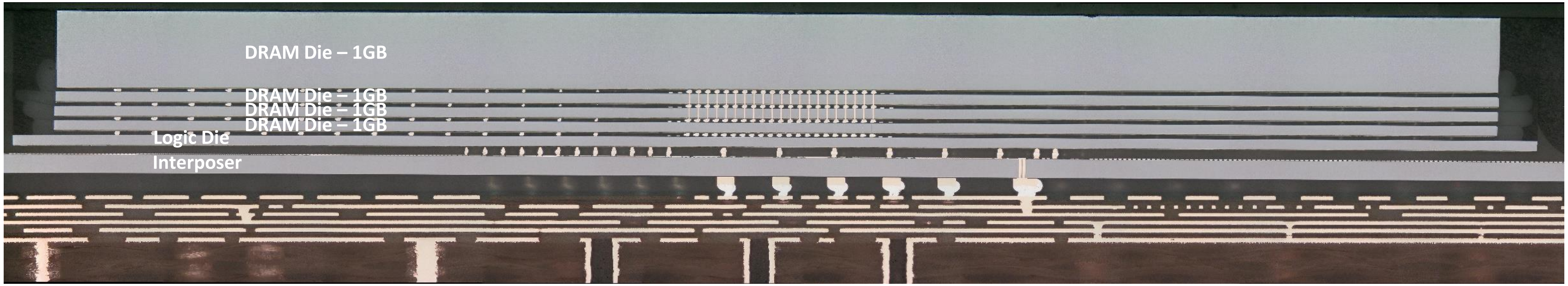
- µBump & TSV pitch: **41µm**
- µBump diameter: **18µm**



TSVs Areas

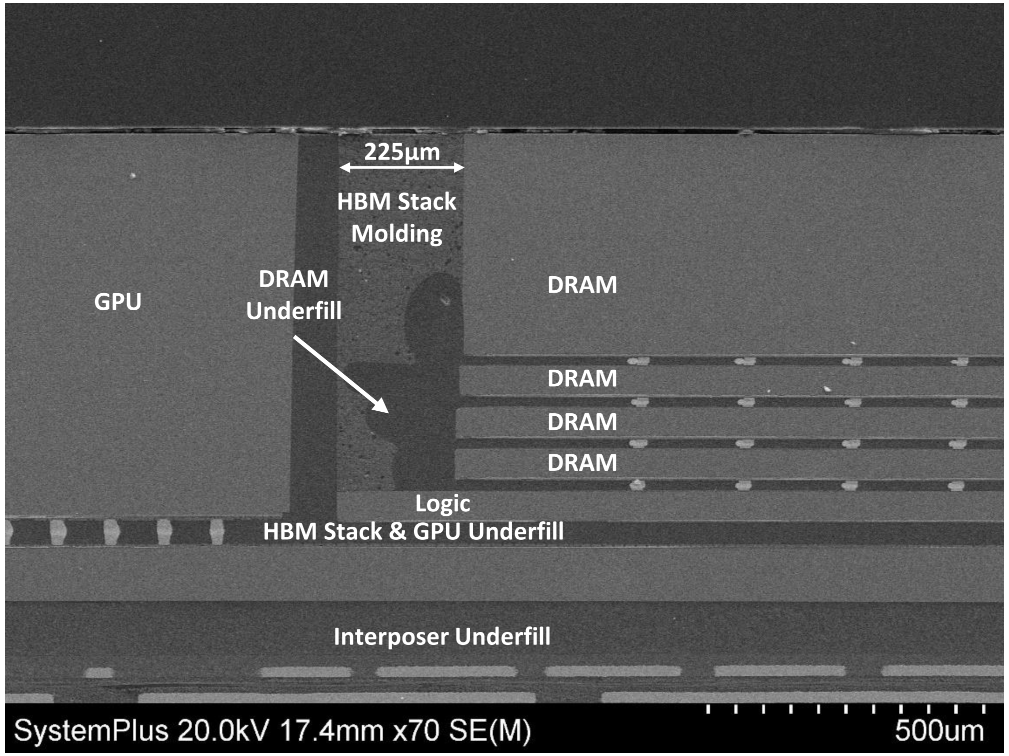
Package Cross-Section – HBM2 Stack

- HBM stacks are flip-chipped on the interposer at the wafer-level though microbumps.
- HBM stacks include 5 dies: 4 1GB DRAM + 1 logic (buffer) die

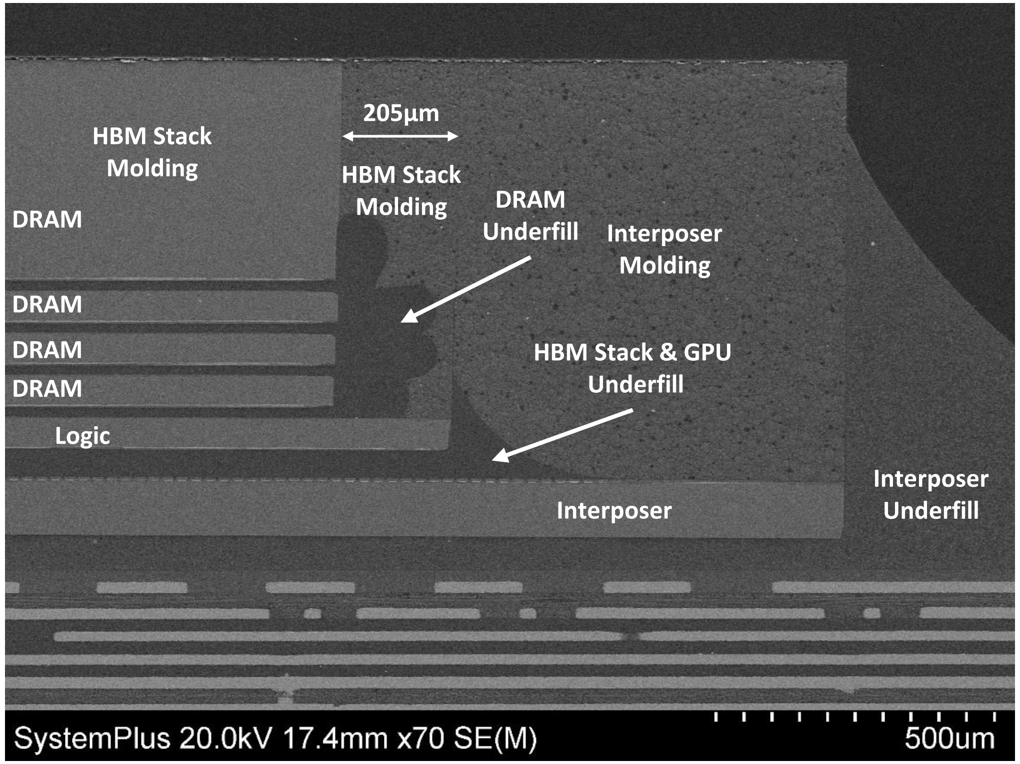


HBM Stack Cross-Section – Optical View
©2018 by System Plus Consulting

Package Cross-Section – HBM2 Stack



HBM Stack Cross-Section – SEM View



HBM Stack Cross-Section – SEM View

- The HBM stack is molded on the side.
- The side mold is 205-225µm wide.
- DRAM dies do not share exactly the same size, they are diced before being bonded together.

Package Cross-Section – HBM2 Stack

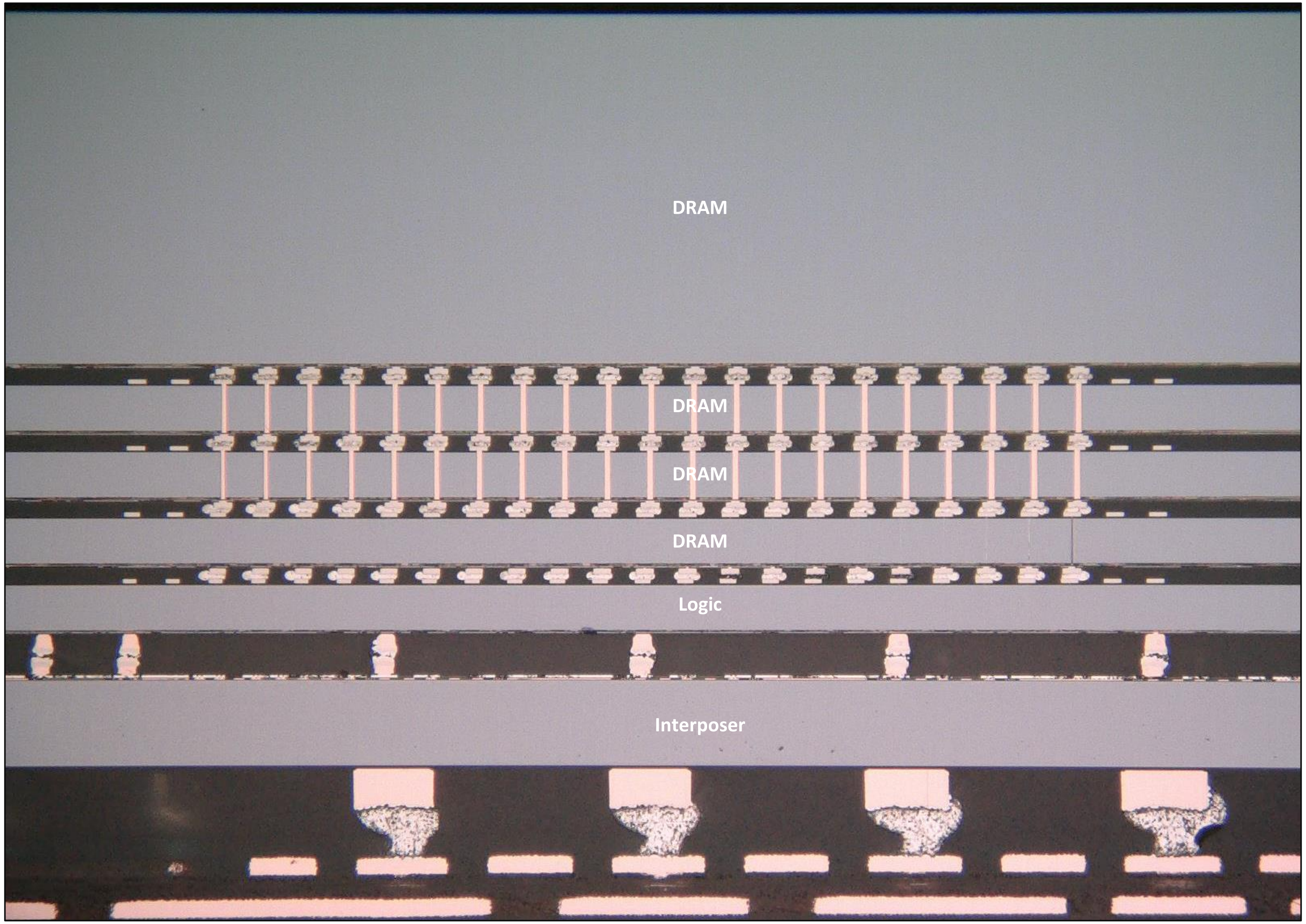
HBM Stack

4x DRAM Die + 1x Logic Die

TSV and microbumps connection

Interposer

TSV, redistribution layers and microbumps connections



HBM Stack Cross-Section – Optical View

Package Cross-Section – HBM2 Stack

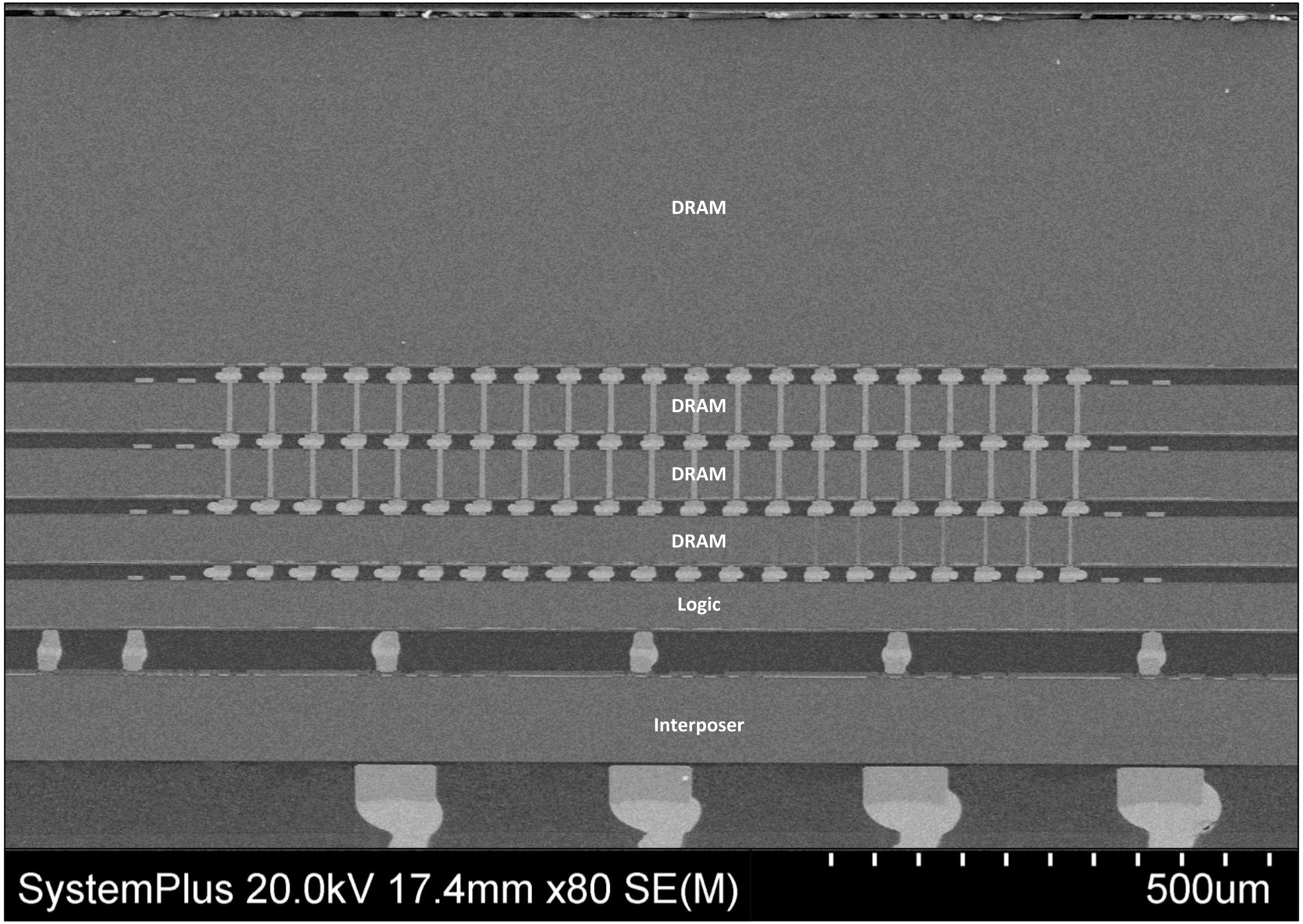
HBM Stack

4x DRAM Die + 1x Logic Die

TSV and microbumps connection

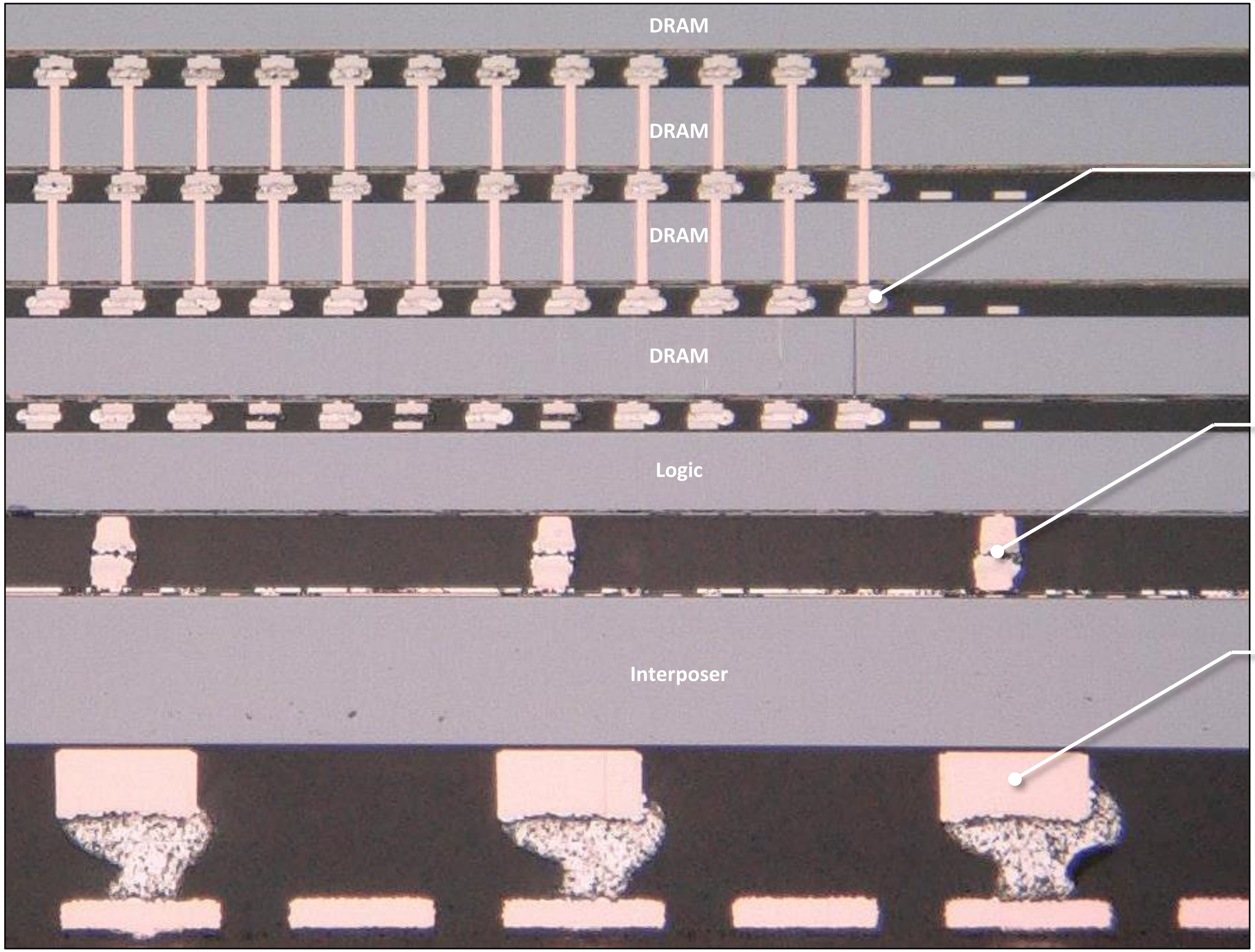
Interposer

TSV, redistribution layers and microbumps connections



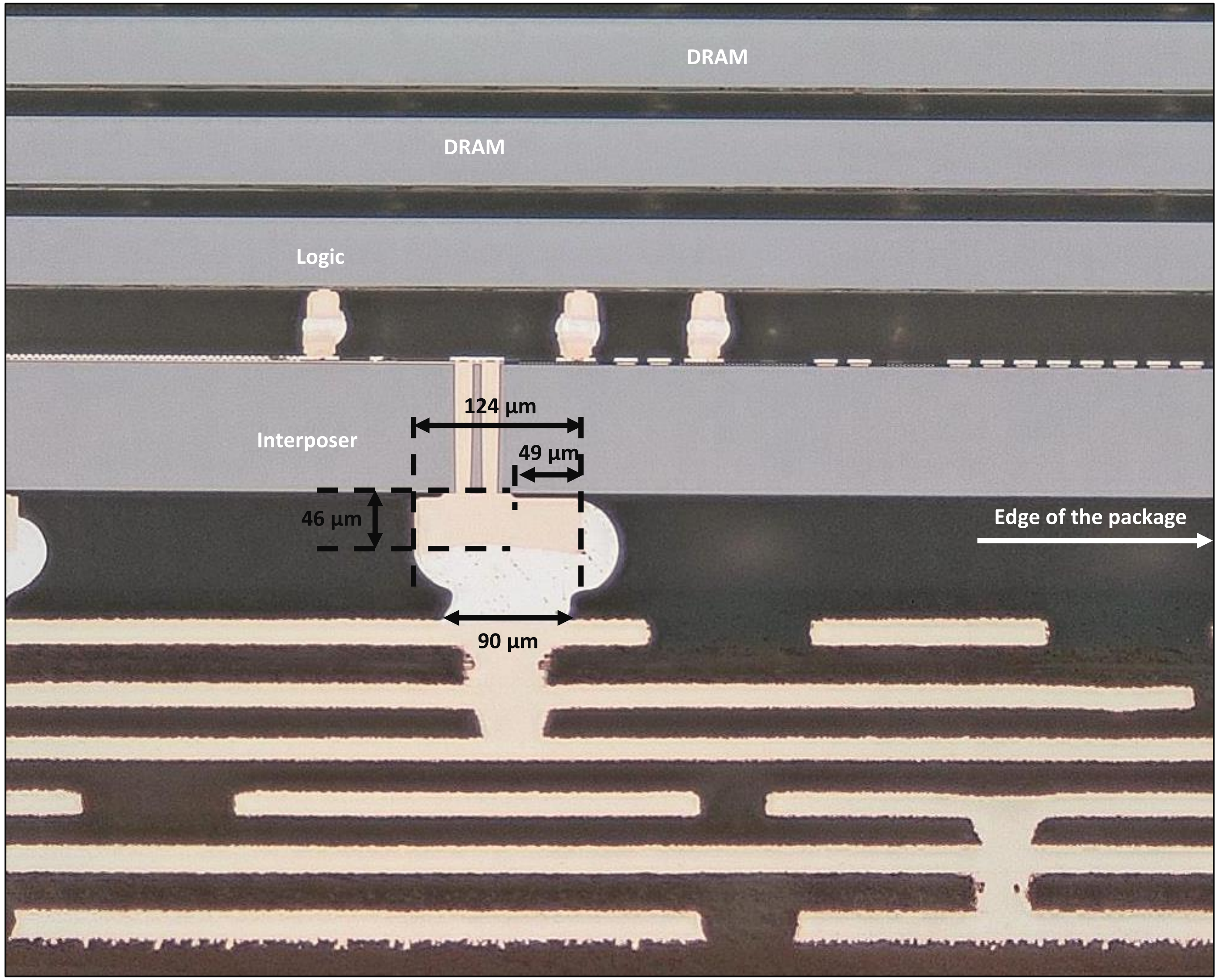
HBM Stack Cross-Section – SEM View

Package Cross-Section – HBM2 Stack



HBM Stack Cross-Section – Optical View

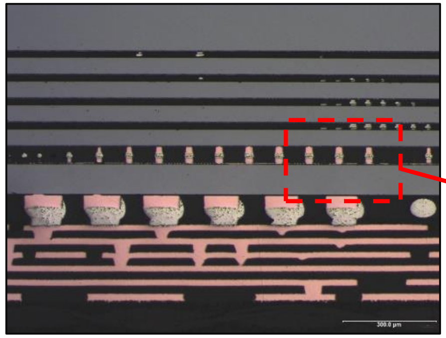
Package Cross-Section – Substrate – Interposer

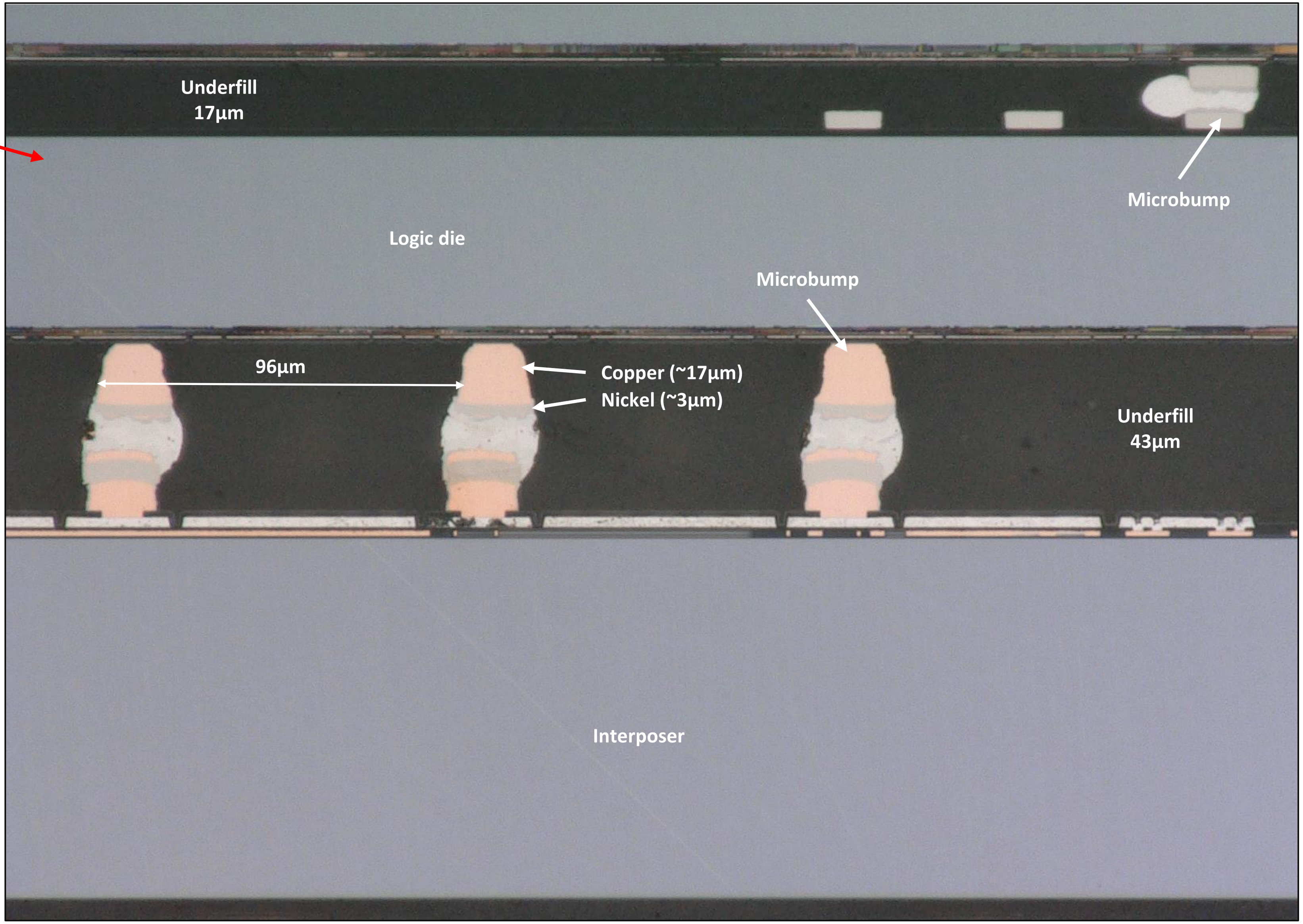


HBM Stack Cross-Section – Optical View

- Under the HBM memory, the bump are larger than under the GPU.
- We assume that the bumps are tweaks to help manage the warpage.
- The bump and the polyimide extends 25 µm in the edge direction.

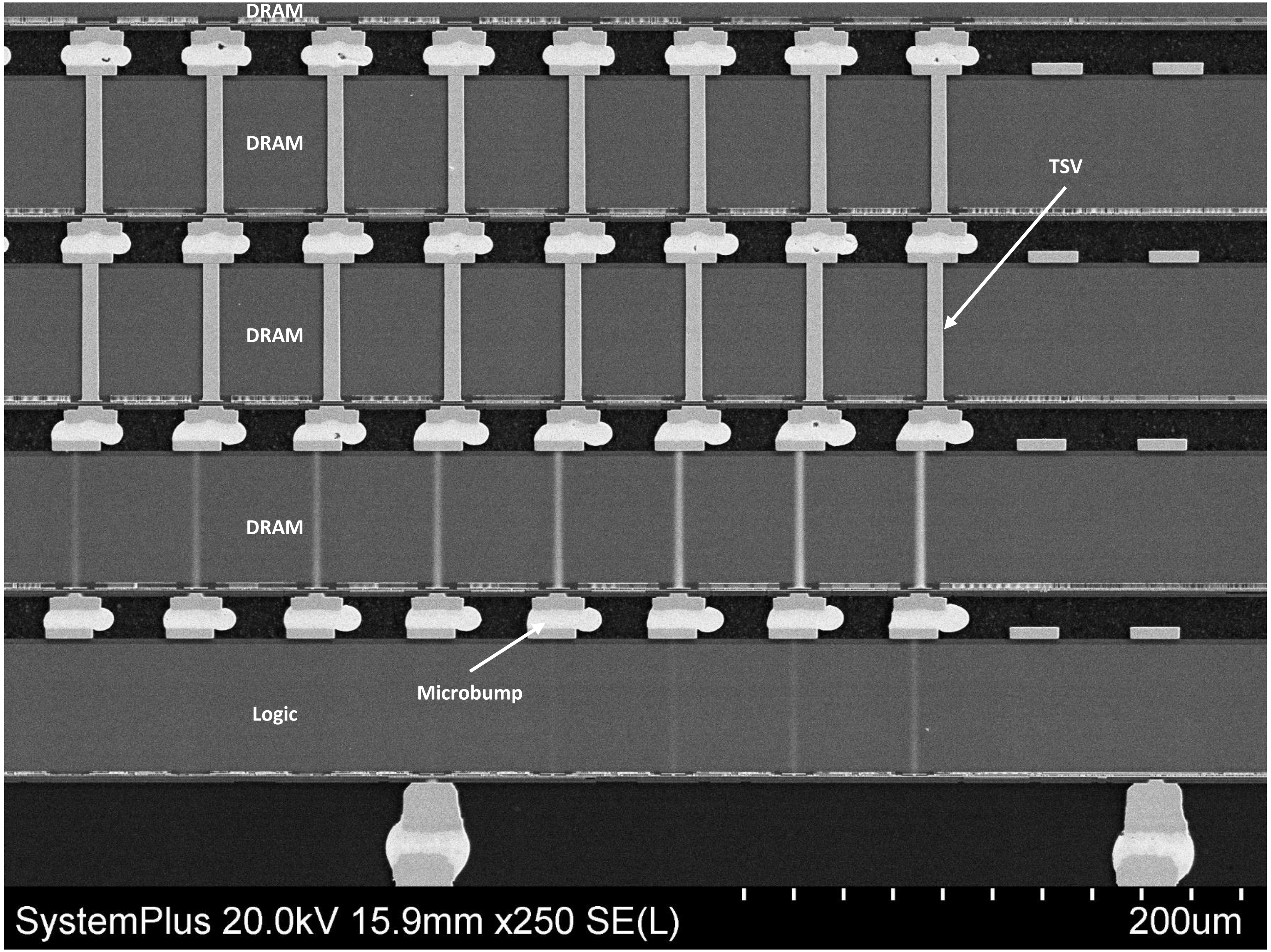
Package Cross-Section – Interposer – HBM2 Stack





HBM Stack Cross-Section – Optical View

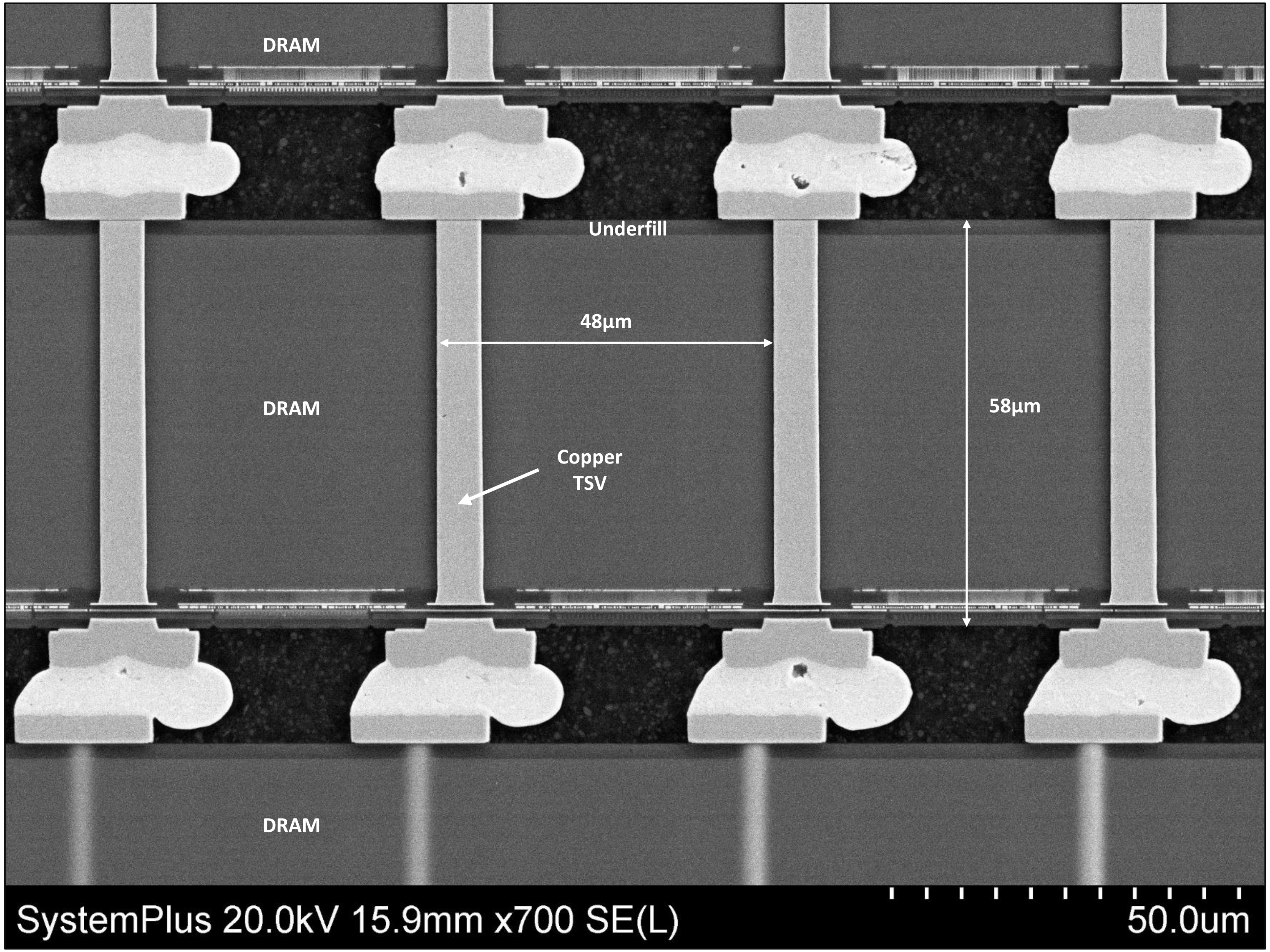
Package Cross-Section – HBM2 Stack



HBM Stack Cross-Section – SEM View

Package Cross-Section – HBM2 Stack

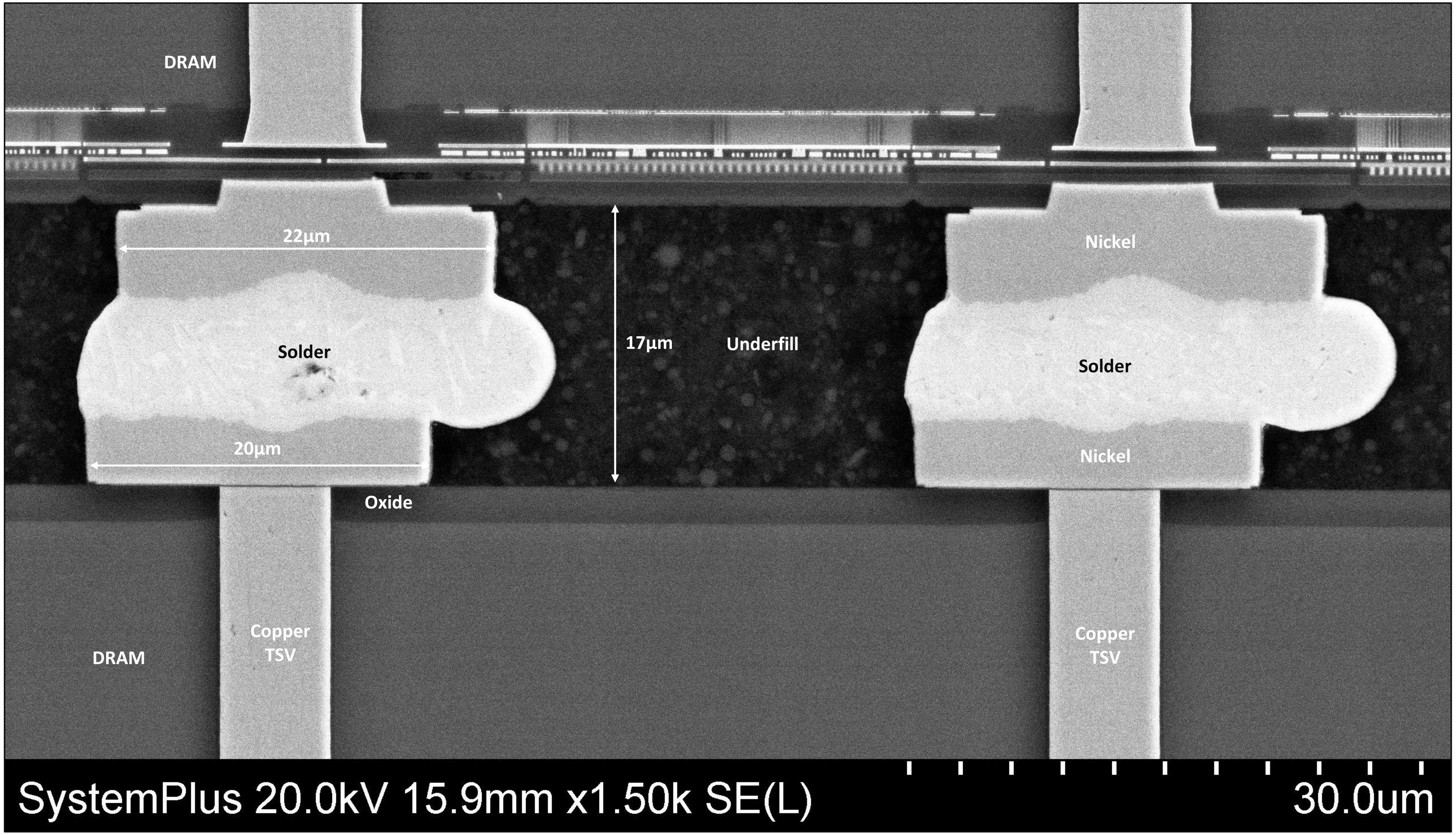
- HBM dies thickness (excepted top die): **56µm**
- HBM stack TSV & micro-bumps pitch: **48µm**
- Underfill thickness: **17µm**



HBM Stack Cross-Section – SEM View

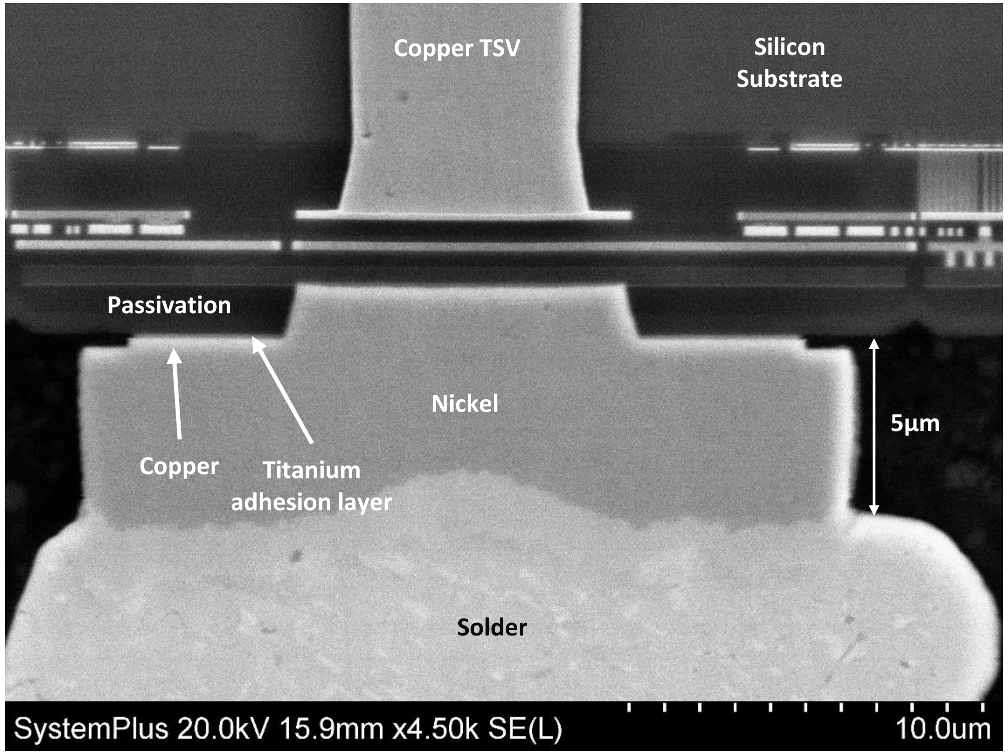
©2018 by System Plus Consulting

Package Cross-Section – HBM2 Stack

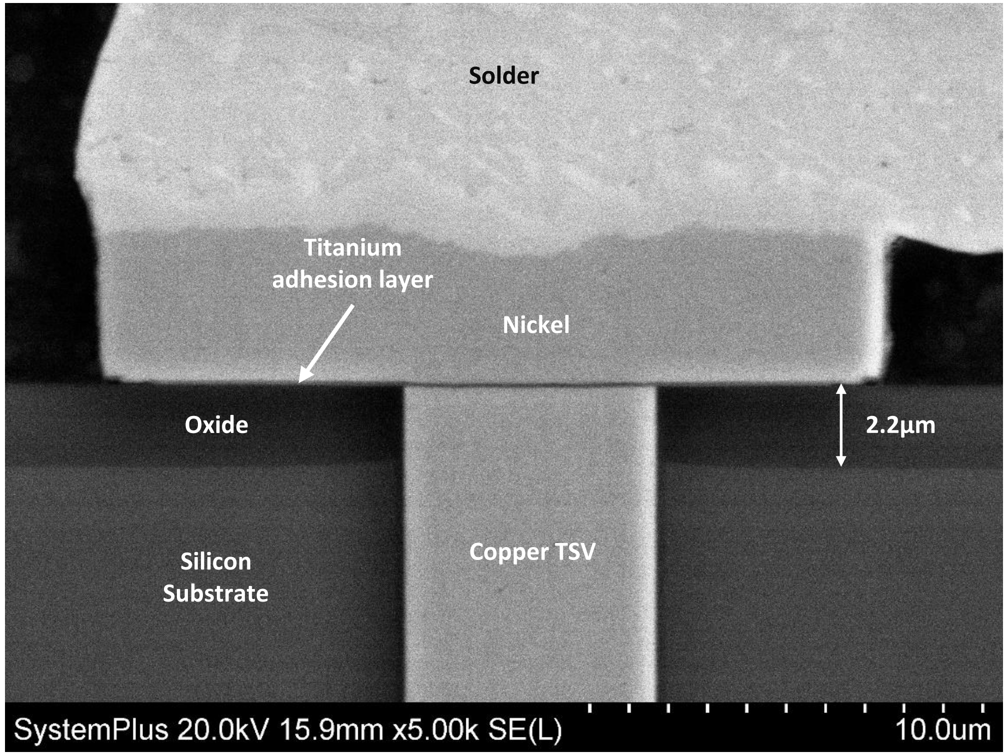


HBM Stack Cross-Section – SEM View
©2018 by System Plus Consulting

Package Cross-Section – HBM2 Stack

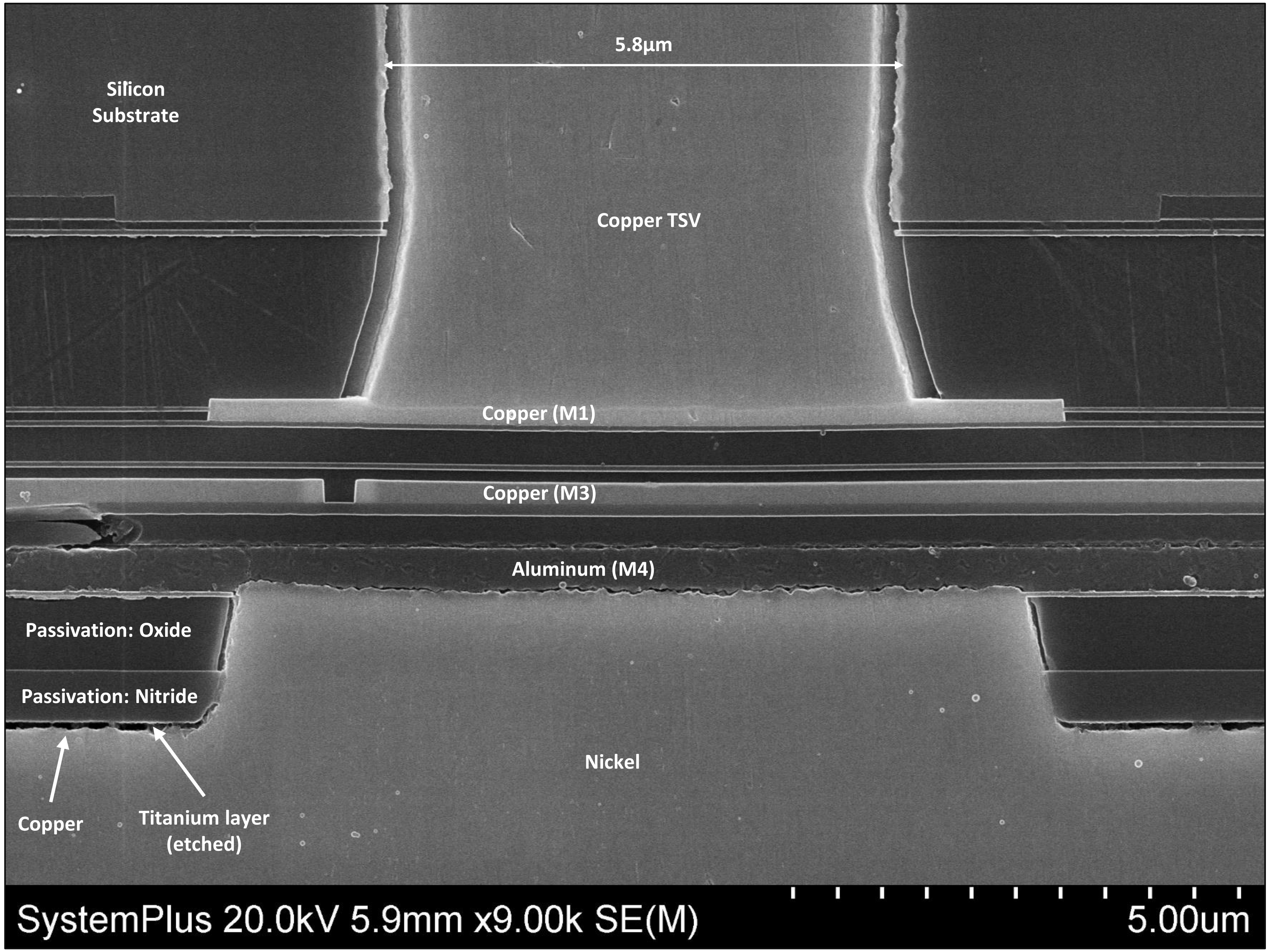


HBM Stack Cross-Section – SEM View



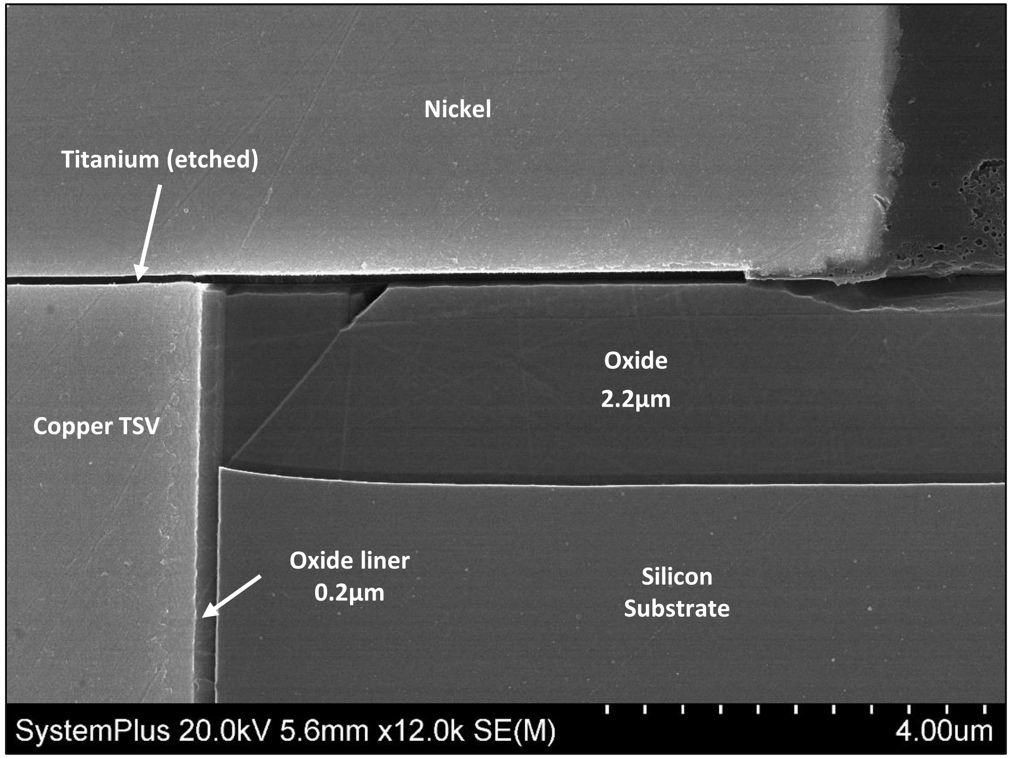
HBM Stack Cross-Section – SEM View

Package Cross-Section – HBM2 Stack



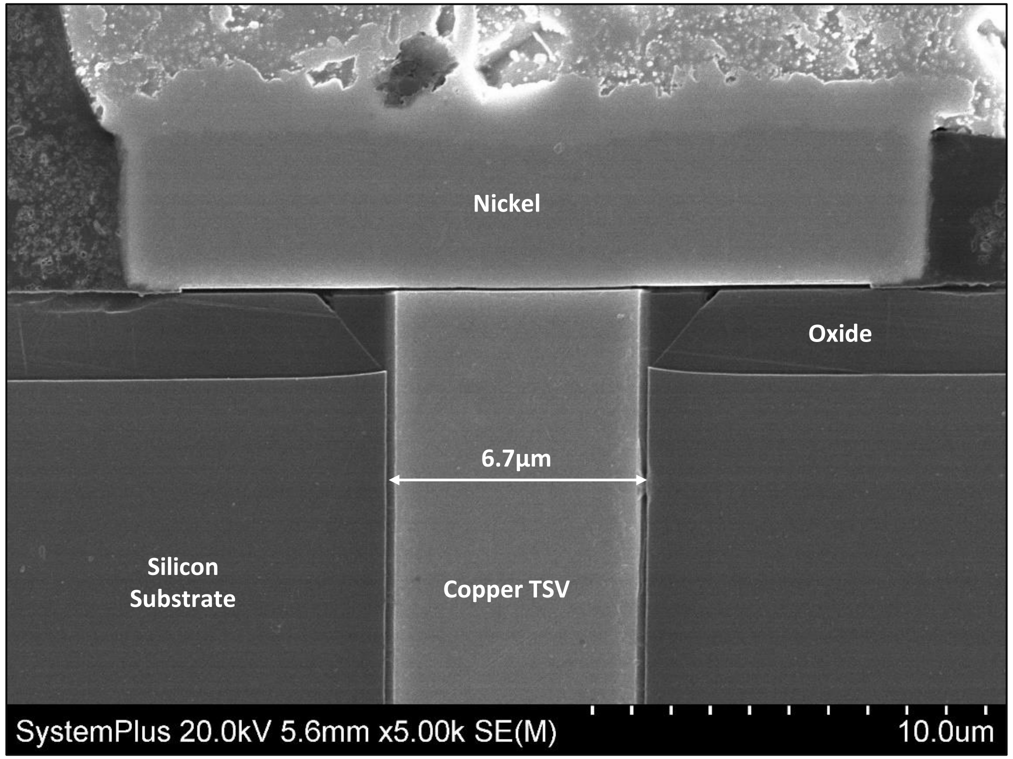
HBM Stack Cross-Section – SEM View

Package Cross-Section – HBM2 Stack



HBM Stack Cross-Section – SEM View

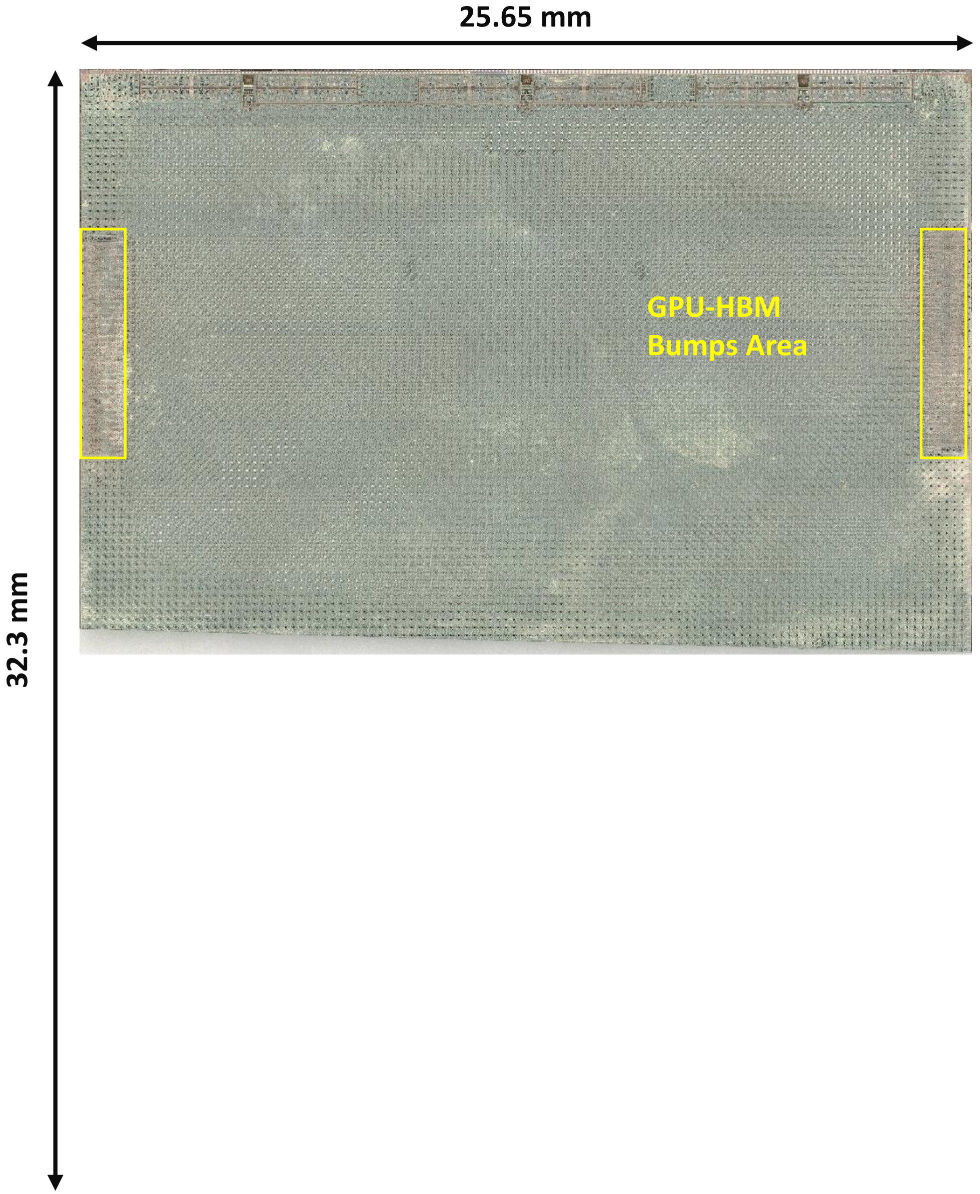
©2018 by System Plus Consulting



HBM Stack Cross-Section – SEM View

©2018 by System Plus Consulting

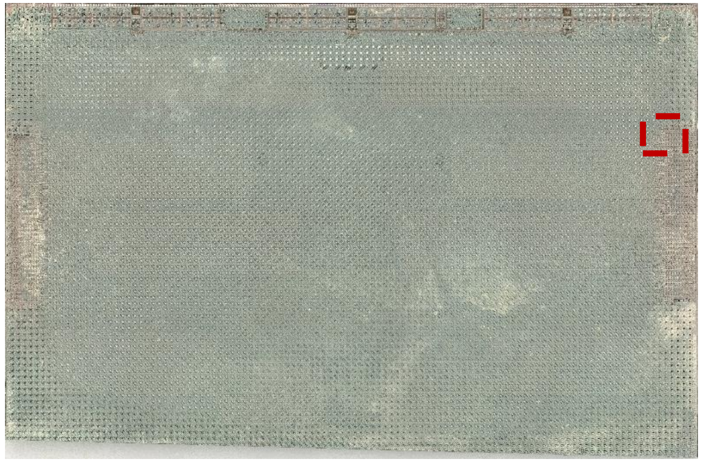
NVIDIA GV100 – GPU Die View and Dimensions



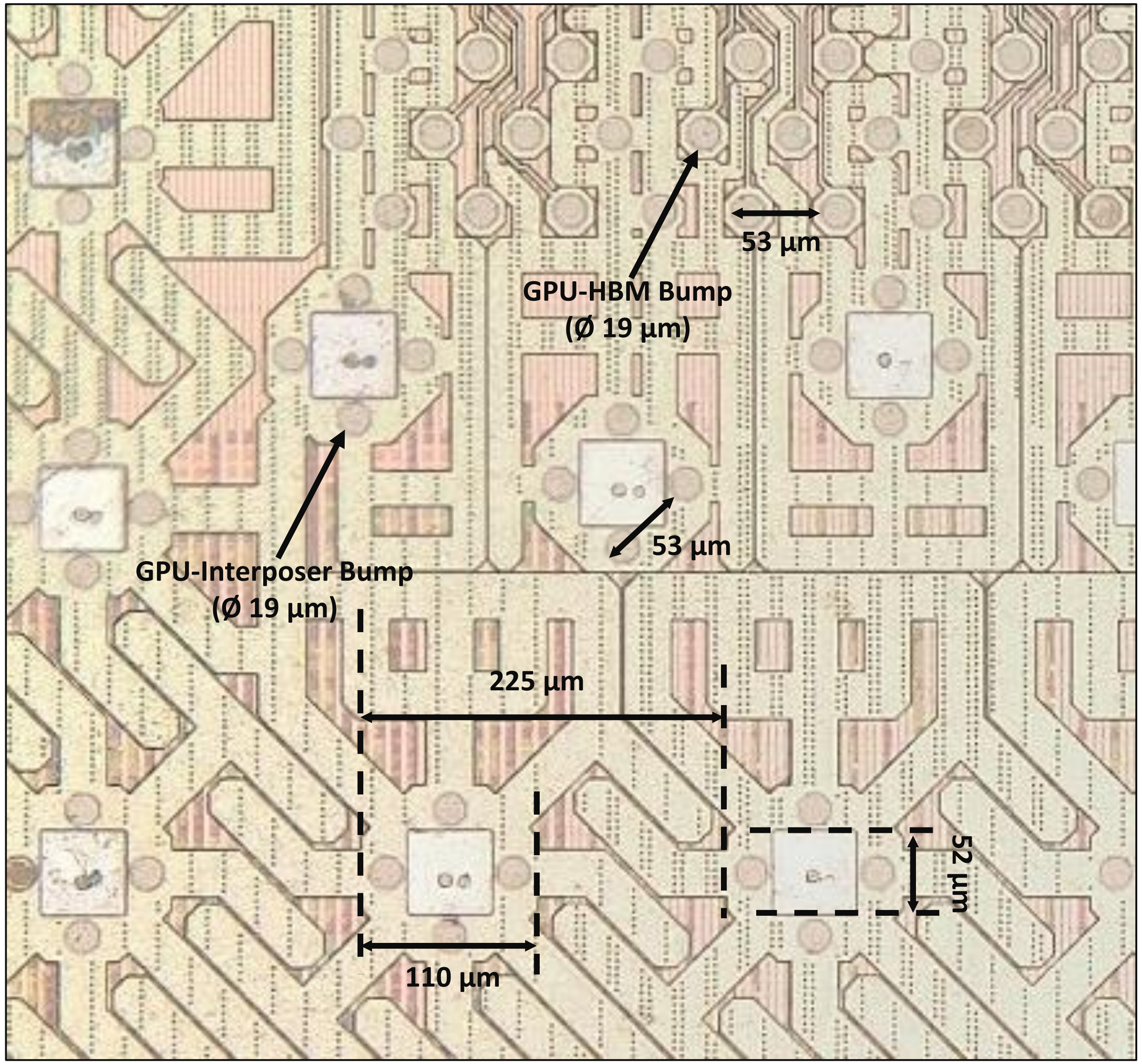
Die Overview

Die Area:	$828.5\ mm^2$ (32.3 x 25.7 mm)
Nb of PGDW per 12-inch wafer:	64
GPU-HBM Bumps Area:	$4 \times 7.32\ mm^2$ (6.10 x 1.20 mm)
GPU-HBM Bumps Number:	2,352
GPU-HBM Bumps Fill Factor:	3.5 %

NVIDIA GV100 – GPU Die Bumps



Die Overview
©2018 by System Plus Consulting

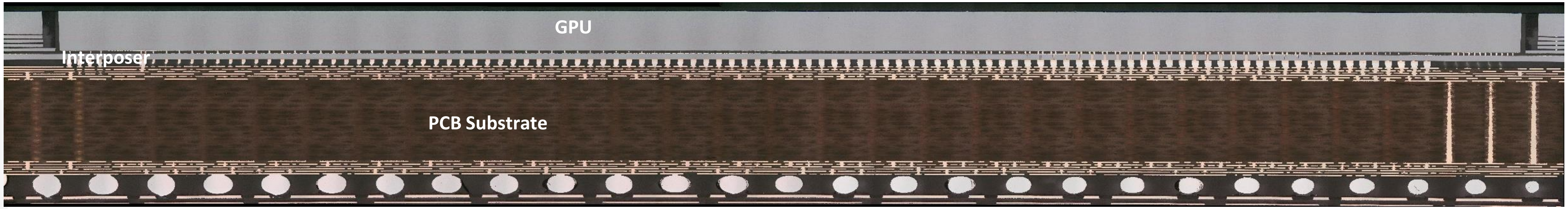


Die Top View – Bumps
©2018 by System Plus Consulting

- GPU-HBM bumps are located at four sites on the die.
- Bump pitch: 53 µm
- Bump diameter: 19 µm
- For the probe testing, each Al pad are opened.

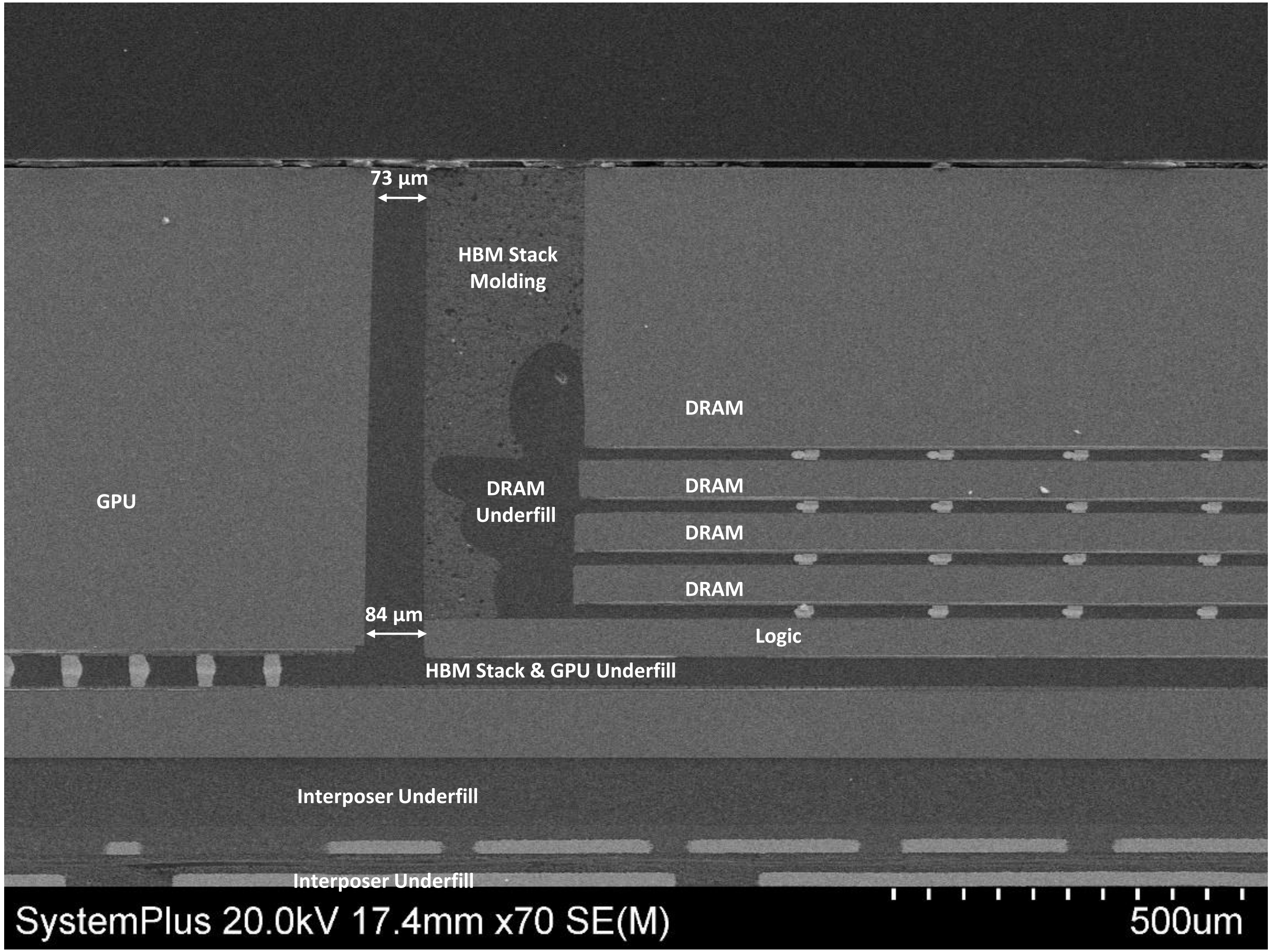
Package Cross-Section – GPU

- The GPU die is flip-chipped on the interposer at the wafer-level though microbumps.



GPU Cross-Section – Optical View

Package Cross-Section – GPU



GPU Cross-Section – SEM View

- The GPU and the HBM stack package are placed side by side in an underfill.
- The space between the dies are 70 – 80 µm wide.

Package Cross-Section – GPU

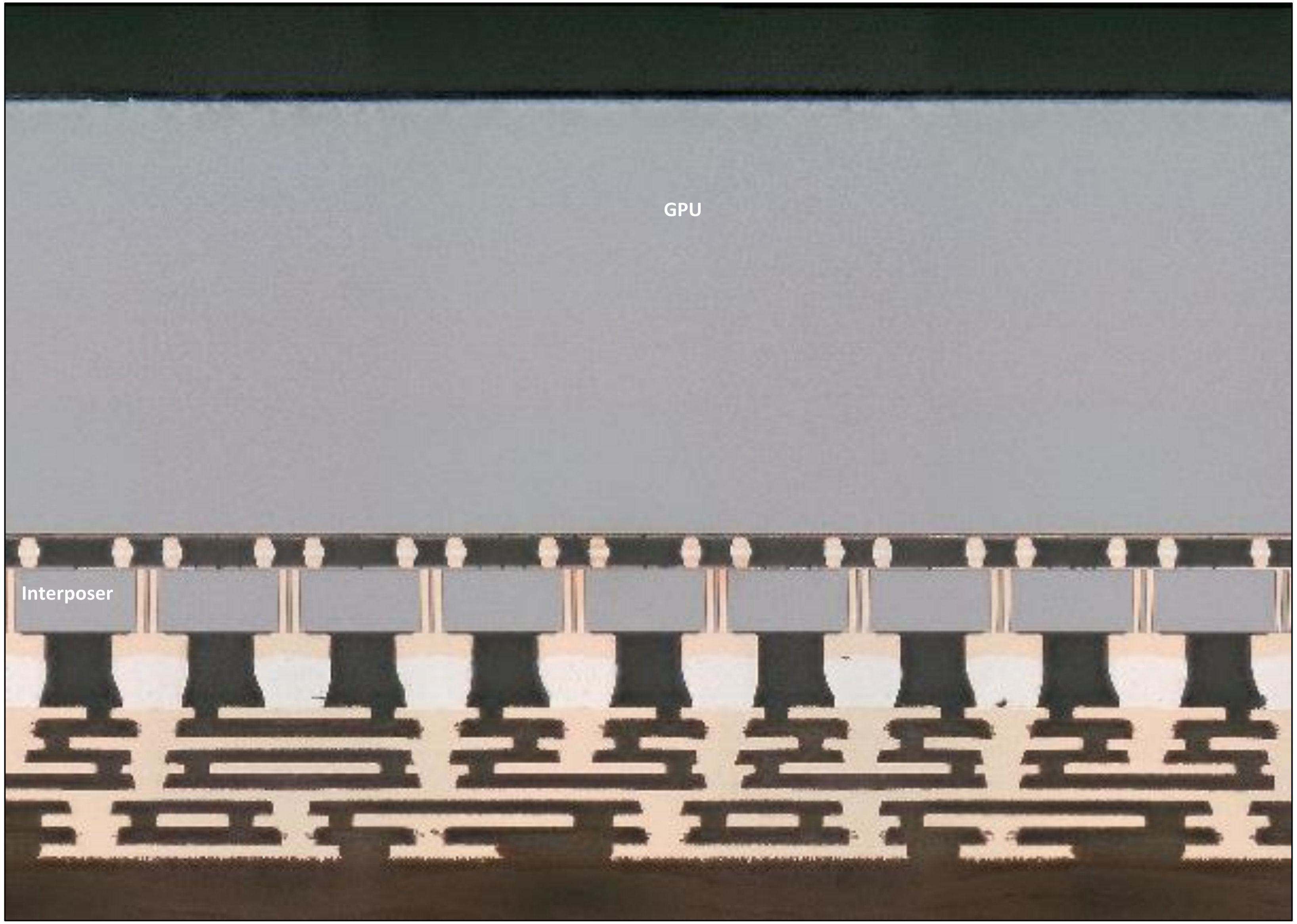
GPU

1x GPU Die

TSV and microbumps connection

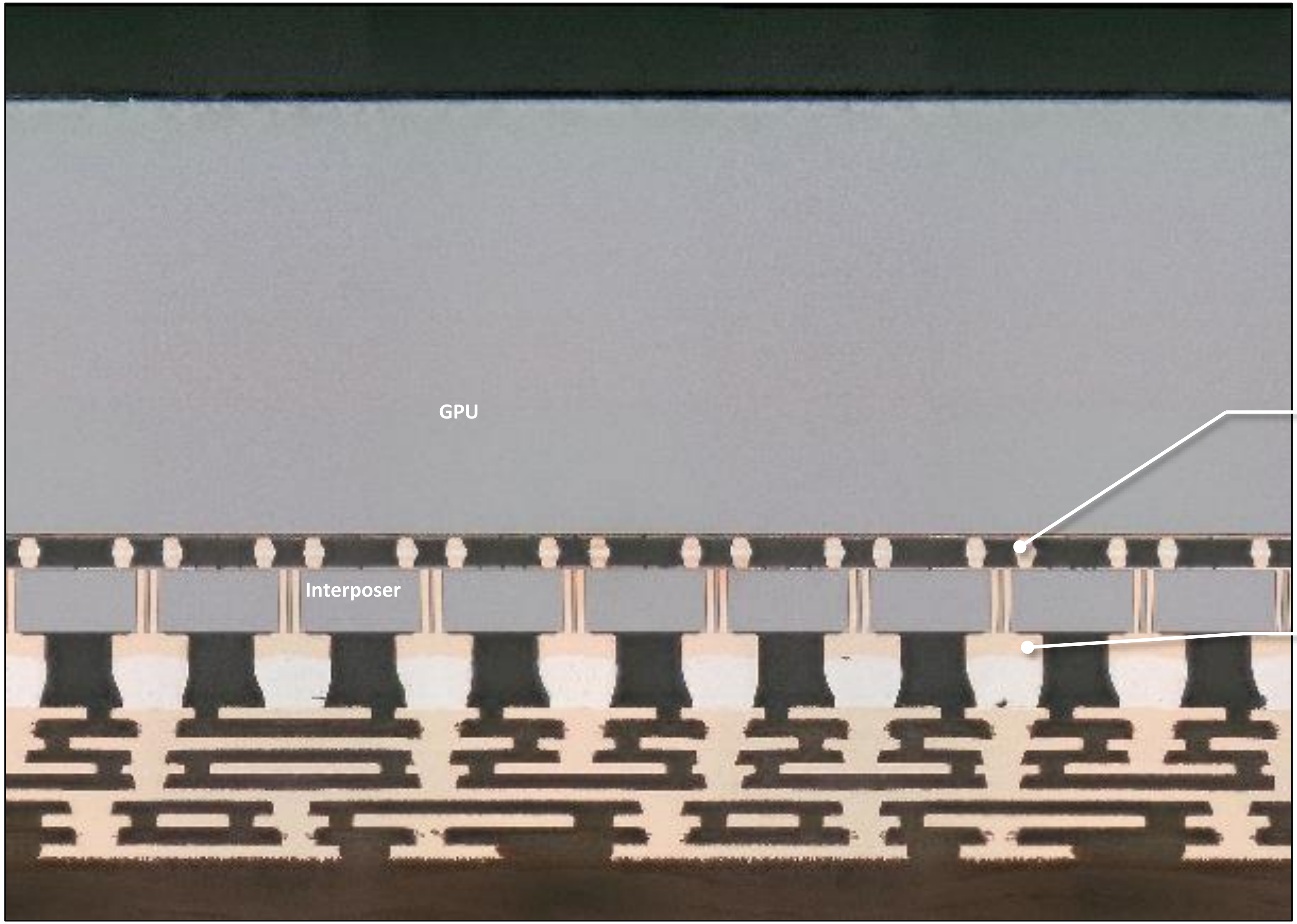
Interposer

TSV, redistribution layers and microbumps connections



GPU Cross-Section – Optical View

Package Cross-Section – GPU

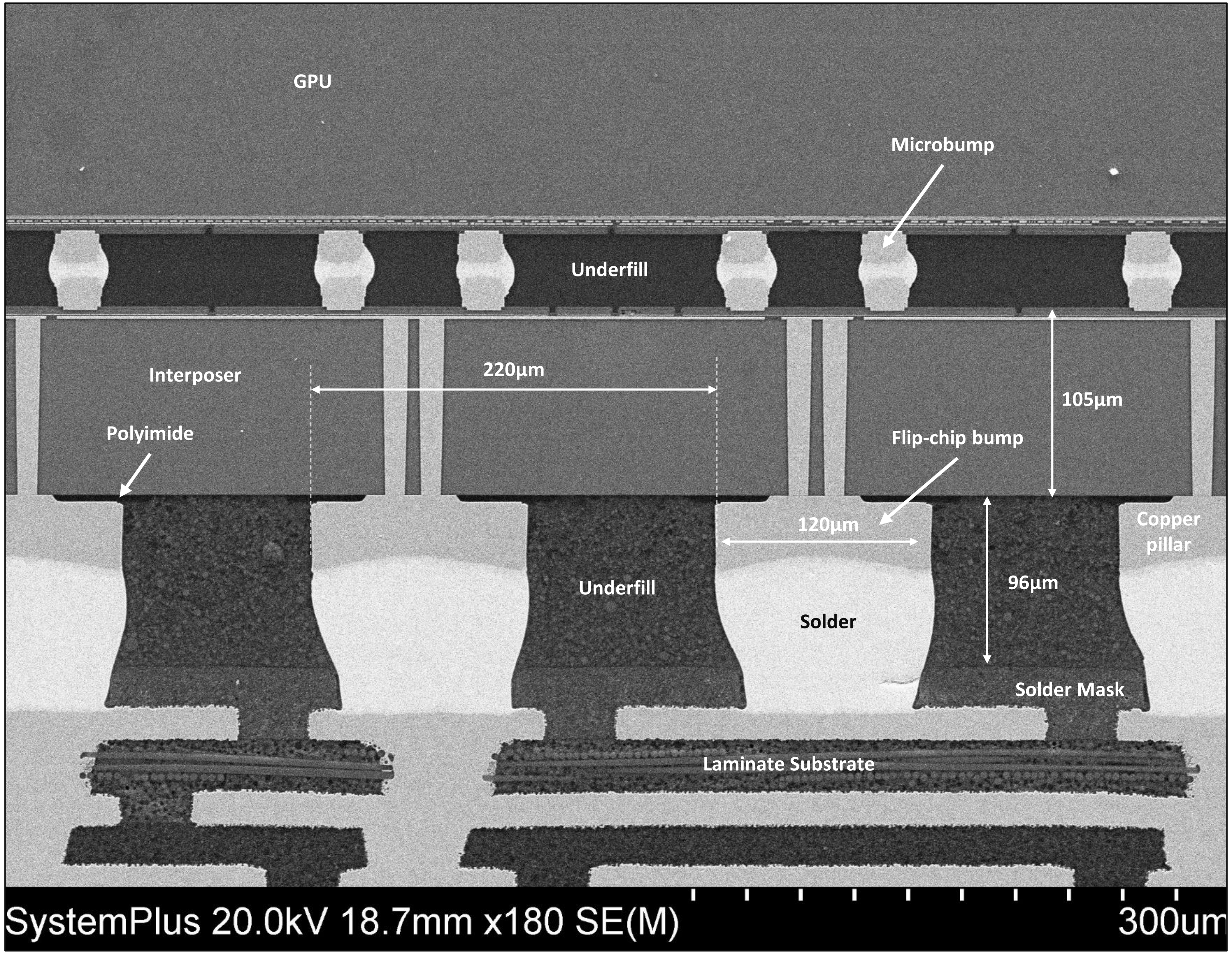


Micro-bump
- ✓ Diameter 25 µm
- ✓ Pitch 78 µm

Micro-bump
- ✓ Diameter 120 µm
- ✓ Pitch 225 µm

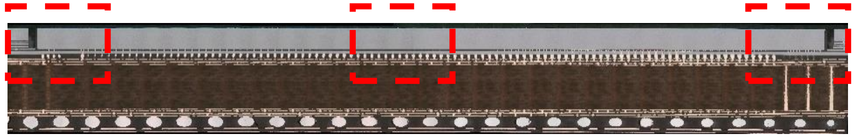
GPU Cross-Section – Optical View

Package Cross-Section – Substrate – Interposer

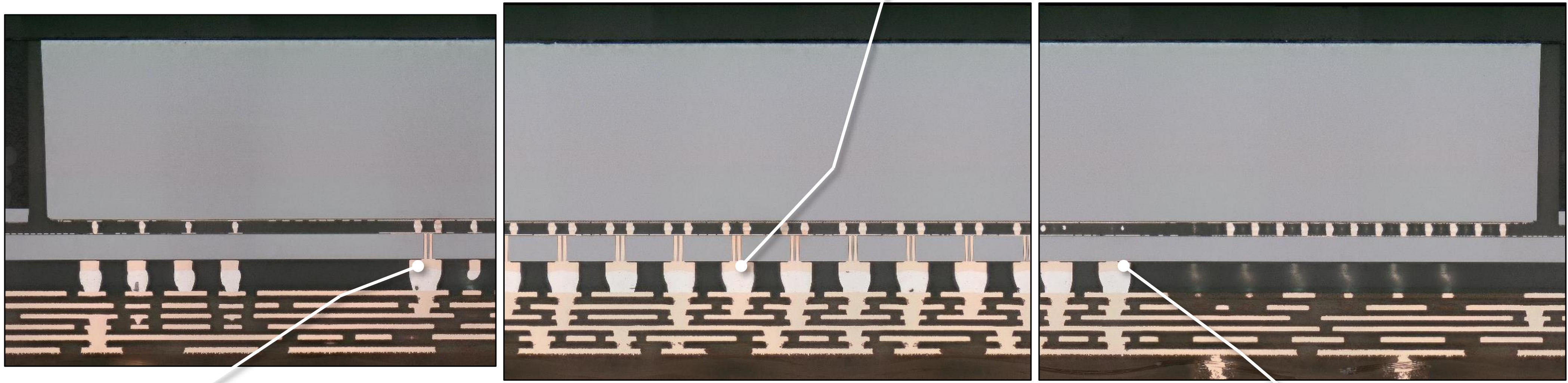


Interposer Cross-Section – SEM View

Package Cross-Section – Substrate – Interposer



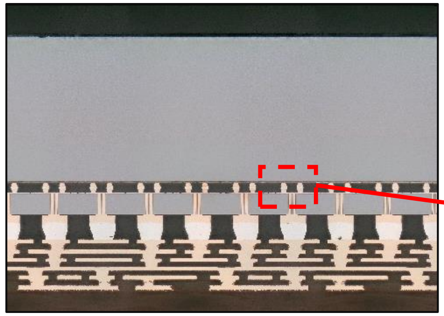
Interposer Cross-Section – SEM View

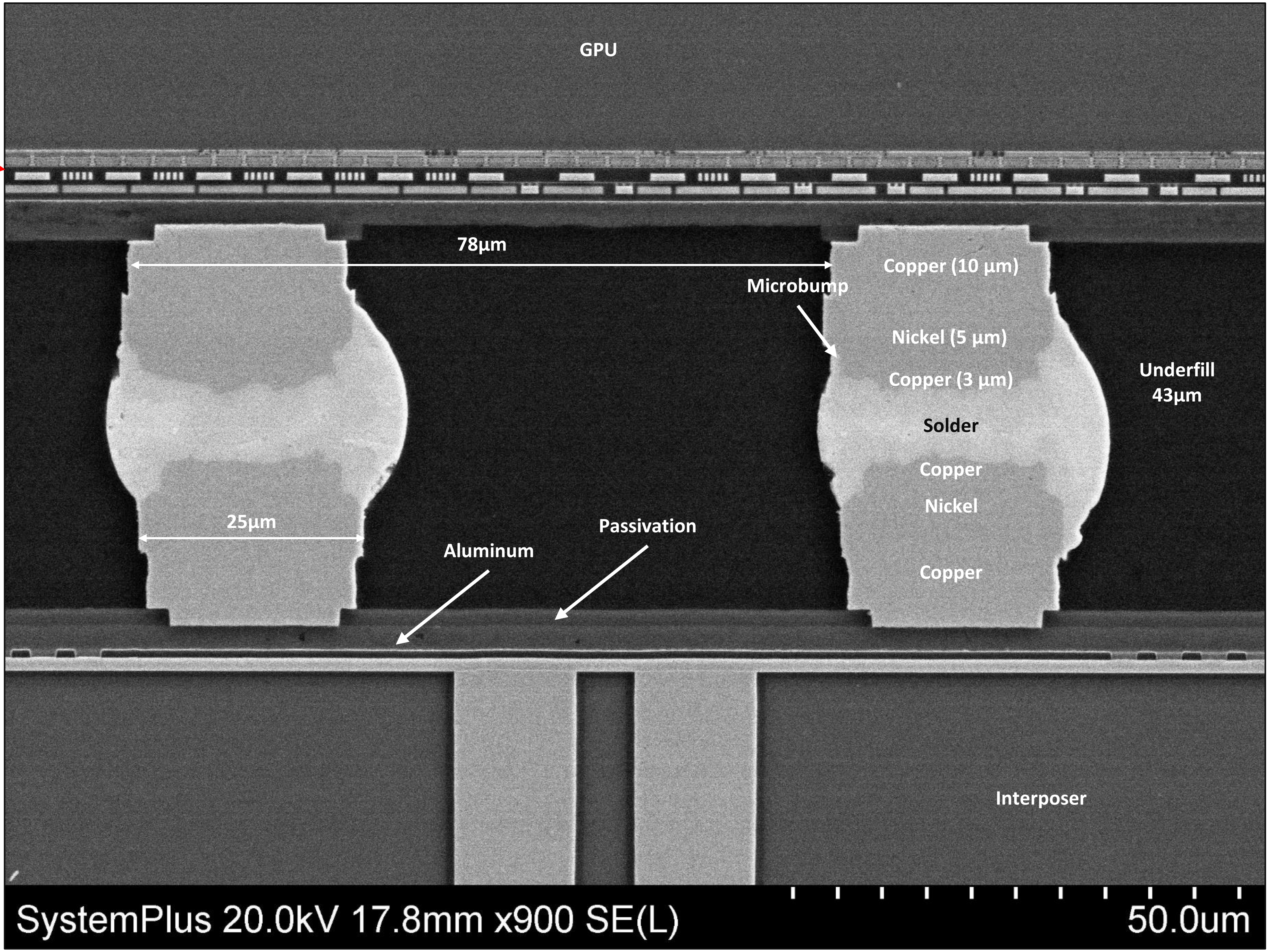


Interposer Cross-Section – SEM View

- Under the GPU, three type of bumps are used.
- The bumps could be regrouped into two group: One in stress region and one in High power region.
- In the stress region, depending on the position of the edge, the bumps is extended in the edge direction.
- In the high power region, the bumps are smaller.
- We assume that the bumps are tweaks to help manage the warpage.

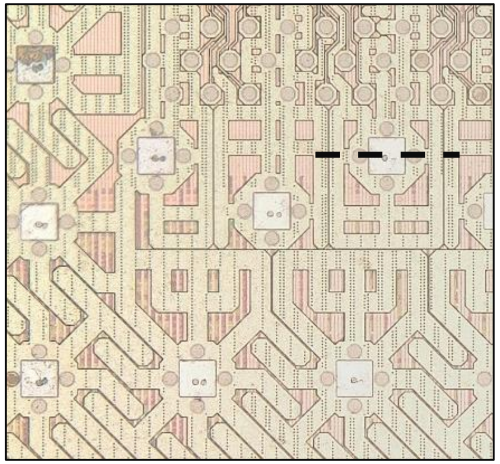
Package Cross-Section – Interposer – GPU





Interposer Cross-Section – SEM View

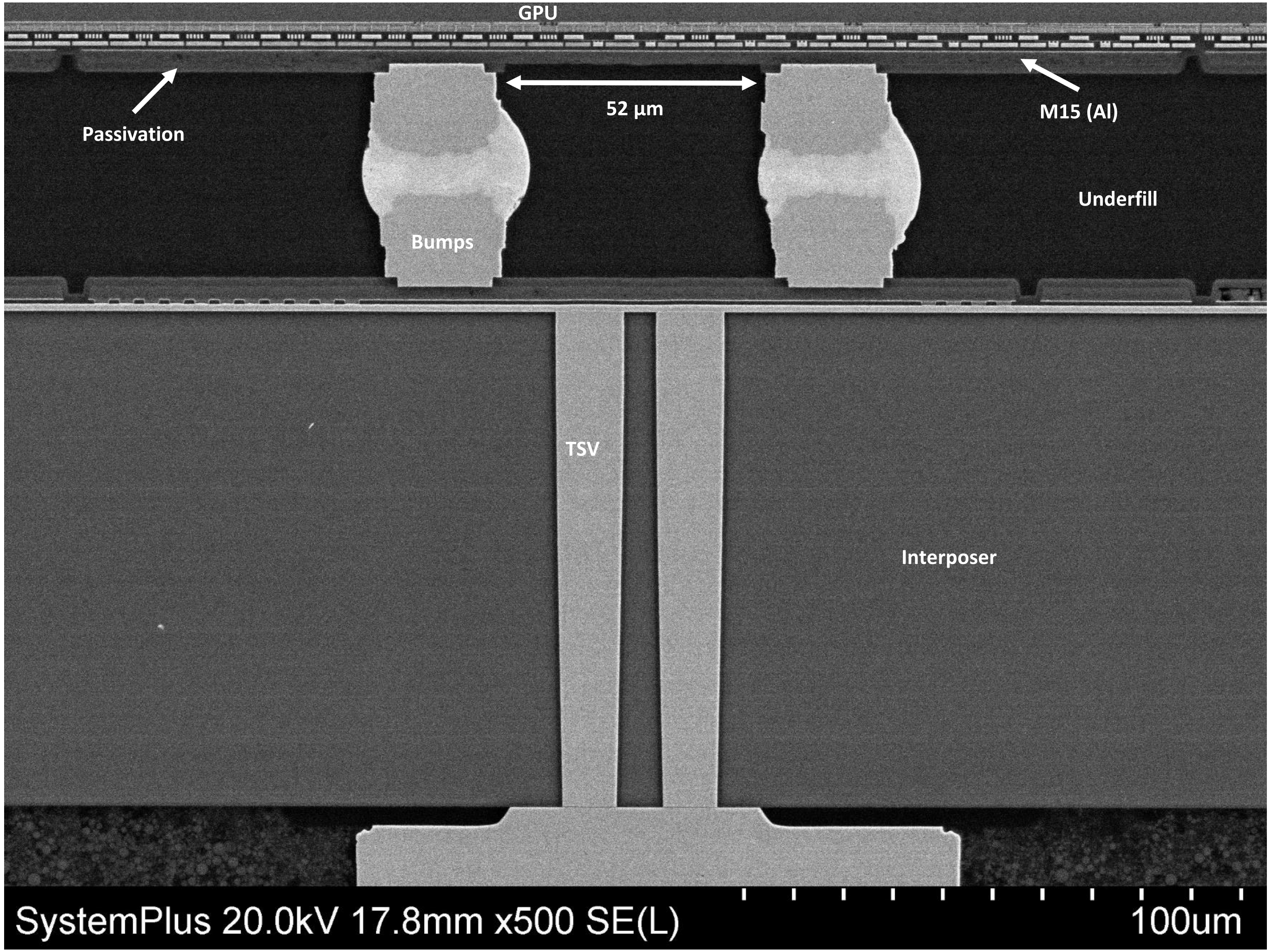
Die Cross-Section –GPU



Die Top View – Bumps

©2018 by System Plus Consulting

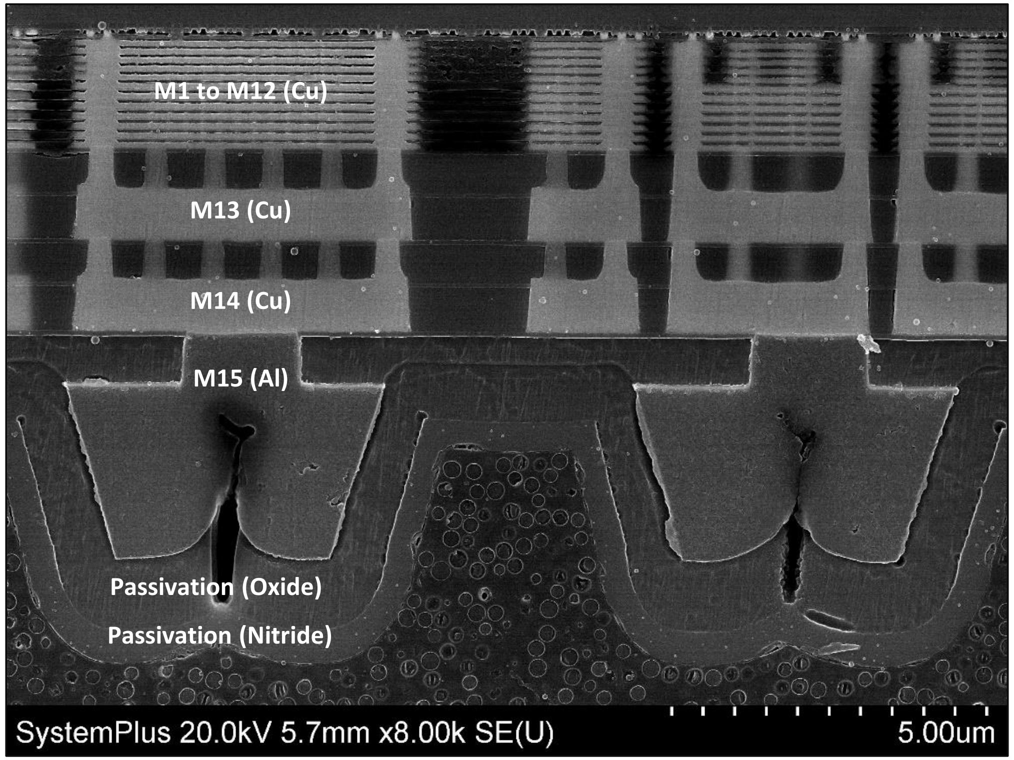
- The passivation on top of the aluminum pads are opened to provide a spot to perform the wafer probe testing.



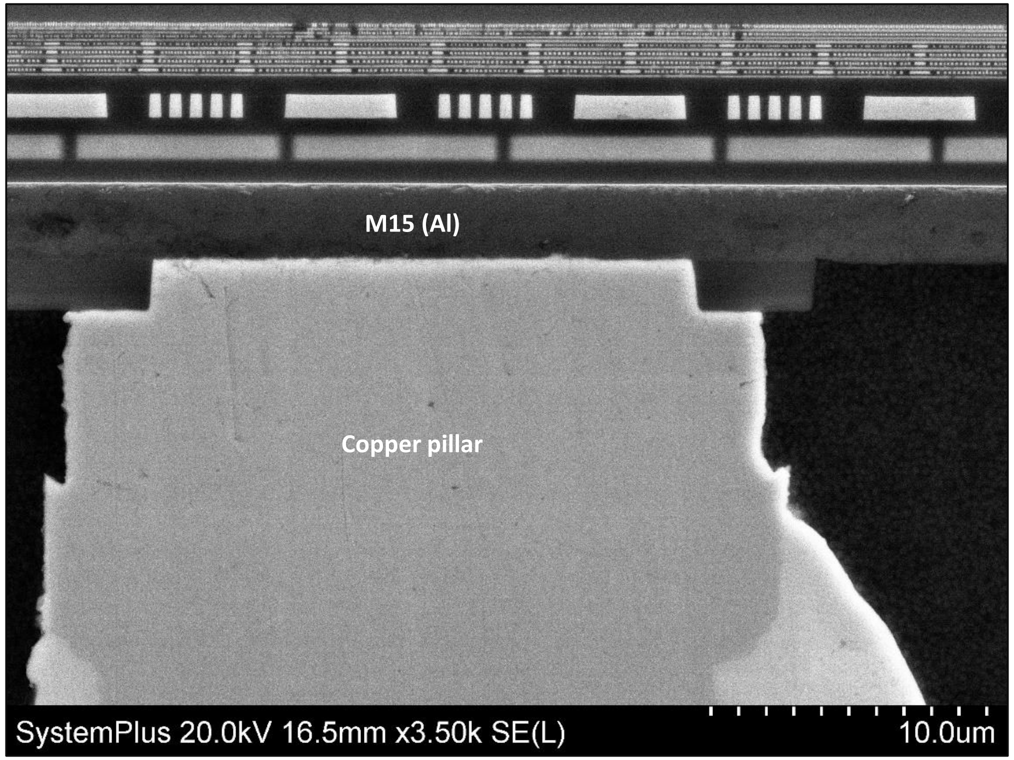
Interposer Cross-Section – SEM View

©2018 by System Plus Consulting

Die Cross-Section – GPU



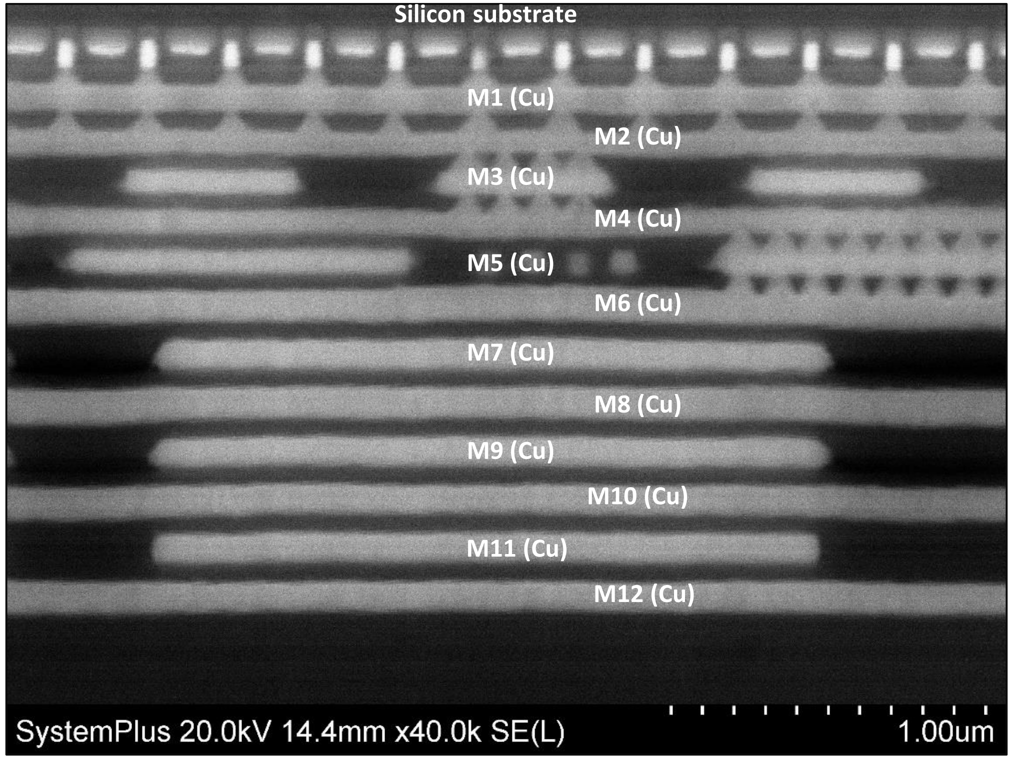
GPU Cross-Section – SEM View



GPU Cross-Section – SEM View

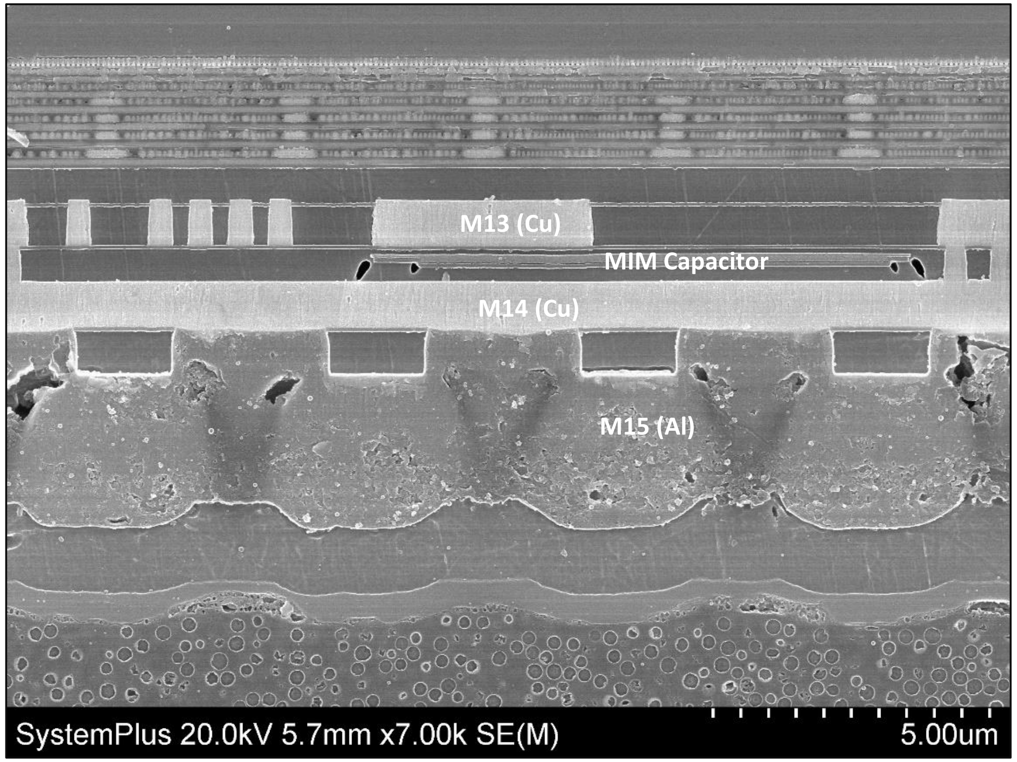
- The GPU uses 15 metal layers (14 Cu + 1 Al)

Die Cross-Section – GPU



GPU Cross-Section – SEM View

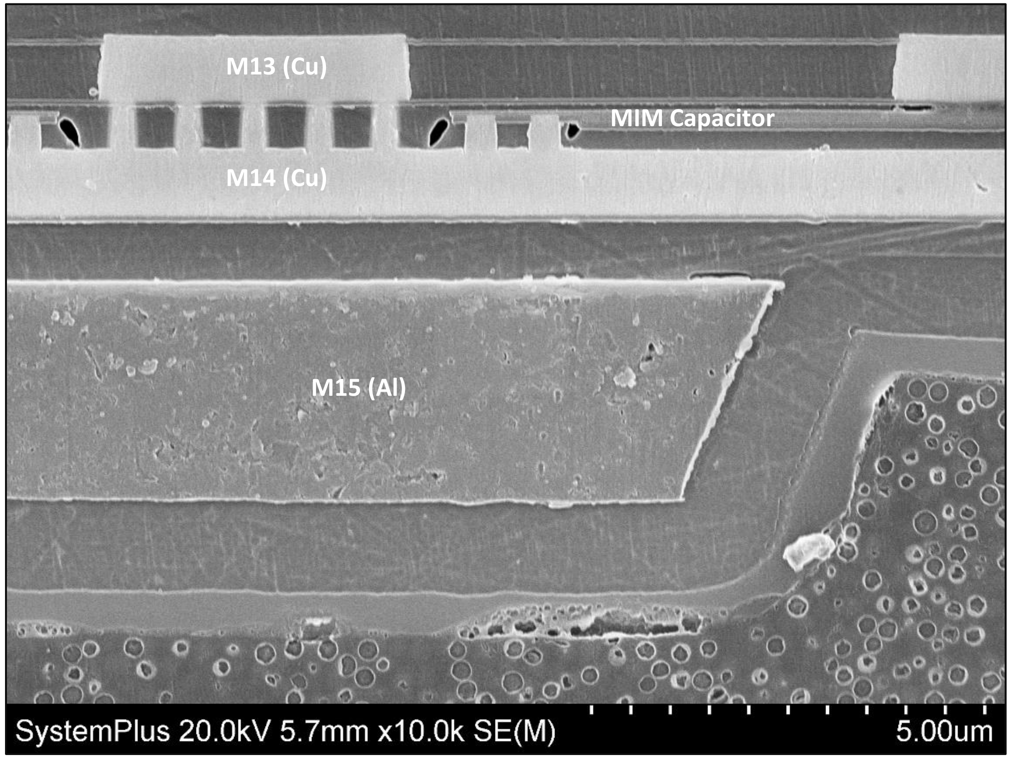
©2018 by System Plus Consulting



GPU Cross-Section – SEM View

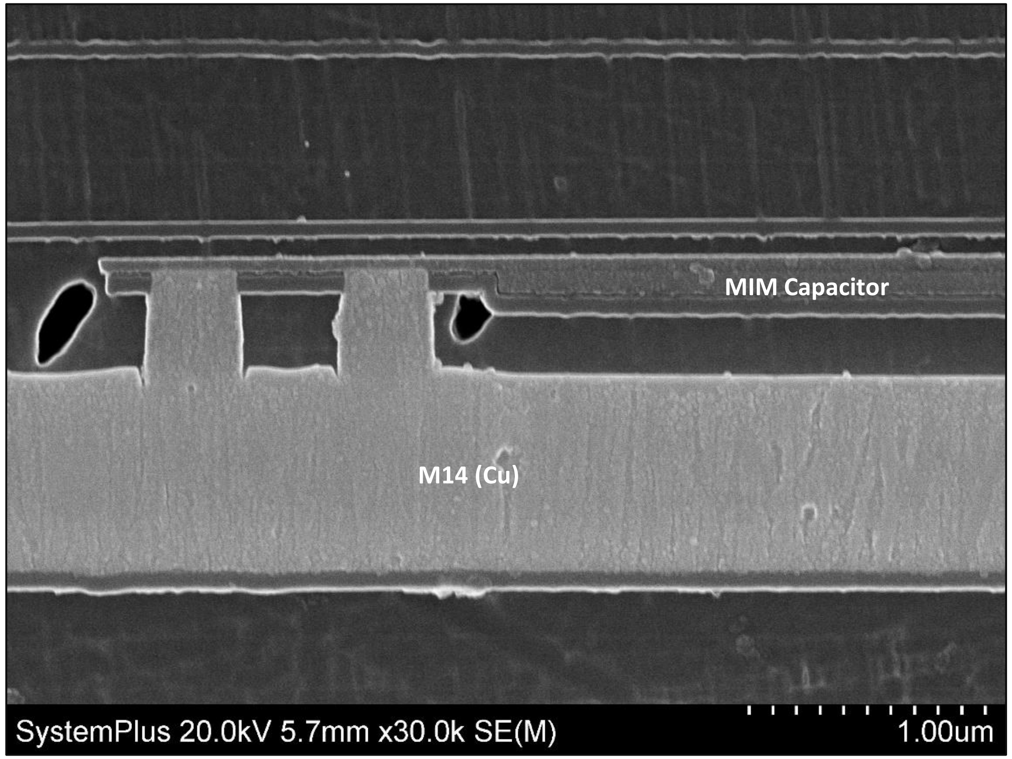
©2018 by System Plus Consulting

Die Cross-Section – GPU



GPU Cross-Section – SEM View

©2018 by System Plus Consulting

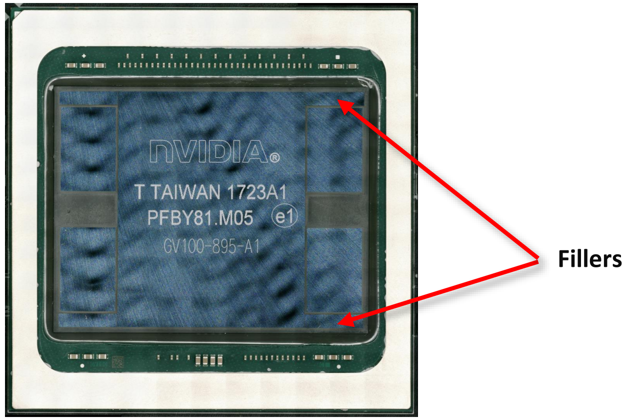


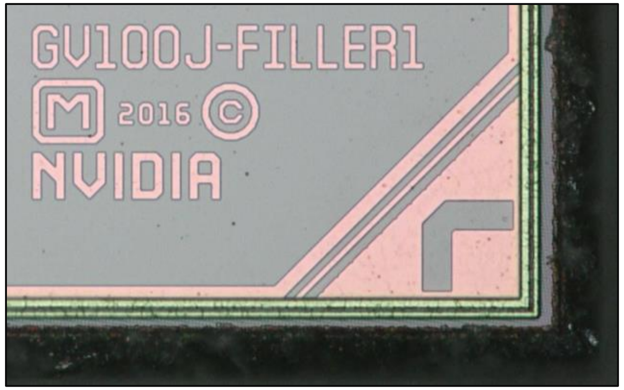
GPU Cross-Section – SEM View

©2018 by System Plus Consulting

- MIM capacitors are present between M13 and M14.

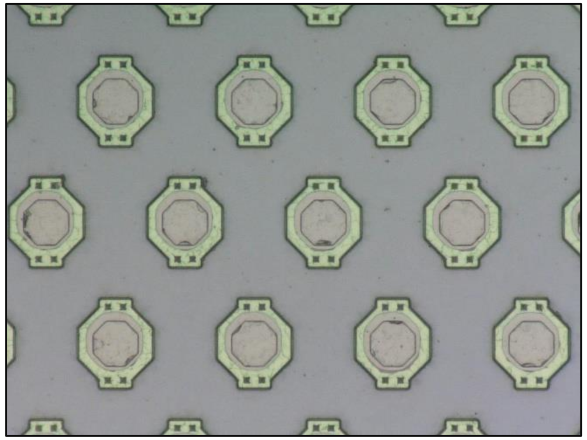
Package Cross-Section – Filler





Die Marking

©2018 by System Plus Consulting

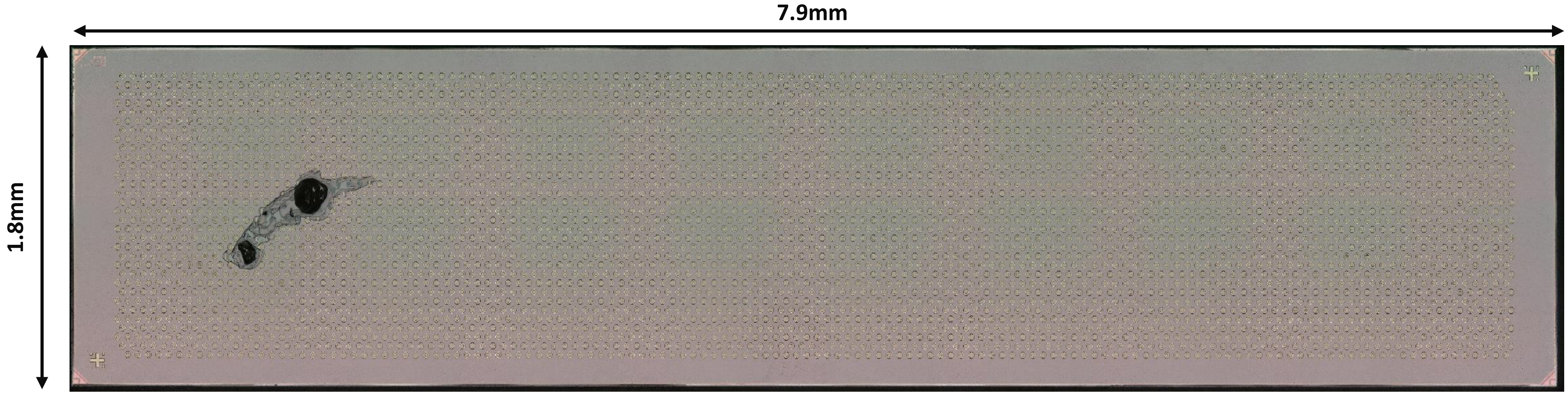


Micro bumps Structure

©2018 by System Plus Consulting

Die Area: 14.22mm² (7.9x1.8mm)

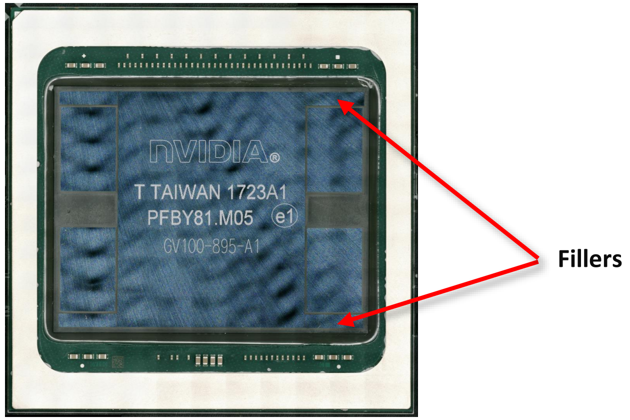
Nb of PGDW per 12-inch wafer: 4,396



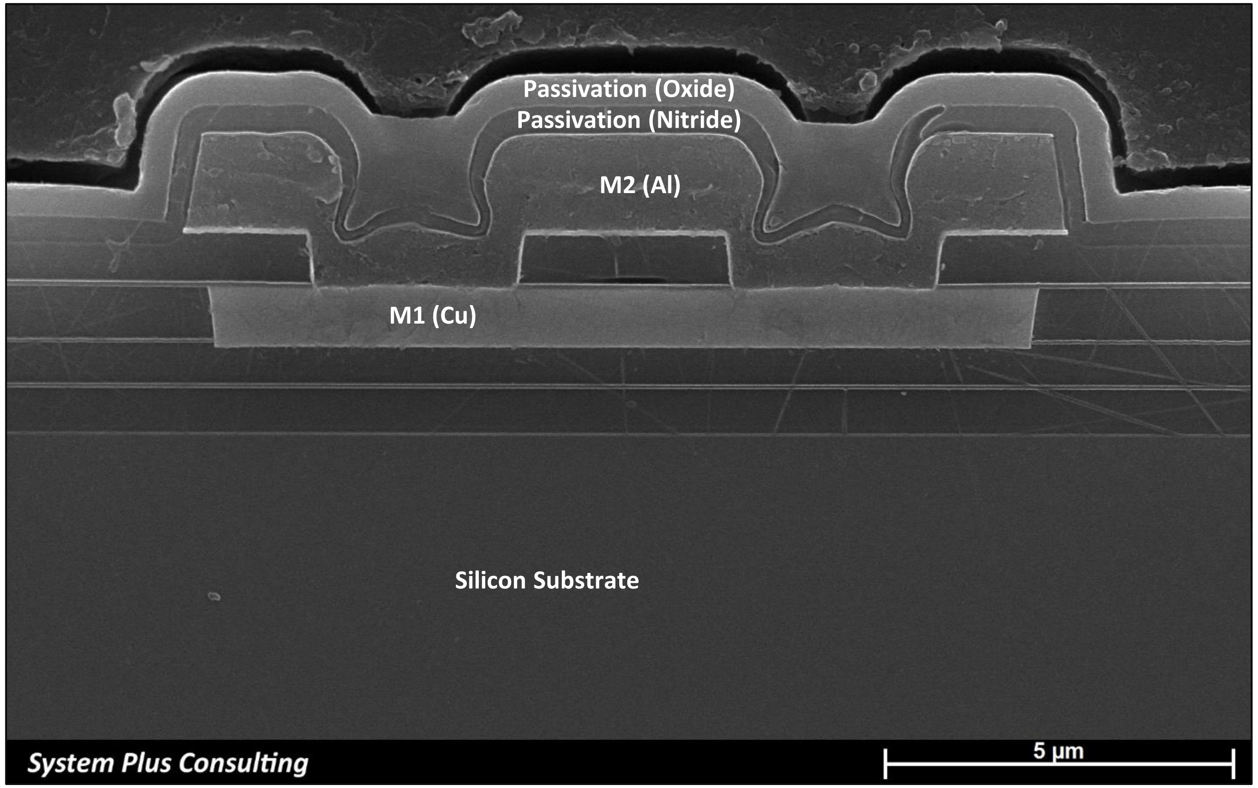
Die Overview

©2018 by System Plus Consulting

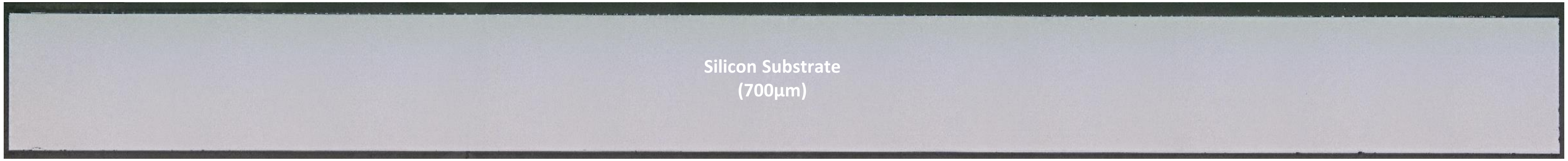
Package Cross-Section – Filler Cross-Section



- The fillers use 2 metal layers and micro bumps connections.

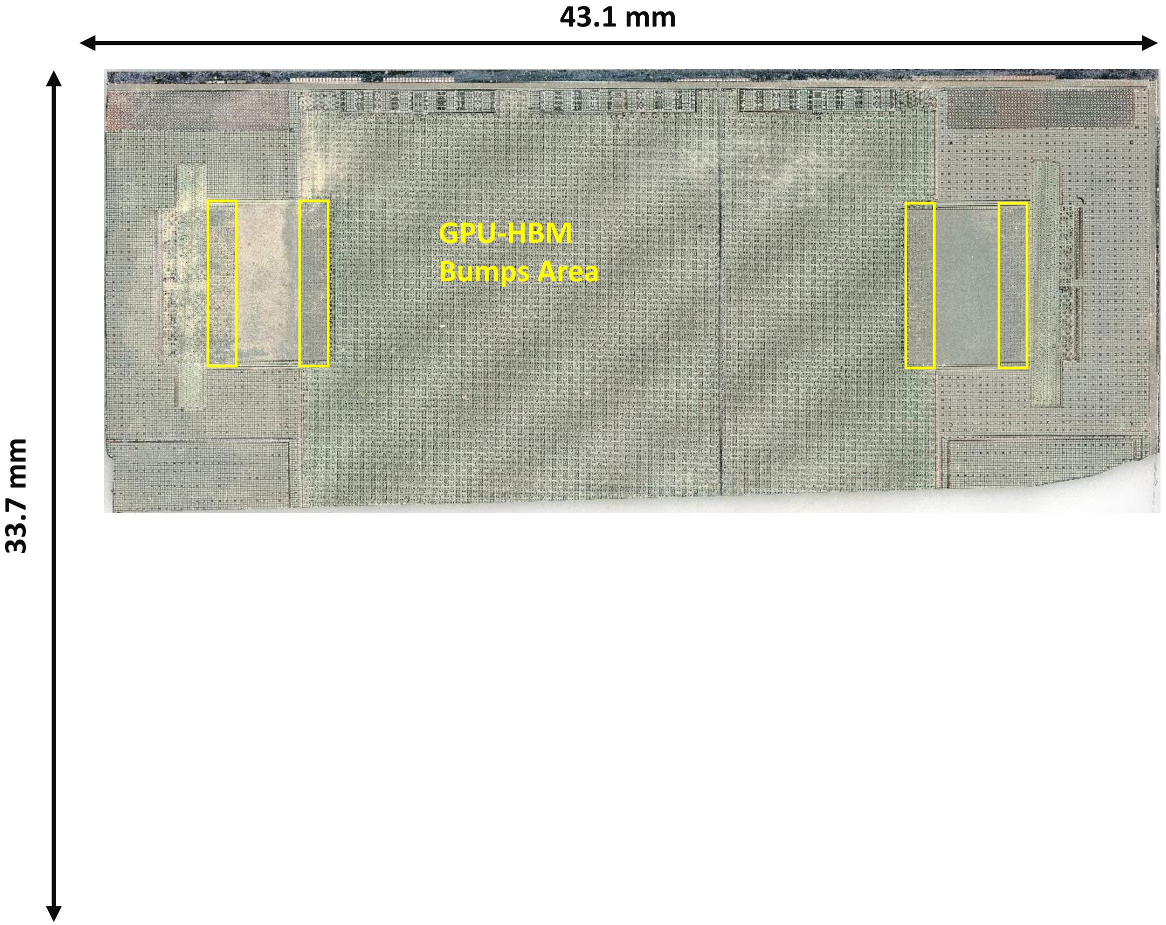


Filler Cross-Section – SEM View



Filler Cross-Section – Optical View

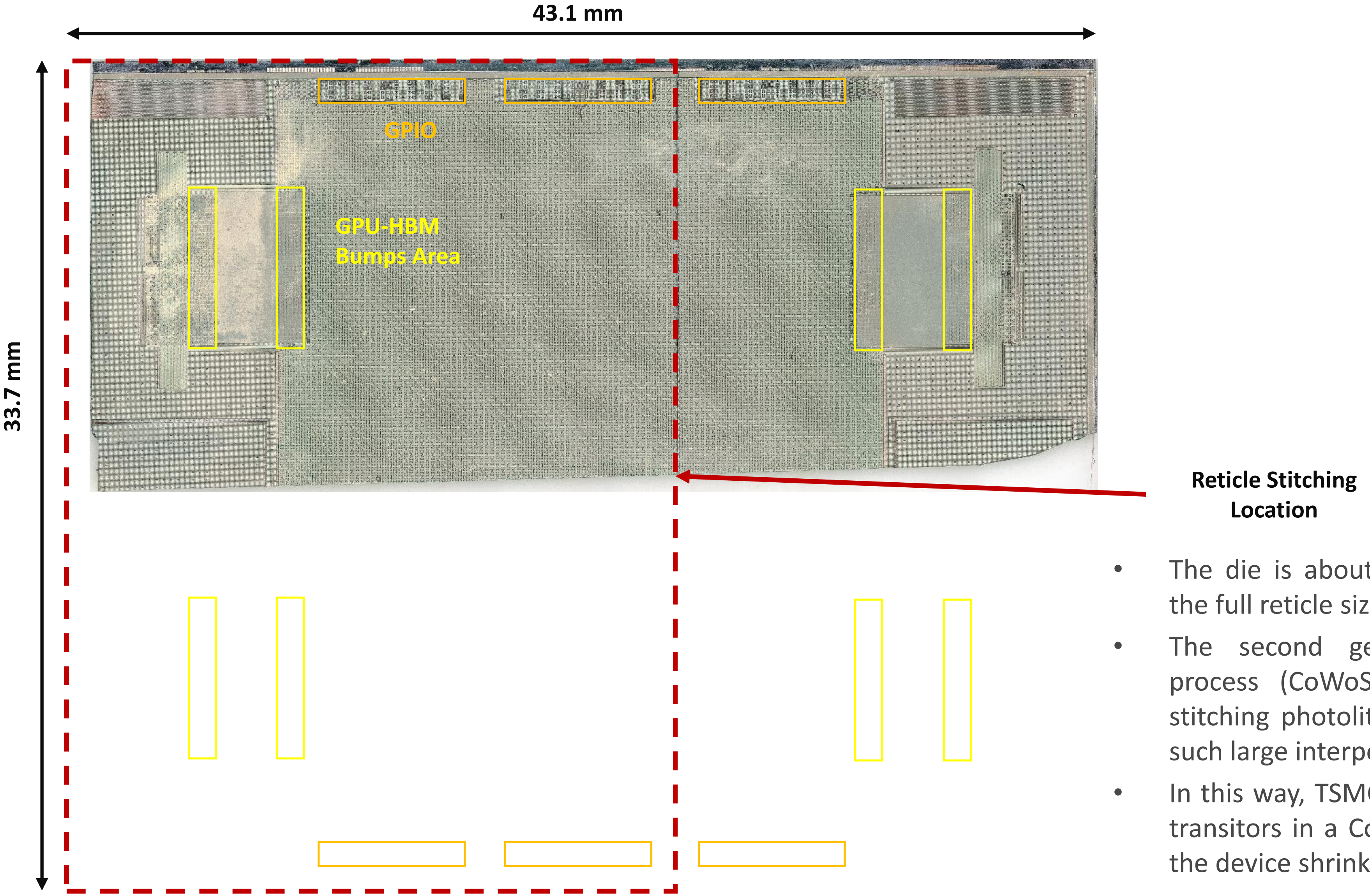
NVIDIA GV100 – Interposer Die View and Dimensions



Die Overview
©2018 by System Plus Consulting

Die Area:	1,452 mm² (43.1 x 33.7 mm)
Nb of PGDW per 12-inch wafer:	36
GPU-HBM Bumps Area:	8 x 7.32 mm² (6.10 x 1.20 mm)
GPU-HBM Bumps Number:	2,352
GPU-HBM Bumps Fill Factor:	4.0 %

NVIDIA GV100 – Interposer Die Overview

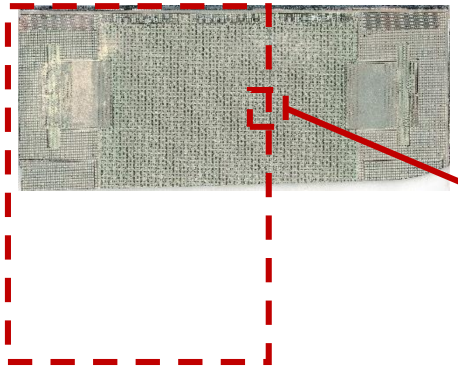


Die Overview

©2018 by System Plus Consulting

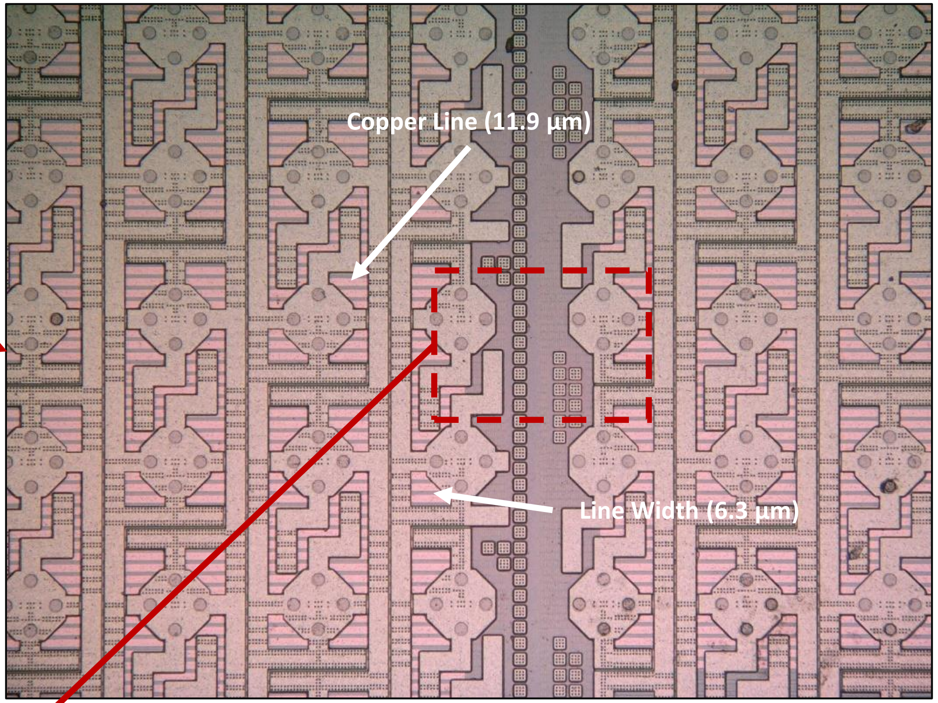
- The die is about 1.5 times large than the full reticle size.
- The second generation of CoWoS process (CoWoS-2) use a two-mask stitching photolithography to fabricate such large interposer.
- In this way, TSMC is able to pack more transitors in a CoWoS package besides the device shrink by Moore's law.

NVIDIA GV100 – Interposer Die Overview



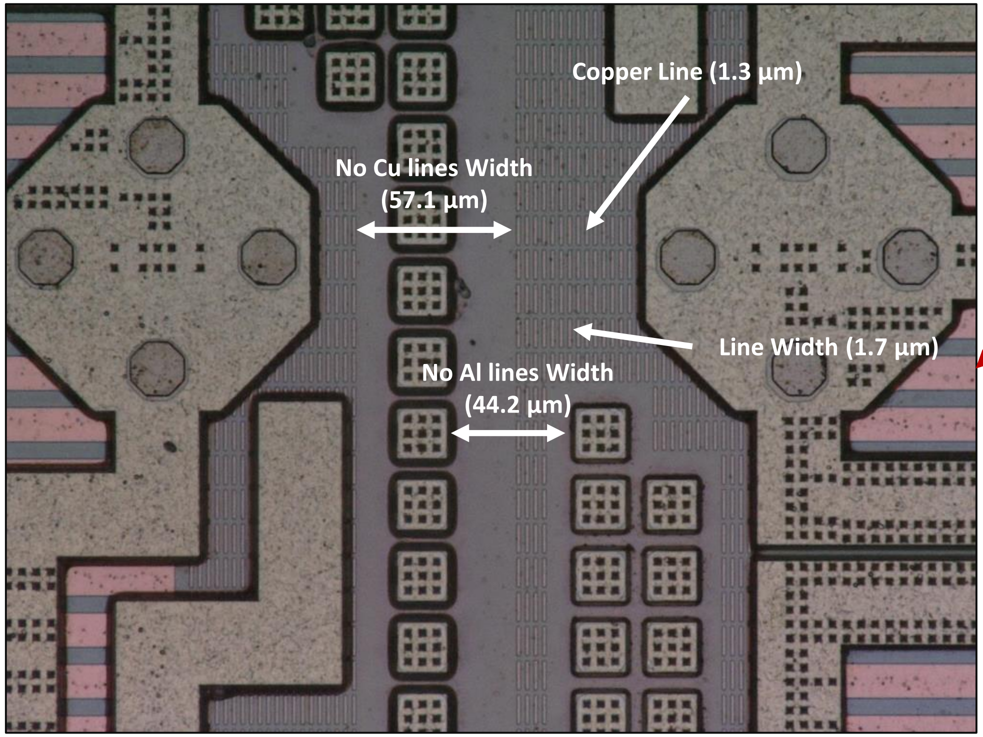
Die Overview

©2018 by System Plus Consulting



Die Overview

©2018 by System Plus Consulting

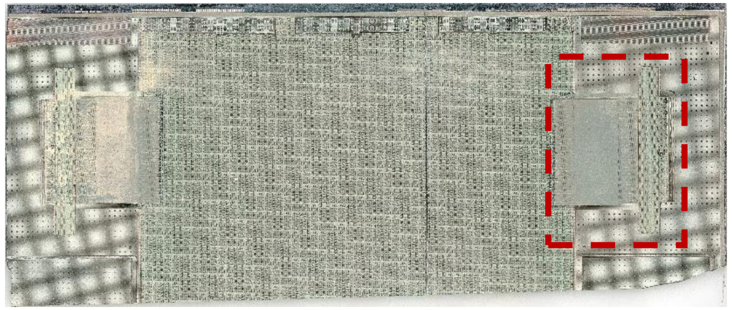


Die Overview

©2018 by System Plus Consulting

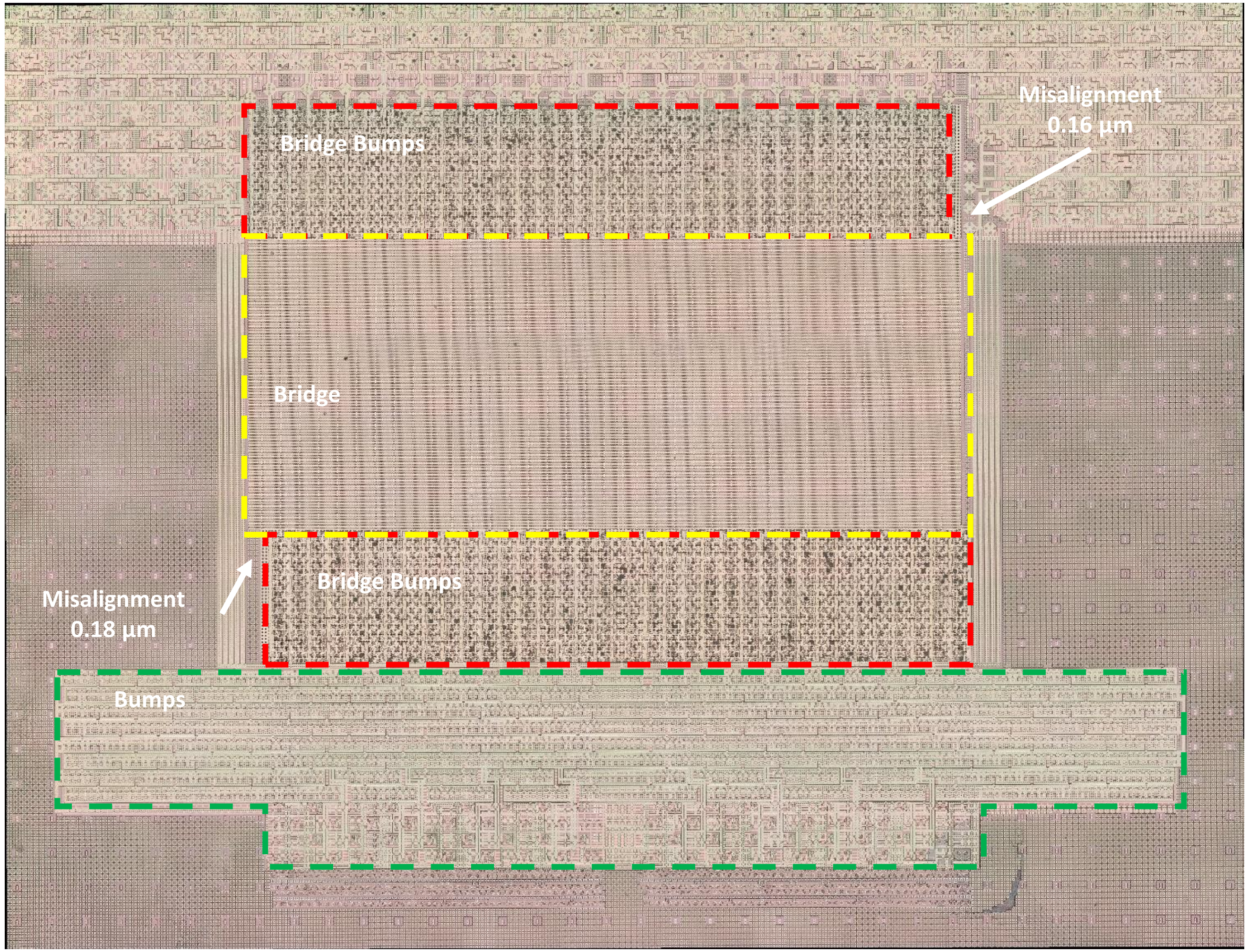
- No Copper or aluminum lines cross the reticle stitching area.

NVIDIA GV100 – Interposer Die Bumps



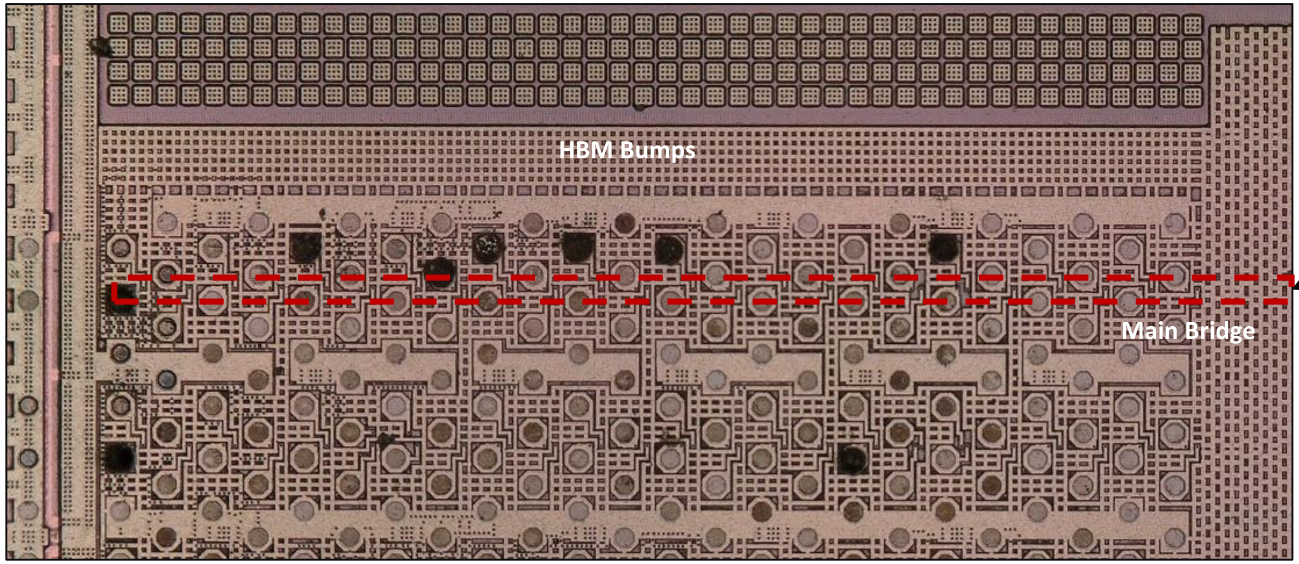
Die Overview
©2018 by System Plus Consulting

- The bridge bumps are misaligned between the GPU and the HBM.



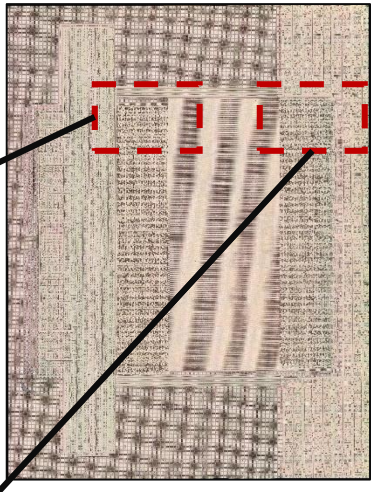
Die Top View – Bumps
©2018 by System Plus Consulting

NVIDIA GV100 – Interposer Die Bumps



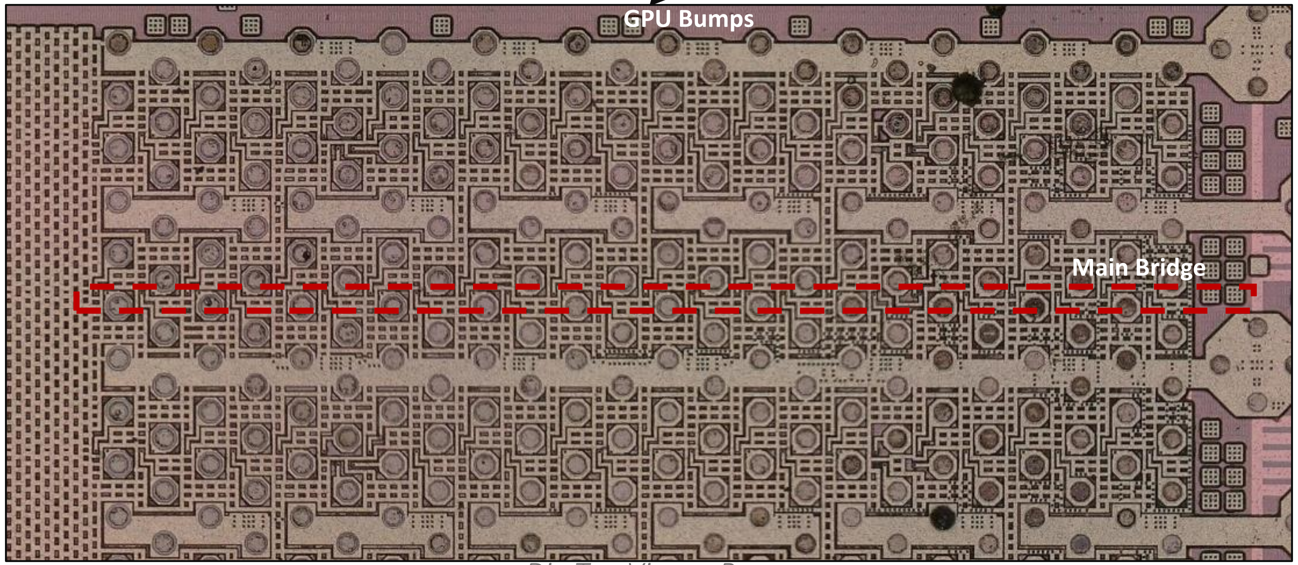
Die Top View – Bumps

©2018 by System Plus Consulting



Die Top View – Bumps

©2018 by System Plus Consulting



Die Top View – Bumps

©2018 by System Plus Consulting

[Overview / Introduction]

[Company Profile & Supply Chain]

[Physical Analysis]

- Summary
- Graphic Card Teardown
- Package
- Dies Size
- Package Opening
- Substrate Cross-Section
- DRAM Dies
- Cross-Section – HBM Stack
- GPU Die
- Cross-Section – GPU
- Filler Die
- Cross-Section – Filler
- Interposer Die
- Cross-Section – Interposer

[Manufacturing Process Flow]

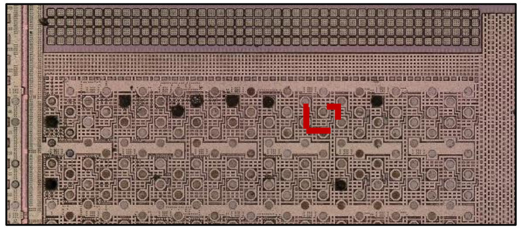
[Cost Analysis]

[Selling Price Analysis]

[Feedback]

[About System Plus]

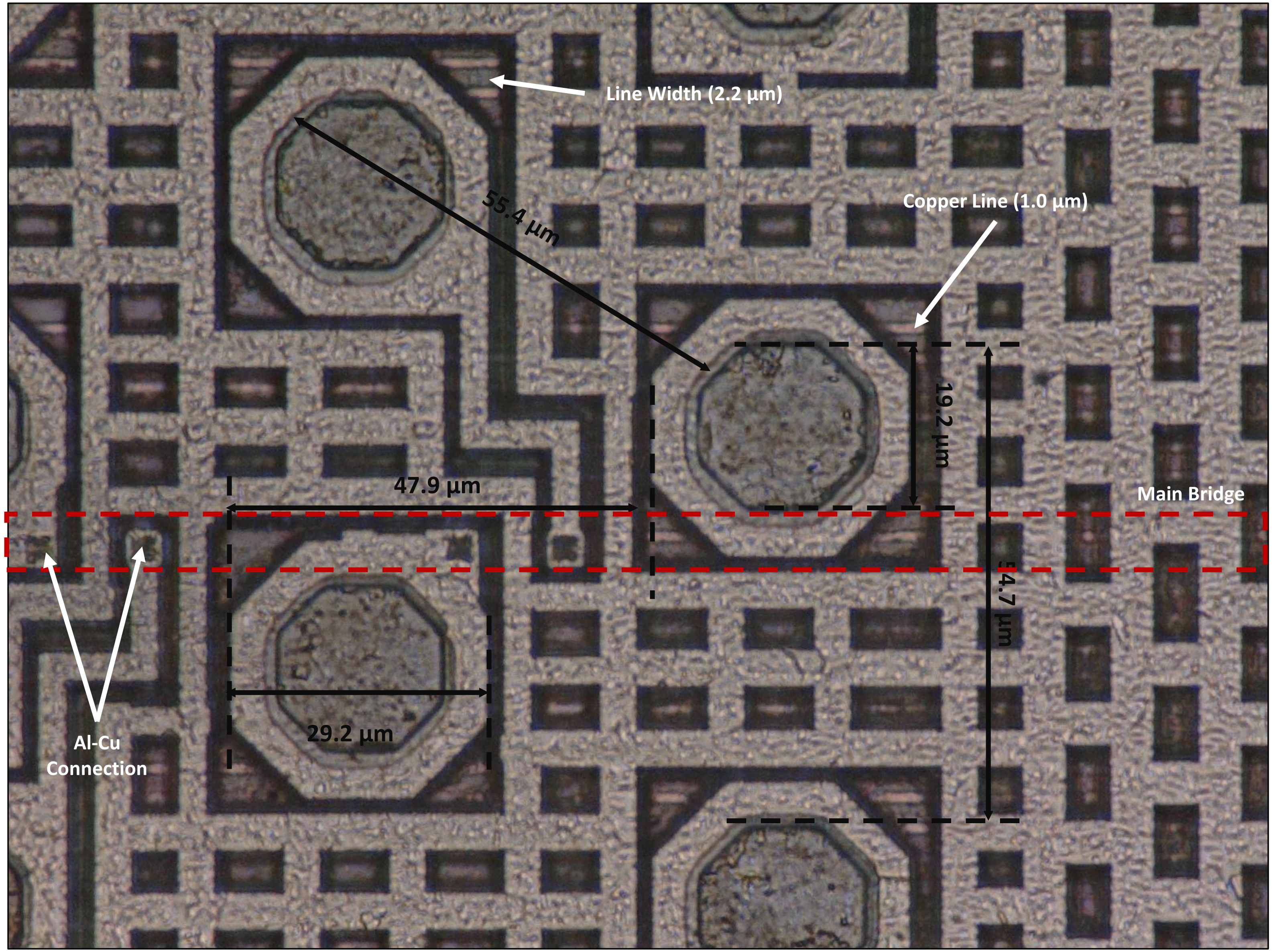
NVIDIA GV100 – Interposer Die Bumps



Die Top View – Bumps

©2018 by System Plus Consulting

- Bump pitch: 55 µm
- Bump diameter: 19 µm

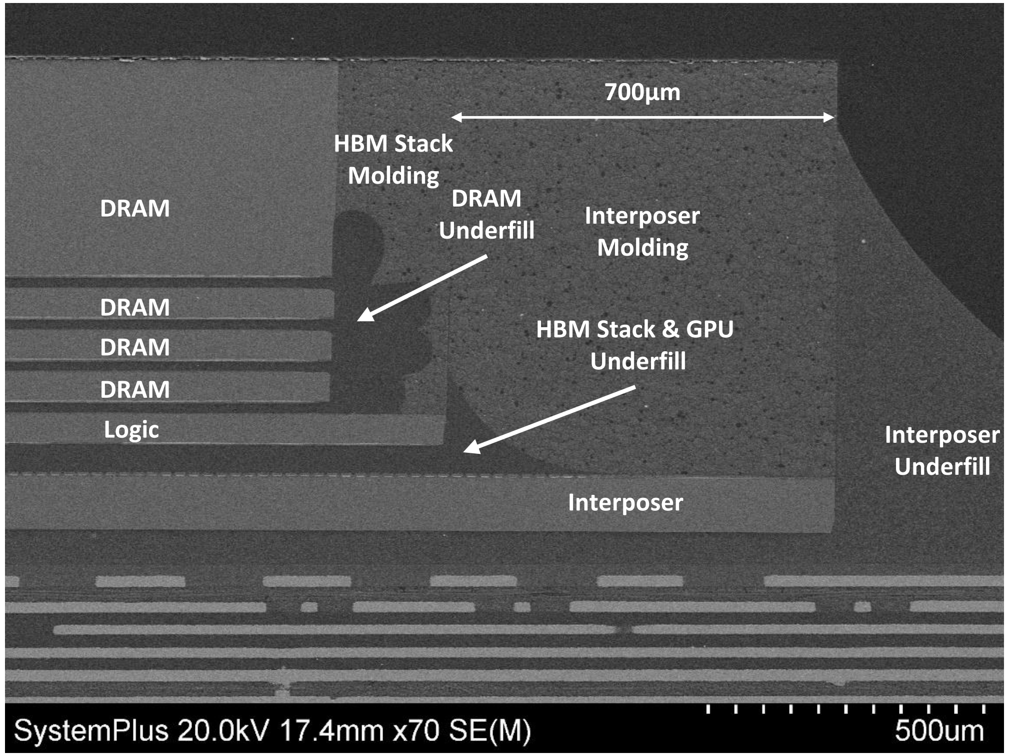


Die Top View – Bumps

©2018 by System Plus Consulting

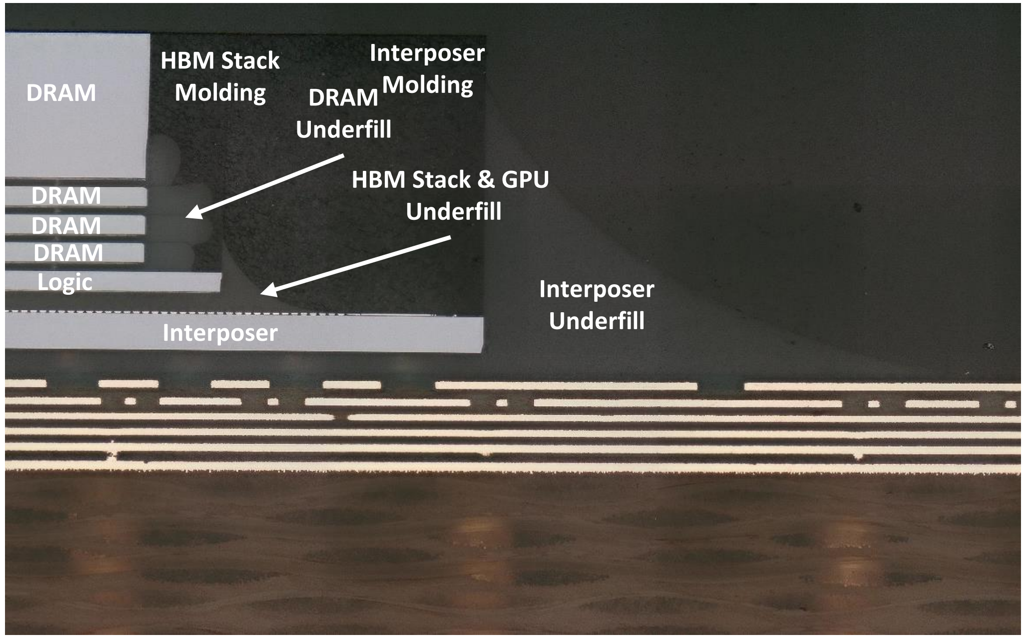
Package Cross-Section – Interposer

- The interposer with GPU and HBM Stacks is molded and grinded.
- 3 different underfills are used :
  - Under the interposer
  - Under GPU & HBM stack
  - Under the DRAM dies in the HBM stack.



Substrate Cross-Section – SEM View

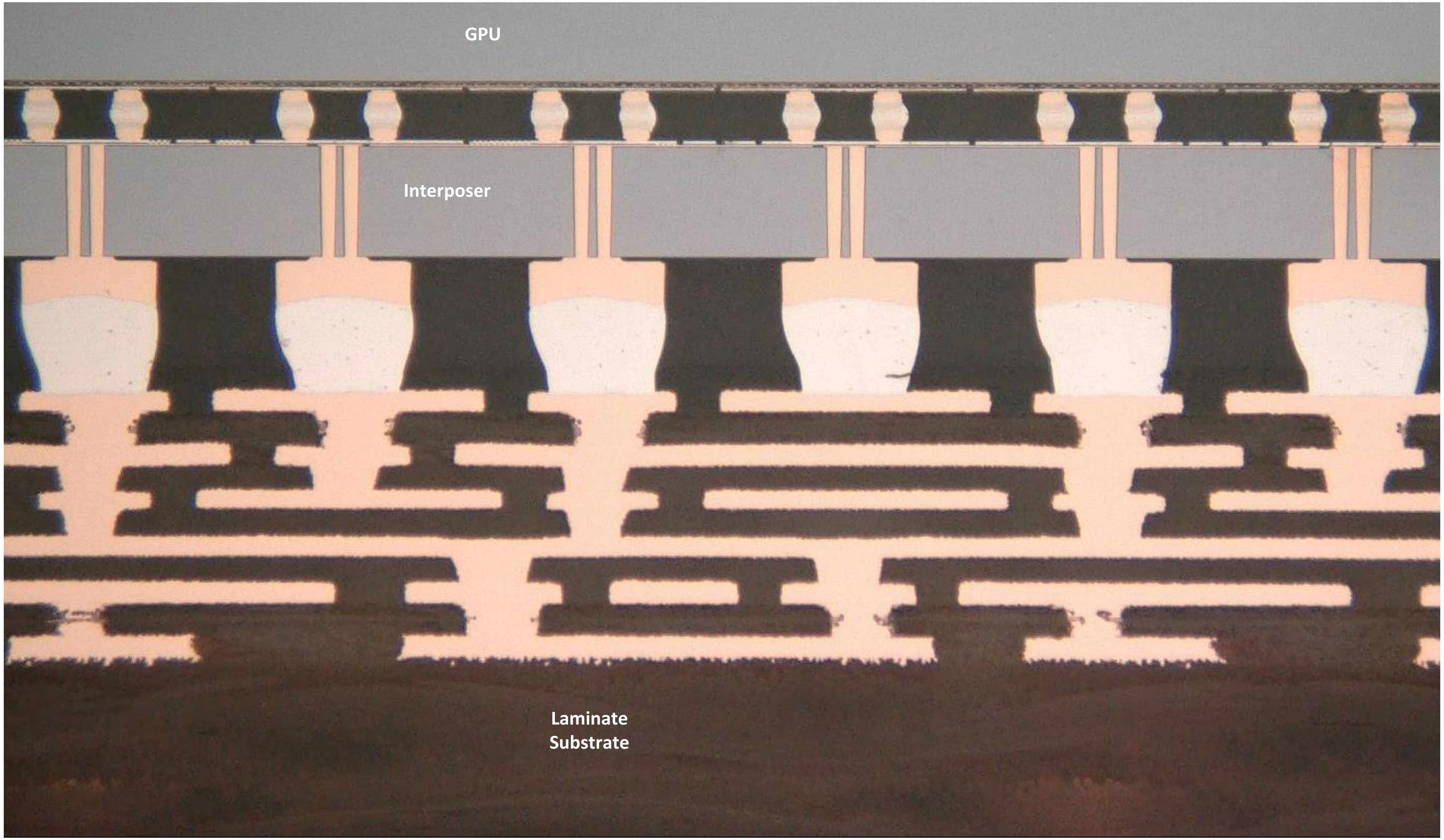
©2018 by System Plus Consulting



Substrate Cross-Section – SEM View

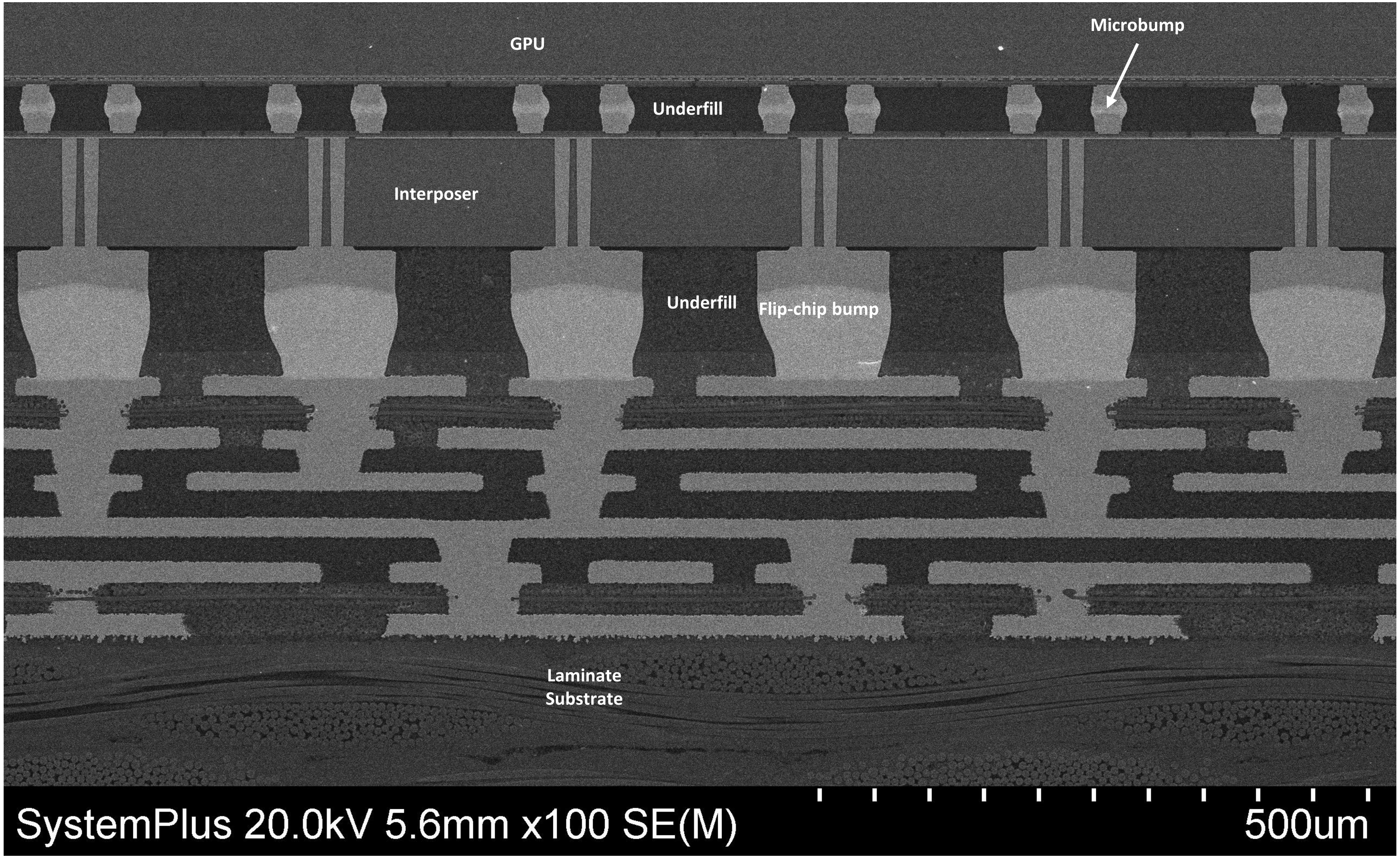
©2018 by System Plus Consulting

Package Cross-Section – Interposer



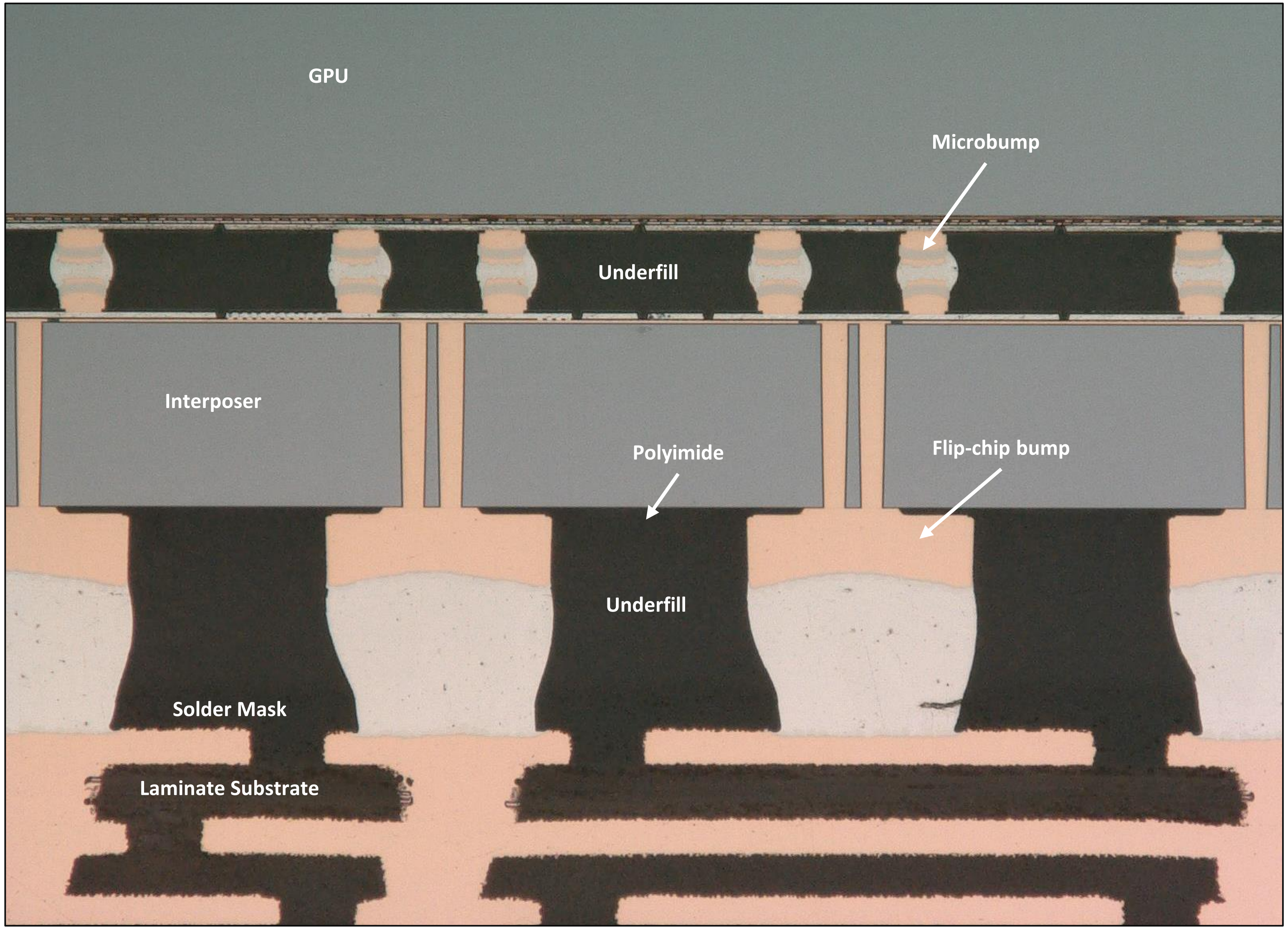
Interposer Cross-Section – Optical View

Package Cross-Section – Interposer



Interposer Cross-Section – SEM View

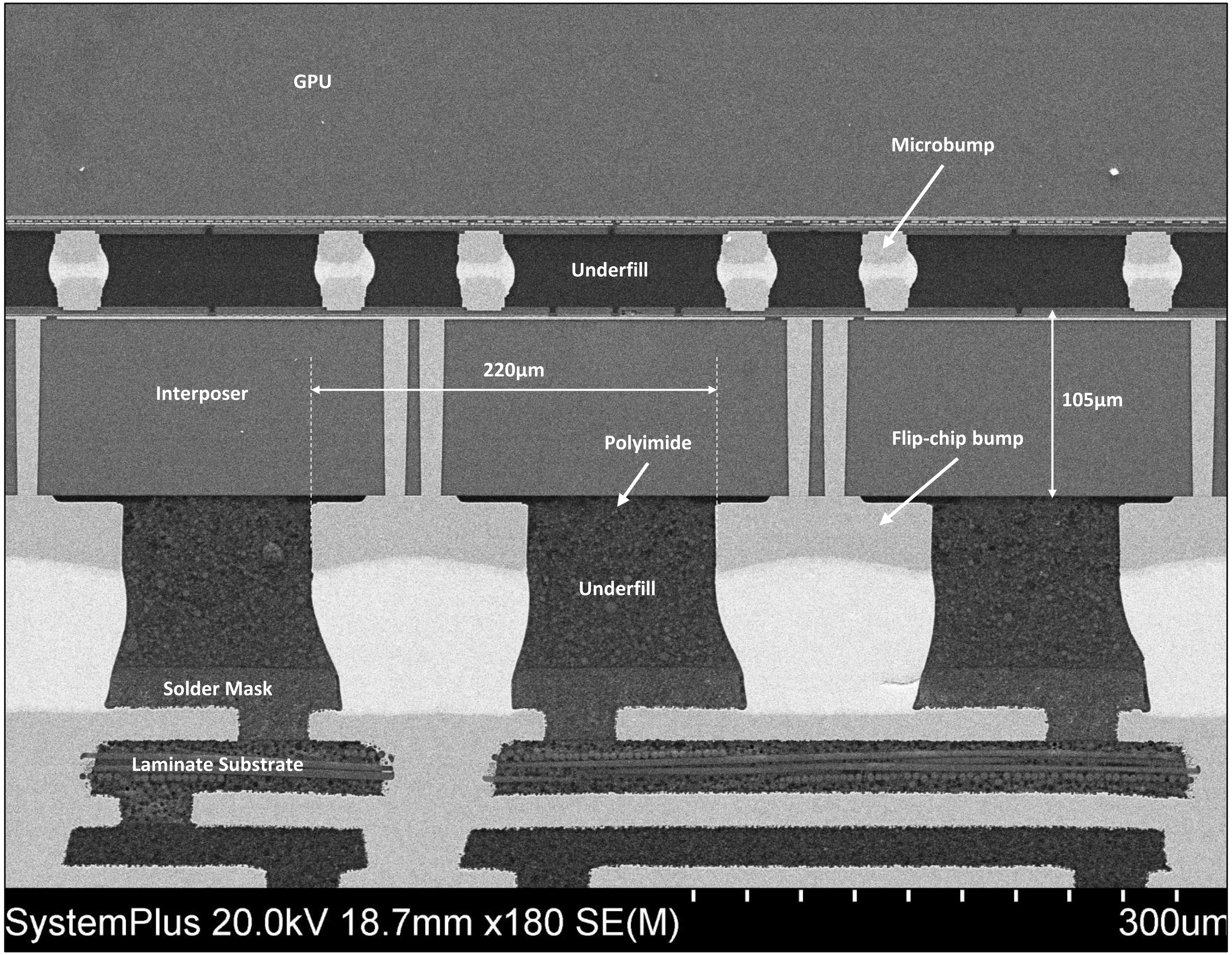
Package Cross-Section – Interposer



Interposer Cross-Section – Optical View

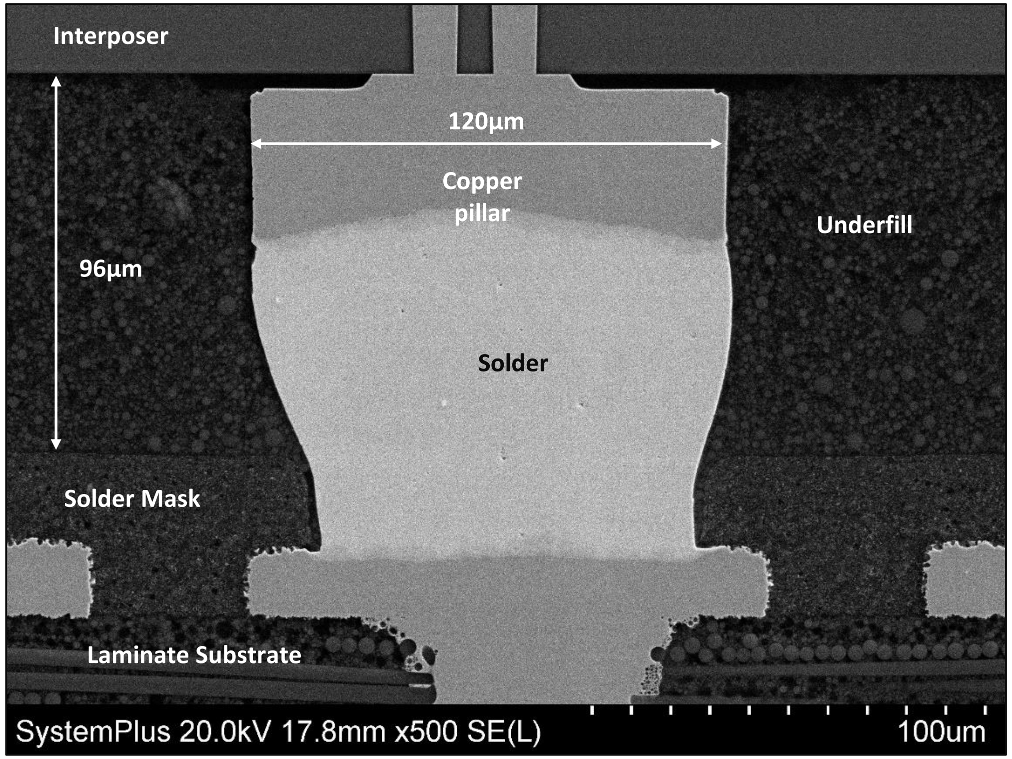
Package Cross-Section – Interposer

- Interposer thickness: 105µm
- Interposer bump pitch: 220µm

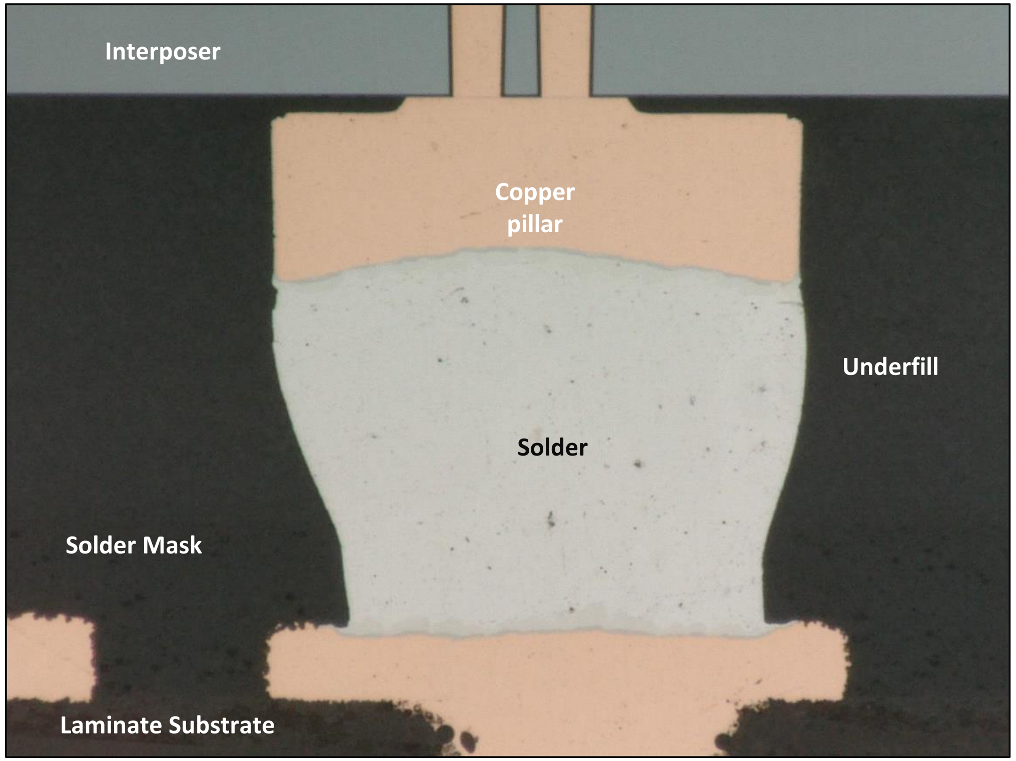


Interposer Cross-Section – SEM View

Package Cross-Section – Interposer



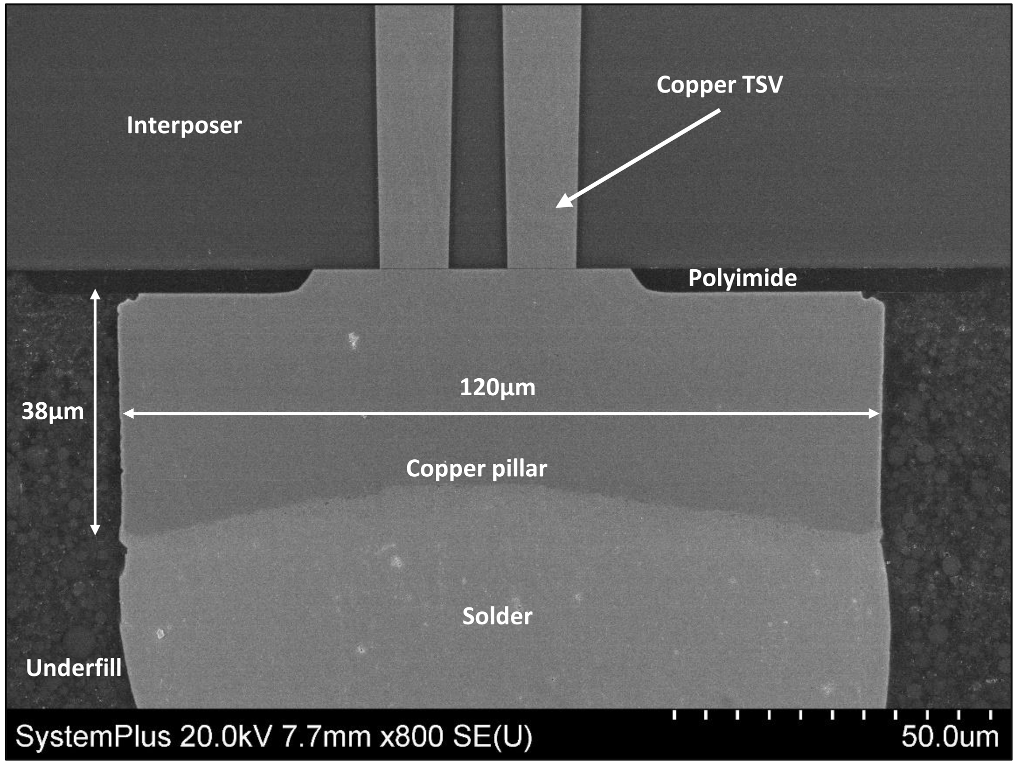
Interposer Cross-Section – SEM View



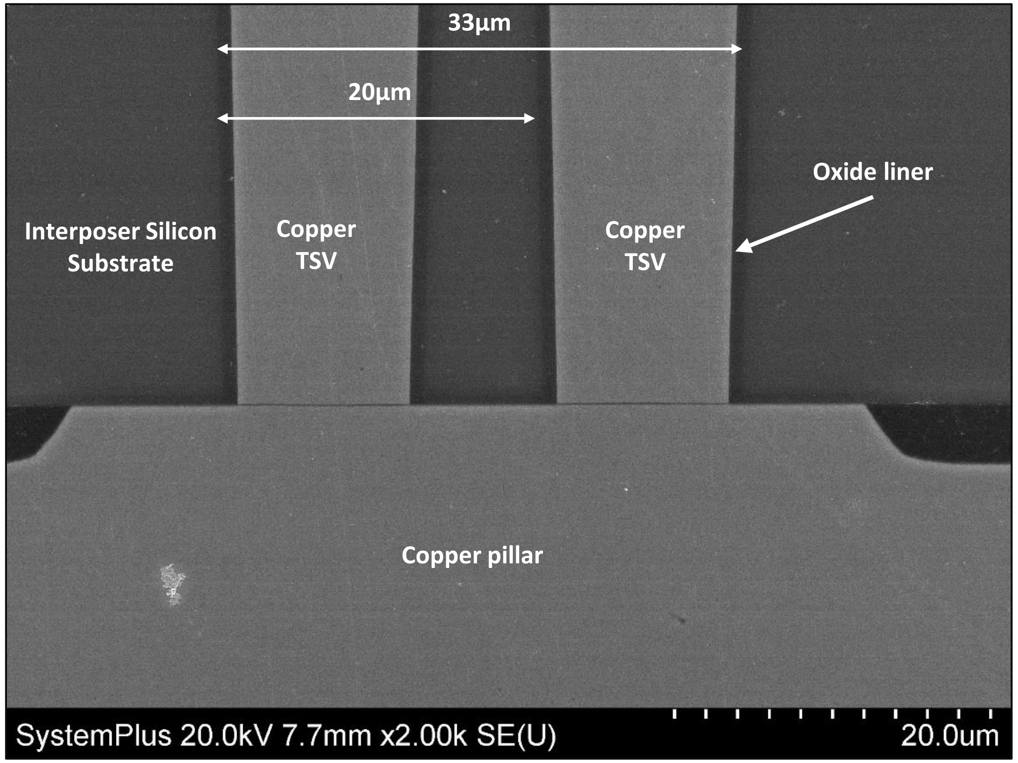
Interposer Cross-Section – Optical View

- Interposer copper pillar diameter: 120µm
- Underfill thickness: 96µm

Package Cross-Section – Interposer

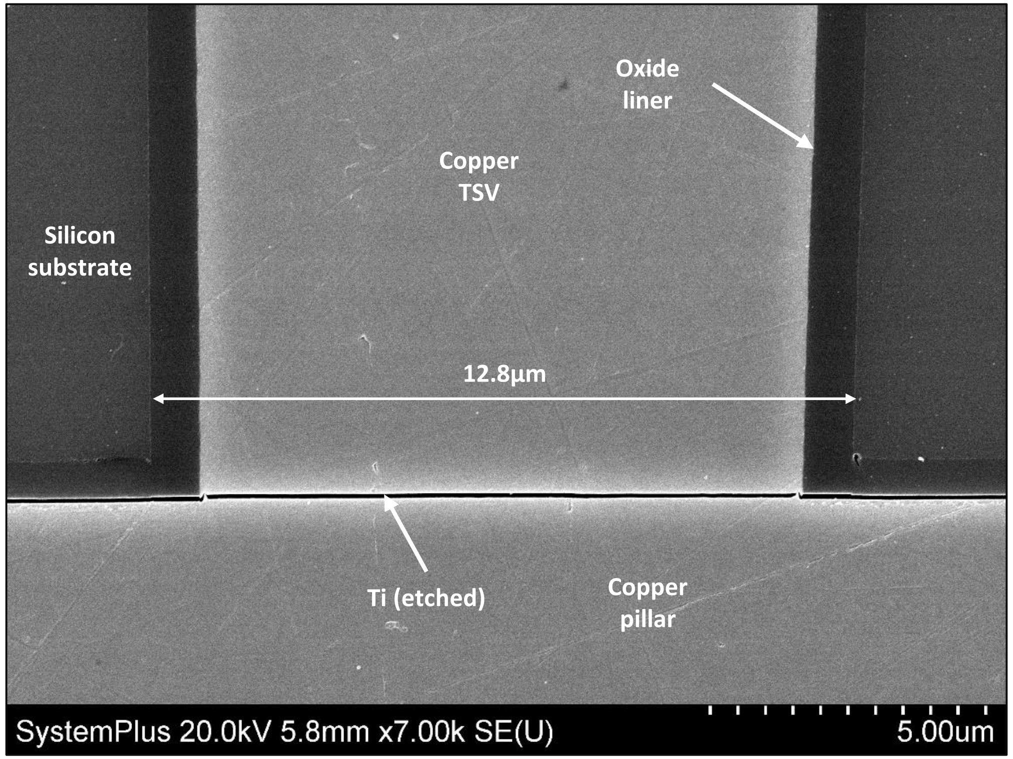


Interposer Cross-Section – SEM View

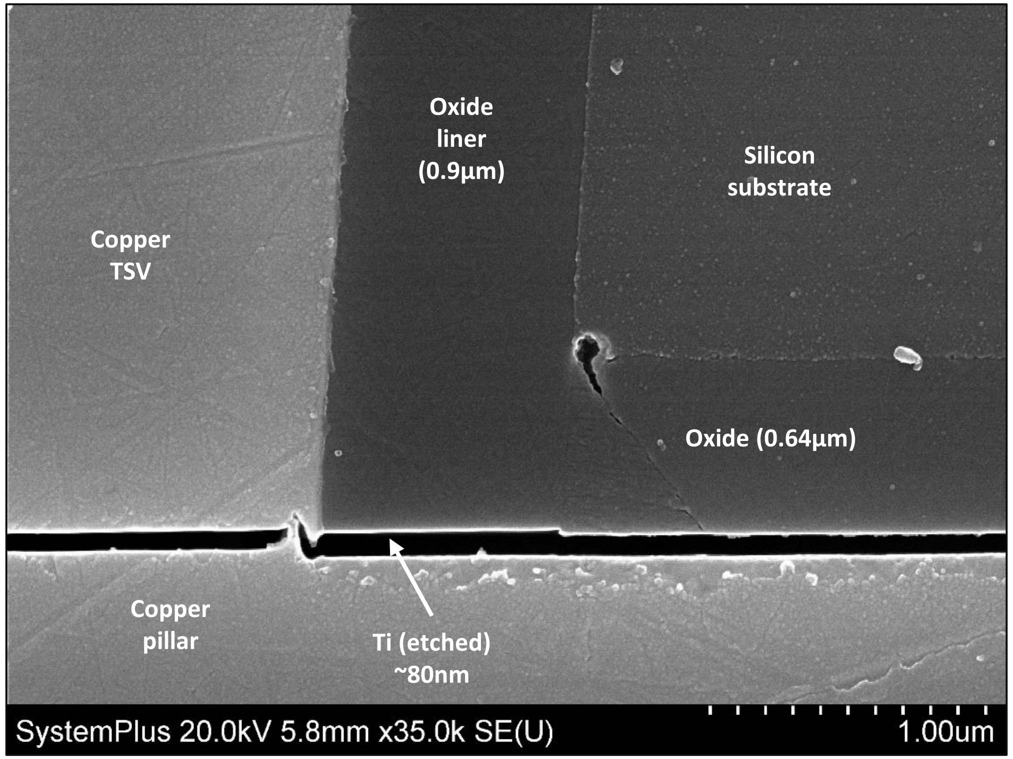


Interposer Cross-Section – SEM View

Package Cross-Section – Interposer

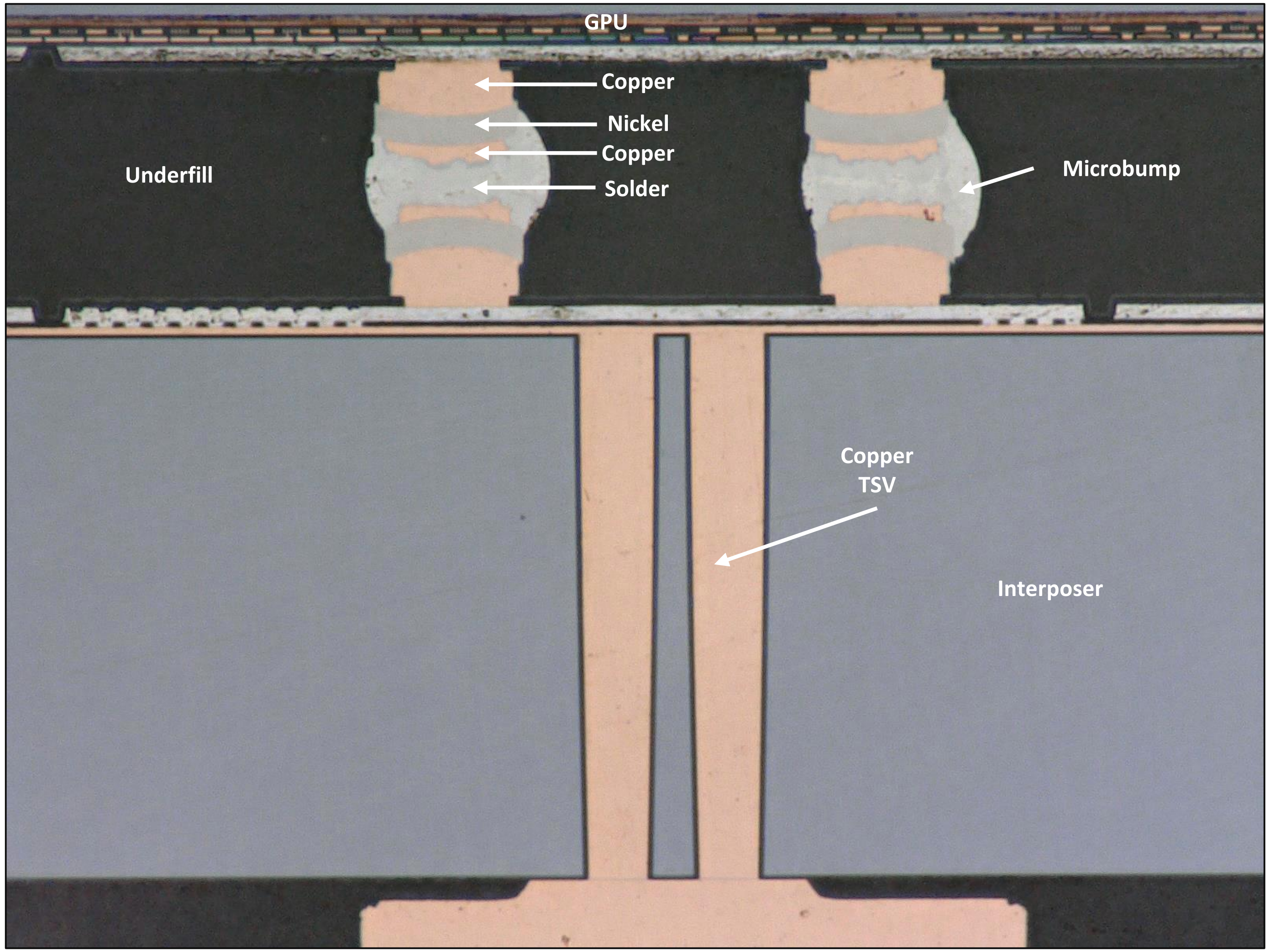


Interposer Cross-Section – SEM View



Interposer Cross-Section – SEM View

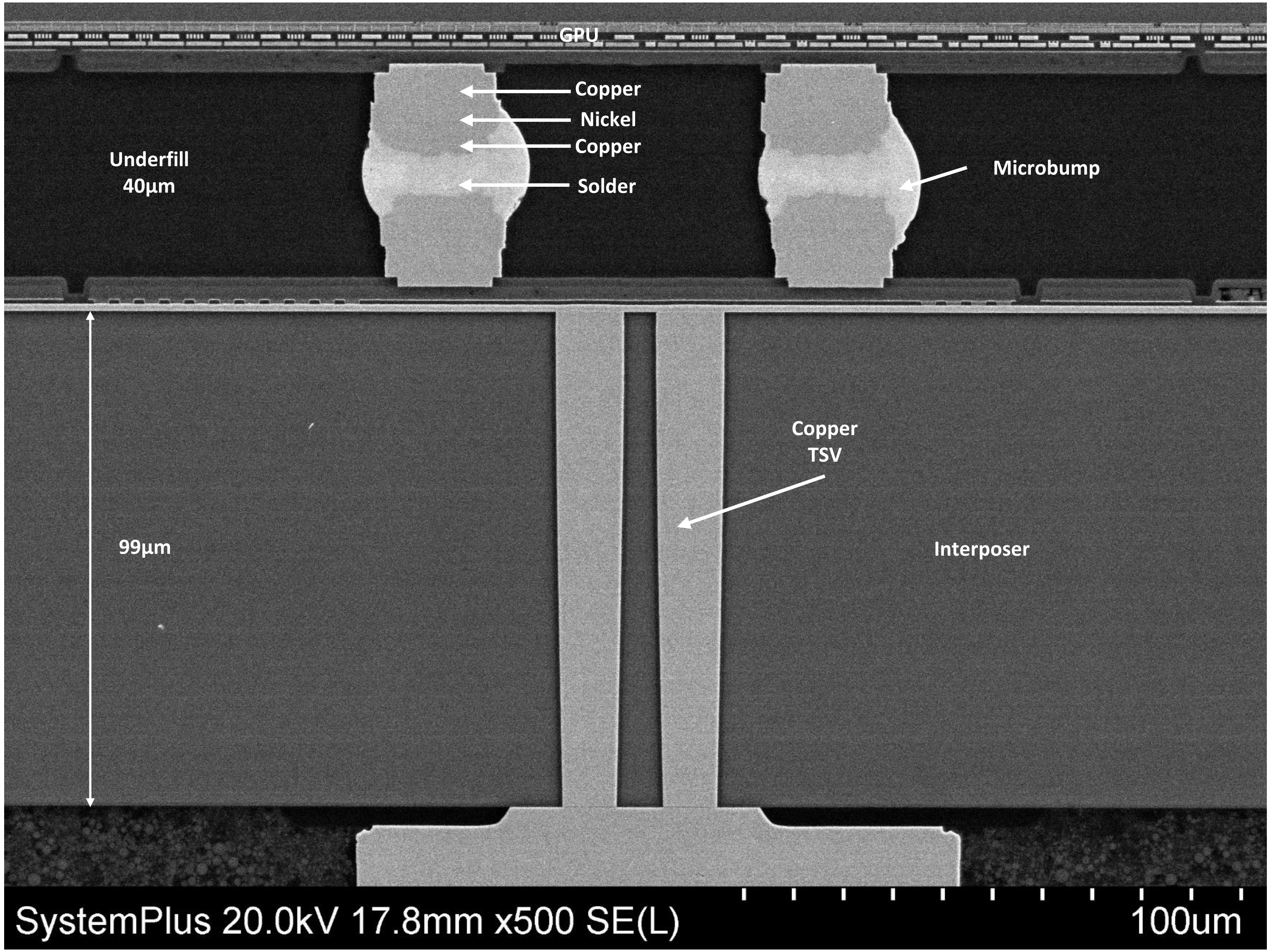
Package Cross-Section – Interposer



Interposer Cross-Section – Optical View

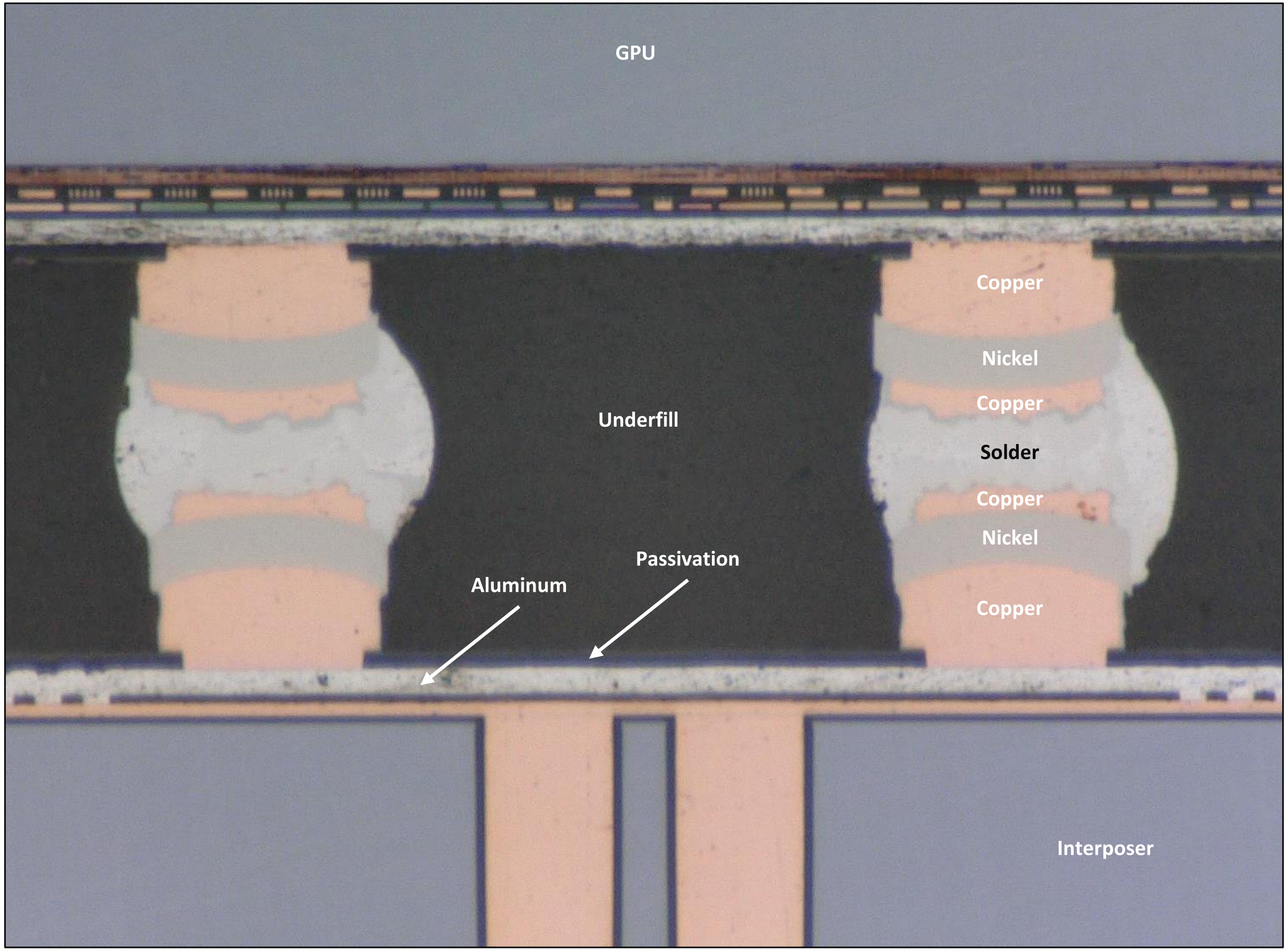
Package Cross-Section – Interposer

- Interposer substrate thickness: **99µm**
- Underfill thickness: **40µm**



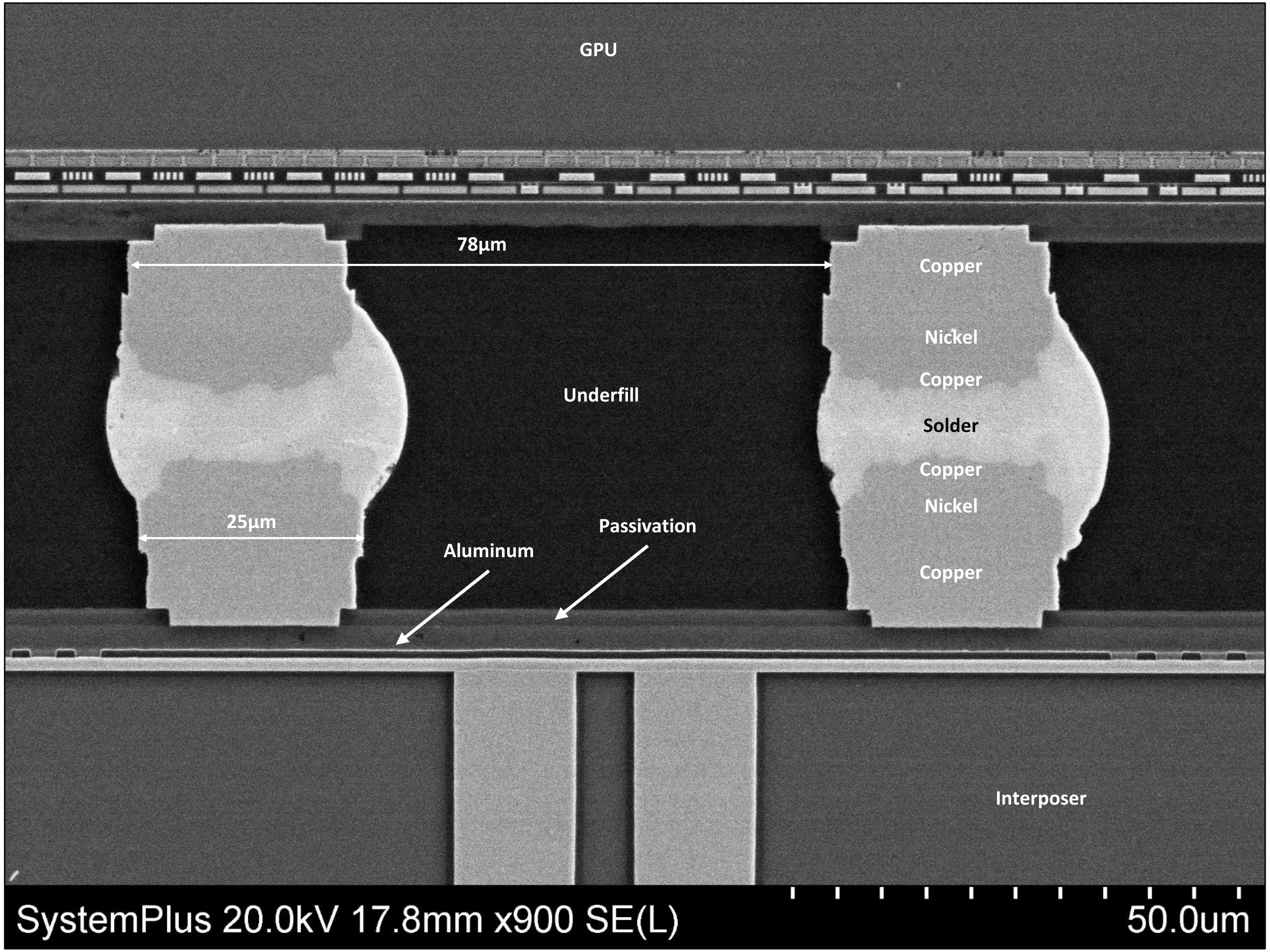
Interposer Cross-Section – SEM View

Package Cross-Section – Interposer



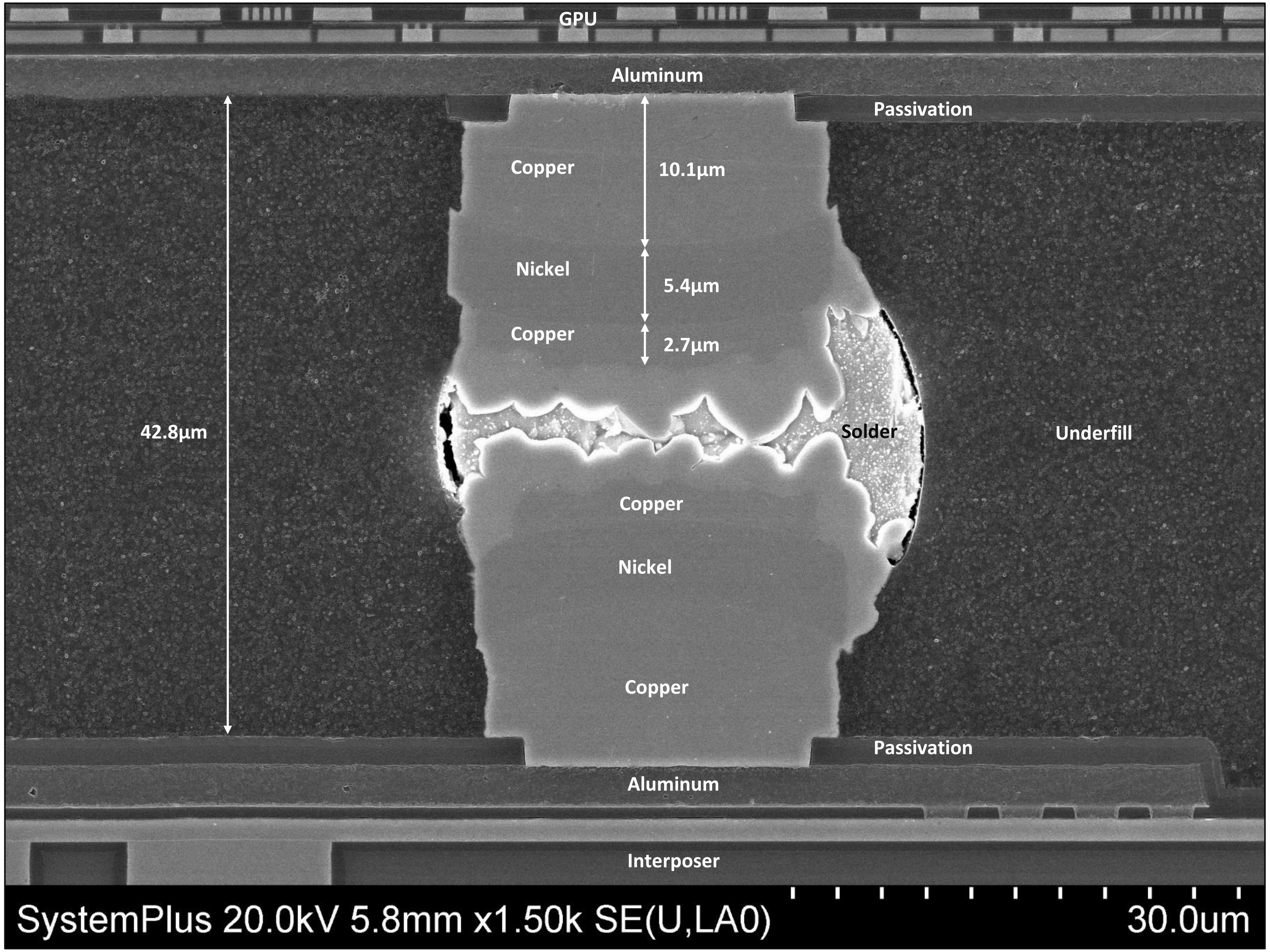
Interposer Cross-Section – Optical View

Package Cross-Section – Interposer



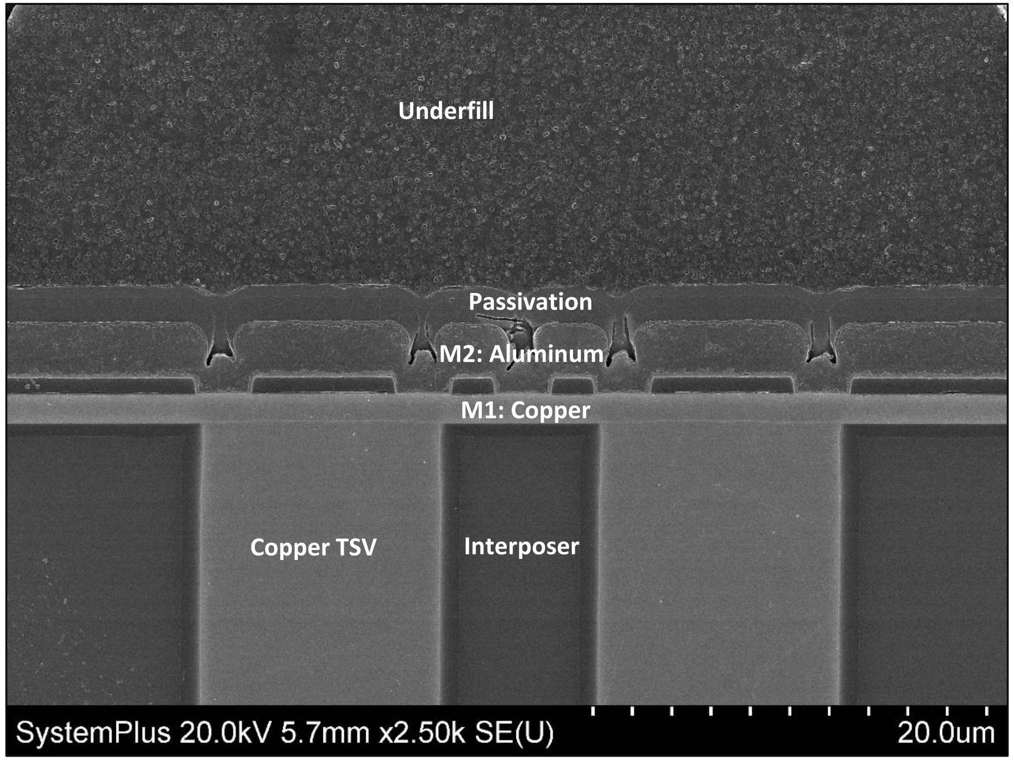
Interposer Cross-Section – SEM View

Package Cross-Section – Interposer

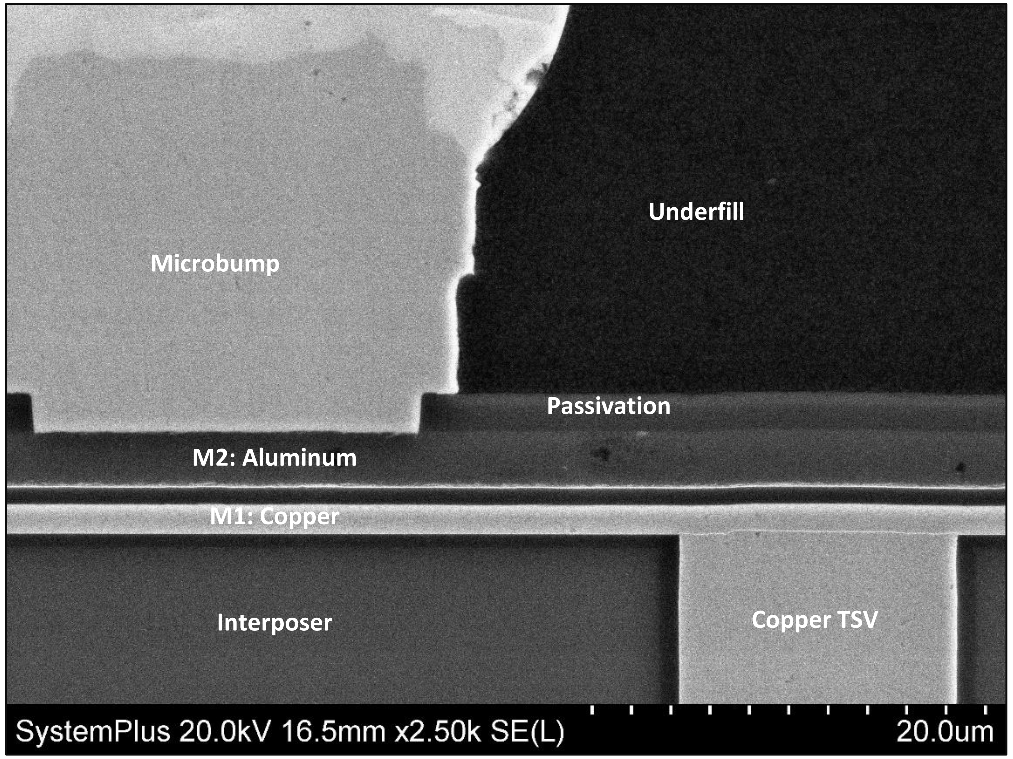


Interposer Cross-Section – SEM View

Package Cross-Section – Interposer



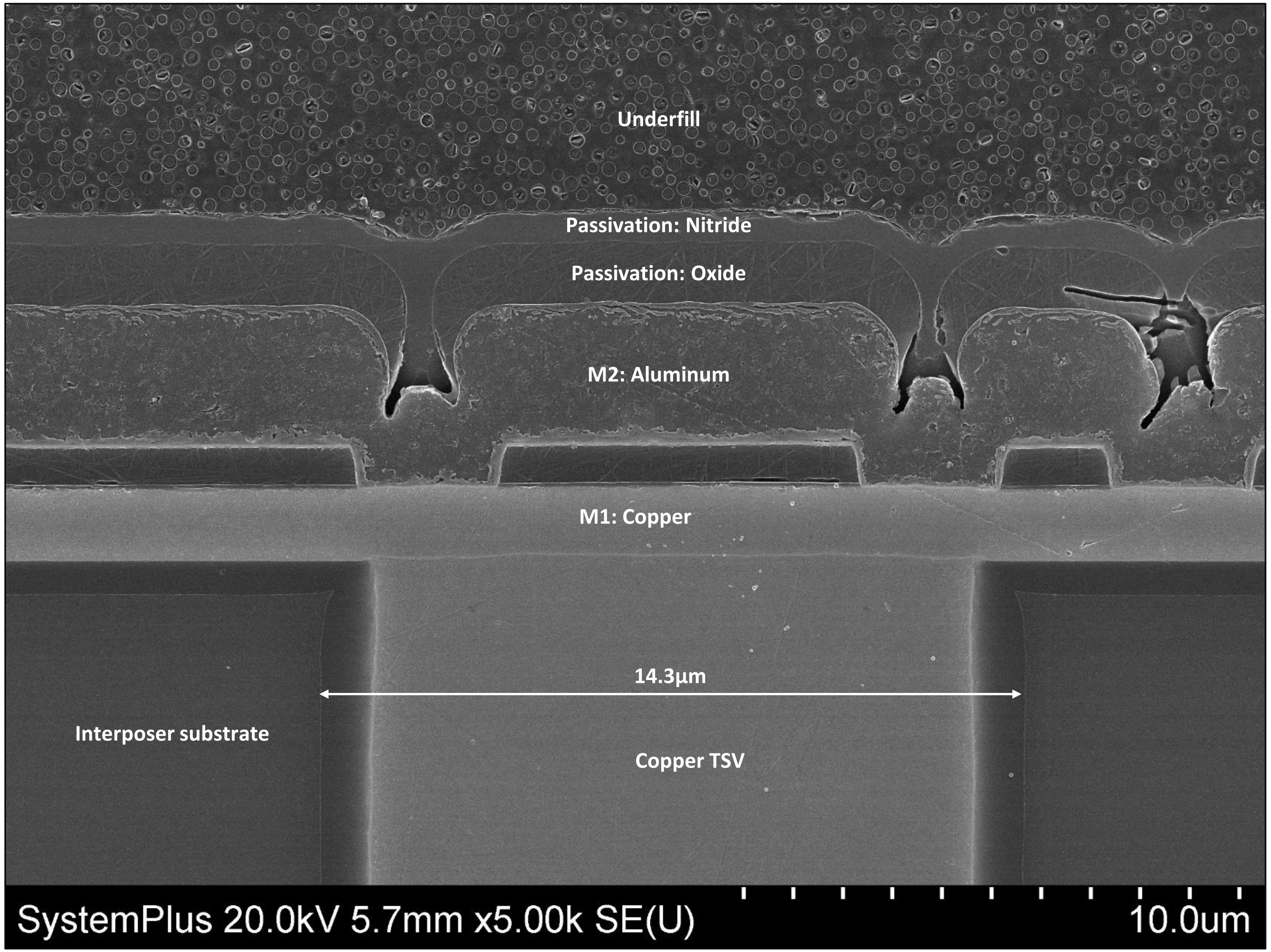
Interposer Cross-Section – SEM View



Interposer Cross-Section – SEM View

- The interposer uses two redistribution metal layers, one in copper and one in aluminum

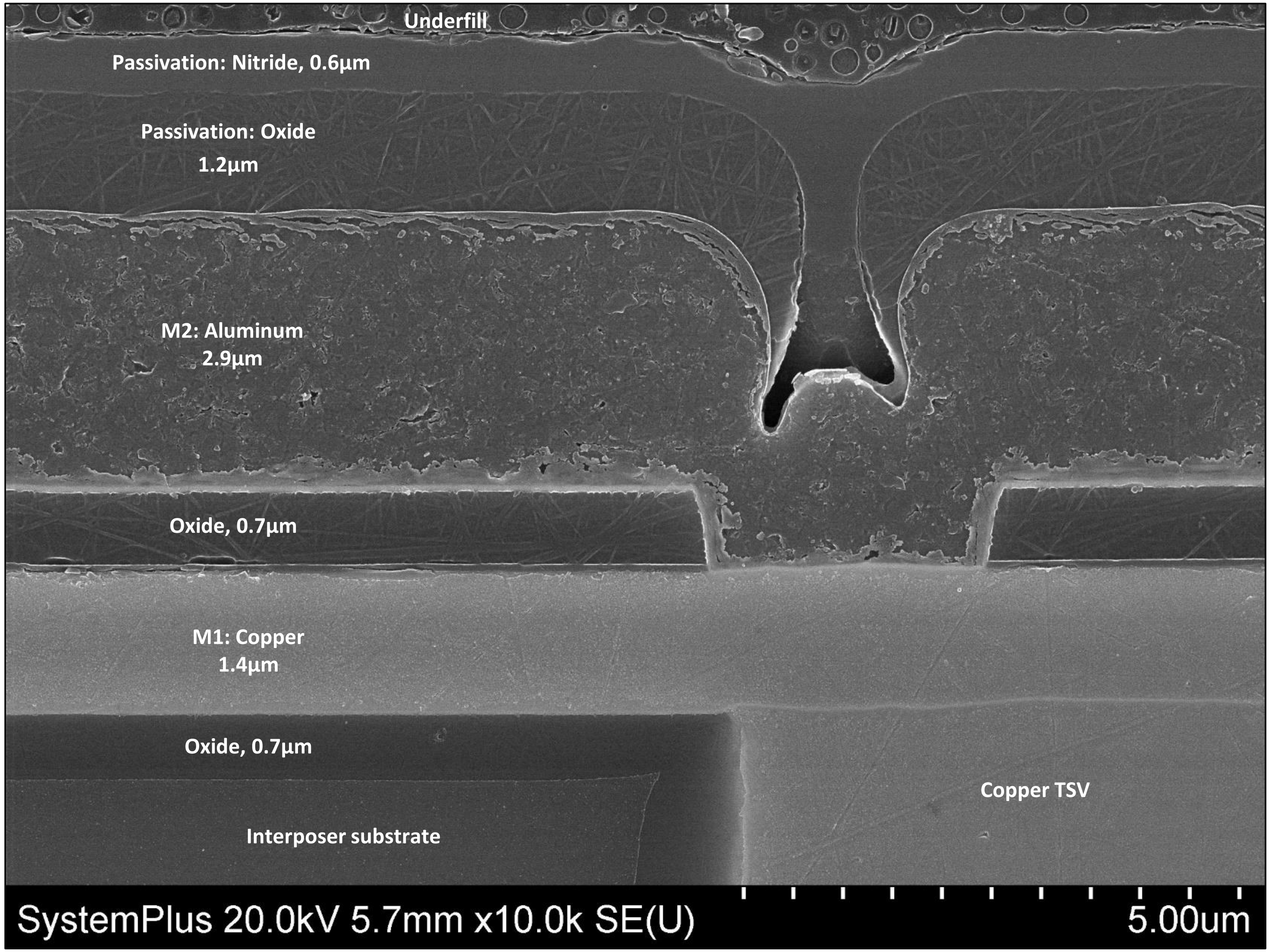
Package Cross-Section – Interposer



Interposer Cross-Section – SEM View

Package Cross-Section – Interposer

- RDL Cu layer thickness: 1.4µm
- RDL Al layer thickness: 2.9µm



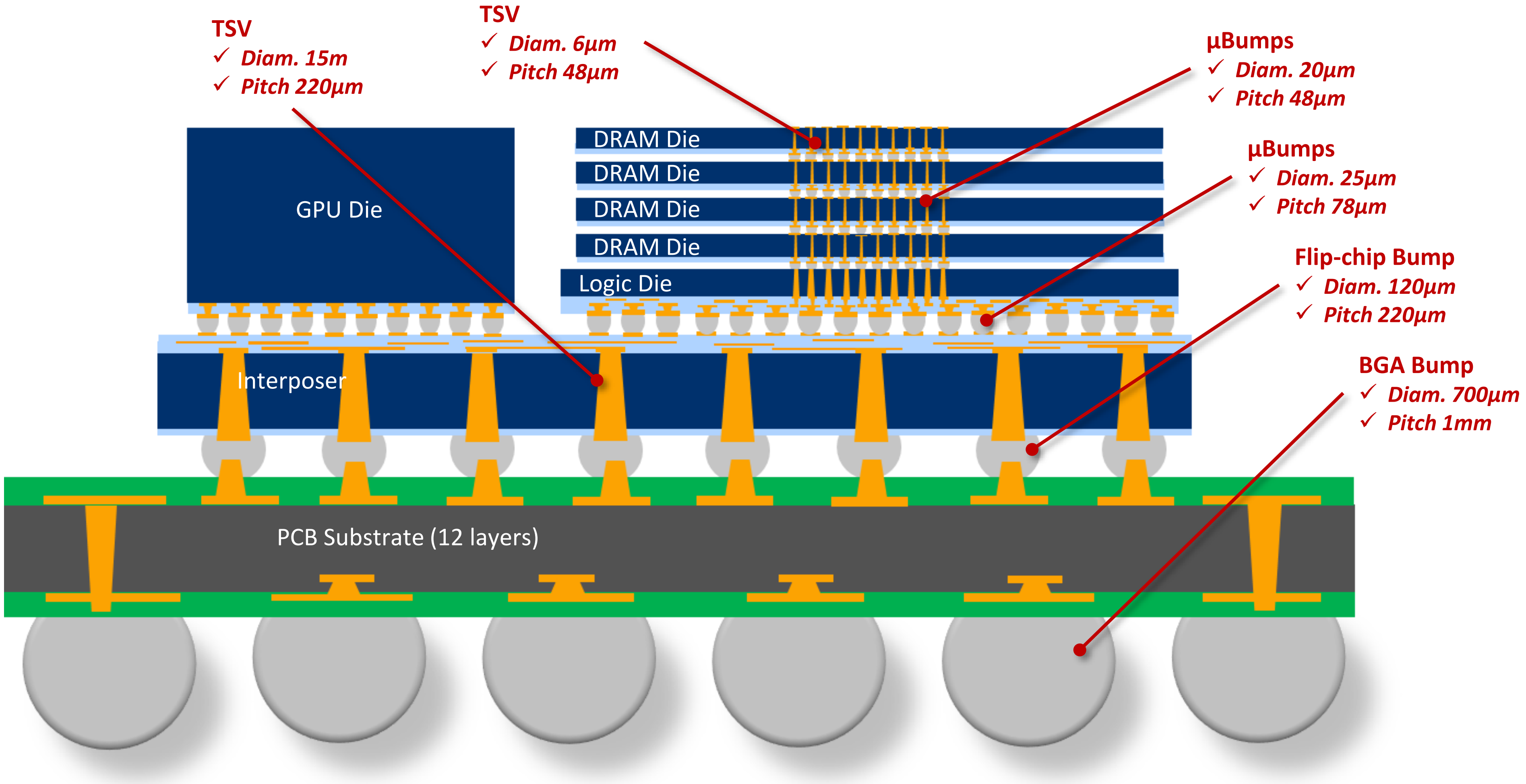
Interposer Cross-Section – SEM View

MANUFACTURING
PROCESS FLOW

Global Overview

Package Structure:

- 3D Packaging: 5 stacked dies with TSV & µBumps (HBM stack).
- 2.5D Packaging: HBM stack and GPU stacked with µBumps and a silicon interposer holding TSV.
- Flip-chip BGA: silicon interposer flip-chipped to a 12-layers PCB substrate



Global Overview

- TSMC CoWoS Process Steps:

CoW

Interposer Process
- TSV
- BEOL Metals
- Microbumping

GPU Pick & Place

HBM Stack Pick & Place

Solder Reflow

Underfill Deposition & Cure

Wafer Molding

Mold Thinning

Interposer backside processing
- Carrier Bonding
- Thinning & Via Reveal
- Bumping
- Debonding & Dicing

oS

Pick & Place Interposer on laminate substrate

Reflow

Underfill Deposition & Cure

GPU Process

- GPU Front-End Process:
  - Substrate: 300mm Silicon wafer
  - Process type: CMOS (Logic, Analog, Memory)
  - Metal layers: 15 (14 Cu + 1 Al)
  - Process: FinFET
  - Technology node: 12nm
  - GPU Area: $828mm^2$

- Test:
  - Test type: Wafer sort (probe test)

- *Note: The process flow of this technology is standard. The above parameters are enough to estimate the manufacturing cost of the die.*

GPU Wafer Fabrication Unit

- We assume that the manufacturing of the GPU is made by TSMC on 300mm wafers.

- Wafer fab unit:
  - Name: TSMC Fab 12 phase 6
  - Wafer diameter: 300mm (12-inch)
  - Capacity: 25,000 wafers / month
  - Year of start: 2013
  - Products: Foundry
  - Location: Hsinchu, Taiwan

- This manufacturing line has been created in 2013.
  - We assume that both clean room and equipment are still in depreciation.

HBM – DRAM & Logic Dies Process

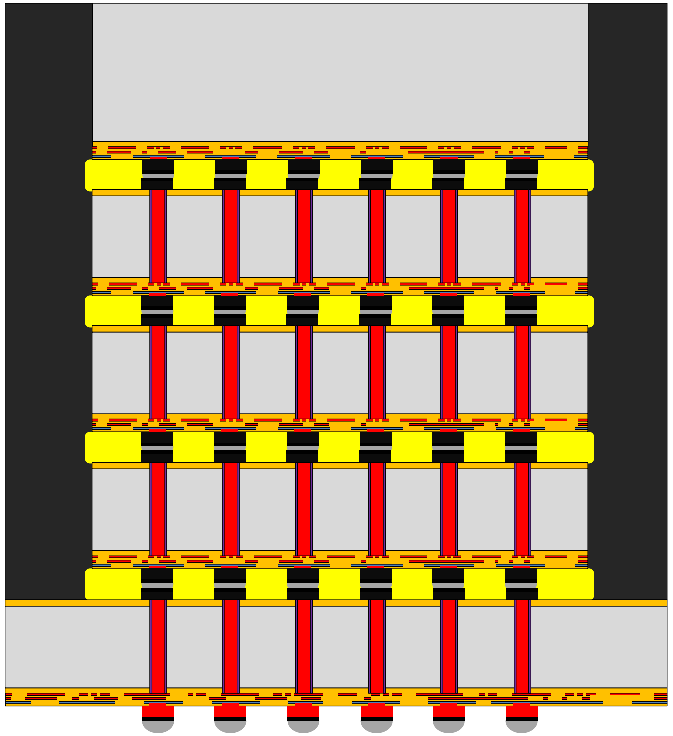
- DRAM Front-End Process:
  - Substrate: 300mm Silicon wafer
  - Process type: DRAM
  - Metal layers: 4 (3 Cu + 1 Al)
  - Technology node: 20nm
  - DRAM Die Area: 86mm²

- Logic Die Front-End Process:
  - Substrate: 300mm Silicon wafer
  - Process type: CMOS
  - Metal layers: 4 (3 Cu + 1 Al)
  - Technology node: 20nm
  - Logic Die Area: 96mm²

- *Note: The process flow of these technology are standard. The above parameters are enough to estimate the manufacturing cost of the die.*

HBM – TSV, Bumping & Stacking Process Flow

- TSV & Bumping Process:
  - Substrate: Silicon 12-inch
  - TSV filling: Bottom-up filling
  - Temporary bonding: TMAT process
    - ✓ Adhesive: Elastomer
    - ✓ Carrier: Silicon



HBM Stacking Structure

Samsung HBM TSV Process (DRAM)
- Lithography
- Oxide Etch : RIE (2µm)
- Silicon Etch : DRIE (50µm depth, 6µm diam)
- PR Strip : Resist Strip
- Oxide Liner: PECVD (SiO2, 0.2µm)
- Barrier/Seed : PVD-Tantalum (Ta, 0.05µm)
- Barrier/Seed : PVD-Copper (Cu, 0.1µm)
- CU filling : Electroplating-Copper (Cu)
- TSV Anneal
- Cu & Ta CMP

Samsung HBM Micro-Bumping Process (DRAM)
- UBM : PVD Titanium (Ti, 0.1µm)
- UBM : PVD Copper (Cu, 0.3µm)
- UBM : Lithography
- UBM : Electroplating-Nickel (Ni, 5µm)
- Bumping : Electroplating Solder (SnAg, 10µm)
- Bumping : Resist Strip
- Bumping : Wet Etching Copper
- Bumping : Wet Etching Titanium
- Temporary Bonding : Edge Trimming
- Temporary Bonding : Spin Coat Elastomer
- Temporary Bonding : bonding to carrier
- Backside Thinning
- TSV Via Reveal : Si etch
- TSV Via Reveal : PECVD Oxide (Passivation)
- TSV Via Reveal : CMP
- UBM : PVD Titanium (Ti, 0.1µm)
- UBM : Lithography
- UBM : Wet Etching Titanium
- UBM : Resist Strip
- UBM : Electroplating-Nickel (Ni, 4µm)
- Carrier debonding
- Vacuum Laminate NCF (Underfill)

Samsung HBM Micro-Bumping Process (Logic)
- UBM : PVD Titanium (Ti, 0.1µm)
- UBM : PVD Copper (Cu, 0.3µm)
- UBM : Lithography
- UBM : Electroplating-Copper (Cu, 17µm)
- UBM : Electroplating-Nickel (Ni, 3µm)
- Bumping : Electroplating Solder (SnAg, 10µm)
- Bumping : Resist Strip
- Bumping : Wet Etching Copper
- Bumping : Wet Etching Titanium
- Temporary Bonding : Edge Trimming
- Temporary Bonding : Spin Coat Elastomer
- Temporary Bonding : bonding to carrier
- Backside Thinning
- TSV Via Reveal : Si etch
- TSV Via Reveal : PECVD Oxide (Passivation)
- TSV Via Reveal : CMP
- UBM : PVD Titanium (Ti, 0.1µm)
- UBM : Lithography
- UBM : Wet Etching Titanium
- UBM : Resist Strip
- UBM : Electroplating-Nickel (Ni, 4µm)

Samsung HBM Stacking Process
- DRAM Die 1 TC Bonding
- NCF Post Cure
- DRAM Die 2 TC Bonding
- NCF Post Cure
- DRAM Die 3 TC Bonding
- NCF Post Cure
- DRAM Die 4 TC Bonding
- NCF Post Cure
- Wafer Molding
- Carrier Debonding
- Dicing

HBM – TSV Process Flow (1/2)

DRAM Die
Logic Die

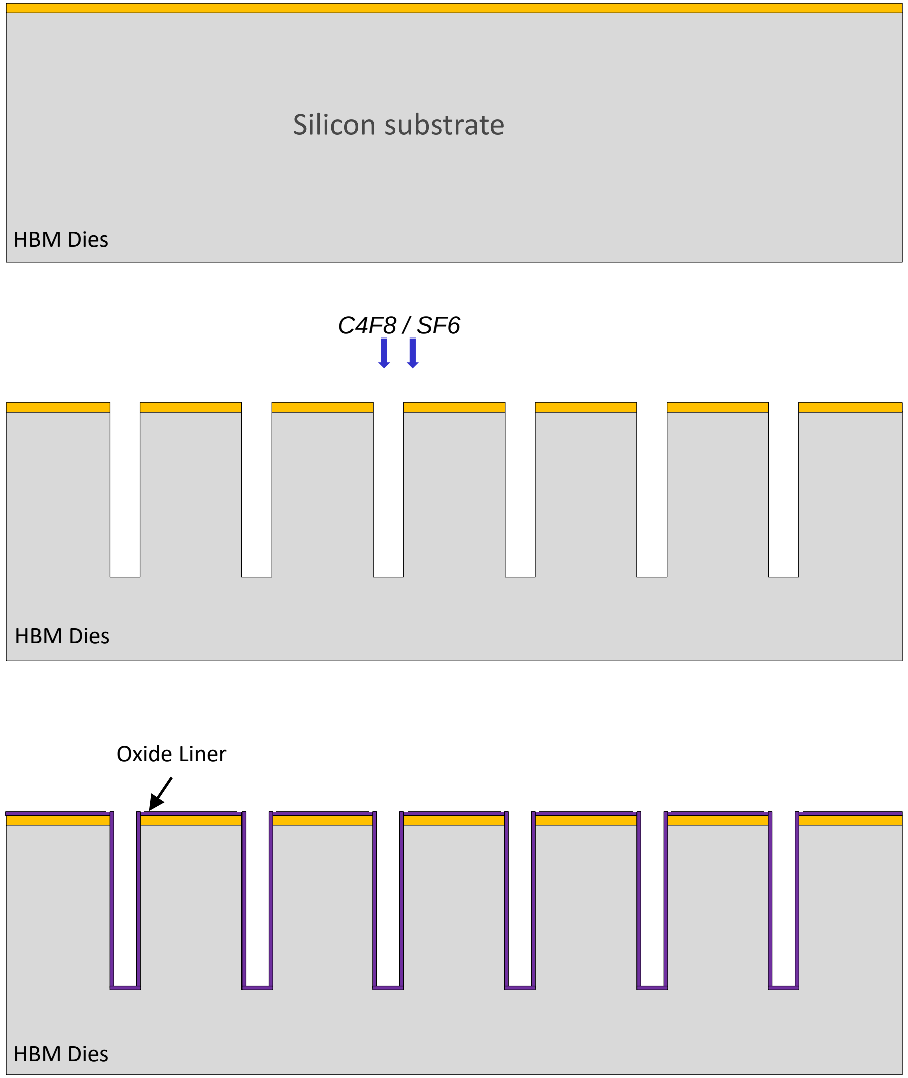
- FEOL

DRAM Die
Logic Die

- TSV etch: DRIE

DRAM Die
Logic Die

- Oxide liner deposition: SiO2 PECVD

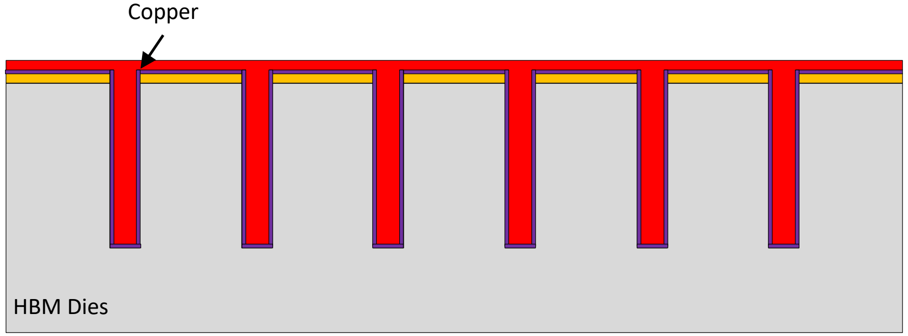


drawing not to scale

HBM – TSV Process Flow (2/2)

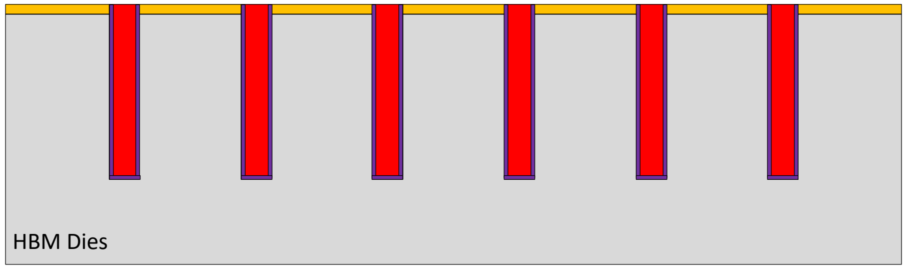
DRAM Die / Logic Die

- TSV barrier/Seed layers deposition
- TSV filling: Copper ECD



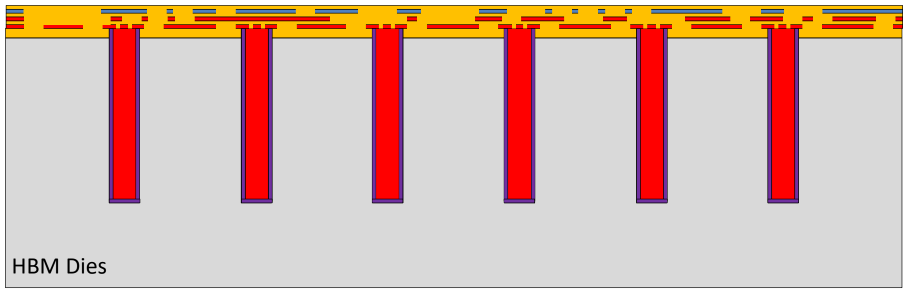
DRAM Die / Logic Die

- CMP



DRAM Die / Logic Die

- BEOL

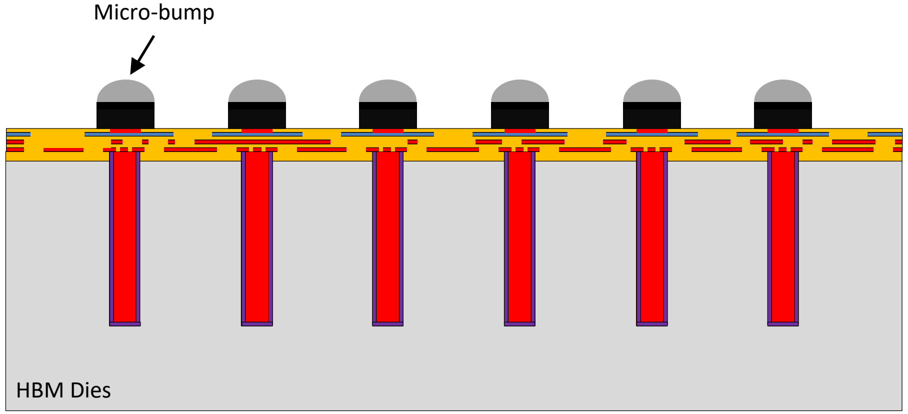


drawing not to scale

HBM – Micro-Bumping Process Flow (1/4)

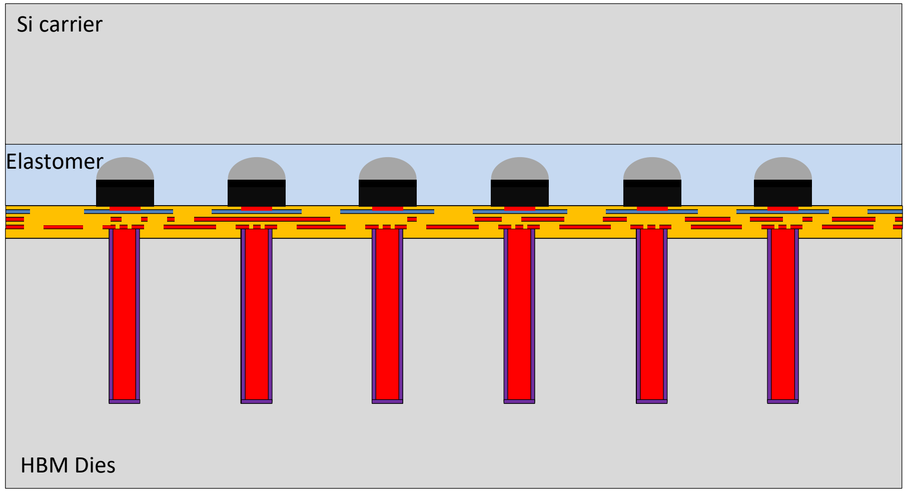
DRAM Die
Logic Die

- UBM: Titanium + nickel
- Bumping



DRAM Die
Logic Die

- Temporary bonding

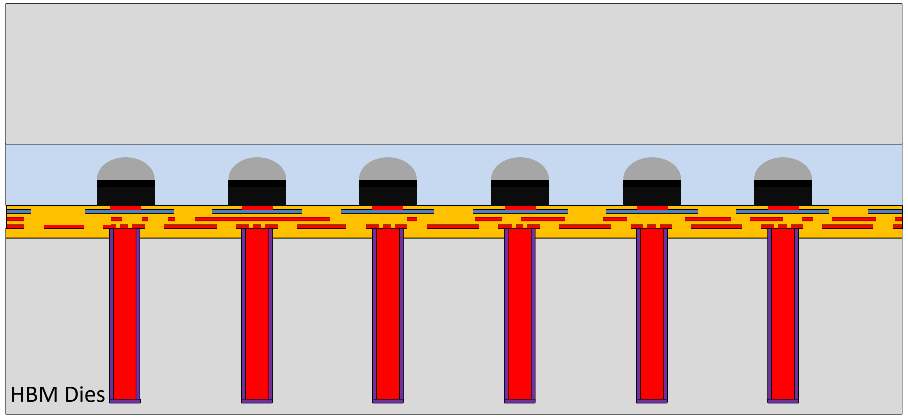


drawing not to scale

HBM – Micro-Bumping Process Flow (2/4)

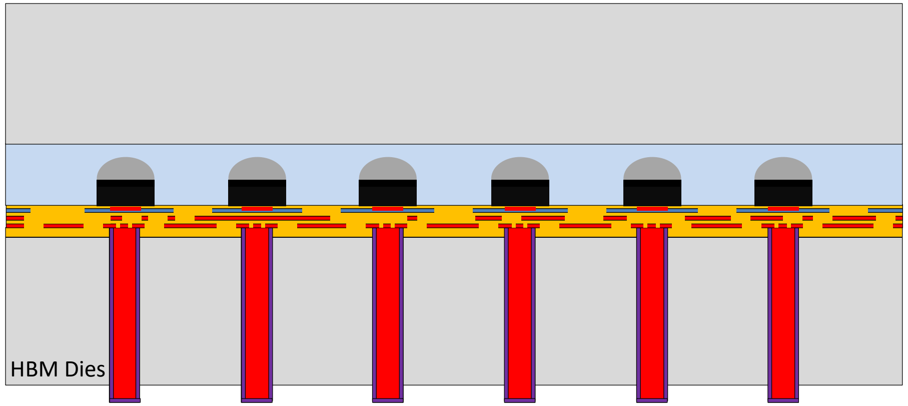
DRAM Die
Logic Die

- Backside thinning



DRAM Die
Logic Die

- Via reveal: Si etch

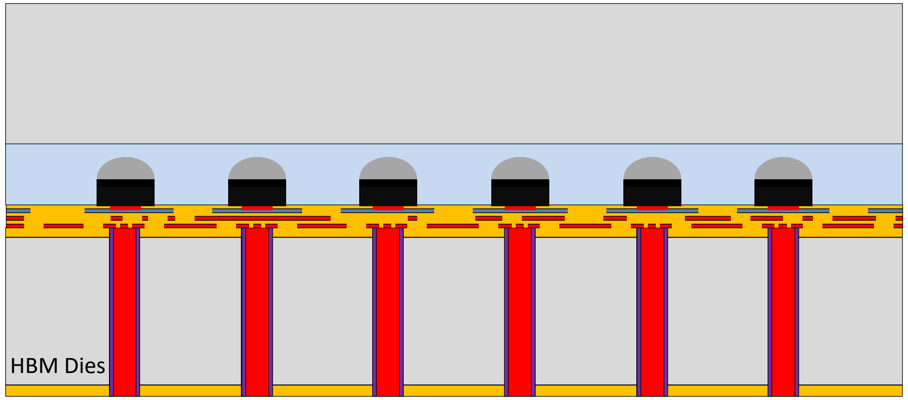


HBM – Micro-Bumping Process Flow (3/4)

DRAM Die

Logic Die

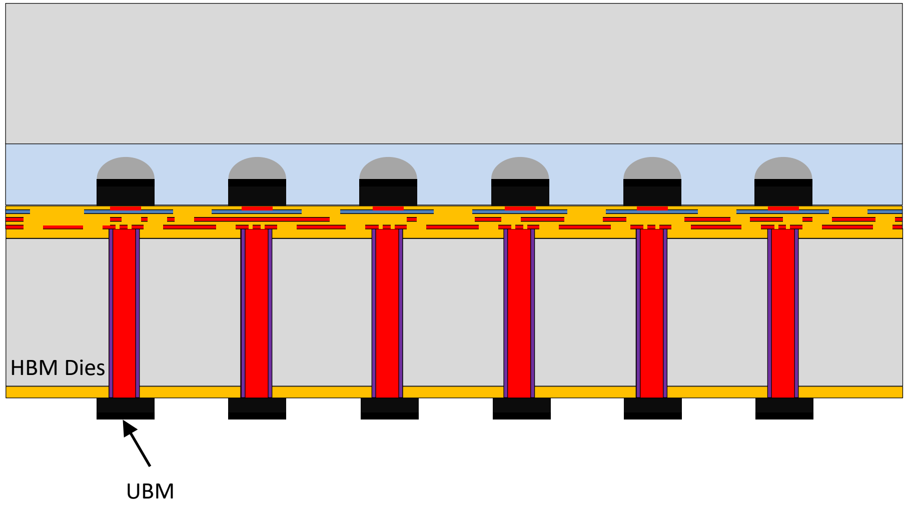
- Via reveal: Oxide Passivation + CMP



DRAM Die

Logic Die

- UBM: Ti + Cu deposition & pattern + Nickel Deposition



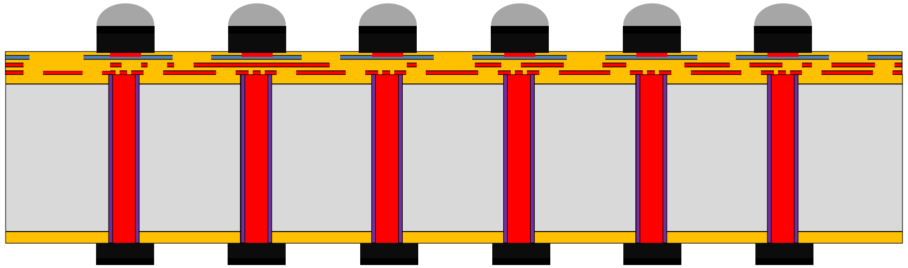
drawing not to scale

HBM – Micro-Bumping Process Flow (4/4)

DRAM Die

Logic Die

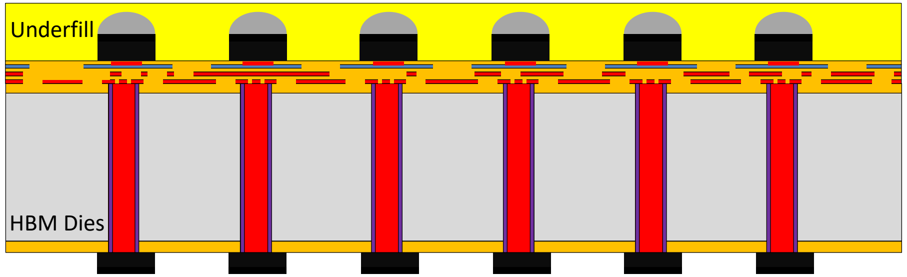
- De-bonding



DRAM Die

Logic Die

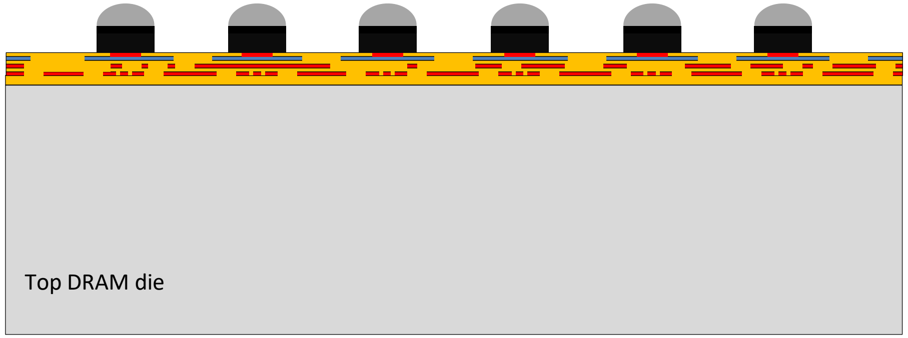
- Vacuum laminate NCF Underfill
- Dicing



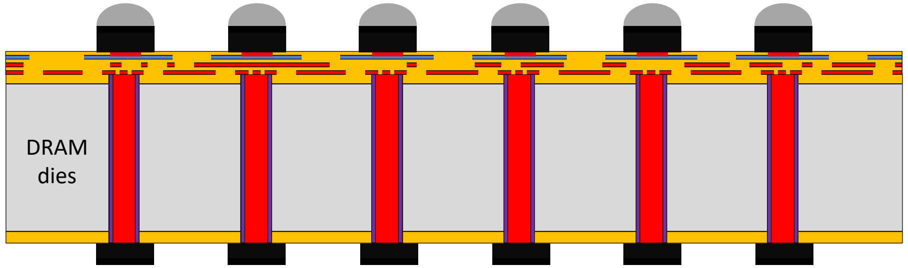
drawing not to scale

HBM Stacking Process Flow (1/4)

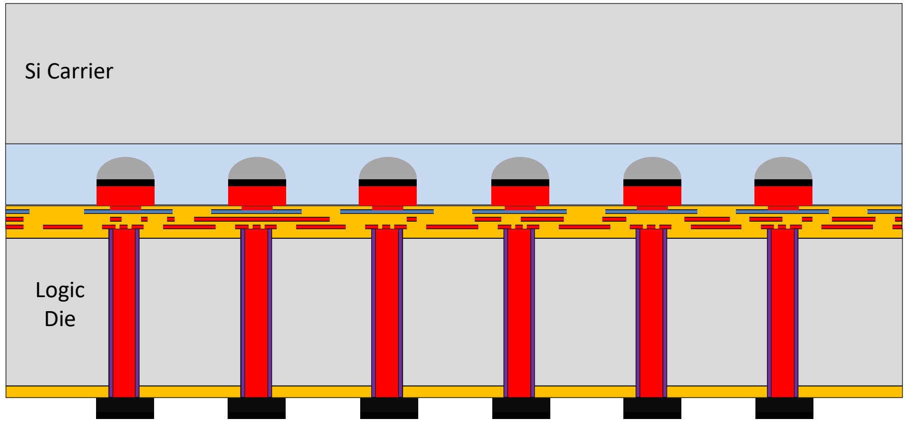
Top DRAM Die
- Die thickness 400µm
- TSVs
- Micro-Bumps



Other DRAM Dies
- Die thickness 50µm
- TSVs
- Micro-Bumps



Logic Dies
- Die thickness 50µm
- TSVs
- Micro-Bumps

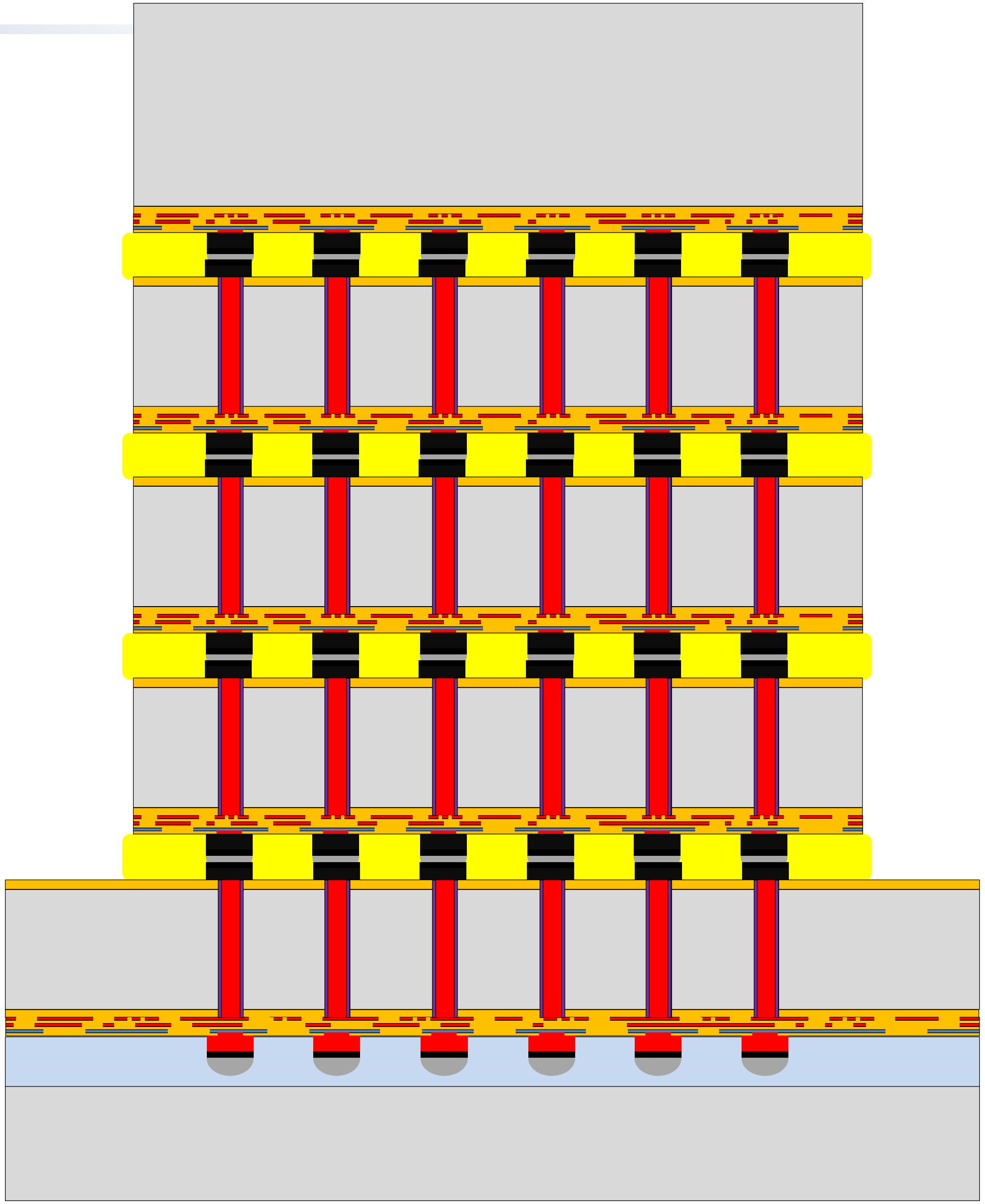


drawing not to scale

HBM Stacking Process Flow (2/4)

HBM Stack

- TCB – DRAM dies Thermocompression Bonding on logic wafer

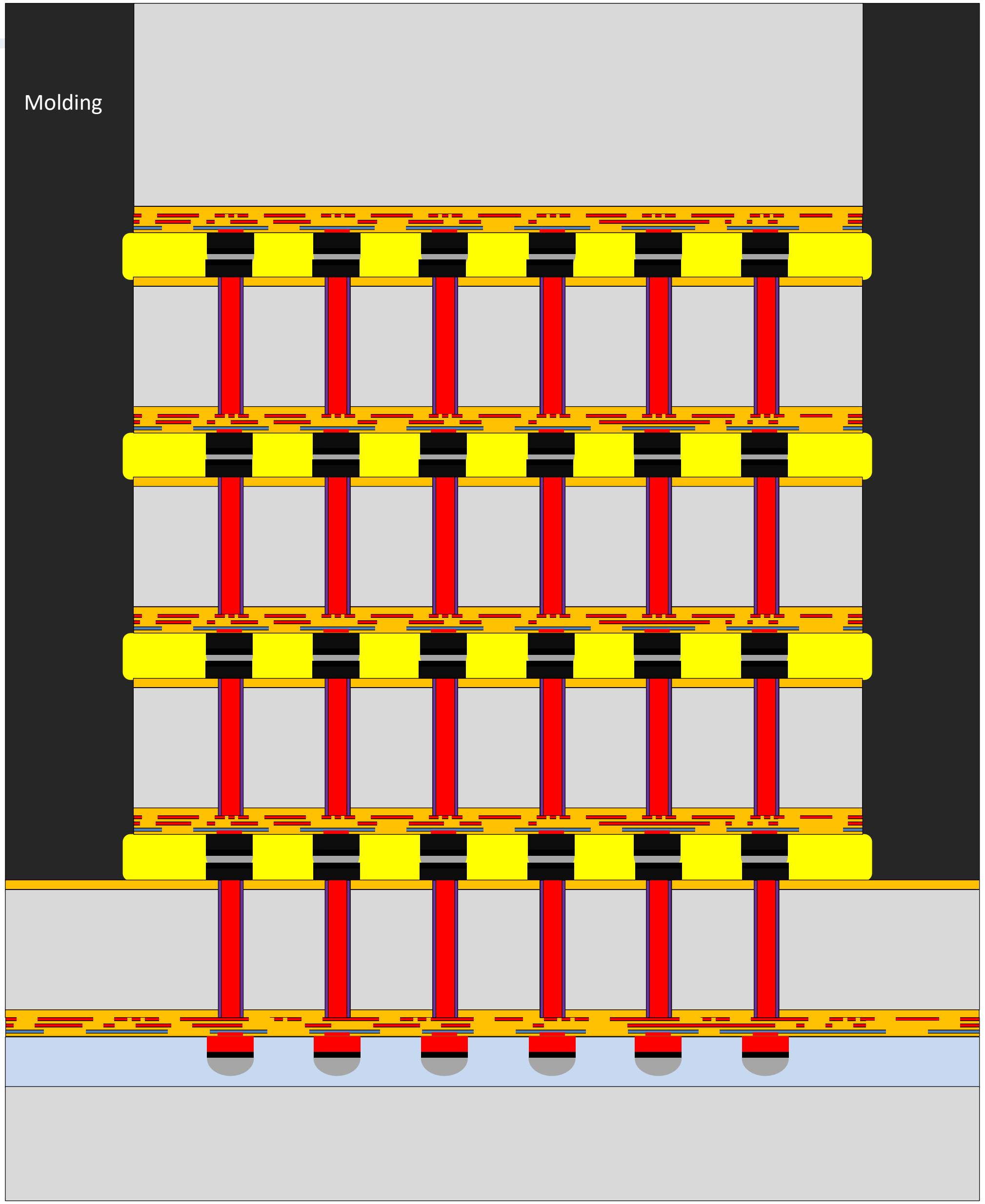


drawing not to scale

HBM Stacking Process Flow (3/4)

HBM Stack

- Wafer Molding

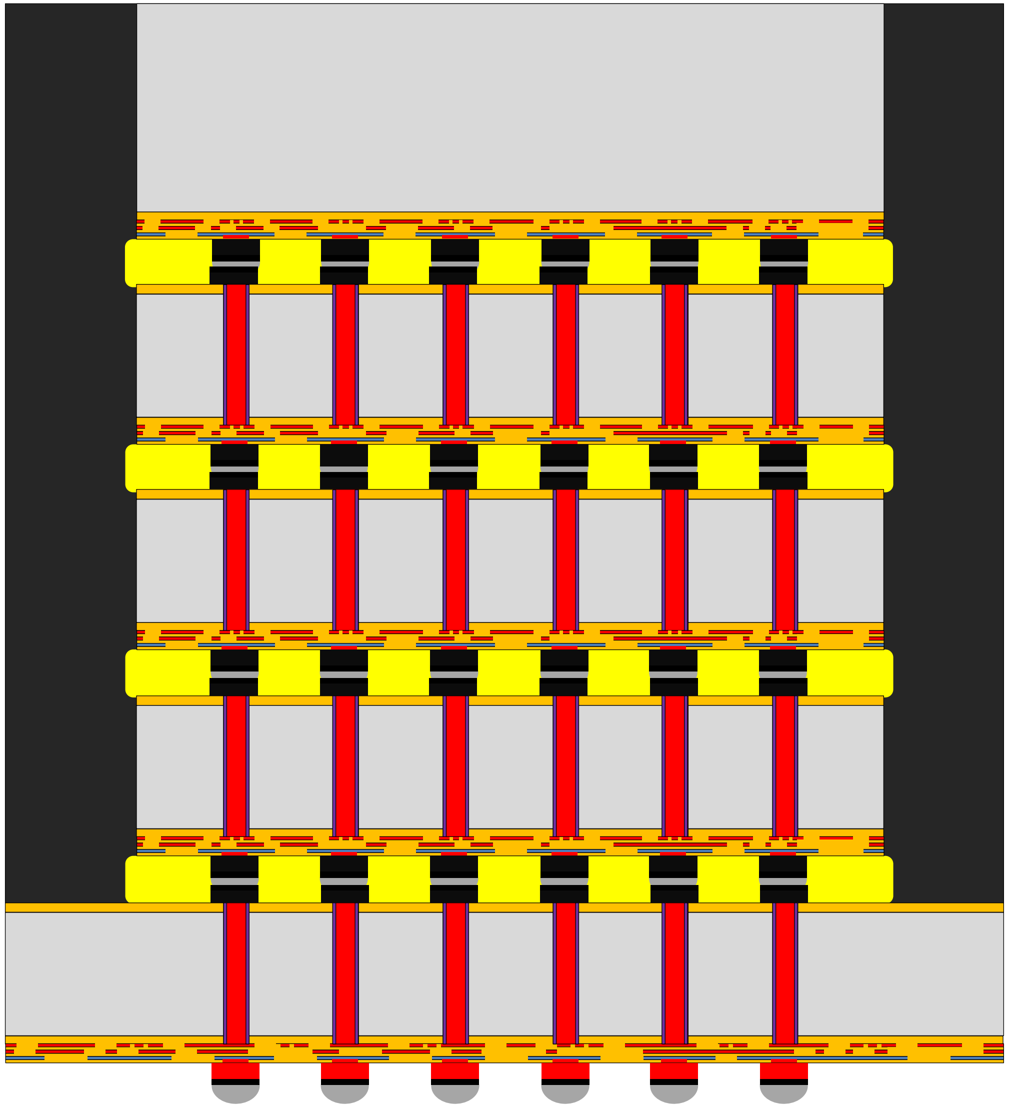


drawing not to scale

HBM Stacking Process Flow (4/4)

HBM Stack

- Debonding & Dicing



drawing not to scale

HBM Wafer Fabrication Unit

- We assume that the manufacturing of the HBM dies are made by Samsung on 300mm wafers.

- Wafer fab unit:
  - Name: Samsung Line 15
  - Wafer diameter: 300mm (12-inch)
  - Capacity: 140,000 wafers / month
  - Year of start: 2006
  - Products: Memory
  - Location: Hwasung, South Korea

- This manufacturing line has been created in 2006.
  - We assume that clean room and equipment are depreciated.
  - We estimate that a residual depreciation of 50% is existing on both equipment and clean room to maintain the efficiency of the line.

Interposer Process Flow

- Interposer Front-End Process:
  - Substrate: 300mm Silicon wafer
  - Process type: TSV via-Middle
  - Metal layers: BEOL 2 ML (1 Cu + 1 Al)
  - Interposer Area: 1,170mm^2

- Test:
  - Test type: Wafer sort (Probe test)

TSMC TSV Process
- Lithography
- Silicon Etch : DRIE (100µm depth, 15µm diam)
- PR Strip : Resist Strip
- Oxide Liner: PECVD (SiO_2, 0.9µm)
- Barrier/Seed : PVD-Tantalum (Ta, 0.05µm)
- Barrier/Seed : PVD-Copper (Cu, 0.1µm)
- CU filling : Electroplating-Copper (Cu)
- TSV Anneal
- Cu & Ta CMP

TSMC Micro-Bumping Process
- UBM : PVD Titanium (Ti, 0.1µm)
- UBM : PVD Copper (Cu, 0.3µm)
- UBM : Lithography
- UBM : Electroplating-Copper (Cu, 9.8µm)
- UBM : Electroplating-Nickel (Ni, 4.7µm)
- UBM : Electroplating-Copper (Cu, 2.8µm)
- UBM : Resist Strip
- UBM : Wet Etching Copper
- UBM : Wet Etching Titanium

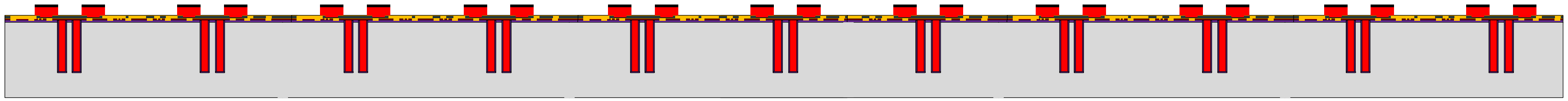
TSMC CoW Process
- GPU Die Bonding
- HBM Stacks Bonding
- Solder Reflow
- CUF Underfill Deposition
- Underfill Curing
- Wafer Molding
- Mold thinning

TSMC Back-side Process
- Temporary Bonding : Edge Trimming
- Temporary Bonding : Spin Coat Elastomer
- Temporary Bonding : bonding to carrier
- Backside Thinning
- TSV Via Reveal : Si etch
- TSV Via Reveal : PECVD Oxide (0.5µm)
- TSV Via Reveal : CMP
- Passivation : Polyimide spin-coating
- Passivation : Lithography
- Passivation : baking
- Copper Pillar : PVD Titanium (Ti, 0.1µm)
- Copper Pillar : Lithography
- Copper Pillar : Wet Etching Titanium
- Copper Pillar : Resist Strip
- Copper Pillar : Electroplating-Copper (Cu, 40µm)
- Bumping : Electroplating Solder (SnAg, 35µm)
- Carrier debonding

Interposer – TSV , BEOL & Microbump Process Flow

Interposer

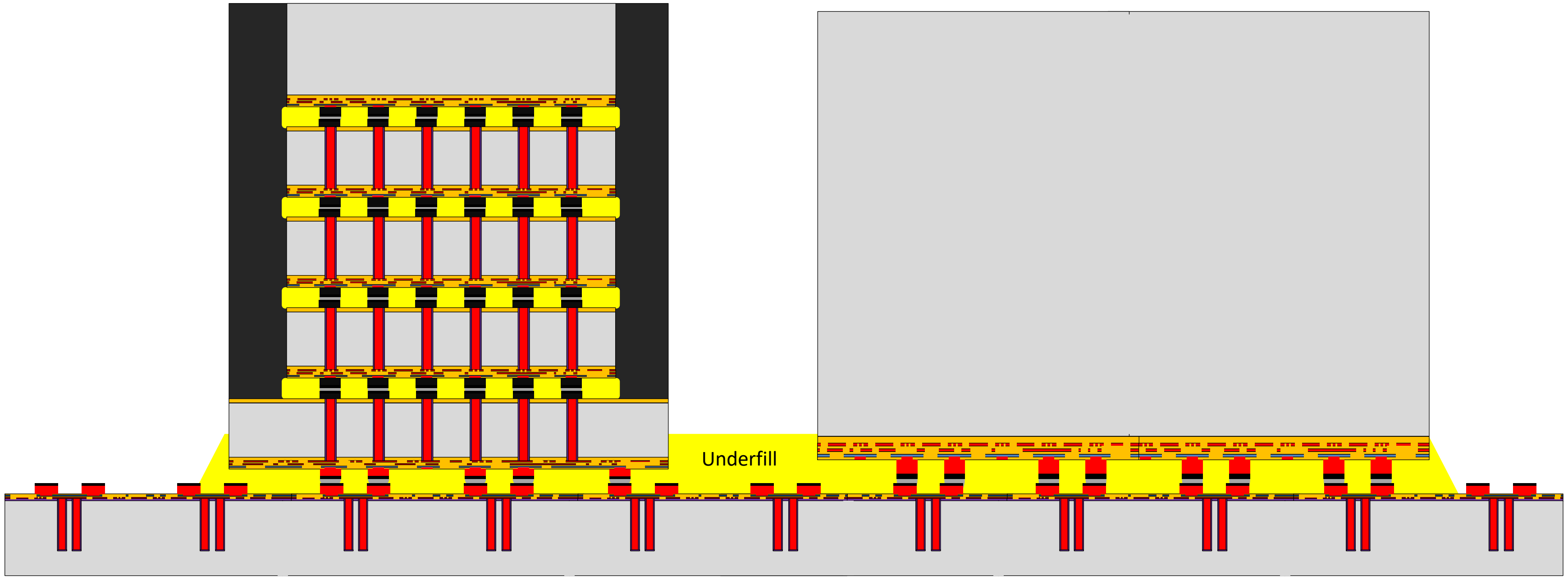
- Si 300mm wafer
- TSV etch: DRIE
- Oxide liner deposition: SiO_2 PECVD
- TSV barrier/Seed layers deposition
- TSV filling: Copper ECD
- CMP
- BEOL (1 Cu + 1 Al)
- Microbump



Interposer – CoW Process Flow (1/7)

Interposer

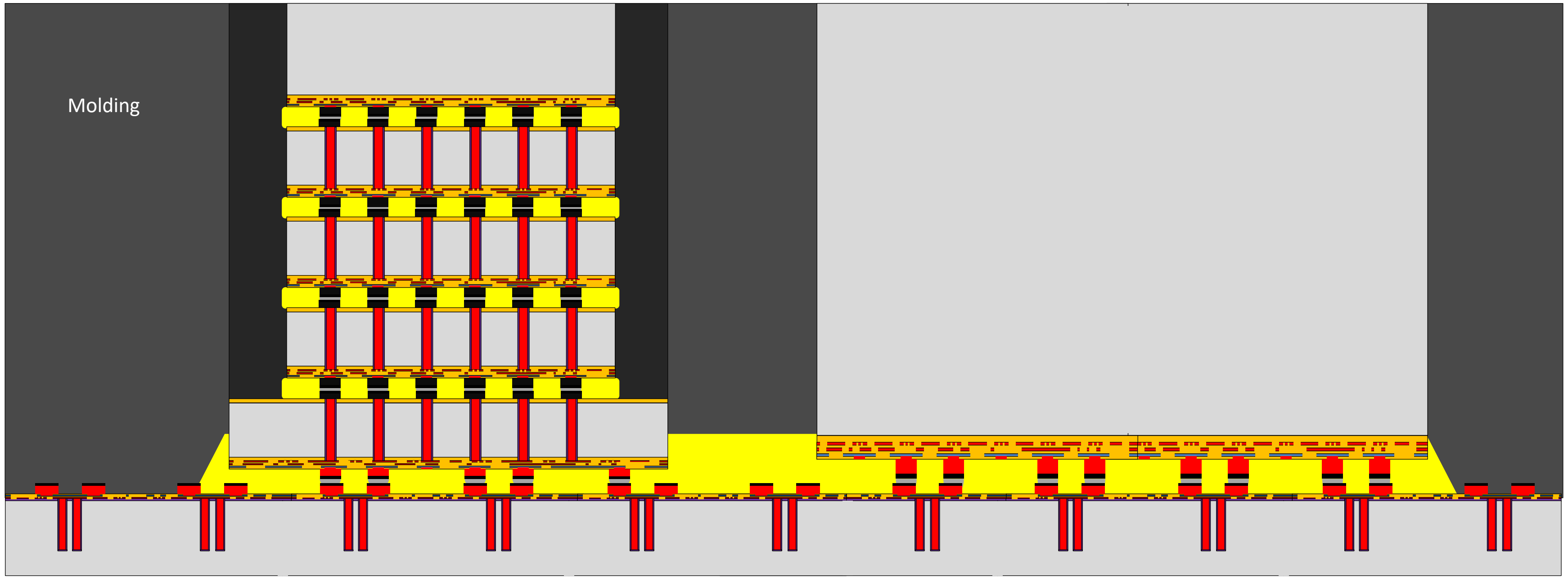
- Pick & Place Dies on interposer
- Reflow
- CUF Underfill Deposition & Curing



Interposer – CoW Process Flow (2/7)

Interposer

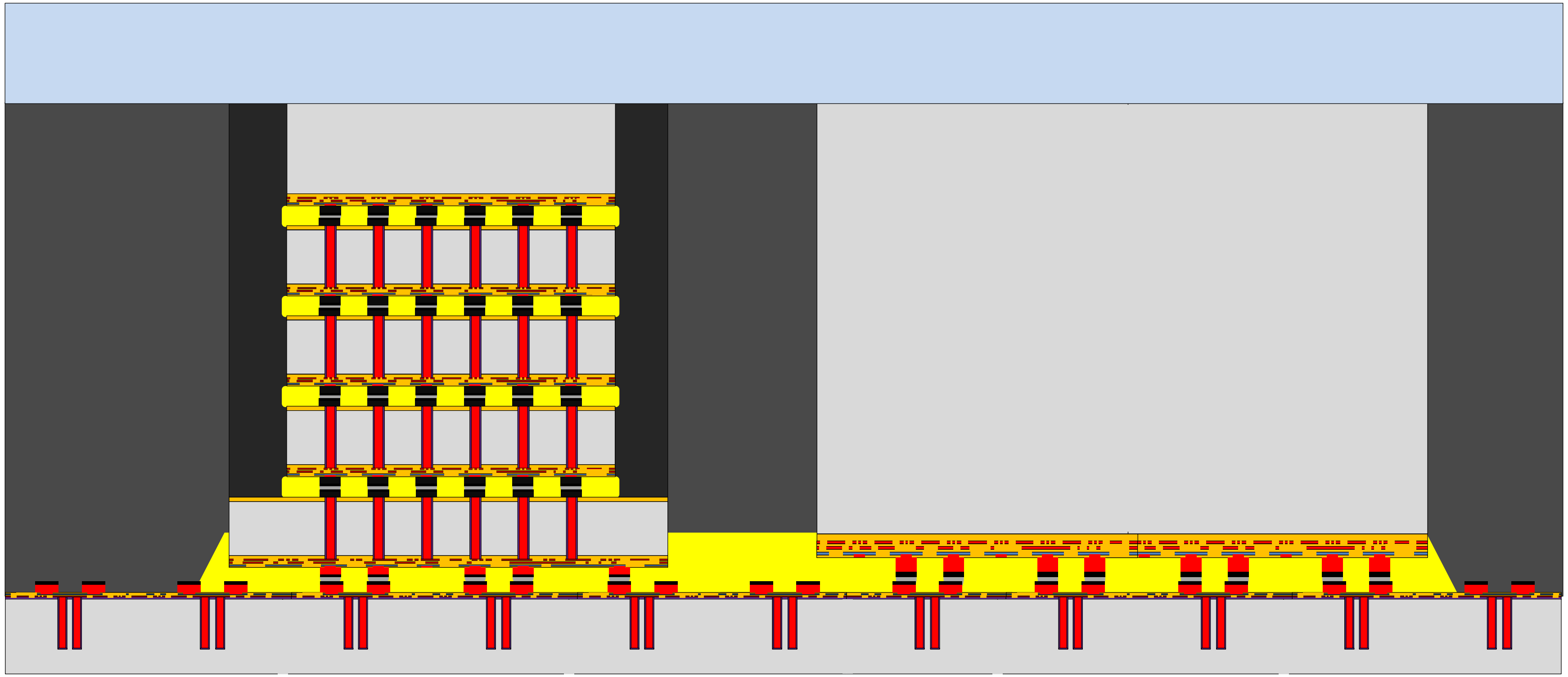
- Interposer Molding
- Mold Thinning



Interposer – CoW Process Flow (3/7)

Interposer

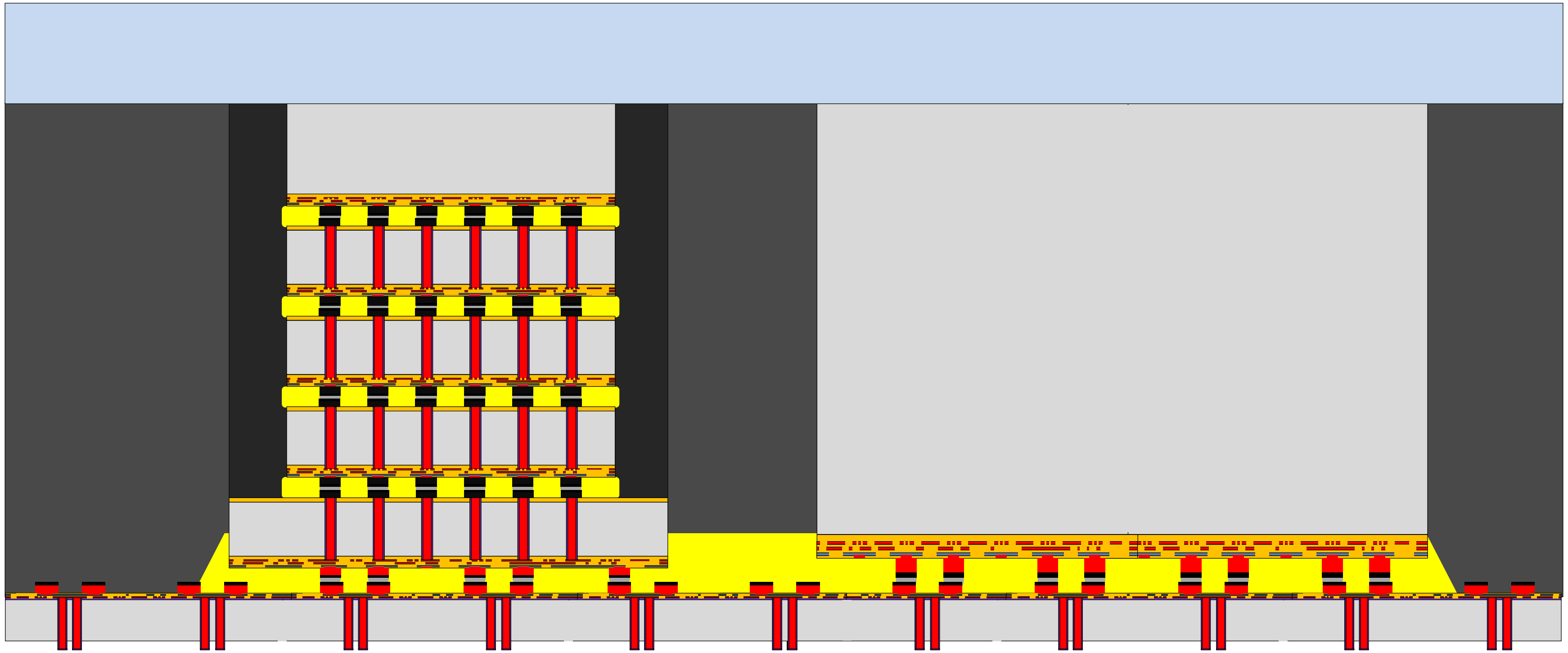
- Carrier bonding



Interposer – CoW Process Flow (4/7)

Interposer

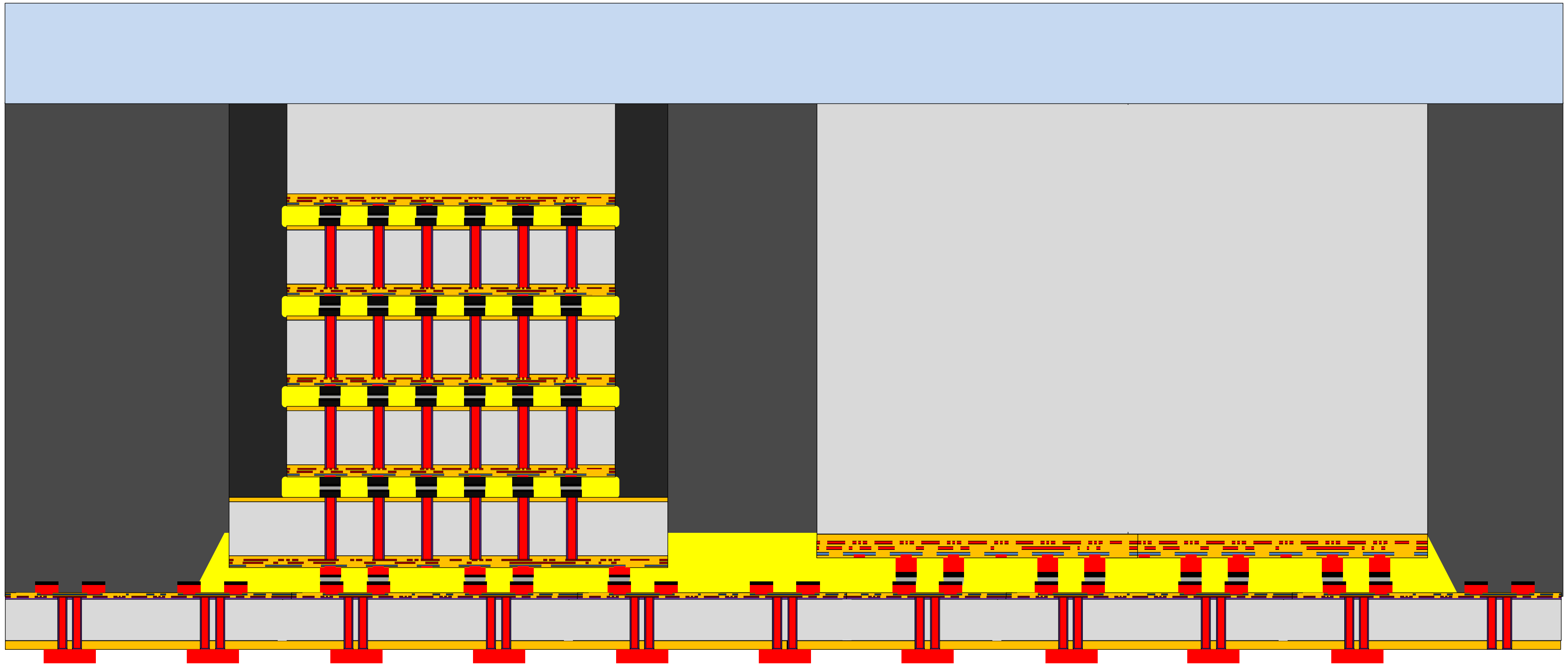
- Backside thinning
- Via reveal



Interposer – CoW Process Flow (5/7)

Interposer

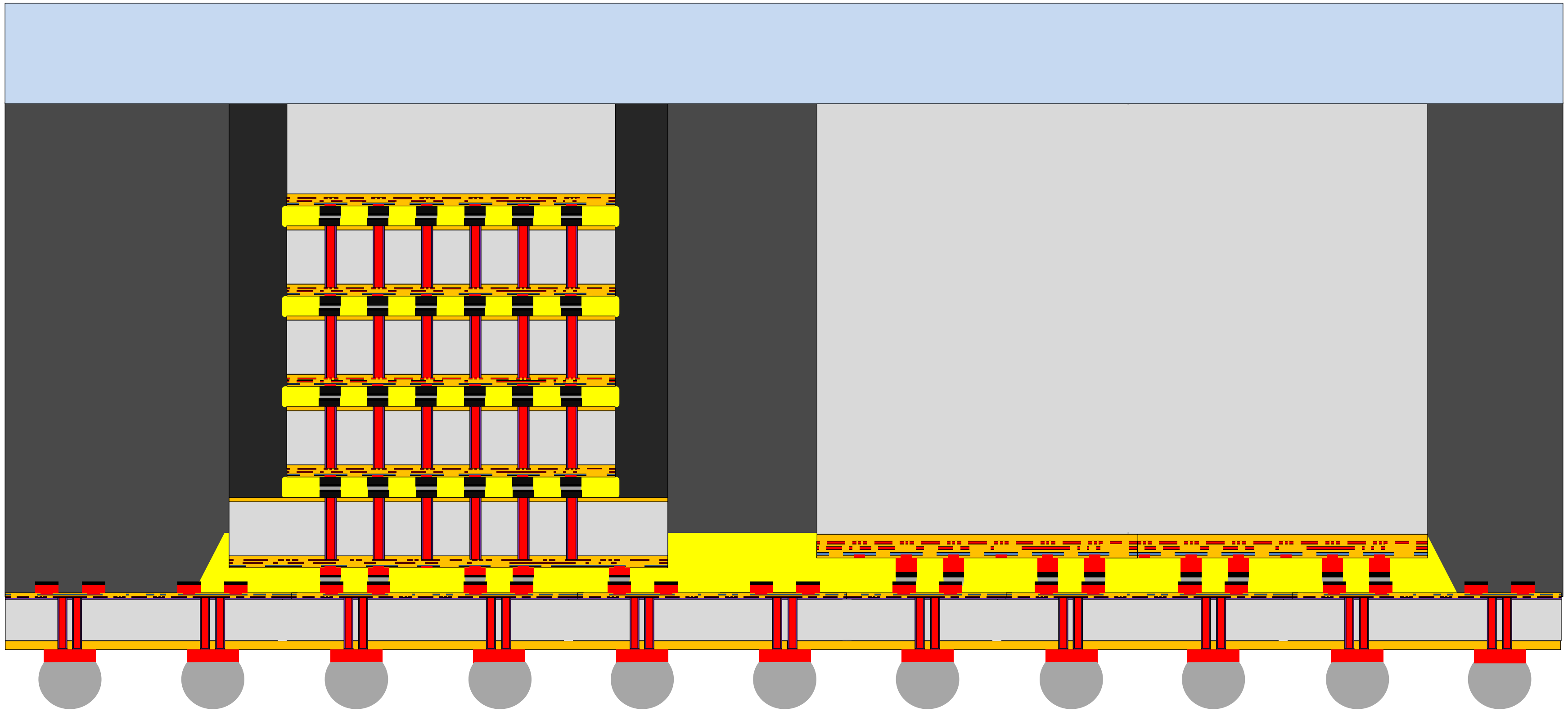
- Via reveal: Oxide Passivation + CMP
- Copper pillar: Polyimide deposition & pattern + Ti deposition & pattern + Copper Deposition



Interposer – CoW Process Flow (6/7)

Interposer

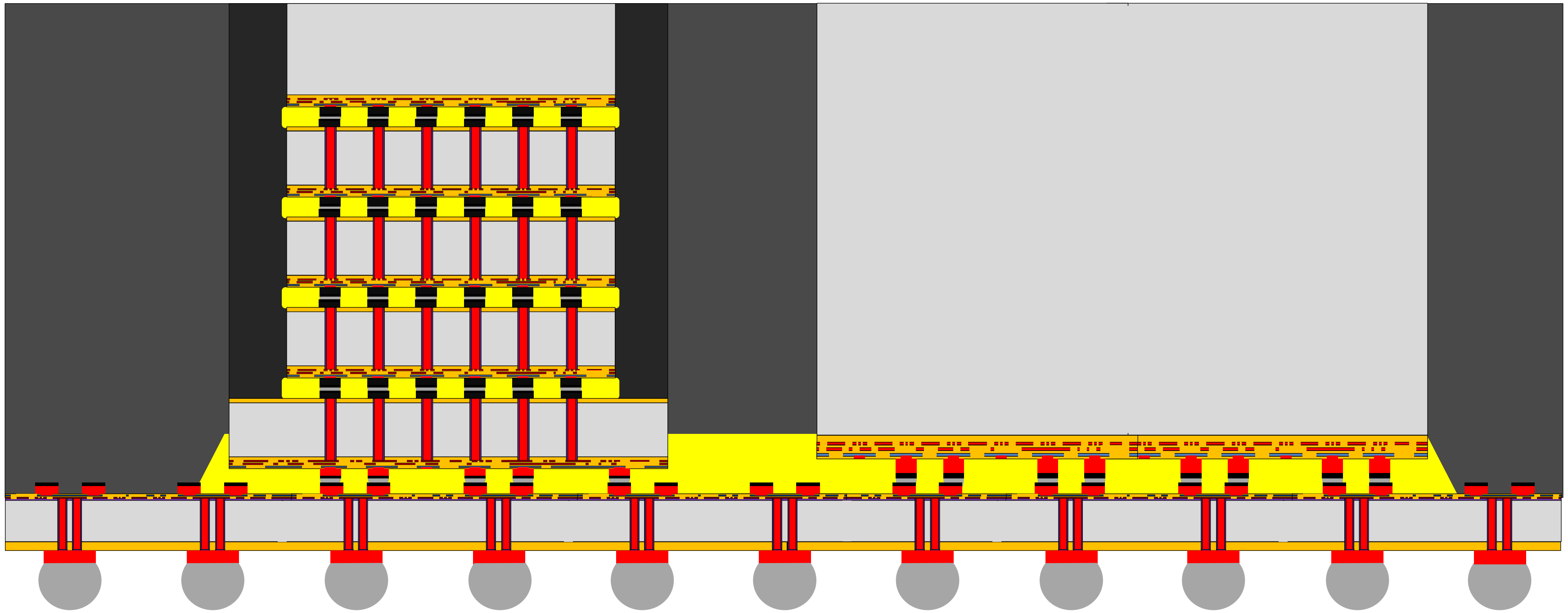
- Bumping



Interposer – CoW Process Flow (7/7)

Interposer

- Carrier debonding
- Dicing



Interposer Wafer Fabrication Unit

- We assume that the manufacturing of the CoW process is made by TSMC on 300mm wafers.

- Wafer fab unit:
  - Name: -
  - Wafer diameter: 300mm (12-inch)
  - Capacity: -
  - Year of start: -
  - Products: Foundry
  - Location: Taiwan

Final Assembly Process Flow

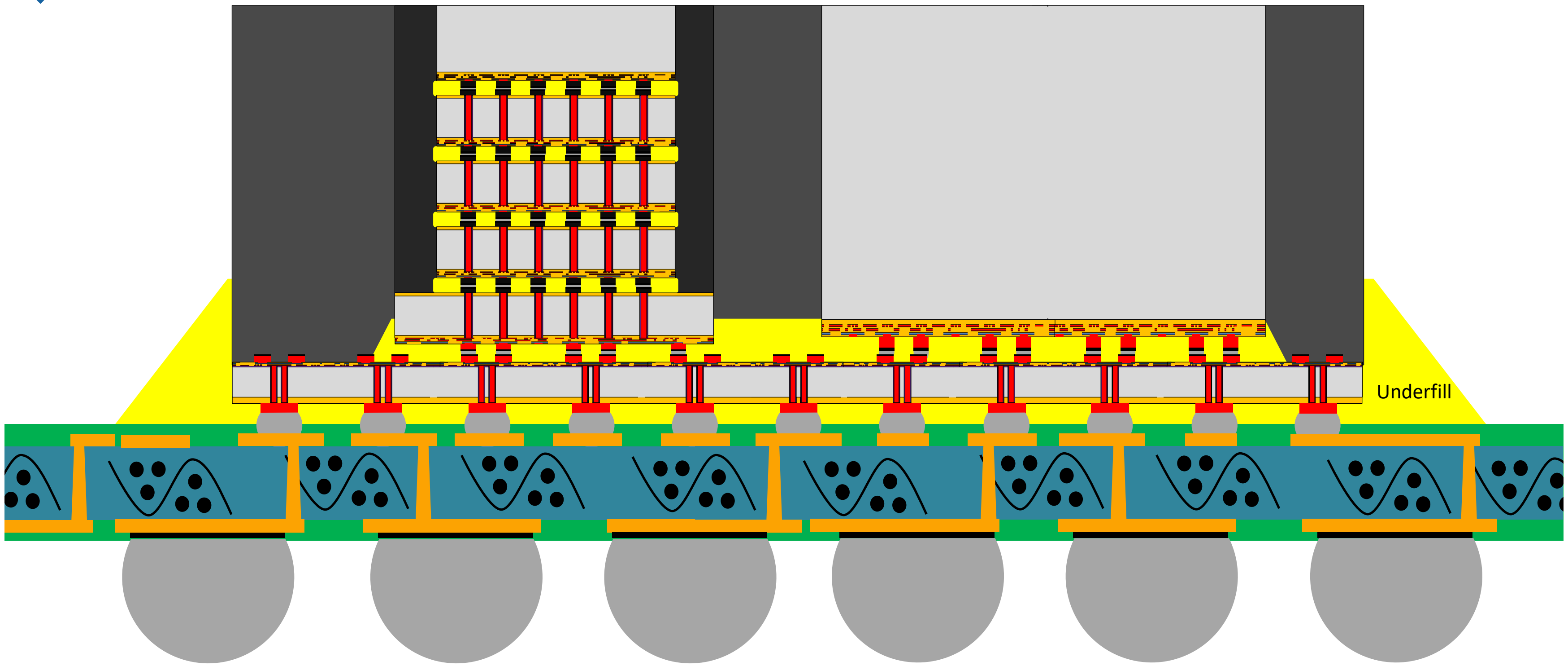
- Packaging Process:
  - Package type: fcBGA-2540 (55x55mm)
  - Substrate: 12-Layer PCB Laminate
  - Process type: 2.5D stacking
  - Special features: Flip-Chip Bonding & Reflow

2.5D Stacking Process (Assembly)
- Interposer Stack Flip-chip Bonding
- Solder Relfow
- CUF Underfill Deposition
- Underfill Curing
- Metal Frame Pick & Place
- LASER Marking
- Dicing

Final Assembly Process Flow (oS)

Interposer

- Passives assembly
- Pick & Place interposer on substrate
- Reflow
- CUF Underfill Deposition & Curing
- Metal Frame deposition
- Laser marking & Dicing



Final Assembly Unit

- We assume that the final assembly is made by TSMC.

- Fab unit:
  - Name: -
  - Capacity: -
  - Year of start: -
  - Products: Assembly
  - Location: Taiwan

COST
ANALYSIS

Summary of the Cost Analysis

The component is designed by NVIDIA and manufactured mainly by TSMC.

Production of the HBM stacks are assumed to be realized by Samsung on 300mm (12-inch) wafers in South Korea.

- The HBM stack wafer cost is estimated at $8,952 (medium yield estimation).
- The HBM stack cost is estimated at $17.08 (medium yield estimation).

Production of the GPU die is assumed to be realized by TSMC on 300mm (12-inch) wafers in Taiwan.

- The GPU wafer cost is estimated at $8,356 (medium yield estimation).
- The GPU die cost is estimated at $220 (medium yield estimation).

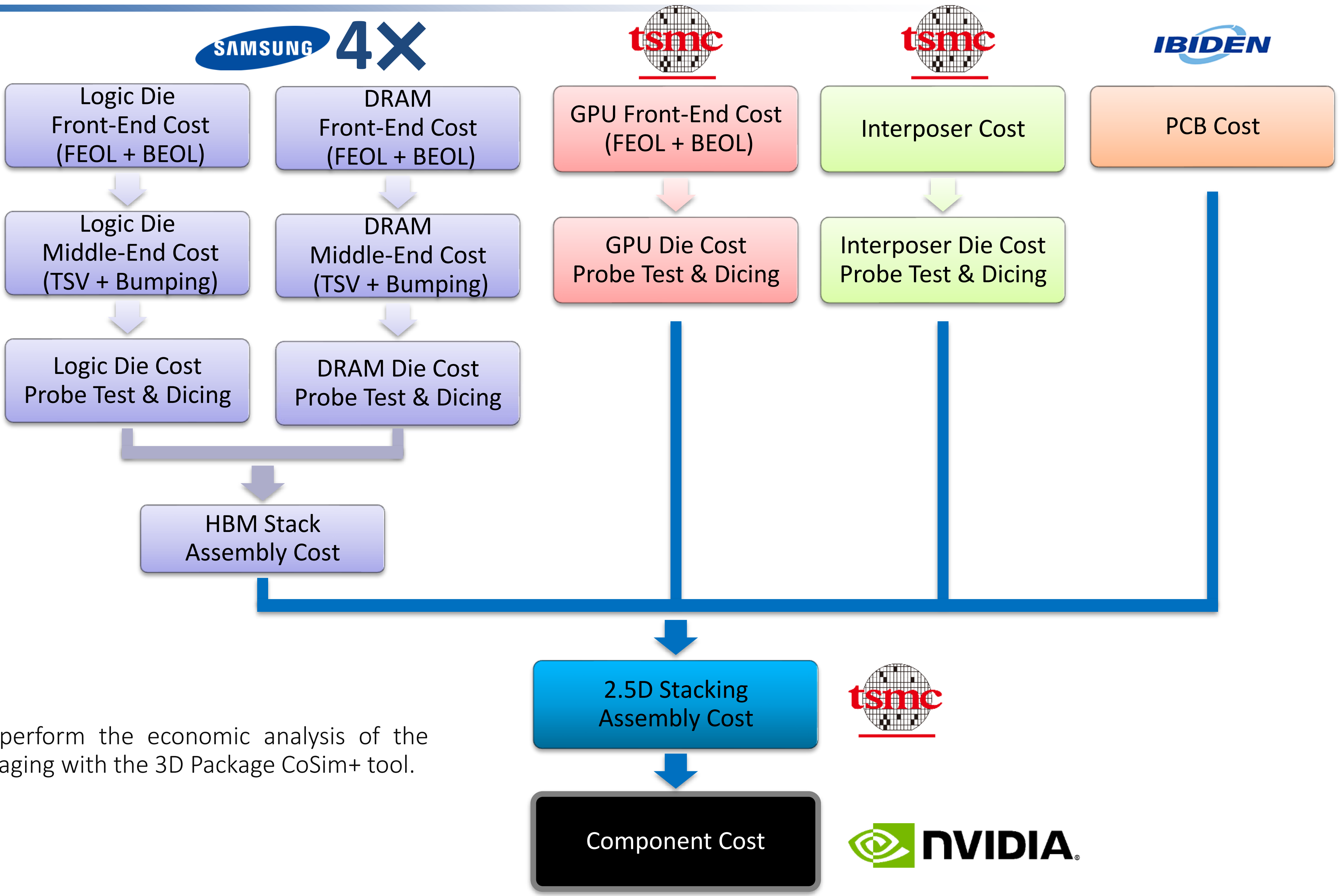
Production of the interposer and Chip-on-Wafer (CoW) assembly are supposed to be realized by TSMC on 300mm (12-inch) wafers in Taiwan.

- The interposer and CoW assembly cost is estimated at $1,593 (medium yield estimation).
- The total CoW stack (including GPU & HBM) wafer cost is estimated at $12,828 (medium yield estimation).
- The CoW stack (including GPU & HBM) cost is estimated at $475 (medium yield estimation).

The Volta component (GPU + HBM + Substrate + Assembly) cost from NVIDIA's perspective ranges from $456 to $626 according to yield hypotheses.

The final price paid by OEMs is estimated to ranges from $1,138 to $1,561 according to yield hypotheses.

Main Steps of Economic Analysis



We perform the economic analysis of the Packaging with the 3D Package CoSim+ tool.

Yield Explanation

The wafers and dies are tested during the process flow. There are 2 types of test :
- The tests on the physical characteristics of the wafer like the thickness of a deposited layer.
- The tests on the electrical functionalities of the die.

The difference is important because with the physical test, a poor result means a problem on a step and all the dies on the wafer are defective, so the wafer is scrapped. Usually these yields are good for mature technologies.

The tests on the dies are different. Each die is tested, one by one or simultaneously using “parallel” tests, and only the defective dies are scraped. During the probe test which is realized on the wafer, the defective dies are marked and are not assembled in package.

In this reverse costing study, 4 yields are used :

Process	Yield	Apply on	Description
Front-End	Middle-End Yield	Wafer	The defective wafers are scraped
Back-End 0	Probe yield	Die	The defective dies are scraped. The number of good dies is function of the probe yield. Only the good dies are assembled in the package.
Back-End 1	Packaging yield	Die + Package	The defective components are scraped
Back-End 1	Final test yield	Die + Package	The defective components are scraped

Yield Hypotheses

	GPU Die		
	Low Yield	Medium Yield	High Yield
Probe Test yield	50.0%	60.0%	70.0%

	HBM Stack		
	Low Yield	Medium Yield	High Yield
Middle-End Yield	93.0%	94.0%	95.0%
Probe Test Yield	70.0%	80.0%	90.0%

	Interposer		
	Low Yield	Medium Yield	High Yield
Middle-End Yield	93.0%	94.0%	95.0%
CoW Yield	90.0%	91.0%	92.0%
Probe Test Yield	70.0%	75.0%	80.0%

	Component		
	Low Yield	Medium Yield	High Yield
Packaging & Final Test Yield	97.0%	98.0%	99.0%

In our simulation, we assume a development and a production ramp up without important technical problem.

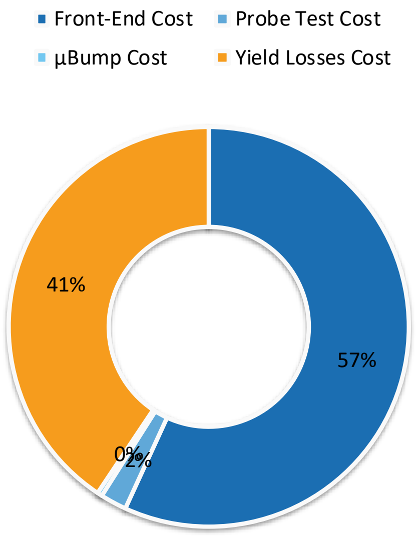
GPU Wafer & Die Cost

	Low Yield		Medium Yield		High Yield	
	Cost	Breakdown	Cost	Breakdown	Cost	Breakdown
Front-End Price	$8,000.00	95.7%	$8,000.00	95.7%	$8,000.00	95.7%
Probe Test Cost	$300.00	3.6%	$300.00	3.6%	$300.00	3.6%
µBump Cost	$55.81	0.7%	$55.81	0.7%	$55.81	0.7%
Total Wafer Cost (including foundry margin)	**$8,355.81**	**100%**	**$8,355.81**	**100%**	**$8,355.81**	**100%**
Nb of potential good dies per wafer	64		64		64	
Nb of good dies per wafer	32		38		44	
Front-End Cost	$125.00	47.9%	$125.00	56.8%	$125.00	65.8%
Probe Test Cost	$4.69	1.8%	$4.69	2.1%	$4.69	2.5%
µBump Cost	$0.87	0.3%	$0.87	0.4%	$0.87	0.5%
Yield Losses Cost	$130.56	50.0%	$89.33	40.6%	$59.35	31.3%
Die Cost (including foundry margin)	**$261.12**	**100%**	**$219.89**	**100%**	**$189.90**	**100%**

The wafer cost for the GPU is estimated to $8,356, including foundry overheads.

The number of good dies per wafer is estimated to range from 32 to 44 according to yield variations, which results in a GPU die cost ranging from $190 to $261.

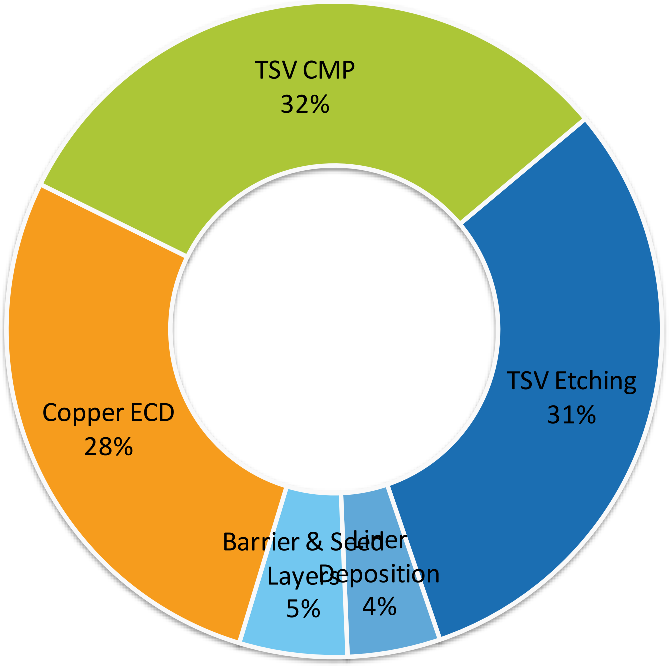
Die Cost Breakdown (Medium Yield)



HBM TSV Manufacturing Cost

	Cost	Breakdown
TSV Etching	$24.73	30.9%
Liner Deposition	$3.67	4.6%
Barrier & Seed Layers	$4.27	5.3%
Copper ECD	$22.11	27.6%
TSV CMP	$25.30	31.6%
TSV Manufacturing Cost	**$80.09**	**100%**

TSV Cost Breakdown



The TSV manufacturing cost for the DRAM dies and the logic die is estimated to $80 per wafer.

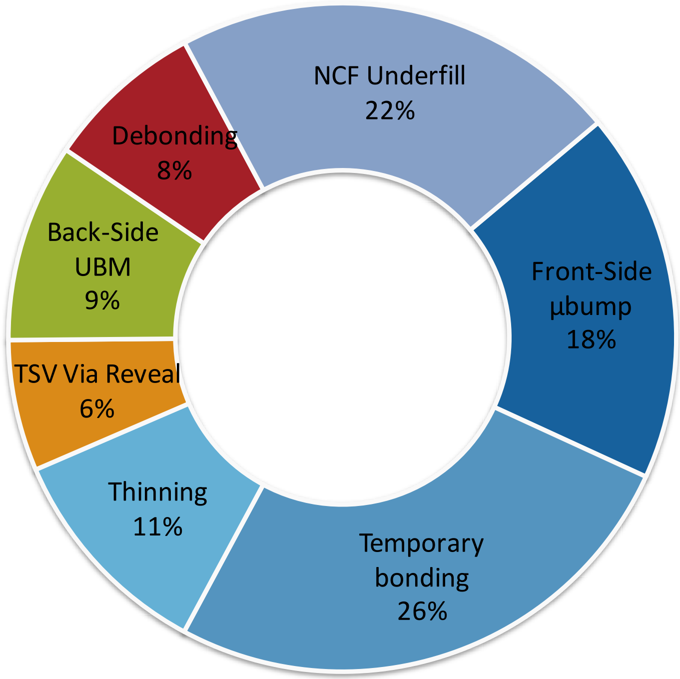
The TSV etching (DRIE) represents 31% of the manufacturing cost.

DRAM Microbumping Cost

DRAM Micro-Bumping Cost	Cost	Breakdown
Front-Side µbump	$33.77	18.0%
Temporary bonding	$48.94	26.0%
Thinning	$20.07	10.7%
TSV Via Reveal	$11.94	6.3%
Back-Side UBM	$18.09	9.6%
Debonding	$14.46	7.7%
NCF Underfill	$40.86	21.7%
DRAM Micro-Bumping Cost	**$188.14**	**100%**

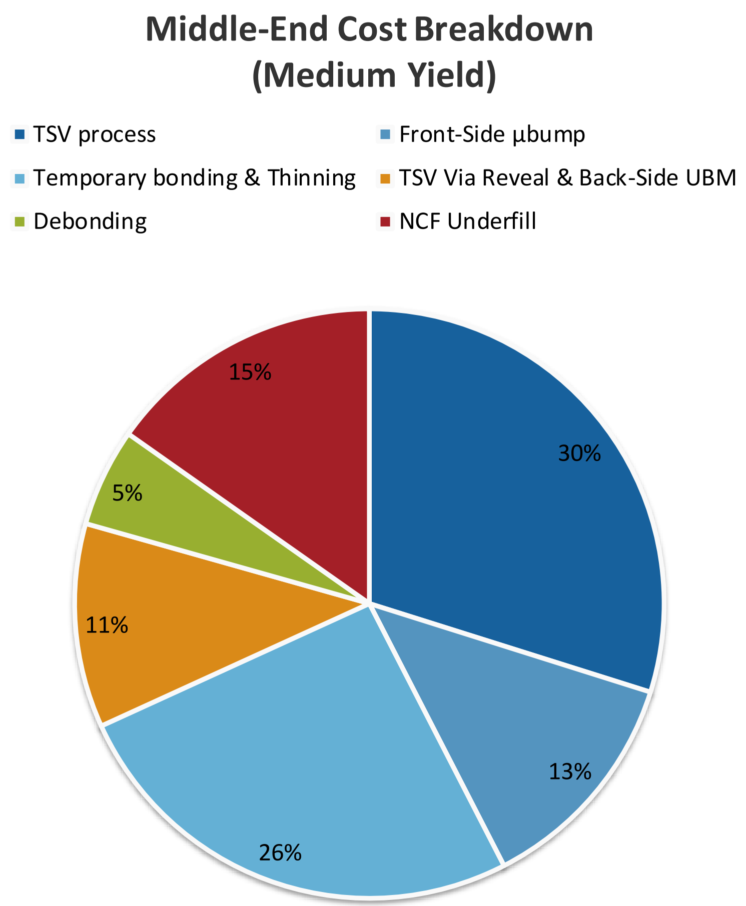
The DRAM micro-bumping manufacturing cost is estimated to $188 per wafer.

DRAM Micro-Bumping Cost Breakdown



DRAM Middle-End Cost (TSV + µBump)

	Cost	Breakdown
TSV process	$80.09	29.9%
Front-Side µbump	$33.77	12.6%
Temporary bonding & Thinning	$69.01	25.7%
TSV Via Reveal & Back-Side UBM	$30.03	11.2%
Debonding	$14.46	5.4%
NCF Underfill	$40.86	15.2%
DRAM Middle-End Manufacturing Cost	**$268.22**	**100%**



The middle-end cost with TSV and micro-bumps is estimated to $268.

The largest portion of the manufacturing cost is due to the TSV process at 30%.

Total DRAM Middle-End Cost (TSV + µBump)

	Low Yield		Medium Yield		High Yield	
	Cost	Breakdown	Cost	Breakdown	Cost	Breakdown
DRAM Front-End cost (FEOL+BEOL)	$1,500.00		$1,500.00		$1,500.00	

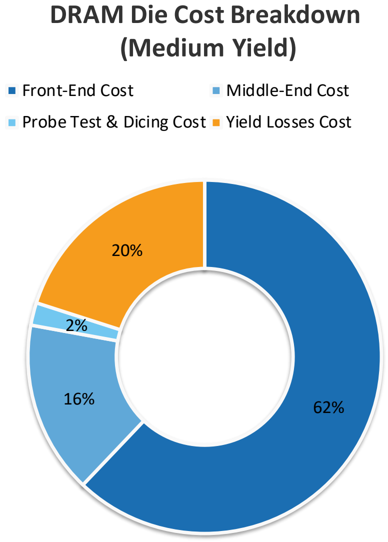
	Low Yield		Medium Yield		High Yield	
	Cost	Breakdown	Cost	Breakdown	Cost	Breakdown
TSV Manufacturing Cost	$80.09	20.0%	$80.09	21.0%	$80.09	22.2%
Micro-Bumping Cost	$188.14	46.9%	$188.14	49.4%	$188.14	52.1%
Yield Losses Cost	$133.09	33.2%	$112.87	29.6%	$93.06	25.8%
Total DRAM Middle-End Cost	**$401.31**	**100%**	**$381.09**	**100%**	**$361.29**	**100%**

By taking into account the yield losses, the total middle-end cost ranges from $361 to $401 according to yield variations.

DRAM Wafer & Die Cost

	Low Yield		Medium Yield		High Yield	
	Cost	Breakdown	Cost	Breakdown	Cost	Breakdown
Front-End Cost	$1,500.00	76.9%	$1,500.00	77.7%	$1,500.00	78.5%
Middle-End Cost	$401.31	20.6%	$381.09	19.7%	$361.29	18.9%
Probe Test & Dicing Cost	$50.00	2.6%	$50.00	2.6%	$50.00	2.6%
Total Wafer Cost	**$1,951.31**	**100%**	**$1,931.09**	**100%**	**$1,911.29**	**100%**
Nb of potential good dies per wafer	728		728		728	
Nb of good dies per wafer	509		582		655	
Front-End Cost	$2.06	53.7%	$2.06	62.1%	$2.06	70.6%
Middle-End Cost	$0.55	14.4%	$0.52	15.8%	$0.50	17.0%
Probe Test & Dicing Cost	$0.07	1.8%	$0.07	2.1%	$0.07	2.4%
Yield Losses Cost	$1.15	30.1%	$0.67	20.1%	$0.29	10.0%
DRAM Die Cost	**$3.83**	**100%**	**$3.32**	**100%**	**$2.92**	**100%**

The number of good dies per wafer is estimated to ranges from 509 to 655 according to yield variations, which results in a die cost ranging from $2.92 to $3.83.



Total Logic Middle-End Cost (TSV + µBump)

	Low Yield		Medium Yield		High Yield	
	Cost	Breakdown	Cost	Breakdown	Cost	Breakdown
Logic Front-End cost (FEOL+BEOL)	$1,500.00		$1,500.00		$1,500.00	

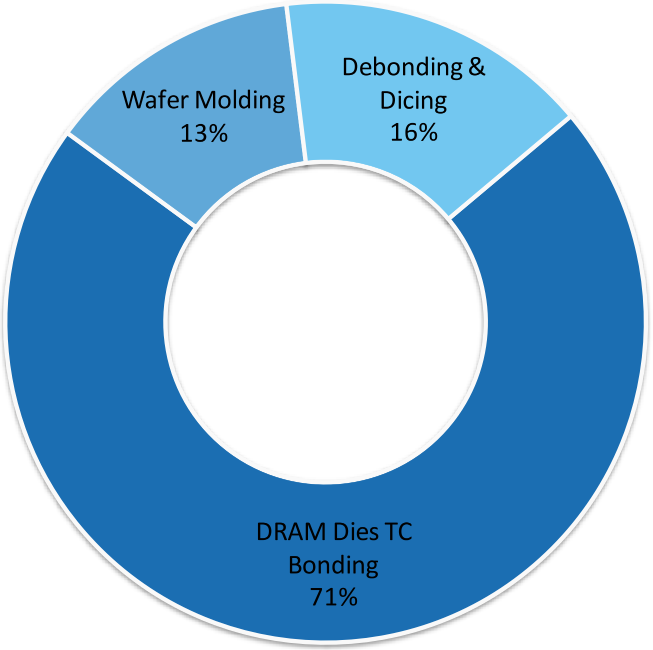
	Low Yield		Medium Yield		High Yield	
	Cost	Breakdown	Cost	Breakdown	Cost	Breakdown
TSV Manufacturing Cost	$80.09	21.7%	$80.09	22.9%	$80.09	24.3%
Micro-Bumping Cost	$158.41	42.9%	$158.41	45.3%	$158.41	48.0%
Yield Losses Cost	$130.85	35.4%	$110.97	31.8%	$91.50	27.7%
Total Logic Middle-End Cost	**$369.35**	**100%**	**$349.46**	**100%**	**$330.00**	**100%**

By taking into account the yield losses, the total middle-end cost ranges from $330 to $369 according to yield variations.

HBM Stacking Cost (TSV + µBump)

	Cost	Breakdown
DRAM Dies TC Bonding	$105.15	71.2%
Wafer Molding	$19.19	13.0%
Debonding & Dicing	$23.42	15.8%
HBM Stacking Cost	**$147.75**	**100%**

HBM Stacking Cost Breakdown



The HBM stacking cost with thermocompression bonding and wafer molding is estimated to $148.

The largest portion of the manufacturing cost is due to the TCB process at 71%.

Total HBM Wafer Cost

	Low Yield		Medium Yield		High Yield	
	Cost	Breakdown	Cost	Breakdown	Cost	Breakdown
Logic Front-End Cost	$1,500.00	80.2%	$1,500.00	81.1%	$1,500.00	82.0%
Logic Middle-End Cost	$369.35	19.8%	$349.46	18.9%	$330.00	18.0%
Total Logic Wafer Cost	**$1,869.35**	**100%**	**$1,849.46**	**100%**	**$1,830.00**	**100%**
DRAM Dies Cost	$7,038.53		$6,954.57		$6,886.46	
HBM Stacking Cost	$147.75		$147.75		$147.75	
Total HBM Wafer Cost	**$9,055.63**		**$8,951.78**		**$8,864.21**	

By taking into account the logic & DRAM cost and the stacking, the total HBM wafer cost ranges from $8,864 to $9,056 according to yield variations.

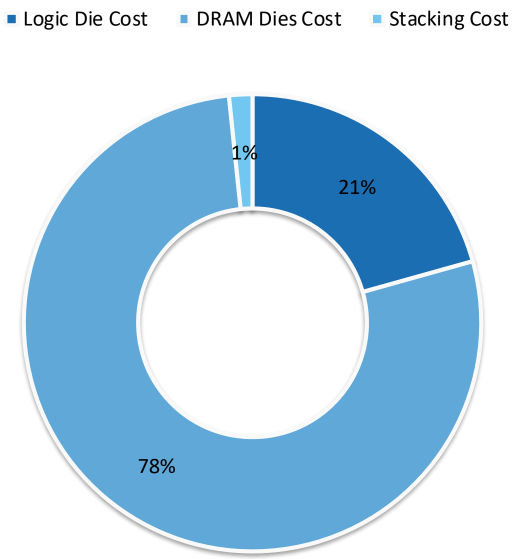
HBM Stack Cost

	Low Yield		Medium Yield		High Yield	
	Cost	Breakdown	Cost	Breakdown	Cost	Breakdown
Total HBM Wafer Cost	**$9,055.63**		**$8,951.78**		**$8,864.21**	
Nb of potential good stack per wafer	656		656		656	
Nb of good stack per wafer	459		524		590	
Logic Die Cost	$4.07	20.6%	$3.53	20.7%	$3.10	20.6%
DRAM Dies Cost	$15.33	77.7%	$13.27	77.7%	$11.67	77.7%
Stacking Cost	$0.32	1.6%	$0.28	1.7%	$0.25	1.7%
HBM Stack Cost	**$19.73**	**100%**	**$17.08**	**100%**	**$15.02**	**100%**
Samsung Gross Profit	$19.73	+50.0%	$17.08	+50.0%	$15.02	+50.0%
HBM Stack Price	**$39.46**		**$34.17**		**$30.05**	

The number of good HBM stack per wafer is estimated to ranges from 459 to 590 according to yield variations, which results in a HBM stack cost ranging from $15.0 to $19.7.

We estimate a gross margin of 50% for Samsung, which results in a HBM stack price ranging from $30 to $39. This corresponds to the selling price to NVIDIA.

HBM Stack Cost Breakdown (Medium Yield)



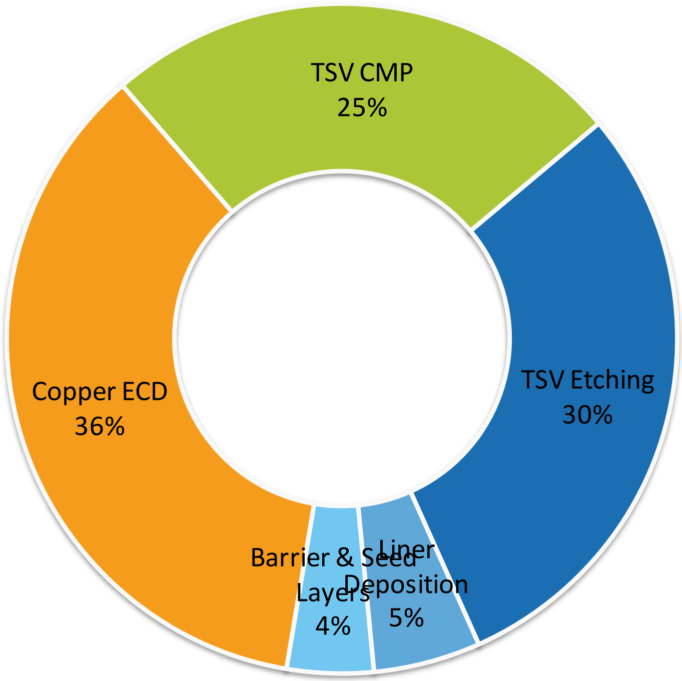
Interposer TSV Manufacturing Cost

	Cost	Breakdown
TSV Etching	$29.03	29.4%
Liner Deposition	$5.10	5.2%
Barrier & Seed Layers	$4.12	4.2%
Copper ECD	$35.50	36.0%
TSV CMP	$24.92	25.3%
TSV Manufacturing Cost	**$98.66**	**100%**

The TSV manufacturing cost for the interposer is estimated to $99 per wafer.

The copper deposition steps represents 36% of the manufacturing cost.

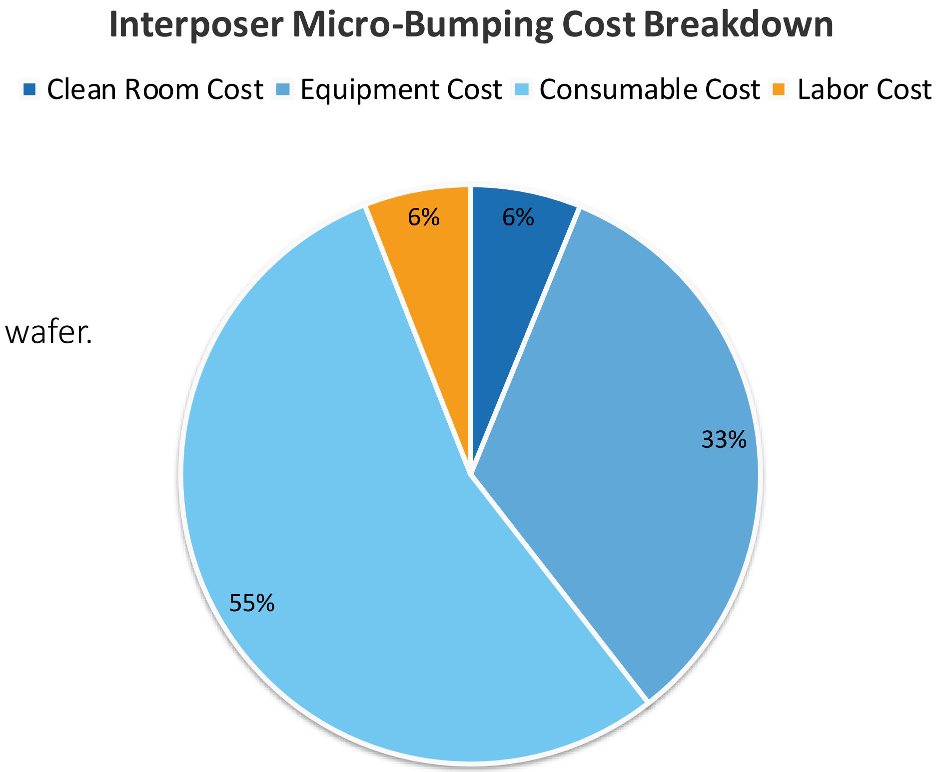
Interposer TSV Cost Breakdown



Interposer Micro-Bumping Cost

Interposer Micro-Bumping Cost	Cost	Breakdown
Clean Room Cost	$2.94	6.2%
Equipment Cost	$15.91	33.3%
Consumable Cost	$26.06	54.6%
Labor Cost	$2.84	6.0%
Interposer Micro-Bumping Cost	**$47.76**	**100%**

The interposer micro-bumping manufacturing cost is estimated to $48 per wafer.



Total Interposer Wafer Cost (TSV + µBump)

Interposer	Low Yield		Medium Yield		High Yield	
	Cost	Breakdown	Cost	Breakdown	Cost	Breakdown
Raw Wafer Cost (Si 300mm)	$75.00	20.1%	$75.00	20.4%	$75.00	20.6%
TSV Manufacturing Cost	$98.66	26.5%	$98.66	26.8%	$98.66	27.1%
Interposer BEOL cost (2 ML)	$125.00	33.6%	$125.00	33.9%	$125.00	34.3%
Micro-Bumping Cost	$47.76	12.8%	$47.76	13.0%	$47.76	13.1%
Yield losses Cost	$26.07	7.0%	$22.11	6.0%	$18.23	5.0%
Total Interposer Wafer Cost	**$372.49**	**100%**	**$368.53**	**100%**	**$364.65**	**100%**

By taking into account the yield losses, the total Interposer Wafer cost ranges from $365 to $372 according to yield variations.
This cost do not take into account the interposer backside process which is realized after the bonding of the dies on the interposer.

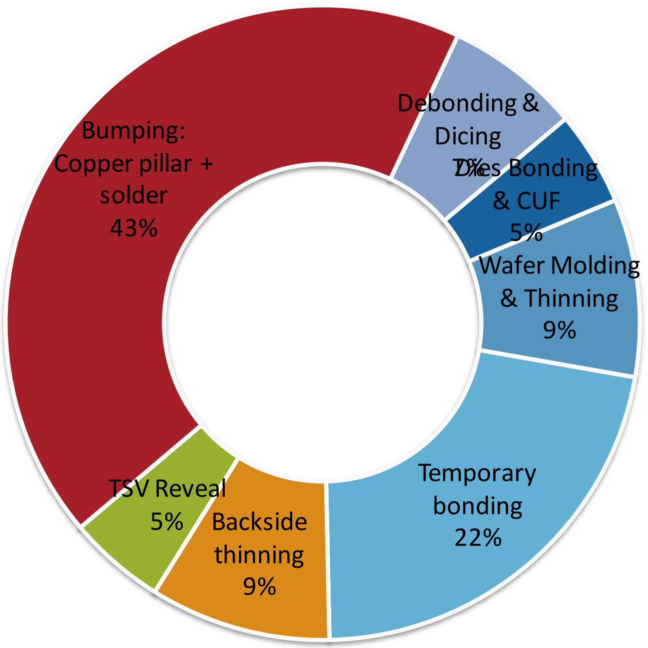
Chip-on-Wafer (CoW) Assembly Cost

	Cost	Breakdown
Dies Bonding & CUF	$10.33	4.8%
Wafer Molding & Thinning	$19.46	9.1%
Temporary bonding	$47.06	21.9%
Backside thinning	$19.81	9.2%
TSV Reveal	$10.65	5.0%
Bumping: Copper pillar + solder	$92.75	43.2%
Debonding & Dicing	$14.78	6.9%
CoW Assembly Cost	**$214.84**	**100%**

The Chip-on-Wafer (CoW) manufacturing cost is estimated to $215 per wafer.

The bumping with copper pillars represents 43% of the manufacturing cost.

CoW Stacking Cost Breakdown



Chip-on-Wafer (CoW) Stack Wafer Cost

	Low Yield		Medium Yield		High Yield	
	Cost	Breakdown	Cost	Breakdown	Cost	Breakdown
Interposer Manufacturing Cost	$372.49	20.5%	$368.53	23.1%	$364.65	26.3%
CoW Cost	$214.84	11.8%	$214.84	13.5%	$214.84	15.5%
Yield Losses Cost	$1,229.01	67.7%	$1,009.82	63.4%	$805.41	58.2%
Total Interposer + CoW Wafer Cost	**$1,816.34**	**100%**	**$1,593.19**	**100%**	**$1,384.90**	**100%**
Foundry Gross Profit	$1,816.34	+50.0%	$1,593.19	+50.0%	$1,384.90	+50.0%
Total Interposer + CoW Wafer Price	**$3,632.69**		**$3,186.38**		**$2,769.80**	
HBM Stacks Cost	$3,945.81		$3,690.05		$3,365.40	
GPU Dies Cost	$6,527.98		$5,937.03		$5,317.34	
Filler Dies Cost	$14.22		$14.45		$14.15	
Total CoW Stack Wafer Cost	**$14,120.69**		**$12,827.91**		**$11,466.68**	

The manufacturing cost of the interposer including the Chip-on-Wafer assembly steps is estimated to range from $1,385 to $1,816 according to yield variations.

By taking into account a gross margin for TSMC (estimated to 50%), the interposer + CoW wafer price is estimated to range from $2,770 to $3,633 according to yield variations.

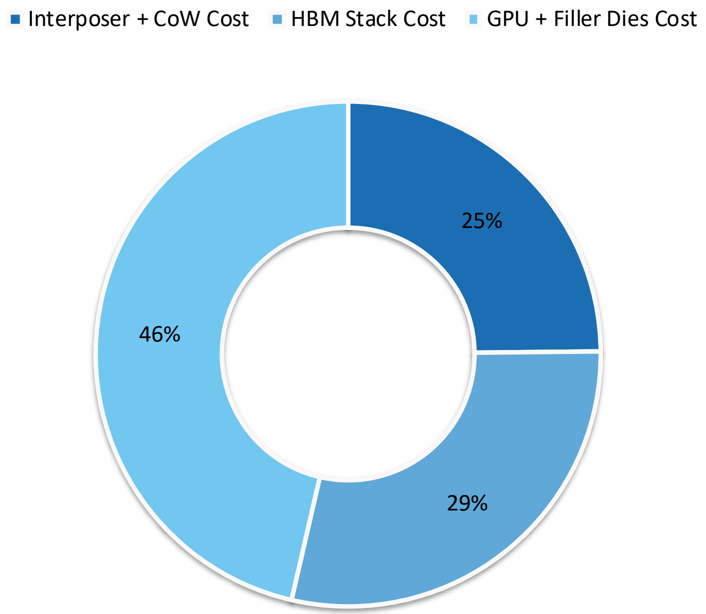
By adding the cost of the HBM Stacks (4) and the GPU die with fillers, the total Chip-on-Wafer stack wafer cost ranges from $11,467 to $14,121 according to yield variations.

Chip-on-Wafer (CoW) Stack Cost

	Low Yield		Medium Yield		High Yield	
	Cost	Breakdown	Cost	Breakdown	Cost	Breakdown
Total CoW Stack Wafer Cost	**$14,120.69**		**$12,827.91**		**$11,466.68**	
Nb of potential good dies per wafer	36		36		36	
Nb of good interposer per wafer	25		27		28	
Interposer + CoW Cost	$145.31	25.7%	$118.01	24.8%	$98.92	24.2%
HBM Stack Cost	$157.83	27.9%	$136.67	28.8%	$120.19	29.3%
GPU + Filler Dies Cost	$261.69	46.3%	$220.43	46.4%	$190.41	46.5%
CoW Stack Cost	**$564.83**	**100%**	**$475.11**	**100%**	**$409.52**	**100%**

The number of good CoW stack per wafer is estimated to ranges from 25 to 28 according to yield variations, which results in a CoW stack cost ranging from $410 to $565.

CoW Stack Cost Breakdown (Medium Yield)



Overview / Introduction

Company Profile & Supply Chain

Physical Analysis

Manufacturing Process Flow

Cost Analysis
- Summary
- Supply Chain
- Yields
- GPU Cost
- HBM Stack Cost
- Interposer Cost
- CoW Assembly Cost
- Component Cost

Selling Price Analysis

Feedback

About System Plus

Component Cost

	Low Yield		Medium Yield		High Yield	
	Cost	Breakdown	Cost	Breakdown	Cost	Breakdown
GPU + Filler Dies Cost	$261.69	41.8%	$220.43	41.8%	$190.41	41.7%
HBM Stack Cost	$157.83	25.2%	$136.67	25.9%	$120.19	26.3%
Interposer + CoW Cost	$145.31	23.2%	$118.01	22.4%	$98.92	21.7%
Package Substrate Cost	$27.23	4.4%	$27.23	5.2%	$27.23	6.0%
Final Packaging & Final Test cost	$15.00	2.4%	$15.00	2.8%	$15.00	3.3%
Yield losses	$18.77	3.0%	$10.56	2.0%	$4.56	1.0%
Component Cost	**$625.83**	**100%**	**$527.89**	**100%**	**$456.31**	**100%**

The package substrate is supposed to be made by Ibiden in Japan.

The Package size is 55x55mm, with a 5+2+5 structure (12 copper layers).

The price of the PCB package substrate is estimated to $27.23.

The final component cost (GPU + HBM + Substrate + Assembly) ranges from $456 to $626 according to yield variations.

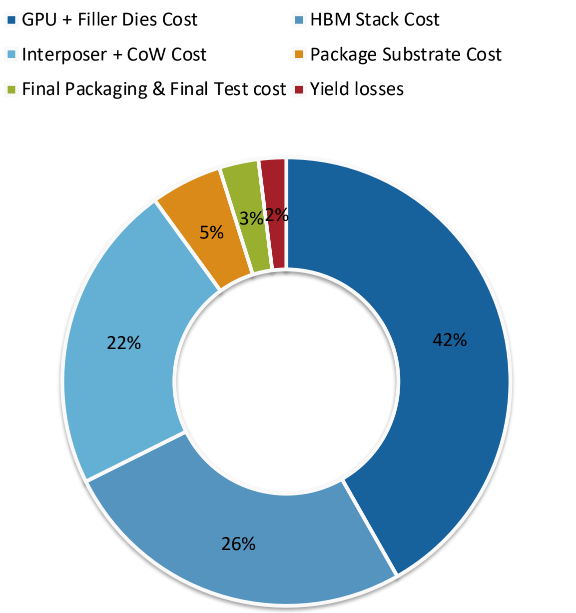
The GPU Die represents 42% of the component cost.

The HBM Stacks (x4) represents 26% of the component cost.

The interposer with CoW process represents 22% of the component cost.

The package substrate represent 5% of the component cost.

Component Cost Breakdown (Medium Yield)



SELLING PRICE

Manufacturer Financial Ratios

Financial ratios of NVIDIA (2018 financial results):

40.1% for Cost Of Sales (59.9% gross margin)
8.4% for G&A
18.5% for R&D
33.0% for Operating income

	Year Ended		
	January 28, 2018	January 29, 2017	January 31, 2016
Revenue	$ 9,714	$ 6,910	$ 5,010
Cost of revenue	3,892	2,847	2,199
Gross profit	5,822	4,063	2,811
Operating expenses			
Research and development	1,797	1,463	1,331
Sales, general and administrative	815	663	602
Restructuring and other charges	—	3	131
Total operating expenses	2,612	2,129	2,064
Income from operations	3,210	1,934	747
Interest income	69	54	39
Interest expense	(61)	(58)	(47)
Other, net	(22)	(25)	4
Total other income (expense)	(14)	(29)	(4)
Income before income tax	3,196	1,905	743
Income tax expense	149	239	129
Net income	$ 3,047	$ 1,666	$ 614

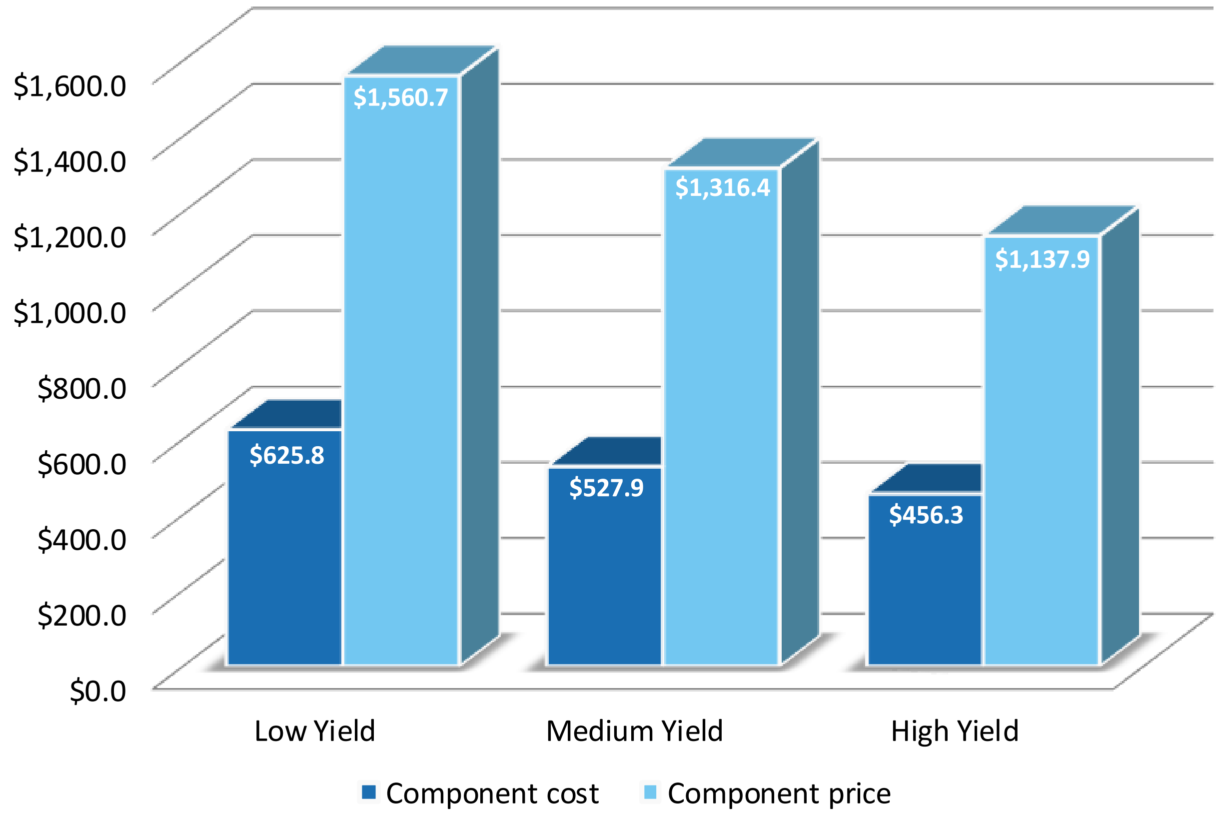
NVIDIA Consolidated Statements of Income
In millions of US Dollar

Estimated Manufacturer Price

	Low Yield		Medium Yield		High Yield	
	Cost	Breakdown	Cost	Breakdown	Cost	Breakdown
Component cost	**$625.8**		**$527.9**		**$456.3**	
NVIDIA Gross Profit	$934.8	+60%	$788.5	+60%	$681.6	+60%
Component price	**$1,560.7**		**$1,316.4**		**$1,137.9**	

We estimate that NVIDIA could realizes a gross margin of 60% on the Volta component, which results in a final component price ranging from $1,138 to $1,561.

Component Cost & Price According to Yield Variation



CUSTOMER FEEDBACKS

Feedbacks

Dear Customer,

Thank you for giving us the opportunity to serve you better.

Please help us by taking only a few seconds to give us your thoughts about the Reverse Costing Report that you have received.
Please note that without any feedback from you, we consider that the report satisfied you.

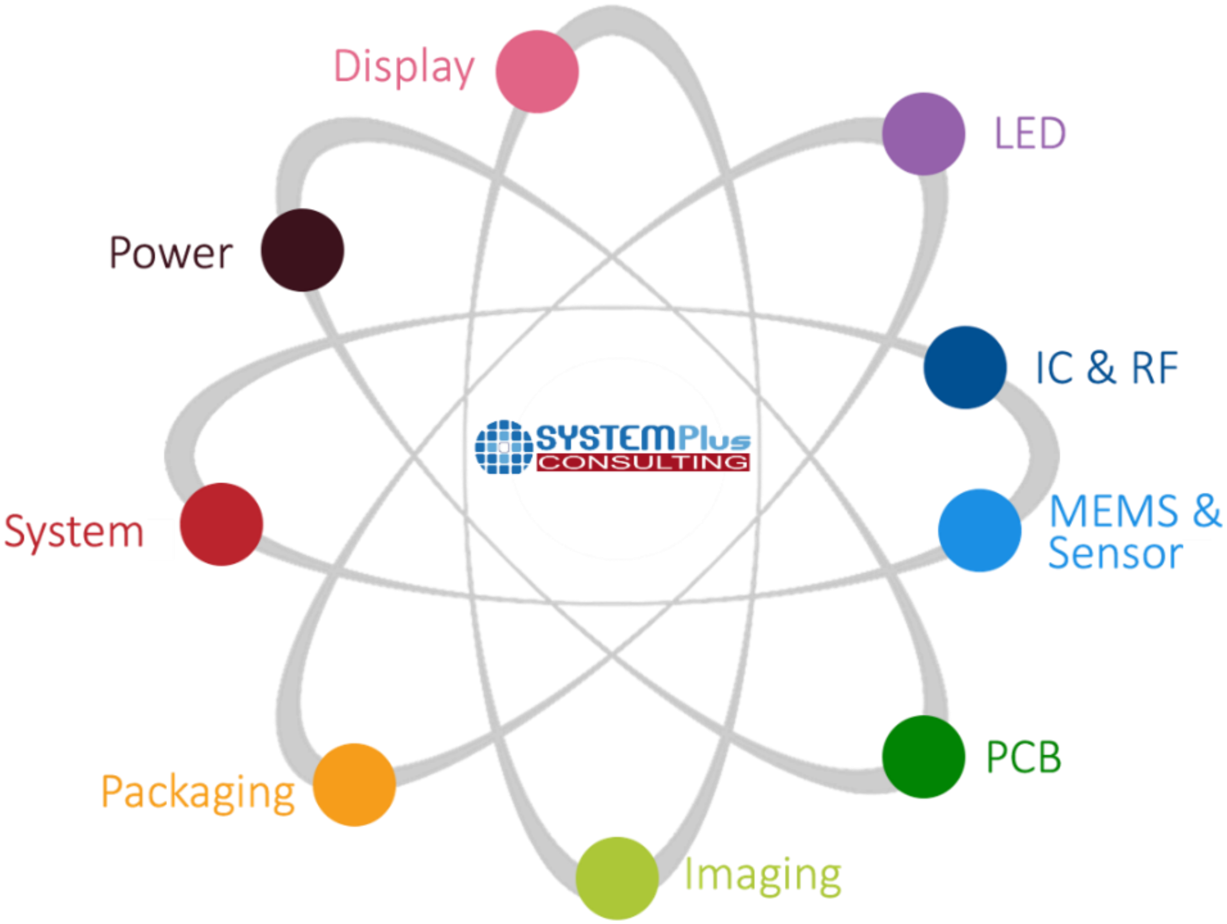
We appreciate to work with you and want to make sure we meet your expectations.

Sincerely,
Wilfried THERON
Quality Manager

Click below to access to
our online Customer Satisfaction Survey.

COMPANY SERVICES

Business Models Fields of Expertise



Contact

Headquarters
22 bd Benoni Goullin
44200 Nantes
FRANCE
+33 2 40 18 09 16
sales@systemplus.fr

Europe Sales Office
Lizzie LEVENEZ
Frankfurt am Main
GERMANY
+49 151 23 54 41 82
llevenez@systemplus.fr

America Sales Office
Steve LAFERRIERE
Phoenix, AZ
WESTERN US
T : +1 310 600 8267
laferriere@yole.fr

Troy Blanchette
EASTERN US
T : +1 704 859 0456
troy.blanchette@yole.fr

Asia Sales Office
Takashi ONOZAWA
Tokyo
JAPAN
T : +81 804 371 4887
onozawa@yole.fr

Mavis WANG
TAIWAN
T :+886 979 336 809
wang@yole.fr

www.systemplus.fr

Legal

DISCLAIMER

System Plus Consulting provides cost studies based on its knowledge of the manufacturing and selling prices of electronic components and systems. The given values are realistic estimates which do not bind System Plus Consulting nor the manufacturers quoted in the report. System Plus Consulting is in no case responsible for the consequences related to the use which is made of the contents of this report. The quoted trademarks are property of their owners.

Reverse Costing® is a deposed brand, by System Plus Consulting.

SERVICES

Reverse costing analysis represents the best cost/price evaluation given the publically available data, and estimates completed by industry experts.

These results are open for discussion. We can reevaluate this circuit with your information.