Intel Crescent Island GPU

Data-Center AI Inference Accelerator

PCB Teardown & Physical Analysis

Kurnal Insights

Version 1.0 - 2026 Q2

Kurnal-insights.com

Table of Contents

Document Control

Version History

Version	Date	Updates	Author
V0	2026/06/07 22:32	Analyze PCB data	Kurnal
V1	2026/06/19	Finish full PCB analysis	Kurnal

About the Author

This report is produced by Kurnal (Kurnal Insights), an independent semiconductor teardown and analysis effort. All findings are based on physical inspection and measurement of the board; estimations are derived analytically from observable geometry.

- **BiliBili:** @Kurnal
- **X (Twitter):** @Kurnalsalts
- **WeChat:** KurnalWeChat
- **Website:** Kurnal-insights.com

1. Overview

Intel Crescent Island is Intel's next-generation data-center GPU, positioned specifically as an AI inference accelerator for cloud and enterprise deployment rather than as a large-scale training part. It is built on Intel's Xe3P graphics architecture and is engineered around three priorities: performance-per-watt, very high memory capacity, and total cost of ownership.

The defining architectural choice of Crescent Island is its memory subsystem. Where competing accelerators from NVIDIA and AMD rely on HBM (High-Bandwidth Memory) stacked on a silicon interposer, Crescent Island instead uses LPDDR5X. This trades peak per-pin bandwidth for a substantially larger, cheaper, and simpler memory pool: the reference design carries 160 GB of LPDDR5X, while partner implementations are expected to scale up to 480 GB. That capacity-first stance is well matched to modern inference and “agentic” AI workloads, where the ability to hold very large models and long context in memory often matters more than raw HBM bandwidth, and where deployment economics dominate.

Equally important is the platform target. Crescent Island is designed to drop into standard, air-cooled PCIe server slots rather than requiring exotic liquid cooling or a proprietary OAM module. The board analyzed here is a standard full-height, short-length PCIe Gen5 x16 card powered through a single 12V-2x6 (12VHPWR) connector – a form factor any datacenter can rack at scale today.

1.1 Purpose & Scope of This Report

This report is a physical-layer teardown of a Crescent Island board. The goal is to reconstruct, from the hardware itself, the engineering decisions Intel and its board partners made and to infer the parts of the design that are not publicly disclosed. The analysis covers the printed-circuit board (mechanical dimensions, layer stack-up, copper weights, surface finish, markings and connectors), the power-delivery network, the GPU package and an estimate of the silicon die beneath it, and the LPDDR5X memory subsystem (package count, capacity, bus width and resulting bandwidth).

Where exact figures are not printed on the board, this report derives them analytically – die size is estimated from the BGA pad field and pad pitch, die-per-wafer from that die size, and memory capacity and bandwidth from package count, package type and the LPDDR5X data rate. These estimates are clearly labelled as such throughout.

1.2 Methodology

- Direct measurement: PCB outline, GPU package outline, BGA pad pitch and connector geometry were measured from top- and back-side imagery of the board.
- Cross-section inspection: layer count, copper thicknesses, inter-layer dielectric spacing and surface metallization were read from microscope cross-sections of the PCB edge.
- Die-size estimation: the sparse central pad region of the package substrate is mapped to the GPU die footprint, then scaled by the measured 0.7 mm pad pitch.
- Die-per-wafer (DPW): computed from the estimated die area on a standard 300 mm wafer assuming a full 1:1 reticle field.
- Memory estimation: capacity and bandwidth derived from the observed LPDDR5X package count and BGA563 package type, combined with published LPDDR5X data rates.
- Comparative benchmarking: each measured/estimated metric is placed against NVIDIA and AMD reference parts (H200, RTX 6000 Ada, RTX 4090, MI210, GB200, etc.) for context.

1.3 Key Findings at a Glance

Dimension	Finding
Architecture	Intel Xe3P, AI-inference-optimized data-center GPU
Form factor	Standard full-height, short-length PCIe Gen5 x16, air-cooled
Board size	235 mm (L) × 110 mm (W); body width 98 mm – shorter than all compared parts
Power input	Single 12V-2x6 (12VHPWR) connector
PCB stack-up	10-layer board; 30 µm outer copper; ENIG surface finish
VRM	18 power phases total (13 core / 3 memory / 2 I/O); est. board power ~350 W
GPU package	69 mm × 70 mm flip-chip BGA, 0.7 mm pitch, ~9,575 balls
Estimated die	~24.5 mm × 31.5 mm ≈ 772 mm² (reticle-class, ~75 dies/wafer)
Memory	LPDDR5X, up to 20 × BGA563 packages, up to 480 GB, 1280-bit bus
Memory BW	~1.36 TB/s (base) to ~1.54 TB/s (max)

2. Executive Summary

Crescent Island is Intel's bet that the economics of AI inference will be won on memory capacity and platform simplicity rather than on peak bandwidth. The teardown supports that thesis at every layer of the board.

Mechanically, the card is notably compact: at 235 mm × 110 mm it is the shortest board in its comparison set, undercutting NVIDIA's H200 PCIe and RTX 6000 Ada (both ~267 mm) and AMD's MI210 (290 mm). The short length, full-height profile, PCIe Gen5 x16 edge and single 12V-2x6 power input all point to a part deliberately built for dense, air-cooled, standard-rack inference servers.

The power-delivery network is substantial but not extreme: 18 total phases (13 for the GPU core, 3 for memory, 2 for I/O/system) feeding an estimated ~350 W board power. The PCB itself is a 10-layer design with heavy 30 µm outer copper for transient current handling and an ENIG (gold) surface finish for reliable, oxidation-resistant bonding – a robust but cost-conscious construction consistent with a high-volume datacenter product.

The GPU silicon is the headline. From the 69 mm × 70 mm flip-chip package and its ~9,575-ball BGA field at 0.7 mm pitch, the GPU die is estimated at roughly 24.5 mm × 31.5 mm, or about 772 mm². That places Crescent Island firmly in reticle-class territory – larger than NVIDIA's AD102 (~616 mm²) and GB202 (~761 mm²) and approaching H100 (~814 mm²). At that die size, a 300 mm wafer yields on the order of 75 candidate dies.

The memory subsystem is where Crescent Island diverges most sharply from the field. Up to 20 LPDDR5X BGA563 packages – 12 on the front around the GPU and 8 on the back – deliver up to 480 GB across a 1280-bit (20-channel) interface. At an 8533 MT/s base data rate this yields ~1.36 TB/s, rising to ~1.54 TB/s at 9600 MT/s. That bandwidth is below HBM-based parts (H100 ~3.36 TB/s, GB200 ~8 TB/s) but is competitive with GDDR7 cards while offering 6–15× their capacity at far lower cost and power.

Bottom line: Crescent Island is not built to win bandwidth benchmarks. It is built to host very large models cheaply, in standard servers, at scale – a capacity-and-TCO play for the inference era. The hardware is internally consistent with that strategy from the die all the way out to the connector.

3. Company Profile & Road-map

3.1 Product Profile

Intel Crescent Island is Intel's next-generation AI inference accelerator targeting cloud and enterprise datacenter deployments. Built on the Xe3P architecture, it is optimized for performance-per-watt, high memory capacity, and cost-efficient AI inference workloads rather than large-scale AI training.

Unlike competing accelerators that rely on HBM memory, Crescent Island adopts LPDDR5X memory technology, enabling significantly higher memory capacity while reducing board complexity, power consumption, and overall system cost. The accelerator is designed for standard air-cooled PCIe server platforms, making it suitable for large-scale inference deployment and agentic AI applications.

Current disclosures indicate a reference configuration featuring 160 GB LPDDR5X memory, with future partner implementations supporting capacities up to 480 GB.

3.2 Data-Center GPU Road-map

Crescent Island is the latest step in Intel's data-center GPU lineage, succeeding the Ponte Vecchio (Data Center GPU Max 1100) generation and the Rialto Bridge program.

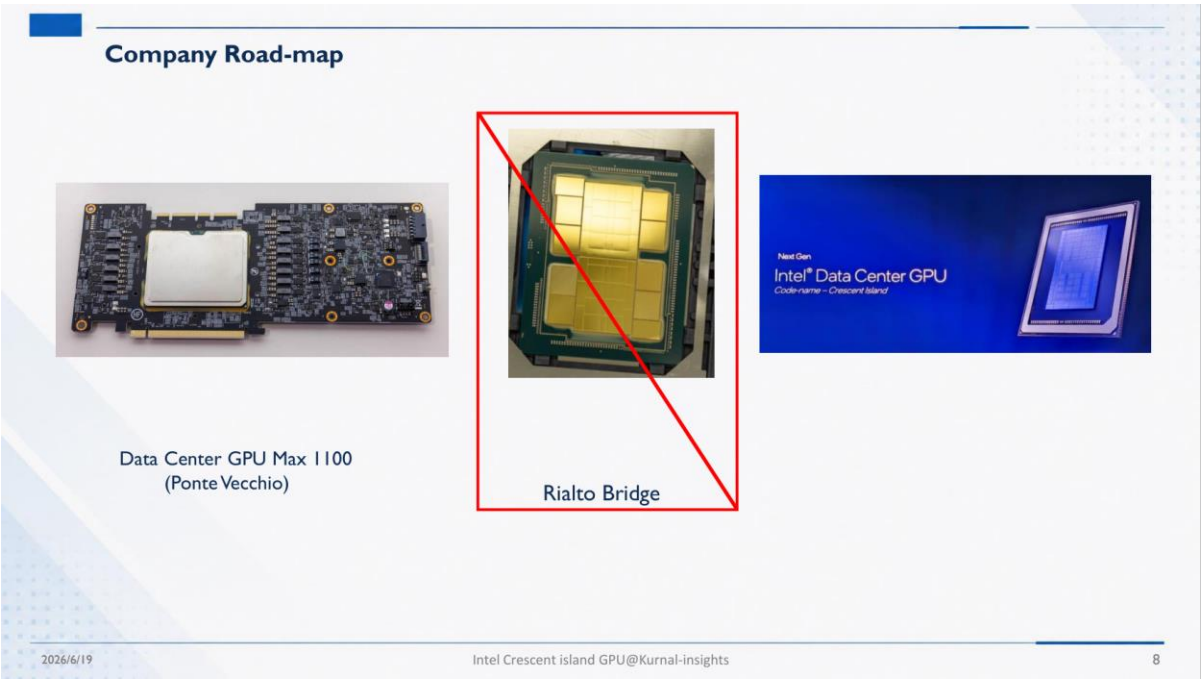

Figure 3-1. Intel data-center GPU road-map: Ponte Vecchio (Max 1100) → Rialto Bridge → Crescent Island.

4. Physical Analysis – PCB

4.1 Board Overview

The board is a standard full-height PCIe add-in card. Front (top) and back views show a GPU package centrally located with memory packages distributed around it on both sides.

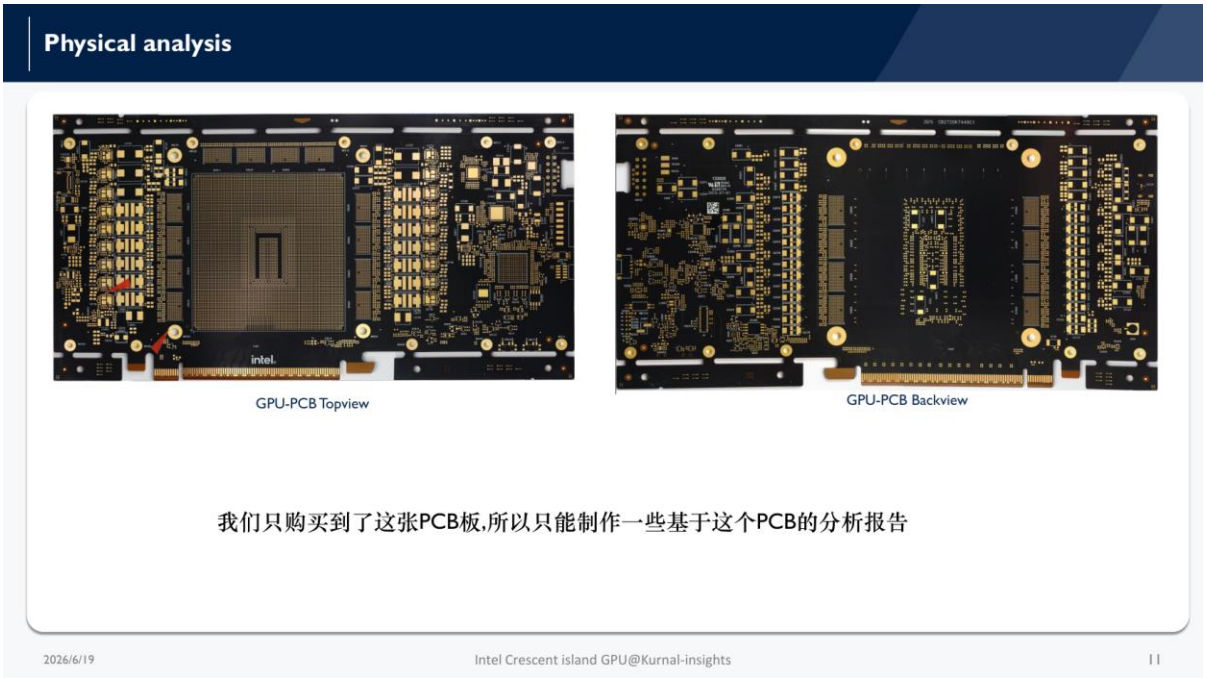

Figure 4-1. GPU-PCB top view (left) and back view (right).

4.2 PCB Dimensions

- **PCB total length:** 235 mm
- **PCB total width:** 110 mm
- **Board body width:** 98 mm (standard full-height)

Additional measured reference points include a 120 mm height reference and a 12 mm bracket region. The board is notably short relative to competing accelerators.

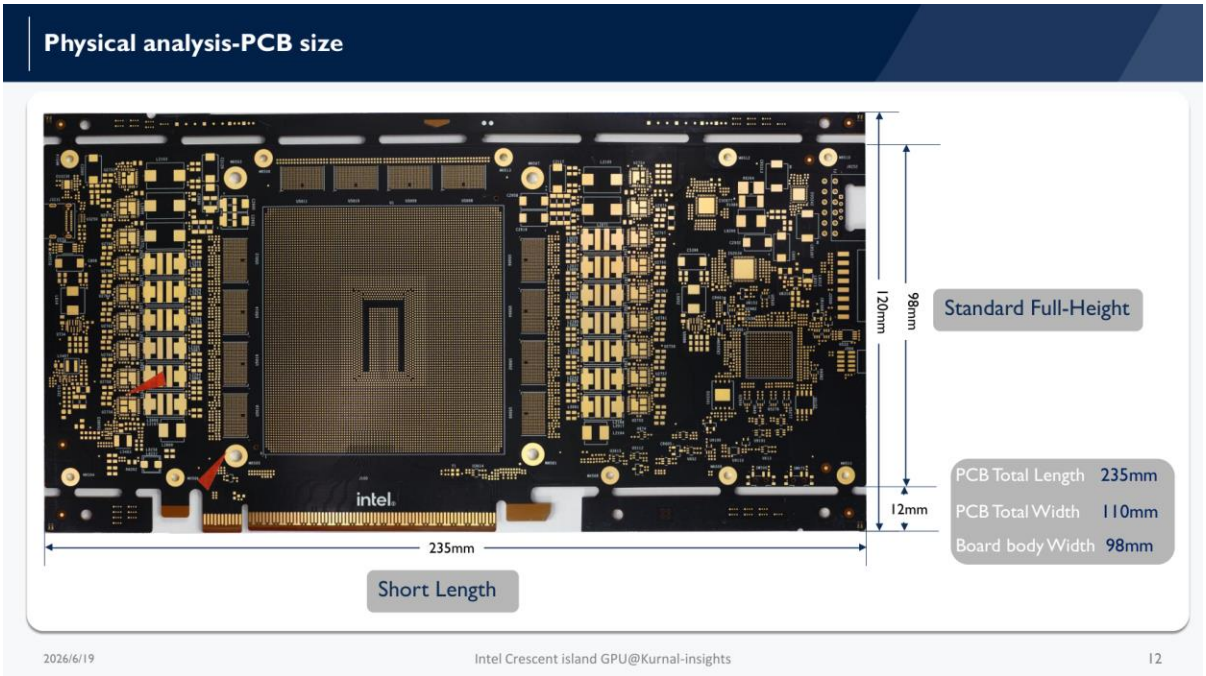

Figure 4-2. PCB mechanical dimensions – 235 mm length, 110 mm width, 98 mm body width.

4.3 Size Comparison vs. Competing Accelerators

Metric	Intel Crescent Island	Intel Ponte Vecchio	NVIDIA H200 PCIe	AMD MI210 PCIe
Length	235 mm	266.7 mm	267 mm	290 mm
Height	110 mm	111.15 mm	112 mm	125 mm

Other references in the source comparison include the NVIDIA RTX 6000 Ada (267 mm × 112 mm) and RTX 4090 FE (304 mm × 137 mm). Crescent Island is the shortest board in the set – advantageous for dense server chassis.

4.4 PCB Mark Analysis

Silkscreen and laser markings on the back side identify the manufacturer and traceability data. Identified mark categories include:

- Manufacturer logo
- Material / flammability rating
- UL file number
- Date code (WWYY format)
- Revision code
- Card part number and customer / Intel P/N
- Data Matrix code: V08261335030016

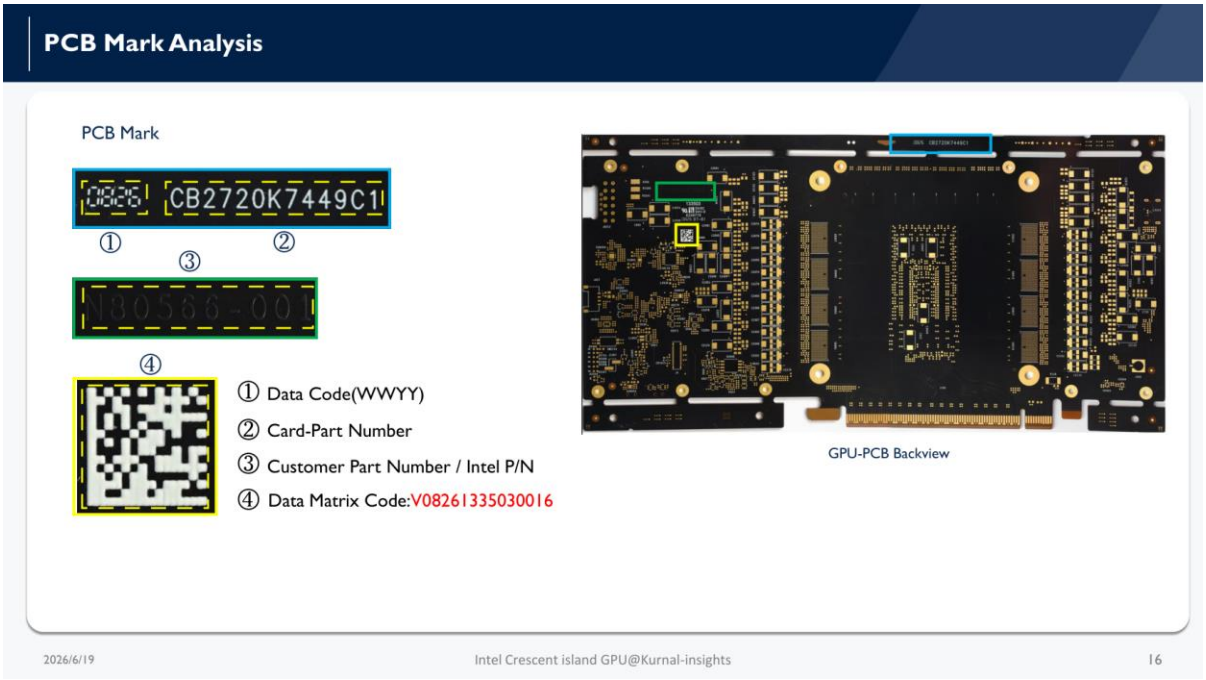

Figure 4-3. PCB back-side markings: date code, part numbers and Data Matrix traceability code.

4.5 Connector & Socket Analysis

The board exposes the following interfaces and on-board controls:

- PCIe 5.0 x16 edge connector (host interface)
- USB-C port
- 12VHPWR / 12V-2x6 power connector
- Jumper set and SMD switch (configuration)
- GPIO header
- SMP socket
- Fan-control header (probable)

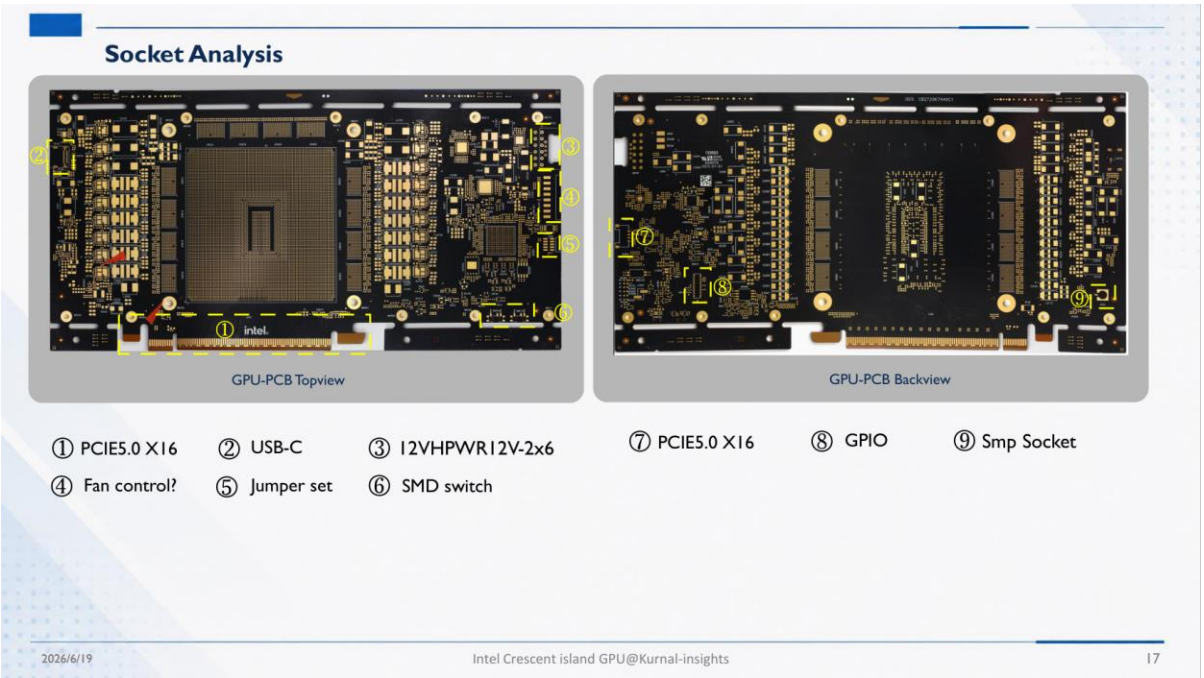

Figure 4-4. Connector and socket map across top and back sides.

4.6 PCB Stack-up Architecture

Edge cross-section imaging reveals the layer construction. Two findings stand out:

- **Copper power layer:** ~30 µm thickness – designed for high GPU transient current and lower IR drop under load spikes
- **Surface metallization (ENIG):** gold layer ~30–50 nm – high-reliability bonding interface, improved contact-resistance stability and oxidation resistance

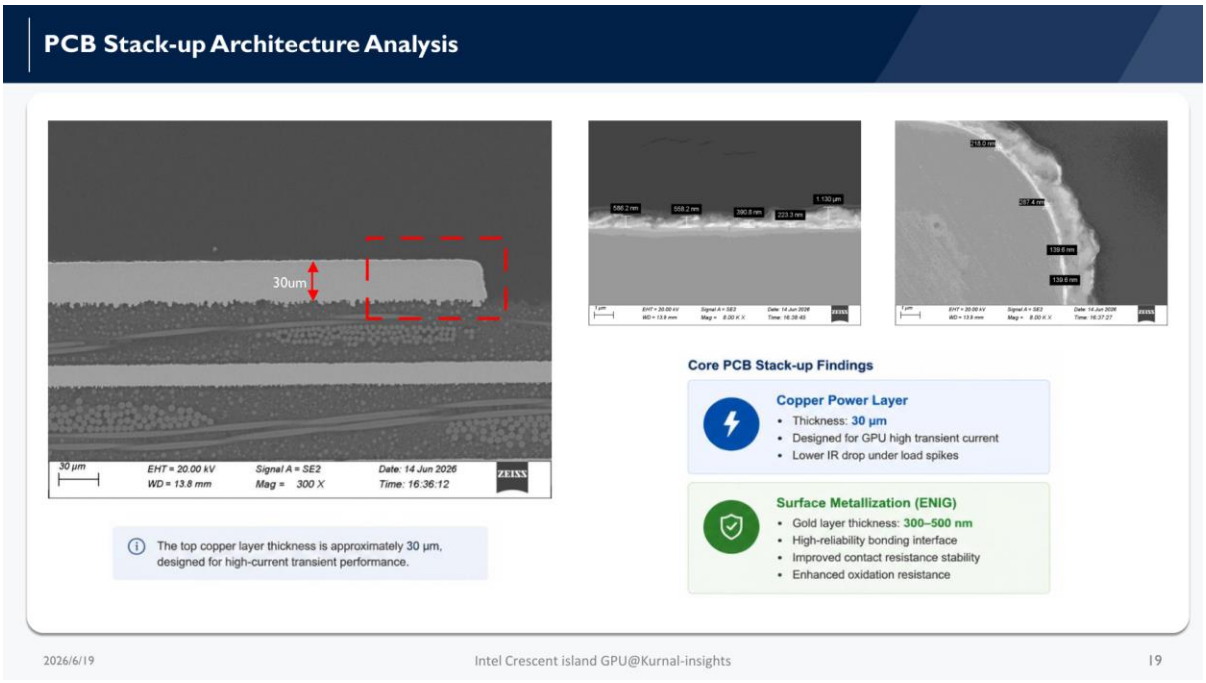

Figure 4-5. PCB edge cross-section – ~30 µm outer copper and ENIG surface finish.

4.7 Inter-layer Dielectric & Copper Spacing

The board is a 10-layer stack-up. Measured per-layer copper thickness and inter-layer dielectric gaps:

Layer	Cu Thickness	Layer Gap
Layer 1	30 µm	47 µm
Layer 2	16 µm	57 µm
Layer 3	20 µm	60 µm
Layer 4	20 µm	60 µm
Layer 5	20 µm	60 µm
Layer 6	20 µm	60 µm
Layer 7	26 µm	47 µm
Layer 8	16 µm	53 µm
Layer 9	30 µm	76 µm
Layer 10	60 µm	60 µm

Heavy outer-layer copper (30 µm on L1, 60 µm on L10) reflects the high current demands of the GPU and memory power rails.

5. Power-Supply Design

The voltage-regulator modules (VRMs) are distributed around the GPU package. Phase counting identifies three distinct power domains:

Power rail	Phases	Description
VDD_Core	13	GPU core power
VDD_Mem	3	LPDDR5X memory power
VDD_IOM	2	I/O and system power
Total	18	Total board power phases

Key observations:

- 13-phase VRM for GPU core power delivery
- 3-phase VRM for LPDDR5X memory
- 2-phase VRM for I/O and system power
- **Estimated total board power:** ~350 W

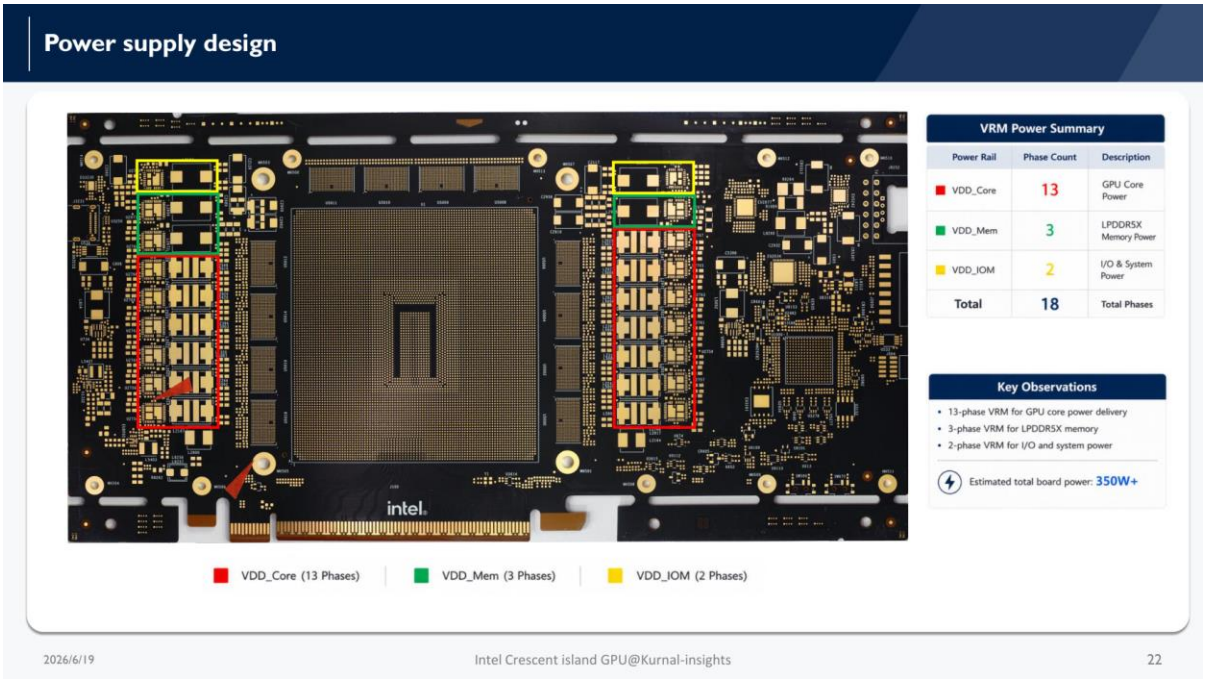

Figure 5-1. Power-delivery layout – VDD_Core (13 phases), VDD_Mem (3 phases), VDD_IOM (2 phases).

6. GPU Analysis

6.1 GPU Package Size

Parameter	Value
Package size	69 mm × 70 mm (length × width)
Pad pitch	0.7 mm (center-to-center)
Pad count	~9,575 total BGA pads
Package type	BGA, flip-chip package

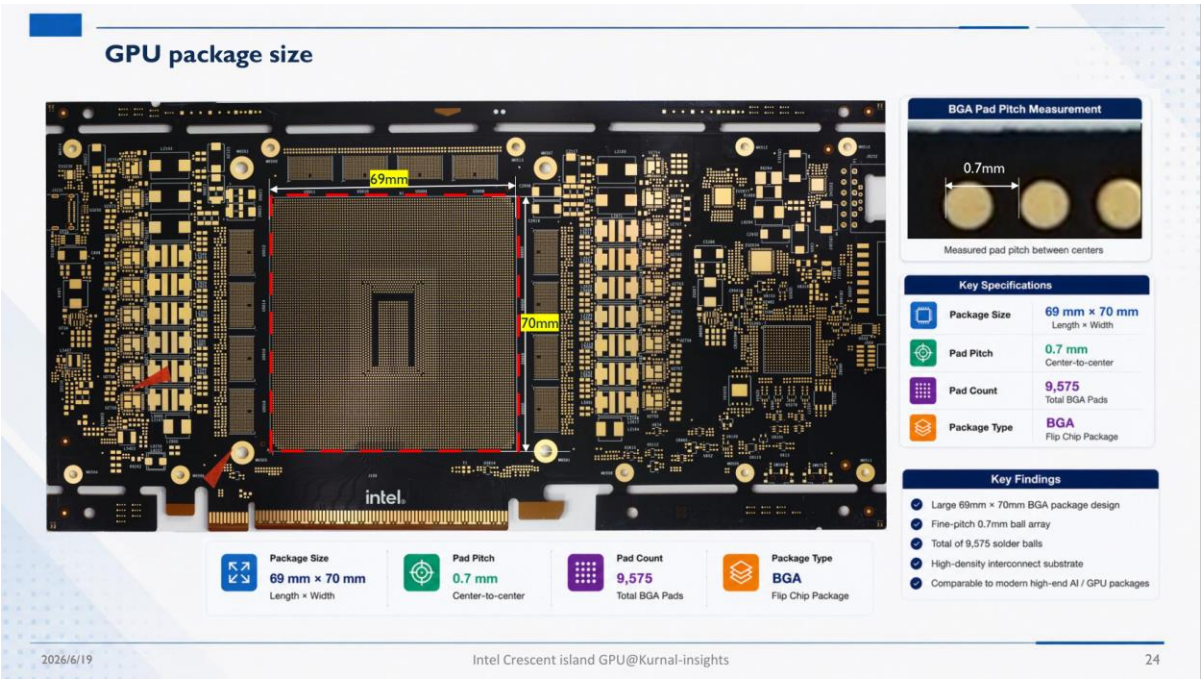

Figure 6-1. GPU package – 69 mm × 70 mm flip-chip BGA with 0.7 mm pad pitch.

6.2 SoC Die-Size Estimation (from BGA sparse-region analysis)

The central BGA sparse region – a low-density pad area on the substrate – corresponds to the GPU die location. Scaling the sparse-region pad count by the measured 0.7 mm pitch yields the die footprint.

Parameter	Value	Note
Width (horizontal)	24.5 mm	35 pads
Height (vertical)	31.5 mm	45 pads
Pad pitch	0.7 mm	center-to-center
Total pad count	9,575	full package
Package size	69 mm × 70 mm	measured

- **Estimated die size:** 24.5 mm × 31.5 mm ≈ 772 mm²

This represents a large, reticle-class GPU die.

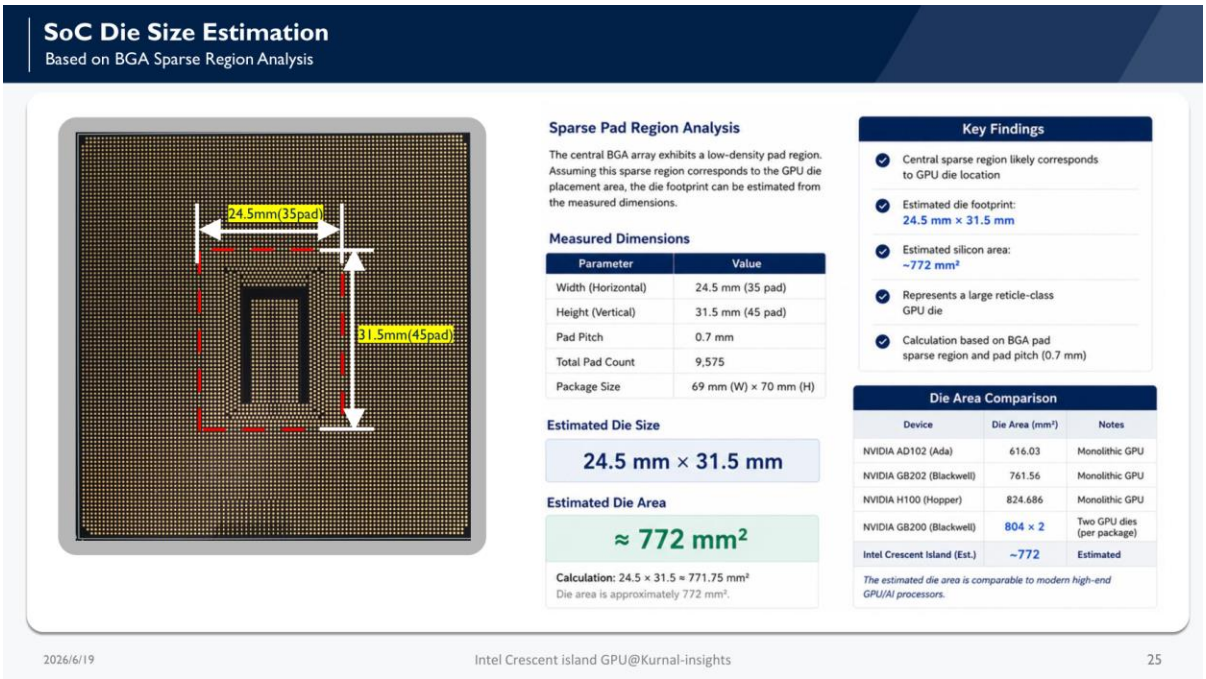

Figure 6-2. SoC die-size estimation – 24.5 mm × 31.5 mm ≈ 772 mm².

6.3 Die Area Comparison

Device	Die area (mm²)	Architecture
NVIDIA AD102 (Ada)	~616	Monolithic GPU
NVIDIA GB202 (Blackwell)	~761	Monolithic GPU
NVIDIA H100 (Hopper)	~814	Monolithic GPU
NVIDIA GB200 (Blackwell)	~800 × 2	Two GPU dies (per package)
Intel Crescent Island (est.)	~772	Estimated

The estimated Crescent Island die is consistent with modern high-end GPU processors.

6.4 Die-Per-Wafer (DPW) Analysis

Using the estimated die size on a 300 mm wafer with a full 1:1 reticle mask:

- **DPW (dies per wafer):** ~75
- **Mask type:** Full mask 1:1 (1:1 scaling)
- **Mask utilization (MFU):** 89.947%
- **Wafer utilization (DWE):** 81.885%
- **Mask-on-wafer efficiency (MOWE = MFU × DWE):** 73.65%

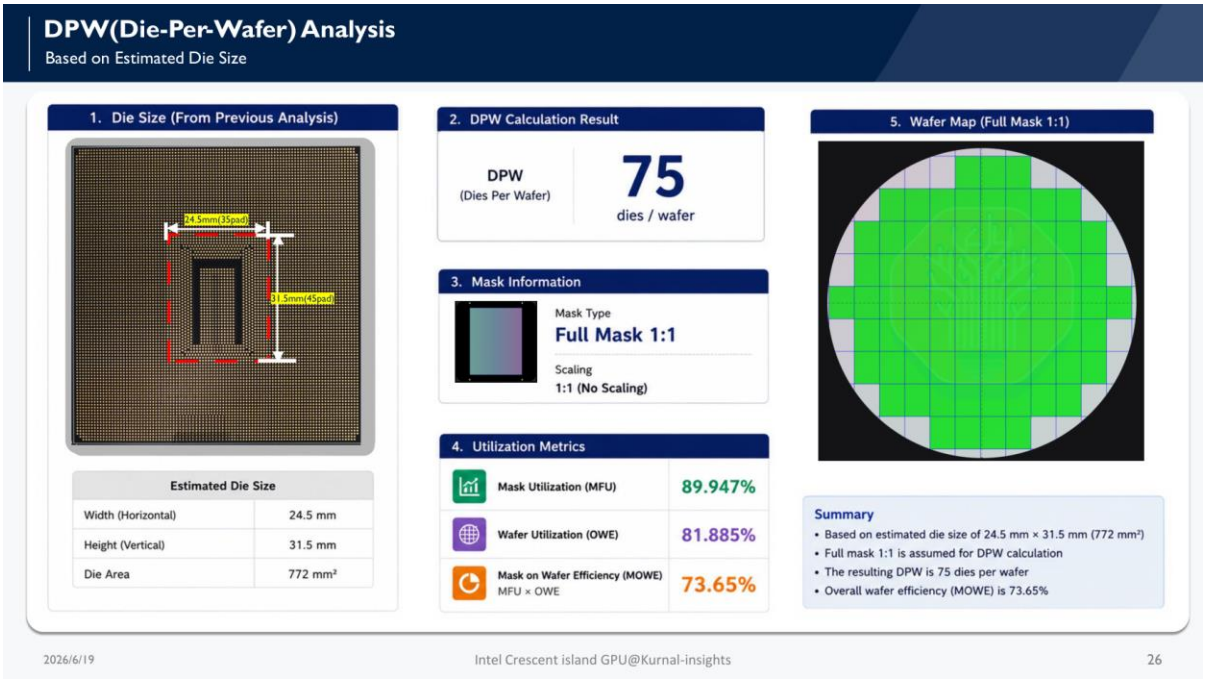

Figure 6-3. Die-per-wafer analysis – ~75 dies/wafer at 73.65% overall efficiency.

7. Memory Analysis

7.1 LPDDR5X Package Placement

LPDDR5X packages are distributed on both sides of the PCB for density:

- 12 LPDDR5X packages on the front side, surrounding the GPU
- 8 LPDDR5X packages on the back side of the PCB
- 20 packages total (dual-side placement)

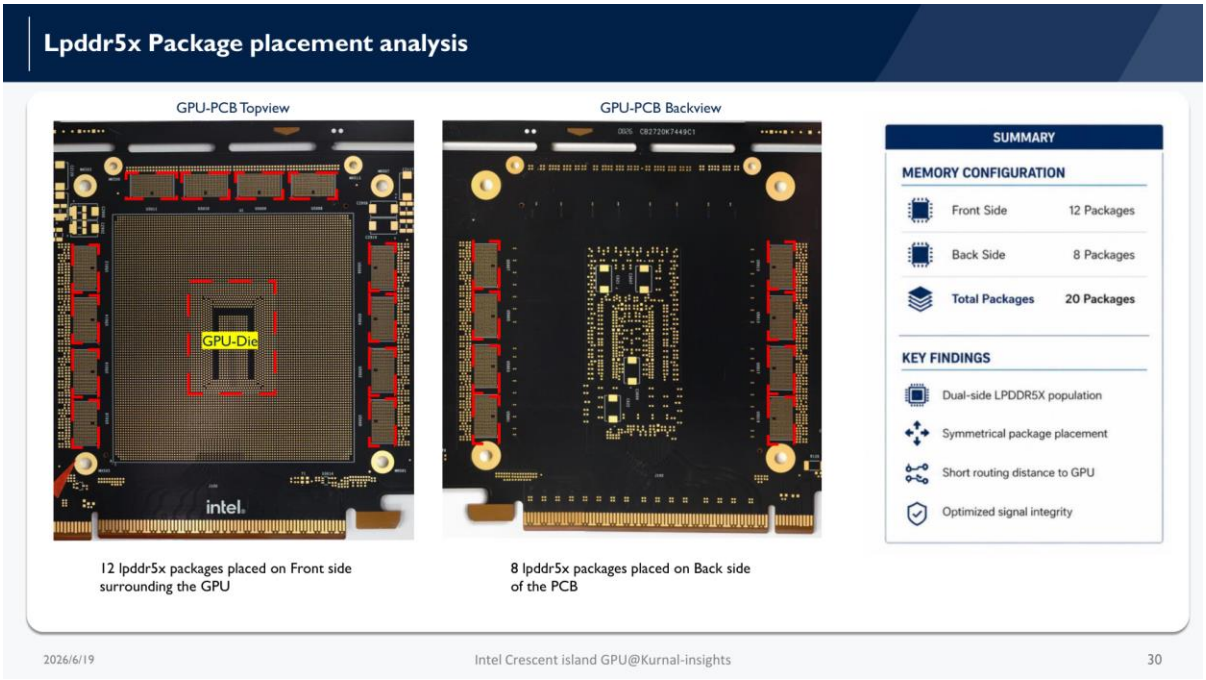

Figure 7-1. LPDDR5X placement – 12 packages front, 8 packages back, surrounding the GPU.

7.2 LPDDR5X BGA563 Package

- **Package type:** BGA563 (common LPDDR5X chip)
- **Package dimensions:** 12.35 mm × 7.00 mm
- **Balls per package:** 563

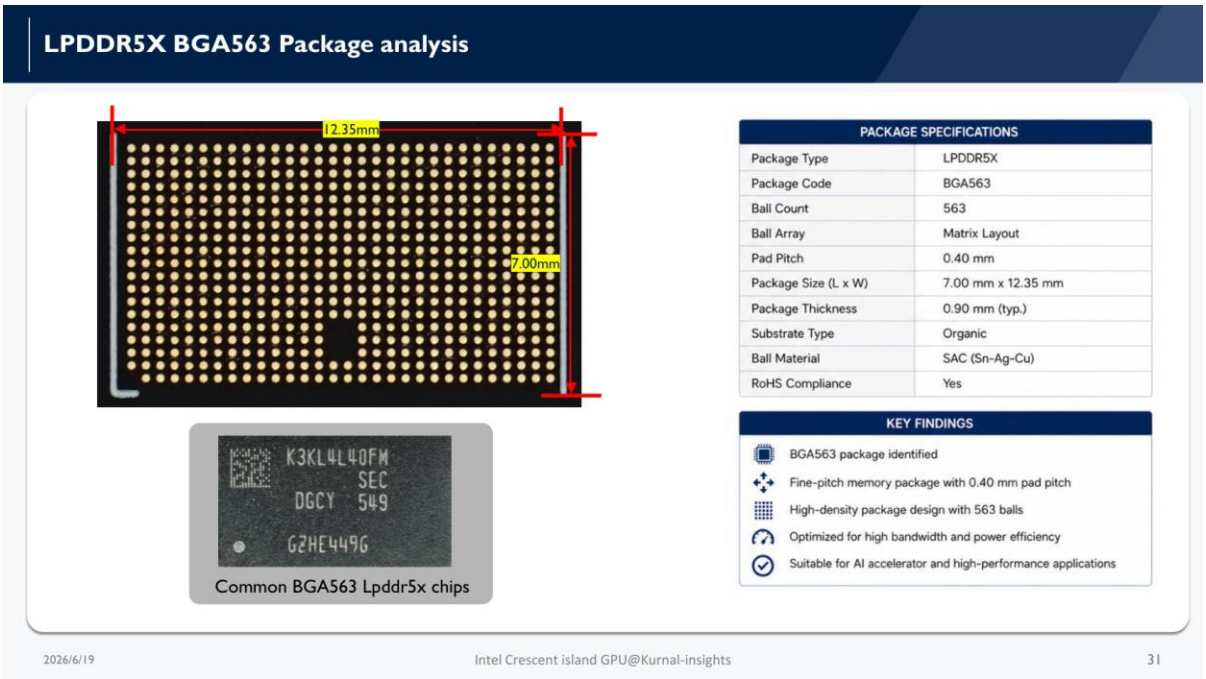

Figure 7-2. LPDDR5X BGA563 package – 12.35 mm × 7.00 mm.

7.3 Memory Capacity Estimation

Parameter	Value
Memory type	LPDDR5X
Package type	BGA563
Packages	20
Balls per package	563
Total capacity	480 GB
Bus width	1280-bit

Architecture highlights:

- Ultra-wide memory subsystem – the 1280-bit bus enables massive bandwidth for AI workloads.
- High-capacity design – 480 GB supports large-scale inference and training.
- Dual-side package placement for optimal density.
- Low-power LPDDR5X optimized for AI and HPC applications.

7.4 Memory Bandwidth Estimation

Bandwidth = Bus Width × Data Rate ÷ 8. With a 1280-bit bus (20 channels):

Scenario	Bandwidth
Base (8533 MT/s)	~1.36 TB/s
Maximum (9600 MT/s)	~1.54 TB/s

7.5 Peak Memory-Bandwidth Comparison

Device	Memory	Capacity	Interface	Bandwidth
NVIDIA GB200	HBM3E	90 GB	8192-bit	8.00 TB/s
NVIDIA H100	HBM3E	80 GB	5120-bit	3.36 TB/s
NVIDIA A100	HBM2E	80 GB	5120-bit	2.00 TB/s
NVIDIA A100	HBM2E	40 GB	5120-bit	1.79 TB/s
NVIDIA RTX 5090	GDDR7	32 GB	512-bit	1.79 TB/s
Intel Crescent Island	LPDDR5X	480 GB	1280-bit	1.36–1.54 TB/s

Key takeaways: the LPDDR5X subsystem delivers up to ~1.54 TB/s – below HBM parts but competitive with GDDR7 – while offering far greater capacity (480 GB) at lower power and cost, optimized for inference and datacenter workloads.

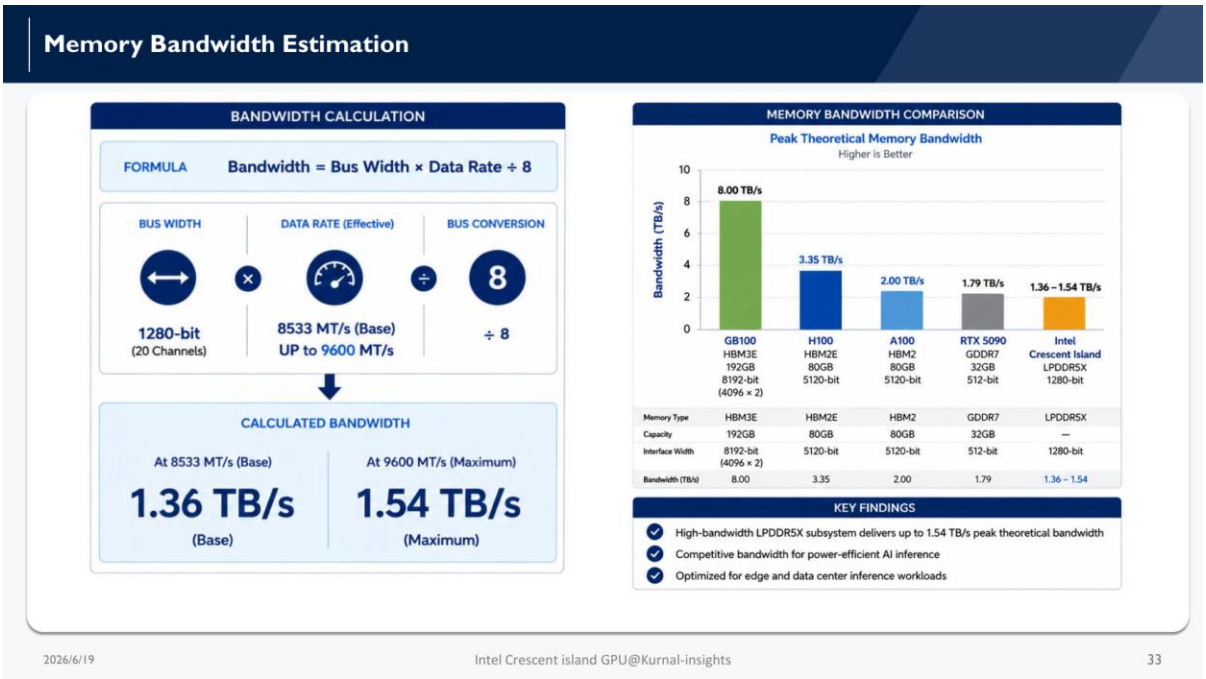

Figure 7-3. Peak theoretical memory-bandwidth comparison.

Disclaimer

This report is an independent physical teardown and analysis produced by Kurnal Insights. Measured dimensions are taken directly from the board; die size, die-per-wafer, memory capacity and bandwidth figures are analytical estimates derived from observable geometry and published component specifications, and may differ from official Intel figures. Competitor specifications are cited for comparison only. This document is provided for educational and informational purposes – Not For Sale.