

IT Hardware

Key message

1. Google's (US) TPU v8 is the first generation TPU to be split into two distinct architectures, 8t and 8i, each optimized for different workload requirements. Mass production is expected to begin across the hardware supply chain in 4Q26F, providing an additional growth driver for the AI server hardware market in 2027F.
2. Although TPU v8 rack designs are differentiated into 8t and 8i configurations, they are largely extensions of the TPU v7 platform. As volume deployments ramp up, suppliers benefiting from specification upgrades in thermal management, power delivery, PCB/ CCL materials, substrates, and optical circuit switch (OCS) solutions are expected to see increased demand.
3. We expect 2027F to be a breakout year for Google TPU deployments. From a chip perspective, we estimate shipments for TPU v8 to increase from around 4.02mn units in 2026F to 9.10mn units in 2027F, up 126% YoY. Total TPU v8 rack shipments are projected to double in 2027F. This substantial volume ramp will drive demand across the AI hardware value chain, including TPU chips, ODM assembly, thermal management solutions, power systems, PCB/ CCL materials, and optical networking components. Key beneficiaries, with TPU-related revenue to account for more than 10% of total sales, include chip partner MediaTek (2454 TT); Celestica (US); power supply provider Delta Electronics (2308 TT); ABF substrate supplier Unimicron (3037 TT); CCL supplier Elite Material Co. (2383 TT); PCB maker Gold Circuit Electronics (2368 TT); liquid cooling supplier Asia Vital Components (3017 TT); and HVLP copper foil play Co-Tech Development (8358 TT).

TPU v8 ramp-up to drive AI-related growth in 2027F

Event

Google's (US) eighth-generation Tensor Processing Unit (TPU) platform (TPU v8) is expected to enter mass production across the hardware supply chain in 4Q26F, providing an additional growth catalyst for the AI server hardware market in 2027F.

Impact

TPU v8, the first TPU generation to be split into two distinct architectures (8t and 8i, each optimized for different workload requirements) to enter mass production across the hardware supply chain in 4Q26F. Google has now advanced to its eighth-generation TPU platform (v8). Unlike TPU v7 (Ironwood), which was designed to support both training and inference workloads within a unified architecture, TPU v8 is the first generation to be explicitly bifurcated into TPU 8t (Zebrafish) and TPU 8i (Sunfish), each optimized for specific workload profiles. TPU 8t (training) targets large-scale model pre-training and embedding-intensive workloads. It retains Google's 3D Torus scale-up architecture and supports up to 9,600 TPUs within a single Superpod. In contrast, TPU 8i (inference) is designed for post-training, sampling, real-time inference, and reasoning workloads. It adopts the new Boardfly inter-chip interconnect (ICI) topology, with each pod supporting up to 1,152 TPUs, including as many as 1,024 active TPUs. Demand for Google TPUs can be broadly divided into internal and external sources. Internal demand is primarily driven by Google DeepMind and Google's own AI services, including the training, post-training, and inference requirements of frontier models such as Gemini. On the external side, Anthropic (US) is currently one of the largest publicly disclosed TPU customers. In 2025, Anthropic announced plans to expand its deployment on Google Cloud TPU infrastructure to as many as one million TPUs, representing more than 1GW of compute capacity, with deployment scheduled throughout 2026. In April 2026, Anthropic further announced a new partnership with Google and Broadcom (US), securing multiple gigawatts of next-generation TPU capacity (TPU v8). The capacity is expected to come online starting in 2027F and will primarily support Claude frontier models.

TPU v8 rack designs are differentiated into 8t and 8i configurations, but both remain largely extensions of the TPU v7 platform. TPU 8t retains the 3D Torus architecture optimized for large-scale training workloads, while doubling ICI bandwidth compared with the previous generation. The platform also introduces the new Virgo scale-out network. In contrast, TPU 8i adopts an entirely new Boardfly topology, with each tray comprising a four-TPU building block. Eight boards are combined into a 32-chip group, and 36 groups are interconnected via optical circuit switch (OCS), forming a pod with up to 1,152 physical TPUs and 1,024 active TPUs. According to current supply chain checks, beyond the TPU accelerators themselves, each L6 board in the TPU v8 platform further integrates power modules and high-speed transceivers/ optics, while also incorporating Google's fourth-generation (Gen4) liquid-cooling solution. As a result, the key upgrade of TPU v8 extends beyond TPU compute performance and represents a broader system-level architectural evolution spanning compute, memory, networking, power delivery, thermal management, and optical interconnects. At present, Celestica (US) remains the primary L10/ L11 system assembly partner for Google TPU platforms. Hon Hai Precision (2317 TT, NT\$251.5, OP) will gain market share in TPU tray and L11 rack assembly for TPU v8, following the assembly of the CPU tray and rack in v7, while Inventec (2356 TT, NT\$59.5, N), currently a TPU L6 supplier, may also expand into system-level integration.

TPU v8 volume ramp to benefit suppliers of thermal solutions, power systems, PCB/ CCL materials, substrates & OCS components.

Liquid cooling thermal solutions. TPU v8 upgrades from chip-level liquid cooling to system-level liquid cooling, driving simultaneous increases in CDU capacity and liquid-cooling content. Based on supply chain checks, cooling capacity rises from approximately 1.4MW in TPU v7 to around 2MW in TPU 8t, expanding demand for liquid-cooling components such as manifolds, quick disconnects (QD), and tubing. We estimate liquid-cooling content per rack (excluding CDU) at US\$45–50k. Key suppliers of TPU 8t cold plates are Asia Vital Components (AVC; 3017 TT, NT\$3,435, OP) and Furukawa Electric (JP), while Cooler Master (TW; unlisted) and Boyd (US) are the primary suppliers for TPU 8i.

Power systems. TPU v8 moves to 10kW PSU and begins adopting HVDC-compatible architectures, significantly increasing power-system value per rack. Under the 50V platform architecture, TPU v8 is estimated to utilize three power shelves with six 10kW PSUs each. Compared with TPU v7's 5.5kW PSU design, PSU content per rack is expected to increase from roughly US\$20k to US\$28–30k, representing a 40–50% uplift. TPU v8 is also designed to support $\pm 400V$ DC, with major beneficiaries including Delta Electronics (2308 TT, NT\$1890, OP), followed by Flex (US) and Lite-On Technology (2301 TT, NT\$279.5, OP).

PCB/ CCL materials & ABF substrates. TPU v8 benefits from simultaneous upgrades in substrate size, layer count, and material specifications. ABF substrates are expected to increase from approximately 80–85mm and 16 layers in TPU v7 to approximately 90–100mm and 18–22 layers in TPU v8, driving an estimated 40–50% increase in ASP. A key beneficiary in ABF substrates is Unimicron Technology (3037 TT, NT\$1,175, OP). Meanwhile, the next-generation TPU 8t and TPU 8i mainboards are expected to feature 34-layer and 40-layer PCB designs, respectively, while CCL materials are upgraded to M8+ grade and paired with HVLP4 copper foil, resulting in an estimated 70–80% increase in CCL and PCB ASP. Key beneficiaries include Elite Material Co. (EMC; 2383 TT, NT\$4,935, OP) in CCL, Gold Circuit Electronics (2368 TT, NT\$1,120, OP) in PCB manufacturing, and Co-Tech Development (8358 TT, NT\$495, NR) in HVLP copper foil.

Optical networking/ OCS components. The architectural divergence between TPU 8t, optimized for scale, and TPU 8i, optimized for latency, is expected to drive incremental demand for optics, fiber connectivity, and OCS. TPU 8t retains the 3D Torus architecture while adding the new Virgo scale-out network, which we estimate could increase the optical attach rate from approximately 1.5 to 2.0 per TPU, a 33% increase versus TPU v7. TPU 8i adopts the Boardfly topology, reducing maximum network diameter from 16 to 7 hops, with an estimated attach rate of 1.25 optics per TPU. As TPU deployments scale, demand for optical transceivers, multi-fiber push-on (MPO)/ high-density fiber connectivity, and OCS solutions is expected to expand further. Key suppliers within the TPU optical interconnect ecosystem are Lumentum (US), the leading supplier of OCS; EZconn (6442 TT, NT\$1,490, NR), which provides optical fiber and cabling infrastructure; Landmark Optoelectronics (3081 TT, NT\$2,695, NR), LuxNet (4979 TT, NT\$569, NR), Truelight (3234 TT, NT\$173, NR) and Elite Advanced Laser (3450 TT, NT\$523, NR), which supply laser diode-related components; and photodiode (PD) vendors Visual Photonics Epitaxy (2455 TT, NT\$552, NR) and GCS Holdings (4991 TT, NT\$488, NR).

Stocks for Action

We expect 2027 to be a year of significant volume ramp-up for Google TPU deployments. From a chip perspective, we estimate shipments for TPU v8 to increase from approximately 4.02mn units in 2026F to 9.10mn units in 2027F, representing 126% YoY growth. Of the total, we estimate TPU 8t (chip developed by MediaTek (2454 TT, NT\$4,920, OP)), will account for around 3.6mn units next year, while the TPU 8i chip, developed by Broadcom, is projected to reach around 5.5mn units in 2027F. Meanwhile, total TPU v8 rack shipments look to more than double in 2027F, driving demand across the TPU ecosystem, including AI accelerators, ODM assembly, thermal solutions, power

systems, PCB/ CCL materials, and optical networking infrastructure. Key beneficiaries include MediaTek, the TPU 8t ASIC design partner, for whom TPU-related business could account for around 40% of 2027F revenue; Broadcom, the TPU v8i ASIC design partner, with TPU-related revenue contribution estimated at around 45%; Celestica, the primary L10/ L11 system integrator, with TPU-related revenue contribution estimated at 25-30%; Delta Electronics, supplying power shelves, PSU, and $\pm 400V$ HVDC solutions, with an estimated contribution of 20-25%; EMC, a key CCL supplier (with market share at 70%), at 20-25%; PCB maker Gold Circuit Electronics, at 15-20%; AVC, an emerging liquid-cooling cold plate supplier, at 10-15%; and HVLP4 copper foil supplier Co-Tech Development, at 5-10%. Other suppliers with more limited TPU-related revenue exposure (below 10%) include Hon Hai Precision in TPU tray and L11 rack-level ODM, Inventec in L6 assembly, Jentech Precision (3653 TT, NT\$6,800, OP) in stiffeners, Fositek (6805 TT, NT\$2,455, OP) in liquid-cooling QD components, Unimicron (3037 TT, NT\$1,175, OP) in ABF substrate; and Lite-On Technology in PSU solutions. Among all beneficiary groups, optical networking suppliers are likely to be the largest beneficiaries of the TPU v8 deployment ramp with 2027F sales contribution of double digit. Key names are Landmark Optoelectronics in laser diodes, EZconn in optical fiber, cabling, and scale-out interconnect solutions, and Lumentum in OCS.

Risks

Chip supply availability; assembly and component bottlenecks.

TPU v8 chip design roadmap & architecture evolution

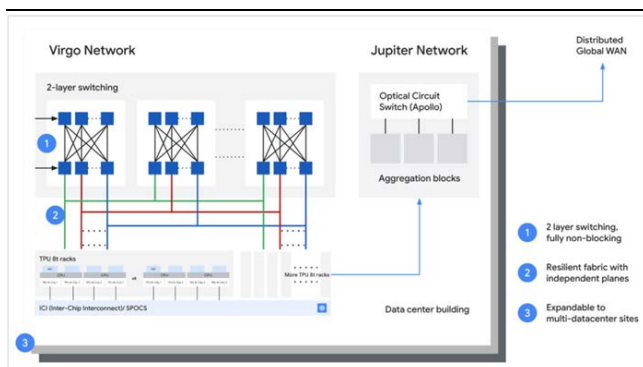
Google's TPU platform has now entered its eighth generation. Unlike the previous Ironwood generation (TPU v7), which handled both training and inference workloads on a unified architecture, TPU v8 marks the first major architectural bifurcation, splitting into TPU 8t (Zebrafish) and TPU 8i (Sunfish), each optimized for distinct workload requirements. TPU 8t (training) is designed for large-scale model pre-training and embedding heavy workloads. It retains Google's proven 3D Torus scale-up architecture, allowing a single Superpod to scale to as many as 9,600 TPU chips. Key enhancements include native FP4 support, the addition of SparseCore acceleration, and the introduction of the new Virgo scale-out network. In contrast, TPU 8i (inference) is specifically optimized for post-training workloads, including model serving, sampling, real-time inference, and reasoning applications. The architecture incorporates 384MB of on-chip SRAM, a dedicated Collectives Acceleration Engine (CAE), and an entirely new Boardfly ICI topology. A single TPU 8i pod can support up to 1,152 TPU chips, with a maximum of 1,024 active TPUs available for inference workloads. Google officially unveiled both TPU 8t and TPU 8i at Cloud Next in April this year, with cloud availability scheduled to begin progressively throughout 2026. This transition marks a fundamental shift in Google's TPU strategy, moving from a unified architecture focused primarily on compute scaling to distinct training and inference platforms tailored for workload-specific optimization.

Figure 1: TPU v8 series is segmented into training (v8t) & inference (v8i) to address the memory wall constraint

Year	2023	2023	2024	2025	2026	2026	2027F	2028F	2028F	2029F	2029F
	v5e	v5p	v6e	v7x	v8t	v8i	v8px	v9	v9	v10	v10
Code name	Viperlite	Viperfish	Ghostlite	Ironwood	Zebrafish (A5921)	Sunfish	Whalefish (Upgraded Sunfish)	Humufish	Triggerfish (A5922)	Merope	Icelfish
ASIC partner	Broadcom	Broadcom	Broadcom	Broadcom	MediaTek	Broadcom	Broadcom	MediaTek	MediaTek	Broadcom/Marvell?	MediaTek
Process node	TSMC N5	TSMC N5	TSMC N4	TSMC N3	TSMC N3	TSMC N3	TSMC N3	TSMC N3/N2	TSMC N3/N2	TSMC N2	TSMC N2
Memory	HBM2E	HBM3	HBM3	HBM3E	HBM3E	HBM3E	HBM4E	HBM4	HBM4E	HBM4E/HBM5?	HBM4E/HBM5?
Memory capacity	16GB	95GB	32GB	192GB	216GB	288GB	576GB?	288GB	?	?	768GB?
2.5D packaging	CoWoS-S	CoWoS-S	CoWoS-S	CoWoS-S	CoWoS-S	CoWoS-S	CoWoS-S	EMIB-T	EMIB-T	CoWoS-L	EMIB-T
Hybrid bonding/ SoIC	N	N	N	N	N	N	N	N	N	Y?	?

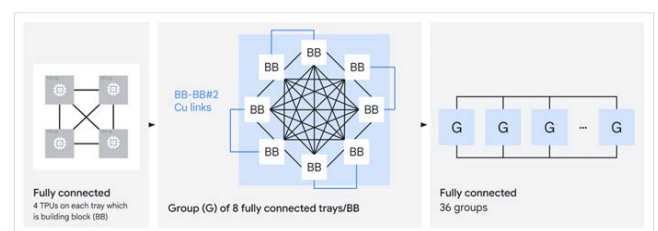
Source: KGI Research

Figure 2: Rack-level connectivity architecture of TPU 8t to the Virgo Fabric Network



Source: Google; KGI Research

Figure 3: TPU 8i adopts a hierarchical Boardfly topology, built from fully-connected four-chip units, which form fully-connected eight-board clusters; thirty-six such clusters are then fully interconnected to create a TPU 8i pod



Source: Google; KGI Research

Figure 4: TPU 8t targets massive-scale training

TPU 8t: The Training Powerhouse

Purpose-built for trillion-parameter scale.



Google

13

Source: Google; KGI Research

Figure 5: TPU 8i targets low-latency real-time inference

TPU 8i: The Low-latency Inference Powerhouse

Zero-wait silicon optimized for reasoning and RL sampling



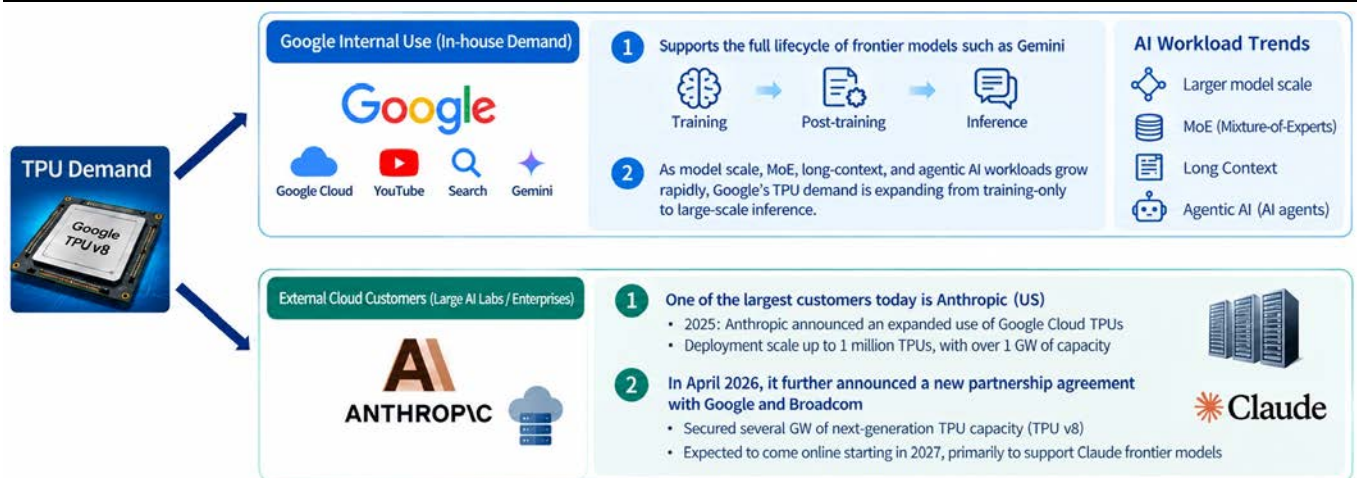
Google

8

Source: Google; KGI Research

Google & external clients drive demand for TPU

TPU demand is comprised of Google's internal consumption and demand from its external cloud customers. Internal demand is driven by Google DeepMind and Google's AI services, including the training, post-training, and inference workloads of frontier models such as Gemini. As model sizes continue to expand and workloads become increasingly complex amid the adoption of Mixture of Experts (MoE) architectures, processing of longer contexts and agentic AI applications, Google's demand for TPUs has expanded beyond training into large-scale inference. On the external front, one of the most important and largely publicly disclosed clients is Anthropic (US). In 2025, Anthropic announced plans to expand its use of Google Cloud TPUs to as many as 1mn units, representing more than 1GW of compute capacity, with deployments scheduled to come online throughout 2026F. In April 2026, Anthropic further announced a new agreement with Google and Broadcom to secure multiple gigawatts of next-generation TPU capacity based on TPU v8, which is expected to begin deployment in 2027F and will primarily support the frontier models of Claude. As a result, TPU demand is evolving from a structure that was dominated by Google's internal consumption toward a dual-engine model driven by both Google's internal use and demand from large external AI labs. This transition is also making TPUs increasingly commercial as an AI accelerator.

Figure 6: Demand for TPU stems primarily from Google & its external clients


Source: KGI Research

Main differences between TPU v8 & TPU v7

The seventh-generation (v7) Ironwood TPU adopts a unified architecture for both training and inference. Each TPU is equipped with 192GB of HBM with a bandwidth of 7.38TB/s, while large scale-up connectivity is enabled through the ICI 3D Torus architecture. In contrast, Google has split workload optimization into two distinct products in TPU v8, namely TPU 8t and TPU 8i. The TPU 8t chip is designed by MediaTek for large-scale, compute-intensive training workloads, with a focus on delivering higher compute throughput and stronger bandwidth scalability. The accelerator features 216GB of HBM, 6.53TB/s of HBM bandwidth, 128MB of SRAM and 12.6 PFLOPS of FP4 performance. It also incorporates native FP4, SparseCore, and Virgo Network architecture. TPU 8i features higher HBM bandwidth and the chip is designed by Broadcom for processing latency-sensitive inference workloads with greater emphasis on memory capacity and communication latency. HBM capacity is increased to 288GB and bandwidth to 8.60TB/s, with a larger SRAM capacity of 384MB. The architecture incorporates Collectives Acceleration Engine (CAE) and Boardfly topology to deliver 10.1 PFLOPS of FP4 performance. Compared to Ironwood (TPU v7), TPU 8t delivers up to 2.7x higher performance-per-dollar for large-scale training workloads, while TPU 8i provides up to 80% higher performance-per-dollar for low-latency inference of large MoE models. Both products improve performance-per-watt by up to 2x. Therefore, the most significant transition from TPU v7 to TPU v8 is not merely an increase in FLOPS. Rather, the key advancement lies in the simultaneous optimization of compute, memory, networking and workload specialization, resulting in substantially superior overall system efficiency.

Figure 7: TPU v8 boasts superior overall system efficiency through simultaneous optimization of compute, memory, networking & workload specialization

	TPU v7 Ironwood	TPU v8t	TPU v8i
Chip design partner	Broadcom	MediaTek	Broadcom
Primary workload	Large-scale training, inference & serving	Large-scale pre-training	Post-training & inference
Network topology	3D Torus	3D Torus	Boardfly
HBM capacity per chip	192 GB	216 GB	288 GB
On-chip SRAM (VMEM)	TBA	128 MB	384 MB
Peak FLOPS per chip	TBA	12.6 PFLOPS	10.1 PFLOPS
HBM bandwidth per chip	7.38 TB/s	6.528 TB/s	8.601 TB/s
HBM generation	HBM3E	TBA	TBA
Host cpu / head node	TBA	ARM-based Google axion	ARM-based Google axion
Single pod / rack	9,216 TPUs	9,600 TPUs	1,152 TPUs (1,024 online)
Total HBM per pod / rack	1.77 PB	~2 PB	331.8 TB
Power consumption per chip	1000w	<800w	1300w
Cooling	Liquid cooling + air cooling	Liquid cooling	Liquid cooling
Substrate	70~80 x 70~80 mm/ 16-18L	80~90 x 80~90 mm/ 18-20L	80~90 x 80~90 mm/ 18-20L
CCL/ PCB	34L/ M7 ; HVLP2	34L-40L/M8+ ; HVLP4	34-40L/M8+ ; HVLP4
Optical transceiver module	800G (Scale-up)	1.6T (Scale-up)	1.6T (Scale-up)
Power	5.5kW	10kW	10kW

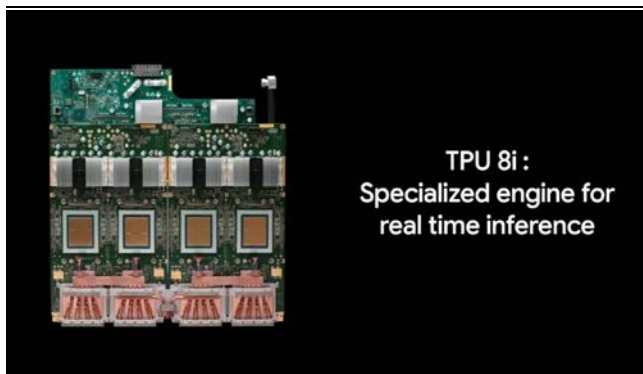
Source: KGI Research

TPU v8 rack design largely built upon that of v7

With TPU v7 Ironwood, the basic compute building block consists of a 4-TPU physical host. Each rack adopts a symmetric right-left design and contains 64 TPUs, equivalent to sixteen 4-TPU hosts. Each layer comprises four TPUs and four Axion CPUs, forming a compute unit, with eight units deployed on each side of the rack. The 64 TPUs within a rack are interconnected through high-speed ICI links and organized into a 3D Torus Cube architecture. When computing requirements exceed the capacity of a single rack, Google leverages OCS to connect multiple 64-chip cubes. For example, four cubes can be combined into a 256-chip pod, while the largest Ironwood SuperPod consists of 144 cubes, each carrying 64 TPUs, amounting to 9,216 TPUs in total. This architecture enables thousands of chips to jointly execute large-scale training and inference workloads. Beyond scale-up connectivity, OCS also enhances system resilience by dynamically rerouting optical pathways around failed cubes or network links, improving overall cluster availability.

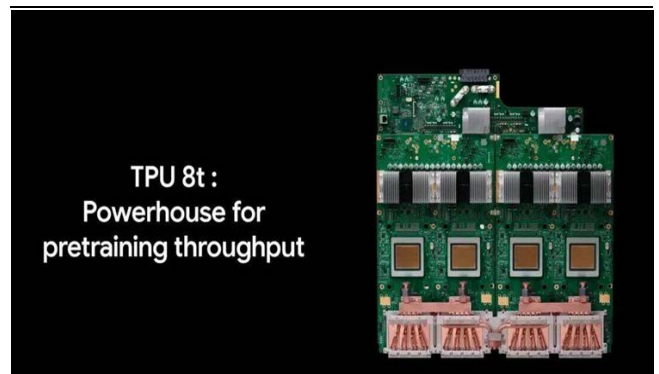
TPU v8 introduces a clear architectural divergence between 8t and 8i, and the rack architecture of 8t and 8i is not entirely the same. TPU 8t retains the 3D Torus architecture optimized for large-scale training workloads, but doubles ICI bandwidth from the previous generation and introduces the Virgo scale-out network. A single Virgo fabric can connect more than 134,000 TPU 8t chips and deliver up to 47Pb/s of non-blocking bisection bandwidth. Combined with JAX and Pathways, the architecture can support logical training clusters consisting of more than 1mn TPUs. In comparison, TPU 8i adopts a new Boardfly topology. Each tray contains four TPUs as a basic building block, while eight trays form a group of 32 chips. A total of 36 groups are then interconnected through OCS, forming a pod containing up to 1,152 physical TPUs and 1,024 active TPUs. This architecture reduces the maximum network diameter of a 1,024-chip system from 16 hops under a 3D Torus architecture to only 7 hops, significantly lowering all-to-all communication latency for MoE and reasoning workloads. According to current supply chain checks, the L6 board in TPU v8 integrates not only the TPU itself, but also power modules and high-speed transceivers/ optics, while adopting Google's fourth-generation (Gen4) liquid-cooling solution. As a result, the key upgrade within TPU v8 extends far beyond improvement in TPU performance. The architecture represents a system-level redesign spanning compute, memory, networking, power delivery, thermal management and optical communications. At present, Celestica remains the primary L10/ L11 system assembler, while we expect Hon Hai to gain market share in TPU tray and L11 rack assembly for TPU v8, following assembly of the CPU tray and rack in v7, while Inventec, currently a TPU L6 supplier, may also expand into system-level integration.

Figure 8: Overview of a TPU v8i tray

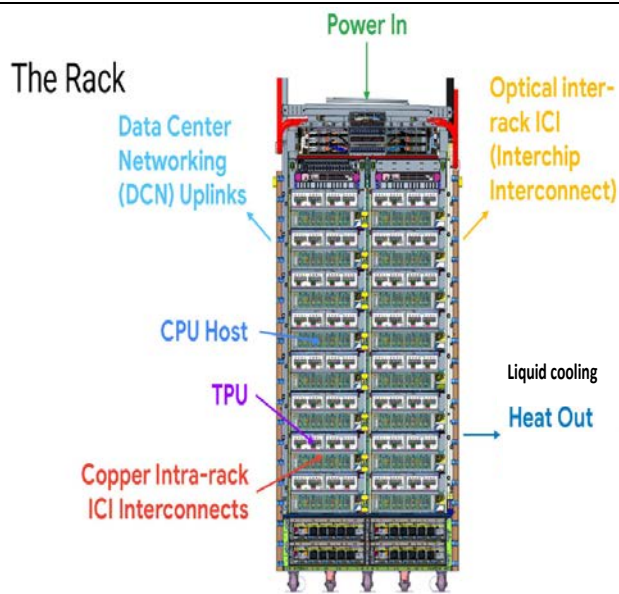


Source: Company data; KGI Research

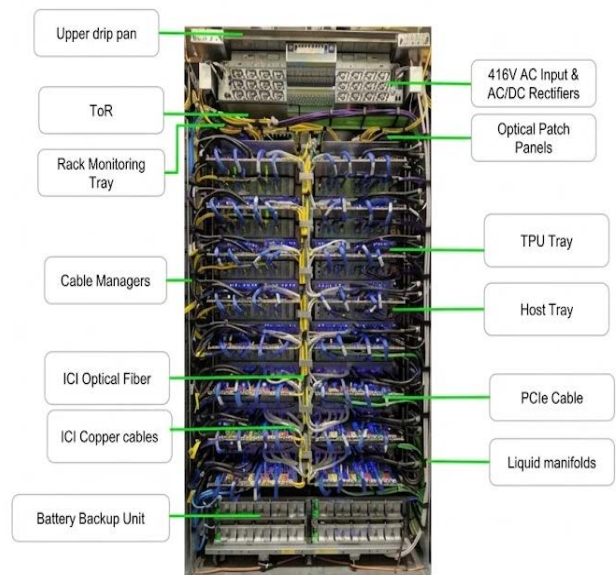
Figure 9: Overview of a TPU v8t tray



Source: Company data; KGI Research

Figure 10: Overview of a TPU v7 rack


Source: Company data; KGI Research

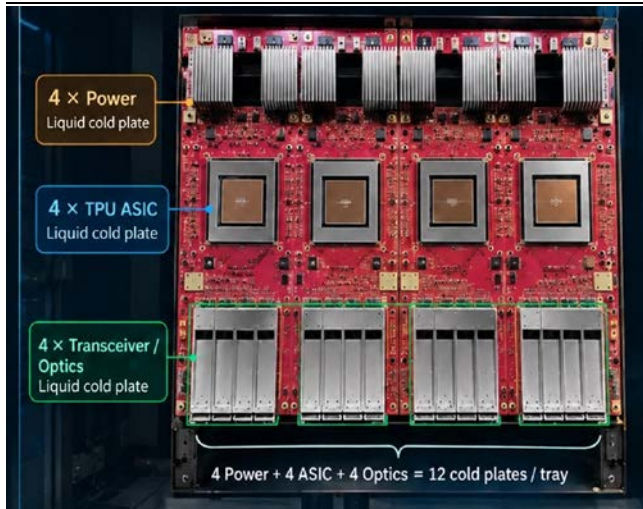
Figure 11: With TPU v7, the basic compute building block consists of a 4-TPU physical host; a rack adopts a symmetrical right-left design and contains 64 TPUs


Source: Company data; KGI Research

TPU v8 thermal solutions evolve from chip-level to system-level liquid cooling

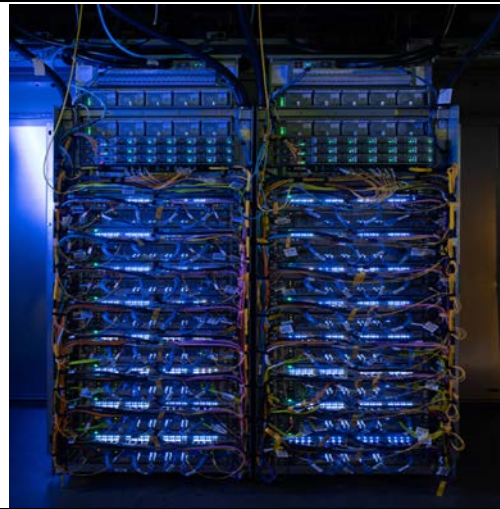
Thermal architecture is one of the most significant changes in the transition from TPU v7 Ironwood to TPU v8. Ironwood adopts Google's third-generation (Gen3) liquid cooling system, in which a CDU delivers coolant to the rack, while manifolds, tubing and cold plates directly remove heat generated by high-power chips. Google has also publicly showcased the Gen3 CDU used in Ironwood SuperPods. With TPU v8, Google has upgraded to fourth-generation (Gen4) liquid cooling, primarily because the compute density and power consumption of both 8t and 8i configurations continue to increase. As a result, thermal requirements are no longer concentrated solely on the TPU chip, but are gradually expanding to encompass the entire high-density computing system. Based on the current supply chain breakdown, each TPU v8 layer requires 12 cold plates, including cooling for four TPUs, four power modules and four transceiver modules. Assuming a rack contains 16 layers, a single rack would require approximately 192 cold plates. Given the higher number of cold plates per layer and the inclusion of QDs, we estimate liquid cooling content value at US\$2,500-3,000 per layer, or US\$45-50k per rack. These estimates do not include CDU content. Google has simultaneously proposed specifications for its next-generation Project Deschutes CDU platform (Gen5), featuring cooling capacity of approximately 2MW, flow rates of 500 GPM, and operating pressure of 80-90 psi, effectively reserving thermal headroom for future AI racks with even higher power density. Given the architectural differences between TPU v8t and v8i, particularly in chip and memory design, the liquid cooling supply chains also differ. For the v8t chip, which is designed by MediaTek, the primary cold plate suppliers are AVC and Furukawa Electric. For the v8i chip, designed by Broadcom, the primary cold plate suppliers are Cooler Master and Boyd. According to market information, TPU v7 racks use 1.4MW CDUs, while TPU v8t racks adopt CDUs of approximately 2MW. A typical deployment architecture pairs one CDU with roughly eight racks. Although CDU capacity should not be interpreted as actual rack-level power consumption, the increase from 1.4MW to 2MW indicates that Google is reserving much greater thermal capacity for next-generation TPU platforms. In addition, the scope of liquid cooling in TPU v8 is expanding beyond cold plates to include manifolds, QD, tubing, rack distribution systems and optical-module cooling.

Figure 12: Each tray of a TPU v8 rack will come with twelve cold plates



Source: Company data; KGI Research

Figure 13: TPU v8 racks to adopt a similar design of sixteen trays & liquid cooling in Gen4 CDU



Source: Company data; KGI Research

Figure 14: TPU v8 has upgraded to fourth-generation (Gen4) liquid cooling CDU

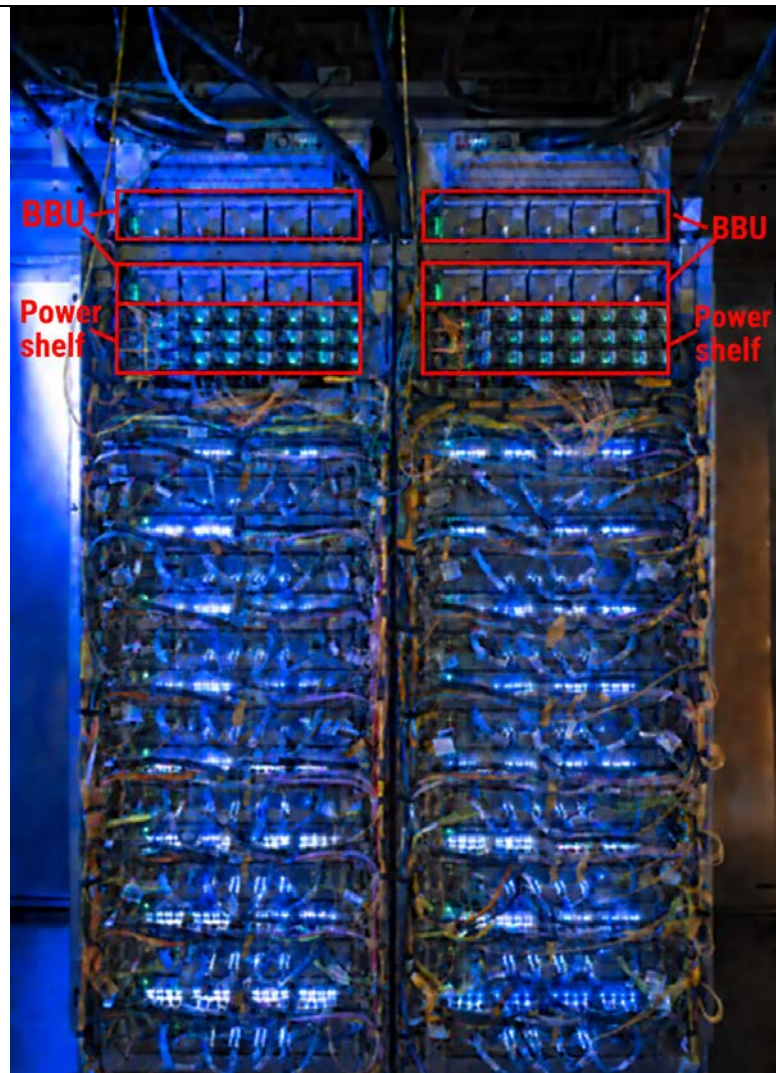


Source: KGI Research

Initial HVDC adoption in TPU v8

As TPU thermal design power (TDP) continues to increase amid model iteration, power system specifications have also advanced in TPU v8. The most notable change is the introduction of HVDC power racks that support $\pm 400\text{V}$ DC architecture, although we believe the existing 50V DC architecture will remain the mainstream spec for this generation. Under 50V DC architecture, we expect each TPU 8i rack to incorporate three power shelves, each containing six 10kW PSUs. By comparison, a TPU v7 rack utilizes six 22kW power shelves, each consisting of four 5.5kW PSUs. Based on our estimates, PSU content value per rack will increase from approximately US\$20k in the v7 platform to US\$28-30k in the TPU 8i platform, representing 40-50% growth. In addition, BBUs and DC-DC power modules are expected to remain core components of the TPU rack power architecture. We believe Delta Electronics is the leading supplier of $\pm 400\text{V}$ DC power racks, with shipments already underway and the firm having captured a high market share in the power shelf segment. Flex is the second supplier, while Lite-On Technology is also believed to have successfully entered the supply chain this year.

Figure 15: TPU power architecture continues to evolve with each generation, with TPU v8 marking the initial adoption of HVDC

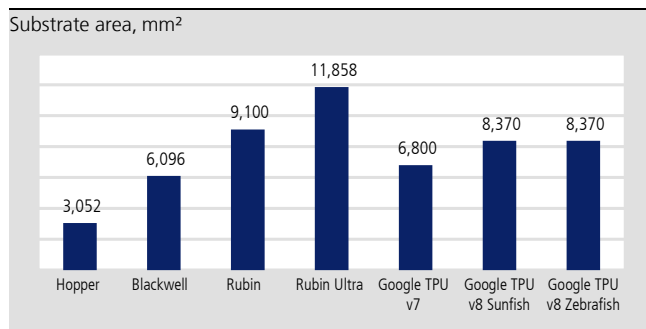


Source: KGI Research

Larger TPU v8 substrates drive both content & shipments growth for PCB supply chain

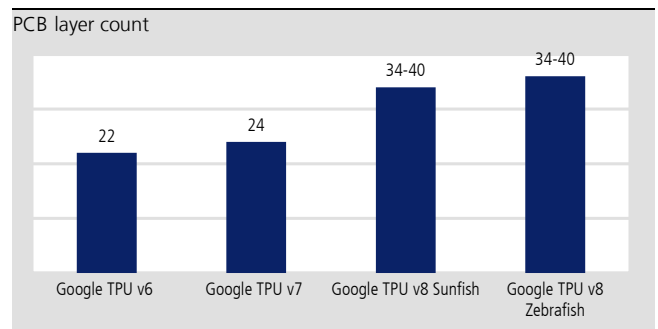
Compared with TPU v7 Ironwood's approximately 80-85mm, 16-layer ABF substrate, TPU v8 adopts a larger approximately 90-100mm substrate to accommodate additional memory, while layer count increases to 18-22 layers. We estimate these upgrades will drive a 40-50% increase in ABF substrate ASP. On the PCB side, content growth is supported by both higher layer counts and material upgrades. The next-generation v8t and v8i motherboards are expected to adopt 34-40-layer PCB designs. CCL specifications are upgraded from M7 to M8+, and copper foil transitions from HVLP2 to HVLP4. These enhancements are expected to increase CCL and PCB ASP by 70-80%. We believe EMC, which commands an up to 70% market share in CCL, will be a key beneficiary of this trend. Other beneficiaries include Gold Circuit Electronics (20-25% PCB market share), a new entrant to the supply chain, ABF substrate supplier Unimicron (3037 TT, NT\$1,175, OP), and Co-Tech Development.

Figure 16: TPU v8 uses larger ABF substrates



Source: Company data; KGI estimates

Figure 17: TPU v8 migrates to higher-layer-count & multi-layer HDI PCBs



Source: Company data; KGI estimates

Scale-up – Optical content per TPU continues to increase

TPU v7 adopted a unified architecture for both training and inference based on a 3D Torus topology. Each cube contains 64 ASICs arranged in a 4x4x4 configuration, corresponding to a single rack, and connected externally through 96 optical fibers. A full pod consists of 144 cubes, equivalent to 9,216 TPUs, requiring 13,824 optical fibers, 13,824 optical transceivers, and 48 OCS. This translates into an average optical attach rate of 1.5 fibers and 1.5 transceivers per TPU.

With TPU v8, Google separates training and inference workloads, with each platform adopting a distinct network topology. For the inference-oriented v8i, Google retains the configuration of four ASICs per tray but replaces the 3D Torus with a Boardfly topology. This reduces the maximum communication distance between any two chips from 16 hops to 7 hops, lowering latency by more than 50%.

Under this architecture, each tray utilizes five outbound optical fibers, and a pod contains 1,152 chips. Consequently, a single pod requires 1,440 optical fibers, 1,440 optical transceivers, and an estimated five OCS units, implying an optical attach rate of 1.25.

For the training-focused v8t, Google retains the 3D Torus architecture and four-ASIC-per-tray configuration. However, Google has yet to disclose the rack-level chip arrangement, which will ultimately determine how the 3D topology is partitioned.

Based on our industry checks, we believe the most likely configuration consists of 32 ASICs arranged in a 4x4x2 topology per rack-level interconnect unit. Under this

assumption, each rack would require 64 outbound optical fibers. Assuming a pod contains 9,600 TPUs, total deployment would comprise 300 racks, 19,200 optical fibers, 19,200 optical transceivers, and 64 OCS units. This implies an optical attach rate of 2.0, or 33% higher than TPU v7.

Details regarding the scale-out topology are limited. Google has yet to disclose the interconnect architecture for v8i. For v8t, the firm has introduced Virgo network fabric alongside its existing Jupiter networking architecture, reducing the switching hierarchy from three layers to two. Overall, continued growth in TPU deployments should drive corresponding increases in demand for optical transceivers, fiber-optic cables, and fiber distribution systems across the scale-out network.

Figure 18: TPU pod topologies

Item	TPU v7 (Ironwood)	TPU v8i (Inference)	TPU v8t (Training)
Scale-up topology			
Topology	3D Torus	Boardfly	3D Torus
Chips per tray	4	4	4
Unit geometry	4×4×4	—	4×4×2/ 4×4×4*
Chips per unit	64	4	32/ 64
External fibers per unit	96	5	64/ 96*
Pod scale			
Physical chips per pod	9,216	1,152	9,600
Usable chips	9,216	1,024	9,600
Spare chips	0	128	0
Optical content per pod			
Fibers	13,824	1,440	19,200/ 14,400*
Optical transceivers	13,824	1,440	19,200/ 14,400*
OCS units	48	5 (est.)	64/ 48*
Assumed ports per OCS	300	300	300
Optical attach rate			
Fibers per TPU	1.50	1.25	2.00/1.50
Transceivers per TPU	1.50	1.25	2.00/1.50
Attach rate vs. v7 (%)	-	-16.7	33.3 /--

* Note: Based on our industry checks, two topology architectures are currently under evaluation for TPU v8t; the final topology configuration has yet to be determined.

Source: KGI Research

Low latency & low power consumption drive Google's extensive adoption of OCS

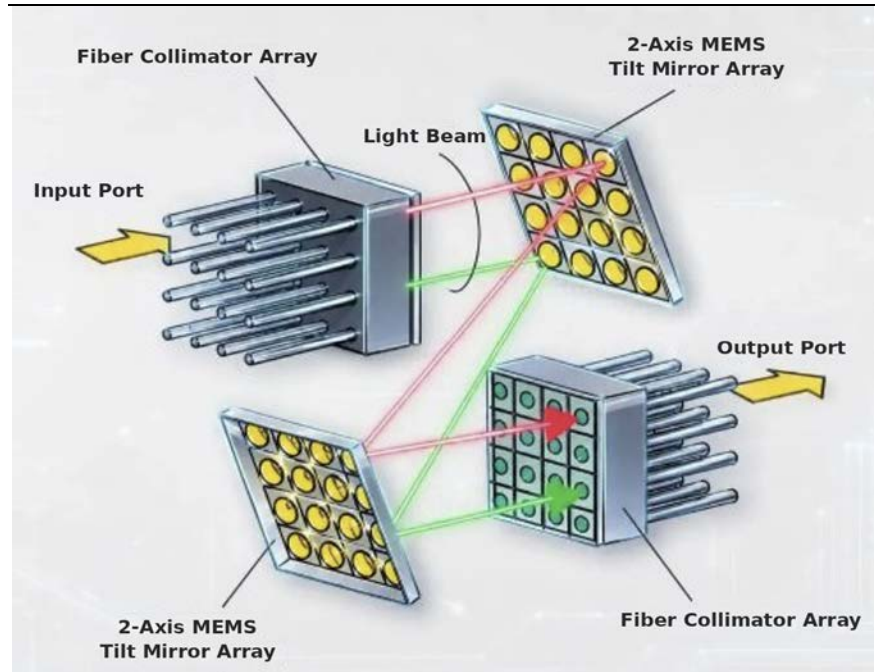
Google deploys OCS because they enable direct optical switching between fiber ports without requiring optical-electrical-optical (OEO) conversion. Key advantages are: (1) lower transmission latency; (2) compatibility with optical transceiver upgrades from 400G/ 800G to 1.6T without requiring replacement of the OCS hardware; and (3) the elimination of SerDes and DSP components, resulting in significantly lower per-port power consumption than conventional electrical switches.

Google's OCS architecture is based on MEMS technology and primarily comprises an input fiber array, a collimating lens array, two opposing 2D MEMS micromirror arrays, an output lens array, and an output fiber array. The MEMS device is fabricated using semiconductor processes that etch dual-axis steerable micromirrors onto silicon wafers. By applying voltage to adjust the mirror angles, optical beams can be redirected dynamically, enabling switching from any input port to a designated output port.

We believe Google is pursuing a dual-track OCS strategy, combining in-house development with solutions supplied by Lumentum, while Taiwanese vendors have relatively limited exposure to the OCS hardware itself.

Nevertheless, we expect rising TPU unit shipments to drive demand for optical transceivers, high-density fiber arrays, MPO cabling, and fiber management products. Key beneficiaries include Taiwan-based optical connectivity supplier EZconn, whose exposure is primarily concentrated in scale-out networking infrastructure, as well as optical transceiver component suppliers like Visual Photonics Epitaxy and GCS Holdings on the PD side, and Landmark Optoelectronics, LuxNet, TrueLight, and Elite Advanced Laser on the LD side.

Figure 19: Optical path architecture of a MEMS-based OCS



Source: Lumentum; KGI Research

Figure 20: Lumentum R300 OCS

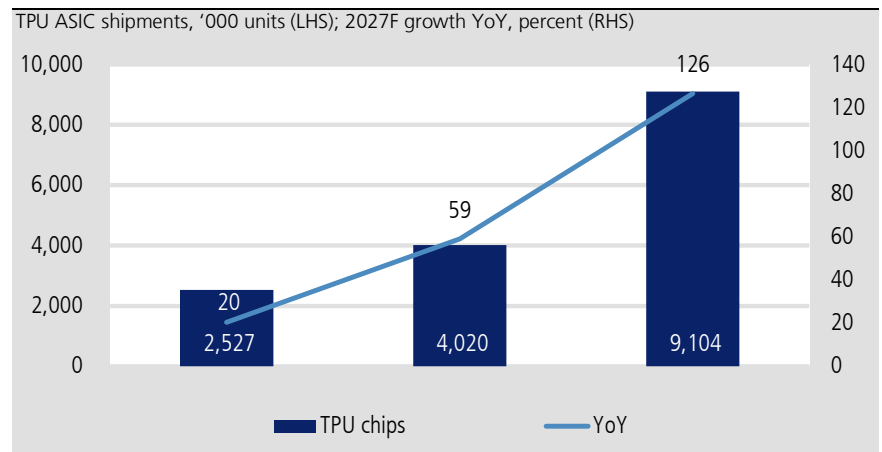


Source: Lumentum; KGI Research

TPU chip shipments to double in 2027F, driving a comparable increase in rack deployments

We expect 2027 to be a year of significant volume ramp-up for Google TPU deployments. From a chip perspective, we project shipments for TPU v8 to increase from approximately 4.02mn units in 2026F to 9.10mn units in 2027F, representing 126% YoY growth. Of the total, the TPU v8t chip (MediaTek) is expected to account for approximately 3.6mn units, while the TPU v8i chip (Broadcom) is projected to reach around 5.5mn units. Meanwhile, total TPU v8 rack shipments look to double in 2027F, driving demand across the TPU ecosystem, including TPU chips, ODM assembly, thermal solutions, power systems, PCB/CCL materials, and optical networking infrastructure. Key beneficiaries include MediaTek, the TPU v8t ASIC design partner, for whom TPU-related business could account for 40% of 2027F revenue, and Broadcom, the TPU v8i ASIC design partner, with TPU-related revenue contribution estimated at around 45%; Celestica, the primary L10/ L11 system integrator, with TPU-related revenue contribution estimated at 25-30%; Delta Electronics, supplier of power shelves, PSUs, and $\pm 400V$ HVDC solutions, with an estimated contribution of 20-25%; EMC, a key CCL supplier, at 20-25%; Gold Circuit Electronics, a leading PCB vendor, at 15-20%; AVC, an emerging liquid-cooling cold plate supplier, at 10-15%; and Co-Tech Development, a supplier of HVLP4 copper foil, at 5-10%. Other suppliers with more limited TPU-related revenue exposure (below 10%) are Hon Hai in rack-level ODM, Inventec in L6 assembly, Jentech Precision in stiffeners, Fositek in liquid-cooling QD components, Unimicron in ABF substrate and Lite-On Technology in PSU solutions. Among all beneficiary groups, optical networking suppliers are likely to be the largest beneficiaries of the TPU v8 deployment ramp with 2027F sales contribution of double digit. Key names include Landmark Optoelectronics in LD, EZconn in optical fibers, cabling, and scale-out interconnect solutions, and Lumentum in OCS.

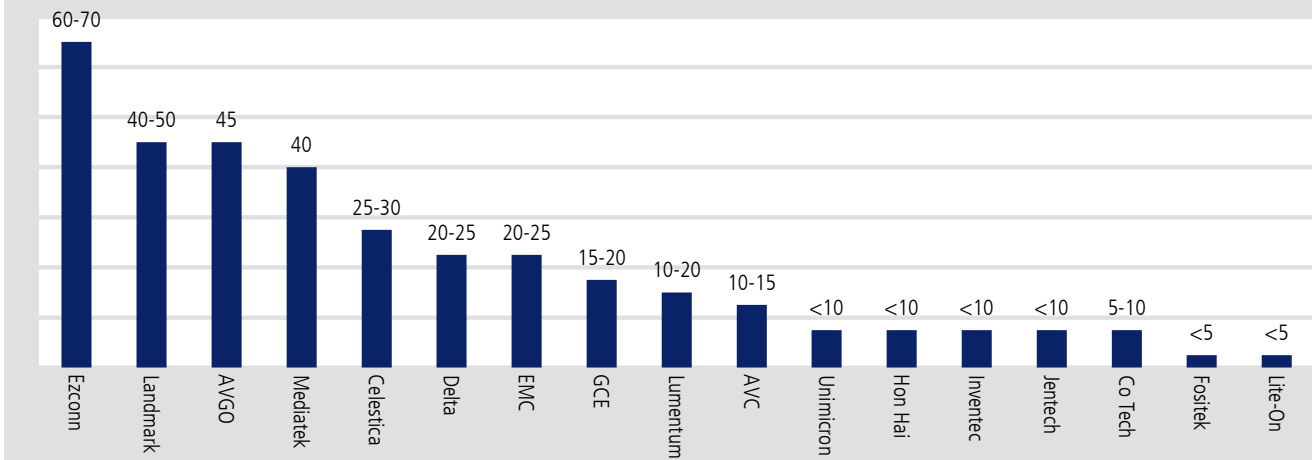
Figure 21: TPU ASIC shipments projected to double in 2027F



Source: KGI estimates

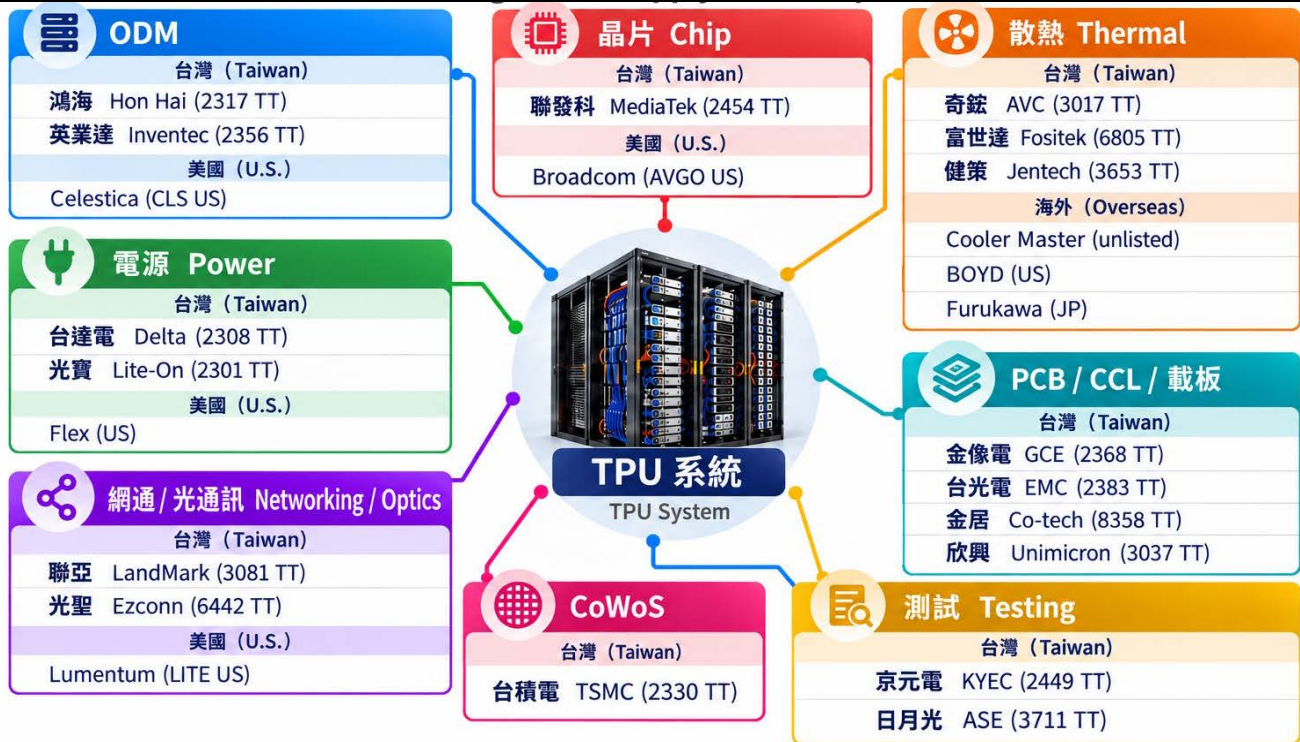
Figure 22: TPU-related to contribute a high proportion of 2027F revenue for EZconn, Landmark, Broadcom, MediaTek, Celestica, Delta, EMC, Gold Circuit, Lumentum & AVC

2027F TPU-related revenue weighting, percent



Source: KGI estimates

Figure 23: Google TPU supply chain key components & Taiwanese players



Source: KGI estimates

KGI Locations

China	Shanghai	Room 1507, Park Place, 1601 Nanjing West Road, Jingan District, Shanghai, PRC 200040
Taiwan	Taipei	700 Mingshui Road, Taipei, Taiwan Telephone 886.2.2181.8888 Facsimile 886.2.8501.1691
Hong Kong		41/F Central Plaza, 18 Harbour Road, Wanchai, Hong Kong Telephone 852.2878.6888 Facsimile 852.2878.6800
Thailand	Bangkok	8th - 11th floors, Asia Centre Building 173 South Sathorn Road, Bangkok 10120, Thailand Telephone 66.2658.8888 Facsimile 66.2658.8014
Singapore		4 Shenton Way #13-01 SGX Centre 2 Singapore 068807 Telephone 65.6202.1188 Facsimile 65.6534.4826
Indonesia		Sona Topas Tower Fl.11 Jl. Jend. Sudirman kav.26 Jakarta Selatan 12920 Indonesia Telephone 62 21 250 6337

KGI Ratings

Rating	Definition
Outperform (OP)	We take a positive view on the stock. The stock is expected to outperform the expected total return* of the KGI coverage universe in the related market over a 12-month investment horizon.
Neutral (N)	We take a neutral view on the stock. The stock is expected to perform in line with the expected total return* of the KGI coverage universe in the related market over a 12-month investment horizon.
Underperform (U)	We take a negative view on the stock. The stock is expected to underperform the expected total return* of the KGI coverage universe in the related market over a 12-month investment horizon.
Not Rated (NR)	The stock is not rated by KGI Securities.
Restricted (R)	KGI policy and/or applicable law regulations preclude certain types of communications, including an investment recommendation, during the course of KGI's engagement in an investment banking transaction and in certain other circumstances.
	<i>Total return = (12M target price - current price) / current price</i>

Disclaimer

All the above named KGI analyst(s) is SFC licensed person accredited to KGI Asia Ltd to carry on the relevant regulated activities. Each of them and/or his/her associate(s) does not have any financial interest in the respectively covered stock, issuer and/or new listing applicant.

All the information contained in this report is not intended for use by persons or entities located in or residing in jurisdictions which restrict the distribution of this information by KGI Asia Limited ("KGI") or an affiliate of KGI. Such information shall not constitute investment advice, or an offer to sell, or an invitation, solicitation or recommendation to subscribe for or invest in any securities or investment products or services nor a distribution of information for any such purpose in any jurisdiction. In particular, the information herein is not for distribution and does not constitute an offer to sell or the solicitation of any offer to buy any securities in the United States of America, or to or for the benefit of United States persons (being residents of the United States of America or partnerships or corporations organised under the laws of the United States of America or any state, territory or possession thereof). All the information contained in this report is for general information and reference purpose only without taking into account of any particular investor's objectives, financial situation or needs. Such information is not intended to provide professional advice and should not be relied upon in that regard.

Some of KGI equity research and earnings estimates are available electronically on www.kgi.com.hk. Please contact your KGI representative for information. The information and opinions in this report are those of KGI internal research activity. KGI does not make any representation or warranty, express or implied, as to the fairness, accuracy, completeness or correctness of the information and opinions contained in this report. The information and opinions contained in this report are subject to change without any notice. No person accepts any liability whatsoever for any loss however arising from any use of this report or its contents. This report is not to be construed as an invitation or offer to buy or sell securities and/or to participate in any investment activity. This report is being supplied solely for informational purposes and may not be redistributed, reproduced or published (in whole or in part) by any means for any purpose without the prior written consent of KGI. Members of the KGI group and their affiliates may provide services to any companies and affiliates of such companies mentioned herein. Members of the KGI group, their affiliates and their directors, officers and employees may from time to time have a position in any securities mentioned herein.