新光投顧

〔投資研究報告〕

2026/4/24

Google 推理專用化重估

目錄

雙軌 TPU 改寫雲利潤與供應鏈

2026/4/24 新光投顧

過去兩年，市場對 AI 基建的討論大多圍繞在同一條主線：大型雲服務商持續擴大資本支出，前沿模型推高訓練需求，最先受惠的是高階 GPU、先進封裝與高頻寬記憶體。這條主線至今仍然成立，但 2026 年開始，另一個更重要的分岔已經浮現。

Google 在 2025 年底的財報會議上已經透露，Google Cloud 年化營收規模超過 700 億美元，Cloud backlog 在第四季末達到 2,400 億美元，營業利益率升至 30.1%，同時 2026 年資本支出指引高達 1,750 億至 1,850 億美元；管理層也直接指出，這些投資會讓折舊與資料中心營運成本顯著上升。換句話說，Google 已進入一個必須更精細管理算力資產回收效率的階段。

就在這個背景下，Google Cloud Next 2026 正式把第八代 TPU 切成 TPU 8t 與 TPU 8i，前者對準大型模型訓練，後者明確對準低延遲推理與 agentic AI 工作負載，且兩款晶片都將在今年稍晚提供。這個動作的意義，在於 Google 已經開始用兩套不同的矽設計與系統路徑，分別對應訓練與推理兩種完全不同的經濟問題。

目錄

一、 產品切法改變資本回收

Google 之所以在此刻把 TPU 正式拆成 8t 與 8i，核心原因是要把 AI 基建的投資回收路徑分拆清楚。根據 Next 2026 的公開資訊，TPU 8t 可在單一 superpod 內擴展到 9,600 顆 TPU 與 2PB 共用高頻寬記憶體，主要服務高同步度、超大規模的訓練工作負載；TPU 8i 則可在單一 pod 直接連接 1,152 顆 TPU，並以更低延遲、更多片上 SRAM 與約 80% 的推理 performance-per-dollar 改善為賣點，對準大量 agent 與近即時推理。這代表 Google 內部已把 AI 基建的需求結構，從單軌訓練邏輯，推進到訓練與推理雙軌並進的邏輯。

當 Cloud capex 已高到足以影響整體自由現金流與折舊曲線時，這種切法會直接影響資本支出的品質：訓練資產承接前沿模型的集中需求，推理資產則去承接高頻、長尾、可持續收費的 serving 流量與 agent 任務。若這兩條收入池都能逐步長大，Cloud 的資本回收會變得更可預測。

但雙軌產品線也帶來新的利用率問題。專用化晶片能提高單一任務效率，代價則是機房彈性調度能力下降。8t 若更偏向訓練最佳化，8i 若更偏向低延遲推理，Google 就必須更依賴雲端控制平面去平衡閒置產能與動態需求。

Google 官方的 AI Hypercomputer 頁面與相關說明，已經把 committed use discounts、spot、Dynamic Workload Scheduler 與多加速器協同排程擺在相當核心的位置；這說明 Google 自己也知道，專用化提升了單位任務效率，同時提高了 fleet-level 管理難度。未來真正值得追的數字，是 Google Cloud 營收相對 capex 的資本密集度是否改善，以及 Cloud 營業利益率能否在折舊抬升下維持穩定。這兩個方向若同時轉好，8i 的商業意義就會明顯放大。

二、 護城河正在往編譯器與控制平面移動

市場第一眼通常會拿 TPU 8i 的推理效率去和 Rubin、Blackwell Ultra 或其他加速器比較，但真正能決定護城河寬度的，是客戶是否願意把工作負載、數據路徑與開發流程交給同一套控制平面管理。Google 目前在做的事情，是把 TPU、GPU、Inference Gateway、Quickstart、Vertex AI、Gemini 與 GKE 連成一套「垂直最佳化 AI 堆疊」。GKE Inference Gateway 與 Quickstart 的 GA 公告顯示，Google 每週對 GPU 與 TPU 進行超過 100 次基準測試，再把這些效能資料和 Gemini、Search、YouTube 背後使用的儲存、網路與軟體最佳化結合，直接提供給客戶做選型與部署決策；同一篇官方文章還指出，透過 prefix-aware routing、disaggregated serving 與相關優化，TTFT 在特定高峰工作負載下最多可改善 96%，throughput 可提升 60%，而且控制台已能直接顯示成本與 latency profiles。這些功能的商業價值很明確：Google 正在把「如何在既定 SLA 下找到最低 serving 成本」這件事，做成平台能力。

這也說明為何 TPU 的封閉性既是阻力，也是黏著來源。多雲企業不會為了單一規格表就輕易搬動整套資料流水線，真正能驅動遷移的，是更低的推理成本、更穩的延遲、更快的部署速度與更完整的管理介面。Google 目前的策略，是

一邊維持 GPU 支援，一邊讓 TPU 與控制平面綁得更緊。Alphabet 法說已提到，近 75% 的 Google Cloud 客戶已用過 vertically optimized AI stack，這些客戶平均使用的產品數是其他客戶的 1.8 倍；同時，生成式 AI 模型相關產品收入在第四季年增近 400%，Gemini Enterprise 的付費席位超過 800 萬，服務超過 2,800 家公司。這些數字共同指向一個現象：Google 的護城河有一大部分正從晶片本身，移向軟硬整合後形成的使用黏著與多產品穿透。未來競爭若持續升溫，真正能守住估值的，會是控制平面與平台滲透率，單次硬體參數已無關勝負。

三、 估值焦點將從 FLOPS 轉向 Revenue per Watt

當推理進入專用化時代，雲端估值模型也需要改寫。過去市場較容易用訓練晶片出貨、GPU 供需與 capex 規模去估算 AI 基建前景，接下來更應該重視系統吞吐、功耗效率與可計費推理流量之間的關係。

Google 在 AI Hypercomputer 頁面講得很直白：推理需求正在變得更長上下文、更複雜推理、更常使用混合專家模型，因此真正該關注的是總體系統運作成本與回覆實用性，而不是處理器本身價格。這種語言的轉變，背後反映的是一套新的利潤模型。當同樣的電力、同樣的機房容量、同樣的 capex 可以產出更多 token、更多 agent 任務與更高品質的回應，雲平台就有機會拿到更高的估值溢價。這也是為什麼未來幾季最值得追的指標，應該包括 Cloud margin 是否能守住約 30%、AI 平台收入與 agent 調用量增速是否快於純基建租賃收入，以及 Google 是否能逐步證明 revenue per watt 正在上升。這些指標比單一晶片參數更接近真實的獲利能力。

這種估值轉軌也會重新排序台美供應鏈。若用配置思維去看，建議把受惠環節分成四組。

第一組是 ASIC 生態與設計服務，因為 TPU 8i 代表推理專用化正在提高客製化加速器的商業正當性，設計服務、NRE 與權利金的長期價值都會提高。

第二組是網通與光互連，因為低延遲推理與 agent 併發讓交換器、1.6T 光模組、LPO/CPO 與 cluster fabric 的重要性上升。

第三組是能源、散熱與高功率配電，因為 Google 的 capex 中約四成投向資料中心與網通設備，而 2026 年決定產能上限的，越來越常是電網、站點與冷卻系統。

第四組才是先進封裝與 HBM。這一組當然仍然重要，尤其對 8t 這類訓練叢集而言屬於剛性需求；只是市場對其重要性的認知已相對成熟，新增 alpha 更可能出現在前面三組的權重上升。若後續出現更多「非 NVIDIA 的 HPC 先進製程占比提升」訊號，這條重排邏輯會更快被市場接受。

四、 接下來最該盯的四個數字

這個主題後續最值得盯的，是四組能直接驗證獲利與估值邏輯的數字。

第一，看 Google Cloud 營業利益率能否守住約 30% 水準。若折舊、電力與資料中心營運成本都在上升，Cloud margin 仍保持韌性，代表 8i 與整體推理專用化已開始改善單位經濟。

第二，看 AI 收入組合，尤其是 Gemini Enterprise、Vertex AI 與 agent 平台的調用量，是否持續快於純算力租賃成長。

第三，看能否找到 revenue per watt 的替代訊號，例如在 capex 顯著上升的背景下，Cloud backlog、AI 相關產品收入與 Cloud 利潤率是否同步上行。

第四，看供應鏈中的 ASIC、交換器、先進封裝與高功率電源訂單，是否呈現比單純訓練 GPU 鏈更高的邊際改善。只要這些數字開始共振，Google 8t / 8i 的意義就會從產品發布，正式走到估值重估。

結論

Google 把 TPU 正式拆成 8t 與 8i，真正改變的是 AI 基建的獲利模型。接下來的競爭焦點會更偏向系統吞吐、功耗效率、延遲控制與平台黏著度。這個趨勢若順利展開，Google Cloud 的估值語言會逐步從單純的基礎設施成長，轉向更高品質的平台型基建；供應鏈的受惠地圖也會從 GPU 單線敘事，擴散到 ASIC、生態軟體、交換器、光互連、先進封裝與電力散熱。真正需要持續驗證的，是 8i 的商業化速度能否快過折舊壓力、Cloud margin 能否維持穩定，以及 TPU 生態的遷移門檻能否被控制平面與軟體工具鏈有效化解。只要這三條線同時爬坡，市場對 AI 基建的估值框架就會被迫重寫。