



〔投資研究報告〕

2025/8/21

AI 資本支出會失速嗎？ 從成本與槓桿看風險

目錄

- 一、推理成本壓力：營收熱不代表能賺錢
- 二、槓桿壓力：誰能跑長跑，誰被債壓垮
- 三、未來半年情境：大者更大

硬體狂潮背後，現金流能不能撐得住？

2025/8/21 新光投顧

過去兩年，全球雲端巨頭與新創把「買 GPU」當成生存賽跑：誰先把產能、電力與機房鎖住，誰就拿到話語權。但投資終究要回到一個問題——**這些資本支出 (CapEx) 能不能靠營收與現金流持續供養？**

本文用最貼近投資報酬的兩個變數做體檢：**推理成本 (需求端毛利)** 與 **金融槓桿 (供給端壓力)**。並把市場分成三群：**超級雲** (Microsoft、AWS、Google、Meta、Oracle)、**次級/專用 GPU 雲** (CoreWeave、Lambda 等)、**AI 應用/租算力使用者** (OpenAI、Anthropic 及各式 App/SaaS)。你將看到：哪一群最穩、哪一群最脆弱、未來半年最可能走向哪個情境，以及值得你放進投資筆記本的可量化觀測指標。

目錄

- 一、 推理成本壓力：營收熱不代表能賺錢
- 二、 槓桿壓力：誰能跑長跑，誰被債壓垮
- 三、 未來半年情境：大者更大

一、 推理成本壓力：營收熱不代表能賺錢

投資重點：看「單位推理成本」與「毛利」，不是只看用戶數。

- **單位成本是分水嶺。** 大模型的「推理」像是每回答一次就燒一次電與記憶體頻寬；如果每千字的推理成本降不下來，即便使用者增長，毛利也難以同步擴大。許多服務開始分級計量與限額，本質上是在修正單位經濟。
- **毛利率才是體溫計。** 對應用型公司，一條簡單的尺：**毛利率長期站不上 50%**，代表「用得越多、賠得越多」；只有當毛利率回升、訂價能反映成本，應用端的現金流才足以回頭支撐更多算力採購。
- **重度用戶吃穿補貼。** 一小群高用量客戶往往吃掉大半資源 (所謂「巨鯨效應」)。當平台被迫對重度用戶漲價或限流，短單會漂亮、**長單才關鍵**——不能把企業級、任務型場景綁進長期合約，才是營收可預期的來源。

給投資人的兩條線

1. 「每千字推理成本」半年內是否再降 30% (無論來自晶片能效、模型壓縮或軟體優化) ；
2. 應用商毛利率能否站上 50% 。

這兩條線若同時改善，需求端才算「健康擴張」；反之，熱度越高、錢燒得越兇。

二、 槓桿壓力：誰能跑長跑，誰被債壓垮

投資重點：看「利息/營收比」與「RPO 覆蓋率」，分辨誰能跑長跑。

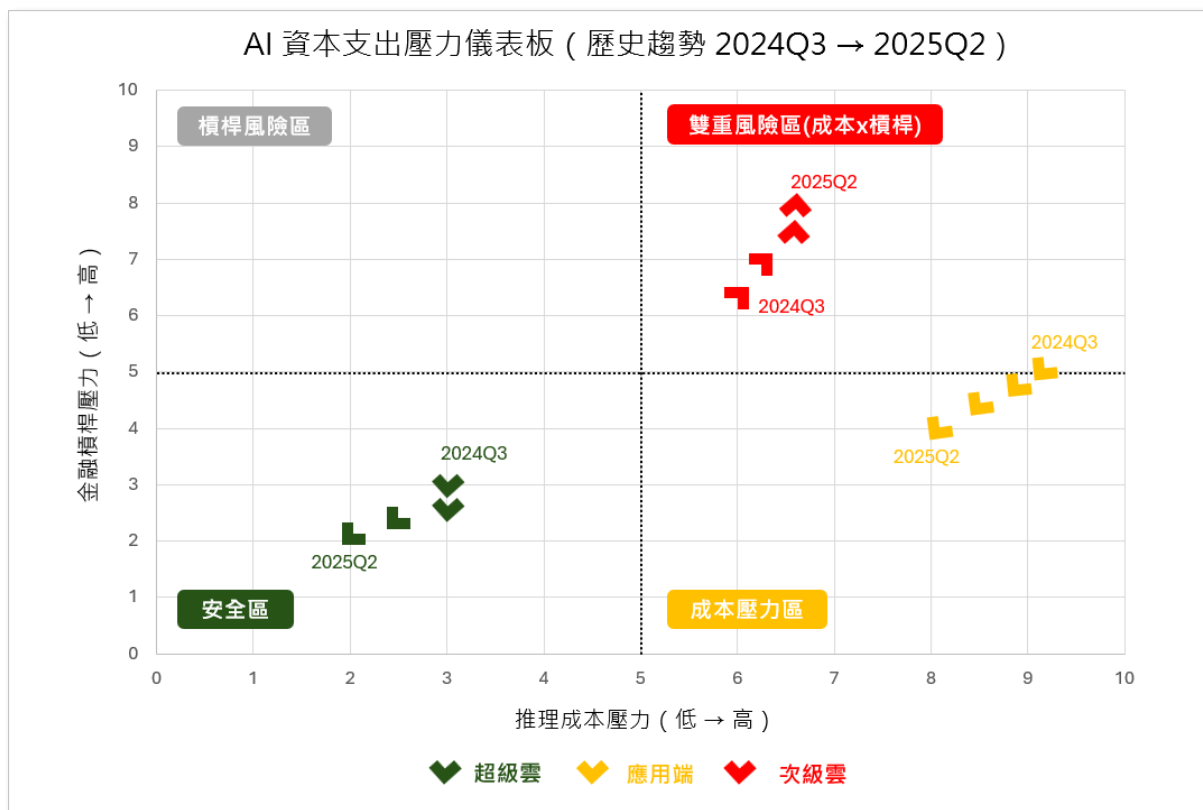
- **超級雲的優勢是資金成本與合約能見度。** 它們手上有廣告、雲訂閱與企業授權的現金奶牛，資金成本低、自由現金流厚，且「多年期合約 (RPO) 」可見度高，能吸收先建後用的波動。
- **次級雲的快與痛並存。** 多靠「資產抵押融資 (以 GPU 抵押) 」與戰略投資加速擴產；只要需求旺、利用率高，現金流就能轉；一旦需求不如預期，**折舊與利息會迅速放大壓力。**
- **電力與散熱是新瓶頸。** 高密度機架把功耗與冷卻推到新台階，帶動資料中心改造。對財務的含義是：**現金流回收更後移、資產更重、錯配風險更大。**

給投資人的兩條線

1. 利息支出/營收比 $\geq 25\%$ ：代表槓桿開始「反咬」；
 2. $RPO (未履約合約) \div 有息債務 < 0.5$ ：可見收入對不上槓桿。
- 這兩條線若同時變差，即使手上 GPU 很多，**擴張也可能踩煞車。**

三、 未來半年情境：大者更大

把上面兩組指標拉到座標軸，就得到一張好用的「推理成本 × 槓桿壓力四象限」如下(數值經過標準化以清楚呈現相對位置)：



製圖：新光證券

- 左下角 (安全區)：**超級雲 (Microsoft、AWS、Google、Meta、Oracle)** 位於此處——成本可控、槓桿低、合約能見度高，而且他們還在往更安全的方向 (左下角) 前進。
- 右下角 (成本壓力區)：**應用端 (OpenAI、Anthropic 及各式 App/SaaS)** 位於此處——商業模式仍在打磨，短期靠訂價與功能分層消化壓力。
- 右上角 (雙重風險區)：**次級雲 (CoreWeave、Lambda 等)** 位於此處——速度快但資產重、槓桿高，對「需求波動」最敏感，而且過去四季，他們的風險正在逐步升高。

接下來六個月，三種走向的機率判讀：

1. 「大者恆大」加速 (機率：中高)

超級雲的資本支出多已簽好產能與電力，集中度上升是大勢；次級雲不會立刻消失，但會被迫更聚焦地區節點、合規或垂直場景。

2. 「利基市場生根」擴散（機率：中）

私有雲即服務、行業專用模型與在地合規，會讓中端與地區性供應商找到生存縫隙；但規模與資金厚度仍難比肩超級雲，屬「慢慢長」。

3. 「需求撐不起供給」觸發雪崩（機率：低）

短期難見全面急凍；真正的系統性風險，多半來自兩顆警報燈同時亮起：

- 推理成本不降 & 毛利起不來 → 需求端回收變慢；
- 利息/營收跳升 & RPO 覆蓋轉弱 → 供給端現金流吃緊。
一旦重疊，次級雲與高槓桿玩家最先受衝擊，其次是高資本密度的供應鏈。

投資人行動清單

- **每季追四指標**：①每千字推理成本（是否持續下降）；②應用商毛利率（是否回到 50% 以上）；③利息/營收比（是否逼近 25%）；④RPO÷債務（是否低於 0.5）。
- **看現金流，不只看故事**：對次級雲與新創，優先檢視**利用率與合約長短**，再談成長故事。
- **逢高檢視資產負債表彈性**：在「大者恆大」階段，**資金成本與合約能見度**就是護城河；對供應鏈標的，也要對照**客戶集中度與交貨排程**，避免被單一客戶的投資節奏綁架。
- **留一手給「效率技術」**：任何能把同樣任務做得更省電、更省記憶體**的軟硬整合**，都是推理成本下行的加速器——這些公司常是**小而美的高彈性受益者**。

AI 投資已進入「回本力比噱頭重要」的新階段。投資人不該只盯晶片規格或新聞聲量，而是看**單位成本曲線與槓桿壓力曲線**，這兩條線決定了資本支出能跑多久，也決定了誰能笑到最後。