GEMINI 429 ISSUES

Gemini 429 錯誤
根因分析與對策建議

429

Gemini Enterprise Agent Platform · 動態共用配額（DSQ）下的容量策略

RESOURCE_EXHAUSTED

2026 年 7 月 8 日

現況:配額一切正常，429 照樣發生

0.02%

系統限制用量

us-central1 圖片輸入 RPM 上限 4,025 萬，實際僅 7,791

無限制

Token 配額顯示值

flash-ga / flash-lite 四項配額全為 DSQ，無固定數字

1.6 萬

每分鐘輸入 Tokens

flash-ga 實際流量（每日約 589 萬），規模很小

結論:瓶頸不在專案配額。控制台看不到任何逼近上限的數字——429 來自模型 × 區域的「共用容量池」瞬時吃緊，配額頁面永遠不會反映這件事。

實際案例：正式環境的 429 日誌

日誌重點（Cloud Logging）

服務 image-recog-data（Cloud Run / asia-east1）

專案 prod-image-rd（正式環境）

錯誤 google.genai ClientError: 429 RESOURCE_EXHAUSTED

場景 護照圖片辨識請求，以及微調模型預測

> “Resource exhausted. Please try again later.”
> ＝ 即付即用（DSQ）架構的官方 429 訊息

事件時間軸（6/18–6/29，共 7 次）

- **6/18** **11:58** 圖片辨識（第 13 張），耗時 2.2s
- **6/24** **13:12** 圖片辨識（第 2 張），耗時 9.5s
- **6/25** **10:32 / 16:24** 圖片辨識 2 次，耗時 10.1s / 11.4s
- **6/26** **17:59** 微調模型預測
- **6/29** **16:23 / 16:33** 圖片辨識（10.7s）＋微調模型預測

初步判讀： 12 天 7 次、全落在台灣上班時段、失敗前多有 9–11 秒等待——符合 DSQ 瞬時吃緊的間歇特徵，偏尖峰型。

根因：動態共用配額（DSQ）

傳統固定配額

額度歸屬　每專案固定數字（舊制示意，與支出無關）

429 條件　自己的用量到頂才會發生，可預期

控制台　用量百分比清楚可見，可提前預警

調整方式　可提出配額提高要求（QIR）逐步擴充

DSQ（Gemini 2.0+ 預設）

額度歸屬　「模型 × 區域」全球所有客戶共用一池

429 條件　池子無餘裕即拒絕，與自身用量無關

控制台　顯示「無限制」、百分比「-」，無從預警

調整方式　沒有固定數字可調，靠架構與 PT 應對

用量層級（Usage Tier）：依組織近 30 天支出分級。本組織月支出>$2,000 ＝ 級別 3（最高級）：Flash 系列基準 3,000 萬 TPM／模型。但基準以全域端點流量為準，且瞬間爆量即使低於基準仍會被節流。

何時容易撞上 429

熱門模型

2.5 Flash 等當紅模型競爭最激烈 ——實例顯示每小時 70–80 個請求就可能被拒，量小不代表安全

熱門區域

us-central1 是最擁擠的區域之一，而我們正在使用它

尖峰時段

美國上班時間全球需求最高，池子餘裕以秒為單位波動

大請求

長 context 單次需一次取得大量容量，比小請求更容易被拒

特徵：間歇性、無規律 ——這一秒 429、下一秒成功。第一步先抓 429 的時間分布，判定是尖峰問題還是持續性容量問題。

429 的官方機制（SLA 觀點）

配額架構	錯誤訊息 / 行為	SLA 計算
即付即用（現況）	`Resource exhausted, please try again later.`	不計入錯誤率，可重試
標準 PT（低於購買量）	容量問題轉為 5XX 回傳	計入 SLA 錯誤率（有保障）
單一可用區 PT（低於購買量）	容量問題轉為 5XX 回傳	不計入 SLA 錯誤率
PT（超過購買量）	超額請求自動改走即付即用（spillover）	依即付即用規則

官方建議的即付即用解法

1 優先使用全域端點

2 指數退避重試

3 平緩流量、消除尖峰

4 訂閱 PT 預留吞吐量

解法選項比較

流量平緩＋退避

指數退避與抖動重試；客戶端限速消除尖峰

成本：零

立即可做，對間歇性尖峰最有效

批次 API

可延後的流量進佇列排隊，而非直接被拒

成本：更低

換取容量彈性，犧牲即時性

多區域分流

429 時自動改打其他允許區域的端點

成本：低

僅限合規邊界內的區域

佈建輸送量（PT）

購買 GSU 預留吞吐量，唯一的容量保證

成本：高

適合關鍵業務的基載流量

採用順序建議：由左至右——先做零成本的重試與平滑，確認仍不足再往右投資。

合規資料夾下的限制：全域端點不可用

全域端點：官方首選，但我們不能用

- 不保證請求在哪個地理位置處理
- 資料落地（data residency）承諾僅涵蓋區域端點
- Assured Workloads 位置設定不含全域端點
- 資源位置組織政策很可能直接擋下global 請求

合規邊界內的替代方案

- 邊界內多區域 failover：us-central1 被拒即改打 us-east4 等允許區域，效果接近全域端點
- 與合規負責人確認「允許區域清單」後實作
- 需要保證容量時，在合規區域內購買 PT

註記：優先隨用隨付（Priority PayGo）僅支援全域端點，在合規邊界內同樣不可用。費用面：Gemini 2.5 全域與區域端點同價；Gemini 3+ 正式版起（2026/7/1）區域端點貴 10%。

不建議的做法

開多個專案分流

- DSQ 池全客戶共用——拆專案是多個專案搶同一池，容量不會變多
- Usage Tier 以「組織」歷史消費計算，拆專案保障底線不變
- 服務條款明文禁止以多專案規避配額（quota circumvention）
- IAM／資源 / 監控／合規審查成本倍增，效益為零

跨合規邊界的代理專案

- 合規要求跟著「資料」走，不是跟著專案走
- 受管制資料借道外部專案＝把違規點搬到架構層
- VPC Service Controls 預設會擋跨專案呼叫
- 例外：非敏感工作負載分流、或去識別化後出境（需合規認定）

建議行動

短期（1–2 週）

1 匯出並分析 429 時間分布：尖峰型或持續型？

2 實作指數退避＋抖動重試

3 客戶端流量平滑，消除瞬間發射尖峰

中期（本季）

1 可延後流量改走批次API

2 與合規確認允許區域清單

3 建立邊界內多區域 failover

長期

1 依業務關鍵性評估PT（GSU 用量規劃）

2 Gemini 3+ 升級時重新評估端點與成本策略（區域+10%）

關鍵訊息

- 429 ≠ 專案配額用罄——是 DSQ 共用容量池的瞬時吃緊，控制台永遠看不到
- 先做零成本的重試與流量平滑，確認不足再投資批次、多區域與PT
- 所有架構調整以合規邊界為前提：全域端點與跨邊界代理都不可行

下一步：本週匯出 429 錯誤的時間分布數據，確定尖峰型或持續型，據此選擇投資路線