

從線上售票看作業系統設計議題

Jim Huang (黃敬群) <jserv.tw@gmail.com>

Jan 8, 2015

注意：

本簡報大量引用網路討論，請斟酌服用



Source:

<https://www.facebook.com/707851765997333/photos/a.707865592662617.1073741827.707851765997333/707865582662618/>

寬宏藝術的聲明

(Jan 6, 2015)

「... 開賣前日，寬宏的網站就持續湧進 22 萬網友進行註冊與試用，2015 江蕙祝福演唱會開賣第一天，更吸引了 35 萬人次同時上網搶票，雖然寬宏使用了速博最大頻寬來服務購票民眾，但真的人數眾多，同時間於全台更有上萬台購票機操作購票多重大量作業下，也因此嚴重影響了售票速度 ... 」

Ant 質疑寬宏聲明造成的誤導



OneStat.com
Number One Real-time Intelligence Web Analytics

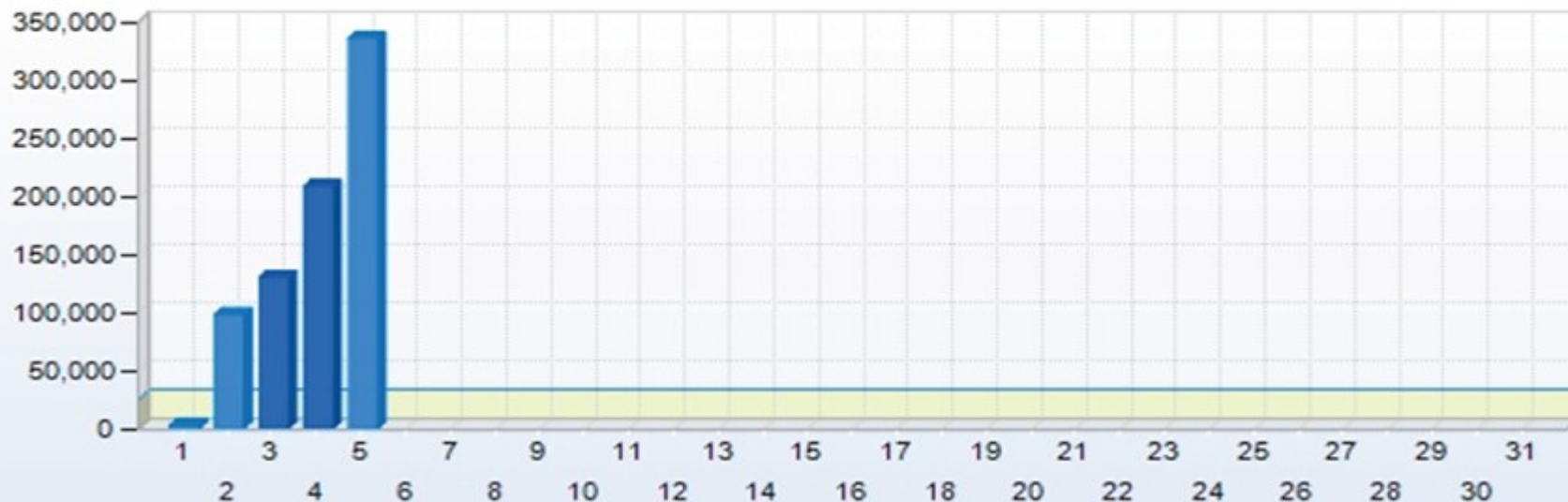
Basic

Account: 221889 - kham Taiwan

Summary

Period: 1 month (1/1/2015 - 1/31/2015)

Weeks Days



Summary

Total pageviews: 771,204

Ant 質疑「35 萬人次同時上網搶票？」

(Jan 7, 2015)

- 圖表的 X 軸是日 (天)。「同時上網」在業界的術語是 concurrent user，但單位至少必須是秒
- 從圖表得知，1/5 日當天根本就不到 35 萬人
- concurrent user 至少必須以「秒」來看，那麼當日 35 萬人換算每秒平均後，也只有 4 人同時在線
$$350,000 \div 24 \div 60 \div 60 = 4.05$$
- 事實上我們都知道不會這麼平均。所以改用峰值推論。假設該日的人數 99% 集中在 1% 的時間 (等同 35 萬人有 99% 人平均落在當日某 15 分鐘一同搶票)，所以推論高峰時有 401 人同時在線。
- 因此，就算用 99% 的峰值推論法，高峰時每秒也只有 401 人同時在線，根本不是官方所稱的 35 萬人同時在線

同時上線的討論

- <Ant Yi-Feng Tzeng> 當日 35 萬，假設同時就 4000 好了。那麼只需 87.5 秒就消耗掉 35 萬，那麼其它時間沒人嗎？怎麼可能，寬宏可是長時間幾乎都不能用
- <Xeon M Freeman> 分析峰值和端點行為完全不能用平均啊，要看瞬間最大承載量。4000 也不會是登入就立馬買好一秒離線
- <Ant Yi-Feng Tzeng> 4000 當然不是一秒就好，但若到了下一秒，當然就是計到下一秒的 4000。注意，4000 是當秒人數，若同一人到了下一秒，則當然計入下一秒的 4000 中 ... 不管是 Web server 或 database，即使塞住，恢復時間也不過數秒內，不可能超過一分鐘。所以你還是無法說明，同時 4000 人時，為何 23 小時服務都還是卡住 (24 小時扣掉 87.5 秒)
- <Ant Yi-Feng Tzeng> 圖表的 35 萬指的是 Page View，不是 unique user

為何每秒僅 401 個同時上線也會出包？

- 問題出於網路頻寬嗎？
- 據網路消息指出
 - 寬宏使用的是單條 100/100 Mbps 的頻寬
 - 寬宏當初的網頁共需 2M 左右的下載量
- 若消息屬實， $401 \times 2\text{M} = 802 \text{ Mbps}$ ，確實遠大於 100 Mbps 的量

為何每秒僅 401 個同時上線也會出包？

- 問題出於**資料庫**嗎？
- 某前輩指出，售票的事務邏輯很複雜，不要與一般的 QPS(每秒查詢處理量) 等同比較，之間相差的量級可能有百倍之有

資料庫使用的討論

- 以 MySQL 在 2010 年針對 5.1 版本進行的 TPS (每秒事務處理量) 效能測試為例
<http://www.percona.com/blog/2010/02/28/maximal-write-throughput-in-mysql/>
- 假設，
 - A. 以當初測試數據裡最嚴謹的數據來看，每秒至少在一萬上下 (10000 TPS)
 - B. 寬宏現在上線的資料庫處理能力與 4 年前的 MySQL 5.1 同等級。
(雖然實際上我認為他們用的是更快的資料庫)
 - C. 寬宏現有硬體不如測試環境上那般好。假設處理效能只達二分之一
 - D. 寬宏售票的事務比測試環境複雜十倍
- 所以， $10000 \div 2(C) \div 10(D) = 500 \text{ TPS}$ 。即使在這麼艱難的條件下，單台資料庫也該能處理 500 TPS，高於先前推估的 401 同時在線人數

資料庫使用的討論

- < 宋維民 > HTTP header 裡面

Server: Microsoft-IIS/6.0

所以合理猜測跑的是 Windows 2003

- <Ant Yi-Feng> 「假設寬宏現在上線的資料庫處理能力與 4 年前的 MySQL 5.1 同等級」，指的是同等級，沒說就是 MySQL。

TonyQ 的解說

- 本質上你可以把所有演唱會位置視為一個格。交易這回事本質上就是把人分到他想要的格子。先不論金流來簡化問題，這個問題的關鍵因素在同時 (concurrent) 在線人數，這種大型演唱會搶票基本上都是同時萬人以上等級
- 我們會碰到的第一個問題是「我們要讓使用者知道哪個格有人、哪個格子沒人」，因為使用者要選位，這件事情就已經夠困難了

TonyQ 分析：訊息問題

- 有沒有玩過線上遊戲，實作一個同時一萬個人在線上（而且一直在講話）的聊天室跟同時實作一百個人在線上的聊天室，訊息量的差異是「指數級」以上的差距。假設後者是 100^2 的話，前者可能就是 10000^6 （只是打個比方啦，數字沒有很精準）。而即時回報座位訊息，大概就像是一萬個人的聊天室。（傳遞的是我選了、沒選的訊息）
- 若仔細觀察，LINE、Cubie、Google hangout 都設定有一百人到兩百人不等得群組上限，背後的原因就在這裡
- 這是第一個，而且坦白說算相對好處理：訊息問題

TonyQ 分析： Lock

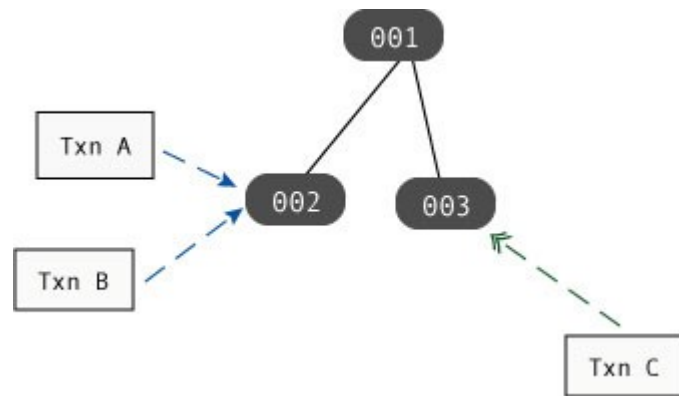
- Lock 是千古難題，概念上不困難，但實作非常困難。 race condition 一直是我們這個領域中最難處理的問題
- 很多人在談的機制都很合理，不論取號也好、各種作法也好。但重點是 concurrent 一萬人以上的情況，你一定要作「分散運算」，否則你在作業系統層面基本上很難處理這麼大量的 concurrent 操作，而且風險也很高（重點！）
- 但分散運算對 lock 來講，根本是先天的天險

Lock 在線上訂票的展現 (1)

[Scott Tsai: 縮減 concurrent write lock contention]

- 若不支援消費者自行劃位，則不用維護一個全域性的「空位數」，可以切成 k 份

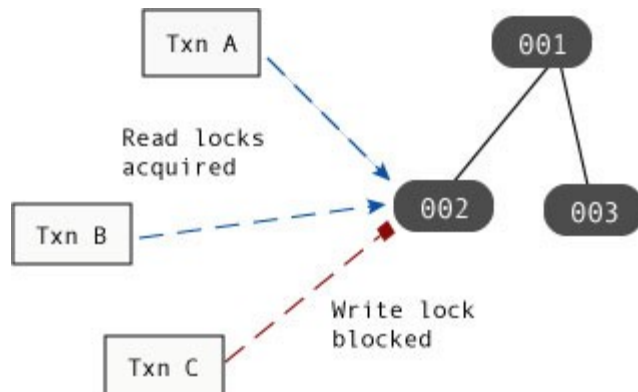
作業系統概念：page fault count 維護 per-cpu counter，要輸出時才加總



Lock 在線上訂票的展現 (2)

[Scott Tsai: 縮減 concurrent write lock contention]

- 若只支援「選座位區塊」，不允許細部劃位，則可把每個區塊空位分給 k 個「同時劃位區」。每個同時劃位區，可同時寫入。
- 縮解單一 write lock 保護的範圍



TonyQ 分析： Lock 與分散式系統

- 原本一台伺服器自己撈記憶體或檔案查就好的事，現在會變成多台伺服器要找某台中央伺服器要資料。兩者的速度與 request 差距是數十倍
- 另一方面，要如何確保這個 lock 在所有機器上都確實處理好？這個機制最後的「源頭」不又回到了 concurrent 10000 對 1 的情況？
- 所以這不是單純的事情，我們還要把「格子」的管理也切開好幾台來分散式管理，但如此一來，又回到原本的問題，你本來是多台對一台的訊息管理，現在是「多台對多台之間的訊息 request」
- 這又是一個架構級的變化，這中間的平衡非常不好抓

TonyQ 總結

- 更不用說，真正面對這個問題的廠商，根本就不覺得自己該把處理這種數萬人的架構當成他最重要的任務。他們的想法就是把問題丟出去給客戶端去承擔（諸如 ibon），甚至坦白說，我覺得他們恐怕認為這個問題無解，這些廠商如果有把這些事情當成真正一回事看，這些系統的設計不會這麼原始、陽春而令人打從心理感到發笑
- 專家從系統面就看得出來有沒有認真花力氣了

回歸架構議題

- <TonyQ> 這個量級已不是單一語言能處理的範圍了。都是架構設計議題
- <Sheng Lih Wang> 股票交易分成前台跟後台，前台是 client-server 架構，讓交易員可以猛打單。後台的撮合就神奇，它分成自家區域撮合跟跨所非同步撮合。一筆交易要成交是多台電腦於多群電腦之間用非同步交易串流交叉演算撮合。
不過股票跟賣票有點不同：它是買方與賣方各設條件，再依照「先券商後證交所」的規則去 post match，相對來說是很規矩的交易

學習電腦科學的你我，應該要能夠用專業看待這些經典議題

延伸閱讀：〈由 12306.cn 談談網站性能技術〉

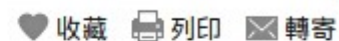
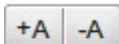
<http://coolshell.cn/articles/6470.html>

真正幸福的人：不是搶到票，是 可以像江蕙一樣選擇人生




撰文者 | 李柏鋒

李柏鋒的擴大機 | 瀏覽數：200+ | 2015-01-08



來源：翻攝自江蕙臉書

 放大顯示