雲端運算即時轉移記憶體預測機率之改善研究

林逸偉 [a*]、李維聰 [a]、楊堯強 [a]、許東榮 [b]
淡江大學電機工程學系 [a]
中山科學研究院 [b]
通信作者[*] ywlin1125@hotmail.com

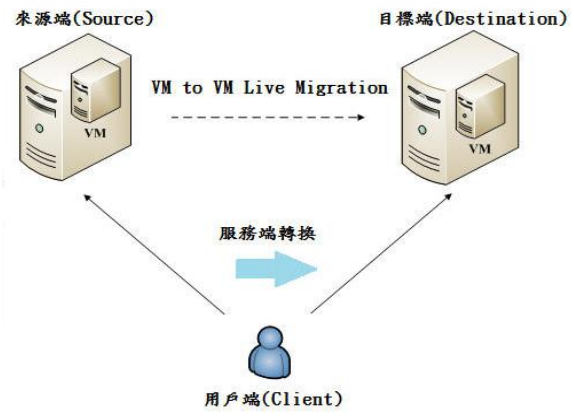
摘要

隨著雲端運算技術的成熟，近年來越來越多的雲端運算應用產生，其中雲端運算應用的系統虛擬化技術，變得相當重要。伺服器虛擬化可以提高伺服器服務效能與提供動態資源調度的特性。但伺服器會發生維修、負載過量或是毀損的情況，造成用戶出現服務中斷的狀況。針對此問題，發展出即時轉移技術，在狀況發生時，能夠持續用戶端使用，而不會感到服務中斷。此篇論文在介紹即時轉移技術中，利用預測記憶體修改機率，判別頁面是否暫停-複製轉移(Stop-and-Copy Phase)。

關鍵字: 即時轉移記憶體預測機率

一、 簡介

虛擬化技術(Virtual Machine)在雲端計算技術的應用上，能夠提升伺服器處理資料數據的安全性和運作效能。在資料數據即時轉移技術中(Live Migration)，是讓運行中的伺服器在動態運行計算的同時，不需關閉虛擬機，依然保持用戶端的各項服務下，將整個虛擬機快速完整的從來源端遷移到目標端，如圖一。先前的論文是在探討總轉移時間與停機時間兩項參數間取得平衡，以減少不必要的資料頁面轉移，來減少總轉移時間。本論文提出一個預測記憶體修改率的模型，用來判斷在各次疊代中，哪些資料頁面適合在暫停-複製階段轉移，減少過多無意義資料頁面的轉移次數。



圖一： 即時轉移示意圖

二、 背景知識與相關研究

2.1 典型的資料轉移

為了實現虛擬機的動態資料轉移可以分成三個階段，推進階段(Push Phase)、暫停-複製階段(Stop-and-Copy Phase) 、拖行階段(Pull Phase)，如圖二。有此三個階段可以完成一個完整的資料轉移，目前主流的提出方法皆是結合其中兩個階段，預複製轉移(Pre-Copy Migration)是結合前兩者，Post-Copy 則是結合後兩者。[1]-[4]

2.1.1 推進階段(Push Phase)

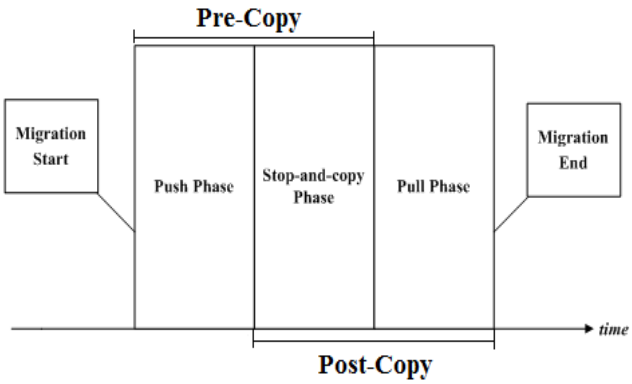
當系統資料開始進行轉移時，來源端(Source)的虛擬機持續運行，並且在運作時間內會持續的傳送資料到目標端(Destination)，為保持與目標端的一致性，在運作期間內，所有更改過的資料都必須要重新傳送。

2.1.2 暫停-複製階段(Stop-and-Copy Phase)

來源端的虛擬機停止運行，並在停止的期間內，傳送資料到目標端的虛擬機，直到系統內容完全轉移到目標端虛擬機後，目標端的虛擬機才會恢復運行。

2.1.2 拖行階段(Pull Phase)

目標端存取到尚未複製的內容，會發生頁面錯誤的(Page Fault)並且從來源端重新傳送系統內容到目標端。



圖二： 虛擬幾轉移階段示意圖

2.2 預複製轉移(Pre-Copy Migration)

預複製轉移的方法可以分為下列六個步驟：

2.2.1 Pre-Migration

確認目標端與來源端的資訊可否相容。來源端會發出需要傳送到目標端的各項資訊，用來選取適合的目標端傳送

2.2.2 Reservation

目標端選擇確認之後，來源端會發出 migration 的要求，待目標端確認完成會發送訊息告知來源端可以開始傳送資料。

2.2.3 Iterative Pre-Copy

來源端開始進行疊代轉移資料頁面的動作，在第一個回合中，所有的頁面將會從來源端傳送到目標端。在接下來的所有回合，只會傳送在上一回合期間內，有更改過的頁面(Dirty Page)。

2.2.4 Stop-and-Copy

來源端停止運行，並且將 CPU state、剩餘的頁面以及在 pre-copy 最後一回合有更改過的頁面資料，也就是具有高修改率(High Dirty Rate)的頁面資料轉移到目標端。

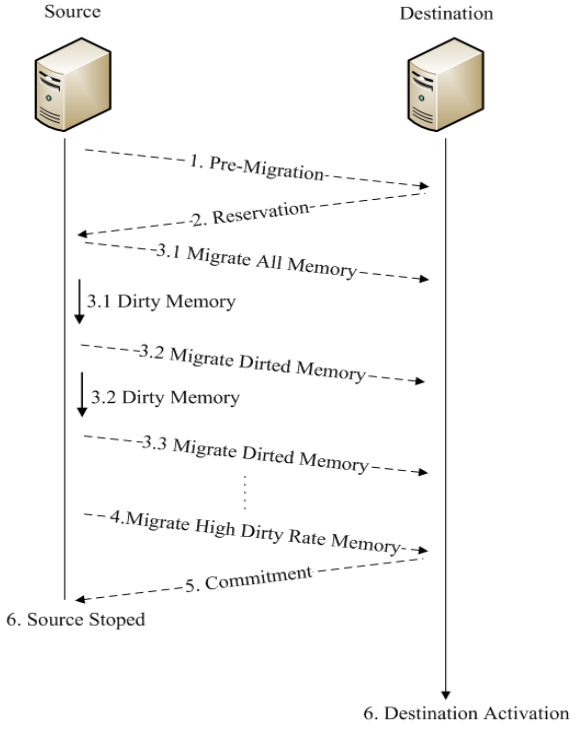
2.2.5 Commitment

目標端通知來源端已經成功收到完整且一致的資料後，發出訊息告知可關閉來源端端。至此可以交手給目標端的虛擬機運行。

2.2.6 Activation

來源端的虛擬機接續來源端的服務運行。

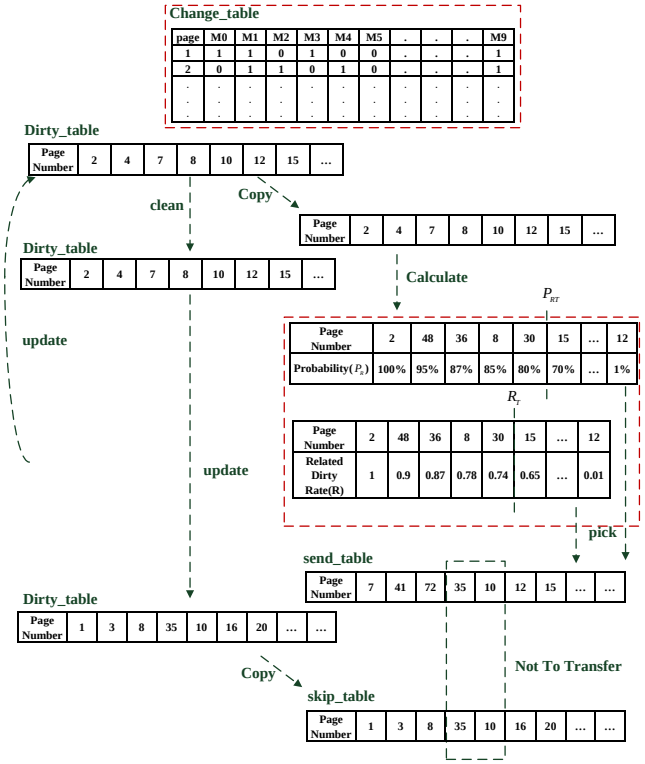
圖三步驟 3.1 表示疊代轉移(Interactive Pre-Copy)中的第一次疊代轉移動作，從來源端轉移資料至目標端時，由於來源端仍在運行，因此會產生出修改頁面(Dirty Page)，圖三步驟 3.2 表示疊代轉移中的二次疊代，依此類推到最後的圖三步驟 4 的暫停-複製階段(Stop-and-Copy Phase)。之後經由圖三步驟 5、6 之後交手給目標端持續原來服務。[1]-[4]



圖三： 預複製轉移(Pre-Copy Migration)階段示意圖

2.3 頁面修改預測機率方法

先前論文利用馬可夫預測模型(Markov Model)與 Bayes 條件機率做整合，可以根據目前的疊代修改，預估出之後各個頁面需要修改的發生機率。而由這些預估的機率，適當的選擇一個臨界值 P_T、R_T來判定頁面是否需要轉移。若頁面每次疊代之後的修改機率都大於臨界值 P_T、R_T，則此頁面不會在這次疊代中進行轉移，如圖四。[1]



圖四：記憶體預測流程示意圖

三、 頁面修改預測機率計算模型

3.1.1 馬可夫預測模型(Markov Model)

在傳統的記憶體頁面(Memory Page)在讀取或未操作時標為 0，所以做出設定為：0、1 分別代表為未改變、改變，藉以提高頁面操作變化的準確性。由於預複製轉移為根據記憶體是否做過修改變更來判斷是否需要轉移，因此讀取與未動作並不直接影響轉移效能；先前論文重新推導此頁面修改機率預測模型。

假設狀態 E_i、E_j，由頁面 E_i 變為 E_j 的狀態機率為本文(1)式。[1]

$$P(E_i \rightarrow E_j) = P(E_i \mid E_j) = P_{ij} \qquad (1)$$

若資料頁面有 n 個狀態，則狀態機率為本文(2)式。

$$P(E_i \to E_j) = P(E_i \mid E_j) = P = \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1n} \\ P_{21} & P_{22} & \cdots & P_{2n} \\ \vdots & \vdots & & \vdots \\ P_{n1} & P_{n2} & \cdots & P_{nn} \end{bmatrix} \quad (2)$$

由於傳統記憶體頁面只有兩種狀態，n=2，記憶體被修改時，所處狀態為 E_0；記憶體未修改，狀態為 E_1。假設已知記憶體初始狀態(k=0)，經過 k 次狀態修改後，在第 k 次疊代處於狀態 的機率，根據馬可夫過程的無後效性及 Bayes 條件機率公式為本文(3)式。

$$\pi_j(k) = \sum_{i=1}^{n} \pi_j(k-1) P_{ij} \quad j=1,2,...,n \quad (3)$$

3.1.1 馬可夫預測模型(Markov Model)例子說明[1]

下圖五是隨機 10 次修改頁面的記錄，「1」表示頁面有修改，「0」表示頁面有無變化。

頁面	1	2	3	4	5	6	7	8	9	10
狀態變化	1	0	1	0	1	1	0	0	0	0

圖五：隨機 10 次頁面修改記錄

我們可以知道總共有 4 個過程跟「1」有關係的，其中只有 1 個「1」開始然後「1」結尾的狀態，根據前述(1)式，我們可以得到下面的結果如(4)式：

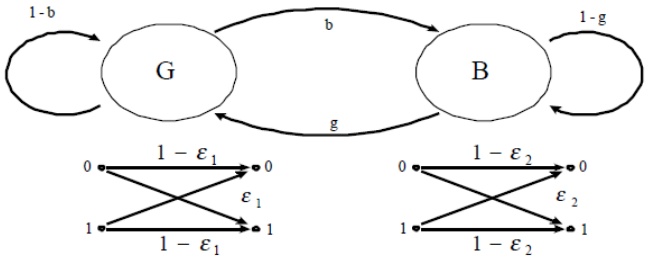
$P_{11} = P(E_1 \to E_1) = P(E_1 \mid E_1) = 1/4 = 25\%$ (4)

根據(2)式而再一次疊代之後的修改率為 25%，如式(5)

$\pi(1) = \pi(0)P = 25\%$ (5)

3.2.1 吉爾伯特-艾略特模型(Gilbert-Elliot Model)

由於先前論文只有針對 Dirty Page 的部分做修改頁面預測，在此我們引用 Gilbert-Elliot (GE) Model 來做討論。[5]



圖六：G-E Model

圖六模型為兩個狀態的馬可夫模型，狀態「G」表示通訊良好的狀態，此狀態對應一個位元錯誤率為 ε_1，狀態「B」表示通訊較差的狀態，此狀態對應一個位元錯誤率為 ε_2 ，且 $\varepsilon_1 << \varepsilon_2$ 。此馬可夫模型的狀態變機率(State transition probability)為：

P_{GG}= P(G｜G)=P(第 i+1 個狀態 G｜第 i 個狀態 G)=1-b

P_{BG}= P(B｜G)=P(第 i+1 個狀態 B｜第 i 個狀態 G)=g

P_{GB}= P(G｜B)=P(第 i+1 個狀態 G｜第 i 個狀態 B)=b

P_{BB}= P(B｜B)=P(第 i+1 個狀態 B｜第 i 個狀態 B)=1-g

因此狀態轉變矩陣(State Transition Matrix)為本文(6)式。[6]

$$S = \begin{bmatrix} P_{GG} & P_{GB} \\ P_{BG} & P_{BB} \end{bmatrix} = \begin{bmatrix} 1-b & b \\ g & 1-g \end{bmatrix} \quad (6)$$

在長時間下，馬克夫模型處於狀態 G 或是 B 的時間比例會接近一常數，而此常數為穩態機率(Steady State Probability)，令 P(G)表是好的狀態的穩態機率，P(B)為壞狀態的穩態機率，可以得到本文(7)、(8)式。[6]

$$P(G) = \frac{g}{b+g} \quad (7)$$

$$P(B) = \frac{b}{b+g} \quad (8)$$

所以根據 Gilbert-Elliot 模型中，我們可以知道每個位元的錯誤機率為本文(9)式。[6]

$$Ave\_BER = P(G)\times\varepsilon_1 + P(B)\times\varepsilon_2 = \frac{g}{b+g}\times\varepsilon_1 + \frac{b}{b+g}\times\varepsilon_2 \quad (9)$$

3.2.1 吉爾伯特-艾略特模型(Gilbert-Elliot Model) 例子說明[1]

根據圖五，我們也可以用 GE Model 來計算更改機率，首先我們可以個別求出每個狀態的機率分別為下面(10)、(11)、(12)、(13)式。

P_{GG}= P(G｜G)= 1-b = 40% (10)

P_{BG}= P(B｜G)= g = 90% (11)

P_{GB}= P(G｜B)= b = 60% (12)

P_{BB}= P(B｜B)= 1-g = 10% (13)

因為在一般的情況下，我們設 ε_1=0.1， ε_2=0.9 來做計算可以得到下面結果 42%，此為我們計算出來的平均更改機率，如(14)式。

Ave_BER=P(G)x ε_1+ P(G)x ε_2 = 42% (14)

四、 結果與分析

4.1 針對 Dirty Page 變化下的分析(P_{BB}、P_{BG})

經由前面章節所介紹的公式，我們運用馬克夫模型透過實際的資料數據，來做以下分析計算。

4.1.1 Random 狀況下的預測機率

亂數取一個頁面變化的 Table 如圖七。

Update / Page Number	U0	U1	U02	U3	U4	U5	U6	U7	U8	U9	U10
1	1	0	0	0	0	1	1	0	1	1	0
2	1	1	1	0	0	1	1	1	0	1	1
3	1	0	0	0	1	1	1	0	1	1	0
4	1	0	0	0	1	1	1	1	1	0	0
5	1	1	0	0	1	0	0	1	0	0	1
6	1	1	0	1	0	0	1	0	1	0	0
7	1	0	1	1	0	1	1	0	0	0	1
8	1	0	1	1	0	0	0	0	0	0	1
9	1	1	1	1	1	0	0	0	1	1	1
10	1	0	0	0	0	0	1	0	0	0	0

圖七：亂數頁面變化 table

利用馬可夫模型，來計算出各頁面在 U10 疊代時頁面變化的預測值為圖八。

Page Number	U10
1	20.00000%
2	50.00000%
3	30.00000%
4	40.00000%
5	10.00000%
6	10.00000%
7	20.00000%
8	10.00000%
9	60.00000%
10	0.00000%

圖八：U10 疊代頁面變化預測值

4.1.2 Locality 狀況下的預測機率

一般來說，記憶體頁面傳輸，齊傳輸資要的變化都會有 Locality 的特性，根據 locality 的方式產生頁面變化的 Table 如圖九。

Update / Page Number	U0	U1	U02	U3	U4	U5	U6	U7	U8	U9	U10
1	1	0	0	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0	1	1	1
3	1	1	0	0	0	0	0	0	0	0	0
4	1	1	1	0	0	0	0	0	0	0	0
5	1	1	1	1	1	1	0	0	0	0	0
6	1	0	0	0	0	0	0	0	1	1	1
7	1	1	1	1	1	1	1	0	0	0	0
8	1	1	1	1	0	0	0	0	1	1	1
9	1	1	1	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	0	1	1	1

圖九：Locality 方式產生的頁面變化 table

同理運用馬可夫模型，來計算出各頁面在 U10 疊代時頁面變化的預測值為圖十。

Page Number	U10
1	0.00000%
2	20.00000%
3	10.00000%
4	20.00000%
5	50.00000%
6	20.00000%
7	60.00000%
8	50.00000%
9	100.00000%
10	80.00000%

圖十：U10 疊代頁面變化預測值

4.2 考慮實際情形的狀下的分析 (P_{GG}、P_{BG}、P_{GB}、P_{BB})

過前面章節所介紹的公式，加上考慮 P_{GB} 的情形下，我們同樣實際運用 Random Page Change 和 Locality Page Change 的例子中來做驗證。透過實際的資料數據，來做以下分析計算。

4.2.1 Random 狀況下的預測機率

經由前面章節介紹的 Gilbert-Elliot 模型，來計算出圖六各頁面在 U10 疊代時頁面變化的預測值為圖十一。

Page Number	U10 Ave_BER
1	47.33%
2	61.43%
3	52.67%
4	53.08%
5	45.00%
6	47.65%
7	50.00%
8	38.57%
9	63.33%
10	32.86%

圖十一：U10 疊代頁面變化預測值

4.2.2 Locality 狀況下的預測機率

同理運用 Gilbert-Elliot 模型，來計算出各頁面在圖八 U10 疊代時頁面變化的預測值為圖十二。

Page Number	U10 Ave_BER
1	23.33%
2	36.67%
3	30.00%
4	36.67%
5	56.67%
6	36.67%
7	65.38%
8	56.67%
9	90.00%
10	76.67%

圖十二：U10 疊代頁面變化預測值

五、 結論與未來方向

經由上面的計算結果，我們可以發現，實際考量資料傳輸狀況的模型，展現出來的預測機率與只考慮 Dirty Page 的狀況下，會有更準確的預測機率。在未來，可以根據考量實際狀況下的預測機率，可以找出更準確的臨界值 Prt 來判斷虛擬機的停機時機，減少過多的停機時間。

參考文獻

[1] 黃致祥,“應用於雲端運送即時轉移之相對式記憶體修改預測機制”淡江大學電機工程學系碩士論文,中華民國一百零一年六月.

[2] Christopher Clark, Keir Fraser, “Live migration of virtual machines”, *NSDI'05 Proceedings of the 2nd conference on Symposium on Networked Systems Design & Implementation*, Vol. 2, pp. 1-14. **(2005)**

[3] Michael R. Hines, Umesh Deshpande, Kartik Gopalan, “Post-copy live migration of virtual machines”, *ACM SIGOPS Operating Systems Review*, Vol. 43, Issue 3, July **(2009)**

[4] Bolin Hu , Zhou Lei , Yu Lei , Dong Xu , Jiandun Li , “A Time-Series Based Pre-copy Approach for Live Migration of Virtual Machines”, *2011 IEEE 17th International Conference on Parallel and Distributed Systems (ICPADS)*,pp. 947 – 952, 7-9 Dec.**(2011)**

[5] Sun Guo-fei, Gu Jian-hua, Hu Jin-hua, Zhao Tian-hai ,“Improvement of Live Memory Migration Mechanism for Virtual Machine Based on Pre-copy”, Computer Engineering, Vol.37, No.13, July(2011), http://en.cnki.com.cn/Article_en/CJFDTOTAL-JSJC201113011.htm

[6] E. N. Gilbert, “Capacity of burst-noise channels,” Bell Syst. Tech.J., vol. 39, pp. 1253-1265, Sept. 1960.