Hadoop 雲端運算平台檔案系統效能之評估與改善

The Evaluation and Improvement of File System Performance for Hadoop Cloud Computing Platform

劉豐豪
國防大學管理學院
資管所副教授
lfh123@gmail.com

蘇品長
國防大學管理學院
資管所助理教授
spc.cg@msa.hinet.net

羅祥福
國防大學理工學院
國科所研究生
superalf@gmail.com

劉亞叡
國防大學管理學院
資管所研究生
duooreo@gmail.com

摘　要

由於 Hadoop 檔案系統處理巨量且大容量的架構設計，導致處理不一致容量大小檔案時，將產生效能問題。而 Hadoop 檔案系統效能問題將可透過虛擬化技術所提供高彈性、高擴充與高可用性的平台服務特性，因而獲得改善。然而，面臨客戶營運規模不斷變化與使用量的鉅變，雲端服務提供者所要面臨的挑戰之一就是在遵循資訊科技服務管理及變更控制管理原則的同時，如何建立評估雲端服務平台持續變更問題的制度，在工程技術面產出具體雲端系統平台調整的衝擊評估(Impact Analysis)，以綜合財務等其他面向的衝擊評估資訊，協助組織更有效率地執行衝擊評估審查，降低資訊架構變更所產生的風險，以增加雲端服務提供者的競爭能力。本研究提出結合組織變更控制管理，且考量技術面底層虛擬化技術與上層 Hadoop 檔案系統效能的跨服務層 Hadoop 效能變更評估機制。模擬結果顯示透過效能模型取得的模擬預估值與實際系統量測的數據極為相近，顯示本研究所提效能變更評估機制除可有效評估雲端系統資訊架構變更要求，有助於雲端服務提供者事先針對 Hadoop 檔案系統效能變更要求執行評估外，並可作為後續雲端系統架構變更執行的審查參考依據。

關鍵詞：Hadoop、HDFS、MapReduce、Performance Model、Change Control Management。

一、緒論

雲端運算（Cloud Computing）風暴已經來襲，這是一場正在改變使用電腦資訊的新革命。根據 NIST 定義，雲端服務區分為如下(表 1-1) [1]。

表 1- 1 雲端服務[1]

服務名稱	服務內容
基礎架構即服務（IaaS）	IaaS 將基礎設備整合起來，提供給個人、公司、機構等租用，減輕接購置與管理的成本。
內容平台即服務（PaaS）	可以提供雲端應用的開發境，但並不掌控作業系統、硬體或運作的網絡基礎架構。
軟體即服務（SaaS）	軟體開發人員或公司都可以自由揮灑創意，面對全世界的使用者提供各式各樣的軟體服務

其中，平台即服務（PaaS）是著重使用者存取各種雲端服務背後的執行環境，能提供高度的可彈性，以動態擴展應用執行時所需要的運行環境。Hadoop 是建置雲端平台的重要參考技術之一，其之所以備受矚目，主要是參考商業化平台 Google 背後的三種關鍵技術[2][3][4]，從而開發相對應的開源軟體，包含針對大規模資料處理而設計的 HDFS(Hadoop Distributed File System) [5]分散式檔案系統及大規模資料處理的 MapReduce [3]程式設計模型。然而，由於 Hadoop 檔案系統是專門設計用來處理巨量且大容量的架構，導致處理不一致容量大小檔案時，將產生效能上的問題。目前 Hadoop 檔案系統效能問題解決方式有哪些。如下表(1-2)[6][7][8][9][10]

表 1- 2 現有解決方案

名稱	*主要貢獻*	*考慮層面*	*檔案處理類型*
HAR(2009) [6]	合併壓縮檔案	Hadoop 層面	小檔案
Liu X (2009) [7]	利用地區相關性合併檔案	Hadoop 層面	小檔案
SequenceFile (2011)[8]	序列化壓縮	Hadoop 層面	小檔案
MapFile (2011)[9]	索引序列化壓縮	Hadoop 層面	小檔案
Zhang(2012) [10]	虛擬機負載平衡	VM 層面	大/小檔案
本研究	雲端資源效能整合模型	Hadoop 層面/VM 層面	大/小檔案

然而，虛擬化技術(Virtualization)有助於平台即服務(PaaS)提供者，提供高彈性(Flexibility)、高擴充(Scalability)與高可用性(Availability)的主機平台服務，如何結合虛擬化技術，改善 Hadoop 檔案系統效能也是一項值得探討的議題。

再者，面臨客戶營運規模與使用需求的不斷增加與鉅變，雲端服務提供者所要面臨的挑戰之一就是遵循資訊科技服務管理(Information Technology Service Management, ITSM)[11]變更控制管理

(Change Control Management)原則的同時，如何建立評估雲端服務平台持續變更問題的制度，在工程技術面產出具體雲端系統平台調整的衝擊評估(Impact Analysis)，以綜合財務等其他面向的衝擊評估資訊，協助組織更有效率地執行衝擊評估審查，降低資訊架構變更所產生的風險，以增加雲端服務提供者的競爭能力。

本研究提出一套可結合變更控制管理，且考量底層虛擬化技術與上層 Hadoop 檔案系統效能的跨層次效能變更評估機制。在因應巨量使用產生大量負載變動的情況下，首先將實際雲端系統特性轉換為各種效能模型，接著透過效能模擬的方式，計算出主要效能指標協助雲端服務提供者事先評估針對系統架構及效能負載的變更要求，最後產出可比較現有架構與架構變更後的效能評估結果，有效率地協助完成變更要求下系統面的衝擊分析。

二、文獻探討

1.Hadoop 檔案系統

在 Hadoop 的運作過程中，Hadoop Distributed File System(HDFS)(圖 2-1)是 Hadoop 的檔案系統。

HDFS 是一個易於擴充的分散式檔案系統，目的為對大量資料進行分析，運作於廉價的普通硬體上，又可以提供容錯功能，給大量的用戶提供總體性能較高的服務。

HDFS 的特性有硬體高容錯能力(Fault Tolerance)，和串流式的資料存取(Streaming data access)，並支援大規模資料集。

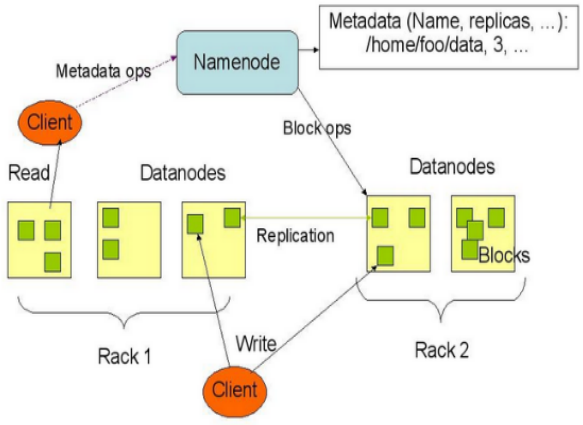

圖 2- 1 HDFS 架構[5]

HDFS 的運作架構[5]有兩個主要節點:

(1)名稱節點(NameNode)整個 HDFS 只有一個名稱節點，負責管理檔案系統的命名空間(namespace)，記錄所有檔案及目錄的 metadata，各項檔案屬性權限等資訊的管理及儲存，記錄檔案的各 Blocks 置放於哪些資料節點

(2)資料節點(DataNode)

可以多個資料節點，處理使用者存取 Block 的請求，並定時地回報 Block 狀態給名稱節點。

2. Hadoop 平行運算架構

Hadoop 的平行運算架構 MapReduce 的執行會影響到檔案系統的效能，MapReduce 是雲端運算的關鍵技術之一，可將要執行的問題，拆解成 Map 和 Reduce 二個部份的方式來執行，並且以達到分散運算平行處理的效果。有許多組織如 Google 或 Yahoo 均有採用的 MapReduce 類似的概念作為執行架構，而儲存的檔案系統也相仿 (Google 是 GFS，而 Yahoo 是 HDFS), 其中配合平行處的輸出/入資料切割,是決定架構效能的重要因素之一。

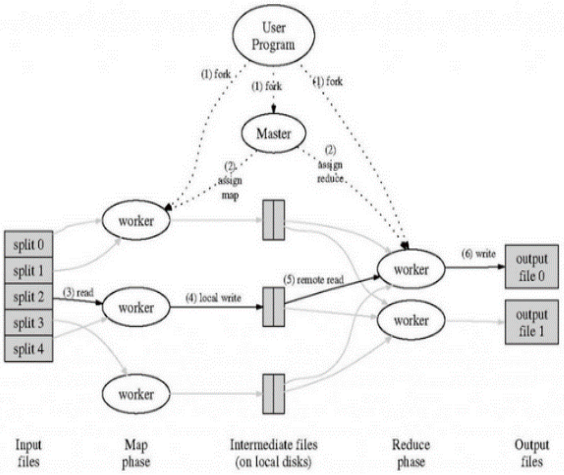

圖 2- 2 分散式運算執行架構[3]

MapReduce 的架構及運作的執行過程(如圖 2-2)[3]

(1)使用者將要執行的 MapReduce 程式複製到 Master 與每一臺 Worker 機器中。

(2)Master 決定 Map 程式與 Reduce 程式，分別由哪些 Worker 機器執行。

(3)將所有的資料區塊，分配到執行 Map 程式的 Worker 機器中進行 Map。

(4)將 Map 後的結果存入 Worker 機器的本地磁碟。

(5)執行 Reduce 程式的 Worker 機器，遠端讀取每一份 Map 結果，進行彙整與排序，同時執行 Reduce 程式。

(6)將使用者需要的運算結果輸出。

目前 MapReduce 的效能問題有 Data locality、Data movement、Data Shuffling 這三個問題分別發生在第三步驟到第五步驟。Data locality 的主要問題點是資料配置(第三步驟)，資料的配置要如何才能提高存取效率，把相同的資料配置在一起，可以使得遠程讀取(Remote Read)的成本降低，進而提高存取效率。Data movement 的主要問題點是預取(第四步驟)，資料要如何預取來提高存取效率，先將一些常用的資料預取在 memory 中，可以減少再去重複讀取 disk 的時間，Data Shuffling 的主要問題點是中介資料移動(第五步驟)，中介資料的移動是靠遠程存取到 reduce，所以如何控制好中介資料移動使得中介資料到 reduce 的流量控制良好，不會讓交通錯亂。

三、效能變更評估機制

1.運作架構

本研究提出一套可結合變更控制管理，且考量底層虛擬化技術與上層 Hadoop 檔案系統效能的跨層次效能變更評估機制。(圖 3-1)為本文的架構圖:

包含雲端系統(Cloud System)、效能變更評估(Performance Change Assessment)、變更控制管理(Change Control Management)。

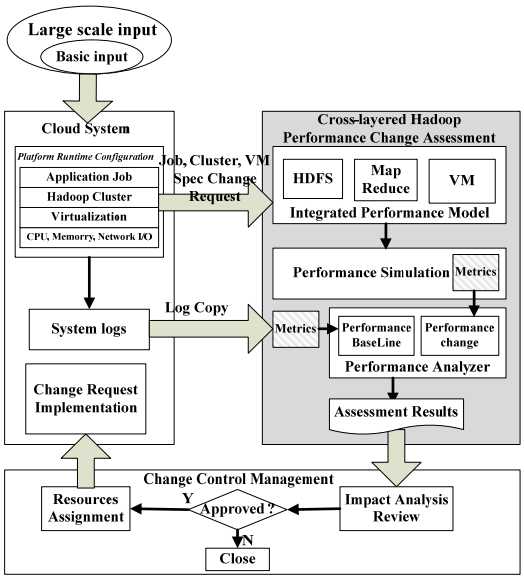

圖 3- 1 管理機制功能架構圖

(1)雲端系統

雲端運算的系統中，實體主機的硬體資源有:CPU、Memory、Network、I/O..等等，在這個硬體資源上架設虛擬化(virtualization)機器，而這虛擬機上面會搭載 Hadoop 雲端運算平台，透過 Hadoop 平台的機制來執行一些大量平行運算的處理，當然如果要執行一些功能，就要在 Hadoop 上運行程式。當大量的資料進來，需要變更控制時雲端系統，會傳送變更需求給效能變更評估。

雲端系統正常運作時會有一個收集 System Log 的機制，用來收集效能的訊息，例如: CPU、Memory、Network、I/O..等等的資料，有了這些系統資料，就可以當作修正 Hadoop 平台的重要資料。

(2)效能變更評估

效能變更評估是本論文著重的部分，效能變更評估有三個階段分別是效能模型(Performance Model)、效能模擬(Performance Simulation)、效能分析(Performance Analyzer)。

第一階段:效能模型主要是由三大種類型(HDFS、MapReduce、VM)所構成，並將從雲端系統中得到的變更需求參數，轉換成數學公式的參數。

第二階段:效能模擬是將所得到的數學參數，透過本文的效能模擬的公式建立一個模擬模型來模擬實際運作情況。

第三階段:效能分析是將從雲端系統所得到的 System Log，與效能模擬模型進行比較分析，產出評估報告。

(3)變更控制管理

是變更控制管理是依據 ITSM(ISO20000)變更管理機制，雲端化之後，服務的需求變得更多樣化了，所以服務管理能力就相當重要，例如:使用者對於雲端化後的問題，提出變更需求。如果沒有一個好的管理機制，服務的改善就會不如預期的順利。所以希望本文提出的效能評估報告可以和管理面的機制作結合。

2. 跨服務層 Hadoop 效能變更評估機制

(1)效能整合模型

效能模型中的參數有 HDFS、MapReduce、VM。三者效能關係非常密切，本研究是為了更加明確的說明效能參數，將參數整理分類後列表出來說明。HDFS 效能模型的參數(表 3-1)主要定義出於記憶體消耗和執行時間的指標值。

表 3- 1 HDFS 效能模型的參數

參數	參數說明
M_{NN}	Namenode 消耗的記憶體
N_{file}	執行的檔案數目
N_{map}	映射數目
M_α	Namenode 自身記憶體消耗
M_β	每個 Block 映射之後所產生的記憶體消耗
HBS	Block 大小
L_i	每個檔案的檔案長度
M_b	每個檔案的 Block 產生的記憶體消耗
M_r	每個檔案的副本產生的記憶體消耗
T_{HDFS}	HDFS 的執行時間
δ_{CN}	Client 端到 Namenode 端的請求時間
δ_{NC}	Namenode 端到 Client 端的回覆時間
δ_{CD}	Client 端到 Datanode 端的尋找時間
$\delta_{metadata}$	每個檔案所消耗的 metadata 時間
δ_{disk}	每個檔案所消耗的硬碟讀寫時間
$f_{network}$	每個檔案所消耗的網路傳輸時間

MapReduce 效能模型的參數(表 3-2)主要是定義出一個完整的任務(Job)(圖 3-2)關的參數值，分別為 Map 執行時間、Copy 執行時間、Sort 執行時間、最後是 Reduce 執行時間。

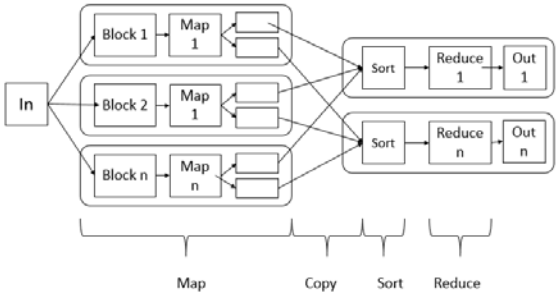

圖 3- 2 Job 任務

表 3- 2 MapReduce 效能模型的參數

參數	參數說明

T_j	作業執行時間
T_m	累計 Map 執行時間
T_c	累計 Copy 執行時間
T_s	排序的執行時間
T_r	Reduce 執行時間
T_{r_rep}	累計 Reduce 輸出的寫入時間
T_{r_proc}	累計 Reduce 處理時間
N_{m_wvs}	Map 數量
N_{m_slots}	在服務器中的 Map 同時執行數量
N_{r_slots}	在服務器中的 Reduce 同時執行數量
$N_{servers}$	叢集中的服務器數目
D_{blk_size}	資料區塊大小
N_{in_blks}	Map 輸入的資料塊數目
P_m	每台服務器核心平均 Map 吞吐量
P_r	每台服務器核心平均 Reduce 吞吐量
N_{racks}	叢集中的機架數
$N_{reduces}$	Reduce 數目
$N_{cp\_threads}$	每台 Reduce 節點複製的 I / O 速度
B_{local}	交換機內部連接的網路頻寬
B_{remote}	跨交換機連接的網路頻寬
S_{cp_thread}	理論上每個 Copy 執行緒的最佳 Copy 速度
S_{r_rep}	理論上每個 Reduce 最佳輸出的複製速度
S_{disk_write}	硬碟寫入速度
r_{map_io}	Map 輸出到輸入大小比率
r_{reduce_io}	Reduce 輸出到輸入大小比率
$N_{replicas}$	在 HDFS 副本數量
N_{rep_locals}	Local 機架副本數量
$N_{rep\_remotes}$	Remote 機架副本數量
$N_{sortpaths}$	合併排序的路徑複製數量
$D_{sortbuf}$	Copy 的排序緩衝區大小

VM 效能模型參數(表 3-3)是由雲端系統傳來的 System Log 所提供的參數值，可以知道變更控制前雲端系統狀況。

表 3- 3 VM 效能模型參數

參數	參數說明
$T_{response}$	回應時間
$IO_{throughputs}$	I/O 吞吐量
r_{cpu}	CPU 利用率
$tf_{network}$	網路流量

(2)效能模擬

本論文是由 Bo Dong(2012)[14]所提出的 HDFS 的計算式，透過下列的計算式可以清楚的知道，HDFS 的記憶體消耗量，和 HDFS 的運算時間。

$$M_{NN} = M_0 + N_{file} + (N_{file} + M_f) \cdot \sum_{i=1}^{N_{file}} \left\lceil \frac{L_i}{HBS} \right\rceil + M_s \cdots (1)$$

公式是 HDFS 的運算時間，透過此運算的運算元，可以推論出 HDFS 的運算時間。此運算式因子會用到 Client 到 NameNode 的請求時間，NameNode 到 Client 的回傳時間，最後是 Client 到 DataNode 的尋找時間，運算式因子裏頭，還有後設資料、硬碟、網路。這些因子可以影響到 HDFS 的執行時間

$$T_{HDFS} = N_{file}(\delta_{CN} + \delta_{NC} + \delta_{CD}) + \sum_{i=1}^{N_{file}} \delta_{metadata} + \sum_{i=1}^{N_m} \delta_{disk} + \sum_{i=1}^{N_m} f_{network}\left(\frac{L_i}{speed_i}\right) \cdots (2)$$

MapReduce 的效能指標，從公式(3)到(6)[15] 中可以透過參數值模擬出 Job 執行的時間。有 Map 運算時間、Copy 運算時間、Sort 運算時間、Reduce 運算時間。

$$T_m = \frac{D_{blk\_size}}{P_m} \times N_{m\_wvs} \cdots (3)$$

$$T_c = \max\left(\frac{MapOutSizeForLocalReduce}{LocalRackCopySpeed}, \frac{MapOutSizeForRemoteReduce}{RemoteRackCopySpeed}\right) \cdots (4)$$

$$T_s = \begin{cases} \dfrac{2 \cdot \left\lceil \frac{N_{reduce}}{N_{servers}} \right\rceil}{S_{disk_{write}}} \cdot \lambda_{N_{sortpaths}}, & \dfrac{N_{in\_blks} \cdot D_{blk\_size} \cdot r_{map\_io}}{N_{reduces} \cdot D_{sortbuf}} > 1 \cdots (5) \\ 0, & ohterwise \end{cases}$$

$$T_r = \max(T_{r\_proc}, T_{r\_rep}) \cdots (6)$$

將 HDFS 的指標與 MapReduce 的指標和 VM 指標，模擬出與實際相符的模型。

(3)效能分析

在本文的效能變更評估機制中，將 Log Copy 和效能評估比較，評估系統在效能變更前和變更後的效能。本文利用在效能模型所定義出來的參數，來分析 Log ，同時也用來分析效能模擬。個別取得變更前的基準，以及變更後的效能結果，來做比較。最後產生出效能結果，假設變更後的結果大於變更前的基準，也就是說在前面所提出的效能變更要求，是可以納入考慮的。這樣的做法是有效的，來提供給變更控制管理來做審查。

四、模擬結果分析與比較

1.模擬實驗環境

實驗環境(表 4-1)架設虛擬機，使用效能基準測試工具為 TestDFIO。來產生出不同資料量的 Input 來取得實際 Hadoop 系統所需的執行時間，再將效能模擬出的模型兩者比較，為了更符合實際情況，將透過一個雲端運算架構式影像資訊隱藏的應用做為測試個案，實測所得數據加以分析，以驗證效能變更評估機制的有效性。

表 4- 1 實驗環境

Server	CPU	Intel Core i5-2450M

		(2.5GHz)
	Memory	8GB
	Disk	500GB
Software	OS	Linux Ubuntu 12.04
	Hadoop	Hadoop 1.0.4
	Java	Sun JDK 6u21

下(圖 4-1)本文將實際狀況與效能模擬出來得 Map 相互比較。使用 Input 資料量為 50MB、100MB、150MB、200MB 為測試值，來測試執行時間，比較本文中效能模擬是否準確。

2.模擬結果分析與比較

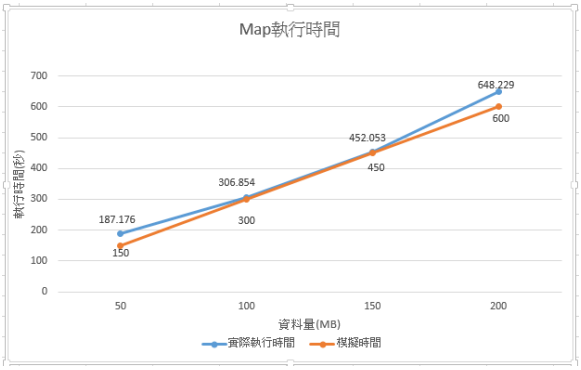

圖 4- 1 Map 執行時間

根據(圖 4-1)可以知道模型計算得到的效能數據與系統實測數據的差距不大，本研究效能模型可以仿真模擬出實際運作平台，協助建置雲端運算平台前的評估。

五、結論與未來研究

本研究提出一套可結合變更控制管理，且考量底層虛擬化技術與上層 Hadoop 檔案系統效能的跨層次效能變更評估機制。在因應巨量使用產生大量負載變動的情況下，首先將實際雲端系統特性轉換為各種效能模型，接著透過效能模擬的方式，計算出主要效能指標協助雲端服務提供者事先評估針對系統架構及效能負載的變更要求，最後產出可比較現有架構與架構變更後的效能評估結果，有效率地協助完成變更要求下系統面的衝擊分析。

模擬結果顯示效能模型模擬取得的預估值與實際系統量測的數據極為相近，顯示本研究所提效能變更評估機制中，代表實際雲端系統的效能模型，可用以模擬雲端系統因應巨量使用產生大量負載變動的情況，將有助於雲端服務提供者事先針對 Hadoop 檔案系統效能變更要求進行評估，作為後續雲端系統架構變更執行的審查參考依據。

致謝

感謝國科會對本研究之補助，計畫編號：NSC 102-2221-E-606-008。

參考文獻

[1] NIST. http://www.nist.gov/

[2] S. Ghemawat, H. Gobioff, and S.-T. Leung, “The Google file system,” in *Proceedings of the nineteenth ACM symposium on Operating systems principles*, New York, NY, USA, 2003, pp. 29–43.

[3] J. Dean and S. Ghemawat, “MapReduce: simplified data processing on large clusters,” *Commun. ACM*, vol. 51, no. 1, pp. 107–113, Jan. 2008.

[4] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber, “Bigtable: A Distributed Storage System for Structured Data,” *ACM Trans. Comput. Syst.*, vol. 26, no. 2, pp. 4:1–4:26, Jun. 2008.

[5] Hadoop Distributed File System (HDFS). http://hadoop.apache.org/hdfs/.

[6] Hadoop Archives. 2009, Hadoop archives guide. http://hadoop.apache.org/common/docs/current/hadoop_archives.html. (visit on 2013/3)

[7] X. Liu, J. Han, Y. Zhong, C. Han, and X. He, “Implementing WebGIS on Hadoop: A case study of improving small file I/O performance on HDFS,” in *IEEE International Conference on Cluster Computing and Workshops, 2009. CLUSTER ’09*, 2009, pp. 1–8.

[8] Sequence File. 2011, Sequence file. http://wiki.apache.org/hadoop/SequenceFile. (visit on 2013/4)

[9] MapFile. 2011, Mapfile api, http://hadoop.apache.org/common/docs/current/api/org/apache/hadoop/io/Map

[10] Z. Zhang, L. Xiao, Y. Li, and L. Ruan, “A VM-based Resource Management Method Using Statistics,” in *2012 IEEE 18th International Conference on Parallel and Distributed Systems (ICPADS)*, 2012, pp. 788–793

[11] ITSM.http://en.wikipedia.org/wiki/IT_service_management

[12] ITIL. http://www.itil-officialsite.com

[13] ISO20000.http://en.wikipedia.org/wiki/ISO/IEC_20000

[14] B. Dong, Q. Zheng, F. Tian, K.-M. Chao, R. Ma, and R. Anane, “An optimized approach for storing and accessing small files on cloud storage,” *Journal of Network and Computer Applications*, vol. 35, no. 6, pp. 1847–1862, Nov. 2012.

[15] J. Han, M. Ishii, and H. Makino, “A Hadoop performance model for multi-rack clusters,” in *2013 5th International Conference on Computer Science and Information Technology (CSIT)*, 2013, pp. 265–274.