分散式檔案系統 Hadoop 與 Ceph 效能比較

卓志遠　楊朝棟*　張廣欽　李偉業
東海大學 資訊工程學系
cyucho@gmail.com, { ctyang, erick , g10035708 }@thu.edu.tw
*通訊作者

摘要

雲端運算[1]是近年來很熱門的領域，借由硬體效能的提升與軟體應用的密切結合，在產官學各界的帶動下，使得雲端服務越完整也越多元，已經到無所不在的境界了。電腦系統的運作主要為程式邏輯與資料儲存，雲端運算的架構改變了程式設計與執行的模式，也改變了資料存取的檔案系統架構，Google 能夠在不到 1 秒的時間內搜尋非常大量的資料[2]、Gmail 能夠對世界 4.25 億用戶[3]，每個用戶提供 15GB 的大量郵件儲存空間，Facebook 有 10 億個註冊用戶[4]，每天產生 300TB 以上的資料量[5]，Dropbox 提供每個使用者最少 2GB 的雲端儲存空間，以上這些雲端服務的例子，沒有使用分散式運算與分散式儲存系統，無法滿足使用者的需求。本論文使用開放原始碼，比較當前較著名的分散式檔案系統 Hadoop 與 Ceph，針對這兩個系統的檔案上傳與下載能力，大小檔案的傳輸能力與容錯能力做比較分析，並提出個別系統的特性及優缺點，讓需要建置分散式檔案系統或有興趣從事與分散式檔案系統相關的研究者一個參考依據。

關鍵詞：雲端運算、分散式檔案系統、Hadoop、Ceph、雲端儲存。

Abstract

Cloud computing is a very popular area in recent years, by enhancing the performance of the hardware and software applications closely, in industry, government and academic communities, driven, making cloud services more complete and the more diverse, has come to the omnipresent realm. Mainly for the operation of the computer system program logic and data storage, cloud computing architecture changed the programming and execution model, also changed the data access file system architecture, Google in less than one second of time searching for a very large number of information, Gmail can the world 425 million users, each providing a large number of e-mail storage space 15G, Facebook has one billion registered users, more than 300TB per day the amount of data generated, Dropbox offers each user at least 2G cloud storage space, examples of these cloud services without the use of distributed computing and distributed storage system, unable to meet the needs of users. This paper will be the premise of open source, compare the current better known Hadoop Distributed File System and Ceph, for these two systems file upload and download capabilities, the size of the file transfer capability and a comparative analysis of fault tolerance, and propose individual system characteristics, advantages and disadvantages, so need to build a distributed file system, or are interested in working with distributed file system-related researchers a reference.

Keywords: Cloud Computing, Distributed File System, Hadoop, Ceph, cloud storage.

1. 前言

網路的出現促使分散式運算的迅速發展，從 1990 年代為了解決大量計算問題而採用的網格計算開始，到現今的雲端運算，而海量資料需要的大量儲存空間是雲端運算平台中的一個重要議題，分散式檔案系統[6,7]是儲存與分析海量資料的當然選擇，根據 IDC(International Data Corporation)的研究報告，2011 年全球數位資料的使用量約為 1.8 ZB，並預測 2020 年的總量為 35.2 ZB 之多[8]。

分散式檔案系統具備高效能、高容錯、高可靠、高可用與高擴充的特性。它除了將資料分散的儲存在資料節點外，每個節點同時也具備分散式計算的能力，所以原本只能在單一主機處理的程式邏輯與資料處理，可以分散到數以萬計的大量計算節點上運行處理，等每個節點計算出結果後，再將處理完成的個別結果結合起來，產生最終的結果，這種運作方式有效率的節省了資料存取與計算的排隊時間，提升了處理效能，例如 Google 搜尋在關聯式資料庫中無法達到這樣的效能與精確度。

若把兩顆硬碟中的其中一顆硬碟故障的機率，應當是一顆硬碟的兩倍，依照這種理論，上千顆硬碟會故障的機率是非常高的，以前的系統把這種情形視為異常，並且利用各種機制如 RAID，HA 來做保護措施，而近期的分散式檔案系統卻把硬體故障視為式系統運作中一個正常的現象，是無法避免的，並且使用複本的機制來解決硬體故障後對整個系統造成的影響，讓使用者完全感覺不到系統問題而能持續正常的提供服務。

分散式檔案系統另一個優點是高度的可擴充性，可以在任何時候增加節點，插上網路線，設定加入整個檔案系統，就能夠很容易的加入並且提供服務。基於分散式檔案系統的優點與重要性，未來

會越來越多系統採用，希望本論文對分散式檔案系統有興趣的使用者能有所助益。

2. 背景

2.1 分散式檔案系統

分散式檔案系統需由三台以上的電腦組成，其中至少有一台管理節點與兩台資料節點，各節點間透過網路連接，使用軟體管理各節點，讓多台用戶端的使用者可以共用檔案和儲存空間，某些大型的分散式檔案系統甚至多達上萬個資料節點。

分散式檔案系統的用戶端並不是直接存取實體硬碟所分割的資料區塊，而是透過網路，使用該分散式檔案系統提供的通訊協定來存取資料。

分散式檔案系統具備資料複製與容錯的功能，即使再多個節點中有某一小部份的節點失效而離線，整個檔案系統仍然可以持續運作而不會造成資料遺失，用戶端也幾乎不會感覺任何異狀。

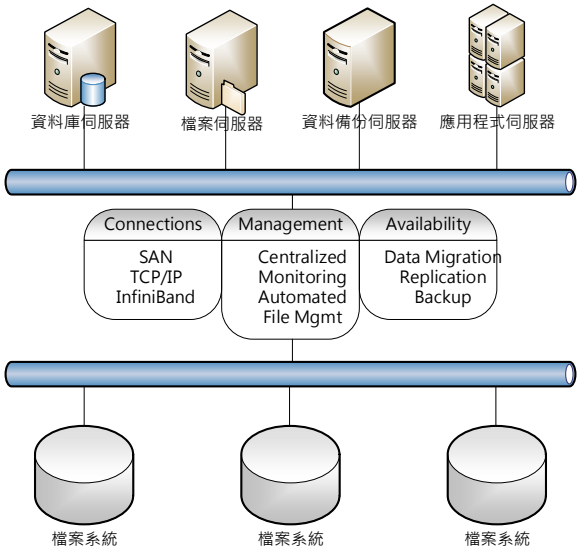

圖 1 分散式檔案系統架構圖

Hadoop[9]是 Apache 軟體基金會所研發的開放式程式碼平行運算的程式撰寫工具和分散式檔案系統，根據 Google 所發表的 MapReduce 與 Google 檔案系統的論文發展而成，Hadoop 的主要元件有 Hadoop Common，HDFS 與 MapReduce。

在 0.20 及之前的 Hadoop 版本，Hadoop Common 包含 HDFS 與 MapReduce，從 0.21 版本開始，HDFS 與 MapReduce 就分離出來為獨立的項目，HDFS 為 Hadoop 的分散式檔案系統，MapReduce 為 Hadoop 的平行計算框架，本論文主要針對 HDFS 進行研究實驗。

西元 2008 年 2 月 19 日，雅虎使用 1 萬個處理器核心的 Linux 電腦叢集安裝 Hadoop 應用程式，另有 Facebook、IBM 等也都使用 Hadoop 為其資料儲存系統。

Ceph[10,11]剛開始是加州大學 Santa Cruz 分校的 Sage Weil 所設計關於自由軟體分散式檔案儲存系統的博士研究論文。在 2007 年畢業之後，便開始全心投入 Ceph 的開發，期望能將 Ceph 在生產環境中使用。它具有可擴充到 PB 及容量、高效能、高可靠性與容錯的特性。Ceph 的檔案系統包含 Ceph Client、Ceph Metadata Daemon、Ceph Object Storage Daemon 及 Ceph Monitor 四個主要的角色，表 1 針對 Hadoop 與 Ceph 的各項特性做比較。

表 1 Hadoop 與 Ceph 的比較

	Ceph	Hadoop
Metadata server	可以擴充多個 MDS，不存在單點故障和瓶頸。	存在單點故障的問題。Namenode 是一個中心伺服器，負責管理檔案系統的命名空間以及用戶端對檔案的存取。
FUSE[12]	支援	支援
介面	POSIX[13]	不完全
資料分佈	檔案被分割，每個資料區塊是一個物件。物件保存在不同的存儲節點伺服器上。	預設的資料區塊大小是 64MB。因而，HDFS 中的檔案是按照 64M 被切分成不同的資料區塊，每個區塊儘可能地存儲於不同的 Datanode 中。
副本	資料複製	資料複製
資料可靠性	由資料的多副本特性提供可靠性	由資料的多副本特性提供可靠性
資料節點故障恢復	當節點失效時，自動遷移資料並重新複製副本。	心跳信號檢測機制。每個 Datanode 節點週期性地向 Namenode 發送心跳信號，當節點失效時，自動遷移資料並重新複製副本。
MDS 故障恢復	當節點失效時，自動遷移資料並重新複製副本。	Namenode 是 HDFS 集群中的單點故障所在。如果 Namenode 機器故障，是需要手工干預的。目前自動重啟或在另一台機器上做 Namenode 容錯移轉的功能還沒實現。
擴充性	可以增加中繼資料伺服器和存儲節點。容量、檔案操作性能與中繼資料操作性能皆可擴充。	在一個集群裡可擴充到數千個節點。一個單一的 HDFS 實例應該能支撐數以千萬計的檔案。
安裝佈署	簡單	簡單
開發語言	C++	Java
實用時機	小檔案	不限
檔案系統規模	中型	大型

除了 Hadoop 與 Ceph 兩個檔案系統外，GlusterFS 也是一個很有名的檔案系統，GlusterFS 是一個由 Red Hat 贊助，開放原始碼的分散式檔案系統，通常可搭配網頁應用程式存取或網路芳鄰作

為後端資料儲存空間，他能夠擴充到 PB 等級的容量，而且能夠處理上千個用戶端請求。GlusterFS 叢集可以在 Infiniband RDMA 或 TCP/IP 連結上建立儲存區塊，能夠在單一的全域命名空間中管理磁碟空間與記憶體。GlusterFS 是一個可堆疊的使用者端空間設計，所以他能夠針對不同的工作負載提供高效率的性能，除了具有分散式儲存檔案複製的優點，還有資料鏡像及容錯的特性。

2.2 雲端儲存

雲端儲存[14]是將資料儲存在網路上，資料中心營運商提供網路空間主機代管服務，企業需求者只需針對企業內部的需求規劃，向資料中心營運商購買或租任網路儲存空間，而資料中心營運商針對使用者的需求規劃儲存資源，如此客戶便可自由的運用這些儲存空間，而這些儲存空間可能被分佈在多部的磁碟伺服器主機上。

採用雲端儲存的企業，只需要負擔租賃實際儲存空間的費用，節省購置儲存設備的成本及日常的維護工作，能夠讓企業專注在其核心業務上。

著名的雲端儲存服務有 Dropbox、Google 雲端硬碟、蘋果 iCloud、國內的華碩雲端與資策會等，這些雲端儲存服務提供大空間的儲存滿足使用者的需求，是近年來一個很大的突破。

3. 系統設計與實作

3.1 Hadoop 的 HDFS 架構

Hadoop 的檔案系統稱為 HDFS[15,16]，是主從架構的設計，在 HDFS 的叢集中包含至少一個名稱節點和兩個以上的資料節點，名稱節點主要管理檔案系統的名稱空間和協調存取使用者端的檔案。他是一種塊狀結構的檔案系統，將單一個檔案分割成許多固定大小的區塊，區塊是指一次讀取或寫入的最小資料量，然後將這些區塊分散的儲存在許多個資料節點中，HDFS 預設的區塊單位是 64M，這比傳統的檔案系統大上許多，傳統的檔案系統通常是 128KB，HDFS 的區塊單位會比傳統檔案系統大的原因是架構不同，傳統檔案系統的檔案都放在同一塊硬碟，對於檔案都搜尋時間會小很多，所以分散式檔案系統為了能夠加快搜尋的速度，需要將檔案區塊設定為較大的區塊，採用區塊儲存可以讓要儲存的檔案比單一節點的儲存空間還大，而且將同一個檔案區塊分別存放在不同的資料節點中，另外以區塊為單位很適合用來做檔案複製，一般重要的檔案會設定複製成三份，如此可避免區塊、磁碟或資料節點的單點故障。

HDFS 的架構，主要區分為 Namenode 與 Datanodes 兩部分，Namenode 存放 Metadata，Datanodes 則是實際存放資料區塊的地方，用戶端寫入資料到 HDFS 時，系統會將要儲存的檔案切割成固定大小的區塊，再將區塊資料儲存到檔案系統內，HDFS 會區分 Rack，將相同的資料區塊儲存在不同的 Rack 中，這種複製的功能來確保資料因單點故障而遺失。用戶端要從 HDFS 內讀取資料時，會先向 Namenode 詢問所要取得的資料放在哪些節點的哪個位置，Namenode 將存放資料的資訊送給用戶端後，再由用戶端向 Datanode 取得資料。

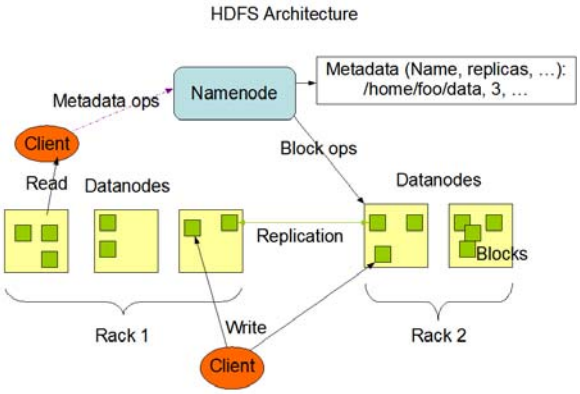

圖 2 Hadoop 架構圖[17]

3.2 Ceph 的 Ceph FileSystem 架構

Ceph 主要組成元素有三個，分別是 Ceph OSD Daemon、Ceph Monitor 與 Ceph Metadata Server。Ceph OSD Daemon 主要負責儲存資料，處理資料複製、復原、資料回填與重新調整，並提供給 Ceph Monitors 一些監測資訊來檢查其他的 Ceph OSD Daemons 節點是否還存活，Ceph 需要至少兩個 ODS 才能運作。Ceph Monitor 監視整個 Ceph 系統的運作狀態。Ceph Metadata Server 紀錄資料存放在 OSD 的位置，用戶端透過 MDS[18]才能取得資料正確的所在位置。

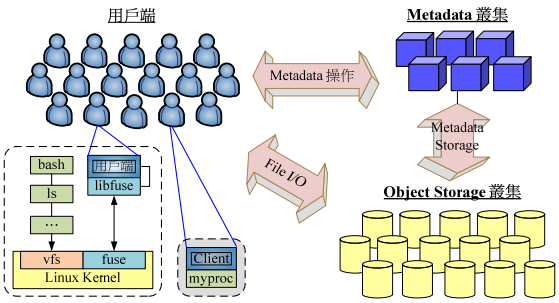

圖 3 Ceph 架構圖[19]

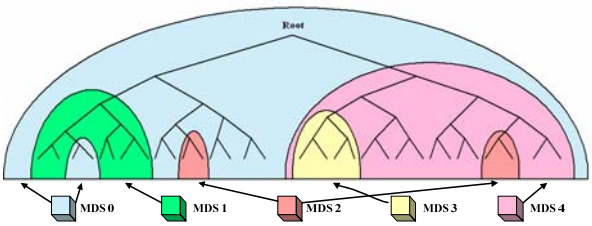

圖 4 Ceph 的名稱空間分區對應至 Metadata 伺服器[20]

4. 實驗方法與結果

4.1 實驗環境

Hadoop 與 Ceph 兩個分散式檔案系統的實驗環境列示在表 2 與表 3，為求實驗的公平性，所有硬體使用一樣的規格，作業系統則全部採用 Ubuntu 12.04 LTS 64 位元伺服器版本。Hadoop 本版為 hadoop-1.1.2、Ceph 版本為 Ceph version 0.61.7。

表 2 Hadoop 系統硬體規格

Hadoop	CPU	RAM	DISK	NETWORK
Namenode	2 顆 2 核	2GB	120GB	1Gbps
Datanode	2 顆 2 核	2GB	120GB	1Gbps

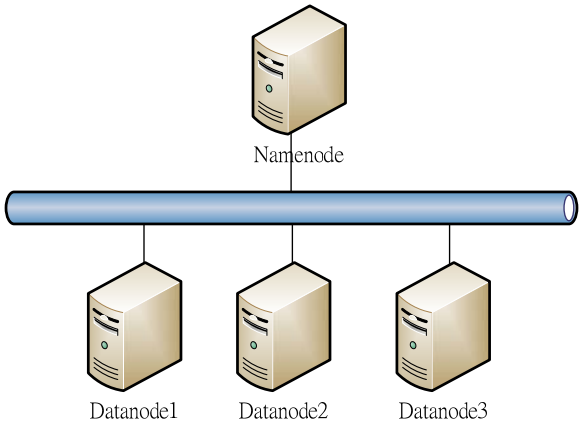

圖 5 Hadoop 實驗環境架構

表 3 Ceph 系統硬體規格

Ceph	CPU	RAM	DISK	NIC
MDS	2 顆 2 核	2GB	120GB	1Gbps
OSD	2 顆 2 核	2GB	120GB	1Gbps
MON	2 顆 2 核	2GB	120GB	1Gbps

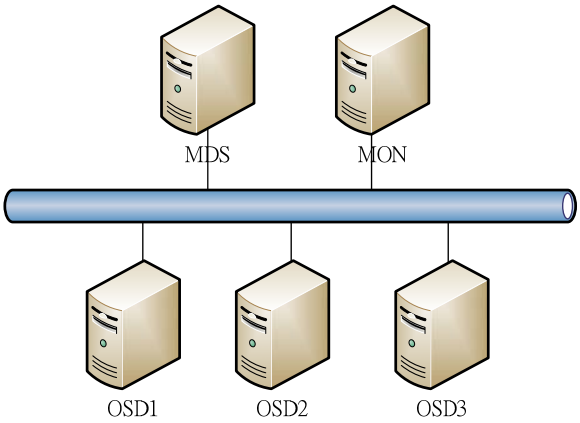

圖 6 Ceph 實驗環境架構

4.2 實驗方法

本實驗分別將 Hadoop 與 Ceph 兩個檔案系統架設完成，利用 time 指令來記錄檔案上傳與下載所需的時間，實驗如下三種狀況下的存取效能。

在用戶端建立單一個檔案，在檔案容量大小分別為 1MB、2MB、4MB、8MB、16MB、32MB、64MB、128MB、256MB、512MB、1GB、2GB、4GB、8GB、16GB、32GB 的上傳與下載效能測試。

小檔案存取實驗，在相同總合容量大小為 4GB，檔案大小分別為 10MB、20MB、30MB、40MB、50MB、60MB、70MB、80MB、90MB、100MB 的上傳與下載效能測試。

大檔案存取實驗，在相同總合容量大小為 4GB，檔案大小分別為 100MB、200MB、300MB、400MB、500MB、600MB、700MB、800MB、900MB、1000MB 的上傳與下載效能測試。

4.3 實驗結果

圖 7 的實驗是不同大小的檔案從用戶端上傳到分散式檔案系統所花費的時間，由圖中可明顯看出，在檔案小於 512MB 時，兩種檔案系統所花費的時間沒有明顯的差距，在 512MB 以後，很明顯的可看出 Ceph 上傳所花費的時間，幾乎比 Hadoop 將近多出一倍的時間，所以由此圖可看出，Hadoop 在處理檔案儲存的效能上是優於 Ceph 的。

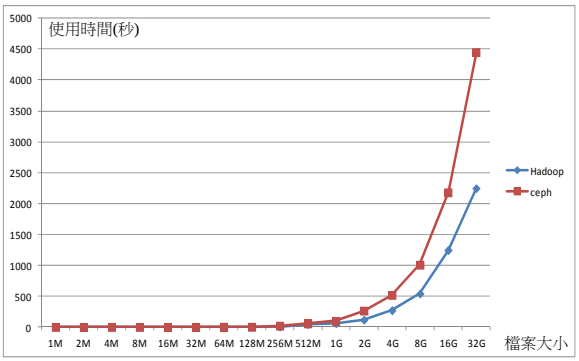

圖 7 Hadoop 與 Ceph 不同檔案大小上傳時間比較

圖 8 的實驗是不同大小的檔案從分散式檔案系統下載到用戶端所花費的時間，由圖中可看出，在 4GB 之後，Hadoop 才有些微的效能優於 Ceph。

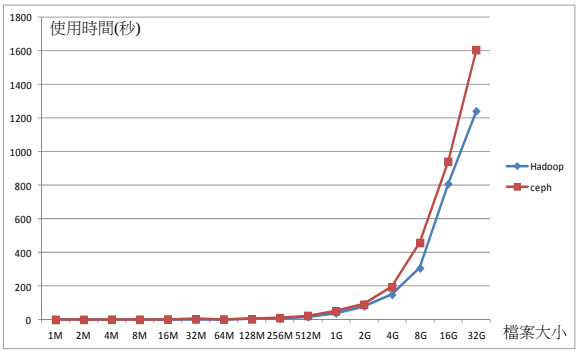

圖 8 Hadoop 與 Ceph 不同檔案大小下載時間比較

圖 9 的實驗是在檔案總容量為 4GB 的前提下，將 4GB 切割為大小 10MB 至 100MB 的小檔，在同時間將這些檔案上傳，這個實驗想要測試在不同檔案大小時對檔案系統的效能是否有影響，由圖中可看到折線圖呈現些微弧狀，檔案大小為 50MB 至 60MB 使用時間最少。

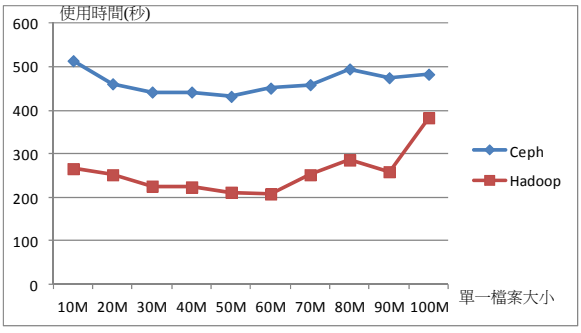

圖 9 Hadoop 與 Ceph 相同 4GB 容量，不同小檔案之檔案大小上傳時間比較

圖 10 的實驗是在檔案總容量為 4GB 的前提下，將 4GB 切割為大小 10MB 至 100MB 的小檔，在同時間將這些檔案下載，這個實驗想要測試在不同檔案大小時對檔案系統的效能是否有影響，由圖中可看到兩條折線圖呈現不規則形狀，Hadoop 在 30MB、40MB 與 90MB 的效能比較好，而 Ceph 在 20MB、70MB、80MB 與 100MB 的效能比較好，之前的實驗都是 Hadoop 的表現比較好，可是這個實驗在 80MB 的時候反而 Ceph 的表現比 Hadoop 好。

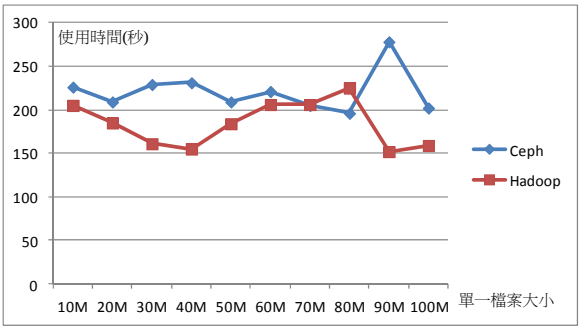

圖 10 Hadoop 與 Ceph 相同 4GB 容量，不同小檔案之檔案大小下載時間比較

圖 11 的實驗是在檔案總容量為 4GB 的前提下，將 4GB 切割為大小 100MB 至 1GB 的大檔，在同時間將這些檔案上傳，這個實驗想要測試在不同檔案大小時對檔案系統的效能是否有影響，由圖中可看到 Ceph 在 500MB 與 600MB 時會有很明顯花費約多二分之一的時間。

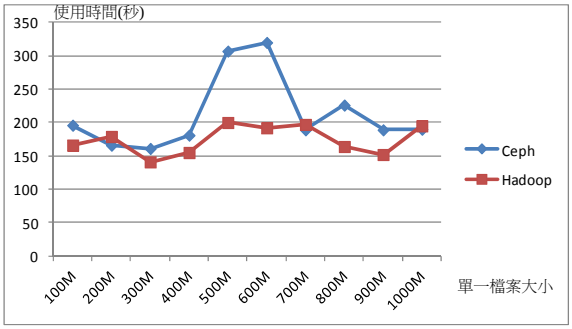

圖 11 Hadoop 與 Ceph 相同 4GB 容量，不同大檔案之檔案大小上傳時間比較

圖 12 的實驗是在檔案總容量為 4GB 的前提下，將 4GB 切割為大小 100MB 至 1GB 的大檔，在同時間將這些檔案下載，這個實驗想要測試在不同檔案大小時對檔案系統的效能是否有影響，由圖中可看到 Hadoop 的表現明顯優於 Ceph。

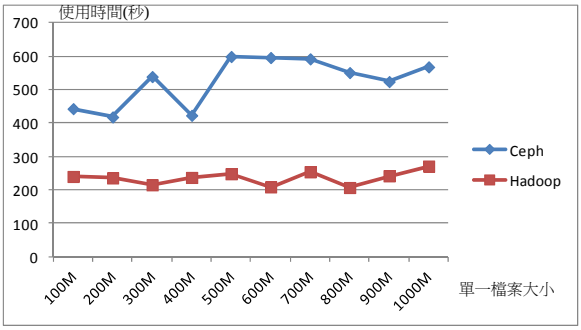

圖 12 Hadoop 與 Ceph 相同 4GB 容量，不同大檔案之檔案大小下載時間比較

5. 結論

經由這次的實驗發現，多家廠商已經採用的 Hadoop 系統在效能上的表現比較出色，儘管在某些情況下會比 Ceph 差，但是在 60 次的實驗中，Ceph 只能夠明顯的 2 次效能比 Hadoop 好，這樣的結果證明 Ceph 雖然有高可靠度、高擴充性的特點，但當這些特點在分散式檔案系統已經變成是必備的條件時，Ceph 勢必必須在檔案的存取效能上有所提升，才能夠符合使用者的需求。

架設分散式檔案系統需要許多電腦、重複的安裝步驟，重複的安裝測試，並且得跟上系統版本的更新，是一件需要耐心的事，在實驗中我儘量將所有的軟硬體規格都配置相同，以減少實驗的誤差，希望未來能針對檔案區塊的大小及檔案複製數量對存取效能的影響再進行實驗。

致謝

本論文之成果為行政院國家科學委員會研究計畫(計畫編號：NSC101-2622-E-029-008-CC3 與 NSC 101-2218-E-029-004-)補助。

參考文獻

[1] Z. Shuai, Z. Shufen, C. Xuebin, and H. Xiuzhen, "Cloud Computing Research and Development Trend," in Future Networks, 2010. ICFN '10. Second International Conference on, 2010, pp. 93-97.

[2] Google Search, http://www.google.com.tw/.

[3] Business Next 數位時代:Gmail 活躍用戶數突破 4.25 億人, http://www.bnext.com.tw/focus/view/cid/103/id/23748.

[4] 維基百科: Facebook, http://zh.wikipedia.org/wiki/Facebook.

[5] TechOrange 科技報橘:當你和 FB 一樣每天要處理 300TB 的資料，你就會知道虛擬化技術的重要！, http://techorange.com/2013/04/16/why-virtual-machine-is-so-attractive/.

[6] Jin Xiong, Yiming Hu, Guojie Li, Rongfeng Tang, and Zhihua Fan, Metadata Distribution and Consistency Techniques for Large-Scale Cluster File Systems, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL.22, NO.5, MAY 2011

[7] H. C. Chao, T. J. Liu, K. H. Chen, and C. R. Dow, "A Seamless and Reliable Distributed Network File System Utilizing Webspace," in Web Site Evolution, 2008. WSE 2008. 10th International Symposium on, 2008, pp. 65-68.分散式檔案系統

[8] J. Shafer, S. Rixner, and A. L. Cox, "The Hadoop Distributed Filesystem: Balancing Portability and Performance," in Performance Analysis of Systems & Software (ISPASS), 2010 IEEE International Symposium on, 2010, pp. 122-133.

[9] Hadoop 官方網站, http://hadoop.apache.org/

[10] Ceph 官方網站, http://ceph.com/

[11] Sage A. Weil, Scott A. Brandt, Ethan L. Miller, Darrell D. E. Long, Ceph: A Scalable, High-Performance Distributed File System, 2007

[12] A. Ismail and L. Shannon, "FUSE: Front-End User Framework for O/S Abstraction of Hardware Accelerators," in Field-Programmable Custom Computing Machines (FCCM), 2011 IEEE 19th Annual International Symposium on, 2011, pp. 170-177.FUSE

[13] N. R. Reizer, G. D. Abowd, B. C. Meyers, and P. R. H. Place, "Using Formal Methods for Requirements Specification of A Proposed POSIX Standard," in Requirements Engineering, 1994., Proceedings of the First International Conference on, 1994, pp. 118-125.POSIX

[14] 黃智霖，「實作一個雲端計算上具資源監控的分散式資料儲存系統」，私立東海大學，碩士論文，民國一〇〇年七月。

[15] Mackey, G., Sehrish, S., Jun Wang,, "Improving metadata management for small files in HDFS", Cluster Computing and Workshops, 2009.CLUSTER '09. IEEE International Conference, pp. 1-4, 2009.

[16] 廖本加，「於雲端計算環境上高可用性儲存系統之實作」，私立東海大學，碩士論文，民國一〇〇年七月。

[17] HDFS Architecture Guide, http://hadoop.apache.org/docs/stable/hdfs_design.html.

[18] J. H. Yun, Y. H. Park, S. J. Lee, S. M. Jang, and J. S. Yoo, "Design and Implementation of A Non-Shared Metadata Server Cluster for Large Distributed File Systems," in Computer Science and its Applications, 2008. CSA '08. International Symposium on, 2008, pp. 343-346.MDS

[19] Ceph Architecture, http://ceph.com/docs/next/architecture/

[20] S. A. Weil, S. A. Brandt, E. L. Miller, D. D. E. Long, and C. Maltzahn, "Ceph: A Scalable, High-performance Distributed File System," in In Proceedings of the 7th Symposium on Operating Systems Design and Implementation (OSDI 2006), 2006, pp. 307-320.