

多項運算任務之高效率 MapReduce 架構

李清晏
國家中山科學研究院
航空研究所
副研究員
chingyen@ms2.hinet.net

楊禮彰
國家中山科學研究院
航空研究所
副研究員
calsard@hotmail.com

曾中文
國家中山科學研究院
航空研究所
助理研究員
demtek@gmail.com

黃英軒
國家中山科學研究院
航空研究所
研發替代役
cenahyh7409@gmail.com

摘要

MapReduce 分散式計算架構可分解一項運算任務成多項運算小任務，並分別交由數台電腦同時進行運算作業，以減少整體運算時間，然而，其未分析多項任務間存在高度關聯性，故本研究主要分析多項任務間變數重複性，以大幅地減少運算成本。實驗分析顯示所提方法可有效應用於六分量氣動力係數轉換模型與產品之保固參數計算模型。

關鍵詞：MapReduce、六分量氣動力係數轉換模型、保固參數計算模型

Abstract

MapReduce can divide a calculation task into several sub-tasks and deliver them into more computers to figure out digits, thereby reducing the entire computation time. However, it does not analyze the relationship among several tasks to cause more computation costs. Consequently, this paper analyzed the repeatability of parameters of multi-tasks to significantly reduce the computation costs. Experimental analysis showed that the proposed method can be used effectively in the six aerostatic force coefficients transformation and the production's warranty coefficients calculation model.

Keywords: MapReduce, Computation costs, Six aerostatic force coefficients transformation, Warranty coefficients calculation

1. 前言

隨著資訊科技蓬勃發展，公家機關或民間企業可藉由各式各樣硬體設備擷取大量數據，例如：感測器(Senor)蒐集氣流壓力參數、

控制器區域網路 (Controller Area Network, CAN) 量測系統等，然而，使用者無法直覺地從大量數據中發現資訊，故如何將大量原始數據昇華至高價值資訊已成為十分重要議題。

至今，單一電腦無法儲存和分析大量數據，故專家學者們提出雲端運算 (Cloud Computing) 技術，其可有效串聯數千台電腦，令所有電腦同時執行各項任務，以大幅地縮短運算時間，且可依資料數量增加進一步擴充電腦數量，綜合上述所言，雲端運算技術具有優異資料處理效能、可擴充性及低廉建置成本。

專家學者已提出不少雲端運算技術，例如：Hadoop、Spark 等，其中 Hadoop 核心架構為 Hadoop 分散式檔案系統 (Hadoop Distribution File System, HDFS) 與 MapReduce 分散式計算技術，前者主要分割與備份資料，再分散儲存於不同電腦，以避免資料毀損，具體而言，HDFS 為一個主從式架構 (Master-Slave) 之檔案系統，其 NameNode 主機管理 DataNode 機器上資料。當使用者讀取資料時，NameNode 主機從最鄰近使用者之 DataNode 設備取出資料。

另一方面，Hadoop 中 MapReduce 分散式計算技術依每部電腦儲存容量大小分解資料，且每部電腦同時計算數據，再由其中一台電腦彙總其它電腦之計算結果，故執行時間遠比單一電腦少，換言之，單一電腦須逐一計算每筆數據，而 MapReduce 架構中每台電腦僅須計算部份資料，因此，MapReduce 有較高運算效能。

雖然 MapReduce 為高效率分散式計算架構，然而，其未考量多項運算任務間存在變數計算重複性，故本研究進一步分析該特性與巧妙地使用四則運算特性，以大幅地減少資料重複計算次數，再有效地分散資料於各個電腦計算，因此，本研究提方法十分適合應用於六分量氣動力係數轉換模型[1]與保固參數計算模型[2]。

2. MapReduce

Dean 與 Ghemawat[3, 4]提出 MapReduce 分散式計算架構，其分解運算任務成數個子任務，並將子任務代入多部電腦同時進行運算，最後，彙總每部電腦所計算結果，以實現高效能運算任務。由於每台電腦同時處理運算任務，故其整體運算時間遠低於單台電腦處理任務時間。

首先，該方法將資料分解成數個小分割 (Splits)，再代入 Map 函數，以排序 (Sorting)、複製 (Duplication) 及合併 (Merge) 資料，其優點為多部電腦同時處理資料，且避免單台電腦因巨量資料重複執行 I/O 問題。完成 Map 函數

後，代入多部電腦所計算結果於 Reduce 函數進行彙總，即完成運算任務。

以空用載具為例(如圖 1)，假設資料庫所記錄更換零件共計 9 件，且電腦數量為 3 台，則每 3 個零件為一組代入各個電腦計算，以統計零件更換次數，第一台電腦統計更換零件的結果為{發動機：2 次}、{火星塞：1 次}；第二台電腦計算更換零件次數，其結果為{發動機：1 次}、{火星塞：1 次}及{化油器：1 次}；第三台電腦統計各零件更換數量結果為{火星塞：2 次}、{化油器 1 次}，最後，彙總各統計結果，即得到每項零件更換總次數。

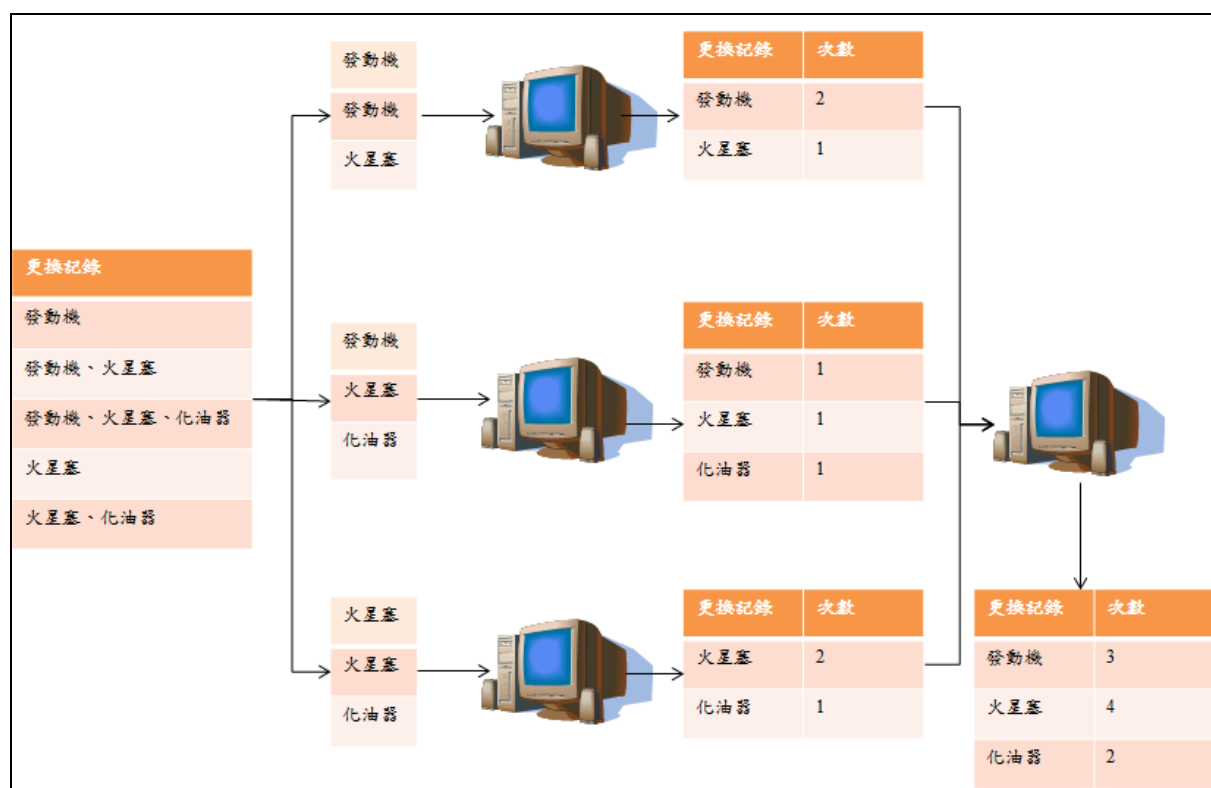


圖 1. MapReduce 流程圖

3. 所提方法

MapReduce 分散式計算架構可分解巨量資料與運算公式，並交付多部電腦執行運算作業，以減少運算時間。我們發現該特性十分適合應用於產品保固參數計算模型與六分量氣動力係數轉換模型，如第四節所示。

圖 2 顯示所提方法流程圖，其中我們可依連乘符號 $\prod$ 與累加符號 $\sum$ 改寫公式，其改寫後公式滿足四則運算定義，換言之，改寫後公式計算結果等於原始公式計算結果，接下來，刪

除相同的小任務，以減少資料重複計算問題，再分別交付各台電腦同時進行運算任務，以減少整體運算時間，最後，指定其中一台彙總結果。

值得一提的是，若某台電腦運算任務之連續加法或乘法數量過多時，則該項任務應再次使用所提方法分解成數個小任務，以避免電腦效能瓶頸問題(Bottleneck)，即完成任務電腦須等待其它未完工電腦，才能進行下一階段工作，換言之，所提方法具有階層性概念，可有效避免瓶頸問題。

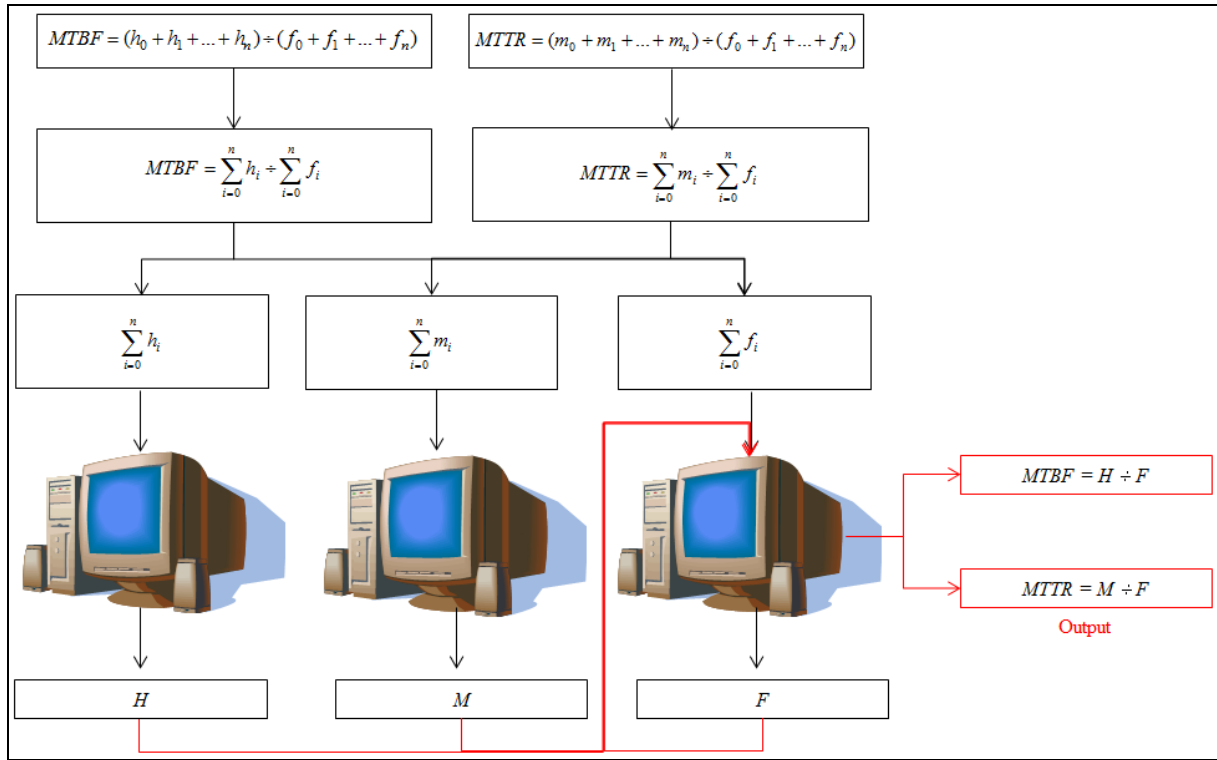


圖 2. 所提方法流程圖

4. 應用暨實驗分析

本節主要分析產品保固參數計算模型[2]與六分量氣動力係數轉換模型[1]，前者主要包含平均故障間隔 (Mean Time between Failures, MTBF) 參數與平均修復時間 (Mean Time to Recovery, MTTR) 參數，即

$$MTBF = (h_0 + h_1 + \dots + h_n) \div (f_0 + f_1 + \dots + f_n) \\ = \sum_{i=0}^n h_i \div \sum_{i=0}^n f_i = \frac{H}{F} = H \times F^{-1} \quad (1)$$

$$MTTR = (m_0 + m_1 + \dots + m_n) \div (f_0 + f_1 + \dots + f_n) \\ = \sum_{i=0}^n m_i \div \sum_{i=0}^n f_i = \frac{M}{F} = M \times F^{-1} \quad (2)$$

其中 h_i 表示第 i 項產品操作時數資訊； f_i 表示第 i 項產品故障次數及 m_i 表示第 i 項產品維修時數資訊。我們可發現重複統計了每項產品故障次數，故應使用所提方法，以降低運算成本。

表 1 顯示我們運用 MapReduce 分散式計算架構減少每台電腦運算成本，其中 K 表示電腦數量。所提方法比傳統單一電腦計算方法有更高運算效能，此外，所提方法不重複計算故障次數，故可有效減少 N 次加法次數。

圖 3 顯示所提方法運用於六分量氣動力係

數運算示意圖，其顯示變數 X_3 (即 $Q \times Aref$) 使用頻率相當高，即其頻率為 $7N$ 次，其中變數 N 為數據數量，故所提方法可減少 $7N$ 次乘法次數，如表 2 所示，另一方面，所提方法為分散式計算架構，故電腦最多負擔 $2N$ 次加法與 $3N$ 次乘法，其運算數量遠少於單一電腦運算數量，此外，當 N 為巨量資料，可使用 K 台電腦以進行 MapReduce 分散式運算，進而大幅地縮減計算時間。

5. 結論

本研究提出一個多項運算任務之高效率 MapReduce 架構，其巧妙地使用四則運算特性分解多項運算任務，並分析變數重複使用頻率，以避免重複計算數據，且十分適合 MapReduce 分散式計算架構，進而實現多部電腦同時進行運算任務，因此，所提方法所花費運算時間比先前計算方式少。我們後續擬研析所提方法運用於各單項產品之故障間隔與修復時間，俾利於提供更詳細的資訊於使用者。

參考文獻

- [1] 黃佩瑩、黃建民，“飛機掛載拋投軌跡模擬風洞試驗技術研析與系統更新驗證”，*新新科技年刊*，第 9 期，pp. 21-30，2013。
- [2] Jones, J. V., *Integrated Logistics Support Handbook*, 3th ed., McGraw Hill

Professional, 2006.

- [3] Dean, J. and Ghemawat, S., “MapReduce: Simplified Data Processing on Large Clusters,” *Proceedings of the 6th USENIX Symposium on Operating Systems Design and Implementation*, vol. OSDI-04, pp. 137-149, 2004.

Simplified Data Processing on Large Clusters,” *Communication of the ACM*, vol. 51, no. 1, pp. 107-113, 2008.

- [4] Dean, J. and Ghemawat, S., “MapReduce:

表 1 產品保固參數計算模型分析

		單一電腦	MapReduce 分散式計算架構		結合 MapReduce 分散式架構 與資料重複性分析	
			K-1 台電腦	第 K 台電腦	K-1 台電腦	第 K 台電腦
MTBF	產品操作時數	N 次加法	N / K 次加法	N / K 次加法	N / K 次加法	N / K 次加法
	產品故障次數	N 次加法	N / K 次加法	N / K 次加法	N / K 次加法	N / K 次加法
	彙總	1 次乘法	2K-2 次加法 與 1 次乘法		2K-2 次加法 與 1 次乘法	
MTTR	產品維修時數	N 次加法	N / K 次加法	N / K 次加法	N / K 次加法	N / K 次加法
	產品故障次數	N 次加法	N / K 次加法	N / K 次加法		
	彙總	1 次乘法	2K-2 次加法 與 1 次乘法			

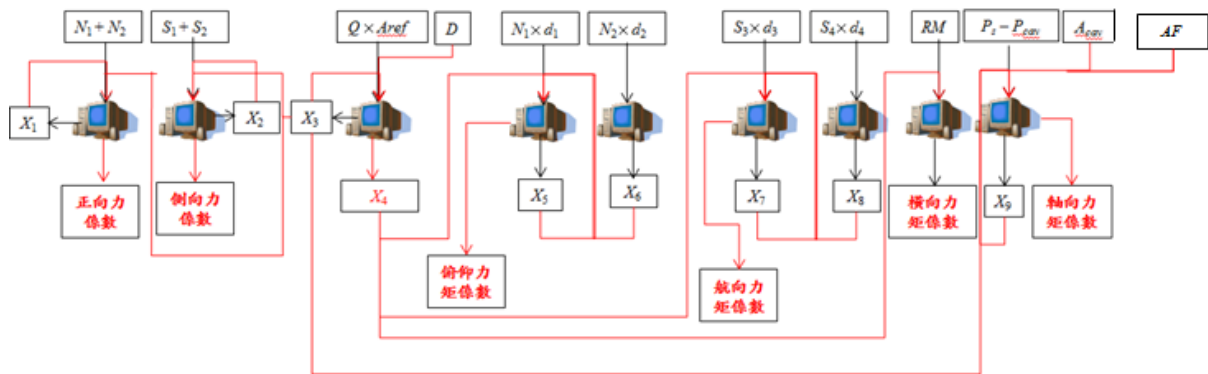


圖 3. 六分量氣動力係數計算環境示意圖

表 2 六分量氣動力係數運算模型分析

	單一電腦	結合 MapReduce 分散式架構與資料重複性分析			
		第一台電腦	第二台電腦	...	第九台電腦
正向力係數	N 次加法與 $2N$ 次乘法	6N 次加法與 15N 次乘法			
側向力係數	N 次加法與 $2N$ 次乘法				
俯仰力矩係數	N 次加法與 $5N$ 次乘法				
航向力矩係數	N 次加法與 $5N$ 次乘法				
橫向力矩係數	$3N$ 次乘法				
軸向力矩係數	$2N$ 次加法與 $5N$ 次乘法				
運算時間	$6N$ 次加法與 $22N$ 次乘法	2N 次加法與 3N 次乘法			