



Power-All 集群檔案系統 (PGFS)

定義全球存儲網格下一代

Power-All 網路有限公司技術白皮書

2008 年 4 月， 1.01 版本

目錄

1. 簡介	3
2. 存儲模式的轉換	4
3. 集群存儲分類	5
4. PGFS 功能	6
5. PGFS 架構	7
6. 數據與服務的高可用性	11
7. 高擴展性	13
8. 線性擴展生產率	14
9. 錯誤處理與數據恢復	15
10. 降低初始投資成本	16
11. 如何使用 PGFS 與當前第三方存儲相結合	17
12. 總結	18
13. 聯絡我們	19

1. 簡介

多媒體與使用者生成內容的極速增長

隨著 Web 2.0 的應用，網路多媒體與使用者生成內容極速增長。高清晰視頻內容及存儲需求正逐年成倍增加。傳統的磁帶，DAS，SAN 和 NAS 已無法滿足此種需求。公司一直在為其保貴資料尋求速度更快、擴展性與可靠性更高的存儲空間。在這種情形下，便形成了從先前的存儲產品的更新換代，即集群存儲。

為什麼需要 PGFS？

摩爾定義僅適用於計算機產業，不幸的是，雖然存儲產業占有計算機產業的一大部分，卻一直處於相對落後的形式。許多數據中心已經開始進行擴展事宜。由於大多數時間節點要等待讀寫緩慢且繁忙的存儲服務器，從而形成一種瓶頸，所以僅僅對於 CPUs 的投入是不夠的。

對於集群存儲，市場存在著不同類型的技術與產品。傳統來講，僅企業存在集群存儲需求，但其昂貴的價格難以令中小型企業所接受，而傳統的 SAN/NAS 又無法滿足它們需求。在這種情況下，Power-All 已開發了一項操作簡單且功能強大的集群文件系統，稱為“**Power-All 全球文件系統 (PGFS)**”。

此白皮書向用戶介紹一項目前正在進行的數據存儲行業的新型模式轉換，PGFS 功能，以及 PGFS 如何在基於低成本的 PC 硬件上提供強大的集群存儲性能。

2. 存儲模式的轉換

從 NAS/SAN 到集群存儲

在過去，人們使用獨立電腦硬盤的數據存儲。而這種狀況被 NAS 和 SAN 所取代。通過使用 NAS 和 SAN，用戶可以通過高速網絡進行數據訪問與共享。NAS/SAN 以其良好的性能與適中的價格被中小型企業所接受，因此 NAS/SAN 成爲了市場主流存儲產品。

對於 NAS/SAN，通常使用單一的 I/O 設備爲網絡客戶提供存儲服務。要擴展 NAS/SAN 的最有效的方法是增加更多的存儲硬件。由於硬件的限制，存儲最大容量有限。由於存儲容量需求的不斷增加，傳統的 NAS/SAN 已逐漸無法滿足此需求。

使用集群存儲的優勢

集群存儲是繼 NAS/SAN 存儲之後的新一代產品，具有更多優勢。首先，摩爾定律描述了計算機硬件曆史的一個重要趨勢：即芯片上可以容納的晶管體數量每隔兩年會翻一番，相應的計算能力也隨之翻番。根據成本效益，不建議初始階段購買硬件作爲未來擴展的備用，

而是當存儲集群不能滿足日益增長的需求時再訂購硬件進行擴展；其次，集群存儲是通過軟件實現的，因此能沖破大量的硬件限制，從而進行擴展；最後，集群可並行 I/O 至用戶端機器，存儲容量及讀寫性能可以線性的同步提高。

通過以上優勢，形成了由先前的 NAS/SAN 到集群存儲的模式轉換。Power-All 的任務是趕超這種模式轉換並通過使用低成本 PC 組件開發網絡全球集群存儲。爲了實現這一目標，Power-All 已開發了一種核心技術，即 PGFS。

3. 集群存儲分類

集群存儲是代表由多個組件構成的存儲池的通用術語。市場上存在著不同類型的集群存儲產品及技術，以下分類為：

集群存儲產品類型

A. 集群 SAN

集群 SAN 是指可擴展 SAN，可在數據塊級進行存儲擴展。集群 SAN 是價格昂貴的企業級產品。

B. 集群 NAS

集群 NAS 是指可擴展 NAS，可在文件系統級進行存儲擴展。通常集群 NAS 通過集群文件系統完成。

集群檔案系統

目前市場有很多不同類型文件系列歸類為集群文件系統，但提供的功能卻不同。如 CFS，主要分為兩種類型：

A. 共用網路分塊設備

對這種類型的 CFS，其目的為支持多個終端同時訪問同一網絡分塊設備。傳統來說，大多數網絡分塊設備技術不允許多終端同時訪問。CFS 提供鎖機制以啓用多終端同時訪問。例如 Redhat 的 GFS 全球文件系統。

B. 分散式資料訪問設備

對於此類 CFS，其目的為將數據進行分布以訪問多重設備。與傳統 NAS 相比，具有更高的擴展力與性能。從原則上講，此類 CFS 性能更高與其容量呈線性關係。如 Sun Microsystems 的 CFS，Power-All 亦屬此類產品。

4. PGFS 功能

什麼是 PGFS

PGFS 全寫：Power-All Global File System，是由 Power-All 網絡有限公司開發的一項專有集群文件系統。PGFS 是 Power-All 全球雲存儲服務（Aspen Cloud Storage）的核心技術。

開發 PGFS 的目標是通過使用標準 PC 硬件組件創建具有高擴展性，高可靠性以及最佳性能的集群式 NAS。

PGFS 主要性能

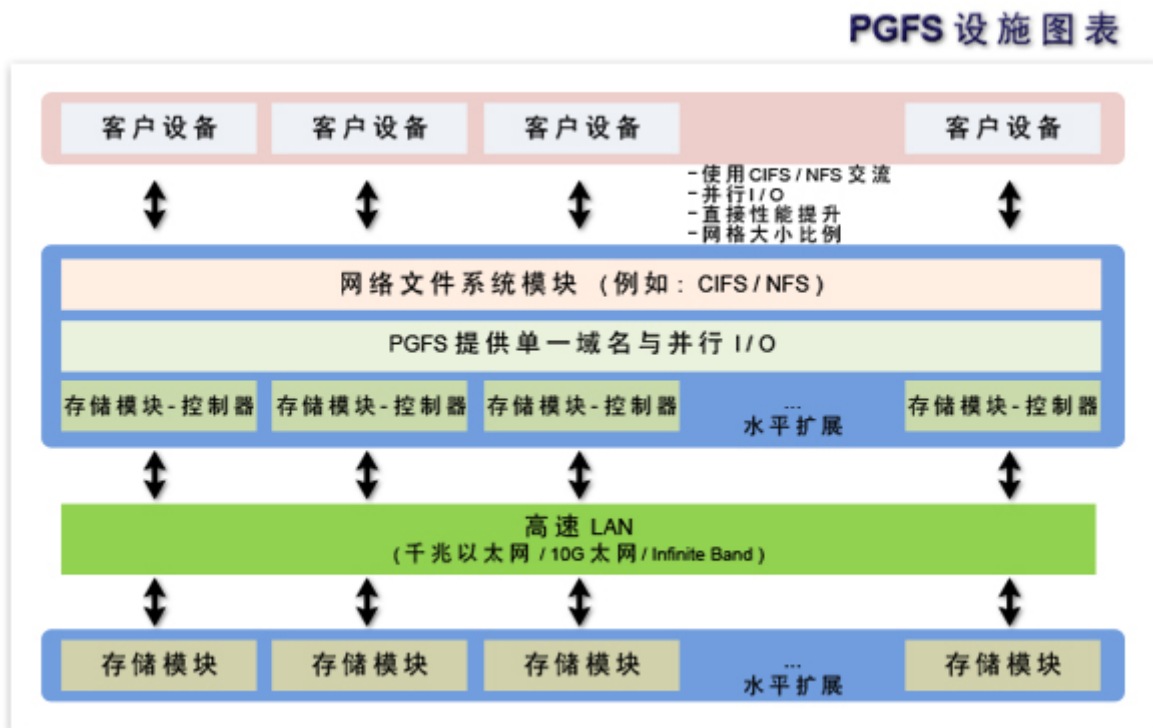
- 可達 PB 級的高擴展性，數據被分布式訪問集群內的所有存儲模塊。
- 數據高可靠性，數據實時複製
- 支持多向複製
- 無單一節點故障，PGFS 所有組件可水平擴展
- 其它許多集群文件系統，在一些組件上會出現單一節點故障或瓶頸故障。如存儲控制器或 meta 數據服務器。PGFS 設計簡潔，無需集中式元數據服務器，不存在以上問題。
- 文件系統的獨立域名使應用集成更加簡易
- 實現性能的線性擴展。許多集成文件系統具有瓶頸限制
- 性能增長與存儲節點增長呈線性關係
- 運行於 IP 網絡之上，說明 PGFS 可以運行於高質量的 WAN 連接之上
- 支持 10G 以太網及 InfiniBand

5. PGFS 架構

PGFS 核心組件

與其它集群文件系統不同的是，PGFS 架構簡潔，易於操作。PGFS 僅包含兩個核心組件，存儲模塊與存儲控制模塊。其架構如圖 1 所示：

圖 1



如圖 1 所示：您可看到 PGFS 由存儲模塊與存儲控制模塊所組建。存儲模塊與存儲控制模塊之間由高速 LAN，如 GE，10GE，或 InfiniBand 連接。

5. PGFS 架構(續上)

存儲模組規則

存儲模塊是指由大量硬盤所組配的裝置。Power-All 開發了基於 CSAN/CNAS 系統的存儲模塊，是針對中小型企業的標準存儲產品。圖 2 為存儲模塊示例：

CSAN/CNAS 系統存儲產品易於操作，即時使用。CSAN/CNAS 系統與 Power-All 存儲管理器（PSM）需提前安裝，PSM 功能齊全，是處理該系統的管理與監測工具。CSAN/CNAS

其自身裝有硬件 RAID5/6 控制器，對數據進行額外保護。

CSAN/CNAS 硬件規格和功能，更多詳情請訪問 <http://www.powerallnetworks.com/>

圖 2

儲存模块图例



5. PGFS 架構(續上)

與存儲模組的區別

存儲模塊是基於 CSAN/CNAS 之上的開發產品，兩者之間的主要區別在於，存儲模塊與 PGFS 服務器端軟件同時預裝。在存儲模塊中，通過 Power-All 存儲管理器（PSM）創建單一邏輯卷（LV）。依據簡明性，我們建議 LV 大小應與 CSAN/CNAS 系統總大小相等。

存儲模組當地檔案系統

原則上，PGFS 是與存儲模塊的當地文件系統相獨立。然而，根據 Power-All 的實驗中，我們發現 XFS 是速度最快且最有效的當地文件系統。XFS 本身支持日志功能，防止突發故障時數據的丟失。XFS 是 CSAN/CNAS 系統唯一支持的內置文件系統，PGFS 的默認存儲模塊為 XFS 系統。

PGFS 伺服器怎樣與 PGFS 控制器相連接

PGFS 服務端軟件運行於所建的邏輯卷上。該軟件是 PGFS 控制器軟件與物理存儲之間的接口。通過使用 IP 層或 InfiniBand RDMA 上的 TCP 協議，從而進行 PGFS 服務器端與 PGFS 控制器之間的連接。

5. PGFS 架構(續上)

存儲控制模組規則

存儲控制模塊是指 PGFS 的控制器。字母“C”代表 Controller，即控制器。它是物理存儲（所有存儲模塊）與用戶應用間的接口。存儲控制模塊具有以下性能：

- 集結多個存儲模塊至單一存儲池內，擁有同一空間名
- 通過所有存儲模塊進行數據分割
- 提供存儲模塊間的多向實時數據複製
- 提供在線增加/減少存儲模塊數量
- 自我修復與監測

由於 PGFS 是專有文件系統，所有存儲控制模塊裝置被提前安裝在 Samba 與 NFS 服務器上。因此服務器支持 CIFS/NFS 通過 Samba/NFS 服務器軟件與 PGFS 進行連接。在圖 1 中，CIFS/NFS 位於 PGFS 連接至客戶應用服務器之上。

在 WAN 或互聯網上運行 PGFS

因為 PGFS 可在 TCP/IP 下運行，理論上講，PGFS 也可通過 WAN/互聯網運行。然而實際上，有許多因素影響其可行性，如：帶寬，成本，以及網絡延遲等。PGFS 通過 WAN 運行時會造成瓶頸障礙。WAN 帶寬始終低於 LAN，因此總是造成瓶頸。通過當前互聯網連接技術，Power-All 不建議通過 WAN/互聯網使用 PGFS 剝離數據。

6. 數據與服務的高可用性

本章向您介紹 PGFS 如何實現數據的高可用性

實時數據複製

PGFS 支持多向數據實時複製。管理員可在不同數據區自定義複製策略。每一數據區為存儲模塊獨立群，以獨立方式運行。參考圖 3，存儲模塊 1 和 2 被 A 區分離，存儲模塊 3 和 4 被 B 區分離。複製規則可配置在所有存儲控制模塊內，以進行兩區間的實時複製。與其它具有 Master-Slave 設計的文件複製工具不同的是，PGFS 提供 Multi-Masters 處理裝置。PGFS

擁有鎖定機制以控制 Multi-Masters 數據同時寫入。

存儲控制模塊根據用戶自定義規則，決定複製數據的區域流向。數據從一個數據區內分離到所有存儲模塊。

除了複製，每一存儲模塊配備了硬件 RAID5/6 控制器，即使 1 或者 2 硬盤失效也能保持存儲模塊在線。

無單一節點故障

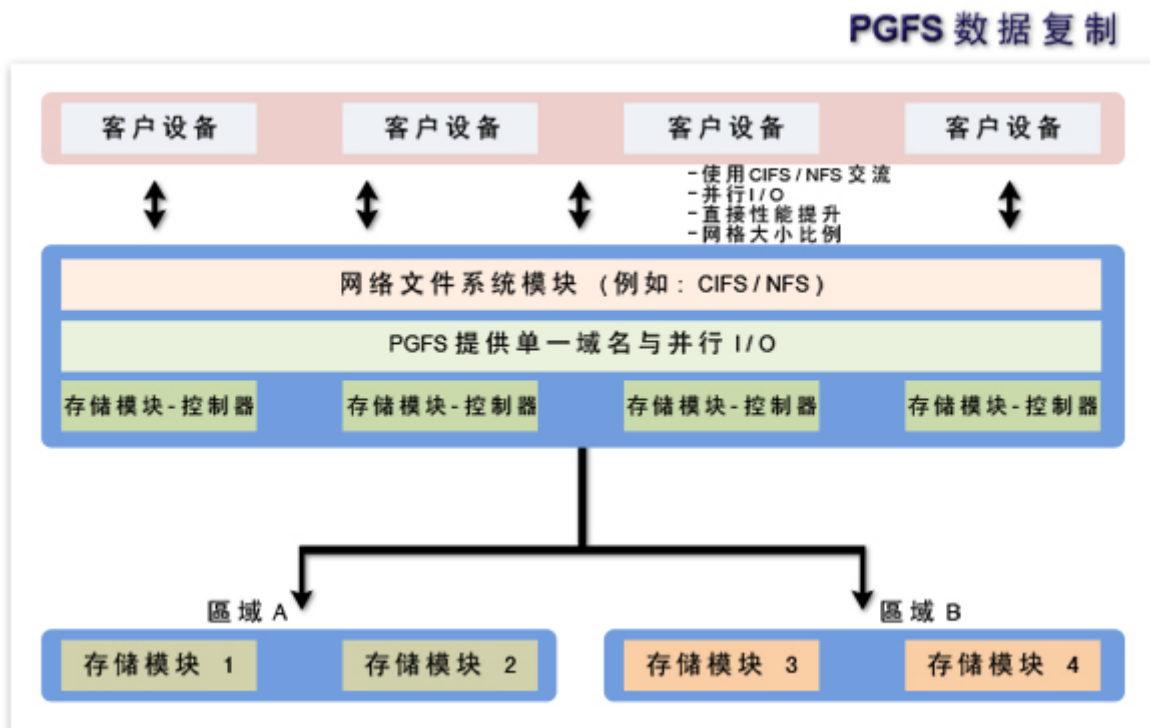
許多集群文件系統為元數據控制器組件。這一設計，會引起性能瓶頸及集群的單一節點故障。PGFS 集群無需任何元數據組件，其組件分為兩部分：存儲模塊和存儲控制模塊。以上兩種組件可在 PGFS 中進行水平擴展，數量可達上萬裝置。

6. 數據與服務的高可用性(續上)

如果其中任一存儲控制模塊失效，將不對其它存儲模塊造成影響。在這種情況下，客戶應用服務器應停止發送 I/O 至失效的存儲控制模塊。

如果存儲模塊下降，存儲控制模塊可在數秒內檢測出失效節點並停止數據複製到此區域。如果區域全部失效，將引起整個集群失效。

圖 3



7. 高擴展性

本節向您介紹 PGFS 如何實現存儲系統的高擴展性

PGFS 是一個操作簡易且功能強大的集群文件系統，無需元數據控制器。PGFS 可在大多數當地文件系統上運行，如 ext3, XFS 等。XFS 是用於所有存儲模塊的默認文件系統。

PGFS 僅包括兩類組件：存儲模塊和存儲控制模塊。以上兩類組件均可進行水平擴展，可達上萬裝置，其總大小可達 PB 級。

爲了擴大存儲總空間，可添加額外存儲模塊。首先，新存儲模塊與後端開關相連接（以太網或 Infiniband，根據存儲模塊接口而定）。其次，修改配置文件複製到所有存儲控制模塊中並進行重裝。所有操作可在線完成。

8. 線性擴展生產率

本節向您介紹 PGFS 如何進行線性性能擴展

本節向您介紹 PGFS 如何進行線性性能擴展

線性擴展量是 PGFS 的主要功能。通過存儲模塊與存儲控制模塊的增加，整體性能呈線性增長。

當 PGFS 增加額外的存儲模塊，其總存儲量與總磁盤容量也隨之增加。然而，一旦存儲控制模塊性能出現瓶頸故障，增加額外的存儲模也無法另其性能提高。在這種情況下，管理員應增加存儲控制模塊的數量。

如何測定哪些元件需要升級？

爲了擴展 PGFS，管理員應根據以下規則確定增加存儲模塊或存儲控制模塊。

- I. 如擴展存儲容量，應增加存儲模塊的數量。
- II. 如磁盤 I/O 出現瓶頸，應增加存儲模塊數量。
- III. 如每一存儲控制模塊的帶寬都被耗盡，應增加存儲控制模塊的數量。

9. 錯誤處理與數據恢復

本節向您介紹 PGFS 如何處理錯誤運行

PGFS 致力於向關鍵應用任務提供一個可靠的存儲空間。錯誤處理是提供 7 × 24 小時不間斷運行環境最爲關鍵的一部分。

硬碟驅動故障案例

每一存儲控制模塊配有硬件 RAID 控制器並預裝在 Power-All 存儲管理器 (PAM) 中。一旦出現任何硬盤驅動問題，PSM 可直接檢測並向管理員發送警報郵件。As there is RAID5/6, Storage Module is still online with 1-2 hard drives.

存儲模組故障案例

一旦整個存儲模塊出現故障，存儲控制模塊可直接檢測並向管理員發送警報郵件。如啓用實時複制功能，服務將不會中斷。存儲控制模塊將停止向失效節點發送流量直至節點恢復正常。

存儲控制模組故障案例

如果存儲控制模塊出現故障，其它存儲控制模塊將保持在線狀態繼續提供服務，所以不會對服務造成中斷。在這種情況下，客戶端應用服務器應具有存儲控制模塊的故障檢測機制並停止向失效節點發送流量。

10. 降低初始投資成本

本節向您介紹 PGFS 如何降低企業存儲成本

摩爾定律 (Moore's Law)

摩爾定律概述計算機硬件曆史的重要趨勢：芯片中晶體管的數量將成倍數增加，每兩年翻一番。

存儲方面，每一 GB 的價格也逐日遞減。另外，硬件公司資產貶值之事也時有發生。基於以上情況，公司購買大批存儲產品以供日後使用是極不劃算的。PGFS 是此類問題的最佳解決方案。PGFS 支持在線存儲擴展，當現存集群存儲接近最大存儲量時，可靈活增加額外存儲模塊。

基於 PC 組件

PGFS 設計可運行於標準 PC 組件。隨著當前 PC 組件的快速變化，存儲模塊將隨其市場產生更具成本效益的硬體組件。PGFS 可兼容不同版本的硬體組件。

11. 如何使 PGFS 與當前第三方存儲相結合

本節向您介紹現存非 Power-All 存儲產品整合到 PGFS：

Power-All 熟悉企業可能擁有當前存儲產品且不希望建立新的 PGFS 後將其閑置。在這種情況下，PGFS 提供特制諮詢解決方案，即將第三方存儲產品與 PGFS 進行整合。

從理論上講，PGFS 可運行於所有文件系統之上。將第三方存儲與 PGFS 整合，通常需要一種作為第三方存儲產品與 PGFS 間的接口的服務器，即存儲耦合器。

12. 總結

使用集群存儲產生的公司效益

存儲產業正經歷由傳統的 NAS/SAN 朝向集群存儲的一種模式轉換。使用集群存儲企業可獲得的效益為：

- 降低存儲成本
- 更有效的企業數據管理
- 性能更高
- 更加可靠

市場上擁有很多集群存儲產品，但大多數是為企業定制的昂貴的硬體組件。Power-All 的 PGFS 正開啓一場存儲革命，通過使用成本低廉的 PC 標準硬件建立具有同等企業性能的存儲產品。是中小型企業存儲的最佳解決方案。

13. 聯絡我們

關於 Power-All 網路有限公司

Power-All是集群存儲行業的領導者之一，在IDC行業擁有堅實的操作經驗。Power-All認為基於PC組件的集群存儲是該行業的發展趨勢並成為下一代主流解決方案。通過與PFFS和其它領先技術的結合，Power-All已開發了全球存儲雲服務，即Aspen Cloud Storage存儲服務。更多詳情，請瀏覽 <http://www.cloudwww.com/>

- 獲得更多資訊，請與 Power-All 聯繫：

地址：*Power-All Networks Limited*
香港新界沙田香港科學園科技大道西五號
企業廣場 5 樓 540 及 541 室

電話：*(852) 2111 8182*

傳真：*(852) 2111 8156*

郵箱：*newgen@powerallnetworks.com*