



## **POWER-ALL GLOBAL FILE SYSTEM (PGFS)**

Defining next generation of global storage grid

Power-All Networks Ltd. Technical Whitepaper

April 2008, version 1.01

**Table of Content**

1. Introduction .....	3
2. Paradigm Shift of Storage Products .....	4
3. Clustered Storage Categories .....	5
4. PGFS Functionalities .....	6
5. PGFS Architecture .....	7
6. High Availability of Data and Service .....	11
7. High Scalability .....	13
8. Linear Scalability Throughput .....	14
9. Error Handling and Data Recovery .....	15
10. Reduce Initial Investment Cost .....	16
11. How to integrate PGFS with existing third party storage .....	17
12. Conclusion .....	18

## 1. Introduction

### Explosive Growth of Multimedia and User Generated Content

With the wide use of Web 2.0, there is an explosive growth of multimedia and user generated content on Internet. High quality of video creates bigger size of content. The requirement of storage is exponentially increasing every year. Traditional tape, DAS, SAN and NAS cannot meet the requirement. Companies always look for fast, scalable and reliable storage for their valuable data. In this situation, there is a major paradigm shift from previous generations of storage products to next generation product, clustered storage.

### Why PGFS is needed?

Moore's law is applied only to the computing industry. Unfortunately storage industry is very much left behind even though they are very much a part of compute infrastructure. Lot of data centers are already running into scaling issues. Investing on more CPUs alone isn't sufficient, because most of the time nodes wait for data to be read or written to a slow and busy storage server creating a bottle neck.

For clustered storage, there are variant types of technologies and products on the market. Traditionally only enterprise requires clustered storage and the price is usually expensive which cannot be afforded by small-to-medium sized companies while traditional SAN/NAS cannot meet their requirement anymore. In this situation, Power-All has developed a simple and powerful clustered file system called **“Power-All Global File System (PGFS)”**.

The purpose of this whitepaper is to introduce you to a new paradigm shift that is currently taking place in the data storage industry, PGFS features, and how PGFS provides powerful clustered storage features based on low cost PC hardware.

## 2. Paradigm Shift of Storage Products

### From NAS/SAN to Clustered Storage

In the past, people use standalone hard drives for data storage. This kind of situation has been faded out by NAS and SAN. By using NAS and SAN, people can access and share the data through high speed network. NAS/SAN itself usually is good in performance and price is affordable by medium sized companies. Therefore, NAS/SAN has become main stream storage products on market.

In NAS/SAN, usually it is a single unit to provide I/O to networked client machines. To scale, NAS/SAN, most common method is adding more hard drives to the unit. Due to hardware limitation, maximum size of storage is greatly limited. In addition, throughput becomes lower when storage size increases. Due to storage size requirement is increasing from time to time, traditional NAS/SAN starts failing to fit requirement.

### Advantages using Clustered Storage

Clustered storage is the next generation of storage after NAS/SAN. There are many advantages using Clustered storage. Firstly, Moore's Law describes an important trend in the history of computer hardware: that the number of transistors that can be inexpensively placed on an integrated circuit is increasing exponentially, doubling approximately every two years. It is not cost effective to buy a powerful hardware on initial stage and reserve it for future expansion. With clustering, hardware can be ordered in future when expansion is required. Secondly, clustering is done by software therefore it can eliminate lot of hardware limitations and able to scale larger. Thirdly, clustering enables parallel I/O to client machines which can increase performance significantly and linear to storage size.

Because of above advantages, there is major paradigm shift from previous NAS/SAN to clustered storage. Power-All's mission is to catch up the paradigm shift and develops global clustered storage on Internet by using low cost PC based components. To achieve this mission, Power-All has developed one of its core technologies called PGFS.

### 3. Clustered Storage Categories

Clustered Storage is a general term representing a storage pool built by multiple units. There are variants of Clustered Storage products and technologies on the market. Below is the classification.

#### Clustered Storage Product Types

##### A. Clustered SAN

Clustered SAN refers to a scalable SAN which can expand the storage at data block level. Usually clustered SAN is expensive and enterprise graded product.

##### B. Clustered NAS

Clustered NAS refers to a scalable NAS which can expand the storage at file system level. Usually clustered NAS is done by Clustered File System.

#### Clustered File System

There are many file systems on the market classified as Clustered File System (CFS), however they are providing different functions. In CFS, there are two main types:

##### A. Share network block device

For this type of CFS, it aims to enable multiple devices for concurrent access to same network block device. Traditionally most network block device technology does not allow multiple devices concurrent access. The CFS aims to provide a locking mechanism enabling multiple concurrent accesses. Example of such CFS is Redhat's Global File System

##### B. Distribute data across multiple devices

For this type of CFS, it aims to distribute data across multiple devices. It enables higher scalability and higher performance compared to traditional NAS. In principle, this type of CFS provides higher performance linear to size. Example of such CFS is Sun Microsystems's Lustre. Power-All's PGFS also belongs to this category.

## 4. PGFS Functionalities

### What is PGFS

PGFS stands for Power-All Global File System which is a proprietary clustered file system developed by Power-All Networks Ltd. PGFS is one of the core technologies to Power-All enabling global cloud storage service, Aspen Cloud Storage.

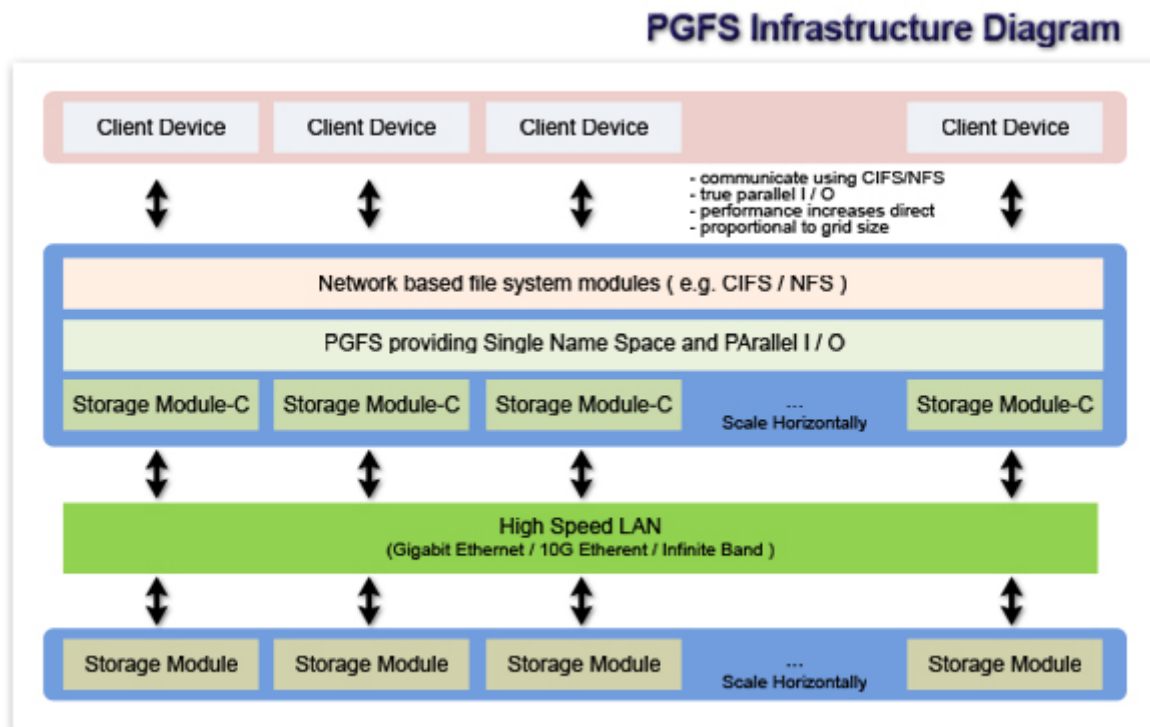
The objective of developing PGFS is to create a highly scalable, highly reliable, and best performance Clustered NAS by using standard PC based hardware components.

### PGFS Key Features

- Highly scalable to Petabytes level. Data can be distributed across all storage modules within cluster.
- High availability of data by real time data replication.
- Multi-directional replication is supported.
- No single point of failure. All components in PGFS can be scaled horizontally. Many other Cluster File Systems, they have bottleneck or single point of failure on some components such as storage controller or metadata servers. PGFS provides a simple design without necessary of centralized metadata servers.
- Single namespace of file system makes application integration easier
- Truly linear scalability in performance. Many clustered file system has bottleneck
- Performance increases linear to number of storage nodes.
- Runs on top of IP network which implies PGFS works over high quality of WAN connection.
- Support 10G Ethernet and InfiniBand

**5. PGFS Architecture**
**PGFS Core Components**

Unlike other clustered file systems, PGFS's structure is simple and easy to use. PGFS only consists of two core components, Storage Module and Storage Module-C. Diagram 1 shows the structure of PGFS.

**Diagram 1**


From diagram 1, you can see PGFS is built by Storage Module and Storage Module-C. In between Storage Module and Storage Module-C, it is high speed LAN such as GE, 10GE, or InfiniBand.

## 5. PGFS Architecture (continued)

### Role of Storage Module

Storage Module refers to the unit equipped with massive hard drives. Power-All developed Storage Module based on CSAN/CNAS System product which is a standard storage product targeted for SMB companies. Diagram 2 shows an example of Storage Module.

As CSAN/CNAS System itself is a ready-to-use storage product, CSAN/CNAS System is preloaded with Power-All Storage Manager (PSM) which is a full featured management and monitoring tool to manage the system. CSAN/CNAS itself is equipped with hardware RAID 5/6 controller for extra protection of data.

For hardware specification and features of CSAN/CNAS, please visit <http://www.powerallnetworks.com/> for more detail.

### Diagram 2

#### Example of Storage Module



## 5. PGFS Architecture (continued)

### **Difference between CSAN/CNAS and Storage Module**

Storage Module is developed based on CSAN/CNAS. The major difference is that Storage Module is pre-loaded with PGFS server side software.

In Storage Module, a single Logical Volume (LV) is created through Power-All Storage Manager (PSM). We recommend LV size should be equal to total size of CSAN/CNAS System in terms of simplicity.

### **Local File System of Storage Module**

In principle, PGFS is independent to Storage Module's local file system. However, from Power-All's experiment, XFS is the fastest and most efficient local file system. XFS itself supports journaling which can prevent data loss due to sudden failure. Since XFS is the built-in and only one file system supported by CSAN/CNAS System, PGFS's Storage Module makes use of XFS by default.

### **How PGFS server communicates with PGFS controller.**

PGFS server side software runs on top of the created logical volume. The software is an interface between PGFS controlling software and physical storage. PGFS server software communicates with PGFS controller using TCP protocol on top of IP layer or InfiniBand RDMA.

## 5. PGFS Architecture (continued)

### Role of Storage Module-C

Storage Module-C refers to PGFS's controller. The character "C" means controller. It is an interface between physical storage (all Storage Modules) and user applications. Storage Module-C has following key features:

- Aggregate multiple Storage Modules into single storage pool with same name space.
- Stripping data across all Storage Modules
- Provide multi-directional real time data replication between Storage Modules
- Provide online increase/decrease number of Storage Modules
- Self healing and monitoring

Since PGFS is a proprietary file system, there by default all Storage Module-C units are pre-configured with Samba and NFS server. Therefore servers supporting CIFS/NFS can connect to PGFS through Samba/NFS server software. On diagram 1, CIFS/NFS is on top of PGFS interfacing to client application servers.

### Running PGFS over WAN or Internet

Since PGFS can be run under TCP/IP, theoretically PGFS works over WAN/Internet as well. However in practical, there are many factors affecting feasibility of running PGFS across WAN/Internet such as bandwidth, cost, and network latency, etc. Actually PGFS across WAN will only create bottleneck. It is reasonable that WAN's bandwidth always lower than LAN, therefore WAN will always be the bottleneck. With current Internet connection technology, Power-All does not recommend using PGFS to strip data across WAN/Internet.

To complement the limitation of current Internet connection technology, Power-All has developed Aspen Cloud Storage and Aspen CDN services which can provide single name space global storage across Internet. For detail, please refer Aspen Cloud Storage and Aspen CDN documents from Power-All website.

## 6. High Availability of Data and Service

This chapter introduces you how PGFS achieves the function of High Availability of data.

### Real Time Data Replication

PGFS supports multi-directional real time data replication. Administrator can define custom replication policies replicating data to different data zones. Each data zone is a separated group of Storage Modules which are running in stripping mode.

By referring diagram 3, Storage Module 1 & 2 are stripped as Zone A while Storage Module 3 & 4 are stripped as Zone B. Replication rule can be configured in all Storage Module-C to do real time replication between two zones. Unlike many file replication tools with Master-Slaves design, PGFS offers Multi-Masters mechanism to replication. PGFS has locking mechanism to control the Multi-Masters data concurrent writing.

Storage Module-C always acts as the commander to decide replicating data to which Zone according to user defined rule. Data is stripped to all Storage Modules within a replicated zone.

In addition to replication, each Storage Module is equipped with hardware RAID 5/6 controller which retains Storage Module online even 1/2 hard drives failed.

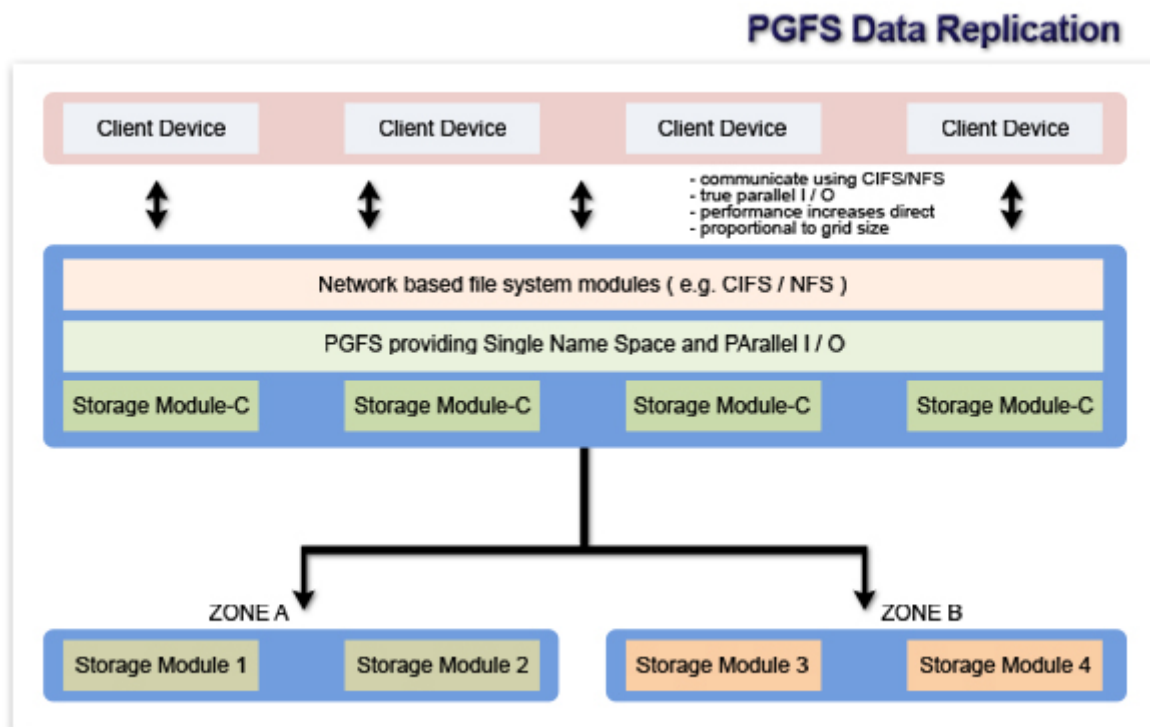
### No Single Point of Failure

In many CFSs, there are components of Metadata Controllers. With this design, usually it results performance bottleneck and single point of failure to the cluster. PGFS does not need any metadata controllers for the cluster. PGFS has only two components, Storage Module and Storage Module-C. Both components in PGFS can be scaled horizontally up to thousands of units.

In case any Storage Module-C is down, there is no interruption to storage system as other Storage Module-C is still running. In this situation, client application servers should have mechanism to stop sending I/O to this failed Storage Module-C.

**6. High Availability of Data and Services (continued)**

In case any Storage Module is down, Storage Module-C can detect its failure within seconds and stop replicating data to that zone. If all zones are down, it results whole cluster is down too.

**Diagram 3**


## 7. High Scalability

This chapter introduces you how PGFS achieves the function of High Scalability of storage system.

PGFS is a simple and powerful cluster file system. It does not require metadata controllers which is most common bottleneck in many CFSs. PGFS can be run on top of most local file system such as ext3, XFS, etc. XFS is the default file system used on all Storage Modules.

PGFS has only two components, Storage Module and Storage Module-C. Both components in PGFS can be scaled horizontally up to thousands of units with total size up to Petabytes level.

To expand total size of storage, additional Storage Module is required. Firstly, new Storage Module is connected to backend switch (either Ethernet or Infiniband, depends on Storage Module interface). Secondly, modify configuration file on all Storage Module-C and reload the service. All actions can be completed while storage remains online.

## 8. Linear Scalability Throughput

This chapter introduces you how PGFS scales in performance linearly.

Linear scalability throughput is one of the key features of PGFS. With increased number of Storage Module and Storage Module-C, the overall performance increases linearly.

When additional Storage Modules are added to PGFS, the total storage size and total disk throughput is increased. However, in case there is bottleneck on Storage Module-C performance, adding additional Storage Module cannot improve performance. In this case, administrator should increase number of Storage Module-C.

### **How to determine upgrading which component?**

To scale a PGFS, administrator should follow below rules to consider adding Storage Module or Storage Module-C:

- I. If expansion to storage size, should increases number of Storage Modules
- II. If bottleneck on disk I/O, should increases number of Storage Modules
- III. If each Storage Module-C's bandwidth is utilized, should increases number of Storage Module-C

## 9. Error Handling and Data Recovery

This chapter introduces you how PGFS handles error.

PGFS aims to provide a reliable storage for mission critical applications. Error handling is very important area to a 7 x 24 non-stop environment.

### **Case for hard drive failure**

Each Storage Module-C is equipped with hardware RAID controller and preloaded with Power-All Storage Manager (PSM). In case any hard drive problem, PSM can detect and send out alert email to administrator. As there is RAID5/6, Storage Module is still online with 1-2 hard drives.

### **Case for Storage Module Failure**

In case the whole Storage Module failed, Storage Module-C can detect the event and send out alert email to administrator. If real time replication is enabled, there is no interruption of service. Storage Module-C will not send traffic to failed node anymore until it is recovered.

### **Case for Storage Module-C Failure**

In case Storage Module-C failed, there is no interruption to service as other Storage Module-C is still online and continues providing services. In this case, client application servers should have mechanism detecting failure of Storage Module-C and stop sending traffic to this failed node.

## 10. Reduce Initial Investment Cost

This chapter introduces you how PGFS can reduce company investment on storage.

### Moore's Law

Moore's Law describes an important trend in the history of computer hardware: that the number of transistors that can be inexpensively placed on an integrated circuit is increasing exponentially, doubling approximately every two years.

In storage, price per GB is dropping per day. In addition, hardware is asset of company which will be depreciated from time to time. Based on this situation, it is not cost effective for a company buying a big storage for future use.

PGFS is designed to solve this situation. PGFS supports online expansion of storage. When existing clustered storage is near maximum size, additional Storage Modules can be added dynamically.

### PC Based Component

PGFS is designed to be run on standard PC components. As PC component changes very fast on market, Storage Modules will follow market with most cost effective hardware components. PGFS is compatible with different aged hardware components.

## 11. How to integrate PGFS with existing third party storage

This chapter introduces you how existing non-PowerAll storage product be integrated into PGFS.

Power-All understands company may have existing storage product and does not want to keep it idle after building a new PGFS. In this case, PGFS offers tailor made consulting solution integrating third party storage product as member of PGFS.

Theoretically PGFS can be run on top of any local file system. To integrate third party storage to PGFS, usually require a Storage Integrator which is a server acts as interface between third party storage product and PGFS.

## 12. Conclusion

### Company Benefits using Clustered Storage

There is a paradigm shift in storage industry in which storage product is moving from traditional NAS/SAN towards clustered storage. With clustered storage, companies can benefit:

- Reduce storage cost
- Better management of company data
- Higher performance
- More reliable

There are many clustered storage products on the market, but most are expensive tailor made hardware components for enterprise. Power-All PGFS enables a new revolution in storage by using low cost PC based standard hardware while remaining all enterprise features. It can bring enterprise graded solution into medium sized companies.

### About Power-All Networks Ltd

Power-All is one of the leaders in clustered storage industry. With solid experience in IDC industry, Power-All believes PC based clustered storage is a trend and next generation of main stream storage solution in the industry. By combining PFFS and other leading technologies, Power-All has developed a global cloud storage service called Aspen Cloud Storage. For more detail, please visit <http://www.cloudwww.com/>

- **For more information, please contact Power-All at:**

*Address: Power-All Networks Limited  
Unit 540 & 541, 5/F,  
Enterprise Place, No. 5 Science Park West Avenue,  
Hong Kong Science Park, Shatin, Hong Kong.*

*Phone: (852) 2111 8182*

*Fax: (852) 2111 8156*

*Email: [newgen@powerallnetworks.com](mailto:newgen@powerallnetworks.com)*