



Power-All 集群文件系统（PGFS）
定义全球存储网络下一代
Power-All 网络有限公司技术白皮书
2008 年 4 月， 1.01 版本

目录

1. 简介	3
2. 存储模式的转换	4
3. 集群存储分类	5
4. PGFS 功能	6
5. PGFS 架构	7
6. 数据与服务的高可用性	11
7. 高扩展性	13
8. 线性扩展生产率	14
9. 错误处理与数据恢复	15
10. 降低初始投资成本	16
11. 如何使用 PGFS 与当前第三方存储相结合	17
12. 总结	18
13. 联络我们	19

1. 简介

多媒体与用户生成内容的极速增长

随着 Web 2.0 的应用，网络多媒体与用户生成内容极速增长。高清晰视频内容及存储需求正逐年成倍增加。传统的磁带，DAS，SAN 和 NAS 已无法满足此种需求。公司一直在为其宝贵数据寻求速度更快、扩展性与可靠性更高的存储空间。在这种情形下，便形成了从先前的存储产品的更新换代，即集群存储。

为什么需要 PGFS？

摩尔定义仅适用于计算机产业，不幸的是，虽然存储产业占有计算机产业的一大部分，却一直处于相对落后的形式。许多数据中心已经开始进行扩展事宜。由于大多数时间节点要等待读写缓慢且繁忙的存储服务器，从而形成一种瓶颈，所以仅仅对于 CPUs 的投入是不够的。

对于集群存储，市场存在着不同类型的技术与产品。传统来讲，仅企业存在集群存储需求，但其昂贵的价格难以令中小型企业所接受，而传统的 SAN/NAS 又无法满足它们需求。在这种情况下，Power-All 已开发了一项操作简单且功能强大的集群文件系统，称为“**Power-All 全球文件系统 (PGFS)**”。

此白皮书向用户介绍一项目前正在进行的数据存储行业的新型模式转换，PGFS 功能，以及 PGFS 如何在基于低成本的 PC 硬件上提供强大的集群存储性能。

2. 存储模式的转换

从 NAS/SAN 到集群存储

在过去，人们使用独立电脑硬盘的数据存储。而这种状况被 NAS 和 SAN 所取代。通过使用 NAS 和 SAN，用户可以通过高速网络进行数据访问与共享。NAS/SAN 以其良好的性能与适中的价格被中小型企业所接受，因此 NAS/SAN 成为了市场主流存储产品。

对于 NAS/SAN，通常使用单一的 I/O 设备为网络客户提供存储服务。要扩展 NAS/SAN 的最有效的方法是增加更多的存储硬件。由于硬件的限制，存储最大容量有限。由于存储容量需求的不断增加，传统的 NAS/SAN 已逐渐无法满足此需求。

使用集群存储的优势

集群存储是继 NAS/SAN 存储之后的新一代产品，具有更多优势。首先，摩尔定律描述了计算机硬件历史的一个重要趋势：即芯片上可以容纳的晶体管数量每隔两年会翻一番，相应的计算能力也随之翻番。根据成本效益，不建议初始阶段购买硬件作为未来扩展的备用，而是当存储集群不能满足日益增长的需求时再订购硬件进行扩展；其次，集群存储是通过软件实现的，因此能冲破大量的硬件限制，从而进行扩展；最后，集群可并行 I/O 至用户端机器，存储容量及读写性能可以线性的同步提高。

通过以上优势，形成了由先前的 NAS/SAN 到集群存储的模式转换。Power-All 的任务是赶超这种模式转换并通过使用低成本 PC 组件开发网络全球集群存储。为了实现这一目标，Power-All 已开发了一种核心技术，即 PGFS。

3. 集群存储分类

集群存储是代表由多个组件构成的存储池的通用术语。市场上存在着不同类型的集群存储产品及技术，以下分类为：

集群存储产品类型

A. 集群 SAN

集群 SAN 是指可扩展 SAN，可在数据块级进行存储扩展。集群 SAN 是价格昂贵的企业级产品。

B. 集群 NAS

集群 NAS 是指可扩展 NAS，可在文件系统级进行存储扩展。通常集群 NAS 通过集群文件系统完成。

集群文件系统

目前市场有很多不同类型文件系列归类为集群文件系统，但提供的功能却不同。如 CFS，主要分为两种类型：

A. 共享网络分块设备

对这种类型的 CFS，其目的为支持多个终端同时访问同一网络分块设备。传统来说，大多数网络分块设备技术不允许许多终端同时访问。CFS 提供锁机制以启用多终端同时访问。例如 Radhat 的 GFS 全球文件系统。

B. 分布式数据访问设备

对于此类 CFS，其目的为将数据进行分布以访问多重设备。与传统 NAS 相比，具有更高的扩展力与性能。从原则上讲，此类 CFS 性能更高与其容量呈线性关系。如 Sun Microsystems 的 CFS，Power-All 亦属此类产品。

4. PGFS 功能

什么是 PGFS

PGFS 全写：Power-All Global File System，是由 Power-All 网络有限公司开发的一项专有集群文件系统。PGFS 是 Power-All 全球云存储服务（Aspen Cloud Storage）的核心技术。

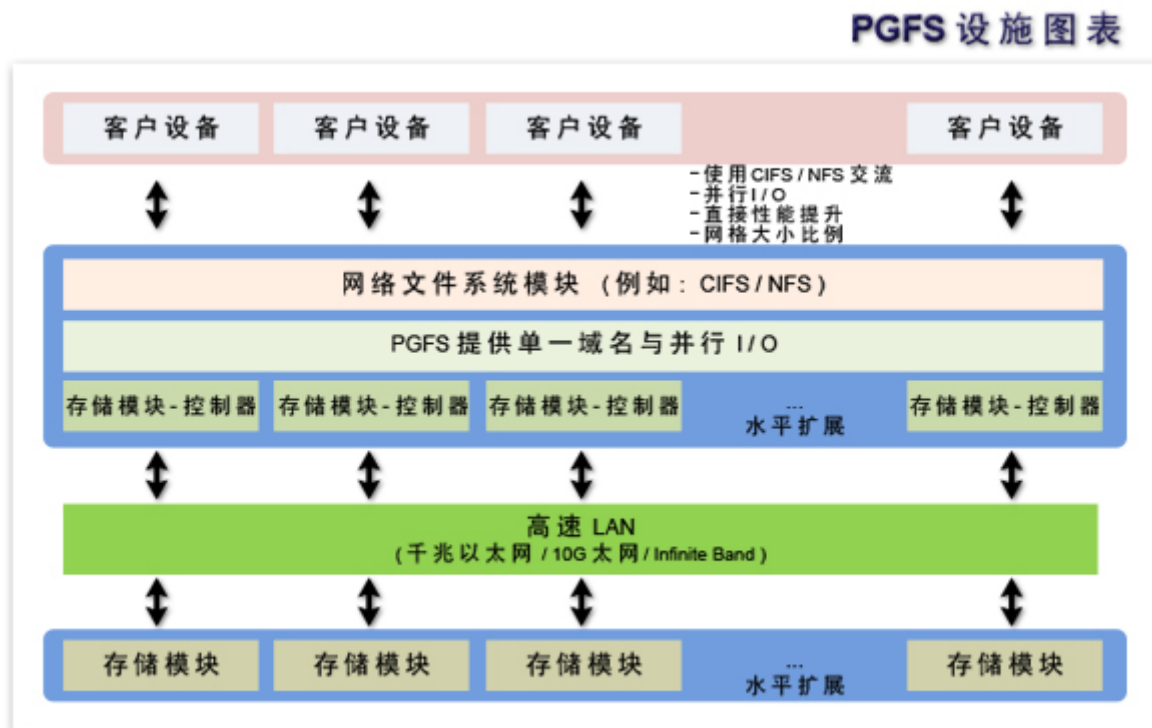
开发 PGFS 的目标是通过使用标准 PC 硬件组件创建具有高扩展性，高可靠性以及最佳性能的集群式 NAS。

PGFS 主要性能

- 可达 PB 级的高扩展性，数据被分布式访问集群内的所有存储模块。
- 数据高可靠性，数据实时复制
- 支持多向复制
- 无单一节点故障，PGFS 所有组件可水平扩展
- 其它许多集群文件系统，在一些组件上会出现单一节点故障或瓶颈故障。如存储控制器或 meta 数据服务器。PGFS 设计简洁，无需集中式元数据服务器，不存在以上问题。
- 文件系统的独立域名使应用集成更加简易
- 实现性能的线性扩展。许多集成文件系统具有瓶颈限制
- 性能增长与存储节点增长呈线性关系
- 运行于 IP 网络之上，说明 PGFS 可以运行于高质量的 WAN 连接之上
- 支持 10G 以太网及 InfiniBand

5. PGFS 架构
PGFS 核心组件

与其它集群文件系统不同的是，PGFS 架构简洁，易于操作。PGFS 仅包含两个核心组件，存储模块与存储控制模块。其架构如图 1 所示：

图 1


如图 1 所示：您可看到 PGFS 由存储模块与存储控制模块所组建。存储模块与存储控制模块之间由高速 LAN，如 GE，10GE，或 InfiniBand 连接。

5. PGFS 架构(續上)

存储模块规则

存储模块是指由大量硬盘所组配的装置。Power-All 开发了基于 CSAN/CNAS 系统的存储模块，是针对中小型企业标准存储产品。图 2 为存储模块示例：

CSAN/CNAS 系统存储产品易于操作，即时使用。CSAN/CNAS 系统与 Power-All 存储管理器（PSM）需提前安装，PSM 功能齐全，是处理该系统的管理与监测工具。CSAN/CNAS 其自身装有硬件 RAID5/6 控制器，对数据进行额外保护。CSAN/CNAS 硬件规格和功能，更多详情请访问 <http://www.powerallnetworks.com/>

图 2

存储模块图例



5. PGFS 架构(續上)

与存储模块的区别

存储模块是基于 CSAN/CNAS 之上的开发产品，两者之间的主要区别在于，存储模块与 PGFS 服务器端软件同时预装。在存储模块中，通过 Power-All 存储管理器（PSM）创建单一逻辑卷（LV）。依据简明性，我们建议 LV 大小应与 CSAN/CNAS 系统总大小相等。

存储模块当地文件系统

原则上，PGFS 是与存储模块的当地文件系统相独立。然而，根据 Power-All 的实验中，我们发现 XFS 是速度最快且最有效的当地文件系统。XFS 本身支持日志功能，防止突发故障时数据的丢失。XFS 是 CSAN/CNAS 系统唯一支持的内置文件系统，PGFS 的默认存储模块为 XFS 系统。

PGFS 服务器怎样与 PGFS 控制器相连接

PGFS 服务端软件运行于所建的逻辑卷上。该软件是 PGFS 控制器软件与物理存储之间的接口。通过使用 IP 层或 InfiniBand RDMA 上的 TCP 协议，从而进行 PGFS 服务器端与 PGFS 控制器之间的连接。

5. PGFS 架构(續上)

存储控制模块规则

存储控制模块是指 PGFS 的控制器。字母“C”代表 Controller，即控制器。它是物理存储（所有存储模块）与用户应用间的接口。存储控制模块具有以下性能：

- 集结多个存储模块至单一存储池内，拥有同一空间名
- 通过所有存储模块进行数据分割
- 提供存储模块间的多向实时数据复制
- 提供在线增加/减少存储模块数量
- 自我修复与监测

由于 PGFS 是专有文件系统，所有存储控制模块装置被提前安装在 Samba 与 NFS 服务器上。因此服务器支持 CIFS/NFS 通过 Samba/NFS 服务器软件与 PGFS 进行连接。在图 1 中，CIFS/NFS 位于 PGFS 连接至客户应用服务器之上。

在 WAN 或互联网上运行 PGFS

因为 PGFS 可在 TCP/IP 下运行，理论上讲，PGFS 也可通过 WAN/互联网运行。然而实际上，有许多因素影响其可行性，如：带宽，成本，以及网络延迟等。PGFS 通过 WAN 运行时会造成瓶颈障碍。WAN 带宽始终低于 LAN，因此总是造成瓶颈。通过当前互联网连接技术，Power-All 不建议通过 WAN/互联网使用 PGFS 剥离数据。

6. 数据与服务的高可用性

本章向您介绍 PGFS 如何实现数据的高可用性

实时数据复制

PGFS 支持多向数据实时复制。管理员可在不同数据区自定义复制策略。每一数据区为存储模块独立群，以独立方式运行。参考图 3，存储模块 1 和 2 被 A 区分离，存储模块 3 和 4 被 B 区分离。复制规则可配置在所有存储控制模块内，以进行两区间的实时复制。与其它具有 Master-Slave 设计的文件复制工具不同的是，PGFS 提供 Multi-Masters 处理装置。PGFS

拥有锁定机制以控制 Multi-Masters 数据同时写入。

存储控制模块根据用户自定义规则，决定复制数据的区域流向。数据从一个数据区内分离到所有存储模块。

除了复制，每一存储模块配备了硬件 RAID5/6 控制器，即使 1 或者 2 硬盘失效也能保持存储模块在线。

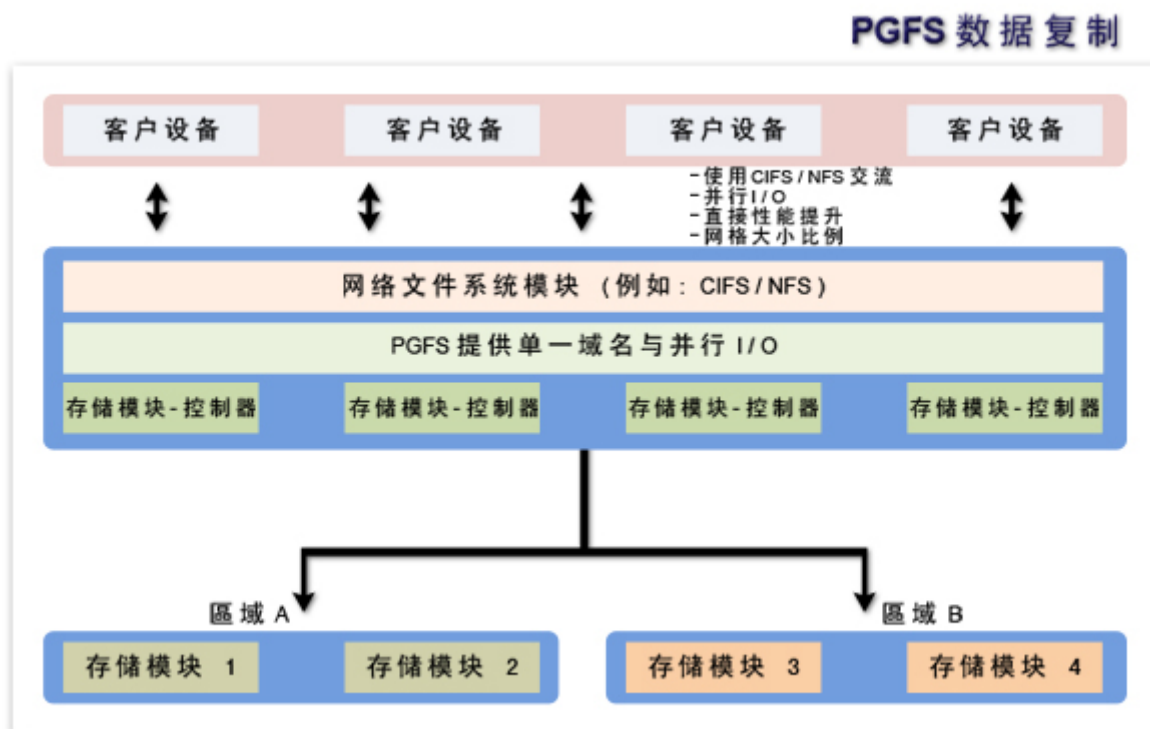
无单一节点故障

许多集群文件系统为元数据控制器组件。这一设计，会引起性能瓶颈及集群的单一节点故障。PGFS 集群无需任何元数据组件，其组件分为两部分：存储模块和存储控制模块。以上两种组件可在 PGFS 中进行水平扩展，数量可达上万装置。

6. 数据与服务的高可用性(續上)

如果其中任一存储控制模块失效，将不会对其它存储模块造成影响。在这种情况下，客户应用服务器应停止发送 I/O 至失效的存储控制模块。

如果存储模块下降，存储控制模块可在数秒内检测出失效节点并停止数据复制到此区域。如果区域全部失效，将引起整个集群失效。

图 3


7. 高扩展性

本节向您介绍 PGFS 如何实现存储系统的高扩展性

PGFS 是一个操作简易且功能强大的集群文件系统，无需元数据控制器。PGFS 可在大多数当地文件系统中运行，如 ext3, XFS 等。XFS 是用于所有存储模块的默认文件系统。

PGFS 仅包括两类组件：存储模块和存储控制模块。以上两类组件均可进行水平扩展，可达上万装置，其总大小可达 PB 级。

为了扩大存储总空间，可添加额外存储模块。首先，新存储模块与后端开关相连接（以太网或 Infiniband，根据存储模块接口而定）。其次，修改配置文件复制到所有存储控制模块中并进行重装。所有操作可在线完成。

8. 线性扩展生产率

本节向您介绍 PGFS 如何进行线性性能扩展

线性扩展量是 PGFS 的主要功能。通过存储模块与存储控制模块的增加，整体性能呈线性增长。

当 PGFS 增加额外的存储模块，其总存储量与总磁盘容量也随之增加。然而，一旦存储控制模块性能出现瓶颈故障，增加额外的存储模也无法另其性能提高。在这种情况下，管理员应增加存储控制模块的数量。

如何测定哪些组件需要升级？

为了扩展 PGFS，管理员应根据以下规则确定增加存储模块或存储控制模块。

- I. 如扩展存储容量，应增加存储模块的数量。
- II. 如磁盘 I/O 出现瓶颈，应增加存储模块数量。
- III. 如每一存储控制模块的带宽都被耗尽，应增加存储控制模块的数量。

9. 错误处理与数据恢复

本节向您介绍 PGFS 如何处理错误运行

PGFS 致力于向关键应用任务提供一个可靠的存储空间。错误处理是提供 7 x24 小时不间断运行环境最为关键的一部分。

硬盘驱动故障案例

每一存储控制模块配有硬件 RAID 控制器并预装在 Power-All 存储管理器 (PAM) 中。一旦出现任何硬盘驱动问题, PSM 可直接检测并向管理员发送警报邮件。As there is RAID5/6, Storage Module is still online with 1-2 hard drives.

存储模块故障案例

一旦整个存储模块出现故障, 存储控制模块可直接检测并向管理员发送警报邮件。如启用实时复制功能, 服务将不会中断。存储控制模块将停止向失效节点发送流量直至节点恢复正常。

存储控制模块故障案例

如果存储控制模块出现故障, 其它存储控制模块将保持在线状态继续提供服务, 所以不会对服务造成中断。在这种情况下, 客户端应用服务器应具有存储控制模块的故障检测机制并停止向失效节点发送流量。

10. 降低初始投资成本

本节向您介绍 PGFS 如何降低企业存储成本

摩尔定律 (Moore' s Law)

摩尔定律概述计算机硬件历史的重要趋势：芯片中晶体管的数量将成倍数增加，每两年翻一番。

存储方面，每一 GB 的价格也逐日递减。另外，硬件公司资产贬值之事也时有发生。基于以上情况，公司购买大批存储产品以供日后使用是极不划算的。PGFS 是此类问题的最佳解决方案。PGFS 支持在线存储扩展，当现存集群存储接近最大存储量时，可灵活增加额外存储模块。

基于 PC 组件

PGFS 设计可运行于标准 PC 组件。随着当前 PC 组件的快速变化，存储模块将随其市场产生更具成本效益的硬体组件。PGFS 可兼容不同版本的硬体组件。

11. 如何使用 PGFS 与当前第三方存储相结合

本节向您介绍现存非 Power-All 存储产品整合到 PGFS：

Power-All 熟悉企业可能拥有当前存储产品且不希望建立新的 PGFS 后将其闲置。在这种情况下，PGFS 提供特制咨询解决方案，即将第三方存储产品与 PGFS 进行整合。

从理论上讲，PGFS 可运行于所有文件系统之上。将第三方存储与 PGFS 整合，通常需要一种作为第三方存储产品与 PGFS 间的接口的服务器，即存储耦合器。

12. 总结

使用集群文件系统产生的公司效益

存储产业正经历由传统的 NAS/SAN 朝向集群存储的一种模式转换。使用集群文件系统企业可获得的效益为：

- 降低存储成本
- 更有效的企业数据管理
- 性能更高
- 更加可靠

市场上拥有很多集群存储产品，但大多数是为企业定制的昂贵的硬件组件。Power-All 的 PGFS 正开启一场存储革命，通过使用成本低廉的 PC 标准硬件建立具有同等企业性能的存储产品。是中小型企业存储的最佳解决方案。

13. 联络我们

关于 Power-All 网络有限公司

Power-All是集群存储行业的领导者之一，在IDC行业拥有坚实的操作经验。Power-All认为基于PC组件的集群存储是该行业的发展趋势并成为下一代主流解决方案。通过与PFFS和其它领先技术的结合，Power-All已开发了全球存储云服务，即Aspen Cloud Storage存储服务。更多详情，请浏览 <http://www.cloudwww.com/>

- 获得更多信息，请与 Power-All 联系：

地址：*Power-All Networks Limited*
香港新界沙田香港科学园科技大道西五号
企业广场 5 楼 540 及 541 室

电话：*(852) 2111 8182*

传真：*(852) 2111 8156*

邮箱：*newgen@powerallnetworks.com*