

Power-All Networks Clustered Storage Area Network: A scalable, fault-tolerant, high-performance storage system.

Power-All Networks Ltd

Abstract:

Today's network-oriented computing environments require high-performance, data storage systems that can satisfy both the data storage requirements of individual systems and the data sharing requirements of work groups and clusters of cooperative systems. The Power-All Networks Clustered Storage Area Network (CSAN), a scalable, fault-tolerant, high-performance clustered data storage system from Power-All Networks Ltd, is a distributed storage system that eliminates the performance, availability, and scalability problems that are present in many traditional distributed file systems. CSAN is a highly modular next generation storage architecture that provides a reliable, network-neutral data storage and retrieval solution. CSAN provides high I/O throughput in clusters and shared-data environments and also provides independence from the location of data on the physical storage, protection from single points of failure, and fast recovery from cluster reconfiguration and server or network outages.

1. Overview

Network-centric computing environments demand reliable, high-performance storage systems that properly authenticated clients can depend on for data storage and delivery. Simple cooperative computing environments such as enterprise networks typically satisfy these requirements using distributed file systems based on a standard client/server model. Current distributed storage systems such as SANS have been successful in a variety of enterprise scenarios but do not satisfy the requirements of today's high-performance computing environments. Today's high-performance computing environments are demanding high performance, high scalability, high fault tolerance and high availability. The CSAN distributed data storage system provides significant performance and scalability advantages over existing distributed storage systems. The name "CSAN" is an amalgam of the terms "Clustered" and "SAN". The "Clustered" portion leverages on today's supercomputer cluster architecture design which comprise multiple compute nodes and aggregate the overall processing power to provide supercomputer performance. In the computer cluster definition, a cluster is a group of homogeneous, whole computer systems running in concert to divide and conquer a computing task and used as a unified computing resource. In the CSAN architecture, multiple whole storage systems run in concert to store and retrieve data and are used as a unified data storage and retrieval system. The "SAN" is an abbreviation for "Storage Area Network". A storage area network (SAN) is a high-speed special-purpose network (or subnetwork) that interconnects different kinds of data storage devices with associated data servers on behalf of a larger network of users. The CSAN effectively works much like a SAN from the users perspective but performance is much higher and more scalable due to the CSAN using supercomputer "cluster" architecture.

The CSAN is a distributed data storage and retrieval system and its has well-known advantages. They decouple computational and storage resources, enabling desktop systems and application server systems to focus on user and application requests. Centralizing storage on the CSAN facilitates centralized system administration, simplifying operational tasks such as backups, storage expansion, and general storage reconfiguration without requiring desktop or server downtime or other interruptions in service. Beyond the standard features required by definition in a distributed file system, it also supports redundancy, which means that fail over services in conjunction with redundant storage devices provide multiple, synchronized copies of critical resources eliminating single points of failure. In the event of the failure of any critical resource, the storage system automatically provides a replica of the failed entity that can therefore provide uninterrupted service. This eliminates single points of failure in the distributed data storage system environment.

CSAN provides significant advantages over other aging data storage systems that preceded it. These advantages will be discussed in more detail throughout this paper, but are highlighted here for convenience. Most importantly, CSAN runs a cluster of storage nodes. This design provides a substantially more efficient division of labor between data retrieval and storage resources. High-availability algorithms are used to ensure accuracy and availability of data stored in the distributed data among the clustered storage nodes. There are also Initiators which are responsible for actual data system I/O and for interfacing with storage devices, which will be explained in more detail in the next section. This division of labor and responsibility leads to a truly scalable data storage system and more reliable recoverability from failure conditions by providing a unique combination of the advantages of cluster computing and distributed retrieval and storage systems.

2. CSAN Functionality and Architecture

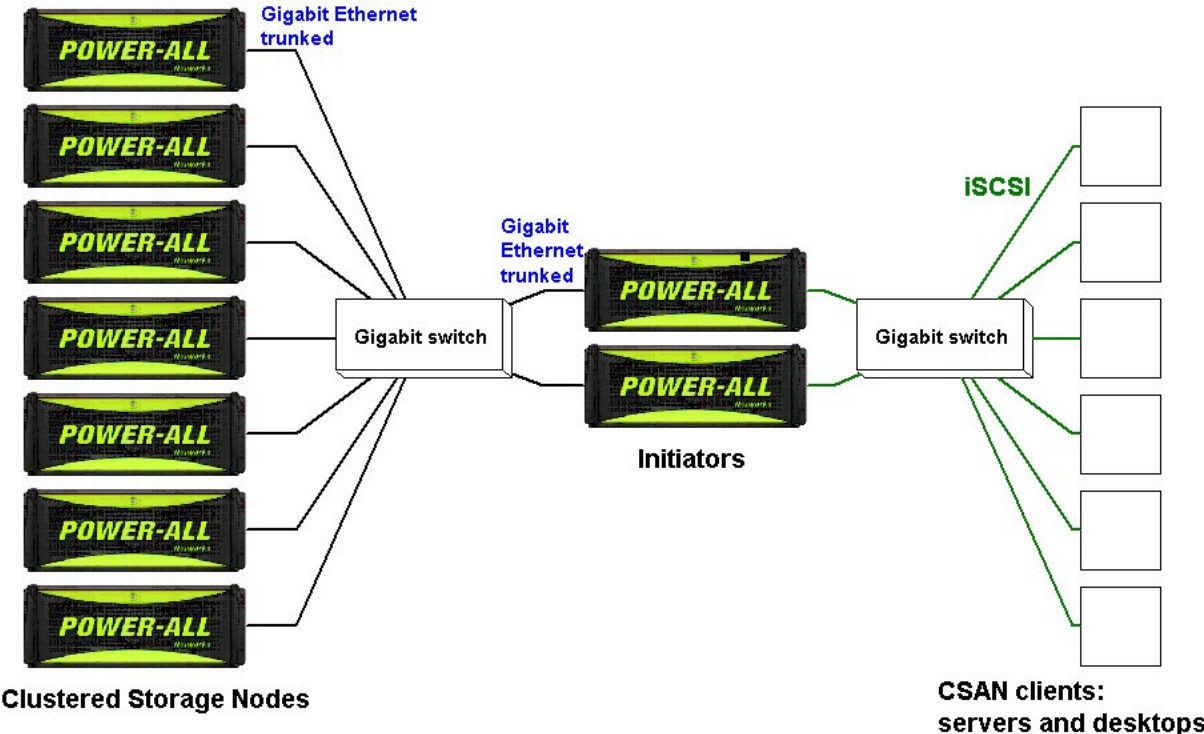


Diagram 2a, CSAN

- The CSAN architecture comprises of the following:
- 1) Clients which store the data in the CSAN.
 - 2) Initiators which interface to the clustered storage nodes and actual input and output of all data passes through the initiators.
 - 3) Clustered storage nodes which are the modular units which actually store the data and use their processing power, cluster-wide, to store and retrieve data.
 - 4) Gigabit switches which are used to connect the clients to the Initiators, and connect the Initiators to the clustered storage nodes. For higher performance, 10-Gigabit switches may be used here.

Like traditional SANs, clients such as servers and desktops connect to the Initiators using iSCSI. iSCSI is currently supported by all the major SAN manufacturers. iSCSI is Internet SCSI (Small Computer System Interface), an Internet Protocol (IP)-based storage networking standard for linking data storage facilities, developed by the Internet Engineering Task Force (IETF). By carrying SCSI commands over IP networks, iSCSI is used to facilitate data transfers over intranets and to manage storage over long distances. The iSCSI protocol is among the key technologies expected to help bring about rapid development of the storage area network (SAN) market, by increasing the capabilities and performance of storage data transmission. Because of the ubiquity of IP networks, iSCSI can be used to transmit data over local area networks (LANs), wide area networks (WANs), or the Internet and can enable location-independent data storage and retrieval.

When an end user or application sends a request, the operating system generates the appropriate SCSI commands and data request, which then go through encapsulation and, if necessary, encryption procedures. A packet header is added before the resulting IP packets are transmitted over an Ethernet connection. When a packet is received, it is decrypted (if it was encrypted before transmission), and disassembled, separating the SCSI commands and request. The SCSI commands are sent on to the SCSI controller, and from there to the SCSI storage device. Because iSCSI is bi-directional, the protocol can also be used to return data in response to the original request. In effect, iSCSI operates like a SCSI device to the end user desktop or server.

The Initiators interface to the clustered storage nodes and actual input and output of all data goes through the initiators. The data from the Initiators is then sent to the clustered storage nodes where it is distributed and stored. When an iSCSI request is sent to the Initiators, the message is translated into a cluster-wide data retrieval or storage request to the cluster storage nodes. In concert, these cluster storage nodes retrieve and send the data to the initiators. High-availability algorithms are used to ensure accuracy and availability of data stored in the distributed data among the clustered storage nodes. The initiators then send the data, using iSCSI, to the end user. In effect, the Initiators put the clustered storage data into iSCSI packets to send to the end user and vice versa. The Initiators, effectively, translate CSAN storage data into iSCSI packets.

The clustered storage nodes work in a cluster environment to process, store and retrieve data. The actual data is stored in these clustered storage nodes. Data is stored and distributed among the clustered storage nodes. Like the supercomputer clusters, the number of clustered storage nodes can easily scale very large to hundreds and more. These clustered storage nodes work in concert for the storage and retrieval of data and aggregately provides extremely high throughput. This division of labor and responsibility leads to a truly scalable data storage system and more reliable recoverability from failure conditions by providing a unique combination of the advantages of cluster computing and distributed retrieval and storage systems.

In contrast, traditional SAN systems use a backplane architecture where all data storage throughput must pass. Also, these traditional SANs have a limited number of processors to handle the throughput of the SAN. Furthermore, due to their backplane, these traditional SANs have limited performance due to the finite throughput of the backplane. These traditional SANs may allow large numbers of hard disks to be attached but overall performance is hindered by its backplane architecture and limited processing power. Therefore, these traditional SANs are not truly scalable data storage systems.

However, the Power-All Networks CSAN uses a clustering architecture, all components can have increased throughput and storage size by adding more units: switches, Initiators and Clustered storage nodes. The clustered storage nodes may each have up to 12 TeraBytes of storage and hundreds to thousands of these nodes may be used in concert. To connect these large numbers of clustered storage nodes, hundreds of switches may be used with Ethernet trunking from the clustered storage nodes and between switches. 10-Gigabit switches may also play a role in such high-throughput architecture to ensure no congestion and simplify the network by reducing the numbers of Gigabit circuits. Up to hundreds of Initiators may be used to connect to hundreds or thousands of clients. In total, this architecture is extremely scalable and throughput scales very well as all components of the cluster contribute. This division of labor and responsibility leads to a truly scalable data storage system

<i>Power-All networks CSAN</i>	<i>Traditional SAN</i>
Clustered architecture	Backplane architecture
Large scalable throughput using many processors	Limited throughput due to limited processing power
High throughput due to cluster architecture	Limited throughput due to backplane architecture
No single point of failure	Multiple single points of failure
Massively scalable and extensible	Limited scalability
Supports iSCSI	Supports iSCSI

3. CSAN Scalability and High Availability

To ensure high availability of the network connection from the CSAN clients to the CSAN Initiators, multiple connections are supported to multiple switches. The Initiators may have several connections to a number of switches to enable link availability (see Diagram 3a). Multiple switches may be added as needed to increase the scale of the connectivity to the CSAN Initiators.

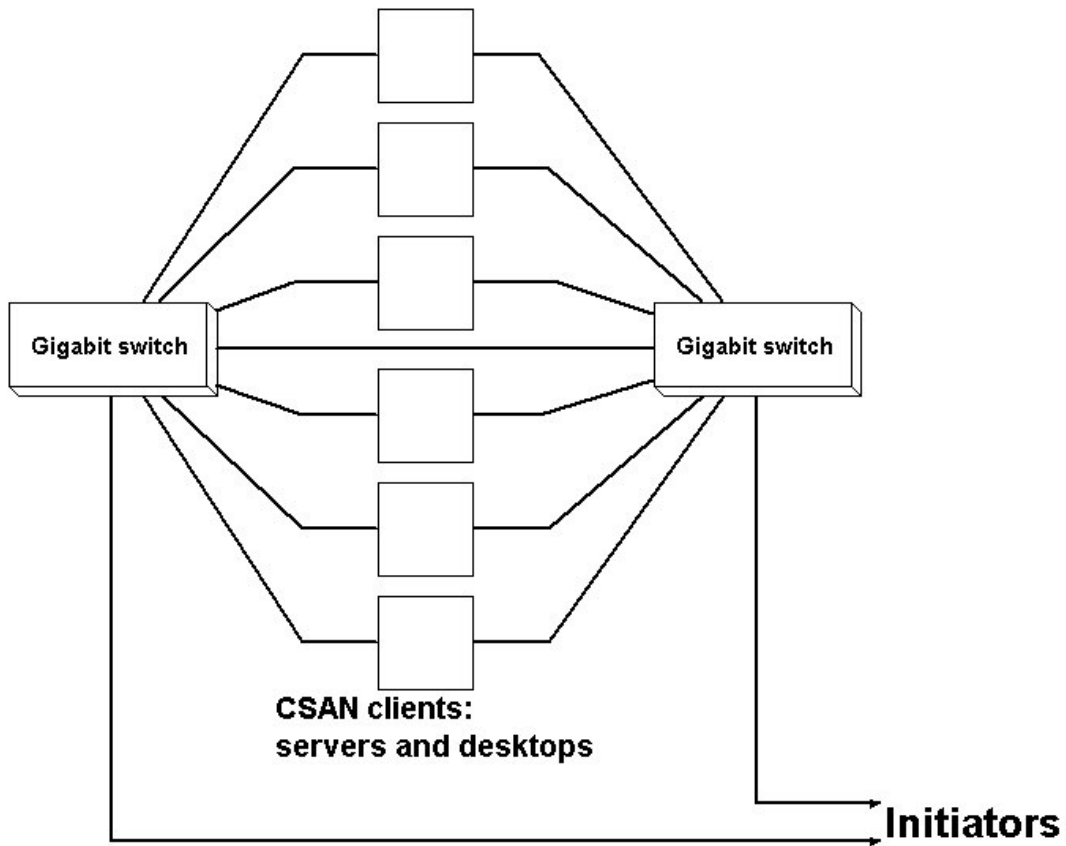


Diagram 3a, Clustered Storage Node availability

As all data passes through the Initiators, they have a large amount of input and output to handle. However, with the CSAN architecture, multiple Initiators may be used to ensure high performance and high scalability. In addition, for higher availability, the initiators may also have a “hot stand-by” Initiator to automatically replace a failed Initiator (see Diagram 3b).



Initiators



**Hot stand-by
Initiator**

Diagram 3b, Initiator availability

The clustered storage nodes connect to the initiators by Ethernet. The Ethernet may be Gigabit Ethernet (trunked or single) or 10-Gigabit Ethernet. The choice of which Ethernet connection depends on a balance between price and performance. Several trunked Gigabit Ethernet have performance which is a factor of the number of trunked Ethernet circuits: for example, four trunked Gigabit Ethernet have about 4 Gigabit per second of throughput. However, when more Gigabit Ethernet circuits are trunked, it may be more economically feasible to use a 10-Gigabit Ethernet circuit.

Furthermore, these clustered storage nodes may connect to multiple switches that connect to the Initiators. This network design ensures higher availability because the switches will not be a single-point-of-failure. The same applies to the Initiators where each Initiators is connected to multiple switches.

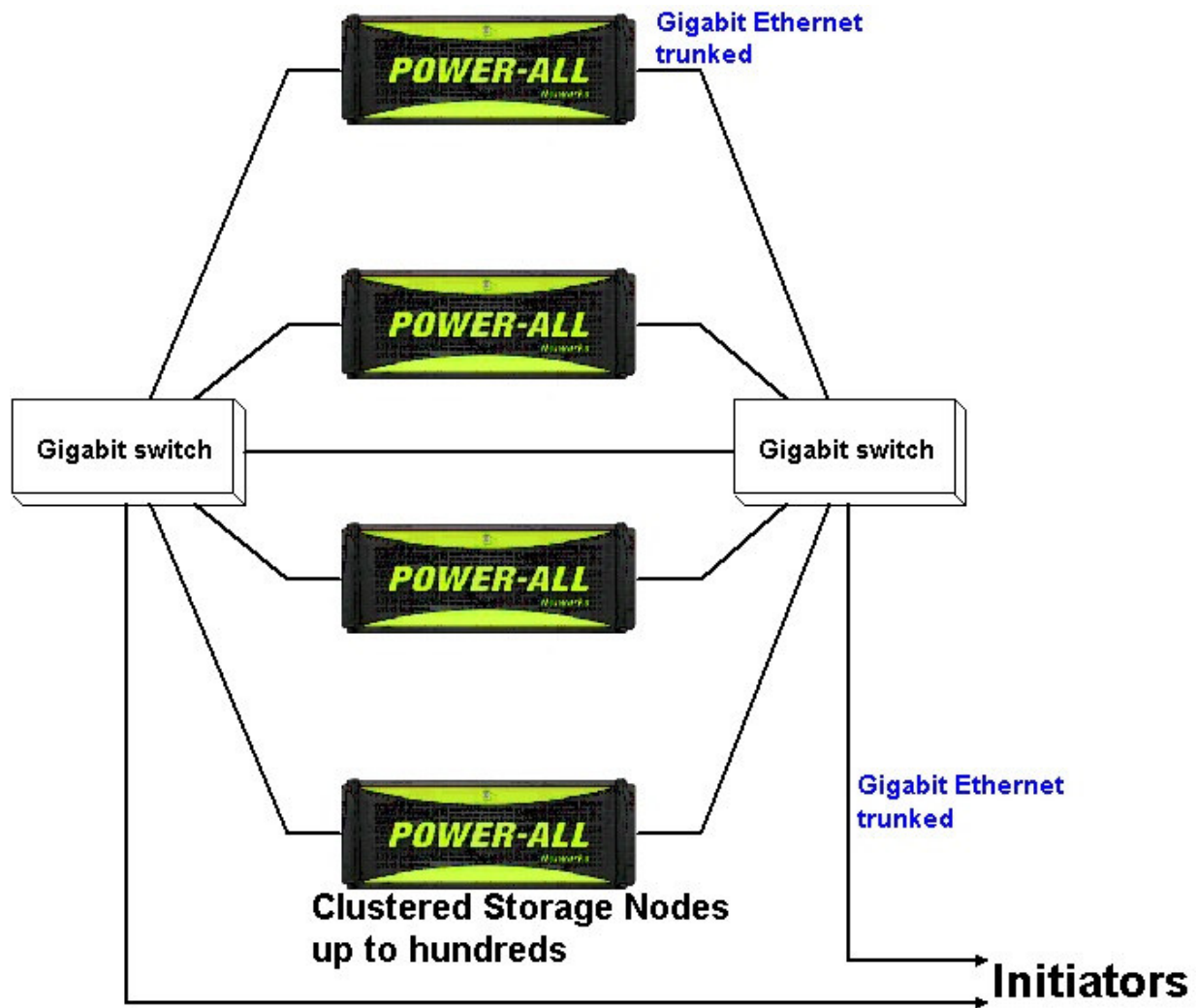


Diagram 3c, Clustered Storage Node network availability

The Power-All Networks CSAN employs “clustering” architecture to the data storage and provides an additional level of fault-tolerance and higher availability than current SAN technologies. There is no single point of failure as data is distributed among the cluster nodes. With built-in algorithms for data accuracy and availability, a failed cluster storage node will not result in loss of data or availability. In fact, individual clustered storage nodes may be taken out of the cluster for maintenance or repairs with no interruption of service nor data loss. Furthermore, for faster recovery of a failed clustered storage node, an extra cluster storage node may be set aside as a “hot stand-by” to automatically replace any failed clustered storage node.



Clustered Storage Nodes



**Hot stand-by
Storage Node**

Diagram 3d, Clustered Storage Node availability

For fault tolerance, each clustered storage node has an array of hard disks in a RAID Level 5 or 6 configuration. RAID, short for Redundant Array of Independent (or Inexpensive) Disks, is a category of disk drives that employ two or more drives in combination for fault tolerance and performance. RAID Level 5 -- Block Interleaved Distributed Parity: Provides data striping at the byte level and also stripe error correction information. If a data disk fails, the parity data is used to create a replacement disk. This results in excellent performance and good fault tolerance. Level 5 is one of the most popular implementations of RAID. RAID Level 6 -- Independent Data Disks with Double Parity: Provides block-level striping with parity data distributed across all disks. This is more fault tolerant than RAID 5 as it can maintain data availability even with two simultaneous hard disk failures. RAID 5 is common to many existing SANs but RAID 6 is not widely available.

4. Summary

The Power-All Networks Clustered Storage Area Network (CSAN), a scalable, fault-tolerant, high-performance clustered data storage system from Power-All Networks Ltd, is a distributed storage system that eliminates the performance, availability, and scalability problems that are present in many traditional distributed file systems. CSAN is a highly modular next generation storage architecture that provides a reliable, network-neutral data storage and retrieval solution. CSAN provides high I/O throughput in clusters and other data environments and also provides independence from the location of data on the physical storage, protection from single points of failure, and fast recovery from cluster reconfiguration and server or network outages.

Future direction of the Power-All Networks CSAN include a high performance architecture where data is tiered

to have a 4 dimensional (4D) architecture in clustering. This 4D clustering of data storage and retrieval is expected to provide a better than exponential increase in throughput with cluster units are added. In addition, future designs will incorporate existing SANs as components of the CSAN architecture. This design would ensure no “forklifting” is needed to move out the old SAN to replace it with a bigger storage system. The use of an existing SAN in the newer CSAN environment maximizes return on investment.

Author: William Kam, CTO. Power-All Networks Ltd.
CSAN white paper version 1.02.
22 June, 2005.
Power-All Networks Ltd Copyright 2005.