SurFS Product Description

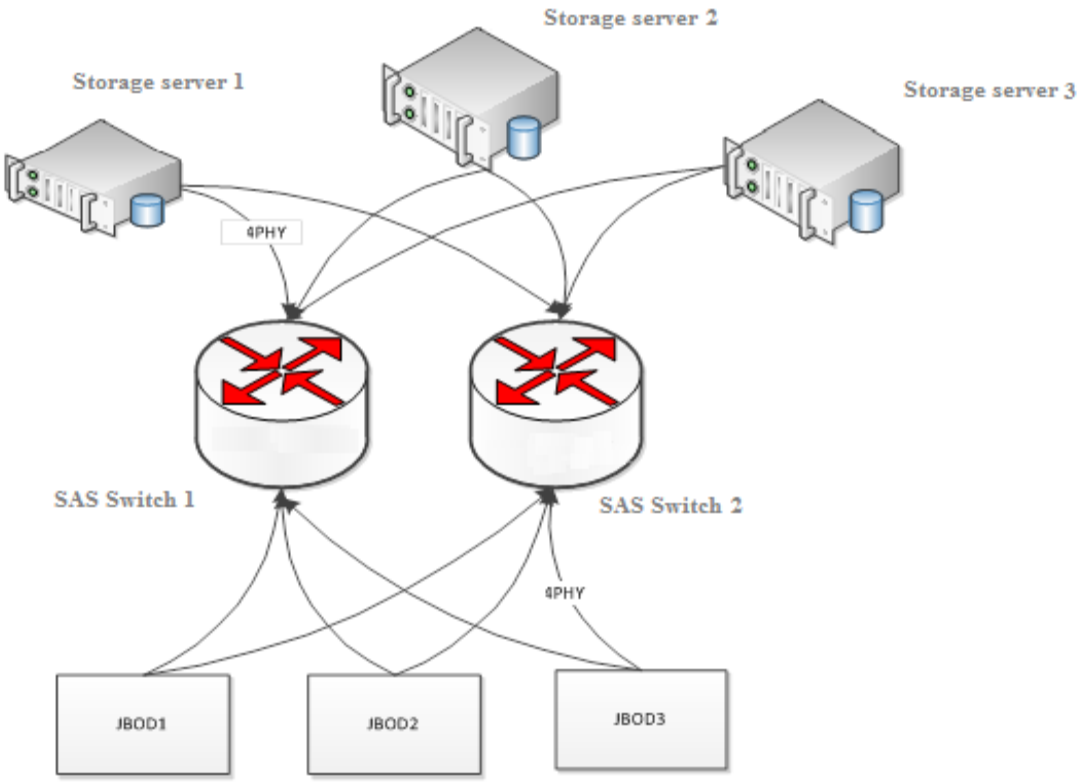
1. ABSTRACT

SurFS — An innovative technology is evolving the distributed storage ecosystem. SurFS is designed for cloud storage with extreme performance at a price that is significantly less expensive than traditional distributed storage systems. It is able to cover enterprise-class SAN and NAS as well.

SurFS uses SAS storage network, separates storage control node from storage medias (such as HDD or SSD), and converges storage control node and compute node, Ultra short I/O path, large bandwidth(24Gb/48Gb) and short latency together with fast re-balancing are the essential factors that constitute an ultra-fast storage system. Much lower redundancy rate with same data durability, lesser storage control node, and cheaper storage network are the essential factors of ultra lower cost. SurFS also incorporates a redundant design that guaranteed high availability and a reliable scale-out storage system. SurFS storage is highly efficient for virtualization and cloud environments, which will enable cloud vendors to take full advantage of this innovative technology. Last but not least, SurFS is 100% compatible with OpenStack.

2. INNOVATIVE HARDWARE ARCHITECTURE

The SurFS distributed storage system is based on a SAS (Serial Attached SCSI) network, instead of traditional Ethernet. All hardware devices are connected to the SAS switch by wide port.



SurFS is build on three main devices:

1) SAS switch: The LSI 6160 SAS-2 switch offers 16 SAS-2 4-wide ports, each narrow port has 6Gbit/s bandwidth, thus a wide port has 24Gb/s of total bandwidth. Its retail price is around $2,000; more advanced and less expensive SAS switches will be available in the near future, e.g. a 68 4-wide ports SAS-3 switch which provides 48Gb/s bandwidth per wide port.
2) JBOD (Just a Brunch Of Disks): JBOD is a storage device which holds many hard drives without any built-in intelligence, such as mainboard, CPU or memory. E.G., the SuperMicro 847E16 JBOD can hold up to 45 3.5" SAS or SATA disks with 4-wide SAS-2 port. JBOD is connected to the SAS switch by a 4-wide port.
3) Storage server: the storage server is a normal x86 server with the SurFS storage control node software installed. It is connected to the SAS switch through HBA card by 4-wide port too. The storage server is recommended to deploy compute nodes on it (such as virtual machines or containers) to constitute a hyper-converged server.

- **SAS Switch**

  SAS switches are the core components of the SurFS storage network. Storage servers communicate with JBODs via the SAS switches. All

devices are connected to the SAS switch by 4-wide port, thus the bandwidth between SAS switch and the storage control nodes is 24Gb/s (SAS-2) or 48Gb/s (SAS-3). SurFS is able to use two SAS switches to implement a redundant design, thus there are two physical links between each storage server and each JBOD. If one physical link is unavailable, then the other one is still in service.

- **JBOD**

  All mounted disks are part of the JBOD, connected by the SAS switch, which means that all disks in one JBOD will share 24Gbit/s (SAS-2) or 48Gb/s (SAS-3) bandwidth.

- **Storage Control Nodes**

  The SurFS storage control node software is running on Linux at the storage server. SurFS storage control node software includes the ZFS file system, management tools, and the NAS server. ZFS is an advanced file system; find more info at http://zfsonlinux.org/.

  NOTE: Every disk in the JBODs could be accessed by every storage control node at the same time, which is a very important feature. E.g., when a server is unavailable, another server will get control of the data within seconds, including the virtual machine image and block devices.

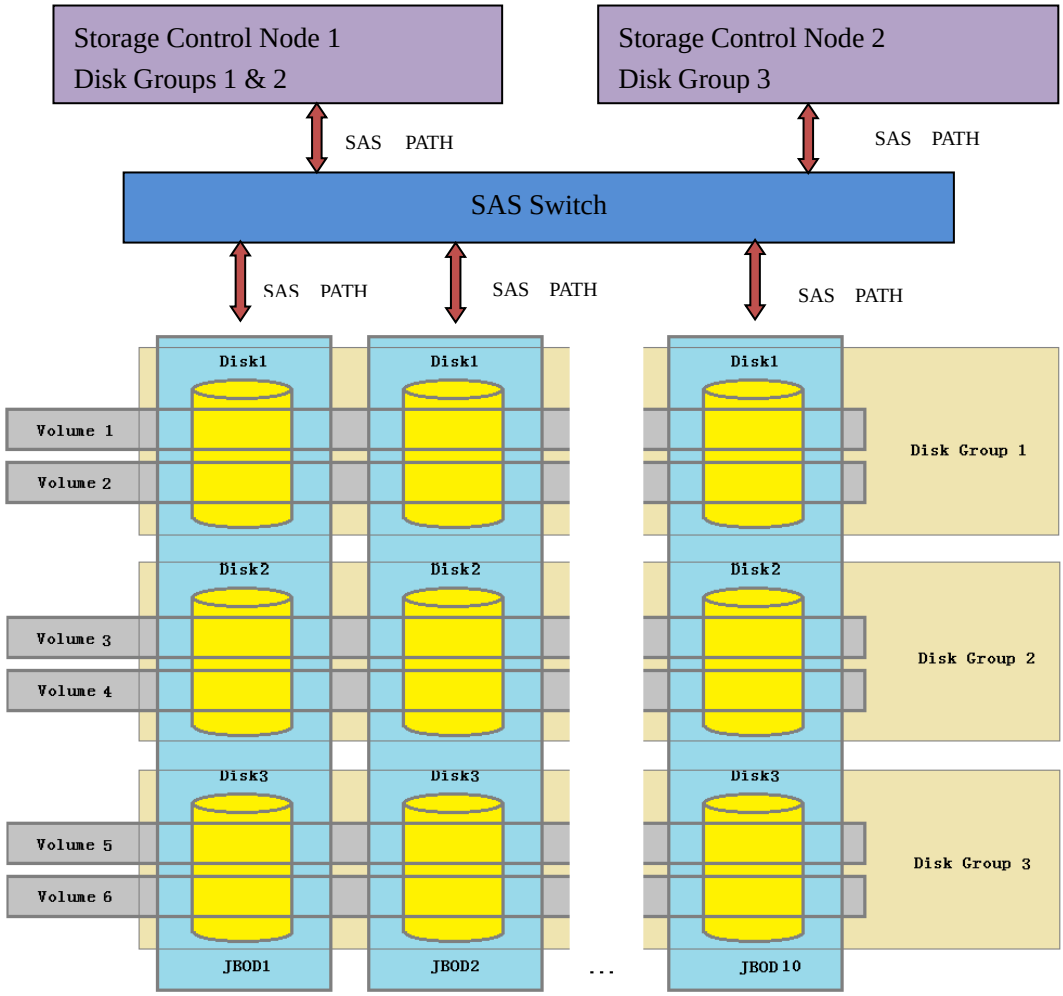
- **Disk Groups**

  A disk group is composed by multiple disks (one disk from each JBOD). The disks at same position of each JBOD compose a disk group, e.g., if there are ten JBODs in the system, all ‘disk 1’ of each JBOD will compose a disk group of ten disks, and all ‘disk 2’ of each JBOD will compose another disk group. Redundant mechanisms (such as RAID and erase code) can be set up for these disk groups as safeguard against both disk failures and device failures.

3. SOFTWARE ARCHITECTURE

- **Storage Server OS**

  The SurFS storage server uses Linux as operating system, CentOS is recommended.



- **SurFS Storage Control Node**

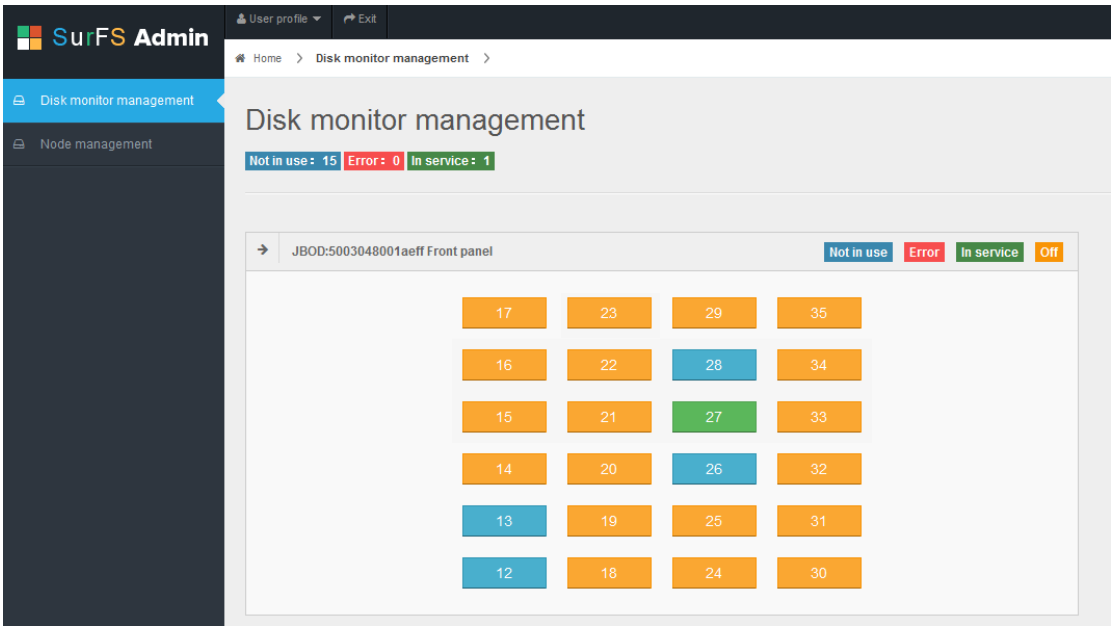
  SurFS storage control node software is installed on each storage server. Each storage control node is allocated to one or more disk groups to build a storage pool. Disk groups are managed by ZFS. One or more volumes are created on each disk group and each volume has a unique ID as shown in the figure above. There are two servers and ten JBODs in the system with four disks in each JBOD. All ‘Disk 1’ in each JBOD compose ‘disk group 1’ and so on. Thus there are three disk groups with ten disks per disk group. ‘storage control node 1’ has the control of ‘disk group 1’ and ‘disk group 2’ and ‘storage control node 2’ has the control of ‘disk group 3’. Each storage control node builds a ZFS file system on all disk groups that it manages.

  ZFS provides redundant storage ability for a disk group. ZFS RAIDZ-1 is similar to RAID 5, which allows one disk failure of a disk group to maintain the data; ZFS RAIDZ-2 is similar to RAID 6, which allows two disk failures of a disk group; ZFS RAIDZ-3 allows three disk failures of a disk group. A statistical evaluation shows that the durability of RAIDZ-3 is very close to ‘three copies’ (both are around 99.999999999%). With the system outlined

above a RAIDZ-3 disk group in 7+3 configuration has a redundancy rate of 10/7=143%, which is much better than ‘three copies’ with a redundancy rate of 300%.

If ‘storage control node 1’ becomes unavailable, then ‘storage control node 2’ will get the control over ‘disk group 1’ and ‘disk group 2’, until ‘storage control node 1’ gets back to service. This fail-over strategy of SurFS guarantees to meet high availability requirements.

SurFS offers a monitoring web UI, where users can see all disk statuses. Users can also check for status of the storage servers.

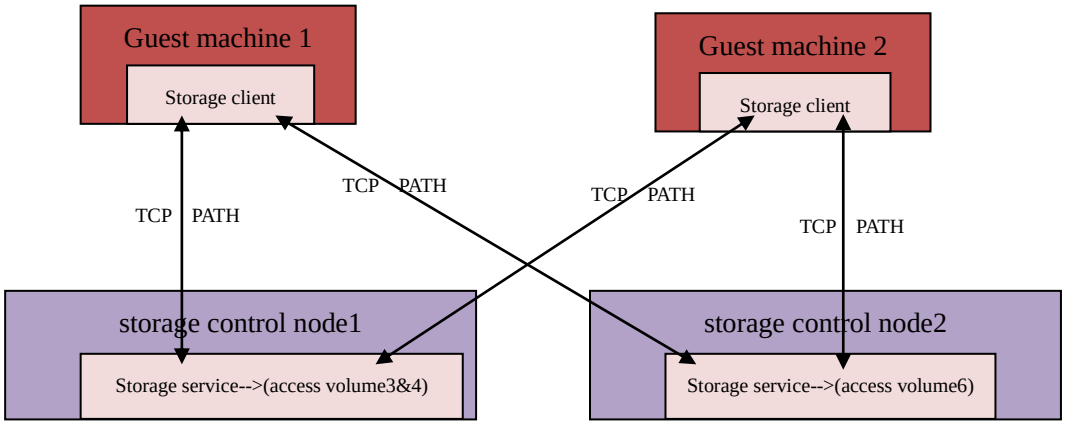


NOTE: ZFS is currently used to manage the raw storage in the storage pool. SurFS is planning to manage disks directly in future to obtain better performance and flexibility.

- **Block Storage Service**

  Each ZFS volume can be exported as a high performance, high durability, high availability and large capacity block device and can be configured with SurFS command line tools.

  Client computers and storage servers are connected by the iSCSI protocol. In a hyper-converged model, the compute node is able to mount this block device directly to gain higher performance than iSCSI.



SurFS offers a set of command line tools used to manage block device services, such as:

- Global information of SurFS block storage
- Pool management
- Volume management
- Snapshot management
- Management of volume export

- **NAS Storage Service**

  SurFS provides a NAS service via the standard CIFS and NFS protocols. SurFS NAS server and SurFS NAS client work together to provide the NAS service. The SurFS NAS server is a part of SurFS storage control node, it manage some or all of the ZFS volumes of its storage control node. The SurFS NAS client works with the compute node. SurFS NAS client and SurFS NAS server can be located in same server or different servers. A SurFS NAS client is able to connect to multiple SurFS NAS servers, the compute node connected with SurFS NAS client is able to use all of space of all connected SurFS NAS servers, so that it is easy to expand the size of NAS storage by adding new volumes to the SurFS NAS server or adding new SurFS NAS servers to the SurFS NAS client. In the system shown in first figure, ‘volume 1’ and ‘volume 2’ can be allocated to the SurFS NAS server of ‘storage control node 1’, allocate ‘volume 6’ to the SurFS NAS server of ‘storage control node 2’ (volume 3, volume 4 and volume 5 are reserved to be exported to block devices). A NAS client which connects both the NAS server of ‘storage control node 1’ and the NAS server of ‘storage control node 2’ can provide NAS storage with a volume size of ‘volume 1’, ‘volume 2’ and ‘volume 6’ summed up.

  The NAS storage offers storage control node load balancing. If a compute

node sends a write request, the SurFS NAS Client needs to determine which SurFS NAS Server to write to, the write strategies are:

1) Round Robin
2) Nearby, finding the SurFS NAS Server with the best bandwidth. In the hyper-converged model, nearby strategy will find the storage node located in same physical server, so that it will be the best performance strategy in this case.
3) Dynamic polling: the SurFS NAS Client will collect the I/O loading status of all SurFS, then taking the node with the lowest load to write to.

NOTE: The SurFS NAS Client has a separated SurFS-NAS-Protocol component https://github.com/surcloudorg/SurFS-NAS-Protocol. It is a separated open source project based on Alfresco-JLAN which uses different license agreement from SurFS.

4. SURFS ADVANTAGES

The SurFS storage architecture offers a much higher performance at much lower costs than existing distributed file systems, because it takes full advantage of the SAS technology, allows RAID or erase code for data durability, separating storage control node from storage medias and unique global storage pool features.

- **Low Cost**

  The RAIDZ-3 model guarantees almost the same data durability as ‘three copies’ but at less than half redundancy rate, thus it has less than half costs.

  Even under the same redundancy rate SurFS is still much less expensive than traditional storage systems because a SAS network is much cheaper than an FC/IB network of the same bandwidth (24Gb for SAS-2, 48Gb for SAS-3). And in the recommended hyper-converged model the storage control node is running on computing server instead of storage device (in most cases, the number of storage devices is much higher than the number of computing servers, and storage control node can share the computing resource with compute nodes).

- **Large Capacity**

  In SurFS storage system, the disks are connected to a SAS storage network by JBOD. A single SAS zone could hold up to 500 hard disks

(exact number depends on the number and type of JBODs), which will get 3PB capacity by using 8TB hard disk and 9+3 RAIDZ-3 redundant strategy. However, multiple SAS zones are allowed, thus the capacity of single cluster can reach 10PB or more.

- **High Performance**

  In SurFS storage system, all the components are connected to SAS switch by 4-wide ports, each port will get a 24Gb or 48Gb bandwidth, which is several times more than a 10GbE network. At the same time, the SAS protocol has much shorter (10 times or more) latency (around 50ns) than TCP/IP.

  There is another reason why SurFS has good performance: SAS is the native interface of all hard drives and SSD, SAS network doesn't need any conversion between different protocols.

- **High Durability**

  RAIDZ-3 can reach the highest standard of data durability of 99.999999999% under the condition that defect disk can be replaced on time.

- **High Availability**

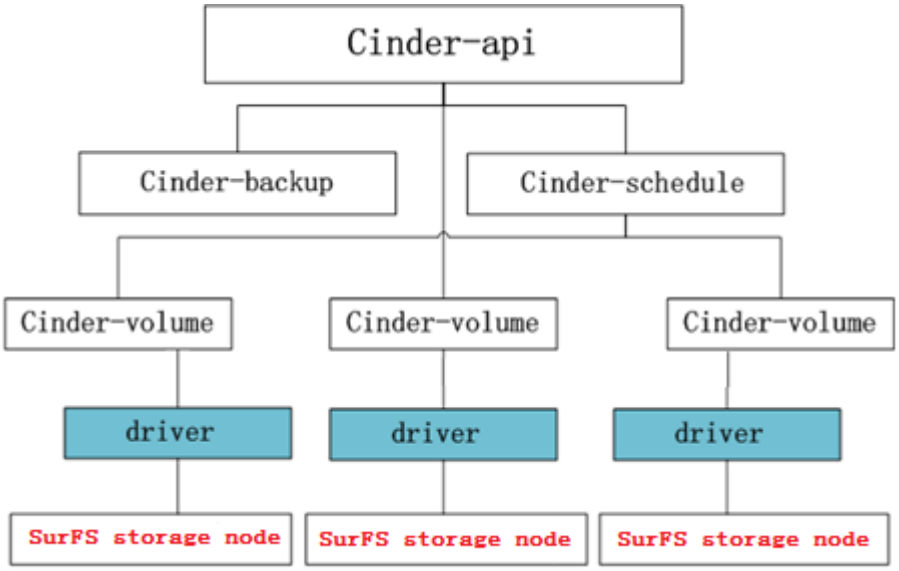
  1) With regard to disk level failures, since all the volumes are organized by ZFS RAID-Z, the volumes will always available as long as the number of failed disks does not exceed the redundancy degree within one disk group.
  2) With regard to the JBOD level failure, it means there is one disk offline for all disk groups. All volumes are still available as long as there are no more than three unavailable JBOD at same time for RAIDZ-3.
  3) With regard to the storage server failure, all the volumes in this node will migrate to other nodes within seconds until the storage server get back to service.
  4) With regard to the SAS switch or SAS channel failure, another SAS network will be used to ensure the availability of whole system.

- **Different access way and unified storage backend**

  Based on SurFS storage, users can implement different kinds of storage services, including NAS storage service, block storage service and object storage service (object storage service does not belong to the SurFS open source project). Because of the unified storage backend, all the above services can get the advantage of low cost, large capacity, high performance, high durability, high availability as well simplifying the system installation and maintenance.

5. SURFS FOR OPENSTACK

SurFS's block device service is optimized for OpenStack. SurCloud has contributed the SurFS cinder-volume drivers to the OpenStack community.



SurFS features:

Type	Features
volume	Create volume
	Clone
	Volume extend
	Delete volume
volume-VM	Attach to VM
	Detach from VM
snapshot	Create snapshot for a volume
	Create a volume from a snapshot
	Delete volume
image	Create volume from a image
	Create an image for a volume
volume migration	Move volume from A to B

Users can optimize the performance by converged architecture. The VMs will access the disks directly through the SAS path to increase the IOPS and bandwidth of I/O throughput.