EMC Isilon OneFS: An Ops Manager's Introduction

http://www.emc.com/collateral/hardware/white-papers/h10719-isilon-onefs-technical-overview-wp.pdf
http://www.emc.com/collateral/TechnicalDocument/docu52911_Isilon-Site-Preparation-and-Planning-guide.pdf

Theory

The Isilon product delivers scale-out NAS using a distributed file system running over a symmetric cluster. The file system employs a mix of mirroring and Reed-Solomon erasure codes as its parity scheme for delivering fault-tolerance. In the language of the CAP theorem, OneFS delivers Consistency and Availability, sacrificing Partitioning in the face of hardware failure (retaining read/write functionality only so long as a simple majority of nodes survive). OneFS leverages the underlying cluster to excel at streaming read/write throughput and concurrency at the cost of transactional performance. Inspired by the *nix philosophy, Isilon clusters support rich flexibility, allowing fine-grained hardware and software customization to support a wide range of workflows along with a rich tool set for optimizing performance. As nodes are added, the cluster's processing, caching, and IO capabilities increase, along with the efficiency and resiliency of its file system layout.

Protection Levels

The administrator specifies a cluster's *Protection Level*: how many simultaneous failures of disks and/or nodes a *DiskPool* within the Cluster can tolerate before data loss begins. OneFS responds to this setting by striping data appropriately. In the event of hardware failure, or the administrator changing the *Protection Level*, the *FlexProtect* job runs, rebuilding the stripes as needed. Choosing a low *Protection Level* increases available capacity while simultaneously increasing the risk of data loss. EMC recommends careful attention to this choice.

Common Ways to Degrade Your Cluster

- Consume more than 80% of the cluster's available space
- Exceed the cluster's resources in terms of CPU, RAM, and/or IOPS
- Redline the cluster and then kick-off a resource-intensive job, like *FlexProtect* or *AutoBalance*
- Set the *Protection Level* below your business' tolerance for risk and for data loss
- Dawdle (or hit typical supply chain / delivery delays) when replacing failed disks or nodes
- Deploy suboptimal power delivery & cable management strategies
- Employ complex configurations: VLANs, identity management, and *Access Zones*

These choices interact synergistically to increase the chance of cluster down time and data loss, i.e. pick two or more to substantially increase the odds of knocking out your cluster

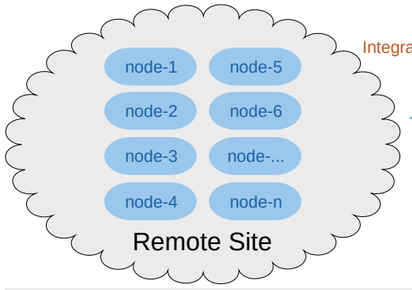
Optional Services

Customers can choose to add supporting services, such as *SyncIQ*, *InsightIQ*, or third-party applications like anti-virus and auditing

InsightIQ
Performance monitoring & reporting station

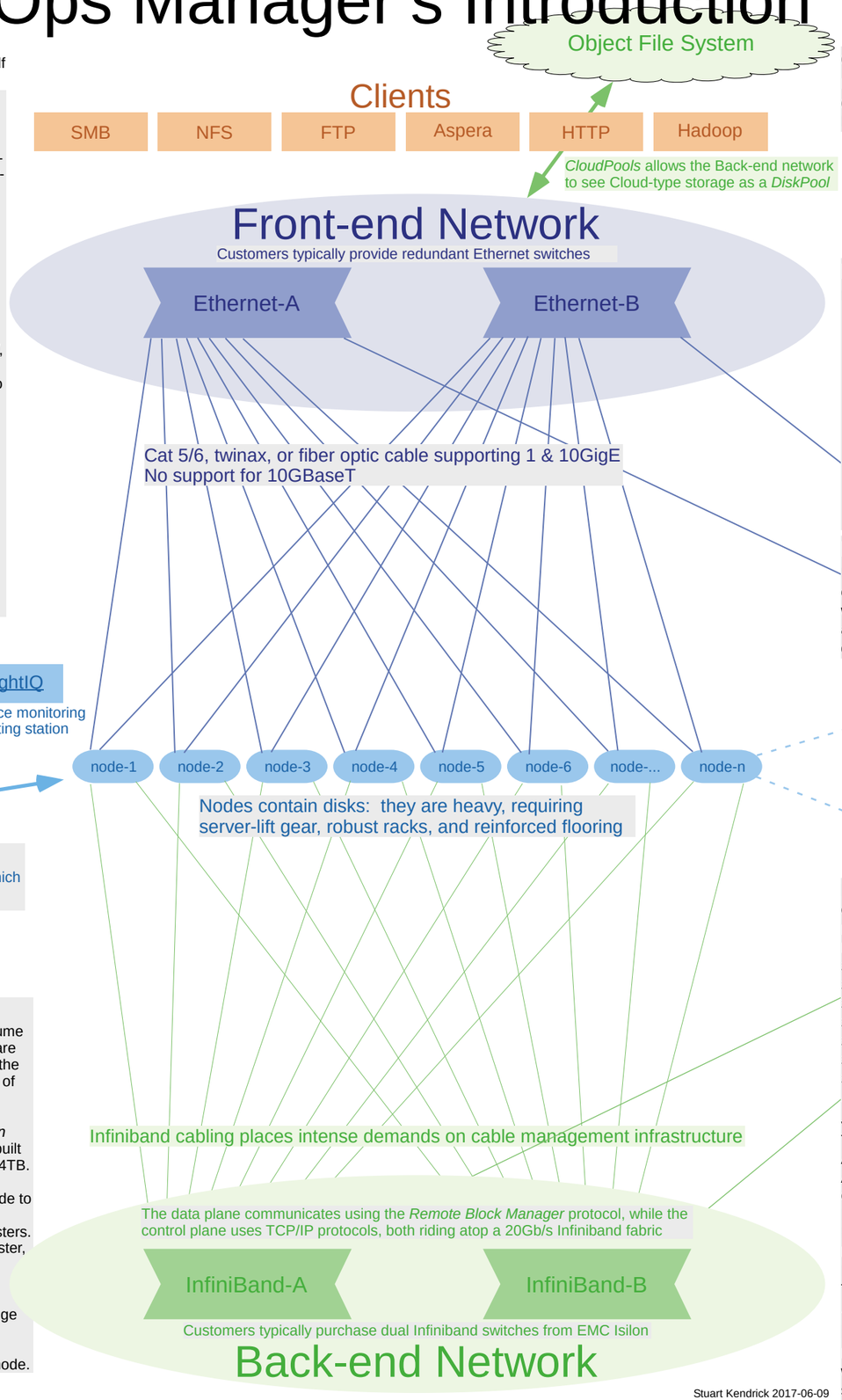
ICAP
Integration with anti-virus servers

Audit
Integration with audit servers



Replication

Replication between clusters uses *SyncIQ*, which leverages each node to parallelize data flow



File System

Each node contributes its data disks to the global file system */ifs* (Isilon File System), with no intermediate volume or partitioning abstraction. Files and their associated parity are chunked into 128K *Protection Groups*, which are striped across multi-drive *DiskPools* spread between multiple nodes according to algorithms which first honor the *Protection Level* settings and then optimize performance. Where needed, these algorithms will mirror instead of stripe; metadata is always mirrored at 8x. Storing one of those metadata mirrors on SSD (*Global Namespace Acceleration*) is a popular performance tweak. As nodes are added and removed, the available space in */ifs* expands and contracts automatically while the *AutoBalance* and *FlexProtect* jobs modify and shuffle *Protection Groups* to continue to meet *Protection Level* guarantees and performance strategies. OneFS contains no in-built limitation to the size of */ifs*, number of files, or breadth/depth of directory trees. File size is currently limited to 4TB.

Each node maintains a cluster-coherent view of the file system in terms of {node, disk, offset}, allowing the node to which a client is attached to initiate its reads & writes, reaching across the Infiniband fabric to other nodes as needed. From the File System point of view, all nodes are peers -- there are no metadata or coordinating masters. File locking and locking coordination is similarly distributed. The administrator can tune read caching on a cluster, directory, or file level to optimize for concurrency (adaptive algorithm), streaming, or random IO.

OneFS makes heavy use of caching and delayed writes to improve performance. All writes pass through the *Journal* on their way to disk; battery backup allows OneFS to treat all writes as synchronous, i.e. to acknowledge commit-to-disk before actually writing bits to spinning rust.

If the cluster drops below quorum, defined as a simple majority of nodes, OneFS places itself into read-only mode.

Customer-contributed services & gear

EMC-contributed services & gear

Operational Services Integration

Isilon clusters require tight integration with DNS servers in order to load-balance clients across nodes (the *SmartConnect* function), depend on a reliable NTP time hierarchy, and rely on NIS/LDAP/Kerberos/Active Directory services for authentication and authorization.

DNS NTP
Clusters rely heavily on robust DNS and NTP infrastructure

LDAP AD KRB NIS
Clusters consult one or more authentication / authorization providers

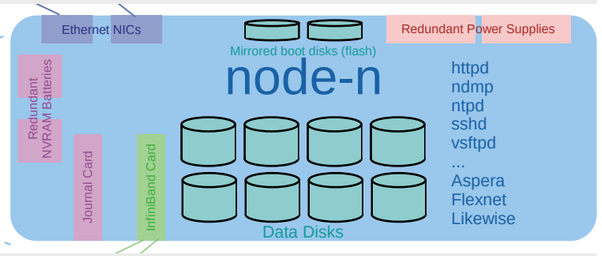
Node Types (Generation 5)

Nodes are configured-to-order with varying mixes of drive types, drive sizes, and RAM

S-Series	SAS drives	IOPS-intensive applications
X-Series	SATA drives	Throughput intensive applications
NL-Series	SATA drives	Near-Line storage: Capacity
HD-Series	SATA drives	Cold storage: Capacity + Density
A-Series	Backup Accelerator	Integrate with FC Data Protection, no disks
A-Series	Performance Accelerator	Increase front-end processing & caching, no disks

Node Design

Nodes consist of lightly customized PC server hardware, equipped with a dual-port Infiniband (IB) card, a *Journal* card (aka NVRAM), 6-256 GB RAM, and redundant batteries to keep uncommitted writes in the *Journal* alive in the event of power loss. Most nodes also contain disks -- SATA, SAS, and/or SSD. Nodes run OneFS (a FreeBSD derivative), standard daemons, and a slew of custom daemons.



Licensed Features

CloudPools	Present Object File Systems to the Back-end Network as *DiskPools*
InsightIQ	File system metrics and performance trending
Isilon for vCenter	VMWare integration
SmartConnect	Client load-balancing across nodes
SmartDedupe	Space conservation within directories
SmartLock	WORM in support of regulatory requirements
SmartPools	Automated tiering between node pools of varying resources
SmartQuotas	Limit utilization based on directories, users, and groups
SnapshotIQ	Copy-On-Write snapshotting strategy
SyncIQ	Replication between Isilon clusters with failover/failback
...	

Job Engine

This scheduler runs numerous cluster maintenance processes

AutoBalance	Redistribute data to more effectively leverage spindles
AVScan	Initiated by ICAP servers
Collect	Return deleted blocks to the free pool
DeDupe	De-duplicates identical blocks within a directory
FlexProtect	Restripe data and parity after disk/node failure, replacement, or addition
IntegrityScan	Sanity-check and repair file system and block layout metadata
MediaScan	Scan drives for media-level errors
TreeDelete	Delete paths in the file system
...	

Data Center Impact

Isilon clusters place particular demands on power and cable management, behaving most reliably when nodes are continuously powered and sophisticated cable management strategies are used to support and route cabling.

Stuart Kendrick 2017-06-09