Data ONTAP GX

Presented at the USENIX FAST 2007 Conference

Mike Eisler, Peter Corbett, Mike Kazar, Dan Nydick, Chris Wagner

email2mre-@yahoo.com

(insert fast2007 between the – and @)

February 16, 2007

Introduction

- Data ONTAP GX is a scalable clustered network file server
- Services NFS and CIFS protocols
- Provides a scalable single system image to both administrators and clients

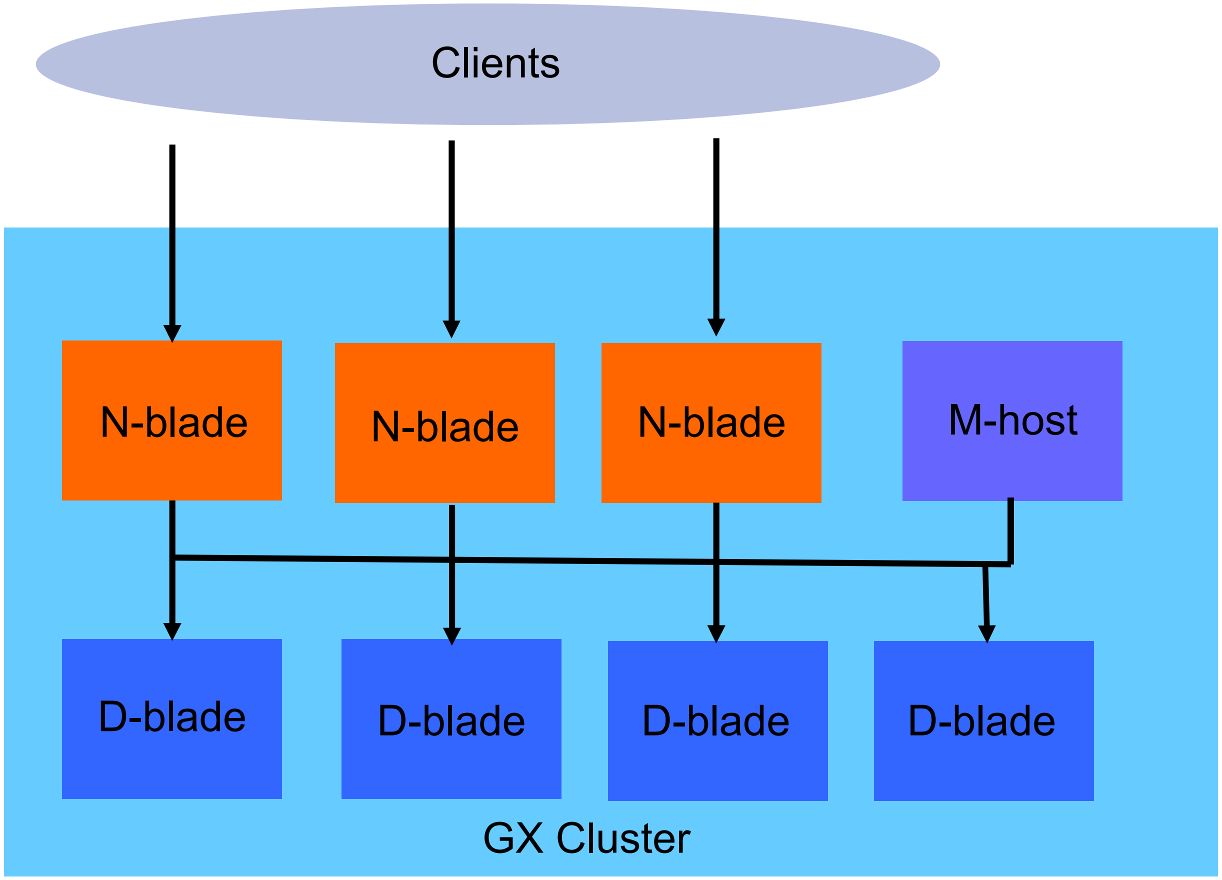
Related Work

- Cluster file servers based on global and distributed lock managers, distributed data and distributed metadata
  - GFS, GPFS, Frangipani (SAN fs), Slice (hashes file names)
- AFS and DFS provide a scalable global namespace
  - But require a non-standard client
- SpinFS
  - Basis for GX
  - Incorporates AFS concepts, but within a scalable cluster

Key Requirements

- Horizontal scalability
  - Can add server nodes to the cluster
  - Keep pace with expanding client compute clusters
  - No need for exotic server hardware
    - Node performance and reliability is important
- Location transparency
  - Transparent data migration among nodes in the cluster
  - Load sharing mirrors of volumes within the cluster
- Global namespace
  - Ability to link volumes from multiple nodes into a hierarchical namespace
- Virtual servers
  - Overlay of multiple virtualized servers and their independent namespaces onto the shared cluster hardware
- Robustness and load balancing
- Support widely used client protocols
  - NFS, CIFS

GX Cluster Block Diagram



Basic Component Modules

- Request processing divided between network facing ***N-blades*** and disk facing ***D-blades***
- **N-blades and D-blades are both software modules, and may run on the same hardware nodes**
- ***M-hosts*** **provide management for the cluster**

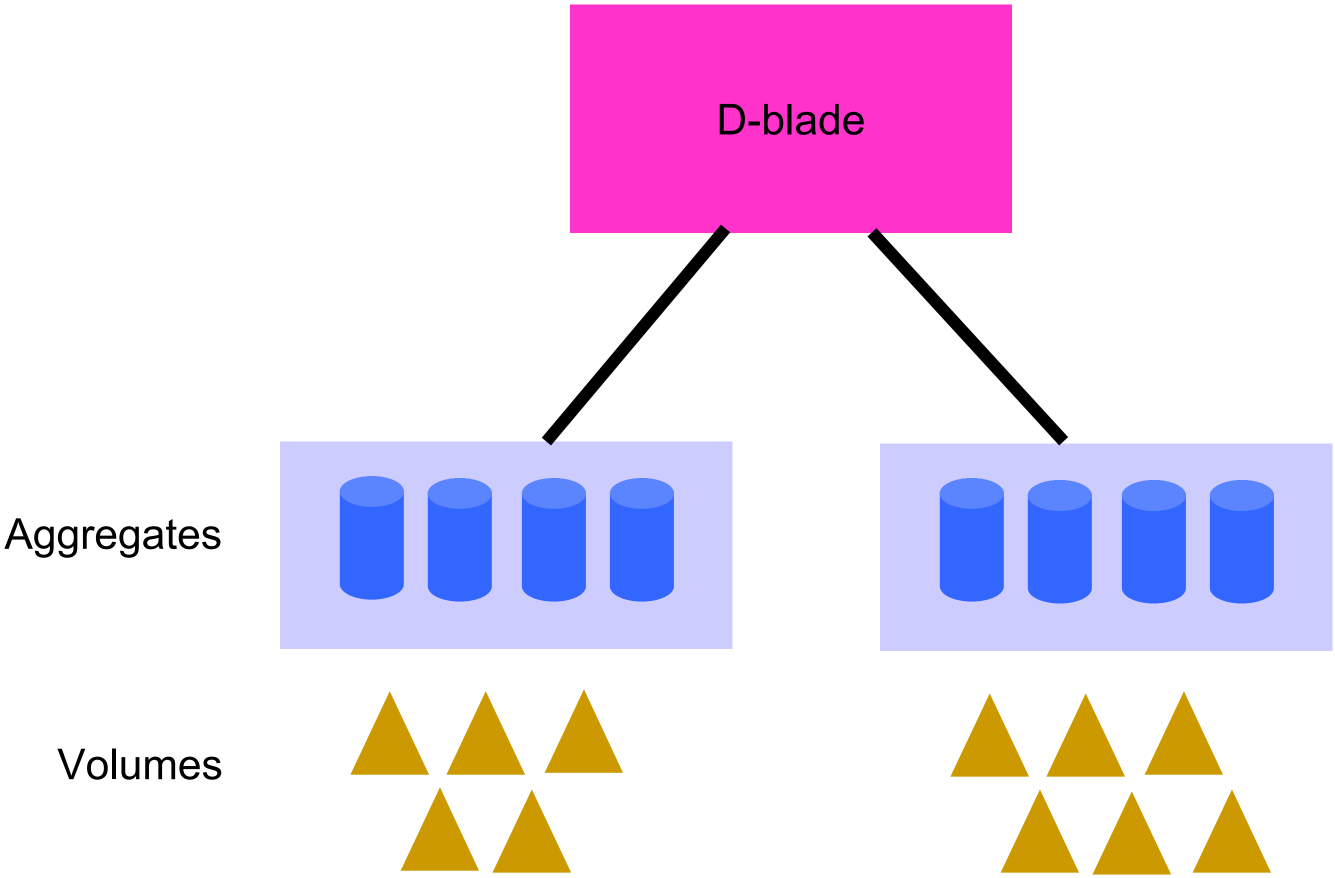
N-blade Roles

- **N-blades:**
  - Terminate client transport connections & sessions
  - Authenticate users / authorize clients (e.g. NFS exports)
  - Process NFS and CIFS protocols
  - Translate to a common internal protocol called *SpinNP*
  - Lookup where to route requests
  - Forward requests to the correct D-blade
  - Route response and callbacks to the correct client
- **N-blades are very nearly stateless and cacheless**
  - Eases moving Vservers among N-blades

D-blade Roles

- D-blades each store data volumes
  - D-blades store all persistent file system state
  - Manage lock state
  - Enforce ACLs
  - Maintain local single-instance data and metadata caches
  - Manage local filesystem instances as shared-nothing data stores
- Multiple volumes are stored in an aggregate
  - Aggregate is a collection of one or more RAID groups
  - Volumes are virtualized within an aggregate
- Each D-blade can control multiple aggregates at any time
- D-blade also handles RAID and storage stacks

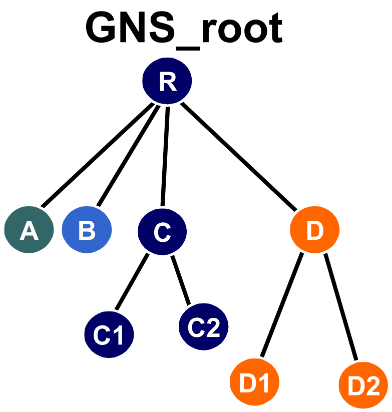
D-blade and Aggregates



Volumes and Junctions

- Each volume is a virtualized container storing a portion of file system namespace that descends from a single root directory
- Volumes are linked together through junctions
- Junctions
  - May appear anywhere in a volume
  - Link to the root of another volume
  - May point to a volume on a different D-blade in the cluster
  - Look like directories to the client
    - Client does not see a referral across a junction

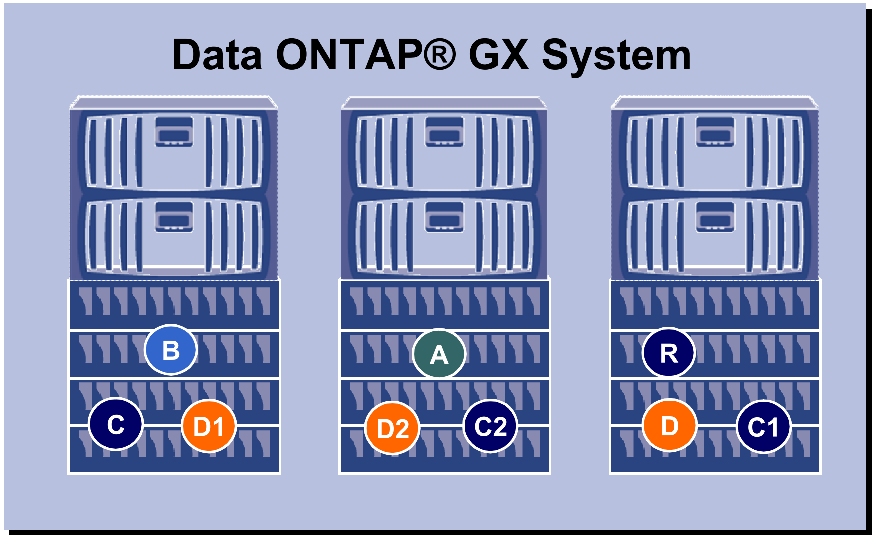
Namespace Across a Cluster



NetApp packages GX in High Availability Pairs

Partners within a pair use Infiniband interconnect

Pairs connect to each other over Ethernet



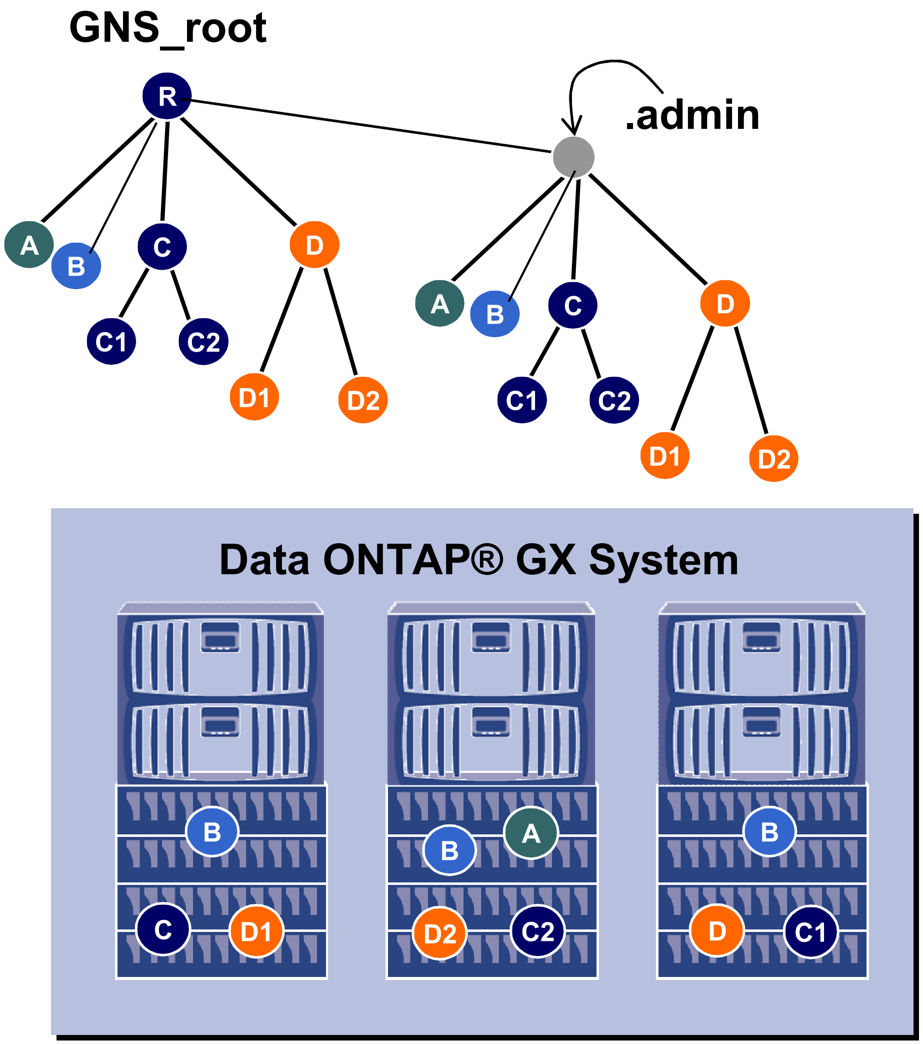
VServers and Namespaces

- A GX cluster exports one or more VServers
  - Often many more
- Each VServer presents its own independent namespace
  - Rooted at a separate root volume
- Each Vserver has its own virtual interfaces (*vifs*)
  - A vif is a network endpoint (IP address)
- Vifs can migrate among N-blades

MSIDs and DSIDs

- MSIDs (Master Data Set Identifiers) identify a group of mirrored volumes
  - MSIDs are present in file handles handed to clients
  - Uniquely specify a version (current or snapshot) of a set of mirrored volumes
- DSIDs (Data Set Identifiers) identify a single volume
  - DSIDs are present in internal file handles presented by N-blades to D-blades
  - Uniquely specify a version (current or snapshot) of a single volume

Load Sharing Mirrors



- Volume B has read-only load sharing mirrors
- Each mirror has the same MSID, but different DSID
- N-blade maps MSID to one of DSIDs and routes to D-blade
- Volume B has one read/write master
  - Same MSID, but different DSID
- Master accessible through /.admin

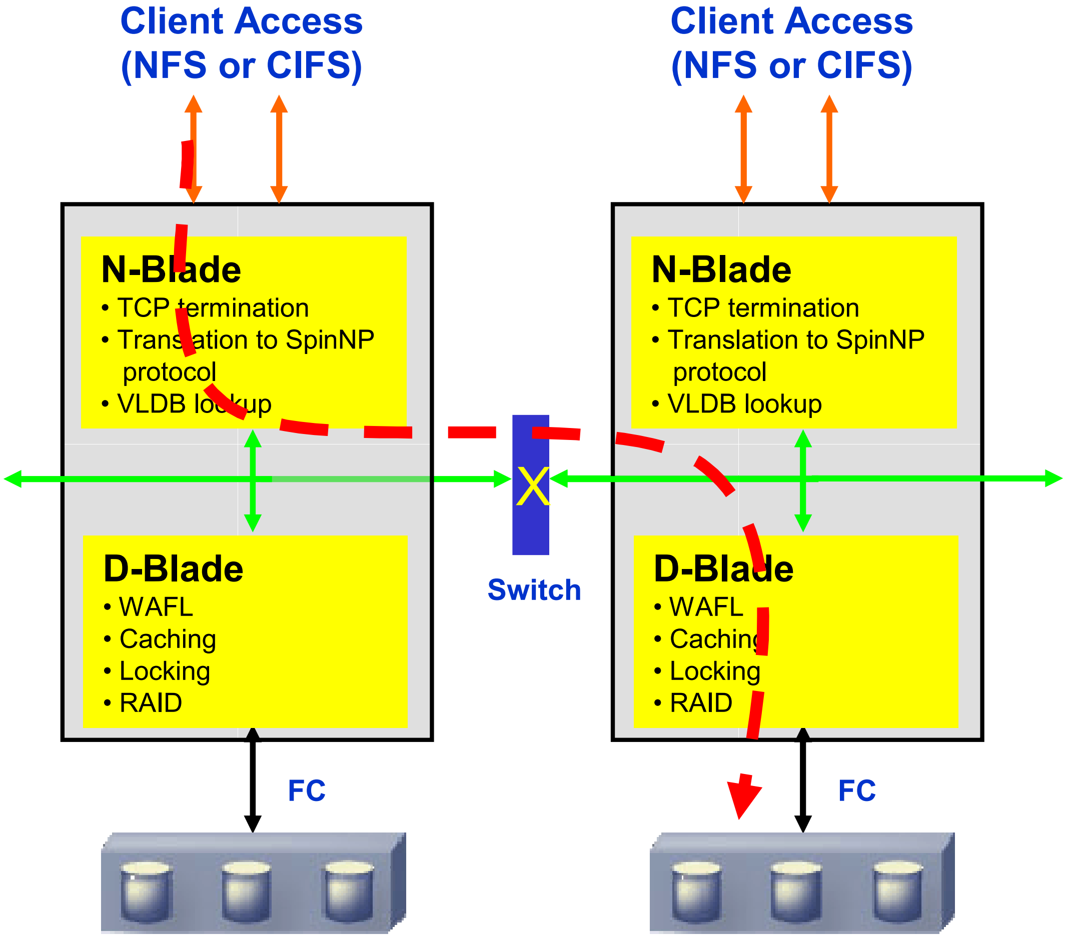
VLDB and VIF Manager Database

- The VLDB (Volume Location Database) records the mappings of:
  - MSIDs to one or more DSIDs
  - DSIDs to D-blade IDs
  - D-blade IDs to IP addresses (cluster network VIFs)
  - Junction mapping:
    - Parent MSID plus Junction ID to child MSID
  - Vserver roots to MSIDs
- VIF manager database records:
  - Current binding of VIFs to N and D blades
    - Client-facing VIFS can move between N-blades as part of Vserver migration or failover
  - Also records failover rules for VIFs

SpinNP

- SpinNP is a network protocol used inside the cluster
- SpinNP has multiple interfaces (application protocols):
  - File ops and file op callbacks
  - Session ops
  - Data protection ops
  - Striped volume ops
- Provides sessions, request-level flow control, security, session recovery
- Has a powerful versioning mechanism
- Used for all the high-bandwidth internal communication
- Suite of tools to compile SpinNP headers and marshal/unmarshal code directly from SpinNP specs

Request Processing



Client Access (NFS or CIFS)

Client Access (NFS or CIFS)

N-Blade
- TCP termination
- Translation to SpinNP protocol
- VLDB lookup

N-Blade
- TCP termination
- Translation to SpinNP protocol
- VLDB lookup

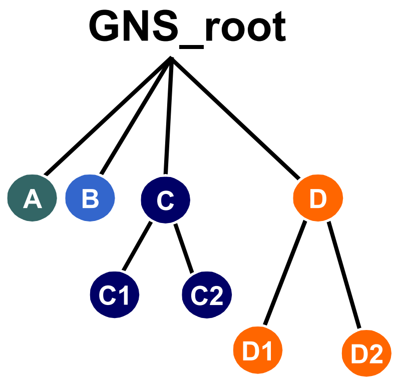
D-Blade
- WAFL
- Caching
- Locking
- RAID

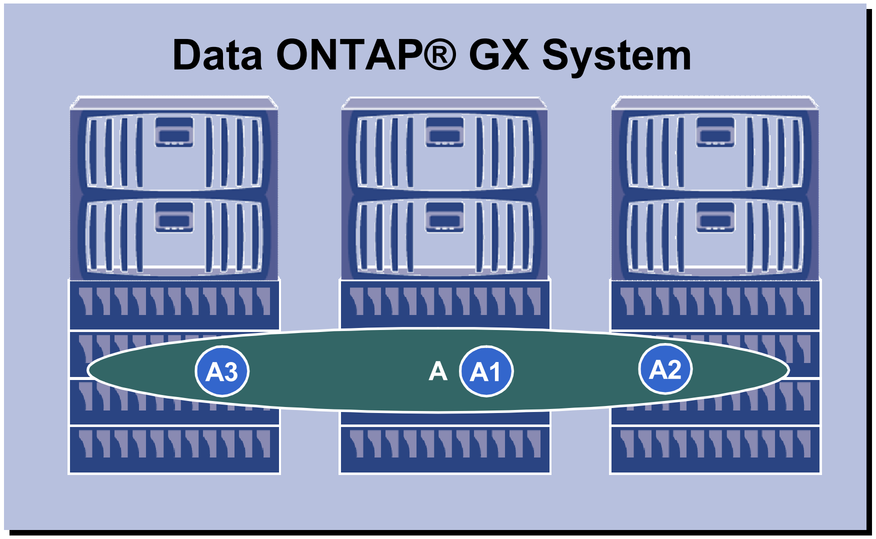
D-Blade
- WAFL
- Caching
- Locking
- RAID

Striped Volumes

- Multiple volumes can be joined together to form a striped volume
- Each component volume holds a disjoint portion of the entire volume
- Component volumes are distributed to different D-blades
- Files are distributed among the component volumes
- Each individual file may be striped across multiple component volumes, depending on its size

Striped Volume Across a Cluster





Single System Image

- Management databases are replicated coherently throughout the cluster
  - VLDB, VIF manager and others
- Contents of these databases are accessible via queries on each node
- Contents are cached at each node for faster lookup on the data path
- Maintain a quorum of nodes that are in the cluster
- Any node in quorum can write a database
- Administer entire cluster through a single management interface

Performance Results

- Wide range of configs tested
  - FAS3050, FAS6070 controllers
  - 2-24 nodes (more nodes are possible; show us the Purchase Order ☺)
  - NFS, CIFS protocols
  - Seq read, seq write, random read, random write
- Achieved over *One Million* operations/sec on SPEC SFS benchmark

Conclusions

- ONTAP GX is a real product running at a number of customer sites
- Achieved linearly scalable performance across clusters of up to 24 nodes
  - Linear scaling expected well beyond 24 nodes
- Provides powerful set of features that go well beyond what a standalone file server offers
  - A key component of our storage and data management virtualization