

Online SAN Re-Engineering and Data Migration Case Study

Danny B. Gross
Member of Technical Staff
Motorola, Inc
6501 William Cannon West
Austin, Texas 78735
danny.gross@motorola.com
(512) 895-4825

Atul Patel
Unix Systems Administrator
Motorola, Inc.
3501 Ed Bluestein Blvd.
Austin, Texas 78721
Atul.C.Patel@motorola.com
(512) 933-5472

Aaron Dahl
Unix Systems Administrator
Motorola, Inc.
3501 Ed Bluestein Blvd.
Austin, Texas 78721
Aaron.Dahl@motorola.com
(512) 933-6310

06/24/02

Abstract

A case study is presented detailing a large-scale data migration project in a highly available, heterogeneous production-manufacturing environment. This paper provides the operational goals and constraints, a report on execution and completion, and solutions to six engineering challenges encountered through the operation:

- a. How to maintain ServiceGuard viability during massive infrastructure reorganization.
- b. How to convert a QuickLoop FCAL environment to Fabric mode with no outage.
- c. How to resolve logical volume split situations that prevent single array operations.
- d. How to manage and plan for a rapidly changing infrastructure.
- e. How to recover the Logical Volume Manager (HP) and Veritas Volume Manager (Solaris) from a massive storage reorganization without rebooting.
- f. How to rapidly re-import modified volume groups within the infrastructure.

Description of the Project

A significant project was conducted in a four-site metropolitan region for the purpose of consolidating data centers and reducing infrastructure costs. The overall project objectives were the following:

- Significant cost avoidance in storage purchase/lease over three years.
- Applications survivability or rapid recovery in major infrastructure disasters.
- Optimization of resources for Database, Storage, and Applications performance
- Net reduction in capital equipment/service costs.

Through this project, development and non-production sites were consolidated into the frameworks of two production-manufacturing sites. These two sites were evaluated for availability and performance, and for storage upgrade requirements to support consolidation and cost savings. This evaluation found that there was significant storage available within the production framework. However, this available storage was either isolated from most of the framework, or located in areas that would impact performance if allocated. An aggressive, innovative plan was needed to re-engineer the framework for storage re-deployment with the following goals:

- Make available 3960 GB in optimized usable storage to the framework.
- Make it possible for all allocated disks to be mirrored between two data centers.
- Make it possible for all applications to survive or recover from the loss of an entire data center within minutes.
- Protect all storage to not less than RAID 5 within a single array when the loss of the other data center or mirror occurs.
- Improve or at least maintain current applications performance when framework is fully allocated.
- Make it possible to consolidate all storage (EMC and HP-XP256) onto two Storage Area Networks (SAN).
- Conduct all disk and SAN operations online with no outage (other than momentary and recoverable) to the application.

Three constraints were to be in place through the conduct of the project:

- The current Network Attached Storage (NAS) (Auspex, Network Appliance) position was to be maintained. This simplified the operation and reduced complexity.
- EMC and HP consolidation would occur separately. This allowed us to focus on one technical knowledge base at a time across multiple operating systems.
- Array groups and disks could be re-deployed within the region from re-engineered disk arrays.
- There was no capital to be provided to support this project, and expense costs were to be very limited.

In creating the final plan, it was found that one site could be easily upgraded, while the other required a significant engineering effort and posed the most risk to production. The remainder of this paper describes the plan and techniques used to successfully accomplish the latter on the HP XP256 framework at the Ed Bluestein facility in Austin, Texas.

The Project Environment.

The following depicts the framework at the Ed Bluestein facility, as it existed prior to the beginning of our project:

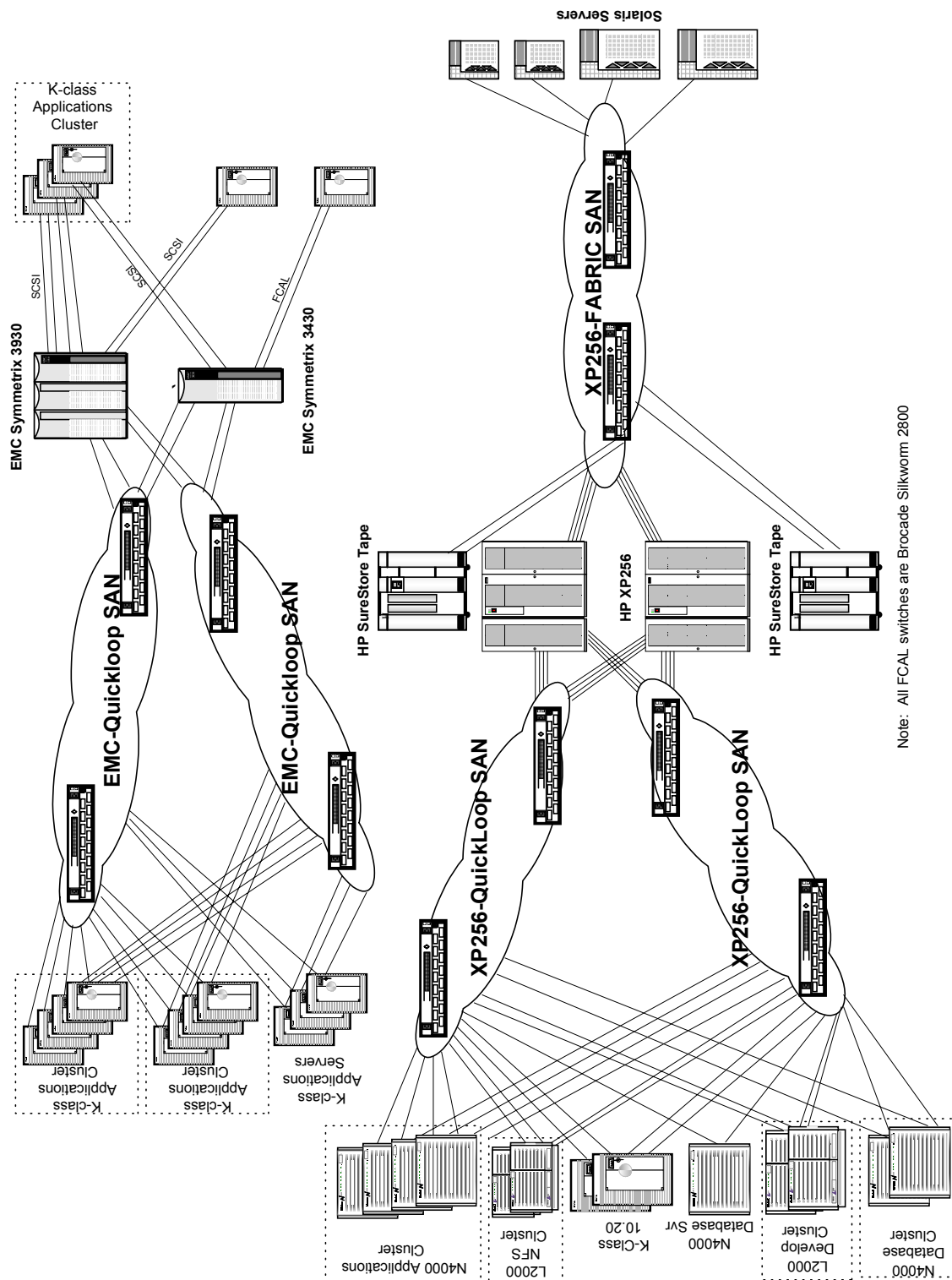


Figure 1. Pre-Project Storage Infrastructure at Ed Bluestein Facility

Within this framework were five SANs –

- 2 EMC-based QuickLoop (2 Brocade Silkorm 2800 Switch each in Data center 2). Connected to this Loop were 2 ports from an EMC3930 and 2 ports from an EMC3430, both in Data center 2.
- 2 HPXP256-QuickLoop (Each Loop had 1 Brocade Silkorm 2800 Switch in each Data center with two Inter-Switch Links (ISL) to the opposite Data center). Connected to these loops were five ports each from two XP256. (one in each data center).
- 1 HPXP256-FABRIC (1 Brocade Silkorm 2800 Switch in each Data center, with two ISL to the opposite Data center). Connected to the Fabric was one port each from two XP256.

There were six ServiceGuard Clusters connected to the SAN environment. Each of these clusters use dual-clusterlock disks:

- 1 2-node N4000 Database Cluster (Each system connected to both HPXP256 QuickLoops)
- 1 2-node L2000 NFS Cluster (Each system connected to both HPXP256 QuickLoops)
- 1 4-node N4000 Applications Cluster (Each system connected to both HPXP256 QuickLoops)
- 2 4-node HPUX 10.20 K-class Application Clusters (Each system connected to both EMC-QuickLoops)
- 1 2-node L2000 Applications (Development) Cluster (Each system connected to both HPXP256 QuickLoops)

There were seven unclustered systems active on the SAN environment:

- Four Enterprise-class Solaris Applications servers (Each system single-connected to XP256-FABRIC)
- Two HPUX 10.20 K-class applications servers (Each system single connected to XP256 QuickLoop)
- One N4000 Database server (Connected to both XP256 QuickLoops)

The following depicts the internal HP XP256 allocation chart at the Ed Bluestein facility at the beginning of the project:

CL1-J / CL2-J	CL1-K / CL2-K	CL1-J / CL2-J	CL1-K / CL2-K	CL1-F / CL2-F	CL1-E / CL2-E	CL1-B / CL2-B	CL1-A / CL2-A		CL1-A / CL2-A	CL1-B / CL2-B	CL1-E / CL2-E	CL1-F / CL2-F	CL1-A / CL2-A	CL1-B / CL2-B	CL1-E / CL2-E	CL1-F / CL2-F
4-8	4-7	4-6	4-5	4-4	4-3	4-2	4-1		2-1	2-2	2-3	2-4	2-5	2-6	2-7	2-8
3:85	3:72	3:5F	3:4C	3:39	3:26	3:13	3:00	ACP 3 1	01:00	01:0F	01:1E	01:2D	01:3C	01:4B	01:5A	01:69
3:86	3:73	3:60	3:4D	3:3A	3:27	3:14	3:01		01:01	01:10	01:1F	01:2E	01:3D	01:4C	01:5B	01:6A
3:87	3:74	3:61	3:4E	3:3B	3:28	3:15	3:02		01:02	01:11	01:20	01:2F	01:3E	01:4D	01:5C	01:6B
3:88	3:75	3:62	3:4F	3:3C	3:29	3:16	3:03		01:03	01:12	01:21	01:30	01:3F	01:4E	01:5D	01:6C
3:89	3:76	3:63	3:50	3:3D	3:2A	3:17	3:04		01:04	01:13	01:22	01:31	01:40	01:4F	01:5E	01:6D
3:8A	3:77	3:64	3:51	3:3E	3:2B	3:18	3:05		01:05	01:14	01:23	01:32	01:41	01:50	01:5F	01:6E
3:8B	3:78	3:65	3:52	3:3F	3:2C	3:19	3:06		01:06	01:15	01:24	01:33	01:42	01:51	01:60	01:6F
3:8C	3:79	3:66	3:53	3:40	3:2D	3:1A	3:07		01:07	01:16	01:25	01:34	01:43	01:52	01:61	01:70
3:8D	3:7A	3:67	3:54	3:41	3:2E	3:1B	3:08		01:08	01:17	01:26	01:35	01:44	01:53	01:62	01:71
3:8E	3:7B	3:68	3:55	3:42	3:2F	3:1C	3:09		01:09	01:18	01:27	01:36	01:45	01:54	01:63	01:72
3:8F	3:7C	3:69	3:56	3:43	3:30	3:1D	3:0A		01:0A	01:19	01:28	01:37	01:46	01:55	01:64	01:73
3:90	3:7D	3:70	3:57	3:44	3:31	3:1E	3:0B		01:0B	01:1A	01:29	01:38	01:47	01:56	01:65	01:74
3:91	3:7E	3:6B	3:58	3:45	3:32	3:1F	3:0C		01:0C	01:1B	01:2A	01:39	01:48	01:57	01:66	01:75
3:92	3:7F	3:6C	3:59	3:46	3:33	3:20	3:0D		01:0D	01:1C	01:2B	01:3A	01:49	01:58	01:67	01:76
3:93	3:80	3:6D	3:5A	3:47	3:34	3:21	3:0E		01:0E	01:1D	01:2C	01:3B	01:4A	01:59	01:68	01:77
3:94	3:81	3:6E	3:5B	3:48	3:35	3:22	3:0F									
3:95	3:82	3:6F	3:5C	3:49	3:36	3:23	3:10									
3:96	3:83	3:70	3:5D	3:4A	3:37	3:24	3:11									
3:97	3:84	3:71	3:5E	3:4B	3:38	3:25	3:12									
	CL1-K / CL2-K	CL1-J / CL2-J	CL1-K / CL2-K	CL1-F / CL2-F	CL1-E / CL2-E	CL1-B / CL2-B	CL1-A / CL2-A		CL1-A / CL2-A	CL1-B / CL2-B	CL1-E / CL2-E	CL1-F / CL2-F	CL1-A / CL2-A	CL1-B / CL2-B	CL1-E / CL2-E	
	3-7	3-6	3-5	3-4	3-3	3-2	3-1		1-1	1-2	1-3	1-4	1-5	1-6	1-7	
ACP 2 0	2:72	2:5F	2:4C	2:39	2:26	2:13	2:00	ACP 2 0	00:00	00:0F	00:1E	00:2D	00:3C	00:4B	00:5A	S P A R E
	2:73	2:60	2:4D	2:3A	2:27	2:14	2:01		00:01	00:10	00:1F	00:2E	00:3D	00:4C	00:5B	
	2:74	2:61	2:4E	2:3B	2:28	2:15	2:02		00:02	00:11	00:20	00:2F	00:3E	00:4D	00:5C	
	2:75	2:62	2:4F	2:3C	2:29	2:16	2:03		00:03	00:12	00:21	00:30	00:3F	00:4E	00:5D	
	2:76	2:63	2:50	2:3D	2:2A	2:17	2:04		00:04	00:13	00:22	00:31	00:40	00:4F	00:5E	
	2:77	2:64	2:51	2:3E	2:2B	2:18	2:05		00:05	00:14	00:23	00:32	00:41	00:50	00:5F	
	2:78	2:65	2:52	2:3F	2:2C	2:19	2:06		00:06	00:15	00:24	00:33	00:42	00:51	00:60	
	2:79	2:66	2:53	2:40	2:2D	2:1A	2:07		00:07	00:16	00:25	00:34	00:43	00:52	00:61	
	2:7A	2:67	2:54	2:41	2:2E	2:1B	2:08		00:08	00:17	00:26	00:35	00:44	00:53	00:62	
	2:7B	2:68	2:55	2:42	2:2F	2:1C	2:09		00:09	00:18	00:27	00:36	00:45	00:54	00:63	
	2:7C	2:69	2:56	2:43	2:30	2:1D	2:0A		00:0A	00:19	00:28	00:37	00:46	00:55	00:64	
	2:7D	2:6A	2:57	2:44	2:31	2:1E	2:0B		00:0B	00:1A	00:29	00:38	00:47	00:56	00:65	
	2:7E	2:6B	2:58	2:45	2:32	2:1F	2:0C		00:0C	00:1B	00:2A	00:39	00:48	00:57	00:66	
	2:7F	2:6C	2:59	2:46	2:33	2:20	2:0D		00:0D	00:1C	00:2B	00:3A	00:49	00:58	00:67	
	2:80	2:6D	2:5A	2:47	2:34	2:21	2:0E		00:0E	00:1D	00:2C	00:3B	00:4A	00:59	00:68	
	2:81	2:6E	2:5B	2:48	2:35	2:22	2:0F									
	2:82	2:6F	2:5C	2:49	2:36	2:23	2:10									
	2:83	2:70	2:5D	2:4A	2:37	2:24	2:11									
	2:84	2:71	2:5E*	2:4B	2:38	2:25	2:12									

Figure 2. HP XP256 Allocation at Ed Bluestein Facility, Pre-Project

Looking at the XP256 allocation, a good amount of available storage space can be seen on the left side (in ACP's 2 and 3), but most of it existed on one path (CL1K/2K) on four array groups. At the beginning of this project, CL1K/2K were FABRIC-mode while the remaining interfaces were QuickLoop Mode. Therefore, the bulk of the available disks were unavailable to the majority of the systems in the framework. Effectively, only the Solaris servers could access these disks initially. Even then, large allocations for a busy application would come from the same RAID-5 array group, hindering performance.

Another problem that existed was that the array performance was imbalanced between ACP pairs:

- All CL1J/2J and CL1K/2K –connected array groups were serviced from ACP’s 2 and 3. Any further allocations would have to be made from the same channels, rather than spread equally across channels.
- All of the 47GB array groups (15 total) belonged to ACP’s 2 and 3, while ACP’s 0 and 1 were fully 36GB groups (15 total).

Further evaluation showed that many volume groups were not allocated in such a way as to optimize performance. As an example, 100% of the Business Copy volumes existed on the same two RAID-5 array groups on one Channel (CL1J/2J)

It was agreed that the arrays required a complete re-engineering and re-allocation with the following in mind:

- 47GB and 36GB array groups would be spread evenly between ACP pairs.
- FCAL paths would be spread evenly between ACP pairs.
- All FCAL connections would be FABRIC mode.
- Allocations would be spread evenly between ACP pairs.
- Logical volumes would be striped within evenly spread physical volume groups.
- SecureManagerXP software would be disabled and replaced with sound operational procedures.

It was decided to re-architect the overall Storage Area Network environment:

- Collapse the five QuickLoop/FABRIC SANs to two FABRIC SANs.
 - Convert all SAN-connected systems to FABRIC mode.
 - Convert all array connections to FABRIC mode.
- Remove all direct-connected systems from the framework.
 - Evaluate applications running on direct-connected systems for upgrade to SAN-connected systems.
 - Isolate and consolidate database and storage services to SAN-connected clustered servers.
- Insure redundant FCAL connections were available from each host to both SAN’s.

The Target Architecture

The target FCAL layout was as follows:

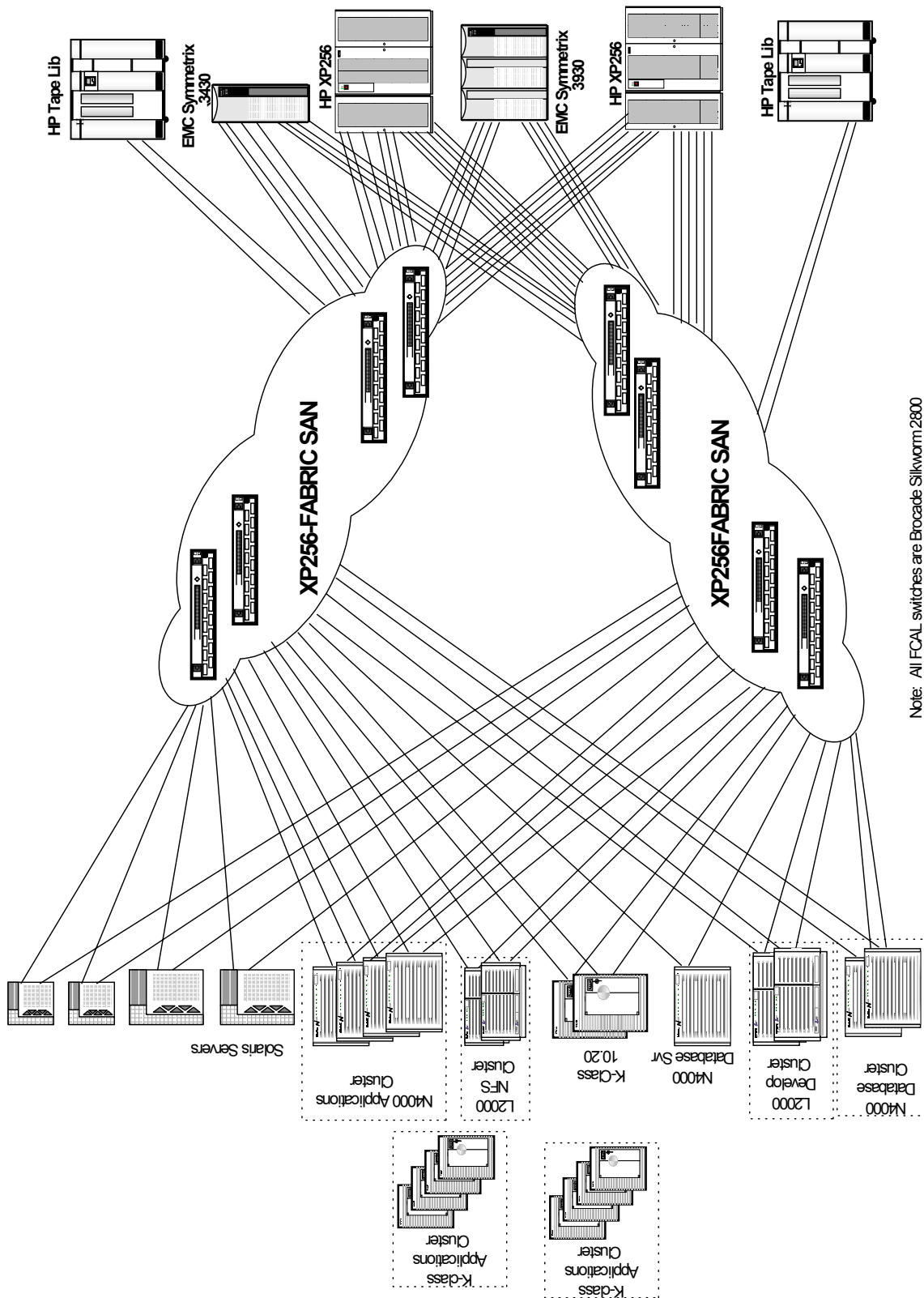


Figure 3. Target Storage Infrastructure at Ed Bluestein Facility

The target XP256 Array Allocation plan was as follows:

CL1-J / CL2-J	CL1-E / CL2-E	CL1-F / CL2-F	CL1-J / CL2-J	CL1-A / CL2-A	CL1-B / CL2-B	CL1-E / CL2-E	CL1-F / CL2-F			CL1-B / CL2-B	CL1-A / CL2-A	CL1-J / CL2-J	CL1-F / CL2-F	CL1-E / CL2-E	CL1-B / CL2-B	CL1-A / CL2-A	CL1-F / CL2-F
4-8	4-7	4-6	4-5	4-4	4-3	4-2	4-1			2-1	2-2	2-3	2-4	2-5	2-6	2-7	2-8
3:75	3:62	3:4F	3:3C	3:2D	3:1E	3:0F	3:00	A C P 3	A C P 1	1:00	1:0F	1:1E	1:2D	1:3C	1:4F	1:62	1:75
3:76	3:63	3:50	3:3D	3:2E	3:1F	3:10	3:01			1:01	1:10	1:1F	1:2E	1:3D	1:50	1:63	1:76
3:77	3:64	3:51	3:3E	3:2F	3:20	3:11	3:02			1:02	1:11	1:20	1:2F	1:3E	1:51	1:64	1:77
3:78	3:65	3:52	3:3F	3:30	3:21	3:12	3:03			1:03	1:12	1:21	1:30	1:3F	1:52	1:65	1:78
3:79	3:66	3:53	3:40	3:31	3:22	3:13	3:04			1:04	1:13	1:22	1:31	1:40	1:53	1:66	1:79
3:7A	3:67	3:54	3:41	3:32	3:23	3:14	3:05			1:05	1:14	1:23	1:32	1:41	1:54	1:67	1:7A
3:7B	3:68	3:55	3:42	3:33	3:24	3:15	3:06			1:06	1:15	1:24	1:33	1:42	1:55	1:68	1:7B
3:7C	3:69	3:56	3:43	3:34	3:25	3:16	3:07			1:07	1:16	1:25	1:34	1:43	1:56	1:69	1:7C
3:7D	3:6A	3:57	3:44	3:35	3:26	3:17	3:08			1:08	1:17	1:26	1:35	1:44	1:57	1:6A	1:7D
3:7E	3:6B	3:58	3:45	3:36	3:27	3:18	3:09			1:09	1:18	1:27	1:36	1:45	1:58	1:6B	1:7E
3:7F	3:6C	3:59	3:46	3:37	3:28	3:19	3:0A			1:0A	1:19	1:28	1:37	1:46	1:59	1:6C	1:7F
3:80	3:6D	3:5A	3:47	3:38	3:29	3:1A	3:0B			1:0B	1:1A	1:29	1:38	1:47	1:5A	1:6D	1:80
3:81	3:6E	3:5B	3:48	3:39	3:2A	3:1B	3:0C			1:0C	1:1B	1:2A	1:39	1:48	1:5B	1:6E	1:81
3:82	3:6F	3:5C	3:49	3:3A	3:2B	3:1C	2:0D			1:0D	1:1C	1:2B	1:3A	1:49	1:5C	1:6F	1:82
3:83	3:70	3:5D	3:4A	3:3B	3:2C	3:1D	3:0E			1:0E	1:1D	1:2C	1:3B	1:4A	1:5D	1:70	1:83
3:84	3:71	3:5E	3:4B											1:4B	1:5E	1:71	1:84
3:85	3:72	3:5F	3:4C											1:4C	1:5F	1:72	1:85
3:86	3:73	3:60	3:4D											1:4D	1:60	1:73	1:86
3:87	3:74	3:61	3:4E											1:4E	1:61	1:74	1:87

CL1-B / CL2-B	CL1-E / CL2-E	CL1-F / CL2-F	CL1-J / CL2-J	CL1-A / CL2-A	CL1-B / CL2-B	CL1-E / CL2-E			CL1-A / CL2-A	CL1-J / CL2-J	CL1-F / CL2-F	CL1-E / CL2-E	CL1-B / CL2-B	CL1-A / CL2-A	CL1-J / CL2-J
3-7	3-6	3-5	3-4	3-3	3-2	3-1			1-1	1-2	1-3	1-4	1-5	1-6	1-7
2:62	2:4F	2:3C	2:2D	2:1E	2:0F	2:00	A C P 2	A C P 0	0:00	0:0F	0:1E	0:2D	0:3C	0:4F	0:62
2:63	2:50	2:3D	2:2E	2:1F	2:10	2:01			0:01	0:10	0:1F	0:2E	0:3D	0:50	0:63
2:64	2:51	2:3E	2:2F	2:20	2:11	2:02			0:02	0:11	0:20	0:2F	0:3E	0:51	0:64
2:65	2:52	2:3F	2:30	2:21	2:12	2:03			0:03	0:12	0:21	0:30	0:3F	0:52	0:65
2:66	2:53	2:40	2:31	2:22	2:13	2:04			0:04	0:13	0:22	0:31	0:40	0:53	0:66
2:67	2:54	2:41	2:32	2:23	2:14	2:05			0:05	0:14	0:23	0:32	0:41	0:54	0:67
2:68	2:55	2:42	2:33	2:24	2:15	2:06			0:06	0:15	0:24	0:33	0:42	0:55	0:68
2:69	2:56	2:43	2:34	2:25	2:16	2:07			0:07	0:16	0:25	0:34	0:43	0:56	0:69
2:6A	2:57	2:44	2:35	2:26	2:17	2:08			0:08	0:17	0:26	0:35	0:44	0:57	0:6A
2:6B	2:58	2:45	2:36	2:27	2:18	2:09			0:09	0:18	0:27	0:36	0:45	0:58	0:6B
2:6C	2:59	2:46	2:37	2:28	2:19	2:0A			0:0A	0:19	0:28	0:37	0:46	0:59	0:6C
2:6D	2:5A	2:47	2:38	2:29	2:1A	2:0B			0:0B	0:1A	0:29	0:38	0:47	0:5A	0:6D
2:6E	2:5B	2:48	2:39	2:2A	2:1B	2:0C			0:0C	0:1B	0:2A	0:39	0:48	0:5B	0:6E
2:6F	2:5C	2:49	2:3A	2:2B	2:1C	2:0D			0:0D	0:1C	0:2B	0:3A	0:49	0:5C	0:6F
2:70	2:5D	2:4A	2:3B	2:2C	2:1D	2:0E			0:0E	0:1D	0:2C	0:3B	0:4A	0:5D	0:70
2:71	2:5E	2:4B											0:4B	0:5E	0:71
2:72	2:5F	2:4C											0:4C	0:5F	0:72
2:73	2:60	2:4D											0:4D	0:60	0:73
2:74	2:61	2:4E											0:4E	0:61	0:74

Figure 4. HP XP256 Allocation at Ed Bluestein Facility, Target

The vast majority of the volumes would be made available via five redundant FCAL ports (CL1A/2A, CL1B/2B, CL1E/2E, CL1F/2F, CL1J/2J) into the FABRIC. Solaris hosts were isolated to one redundant port (CL1K/2K), which was assigned the last LUN on each array group. In this way, we were able to segregate Hewlett Packard and Sun LUNs simply and rapidly without using the SecureManagerXP software.

The Implementation Plan and Execution Report

The implementation plan was broken into three parts:

Phase I

Complete consolidation efforts of non-production systems into two production sites. From this phase, 14 36GB and six 47 GB array groups were available from other arrays for re-deployment. This phase was accomplished in two months.

Phase II Planned Time: 12 days
Actual Time: 15 days

- a. The Phase II Plan.
 1. Split mirrors between Data centers 1 and 2 in our target-manufacturing environment so that primary side existed in Data center 1. **Planned time: 8 hours. Actual time: 18 hours.**
 2. Re-engineer/overhaul/rebuild array, re-architect the FCAL and QuickLoop framework in Data center 2. **Planned time: 2 days. Actual time: 4 days.**
 3. Re-establish connectivity between FABRIC SAN and all hosts. **Planned time: 2 days. Actual time: 3 days.**
 4. Re-mirror/reorganize the logical volumes. **Planned time: 10 days. Actual time: 6 days.**
- b. Phase II Execution Report.
 1. Riskiest part of entire project (mirror split and volume group reduction) completed on time, with no impact or interruption to production operations.
 2. 264 logical volumes on 624 disks were modified from 18 systems to run in Data center 1, freeing the Data center 2 array for re-engineering work.
 3. We identified three volume groups that would have prevented the successful failover of the systems in the event of the failure of data centers 1 or 2. These volume groups were allocated in a way that required the existence of both XP256 arrays, even when the mirrors were split. This situation was corrected.
 4. We identified more than 20 logical volumes that were allocated on the XP256 in such a way that could hamper the performance of the system. Each of these situations was resolved during the re-mirroring operation.

5. We had to relocate approximately 56 GB of data from live database systems to the Data center 1 XP256 to support running the factory from a single array. This operation was conducted with no database interruption or performance impact.
6. One non-production system panicked, and one production system was scheduled for a reboot as a result of the QuickLoop-FABRIC switch. This was due to an out-of-date patch on both systems. There was no impact to production operations as a result of either of these issues.

Phase III Planned Time: 12 days

Actual Time: 6 days

a. The Phase III Plan.

1. Split mirrors between Data centers 1 and 2 in our target-manufacturing environment so that primary side existed on the previously re-engineered XP256 in Data center 2, over the Fabric SAN. **Planned time: 8 hours. Actual time: 4 hours.**
2. Re-engineer/overhaul/rebuild XP256 array, re-architect the FCAL and QuickLoop framework in Data center 1. **Planned time: 2 days. Actual time: 2 days.**
3. Re-establish connectivity between FABRIC SAN and all hosts. **Planned time: 2 days. Actual time: 1 day.**
4. Re-mirror/reorganize the logical volumes. **Planned time: 10 days. Actual time: 3 days.**

b. Phase III Execution Report.

1. Re-engineering of XP256 and FCAL frameworks went much more smoothly in Data center 1.
2. 264 logical volumes were split in 4 hours, a tremendous improvement over the Phase I split.
3. No outage or impact to factory systems during operations. However, after the close of operation, an Inter-Switch Link (ISL) failed between two core Brocade 2800 Switches, causing degradation in performance to four Solaris systems. This situation was rapidly identified, then recovered by disabling the affected ISL link on the Fabric.
4. We completed the overall project two days ahead of schedule.

Engineering Challenges and Solutions

During the planning and conduct of this project, we found that there were six tremendous challenges to its successful and timely completion:

Challenge 1:

All of our SAN-connected ServiceGuard clusters utilize dual-clusterlock disks, one on each of the two XP256 arrays. During the split operation, loss of contact was to occur with one of these disks. During the reorganization of the array, probability was high that the neither clusterlock disk would belong to the volume groups specified in the cluster configuration. How were we to execute this operation with no outage, with the following issues adding to the problem?

- Modification of the clusterlock disk in the ServiceGuard Cluster Configuration requires that the cluster be halted.
- While we could reboot a cluster member if necessary, we could not incur any outage to any entire cluster through this operation.

Solution:

- a. We first identified a mechanism by which ServiceGuard reports the loss of a clusterlock disk, and then recovers. We found that the cmcld daemon checks hourly for the existence of the configured clusterlock disks. When it finds an inaccessible disk, it prints the message “WARNING: Cluster lock on disk /dev/dsk/cXtYdZ is missing!” to the syslog file every hour until the condition is recovered. If the disk becomes accessible, cmcld immediately reports, “Cluster lock /dev/dsk/cXtYdZ is back on-line”
- b. We next identified a means by which to re-write the initialization data to a recovered clusterlock disk. There is a downloadable (and unsupported) program, “cminitlock” that was designed to recover this area on a failed clusterlock disk.
- c. In order for us to isolate the first XP256 array for re-engineering, it was necessary to remove the physical volume for the clusterlock on that array from its volume group. We discovered that ServiceGuard had no interest in the reduction of the physical volume from its volume group, other than to print the warnings identified previously.
- d. When we brought the re-engineered array back on line, and established the host connection into the Fabric, it was necessary to change the device that held the clusterlock. The solution was simple:

- 1) Identify the target (desired) location of the new clusterlock disk, and derive its device major and minor numbers for both the character and block devices.
- 2) Remove both character and block device files for the previously configured clusterlock disk using the command “rmsf /dev/dsk/cXtYdZ /dev/rdisk/cXtYdZ”.
- 3) Make block and character device files using the previously configured name with the target location major and minor numbers using the command “mknod”.

Example:

A Cluster is configured with clusterlock on the disk c8t7d5. The desired location of the disk is determined to be c27t0d4. The block device listings for both are as follows:

```
brw-r----- 1 bin    sys      31 0x087500 Jan 30 18:36 c8t7d5  ←Configured
brw-r----- 1 bin    sys      31 0x1b0400 Jan 30 18:36 c27t0d4  ←Desired
```

Following the instructions above, we change the configured to the desired:

```
rmsf /dev/dsk/c8t7d5 /dev/rdisk/c8t7d5
mknod /dev/dsk/c8t7d5 b 31 0x1b0400
mknod /dev/rdisk/c8t7d5 c 31 0x1b0400
```

- e. Next, we pvcreate the disk, and extend back into the volume group to which it belongs using “vgextend –g <VG Name> <Configured Clusterlock disk>”. For our example, the disk is /dev/dsk/c8t7d5.
- f. The final step is to recover the clusterlock on our relocated disk. To do so, execute the command “cminitlock –v <VG Name> <PV Name>”. Once complete, the cmeld daemon will report that the clusterlock is recovered.
- g. At this point, the cluster is fully functional with the relocated clusterlock disk. There are a few follow-on recommendations regarding this operation:
 - 1) The recreated device files may be destroyed the first time the command “insf –e” is executed. This command will re-install the device files, potentially over-writing the manually created devices.
 - 2) We recommend that the cluster be modified to reflect the true clusterlock configuration on a node-by-node basis at a more convenient time.

Challenge 2:

All of our HPUX SAN-connected hosts had redundant FCAL cards in QuickLoop Mode. Each of these cards was capable of supporting FABRIC mode. Our Solaris hosts had redundant FCAL cards, only one of which was active on the FABRIC SAN. How were we to complete the conversion/consolidation of the SAN's to FABRIC mode with no outage, with the following issues adding further complexity to the issue:

- The QuickLoop-Mode Brocade Silksworm 2800 Switches required firmware upgrade to support FABRIC mode.
- While we could reboot a single host if necessary, many hosts were running critical applications and could not be rebooted.

Solution:

- a. The overall strategy of Phase II provided for a unique opportunity to conduct extensive operations on the SAN. By reducing mirrors to one data center, then removing the primary and pmlink for the mirror as well as the pmlink for the remaining primary, it was possible to completely rebuild the SAN infrastructure between the buildings. This provided the opportunity to upgrade to FABRIC mode. The following components were in place prior to the conduct of this operation:
 - 1) N- and L- class systems: A5158A Fibre Channel PCI Tachyon TL Adapter. Driver and PHKL_23939 installed.
 - 2) K-class systems: J6685A Tachlite adapters. Driver and PHKL_23939 installed.
 - 3) Sun Systems: JNI 64-bit Fibre Channel Adapters. JNI drivers and prerequisite patches installed.
- b. The Solaris systems were already configured and connected to the Fabric. The HP installed base supported both Fabric and QuickLoop Mode, but was connected to a QuickLoop switch framework. The Brocade Switches required a firmware upgrade to fully support the Fabric mode. We proceeded as follows:
 - 1) Re-engineered the XP256 in data center 1, and set the ports to Fabric Mode.
 - 2) Upgraded the firmware on the isolated Brocade switches, changed the topology to Fabric and installed zone configuration.

- 3) Re-designed the Fabric to support the consolidated switch framework. Isolated storage to a “core” switch, hosts to a cascaded switch.
- 4) Our biggest concern was whether the HP Tachyon adapters would automatically convert to Fabric Mode without system restart when the port was re-enabled on the new Fabric switch. We found that the adapters correctly negotiated the topology with the switch with no problems. Effectively, the hosts were converted from QuickLoop to Fabric mode merely by disabling and re-enabling a switch port in Fabric mode.
- 5) Once Phase II was complete, data center 2 was re-engineered and upgraded in exactly the same way.

Challenge 3:

When we initiated the logical volume split operation in Phase II, we discovered several logical volumes that, when split, occupied LUN's on both XP256 arrays. This occurred even though we had correctly specified the array in Data center 1 with the lvreduce command. How could we continue to isolate the XP256 arrays without impacting or halting the critical application that was served from these disks?

Solution:

The crux of the problem was that some of the logical extents of the volume were mirrored to disks on the same XP256 array. As a result, when split, the logical volume occupied extents on both XP256 arrays. We noted that remaining logical extents were allocated very unevenly between the physical volumes, and that these physical volumes were distributed and striped. The solution to this issue was to use the pvmove command to relocate the extents from Data center 2 to Data center 1. This command works by establishing a mirror on the target disk, synchronizing the extent, then splitting the extent from the source. As a result, even live filesystems can be relocated with no outage. We realized that the resulting allocation may not be optimal, but, when the re-engineered array was brought back on line, the re-mirroring operation would correct this problem.

Challenge 4:

Any complex infrastructure can be difficult to manage with respect to changes. This project required 22 days of work that imposed risk on the production systems by splitting mirrors, online re-engineering, etc. Through this period, multiple projects were placed on hold, but several required immediate attention. How could we manage the infrastructure, accomplish the high-priority changes and even plan for the unknown through the riskiest parts of the project with the following issues:

- Several new systems required deployment to the SAN infrastructure.
- Several applications urgently needed additional disk space.

Solution:

In order to accomplish this project successfully, we established several management-supported rules for any change to be made during the conduct of the operation:

- No change would be appropriate to any volume group that held a clusterlock disk for a ServiceGuard cluster until the completion of the project.
- No change would be appropriate that would risk the loss of a quorum in a ServiceGuard cluster. This included any network modifications on existing cluster members, kernel changes, etc.
- No change would be appropriate which would risk our ability to manipulate the FCAL infrastructure as necessary within the plan.
- New allocations of LUN's on the SAN would not be appropriate until after the Phase II array re-engineering was complete. At this point, allocations would be possible, but not mirrored to the alternate data center until later in the project.

Challenge 5:

HPUX systems utilize Logical Volume Manager (LVM) and MirrorDisk/UX to mirror logical volumes between XP256 arrays, and utilize physical volume links for FCAL path redundancy. Solaris systems utilize Veritas Volume Manager to mirror disks in the same way and maintain redundant paths via Dynamic Multi-Path (DMP). The reorganization of the XP256 and restructuring of the SAN environment would impose dramatic changes on these systems. How could we without rebooting or impacting HPUX or Solaris systems accomplish all of these changes?

Solution:

HPUX Logical Volume Manager:

- 1) After removing the mirror from the appropriate target XP256 array, remove the pvlink and primary disk for the (former) mirror from the volume group using the vgreduce command. (Note: remove the pvlink first to avoid link switches). This will isolate the data to one XP256, freeing the other for re-engineering efforts.

- 2) Remove the pmlink for the remaining volumes in the volume group, insuring that the I/O remains on the correct switch. Once complete, the system will be utilizing a single FCAL interface for all disk traffic. At that point, the secondary switch framework can be completely dismantled, upgraded, and reset as necessary with no impact to the production system.
- 3) Remove the device name for the former mirror, its pmlink, and the reduced primary pmlink using the “rmsf” command. This act is purely for good housekeeping. Once the new Fabric is enabled, the “old” disks formerly used for mirroring will not be needed.
- 4) After re-organization complete, rescan the infrastructure using the “ioscan -fnC disk” command.
- 5) Re-install device files using the “insf -e” command, and verify correct device count. Run “xpinfo” and verify the appropriate layout is seen from the host.
- 6) Extend new primary and pmlink into volume group using the “vgextend” command, and re-mirror using “lvextend”.

Solaris (Veritas Volume Manager):

Solaris/Veritas Volume Manager proved to be trickier than HP-UX to recover from the re-organization. If we were to follow the same logic as with HP-UX, we would remove the DMP volumes and mirrors, and disable the paths. This can be done on line. However, in order to re-enable a path, the system must be rebooted. Therefore, a more extensive, somewhat riskier method is used for Solaris:

- 1) Remove the mirrors from the target XP256 array using the command “vxassist remove mirror <vol> <diskname1> <diskname2> ...”
- 3) Reduce the disk from the disk group using the command “vxdg -g <disk group> rmdisk <disk>”. This effectively removes both the primary disk and its Veritas equivalent of a pmlink.
- 3) Remove the host from the Fabric zone, or disable the ISL port to the target XP256 on the Brocade SilkWorm Switch with “port disable <port number>”. (This isolated all I/O traffic to the single array, and allowed the re-structuring of the Fabric and array in data center 2). At this point, the system may begin to report in /var/adm/messages and dmesg that disks are powered down or

taken offline. The command “vxdisk list” may show disks listed but “unavailable” or “formatting”.

- 4) After array re-organization is complete, re-enable the zoned-out or disabled ISL port on the Brocade switch with “port enable <port number>”. Rescan the infrastructure from the Solaris host using the “drvconfig” command, and re-install device files using “disks”. Verify that appropriate disks are seen using the “format” command.
- 5) The first Solaris system on the SAN to scan the newly available XP256 disks will complain about “unlabeled” disks. Use the format command to apply a label to each new disk. (Ex: “format -f labelfile <disk>” where labelfile is a text file containing the format command “l”). NOTE: This operation is only necessary once. When complete, other Solaris systems will not complain about unlabeled disks.
- 6) Run the command “vxdctl enable”. This command causes the Volume Manager to re-scan its infrastructure, and rebuild the DMP volume matches. Active disk groups and mounted filesystems are unaffected by this command. NOTE: This step is required on each Solaris system attached to the SAN.
- 7) Initialize each new disk using the “vxdisksetup -i <disk>” or “vxdiskadm” command. NOTE: This is only necessary once, and will be seen by all other Solaris systems. “vxdisk list” will show the status of each disk initialized as “online”.
- 8) Extend appropriate disks into the disk groups using “vxdg -g <disk group> adddisk <disk name>”
- 9) Re-mirror the logical volumes using “vxassist -g <disk group> mirror <vol> <disk name1> <disk name 2>...”

Challenge 6:

There were initially over 90 volume groups with more than 624 physical volumes modified throughout the infrastructure. How could we rapidly manage and transfer these changes with zero error to all the hosts in the infrastructure?

Solution:

1. SecureManagerXP was turned off in the XP256 arrays to simplify and speed disk re-organization.

2. All disk information is managed via a single file that holds the following data:

- a) Volume Group assigned to LUN.
- b) Volume Group minor number.
- c) Physical Volume Group assigned to LUN.
- d) LUN.
- e) Information field (currently active on host, etc).

EXAMPLE:

```
vg_admin,0x160000,vg_admingrp5,0:3f,system1
vg_admin,0x160000,vg_admingrp5,1:3f,system1
vg_admin,0x160000,vg_admingrp6,2:3f,system1
vg_admin,0x160000,vg_admingrp6,3:3f,system1
vgapp1,0x170000,vgapp1grp1,0:43,system2
vgapp1,0x170000,vgapp1grp1,1:43,system2
```

3. The following policies were put into place:

- a) SAN-connected HP-UX hosts have maxvgs set to 96. This makes it possible for any host in the infrastructure to activate any volume group without minor number conflicts.
- b) Any change to any LUN would be managed via the central file. This insures that no LUN is assigned to more than one Volume Group.
- c) Every LUN assigned will have a mirror in the identical position in the opposite building. Therefore, only one entry per LUN is needed.
- d) Every Volume Group (with the exception of vg00) would be managed via the central file, and will be assigned a unique minor number.
- e) The centralized file is shared or copied on every SAN-connected host to avoid conflicts.

4. With the centralized map file and the output from the command “xpinfo” or “inquiry256”, a shell script was written that could parse the appropriate “vgimport” command for the desired volume group.

Summary.

Utilizing good systems management techniques, it is possible to consolidate diverse Storage infrastructures into simplified, redundant Fabrics without incurring outage to the client systems. Such efforts can save real dollars by fully utilizing and optimizing the available storage systems. By taking advantage of the Operating Systems and available Storage Management tools available, engineering challenges posed by complex infrastructures can be overcome with great success.