

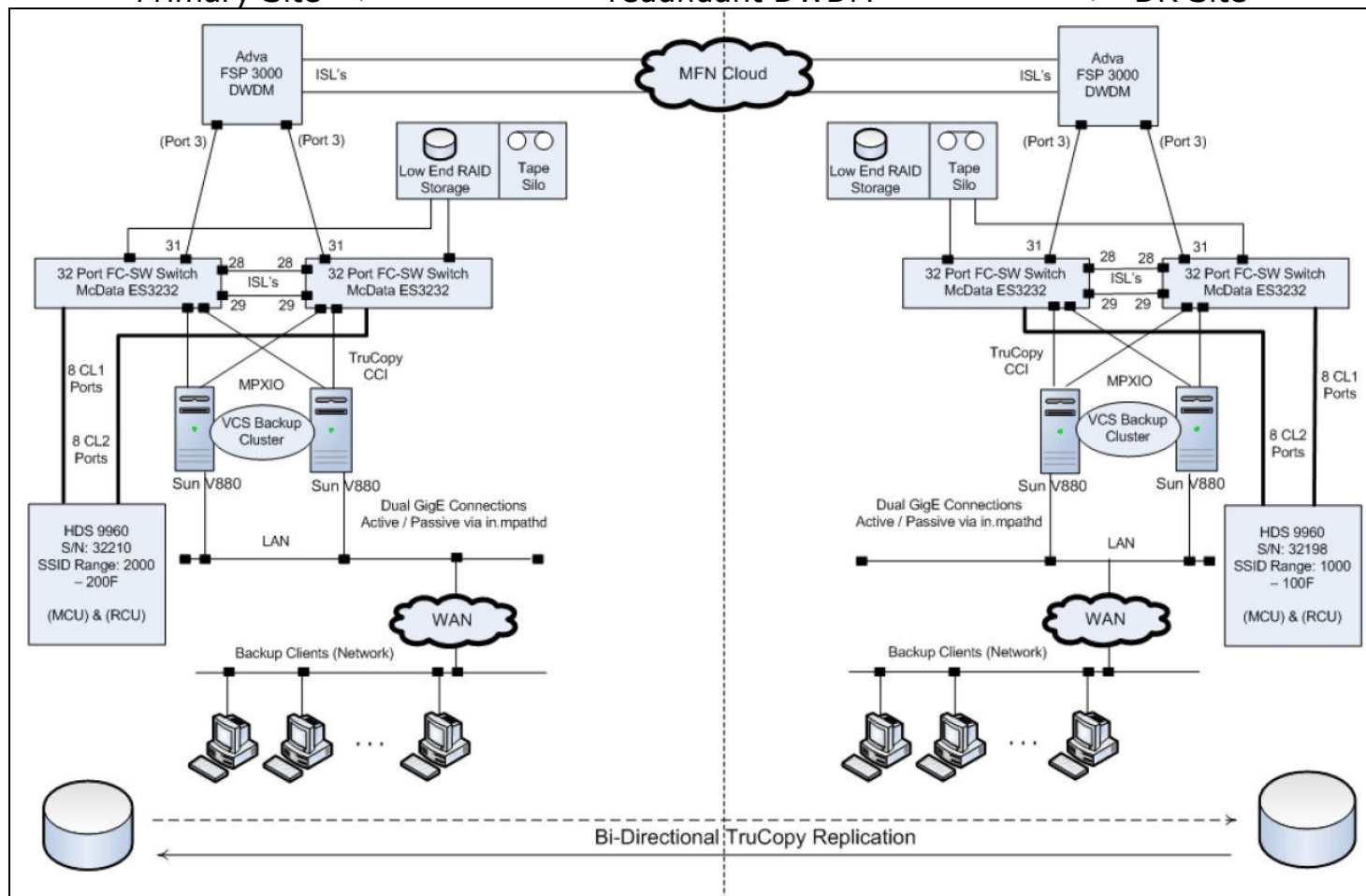
**The New York Merchantile Exchange's (NYMEX)
first Storage Area Network (both Local & Multi-Site Distance SAN)**

Architect.....: Noel Milton Vega
Implementation...: Noel Milton Vega
Documentation....: Noel Milton Vega

05/2004

The New York Mercantile Exchange's (NYMEX) inaugural Local & Wide Area/Distance SAN

Primary Site <----- redundant DWDM -----> DR Site



Milton Vega (Rensselaer Technology Group, LTD. V 0.1 (04/16/2004)

Figure1: Solution: Designed for local clusters, and simultaneous remote Sync/Async data replication for Disaster Recovery/Disaster Tolerance. The two horizontal lines at the top of the diagram represent diverse 2Gbps Dense Wave Division Multiplexing (DWDM) optic paths that carried the TrueCopy traffic. From start to end, each optical path took divergent geographical routes by design. Because the paths were asymmetric, their fibre optic distances differed, and this had several implications that needed to be accounted for: (1) higher end-to-end latency over the longer path, which can be noticeable for synchronous replication (but asynchronous as well); (2) The FibreChannel switches purchased (4 of them) would need to have enough Buffer Credits for the longest of the distances involved (i.e. for the worst case, since enough buffer credits ensures continuous streaming of data); and (3) asymmetric path (request over one path; response over the opposite path).

While designing the distance portion of this SAN, I wrote a white paper to myself that talked about the effects that physical and logical distances, bit-insertion rates, etc., imposed on switch Buffer credit requirements and performance. That paper can be viewed here: <http://doc5.computingarchitects.com>

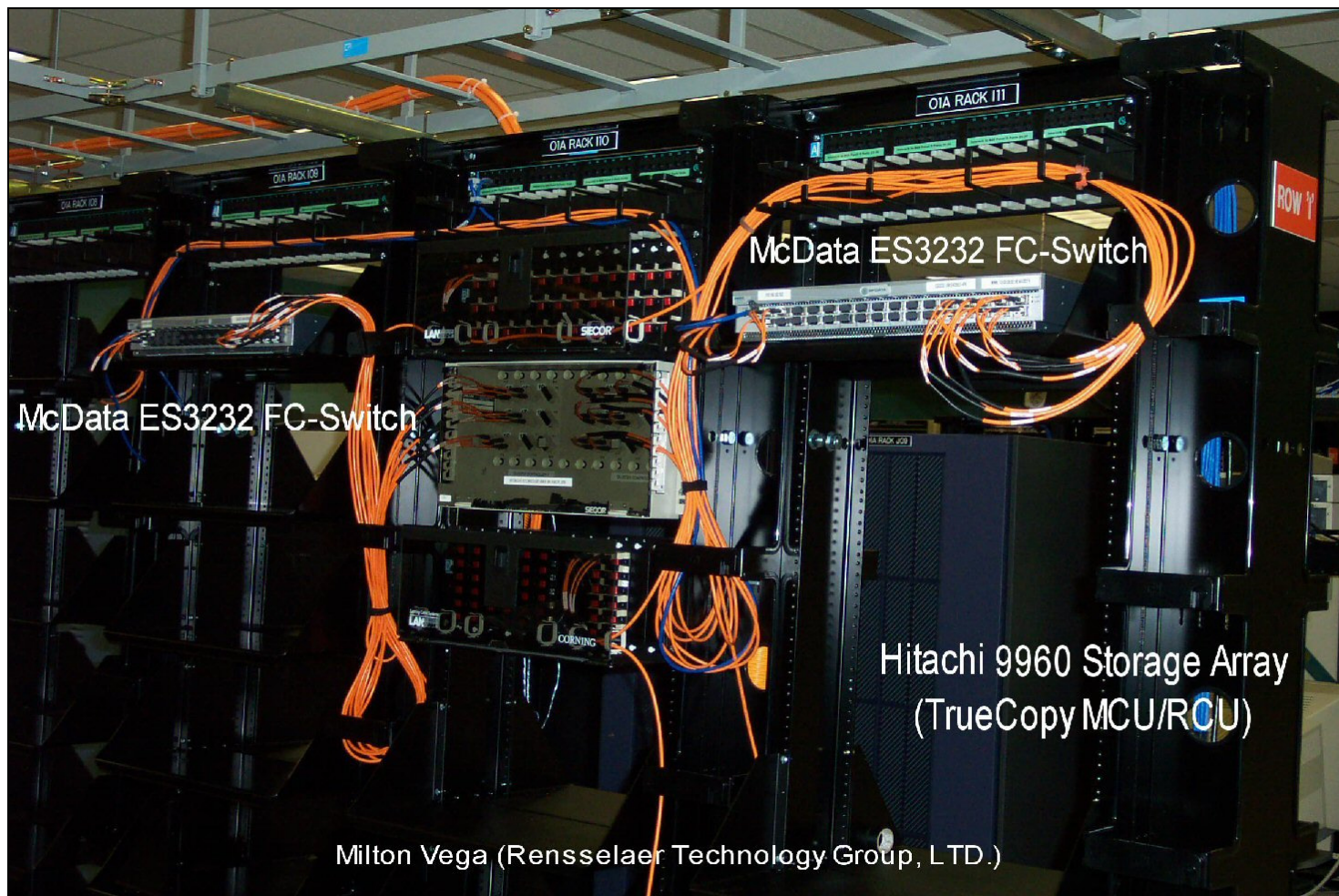


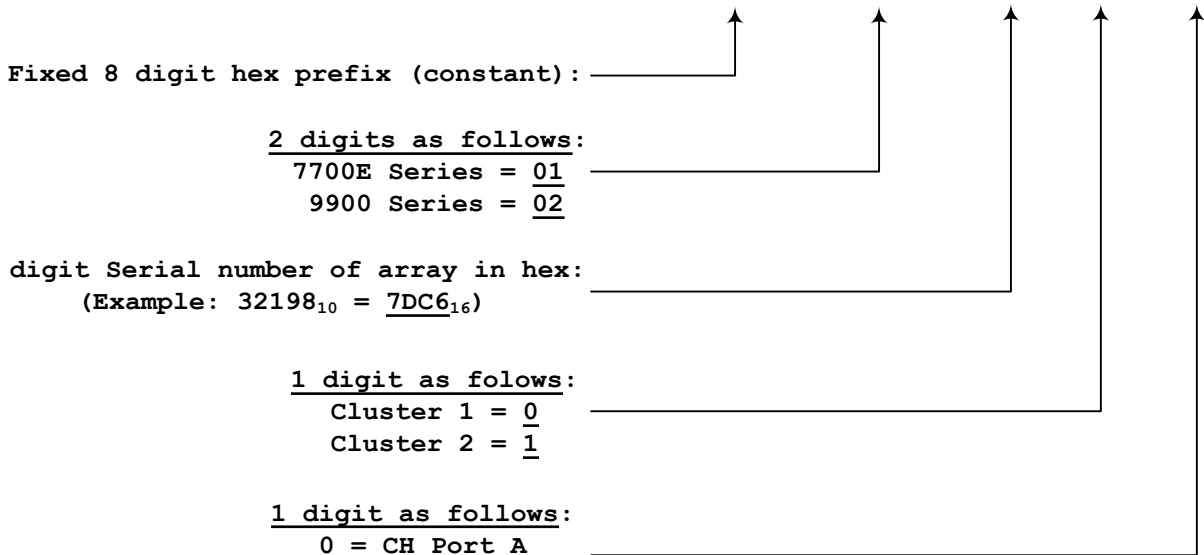
Figure 2: A photo of one datacenter (of two), in this TrueCopy enabled multi-site SAN solution. The photo illustrates the redundant ES3232 McData switches (which were temporarily connected via ISL's), as well as the HDS9960 array (the bluish-purple frame in the background towards the right). The design at the opposite end of the DWDM/ISL link (i.e. at the other datacenter 92Km away) is configured symmetrically, right down to the ports and LUN's used (as suggested in the diagram above)

Here, the integration of two McData ES3232 Fibre Channel switches serves the following purposes:

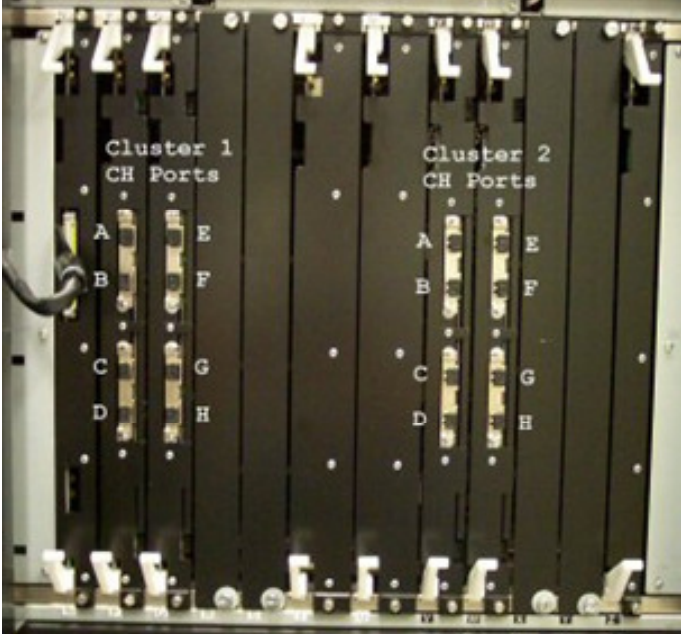
- (1) Front-ending the Hitachi 9960 to efficiently use Array ports by fanning them out in a 1 - N configuration to clients through the Fibre Channel switch. The switch on the left services the Cluster-1 ports of the array, while the switch on the right covers the Cluster-2 ports of the array. The switch provides F-Port to N-Port FC BB_Credits.
- (2) Ports 29 and 30 on each switch are configured as 2Gbps E-Ports to connect them to each other as local ISL's. Ports 31 of each switch, also configured as E-Ports, each connect to physically different DWDM equipment and circuit ID's (for redundancy - see diagram above). These DWDM connections create remote ISL's with two corresponding ES3232 FC switches at the downtown Manhattan site. Each Fibre Channel switch provides 60 EE_Credits, sufficient to cover the 96km linear fibre distance between the two sites, for uninterrupted streaming of data. The final result is a 4 switch, 128 port highly available multi-switch fabric, replicating mission critical data between sites for both Active/Active purposes, as well as disaster recover (DR) purposes.
- (3) Inter-site Hitachi TrueCopy (i.e. remote data replication) traffic commute through the DWDM ISL links. At least 2 logical paths between two Hitachi 99xx arrays are needed for reliable TrueCopy traffic going in one direction. Reliable Bi-Directional TrueCopy traffic therefore requires four. Since initially only two physical DWDM paths were allotted for FibreChannel traffic, the switches also serve to fan-out these remote ISL connections to create 4 logical inter-Site Fibre Channel paths out of two physical DWDM lines; thus allowing for (i.e. facilitating) Bi-Directional TrueCopy traffic.

Determining CH Port WWPN's of Hitachi HDS 7700E & 9900 series storage arrays.

General Format (Hex number): 500060E8 [01|02] [SSSS] [0|1] [0-F]



- 0 = CH Port A
- 1 = CH Port B
- 2 = CH Port C
- 3 = CH Port D
- 4 = CH Port E
- 5 = CH Port F
- 6 = CH Port G
- 7 = CH Port H
- 8 = CH Port I
- 9 = CH Port J
- A = CH Port K
- B = CH Port L
- C = CH Port M
- D = CH Port N
- E = CH Port O
- F = CH Port P



Noel Milton Vega

Figure 3: The WWPN of each HDS Array front-end CH ports can be determined using the algorithm above. Although it is possible to also determine the WWPN's by connecting the ports to a FibreChannel switch, when designing and building a SAN (two distinct steps), it is useful to know such information before hand: (1) its allows you to design and debug the Zone-Port diagram on paper first (see below); and (2) it allows you to pre-configure the FibreChannel switch Zone databases with Zones before anything is connected.

Version 10.0										
McData3232-SN:S405392-IPK IP Address: 192.168.102.191 / 255.255.254.0 / Detrouter: 192.168.102.52										
INPUTS						FAN IN(S)/OUT(S)				
SW Port#	Speed 1/2Gbps	Device Name/ID	Device Type	Device Adapter Id	Device Adapter WWPN	Device Name/ID	Device Type	Device Adapter Id	Device Adapter WWPN	SW Port#
0	F-Port / 2Gbps	SN 32210	HDS9960	CL1P-A (mode00)	500060e8027d4200	db02.prod.nymex.com	Sun v880	PCI Slot 7	21:00:00:E0:8B:07:6F:AC	McData3232-SN:S405392-IPK (port 8)
1	F-Port / 2Gbps	SN 32210	HDS9960	CL1P-B (mode00)	500060e8027d4201	c2202.prod.nymex.com	IBM x445	Adapter1 - Slot5	21:00:00:E0:8B:0F:D4:75	McData3232-SN:S405392-IPK (port 10)
2	F-Port / 2Gbps	SN 32210	HDS9960	CL1P-C (mode00)	500060e8027d4202					
3	F-Port / 2Gbps	SN 32210	HDS9960	CL1P-D (mode00)	500060e8027d4203					
4	F-Port / 2Gbps	SN 32210	HDS9960	CL1Q-E (mode00/RCLL-T)	500060e8027d4204	S/N 32198	HDS9960	CL1Q-E (mode00)	500060e8027dc604	McData3232-SN:S405762-1NE (port 4)
5	F-Port / 2Gbps	SN 32210	HDS9960	CL1Q-F (mode00/RCLL-T)	500060e8027d4205	S/N 32198	HDS9960	CL1Q-F (mode00)	500060e8027dc605	McData3232-SN:S405762-1NE (port 5)
6	F-Port / 2Gbps	SN 32210	HDS9960	CL1Q-G (mode00/Init)	500060e8027d4206	S/N 32198	HDS9960	CL1Q-G (mode00)	500060e8027dc606	McData3232-SN:S405762-1NE (port 6)
7	F-Port / 2Gbps	SN 32210	HDS9960	CL1Q-H (mode00/Init)	500060e8027d4207	S/N 32198	HDS9960	CL1Q-H (mode00)	500060e8027dc607	McData3232-SN:S405762-1NE (port 7)
8	F-Port / 1Gbps	db02.prod.nymex.com	Sun v880	PCI Slot 7	21:00:00:E0:8B:07:6F:AC	S/N 32210	HDS9960	CL1P-A (mode00)	500060e8027d4200	McData3232-SN:S405392-IPK (port 0)
9	F-Port / 1Gbps	db02.prod.nymex.com	Sun v880	PCI Slot 8	21:00:00:E0:8B:07:55:AE	S/N 32210	HDS9960	CL2V-A (mode00)	500060e8027d4210	McData3232-SN:S405392-IPK (port 22)
10	F-Port / 2Gbps	c2202.prod.nymex.com	IBM x445	Adapter1 - Slot5	21:00:00:E0:8B:0F:D4:75	S/N 32210	HDS9960	CL1P-B (mode00)	500060e8027d4201	McData3232-SN:S405392-IPK (port 11)
11	F-Port / 2Gbps	c2202.prod.nymex.com	IBM x445	Adapter0 - Slot6	21:00:00:E0:8B:0F:4D:76	S/N 32210	HDS9960	CL2V-B (mode00)	500060e8027d4211	McData3232-SN:S405392-IPK (port 23)
12										
13										
14										
15										
16										
17										
18										
19										
20										
21										
22	F-Port / 2Gbps	S/N:32210	HDS9960	CL2V-A (mode00)	500060e8027d4210	db02.prod.nymex.com	Sun v880	PCI Slot 8	21:00:00:E0:8B:07:55:AE	McData3232-SN:S405392-IPK (port 9)
23	F-Port / 2Gbps	S/N:32210	HDS9960	CL2V-B (mode00)	500060e8027d4211	c2202.prod.nymex.com	IBM x445	Adapter0 - Slot6	21:00:00:E0:8B:0F:4D:76	McData3232-SN:S405392-IPK (port 11)
24	F-Port / 2Gbps	S/N:32210	HDS9960	CH2W-E (mode00/RCLL-T)	500060e8027d4214	S/N:32198	HDS9960	CH2W-E (mode00)	500060e8027dc614	fcsw02-1NE (port 24)
25	F-Port / 2Gbps	S/N:32210	HDS9960	CH2W-F (mode00/RCLL-T)	500060e8027d4215	S/N:32198	HDS9960	CH2W-F (mode00)	500060e8027dc615	fcsw02-1NE (port 25)
26	F-Port / 2Gbps	S/N:32210	HDS9960	CH2W-G (mode00/Init)	500060e8027d4216	S/N:32198	HDS9960	CH2W-G (mode00)	500060e8027dc616	fcsw02-1NE (port 26)
27	F-Port / 2Gbps	S/N:32210	HDS9960	CH2W-H (mode00/Init)	500060e8027d4217	S/N:32198	HDS9960	CH2W-H (mode00)	500060e8027dc617	fcsw02-1NE (port 27)
28	E-Port / 1Gbps	A3-UNYMAR-B120 (6th. Fl.)	FSP-3000	Adva Port-3	N/A (DWDM ISL)	???	FSP-3000	Adva Port-3	N/A (DWDM ISL)	McData3232-SN:S405762-1NE (port 30)
29	E-Port / 2Gbps	ES3232-SerNumTBD-IPK	ES3232	Port 29	N/A (LOCAL IPK ISL)					
30	E-Port / 2Gbps	ES3232-SerNumTBD-IPK	ES3232	Port 30	N/A (LOCAL IPK ISL)					
31	E-Port / 1Gbps	A3-UNYMAR-B300 (2nd. Fl.)	FSP-3000	Adva Port-3	N/A (DWDM ISL)	???	FSP-3000	Adva Port-3	N/A (DWDM ISL)	McData3232-SN:S405762-1NE (port 31)

Figure 4: Zone-Port for one of the four (qty. 4) ES3232 McData switches. (Zoom in to see detail).

	A	B	C	D	E	F	G
1	CU:LDEV	EMUL	TYPE	~SIZE (MB)	Host Usage	CHA Port / LUN Mapping	Filesystem
2	00:00	OPEN-9	REGULAR	7042.5	db01.prod (Sun)	CH1A & 2A as LUN 00 (LUSE Head)	/u01
3	00:01	OPEN-9	REGULAR	7042.5	Luns 00-09	CH1A & 2A as LUN 00 (LUSE Disk)	/u01
4	00:02	OPEN-9	REGULAR	7042.5		CH1A & 2A as LUN 00 (LUSE Disk)	/u01
5	00:03	OPEN-9	REGULAR	7042.5		CH1A & 2A as LUN 00 (LUSE Disk)	/u01
6	00:04	OPEN-9	REGULAR	7042.5		CH1A & 2A as LUN 00 (LUSE Disk)	/u01
7	00:05	OPEN-9	REGULAR	7042.5	C22PROD (x445MP)	CH1B/2B as LUN 05 (LUSE Head)	G:/
8	00:06	OPEN-9	REGULAR	7042.5	Luns 04-07	CH1B/2B as LUN 05 (LUSE Disk)	G:/
9	00:07	OPEN-9	REGULAR	7042.5	C22QA (x445MP)	CH1D/1B/2B as LUN 00 (LUSE Head)	F:/
10	00:08	OPEN-9	REGULAR	7042.5	Luns 00-03	CH1D/1B/2B as LUN 00 (LUSE Disk)	F:/
11	00:09	OPEN-9	REGULAR	7042.5		CH1D/1B/2B as LUN 00 (LUSE Disk)	F:/
12	00:0a	OPEN-9	REGULAR	7042.5		CH1D/1B/2B as LUN 00 (LUSE Disk)	F:/
13	00:0b	OPEN-9	REGULAR	7042.5		CH1D/1B/2B as LUN 00 (LUSE Disk)	F:/
14	00:0c	OPEN-9	REGULAR	7042.5		CH1D/1B/2B as LUN 00 (LUSE Disk)	F:/
15	00:0d	OPEN-9	REGULAR	7042.5			
16	00:0e	OPEN-9	REGULAR	7042.5			
17	00:0f	OPEN-9	REGULAR	7042.5			
18	00:10	OPEN-9	REGULAR	7042.5			
19	00:11	OPEN-9	REGULAR	7042.5			
20	00:12	OPEN-9	REGULAR	7042.5			
21	00:13	OPEN-9	REGULAR	7042.5			
22	00:14	OPEN-9	REGULAR	7042.5			
23	00:15	OPEN-9	REGULAR	7042.5			
24	00:16	OPEN-9	REGULAR	7042.5			
25	00:17	OPEN-9	REGULAR	7042.5			
26	00:18	OPEN-9	REGULAR	7042.5			
27	00:19	OPEN-9	REGULAR	7042.5			
28	00:1a	OPEN-9	REGULAR	7042.5			
29	00:1b	OPEN-9	REGULAR	7042.5			
30	00:1c	OPEN-9	REGULAR	7042.5			
31							

Figure 5: In the HDS9900 series (9960/9970/9980/9990) upon initial configuration, RAID5 PARITY GROUPS are created using either (3+1 = Basic 4) or (7+1 = Basic 8) configurations. Next, from those PARITY GROUPS, smaller logical devices are created by splitting each up according to the Open Emulation desired. Above we see a parity group formed by using four 72GB drives, which was subsequently split (at array initialization time) using an OPEN-9 Emulation. The resulting CU:LDEV devices could be assigned to hosts by mapping them to CH ports, and implementing appropriate LUN masks (via Santinel).

Some Tools Used:

- San Pilot & CLI (McData ES3232).
- Remote Console 4.x and HDS/StoreEdge 9960 Service Processor Laptop.
- Santinel for LUN Masking with the HDS 9960.
- Hitachi Command Control Devices (CCD) and horcm (similar to EMC's gatekeeper and ECC commands)
- Hitachi DataLink Manager (HDLM) (also MPXIO)
- Hitachi Cruise Control (and optimizer similar to EMC's SymOptimizer).
- LUSE devices (Logical Unit Storage Extension) (similar to EMC's Metadevice).
- Qlogic 23xx HBA's (Some I flashed with FCODE for use with Sun SPARC servers; and others I flashed with BIOS Code for use with Windows & Linux x86 based servers).

Parenthetical note: The HDS9960 arrays above were acquired through Sun, as StoreEdge 9900 arrays. The firm's plans were to transform to a Windows based trading platform, and because they did not recognize that these Sun re-badged Hitachi arrays could also be used for Windows, Linux (etc.) hosts, the firm's development roadmap had no plans to use these arrays, even though they were practically brand new, and cost a significant amount to acquire. With no documentation (except those on a micro-code CD I found) and with no support contract from Sun or Hitachi, I convinced the, then VP of Technology, to allow me to divert some of my consulting time to investigating and building the local and distance SAN solution above.

To do this, it was necessary for me to take a proof-of-concept approach with these arrays, before I could convince the firm to purchase FibreChannel switches, use DWDM lines, or do anything that would incur collateral cost to implementing these arrays (which, again, they had no plans for). Towards that end I had to do things like, get the two arrays communicating/replicating via FC-AL using a simple back-to-back 50µm cable; then get them replicating via NISHAN IPS4000 FCIP switches (i.e. so the IP infrastructure could be used to test site-to-site replication without using DWDM lines), etc. Once the proof-of-concept tests were demonstrated to work, the firm allowed me to design the FibreChannel fabric based infrastructure above.