CASPUR Storage Lab

Andrei Maslennikov
CASPUR Consortium

Catania, April 2002

Contents

- Reasons to build a lab
- Agenda
- Components
- Testing plans
- First results - Protocols
- First results - SCSI over IP
- First results - GFS
- Final remarks

Why should we build a lab?

- Objective: inventory and comparative studies for both current and new storage solutions.

- General issues to look at:

  - True data sharing across architectures
  - Best performance, scalability
  - Efficient remote data access
  - Performance and reliability, but possibly cheaper components

- Questions that we have to answer in a short run:

  We soon will have to upgrade our NAS services (need scalable NFS, must migrate to OpenAFS):

  - Can we replace our NetApp F760 (NFS) and Sun

Agenda

- Large file serving across architectures

- File serving off Linux-based servers: performance,
  limitations, hardware issues

- File serving for Linux clients: new solutions

- Data access over WAN

- New disk and tape media

Components

- High-end base Linux unit for both servers and clients
  - SuperMicro Superserver 6041G with:
    2 x Pentium III 1000 MHz
    2 GB of RAM, dual channel 160 MB SCSI on board
    SysKonnect 9843 Gigabit Ethernet NIC
    Qlogic QLA2200 Fibre Channel HBA
    System disk: 15000 RPM (Seagate)

- Network
  - NPI Keystone 12-port switch (throughput 12 Gbit)
  - Myricom Myrinet 8-port switch, 4 nodes attached

- Wide Area Lab: in collaboration with CNAF(INFN)
  - 2 identical Dell Poweredge 1650 servers

Components -2

Disks:

scsi : several 15K RPM local units

scsi-fc : 7.5K and 10K RAID systems (DotHill)

fc-fc : 10K RAID 256MB cache (on loan from DotHill)

fc-fc : 15K RAID 1GB cache (on loan from IBM, rriving)

ide-fc : RAID (Infortrend base with IBM disks, just rdered)

Tapes:

4 LTO fc Ultrium drives via SAN

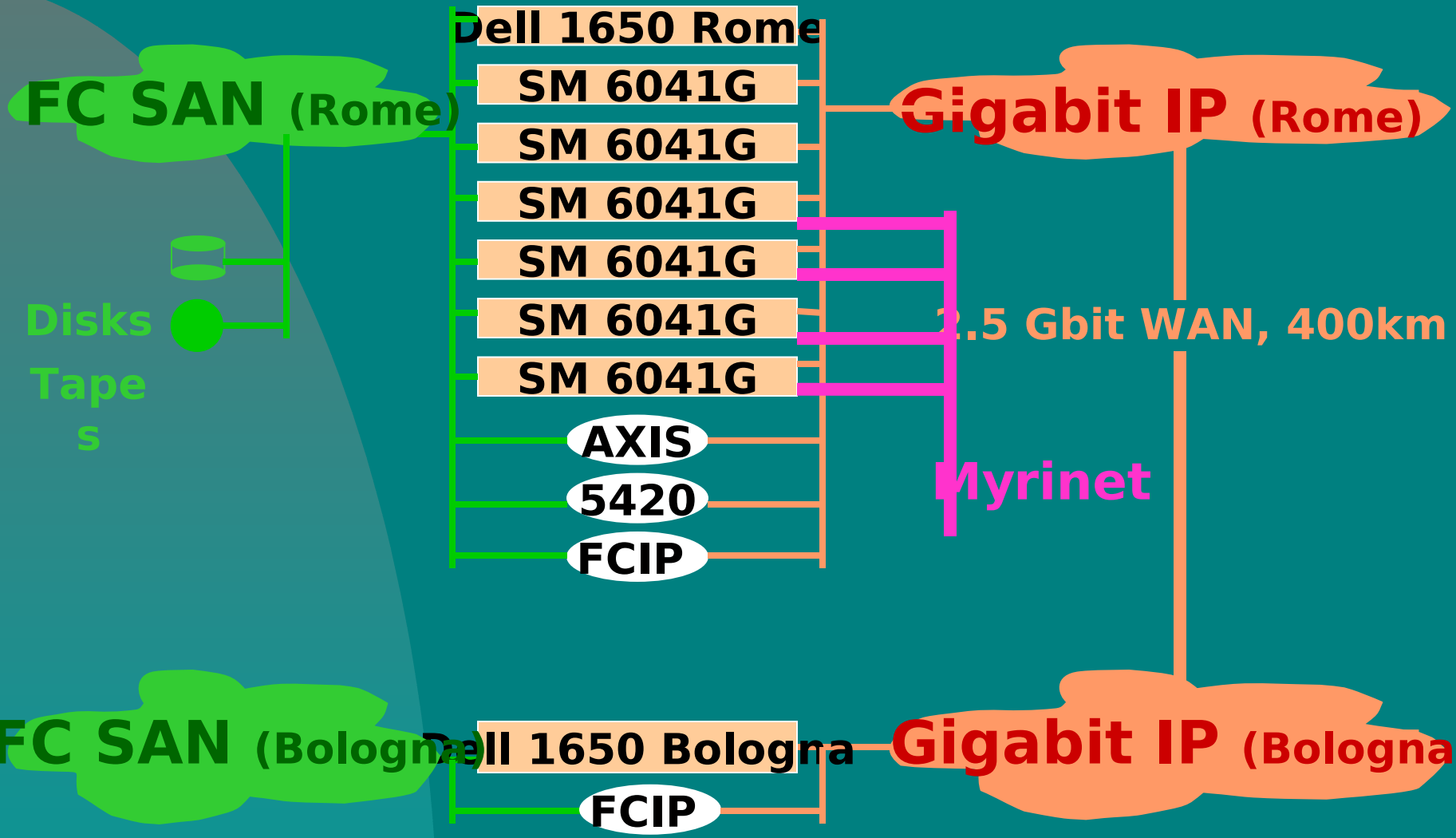
2 AIT-3 fc drives via SAN (on loan from ENEA, just rrived)

SCSI / IP appliances:

CISCO SN5420 appliance (Fibre Channel / iSCSI) – on loan rom CISCO, now bought it

DotHill Axis appliance (Fibre Channel / Ipstor) – on loan rom DotHill

CASPUR / CNAF Storage Lab



Testing Plans

Series 1. Comparison of the file transfer methods for large files

- Setup : One server with a local disk, several clients on the network.
- Goals : Benchmark several most commonly used file transfer methods:
  NFS, AFS, AFS-cacheless(Atrans), RFIO, ROOT, GridFTP,
  both on LAN and over WAN. Use large files (>1 GB).

Series 2. Study of SCSI-over-IP solutions

- Setup : Fibre channel devices (tapes, disks), FC / IP appliances,
  Study the case of multiple clients accessing the same server
  tcp offload-capable NICs, clients on LAN and WAN
  Study other Linux file systems for large file hosting on the server
- Goals : Provide client access over IP for native fibre channel devices,
  (64 bit, volume manager etc).
  in a variety of ways (Ipstor, iSCSI, and others). Study SAN
  interconnection on the WAN (FCIP, iFCP, SoIP

Testing Plans, 2

Series 3. Study of serverless disk sharing

- Setup : Fibre channel disk devices accessible from several clients on the LAN
- Goals : Configure and study: Sistina Global File System, IBM Sanergy.
  For DMEP-capable devices, try hardware locking (with GFS).
  See if GFS may be used for HA configurations (mail, web, dns etc).

Series 4. Scalable NFS server based on IBM GPFS

- Setup : Several server nodes with local disk interconnected with a fast,
  low-latency network; several client nodes.
- Goals : Configure IBM GPFS, benchmark peak performance on the clients.
  Benchmark also the aggregate perfomance of the multinode
  server complex. Calculate the costs

Testing Plans, 3

Series 5. Study of the new media (AIT-3, ide-fc)

- Setup : New media (AIT-3 tapes, ide-fc RAID systems), test machines.

- Goals : Configure systems and run a series of stress tests.

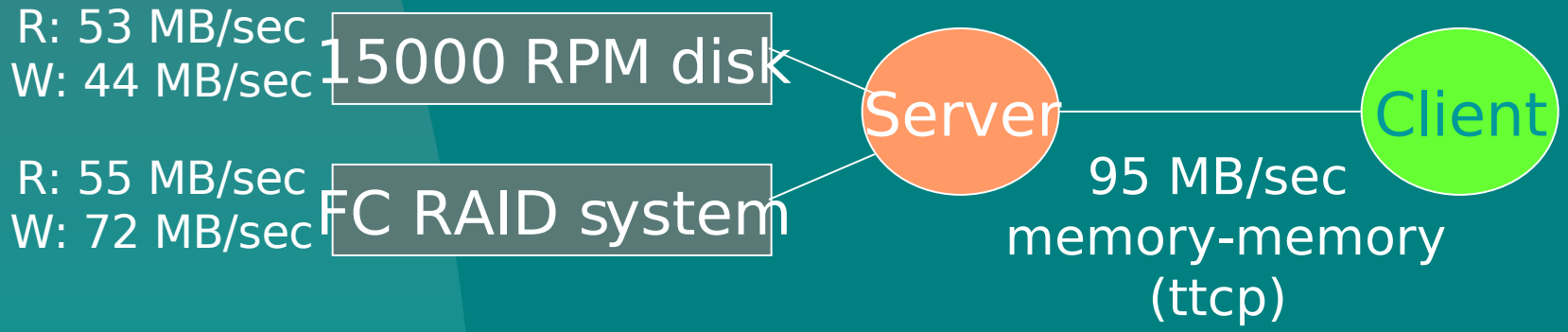
Benchmark the performance.

First results – Series 1 (Protocols)

Participated:

- CASPUR : A.Maslennikov, G.Palumbo.
- CERN : F.Collin, J-D.Durand, G.Lee, F.Rademakers, R.Többicke.

Hardware configuration:



Series 1 - details

Some settings:

- Kernel: 2.4.17 (no kswapd problem)
- AFS : cache was set up on ramdisk (400MB), chunksize=256 KB
- NFS : version=3, rsize=wsize=8192
- used ext2 filesystem on servers

Problems encountered:

- Two highly performant cards on the same PCI bridge interfere visibly.
  There are indications that FC HBA and Gigabit NIC, when used together,
  may hit the ceiling even when they are using two different bridges.

Series 1 - more detail

Write tests:

- Measured average time needed to transfer 1 GB from memory on the client to the disk of the file server, including the time needed to run “sync” command on both client and the server at the end of operation:

dd if=/dev/zero of=<filename on server> bs=1000k count=1000

T=Tdd + max(Tsyncclient, Tsyncserver)

For RFIO, this was done via a named pipe;

For ROOT, 1GB file on client was first put in memory with “cat” command

Series 1 - more detail

Read tests:

- Measured average time needed to transfer 1 GB file from a disk on the server
  to the memory on the client (output directly to /dev/null).

- Reading was done in a loop over groups of 10 different files of 1GB each,
  so it was guaranteed that neither client nor server had any part of the file
  in the memory, at the moment when the file was read.

Series 1- current results (MB/sec)

	RAID read	RAID write	15K read	15K write
Pure disk	55.0	72.5	53.0	43.6
AFS	21.1	19.5	21.8	16.1
NFS	23.3	33.2	25.7	26.7
AFS(Atrans)	29.4	24.3	29.6	20.8
RFIO	42.1	23.6	50.0	19.7
ROOT	29.7	25.6	43.3	20.1

Next steps: - **LAN measurements with a new extrafast IBM disk**

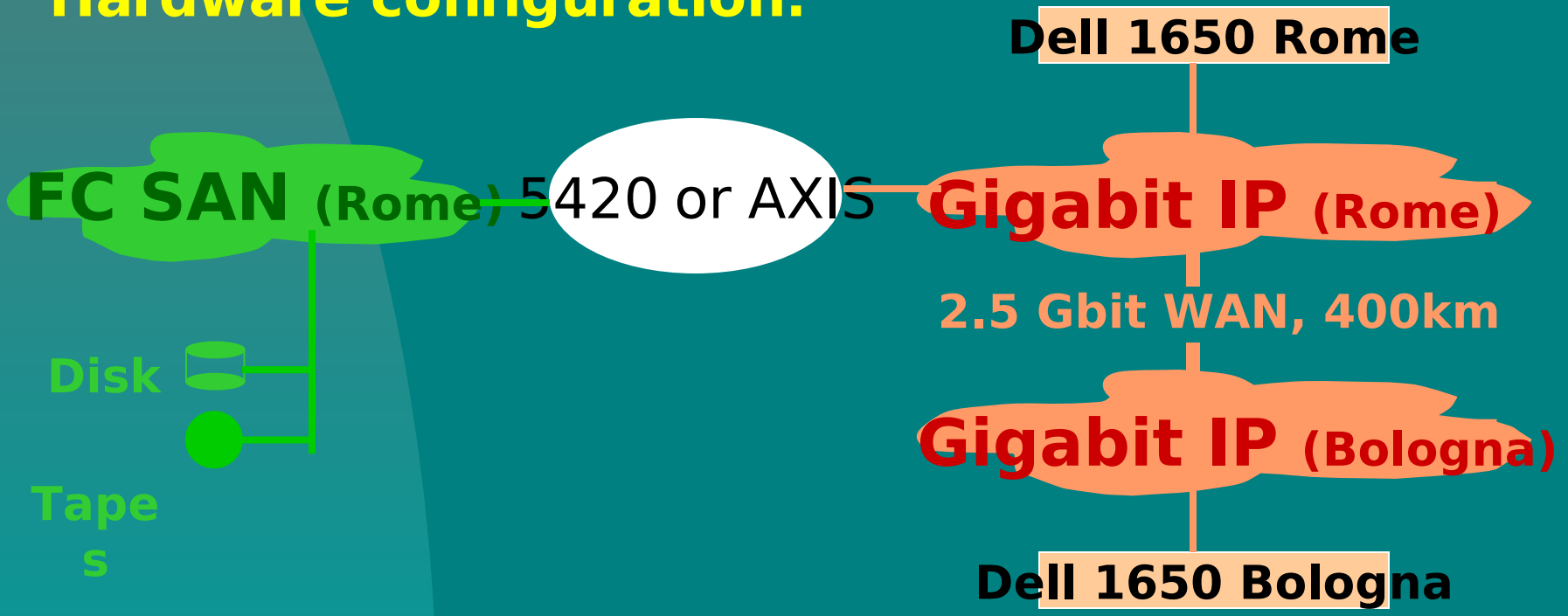
- **Will try XFS filesystem on server**
- **WAN measurements**
- **Aggregate max speeds**
- **GridFTP / bbftp benchmarks**

Series 2 (SCSI over IP)

Participated:

- CASPUR : M.Goretti, A.Maslennikov, G.Palumbo.
- CNAF : PP.Ricci, F.Ruggieri, S.Zani.

Hardware configuration:



Series 2 - details

TCP settings for WAN:

- With default TCP settings, we have obtained these speeds on the WAN link between the two Dell 1650 servers:

11 MB/sec (TCP, ttcp test)
~100 MB/sec (UDP, netperf test)

We used then the B.Tierney's cookbook, and got advice from L.Pomelli (CISCO), and S.Ravot(CERN). In the end, TCP window size was set to 256 Kbytes (we tried different values), and our best results were:

65 MB/sec on kernel 2.4.16 (TCP, ttcp test)
15 MB/sec on kernel 2.2.16 (TCP, ttcp test)

- Obviously, AXIS performance on WAN was expected to be poor, because

Series 2 - more detail

What was measured:

- Write tests: average time needed to transfer 1 GB from memory on the client to the iSCSI or ipstor disk or tape, including the time needed to run "sync" command on the client at the end of operation.

- Read tests: average time needed to transfer 1 GB file from iSCSI or ipstor disk or tape to the memory on the client.

Like in the Series 1 tests, reading was done in a loop over several different files of 1GB each.

Series 2- current results (MB/sec)

	LAN read	LAN write	WAN read	WAN write
5420 Disk	34	30	7	20
5420 Tape	15	15	7	14
5420 4 Tapes		60		40
AXIS Disk	29	29	7.5	9.2
AXIS Tape	15	15		4.3
AXIS 4 Tapes		38		9.6

R/W speed on native Fibre Channel HBA: this disk: 56/36, this tape: 15/15.

Notes: **- New CISCO firmware may further improve the aggregate speed on 5420**

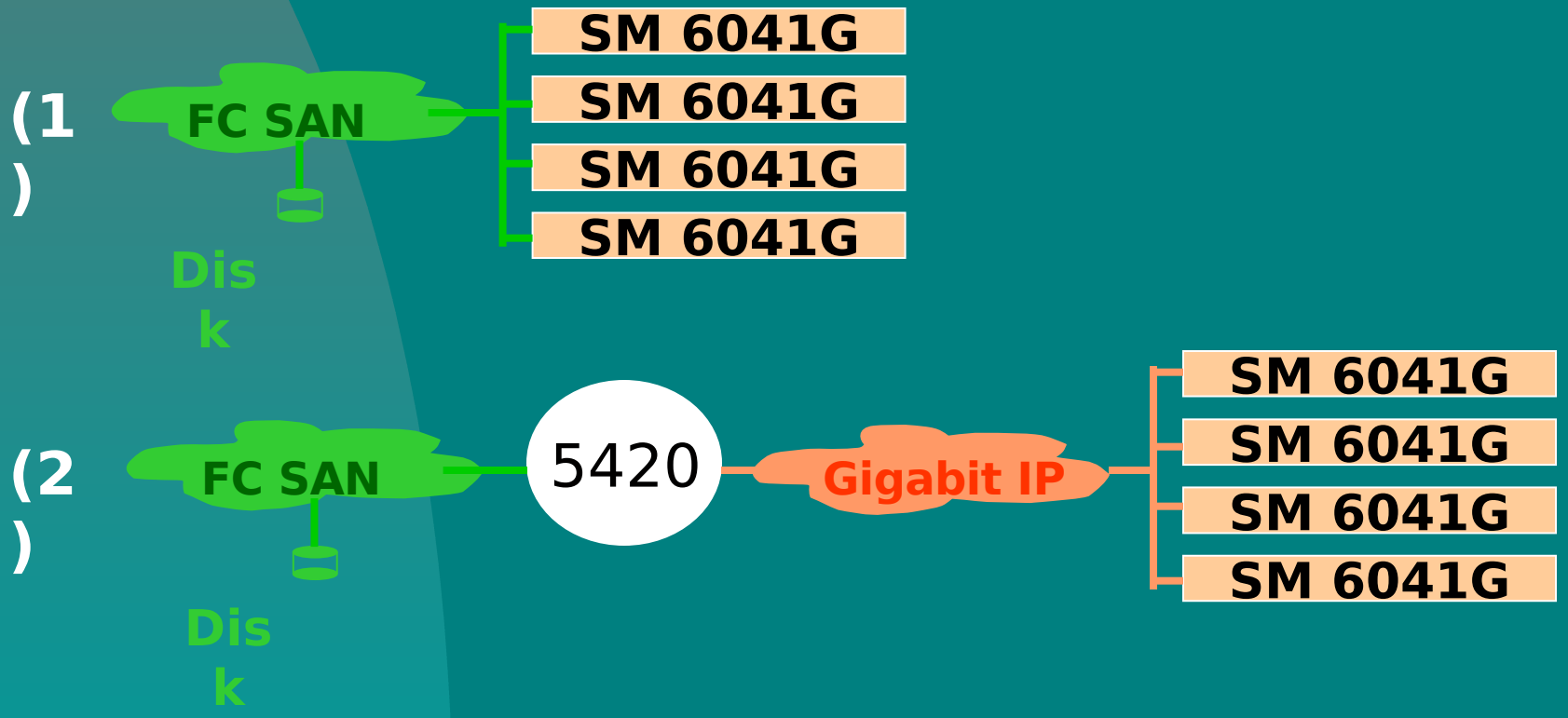
- Waiting for AXIS sw upgrade to repeat the WAN tests with kernel 2.4.x

Series 3 (Global File System)

Participants:

- CASPUR : A.Maslennikov, G.Palumbo.

Hardware configuration (2 variants):



Series 3 - details

GFS installation:

- Requires kernel 2.4.16 (may be downloaded from Sistina together with the trial distribution). On fibre channel, everything works out of the box.

- CISCO driver required recompilation. Compiled smoothly but would not work with Sistina kernel (we used the right 2.4.16 source tree, complete with their patches).

- Found a workaround: rebuilt kernel with 2.4.16 source + Sistina patches. Then CISCO driver compiled and loaded smoothly, but Sistina modules would not load. Hacked them with the “objcopy”. All then worked automagically.

Series 3 – GFS current results (MB/sec)

NB: - Out of 4 nodes:
1 node was running the lock server process
3 nodes were doing only I/O

	FC read	FC write	Lock server CPU
1 client	42	35	0%
2 clients	33	36	30%
3 clients	37	27	75%

	5420 read	5420 write	Lock server CPU
1 client	20	32	0%
2 clients	24	27	30%
3 clients	27	26	60%

R/W speed on native Fibre Channel HBA:
this disk: 56/36

Next steps: - Will repeat benchmarks with the disk from Series 1 test, and
compare them with those for the other methods
- Will be exploring hardware DMEP

Final remarks

- We will continue with the tests, and any comment
  is very welcome

- Vendors see these activities with a good eye, so new
  hardware may be arriving for tests, at no charge.

- All of a sudden, it may become a big job (it is already!)
  Could we join the forces? We opt for setting up a
  storage working group