GridPP
UK Computing for Particle Physics

Tier1 Status Report

Andrew Sansum

Service Challenge Meeting

27 January 2004

CCLRC

Overview

What we will cover:

Andrew	Tier-1 local infrastructure for Service Challenges
Robin	Site and UK networking for Tier 1
John (today)	GRIDPP Storage Group plans for SRM
John (Tomorrow)	Long range planning for LCG

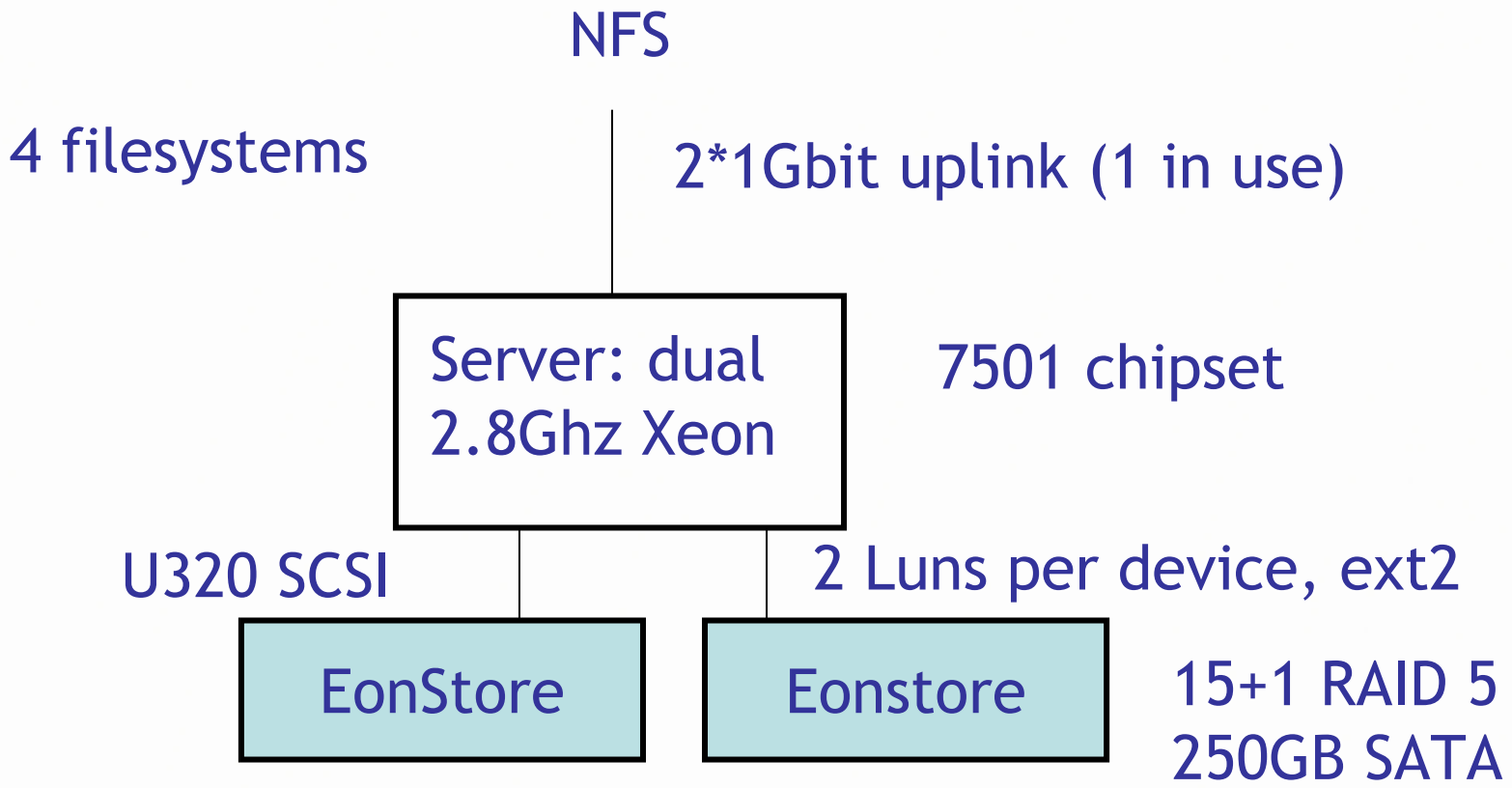
SC2 Team

- Intend to share load among several staff:
  - RAL to CERN Networking: Chris Seelig
  - LAN and hardware deployment: Martin Bly
  - Local System Tuning/RAID: Nick White
  - dCache: Derek Ross
  - Grid Interfaces: Steve Traylen
- Also expect to call on support from:
  - GRIDPP Storage Group (for SRM/dCache support)
  - GRIDPP Network Group (for end to end Network Optimisation)

Tier1A Disk

- 2002-3 80TB
  - Dual Processor Server
  - Dual channel SCSI interconnect
  - External IDE/SCSI RAID arrays (Accusys and Infortrend)
  - ATA drives (mainly Maxtor)
  - Cheap and (fairly) cheerful
  - 37 servers
- 2004 (140TB)
  - Infortrend Eonstore SATA/SCSI RAID Arrays
  - 16*250GB Western Digital SATA per array
  - Two arrays per server
  - 20 servers

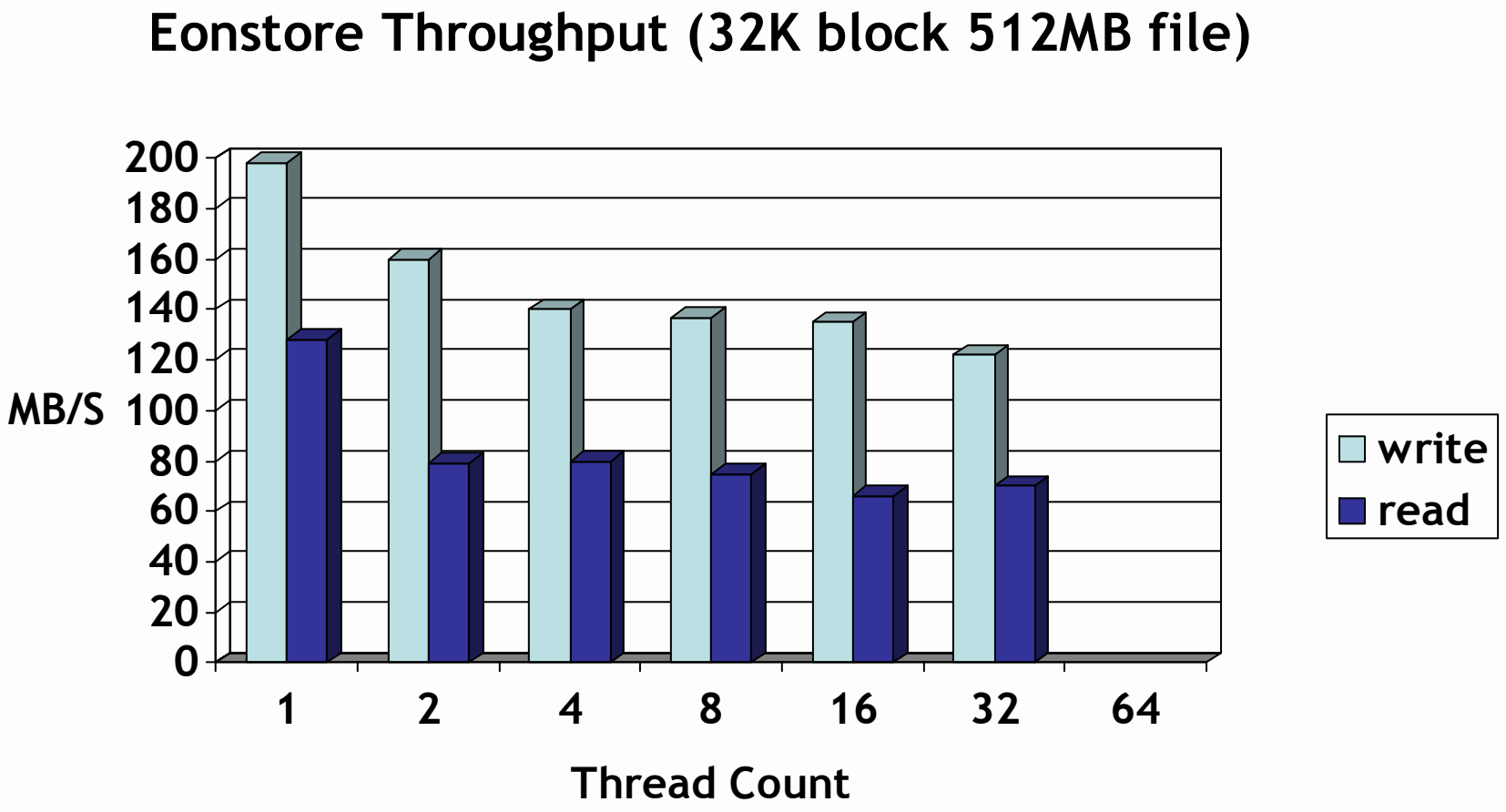
Typical configuration



Disk

- Ideally deploy production capacity for Service Challenge
  - Tune kit/sw we need to work well in production
  - Tried and tested – no nasty suprises
  - Don't invest effort in one off installation
  - OK for 8 week block provided no major resource clash, problematic for longer than that.
  - 4 servers (maybe 8) potentially available
- Plan B – deploy batch worker nodes with extra disk drive.
  - Less keen – lower per node performance …
  - Allows us to retain the infrastructure

Device Throughput



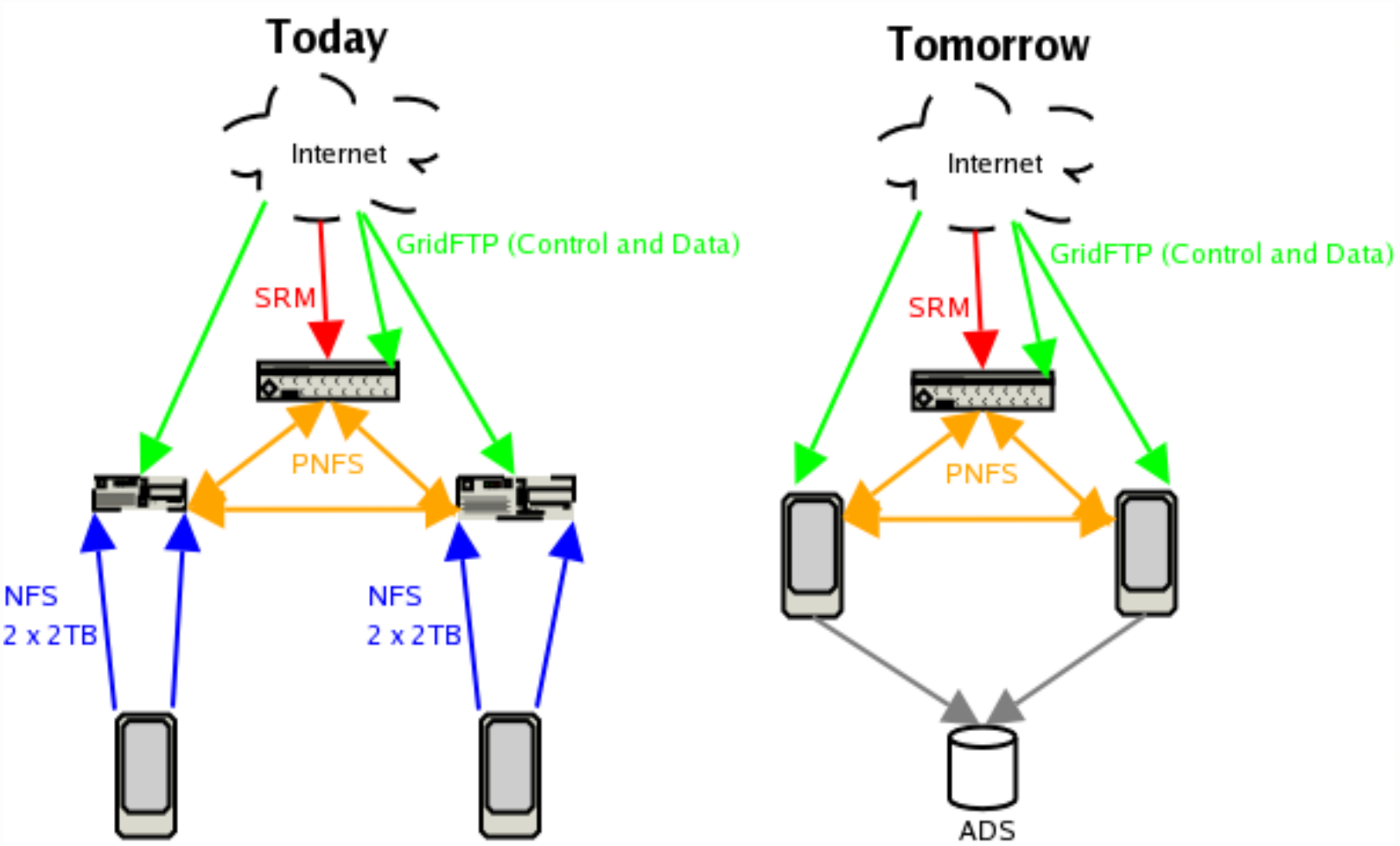
Performance Notes

- Rough and ready preliminary check, – out of the box. Redhat 7.3 with kernel 2.4.20-31. Just intended to show capability – no time put into tuning yet.
- At low thread count system caching is impacting benchmark. With larger files 150MB is more usual for write.
- Read performance seems to have a problem,probably a hyper-threading effect – however pretty good at high thread count.
- Probably can drive both arrays in parallel

dCache

- dCache deployed as production service (also test instance, JRA1, developer1 and developer2?)
- Now available in production for ATLAS, CMS, LHCB and DTEAM (17TB now configured – 4TB used)
- Reliability good – but load is light
- Work underway to provide Tape backend, prototype already operational. This will be production SRM to tape for SC3
- Wish to use dCache (preferably production instance) as interface to Service Challenge 2.

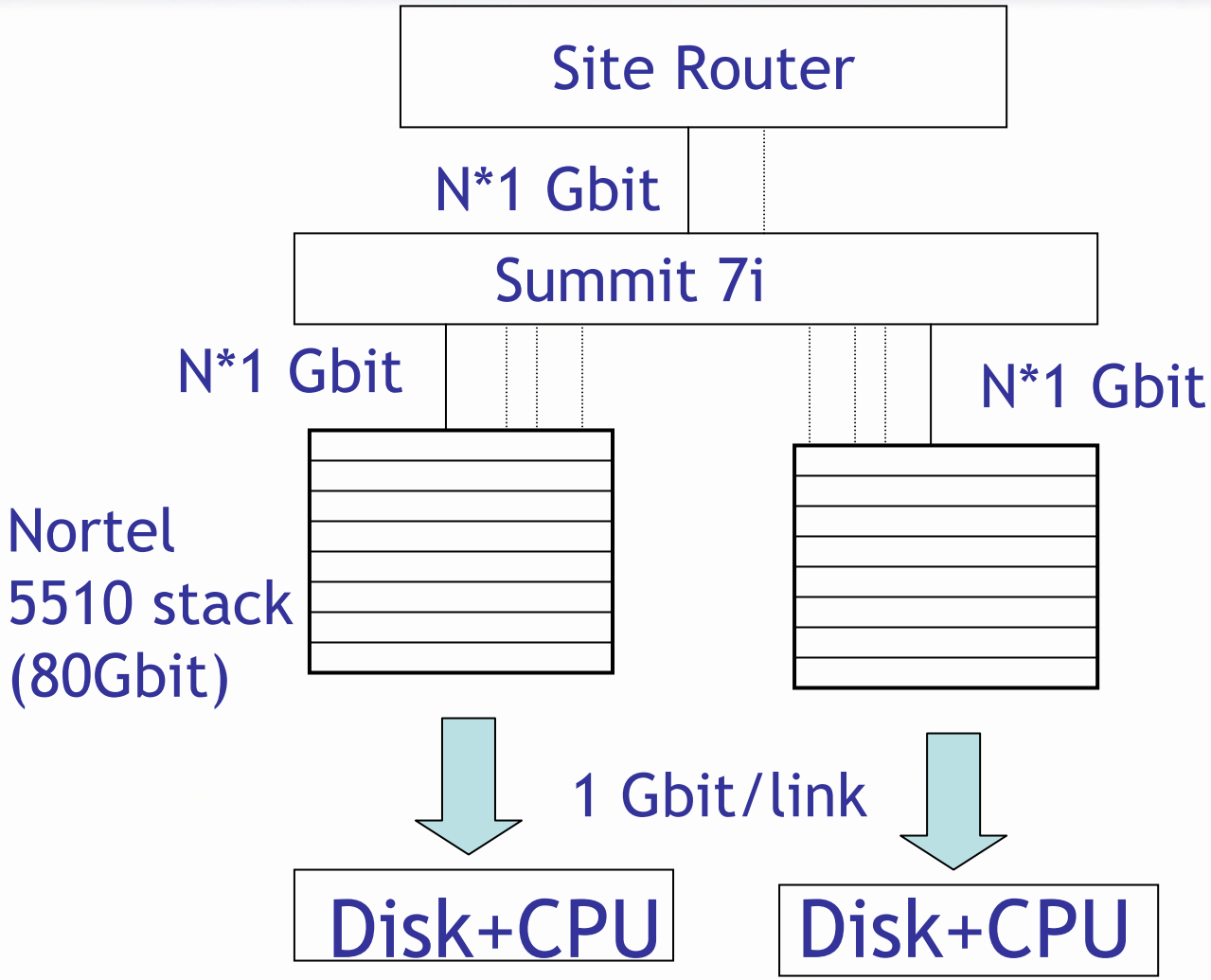
Current Deployment at RAL



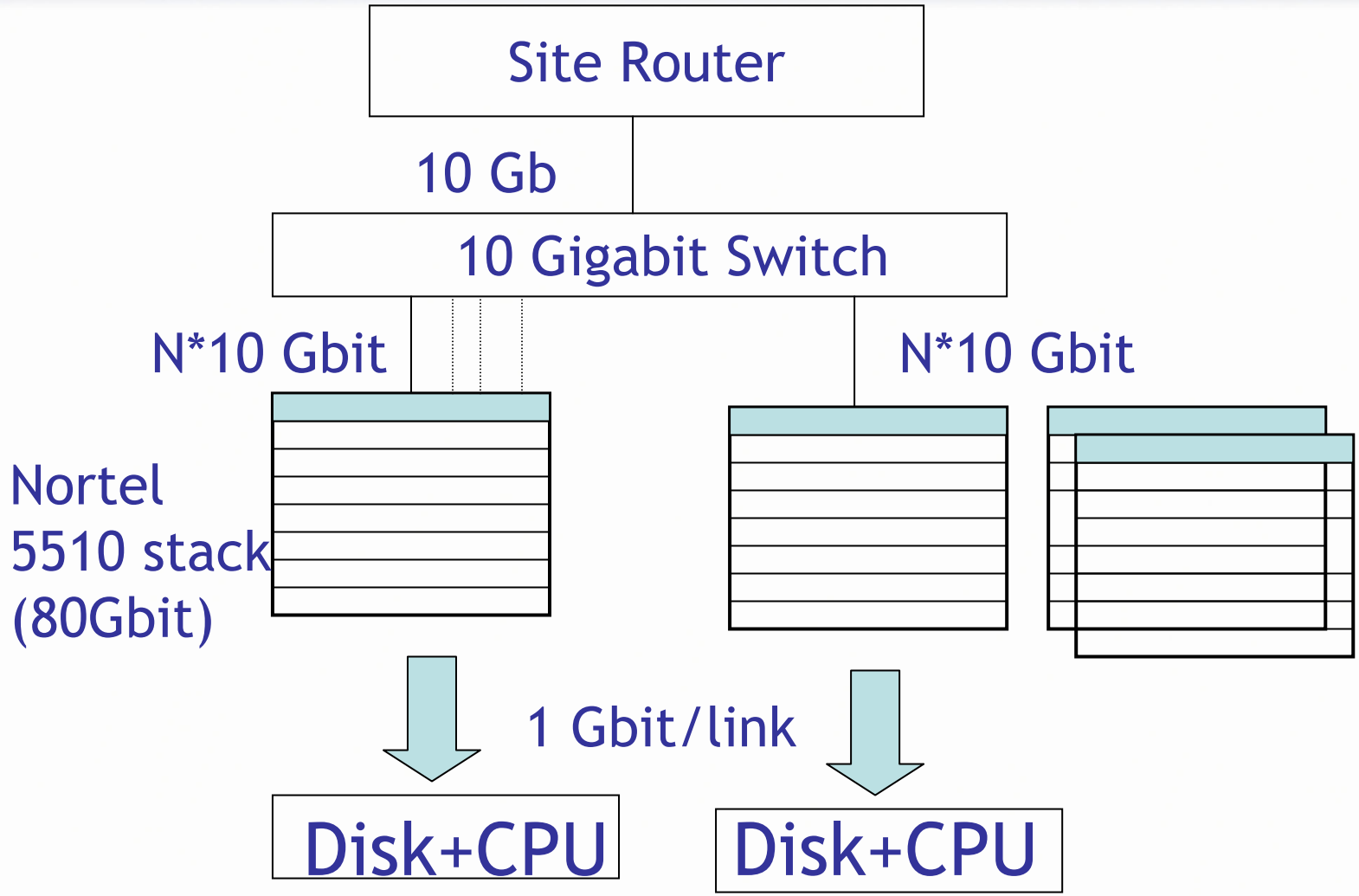
LAN

- Network will evolve in several stages
- Choose low cost solutions
- Minimise spend until needed
- Maintain flexibility
- Expect to be able to attach to UKLIGHT by March (but see Robin's talk).

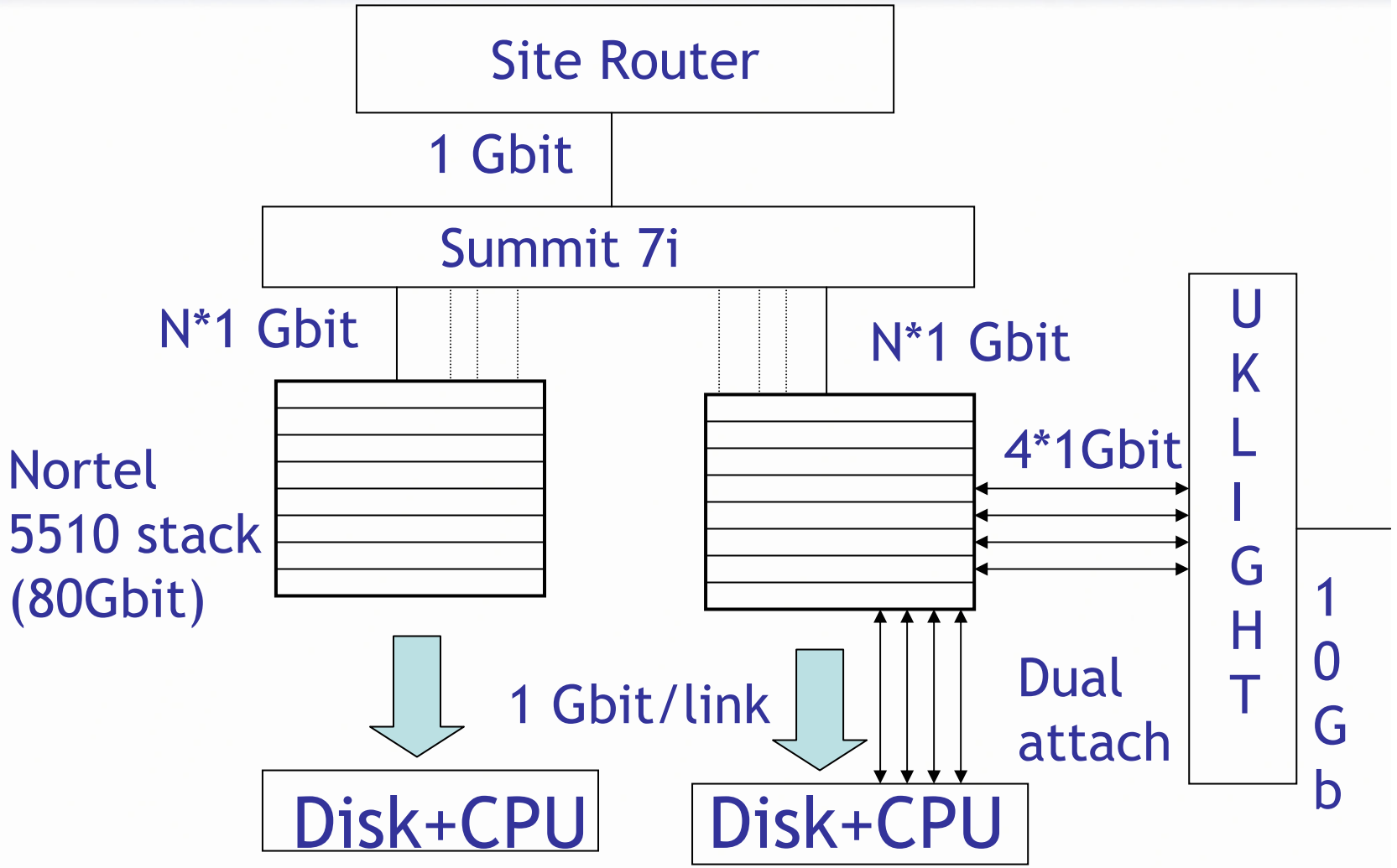
Now (Production)



Next (Production)

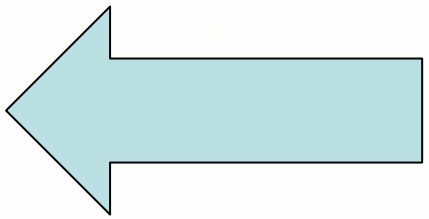


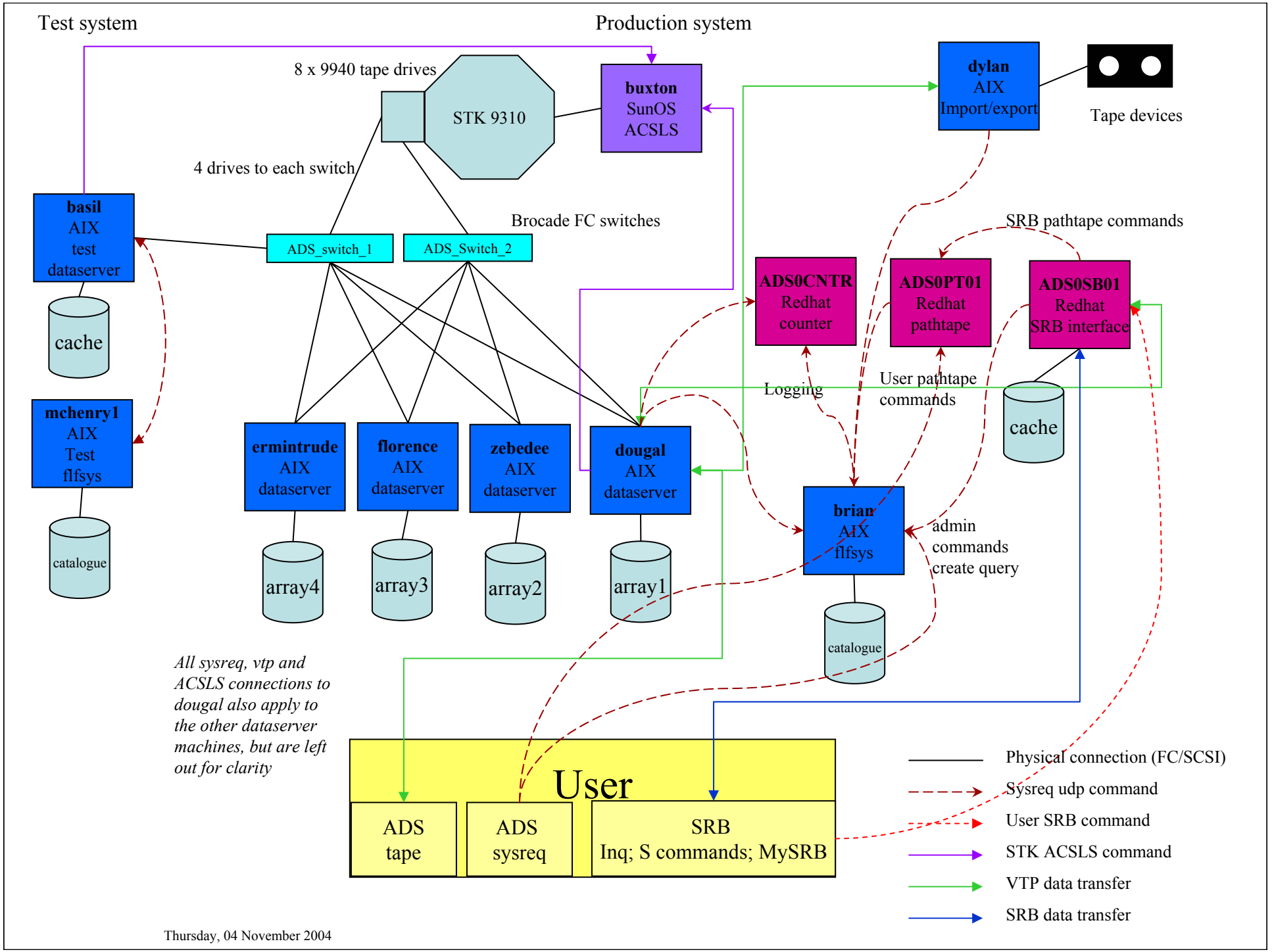
Soon (Lightpath)

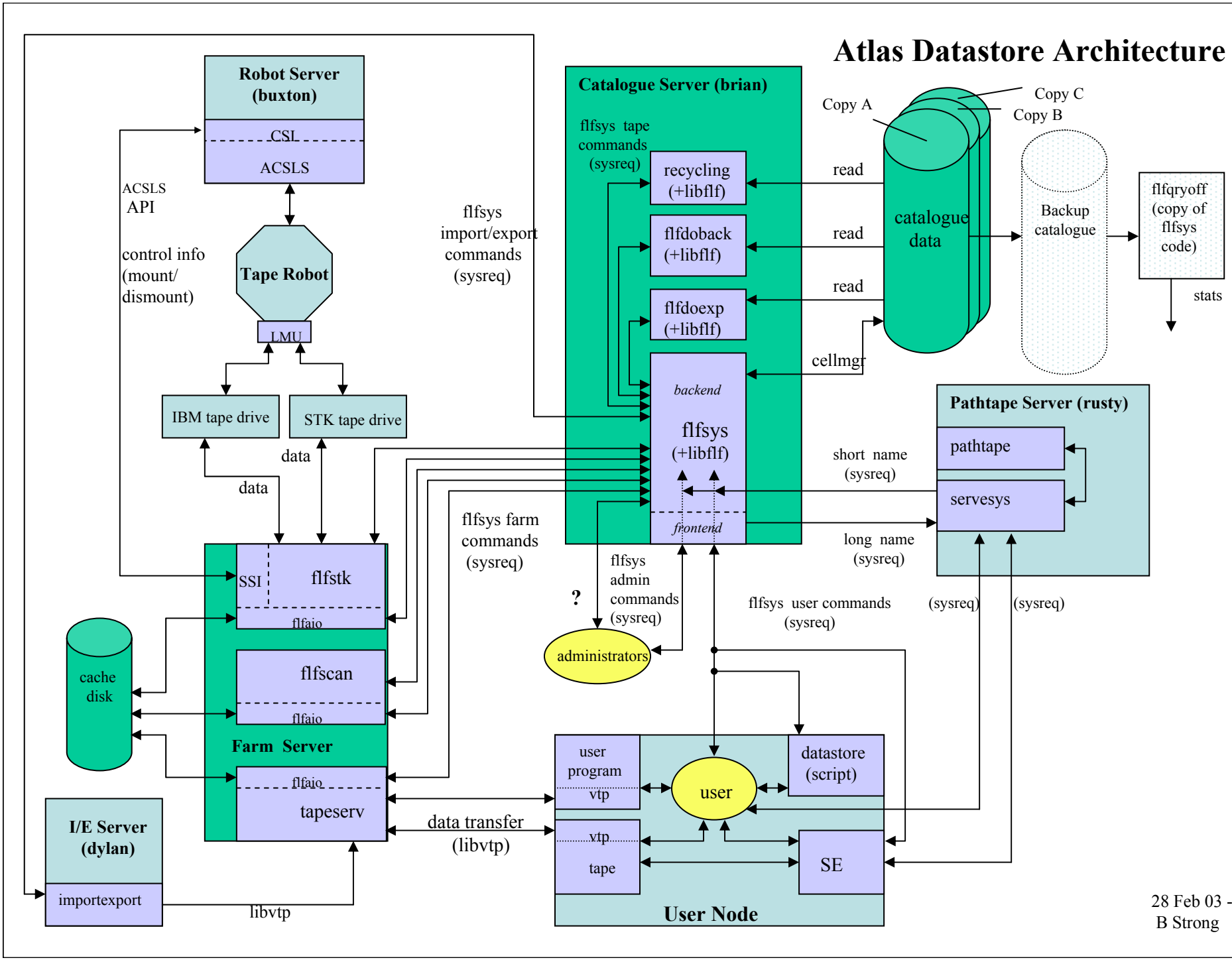


MSS Stress Testing

- Preparation for SC 3 (and beyond) underway (Tim Folkes). Underway since August.
- Motivation – service load has been historically rather low. Look for "Gotchas"
- Review known limitations.
- Stress test – part of the way through the process – just a taster here
  - Measure performance
  - Fix trivial limitations
  - Repeat
  - Buy more hardware
  - Repeat



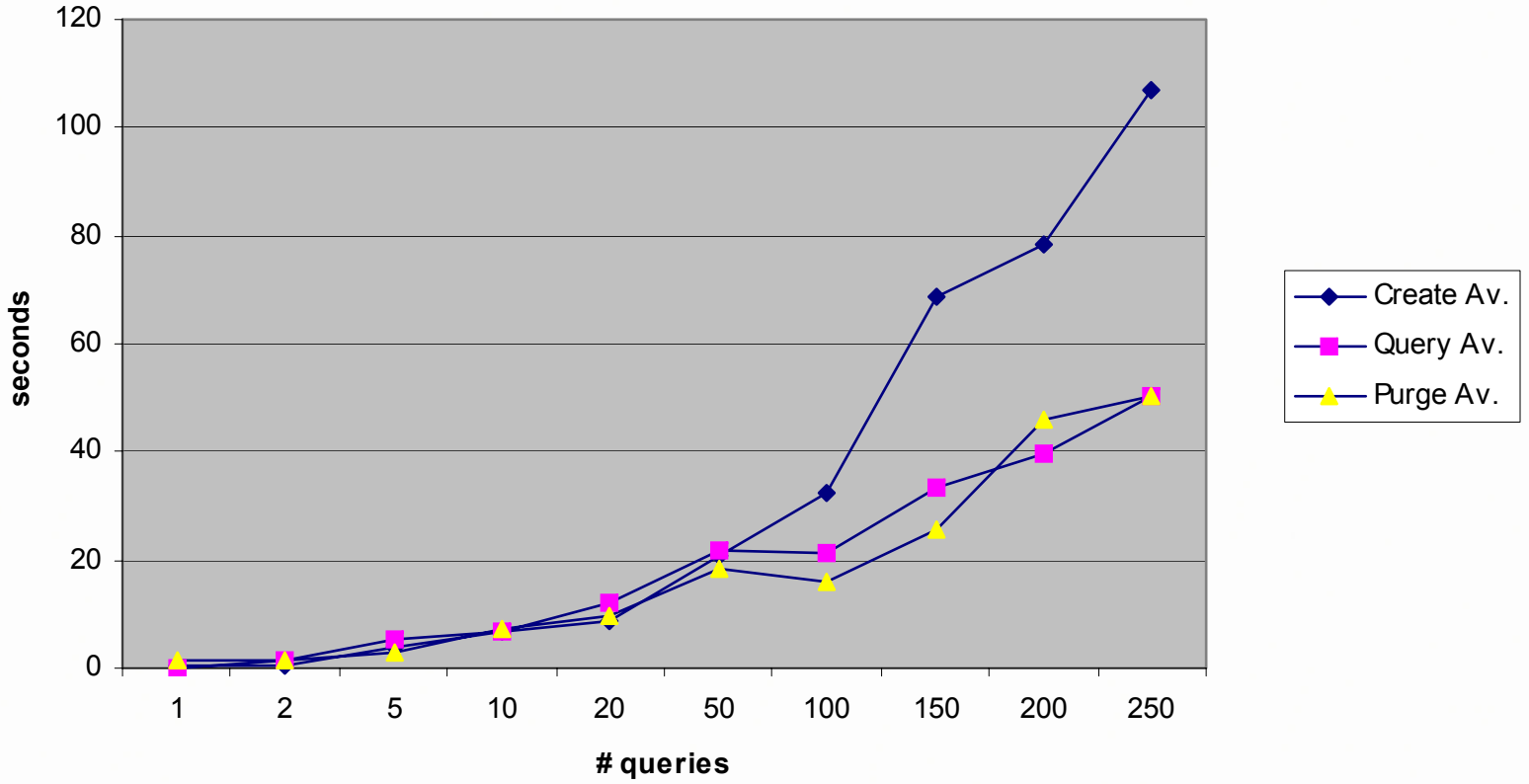

Test system
Production system
8 x 9940 tape drives
STK 9310
buxton
SunOS
ACSLS
dylan
AIX
Import/export
Tape devices
4 drives to each switch
Brocade FC switches
basil
AIX
test
dataserver
ADS_switch_1
ADS_Switch_2
SRB pathtape commands
ADS0CNTR
Redhat
counter
ADS0PT01
Redhat
pathtape
ADS0SB01
Redhat
SRB interface
cache
Logging
User pathtape
commands
cache
mchenry1
AIX
Test
flfsys
catalogue
ermintrude
AIX
dataserver
florence
AIX
dataserver
zebedee
AIX
dataserver
dougal
AIX
dataserver
brian
AIX
flfsys
admin
commands
create query
array4
array3
array2
array1
catalogue
All sysreq, vtp and ACSLS connections to dougal also apply to the other dataserver machines, but are left out for clarity
User
ADS
tape
ADS
sysreq
SRB
Inq; S commands; MySRB
Physical connection (FC/SCSI)
Sysreq udp command
User SRB command
STK ACSLS command
VTP data transfer
SRB data transfer
Thursday, 04 November 2004


Atlas Datastore Architecture
Robot Server (buxton)
CSL
ACSLS
ACSLS API
control info (mount/ dismount)
Tape Robot
LMU
IBM tape drive
STK tape drive
data
data
flfsys import/export commands (sysreq)
Catalogue Server (brian)
flfsys tape commands (sysreq)
recycling (+libflf)
flfdoback (+libflf)
flfdoexp (+libflf)
backend
flfsys (+libflf)
frontend
Copy A
Copy C
Copy B
read
read
read
catalogue data
cellmgr
Backup catalogue
flfqryoff (copy of flfsys code)
stats
Pathtape Server (rusty)
pathtape
servesys
short name (sysreq)
long name (sysreq)
flfsys farm commands (sysreq)
flfsys admin commands (sysreq)
?
administrators
flfsys user commands (sysreq)
(sysreq)
(sysreq)
SSI
flfstk
flfaio
flfscan
flfaio
cache disk
Farm Server
flfaio
tapeserv
data transfer (libvtp)
I/E Server (dylan)
importexport
libvtp
user program
vtp
vtp
tape
user
datastore (script)
SE
User Node
28 Feb 03 - 2
B Strong

Code Walkthrough

- Limits to growth
  - Catalogue able to store 1 million objects
  - In memory index currently limited to 700,000
  - Limited to 8 data servers
  - Object names limited to 8 char username and 6 char tape
  - Maximum file size limited to 160GB
  - Transfer limits: 8 writes per server, 3 per user
- All can be fixed by re-coding, although transfer limits are limited by hardware performance

Catalogue Manipulation

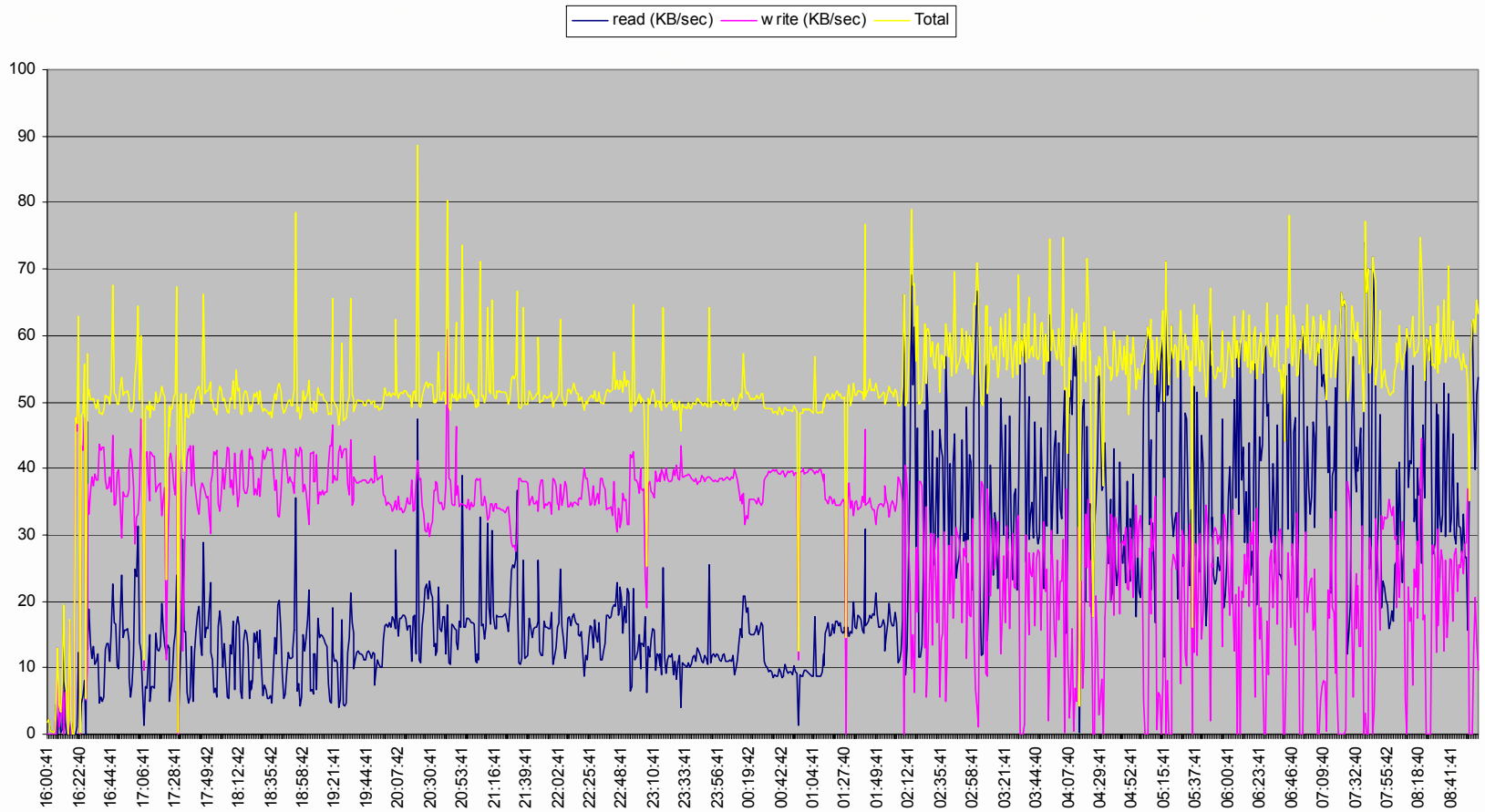


Catalogue Manipulation

- Performance breaks down under high load:
- Solutions:
  - Buy faster hardware (faster disk/CPU etc)
  - Replace catalogue by Oracle for high transaction processing and resiliance.

Write Performance

Single Server Test



Conclusions (Write)

- Server accepts data at about 40MB/s until cache fills up.
- Problem balancing write to cache against checksum followed by read from cache to tape.
- Aggregate read (putting out to tape) only 15MB/s until writing becomes throttled. Then read performance improves.
- Aggregate throughput to cache disk 50-60MB/s shared between 3 threads (write+read).
- Estimate suggests 60-80MB/s -> tape now. Buy more/faster disk and try again .

Timeline

- Expect to have UKLIGHT in time for March start, but schedule for end to end network availability needs to be finalised.
- At least 2 options for disk servers – we prefer to use production servers – but depends on timetable.