



System Architecture of Storage Arrays

x y r a t e x •

Notices

The information in this document is subject to change without notice.

While every effort has been made to ensure that all information in this document is accurate, Xyratex accepts no liability for any errors that may arise.

© 2007 Xyratex (the trading name of Xyratex Technology Limited). Registered Office: Langstone Road, Havant, Hampshire, PO9 1SA, England. Registered number 03134912.

No part of this document may be transmitted or copied in any form, or by any means, for any purpose, without the written permission of Xyratex.

Xyratex is a trademark of Xyratex Technology Limited. All other brand and product names are registered marks of their respective proprietors.

Tim Courtney

Issue 2.0 | October, 2007

Why the SBOD?

When the original Fibre Channel Arbitrated Loop (FC-AL) specification was drafted it was assumed that the disks used in a system would be connected using a simple daisy chain approach into a logical loop, this is the original Just a Bunch Of Disks (JBOD) system as shown in figure 1.

The JBOD has the big advantage that it is really simple to design and is not too complex to manufacture either. However, there is a problem with reliability. Once one of the devices on the loop fails or one of the links in the loop is interrupted there is no access to any of the data on any of the disks in the system. Looking at a fully populated loop there is a statistical mean time to failure of a disk of about a year and so failures are going to be seen.

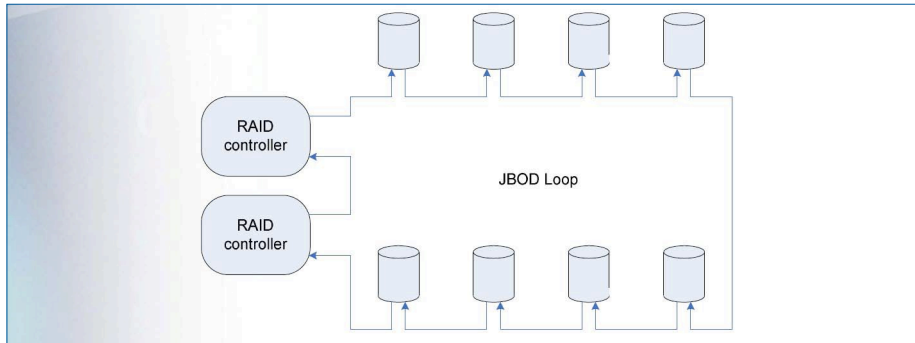


Figure 1: JBOD topology

To overcome this problem, centralised port by-pass switching was introduced to connect the devices in a hubbed arrangement such that they still formed a logical loop but also such that any device can be switched out of the loop using the port by-pass switches; this is shown in figure 2. Obviously this system is able to adapt around failures and data on a non-failed device can be accessed even when there is a fault on one of the other devices or links. If all the port by-pass circuits are collected together in a central location then the system is known as a Managed Bunch Of Disks (MBOD).

The MBOD has the definite advantage over the JBOD that there is a central device in the system that can be used for monitoring operating conditions as well as perform port-by-pass. This device can collect statistics about code violations, CRC errors and similar link characteristics. These can be relayed back to a user accessible enclosure device for collation and analysis.

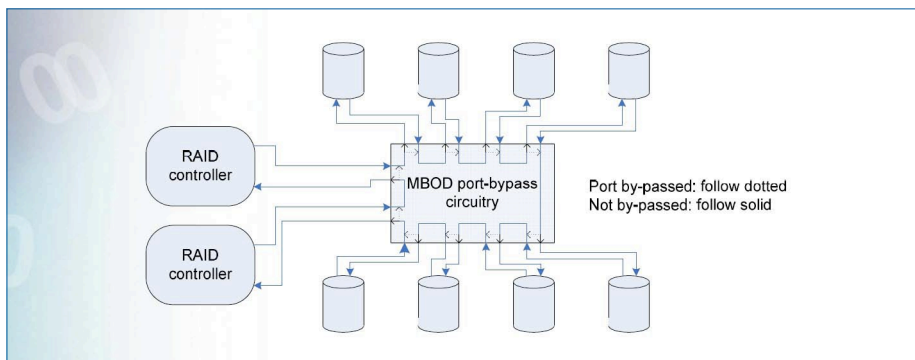


Figure 2: MBOD topology

The MBOD and JBOD both share a performance limitation stemming from their use of a logical loop topology to connect the devices in the FC-AL system. To allow access to the disks in the event of RAID controller failure two controllers are placed in the system. Since they are connected as a logical loop and considering that there can be only one active connection present on that loop, only one of these controllers can be actively communicating at any time. Therefore if controller A is sending data to disk X, controller B cannot send data to disk Y and so could be idle. In addition to this, any communication datagram must travel around the full logical loop in order to get from A to X and back to A. In the case of large loops and small datagrams this is a significant delay. With the MBOD the situation can be worse still as re-timing can be used as the signal traverses the port by-pass circuitry to remove any jitter problems that might be introduced. Re-timing introduces delay and this increases the loop delay experienced by datagrams

To overcome these limitations the Switched Bunch Of Disks was introduced in 2004. In this system the port by-pass circuitry of the MBOD is replaced by a more intelligent switch that is able to sustain multiple simultaneous links between pairs of switch ports. In the case of a system containing dual RAID controllers as described above, two simultaneous connections could exist, one between A and X the other between B and Y, thus removing the idle time from controller B. The SBOD system is shown in figure 3 with the default loop-back situation when no connection is established on the left and then the 2 simultaneous connection situation shown on the right.

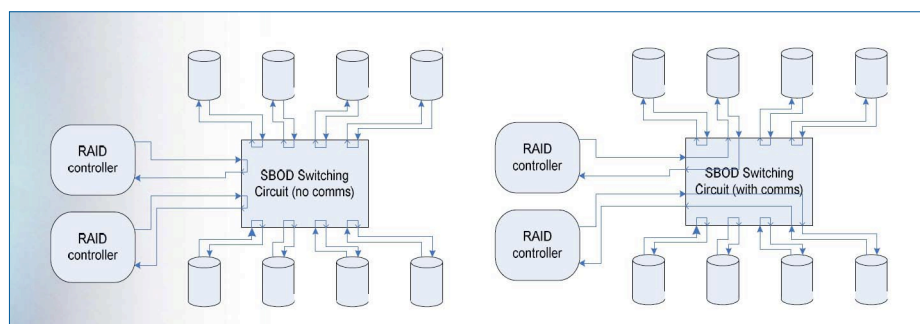


Figure 3: SBOD topology

To keep the costs low the switch unit is not a fully functional Fibre Channel Switched Fabric device, it is solely designed for use within an FC-AL system. The switch is invisible to the FC-AL protocol but intercepts some of the basic communications protocols to ensure that all the attached devices are serviced in a fair manner. The switch also has a responsibility to ensure that initialisation of the system occurs correctly.

In the initialisation phase there remain the 126 addresses available in a JBOD FC-AL, the use of a switch does not extend the address space available, this is dissimilar to the operation of a switched fabric device. Since the switch must operate by connecting devices on-demand it must be responsible for ensuring that an alternative to the FC-AL fairness algorithm is employed to prevent starvation of access.

One such SBOD technology that is used by Xyratex is the Emulex InSpeed SOC320 20-port switching technology.

When the switch makes a connection between two of its ports it creates a virtual loop consisting of the devices attached to each port of the switch. Unlike in the JBOD and MBOD, the other devices in the system do not add

to the delay experienced by the virtual loop. Whilst the introduction of a switch does add cross-switch latency to the delay, this will normally be significantly lower than the delay removed by removing the need to traverse every node in the system.

Looking at a system with 16 disks in an enclosure linked to two RAID controllers there is obviously one device per port of the switch. Therefore when a connection is made a virtual loop is created with just 2 devices in the loop rather than the 18 devices that are there for the JBOD and MBOD. For this MBOD there are 18 hub delays and 18 port delays in the loop, for the SBOD there are just 2 port delays and 2 cross-switch delays in the system, this is a significant latency reduction, coupling this with the doubling of available bandwidth coming from having two simultaneous connections we see a significant performance improvement with the SBOD.

At Xyratex we have modelled disk drives, RAID controllers and switches to determine their relative performance when confronted with particular I/O traffic profiles. When a sequential stream is used the SBOD outperforms the MBOD by at least a factor 2 for all of the configurations tested. These included testing large and small systems (8 to 48 disks in the loop) and large and small access sizes (512 bytes to 64kbytes per access).

The worst case system for the SBOD is when there are only a small number of disks in the system since the delay removed through use of the SBOD rather than an MBOD is minimized and the switch latency remains constant. The worst case access size for the SBOD is large accesses as the data bandwidth outweighs command and access bandwidth used. For this case we still see that the SBOD outperforms the MBOD by a factor 2-04.

When the situation is reversed and a large number of disks are used then the SBOD can outperform the MBOD by a factor 4. These sequential I/O requests are similar in nature to the requests produced by a sequential run of the IOMeter benchmark and have been compared with physical IOMeter results to verify the validity of our modelling technique.

Since sequential data is not the end of the story, we have also run tests using a mix of I/O request streams. We have used the command profiles published by the Storage Performance Council as the SPC Benchmark-1™ benchmark. The overall command profile for this benchmark includes both sequential and random I/O streams and also contains I/O streams with a self-similar access profile that has the same level of randomness no matter what level it is viewed at. Further information can be obtained from the SPC-1 specification documentation.

Our results show that the performance of a system in the light of random I/O requests is mainly limited by the performance of the disks in the system but that there are still times when the bandwidth in the system is a limiting factor. This means that the SBOD will always perform as well as or better than the MBOD, in some circumstances this can be a factor 2 improvement.

The trends observed in these simple tests show that as the number of disks in the system increases, available bandwidth becomes an ever more important limiting factor and so the SBOD will further outperform a similar MBOD. With the move away from 3.5 to 2.5 inch drives, more pack into a given enclosure and so the SBOD advantage becomes increasingly obvious.

But what about the relative costings? We know that the SBOD is able to far outperform the MBOD but this must come at a sensible cost. In the system the bulk of the cost is built from the disk drives and so we see that the use of an SBOD rather than an MBOD introduces in the order of 5% cost increment. This therefore represents a good trade-off in terms of large increase in performance for small increase in cost.

UK HQ

Langstone Road
Havant
Hampshire PO9 1SA
United Kingdom

T +44(0)23 9249 6000

F +44(0)23 9249 2284

www.xyratex.com



ISO 14001: 2004 Cert No. EMS91560

©2007 Xyratex (The trading name of Xyratex Technology Limited). Registered in England & Wales. Company no: 03134912. Registered Office: Langstone Road, Havant, Hampshire PO9 1SA, England. The information given in this brochure is for marketing purposes and is not intended to be a specification nor to provide the basis for a warranty. The products and their details are subject to change. For a detailed specification or if you need to meet a specific requirement please contact Xyratex.

