Switched FC-AL: An Arbitrated Loop Attachment for Fibre Channel Switches

Vishal Sinha
sinha@cs.umn.edu
Department of Computer Science and Engineering
University of Minnesota Minneapolis, MN 55455
7481 Lee Hwy, Falls Church, VA 22042
Tel: +1-703-698-9749
Fax: +1-703-245-4545

David H. C. Du
du@cs.umn.edu
Distributed Multimedia Research Center
Computer Science and Engineering Department
University of Minnesota, Minneapolis, MN 55455
Tel: +1-612-626-7523

Introduction

Fibre Channel Arbitrated Loop (FC-AL) is a loop architecture. It can support any combination of hosts and storage devices, up to a maximum of 127 nodes. With 100 Mbytes/sec bandwidth, the loop structure enables rapid exchange of data between the devices.

A major problem with the existing FC-AL is scalability. Theoretically, a maximum of 127 devices can be attached to a loop. However, past research has indicated that the loop can be saturated by as few as 32 devices. This number is expected to decrease rapidly, as the disks become faster and faster. To overcome this limitation, we investigated the possibility of extending the existing FC-AL protocol.

Switched FC-AL protocol aims at using a switch to accommodate Arbitrated Loop devices without requiring a true fabric connection. Ideally, both fabric and loop devices should be able to share a single switch and enjoy their own 100Mbytes/s segments on each switch port, and yet be logically grouped together for high-speed transactions.

We designed and simulated two different approaches to implement the switched FC-AL protocol. In the first approach, which is a circuit switched approach, the initiator establishes a connection with the target on the remote loop before initiating the data transfer. This approach is useful for applications that have hard Quality of Service (QOS) requirements. In the second approach, which is a packet switched approach, the exchange of data takes place in hops. The initiator needs to establish a local connection only. Each of these approaches has it's own pros and cons. The objective of our research is to investigate the design issues and trade-offs of both these approaches.

Switched FC-AL design:

The two design considerations are the circuit switched approach, in which the Initiator wins arbitration on it's local loop and establishes a connection with the Target on the remote loop via the switch. The second approach is based on the packet switched concept, in which the Initiator wins arbitration on its local loop and transmits data to the Switch Connected Node (SCN) attached to the switch. The SCN buffers the data and inturn arbitrates on the remote loop to deliver the data to the Target. A detailed description of the two approaches are presented below.

Circuit switched approach:

Figure 1 illustrates this approach. The Initiator X, sends out an arbitrate primitive ARB(X) to gain access to the loop. Upon winning the arbitration, by receiving the ARB(X) primitive back, it sends out the OPEN(Y) primitive. If the target Y is not on the local loop, the Switch Connected Node, SCN(L_1) intercepts the OPN(Y) primitive and sends a request to the switch to access the remote loop. If the remote loop is not busy, the switch sends an acknowledgement back to the SCN(L_1) and a connection request is sent to SCN(L_n) via the switch. If the remote loop is busy, the switch sends a negative acknowledgement back to the SCN(L_1) and a connection tear down request is sent to Initiator X by SCN(L_n). Else, the SCN(L_n) becomes the initiator of this remote request on the local loop Ln. It next proceeds to establish a connection with the Target Y. If successful, the Target Y responds with R_RDY primitive which gets routed all the way back to the Initiator X . At this stage the connection is established and all the data transfer takes place directly between the Initiator and the Target using the usual FC-AL protocol. Tear down of the connection can be initiated by either the Initiator or the Target by sending the CLS primitive.

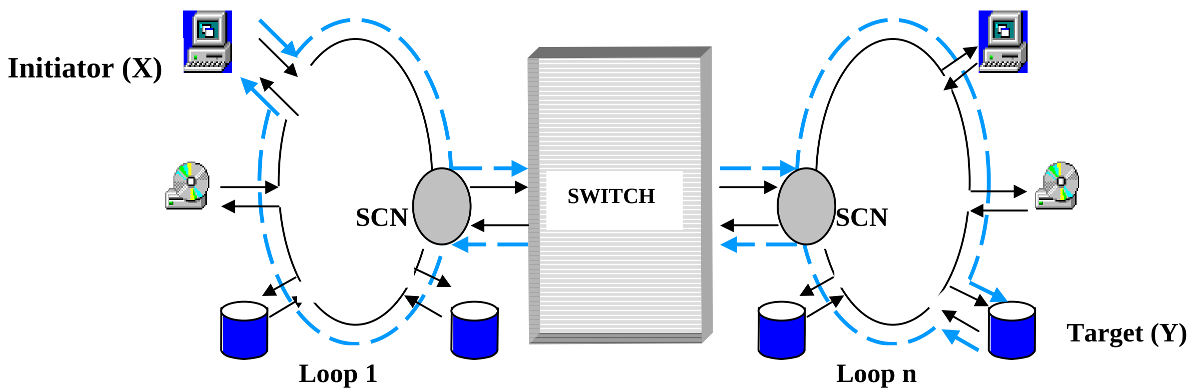


Figure 1: The simulation model for the circuit Switching based approach.

Packet switched approach :

It follows the store-and-forward technique of data transfer. In this approach, the Initiator X, sends out the arbitrate primitive ARB(X) to gain access to the local loop. On winning the arbitration, it send out the OPEN(Y) primitive. If the target Y is not on the local loop, the Switch Connected Node, SCN(L_1) intercepts the OPN (Y) primitive and sends a request to the Switch to access the remote loop.

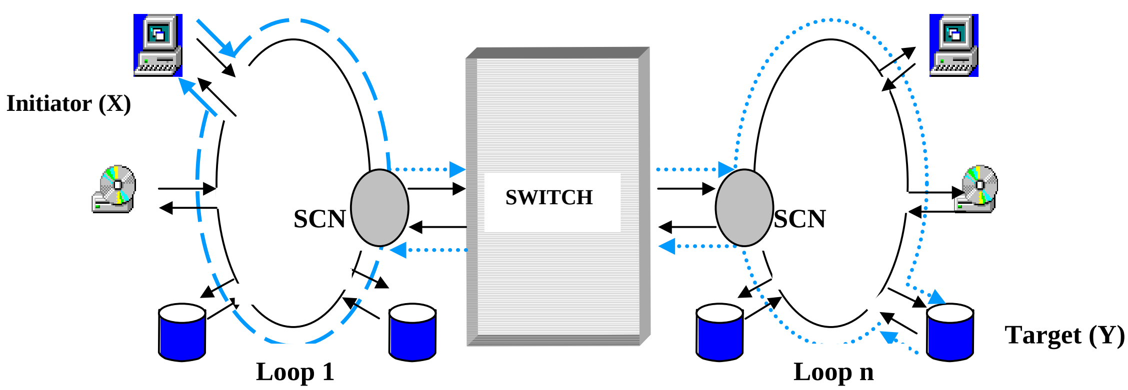


Figure 2: The simulation model of the packet switching based approach

At the same time, it sends a local acknowledgement back to the Initiator in terms of R_RDY. At this stage, SCN (L_1) becomes the local target for the global request in loop L1 and all the data transfer takes place directly between the Initiator X and the SCN(L_1) using the FC-AL protocol. If the remote loop is not busy, the switch sends an acknowledgement back to the SCN (L_1) and a connection request is sent to SCN(L_n) via the switch. The SCN (L_n) next proceeds to establish a connection with the Target Y. If successful, the Target Y responds with R_RDY primitive which gets routed back to SCN

(L_1). Now data transfer takes place directly between the SCN (L_1) and Target Y. The advantage of this approach lies in the fact that both the stages of data transfer can take place asynchronously there by providing some means of parallelism. Also, the acknowledgement sent is segment-by-segment, therefore a negative acknowledgement from the switch does not need to tear down the connection.

Simulation Model

The detailed simulation model is shown in Figure 3. Each loop consists of one host, one Switch Connected Node (SCN)[1] and a variable number of disks. The number of disks attached to a loop is one of the parameters of our study. SCN has been given the highest priority and does not implement the fairness algorithm. All other nodes, including the host are assigned priority randomly. Table 1 shows the values of the parameters used for the simulation.

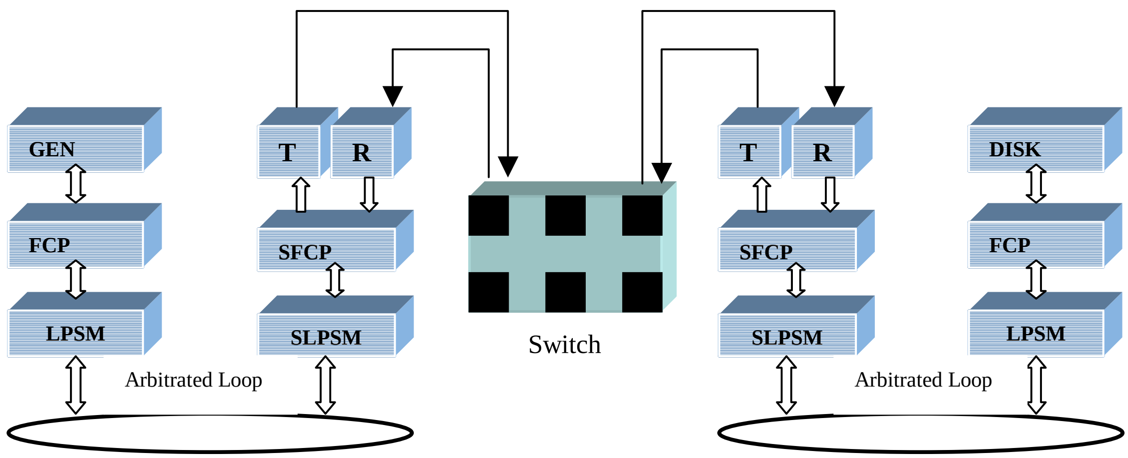


Figure 3. The detailed simulation model used for both the schemes.

Parameters	Default Values	Description
Propagation delay	3.5ns	The propagation delay between two nodes
Per node delay	6 words time	The delay for the interface to forward the frame.
Bandwidth	100 Mbytes/sec	The FC-AL link bandwidth
Maximum outstanding commands	8 times the number of disks on the loop.	The maximum number of commands allowed at the initiator. Used for the stress test.

Table 1: Switched FC-AL simulation model parameters

Simulation Results

We studied the performance of switched FC-AL under two different load scenarios namely, light load and heavy load. In the first scenario, the system is under extremely light load. The host generates a request and waits till it gets serviced, before generating the next request. Thus, the storage subsystem has only one outstanding command at a time. This allowed us to compare the total latency of both the approaches as the percentage of global traffic is varied. It also gave us an idea of the latency

[1] SCN is the interface between the loop and the switch. It's like a normal node with extra buffers.

overhead involved. We also studied the effect of the number of loops on the total latency of the system.

In the second scenario (heavy load), we studied the scalability of the existing FC-AL protocol and compared it with that of switched FC_AL. We heavily loaded the loop by setting the number of outstanding commands per disk to 8, and increased the number of disks per loop. The purpose of this test was to discover the number of disks (all the three types) needed to saturate the loop and to find out how the two approaches of switched FC_AL perform near saturated conditions.

Performance under Light Load

In the light load scenario, the target disk and loop for any command is chosen randomly according to a uniformly distributed random number. We use 64KB as the request size to represent a normal request. We want to see how the loop latency and total latency change as the percentage of local and global traffic changes and also the effect of number of loops on the total and loop latency of the storage subsystem.

Percentage Global Traffic	Maximum circuit_1 establish time	Maximum circuit_2 establish time
0	0.6875	0.697754
10	1.584229	1.805176
20	2.085144	2.28125
30	2.462872	2.783691
40	2.754395	2.988281
50	3.386963	3.453369

Table 5: Maximum circuit establish time for both the phases of circuit switched approach.

Table 5 shows the effect of increasing global traffic on the circuit establish time for circuit switched approach when the number of disks on each loop is 8 and the request size is 64KB. Evidently, the circuit establish time[2] increases exponentially with the percentage of global traffic. To understand this behavior, we counted the number of attempts needed by an initiator to establish a circuit. Graph 2 shows the observed behavior.

	Circuit switched approach				Packet switched approach			
%global traffic	Disk time	Loop time	Total time	# of re attempts	Disk time	Loop time	Total time	# of attempts
0	4.86336	0.822962	5.686322	0	4.86336	0.822964	5.686324	0
10	4.969455	0.851242	5.820697	1.964646	4.917598	0.89883	5.816428	1.006711
20	5.078583	0.880659	5.959242	4.49596	5.010529	0.993514	6.004043	1.013423
30	5.091376	0.926795	6.018171	10.15657	5.053116	1.123394	6.17651	1.020134
40	5.112677	0.985283	6.09796	18.69596	5.067917	1.251087	6.319004	1.026846
50	5.149949	1.050099	6.200048	28.51717	5.098782	1.379847	6.478629	1.033557

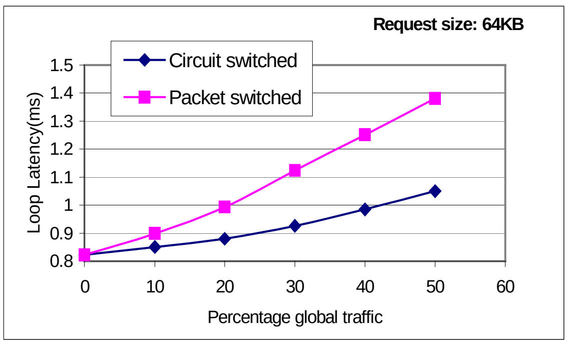
Table 6: Average latencies of both approaches with varying percentage of global traffic.

latency. The disk access time consists of the command queuing time, disk seek time, disk rotation latency and the data transfer time. The loop latency consists of the time to win

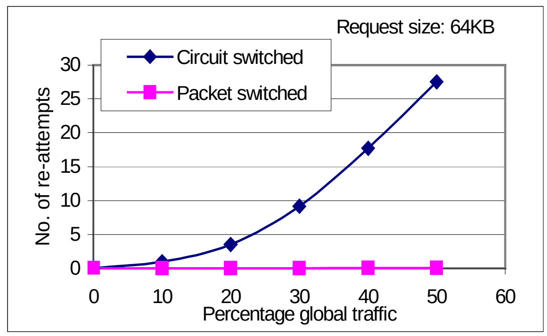
[2] Circuit establish time is defined as the total time an initiator takes to win arbitration, open the target and get the R_RDY back.

the arbitration and the transmission time of control and data frames. In order to avoid concurrent data fetch and transfer feature of FC-AL, we restricted our request size to 64KB.

From Table 6, we found that the disk access time dominates the total time for any percentage of global traffic. This is because the bandwidth of FC-AL loop is 100 MB/s, and it takes just 0.64 ms to transfer the 64KB data, whereas the average seek time alone of the disk is 2.475 ms. This clearly indicates the need for dividing a transaction into two phases. In the first phase, a host can send the request to the disk and then relinquish control over the loop. In the second phase, the disk, after fetching the data, can transfer the data back to the host, thereby allowing multiple I/O operations to fully utilize available bandwidth of FC-AL at the same time.



Graph 1: Comparison of loop latency of both the approaches.



Graph 2: Comparison of "number of re-attempts needed " for both the approaches.

We also observe that both disk access time and loop latency increase as the percentage of global traffic increases. The increase in disk access time is primarily due to the increase in command queuing time. As the percentage of global traffic increases, the probability of two requests being generated for a particular disk increases and this contributes to the increased queuing time.

Graph 1 shows the loop latency of both the approaches. The rapid increase in loop latency, in the case of circuit switched approach, is due to the exponential rise in number of attempts needed to establish the circuit. In the packet switched approach, the increase though exponential, is very small, and is mainly contributed by the additional store and forwarding time of the frames at the SCN. In the light load scenario, the circuit switched approach exhibits better total request servicing time than the packet switched approach.

We also investigated the effect of increase in the number of loops on the total request service time. The percentage of global traffic was set to 40% and request size to 64KB. From table 7, we observe that as the number of loops increases, the number of re-attempts needed to establish a connection increases more rapidly for the circuit switched approach than the packet switched approach. So the increase in total latency is more for the former than in the latter.

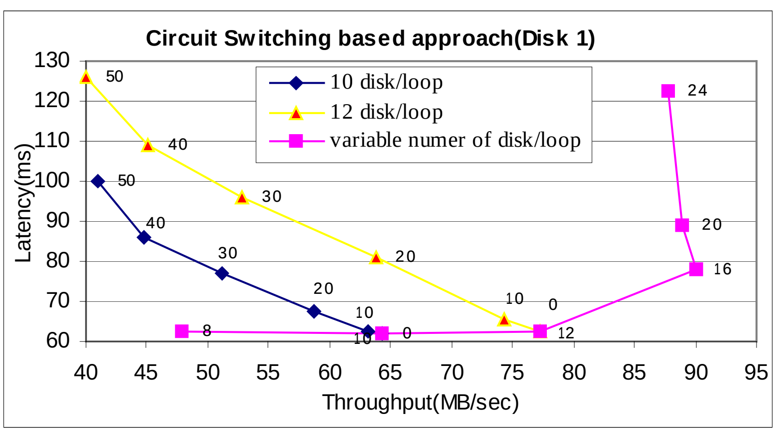
	Circuit switched approach				Packet switched approach			
No. of loops	**Disk time**	**Loop time**	**Total time**	**# of attempts**	**Disk time**	**Loop time**	**Total time**	**# of attempts**
2	**5.029166**	**0.980776**	**6.019942**	**45**	**5.00825**	**1.210549**	**6.298799**	**26**
4	**5.122677**	**0.985283**	**6.107959**	**17519**	**5.067917**	**1.241087**	**6.319005**	**29**
8	**5.12838**	**1.014671**	**6.143051**	**42740**	**5.107891**	**1.272409**	**6.372301**	**49**

Table 7: Average latencies of both approaches with number of loops.

Performance under Heavy Load

With the advancement in disk technology, fewer disks are now capable of saturating an arbitrated loop. To investigate this scenario, we studied the performance of FC-AL with three different types of disks. At the beginning of the simulation, the commands were generated one after the other until the number of outstanding commands reached a pre-defined upper limit. This value was set to be the number of disk on the loop, multiplied by 8. We maintained the same number of outstanding commands throughout the entire simulation. When one command is completed, a new command is generated right away. We used two uniformly distributed random number sequences, which are generated with different seeds. They are used to determine the target disk and loop for any command.

The simulation results of circuit switched and packet switched approaches for disk1 are shown in graph 1 and 2 respectively. The graphs depict three series. The first one, from the right, shows the number of disks needed to saturate the loop with a given 64KB request size while the next two series show the throughput vs. latency behavior of the two approaches before saturation. Each of the data points in a graph is labeled with a number indicating either the number of disks per loop (for first one) or the percentage of global traffic (for series 2 and 3). The percentage of global traffic is varied from 0 to 50, in steps of 10.

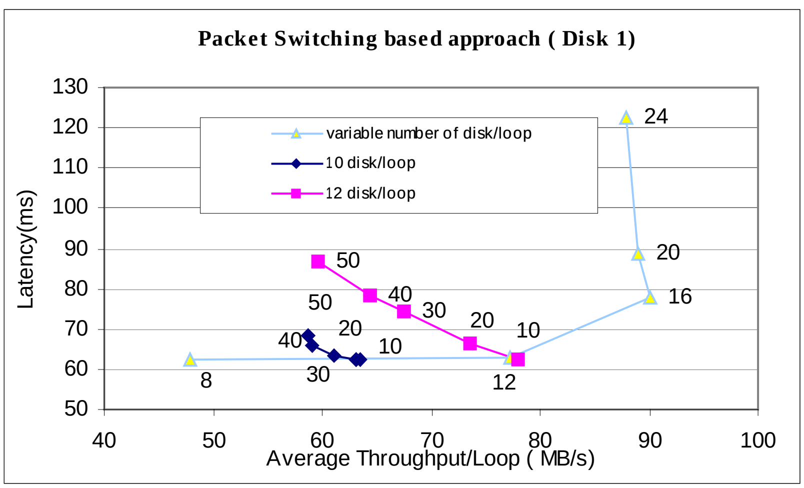


Graph 3: Throughput vs. Latency behavior of the circuit switching based approach.

From graph 3, we can see that the FC-AL loop starts getting saturated with 16 disks of type 1. The maximum throughput obtained is 90.04 Mbytes/sec. After saturation, with more disks attached to the loop, longer latency is observed.

We investigated the performances of both the circuit switched and the packet switched approaches, when the number of disks on the loop is 12 and 10. From graph 3 and 4, we observe that as the percentage of global traffic increases, the throughout decreases while the latency increases. This is mostly due to the following two factors: increased arbitration overhead and blocking. Since we need to win arbitration on both the loops for each global request, the arbitration time is doubled. The second factor contributing to this delay is the number of blocked requests. In case of circuit switched approach, a request is blocked if the remote loop is busy servicing another global request while in case of packet switched approach, a request is blocked for the duration in which data is transferred between the host and the target SCN. Since the blocking time is much longer for the circuit switched approach, it incurs steep rise in latency as the percentage of global request increases. As a result, the packet switched approach shows better throughput and latency characteristics than circuit switched approach.

We also observe that the latency of the packet switched approach for 10 disk/loop is 68.52 ms while that of 12 disk/loop is 86.52ms. This big difference is due to the fact that as the number of disk on a loop increases, not only does the propagation delay and per node delay increase but so does the number potential initiators on the loop[3]. Hence, the number of L-Ports arbitrating at a time increases forcing the arbitration latency to increase.



Graph 4: Throughput vs. Latency behavior of the packet switched approach.

Conclusion and Future Work

In this paper, we presented some initial results of our study on the performance of the circuit and packet switched approaches to implement switched FC-AL. The light load test indicated that even with vast improvement in the disk technology, the disk is still a bottleneck and we need to multiplex several I/O operations to fully utilize the available

[3] In our implementation, a disk is an initiator in the second phase of data transfer.

bandwidth. We also see that with an increase in the percentage of global traffic the number of attempts needed to establish a circuit increases rapidly. This is the main limitation of the circuit switched approach. Packet switched approach, on the other hand needs very few attempts to send the request to the remote loop but it's main overhead is the store and forwarding time. Also, with an increase in the number of loops connected to the switch, there is a marginal increase in the overall latency for both the approaches.

The scalability issue of the switched FC-AL protocol was studied using the heavy load test. It is evident that with faster disk, the number of disks needed to saturate the loop is very low. Hence, to use large numbers of disks and still have better throughput vs. latency behavior, the switched topology is a good option. For both the approaches, the aggregate throughput for as much as 50% of global traffic was quite high but clearly the packet switched approach outperformed the circuit switched approach under our set of assumptions.

We next plan to investigate the effect of variable number of hosts per loop on the performance of both approaches. Other related issues like larger request size, inverse priority of SCN and higher bandwidth of the loop also need to be studied. Finally, a comparison of switched FC-AL and FC-AL 3 will be worth doing.

References:

[1] ANSI X3.272-199x, "Fibre Channel - Arbitrated Loop (FC-AL), Revision 4.5", American National Standard Institute, Inc., June 1, 1996.

[2] ANSI X3.272-199x, "Fibre Channel - Arbitrate d Loop (F C-AL-2), Revision 5.1" , American National Standard Institute, Inc., March 26, 1996.

[3] ANSI X3.230-1994, Fibre Channel Physic al and Signaling Interfac e (F C-PH)" American National Standard Institute, Inc., 1994.

[4] David H.C. Du, Jenwei Hsieh, Taisheng Chang, Yuewei Wang and Sangyup Shim, Performance Study of Emerging Serial Storage Interfaces: Serial Storage Architecture (SSA) and Fibre Channel - Arbitrated Loop (F C-AL)", Submitted .

[5] Chris Ruemmler and John Wilkes , "An introduction to disk drive modeling" IEEE computer, march 1994, pp. 17-28.

[6] Gadzoox Network's white paper on the Denali Fibre Channel Switch. http://www.gadzoox.com/links/products/switch/denali.html

[7] Vixel's white paper "Arbitrated Loop Attachment for Fibre Channel Switches" http://www.vixel.com/whitepapers/wpaper2.html, July 1998