X3T9.2/91-99

TO: X3T9.3 FCS Working Group and X3T9.2 SCSI Committee
FROM: Gary R. Stephens, IBM
Date: June 24, 1991

SUBJECT: Fabric Loop Meeting, June 18, 1991

The meeting was convened at 5 PM at the Sofitel Hotel, Minneapolis, MN. An attendance list was passed around. The list is attached. This is the third meeting of this special subject working group. The first meeting was held in St. Petersburg. The name is being changed to the Fabric Loop to indicate its focus on the Seagate loop proposal rather than a general low cost distributed fabric.

An agenda was displayed and modified as shown below:

1. Introductions
2. Develop Meeting Agenda
3. Mission Statement
4. Review May 15, 1991 Minutes
5. Physical Layer - Bruce Johnson, Seagate
6. Bad Frame Management - Horst Truestedt, IBM
   A. Hop Count Primitive
   B. Special Frame Addressing
7. Bad Frame Management - Giles Frazier, IBM
   A. Expiration Timer Header
8. Performance Analysis - Marti Miller, NCR.
9. Self Configuring Fabric Elements - Ken Hardwick, Network Systems
10. Logical Layer
11. Frame Size - Minimum
12. Future Meeting Schedule

The minutes from Harrisburg were distributed. No changes were noted.

The mission statement was presented by Wayne Sanderson which indicates a focus on the fiber loop. The distinction between the Seagate proposal and the CANSTAR distributed fabric proposal needs to be discussed. This may lead to a split into two groups based on this distinction. This is scheduled for the July meeting.

Bruce Johnson, Seagate, gave his updated presentation for the Fiber Channel Loop proposal. Bruce continued his education on the proposal. The proposed frame size is 2148 bytes which matches the maximum permitted by Fiber Channel. The general rules for L_Port management were presented (Slide 1). Bruce then displayed two possible implementations: one a double loop; and a string array arranged to give the shortest possible links between the FL_Port and the first and second L_Ports (Slides 2-3).

Bruce then discussed his proposal for hot plug connections using some industry available parts and some of his own ideas. No decisions were made in this area (Slides 4-5).

Bruce described his equalizer function and passed around a sample functioning circuit. Test results for various frequencies from 50 MHz to 1000 MHz were shown (Slides 6-7).

The next part of Bruce's presentation focused on the transceiver/ FC protocol chip interface. He showed two potential implementations. (See Slides 8a, 8b and 9.)

Slides 10a, 10b and 11-13 focused on costs. Slide 11 indicated that considering LSI costs alone in comparing to other interfaces was not the best measure of this proposal. Rather, total costs for LSI, board space, power, connectors, cabling and MTBF improvement should be considered in the comparison.

369

Horst Truestedt, IBM, opened his presentation with new information that it is possible to "decrement" the Hop Count as a 10-bit construct rather than decoding, subtracting and re-encoding the 10-bit field.

Horst then made a presentation comparing the proposed Hop Count primitive to a proposal to place duplicate addresses bytes within the S_ID and D_ID fields. It was noted that since the two bytes were within the same word, a corruption in the duplicate address case would probably corrupt both addresses and therefore leave a circulating frame on the loop. Sufficient problems arose in discussion to tend back toward the original Hop Count proposal. It was noted that the Hop Count has been successfully implemented in other networks. (See Slides 1-6.)

Giles Frazier provides a third proposal for handling frames circulating too long. The proposal entails use of the optional Expiration Security Header in all frames. This would require all communication to contain such a header from whatever source. This may be difficult to mandate. His premise is that the Hop Count is not FC-PH compliant. However, since the Hop Count is a fabric frame, and never exits the fabric, it is defined within the standard. No change to N_ports is required to support this frame since true N_Pot devices do not attach to the Fabric Loop. (See Slides 1-8.)

Marti Miller, NCR, gave a short presentation on the preliminary results of a performance model he is constructing. He solicited inputs for other items and was deluged by things to measure. Marti will provide formal charts at a later meeting.

Attached is an independent analysis from Jeane Chen, IBM, which was presented in an FDDI meeting later in the week. It is included for information. It is possible that Marti and Jeane can find some synergism on a common model.

A brief discussion occurred on self-configuration. No significant items resulted from that discussion.

Item 10 was bypassed in the interest of time. Item 11 appeared covered in Bruce Johnson's proposal that the maximum frame length be the design point for the Fabric Loop.

Future meetings were discussed. Dedicated meetings were postponed at this time to see how the interest and need progresses. It was agreed to continue meetings in the evenings during the plenary and work group weeks. The idea was to pick evenings which were likely to overlap the SCSI and Fiber Channel attendees. Wednesday's for work group weeks and Tuesdays for plenary weeks seemed appropriate for now. The time selected was 5-8 PM.

The next meeting, then, is scheduled for Wednesday, July 17, 1991, 5 PM to 8 PM in Valley Forge(King of Prussia).

The proposed agenda is as follows:

Distinction between the Loop and a distributed fabric

Are these separate topics for separate interest groups?

CANSTAR Cost Presentation on the Distributed Fabric Sub-Element

A discussion of the FC Loop vs a Standard N_port.

Bruce Johnson to continue to educate and expose the group to the FC Loop.

Other items may be added by contacting Wayne Sanderson.

The meeting was adjourned.

Fabric Loop Meeting Attendees, June 18,991

* New Attendee

I = IPI
S = SCSI
F = FCS
H = HIPPI

NAME		COMPANY	INTEREST
John Aguilar	*	CDC	I
Dal Allan		ENDL	SIF
K. Annamalai		Gazelle	SIF
Bill Burr		NIST	?
Kurt Chan		HP	SF
K. C. Chennappan	*	IBM	SIF
Chris A. Ciufo	*	AMD	SF
Roger Cummings	*	StorageTek	SIF
Giles Frazier		IBM	SIF
Marc Friedmann	*	AMCC	F
Edward A. Gardner	*	DEC	S
Doug A. Grieg	*	Tandem	F
Ken Hardwick	*	Network Systems	HF
George Hopkins	*	Cray Research	HIF
Gerald Houlder	*	Seagate	S
Brian Johnson	*	Seagate	S
Soon Kang	*	CDC	IFH
Nobukazu Kume	*	Furukawa	F
Larry Lamers	*	MAXTOR	S
John Lohmeyer		NCR	S
Jim Luttrull		Fujitsu America	SI
Kumar Malavalli		CANSTAR	SIFH
Gerald Marazas		IBM	SF
Bill Medlinski	*	Panasonic	S
Marti Miller		NCR	SF
Mike Miller		Seagate	SI
Gene Milligan		Seagate	SI
Charles Monia (sp?)	*	DEC	SF
Chris Nieves	*	AST Research	S
Vit Novak	*	Sun	S
Ken Ocheltree	*	IBM	F
Tom Palkert	*	AMCC	?
Vit Patel		Seagate	SI
George Penokie	*	IBM	S
Grover Phillips		NCR	S
Jerry Radcliffe	*	IBM	F
Matt Rooke		IBM	SI
Paul Scott	*	Cypress Semiconductor	F
Jim Schuessler	*	NSC	SF
Jim Smith		Tandem	SIF
Joe Soriano (sp?)	*	NSC	F
David Steele		NCR	SF
Gary Stephens		IBM	SF

371

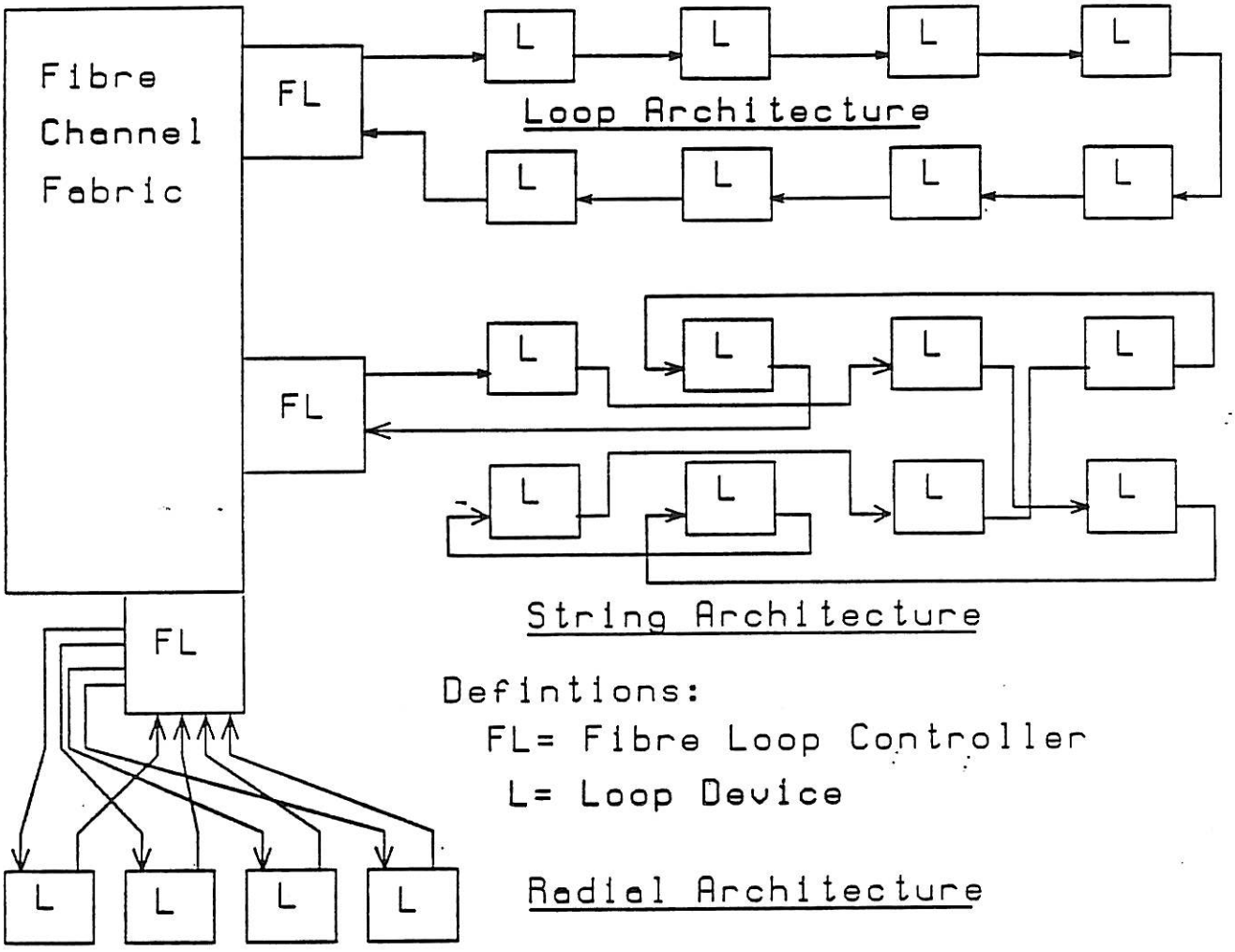
Fabric Loop Meeting Attendees, June 18,991

* New Attendee

I = IPI
S = SCSI
F = FCS
H = HIPPI

NAME	COMPANY	INTEREST
Arlan Stone	UNISYS	SF
Pete Tobias	* Tandem	S
Don Tolmie	LANL	F
Horst Truestedt	* IBM	SIFH
Alfonso Vaca	* Condumex	S
Lynn Whitfield	* Sun Micro	SIF
Carl Zeitler	IBM	SIF

END OF MINUTES



- Packetized Frames Up to 2148 Bytes Long.
- Bad Frame Management by a Decrementing Hop Count.
- Transmission When Idle Received & Input FIFO Empty.
- Frame Removed Upon Destination Address Match.
  -Bad Frame Removed When Hop Count is Invalid or Zero.
- Frame Passed On Without Destination Address Match.
  -Hop Count Decremented Each Time Frame is Passed On.

Fibre Channel Loop Overview

fcloop
6-6-91

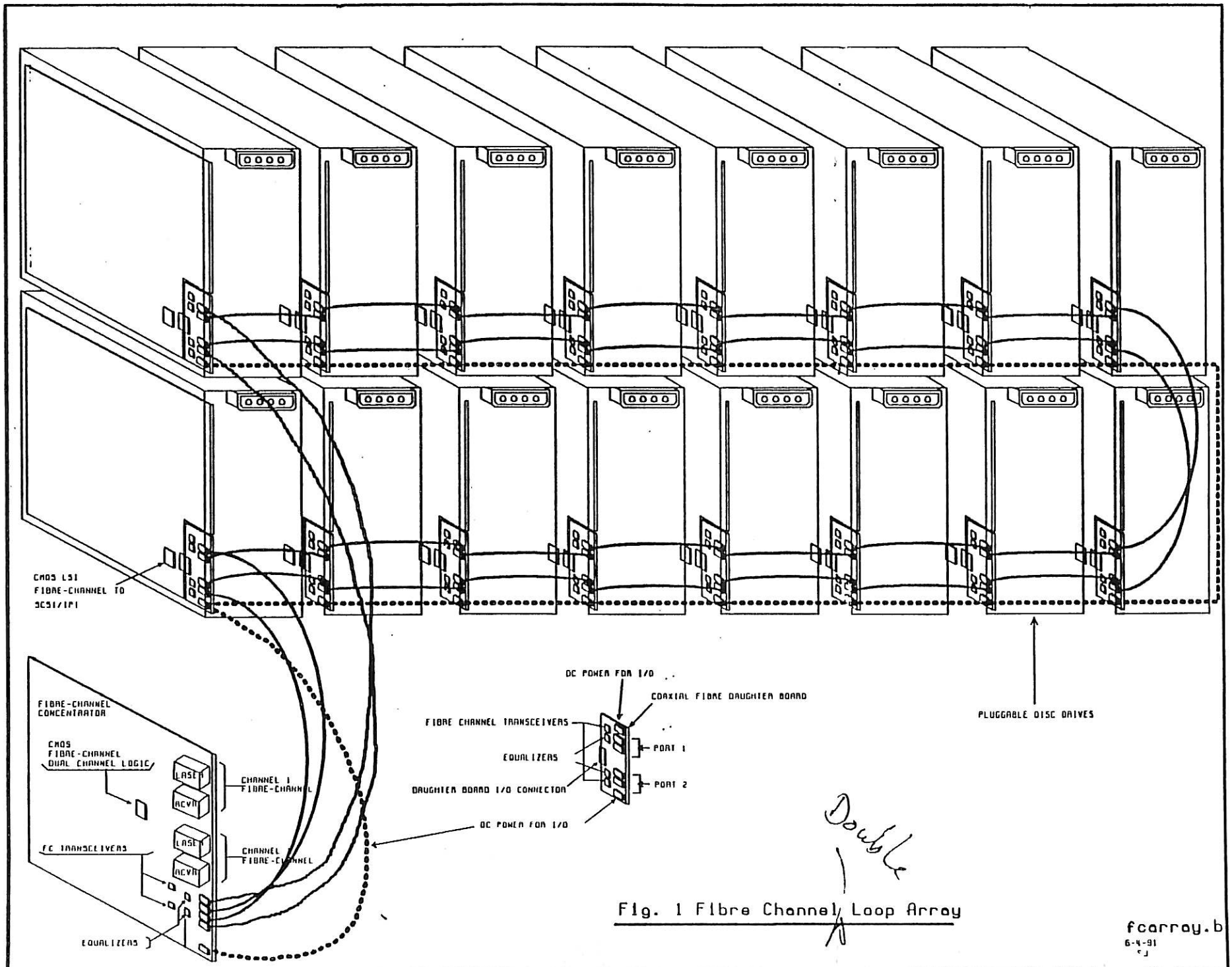


Fig. 1 Fibre Channel Loop Array

fcarray.b
6-4-91

Double

379

BJ-2

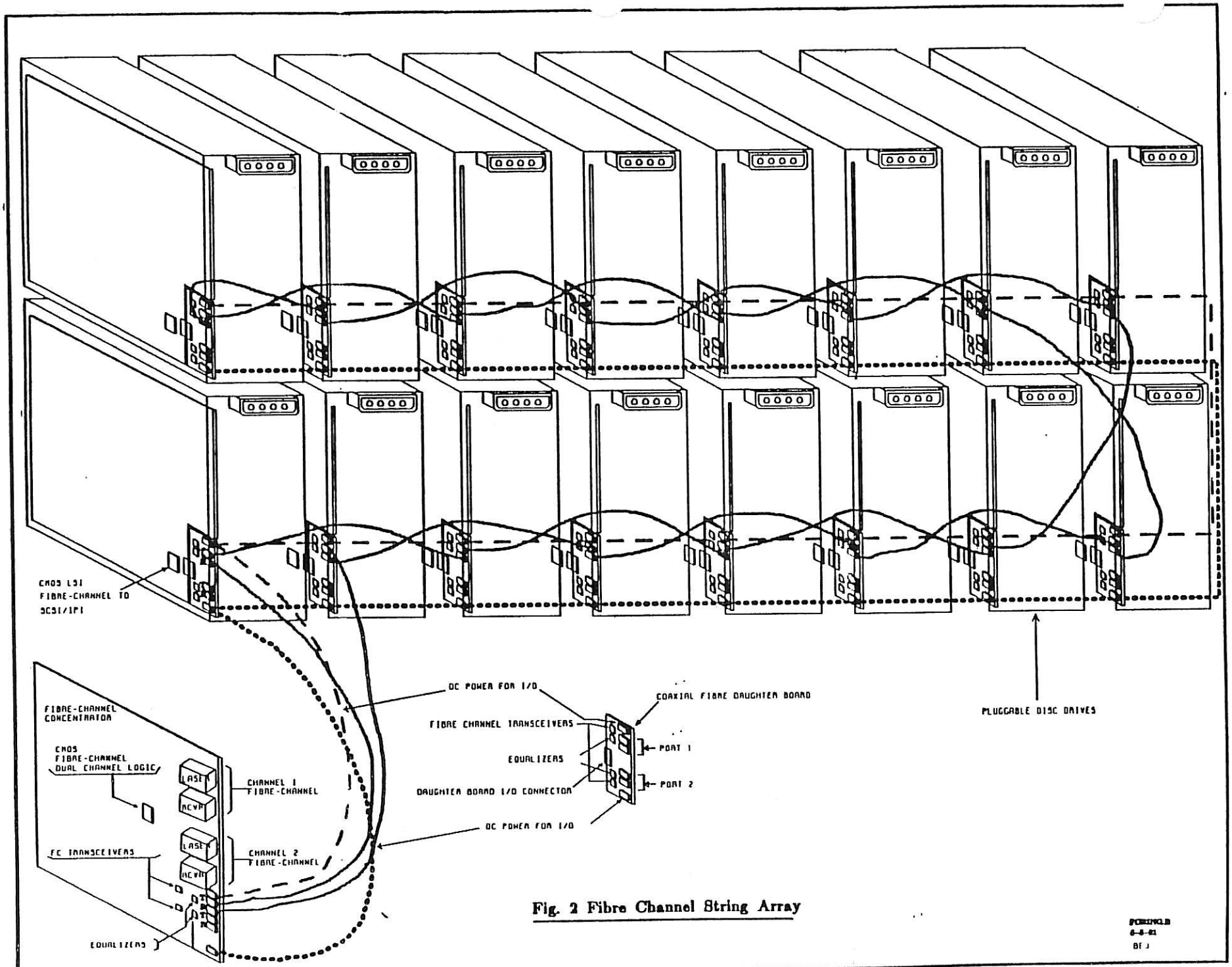


Fig. 2 Fibre Channel String Array

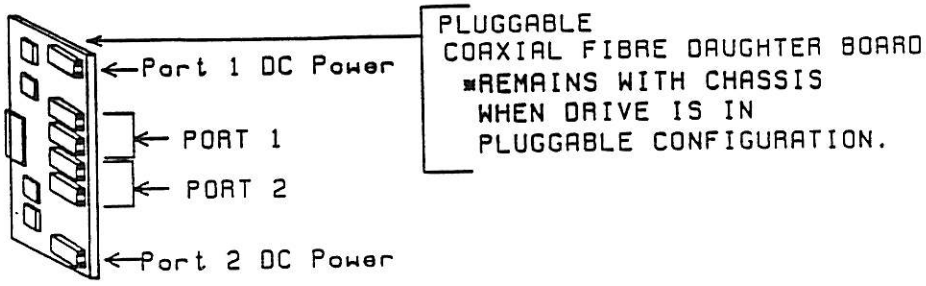


Fig. 1 Pluggable Daughter Board

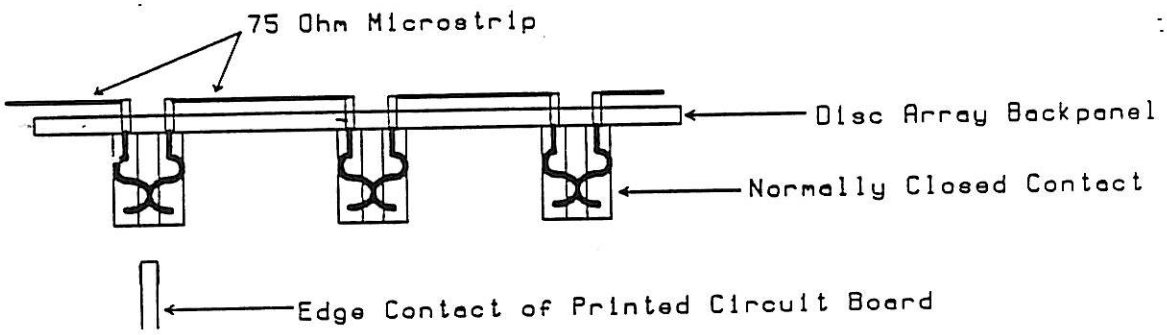


Fig. 2 Normally Closed Contact

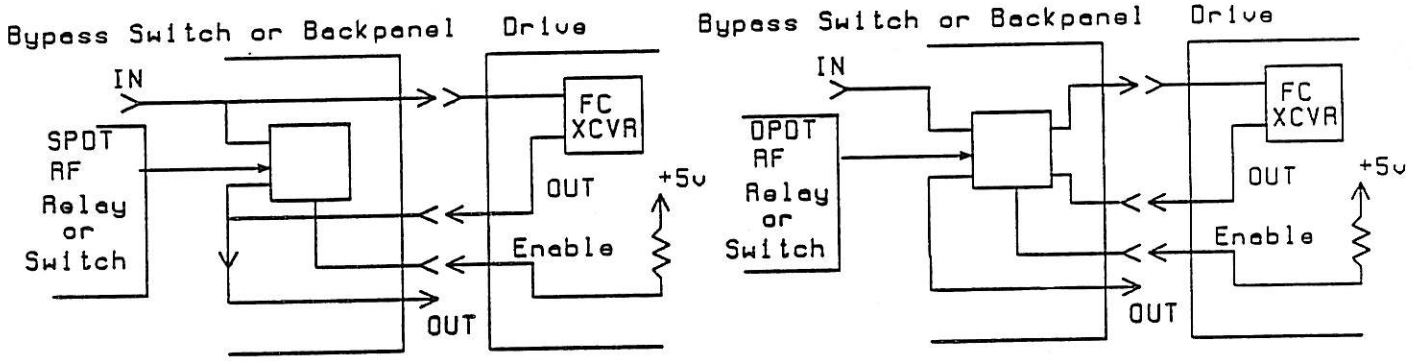


Fig. 3 RF Relay Bypass Switch

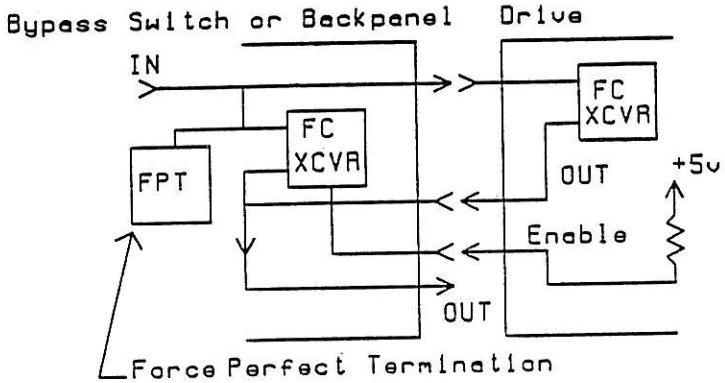


Fig. 4 Redundant Transceivers

fcresloop
bej/5-28-91

BJ-4

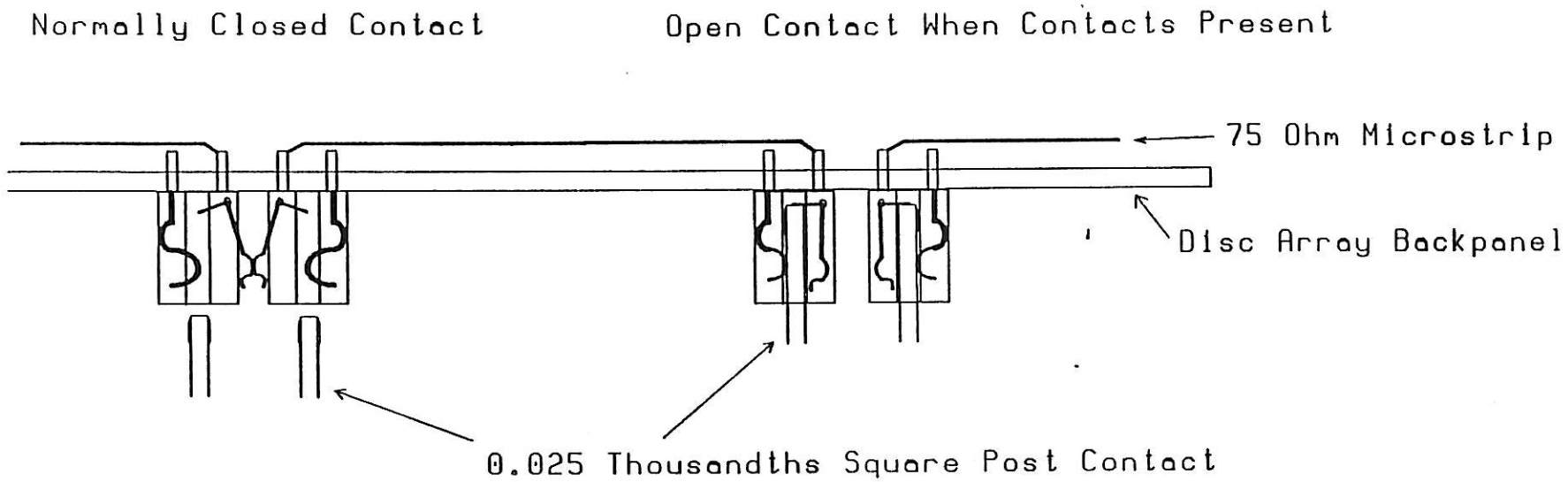


Fig. 1 Resilient Connector For Fibre Channel Loop

fcresloop

bej/6-17-91

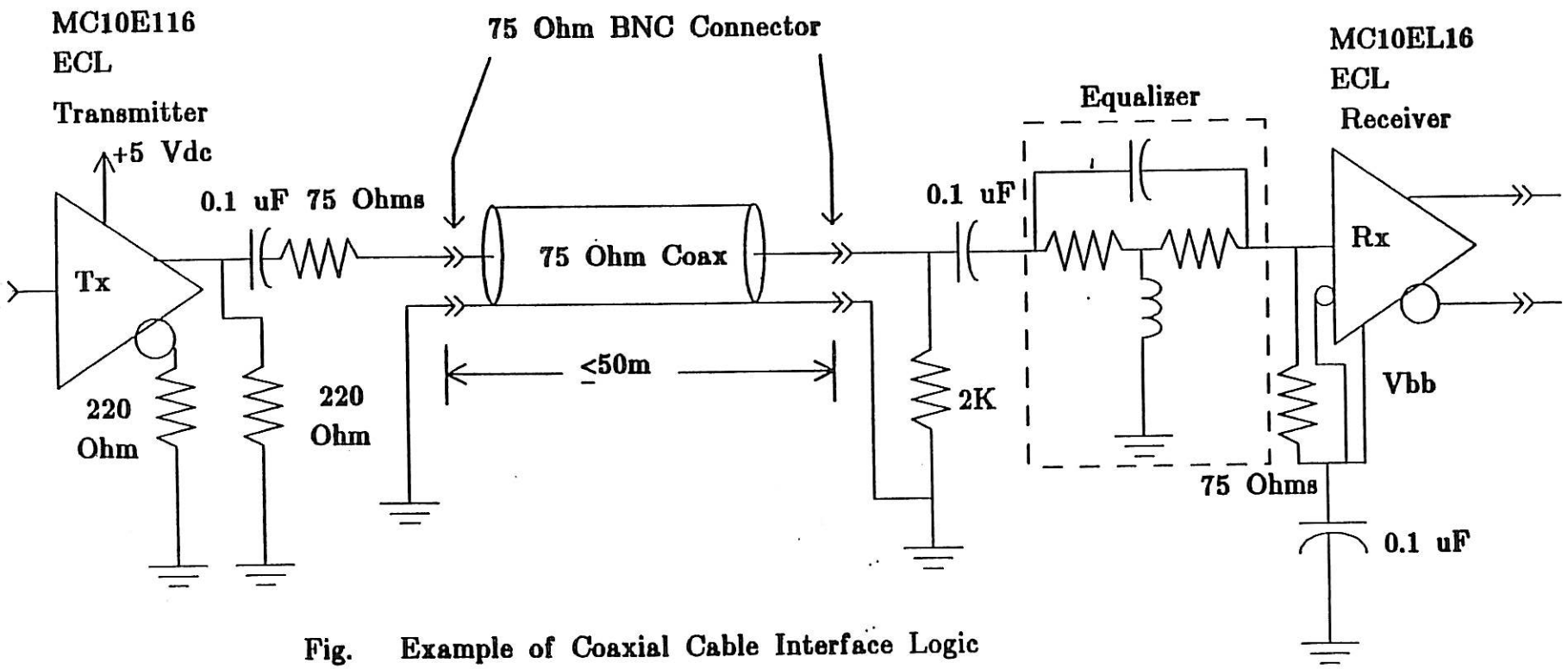


Fig. Example of Coaxial Cable Interface Logic

- Simple, Low-Cost, Yet High-Performance Capability

coaxfc.a

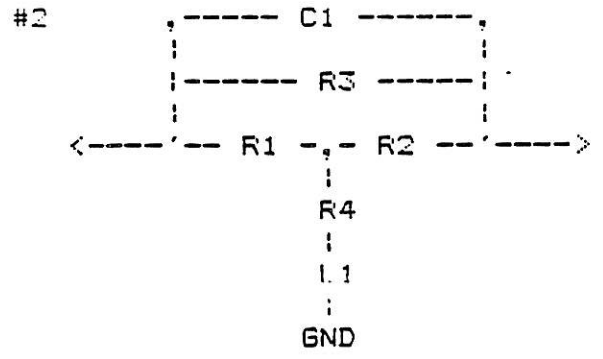
bej/4-4-91

TEST Results of Equalizer

IBM TS 68

(EQ.	VSWR	ATTEN	PHASE	DELAY(ns)
50.000	1.213	-3.8490	6.9348	*****
100.000	1.273	-3.1621	10.4561	0.000
150.000	1.307	-2.5701	11.6454	0.000
200.000	1.319	-2.1188	11.8646	0.000
250.000	1.315	-1.7709	11.6872	0.010
300.000	1.304	-1.4956	11.3267	0.020
350.000	1.289	-1.2739	10.8765	0.025
400.000	1.273	-1.0935	10.3870	0.027
450.000	1.257	-0.9454	9.8881	0.028
500.000	1.242	-0.8231	9.3982	0.027
550.000	1.228	-0.7215	8.9277	0.026
600.000	1.214	-0.6365	8.4825	0.025
650.000	1.202	-0.5649	8.0649	0.023
700.000	1.191	-0.5042	7.6755	0.022
750.000	1.181	-0.4525	7.3136	0.020
800.000	1.171	-0.4081	6.9778	0.019
850.000	1.163	-0.3698	6.6665	0.017
900.000	1.155	-0.3366	6.3780	0.016
950.000	1.148	-0.3076	6.1102	0.015
1000.000	1.141	-0.2821	5.8616	0.014

(EANALYZE(Y OR N)?



C1= 51.000 pf

L1= 64.000 nh

R1= 75.000 Ohms

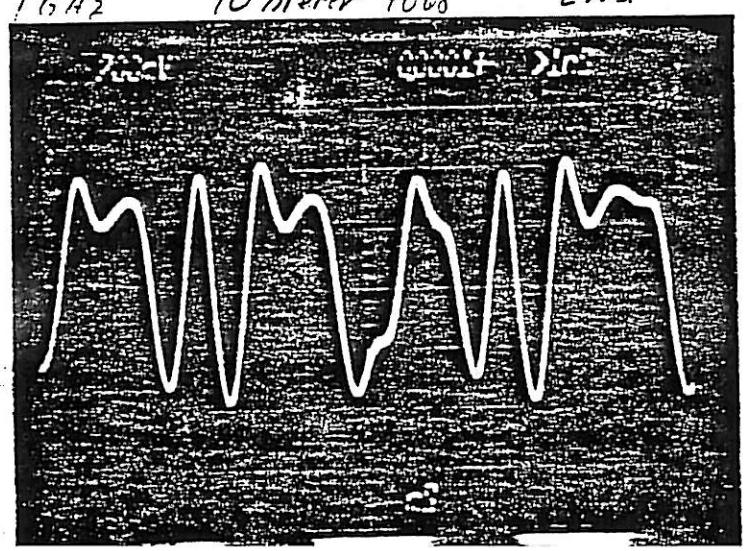
R2= 75.000 Ohms

R3= 33.000 Ohms

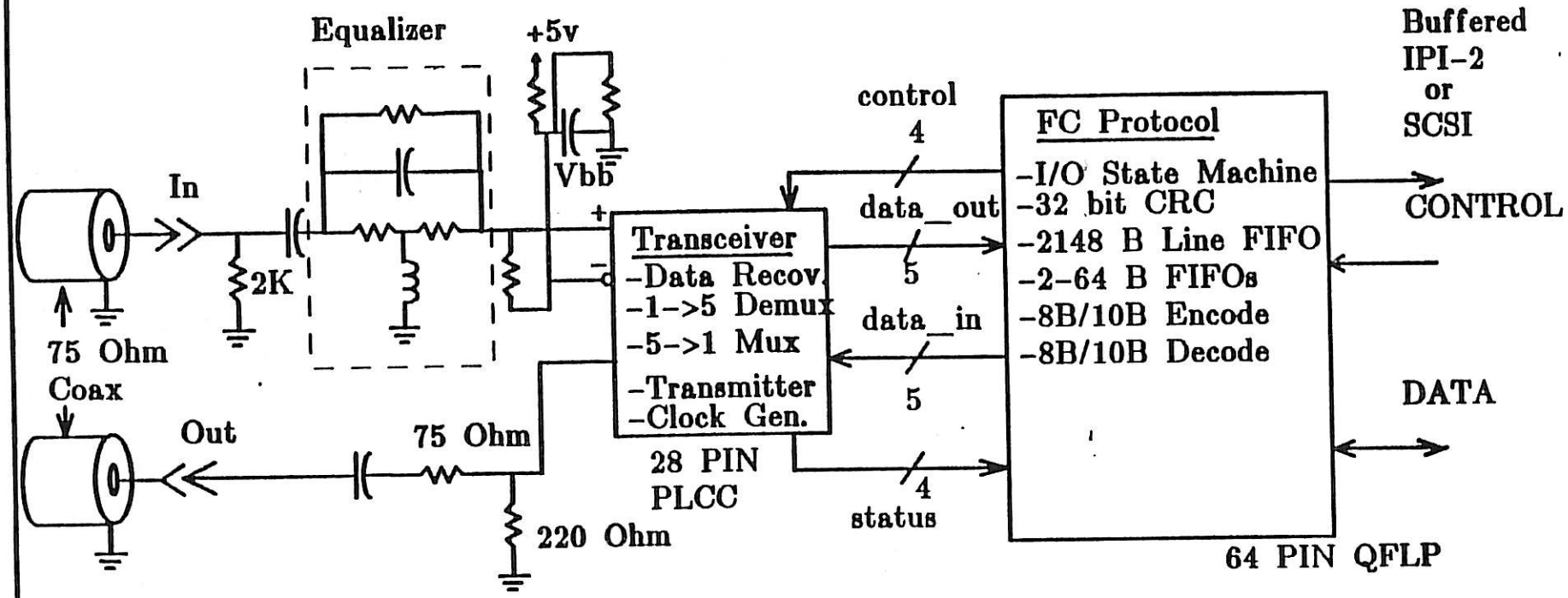
R4= 86.000 Ohms

Select element to ch
or pad value (P),<CR

6-14-91 Vern's
Filter



379



1. FC Applications At 25 MB/sec or Greater:
   - -Transceiver is Separate, ECL or GaAs part.
   - -Transceiver Could Be Upgraded to 50 MB/sec.
   - -FC Protocol is Separate CMOS Array.

2. FC Applications at 12.5 MB/sec or Lower:
   - -One CMOS or BiCMOS Chip.
   - -TI Claims to Have 50 Mbit/sec Transceiver with < 75 mw!

Fig. 1 Fibre Channel Coaxial Cable Interface Electronics

6-5-91
fclsi.c

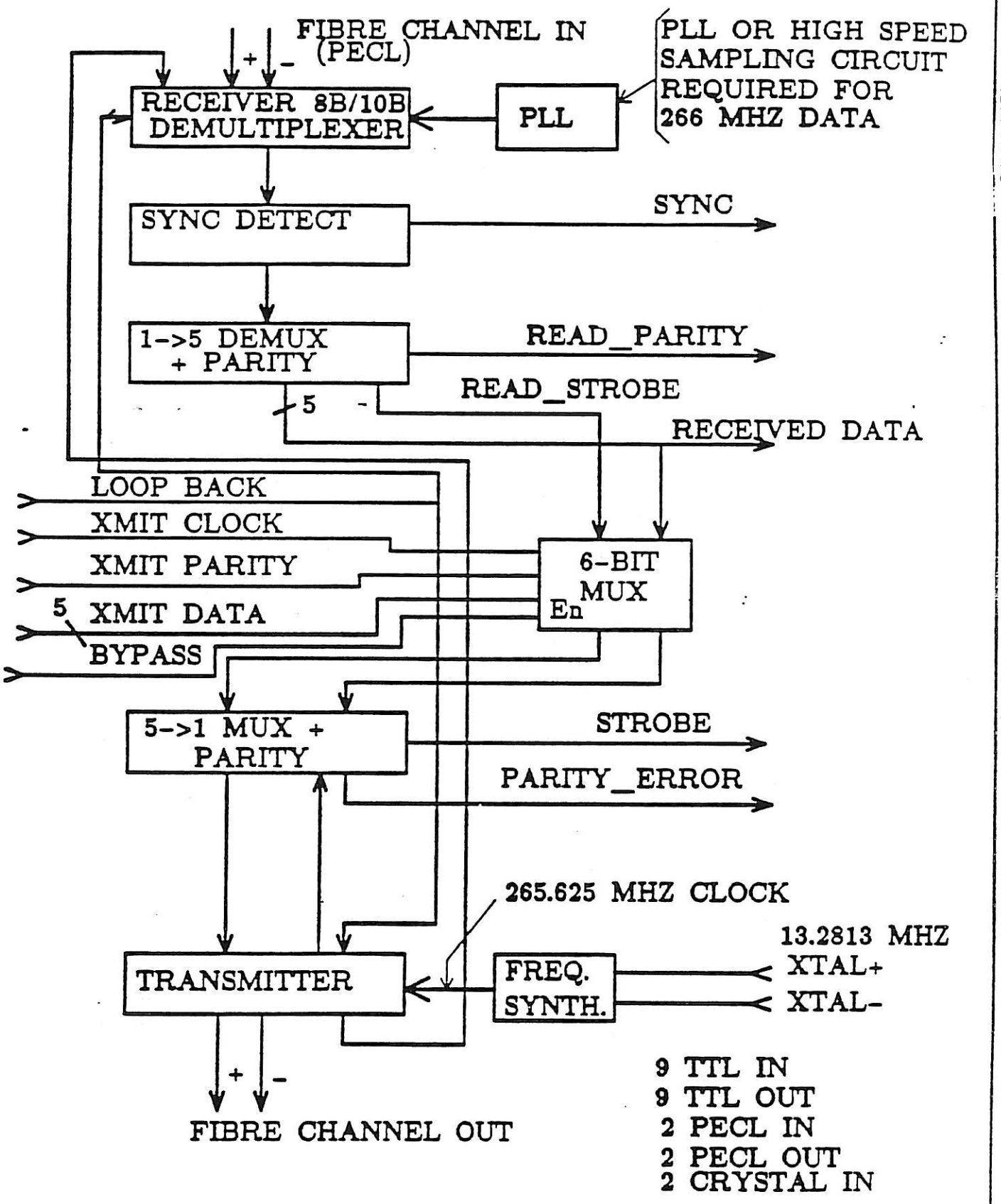


FIG.2 FIBRE CHANNEL TRANSCEIVER
25 MBYTES/SEC

fclsi.a
2-6-91

BJ--8b

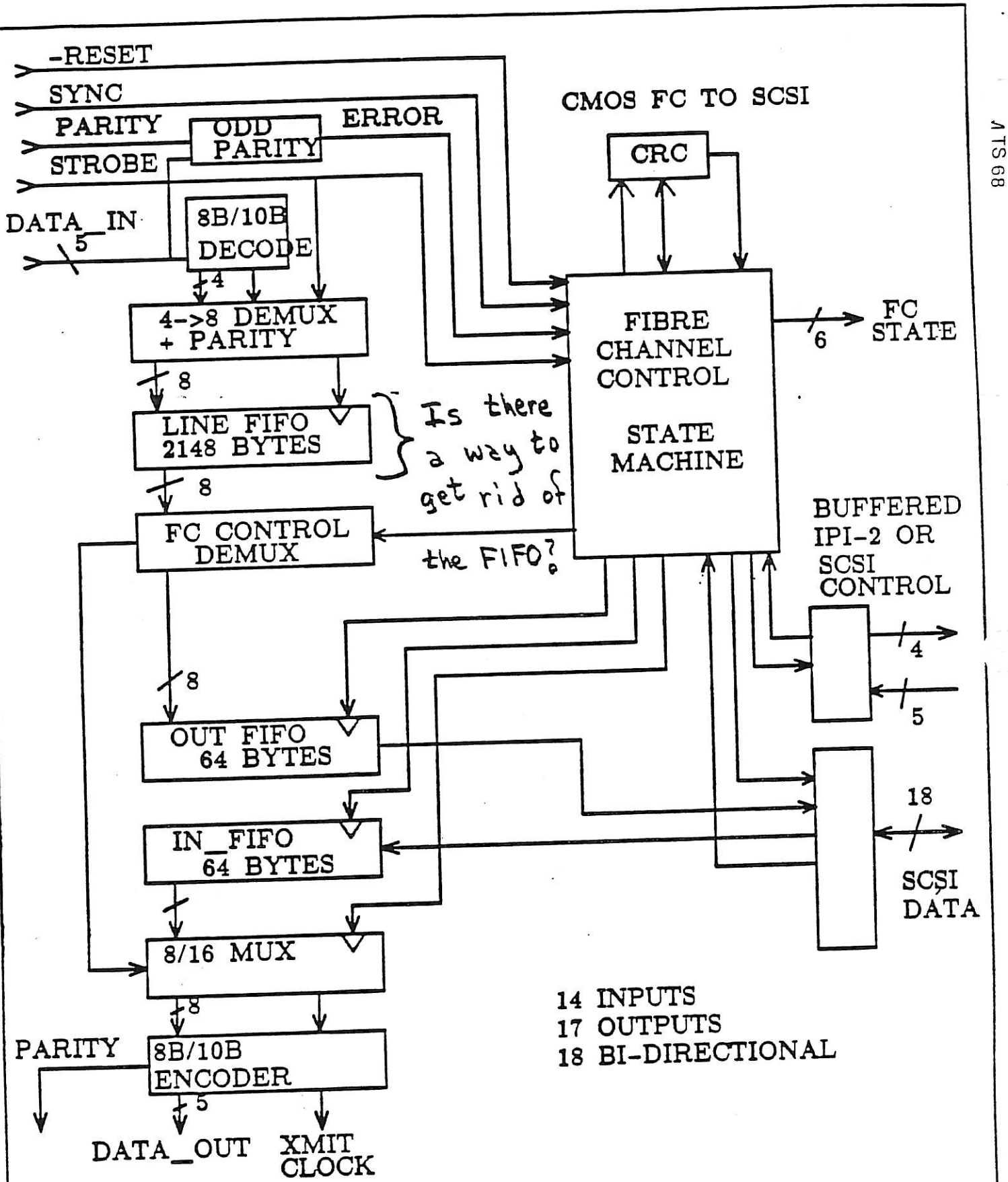


FIG. 3 FIBRE CHANNEL CONTROL AND SCSI LSI

fclsi.b
2-12-91

ATS 68

BJ-9

* SCSI-2 Versus Fibre Channel Costs (Drive Cost)

-Assume High Volume (>100 K/year)

1. SCSI-2 Differential, 20 MB/sec

Item.	Quantity	Description	Total Cost
1.	27	75176B XCVR	$12.15
2.	1	68 Pin Connector	3.50
3.	2	20 pin socket	.80
4.	1	6.8 uf Cap	.35
			$16.80

Does Not Account For Board Space Costs.

2. Optical Fibre Channel, 25 MB/sec

Item.	Quantity	Description	Total Cost
1.	1	FC XCVR	$ 9.45
2.	1	FC Protocol	10.00
3.	1	Laser Driver	2.00
4.	1	Laser Diode	40.00
5.	1	PIN Diode	20.00
6.	1	Trans-Imp Amp.	2.00
7.	1	10H116	2.00
8.	_	Miscellaneous	4.00
			$89.45

3. Coaxial Cable Interface Version Of Fibre Channel 25 MB/sec

Item.	Quantity	Description	Total Cost
1.	1	FC XCVR	$ 9.45
2.	1	FC Protocol	10.00
3.	2	Header	.10
4.	_	Miscellaneous	1.30
			$20.85

Lowest Projected Cost

CONCLUSION: The Coaxial Cable Interface Can Provide a Cost Effective Implementation of Short Distance Fibre Channel.

fccosts/6-5-91

IBM TS 68

BJ-10a

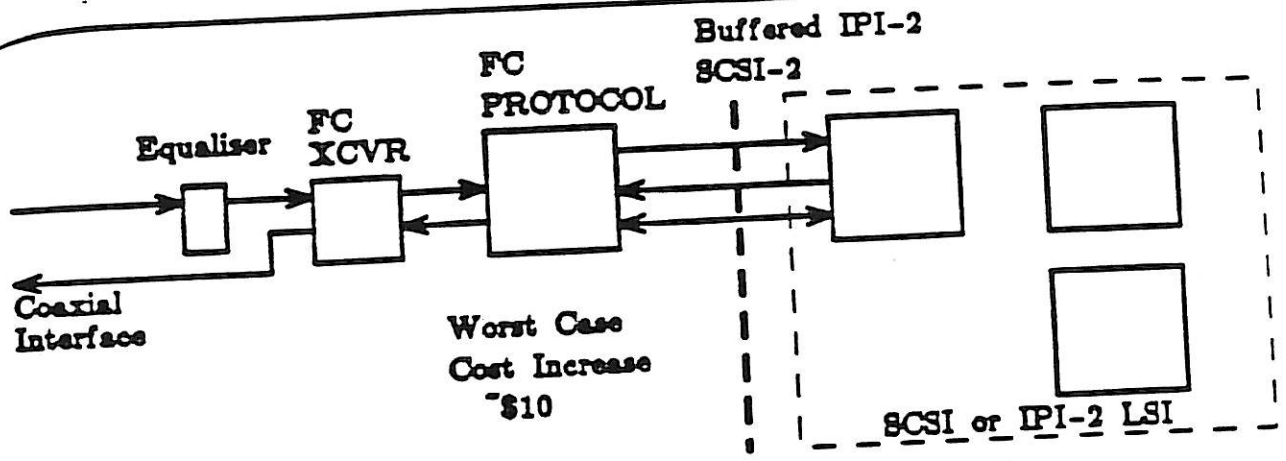


Fig. Stage 1 Fibre Channel LSI Integration (25 MB/Sec)

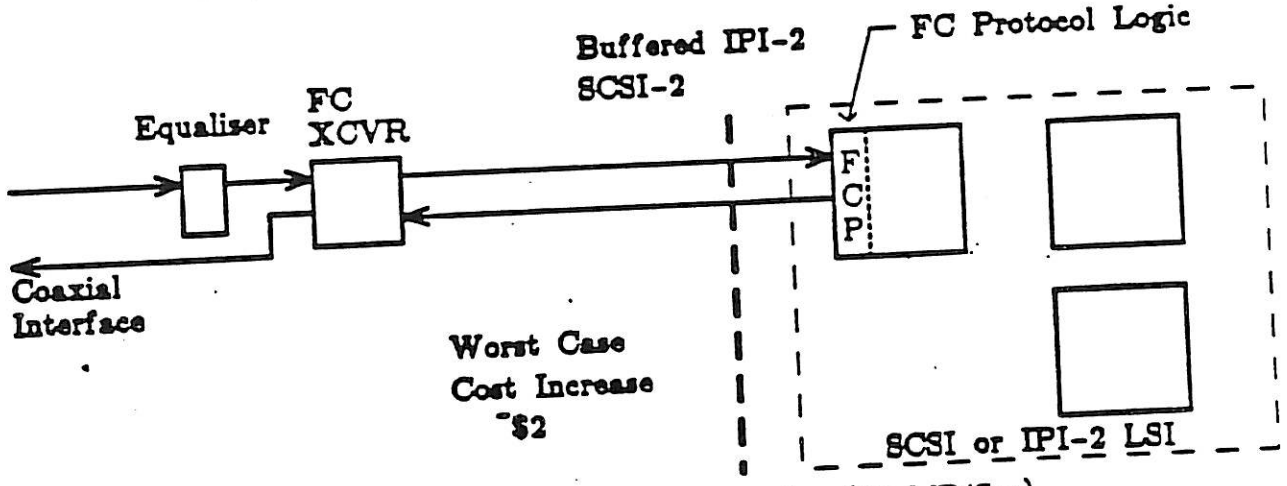


Fig. Stage 2 Fibre Channel LSI Integration (25 MB/Sec)

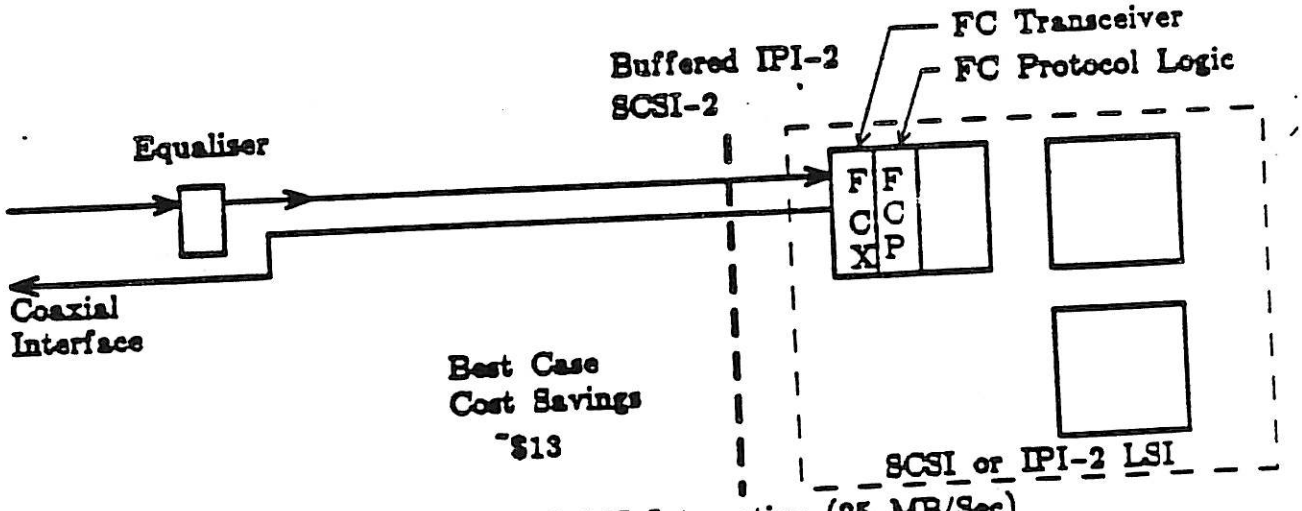
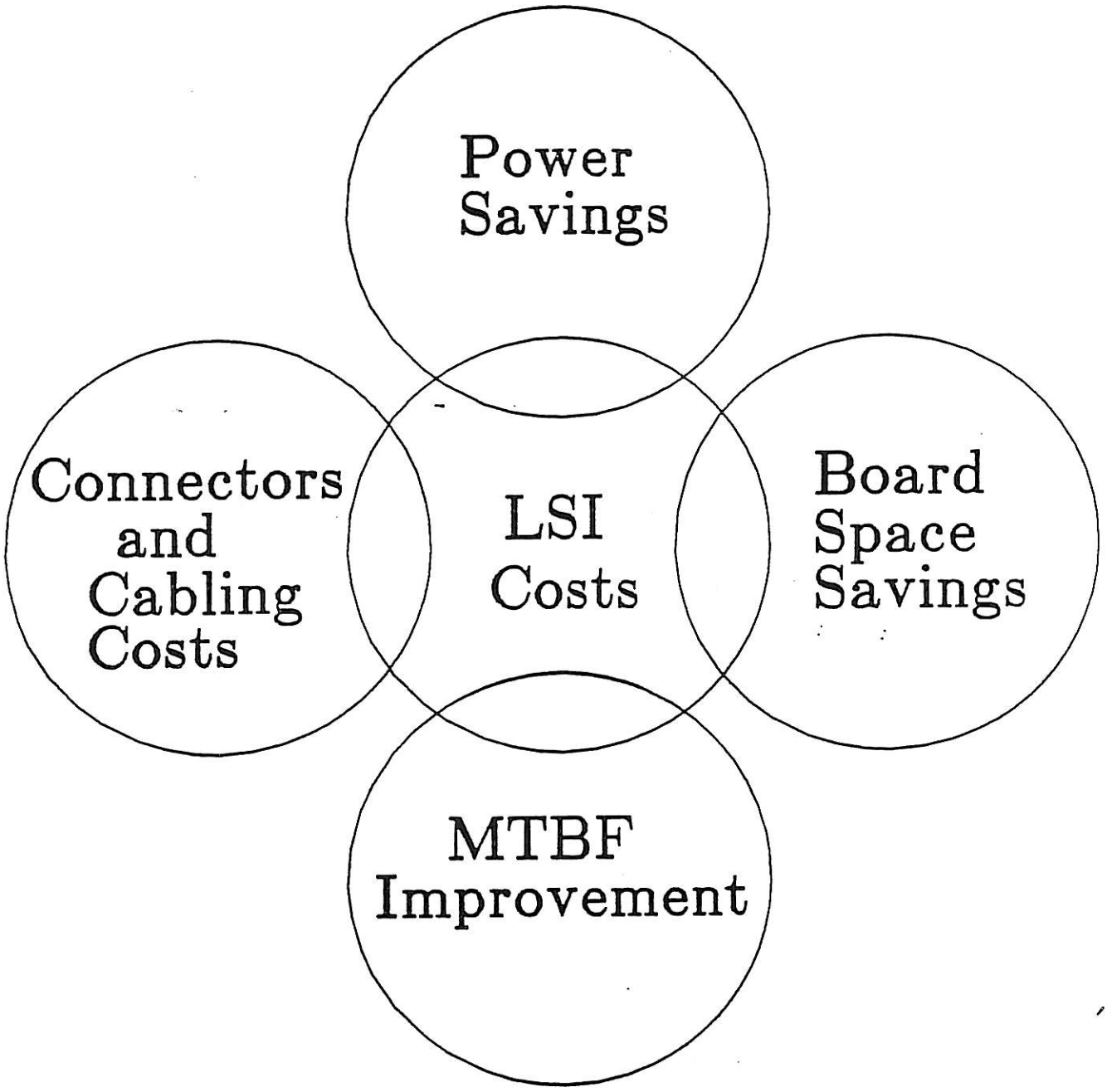


Fig. Stage 3 Fibre Channel LSI Integration (25 MB/Sec)

CONCLUSION: Fibre Channel Should Lower Cost with Future LSI Integration.

fcsvolv/4-8-91

TS 68



FC System Cost Parameters

fcsyscost
6-5-91

BJ-11

Data Rate	Estimated Fibre Channel	Measured IPI-2	Measured SCSI-1 75176 XCVRs	Estimated SCSI-2 75176 XCVRs
5 MB/Sec	75 mW with Serial TI Transceiver 0.6 Watts Total	22-75176 XCVRs 6.05 Watts Max. for XCVRs	12.95 Watts / Ready 17.55 Watts R/W Total	———
10 MB/sec	300 mW with BiCMOS Transceiver 1.0 Watts Total			18-75176 XCVRs 4.95 Watts Max. for XCVRs
20 MB/Sec	300 mW with BiCMOS Cypress XCVR 800 mW wiht Bipolar XCVR 1.2-->1.5 Watts Total	22-75ALS176 XCVRs	———	22-75176 XCVRs 7.425 Watts Max. for XCVRs
50 MB/Sec	1.0 Watt with Bipolar Transceiver 2.0 Watts Total	———	———	45-75176 XCVRs 12.375 Watts Max. for XCVRs.

Not Complete!

Fig. Fibre Channel Power Comparison to IPI-2 and SCSI.

fcpower
6-10-91

Fibre Channel Reliability Improvement

- For SCSI-2,

  (27 XCVRs) * (.00754 Failures/Million Hours) = 0.20358 $\frac{\text{Failures}}{\text{Million Hours}}$

- Approximately 30% of SCSI I/O Failure Rate.

- Ignores 216 I.C. Pins That Must Be Soldered Correctly.

- Assume that Fibre Channel Transceiver is 8 Times More Likely to Fail.

  → Single Transceiver is Still 3 Times More Reliable.

Fibre Channel Space Savings

- Single 28 Pin Transceiver Should Take '1/8 the Space of Differential SCSI.

- Could Eliminate an Unjustified LSI Effort or Tight Packaging.

fcfcrel
6-10-91

Low-Cost Fabric

PROBLEM

Avoid continuously looping frames for the following reason.

1. Destination is OFF-Line (normal)

2. Corrupted Destination Address (error)

SOLUTION

Add one of the following to each Loop element with an attempt at minimizing the Loop element from a normal end-node.

1. Add a HOP COUNT preceding the SOF

2. DUPLICATE Address and CHECK for Source Address

Low-Cost Fabric

HOP COUNT

Can be done (i.e., a 10-bit counter can be implemented in hardware to decrement a counter and verify a fixed amount). This requires that every loop element needs to decrement and check a value to determine if the hop count has been exhausted.

PROBLEM:

1. Requires a new HOP COUNT Primitive.

2. Requires a new detector in the Fabric port and the Loop element (aside from the normal address check). Also, a 10-bit counter and comparator is needed.

DUPLICATE Address and CHECK for Source Address

1. In the three-byte address field choose the address such that the two right-most bytes are duplicated using the 10-bit codes that provide even disparity (there are 72 such addresses). This combined with the domaine address of the CANSTAR proposal would give a one byte domaine and a duplicate address in the other two bytes. These address would always be used on a loop, but they may be used in a normal end-node as well.

2. Each Loop element checks to make sure that any source or destination address on the loop has the right-two bytes (10-bits) are identical. If they are not, the frame is discarded.

3. Each Loop element checks its own source address. If it is found, the frame has traversed the loop and is discarded. This includes a Fabric Port which would recognize that a non-duplicate address (or another domaine byte) is the source or it would remember that it placed this frame onto the loop. NOTE: this may restrict the number of Fabric ports on a loop to one.

Low-Cost Fabric

DISCARD POLICY (Duplicate Address):

D-ID	S-ID	Comments
D XX	D YZ	Source outside of Loop (F-port discards)
D YZ	E XX	Source on Loop - going to Fabric (Discarded by Fabric or Source)
D YZ	D VX	Discarded by anyone on Loop (Frame should not be on Loop)
D XX	D YY	Discard by Source if seen

Low-Cost Fabric

PROBLEMS (Duplicate Address):

1. No two F-ports supported
2. Error hits both of the duplicate bytes
3. Source sends a Frame to an OFF-line or non-existing Destination and then goes OFF-line.
4. Loop Address assignment
5. Address check and discard (especially in F-port)

TO: X3T9.3 FIBRE CHANNEL COMMITTEE
DATE: JUNE 19, 1991
SUBJECT: LOW COST FABRIC WORK GROUP

A PROPOSAL TO DISCARD UNWANTED FRAMES IN THE LOW COST FCS LOOP

KC Chennappan
Giles Frazier
Jerry Rouse
IBM AUSTIN
11400 Burnet Road
Austin, Texas 78758
(512) 823–0000

THE PROBLEM:

A simple method of eliminating unwanted circulating frames is needed.

TWO PROPOSED SOLUTIONS

1) INSERT A "HOP COUNT" IN FRONT OF ALL FRAMES
   ADVANTAGES
   - Does not require optional headers or timers
   - Guarantees elimination of all unwanted frames

   DISADVANTAGES
   - Can't directly connect a loop to F_PORT or N_PORT
   - Requires design of new frame delimiter hardware
   - Requires basic changes to FC_1

2) USE A SPECIAL ADDRESSING SCHEME
   ADVANTAGES
   - Does not require new frame delimiter hardware
   - Does not require timers or optional headers
   - Does not require basic changes to FC_PH

   DISADVANTAGES
   - Does not guarantee elimination of all unwanted frames

A NEW PROPOSAL: USE A SIMPLE FORM OF EXPIRATION_SECURITY HEADER

ADVANTAGES
- Does not require new delimiter hardware
- Allows direct connection to fabrics and N_PORTS
- Does not require basic changes to FC_PH

DISADVANTAGES
- Requires use of optional header & timer

AN OUTLINE OF SOME PRELIMINARY WORK IS GIVEN HERE

WHATEVER SOLUTION WE CHOOSE, STAYING WITHIN FC_PH SHOULD RESULT IN:

1) L_PORT designs very similar to N_PORT designs
2) No new "Loop Architecture" specification
3) More L_PORT users

FRAME CHARACTERISTICS

* DATA FRAMES

  - These frames carry user data and link control protocol
  - Source of these frames may be within a loop or from a fabric
  - These frames have expiration_security headers

* "LOOP_CONTROL" FRAMES

  - These frames are used only for loop error recovery and initialization
  - Any L_PORT may send these frames
  - All L_PORTs must accept and process these frames
  - These frames are local to a loop
  - These frames do not contain optional headers or payloads.
  - These frames carry a simple FC_4 protocol in the R_CTL and TYPE fields
    - CATEGORY 000000 = LOOP RESET FRAME
    - CATEGORY 000001 = LOOP RESET RESUME FRAME
    - OTHER CATEGORIES RESERVED

L_PORT CHARACTERISTICS

ADDRESSING

* UNIQUE L_PORT ADDRESS
  - This address is used as the S_ID for the port in all frames sent by the port.

* COMMON "LOOP_CONTROL" ADDRESS
  - This address is used to broadcast Loop_control frames to all L_PORTS

OPERATIONAL STATES

* NORMAL STATE
  - This is the usual state in which L_PORTS pass, accept, or generate data frames and discard bad frames.

* LOOP RESET STATE
  - This state is entered during any L_PORT initialization. It is initiates a loop time synchronization cycle.

* LOOP RESET RESUME STATE
  - This state is used to exit a loop time synchronization cycle

L_PORTS RETAIN ALL ORIGINAL DESIGN GOALS

CHEAP
HOT PLUGGABLE
CLASS 2 ONLY
INTERCONNECTABLE AS PEERS

LOOP CONTROL PROTOCOLS

OBTAINING THE TIME

* When any L_PORT is connected, it must obtain the time

* First, the L_PORT sends a frame to the timeserver. If it responds then the port uses the time it obtains. (This step could be optional for closed loops without a timeserver.)

* If no timeserver is present (closed loops), then the port sends a Loop_Reset broadcast frame. This frame forces all L_PORTS to reset their times.

  - A protocol similar to the FCS "Link Recovery Protocol" is followed:
    1. Transmit "Loop Reset" frame
    2. If "Loop Reset" is recognized, then transmit "Loop Reset Resume"
    3. If "Loop Reset Resume" is recognized, return to normal state. (—FC_PH Rev. 2.1 sec. 27.6)

  - This protocol sets all L_PORTS in a loop to time zero without a timeserver.
  - Frames in transit during the reset cycle proceed to their destinations

  - Expired frames are discarded.

ERROR RECOVERY PROTOCOL

* This protocol is entered only after other error recovery such as sequence retransmission

1. Resynchronize the time. (See above)
2. Execute the FC_PH link recovery protocol using LR and LRR primitive sequences.

MANY MORE PROBLEMS MUST BE SOLVED, ONLY A PRELIMINARY OUTLINE IS GIVEN HERE. SOME AREAS TO WORK ON ARE:

- Can the loop use existing FC_PH buffer to buffer flow control?

- Can L_PORTS of this type be made as inexpensively as other designs?

- Many others... (Some details are included on the next few pages.)

L_PORT FRAME PROCESSING

```
I   IAL STATE
  .HILE (D_ID MATCHES OWN D_ID):
    IF (D_ID /= FFFFFX)      /*All user data frames satisfy this criterion*/
    THEN IF {(OPTIONAL TIME HEADER) AND (MYTIME < EXPTIME)}
         THEN (PASS OR KEEP BASED ON D_ID)
         ELSE (DISCARD)
         ENDIF;
    ELSE IF (D_ID /= FFFFF9)    /*Assume FFFFF9 is Loop_control address*/
         THEN (PASS)            /*Pass all frames to all fabric servers*/
         ELSE IF (LOOP_RESET FRAME)      /*Begin loop reset procedure.*/
              THEN (PASS FRAME & ENTER RESET STATE)
              ELSE (DISCARD FRAME) /*First frame to loop reset address must
              ENDIF;                                  be a Loop_Reset frame*/
         ENDIF;
    ENDIF;
  ENDWHILE;

LOOP RESET STATE    /*A loop reset frame was sent, await Loop Reset Resume.*/
    IF (D_ID = FFFFF9)                /*Assume FFFFF9 = Loop_control address*/
    THEN IF (LOOP_RESET FRAME & S_ID = OWN ID)
                              /*L_PORT sending the reset awaits its return*/
         THEN (DISCARD FRAME
               SEND LOOP_RESUME FRAME
               ENTER "LOOP RESUME" STATE)
         ELSE IF (LOOP_RESUME FRAME)  /*Other L_PORTS wait for resume frame*/
              THEN (SET MYTIME = 0 & ENTER NORMAL STATE)
              ELSE (PASS FRAME)
              ENDIF;
         ENDIF;
    ELSE (PASS OR KEEP FRAME BASED ON D_ID)   /*Flush valid frames to their*/
    ENDIF;                                                  destinations*/

LOOP RESUME STATE /*The L_PORT sending the reset enters this state*/
    IF (ADDRESS = FFFFF9)   /*Assume FFFFF9 = Loop_control address*/
    THEN IF (LOOP_RESUME FRAME)
         THEN (DISCARD
               SET MYTIME = 1
               ENTER NORMAL STATE)
         ELSE (DISCARD FRAME)
         ENDIF;
    ELSE (DISCARD FRAME)  /*Discard any unwanted circulating frames*/
    ENDIF;
```

Note:

A. The above protocol is very similar to the "Link Recovery" protocol of FC_PH:
   1. Transmit LR
   2. Wait for LR, then transmit LRR
   3. Wait for LRR, then resume normal operation

B. L_PORT Loop_Control (broadcast) address = FFFFF9.

C. Frames to FFFFF9 have two meanings (distinguished by category bits):
   1. Loop Reset
   2. Loop Reset Resume

IBM9664656

L_PORT INITIALIZATION PROCEDURE

```
SEND "WHATTIME" FRAME TO TIMESERVER FFFFFB
IF (TIMESERVER FFFFFB ANSWERS REQUEST)
THEN (SET MYTIME = SERVERTIME)
ELSE
     SEND "LOOP_RESET" FRAME TO FFFFF9
     ENTER "LOOP RESET" STATE
ENDIF;
```

FRAME FORMATS

DATA FRAMES

DF_CTL (22) = 1 (EXPIRATION_SECURITY HEADER PRESENT)
SET 1st 4 BYTES TO MYTIME + 4 SEC, LAST 12 BYTES TO 0

LOOP_CONTROL FRAME FORMATS

R_CTL: DL=LINK DATA; CATEGORY = 000000 - LOOP_RESET
= 000001 - LOOP_RESUME
(RESERVE OTHER VALUES)

D_ID = FFFFF9
S_ID = OWN UNIQUE LOOP ADDRESS
TYPE = "L_PORT PROTOCOL" (ONE MORE FC_4 PROTOCOL)
NO OPTIONAL HEADERS, NO PAYLOAD

TIME GRANULARITY CALCULATIONS

- FRAME LIFES ARE ALLOWED TO BE SEVERAL LOOPAROUND TIMES

- MAXIMUM LOOPAROUND TIME:
  = 64 L_PORTS * {MAX L_PORT DELAY + MAX INTERPORT XMSN TIME}
  = 64 * {(2300 BYTES/25E6 BY/SEC) + 30 uS }
  (ASSUMES 2 K FRAME AND 10 KM INTERPORT DISTANCE)
  = 64 * { 91uS + 30 uS}
  = 7.5 mS

- CONCLUSION:
  IF WE ALLOW A 2 SECOND FRAME LIFE, THEN WE WILL NEVER DISCARD A FRAME PREMATURELY

IBM 966K 6856

A FEW ERROR CASES

ED LOOP (NO FABRIC--ONLY L_PORTS PRESENT)

ERROR	RECOVERY
BAD D_ID	DISCARDED AFTER EXP TIME
DEAD D_ID	DISCARDED AFTER EXP TIME
L_PORT WITH WRONG TIME	HARDWARE FAILURE--USE LR, LRR PRIMITIVES
BAD "RESET TIME" FRAME	TIME IS RESET ANYWAY
L_PORT RESETS TIME & DIES	NEXT NODE IN LOOP TAKES OVER RESET PROTOCOL
MULTIPLE L_PORTS SEND RESET FRAMES SIMULTANEOUSLY	PROTOCOL PROCEEDS NORMALLY

OPEN LOOP (FABRIC(S) PRESENT)

ERROR	RECOVERY
BAD D_ID	DISCARDED AFTER 4 SEC
DEAD D_ID	DISCARDED AFTER 4 SEC
L_PORT WITH WRONG TIME	HARDWARE FAILURE--USE LR, LRR PRIMITIVES
BAD "RESET TIME" FRAME	TIME IS RESET ANYWAY
L_PORT RESETS TIME & DIES	NEXT NODE IN LOOP TAKES OVER RESET PROTOCOL
MULTIPLE L_PORTS SEND RESET FRAMES SIMULTANEOUSLY	PROTOCOL PROCEEDS NORMALLY

Performance of a Gbps Buffer Insertion Ring with Fairness

Dr. Jeane S.-C. Chen

IBM Research Division

FFOL Presentation, June 1991

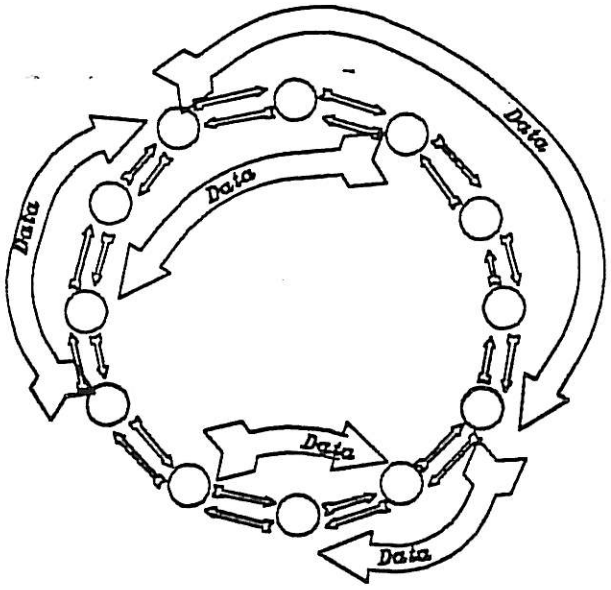
IBM

Outline

- Buffer insertion with spatial reuse
- Flow-based global fairness algorithm
- Performance results for various scenarios
  - Saturation analysis
  - Delay-throughput study
  - The effect of fairness algorithm
- Summary

Spatial Reuse

- Multiple nodes can transmit simultaneously
- Packets are removed by their destinations



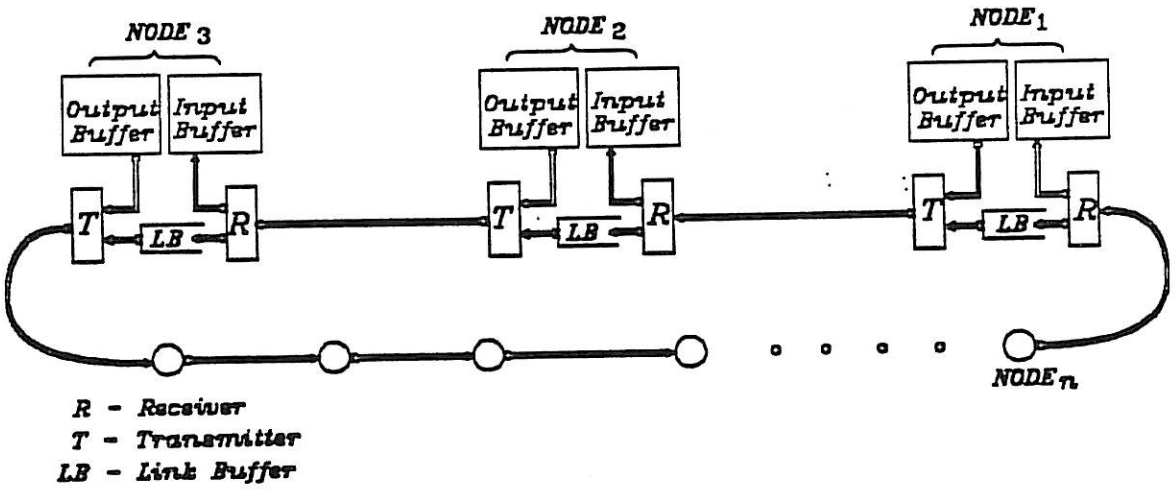
Throughput Gain of Spatial Reuse

- For a full-duplex ring with n nodes, uniform destination distribution, and shortest-path routing
  - maximum distance $\frac{n}{2}$
  - average distance $\frac{n}{4}$

---> Potentially 4 times the link bandwidth in each direction

Buffer Insertion Ring with Spatial Reuse

- Local Access Decision
  - Transmit whenever the insertion buffer is empty
- Ring traffic has non-preemptive priority
- Cut-through via intermediate node



Advantages

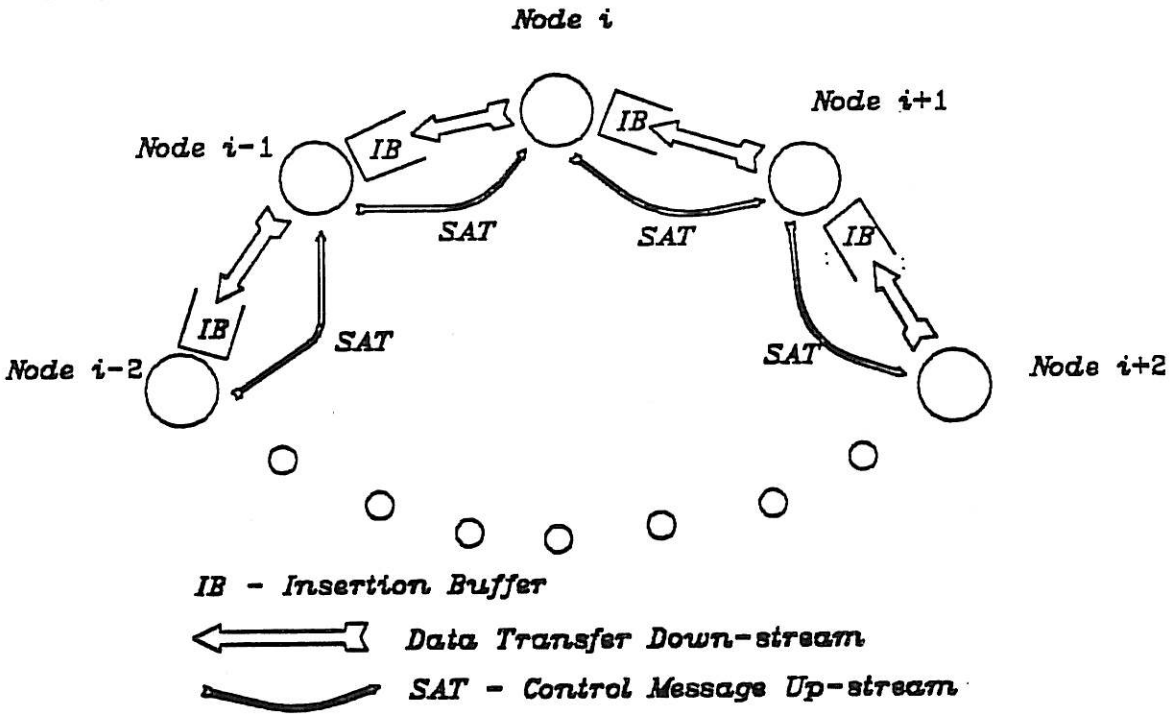
- Immediate access under light load
- Single active node has access to full capacity
- Variable size packets

Problems

- Large delay bounds
- Fairness (up-stream priority)

Flow-based Global Fairness Algorithm

- A single control message: **SAT**
- SAT is used for regulating the access into the ring
- SAT rotates in opposite direction to data
- SAT creates a *global cycle*



- Each node has a given *quota* for transmission in each global cycle

$$Q_{min} \leq quota \leq Q_{max}$$

Flow-based Global Fairness Algorithm

The SATisfied condition

- The node is SATisfied if:
  1. it has sent Q_{min} bytes between two successive SAT visits, **or**
  2. its output queue is empty.

The transmission condition

- The node can transmit if:
  1. it has transmitted less than Q_{max} bytes since the last SAT visit, **and**
  2. the IB is empty.

The SAT algorithm

- When the SAT is received:
  1. if SATisfied, then forward the SAT, else
  2. hold until SATisfied, then forward the SAT.
- After forwarding the SAT renew the quota to Q_{max}.

System configuration

- Transmission rate: 1 Gbps
- Topology: Dual ring
- Number of stations: 20
- Fiber length: 20 km
- Stations are equally spaced

Traffic Profile and Routing

- Poisson arrival
- Hyperexponential packet length
  - CV: 2
  - mean packet length: 1 Kbytes
  - maximum packet length: 4.5 Kbytes
- Shortest path routing
- No transmission from a node to itself

Max Aggregate Throughput:

- Even number of stations

$$8 - \frac{8}{n} \times \textit{link bandwidth}$$

- Odd number of stations

$$8 - \frac{8}{n+1} \times \textit{link bandwidth}$$

Parameter Definitions

- System delay (T_{sys}): delay experienced from arrival until delivered to destination
- Access delay (T_{acc}): delay in HOL(Head-Of-the-Line) of input queue until access of the ring
- Transmission delay (T_{tx}): time required to transmit a packet
- Buffering delay (T_{buf}): delay experienced in insertion buffers
- Propagation delay (T_{prop}): delay experienced in propagation
  - propagation time is 5 μS / km

- $T_{sys} = T_{acc} + T_{tx} + T_{buf} + T_{prop}$

- SAT cycle: duration between two consecutive SAT visits

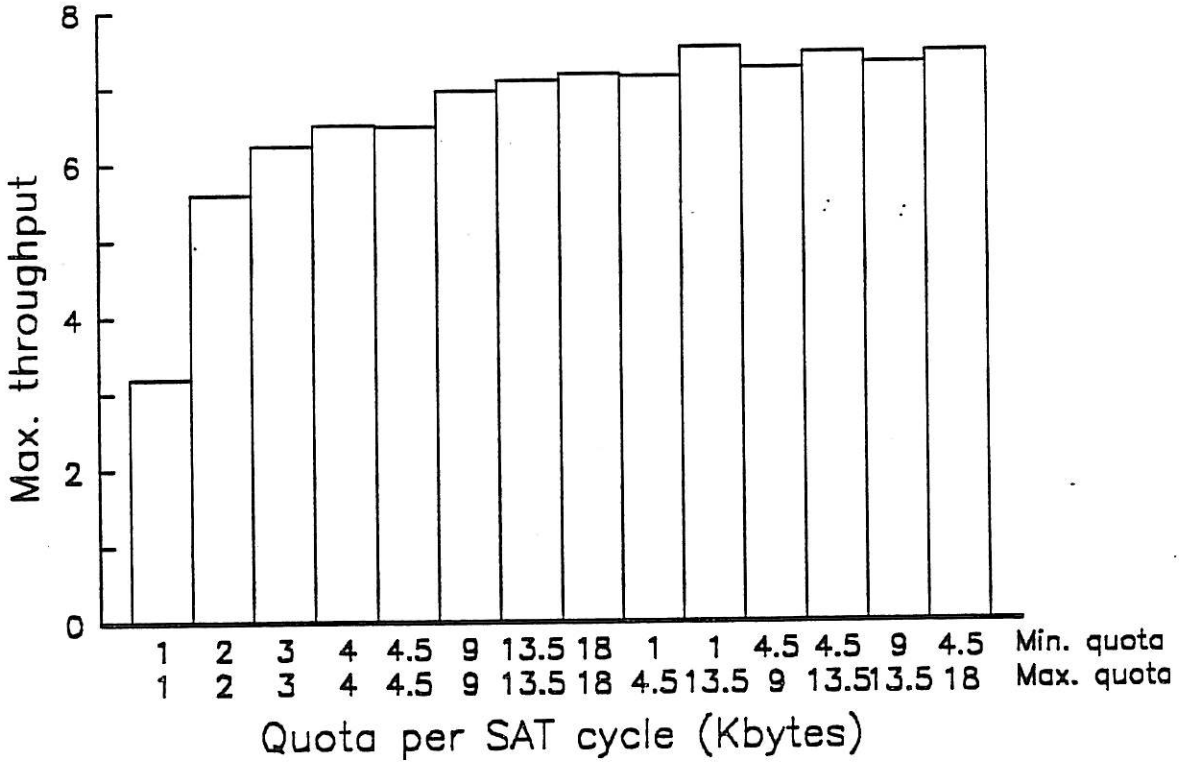
Simulation Scenarios

Saturation analysis

- Uniform traffic
- Stations are fully loaded
- Results:
  - → Maximum aggregate throughput vs. different quotas
  - → SAT cycle length vs. different quotas

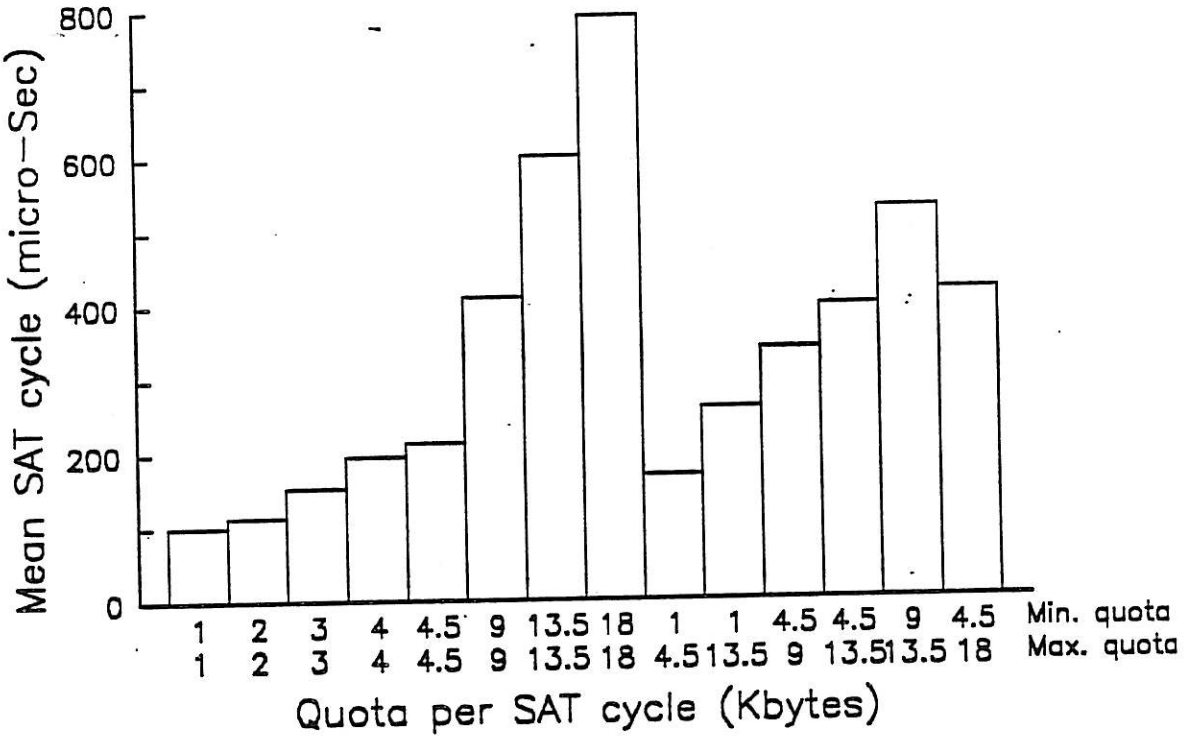
Results: Saturation Analysis

Maximum Throughput for Different Quotas



Results: Saturation Analysis

Mean SAT Cycle time for Different Quotas



Simulation Scenarios

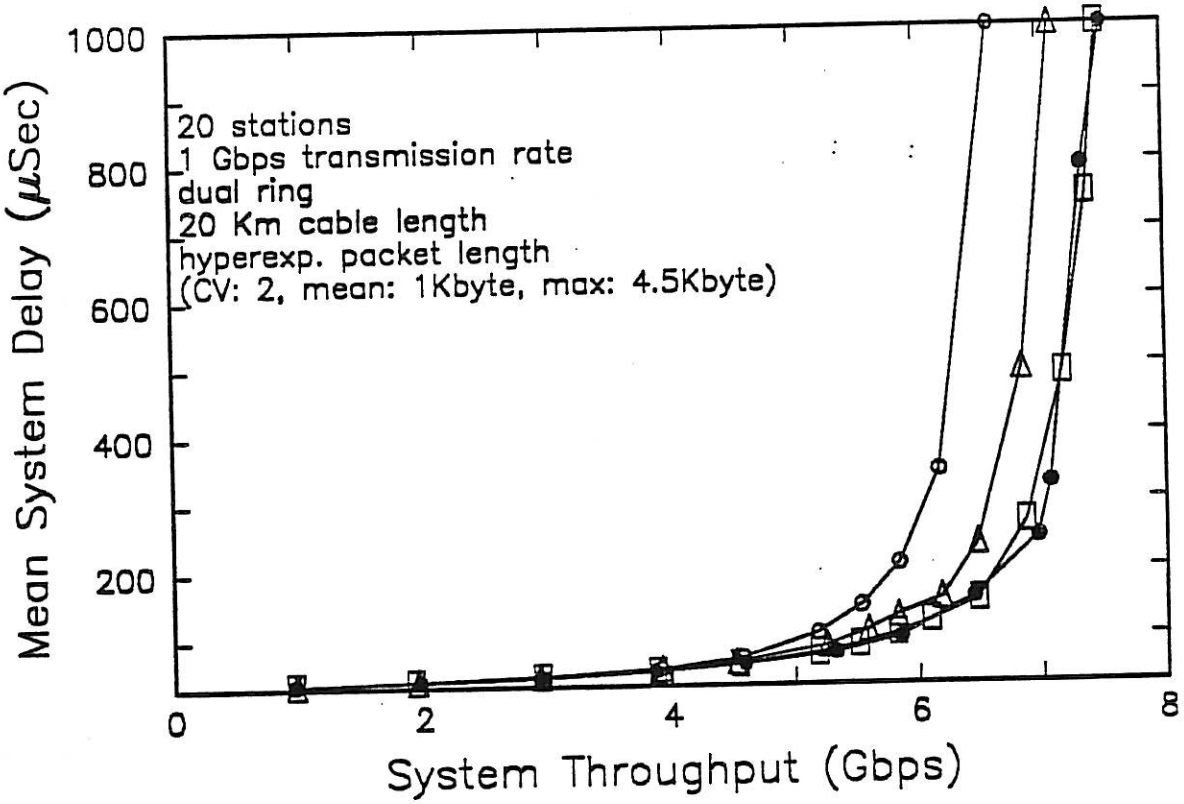
Uniform traffic

- Uniform input rate
- Uniform destination
- Results:
  - System delay vs. throughput
  - Access delay vs. throughput
  - SAT cycle vs. throughput

Results: Uniform Traffic

Mean System Delay

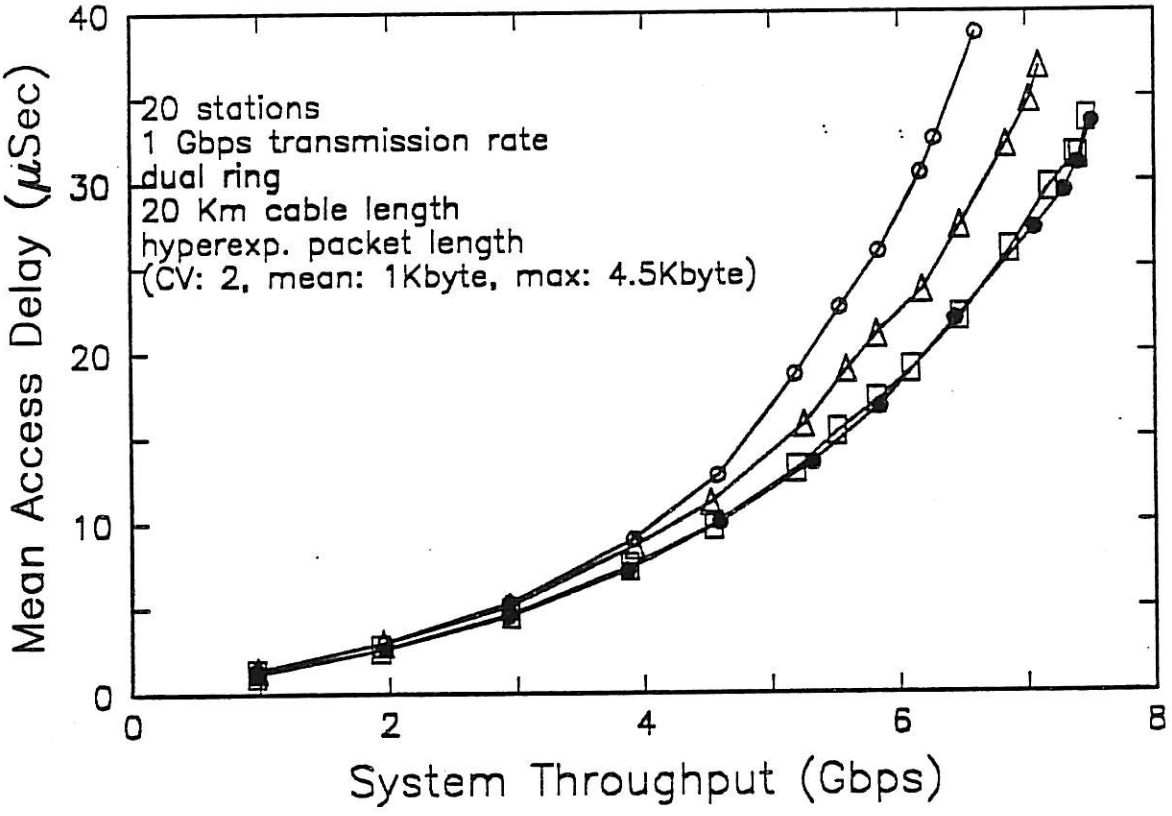
⊛ : No Fairness
○ : Q_{max} = 4.5 Kbyte, Q_{min} = 4.5 Kbyte
△ : Q_{max} = 4.5 Kbyte, Q_{min} = 1 Kbyte
□ : Q_{max} = 13.5 Kbyte, Q_{min} = 1 Kbyte



Results: Uniform Traffic

Mean Access Delay

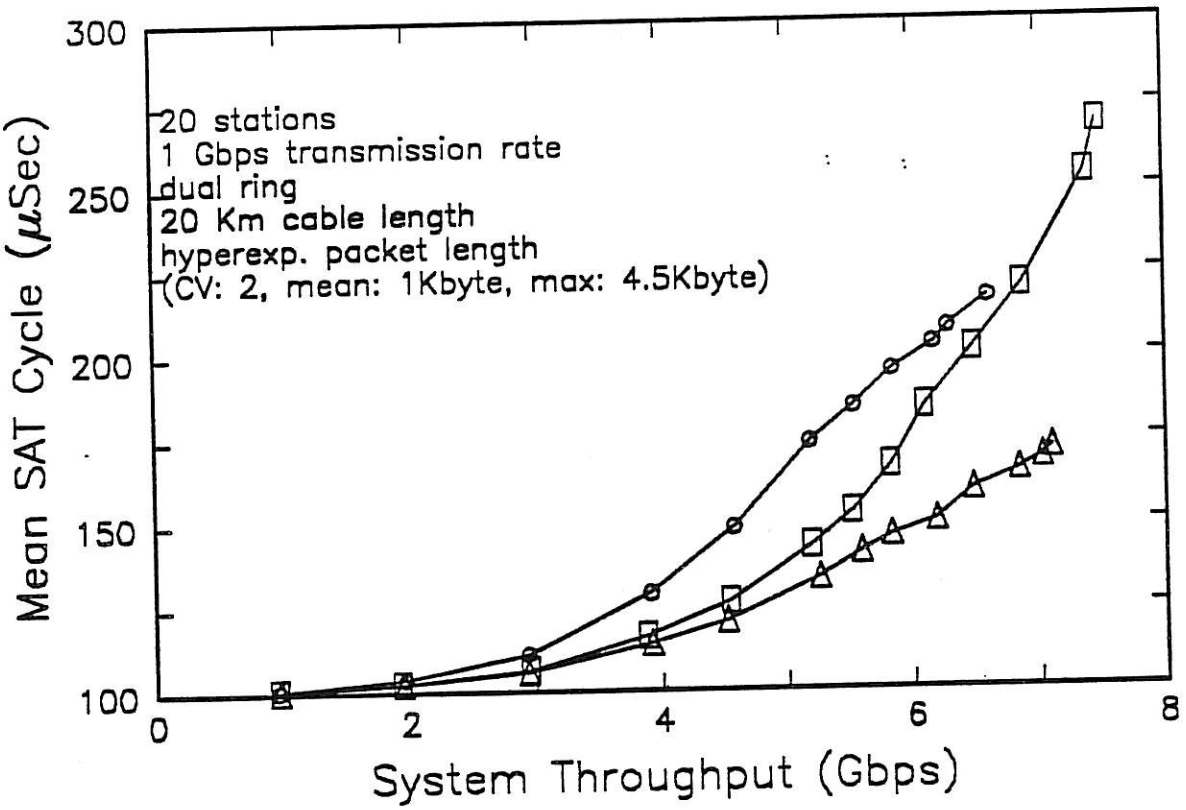
⊛ : No Fairness
○ : Q_{max} = 4.5 Kbyte, Q_{min} = 4.5 Kbyte
△ : Q_{max} = 4.5 Kbyte, Q_{min} = 1 Kbyte
□ : Q_{max} = 13.5 Kbyte, Q_{min} = 1 Kbyte



Results: Uniform Traffic

Mean SAT Cycle Time

⊛ : No Fairness
o : Q_{max} = 4.5 Kbyte, Q_{min} = 4.5 Kbyte
Δ : Q_{max} = 4.5 Kbyte, Q_{min} = 1 Kbyte
▯ : Q_{max} = 13.5 Kbyte, Q_{min} = 1 Kbyte



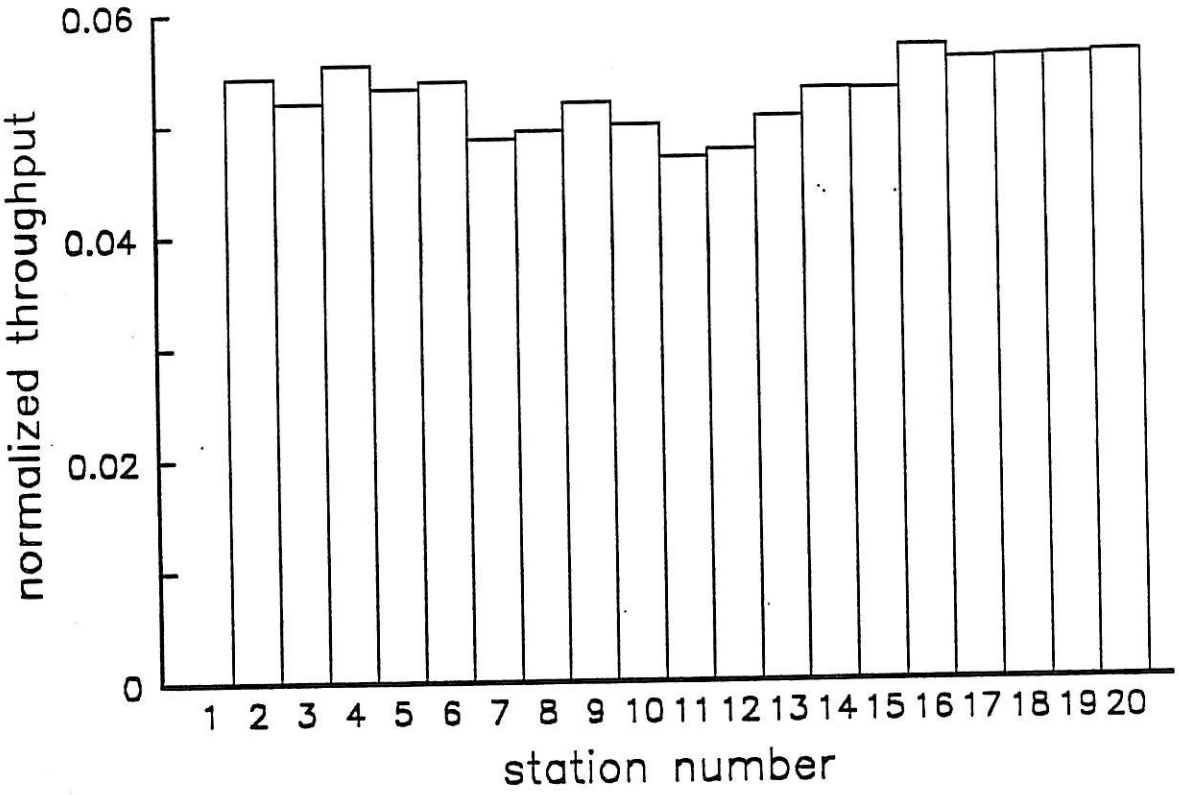
Simulation Scenarios

Fairness evaluation: Non-uniform destination distribution

- Station #1 never chosen as destination
- Other stations have uniform destination
- Uniform input
- Fully loaded
- Results:
  - Throughput distribution over user population
  - Access delay over user population

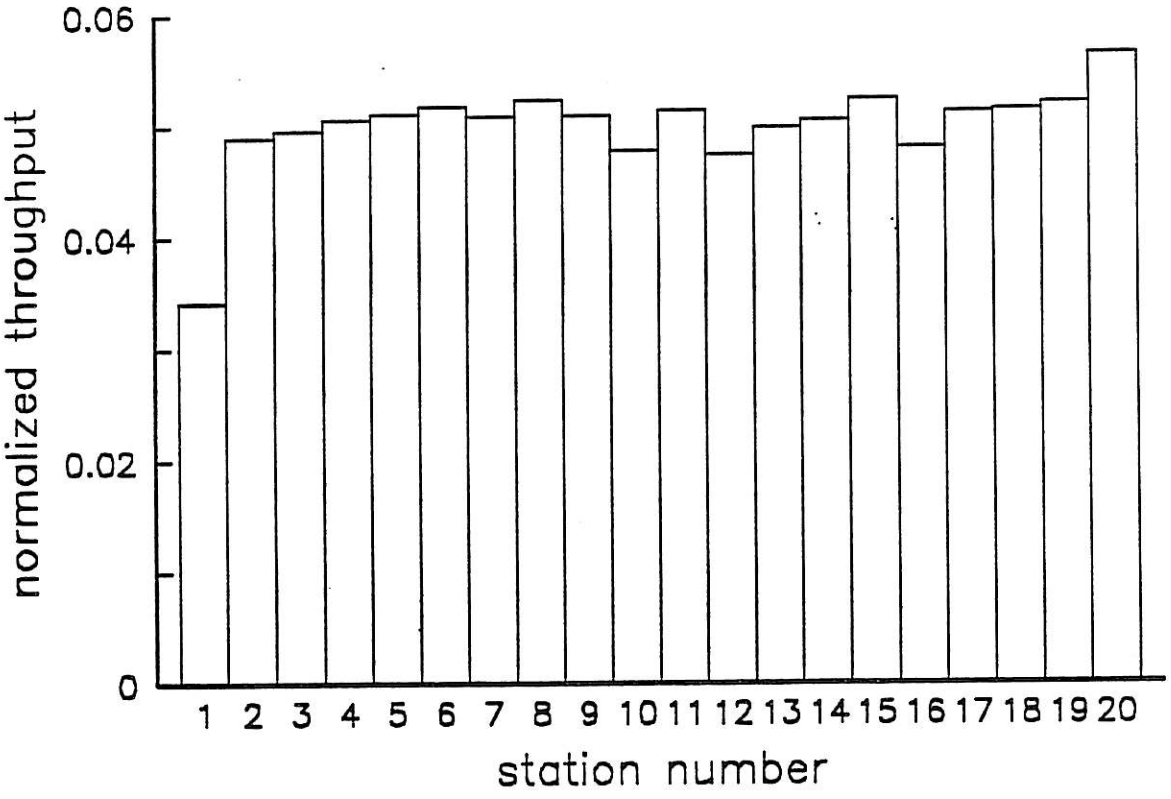
Results: Fairness Evaluation

- Station #1 never chosen as destination
- No fairness
- Aggregate throughput: 7.45 Gbps



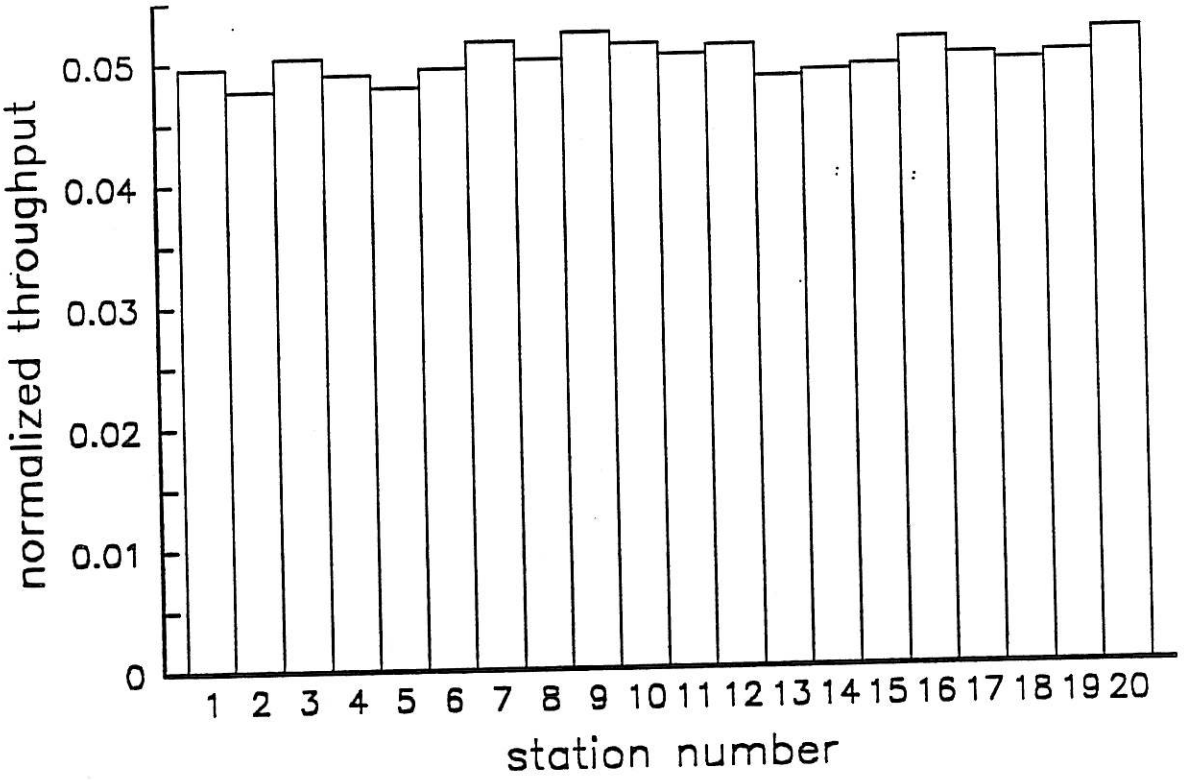
Results: Fairness Evaluation

- Station #1 never chosen as destination
- Q_{min}: 1 Kbytes, Q_{max}: 4.5 Kbytes
- Aggregate throughput: 6.9 Gbps



Results: Fairness Evaluation

- Station #1 never chosen as destination
- Q_{min}: 4.5 Kbytes, Q_{max}: 4.5 Kbytes
- Aggregate throughput: 6.51 Gbps



SUMMARY

- Performance results for Gbps LAN

  1. Buffer insertion ---> immediate access
  2. Spatial reuse ---> increase aggregate throughput
  3. A global control algorithm ---> fairness