光纤通道交换机在强实时约束下的分组调度

林 强 熊华钢 张其善

(北京航空航天大学电子信息工程学院 北京 100083)

摘 要 以光纤通道交换网络强实时约束下的性能研究为背景,采用实时通信中的周期性任务模型,提出了负载匹配的加权轮循分组调度,导出了在该方法下网络消息集严格实时的充要条件,以最差情形下强实时的网络可达负载率为性能衡量指标推证了采用该算法的优越性并通过仿真进行了验证.

关键词 光纤通道;交换结构;强实时约束;分组调度;负载率

中图法分类号 TP393

Packet Scheduling for Fibre Channel Switched Fabric Under Hard Real Time Constraints

LIN Qiang XIONG Hua-Gang ZHANG Qi-Shan

(*School of Electronic and Information Engineering, Behang University, Beijing* 100083)

Abstract Fibre channel is a computer communications protocol designed to meet many requirements related to the ever increasing demand for high performance information transfer. Switched fabric is one of basic fibre channel topology. Fixed-length packet scheme is mainly used in fibre channel switches. Taking the study on fibre channel switched fabric under hard real-time constraints as background, a new packet scheduling algorithm is proposed based on Weighted Round Robin. The necessary and sufficient conditions of guaranteeing message deadlines under the packet scheduling algorithm has been derived. Taking the achievable utilization of the worst case as the main index, the merit of the algorithm has also been derived and tested with simulation results.

Keywords fibre channel; switched fabric; hard real-time constraints; packet scheduling; utilization

1 引 言

光纤通道(fibre channel)协议是由 ANSI(American National Standards Institute)提出的新型串行通信标准. 光纤通道交换式结构(Fibre Channel Switched Fabric,FC-SW)提供了在交换式网络中把多个节点连接在一起的方法,是光纤通道中一种重要的拓扑结构. 商用领域中,FC-SW 的应用目前多集中于存储局域网(storage area network),在研究方法上侧重于平均消息延时、平均队列长度和吞吐量等网络平均统计特性方面的研究[1~5],但是在实时通信领域中,要求网络能够提供强实时约束下的消息传输. 如果消息传输超过时限,那么产生的延迟有可能会导致灾难性的后果.

由于光纤通道帧格式中数据域较大(2KByte),

收稿日期:2005-05-26;修改稿收到日期:2005-11-13. 本课题得到国家自然科学基金(10477005)资助.林 强,男,1973 年生,博士研究生,主要研究方向为宽带通信网络性能优化与仿真评价. E-mail: buaa.linqiang@163.com. 熊华钢,男,1961 年生,教授,博士生导师,主要研究领域为嵌入式计算机网络、电子综合化和数字通信. 张其善,男,1936 年生,教授,博士生导师,主要研究领域为信息处理、卫星通信和智能交通系统.

所以目前光纤通道交换机大都采用定长分组，将消息帧在交换节点中重新打包成信元，通过时隙在入线与出线间进行数据交换，参考文献[6,7]给出了交换内核采用 ATM 定长分组的设计方案. 在网络或分布式系统用于端到端调度时，WRR 是一个很好的实际选择，目前部分形式的 WRR 算法已经在 ATM 网络中得到实现[8]. 在上述研究背景下，本文根据 FC-SW 协议，提出了强实时约束条件下负载匹配的加权轮循分组调度算法；导出了在该方法下网络消息集严格实时的充要条件，以最差情形下保证消息集实时传输的网络可达负载率作为性能评价指标，与其它常见加权轮循分组调度算法进行了理论比较和仿真验证.

2　FC-SW 交换节点模型

2.1　消息模型

消息模型是采用实时通信中的周期性任务模型(periodic task model)[8]，交换机用交换节点表示，每个交换节点中有 n 条入线和 n 条出线，每个入线端均有一个实时消息流需要传送至出线；消息流优先级相同，按最大允许延迟时间由小到大顺序排列为 $S_1, S_2, \cdots, S_n$，它们组成一个消息集合 M，即

$$M = \{S_1, S_2, \cdots, S_n\} \tag{1}$$

对于消息流 S_i 和交换节点，有如下规定：

(1)消息流产生周期为 P_i. 它表示消息流 S_i 的消息产生周期. 对于非周期性消息，则表示消息产生最小时间间隔；

(2)消息流最大长度为 C_i. 它表示第 i 个消息流的传输时间，包括网络协议规定的分隔符、帧头、信息域和校验域等帧的全部内容；

(3)输出链路的轮长为 RL. 它表示交换节点在分组调度中每次轮循所允许的最大时隙数，分组调度轮循周期的上限；

(4)消息流最大允许延迟时间等于消息流产生周期 P_i；

(5)消息流产生周期 P_i 的最小值用 $P_{\min}$ 表示；

(6)消息流与交换节点参数均以时隙为基本时间单位进行归一化，时隙归一化结果为 1.

消息由一个二维数组表示：

$$S_i = (C_i, P_i) \tag{2}$$

消息流负载率 U_i 定义为

$$U_i = C_i / P_i \tag{3}$$

网络总的负载率为

$$U = \sum_{i=1}^{n} U_i \tag{4}$$

2.2　队列调度模型

进入交换节点的消息帧根据时隙大小重新打包，如图 1 所示. 设每个入线队列 Q_i 对应消息集 M 中一个消息 S_i 并假定其工作在最差情形下，即 n 条入线同时竞争一条出线. 在连接建立期间，各个交换上的调度程序为新连接 i 分配一个权值 wt_i.

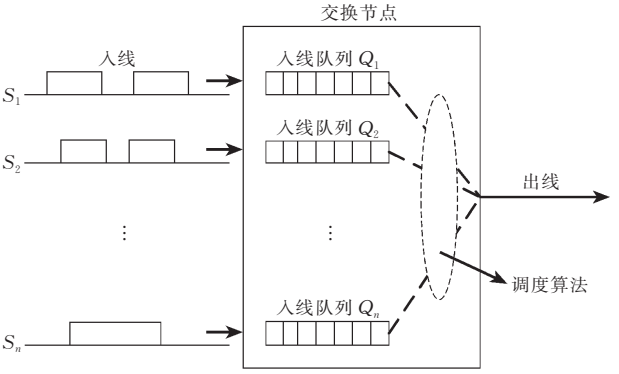

图 1　队列调度模型

3　实时分组调度

3.1　实时限制条件

在强实时条件下，分组调度算法为每个连接所分配的权值应同时保证消息传输的时限约束条件和轮循权值约束条件.

轮循权值约束条件：

忽略算法的建立开销，在分组调度中源节点用于发送消息的权值应满足

$$\sum_{i=1}^{n} wt_i \le RL < P_{\min} \tag{5}$$

时限约束条件：

对于任意时间间隔 t，用 $X_i(t)$ 表示输入队列 Q_i 发送消息的最小时间量. 消息集合 M 中每个消息在其最大允许延迟时间内，应有足够发送该消息的时间，因此对于任意消息流 S_i 应有

$$X_i(P_i) \ge C_i, \quad i = 1, 2, \cdots, n \tag{6}$$

上式称为消息的时限约束条件.

对于特定的消息流，如果所采用的分组调度算法既能满足轮循权值约束条件又能满足时限约束条件，那么在该分组调度算法下对该消息流可实现实时传输.

3.2　分组调度方法

对于轮循中的任意时间间隔 t，入线队列 Q_i 发

送消息的最小时间量可以表示为

$$X_i(t) = \lceil t/RL \rceil wt_i \tag{7}$$

从式(7)可知，$\lceil P_i/RL \rceil wt_i$ 为任意时间间隔 $(t_0, t_0+P_i]$ 内可用于发送消息的时间，C_i 为消息流传输时间，将两者相比较，在满足时限约束条件下取连接 i 的最小权值，则可得到 wt_i 的负载匹配的确定方法：

$$0 \le wt_i - \frac{C_i}{\lceil P_i/RL \rceil} < 1 \tag{8}$$

$\lceil P_i/RL \rceil$ 表示对括号内的 P_i/RL 取整，以下类同.

3.3 强实时的充要条件

引理 1. 对于任意给定的消息集 M，如果在分组调度中，入线连接权值 wt_i 满足 $0 \le wt_i - \frac{C_i}{\lceil P_i/RL \rceil} < 1$，则 $X_i(P_i) \ge C_i$.

证明. 根据式(8)可推得

$$X_i(P_i) = \lceil P_i/RL \rceil wt_i \ge \lceil P_i/RL \rceil \frac{C_i}{\lceil P_i/RL \rceil} = C_i.$$

证毕.

定理 1. 对于任意给定的消息集 M，如果入线连接权值 wt_i 满足 $0 \le wt_i - \frac{C_i}{\lceil P_i/RL \rceil} < 1$，则当且仅当 $\sum_{i=1}^{n} wt_i \le RL$ 时，M 中所有消息满足最大允许延迟时间要求.

证明. 由式(6)和引理 1 可证.

4 系统负载率分析

在分组调度算法作用下，如果网络负载率小于 U_A 时，网络中所有消息的实时性都能得到保证，那么称 U_A 为网络在该带宽分配方法下的可达负载率.

引理 2. 对于消息集 $\{S_n\}$，如果在分组调度中以下条件成立：$0 \le wt_i - \frac{C_i}{\lceil P_i/RL \rceil} < 1$，并且消息集负载率满足 $U < \frac{\lceil P_{\min}/RL \rceil}{\lceil P_{\min}/RL \rceil + 1}\left(1 - \frac{n}{RL}\right)$，那么 $\sum_{i=1}^{n} wt_i \le RL$.

证明. 由已知条件可得

$$\begin{aligned}\sum_{i=1}^{n} wt_i &< \sum_{i=1}^{n} \frac{C_i}{\lceil P_i/RL \rceil} + n \\ &= \sum_{i=1}^{n} \frac{C_i}{P_i} \frac{P_i/RL}{\lceil P_i/RL \rceil} RL + n \\ &\le \sum_{i=1}^{n} U_i \frac{\lceil P_i/RL \rceil + 1}{\lceil P_i/RL \rceil} RL + n.\end{aligned}$$

因为上式右边为 P_i/RL 的减函数，所以可以推出：

$$\begin{aligned}\sum_{i=1}^{n} wt_i &< \sum_{i=1}^{n} U_i \frac{\lceil P_{\min}/RL \rceil + 1}{\lceil P_{\min}/RL \rceil} RL + n \\ &\le \frac{\lceil P_{\min}/RL \rceil + 1}{\lceil P_{\min}/RL \rceil} RL \sum_{i=1}^{n} U_i + n \\ &= \frac{\lceil P_{\min}/RL \rceil + 1}{\lceil P_{\min}/RL \rceil} RL \cdot U + n \\ &< \frac{\lceil P_{\min}/RL \rceil + 1}{\lceil P_{\min}/RL \rceil} \times \\ &\quad \frac{\lceil P_{\min}/RL \rceil}{\lceil P_{\min}/RL \rceil + 1} RL\left(1 - \frac{n}{RL}\right) + n = RL.\end{aligned}$$

证毕.

定理 2. 对于消息集 $\{S_n\}$，如果分组调度中输入连接权值 wt_i 满足 $0 \le wt_i - \frac{C_i}{\lceil P_i/RL \rceil} < 1$，则网络最坏情况下的可达负载率满足

$$U_A^* < \frac{\lceil P_{\min}/RL \rceil}{\lceil P_{\min}/RL \rceil + 1}\left(1 - \frac{n}{RL}\right) \tag{9}$$

证明.

(1)由定理 1 和引理 2 知式(9)成立意味着消息实时性得到保证.

(2)下面证明对于任意给定实数 ξ，$0 < \xi < 1$，至少存在一个消息集 $\{S_i\}$ 的网络负载率 $U \le \frac{\lceil P_{\min}/RL \rceil}{\lceil P_{\min}/RL \rceil + 1}\left(1 - \frac{n}{RL}\right) + \xi$，使得 $\sum_{i=1}^{n} wt_i > RL$，即不满足协议限制条件.

对于给定的 RL，ξ，假设 $P_{\min} \ge 2RL$，$\alpha = \lceil P_{\min}/RL \rceil$，$\xi' = \min(\alpha + 1 - P_{\min}/RL, \xi)$，则 $\alpha \ge 2$，$0 < \xi' < 1$.

构造一个消息集，具有下列参数：

$$\begin{aligned}P_1 &= P_{\min}, \\ C_1 &= \xi' P_{\min}, \\ P_2 &= (\alpha + 1 - \xi')RL, \\ C_2 &= P_2 \frac{\alpha}{\alpha + 1}\left(1 - \frac{2}{RL}\right),\end{aligned}$$

则

$$\begin{aligned}U_1 &= \xi', \\ U_2 &= \frac{\alpha}{\alpha + 1}\left(1 - \frac{2}{RL}\right), \\ U &= U_1 + U_2 = \xi' + \frac{\alpha}{\alpha + 1}\left(1 - \frac{2}{RL}\right) \\ &\le \frac{\alpha}{\alpha + 1}\left(1 - \frac{2}{RL}\right) + \xi \\ &= \frac{\lceil P_{\min}/RL \rceil}{\lceil P_{\min}/RL \rceil + 1}\left(1 - \frac{2}{RL}\right) + \xi, \\ wt_1 &= \left\lceil \frac{C_1}{\lceil P_1/RL \rceil} \right\rceil + 1 \ge \left\lceil \frac{U_1 P_1}{\lceil P_1/RL \rceil} \right\rceil + 1\end{aligned}$$

$$= \left[\frac{\xi'}{\alpha}P_{\min}\right]+1,$$

$$wt_2 = \left[\frac{C_2}{\lceil P_2/RL\rceil}\right]+1 \geq \left[\frac{U_2P_2}{\lceil P_2/RL\rceil}\right]+1$$

$$= RL-2-\frac{\xi'}{\alpha+1}(RL-2)+1,$$

$$\sum_{i=1}^{2} wt_i \geq RL+\left(\left[\xi'\frac{P_{\min}}{\alpha}\right]-\frac{\xi'RL}{\alpha+1}\right)+\frac{2\xi'}{\alpha+1} \quad (10)$$

因为

$$\left[\xi'\frac{P_{\min}}{\alpha}\right]-\frac{\xi'RL}{\alpha+1} > \left[\xi'\frac{P_{\min}}{\lceil P_{\min}/RL\rceil}\right]-\frac{\xi'RL}{P_{\min}/RL}$$

$$\geq \left[\xi'\frac{P_{\min}/RL}{\lceil P_{\min}/RL\rceil}RL\right]-\frac{\xi'RL}{P_{\min}}RL$$

$$\geq \lceil\xi'RL\rceil-\frac{\xi'RL}{P_{\min}}RL$$

$$= \frac{\lceil\xi'RL\rceil}{\xi'RL}\xi'RL-\frac{RL}{P_{\min}}\xi'RL$$

$$\geq \left(\frac{\lceil\xi'RL\rceil}{\xi'RL}-\frac{1}{\alpha}\right)\xi'RL \quad (11)$$

因为 $1\leq\frac{\xi'RL}{\lceil\xi'RL\rceil}\leq 2$，所以 $\frac{1}{2}\leq\frac{\lceil\xi'RL\rceil}{\xi'RL}\leq 1$. 又由已知条件得 $\frac{1}{\alpha}\leq\frac{1}{2}$. 将这两个结果代入式(11)得

$$\left[\xi'\frac{P_{\min}}{\alpha}\right]-\frac{\xi'RL}{\alpha+1} > 0 \quad (12)$$

将式(12)代入式(10)得

$$\sum_{i=1}^{2} wt_i \geq RL+\left(\left[\xi'\frac{P_{\min}}{\alpha}\right]-\frac{\xi'RL}{\alpha+1}\right)+\frac{2\xi'}{\alpha+1} > RL$$

证毕.

对于一般交换节点，由定理 2 可知：当 $P_{\min}$ 趋近于 RL 时，U_A^* 趋近于最小值. 由于节点端口数和时隙带宽有限，对于大块(视频、音频)数据流的传送，$n \ll RL$，此时 $\min(U_A^*)\approx\frac{1}{2}$，即最坏情况下消息集强实时条件下的网络可达负载率约为 50%. 在工程设计的初始阶段，消息集合可能并不完备，此时网络负载率将成为系统设计所要考虑的主要因素，根据上述讨论结果，只要负载率小于 50%，可以保证所有同优先级消息的实时性要求.

5　计算机仿真实例

加权轮循分组调度中，权值分配的方法很多. 本文通过 C++ 建模，基于事件的调度方式对光纤通道交换型网络中的多个消息集进行了离散事件仿真实验. 将负载匹配的加权轮循分组调度与全负载和负载均衡两种情况下的加权轮循分组调度进行了实验比较. 全负载型权值分配的设计思想是期望源节点每次都能占用全部信道发送全部消息；负载均衡型权值分配的设计思想是将信道资源平均分配给每个源节点. 实验中，将消息延迟时间率作为衡量网络强实时性的指标，消息延迟时间率定义为消息的实际传输延迟时间与消息的最大允许延迟时间的比率. 最大延迟时间率、最小延迟时间率和平均延迟时间率分别表示消息传输的最大延迟时间、最小延迟时间和平均延迟时间与消息的最大允许延迟时间的比率. 实验结果如图 2～图 4 所示，表 1 给出了实验中的一组消息集. 表 2 给出了实验对比结果.

表 1　节点消息流特征

消息名	消息长度(Byte)	发送周期(ms)
A	1000	0.19
B	2000	0.38
C	2000	0.44
D	500	0.12
E	1000	0.19
F	1000	0.19
G	200	0.12
H	2000	0.44
I	200	0.12
J	1000	0.19

表 2　采用不同方式分配权值仿真结果比较

比较参数	仿真时间(ms)	传输消息总数	网络通信负载率(%)	最小延迟时间率(%)	最大延迟时间率(%)	平均延迟时间率(%)	超过最大允许延迟周期消息数
全负载	100	4491	70.33	6.734	105.224	19.068	16
负载均衡	100	4488	70.29	6.748	108.213	18.264	4
负载匹配	100	4489	70.28	6.727	95.155	19.351	0

从表 2 的结果可以看出，对于同一消息集，在相同的仿真时间内，三种权值分配方法的最大延迟时间率差别较大. 全负载和负载均衡的权值分配均使得消息的实际传输延迟时间有可能超过最大允许时间(最大延迟时间率分别为 105.224%，108.213%，超过最大允许延迟的周期消息数分别为 16 和 4)，而负载匹配的权值分配实验中，没有消息的实际传输延迟时间超过最大允许时间(最大延迟时间率为 95.155%，超过最大允许延迟的周期消息数为 0). 仿真实验结果表明基于负载匹配的加权轮循可以更好地满足交换网络端对端调度时的实时性要求.

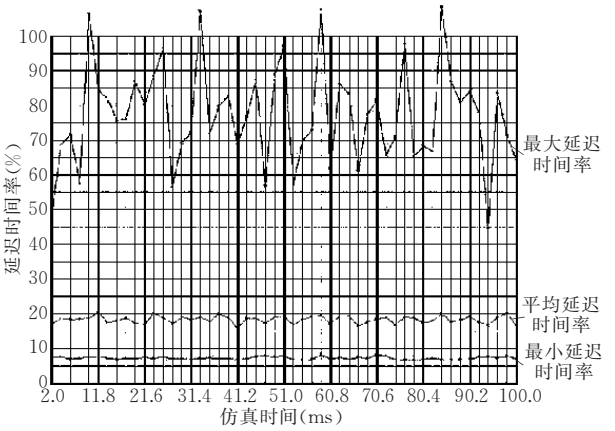

图 2 负载均衡分配权值情况下延迟时间率曲线

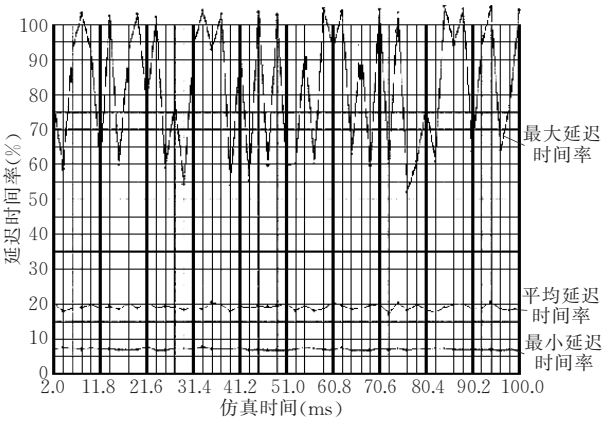

图 3 全负载分配权值情况下延迟时间率曲线

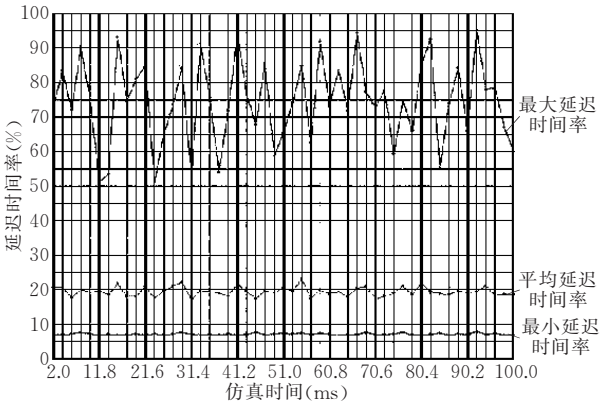

图 4 负载匹配分配权值情况下延迟时间率曲线

6 结束语

本文根据光纤通道交换型网络协议,在强实时约束的研究背景下,提出了以保证消息实时传输为目标的基于负载匹配的加权轮循分组调度;对该算法实时性能进行了理论分析,求出了采用此算法后在最坏情况下的可达负载率,为实时系统中交换网络的工程设计与优化提供了理论依据. 本文的研究成果是以光纤通道交换式网络环境为背景进行的,其设计思想和结论不失一般性,也适用于采用其它协议的交换网络结构在强实时约束下的优化设计.

参 考 文 献

1 Cherkasova L., Kotov V., Rokicki T.. Designing fibre channel fabrics. In: Proceedings of the Computer Design: VLSI in Computers and Processors, Austin, Texas, USA, 1995, 346～351

2 Cherkasova L., Kotov V., Rokicki T.. Fibre channel fabrics: Evaluation and design. In: Proceedings of the System Sciences, Maui, Hawaii, USA, 1996, 1: 53～62

3 Molero X., Silla F., Santonja V., Duato J.. On the switch architecture for fibre channel storage area networks. In: Proceedings of the Parallel and Distributed Systems, KyongJu City, Korea, 2001, 484～491

4 Varma A., Sahai V., Bryant R.. Performance evaluation of a high-speed switching system based on the fibre channel standard. In: Proceedings of the High Performance Distributed Computing, Spokane, Washington, USA, 1993, 144～151

5 Varma A., Murthy S., Bryant R.. Using camp-on to improve the performance of a fibre channel switch. In: Proceedings of the Local Computer Networks, Minneapolis, Minnesota, USA, 1993, 247～255

6 Anzaloni A., Agnitelli N., Avaltroni F.. Fiber channel FCS/ATM interworking: Design and performance study. In: Proceedings of the Global Telecommunications Conference, Houston, Texas, USA, 1994, 3: 1801～1807

7 Anzaloni A., De Sanctis M., Avaltroni F.. Fibre channel (FCS)/ATM interworking: A design solution. In: Proceedings of the Global Telecommunication Conference, Houston, Texas, USA, 1993, 2: 1127～1133

8 Liu Jane W. S.. Real-Time Systems. Beijing: Higher Education Press, 2002

LIN Qiang, born in 1973, Ph. D. candidate. His research interests include optimization and evaluation of digital communication network.

XUA Hua-Gang, born in 1961, Ph. D., professor. His research interests include embedded network, digital communication and integrated avionics system.

ZHANG Qi-Shan, born in 1936, professor, Ph. D. supervisor. His research interests include information processing, satellite communication and intelligent traffic system.

Background

Fibre Channel is a new serial communication protocol approved by ANSI. It is widely used in the domains of network and high speed bus gradually with its good-compatibility, high-speed and long-distance. The original fibre channel protocol cannot be directly exploited for avionics real-time systems because its extension ensures the delivery of messages rather than guaranteeing a deterministic latency. Suggestion have been made by fibre channel working group members on how to extend fibre channel to support real-time applications in avionics environments(FC-AE), but the problem of providing real-time capability in fibre channel network for general purpose is not solved satisfactorily and remains to be answered.

The authors have made researches on the fields of Avionics "Unified Network" design, the real-time performance and reliability of fibre channel. According to protocol and periodic task model, The problem to deliver periodic messages on time with minimum amounts of resources and good utilization of every resource is studied. A new packet scheduling algorithm is proposed based on node load matching. The necessary and sufficient conditions of guaranteeing message deadlines under the packet scheduling algorithm has been derived. The bound of the worst case achievable utilization for the algorithm has also been derived and formally proved. As a result, the above mentioned works lay a theory foundation for engineering applications of fibre channel switched fabric under hard real time constraints.