基于时间触发光纤通道网络的交换调度算法

白 焱[1,2]，孙万录[3*]，宋 平[3]，李 伟[3]

(1. 中国科学院沈阳计算技术研究所有限公司 沈阳 110042；2. 空装驻沈阳地区第一军事代表室 沈阳 110051；
3. 沈阳航盛科技有限责任公司 沈阳 110051)

【摘要】提出了一种基于时间触发的光纤通道网络数据交换调度算法，在基于端口序号进行轮询调度 (vp-RRM) 算法的基础上增加了流量自适应机制。该算法对光纤通道网络中的传输数据按 TT、RC、BE 等业务类型分队列缓存，将队列长度与交换调度的优先级建立关联，可明显改善非均匀业务流的交换调度效率。经仿真及实验验证，该算法吞吐量性能在非均匀业务流下较 vp-RRM 明显提升，更加适用于 TTFC 网络的事件触发业务的交换调度。

关 键 词 数据交换； 光纤通道； 调度算法； 时钟触发
中图分类号 TP301.6; V19 **文献标志码** A doi:10.12178/1001-0548.2021275

A Switching Scheduling Algorithm Based on Time Triggered Fibre Channel Network

BAI Yan[1,2], SUN Wanlu[3*], SONG Ping[3], and LI Wei[3]

(1. Postdoctoral of Shenyang Institute of Computing Technology, Chinese Academy of Sciences Shenyang 110042;
2. The First Military Representative Office of Air Equipment in Shenyang Shenyang 110051;
3. Shenyang Hangsheng Technology Co., Ltd. Shenyang 110051)

Abstract In this paper, a data exchange scheduling algorithm based on time trigger fibre channel network is proposed. Based on variable-length priority round Robin matching (vp-RRM), a traffic adaptive mechanism is added. The algorithm queues and caches the transmission data in the fibre channel network according to the service types such as time trigger (TT), rate-constrained (RC) and best-effort (BE), and correlates the queue length with the priority of switching scheduling, which can significantly improve the switching scheduling efficiency of non-uniform service flow. Through simulation and experimental verification, the throughput performance of the algorithm is significantly improved compared with VP-RRM under non-uniform traffic flow, and it is more suitable for the exchange scheduling of event triggered services in time trigger fiber channel (TTFC) network.

Key words data exchange; fiber-channel; scheduling algorithm; time trigger

时间触发光纤通道 (time trigger fiber channel, TTFC) 具有全局时钟和预定义的传输时间表[1]，通过解决一部分光纤通道 (fiber channel, FC) 网络数据交换的冲突问题，改善了重要业务的传输确定性和实时性，提升了航电任务系统的整体性能，成为了航电网络通信的发展方向。

TTFC 网络支持时间触发 (time trigger, TT) 和事件触发 (events trigger, ET) 多种优先级业务。TT 业务具有最高优先级，通过离线通信调度保证其准确性，TT 业务可在定义的延迟和抖动范围内在 TTFC 交换网络上无冲突地精确传送[2]。ET 业务按优先级从高到底可划分为流量控制 (rate-constrained, RC) 和尽力而为 (best-effort, BE) 两种不同的优先级业务[3]。TTFC 交换机是 TTFC 网络的关键组件，TTFC 交换机在传统 FC 交换机基础上增加了时钟触发 TT 业务的数据交换。传统 FC 交换机的实现由于受硬件的交换加速能力限制，大多采用无阻塞的交叉开关矩阵 (crossbar) 式直通交换结构[4]，并根据信元缓存位置的不同分为输入和输出两种排队方式。而输入队列会出现线头 (head of line, HOL) 堵塞现象，通常采用虚拟输出队列 (virtual output queue, VOQ) 的方法来解决 HOL 问题[5]。

收稿日期：2021－10－20；修回日期：2021－12－07
作者简介：白焱 (1987－)，男，博士，主要从事计算机技术与应用方面的研究.
*通信作者：孙万录，E-mail：sunwanlu@163.com

为提升交换转发效率，降低转发延迟，减少交换调度开销，FC 交换机采用针对多优先级的变长调度算法来实现交换调度。OSP (orthogonal subspace projection)、p-iDRR (prioritized iDRR) 等算法虽支持多优先级，但算法设计时采用固定长度的信元，其应用变长信元交换时会引入额外开销。RRM 算法在输出端指针同步时会增加性能的损耗等，这些问题在 vp-RRM 算法中已得到了很好地解决。

在 TTFC 网络中，由于 TT 业务的加入，占用了交换端口的部分流量，同时 TT 业务对各端口流量的占用不同，导致在进行 ET 业务数据交换时，端口流量也不均匀。目前对于 TTFC 网络的研究，大多从网络模型仿真的角度对网络的实现算法进行分析，暂没有文献结合 TTFC 网络特点，对 ET 业务的交换调度效率进行深入分析[6]。vp-RRM 算法基于端口序号进行轮询调度，每个队列的输出是无差别的，但流量的非均匀分布会导致个别队列等待时间较长，吞吐性能变差[7]。因此，本文在加权轮询调度算法的基础上，提出了基于流量自适应的多优先级变长轮询调度算法 (traffic adaptive variable-length priority round robinmatching, tavp-RRM)，该算法针对非均匀流量状态下的交换调度算法进行改进，使流量大的端口得到更多的调度机会，以此提升网络的吞吐效率。

1　tavp-RRM 算法

tavp-RRM 算法可基于端口流量队列的长度自适应调整其请求优先级，使负载量大的端口优先进行数据交换，从而提升 TTFC 网络的吞吐率。其优点包括：支持非定长信元、非固定时隙、支持单播和广播队列、采用流水线工作和使用 VOQ 等。

1.1　tavp-RRM 算法原理

算法原理如下：

1) 为了提升 TTFC 网络的有序性，广播和组播业务均规划为 TT 业务，ET 业务仅处理单播业务，因此在本文中，不讨论该算法对组播和广播业务的适用性。

2) 支持多优先级调度。TTFC 网络的 ET 业务按优先级分为 RC 和 BE 两种业务。如图 1 所示，业务信元传输至 TTFC 交换机后，将依据其优先级存入各个 VOQ 中。VOQ 向输出端口发起请求时，VOQ 的优先级会根据其队列长度及等待时间进行调整。优先级从高到低依次为 P3、P2、P1、P0，RC 业务的起始优先级为 P2，可提升的优先级为 P3，BE 业务的起始优先级为 P0，可提升的优先级为 P1。

3) 输出端口同时收到多个输入端口的请求时，对优先级最高的输入端口先进行授权。

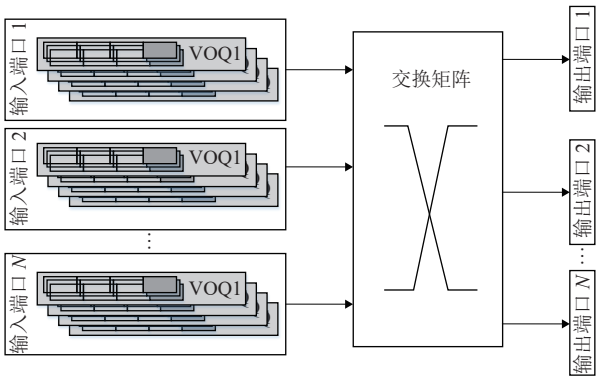

图 1　VOQ 架构示意图

tavp-RRM 调度算法在研发阶段即注重硬件的可实现性，经过多次迭代，完成输入、输出端口的最优匹配[8]。通过在每个输入、输出端口设置多个仲裁器，通过循环优先级仲裁来依次匹配所有有效的输入和输出，以保证每一次迭代的独立性。

1.2　tavp-RRM 算法执行过程

tavp-RRM 算法的执行过程与 RRM 类似，分为“请求–授权–接受”，步骤如下：

1) 请求

输入端口接收到信元后，会向信元的输出端口发起包含当前数据优先级的请求，请求的优先级由 3 种因素决定，分别为 ET 业务的优先级、当前 VOQ 的队列长度及当前 VOQ 的等待时间。图 2 给出了 ET 业务 VOQ 队列优先级的确定流程。

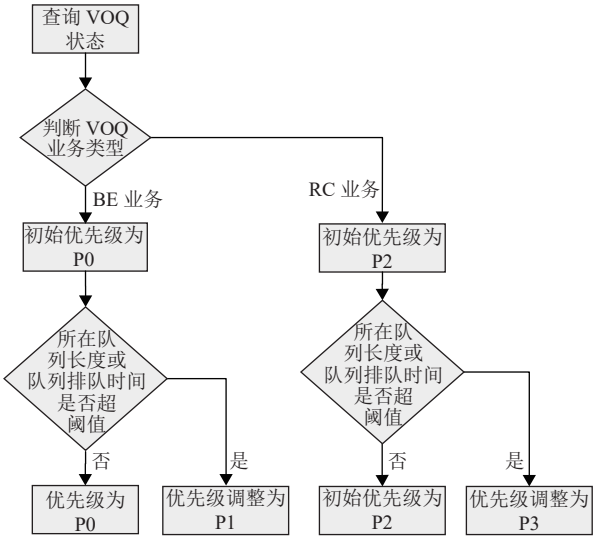

图 2　ET 业务 VOQ 队列优先级的确定流程

基于以下原则来确定阈值：

① 队列长度的阈值点距离队列满不小于 2 个 FC 最长帧的空间，防止阈值生效太晚，从而导致对外部输入端产生流控；

② 队列长度的阈值点应大于队列半满的位置，避免阈值被频繁触发；

③ 队列排队时间阈值应小于系统可接收的最大延迟。

本文在仿真及实验验证过程中，以队列深度的 3/4 作为队列长度阈值，以 100 μs 作为队列排队时间阈值。

2) 授权

每个输出端口处设有 1 个调度器，调度器通过轮询调度算法来匹配多个端口的请求，在均匀业务状态下具有良好的公平性。由于输入端的业务数据按照业务类型 (RC 或 BE) 分配队列缓存，因此每个输出端口对不同类型的业务通过轮询指针 r0 和 r1 分别维护其调度状态。r1 对应 RC 业务，优先级为 P2 或 P3；r0 对应 BE 业务，优先级为 P0 或 P1。调度器进行调度时，先过滤出当前最高优先级的请求，在根据 r0 或 r1 的当前位置按顺序开始轮询查找，对找到的第一个请求进行授权，并通过 r0 或 r1 更新到当前授权位置的下一个位置处。

3) 接受

输入端口会同时向多个输出端口发起调度请求，当输入端口同时接收到多个端口的授权时，同一时刻只能接受一个授权。因此，输入端口需要对接收到的授权进行仲裁来选择最终接受方。输入端口的接受仲裁对不同类型的业务通过轮询指针 c0 和 c1 分别维护其仲裁状态。c1 对应 RC 业务，优先级为 P2 或 P3；c0 对应 BE 业务，优先级为 P0 或 P1。仲裁器进行仲裁时，先过滤出当前最高优先级的请求，在根据 c0 或 c1 的当前位置按顺序开始轮询查找，对找到的第一个授权进行接受，并通过 c0 或 c1 更新到当前授权位置的下一个位置处。图 3 为一个 4×4 端口的调度过程，在请求授权阶段，输出端口 1 在优先级 P2、P3 之间选择高优先级 P3 对应的输入端口 3；输出端口 2 在 3 路 BE 业务中选择输入端口，输入端口 1 由于其队列长度大于门限值，请求将优先级调整为为 P1，因此选择高优先级 P1 对应的输入端口 1；输出端口 3 在两个相同优先级 P0 端口间选择，此时输出端口 3 对应 BE 业务的 r0=3，按照轮询机制，最终选择输入端口 4 进行授权；输出端口 4 在优先级 P1、P0 间选择高优先级 P1 对应的输入端口 1。

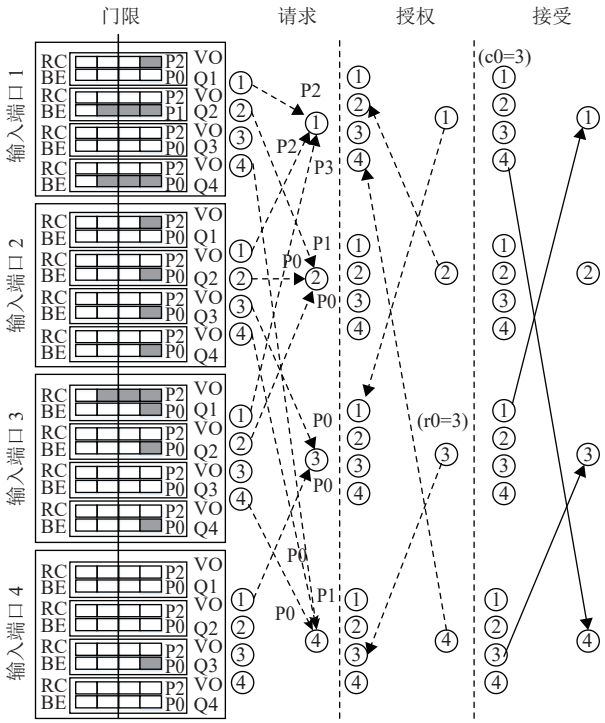

图 3 4×4 端口的调度过程

接受阶段，输入端口 1 同时收到输出端口 2 和输出端口 4 的授权，由于两个授权的优先级相同均为 P1，按照轮询机制从当前 BE 的指针 c0=3 启动轮询。输入端口 1 优先查询到输出端口 4 的授权，因此选择输出端口 4 的授权进行接受。输入端口 3 和 4 由于仅接收了 1 个授权，直接接受即可。

2 仿真分析

使用计算机仿真可以大幅降低 crossbar 调度算法性能分析的难度，本节将通过计算机仿真来分析 tavp-RRM 算法的吞吐量。吞吐量是指输出端口的平均利用率，即在一个时隙内平均发送的信元数量。在分析吞吐量性能时，本文将 vp-RRM 算法作为比较的对象，vp-RRM 算法和 tavp-RRM 的差异在于对优先级的管理不同，下面选取均匀业务流和非均匀业务流对两种算法的吞吐量性能进行对比。

1) 均匀业务流

输入端负载均相同，即$\gamma_i = \gamma$，所有业务流在 VOQ 中均匀分布。均匀业务流模型下，tavp-RRM 和 vp-RRM 的吞吐量均为相同的恒定值：

$$\gamma_{i,j} = \gamma \frac{1}{N} \quad 0 \leqslant i,\ j \leqslant N-1 \tag{1}$$

2) 非均匀业务流

采用如下两种非均匀业务流模型来研究 tavp-RRM 算法的吞吐量。

模型 1：输入端负载均相同，即$\gamma_i=\gamma$，业务流在 VOQ 中的分布方式为：

$$\gamma_{i,j}=\begin{cases}\gamma\delta & j=i\\ \gamma(1-\delta) & j=(i+1)\bmod N\end{cases}\quad 0\leqslant i,j\leqslant N-1 \tag{2}$$

式中，$\gamma_{i,j}$表示从输入端 i 到输出端 j 的负载；δ为非均衡 (unbalanced) 因子($0\leqslant\delta\leqslant 1$)。

业务流模型 1 的算法吞吐量对比如图 4 所示，根据式 (2) 可推算得知，1 个输出端口同时只转发两个输入端口的数据，在交换调度时竞争较少，转发效率本来较高，基于端口流量队列的长度自适应调整其请求优先级的功能 (traffic adaptive, TA) 作用不明显，vp-RRM 和 tavp-RRM 仿真结果基本一致。

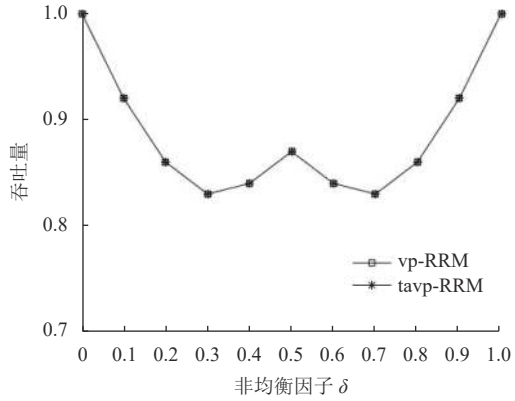

图 4　业务流模型 1 的算法吞吐量对比

模型 2：输入端负载均相同，即$\gamma_i=\gamma$，业务流在 VOQ 中的分布方式为[9]：

$$\gamma_{i,j}=\begin{cases}\gamma\left(w+\dfrac{1+w}{N}\right) & j=i\\ \gamma\dfrac{1-w}{N} & j\neq i\end{cases}\quad 0\leqslant i,j\leqslant N-1 \tag{3}$$

式中，w为非均匀 (nonuniform) 因子($0\leqslant w\leqslant 1$)，当$w=0$时，该业务流就是均匀业务流。业务流模型 2 的算法吞吐量对比如图 5 所示。

图 5 展示了在 16×16 的 crossbar 中 tavp-RRM 和 vp-RRM 算法的吞吐量。从图中可以看到，当 w=0 时，业务流为均匀业务流，tavp-RRM 算法的吞吐量与 vp-RRM 算法一致；当 w=1 时，业务流集中到同一个端口，达到满额负荷，tavp-RRM 算法的吞吐量与 vp-RRM 算法相同。其他状态下 tavp-RRM 算法的吞吐量性能始终优于 vp-RRM，且 w 在 0.5 附近时，性能差异达到最大。图 5 的仿真结果说明 tavp-RRM 算法相对于 vp-RRM 算法，在多个输入端口向同一个输出端口发生转发竞争，且流量不均匀时，具有更高的交换调度效率。

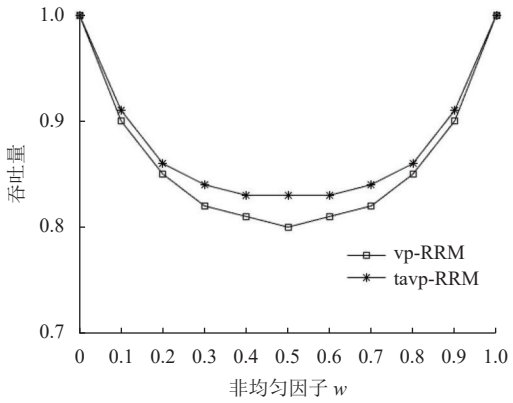

图 5　业务流模型 2 的算法吞吐量对比

3　实验验证

本文采用复旦微电子公司的 JFM7K325TFPGA 设计了 4×4 端口，速率为 2.125 Gb/s 的 TTFC 交换机。为了进行算法对比测试，TA 功能支持通过软件配置进行使能或关闭。

通过对交换机的 TT 业务转发进行规划，使交换机各端口的 TT 业务占用不同的带宽。TT 业务在端口 1 至端口 4 的输入带宽占用分别为 9%、12%、15% 和 18%，输出带宽在非输入端口均匀分布。

使用 JSDU 公司的 XgigLoadTester 测试仪对 TTFC 交换机进行 ET 业务的吞吐量测试，由于测试仪不支持 TT 业务注入，为了在测试中模拟 TT 业务引入的非均匀性，TTFC 交换机采用如下设计：初始化 TTFC 交换机各端口的 TT 调度配置，并在 TT 业务的转发时隙，强制 TT 转发路径处于忙状态，时隙内不进行 ET 业务的转发[10]。

具体测试说明如下：

1) 各端口激励数据均采用帧长为 1024 B 的 ET 帧；

2) 测试拓扑采用图 6 所示的不带自环的全网络拓扑模型。受测试仪功能限制，VOQ 分布采用均匀分布，通过 TT 业务的影响引入非均匀因素；

3) 各端口测试激励初始为 100% 负载；

4) 分别在 TA 功能使能和关闭的状态下进行测试。

各端口在未开启 TA 功能和开启 TA 功能后的对比如表 1 所示。tavp-RRM 算法的实测吞吐量性能优于 vp-RRM，吞吐量提升在 5% 以上，表明 TA

功能对于非均匀流量环境下的吞吐量有改善作用。

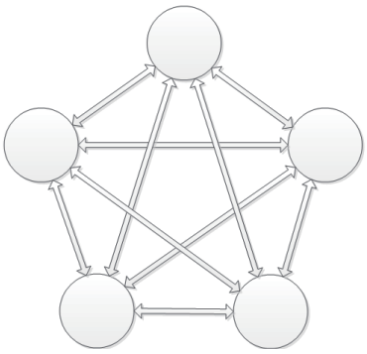

图 6 测试拓扑模型

表 1 实测结果对比

端口序号	端口吞吐量(未开启TA功能) / MB·s^{-1}	端口吞吐量(开启TA功能)/ MB·s^{-1}
1	291.426	312.749
2	281.818	302.439
3	272.211	292.129
4	262.603	281.818

4 结束语

本文根据 TTFC 网络中 ET 业务流量不均匀的特点，提出一种多优先级交换调度方案，基于端口流量队列长度进行优先级的自适应调整，使带宽资源紧张的 VOQ 得到更多的授权机会，从而改善 TTFC 网络的吞吐量性能，提高交换调度效率，更加适用于 TTFC 网络的事件触发业务的交换调度。

参 考 文 献

[1] KIRNER R, PUSCHNER P. A quantitative analysis of interfaces to time-triggered communication buses[J]. IEEE/ACM Transactions on Networking, 2021, 29(4): 1786-1797.

[2] FINZI A, ZHAO L X. Impact of AS6802 synchronization protocol on time-triggered and rate-constrained traffic[C]//32nd Euromicro Conference on Real-Time Systems (ECRTS 2020). Dagstuhl: [s.n.], 2020, 165: 17: 1-17, 22.

[3] SAE Technical Standard. SAE AS6802[S]. SAE Aerospace Standard. 2011(11): 1-108

[4] 彭来献, 田畅, 郑少仁. 高速交换网络的建模与仿真[J]. 系统仿真学报, 2003, 15(10): 1474-1476.
PENG L X, TIAN C, ZHENG S R. Modeling and simulation for high-speed switching fabrics[J]. Journal of System Simulation, 2003, 15(10): 1474-1476.

[5] LIU Y L, ZHANG H, JIANG N. Design and performance analysis of fibre channel network switch based on link aggregation optimization port circuit[J]. Journal of Nanoelectronics and Optoelectronics, 2020, 15(9): 1137-1145, 1149.

[6] MA X F, HAMDULLA A. Hybrid scheduling technology of time-triggered ethernet switches: A review[J]. Journal of Physics: Conference Series, 2020, 1673(1): 012024.

[7] 张建东, 吴勇, 史国庆, 等. 光纤通道交换网络 WRR 实时调度算法分析[J]. 航空学报, 2012, 33(2): 306-314.
ZHANG J D, WU Y, SHI G Q, et al. Analysis of fibre channel switching network WRR real-time scheduling algorithm[J]. Acta Aeronautica et Astronautica Sinica, 2012, 33(2): 306-314.

[8] 孙雪, 曹素芝, 许辉. FC 交换机中多优先级变长 CROSSBAR 调度策略[J]. 光通信技术, 2018, 10(15): 15-19.
SUN X, CAO S Z, XU H. Multi-priority variable-length CROSSBAR scheduling strategy in FC switch[J]. Optical Communication Technology, 2018, 10(15): 15-19.

[9] 彭来献, 田畅, 赵文栋. 一种具有 $O(\log N)$ 信息复杂度的高速 crossbar 调度算法[J]. 电子学报, 2016, 11: 2024-2029.
PENG L X, TIAN C, ZHAO W D. A new scheduling algorithm with $O(\log N)$ control messages complexity for high-speed crossbars[J]. Acta Electronica Sinica, 2016, 11: 2024-2029.

[10] 谢军, 涂晓东, 孟中楼. 多用途光纤通道交换机的设计与实现[J]. 计算机研究与发展, 2011, 48: 335-339.
XIE J, TU X D, MENG Z L. Design and implementation of multi-purposefibre channel switch[J]. Journal of Computer Research And Development, 2011, 48: 335-339.

编 辑 叶 芳