星载 Clos 网络的全分布式容错调度算法

刘 凯[①③] 晏 坚[*②] 高晓琳[①④] 陆建华[①]
[①](清华大学电子工程系 北京 100084)
[②](清华大学宇航中心 北京 100084)
[③](清华大学深圳研究生院 深圳 518055)
[④](北京航天飞行控制中心 北京 100094)

摘 要：针对星载交换结构受空间辐射影响造成的可靠性严重下降问题，该文提出了一种支持全分布式调度的三级 Clos 网络及其全分布式容错(Fully Distributed Fault Tolerant, FDFT)调度算法，以提高星载交换结构在交叉点故障下的容错能力。该 Clos 网络的中间级和输出级采用联合输入交叉点队列，以支持 Clos 网络和交换单元内部的全分布式调度。FDFT 采用一种分布式故障检测算法获得交叉点故障信息。基于对交叉点故障影响范围的分析，FDFT 在输入级采用一种容错信元分发算法，实现无故障路径的负载均衡。理论分析证明，当任一输入/输出级交换单元故障个数不超过 $(m-n)$ 或所有中间级交换单元故障个数不超过 $(m-n)$ 时，其中 m, n 分别为输入级交换单元输入、输出端口数，FDFT 能够达到 100%吞吐率。仿真结果进一步验证，故障随机发生情况下，FDFT 能够抵抗比故障任意发生情况下更多的故障，且在不同的业务场景下具有良好的吞吐率和时延性能。

关键词：星载交换；Clos 网络；全分布式；容错调度；负载均衡
中图分类号：TN927 **文献标识码**：A **文章编号**：1009-5896(2016)06-1377-08
DOI: 10.11999/JEIT150944

Fully Distributed Fault Tolerant Scheduling for Onboard Clos-network Switching

LIU Kai[①③] YAN Jian[②] GAO Xiaolin[①④] LU Jianhua[①]
[①](*Department of Electronic Engineering*, *Tsinghua University*, *Beijing* 100084, *China*)
[②](*Tsinghua Space Center*, *Tsinghua University*, *Beijing* 100084, *China*)
[③](*Graduate School at Shenzhen*, *Tsinghua University*, *Shenzhen* 518055, *China*)
[④](*Beijing Aerospace Control Center*, *Beijing* 100094, *China*)

Abstract: For an onboard switching, serious decline in the reliability is induced by the harsh space radiation environment. In this paper, a 3-stage Clos-network supporting fully distributed scheduling and a Fully Distributed Fault Tolerant (FDFT) scheduling algorithm are proposed to improve fault-tolerant ability of an onboard switching. Combined input and output queued architecture is employed in the central and output stages of the proposed Clos-network to support fully distributed scheduling in both the network and switching elements. In FDFT, a distributed fault detection algorithm is employed to obtain the crosspoint fault information. Based on the analysis of the influence of the faults, a fault-tolerant cell dispatching algorithm is proposed in the input stage which achieves load-balancing to fault-free paths. Theoretical analysis demonstrates that 100% throughput is achieved when no more than $(m-n)$ crosspoint faults occur in any input/output module or in all central modules, where m and n are the number of inputs and outputs of input module, respectively. Furthermore, simulation results indicate that, in the case of faults occurring randomly, FDFT tolerates much more faults, and exhibits a good performance in terms of throughput and average cell delay under different traffic scenarios.

Key words: Onboard switching; Clos-network; Fully distributed; Fault-tolerant scheduling; Load balancing

1 引言

为了满足下一代宽带卫星网络对带宽密集型业务的支持[1]，星载交换技术(OnBoard Switching, OBS)受到越来越多的重视。相比于传统的透明转发，OBS 能够提高系统的带宽利用率，通过一跳建立全互联卫星网络[2,3]。作为 OBS 的核心技术，星载交换结构用于提供上行链路与下行链路的交换通路，其性能决定了 OBS 的容量、吞吐率、时延等性能。相比于地面交换结构，星载交换结构会受到空间辐射影响，如总剂量效应、单粒子翻转、单粒子

收稿日期：2015-08-19；改回日期：2016-01-20；网络出版：2016-03-14
*通信作者：晏坚 yanjian_ee@tsinghua.edu.cn
基金项目：国家自然科学基金(91338108, 91438206)，中国电子科技集团校企合作基金(空间互联网关键技术)
Foundation Items: The National Natural Science Foundation of China (91338108, 91438206), China Electronics Technology Group School-Enterprise Cooperation Foundation (Key Technology of Space Internet)

功能中断等[4]。空间辐射效应会降低星载交换结构的可靠性。因此，星载交换结构需要比地面交换结构更高的可靠性。另外，空间辐射效应的影响范围随器件的工作频率升高而扩大[4]。在交换结构中，工作频率主要体现在调度算法的时间复杂度和队列加速比(队列的读取速率与输入数据速率的比值)。因此，为降低空间辐射效应的影响，需要研究低加速比和低时间复杂度的容错调度算法。

现有的星载交换结构主要有共享存储器，crossbar, knock-out 和 Clos 网络等[3,5]。共享存储器，crossbar 和 knock-out 等交换结构的单通路特性使得任意一条路径故障都会造成相应的输入/输出间无法交换，可靠性低。Clos 网络的任意输入/输出间存在多条路径，提高了可靠性，因此 Clos 网络更适用于星载交换结构。

Clos 网络内各交换单元间的分布式调度能够降低调度算法的时间复杂度。根据 Clos 网络三级是否采用缓存，Clos 网络分为三级无缓存(Space-Space-Space, S^3)型、输入级和输出级有缓存(Memory-Space-Memory, MSM)型、三级有缓存(Memory-Memory-Memory, MMM)型和中间级、输出级有缓存(Space-Memory-Memory, SMM)型。S^3 型 Clos 网络需要复杂的集中式调度算法解决端口竞争及路径选择问题[6]。另外，为实现故障条件下的路径选择，调度算法的复杂度会进一步增大[7]。MSM 型 Clos 网络的调度算法需要解决中间级的端口竞争，无法实现交换单元间分布式调度，并且输入级和输出采用共享存储的结构，加速比大于 1，不适用于星载交换结构[8]。MMM 型 Clos 网络能够实现交换单元间分布式容错调度，但每一级都添加缓存会造成实现成本上升[9,10]。文献[11]提出一种 SMM 型 Clos 网络，输入级采用 DSRR(Desynchronized Static Round-Robin)信元分发算法，达到同 MMM 型 Clos 网络相同的吞吐率和交换单元间的分布式调度，因此 SMM 型 Clos 网络更适用于星载交换结构。

交换单元内部的分布式调度能够进一步降低空间辐射效应的影响。SMM 型 Clos 网络中间级和输出级可采用不同的队列结构。文献[11]采用输出排队(Output Queuing, OQ)，但 OQ 的加速比大于 1。为降低加速比，文献[12]提出采用输入排队(Input Queuing, IQ)，但 IQ 需要在交换单元内部采用集中式调度，时间复杂度高。文献[13,14]提出采用交叉点排队(Crosspoint Queuing, CQ)，但 CQ 需要较大的交叉点队列空间以满足吞吐率要求(达到 100%吞吐率需要交叉点队列长度大于 32cells)，提高了星载器件的实现复杂度。另外，现有的 SMM 型 Clos 网络调度算法不能规避故障路径，无法发挥 Clos 网络的容错能力。

联合输入交叉点队列(Combined Input and Crosrpoint Queued, CICQ)具有分布式调度和低交叉点队列长度的特点[15]。本文提出在中间级和输出级采用 CICQ 结构的 CICQ-SMM Clos 网络，以实现 Clos 网络和交换单元内部的全分布式调度。为抵抗交叉点故障的影响，提出一种全分布式容错(Fully Distributed Fault Tolerant, FDFT)调度算法。FDFT 通过分布式故障检测(Distributed Fault Detection, DFD)算法得到各级交换单元的故障信息，在输入级采用 FT-DSRR(Fault Tolerant Desynchronized Static Round-Robin)容错信元分发算法，实现无故障路径的负载均衡。理论和仿真分析证明，FDFT 在不同的业务场景和故障模型下具有良好的性能和容错能力。

2 CICQ-SMM Clos 网络及故障模型

如图 1 所示，CICQ-SMM Clos 网络 $C(n,m,k)$ 的输入级包括 k 个 $n\times m$ 的输入模块(Input Module, IM)；中间级包括 m 个 $k\times k$ 的中间模块(Central Module, CM)；输出级包括 k 个 $m\times n$ 的输出模块(Output Module, OM)。每个 IM/OM 连接 n 个输入/输出，相邻两级的交换单元通过唯一通路连接。交换结构的尺寸为 $N=nk$。IM 采用空分结构，CM 和 OM 采用 CICQ 结构。到达的分组在输入端分割成定长信元(Cell)，通过 CICQ-SMM Clos 网络后在输出端口进行重组，交换一个信元所需的时间称为一个时隙(Slot)。为方便后续讨论，本文采用如表 1 所示的符号表示。

在一个交换单元(Switching Element, SE)内，任意输入/输出间的交换关系由相应的交叉点状态决定。空间辐射效应会造成交换单元内交叉点处于错误状态，导致交换错误，该故障称为交叉点故障(Crosspoint Fault)。在 Clos 网络中，交叉点故障发

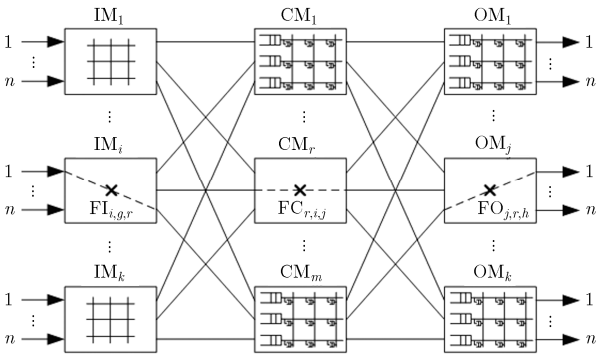

图 1 CICQ-SMM Clos 交换结构及不同交叉点故障

表 1 符号表示

符号	定义
IM_i	第 i 个 IM，$1 \le i \le k$
CM_r	第 r 个 CM，$1 \le r \le m$
OM_j	第 j 个 OM，$1 \le j \le k$
$\mathrm{I}_{i,g}$	IM_i的第 g 个输入，$1 \le g \le n$
$\mathrm{O}_{j,h}$	OM_j的第 h 个输出，$1 \le h \le n$
$\mathrm{IL}_{i,r}$	IM_i的第 r 个输出，连接 CM_r
$\mathrm{CL}_{r,j}$	CM_r的第 j 个输出，连接 OM_j
$\mathrm{FI}_{i,g,r}$	发生在 IM_i的(g,r)交叉点的故障
$\mathrm{FC}_{r,i,j}$	发生在 CM_r的(i,j)交叉点的故障
$\mathrm{FO}_{i,g,r}$	发生在 OM_j的(r,h)交叉点的故障
$\mathrm{P}_{i,g,r,j,h}$	路径 $\mathrm{I}_{i,g}\rightarrow\mathrm{IL}_{i,r}\rightarrow\mathrm{CM}_r\rightarrow\mathrm{CL}_{r,j}\rightarrow\mathrm{O}_{j,h}$

生在不同位置会造成不同的影响。如图 1 所示，当 IM_i中发生故障 $\mathrm{FI}_{i,g,r}$时，$\mathrm{I}_{i,g}$ 无法通过 CM_r到达 $\mathrm{O}_{j,h}$ $(1 \le j \le k,\ 1 \le h \le n)$，即路径 $P_{i,g,r,j,h}$ $(1 \le j \le k,\ 1 \le h \le n)$为故障路径；当 CM_r 中发生故障 $\mathrm{FC}_{r,i,j}$ 时，$\mathrm{I}_{i,g}$ $(1 \le g \le n)$无法通过 CM_r 到达 $\mathrm{O}_{j,h}$ $(1 \le h \le n)$，即路径 $\mathrm{P}_{i,g,r,j,h}(1 \le g \le n, 1 \le h \le n)$为故障路径；当 OM_j 中发生故障 $\mathrm{FO}_{i,g,r}$ 时，$\mathrm{I}_{i,g}(1 \le j \le k, 1 \le g \le n)$无法通过 CM_r 到达 $\mathrm{O}_{j,h}$，即路径 $\mathrm{P}_{i,g,r,j,h}(1 \le j \le k, 1 \le g \le n)$为故障路径。

3　全分布式容错调度算法

为进行容错调度，首先需要设计故障检测算法以获得故障信息。其次，由于信元达到中间级后的路径是唯一的，因此 IM 的信元分发算法需具有故障路径选择能力以避免交叉点故障(包括发生在输入级、中间级和输出级的交叉点故障)造成的故障路径。另外，为降低突发业务对中间级和输出级调度的影响，IM 的信元分发算法需具有良好的负载均衡特性。因此，故障检测算法和 IM 的容错信元分发算法是容错调度算法的关键。

如图 2 所示，全分布式容错(Fully Distributed Fault Tolerant, FDFT)调度算法包括分布式故障检测(Distributed Fault Detection, DFD)算法、输入级 FT-DSRR(Fault Tolerant Desynchronized Static Round-robin)信元分发算法+中间/输出级的 CICQ 调度算法。FDFT 采用周期故障检测，检测周期为 T。在一个周期内，故障检测时间为 T_D，信元调度时间为 T_S。CICQ 调度借鉴已有的稳定调度算法，即能够达到 100%吞吐率的调度算法，如基于 Round-Robin 的 RR-RR 算法[15]。

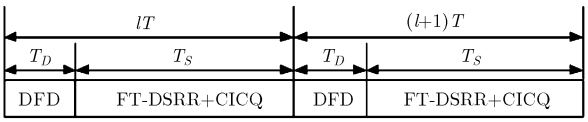

图 2　DTFT 算法的周期过程

3.1 DFD 故障检测算法

分布式故障检测算法包括交换单元内故障检测和级间故障信息反馈。

(1)交换单元内故障检测，以大小为 $N_\mathrm{I} \times N_\mathrm{O}$ 的交换单元为例说明。

第 l 个检测周期内，当 $t \in [lT, lT + T_D), l \in \mathbb{N}$ 时，任一输入端 $p, 1 \le i \le N_\mathrm{I}$，向输出端 $q=(i+t) \bmod N_\mathrm{O} +1$ 发送检测信元 C_D，如果输出端正确接收，则交叉点(p,q)无故障，反之，该交叉点故障。

(2)级间故障信息反馈。

后一级交换单元的故障信息向前一级进行反馈，最终反馈到 IM, IM 更新各自的故障信息。

3.2 FT-DSRR 容错信元分发算法

以 IM_i为例说明。

在 IM_i 中，为实现无故障路径选择定义如下变量。状态向量 $\boldsymbol{S}_i = [s_r]_m (1 \le r \le m, s_r \in \{0,1\})$ 表征端口状态，$s_r = 1$ 表征 $\mathrm{IL}_{i,r}$ 被占用。$\mathbf{FI}_i = [fi_{g,r}]_{n \times m}$ $(1 \le g \le n, 1 \le r \le m, fi_{g,r} \in \{0,1\})$ 记录输入级故障信息，$fi_{g,r}=1$ 表征故障 $\mathrm{FI}_{i,g,r}$ 发生。$\mathbf{FC}_i = [fc_{r,j}]_{m \times k}$ $(1 \le r \le m, 1 \le j \le k, fc_{r,j} \in \{0,1\})$ 记录中间级故障信息，$fc_{r,j}=1$ 表征故障 $\mathrm{FC}_{r,i,j}$ 发生。$\mathbf{FO}_i = [fo_{r,l}]_{m \times N}$ $(1 \le r \le m, 1 \le l \le N, fo_{r,l} \in \{0,1\})$ 记录输入级故障信息，$fo_{r,l}=1$ 表征 $\mathrm{FO}_{j,r,h}(\ j = \lfloor (l-1)/n \rfloor + 1, h = ((l-1) \bmod n) + 1)$发生。各级故障信息由 3.1 节中所述的 DFD 算法得到。当 $fi_{g,r}=fc_{r,j}=fo_{r,l}=0$ 时，路径 $\mathrm{P}_{i,g,r,j,h}$ 为无故障路径。为保证不同输入端口的公平性，定义主指针(primary pointer)PI_i 为 $\mathrm{I}_{i,g}$ 按照轮询方式分配最高优先级。具体的 FT-DSRR 调度过程如下：

(1)在时隙 t 开始时，$\boldsymbol{S}_i = [0]_m$；

(2)FOR $g = \mathrm{PI}_i : (\mathrm{PI}_i + n - 1) \bmod n$

　　输入端 $\mathrm{I}_{i,g}$ 有信元到达，其目的地址为 $\mathrm{O}_{j,h}$;

　　FOR $p = 1 : m$

　　令 $r = (g + \mathrm{offset} + p) \bmod m$，如果路径 $\mathrm{P}_{i,g,r,j,h}$ 为无故障路径，且 $s_r = 0$，则发送信元到 $\mathrm{IL}_{i,r}$，并置 $s_r = 1$; BREAK;

(3)更新 $\mathrm{offset}=(\mathrm{offset}+1) \bmod m$，$\mathrm{PI}_i=(\mathrm{PI}_i+1) \cdot \bmod n$。

4 性能分析

4.1 T_S的选择

如图 2 所示，令网络带宽利用率为γ，则$\gamma = T_S/(T_S + T_D)$。根据 DFD 过程，$T_D = \max\{m,n,k\}B_C/R_L$，其中 R_L为输入数据速率，B_C为一个 cell 的长度。因此，为提高带宽利用率，需增大 T_S。另一方面，T_S的增大会增加 T_S内发生新故障的概率，从而造成 DFD 的漏检。

不失一般性，假设不同故障发生相互独立，令$r_{cp}(t) = \exp\{-\lambda_{cp}t\}$为交叉点在时间段 t 内的可靠性，λ_{cp}为交叉点的故障率，$N_{cp}(l)$为第 l 个周期 DFD 检测得到的无故障交叉点个数。对于给定的漏检率P_D，第 l 个周期内 DFD 检测正常的条件是$[r_{cp}(t)]^{Ncp(l)} \ge 1 - P_D$。为提高带宽利用率，选择第 l 个周期内的$T_S(l)$为保证该周期内故障检测正常的调度时间最大值，则 $T_S(l)$满足：

$$T_S(l)=\arg\max_t\left\{t:\left[r_{cp}(t)\right]^{N_{cp}(l)} \ge 1-P_D\right\}=-\frac{\ln(1-P_D)}{\lambda_{cp}N_{cp}(l)} \tag{1}$$

又因为 $N_{cp}(l) \le (r^2m + 2mnr)$，所以$[r_{cp}(t)]^{N_{cp}(l)} \ge [r_{cp}(t)]^{(r^2m+2mnr)}$，因此，在 FDFT 中选择 T_S为

$$T_S = \min\{T_S(l)\} = -\frac{\ln(1-P_D)}{\lambda_{cp}(k^2m + 2mnk)} \tag{2}$$

令 $P_D = 10^{-5}, \lambda_{cp} = 7.605\times10^{-7}\,day^{-1}$ 为宇航级 FPGA Virtex-4VQ 在同步轨道下的单粒子翻转率[4]。表 2 为不同交换尺寸下的 T_S，可得当链路速率达到 Gbps 量级时，$T_S \gg T_D$，即 $\gamma \approx 1$。

表 2 不同尺寸下的 T_S

(n,m,k)	(4,7,4)	(16,20,8)
T_S	0.939 h	2.9586 min

4.2 无故障路径的负载均衡

令 Clos 网络的业务矩阵为 $\boldsymbol{\Lambda} = [\lambda_{i,g,j,h}]_{N\times N}$，其中$\lambda_{i,g,j,h}$为 $I_{i,g}$→$O_{j,h}$的业务到达率。如图 3 所示，假设输入端$I_{i,g}$与输出端$O_{j,h}$间存在$m_{i,g,j,h}$条故障路径。不失一般性，假设路径 $P_{i,g,r,j,h}(1 \le r \le m_{i,g,j,h})$为无故障路径，其余为故障路径。定义分发比例$\eta_{i,g,r,j,h}$为信元从 $I_{i,g}$经 $P_{i,g,r,j,h}$前往 $O_{j,h}$的概率。通过在 IM 中维持主指针(primary pointer)以轮询方式为每个输入端口分配最高优先级，FT-DSRR 将输入业务均匀分配到无故障路径，即实现无故障路径上的负载均衡。因此 FDFT 的$\eta_{i,g,r,j,h}$满足

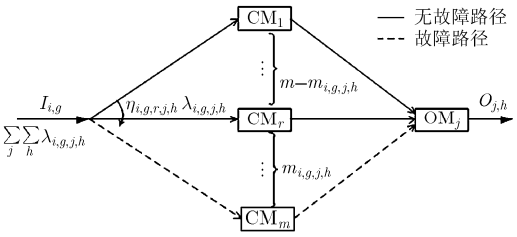

图 3 无故障路径间的负载均衡

$$\eta_{i,g,r,j,h} = \begin{cases} \dfrac{1}{m - m_{i,g,j,h}}, & 1 \le r \le m - m_{i,g,j,h} \\ 0, & m - m_{i,g,j,h} < r \le m \end{cases} \tag{3}$$

4.3 FDFT 的稳定性及容错能力分析

假设输入业务为可允许业务(admissible traffic)，即$\forall i,g, \sum_j\sum_h(\lambda_{i,g,j,h}) \le 1$, $\forall j,h, \sum_i\sum_g(\lambda_{i,g,j,h}) \le 1$，且中间/输出级采用可允许下的稳定的调度算法，输入级信元分发算法在故障条件下稳定的条件如定理 1 所示。

定理 1 对于 CICQ-SMM Clos 网络 C(n,m,k)($m \ge n$)，任一信元分发算法在故障条件下稳定的充分条件是$\forall i,g,j,h$, $\eta_{i,g,r,j,h}$满足：

$$\sum_{g=1}^{n}\sum_{j=1}^{k}\sum_{h=1}^{n}\eta_{i,g,r,j,h}\lambda_{i,g,j,h} \le 1 \tag{4}$$

$$\sum_{i=1}^{k}\sum_{g=1}^{n}\sum_{h=1}^{n}\eta_{i,g,r,j,h}\lambda_{i,g,j,h} \le 1 \tag{5}$$

$$\sum_{r=1}^{m-m_{i,g,j,h}}\eta_{i,g,r,j,h} = 1 \tag{6}$$

$$\eta_{i,g,r,j,h} = 0, \quad m - m_{i,g,j,h} < r \le m \tag{7}$$

证明 式(6)为保证全部业务传输的条件，式(7)为保证业务不经过故障路径的条件。如图 3 所示，CM_r的输入端口 $IL_{i,r}$的业务到达率λ_{IL}为

$$\lambda_{IL} = \sum_g\sum_j\sum_h\eta_{i,g,r,j,h}\lambda_{i,g,j,h} \tag{8}$$

由于 CM 采用可允许业务下稳定的调度算法，因此，当$\lambda_{IL} \le 1$时，中间级的调度稳定。此时，CM_r的输出端口 $CL_{r,j}$的业务负载λ_{CL}为

$$\lambda_{CL} = \sum_i\sum_g\sum_h\eta_{i,g,r,j,h}\lambda_{i,g,j,h} \tag{9}$$

由于 OM 采用可允许业务下稳定的调度算法，因此，当$l_{CL} \le 1$时，输出级的调度稳定。 证毕

由于无容错能力的算法(DSRR[11,14], MFRR[12], SDIB[13]等)无法区分故障路径和无故障路径，因此其无法保证式(7)所示的条件。所以即使仅有一个故障发生时，无容错能力的调度算法无法使调度达到稳定。另一方面，由于 DSRR 算法的分发系数满足$\lambda_{i,g,j,h} = 1/m$，所以根据式(3)，无故障条件下 FDFT

能够达到与 DSRR 相同的性能。

由定理 1 可得 FDFT 的容错能力如推论 1 和推论 2 所示。

推论 1　对于 CICQ-SMM Clos 网络 C(n,m,k) ($m \ge n$), FDFT 稳定的充分条件是

$$\max_{i,j,g,h}\{m_{i,g,j,h}\} \le m-n \tag{10}$$

证明　由式(1)可知，FDFT 的$\lambda_{i,g,j,h}$满足式(4)和式(5)。另一方面，

$$\sum_{g=1}^{n}\sum_{j=1}^{k}\sum_{h=1}^{n}\eta_{i,g,r,j,h}\lambda_{i,g,j,h} \le \frac{\sum_{g,j,h}\lambda_{i,g,j,h}}{m-\max\limits_{g,j,h}\{m_{i,g,j,h}\}} \le \frac{n}{m-\max\limits_{g,j,h}\{m_{i,g,j,h}\}} \tag{11}$$

$$\sum_{i=1}^{k}\sum_{g=1}^{n}\sum_{h=1}^{n}\eta_{i,g,r,j,h}\lambda_{i,g,j,h} \le \frac{\sum_{i,g,h}\lambda_{i,g,j,h}}{m-\max\limits_{i,g,h}\{m_{i,g,j,h}\}} \le \frac{n}{m-\max\limits_{i,g,h}\{m_{i,g,j,h}\}} \tag{12}$$

当式(10)成立时，式(11)和式(12)成立。由定理 1 可知，此时 FDFT 算法稳定。　　证毕

在实际情况中，更关心的是在 CICQ-SMM Clos 网络稳定条件下，FDFT 能够容忍的交叉点故障个数。

推论 2　当交叉点故障只发生在输入级和输出级时，当每一个 IM/OM 的交叉点故障个数不超过 $(m-n)$ 时，FDFT 一定能够到达 100%吞吐率。当交叉点故障只发生在中间级时，当所有 CM 的交叉点故障个数不超过 $(m-n)$ 时，FDFT 一定能够到达 100%吞吐率。

证明　如图 3 所示，一个交叉点故障只能造成输入端 $\mathrm{I}_{i,g}$ 与输出端 $\mathrm{O}_{j,h}$ 间的一条路径故障，即 $m_{i,g,j,h}$ 加 1。另外，输入端 $\mathrm{I}_{i,g}$ 与输出端 $\mathrm{O}_{j,h}$ 间的 m 条路径仅经过 IM_i 和 OM_j 但经过所有 m 个 CM。因此当 IM_i 或 OM_j 中的交叉点故障个数不超过 $(m-n)$ 时，或当所有 CM 的交叉点故障个数不超过 $(m-n)$ 时，$\max\limits_{i,j,g,h}\{m_{i,g,j,h}\} \le m-n$ 一定成立。由推论 1 可知，当 IM_i 或 OM_j 中的交叉点故障个数不超过 $(m-n)$ 时，或当所有 CM 的交叉点故障个数不超过 $(m-n)$ 时，全分布式容错调度算法能够达到 100%吞吐率。证毕

推论 2 含义是 FDFT 算法能够抵抗的故障数量的下界，即当故障数量不超过推论 2 所限定的数量时，FDFT 总能够达到 100%吞吐率。值得注意的是，该结论在任意情况下都成立，包括导致交换无法达到 100%吞吐率所需最少故障数量的最坏情况，如故障只发生在一对输入—输出间的路径上导致该输入—输出无法正常交换。如下一节所示，在故障随机发生的条件下 FDFT 能够容忍更多的交叉点故障。

4.4 FDFT 的复杂度分析

根据 FT-DSRR 所需的故障信息，每个 IM 需要 $[m(n+k+N)]$ bit 的空间存储故障信息。

在 FT-DSRR 中，IM 每个输入需要轮询寻找未被占用的无故障输出。在最坏情况下，每个输入需要轮询所有输出，时间复杂度为 $O(\lg m)$。因此，在最坏情况下，FT-DSRR 的时间复杂度为 $O(n\lg m)$。

5　仿真结果

为衡量 CICQ-SMM Clos 网络及 FDFT 的性能及容错能力，在无故障条件下，同 IQ-SMM Clos 网络[12]及 OQ-SMM Clos 网络[11]的性能进行比较，仿真中 CICQ 采用 RR-RR 算法[15]，IQ 采用 iSLIP 算法[16]，分别表示为 RR-RR 和 i-SLIP。在故障条件下，同输入级采用 DSRR 算法[11,14]的 CICQ-SMM Clos 网络的性能进行比较。仿真中 Clos 网络尺寸为 $n=r=4$,$m=7$，交叉点队列长度 C_{XP} 分别为 1 和 4, iSLIP 的迭代次数 i 分别为 1 和 4。另外，由于中间级缓存会造成信元乱序，仿真中在输出端添加重排序缓存。

采用两种故障产生模式。故障模式 1：为衡量发生在不同位置的故障影响，故障只在某一级均匀随机产生，令输入级发生的故障个数为 NF_I，中间级发生的故障个数为 NF_C。故障模式 2：为衡量 Clos 网络能够容忍的最大故障个数，故障在所有交换单元均匀随机产生，令产生的总故障个数为 NF_T。仿真采用的业务模型为 Bernoulli 均匀业务，突发均匀业务和不均衡业务。突发业务的突发长度 B 分为 16 cell 和 32 cell，一次突发中目的地址相同。令输入业务负载为 $\lambda, 0 \le \lambda \le 1$，不均衡业务满足：

$$\lambda_{i,j,g,h}=\begin{cases}\lambda\left(\omega+\dfrac{1-\omega}{N}\right), & i=g, j=h\\ \lambda\dfrac{1-\omega}{N}, & \text{其他}\end{cases} \tag{13}$$

其中ω为非均衡因子。仿真性能通过吞吐率和信元时延衡量。吞吐率为 $\lambda=1$ 时接收数据量与发送业务量的比值。信元时延为通过时隙归一化后的时延，单位为时隙。对于给定信元大小 B_C 和数据速率 R_L，数据的实际时延为 $T_C B_C / R_L$，其中 T_C 为仿真得到的时延。

5.1 Bernoulli 均匀业务下性能分析

图 4 为 Bernoulli 均匀业务无故障条件下算法时延性能。当 iSLIP 迭代次数 $i=1$ 时，IQ- SMM Clos 网络无法达到 100%吞吐率。C_{XP}=1 和 4 时，随负载增大，FDFT 都可以达到 100%吞吐率，时延性能介于 OQ-SMM Clos 网络的时延和 IQ-SMM Clos

网络的时延性能之间。图 5 为 Bernoulli 均匀业务故障模式 1 下算法时延性能，其中 C_{XP}=4。随 NF_I和 NF_C增加，FDFT 时延增加缓慢，而 DSRR 算法时延迅速增大，且无法达到 100%吞吐率。另一方面，根据推论 2，当 NF_I>12 或 NF_C>4 时，FDFT 无法达到 100%吞吐率。但在故障随机发生的情况下，当 NF_I>12 或 NF_C>4 时，FDFT 仍能达到 100%吞吐率。因此，当故障随机发生时，FDFT 能够抵抗更多的交叉点故障。图 6 为 Bernoulli 均匀业务故障模式 2 下算法的吞吐率性能。当 NF_T=15(对应一个 SE 发生一个故障)，DSRR 算法吞吐率约为 85%，而 FDFT 算法吞吐率为 100%。

5.2 突发业务下性能分析

图 7 为突发业务无故障条件下算法时延性能。C_{XP}=1 和 4 时，FDFT 具有相同的时延性能，介于 IQ-SMM Clos 网络的时延和 OQ-SMM Clos 网络的时延之间。随负载增大，FDFT 的时延接近 OQ-SMM Clos 网络的时延。Bernoulli 业务可看作突发长度为 1(B=1)突发业务。结合图 4，无故障条件下，随着突发长度的增大信元时延增大。图 8 为突发业务故障模式 1 下算法时延性能，其中 C_{XP}=4。随 NF_I和 NF_C增加，DSRR 算法的时延迅速增大，且无法达到 100%吞吐率。随 NF_I和 NF_C增加，FDFT 算法时延增加缓慢。同样地，在突发业务下，当故障随机发生时，FDFT 能够抵抗更多的交叉点故障。随着业务突发长度增大，信元时延显著增大。图 9 为突发业务故障模式 2 下算法的吞吐率性能。当 NF_T=15(对应一个 SE 发生一个故障)，DSRR 算法吞吐率约为 85%，而 FDFT 算法吞吐率仍大于 95%。受业务突发特性影响，随业务突发长度增大，FDFT 吞吐率下降且随故障数量变化剧烈。另一方面，比较图 6 和图 9 可得 DSRR 吞吐率变化相同，因此 DSRR 的吞吐率主要受故障数量影响。

5.3 非均匀业务下性能分析

图 10 为非均匀业务下算法的吞吐率性能。无故障发生时，CICQ-SMM Clos 网络吞吐率介于 IQ-SMM 和 OQ-SMM 之间。当 NF_{T}=20 时，DSRR 算法吞吐率小于 85%，而 FDFT 的吞吐率仍大于 97%。

6 结束语

受空间辐射环境影响，星载交换结构需要具有比地面交换结构更高的可靠性。相比于其他交换结构，Clos 网络的多通路特性能够提供更高的可靠性。本文提出一种 CICQ-SMM Clos 网络及 FDFT 调度算法。CICQ-SMM Clos 网络支持 Clos 网络和交换单元内部的全分布式调度。FDFT 包括分布式故障检测算法、输入级 FT-DSRR 信元分发算法和中间/输出级的 CICQ 稳定调度算法。FDFT 能够实现无故障路径的负载均衡。理论证明，FDFT 能够达到 100%吞吐率的充分条件是一个 IM/OM 的故障个数不超过 $(m-n)$ 或所有 CM 的故障个数不超过 $(m-n)$。仿真分析表明，无故障条件下，当交叉点队列长度 $C_{\mathrm{XP}} \geq 1$ 时，FDFT 是稳定的；当故障随机发生时，FDFT 仍具有良好的性能(如当每个交换单元发生有一个故障随机发生时，FDFT 在均匀业务下的吞吐率为 100%)。该算法以信元为调度粒度，面临信元乱序的问题，设计具有保序能力的全分布式容错调度算法将作为本文的后续工作。

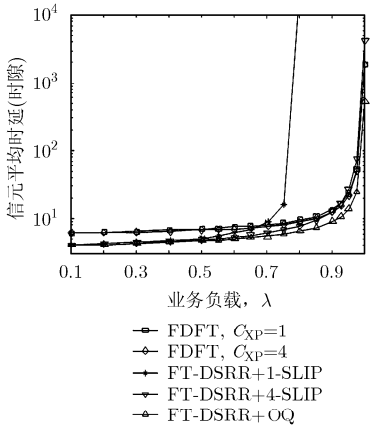

图4 Bernoulli业务无故障下算法时延性能

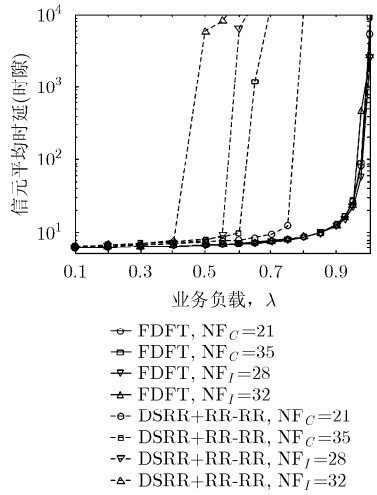

图5 Bernoulli业务故障模式1下算法时延性能

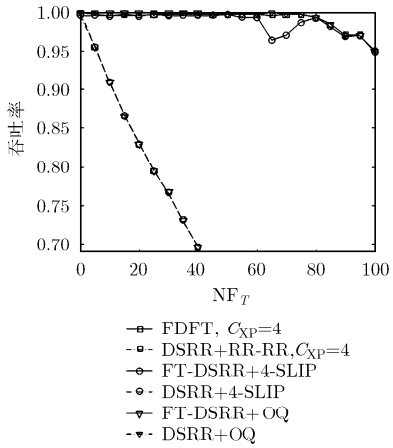

图6 Bernoulli业务故障模式2下算法吞吐率性能

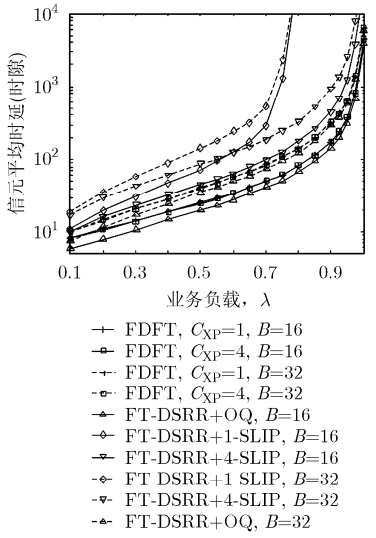

图7 突发业务无故障下算法时延性能

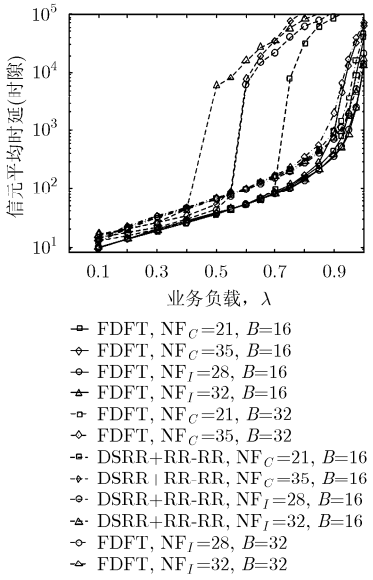

图8 突发业务故障模式1下算法时延性能

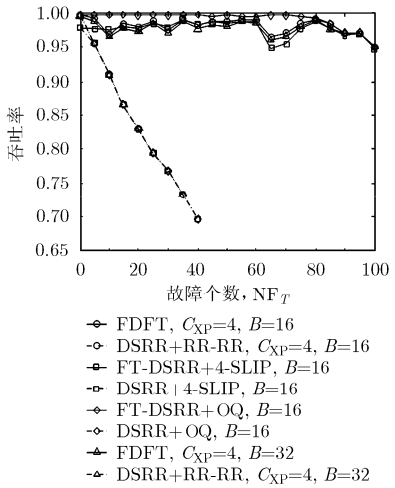

图 9 突发业务故障模式 2 下算法时延性能

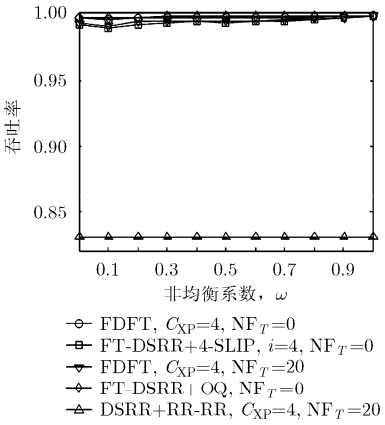

图 10 非均匀业务下算法吞吐率性能

参 考 文 献

[1] BOTTA A and PESCAPE A. On the performance of new generation satellite broadband internet services[J]. *IEEE Communications Magazine*, 2014, 52(6): 202–209. doi: 10.1109/MCOM.2014.6829965.

[2] JAFF E, PILLAI P, and HU Y. IP multicast receiver mobility support using PMIPv6 in a global satellite network[J]. *IEEE Communications Magazine*, 2015, 53(3): 30–37. doi: 10.1109/MCOM.2015.7060479.

[3] COURVILLE N, BISCHI H, and ZENG J. Critical issues of onboard switching in DVB-S/RCS broadband satellite networks[J]. *IEEE Wireless Communications*, 2005, 12(5): 28–36. doi: 10.1109/MWC.2005.1522101.

[4] SIEGLE F, VLADIMIROVA T, ILSTAD J, *et al*. Mitigation of radiation effects in SRAM-based FPGAs for space applications[J]. *ACM Computing Surveys*, 2015, 47(2): 37:1–37:34. doi: 10.1145/2671181.

[5] 张茂森, 邱智亮, 高雅, 等. 星上 Clos 交换网络的分治调度算法[J]. 电子与信息学报, 2012, 34(11): 2734–2740. doi: 10.3724/SP.J.1146.2012.00553.

ZHANG M, QIU Z, GAO Y, *et al*. Divide-and-conquer dispatching scheme for satellite clos-network switches[J]. *Journal of Electronics & Information Technology*, 2012, 34(11): 2734–2740. doi: 10.3724/SP.J.1146.2012.00553.

[6] KAO Y H and CHAO H J. Design of a bufferless photonic clos network-on-chip architecture[J]. *IEEE Transactions on Computers*, 2014, 63(3): 764–776. doi: 10.1109/TC.2012.250.

[7] YANG Y and WANG J. A fault-tolerant rearrangeable permutation network[J]. *IEEE Transactions on Computer*, 2004, 53(4): 414–426. doi: 10.1109/TC.2004.1268399.

[8] GAO Y, QIU Z, ZHANG M, *et al*. Distributed weight matching dispatching scheme in MSM clos-network packet

switches[J]. *IEEE Communications Letters*, 2013, 17(3): 580–583. doi: 10.1109/LCOMM.2013.012213.122552.

[9] 高雅, 邱智亮, 张茂森, 等. 基于帧填补的 MMM Clos 网络按序分组交换算法[J]. 电子与信息学报, 2012, 34(11): 2715–2720. doi: 10.3724/SP.J.1146.2012.00617.
GAO Y, QIU Z, ZHANG M, *et al*. Padded-frame based in-sequence dispatching scheme for memory-memory-memory (MMM) clos-network[J]. *Journal of Electronics & Information Technology*, 2012, 34(11): 2715–2720. doi: 10.3724/SP.J.1146.2012.00617.

[10] 杨君刚, 刘增基, 雒晓卓. 一种新型的三级Clos网络分布式容错调度机制[J]. 解放军理工大学学报: 自然科学版, 2011, 12(3): 217–222.
YANG J, LIU Z, and LUO X. New distributed fault tolerance scheduling algorithm in three-stage clos network[J]. *Journal of PLA University of Science and Technology* (*Natural Science Edition*), 2011, 12(3): 217–222.

[11] LI X, ZHOU Z, and HAMDI M. Space-memory-memory architecture for clos-network packet switches[C]. Proceedings of IEEE International Conference on Communications, Seoul, 2005, (2): 1031–1035. doi: 10.1109/ICC.2005.1494505.

[12] YU H, RUEPP S, and BERGER M S. Out-of-sequence prevention for multicast input-queuing space-memory-memory clos-network[J]. *IEEE Communications Letters*, 2011, 15(7): 761–763. doi: 10.1109/LCOMM.2011.051011.102535.

[13] KLEBAN J and SUSZYNSKA U. Static dispatching with internal backpressure scheme for SMM clos-network switches [C]. Proceedings of IEEE Symposium on Computers and Communications, Madeira, 2013: 654–658. doi: 10.1109/ISCC.2013.6755022.

[14] ZHANG M, QIU Z, and GAO Y. Space-memory-memory clos-network switches with in-sequence service[J]. *IET Communications*, 2014, 8(16): 2825–2833. doi: 10.1049/iet-com.2013.0844.

[15] ROJAS-CESSA R, OKI E, JING Z, *et al*. CIXB-1: combined input-one-cell-crosspoint buffered switch[C]. Proceedings of IEEE Workshop on High Performance Switching and Routing, Dallas, 2001: 324–329. doi: 10.1109/HPSR.2001.923655.

[16] MCKEOWN N. The iSLIP scheduling algorithm for input-queued switches[J]. *IEEE/ACM Transactions on Networking*, 1999, 7(2): 188–201. doi: 10.1109/90.769767.

刘 凯： 男，1986 年生，博士生，研究方向为星载交换技术、卫星网络等.

晏 坚： 男，1975 年生，副研究员，研究方向为星载交换技术、空间网络协议等.

高晓琳： 女，1979 年生，博士生，研究方向为空间网络协议.