一种多级多平面分组交换结构中的带宽保证型调度算法

马祥杰[①] 李秀芹[②] 兰巨龙[①] 张百生[①]
[①](解放军信息工程大学信息工程学院 郑州 450002)
[②](华北水利水电学院信息工程学院 郑州 450011)

摘 要：多级多平面分组交换结构 MPMS 以其优异的可扩展性正成为新一代交换路由设备的交换核心。但 MPMS 结构中的调度算法却往往比较复杂。该文提出了一种 MPMS 结构的带宽保证型调度算法 BG-CRRD，该算法将分组流预留带宽信息引入判决机制，仿真实验表明，BG-CRRD 在 Bernoulli 均匀流量条件下可以获得 100%的吞吐率，在非均匀流量条件极坏情况下获得高达 92%的吞吐率，在过载情况下根据预定带宽分配输出链路带宽。

关键词：调度算法；多级多平面交换结构；并行轮转匹配；iSLIP；带宽保证
中图分类号：TP393 **文献标识码**：A **文章编号**：1009-5896(2009)06-1475-04

A Novel Scheduling Scheme with Bandwidth Guarantees in the Multiple-Plane and Multiple-Stage Packet Switching Fabric

Ma Xiang-jie[①] Li Xiu-qin[②] Lan Ju-long[①] Zhang Bai-sheng[①]
[①](*Information Engineering Institute*, *PLA Information Engineering University*, *Zhengzhou* 450002, *China*)
[②](*North China University of Water Conservancy and Electric Power*, *Zhengzhou* 450011, *China*)

Abstract: The multiple-plane and multiple-stage (MPMS) switching fabric has attractive scalability features that make it appealing as an alternative for scalable routers. However, scheduling packets in MPMS fabric is complex. In this paper, a novel scheduling scheme is proposed with bandwidth guarantees for the MPMS fabric. It can deliver 100% throughput under Bernoulli uniform traffic, 92% throughput in the worst case under nonuniform traffic and allocate bandwidth according to reserved bandwidth under overloaded traffic.

Key words: Scheduling scheme; Multiple-Plane and Multiple-Stage (MPMS); Concurrent Round Robin Dispatching (CRRD); iSLIP; Bandwidth guarantee

1 引言

当前互联网中正在兴起的高带宽网络业务，如 IPTV[1]，电子科学[2]等，对交换路由设备的交换容量和 QoS 保证能力提出新的挑战。MPMS 交换结构采用并行分布式结构，可以很好地解决交换容量的扩展问题[3]。而交换结构保证 QoS 主要依赖于调度算法的支持，但 MPMS 中的调度问题非常复杂。解决输入竞争和输出竞争的端口匹配是一个时间复杂度为 $O(N^3 \log N)$ 的调度问题[4]，这对于高速率和多端口的 MPMS 是难以实现的。RD 算法最早用于 ATLANTA 交换芯片[5]解决多级调度，它采用随机选择机制进行调度判决，但要获得 100%吞吐率其内部加速比高达 1.6。为了克服 RD 算法的加速问题，Oki E 等人提出一种基于类似 iSLIP[6]优先级指针机制的 CRRD 算法[7,8]，并证明了 CRRD 算法不需要内部加速即可在均匀流量下获得 100%的吞吐率，但在非均匀流量下 CRRD 仅能获得 63%的吞吐率。本文基于已经研究的一种具有良好可扩展性的多级多平面分组交换结构的图论模型[9]，提出一种新型可提供带宽保证的 BG-CRRD 算法，该算法在调度判决机制中引入了分组流预留带宽信息，公平优先级列表中的对应端口数量和预留带宽成正比。仿真实验表明，BG-CRRD 算法在均匀流量条件下获得 100%的吞吐率；在非均匀流量条件下极坏情况获得 92%的吞吐率；在过载流量情况下能够按照预留带宽为各分组流分配输出链路带宽。

2008-04-23 收到，2008-11-26 改回
国家 973 重点基础研究发展计划(2007CB307102)资助课题

2 MPMS 的带宽保证型调度算法 BG-CRRD

2.1 BG-CRRD 算法中的参数定义

为了便于表述，定义 BG-CRRD 算法中的有关参数如下：n 为每个 ISU/OSU 的输入端口/输出端口数量。k 为 MPMS 中的交换单元 ISU/OSU 数量。P, p 为 MPMS 中的 SP 数量和第 p 个 SP。$f^{\text{mpms}}_{v_i^0 v_j^6}(i,j=1,2,\cdots,nk)$ 为 MPMS 中从 v_i^0 到 v_j^6 的分组流。$R^{\text{mpms}}_{v_i^0 v_j^6}(i,j=1,2,\cdots,nk)$ 为 MPMS 中 $f^{\text{mpms}}_{v_i^0 v_j^6}$ 预留的带宽。$R^{p}_{v_i^1 v_s^2 v_t^3}$ 为 MPMS 中第 p 个 SP 上流经 v_i^1，v_s^2 和 v_t^3 的带宽分量。$\text{FPL}(\vec{e}^{\,2}_{(p,s,t)})$ 为 MPMS 中 $\vec{e}^{\,2}_{(p,s,t)}$ 对应的仲裁器公平优先级列表 FPL。$N(\vec{e}^{\,2}_{(p,s,t)})$ 为 $\text{FPL}(\vec{e}^{\,2}_{(p,s,t)})$ 中总顶点标识数量，取值为 MPMS 端口数 nk 的整数倍。$N(v_i^1, \vec{e}^{\,2}_{(p,s,t)})$ 为 $\text{FPL}(\vec{e}^{\,2}_{(p,s,t)})$ 中顶点 v_i^1 的标识数量。$P(v_i^1)$ 为 MPMS 中 v_i^1 的仲裁器优先级指针。$P(\vec{e}^{\,2}_{(p,s,t)})$ 为 MPMS 中 $\vec{e}^{\,2}_{(p,s,t)}$ 的仲裁器

优先级指针。$P(\vec{e}^3_{(p,t,u)})$ 为 MPMS 中 $\vec{e}^3_{(p,t,u)}$ 的仲裁器优先级指针。

2.2 BG-CRRD 的算法原理与设计

根据 MPMS 模型参数定义，分组流 $f^{\text{mpms}}_{v_i^0 v_j^6}$ 分流到第 p 个 SP 上流量在 m 条 $\vec{e}^2_{(p,s,t)}(t=1,2,\cdots,m)$ 上平均分配，每条有向边带宽分量为 $R^p_{v_i^1 v_s^2 v_t^3}$，根据 PPS 结构中的结果[3]可知：

$$R^p_{v_i^1 v_s^2 v_t^3}=\frac{1}{Pm}R^{\text{mpms}}_{v_i^0 v_j^6} \tag{1}$$

BG-CRRD 算法调度决策基于仲裁器中的 FPL 列表。与 CRRD 算法不同，为了能够保证预留带宽 $R^{\text{mpms}}_{v_i^0 v_j^6}(i,j=1,2,\cdots,nk)$，BG-CRRD 的 FPL($\vec{e}^2_{(p,s,t)}$) 列表中各顶点标识数量 $N(v_i^1,\vec{e}^2_{(p,s,t)})$ 与分流到 $\vec{e}^2_{(p,s,t)}$ 上的流量带宽 $R^p_{v_i^1 v_s^2 v_t^3}$ 成正比。如果 FPL($\vec{e}^2_{(p,s,t)}$) 列表空间中包含 $N(\vec{e}^2_{(p,s,t)})$ 个顶点标识，那么 FPL($\vec{e}^2_{(p,s,t)}$) 中 $N(v_i^1,\vec{e}^2_{(p,s,t)})$ 的计算方法如式(2)：

$$\begin{aligned}N(v_i^1,\vec{e}^2_{(p,s,t)})&=\frac{R^p_{v_i^1 v_s^2 v_t^3}}{\sum_{i=(s-1)n+1}^{sn}R^p_{v_i^1 v_s^2 v_t^3}}N(\vec{e}^2_{(p,s,t)})\\&=\frac{R^{\text{mpms}}_{v_i^0 v_j^6}}{\sum_{i=(s-1)n+1}^{sn}R^{\text{mpms}}_{v_i^0 v_j^6}}N(\vec{e}^2_{(p,s,t)})\end{aligned} \tag{2}$$

BG-CRRD 采用两个匹配阶段分隔输入竞争和输出竞争问题：第 1 阶段完成各 ISU 内部匹配，解决输入竞争；第 2 阶段完成 ISU 和 MSU 之间的匹配，解决输出竞争。类似于 iSLIP 算法，两个匹配阶段采用“请求-响应-接受”3 步迭代的方法以易于硬件实现。第 1 阶段的在 ISU 中的算法运行结果实现 v_i^1 与 $\vec{e}^2_{(p,s,t)}$ 之间的匹配，匹配成功的 v_i^1 数量为 $\min(d(\overline{v}_i^1),m)$，其中 $d(\overline{v}_i^1)$ 是匹配前顶点子集 V^1 中的非空顶点数。第 2 阶段在 ISU 与 MUS 间的匹配主要实现第 1 阶段匹配成功的 v_i^1 与 $\vec{e}^3_{(p,t,u)}$ 之间的关联。经过两个阶段的匹配后成功关联的 DEM-MUX 对(v_i^1， v_j^5)完全解决了输入冲突和输出冲突问题， v_i^1 即可通过第 1 阶段选定的 $\vec{e}^2_{(p,s,t)}$ 和第 2 阶段选定的 $\vec{e}^3_{(p,t,u)}$ 向 v_j^5 发送数据分组了。

BG-CRRD 算法的 2 次匹配具体设计如下：

(1)第 1 阶段：ISU 内部的匹配

(a)第 1 次迭代

步骤 1(请求)：所有非空的 v_i^1 向所有与 v_s^2 邻接的 $\vec{e}^2_{(p,s,t)}$ 仲裁器发送请求；

步骤 2(响应)：所有收到请求的 $\vec{e}^2_{(p,s,t)}$ 仲裁器从 FPL ($\vec{e}^2_{(p,s,t)}$) 列表的指针位置 $P(\vec{e}^2_{(p,s,t)})$ 开始按照轮转方式选择一个非空 v_i^1 请求，并向选择的 v_i^1 返回响应；

步骤 3(确认)：所有收到响应的 v_i^1 仲裁器从位置 $P(v_i^1)$ 开始按照轮转方式选择一个 $\vec{e}^2_{(p,s,t)}$ 的响应，并向选择的 $\vec{e}^2_{(p,s,t)}$ 仲裁器发送确认。

(b)第 i 次迭代($i>1$)

步骤 1(请求)：所有上一次迭代中未匹配 v_i^1 继续向未匹配的仲裁器 $\vec{e}^2_{(p,s,t)}$ 发送请求；

步骤 2(响应)：所有收到请求的未匹配 $\vec{e}^2_{(p,s,t)}$ 仲裁器从 FPL ($\vec{e}^2_{(p,s,t)}$) 列表的指针位置 $P(\vec{e}^2_{(p,s,t)})$ 开始按照轮转方式选择一个非空 v_i^1 请求，并向选择的 v_i^1 返回响应；

步骤 3(确认)：所有收到响应的未匹配 v_i^1 仲裁器从位置 $P(v_i^1)$ 开始按照轮转方式选择一个 $\vec{e}^2_{(p,s,t)}$ 的响应，并向选择的 $\vec{e}^2_{(p,s,t)}$ 仲裁器发送确认。

(2)第 2 阶段：ISU 与 MSU 之间的匹配

步骤 1(请求)：在第 1 阶段中完成匹配的 v_i^1 向 $\vec{e}^3_{(p,t,u)}$ 仲裁器发送请求；

步骤 2(响应)：所有收到请求的 $\vec{e}^3_{(p,t,u)}$ 仲裁器从位置 $P(\vec{e}^3_{(p,t,u)})$ 开始按照轮转方式选择一个 v_i^1 请求，并向选择的 v_i^1 返回响应；

步骤 3(发送分组)：所有收到响应的 v_i^1 在下一时隙向与 $\vec{e}^3_{(p,t,u)}$ 关联的 v_j^5 发送对应的数据分组。

(3)优先级指针更新

(a)优先级指针 $P(\vec{e}^2_{(p,s,t)})$ 的更新：如果 $\vec{e}^2_{(p,s,t)}$ 在第 1 阶段某一次迭代中得到匹配并且在第 2 次匹配中得到 MSU 的响应，那么指针 $P(\vec{e}^2_{(p,s,t)})$ 更新到下一位置： $P(\vec{e}^2_{(p,s,t)})=(P(\vec{e}^2_{(p,s,t)})+1)\bmod N(\vec{e}^2_{(p,s,t)})$；

(b)优先级指针 $P(v_i^1)$ 和 $P(\vec{e}^3_{(p,t,u)})$ 的更新：如果顶点 v_i^1 和有向边 $\vec{e}^3_{(p,t,u)}$ 在第 1 阶段第 1 次迭代中得到匹配并且在第 2 次匹配中得到 MSU 的响应，那么指针 $P(v_i^1)$ 和 $P(\vec{e}^3_{(p,t,u)})$ 更新到下一位置：

$P(v_i^1)=(P(v_i^1)+1) \bmod m$, $P(\vec{e}^3_{(p,t,u)})=(P(\vec{e}^3_{(p,t,u)})+1) \bmod nk$

BG-CRRD 算法在第 1 阶段 ISU 内部匹配中的多次迭代过程进行到没有新的 v_i^1 和 $\vec{e}^2_{(p,s,t)}$ 获得匹配时即可结束。BG-CRRD 算法的优先级指针 $P(\vec{e}^2_{(p,s,t)})$，$P(v_i^1)$ 和 $P(\vec{e}^3_{(p,t,u)})$ 的更新方法继承了 CRRD 和 iSLIP 算法类似机制，具有各仲裁器指针去同步功能，保证了低迭代次数和高吞吐率。

3 BG-CRRD 算法的仿真实验与结果分析

本文在 Bernoulli 均匀流量，非均匀流量以及过载流量条件下分别对 BG-CRRD 算法和 CRRD 算法的性能进行了仿真实验。实验中 BG-CRRD 和 CRRD 在第 1 阶段中 ISU 采用多次迭代过程，第 2 阶段 ISU 和 MSU 之间采用 1 次迭代过程。数据分组长度为定长 512 byte，平均时延不包括分组分段和重组以及链路传输造成的延迟。

3.1 Bernoulli 均匀流量实验

实验中选用 MPMS($P=3,n=8,k=8,m=8$) 结构，1~3 次迭代 BG-CRRD 和 1，4 次迭代 CRRD 在 Bernoulli 均匀流量下的平均时延如图 1 所示。从图中可以看出，BG-CRRD 和 CRRD 算法在 Bernoulli 均匀流量条件下均可获得 100%吞吐率。CRRD 需要 4 次迭代即可收敛，获得最小平均时延。BG-CRRD 迭代次数增加至 3 次时获得最优平均时延性能。同时流量强度 λ 取值在 0.6 至 0.9 时 BG-CRRD 的最优平均时延明显优于 CRRD。

图 2 给出了 BG-CRRD 算法在不同数量交换平面的平均

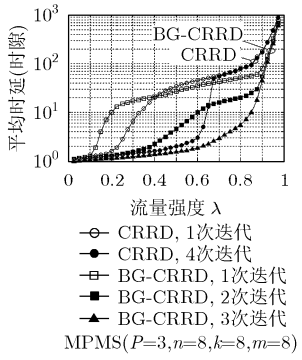

图 1 BG-CRRD 和 CRRD 算法在 Bernoulli 均匀流量下的平均时延

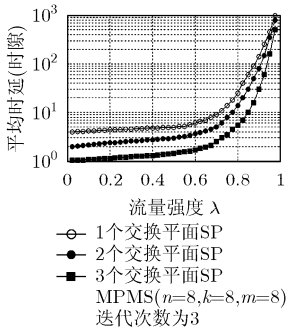

图 2 BG-CRRD 算法对于不同数量 SP 在 Bernoulli 均匀流量下的平均时延

时延性能。该实验采用的多级多平面分组交换结构为 MPMS ($n=8,k=8,m=8$)。仿真结果表明，BG-CRRD 算法的时延性能随着交换平面 SP 数量的增加而提高，当交换平面数量增加至 3 时获得最优平均时延。

3.2 非均匀流量实验

文献[7]给出了交换调度算法常用的非均匀流量模型。假设 v_i^0 的流量速率为 R，那么非均匀流量条件下对于 MPMS 结构的顶点对(v_i^0 , v_j^6)之间的流量速率 $R_{v_i^0 v_j^6}^{\text{mpms}}$ 计算方法如式(3)所示：

$$R_{v_i^0 v_j^6}^{\text{mpms}} = \begin{cases} R\left(w+\dfrac{1-w}{nk}\right), & i=j \\ R\dfrac{1-w}{nk}, & i\neq j \end{cases} \tag{3}$$

图 3 给出了 BG-CRRD 和 CRRD 算法在非均匀流量条件下的吞吐率性能。实验采用 MPMS($P=3,n=8,k=8,m=8$)。从图中可以看出，对于 $w=0$ 的 Bernoulli 均匀流量和 $w=1$ 的定向流量，所有迭代次数的 BG-CRRD 和 CRRD 算法都获得 100%的吞吐率。在极坏情况下，CRRD 的 4 次迭代在 $w=0.45$ 时的吞吐率为 63%，BG-CRRD 的 3 次迭代在 $w=0.4$ 时的吞吐率为 92%，这说明在极坏情况下 BG-CRRD 的收敛吞吐率比 CRRD 高出 29%。

3.3 过载流量带宽分配实验

在正常的 Bernoulli 均匀流量和非均匀流量的仿真实验中，所有流向输出链路的带宽之和不能超过该链路的最大带宽。但是实际的交换结构存在输出链路带宽过载的情况，如果能够按照实际预定带宽比例为各分组流分配带宽无疑可以最大限度地保证业务带宽需求和带宽分配的公平性。

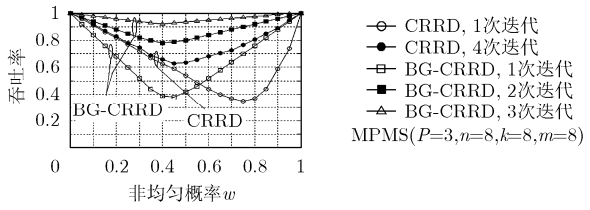

图 3 BG-CRRD 和 CRRD 算法在非均匀流量下的吞吐率性能

该实验采用 MPMS($P=3,n=2,k=2,m=2$)，分组流 $\text{Flow}_{v_1^0 v_1^6}^{\text{mpms}}$,$\text{Flow}_{v_2^0 v_1^6}^{\text{mpms}}$,$\text{Flow}_{v_3^0 v_1^6}^{\text{mpms}}$ 和 $\text{Flow}_{v_4^0 v_1^6}^{\text{mpms}}$ 的预留带宽比例分别为 40%, 30%, 20%和 10%。本文采用的流量速率矩阵 $\boldsymbol{\Gamma}=[R_{v_i^0 v_j^6}^{\text{mpms}}]$ 如式(4)所示，其中 91%的流量强度分配给分组流 $\text{Flow}_{v_1^0 v_1^6}^{\text{mpms}}$,$\text{Flow}_{v_2^0 v_1^6}^{\text{mpms}}$,$\text{Flow}_{v_3^0 v_1^6}^{\text{mpms}}$ 和 $\text{Flow}_{v_4^0 v_1^6}^{\text{mpms}}$，9 %的流量强度分配给其余分组流。

$$\boldsymbol{\Gamma}=[R_{v_i^0 v_j^6}^{\text{mpms}}]=\begin{bmatrix} 0.91 & 0.03 & 0.03 & 0.03 \\ 0.91 & 0.03 & 0.03 & 0.03 \\ 0.91 & 0.03 & 0.03 & 0.03 \\ 0.91 & 0.03 & 0.03 & 0.03 \end{bmatrix} \tag{4}$$

图 4 给出了 BG-CRRD 和 CRRD 算法在过载流量 $\boldsymbol{\Gamma}=[R_{v_i^0 v_j^6}^{\text{mpms}}]$ 下带宽分配情况。当流量强度 λ 小于 27.5% (1/(4×91%)=27.5%)时属于非流量过载区，BG-CRRD 和 CRRD 都为 4 个分组流平均分配链路带宽。但当流量强度 λ 大于 27.5%出现流量过载，BG-CRRD 根据预留带宽比例 40%, 30%, 20%和 10%为 4 个分组流分配带宽，而 CRRD 仅为 4 个分组流平均分配带宽，每个分组流带宽为 25%。

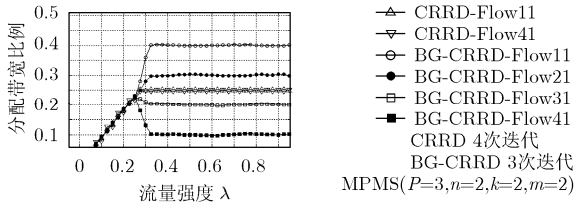

图 4 BG-CRRD 和 CRRD 算法在过载流量 $\boldsymbol{\Gamma}=[R_{v_i^0 v_j^6}^{\text{mpms}}]$ 下的带宽分配性能

4 结束语

本文在多级多平面分组交换结构 MPMS 模型的基础上提出一种新型可提供带宽保证的调度算法 BG-CRRD。该算法通过在仲裁器调度判决机制中引入各分组流预留带宽信息，从而实现了交换过程对业务流带宽的保证。仿真实验表明，BG-CRRD 算法在 Bernoulli 均匀流量条件下可以获得 100%的吞吐率，仅需要 3 次迭代即可收敛至最优的时延性能；在非均匀流量条件下可以获得的吞吐率性能明显优于 CRRD 算法，极坏情况下可以获得的吞吐率高达 92%；在过载流量条件下可以根据预定带宽在各分组流间分配输出链路带宽，从而保证业务流的带宽需求和带宽分配的公平性。

参 考 文 献

[1] Cherry S. The battle for broadband (Internet protocol television) [J]. *IEEE Spectrum*, 2006, 42(1): 24–29.

[2] Newman H B, Ellisman M H, and Orcutt J A. Data-intensive E-science frontier research [J]. *Communnication of the ACM*, 2005, 46(11): 68–77.

[3] Ma Xiangjie and Lan Julong. Emulating output queueing with the central-stage buffered Clos packet switching network [C]. IEEE Conference on High Performance Switching and Routing, Shanghai, China, May, 2008: 98–103.

[4] Mekkittikul A and McKeown N. A practical scheduling algorithm for achieving 100% throughput in input-queued switches [C]. IEEE INFOCOM Proceedings, San Francisco, USA, 2006: 792–799.

[5] Chiussi F M, Kneuer J G, and Kumar V P. Low-cost scalable switching solutions for broadband networking: The ATLANTA architecture and chipset [J]. *IEEE Communnication Magzin*e, 1997, 5(2): 44–53.

[6] McKeown N. The iSLIP scheduling algorithm for input-queued switches [J]. *IEEE/ACM Trans. on Networking*, 1999, 7(2): 188–200.

[7] Oki E, Jing Z, and Chao H J. Concurrent rounrobin dispatching scheme for Clos-network switches [J]. *IEEE/ACM Trans. on Networking*, 2001, 10(6): 830–844.

[8] Chiussi F, Gerla M, and Sivaraman V. Traffic shaping for end-to-end delay guarantees with edf scheduling [C]. International Workshop on Quality of Service, Pittsburgh, USA, June, 2006: 2056–2066.

[9] 马祥杰, 李秀芹, 兰巨龙等. 一种新型可扩展的多级多平面分组交换结构的图论模型与性能分析[J]. 电子与信息学报, 2009, 31(5): 1026-1030.
Ma Xiang-jie, Li Xiu-qin, and Lan Ju-ling, e*t al.*. Graphic model and performance analysis of a novel scalable multiple-plane and multiple-stage packet switching fabric[J]. *Journal of Electronics & Information Technology*, 2009, 31(5): 1026–1030.

马祥杰： 男，1977 年生，博士，从事高速网络交换结构与调度策略的研究.

李秀芹： 女，1967 年生，副教授，从事高速网络交换结构与 QoS 保证策略的研究.

兰巨龙： 男，1962 年生，教授，博士生导师，从事交换路由理论与技术的研究.

张百生： 男，1969 年生，高级工程师，从事交换理论与技术的研究.