ISSN 1000-0054 清华大学学报（自然科学版）2011 年 第 51 卷 第 11 期 12/27
CN 11-2223/N J Tsinghua Univ (Sci & Tech), 2011, Vol. 51, No. 11 1680-1685

用于构建数据中心的双列循环移数结构

张小平， 段武清， 李孟涵， 张 超

（清华大学 计算机科学与技术系，清华信息科学与技术国家实验室(筹)，北京 100084）

摘 要：目前数据中心的结构多采用 server-centric 或树状结构设计。Server-centric 结构由于交换节点之间缺乏直接相连的链路，影响了服务器之间通信的路径多样性。树状结构采用层次结构，不利于同层各交换节点之间的数据交换。该文介绍了一种用于构建数据中心的新型结构，即双列循环移数结构(two line barrel shifter，TLBS)。理论分析和仿真实验结果显示：该结构可克服树状结构中同层节点数据交换不便的缺点，同时具备丰富的数据路径多样性。该结构网络直径低，扩展性较强，是一种理想的数据中心构建结构。

关键词：数据中心；循环移数结构；网络直径

中图分类号：TP 393.4 **文献标志码**：A

文章编号：1000-0054(2011)11-1680-06

Two line barrel shifter for data centers

ZHANG Xiaoping, DUAN Wuqing, LI Menghan, ZHANG Chao

(Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract: Many of today's data centers use server-centric or tree-like structures. Sever-centric systems have few direct links between switch nodes, which influences the path diversity for the communication between servers. Tree-like systems use tier structures, which limits data switching between switch nodes in the same tier. This paper introduces a data center architecture, named the two line barrel shifter (TLBS). Tests indicate that this architecture has good path diversity and provides convenient data switching between nodes in the same tier in a tree-like structure. In addition, this architecture has low diameters and strong scalability, so it is an ideal data center architecture.

Key words: data center; barrel shifter; network diameter

近年来，数据中心已成为许多重要 Internet 业务的支撑，相关技术是当前的研究热点[1]。一个数据中心通常是由多台服务器通过某种结构连接在一起共同对外提供各种服务的集合体，其结构可采用多种设计方案[1-9]，典型的结构有：server-centric[1-2]和树状结构[4,6]。文[1-2]采用以服务器为中心的结构，该类结构需要在每个服务器上安装交换线卡，提高了成本；而且由于交换节点之间缺乏直接相连的链路，影响了服务器之间通信路径的多样性，进而影响分组延时和系统容错性。文[4]给出了传统数据中心的经典结构，该结构采用树状结构设计；文[6]对文[4]进行了改进，采用三层胖树(fat tree)结构，该结构无需在顶层采用高性能交换节点且增加了任意交换节点间的可用路径数；这两者都有利于层间数据交换但不利于同层数据交换。

本文提出一种用于构建数据中心的新型结构，即双列循环移数结构(two line barrel shifter，TLBS)。该结构可克服树状结构中同层节点数据交换不便的缺点，同时具备丰富的数据路径多样性。

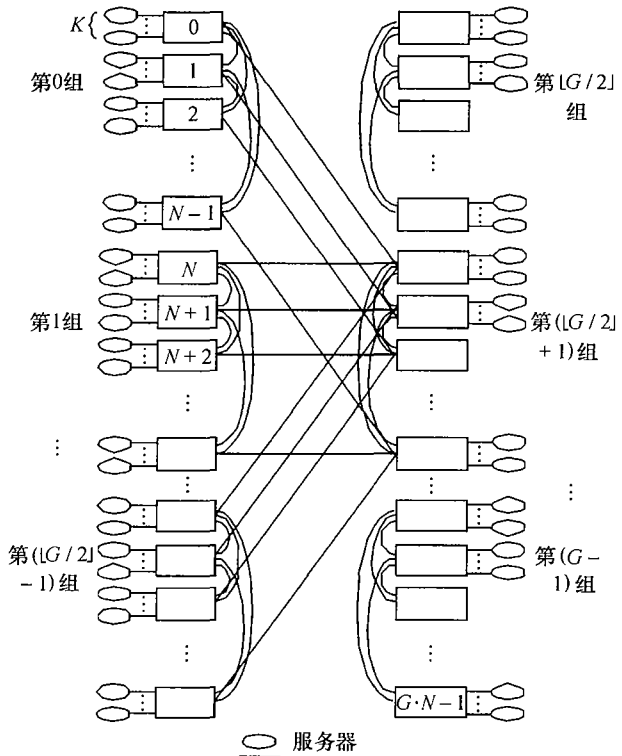
1 拓扑结构及节点编号

数据中心由交换结构和服务器组成。本文把双列循环移数结构连接的数据中心抽象为由交换节点和服务器节点构成的模型，其拓扑结构如图 1 所示(仅给出部分连线，每条连线表示双向全双工链路)。图 1a 为全局结构，图中交换节点分为对称的两列，各列内又进行分组。如果将每个组作为一个整体，则组之间的连接是一个完全二分图。图 1b 为组内结构，其内交换节点之间采用模 2 加(减)的循环移数结构连接，该循环移数结构的连接方式可参考文[10]。本文将属于同一组的交换节点的整体称为一个交换节点组。

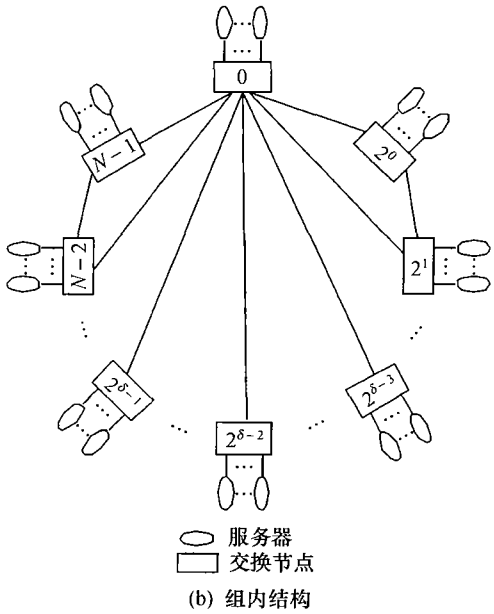
收稿日期：2011-03-18

基金项目：国家自然科学基金资助项目 (60903184)；
国家“八六三”高技术项目 (2008AA01A323)

作者简介：张小平(1975—)，男(汉)，内蒙古，副教授。
E-mail：zhxp@tsinghua.edu.cn



(a) 全局结构



(b) 组内结构

图 1 TLBS 的结构图

设 TLBS 中交换节点总共有 G 组，每组 N 个节点，每个节点连接 K 个服务器。用 n_i 表示编号为 i 的节点，其中 $i\in[0,G\cdot N-1]$，则 TLBS 中与 n_i 相连的节点可分为两类：1）与 n_i 同组的交换节点，2）与 n_i 不同列的交换节点。将所有与 n_i 同组的交换节点用集合 A 表示，则 A 可表示为

$$A=\begin{Bmatrix} n_j \mid j=\lfloor i/N\rfloor\cdot N+(i+2^\delta)\bmod N \\ \text{或}\quad \lfloor i/N\rfloor\cdot N+(i-2^\delta)\bmod N \end{Bmatrix}.$$

同样，将所有与 n_i 不同列的交换节点用集合 B 表示，则 B 可表示为

$$B=\begin{Bmatrix} n_j \mid j=(\beta+\lfloor G/2\rfloor)\cdot N+i\bmod N, & \lfloor i/N\rfloor<\lfloor G/2\rfloor \\ \text{或}\quad \beta\cdot N+i\bmod N, & \lfloor i/N\rfloor\geqslant\lfloor G/2\rfloor \end{Bmatrix}.$$

TLBS 中节点之间的连接关系为：对 $\forall i\in[0,G\cdot N-1]$，$\forall\delta\in[0,\lceil\log_2 N\rceil-1]$，$\forall\beta\in[0,\lfloor G/2\rfloor-1]$，以 n_i 为起点、n_j 为终点建立一条数据通路，其中 $j\in A\cup B$。

定义 1 将同一个交换节点组中与 n_i 相连的节点集合 A 划分为互不相交的两个集合 C 和 D，其中 C 和 D 划分 A 的规则如下：

$$C=\begin{Bmatrix} n_j \mid j=\lfloor i/N\rfloor\cdot N+(i+2^\delta)\bmod N,\ (j-i+N)\bmod N\leqslant\lfloor N/2\rfloor \\ \text{或}\quad \lfloor i/N\rfloor\cdot N+(i-2^\delta)\bmod N,\ (j-i+N)\bmod N>\lfloor N/2\rfloor \end{Bmatrix}.$$

$$D=\begin{Bmatrix} n_j \mid j=\lfloor i/N\rfloor\cdot N+(i-2^\delta)\bmod N,\ (j-i+N)\bmod N\leqslant\lfloor N/2\rfloor \\ \text{或}\quad \lfloor i/N\rfloor\cdot N+(i+2^\delta)\bmod N,\ (j-i+N)\bmod N>\lfloor N/2\rfloor \end{Bmatrix}.$$

将以 n_i 为起点、$n_j\in C$ 为终点的数据通路称为顺时针链路，将以 n_i 为起点、$n_j\in D$ 为终点的数据通路称为逆时针链路。

定义 2 如果 n_i 到 n_j 为顺时针链路，将其跨度定义为 $((j-i)+N)\bmod N$；如果 n_i 到 n_j 为逆时针链路，则将其跨度定义为 $((i-j)+N)\bmod N$。

定义 3 将跨度为 2^δ 的顺时针链路称为 $+\delta$ 维链路，将跨度为 2^δ 的逆时针链路称为 $-\delta$ 维链路，其中，$\pm\delta$ 称为该链路的维度。

定义 4 若从 n_i 到 n_j 的某条路径可以由若干维度严格递减（递增）的 $+\delta(-\delta)$ 维链路组成，则称该路径为从 n_i 到 n_j 的组内特征路径。由于 $+\delta$ 或 $-\delta$ 维链路的出现当且仅当二进制数 $[((j-i)+N)\bmod N]_2$ 或 $[((i-j)+N)\bmod N]_2$ 的第 δ 位为 1，因此称该二进制数为该组内特征路径的特征。

TLBS 中将交换节点和服务器采用一个二进制数编号，如图 2 所示。该编号分为 3 段：组号、组内编号和局部编号。前两段表示交换节点编号，3 段

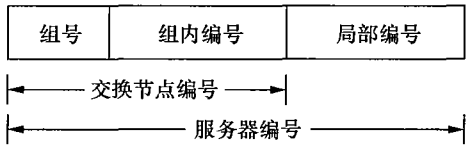


图 2 交换节点和服务器编号

合在一起为服务器编号。

2 内部路由算法

目前数据中心通常不采用 OSPF[11]、IS-IS[12]、RIP[13]和 BGP[14]等现存的路由算法,原因是多方面的。首先,这些算法通常要求在每个交换节点上运行一个协议栈,并交换和更新用于描述路径长度或链路状态的信息,因而开销较大。其次,路由计算和收敛时间限制了这些算法在具有较大数量交换节点的网络系统中的应用[15]。另外,数据中心中通常存在多条等长路径到达目的节点,但这些算法并没有考虑在多条等长路径上进行流量均衡。最后,数据中心采用的交换结构通常具有一定的规则性,利用该特点可简化算法设计并提高系统性能。因此,目前数据中心通常采用自行设计的路由算法[2,6]。

本文根据 TLBS 的结构特征设计了简单且利于硬件实现的分段特征路由算法。该算法利用特征序列表示数据的转发路径。特征序列的结构如图 3 所示。

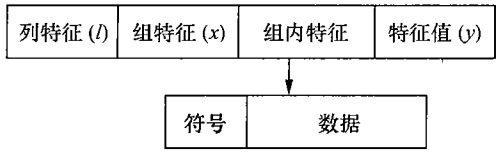


图 3 特征序列结构

定义 5 特征序列是一个用于在分段特征路由算法中表示分组转发路径的二进制数。

特征序列的结构分 4 段:列特征、组特征、组内特征和特征值。列特征表示目的服务器所在的列,宽度为 1。组特征表示目的服务器所在的交换节点组。列特征和组特征由目的服务器的组号确定,如图 4 所示。组内特征表示到达目的服务器的组内特征路径,它由符号和数据组成,符号为 0 表示该路径由顺时针链路组成,符号为 1 表示该路径由逆时针链路组成;数据为该组内特征路径的特征,由源服务器和目的服务器的组内编号按表 1 规则计算得到。表 1 中 a、b 分别为源服务器和目的服务器的组内编号,当 a、b 和 N 的关系满足计算约束条件时,采用前面与之对应的公式计算得到数据。特征值是目的服务器的局部编号。

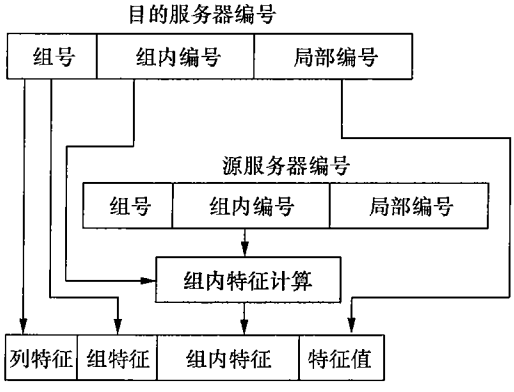


图 4 特征序列的计算

表 1 组内特征的计算方式

符号	数据	计算约束条件
0	$[((b-a)+N) \bmod N]_2$	$(b-a+N) \bmod N \leqslant \lfloor N/2 \rfloor$
1	$[((a-b)+N) \bmod N]_2$	$(b-a+N) \bmod N > \lfloor N/2 \rfloor$

用 n_i 表示 TLBS 中的任意节点,分段特征路由算法中分组转发过程分 3 步:

步骤 1 当 n_i 收到分组后,判断列特征 l 和组特征 x 是否均与本节点相同,如果相同,则转步骤 2;否则,判断 l 是否与本节点的列号相同,如果相同,则将分组转发到 n_j 然后继续执行步骤 1,其中:

$$j=\begin{cases}(\lfloor i/N \rfloor+\lfloor G/2 \rfloor)\cdot N+i, & \lfloor i/N \rfloor<\lfloor G/2 \rfloor \\ (\lfloor i/N \rfloor-\lfloor G/2 \rfloor)\cdot N+i, & \lfloor i/N \rfloor\geqslant\lfloor G/2 \rfloor\end{cases}$$

如果列号不同,则将分组转发到 n_z,$z=xN+i$,然后执行步骤 2。

步骤 2 当 n_i 收到分组后,判断组内特征的数据位是否全部为 0,若是则转步骤 3;否则,查找组内特征的数据中最左边的 1 所在的位置,设该位置为 δ,然后根据组内特征的符号是 0 或 1 分别用 n_i 的第 $+\delta$ 或 $-\delta$ 维链路转发分组,然后继续执行步骤 2。

步骤 3 当 n_i 收到分组后,将其转发给局部编号为 y 的服务器。

该算法有两个特点:1) 交换节点组内部在随机均匀流量模型下能随节点数增加保持吞吐率为 4,因此 TLBS 的组内连接不会因为系统扩展成为数据交换的瓶颈,具体证明见定理 1。2) 路由计算只采用加减和取模运算,方式简单,有利于硬件实现。

定理 1 TLBS 结构中,组内交换节点构成的循环移数结构在随机均匀流量模型下的吞吐率为 4。

循环移数结构是一种直连交换结构。证明之前,先给出直连交换结构吞吐率的定义。

定义 6 直连交换结构中,能使内部链路不达到饱和的节点最大注入流量称为吞吐率,记为 Θ。当注入单位速率时,$\Theta=1/\gamma_{\max}$,其中 $\gamma_{\max}$ 为链路在单位注入速率下的最大负载。

证明：向 n_i 注入单位 1 的随机均匀流量，该流量通过顺时针链路和逆时针链路往外发送。两类链路发出的流量在组内特征上仅表现为符号位不同，因此只需考察其中一类。第一类流量其组内特征的数据部分构成一个从$[0]_2$ 到$[\lfloor N/2 \rfloor]_2$ 共$\lfloor N/2 \rfloor$个递增的二进制数数列。根据组内特征的定义，该数列中所有数的第 δ 位之和表示了$\lfloor N/2 \rfloor$个流对 TLBS 中维度为$+\delta$ 的链路的负载。对于一个二进制数，其 δ 位发生变化必须将该数加上一个大小为 2^δ 的数。由于该数列递增的值为 1，因此数列中各数的第 δ 位发生变化必然在其之前有 2^δ 个数其 δ 位未发生变化。因此，该数列中，第 δ 位的变化规律为 2^δ 个连续的 0 或连续的 1 交替出现。图 5 给出了第 δ 位和第$\delta+1$ 上 1 出现的规律。下面证明第 δ 位出现 1 的总次数不少于第 $\delta+1$ 位。根据$\lfloor N/2 \rfloor$的大小分别讨论。当$\lfloor N/2 \rfloor<2^\delta$ 时，δ 位和$\delta+1$ 出现 1 的个数都为 0；当 $2^\delta\leqslant\lfloor N/2 \rfloor<2^{\delta+1}$时，$\delta$ 位出现 $2^{\delta+1}-\lfloor N/2 \rfloor\geqslant 0$ 个 1，而$\delta+1$ 位出现 0 个 1；当 $2^{\delta+1}\leqslant\lfloor N/2 \rfloor<2^{\delta+1}+2^\delta$ 时，δ 位出现 2^δ 个 1，而 $\delta+1$ 位出现$\lfloor N/2 \rfloor-2^{\delta+1}<2^\delta$ 个 1；当 $2^{\delta+1}+2^\delta\leqslant\lfloor N/2 \rfloor\leqslant 2^{\delta+1}+2^{\delta+1}$时，$\delta$ 位出现$\lfloor N/2 \rfloor-2^{\delta+1}$个 1，$\delta+1$ 位也出现$\lfloor N/2 \rfloor-2^{\delta+1}$个 1；当$\lfloor N/2 \rfloor>2^{\delta+1}+2^{\delta+1}$后，如果排除 δ 位和 $\delta+1$ 之前都出现过的 $2^{\delta+1}$ 个 1，后面将重复出现之前的情况。综上，δ 位出现 1 的次数不小于 $\delta+1$ 位。因此，第 0 位出现 1 的次数不小于其他位，即 $+0$ 维边上的负载最大。当 N 较大时，$+0$ 维上出现 1 的次数近似为 $N/4$。由于注入的流量为均匀流量，因此每个流的最大值为 $1/N$。取遍所有 n_i 可得每条 $+0$ 维链路上的负载为 $(1/N)\cdot\sum_{i=0}^{N-1}(N/4)\cdot(1/N)=\frac{1}{4}$。因此 $\gamma_{\max}=1/4$，$\Theta=1/\gamma_{\max}=4$。

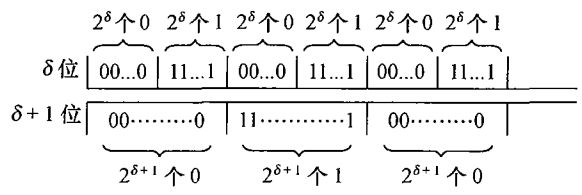


图 5 第 δ 和 $\delta+1$ 位上 1 的出现规律

3 性 能

3.1 路径长度

TLBS 中，组之间的连接为一个完全二分图，因此寻找目的交换节点组至多需要经过 1 个交换节点。交换节点组内部的路径长度则由定理 2 给出。

定理 2 TLBS 结构中，交换节点组内部的最长路径为$\lceil\log_2 N\rceil-1$，平均路径长度为$(\lceil\log_2 N\rceil-1)/2$。

证明：由组内特征的计算过程可知，对于两节点 n_i 到 n_j，其组内特征的绝对值满足 $|i-j|\leqslant\lceil N/2\rceil$，$\lceil N/2\rceil$最多需要$\lceil\log_2 N\rceil-1$ 位二进制数表示，因此组内特征路径最多由$\lceil\log_2 N\rceil-1$ 条链路组成。令 $\delta_{\max}=\lceil\log_2 N\rceil-1$。在交换节点组内部，对一个给定的交换节点 n_i，其他节点与 n_i 的距离由组内特征中 1 的个数确定。与 n_i 距离相等的节点数可用组合数 $2C(\delta_{\max},m)$表示，其中 m 为组内特征中 1 的个数。该类节点与组内节点总数的比例由二项式定理可知为

$$C(\delta_{\max},m)/2^{\delta_{\max}}.$$

由组合数的性质得

$$\frac{\sum_{m=0}^{\delta_{\max}} m\,C(\delta_{\max},m)}{2^{\delta_{\max}}}=\frac{\delta_{\max}(2^{\delta_{\max}}/2)}{2^{\delta_{\max}}}.$$

因此，组内两节点的平均距离为 $\delta_{\max}/2$。

由定理 2，TLBS 中连接任意两个交换节点最多经过 $\delta_{\max}+1$ 个中间节点，平均经过$\lceil\delta_{\max}/2\rceil+1$ 个中间节点。因此，TLBS 中的路径长度主要取决于组内路径长度。由于组内路径长度以 $O(\log_2 N)$ 增长，因此 TLBS 是一种低直径结构。

3.2 路径多样性

表 2 给出了 TLBS 与其他典型结构的路径多样性比较，其中 k'为 Fat Tree 中的交换节点接口数，k 为 BCube 的递归层数。从表 2 中可以看出，BCube 中不同服务器间同时存在的链路数目最多，但服务器自身的总线带宽限制了这些链路的可用性[2]。此外，BCube 中直接与服务器相连的交换节点间的链路数和同层交换节点之间相连的链路数均为 0。Tree、Fat Tree 结构中服务器之间的链路数与 TLBS 相等，但交换节点之间的链路数均不如 TLBS。

表 2 TLBS 与其他典型结构的路径比较

结构名称	链路数目		
	服务器间	连接服务器的交换节点间	同层交换节点间
Tree	1	2	0
Fat Tree	1	$k'/2$	0
BCube_k	$k+1$	0	0
TLBS	1	$\lceil\log_2 N\rceil+\lceil G/2\rceil$	$\lceil\log_2 N\rceil$

3.3 可扩展性

TLBS中组内交换结构的可扩展性已由定理1给出。组的扩展体现在：如果不考虑列的对称性，可以在任意一列上加入若干个组而不影响之前的结构和运行状态。此外，TLBS结构中所有交换节点结构相同。设节点的接口数为 L，则可连接的服务器数目用 $G\cdot K\cdot 2^{L-K-G/2}$ 表示，该数目呈指数增长。因此，TLBS能用接口数较少的交换节点连接众多服务器。综上可见，TLBS扩展性较强。

3.4 兼容性

TLBS中的节点编号与IP地址编制方式不一致，该问题将导致现有的上层应用不能在TLBS上运行。为此，建立节点编号与IP地址 $p_1.p_2.p_3.p_4$ 之间的映射规则，如表3所示。表3中的A表示 p_1 采用A类地址。转发中，利用该规则进行IP和节点编号之间的转换。此外，将特征序列附加在IP头部的后面供转发过程使用，如图6所示。该方法保证TCP/IP应用在TLBS上的正常运行。

表3 节点编号与IP地址映射表

节点类型	编号	p_1	p_2	p_3	p_4
交换节点	i	A	$\lfloor i/N \rfloor$	$i \bmod N$	1
服务器	i	A	$\lfloor i/(N\cdot K) \rfloor$	$\lfloor i/K \rfloor \bmod N$	$i \bmod K+1$

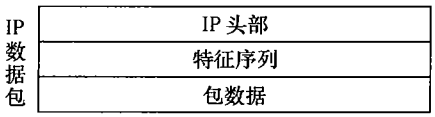


图6 特征序列在IP数据包中的位置

TLBS中任意编号的交换节点可以作为数据中心与外部的数据接口。这些交换节点承担网关功能，负责内外地址转换及数据转发。

4 仿真实验

图7给出了数据流经过的交换节点跳数及延时对比情况。图7a表示10 000个数据流在TLBS、Fat Tree和BCube 3种结构中经过的平均交换节点跳数对比情况。其中，每个交换节点连接24个服务器，数据流采用随机方式生成，当数据流通过服务器转发时，其转发跳数以2倍计算，这主要是因为服务器转发采用交换线卡，其效率相对低于专用交换节点。从图7可以看出，TLBS中数据流经过的跳数最少，其次为Fat Tree。TLBS跳数最少的原因主要在于不同组交换节点最多经过1跳可达，而同组内的交换节点最多经过 $\lceil \log_2 N \rceil - 1$ 可达。图7b给出了TLBS、Tree、Fat Tree和BCube 4种结构中分组延时随交换结构负载增加而变化的情况。仿真中共设置256个服务器，各服务器的最大接口容量为1。图7b显示：BCube和Tree结构当负载达到0.4后，分组延时急剧增加；Fat Tree结构当负载达到0.5后，分组延时也急剧增加；而TLBS当负载达到0.6时，仍然具有较低的分组延时，性能最优。这主要因为分组延时主要与分组经过的跳数和交换节点接口数相关，当连接相同数目服务器时，TLBS不仅跳数较少，且其交换节点内的交换接口数也较少。

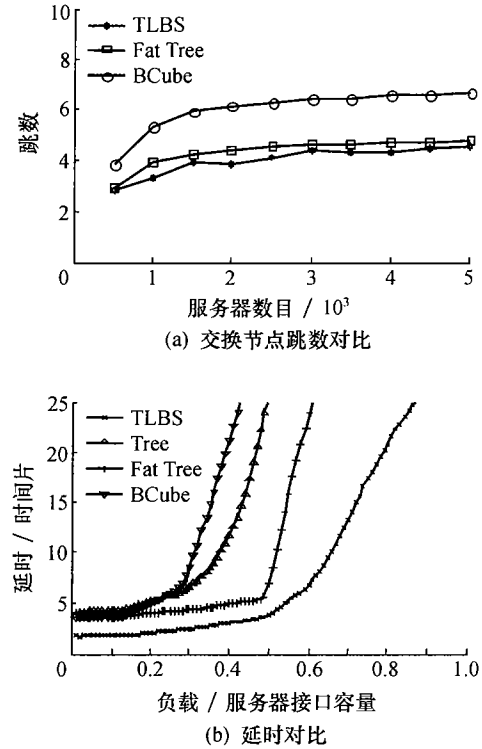


图7 数据流经过的交换节点跳数及延时情况对比

图8给出了TLBS、Tree、Fat Tree和BCube 4种结构所需交换节点数量对比及不同结构中故障节点对数据流量影响情况对比。图8a为所需交换节点数量对比情况。可以看出，要实现相同数目服务器的连接，TLBS需要的交换节点数最少(仍然设每个节点连接24个服务器)，其次为Tree，Fat Tree需要的交换节点数最多。原因主要在于TLBS中每个交换节点都直接与服务器相连，交换节点利用率相对较高；BCube中虽然每个交换节点也与服务器相连，但由于其结构特点，每增加一层，都要相应地增加大量交换节点，因而所需要的交换节点多于TLBS。图8b为故障节点对数据流的影响情况，对比在多对多(all-to-all)的数据流量模式下进行，且故障节点随机生成。可以看出，随着故障节点的

增加，TLBS在4种结构中表现出最佳性能，原因主要在于TLBS结构比其他结构具备更为丰富的数据路径多样性，当故障发生后，数据流可以有更多的路径选择绕过故障节点。

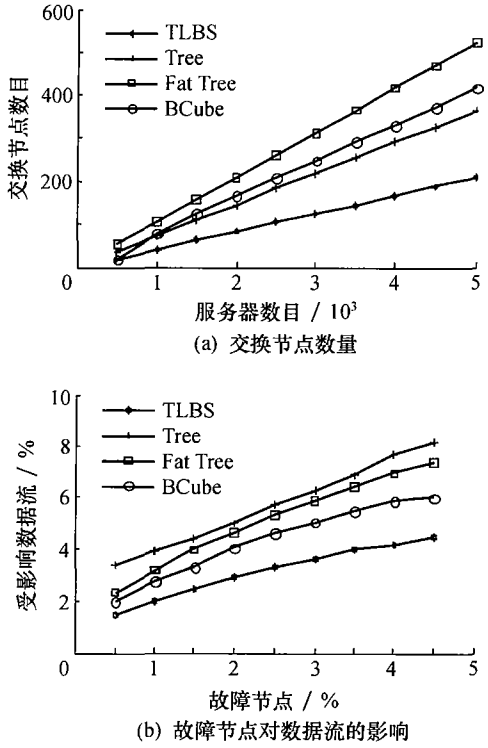


图8 所需交换节点数量及故障节点对数据流的影响对比

5 结论及下一步工作

本文给出了一种新型的数据中心结构TLBS，并设计了分段特征路由算法。该结构可克服树状结构中同层节点数据交换不便的缺点，同时具备丰富的数据路径多样性。此外，该结构网络直径低，扩展性较强，因此是一种理想的数据中心构建结构。

下一步的研究工作将用实际应用数据流进一步验证TLBS的性能特点，同时研究支持1对多、多对多等数据流量以及具有容错能力的路由算法和策略。此外，由于TLBS中使用的链路较多，而链路是传统交换结构成本的重要组成部分，因此下一步的研究工作还将包括TLBS中链路成本与其他结构中链路成本的对比分析。

参考文献 (References)

[1] GUO Chuanxiong, WU Haitao, TAN Kun, et al. DCell: A scalable and fault-tolerant network structure for data centers [C]// Proc of the ACM SIGCOMM 2008 Conf Data Communication. Seattle, Washington, USA: ACM Press, 2008: 75 - 86.

[2] GUO Chuanxiong, LU Guohan, LI Dan, et al. BCube: A high performance, server-centric network architecture for modular data centers [C]// Proc of the ACM SIGCOMM 2009 Conf Data Communication. Barcelona, Spain: ACM Press, 2009: 63 - 74.

[3] Mysore R N, Pamboris A, Farrington N, et al. PortLand: A scalable fault-tolerant layer 2 data center network fabric [C]// Proc of the ACM SIGCOMM 2009 Conf Data Communication. Barcelona, Spain: ACM Press, 2009: 39 - 50.

[4] Cisco. Data Center: Load Balancing Data Center Services [Z]. San Jose, CA, USA: Cisco Systems, 2004.

[5] Costa P, Zahn T, Rowstron A, et al. Why should we integrate service, servers, and networking in a data center? [C]// Proc of the 1st ACM Workshop: Research on Enterprise Networking. Barcelona, Spain: ACM Press, 2009: 111 - 118

[6] Al-Fares M, Loukissas A, Vahdat A. A scalable, commodity data center network architecture [C]// Proc of the ACM SIGCOMM 2008 Conf Data Communication. Seattle, Washington, USA: ACM Press, 2008: 63 - 74.

[7] Greenberg A, Lahiri P, Maltz D A, et al. Towards a next generation data center architecture: Scalablility and commoditization [C]// Proc of the ACM Workshop on Programmable Routers for Extensible Services of Tomorrow (PRESTO08). Seattle, Washington, USA: ACM Press, 2008: 57 - 62.

[8] Greenberg A, Hamilton J R, Jain N, et al. VL2: A scalable and flexible data center network [C]// Proc of the ACM SIGCOMM 2009 Conf Data Communication. Barcelona, Spain: ACM Press, 2009: 51 - 62.

[9] Kant K. Data center evolution [J]. *Computer Networks: The International Journal of Computer and Telecommunications Networking*, 2009, **53**(17): 2939 - 2965.

[10] 张小平. 可扩展路由器体系结构关键技术研究 [D]. 北京：清华大学，2008.
ZHANG Xiaoping. Research on the Key Technology of Scalable Router Architecture [D]. Beijing: Tsinghua University, 2008. (in Chinese)

[11] RFC 2328. OSPF, Version 2 [S]. Westford, MA, USA: Network Working Group, 1998.

[12] RFC 1142. OSI IS-IS Intra-Domain Routing Protocol [S]. Littleton, MA, USA: Digital Equipment Corporation, 1990.

[13] RFC 2453. RIP, Version 2 [S]. Washington, DC, USA: The Internet Society, 1998.

[14] RFC 4271. A Border Gateway Protocol 4 (BGP-4) [S]. Washington, DC, USA: The Internet Society, 2006.

[15] Moy J. OSPF: Anatomy of an Internet Routing Protocol [M]. Upper Saddle River, NJ, USA: Pearson Education, 1998.