

DPClos: 基于3级Clos结构的集装箱数据中心网络设计

李德顺, 申彦明, 李克秋*

(大连理工大学 计算机科学与技术学院, 辽宁 大连 116024)

摘要: 集装箱数据中心因其部署便捷在应急数据处理中发挥着基础性作用. 为满足集装箱数据中心对网络带宽和容错性的需要, 提出了一种基于3级Clos结构的集装箱数据中心网络, 记为DPClos. DPClos能够充分利用Clos结构和双端口服务器的特点, 为集装箱数据中心提供2倍网络带宽容量. DPClos网络的平均路径长度小于4, 且不同层次链路之间负载均衡. 双层网络结构赋予DPClos良好的容错性能, 当20%设备故障时数据流的成功到达率仍超过85%. 理论分析和实验表明, DPClos结构能够满足集装箱数据中心对网络性能的需求.

关键词: 集装箱数据中心; 数据中心网络; 网络带宽; 容错性

中图分类号: TP393

文献标识码: A

doi: 10.7511/dllgxb201606013

0 引言

随着云计算、大数据和物联网技术的发展, 传统数据中心网络面临着诸多问题^[1-6]. 集装箱数据中心因具有移动便捷、部署迅速和适应性强等特点, 在应急通信、军事勘探、灾难恢复和野外科考等领域发挥着基础性作用, 受到了业界的广泛关注^[4,7-8].

目前学界集中于对大规模数据中心网络的研究^[1-3,5], 以满足数据海量式增长的需求, 而对集装箱型数据中心网络关注较少^[4,7]. 数据中心网络可分为交换机为中心和服务器为中心两种结构. 交换机为中心的结构以Fat-tree^[2]和Jellyfish^[5]为代表, 服务器为中心的网络以DCell^[3]和BCube^[4]为代表. Li等^[6]提出的FiConn结构使用服务器的备用网卡, 属于以服务器为中心的网络. 上述网络针对大规模数据中心设计, 难以满足集装箱数据中心的需求. 集装箱数据中心存在着两个特点: (1) 带宽要求高. 作为应急部署的基础设施, 集装箱数据中心需要高容量的网络结构, 以满足服务器之间的通信和数据传输需求. (2) 部署环境差. 集装箱数据中心的部署环境可能导致设备的物理损伤和故障, 而一体化设计的集装箱数据中心在应用中维护困难. 集装箱数据中心需要考

虑网络带宽高和容错性强的结构, 而现有设计不能满足带宽和容错性的需求, 为此本文提出集装箱数据中心网络DPClos, 该网络采用2层对折的3级Clos结构^[9]和双端口服务器构建以满足集装箱数据中心对网络带宽和容错性的需求.

1 网络结构

1.1 Clos结构及特性

1.1.1 Clos结构 Clos结构^[9]源于电话网络, 其目的是以最少的交换开关提供无阻塞的电话通信. 3级Clos是最常用的交换结构, 记为 $C(n_1, r_1, m, n_2, r_2)$. 其中, 第1级为输入级, 由 r_1 个 $n_1 \times m$ 交换开关组成; 第2级为中间级, 由 m 个 $r_1 \times r_2$ 交换开关组成; 第3级为输出级, 由 r_2 个 $m \times n_2$ 交换开关组成. 输入级和输出级的每个交换开关有且仅有一条链路与中间级的每个交换开关相连. 当 $r_1 = r_2 = r$ 且 $n_1 = n_2 = n$ 时, 该结构是对称的3级Clos结构, 简记为 $C(n, m, r)$. 其交换链接如图1所示.

1.1.2 性质特点 Clos结构以交换容量大、扩展性好和可靠性高等特点广泛应用于网络设备中. 在 $C(n, m, r)$ 中, 当 $m \geq 2n - 1$ 时, 该结构是严

收稿日期: 2016-06-09; 修回日期: 2016-09-28.

基金项目: 国家杰出青年科学基金资助项目(61225010); 国家自然科学基金资助项目(61173160); 中央高校基本科研业务费专项资金资助项目(DUT15TD29).

作者简介: 李德顺(1983-), 男, 博士生, E-mail: lideshunlily@qq.com; 申彦明(1976-), 男, 教授, 博士生导师; 李克秋*(1971-), 男, 教授, 博士生导师, E-mail: keqiu@dlut.edu.cn.

格无阻塞的,即当一个新的输入级到输出级的连接请求到达时,不需要对网络中现有连接进行任何改动,即可满足该连接请求.当 $m \geq n$ 时, $C(n, m, r)$ 是可重排无阻塞的,即当一个新的输入级到输出级的连接请求到达时,可以通过对网络中已有连接进行调整来满足该连接请求.

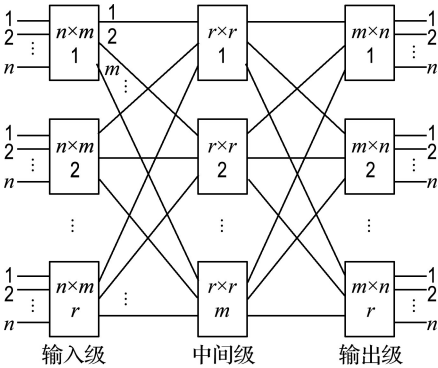


图 1 $C(n, m, r)$ 交换结构

Fig. 1 The exchange architecture of $C(n, m, r)$

Clos 结构的无阻塞特性使其成为数据中心网络的热点,在其上所设计的网络有 Fat-tree、PortLand、VL2 和 Elastic tree 等.这类网络的共同特点是采用折叠的 5 级 Clos 结构来提供可扩展的高性能数据中心网络.

1.2 第 1 类 DPClos 网络

DPClos 通过部署双层网络和服务端备用端口来满足集装箱数据中心的网络带宽和容错性需求.用 k 来表示交换机的端口数量.图 2 以 $k=4$ 为例展示了第 1 类 DPClos 集装箱数据中心网络结构,记为 FDPClos.该结构包括 1 层服务器和对称的上下层交换机网络,分别为上下核心层和上下边界层.上下层交换机网络均是由 3 级 $C(k/2, k/2, k/2)$ 沿中间级对折形成.输入级和输出级处于等价位置,组成交换机网络的边界层;中间级构成交换机网络的核心层.在 FDPClos 中,上下边界层交换机使用一半数量的端口连接到服务器.

在 FDPClos 中,每层交换机从 1 开始自左向右进行编号.分别用 Cid 和 Eid 表示核心层和边界层交换机编号.每台服务器按照所连接的上边界交换机为分组从 1 开始自左向右进行编号,该组编号即为对应的上边界交换机的编号 Eid .用 Sid 表示服务器编号.

FDPClos 上下两层网络分别采用 10.0.0.0/8 和 20.0.0.0/8 地址段.上层核心交换机的地址

格式为 10.4.1. Cid ,其中 Cid 为核心交换机的编号.上层边界交换机的地址格式为 10.0. Eid .1,其中 Eid 为边界交换机的编号. Eid 对应的服务器地址格式为 10.0. Eid . s ,其中 $s = Sid + 1$.下层网络采用与上层类似的地址结构,其地址段为 20.0.0.0/8.例如,下层第 1 台核心交换机地址为 20.4.1.1,第 1 台边界交换机地址为 20.0.1.1,该交换机所连接的 $Sid=1$ 服务器地址为 20.0.1.2.

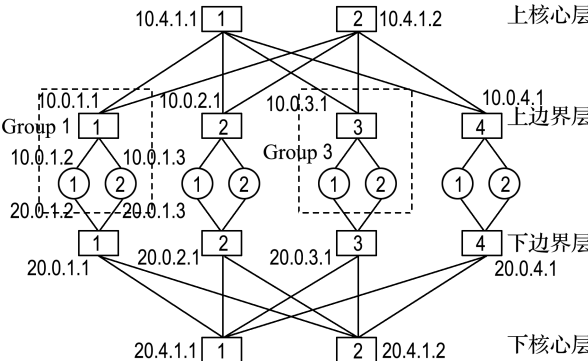


图 2 第 1 类 DPClos 结构 (FDPClos)

Fig. 2 The first class of DPClos architecture (FDPClos)

1.3 第 2 类 DPClos 网络

第 2 类 DPClos 采用与 FDPClos 相同的交换机网络结构,其交换机之间的连接和编号与 FDPClos 一致.第 2 类 DPClos 结构与 FDPClos 的不同之处在于服务器和下层边界交换机采用错位连接的方式,该结构记为 SDPClos.图 3 以 $k=4$ 为例展示了 SDPClos 网络连接.在 SDPClos 中, $Eid=i$ 的下边界交换机与服务器的连接方式为当 $i \leq k/2$ 时,下边界交换机 i 连接组 1 到组 $k/2$ 中 $Sid=i$ 的服务器;当 $i > k/2$ 时,下边界交换机 i 连接组 $k/2+1$ 到组 k 中 $Sid=i-k/2$ 的服务器.SDPClos 采用错位连接的方式不但可以缩短平均路径长度,而且能够避免 FDPClos 中编号相同的边界交换机同时故障时该组服务器被全部断开的情况.

SDPClos 中交换机采用与 FDPClos 相同的网络地址配置,服务器在 SDPClos 上层网络中的地址配置和 FDPClos 中对应的配置相同.方便起见,图 3 仅显示了服务器在上层边界交换机中的编号.用 Eid 代表下层边界交换机的编号,则其对应的服务器地址格式为 20.0. Eid . s ,其中 $s = Sid + 1$, Sid 是服务器在相应的交换机 Eid 中的编号.例如,交换机 20.0.1.1 所连接的两台服

器地址分别为 20.0.1.2 和 20.0.1.3。

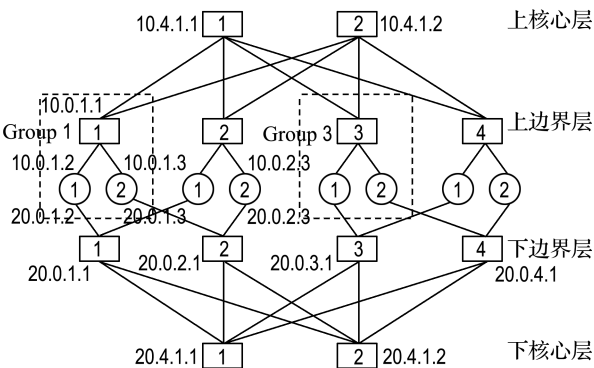


图 3 第 2 类 DPClos 结构(SDPClos)
Fig. 3 The second class of DPClos architecture (SDPClos)

在 DPClos 中,上下层交换机网络均能独立地保证任意两台服务器之间严格无阻塞通信。两类 DPClos 都是以交换机为中心的网络结构,服务器不参与数据的转发。在路由选择上,同一台边界交换机互联的服务器通过该边界交换机进行数据转发;非共享边界交换机的服务器之间可通过任意一台核心层交换机进行数据转发。在路由调度过程中,可以通过当前端口的负载来选择上下层网络和所路由的核心交换机。

2 DPClos 性质

DPClos 网络的性质特点包括设备数量、路径长度、瓶颈度和对剖带宽等。

2.1 设备数量

对于 DPClos 网络中交换机和服务器的数量,有如下定理:

定理 1 在 DPClos 网络中,交换机的数量 $N=3k$,服务器的数量 $S=k^2/2$,其中 k 为交换机的端口数。

证明 在 $C(n,m,r)$ 中,交换开关的数量为 $n+m+r$ 。因此,在交换机网络 $C(k/2,k/2,k/2)$ 中,交换机的数量为 $3k/2$ 。DPClos 使用独立的两层交换机网络,所以 DPClos 中交换机总数 $N=3k$ 。上层边界交换机数量为 k ,且每个交换机连接 $k/2$ 台服务器。因此,DPClos 中服务器数量 $S=k^2/2$ 。

表 1 列出了不同端口交换机构成的 DPClos 网络中设备的数量,其中 N 、 S 和 L 分别为交换机、服务器和连线数量。从表 1 中可以看出,当 $k=48,56$ 和 64 时分别可以构造拥有 $S=1\,152$,

1 568 和 2 048 台服务器的网络。这表明 DPClos 结构具有满足集装箱数据中心网络的扩展性。对比其他结构,DPClos 中服务器/交换机数量为 $S:N=k:6$;在 Fat-tree 中,该比例为 $k:5$;在文献[10]中该比例为 $k:10$ 。

表 1 DPClos 网络中的设备数量
Tab. 1 Number of equipments in DPClos

k	N	S	L	k	N	S	L
12	36	72	288	56	168	1 568	6 272
16	48	128	512	64	192	2 048	8 192
24	72	288	1 152	72	216	2 592	10 368
32	96	512	2 048	84	252	3 528	14 112
48	144	1 152	4 608	96	288	4 608	18 432

2.2 路径长度

对于 DPClos 中服务器之间的路径长度,有如下定理:

定理 2 在 FDPClos 中,服务器之间的平均路径长度为 $4-\frac{2(k-2)}{k^2-2}$;在 SDPClos 中,服务器之间的平均路径长度为 $4-\frac{4(k-2)}{k^2-2}$ 。

证明 在 DPClos 网络中,所有服务器都处于对等位置,因此考虑第 1 台服务器到其他各服务器的平均路径长度。在 FDPClos 中,第 1 台服务器到同一组其他服务器的路径长度为 2,数量为 $k/2-1$;第 1 台服务器到该组外其他服务器的路径长度为 4,数量为 $k^2/2-k/2$ 。因此,在 FDPClos 中,服务器之间的平均路径长度为 $\frac{2(k/2-1)+4(k^2/2-k/2)}{k^2/2-1}$,即 $4-\frac{2(k-2)}{k^2-2}$ 。同理可以分析 SDPClos 中服务器之间的平均路径长度为 $4-\frac{4(k-2)}{k^2-2}$ 。

由定理 2 得出,DPClos 网络的平均路径长度不大于 4,且 SDPClos 网络的平均路径长度小于 FDPClos 网络的。图 4 显示了 DPClos 网络的平均路径长度 d 随交换机端口数量变化的关系。从中可以看出,随着交换机端口数量的增长,DPClos 网络的平均路径长度接近 4。这表明,DPClos 网络满足集装箱数据中心的数据传输延迟小的需求。

2.3 瓶颈度

瓶颈度(bottleneck degree)是指网络在 all-to-all 流量模式下,单个链路所负载的流量最大值。对于 DPClos 网络的瓶颈度,有如下定理:

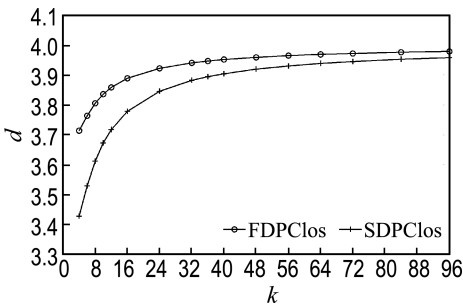


图 4 平均路径长度

Fig. 4 The average routing path length

定理 3 DPClos 网络的瓶颈度为 $(k^2-2)/4$.

证明 用 S 来表示网络中服务器的数量,用 SE 和 EC 分别来表示服务器和边界交换机之间、边界交换机和核心交换机之间的链路.忽略下层网络,仅分析上层网络在 all-to-all 流量模式下的链路负载.在上层网络中,服务器和边界交换机之间链路 SE 的负载为 $S-1$;第 1 台上边界交换机与核心交换机之间的总流量为 $(k/2) \cdot (S-k/2)$,其间链路数量为 $k/2$.因此,每条链路 EC 负载的流量为 $S-k/2$.在 FDPClos 中,下层网络对流量具有均分作用,因此链路 SE 承载流数量为 $(S-1)/2$;链路 EC 承载流数量为 $(S-k/2)/2$.故此,FDPClos 网络的瓶颈度为 $(S-1)/2$,其中 $S=k^2/2$.同理可以分析 SDPClos 的瓶颈度产生在链路 SE 上,为 $(k^2-2)/4$.

□

定理 3 表明,FDPClos 和 SDPClos 具有相同的瓶颈度.图 5 显示了 DPClos 网络在 all-to-all 流量模式下不同层次的链路上流数量分布 F_d .从图 5 可以看出,对于给定的交换机端口数 k ,DPClos 不同层次的链路在 all-to-all 流量模式负载近似均衡,不同链路负载差距很小.

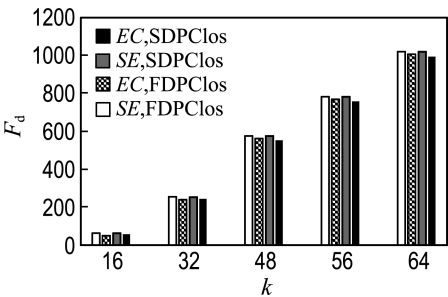


图 5 All-to-all 流量模式下链路的流数量分布

Fig. 5 The number of flows on a single link in all-to-all traffic model

2.4 对剖带宽

对剖带宽(bisection width)是指将网络均分成两部分所需断开链路的最小数量.对于 DPClos 网络的对剖带宽,有如下定理:

定理 4 DPClos 网络的对剖带宽为 $k^2/2$.

证明 折叠 $C(k/2,k/2,k/2)$ 结构的对剖带宽为 $k^2/4$.DPClos 网络采用了独立的上下层 Clos 结构,其对剖带宽为 $k^2/2$.

□

DPClos 是一种高带宽的数据中心网络,该网络能够提供 2 : 1 的网络带宽需求,即在任意网络连接下,两台服务器之间都可以拥有 2 条独立的链路连接.这不仅提高了网络的带宽,满足大数据传输的网络容量需求,而且提高了网络的容错性.相比之下,Fat-tree 提供了 1 : 1 的网络带宽需求,文献[10]提供了 2 : 1 的网络带宽需求,而其他数据中心网络均不能达到 1 : 1 的网络带宽需求.

表 2 展示了不同数据中心网络的性能指标,其中 n 表示层次结构中的层次数, k 表示交换机端口数, S 表示服务器数量.从表 2 可以看出,相比于其他数据中心网络,DPClos 网络具有更高的网络带宽,更小的网络直径和瓶颈度.除此之外,DPClos 中的服务器之间还存在着大量的并行路径.这些特点有利于提高基于 DPClos 结构的集装箱数据中心的网络性能和容错性.

表 2 数据中心网络结构比较

Tab. 2 Comparison of data center network architectures

	服务器端口数	网络直径	对剖带宽	瓶颈度
Fat-tree	1	6	$S/2$	$S-1$
DCell	$n+1$	$<2\log_k S-1$	$\frac{S}{4\log_k S}$	$S\log_k S$
BCube	$n+1$	$n+1$	$S/2$	$(k-1)\frac{S}{k}$
FiConn	2	$<2^{k+1}-1$	$\frac{S}{4 \cdot 2^k}$	$>S \cdot 2^k$
DPillar	2	$n+\lfloor \frac{n}{2} \rfloor$	$(\frac{k}{2})^n$	$\frac{3n(S-1)}{2}$
文献[10]	2	6	S	$\frac{S-1}{2}$
DPClos	2	4	S	$\frac{S-1}{2}$

3 模拟实验

对所提出的集装箱数据中心网络 DPClos 进行实验,通过数值实验观察该结构在容错性、传输延迟和吞吐量方面的性能.

3.1 实验设定

实验采用 $k=48$ 的 DPClos 结构,该网络由 144 台交换机和 1 152 个服务器组成.采用文献[11-12]中基于时间片的路由方法.数据流设定为单包流,即交换机和服务器的每个端口可以在一个时间片内完成一个单位的数据流发送.交换机以存储转发模式工作:当接收到数据流时交换机将其存储到等候队列;每个时间片内,交换机将队首的数据流发送到下一跳的节点,而该队列中其他数据流被延迟一个时间片.

容错性:按照一定的故障率随机设定链路和交换机故障,观察数据流的到达情况.对每一个设备故障率,随机生成 10 组,每组 10 000 个数据流,观察数据流的到达率.该故障率的数据流到达率为 10 组数据流到达率的平均值.对于每个故障率,随机选择 10 组故障设备,最终结果为 10 组数据的平均值.

传输延迟:对于给定的数据流量,计算组内数据流到达的平均时间作为该组数据流的传输延迟.统计各组数据流的传输延迟,取其平均值作为该组数据流的传输延迟.

吞吐量:吞吐量的计算方式为数据流量除以本组数据流中最大的数据传输时间.统计各组数据流中的吞吐量,取其平均值作为该组数据流的吞吐量.

在传输延迟和吞吐量上,考察两种流量模式,即随机模式和 Incast 模式.在随机流量模式中,随机选择每个数据流的源服务器和目的服务器.在 Incast 流量模式中,每个目的服务器从其他 10 个源服务器中接收数据,源服务器和目的服务器随机选择.考察数据流数量从 500 增长到 50 000 时传输延迟和吞吐量的变化情况.对给定的数据流数量,随机生成 10 组数据流进行测试.实验结果为 10 组数据流的平均值.

3.2 容错性

分别对 FDPClos 和 SDPClos 进行链路故障和交换机故障的容错实验,观察故障率 F_f 从 0 增长到 20%时数据流的到达情况.图 6 显示了在链路/交换机随机故障下数据流的平均到达率 R .

图 6 表明,随着设备故障率的增大,FDPClos 和 SDPClos 中数据流的到达率逐渐降低.从图中可以看出,FDPClos 和 SDPClos 具有相近的链路故障容错性.当链路故障小于 5%时,FDPClos 和 SDPClos 中数据流的到达率均大于 99%;当链路

故障率达到 20%时,数据流的到达率大于 87%.这说明,DPClos 网络结构在链路故障率极高的情况下,仍保持较高的数据流到达率.文献[13]显示,当 20%链路故障发生时,S2、SWDC 和 Jellyfish 的数据到达率分别为 85%、70%和 59%.

DPClos 网络也具备良好的交换机故障容错性.从图 6 中可以看出,当交换机故障率小于 5%时,数据流在 FDPClos 和 SDPClos 中的到达率均大于 99%.当交换机故障率达到 20%时,数据流在 FDPClos 和 SDPClos 中的到达率分别为 85%和 87%.这表明,DPClos 网络在 20%交换机故障率下仍然能够满足超过 85%服务器之间的数据通信.

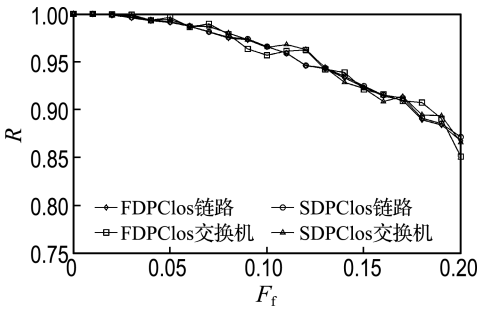


图 6 数据流到达率和链路/交换机故障率关系
Fig. 6 The flow success rate versus the fraction of link/switch failure

3.3 传输延迟

在随机和 Incast 流量模式下观察 FDPClos 和 SDPClos 中平均传输延迟 D 与数据流数量 N_d 的关系.图 7 显示了在这两种流量模式下数据流的延迟情况.

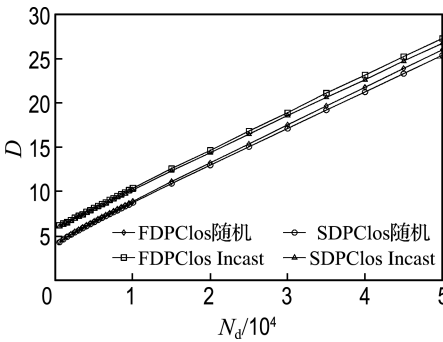


图 7 随机和 Incast 流量模式下 DPClos 网络的数据流延迟
Fig. 7 The delay of flows for random and Incast traffic in DPClos

在 FDPClos 和 SDPClos 中,数据流的平均路由路径长度为 3.96 和 3.92.从实验结果可以看

出,在两种流量模式下数据流的延迟均随着流数量增长而呈线性增加,且 Incast 流量模式下的平均延迟大于随机流量模式.在同一种流量模式下,数据流在 FDPClos 中的延迟略大于其在 SDPClos 中的延迟,这源于 SDPClos 具有更短的平均路由路径.

3.4 吞吐量

在随机和 Incast 流量模式下,观察 DPClos 网络的吞吐量 T 与数据流数量的关系.图 8 显示了 FDPClos 和 SDPClos 网络的吞吐量与数据流数量之间的变化关系.

从图 8 可以看出,DPClos 网络的吞吐量在这两种流量模式下均随着数据流数量的增长而增加.当数据流数量较少时,网络的吞吐量随着数据流数量近似呈线性增长;当数据流数量较大时,网络吞吐量增长速率有所降低,并趋于饱和.在同一种网络流量模式下,FDPClos 和 SDPClos 吞吐量相近.在两种网络中,随机流量下的吞吐量均大于 Incast 流量模式下的吞吐量.

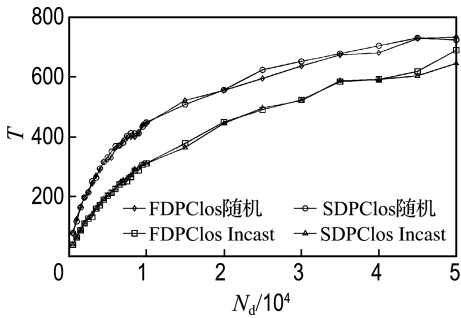


图 8 随机和 Incast 流量模式下 DPClos 网络吞吐量

Fig. 8 The throughput of DPClos for random and Incast traffic

3.5 实验结论及分析

分析网络结构和实验结果可知:在容错性方面,FDPClos 和 SDPClos 都具有较高的链路和交换机故障容错性,这源于 DPClos 中的双层交换机网络结构所提供的大量并行路径.

在延迟方面,SDPClos 在两种流量模式下的延迟均略小于 FDPClos. SDPClos 的优势来源于下层边界交换机的错位连接.这种连接使得网络中更多服务器之间的路径长度为 2.

在吞吐量方面,两种结构在相同流量模式下的性能相近.随机流量和 Incast 流量模式下网络性能差异来源于这两种流量模式的特性:Incast

模式会导致路由路径下游交换机产生拥塞,从而引起延时增加和吞吐量降低;随机流量则不存在这种拥塞特性.

FDPClos 和 SDPClos 采用相同的设备,具有相近的网络性能,但 SDPClos 能够提供更短的平均路由路径长度.这两种结构都能满足集装箱数据中心对网络性能和容错性的需求.

4 结 语

本文研究了高带宽高容错性的集装箱数据中心网络体系结构.通过分析集装箱数据中心的应用需求和现有数据中心网络的特点,提出了一种基于折叠的 3 级 Clos 的集装箱数据中心网络结构,简记为 DPClos.理论分析表明,DPClos 能够满足集装箱数据中心的网络性能需求.实验结果验证了 DPClos 结构在容错性、传输延迟和吞吐量方面的性能.在今后的工作中,将进一步完善 DPClos 网络配置和路由调度,并致力于应用部署.

参考文献:

- [1] YI Xiao-meng, LIU Fang-ming, LIU Jiang-chuan, *et al.* Building a network highway for big data: architecture and challenges [J]. **IEEE Network**, 2014, **28**(4):5-13.
- [2] Al-Fares M, Loukissas A, Vahdat A. A scalable, commodity data center network architecture [C] // **2008 ACM SIGCOMM Conference on Data Communication, SIGCOMM 2008**. New York: Association for Computing Machinery, 2008:63-74.
- [3] GUO Chuan-xiong, WU Hai-tao, TAN Kun, *et al.* DCell: a scalable and fault-tolerant network structure for data centers [C] // **2008 ACM SIGCOMM Conference on Data Communication, SIGCOMM 2008**. New York: Association for Computing Machinery, 2008:75-86.
- [4] GUO Chuan-xiong, LU Guo-han, LI Dan, *et al.* BCube: a high performance, server centric network architecture for modular data centers [C] // **2009 ACM SIGCOMM Conference on Data Communication, SIGCOMM 2009**. New York: Association for Computing Machinery, 2009:63-74.
- [5] Singla A, Hong C Y, Popa L, *et al.* Jellyfish: Networking data centers randomly [C] // **9th USENIX Symposium on Networked Systems Design and Implementation, USENIX NSDI 2012**. Berkeley:USENIX, 2012:225-238.

- [6] LI Dan, GUO Chuan-xiong, WU Hai-tao, *et al.* FiConn: Using backup port for server interconnection in data centers [C] // **2009 International Conference on Computer Communications, IEEE INFOCOM 2009**. Piscataway: IEEE Communications Society, 2009: 2276-2285.
- [7] WU Hai-tao, LU Guo-han, LI Dan, *et al.* MDCube: a high performance network structure for modular data center interconnection [C] // **5th International Conference on Emerging Networking Experiments and Technologies, ACM CoNEXT 2009**. New York: Association for Computing Machinery, 2009: 25-36.
- [8] 陆菲菲, 朱桂明, 陶志荣, 等. MDCent: 一种高可扩展, 高吞吐量的模块间互连结构[J]. 计算机研究与发展, 2015, **52**(5): 1127-1136.
- LU Fei-fei, ZHU Gui-ming, TAO Zhi-rong, *et al.* MDCent: a modular data center interconnection with high scalability and high performance [J]. **Journal of Computer Research and Development**, 2015, **52**(5): 1127-1136. (in Chinese)
- [9] Clos C. A study of non-blocking switching networks [J]. **Bell System Technical Journal**, 1953, **32**(2): 406-424.
- [10] 王 聪, 王翠荣, 王兴伟, 等. 面向云计算的数据中心网络体系结构设计[J]. 计算机研究与发展, 2012, **49**(2): 286-293.
- WANG Cong, WANG Cui-rong, WANG Xing-wei, *et al.* Network architecture design for data centers towards cloud computing [J]. **Journal of Computer Research and Development**, 2012, **49**(2): 286-293. (in Chinese)
- [11] LI Da-wei, WU Jie. On the design and analysis of data center network architectures for interconnecting dual-port servers [C] // **2014 International Conference on Computer Communications, INFOCOM 2014**. Piscataway: IEEE Communications Society, 2014: 1851-1859.
- [12] LI Da-wei, WU Jie. Dual-centric data center network architectures [C] // **44th International Conference on Parallel Processing, ICPP 2015**. Piscataway: IEEE Communications Society, 2015: 679-688.
- [13] YU Ye, QIAN Chen. Space shuffle: A scalable, flexible, and high-bandwidth data center network [C] // **22nd International Conference on Network Protocols, ICNP 2014**. Piscataway: IEEE Communications Society, 2014: 13-24.

DPClos: Design of network architecture for container data center based on 3-stage Clos structure

LI De-shun, SHEN Yan-ming, LI Ke-qiu*

(School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China)

Abstract: Owing to the convenience of deployment, container data center plays an essential role in emergency application of harsh environments. To satisfy the requirements of high network bandwidth and high fault tolerance, a novel container data center network based on the 3-stage Clos structure, DPClos is proposed. DPClos architecture takes advantage of Clos structure and the dual-port servers, which can provide double capacity of network bandwidth. The average path length is less than 4 and traffic load is balanced among links of different levels in DPClos. Bilayer structures benefit such fault tolerance of DPClos as follows: the flow success rate is more than 85% with a fraction of 20% equipments failure. Theoretical analyses and experimental results show that DPClos can meet the requirements of performance in container data center network.

Key words: container data center; data center network; network bandwidth; fault tolerance