

DCNS: 一种高可用性的数据中心网络

冷 飞 徐进华 栾仕喜

(苏州大学信息中心 苏州 215104)

摘 要 近年来,随着云计算技术的飞速发展,数据中心网络作为底层基础设施,发挥着越来越重要的作用。数据中心网络能够为各种云计算环境提供良好的服务。这几年在研究云计算的同时,数据中心也是一个重要的研究热点。学习了一些新的拓扑结构,主要包括 Fat-tree、DCell 以及 BCube 等。而这些方案在系统扩展太快或者太慢的情况下,都在一定程度上有一些性能的瓶颈或者在路由以及结构的构建上有比较高的代价。提出了一个代价有效并且具有很强扩展性和容错性的数据中心网络——DCNS。其混合了 DCell 以及 BCube 的优点,避免了它们各自的缺点。之后提出了一个针对 DCNS 的容错路由机制。最后,使用一个合适的测试环境来测试该数据中心网络架构的有效性。实验表明,DCNS 能较好地满足数据中心对网络结构的要求,并且能够保证系统的可用性。

关键词 数据中心网络,失效恢复,网络结构,可用性

中图分类号 TP319

文献标识码 A

DCNS: A High Available Data Center Network Topology

LENG Fei XU Jin-hua LUAN Shi-xi

(Information Center, Soochow University, Suzhou 215104, China)

Abstract In recent years, with the rapid development of cloud computing technology, the data center network which acts as the underlying infrastructure of cloud computing, plays an increasingly important role. But there are some failures in the network, which requires a strong and effective recovery mechanism to ensure data center available. The network structure plays an important role in the availability of the data center network. Data center network can provide various cloud computing services. As a result, data center networking has recently been a hot research topic in both academia and industry. A fundamental challenge in this research is the design of the data center network that interconnects the massive number of servers, and provides efficient and fault-tolerant routing service to upper-layer applications. In response to this challenge, the research community has begun exploring novel interconnect topologies including Fat-Tree, DCell, and BCube, etc. Extension of the proposed solutions is either too fast (i. e., double exponentially) or too slow, they suffer from performance bottlenecks, or can be quite costly in both routing and construction. This paper proposed a cost-effective, gracefully scalable, and fault-tolerant data center interconnect termed DCNS that combines the advantages of both DCell and BCube architectures while avoiding their limitations. We then proposed fault-tolerant and routing mechanisms for DCNS. Finally, we proposed a comprehensive benchmarking environment that can be used for accurately and practically evaluating and testing the proposed data center architecture. Experimental results show that DCNS can better meet requirements of the data center network architecture and ensure system availability.

Keywords Data center network, Failure recovery, Network structure, Availability

数据中心基础设施的设计在最近几年内已经成为了工业界及学术界的一个重要的研究热点,主要是因为其在应对一些快速增长的基于 Internet 的应用程序(包括搜索(Google)、在线视频(YouTube)、社会网络(Facebook)以及大规模计算(数据挖掘以及生物信息))的数据中心时起到了更为重要的作用。

特别是云计算(Amazon 的 EC2, Google 的 APP Engine 以及 Microsoft 的 Azure)的发展,使得对计算及基础设施的要求达到了高峰,其要求能够提供一种可扩展、灵活以及代价

有效的方法来满足各种变化的 IT 需求,不管是企业还是一般的公共服务都有这样的需求^[2,11]。大规模的数据中心能够为云提供核心设施,其占总代价的 45%^[2]。这就不可避免地需要获取一些数据中心的基础设施,包括数据中心网络的权利,从而使得对于基础设施的部署和维护是代价有效的。

在考虑数据的可用性以及安全性上,数据中心就会比以前的阶段上具有更重要的作用。这可以从数据中心的花费上得知,从新建立的标准可以看到数据中心的重要性也在提高。从工业上统计,美国的数据中心 2009 年的市场从 2005 年的

到稿日期:2013-03-12 返修日期:2013-06-04 本文受江苏省自然科学基金项目(102042)资助。

冷 飞(1974—),男,博士,实验师,主要研究方向为计算机技术、现代信息技术与信息安全,E-mail:lengfei_sz@163.com;徐进华(1972—),男,硕士,副教授,主要研究方向为嵌入式、管理信息系统;栾仕喜(1988—),男,硕士,主要研究方向为计算机网络。

162 亿美元增加到了现在的 390 亿美元。这样的市场趋势其实对于全世界都是一样的,例如打算建立一个能够基于云计算的 Internet 数据中心来为各种服务提供服务链,整个提供链将会被销售到海外的市场,其能够获得的价值以及市场机遇到 2014 年估计会达到 1000 亿美元^[7]。

传统的数据中心网络架构主要是基于上层机架(TOR)交换机与底层机架(EOR)交换机进行相互连接,并且它们之间通过核心交换机来转换连接。该方法不仅代价昂贵,同时也会导致核心网络的带宽消耗大。这就促使很多研究者提出了一些新的拓扑架构,包括 Fat-Tree^[2,11,15]、DCell^[3]、BCube^[4]、MDCube^[8]以及 Clos 网络^[1]。

虽然这些研究提供了一些正确的方向,但是仍然有一些没有解决的挑战问题,主要包括扩展性上的问题及容错上的问题。数据中心网络(Data Center Network, DCN)^[17]作为底层基础设施,发挥越来越重要的作用,成为云计算领域的研究热点。首先,在当前云计算环境下,数据中心网络的结点数目呈指数级增长。据不完全统计,到 2006 年底,运行在 Google 的集群上的服务器超过 450000 台^[17]。另外在应对节点或者链路的失效上对数据中心网络提出了更高的要求,其必须具有很强的容错性;因为在云计算环境下节点的失效是正常现象而不是异常情况。例如,在 Google 和 Amazon 数据中心网络上,数小时的结点或链路失效也会带来数百万美元的损失。这些应用对数据中心网络的构建与维护提出了 3 种要求:可扩展、容错和高带宽吞吐量。

本文在分析现有的网络结构对数据中心网络中节点失效的影响的基础上,针对数据中心网络在云计算环境下的重要挑战,提出一种新的数据中心网络结构 DCNS。在 DCNS 结构下能够使用一些普通的交换机提供一种灵活的扩展方式,并且混合了 DCell 以及 BCube 架构的优点,从而避免它们各自的缺点。

1 相关工作

对于数据中心网络架构的研究,在学术界以及工业界已经有了很多的研究成果。传统数据中心网络采用的架构是树形分层结构^[9](如图 1 所示),其主要包括边沿层、聚集层、核心层 3 层结构。在树形的末端,一个机架上通常存放数 10 台服务器,这些服务器通过边沿层的交换机连接到网络上。在汇聚层和核心层,为了提供尽可能高的性能,采用了造价极高的高端交换机(10 GE 级别)形成高连通网络拓扑。现有的这些研究成果大部分是基于该传统结构的,根据其演化的来源,主要可以分为从并行计算上进行演化以及在 Internet 交换机和路由器上进行演化两种情况。

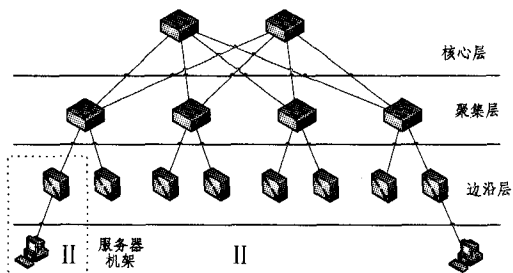


图 1 传统数据中心网络架构

1.1 并行计算演化的数据中心网络架构

在最近的研究中所提出的可扩展、容错的数据中心网络

架构,包括 DCell^[3]、FiConn^[6]、BCube^[4]以及 MDCube^[8],都是基于一些来源于并行计算^[10]的交叉网络的文献,还包括一些以前的研究^[12-14]。

DCell 是一种递归定义架构。在 DCell 中的服务器具有多个端口。每个服务器不仅和一个 mini-switch 连接,还通过通讯链接与其他的服务器进行连接。DCell₀ 是一个基本的块,从而来构建一个更大的 DCell。其有 n 个服务器并且它们将会与 $n-1$ 个交换机的端口进行连接。而 DCell_k 由 $a_{k-1}+1$ 个 DCell_{k-1} 组成,其中 a_{k-1} 代表 DCell_{k-1} 下的服务器数量。因此,DCell 架构的规模将是指数级的。

FiConn 是另外一个递归架构的例子。它要求每个机器必须具有两个端口,并且能够容纳百万级别规模的服务器。FiConn 的基本单元是 FiConn₀,其与 DCell₀ 类似。在 FiConn 中每个计算机将会用一个端口连接 FiConn₀ 的一个交换机,其被叫做 0 层端口。而该计算机的备份端口用于连接其他的计算机。假设在 FiConn_{k-1} 中有 b 个可用的备份端口。为了构造 FiConn_k,必须使用 $b/2$ 个端口来连接 $b/2$ 个其他的 FiConn_{k-1}。因此,一个 FiConn_k 将会有 $(b/2+1)$ 个 FiConn_{k-1}。

BCube 是一个服务器中心网络架构。在 BCube 中主要有两种类型的设备——具有多个网络端口的服务器以及用于连接不同层次服务器的 mini-switch。一个 BCube_k 是递归地从 n 个 BCube_{k-1} 以及 n^{k-1} 个具有 n 个端口的交换机形成。在 BCube 结构中,一个交换机从来都没有和其他的交换机进行直接的链接,它们只是用来转发请求。BCube 由很多的电线、交换机和高性能的电缆组成。

MDCube^[8]是基于 BCube 的模块数据中心(MDC)进行扩展,从而支持更多的服务器。其使用 MDC 作为构建 MDCube 的基本块。在这样的架构中,每个容器具有 m 个元组。一个容器将会与那些在标签上只有一个不同的容器进行连接。

PortLand 架构是一个 multi-rooted 的网络拓扑,是 Fat-Tree 的演化。但是它具有一些性能瓶颈,并且容错拓扑较弱。如果中心架构管理器出错,则整个 PortLand 模式将会出错。

1.2 Internet 交换机和路由器演化的数据中心网络

有很多的数据中心网络架构都是基于 Internet 交换机和路由器的原始设计。这些结构主要包括 VL2^[11]和 DOS^[16]。

一个 VL2 网络是分配到一个 Clos 拓扑的多个交换机上设计的,从而能够支持多种大量路径。可以使用灵活的负载均衡来扩展多个服务器之间的目标独立流量。VL2 能够对任何两个机器提供相同的高容量。但是机器数量的扩展被最大端口数量和中间交换机所限制。

DOS 使用基于波导阵列路由器的坡长特征为数据中心设计了一个可伸缩的光学交换机。它们采用多个坡长的优点来描述,而这个架构将会比电子交换机具有更好的性能。

2 DCNS 数据中心网络结构

2.1 设计动机

在一般的情况下,一个数据中心网络架构必须综合考虑扩展性、可用性、复杂性以及容错性来进行完整的衡量。但是在实际生活中,这几个请求之间会有一定的冲突。因此在数据中心网络架构的设计中需要对它们进行一些折中。例如,

扩展性和容错性会在很大程度上影响系统的性能以及整个数据中心网络的代价,因此其很重要。即使具有一些不同点,以前的架构或者算法可以根据它们的扩展性和容错性被分成3类(在固定配置下服务器的数量): $O(n^c)^{[1,2,11,15]}$, $O(n^k)^{[4]}$ 以及 $O(n^{(2^k)})^{[3,6]}$,其中 c 是一个常数; n 是一个交换机以及路由器的端口数; k 代表节点或者服务器的度或者网络层次的数量。

可扩展及容错的数据中心网络架构必须通过接口卡以及中间交换机等使用最小代价来组成百万数量的服务器。另外,其必须在一个容错环境下产生平缓的性能。很多传统的数据中心网络虽然能够提供很好的扩展性($O(n^{(2^k)})$),但是却很难产生同等扩展的性能。在实际情况下,可能会使得部分数据中心网络空闲,从而可以进行简单的扩展。在这样的数据中心中通常会有很高的无效性,特别是对于那些指数规模的数据中心网络,例如将一千服务器数据中心网络扩展到一百万的服务器规模。

在另外一种情况下, $O(n^c)$ 和 $O(n^k)$ 的扩展速度是独立于实际的,以下是一些观察的结果。

- 最大允许的节点度(k)一般都小于6。这实际上是限制了实际中服务器能够拥有的接口卡必须小于6。
- 交换机的端口数(n)有明显的限制并且更倾向于选择小的值(4或者8)。因为交换机的价格受端口数的影响是相当大的^[9]。一个小端口的交换机比大端口数的交换机具有更好的代价有效性。

从上面的观察上可知,类似于BCube这样的指数级规模的数据中心网络都存在扩展性问题。例如,以 $n=4$ 以及6层来构造一个数据中心,则可以拥有 $4^6=4096$ 个服务器,而对于现在数据中心网络的需求还差很大。

本文从各个角度来分析扩展性和容错性问题。主要是判断对于小数量端口的中间交换机的中心网络其服务器的限制是否为 $O((an)^k)$,其中 a 是一个常数,用于与 n 进行比较。

2.2 DCNS 架构

首先将充分利用 $O((an)^k)$ 的优点,我们将使用一个特殊的情况即 $a=n$ 。本文提出了一种新的互连网络架构——DCNS,其扩展的速度是 $O(n^{2k})$ 。

DCNS的第1层包括 n 个节点和一个具有 n 个端口的交换机。其与DCell₀是类似的。从第2层开始,一个 k 层($k \geq 2$)的DCNS将由 n^2 个 $k-1$ 层的DCNS组成。在实际中,一个 $k(k \geq 2)$ 层的DCNS可以被看成一个 $n^2 * n^{2k-3}$ 的矩阵,其中每行是一个 $k-1$ 层的DCNS,其由 n^{2k-3} 个节点组成。另外,其可以被看成具有 n^{2k-3} 列,并且每个列包含 n^2 个节点,而其就是拥有 n^2 个 $k-1$ 层的DCNS。在这个具有 n 个端口的 n 个交换机上,基于列的连接被引入来连接 n^2 个节点。这 n 个交换机的连接模式如表1所列。其拥有两种不同类型的连接模式,它们经常会交织在一起。图2是具有4个端口的交换机的 k 层($k \geq 2$)DCNS的前面两列的连接模式(这里只给出第 k 层交换机的模式)。

表1 相应交换机中第 i 列的目标行数量

第 x 个交换机	第 j 个端口, $(i \bmod 2)=1$	第 j 个端口, $(i \bmod 2)=2$
	$(x-1)n+j$	$(j-1)n+x$

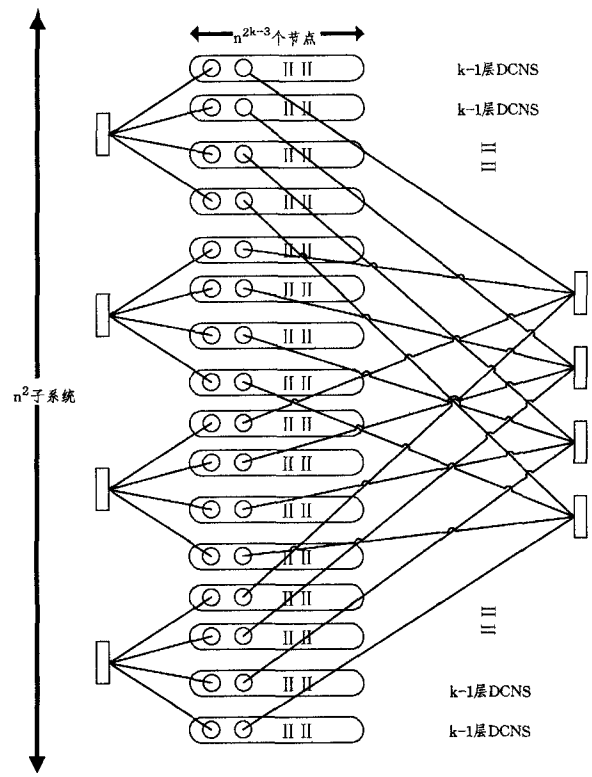


图2 具有4个端口交换机的 $k(k \geq 2)$ 层DCNS的前两列的连接模式

2.3 路由及容错算法

DCell以及其他的架构证明本地重路由算法的效率较低^[5]。一个容错的路由路径可以通过一个特定的层次网络来转换到重复路径长度为 $O(2^k)$ 的网络中。相反,在DCNS中的路由算法更为有效,其容错的路径能够在 $k(k > 1)$ 层的DCNS中只传输一次。

在 k 层,DCNS可以使用 k 个坐标来标记 $(C_k, C_{k-1}, \dots, C_1)$,其中 C_j 代表该节点在一个 j 层的DCNS中的第 C_j 行。本文DCNS使用一个基于行的层次路由。因此给定一个节点对,源节点 $(S_k, S_{k-1}, \dots, S_1)$ 和目标节点 $(D_k, D_{k-1}, \dots, D_1)$ 的路径可以通过以下 k 个步骤来建立。其中每个步骤中只有一个坐标被使用: $(S_k, S_{k-1}, \dots, S_1) \rightarrow (D_k, ?, \dots, ?) \rightarrow \dots \rightarrow (D_k, D_{k-1}, \dots, ?) \rightarrow (D_k, D_{k-1}, \dots, D_2, ?) \rightarrow (D_k, D_{k-1}, \dots, D_1)$ 。其中“?”代表未知的值。

每一步可能需要更多的转换。例如一个路由转换 $(\dots, S_j, \dots)$ 到 $(\dots, D_j, \dots)$ 。我们描述一个从第 S_j 行到第 D_j 行的容错路由路径。首先, S_j 和 D_j 可以用式(1)表示。

$$F(z) = \begin{cases} n - (z \bmod n), & \text{if } (z \bmod n \neq 0) \\ 0, & \text{if } (z \bmod n = 0) \end{cases} \quad (1)$$

$$S_j = \left\lceil \frac{S_j}{n} \right\rceil * n - F(S_j) \quad (2)$$

$$D_j = \left\lceil \frac{D_j}{n} \right\rceil * n - F(D_j) \quad (3)$$

而第 S_j 行和第 D_j 行之间的距离可以由式(4)表示。

$$D_j - S_j = \left(\left\lceil \frac{D_j}{n} \right\rceil - \left\lceil \frac{S_j}{n} \right\rceil \right) * n - [F(D_j) - F(S_j)] \quad (4)$$

如果没有连接错误,通过 $[F(D_j) - F(S_j)]$ 和 $(\lceil D_j/n \rceil - \lceil S_j/n \rceil)$ 对相邻的列使用0或者3次的转换就可以建立一个第 S_j 行到第 D_j 行的容错路径。当 $(\lceil D_j/n \rceil - \lceil S_j/n \rceil) \neq 0$ 时,通过第 i 个列的转换是需要的,其中 $(i \bmod 2) = 0$ 。同时,若 $[F(D_j) - F(S_j)] \neq 0$,第 i 个列的一次转换是有必要的,其

中 $(i \bmod 2)=1$ 。另外,当上面的两种转换都是必要的时候,第 i 个列到第 $(i \pm 1)$ 个列的转换是必要的。因此,在第 $j(j > 1)$ 层的 DCNS 中最多有 3 次转换来加强一个坐标。一步路由的时间复杂度为 $O(1)$,在 k 层的 DCNS 中第 k 层的容错路由事件复杂度为 $O(k)$ 。

在 DCNS 架构中,容错路由可以通过网络中各种层次发现的引导路由的交叉路径来实现。在 DCNS 中有两种不同的基于列的连接模式,而这两种模式一般情况下会相互交织在一起。在一个 $k(k \geq 2)$ 层的 DCNS 网络中有 $n^2 * n^{2k-3}$ 个节点,则任何行对中最少具有 $(n^{2k-3}/2)$ 个实体和并行路径。换言之,必须在 $2m$ 列中有一个远离当前列的交叉路径。 m 决定了网络中哪一层已经被选为引导。

DCNS 的容错路由算法如算法 1 所示,其由 k 个单步路由组成。每一个单步路由是递归定义的。其返回一个最大跳数为 $2^{\text{RetryTimes}}$ 的合理路径。在给定的多个合理路径中,选择在序列中的第一个可用路径。举例来说,它经常从目前交换机的下一个端口开始。对于每个单步路由,RetryTimes 值缓慢增长来保证合理路径的最短长度。DCNS 的整个路由算法可以被看作是约束的宽度有效搜索(BFS)。与 BCube 采用的传统 BFS 不同,其复杂度为 $O(|V| + |E|) = O((k' + 1)|V|)$,DCNS 的容错路由算法的复杂度为 $O(k * 2^{\text{RetryTimes}}) = O(|V| * kn^{2(\text{RetryTimes}-k)+1})(k' = 2k-1)$,这是由其无回滚路由模式所决定的。给定一个 n 和 RetryTimes 的小值,这样的复杂度在实际上是很小的,而且比以前算法例如 DCell 以及 BCube 的性能会好些。

算法 1 DCNS 的容错路由算法

```

1. Single-Step-Routing( $S_j, D_j, j, \text{RetryTimes}$ ) {
2.   if(RetryTimes=0)
3.     return null;
4.   if( $S_j = D_j$ )
5.     return 一个 0 跳的路径;
6.   else {
7.     在序列中通过直接连接  $k$  层交换机来尝试 1 跳的路径,如果成功,return 一个一跳路径;
8.     首先通过直接连接第 1 层的交换机来尝试两跳路径,然后再通过一个  $j$  层交换机,如果成功,return 一个两跳路径;
9.     使用一跳以及两跳中可达的节点作为中介,尝试进行 Single-Step-Routing(intermediate,  $D_j, j, \text{RetryTimes}-1$ ),如果成功,return 一个混合路径;
10.    if(all fail)
11.      return null;
12.    }
13. }
Routing(Src, Dest, RetryTimes) {
1. 调用 Single_Step_Routing( $S_j, D_j, j, \text{RetryTimes}$ )  $k$  次;
2. for 每个单跳路由
3.   缓慢增加 RetryTimes 的值,从而保证合理路径中的最短长度
4. }
```

3 实验结果及分析

本文提出的 DCNS 数据中心架构具有高扩展、高容错以及高带宽吞吐量的优点。本节通过一些实验来验证该结构的优点,将其与 DCell 以及 BCube 的以下两种性能指标进行对比:

(1) 平均路径长度,在不同规模和组件的不同失效比率情

况下,求出每个节点与其他节点之间最短路径长度的平均值。

(2) 平均带宽吞吐量,即在单位时间内接收者接收到的数据量,在本文实验的 all to all 通信模式中,平均带宽吞吐量是指所有接收者单位时间内接收到的数据量。

在对这 3 种结构的平均路径长度进行对比时,主要是比较在节点完好情况下不同规模($n=4, 0 \leq k \leq 3$)的平均路径长度。另外一种情况是考虑每个结构都具有相同规模的服务器(4096),不同的失效比例下的平均路径长度。

而在对比 3 种结构的平均带宽吞吐量上,数据中心的链路带宽被设置为 1Gb/s,同时每个服务器需要给其他所有结构发送 10Gb 的数据。首先在服务器规模为 16、无部件失效的情况下,对比 3 种结构完成数据通信所需的时间,然后在服务器不同失效比率、结点数目为 4096 的情况下,对比 3 种结构的带宽吞吐量。

3.1 平均路径长度(APL)测试

随着 DCNS 层次的不断提高,支持的服务器数量也不断增加。为验证 DCNS 的可扩展性,实验分两种情况对这 3 种结构的平均路径长度进行测试对比。(1)在整个数据中心完好的情况下平均路径长度随规模变化的情况;(2)在服务器不同失效比例下平均路径长度的变化。图 3 表示在无失效情况下,随着规模的增大,3 种结构所支持的服务器数量和平均路径长度都呈现不断增加的趋势。当 $k=0$ 时,服务器之间的通信都是通过交换机来进行的,因此它们的平均路径长度将都是 1。但是随着 k 的增大,DCell 的平均路径长度的增长速度比其他两种的快,DCNS 和 BCube 的增长会相对慢些。另外,在初始阶段平均路径长度 DCNS 会比 BCube 大,但随着 k 的增长逐渐接近,当 $k=3$ 时,已接近相等。

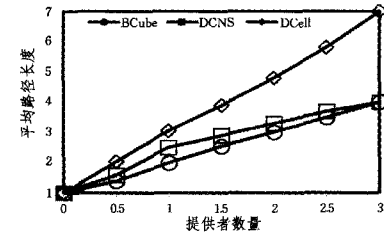


图3 平均路径长度随 k 的变化($n=4$)

图4是在不同失效比例下的平均路径长度。从图4结果可知,随着失效率的增加,所有架构的平均路径长度都会变大。但是 DCNS 相对于其他架构其平均路径长度最小。BCube 的平均路径长度与 DCNS 的比较接近。而 DCell 只有一层的交换机,则所有的路由都依赖服务器,因此其效果最差。而 BCube 全部依赖于交换机,所以效果次之。但是 DCNS 中服务器之间有更多的并行链路,既可以通过交换机中继,也可以通过服务器直接通信,所以在服务器失效比例增加时,其平均路径长度最短。

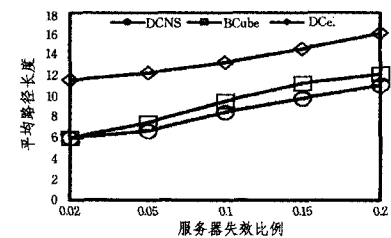


图4 平均路径长度随服务器失效比例的变化趋势

从上面的两个实验结果可知,不论在何种情况下,3 种结构中 DCNS 的平均路径长度都较小,并且具有较好的扩展能力。

平均带宽吞吐量是横向数据中心网络中服务器之间通信的核心因素。因此,本文通过一些实验来检测服务器之间通信的平均带宽吞吐量,以衡量数据中心网络的容错性。

图5是3种结构在无节点失效下间隔50s的情况下平均带宽吞吐量随时间的变化趋势。由图5可知,在初始化状态下,DCNS的带宽吞吐量最大,而另外的两种结构是近似的。随着时间的变化,所有结构的平均带宽吞吐量都会下降,但它们通信结束的时间不一样。在DCNS中不仅有通过交换机中继的路径,也存在服务器之间的直接路径,因此其将最早结束通信。BCube中仅存在通过交换机中继的路径,所以次之。而DCell仅存在服务器之间的直接路径,因此将最晚结束。

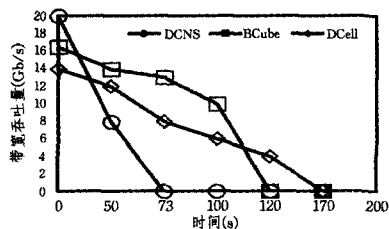


图5 带宽吞吐量随时间的变化情况

图6是服务器在不同失效比例的情况下,3种结构的平均带宽吞吐量的变化趋势。从图6可以看出,一旦服务器的失效比例增加,则系统的平均带宽吞吐量会减少。随着服务器失效比例增加,3种结构的带宽吞吐量都在减少。DCNS服务器之间有更多的并行路径,其平均带宽吞吐量仍然最大,而BCube次之,DCell由于较为均衡的负载,使其具有最低的吞吐量。总体来看,DCell的下降趋势比较平缓而其他两种都相对较大,而BCube和DCNS则相对明显一些,但并不影响它们的优势,仍然具有很好的效果。

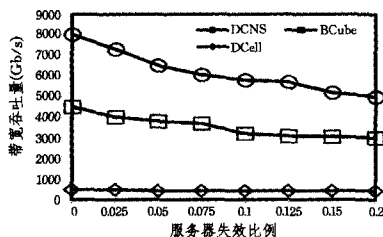


图6 带宽吞吐量随服务器失效比例的变化情况

所以,从上面的两个实验结果可以知道,不管是在服务器完整还是有节点失效的情况下,本文提出的DCNS结构都具有较高的平均带宽吞吐量,因此该结构具有更好的容错性。

从以上的所有实验可以知道,DCNS相比于其他两种结构将具有更好的扩展性以及容错性。因此,其将会显著提高数据中心网络的可用性。

结束语 数据中心网络的可扩展性和容错性,以及高带宽吞吐量的性能需求给拓扑结构的设计带来了巨大的挑战。层次式结构被广泛应用于大规模数据中心网络拓扑结构的构造中,其中,BCube结构和DCell结构能很好地解决其中的大部分问题,但也各自存在缺陷。本文提出的数据中心网络结构DCNS能够使得服务器之间存在多条并行路径,从而使DCNS在部分服务器节点失效的情况下,仍能保持较短的平均路径长度和较高的带宽吞吐量。通过实验验证,DCNS不仅具有较好的可扩展性,而且与已有的DCell和BCube相比能提供更好的容错性和更高的带宽吞吐量,能较好满足数据中心对网络结构的要求以及对于节点可用性的要求。

- [1] Greenberg A, Hamilton J R, Jain N, et al. VL2: a scalable and flexible data center network[J]. ACM SIGCOMM Computer Communication Review, ACM, 2009, 39(4): 51-62
- [2] Greenberg A, Hamilton J, Maltz D A, et al. The cost of a cloud; research problems in data center networks [J]. ACM SIGCOMM Computer Communication Review, 2008, 39(1): 68-73
- [3] Guo C, Wu H, Tan K, et al. Dcell: a scalable and fault-tolerant network structure for data centers[J]. ACM SIGCOMM Computer Communication Review, ACM, 2008, 38(4): 75-86
- [4] Guo C, Lu G, Li D, et al. BCube: a high performance, server-centric network architecture for modular data centers [J]. ACM SIGCOMM Computer Communication Review, ACM, 2009, 39(4): 63-74
- [5] Sarkar C. Seminar Report on Data Center Network Design, Technical Report[OL]. http://www.cse.iitb.ac.in/~chayan/seminar/final_report_dcn.pdf, 2013
- [6] Li D, Guo C, Wu H, et al. FiConn: Using backup port for server interconnection in data centers[C]// INFOCOM 2009. IEEE, 2009: 2276-2285
- [7] Wu H, Lu G, Li D, et al. MDCube: a high performance network structure for modular data center interconnection[C]// Proceedings of the 5th international conference on Emerging networking experiments and technologies. ACM, 2009: 25-36
- [8] IT247 [OL]. <http://www.it247.com/category/Network-Storage-Comms-Routers-Switches-Security-Routers-Switches.html>
- [9] Bhuyan L N, Agrawal D P. Generalized hypercube and hyperbus structures for a computer network [J]. IEEE Transactions on Computers, 1984, 100(4): 323-333
- [10] Al-Fares M, Loukissas A, Vahdat A. A scalable, commodity data center network architecture [J]. ACM SIGCOMM Computer Communication Review, ACM, 2008, 38(4): 63-74
- [11] Hamdi M, Hall R W. RCC-Full: An effective network for parallel computations [J]. Journal of Parallel and Distributed Computing, 1997, 41(2): 139-155
- [12] Hamdi M. A class of recursive interconnection networks: architectural characteristics and hardware cost [J]. Circuits and Systems I: Fundamental Theory and Applications, IEEE Transactions on, 1994, 41(12): 805-816
- [13] Hamdi M, Hall R W. Image Processing on Augmented Mesh-Connected Parallel Computers [J]. Journal of Computer and Software Engineering, 1994, 2(3): 329-348
- [14] Niranjan Mysore R, Pamboris A, Farrington N, et al. PortLand: a scalable fault-tolerant layer 2 data center network fabric[J]. ACM SIGCOMM Computer Communication Review, ACM, 2009, 39(4): 39-50
- [15] Ye X, Yin Y, Yoo S J B, et al. DOS: A scalable optical switch for datacenters[C]// Proceedings of the 6th ACM/IEEE Symposium on Architectures for Networking and Communications Systems. ACM, 2010: 24
- [16] Ghemawat S, Gobiolf H, Leung S T. The Google file system[J]. ACM SIGOPS Operating Systems Review, ACM, 2003, 37(5): 29-43
- [17] 邓维, 刘方明, 金海, 等. 云计算数据中心的新能源应用: 研究现状与趋势[J]. 计算机学报, 2013, 36(3): 582-598
- [18] 廖彬, 于炯, 孙华, 等. 基于存储结构重配置的分布式存储系统节能算法[J]. 计算机研究与发展, 2013, 50(1): 3-18
- [19] 熊润群, 罗军舟, 宋爱波, 等. 云计算环境下 QoS 偏好感知的副本选择策略[J]. 通信学报, 2011, 32(7): 93-102
- [20] 黄志宏, 巫莉莉, 钟焯荣, 等. 聚合增强型以太网环境下的数据中心建设[J]. 重庆理工大学学报: 自然科学版, 2013, 27(8): 70-75