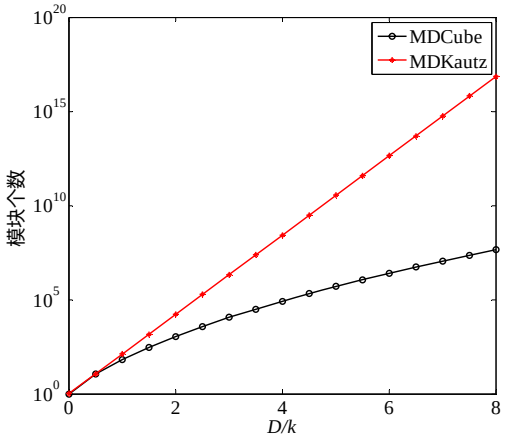

Fig.7 Number of modules is a function of each parameter

图 7 模块个数随相应参数的变化关系

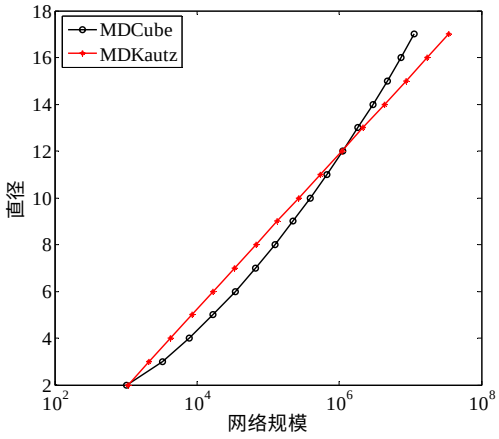

Fig.8 Diameter is a function of network size

图 8 网络直径随网络规模的变化关系

定理 **4**. $M(n,m,k)$网络的二分宽度为 $\frac{n^{k(m+1)}}{k2^{k+1}}\left(\frac{1}{2}n^{m+1}+1\right)$.

证明:由 Kautz 图的性质 3 和 $d=\frac{1}{2}n^{m+1}$,容易得证.

在便于分析,同时不影响分析结论的前提下,假设 MDCube 在各维度上互连的模块数量相等,则其二分宽度可以表示为 $\begin{cases}\frac{N^2-1}{4}N^{D-1}, & N\text{为奇数}\\ \frac{N^2}{4}(1+N^{D-1}), & N\text{为偶数}\end{cases}$,其中, $N=\frac{n^m(m+1)}{D}+1$ 表示每个维度包含的模块个数. □

图 9 给出了单个模块的规模相同时,两种结构的二分宽度随网络规模的变化情况.可以看出,两种结构的二分宽度均随网络规模的增大而增大,MDKautz 的二分宽度明显优于 MDCube.

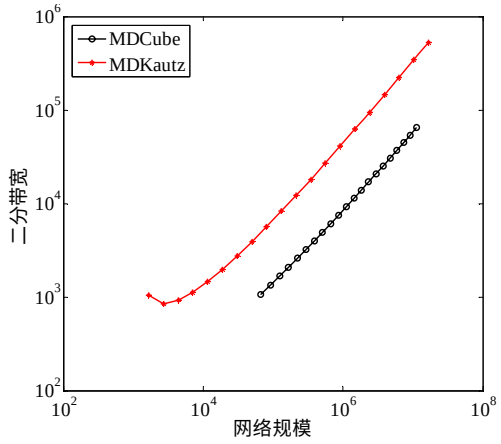

Fig.9 Bisection bandwidth is a function of network size

图 9 二分宽度随网络规模的变化关系

4.2 流量分布与ABT

All-to-All 通信模式广泛应用于 MapReduce 等应用中.ABT(aggregate bottleneck throughput)是数据中心网络中衡量 all-to-all 通信的一个评价指标,表示网络内各数据流获得的最小带宽与网络内所有数据流总和的乘

积,反映了数据中心网络在 all-to-all 通信模式下的网络容量.ABT 较大,也意味着 all-to-all 通信的时长更短.在 all-to-all 通信模式下,数据中心网络中的流量由模块内数据流和模块间数据流两部分组成,模块内数据流指每台服务器与同一模块内的其他所有服务器进行通信时产生的数据流;模块间数据流指每台服务器与其他所有不同模块内的所有服务器进行通信产生的网络流.由此,MDKautz 网络和 MDCent 网络中数据流的分布情况可用如下定理表示.

定理 **5**. 假设整个 MDKautz 网络都处于 all-to-all 的流量模式下,即网络中任意两台服务器之间存在一对数据流.若采用 MDKautzRouting 算法,则每条普通链路上承载的数据流的数量为

$$\left(2m+1-\frac{2m}{n}+\frac{n-1}{n}(k-1)(m+1)\right)\frac{(N-t)}{n}.$$

每条高速链路上承载的数据流的数量为 $(k-1)(m+1)\frac{(N-t)}{n}$,其中,$t$ 和 N 分别为 CH 模块和 MDKautz 网络中服务器的数量.

证明:根据 CH_m 结构的性质 3 可知,单个 CH_m 模块的服务器个数 $t=n^m(m+1)$,交换机个数 $g=n^{m+1}$,由 $M(n,m,k)$ 结构的定义可知:网络包含的模块总数为 M,网络规模 $N=tM$.有向普通链路的个数为 $2nN$,有向高速链路的个数为 $\frac{1}{2}gM$.在 all-to-all 流量模式下,MDKautz 网络中所有模块间数据流的数量为 $M(M-1)t^2$.

在 CH_m 网络中,所有服务器到任意交换机 A 的路径长度可以分成 i 组,第 G_i 组包括所有距离 A 为 $2i-1$ ($i\in[1,m+1]$)跳的服务器.每组中的服务器数量可表示为 $(m+1)C_m^i(n-1)^i$,则服务器到交换机的平均路径长度为

$$h_1=\frac{1}{(m+1)n^m}\sum_{i=0}^{m}[(2i+1)(m+1)C_m^i(n-1)^i]=2m+1-\frac{2m}{n}.$$

同理,计算出任意两台交换机之间的平均路径长度 $h_2=\frac{1}{n^{m+1}-1}\sum_{i=1}^{m+1}[2iC_{m+1}^i(n-1)^i]=\frac{2(n-1)t}{n(t-1)}(m+1)$.

在 $M(n,m,k)$中,任意两台服务器节点之间最多经过包括源、目的模块在内的 $k+1$ 个模块,由于中间模块通常为几十到数百不等,所以本文忽略目的模块中服务间的数据流,并认为经过的中间模块的平均数量为 $k-1$,则平均每条模块间数据流所要经过的普通链路的数量近似可表示为 $2h_1+(k-1)h_2$.由于 CH 中所有链路的使用都是平等和均衡的,因此,每条普通链路上承载的流的数量为

$$\frac{(2h_1+(k-1)h_2)M(M-1)t^2}{2nN}=\left(2m+1-\frac{2m}{n}+\frac{n-1}{n}(k-1)(m+1)\right)\frac{(N-t)}{n}.$$

同理,每条高速链路上的流的数量为 $\frac{kM(M-1)t^2}{\frac{1}{2}gM}=2k(m+1)\frac{(N-t)}{n}$. □

由定理 5 可知,普通链路和高速链路传输的数据流的数量不同.因此,要实现无阻塞地 all-to-all 通信,二者所需提供的链路带宽也不尽相同.高速链路和普通链路所承载的数据流数量之比,决定了 all-to-all 通信时模块间互联结构与模块内互联结构所需提供的聚合带宽,也决定了整个模块化数据中心网络是否存在性能瓶颈.由定理 5 可得出,这一比率为 $r=\frac{m+1}{2m+1-\frac{2m}{n}+k\frac{n-1}{n}(m+1)}$.

因此,对于一个由 16 512 个 CH 模块构成的 $M(2,7,2)$网络,假设普通链路的容量为 1,那么 all-to-all 流量模式下其所需的高速链路容量为 2.7Gbps,即高速链路只需提供 2.7Gbps 的带宽即可避免成为瓶颈,实现无阻塞的 all-to-all 传输.而目前普通商用交换机单个高速端口的带宽为 10Gbps,因此,高速链路不会成为整个 MDKautz 网络的性能瓶颈.

从定理 5 的证明过程中还可以得出以下推论.

推论 **1**. MDKautz 网络的瓶颈链路为模块内普通链路,故其 ABT 为 $N\left(\frac{2m+1}{n}+\frac{(k-1)(n-1)(m+1)-2m}{n^2}\right)^{-1}$.

图 10 给出了 MDKautz(k=2)与 MDCube(D=2)的 ABT 随网络规模的变化规律,二者的 ABT 均随网络规模的增大呈线性增长趋势,当 n>6 时,MDKautz 的 ABT 大于 MDCube 的 ABT.

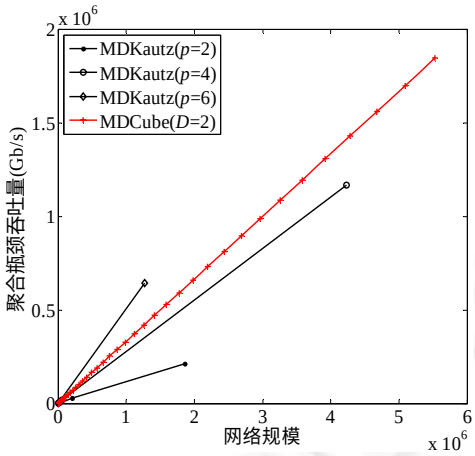

Fig.10 ABT is a function of network size

图 10 ABT 随网络规模的变化关系

5 模拟实验

在超大规模数据中心中,服务器和交换机都不可避免地遇到各种故障,而且故障发生时往往无法即刻定位并修复.因此,我们更关注存在网络失效的情况下模块内、外的容错路由策略是否能够确保 MDKautz 网络的性能呈现平缓下降的趋势.在本节中,我们在 Eclipse6.0 平台下编写了模拟程序,选取 k=2 时的 MDKautz 网络作为评估对象,选取 D=2 时的典型 MDCube 网络作为比较对象.令模块内部普通链路的速率为 1Gbps,而不同模块间高速链路的速率为 10Gbps.

第 1 个实验测试不同规模的 MDKautz 网络中 2 个模块间通信的 ABT.该实验模拟 MapReduce 的 reduce 过程,即每个 reducer 从所有 mapper 中取回数据,产生一个 all-to-all 的流量模式.在该流量模式下,当服务器/交换机的失效率从 0%达到 20%时,观察 2 个模块间通信时网络的 ABT 的变化情况.假设模块内部普通链路的速率为 1Gbps,而不同模块间高速链路的速率为 10Gbps(如图 11 所示).

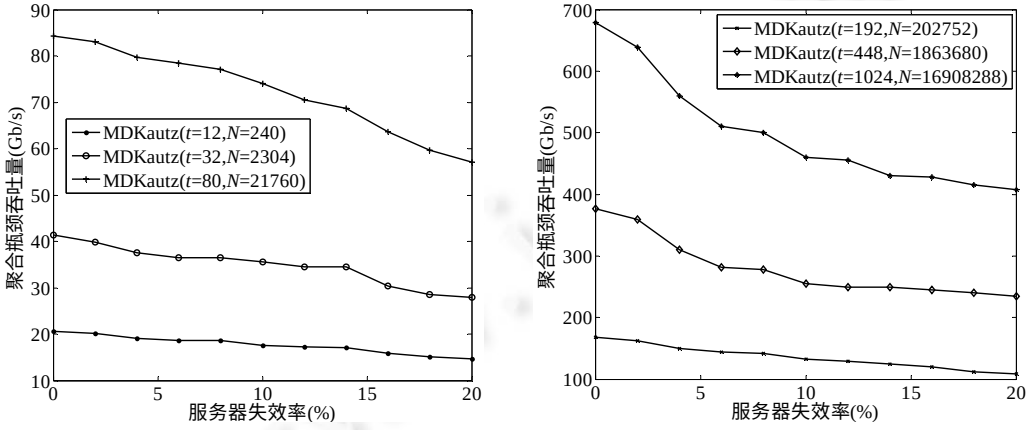

Fig.11 ABT of two modules in MDKautz under server failures

图 11 MDKautz 中 2 个模块间通信的 ABT 随服务器失效率的变化曲线

从图 11 中可以看出:单个模块的规模 t 越大,MDKautz 网络中 2 个模块之间的 ABT 越大.因为当服务器的网络端口数一定时,单个模块中交换机的数量随规模的增大而增加,因此能够提供给模块间的通信链路也随之增多.与此同时,随着链路故障率的上升,MDKautz 网络的性能始终是平缓下降的,这是因为 MDKautz 网络的模块内部和模块之间均存在多条并行路径,因此任意一对服务器间存在多条富连接,在网络失效的情况下,冗余的通信路径确保任意两台服务器间总有一条连通的链路.根据 ABT 还可以算出每台服务器实际获得的吞吐量,以规模为 2 304 为例,当链路失效率为 0 时,2 个模块间(共 64 台服务器)的最大可达聚合吞吐量为 45.304.因此,在网卡线速为 1Gbps 的条件下,每台服务器实际获得的吞吐量为 0.71Gbps.

在目前提出的模块间互连结构中,MDCube 结构拥有最优网络容量和容错性能,因此,第 2 个实验比较单个模块规模相同、网络规模不同的 MDKautz 和 MDCube 网络中 2 个模块间通信的 ABT.结果表明,图 12(a)、图 12(b)中,MDKautz 的 ABT 大于 MDCube,而图 12(c)、图 12(d)正好相反.这是因为单个模块的规模相同时,模块内的交换机数量决定了模块间互连的链路数,从而决定了网络容量的大小.在图 12(a)的参数配置下,MDKautz 中单个模块对外能够提供的交换机端口数均大于 MDCube,所以其相应的 ABT 也大于 MDCube,而图 12(b)中,MDKautz 的单个模块对外能够提供的交换机端口数均小于 MDCube,所以其相应的 ABT 也比 MDCube 小.同时,当 20%的链路失效率发生时,MDCube 的 ABT 性能平均下降 50%,而 MDKautz 只有 38%.因此,随着网络失效率的增加,MDKautz 的性能下降更为平缓.

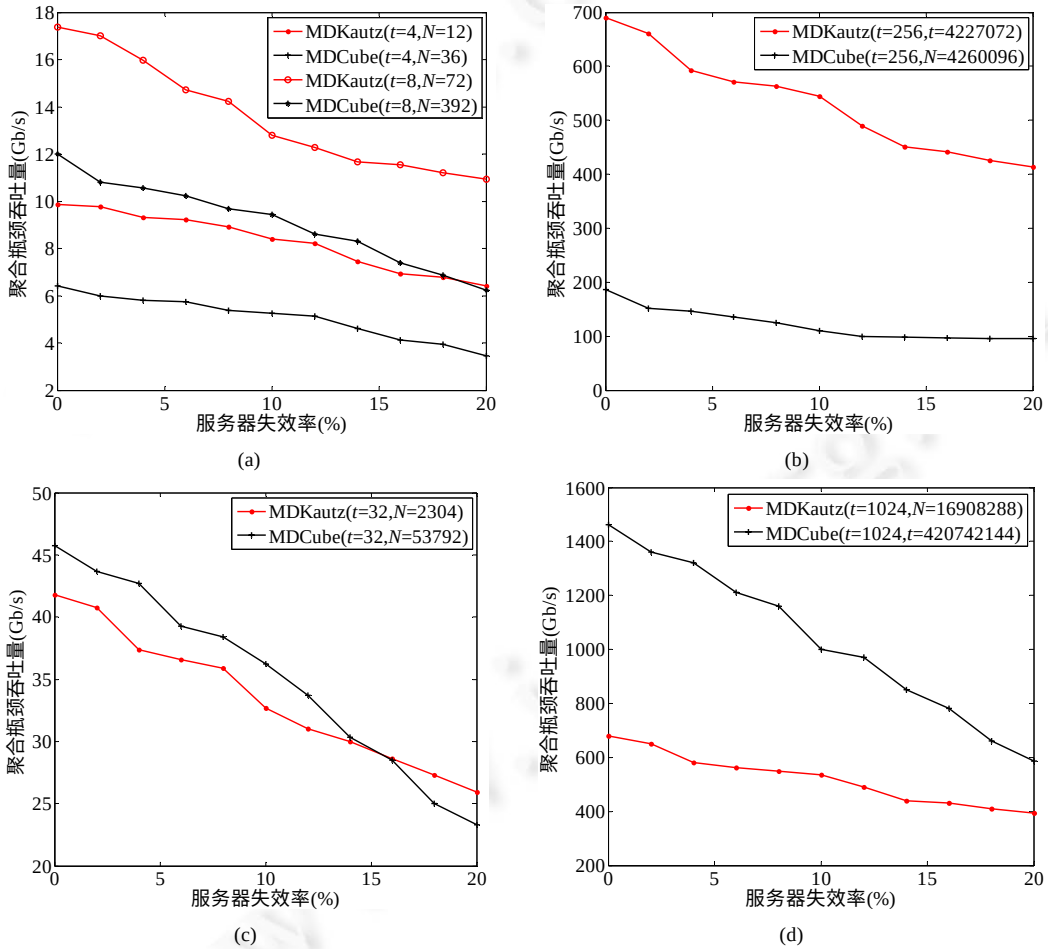

Fig.12 Comparison between MDKautz and MDCube in the ABT of two modules

图 12 MDKautz 与 MDCube 中 2 个模块间通信的 ABT 对比

6 结 论

本文针对超大规模MDC网络的互连问题提出了一种新型网络结构MDKautz.MDKautz利用模块内大量未被使用的交换机预留高速端口将模块以Kautz图互连,构造出具有高带宽、高容错和灵活可持续扩展性的超大规模数据中心网络.模块内外松耦合的互连设计和路由算法为模块之间提供了较高的聚合带宽;多条模块间不相交平行多路径确保网络具有良好的容错能力;且能够通过少量地调整现有网络结构,使网络的扩展不受限于模块内的可用高速端口数,从而进一步满足大规模数据中心对无损扩展和持续扩展的需求.MDKautz 能够广泛适用于连接各种模块内的互联结构,且均能达到较理想的网络性能.数学分析和模拟实验结果表明:MDKautz结构具有良好的拓扑特性和通信性能,能够以较小的直径构建出较大规模的网络,同时具有更大的二分带宽,对构建超大规模MDC网络是可行的.

References:

[1] Katz RH. Tech titans building boom. IEEE Spectrum, 2009,46(2):40−54. [doi: 10.1109/MSPEC.2009.4768855]

[2] Ghemawat S, Gobioff H, Leung S. The google file system. In: Proc. of the SOSP 2003. New York: ACM Press, 2003. 29−43. [doi: 10.1145/945445.945450]

[3] Shvachko K, Kuang H, Radia S, Chansler R. The hadoop distributed file system. In: Proc. of the MSST 2010. Washington: IEEE Computer Society, 2010. 1−10. [doi: 10.1109/MSST.2010.5496972]

[4] Weil SA, Brandt SA, Miller EL, Long DDE, Maltzahn C. Ceph: A scalable, high-performance distributed file system. In: Proc. of the OSDI 2006. Berkeley: USENIX Association, 2006. 307−320.

[5] Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters. In: Proc. of the OSDI 2004. Berkeley: USENIX Association, 2004. 27−39. [doi: 10.1145/1327452.1327492]

[6] Isard M, Budiu M, Yu Y, Birrell A, Fetterly D. Dryad: Distributed data-parallel programs from sequential building blocks. In: Proc. of the EuroSys 2007. New York: ACM Press, 2007. 59−72. [doi: 10.1145/1272996.1273005]

[7] Meisner D, Sadler CM, Barroso LA, Weber WD, Wenish TF. Power management of online data-intensive service. SIGARCH Computer Architecture News, 2011,39(3):319−330. [doi: 10.1145/2000064.2000103]

[8] Wu H, Lu G, Li D, Guo C, Zhang Y. MDCube: A high performance network structure for modular data center interconnection. In: Proc. of the CoNEXT 2009. New York: ACM Press, 2009. 25−36. [doi: 10.1145/1658939.1658943]

[9] Bhuyan LN, Agrawal DP. Design and performance of generalized interconnection networks. IEEE Trans. on Computers, 1983, c-32(12):1081−1090. [doi: 10.1109/TC.1983.1676168]

[10] Dally WJ, Towles B. Principles and Practices of Interconnection Networks. San Francisco: Morgan Kaufmann Publishers Inc., 2003.

[11] Al-Fares M, Loukissas A, Vahdat A. A scalable, commodity data center network architecture. In: Proc. of the SIGCOMM 2008. New York: ACM Press, 2008. 63−74. [doi: 10.1145/1402958.1402967]

[12] Massively Scalable Data Center (MSDC) Design and Implementation Guide. Cisco Systems, Inc., 2014.

[13] Niranjan MR, Pamboris A, Farrington N, Huang N, Miri P, Radhakrishnan S, Subramanya V, Vahdat A. PortLand: A scalable fault-tolerant layer 2 data center network fabric. SIGCOMM Computer Communication Review, 2009,39(4):39−50. [doi: 10.1145/1594977.1592575]

[14] Greenberg A, Hamilton JR, Jain N, Kandula S, Kim C, Lahiri P, Maltz DA, Patel P, Sengupta S. VL2: A scalable and flexible data center network. Communications of the ACM, 2011,54(3):95−104. [doi: 10.1145/1897852.1897877]

[15] Ramachandran K, Kokku R, Mahindra R, Rangarajan S. 60GHz Data-Center networking: Wirless⇒Worry less? Technical Report, Princeton: NEC Laboratories America, 2008.

[16] Wang G, Andersen DG, Kaminsky M, Papagiannaki K, Ng TSE, Kozuch M, Ryan M. c-Through: Part-Time optics in data centers. SIGCOMM Computer Communication Review, 2010,40(4):327−338. [doi: 10.1145/1851275.1851222]

[17] Farrington N, Porter G, Radhakrishnan S, Bazzaz HH, Subramanya V, Fainman Y, Papen G, Vahdat A. Helios: A hybrid electrical/optical switch architecture for modular data centers. In: Proc. of the SIGCOMM 2010. New York: ACM Press, 2010. 339−350. [doi: 10.1145/1851182.1851223]

[18] Guo C, Wu H, Tan K, Shi L, Zhang Y, Lu S. DCell: A scalable and fault-tolerant network structure for data centers. In: Proc. of the SIGCOMM 2008. New York: ACM Press, 2008. 75−86. [doi: 10.1145/1402958.1402968]

[19] Li D, Guo C, Wu H, Tan K, Zhang Y, Lu S, Wu J. Scalable and cost-effective interconnection of data-center servers using dual server ports. IEEE/ACM Trans. on Networking, 2011,19(1):102−114. [doi: 10.1109/TNET.2010.2053718]

[20] Guo C, Lu G, Li D, Wu H, Zhang X, Shi Y, Tian C, Zhang Y, Lu S. BCube: A high performance, server-centric network architecture for modular data centers. SIGCOMM Computer Communication Review, 2009,39(4):63−74. [doi: 10.1145/1594977.1592577]

[21] Huang F, Lu X, Li D, Zhang Y. A fault-tolerant network architecture for modular datacenter. Int'l Journal of Software Engineering and Its Applications, 2012,6(2):93−106.

[22] Bhuyan LN, Agrawal DP. Generalized hypercube and hyperbus structures for a computer network. IEEE Trans. on Computers, 1984,c-33(4):323−333. [doi: 10.1109/TC.1984.1676437]

[23] Li D, Xu M, Zhao H, Fu X. Building mega data center from heterogeneous containers. In: Proc. of the ICNP 2011. Washington: IEEE Computer Society, 2011. 256−265. [doi: 10.1109/ICNP.2011.6089059]

[24] Lu FF, Luo XG, Xie XH, Zhu GM, Pu XC. Researching on constant degree network for massively data center. Journal of Computer Research and Development, 2014,51(11):2437−2447 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2014.20130165]

[25] Zhang Y, Lu X, Li D. SKY: Efficient peer-to-peer networks based on distributed Kautz graphs. Science in China Series F: Information Sciences, 2009,52(4):588−601. [doi: 10.1007/s11432-009-0016-x]

附中文参考文献:

[24] 陆菲菲,罗兴国,谢向辉,朱桂明,濮小川.面向大规模数据中心的常量度数互连网络研究.计算机研究与发展,2014,51(11):2437−2447. [doi: 10.7544/issn1000-1239.2014.20130165]

陆菲菲(1981－),女,江苏无锡人,博士,工程师,CCF 专业会员,主要研究领域为分布式计算,数据中心网络.

郭得科(1980－),男,博士,教授,CCF 高级会员,主要研究领域为分布式系统,软件定义网络,数据中心网络.

谢向辉(1958－),男,博士,研究员,博士生导师,CCF 专业会员,主要研究领域为计算机系统结构,分布式计算.

朱桂明(1980－),男,博士,工程师,主要研究领域为分布式计算,数据中心网络.