针对可扩展交换网络的顽健性评价方法

杨光辉[1]，吴建平[1]，赵有健[1]，孙书韬[2]
(1. 清华大学 计算机科学与技术系，北京 100084；2. 中国传媒大学 计算机学院，北京 100024)

摘 要：分析了现有指标不能评价大规模可扩展交换网络顽健性的问题，结合可扩展交换网络拓扑特性和故障模型，提出了一种基于故障影响的顽健性评价方法，并进一步提出了该方法评价指标的优化算法。通过实验比较了故障影响方法与现有方法对于可扩展交换网络的评价效果，结果表明故障影响方法可以有效地评价大规模可扩展交换网络的顽健性。

关键词：可扩展路由器；可扩展交换网络；顽健性评价方法；故障影响；直连交换网络
中图分类号：TP393.1 **文献标识码**：A **文章编号**：1000-436X(2012)05-0001-11

Robustness measurement for scalable switch fabric

YANG Guang-hui[1], WU Jian-ping[1], ZHAO You-jian[1], SUN Shu-tao[2]
(1. Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China;
2. School of Computer Science, Communication University of China, Beijing 100024, China)

Abstract: Traditional robustness measurements are not suitable for large-scale scalable switch fabrics. Therefore, a novel robustness measure named failure influence was introduced, and improved algorithms for its metrics were also proposed. Experimental results showed that failure influence measurement can represented subtle robustness differences among various scalable switch fabrics. Failure Influence measurement was proved to be an appropriate robustness measurement for large-scale scalable switch fabrics.

Key words: scalable router; scalable switch fabric; robustness measurement; failure influence; direct network

1 引言

随着规模和用户持续高速增长，互联网已经演变成为一个全球性通用网络。互联网技术也因受到需求的驱动得到了长足的发展，各种应运而生的新型业务和协议已经在互联网中大规模的部署。互联网中的流量正以指数级别的速率增长。这些状况对互联网核心路由器提出了许多新的要求，包括提供更多更高速的端口设备以及更大的交换容量等。目前互联网的核心路由器一般采用集中式体系结构。这种体系结构虽然便于对需要转发的数据流量进行调度，但限制了路由器交换容量的扩展以及端口的数量增长[1]。为了解决这一问题，基于分布式体系结构的可扩展路由器被提出并得到越来越多的关注和研究[2]。

交换网络是路由器的关键部件。目前大多数商用路由器中，交换网络被设计为固定专用部件，扩展性较差。例如，共享总线交换结构中各个端口所共享的总线带宽是扩展的瓶颈。基于交叉开关（crossbar）的交换结构的扩展受到 crossbar 规模限

收稿日期：2010-10-14；修回日期：2011-04-10
基金项目：国家自然科学基金资助项目（60903184，60173167）；国家高技术研究发展计划(“863”计划) (2008AA01A324，2008AA01A323)
Foundation Items: The National Natural Science Foundation of China (60903184, 60173167); The National High Technology Development Program of China(863 Program) (2008AA01A324, 2008AA01A323)

制，扩展代价较大，接近于 $O(N^2)$级别（N 为交换结构的端口数量）。交换网络的扩展性直接决定了路由器的扩展性，具备较好规模扩展能力的交换网络称为可扩展交换网络。可扩展交换网络可以分为多级互连网络[3,4]（MIN, multistage interconnection network）、直连交换网络[5~7]（DN, direct switch network）、以及混合交换网络[8]。由于混合交换网络结构复杂，在路由器实际设计中少有采用，因此在本文中可扩展交换网络特指 MIN 和 DN 2 类。

无论 MIN 或 DN 都可抽象为连接图。连接图是将一定数量的节点（交换单元）通过边（链路）连接起来的抽象图。下文中将边和节点视为连接图的组件。MIN 抽象图和 DN 抽象图存在一定区别。MIN 抽象图中的节点可以分为 2 类：一类为输入/输出端口节点（简称端口节点），另一类为交换单元节点。外部数据流量通过输入端口节点注入路由器交换网络，再通过交换单元节点转发到输出端口节点。与 MIN 抽象图不同，DN 抽象图中的每个节点既是端口节点，又是交换单元节点。DN 节点可以接收目的是本节点的数据流量，也可将目的是其他节点的数据流量根据策略进行转发。随着组件数量的不断增大，整个交换网络的联合故障概率必然会随着规模扩展而不断变大。因此可扩展交换网络的设计必须考虑可能发生的组件故障对交换网络的影响。

故障对网络的影响一般采用基于顽健性的评价指标。对于一个网络，顽健性是当部分网络拓扑组件发生故障后，该网络保持基本功能的能力。根据基本功能的不同，顽健性评价方法可以分为 2 类[9]：一类方法的评价指标关注图的连通性，另一类的评价指标关注节点间距离的变化。可扩展交换网络的拓扑不仅可以决定一个交换网络的吞吐率、延迟等性能的上限，其顽健性也直接决定着可扩展路由器可靠性[10]。因此，合适的评价方法可以区别出不同可扩展交换网络顽健性差别，这对于可扩展交换网络的设计具有重要意义。

MIN 类可扩展交换网络的顽健性适宜采用基于连通性指标的评价方法；DN 类可扩展交换网络的顽健性则更适合采用基于距离指标的评价方法。目前研究中基于距离的评价指标多采用直径以及直径变化[11]等全局信息，这些指标应用于 DN 的顽健性评价存在以下问题：第一，直径类评价方法不能在少数组件随机发生故障的情况下对 DN 进行顽健性评价；第二，现有评价方法的算法复杂度较高。可扩展交换网络的顽健性评估需要根据评估的规模重复运行评估算法，此时评价方法的复杂度问题尤其突出。

本文内容安排如下。第 2 节中研究可扩展交换网络的拓扑特点以及随机故障模型，并建立模型比较 DN 在不同故障模型下的顽健性差别。提出一种新的基于故障影响（failure influence）的顽健性评价方法，故障影响包含的 2 种评价指标:故障影响范围（FIS, failure influence scope）和故障影响强度（FII, failure influence intensity）。第 3 节提出故障评价方法的优化算法。第 4 节分析和对比一些主流可扩展交换网络的故障影响属性。第 5 节是结束语。

2 针对可扩展直连网络故障影响的顽健性评价方法

失效、故障和错误的概念在容错计算领域中有着详细的区分。失效是指各种物理现象，如链路中的高斯噪声、导体中的电子迁移、供电系统异常等现象。故障是失效的表现形式[12]，如供电异常（失效）导致某寄存器无法保存数据，则这种行为定义为寄存器故障。这个故障致使计算机无法使用该寄存器导致计算结果偏离正常，最终表现为计算机错误。本文的研究主要针对组件层面（交换节点、链路等），以下统一采用故障进行表述。针对可扩展交换网络的顽健性评估方法必须考虑 2 个问题：首先，评价方法必须基于可扩展交换网络的应用场景；其次，评价方法必须可以区别不同可扩展交换网络顽健性的差别。

2.1 可扩展直连交换网络顽健性评价的需求分析

2.1.1 随机故障模型

故障模型对故障产生的类型、故障持续的长短和故障在系统中的分布以及系统中设备出现故障所服从的统计规律等进行具体的规定[13]。可扩展交换网络在运行过程中可能产生多种故障，例如硬件故障，软件运行时出现的漏洞导致的异常，流量导致的缓存溢出异常等。这些来源不同的故障从抽象图的角度可以归纳为节点故障、边故障、混合故障（边故障和节点故障同时存在）。根据故障持续时间可以划分为瞬态故障、永久故障、间歇型故障等。如果在故障产生后不能利用设备的冗余资源进行恢复，那么值守人员对该故障进行人工恢复的时间一般为小时级别甚至更长，这种故障定义为永久故障。

2 类最常使用的故障模型包括界限模型（bounded model）和概率模型（probabilistic model）[14]。界限模型设定了发生故障的节点/边的上界值，以最坏情况进行分析。概率模型中故障的发生概率是随机且相互独立的。本文针对的随机故障模型属于概率模型。在随机故障模型中假设基于严格正交拓扑的可扩展交换网络抽象图中的节点或边发生故障的概率是相同的，并且各个节点或边的故障概率相互独立。随机故障模型常用于可扩展交换网络的顽健性研究中[9]。系统的一次随机故障记为 $f(i)$，其中，i 为故障组件的数量。根据发生故障的组件不同，随机故障模型又可分为随机节点故障模型（记为 $f_v(i)$）、随机边故障模型（记为 $f_e(i)$）、随机混合故障模型（记为 $f_h(i)$）。随着微电子技术的进步以及路由器的广泛应用，路由设备专用芯片已经非常成熟，多个组件同一时刻发生故障的概率非常小。因此，小规模的组件同时发生故障的情况在可扩展交换网络随机故障模型中占主要地位。

2.1.2　DN 和 MIN 顽健性评价方法的区别

将 DN 和 MIN 抽象为连接图，二者最大的不同在于 DN 的交换单元和输入输出端口结合并采用严格正交拓扑连接，而 MIN 的输入输出端口则分离且其交换单元以分级方式连接。DN 和 MIN 在顽健性方面也有明显的差别。以 24 节点 Benes（MIN，图 1(a)）和 25 节点 2D-Torus (DN，图 1(b))为例。该 Benes 网络的交换单元节点分为 6 级，每一个输入输出端口对之间都存在多条路径。例如，如果有数据分组需要从 $C_0(6)$转发到 $C_6(7)$，图 1(a)存在 8 条路径可选。如果节点 $G_1(2)$发生故障（在图 1(a)中以矩形阴影表示），那么仍有其他长度为 6 的路径可选。对于 MIN，只要输入输出端口对是可达的，从源到目的节点的路径长度将不受影响。因此，基于连通指标的评价方法相比于基于路径长度指标的顽健性评价方法更适用于 MIN。通过观察易得，DN 的任意端口对间的路径数目远大于 MIN，这使得少量的故障很难对 DN 的连通性产生影响。但由于路径长度是不等的，因此基于路径长度指标的顽健性评价方法更适宜于 DN。

基于连通性指标的顽健性评价方法的研究成果较多，例如文献[8,15]中提出了适用于 MIN 的连通性评价指标。而基于路径长度指标的顽健性评价方法较少，下文将分析现有基于路径长度指标是否适用于可扩展交换网络的顽健性评价。网络的顽健性依赖于故障模型，不同故障模型会对网络的顽健性产生不同的影响。根据本文的调研，现有文献中较少结合随机故障模型对 DN 进行顽健性分析。下文将通过定义和证明来分析 DN 在随机故障模型下的顽健性。

定义 1　在图 $G(V,E)$中，V 为节点集合，E 为边的集合。图 $G(V,E)$的 K 子集定义为 G 中任意的 K 个节点，且 K 个节点之间有边连接，亦称为 G 的 K 子图。V_k 是 G 中非 K 子集节点且与 K 子集的节点有边连接的节点数目，亦称为 K 子集的邻居节点数。S_k 是 G 中 K 子集的总数。如果 G 中所有节点的度等于 d，且满足对任意 $1<K<N/2$ 存在 $V_K - d \geq 2$，那么图 G 称为一个 δ-规则图。

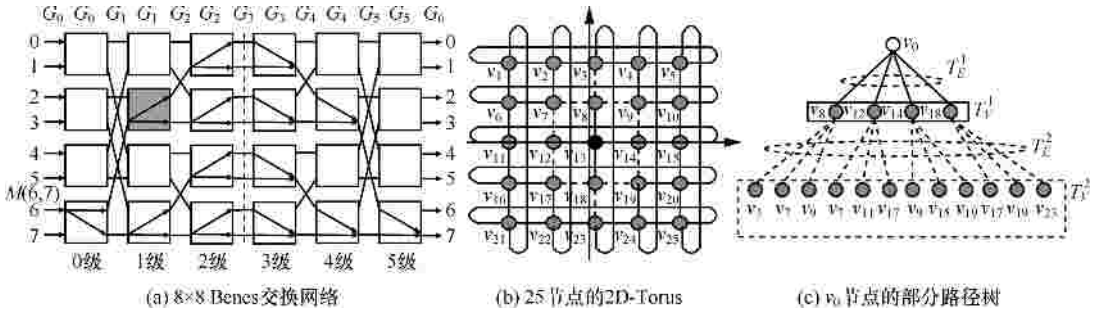
定义 2　图 G 为非连通图当且仅当图中存在子集被从 G 中隔离开。其中几种事件的概率如下。

$P(i)$：图 G 中发生 i 个组件故障后变成非连通图的概率。

$Q(i)$：在 i-1 个故障时是连通图，然后移除一个节点后，变为非连通图的概率。

$Q_k(i)$：图 G 在 i-1 个故障时是连通图，产生任意一个节点故障，在 i 个故障后从 G 中隔离出一个 K 子图概率。

上述几种概率满足以下 2 个关系：① $P(i) = Q(i)\prod_{j=1}^{i-1}(1-Q(j))$，这个公式的含义是在发生第 i 个



(a) 8×8 Benes交换网络　(b) 25节点的2D-Torus　(c) v_0 节点的部分路径树

图 1　以 Benes 网和 2D-Torus 网举例说明 MIN 和 DN

故障后变为非连通图的概率等于发生 1 到 i-1 个故障都不会变成非连通图的概率和第 i 个故障后正好变成非连通图概率的乘积。② $Q(i)=\sum_{K=1}^{Max}Q_K(i)$，这个公式的含义是在第 i 个故障后变为非连通图的概率等于在第 i 个故障后隔离出所有可能的 K 子图概率之和，其中，Max 代表在 i 故障下 K 子图所有可能值。

设 $C(G)=\dfrac{Q_1(d)}{\sum_{K=2}^{N}Q_K(V_K)}$，$C(G)$为隔离一个节点与隔离一个 K 子图的概率比值。$C(G)$的值越大，则 G 中单个节点被隔离的情况占的比率越大，多个节点因故障被隔离的概率越小。这也可以说明 G 的连通性越好。

分别记只有节点发生随机故障、只有边发生随机故障以及边和节点都发生随机故障的 $C(G)$为 $C_v(G)$、$C_e(G)$以及 $C_h(G)$。假设图 G 的节点数 N 远大于 δ，首先分析 $C_v(G)$。在只有节点发生随机故障时，$Q_1(d)=N/C_N^d$，并且 $Q_K(V_K)=S_K/C_N^{V_K}$。在 N 远大于 δ的假设下，$C_N^d=P_N^d/d!=(N\times(N-1)\llcorner(N-d+1))/d!\approx N^d/d!$。利用斯特林公式对上式中的阶乘进行化简。可以得出$d!\approx\sqrt{2\pi d}\,(d/e)^d$。从定义 2 分析可以得出 S_k 是 $O(N)$级别的数，而 V_k 是 $O(d)$ 级别的数，可以推出 S_k 远大于 V_k。当 K>1 时，可得：

$$Q_K(V_K)\approx Q_1(d)/(Ne)^{(V_K-d)} \tag{1}$$

由定义 2 可知，$V_K-d\geq 2$，且 N 远大于 δ，将式(1)代入 $C_v(G)$可以得到：

$$C_v(G)=\frac{Q_1(d))}{\sum_{K=2}^{N}Q_K(V_K)}\approx\frac{(Ne)^{(V_K-d)}}{N}\geq Ne^{(V_K-d)}\geq Ne^2>N \tag{2}$$

根据式(1)可知，在随机节点故障模型下 $C_v(G)$ 随着 N 增大而增大，G 的连通性也随之增强。因此，随着规模增大，连通性指标对于 δ-规则图的评价效果逐渐变差。下面的分析主要针对 $C_e(G)$和 $C_h(G)$，化简中用到的技巧与式(1)的推导过程类似，为了节省篇幅，下文中简要给出 $C_e(G)$和 $C_h(G)$的分析过程。

$$Q_K(V_K)\approx\frac{Q_1(d)S_KC_{Nd}^d}{NC_{Nd}^{V_K}} \tag{3}$$

$C_e(G)$推导中采用随机边故障模型，因此式(2)中加入组合数 C_{Nd}^d，表示从所有的 Nd 个边中选取 δ个边，代表单个节点被隔离的情况。K 子图被隔离的情况与式(1)中表述相同。应用$V_K-d\geq 2$ 和 N 远大于 δ，易得出 $C_{Nd}^{V_K}/C_{Nd}^d>C_N^{V_K}/C_N^d$。采用类似式(1)的化简过程，将式(2)代入 $C_e(G)$中，可以得出 $C_e(G)>C_v(G)>N$。

在混合故障模型下产生故障的组件既有节点也有边。式(4)表示 l 个边故障，其中，

$$Q_K(V_K)\approx\frac{Q_1(d)S_KC_{Nd}^tC_N^{d-t}}{NC_{Nd}^lC_N^{V_K-l}},0<l<V_K,0<t<d \tag{4}$$

通过组合数之间的比较很容易得出式(5)，在此不再做详细的证明。

$$\frac{C_N^{V_K}}{C_N^d}<\frac{C_{Nd}^lC_N^{V_K-l}}{C_{Nd}^tC_N^{d-t}}<\frac{C_{Nd}^{V_K}}{C_{Nd}^d} \tag{5}$$

考虑到混合随机故障模型的定义，其中，故障应包含所有边故障和节点故障的组合，则由式(4)和式(5)可得式(6)。

$$C_h(G)=\frac{N\dfrac{1}{d}\sum_{l=1}^{d-1}C_{Nd}^lC_N^{V_K-l}}{S_K\dfrac{1}{V_K}\sum_{i=2}^{V_K-1}C_{Nd}^tC_N^{d-t}},0<l<V_K,0<t<d \tag{6}$$

结合式(4)和式(5)，利用和推导式(2)相同的化简方法，可得 $C_v(G)<C_h(G)<C_e(G)$。根据式(2)、式(4)和式(6)可以得出：无论在随机边故障模型、随机节点故障模型还是随机混合故障模型，上述证明说明随机节点故障模型对顽健性的影响最大，随机混合故障模型次之，随机边故障模型对 δ-规则图的顽健性影响最小。这说明图的顽健性研究只需基于随机节点故障模型和随机边故障模型，这 2 种模型下的顽健性评价效果将是混合故障模型下评价效果的上下界。

2.1.3 现有评价尺度对于 DN 的不足

在节点规模相同情况下，MIN 任意输入输出端口对间的路径数量远小于 DN。采用已有文献[8,15]提出的评价方法，例如网络可靠性、终端可靠性和广播可靠性等基于连通指标的顽健性评价方法能够很好地对 MIN 类的网络进行评价。已有研究中针对 DN 的顽健性评价一般采用直径类指标。直径为图 G 中任意节点对之间距离的最大值。直径类指标的评价效果依赖于故障模型。在以一定故障数来产生最大破坏能力的界限模型下，直径类

指标能够较好显示出顽健性的变化[11]。但如果在随机故障模型下，直径类指标并不能对 d-规则图进行有效的顽健性评价。

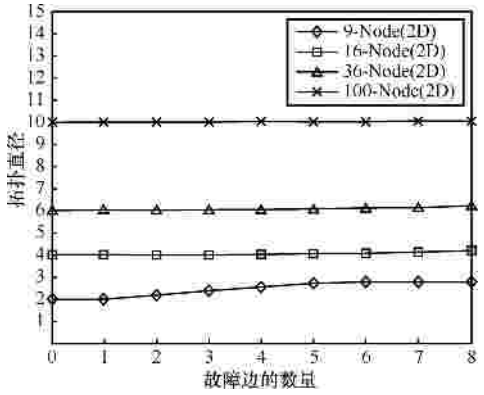
以 2 种典型的 DN 即 2D-Torus 和 3D-Torus 为例，在随机故障模型下验证直径增量指标的评价效果。由 2.1.2 节的证明可知混合故障模型的顽健性处于随机节点故障模型和随机边故障模型之间，因此只选择后 2 种故障模型进行研究。图 2(a)和图 2(b)分别说明了 2D-Torus 在随机边故障模型和随机节点模型下直径的变化。在节点规模达到 100 的时候，直径随着故障数量增加的变化非常小。这种趋势在 3D-Torus 上更加明显（如图 2(c)和图 2(d)所示）直径变化趋势已经成为水平直线。上述实验结果表明直径类指标无法对 DN 进行顽健性评价。因此必需在随机故障模型下设计一种有效的 DN 顽健性评价方法。

2.2　基于故障影响顽健性评价方法

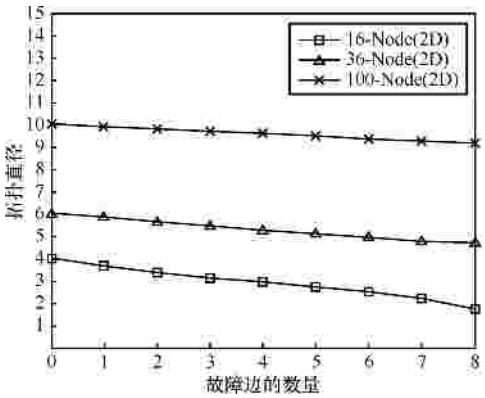
2.2.1　故障影响评价方法的设计动机

目前，研究和应用较多的 DN 包括 n 维全连接网络、链形网络、回环网络(torus)和超立方网络等。DN 一般采用严格正交拓扑，原因在于：首先，当网络规模持续增大时，严格正交拓扑易于保持固有的良好属性。其次，统一设备规格可以使制造、维护以及更新设备的成本大大降低。少量故障对 DN 造成大规模节点被隔离的概率非常低。更可能的情况是，少量组件故障只是某些节点对间的最短路径长度增加，这类似于一场地震。地震会对地面建筑或人员造成较大的破坏，距离震中越近的区域这种破坏效果越明显。地震可以用等震线在地图中表示，这种方法可以较为清晰地显示一场地震的破坏力及其影响范围。

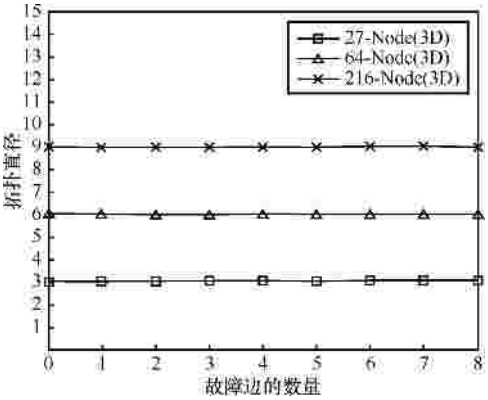
借鉴地震破坏力的评价方法，本文针对基于 DN 的可扩展交换网络提出了一种新的顽健性评价方法，命名为故障影响。这种方法较直径类的评价方法不同之处在于评价指标的选取。直径类指标关注于发生故障后图的某种全局特征的变化。但少量组件故障很可能难以导致全局特征的变化。故障影响评价方法将故障抽象视为震中，并评价该故障对于整个网络产生的破坏力。与震级和等震线的概念类似，故障影响包含 2 种不同评价指标：故障影响强度(FII, fault influence intensity)和故障影响范围(FIS, fault influence scope)。下文将对故障影响的评价指标进行描述。



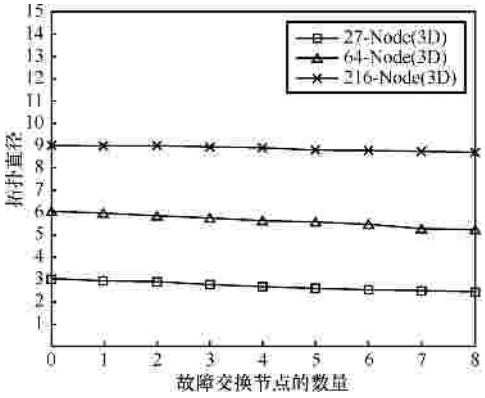
(a)在 $f_e(i)$下的 2D-Torus



(b)在 $f_v(i)$下的 2D-Torus



(c)在 $f_e(i)$下的 3D-Torus



(d)在 $f_v(i)$下的 3D-Torus

图 2　多种规模 Torus 网络在随机故障下的直径变化趋势

2.2.2 故障影响评价方法的基本定义

针对随机故障模型，在文中采用 $f_e(G,k)$代表图 G 中同时产生故障的 k 个边。同理，$f_v(G,k)$代表图 G 中同时产生故障的 k 个节点，$f_h(G,k)$代表图 G 中同时产生故障的 k 个节点或边，$f(G,k)$代表图 G 中同时产生的 k 个任意来源故障。在 δ-规则图 $G(V,E)$ 中，当发生一次 k 规模故障 $f(G,k)$后，G' 为从 G 移去故障节点和边后的图。

定义 3 节点的级和边的级。在 δ-规则图 $G(V,E)$中，任选一个节点 v_0 作为起始节点，v_0 的级定义为 0，记为 $N_l(v_0)=0$。假设 v_i (i?0, v_i∈V)距离 v_0 的最短路径长度为 k，则 v_i 的级为 k，记为 $N_l(v_k)=k$。图 $G(V,E)$中所有级为 k 的节点组成的集合记为 T_V^k。$G(V,E)$中的 v_a 和 v_b 之间如果存在一条边，那么这条边记为 e_{ab}，e_{ab} 的级记为 $E_l(e_{ab})$，$E_l(e_{ab})=\max(N_l(v_a), N_l(v_b))$。所有级为 l 的边组成的边的集合记为 T_E^i。图 1(c)以虚线框标识了 $T_E^1, T_E^2, T_V^1, T_V^2$。

定义 4 最短路径长度矩阵 (OM, one to all shortest path matrix)。在 d-规则图 $G(V,E)$中，节点 v_i 到 G 中其余所有节点 v_j ($i \neq j$, v_j∈V)的最短路径的长度值形成 N-1 个元素的一维向量，记为 $O_M(G, v_i)$。假设当 $G(V,E)$中产生故障 $f(G,k)$后拓扑变为图 $G'(V', E')$，对于 v_i(v_i∈V，v_i∈V)，如 $O_M(G, v_i)$?$O_M(G', v_i)$，则 v_i 定义为 G 中被故障 $f(G,k)$影响的节点。其中，$O_M(G,v_i)$和 $O_M(G', v_i)$只比较 v_i 到正常节点的距离变化。

定义 5 在 δ-规则图 $G(V，E)$中，当发生一次 k 规模故障 $f(G,k)$后，G 的拓扑变为 G'。G' 中被故障 $f(G,k)$影响的节点的最大数量称为 $f(G,k)$的故障影响范围，记为 $F_{\mathrm{FIS}}(f(G,k))$。

定义 6 G' 中被故障 $f(G,k)$影响的节点集合记为 $V_{f(G,k)}$。对于 $v_i \in V_{f(G,k)}$，$\max(O_M(G',v_i)-O_M(G,v_i))$ 为 2 个 N-1 个元素的一维矩阵差值中的最大值，记为 $M_{\max}(G',v_i)$。其中，$O_M(G, v_i)$和 $O_M(G', v_i)$只比较 v_i 到正常节点的距离变化。$F_{\mathrm{FII}}(f(G,k))$ 的值为 $\max(M_{\max}(G',v_i)) \mid \forall v_i \in V_{f(G,k)}$。

以图 3 为例，节点 15 在某一时刻产生了故障，即 $f_V(G,1)=\{15\}$。由于节点的规模较小，通过观察就可以得出在其余节点中，节点最短路径长度矩阵发生变化的节点是$\{v_4,v_7,v_8,v_9,v_{14}\}$。根据定义 5，$FIS(f_V(G,1))=5$。$\{v_4,v_7,v_8,v_9,v_{14}\}$这些受影响点的 $M_{\max}(G',v_i)$值则分别为$\{1,1,1,1,1\}$。因此 $F_{\mathrm{FII}}(f(G,k))=1$。

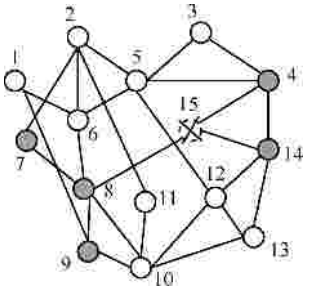


图 3 网络故障

2.2.3 评价方法的算法简化

图的顽健性评价算法一般具有很高的复杂度，其中，大部分评估算法都不具备多项式级别的评估算法。如 Minimum m-Degree[16]、Fragmentation[17]、Persistence[18]等评价估算法，在文献[9]中已经归纳并证明了多数顽健性评价尺度的评估算法是 NP-hard 难度。在实际的网络运营环境中，运营商会根据需要变更可扩展路由器的规模。规模扩展的灵活性要求顽健性评价方法应能评价顽健性随规模变化的趋势，这需要对可能应用到的规模都运行评价算法进行顽健性分析，此时算法复杂度的重要性更加凸显。因此，简化评价算法复杂性非常重要。

FIS 和 FII 都是基于距离的评价指标。如果基于经典最短距离算法 Dijsktra，节点的 OM 算法的复杂度应为 $O(N^4)$。考虑所有可能的节点规模，OM 算法的复杂度增加为 $O(N^5)$。FIS 和 FII 都是针对 δ-规则图设计的。文献[11]证明，在一个严格正交的网络中，故障对其邻居节点集合(记为 $V_N(f(G,k))$)的连通性产生最大的影响。这个结论可以应用于简化 FII，如式(7)所示。假设 $V_N(f(G,k))<<N$，FII 的评价算法的复杂度可以降低为 $O(N^4)$。

$$F_{\mathrm{FII}}(f(G,k)) = \max(M_{\max}(G',v_i)) \mid \forall v_i \in V_N(f(G,k)) \tag{7}$$

3 直连交换网络的故障影响分析

3.1 几种流行的直连网络

常用的流行直连交换网络包括：环型拓扑、星型拓扑、立方体型（cubes）、回环图型等。首先对结构相对简单的环型和回环型拓扑进行随机单故障模型的故障影响分析。从定义 1 可以得出，虽然环型拓扑是点对称的，但因为 $V_K - d < 0$，所以环型不属于 d-规则图。环型和回环型的故障影响分析由于其拓扑特征而具有相似性，环型拓扑或回环拓

表 1　环型拓扑和回环型拓扑的 FIS 和 FII 属性

拓扑类型	FIS		FII		度
	单边	单节点	单边	单节点	
环型	$2\times\lfloor(N-1)/2\rfloor$	$2\times\lfloor(N-3)/2\rfloor$	$\sum_{i=1}^{FIS\_ring} 2N-4i$	$Ni-\sum_{i=2}^{(FIS/2)+1} 2i$	2
2D-Torus	$2\times\lfloor(N^{1/2}-1)/2\rfloor$	$4\times\lfloor(N^{1/2}-3)/2\rfloor$	$N^{1/2}-2$	$N^{1/2}-4$	4
3D-Torus	$2\times\lfloor(N^{1/3}-1)/2\rfloor$	$6\times\lfloor(N^{1/3}-3)/2\rfloor$	$N^{1/3}-2$	$N^{1/3}-4$	6

扑中的环可以按照节点个数的奇偶性分为 2 类：偶环和奇环。通过观察和归纳易得出环的边故障影响范围。为节省篇幅，下文直接给环型和回环型的单边故障影响的研究结果，如表 1 所示。

文献[19]对严格正交拓扑进行分析后得出结论：在严格正交连接网络中，如果节点的度固定且内部链路带宽固定，则其吞吐率将随节点规模增大而最终下降。这个结论说明，如果要实现理论意义上的规模无限可扩展，节点度必然需要随着节点规模扩大而增长，否则内部带宽将成为吞吐率增长瓶颈；另一方面，节点的度不能过大，否则将导致扩展成本增大。为了平衡这对矛盾，本文前期工作在文献[1]中提出了渐进最小度扩展的可扩展交换网络 P2i。P2i 具有直径小、灵活扩展粒度、等分带宽较大等优点。

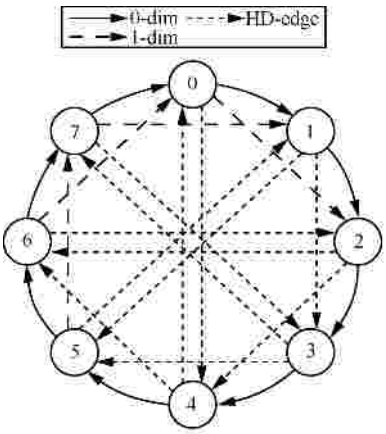
3.2　P2i 的故障影响分析

3.2.1　拓扑概述

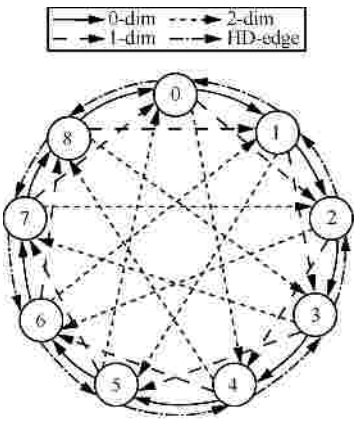
P2i 可以抽象为连接图 $G=(V,E)$。设 N 为节点的总数，所有节点从 0 至 $N-1$ 编号。编号 v 的节点拥有 $d=\lceil \mathrm{lb}N \rceil$ 条出边和 d 条入边。节点 v 的各条出边依次称为 0 维边至 $d-1$ 维边。i 维边（记为 i-dim）连接编号为 $(v+2^i) \bmod N$ 的节点。$d-1$ 维边亦记为 HD 边，其余出边记为 non-HD 边。图 4 为 8 节点和 9 节点的 P2i。

定义 7　跨度(span)：在一个 N 节点的 P2i 中，如果在 v_a 和 v_b 之间存在一条边 e_{ab}，则 e_{ab} 的跨度记为 $d_{\mathrm{span}}(e_{ab})$，且 $d_{\mathrm{span}}(e_{ab})=((b-a)+N) \bmod N$。$v_a$ 和 v_b 间如存在一条最短路径 P，则 v_a 和 v_b 的跨度记为 $d_{\mathrm{span}}(v_a,v_b)$，并且 $d_{span}(v_a,v_b)=\sum d_{span}(e_{xy}) \mid \ \forall e_{xy} \in P$。

定义 8　在 N 节点的 P2i 中，v_a 和 v_b 之间的最短路径如果只包含 non-HD 边，那么这种最短路径记为 NHSP。如果只包含 HD 边，那么这类最短路径记为 HDSP。如果最短路径中既有 HD 边又有 non-HD 边，则记为 HYSP。HYSP 的路径长度应大于等于 2。如果 v_a 和 v_b 间存在多于 2 条的最短路径，那么这些路径互相称为对称路径。



(a) 8节点P2i结构



(b) 9节点P2i结构

图 4　一个 8 节点 P2i 和一个 9 节点 P2i

3.2.2　P2i 故障影响的相关结论

结论 1　任取一个 N 节点的 P2i 中的节点 v_0，从 v_0 到其他节点的一条最短路径中，跨度为 $2^i\ (i<d-1)$ 的边最多只能包含一次。

结论 2　可以通过反证法来证明。当路径长度为 1 的时候，结论显然成立。跨度为 $2^i\ (i<d-1)$ 的边属于 non-HD。当路径长度大于 2 的时候，如果

跨度为 2^i $(i<d-1)$ 的边出现了 2 次，则必然能找到一条跨度为 $d_{\text{span}}((i+1)\text{-dim})=2$，$d_{\text{span}}(i\text{-dim})$的 $i+1$ 维边将代替两条 i 维边，形成一条新的最短路径 P'，并且 P'的路径长度小于 P，则 P 为最短路径的假设不成立。据此，结论 1 得证。

结论 3 在一个 N 节点 P2i 中，假设从 v_a 到 v_b 存在一条路径长度大于 1 的最短路径 P，并且 P 是 NHSP 或者 HYSP 中的任一种，则 P 必然存在一条对称路径 P'，并且 P'和 P 是边不相交的。

在一个 N 节点规模的 P2i 中的最短路径，如图 4(a)的 8 节点 P2i 中，从 v_0 到 v_3 的一条最短路径 $v_0 \to v_2 \to v_3$，这条路径可以用一个节点序列表示为(v_0, v_2, v_3)，也可以表示为一个边的序列$(e_{0,2}, e_{2,3})$或者一个跨度值的序列(1,0)。如果 P 为 NHSP 并且 P 的路径长度大于 1，则跨度值序列至少有 2 个位置的元素必然是不同的。在上述例子中跨度值的序列(1,0)，必然还存在着另外一种跨度值的序列(0,1)。路径长度越大，跨度值的排列越多，这意味着对称路径的数量也就越多。由于跨度值排列数不同，则节点序列不同，因此边的序列也是不相同的，即为边不相交的。结论 2 得证。

结论 4 在一个 N 节点 P2i 中，non-HD 的随机单边故障的故障影响范围为 1，P2i 的随机单边故障的故障影响范围等于 HD 边的单边随机故障的影响范围。

任取一个 N 节点 P2i 中的节点 v_0。在随机单故障模型下，如果故障发生在 v_0 的 1 级边，则 $O_M(v_0)$必然会改变。如果故障边为非 1 级边，则故障边可以分为 2 种：HD 边和 non-HD 边。如果故障边为 non-HD 边，根据结论 2，$O_M(v_0)$不会变化。则 non-HD 边在随机单故障模型下故障影响范围为 1。如果在故障边为 HD 边的情况下，由于 P2i 边的不对称性，HD 边的故障影响范围大于等于 1。因此，根据定义 6 可知 P2i 的随机单边故障的故障影响范围等于 HD 边的故障影响范围。

结论 5 不失一般性，任取一个 N 节点 P2i 中的节点 v_0，HD 边故障只可能会影响到 v_0 到其他节点的 HDSP 最短路径的长度。

v_0 到其他节点的最短路径中包含 HD 边的只可能是 HDSP 和 HYSP 二者之一。假设从 v_0 到 v_b 存在一条 HYSP 最短路径 P。显然 HYSP 的路径长度大于等于 2，根据结论 2，P 存在一条对称路径 P'，$O_M(v_0)$不会改变，因此结论 4 得证。

3.2.3 P2i 故障影响的优化算法

任取一个 N 节点 P2i 的节点 v_0，存在一个正整数 k，当 HDSP 的路径长度大于或等于 $k+1$ 时，v_0 通过这条 HDSP 路径连接了 v_k，则 v_0 到 v_k 必然存在一条长度小于等于 $k+1$ 的 NHSP，k 称为 P2i 的阈值。根据结论 3 可知阈值就是 P2i 在单随机故障模型下的 FIS 值。据此只需采用启发式算法求得 P2i 中的阈值即可得到该 P2i 的 FIS 值。上述启发式算法可以由以下的伪代码进行描述。其中，$SPF(v_s,v_d)$表示从 v_s 到 v_d 最短路径的长度，$v_{s,HD}$ 表示 v_s 第 $d-1$ 维边所连接的节点。k 是一个初始值为 0 的变量，FIS 的初值为 0。

P2i故障影响范围(FIS)优化算法

1) $G' = G - HD$
2) $k \leftarrow \text{SPF}_{G'}(v_s, v_{s,HD})$
3) if $FIS \geq k$ then $FIS \leftarrow 1$
4) *else*
5) while $FIS < k$
6) $FIS \leftarrow FIS + 1$
7) $v_s \leftarrow v_{s,HD}$
8) $k \leftarrow \text{SPF}_{G'}(v_s, v_{s,HD})$
9) end while
10) $FIS \leftarrow FIS - 1$
11) end if

4 实验结果及比较

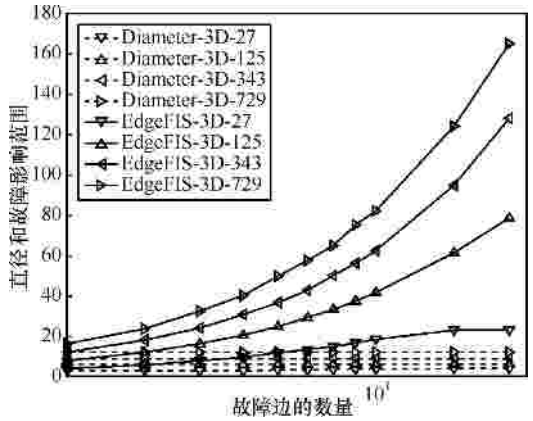
在第 2 节中提到，顽健性评价方法首先必须基于可扩展交换网络的应用场景；其次，该方法必须可以区别不同可扩展交换网络的顽健性差别。本节将通过 3 个实验来检验故障影响评价方法是否满足上述 2 个标准。根据 2.1.1 节的论述，本节所有实验都基于随机故障模型以符合可扩展交换网络的应用场景。下文首先通过敏感性实验来检验故障影响评价方法是否能在随机故障模型下对大规模可扩展交换网络进行顽健性评价。其次，用单故障实验和相近连接网络故障属性实验来检验故障影响评价方法能否区别不同可扩展交换网络的顽健性差别。

4.1 敏感性实验结果及分析

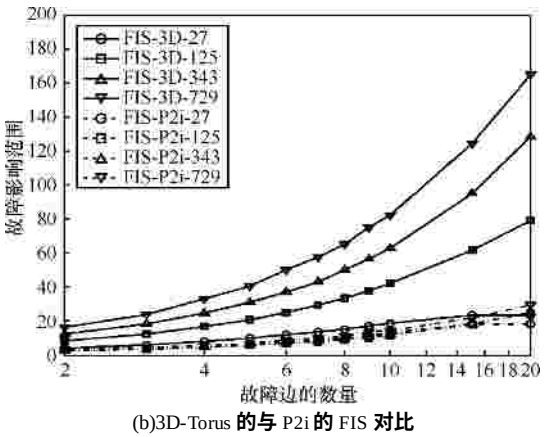
实验采用随机从所有组件中选取故障组件的方式。为了近似地实现随机性，实验中采用随机故障分布模式。例如，假设节点的随机故障概率为 1%，在 1 000 个节点规模时，应有 10 个节点发生故障。随

机从 1 000 个节点中选取 10 个节点，称为一次故障分布模式。第 4 节所有实验均随机选取 10 000 种故障分布。本节实验只选取 Edge-FIS 评价指标对不同种类 DN 进行分析，目的在于对比故障影响评价指标和直径类评价指标在少量组件故障下能否敏感反映顽健性的变化。

图 5(a)的实验分别选择了 27 节点（记为 3D-27 横坐标采用 log 坐标）、125 节点、343 节点、729 节点的 3D-Torus。在发生故障的组件逐渐增加的情况下对比直径的增量和 Edge-FIS 指标的变化。图 5(a)中各条虚线为直径增量评价结果。实验结果表明，直径类指标几乎毫无变化，在图 5(a)中显示为近似水平直线，显然无法反映网络顽健性的变化。相反地，从图 5(a)中可以看出，即使网络规模增大到 729，故障影响指标的变化依旧明显。这表明故障影响指标能够敏感的反映大规模可扩展交换网络顽健性的变化。



(a)直径与边的 FIS 对比



(b)3D-Torus 的与 P2i 的 FIS 对比

图 5　直径与故障影响方法的敏感性对比以及不同网络的故障影响对比

图 5(b)的实验（横坐标采用 log 坐标）采用与图 5(a)相同的故障模型，以 3D-Toru 和 P2i 为例，对边故障影响范围进行对比。实验选择了 27、125、343、729 节点规模。图 5(b)中各条虚线为 P2i 各个节点规模下的边故障影响属性。即使在发生故障的边从 2 逐渐增加到 20 后，P2i 的边故障影响属性依然远优于相应规模的 3D-Torus。尤其值得注意的是，各个规模 P2i 的边故障影响属性(图 5(b)中虚线)密集在一起。这说明即使节点规模扩大后，大规模 P2i 拥有与小规模 P2i 近似的边故障影响属性。从故障影响指标来看，P2i 非常适宜进行规模扩展。

4.2　单故障实验结果及分析

单故障是指在某一时刻只有一个组件发生故障。这种故障模型在实际应用中较为常见，如对某个节点的维护和升级等情况。由于 DN 采用严格正交拓扑，点对称特性使得 DN 的单故障模型等同于随机单故障模型。在单故障模型下，故障影响评价方法包括单边故障影响范围和强度、单节点故障影响范围和强度等评价指标。本节实验在 DN 中选择了 2D-Torus、3D-Torus 和 P2i 3 种典型的网络，节点规模的最大值设为 1 000。

实验结果如图 6 所示，它们分别给出了不同 DN 的单边故障影响范围、单节点故障影响范围、单边故障影响强度、单节点故障影响强度随节点规模的变化情况。由图 6(a)和图 6(b)可以看出，在单随机故障模型下，P2i 在大多数规模下拥有较 3D-Torus 更好的单边故障影响范围和单节点故障影响范围属性，且远优于 2D-Torus。图 6(b)说明，在随机单节点故障模型下，P2i 拥有非常良好的故障影响范围属性。这意味着如果 P2i 规模增加或减少一个节点，对整个交换网络的影响较小，这种性质非常利于交换网络进行平滑的规模扩展。3D-Torus 的故障影响强度属性远优于 2D-Torus，且在节点规模小于 500 的情况下，拥有较 P2i 更好的单边故障影响强度属性和单节点故障影响强度属性。其原因可能主要源于 P2i 并非边对称结构，某些维度的边故障在特殊规模下对故障影响强度属性的影响更大。

4.3　相近连接网络的故障属性实验及分析

P2i 和 Torus 连接和扩展方式不同。Torus 的节点度固定，而 P2i 的节点度随着规模增大而渐增。通过 4.1 节和 4.2 节的实验可知两者的故障影响属性区别明显。P2i 和 Hypercube 具有相似的拓扑。它们具有近似的直径和节点度并且节点度都会随着规模而变化。本节实验以这 2 种相近连接网络为

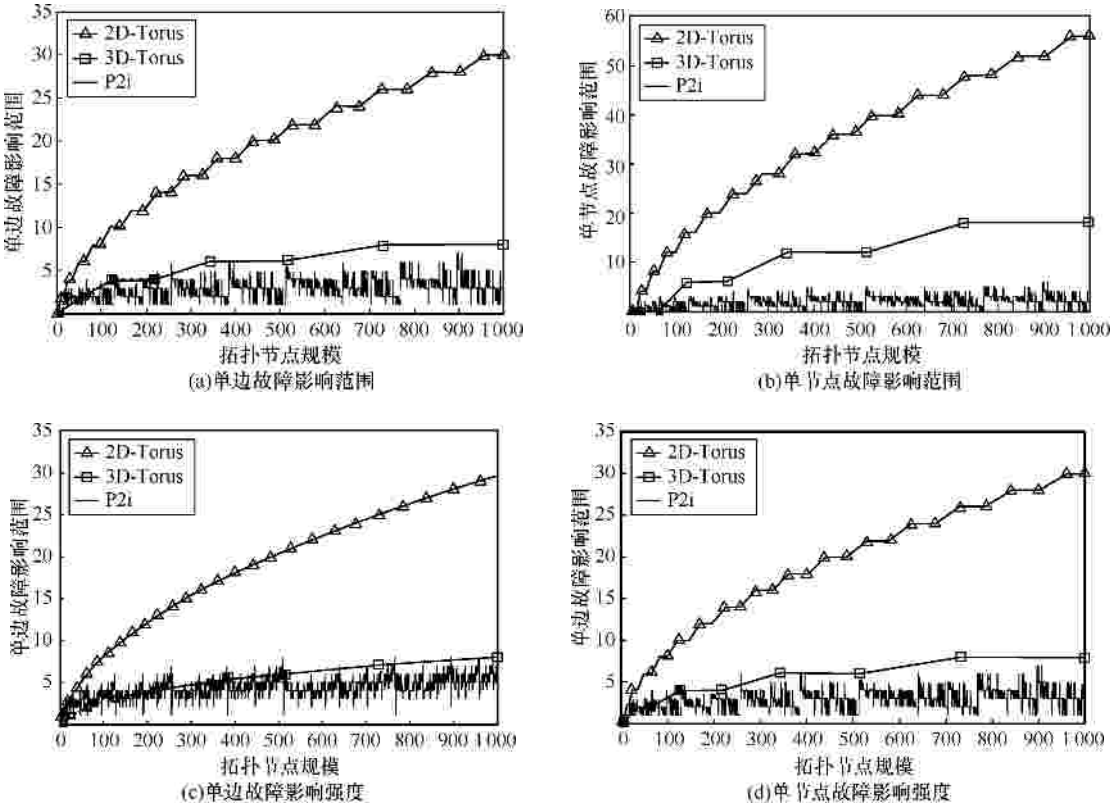


图 6 在单节点和单边故障下的不同网络的故障影响属性对比

表 2 P2i 和 Hypercube 的平均边故障影响（AFII）和最大的边故障影响（MFII）比较

边故障数量	H64-AFII	H64-MFII	P64-AFII	P64-MFII	H128-AFII	H128-MFII	P128-AFII	P128-MFII
5	2	2	1.18	1.674	2	2	1.14	1.58
10	2.001	2.004	1.212	2.063	2	2	1.16	1.91
15	2.003	2.05	1.26	2.366	2	2	1.17	2.02
20	2.006	2.134	1.31	2.67	2	2	1.19	2.25
30	2.038	2.818	1.508	3.45	2.01	2.04	1.24	2.63

表 3 P2i 和 Hypercube 的平均节点故障影响（AFII）和最大的节点故障影响（MFII）比较

节点故障数	H64-AFII	H64-MFII	P64-AFII	P64-MFII	H128-AFII	H128-MFII	P128-AFII	P128-MFII
2	0.536	0.536	0.523	0.523	0.36	0.36	0.36	0.36
4	1.638	1.638	1.01	1.232	1.61	1.62	0.98	1.045
6	1.938	1.938	1.15	1.65	1.88	1.88	1.07	1.28
8	1.99	1.996	1.21	2.025	1.98	1.98	1.09	1.58
10	2.01	2.04	1.27	2.362	2	2	1.13	1.865

例，在随机故障模型下，使用故障影响强度指标 FII 对这对相似交换网络进行顽健性评价及对比。

表 2 和表 3 是在随机边故障和随机节点故障模型下，64 节点和 128 节点规模的 P2i 和 Hypercube 的 FII 属性。FII_E 和 FII_V 分别表示发生故障的边和节点的个数。在随机故障模型下，FII 的最大值和平均值分别记为 MFII、AFII。例如，64 节点 Hypercube 的 MFII 和 AFII 分别记为 H64-MFII 和 H64-AFII。类似地，64 节点 P2i 的相关参数记为 P64-AFII 和 P64-AFII。从表 2 和表 3 中的数据可以得出，P2i 的 AFII 属性优于 Hypercube，而 MFII 属性较 Hypercube 差。这种差别可能缘于 P2i 边的不对称特性。P2i 各维度的边跨度不同，0 维边故障的影响强度高于其他维度的边。如果采用节点度和

拓扑直径的概念，则无法对 P2i 和 Hypercube 的顽健性进行区分。而表 2 和表 3 的实验结果则表明，故障影响指标可以有效评价 P2i 和 Hypercube 的顽健性的区别。

5　结束语

可扩展交换网络的顽健性评价是可扩展路由器研究中的一个重要问题。基于 DN 的严格正交网络具备良好的扩展性并被广泛应用到可扩展交换网络的设计中。在随机故障模型下，现有顽健性评价方法不能有效地对 DN 顽健性进行评价。本文总结了基于 DN 的可扩展交换网络的拓扑特征，将图论中的顽健性评价方法引入可扩展交换网络的顽健性评价方法中，提出了基于故障影响的顽健性评价方法。这种方法包括 2 种评价指标，即故障影响范围和故障影响强度。实验和分析表明，在随机故障模型下，基于故障影响的顽健性评价方法可以有效地对 DN 的顽健性进行评价，并可以对相似拓扑的顽健性进行区分。

参考文献：

[1] LIU Z, ZHANG X, ZHAO Y, *et al*. An asymptotically minimal node-degree topology for load-balanced architectures[A]. Proc of the IEEE GLOBECOM 2008[C]. New Orleans: IEEE, 2008. 1-6.

[2] TSE H. Switch fabric design for high performance IP routers: survey[J]. Journal of Systems Architecture, 2004, 51(10):571-601.

[3] KESLASSY I, CHUANG S, YU K, *et al*. Scaling internet routers using optics[A]. Proc of the 2003 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications[C]. Karlsruhe, Germany, ACM, 2003. 189-200.

[4] SUDAN R, MUKAI W. Introduction to the Cisco CRS-1 Carrier Routing System[R]. Cisco Systems California: Cisco Inc, Jan 1994.

[5] DALLY W. Performance analysis of *k*-ary *n*-cube interconnection networks[J]. IEEE Transactions on Computer, 1990, (39): 775-785.

[6] DALLY W. Scalable Switching Fabrics for Internet Routers[R]. Computer Systems Lab, Stanford University: Stanford University and Avici Inc, 1999.

[7] ZHAO Y, YUE Z, WU J. Research on next-generation scalable Routers implemented with H-Torus topology[J]. Journal of Computer Science and Technol, 2008, 23: 684-693.

[8] DUATO J, YALAMANCHILI S, Ni M. Interconnection Networks: An Engineering Approach[M]. San Francisco, USA: Morgan Kaufmann, 2003.

[9] BRANDES U, ERLEBACH T. Network Analysis: Methodological Foundations[M]. New York, USA: Springer-Verlag, 2005.

[10] DALLY W, TOWLES B. Principles and Practices of Interconnection Networks[M]. San Francisco, USA: Morgan Kaufmann, 2003.

[11] KRISHNAMOORTHY V, THULASIRAMAN K, SWAMY M N S. Incremental distance and diameter sequences of a graph: new measures of network performance[J]. IEEE Trans Computer, 1990, 39: 230-237.

[12] GARTNER F C. Fundamentals of fault-tolerant distributed computing in asynchronous environments[J]. ACM Computer Survey, 1999, 31: 1-26.

[13] KOREN I, KRISHNA M. Fault Tolerant Systems[M]. San Francisco, USA: Morgan Kaufmann, 2007.

[14] ZHANG Z P. Fault Tolerant Routing Algorithms of Regular Networks and Reliable Multicast[D]. Changsha: Central South University, 2005.

[15] CHENG X, IBE O C. Reliability of a class of multistage interconnection networks[J]. IEEE Trans Parallel Distribute System, 1992, 3: 241-246.

[16] BOESCH F, THOMAS R. On graphs of invulnerable communication nets[J]. IEEE Transactions on Circuits Theory, 1970, 17: 183-192.

[17] TANGMUNARUNKIT H, GOVINDAN R, JAMIN S, *et al*. Network topologies, power laws, and hierarchy[J]. ACM Sigcomm Computer Communication Review, 2002, 32: 76-76.

[18] BOESCH F T, HARARY F, KABELL J. Graphs as models of communication network vulnerability: connectivity and persistence[J]. Networks, 1981, (11): 57-63.

[19] ZHANG X P, LIU Z H, ZHAO Y J, *et al*. Scalable router[J]. Journal of Software, 2008, 19(6): 1452-1464.

作者简介：

杨光辉（1981-），男，河南长垣人，清华大学博士生，主要研究方向为可扩展路由器体系结构、交换网络和容错交换机制。

吴建平（1953-），男，山东巨野人，博士，清华大学教授、博士生导师，主要研究方向为计算机网络体系结构、计算机网络协议测试和形式化技术。

赵有健（1969-），男，甘肃会宁人，博士，清华大学教授、博士生导师，主要研究方向为高速路由器硬件体系结构、高速大容量交换结构、调度算法和混洗交换高速背板。

孙书韬（1967-），男，辽宁朝阳人，博士，中国传媒大学副教授，主要研究方向为计算机网络、数字媒体分析和调度算法。