文章编号 1004-924X(2009)12-3077-07

基于双向转发检测协议的光突发交换链路快速故障检测方案

付明磊[1],乐孜纯[2]
(1.浙江工业大学 信息工程学院,浙江 杭州 310014;2.浙江工业大学 理学院,浙江 杭州 310014)

摘要:由于光突发交换(OBS)网络通常采用单向资源预留机制,一般的链路故障检测方法难以在 OBS 网络中使用,因此本文基于双向转发检测(BFD)协议设计了符合 OBS 数据传输机制的链路故障检测方案,并根据实际网络需要设定了 BFD 报文的发送周期与检测时间,从而满足了 OBS 网络快速故障检测的要求。以 NSFNET 为仿真网络,实现了业务流量发生器与节点故障发生器,对 BFD 在单链路场景与网络场景下的链路故障检测性能进行了仿真。仿真结果显示,在单链路场景下,OBS 网络平均丢包率<0.001;在网络场景中,由于存在资源冲突、固定偏置时间等 OBS 网络自身的因素,OBS 网络平均丢包率接近 0.1。

关 键 词:双向转发检测;光突发交换;故障检测;网络生存性
中图分类号:TN929.1 **文献标识码**:A

Fast failure detection scheme for OBS links based on BFD

FU Ming-lei[1], LE Zi-chun[2]

(1. *College of Information Engineering, Zhejiang University of Technology, Hangzhou* 310014, *China*;
2. *College of Science, Zhejiang University of Technology, Hangzhou* 310014, *China*)

Abstract: Optical Burst Switching (OBS) network is a solution to realize all-optical internets. However, common link failure detection mechanisms are hardly used for OBS networks due to its special one-way resource reservation mechanism in the OBS. To meet the need of OBS data transmission, this paper designs a new fast failure detection scheme based on the Bidirectional Forwarding Detection (BFD) protocol. It sets both the transmitting period and detection time according to the practical requirements of OBS networks to achieve the fast failure detection. In the simulation part, the NSFNET is chosen as the simulation network to impletment both the OBS traffic generator and node failure generator. Moreover, the performance of link failure detection for BFD is simulated in the single path scenario and network scenario. Simulation results show that the average packet drop rate is lower than 0.001 in the single path scenario, While it is nearly 0.1 in the network scenario for the restrictions of OBS networks such as resource contention and fixed offset time.

Key words: Bidirectional Forwarding Detection (BFD); Optical Burst Switching (OBS); failure detection; network survivability

收稿日期:2008-11-07;**修订日期**:2009-02-16.
基金项目:浙江省自然科学基金资助项目(No. Y1080172)

1 引　言

随着DWDM技术的广泛应用，由各种基于IP(Internet Protocol)的新业务所引起的带宽瓶颈问题已经由传输链路转移到了交换节点，并由此引发了人们对各类交换技术的研究。与传统的电交换技术相比，光交换技术在器件体积、功耗以及交换效率等方面具有较明显的优势，其中光线路交换(Optical Circuit Switching，OCS)、光突发交换和光分组交换(Optical Packet Switching，OPS)被认为是3种最主要的交换技术。

OCS类似于电路交换，具有技术成熟、容易实施等优点，但是它的交换粒度大、无法实现统计复用，因此带宽利用率低。OPS是较理想的光交换技术，它具有交换粒度小、带宽利用高等优势，但是现有的光学技术无法提供其必需的高速随机存储缓存单元(Random Access Memory，RAM)，也难以实现光同步信号提取等关键技术，因此OPS的实施难度很大。OBS介于OCS与OPS之间，是一种依靠现有光学与电学技术能够实现的过渡性的光交换技术[1-6]。

在OBS网络中，控制信道与数据信道在空间上是分离的。在数据信道中，数据包以突发数据分组(Burst Data Packet，BDP)的形式传输。而对于每一个BDP，在控制信道中都对应一个突发控制分组(Burst Control Packet，BCP)，并且BCP比BDP提前一个偏置时间(Offset Time)传输。这样的数据传输方式使得BCP能够提前为BDP配置必需的路由交换器件(如Optical Cross Connect，OXC)，从而避免了数据传输过程的O/E/O转换。因此，与一般具有存储-转发功能的网络不同，OBS网络中通常采用的是一种单向资源预留协议，如常见的JIT(Just In Time)协议和JET(Just Enough Time)协议。但是这类协议一般不具备应答功能，这使得一般的故障检测机制，如传统的Hello机制，难以在OBS网络中使用[7-8]。而对于数据传输速率达到G比特的OBS网络，较长的故障检测和恢复时间意味着大量数据的丢失，因此选择一种有效而快速的故障检测机制对于维护OBS网络生存性是非常必要的。

本文将双向转发检测(BFD)协议应用于OBS网络链路的故障检测中。BFD的运行不依赖于回声报文(ECHO)，这使得它非常适合OBS网络单向资源预留的特点，并且相对于Hello机制的检测时间(例如：OSPF(Open Shortest Path First)需要2 s的检测时间，IS-IS(Intermediate System to Intermediate System)需要1 s的检测时间[9]，BFD的检测时间<30 ms，并且可以根据实际网络需要设定发送周期与检测时间，以满足OBS网络链路快速检测的要求，从而有效地提高了OBS网络生存性。

2 BFD的报文格式与发送、检测时间

双向转发检测(BFD)是一套用来实现快速检测的国际标准协议，提供一种轻负荷、持续时间短的检测，而且它适合所有的媒体类型、封装、拓扑结构和路由协议。BFD不但可以检测和判断传输链路、光接口和设备端口的中断故障，还可以检测和判断传输层、链路层、IP层乃至应用层存在的误码、丢包等软故障。而且BFD技术不依赖于其他协议或者应用，可以运行在任何层面，采用硬件实现，不影响设备性能[1,9-11]。

2.1 BFD的控制报文格式

BFD发送的检测报文是UDP(User Datagram Protocol)报文，它的控制报文格式如图1所示。

Vers	Diag	Sta	P	F	C	A	D	R	Detect mult	Length
My discriminator										
Your discriminator										
Desired min TX interval										
Required min RX interval										
Required min echo RX interval										

图1　BFD的控制报文格式

Fig. 1　Format of BFD message

其中VERS表示BFD协议版本号，目前为1；DIAG表示诊断编码，说明本地BFD系统最后

一次会话状态变化的原因；STA 表示 BFD 本地状态；P 表示参数发生改变时，发送方在 BFD 报文中置 1，接收方必须立即响应该报文；F 表示响应 P 标志置位的回应报文中必须将 F 标志置 1；C 表示转发/控制分离标志，一旦置位，控制平面的变化不影响 BFD 检测；A 表示认证标识，置位代表会话需要进行验证；D 表示查询请求，置位代表发送方期望采用查询模式对链路进行监测；R 是预留位；Detect MULT(DM)表示检测超时倍数，用于计算检测超时时间；Length 表示报文长度，以字节计；My Discriminator 表示 BFD 会话连接本地标示符；Your Discriminator 表示 BFD 会话连接远端标示符；Desired Min TX Interval(DMTI)表示本地支持的最小 BFD 报文发送间隔；Required Min RX Interval(RMRI)表示本地支持的最小 BFD 报文接收间隔；Required Min Echo RX Interval 表示本地支持的最小 Echo 报文接收间隔。

2.2　发送周期与检测时间的计算

BFD 协议描述了实现双向检测的机制，可分为两种模式：异步模式和查询模式，另外还有一种辅助功能回声功能。但是考虑到 OBS 网络的资源预留特点，本文选择 BFD 协议工作在异步模式。在异步模式下，OBS 网络的核心路由器之间相互周期性地发送 BFD 控制报文，如果某个核心路由器在检测时间内没有收到对端发来的 BFD 控制报文，就宣布两者之间的连接失效。发送周期与检测时间的计算采用下面的方法。

对于发送周期，由于要考虑到链路的抖动，需要一个允许的范围。如果 Detect MULT 为 1，那么发送周期选择本地端的 DMTI 与接收到的来自对端的 RMRI 两者间的最大值，波动范围为 70％至 90％。否则，发送周期选择为本地端的 DMTI 与接收到的来自相邻节点的 RMRI 两者间的最大值，波动范围为 90％至 100％。

对于检测时间，由于检测的位置是在对端，所以对端在计算检测时间时需要用到本端的检测倍数，即检测时间为本地端的 DMTI 与接收到的来自对端的 RMRI 两者间的最大值，其检测倍数为接收远端的 DM。

3　BFD 在 OBS 链路故障检测中的应用

本文以一个简单的 6 个节点的 OBS 网络为例(如图 2 所示)，简要介绍 BFD 在 OBS 网络中的检测过程。假设设定每个节点的检测倍数 DM＝3，最初的 DMTI＝10 ms，RMRI ＝10 ms，由检测时间公式，得到检测时间＝30 ms。OBS 核心路由器 A、B、C、D、E 和 F 的 My Discriminator 分别设为 1、2、3、4、5 和 6，并且假设核心路由器 E 出现故障。

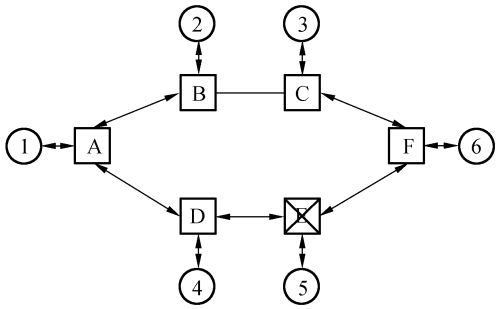

图 2　6 节点 OBS 网络拓扑图

Fig. 2　Topology of six-node OBS network

3.1　OBS 链路连通的检测过程

首先，以核心路由器 A 与 D 之间的 BFD 检测为例，说明 BFD 检测连通的过程。

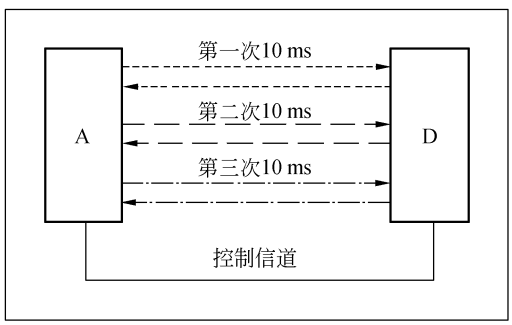

图 3　核心路由器 A 与 D 之间 BFD 检测

Fig. 3　BFD detection between core router A and D

如图 3 所示，A 第一次发送，BFD 报文格式为：VERS：1；DIAG：0；STA：0 ；D：0 ；P：0 ；F：0；DM：3；Length：24；My Discriminator：1；Your

Discriminator: 0; DMTI: 50 (ms); RMRI: 50 (ms); Required Min Echo RX Interval: 0。

D 第一次发送,BFD 报文格式为: VERS: 1; DIAG: 0; STA: 0 ; D: 0 ; P: 0 ; F: 0; DM: 3; Length: 24; My Discriminator: 4; Your Discriminator: 0; DMTI: 50 (ms); RMRI: 50 (ms); Required Min Echo RX Interval: 0。

A 第二次发送,BFD 报文格式为: VERS: 1; DIAG: 0; STA: 1 ; D: 0 ; P: 0 ; F: 0; DM: 3; Length: 24; My Discriminator: 1; Your Discriminator: 4; DMTI: 50 (ms); RMRI: 50 (ms); Required Min Echo RX Interval: 0。

D 第二次发送,BFD 报文格式为: VERS: 1; DIAG: 0; STA: 1 ; D: 0 ; P: 0 ; F: 0; DM: 3; Length: 24; My Discriminator: 4; Your Discriminator: 1; DMTI: 50 (ms); RMRI: 50 (ms); Required Min Echo RX Interval: 0。

核心路由器 A 与 D 连通成功。如果设置参数没有变化,那么两个核心路由器将每隔一定的时间发送第二次发送的控制报文。

3.2 OBS 链路故障的检测过程

接下来,以核心路由器 D 与 E 之间的 BFD 检测为例,说明 BFD 检测不连通的过程,如图 5 所示。核心路由器 D 每隔 10 ms 发送一次 BFD 控制报文,发送 3 次后,在检测时间 30ms 内仍然没有收到 E 发来的 BFD 控制报文,这说明核心路由器 E 出现故障,因此在进行路由选择的时候就不能选择 DE 链路。D 在接下来的控制报文中会把 DIAG 位置 1,表明检测超时,具体的 BFD 报文格式如下。

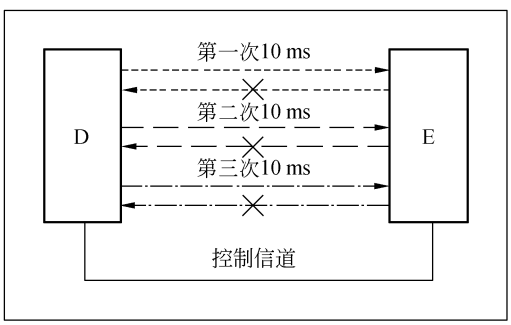

图 4 核心路由器 D 与 E 之间 BFD 检测

Fig. 4 BFD detection between core router D and E

D 连续发送了 3 次同样的控制报文,BFD 报文格式为: VERS: 1; DIAG: 0; STA: 0 ; D: 0 ; P: 0 ; F: 0; DM: 3; Length: 24; My Discriminator: 4; Your Discriminator: 0; DMTI: 50 (ms); RMRI: 50 (ms); Required Min Echo RX Interval: 0。

4 BFD 快速故障检测方案的仿真过程

4.1 OBS 网络拓扑与业务流量模型

本文选择一张典型的 NSFNET (National Science Foundation Network)网络作为仿真网络的拓扑图,这是一个具有 14 个节点、21 条链路的 Mesh 网络拓扑,如图 5 所示。

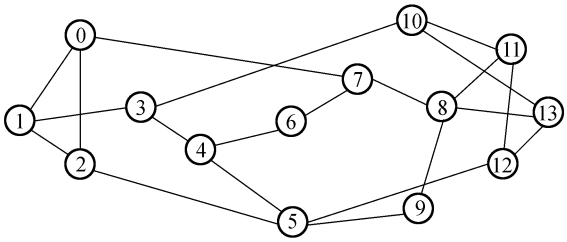

图 5 NSFNET 网络拓扑

Fig. 5 Topology of NSFNET

每个网络节点都装备一个业务流量发生器,如图 6 所示,BFD 信道负责传送节点之间的 BFD 控制报文,用来检测节点之间的连通性;BCP 信道用来传送 BCP 数据包,它负责为 BDP 进行路由选择、波长预留等;BDP 信道主要传送突发数据包。BFD 信道和 BCP 信道传送的数据需要在电学器件中进行处理,而 BDP 信道传送的数据全部在光学器件进行处理。

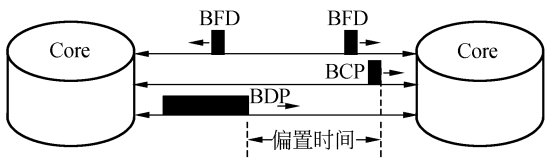

图 6 节点业务流量发生器模型

Fig. 6 Model of node traffic generator

NSFNET 的 14 个节点与其相邻节点之间相互周期性地发送 BFD 控制报文,从而不断地更新网络的连接情况。同时业务流量发生器以 ON/OFF 开关函数产生 BDP 及其对应的 BCP。ON

与 OFF 的状态时间（State Time）均服从 Pareto 分布。

BCP 和 BDP 的报文格式，到目前为止，国际上没有给出严格的、统一的定义。在本文，BCP 负责携带目的地址、BDP 的编号和字节长度信息；BDP 除了携带净荷（Payload）信息外，负责携带目的地址、编号、BDP 字节长度信息。BCP 采用常用的 Dijkstra 算法进行路由选择，并采用 First-Fit 算法进行波长选择和预留。为了简化波长预留计算，本文采用固定偏置时间，并且要求偏置时间大于 BFD 检测时间。

4.2　节点故障发生器

为了仿真节点故障，本文在路由的中间节点设置故障发生器，以检验 BFD 的故障检测性能。故障发生器能够通过调整随机数发生器的参数，选择故障节点的个数及故障发生概率。表 1 表示源节点为 1，目的节点为 12 的路由，在不同轮次下的中间节点故障发生情况。如表 1 所示，“√”表示节点正常，“×”表示节点故障，而且故障节点会在一次路由选择及数据传输过程结束后恢复。

表 1　节点故障发生器示例

Tab. 1　Example of node failure generator

Round	1	3	10	13	12
1	√	√	√	√	√
2	√	×	√	√	√
3	√	√	×	×	√
4	√	√	√	√	√
5	√	×	√	×	√

节点故障发生器适合单链路仿真的场景。但是在进行网络仿真过程中，本文发现节点故障发生器能够引起网络丢包率急剧增加。原因在于：节点故障发生器产生的故障节点与其他路由的源节点或者目的节点重合，导致业务流量不能正常的输入、输出。因此在进行网络仿真过程中，故障节点的产生范围不再选择路由的中间节点，而选择在非源节点与非目的节点的集合中随机产生故障节点，产生机理与节点故障发生器一致。

5　仿真结果

本文分两种场景仿真 BFD 协议在 OBS 网络中的快速故障检测性能。第一种场景选择单链路仿真，即随机选择源节点与目的节点，并给中间节点设置故障发生器，通过统计发送端与接收端的业务流量来评估 BFD 协议的性能；第二种情况为网络仿真，模拟真实 OBS 网络运行情况，节点之间同时进行通信，通过统计所有节点的发送与接收业务流量来评估 BFD 的性能。

5.1　单链路仿真

本文首先仿真 BFD 协议在单链路场景的性能。为了更清楚的反应源-目的节点间的业务流量变化，本文延长了链路传输时延。如图 7 所示，由于节点故障发生器使得中间节点产生故障节点，在目的节点输入的业务流量并不能与源节点输出的业务流量具有相同的流量曲线，但是总体流量接近。

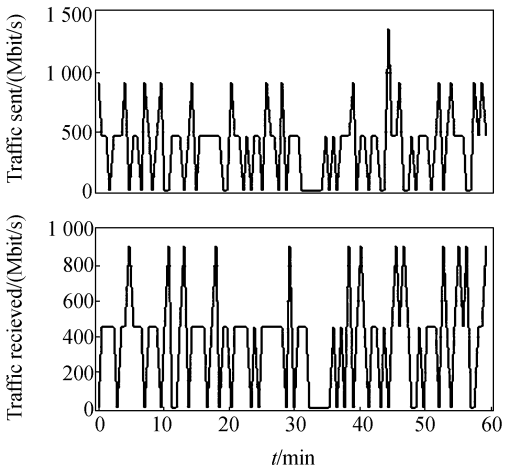

图 7　单链路实时仿真结果

Fig. 7　Simulation result of single path scenario

图 8 是单链路源-目的节点业务流量的统计结果。平均输出、输入曲线的比较结果显示：在单链路场景下，网络的丢包率很低，平均丢包率＜0.001。输出与输入曲线基本能够拟合。这表明：由于 BFD 能够快速检测故障节点，使得 BCP 能够选择其他路由传递 BDP，从而有效降低网络的丢包率。

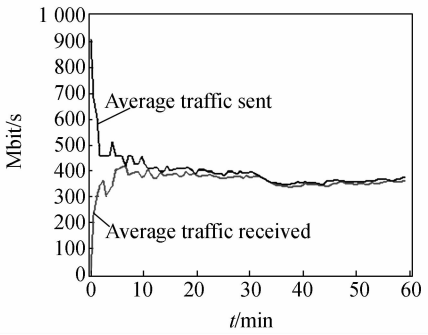

图 8　单链路源-目的节点业务流量比较

Fig. 8　Compare of source and destination traffic for single path scenario

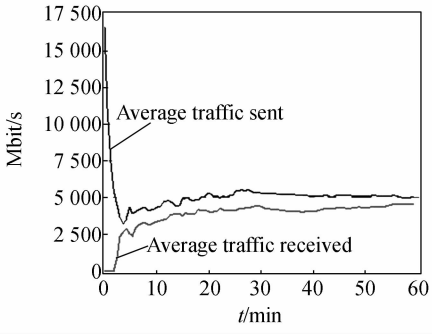

图 10　网络各节点输出、输入业务流量比较

Fig. 10　Compare of source and destination traffic in network scenario

5.2　网络仿真

与单链路场景相比，网络仿真场景仿真复杂度增加。一方面体现在源-目的节点对数目增加，从而导致数据统计量加大；另一方面，节点故障发生器与单链路场景相比也要改变。图 9 是一组网络实时仿真结果，节点输出与输入的流量变化曲线显示：虽然曲线变化相似，但是总体流量存在差异。

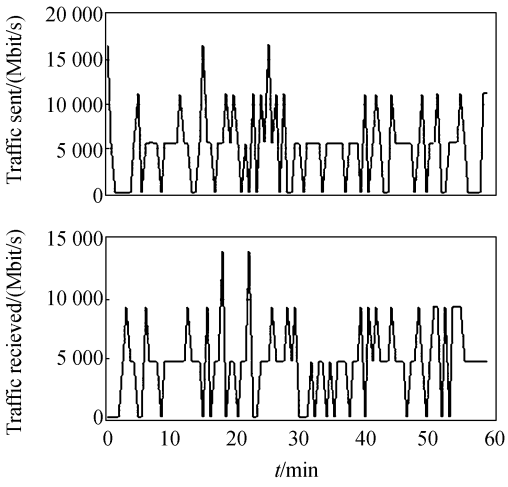

图 9　网络实时仿真结果

Fig. 9　Simulation result of network scenario

图 10 更加清楚地表示了网络各节点输出、输入业务流量的差异。平均输入曲线虽然与平均输出曲线具有相似的变化率，但是依然存在较明显的丢包现象，平均丢包率接近 0.1。

经分析，丢包现象的原因主要有两个方面。首先，本文在仿真场景设置时采用了 BCP 与 BDP 的固定偏置时间，以简化业务流量发生器部分，但是在中间节点发生故障导致资源冲突时，BCP 无法为 BDP 预留波长，引起丢包；第二个方面是由于节点故障发生器使得某些路由无法被选择，如 NSFNET 中的节点 9，如果故障发生器使得节点 5 和节点 8 同时发生故障，那么无论节点 9 是源节点或者目的节点都无法发送或者接收数据，导致丢包。

以上分析的原因，从根本上说属于 OBS 网络自身问题，可以从改善网络调度性能、增加 FDL（光纤延迟线）等方法解决。

6　结　论

光突发交换（OBS）网络的生存性问题对于 OBS 的实用化具有重要的影响。本文尝试将双向转发检测（BFD）协议应用在 OBS 链路的故障检测中。BFD 自身的传输特点非常适合 OBS 的单向资源预留协议，并且它的快速检测能力能够有效降低 OBS 网络的丢包率。在一个 6 点 OBS 网络为例介绍了 BFD 的应用过程后，本文以 NSFNET 为仿真网络，对 BFD 在单链路场景与网络场景下的性能进行了仿真。仿真结果显示：在单链路场景下，BFD 能够快速发现故障节点，从而有效降低丢包率，平均丢包率<0.001；在网络场景中，由于存在资源冲突、固定偏置时间等 OBS 网络自身的因素，使得网络存在一定的丢包率，平均丢包率接近 0.1，但是在总体上，网络仿真的结果也能够反映 BFD 的快速故障检测能力。因此，BFD 是一种适合 OBS 网络的快速故障检测协议。

在 OBS 网络实际应用的情况下，由于 BFD 控制报文开销小，可以考虑在 OBS 控制信道中传输 BFD 控制报文，从而节省宝贵的 OBS 网络资源。另外，BFD 故障检测协议的使用必须要配合路由层面的链路恢复技术，如快速重路由（Fast Re-Route）技术，从而有效保护 OBS 网络中的数据传输。

参考文献：

[1] 韦乐平，张成良. 光网络一系统、器件与联网技术[M]. 北京：人民邮电出版社，2006：389-449.
WEI L P, ZHANG CH L. *Optical Networks-System, Devices and Networking Technology* [M]. Beijing: Posts & Telecom Press, 2006:389-449.

[2] 杨俊波，苏显渝. (3,3,2)矩形 CC 榕树网光学实现方法[J]. 光学 精密工程，2007,15(8):1220-1228.
YANG J B, SU X Y. Optical implementation of (3, 3, 2) rectangular CC-Banyan network [J]. *Opt. Precision Eng.*, 2007,15(8):1220-1228. (in Chinese)

[3] 查英，孙德贵，刘铁根，等. 扩展 BANYAN 网络的可重构无阻塞 8×8 矩阵光开关[J]. 光学 精密工程，2007,15(1):51-56.
ZHA Y, SUN D G, LIU T G, *et al*.. Rearrangeable nonblocking 8×8 optical matrix switch with extended BANYAN network [J]. *Opt. Precision Eng.*, 2007,15(1):51-56. (In Chinese)

[4] QIAO C, YOO M. Optical burst switching - a new paradigm for an optical internet [J]. *Journal of High Speed Networks*, 1999,8(1):69-84.

[5] CHEN Y H, TUMER J S, MO P F, Optimal burst scheduling in optical burst switched networks [J]. *Journal of Lightwave Technology*, 2007,25(8):1883-1894.

[6] ROSBERG Z, ZALESKY A, VU H L, *et al*.. Analysis of OBS networks with limited wavelength conversion [J]. *IEEE/ACM Transactions on Networking*, 2006,14(5):1118-1127.

[7] TACCA M, WU K, FUMAGALLI A. Local detection and recovery from multi-failure patterns in MPLS-TE networks [C]. *Proceedings of IEEE International Conference on Communications*, *Istanbul*, 2006:658-663.

[8] SU Y D. An integrated design of fast LSP data plane failure detection in MPLS-OAM [C]. *Proceedings of Third International Conference on Next Generation Web Services Practices*, *Seoul*, 2007:23-27.

[9] 陈利兵. BFD 技术在 IP 承载网中的应用[J]. 现代电信科技，2008(1):61-64.
CHEN L B. Application of BFD in IP carrying network[J]. *Modern Science & Technology of Telecommunications*, 2008(1):61-64. (In Chinese)

[10] http://www.huawei.com/cn/products/datacomm/detailitem/[OL].

[11] http://www.cisco.com/en/US/docs/ios/12_0s/feature/guide/fs_bfd.html[OL].

作者简介：

付明磊（1981－），男，河北唐山人，博士研究生，分别于 2004 年和 2007 年在浙江工业大学获得学士和硕士学位，主要从事光突发交换网络中拥塞控制技术、节点硬件实现及算法研究等方面的研究。E-mail：fuml_zjut@yahoo.com.cn

导师简介：

乐孜纯（1965－），女，浙江杭州人，教授，博士生导师，1987 年于浙江大学获得学士学位，1998 年于中科院长春光学精密机械与物理研究所获得博士学位，主要从事光纤通信网络组网技术、微结构光电子器件等方面的研究。E-mail：lzc@zjut.edu.cn