

Shortest Path Prioritized Random Deflection Routing (SP-PRDR) in Optical Burst Switched Networks

Craig Cameron, Andrew Zalesky and Moshe Zukerman

ARC Special Research Centre for Ultra-Broadband Information Networks (CUBIN*)

Department of Electrical and Electronic Engineering,

The University of Melbourne, VIC 3010, Australia

Email: {c.cameron, a.zalesky, m.zukerman}@ee.mu.oz.au

Abstract

Optical Burst Switching (OBS) aims to provide higher utilization and flexibility than circuit switching at a lower cost and complexity than the current optical paradigm of multiple Optical-Electronic-Optical conversions. We introduce a new routing protocol for Optical Burst Switching, Shortest Path Prioritized Random Deflection Routing (SP-PRDR), that aims to lower blocking probabilities for all ranges of input loads, topologies and routing matrices while only using state information from traditional Internet Protocol technologies. We show, through analysis and simulation, that blocking probability in OBS networks is significantly reduced by SP-PRDR, with negligible impact on average delay. Additionally, unlike other schemes, we show that the worst case blocking probability of SP-PRDR is provably upper-bounded by the blocking probability of standard OBS.

1 Introduction

Optical Burst Switching (OBS) has been recently introduced as an architecture that may enable low cost, all optical, deployment of future ultra-broadband networks. Packets arriving at an OBS ingress node that are destined for the same egress OBS node and belong to the same Quality of Service class are aggregated and sent in discrete bursts, at times determined by the burst assembly policy. At intermediate nodes, the data within the optical signal is not processed but instead, the whole burst is transparently switched according to directives contained within a control packet preceding the burst. At the egress node, the burst is subsequently de-aggregated and forwarded electronically. Due to the lack of buffers within the network, contention between

bursts can cause high blocking probabilities and subsequent high loss rates even for moderate network utilizations.

Currently, most traffic on the Internet is sent using control protocols that retransmit lost data¹. Assuming this dominance will continue, at least in the near future, minimizing loss probability when designing new network architectures and protocols is a very important goal. In this paper, we focus on one such technique that aims to reduce the loss probability by increasing the round trip time: Deflection Routing. We borrow the definition of deflection routing from an earlier paper [5]: when a control packet arrives at a node and the link corresponding to the next-hop entry in the routing table is fully occupied, it is redirected onto a different unoccupied output link or dropped if all output links are occupied.

Deflection routing was first studied on simple, uniform topologies such as ShuffleNet [7,9] and the Manhattan Street Network [12]. While these topologies are amenable to mathematical analysis, deflection protocols are highly sensitive to topological structure and therefore simulations using more realistic and complex topologies are often necessary [15]. For example, deflection routing performs well on the Manhattan Street Network and ShuffleNet, even under heavy load [5, 10], but recent simulations using a variety of complex networks show that deflecting rather than dropping can cause higher blocking probability for high loads [17, 22] but much lower blocking probabilities for networks with few wavelengths and under light load [6, 17, 19, 20].

The choice of which output links and wavelengths to use for deflected bursts is critical to the overall performance of the network. There are several main types of deflection protocols - fixed alternate routing [6, 10, 19, 20, 22], dynamic traffic-aware [11, 16] and random [17]. The more popular approach is to deflect along a fixed alternate path, either on

*CUBIN is an affiliated program of National ICT Australia

¹Longitudinal study of Internet traffic in 1998-2003, M. Fomenkov, K. Keys, D. Moore and k claffy, http://www.caida.org/outreach/papers/2003/nlanr/nlanr_overview.pdf

a “hop-by-hop” basis [10, 19, 20] or by storing at each node both the complete primary path and the complete alternate path from itself to every possible destination node in the network [6]. While the latter becomes clearly infeasible as the network grows, even the “hop-by-hop” approaches have a significant disadvantage: the choice of a good alternate path is very difficult due to tight coupling between subsequent blocking probabilities, traffic matrices and network topology. In previous papers, the algorithms to chose explicit fixed alternate routes vary considerably, from simple heuristics such as the link-disjoint next-shortest path [22] to the more complex, yet still ad-hoc, approach of requiring alternate routes to return to the original “next-hop” node in less than three hops [19, 20]. In some cases, the algorithm is not given at all [6, 10]. Traffic-aware deflection is even more tightly coupled to transient traffic matrices and the network topology and is consequently prone to instability: continual oscillation between congested and uncongested link states [1]. In most cases, the majority of bursts are carried on the primary routes, therefore to effectively spread the load robustly, the primary routing algorithm must be also dynamic and traffic dependent. Congestion-based routing is outside the scope of this paper, however our deflection algorithm can be implemented with any primary routing algorithm and therefore is also applicable.

Fixed deflection routing has also been shown to destabilize OBS networks for high loads, yielding higher blocking probabilities than if bursts were not deflected but simply dropped [17, 22]. As link utilizations in the internet core have been shown to vary widely even over moderate time scales [3], possible instability in OBS networks is a major concern. Two approaches have been suggested to stabilize deflection routing schemes: wavelength reservation [22] and sender check/retransmission messaging functions [17]. The former approach, in a similar way to classical trunk reservation in circuit switched networks, limits the amount of deflection at high load by reserving wavelengths on each link for the exclusive use of primary bursts. The latter approach avoids deflections for one hop paths by holding bursts at the sender until a wavelength is free. Bursts that are deflected back to their sender are also not transmitted immediately but similarly held.

In this paper, we introduce a new routing algorithm that combines standard Internet Protocol next-hop forwarding with prioritized random deflection for control packets and corresponding bursts: Shortest Path Prioritized Random Deflection Routing (SP-PRDR). We show, through analysis and simulation, that by deflecting otherwise blocked bursts to a random unused output link and reclassifying these deflected bursts with a strictly lower and preemptable priority label, the burst loss rate for all ranges of input loads is strictly decreased compared to the no-deflection case. Unlike all previously proposed deflection schemes, SP-PRDR

only requires primary path next-hop state information at each node and no additional control messaging, yet yields significant performance improvements.

2 Algorithm

Many variants of Optical Burst Switching have been proposed, each with different reservation, contention and routing protocols. In this paper, we add a deflection routing algorithm to the simplified version of the Just-Enough-Time (JET) protocol [21]. JET has one main feature: *delayed reservation*. After the burst is generated at the ingress node, a control packet is sent to reserve bandwidth on each link for the duration of the burst after which the bandwidth is freed and therefore available for future bursts. The burst is then held at the ingress node for a fixed amount of time, equal to or greater than the total processing time of the control packet along the entire path, before being sent into the network. If a control packet is unable to schedule a burst at an intermediate node due to the corresponding output link being fully reserved anytime during the desired period, a contention is said to occur and the control packet and the corresponding burst must be re-routed or dropped (blocked). To simplify notation, as the burst always follows the path of the control packet, we ignore the control packet and refer only to the bursts. In this section, we outline a new approach that re-routes bursts in cases of contention.

2.1 Node Architecture

Technology trends point to network intelligence moving up to IP [8]. In this paper we use the Smart Routers, Simple Optics architecture in which each network node is both an IP router and an optical layer cross-connect [8]. We therefore route traffic in the case of no contention on a hop-by-hop basis with next-hop paths chosen by an unweighted shortest path algorithm, as in OSPF [13]. It is important to note that contention between generated bursts and cross-traffic bursts can be a significant fraction of the overall loss probability. However, as bursts can be buffered on the edges in electronic form, an optimal initial offset time can be chosen to minimize these loss events. As this optimization is independent of routing, it can be implemented in parallel with any routing scheme and therefore will not be considered further in this paper.

2.2 Routing Notation

Routing algorithms can be most succinctly described by representing a network as a standard graph. Using the notation from [2], we define a network as a unidirectional *graph* $G = (N, A)$ with a finite nonempty set N of *nodes* and a collection A of pairs of distinct nodes from N . Each pair

of nodes in A is called an *ordered arc* or a *link*. Note that (n_a, n_b) and (n_b, n_a) are different arcs. A *walk* is a sequence of nodes (n_1, n_2, \dots, n_k) , such that each of the pairs $(n_1, n_2), (n_2, n_3), \dots, (n_{k-1}, n_k)$ are arcs of G . Unlike [2], we denote a *path* as equivalent to a walk.

The routing for each burst follows the reservations and therefore routing of the initial control packet. In the following, we refer to bursts being dropped and deflected, abstracting away the implementation details - control packets being dropped (bursts dropped) and reservations being made for different output links (bursts deflected).

2.3 Shortest Path Random Deflection Routing

For each node, n_i , the shortest path to every other node, $n_j, j \neq i$, is calculated and used to route bursts in the contention-free case. We assume a distributed algorithm, such that each node only contains enough information to route a burst to the corresponding next node. In the case of contention, the contending burst is deflected to a random free output link - the *deflection link*. We assume the set of free links is known and that each link in the set has an equal probability of being chosen. If all output links have previously established overlapping reservations and therefore are not free, the burst is dropped. Upon arrival at the recently chosen random node, the burst is then routed using the shortest path from that node to the burst's destination. For example, let the network topology be NSFNET T3, as shown in Figure 1. The shortest path between nodes 2 and 6 is $(2, 7, 6)$. Let a burst from node 2 experience a contention on link $(2, 7)$ and let both links $(2, 1)$ and $(2, 3)$ be free. If link $(2, 3)$ is chosen as the deflection link, the new complete path is $(2, 3, 5, 6)$, as the shortest path from node 3 to node 6 is $(3, 5, 6)$.

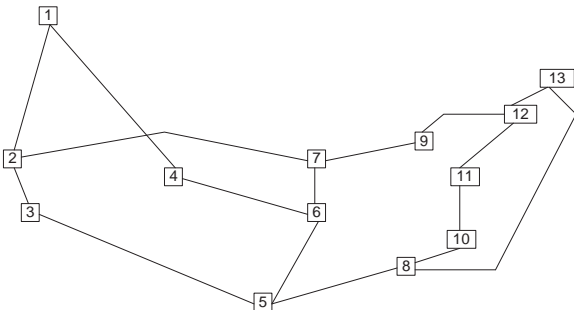


Figure 1. Sample OBS Network - NSFNET T3 comprising 13 OXCs and 32 directed links.

2.4 Shortest Path Prioritized Random Deflection Routing (SP-PRDR)

As discussed in the introduction, Shortest Path Random Deflection Routing has been shown to be unstable at high loads [17]. To solve this stability problem, and thereby achieve blocking probabilities less than or equal to Shortest Path Routing without deflection, we explicitly differentiate between deflected and undeflected bursts by assigning different priorities to the corresponding reservations: P_{low} for the former and P_{high} for the latter. In the case that a future non-deflected control packet with priority P_{high} experiences contention, if there is an overlapping reservation of priority P_{low} , this reservation is preempted, else the control packet is deflected and its priority lowered to P_{low} . Note that if a low priority reservation is preempted, any downstream reservations will be then unused and may unnecessarily block other deflected bursts. It is very important to note these surplus reservations can be preempted by higher priority bursts and therefore do not introduce significant inefficiency. Indeed, in preliminary experiments, having more than two priority levels was found to give very minimal performance improvement, therefore, only two are used in this paper.

Furthermore, it is important to note that the SP-PRDR algorithm allows header packets from P_{high} bursts to preempt a P_{low} bursts even after the P_{low} burst has arrived at the node and is in the process of being switched to an output link. The probability of this occurring increases with the number of hops already traversed by the P_{high} bursts. However deflected bursts have, by definition, a longer average path length and therefore the time between the header packet and the burst arriving is usually shorter than undeflected bursts, reducing the chance of this scenario, especially for small burst sizes. In preliminary experiments, allowing or disallowing "in-flight" preemption had little effect and therefore the default algorithm described above was used.

As every deflected burst can be preempted by a nondeflected burst, the blocking probability using this new algorithm is strictly less than or equal to that of the no deflection case where a burst experiencing contention is immediately dropped. In other words, by deflecting the burst and lowering its priority, it is possible for that burst to be successfully received while having no negative impact on the blocking of the nondeflected bursts. The authors believe that the possibility of using explicit priority levels when reserving bandwidth in OBS networks is a significant unexplored research area and intend to explore other novel algorithms in future work.

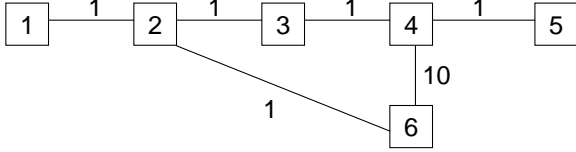


Figure 2. Sample network with weighted links.

2.5 Offset Time

If all links have equal weight, it is also possible to choose an appropriate offset time for systems using SP-PRDR that avoids the insufficient offset time problem outlined in [6]. Let the time taken to process a control packet reservation request at each node be δ , the initial offset time be T , the diameter of the network be D and the number of allowed deflections be d . Now, as mentioned above, after a control packet is deflected from node n_a to a random output node n_b , it is subsequently routed along the shortest path from this new node to its destination. Therefore, if all paths are weighted equally, the length of this new path can not be more than the length of the old path plus two: in the worst case, the control packet loops back from n_b to n_a . Therefore, a suitable minimum offset time is

$$T = (D + 2d) * \delta. \quad (1)$$

If a weighted shortest path algorithm is used, Equation 1 is no longer valid. Note that the diameter, D , is equal to the maximum of the shortest path distances between all possible pairs of nodes of a graph, where the distance between two nodes n_a and n_b in a weighted graph is the sum of weights of the edges of a shortest path between them. For a weighted graph, this is no longer equal to the length of the path. Additionally, replacing D with a new quantity: the maximum hop count of all weighted shortest paths, D^* , also is incorrect. A counter example is given in Figure 2. The weighted shortest path from node 1 to node 5 is (1, 2, 3, 4, 5). Now let a burst from node 1 to node 5 be deflected to node 6 upon arrival at node 4. The complete path is now (1, 2, 3, 4, 6, 2, 3, 4, 5). In this case $D^* = 4$ but the length of the path is now 9. Therefore if weights are used in the shortest path routing computation, additional safeguards, such as time to live hop counters, may be needed.

2.6 Fibre Delay Lines

JET also includes an optional feature: *burst arrival postponing* that further improves blocking probabilities by using Fibre Delay Lines (FDLs) at intermediate nodes. While it has been shown that FDLs decrease the blocking probability [4], they currently require complex hardware and electronic

controls yet are only able to provide on the order of $10\mu s$ delay as the speed of light requires extremely long fibre spools. It is important to note that, while we do not consider FDLs in this paper, our new routing algorithm can be easily adapted to leverage FDL blocking probability gains by randomly deflecting only if the FDL is occupied or the output link is still fully reserved after the FDL delay.

2.7 Loop Removal

A given path P is said to have a routing *loop* if there exists a duplicate node in the path. In traditional buffered packet switched network, loops waste network resources unnecessarily and all routing algorithms strive to be loop-free. In OBS networks, loops can be viewed in a similar way to a FDL but instead of a dedicated fibre, a blocked burst uses the network as a buffer. To the best of the authors' knowledge, all previously introduced OBS routing schemes disallow loops. We therefore also introduce a modified version of SP-PRDR, Shortest Path Prioritized Random Deflection Routing- No Loops, or SP-PRDR-NL, that removes loops from the network. In this new algorithm, when a burst is deflected, the node at which the deflection occurred is added to the header of the control packet. When a control packet arrives at a node, the next-hop node is calculated, the header scanned, and the control packet (and subsequent burst) discarded if the next-hop node is found in the header.

3 Simulation

We use the simulation model from [22] to illustrate the benefit of SP-PRDR for two different routing matrices, consisting of approximately 20 Origin-Destination (O-D) pairs with equal and constant input loads, chosen to represent a variety of path lengths and link-sharing degrees. We use the NSFNET T3 topology, with 80 wavelengths per directed link, as shown in Figure 1 and the full paths are shown in Table 1. Care was also taken to represent different network states: unbalanced (Routing Matrix 1) and balanced (Routing Matrix 2), as shown in Table 2.

We used the same simulation framework from [22] to simulate the performance of the network. This framework introduces several approximations in an attempt to accurately model OBS networks while maintaining feasible simulation times. Bursts are generated by independent Poisson processes, burst transmission times on each link are independent and exponentially distributed with a common mean and deflected bursts are generated according to independent Poisson processes, not more complex and accurate two-state Markov modulated Poisson processes.

Routing Matrix 1		Routing Matrix 2	
Label	Path	Label	Path
A_1	(1,4,6)	A_2	(1,4,6)
B_1	(2,7,9,12,13)	C_2	(8,5,3,2,1)
C_1	(8,5,3,2,1)	D_2	(9,12,11,10)
D_1	(9,12,11,10)	E_2	(13,8,5,6)
E_1	(13,8,5,6)	F_2	(7,6,5,8,10)
F_1	(7,6,5,8,10)	G_2	(3,2,7,9)
G_1	(3,2,7,9)	H_2	(12,9,7,2)
H_1	(12,9,7,2)	J_2	(4,6,7)
I_1	(11,10,8,5,3)	K_2	(8,10,11)
		L_2	(11,12,13)
		M_2	(12,11,10,8)

Table 1. Global Routing table for simulated Origin-Destination pairs. All traffic flows are bi-directional.

Flows	Routing Matrix 1	Routing Matrix 2
0	0	0
1	6	2
2	6	13
3	2	1
4	1	0

Table 2. Number of bidirectional O-D pair flows for each link in network.

4 Results

We compared the burst loss probability performance, the average number of hops taken for successful burst transmissions and the link utilizations of Shortest Path Prioritized Random Deflection Routing and Shortest Path Random Deflection Routing with a maximum of one and two deflections allowed and standard Shortest Path with no deflection. The results for both Routing Matrices are plotted below in Figure 3 to Figure 9.

Simulations were run for sufficient time to ensure that error bars corresponding to 95 percent confidence intervals, using a Gaussian approximation across 5 independent runs, were mostly smaller than the shapes used to mark the data points, except at the lowest value of load. Due to computational time constraints, probabilities lower than 10^{-5} are not precise and probabilities lower than 10^{-6} are not included in the results.

Confirming what was previously noted in [17], the loss probability for random deflection without priority is even higher than the fixed, non-deflecting case, for both routing matrices, as shown in Figure 3 and Figure 4. For the balanced case, the instability is even more apparent as de-

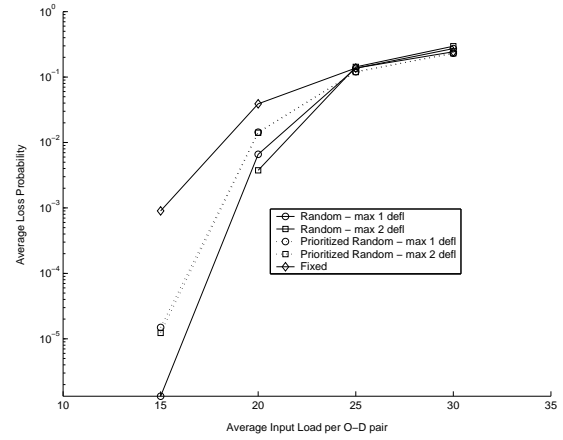


Figure 3. Average O-D Pair Blocking Probability, Routing Matrix 1.

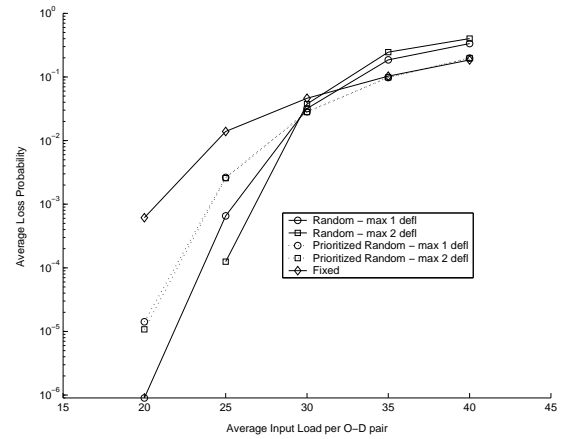


Figure 4. Average O-D Pair Blocking Probability, Routing Matrix 2.

flected bursts no longer are sent over under-utilized links and therefore have a much higher chance of blocking non-deflected bursts. However, SP-PRDR significantly reduces the loss probability for low input loads, while is equal to the non-deflection case for higher loads.

This dependence on load can be seen more clearly in Figure 5 and Figure 6. As the input load increases, many more deflections occur in the non-prioritized case, while bursts are preempted and consequently dropped for SP-PRDR.

Both deflection schemes utilize the network more than the non-prioritized case, as shown in Figure 7, Figure 8 and Figure 9. This is intuitively obvious as both schemes deflect bursts instead of dropping, adding additional traffic to the network. However, by using SP-PRDR, this additional

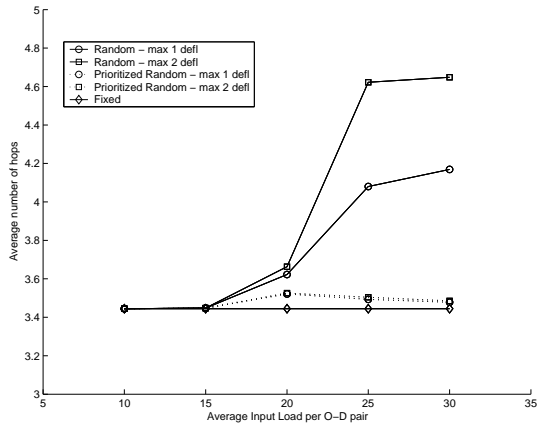


Figure 5. Average Number of Hops for successful transmissions, Routing Matrix 1.

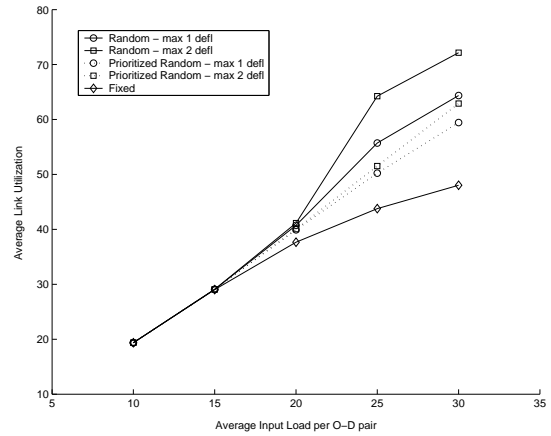


Figure 7. Utilization averaged over all links, Routing Matrix 1.

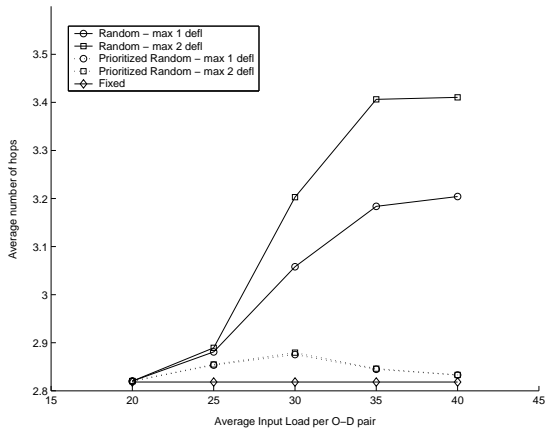


Figure 6. Average Number of Hops for successful transmissions, Routing Matrix 2.

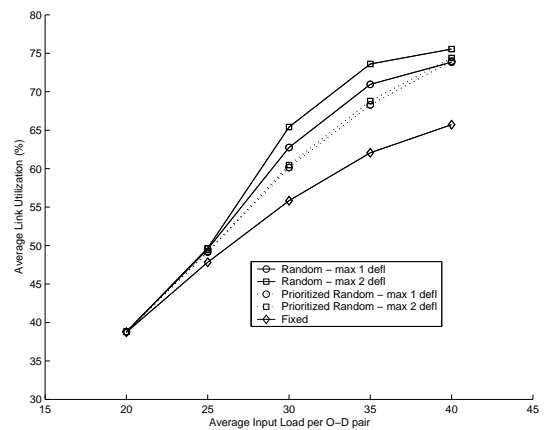


Figure 8. Utilization averaged over all links, Routing Matrix 2.

traffic, while adding to overall network utilization, can be preempted by the non-deflected bursts and therefore does not induce instability and consequent higher loss.

5 Conclusion

We introduced a new routing protocol for Optical Burst Switching, SP-PRDR, that only requires state information from traditional Internet Protocol technologies. We showed, through analysis and simulation, that the blocking probability in OBS networks is significantly reduced by SP-PRDR, with negligible impact on average delay, and that worst case performance is upper-bounded by the blocking probability of standard OBS, due to the convergence of SP-PRDR to

fixed, non-deflection routing at high load.

6 Acknowledgment

This work was supported by the Australian Research Council.

References

- [1] D. Awduche, A. Chiu, A. Elwalid, I. Widjaja, and X. Xiao. Overview and principles of internet traffic engineering. *RFC 3272*.
- [2] D. Bertsekas and R. Gallager. *Data Networks*. Prentice Hall, 2nd edition, 1992.

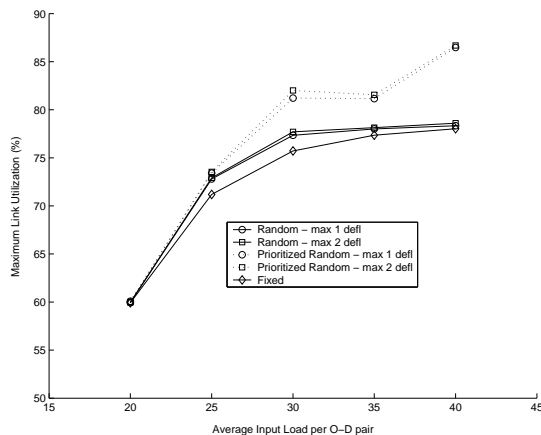


Figure 9. Maximum Link Utilization, Routing Matrix 2 (5-8).

- [3] S. Bhattacharyya, C. Diot, J. Jetcheva, and N. Taft. Geographical and temporal characteristics of inter-POP flows: View from a single POP. *European Transactions on Telecommunications*, 13(1):5–22, Feb. 2002.
- [4] I. Chlamtac, A. Fumagalli, and C.-J. Suh. Multibuffer delay line architectures for efficient contention resolution in optical switching nodes. *IEEE Transactions on Communications*, 48(12):2089–2098, Dec. 2000.
- [5] A. K. Choudhury and V. O. K. Li. An approximate analysis of the performance of deflection routing in regular networks. *IEEE Journal on Selected Areas in Communications*, 11(8):1302–1316, Oct. 1993.
- [6] C.Hsu, T. Liu, and N. Huang. Performance analysis of deflection routing in optical burst-switched networks. In *Proc. IEEE INFOCOM 2002*, pages 66–73, 2002.
- [7] F. Forghieri, A. Bononi, and P. R. Prucnal. Analysis and comparison of hot-potato and single-buffer deflection routing in very high bit rate optical mesh networks. *IEEE Transactions on Communications*, 43(1):88–98, Jan. 1995.
- [8] G. Hjalmtysson, J. Yates, S. Chaudhuri, and A. Greenberg. Smart routers-simple optics: an architecture for the optical Internet. *Journal of Lightwave Technology*, 18:1880–1891, Dec. 2000.
- [9] M. G. Hluchyj and M. Karol. Shufflenet: An application of generalized perfect shuffles to multihop lightwave networks. In *Proc. IEEE INFOCOM 1988*, pages 379–390, 1988.
- [10] S. Kim, N. Kim, and M. Kang. Contention resolution for optical burst switching networks using alternative routing. In *Proc. IEEE International Conference on Communications*, volume 5, pages 2678–2681, Apr. 2002.
- [11] S. K. Lee, K. Sriram, H. S. Kim, and J. S. Song. Contention-based limited deflection routing in OBS networks. In *Proc. GLOBECOM*, Dec. 2003.
- [12] N. Maxemchuk. Regular mesh topologies in local and metropolitan area networks. *AT&T Technical Journal*, 65:1659–1685, Sept. 1985.
- [13] J. Moy. OSPF version 2. *RFC 2328*, 1998.
- [14] J. Padhye, V. Firoiu, D. F. Towsley, and J. F. Kurose. Modeling TCP Reno performance: A simple model and its empirical validation. *IEEE/ACM Transactions on Networking*, 8(2):133–145, Apr. 2000.
- [15] V. Paxson and S. Floyd. Why we don't know how to simulate the internet. In *Proc. 1997 Winter Simulation Conference*, pages 1037–1044, 1997.
- [16] G. P. V. Thodime, V. M. Vokkarane, and J. P. Jue. Dynamic congestion-based load balanced routing in optical burst-switched networks. In *Proc. Globecom 2003*, pages 2628–2632.
- [17] X. Wang, H. Morikawa, and T. Aoyama. Burst optical deflection routing protocol for wavelength routing WDM networks. *Optical Networks Magazine*, 3(6):12–19, Nov. 2002.
- [18] Y. Xiong, M. Vandenhouste, and H. C. Cankaya. Control architecture in optical burst switched WDM networks. *IEEE Journal on Selected Areas of Communication*, 18(10):1838–1851, Oct. 2000.
- [19] S. Yao, B. Mukherjee, S. J. B. Yoo, and S. Dixit. All-optical packet-switched networks: a study of contention-resolution schemes in an irregular mesh network with variable-sized packets. In *Proc. OptiComm 2000*, Oct. 2000.
- [20] S. Yao, B. Mukherjee, S. J. B. Yoo, and S. Dixit. A unified study of contention-resolution schemes in optical packet-switched networks. *Journal of Lightwave Technology*, 21(3):672–683, Mar. 2003.
- [21] M. Yoo and C. Qiao. Just-enough-time (JET): a high speed protocol for bursty traffic in optical networks. *1997 Digest of the IEEE/LEOS Summer Topical Meetings*, pages 26–27, Aug. 1997.
- [22] A. Zalesky, H. L. Vu, Z. Rosberg, E. W. M. Wong, and M. Zukerman. Reduced load erlang fixed point analysis of optical burst switched networks with deflection routing and wavelength reservation. In *Proc. First International Workshop on Optical Burst Switching (WOBS)*, Oct. 2003.