Terabit Burst Switching
Progress Report (4/01-6/01)

Jonathan S. Turner
jst@cs.wustl.edu

WUCS-01-23

August 10, 2001

Department of Computer Science
Campus Box 1045
Washington University
One Brookings Drive
St. Louis, MO 63130-4899

Abstract

This report summarizes progress on Washington University's *Terabit Burst Switching* Project, supported by DARPA and Rome Air Force Laboratory. This project seeks to demonstrate the feasibility of *Burst Switching*, a new data communication service which can more effectively exploit the large bandwidths becoming available in WDM transmission systems, than conventional communication technologies like ATM and IP-based packet switching. Burst switching systems dynamically assign data bursts to channels in optical data links, using routing information carried in parallel control channels. The project will lead to the construction of a demonstration switch with throughput exceeding 200 Gb/s and scalable to over 10 Tb/s.

This work is supported by the Advanced Research Projects Agency and Rome Laboratory (contract F30602-97-1-2703).

Terabit Burst Switching

Progress Report (4/01-6/01)

Jonathan S. Turner
jst@cs.wustl.edu

This report summarizes progress on the Terabit Burst Switching Project at Washington University for the period from March 1, 2001 through June 30, 2001.

1. Prototype Burst Switch Progress

The following paragraphs summarize status and progress on the various components being developed for the prototype burst switch. Figure 1 shows the overall structure of the prototype and details the location of each component in the system architecture.

- *PC Boards and Physical Design.* There are six different printed circuit board designs that are required for the burst switch prototype. We have completed the layout of the IO Board, BSE Control board and the backplane for the burst switch. We expect to complete the BSE Datapath Board by September 1, the Miscellaneous Board by October 1 and the ATM Interface Board by December 1. We expect to assemble and begin testing the system by the end of the year.
- *Crossbar* (XBAR). The crossbar is the principal component of the burst switch datapath. It accepts 256 1 Gb/s data streams and switches each to one of 256 outputs for an aggregate data rate of 256 Gb/s. It uses a bit-sliced organization with nine parallel planes for carrying the data. Control inputs allow an input port to be selected for each output port, and an input with null data is selected when an output is unused. Successive groups of 32 outputs have independent control sections, enabling different Burst Processors to manage connections to the outputs they are responsible for without contention from other Burst Processors. The crossbar is being implemented using 36 Xilinx Virtex FPGAs. The VHDL for the modified design is completed, and has been simulated and synthesized.
- *Synchronization Chip* (SYNC). The SYNC circuit accepts data from Serializer/ Deserializer (SERDES) chips, delays it for up to 50 μs, and passes it on to the crossbar chips. Data returned from the crossbar chips is returned to the SYNC circuit and passed to the SERDES for transmission over the optical fibers. The SYNC chip is being implemented using Xilinx Virtex FPGAs. Each chip implements four channels. The VHDL for the SYNC chip has been completed, simulated and synthesized.
- *Burst Storage Unit* (BSU). The BSU provides an interface between the crossbar and the memory in which bursts are stored when they cannot be sent directly to the outgoing links. Each BSU supports 32 channels and has an aggregate throughput of 4 Gb/s. It uses a 128 bit wide memory, made up of four 1 MB static RAM chips, plus two additional memory chips

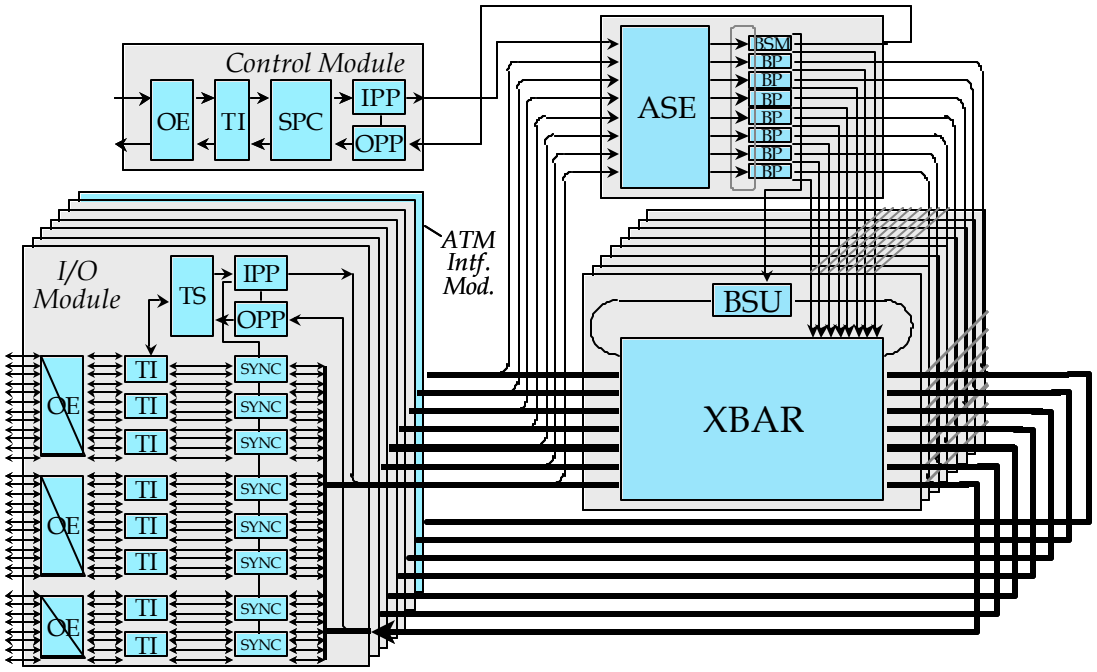

Figure 1. Prototype Burst Switch

which hold linked list pointers to enable the BSU to manage the memory in a flexible fashion. It is being implemented using an FPGA, to minimize risk and provide flexibility for alternative implementations. Arriving data enters through one of 32 shift registers and is then multiplexed through to the memory interface where it is stored. Departing data passes through a similar data path.

- *Burst Processor* (BP). The BP is the most important single component of the burst switch. Each BP is responsible for managing 31 outgoing channels from the crossbar. It maintains a schedule for those outgoing channels, and assigns incoming bursts to places in the schedule, using the information it receives in *Burst Header Cells* that come to it through the ATM Switch Element (ASE). The BP also communicates with the Burst Storage Manager (BSM) through a local control ring and has connections that can be used to communicate with upstream and downstream neighbors in a multistage configuration.
- *Burst Storage Manager* (BSM). The BSM schedules the storage of bursts in the BSU. This component is not required in Phase 1, but will be included in Phase 2. The physical hardware configuration of the BSM is identical to the BP; just the programming of the FPGAs is different. The BSM requires separate channel managers for its input and output interfaces. In addition, it has a controller to manage the storage within the BSU.

  For Phase 2, we have decided to use a simplified version of the general storage management data structure that we have developed. It involves a differential search tree, but we allocate storage to a burst from the time a burst header cell is received, rather than waiting until the burst arrives. There is little performance penalty incurred by this, in systems where there is only a small variation in the time between arrival of a Burst Header Cell and arrival of the corresponding burst.

- *Time Stamp Chip* (TS). The TS chip adds a system-wide timestamp to arriving BHCs and provides delay compensation on both the input and output sides of the system. On the input side, this is intended to enable compensation of known variable delays associated with different channels (in a system with WDM links, such delay variations are caused by the wavelength dependence of the speed of light). On the output side, it compensates for varying delays that BHCs experience when passing through the system. The TS also converts between conventional time units used on the external links and internal time units based on the switch's internal clock frequency. This allows various internal components to perform timing in terms of clock ticks and reduces the number of components that require precise timing calibration.

  The TS chip is being implemented in an FPGA. The logic has been designed, simulated, placed and routed and the device is expected to run at the required clock rates (125 MHz for the interface to the transmission circuits and 50 MHz for other parts).

- *ATM Interface Module*. The ATM interface module is an IO card that allows data to be received from an ATM switch and converted into a burst that is suitable for transmission through a burst switching network. Details of the AIM appear in an earlier progress report [TU99d].

- *ATM Switch Components*. Three components from the Washington University Gigabit ATM switch are being used within the burst switch prototype. The next section describes progress on the 160 Gb/s configuration of that switch that is now being assembled. It includes descriptions of the ATM components being used in the prototype burst switch, and their status.

- *Transmission Interfaces* (TI). Transmission formatting will be provided using quad gigabit serial link components made by AMCC. Each of these components has four gigabit transmitters and four gigabit receivers. The chips encode the data for transmission using a 4B/5B line code, decode it on reception and recover clock from the received bit stream. Each IO module will have eight of these components. Samples of these components have been obtained and evaluated in a test fixture.

- *Optoelectronics* (OE). The optical interfaces will be implemented using VCSEL array devices that handle 12 serial data channels at data rates of 1.25 Gb/s and distances up to 500 meters. The specific devices that we plan to use are the Siemens Parallel Optical Link components (PAROLI).

2. 160 Gb/s ATM Switch

The following paragraphs summarize status and progress on the various components being developed for the 160 Gb/s ATM switch being constructed as part of this project. Several of these components are common with the burst switch. Figure 2 shows the overall structure of the prototype and details the location of each component in the overall architecture.

- *PC Boards and Physical Design*. The designs for all the printed circuit boards required for the system have been completed, and all boards have been fabricated or are in the process of being fabricated. Specifically, the quad-OC-12 line cards and dual G-link line cards have been fabricated and tested. Initial testing of the IO Board revealed a flaw in the Switch Element chip reported earlier. This has made it necessary to tie off eight signals on each

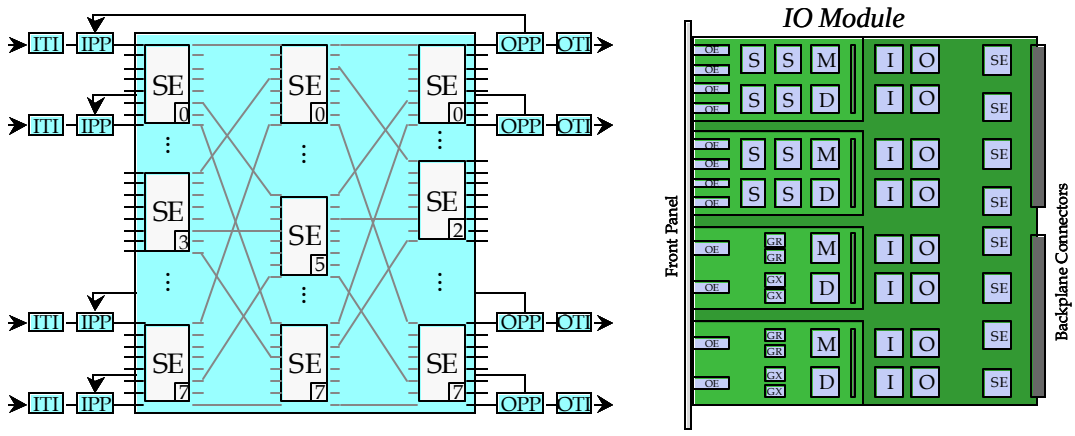

Figure 2. 160 Gb/s ATM Switch

Switch Element chip by adding a resistor to ground. While we considered manually adding these resistors to the original board design, we concluded that because of the density of the solder balls on the BGA packages, it would be difficult to do this reliably. Consequently, we made the required changes to the IO board design and are having the modified design fabricated. The Switch Element board and backplane designs have been completed and these boards are being fabricated now.

We have recently encountered a problem with one of our suppliers. The company we have been using to solder the components to our printed circuit boards has changed their business plan and is no longer accepting complex, high performance board designs like the ones that we require. We are currently investigating alternative suppliers and expect to be able to establish a new working relationship with one of these companies.

- *ATM Switch Element* (ASE). This chip is a revised version of a chip that was developed in an earlier project. The new chip implements four priority classes, doubles the cell buffering of the previous chip and corrects timing flaws that limited the operational frequency of the original chip. As reported earlier, the chip has a design flaw that disables one of the eight input ports, but we do not expect this to prevent us from achieving our primary objectives.
- *ATM Input Port Processor* (IPP). The IPP is a modified version of a component developed for an earlier project. The new chip provides a larger VPI/VCI lookup table (4096 entries instead of 1024) and allocates those entries more flexibly. It also implements features for reliable multicast and provides more extensive support for traffic monitoring. The IPP has been implemented in a .35 micron ASIC process. As reported earlier, the circuit has a design flaw that prevents us from using the reliable multicast feature. We have concluded that there are no realistic options available to us for correcting the design, so are proceeding with the current chips.
- *ATM Output Port Processor* (OPP). This chip was developed in an earlier project. The required die (fabricated in a .7 micron ASIC process) have been packaged in ball grid array packages (rather than the original pin grid array package) to make them compatible with other components in the system. The repackaged chips have been tested and work correctly.

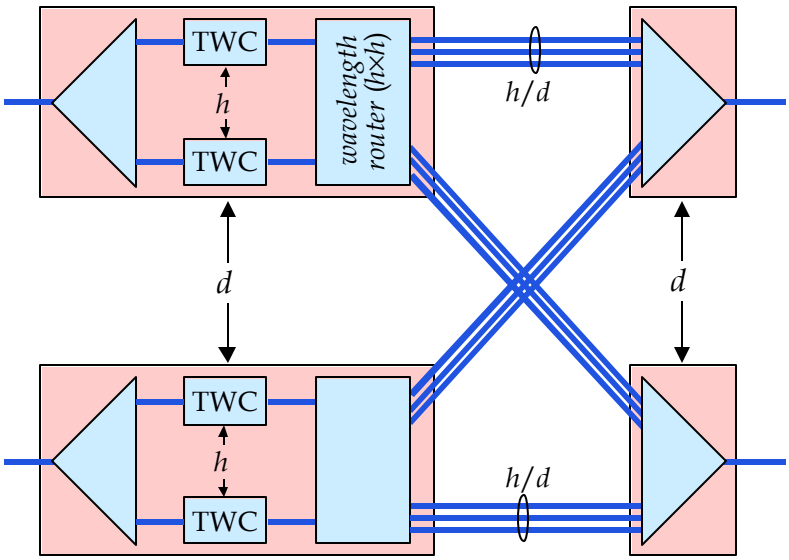

Figure 3. Wavelength Switch Using Tunable Wavelength Converters (TWC) and Passive Wavelength Routers (AWGN)

- *Dual G-link Line Card*. This card multiplexes a pair of 1 Gb/s links onto a single core switch port, using an FPGA to perform the input-side multiplexing and output-side demultiplexing. This card is now complete, tested and all 24 planned units have been produced.
- *Quad OC-12 Line Card*. This card multiplexes four OC-12 links onto one switch port. It uses an FPGA to do the required multiplexing and demultiplexing. This board is now complete, tested and all 24 planned units have been produced.
- *OC-48 Line Card*. This card terminates a single OC-48 link. The schematic for this board has been completed, but the layout has been deferred until the burst switch boards are completed. We expect to complete the layout by the end of the year and fabricate the boards in the first quarter of 2002.

3. Burst Switch Architecture Studies

Our last report discussed two new optical datapath designs for a Burst Switch Element based on tunable lasers. We have been continuing to investigate the second of these two designs which uses tunable lasers and passive wavelength routers to provide space-division switching.

The design is shown in Figure 3. Note that the tunable wavelength converters are the only active components. Since the wavelength routers have h inputs and h outputs, h/d fibers connect each input section with each output section. For h=256 and d=8, there will be 32 fibers connecting each input section with each output section. In this design, the tunable wavelength converters serve two purposes. First, they provide the required space switching. By tuning the laser to one wavelength in the appropriate set of h/d wavelengths, we can "steer" the signal to the desired output port. At the same time, we need to avoid wavelength conflicts on the output links of the system, so the choice of output wavelengths is constrained. The implication of this, is that the burst switch design in Figure 3, is not a nonblocking design. That is, there may be

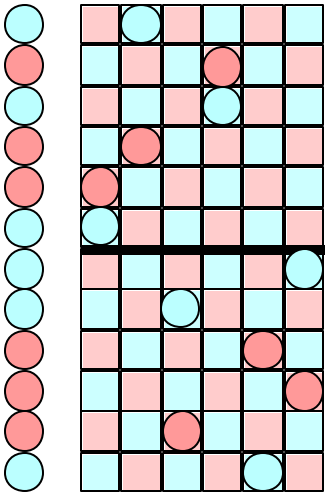

Figure 4. Puzzle Formulation of Wavelength Routing Problem

situations where all of the wavelengths that can be used to get to a desired output, are in use, causing blocking to occur, even when there are free wavelengths available on the outgoing link. However, our studies of systems of this type indicate that for practical values of h and d (in particular, when h is much larger than d), the impact of blocking on system performance can be acceptably small.

The likelihood of blocking in these systems is largely determined by the pattern of interconnections used to connect the input sections with the output sections. If we select this pattern appropriately, we can dramatically increase the number of wavelengths that will be available to route from a set of inputs to any given output. While any individual input channel has h/d wavelengths it can use to reach an output, a pair of input channels may have close to $2h/d$ wavelengths that they can use to reach a given output. We can reduce the likelihood of blocking by using an interconnection pattern for which different input channels share only a few wavelengths for reaching any given output.

The problem of finding the best interconnection pattern can be stated in an alternative way, which makes the key issues easier to grasp. The re-statement takes the form of a puzzle, as illustrated in Figure 4. The puzzle consists of a game board with colored squares and colored tokens. One sets up the puzzle by placing the tokens alongside the game board, one token for each row of the game board. One solves the puzzle by placing each token in a square in its row. Tokens must be placed on squares of matching colors and no two tokens of the same color may appear in the same column. Figure 4 shows both the initial setup for the puzzle (the tokens that appear just to the left of the game board) and a corresponding solution (the tokens that appear on the game board).

This particular puzzle corresponds to routing wavelengths through a burst switch element with 2 inputs and outputs and 8 wavelengths on each link. Each row in the game board corresponds to a different input channel of the burst switch element, and each column corresponds to a wavelength. The color of a token corresponds to the output that the given input channel is to be routed to, and placing a token in a given column corresponds to tuning the tunable wavelength converter for that input channel to the wavelength that corresponds to that column. The color of

a square corresponds to the output that will be reached by using the wavelength corresponding to the square's column. Thus, placing a token on a square of the same color corresponds to selecting a wavelength that goes to the desired output. Whenever the puzzle has a solution, it means that there is a way to route the input signals to the output channels that are specified by the tokens placed in each row. If the puzzle does not have a solution, then there is no way to route the channels.

We have recently shown that for any puzzle that has a solution, we can find the solution by solving a maximum size matching problem in a bipartite graph. There are efficient algorithms for solving this problem. The worst-case time for the best general algorithm is $O(h^{5/2})$. We expect that there may be opportunities to improve this further by exploiting the special structure of the problem. When the puzzle does not have a solution, we can use the maximum size matching solution to place the largest possible number of tokens on the board, or equivalently, to route the largest possible number of wavelengths through the switch.

While it is known that the puzzle does not always have a solution when there are more than $h-(d-1)$ tokens of any one color, in the optical switching context, h is typically much larger than d, and consequently this result is not necessarily a serious problem. It just means that each fiber must have a few "extra" wavelengths that are not actually used, but which provide the routing flexibility needed to avoid blocking. In fact, even a switch that is subject to blocking may be acceptable, if the probability of blocking is sufficiently low. We have performed a simulation study of the blocking performance of a burst switch element using the proposed datapath design. The random arrival of bursts at a switch element means that some bursts will be blocked due to contention at the output link, even if a nonblocking datapath is used. Our study showed the effect of the blocking datapath on the fraction of bursts that are lost at a burst switch element. The results are shown in Figure 5. For h=256, the nonblocking switch can operate at a load of 75% with a burst rejection probability of 1 in a million. The blocking switch can operate at a load of about 62% with the same burst rejection probability, meaning that it can provide about 83% of the throughput of the nonblocking switch. This result was obtained with a random game board configuration. With a systematically constructed game board, we can achieve 87% of the throughput of the nonblocking switch.

There are two further improvements that we are now investigating. First, the above results were obtained using a heuristic algorithm for selecting the wavelength used by arriving bursts. We are now investigating whether it is possible to improve on the performance using a different routing algorithm. The recent result showing that the puzzle problem can be solved using bipartite matching, provides a tool for determining how close the current algorithm is to theoretically optimal performance. The second improvement has to do with buffering. The results in Figure 4 are for a system with no ability to buffer bursts. While the addition of buffering will improve the performance of both the blocking and the nonblocking designs, we expect it to have a bigger impact on the blocking design, further reducing the performance gap.

Burst switching systems with large numbers of channels per link can provide excellent statistical multiplexing performance with little or no buffering. However, buffering can improve performance further. Since optical buffering remains a very expensive option, electronic memory remains the most cost-effective approach, even with the added expense of the electro-optic conversions needed to buffer bursts electronically. Since only a small fraction of the data that passes through a system requires buffering, the added cost of the buffers can be kept acceptably small. We have recently realized that there is an additional benefit that can be

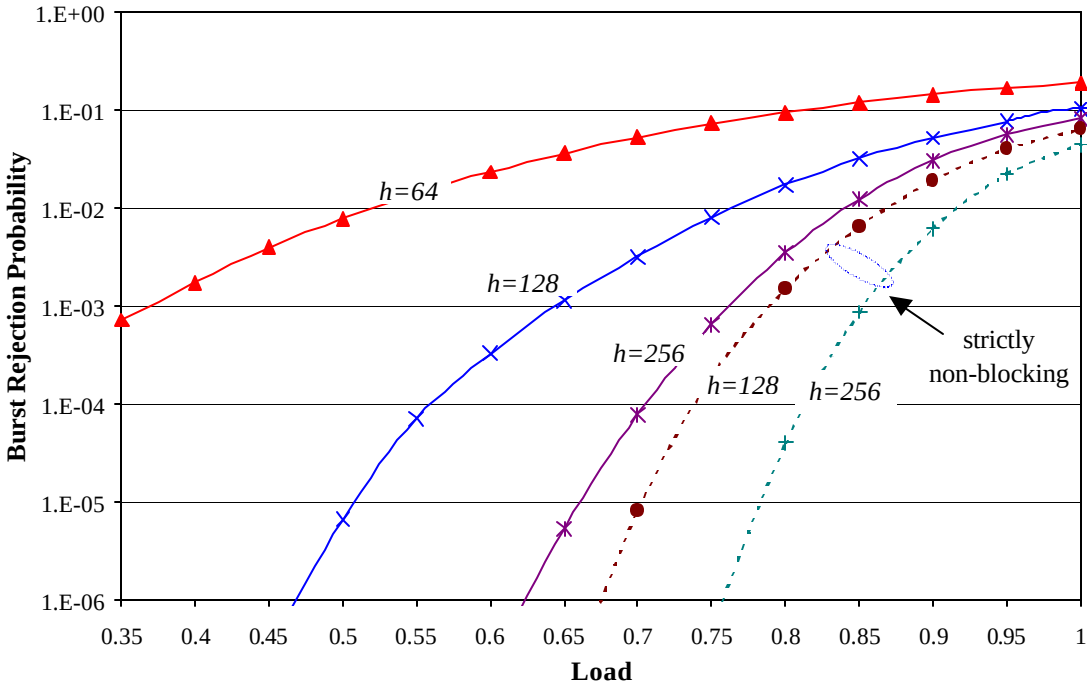

Figure 5. Burst Rejection Performance of Burst Switch Elements (d=8)

obtained in burst-switched networks that use electronic buffering. Since each time a burst it is stored, the optical signal is fully regenerated on output, we can extend the reach of burst-switched networks, without requiring optical regeneration. To do this in a systematic way, a field would be added to each Burst Header Cell, that provides a measure of how far the burst has traveled since it was last regenerated (more precisely, it records how much signal degradation that burst has potentially been subjected to). The electronic control subsystem of each burst switch would update this field, ensuring that bursts are regenerated from time to time as they pass through a large network. This effectively eliminates any constraint on the range of a burst-switched network, and relaxes the requirements on the optical components used in transmission and switching. Regeneration should only be needed at a relatively small fraction of the switching stages, meaning that such systems would still retain the essential advantage of optical networks over electronic networks, in reducing the number of electro-optic conversions required.

REFERENCES

[EA99] Ramamirtham, Jeyashankher and Jonathan Turner. *Design of Wavelength Converting Switches for Optical Burst Switching*. Submitted to *Infocom 2002*, 7/01.

[TU98a] Turner, Jonathan S."Terabit Burst Switching," Washington University Technical Report, WUCS-98-17, 1998.

[TU98b] Turner, Jonathan S. "Terabit Burst Switching Progress Report (12/97-3/98)," Washington University Technical Report, WUCS-98-16, 1998.

[TU98c] Turner, Jonathan S. "Terabit Burst Switching Progress Report (3/98-6/98)" Washington University Technical Report, WUCS-98-22, 1998.

[TU98d] Turner, Jonathan S. "Terabit Burst Switching Progress Report (6/98-9/98)" Washington University Technical Report, WUCS-98-30, 1998.

[TU98e] Turner, Jonathan S. "Terabit Burst Switching Progress Report (9/98-12/98)" Washington University Technical Report, WUCS-98-31, 1998.

[TU99a] Turner, Jonathan S. "Terabit Burst Switching," *Journal of High Speed Networks*, vol. 8, no. 1, 1999.

[TU99b] Turner, Jonathan S. "WDM Burst Switching," *Proceedings of INET*, San Jose, CA, 6/99.

[TU99c] Turner, Jonathan S. "WDM Burst Switching for Petabit Capacity Routers," *Proceedings of Milcom*, Atlantic City, NJ, 11/99.

[TU99d] Turner, Jonathan S. "Terabit Burst Switching Progress Report (1/99-6/99)" Washington University Technical Report, WUCS-99-21, 1999.

[TU99e] Turner, Jonathan S. "Terabit Burst Switching Progress Report (7/99-12/99)" Washington University Technical Report, WUCS-99-32, 1/2000.

[TU00a] Turner, Jonathan S. "WDM Burst Switching for Petabit Data Networks" *Proceedings of the Optical Fiber Conference*, 3/2000.

[TU00b] Turner, Jonathan S. "Terabit Burst Switching Progress Report (1/00-6/00)" Washington University Technical Report, WUCS-00-18, 7/2000.

[TU00c] Turner, Jonathan S. "Terabit Burst Switching Progress Report (7/00-9/00)" Washington University Technical Report, WUCS-00-28, 10/2000.

[TU01] Turner, Jonathan S. "Terabit Burst Switching Progress Report (10/00-3/01)" Washington University Technical Report, WUCS-01-09, 5/2001.