Terabit Burst Switching Progress Report (1/00–6/00)

Jonathan S. Turner
jst@cs.wustl.edu

WUCS-00-18

August 21, 2000

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.

[0] This work is supported by the Advanced Research Projects Agency and Rome Laboratory (contract F30602-97-1-0273).

Terabit Burst Switching
Progress Report (1/00–6/00)

Jonathan S. Turner
jst@cs.wustl.edu

This report summarizes progress on the Terabit Burst Switching Project at Washington University for the period from January 1, 2000 through June 30, 2000.

1. Prototype Burst Switch Progress

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

- *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 72 Xilinx Virtex FPGAs. The VHDL for the crossbar is completed and has been fully simulated.

- *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.

- *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 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

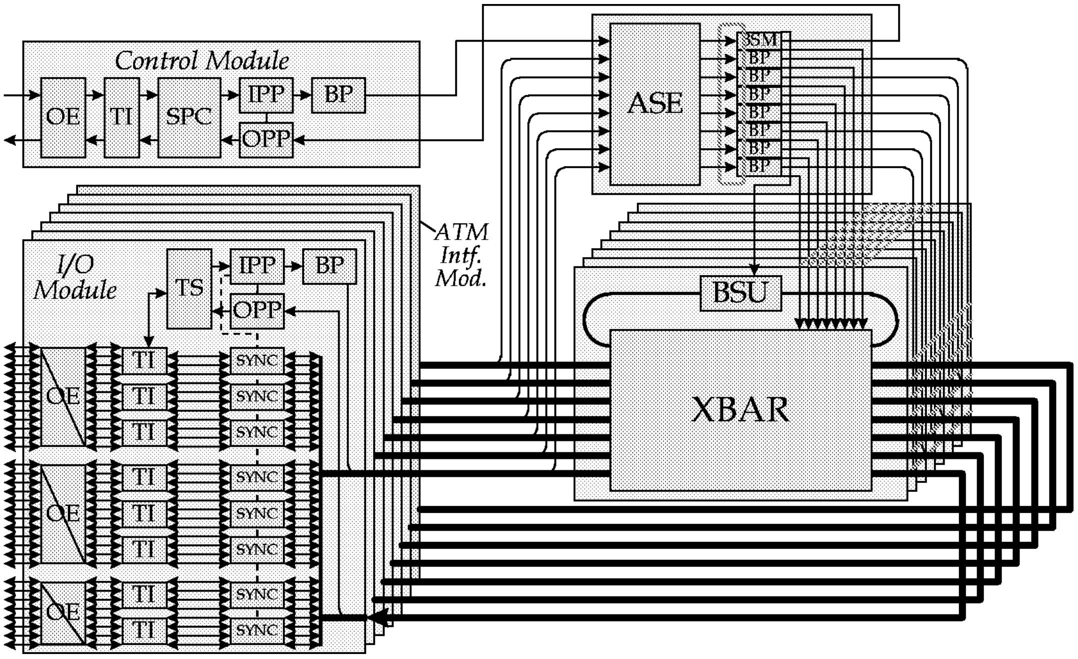

Figure 1: Prototype Burst Switch

for alternative implementations. Figure 2 is a block diagram of the BSU FPGA. 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.

  Figure 3 shows the organization of the Phase 1 BP. It is being implemented using a pair of FPGAs. Each has an external memory that can be used for either data storage or for control information. The Phase 1 BP logic includes a horizon link scheduler and a crossbar controller for managing the creation and removal of connections in the crossbar at the appropriate times. The Phase 1 BP logic has been designed, synthesized, simulated, placed and routed. Timing verification shows that it will run at the designed clock rate (50 MHz). The logic uses about 30% of the logic blocks in the two FPGAs.

  Figure 4 shows the Phase 2 BP. It will use the same physical hardware as the Phase 1 BP. We will merely re-program the FPGAs to provide the required expanded functionality. The only completely new blocks are the ring interfaces that connect to the control ring. The control ring has been split into two parts. One is for communicating with the Burst Storage Manager, and the other enables the BPs within a BSE to exchange status information. Only the first part is needed in Phase 2. The Phase 2 BP maintains many of the same components as in

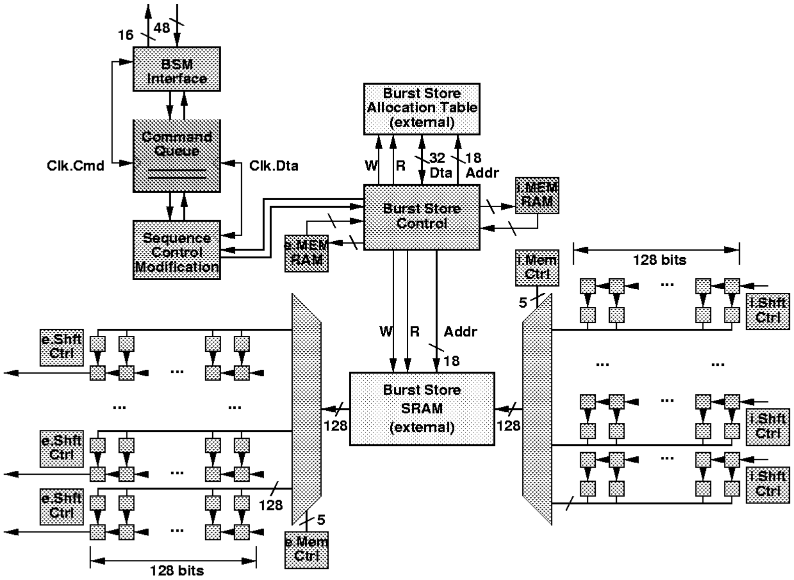

Figure 2: Burst Storage Unit Controller Block Diagram

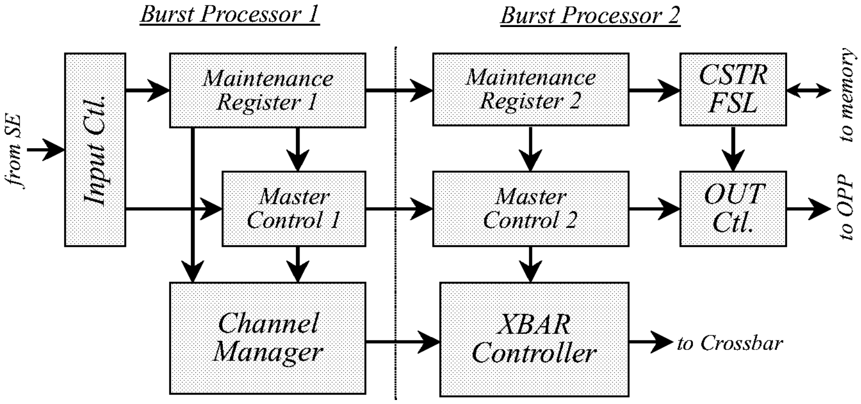

Figure 3: Phase 1 Burst Processor

Phase 1, but significant changes are needed to the Channel Manager and Master Control 1. The modified channel manager generates a storage request, if an arriving burst cannot be directly switched through to its output. On receiving a reply, it either completes scheduling of the selected outgoing channel (if the BSM accepts the storage request) or cancels the scheduling operation. While the BP is waiting for the reply, the tentatively selected channel is unavailable for selection by other bursts.

- *Burst Storage Manager* (BSM). The BSM schedules the storage of bursts in the BSU. This component is not required in Phase 1, but the design is now under way in preparation for Phase 2. A block diagram is shown in Figure 5. 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

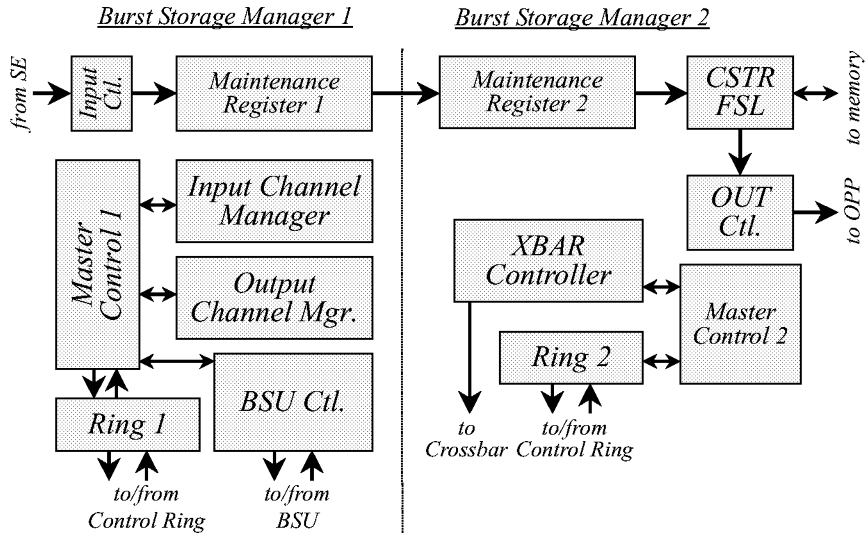

Figure 4: Phase 2 Burst Processor

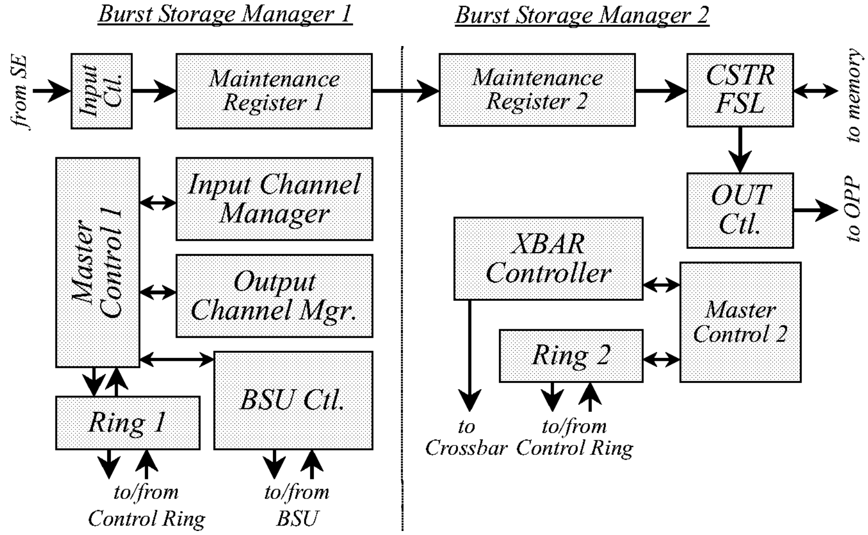

Figure 5: Burst Storage Manager

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).

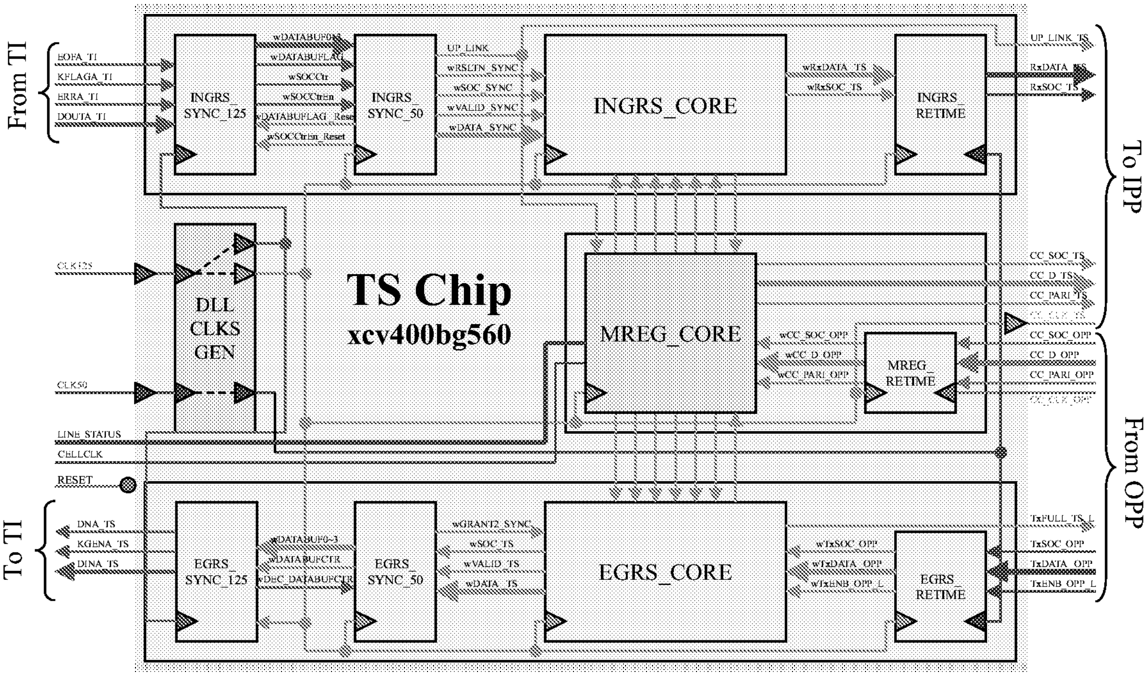

Figure 6: Timestamp Chip

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.

A block diagram of the TS chip is shown in Figure 6. It contains two main data paths, an *ingress path* and an *egress path.* It also has a separate pair of interfaces used for control purposes. 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 were given in a previous progress report [11].

- *ATM Switch Components.* Three components from 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

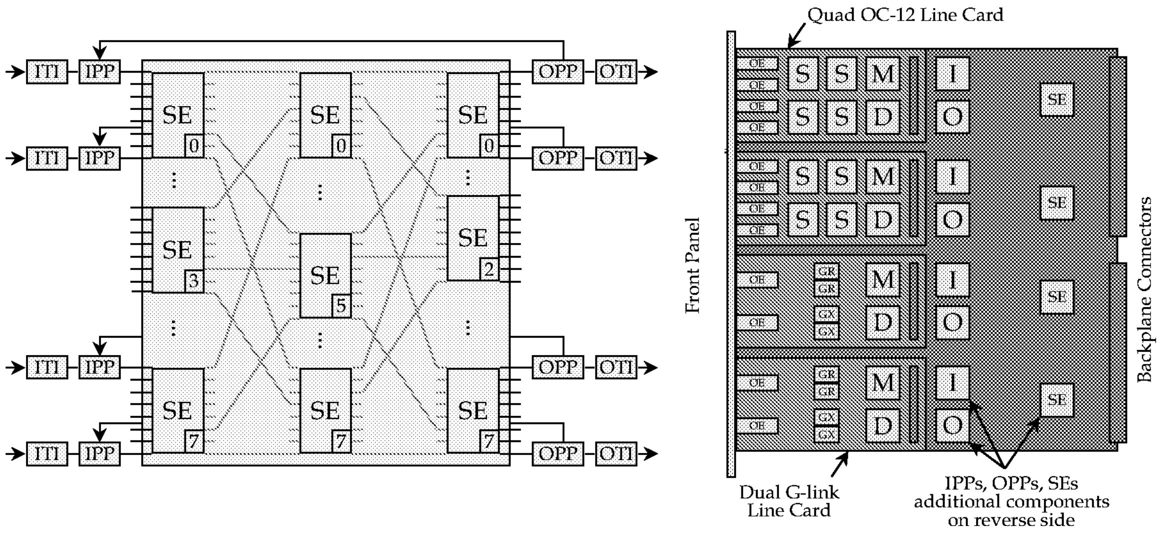

Figure 7: 160 Gb/s ATM Switch

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). Samples of these components have been obtained and evaluated in a test fixture.

- *PC Boards and Physical Design.* There are six different printed circuit board designs that are required for the burst switch prototype. The design of these circuit boards is consuming most of our efforts currently. Schematics have been completed for the BSE Datapath Board, the BSE Control Board, and the IO Board. These three are now ready for layout. The backplane board schematic is also nearly complete and should be ready for layout shortly. The high level design of the Miscellaneous Board and ATM Interface board are completed, but the schematics are not yet done. We have also designed a test board for the burst switch to verify the operation of memory interfaces between the FPGAs and off-chip SRAM components. This board is currently in layout.

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 1 shows the overall structure of the prototype and details the location of each component in the overall architecture.

- *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. The ASE is being implemented in a .35 micron ASIC process. The chip is now in fabrication and is expected back from the foundry by the end of August.
- *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 is being implemented in a .35 micron ASIC process. The chip is in fabrication and is expected back from the foundry by the end of August.
- *ATM Output Port Processor* (OPP). This chip was developed in an earlier project. The required die (fabricated in a .7 micron ASIC process) are on hand. These chips are being packaged in a ball grid array (rather than a pin grid array) to make them compatible with other components in the system. Packaged chips are expected back from the packaging house by the end of August.
- *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 currently in fabrication and will be available for testing in September.
- *Quad OC-12 Line Card.* This card multiplexes four OC-12 links onto one switch board. The FPGAs to do the required multiplexing and demultiplexing functions have been designed and simulated. The PC board layout for this board is completed and it should go to fabrication by the end of August.
- *OC-48 Line Card.* This card terminates a single OC-48 link. Although not part of the original project plan, we believe it will be feasible to include it in the project, given the recent availability of integrated OC-48 framer components.
- *PC Boards and Physical Design.* In addition to the line cards mentioned above, there are three other PC boards required to implement the WUGS-160 design. The IO Board is now being fabricated. The Backplane schematic is nearly complete and the high level design of the Center Stage Board is complete. We have also designed two additional PC boards to facilitate the testing of the various system components. One is an adapter that allows the Dual G-link and Quad OC-12 line cards to be tested in an existing 20 Gb/s ATM switch. This will allow these components to be checked out in isolation before testing them with in a WUGS-160 IO Boards. We have also designed a test backplane that will allow each of the IO Boards to be

tested in isolation (using the test backplane, an IO Board can function as a complete 20 Gb/s switch).

References

[1] Eatherton, Will. *Hardware-Based Internet Protocol Prefix Lookups.* Washington University Electrical Engineering Department, MS thesis, 5/99.

[2] Patel, Jay and Y Sliberberg. "Frequence Selective Optical Switch," U.S. Patent 5414540. See also `http://www.phys.psu.edu/faculty/PatelJ`.

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

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

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

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

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

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

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

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

[11] Turner, Jonathan S. "Terabit Burst Switching Progress Report (1/99-6/99)," Washington University Technical Report, WUCS-99-21, 8/99.

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

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