Terabit Burst Switching
Progress Report (7/00-900)

Jonathan S. Turner
jst@cs.wustl.edu

WUCS-00-28

December 19, 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.

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

Terabit Burst Switching

Progress Report (7/00-900)

Jonathan S. Turner
jst@cs.wustl.edu

This report summarizes progress on the Terabit Burst Switching Project at Washington University for the period from July 1, 2000 through September 30, 2000. As per DARPA's recent direction, we are reducing the project budget by $300,000. This reduction will prevent us from completing phase 3 of the planned burst switch prototype. Phase 3 primarily involves improvements to the logic for the Burst Processor and Burst Storage Manager to improve the efficiency with which bursts are handled. We still expect to complete both phase 1 and phase 2.

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.

- *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. In addition, we have designed a test board to enable us to prototype certain critical circuits before committing the design for the full system. The test board is currently being fabricated and will be checked out in the fourth quarter. Three boards are currently in layout. These three are the backplane, the BSE Control Board, and the IO Board. The backplane and IO Board should be completed by the end of the fourth quarter. The BSE Datapath Board is ready for layout, and will begin layout, as soon as the staff resources are available. The ATM Interface board is also ready for layout. The Miscellaneous Board schematic is to be completed following evaluation of the test board, which includes a copy of the FPGA programming circuit included on the Miscellaneous Board.
- *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. This is a change from

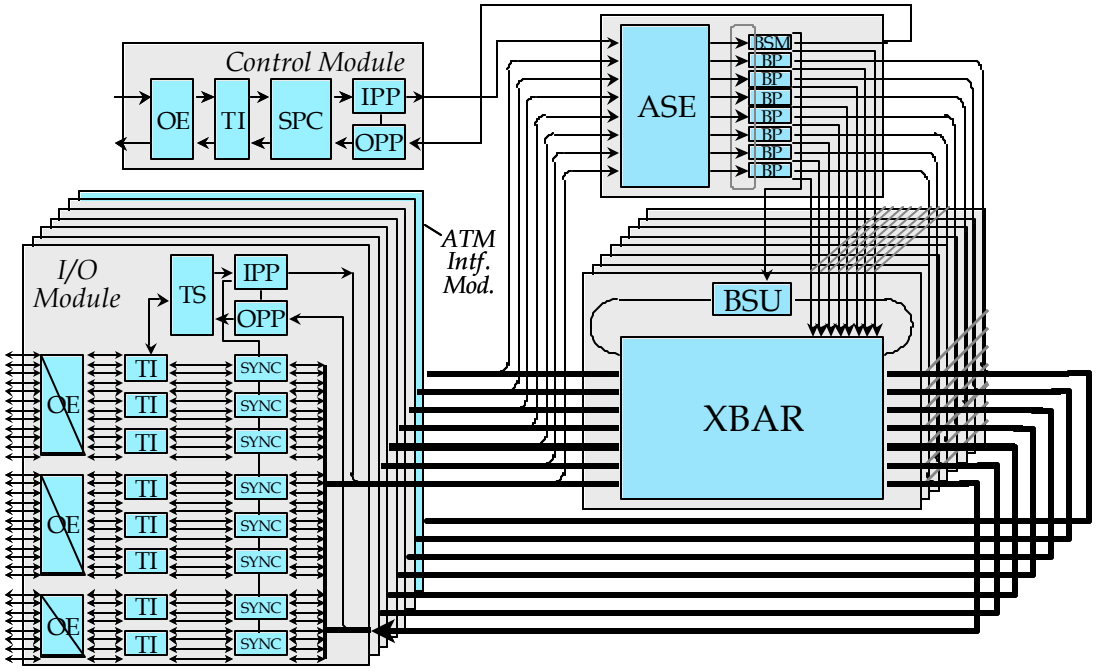

Figure 1. Prototype Burst Switch

our earlier plan, which was to use 72 smaller FPGAs. While the original approach was less expensive, simulations have shown that the fanout associated with the larger number of chips would prevent reliable full-speed operation. The VHDL for the modified design is completed, 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 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 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.

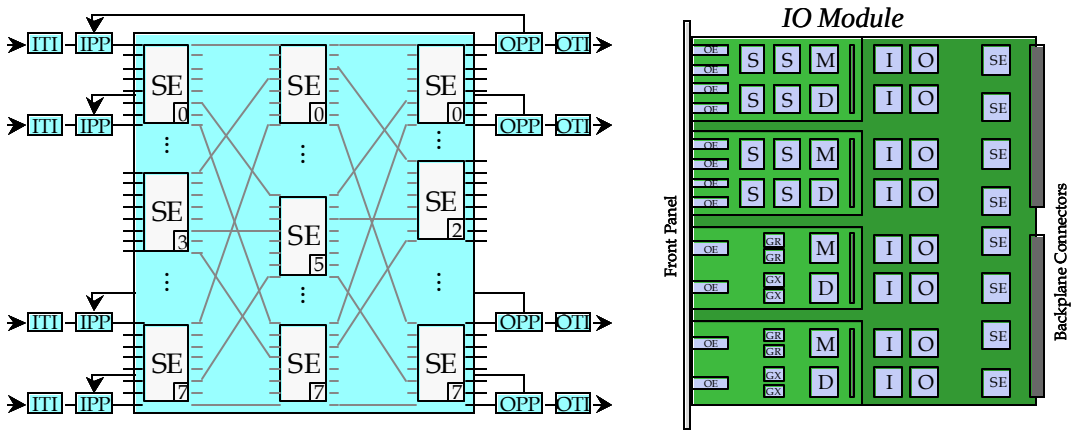

Figure 2. 160 Gb/s ATM Switch

- *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. All boards are either being fabricated now or have come back from fabrication. Specifically, the dual G-link line cards have been completed and checked out in an existing 20 Gb/s ATM switch. Two sample quad OC-12 line cards have also been fabricated and tested successfully. Photographs of the two line cards are shown in Figure 3.

  The test backplane has been completed and checked out. We have also received the first prototype IO Board from fabrication and have begun testing on that (a photograph appears in Figure 4) . There are some initial problems that have been uncovered in the early testing of this board. First, there was an error in the resistor values used for signal termination on a substantial number of signals on the board. This should not prevent further testing and can be easily corrected when

Figure 3. Dual G-link Line Card and Quad OC-12 Line Card

subsequent copies of the board are produced. The problem was traced to a CAD tool configuration problem and that problem has been corrected.

The second problem is more serious. Initial testing of the IO board showed that the Switch Element component appears to have an internal power/ground short. Our initial investigation has revealed an error in the wire-bonding data used in packaging the chip. Assuming there are no additional problems with these components, the problem can be handled without serious difficulty. However, there will be some delay, as our supply of the custom packages used for these components has been used up. We are now arranging to have additional packages fabricated and will then have some of our additional spare chips packaged, so that we can proceed with testing.

The remaining PC boards for the WUGS-160 include the Backplane and the Center Stage Board. These are both in fabrication now.

- *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 fla ws that limited the operational frequency of the original chip. The ASE is being implemented in a .35 micron ASIC process. The chip has been completed, but initial testing, indicates an internal power/ground short, apparently the result of faulty wire-bonding data (see previous item). A photograph of the switch element is shown in Figure 5.
- *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. The chip is back from fabrication and will be tested in the IO Module when revised versions of the Switch Element chip are available. A photograph of the IPP is shown in Figure 5.
- *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

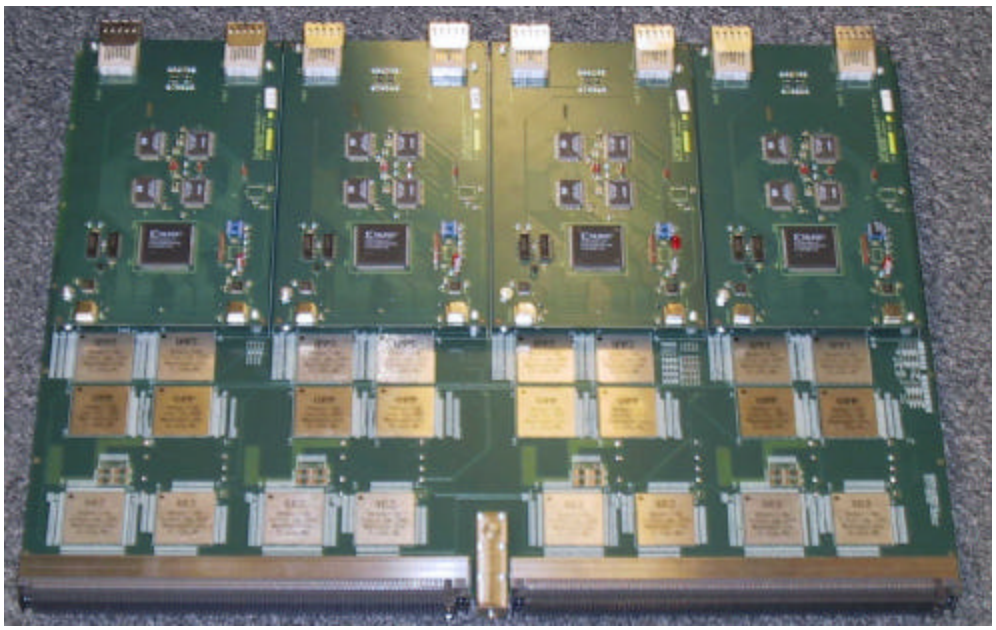

Figure 4. IO Module

compatible with other components in the system. The chips are in hand and should be tested by the end of the fourth quarter.

- *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 and tested.
- *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. This board is now complete and tested.
- *OC-48 Line Card*. This card terminates a single OC-48 link. Although not part of the original project plan, we are proceeding with a design now. The main elements of the design have been completed and a schematic should be completed by the end of the fourth quarter.

3. **Burst Switch Architecture Studies**

Recent dramatic progress in tunable lasers, appears to hold considerable promise for the design of practical burst switching systems. Our earlier studies of the feasibility of an all-optical datapath for a burst switch left us rather pessimistic about the near-term prospects. However, the progress in tunable lasers introduces new architectural options that appear fairly promising. Figure 6 shows one option for an optical crossbar that may be suitable for use in a Burst Switch Element with an all-optical datapath. Incoming WDM signals at left, are demultiplexed and converted to a new wavelength by a tunable wavelength converter. The optical signals then pass through an AWGN-type wavelength router, and from there to one of a set of parallel optical crossbars. Outputs

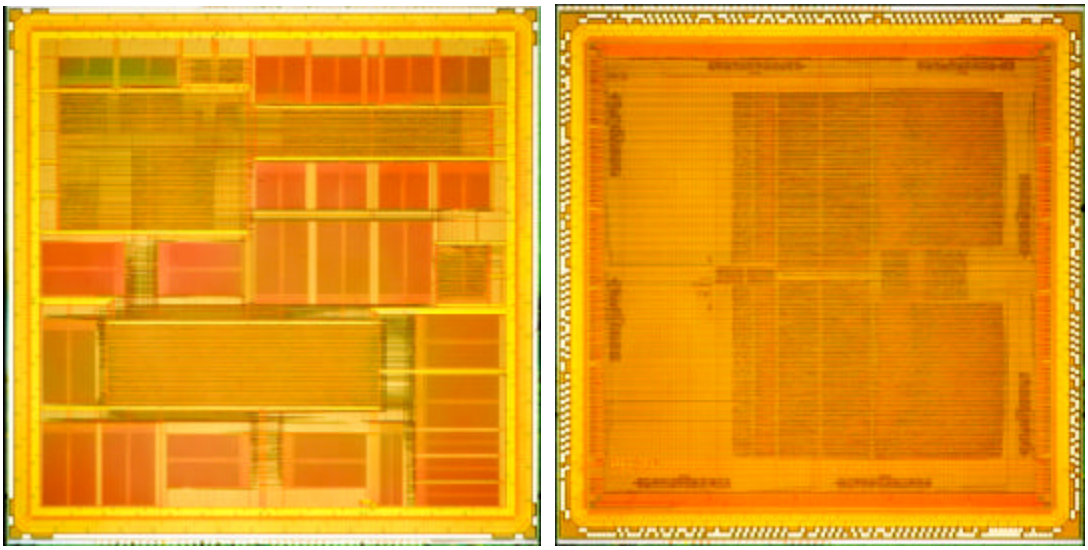

Figure 5. Input Port Processor (left) and Switch Element Chips

from the different optical crossbars are combined using passive optical couplers to produce the desired WDM output signal.

The complexity of this design is significantly smaller than other alternatives, but the design is not non-blocking. Our initial evaluation shows that for configurations of practical interest, the probability of blocking is very small, but further study is needed to determine the impact of the blocking characteristics on the burst loss probability of a Burst Switch Element that uses this approach.

REFERENCES

[EA99] Eatherton, Will. *Hardware-Based Internet Protocol Prefix Lookups*.. Washington University Electrical Engineering Department, MS thesis, 5/99.

[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.

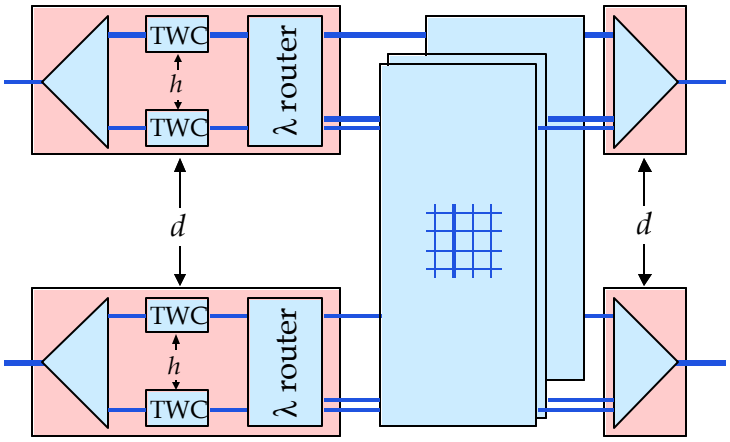

Figure 6. Wavelength Switch Using Tunable Wavelength Converters (TWC)

[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.