Terabit Burst Switching
Progress Report (10/00-3/01)

Jonathan S. Turner
jst@cs.wustl.edu

WUCS-01-09

May 1, 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 (10/00-3/01)

Jonathan S. Turner
jst@cs.wustl.edu

This report summarizes progress on the Terabit Burst Switching Project at Washington University for the period from October 1, 2000 through March 31, 2001.

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 continues to consume most of our efforts. In addition, we have designed and implemented a test board to enable us to prototype certain critical circuits before committing the design for the full system. The test board has been completed and tested, and we have successfully verified all the critical circuits. We have completed the layout of the IO Board and the backplane for the burst switch, although these have not yet been submitted for fabrication. Three other boards are currently in layout. These three are the BSE Control Board, the BSE Datapath Board and the ATM Interface Board. The first two of these should be completed by the end of the second quarter of 2001. The Miscellaneous Board schematic has been completed and will begin layout by the end of May 2001. This is the last of the circuit board designs needed for the system, so we expect things to start coming together quickly in the next six months.
- *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.

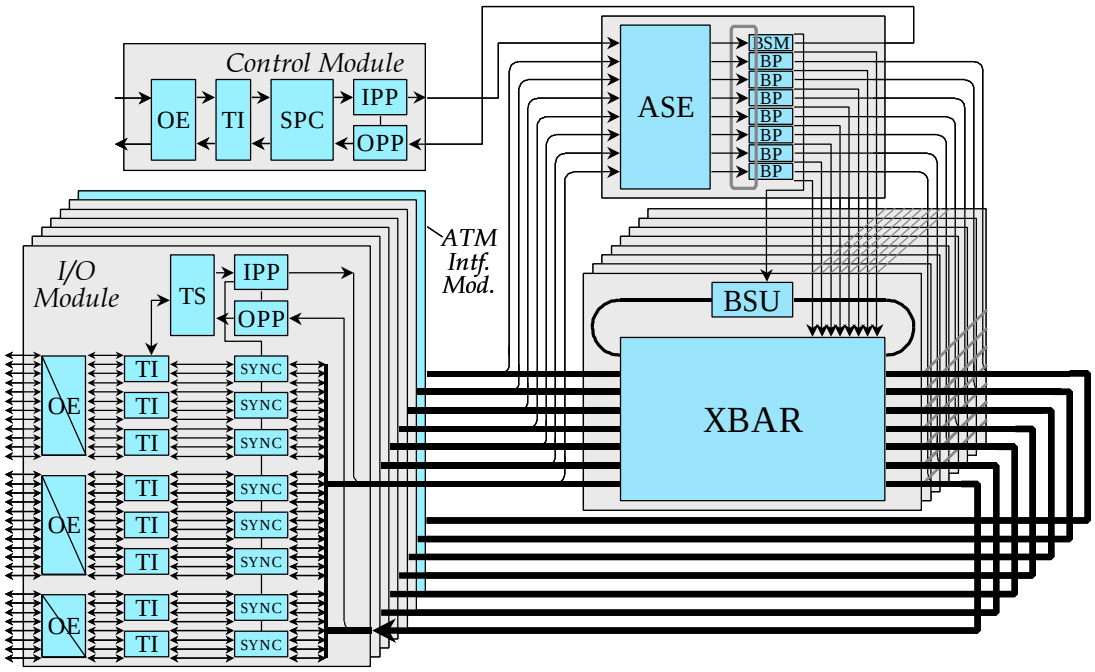

Figure 1. Prototype Burst Switch

- *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 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.

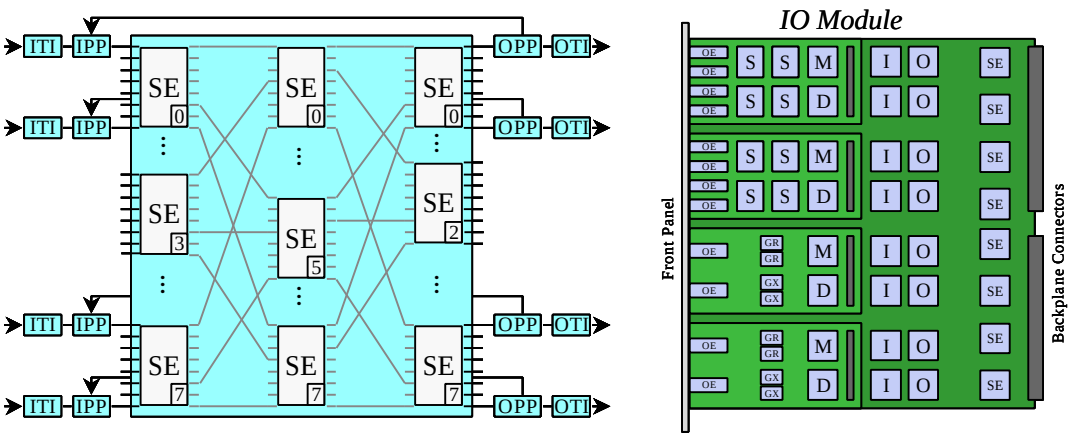

Figure 2. 160 Gb/s ATM Switch

- *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, and most of the boards either 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. The prototype IO Board for the system has also been completed and tested, using a special test backplane that was developed to allow stand-alone testing of the IO boards. A photograph of the IO board in the test backplane appears in Figure 3.

  The testing of the IO boards has uncovered several problems in both the Switch Element chips and the Input Port Processor chip. None of the problems are show stoppers, but they will place some limits on the system capabilities. These problems are detailed further below.

Figure 3. WUGS 160 IO Board Operating in Test Fixture

The remaining PC boards for the WUGS-160 include the Backplane and the Center Stage Board. The backplane is now being fabricated. The Center Stage board layout is complete, and it will be released for fabrication shortly.

- *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.

  Our testing of the Switch Element chip revealed that an error in the wire-bonding data used in packaging the chip has led to a power/ground short. We have been able to work around

the problem by severing these wire bonds, slightly reducing the power-supply current to the chip (not enough to interfere with its operation) We also discovered a design flaw that disables one of the eight input ports to the Switch Element chips. This is unfortunate, as it effectively reduces the number of usable switch ports from the intended 64 to 49. However, it does not otherwise affect the functionality of the system, and we have concluded that the best course of action at this point is to proceed with system integration and testing, using the reduced number of ports.

- *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.

  Our testing of the IPP has revealed a design flaw in the circuitry that implements the reliable multicast mechanism. While the standard cell forwarding functions work correctly, it appears we will not be able to demonstrate the reliable multicast features, without refabricating the chip. We are continuing to evaluate possible options for getting the problem corrected, but in the meantime 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.

- *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 layout of the board is completed, and we plan to send it off for fabrication by the end of the second quarter. We plan to produce eight units.

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 for these systems. However, the progress in tunable lasers introduces new architectural options that appear fairly promising. Our last report mentioned a new optical datapath design based on tunable lasers. We have subsequently developed two simpler designs that appear even more promising. The first of these is shown in Figure 4.

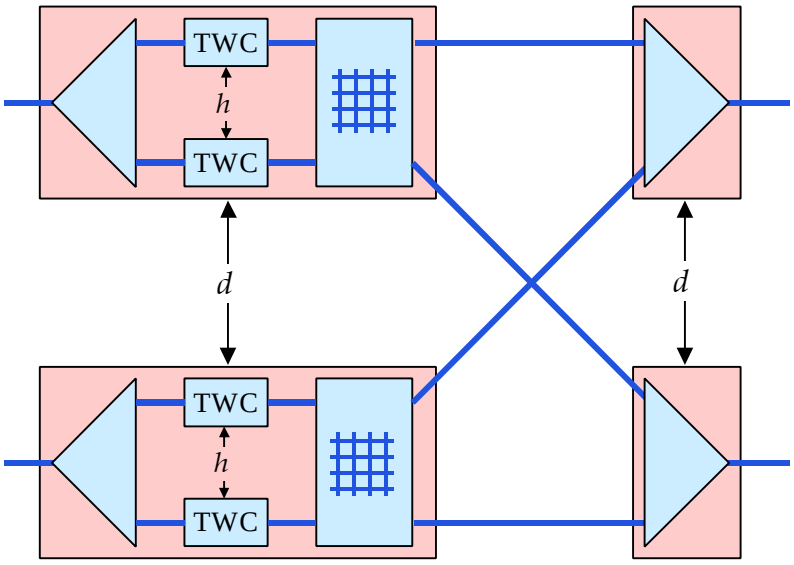

Figure 4. Wavelength Switch Using Tunable Wavelength Converters (TWC), Optical Crossbars and Passive Multipexors and Demultiplexors

The optical crossbars in each input section have h inputs and d outputs, where h is the number of wavelengths per link and d is the number of inputs and outputs of the burst switch element. Typically h will be fairly large (typical values would range from 64 to 256), while d will be relatively small (8 or 16 possibly). Fortunately, each $h \times d$ crossbar can be decomposed into a set of $d \times d$ crossbars, followed by a set of passive multiplexors, so systems of practical interest can be built using a crossbar technology capable of producing 8×8 crossbars, for example. One thing to note about this design is that the crossbars really implement both a space division switching function and a multiplexing function. If several of the input channels on a given input link are routed to the same output link, they will be multiplexed onto the connecting link between the input and output sections. SOA-based crossbar technologies can implement this combined switching and multiplexing function.

An alternative design for a burst switch element is shown in Figure 5. This design uses a passive wavelength router (AWGN-type) in place of the optical crossbars used in the first switch design. Thus, 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 5, is not a nonblocking design. That is, there may be 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 initial study 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 negligible.

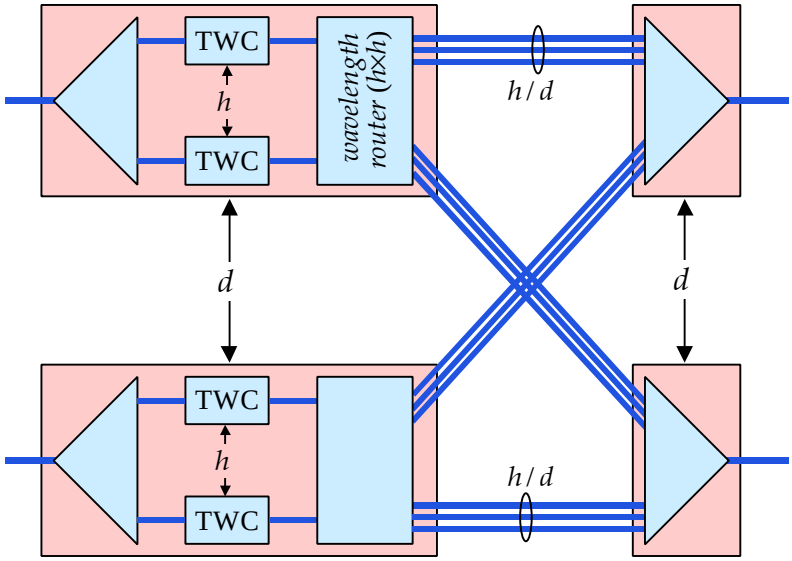

Figure 5. Wavelength Switch Using Tunable Wavelength Converters (TWC) and Passive Wavelength Routers (AWGN)

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 somewhat easier to grasp. The re-statement takes the form of a puzzle, as illustrated in Figure 6. 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 6 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, 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

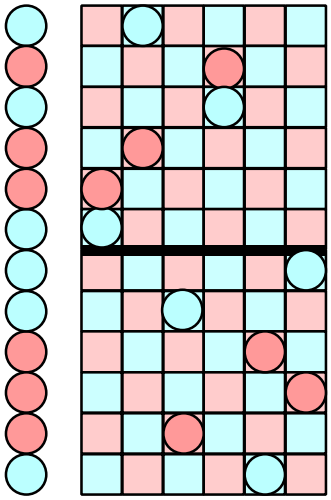

Figure 6. Puzzle Formulation of Wavelength Routing Problem

placed in each row. If the puzzle does not have a solution, then there is no way to route the channels.

For larger burst switch elements, the puzzle dimensions will also be larger. In general, we model a $d \times d$ burst switch element with h wavelengths per link, using a game board with h columns and dh rows. The number of different colors is d and there can at most h tokens of each color. The game board can be logically divided into d separate *sections*, which correspond to the different input sections of the burst switch element. This is indicated in figure 6 by the heavy line separating the top and bottom halves of the game board.

The pattern of colored squares in the game board corresponds to the pattern of the interconnections between the input sections of the burst switch element and the output sections. It is easy to see that the game board shown in Figure 6 corresponds to a poor interconnection pattern, since there is no way to solve the puzzle, if the original setup has tokens of one color in even-numbered rows, and tokens of the other color in odd-numbered rows. With this setup, at most half of the tokens can be placed on the game board, since only half of the columns can be used for tokens of each color. Game boards in which different rows have exactly the same pattern of colors are more likely to have no solution than game boards in which the color patterns of different rows are very different.

The problem of finding good interconnection patterns corresponds directly to choosing the color pattern for the game board. We would like a color pattern that guarantees that the puzzle always has a solution. For a given game board, define $C_{i,j}$ to be the set of columns with squares of color j in row i. It is easy to show that the puzzle always has a solution if for all sets R of $\leq h$ rows, and all colors j,

$$|R| \leq \left| \bigcup_{i \in R} C_{i,j} \right|$$

When selecting our design for a game board, we would like to find designs that satisfy this inequality, or at least come close. When constructing a game board, there are certain constraints we need to satisfy. In particular, within each section of the game board, each row is a rotation of the previous row. This constraint on the game board models the routing characteristics of the

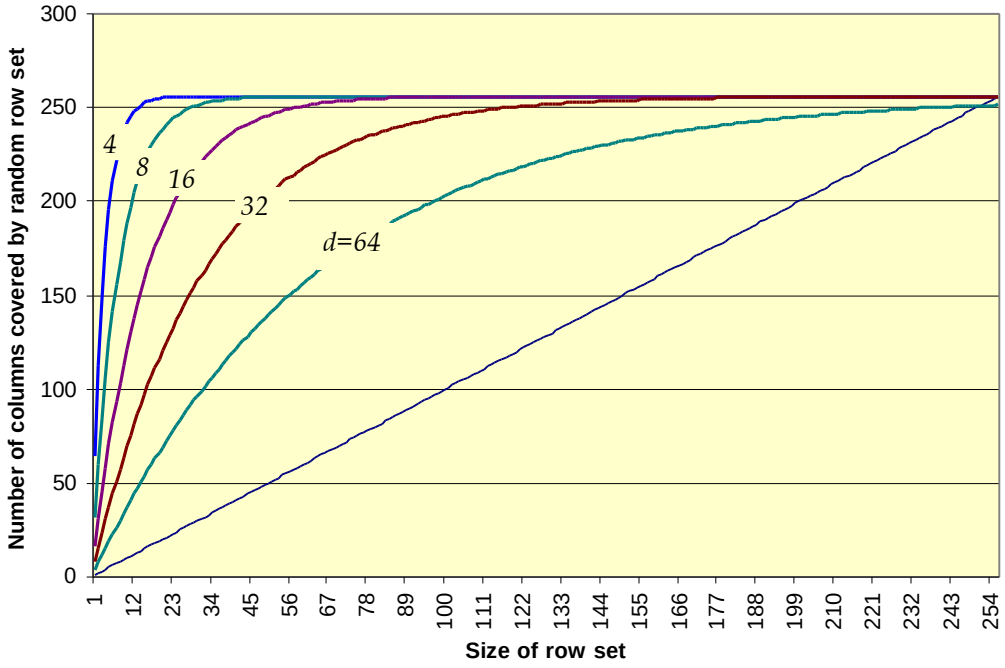

Figure 7. Number of Columns Covered by Random Row Sets of all Sizes (h=256)

AWGN wavelength router. Thus, when designing the game board, we just need to select a color pattern for one row in each section.

It's actually easy to show that no game board can be designed that has the ideal properties that we want. To see this, consider an arbitrary game board and some color (call it "blue"). There are exactly h blue squares in any column of the game board, meaning that there are $dh–h$ squares that are not blue. If we select any h rows from among the $dh–h$ rows that don't have blue squares in the given column, then we clearly cannot place h blue tokens in these rows, since none of the rows can use the given column. Similarly, if we consider any $i \leq d–1$ columns, there must be at least $(d–i)h$ rows that do not contain blue squares in any of these columns. So, we cannot place blue tokens in more than $h–i$ of these rows. In general, this implies that we cannot expect to construct a game board that will guarantee our ability to place more than $h–d+1$ tokens of the same color. Or, equivalently, we cannot expect to construct a switch of the type in Figure 5 that guarantees our ability to route more than $h–d+1$ channels to a given output.

Fortunately, in the optical switching context, h is typically much larger than d, and so 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 done a preliminary study to evaluate this possibility. Figure 6 shows one result from this study. These results are for a game board with 256 columns (corresponding to a system with 256 wavelengths per fiber) and different numbers of colors. A game board was constructed randomly, and then random row sets of varying sizes were selected, and the number of columns covered by each row set was determined (for a specfic randomly selected color. The plot shows that for the most realistic values of d (16 and less), the number of columns available for placing tokens in a row set far exceeds the minimum

needed. This provides strong evidence that blocking probabilities can be expected to be low for all but the very highest traffic loads. Explicit evaluation of blocking probabilities in systems of this type is now being carried out.

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.

[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.