

BPX 8600 Architecture and Performance

Document ID: 5284

- Introduction
- Crosspoint Matrix Concept Description
- Buffering Strategies and Blocking Performance
- The Port Speed Issue: Clos' Rule
- BCC-4
- Asymmetrical Crosspoint Switching Matrix
- Crosspoint Arbitration
- BPX Switch Performance
- Oversubscription
- Multicast
- Recommendation
- NetPro Discussion Forums – Featured Conversations
- Related Information

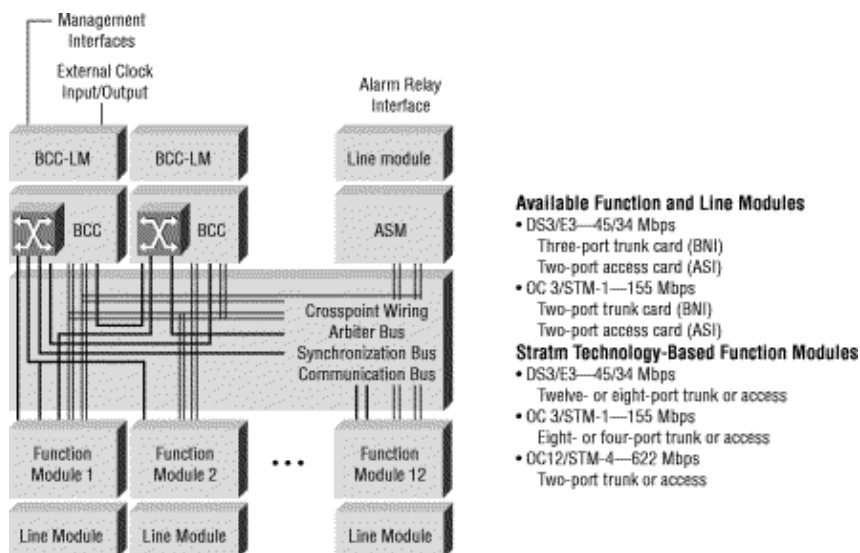
Introduction

This document provides guidelines and limitations on the use of the BCC-3 BPX Controller Card, and focuses exclusively on the switching architecture. The BPX switching architecture is based on a crosspoint switch design. The document targets any audience that has an interest to understand the BPX architecture.

The basic architectural characteristics of the BPX broadband shelf, shown in Figure 1, include a 15-slot chassis:

- 12 slots implement trunk interfaces to other BPX/IGX/MGX or ATM UNI/NNI interfaces.
- Two slots are reserved for the redundant broadband control cards (BCCs) and combine both the switching fabric and the control subsystem.
- One slot is for the alarm status monitor (ASM) card.

Figure 1. BPX Broadband Shelf



Crosspoint Matrix Concept Description

The heart of the BPX 8600 is a crosspoint matrix switching fabric, which is basically a space-switching device (a single stage that connects input with output). The crosspoint matrix is an independent subsystem on the BCC card. This section discusses the first-generation BPX crosspoint matrix.

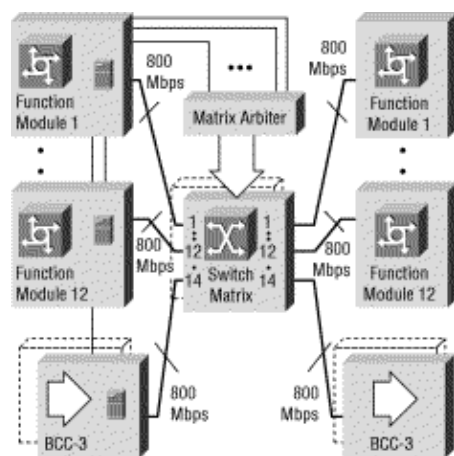
The primary function of the switch fabric is to pass traffic between interface cards. A crosspoint switching matrix performs better than bus-based products, when it operates at broadband speeds. The crosspoint switching matrix is a single-element, externally buffered matrix switch. The BCC cards available prior to BPX Release 8.4.1, such as the BCC-3, are 16 x 16 matrices. Each of the 16 crosspoint matrix ports is a full-duplex-capable, 800-Mbps link. Only 14 of the 16 ports to the crosspoint matrix, are used: two by the redundant BCCs, and the 12 that remain for the 12 function modules on the BPX broadband shelf. Each interface slot in the BPX 8600 connects to a redundant switching matrix with a redundant, full-duplex, 800 Mbps serial interface. If there is a control card failure, the redundant card can control traffic without cell loss.

An overview of the crosspoint matrix operation is shown in Figure 2.

1. Every 687.5 ns, the crosspoint matrix arbiter polls the 14 connected cards for the internal destination of the next cell to transmit.
2. The crosspoint matrix:
 - a. Checks the requests
 - b. Verifies that there are no conflicts
 - c. Configures the crosspoint to serve all requests
 - d. Grants the cards permission to send cells to the serial 800 Mbps crosspoint port
3. The cell is switched to the destination egress card.

The function modules also implement onboard arbitration functions.

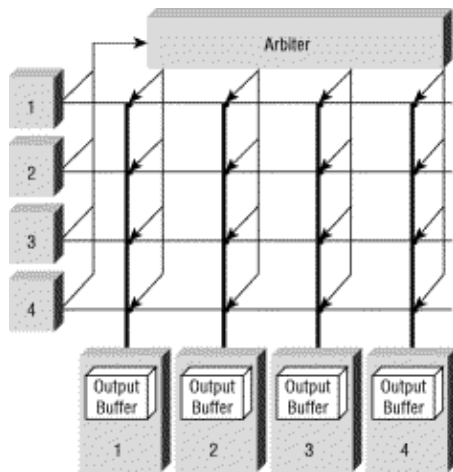
Figure 2. First Generation BPX 8600 Switch Architecture



Buffering Strategies and Blocking Performance

A switching element makes it possible for input to reach output when non-contending requests arrive.

Figure 3. Four-Port Crosspoint Architecture



Nonblocking, in ATM switch architectures, refers to the treatment of non-correlated, statistical, Bernoulli traffic (a sequence of cells with no relation to each other). The term, nonblocking, is only theoretically relevant, and it is more important to analyze how the switch architecture handles real-world traffic patterns.

The Bernoulli traffic assumption can be used for ports that have thousands of logically multiplexed user connections. You can assume that trunks between switches in large networks with many users operate this way. Thus, in the traditional trunk card design of a BPX, the Broadband Network Interface (BNI) card relies nearly exclusively on egress buffering (up to 32,000 cells can be buffered for every trunk in the egress direction).

However, on an ATM User-Network Interface (UNI), you cannot assume that user traffic is uncorrected Bernoulli traffic. The frame-oriented, higher-layer protocols that feed long frames into the convergence, adaptation, and segmentation layers, such as TCP/IP, lead to long bursts of correlated cells. These cells head toward the same destination, which is the same output port in the switch fabric. When contention occurs, it affects the size of the egress buffer, which tries to accommodate these long bursts. The buffer size is the factor that determines whether an ATM switch architecture is lossy and is blocking or nonblocking.

Therefore, the egress buffer is a critical resource in the switch and in the network. Intelligent flow control algorithms, which rely on feedback messages that accurately reflect the use of resources, must work on top of the egress buffering architectures to avoid cell loss under high load.

Therefore, ATM service switch architecture mechanisms must do these:

- Control long, correlated cell bursts on ingress ports.
- Prevent the drop of cells, other than in the most extreme network overload situations.
- Prevent cell bursts from uncontrolled flowing toward the egress buffers.

The Port Speed Issue: Clos' Rule

The blocking behavior on a switch is affected by the traffic volume and the speed of the port in and out of the crosspoint matrix. Clos' Rule, developed in 1953 by C. Clos of Bell Labs, uses three stages to convert different switching architectures into non-blocking networks. One of these stages uses the formula $k^* = 2n - 1$ to determine if the switch is nonblocking. A simple generalization turns Clos' rule to $k^* = 2k$. This means that if a switch must handle input lines of speed k , the switching stage needs to run at twice that speed to guarantee nonblocking performance.

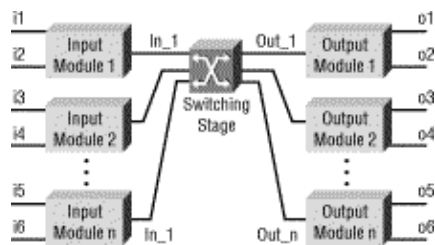
While most switch architectures do this for T3 speeds, high-density OC-3 cards push many architectures beyond the limits of nonblocking claims. In fact, OC-12 interfaces turn all ATM service switches that exist into blocking architectures. This is not the case for the BPX switch with the next-generation BCC-4. The 1.6

Gbps allocated in the egress direction exceed Clos' rule for BXM cards that use one of the two OC-12 ports. This is why OC-12 trunks, where nonblocking behavior is important, only use one OC-12 port on the BXM card.

As shown in Figure 4, typical switch architectures are not reliable to deliver low blocking when port speeds and traffic loads increase. A typical ATM switch uses an architecture where the IN_n port speed is equal to the Out_n port speed. This is typically around OC-12 speeds, which are more than 622 Mbps. For example, if ports i1, i3, and o1 are OC-12 ATM ports that run at 622 Mbps, there are two major problems:

- If ports i1 and i3 experience even brief bursts with cells that attempt to reach port o1, an architecture that relies exclusively on output buffering drops cells immediately. The Out_1 link runs at a lower speed than the aggregated traffic of the two ingress ports and cannot accommodate the cells. Since the ingress cards do not have buffers able to cope with this high-speed burst, cells are dropped. Therefore, every contention situation for an egress port leads to cell loss and requires ingress buffering. However, primitive ingress buffering implementations can cause head-of-line (HOL) blocking. The same cell loss can occur when high-density cards attempt to pass cells at equal or greater OC-12 speeds than the Out_n links can accommodate.
- The only OC-12 traffic this architecture can accommodate is simple port-to-port forwarding, such as traffic from ports i1 to o1. In this scenario, the output buffer allocated on the egress card is not used efficiently, given the speed of the links involved. All the traffic that the Out_n links forward to the card can be immediately forwarded to the outgoing OC-12 port.

Figure 4. Switch Architecture and Blocking



With the enhanced control cards (BCC-4) in Release 9.0, the BPX switch implements a switching architecture with 800 Mbps IN_n links, and 1.6 Gbps (2 x 800 Mbps) for the Out_n links with the new 16 x 32 crosspoint matrix chip. This architecture is more successful in OC-12 traffic switching. Therefore, the enhanced BCC-4 provides better service than the BCC-3 cards. This is especially true when multiple OC-12 traffic switching is required in networks where the Bernoulli traffic patterns cannot be assumed.

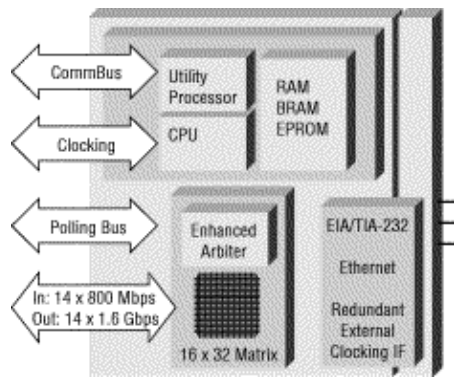
BCC-4

The BCC-4 is an enhanced control card for the BPX switch and provides the best performance of BPX technology in nodes equipped with BXM function modules.

This next-generation BCC provides enhanced processing power for general administrative node functions, but the real benefit is the fact that it provides the BPX switch with a 16 x 32 switching matrix. Some minor modifications have been made to the arbitration scheme to handle multicast traffic more efficiently.

From an architectural point of view, the BCC-4 card is similar to the BCC-3 control card that exists (see Figure 5). The CPU runs the software subsystem responsible for broadband shelf administration. An onboard Stratum 3-quality clock can be used for high-quality plesiochronous node operation or distributed as a reference, or the node can use any interface or the redundant BCC clocking port signals as a clocking reference.

Figure 5. BCC-4 Architectural Overview



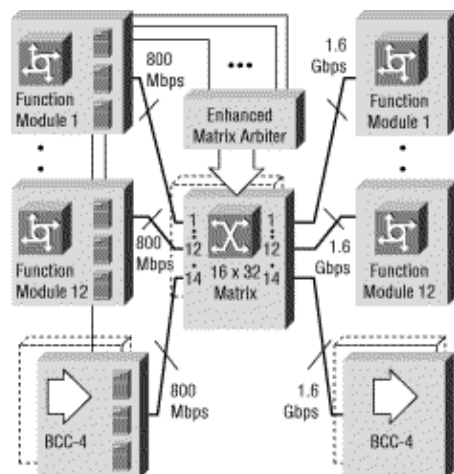
Asymmetrical Crosspoint Switching Matrix

The major innovation introduced by the BCC-4 is the asymmetrical crosspoint switching matrix. As Figure 6 illustrates, this only represents a minor modification to the architecture of the BPX switch as it is presented in Figure 2. The function modules still transmit their cells to the crosspoint matrix over an 800 Mbps link, but in receive direction. A 2 x 800 Mbps (= 1.6 Gbps) link receives cells from the crosspoint matrix. This leads to enhanced blocking behavior for high-speed links (OC-12) or high-density cards, such as the eight-port OC-3 BXM card.

In addition, switch latency is improved. The combination of advanced arbitration logic on BXM cards and the 16 x 32 crosspoint matrix switch delivers 19.2 Gbps of peak switch throughput.

Note: References [2] and [3] provide an exhaustive analysis of this switch architecture.

Figure 6. BPX Switch Architecture With BCC-4



Crosspoint Arbitration

As mentioned earlier, the BCC-4 introduces a new, more elaborate arbitration dialog with the function modules. The installed BCC-4 maximizes the the BXM module use of the 16 x 32 crosspoint switching matrix and the interworking of the advanced arbitration logic. Advanced arbitration is a key system requirement because BXM cards can be configured to oversubscribe the 800-Mbps links towards the crosspoint switching matrix. Oversubscription is a strength of the BPX architecture, because it enables cost-efficient service implementation for ATM access. For backward compatibility reasons with ATM Service Interface (ASI) and BNI cards, the BCC-4 implements full interworking with these function modules. Therefore, it fully supports a mixture of all types of function modules and control cards in one switch.

Full interworking between all cards that exist and future cards is ensured in the BPX switch to maximize investment return for customers. Therefore, four possible combinations of crosspoint type and arbitration are possible.

BPX Switch Performance

Crosspoint throughput is not limited to 58.6 percent. This result applies to the simplest kind of arbitration and to a basic, single-line, symmetric crosspoint fabric. The BPX switch uses advanced arbitration techniques and, with the BCC-4, a double-line, asymmetric crosspoint fabric. The simulation results presented here complement the theoretical analysis, because they take into account the details of the switch arbitration mechanism and show the distinct performance advantage of using the combination of advanced arbitration and double output lines.

Note: References [2] and [4] provide a theoretical analysis of crosspoints with various arbitration techniques, and References [3] and [5] give an analysis of asymmetric fabrics.

What is the nonblocking throughput of the BPX architecture with the different BCC and function module combinations? There are two commonly understood definitions of nonblocking. The simple crosspoint is classed as nonblocking because of the potential to send cells from all cards simultaneously. In addition, nonblocking is used in a more conservative sense to mean the saturation throughput.

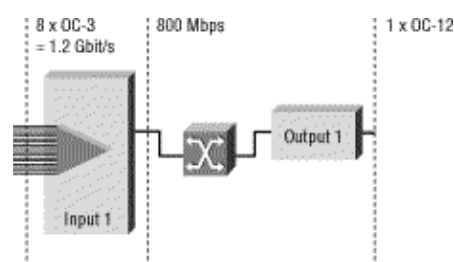
In both cases, the loading model for the switch simulations is Bernoulli traffic with traffic evenly distributed across all input ports, each port having traffic evenly distributed for each destination. The model is commonly applied in the switch performance literature. Other loading models can produce slightly different results. However, extensive simulations with a variety of traffic models indicate that the saturation limits that bind the function, shown here, are relatively independent of traffic model.

Oversubscription

Oversubscription, achieved with BXM-based function modules (two-port OC-12 and eight-port OC-3), is a benefit for service providers who offer cost-effective ATM services (see Figure 7). Large buffers and elaborate arbitration schemes are required in order to support oversubscription without cell loss.

Service nodes behave in a fully nonblocking way for trunks, which is why BXM function modules support only one OC-12 trunk or four OC-3 trunks. However, an access switch must provide aggregation of user traffic toward the switching matrix. It is statistically unlikely for all user ports on one multiport function module to show peak activity, but the ingress buffering on the BXM card is able to cope with these extremely rare activity peaks.

Figure 7. Supporting



Multicast

The enhanced 16 x 32 crosspoint matrix is optimally suited to accommodate the characteristics of multicast

traffic distribution, which is always output biased and creates more traffic in the egress direction of the crosspoint matrix. It is the crosspoint arbiter that replicates the cells from one input port to a number of output ports. In a second pass, BXM cards can implement logical multicasting to replicate one cell that arrives from the crosspoint matrix to different virtual connections (typically different VPs).

Recommendation

In network applications where multiple OC-12 trunks or ports are used to aggregate traffic to only one or two destination OC-12 ports, the recommendation is to use BCC-4 control cards. This non-Bernoulli traffic pattern is best served with the 16 x 32 crosspoint matrix of the BCC-4 Control Card.

NetPro Discussion Forums – Featured Conversations

Networking Professionals Connection is a forum for networking professionals to share questions, suggestions, and information about networking solutions, products, and technologies. The featured links are some of the most recent conversations available in this technology.

NetPro Discussion Forums – Featured Conversations for WAN Switching
Network Infrastructure: WAN Routing and Switching

Related Information

- [Cisco WAN Switching Solutions](#)
- [Guide to New Names and Colors for WAN Switching Products](#)
- [Downloads – WAN Switching Software](#)
- [Technical Support & Documentation – Cisco Systems](#)

[Contacts & Feedback](#) | [Help](#) | [Site Map](#)

© 2008 – 2009 Cisco Systems, Inc. All rights reserved. [Terms & Conditions](#) | [Privacy Statement](#) | [Cookie Policy](#) | [Trademarks of Cisco Systems, Inc.](#)

Updated: Apr 17, 2009

Document ID: 5284
