PM7326

TRAFFIC MANAGEMENT AND SWITCHING USING THE VORTEX CHIP SET:

S/UNI-APEX

TECHNICAL OVERVIEW

RELEASED

ISSUE 2: OCTOBER 2000

REVISION HISTORY

Issue No.	Issue Date	Details of Change
2	October 2000	Added Additional Reading section. Corrected minor typos as indicated by change bars.
1	August 1999	Document created.

CONTENTS

1 REQUIRED READINGS 1

2 PURPOSE AND SCOPE OF THIS DOCUMENT 2

3 OVERVIEW 3

4 THE PARALLEL BUS SPECIFICATIONS 6

5 REQUIREMENTS OF A SINGLE STAGE SWITCH/MUX 8

5.1 HOW THE INDIVIDUAL CHANNELS REACH THE S/UNI-APEX....................8

5.2 SWITCHING OVERVIEW FOR ATM, FRAME, AND PACKET TRAFFIC....................10

5.3 LINK LAYER FRAMING, INTERWORKING, AND SAR FUNCTIONS 11
5.3.1 ATM 11
5.3.2 FRAME RELAY....................12
5.3.3 IP OVER PPP....................15

5.4 INGRESS ADDRESS RESOLUTION OR FLOW CLASSIFICATION 15

5.5 INGRESS POLICING....................17

5.6 INGRESS CONGESTION MANAGEMENT....................17

5.7 EGRESS QUEUING, CLASS SCHEDULING, AND PORT SCHEDULING....................21
5.7.1 EGRESS CELL SCHEDULING....................22
5.7.2 EGRESS PACKET SCHEDULING....................23

5.8 EGRESS TRAFFIC SHAPING 24

6 SPECIAL TOPICS 26

6.1 THE S/UNI-APEX MICROPROCESSOR PORT....................26

6.2 MULTICAST SUPPORT IN THE S/UNI-APEX 26

6.3 EMBEDDED INTER-DEVICE COMMUNICATIONS CHANNEL 27

6.4 ATM OAM HANDLING USING THE S/UNI-ATLAS AND S/UNI-APEX 27

6.5 ATM SIGNALING AND IP ICMP PROCESSING....................30

6.6 INGRESS ADDRESS RESOLUTION USING THE S/UNI-APEX....................31

6.7 INGRESS ADDRESS RESOLUTION USING THE S/UNI-ATLAS 31

6.8 EGRESS CELL IDENTIFIER, SWITCH TAG, AND VCI/VPI MAPPING....................32

6.9 HANDLING AAL2 TRAFFIC 34

6.10 OC-12 (622 MBPS) ARCHITECTURES 35

7 GLOSSARY....................38

FIGURES

FIGURE 1 - TYPICAL APPLICATION OF THE VORTEX CHIP SET 4

FIGURE 2 - 3G WIRELESS BASE STATION 9

FIGURE 3 - HIGH DENSITY FRAME RELAY PORT CARD 10

FIGURE 4 - SIMPLE FRAME RELAY MULTIPLEXER WITHOUT ATM INTERWORKING 14

FIGURE 5 - 2048 CHANNEL IP EDGE ROUTER...15

FIGURE 6 - EPD/PPD THRESHOLD HIERACHY 20

FIGURE 7 - EGRESS QUEUING STRUCTURES 22

FIGURE 8 - TRAFFIC SHAPING ON THE WAN PORT 25

FIGURE 9 - THE EMBEDDED CONTROL CHANNEL..27

FIGURE 10 - LOOP PORT TO WAN PORT USER AND RDI CELLS 29

FIGURE 11 - WAN PORT TO LOOP PORT TO USER AND RDI CELLS 30

FIGURE 12 - TRANSMIT CELL FORMAT OPTIONS 33

FIGURE 13 - 3G WIRELESS BASE STATION 34

FIGURE 14 - OC-12 MULTIPLEXER ARCHITECTURE 36

FIGURE 15 - OC-12 SWITCH PORT ARCHITECTURE...37

1 REQUIRED READINGS

Before reading this document we encourage the reader to review a short paper entitled *VORTEX Chip Set Introduction* (document number PMC-990712). This introductory paper highlights the features and applications of the VORTEX chip set and will greatly assist the reader in understanding the system context in which the S/UNI-APEX typically functions.

The reader is also encouraged to review a related document titled *S/UNI-VORTEX and S/UNI-DUPLEX Technical Overview* (document number PMC-981025). The S/UNI-VORTEX and S/UNI-DUPLEX Technical Overview describes how these system interconnect devices connect up to 2048 ports or channels to the S/UNI-APEX and S/UNI-ATLAS. It will likely be much easier to understand the S/UNI-APEX technical overview if the S/UNI-VORTEX and S/UNI-DUPLEX technical overview is read first.

Additional Reading

Readers interested in detailed device level information about the S/UNI-APEX should read the *S/UNI-APEX Data Sheet* (document number PMC-981224). This document contains in-depth information about all functions of the S/UNI-APEX as well as register-level configuration information, electrical characteristics and other details.

The *S/UNI-APEX Hardware Programmer's Guide* (document number PMC-991454) provides practical information about the configuration and operation of the S/UNI-APEX including guidelines for determining specific configuration parameters.

All documentation for the S/UNI-APEX and related devices is available from the PMC-Sierra web site: www.pmc-sierra.com .

2 PURPOSE AND SCOPE OF THIS DOCUMENT

The VORTEX chip set consists of four devices:

PM 7350 S/UNI-DUPLEX – Dual Serial Link, PHY Multiplexer

PM 7351 S/UNI-VORTEX – Octal Serial Link Multiplexer

PM 7324 S/UNI-ATLAS – ATM Layer Device

PM 7326 S/UNI-APEX – ATM/Packet Traffic Manager and Switch

The VORTEX chip set creates an integrated device family designed to satisfy requirements of the newest and fastest growing network access applications:

- Digital Subscriber Line Access Multiplexers -- DSLAMs.
- Third generation digital wireless base stations and base station controllers.
- Multi-service access multiplexers.

This Technical Overview describes how the S/UNI-APEX fulfils the switching and traffic management requirements across this wide range of communication equipment. The document answers the following questions:

- What function does the S/UNI-APEX fulfill in the overall system architecture of these applications?
- What are the overall capabilities and limitations of the S/UNI-APEX?
- How do the other components of the VORTEX chip set inter-work with the S/UNI-APEX?

This document is a companion of, but subordinate to, the *S/UNI-APEX Data Sheet* (document number PMC-981224). If there appear to be differences, contradictions, or omissions in this Technical Overview the reader is advised that the data sheets take precedence.

3 OVERVIEW

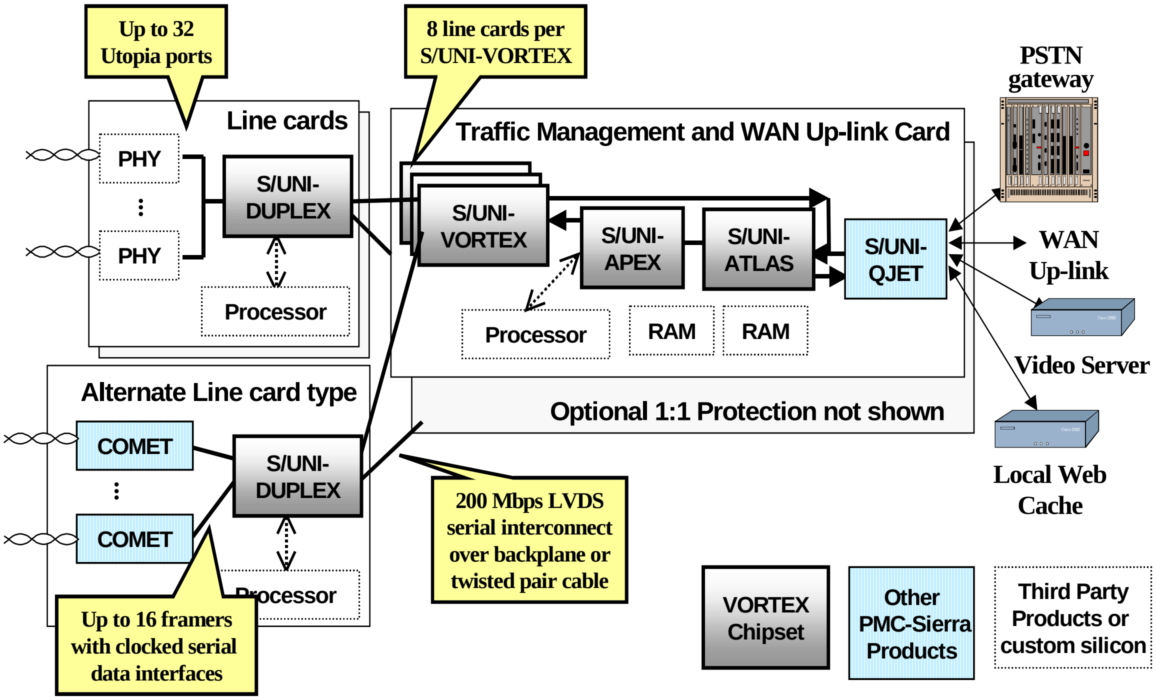
The S/UNI APEX is an ATM and packet capable traffic management and switching device providing the following features:

- Any port to any port cell and packet switching for up to 64K active connections.
- Highly configurable, packet aware management of up to a 16 Mbyte data buffer for effective and fair congestion management under high traffic loads.
- Hierarchical scheduling across three physical bus interfaces capable of supporting 2048 line ports, 4 WAN ports, and a high speed microprocessor port respectively. Each port provides 4 classes of service.
- Per connection cell or whole packet queuing for improved fairness of the bandwidth allocated among connections within a single class of service.
- Traffic shaping on the WAN ports.

As an integral member of the VORTEX chip set the S/UNI-APEX is ideally suited for applications such as multi-service access concentrators (e.g. DSLAMS), edge switches/routers, and 3rd generation wireless base stations and base station controllers.

These applications share a common problem -- efficiently multiplexing a large number of lower speed ports or channels into a small number of high speed ports. Typically, a number of line-side ports (modems, ATM PHYs, or radio channels) are terminated on each line or radio card. Numerous line cards are then slotted into one or more shelves and backplane traces or inter-shelf cables are used to connect the line cards to a centralized (often 1:1 protected) common card, hereafter referred to as the core card. The core card normally includes one or more high speed WAN up-link ports that transport traffic to and from a high speed broadband network. A block diagram of a 1:1 redundant access multiplexer is shown in Figure 1.

Figure 1 - Typical Application of the VORTEX Chip Set


Up to 32 Utopia ports
8 line cards per S/UNI-VORTEX
PSTN gateway
Line cards
PHY
PHY
S/UNI-DUPLEX
Processor
Traffic Management and WAN Up-link Card
S/UNI-VORTEX
S/UNI-APEX
S/UNI-ATLAS
S/UNI-QJET
Processor
RAM
RAM
WAN Up-link
Video Server
Local Web Cache
Alternate Line card type
COMET
COMET
S/UNI-DUPLEX
Processor
Optional 1:1 Protection not shown
200 Mbps LVDS serial interconnect over backplane or twisted pair cable
Up to 16 framers with clocked serial data interfaces
VORTEX Chipset
Other PMC-Sierra Products
Third Party Products or custom silicon

In this type of equipment the majority (perhaps all) user traffic goes from WAN port to line port, or from line port to WAN port. Although the individual ports on the line cards are often relatively low speed interfaces such as T1, E1, or xDSL, there may be many ports per line card and many line cards per system, resulting in hundreds or even thousands of lines terminating on a single WAN up-link. In the upstream direction (from line card to WAN up-link), the equipment must have capacity to buffer and intelligently manage bursts of upstream traffic simultaneously from numerous line cards.

In the downstream direction the equipment must handle a similar issue, the “big pipe feeding little pipe” problem. When a large burst of traffic destined for a single line port is received at the high speed WAN port it must be buffered and managed as it queues up waiting for the much lower speed line port to clear.

The S/UNI-APEX can support aggregate traffic of 700 Mbps ingress (data written into its data buffer from any port), and 700 Mbps egress (data read out of its data buffer to any port). When supporting an OC-12 (622 Mbps) WAN port the S/UNI-APEX is capable of buffering and multiplexing/demultiplexing a single direction of traffic. In OC-3 designs it is capable of handling full duplex traffic of up to two OC-3’s with a small amount of speed-up. OC-3 and multiple DS-3 type WAN port configurations are readily supported with significant speed-up provided by the S/UNI-APEX.

When evaluating system switching architectures, one must consider the full set of system level requirements:

- Congestion management, and scheduling: i.e. implementing the data-path.
- Inter-processor communication: i.e. supporting an inter-card control channel.
- Allowing for co-processing (i.e. add/drop of data, signaling, and OAM traffic to a processing engine).

The remainder of this Technical Overview discusses these requirements and describes how each is satisfied by the S/UNI-APEX for cell (fixed length) and packet (arbitrary length) traffic.

4 THE PARALLEL BUS SPECIFICATIONS

Throughout this document references are made to three related bus interface specifications.

Utopia Level 2 – The industry standard ATM Forum PHY to ATM layer bus specification. Also referred to as Utopia L2. It is available from the ATM Forum web site at www.atmforum.com.

SCI-PHY Level 2 – Also referred to as SCI-PHY L2, it is a backward compatible extension to the Utopia bus specification to allow the following enhancements:

- 32 PHY maximum instead of 31.
- Extended cells: SCI-PHY supports user defined bytes prepended or postpended to the standard 53 byte cells allowed by Utopia. This particularly useful when switching tags are to be attached to the cell by an ATM layer device such as the S/UNI-ATLAS. SCI-PHY allows cells to be up to 64 bytes long.

Any-PHY – PMC-Sierra's extension to the Utopia bus specification to allow the following enhancements:

- Addressing for an unlimited number of logical PHYs.
- Fixed cell or arbitrary length packet transfers. Only cell mode transfers are used by the VORTEX chip set, including the S/UNI-APEX.
- Extended cells: This includes optional prepend or postpend cell extensions plus the additional overhead bytes used for in-band PHY addressing and selection.
- In the transmit direction, separation of PHY status polling (using bus pins) from PHY selection during start of cell or packet transfer (using in-band addressing).
- In the receive direction, cell overhead bytes can be used to identify the source PHY of the cell.
- Relaxed decode response time to simplify device status polling circuitry.

A comparison of the bus signals for these first three bus types is shown in Table 1.

POS-PHY – A predecessor to the Any-PHY bus operating in packet mode, but with a limited number of logical PHYs.

Table 1 - Comparison of Bus Signals

UTOPIA Level 2 Bus Slave	SCI-PHY Level 2 Bus Slave	Any-PHY Bus Slave	UTOPIA Level 2 Bus Slave	SCI-PHY Level 2 Bus Slave	Any-PHY Bus Slave
TxClk	TFCLK	TCLK	RxClk	RFCLK	RCLK
TxEnb*	TWRENB	TENB	RxEnb*	RWRENB	RENB
TxAddr[4:0]	TADDR[4:0]	TADR[4:0]	RxAddr[4:0]	RADDR[4:0]	RADR[4:0]
n/a	TAVALID	TADR[5]	n/a	RAVALID	RADR[5]
TxData[15:0]	TDAT[15:0]	TDAT[15:0]	RxData[15:0]	RDAT[15:0]	RDAT[15:0]
TxPrty	TPRTY	TPRTY	RxPrty	RPRTY	RPRTY
TxClav	TCA	TPA	RxClav	RCA	RPA
TxSOC	TSOC	TSOP	RxSOC	RSOC	RSOP
n/a	n/a	TSX	n/a	n/a	RSX

5 REQUIREMENTS OF A SINGLE STAGE SWITCH/MUX

ATM, frame relay, and packet (e.g. IP) multiplexers and switches share many common features and implementation requirements. In the following discussion we identify common processing requirements for all three data types and discuss how each requirement is handled by the S/UNI-APEX, the S/UNI-ATLAS, or corresponding support circuitry.

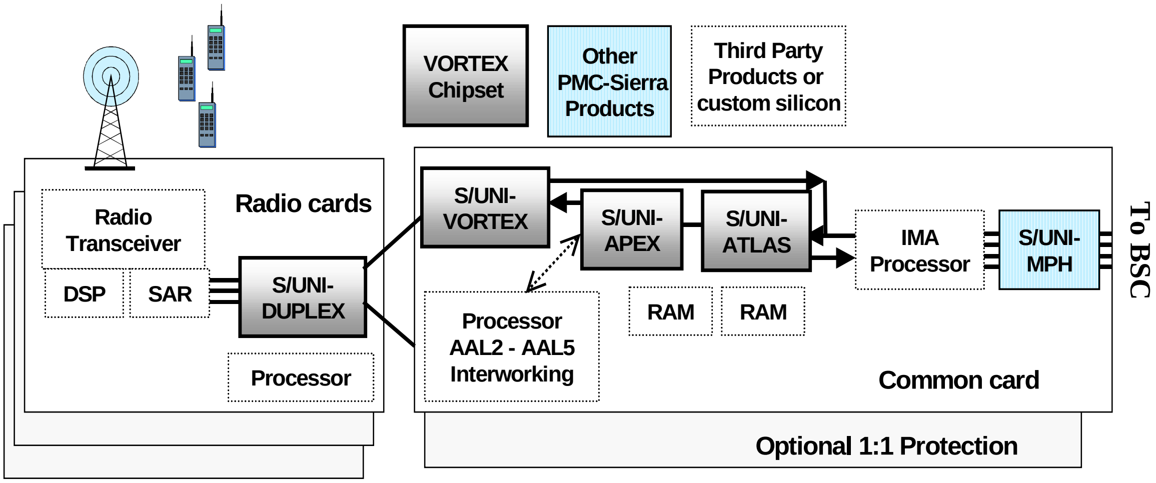
5.1 How the Individual Channels Reach the S/UNI-APEX

The S/UNI-APEX does not *directly* schedule traffic into fixed rate Time Division Multiplexed (TDM) channels such as T1, DS-3, etc.. The S/UNI-APEX assumes that between it and the transmission system are physical layer, transmission convergence (TC) layer, and link layer devices that:

- Frame the transmission signal and identify the data bits.
- Demultiplex the individual channels (if there are more than one) and provide a separate data stream for each channel.
- ATM cell delineation or HDLC framing, checking for transmission errors and discarding any data units that may have had their routing information corrupted, and inserting/extracting cells or packets to/from each channel.
- Insert/remove Idle cells or idle flags when there is no data.
- If the data arrives in variable length packets, reassemble/segment the packet into ATM sized cells (typically using the AAL5 protocol).
- Provide indications to the S/UNI-APEX (via bus signaling) when it is possible to insert/extract additional data to/from the individual channels.

The architecture shown in Figure 1 implements the data path of a high fan-in multiplexer or switch port in which each physical line is a single channel carrying ATM data streams. The architecture shown in Figure 2 implements the data path of a third generation (3G) wireless base station in which each radio channel is a single channel carrying ATM data streams.

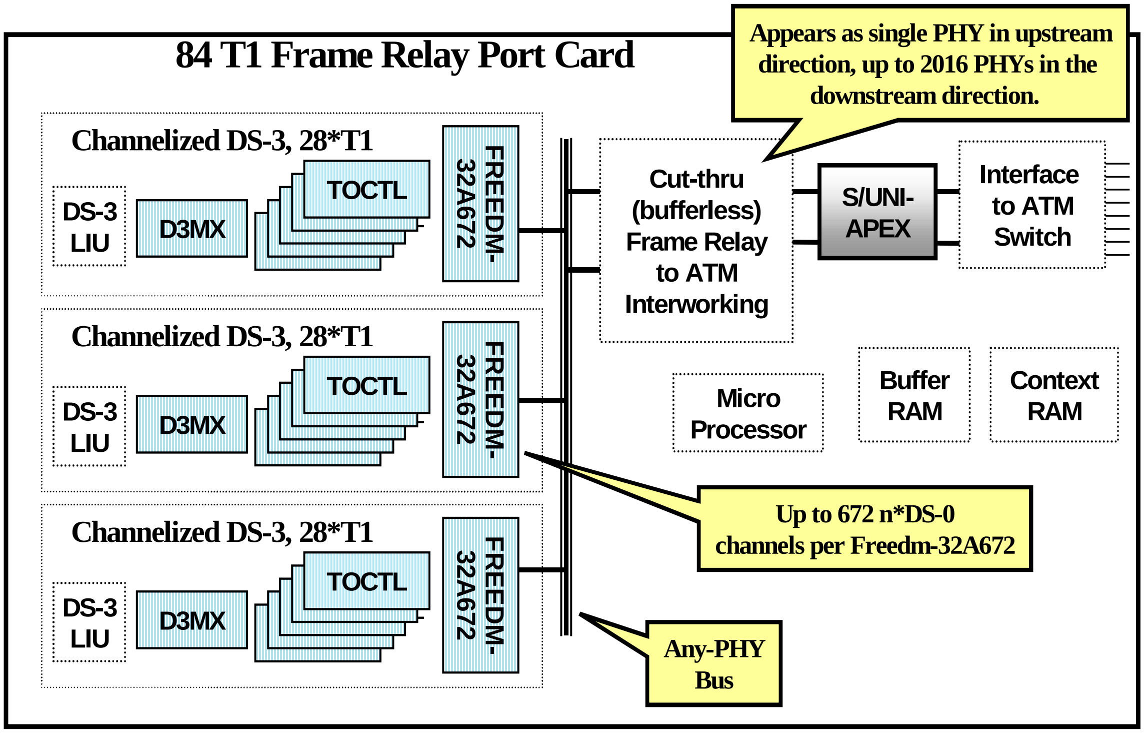
Figure 2 - 3G Wireless Base Station



In both of these applications each PHY or radio channel is connected to the S/UNI-APEX via the S/UNI-VORTEX, which acts as a *PHY Proxy*. In the upstream direction (toward the WAN) the S/UNI-VORTEX appears as a single PHY by marking each cell with its source PHY ID and then multiplexing all upstream traffic into a single stream. In the downstream direction the S/UNI-VORTEX functions as a multi-port PHY slave. Each S/UNI-VORTEX provides individual transmit and receive bus control signals for up to 256 ports and 8 control channels. If the reader is not familiar with the functionality of the S/UNI-VORTEX and S/UNI-DUPLEX devices it is highly recommended to pause now and review the *S/UNI-VORTEX and S/UNI-DUPLEX Technical Overview* (document number PMC-981025).

There are also system configurations in which the S/UNI-APEX processes multiplexed traffic carried over DS-3s, OC-3, etc.. Figure 3 shows a high density frame relay card that terminates 3 DS-3 signals and brings up to 2016 individual n*DS-0 channels to the S/UNI-APEX. In this configuration the HDLC processing performed by PMC-Sierra’s Freedm-32A672 device decouples each individual channel’s bit rate from the actual data rate on that channel. The AAL5 SAR in Figure 3 functions as a single PHY (with tagged cells) in the upstream direction and a 2016 multi-PHY in the downstream direction, and hence is functionally analogous to the S/UNI-VORTEX.

Figure 3 - High Density Frame Relay Port Card



5.2 Switching Overview for ATM, Frame, and Packet Traffic

Regardless of the length of the fundamental data unit being handled, the following steps are required to route the data through the switching equipment from incoming port to outgoing port. Each of these steps is discussed more fully in later sections.

- TC layer and link layer processing: ATM cell delineation or HDLC framing, transmission error detection, idle cell or fill pattern generation/removal, and segmentation/reassembly (packet only). Link layer processing is discussed in Section 5.3.

- Address resolution lookup (ARL) must be performed to uniquely identify the individual virtual connections (VCs) to be switched. Typically there will be a device between the S/UNI-APEX and the Link Layer processing that can perform the address resolution function, although the S/UNI-APEX can provide a limited ARL function internally. Address resolution typically combines the physical port number on which the data unit arrives with the value of an address field within the data unit and generates a “switch tag” or *Ingress Connection Identifier* (ICI). The ICI is used by the S/UNI-APEX to efficiently reference the data structure that defines how to handle the data stream. ARL is discussed further in Section 5.4

- Perform ingress *policing* on each VC to measure traffic and mark data units that exceed the user's expected service rate. Policing is discussed further in Section 5.5.

- Perform *congestion management* on the ingress traffic to ensure heavy traffic volumes impact low priority traffic first. Congestion management is discussed further in Section 5.6.

- Buffer data units until their destination port is ready to accept them, and then schedule the egress data units onto the appropriate bus (line-side bus, WAN-side bus, or microprocessor/control bus). Scheduling typically involves sophisticated algorithms that support various levels of QoS. Egress traffic management is discussed in Section 0.

- Optionally shape traffic (alter its emitted rate) to meet a specified traffic pattern. Shaping is discussed in Section 5.8.

Besides these common functions, there are certain unique traffic handling requirements that warrant special handling. These in include Operations, Administration, and Maintenance (OAM) traffic, signaling, and AAL2. These are discussed in Sections 6.2, 6.5, and 6.2 respectively.

5.3 Link Layer Framing, Interworking, and SAR Functions

Although it is true that the S/UNI-APEX's fundamental switching unit is a fixed length cell, the S/UNI-APEX is also fully "packet aware", meaning, it can treat groups of cells as a single packet. In order for the S/UNI-APEX to function as a packet aware device, packet data unit boundaries must have been externally identified using the AAL5 convention of marking the last cell of a packet with an End of Message (EOM) bit in the Payload Type Identifier (PTI) field of the standard ATM header.

The remainder of this section discusses the Link Layer processing required for the various traffic types entering the switch on the line or WAN ports. Link Layer processing must be performed before the traffic reaches the S/UNI-APEX.

5.3.1 ATM

Because of its simplicity, ATM TC and Link Layer processing is typically integrated into the Layer 1 device (i.e. the framer). Alternatively, the cell delineation, idle cell extraction/generation, etc. can be handled by an external device such as PMC-Sierra's S/UNI-DUPLEX device. The VCI/VPI fields within the ATM cells identify the virtual connection that each cell is associated with. Therefore, ATM traffic requires no special treatment to be compatible with the S/UNI-ATLAS and S/UNI-APEX devices.

5.3.2 Frame Relay

With respect to frame relay services, the first question to ask is “Do I need a pure frame relay switch, or is this to be an ATM switch with frame relay ports?” The difference is subtle but important since it determines the degree of interworking performed on the frame relay service as it enters the switch.

Frame Relay to ATM Interworking:

For example, Figure 3 shows a frame relay port card where the frame relay to ATM interworking is located between the PMC-Sierra Freeedm-32A672 HDLC controller and the S/UNI-APEX. In this example Link Layer processing of the upstream traffic (from the DS-3s to the ATM switch) is handled as follows:

1. DS-3 LIU converts the DS-3 signal from analog to digital
2. D3MX frames the DS-3 and produces 28 independent T1 (1.544 Mbps) streams
3. Each T1 framer (8 T1 framers per TOCTL device) processes a T1 stream, removes the overhead, and provides the Freedm-32A672 with a serial data stream.
4. The Freedm-32A672, based on user programming, collects individual 64 kbps channels into n*DS-0 streams. The Freedm performs HDLC processing on each stream, extracts the packet data units, performs the HDLC CRC-32 calculation, and presents the data to the AAL5 SAR in user programmable sized data blocks. If, at the end of the packet, the Freedm detects a CRC-32 error it will notify the SAR/interworking device coincident with the end of packet indication. See the *Freedm-32A672 Data Sheet* for a detailed description of its functionality.
5. The AAL5 SAR is a bus master to the Freedm and a bus slave to the S/UNI-APEX. In the upstream (toward the switch) direction each Freedm appears as a single physical PHY. This is also true of the AAL5 SAR – it appears as a single PHY device to the S/UNI-APEX master.
6. The AAL5 SAR and interworking functions are a customer provided. The SAR/interworking device takes each data block from the Freedm, performs the frame relay interworking, creates the 48 byte ATM user data blocks, adds the 5 byte ATM header field and 16 bit ICI, and passes the resultant cells to the S/UNI-APEX. AAL5 encapsulation and interworking between frame relay and ATM, as defined by Frame Relay Forum specification FRF.5 and FRF.8 is the responsibility of the SAR/interworking device. The frame relay channel ID and DLCI address is mapped to a 16 bit ICI value used by the S/UNI-APEX as a switch tag. The interworking function can be performed without buffering more than a few cells per HDLC channel.
7. At the end of the packet the Freedm will have indicated whenever an errored CRC was detected. If this occurs the SAR/interworking device will set the AAL5 packet length field (always present in the last AAL5 cell) equal to 0. As described in Section 5.6, the S/UNI-APEX can be programmed to buffer the entire packet before scheduling it out to its destination port. The S/UNI-APEX can also be programmed to discard the entire AAL5 packet when the AAL5

length field equals 0. In this way the packet is only buffered once at the S/UNI-APEX -- neither the Freedm nor the SAR device need buffer the entire packet.

8. Once the packet is buffered in the S/UNI-APEX the individual cells will be sent to the ATM switch fabric where they will be routed to the appropriate egress port.

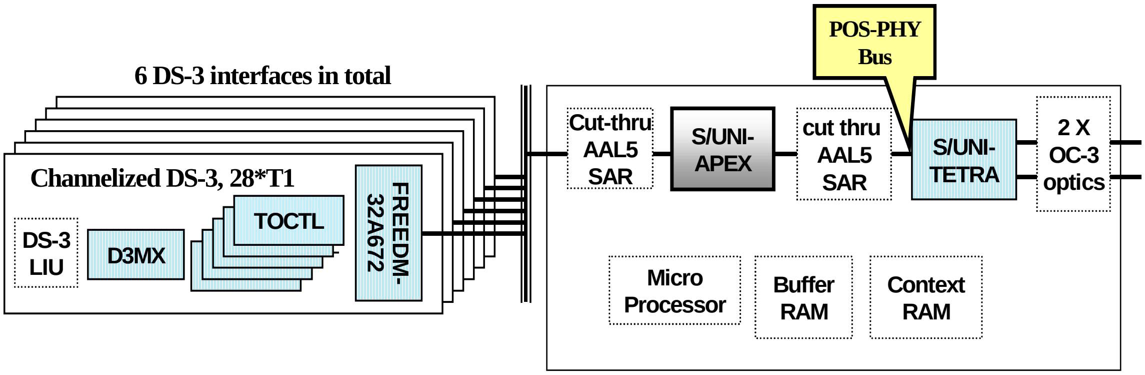
In the downstream direction (from the switch to the DS-3) Link Layer processing is handled as follows:

1. The SAR/interworking device appears as a 2016 port AnyPHY bus slave to the S/UNI-APEX. The S/UNI-APEX polls the 2016 channels and is programmed to schedule whole packets (a cell at a time, but packet contiguous) to the appropriate channels as they become available.
2. Each Freedm presents up to 672 ports (channels) to the SAR. The SAR is responsible to poll the Freedm devices and pass the per-channel flow control information back to the S/UNI-APEX in the form of TPA status indication for each of the (up to) 2016 channels. With this architecture, each n*DS-0 channel provides its own back-pressure indication all the way back to the S/UNI-APEX where the traffic buffering and scheduling is being managed. Since the S/UNI-APEX will only schedule traffic to a channel if that channel is ready to accept it this architecture eliminates the potential for head of line blocking inside the SAR and Freedm devices. It also greatly reduces buffering requirements in the SAR.
3. The SAR/interworking device receives the packet from the S/UNI-APEX in a stream of contiguous packets with no interleaving of packets from other channels. Therefore the packet reassembly and frame relay interworking functions require only shallow buffering and are greatly simplified. As cells arrive the SAR strips off the 5 ATM header bytes, performs the FRF.8 or FRF.5 interworking, and passes the resulting data blocks to the appropriate Freedm-32A672.
4. The Freedm-32A672 performs the HDLC generation on each stream, does the CRC-32 calculation, and inserts the data into the destination n*DS-0 channels in the appropriate T1 channel. When no user packets are being sent on a n*DS-0 channel the Freedm-32A672 inserts idles flags into the channel.
5. Each T1 framer processes a T1 stream, adds the overhead, and provides the D3MX with a T1 stream.
6. The D3MX multiplexes the 28 T1s into the DS-3.
7. The DS-3 LIU converts the DS-3 signal from digital to analog.

Switching Frame Relay to Frame Relay

In the previous section we described the basic Link Layer processing of a frame relay interface to the S/UNI-APEX using ATM interworking. This would be used when frame relay is being carried over ATM, i.e. the lines are carrying packet traffic but the up-link from the switch is carrying ATM cells. We now compare this to an example where both the line side and up-link sides of the switch are carrying packet traffic. Refer to Figure 4.

Figure 4 - Simple Frame Relay Multiplexer Without ATM Interworking



6 DS-3 to 2 OC-3 Frame Relay Multiplexer

The architecture of Figure 4 is similar to that of Figure 3, with the key differences noted below:

- This example (Figure 4) represents a standalone frame relay multiplexer while Figure 3 shows a frame relay port card in an ATM switch. Where the previous example had an interface to the ATM switch fabric, this example has a dual OC-3 Packet Over SONET (POS) or Frame Relay up-link.

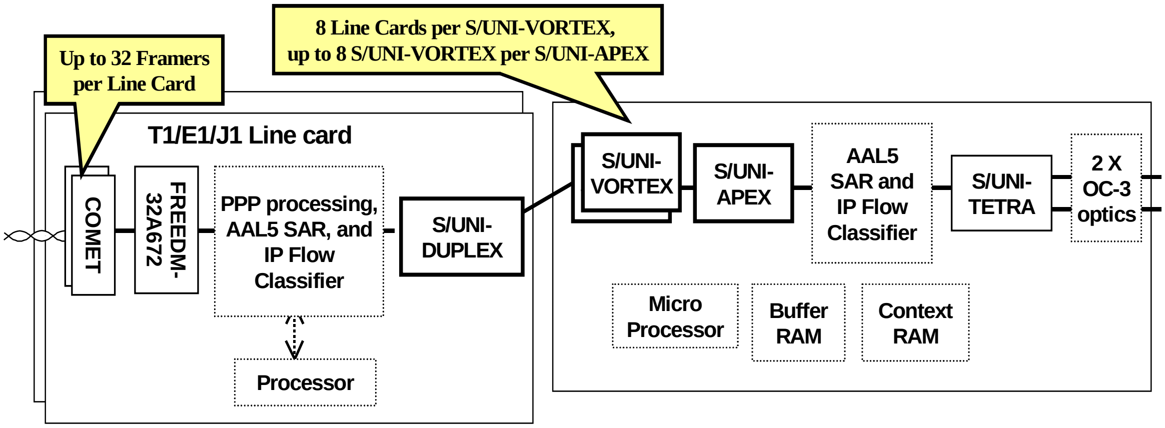
- In this example the line side frame relay to ATM interworking is eliminated. In the ingress direction the cut-thru SAR can directly generate the per-cell switch tag (ICI) for the S/UNI-APEX from the frame relay DLCI field and the HDLC channel number. The ATM cell's VCI and VPI fields can be left undefined since the S/UNI-APEX only requires an ICI. Alternatively, the DLCI could be mapped into the VPI/VCI and the limited address resolution capability of the S/UNI-APEX could by utilized to generate the cell's ICI (see Section 6.6, Ingress Address Resolution Using the S/UNI-APEX). In either case, AAL5 encapsulation, including setting length = 0 when the CRC-32 is bad, should be used to allow the S/UNI-APEX to perform packet discard of errored packets, as described in the previous example. Depending on the service being offered there may also be a need to include ingress packet policing at this point. The packet policing would be analogous to what is performed by the S/UNI-ATLAS for ATM traffic.

- In the WAN side, an AAL5 functionality analogous to the line-side is needed although it is greatly simplified as there will be only two HDLC channels, one per OC-3. The POS-PHY bus interface on PMC-Sierra's S/UNI-TETRA quad OC-3 PHY allows for arbitrary length transfers, similar to the Any-PHY bus used by the Freedm-32A672. The S/UNI-TETRA provides an integrated HDLC controller for each of its OC-3 interfaces, and indicates CRC-32 errors when the end of packet is reached. Therefore the AAL5 functionality on the WAN side will correspond closely to that of the line side.

5.3.3 IP over PPP

The basic architecture needed to implement an IP edge router using the S/UNI-APEX is similar to the frame relay architecture discussed previously. Refer to Figure 5 - 2048 Channel IP Edge Router. This figure also provides an example of how the S/UNI-VORTEX and S/UNI-DUPLEX can be used to construct a multi-card architecture to process packets, although this technique of placing the SAR on the line cards rather than the core card could have been used in the frame relay examples as well.

Figure 5 - 2048 Channel IP Edge Router


8 Line Cards per S/UNI-VORTEX,
up to 8 S/UNI-VORTEX per S/UNI-APEX
Up to 32 Framers per Line Card
T1/E1/J1 Line card
COMET
FREEDM-32A672
PPP processing, AAL5 SAR, and IP Flow Classifier
S/UNI-DUPLEX
Processor
S/UNI-VORTEX
S/UNI-APEX
AAL5 SAR and IP Flow Classifier
S/UNI-TETRA
2 X OC-3 optics
Micro Processor
Buffer RAM
Context RAM

In this example, the line side IP packets arrive and leave the edge router encapsulated in PPP frames. The PPP protocol is terminated on ingress and generated on egress from the router. This also includes processing the Layer 2 Control Protocol (LCP) packets that manage the PPP links. The LCP processing can be done locally by the processor on each line card or the LCP packets can simply be tagged and switched to the microprocessor port on the S/UNI-APEX for processing by the core card microprocessor. On the WAN side the packets arrive on Packet over SONET (POS) as was discussed in the frame relay example.

Besides terminating the PPP or POS in the ingress direction, the Link Layer processing should include segmentation of the packets into AAL5 cells, as described in the frame relay examples. Note, however, that full interworking of IP with ATM is not required. Rather, the S/UNI-VORTEX, S/UNI-DUPLEX, and S/UNI-APEX are used purely to transport and switch packets from ingress ports to egress ports.

5.4 Ingress Address Resolution or Flow Classification

Whether it is switching AAL5 packets or individual ATM cells, the S/UNI-APEX requires each cell to enter the device carrying a16 bit switch tag, also called an Ingress Connection Identifier (ICI). The ICI can be prepended to the cell, encoded into the VCI/VPI fields, or held in the HEC/UDF fields. As the cell enters the device its ICI is used as a direct index into the S/UNI-APEX's context memory, which is capable of storing information for 64K connections. The context for each ICI fully defines how the cells tagged with that ICI are processed by the S/UNI-APEX.

Generation of the ICI can be performed anywhere along the ingress path. The S/UNI-VORTEX and S/UNI-DUPLEX do not require an ICI but are capable of passing extended length cells to the S/UNI-APEX. Therefore the ICI can be generated on the line card and prepended to the cell before it is sent to the core card.

Looking at the previous examples, we see that the ICI is generated as follows:

- In the access multiplexer example shown in Figure 1 on page 4 the ICI is generated by the S/UNI-ATLAS's ARL (Address Resolution Lookup) function. S/UNI-ATLAS ARL is discussed more fully in Section 6.7. The ARL uses the PHY ID and the VCI/VPI on the line side and the physical WAN port number and the VCI/VPI on the WAN side. In both cases the mapping between VPI/VCI and ICI will be programmed into the S/UNI-ATLAS by the local microprocessor based on the PVCs or SVCs currently in effect. The context associated with each ICI will also depend on the current PVCs or SVCs active in the multiplexer. The ICI context defines the VC's switching, class of service, scheduling weight, etc.. ICI context is established via the microprocessor port on the S/UNI-APEX, as discussed in Section 6.1.

- The 3G wireless base station example shown in Figure 2 on page 9 is similar to the access multiplexer except on the line side the PHY ID corresponds to a radio card. Also, on the WAN side the S/UNI-ATLAS only uses the VCI/VPI since there is only one up-link port via the IMA processor.

- In the frame relay example of Figure 3 on page 10 interworking with ATM is performed by the customer provided SAR/interworking device. In the line side ingress direction this interworking can include the generation of the appropriate VCI/VPI address fields derived from the channel number and frame relay DLCI address fields. The 16 bit ICI should also be generated at this time. For the switch side ingress direction (cells going from the switch to the S/UNI-APEX) the ICI is generated as an ECI (egress cell identifier) by the far end port. See Section 6.7 for a discussion of egress cell identification capabilities of the S/UNI-APEX.

- In the frame relay switching without ATM interworking example (Figure 4, page 14) the ICI is generated coincident with the AAL5 SAR function, which is performed on core card for both the line side and WAN side ingress traffic. As in the previous frame relay example, the ICI is derived from the channel number (only the line side has multiple channels, there is only one channel from the WAN side) and the DLCI field of the frame header.

- The IP edge router example of Figure 5 on page 15 represents a scenario where the line-side ICI is generated on the line card and passed transparently through the S/UNI-VORTEX and S/UNI-DUPLEX. The lookup will likely be based on some form of flow classification where each packet's IP source and destination addresses, ports, and perhaps the DiffServ bits (to identify class of service) are used to dynamically generate an ICI.

As each cell enters the S/UNI-APEX it uses the ICI to read the associated connection context from external memory. This context is fully configurable for each ICI and determines precisely how all cells associated with that ICI are handled.

5.5 Ingress Policing

PMC-Sierra's S/UNI-ATLAS device is capable of feature rich ATM policing, and is an integral part of the VORTEX chip set. The S/UNI-ATLAS can handle the policing function for ATM-centric equipment such as DSLAMs, multi-service access multiplexers, and 3G wireless base stations and base station controllers. Policing of frame relay services can be performed by the S/UNI-ATLAS if frame relay to ATM interworking is being performed. However, policing of IP traffic is not within the scope of the S/UNI-ATLAS device. A customer supplied device is required if policing of IP traffic is to be implemented.

Policing of ingress traffic is performed to ensure the actual traffic pattern fits within the contracted traffic profile. Received cells that do not fall within the traffic contract are marked by the S/UNI-ATLAS by setting the CLP bit to 1. The S/UNI-ATLAS implements per-VC dual leaky bucket policing or GFR policing on a per-VC basis. The S/UNI-ATLAS also maintains per VC traffic (cell) counters that can be used to measure traffic flow. Refer to the *S/UNI-ATLAS Data Sheet* for details.

5.6 Ingress Congestion Management

Congestion management is the process of managing shared resources under heavy load. The S/UNI-APEX is a shared memory switch capable of storing up to 256K cells of traffic in an external 16 Mbyte DRAM. This external memory is shared by all active connections. Using a large centralized shared buffer space reduces system-wide buffer requirements due to the statistical gain of combining traffic patterns across a large number of users. However, sharing requires congestion management features that ensure misbehaving users do not consume too many resources. To this end, the S/UNI-APEX provides highly configurable congestion management features that support a wide range of service types.

Congestion management decisions are made by the S/UNI-APEX immediately upon receiving an ingress cell. In its simplest form, congestion management is a binary decision – should I store this cell in the shared memory, or should I discard it? Provided below is an overview of how the congestion management process is configured.

When the S/UNI-APEX receives a cell or packet it checks the discard threshold for the destination direction, port, class, and VC. Only if a cell/packet has passed all thresholds without being discarded will it be permitted entry into the data memory. For cell based traffic this decision is made on each cell received. For packet traffic this decision is made at the first cell in the packet and applies to all subsequent cells on that packet.

The system programmer can allocate buffer resources based on the cell's destination, which is comprised of the following components:

- direction or output bus (the WAN bus, line-side bus, or microprocessor bus),
- port number (1 of 4 ports on the WAN bus or 1 of 2048 ports on the line-side bus, the microprocessor bus is treated as a single port)
- class of service (four classes in total with the buffers of each class shared across all VCs assigned to that class), and
- VC (the 64K ICI values).

For each destination component the congestion control has four zones of operation:

1. Plenty of resources available for this component of the destination, no discard.
2. CLP1Threshold exceeded, discard all CLP1 cells or packets, accept all CLP0 cells or packets. For VCs configured in packet discard mode, if a CLP0 packet has already been started the remainder of that packet will be admitted up to the maximum threshold value. If the max threshold is reached (i.e. the corresponding direction, port, class, or VC buffer space is completely full) packet discard will be invoked. If the port is configured in EPD/PPD congestion mode (i.e. WFQ applied to packet traffic) partial packet discard is used. If the port is configured in frame contiguous mode (whole packet are scheduled) the entire packet is discarded.
3. CLP0Threshold exceeded, discard all cells/packets except 1) CLP0 or CLP1 packets that have already been started (packet mode VCs), and 2) CLP0 cells that have not met their CLP0MinThreshold allocation (cell or packet mode). The CLP0MinThreshold is discussed further below.
4. maxThreshold exceeded, resources in this component are exhausted, discard all traffic. Note that if packet discard mode is configured the last cell of the packet (containing the EOM bit) will be stored even if maxThreshold has been exceeded. This is true unless the device maxThreshold has been exceeded, which implies there is no buffer space remaining at all.

On a per-VC basis, the S/UNI-APEX can be programmed to use cells or packets as the fundamental unit of traffic discard, as shown in Figure 6. Packet discard is always preferred when the underlying traffic is frames over AAL5 since this maximizes the "good-put" of the S/UNI-APEX. Put another way, if you are sending packets there is no use in discarded just one cell from a packet, you might as well discard the whole packet. The handling of Early Packet Discard and Partial Packet Discard are fully described in the *S/UNI-APEX Data Sheet*.

For each VC's CLP0 traffic there is also a minimum buffer allocation threshold that can be set. The CLP0MinThreshold parameter is an important value because it modifies the per-VC congestion management behavior under heavy load. Each VCs CLP0MinThreshold value defines the number of reserved CLP0 cells reserved to the VC even if the CLP0Threshold has been reached on one or more of the various destination components (direction, port, class, VC).

The interplay of the minThreshold values and the CLP0Threshold values is best explained with some examples. In what follows we simplify things by setting the CLP0 and CLP1 thresholds equal to the CLP0Threshold (i.e. in these examples there is only one active threshold, the CLP0Threshold).

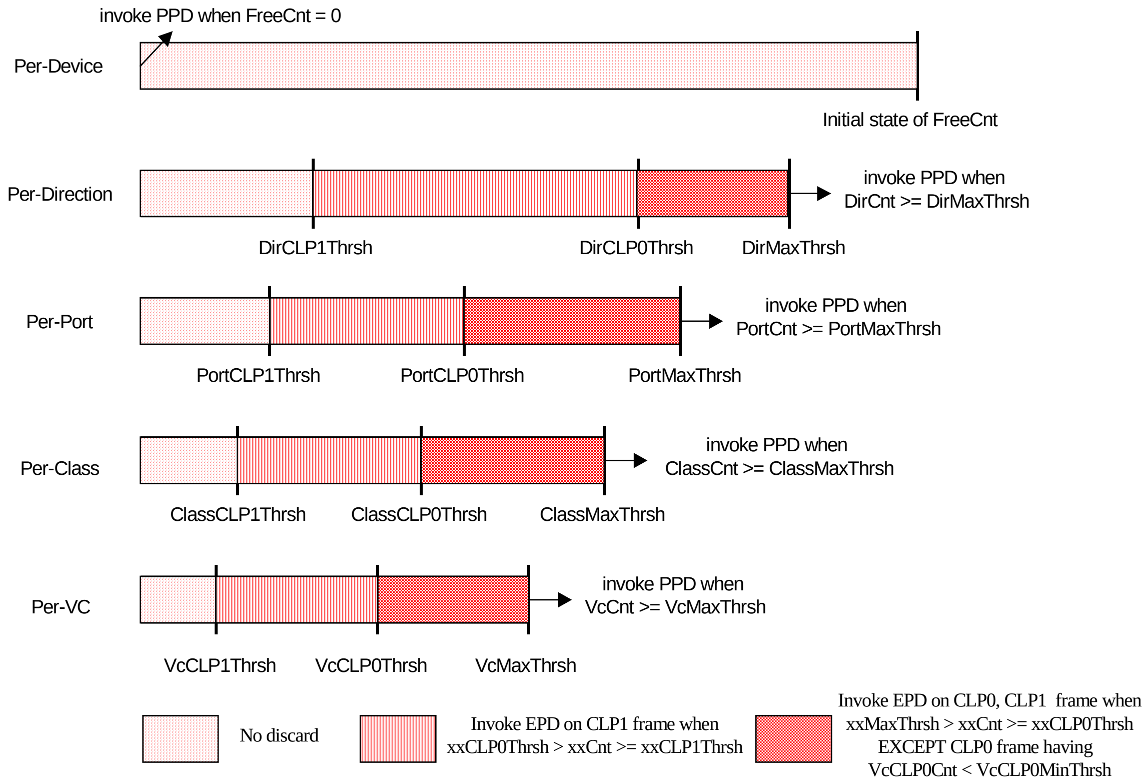
Suppose a port has two Class 1 VCs assigned to it, VC#1 and VC#2. Suppose VC#1 is the only VC active and it has consumed all the cells up to the port's CLP0Threshold. If VC#2 becomes active the VC#2 cells will begin to arrive at the S/UNI-APEX. These VC#2 cells will fail their destination port's PortCLP0Threshold test but still be admitted up to VC#2's minThreshold value[1]. Meanwhile, until enough cells have been sent out on the port to bring the port below its CLP0Threshold, cells arriving from VC#1 will be discarded.

Taking the previous example further, suppose other VCs begin to consume Class 1 buffers until the Class#1 classCLP0Threshold is reached. In this scenario newly arrived VC#1 cells will be discarded based on the classCLP0Threshold, while VC#2 cells are admitted based on its minThreshold. Over time, cells in this class will be sent out and others admitted, but let's assume that in aggregate the class continues to be congested (more cells arrive than sent). In this case all active VCs in Class#1 will eventually converge on their minimum threshold levels. The sum of the minTheshold values defines the class's fully congested maximum threshold, although in reality the sum cannot be greater than the programmed value of classMaxThreshold.

From the above examples it is clear that as traffic bursts into the S/UNI-APEX there are many possible interactions between the various maximum thresholds and the per VC minimum threshold. However, a predictable steady state congestion management algorithm is always possible as long as the sum of all active VCs' minThreshold values is less than the total buffer space. Once the minThreshold values are set, the software can then focus on setting the port, class, and direction maximum thresholds to allow for optimal resource allocation during partially congested states, where optimal is defined by the nature of the service classes being implemented by the equipment.

[1] We are assuming that the port's maximum threshold is not reached, of course. If it is then unconditional discard would result regardless of the minThreshold.

Figure 6 - EPD/PPD Threshold Hierachy



To illustrate how congestion management could be used at the service level, consider the example of a service provider wishing to implement four service classes: CBR, Voice Over IP (VoIP) carried over VBR-RT, GFR, and UBR:

- VCs belonging to the CBR service would be assigned to the highest priority class and would not require a large buffer space. The congestion threshold for the CBR class of service would be quite low, as would the CLP0 threshold on a per VC basis. The per VC minThreshold may not required for CBR because each class has its own memory space (the sum of the four class maxThreshold values will not exceed the buffer space). However, setting each VC's minThreshold equal to its maxThreshold would be necessary if per port thresholds might cause CBR traffic to be discarded (see below). The CLP1 threshold for the CBR VCs could be zero, meaning these policed cells are always discarded.

- The VBR-RT traffic would be assigned the second priority class. Since this traffic is latency sensitive but bursty it would require a larger max threshold than CBR. Depending on the service model, this traffic might be tightly policed to ensure misbehaving VCs do not drive latency onto the other VCs. Hence the CLP1 threshold might be zero or very small. As with CBR traffic, the minThreshold would only be required if per port thresholds are used.

- The GFR and UBR services could share the third class of service, which would have a ClassMaxThreshold that consumes the remainder of the cell buffers. Here the CLP1 thresholds on each VC could be high, but less than the CLP0 thresholds. The GFR VCs would have a non-zero minThreshold, but the sum of the minThresholds would not exceed the ClassMaxThreshold (otherwise the GFR VCs could not be assured their full minThreshold). The UBR VCs would have a zero minThreshold, so under maximum congestion the GFR VCs take priority over the UBR traffic. Even under partial congestion, no single VC should be allowed to consume more than a certain portion of the total buffer space, and this can be assured by setting the VC's VCCLP0Threshold and VCmaxThreshold appropriately. Layered on top of this could be a per-port congestion level used to ensure that a single port does not consume too many buffers across the aggregate traffic of all its VCs.

To support ABR services, optional EFCI marking can be enabled on a per-VC basis. Marking of EFCI can be programmed to occur when either the CLP1 threshold or the CLP0 threshold has been exceeded on the associated direction, port, class, or VC.

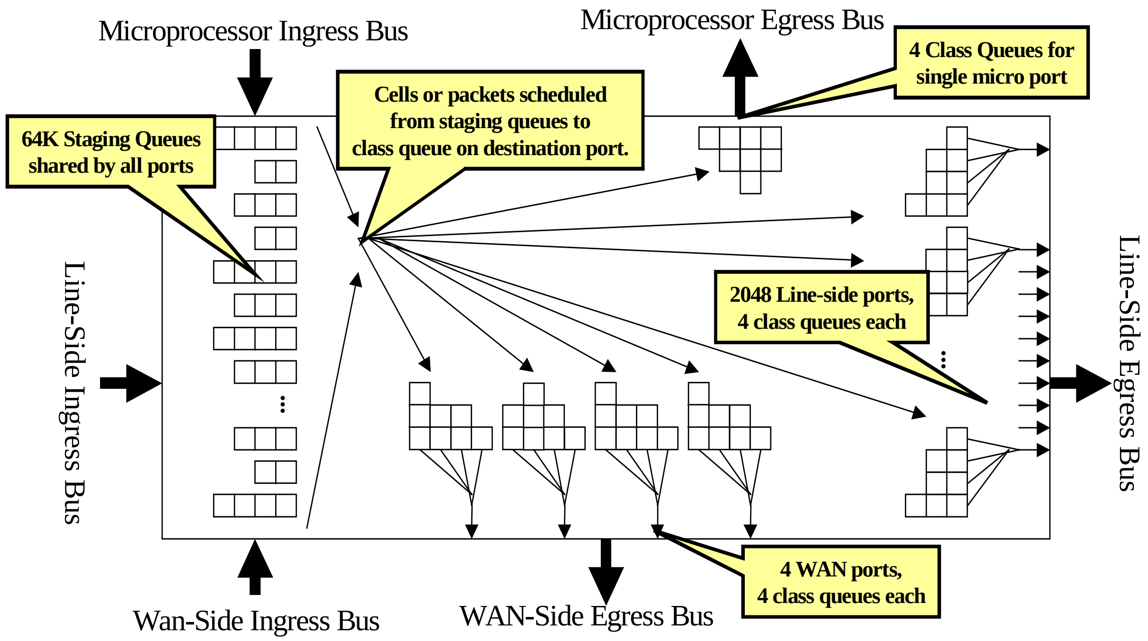
The S/UNI-APEX assumes that traffic policing and cell/packet counting is performed outside of the device. However it does assist in the traffic monitoring task in two ways:

1. The S/UNI-APEX counts discarded cells and provides identification of the ICI of the last cell discarded. There are three discard counters implemented in the device: a CLP0 congestion discard count, a CLP1 congestion discard count, and an "other discard" count. Discards not due to congestion are grouped in the "other" category and include discards due to packet reassembly timeouts, cells received (and discarded) on disabled VCs, or cells discarded due to a connection tear-down.
2. Two cell emission counts are maintained for each VC of the 64K VCs, one counter for CLP0 cells and one for CLP1 cells. Ingress cell counts are not provided since they will have been counted by the external policing device.

5.7 Egress Queuing, Class Scheduling, and Port Scheduling

Once a cell has successfully passed through the congestion management process without being discarded it is queued in the external DRAM where it awaits egress scheduling. The destination class and port is defined by the connection's context, as determined by the ICI context lookup. Refer to Figure 7.

Figure 7 - Egress Queuing Structures


Microprocessor Ingress Bus
Microprocessor Egress Bus
4 Class Queues for single micro port
Cells or packets scheduled from staging queues to class queue on destination port.
64K Staging Queues shared by all ports
Line-Side Ingress Bus
Line-Side Egress Bus
2048 Line-side ports, 4 class queues each
4 WAN ports, 4 class queues each
Wan-Side Ingress Bus
WAN-Side Egress Bus

5.7.1 Egress Cell Scheduling

This section discusses the cell scheduling capability of the S/UNI-APEX. Cell-at-a-time scheduling will be used for ATM based architectures such as shown in Figure 1, Figure 2, and Figure 3.

The S/UNI-APEX maintains 64K (65535) individual queues, one per VC (ICI). Therefore each connection has its own per-VC queue where cells await their turn to be scheduled into the appropriate class of service on the destination port. Scheduling from the staging queues into the class of service queues is performed a cell at a time using a software configurable weighted fair queuing algorithm. The weight associated with each connection determines the maximum number of cells that can be placed in the destination class queue at any given time. In general, VCs expected to sustain a higher bandwidth than the other VCs sharing the same class of service on the same destination port should be given a proportionately higher weight. This, combined with the per-VC queues, ensures that high bandwidth VCs receive proportionally more bandwidth compared to lower bandwidth VCs.

Scheduling from the four class queues into their associated line-side or WAN-side port (each port has 4 class queues) is performed with simple priority – lower priority classes must wait until higher priority cells are sent. There is, however, an optional programmable minimum bandwidth guarantee for the lower priority classes. In this case the lower priority class maintains a counter which is incremented each time it has a cell to send but is not allowed to because of a higher priority class. When the counter reaches the programmed limit the lower priority class is allowed to override the higher class for a single cell transfer, at which time the counter is reset. This allows for a minimum bandwidth guarantee and is useful for GFR type services.

The scheduling of ports onto their associated physical egress bus differs depending on the bus type:

- The microprocessor bus does not perform class scheduling … it is up to the microprocessor to determine which of the four class queues it wishes to read.
- The four ports of the WAN bus are scheduled onto the bus using a configurable weighted interleaved round robin scheduling algorithm described in *the S/UNI-APEX Data Sheet*. WAN port scheduling can also be shaped, as discussed below.
- Scheduling to the 2048 loop ports is based on a highly configurable algorithm that has been optimized to reduce unnecessary device polling while supporting a wide range of port speeds. This is fully described in *the S/UNI-APEX Data Sheet*.

5.7.2 Egress Packet Scheduling

This section discusses the packet scheduling capability of the S/UNI-APEX. Packet contiguous scheduling will be used for frame based architectures such as shown in Figure 4 and Figure 5. Note, however, that there is no inherent restriction preventing mixing cell and packet based scheduling. For example one group of VCs and ports could be operating in cell mode, while the remaining VCs and ports operate in packet mode.

To perform VC merge or frame switching on a particular VC, the staging queue for that ICI can be configured to act as an AAL5 packet reassembly queue. In this mode the entire packet will be collected in the staging queue before it is scheduled onto its destination class queue. As cells arrive the S/UNI-APEX monitors the PTI bit of the ATM header, looking for the End of Message (EOM) indication. Once the EOM cell is found the entire staging queue (which will contain the entire AAL5 packet) is transferred to the destination class queue. As discussed in Section 5.3.2, the staging queue can also be configured to evaluate the AAL5 length field of the EOM cell and discard the entire packet if the length field is set to 0. Programmable reassembly thresholds are available to discard packets that have grown too large due to a missing EOM cell. As well, a programmable packet reassembly watchdog operates in the back ground to flush queues (i.e. discard cells) that have been stranded by a missing or corrupted EOM cell.

Egress ports that are receiving frame contiguous traffic can also be programmed to function in packet-contiguous mode. In this mode, once a class queue gains permission to transmit on the port it will continue to transmit the entire packet a cell at a time, regardless of whether higher priority traffic arrives during the transmission. Another option is to operate the port in packet fragmentation mode. In this mode lower priority classes will pause (i.e. fragment) their packet transfer if a higher priority packet arrives. See the *S/UNI-APEX Data Sheet* for more details.

Fragmentation mode is of value when the S/UNI-APEX is being used in frame and IP switching equipment where packet QoS via packet fragmentation is supported. The external SAR function is responsible for properly terminating the partial packets and maintaining the higher level packet fragmentation protocol.

5.8 Egress Traffic Shaping

Traffic shaping is available on the four WAN ports, but not on the loop ports. A maximum of four out of the sixteen WAN port classes (four ports, four classes per port) can have shaping applied to their output. Every VC connected to a shaped class will have shaping applied to it, but each VC can have a unique shape rate, as described below. Classes that are not shaped can coexist on the same port as classes that are shaped, and there can be more than one shaped class on a single port.

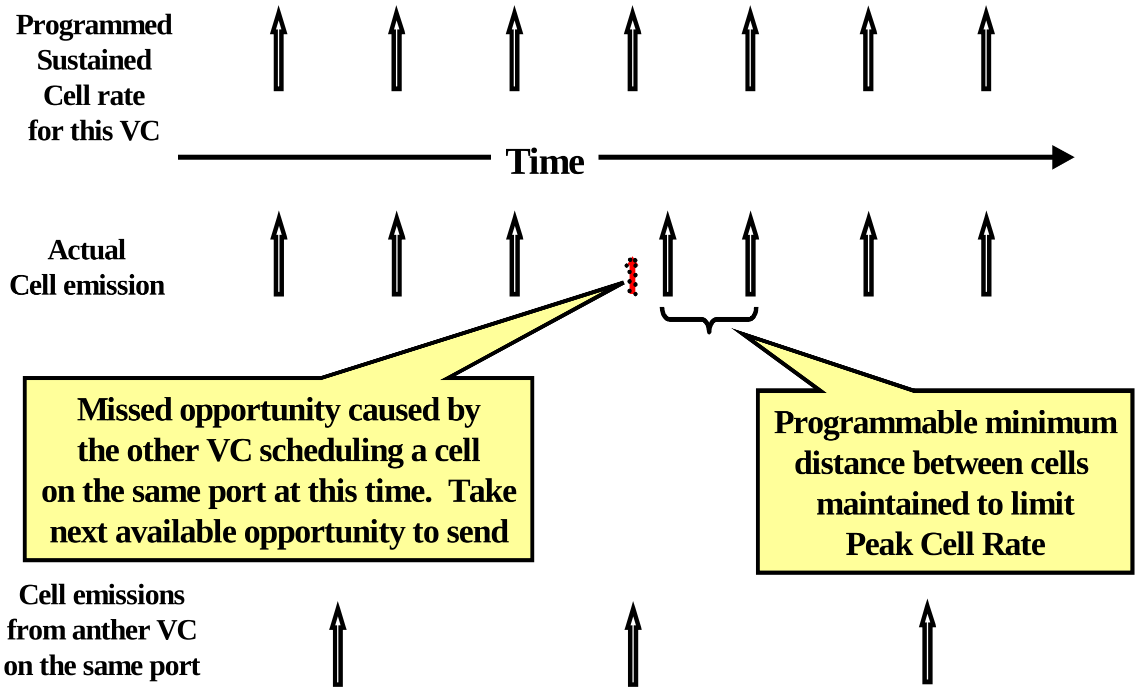
User programming defines how each of the four shapers function, and each shaper is independent of the other three shapers. Programming determines the port and class to which the shaper is assigned, and the value of the fundamental time step (tmin) used by the shaper. tmin is the minimum time increment between successively scheduled cells from the same VC. Per VC programming determines how many of these minimum time periods will be inserted between the VC's cells as they scheduled by the shaper.

User programming of each VC's context table entry determines how that VC will be shaped by its assigned shaper. The key per-VC parameters are as follows:

1. The port and class (i.e. the shaper) to which the VC is assigned.
2. The desired number of tmin periods to insert between cells from this VC emitted on the output port.
3. The minimum number of tmin periods to insert between cells. This value limits how closely cells from a single VC can be scheduled when the optimal scheduling could not be achieved due to contention at the output port. The minimum cell spacing to use is defined by the peak cell rate limit desired for the VC. This is shown in Figure 8.

User programming can also define the action of the shaper when overall egress congestion (i.e. too much traffic being sent to a single port) is causing all VCs on that port to experience shaping delay due to port contention. Based on congestion thresholds set by the user the shaper will temporarily increase the fundamental shaping time unit, tmin, thereby causing each VC to schedule cells less frequently. This will eventually relieve the congestion, at which point tmin will be reset to its previous value. Since all VCs on the port experience the same relative decrease in scheduling frequency, the negative impact of the contention is distributed fairly across all VCs on the congested port. See the *S/UNI-APEX Data Sheet* for details.

Figure 8 - Traffic Shaping on the WAN Port


Programmed
Sustained
Cell rate
for this VC
Time
Actual
Cell emission
Missed opportunity caused by the other VC scheduling a cell on the same port at this time. Take next available opportunity to send
Programmable minimum distance between cells maintained to limit Peak Cell Rate
Cell emissions
from anther VC
on the same port

6 SPECIAL TOPICS

The following topics are of interest to system architects wishing to take full advantage of specific features of S/UNI-APEX.

6.1 The S/UNI-APEX Microprocessor Port

The microprocessor port on the S/UNI-APEX is used to:

- Access device registers for initializing the device.
- Provide indirect access to the context memory (via configuration registers) for configuration of ports, classes, congestion thresholds, shape rates, etc..
- Provide indirect access to the context memory for setup, tear down, and configuration of each connection.
- Access four class of service queues from which cells can be read.
- Access a two cell (double buffered) cell write buffer through which cells are written into the S/UNI-APEX. Each cell write includes an ICI prepend. The ICI determines how the cell is handled by the S/UNI-APEX just as it does for the WAN and line-side ingress ports.

The four microprocessor class of service queues are a valid destination for any of the 64K connections. As is the case for WAN and line-side ports, per-VC queuing is combined with the four class of service queues to provide congestion management and QoS aware scheduling for all traffic switched to the microprocessor port. The microprocessor port also includes an integrated CRC-32 hardware assist that can be used to simplify the processing requirements for AAL5 packet transfers to and from the microprocessor. A CRC-10 hardware assist is also provided to simplify OAM cell processing.

6.2 Multicast Support in the S/UNI-APEX

Multicast is supported via the cell transfer and duplication capability of the microprocessor port. There is a multicast procedure that fits the requirements of many low bandwidth multicast applications such as signaling or paging services in wireless applications.

We take advantage of the fact that the microprocessor attached to the S/UNI-APEX can simply change the switch tag (ICI) of an existing cell in its microprocessor port transmit FIFO. This allows the microprocessor to direct the same cell to numerous VCs without having to re-write the entire cell into the S/UNI-APEX. The FIFO is double buffered, so the procedure works as follows:

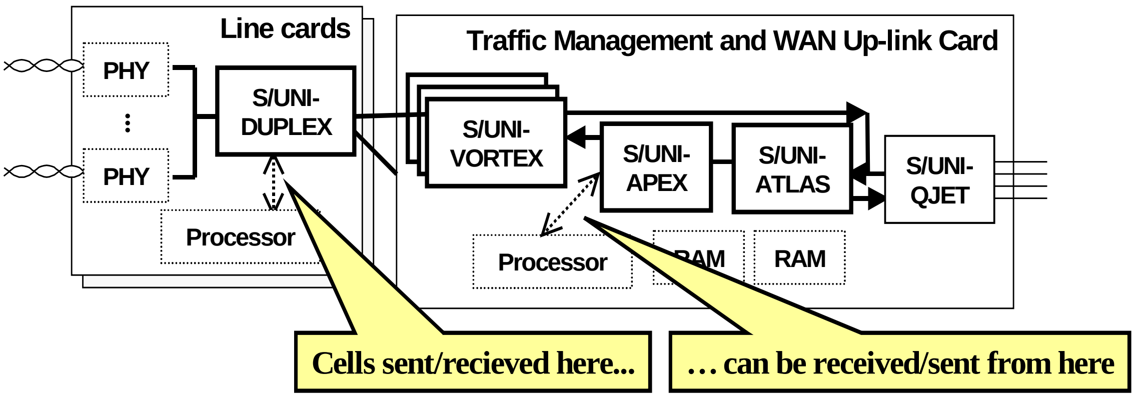
- uProcessor writes a copy of the cell into FIFO (1st buffer), including the first destination's ICI value. This cell is automatically sent to first multicast destination when the last word of the cell is written
- uP writes the same cell (with next ICI destination) into FIFO (2nd buffer), and the cell is automatically sent to the second multicast destination with the write of last word of the cell.
- For all subsequent copies, uP only has to write the next ICI value (one word) and then write the last FIFO word, which triggers the send.
- Individual writes for switch tag and last FIFO words are ~6 uP cycles each at a maximum 66Mhz microprocessor bus rate.

Of course multicast traffic takes bandwidth away from the non-multicast traffic (i.e. all traffic shares the bus connecting the S/UNI-APEX to the S/UNI-VORTEX devices) but the S/UNI-APEX does ensure that the multicast traffic is queued in the appropriate class queue and doesn't override higher priority traffic.

6.3 Embedded Inter-device Communications Channel

As discussed fully in the *S/UNI-VORTEX and S/UNI-DUPLEX Technical Overview* document referenced in the Required Readings section on page 1, the S/UNI-VORTEX and S/UNI-DUPLEX provide an addressable embedded communication channel that can be used to create an inter-processor communication channel between the core card and every line card. The S/UNI-APEX participates in this communication via one or more of its class queues on the microprocessor port. Refer to Figure 9.

Figure 9 - The Embedded Control Channel



Cells sent/recieved here...

… can be received/sent from here

6.4 ATM OAM Handling Using the S/UNI-ATLAS and S/UNI-APEX

The S/UNI-ATLAS is a fully featured ATM OAM processor capable of handling F4 and F5 FM and PM processing compliant with the 1999 version of the I.610 standard. The S/UNI-ATLAS handles most OAM cell processing without microprocessor intervention.

In architectures where the S/UNI-ATLAS is not used, or should the I.610 OAM standard change in some way, the microprocessor port on the S/UNI-APEX provides an fully buffered path for OAM cell traffic to be directed to the microprocessor for processing in software. On a per VC basis the S/UNI-APEX can be programmed to treat the VC as a VCC or a VPC. The difference is in how the S/UNI-APEX detects OAM cells for redirection to the microprocessor port. F4 OAM flows at the VPC level are detected by monitoring the cell's VCI value. F5 flows at the VCC level are detected by monitoring the PTI field. See the I.610 standard for details.

Software based processing of OAM cells via the S/UNI-APEX microprocessor port is normally not required if the S/UNI-ATLAS is used, and hence is outside the scope of this document. For the remainder of this discussion we are assuming that the S/UNI-ATLAS is handling all OAM processing and that the S/UNI-APEX is programmed to treat OAM cells like any other user cells on that VC.

There is one subtlety in the OAM processing of the S/UNI-ATLAS that requires further discussion. It is assumed the reader is familiar with the I.610 OAM protocols. Refer to Figure 10 and Figure 11, which show details of the interconnection between the S/UNI-ATLAS and the S/UNI-APEX.

If programmed to do so, the S/UNI-ATLAS can detect incoming AIS OAM cells on its ingress input ports and automatically generate the response RDI cells. However, the S/UNI-ATLAS sends these automatically generated RDI cells out on the egress output port since in a typical S/UNI-ATLAS configuration[1] the PHY device is connected to the ingress input and egress output ports. However, in the typical S/UNI-APEX to S/UNI-ATLAS configuration only the WAN PHY is connected in this fashion, while the line-side traffic is handled by the S/UNI-VORTEX which is connected as shown in the figures.

As shown in Figure 11, RDI cells heading to the WAN port are not an issue, they naturally flow back to the WAN port as expected. However, Figure 10 shows that loop-side RDI cells do require special handling. In order to capture these automatically generated RDI cells and route them to the appropriate loop ports the S/UNI-APEX must be connected to the S/UNI-ATLAS egress output port as shown. This is in addition to the connection from the S/UNI-APEX to the S/UNI-ATLAS's ingress output port, which is the normal flow taken by user cells heading to the WAN port from the loop port.

This bus configuration works because the S/UNI-ATLAS can be programmed such that, on a per VC basis, the ICI value added to user cells sent over the Ingress Output port is identical to the ICI value added to OAM cells automatically generated and sent over the Egress Output port. Even though the cells arrive at the S/UNI-APEX on different

[1] A typical application for the S/UNI-ATLAS is on a switch port card where the S/UNI-ATLAS sits between one or more PHY devices and a switch fabric. In the typical S/UNI-APEX – S/UNI-ATLAS configuration the S/UNI-ATLAS is being used "twice", once to process the line-side ingress OAM traffic, and once to process the WAN side ingress OAM traffic. While this is an efficient use of the S/UNI-ATLAS device, it does create the non-standard configuration being described here.

physical ports, the identical ICI values ensures that the cells are routed to the same loop port.

Although we have been discussing how AIS/RDI cell flows are handled, the same bus configuration can be used to ensure that automatically generated backward reporting PM cells are properly routed to the loop ports. The *S/UNI-ATLAS Data Sheet* discusses the automatic generation of RDI and PM flows in detail. Since Loopback cells are not automatically generated by the S/UNI-ATLAS they do not present a problem.

Figure 10 - Loop port to WAN port User and RDI Cells

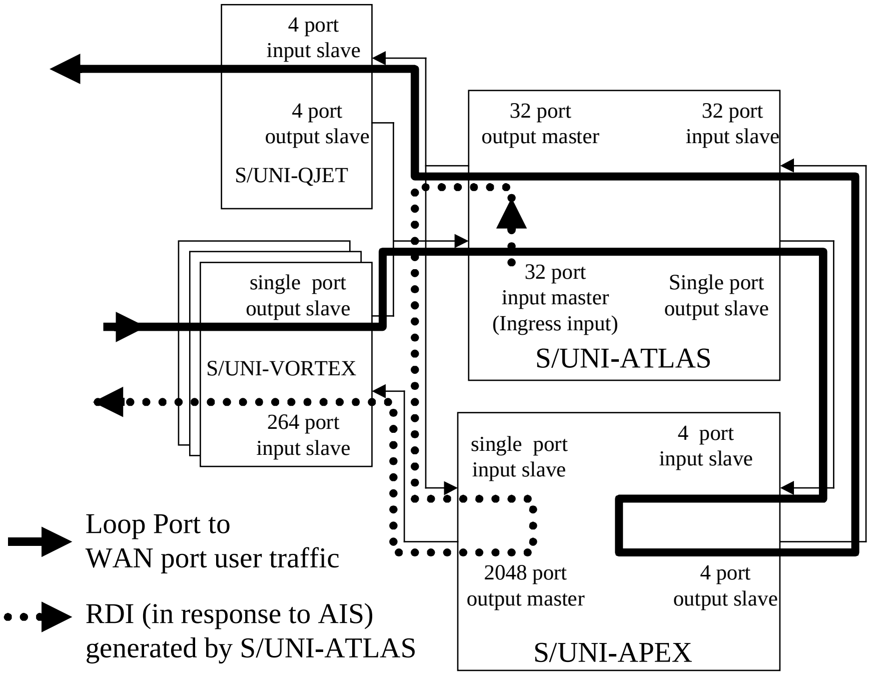
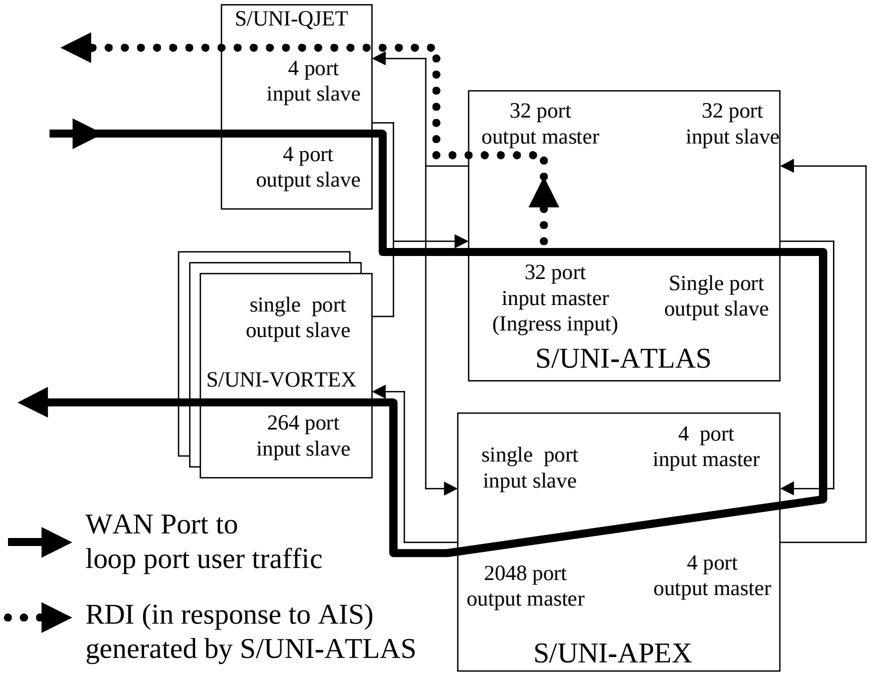


Figure 11 - WAN port to Loop port to User and RDI Cells



6.5 ATM Signaling and IP ICMP Processing

In ATM-centric equipment, centralized software processing of in-band ATM signaling, such as required for SVCs, can be readily accommodated by the fully buffered cell read/write capability of the S/UNI-APEX's microprocessor port. SVC ingress traffic will be identified by its VCI/VPI value and tagged, by the S/UNI-ATLAS for example, with an ICI that results in the cell being switched to a high priority class queue on the S/UNI-APEX microprocessor port. The signaling cell will remain buffered there until the microprocessor can retrieve and process it. Similarly, signaling response cells can be generated by the microprocessor and sent to the appropriate port by prepending the appropriate ICI value to the cell.

In IP-centric equipment routing table updates and other important information will be sent to and received from the equipment's control system in ICMP (Internet Control Message Protocol) packets. Incoming ICMP packets will be classified, tagged with the appropriate ICI, and SARed like all other packets. The ICI associated with the ICMP AAL5 cells will result in the cells being routed to the microprocessor port on the S/UNI-APEX. By putting the microprocessor port's class queue into packet mode the transfer of ICMP packets to the microprocessor will be performed packet-contiguous, greatly simplifying the process of reassembling the ICMP packet in software. ICMP cells sent from the microprocessor will also be in AAL5 format, with the prepended ICI set to route the ICMP packet to the correct egress port where it will reassembled into a packet by the external SAR before being sent to the PHY device. As discussed in Section 5.7.2 -

Egress Packet Scheduling on page 23, the egress port should be placed in packet mode to greatly simplify the external SAR function.

6.6 Ingress Address Resolution Using the S/UNI-APEX

As a software configurable option, the S/UNI-APEX can use a subset of the ingress cell's ATM header's VCI/VPI field in place of the ICI. This feature may be useful in non-ATM applications where the VCI/VPI values are arbitrary and have no meaning outside of the equipment, or in ATM applications in which a restricted VCI or VPI address space is sufficient. The mapping of VCI/VPI to ICI is fixed and defined as follows:

> if the VPI value is less than 0xFFF then the ICI is the VPI field, right justified and filled out to 16 bits.
>
> Otherwise the ICI is the VCI field.

6.7 Ingress Address Resolution using the S/UNI-ATLAS

In a typical configuration such as shown in Figure 1, the S/UNI-ATLAS ingress input port is connected to one or more S/UNI-VORTEX devices that provide the S/UNI-ATLAS with a stream of ATM cells to process. The S/UNI-VORTEX either prepends each cell in this steam with a unique PHY ID value (hereafter called the "prepend approach"), or the PHY ID is inserted in the HEC/UDF field (hereafter called the "HEC/UDF approach"). The former approach can be used if all other devices connected to the ingress bus support extended length ATM cells, the latter approach is used if the other devices on the bus are standard Utopia L2 devices. Since there is often a standard WAN up-link PHY device sharing this bus with the S/UNI-VORTEX devices (such as shown in Figure 10), the latter situation is common.

As discussed in Section 5.4, the first task of the S/UNI-ATLAS while processing the cells arriving on its ingress input port is to perform the address resolution function and generate a 16 bit switch tag or Ingress Connection Identifier (ICI). The ICI is used by the S/UNI-ATLAS and the S/UNI-APEX to quickly lookup the context associated with the cell. Programming of the address resolution function is fully described in the *S/UNI-ATLAS Data Sheet* and the *S/UNI-ATLAS Programmer's Guide*, both of which are available from the PMC-Sierra web site. For the remainder of this section we discuss a couple of subtle issues that may not be immediately obvious when implementing the S/UNI-ATLAS's lookup function[1].

The S/UNI-ATLAS documentation describes how it can include the physical port ID in its primary address lookup. The physical port ID discussed in that documentation is the actual port number (the bus address) from which the cell has been transferred – it is not the PHY ID added by the S/UNI-VORTEX and S/UNI-DUPLEX. Since the physical port

[1] The reader will need to be somewhat familiar with the S/UNI-ATLAS lookup function for the remainder of Section 6.7 to be of value.

ID is redundant with a portion of the PHY ID added by the S/UNI-VORTEX the programmer can set up the S/UNI-ATLAS to either ignore the physical port ID or ignore the high order bits of the embedded PHY ID. Note, however, that if an external device such as a WAN PHY is sharing the bus with the S/UNI-VORTEX devices then using the physical port is preferred since PHYs do not typically generate anything useful in the HEC/UDF field (more on this below).

The S/UNI-ATLAS is capable of including any part of the cell prepend or header fields in its address lookup function. A typical approach is to set up the S/UNI-ATLAS to use the physical port ID and the embedded PHY ID as the primary search key[1], while the VCI/VPI is used as the secondary key. As mentioned previously, if a standard WAN PHY is sharing the bus with the S/UNI-VORTEX then the PHY ID must be embedded in the HEC/UDF field in order to keep the bus Utopia L2 compliant. However, the S/UNI-ATLAS must perform the same address resolution procedure on every cell arriving on its ingress port, so cells coming from the WAN PHY will also have their HEC/UDF field interpreted as if they contain a PHY ID value. But what is the value of the HEC/UDF field for cell arriving from the WAN PHY?

If the PHY can be programmed to always generate a constant value in the HEC/UDF field then there will only be a single primary lookup needed for cells arriving from the WAN PHY. However, what if the WAN PHY inserts random values or multiple values in the HEC/UDF field? In this case we need to populate the primary search table with multiple pointers to the same secondary search tree, one pointer for each HEC/UDF value that the WAN PHY is capable of generating[2]. This ensures that all primary lookups on cells from the WAN PHY will resolve to the same secondary search tree regardless of the HEC/UDF value generated by the WAN PHY.

6.8 Egress Cell Identifier, switch tag, and VCI/VPI mapping

In the transmit direction the WAN bus and line-side bus of the S/UNI-APEX can be configured to modify the cells as they leave the device. Figure 12 shows the options for a SCI-PHY (extended Utopia L2) bus format. If the S/UNI-APEX is configured for the Any-PHY bus format an additional Word 0 is included for the in-band addressing defined by the Any-PHY specification. See the *S/UNI-APEX Data Sheet* for details.

The egress SCI-PHY bus can be configured for 26, 27, 28 or 29 word cell lengths. The HEC/UDF field (word 5), if it exists in the egress direction, is always overwritten by the Egress Cell Identifier (ECI) value; there is no option available that passes it through the S/UNI-APEX transparently.

[1] The primary search key is used to directly lookup the address of the secondary search table. The primary key can be a combination of the physical PHY ID (or not) plus a fixed length field taken from anywhere in the cell prepend (if it exists) or cell header fields.

[2] Typically less than 8 bits out of the HEC/UDF field are significant when the physical port ID is being used as well, in the worst case the WAN PHY will require 256 primary search table locations.

Because a single 32 bit field in the VC's context record is used to determine the value of the ECI, switch tag, and VCI/VPI fields the options for setting their values are limited to the following configurations:

1. ECI=ICI, switch tag defined in the record, no change to VCI/VPI
2. ECI and switch tag defined by the record, no change to VCI/VPI
3. ECI=ICI, switch tag defined by the record, VPI defined by the record, VCI unchanged
4. ECI=ICI, no switch tag allowed, VCI and VPI defined by the record

The "no prepends" cell length option (26 or 27 word cells) is used when interfacing to standard Utopia L2 devices. UL2 interfaces typically include the HEC/UDF field (i.e. 27 word cell) and it will contain the ECI value defined by whatever configuration (1-4 above) is used.

The ECI prepend or switch tag prepend should be used when interfacing the S/UNI-APEX to the egress input port of the S/UNI-ATLAS since the S/UNI-ATLAS is expecting cell arriving on this port to be tagged with a 16 bit context identifier. If the ability to overwrite the VCI/VPI is desired then the ECI field must be used and its value will be the same as the ICI value.

A bus configuration that contains a switch tag and ECI prepend might be required if the S/UNI-APEX is acting as an ingress switch port controller for a switch fabric that is expecting all cells to contain a switch tag when they enter the switch fabric. In this type of architecture the switch tag may be stripped off as cells leave the far end of the fabric. However, the ECI field is still available to act as the ICI value for the far end port card.

Figure 12 - Transmit Cell Format Options

	Bits 15-8	Bits 7-0
Word 1 (Optional)	**Switch Tag Prepend**	
Word 2 (Optional)	**ECI Prepend**	
Word 3	**H1**	**H2**
Word 4	**H3**	**H4**
Word 5 (Optional)	**ECI Prepend**	
Word 6	**PAYLOAD1**	**PAYLOAD2**
Word 7	**PAYLOAD3**	**PAYLOAD4**
	• • •	• • •
Word 29	**PAYLOAD47**	**PAYLOAD48**

6.9 Handling AAL2 Traffic

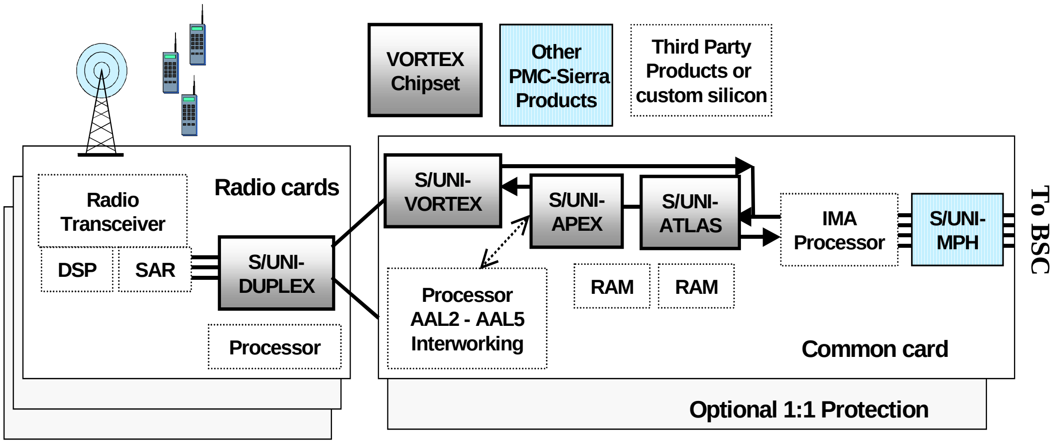
AAL2 traffic presents a special case because voice traffic for several physical ports (line-side ports) may be multiplexed into a single AAL2 cell. The S/UNI-APEX treats individual cells as the smallest unit of switching. Therefore, to use the S/UNI-APEX in an AAL2 switching configuration it is necessary to demultiplex the AAL2 cell into several ATM cells, each containing the traffic for a single port. Once each channel of the AAL2 traffic is contained in its own cell the S/UNI-APEX will be able to perform the QoS aware buffering and traffic management for each channel.

The basic concept is that all AAL2 traffic is sent to the microprocessor port first where it is demultiplexed into per-channel cell streams that are tagged with the appropriate ICI by the software and sent back into the S/UNI-APEX for switching. In the other direction the individual cells can be sent to the microprocessor where they are combined into multi-channel AAL2 cells and sent back to the S/UNI-APEX for buffering before being sent to the WAN up-link.

It should be noted that this approach works well if the net bandwidth of all traffic is low enough that the S/UNI-APEX and microprocessor do not become bottlenecks. A higher performance solution is to insert the AAL2 demultiplexer between the WAN PHY and the S/UNI-APEX, while placing an AAL2 multiplexer function[1] on each line card.

Using the wireless base station example shown in Figure 13, the following represents an AAL2 processing scenario that takes advantage of the high speed microprocessor port of the S/UNI-APEX and allows the AAL2 multiplexing/demultiplexing to be done in software.

Figure 13 - 3G Wireless Base Station



Downstream traffic flow:

[1] As long as the entire AAL2 cell is being sent to the WAN up-link there is no reason why the line card cannot construct it and send it to the S/UNI-APEX for buffering.

Voice: MPH→IMA→Atlas→Apex→microprocessor for AAL2 demultiplexing→Apex→Vortex→Duplex→SAR→DSP→Radio

Data: MPH→IMA→Atlas→Apex→Vortex→Duplex→SAR→Radio

Signaling:
Base station controller→MPH→IMA→Atlas→Apex→Core card microprocessor, or
Base station controller→MPH→IMA→Atlas→Apex→Vortex→ Duplex→Radio card microprocessor, or
Core card microprocessor→Apex→Vortex→Duplex→Radio card microprocessor

Upstream traffic flow:

Voice: Radio→DSP→SAR→Duplex→Vortex→Apex→microprocessor to multiplex individual channels into AAL2 cells→Apex→Atlas→IMA→MPH

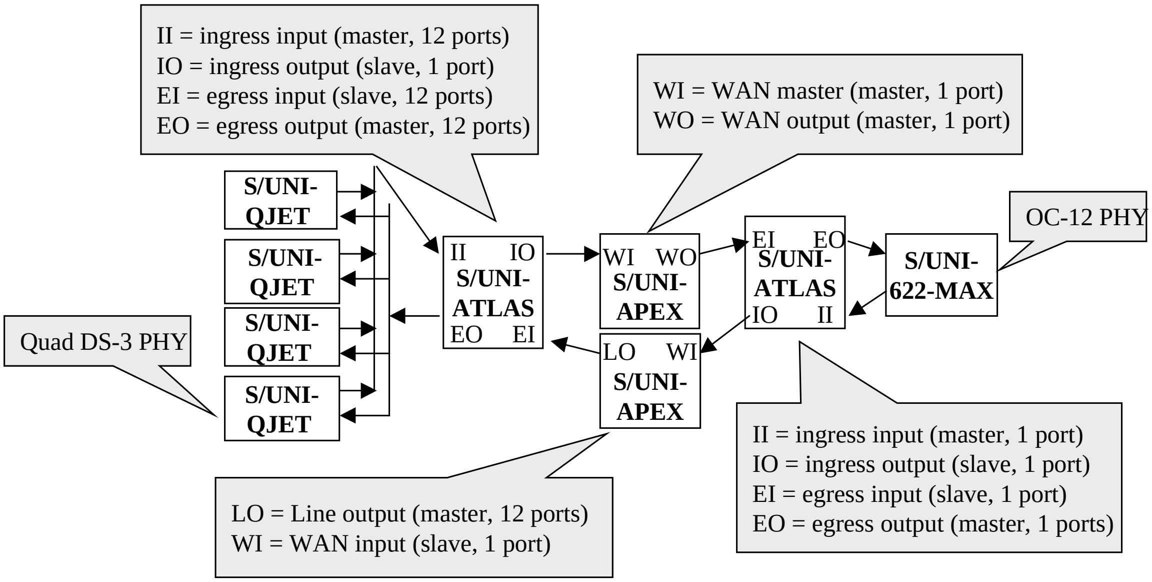
Data: Radio→SAR→Duplex→Vortex→Apex→Atlas→IMA→MPH

Signaling:
Core card microprocessor→ Apex→Atlas→IMA→MPH→Base station controller, or
Radio card microprocessor→ Duplex→Vortex→Atlas→Apex→ Atlas→IMA→MPH→Base station controller, or
Radio card microprocessor→ Duplex→Vortex→Atlas→Apex→Core card microprocessor

6.10 OC-12 (622 Mbps) Architectures

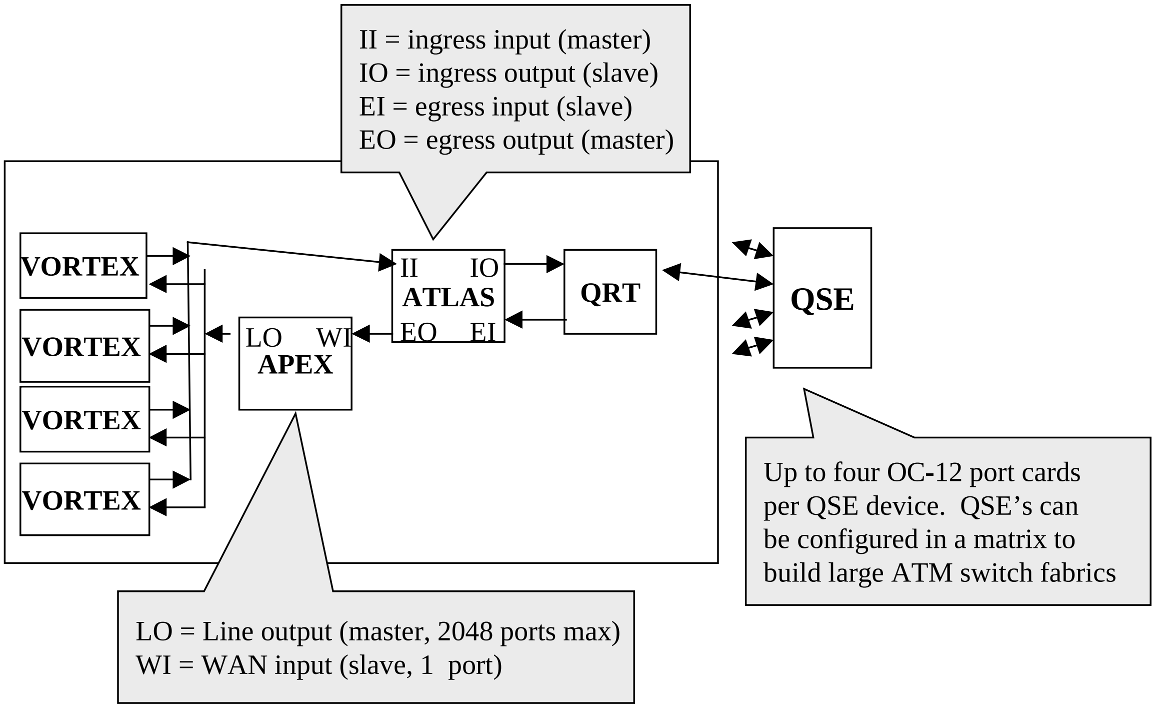
The S/UNI-APEX is capable of processing an OC-12 worth of traffic, with a small amount of speed up, but only in one direction. If line to line switching is not a requirement (i.e. it is purely a multiplexer), two S/UNI-APEX devices can be used to implement a high fan-in OC-12 multiplexer. An example of a twelve DS-3 to OC-12 multiplexer with full I.610 compliant OAM processing is shown in Figure 14.

Figure 14 - OC-12 Multiplexer Architecture


II = ingress input (master, 12 ports)
IO = ingress output (slave, 1 port)
EI = egress input (slave, 12 ports)
EO = egress output (master, 12 ports)
WI = WAN master (master, 1 port)
WO = WAN output (master, 1 port)
OC-12 PHY
S/UNI-QJET
S/UNI-QJET
S/UNI-QJET
S/UNI-QJET
Quad DS-3 PHY
II IO S/UNI-ATLAS EO EI
WI WO S/UNI-APEX
LO WI S/UNI-APEX
EI EO S/UNI-ATLAS IO II
S/UNI-622-MAX
II = ingress input (master, 1 port)
IO = ingress output (slave, 1 port)
EI = egress input (slave, 1 port)
EO = egress output (master, 1 ports)
LO = Line output (master, 12 ports)
WI = WAN input (slave, 1 port)

If the requirement is for a high port count switch port on an ATM switch, then only a single S/UNI-APEX, used as the egress buffer manager, is required. In the ingress direction the S/UNI-VORTEX devices appear as single PHY devices, so an existing ingress switch port controller (for example PMC-Sierra's QRT device) can be used to perform the ingress buffering and switch fabric interface. This is shown in Figure 14.

Figure 15 - OC-12 Switch Port Architecture


II = ingress input (master)
IO = ingress output (slave)
EI = egress input (slave)
EO = egress output (master)
VORTEX
VORTEX
VORTEX
VORTEX
II
IO
ATLAS
EO
EI
QRT
QSE
LO
WI
APEX
Up to four OC-12 port cards per QSE device. QSE's can be configured in a matrix to build large ATM switch fabrics
LO = Line output (master, 2048 ports max)
WI = WAN input (slave, 1 port)

7 GLOSSARY

ARL — Address resolution lookup. ARL is typically performed on traffic as it enters the switch. ARL is the process of taking the source port number of the data unit plus an address fields or fields within the data unit, and performing a lookup to generate a short (typically 16 bit) switch tag or ingress connection identifier (ICI). (See ICI below).

back-pressure — In this document the term back-pressure refers to an indication from the receiver to the transmitter that the receiver's buffer (normally a cell FIFO) is becoming full and the transmitter should hold off transmitting any more cells until the back-pressure indication is de-asserted. For example, on the Utopia bus back-pressure from the PHY to the bus master is indicated via the TCA bus signal.

BOM — Beginning of Message. Normally used to mark the first segment or cell of a multi-cell packet.

cell — A fixed length unit of data transfer. When used in ATM systems cells are standardized as 53 bytes long – 48 bytes of user data and 5 bytes of overhead. Many devices can operate on longer cells (to allow for system overhead).

congestion — A state in which the switch has experienced a long period of ingress or incoming traffic exceeding outgoing or egress traffic.

congestion management — When a switch is in congestion it will typically begin discarding low priority traffic in order to protect the throughput of higher priority traffic. Without congestion management it is impossible for a switch to guarantee quality of service.

context — In switching, an individual VCs context refers to the collection of control information that defines how that virtual connection or data flow is handled. Typical parameters in the VC context are outgoing port ID, queuing class, scheduling weight, etc.

core card — A printed circuit board consisting of the circuitry necessary to multiplex and/or switch data traffic to and from the line cards and the up-link port. Also called the WAN card in this document.

DSLAM — Digital Subscriber Line Access Multiplexer. A C.O. located access concentrator terminating local loops.

EIC channel

Embedded Inter-device Communications channel. A communication channel accessed via the microprocessor ports of the S/UNI-VORTEX and S/UNI-DUPLEX, or via the parallel bus of the S/UNI-VORTEX. It is used to send packets of information between the devices via an embedded channel in the high speed serial link.

EOM

End of Message. Normally used to mark the last segment or cell of a multi-cell packet.

head of line blocking

Head of line blocking occurs when a cell at the head of a queue cannot proceed, and it is stopping cells behind it from proceeding even though in theory those cells could proceed if the blocking cell was not in the way.

ICI

Ingress Connection Identifier, also called a switch tag. A short (two bytes is typical) index value attached to a data cell used used as the direct lookup address of the data structure which defines the switching context under which the cell should be handled.

ICMP

Internet Control Message Protocol. The protocol used by the IP layer to report errors and provide management layer information to IP layer entities.

line card

A printed circuit board on which resides the circuitry necessary to terminate one or more lower speed transmission interfaces.

LVDS

Low Voltage Differential Signal. IEEE 1596.3 standard defining a 4 wire serial transmission format suitable for backplane and short cable transmission. LVDS is the physical layer used between the S/UNI-VORTEX and S/UNI-DUPLEX.

PHY

A layer 1 (physical or transmission layer) device capable of transmitting and receiving a signal carrying cell structured traffic. An example of a PHY is an ADSL modem with a Utopia bus interface.

Policing

Policing is typically performed on ingress (into the switch) traffic in order to measure and record per VC traffic flow and optionally mark data units that exceed the user’s expected service profile. Marked data units may be handled differently by the switch. For example, marked data units may have a lower priority and hence be discarded first under heavy congestion.

POS

Packet Over SONET. An industry standard that defines how byte aligned HDLC encapsulated packets can be mapped directly into a SONET payload.

QoS

In its most general sense, Quality of Service means that some user traffic is handled differently than the rest of the traffic. The impact that QoS requirements have on system implementation is often most significant under heavy load - the type of situation where the traffic is piling up waiting for a slow modem or a congested WAN up-link.

RCA

Receive Cell Available. This is a standard Utopia bus signal that defines the status of a bus slave's receive FIFO. Remember that Utopia defines all signals with respect to the bus master, so RCA asserted means the slave has at least one cell for the bus master to receive.

Speed-up

In this document speed-up is used loosely to describe the bandwidth head-room or spare capacity that the S/UNI-APEX has with respect to its expected WAN port speed. Speed-up is important because it allows the switch to handle traffic bursts.

TC Layer

Transmission convergence layer. Formally define as part of the physical layer, the TC layer maps the data stream onto the physical channel. In ATM systems the TC layer is defined in the I.432 specification.

TCA

Transmit Cell Available. This is a standard Utopia bus signal that defines the status of a bus slave's transmit FIFO. TCA asserted means the slave has room for at least one more cell. The bus master is responsible for polling the TCA line and sending the PHY a cell only when it has room to accept it.

Utopia

A parallel bus specification and standard defined by the ATM forum. It is available from the ATM Forum web site at ftp://ftp.atmforum.com/pub/approved-specs/af-phy-0017.000.pdf.

WAN up-link

A high speed transmission interface used to transport traffic to and from the switch or multiplexer.

WAN card

A printed circuit board consisting of the circuitry necessary to multiplex and/or switch data traffic to and from the line cards and the up-link port. Also called the core card in this document.

NOTES

CONTACTING PMC-SIERRA, INC.

PMC-Sierra, Inc.
105-8555 Baxter Place Burnaby, BC
Canada V5A 4V7

Tel: (604) 415-6000

Fax: (604) 415-6200

Document Information:	document@pmc-sierra.com
Corporate Information:	info@pmc-sierra.com
Application Information:	apps@pmc-sierra.com
Web Site:	http://www.pmc-sierra.com

None of the information contained in this document constitutes an express or implied warranty by PMC-Sierra, Inc. as to the sufficiency, fitness or suitability for a particular purpose of any such information or the fitness, or suitability for a particular purpose, merchantability, performance, compatibility with other parts or systems, of any of the products of PMC-Sierra, Inc., or any portion thereof, referred to in this document. PMC-Sierra, Inc. expressly disclaims all representations and warranties of any kind regarding the contents or use of the information, including, but not limited to, express and implied warranties of accuracy, completeness, merchantability, fitness for a particular use, or non-infringement.

In no event will PMC-Sierra, Inc. be liable for any direct, indirect, special, incidental or consequential damages, including, but not limited to, lost profits, lost business or lost data resulting from any use of or reliance upon the information, whether or not PMC-Sierra, Inc. has been advised of the possibility of such damage.

PMC-1981024 (R2) Issue date: October 2000