11

The Design of BTOP - An ATM Bulk Transfer Protocol

Liam Casey

BNR, PO Box 3511, Station C, Ottawa, Canada K1Y 4H7.
liam@bnr.ca

Abstract

Broadband networks have the raw speed to achieve large data transfers that appear instantaneous to users (i.e. that take less than a second). To realize such speeds in practice requires new protocols. BTOP (Bulk Transfer Oriented Protocol) is such a protocol, designed to be simple and efficient. Not only is BTOP intended to run directly within the AAL-5 layer, BTOP is inspired by the simplicity of AAL-5. AAL-5 was developed in reaction to the perceived complexity of AAL-3 and AAL-4. BTOP's simplicity derives from it only performing those functions that the ATM layer and AAL-5 common part sublayer do not. This paper gives an overview of BTOP and addresses the issues in designing a protocol to specifically exploit the capabilities of ATM.

Keyword Codes: C.2.2;
Keywords: Network Protocols;

1. INTRODUCTION

ATM is often touted as being ideal for bulk transfer applications such as image transfer. A high resolution image can occupy 2 Mbytes or more. Frequently discussed applications for ATM based Broadband networks are medical imaging and advertising pre-press. In the medical imaging arena there is a strong desire to be able to store, retrieve and transfer high resolution medical images in support of a wide range of activities. Within hospitals, for example, the physical transfer of X-rays between the radiology department and the emergency department could be eliminated by providing ATM connectivity between a radiology image server and "electronic light boxes" in the emergency department. Teleradiology applications include remote consultations whereby an image can be jointly viewed and discussed by a specialist in one location (e.g. a tertiary care facility) and a practitioner in another (a primary care facility).

G. Neufield et al. (eds.), *Protocols for High Speed Networks IV*

Applications for the high speed transfer of images are also seen in advertising and civil engineering, where the production of artwork may be geographically distant from clients who want to browse through it. For examples of other image transfer applications see [1].

Images are not the only possible bulk transfer items. If a bulk transfer service were in place offering 10 Mbyte transfers in less than a second, it would be used for transferring large documents and design databases. Such applications currently may not need sub-second responsiveness (waiting 30 seconds to obtain a 1 Mbyte Postscript file from the other side of the continent, using say a T1 circuit, does not add much to the 5 minutes it takes to print the document and the longer time to read it). On the other hand, waiting a minute or more between viewable images on an electronic light box would have a definite productivity impact. Hence image transfer has been chosen as the archetypal bulk transfer application for ATM.

Sub-second bulk transfers cannot be performed with the wide area circuits available to users to-day. But it is not a foregone conclusion that they will be feasible even when an ATM based Broadband ISDN infrastructure is in place. High bandwidth is a necessary, but not a sufficient condition. A careful choice of protocol stacks is required to match ATM's bandwidth. Existing protocols, designed when 9.6 kbps was considered a very high transmission speed and 20 kbytes was the size of a computer's main memory, are not necessarily up to the task. We have designed BTOP (Bulk Transfer Oriented Protocol) explicitly for sub-second bulk transfers over ATM. The overriding design goal of BTOP has been simplicity, so that its execution speed can match ATM's transfer speed.

In the next section we summarize ATM protocol architecture. As an ATM cell is too small to deal with, individual cells are aggregated into AAL-5 PDUs (packets) with a maximum size of 64 Kbytes. In section 3 we first define Application Data Units which can consist of multiple AAL-5 PDUs. We then describe how BTOP transfers an Application Data Unit as a back-to-back sequence of AAL-5 PDUs, acknowledged by a single cell at the end.

Sections 4 and 5 summarize the protocol performance and the underlying assumptions we have made of ATM, such as a very low cell loss ratio. In section 6 we discuss some of the design guidelines we adopted. ATM provides an unique set of services and we show that BTOP's simplicity arises from the design decision not to duplicate services in higher layer protocols. Finally we look at implementation issues, then briefly contrast BTOP with existing protocols.

BTOP is in the definition phase. Our purpose here is to describe the design considerations of a truly high speed protocol. Not all the supporting assumptions we make will be true of initial, limited, ATM deployments. But, unless issues of protocol throughput are addressed, the practicability of new, data transfer intensive applications, justifying ATM's widespread adoption, is in doubt .

2. BROADBAND PROTOCOLS

The ITU-T (formerly CCITT) has been the primary source of Broadband protocols. This Standards body has determined that ATM will be the basic transport mode of Broadband ISDN. The ATM Forum exists to promote the use of ATM in both public and private networks. It bases the protocols it develops on ITU-T recommendations. Broadband protocols do not fit exactly into the ISO seven layer OSI stack. There are two general areas of difference: the 3-plane model and the unspecified upper layers.

2.1 The three planes of BISDN

The ITU-T has adopted a 3 plane model for protocol stacks using ATM, depicted below in Figure 1 [2]. The ATM Forum has also adopted this model [3].

Briefly the three planes are:

U-plane: The User plane is where applications usually work. It contains the protocols for the transfer of normal user data. BTOP is in the U-plane.

C-plane: The Control plane contains the signalling protocols for the "out of band" establishment of connections. In BISDN, ATM connections, called VCCs (Virtual Channel Connections), are established using the C-plane Q.2931 (formerly Q.93B)/ATM Forum UNI Signalling [3] protocol stacks.

M-plane: The Management plane manages the exchange of information between the C-plane and the U-plane and generally provides management functionality. For example, detecting and reporting physical link failures is an M-plane responsibility

For more details see [3] or a text such as [4].

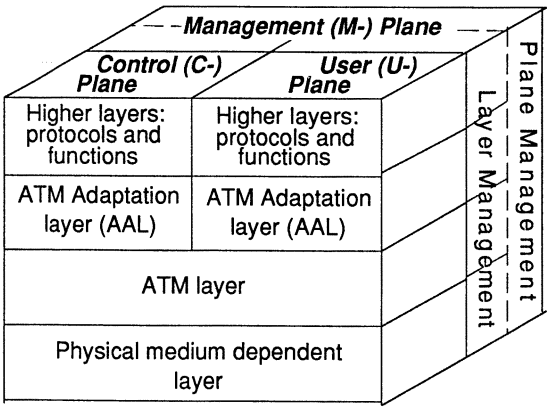

Figure 1: B-ISDN Protocol Model for ATM

Designing a protocol within the context of the three plane model is considerably simpler than designing classic, single stack protocols. For example, although BTOP is connection oriented, unlike existing transport protocols such as TCP [5], it does not have its own connection establishment phase. Applications using BTOP are responsible for establishing the VCC to be used. They may do this using C-plane signalling facilities. Alternatively applications may use permanent virtual connections (PVCs) pre-established using M-plane facilities. The latter option may be appropriate for "dumb" terminals such as electronic light boxes.

2.2 Layering - where BTOP fits.

As shown above in figure 1, there are several layers defined for both the C-plane and U-plane protocols. A particular protocol stack will consist of a Physical Medium Dependent layer, an ATM layer, an ATM Adaptation layer and higher layers. The ITU-T has defined Q.2931 as a higher layer in the C-plane but has not defined U-plane higher layers. No correspondence between the top of the ATM Adaptation layer and a position in the OSI stack has been stated. This is not surprising since the purpose of the Adaptation layer is to adapt the service offered by the ATM layer, viz. the basic relaying of cells, into the service required by the Upper layer. Thus different kinds of Upper layers require different Adaptation layers.

The ITU-T has specified 4 ATM adaptation layers, AAL-1, AAL-2, AAL-3/4 and AAL-5, each for use in providing different services. AAL-5 is an ATM Adaptation layer intended for the transfer of multi-cell data packets (called PDUs henceforth). As depicted below in figure 2, AAL-5 is itself divided into sublayers.

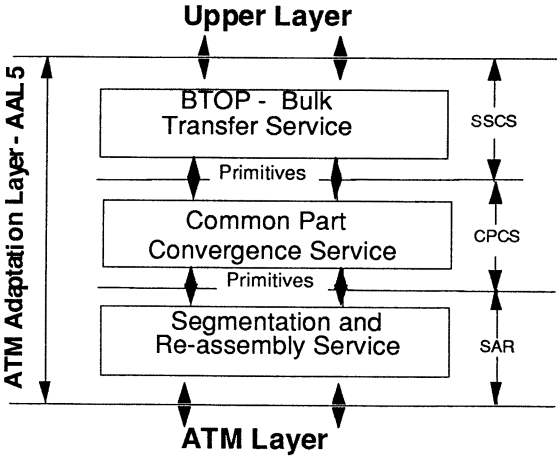

Figure 2: BTOP positioned as AAL-5 sublayer

Looking from the bottom up (as when cells are received) the sublayers and their functions are:

SAR — The Segmentation and Re-assembly Service. On the reception side this sublayer takes the contents of cells delivered by the ATM layer and concatenates them into bigger PDUs.

CPCS — The Common Part Convergence Service. On reception this sublayer takes the concatenated cell PDU (called AAL-5 CPCS PDU or basic AAL-5 PDU) and checks that it has been received correctly. It does this by examining a trailer that is carried in the last 8 bytes of the last cell of the PDU (cf. figure 3). This trailer was generated by CPCS on the sending side. The CPCS function is common to all classes of AAL-5.

SSCS: — The Service Specific Convergence Service. The functionality of this sublayer depends on the nature of the Upper layer. It implements all other services that specific Upper layers assume will be provided.

The ITU recommendation I.363 [6] leaves the SSCS unspecified - different SSCS are to be defined for different services. BTOP is the protocol of a Bulk Transfer (BT) service. Bulk Transfer is a "specific" Service Specific Convergence Sublayer of AAL-5. Other, already defined SSCSs include the null one (when the common CPCS provides all the services needed by the Upper layer) and SSCOP, which has been defined to support signalling applications.

In the case of BTOP the Upper layer is the Application layer (unless compression or special encoding are required in which case a Presentation layer would be interposed between Application and BTOP). The Bulk Transfer Convergence Sublayer transfers large Application layer PDU's without requiring any intervening layers between it and the Application.

We attach great importance to placing BTOP within the context of BISDN planes and protocol levels. Only by examining the total picture of the services each plane and level offers can we minimize the functionality that we need to provide in BTOP. For example, BTOP does not provide any flow control because this is provided by the ABR service of the ATM layer (see section 6).

3. OVERVIEW OF BTOP

3.1 Application Data Units

A Bulk Transfer is the transmission of a large Application Data Unit from a Source application to a Destination application. An Application Data Unit is a chunk of data that makes sense at the application level. The length of transport PDUs (packets) in most networks exhibits a bimodal distribution. There are short packets, carrying either protocol information (e.g. acks) or application control (e.g. a get file request), and there are maximum size packets. A sequence of maximum size packets are used to effect a large data transfer,

such as a file or an image. What is really being transferred is an Application Data Unit, although existing Transport protocols do not recognize any unit above the transport PDU.

In [7] an Application Data Unit is defined as an aggregate of data that an application can process out of order. Our definition is similar: the aggregate of data that an application processes as a complete unit, in a single processing step. An image, for example, would be considered one Application Data Unit. For the applications most likely to use BTOP, the transfer of an Application Data Unit will be triggered by some form of user request (e.g. pushing a button to view the next set of X-rays).

Although Application Data Units could be of any size, for practical reasons we have limited them to 16 Mbytes. 16 Mbytes meets current needs, for example an uncompressed 2000 x 2000, 24 bit colour image occupies 12 Mbytes. A protocol to deal with multiple Application Data Units could be defined but we have chosen not to elaborate upon it in this paper (c.f. NETBLT [8] which deals with multiple bursts).

3.2 Source-Destination Associations

Before any transfers can take place there has to be a VCC established between the Source and Destination. In general we assume that the Source and Destination applications will have some form of association established between them (with its own protocol stack) to signal between themselves what information is to be transferred and when the transfer is to occur.

The nature of the signalling association is not of particular concern to BTOP. It could be "user to user" signalling in the C-plane. If in the U-plane, the protocol would probably run on a different VCC.

BTOP is not intended for the unsolicited transfer of indeterminate amounts of data. We assume that the Destination knows the size of an Application Data Unit before it is transmitted. In general we suppose that there will have been a signalling exchange between Source and Destination prior to invoking Bulk Transfer service. This enables the Destination to prepared itself for the reception of the Application Data Unit.

One particular scenario has Bulk Transfer taking place during the service phase of a remote procedure call (RPC) made from the Destination, invoking the Source. If a protocol such as ROOP [9] were used, then it might be possible that the same VCC could be used for both RPC "signalling" and bulk transfer.

3.3 Protocol and BT-PDU types.

BTOP packets (BT-PDUs) come in two classes, data and control. Both classes of BT-PDU have the format of AAL-5 CPCS PDUs. The class of BT-PDU is indicated by the CPI (Common Part Indicator) field in the CPCS-PDU Trailer (see figure 3). Four CPI values will have to be reserved to indicate different data and control types. Control BT-PDUs carry no payload and so have a Length of zero and fit into a single ATM cell. The use of the CPCS-UU (CPCS User to User Indication) is described below.

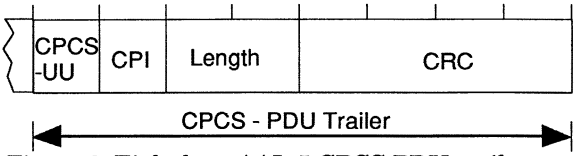

Figure 3: Eight byte AAL-5 CPCS PDU trailer

When a source BT entity has an Application Data Unit to send, it segments it into DATA PDUs. Each DATA PDU, except usually the last, is maximally sized for the VCC being used between the Source and the Destination. The DATA PDUs are each given a decrementing sequence number, carried in the CPCS-UU field. The last DATA PDU has a UU value of zero. All DATA PDUs are passed down to the ATM layer to be transmitted back-to-back. That is, the entire Application Data Unit is transmitted as a single burst of cells, subject only to traffic management policies applied at the ATM layer (see below). The entire burst is acknowledged with a single Received OK control PDU (1 cell) after the last DATA PDU been received. Figure 4 depicts the normal BTOP message sequence when the Application Data Unit occupies n DATA PDUs.

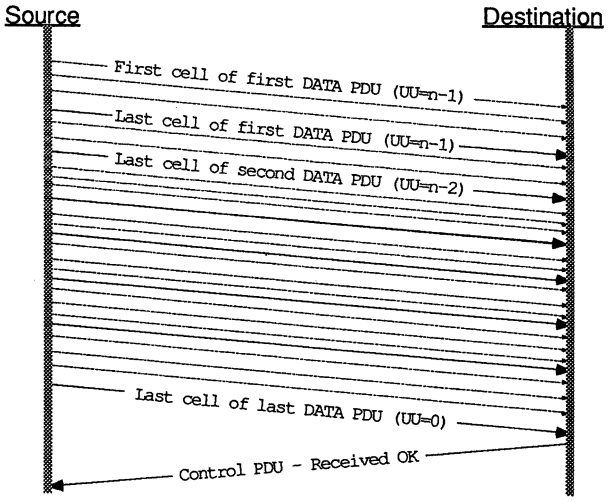

Figure 4: Cell sequence chart for error free Bulk Transfer

If there is a missing or corrupt cell in a data PDU then the destination uses a RESEND control PDU to request that the entire data PDU be retransmitted. The RESEND control PDU is a single cell AAL-5 PDU containing the sequence number of the corrupted data PDU in the UU field. (Note that under the

assumptions of cell loss rate described below, losing two or more DATA PDUs in a burst will be extremely rare - the overhead of sending a separate RESEND PDUs for each lost PDU is not significant). The Source is expected to respond with the missing data PDU. Re-transmitted data PDUs have a different CPI value (RDATA) from original DATA PDUs, to avoid potential confusion with delayed responses.

As discussed later (in the section on Implementation Issues) rare occurrences, such as the loss of the Received OK control PDU can be protected against by timers. For full details of the BTOP protocol see [10].

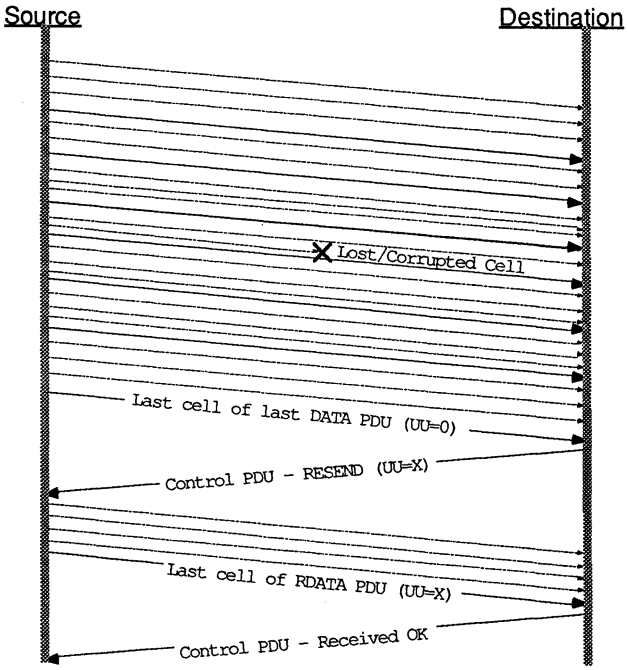

Figure 5: Cell sequence chart showing resent data PDU

3.4 BTOP APIs

Application Programming Interfaces (APIs) could be developed for applications to access BTOP's Bulk Transfer service. A basic API would offer Bulk_Send and Bulk_Recieve primitives, for the Source and Destination respectively. This basic API would be applicable where there are in-memory buffers to hold the entire burst. As noted above the maximum size of Application Data Unit that BTOP treats as a single burst is 16Mbytes. 16Mbytes is not beyond the capacity of today's workstations to store directly in primary memory. Simple devices, such as electronic light boxes in medical imaging

applications, might use the Bulk_Recieve primitive to transfer an incoming image directly from the network into a display frame buffer.

The basic API for Bulk Transfer service might look like:

```
type Ind_Code is (OK, NoVCC, BufferTooBig, ...etc.);

type Buf_Address is new System.Address;

type Buf_Length  is range 1..16M;          -- In bytes

type Up_Call is access procedure(Result : Ind_Code);

procedure Bulk_Send_Request
               (B_Handle :In  Buf_Address;
                B_Size   :In  Buf_Length;
                Result   :Out Ind_Code);

procedure Bulk_Recieve_Request
               (B_Handle :In  Buf_Address;
                B_Size   :In  Buf_Length;
                Indication  : Up_Call);
```

The basic Bulk Transfer service copies data from the Source buffer onto the ATM network as cells. At the Destination the data is moved into the buffer specified by the Destination as it arrives. A call-back routine is used to notify the Destination application of the completion of the transfer. The Source application is assumed to block while waiting for the transfer to complete, although an asynchronous version of Bulk_Send could be provided also. If any retransmissions are needed, the Bulk Transfer service re-transmits the required data directly from the buffer without any intervention from the application.

Since BTOP processes an Application Data Unit in the predictable access pattern of first byte to last byte, there is no impediment to an implementation using it for data that is held on secondary storage and pre-loaded piecemeal into memory as needed. The BTOP protocol itself would not be affected, but extra API hooks would be needed to handle a chain of smaller buffers and to allow BTOP to re-access any data that has to be re-transmitted because of cell losses. Note though that re-transmissions will only occur after the first pass through the Application Data Unit is completed, and that the response time in providing data then is not critical to overall throughput.

4. SUPPORTING ASSUMPTIONS

BTOP relies on a number of properties particular to ATM and AAL-5. These are listed below.

4.1 End-to-End ATM Service.

Today ATM cards are available for PCs and workstations. ATM to the desktop is strongly emphasized in the work of the ATM Forum. Rather than carry extra network protocol overheads to cope with non-ATM stations

accessing a network though routers and gateways, BTOP has been designed for deployment where ATM service is end-to-end.

4.2 AAL-5 CPCS Properties

The CPCS service used by the BT sublayer Source and Destination entities is the Non-assured Message Mode service [6]. It, in conjunction with the underlying ATM layer, has the following properties:

- No re-ordering: The delivery order of PDUs on a particular VCC is the same as the sending order (i.e. misordering is negligible).
- No duplicates: CPCS does not duplicate PDUs and the underlying ATM layer does not duplicate cells (i.e. the duplication rate is negligible).
- Notification of corrupt PDUs. Cells may be corrupted or lost (infrequently). CPCS detects the resulting errors in PDUs. An option in the AAL-5 specification allows CPCS to deliver PDUs that are corrupt. BTOP assumes that it is notified when a PDU has arrived (see below in the section on Implementation Issues).

4.3 VCC and Source-Destination Association management.

There is no provision in the BTOP protocol for detecting the loss of a connection or the termination of either end. Rather the Burst Transfer service relies upon M-plane notifications.

- Should either the Source or Destination fail, it is a Layer Management function to close down the VCC.
- When the VCC is closed down or fails (including underlying transport failure) it is a Layer Management function to notify both the Source and the Destination, including the Burst Transfer entities.

5. PROTOCOL CHARACTERISTICS

5.1 Burst size

The maximum payload length of an AAL-5 PDU is 65535 bytes (64 Kbytes less one byte). The size of the UU field limits the total number of DATA PDUs in a burst to 256. Thus BTOP can handle Application Data Units of just less than 16 Mbytes. 16 Mbytes provides a much larger unit of transfer than current protocols, while remaining within the capacity of current network and workstation technology.

5.2 Error Handling

BTOP is designed on the assumption that the cell loss ratio (CLR) of the underlying ATM transport is very low. Assuming a CLR of less than 1.7E-10, (c.f. [11]), fewer than one in 4 million data PDUs will be corrupted. With 256 data PDUs in a maximum sized Application Data Unit (16 Mbyte) we would expect fewer than one in every 16 thousand Bulk Transfers to encounter a cell loss on the first pass.

The error checking and recovery unit in BTOP is the AAL-5 CPCS PDU. A 32bit CRC is provided in the CPCS sublayer that protects the entire payload of the AAL-5 PDU including UU and CPI fields.

BTOP uses selective retransmission. The Source only services RESEND requests once it has transmitted the entire first pass burst. A design option was to have the Destination issue the RESEND control PDU as soon as it detected a corrupted DATA PDU. But since loss of a DATA PDU will be very infrequent, this optimisation has been ignored in favour of simplifying state machines and timer requirements. Thus the retransmission of each missing DATA PDU is requested and acknowledged serially, after the last DATA PDU has been received.

5.3 Performance

In normal operation one Received OK PDU (a single cell) acknowledges the entire transfer (i.e. up to 16 Mbytes of application data). The transaction is completed in the transfer time plus one round trip delay (RTD).

If a cell is lost in the transfer then, with the selective retransmission strategy described above, another round trip delay will be added before the transfer is completed.

With these delay characteristics BTOP is well suited for wide area transfers, as well as local area transfers. In the wide area the RTD can be a significant portion of the transfer time even for relatively large Application Data Units (e.g. a 3000km separation of Source and Destination gives a RTD of ~ 33msecs, which is more than the transfer time for 0.5 Mbytes at 150 Mb/s). In the wide area BTOP will perform significantly better than protocols that require multiple round trips.

6. DESIGN CHOICES

Not only is BTOP intended to run directly on top of the AAL-5 CPCS, BTOP is inspired by the simplicity of AAL-5. AAL-5 was developed by the computer industry in reaction to the perceived complexity of AAL-3 and AAL-4 [12]. To achieve the goals of simplicity and efficiency, BTOP's design was developed with the following three guidelines:

- focus on one function
- do not perform functions that are already done in other parts of the protocol stack
- exploit the special characteristics of ATM

BTOP is focused on just the transfer of large Application Data Units. It is proposed that other functions such as remote procedure calls use other protocols, cf. [9]. Below we look at the application of the second and third guidelines.

6.1 Exploiting Connection Orientation

An important aspect of BTOP's simplicity arises from it being connection oriented. ATM is inherently connection oriented (53 byte cells are too short to carry a full source and destination address in each one!) and BTOP exploits that fact. AAL-5 offers a very simple end-to-end connection oriented protocol. The AAL-5 connection is directly provided by the ATM layer VCC (Virtual Channel Connection) between Source and Destination.

Any connection oriented protocol has the overhead of connection establishment. As noted above for BTOP a VCC needs to be pre-established between Source and Destination using C-plane or M-plane facilities. The virtue of connection orientation is that there is a lot of information that needs to be exchanged between Source and Destination only once, not with every PDU transfer. For BTOP the following assumptions are made:

- At or before connection establishment there is agreement on the maximum size of AAL-5 PDU and on the maximum transfer rate.
- At or before connection establishment the Source and Destination have satisfied themselves as to the identities of their respective partners (i.e. authentication, if required, is performed as part of VCC setup, not on a per transfer basis). Likewise any encryption key exchange occurs at or before connection establishment.
- During connection establishment the network is informed of the class of service and QOS required (e.g. ABR).

To summarize, all fixed attributes for Bulk Transfer between Source and Destination are set before or during connection establishment. No fields are used in BT_PDUs to convey these attributes. Consequently no processing is needed to generate the fields on PDU transmission, nor to check their validity upon reception.

6.2 No multiplexing

Multiplexing is cheap at the ATM layer. There is generous provision for multiple VCCs between two end systems. Cell flows that are part of different associations between entities on the end systems can be given their own separate VCCs, each tailored to their own traffic characteristics. There is no need for multiplexing different streams of data at protocol layers above the ATM layer, and BTOP makes no provision for it. There is only one Service Access Point (SAP) for the BT Service per VCC. The single SAP is identified by the VCC. There is no sublayer addressing.

Neither is there any overlapping of transfers. For a given instance of the Bulk Transfer Service (i.e. a given VCC) there can be only one outstanding Bulk Transfer in progress at a time.

These design decisions result in there being no need for sub-address fields and transaction identifiers in BT-PDUs.

6.3 No Destination Flow Control

The basic capability of BTOP is to transfer up to 16Mbytes of data from memory to memory. Modern systems that want to perform large data transfers can easily afford to buffer the whole burst in memory. There is little point in buying high speed network interfaces if the attached system can not keep up with the line rate. Hence, while sender and receiver can agree on a maximum transmission rate when a connection is established, BTOP does not provide for destination flow control during a Bulk Transfer (although the definition of ABR might permit it - see below).

6.4 Network Flow Control via ABR

The ATM Forum's Traffic Management Subworking group is working towards defining an ATM layer service called ABR (Available Bit Rate) service. ABR is intended to be the ATM analogue of the service that a MAC layer offers in a LAN. In a LAN the available bandwidth is not permanently divided between all stations, rather each station gets the full bandwidth when it needs it. The media access control mechanism determines access opportunities when multiple stations contend for the bandwidth.

At the time of writing, the specification of the ABR service has yet to be nailed down, and no mechanism has been chosen for its implementation. However the general nature of ABR is clear. Stations will be able to use a substantial part of their access bandwidth to transmit bursts. In threatened congestion situations, ABR delays the transmission of cells rather than discarding them. Some form of protocol will exist to control the flow of cells in the ATM network. Thus, unlike CBR (Constant Bit Rate) and VBR (Variable Bit Rate) services, ABR will not give a tight guarantee of cell delay and cell throughput. ABR will however offer a cell loss rate close to that of the underlying medium, similar to CBR.

Given that ABR will be provided at the ATM layer, not only is it unnecessary to provide a flow control service within BTOP, it would be very undesirable to do so. Any flow control mechanism installed above ABR is likely to interfere with it in ways that are difficult to analyse and will certainly decrease efficiency. Hence BTOP relies on the ATM layer ABR service for flow control. The BTOP protocol itself does not do anything to regulate its rate, or avoid congestion during a burst.

Note that, under the assumption that the Destination can receive a burst at full rate, there is no impediment to using BTOP with CBR service. Of course, since CBR bandwidth is reserved for the lifetime of the VCC, either a VCC will have to be established and held just for each Bulk Transfer or bandwidth will be wasted between bursts.

6.5 Using large packets

In legacy networks the maximum size of packets is quite small, as low as 256 bytes. This arose in part from concern about the buffer space needed at source and destination, and in part from trying to optimize packet size for the likely transmission error rates. Such small packets cause excessive processing overhead and limit overall throughput.

Since ATM networks have a very low inherent transmission error rate, the maximum (payload) size of AAL-5 PDUs has been set at 65535 bytes. At this size the overhead is manageable (265 PDUs per second on a 150 Mb/s link) even if a processor has to get involved with each PDU (but see below) .

7. IMPLEMENTATION ISSUES

7.1 Hardware realization

BTOP is designed with future so-called "desktop area networks" in mind. To avoid interrupting a high performance CPU and moving data through main memory, bulk transfers may be direct from peripheral device to device (e.g. from disk controller to frame buffer). BTOP is intended to be fully realizable in hardware.

The key to hardware implementation is simplicity. Bulk Transfer across a network should be no more complex than say a chained SCSI transfer from disk to main memory. As we described above, connection orientation and the elimination of options has allowed us to produce a protocol that does not add any more fields to the basic AAL-5 PDU. BTOP needs to process just two otherwise unused AAL-5 fields.

Basic AAL-5 is already encapsulated in hardware. The state machines for BTOP should be easy to integrate (we do not expect a silicon implementation to reflect the layering of figures 1 and 2). We have eschewed high efficiency in low-runner situations in order to simplify implementation. For example the PDU re-transmission scheme, with retransmitting beginning only after the entire burst has been sent (and dealing with only one PDU at a time), was chosen so that Destination hardware can be set up for direct copying from cell buffer to final destination.

7.2 Dealing with special cell losses

If Application Data Units are going to be built up "in situ", then it is important that the Destination BTOP state machine have an accurate view of the sequence number of an incoming data PDU before the trailer arrives. This allows the calculation of the final location into which the contents of each cell of the data PDU is to be directly copied. To do this the state machine needs to be notified of the end of each data PDU even if the final (specially marked) cell is lost or corrupted.

The corruption of intermediate DATA PDUs due to loss or corruption of their final cells does not cause particular problems. When the final cell of an intermediate PDU is lost, the first or second cell of next data PDU should cause the SAR layer to indicate "Re-assembly Buffer Overflow", from whence it can be deduced that the final cell was lost. (Note that in all likelihood the following DATA PDU will not be received correctly, resulting in the retransmission of two DATA PDUs).

Loss of the final cell of the final DATA PDU is more problematic. BTOP requires that a control PDU (either Received OK or RESEND) be generated by

the Destination after a burst has been received. Completion of the burst is signalled by the final cell of the last PDU. When the final cell of the last DATA PDU or a RDATA PDU is lost there are no more cells to trigger Re-assembly Buffer Overflow and so the normal notification of PDU completion would not occur. The final cell of the last PDU is a special cell in BTOP.

Likewise the cell that constitutes a control PDU is a special cell: its loss, if not provided for, will cause a Bulk Transfer to "hang".

The simplest implementation approach is to ignore the loss of these special cells and let the Bulk Transfer remain in an uncompleted state until a user intervenes and resets the ATM connection. This approach is very attractive if ATM connections meet the specified cell loss ratio of 1.7E-10 mentioned above. With such a low loss rate and performing a Bulk Transfer every second (i.e. generating two special cells per second), a connection will hang only once every 95 years. This is far less frequent than most disk controllers hang.

A more conservative approach requires timers. Dealing with timers adds expense to an implementation, especially if the timers have to be manipulated on every cell. AAL-5 allows an optional SAR re-assembly timer, to detect partially filled PDUs that are not going to be completed. Any SAR re-assembly timer value must be very loose or it will interact negatively with ABR service, which delays cells rather than lose them. Using a SAR re-assembly timer gives almost complete coverage for detecting the last DATA PDU, but it does not cover single cell control PDUs or last DATA PDUs that happen to consist of a single cell. Our recommendation is to use a subsequent application request to send or receive an Application Data Unit as an opportunity to detect a hung connection. If when a new request is received, it appears to the BT layer entity that a current transfer could be hung then it should set a long timer. Should the timer expire the BT layer entity should tidy up and then proceed to process the new request.

8. EXISTING PROTOCOLS

Existing protocols that are used for Bulk Transfer are neither tailored to, nor appropriate for, ATM. They were designed for connectionless transports with small packet sizes, high packet loss rates and low bandwidth-delay products.

TCP [5] is the protocol used by FTP. FTP is the most commonly used bulk transfer protocol today. TCP establishes its own connection on top of connectionless service (IP). In a Broadband environment, IP in turn would be carried on top of AAL-5 connections either directly or over a LAN emulation layer. This layering is very inefficient and completely unnecessary in an environment that is end-to-end ATM. There are so many protocol fields to be filled in that even a simple "ack" will not fit in a single cell. BTOP is considerably simpler.

With TCP, retransmissions are of the "go-back-n" variety. This is not efficient in Broadband networks with their high bandwidth delay product.

BTOP uses selective retransmission. TCP also has a slow start windowing scheme. This mixes up network congestion control, error recovery and destination flow control. Slow start can double or triple the transfer time of a burst in a wide area network. There are also questions as to how it will interact with ABR flow control. BTOP leaves congestion control to the underlying ABR service.

XTP [13] was designed for high speed protocol processing but is too general. In [14] Kure and Sorteberg examine the problems of running XTP over ATM. XTP is not targeted just at bulk transfers. It has both rate and flow control schemes, and a priority scheme for handling interleaved packets on the same connection (which it manages itself). All of these features are present in ATM with ABR service and so BTOP does not duplicate them. Duplicating them in XTP result in extra processing and PDU space overheads. XTP was also designed to work over a number of link layers, with concomitant protocol overhead. However both XTP and BTOP have the right flavour of retransmission scheme (i.e. retransmit missing blocks).

NETBLT [8] is explicitly targeted at the rapid transfer of a large quantity of data between computers. The main concerns of the protocol are to control the rate of the sender so as to avoid network congestion and client overrun. It is intended to run over IP and performs its own connection establishment and maintenance. It too uses selective re-transmission. In some ways BTOP can be regarded as the re-engineering of NETBLT's burst transfer for the ATM environment - discarding IP, and exploiting ATM's connection oriented services.

In [1] Turner and Peterson specifically study the transfer of images. Their work is predicated on there being a significant rate of packet loss in a network (1% or more). They show that compression/encoding schemes that allow reconstruction of missing parts of an image can be more effective than retransmitting missing packets. In BTOP we too have attempted to do an end-to-end design but we have not considered interpolated reconstruction necessary given the relatively lossless nature of ATM. BTOP is not confined to just images. In the image domain it will work with any compression scheme but is particularly appropriate for applications that are intolerant of any artefacts in images.

9. SUMMARY

ATM will provide high bandwidth over long distances. This could be exploited to give sub-second transfer times for large Application Data Units, such as an uncompressed image, but to do so will require new protocols. BTOP is such a protocol that exploits the unique properties of ATM. Existing protocols were designed for considerably less powerful networking technology and are inefficient when run over ATM networks.

BTOP has been designed to reduce the number of layers involved in performing a Bulk Transfer. It offers service directly to applications and runs on top of basic (CPCS) AAL-5. AAL-5 runs on top of the ATM layer. BTOP

achieves high efficiency and low overhead by not performing any function done in another protocol layer.

BTOP operates in the context of the ATM three plane model. BTOP relies on the C-plane functionality to establish and tear down the VCC it uses. BTOP does not have any "connection establishment" phase of its own. BTOP exploits connection establishment to set up all required parameters, authentication etc. rather than performing these functions for every transfer, or, even worse, for every constituent packet. The ATM layer is assumed to provide ABR service to take care of network congestion. BTOP uses ABR, it does not compete with it.

BTOP transfers Application Data Units across an ATM network as a sequence of AAL-5 PDUs. A positive acknowledgement is returned when all PDUs have been successfully received. Selective re-transmission of an AAL-5 PDU is invoked if any of its constituent cells should go missing or be corrupted.

BTOP is the first of what we hope will be a growing number of protocols that attempt to match an application area directly with the capabilities of ATM [15]. We regard this as a more worthwhile activity than contorting ATM to handle old protocols past their prime.

REFERENCES

[1] Image Transfer: An End-to-End Design, C. Turner and L Peterson, Proc SIGCOMM '92 (Baltimore) as ACM SIGCOMM Vol 22 No 4 Oct 1992 pp 258-268.

[2] CCITT (now ITU-T) Recommendation I.121: Broadband Aspects of ISDN, IXth Plenary Assembly (Melbourne) Nov 1988.

[3] ATM User-Network Interface Specification, Version 3.0, The ATM Forum, Mountain View CA, Sept 93, (to be published by Prentice Hall).

[4] *Asynchronous Transfer Mode: Solution for Broadband ISDN*, Martin de Prycker, Ellis Horwood series in computer communication and networking, Chichester, 1991.

[5] Transmission Control Protocol - RFC 791, Defense Advanced Research Projects Agency, Arlington Va, Sep 1981

[6] AAL Type 5, Draft Recommendation text for section 6 of I.363, ITU-T SWP XVIII/8-5, Geneva Jan 1993.

[7] Architectural Considerations for a New Generation of Protocols, D Clark and D. Tennenhouse, Proc SIGCOMM '90 (Philadelphia) as ACM SIGCOMM Vol 20 No 4 Sept 1990 pp 200-208.

[8] NETBLT: A Bulk Data Transfer Protocol - RFC 998, D Clark, M Lambert and L Zhang, MIT, Mar 1987.

[9] ROOP - Remote Operations straight on top of AAL-5, L Casey, ATM_FORUM/94-0083, (Lake Tahoe) Jan 1994.

[10] BTOP Specification, L. Casey BNR CRL Internal Report 94123, Jun 1994.

[11] TA-NWT-001110: Broadband ISDN Switching System Generic Requirements, Issue 2 Bellcore, Aug 1993.

[12] The Development of ATM Standards and Technology: A Retrospective, R.Vickers, IEEE Micro Vol 13 No 6, Dec 1993, pp 62-73

[13] XTP Protocol Definition, Rev 3.1. G. Chesson et al, Protocol Engines Inc, Santa Barbara, 1988

[14] XTP over ATM, Ø. Kure & I. Sorteberg, Computer Networks and ISDN Systems Vol 26, 1993, pp 253-262.

[15] BTOP - A Bulk Transfer protocol over AAL-5, L Casey, ATM_FORUM/94-0082, (Lake Tahoe) Jan 1994.