A Multiple Time Scale, Aggregate Source Model for Edge Traffic in the Next Generation Network

Paul Plantinga, Hu Hanrahan
Centre for Telecommunication Access and Services
University of the Witwatersrand, Johannesburg
Email: {p.plantinga, h.hanrahan}@ee.wits.ac.za
Tel: +27 11 7177250; Fax: +27 11 4031929

Abstract—Edge or core-ingress traffic in the Next Generation Network (NGN) is generated by the hierarchical aggregation of multiple, heterogeneous client flows. We model the edge traffic arrival process for an aggregated multi-service user group. The model quantifies the relationship between traffic shaping factors: user group behaviour and service properties, multiplexing buffering and scheduling architecture, and the resulting multiple-time-scale burst characteristics of the aggregated, edge traffic. An initial model analysis shows significant impact due to Class-based queueing (CBQ) at the voice burst time-scale, and a predictable relationship between changing user-group traffic mix and aggregate traffic burstiness. Further usage of modelling includes determining quantifiable impacts of additional architectural variations and the performance implications of multi-timescale burstiness in queueing systems.

Index Terms—Multi-scale, Source Model, Aggregate, Arrival Process

I. INTRODUCTION

An important aim in the NGN is to optimise and manage resource allocation in the core/transport network such that specific resource layer QoS guarantees may be given for core tunnels. The traffic in each tunnel consists of source contributions that are aggregated into forwarding classes (FC) according to allocation parameters including priority and packet size. Each forwarding class requires a performance-guaranteed, isolated network path during multiplexing. Network resource management responds to changes in aggregate arrival traffic, at multiple time scales, in multiple classes to maintain stable, predictable and efficient transport network performance.

An element of a network’s performance is described from a QoS perspective. We can predict specific levels of QoS by modelling the packet loss probability, delay, and jitter of a buffer network transporting a modelled aggregate traffic arrival process. The aim of this paper is to parsimoniously describe the edge traffic arrival process for a defined user-access-group, such that a core network resource control system may tune, re provision or reconfigure its architecture and control the admission of flows as the user-group profile changes. We reduce the complexity of the multi-time-multi-layer user demand profile to representative busy periods, at multiple timescales (within the busy period), at the packet multiplexing layer.

The Centre is supported by Telkom SA Limited, Siemens Telecommunications, Sun Microsystems SA and the THRIP Programme of the Department of Trade and Industry

The objective of this paper is to quantify how the following factors influence the multi-timescale properties of aggregated source flows: (1) the implementation of existing (or proposed) services in a multiservice user environment, (2) the relative penetration of a service as a percentage of aggregate traffic mix, and (3) scheduling and shaping variables in flow multiplexing. Specifically, the profile aims to describe the burstiness of edge traffic for multiple forwarding classes, at multiple time scales.

Section II introduces the components driving an edge traffic profile and the scope within the NGN context. In Section III we first discuss factors driving the source traffic in part A and second, factors shaping the aggregated traffic in part B. Our simulation implementation using ns2 [1], and time-variance analysis of resulting trace files are discussed in Section IV.

II. THE EDGE TRAFFIC PROFILE WITHIN AN END-TO-END NGN TRAIL

Figure 1 shows the components of an end-to-end NGN trail connecting network customers. A trail extends across subnetworks and core transport networks and is constructed from several source and multiplexing objects which may be modelled independently. An end-to-end trail model may be used to determine an individual service class and trunk traffic characteristics at different points on the route, and predict the impact of resource management or admission control on QoS in the core and distribution networks. In addition, the model components may be implemented as a data model abstraction to management interfaces similar to the proposed implementation of DMTF’s Common Information Model (CIM) Device and Network
Models [2]. The Logical Port and QoS Service abstractions in [2] are associated with statistics classes that collect information on traffic flowing through network nodes. We envisage a similar but more extensive contribution to the network management platform using aggregate traffic models.

We model the trail from User Group to Core network edge (PE) in Figure 1 to determine the characteristics of the arrival process into core network tunnels. A trail is constructed from path or interface sections linking, for example, CE and PE. Input and output interfaces queue, multiplex, transmit and receive aggregates or individual source traffic flows. Local cross traffic may join a path but is modelled and serviced independently.

We isolate traffic destined for a single remote edge that is not local to any of the access or distribution networks. This assumption allows us to analyse the traffic arriving at, or departing from, a single PE interface in Figure 1. The interface is further sub-divided into forwarding classes for which we determine the specific arrival processes.

The trail components up to the core network edge are derived from the domains in Figure 1:

1) The multiservice source model determines the individual traffic generated by a multiservice user and group.
2) In the network-layer multiplexing model we account for multiple levels of aggregation in the access and distribution networks. Multiplexing dynamics, including QoS scheduling and shaping, within the network layer queueing subsystem, are abstracted to determine the traffic departure process and QoS loss contribution.
3) The multiplexed traffic arrival process can be modelled at the terminating side of each inter-node interface but we focus on the edge-terminating path in Figure 1. The parameters of the abstraction of this arrival process are determined by the multi-scale properties found to affect performance of the core network under specific conditions (e.g. utilisation levels, device configuration)[3][4][5]. In addition we add a QoS budget parameter to the profile such that the transport network can calculate the further allowable QoS loss in the core.

Components 1 and 2 shape the complex traffic arrival process of 3, which is modelled using statistical methods. We discuss the details of these components in the following sections.

III. A COMPOUND TRAFFIC SOURCE OBJECT

In his paper on the abstraction of services and network technologies to support telecommunications legislation [6] Hanrahan identifies Service Groups (SG) as a means of defining benchmark service and bearer capabilities. We extend this description to define a compound, aggregate, user-group-access-object up to the core edge. The object includes four domains: multiservice user, user group, access and distribution networks. The domains drive traffic variability and their operation is transient as new technologies or services are implemented.

![Fig. 2. Compound User Group Model](image)

**A. The Multiservice User and User Group Domains**

The multiservice user may start a simultaneous session of two or more services as in the User Layer of Figure 2. These services may be voice (at different qualities), data (FTP, HTTP), video etc. We focus on the implementation of VoIP (G.726) and essential data services (HTTP, FTP, Email) with variable session length distributions and complexity. The properties of the Subscriber Group Layer are defined during a busy period (whenever that may be) from which daily, weekly and monthly variations in access are a simple derivation. The arrival process of simultaneous web-browsing-voice sessions (mixed with single service sessions of variable percentage) is assumed to be Poisson within this period. The arrival distribution of connections within a service in the User and Application Command Layer is determined by the user commands (e.g. link to new page), QoS response (e.g. restart connection), and application specific properties (e.g. an HTTP session may open single connection or multiple simultaneous connections to the same web page). The packet or burst arrival process is known to be self-similar at large time scales [7][8], and caused by connection interarrivals at the User and Application Command Layer. However, data burst arrival processes at finer time scales (below the TCP RTT) are shaped by closed-loop control at the Transport Layer [7]. Voice packet and burst arrival processes are determined by the encoding type, packetisation interval and silence suppression. The User and Subscriber Layers are easily extended to describe additional services. Any exten-
sion at the higher layers would generate a greater number of QoS and connection related variables in the User and Application Command and Connection Layers.

These parameters are used to instantiate specific user group types and may additionally include source-destination addresses as part of the traffic load, origin-destination matrix. The aggregated origin component is then defined by all source traffic, destined for the same single originating-terminating edge pair, that is aggregated and modulated up to the core network edge. The multiplexing process is discussed in the following section.

B. Access and Distribution Network Domains

The access and distribution networks transport and aggregate multiple source traffic flows and classes. The interaction of these flows under multiplexing and QoS control modulates the traffic departure process. We determine the characteristics of this aggregate’s pattern by analysing the processes shaping the flow and its forwarding classes.

**Access Abstraction:** An SG-4 (web browsing and voice services) [6] user may use wireline or wireless access. We envisage that the steady state properties of different access technologies can be abstracted by its transmission serialisation process and propagation delay parameters. The serialisation process and delay is typically dependent on bit-rate, error feedback and correction, processing delays, channel sharing and interference components. Serialisation can be modelled as the interaction of these components in the transmission process. In the wireline case we assume that most of the parameters have negligible effect, and serialisation is dependent only on the bit-rate of the channel. Alternatively, the access network can be defined by its bottleneck bit-rate according to the maximum reasonable demand to support the implemented services. For example; in the case of an SG-4 user (web browsing and voice) [6], this bottleneck is chosen close to 130kb (dependent on allowed download time for web objects). This bit-rate demand is sufficient to define the access component in this paper.

**Multiplexing Abstraction:** The traffic departure process resulting from the statistical multiplexing of multiple source traffic flows can be modelled analytically. However, analytical modelling methods become intractable when the multiplexing process is controlled and shaped by QoS components such as class-based queueing, random early detection (RED), fragmentation, policing and shaping at network interfaces. In addition, closed-loop control in TCP adds feedback complexity to the model. We therefore implement Object-Oriented (OO) modelling and simulation of these multiplexing, QoS control and congestion feedback components to determine their effect on the traffic departure process (and QoS). The usefulness of multiplexing abstraction extends into network resource management, as in the CIM, and is also used to propose possible implications of multi-timescale burstiness in the following section.

We use a CIM-like OO description in Figure 3 to model the inter-node domain from output interface of originating node to input interface of destination node for simulation implementation in Section IV-B. The Modules or components of an inter-node domain are generically modelled at the network layer, packet switching and queueing subsystem. The inter-node domain consists of an Ex_Switch and In_Switch interface. The Ex_Switch interface of, for example, the CE originating node typically contains modules. Each module in turn contains possible QoS components: Class-based queueing Scheduler with WRED, Policing of priority flows, Shaping of data flows, and Fragmenting and interleaving. The In_Switch modules of, for example, the PE terminating node may include an ingress Queue with FIFO Scheduler and a Classifier. The model is defined as an inter-node interface, meaning that the aggregated traffic abstraction is analysed before space switching.

By using the OO model in a simulation implementation we are able to determine the modules of the above multiplexing abstraction and elements of the user source model (of Section III-A) that affect the traffic profile significantly, according to the multi-scale burst properties discussed next.

IV. Time-Variance Analysis of Edge Traffic

A time-variance analysis of a traffic arrival or departure processes is insufficient as a model for analytical queueing approximations due to the inaccuracy of its curve estimate [8]. However, analysis using wavelet models is shown to be accurate [7][9], and the time-variance derivative is a close approximation to the Haar wavelet transform used in [9]. We therefore use the time-variance derivative to determine edge traffic arrival process response to changes.
in multiplexing control and the source group characteristics. We first discuss the interest in, and prove the need for, multi-time-scale properties with regard to performance management, followed by simulation and time-variance modelling.

A. Performance impacts of multi-scale variance

New research [7][10] investigates the variance of aggregated traffic at multiple time scales, confirming the existence of self-similarity over large time scales, and also the presence of small time scale multi-fractals (below a lower time cut-off of the TCP round trip time (RTT)). Whilst self-similar processes have a linearly decaying variance with slope between -1 and 0 on a log scale, fine time scale multi-fractals show greater variance and non-linear time-variance decay relationships (easily illustrated in the wavelet domain [9][3]).

The performance impacts of long-range dependence are well studied and known to affect performance at intermediate and high utilisations, whilst current research [3] predicts a significant queueing backlog, at low and intermediate utilisation, due to fine-timescale multi-fractals [3][4]. The analysis is based on sample packet or byte counts depending on its relevance to the device operation. We propose possible implications on resource management given multi-timescale information at a small to intermediate aggregation level:

1) Call admission control is based on predicting the resulting QoS change due to an change in the bit-rate load. We extend the possible condition-of-entry parameters to use the multi-timescale arrival process properties. The incoming flow may be denied based on how it will change the multi-scale properties of the class or trunk at that utilisation level.

2) Class-based scheduling is used to isolate the transport of similar flows. A flow may need to be reassigned to a new class or existing class’s scheduling modified based on the link utilisation. For example, a higher utilisation meaning that larger timescale, self-similar properties are more dominant [3], will introduce different delay and overflow probabilities for this class.

3) Router interface configuration requires input or output buffer sizing (possibly as shared memory). Queueing theory is used to determine optimum buffer sizes based on traffic arrival patterns and delay/overflow limits. Much of the router processing overhead is dependent on packet address lookup delay rather than packet size. The queue backlog is therefore more dependent on packet count processes. The necessary interface bandwidth is dimensioned and allocated according to average bit rates of each traffic class at a coarse time scale.

4) Router interface configuration includes shaping or policing parameters based on the predicted excess and committed byte burst sizes. The burst-time-scale of interest may then be influenced by likely utilisation levels and class of traffic.

In future we will include a QoS loss component through access and distribution networks, such that the resource manager may allow for a further maximum QoS loss between edge nodes as part of the overall QoS budget. The resource manager will be able to predict a certain performance of available network resources, components or configuration for certain edge arrival processes. The manager is also able to extend or reconfigure the network to provision for a change in the arrival process caused by a changing user-group defined in Section III. These possible application areas provide a framework for analysing the simulation traces.

B. Simulation of Source Traffic Aggregation by implementation of the OO Element Model

In implementing a simulator we required a flexible platform to which we could add generic components that are not provisioned in the existing simulator model, and to allow for flexible user configuration of element architecture as new technologies emerge. The existing implementation of the Network Simulator, ns2 [1], fitted well with our node-to-node interface abstraction in Section III-B. The Link object was modified to allow for a full interface implementation as in Figure 3. Our simulations were implemented using the subscriber group source model of Section III. The existing ns2 source models are consistent with stack down processing of packets that characterise the load transforms [11] at the user terminal. An HTTP agent class generates web browsing sessions according to Table I [12]. The voice application is easily configured for multiple codec types and ON/OFF burst distributions. We base our voice agent implementation on parameters in Table I, from the model in [13]. The simulation topology in Figure 4 shows the implemented system and optional extension as dotted lines. It is pos-
sible to implement further aggregation levels and topologies (e.g. ring) by explicitly defining the interconnection of nodes.

C. Time-Variance-Derivative Analysis of Edge Traffic Traces

We determine qualitative relationships between (1) the edge aggregated traffic profile (per class or trunk) and (2) the system variables: user and user-group characteristics and multiplexing architecture. The profile describes traffic crossing the edge interface in a down or upstream direction. Due to the shape of the simulated topology (client-server asymmetry) the downstream flows carry the majority of bytes of TCP traffic whilst the voice traffic is symmetrical. We therefore also model the downstream edge traffic profile as it presents a more interesting byte-count process. The simulation traces are sampled at system steady state: the average number of arrivals equals the average number of departures. The following results were determined at an aggregation with average bit-rate approximately 3Mbit/sec, as an initial investigation into the phenomenon.

We use the time-variance derivative as this visualisation method replicates the Haar wavelet transform and is similar to the one-to-one relationship between wavelet and time domains proven in [9]. This graphical method is useful to detect patterns in traffic behaviour as the multiplexing and source variables change. In addition, we inherently include analysis of flow level dynamics by sampling at multiple time scales up to the trace length.

The simulation scenarios represent a likely daily and weekly change in usage patterns, and are used to determine the effect of CBQ on the arrival process. The packet arrival process trace is collected at the PE terminating path (of Edge 2) and is sampled at an initial fine time scale period of 7 ms. A sample is either a byte or packet count for the initial fixed interval denoted by time-scale: \( m = 1 \). A more coarse time scale series function \( (m = 2) \) is constructed from the initial mother function \( (m = 1) \) by calculating the average count over non-overlapping blocks of size 2. This operation is repeated for increasing \( m \), generating multiple arrival process time-series at increasing timescales. We are able to determine the variance of each time-series and plot the variance vs. \( m \) relationship. This relationship is used in self-similar analysis to determine the Hurst parameter from the log-scale gradient. However, we take the derivative of this relationship to approximate the wavelet-coefficient analysis method used in [7] and [9]. Our analysis does not investigate self-similar, long-range properties since the effect is well known at large time-scales, and the relatively small simulation time sample does not allow for an accurate analysis of long-range data.

We compare the effect of FIFO and CBQ on the multi-scale properties for the aggregate arrival process of a single user group, \( A \) \( (N = 400) \), in Figure 4. The parameters of \( A \) are defined by: Voice traffic at 20 Erlang with average ON bit-rate of 18.252 kb/s giving an aggregate average bit-rate of 0.365 Mb/s. Data traffic at 200 Erlang (due to long holding times), with average bit-rate of 14 kb/s giving an aggregate average bit-rate of 2.8 Mb/s.

Figure 5 shows the log-scale, Variance-Time Derivative of the byte count process up to \( m = 8 \) (fine-mid timescales). In this case we determine the effect of CBQ between \( m = 5 \) \( (0.112s) \) and \( m = 9 \) \( (1.792s) \). The CBQ curve shows a drop in variability gradient between 0.224 \( (m = 6) \) and 0.896 \( (m = 8) \) seconds. This drop is due to the priority queuing of voice packets distributing data traffic in the OFF periods of voice traffic, which are Poisson in size and arrival. The Poisson distribution of data causes a faster decay near the voice ON/OFF burst time scale. An increasing gradient between \( m = 2 \) and 7 indicates a quadratic variance decay function, possibly characteristic of multi-fractal burstiness [9]. The lower time-scale peak in CBQ is due to sampling at an interval smaller than the voice ON/OFF burst periods, thereby capturing data and voice bursts in separate samples.

The curves in Figure 6, from a lower order interpolation, show a reduction in variability decay of the byte count process in the mid-range component \( (0.5s \text{ to } 114s) \) of the CBQ case. The packet count process, however, showed marginal differences in decay rate. In the low-mid time scales CBQ forces faster variability decay, but at large time scales the same burst distribution of traffic of each class is transferred in FIFO and CB queueing, thus the decay in variability becomes similar.

![Fig. 4. Simulation Topology](image)

![Fig. 5. Variance-Time Derivative Plot of Byte Count at Low Time Scales](image)
We modify the parameters of group A to generate the Variance-Time Derivative curves in Figure 7. The plot shows the effect of different percentages of each service, representing different busy periods of the day, for the same size user group, using CBQ. The three scenarios are summarised in Table II.

The lower bit-rate (and resulting lower level of statistical multiplexing) in the Low Data scenario results in relatively slow variability decay, but the dominance of voice traffic between 0.224 seconds (m = 6) and 0.896 seconds (m = 8) is apparent and expected. A peak at m = 11 is due to the low level of multiplexing and therefore the high relative impact of a small amount of data traffic. More data is introduced in the next two scenarios reducing the voice dominance between m = 6 and 8, and causing a lower rate of variance decay towards the tail.

**TABLE II**

<table>
<thead>
<tr>
<th>Relative</th>
<th>Erlang Voice</th>
<th>Erlang Data</th>
<th>Bottleneck Utilisation, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mix</td>
<td>(Avg. agg. rate, Mb/s)</td>
<td>(Avg. agg. rate, Mb/s)</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>20 (0.365)</td>
<td>5 (0.07)</td>
<td>14.3</td>
</tr>
<tr>
<td>Mid</td>
<td>20 (0.365)</td>
<td>200 (2.8)</td>
<td>105</td>
</tr>
<tr>
<td>High</td>
<td>0.5 (0.01)</td>
<td>200 (2.8)</td>
<td>93.7</td>
</tr>
</tbody>
</table>

Based on current work [3][4][5] and further analysis, we can determine the allowed occupancy and properties of aggregated traffic on a class or trunk for admission control, optimal buffer sizes, and optimal shaping and policing limits. Further dimensioning and configuration analysis does require a more quantitative investigation to produce mathematical models of the multi-scale arrival process.

**V. CONCLUSIONS**

We model the source and subscriber group properties for a heterogeneous service profile, and the aggregation of source traffic up to the core network edge. The paper analyses the multiplexing of multiservice, upstream client, and downstream server flows under different aggregation architectures (class-based and FIFO queueing); and different source subscriber group properties with varying percentages of voice and data traffic. We characterise the edge traffic in the low to mid range timescale - to add to past work on long range self-similarity - and show a distinct effect due to CBQ at voice burst time-scale, with an expected trend as the number of data users increases. This edge traffic profile will be used as a future input to resource management in the core and distribution inter-networks.

**REFERENCES**


**Biography**

Paul Plantinga holds a BSc(Eng.) degree in Electrical Engineering from the University of the Witwatersrand. He is presently pursuing his MSc(Eng.) degree at the University of the Witwatersrand's CeTAS Laboratory.

Hu Hanrahan is Professor of Communications Engineering at Wits University. He leads the Centre for Telecommunications Access and Services (CeTAS), a research and advanced teaching centre devoted to improving knowledge and practice in the evolving telecoms access networks and telecoms services.