Scholars Journal of Engineering and Technology (SJET) **ISSN 2321-435X (Online)**
Sch. J. Eng. Tech., 2015; 3(7):658-663 **ISSN 2347-9523 (Print)**

(An International Publisher for Academic and Scientific Resources)
www.saspublisher.com

Research Article

Review of Differentiated-Services Architecture for QoS-Aware and Delay-Sensitive Campus Area Network (Case Study of Ahmadu Bello University, Zaria)

Patrick N.J., Usman A.D, Tekanyi A.M.S.
Department of Electrical and Computer Engineering, Ahmadu Bello University Zaria, Nigeria

***Corresponding author**
Patrick N.J
Email: nyabvou@gmail.com

Abstract: University networks are mostly designed to support best-effort data applications, but demands by the university's faculty, researchers, students, administrators, and staff to discover, learn, reach out, and serve society expect the networks to run voice, video and other multimedia traffic over it as well. In today's networks, the widespread use of real-time and multimedia traffic applications demand special service guarantee in terms of throughput, delay, and delay variance, thus making quality of service (QoS) a key problem. For multimedia packets to travel seamlessly on the network and for the network to meet the demand for higher performance by applications; Voice, video and other multimedia packets must be given priority in term of packet delay and packet delay variance over other, less-time-sensitive traffic, such as e-mail or Web browsing. In this paper we present the review of development of a differentiated services computer networking architecture for packet delay and packet delay variance reduction in a campus area network. Using management tools that can be used to provision and monitor set of routers in a campus network in a coordinated manner, which can classify and apportion network traffic give preferential treatment to applications identified as having more demand requirements. This scheme is capable of differentiating between delay sensitive and best-effort traffic and route packets accordingly.
Keywords: Best-effort, differentiated services, quality of service, packed delay, packet delay variance, network architecture.

1.0 INTRODUCTION

Information technology is strategically important to the goals and aspirations of business enterprises, government entities or educational institutions particularly universities. It is the cornerstone that enables the university's faculty, researchers, students, administrators, and staff to discover, learn, reach out, and serve society. On campus area networks (CANs) such as Ahmadu Bello University (ABU) network, many real-time applications, such as video and audio streaming, are available for experimental, practical and other uses; the numbers of real-time applications on such network are on the increase. Voice, video and data applications demand different types of performance assurance and service differentiation, and so quality of service (QoS) provisioning is one of the important goals in the design of such networks. However, the mechanism to guarantee and support QoS for real-time applications on most networks has not been achieved yet; only best-effort service is available. As a result, real-time applications must tolerate some degradation of QoS in terms of packet loss, delays, and delay variance for messages transmitted over the networks.

1.2 Problems in Campus Network Management

Packet-based networks, are networks in which message gets broken into small data packets that seek out the most efficient route as circuits become available for efficient transmission of the message. Packets sent from a source may traverse different paths to arrive at the final destination; the packets that are routed over separate paths are reassembled at the destination. Transmission rates of the various paths may vary depending upon the usage of the network paths over which the packets are being transported [1]. During heavy traffic conditions packets may be delayed and lost, Packet delays and losses causes poor performance of the network and is more obvious with voice and other multimedia streaming communication. Streaming packets in a network share the network bandwidth with conventional non-streaming packets (such as data associated with electronic mail, file transfer, web access, and other traffic) [2]. Voice data and other multimedia packets that are lost or delayed due to inadequate or unavailable capacity of networks may result in gaps, silence, and clipping of audio at the receiving end thus affecting the quality of service of the network.

The protocols used in today's networks have been mainly optimized to provide connectivity. In these networks, the main problem was to reach the destination even if transmission quality was poor. Due to this fact, most IP based networks provide today a best-effort service, i.e. a context where the network does its best to transmit information as quickly as possible but does not provide any guarantee on the timeliness or even the actual delivery of this information. In today's electronic trade context, there is a real demand for a minimum level of performance to be guaranteed to mission critical applications. Corporate administrators are in need of simple and comprehensive mechanisms to deliver services. Introducing such mechanisms between users and the network means that, contrary to the current best-effort networks which handle all data equally, the network now needs to discriminate between various kinds of data traffic to provide multiple service levels. In this context, a data flow is conceived as a set of transfer units considered related to each others, according to some discriminating criteria on quantitative observations such as generating sources, users, applications, or destinations.

1.3 Need for Quality of Service (QOS)

The concept of quality of service refers "to the ability of a network to provide improved service to selected network traffic over various underlying technologies" [3]. Due to the demands for new types of services, universities campuses are facing new challenges related to network infrastructure. Today's networks need to support multiple kinds of traffic over single network links. Different kinds of traffic demand different treatments from the network; we cannot have separate network connections for each kind of traffic. Therefore much of the bulks of network traffic have to flow through lines where high priority traffic and other classes of traffic have to share the bandwidth. We can only differentiate at places where the traffic flows through active network elements which have the capability to differentiate. Examples of such entities are routers, switches and gateways.

Knowing that networks are mostly design to meet certain expectation, design goals, and other critical factors and features such as scalability, robust, efficiency and security which are dependent on network parameters such as bandwidth, channel utilization, buffer problems, productivity (throughput and effective capacity), responsiveness (delay, round trip time and queue size), and losses (packet loss rate and frame retries). A need exists for the development of methods and ways to continuously improve the quality level of networks most especially network packet delays, Packet delay variance and packet losses to guarantee better performance and assure Quality of service (QoS) of such networks. Three important parameters used to measures QoS of a network include delay, jitter or delay variance and packet loss [4].

i. Delay: is the time taken by a packet to move from point-to-point in a network. Delay can be measured in either one-way or round-trip. VoIP typically tolerates delays up to 150 ms before the quality of the call become unacceptable [4].
ii. Packet Delay Variance: is the variation in delay over time from point-to-point. If the delay of transmissions varies too widely in multimedia information, the information quality is greatly degraded.
iii. Packet loss: is the loss packet along the data path. The non-arrival of some packet, at their destination severely degrades the message quality.

Table1.1 QoS Requirements for Different Applications [5]

Traffic type	Bandwidth	Packet loss (max)	Delay (max)	Jitter (max)
Interactive voice (G.711)	12-106 kbit/s	1%	150 ms	30 ms
Streamed video (MPEG-4)	0.005-10 Mbit/s	2%	5000 ms	Insensitive
Streamed audio (MP3)	32-320 kbit/s	2%	5000 ms	Insensitive
Data	Variable	Sensitive	Insensitive	Insensitive

Therefore, the need to design networks which has the following functionalities

i. Can deliver multiple classes of service – that is they should be QoS conscious.
ii. Is scalable – so that network traffic can increase without affecting network performance
iii. Can support emerging network intensive, mission critical applications

In order to meet these functionalities the network should implement service models so that services are specific to the traffic they service.

1.4 Network Services Models

New QoS aware network services models are been proposed and ways to implement and improve commonly used service model are been develop on regular basis. The commonly use service model and establish QoS network service models include:

Best Effort Services Model:

Best Effort is the common service provided for traffic transportation. It's a network policy for which no special QoS model is implemented; therefore all kind of traffic is treated equally. As the name imply. This model delivers the entire packet to the destination without guarantees of delay, packet delay variance,

packet loss, etc. There is no differentiation between the kinds of traffic, no classification or prioritization and all packets receive the same treatment independent of its content. Best effort is the treatment that packets get when no predetermined treatment is specified for them. If there is congestion on the medium, the caching function can be used to store packets temporarily and when the situation is solved packets will be forwarded. In the event of extreme congestion, when the network device cannot handle the situation, packets can be dropped indiscriminately. Best-effort service is not suitable to guarantee end-to-end QoS for all kind of traffic. To provide end-to-end QoS two models have been deployed; integrated services and differentiated services model. End-to-end QoS means that the network provides the level of service required by traffic throughout the entire network, from one end to the other.

Integrated Services Model:

The basic concept behind integrated services model is that an application request specific treatment from the network device that makes decision in forwarding traffic (such as router), and the network device confirm that it can provide the required resources, and they come to an agreement before any data is sent. Integrated services model QoS is achieved through an agreement of specific treatment for a given type of traffic before it is sent.

An integrated service uses an explicit signaling mechanism from applications to network devices. Signaling is used to reserve and release resources in the network. QoS signaling allows network node to communicate with its neighbors to request specific treatment for a given traffic type. The application requests a specific service level, for example, its bandwidth and delay requirements. After the network devices have confirmed that it can meet these requirements, the application is assumed to only send data that requires that level of service.

Applications in an integrated services environment use the Resource Reservation Protocol (RSVP) to indicate their requirements to the network devices. The network devices keep information about the flow of packets, and ensure that the flow gets the resources it needs by using appropriate queuing (prioritizing traffic) and policing (selectively dropping other packets) methods [4]. Two types of services provided in an integrated services environment are as follows:

i. Guaranteed Rate Service - This service allows applications to reserve bandwidth to meet their requirements. The network uses weighted fair queuing (WFQ) with RSVP to provide this service.
ii. Controlled Load Service - This service allows applications to request low delay and high throughput, even during times of congestion. The network uses RSVP with weighted random early detection (WRED) to provide this kind of service.

Integrated services are very networking consuming resources, because it requires RSVP on all network devices. This characteristic makes it currently not used as much as differentiated services.

Differentiated Services Model:

Differentiated services mechanism technique treats packets with different level of requirements depending on their source, destination and/or the kind of traffic they are carrying. To accomplish this, packets are first divided into classes by marking the type of service (ToS) byte in the IP header. A 6-bit bit-pattern (called the Differentiated Services Code Point [DSCP]) in the IPv4 ToS Octet or the IPv6 Traffic Class Octet is used. The network tries to provide level of service based on the quality of service defined in the header of each packet. Packets are usually classified or marked by edge network devices according to previous defined criteria such as source, destination and kind of traffic. Classification and marking are the fundamental base of differentiated services as they help implement priority in the network. At first packet is identified, therefore depending on that identification it is given priority over other packets or different treatment from them. The process of classification of packet is the process of analyzing and sorting packets according to their contents between different categories. It means that each packet is designate as belonging to voice category, data category or multimedia category, etc. Each category has different level of quality requirements. Marking process will check the category that the packet belong to and therefore put a mark or level of priority within the head of packet. When packets are classified at the edge of the network, specific forwarding treatments, formally called Per-Hop Behavior (PHB), are applied on each network element, providing the packet the appropriate delay-bound, jitter-bound, bandwidth, etc.

1.5 Related Works

There have been several works on QoS-aware and delay-sensitive networks, each with different advantages and disadvantages. Specifically, [6] developed an architecture in which packets carry as much information as possible, while routers process packets as detailed as possible, using Load Adaptive Routers. The developed architecture provides better protection to data flows in terms of packet loss, than best-effort and differentiated service architecture when malicious data flows exist. However the work introducing complexity to routing design and also increasing network processing time resulting in increasing packet delay.

[7] Worked to model an adaptive buffer sharing scheme for consecutive packet loss and packet delay reduction in broadband networks. They developed an algorithm for optimization of adaptive buffer allocation scheme for broadband network which reduced networks packet delays by systematically dropping few packets of messages that could tolerant loss of few packets but not consecutive packets drop in data-stream. The work however did not addressing the effect of latency due to large buffers.

[8] Presented queue management and scheduling in network performance, by analyzing the performance bottleneck of different queuing policies. They proposed the use of Active Queue Management (AQM) to replace drop-tail queue management in order to improve network performance in terms of delay, link utilization, and packet loss rate and system fairness. The work shows that AQM improved network performance by reducing packet loss rate and packet delay. Their work however is based on simulated finding.

[9] Presented the effects of finite buffer sizes on the throughput, packet losses and packet delay in different networks. He used the established effect of finite buffer sizes to analytically characterize network throughput, packet losses and packet delay to estimates and model the performance measures of a network. The work captures the vital trends of the network, yields better estimates of the throughput, packet losses and delay performance measures. However, large buffers have an adverse effect on the latency.

[10] Worked to develop an application that would improve performance of the network by controlling packet loss using ways to control network congestion. Implement the Stable Token-Limited Congestion Control (STLCC) mechanism for controlling inter-domain congestion and improve network performance. The results revealed that the application was able to control network congestion by controlling packet loss, thus improving performance of the network.

[11] Developed differentiated service (DiffServ) architecture for QoS support and routing for delay-sensitive and best-effort services in IEEE 802.16 mesh network. They developed new cross-layer routing metric namely, expected scheduler delay (ESD), using efficient distributed scheme to calculate ESD and route the packets using source routing mechanism. This scheme was capable of differentiating between delay sensitive and best-effort traffic and route packets accordingly. The work was able to establish that ESD metric was better compared to hop count metric in terms of delay for the service model that contained both delay sensitive and best effort service. However level of improvement of the architecture depended on the expected scheduler delay performance.

DESIGN ISSUES

2.1 Campus Area Network Resource Control

A typical picture of network resource control involves a management tool, a policy repository, a policy decision point and a policy enforcement entity. The network administrator uses the management tool to populate the policy repository with a number of policy rules that regulate access/use of network resources. These rules could specify for instance, the service category to be employed for a particular application, how much bandwidth is allocated to a particular flow or type of service category, etc.

The administrator-specified rules are stored in the policy repository in a well-understood format or schema. The decision entity downloads the policy rules. The enforcement entity is the router, which encounters packets flowing across the network. It queries the decision entity for specific actions that are to be applied in conditioning the packet stream. Some of the directory clients are management tools, which populate and maintain the policy repository in the directory. Others are enforcement entities that apply policy rules by dropping, marking, reshaping or otherwise conditioning the packet stream. Examples of such clients include routers, firewalls, and proxy servers.

The QoS Manager (Policy Decision Point)

This component is based on client-server architecture. All the Routers, on start up, will connect to the QoS Manager to set up basic configurations and keep updating the information henceforth at regular intervals. The Manager then waits for incoming requests from the routers, mainly for priority resolution of the traffic.

Differentiated Service enabled Routers:

The Differentiated Service enabled routers, or the Policy Enforcement Points, are where the policy decisions are implemented. As the QoS architecture we have adopted is differentiated service, this component must have the functionalities to support differentiated service.

Because edge routers and backbone routers in a network do not necessarily perform the same operations, the QoS tasks they perform might differ as well. In general, edge routers perform the following QoS functions:

- Packet classification
- Admission control
- Configuration management

Backbone routers perform the following QoS functions:

- Congestion management
- Congestion avoidance

The Directory Service:

For storing and look-up of local policies, a system needs to be used for facilitating network management and account generation. The Policy Repository can be visualized to be a Directory Services the policy information is not expected to change at a fast rate. As the number of updates are expected to be much less than the number of reads. Using a directory service that will suit the needs of a campus will augment a system that can encourage uniformity in use. QoS Manager uses this information to allocate bandwidth to end users. e.g., delay sensitive traffic can be assigned a higher priority than non-delay sensitive traffic. The attributes can be set in whichever manner the administrator feels would serve the purpose.

3.1 The Differentiated Service Architecture

The differentiated service approach to providing quality of service in networks employs a small, well-defined set of building blocks from which a variety of aggregate behaviors may be built. A small bit-pattern in each packet, in the IPv4 Type of Service (ToS) octet is used to mark a packet to receive a particular forwarding treatment, or per-hop behavior, at each network node. For true QoS, the entire IP path that a packet travels must be differentiated services enabled. An example service policy— Expedited Forwarding (EF) gets 10 percent, gold 40 percent, silver 30 percent, bronze 10 percent, and best effort traffic (default class/PHB) the remaining 10 percent of the bandwidth. Gold, silver, and bronze could be mapped to Assured Forwarding (AF) classes AF1, AF2, and AF3 for example. This can be enforced in any part of the network, including end-to-end.

Typically, the differentiated service boundary node performs traffic conditioning. A traffic conditioner typically classifies the incoming packets into pre-defined aggregates, meters them to determine compliance to traffic parameters (and determines if the packet is in profile, or out of profile), marks them appropriately by writing/re-writing the DSCP, and shapes (buffers to achieve a target flow rate) or drops the packet in case of congestion. A differentiated service internal node enforces the appropriate PHB by employing policing or shaping techniques, and sometimes re-marking out of profile packets, depending on the policy.

3.2 Analysis

[12] Suggested a general approach for expected delay calculations for AF mechanisms. [13] Extend this analytical approach for a threshold dropping queue with Poisson arrivals to the N drop-precedence case.

In an N drop-precedence threshold queue as shown in Figure 3.1, there are N flows (each flow corresponds to a level of drop precedence) arriving at the queue. A packet is discarded at its arrival when its corresponding buffer threshold has been reached or exceeded.

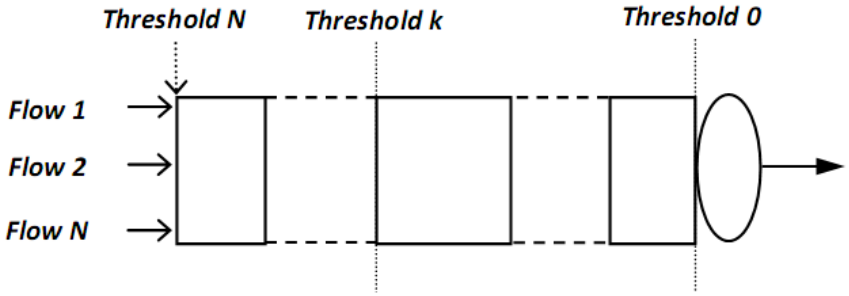

Fig- 1: Threshold Dropping Queue with N Drop Precedence

The analysis has the assumption that the incoming traffic flows are Poisson. We introduce the following terms:

- The arrival rate of the i^{th} priority flow is λ_i.
- The packet service times are exponentially distributed service times with mean $1/\mu$.
- The loads of the i^{th} priority flow and the aggregation are ρ_i and ρ respectively
- The buffer threshold of the i^{th} priority flow is L_i, packets (L_o is *0*)
- At steady-state, the probability that there are n packets in the system is $\Pi(n)$
- $\alpha(n)$ is the acceptance probability of a packet which arrives to the queue seeing *n* other packets already in the system
- $\alpha_i(n)$ is the acceptance probability of an i^{th} priority packet which arrives to the queue seeing *n* other packets already in the system. For a threshold queue, this probability can be determined as

$$\alpha_k(n)=\begin{cases}1 & if \quad n < l_k \\ 0 & if \quad l_k \le n\end{cases} \qquad (3.1)$$

- p_i is the ratio of the i^{th} priority flow's load to the overall load. Hence, p_i is the ratio of λ, over the sum of all arrival rates.

It is important to notice that the lower the drop precedence of a flow, the higher the priority of the flow (eg. the 1st priority flow has the lowest drop precedence and a buffer threshold of L_N, which is the buffer size of

the queue). From the definition of $\alpha(n)$ and $\alpha_i(n)$ we have

$$\alpha(n) = \sum_{i=1}^{N} p_i \alpha_i (n) \qquad (3.2)$$

$$\alpha(n) = \begin{cases} P_1 + \ldots + P_N & if \quad n < L_1 \\ P_2 + \ldots + P_N & if \quad L_1 \leq n < L_2 \\ \ldots\ldots \quad \ldots\ldots \quad \ldots\ldots & \\ P_k + \ldots + P_N & if \quad L_{K-1} \leq n < L_k \\ \ldots\ldots \quad \ldots\ldots \quad \ldots\ldots & \\ P_N & if \quad L_{N-1} \leq n < L_N \\ 0 & if \quad L_N = n \end{cases} \qquad (3.3)$$

It can be seen that this threshold queue can be modeled as a birth-death process. For a state n, the birth rate is p*p*α(n) while the death rate is μ. The steady-state distribution of buffer content is:

$$\Pi(n) = \Pi(0)\rho^n \prod_{i=0}^{n-1} \alpha(i) \qquad (3.4)$$

with the probability that the buffer is empty Π(0)

$$\Pi(0) = \left[\sum_{n=0}^{L_N} \rho^n \prod_{i=0}^{n-1} \alpha(i)\right]^{-1} \qquad (3.5)$$

From (4.3) and (4.4), we obtain:

$$\Pi(n) = \Pi(0) \prod_{j=1}^{k-1} \left[\sum_{i=j}^{N} p_i\right]^{L_j - L_{j-1}} \left[\sum_{i=k}^{N} p_i\right]^{n - L_{k-1}} \quad \text{if } L_{K-1} < n \leq L_k \qquad (3.6)$$

The loss probability of the i^{th} priority flow is determined as:

$$\text{Loss}_i = 1 - \sum_{n=0}^{L_N} \Pi(n) \alpha i(n) \qquad (4.7)$$

Clearly, when a packet arrives at the queue which already has n packets, it has a delay of n packets service times plus its own service time. Therefore, the mean delay of the i^{th} priority flow (excluding rejected packets) is:

$$\text{Delay}_i = \frac{1}{\mu} \frac{\sum_{n=0}^{L_N - 1} (n+1)\Pi(n)}{\sum_{n=0}^{L_N - 1} \Pi(n)} \qquad (4.8)$$

From equation (4.8), the mean delay of the i^{th} priority flow is affected by the packet service time's 1/μ and delay of n packets service times at the queue which already has n packets.

4.0 CONCLUSION

This paper presents the review, of development of a differentiated services computer networking architecture architecture, for packet delay and packet delay variance in a campus area network. Using management tools that can be used to provision and monitor set of routers in a campus network in a coordinated manner, which can classify and apportion network traffic give preferential treatment to applications identified as having more demanding requirements. This scheme is capable of differentiating between delay sensitive and best-effort traffic and route packets accordingly.

REFERRECES

1. Abaye A, Lo WF, Carsten RR, Robertson B, Briere D; Performance modeling of a communications system. IEEE 19th Conference On Local Computer Networks, 2006; 1-13,
2. Obiniyi A, Soroyewun M, Abur M; New Innovations in Performance Analysis of Computer Networks: A Review. Computer Networks Review, 2014.
3. Wooldridge M; A response to Franklin and Graesser Intelligent Agents III Agent Theories Architectures and Languages. Springer, 1997; 47-48.
4. Tavares ÁS; Electronics and Communication Engineering, 5th, Portugal, 2011: 24-58.
5. Hutcheson L; FTTx Current status and the future. Communications Magazine, IEEE, 2008; 46(7): 90-95.
6. Wang Y, Krishnamurthy A, Qian L, Dauchy P, Conte A; A-Serv a novel architecture providing scalable quality of service [Internet applications]. IEEE Global Telecommunications Conference, 2004.
7. Kausha S, Sharma R; Modeling and analysis of adaptive buffer sharing scheme for consecutive packet loss reduction in broadband networks. International Journal of Computer Systems Science and Engineering, 2007; 4(1): 8-15.
8. Olawoyin L, Faruk N, Akanbi L; Queue management in network performance analysis. International Journal of Science and Technology, 2011; 1: 215-218.
9. Torabkhani N; Modeling and Analysis of the Performance of Networks in Finite-Buffer Regime. Georgia Institute of Technology, Georgia, 2014.
10. Sonawane RR, Varalaxmi G; A Novel Token Based Approach Towards Packet Loss Control. International Journal of Research in Engineering and Technology, 2013; 2 (10): 522-525
11. Bhakta I, Chakraborty S, Mitra B, Sanyal DK, Chattopadhyay S, Chattopadhyay M; A diffServ architecture for QoS-aware routing for delay-sensitive and best-effort services in IEEE 802.16 mesh networks. Journal of Computer Networks and Communications, 2011; 2011: 1-13
12. Bolot JC, May M, Jean-Marie A, Diot C; Simple Performance Models of Differentiated Services Schemes for the Internet. Proceedings of IEEE INFOCOM, 1999;3: 1385 - 1394
13. Nguyen, Long V, Tony Eyers, Joe Chicharo F; Differentiated service performance analysis. Fifth IEEE Symposium on Computers and Communications, 2000; 328-333.