

Comparison of NGN QoS control Models - distributed or centralized

Jongtae Song and Soon Seok Lee
ETRI
Republic of Korea

Abstract - The QoS control can be implemented either in the centralized or distributed model. In the centralized model, the call setup signaling and resource control functions are implemented in the control/signaling servers while distributed model implements the same functions in the transport equipment. Selecting the appropriate control model is critical when deploying the real network. This paper identifies the control procedure of both models and provides a performance comparison on both models. Based on our analysis, the performance of distributed model is better when traffic load is high and traffic is uniformly distributed. When the majority of calls are initiated or destined to one edge node, centralized model has better performance.

Index Terms—NGN, QoS control, distributed, centralized

I. INTRODUCTION

The Internet service becomes available in many industry sectors such as health, finance, entertainment, education, and government. The demand of the new service is increasing as more people have experienced the Internet service. The service requirement from the customer changes dramatically.

In the past decades, the progress of technology creates new services. In the future, however, the customers' needs for new service will be the main driving force of the technology development. Therefore, the forecasting the direction of the service evolution is important to decide the direction of the technology development. The industry power shifted from providers to customers [1][2]. Customers no longer satisfy with the typical services designed by the service providers. Personalization of the service is one of the key requirements for the future services.

The transport network should be flexible enough to meet this service requirement. For satisfying the various service requirements in the common transport framework, continuous effort has been made. One of important effort is having independent architecture between service and transport functions. By having independent architecture between service and transport function, network can support the various QoS request from any service provider. The independency of service and transport has been actively discussed in standard bodies.

ITU-T defines the QoS control functions based on its NGN architecture [8]. The functions of NGN architecture are divided into two strata - Service Stratum and Transport Stratum. The transport is concerning about the delivery of packets of any kind generically, while the services are concerns about the packet payloads, which

may be part of the user, control, or management plane. Under the concept of the independence of a service and transport functions, the network resource and quality are guaranteed by the network side upon request from the service stratum. Service Stratum is responsible for the application signaling. The service stratum can be a simple application server or a full-blown system such as IMS (IP Multimedia Sub-system). Transport Stratum is responsible for reliable data packet forwarding and traffic control.

Transport control function is located in Transport stratum interfacing with the Service stratum. It makes decision on the admission of the requested service based on the network policy and the resource availability. It also controls the network element to allocate the resource once it is accepted. Resource and Admission Control Functions (RACF) is responsible for the admission decision and resource control of the transport function. Details of RACF mechanism can be found in [3][9].

Functional entities for QoS control are defined in [9]. Physical implementation of the entities can be different based on control models. The example of control architecture is distributed and centralized architecture. In the distributed model, signaling and resource control functions are implemented in the transport nodes. In the centralized architecture, the control functions are implemented in the centralized control servers.

The choice of the control model depends on many factors such as network size, number of customers, dynamicity of the QoS request, performance of the control servers, and so on. However, there was no serious research work for analyzing the pros and cons of both architectures. In [6], the distributed and centralized structure was studied and proposed a new hybrid model for combining the distributed and centralized model. But its focus was on the software performance in terms of data access. Several architectures have been proposed without performance analysis [4][7].

In this paper, we analyze the two QoS control models and investigate the performance in both models. In section II, we will provide overview of NGN QoS control and explains the two architectures - centralized and distributed. In section III, we will analyze the two architectures. In section IV, we will provide the simulation result, and conclude in section V.

II. OVERVIEW ON NGN QOS CONTROL ARCHITECTURE

A. Review on NGN QoS control

Resource control in NGN assumes the dynamic call-by-call control. The call request is generated dynamically by

CPE (customer premises equipment) and it may create control overhead. In the QoS control architecture, the edge nodes play an important role. They are responsible for enforcing the traffic, monitoring the traffic conditions, and gate control. For the dynamic resource control, several control architectures have been introduced. Recently, SBC (Session Border Controller) obtained attention as a possible solution of reducing the control overhead. SBC integrates service and transport functions and performs both call setup signaling proxy and media traffic controls such as media packet scheduling, shaping, and gate control. By integrating the functions, it may efficiently perform the resource control tasks.

SBC model can further classified into two control models. One is centralized model, the other is distributed models¹. In the distributed model, edge node contains the signaling and traffic control function. In the centralized model, centralized control server is responsible for the signaling and policy level decision and the edge node performs the traffic control of the media flow.

Similar type of architectural alternative can be found in many other transport technologies. For example, the framework of distributed and centralized resource control in MPLS network is under development in ITU-T.

Based on ITU-T NGN definition, the distributed and centralized model can be explained as Figure 1. Related control functions are Service Control Function (SCF), Resource and Admission Control function (RACF), and Policy Enforcement function (PE) function. SCF is responsible for call setup signaling. RACF is responsible for resource control, and the PE responsible for the traffic enforcement. In the distributed model, the edge node integrates the function of SCF, RACF, and PE. In the centralized control server integrates the SCF and RACF functions. The edge node of centralized model has PE functions which performs media traffic control.

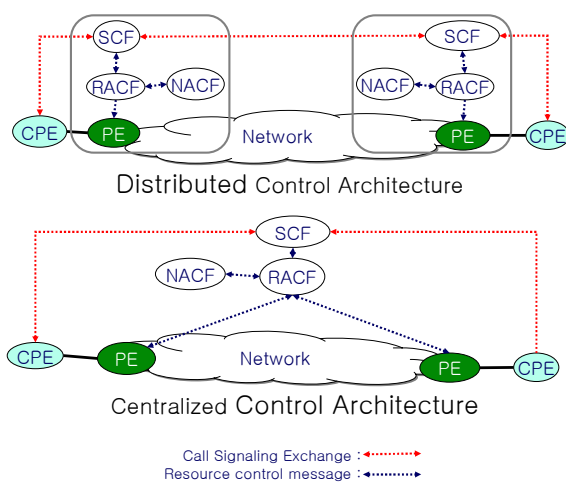


Figure 1 Control Architecture for Distributed and Centralized model

Based on physical implementation of the functional

¹ Centralized model in this paper often referred as distributed model in SBC related material because SBC functions (signaling and traffic control) are distributed to multiple network element. Distributed model is referred as integrated model.

entities, the performance of two control models can be different. In this paper, we will investigate the performance of the two architectures.

III. ANALYSIS ON QoS CONTROL ARCHITECTURES

NGN QoS control assumes the call by call control. The optimal QoS control models can be decided based on its control overhead. The control overhead depends on the network condition. In this section, we will identify the control procedure of two models and analyze them.

B. Distributed and Centralized QoS control

In the distributed architecture, call setup signaling, resource control, and policy control are done in the edge nodes. The signaling and resource control operation need to be done in both source and destination sides.

In the centralized model, call setup signaling and resource control are done in call server and resource control server respectively, and policy setup is done in the edge nodes. Gate control command should be delivered from the resource control server to the edge node.

Figure 2 and Figure 3 shows the control procedure of distributed and centralized models. Several control procedures need to be done during the call by call procedure. The processing time for each procedure is defined as follows.

T_s : processing time for signaling

T_r : processing time for resource control

T_p : processing time for policy activation

During T_s , call setup signaling such as SIP is processed in the call setup servers. During T_r , the resource and policy check for the requested call is performed. During T_p , gate control and traffic control parameter of the call is installed in the edge nodes.

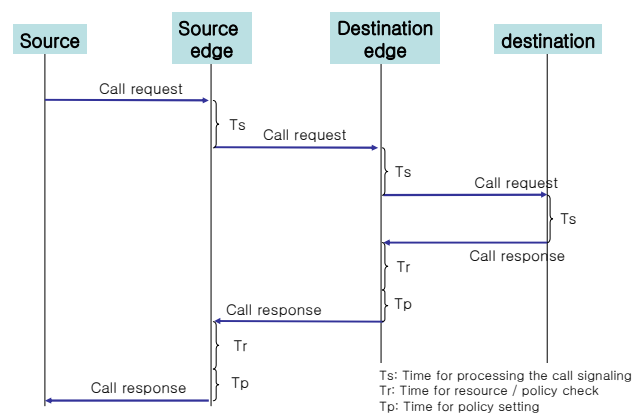


Figure 2 QoS control procedure for distributed model

As indicated in the Figure 2 and Figure 3, the centralized model needs the call server and resource control server (RACF) for the centralized processing. Distributed model doesn't require the separate server, because the control function is implemented in the edge node. According to the design principle of independency between service and transport, centralized model is

appropriate.

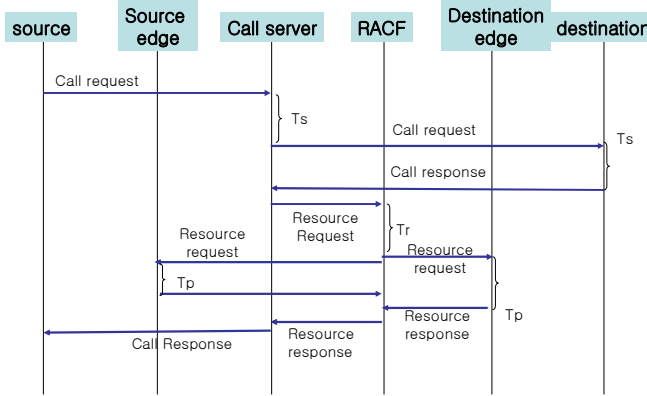


Figure 3 QoS control procedure for centralized model

In the centralized model, the network policy information can pre-provisioned in the network edge. The incoming media flow is mapped into the LSP tunnel. In this case, the resource control does not require the interaction between the resource control server and the edge node. In this case, policy enforcement procedure (noted as T_p in the centralized model) is not required.

The goal of this paper is to investigate the call by call resource control overhead in the two control models – distributed and centralized. The performance metrics are the processing delay for the end-to-end resource control. The transport delay of the message exchange depends on network topology and location of the servers. To focus on the effect of the control overhead, we ignore the transport delay.

First, we define the performance parameters.

N : Set of all the edge nodes in the network

$r_{i,j}$: call request rate from edge node i to edge node j

r_i : call arrival rate to edge node i

r : overall call arrival rate

t_p : average processing time for policy setting

t_s : average processing time for call signaling

t_r : average processing time of resource control

Note that t_p , t_s , and t_r depend on the processing power of the control entities.

In the centralized model, the call arrival rate to node i for policy setup (T_p) is

$$r_i = \sum_{j \in N} r_{i,j} + \sum_{j \in N} r_{j,i}$$

The request rate to signaling server and resource control server for processing the call setup signaling and resource control (i.e., T_s and T_r) is

$$r = \sum_{i \in N} \sum_{j \in N} r_{i,j}$$

In the distributed model, each node receives two requests when setup the call. First request is for processing the call setup signaling whose average processing time is t_s . The

second request is for resource control and policy setup whose processing time is $t_r + t_p$. The total request rate to node i is $2r_i$. The average processing time in node i is $\frac{t_s + t_r + t_p}{2}$. Similar message exchange is done when

terminating the call. Therefore the overall request rate for both setup and releasing is double of the call request.

C. Comparison of two control models

To provide stable control, the processing power of the control servers or edge nodes should be able to handle the control overhead generated by the call request.

In the distributed model, the arrival rate of the request needs to be less than the average processing rate in each node. Therefore,

$$\max_{i \in N} \{4r_i\} < \frac{2}{t_s + t_r + t_p} \quad (1)$$

In the centralized model, the request processing rates for the call server and resource control server are $1/t_s$ and $1/t_r$. Therefore

$$2r < \frac{1}{t_s} \text{ and } 2r < \frac{1}{t_r}$$

The edge node will be processing the policy setup (T_p)

$$\max_i \{2r_i\} < \frac{1}{t_p} \quad (2)$$

Under a given network condition, distributed and centralized model can be selected based on its control overhead. One of choice is to select the control model whose control overhead is small. Note that (1) is sufficient condition of (2), therefore, processing overhead of edge node in the centralized model is always smaller than that in the distributed model. Condition (1) is not critical when selecting the control model. The critical utilization of distributed and centralized model is defined as

$$\rho_{dist} = 2 \max_i \{r_i\} (t_s + t_r + t_p)$$

$$\rho_{cent} = 2r \max \{t_s, t_r\}$$

Distributed model perform better when $\rho_{dist} < \rho_{cent}$, i.e.,

$$\max_i \{r_i\} (t_s + t_r + t_p) < r \max \{t_s, t_r\}$$

Based on this result, the traffic distribution is critical when selecting the control model. When $\max \{r_i\}/r$ is high, i.e., majority of traffic is initiated from (or destined to) a single edge node, the centralized model performs better. When we assume that t_p , t_s , and t_r are same in both modes, more than 1/3 of traffic is concentrated in a single node, the centralized model will have better performance.

IV. NUMERICAL RESULT

We investigate the network overhead in terms of control processing delay. We will consider two scenarios. First scenario is uniform distribution case when control requests are uniformly distributed among all the edge nodes. Second one is hot spot distribution case when the

majority of calls are crowded in an edge node. An emergency situation or a special event may cause this situation.

In the simulation, we focus on the control processing time and ignore the propagation delay of the message exchange. The total number of node is 20. The processing rates are $1/t_p = 1/t_s = 1/t_r = 20$ (requests/sec). We examine the processing delay for the different call request rate r (calls/sec).

D. Uniform Distribution case

In uniform distribution case, the source and destination of call request are uniformly distributed to all the edge nodes. Figure 4 shows the average processing delay in the uniform distribution model. In the low call request, centralized model has smaller processing delay. As the request rate increases, the processing delay of centralized model increases more rapidly.

In the distributed model, the control overhead is distributed evenly among all the edge nodes. The processing delay increases slowly as the request rate increases.

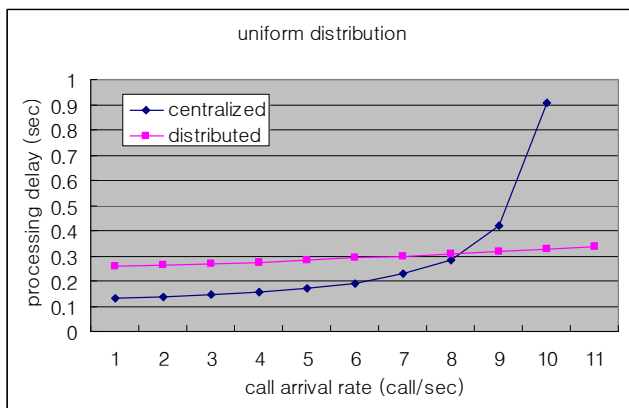


Figure 4 Average processing delay in uniform distribution

E. Hot Spot case

Hot spot distribution assumes that the majority of calls are crowded in one edge node. Figure 5 and Figure 6 show the average processing delay in the hot spot distribution. Figure 5 shows the case that 30 percent of calls go through the hot spot node. In this case average processing delay increases almost same rate as the call request increases. In the distribution model, most processing delay occurs at the hot spot node and the performance of hot spot node is worse than the average.

The performance of distributed model becomes worse as the degree of hot spot increases. In the 50 percent hot spot case in Figure 6, the performance of distribution model is worse than that of centralized case.

V. CONCLUSION

In this paper, we identify the three major parts of QoS processing – signaling, resource control, and policy setup. We presented the decision rule based on the traffic matrix and the average processing delay on each processing.

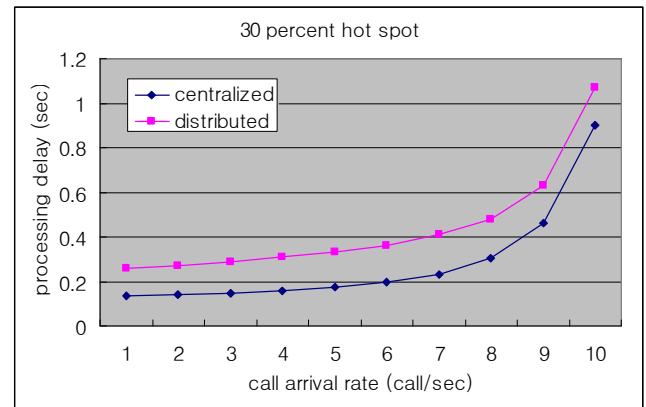


Figure 5 Average processing delay in 30% hot spot

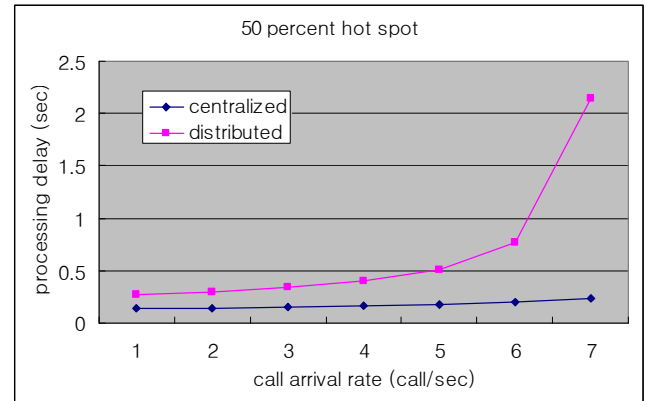


Figure 6 Average processing delay in 50% hot spot

Overhead of QoS control depends on the implementation details such as hardware implementation, processing power of servers, or efficiency of processing algorithm. Assuming the same processing power of both models, distributed model perform better when traffic distribution is uniform and request rate is high. Under the hot spot distribution, centralized model perform better.

The public network is designed to balance the traffic load among the network nodes and significant traffic unbalance may not happen. However, in the special cases such as an emergency situation, a special event, special topologies, or a military operation, the majority traffic can be crowded in one region. In this case,

Network providers have unique network deployment condition. Also in each region of network (e.g., access, metro, and core network) has different control overhead, dynamics, and traffic distribution. Control model needs to be selected considering the unique characteristics of the target network.

REFERENCES

- [1] Seong Wook Kang and Jae Yoon Kim, "Coming u-Health era"(in Korean), CEO information, Samsung Economic Research Institute(SERI), no. 602, May 2007
- [2] World Economic Forum, "Shaping the Global Agenda: The Shifting Power Equation", Annual Meeting Report, Jan. 2007
- [3] Jongtae Song, Mi Young Chang, Soon Seok Lee, and Jinoo Joong, "Overview of ITU-T NGN QoS control", IEEE Communication Magazine, vol. 45, no. 9, pp.116-123, Sept. 2007.

- [4] Jongtae Song, et al, "Scalable network architecture for flow-based traffic control", ETRI Journal, April 2008.
- [5] Jongtae Song, Mi Young Chang, Soon Seok Lee, and Jinoo Joung, "Overview of ITU-T NGN QoS control", IEEE Communication Magazine, vol. 45, no. 9, pp.116-123, Sept. 2007.
- [6] Bruno Ciciani et al, "A Hibrid Distributed Centralized System Structure for Transaction Processing", IEEE Trans. on Software Engineering, Vol.16, No. 8, August 1990.
- [7] M.A. Raza et al, "Dynamic Resource Management in Hybrid Architecture for OBS Control Plane", ICEE'07, April 2007
- [8] ITU-T recommendation Y.2012 "Functional Requirements and Architecture of the NGN", August 15, 2006.
- [9] ITU-T recommendation Y.2111 "Resource and Admission Control Functions in NGN", Sept. 1, 2006