A QoS Guarantee Strategy for Multimedia Conferencing based on Bayesian Networks

Junfei Huang and Guochu Shou
State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications, Beijing 100876, P. R. China.
Email: junfei.huang@outlook.com, gcshou@bupt.edu.cn.

Abstract—Service Oriented Architecture (SOA) is commonly employed in the design and implementation of web service systems. The key technology to enable media communications in the context of SOA is the Service Oriented Communication. To exploit the advantage of SOA, we design and implement a web-based multimedia conferencing system that provides users with a hybrid orchestration of web and communication services. As the current SOA lacks effective QoS guarantee solutions for multimedia services, the user satisfaction is greatly challenged with QoS violations, e.g., low video PSNR (Peak Signal-to-Noise Ratio) and long playback delay. Motivated by addressing the critical problem, we firstly employ the Business Process Execution Language (BPEL) service engine for the hybrid services orchestration and execution. Secondly, we propose a novel context-aware approach to quantify and leverage the causal relationships between QoS metrics and available contexts based on Bayesian networks (CABIN). This approach includes three phases: (1) information discretization, (2) causal relationship profiling, and (3) optimal context tuning. We implement CABIN in a real-life multimedia conferencing system and compare its performance with existing delay and throughput oriented schemes. Experimental results show that CABIN outperforms the competing approaches in improving the video quality in terms of PSNR. It also provides a one-stop shop controls both the web and communication services.

Index Terms—service oriented architecture; multimedia conferencing; Bayesian networks; context-awareness; causal relationship

I. INTRODUCTION

Service Oriented Architecture (SOA) is a collection of development principles for distributed service systems [1-2]. Web service is a major technology to implement SOA and it enables loosely coupled SOA and interoperable solutions across heterogeneous platforms and systems [3]. Service Oriented Communications (SOC), which uses the web services to enable communications, has already become a hot research issue in both industry and research domains [4-5]. The idea of using web services to support multimedia services is attractive as it can be easily integrated with other SOA-based business systems [6-7]. Furthermore, the web-based multimedia services enjoys better scalability, manageability and less cost for the development of multimedia systems [8]. Recent advances in web services technology have made it practical to enable web-based multimedia communications. With the convergence of Internet and Telecom networks, the integration of web and communication services is becoming an inevitable trend for the service providers [9-10].

Multimedia conferencing serves as the basis of many prevalent applications, e.g., audio/video conferencing, multi-player online games and remote learning [11]. The traditional multimedia conferencing systems are expensive and challenging to develop and maintain since they are typically monolithic systems [8]. Motivated by the principles of SOA, we have designed and implemented a web-based multimedia conferencing system [12] (see Fig. 1) based on reusable components, e.g., the services orchestration engine and telecom service engine. However, the system users’ satisfaction is still greatly challenged with QoS violations (e.g., low video PSNR and long playback delay) because:

1) On one hand, the multimedia communication services are characterized by stringent performance requirements, e.g., high-throughput, low-latency and error-resilience; On the other hand, the current SOA stack lacks of effective QoS guarantee mechanisms for multimedia services [3,6,8].

2) It is difficult to trace and identify the causal relationships between QoS metrics and available contexts\(^\dagger\). Furthermore, even if the relationship is determined, it is still a challenging task to carry out context adaption strategies due to the heterogeneity of conferencing participants’ networks environments, hardware, QoS requirements, etc.

3) The number of tunable contexts is very limited at the Internet Service Provider (ISP) side, as many parameters are under the control of Telecom operators and hardware manufacturers [13].

To address the first problem, the BPEL (Business Process Execution Language) service engine, which can describe and organize the services in a standardized manner, is employed for the hybrid services orchestration. Moreover, as the current data transmission protocols (e.g., SOAP and Restful) for web services are not capable of supporting the multimedia streaming well, the RTP (Real-time Transmission Protocol) is employed for media delivery.

On tackling the remaining two problems, the end-to-end QoS management [14] has proved to be a promising solution and the existing schemes can be generally divided into

\(^\dagger\)By the term ‘contexts’, we mean the available system information of application-level metrics (e.g., the user-perceived playback delay and stored video in the playback buffer) or network-level parameters (e.g., available bandwidth, packet loss rate or round trip time) obtained through specific monitoring or sensing mechanisms.
two branches: active probing (e.g., Keynote [15] and Gomez [16]) and passive monitoring (e.g., the CEM [17]). A major problem with both the active and passive schemes is that they either focus on the network level or application level metrics and adapt the tunable contexts accordingly. Therefore, many problems arise from this isolated working manner, e.g., the bandwidth-aware strategies aim at maximizing the throughput of the applications while the upper layer application can not directly benefit form that as it is a typical Constant Bit Rate (CBR) one. In this paper, we propose to analyze and exploit the relationships between the target QoS metrics and available contexts in a comprehensive way, i.e., examining the intrinsic causality between QoS metrics and contexts so as to find out the optimal adaption strategies. This research makes the following contributions:

1) We design and develop a SOA-based multimedia conferencing system that employs a novel hybrid web and communication services orchestration mechanism. Based on the implemented conferencing system, we conduct measurements of the QoS metrics and contexts.

2) We profile the causal relationships between the QoS metrics and contexts through the measurements. A context-aware approach based on Bayesian networks (CABIN) is proposed to quantify and leverage such relationships and it includes three key phases: (1) information feedback and training, (2) QoS-to-context mapping, and (3) optimal context adaption.

3) We evaluate the performance of CABIN in the multimedia conferencing system. Experimental results show that: (1) CABIN improves the average video PSNR (Peak-Signal-to-Noise Ratio) by up to 3.13 and 4.16 dB compared to the throughput oriented (TOR) [25-26] and delay oriented (DON) [27-28] schemes, respectively; (2) CABIN exhibits good performance in preventing the playback buffer starvation and reduces the average playback delay by up to 21.3ms compared to the TON scheme; (3) CABIN effectively exploits the available capacity for maximizing the video streaming throughput and increases the mean value by up to 77.5Kbps compared to the DON sheme.

The remainder of this paper is structured as follows. In Section 2, we briefly review the related work and discuss our research motivation. In Section 3, we describe the overall design and key components of CABIN in detail. The implementation and evaluation of the proposed CABIN in real multimedia conferencing system are provided in Section 4. Conclusions are given in Section 5.

II. RELATED WORK AND RESEARCH MOTIVATION

In this section, we firstly review the related work to this paper. Then, we discuss our research motivation from the practically observed contexts and QoS metrics.

A. Related Work

The related work to this paper can be classified into two categories: (1) service oriented communications, and (2) end-to-end QoS management solutions.

Decneut et al. [7] discuss the use of multimedia Web services in heterogeneous multimedia environments. The work is targeted at the distribution of multimedia data for a variety of client configurations and network conditions. The transmission of multimedia content is achieved using the now obsolete DIME attachment scheme over HTTP, TCP or UDP channels. A multimedia Web service is only allowed to send multimedia content within one single package. That means streaming data is not supported, or it has to be transmitted as if the packets are a big chunk of multimedia content. Zhang et al. [33] propose a framework based on SOAP-oriented component for multimedia Web service. In this work a multimedia Web service is separated into a control flow and a data flow. The control flow is transmitted via ordinary Web services whereas the data flow is managed by the existing multimedia infrastructure. The study proposed a SOAP enhancement so that multimedia data can be alternatively transmitted using SOAP. The enhancement includes the use of ‘boxcarring’ and batching of messages. However, since the transfer of each part of a big chunk of multimedia data requires a separate Web service request, the transfer may be too costly and inefficient
when streaming multimedia is transmitted using one separate request per packet.

Until now, much research has been done on the end-to-end QoS management. Active approaches require the injection of probe packets into the network. The pioneering active approach [29] traceroutes between 37 participating sites are collected and analyzed to characterize the end-to-end performance issues. The authors of [30] propose to detect path outage among hosts using ping and localizes the observed path outage with traceroute. PlanetSeer [31] relies on active probes to diagnose the root cause of Internet path failures that are detected by passively monitoring the end-users of a content distribution network service deployed on PlanetLab. Commercial systems, e.g., the Keynote [15] and Gomez [16] are also available to detect issues from the end-users perspective by active probing. All these work employ active probing while “Argus” purely depends on passive monitoring. Although the DiffServ architecture [32] supports end-to-end QoS guarantee, only a limit number of static QoS classes are provided. Obviously, it can not meet the active fine-grained QoS demands of diverse services. Moreover, end-to-end QoS guarantee by DiffServ may require the supports of underlying hardwares, which will introduce overhead to those core devices and result in low efficiency. In addition, fine-grained QoS quantitative guarantee is difficult to be implemented because these fine-grained service anomalies are hard to be traced among ISPs due to commercial security problem.

B. Research Motivation

In this subsection, we discuss the research motivation from two aspects: (1) the practically observed values of the QoS metrics and contexts, and (2) the causal relationships between them after the training and studying process.

1) Practically Observed QoS Metrics and Contexts: First, we depict the real-data measurements of the QoS metrics (video PSNR) and available contexts (video traffic rate, available bandwidth, etc.) from the multimedia conferencing system in Fig. 2. The experiments lasted for two months and we collected the data from more than 800 conference sessions for the campus and the Internet users. It can be observed from Fig. 2 that the causal relationships between the video traffic rate and other variables are ambiguous, as no variable changes regularly along with the video traffic rate. To further study the relationships, the raw data needs to be refined and processed. In this work, the Gaussian fitting is used to achieve the goal. Next, we shall present and analyze the fitting results.

2) Causal Relationships: After studying on the continuous context samples collected from the experimentations, we found that each context follows an approximate Gaussian distribution, which is a summation of Gaussian distributions. Due to the lack of space, we only present the video PSNR results in Fig. 3(a) to show that the contexts and QoS metrics in the experiments of this study approximately followed some Gaussian distribution. The other was that some context/QoS metric was approximated or fitted by a sum of Gaussian distributions, instead of one Gaussian distribution. This manner can reduce the fitting residuals as much as possible. In this study, the best fitting was marked out if it had the minimum root mean square error, which is often consistent with the sum of square error. It should be noted that the probability density distributions in Fig. 3(a) are inconsistent with what are expected to be obtained, since the contexts/QoS metrics are re-arranged according to the increment of video traffic rate depicted in Fig. 2. Fig. 3(b) presents the mapping of the values after the Gaussian fitting process and the relationship between them can be observed. However, these causal relationships represent time-varying characteristics as the video sessions advance. In the next section, we will describe how to determine and exploit the relationships in the proposed solution.

III. Proposed Solution

In this section, we outline the over design of the proposed solution and describe the functionality of its major components in detail. The system framework includes working components in both the source and the client side. The information monitor in the client side is responsible for providing the feedback information of all QoS metrics and contexts. The QoS metrics and contexts (e.g., available bandwidth and playback buffer size) can be obtained through specific communication protocols between server and clients. Estimating network status based on end-to-end measurements has been an active research for many years and numerous algorithms have been proposed
to achieve the goal. In this paper, the pathChirp [18] algorithm is employed to estimate the end-to-end available bandwidth with high accuracy and efficiency. Next, we will describe the three key phases of CABIN in detail.

A. Information Discretization

In order to exploit the causal relationships, the feedback QoS metrics/contexts must be properly trained. Fuzzy logic is a multi-valued logic that maps imprecise terms into crisp values. A fuzzy controller is composed of the inference system that includes a rule set, the input membership functions and the output variable. The input values go through a process of fuzzification, where they are converted in terms of the membership functions of the fuzzy sets. These sets are defined over the range of the fuzzy input values, and linguistically describe the variables universe of variation. However, fuzzy set theory requires experts to subjectively determine fuzzy sets and their membership functions for a continuous concept. In other words, these fuzzy sets and membership functions may vary from expert to expert. Hence, discretization results may be inconsistent from expert to expert from time to time. To eliminate subjectivity in determining fuzzy sets and membership functions and to assure consistency of discretization result, we propose an adaptive method to discretize a continuous context in this study. Formally, the probability density \( f(x) \) of an approximate Gaussian distribution can be defined as:

\[
f(x) = \sum_{i=1}^{n} a_i \cdot \exp \left( - \frac{(x - b_i)^2}{c_i} \right),
\]

in which \( n \) is the total number of sample values, \( a_i \) is the coefficient of the \( i \)th term of the probability density; \( b_i \) and \( c_i \) are expectation value and standard deviation of the \( i \)th term, respectively. In this study, a continuous context \( c \) is thought to have \( n \) discrete values if the probability density of its approximate Gaussian distribution has \( n \) terms because each term shows an identical numerical characteristic, i.e., \( a_i, b_i \) and \( c_i \) in the above formal definition. That is, the values covered by a term are more close in semantic than those covered by other terms. Once training data, a set of samples of a target context is collected. The QoS metrics/contexts training algorithm (see Algorithm 1) is referred to make them discrete.

\begin{algorithm}
\caption{Information Discretization.}
\begin{algorithmic}
\STATE \textbf{Input}: Sample values \( S \) of QoS metrics/contexts.
\STATE \textbf{Output}: Discrete value set \( DS \) of sample \( S \).
\STATE \textbf{Initialize}: \( D \leftarrow 0, M \leftarrow 0, DS \leftarrow S \);
\STATE 1: \( \{D, M\} \leftarrow \text{Dis-Mem}(S); \)\p \text{call the discretization and membership function}\#
\STATE 2: \textbf{for all} \( s \in S \) \textbf{do}
\STATE 3: \( f_D(s) \leftarrow \max\{f_D(s) | D \in D, f_D \in M\}; \)
\STATE 4: \( s \leftarrow D; \) \( \rho \) \text{s is discretized as discrete value} \( D^\rho \)
\STATE 5: \( s \leftarrow D; \)
\STATE 6: \textbf{end for}
\STATE 7: \textbf{return} \( DS; \)
\STATE Procedure \text{Dis-Mem}(S):
\STATE 1: \( PD \leftarrow \text{K-DENSITY}(S); \) \( \rho \) \text{estimate the probability density by kernel density estimation}\#
\STATE 2: \( \{GF, MSRM\} \leftarrow \text{GAUSSIAN-FITTING}(PD); \) \( \rho \text{do Gaussian fittings on} PD\rho \)
\STATE 3: \( G \leftarrow GF_{\text{max}}\left(\text{RMSE}, |\text{RMSE}|; \text{RMSE}, 1 \leq i \leq |\text{RMSE}|\right); \)
\STATE 4: \( \sigma_{\text{max}} \leftarrow 0; \)
\STATE 5: \( 	ext{temp} \leftarrow 0; \)
\STATE 6: \textbf{for each} \( t \in G \) \textbf{do}
\STATE 7: \( \text{temp} \leftarrow \text{temp} + 1; \)
\STATE 8: \( M[t] \leftarrow t; \)
\STATE 9: \( D[t] \leftarrow \text{cnt}; \)
\STATE 10: \( \sigma_{\text{max}} \leftarrow \max\{\sigma_{\text{max}}, \sigma_{\text{cnt}}\}; \)
\STATE 11: \textbf{end for}
\STATE 12: \textbf{for each} \( m \in M \) \textbf{do}
\STATE 13: \( m \leftarrow \frac{\sigma_{\text{max}}}{\sigma_{\text{cnt}}}; \) \( \rho \text{ normalize the coefficient} \ a \) (see Equ. (1)) of \( m \) by dividing \( \sigma_{\text{max}}\rho \)
\STATE 14: \textbf{return} \( \{D, M\} \)
\end{algorithmic}
\end{algorithm}

B. Causal Relationship Profiling

After the available contexts have been trained to be in the right form, the QoS-to-context mapping is performed to find out causal relationships between a QoS metric and these contexts in a real-time manner. A QoS metric is thought to be caused by a context if it changes along with the context, i.e., a causal relationship exists in the QoS metric and corresponding context. The Bayesian network [19] is employed in this paper for studying and modeling these causal relationships.

In the proposed solution, the QoS metric to be studied is considered as a child node, and its contexts are regarded as ancestor nodes. Bayesian network structure learning algorithms are employed here to construct a directed acyclic graph, which qualitatively models the causal relationships between the QoS metrics and the corresponding contexts.
Formally, a Bayesian network is a directed acyclic graph with a conditional probability table associated with each discrete node. A conditional probability distribution is associated with a continuous variable. Node $v_a$ is a parent of node $v_b$, which in turn is a child of node $v_a$, if a directed arc from $v_a$ to $v_b$ exists. $G = (V, E)$ denotes the network topology: $V$ denotes the set of nodes and $E$ represents the set of directed arcs. By this catenation connection, the conditional probability table space will be saved exponentially since only parents instead of all ancestors are considered to compute conditional probability. The conditional probability of $v_i$, given its ancestors has been encoded in the probability tables of its parents $p_{v_i}$. For instance, we assume the mean number of discrete values of a node $v_i$ is $e_v$, the mean number of parents is $e_a$ and the mean number of ancestors is $e_a$. Thus, Bayesian network only requires $v_i$ slots instead of $e_v$ to store the conditional probability table for node $v_i$, i.e., the space is saved about $(1 - \bar{r}^{a-a})$, which approaches to one exponentially. In order to study the contexts for a target QoS metric using Bayesian network, the QoS metric to be evaluated is considered as a child node, and its corresponding contexts are considered as ancestor nodes. BN structure learning algorithms are referred to construct a direct acyclic graph, which represents the causal relationships between the QoS metric and corresponding contexts. The $K2$ algorithm [20] is applied to learn network structure in this paper. Parents of a node $v$ can only be chosen from the nodes before $v$ in the order. The following three paragraphs convey a node-ordering method.

![Fig. 4. The resulting Bayesian network structure for the QoS-to-context mapping.](image)

The following steps are taken to exploit the causal relationships between a QoS metric and its contexts. The parent contexts are the causes of a QoS metric. More specifically, parent nodes are direct causes of a child node while other ancestor nodes are the indirect causes. Therefore, the QoS-to-context mapping phase analyzes the relationships in a systematic manner. The resulting Bayesian network structure for the considered QoS metrics/contexts in this paper is presented in Fig. 4.

### C. Optimal Context Tuning

Bayesian network parameter training is used to learn the conditional probability tables from training data after the Bayesian network structure is constructed by the the QoS-to-context mapping. In this study, we use Bayesian parameter updating from complete data, i.e., the samples of contexts and QoS metrics. A Bayesian network is completely defined once both its structure and parameters are determined. Then, given the observed data, the Bayesian network inference can be applied to compute the marginal on the specified query nodes. The marginal on a query node $H$ given observed data $E$ can be viewed as the conditional probability $P(H|E)$, which is a quantitative metric. Therefore, we use the marginal on a QoS metric node given the observed contexts to guarantee the QoS metric.

The optimal context adaption is to carry out with a context value that guarantees the QoS metric to the target value with probability $p$. When a user picks up a specific value of the QoS metric, the causal contexts of the QoS metric are tuned to appropriate values with regard to the specific QoS metric value. The optimal context adaption steps are summarized in Algorithm 3. Only parent contexts will be tuned and other contexts will be left untouched. It makes sense by the theory that every set of nodes in a Bayesian network is conditionally independent of QM when conditioned on the Markov blanket QM of the node QM, which consists of its parents, children and the other parents.

### IV. IMPLEMENTATION AND EVALUATION

In this section, we evaluate the efficacy of CABIN in our real multimedia conferencing system by comparing it with existing QoS guarantee solutions. First, we describe the evaluation methodology that includes the experimental setup, performance metrics, compared approaches and evaluation
scenario. Then, we present and discuss the evaluation results in detail.

A. Implementation Issues

In this section, we describe the architecture of our multimedia conferencing system and the system implementation of CABIN. Our multimedia conferencing system includes some real-time communication services, e.g., the multiple party call control service and short message service. The system also contains some non real-time Web services, e.g., the user authentication and charging services. Therefore, the system must provide the capability to execute those hybrid services. We have developed the SOA based multimedia conferencing system and the system framework is presented in Fig. 5. Here, the Enterprise Service Bus (ESB) is based on Servicemix [21], which is an open source project. The Telecom service engine kernel is based on another open source project Mobicents [22], which is based on the JAIN SLEE specification and can be easily connected to different application platforms or network elements via a resource adaptor mechanism. In the system design, the BPEL is used for automatically translating the multimedia conferencing processes and properties into a corresponding formal model.

The IBM X3800 8866 serves as the SIP server while the AINOL HD media server and the Radisys CMS9000 are alternate in user for the media process. The conferencing system includes the audio/video services, shared contents and instant messages as depicted in the user interface. Interested readers could refer to [12] for the detailed introductions to our system. Our implementation includes modifications at the media server and the client. The main components of the proposed CABIN are implemented on the Enterprise Service Bus. We significantly extend and modify the system (about 4000 lines of JAVA code and 2000 lines of C++ code) to implement the CABIN’s components. The function of video traffic rate adaption is realized by the JAVA interface provided by the media server manufacturer.

B. Evaluation Methodology

1) Experimental Setup: We adopt the Exata [23] as the network emulator. Exata is advanced edition of QualNet [24] in which we can perform semi-physical emulations. In order to implement the realistic emulations of practical network environments, we set the number of nodes and their locations based on the network topology as depicted in Fig. 5. The real computers are connected to the emulation server through the Exata 2.1 connection manager. The parameters of the injected background traffic are listed in Table 1. The video traffic rate is adapted every 2.5 seconds in all the experiments.

2) Performance Metrics:
- PSNR (Peak Signal-to-Noise Ratio) is a standard metric of video quality and is a function of the mean square error between the original and the received video frames. If a video frame is dropped or past the deadline, it is considered to be lost and is concealed by copying from the last received frame before it.

- Playback delay. The playback delay process corresponds to the time left to playback the video data at the tail of the playback buffer or, equivalently, the buffered video length plus the time left to start video playback.
- Throughput. The throughput is the bandwidth used for the video application and it represents the capability of the competing approaches to effectively exploit the available bandwidth. This network-level metric is measured by the Exata Analyzer.

3) Compared Approaches: We compare the proposed solution with the following context-aware approaches:
- Throughput oriented approach [25-26]. The throughput oriented (TON) adaption aims at maximizing the application’s throughput by adapting the video traffic rate. In this work, we assume the TON scheme adjusts the video encoding rate according to the average ‘loss-free’ bandwidth of all the participants.
- Delay oriented approach [27-28]. The delay oriented (DON) scheme adjusts the video traffic rate based on the feedback info of playback buffer size from the clients. Playback buffer starvation occurs when a video frame is requested by the application layer for playback before being fully received. From a user QoS perspective, playback buffer starvation is undesirable because it results in video frame freezes and re-buffering. Therefore, the constraint for the video traffic rate is equal to prevent the buffer underflow.

4) Evaluation Scenario: In order to fully evaluate the performance of the compared approaches, we conduct experiments with different numbers of participants: 4, 8, 12 and 16. For the standard deviation results, we repeated each round of experimentation more than 5 times and obtain the average values with a 95% confidence interval.

C. Experimental Results

1) PSNR: As depicted in Fig. 6, CABIN achieves higher average PSNR values and lower deviations than the TON and DON schemes in all the evaluation scenarios. It can be observed that the superiority of CABIN over the competing schemes becomes more distinct as the number of participants increases. The results indicate the relationships between the QoS metrics and corresponding contexts become more complex in case of many online users. The average video quality degrades with the increase in the number of users for all the approaches. Indeed, due to the capability limitations of the media server, the user-perceived video quality is a challenging problem to be tackled in our future work. In general, CABIN
improves the average video PSNR by up to 3.03 and 4.16 dB compared to the TON and DON scheme and the video PSNR per frame is shown in Fig. 7.

Fig. 6. Average PSNR values for all the competing approaches in different experimental scenarios.

Fig. 7. Instantaneous PSNR value of all the competing approaches for each video frame indexed from 1 to 300.

2) Playback delay: The playback delay is defined as the time left to playback the video data at the tail of the playback buffer. In Fig. 8, we plot the average playback delay and DON maintains the lowest latency of all the competing approaches. In order to have a microscopic view of the results, we show the playback delay process of [0, 250] seconds in Fig. 9. It can be observed from Fig. 9 that the DON is able to regulate the playback buffer contents based on its size by adjusting the video traffic rate to maintain enough data to compensate for the time-varying networks status. Furthermore, it avoids the playback buffer starvation and re-buffering. The buffering behavior is also correlated with the available bandwidth. Although the performance of CABIN is generally inferior to that of DON, it still exhibits good performance in preventing the playback buffer starvation.

Fig. 8. Average playback delay of all the participants in the experimental scenarios.

Fig. 9. Instantaneous value of playback delays during the interval of [0, 250] sec.

3) Throughput: Fig. 10 plots the average throughput values for all the approaches under different scenarios. The TON approach achieves the highest throughput as it generally adjusts the video traffic rate to larger values than the competing schemes. However, the real-time video applications can not directly benefit from the throughput gains as the end-to-end delay increases simultaneously. For real-time video applications, delay is another key performance since each video packet is associated with a decoding deadline. The overdue packets will be discarded at the client side since they can no longer contribute to the decoding process. Fig. 11 illustrates the instantaneous throughput values of all the competing schemes during the interval of [0, 300] second.

D. Discussion

The evaluation results demonstrate that the proposed CABIN outperforms the TON and DON schemes in improving the target QoS metric (i.e., the video PSNR). Due to the lack of space, we do not present the results of computational cost of CABIN and its competing models in this section.
Indeed, the performance gain of CABIN is at the sacrifice of computation and network overhead. However, with the advances in the system hardware and software, this overhead is negligible comparing to the service creation and execution time. On the other hand, the major flaw with both TON and DON schemes is that they are highly dependent on one specific context. Therefore, the accuracy of obtained contexts imposes significant impact on their performances. It is well known the networks status monitoring is still a challenging problem remains to be largely explored. As CABIN takes all the available contexts into consideration in a comprehensive manner, it is less single-context-dependent and thus achieves better results.

V. Conclusion

In this paper, we have presented the CABIN, a hybrid web and communication service framework to comprehensively exploit the causal relationships between the QoS metrics and corresponding contexts. CABIN offers a one-stop-shop integrating web and communication services, and a flexible scheme to orchestration them. The performance of CABIN is evaluated in our real multimedia conferencing system [35][36] and experimental results demonstrate its superiority over existing approaches. As future work, we will consider evaluating the performance of CABIN in other multimedia conferencing services [45], e.g., the audio service.

REFERENCES

[1] S. Tonnies, B. Kohnecke, P. Hennig and W. T. Balke, “A service oriented architecture for personalized rich media delivery,” in Proc. of the IEEE SCC, 2009.

[2] S. Heinzel et al., “A scalable service-oriented architecture for multimedia analysis, synthesis and consumption,” International Journal of Web and Grid Services, vol. 5, no. 3, pp. 219-260, 2009.

[3] J. Wu, B. Cheng, C. Yuen, Y. Shang, J. Chen, “Distortion-Aware Concurrent Multipath Transfer for Mobile Video Streaming in Heterogeneous Wireless Networks,” IEEE Transactions on Mobile Computing, vol. 14, no. 4, pp. 688-701, 2015.

[4] W. Chou, L. Li and F. Liu, “Web Services Methods for Communication over IP,” in Proc. of the IEEE ICWS, 2007.

[5] I. Brunkhorst, S. Tonnies and W. T. Balke, “Multimedia content provisioning using service oriented architectures,” in Proc. IEEE ICWS, 2008.

[6] J. Wu, C. Yuen, B. Cheng, Y. Shang, J. Chen, “Goodput-Aware Load Distribution for Real-time Traffic over Multipath Networks,” to appear in IEEE Transactions on Parallel and Distributed Systems.

[7] I. Brunkhorst, S. Tonnies and W.T. Balke, “Multimedia content provisioning using service oriented architectures,” in Proc. of the IEEE ICWS, 2008.

[8] J. Wu, B. Cheng, Y. Shang, J. Huang, J. Chen, “A Novel Scheduling Approach to Concurrent Multipath Transmission of High Definition Video in Overlay Networks,” Journal of Network and Computer Applications, vol. 44, pp. 17-29, 2014.

[9] D. Lam and D. Rossiter, “A Web Service Framework Supporting Multimedia Streaming,” IEEE Trans. Services Computing, 2012.

[10] J. Wu, X. Qiao, Y. Xia, C. Yuen, and J. Chen, “A Low Latency Scheduling Approach for High Definition Video Streaming in a Heterogeneous Wireless Network with Multi-homed Clients,” Multimedia Systems, vol. 21, no. 4, pp. 411-425, 2015.

[11] F. Paolo and C. Venezia, “Communication Web Services and JAIN-SLEE Integration Challenges,” International Journal of Web Services Research, vol.5, no.4, pp. 59-78, 2008.

[12] J. Wu, Y. Tian, B. Cheng, Y. Shang, “Comprehensively Context-aware Approach to Guarantee Multimedia Conferencing Services,” China Communications, vol. 10, no. 9, pp. 53-64, 2013.

[13] C Fu, F Khendek, and R. Gilmore, “Signaling for multimedia conferencing in 4G: the case of integrated 3G/MANETs,” IEEE Commun. Mag., vol. 44, no. 8, pp. 90-99, 2006.

[14] J. Wu, C. Yuen, N.-M. Cheung, J. Chen, “Delay-Constrained High Definition Video Transmission in Heterogeneous Wireless Networks with Multi-homed Terminals,” to appear in IEEE Transactions on Mobile Computing.

[15] X. Lin, B. Cheng and J. Chen. “Context-aware end-to-end QoS Qualitative Guarantee and Quantitative Guarantee based on Bayesian Networks,” Computer Communications, vol. 17, no. 6, pp. 2132-2144, 2010.

[16] H. Yan et al. “Args: End-to-End Service Anomaly Detection and Localization From an ISPs Point of View,” in Proc. of the IEEE INFOCOM, 2012.

[17] J. Wu, Y. Shang, X. Qiao, B. Cheng, J. Chen, “Robust Bandwidth Aggregation for Real-time Video Delivery in Integrated Heterogeneous Wireless Networks,” Multimedia Tools and Applications, vol. 74, no. 11, pp. 4117-4138, 2015.

[18] J. Wu, B. Cheng, C. Yuen, N.-M. Cheung, J. Chen, “Enabling Adaptive High-Frame-Rate Video Streaming in Mobile Cloud Gaming Applications,” to appear in IEEE Transactions on Circuits and Systems for Video Technology.

[19] Keynote systems, http://www.keynote.com/

[20] J.-Y. Wu, X.-Q. Qiao, B. Cheng, J.-L. Chen, Y.-L. Sun, “End-to-end QoS Guarantee for Delay-Sensitive Mobile Multimedia Conferencing,” Chinese Journal of Computers, vol. 36, no. 7, pp. 1399-1412, 2013.

[21] Gomez systems, http://www.gomez.com/.

[22] J. Wu, Y. Shang, C. Yuen, B. Cheng, and J. Chen, “TRADER: A Reliable Transmission Scheme to Video Conferencing Applications over the Internet,” Journal of Network and Computer Applications, vol. 44, pp. 161-171, 2014.

[23] D. Cheffines, F. Bustamante and Z. Ge, “Crowdsourcing service-level network event monitoring,” ACM SIGCOMM Computer Communication Review, no. 4, vol. 40, pp. 387-398, 2010.

[24] V. Ribeiro, R. Riedi, R. Baraniuk, J. Navratil and L. Cottrell, “pathChirp: Efficient available bandwidth estimation for network paths,” Passive and Active Measurement Workshop, 2003.

[25] J. Wu, C. Yuen, M. Wang, J. Chen, “Content-Aware Concurrent Multipath Transfer for High-Definition Video Streaming over Heterogeneous Wireless Networks,” in Proc. of the IEEE ICWS, 2007.

[26] G. F. Cooper, et al., “A Bayesian method for the induction of probabilistic networks from data,” Machine Learning.

[27] ServiceMix, http://servicemix.org.

[28] Mobicets Project, https://mobicents.dev.java.net/.
[29] J. Wu, J. Yang, X. Wu, and J. Chen, “A Low Latency Scheduling Approach for High Definition Video Streaming over Heterogeneous Wireless Networks,” in Proc. of IEEE GLOBECOM, 2013.
[30] Exata, http://www.scalable-networks/exata
[31] QualNet, http://www.scalable-networks/qualnet
[32] J. Wu, B. Cheng, C. Yuen, N.-M. Cheung, J. Chen, “Trading Delay for Distortion in One-Way Video Communication over the Internet,” to appear in IEEE Transactions on Circuits and Systems for Video Technology.
[33] K. Miller, E. Quacchio, G. Gennari, and A. Wolisz, “Adaptation algorithm for adaptive streaming over HTTP,” In Proc. of the IEEE International Packet Video Workshop, 2012.
[34] J.-Y. Wu, B. Cheng, G.-S. Nan, J.-L. Chen, “A Novel Joint Source-Channel Coding scheme for Mobile Video Delivery in Heterogeneous Wireless Networks,” Chinese Journal of Computers, vol. 38, no. 2, pp. 439-454, 2015.
[35] L. De Cicco, S. Mascolo, and V. Palmisano, “Feedback Control for Adaptive Live Video Streaming,” In Proc. of the ACM Multimedia Systems Conference, 2011.
[36] R. Mok, X. Luo, E. Chan, and R. Chang, “QDASH: a QoE-aware DASH system,” In Proc. of the ACM Multimedia Systems Conference, 2012.
[37] J. Wu, J. Yang, Y. Shang, B. Cheng, J. Chen, “SPMLD: Sub-Packet based Multipath Load Distribution for Real-time Multimedia Traffic,” Journal of Communications and Networks, vol. 16, no. 5, pp. 548-558, 2014.
[38] V. Paxson, “End-to-end routing behavior in the Internet,” ACM SIGCOMM Computer Communication Review, no. 5, vol. 36, pp. 41-56, 2006.
[39] J. Wu, Y. Shang, J. Huang, X. Zhang, B. Cheng, J. Chen, “Joint Source-Channel Coding and Optimization for Mobile Video Streaming in Heterogeneous Wireless Networks,” EURASIP Journal on Wireless Communications and Networking, pp. 1-16, 2013.
[40] N. Feamster, D. Andersen, H. Balakrishnan, and M. Kaashoek, “Measuring the effects of internet path faults on reactive routing,” in Proc. of the ACM SIGMETRICS, 2003.
[41] M. Zhang, C. Zhang, V. Pai, L. Peterson and R. Wang, “PlanetSeer: Internet path failure monitoring and characterization in wide-area services,” in Proc. of the USENIX OSDI, 2004.
[42] J. Wu, Y. Shang, B. Cheng, B. Wu, J. Chen, “Loss Tolerant Bandwidth Aggregation for Multihomed Video Streaming over Heterogeneous Wireless Networks,” Wireless Personal Communications, vol. 74, no. 2, pp. 1265-1282, 2014.
[43] S. Blake, et al., An architecture for differentiated services, IETF RFC 2475, 1998.
[44] J. Zhang, L. Zhang, F. Quek, and J. Chung, “A service-oriented multimedia componentization model,” International Journal of Web Services Research, vol. 2, no. 1, pp. 54-76, 2005.
[45] J. Wu, B. Cheng, Y. Shang, C. Yuen, J. Chen, “A Novel Transmission Scheme to Inter Destination Video Synchronization,” IET Communications, vol. 9, no. 5, pp. 603-612, 2015.