Dynamic IMS Reconfiguration using Session Migration for Power Saving

Satoshi Komorita, Manabu Ito, Yoshinori Kitatsuji, and Hidetoshi Yokota
KDDI R&D Laboratories, Inc.
Fujimino-shi, Saitama, JAPAN
e-mail: {sa-komorita, mn-itou, kitaji, and yokota}@kddilabs.jp

***Abstract*— Legacy telecommunication services have been tending to shift to All-IP networks from closed and circuit switched networks by the deployment of high bandwidth and stable mobile broadband networks such as Long Term Evolution (LTE). The shift to an All-IP network leads to the integration of telecommunication services and internet services, which are called Over-The-Top (OTT) services. The trend towards increasing and unpredictable traffic makes it important for the IP Multimedia Subsystem (IMS), which is a call control system running over the All-IP network, to be flexible and reliable. With that goal, the authors have proposed an IMS reconfiguration mechanism using call session state migration, which allows the IMS to distribute live call sessions as required by the failure of IMS servers and the quantity of calls in progress. Furthermore, in this paper we propose a model to determine the configuration of the IMS dynamically from the perspective of electrical power consumption and performance. A formula based on the model calculates which IMS servers should be active and which call sessions should move to which IMS servers to reduce the overall power consumption, using linear programming. Moreover, this paper implements the proposal and shows its results.**

Keywords-IMS; SIP; Session Migration; Power Saving; Linear Programming; Optimization; CPLEX

I. INTRODUCTION

In order to provide voice over IP (VoIP) service, mobile network operators (MNOs) are developing the IP multimedia subsystem (IMS) [1], which is a call control system over an IP network (adopted in GSMA [2]), to ensure interconnectivity between MNOs around the world. MNOs also plan to launch the Rich Communication Suite (RCS) [3], which will enable mobile phone users to use instant messaging, live video sharing, and file transfer over the IMS. In addition, through leveraging their IMS development, MNOs are now looking into the possibility of integrating Over-The-Top (OTT) services such as Google services, Facebook, and Skype.

The support of these future services in addition to legacy telecommunication services such as voice and Short Message Service (SMS) has resulted in an increase in signaling traffic in the IMS. It is difficult to estimate the required system capacity for the increase because OTT services are created by third parties outside MNO's awareness. To continue to deal with increasing and unpredictable traffic without investing in extra capacity, enhanced flexibility is desirable in the operation of the call session control functions (CSCFs) of which the IMS is composed. For example, in case of server failure or disaster, it is desirable to ensure reliability by continuing the processing of CSCFs on other servers, thus preventing service disruption. On the other hand, it is desirable to minimize the number of running servers, to decrease power consumption.

In order to realize the desired flexibility, we have proposed call session state migration [4], which allows call/service sessions to continue to be processed on another CSCF. The proposed method mainly involves a migration of the session states between CSCFs and a routing mechanism for control messages in the IMS without extensions to the standard procedure for control of a call/service session. However, the MNOs still need to decide which sessions should move to which CSCFs and which server can be stopped, even if the method provides for flexibility in the IMS configuration. This makes IMS reconfiguration more difficult as the number of CSCFs increases.

Thus, in this paper, we propose a model of IMS reconfiguration to determine the IMS configuration dynamically and easily so as to minimize power consumption and provide enough performance. The model incorporates the power consumption of the IMS so that the IMS can provide enough performance, while considering the power consumption of the reconfiguration, including all session state migration and running CSCFs. We solve the model and determine the reconfiguration by using linear programming [5]. Furthermore, we implement our proposal and the measurement metrics required for our model, and show the result of our proposal in a real environment.

The rest of this paper is organized as follows. Section 2 describes the IMS architecture and the call session state migration. Section 3 proposes a model of the IMS reconfiguration and its formula for linear programming. Section 4 shows our implementation and results of the experiment. Finally, Section 5 concludes this paper.

II. CALL SESSION STATE MIGRATION

A. IMS Architecture and Call Session Control

Fig. 1 shows the basic IMS network configuration. The following IMS components are located in the IMS core network that an MNO manages and operates: an Home Subscriber Server (HSS), which is a database server for managing subscribers, an Serving Call Session Control Function (S-CSCF), which is the main Session Initiation Protocol (SIP) [6] server for call control, a Proxy CSCF(P-

 ISBN: 978-1-61208-279-0

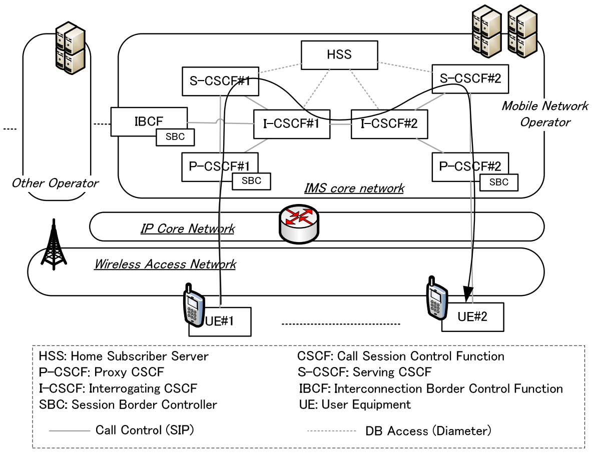

Figure 1. Basic IMS Architecture

CSCF) that communicates with the User Equipment (UE) directly and establishes a secure connection to it, an Interrogating CSCF (I-CSCF), which is a kind of resolver of SIP message routing, and an Interconnection Border Control Function (IBCF) that is a gateway for other MNO's IMSs. Furthermore, an Session Border Controller (SBC) providing a variety of functions for security and connectivity (e.g., access control, topology hiding, NAT traversal, protocol interworking, and media monitoring) is defined and often integrated into the P-CSCF and the IBCF in a real operation. A UE connects with the IMS through the IP core network and wireless access networks such as EV-DO [7] and Long Term Evolution (LTE) [8].

In the IMS, SIP is used for call control between CSCFs and UEs. First, a UE registers with the IMS before obtaining IMS services. The UE conducts its registration with an S-CSCF via a P-CSCF and an I-CSCF. The S-CSCF assigned by the I-CSCF verifies the UE based on its authentication information stored in the HSS. After that, the UE uses IMS services by sending and receiving SIP messages to/from a correspondent UE via the P-CSCF, S-CSCF, and I-CSCF. For example, when the UE makes a call, the UE sends an INVITE message to the correspondent UE and establishes a session to communicate the required information. The exchanged SIP messages take, as shown in Fig. 1, the following path: UE#1 (caller), P-CSCF#1, S-CSCF#1, I-CSCF#1, I-CSCF#2, S-CSCF#2, P-CSCF#2, and UE#2 (callee). Note that the assigned P-CSCF and S-CSCF of UEs hold states of register and sessions and each SIP message includes route information to their CSCFs, thus once CSCFs begin to handle the session of the UEs, it is not possible to change CSCFs to process the sessions in the standard procedures.

B. *Call Session State Migration and Determinaton of IMS Reconfiguration*

We have proposed a call session state migration mechanism [4], which allows sessions to continue to be processed on another CSCF in order to cope with server failure and dynamically change the number of physical servers to match the network and resource usage situation.

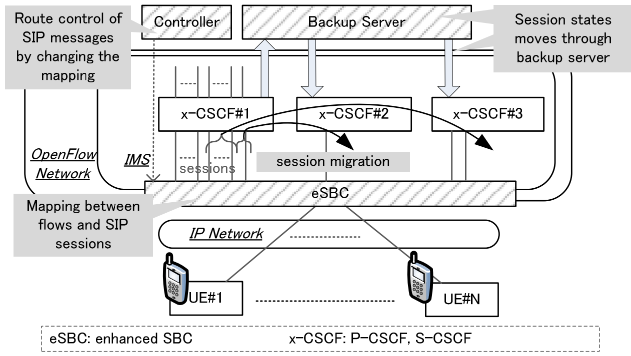

Figure 2. Call State Session Migration

The method mainly consists of the migration of session states between CSCFs and route control of SIP messages as shown in Fig. 2. The session states are migrated through a backup server in order to backstop server failure. The routing of SIP messages is controlled by using OpenFlow [9] to hide changes in the IP address of CSCFs. OpenFlow is a flow-based routing protocol, and we enhanced SBC to map between SIP sessions and flows of the OpenFlow network. This mechanism supports granular session migration to realize effective resource use.

While the session migration mechanism has been proposed, MNOs still need to decide which sessions should move to which CSCFs and which server can be stopped. If there is obvious trouble such as a server failure, MNOs can decide easily. However, if there are a few hundred servers in the IMS, it is difficult to estimate outcomes and quickly make a decision appropriate to the situation. Thus, a model to determine the IMS configuration dynamically is needed to support the operation of a flexible IMS.

C. *Related Work*

The study of optimized configuration in networks and SIP servers has a long history of improving performance such as traffic throughput, and maximizing the amount of processing using limited resources. Some studies [10][11] tried load balancing of processing to improve the performance of SIP servers. Another investigation [12] tried to propose an effective route for reducing message delay. Analysis of the IMS, which has additional processing compared with basic SIP procedures, has also been undertaken. N. Rajagopal *et al.*[13] analyzed IMS processing based on queuing of SIP messages and estimated the overload of each CSCF based on the SIP processing sequence.

The above works focused mainly on calculating the capability of the IMS in advanced before operation. On the other hand, the concept of effective resource use on demand has been also proposed, based on the perspective of flexible resource balancing using the recent development of virtualization technologies, rather than improvement in performance using limited resources [14]. This study proposed live session migration between physical servers by using a virtualized machine's live migration on demand. V. Petrucci *et al.* [15] have proposed the concept of calculating the required minimum number of servers and stopping the extra servers to save power in cluster environments.

However, in dynamic IMS reconfiguration, we need to consider the specific IMS configuration and the cost of the reconfiguration, and in particular, live session migration. In addition, some metrics for the calculation must be measured in a real environment to evaluate their adequacy.

III. PROPOSED MODEL OF DYNAMIC IMS RECONFIGURATION

A. Modeling Aproach based on Power Consumption

As an index to characterize IMS reconfiguration, we focus on the reduction of power consumption of the IMS because the purpose of a flexible IMS is to supply sufficient IMS capacity on demand while using minimum facilities. In our approach, we estimate all power consumption required for running the IMS and executing the reconfiguration, and use a linear programming method to try to reduce the total amount while still providing stable IMS service.

In contrast to a static IMS configuration, a dynamically reconfigurable IMS can change its configuration as many times as needed depending on the momentary situation, such as the number of calls. Thus, we also consider the cycle of IMS reconfiguration. Otherwise, even if the number of servers and their power consumption could be reduced on a temporary basis, the total power consumption might actually increase if the reconfiguration and call session state migration happened frequently.

B. Reference Model of IMS Reconfiguration

Fig. 3 shows a reference model of IMS reconfiguration, consisting of two IMS configurations and the transition between them. In the IMS, there are several nodes, which are fundamental physical servers. An Functional Entity (FE) is a running and session-processing entity such as a P-CSCF on a node. Here, for simplicity, just one FE will run on one node and just one type of FE shall exist in this model so that we can consider this model for each type of FE independently. In addition, nodes in the IMS can be turned off and on for power saving.

The following parameters are introduced for nodes and FE of the reference model.

$$
\begin{aligned}
&N = \text{a set with all nodes}\\
&B_n = \text{base power consumption by node\#}n,\ n \in N\\
&y_n = \begin{cases} 1 & \text{if node\#}n \text{ is on} \\ 0 & \text{if node\#}n \text{ is off} \end{cases},\ \{y_n\}\ \text{binary (integer)}\\
&M_n = \text{total number of sessions on node\#n}\\
&K_n = \text{capacity of FE on node\#}n\\
&\tau = \text{base power consumption of FE}\\
&\varepsilon = \text{power consumption for running a single session}\\
&Z \equiv \text{a state of configuration given by } \{y_n, M_n\}, n \in N\\
&R = \text{runing time of state Z}
\end{aligned} \tag{1}
$$

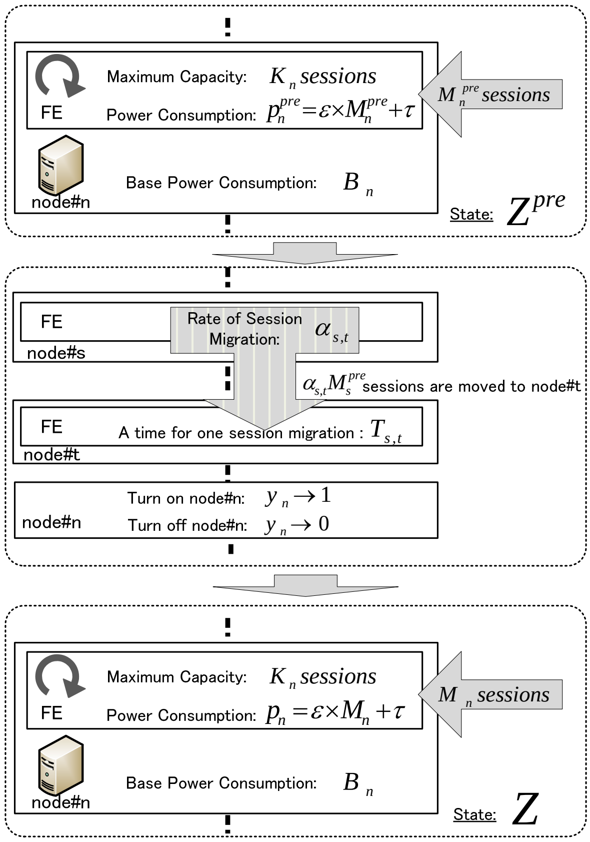

Figure 3. Reference Model of IMS Reconfiguration

Then, the IMS executes the reconfiguration by performing session migrations and turning nodes on/off. Previous state Z^{pre} makes a transition to the next state Z. The parameters for the transition are as follows:

$$
\begin{aligned}
&\alpha_{s,t} \equiv \text{fraction of sessions moved from FE on node\#}s\\
&\qquad \text{to FE on node\#}t, s \in N, t \in N\\
&C_{s,t} \equiv \text{power consumption of moving a single session}\\
&\qquad \text{from FE on node\#s to node\#t. Note that } C_{n,n} = 0\\
&T_{s,t} \equiv \text{time for moving a single session}\\
&\qquad \text{from FE on node\#}s \text{ to node\#t}\\
&\theta_n = \left| y_n - y_n^{pre} \right|,\ y_n^{pre} = \text{previous on/off state of node\#}n\\
&h_n = \text{power consumption of switching state at node\#}n
\end{aligned} \tag{2}
$$

After the reconfiguration, the number of sessions can be calculated as below.

$$
M_n = \sum_{s \in N} \alpha_{s,n} M_s^{pre} \tag{3}
$$

In addition, several constraints can be considered for the proposed reference model as below. Here, we defined thresholds:

 ISBN: 978-1-61208-279-0

$$U \equiv \text{maximum utilization rate of FE}$$
$$W \equiv \text{maximum authorized time for reconfiguration} \tag{4}$$

C. Formulation for Linear Programming

To obtain the most suitable reconfiguration by using linear programming, in this section we will formulate the reference model of the previous section. Here, the objective function to be minimized by linear programming is the power consumption required for running the IMS and executing the reconfiguration.

The power consumption of the reference mode falls into three categories. The first is running power, which is consumed by running nodes and processing sessions. It depends on their running time, thus it is given by:

$$PCR = R\sum_{n\in N} y_n \left[B_n + \varepsilon \times M_n + \tau\right] \tag{5}$$

Secondly, the power is consumed by switching node states such as booting and shutdown. The total power consumed for them is given by:

$$PCS = \sum_{n\in N} h_n \theta_n \tag{6}$$

Thirdly, the migration of sessions from one FE to another consumes power in signaling overhead. The total power consumed for migrating sessions between nodes is given by:

$$PCM = \sum_{s\in N}\sum_{t\in N} C_{s,t}\alpha_{s,t} M_s^{pre} \tag{7}$$

Then, the objective function to be minimized can be expressed as the sum of the above:

$$OF = PCR + PCS + PCM \tag{8}$$

In addition to the objective function, there are several constraints for the linear programming to take into account. The number of sessions cannot be larger than the FE capacity and its utilization to provide stable IMS capacity. The constraint is given by:

$$M_n \le K_n U \tag{9}$$

The session transfer time is assumed to be lower than a given threshold in order to satisfy quality-of-service requirements because a large reconfiguration has an impact on stability of the IMS service. The constraint to suppress excessively large reconfigurations is given by:

$$\sum_{s\in N}\sum_{t\in N} T_{s,t}\alpha_{s,t} M_s^{pre} \le W \tag{10}$$

Finally, the ratios of sessions to be migrated and to be kept on the same node should sum to one. This constraint is given by:

$$\sum_{s\in N}\sum_{t\in N}\alpha_{s,t} = 1 \tag{11}$$

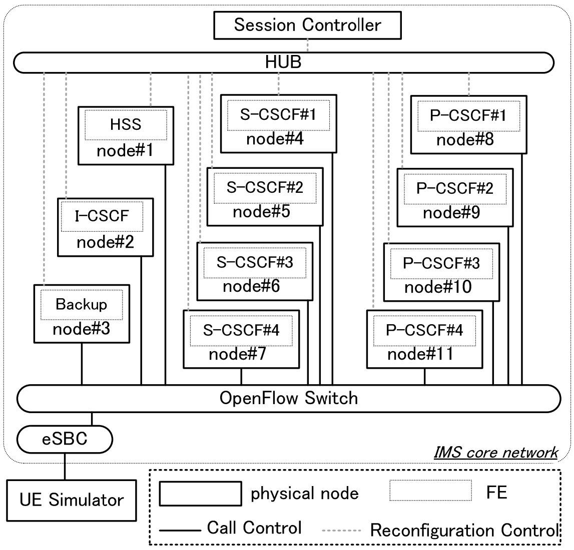

Figure 4. Experimental Network Configuration

D. Cycle of IMS Reconfiguration

In order to reduce the power consumption of the reference model in a certain period of time, the cycle of the reconfiguration is also important because the IMS can execute the reconfiguration as many times as needed in the period. The objective function (8) expresses one reconfiguration from state Z^{pre} to state Z and the current state Z during R. Thus, the total objective function during a certain period which includes state $Z \in Z_{total}$ is as described below.

$$OF_{total} = \sum_{Z\in Z_{total}} OF \tag{12}$$

Z_{total} is a set of state Z during the period. However, it is difficult to solve OF_{total} all at once. Thus, the system evaluates the objective function (8) periodically and executes a reconfiguration based on the result step by step. If the following equation (13) is satisfied, the reconfiguration should be executed. The IMS also executes the reconfiguration if constraint (9) is not satisfied in the current state due to changing of M_n.

$$OF \le R\sum_{n\in N} y_n^{pre}\left[B_n + \varepsilon \times M_n^{pre} + \tau\right] \tag{13}$$

This means that if the power consumption of the reconfiguration (6) + (7) is smaller than the expected reduction going from the previous state Z^{pre} to the next state Z during R, the reconfiguration should be done. However, note that this depends on value of the expected R. Thus, to improve the accuracy of the model, an estimation of R based on the call rate in an actual environment is desirable.

IV. IMPLEMENTATION AND EVALUATION

A. Implementation and Experiment Configuration

To solve our objective function describing the reference model, we used CPLEX [5], which can solve large linear problems. In addition, we implemented the IMS on real

TABLE I. SPECIFICATIONS OF EXPERIMENTAL NETWORK COMPONENTS

Node	Specifications			
	Hardware	**CPU**	**Memory**	**OS**
P-CSCF, I-CSCF, S-CSCF, eSBC, UE Simulator	Generic PC	Core2Duo E8600 3.33GHz	4 GByte	Ubuntu 10.04 LTS
HSS	Generic PC	Xeon E5450 3GHz	12 GByte	(cont.)
Session Controller	Generic PC	Core2Duo E8400 3GHz	8 GByte	(cont.)
OpenFlow Switch	NEC UNIVERGE PF5240 (OpenFlow Ver.1.0.0)			

machines to obtain the parameters required for the reference model and to verify our proposal.

Fig. 4 shows our experimental network configuration. In the IMS network, CSCFs and an eSBC are connected via an OpenFlow switch. A UE Simulator connects to the IMS through the eSBC. A session controller connects to the CSCFs, OpenFlow switch, and eSBC via a HUB. An HSS also connects to the CSCFs via the OpenFlow switch. Table. I shows the specifications of the components. The P/I/S-CSCFs and HSS were built based on the Open IMS core [16], which is an open-source SIP server. We implemented the required modules of the CSCFs for our proposed mechanism and the required software for the session controller and eSBC. We simulated a large UE load by using ims_bench [17], which is load testing software for the IMS (also open-source). The controller periodically collects the number of sessions from each CSCF. Then on the basis of its information, the controller calculates the objective function and executes the reconfiguration if the result satisfies the criterion in (13).

B. *Measurements*

We executed the basic calls and their session migration to obtain the parameters required for evaluating the objective function. UEs registered with four S-CSCFs through four P-CSCFs, and then the UEs continued to make calls to keep certain sessions at each P-CSCF. Here the duration of each call was 120 seconds. Then, we performed call session migration between P-CSCFs and turned the nodes on/off. The result is shown in Table II. The values are averages for four P-CSCFs. Note that here we adopted a P-CSCF as the target FE and measured the parameters. Of course, we can measure S-CSCF in the same way.

On basis of the obtained parameters, we wrote scripts for CPLEX for the objective function and the constraint conditions. Then we verified the behavior of the dynamic determination of the reconfiguration and measured the power consumption of each P-CSCF while performing calling, call session migration, and turning nodes on/off. First, UEs made calls through four P-CSCFs randomly, and then increased the number of calls to exceed the upper limit of the P-CSCF. We observed the behavior of the reconfiguration caused by the constraint condition. Secondly, the UEs reduced the number of sessions enough to turn off one node and observed the behavior of the reconfiguration. Then, the UEs increased the number of sessions from the reduced number enough to turn on an additional node, and we observed the behavior in the same way. Finally, the UEs cycled through a decrease and increase, and we evaluated the reduction in power consumption compared with the existing static IMS configuration, which was simply traditional operation where no reconfiguration was executed and the four nodes were kept running.

TABLE II. PARAMETERS OF P-CSCFS

params	values	params	values
B_n	48.2 W	$C_{s,t}$	3.10×10^{-3} J*
τ	0.137 W	$T_{s,t}$	1.86×10^{-4} s
ε	1.06×10^{-3} W	h_n	636 J*
K_n	18,000 sessions	W	2 s
U	0.8	R	100 s

* J: W·s (watt second)

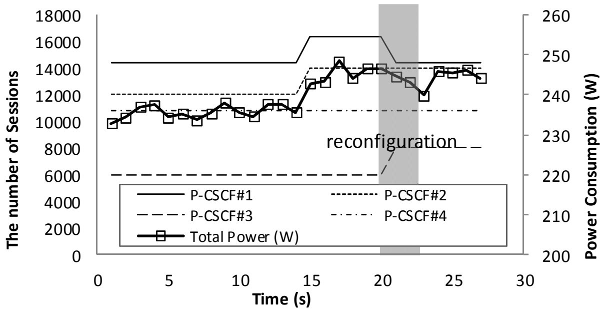

Figure 5. Load Balancing Session

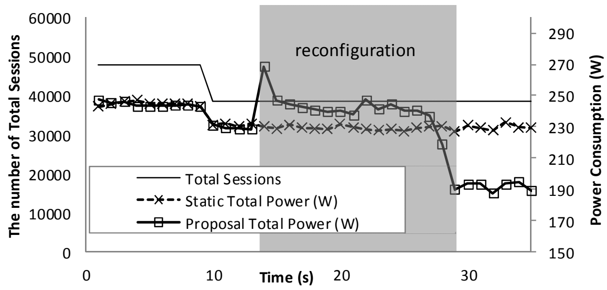

Figure 6. Turning off a node

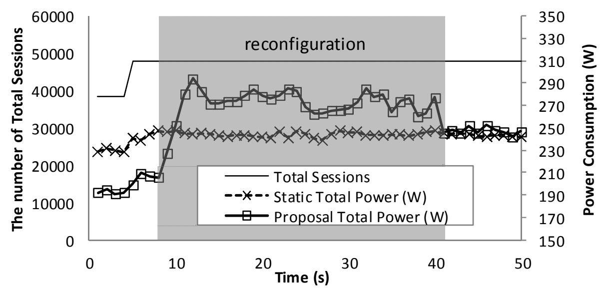

Figure 7. Turning on a node

 ISBN: 978-1-61208-279-0

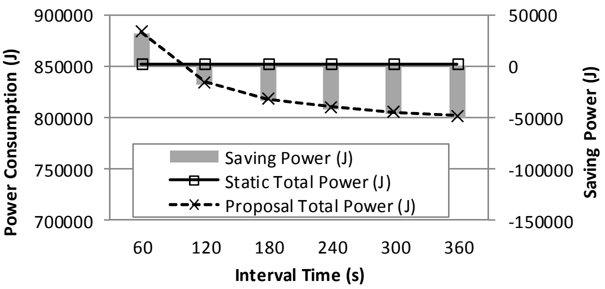

Figure 8. Comparison during reconfiguration cycle

C. *Results and Discussion*

Fig. 5 shows the load balancing behavior of the reconfiguration when the number of sessions exceeds the upper limit of the P-CSCF, which is 14,400 sessions (18,000 $\times$ 0.8). The reconfiguration began to be executed with periodical timing of calculating criterion in (13). The excess sessions on P-CSCF#1 were migrated to P-CSCF#3. Fig. 6 shows the behavior of the reconfiguration when the number of sessions was reduced. "Static Total Power" in the figure is the power consumption in the static IMS configuration, as in traditional operation. In our proposal, all sessions of a P-CSCF on a node were migrated to other P-CSCFs and the node was turned off. On a temporary basis, its power consumption increased, but it became smaller than the static power consumption after the reconfiguration. Fig. 7 shows the behavior of the reconfiguration when the number of sessions increased. In the beginning, the power consumption in our proposal was small because the number of running nodes was small. After that, it increased temporarily to equal the static power consumption for turning on a node and processing more sessions. Fig. 8 shows the power saving with our proposal compared with the static power consumption while running for one hour. During the period, we repeated the behaviors of Fig. 6 and Fig. 7. The interval time is the time for one cycle of reducing and increasing sessions (R = (interval time)/2). When the cycle of reconfiguration was long enough, although that depends on the situation, the power consumption was reduced by 5.9%. The amount of reduction becomes larger depending on the situation such as the changing number of nodes and sessions.

Thus we verified the value of dynamic determination and reconfiguration to reduce power consumption based on our proposed model and formula for linear programming in a real environment. To use linear programming to obtain an optimized solution, we approximated the non-linear values as accurately as possible with linear values calculated from their averages. This does not have a large impact on our proposal, and could be made more precise by formulating the non-linear values in a more rigorous manner. Also we can apply our proposal to larger configurations because the time to solve the linear problem was about 5 seconds, and larger problems can be solved within practical time constraints.

V. CONCLUSION

We have focused on the dynamic determination of IMS reconfiguration in a flexible IMS environment that can change its configuration by live session migration and turning nodes on/off. In order to realize this, we proposed a model of the reconfiguration and formulation to be solved by linear programming with the goal of saving power. Furthermore, we implemented a dynamic reconfigurable IMS that used our proposal, and verified that the reconfiguration reduced power consumption.

ACKNOWLEDGMENT

This work has been supported by the Japanese Ministry of Internal Affairs and Communications funded project.

REFERENCES

[1] 3GPP TS 24.229 V9.13.0, "IP multimedia call control protocol based on Session Initiation Protocol (SIP) and Session Description Protocol (SDP); Stage 3," 2012.

[2] Global System for Mobile communications Association: www.gsma.com

[3] GSMA, "Rich Communication Suite 5.1 Advanced Communications Services and Clients specification," 2012.

[4] M. Ito, S. Komorita, Y. Kitatsuji, and H. Yokota, "OpenFlow-based Routing Mechanism for Call Session State Migration in the IMS," Proc. 7th WSEAS International Conference on CEA, 2013.

[5] ILOG CPLEX 11.0 User's Manual, 2007.

[6] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, et al., "SIP: Session Initiation Protocol, Internet Engineering Task Force," RFC 3261, 2002.

[7] 3GPP2 X.S0011-001-C v3.0, "cdma2000 Wireless IP Network Standard: Introduction," 2006.

[8] 3GPP TS36.300, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2," 2009.

[9] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, et al., "OpenFlow: Enabling Innovation in Campus Networks," *ACM SIGCOMM Computer Communication Review*, vol.38, No.2, 2008, pp. 69-74.

[10] H. Jiang, A. Iyengar, E. Nahum, W. Segmuller, A. Tantawi, and C. Wright, "Load balancing for SIP server clusters," IEEE INFOCOM, 2009, pp. 2286-2294.

[11] V. Hilt and I. Widjaja, "Controlling overload in networks of SIP servers," IEEE International Conference on Network Protocols (ICNP), 2008, pp. 83-93.

[12] H. Fathi, S. Chakraborty, and R. Prasad, "Optimization of SIP session setup delay for VoIP in 3G wireless networks," IEEE Transactions on Mobile Computing, Vol. 5, Number. 9, 2006, pp.1121-1132.

[13] N. Rajagopal and M. Devetsikiotis, "Modeling and Optimization for the Design of IMS Networks," IEEE ComSoc Proceedings of the 39th annual Symposium on Simulation, 2006, pp. 34-41.

[14] M. Fakhfakh, O. Cherkaoui, I. Bedhiaf, and M. Frikha, "High availability in IMS virtualized network," First International Conference on Communications and Networking, 2009, pp. 1-6.

[15] V. Petrucci, O. Loques, and D. Mossé, "Dynamic optimization of power and performance for virtualized server clusters," Proceedings of the 2010 ACM Symposium on Applied Computing, 2010, pp. 263-264.

[16] The Open Source IMS Core Project, Available at: http://www.openimscore.org/

[17] IMS Bench SIPp, Available at: http://sipp.sourceforge.net/ims_bench/