Computer Networks xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

Energy-aware joint management of networks and Cloud infrastructures

Bernardetta Addis [a], Danilo Ardagna [b], Antonio Capone [b,*], Giuliana Carello [b]

[a] LORIA (UMR 7503 CNRS), Université de Lorraine, INRIA Nancy – Grand Est, France
[b] Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Italy

ARTICLE INFO

Article history:
Received 24 July 2013
Received in revised form 8 April 2014
Accepted 20 April 2014
Available online xxxx

Keywords:
Green IT
Cloud service
Energy management
Green networking
Optimization
Mathematical programming
Renewable energy

ABSTRACT

Fueled by the massive adoption of Cloud services, the CO_2 emissions of the Information and Communication Technology (ICT) systems are rapidly increasing. Overall service centers and networks account for 2–4% of global CO_2 emissions and it is expected they can reach up to 10% in 5–10 years. Service centers and communication networks have been managed independently so far, but the new generation of Cloud systems can be based on a strict integration of service centers and networking infrastructures. Moreover, geographically-distributed service centers are being widely adopted to keep Cloud services close to end users and to guarantee high performance. The geographical distribution of the computing facilities offers many opportunities for optimizing energy consumption and costs by means of a clever distribution of the computational workload exploiting different availability of renewable energy sources, but also different time zones and hourly energy pricing. Energy and cost savings can be pursued by dynamically allocating computing resources to applications at a global level, while communication networks allow to assign flexibly load requests and to move data.

Even if in the last years a quite large research effort has been devoted to the energy efficiency of service centers and communication networks, limited work has been done for exploring the opportunities of integrated approaches able to exploit possible synergies between geographically distributed service centers and networks for accessing and interconnecting them. In this paper we propose an optimization framework able to jointly manage the use of brown and green energy in an integrated system and to guarantee quality requirements. We propose an efficient and accurate problem formulation that can be solved for real-size instances in few minutes to optimality. Numerical results, on a set of randomly generated instances and a case study representative of a large Cloud provider, show that the availability of green energy have a big impact on optimal energy management policies and that the contribution of the network is far from being negligible.

1. Introduction

The debate on climate change and carbon dioxide emission reduction is fostering the development of new "green policies" aimed at decreasing the environmental impact of human activities. How to contrast global warming and how to enhance energy efficiency have been put on top of the list of the world global challenges. ICT plays a key role in this greening process. Since the beginning, ICT applications have been considered as part of the solution as they can greatly improve the environmental performance of all the

* Corresponding author. Tel.: +39 02 2399 3449; fax: +39 02 2399 3413.
E-mail address: capone@elet.polimi.it (A. Capone).

http://dx.doi.org/10.1016/j.comnet.2014.04.011
1389-1286/

other sectors of the world economy. More recently, the awareness of the potential impact of the carbon emissions of the ICT sector itself has rapidly increased.

Overall, the combination of the energy consumption of service centers and communication networks accounts for 2–4% of global CO_2 emissions (comparable, e.g., to the emissions due to the global air traffic) and it is projected to reach up to 10% in 5–10 years, fueled by the expected massive adoption of Cloud computing [1,2]. Service centers investment grew by 22.1% during 2012 and it is expected it will further grow by another 14.5% in 2013 [1]. So, one of the main challenges for Cloud computing is to be able to reduce its carbon and energy footprints, while keeping up with the high growth rate of storage, server and communication infrastructures.

Even if computing and networking components of the system have been designed and managed quite independently so far, the current trend is to have them more strongly integrated for improving performance and efficiency of Cloud services offered to end users [3]. The integration of computing and networking components into a new generation of Cloud systems can be used not only to provide service flexibility to end users, but also to manage in a flexible way resources available in geographically-distributed computing centers and in the network interconnecting them.

A key enabler of Cloud/network cooperation is the use of geographically distributed service centers. Distributed Cloud service provisioning allows to better balance the traffic/computing workload and to bring computing and storage services closer to the end users, for a better application experience [4]. Moreover, from an energy point of view, having geographically distributed service centers allows Cloud providers to optimize energy consumption by exploiting load variations and energy cost variations over time in different locations.

The level of flexibility in the use of resources in different service centers depends on the application domains and basically comes from the geographic distribution of virtual machines hosting service applications, the use of dynamic geographically load redirect mechanisms, and the intelligent use of storage systems with data partitioning and replication techniques. This flexibility in service centers management also has a relevant impact on the communication network mainly for two reasons. First, since the service requests from users are delivered through an "access network" to service centers hosting the virtual machines, moving dynamically the workload among service centers can completely change the traffic pattern observed by the network. Second, an "interconnection network" is used to internally connect service centers and redirect end-user request or even move virtual machines and data with, again, a non-negligible impact on the traffic load when reconfiguration decisions are taken by the Cloud management system. Access and interconnection networks can be actually implemented in several different ways as private Intranets or using the public Internet, depending on the Cloud provider policies and the specific application domains. In any case, there is a large number of possible interaction models among telecommunications and Cloud providers that regulate their service agreements from a technical and economical perspective.

The main contribution of this paper is to explore the possibility of jointly and optimally managing service centers and the network connecting them, with the aim of reducing the Cloud energy cost and consumption. We develop an optimization framework that is able to show the potential savings that can be achieved with the joint management and to point out the relevant parameters that impact on the overall system performance. Concerning the network, we propose two representations, a high level approximated model and a detailed one. We show that the energy consumption obtained with the approximated version is very close to the one obtained with the detailed network representation and that the approximated version requires significantly smaller computational effort. Thus, it is suitable to assess the importance and impact of the joint optimization. The resource utilization and load allocation scheduling is performed on a daily basis assuming a central decision point (one of the service centers) and the availability of traffic patterns for different time periods. However, since the computational time is quite small (order of a few minutes), the time period length can be decreased, and the granularity can be finer, so as to follow better unpredictable traffic variation.

The approach proposed is based on a Mixed Integer Linear Programming (MILP) model which is solved to optimality with a state-of-the-art solver. The model assumes a Cloud service provider adopting the PaaS (Platform as a Service) approach and optimizes the load allocation to a set of geographically distributed Service Centers (SCs) where virtual machines (VMs) are assigned to physical servers in order to serve requests belonging to different classes. The goal is minimize the total energy cost considering the time-varying nature of energy costs and the availability of green energy at different locations. The traffic can be routed to SCs using a geographical network whose capacity constraints and energy consumption are accounted for. We formally prove that the problem is NP-hard since it is equivalent to a Set Covering Problem.

We present a set of numerical results on a realistic case study that considers the real worldwide geographical distribution of SCs of a Cloud provider and the variations of energy cost and green energy availability in different world regions. Moreover, we show some results to characterize the scalability of the optimization framework proposed and the sensitivity to workload and green energy prediction errors.

Even if a large literature on the improvement of energy efficiency of service centers (see e.g. [5]) and communication networks (see e.g. [6]) exists, very few studies have considered so far the cooperation between Cloud and network providers for a joint energy management, and, to the best of our knowledge, no one has proposed a joint optimization framework able to exploit low energy modes in physical servers and networking devices (see Section 5 for a detailed analysis of related work). From a practical point of view, we assume that the economic advantages coming from the energy savings can be managed through service agreements between Cloud and network operators