Green Cloud: Towards a Framework for Dynamic Self-Optimization of Power and Dependability Requirements in Cloud Architectures

Rami Bahsoon
School of Computer Science, The University Of Birmingham
Edgbaston, B15 2TT, Birmingham, UK
r.bahsoon@cs.bham.ac.uk

Abstract. *I report on the activities of the ongoing EPSRC/University of Birmingham Bridging the Gap Fellowship on Green Cloud. The initiative is multidisciplinary; it involves the School of Computer Science, The School of Electrical, Electronic and Computer Engineering, Institute for Energy Research and Policy, the Department of Economics at Birmingham University, external collaborators at MIT, and industrial partners. The initiative is aimed at a framework for dynamic self-optimization of the cloud architecture taking into account the tradeoffs involved in maintaining acceptable dependability requirements with minimal power at runtime. The research pioneers research in dynamic optimization explicating green-aware software architectures and engineering, self-management of the cloud in relation power and dependability. This initiative feeds into the long-term and green-aware vision of helping in reducing power consumption and CO2 emissions in ICT infrastructures.*

1. Introduction

Cloud computing is claimed for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort. The popularity of the cloud is rapidly increasing: many computing services have now moved to the cloud and many new applications are emerging either as standalone or in orchestration with other services. As a result, the cloud architecture is *dynamically* scaling up to accommodate such growth. Such growth necessitates dynamic approaches for maintaining and monitoring an acceptable level of Quality of Service (QoS). From the architecture point of view, scaling up such systems does certainly introduce additional computational power, as meeting *QoS requirements* such as *scalability, availability, reliability, real time performance, fault-tolerance, openness and security* could imply the need for additional computational resources: this may, for example, entail hardware and software redundancy to be deployed in realization to some architectural mechanisms and tactics, like load balancing, replication, migration transparency, and so forth. The situation will lead to an uncontrolled growth of computational power. Such growth, if left unmanaged, is expected to contribute to the degradation of our ecosystem and more CO_2 emissions as we move and heavily depend on the cloud. For example, it would be expecting that more servers/nodes to be deployed and consequently more cooling power would be required. Meanwhile, meeting QoS and dependability requirements is critical and can't be neglected in favor of power savings.

In this position statement, we report on the activities of the ongoing EPSRC/University of Birmingham Bridging the Gap Fellowship on Green Cloud. The initiatives is multidisciplinary and aimed at a framework for *dynamic self-optimization* of the cloud architecture taking into account the tradeoffs involved in maintaining acceptable dependability requirements with minimal power at runtime.

2. Activities Leading to Green Cloud

In this section, we report on the activities for realizing the framework.

2.1 ***Power data analyzer in the cloud in relation to QoS.*** With colleagues in Electrical and Electronic Engineering, we are investigating mechanisms for *measuring, logging, control, accuracy and calibration of the power as the cloud architecture dynamically evolves in response to QoS demands and provision*. The investigations leverages on the state-of-art and state-of-practice in Dynamic Power Management to benefit the case of the cloud. For example, realizing scalability requirements, more replica nodes are expected to be deployed at runtime and more computing resources will be used to address scalability transparently. Meanwhile, other emerging dependability requirements (e.g. performance and throughput) may need to be rectified as we scale up; henceforth, consuming additional power. The fundamental premise is that the cloud (and its components) experiences no uniform workloads exhibiting variation in QoS requirements during its operation. Such an assumption is valid for most systems, both when considered in isolation and when internetworked as for the case of the cloud. A second assumption is that it is possible to monitor, with a certain degree of confidence, the fluctuations of QoS and their power. Such observation and prediction should not consume significant energy, however.

2.2 Modeling power and Quality of Service (QoS) demands and new meters for QoS per cloud power value. Dynamic changes in requirements such as scalability, availability, security, and performance requirements may suggest additional hardware/software resources, which need to be deployed at runtime. This may effectively translate to additional computing power. As result, power is best modeled with respect to QoS demands of a runtime instance. With colleagues at *the School of Electrical, Electronic and Computer Engineering, Institute for Energy Research and Policy, and the Department of Economics, we are investigating ways for formulating models for expressing QoS demands and provision-value per power usage. Vice versa, we are proposing new meters for QoS-per-power value. Such meters will show the value of QoS on the expense of power sacrificed (and vice versa).* We expect to further refine these meters to include *individual QoS quality-per-power value* (e.g. performance-per-power value, availability-per-power value etc.) and how they can be expressed in isolation or when combined. Such meters will be useful for performing what-if analyses when trying to match resource provision to power demands (see 2.3), for facilitating sensitivity and tradeoffs analyses-either statically or dynamically (at runtime).

2.3 Economics-inspired approach for self-optimizing the cloud architecture. We argue that dynamic change and evolution of the cloud architectures is a value-seeking and value-maximizing process, where the architecture is undergoing a dynamic change (at runtime) and seeking value. We treat QoS provision and their power consumption as scare resources, which need to be dynamically optimized. Performing a runtime search for best architectural instances, which address the QoS requirements with minimal power is a problem, which is appealing to "dynamic" or "on line" Search-based Software Engineering (SBSE), where the optimization problem is rapidly changing and the current best solution must be continually adapted. The goal is to maximize the value added and select the optimal execution plans.

The highly dynamic nature of the cloud architecture requires a simple, scalable, and inexpensive runtime optimization technique. This is necessary as the search for the optimal runtime instance addressing the tradeoffs between power and QoS is continuously active and may be initiated at various time intervals of the execution. This is not only for seeking to find an optimal solution to the said problem, but rather, for seeking to improve upon the current runtime situation. Classical and static optimization techniques may be ineffective and expensive to use in this dynamic setting for the representation of the problem and the definition of the fitness function are mere active at runtime and very volatile with respect to time. Furthermore, the problem entails judging the tradeoffs not only from a technical perspective but also from an economics driven one. We are investigating how economics-inspired approaches, based on market-control theory and/or game theory, can address the problem of dynamic runtime optimization and self-adaptation of the cloud architecture in addressing such tradeoffs. Together with colleagues at the Department of Economics, we will formulate models and scenarios, utilizing the meters developed in 2.2. We are investigating how classical market based theory and it various simple concepts and scenarios like supply, demand, inflation, recession, and equilibrium are mapped to the case of matching supply and demand of power vs. that of QoS(individual or combined). We have reported on some preliminary results on the analogy for the case of standalone software architectures [2]. The relevance to the cloud case looks to be promising.

2.4 Implementation framework using Dynamic Data Simulation Systems (DDAS). As the dynamic optimization is merely highly active at runtime and of high runtime refresh rate (due to the dynamicity/scalability of the cloud topology). We are investigating how Dynamic Data Driven Simulation Systems (DDDAS) paradigm, which the University of Birmingham is a key contributor to the field, can benefit the cloud. In particular, we are investigating how elements of DDDAS – *including measurement, simulation, control, and feedback* can extend the cloud architectural style to assist in the problem of dynamic runtime self-optimization of the cloud in relation to power and QoS. This layer is expected to utilize the analyzer of 2.1 and the economics-inspired models developed in 2.2 & 2.3. One important aspect of DDDAS in this setting is the ability to use *runtime simulation* (through the use of high performance computing) in order to improve the prediction and the self-management process of the runtime configuration through symbiotic control and feedback loops. To achieve this objective, the simulation will require mining data stored from previous runtime instances to inform and tune the prediction and the execution plans to self-adaptation. In particular, the simulation

- will identify scenarios and possible moves leading to robust and stable runtime configurations in the cloud topology showing efficiency of power use and acceptable QoS. Simulation can also predict situations leading to QoS sacrifices in relation to power savings.
- can predict the rippling impact of such sacrifices on the robustness of individual nodes/services and the cloud as whole.
- shall provide the basis for dynamic analysis for QoS value sacrifices with respect to power savings. Simulation can also determine the long-term implications of favoring QoS requirements and policies over power savings for some instances, where QoS can't be compromised. These implications can be in relation to cost and stability of the cloud.
- can also have static use: For example, the simulation can inform the cloud management decisions and long-term policies related to cloud service provision, management, deployment, and capacity restriction or leasing in relation to various cloud services – ranging from software-, platform-; infrastructure-, data- services.

The DDDAS layer is expected to form a standalone and independent layer, which can be integrated in the cloud to make it green-aware and energy efficient.

3. Expected Impact

The research is expected to initiate a multidisciplinary dialogue, foster networking, and research collaboration among some researchers in dynamic power management, green-aware computing, system dependability, simulation, and economics research. This initiative is a stepping point towards multidisciplinary proposal(s), targeting timely research programmes such Energy Futures and Digital Economy. If successful, research will raise the understanding of evolution trends in dynamic systems and improve their quality and robustness through dependability and power measurement and control. More widely, we hope the research results will feed into long-term vision of helping in reducing power consumption and CO2 emissions in ICT infrastructures, which is in line with the University of Birmingham research strategy on Energy. Nationally, funding bodies such as EPSRC has recently announced a call for the submission of ideas for Energy Networks Grand Challenges, which such programme directly addresses. Internationally, the U.S. President's Council of Advisors on Science and Technology (PCAST) released a major report on U.S. government investments in Networking and Information Technology Research and Development (NITRD), calling for major increases in both the ambitiousness of and funding for research, with particular emphasis on software and linkages between computing systems and the physical environment, which is inline with our proposed research. This work programme will set the seed for an interest specialist group and an institute for research explicating power and cloud at Birmingham. We will seek to widen interest in the output within the software architecture practitioners and middleware communities. We intend to organise workshops, to engage into fruitful discussions that may lead to future multidisciplinary collaborations with researchers on system power, distributed software architectures, dependability and environmentalists and conduct research visit to our industrial collabo-

rators, like Vodafone, Google, and/or middleware players e.g., MS and IBM.

4. References

[1] Nallur, V. R. Bahsoon, and X. Yao (2009). Self-Optimizing Architecture for Ensuring Quality Attributes in the Cloud. In Proceedings of the Joint Working IEEE/IFIP Conference on Software Architecture 2009 & European Conference on Software Architecture 2009, Cambridge, UK

[2] G. Rangaraj and R. Bahsoon(2010). A Market-based Approach for Self-Managing Power in Software Architectures – in submission. Technical Report, School of Computer Science, University of Birmingham, CSR-10-01.

[3] V. Nallur and R. Bahsoon(2010). Design of a Market-Based Mechanism for Quality Attributes Tradeoffs of Services in the Cloud, To appear, in the Proceedings of the 25th ACM Symposium of Applied Computing (ACM SAC 2010), 2010.

[4] EDSER 1-8: Proceedings of the Workshops on Economics-Driven Software Engineering Research: In conj. with the 21st through 28th International Conference on Software Engineering (1999 - 2006)