I. Research Overview and Outcome

Problem Statement:
- Cloud computing presents the illusion of infinite capacity of computational resources. In the case of Infrastructure as a Service (IaaS) clouds, these resources are typically offered in bundles with specific amounts of CPU, Memory, and Network.
- Solution developers are thus presented with the problem of ensuring the Quality of Service (QoS) non-functional requirements of an application by mapping the application’s components to one or more of these bundles, and to change this mapping if the workload changes.
- That is: Given an app model, which includes its architecture, QoS Constraints, and its workload, compute which IaaS resources are needed to fulfill the QoS requirements.

Proposed Solution:
- The key idea of our solution is a Mapping-as-a-Service approach in which users provide a descriptive model of their application and get back mappings in various IaaS providers. These mappings emphasize the assurance of the Quality of Service metrics from the application’s model. An initial mapping is done based on heuristics, and then we monitor application performance to provide scaling suggestions via a callback interface. Underneath the service’s API, the implementation accepts different resource usage prediction methods and allows allocation on different IaaS providers.
- The main technical challenges of this work were the experimentation with different application model representations, and the implementation of a resource usage prediction engine. As a cloud application model, we propose the use our Distributed Ensemble of Virtual Appliances (DEVA) model [1]. To be able to experiment with different matching approaches, we use a rule-based engine similar to the one proposed in [2].

Remarks and Future Work:
- We present the design of an autonomic solution for cloud application resource mapping and scaling based on monitoring of QoS constraints.
- We provide details on a prototype implementation and how we deal with the technical challenges.
- We assess the validity of the idea by presenting experiments on functionality and scalability.
- For future work, we intend to enlarge the experiment size and to extend the solution so that it can do horizontal scaling as well as vertical scaling.

Mapping model:
- For the resource usage prediction, we develop a rule-based system that does a match between application QoS constraints to what we call intermediate-form resources. These intermediate-form resources are estimated using an approach comprising multivariate linear regression and heuristics.
- The heuristic takes the estimated parameters from the regression and tries to fit known IaaS resource bundles that would ensure the QoS.
- We then recommend the tuples with the best fit for various IaaS providers and start monitoring for workload changes.

References:

II. International Experience

Barcelona Supercomputing Center
- Had the opportunity to work with and get feedback from successful HPC and cloud computing researchers.
- This experience adds value for when I’m in the job hunting process.
- Obtained a good understanding of the research topics of interest in BSC.
- This experience opens the door for further collaborations.

Barcelona – the Capital of Catalonia
- Barcelona is a vibrant and cosmopolitan city with enough sightseeing to have you busy for months.
- From the very busy and touristic Las Ramblas street, to the Ciutadella Park, there is a public space for anyone.
- Had the opportunity to immerse myself in a different culture, to live in Barcelona, and to visit Madrid.

III. Acknowledgement

The material presented in this poster is based upon the work supported by the National Science Foundation under Grant No. OISE-0730065. This work was also supported by the US Department of Education under P200A090061. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Department of Education.