Name: Prashanth Mallyampatti
Unity ID: pmallya
Student ID: 200250501

CSC724- Advanced Distributed Systems
Paper Review

Agile: elastic distributed resource scaling for Infrastructure-as-a-Service

Summary:

The paper addresses the problems of keeping up the load changes and the ability to the know the future demands of IaaS clouds. The paper proposes AGILE- a tool which addresses the above problem using medium-term resource demand prediction and dynamic VM cloning to reduce the initial application load time. The matter of deciding when and how much to allocate the resources without any SLO violations is a common question is dynamically changing workloads. AGILE uses wavelet transforms and Markov model to predict the future demands and discusses VM cloning to apply these predictions dynamically.

AGILE is an elastic distributed resource scaling system for IaaS cloud infrastructures using online profiling and polynomial curve fitting. Coming to the AGILE system design, it has three main components. The first being, Medium-Term Resource demand prediction using Wavelets. Here the resources are predicted using sliding window D of recent resource usage data. The data which is sampled wavelet transforms are applied predicting the resource demand over a window W. This window W is determined by the VM cloning time(typically 2 minutes), which means that if there's a requirement of resources in future, a VM has to be cloned which takes 2 minutes, hence a prediction has to be made for a window size of W= 120 seconds. The original resource demand is converted to wavelet signals, decomposed it into increasing scales and predict the wavelet values using Markov model. Second being, Online resource pressure modeling, where AGILE picks an appropriate resource allocation by using a application agnostic resource pressure model to map the application's SLO violation rate target. Here resource pressure is the ratio of resource usage to allocation. This model is generated dynamically at runtime with some pre-determined models to map (to avoid frequent model training). The last technique used by AGILE is Dynamic server pool scaling. When there's a prediction of resource demand increase AGILE uses Pre-copy live cloning to clone a VM. AGILE two features to minimize the cloning overheads.

AGILE has been implemented on top of KVM virtualization platform and evaluated using RUBiS Webserver & Database, Apache Cassandra key-value store and Google cluster data. How to predict the accuracies has been discussed and focuses on PRESS to compare. True positive and False positive rates have been discussed. The experimental results on various aspects such as prediction accuracy, overload handling, overheads, and dynamic copy-rate have been discussed. The results are promising in all aspects with very minimal overheads when compared to PRESS and other methods.

Strong Points:

1) AGILE doesn't require any prior knowledge of the application.
2) AGILE generates the resource pressure model dynamically (application agnostic), which is very much required in changing workload environments.
3) Features to minimize cloning time seems noteworthy.
4) The peak performance attained immediately after the start adds to a strong point of AGILE.
5) Overheads of AGILE are acceptable for the performance it gives.

Weak Points:

1) The length the prediction window W (which is set to 2 minutes) could have been made dynamic, as there's no guarantee that a VM would clone within 2 minutes.
2) AGILE could have been made more intelligent to dynamically see workload changes, as currently AGILE is periodically triggered.
3) The paper could have discussed briefly about Markov model stating its pros and cons.