ACE: Automated Capacity Evaluation for HP ePrint

Vipul Garg, Ludmila Cherkasova*, Swaminathan Packirisami, Jerome Rolia*
HP PPS R&D Bangalore, India and *HP Labs Palo Alto, CA, USA
E-mail: {firstname.lastname}@hp.com

The HP cloud ePrint Service [1] allows customers to print from anywhere to an HP ePrint-enabled printer accessible via the Internet. As ePrint is a hosted service, it must provide customers with a high quality of service while keeping the costs of supporting the load as low as possible. The jobs entering the ePrint Service are diverse and complex in nature. The jobs vary largely in terms of their formats, sizes, originator clients, etc. With rapidly increasing growth in client requests, it has become imperative to understand the new service's characteristics and the service dynamics over time in order to efficiently support and scale the ePrint System. With the complexity of offered cloud services increasing and application requirements for QoS growing, the research challenge is to design an integrated framework of measurements and system modeling to support performance analysis of these services.

We have developed a new tool, called ACE (Automated Capacity Evaluation), which enables a QoS-driven capacity planning for ePrint service (see [2] for more details). A user interaction with ACE is performed via a GUI written in Python: given a response time goal and a workload for processing, ACE estimates the amount of resources required by the ePrint System for job processing with given performance goals.

ACE consists of the following main components:

- *Workload Profiler:* The ePrint system collects a diverse variety of metrics to monitor service behavior: *i)* server metrics such as CPU utilization, memory usage, and network bandwidth; *ii)* event logs that record processed jobs and related activities over time. The Profiler determines the ePrint job counts and response times observed for each job type (e.g., format, size, etc.) for subsequent time intervals (e.g., 10 minute intervals) along with service-node CPU and network utilizations.

- *Regression-based Solver:* Using the amount of processed work (defined by different types of jobs and their counts) and the CPU utilization values from the *Profiler*, the *Solver* uses a step-wise linear regression to derive the CPU and network bandwidth costs (resource demands per job type). It determines this cost for a given hardware and VM's with different OS's, as specified by the architecture.

- *Key Performance Indicator (KPI) Evaluator:* Given a response time KPI, the Evaluator estimates *i)* maximum service-node CPU utilization (that should be used in the capacity planning process), and *ii)* a possible percentage of jobs violating a given KPI per time interval.

- *Predictor:* Given the ePrint QoS goals, the Reporter computes the number of service-nodes needed to satisfy a given system KPI. The Predictor can evaluate capacity for two different scenarios: *i)* predicting a required capacity for a scaled *Past Workload*. The *Predictor* computes a weekly trend for total job counts at the current time and for next $\boldsymbol{n}$ number of weeks, using regression models of the following form: $\boldsymbol{predicted\ counts = constant + n * a}$, where a is a regression coefficient and then predict the capacity for this workload keeping the ratio of different job types in the workload mix as the same which is observed in the production environment; b) predicting a required capacity for a *User-defined Workload* (a customized mix of job types). The user can change the percentages of different formats and their sizes, to basically change the ratio of different job types in the workload, as he desires and observe the capacity of the system. The *Predictor* takes into account the user-entered values for the formats and their sizes, along with the throughput expected to predict the resource demands.

- *Reporter:* It provides *i)* compact workload statistics for the ePrint service: percentage of different jobs types, percentage of jobs violating KPIs, etc., and *ii)* an in depth analysis of the production system behavior.

Figures 1 shows the architecture diagram of the ACE Tool illustrating the relationship between different components at a high level. Figure 2 shows a screenshot of the ACE Tool showing the Job Counts of the top 3 job types during a particular monitoring time (all the values in the chart have been normalized by a random value).

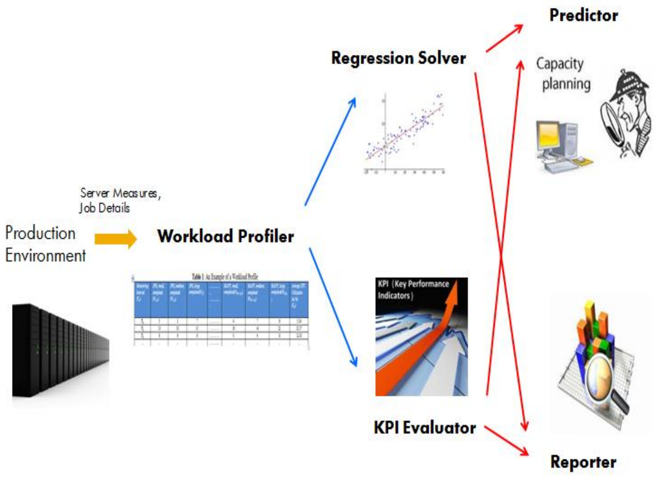
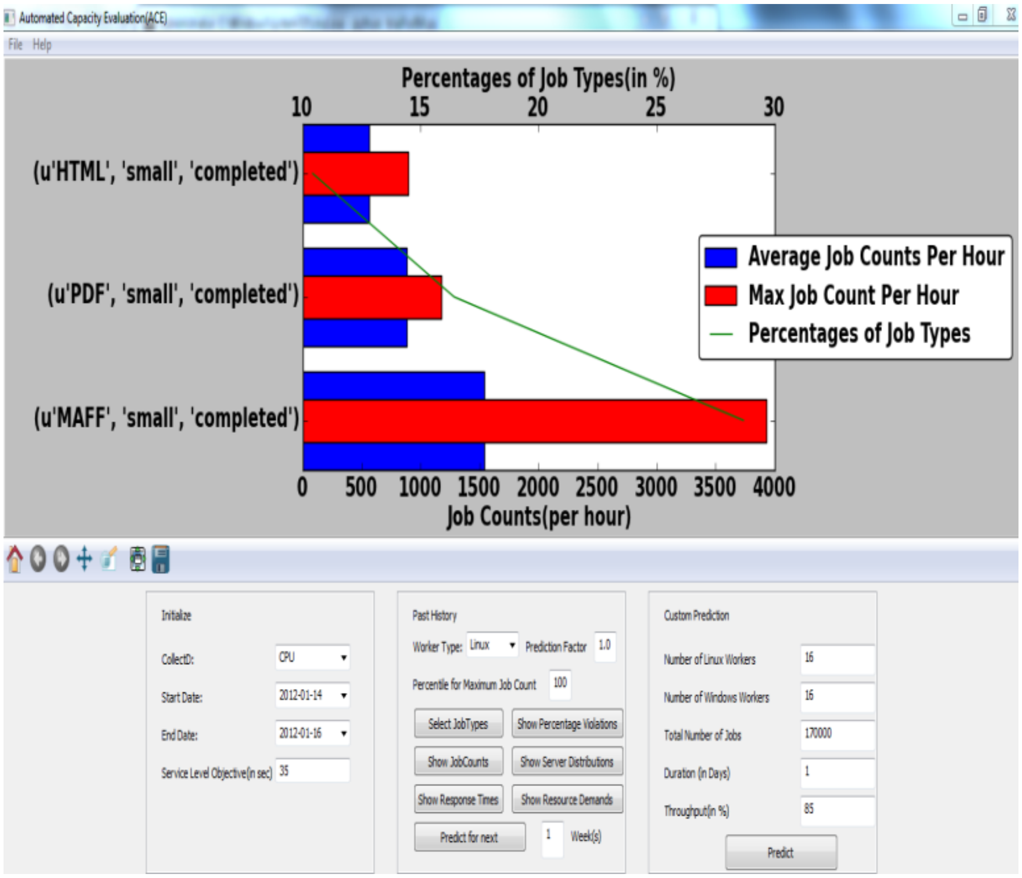


Figure 1. ACE Tool Architecture.


Automated Capacity Evaluation(ACE)
File
Help
Percentages of Job Types(in %)
10
15
20
25
30
(u'HTML', 'small', 'completed')
(u'PDF', 'small', 'completed')
(u'MAFF', 'small', 'completed')
Average Job Counts Per Hour
Max Job Count Per Hour
Percentages of Job Types
0
500
1000
1500
2000
2500
3000
3500
4000
Job Counts(per hour)
Select JobTypes
Show Percentage Violations
Show JobCounts
Show Server Distributions
Show Response Times
Show Resource Demands
Predict for next
Predict

Initalize

CollectD: CPU

Start Date: 2012-01-14

End Date: 2012-01-16

Service Level Objective(in sec) 35

Past History

Worker Type: Linux Prediction Factor 1.0

Percentile for Maximum Job Count 100

Predict for next 1 Week(s)

Custom Prediction

Number of Linux Workers 16

Number of Windows Workers 16

Total Number of Jobs 170000

Duration (in Days) 1

Throughput(in %) 85

Figure 2. ACE Tool Screenshot (Job Counts)

The Initialize panel specifies the resource used for reporting when showing resource demands, the date range of historical data to be used for input for the tool and a response time oriented service level objective. The historical data determines a rate of growth for the workload using the Predictor algorithm as described above. The Past History panel specifies, for each type of worker, a user entered Prediction Factor that scales the rate of growth upwards or downwards. This factor is based on expected business conditions. Predict for next states how many weeks into the future to predict. When selected another screen appears that describes the required configuration. The Custom Prediction panel allows a user to enter the number of workers and then reports on the number of jobs that can be supported while satisfying the service level objective. When the Predict button is used another panel appears that allows for additional configuration of the workload, e.g., selection of job types and/or changes from observed workload job mix.

The ACE tool offers a practical solution for the automated evaluation of required capacity that is needed for processing a diverse ePrint workload in production environment while satisfying QoS requirements.

Related work: Solutions provided by cloud computing environments dynamically adjust the number of VMs in response to a user provided function [3][4]. Functions are typically based on current throughput or on the number of users. We have developed and described a function for the HP's e-Print service that is novel because it supports the management of e-Print according to a KPI based on a percentile of response times – the KPI was already defined for the e-Print service. Our approach helps to predict the cost of the resources needed to support future workloads.

References

[1] HP ePrint, http://www.hp.com/ePrint

[2] V. Garg, L. Cherkasova, S. Packirisami, J. Rolia: Workload Analysis and Demand Prediction of HP ePrint Service. To appear in Proc. of IM'2013.

[3] Amazon Elastic Cloud: Auto Scaling, http://aws.amazon.com/ec2/#features

[4] Case Studies, http://www.rightscale.com/customers/case_studies.php