Workload Analysis and Demand Prediction for the HP ePrint Service

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

Abstract — **The proliferation of service oriented computing boosted by advances in cloud computing has led to design of new services that combine the power of both trends. The HP ePrint Service 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. Understanding the load and the nature of this new service is crucial for properly designing the service's support infrastructure under rapidly growing user demand. With the complexity of services increasing and application requirements for QoS growing, the research challenge is to design an integrated framework of workload analysis combined with system measurement and modeling techniques that support performance analysis for the service. In this work, we present a detailed workload analysis of ePrint and a new performance tool for the automatic evaluation of required capacity for processing its diverse workload in a production environment while satisfying QoS requirements.**

I. Introduction

Printing, traditionally, has been a tedious task for the end user. The printing lifecycle includes finding a computer, locating a printer the computer is able to access, possibly installing or updating appropriate drivers, transferring the document to the computer, and only then printing the document. Two years ago, Hewlett-Packard produced a new generation of printers that connect to and print content directly from the Internet. The ePrint service [1] provides each of these new printers with an ePrint email address. By emailing a document to a printer's ePrint email address, any authorized Internet connected device can cause a document to be printed, eliminating the hassle of finding a remote computer and printer.

Since its release, ePrint service has experienced substantial growth in the number of registered printers and the number of print jobs which are sent to these printers. The jobs entering the ePrint System 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, its workload properties, access patterns, and the service dynamics and evolution over time, in order to efficiently support and scale the ePrint system.

Existing studies of Internet, media, and enterprise workloads [2, 3, 4] indicate that client demands are highly variable and that it will not be economical to provision the service using "peak" demands. Recently, there have been a few efforts to analyze cloud service provider models and perform comparison studies. Garfinkel [5] and Walker [6] devote their attention to the performance analysis of Amazon Web Services [7]. Some other industry reports [8, 9] present a comparison of cloud service features across different companies such as Amazon and Rackspace. A related problem is considered in [10], where the authors analyze the properties of the applications and the accompanying cost to help in making a decision whether to move or not to move the application in the cloud. The authors show that the specific application characteristics such as workload intensity, growth rate, storage capacity and software licensing cost create a complex, intricate cost function influencing the decision outcome. The application horizontal scalability is another important decision factor. Therefore, understanding workload characteristics of a new service is critical for its efficient support and management.

In this work, using 1-year of recorded user requests to the ePrint service and details of their processing by the system, we characterize properties of this new service and its evolution over time. We offer a workload analysis tool that characterizes the ePrint service and workload profile in a way useful to service providers to get critical insights into user demand and its dynamics over time. The tool offers a practical approach for automatically computing the capacity needed to process the diverse and growing ePrint workload. The same techniques may also relevant for the study of other large-scale cloud-based services.

This paper initially explains the core architecture of HP ePrint system that converts any document entering the ePrint System to a printable format. Then we describe the print job traces which we have collected and analyzed in the paper. The remaining sections present our static and dynamic analysis of the workload as well as outline our approach for automated evaluation of the required system capacity for processing the growing and varying ePrint workload in production environment while satisfying QoS requirements. Finally we offer summary and concluding remarks.

II. Architecture

The HP ePrint service is a complex system with *three main components* as illustrated in Figure 1:

1) *OnRamp Servers* that receive all types of print jobs like Simple Internet Printing Services (SIPS) jobs, ePrint jobs, etc., by Email or from Mobile or Cloud-based applications;

2) *Worker Nodes* that transform jobs into a format suitable for printing, where different types of jobs are processed by different types of worker nodes: Linux or Windows; and

3) *OffRamp Servers* that transmit jobs to printers.

There are several servers with different roles and multiplicities for handling the job that arrives for processing. Figure 1 shows the request flow through the ePrint System. Print job requests are load balanced globally and sent to the OnRamp and Email servers. Then the OnRamp or Email server requests a worker node for the job. The *master server*

978-3-901882-50-0 ©2013 IFIP

determines which worker will process the job (depending on their availability and load), and the OnRamp or Email server sends the job to the assigned worker. There are two types of workers (based on the document format): Linux and Windows. As the name suggests, Windows workers process the Office Documents like DOC, DOCX, etc., whereas Linux workers process the formats like PDF, HTML, JPG, etc. The worker node converts the document to be printed to a printable format; this process is referred to as *rendering*. Multiple rounds of conversions may be necessary depending on the input file format. Once the job has been processed, the worker sends a notification to an XMPP server. XMPP servers operate the communication protocol that interacts with ePrint enabled printers. The XMPP server attempts to notify the printer that it has a document ready to be printed. Once the printer responds to the XMPP server, the worker passes the job to the OffRamp server that sends the print job to the designated printer.

The HP ePrint System is a scalable system with multiple parallel legs of servers. Figure1 shows one leg with some number of servers in each tier. To scale the system more servers can be added per leg or more legs can be added.

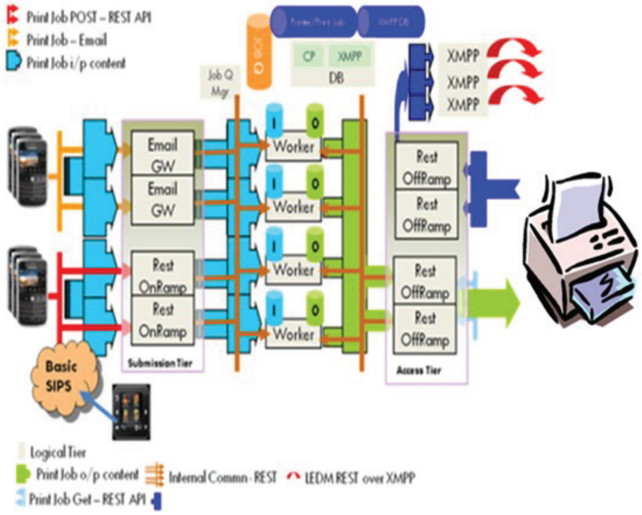


Figure 1. High-level view of HP ePrint System.

First, we analyze which components may become the *bottleneck* of ePrint system. The OnRamp and OffRamp servers do not actually process jobs. They only store data for brief periods as job data is moved in and out of the systems. The main functionality of the XMPP servers is to provide a direct connection to every online printer. The XMPP servers have to handle thousands of concurrent client connections. These servers must continuously connect with the ever growing number of e-Print users. Test measurements reveal that a single 2-core XMPP server with 6GB of RAM can handle 45,000 concurrent connections and a 16-core XMPP server with 64 GB of RAM can handle 750,000 concurrent connections. The connection requirements are statically defined by the memory footprint needed to keep the connection state. Therefore, the additional XMPP servers must be provisioned in proportion to the number of new printers being shipped and registered each month. The number of worker nodes that do job processing and store data for the longest time is a performance concern. Thus the challenge is to determine the adequate number of worker nodes needed for processing the increasing and varying number of print jobs in the system over time.

III. Trace Description

The HP ePrint system is instrumented to collect a variety of metrics and events to monitor a service behavior and performance. There are detailed traces of events for tracking the state and processing time of each print job through the system.. We have collected the hourly traces of the jobs entering into the system over the past year. Each job has been categorized on the basis of the following three characteristics:

1. **Formats** : JPG, PDF, HTML, TXT, etc.
2. **Size** : Small, Medium, Large (*thresholds are different for different formats*)
3. **State** : *Completed* : The job was successfully completed; *ErrorProcessed* : The job was rendered but was not printed, e.g., the remote printer went offline after rendering, empty paper tray, etc.; and *Error* : The job was not rendered at all, e.g., due to a bad format, image too small to print, password protected content, or destination printer does not exist, etc.

The analyzed trace has a total of 5.9 million job records from the time period between May, 2011 to July, 2012.

IV. Workload Analysis

A. Static Workload Analysis

In this section, we present the main statistics and characteristics of the overall workload in the entire trace. There are 20 major formats of the documents that arrive into the system. Prominent document formats processed by the Linux workers are PDF, HTML, JPG, PNG, etc. Additionally, the Linux workers support MAFF which is a special format for instruction pages that are printed when a user first registers a printer for the ePrint service. Document formats of the Windows workers constitute TXT, DOC, PPT, XLS, etc.

Figure 2 shows the percentage of major formats in the entire workload. The chart shows that the Linux formats are more dominant than the Windows formats. They constitute around 80% of all the jobs entering the system. The jobs with MAFF format that correspond to printer registrations represent the second largest fraction after the PDF jobs. The top five formats account for 90% of all the print jobs.

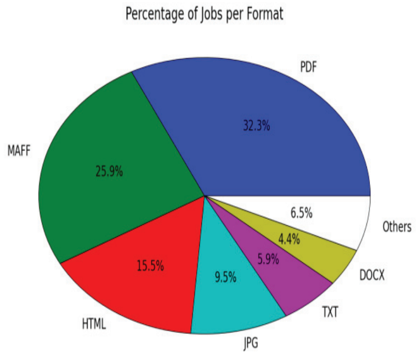


Figure 2. Percentage of most popular job formats.

However, jobs with the same format may vary significantly in size: ranging from few bytes to MB's. The job size does impact the resource requirements and processing time. We have created individual thresholds for each format to further categorized them into *small, medium* and *large* jobs. Table I shows the percentage of the jobs of different formats and the corresponding sizes for top 15 formats. As we can see the majority of processed jobs are of small size, followed by the medium size jobs.

TABLE I. PERCENTAGE AND CHARACTERIZATION OF JOBS BASED ON FORMAT AND SIZE (FOR TOP 15 JOB TYPES)

	% of Total	Small (%)	Medium (%)	Large (%)
PDF	32.3	92.8	6.6	0.5
MAFF	25.9	99.9	0.1	0.0
HTML	15.5	96.3	2.4	1.3
JPG	9.5	81.6	18.3	0.2
TXT	5.9	99.2	0.7	0.1
DOCX	4.4	94.7	4.8	0.5
DOC	3.6	94.1	5.3	0.6
XLS	0.9	94.8	4.8	0.4
PNG	0.7	73.1	24.9	2.0
XLSX	0.6	97.4	2.4	0.2
GIF	0.2	94.1	5.1	0.7
PPTX	0.1	80.7	18.3	1.0
PPT	0.1	74.2	24.7	1.1
TIF	0.1	90.5	9.2	0.3
BMP	0.04	79.7	18.2	2.1

Furthermore, there are *three types* of final job states observed in the HP ePrint system. There are jobs which are rendered successfully and printed at the printer: we call them *Completed* jobs. *Error processed* jobs are rendered successfully in the ePrint system but could not be printed at the remote printer. Causes include: the printer is offline after document rendering, there is a paper jam at the printer, network errors, a paper tray is empty, etc. The rest of the jobs are *Error* jobs and do not get rendered at all. The major causes for these include bad format, image too small to print, password protection, device not being registered, invalid email, etc. Table II shows the percentage of jobs based on their final processing states.

TABLE II. PERCENTAGE OF JOBS BASED ON A FINAL STATE

Completed (%)	*Error* (%)	*ErrorProcessed* (%)
80.9	9.4	9.7

For accurate characterization of workload mix, per job resource consumption, and overall system capacity planning, it is important to characterize the jobs processed by ePrint service using the three described attributes: *format, size,* and *state.*

Table III shows the client geographical distribution and their requests issued to the HP ePrint System. It shows that almost 3/5th of the client base for ePrint is located in America. It also shows that clients in America have a slightly higher tendency to print through HP web connected printers.

TABLE III. CLIENT DISTRIBUTION

	America (%)	Europe and Middle East Asia (%)	Rest of the World (%)
Web Capable Installed Base	58	34	8
Total Web Connected Printer Pages	64	30	6

Finally, we characterize each job using all the three categories: format, size and final state. Figure 3 shows the percentage of jobs covered by the top 10 job types excluding the jobs with MAFF format (registration pages). The jobs with PDF, HTML and JPG formats are the most dominant among all. These document formats are generally associated with SIPS Apps and Mobile Apps. Office Documents like TXT, DOC, DOCX, etc., are relatively small in numbers. The figure shows that the majority of the jobs have *small* size. Also, the majority of jobs are in the *completed* state. At the same time, the *small* size PDF documents exhibit considerable percentage of *Error* and *ErrorProcessed* final states.

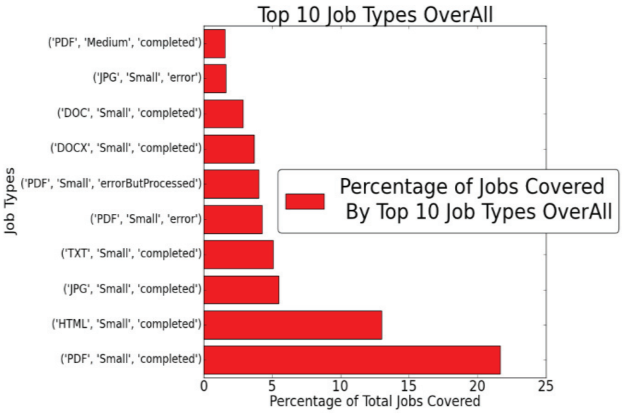


Figure 3. **Percentage of top 10 job types (Exluding MAFF).**

B. Analysis of Workload Dynamics and Evolution Over Time

Since the inception of HP ePrint System two years ago, there has been a tremendous increase in the number of jobs entering the HP ePrint System. Figure 4 shows the number of jobs arriving to the ePrint System per day normalized by the lowest count of the jobs coming in 1 day. The number of jobs has increased almost 4 times over the past year.

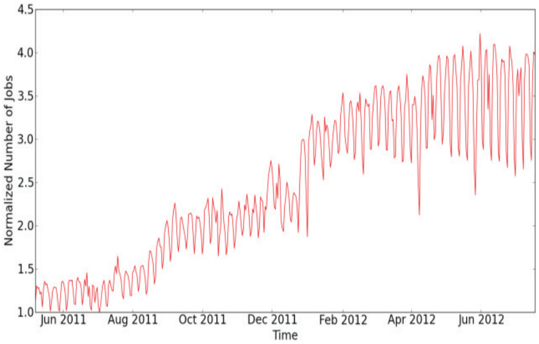


Figure 4. **Normalized number of jobs per day.**

Figure 5 shows more details of this growth trend via presenting *completed* and *error* (aborted/cancelled) jobs over time (normalized by the lowest count of the aborted jobs coming in 1 day). In addition, one can observe the dumb-bell pattern pronounced in both figures. This pattern represents a weekly trend: the number of jobs tends to rise during weekdays and then fall off during weekends. This variance gets larger in the later portion of the trace, stressing increasing burstiness of the user arrivals and a difference in resource requirements of the service during weekdays and weekends.

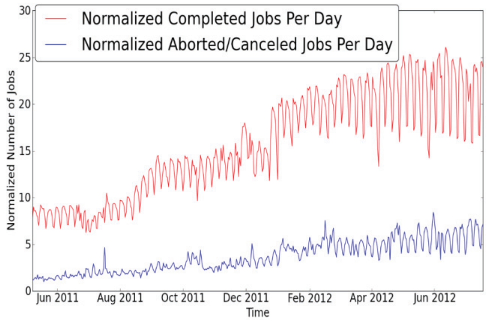


Figure 5. Normalized number of completed and error/aborted/canceled jobs per day.

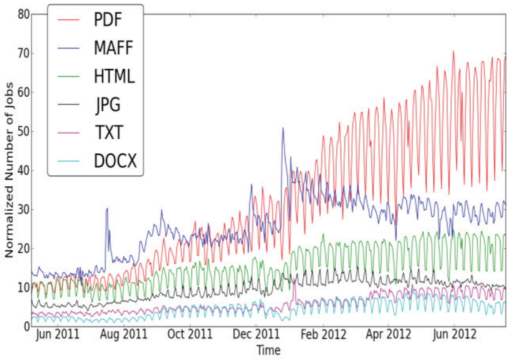


Figure 6. Normalized number of jobs per format per day.

Figure 6 shows the number of print jobs for popular formats per day normalized by the lowest count of DOCX jobs coming in 1 day. The number of these jobs is steadily increasing over the past 1 year. The MAFF format follows a slightly different pattern. This format represents the user registration jobs that are printed by the remote printer automatically when it first time connects to ePrint service. This page contains the printer email address and other relevant information. There is a steady rise of MAFF jobs over the year that reflects the increasing number of registered printers. There is a large spike of MAFF jobs at the end of December - beginning of January (Christmas and New Year holidays). People tend to purchase more printers during the holidays and it leads to a higher number of printer registrations and MAFF jobs during this time. In fact, almost all job formats experience a spike during this time of the year, which suggests people print more during this time period.

Figure 7 shows the trends for top 5 job types, where in addition, we consider the job size and its final state (these jobs are listed in Figure 3). This figure has been normalized by lowest job count in 1 day of the corresponding job type in the bottom of the legend. The figure clearly shows very high increase in the jobs with PDF format. In June, 2011 the numbers of PDF and HTML jobs were similar, but over the last year, the volume of PDF jobs has significantly increased.

Figure 8a shows the percentage of top 5 job formats over time. We can see that the percentage of the PDF documents in the workload mix is slightly rising over the past year, and getting a quite stable portion in the overall mix during the second half of the year. The remaining job types show a quite stable percentage over time in the analyzed time period.

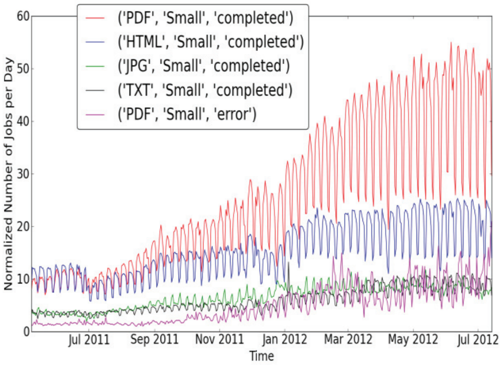
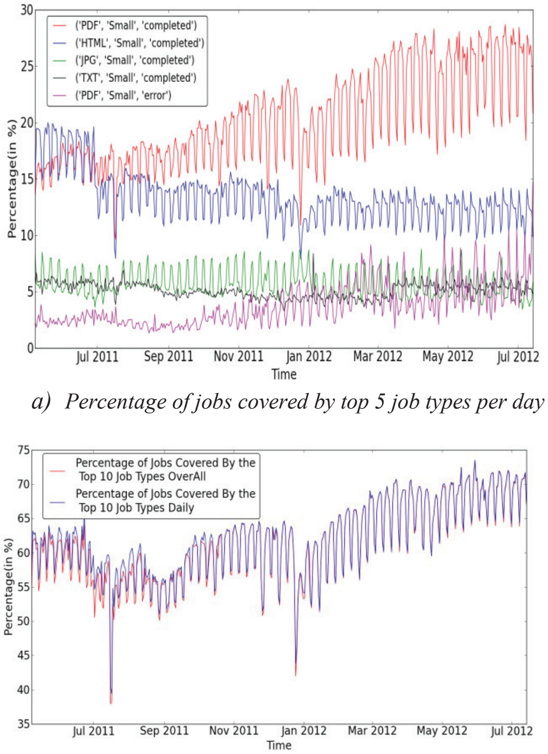


Figure 7. Normalized number of top 5 job types per day (excluding MAFF).

To better understand the dynamics and evolution of workload mix over time, we identified the top 10 job types in the overall trace and then computed the top 10 job types during each day of the trace. We then calculate the percentage of the jobs covered by these two groups during each day. Figure 8b shows that these two groups are very close, and they cover 65-70% of the jobs. If we include MAFF jobs into these groups, we get that approx. 80% of jobs are covered by a small group of job types and that they represent a quite stable workload mix over time, especially over last half of the trace.



a) Percentage of jobs covered by top 5 job types per day

b)Percentage of jobs covered by top 10 job types (10 top jobs overall and 10 top jobs daily)

Figure 8. Analysis of workload mix properties over time.

This behavior might suggest that the users became accustomed to the ePrint service and are printing the known content (from SIPS or Mobile Apps). Such analysis is very useful since it enables simple and intuitive predictive models for projecting performance and required resources with stationary workload mix but varying load over time**.**

Figure 9 shows the average and maximum normalized number of jobs per hour, for the top 10 job types (including MAFF) for 1 week. All values have been normalized by the average job count per hour for *('JPG','Small','Error')*. There is a significant difference in hourly load. To understand this phenomena in a better way, we have analyzed the difference in the load over different weekdays/weekends as well as over different times of the day. Figure 10 shows the average job counts of top 10 job types (including MAFF) during weekdays and weekends normalized by the job count during weekdays for *('MAFF','Small','errorButProcessed')*. Most of the job types have higher average job counts on weekdays than on weekends. The increase in JPG jobs during weekends might be attributed to the SIP Apps like Disney and DreamWorks and suggests that people print some entertainment materials over weekends. However, this does not explain all the differences seen in Figure 9.

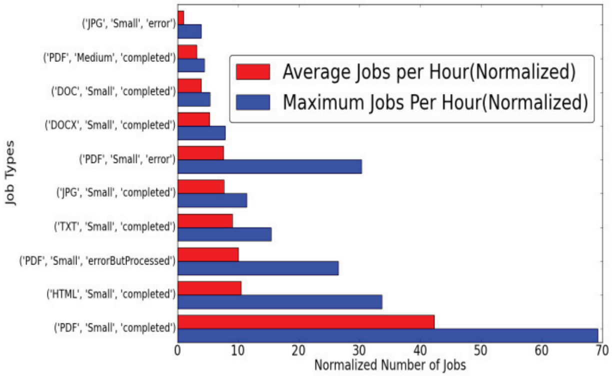


Figure 9. Normalized Average and maximum job counts per hour for the top 10 job types (for 1 week).

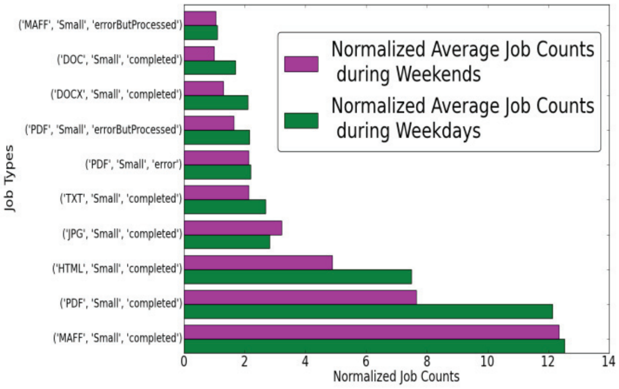


Figure 10. Normalized Average number of jobs for top 10 job types during weekdays and weekends.

Figure 11 reflects a similar analysis for different 12 hour time periods in the day. The values in the figure have been normalized by the job count during day time for *('MAFF','Small','errorButProcessed')*. Basically, we analyze the diurnal pattern: can we observe a difference in the load and workload mix during day or night time for ePrint service. The daily hours were divided into two complementary periods: **i)** from 12 AM to 12 PM, and **ii)** from 12 PM to 12 AM.

We can clearly observe that the load (job counts) during the second half of the day is much higher. It is almost twice as high compared to the load (job counts) during the first half of the day. The explanation behind this diurnal pattern is that the majority of the users of the HP ePrint Service are from US (which is the case as shown in Table III). Thus, they print more documents in the second half of the day rather than the first half that includes the US night time. This diurnal pattern explains the observed difference between average and maximum job counts per hour for some popular job formats.

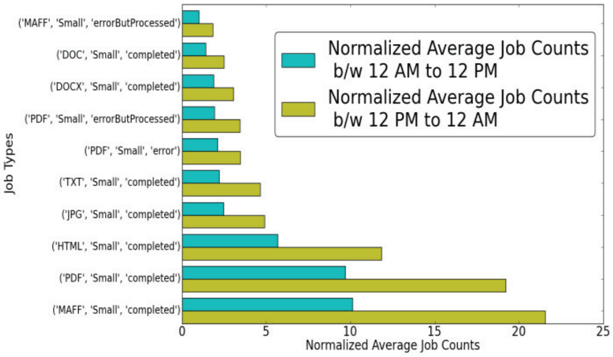


Figure 11. Normalized Average number of jobs for top 10 job types during a different times of day.

V. A Tool for Demand and Resource Prediction

Understanding the load, workload mix, access patterns, and workload trends over time for ePrint service is crucial for properly designing the support infrastructure under the growing user demand. Our next goal is to develop a practical solution for the automatic evaluation of required capacity that is needed for processing a diverse ePrint workload in production environment while satisfying QoS requirements. Currently, we are working on the new tool, called ACE that stands for Automated Capacity Evaluation, which *predicts the required number of worker nodes for processing a given job workload mix under varying load assumptions.* We briefly discuss below the two main components that form the basis of this tool: *Workload Profiler* and *Regression-based Solver*.

Workload Profiler: The ePrint system collects a large number of metrics to analyze its performance. The Profiler extracts the job counts for different formats, size, and final states in a given monitoring interval T_k, along with the system metrics for the same interval. A fragment of the Workload Profile is shown in Table IV describing a total of M different types of jobs denoted by N_i and CPU Utilizations denoted by C_k.

TABLE IV. AN EXAMPLE OF A WORKLOAD PROFILE

Monito-ring interval (T_k)	*PDF, small, completed ($N_{1,k}$)*	*PDF, small, error ($N_{2,k}$)*	*… … … …*	*JPG, small, completed ($N_{M-1,k}$)*	*JPG, small, error, ($N_{M,k}$)*	*Avg. CPU Utilization (in %) (C_k)*
$\boldsymbol{T_1}$	3	7	…	7	6	5.34
$\boldsymbol{T_2}$	10	15	…	14	22	32.57
$\boldsymbol{T_3}$	6	9	…	4	8	12.03
…						

Regression-based Solver: Table IV shows a relation between the amount of processed work (defined by different types of jobs and their counts) and the CPU utilization values measured for this interval. Therefore, we can apply a regression technique for determining the CPU processing requirements of different jobs. We use the Step-wise Non-negative Least Squares Regression to approximate the *CPU cost* $\boldsymbol{A_i}$ for each job type *i* from the following equation:

$$A_0 + \sum_i A_i N_{i,k} = C_k * T_k, \qquad where\ \ i = 1,2, \dots \dots M$$

where $N_{i,k}$ is the number of jobs of type i processed in monitoring interval T_k ; C_k is the average CPU utilization measured duringT_k, and A_0 is the average CPU overhead for keeping the idle system up. We employ the same approach for estimating network bandwidth per job type. Moreover, using this approach we can evaluate the job CPU processing cost on different hardware, or different VMs: Linux or Windows as specified by the architecture. Figures 12 and 13 show measured and predicted CPU and network bandwidth demands based on data from an ePrint test system subjected to controlled workloads. The figures show that the regression method does a very good job at estimating resource demands for given workload mixes.

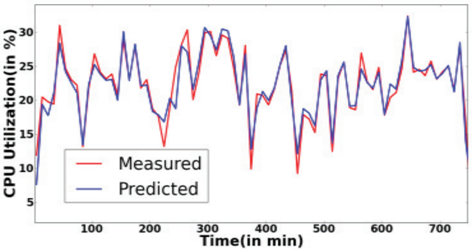


Figure 12. Predicted vs measured CPU utilization for a Linux worker node.

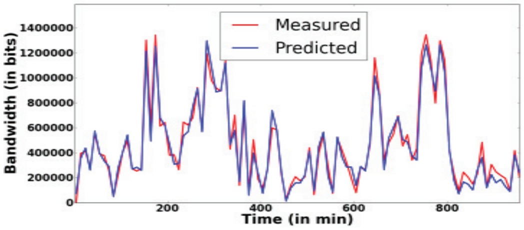


Figure 13. Predicted vs measured bandwidth for a Linux worker node.

Using the derived CPU cost for different job types, we can determine how much load each worker node can process for a given workload while satisfying QoS requirements. For QoS guarantees, we empirically derive a level of CPU utilization that supports desirable job processing latencies and percentage of allowable violations. Using the determined CPU utilization level per worker node, and knowledge that jobs are divided approximately equally across servers we predict the required number of servers for processing a given workload.

CONCLUSION AND FUTURE WORK

This paper presents a workload analysis and capacity planning tool for the new ePrint service. The new tool and the proposed approach supports useful “what if” analysis, capacity management, and predictions of required capacity over time. For efficient support of the service it is important to understand how the service is used by the customers. In particular, what are the most popular documents formats (for providing the adequate load balancing and provisioning support by the Linux and Windows worker nodes); what are the user access patterns during different time of the day (for more predictable dynamic resource allocations), etc. The tool and results are under the evaluation by the e-Print operations team.

The applicability of the proposed capacity planning solution depends on the monitoring support offered by the cloud service providers. Our method of demand estimation and response time modeling is based on the collected CPU utilization measurements at the worker nodes during workload processing. For many cloud environments this represent a challenging task because not all cloud providers report actual CPU utilization values per VMs. Unfortunately, at the user level (within VM) the CPU utilization cannot be measured accurately. Currently, we are working on a new empirical approach for capacity planning in such cloud environments where we only use an available variety of service metrics and workload properties available at the application (service) level.

ACKNOWLEDGEMENTS

Many thanks to our HP colleagues: Subu Iyer for his active support of this project and the ePrint team for providing detailed measurements as well as insights into the service’s design and behavior.

References

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

[2] M. Arlitt and C. Williamson. Web Server Workload Characterization: The Search for Invariants. Proc. of the ACM SIGMETRICS , 1996.

[3] L. Cherkasova and M. Gupta. Characterizing Locality, Evolution, and Life Span of Accesses in Enterprise Media Server Workloads. *Proc. of NOSSDAV*, May 2002.

[4] D. Gmach, J. Rolia, L. Cherkasova and A. Kemper. Capacity Management and Demand Prediction for Next Generation DataCenters. *Proc. of ICWS,2007*

[5] S. Garfinkel. “An Evaluation of Amazon s Grid Computing Services: EC2 , S3 and SQS”. Harvard University, Tech. Rep. TR-08-07.

[6] E. Walker. “Benchmarking amazon EC2 for high-performance scientific computing”. USENIX Login, 2008.

[7] Amazon web services, http://aws.amazon.com/

[8] Rackspace Cloud Servers versus Amazon EC2: Performance Analysis. http://www.thebitsource.com/featuredposts/rackspace-cloud-servers-versus-amazonec2-performance-analysis/

[8] VPS Performance Conversion, http://www.journal.uggedal.com/vps-performance-comparison/.

[10] B. C.Tak, B. Urgaonkar, and A. Sivasubramaniam: “To Move or Not to Move: The Economics of Cloud Computing”. HotCloud, 2011.