REAL WORKLOAD CHARACTERIZATION AND SYNTHETIC WORKLOAD GENERATION

Keerthi Bangari[1], Chittipothula C Y Rao[2]

[1]*Assistant Professor, Department of Computer Science & Engineering, Geethanjali College of Engineering and Technology (A), Cheeryal (V)*
Keerthi2112@gmail.com
[2]*Assistant Professor, Department of Computer Science & Engineering, Geethanjali College of Engineering and Technology (A), Cheeryal (V)*
chinnay6cs820@gmail.com

Abstract

In computing, the workload is the amount of processing that the computer has been given to do at a given time. Workloads are two types namely synthetic and real workload. Real workloads are not publicly available and some workloads are available in the internet like Google trace, world cup 98 trace and Clark Net trace. Synthetic workloads are generated based on our experiments. The real trace is downloaded from Google cluster data which consists of two workloads. In first trace it refers to 7 hours period with set of tasks. In second trace it refers to 30 days period of work. Each dataset is packed as set of one or more files, each provided in compressed Common Separated Values(CSV) format. In this paper we are analyzing the Google cluster data version 2 trace in IBM SPSS statistics and generating another workload called synthetic workload with the same characteristics and behavior of real workload based on formulae which is generated using linear regression in IBM SPSS statistics.

Keywords: *Real Workload; IBM SPSS Statistics; Characterization of Real Workload; Application Benchmarks*

----------***----------

1. INTRODUCTION

Cloud Computing is a term related to technology which allows to use the internet and it maintains the data and applications in the central remote servers. It allows the users to access their files in the cloud with the help of internet. The data access in the cloud computing is more efficient by centralizing the data processing ,storing the data and bandwidth. Simple examples of Cloud Computing is Yahoo Email, GMAIL etc. The internet connection is need to sending the mails. In the cloud there is inbuilt Email management software. And it is controlled by the service providers of cloud such as yahoo, GMAIL etc. Cloud Computing composed into three types of service models and four types deployment models.

In public cloud , the management has the same requirements and follow to shares the infrastructures so has to need of appliances. It is attractive based on the cost as the resource (storage, workstations) used and shares in the communities and exploited already. In Cloud Computing models applications are provided in open source for the users. The user allows to access several important resource on cloud, such as: Applications, Software's and Storing. The users has to store the important information in the public clouds this is , one major advantage for storing information in the cloud. computing machines, these includes: Installing the resources, there configurations; and Storaging. private cloud is a word to specify the cloud platforms which implements with in the private firewalls,it is in information technology control. Hybrid Cloud is combination of both public and private cloud. Service models are reference models on which cloud computing is based. Software applications are hosted by the vendors in the world wide web and the users are allowed to download the application from internet it is called as software as a service(saas). platform as a service (paas) is standard for deliver the operating systems and services which is related across the internet without downloading or installation of services. Infrastructure as a service(Iaas) includes the outsourcing the equipments are used to supports the operation, includes the storing the data, Computer hardware, Servers and network components.

Synthetic workloads are part of performance evaluation in cloud computing. To do performance evaluation on cloud we need experimental platforms, workloads and application benchmarks. A custom testbed can allows to skips the cost of experiments and to has a huge control accomplishments, where it has a cost in terms of system configuration. At server level virtualization is needed. Virtualization commonly addressed as Hypervisors or VirtualMachineMonitor (VMM). Some of important hypervisors are Xen , VMWare ESXi and Kernel Virtual Machine (KVM).Here some platforms are present in the internet for deploying the custom clouds. Some of open

source platforms such as Eucalyptus and OpenStack but these open source platforms are still under development. In opposite to real infrastructure, the software is used to simulating the functions of a cloud platforms, it includes allocation and deallocation of resources. We need to choose an simulator which is available in the past and convert it to our related work. Whereas, using a simulator refers to prepare the software but, inspite of this it has a lot of advantages. The estimated process can reduced in several ways and allows to test set of algorithms and no need to re-configuring all the frameworks everytime. Experiments performed in real infrastructure can takes several hours, but in an event based simulators, it takes only in minute. Simulator allow user to collect any type of data about systems state or evaluation metric. Some of research based cloud simulator are Cloud Sim, Green Cloud, and GroudSim[1].

The aim of this paper is to generate the synthetic workload with the same characteristics and behavior of real workload. First we could download the already existing real trace from Google cloud datacenter. In Google cloud datacenter two real traces are available namely Google cluster trace version 1 and Google cluster trace version 2.In trace version 1 it have only 7 days period of data with time, jobid, taskid and jobtype where as in trace version 2 it have 30 days period of data with Starttime, Endtime, Taskindex, CPU usagemean, Memory usagemean, RAM request and Resource request. We could download Google cluster trace version 2 trace and analyzed the characteristics and behavior in IBM SPSS statistics. Based on the linear regression in IBM SPSS synthetic workload is generated with same statistical similarities of real trace.

2. WORKLOAD SPECIFICATION

The name workload specifies each user request generates an arrival of timestamp. In class of cloud computing the workload models can have different workload attributes. Workloads are two types namely synthetic and real workload. First we have to known that the workload either it is a real trace or synthetic trace.

2.1 Real Workload

Real workloads are not available in the internet this is the main drawback for research on workloads in cloud. For this reason synthetic workload can be generated such that what are the experiments we need based on that requirements real workloads is analyzed and another workload is generated known as synthetic workload with similar behavioral characteristics of real workloads. Some of real workloads available on internet such that ClarkNet ,Worldcup trace 98 , Google cluster trace data.

ClarkNet[2] is a real workload trace which is available in the internet which maintains the useful information of each and every HTTP request and it is collected from two weeks. These HTTP requests received at the ClarkNet trace World Wide Web Server. It is an open source access provider for Metro baltimore Washington DC area.

The data of first log was collect from 00:00:00 28 th of August, 1995 over 23:59:59 3rd of September, and in 1995,it is total 7 days period of information. The data of second log was collect from 00:00:00 4th of September, and in 1995 over 23:59:59 10th of September, 1995, it is total 7 days period of information. In two week's of period we have a 3,328,587 requests. Timestamps have a resolution of 1 second.

It shows cyclic pattern where the workload is more at daytime when compared to night. And the workload is less in weekends when compared to weekdays.

World Cup 98 trace[3] is a real workload which is available in the internet. It maintains all the requests which is made from the World Cup of year 1998 placed in Worldcup 98 trace Web site between 30th of April, and in the year 1998 and 26th of July, in the year 1998.The WorldCup 98 gtrace has received the 1,352,804,107 requests in this period of time.

Google Cluster trace[4] data consists two traces namely first dataset version 1 and second dataset version 2.

The first dataset[5], it provides traces from Borg cell it were taken over 7 hours period of time. Google cluster trace version 1 maintains set of tasks, each and every task will be run on a separate machines. These tasks occupies the one or more cores which is divided into units of memory space in the Computer system. Each task has a unique job; and a job can have two or more tasks. The Google trace data has anonymized in many ways: The Google trace has number of tasks or it is also called as job names ,it has only numerical identifiers; the tasks has only the start time and the consumption of cpu using the linear transformation. By using the transformation of Google cluster data, researchers can able to do workload characterizations and generation.

Here the data can be formed as the blank separated columns. Execution of each and every row reports of single task during the five minutes period of time. The following fields contained by the trace 1.

- Time (int) – Time specifies the start of the data collection in seconds.

- JobID (int) –The task belongs to the unique identifier of job which is also called as parentID.
- Task ID (int) -Execution task of unique identifier.
- Job Type (0, 1, 2) - A categorisation of a work in types of jobs.
- Normalised Task per Cores(float) - The task uses the normalised values of a avg number of cores .
- Normalised Task Memory(float) - The task uses the normalised values of avg memory used by the tasks.

Google cluster trace 2[6] is another workload available in Google data centre. Google cluster is made up of set of machines and it is packed into racks. A cell maintains a group of machines, where single cluster maintained by each cell. And common cluster management shared by each cell. Here one or more tasks is called a job. Each of which is accompanied by set of resource requirements used for scheduling tasks onto the machines. Each and every task maintains a Linux programs, where it mainatins the multiple processes, runs on a single machine. In trace 2 it has 30 days period of data. A trace consists of several datasets. A dataset is composed into a single table, indexed by primary key that typically includes timestamp. Each dataset is packaged as set of one or more files, each dataset is provided in a compressed CSV format.

In trace 2 the jobs and tasks described by the following tables:

- Task event table
- Job event table
- Task resource table.
  In our work we are merging both the task event table and task resource table.

In task event table it contains the following fields:

- timestamp
- missing info
- job ID
- task index - within the job
- machine ID
- event type
- user name
- scheduling class
- priority
- resource request for CPU cores
- resource request for RAM
- resource request for local disk space
- different-machine constraint

In task resource table it contains the following fields:

- start time of the measurement period
- end time of the measurement period
- job ID
- task index
- machine ID
- mean CPU usage rate
- canonical memory usage
- assigned memory usage
- unmapped page cache memoryusage
- total page cache memoryusage
- maximum memory usage
- mean disk Input out time
- mean local disk space used
- maximum CPU usage
- maximum disk IO time
- cycle per instruction (CPI)
- memory access per instruction (MAI)
- sample portion
- aggregation type (1 if maximums from subcontainers were summed)
- sampled CPU usage: mean CPU usage during a random 1s sample in the measurement
- period (only in v2.1 and later)

The combination of both the task event and task resource table contains the following fields:

- start time of the measurement period
- end time of the measurement period
- job ID
- task index
- machine ID
- mean CPU usage rate
- canonical memory usage
- assigned memory usage
- unmapped page cache memoryusage
- total page cache memory usage
- maximum memoryusage
- mean disk input output time
- mean local disk space used
- maximum CPU usage
- maximum disk IO time
- cycles per instruction (CPI)
- memory accesses per instruction (MAI)
- sample portion
- aggregation type (1 if maximums from subcontainers were summed)
- sampled CPU usage: mean CPU usage during a random 1s sample in the measurement period (only in v2.1 and later)
- timestamp
- missing info

- job ID
- task index - within the job
- machine ID
- event type
- user name
- scheduling class
- priority
- resource request for CPU cores
- resource request for RAM
- different-machine constraint

2.2 Synthetic Workload

The synthetic workloads[7] are generated based on the characteristics and behaviour of real workload. First analyse the characteristics of real workload by using some tools called R software and IBM SPSS statistics. Here the synthetic workloads are generated because the real workloads are available in less number and some of the traces are available in internet called Clark net trace and World Cup trace 98.Based on requirements of our project the real workload is statistically analysed and Synthetic workload is generated.

R Software and IBM SPSS statistics tools are used for analysing the synthetic workload.

- R Software: R programming is software which is publicly available in the internet for mathematical analysing of data and for graphics. It can compile and run on different types of Unix platforms, Windows and Mac operating systems. It provides a huge varieties of mathematical and graphical technique, such that different models of linear and non-linear, simple mathematical tests, time series based analysis, classification of data, clustering of data etc. In R programming it is easy to extend the services. It is a language called interpreted .The users can access from command line interpreter. Matrix arithmetic was supported by the R programming. Data structures of R includes vector, array, data frame (same as the tables which is in relational database) and lists. R extends the object system include objects from other objects: regression model, time series and geospatial coordinates. R is a software which is used for data modifications, calculations and graphics.
- IBM SPSS Statistic: It is loaded with the robust analytical techniques and one major benefit is time saving capability and it will help us to analyse the data quickly and easily. In IBM SPSS Statistics there is different versions are available and each product identifies the total pre processed data, from planning to collecting of data to analysing the data, reportings and deployment.

In our work IBM SPSS statistics 22 are used to analyse the real workload.

Here are the some tools used to generate the synthetic workload.

- Faban[8]: Faban is a open source tool which used to develop and run the benchmarks. It supports multi-tier environment such that benchmarks can be run in the set of machines. Simple micro benchmarks are developed and run in the target machines.

The major components of Faban are, the Faban Driver Framework and Faban Harness.

1. Faban Driver Framework is an Application Programme Interface based Structure. It uses to establish the new benchmarks . It gives the in built support for several servers i,e Oracle server, Java,SQL etc. It helps to provide an interface to add other servers.
2. Faban Harness it is a tool to running of slave benchmarks. It maintains a container to benchmark of host and allows the new benchmark to be arranged in a fast manner. It allows to use an web interface in a easy way and to form a queue runs, it adds the huge performance to view, compares and runs the graph outputs.

- Olio[9]: Olio is toolkit and it is available publicly in the internet and it is used to evaluation of the sufficiently, use and efficiency of web technology and it used to provide three initial implementations : Preprocessor Hypertext, Java Enterprise Edition and ROR. This toolkit helps to load the applications to measure the efficiency of a Computer.
- Apache JMeter[10]: Apache JMeter application is publicly available in the internet and it is programmed totally in java so it is called as pure java application .It is used to design to load the applications and to check the working behavior and performance efficiency of an application .It is used to test other for services but mainly to test the web applications.
- Rain[11]: Rain is a Emperical-based workload generation toolkit that uses Constant and statistical distributions to specify the different sources of workloads.
- Vmware VMmark[12]:VMmark is a benchmark tool used to measure the efficiency and scalable of applications which is running in the virtualized environment. VMmark has extensive hardware and software requirements compared to the aforementioned tools. VMmark enables users to measure, view, and compare virtual datacenter performance.It utilizes two previously discussed toolkits, Rain and Olio. VMmark is well documented, but this research does not utilize VMmark due to time constraints.

3. PROPOSED METHODOLOGY

Figure 1 shows overview of our work for real workload characterization and the generation of synthetic workload using IBM SPSS statistics.

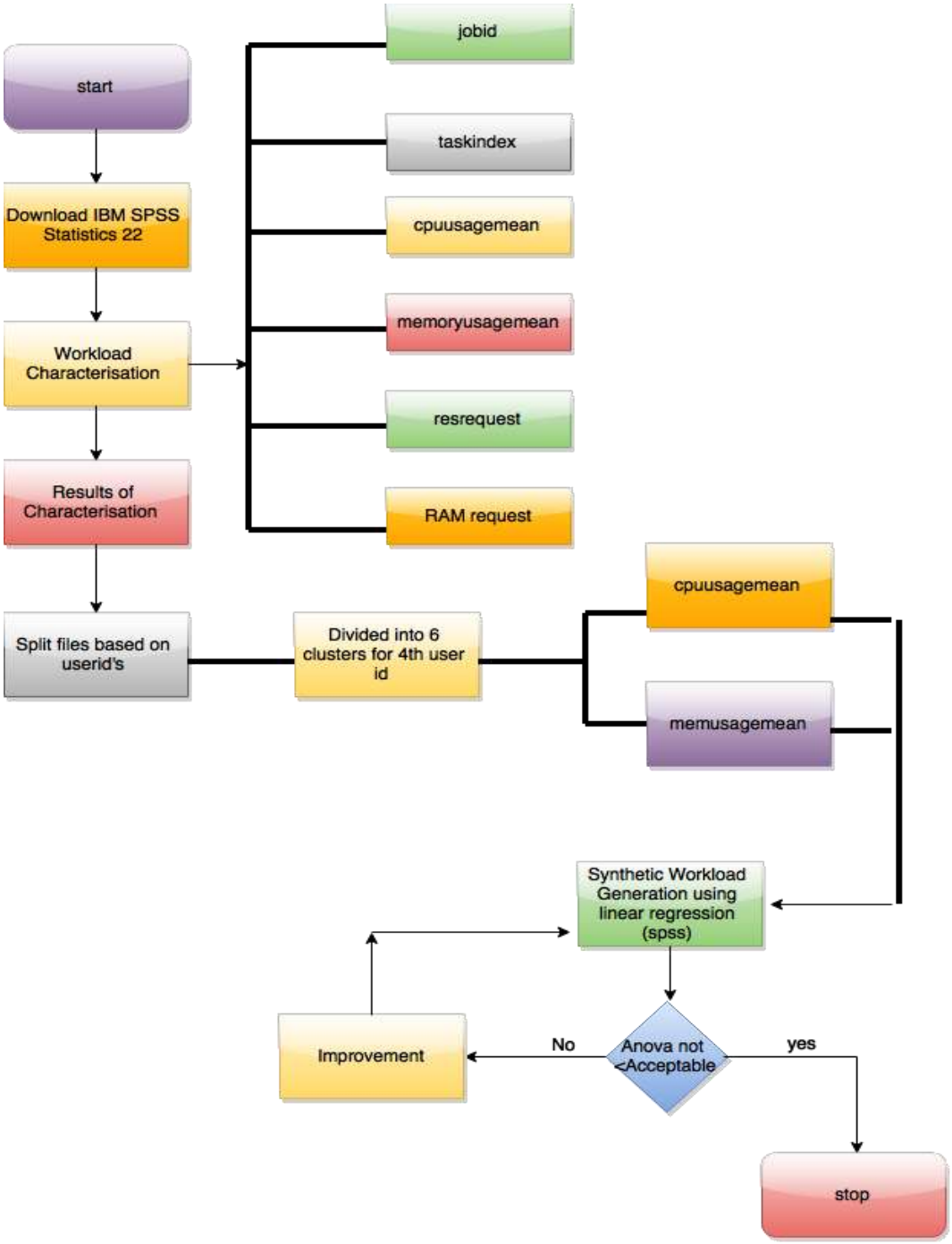


Fig 1: Proposed methodology for synthetic workload generation.

In proposed methodology we have two modules:

- Module 1: Characterization of real workloads.
- Module 2: Linear Regression analysis using SPSS statistics.

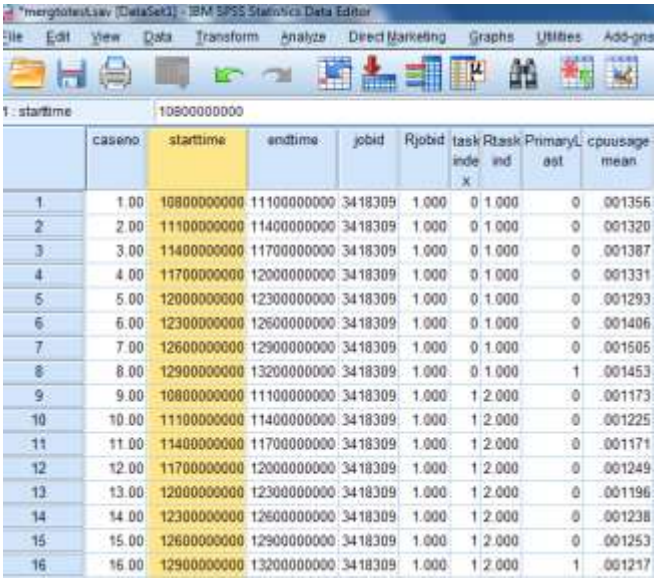
In Module 1, the real workload trace file is downloaded, and subjected to statistical analysis. This is done through IBM SPSS Statistics. The workload characteristics of real workload, Google Trace Data are identified and the computations are done using IBM SPSS Statistics and plots are generated. With these plots or graphs, the distribution of the workload characteristics is found out. As the characteristics and its distribution is available for real workloads, this can be used to generate the synthetic workloads.

In Module 2, The synthetic workload can be generated by using the Linear regression in spss. In spss first the real workload is split into files based on the userid . After splitting into files these files are divided into clusters. Here I am used to split 4th userid into 6 clusters. By using the formulae generated by the linear regression for each cluster different values are generated. Based on these generated values for cpuusage and memoryusage mean the graphs are generated for synthetic workload.

3.1. Characterization of Real Workloads

In Module 1, first download workload file, Google Trace data version 2.Fields which are contained by the google trace version 2 is mentioned in workload specification. And the real data characteristics is analysed in the tool called IBM SPSS statics.

- First the downloaded trace has to be loaded into the SPSS statistics. The snapshot shows the file loaded into the SPSS.



	caseno	starttime	endtime	jobid	Rjobid	taskindex	Rtaskind	PrimaryLast	cpuusagemean
1	1.00	10800000000	11100000000	3418309	1.000	0	1.000	0	.001356
2	2.00	11100000000	11400000000	3418309	1.000	0	1.000	0	.001320
3	3.00	11400000000	11700000000	3418309	1.000	0	1.000	0	.001387
4	4.00	11700000000	12000000000	3418309	1.000	0	1.000	0	.001331
5	5.00	12000000000	12300000000	3418309	1.000	0	1.000	0	.001293
6	6.00	12300000000	12600000000	3418309	1.000	0	1.000	0	.001406
7	7.00	12600000000	12900000000	3418309	1.000	0	1.000	0	.001505
8	8.00	12900000000	13200000000	3418309	1.000	0	1.000	1	.001453
9	9.00	10800000000	11100000000	3418309	1.000	1	2.000	0	.001173
10	10.00	11100000000	11400000000	3418309	1.000	1	2.000	0	.001225
11	11.00	11400000000	11700000000	3418309	1.000	1	2.000	0	.001171
12	12.00	11700000000	12000000000	3418309	1.000	1	2.000	0	.001249
13	13.00	12000000000	12300000000	3418309	1.000	1	2.000	0	.001196
14	14.00	12300000000	12600000000	3418309	1.000	1	2.000	0	.001238
15	15.00	12600000000	12900000000	3418309	1.000	1	2.000	0	.001253
16	16.00	12900000000	13200000000	3418309	1.000	1	2.000	1	.001217

Fig 3.1 Trace loaded into the SPSS Statists

- By using the IBM SPSS statics tool the Google trace data can be sorted per job id and task index. Each task starts at the point 10800000000 and ends at 13200000000.

- SPSS RANK can be used to create a variable holding the rank numbers of the values of some other variable.

	starttime	endtime	jobid	Rjobid
1	10800000000	11100000000	3418309	1.000
2	11100000000	11400000000	3418309	1.000
3	11400000000	11700000000	3418309	1.000
4	11700000000	12000000000	3418309	1.000
5	12000000000	12300000000	3418309	1.000
6	12300000000	12600000000	3418309	1.000
7	12600000000	12900000000	3418309	1.000
8	12900000000	13200000000	3418309	1.000
9	10800000000	11100000000	3418309	1.000
10	11100000000	11400000000	3418309	1.000
11	11400000000	11700000000	3418309	1.000
12	11700000000	12000000000	3418309	1.000
13	12000000000	12300000000	3418309	1.000
14	12300000000	12600000000	3418309	1.000
15	12600000000	12900000000	3418309	1.000
16	12900000000	13200000000	3418309	1.000
17	10800000000	11100000000	3418314	2.000
18	11100000000	11400000000	3418314	2.000
19	11400000000	11700000000	3418314	2.000
20	11700000000	12000000000	3418314	2.000

Fig 3.2 Shows the rank in spss

caseno	starttime	endtime	jobid	Rjobid	taskindex	Rtaskind	PrimaryLast
1.00	10800000000	11100000000	3418309	1.000	0	1.000	0
2.00	11100000000	11400000000	3418309	1.000	0	1.000	0
3.00	11400000000	11700000000	3418309	1.000	0	1.000	0
4.00	11700000000	12000000000	3418309	1.000	0	1.000	0
5.00	12000000000	12300000000	3418309	1.000	0	1.000	0
6.00	12300000000	12600000000	3418309	1.000	0	1.000	0
7.00	12600000000	12900000000	3418309	1.000	0	1.000	0
8.00	12900000000	13200000000	3418309	1.000	0	1.000	1
9.00	10800000000	11100000000	3418309	1.000	1	2.000	0
10.00	11100000000	11400000000	3418309	1.000	1	2.000	0
11.00	11400000000	11700000000	3418309	1.000	1	2.000	0
12.00	11700000000	12000000000	3418309	1.000	1	2.000	0
13.00	12000000000	12300000000	3418309	1.000	1	2.000	0
14.00	12300000000	12600000000	3418309	1.000	1	2.000	0
15.00	12600000000	12900000000	3418309	1.000	1	2.000	0
16.00	12900000000	13200000000	3418309	1.000	1	2.000	1

Fig 3.3 shows the primary last in spss

- In Google trace2 file it has CPU Usage Mean coloumn after preprocessing the data it shows that how much CPU Usage mean used for each job.

cpuusagemean	cpuusagemean_sum
.001356	.01105100
.001320	.01105100
.001387	.01105100
.001331	.01105100
.001293	.01105100
.001406	.01105100
.001505	.01105100
.001453	.01105100
.001173	.00972200
.001225	.00972200
.001171	.00972200
.001249	.00972200
.001196	.00972200
.001238	.00972200
.001253	.00972200
.001217	.00972200
.000233	.00173180
.000218	.00173180
.000213	.00173180

Fig 3.4 shows the CPU Mean usage sum

- In memoryusage mean,after preprocessing the data it shoes how much memory used for each job.

memoryusasge	memoryusasge_sum
.06787000	.54296000
.06787000	.54296000
.06787000	.54296000
.06787000	.54296000
.06787000	.54296000
.06787000	.54296000
.06787000	.54296000
.06787000	.54296000
.06787000	.54296000
.06787000	.54296000
.06787000	.54296000
.06787000	.54296000
.06787000	.54296000
.06787000	.54296000
.06787000	.54296000
.06787000	.54296000
.06787000	.54296000
.06787000	.54296000
.08044000	.64352000
.08044000	.64352000
.08044000	.64352000

Fig 3.5 shows the memory usage sum for each job

- In below screenshot, it shows how much percentage of cpuusage mean and memory usage mean used by each user

username	utilizationpercentage	V2
70s3v5qRyCO/1PCd6fVXnrW8FU/w+5CKRSa72xgcIo=	01084800	1
70s3v5qRyCO/1PCd6fVXnrW8FU/w+5CKRSa72xgcIo=	01056000	1
X1fiQH40IsOBlZKZcc3oourOrsvg/qzZLzBaGgO6iKg=	01056359	2
jWsqrjTWzmOBlm1mmGRz6DcxPGwwxh832jOwL7GTjI=	04259200	3
jWsqrjTWzmOBlm1mmGRz6DcxPGwwxh832jOwL7GTjI=	04137600	3
/fk1fVcVxZ6iM6gHZzqblyq56m5zrmHfpdcZ/zzkq4c=	01216263	4
/fk1fVcVxZ6iM6gHZzqblyq56m5zrmHfpdcZ/zzkq4c=	00944166	4
/fk1fVcVxZ6iM6gHZzqblyq56m5zrmHfpdcZ/zzkq4c=	00911543	4
70s3v5qRyCO/1PCd6fVXnrW8FU/w+5CKRSa72xgcIo=	00938400	1
X1fiQH40IsOBlZKZcc3oourOrsvg/qzZLzBaGgO6iKg=	00932978	2
X1fiQH40IsOBlZKZcc3oourOrsvg/qzZLzBaGgO6iKg=	00891851	2
jWsqrjTWzmOBlm1mmGRz6DcxPGwwxh832jOwL7GTjI=	03996800	3
/fk1fVcVxZ6iM6gHZzqblyq56m5zrmHfpdcZ/zzkq4c=	01034602	4
/fk1fVcVxZ6iM6gHZzqblyq56m5zrmHfpdcZ/zzkq4c=	01070934	4

Fig 3.6 shows the utilization percentage

- Simulating the data specifies the what kind of distribution followed by rjobid, cpuusagemean, memoryusagemean etc .
-

Fig 3.7 shows the simulation

- We know by simulating the data that , the RAM request follows normal distribution. Similarly, we find distributions for other parameters of intrest.

- The distribution which are generated in SPSS are useful to enable us to generate workload with similar distribution in faban workload generator.

3.2 Analasyis of real workload using SPSS Stastistics

Linear regression is used to predict the variable value based on the some other value. The value which we predict it is known as dependent variable and also called as outcome variable. The other variable's value is known as the independent variable is called as predictor variable.

Based on the variable called cpuusage mean and memoryusage mean which is present in real workload these are called as independent variables. By using the linera regession in SPSS statistics we generate the dependent variables called synthetic workload of cpuusage mean and memoryusage mean.

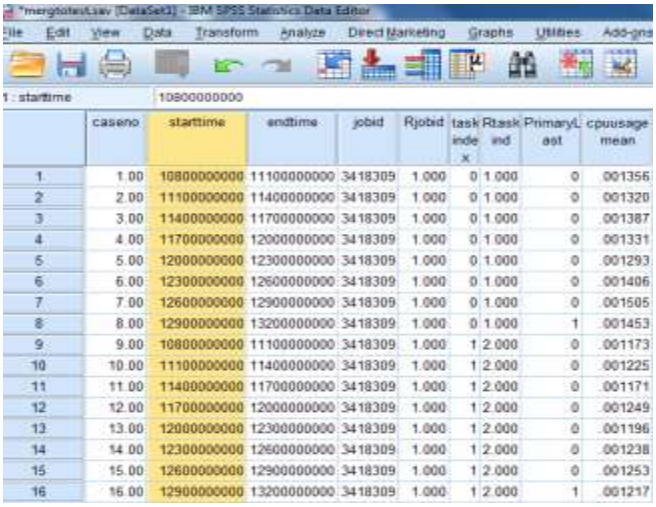
3.2.1 Process to generate the synthetic workload using linear regression in SPSS statistics

In SPSS first the real workload is split into files based on the userid. After that the files are divided into clusters. Here we used to split 4th userid into 6 clusters. By using the equation in linear regression the new values(synthetic workload for cpuusageusage and memoryusage mean) are predicted,these values are known as dependent variables and it is generated based on the values of indepent variables. Graphs are generated for real and synthetic workloads of cpuusage mean and memoryusage mean which shows the difference between the characteristics and behavior of real and synthetic workload.

In SPSS the similar userid files are sperated.There are 81 userid's where each userid is placed in one separated file.we took the 4 th userid among the 81 users where it contains more files.The 4 th userid is divided into 6 clusters by using K-means cluster in SPSS.

Based on the equation the synthetic workload for cpuusagemean and memoryusagemean is calculated.

cpuusage=c+m.jobid+n.memoryusage.
where **c** is constant
m is RAM request
n is CPU request



	caseno	starttime	endtime	jobid	Rjobid	task index	Rtask ind	Primary last	cpuusage mean
1	1.00	10800000000	11100000000	3418309	1.000	0	1.000	0	.001356
2	2.00	11100000000	11400000000	3418309	1.000	0	1.000	0	.001320
3	3.00	11400000000	11700000000	3418309	1.000	0	1.000	0	.001387
4	4.00	11700000000	12000000000	3418309	1.000	0	1.000	0	.001331
5	5.00	12000000000	12300000000	3418309	1.000	0	1.000	0	.001293
6	6.00	12300000000	12600000000	3418309	1.000	0	1.000	0	.001406
7	7.00	12600000000	12900000000	3418309	1.000	0	1.000	0	.001505
8	8.00	12900000000	13200000000	3418309	1.000	0	1.000	1	.001453
9	9.00	10800000000	11100000000	3418309	1.000	1	2.000	0	.001173
10	10.00	11100000000	11400000000	3418309	1.000	1	2.000	0	.001225
11	11.00	11400000000	11700000000	3418309	1.000	1	2.000	0	.001171
12	12.00	11700000000	12000000000	3418309	1.000	1	2.000	0	.001249
13	13.00	12000000000	12300000000	3418309	1.000	1	2.000	0	.001196
14	14.00	12300000000	12600000000	3418309	1.000	1	2.000	0	.001238
15	15.00	12600000000	12900000000	3418309	1.000	1	2.000	0	.001253
16	16.00	12900000000	13200000000	3418309	1.000	1	2.000	1	.001217

Fig 3.2.1.1 Analysed real workload.

The below screenshot shows the real data for both cpuusage and memoryusage mean.

cpuusagemean	cpuusagemean_sum	N_BREAK	memory usage	memoryusasge_sum	N_BREAK_1	maxmemoryusage	maxmemoryusage_sum
001356	.01105100	8	.067870	54296000	8	.067870	54296000
001320	.01105100	8	.067870	54296000	8	.067870	54296000
001387	.01105100	8	.067870	54296000	8	.067870	54296000
001331	.01105100	8	.067870	54296000	8	.067870	54296000
001293	.01105100	8	.067870	54296000	8	.067870	54296000
001406	.01105100	8	.067870	54296000	8	.067870	54296000
001505	.01105100	8	.067870	54296000	8	.067870	54296000
001453	.01105100	8	.067870	54296000	8	.067870	54296000
001173	.00972200	8	.067870	54296000	8	.067870	54296000
001225	.00972200	8	.067870	54296000	8	.067870	54296000
001171	.00972200	8	.067870	54296000	8	.067870	54296000
001249	.00972200	8	.067870	54296000	8	.067870	54296000
001196	.00972200	8	.067870	54296000	8	.067870	54296000
001238	.00972200	8	.067870	54296000	8	.067870	54296000
001253	.00972200	8	.067870	54296000	8	.067870	54296000
001217	.00972200	8	.067870	54296000	8	.067870	54296000

Fig 3.2.1.2 Real workload for cpu and memory usage mean.

The synthetic workload for cpuusagemean can be generated based on the equation

- cpuusage=c+m.jobid+n.cpuusage.

caseno	starttime	endtime	cpuusage1	cpuusagemean	err
5178.00	11100000000	11400000000	.11629222	.10170000	-.01
5179.00	11400000000	11700000000	.11629222	.12870000	.01
5180.00	11700000000	12000000000	.11629222	.12550000	.01
5181.00	12000000000	12300000000	.11629222	.11730000	.00
5182.00	12300000000	12600000000	.11653012	.13040000	.01
5183.00	12600000000	12900000000	.11656062	.10510000	-.01
5184.00	12900000000	13200000000	.11649962	.10280000	-.01

Fig 3.2.1.3 Screenshot for real and synthetic workload generation for cpu usage.(cluster1)

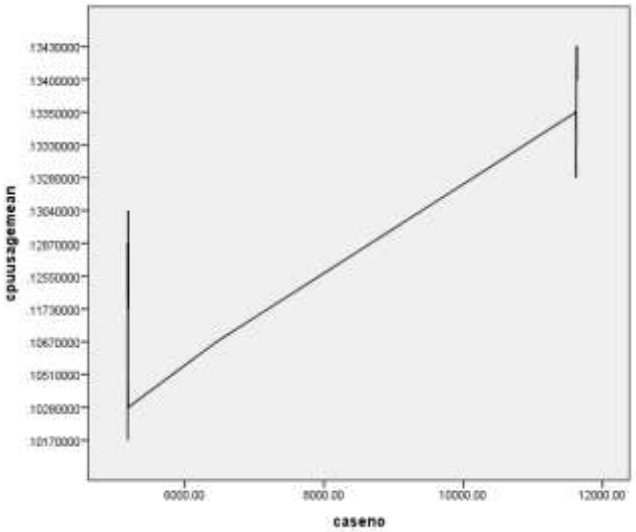


Fig 3.2.1.3 (a) Graph generated for realworkload Cpuusagemean.

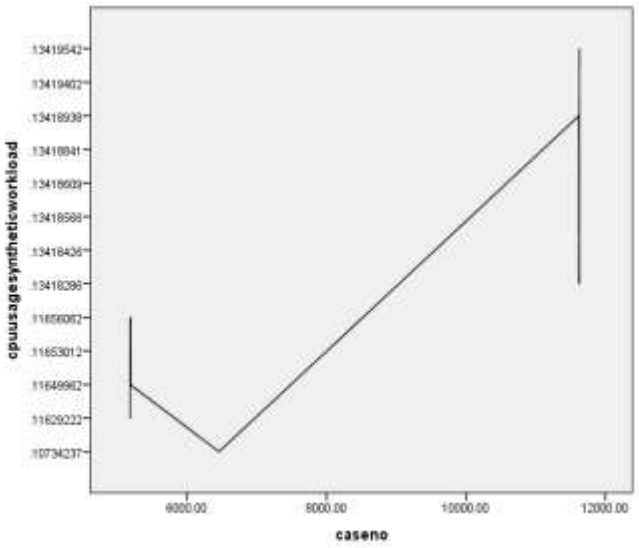


Fig 3.2.1.3(b) Graph generated for synthetic workload cpuusage

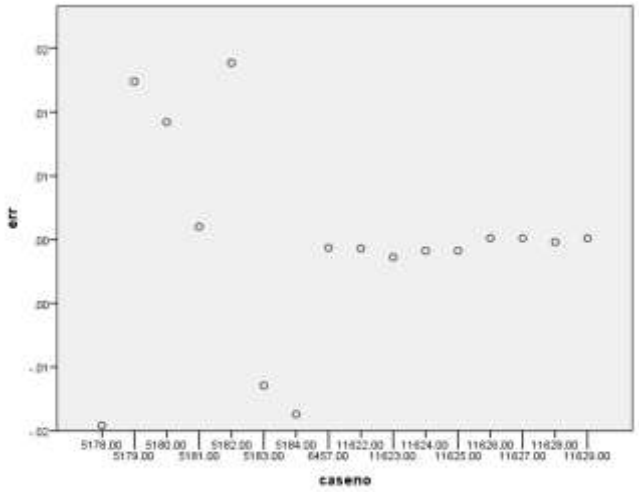


Fig 3.2.1.3(c) Error rate for real and synthetic workload cpuusage mean

The error rate is difference between the real and synthetic workload of cpuusage mean .

caseno	starttime	endtime	cpuusage1	cpuusagemean	error
305.00	10800000000	11100000000	.0558979991	.058650	.00275200
306.00	11100000000	11400000000	.0558979991	.058170	.00227200
307.00	11400000000	11700000000	.0558979991	.056520	.00062200
308.00	11700000000	12000000000	.0558979991	.057070	.00117200
309.00	12000000000	12300000000	.0558979991	.054990	-.00090800
310.00	12300000000	12600000000	.0558979991	.051820	-.00407800
311.00	12600000000	12900000000	.0558979991	.052860	-.00303800
312.00	12900000000	13200000000	.0558979991	.054320	-.00157800
313.00	10800000000	11100000000	.0559213091	.056700	.00077869
314.00	11100000000	11400000000	.0559213091	.056400	.00047869

Fig 3.2.1.4 Screenshot for real and synthetic workload generation for cpu usage.(cluster2)

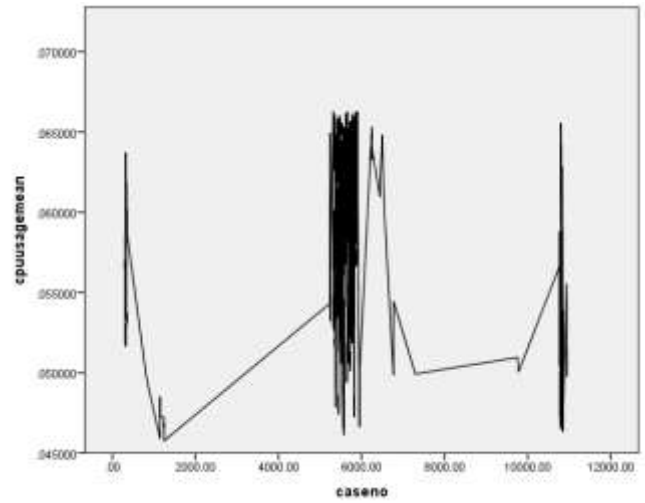


Fig 3.2.1.4(a) Graph generated for real workload cpuusage mean.

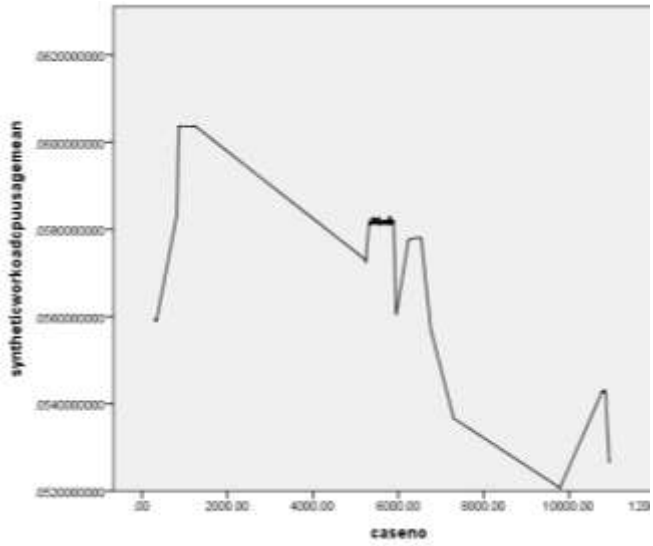


Fig 3.2.1.4 (b) Graph generated for Synthetic workload cpuusage mean.

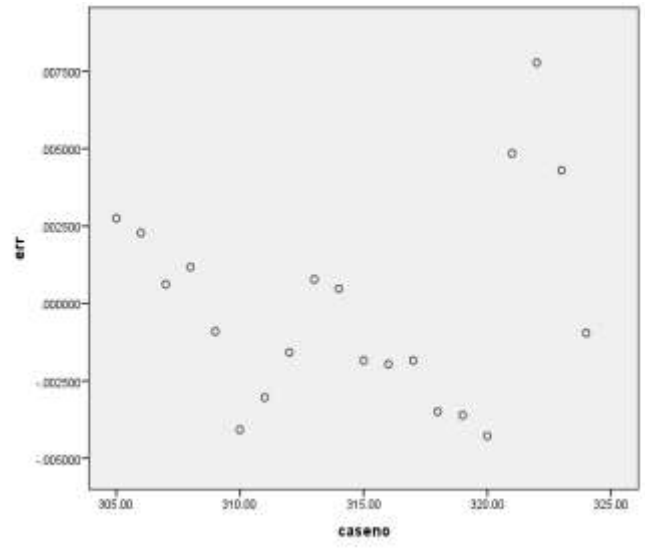


Fig 3.2.1.4(c) Error rate for real and synthetic workload cpuusage

The above fig shows the error rate is difference between the real and synthetic workload of cpuusage mean .

caseno	starttime	endtime	memoryusasge	mem	errmem
5178.00	11100000000	11400000000	29150000	2418519	.0584481
5179.00	11400000000	11700000000	29150000	3048259	-.0133259
5180.00	11700000000	12000000000	29150000	2972675	-.0057675
5181.00	12000000000	12300000000	29150000	2778991	.0136009
5182.00	12300000000	12600000000	29540000	3088413	-.0134413
5183.00	12600000000	12900000000	29590000	2490827	.0468173
5184.00	12900000000	13200000000	29490000	2436501	.0512499
6457.00	11700000000	11728000000	00627900	1772540	-.1709750
11622.00	10800000000	11100000000	00552400	-.0001141	.0056381
11623.00	11100000000	11400000000	00547000	-.0017675	.0072375
11624.00	11400000000	11700000000	00544000	-.0005865	.0060265
11625.00	11700000000	12000000000	00541700	-.0005865	.0060035
11626.00	12000000000	12300000000	00546300	.0017755	.0036875
11627.00	12300000000	12600000000	00550800	.0017755	.0037325
11628.00	12600000000	12900000000	00560000	.0019669	.0046331
11629.00	12900000000	13200000000	00562300	.0017755	.0038475

Fig 3.2.1.6 Screenshot for real and synthetic workload generation for memory usage.(cluster1)

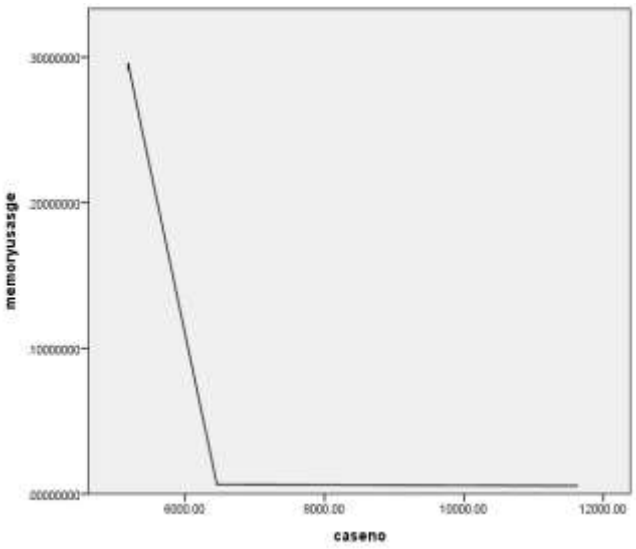


Fig 3.2.1.6 (a) Real workload generation for mem usage.(cluster1)

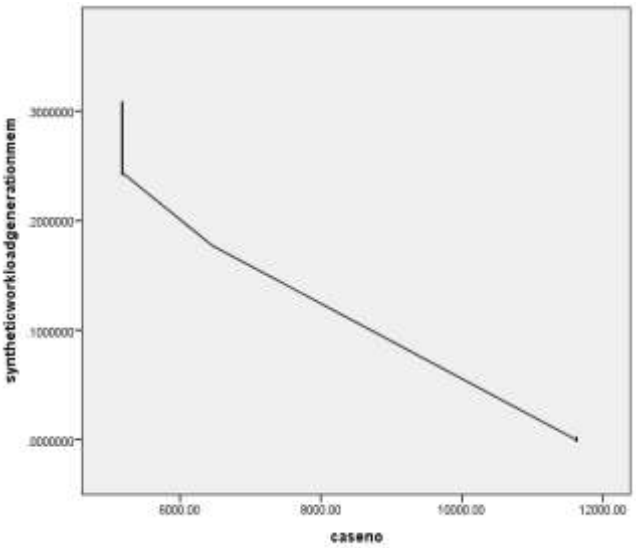


Fig3.2.1.6(b) Synthetic workload generation for mem usage.(cluster1)

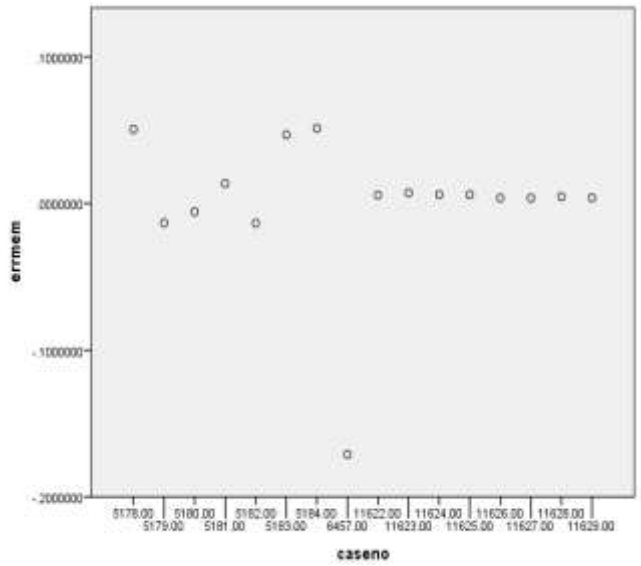


Fig 3.2.1.6(c) Error rate for real and synthetic workload memory usage

cpuusagemean	cpuusagemean_sum	N_BREAK	memory usage	memoryusage_sum	N_BREAK_1	maximemoryusage	maximemoryusage_sum	N_BREAK_2
001356	01105100	8	067870	54296000	8	067870	54296000	8
001320	01105100	8	067870	54296000	8	067870	54296000	8
001367	01105100	8	067870	54296000	8	067870	54296000	8
001331	01105100	8	067870	54296000	8	067870	54296000	8
001293	01105100	8	067870	54296000	8	067870	54296000	8
001406	01105100	8	067870	54296000	8	067870	54296000	8
001505	01105100	8	067870	54296000	8	067870	54296000	8
001453	01105100	8	067870	54296000	8	067870	54296000	8
001173	00972200	8	067870	54296000	8	067870	54296000	8
001225	00972200	8	067870	54296000	8	067870	54296000	8
001171	00972200	8	067870	54296000	8	067870	54296000	8
001249	00972200	8	067870	54296000	8	067870	54296000	8
001196	00972200	8	067870	54296000	8	067870	54296000	8
001238	00972200	8	067870	54296000	8	067870	54296000	8
001253	00972200	8	067870	54296000	8	067870	54296000	8
001217	00972200	8	067870	54296000	8	067870	54296000	8

Fig 3.2.1.7 Screenshot for real and synthetic workload generation for memory usage.(cluster2)

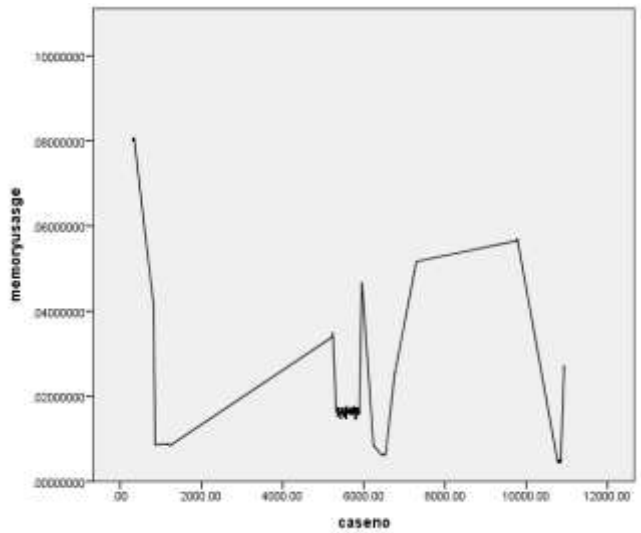


Fig 3.2.1.7(a) Real workload generation for mem usage (cluster2).

The above fig shows the generation of real workload by using the values of memory usage.

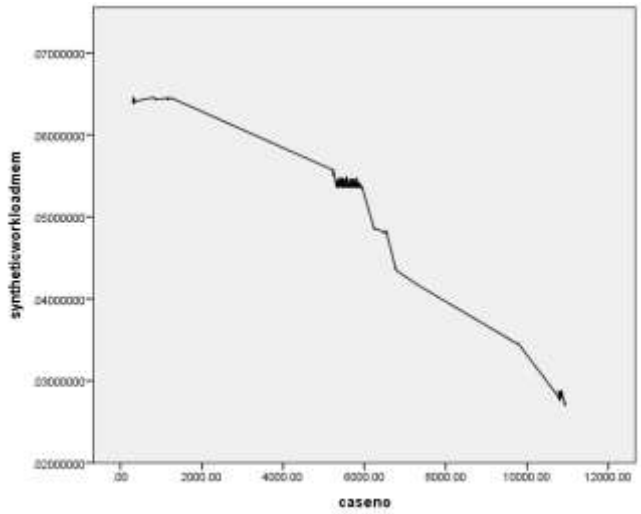


Fig 3.2.1.7(b) Synthetic workload generation for memory usage.(cluster2)

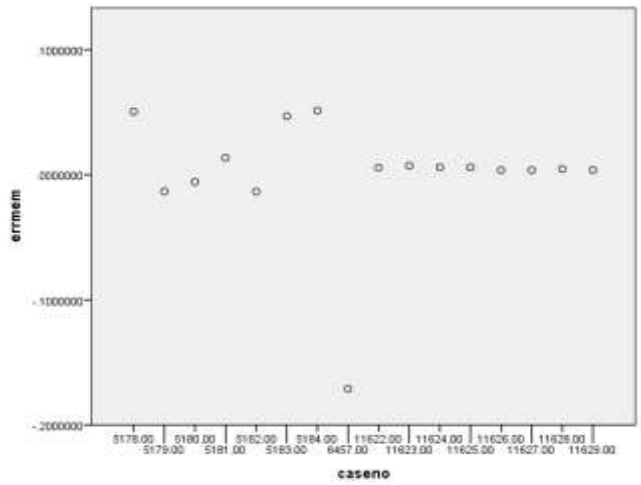


Fig 3.2.1.7(c) Error rate for real and synthetic workload memory usage

4.1 Comparison between Real and Synthetic Workload

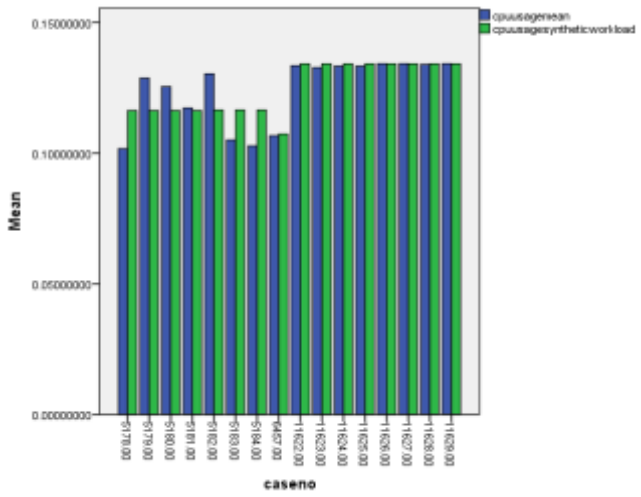


Fig 4.1(a) Comparison between cpu usage real workload and synthetic workload (cluster1)

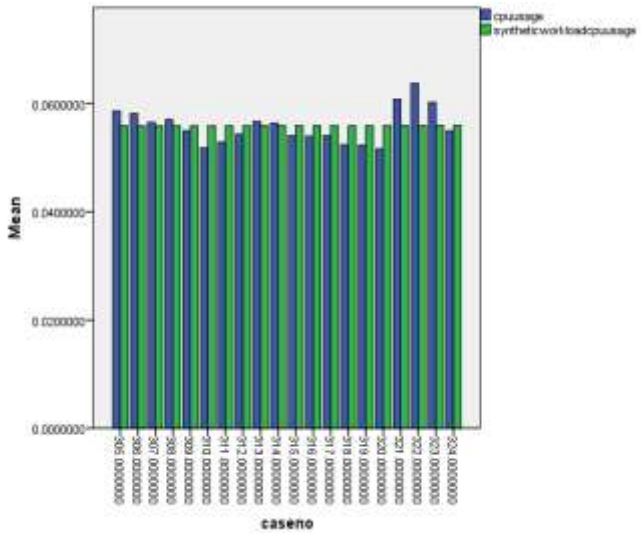


Fig 4.1(b) Comparison between cpu usage real workload and synthetic workload (cluster2)

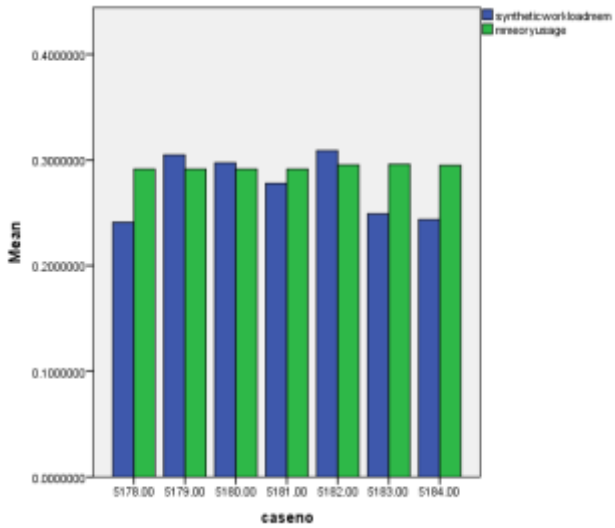


Fig 4.1(c) Comparison between memory usage real workload and synthetic workload (cluster1)

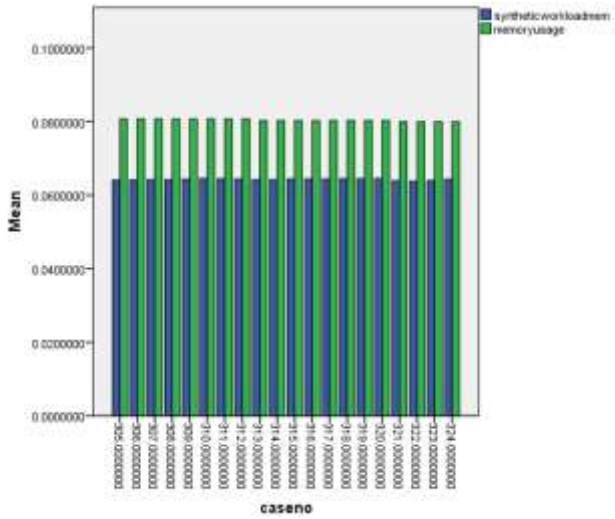


Fig 4.1(d) Comparison between memoryusage real workload and syntheticworkload(cluster2)

4.2 Evaluation of Generated Workload

The Google trace data version2 is downloaded from Google data centre. It contains the several contents in the table such as cpu usage mean, memory usage mean, jobid, RAM request, CPU request etc. The data preprocessed in the tool called IBM SPSS statisticts 22. After preprocessing the data some graph are generated based on the characteristics and behaviour of the real workload. From the graphs the statistical distribution of synthetic workload is matches with the real workload. The Correlation coefficient shows the similarity between the synthetic workload generation and real workload. Below diagram shows the correlation coefficient between the cpu usage and memory usage.

	Synthetic Workload cpuusagemean	Cpu usage mean
Pearson Correlation	1	.781**
Sig. (2-tailed)		.000
N	16	16
Pearson Correlation	.781**	1
Sig.(2-tailed)	.000	
N	16	16

**. Correlation is significant at the 0.01 level (2-tailed).

Correlations

		Mem	Memory usasge
Synthetic Mem	Pearson Correlation	1	.941**
	Sig. (2-tailed)		.000
	N	16	16
Memory Usasge	Pearson Correlation	.941**	1
	Sig. (2-tailed)	.000	
	N	16	16

**. Correlation is significant at the 0.01 level (2-tailed)

5. APPLICATIONS FOR SYNTHETIC WORKLOAD GENERATION.

5.1 Capacity Planning

In capacity planning the resources are provided by the system such that those are somewhat relevant to the quality levels. Here the services are used by the processes in the view of forthcoming demands. For the successful execution the resources has to be designed to reduce the costs and to grasp the stable service-level agreement.

The tools of Capacity planning plans to checking and calculating workloads such that the values are generated synthetically and it allows the analysis of resource utilization.

5.2 Performance of Computer System Using Synthetic Workloads

Energy efficiency is a used for system design, It has been used in two domains called embedded and enterprise. It focus on methods of sum of power benchmarks. that has been released recently. Consider, that EnergyBench was released by EEMBC and SPECpower was released by SPEC to calculate the energy efficiency of systems.. The SWEEP is a structure to generate the synthetic workloads with same characteristics and behavior of real workload. SWEEP is used to form an advanced range of synthetic workloads while it differs in the instructions. The usage of SWEEP is to calculate the performance of fiscal computer systems in workloads. Based on the characteristics of workload it specifies that the performance of a computer.

6. CONCLUSION

The real workload is downloaded from Google cluster data center. It has two versions Google trace version 1 and version 2 which is publicly available. We downloaded the Google cluster trace 2 called as real workload and preprocessed the characteristics of real workload in IBM SPSS statistics and are used to generate the synthetic workload such that, the generated workload should has similar characteristics and behavior of the real workload. We have been able to generate the synthetic workload which we are going to use in research resource provisioning , load balancing, performance testing, energy management and other related areas of research problems being solved.

REFERENCES

[1] Lorido-Botrán, Tania, José Miguel-Alonso, and Jose Antonio Lozano. "Auto- scaling techniques for elastic applications in cloud environments." Department of Computer Architecture and Technology, University of Basque Country, Tech. Rep. EHU-KAT-IK-09 12 (2012): 2012.

[2] ClarknetTrace. ftp://ita.ee.lbl.gov/html/contrib/ClarkNet-HTTP.html accessed on 4/5/2014

[3] World Cup 98 Trace (From the Internet Tra_c Archive).http://ita.ee.lbl.gov/html/contrib/WorldCup.html, 2012. [Online; accessed 13-September-2012].

[4] Hellerstein, Joseph L., W. Cirne, and J. Wilkes. "Google cluster data." Google research blog, Jan(2010).

[5] Chen, Yanpei, et al. "Analysis and lessons from a publicly available google cluster trace." EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2010-95 94 (2010).

[6] Chen, Y., Ganapathi, A. S., Griffith, R., & Katz, R. H. (2010). Analysis and lessons from a publicly available google cluster trace. EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2010-95, 94.

[7] Mao, Ming, and Marty Humphrey. "Auto-scaling to minimize cost and meet application deadlines in cloud workflows." Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis. ACM, 2011.

[8] Documentation, Faban. "Sun Microsystems. 2009."

[9] Halili, Emily H. Apache JMeter: A practical beginner's guide to automated testing and performance measurement for your websites. Packt Publishing Ltd, 2008.

[10] Halili, E. H. (2008). Apache JMeter: A practical beginner's guide to automated testing and performance measurement for your websites. Packt Publishing Ltd.

[11] A Beitch, B Liu, T Yung, R Griffith of California, Tech, 2010 - techreports.lib.berkeley.edu

[12] Marija J. Norušis, and SPSS Inc. SPSS professional statistics 6.1. Prentice Hall, 1994.

[13] Arshdeep Bahga, Vijay Krishna Madisetti, " Synthetic Workload Generation for Cloud Computing Applications", Journal of Software Engineering and Applications, 2011, pp 396-410

[14] Lee Gillam, Bin Li, John O'Loughlin, Anuz PranapSingh Tomar, "Fair Benchmarking for Cloud Computing Systems", University of Surrey, March 2012

[15] Raoufehsadat Hashemian, Diwakar Krishnamurthy,Martin Arlitt, "Web Workload Generation Challenges-AnEmpirical Investigation", HP Laboratories.

[16] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A view of cloud computing," Communications of the ACM, vol. 53, no. 4, pp.50–58, 2010.

[17] Pavlos Kranas, Andreas Menychtas, Vasileios Anagnostopoulos, Theodara Varvarigou,"ElasS: An innovative Elasticity as a Service framework for dynamic management across the cloud stack layers". Sixth International Conference on Complex, Intelligent, and Software Intensive Systems 2012, pp 1042-1049

[18] Zhiming Shen, Sethuraman Subbiah, Xiaohui Gu, JohnWilkes, "CloudScale: Elastic Resource Scaling for Multi-Tenant Cloud Systems", In Proc SOCC'11 2011

[19] Sourav Dutta, Sankalp Gera, Akshat Verma, Balaji Vishwanathan, " SmartScale: automatic Application Scaling in Enterprise Clouds". IEEE Fifth International Conference on Cloud Computing,2012, pp 221-228.

[20] B Urgaonkar, P. shenoy, and et al. " Resource over booking and application profiling in shared hosting platforms." In Proc. OSDI, 2002.

[21] Martin Arlitt, Tai Jin "Workload Characterization of the 1998 World Cup Web Site", Internet Systems and Applications Laboratory HP Laboratories Palo Alto HPL-1999-35(R.1) September, 1999

[22] L. Devroye. Non-Uniform Random Variate Generation. Springer-Verlag, New York NY, 1986.

[32] J. E. Gentle. Random Number Generation and Monte Carlo Methods.Springer, New York NY, 1998.

[33] L. S. Cassandras, Christos G. Introduction to Discrete Event Systems.Springer,2nd ed., 2008.

[34] C. A. Chung. Simulation Modelling Handbook: A Practical Approach. Routledge, USA, 2003.
[35] B. R. et al. I-cluster : Reaching top 500 performance using mainstream hardware. Technical Report ,HPL-2001-206,HP-Labs, 2001.
[36] R. Henderson and D. Tweten. Portable batch system : Requirements specification. Technical report, NAS, NASA Ames Research Center, 1998.
[37] Veridian-System. Openpbs release 2.3 - administrator guide. http://gradea.uci.edu/pbs-doc/pbsadmin guide.pdf, 2000.
[23] D.Feitelson. Job scheduling in multiprogrammed parallel systems. IBM Research Report RC 19790 (87657), 1997
[24] A. Downey. A parallel workload model and its implicationsfor processor allocation. 6th Intl. Symp. High PerformanceDistributed Compute, 1997.