An Approach for Minimizing the Response Time and Improving Availability of Web Services

M. Swami Das, A. Govardhan, D. Vijaya Lakshmi

Abstract*: The worldwide use of the Web-based application is increasing rapidly in various domains like E-commerce, banking etc. The Web users use mobiles, smart devices, laptops and PC. The devices use communication protocols with the Internet based web application. Web services are APIs, design application use of SOA Architecture, SOAP, UDDI and WSDL specifications. In this paper, we have discussed the basic elements, the applications to require high-quality parameters related to computer networking, operating system, software related parameters, response time and availability. The minimum response time to invoke operations with use of Optimized Multi-level Shortest Remaining Time CPU scheduling algorithm to minimize the waiting time to achieve high availability of services even in failure of the system the recovery procedures by providing backup, elastic and Fault-tolerant services. We have used the QWS dataset, Dream set and Grid dataset for experiments. The experiments on this dataset improved performance minimizing response time (RT) and increased availability

Index Choice: Web service, QoS, Response Time, availability, operating systems, FTS, Performance, software

I. INTRODUCTION

The growth of web based applications rapidly, the user require to use Web-based applications, nowadays; tremendously increasing the use of applications like E-commerce, Web applications, banking etc. Business needs are growing the use of Web application by users and clients seek a high Quality of Service(QoS) which include response time, availability. Web services use of Simple Object Access Protocol (SOAP) is message communication, Universal Description, Discovery, and Integration (UDDI) is the registry of services, Extensible Markup language (XML) and other elements [6]. Web users can access the Web applications using communication protocols with the Internet and the business services are available to Web servers. It is necessary to authenticate and identify the genuine user needs to access web services with high-quality parameters services to minimize response time and more availability. In dynamic environment web applications, the service providers ensure the demands of client specifications. The response time is a significant measure in quality of services, is measured in milliseconds (ms). The Web services control is based on the QoS manager to minimize the network traffic, policies and most efficient algorithms are required. The Web server accepts the request, execute the request and send back a response to the client with minimum time [8].

For example; in multimedia applications, the users want to access Web applications with minimum response time with quality parameters of functional resources [1][4][5]. High-quality services provided in enterprise applications to the web users, the services without loss of data the system will provide and guarantee to backup services, security, reliability, and other quality parameters. Availability is an important parameter is provided by backup devices and Fault-tolerant services. Increase availability to authenticated users, the user demands the high-quality services, which include reliability, throughput, response time, best practices and others. For building Web service dependable applications services, WS by different service providers, to identify the fault and resolve the issues immediately without any delay [7][12]. Web service technology use of applications to access on-demand services "*pay as you*" manner, it will be connected to cloud-based web data centers the cloud services and virtualization, these services need to execute parallel applications which are running, at data centers use of the CPU schedulers, this can analyze the problems such as response time and availability[9].

Web service(WS) is communicated to cloud services consists of services such as Platform as Service (PaaS), Infrastructure as Service (IaaS) and Software as Service (SaaS). Now these days mobile phones, smart devices can be connected IoT controlled web-based cloud applications. The remaining paper is organized section 2.related work, section 3. Web service Architecture and Quality parameters, section 4. IoT based control system, section 5. Is a proposed method to improve performance, section 6. Result and discussion section 7. Conclusion & Future Scope.

II RELATED WORK

Hadoop is a framework suggested by ZujieRen [2] use of software library in distributed processing to avoid dataset in large applications are handled by the server which connects to thousands of machines, failure in the application layer and provides more availability use of Hadoop common mode Hadoop (HDFS), Hadoop map reduce, is a Parallel processing large dataset, Hadoop Yarn is scheduling user application with throughput, and successful message delivery over communication channels measured in (BPS). The job slot allocation process is based on priority in the queue, CPU scheduling algorithms, SJF, FCFS, and Round robin used.

Revised Manuscript Received on December 28, 2018.

M. Swami Das, Associate Professor& Research Scholar, Department of CSE, MREC & JNTUH, Hyderabad, TS, India

A. Govardhan, Professor & Rector, Jawaharlal Nehru Technological University Hyderabad, Kukatpally Hyderabad, TS, India

D. Vijaya Lakshmi , Professor & Head, Department of IT, MGIT, Hyderabad, TS, India

The William Stallings[3] suggested a Module of high-speed internet consists of data communication systems the design of application consists of terminals, Personal Computers, Workstations connected in LAN, to handle the load of the network, performance actions, queuing analysis for statistical concepts are most useful.

J. Zhu [4] proposed a Web service positioning method, which combines the advantages of network coordination approach, and collaborative filtering approach. The Wireless Session Protocol(WSP) is a method that finds the landmark of network locations, periodically to monitor the Web server and add the value(network distance like Euclidean distance) to the application and find the values minimum , maximum, mean and standard deviation of RT values from plant lab dataset 359400 and response time values over 200 users.

The A.E Yilmaz [5] proposed a model Genetic algorithm based Simulated Annealing, and Genetic algorithm with Harmony Search. A Hybrid Genetic algorithm use of single and multi-objective approaches.The hybrid GA use heuristic simulated annealing, Harmony searches to optimize quality parameters such as cost, response time and reliability. GA accepts the several execution plans, which may be parallel or serial based on a scenario, the solution has the lowest fitness values, serial, parallel aggregations values use of a selection of the execution plan of web service.

Mathew [10] proposed model for availability of service that operates in percentage of time during the user operation invocation of services. The users and the loss of availability is outage due to unavailable network communication failed, the servers heavily loaded, and Service is not available because of system damage, the necessary steps to improve availability of resources by scalable design of application. This means the system resources, which are available replica of web servers, which is probably assure high availability by FTS. The policies are used for disaster recovery and restore the normal operations to web users.

Problem Definition: Minimizing the response time, and availability of services to be maximized. In table.1 shows the notations are used in the paper.

Notation /symbol	Definition- Description
RT	Response time
RCT	Response completion time
URT	User request time
TP	Throughput
A_v	Availability
R	Reliability
E_m	Error Messages
T_m	Total messages
L	Latency
$r_0,r_1,r_2,\ldots,r_n$	User requests
F(X)	Web page request function
N	Servers
P	Utilization of each server
Np	Utilization of entire system
γ_{max}	Maximum Input rate
X'	Variance- Mean packets
MTTF	Mean time to failure
MTTR	Mean time to Repair
MTBF	Mean time between Failure

Table.1.Notations used

III Web service Architecture and Quality Parameters

Web service performance improvement parameters such as response time, availability and reliability. Mostly used architecture is three-tier architecture consists of model, view, and controller. The important elements are Model elements, quality parameters, response time, latency, availability, service reliability and performance parameters are discussed.

Model, elements and quality parameters

The generalized model of web-based applications is shown in Figure.1. It has a client, web server and database server. The quality architecture is proposed by M Swami Das [22] in Quality Manager one of the significant parameters is response time and availability. The following diagram discusses basic elements are used web service and quality parameters are response time, availability and service reliability.

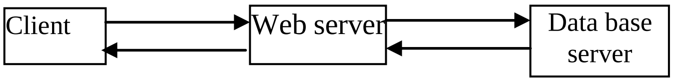

Fig 1: the web server architecture

Basic Elements

The basic elements of Web based system consists of SOAP, WSDL, HTTP, XML, and programming languages are used building of web applications. In software development process use of best practices, M Swami Das[23] improve the performance software applications.

In web-based system applications, the significant role by minimum response time and availability performance. The response time factors by considering the network layer, and transport layer parameters, end to end communication between client and server systems also view of protocols used for message communications like HTTP request, HTTP responses, browsers with modern features like flash player and plug-in, and the behavior of service responses, use of parallel downloading and HTTP pipelining process.

To reduce download cost, high availability of service, minimum delay, and secured internet services the designer must follow quality standards that will improve the performance of web-based applications.

Web service stack architecture shown in figure 2. which has client, server, protocols, encoding and transportation protocols, which influences some of the quality parameters like response time by identify the factors like a minimum response, reduce the overhead of network communication, best services, security protocols, encryption, decryption, RMI over SSL security of web application.

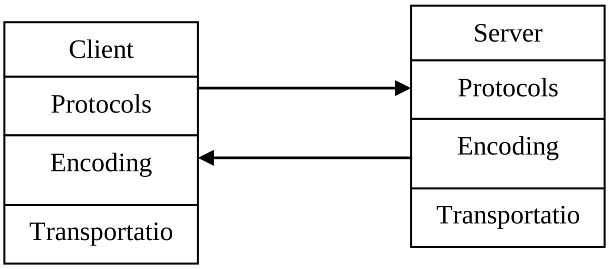

Fig 2: Web service stack

Web service quality parameters:

Quality of services proposed by Marc Oriel[24] defined quality standards, models, and software components to maintainability, and portability. The quality hierarchy in software and web services. They suggested that the quality parameters high priority and more availability, response time, functional correctness, security, and minimizing the cost of web applications based on the analysis. Latency, throughput, availability of service with the use of Fault-tolerant systems and reliability. Web page request waiting time for loading, if slow response transactions of e-commerce slow as a result sales was reduced due to unavailability.

Response Time and Latency

Response time is the amount of total time taken to respond to service. The response time of terminal, where user communication time between terminal users and server is calculated by equation 1

$$RT = RCT - URT \qquad (1)$$

The response time is measured by considering the network traffic, submission, and process of web service. RT depends on the complexity of third-party component services. If complexity increases proportional to failure are also more. Each component such as image, video, text and other elements used to load the web page. Web service delivery, in network, transport and Interdependencies issue. Web site application infrastructure endpoint services use of Communication of components in configuration systems. Latency is mean time to failure and repair when critical situations due to failures and demand of Web service in peak hours to access the resources for example. If university announces the online results at a moment more students to accessed web application.

The client demand the high-quality services, like minimum RT, and others. The Internet which provides the communications between the clients and web-based systems. The traffic analyzer plays a significant role in access web applications with use of XML messages.

Locating the web service by service provider with use of UDDI, Network traffic to load web page and relevant application. The analyzer has sniffer with live packets, HTTP, TCP/IP, SSL, record the headers, response time, message data packet processing format, and connections of web systems.

The web-based system which uses various protocols HTTP, SMTP, TCP/IP, XML and others, Let requests for r_0, r_1..... r_n, then the header, source code download, identifier, IP, static and dynamic web pages, request function is denoted by equation.2. The figure 3.show the user request from a client to server use of transport protocol. Transport layer establishes an end to end communication, the packet analyzer plays a significant role in a response time web systems. The packet analysis, session layer used for establishing sessions between client and server. The response depends on type of network, web pages, web server availability, HTTP traffic, overhead, load balancing and other factors.

$$F(X) = \quad F(X_n) + h/2 \qquad (2)$$

where H step size and X is the function

Client address, server address with HTTP connections can be active, passive modules request type, transaction-oriented requests, and database requests. The system will use single and multiple, processors based on demand, traffic and availability of service. Some of elements influence response time of Web services, HTTP, used for message communications, secure protocols Master Data Service(MDS) is used for waiting security members is applied at time intervals, caution when overlapping user and allows the data user can access. Secure Hash Algorithm (SHA) is the secure hash cryptographic algorithm used hash functions for digital security.

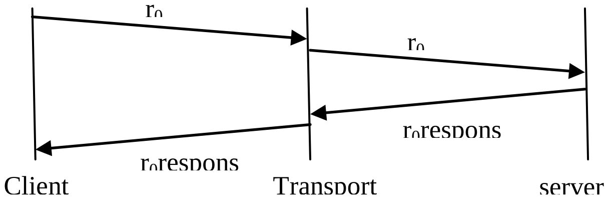

Fig 3: The response server to cleint communication transport protocol.

Data Encryption Standard (DES) is a symmetric key encryption algorithm, by plain text into cipher text by series of mathematical operations.

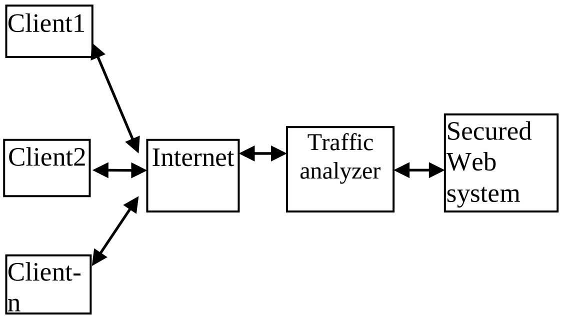

Fig 4: web secured system with a traffic analyzer

The Web secured system in which the clients can use secure web-based applications. The Figure.4 it has clients, internet, traffic analyzes and secured web system.

Latency (network delay), the time taken for the server to process a given request is called latency, (measured in ms), the latency is represented by equation 3. Which is used round-trip delay of the network and the aim is to minimize the latency.

$$L = RT - URT \qquad (3)$$

Most of useful for multimedia applications use of audio, video data streams, it does not affect bandwidth, latency performance depends on several implementation of software components like middleware technologies, Simple Object Access Protocol (SOAP), Java RMI, CORBA, UDDI and Network Protocols.

Throughput: Throughput is measured as the average rate of successful communication messages received in network channel (is measured invocations per second) and described in equation 4.

$$\text{Throughput (TP)} = TI/PT \qquad (4)$$

Where TI is total invocations, PT: period of time

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Availability –Service Reliability

Availability is the number of successful invocations over application. The availability of service is expected all the time (i.e. 24X7), the availability is denoted by equation 5.

$A_v = S_i / T_i$ (5)

Where A_v – availability,Si- number of successful invocations, Ti- Total invocations

Availability is the process where guarantee of service in network communication from client to the web server. The probability of systemup is based on outputs, schedules and workload. The systemup availability is by equation 6.

$SU_a = SU_t / (SU_t + SD_t)$ (6)

Where SU_a is System up availability, $SU_t$is system up time, SD_t– system downtime

The results show high availability is essential for any organizations and waiting time to be minimized

Reliability: The availability of the system even failures recovery steps by Fault-tolerant systems, reliability to be increased (i.e. failures is very less). Reliability is denoted by equation 7. Is the failures are measured over a given period. The software is design to maintain high reliability with minimum failures. This is measured in percentage.

$Reliability(R)=E_m/ T_m$ (7)

Service Reliability is provided based on Service Level Agreement (SLA). The web bases system is capable to provide applications at required level of quality.

Performance Parameters

The performance of web service can be measured by Tarek F. etl[7], Feedback control theory of users, response time of clients to invoke operations, CPU scheduling and efficient scheduling algorithms, throughput, Load balancing (maximum utilization of load), number of host sites. The server needs maintain important clients and server need to adopt QoS with minimum communication delay.

In the design of quality Service use prioritize the request, policies for clients, communication channel (wire communication, wireless communication), web traffic and data encoding (if the data is multimedia application including audio, video, and text etc.). Applications that meet the real-time systems, most sophisticated high-performance algorithms to meet the quality service design use of operating system, CPU scheduling, resource allocations, hierarchical allocations and the performance is for distributed applications by considering the web caching, hit ratio, feedback control system (thread scheduling) pipeline for multimedia application. Server computing services based on the client requests the message to the web server, the server receives the request and handles the requests based on the scheduling algorithms. The delay is called time spent of web server to ready schedule to process request. The sequence of instructions used in Software applications is sequential, iterative, loops and conditional. The instructions may be sequential, serial or parallel, the cyclomatic complexity is used to measure the quality of software application. The designer must ensure to avoid the system by quality management[22]. The operating system use of CPU scheduling, resource allocation, hierarchical scheduling, process capability and middleware etc. The tuning parameters of service utilization in equation 8, response time in equation 9, reliability is in equation 10. and throughput by equation 11.

Parameters need to be improved:

Server utilization (maximized)

Server Utilization = Max (WSi) (8)

Where Wsi is a Web server for i is 1 to n

Response time (minimized)

Response time = Min (WSi) (9)

Where WSi is a Web server for i is 1 to n

Reliability (maximized)

Reliability = Max (WSi) (10)

Where Wsi is a Web server for i is 1 to n

Throughput (Maximized)

Throughput= = Max (WSi) (11)

Where Wsi is a Web server for i is 1 to n

IV. IOT BASED QUALITY CONTROL SYSTEM

IoT based Quality Control System is a new approach used quality control systems in web based applications. The IoT system has sensors to read the data which will regularly monitor the web server performance based on the policies, guidelines, inputs (quality parameters) and real-time values. Figure 5.showsIoT quality control system has users, internet, web server with quality management and database server. Web server is a combination of structure, model, QoS parameters to establish the relationship between a client request services to web service is verify the design by feedback control system. The IoT based approach most important component is the quality manger which evaluate, predict and required level quality bases of quality parameters inputs used by SLA.

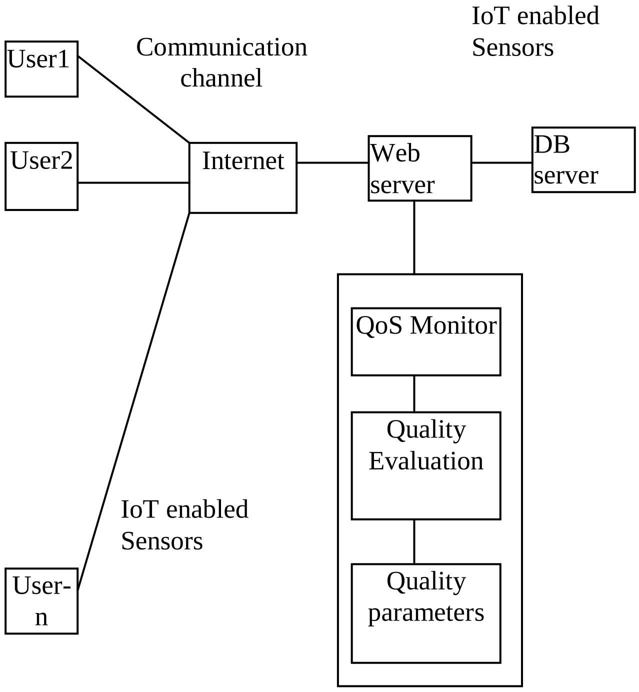

Fig 5: IoT based Quality Control System.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

The response time depends on network topology, media, type of systems, applications, availability and scheduling algorithms on arrival requests. The Operating system use of efficient scheduling algorithms used to improve the performance, CPU cycles, disk, network bandwidth, bottleneck problem to resolve, and utilization of services to be improved. Batch scheduling methods are used in mainframe systems for fast response.

CPU Scheduling is handled by scheduler [16, 19, 20] in which the number of tasks in queue, the CPU executes the tasks from scheduler is two modes, one is preemption (for example Round robin) is service goes another request which uses CPU time sharing, I/O bound time, and another is not Preemption means the processing job which gets completed requests during operations (for example FCFS, SJF). It can achieve the user service and system performance, priority multiprogramming, time slice to the requested process according to CPU scheduling, and re-ordering the process, will improve the performance and throughput, response ratio arrival and service time.

The response time completion time of server, operating system of types long, medium and short-term schedulers. For long-term scheduling the request arriving and processing by server, medium in which events affect cause by execution,(use process preemption), and short-term will use event handling, process dispatch the process is context save and new process arrivals. The performance is by equation 12. Arrival time by equation 13., and service rate by equation 14.

Let

α- Mean time arrival rate, ω- mean time execution rate, p - performance

Performance = Mean time to arrival /Mean time to execution (12)

If p>1, work system is exceeds the capacity, If P<1 means capacity exceeds, work directed interval and stead state when t=0, rescheduling policy,

Arrival time F (t) = 1- $e^{-\text{mean arrival rate, time}}$ (13)

Service rate S (t) = 1- $e^{-\text{mean execution rate,time}}$ (14)

Real-time schedules are two special qualities priorities and periodic, these are used in real-time applications to meet the deadlines of all process. The performance analyzer use of scheduling policies, because workload is directed to schedule to estimate execution time for long running jobs, attacks, delay, cost, relevant events and actions. To find server overhead by the meantime to arrival, and mean execution rate(requests per second).

QoS Management to ensure the Web server to provide the quality guarantee service by virtual server by maximizing the request rate, maximizing the bandwidth, request prioritization, load balance Data center allocations by quest rate, bandwidth rate, target allocation, target utilization and load policy.

V PROPOSED METHODS TO IMPROVE THE PERFORMANCE

The proposed systems moved to single processor to multi-processor and multi-tasking operations which are called distributed and parallel processing, the best use of services in applications is definite improve the performance. The web service queue models (i.e single and multi server models), and case study of scheduling algorithms use of FCFS, SJF, SRTF and RR and FTS used to attain high reliability and availability of services used in real time applications.

Web service queuing –Service Models

Distributed computing is loosely coupled components server handles small processing applications at client places, and coordination is well defined. The web server will handle a large applications where jobs are well defined. The processors are tightly coupled to target data systems which is shared by different processors. It is necessary to communicate different processors to parallel processing use of same machine instructions at communication very fast shared secondary drives, the advantage improve performance. The speed depends on applications, high fault-tolerant systems, if one processor fails then re-scheduled to another processor, this reduced throughput and increase availability. The demand and increasing computational services by adding more processors and maintenance cost are lower than new processors by load balancing. For parallel processing is mutual exclusion the to access more than one process in critical regions, deadlocks can be ignored, detected and recovered, preventive measures in software design avoid deadlocks[15]. The main of the process in queue of two type methods one single server queue and another multi-server queue is discussed below.

Single Server Queue

The client request a queue one web server will process the request. The drawback is RT is very poor, if the server is failed, there will no alternative another server to process for further request. To overcome this is by introducing a multi-server queue model is described section 5.1.2. The response time is increased due to more loads on network systems. The single queuing server model is shown in Figure 4. For example let us assume γ= arrival rate, W= items in the queue, Tw = waiting time, Server, Ts = service time, P= utilization, X. single server queue of web requests arrival rate is described in equation 15.

Mean ($\gamma_{\max)}$ = 1/Ts (15)

Where arrival rate maximum (γmax) which depends on Service Time (Ts), the depends characteristics of information, queue size and dispatching algorithms FIFO, FCFS.

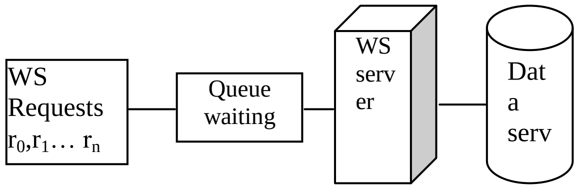

Fig.4: Single Queuing and Single server Model

Multi Server Queue:

The Multiserver queue model has web service scheduled by multiple requests to sharing queues, arrival request at least one server is available, then immediately dispatched request to the server. Multiple servers shown in Figure.5, it has client requests,

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

multiple queuing waiting with multiple web servers the Maximum utilization is 100%. Among the ‘N’ servers at least one server is available for web operations. Maximum utilization of servers is described in equation 16. Maximum input rate is described in equation 17.

Let N= servers, P= utilization of each server, Np: utilization of the entire system, U= traffic intensity.

Maximum utilization= NX Utilization percentage (16)

Maximum input rate is

γ_{max}= N/ (T_s) (17)

Where N is sample size, T_s is the service time and multiple queue servers to the services.

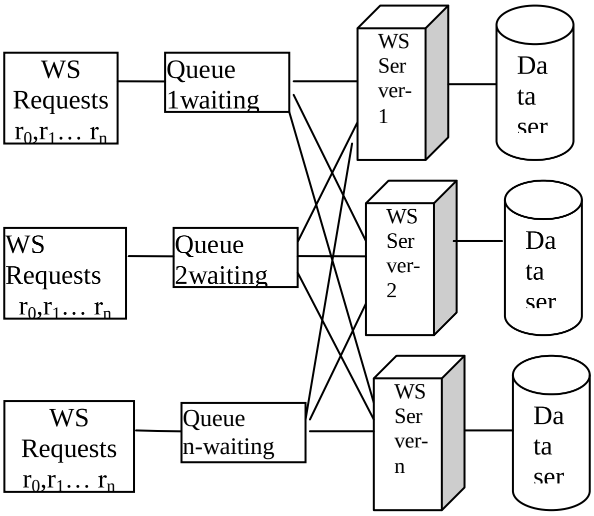

Fig 5: Multiple Server Queue of WS requests

Estimation model of queuing analysis input parameters mean and standard deviation of arrival rate and service time of the new system. Collection of terminals connected with network systems, multiplexers to share the load of active systems.

The sampling rate of packets are generated from terminal of the size of packets, during the period of time, Estimate mean of packets is described in equation 18., and variance is equation 19[6].

$$X' = \frac{1}{N}\sum_{i=1}^{n} X_i \qquad (18)$$

For i {1 to n}, N sample size, X_i i th input packets

Variance (X) = σ2 / N

(19)

The existing Model is Multi-Queue processor Model with Shortest Job first is shown in figure 6. The proposed enhanced model which is Multi-Queue processor with Shortest remaining time which improves the performance by minimizing the waiting time is shown in figure 7., proposed algorithm 1. And optimized multi-level shortest remaining time CPU scheduling algorithm.

Priorities are low and high we can take numerical value by setting priority 1- High, and 7 Minimum priorities among process (1 to 7) or binary value (0 represents low probity and 1 represents high priority)

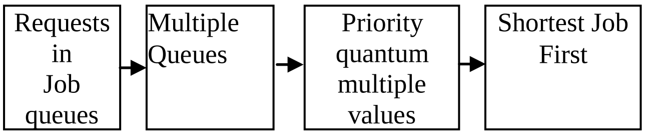

Fig 6: Exiting Approach for CPU scheduling algorithm

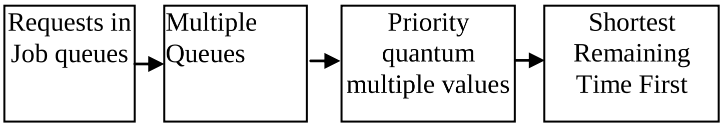
Requests in Job queues
Multiple Queues
Priority quantum multiple values
Shortest Remaining Time First

Fig7: Proposed Approach Optimized Multi-level -Shortest Remaining Time CPU scheduling algorithm

Algorithm 1. The optimized Multilevel Web service CPU Scheduling algorithm

```
Input: Jobs, $J_1$, $J_2$,..$J_n$
Job priority $Jp_1$. $Jp_2$, ..$Jp_n$
Arrival time $Ja_1$, $Ja_2$,.. $Ja_n$
Burst time    BT ($J_1$), BT ($J_2$),…. BT ($J_N$)
Output:
Average waiting time (AW)
Average Turnaround time (AT)
Average response time (AR)
Data structures:
Queues $Q_1$, $Q_2$,,. $Q_N$ (Multiple queues)
Begin
step1) Initialize the variables
Step2) Read Jobs, priority, arrival time and burst time
Step3) Initialize the time quantum of Queue
    Let us assume the all the jobs are entered into the first queue
  If (burst time of Job ≤ time quantum of the job) then
        Execute the job in queue
else
      Process job to next queue
          If the processes in the last queue i.e 'N'
     Sort the jobs based on shortest remaining time next in queue.
  // this improves the Turn around time, response time and waiting time
Step4) Repeat the step 3 until all jobs are executed
Step5) Print the Average waiting time (AW), average Turnaround time (AT)
And average response time (AR)
Step6) stop
```

Case study:Using algorithm 1. Optimized Multi-level CPU scheduling algorithm1. With use of FCFS, SJF, SRTF and RR scheduling methods simultaneously described in table 2. Shortest Job First in table 3., Shortest Remaining Time First in table 4., and Round Robin in Table 5**.** Consider an example to explain the proposed concept Let us take four processes P_1, P_2, P_3, P_4 with arrival time 0,3,10,12 simultaneously and estimate time of 10,5,3,1 simultaneously.

Finding the waiting time by FCFS:

Waiting time=Starting time-Arrival time.

	Arrival time	Estimate time	Starting time	Waiting time
P_1	0	10	0	0
P_2	3	5	10	7
P_3	10	3	15	5
P_4	12	1	18	6

Table 2. Gantt Chart for FCFS

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Average waiting time=4.5

Finding the waiting time by SJF:

Waiting time=Starting time-Arrival time.

Process	Arrival time	Estimate time	Starting time	Waiting time
P_1	0	10	0	0
P_2	3	5	10	7
P_3	10	3	15	5
P_4	12	1	18	6

Table 3. Gantt Chart for SJF

Average waiting time=4.5

Finding the waiting time by Shortest Remaining Time First (SRTF):

Waiting time=Turn around time-Estimate time.

Turnaround time=Completion time-Arrival time.

	Arrival time	Estimate time	Starting time	Starting time	Completion time	Waiting time	Turnaround time
P_1	0	10	0,8,14	0,8,14	19	9	19
P_2	3	5	3	3	8	0	5
P_3	10	3	10	10	13	0	3
P_4	12	1	13	13	14	1	2

Table 4. Gantt chart for SRTF

Average waiting time=2.5

Round Robin algorithm to find Waiting Time

Waiting time=Turn around time-Estimate time.

Turnaround time=Completion time-Arrival time.

	Arrival time	Estimate time	Starting time	Completion time	Waiting time	Turnaround time
P_1	0	10	0,2,6,10,16	18	8	18
P_2	3	5	4,8,14	15	7	12
P_3	10	3	12,18	19	6	9
P_4	12	1	15	16	3	4

Table 5. Gantt chart for Round Robin

Average waiting time=6

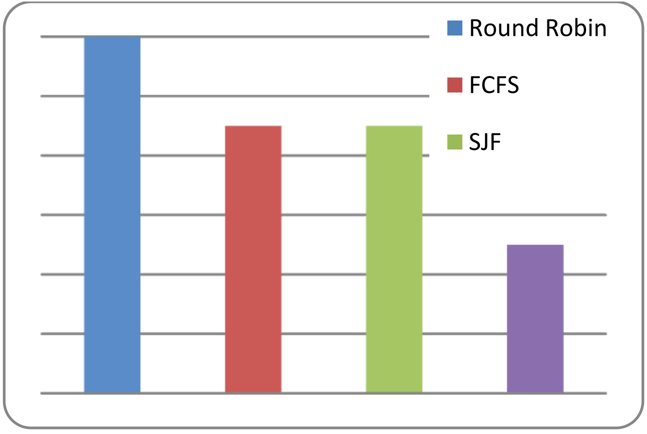

Fig 8: The response time of WS by Multilevel RR, FCFS, SJF and SRTF (Proposed Method).

Availability by Optimized Multi-level -Shortest Remaining Time CPU scheduling algorithm which improves the performance by a case study in results table 1. using FCFS, table 2. SJF, table 3., SRTF, and table 4. Round robin among these improve performance response time is shown in Figure 8.

Use of Fault-Tolerant Systems for High availability

The availability is a UDDI service by web service provider, and it is the most important and heart of the web service. Due to unavailability web systems, the users unhappy due to poor service. To provide high availability by planning, re-designing, analyzing, predict failures by eliminating single point, the suggestions recommendations as 1) Presentation layer is dynamic content caching, improve UI technology relevant functional feature 2) the business layer is the SOA will help to achieve highly available due to service failures 3) database layer is a high volume of data, bottleneck.

The designer and developer follow the best practices[23] in the development of software applications. The software component services are interoperable and distributed systems.

A distributed file system which depends on transparency, and file sharing semantics, modification of file visible to others. Fault tolerance is the system availability due to system faults and failures. Performance considered by efficiency and scalability, if the failure of service due to various causes to recovery and backup, Fault-tolerant services provided without any time delay.Web service functional features that improve the quality[11,12,19], high availability by providing elastic, and Fault Tolerance Web server is shown in Figure 9., which will provide high availability. For example.eBay, Flipkart, Amazon, etc.

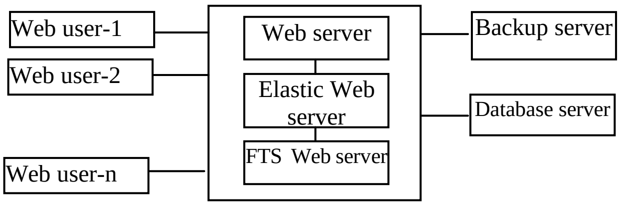

Fig 9: Service providers with high Infrastructural Facilities

The web service provider use of multiple WS service providers to resolve the issues, service not available due to failure, the WSDL and WS time transactions and default failures to service providers. The defect failures are identified and backup recovery procedures in code, Transaction state will be recorded, that follows (Atomic, Consistency, Isolation, and Durability), Fault-tolerant service will provide a better method, but it is too costly, For example in real time Amazon Web service EC2, S3, are available fault tolerant services, and elastic load balancing applications [10] [13]. The failure of Web services because of design failure, and other reasons. Load balancers, health checkups, and FTS used to attain high availability. The systems running the workload, in subnet, provide the same functionalities at multiple zones, provide auto -elastic scaling, MTFS and recovery use of FTS.

To design web service applications loose coupling and high scalability, and some of the most important concepts in high availability are discussed here.

Mean Time to Failure (MTTF) is the system is a time of failure, an outage of the system. Mean time to repair(MTTR) is the time taken when considering the failure time to till repair completion time(i.e. recovery time of available web services), the amount of time spent to bring back services by providing fault-tolerant services.[13].The availability is represented by equation 20.

Availability= MTTF/ (MTTF+MTTR) (20)

Increase MTTF,MTTR decrease the redundant software and hard ware.

Failure of Web service is due to failure service, crash, disk failure, communication failure and heavily loaded system etc. The availability of web services, in real time critical applications availability use MTBF.

Meantime Between Failure: Is the time of failure and recovery time or measured the sum of MTTF and MTTR, This is a most critical metric for real-time applications. The MTBF is represented by equation 21.

MTBF = MTTF + MTTR (21)

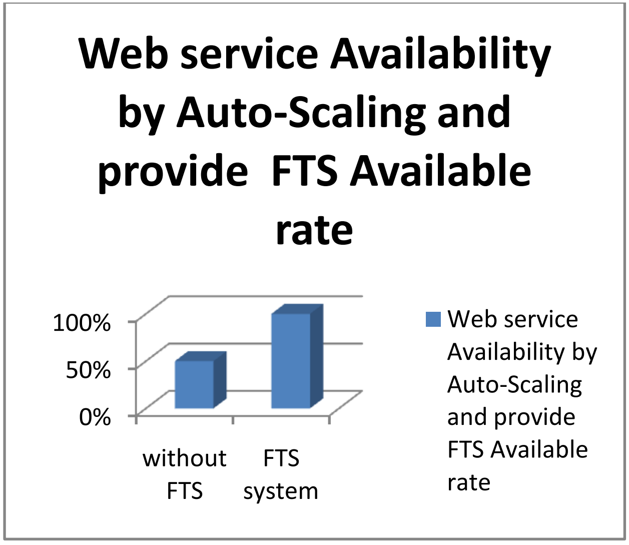

Figure 10. Web service availability by Auto scaling and FTS services improve performance

The Availability of Service by equation 20. by minimizing MTBF in equation 21, by auto-scaling and providing context switching Fault tolerant service will improve the performance.

VI RESULTS AND DISCUSSIONS

The web service queuing model single queue into multiple queues with priority quantum use of Optimized Multi-level Shortest remaining time CPU scheduling algorithm got waiting time 2.5 seconds compared with other methods round robin, SJF, FCFS. is shown in Table 2,3,4 and 5 and Algorithm 1. The Response time, availability, throughput, successability and reliability of QWS data normalized values minimum, average and maximum measurements shown in table 6. and in Figure 11. The values of RT, availability ,successability and throughput normalized values lying between 0 and 1.

QoS parameters	Units Measurement	Minimum	Average	Maximum
RT_N(Response time)	Milliseconds	0.007415	0.0769	1
AV_N(Availability)	Percentage	0.07	0.811456	1
TH_N(Throughput)	Invocations/sec	0.00232	0.209641	1
SUCC_N(successability)	Percentage	0.08	0.838871	1
REL_N(Reliability)	Percentage	0.370787	0.784083	1

Table 6. QWS normalized data minimum, average and maximum values lying between 0 and 1

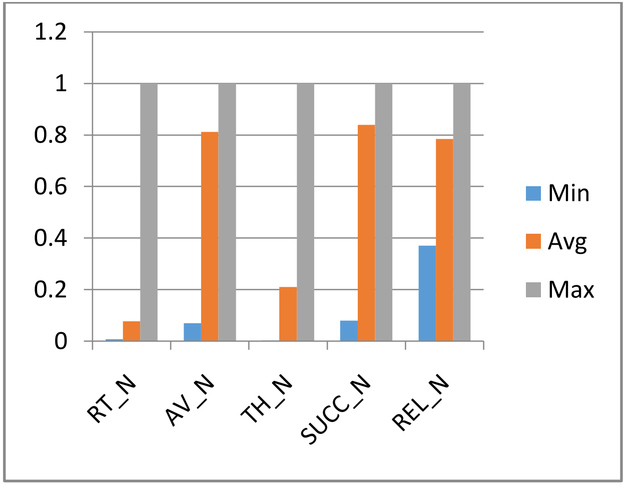

Fig11: QWS data the comparisons of Minimum, Average, Maximum values of Response time, Availability, Throughput, Successability, and Reliability.

The response time of Web-based application dream set data conducted experiments by using R language it shows in figure 12. At users 180th user the RT is maximum and initially is low for dream set data

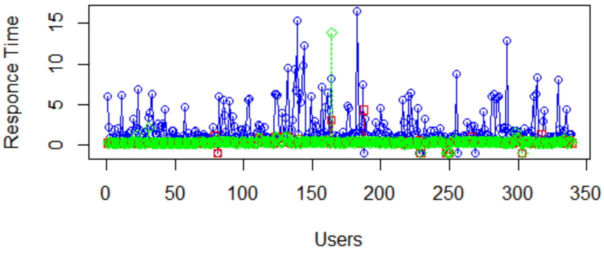

Fig 12: The response time of Dream set data of 335 users, and request and response time

Cumulative distribution of RT from different users QWS dataset is shown in Figure 13.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

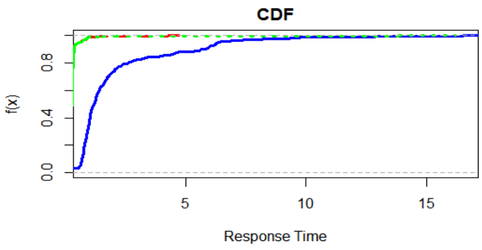

Fig 13: The cumulative RT from different users QWS dataset

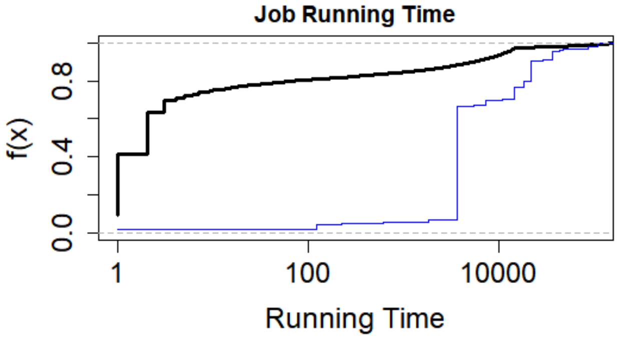

Fig 14: The Cloud -Grid; 10, 20,195 records, running and job running time.

The Figure 14. has cloud data set 1020195 records, the behavior of cloud waiting time and request time, in initial step the waiting time and RT is minimum as the requests are increasing that is more than 8000 requests then waiting time is also increased proportionally to the request time of applications.

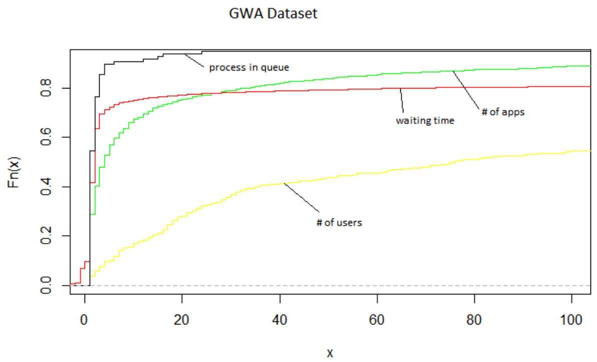

Fig 15: In X- Axis applications users, waiting time, applications and process queues, and Y Axis range of values.

The users increasing the use of applications waiting time increasing, and application usage is also increasing, the process queue waiting time reached to maximum, when running maximum utilization of resources with all the service users shown in figure.15 using Grid dataset [21]

VI CONCLUSION AND FUTURE SCOPE

The use of Web applications worldwide increasing a lot, the user demands to use applications at high quality including minimum response time, maximum availability by auto-scaling and FTS. Use of high-speed communications, due to delay and un-availability of services. The service provider of Web service will lose business opportunities. Service availability improves the reliability as the user Multiple Queue Shortest Remaining CPU algorithm which will reduce the Turnaround time(TA), waiting time(WT) and improves the availability comparing the other scheduling methods FCFS, SJF and Round robin is shown Figure 8. The availability of Service is improved is by auto-scaling and FTS is shown in Figure 10. The experimental results conducted using R Language on QWS dataset [26], dreamset data [25], and Grid data [21] the results are shown in graph figure 12. dream set, figure 13. QWS dataset. Figure.11, cloud dataset, figure 14, and figure.15 use of grid dataset. As recommends that the high availability by auto-scaling systems will improve performance by reducing Mean time to Failure and Repair is optimized and provide high availability. As long as jobs are increasing the load on the system waiting time is also increased, to minimization of the waiting time by proposed algorithm Optimized Multilevel Web service CPU Scheduling algorithm 1, and Model in figure 7.and figure 9. describe high availability and improve performance. In Future IoT base Quality control system for web based applications will definitely improve the overall quality using communication sensors, QoS manage and applications.

ACKNOWLEDGMENT.

Thanks to Dr. Eyhab Al-Masri for proving QWS dataset 2507 records, Thanks to Prof. Mingdong Tang, for providing WS-DREAM github.com and Planet Lab, thanks to Dr. Franck Cappello and to Dr. Olivier Richard for providing cloud dataset. And Thanks to Hui Li, DrorFeitelson, at Imperial College London, for providing Grid dataset by e-Science Group of HEP, Grid Workloads Archive

REFERENCES

1. Jianbin Wei, and Cheng-ZhongXu,"Measuring Client-Perceived Page view Response Time of Internet Services",pp.773-785 IEEE Transactions on Parallel and Distributed Systems, Vol. 22, No. 5,(2011)
2. ZujieRen, Jian Wan, Weisong Shi, XianghuaXu, and Min Zhou,"Workload Analysis, Implications, and Optimization on a Production Hadoop Cluster: A Case Study on Taobao", pp. 307-321, IEEE Transactions on Services Computing, Vol. 7, No. 2, (2014)
3. William Stallings, “High-speed Networks and Internets Performance and Quality of Service”, pp.183-247, Pearson Education Publishers, (2002)
4. J. Zhu, Y. Kang, Z. Zheng and M. R. Lyu, "WSP: A Network Coordinate Based Web Service Positioning Framework for Response Time Prediction," pp. 90-97.doi: 10.1109/ICWS.2012.81, IEEE,19th International Conference on Web Services, Honolulu, (2012)
5. A. E. Yilmaz and P. Karagoz, "Improved Genetic Algorithm Based Approach for QoS Aware Web Service Composition,", pp. 463-470.doi: 10.1109/ICWS.2014.72, IEEE International Conference on Web Services, Anchorage, AK, (2014)
6. Balazs Simon, Balazs Goldschmidt, and KarolyKondorosi," A Performance Model for the Web Service Protocol Stacks", pp. 644-657, IEEE Transactions on Services Computing, Vol. 8, No. 5, (2015)
7. 7.Tarek F. Abdelzaher, Kang G. Shin, and Nina Bhatti,"Performance Guarantees for Web Server End- systems: A Control-Theoretical Approach", pp. 80-96, IEEE Transactions on Parallels and Distributed Systems, vol.13, No.1, (2002)
8. Chen Hou and Qianchuan Zhao, “Optimization of Web Service-Based Control System for Balance between Network Traffic and Delay”, pp. 1-11", IEEE Transactions on Automation Science and Engineering, (2017)

9. Song Wu, Like Zhou, Huahua Sun, Hai Jin, and Xuanhua Shi, "Poris: A Scheduler for Parallel Soft Real-Time Applications in Virtualized Environment” pp. 841-854, ", IEEE Transactions on Parallel and Distributed Systems, Vol. 27, No. 3, (2016)
10. Sajee Mathew, “Architecting for High Availability", pp. 1-111, AWS Summit 2013 Navigating the Cloud, (2013)
11. Kranti Pore, “How to Achieve Website High Availability in a Distributed Enterprise Environment", http://www.bitwiseglobal.com/blogs/website-high-availability-in-distributed-enterprise-environment/
12. P. M. Melliar-Smith and L. E. Moser, "Conversion Infrastructure for Maintaining High Availability of Web Services Using Multiple Service Provider spp. 759-764.doi: 10.1109/ICWS.2015.110" 2015 IEEE International Conference on Web Services, New York, NY, (2015)
13. 13.http://blog.fosketts.net/2011/07/06/defining-failure-mttr-mttf-mtbf/
14. 14. Andrew S. Tanenbaum, “Modern Operating Systems”, pp. 71-151, Prentice Hall India, 2nd Edition,(2001)
15. 15.Achyut S Godbole, “Operating Systems”, pp.404-420, 2nd Edition Tata McGraw Hill Publishers, (2005)
16. 16. D M Dhamdhere, “Operating Systems: A concept based Approach”, pp.339-735, Tata McGraw Hill publications, (2002)
17. 17. Gary Nutt, NabenduChaki, and SarmsisthaNeogy, “Operating Systems”, pp. 42- 54, 3rd Edition, Pearson Publications, (2004)
18. 18. Parag K. Lala, “Fault-Tolerant and Fault Testable Hardware Design”, BS Publications (2002), pp. 1-11
19. 19. Dhananjay M. Dhamdhere, “Operating systems: A Concept-based approach”, McGraw Hill Education publishers,3rd Edition, (2009), pp. 760-783
20. 20.Andrews S. Tanenbaum, Herbert Bos, “Modern Operating Systems”, Pearson publishers(2016), pp. 148-165
21. 21. Grid dataset http://gwa.ewi.tudelft.nl/
22. 22. M. Swami Das, A. Govardhan, and D. Vijaya Lakshmi. 2015. QoS of Web Services Architecture. In Proceedings of the International Conference on Engineering & MIS 2015 (ICEMIS '15). ACM, New York, NY, USA, article 66, pp. 1-8
23. 23. M. Swami Das, A. Govardhan, and D. Vijaya Lakshmi. Best practices for web applications to improve performance of QoS. In Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies (ICTCS '16). ACM (2016), NewYork, NY, USA, Article123, pp.1-9
24. 24. Marc Oriol, Jordi Marco, and Xavier Franch, "Quality models for web services: A systematic mapping ", Information and Software Technology, (2014), pp.1-16
25. 25. https://github.com/wsdream/wsdream-dataset
26. 26. QWS Data set http://www.uoguelph.ca/~qmahmoud/qws/