International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 24 Issue 11 – AUGUST 2017.

An Enhanced Load Balancing Strategy in Cloud with Energy Constraints

SAMSANI VENKATA SIVA NARAYANA[#1] and T.NAGARAJU[*2]

[#]*PG Scholar, Kakinada Institute Of Engineering & Technology Department of Computer Science & Engineering, JNTUK,A.P, India.*

[*] *Assistant Prof, Dept of Computer science& Engineering, Kakinada Institute of Engineering & Technology, JNTUK, A.P, INDIA.*

***Abstract*— In this paper we enhance the existing Cloud Load balancing Strategy and provide an efficient energy saving concept for the Cloud Servers by considering the Energy factor. A system uses a significant amount of energy even when idle or lightly loaded. A widely reported solution to resource management in large data centers is to concentrate the load on a subset of servers and, whenever possible, switch the rest of the servers to one of the possible sleep states. We propose a reformulation of the traditional concept of load balancing aiming to optimize the energy consumption of a large-scale system: distribute the workload evenly to the smallest set of servers operating at a n optimal energy level, while observing QoS constraints, such as the response time. Our model applies to clustered systems; the model also requires that the demand for system resources to increase at a bounded rate in each reallocation interval. In this paper we report the VM migration costs for application scaling.**

Index Terms*— load balancing, application scaling, idle servers, server consolidation, energy proportional systems.

I. Introduction

Cloud computing appears to be a highly disruptive technology, which is gaining momentum. It has inherited legacy technology as well as new ideas on large-scale distributed systems. The concept of cloud computing addresses the next evolutionary step of distributed computing. The goal of this computing model is to make a better use of distributed resources, put them together in order to achieve higher throughput, and be able to tackle large-scale computation problems. The computing power nowadays is easily available for massive computational processing [1].

The concept of “load balancing” dates back to the time the first distributed computing systems were implemented in the late 1970s and early 1980s. It means exactly what the name implies, to evenly distribute the workload to a set of servers to maximize the throughput, minimize the response time, and increase the system resilience to faults by avoiding overloading one or more systems in the distributed environment [2].

Load balancing is the key feature and one of the central issues in cloud computing. It is a technique in which distribution of the dynamic local workload equally across the different clusters in cloud in order to avoid the situation where few nodes are overloaded while few are remains idle. Cloud computing is a client-server architecture composed by large and power-consuming data centers designed to support the efficient and scalable output required by consumers [3]. The cloud computing having a on demand service, broad network access, resource pooling, rapid elasticity, measured service are the essential characteristics. The dynamic workload results some systems overload and some systems remains unused, therefore it is necessary to distribute this load efficiency for preventing such odd resource utilization. Hence the concept load balance comes into existence that can distribute load across different computer clusters, network links, disk driver and some other resources to improve the throughput and optimal resource utilization, avoid overloading and minimizing response time [4].

Scaling is the method that allocates additional resources to a cloud application in response to a request consistent with the SLA. We can distinguish two scaling methods, i.e. Horizontal and Vertical scaling. Horizontal scaling is the most common method of scaling on a cloud system; it can increase the number of Virtual Machines (VMs) when the load of nodes increases and reduce the number when the load decreases. Large farms of computing and storage platforms have been assembled and a fair number of Cloud Service Providers (CSPs) offer computing and storage services based on three different delivery models SaaS (Software as a Service), PaaS (Platform as a Service), and IaaS (Infrastructure as a Service). Reduction of energy consumption thus, of the carbon footprint of cloud related activities, is increasingly more important for the society [5]. Indeed, as more and more applications run on clouds, more energy is required to support cloud computing than the energy required for many other human related activities.

While most of the energy used by data centers is directly related to cloud computing, a significant fraction is also used by the networking infrastructure used to access the cloud. This fraction is increasing, as wireless access becomes more popular and wireless communication is energy intensive. In this paper we are only concerned with a single aspect of energy optimization, minimizing the energy used by cloud servers [6].

The strategy for resource management in a computing cloud we discuss is to concentrate the load on a subset of servers and, whenever possible, switch the rest of the servers to a sleep state. In a sleep state the energy consumption is very low. This observation implies that the traditional concept of load balancing could be reformulated to optimize the energy consumption of a large-scale system as follows: dis- tribute evenly the workload to the smallest set of server's op- erating at an optimal energy level, while observing QoS con- straints, such as the response time. An optimal energy level is one when the normalized system performance, defined as the ratio of the current performance to the maximum performance, is delivered with the minimum normalized energy consumption, defined as the ratio of the current energy consumption to the maximal one [7].

II. LITERATURE SURVEY

A. Data replication and power consumption in data grids

While data grids can provide the ability to solve large-scale applications which require the processing of large amounts of data, they have been recognized as extremely energy inefficient. Computing elements can be located far away from the data storage elements. A common solution to improve availability and file access time in such environments is to replicate the data, resulting in the creation of copies of data files at many different sites. The energy efficiency of the data centers storing this data is one of the biggest issues in data intensive computing. Since power is needed to transmit, store and cool the data, we propose to minimize the amount of data transmitted and stored by utilizing smart replication strategies that are data aware. In this paper we present a new data replication approach, called the sliding window replica strategy (SWIN), that is not only data aware, but is also energy efficient. We measure the performance of SWIN and existing replica strategies on our Sage green cluster to study the power consumption of the strategies. Results from this study have implications beyond our cluster to the management of data in clouds [8].

B. Server workload analysis for power minimization using consolidation

Server consolidation has emerged as a promising technique to reduce the energy costs of a data center. In this work, we present the first detailed analysis of an enterprise server workload from the perspective of finding characteristics for consolidation. We observe significant potential for power savings if consolidation is performed using off-peak values for application demand. However, these savings come up with associated risks due to consolidation, particularly when the correlation between applications is not considered. We also investigate the stability in utilization trends for low-risk consolidation. Using the insights from the workload analysis, two new consolidation methods are designed that achieve significant power savings, while containing the performance risk of consolidation. We present an implementation of the methodologies in a consolidation planning tool and provide a comprehensive evaluation study of the proposed methodologies [9].

C. Performance and power management for cloud infrastructures

A key issue for Cloud Computing data-centers is to maximize their profits by minimizing power consumption and SLA violations of hosted applications. In this paper, we propose a resource management framework combining a utility-based dynamic Virtual Machine provisioning manager and a dynamic VM placement manager. Both problems are modeled as constraint satisfaction problems. The VM provisioning process aims at maximizing a global utility capturing both the performance of the hosted applications with regard to their SLAs and the energy-related operational cost of the cloud computing infrastructure. We show several experiments how our system can be controlled through high level handles to make different trade-off between application performance and energy consumption or to arbitrate resource allocations in case of contention [10].

D. Elastic- Tree: saving energy in data center networks

Networks are a shared resource connecting critical IT infrastructure, and the general practice is to always leave them on. Yet, meaningful energy savings can result from improving a network's ability to scale up and down, as traffic demands ebb and flow. We present ElasticTree, a network-wide power1 manager, which dynamically adjusts the set of active network elements -- links and switches--to satisfy changing data center traffic loads. We first compare multiple strategies for finding minimum-power network subsets across a range of traffic patterns. We implement and analyze ElasticTree on a prototype testbed built with production OpenFlow switches from three network vendors. Further, we examine the trade-offs between energy efficiency, performance and robustness, with real traces from a production e-commerce website. Our results demonstrate that for data center workloads, ElasticTree can save up to 50% of network energy, while maintaining the ability to handle traffic surges. Our fast heuristic for computing network subsets enables ElasticTree to scale to data centers containing thousands of nodes. We finish by showing how a network admin might configure Elastic Tree to satisfy their needs for performance and fault tolerance, while minimizing their network power bill [11].

E. Energy-efficient protocol for cooperative networks

In cooperative networks, transmitting and receiving nodes recruit neighboring nodes to assist in communication. We model a cooperative transmission link in wireless networks as a transmitter cluster and a receiver cluster. We then propose a cooperative communication protocol for establishment of these clusters and for cooperative transmission of data. We derive the upper bound of the capacity of the protocol, and we analyze the end-to-end robustness of the protocol to data-packet loss, along with the tradeoff between energy consumption and error rate. The analysis results are used to compare the energy savings and the end-to-end robustness of our protocol with two non-cooperative schemes, as well as to another cooperative protocol published in the technical literature. The comparison results show that, when nodes are positioned on a grid, there is a reduction in the probability of packet delivery failure by two orders of magnitude for the values of parameters considered. Up to 80% in energy savings can be achieved for a grid topology, while for random node placement our cooperative protocol can save up to 40% in energy consumption relative to the other protocols. The reduction in error rate and the energy savings translate into increased lifetime of cooperative sensor networks [12].

III. EXISTING SYSTEM

An important strategy for energy reduction is concentrating the load on a subset of servers and, whenever possible, switching the rest of them to a state with low energy consumption [13]. This observation implies that the traditional concept of load balancing in a large-scale system could be reformulated as follows: distribute evenly the workload to the smallest set of servers operating at optimal or near-optimal energy levels, while observing the Service Level Agreement (SLA) between the CSP and a cloud user. An optimal energy level is one when the performance per Watt of power is maximized [14].

In order to integrate business requirements and application level needs, in terms of Quality of Service (QoS), cloud service provisioning is regulated by Service Level Agreements (SLAs): contracts between clients and providers that express the price for a service, the QoS levels required during the service provisioning, and the penalties associated with the SLA violations. In such a context, performance evaluation plays a key role allowing system managers to evaluate the effects of different resource management strategies on the data center functioning and to predict the corresponding costs/benefits [15].

A. DISADVANTAGES OF EXISTING SYSTEM:

On-the-field experiments are mainly focused on the offered QoS, they are based on a black box approach that makes difficult to correlate obtained data to the internal resource management strategies implemented by the system provider [16]. Simulation does not allow conducting comprehensive analyses of the system performance due to the great number of parameters that have to be investigated.

IV. PROPOSED SYSTEM

There are three main contributions of this paper that follow:-

1) A new model of cloud servers that is based on different operating regimes with various degrees of energy efficiency (i.e. processing power versus energy consumption).

2) A novel algorithm that can perform load balancing and application scaling for maximizing the number of servers operating in the energy-optimal regime and comparison of techniques analysis for load balancing and application scaling using three different sizes of clusters and two different average load profiles.

3) The objective of the algorithms is to ensure that the largest possible number of active servers operate with their respective optimal operating regime.

A. PROPOSED APPROACH

A new model of cloud servers that is based on different operating regimes with various degrees of \energy efficiency" (processing power versus energy consumption); A novel algorithm that performs load balancing and application scaling to maximize the number of servers operating in the energy-optimal regime; and analysis and comparison of techniques for load balancing and application scaling using three differently-sized clusters and two different average load profiles. The objective of the algorithms is to ensure that the largest possible number of active servers operate within the boundaries of their respective optimal operating regime.

After load balancing, the number of servers in the optimal regime increases from 0 to about 60% and a fair number of servers are switched to the sleep state. There is a balance between computational efficiency and SLA violations; the algorithm can be tuned to maximize computational efficiency or to minimize SLA violations according to the type of workload and the system management policies.

1) System Architecture

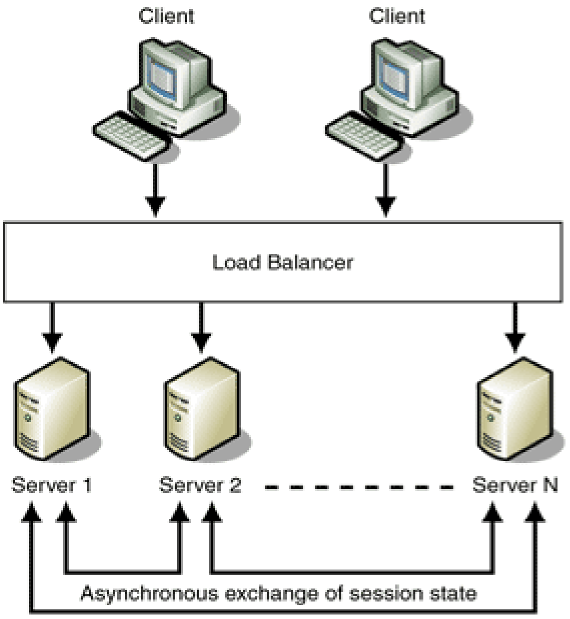


Fig 1: Load Balancing Architecture.

B. SYSTEM MODULES

1) Load Balancing in Cloud Computing

Cloud computing involves web services, visualization, software, networking and distributing computing. Cloud can have several elements such as client data center and multiple distributed servers. It includes on-demand service, high availability, fault tolerance, flexibility, scalability, reduced cost for owners, reduced overhead for clients or users etc. The effective establishment of load balancing algorithm can solve such a problem; the load is basically a CPU utilization, memory capacity, request delay or load in the network.

Load balancing is the technique that can distribute the load among the various nodes of different distributed system for improving the job response time and resource utilization that can avoid the situation where some nodes are heavily loaded and others are idle or in a sleep state. All processor every node or system in the network does the same amount of work at any specific time. The main aim is to use effective load balancing algorithm that can minimize the latency and maximize the throughput on the cloud environment.

2) *Energy Efficiency of a System*

We can measure the energy efficiency of the different system as a ratio of performance per watt of power. From few decades performance of computing system is increased much more rapidly than its energy efficiency. The idle or lightly loaded system consumes less energy or should be nearer to zero and it increased linearly with the system workload. But practically the idle system can consume more than half energy of full load system. We have an optimal energy consumption regime that is far from the typical operating regime of data center servers. When energy proportional system is in idle state is consumes no energy or very little energy and increases gradually as load increases.

3) *Resource management policies for large-scale data centers*

These policies can be loosely grouped into five classes

(i) Admission control
(ii) Capacity allocation
(iii) Load balancing
(iv) Energy optimization and
(v) Quality of service (QoS) guarantees.

To prevent the system from accepting the workload in violation of high level system policies is explicit goal of an admission control policy; system not accepting the additional workload that can prevent it from completing the work that is already done or in progress We have a knowledge of the global state of the system for limiting the workload. In a dynamic system, this knowledge is when available, is at best case. Allocating the resources for individual instances is known as Capacity allocation.

4) *Server Consolidation*

We can measure the energy efficiency of the different system as a ratio of performance per watt of power. From few decades performance of computing system is increased much more rapidly than its energy efficiency. The idle or lightly loaded system consumes less energy or should be nearer to zero and it increased linearly with the system workload. But practically the idle system can consume more than half energy of full load system. We have an optimal energy consumption regime that is far from the typical operating regime of data center servers. When energy proportional system is in idle state is consumes no energy or very little energy and increases gradually as load increases.

5) *Energy-aware Scaling Algorithms*

The main aim of this algorithm is to ensure that the numbers of active server out of all are operating with the optimal operating regime.

The actions implementing this policy are

(a) Migration of VMs from a server operating in the low regime and then it switch the server to a sleep state system.

(b) Switching idle servers to a sleep state system and reactivate servers in a sleep state when the data center load increases;

(c) Migration of VMs from an overloaded server, a server operating in the high regime with applications predicted to increases their demands for computing.

6) *Proposed Algorithm*

Input:

Workload (W) ->W1, W2, W3.....
Resource (R) -> R1, R2, R3...

Output:

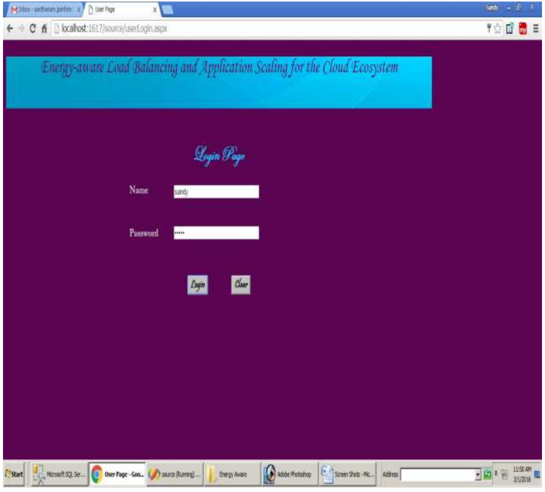
Migration List (M) -> M1, M2, M3...

Energy Efficiency Algorithm

```
1) Start
2) Extract Total workload list
W (Z) -> W1, W2, W3....Wn
3) Access total Resource list
R (Z) -> R1, R2, R3.....Rn
4) User uploads file
U (Z): F (Z) -> Un, Fn
5) Check first cloud server workload
6) Limitation of server depends on energy level
7) If server is going to energy threshold
8) File migrates to another server->HOT SPOT Process.
9) Check remaining server workload.
10) Find Min (workload Resources) ->optimization
11) Manage workload of every server ->Green Computing.
12) Check energy level
13) End.
```

C. *IMPLEMENTATION*

The results of the proposed system are shown below.



localhost:1617/source/userLogin.aspx

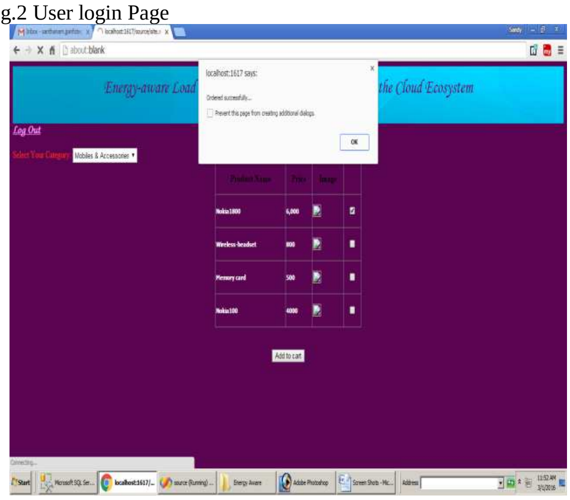
Energy-aware Load Balancing and Application Scaling for the Cloud Ecosystem

Login Page

Name: sandy

Password: •••••

Fig.2 User login Page



about:blank

Energy-aware Load ... the Cloud Ecosystem

localhost:1617 says:

Ordered successfully...

□ Prevent this page from creating additional dialogs

Select Your Category: Mobiles & Accessories

Product Name	Price	Image	
Nokia1800	6,000		☑
Wireless-headset	800		□
Memory card	500		□
Nokia100	4000		□

Fig.3 Order Confirmation

V. CONCLUSION

The average server utilization in large data-centers is 18% [21]. When idle the servers of a data center use more than half the power they use at full load. The alternative to the wasteful resource management policy when the servers are always on, regardless of their load, is to develop energy-aware load balancing policies. Such policies combine dynamic power management with load balancing. There are ample opportunities to reduce the energy necessary to power the servers of a large-scale data center and shrink the carbon footprint of cloud computing activities, even though this is only a fraction of the total energy required by the ever increasing appetite for computing and storage services. To optimize the resource management of large farms of servers we redefine the concept of load balancing and exploit the technological advances and the power management functions of individual servers.

In the process of balancing the load we concentrate it on a subset of servers and, whenever possible, switch the rest of the servers to a sleep state. From the large number of questions posed by energy-aware load balancing policies we discuss only the energy costs for migrating a VM when we decide to either switch a server to a sleep state or force it to operate within the boundaries of an energy optimal regime. The policies analyzed in this paper aim to keep the servers of a cluster within the boundaries of the optimal operating regime. After migrating the VMs to other servers identified by the cluster leader, a lightly loaded server is switched to one of the sleep states. There are multiple sleep states; the higher the state number, the larger the energy saved, and the longer the time for the CPU to return to the state C0 which corresponds to a fully operational system. For simplicity we chose only two sleep states C3 and C6 in the simulation. If the overall load of the cluster is more than 60% of the cluster capacity we do not switch any server to a C6 state because in the next future the probability that the system will require additional computing cycles is high. Switching from the C6 state to C0 requires more energy and takes more time.

On the other hand, when the total cluster load is less than 60% of its capacity we switch to C6 because it is so unlikely that for the next interval and the interval after that system needs extra computational unit. The simulation results reported in Section 5 show that the load balancing algorithms are effective and that low-cost vertical scaling occurs even when a cluster operates under a heavy load. The larger the cluster sizes the lower the ratio of high cost in-cluster versus low-cost local decisions. The QoS requirements for the three cloud delivery models are different thus, the mechanisms to implement a cloud resource management policy based on this idea should be different. To guarantee real-time performance or a short response time, the servers supporting SaaS applications such as data streaming or on-line transaction processing (OLTEP) may be required to operate within the boundaries of a sub-optimal region in terms of energy consumption.

There are cases when the instantaneous demand for resources cannot be accurately predicted and systems are forced to operate in a non-optimal region before additional systems can be switched from a sleep state to an active one. Typically, PaaS applications run for extended periods of time and the smallest set of serves operating at an optimal power level to guarantee the required turnaround time can be determined accurately. This is also true for many IaaS applications in the area of computational science and engineering. There is always a price to pay for an additional functionality of a system, so the future work should evaluate the overhead and the limitations of the algorithms required by these mechanisms.

REFERENCES

[1] Shunmei Meng, Wanchun Dou, Xuyun Zhang, Jinjun Chen, "KASR: A Keyword Aware Service Recommendation Method on MapReduce for Big Data Applications," IEEE Transactions On Parallel and distributed system, TPDS-12-1141-2013.

[2] J. Baliga, R.W.A. Ayre, K. Hinton, and R.S. Tucker." Green cloud computing: balancing energy in processing, storage, and transport." Proc. IEEE, 99(1):149-167,2011.

[3] A. Beloglazov, R. Buyya "Energy efficient resource management in virtualized cloud data centers." Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Comp., 2010.

[4] A. Beloglazov and R. Buyya. "Managing overloaded hosts for dynamic consolidation on virtual machines in cloud centers under quality of service constraints." IEEE Trans. on Parallel and Distributed Systems, 24(7):1366- 1379, 2013.

[5] M. Elhawary and Z. J. Haas. "Energy-e_cient protocol for cooperative networks." IEEE/ACM Trans. on Net- working, 19(2):561{574, 2011.

[6] A. Gandhi, M. Harchol-Balter, R. Raghunathan, and M.Kozuch. "AutoScale: dynamic, robust capacity management for multi-tier data centers." ACM Trans. On Computer Systems, 30(4):1{26, 2012.

[7] A. Gandhi, M. Harchol-Balter, R. Raghunathan, and M.Kozuch. "Are sleep states effeective in data centers?" Proc. Int. Conf. on Green Comp., pp. 1{10, 2012.

[8] D. Gmach, J. Rolia, L. Cherkasova, G. Belrose, T. Tucricchi, and A. Kemper. "An integrated approach to resource pool management: policies, efficiency, and quality metrics." Proc. Int. Conf. on Dependable Systems and Networks, pp. 326{335, 2008.

[9] V. Gupta and M. Harchol-Balter. "Self-adaptive admission control policies for resource-sharing systems. " Proc. 11th Int. Joint Conf. Measurement and Modeling Computer Systems (SIGMETRICS'09), pp. 311{322, 2009.

[10] B. Heller, S. Seetharaman, P. Mahadevan, Y. Yiakoumis, P. Sharma, S. Banerjee, and N. McKeown. "Elastic- Tree: saving energy in data center networks." Proc. 7th USENIX Conf. on Networked Systems Design and Im- plementation, pp. 17{17, 2011.

[11] E. Le Sueur and G. Heiser. "Dynamic voltage and frequency scaling: the laws of diminishing returns." Proc. Workshop on Power Aware Computing and Systems, HotPower'10, pp. 2{5, 2010.

[12] D. C. Marinescu, A Paya, and J.P. Morrison. \Coalition formation and combinatorial auctions; applications to self-organization and self-management in utility computing." http: //arxiv.org/pdf/1406.7487v1.pdf, 2014.

[13] A. Paya and D. C. Marinescu. "Energy-aware load balancing policies for the cloud ecosystem." http: //arxiv.org/pdf/1307.3306v1.pdf, December 2013.

[14] B. Urgaonkar and C. Chandra. „Dynamic provisioning of multi-tier Internet applications." Proc. 2nd Int. Conf, on Automatic Comp., pp. 217{228, 2005.

[15] H. N. Van, F. D. Tran, and J.-M. Menaud. "Performance and power management for cloud infrastructures." Proc. IEEE 3rd Int. Conf. on Cloud Comp., pp. 329{336, 2010.

[16] S. V. Vrbsky, M. Lei, K. Smith, and J. Byrd. „Data replication and power consumption in data grids." Proc IEEE 2nd Int. Conf. on Cloud Computing Technology and Science, pp. 288{295, 2010.

AUTHOR PROFILE

SAMSANI VENKATA SIVA NARAYANA, is a student of Kakinada Institute Of Engineering & Technology affiliated to JNTUK, Kakinada pursuing M.Tech (Computer Science). His Area of interest includes Cloud Computing and its objectives in all current trends and techniques in Computer Science.

T.NAGARAJU, M.TECH, is working as Assistant Professor, Department of Computer science & Engineering, Kakinada Institute of Engineering & Technology, JNTUK, A.P, INDIA.