Implementation of multiple operating systems in visualized platform for hybrid Cloud computing

N.STAALINPRASANNAH Final ME., CSE

Mahendra Institute of Engineering & Technology Mahendrapuri, Thiruchengode stalinmecse@gmail.com S.SURIYA II MCA Vivekanandha Business School for Women Elayampalayam, Tiruchengode ssuriyamca@gmail.com 8807520734 T.RANJITHA II MCA Vivekanandha Business School for Women Elayampalayam, Tiruchengode ranji.thirumca@gmail.com 9787117477

ABSTRACT

9894912923

Virtualization is critical to cloud computing because it simplifies the delivery of services by providing a platform for optimizing complex IT resources in a high performance manner, which is what makes cloud computing so cost effective. Virtualization can be applied very broadly to just about everything you can imagine including virtual memory, virtual networks, virtual storage, hardware, virtual operating systems and virtual applications.

Keywords

Virtualization, Hybrid computing, Vmware, ESX, ESXi

1. INTRODUCTION

Cloud Computing [1-3] is becoming an increasingly popular enterprise model in which computing resources are made available on-demand to the user as needed. The unique value proposition of Cloud Computing creates new opportunities to align IT and business goals. Cloud computing is essentially a powerful computing paradigm in which tasks are assigned to a combination of connections, software and services accessed over a network. The vast processing power of Cloud Computing is made possible though distributed, large-scale computing clusters, often in concert with server virtualization software, like VMware ESX Server [4] and Xen [5], and parallel processing. This network of servers and connections is collectively known as the Cloud. The computing resources may be maintained within the client enterprise, or made available by a service provider. Computing at the scale of the Cloud allows users to access supercomputer-level computing power. Users can access the enormous and elastic resources whenever they need them. For this reason, Cloud Computing is also described as on-demand computing.

The on-demand model was developed to overcome the common challenge to an enterprise of being able to meet fluctuating demands efficiently. The model of Cloud Computing has evolved from following concepts of utility computing, autonomic computing, grid computing, and software as a service (SaaS) [6]. Utility computing, for example, is an on-demand approach that combines outsourced computing resources and infrastructure management with a usagebased payment structure. Because an enterprise demand on computing resources can vary drastically from one time to another, maintaining sufficient resources to meet peak requirements can be costly. Conversely, if the enterprise cuts costs by maintaining only minimal computing resources, there may not be sufficient resources to meet peak requirements.

On the other hand, scalability or ability to expand and add resources dynamically is critical to the success of many enterprises currently involved in doing business on the Web and in providing information that may suddenly become heavily demanded. While there are many strategies that IT organizations can undertake to service more customer demands, the way they are designed and implemented can make or break these businesses. Cloud Computing offers a powerful environment to scale web applications without difficulty. In fact, Cloud Computing can provide different resources on-demand for many of the typical scaling points that a web application needs including servers, storage and networking. The on-demand nature of Cloud Computing combined with the pay-as-you-go model means that as the application demand grows, so can the resources required to service that demand. In this situation, the capacity equals exactly the demand as long as the application is designed properly and its architecture is amenable to scaling well.

In this paper, we will present a dynamic scaling scenario with novel design of web applications deployed in virtual machine instances that are dynamically installed on a Cloud to illustrate the powerful scaling capabilities of the Cloud Computing environment. We will first describe Cloud Computing and the various scaling indicators for scalable web applications. We then discuss the scalability capabilities of the Cloud and its use of virtualization technologies. We will then present our novel architecture design of a scaling scenario with an on-line application and a dynamic scaling algorithm based on number of login users to the web application.

The outline of the rest of the paper is organized as follows. Section 1 gives an introduction. Section 2 describes the cloud computing architecture and virtualization. Section 3 illustrates the architecture design of a scaling scenario for an on-line web application. A scaling algorithm based on number of login users will be described. Related work will be discussed in Section 4. Finally, Section 5 concludes the paper.

2. CLOUD COMPUTING, VIRTUALIZATION AND SCALABILITY OF SERVICES

Cloud Computing provides the ability to add capacity as needed, typically with very small lead times. Clearly, Cloud Computing provides a new compelling mechanism for dealing with application services that need to be scalable. A brief introduction of Cloud Computing, its key delivery technology of virtualization, the scalability and the scaling indicators of web applications in a Cloud will be given in the following sections.

2.1 Cloud Computing

Cloud Computing is a way to deliver services over New advances in virtualization the network. technology [7, 8], processors, disk storage, broadband Internet access and fast, inexpensive and powerful servers have all combined to make Cloud Computing a realistic and compelling paradigm. Cloud Computing basically use virtualization technique to turn computer resources into virtual guest machines. These virtual machines usually reside on some networked physical servers in a hosting environment. However, the virtual guest machines can be moved around, thus breaking direct hardware dependency associated with physical machines. With hardware dependency no longer an issue, the guest system can be insulated from hardware breakdowns, and be automatically moved to another piece of available hardware. A typical Cloud Computing environment with automated provisioning capabilities, advanced virtualization technologies, and virtual machines hosting on physical servers offering Cloud application services to users directly is illustrated in Fig. 1.

In Cloud Computing, new applications are made available by highly efficient virtualized computing resources that can be rapidly scaled up and down in a secure way to deliver a high quality of service. Users can gain access to their applications from anywhere through their connected devices. The user sees only the service and not the implementation or infrastructure required for its delivery. Examples of Cloud services include technology services such as storage, data protection, applications, business processes and even business and consumer services such as email and office applications. Cloud Computing allows users and companies to use the services and storage that they need, when they need them and, as wireless broadband connection options grow, where they need them. Customers can be billed based upon server utilization, processing power used or bandwidth consumed.

Figure 1. Typical Cloud Computing Environment.

As a result, Cloud Computing has the potential to overturn the software industry entirely, as applications are purchased, licensed and run over the network instead of a user desktop. This shift will put data centers and their administrators at the center of the distributed network, as processing power, bandwidth and storage are all managed remotely. The following are typical types of Cloud Computing services depending on nature of offerings:

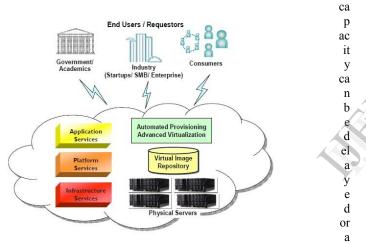
- **2.1.1.** Application Services Any web application is a Cloud application service in the sense that it resides in the Cloud. Google, Amazon, Facebook, Twitter, Flickr, and virtually every other Web 2.0 application is a Cloud application in this sense.
- **2.1.2. Platform Services** One step up from pure utility computing are Cloud platform services like Google Apps and Google Apps Engine, and Salesforce's force.com, which hide virtual machine instances behind higher-level APIs.
- **2.1.3.** Infrastructure Services Amazon Elastic Compute Cloud (EC2) is a typical Cloud infrastructure service which provides raw virtual machine instances, storage, and computation at pay-as-you-go utility pricing, and is currently the leading provider in this category. Developers are the typical target of this kind of Cloud Computing services.

2.2. Virtualization Technology

Virtualization technology [7, 8] is the key delivery technology in Cloud Computing. In a Cloud, virtualization refers primarily to platform virtualization or the abstraction of physical IT resources from the users and applications using them. Virtualization allows servers, storage devices, and other hardware to be treated as a pool of resources rather than discrete systems, so that these resources can be allocated on demand. Platform virtualization is performed on a given hardware platform by a control software, called a hypervisor or virtual machine monitor. This software creates a simulated computer environment, called a virtual machine, for its guest software. The guest software, which is often itself a complete operating system, runs just as if it were installed on a stand-alone hardware platform. The current leading virtualization and software providers include VMware [4], Xen [5], KVM [9], Force.com [10] and Microsoft Virtualization [11].

As a means of encapsulation of physical resources, virtualization solves several core challenges of datacenter managers and delivers specific advantages, including:

2.2.1. Higher Utilization Rates: Through virtualization, workloads can be encapsulated and transferred to idle or underused systems. This means that existing systems can be consolidated, so purchasing additional server



voided.

- **Consolidation:** 2.2.2. Resource Virtualization allows for consolidation of multiple IT Beyond server storage resources. and consolidation, virtualization provides an opportunity to consolidate the systems architecture, application infrastructure, data and databases, interfaces, networks, desktops, and even business processes, resulting in cost savings and greater efficiency.
- **2.2.3.** Lower Power Usage/Costs: Using virtualization to consolidate makes it possible to cut total power consumption and save significant costs.
- **2.2.4. Space Savings:** Server sprawl remains a serious problem in most datacenters, but datacenter expansion is not always an option, with expensive building costs and cooling costs. Virtualization can alleviate the strain by

consolidating many virtual systems onto fewer physical systems.

2.3. Scalability of Applications

Scalability is critical to the success of many organizations currently involved in doing business on the web and in providing information that may suddenly become heavily demanded. Scalability is a measure of the ability of an application to expand to meet enterprise business needs.

Generally, enterprise can scale a given application by adding more or larger resources when needed. Resources can be many things including servers, processors, storage, and networking bandwidth. Scalable applications are able to operate normally as they grow and can have more resources added at any time to service more customer demands. Applications that are not scalable may encounter performance and service availability problems as demand increases. These kinds of non-scalable applications may not be able to take advantage of more resources.

2.4. Scaling Indicators

Although different web applications may not perform in the same way, there are very common scaling points where resources become constrained. For example, it is very unlikely that a web application can grow to meet business needs on a single server. At some point in time, it is required to add more servers to meet increasing demands in order to meet service quality requirements. The same is generally true for storage, networking, and other scaling point.

In order to scale the application in dynamic manner, it is common to use scaling indicators at the scaling point to monitor and track the performance. For web applications, typical scaling indicators at the web server may include:

- Number of concurrent users.
- Number of active connections.
- Number of requests per second.
- Average response times per request.

Once a scaling indicator is selected for use to scale the application, samplings of the scaling indicator are collected in real-time, and statistics is calculated periodically. Based on the historical trends and predictions derived from the statistics of the scaling indicator, scaling rules can be defined to scale up or down the amount of web application instances.

3. SCALING WEB APPLICATIONS IN A CLOUD

In the following sections, we will present our scaling scenario, the novel architecture design, and the dynamic scaling algorithm for scaling a web application installed in virtual machine on a Cloud. The scenario is based on using a front-end load balancer to dynamically route user requests to back-end web servers that host the web application. The number of web servers should automatically scale according to the threshold on the number of current active sessions in each web server instance to maintain service quality requirements.

3.1. Architecture Design

We consider a scenario to provide quality of services on a given web application over the Internet at any time since there may have potentially unlimited number of users that could access the application unpredictably. This type of workload usually demands a short response time and high level of reliability and availability from the application.

Thus, the web application should be made available without downtime and should provide the fastest response time regardless the amount of concurrent users accessing the system.

The main problem with such web applications is the inability to plan ahead or even predict the amount of users that will be accessing. A solution is to scale the web application in a dynamic manner and let the amount of web servers and web application components to grow (or shrink) on demand. A scalable architecture which effectively deals with this scenario is illustrated in Fig. 2 using a virtualized Cloud Computing environment. The architecture design includes a frontend load-balancer, a number of web app virtual machines, a Provisioning sub-system, and a Service Monitor sub-system with a dynamic scaling algorithm.

To control the action of provisioning or deprovisioning web server virtual machine instances, a dynamic scaling algorithm based on relevant threshold or scaling indicator of the web application is developed. The scaling indicator that we select here is the number of active sessions or logon sessions in each web application. This scaling indicator is found to be critical to the performance of our web application deployed on an Apache HTTP server in a Linux operating system according to our test results of a performance and scalability study. In the study, we have discovered that the web application can only support up to about 40,000 active, concurrent sessions, and when the active sessions increase beyond this threshold number, the web server system starts failing to respond to user requests and then freezes up eventually.

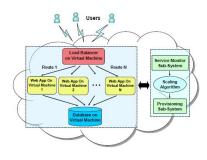


Figure 2., Architecture to Scale Web Applications in a Cloud.

3.2. Load-Balancer and Web-App Virtual Appliances

Use of a load balancer such as Apache HTTP Load Balancer will allow incoming HTTP request to be routed into web servers that host the web application. Since the Apache HTTP Load-Balancer configuration can be updated while the system is running, this allows the system to automatically and dynamically add new web server instances. The additional web servers enable the system to scale and thus provide better response time for incoming HTTP requests.

To simplify provisioning of additional resources, the web application and its corresponding web server is pre-installed in a virtual machine appliance image template which is made available in the image repository of the Cloud. Upon system command, new instances of web applications can be readily provisioned and created on-demand using the corresponding appliance image template.

3.3. Service Monitor and Provisioning Subsystems

The Service Monitor sub-system is responsible in gathering individual scaling indicators from the web applications and then calculating their moving average. In our current design, numbers of active sessions from individual web applications are used as the scaling indicators. A monitoring agent is installed in each web application to track the number of active sessions and to forward the number periodically to the Service Monitor subsystem.

Based on the moving average of the scaling indicator, a dynamic scaling algorithm, to be given in next section, is used to trigger a scaling event to the Provisioning subsystem. Depending on the updated statistics, action to scale up or down may be initiated. Scale up or down means that an event will be triggered instructing the Provisioning subsystem to start or shutdown web-server virtual machine instances running in the cloud.

The Provisioning sub-system is basically constructed from the IBM Tivoli Provisioning Manager (TPM) software product [12]. This TPM software component can be used in general to automate manual tasks of provisioning and configuring servers, operating systems, middleware, applications, storages and network devices in the server clusters.

3.4. Image-Based Provisioning

Image-based provisioning is a deployment and activation mechanism that clones a "golden" virtual image to create new virtual machine instances. One of the challenges with cloning virtual images is the handling of operating system, network, and application specific customization.

Automating the provisioning of new virtual machines from a "golden" image template [13] can be accomplished by adding automation capabilities into the template image, combined with external automation scripts that control the deployment.

In our work, we perform the automated imagebased provisioning of Linux-based virtual appliances based-on Xen hypervisor technology [5] in Red Hat Enterprise Linux system [13]. We utilize a simplified provisioning process that leverages the Linux disk image mount utility on the cloned Xen image and perform host name and network fixups before booting. The whole process can be summarized in two simple phases.

3.5. SCALING ALGORITHM

As mentioned previously, the scaling algorithm is implemented in the Service Monitor sub-system, and is used to control and trigger the scale-up or down in the Provisioning sub-system on the number of virtual machine instances based on the statistics of the scaling indicator. Our dynamic scaling algorithm, as given in Fig. 3 in pseudo procedures, is based on the scaling indicator Ai in each virtual machine instance in the Cloud. For the sake of illustration, we choose a scaling indicator in our implementation that corresponds to the number of active sessions in the web application of each instance.

The algorithm first determines the current web application instances with active sessions above or below given threshold numbers. If all instances have active sessions above the given upper threshold, a new web application instance will be provisioned, started, and then added to the front-end load-balancer. If there are instances with active sessions below a given lower threshold and with at least one instance that has no active session, the idle instance will be removed from the load-balancer and be shut down from the system. In each case, the load factors for all active instances will be recalculated and then applied to the load-balancer to re-distribute the request workloads to each instance evenly.

Dynamic Scaling Algorithm for Virtual Machine Instances in the Cloud.

For an instance i in NInstance If (Ai/SMax >= TUpper) then Increment N Exceed If (Ai/SMax < TLower) then Increment N Below Record and sort all indexes J in ascending of A_i/S_{Max} If (NExceed == NInstance) then *Provision and start a new instance* Increment number of instances: NInstance

Add new instance to Load-Balancer If $(N_{Below} \ge 2)$ then Set m equal first index in J If $(\mathbf{A}_{\mathbf{m}} == 0)$ then Remove instance m from Load-Balancer Shutdown instance m Decrement number of instances: NInstance Decrement N_{Below} and remove index m from J If $(N_{Below} \ge 2)$ then Set n equal first index in J Remove instance **n** temporary Decrement number of instances: NInstance Set normalized load factor $L_n = 0$ For an instance **i** in **N** Instance Evaluate normalized load factor: Li = (1-Ai/SMax) / Sum k=1,Ninstance [1-Ak/SMax] Apply new load factors L_i to Load Balancer

where Ai: Number of active sessions in instance i S Max: Maximum sessions per instance (e.g.

40,000)**Tupper** : Session upper-threshold (e.g.80% or 0.8)

T Lower : Session lower-threshold (e.g. 60% or 0.6)

N_{Instance}: Number of existing instances

 N_{Exceed} : Number of instances exceeding session upper-threshold

 N_{Below} : Number of instances below session lowerthreshold

4. VMWARE ESXI FOR HYBRID COMPUTING

4.1 BENEFITS OF THE VMWARE ESXI HYPERVISOR ARCHITECTURE

The hypervisor architecture of VMware vSphere plays a critical role in the management of the virtual infrastructure. The introduction of the bare-metal ESX architecture in 2001 significantly enhanced performance and reliability, which in turn allowed customers to extend the benefits of virtualization to their mission-critical applications. The removal of the Linux based console operating system (COS or 'service console") with the new ESXi architecture represents a similar leap forward in reliability and virtualization management. Less than 5% of the size of ESX, the new vSphere ESXi architecture improves hypervisor management in the areas of security, deployment and configuration, and ongoing administration.

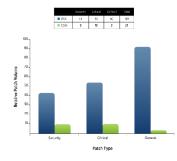


Figure 3: VMware ESX and ESXi

4.1.1. Improve Reliability and Security - The ESX architecture available in releases prior to vSphere 5.0 relied on a Linux-based COS for serviceability and agent-based partner integration. In the new, operatingsystem independent ESXi architecture. the approximately 2 GB COS has been removed and the necessary management functionality has been implemented directly in the core Kernel. Eliminating the COS drastically reduces the install footprint of the vSphere ESXi hypervisor to approximately 150 MB improving security and reliability by removing the security vulnerabilities associated with a general purpose operating system.

4.1.2. Streamline Deployment and Configuration. The new ESXi architecture has far fewer configuration items greatly simplifying deployment and configuration and making it easier to maintain consistency.

4.1.3 Reduce Management Overhead. The API-based partner integration model of the ESXi architecture eliminates the need to install and manage third party management agents. You can automate routine tasks by leveraging remote command line scripting environments such as vCLI or PowerCLI.

4.1.4 Simplify Hypervisor Patching and Updating. Due to its small size and limited components, the ESXi architecture requires far fewer patches than early versions, shortening service windows and reducing security vulnerabilities. Over its lifetime, the ESXi architecture requires approximately 10 times fewer patches than the ESX hypervisor running with the COS.

4.2. VMWARE ESXi for Hybrid Computing

4.2.1. VMware ESXi is Free

I have worked for SMBs before and, typically, the #1 factor when making IT decisions is cost. Fortunately, VMware ESXi is Free so that removes much of the cost concerns. However, you should be aware of the cost to upgrade to the various levels of the vSphere suite and budget for that cost in your IT budget in the future. This is true, based on my experience, because your virtualization project will be a success and the demand for new servers and applications will just continue to grow (not shrink).

4.2.2. Easy Upgrade to VMware vSphere

VMware ESXi is the hypervisor of VMware's vSphere and vSphere is, undeniably, the highest performance and most full-featured virtualization platform available today. VMware has been offering virtualization products for more than 10 years – longer than any other company offering virtualization today. This means that VMware has more experience and their platform is the most mature.

4.2.3. Higher consolidation ratio

If you are going to consolidate your physical servers into virtual servers (P2V), you want to be able to put the greatest quantity of virtual servers on your physical server (the ESXi host). This way, you are going to get the most for your money and have the lowest TCO.

4.2.4. Memory over Commitment and Transparent Page Sharing

Speaking of features that result in the VMware solution costing less, memory over commitment and transparent page sharing are two of the most important ESX features. Again, I refer to the blog post entitled Cheap Hypervisors: A Fine Idea

4.2.5. Most number of complementary applications that you can upgrade to in the future

One day, when you do choose to upgrade from the Free VMware ESXi to vSphere, you will have a ton of VMware options to choose from. For example:

- ✓ Cloud computing
- ✓ Chargeback of resources
- ✓ Dynamic Resource Scheduling with DRS
- ✓ High Availability with VMHA and VMFT
- ✓ Greater Uptime with VMotion and SVMotion
- ✓ Virtualization Security
- ✓ Distributed Virtual Switch (dvSwitch) and, optionally, the Cisco Nexus 1000V
- ✓ VMware Data Recovery for easy backup and restore of virtual machines

4.2.6. Choose the right virtualization platform the first time

The old saying is "no one ever got fired for buying IBM". I think that this saying applies today to VMware. What this comes down to is you want to choose the right virtualization solution the first time. You do not want to implement one company's virtualization option, then have to convert to VMware later.

4.2.7. The most "grassroots" support and education options available

I talked about how VMware has been making virtualization solutions longer than other companies. What this means is that there are many more websites, blogs (like VirtualizationAdmin.com), books, training videos, and consultants available to help you implement VMware's solutions. Face it, you expect to be able to Google a virtualization topic or issue you are having and find a solution

4.2.8. Less Infrastructure and Reduced Energy Costs

VMware's solutions are similar to other virtualization platforms in the sense that, by consolidating servers, they are going to reduce infrastructure. Think about it – less physical servers equal less power, less cooling, less space, and less datacenter infrastructure.

5. RELATED WORK

In the last decade, the goals of distributed systems have been centered on the decoupling of interfaces from implementation, the hosting models, subscription-based computing, the service orientation architecture (SOA) [10], and social collaboration. Recently, Internet-hosted distributed, multi-tenant [11] applications with connectivity to internal business applications, referred as software as a service (SaaS) [6], are gaining popularity. Businesses are leveraging datacenters hosted by third-party providers with Internet-accessible, virtualized Cloud Computing environment to alleviate concerns about hardware, software, maintenance, availability, reliability, and scalability.

Particularly, scalability has been one of the principal topics discussed regarding Cloud Computing. Scalability is the ability of an application to be scaled up to meet demand through replication and distribution of requests across a pool or farm of servers. Dynamic scalability of web applications in virtualized Cloud Computing as outlined in the present work has not been much discussed. Most of the previous works [11-13] on web scalability have been reported through implementation of static load balancing solutions with server clusters.

6. CONCLUSION

In summary, we have presented a scaling scenario to address the dynamic scalability of web applications on a virtualized Cloud Computing environment. The system is constructed with a front-end load balancer to route and balance user requests to web applications deployed on web servers in virtual machine instances in the Cloud. A dynamic scaling algorithm for automated provisioning of the virtual machine resources based on number of active sessions has also been introduced. Our work has demonstrated the compelling benefits of the Cloud which is capable of handling sudden load surges, and delivering IT resources on-demands to users in a better and cheaper way. The automated provisioning, dynamic allocation and rapid scaling capabilities of the Cloud are the essential elements in providing higher resource utilization, thus reducing infrastructure and management costs.

7. REFERENCES

- [1] G. Gruman, "What cloud computing really means", InfoWorld, Jan.2009.
- [2] R. Buyya, Y. S. Chee, and V. Srikumar, "Market-Oriented Cloud Computing: Vision, Hype, and Reality for Delivering IT Services as Computing Utilities", Department of Computer Science and Software Engineering, University of Melbourne, Australia, July 2008, pp. 9.
- [3] D. Chappell, "A Short Introduction to Cloud Platforms", David Chappell & Associates, August 2008.
- [4] VMware ESX Server, VMware Inc., http://www.vmware.com/products/vi/esx/
- [5] Xen Hypervisor, http://www.xen.org/
- [6] E. Knorr, "Software as a service: The next big thing", InfoWorld, March 2006.
- [7] Virtualization Technology, http://www.kernelthread.com/publications/virtualiz ation/
- [8] VMware Inc., "Understanding Full Virtualization, Paravirtualization, and Hardware Assist", VMware, 2007,http://www.vmware.com/files/pdf/VMware_p aravirtualization.pdf.
- [9] Kernel-based Virtual Machine (KVM), http://www.linuxkvm.org/page/Main Page
- [10] Force.com, http://www.salesforce.com/platform/
- [11] Microsoft Virtualization, http://www.microsoft.com/virtualization/
- [12] IBM Tivoli Provisioning Manager Products, http://www.ibm.com/software/tivoli/products/provmgr/
- [13] L. He, S. Smith, R. Willenborg and Q. Wang, "Automating deployment and activation of virtual images", IBM developerWorks, Aug. 2007. http://www.ibm.com/developerworks/websphere/te chjournal/0708_he/0708_he.html

ABOUT AUTHORS

N.STAALINPRASANNAH received the Master of Computer Technology degree in 2005, Periyar University, majoring in Computer Technology. He is the recipient of computer society of India. Currently, He is pursuing ME., Computer Science and Engineering degree at Mahendra Institute of Engineering &Technology Affiliated to Anna University.



S.SURIYA received the Bachelor of Computer Science degree in 2010, Thiruvalluvar University, majoring in Computer Science. She is the recipient of computer society of India Scholarship. Currently, she is pursuing her MCA

degree at Vivekanandha Business School for Women Affiliated to Anna University.

T.RANJITHA received the Bachelor of Mathematics (Computer Application) degree in 2010, Periyar University, majoring in Computer Science. Currently, she is pursuing her MCA degree at Vivekanandha Business School for Women Affiliated to Anna University

