1. INTRODUCTION

Datacenters are an integral part of today’s Information Technology (IT) infrastructure and internet services. Business organizations and corporations around the world rely heavily on datacenters for their daily operations. Datacenter is a facility hosting massive number of interconnected servers as physical server clusters or virtualized server clusters along with support infrastructures (Figure. 1.1), providing computational and storage capacities and serving users with needs, like interactive computation, batch computation, and real-time transactions etc. Compute infrastructure consists of server clusters with interconnected servers, which in-turn could host virtual machines (VM). Servers consume power to process user workload requests, which leads to heat generation and increase in temperature of server and also the datacenter as a whole. Support infrastructure ensures server’s power supply and maintains server temperatures within operational limits. Support infrastructure includes emergency power generation units, power distribution units, uninterruptible power supplies and the cooling infrastructures. Cooling infrastructure includes server fans, computer room air conditioners and chillers.

![Figure. 1.1: Virtualized datacenter](image)
1.1 ENERGY CONSUMPTION

Primary goal of a datacenter is to provide good performance to the customer base. Secondary and equally important goal is to optimally minimize cost to operate the datacenter or server clusters.

With the emergence of cloud computing and growth in big data, the number of servers in a datacenter will rise even higher. At such scales, the power consumed by the servers becomes a significant operational expense. In terms of cloud service provider’s operational expense, energy cost is one of the main contributing entities [1] [2].

1.2 RESEARCH PROBLEMS

This thesis tackles the research challenges with respect to energy-efficient VM scheduling, dynamic VM consolidation, dynamic server provisioning using server power state transitioning and use of computer room air conditioner controls to optimally use cooling power to reduce datacenter energy consumption. List of problems addressed are as follows:

- Typical datacenter consists of multitude of server models or server configurations catering to different user or application and performance needs. At any point in time, a datacenter may host 3-5 server generations with a few hardware configurations per generation with respect to processor speeds, memory, storage, power consumption, and networking subsystems. Hence, it is common to have 10 to 40 configurations throughout the datacenter [39]. Servers of a given type are modeled by their processing capacity, thermal characteristics as well as their power consumption. Also servers have power and latency time overheads when transitioning between power states (on, off etc.). Server’s energy efficiency is driven by performance and power consumption and is influenced by server heterogeneity. So, a good solution design should account for this heterogeneity aspect to meet user needs and at reduced energy cost. Some servers exhibit good performance characteristics at higher power consumption characteristics. Some servers comparatively have better
performance, and consume less power and so on. An approach to rank servers based on energy efficiency with respect to heterogeneity is proposed in this work thereby accounting for both performance and optimal energy consumption.

- With virtualization, a server can be configured to host multitude of VMs, each of these VMs, which from application or users perspective behaves like a physical server. Operationally, virtualization technology helps scale up and down resource capacities to virtual machines and also helps consolidate servers by selective migration of VMs between physical servers. This consolidation approach reduces the number of physical servers by either switching off or transitioning to low power sleep state when un-used, thus reducing datacenter’s overall power consumption. An approach has been outlined in this thesis that specifies when VMs should be identified for migration, identify which VM to migrate, and to which server should the VM be migrated.

- Server processors have various levels of power state operations that control server power consumption. In many of the current approaches, sleep state transition overheads (transition time and power consumption overhead) are not considered in virtualized server cluster environment, and even if considered are limited to single server. A combined dynamic server provisioning and VM consolidation approach considering the transition overheads has been proposed to optimize energy consumption by the system and avoid violations of the QoS requirements.

- In a datacenter, cooling infrastructures consume power (cooling power) to dissipate heat generated by computational systems and maintain temperatures within operational limits. This cooling power is nearly on par with the power consumed by computational systems itself. So, a holistic datacenter level energy efficiency improvement approach has been proposed to reduce datacenter cooling power consumption dynamically.
1.3 OBJECTIVES

To deal with the challenges associated with the above research problems, the following main and specific objectives have been delineated:

Main objective

- To propose power management approaches to improve energy efficiency of virtualized server clusters

Specific objectives

- To propose
  - a VM scheduling approach which uses performance-power heuristic to rank heterogeneous servers and a VM consolidation and deconsolidation approach to reduce power consumption
  - a power management technique approach in using low power sleep state server transitions
  - a combined compute and cooling power management control approach to improve datacenter energy efficiency
  - a combined Dynamic Power Management (DPM) and Dynamic Voltage Frequency Scaling (DVFS) approach to improve energy efficiency in a Software as a Service (SaaS) workload delivery model

1.4 CONTRIBUTIONS

The contributions of this thesis are broadly divided into the following categories: classification and analysis of the area, algorithms for dynamic VM placement and consolidation, algorithms for server state transitions, algorithm for dynamic temperature controls, and algorithms for Dynamic Power Management (DPM) and Dynamic Voltage Frequency Scaling (DVFS) power management approach.
The key contributions are:

**Contribution 1:** Taxonomy and survey of the state-of-the-art power management approaches.

**Contribution 2:** Algorithms for dynamic VM placement and consolidation.

**Contribution 3:** Algorithms for server sleep state transitions

**Contribution 4:** Algorithm for dynamic computer room air conditioner thermal controls.

**Contribution 5:** Algorithm for dynamic Power Management (DPM) and Dynamic Voltage Frequency Scaling (DVFS) combined power management approach.

**1.5 THESIS STRUCTURE**

The outline of this work is structured according to the depiction in Figure. 1.2, which shows the relationship between the Chapters, the approach adopted and scientific contributions made.
The remainder of this thesis is organized as follows:

- **Chapter 2** presents the background on power management approaches and a survey on research works done pertaining to the challenges discussed. This Chapter is derived from our paper [JNCA, 2015] MohanRaj V.K, and Shriram R, “Power management in virtualized datacenter – a survey”, Elsevier Journal of Network and Computer Applications (JNCA), (in Review), 2015.

- **Chapter 3** presents an algorithm for distributed dynamic server and VM provisioning and consolidation controls. This Chapter describes Contribution 2 in detail and is derived from our paper [JCSS, 2015]. MohanRaj V.K, and Shriram R, “Heterogeneity and thermal aware adaptive heuristics for energy efficient consolidation of virtual machines in

- **Chapter 4** presents the approach in using server sleep state transition controls. This Chapter describes Contribution 3 in detail and is derived from our paper [COMSNETS, 2012 and JCSS, 2015]


- **Chapter 5** presents datacenter dynamic thermal controls. This Chapter describes Contribution 4 in detail and is derived from our paper [JCSS, 2015].


- **Chapter 6** presents an algorithm for distributed dynamic server and VM provisioning with DPM state transition and DVFS controls. This Chapter describes Contribution 5 in detail. This Chapter is derived from our paper [TJEECS, 2014].

• **Chapter 7** presents the implementation setup, evaluation approach and experimentation results of approaches discussed in Chapters 3, 4, 5 and 6.

• Chapter 8 concludes the thesis with a summary of the main findings, discussion of future research directions, and final remarks.