

Chapter 1

Introduction

We live in a digital universe where massive volumes of digital data are generating in day-to-day activity. The traditional storage infrastructures are overwhelmed with this data and not suitable to store and manage. This problem is the fundamental factor for the paradigm shift from traditional IT solutions to cloud computing in various sectors including health, banking, commerce, defense, research and education, and entertainment. Cloud computing emerged as the prominent platform to handle this problem and it offers computing resources on a pay-per-use basis.

A workflow is a set of tasks that are organized in a sequence to perform the execution of real-time applications. It improves the application management capabilities and reduces the execution complexities. It is most common to model workflow applications as Directed Acyclic Graph (DAG). A sample DAG with ten tasks is shown in Figure 1.1. Each node of the DAG refers to the computational tasks, and edges refer to the precedence constraints between the tasks. A task T_i has input data represented as $f_{i.in}$ and once T_i completes its execution then its generated data $f_{i.out}$ is transferred to its successor(s) as shown in Figure 1.1.

Usually, most of the scientific applications including physics, earth science, astronomy, computer science, bioinformatics, etc., comprise a lot of tasks with internal dependencies [1]. These are highly complex applications with varying task loads. It is common to model scientific workflows as DAGs. Cloud computing offers various services to deal with these applications with distinct QoS levels of high availability and computing capacity [2].

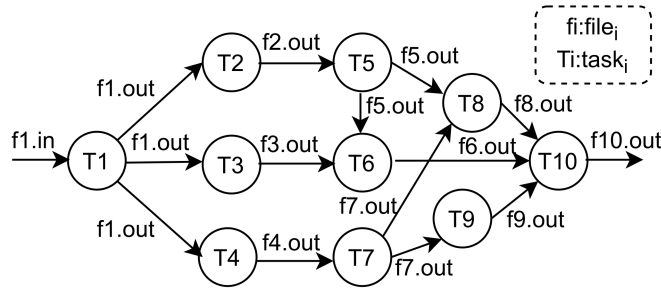


FIGURE 1.1: A sample workflow

Especially, the IaaS clouds provides an easily attainable, flexible, and scalable computing infrastructure for the deployment of workflow applications. The IaaS clouds are cost-effective and users can lease an unlimited number of computing resources like virtual machines without owning them.

Due to the high computing requirements for scientific workflow applications, data centers consume more electrical energy while executing workflows. The service providers practice different techniques to make sure that minimal energy costs. The scheduling algorithms play a key role in efficiently handling the workflow tasks and VM consolidation to get benefits from the cloud features and as well energy saving.

This thesis addresses the energy-aware workflow tasks scheduling and VM consolidation in IaaS clouds. It examines novel scheduling and VM consolidation approaches that address the major challenges derived from specific characteristics of clouds. This is achieved by conducting a detailed survey on the energy-aware scheduling algorithms for workflows in clouds. Furthermore, a new set of new approaches were proposed which are highly efficient in optimizing energy consumption.

1.1 Background

This section introduces a brief overview of the basic concepts like cloud computing, scientific workflows, and VM consolidation which are related to the research problem addressed in the thesis.

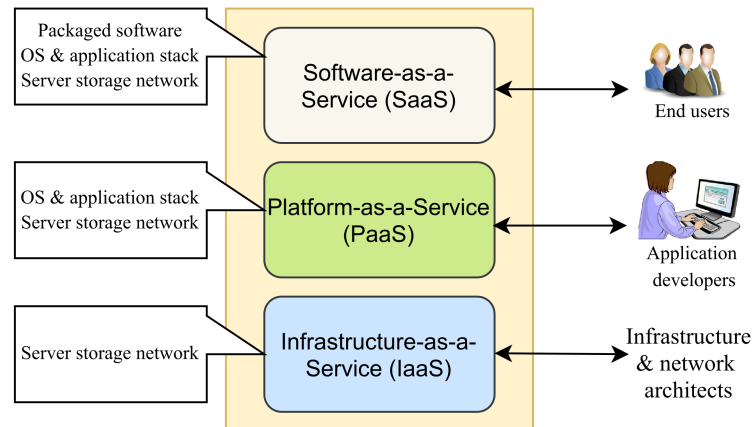


FIGURE 1.2: Classical cloud service offerings

1.1.1 Cloud computing

Cloud computing delivers computing resources such as physical servers, VMs, storage, network, etc., over the Internet. It depends on shared computing infrastructure instead of having on-premise servers. The NIST [3] proposed a definition for this paradigm as “a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction”.

Classically cloud services available to the users in three models such as Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). Figure 1.2 depicts the classical cloud service offerings to the users. SaaS is known as “on-demand software” where the service providers host applications that are accessible to the end-users. Users are enabled through the Internet to access these services. SaaS users get benefits from no need of purchasing and maintaining complex software. PaaS cloud platform is created for the programmer to develop, test, run, and manage the applications. IaaS is a computing infrastructure managed over the internet. The primary benefit of using IaaS is that it helps users free from the cost and complexity of purchasing and managing the physical servers. Users are allowed directly to access the core infrastructure of cloud computing and enable them to build applications from scratch.

1.1.2 Scientific workflows

The workflow concept has its origin in commercial business processing, which aims to 1) automate to pass tasks and information among participants and 2) optimize the business process in an organization. The business workflow research managed by Workflow Management-system Coalition¹ (the WfMC established in 1993). A WfMC can serve as a monitoring tool, and introducing WfMC in an enterprise can get increased management information [4].

The workflow concept extends to computational problems in science [5]. To execute large-scale scientific applications across the distributed sources, a scientific workflow system is a necessary tool [6]. The main goal of scientific workflows is to automate the scientist's repetitive data management and analysis tasks. The typical phases of scientific workflows are data access, scheduling, generation, transformation, aggregation, analysis, visualization, and data publication. Business workflows are task-oriented that can be processed in a specific sequence. These workflows optimize the business process. Usually, tasks, documents, etc., can undergo modifications, but such modified workflow objects are identifiable. These have complex control flow and complex process composition and often a separation between data flow and control flow. Despite this, business workflows and scientific workflows adopt the same concepts but have separate requirements, so they must be considered separately. Business workflows are out of the scope of this study. We focused on scientific workflows, and from now on, we will use “workflow” and “scientific workflow” interchangeably.

Different institutes or individuals in various scientific research fields are managing scientific workflows. Each domain has its software requirements to run the workflow tasks. Cloud computing is a suitable option to serve these needs. It is accessible to everyone, so scientific applications can run cost-effectively on clouds. Scientists can configure virtual machine images according to their workflow requirements. Further, they run applications efficiently across the heterogeneous resources of a cloud by making use of scheduling algorithms and workflow management systems to produce results in a reasonable time frame [7].

To express and process workflows, there are many models available such as DAGs, streaming pipeline, superscalar, and best effort. We limited the scope of this thesis to the

¹<http://www.wfmc.org>

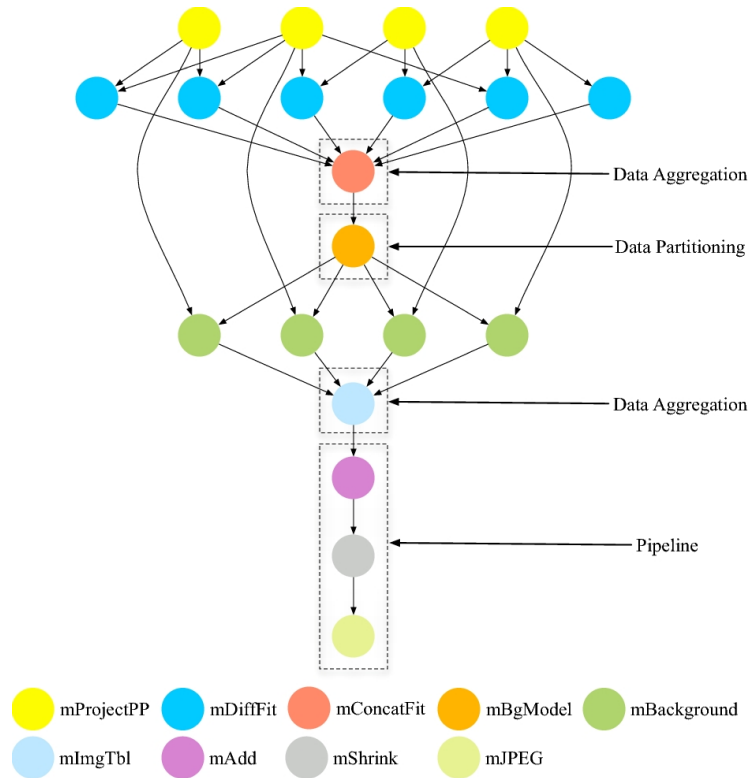


FIGURE 1.3: Montage

applications modeled as DAGs, which by definition have no conditional dependencies or no cycles. Different workflow management systems for instance DAGMan [8], Cloudbus WfMS [9], Pegasus [10], ASALON [11] etc., are support the execution of workflows modeled as DAGs.

Scientific workflows are computationally complex problems that need to be processed effectively in distributed environments such as clouds, which enables users to express complex problems in a simple way. For example, we presented five different real-world scientific workflows Montage, CyberShake, Epigenomics, LIGO and Sipt.

Montage [12] is a toolkit that enables researchers in assembling sky images in FITS (Flexible Image Transport System) in custom mosaics. It is an I/O-intensive astronomy application. Montage is a convenient and flexible tool to manage large image files with high accuracy and performance. Other key features such as portability, scalability, and open availability of this tool attracting researchers. For example Caltech astronomers at California Institute of Technology and even major projects like Spitzer Space Telescope

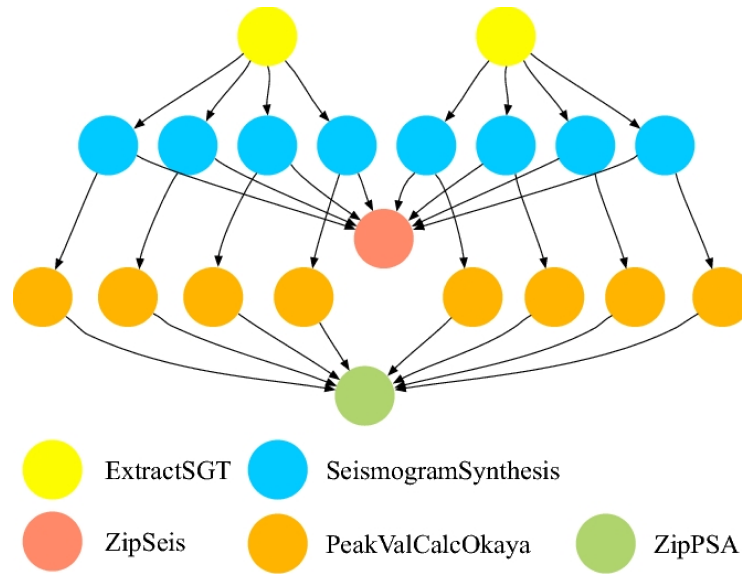


FIGURE 1.4: CyberShake

Legacy teams using Montage workflow. A sample topological structure of Montage workflow is depicted in Figure 1.3.

Cybershake [13] is a Probabilistic Seismic Hazard Analysis to characterize earthquake hazards. It is a memory-intensive and as well data-intensive application. This model was used in Sothern California to study Seismic Hazards. The structure of this workflow shown in Figure 1.4

Epigenomics [14] is a CPU-intensive workflow used in bioinformatics applications that enable the automation of several genome sequencing operations. A sample topological structure of this workflow is depicted in Figure 1.5.

SIPHT is sRNA Identification Protocol using High-throughput Technology program [15] for bioinformatics applications. The SIPHT empowers high-throughput, kingdom-wide prediction and functional annotation of bacterial sRNA-encoding genes. A sample topological structure of this workflow is shown in Figure 1.6.

LIGO the laser interferometer gravitational-wave observatory (LIGO) is an application in astrophysics [16]. It is a memory-intensive application in astrophysics [16]. This pioneering project detects gravitational radiation generated during the most violent events in the universe, like exploding stars, gamma-ray bursts, a collision of black holes, etc., or the explosion of supernovae. The radiation produced by these events has the potential to enter

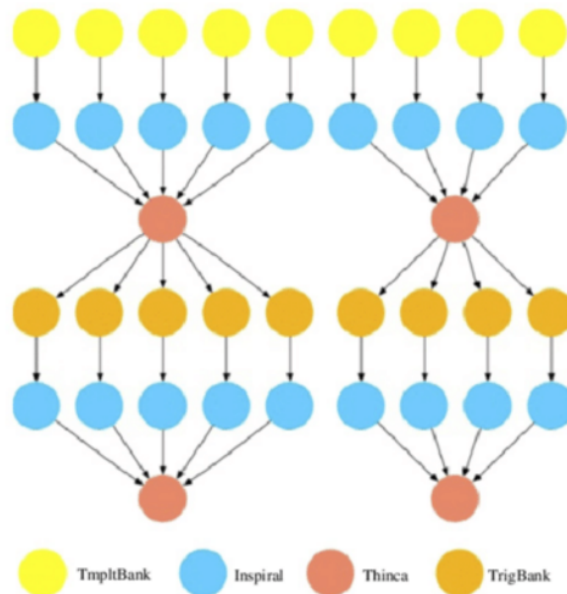


FIGURE 1.7: LIGO

into the vicinity of the Earth. A sample topological structure of this workflow is shown in Figure 1.7.

These applications have taken from different domains and provide a comprehensive view of sophisticated workflow technologies. Every workflow discussed above has its requirements in data and computing capacity. The characterization of these workflows is introduced in [17].

1.1.3 Virtual machine consolidation

VM consolidation is one of the power-saving approaches in cloud data centers by minimizing the number of servers by migrating and consolidating VMs into a minimal number of servers [18]. It optimizes energy utilization by placing the VMs on the most energy-efficient servers and improving the resource utilization of active servers. VMs are migrated for loading balancing or to switch off the underloaded servers. It switches off the idle/underloaded servers after migrating all of its VMs to a suitable power-efficient server(s).

The VM consolidation is performed in two ways 1) static VM migrations and 2) dynamic VM migrations. A new VM is placed on a physical server in a static approach but no migrations happen in reality. Whereas in dynamic VM consolidation migration of VM from source to destination servers happen whenever needed. Basically, VM consolidation can be carried by categorizing the servers based on the workload. Depending on the server workload it is categorized into one of three types: overloaded, underloaded, and normally loaded servers. In the context of VM migrations, any server is a source host for migrating VMs and any server is a potential destination host server [19].

1.2 Motivation

The increased demand for accessing cloud-based applications has led to significant growth in cloud data center traffic. Kimberly Mlitz [20] forecasts 20.6 zettabytes of IP traffic per year by 2021. The expected data center IP traffic between 2013 and 2021 is depicted in Figure 1.8. This fast growth in IP traffic resulting in high energy consumption in cloud data centers, hence it is required to develop energy-aware techniques. Most notable corporates including Google, Amazon, Microsoft, etc., have been utilizing online services to provide effective and reliable services to end-users by virtue of the cloud computing and storage models. According to Gartner, Inc., [21] the end-user spending on public clouds is anticipated to rise 18% by the end of 2021 over the previous year. The growth trend of different cloud-based events is studied including conferencing and telephony/messaging and is expected a 24.3% and 8.9% increase respectively. These increasing energy costs create a negative effect on the cloud infrastructure costs. As a consequence, a rise in energy utilization affects the total ownership cost and mitigates the ROI. Additionally this over energy utilization releasing green gases into the environment. It is estimated that the total greengases emitting by the data centers into the environment are around 3.2% of the global emissions by 2025 and energy utilization shares about 20% of the whole energy consumption in the world. Further, corban emissions of data centers due to over energy consumption will reach 14% of the world's emissions in two decades i.e by the year 2040 [22]. Hence, energy-efficient models are essential to minimize energy costs and improve the environment.

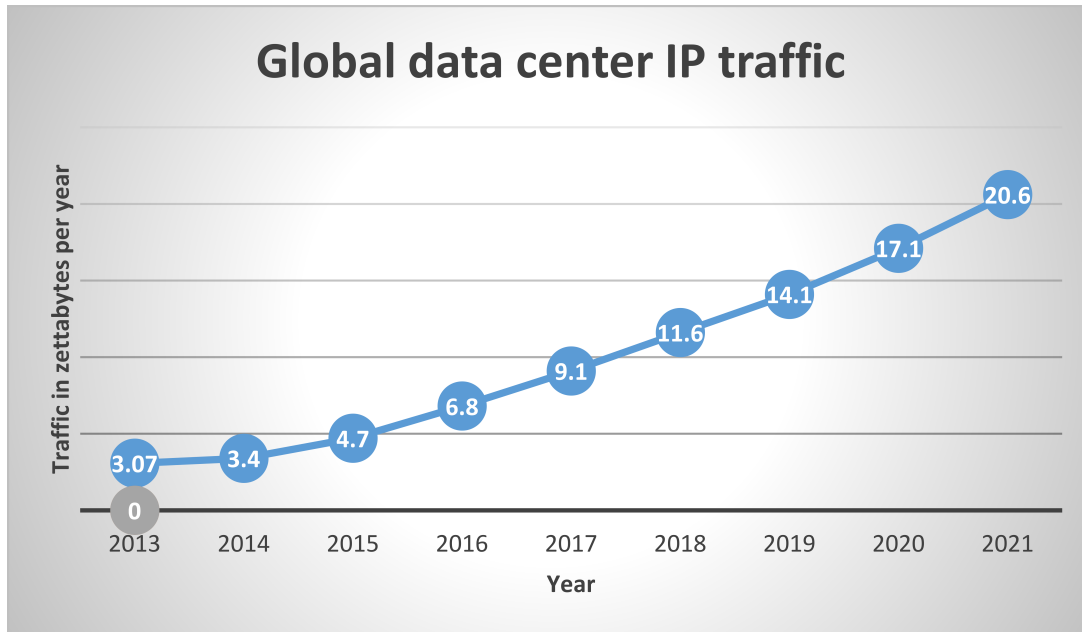


FIGURE 1.8: Global data center IP traffic from 2013 to 2021

The unnecessary use of cloud resources leads to excessive energy utilization. An active but idle server utilizes up to 70% of its peak power [23]. Hence, it is necessary to limit the data center resource utilization in the real world without compromising performance. Therefore, this thesis studies workflow scheduling and VM consolidation algorithms that aim to optimize energy consumption without violating the QoS requirements.

1.3 Thesis contributions

The primary contributions of this thesis are discussed in this section. Based on the thesis objectives defined previously the contributions are as follows:

- A review of the energy-aware workflow scheduling algorithms for IaaS clouds. Which discusses state-of-the-art techniques within the scope of this thesis in detail.
- A heuristic-based energy-efficient and cost-aware scheduling algorithm with deadline constraints.
- A heuristic-based scheduling algorithm for energy and reliability optimization.

- A meta-heuristic-based virtual machine consolidation algorithm.
- A hybrid scheduling algorithm with virtual machine consolidation.

1.4 Thesis organization

The primary chapters of this thesis are organized as illustrated in Figure 1.9. Chapters 2 to 6 are derived from various conference and journal articles published during the Ph.D. candidature. The rest of the thesis is organized as follows:

- Chapter 2 presents a detailed survey on the energy-aware workflow scheduling algorithms for IaaS clouds. This chapter studies existing algorithms for energy-efficient workflow scheduling in different perspectives and is acquired from the following source:
 - This chapter is derived from: Medara, R., Singh, R.S. A Review on Energy-Aware Scheduling Techniques for Workflows in IaaS Clouds. *Wireless Pers Commun* (2022). <https://doi.org/10.1007/s11277-022-09621-1>

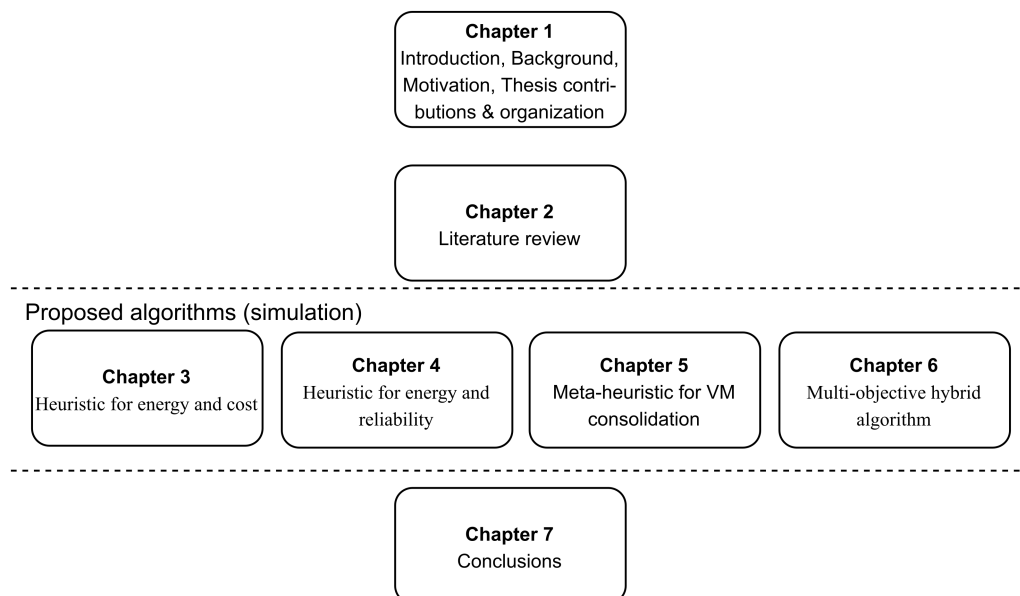


FIGURE 1.9: Thesis organization

- Chapter 3 presents a deadline-based scheduling algorithm for energy and cost optimization scheduling (ECWS). It evaluates the energy efficiency of the cloud resources (VMs) by using a parameter called the ratio of effectiveness (RE). The proposed scheduling algorithm removes VMs which are inefficient in power utilization based on the RE factor by considering the workflow's deadline. By removing power inefficient VMs from the schedulable VMs list, the ECWS algorithm saves valuable electrical energy. Then it maps the tasks to the energy-efficient VMs without compromising the performance. Further, it uses a slack reclaiming approach to save more energy. This chapter is acquired from the following source:
 - This chapter is derived from: Medara, Rambabu, Ravi Shankar Singh, and Mahesh Sompalli. "Energy and cost aware workflow scheduling in clouds with deadline constraint." *Concurrency and Computation: Practice and Experience*: e6922.
- Chapter 4 presents a heuristic-based scheduling algorithm for energy and reliability optimization. It comprises five sub-algorithms to manage the cloud resources in reducing energy utilization and maximizing system reliability without compromising the performance. This chapter is acquired from [1].
 - Medara, Rambabu, and Ravi Shankar Singh. "Energy Efficient and Reliability Aware Workflow Task Scheduling in Cloud Environment." *Wireless Personal Communications* (2021): 1-20.
- Chapter 5 presents a meta-heuristic-based virtual machine consolidation algorithm. VM consolidation implements VM migrations for load balancing and consolidates VMs to switch off idle servers. A nature-inspired algorithm "water wave optimization (WWO)" algorithm is used for implementing VM consolidation. This chapter is acquired from [24]
 - Medara, Rambabau, et al. "Energy Efficient Virtual Machine Consolidation Using Water Wave Optimization." *2020 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2020.
- Chapter 6 presents a hybrid scheduling algorithm with virtual machine consolidation. This algorithm runs in two phases scheduling and VM consolidation. We combined

a heuristic scheduling approach with a meta-heuristic algorithm to address the aimed objectives. The scheduling phase manages resources in an energy-efficient way. A discrete WWO algorithm is used to implement the VM consolidation phase. This chapter is acquired from [25]

- Medara, Rambabu, and Ravi Shankar Singh. "Energy-aware workflow task scheduling in clouds with virtual machine consolidation using discrete water wave optimization." *Simulation Modelling Practice and Theory* 110 (2021): 102323.
- Chapter 7 discusses the conclusions of the thesis and future research directions.