JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

SCIENTIFIC WORKFLOW SCHEDULING IN CLOUD CONSIDERING COLD START AND VARIABLE PRICING MODEL

[1]Panchashree, [2]Athmaranjan K, [3]Manisha
[1]Final year B.E. Student, [2]Associate Professor, [3]Final year B.E. Student
[1]Information Science & Engineering
[1,]Srinivas Institute of Technology, Mangaluru, India

Abstract : Cloud computing has become an essential platform for running scientific workflows because of its flexibility and cost-efficiency. Scientific Cloud Service Providers (SCSPs) serve as intermediaries by renting virtual machines (VMs) from Infrastructure-as-a-Service (IaaS) providers to fulfil workflow execution requirements. These providers generate profit when workflows are completed within their specified deadlines. However, two major challenges affect this profitability: the cold start problem and the complexity of managing various VM pricing models such as reserved, on-demand, and spot instances. This paper introduces a hybrid scheduling framework that combines initial resource planning using historical workload data with real-time adjustments based on current workload changes. In the planning phase, the system allocates VMs using reserved and spot instances according to predicted demands. During execution, it dynamically scales by adding more VMs through on-demand or spot instances when there are sudden increases in tasks. Additionally, the framework uses a dependency-aware scheduling method that considers both cold start delays and the fluctuation in spot instance prices. Experimental results using real-world benchmark datasets show that this approach performs better than existing methods, offering up to 20% improvement over cold-start-based techniques and 15% better efficiency than methods focusing only on pricing model optimization.

IndexTerms **- scientific workflow, cold start, pricing model, spot instances, VM renting**

I. INTRODUCTION

Cloud computing has become a critical platform for executing a wide range of workloads and workflows due to its inherent flexibility, scalability, and cost-effectiveness. Users typically submit workflows to cloud environments, which then provision the necessary computational resources for their execution [1], [2]. This research focuses specifically on scientific workflows, and refers to cloud infrastructure providers as Scientific Cloud Service Providers (SCSPs).

The SCSP primarily functions as an intermediary, receiving scientific workflows from users and renting virtual machines (VMs) from Infrastructure-as-a-Service (IaaS) providers. This model allows SCSPs to avoid the capital expenditures (CapEx) associated with owning physical infrastructure. The main goal of the SCSP is to generate profit, which it does by charging users subscription or usage fees and strategically renting resources from IaaS providers. Notable examples of IaaS providers include Amazon Web Services (AWS EC2), Microsoft Azure, and Google Cloud Platform (GCP), while SCSPs include Jetstream, Chameleon Cloud, Open Science Grid, CloudLab, and SciServer.

In addressing this objective, two significant challenges are highlighted. The first challenge is the cold start problem, which complicates workflow execution. Cold start refers to delays encountered during the initial setup before task execution begins, typically accounting for around 20% of total execution time [3]. These delays can degrade performance, lead to inefficient resource utilization, and increase the likelihood of missing deadlines. This problem becomes especially important in online workflow scheduling, where task arrivals are unpredictable and dynamic. Efficient scheduling strategies are needed to balance cost and execution time. This issue is particularly crucial because studies have shown that approximately 20% of functions are invoked around 99% of the time [3]. By strategically using caching mechanisms and optimizing cold start handling, significant improvements in execution time and resource utilization can be achieved.

The second challenge involves the strategic management of machines rented under different pricing models provided by IaaS providers [4]. These providers offer various VM types with different memory and computational capacities, available through reserved, on-demand, or spot pricing models [5]. Reserved instances are booked in advance based on historical usage data, providing predictable costs but risking either underutilization or insufficient resources during unexpected workloads. On-demand instances, although more expensive, accommodate real-time demand spikes. Spot instances, offered at significantly discounted prices, are available based on real-time supply and demand, but their availability is not guaranteed. Spot instances are at risk of being revoked if the market price exceeds the SCSP's bid, potentially interrupting tasks and requiring reallocation and re-execution. Despite this uncertainty, spot instances remain attractive for non-critical or fault-tolerant tasks due to their lower cost [6].

The primary challenge lies in balancing these different VM pricing options. Reserving too many VMs can waste money if they are underutilized, though they help mitigate cold start delays because caching works more effectively with reserved VMs. On the other hand, fewer reserved VMs may require more expensive on-demand VMs. Spot VMs are cheaper but come with risks of revocation if prices increase. Bidding higher to secure spot VMs raises costs, while bidding lower saves money but risks losing the

VM. This paper addresses this issue by helping SCSPs maximize profit by: (a) managing multiple user workflows, (b) leveraging spot price fluctuations, (c) minimizing cold start costs, and (d) using historical data and dynamic price changes for better planning.

To tackle these challenges, we propose a hybrid two-phase scheduling strategy. In the first phase, we develop an initial schedule based on historical workload data (referred to as the predicted phase). In the second phase, adjustments are made dynamically in real-time (referred to as the actual/predicted phase) to accommodate deviations from the predicted workload. The main challenges in real-time scheduling include managing cold start delays, adjusting reserved bookings dynamically, and efficiently utilizing on-demand and spot instances despite the risk of spot instance revocations. These challenges are addressed in our real-time scheduling framework.

The key contributions of this work are as follows:

- We propose an integrated scheduling framework that addresses cold start delays, task dependencies, and VM pricing strategies—unlike previous approaches that treat these concerns separately. To our knowledge, this is the first work to incorporate all three major VM pricing models (reserved, on-demand, and spot) into a unified scheduling strategy.
- We introduce a hybrid two-phase scheduling strategy: an offline phase that plans the use of reserved VMs based on predicted workloads and spot availability, and a real-time phase that dynamically provisions on-demand and spot instances in response to actual workflow demands.
- We incorporate deadline- and dependency-aware scheduling by proportionally distributing workflow deadlines across tasks. This minimizes cold start overheads and reduces the risk of deadline violations, directly contributing to higher profit margins.
- We develop a cost-aware VM pool management strategy that includes a priority-based VM selection mechanism and a reward-guided spot bidding policy. These strategies optimize task placement while mitigating the risks of spot instance revocation.
- Experimental evaluations using real-world scientific workflows and pricing traces demonstrate consistent performance improvements over state-of-the-art baselines. Our framework remains robust under varying workload intensities, spot availability patterns, and pricing dynamics, achieving up to 80.394% profit even with up to 40% prediction error.

II. Applications of Scientific Workflow Scheduling in Cloud Environments

The following are some applications of scientific workflow scheduling in cloud environments, particularly considering cold start issues and variable pricing models:

2.1 Genomic Data Analysis

Genomic data analysis involves processing and interpreting vast amounts of DNA and RNA sequencing data to understand genetic information, mutations, and biological functions. It typically includes steps such as sequence alignment, variant detection, and functional annotation, organized into scientific workflows. These workflows demand intensive computation and large-scale data handling, making cloud computing a practical solution. However, the cold start problem in provisioning virtual machines (VMs) can delay workflow execution, especially for time-sensitive tasks. Variable pricing models, such as spot and on-demand instances, further complicate cost and resource planning. Intelligent scheduling frameworks can predict workload demands and provision suitable VMs in advance to minimize delays. Cost efficiency is achieved by leveraging low-cost spot instances for background processing and switching to on-demand VMs for urgent jobs. This ensures faster, more affordable genomic analysis crucial for medical research and personalized healthcare.

2.2 Climate Modeling and Simulation

Climate modeling simulates complex interactions among the atmosphere, oceans, and land surfaces to forecast climate behavior and predict weather patterns. These simulations consist of multi-step scientific workflows requiring large-scale data input and continuous computation over extended periods. Running these workflows on cloud platforms offers scalability and flexibility, especially during intensive modeling phases. However, VM cold starts can slow down early-stage simulations, while unpredictable spot instance pricing can affect budget control. Efficient scheduling is essential to pre-provision VMs and balance cost-performance tradeoffs across different pricing models. Reserved instances can support predictable workloads, while on-demand instances help handle computational surges caused by real-time weather updates. This hybrid approach improves responsiveness in weather prediction systems and reduces costs in long-term climate forecasting. As a result, climate scientists can produce more accurate, timely results to support environmental planning and disaster preparedness.

2.3 Drug Discovery and Molecular Simulation

Drug discovery involves simulating molecular interactions, docking experiments, and chemical property predictions to identify potential therapeutic compounds. These simulations are implemented as scientific workflows composed of computationally intensive tasks, often repeated multiple times for different compounds. Cloud-based infrastructure offers a viable solution to handle the computational load while accommodating changes in simulation demand. Cold start delays can disrupt rapid testing phases, especially when new simulations are launched dynamically. Spot instances can be cost-effective for large-scale, non-urgent screenings but may be interrupted unpredictably. Intelligent scheduling allows fallback to on-demand instances during such interruptions to maintain workflow continuity. By predicting usage patterns, the system can reserve VMs in advance to avoid delays and optimize resource usage. This approach speeds up the drug discovery process while keeping research costs manageable.

2.4 Astrophysics Data Processing

Astrophysics workflows involve processing vast amounts of data from space-based and ground-based observatories to study cosmic phenomena like black holes, quasars, and gravitational waves. These workflows require high-throughput computation and large-scale storage for tasks such as signal filtering, data classification, and pattern recognition. Cloud computing provides the necessary scalability to handle sudden data influxes from observation events. However, cold start latency in VM provisioning can be critical when analysing real-time astronomical events. Spot instance interruptions can also lead to incomplete analysis if not properly managed. Efficient scheduling frameworks pre-allocate VMs for predictable tasks and dynamically scale using on-demand instances

during observation surges. This ensures minimal downtime and accurate analysis under variable cloud conditions. As a result, astrophysicists can achieve faster, more reliable insights from massive cosmic datasets.

2.5 Earthquake and Disaster Prediction Systems

Disaster prediction systems rely on scientific workflows that analyze seismic data and environmental indicators to forecast earthquakes, tsunamis, and other natural disasters. These workflows must operate with low latency and high reliability, as delays can impact timely warnings and response. The cloud provides a dynamic platform to manage these urgent and data-intensive computations, especially during unpredictable activity spikes. Cold start problems in VM provisioning can hinder immediate analysis when a seismic event is detected. Likewise, relying solely on low-cost spot instances risks interruption during critical computations. A robust scheduling model can reserve core VMs for continuous monitoring and use on-demand instances to handle surges in processing needs. This hybrid approach reduces response time and ensures consistent system availability during emergencies. Ultimately, it enhances public safety by enabling faster, data-driven decision-making in disaster management.

Additionally, by incorporating historical data patterns and real-time sensor input, the scheduler can prioritize tasks dynamically, ensuring that the most critical predictions are processed first with minimal delay. Integrating fault-tolerant mechanisms within the scheduling framework helps maintain workflow continuity even during VM interruptions or cloud infrastructure failures, thereby increasing the reliability of life-saving early warning systems.

III. CHALLENGES AND LIMITATIONS

Scientific workflow scheduling in the cloud is complicated by the cold start problem, where initializing virtual machines introduces delays that affect time-sensitive tasks. Accurately predicting workload demands for advance VM reservations is difficult, especially for irregular or bursty scientific workloads. Dynamic pricing models, such as spot instances, offer cost savings but come with reliability concerns due to possible instance termination. Ensuring deadline adherence while balancing performance and cost across heterogeneous VM types is a significant scheduling challenge. Moreover, maintaining workflow dependencies during dynamic scaling adds complexity to the scheduling logic.

The unpredictability of spot market prices can hinder consistent cost savings, making resource planning challenging. Reliance on historical workload data may not accurately capture future demand, potentially resulting in under- or over-provisioning of resources. Incorporating fault-tolerant mechanisms adds overhead, which can impact overall system efficiency. Furthermore, existing frameworks often lack fine-grained control over VM boot times and real-time pricing fluctuations, reducing their responsiveness to sudden changes in workload requirements.

IV. FUTURE DIRECTIONS

Future research can focus on integrating machine learning techniques to improve workload prediction accuracy and optimize VM provisioning decisions in real time. By leveraging deep learning models trained on historical and streaming data, the scheduler can better anticipate resource demands and minimize cold start delays. This would enhance both cost efficiency and deadline compliance in dynamic scientific workflow environments.

Another promising direction involves developing adaptive hybrid scheduling algorithms that intelligently combine reserved, spot, and on-demand instances based on evolving workflow requirements and market conditions. These algorithms should dynamically switch strategies during execution to maximize resource utilization and maintain service-level agreements. Further, incorporating user-defined QoS (Quality of Service) preferences can enable more flexible and personalized scheduling outcomes.

Future systems should also emphasize greater interoperability and standardization across cloud platforms to support more portable and reusable scientific workflows. Enhancing support for real-time monitoring and feedback loops can help continuously refine scheduling decisions and respond swiftly to runtime uncertainties. Incorporating edge computing capabilities may further reduce latency and enhance performance for geographically distributed scientific applications. Moreover, fostering collaboration between cloud providers, researchers, and developers can lead to the creation of unified frameworks that better address cold start issues and pricing variability.

Conclusion and Future work

This paper addresses key challenges in scheduling scientific workflows in cloud environments, focusing on mitigating cold start delays and managing various VM pricing models (reserved, on-demand, and spot instances) cost-effectively. We proposed a unified hybrid scheduling framework that combines historical workload predictions with real-time adjustments to improve the profitability of Scientific Cloud Service Providers (SCSPs). The hybrid two-phase approach plans VM reservations based on predicted workloads and spot instance availability, while dynamically adapting to actual demand by provisioning additional on-demand or spot instances. This novel concept for workload scheduling includes cost-aware VM management, utilizing a priority-based VM selection mechanism and a reward-driven spot bidding strategy, which effectively balances cost reduction with the risk of spot instance revocation. Extensive experiments with real-world workflows and pricing data demonstrated the framework's scalability and robustness, consistently outperforming baseline methods even with significant prediction errors. Future work will focus on dynamic hyperparameter tuning and the integration of advanced reward mechanisms to further optimize both profit and efficiency.

REFERENCES

[1] J. Yu and R. Buyya, "A comprehensive survey on workflow scheduling algorithms for cloud computing," Concurrency and Computation: Practice and Experience, vol. 29, no. 12, p. e4041, 2017.

[2] M. A. Rodriguez and R. Buyya, "A classification and survey of scheduling algorithms for scientific workflows in IaaS cloud environments," Concurrency and Computation: Practice and Experience, vol. 29, no. 8, p. e4041, 2017.

[3] M. Shahrad, R. Fonseca, I. Goiri, G. Chaudhry, P. Batum, J. Cooke, E. Laureano, C. Tresness, M. Russinovich, and R. Bianchini, "Analyzing and optimizing serverless workloads at a major cloud provider," in Proceedings of the 2020 USENIX Conference on Usenix Annual Technical Conference, USENIX ATC'20. USA: USENIX Association, 2020.

[4] *J. Zhao, H. Li, C. Wu, Z. Li, Z. Zhang, and F. C. M. Lau, "Dynamic pricing strategies and profit maximization for cloud services with geographically distributed data centers," in IEEE INFOCOM 2014 - IEEE Conference on Computer Communications, 2014, pp. 118–126.*

[5] D. Kumar, G. Baranwal, Z. Raza, and D. P. Vidyarthi, "A comprehensive survey on spot pricing in cloud computing," Journal of Network and Systems Management, vol. 26, no. 4, pp. 809–856, Oct. 2018. [Online]. Available: https://doi.org/10.1007/s10922-017-9444-x

[6] V. K. Singh, S. Shivendu, and K. Dutta, "Effect of spot instance similarity and substitution in the cloud spot market," Decision Support Systems, vol. 159, p. 113815, 2022.