Joint Optimization of Resource Provisioning In Cloud Computing

Manika Manwal,
Asst. Professor, Department of Comp. Sc. & Info. Tech., Graphic Era Hill University, Dehradun, Uttarakhand India 248002

Abstract

Virtualization is utilized by cloud computing to effectively supply resources. Virtual Machines (VMs) are requiring more and more bandwidth, but prior study has not adequately addressed the issue of both VM and bandwidth allocation. A combined strategy combining bandwidth distribution and virtual machines (VMs) is needed to effectively supply resources. Demand in real life is also unpredictable. Resource reservations are permitted by service suppliers. However, we employ stochastic programming to account for this risk into consideration due to risks associated with over- and under-provisioning. We use a scenario tree reduction method to condense the problem space in order to increase the tractability of the stochastic optimization while maintaining its effectiveness as a heuristic. We also conduct a sensitivity analysis to determine how sensitive our answer is to parameter variations. We are using a deterministic equivalent formulation to analyses past demand data and discover that our answer is both optimum and responsive to alterations in parameter values. Additionally, this research demonstrates how consumers and suppliers can both benefit from price sensitivity analysis to maximise cost effectiveness.

DOI Number: 10.48047/nq.2021.19.5.NQ21081 NeuroQuantology 2021; 19(5): 330-335

1. INTRODUCTION

The recent data explosion has increased requirementfor large data handling in modern data centres, which are typically dispersed across various geographic areas. Data centre resizing (DCR), which modifies the number of enabled computers via task allocation, has been suggested to lower processing costs. For fine-grained load balancing and high parallel data access efficiency, big data service systems, for instance, have a distributed file system underlying them that spreads data chunks and their copies throughout data centres. A few new studies try to enhance data locality by putting tasks to the servers that house the input data in order to prevent remote data loading in an effort to reduce transmission costs.Even though, these approaches fall short of achieving cost-effective big data processing due to the flaws such as: data locality may lead to resource waste, network links vary in transmission rates and costs depending on their special characteristics, and the current data centre routing strategy does not benefit from the link diversity of data centre networks.Through combined work assignment, data placement, and route optimization in geographically dispersed data centres, this book examines the cost reduction issue for handling large data. The Service-Level Agreement (SLA) between a service provider and the consumer, as well as the transfer cost, energy use, and Quality-of-Service (QoS) of large data duties, are all taken into account.The goal is to reduce the total processing and transmission costs by optimising big data location, routing, job assignmentand DCR. We suggest a two-dimensional Markov chain and deduce the anticipated time for task completion in closed form to characterise the rate-constrained calculation and transmission in big data processing process. We linearize MINLP as a mixed-integer linear programming (MILP) problem that can be resolved utilizing a commercial algorithm, to lower the high computational complexity of the issue. We demonstrate the high effectiveness of our suggested joint-optimization based method through in-depth numerical studies.

2. LITERATURE SURVEY

In this article, researchers show the overall power consumption features of big server clusters (up to 15,000 servers) for various application classes over the course of about 6 months. Their ability to assess chances for optimizing the usage of distributed power capacity of datacenter and evaluate the dangers of oversubscribing it is made possible by these findings. Researchers discover that there is a significant difference (7–16%) among obtained and predicted collective peak power consumption at the cluster level, even in well-tuned applications (thousands of servers). This headroom can be used to implement extra computing resources while maintaining the same electricity limit with little chance of going over.Researchers discover that there are substantial chances for energy and power savings, but these opportunities are bigger at the cluster-level (involving hundreds of servers) as compared to the rack-level (tens). Lastly, researchers make the case that systems should be power-efficient throughout the action spectrum rather than just at the very top of their game. Due to the possibility that the cost of constructing sizable datacenters could be higher than the energy'scost over the facility's lifespan, choices regarding power supply for such systems might have a significant economic effect.Intelligent power provisioning also has a significant strategic effect because it may enable a current facility to support business development within a specific power budget, despite the fact that building a new data centre can take tens of months [1].

Additional uses of this work include production planning, proactive scheduling with setup expenses, and dynamic placement of Web apps. Researchers demonstrate that, even for very limited cases, the reconfiguration issue is NP-hard. Then, researchers create methods that use servers with load capacities that are increased by O(1), specifically, by factor 1 + for any small 0 1 when the number of servers is fixed, and by factor of 2 + for an unlimited number of servers, for some (0, 1). To the best of our understanding, this is the first time the data migration issue has been examined in this specific variant. Video on Demand (VoD) programmes are now widely used in business, entertainment, and information retrieval uses in libraries.A network connects customers in a VoD system to a collection of servers that house a sizable catalogue of video programmes. Allocate a unit load capability (or data stream) on a server that has a video'scopy in order to transmit it to a client. Therefore, the system must modify the location of file duplicates and distribution of load capability to these copies in order to keep high throughput. As researchers use bandwidth and other resources on both the source and target computer, file replications are very expensive.To maximise system efficiency, this expense must be reduced [2].

In this article, researchers suggest a mechanism that maintains the "semantic" character of LOD while enabling it to benefit from current large-scale datastores. Researchers investigate how to adjust the basic semantic operation to satisfy the demands of distributed and parallel data processing and show how to distribute RDF-based semantic models across various storage sites.Future research is going to concentrate on platform endurance testing with tens of billions of triples of load, in addition to usability and efficiency comparisons with competing products. RDF-encoded data can be stored and queried using the suggested storage option. Researchers aim for a distributed tolerant of fault system that can grow in terms of the volume of data that needs to be handled.Since each RDF triple serves as Key for a Key-value combination, researchers use a dispersed, ordered Key-value database (Seeger, 2009). In the meantime, value component of the Key-value combinations offers simple way to link information with specific triples. This offers a wonderful opportunity to increase the strength of the distributed data store, for instance by integrating new application-specific features directly into the storage system, which would be more effective than doing so outside the storage system. The semantic layer also adds a new consistency component because numerous data sources frequently use various references to denote the same physical entity. Finding and combining numerous references for semantic coherence is a crucial and essential step towards using LOD-enabled Big Data [3].

Typically, data centres have three primary components. Customers are served by IT equipment, which is supported by electricity infrastructure and cooling infrastructure that eliminates heat produced. In order to simulate the energy flows in a data centre and improve its overall performance, this study provides a new methodology. Historically, IT workload management was mainly handled separately from supply-side limitations like electricity or cooling availability. This study improves the general realizability of data centre operations by reducing costs and environmental effect by integrating energy supply such as renewable supply and dynamic pricing and cooling supply such as chiller and exterior air conditioning.More precisely, researchers develop an IT workload management strategy that distributes IT resources and schedules IT tasks inside a data center in

accordance with power supply that varies over timeand cooling effectiveness. Researchers also forecast demand for IT and renewable energy. Utilizing traces from actual data centres and manufacturing systems, we put our strategy into practiseand assessed its effectiveness. The findings show that, while maintaining business objectives and Service Level Agreements, their method can cut ongoing power expenses and the usage of non-renewable energy by up to 60% in comparisonwith current, non-integrated methodologies. Due to the rising demand for IT services andapplications, data centres are mushrooming all over the globe and consume a significant quantity of electricity.The use of alternative cooling methods and on-site renewable energy sources in data centre architecture is one growing option [4].

The amount of power used in a datacenter has a big effect on its one-time building expenses as well as its ongoing electricity bill (Op-ex) (Cap-ex). Throttling devices or task shaping—both of which have the potential to degrade performance—have been the mainstays of existing work aimed at reducing these expenses. In order to achieve this cost minimization, researchers present in this article a new energy buffer control (eBuff) that is accessible in datacenters in the shape of UPS batteries. According to intuition, eBuff saves energy in Backup batteries during "valleys"—when demand is at its lowest—which can be depleted during "peaks"—when demand is at its highest. It is a good idea to have a backup plan in place, especially if you are planning to travel. Additionally, repeated outages can hasten the early failure of UPS batteries.Given the worries about battery longevity and datacenter availability, researchers perform a thorough study of battery operation to identify practical working zones. Researchers create peak reduction algorithms that merge UPS battery knob with currently used throttling-based methods to reduce server power expenses using the insights gained from this research. Researchers provide details about the Op-ex benefits provided by eBuff for a variety of workload peaks/valleys, UPS supply, and application SLA limitations using an experimental platform. Researchers discover that eBuff can be used to achieve peak power reductions of 15–45%, which translates to Op-ex savings of 6–18% across this range. By providing for tighter overbooking of electricity infrastructure components, eBuff can also contribute to lower Cap-ex expenses. Researchers calculate the size of such Cap-ex savings.To their understanding, this is the first study to handle the high power demand issue by utilising stored energy, which usually lies dormant in datacenters [5].

3. PROPOSED SYSTEM

We were the initial ones to take into account data placement, data routing, and job assignment jointly with the big data processingcost reduction issue. We suggest a two-dimensional Markov chain and deduce the anticipated time for task completion in closed form to characterise the rate-constrained calculation and transmission in large data processing process. We define the cost minimization issue using mixed integer nonlinear programming (MINLP), built on the closed-form equation, to address the following queries: How to resize data centres to achieve the goal of minimizing operation costs. How these data chunks should be placed in the servers. How to spread duties among servers without going against the resource limits.We linearize MINLP as a mixed-integer linear programming (MILP) problem in order to reduce the high computational complexity of the issue and make it easier to answer with a commercial solver. We demonstrate the high effectiveness of our suggested method, which is based on joint optimization, through comprehensive numerical studies. We start by outlining the restrictions on where to put data and tasks, load distant data, and QoS. Next, we provide a mixed-integer nonlinear programming form of the full statement of the cost reduction problem. The rate-constrained calculation and transfer in large data processing are described by two-dimensional Markov chain process. Create the cost-minimization issue using a MINLP approach based on the closed-form equation.For purpose of resolving the MINLP's difficulty, linearize it as a MILP issue.

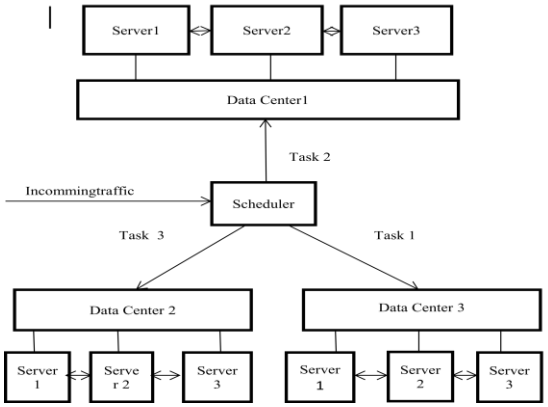


Fig: System Architecture

The suggested method has several benefits, including the following:

- Low cost for strong computational input.
- Simplifying system operations improves system dependability.
- There is minimal energy consumption.
- The issue of data placement, task assignment, and data movement can all be jointly optimised;
- Work completion time is kept to a minimum.

Following section explains several stages that are involved in putting the suggested technique into practice:

Data Uploading

Choose the large data and put it in the hadoop system so that 'map reduce' can be performed on it. The VM server address should be loaded with the info. Data division is carried out after file uploading in preparation for further processing.

Segmentation

By dividing packets contained in incoming Ethernet frames into distinct buffers, packet segmentation enhances network efficiency. When a data packet is bigger than the network's highest allowed transmission unit, segmentation may be necessary. The packet processing engine was created especially to handle network data. Equipment at the periphery of the network processes data more slowly than central network equipment.Lastly, a client endpoint's terminals are connected by the access network. Any component of the network can have a packet processing device installed, including high-end core routers and LAN switches. The system's configurable components, or NPs, are what give it its versatility. Additionally, a number of stacked network protocols ensure that it can accomplish the speed.

Task Assignment

The data centre should be chosen based on the number of computers housed there and their computing and storing capabilities. The identification of the data centre is crucial for reducing the running costs of the computers housed in each data centre. When each data centre has more computers available, within the same data center, data chunks may be stored. The distribution of duties won't change even if there are more computers. The data center with the highest number of enabled computers should be allocated the task. Task assignment is profoundly impact running expenses of data centre. For efficient data handling, tasks are distributed to data centres based on their proximity. Big data activities will demand a certain amount of storage for each data chunk.

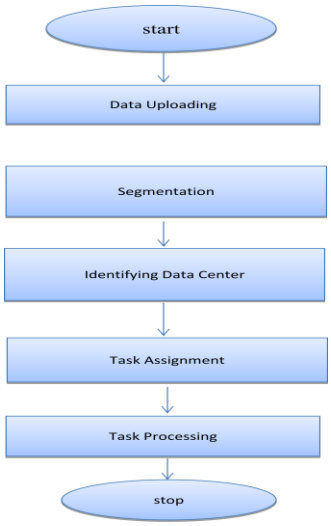


Fig: Data Flow Diagram

Data Loading

In order to reduce transmission costs while maintaining user experience, a data location on the computers and the quantity of load capability given to each file duplicate is necessary. Volley is used by cloud services to send records of datacenter queries. On the basis of customer locations and trends of data access, Volley examines the records and generates suggestions for migration back to the cloud service. Create Min Copy sets, a data replication placement strategy that separates data replication from dissemination for enhancing characteristics of data longevity in dispersed data centres. In recent times, Jin et al. proposed a combined optimization strategy that concurrently maximises energy savings by optimising the location of virtual machines (VMs) and network traffic routing.

Processing of Task

The high computational server shouldn't handle a data block with a small population. Because it raises server operating costs, storage waste, and transfer costs. The number of computers in the data centers that can manage the workload determines how much data is processed.

Evaluation Process

Utilizing MILP formulation, we demonstrate the performance outcomes of our combined optimization method. Consider various total server counts when evaluating server costs, connectivity costs, and overall costs.

4. RESULT

Virtualization is used by cloud computing to effectively supply resources, although provisioning

bandwidth and VMs simultaneously has not been the subject of earlier study. This calls for a collaborative strategy that integrates VMs and bandwidth allocation, makes use of stochastic programming, shrinks the problem space using a scenario tree reduction technique, and analyses price sensitivity. To address issues like data location, task assignment, and data routing, the cost reduction issue of large data processing is tackled in a way called mixed integer nonlinear programming (MINLP).

The significant computational difficulty of solving MINLP is addressed in this study through a joint-optimization based approach. A scenario tree reduction method is used to determined an appropriate heuristic, a stochastic programming formulation is used to determine how to best reserve virtual machines and bandwidth in a cloud computing environment, and a sensitivity analysis is used to determine how sensitive the solution is to parameter changes. The two-dimensional Markov chain process calculates the anticipated task time of completion in closed form and represents rate-constrained calculation and transmission in large data processing.

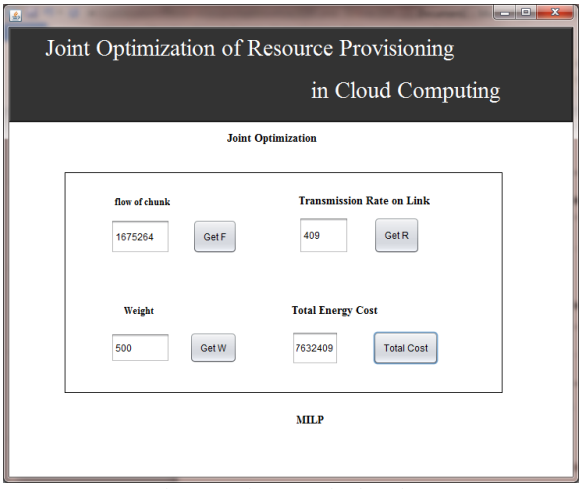
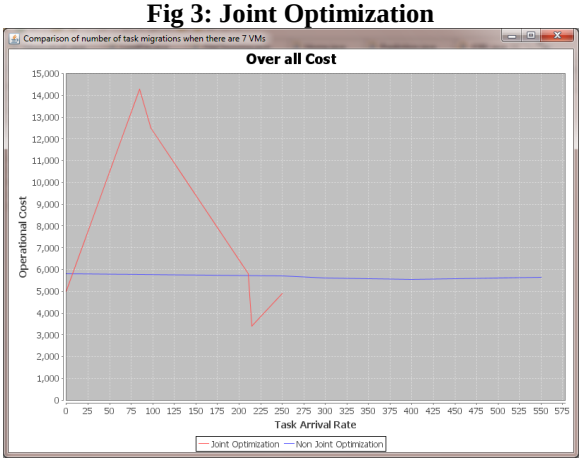


Fig 3: Joint Optimization



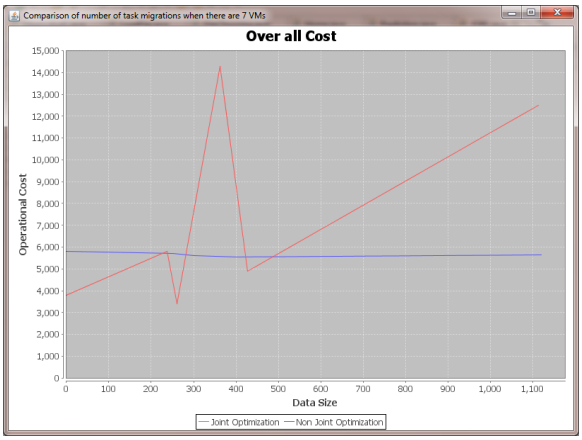


Fig 4: Performance Analysis

5. CONCLUSION

In this article, we've given a model for stochastic programming that will help you secure virtual computers and data in a cloud computing environment in the best possible way. Despite unclear demand, the formulation chooses the best course of action when reserving resources over numerous time periods. The interdependence of VM and bandwidth is used to demonstrate the need for combined management. Since the issue space in a real-life application can easily grow to be very big, we used a scenario tree reduction method for identifying a workable heuristic. We have discovered that, despite the fact that shrinking the scenario tree raises overall costs, the algorithm still provides an acceptable cost-performance balance.Lastly, in order to determine how sensitive the answer is to parameter changes, we conducted a sensitivity analysis of the stochastic programming problem. We have demonstrated that sensitivity analysis is helpful to both customers and providers in establishing system parameters because the optimization has sizable intervals.

6. FUTURE ENHANCEMENT

Future research should increase the issue formulation's reality by taking into account elements like random network latency and VM migration. The model could be expanded for taking service structure and interactions between VMs into account. It should be thought about making additional improvements to processing speed, like using a distributed strategy. The sensitivity analysis findings are encouraging, but they can be expanded to look more closely at how system components interact, like the importance of limits and the importance of variables.

REFERENCE

[1] X. Fan, W.-D. Weber, and L. A. Barroso, "Power Provisioning for A Warehouse-sized Computer," in Proceedings of the 34th Annual International Symposium on Computer Architecture (ISCA). ACM, 2007, pp. 13–23.

[2] H. Shachnai, G. Tamir, and T. Tamir, "Minimal Cost Reconguration of Data Placement in a Storage Area Network"- 2010

[3] B. Hu, N. Carvalho, L. Laera, and T. Matsutsuka, "Towards big linked data: a large-scale, distributed semantic data storage"-2012

[4] Z. Liu, Y. Chen, C. Bash, A. Wierman, D. Gmach, Z. Wang,M. Marwah, and C. Hyser, "Renewable and Cooling Aware Workload Management for Sustainable Data Centers," in Proceedings ofInternational Conference on Measurement and Modeling of Compute Systems (SIGMETRICS). ACM, 2012, pp. 175–186.

[5] S. Govindan, A. Sivasubramaniam, and B. Urgaonkar, "Benefits and Limitations of Tapping Into Stored Energy for Datacenters" – 2011

[6] R. Raghavendra, P. Ranganathan, V. Talwar, Z. Wang, and X. Zhu,"No "Power" Struggles: Coordinated Multi-level Power Management for the Data Center," in Proceedings of the 13th InternationalConference on Architectural Support for Programming Languages andOperating Systems (ASPLOS). ACM, 2008, pp. 48–59.

[7] L. Rao, X. Liu, L. Xie, and W. Liu, "Minimizing Electricity Cost:Optimization of Distributed Internet Data Centers in a MultiElectricity-Market Environment," in Proceedings of the 29th International Conference on Computer Communications (INFOCOM). IEEE,2010, pp. 1–9.

[8] Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. L. Andrew,"Greening Geographical Load Balancing," in Proceedings of International Conference on Measurement and Modeling of ComputerSystems (SIGMETRICS). ACM, 2011, pp. 233–244.

[9] R. Urgaonkar, B. Urgaonkar, M. J. Neely, and A. Sivasubramaniam, "Optimal Power Cost Management Using Stored Energyin Data Centers," in Proceedings of International Conference onMeasurement and Modeling of Computer Systems (SIGMETRICS).ACM, 2011, pp. 221–232.

[10] B. L. Hong Xu, Chen Feng, "Temperature Aware WorkloadManagement in Geo-distributed Datacenters," in Proceedings ofInternational Conference on Measurement and Modeling of ComputerSystems (SIGMETRICS). ACM, 2013, pp. 33–36.

[11] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," Communications of the ACM, vol. 51, no. 1,pp. 107–113, 2008.

[12] S. A. Yazd, S. Venkatesan, and N. Mittal, "Boosting energy efficiency with mirrored data block replication policy and energyscheduler", SIGOPS Oper. Syst. Rev., vol. 47, no. 2, pp. 33–40, 2013.

[13] Marshall and C. Roadknight, "Linking cache performance to user behaviour" Computer Networks and ISDN Systems, vol. 30, no. 223, pp. 2123 – 2130, 1998.

[14] H. Jin, T. Cheocherngngarn, D. Levy, A. Smith, D. Pan, J. Liu and N. Pissinou, "Joint Host-Network Optimization for EnergyEfficient Data Center Networking," in Proceedings of the 27thInternational Symposium on Parallel Distributed Processing (IPDPS), 2013, pp. 623–634.

[15] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs, "Cutting the Electric Bill for Internet-scale Systems," in Proceedings of the ACM Special Interest Group on Data Communication(SIGCOMM). ACM, 2009, pp. 123–134.