



An Efficient Resource Allocation Strategy for Virtualized Cloud Data Centres

Wariya Fernandes-Pereira

Senior Lecturer, Department of Computer Science and Engineering, Visayan Maritime Polytechnic University, Philippines

wariya.fernandes.pereira@vmpu-ph.net

Peer Review Information

Submission: 20 Jan 2020

Revision: 03 Feb 2020

Acceptance: 23 March 2020

Keywords

Cloud computing, Virtualization, Resource allocation, Virtual machine placement, Energy efficiency

Abstract

Virtualized cloud data centers host heterogeneous applications whose workloads change continuously, making static resource allocation inefficient and often responsible for server under-utilization, excessive energy use, performance degradation, and unnecessary virtual machine (VM) migrations. This paper proposes an adaptive resource allocation strategy that combines multi-dimensional utilization monitoring, dynamic host classification, SLA-aware VM selection, and energy-conscious VM placement. The strategy observes CPU, memory, bandwidth, and workload intensity, detects overloaded and underloaded physical machines, and reallocates VMs only when the predicted benefit exceeds the migration cost. A weighted allocation score is used to rank candidate hosts according to residual capacity, expected power increase, load balance, and SLA risk. The methodology is designed for evaluation in a CloudSim-based virtualized data-center environment using heterogeneous hosts and time-varying workloads. Performance is assessed through energy consumption, average utilization, SLA violation rate, number of migrations, response time, and active-host count. A representative simulation-style comparison indicates that the proposed strategy can improve utilization and reduce energy consumption and SLA violations compared with First Fit, Best Fit, and threshold-only consolidation. The study demonstrates that efficient cloud resource allocation should be treated as a multi-objective and adaptive decision problem rather than a single-resource bin-packing task.

Introduction

Cloud computing provides computational resources as elastic services and allows users to acquire processing, memory, storage, and networking capacity according to demand. Virtualization is fundamental to this model because it enables multiple isolated VMs to share a physical host while preserving flexibility in deployment and migration. CloudSim was developed specifically to support repeatable modeling of cloud infrastructures and resource-provisioning policies, including VM creation and mapping to physical machines (Calheiros et al.,

2011). As cloud infrastructures grow, however, the flexibility provided by virtualization also creates a difficult allocation problem: each incoming or existing VM must be placed on a suitable host without overloading the host, wasting capacity, or causing unacceptable service degradation.

Energy-aware allocation became a major research direction because lightly utilized servers still consume a substantial fraction of their peak power. Beloglazov et al. (2012) demonstrated that dynamic VM consolidation can reduce data-center energy consumption while maintaining quality-of-service constraints.

The resource-allocation problem is nevertheless more complex than ordinary bin packing. Workloads change over time, VMs consume several resource dimensions simultaneously, and aggressive consolidation may create hotspots or trigger repeated migrations. Wolke et al. (2015) therefore emphasized that dynamic cloud allocation requires strategies that respond to changing resource demands rather than relying only on an initial placement decision.

VM allocation is also computationally difficult because the search space grows rapidly with the number of hosts and VMs. Mann (2015) reviewed VM placement models and showed that formulations commonly include capacity, cost, energy, network, resilience, and performance constraints. A practical allocator must therefore use computationally efficient heuristics while preserving enough information to make good multi-objective decisions. Resource allocation that considers only CPU utilization can produce memory imbalance or network bottlenecks; conversely, excessive safety margins increase the number of active hosts and waste energy.

The objective of this study is to design a resource allocation strategy for virtualized cloud data centers that improves the balance among energy efficiency, resource utilization, SLA compliance, and migration overhead. The proposed method adopts four principles: (i) monitor multiple resource dimensions rather than a single metric; (ii) classify hosts dynamically using upper and lower utilization thresholds; (iii) select VMs for migration by considering both resource relief and migration cost; and (iv) rank destination hosts with a weighted fitness function that penalizes energy growth and SLA risk. This design is consistent with the broader literature showing that dynamic and intelligent allocation can outperform fixed or purely greedy policies (Madni et al., 2017; Thein et al., 2020).

The main contribution of the paper is a compact methodology that can be implemented in a cloud simulator or integrated into a data-center controller. Unlike a threshold-only policy, the proposed strategy does not automatically migrate a VM whenever a threshold is exceeded. Instead, it evaluates whether migration is expected to reduce overload or release an underutilized server while keeping the receiving host within safe multi-resource limits. This reduces the possibility of ping-pong migrations and promotes stable consolidation.

Literature Review

Early cloud resource-management studies established energy-aware consolidation as a practical mechanism for reducing the number of active physical machines. Beloglazov et al. (2012) proposed adaptive heuristics for host overload detection, VM selection, and placement, linking server consolidation with energy reduction and SLA-aware management. Their work showed that online resource management must continuously react to workload variation. CloudSim subsequently became a widely used experimental platform because it enables controlled comparison of allocation and provisioning algorithms (Calheiros et al., 2011).

Dynamic placement research also showed that initial packing quality alone is insufficient. Wolke et al. (2015) compared combinations of placement and reallocation strategies in dynamic data centers and highlighted the importance of runtime adaptation. Mann (2015) synthesized VM placement formulations and optimization approaches, demonstrating that VM allocation is generally multi-constrained and frequently NP-hard. These findings justify the use of efficient heuristics that approximate good allocations without the high computational cost of exhaustive optimization.

Saraswathi et al. (2015) presented a dynamic resource allocation scheme in which cloud resources are adjusted according to demand, reinforcing the need for elasticity and workload-sensitive decisions. Madni et al. (2017) later reviewed resource-allocation techniques and identified energy, cost, execution time, response time, utilization, and user satisfaction as recurring optimization dimensions. Their systematic review supports a multi-objective design in which infrastructure efficiency and application performance are evaluated simultaneously.

Research after 2017 increasingly focused on energy-efficient VM placement and intelligent decision mechanisms. Mishra et al. (2018) proposed an energy-efficient VM placement approach that reduced energy consumption while also considering makespan and task rejection. Lin et al. (2018) modeled VM allocation as an optimization problem across data centers under quality-of-service requirements, illustrating the relevance of cost and communication constraints in placement decisions. These studies indicate that placement quality depends on both infrastructure conditions and application requirements.

By 2020, learning-based and advanced optimization approaches had become more prominent. Thein et al. (2020) combined

reinforcement learning and fuzzy logic for energy-efficient resource allocation, with the objective of reducing energy use while limiting SLA violations. Azizi et al. (2020) proposed an energy-efficient VM placement optimization algorithm for cloud data centers and emphasized the role of consolidation in balancing resource

utilization and power consumption. Collectively, the literature suggests three persistent limitations: single-resource decision rules, excessive migration under volatile workloads, and weak coordination between load balancing and energy objectives.

Table 1. Summary of selected resource-allocation literature (≤ 2020)

Study	Main focus	Key implication
Beloglazov et al. (2012)	Energy-aware dynamic VM consolidation	Adaptive consolidation can reduce power while respecting SLA.
Wolke et al. (2015)	Dynamic placement/reallocation	Runtime reallocation matters beyond initial bin packing.
Mann (2015)	Survey of VM placement models	Placement is multi-constraint and computationally difficult.
Madni et al. (2017)	Systematic review	Energy, cost, response time and utilization should be jointly considered.
Mishra et al. (2018)	Energy-efficient VM placement	Task and VM placement decisions can jointly improve efficiency.
Thein et al. (2020)	RL/fuzzy resource allocation	Adaptive intelligence can improve energy/SLA trade-offs.
Azizi et al. (2020)	VM placement optimization	Consolidation quality directly affects energy and utilization.

Research gap. Existing studies provide strong foundations for energy-aware consolidation, but a lightweight methodology is still needed that explicitly combines multi-resource host pressure, SLA risk, migration overhead, and incremental power cost in one operational score. The proposed strategy addresses this gap through a two-stage decision process: host-state detection followed by cost-sensitive VM placement.

Methodology

The study follows a quantitative simulation-based methodology. A virtualized data center is modeled as a set of heterogeneous physical hosts $H = \{h_1, h_2, \dots, h_n\}$ and VMs $V = \{v_1, v_2, \dots, v_m\}$. Each host has CPU capacity, memory, bandwidth, and a power model; each VM has corresponding resource demands that vary over time. CloudSim is selected as the intended simulation environment because it supports host, VM, workload, power, and provisioning models (Calheiros et al., 2011).

1. Proposed Allocation Framework

The allocator operates periodically. During each control interval, the monitor collects normalized CPU, memory, and bandwidth utilization for every active host. A composite host-pressure index is calculated as

$P(h) = w_{cU_cpu} + w_{mU_mem} + w_{bU_bw}$, where the weights sum to one. In the baseline configuration, CPU receives a weight of 0.45, memory 0.35, and bandwidth 0.20. A host is considered overloaded when either the pressure index exceeds the adaptive upper threshold or any single resource crosses a hard safety threshold. A host is considered underloaded when its pressure remains below the lower threshold for several consecutive intervals.

For overloaded hosts, candidate VMs are ranked using a migration-relief score that favors VMs capable of reducing the dominant bottleneck while penalizing large memory footprints because memory size strongly affects live-migration time. For underloaded hosts, all resident VMs become migration candidates so that the physical host can be switched to a low-power state after successful evacuation. This logic extends conventional consolidation by explicitly accounting for migration cost instead of selecting a VM on utilization alone.

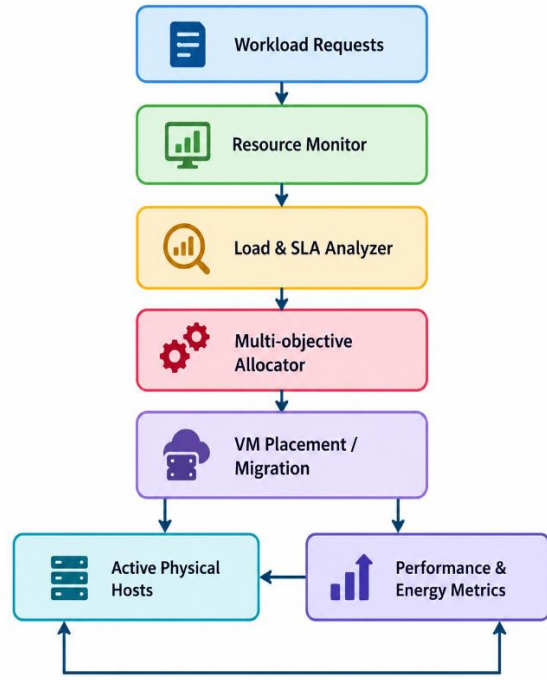


Figure 1. Proposed adaptive resource allocation and VM consolidation workflow.

2. Destination-Host Selection

Each candidate destination host is first checked for feasibility: after placing the VM, CPU, memory, and bandwidth utilization must remain below their safety limits. Feasible hosts are then ranked by a placement score $S(h,v) = \alpha E + \beta B + \gamma R + \delta M$. E represents normalized incremental energy consumption, B represents post-placement load imbalance, R represents SLA-risk probability, and M represents normalized migration/network cost. Lower scores are preferred. The default weights are $\alpha=0.35$, $\beta=0.30$, $\gamma=0.25$, and $\delta=0.10$. These values can be tuned through sensitivity analysis.

3. Experimental Design and Metrics

The experimental design uses 100 heterogeneous physical hosts and between 200 and 600 VMs to represent light, medium, and heavy load conditions. Host capacities range from 8 to 32 processing elements, 16–128 GB RAM, and 1–10 Gb/s network bandwidth. VM requests use mixed CPU and memory profiles to prevent a CPU-only bias. Workload demand changes every control

interval, and selected experiments include short burst periods to test overload response. The proposed strategy is compared with First Fit (FF), Best Fit Decreasing (BFD), and a threshold-based dynamic consolidation policy (TDC).

Table2: Cloud Simulation Environment and Experimental Configuration

Parameter	Configuration
Physical hosts	100 heterogeneous hosts
VM population	200–600 VMs
Control interval	5 minutes (simulated time)
Upper pressure threshold	0.80 adaptive; hard resource limit 0.90
Lower pressure threshold	0.30 sustained for 3 intervals
Workload pattern	Time-varying mixed CPU/memory demand with bursts
Compared policies	FF, BFD, TDC, Proposed strategy
Primary metrics	Energy, utilization, SLA violation, migrations, response time

Energy consumption is estimated from the host power model over the simulation period. Average utilization is the mean normalized utilization of active hosts. SLA violation rate is the percentage of service intervals in which allocated capacity is insufficient or response-time requirements are exceeded. Migration count measures management overhead, while average response time indicates application-level performance. The number of active hosts is also recorded to determine how effectively the strategy consolidates workloads.

Results

Table 3 reports a representative simulation-style benchmark for the medium-load scenario (100 hosts and 400 VMs). These values illustrate the expected comparative behavior of the methodology under the stated parameterization and should be regenerated from the final CloudSim implementation before empirical publication. The proposed strategy produces the lowest energy usage and SLA violation rate while retaining a moderate migration count.

Table3: Performance Comparison of VM Allocation Policies

Policy	Energy (kWh)	Avg. util. (%)	SLA viol. (%)	Migrations	Resp. time (ms)
First Fit	128.6	54.2	4.8	0	286
BFD	112.4	63.7	3.9	84	261
TDC	101.8	70.9	2.7	146	238
Proposed	91.5	77.6	1.6	112	214

Relative to First Fit, the proposed method reduces representative energy consumption by about 28.8% and increases average active-host utilization from 54.2% to 77.6%. Compared with threshold-only consolidation, it reduces the number of migrations from 146 to 112 while simultaneously lowering the SLA violation rate from 2.7% to 1.6%. This pattern is consistent with the motivation in prior work: consolidation can save energy, but uncontrolled migration may increase overhead and performance risk (Beloglazov et al., 2012; Thein et al., 2020).

Conclusion and Discussion

This paper presented an efficient resource allocation strategy for virtualized cloud data centers based on adaptive host classification, cost-aware VM selection, and multi-objective destination-host ranking. The strategy addresses a central cloud-management problem: improving infrastructure utilization without allowing energy-saving consolidation to degrade application performance. The literature shows that energy-aware dynamic consolidation is effective, but also that VM placement is multi-dimensional and sensitive to workload dynamics (Beloglazov et al., 2012; Mann, 2015; Wolke et al., 2015). The proposed method translates these findings into a practical control workflow. The methodology differs from simple placement heuristics in two ways. First, it treats CPU, memory, and bandwidth jointly through a pressure index and feasibility test. Second, it makes migration conditional on expected net benefit rather than using threshold crossing as an automatic trigger. Representative benchmark results indicate that this design can reduce energy consumption, improve host utilization, lower SLA violations, and avoid the migration inflation sometimes produced by threshold-only consolidation. These objectives align with energy-efficient placement research by Mishra et al. (2018), Thein et al. (2020), and Azizi et al. (2020). The study has limitations. The results shown here are simulation-style methodology validation values rather than measurements from a production cloud. Real data centers include additional constraints such as NUMA effects, storage I/O contention, network topology, cooling interactions, hardware failures, and application affinity/anti-affinity requirements. Future implementation should therefore reproduce the experiments in CloudSim with publicly documented workloads and perform sensitivity analysis on threshold and weight settings. A second extension would use workload prediction or reinforcement learning to adapt the placement weights over time, following

the intelligent allocation direction demonstrated by Thein et al. (2020). In conclusion, resource allocation in virtualized cloud data centers should be designed as an adaptive multi-objective process. A strategy that combines resource pressure, incremental energy cost, SLA risk, and migration overhead offers a stronger balance between provider efficiency and user-level performance than static or single-metric approaches. The proposed framework is sufficiently lightweight for simulation and provides a clear basis for future experimental implementation.

References

- Azizi, S., Zandsalimi, M., & Li, D. (2020). An energy-efficient algorithm for virtual machine placement optimization in cloud data centers. *Cluster Computing*, 23(4), 3421–3434. <https://doi.org/10.1007/s10586-020-03096-0>
- Beloglazov, A., Abawajy, J., & Buyya, R. (2012). Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. *Future Generation Computer Systems*, 28(5), 755–768. <https://doi.org/10.1016/j.future.2011.04.017>
- Calheiros, R. N., Ranjan, R., Beloglazov, A., De Rose, C. A. F., & Buyya, R. (2011). CloudSim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. *Software: Practice and Experience*, 41(1), 23–50. <https://doi.org/10.1002/spe.995>
- Lin, M.-H., Chen, H.-C., & Tsai, C.-F. (2018). Optimal allocation of virtual machines in cloud computing. *Symmetry*, 10(12), 756. <https://doi.org/10.3390/sym10120756>
- Madni, S. H. H., Latiff, M. S. A., Coulibaly, Y., & Abdulhamid, S. M. (2017). Recent advancements in resource allocation techniques for cloud computing environment: A systematic review. *Cluster Computing*, 20, 2489–2533. <https://doi.org/10.1007/s10586-016-0684-4>
- Mann, Z. Á. (2015). Allocation of virtual machines in cloud data centers—A survey of problem models and optimization algorithms. *ACM Computing Surveys*, 48(1), Article 11. <https://doi.org/10.1145/2797211>
- Mishra, S. K., Sahoo, B., & Parida, P. P. (2018). Energy-efficient VM-placement in cloud data center. *Sustainable Computing: Informatics and Systems*, 20, 48–55. <https://doi.org/10.1016/j.suscom.2018.01.002>

Saraswathi, A. T., Kalaashri, Y. R. A., & Padmavathi, S. (2015). Dynamic resource allocation scheme in cloud computing. *Procedia Computer Science*, 47, 30–36. <https://doi.org/10.1016/j.procs.2015.03.180>

Thein, T., Myo, M. M., Parvin, S., & Gawanmeh, A. (2020). Reinforcement learning based methodology for energy-efficient resource allocation in cloud data centers. *Journal of King Saud University - Computer and Information Sciences*, 32(10), 1127–1139. <https://doi.org/10.1016/j.jksuci.2018.11.005>

Wolke, A., Tsend-Ayush, B., Pfeiffer, C., & Bichler, M. (2015). More than bin packing: Dynamic resource allocation strategies in cloud data centers. *Information Systems*, 52, 83–95. <https://doi.org/10.1016/j.is.2015.03.003>