Abstract

Cloud computing has revolutionized computational resource access but introduced complex management challenges that traditional static allocation approaches cannot address, necessitating sophisticated strategies for dynamic workloads, energy efficiency, and service quality guarantees. This thesis presents a comprehensive framework for intelligent cloud resource management addressing three fundamental challenges through interconnected solutions that coordinate optimization across multiple time horizons while maintaining service level agreements and system stability. The first contribution develops proactive resource management through optimal VM placement, employing machine learning-based workload prediction that identifies temporal patterns through clustering and develops specialized models for each workload category, combined with statistical-stochastic frameworks treating resource consumption probabilistically for risk-aware allocation decisions, integrated with heuristic-based and game-theoretic VM placement strategies that achieve superior energy efficiency through optimal consolidation. The second contribution addresses temporal limitations through QoS-aware load balancing implementing probabilistic overload detection using stochastic resource modeling to anticipate deficit situations before SLA violations occur, coupled with intelligent migration optimization algorithms addressing VM selection and destination placement while incorporating cumulative SLA violation tracking for long-term service quality impact assessment, successfully transforming reactive resource management into proactive load balancing. The third contribution completes the framework through energy-efficient dynamic VM consolidation using Resource-Optimized VM Consolidation that employs stochastic load imbalance detection to identify consolidation opportunities while maintaining load balancing compatibility, resource intensity-aware VM distribution through game-theoretic optimization achieving Nash equilibrium solutions balancing energy efficiency with system resilience, and multi-objective optimization integrating energy minimization with migration frequency control. Comprehensive experimental validation using real-world workload traces demonstrates the integrated framework's effectiveness: proactive placement achieves 12% energy reduction with superior consolidation efficiency, SLA-aware load balancing delivers 10-39% reduction in migrations and 17-54% fewer overloaded hosts while maintaining utilization efficiency, and energy-efficient consolidation provides 21-65% reduction in active hosts, 17-46% energy decrease, and 37-70% fewer load imbalances compared to existing approaches, with coordinated optimization achieving simultaneous energy efficiency, service quality maintenance, and resource utilization effectiveness across diverse workload scenarios. The research establishes a stochastic modeling foundation enabling sophisticated decision-making under uncertainty, successfully bridging theoretical optimization with practical deployment requirements through modular design that provides measurable improvements across all performance dimensions, demonstrating that effective cloud resource management requires coordinated optimization rather than isolated solutions and that aggressive energy optimization can be achieved without

compromising system resilience or service quality, ultimately advancing sustainable and efficient cloud computing infrastructure management while enabling providers to deliver operational excellence through simultaneous optimization of competing objectives.