Scholar Publishing Group
International Journal of Neural Network
https://doi.org/10.38007/NN.2025.040107
ISSN 2790-2692 Vol. 4, Issue 1: 57-64

Methods of Load Optimization for Computer Systems Based on Physical Principles

Buqin Wang

Meta Platforms / Infrastructure, Menlo Park, CA, 94025, US

Keywords: Computer system; Load optimization; Principles of Physics; Dynamic scheduling; Heterogeneous computing

Abstract: Traditional computer system load optimization strategies have exposed many shortcomings in dealing with continuously changing workloads, power management, and heat dissipation challenges. This article proposes a new computer system load optimization method based on principles of physics. By real-time monitoring and load prediction, combined with principles of thermodynamics and fluid mechanics, load distribution is optimized to ensure efficient allocation of computing resources while preventing overheating. A collaborative utilization method for heterogeneous computing resources has been proposed to balance power consumption and computing resources, fully tapping into the unique performance of various resources to enhance overall efficiency. Through temperature related load distribution strategies, flexible adaptation between heat dissipation and load has been successfully achieved, thereby reducing energy consumption and improving system reliability. This has opened up innovative ideas and means for improving the efficiency of computer systems, reducing energy consumption, and suppressing temperature rise.

1. Introduction

With the continuous advancement of computer technology, the task of optimizing system load has become increasingly complex. Traditional methods often focus on software level resource scheduling, ignoring the physical characteristics of hardware itself, resulting in uneven allocation of computing resources, high power consumption, and heat dissipation problems. Especially in handling high-performance computing and large-scale distributed systems, the dynamic changes in system load are closely related to the thermodynamic and electrical characteristics of hardware. The load optimization method based on physics principles can adjust the configuration of computing resources, balance power consumption and temperature, improve the efficiency and stability of system operation, enhance the load optimization effect of computer systems, and promote the evolution of systems towards high efficiency and low energy consumption through thermodynamic, fluid dynamics, and electrical principles.

2. The relationship between physics principles and computer systems

The design and operation of computer systems are deeply influenced by the principles of physics,especially in key areas such as thermal management, power control, and resource scheduling.The working process of computer hardware involves physical phenomena such as heat conduction, current flow, and airflow movement, which directly affect the efficiency, reliability, and energy utilization of system operation.

The principles of thermodynamics are crucial in computer systems.The processor and other components generate a large amount of heat during operation, and an increase in temperature not only weakens hardware performance but may also cause malfunctions.By applying thermodynamic principles, the design of heat sinks, fans, and liquid cooling systems can be improved to efficiently dissipate heat and ensure that hardware is at a safe temperature level.At the same time, uneven distribution of hardware loads may cause local temperature increases, so it is necessary to dynamically adjust task allocation through load balancing strategies to avoid overheating situations.

The principles of fluid mechanics affect the air flow and heat dissipation inside computers.By adjusting the airflow path and fan configuration, the cooling performance can be enhanced, heat accumulation can be alleviated, and the system can be ensured to remain stable under heavy load conditions.Fluid dynamics provides a scientific basis for the construction of heat dissipation systems, helping to improve the cooling capacity of computer systems.

In terms of power management, electrical principles play a key role.The power consumption of computer hardware is closely related to changes in current. Excessive power consumption only leads to energy waste and can also cause excessive heat generation.Dynamic Voltage Frequency Adjustment (DVFS) technology is based on electrical principles, which adjusts the voltage and frequency of the processor through real-time changes in the load, thereby achieving a balance between performance and energy consumption.In summary, the principles of physics play an important role in the design and performance optimization of computer systems, effectively promoting system performance, stability, and energy efficiency.

3. Limitations of Current Computer System Load

3.1 The mismatch between static optimization and dynamic changes

Most existing load optimization methods are based on static optimization models, assuming that the load of the computer system is relatively stable over a period of time. However, in reality, load fluctuations are significant, especially in the process of executing diverse tasks, where load changes are particularly noticeable. This static optimization strategy assumes that the load distribution is balanced or predictable, which often does not match the dynamic characteristics of the system during actual operation, resulting in inaccurate resource allocation.

Table 1. Reasons and Effects of Problems in Static Optimization and Dynamic Changes

Problem	Reason	Effect
High load fluctuation	The system load fluctuates continuously with factors such as task types and user demands.	Static optimization cannot predict and adapt to changes in load, resulting in inefficient system operation.
Static assumptions do not conform to reality	Static optimization often assumes stable or periodic load changes, ignoring dynamic loads.	When the load suddenly increases, the system may experience resource overload or idle, which reduces performance.
Lack of real-time adjustment mechanism	The current system relies heavily on pre-set load optimization strategies and lacks dynamic adjustment mechanisms.	Dynamic load cannot be adjusted in real time, resulting in resource waste or partial node overload.

As shown in Table 1, static optimization methods cannot adjust according to real-time changes in load, which may lead to excessive or insufficient resource allocation, thereby affecting the overall performance and response speed of the system. For example, when the system encounters a sudden increase in load, the static optimization model often cannot respond immediately, which may lead to the computing nodes being unable to withstand the pressure and form performance barriers. Meanwhile, when the load is low, static optimization may also result in resource waste and increase energy consumption. In addition, static optimization models do not take into account the sudden nature of tasks, which makes the system unable to quickly self adjust when facing high load challenges, thereby reducing computational efficiency.

3.2 The difficult balance between computing resources and power consumption

In modern computer systems, enhanced computing resources often mean higher energy consumption. In order to improve system performance, it is usually necessary to increase computing resources, such as increasing processor frequency, increasing the number of cores, etc. However, the implementation of these suggestions usually leads to a significant increase in energy consumption, and how to balance the relationship between computing resources and power consumption has become an urgent problem to be solved.

Table 2. Reasons and impacts of issues with computing resources and power consumption

Problem	Reason	Effect
High performance demands increase computing resources	When handling high load tasks, more computing resources are required (such as more cores and higher frequencies).	The increase in computing resources is accompanied by an increase in power consumption, which increases the difficulty of heat dissipation and may lead to overheating.
Difficulty balancing power consumption and performance	Improving performance usually requires higher power consumption, and the two are positively correlated.	Excessive power consumption not only leads to energy waste, but also increases the risk of hardware overheating and reduces system stability.
Power management restrictions	The current hardware and management strategy cannot flexibly adjust power consumption according to load changes.	The system loses control of power consumption under high load, resulting in a sharp increase in energy consumption and excessive heat generation.

As shown in Table 2, when processing high-performance tasks, expanding computing resources inevitably leads to an increase in power consumption. Under high load conditions, the power consumption and heat dissipation requirements of the system significantly increase, and the current power control schemes often cannot effectively solve this problem. Excessive power consumption not only leads to energy abuse, but also promotes heat generation, thereby increasing the working pressure of the cooling system, which may cause system overheating or even hardware failure. In addition, in the hardware design phase, power management functions are often limited, making it difficult to flexibly adjust the matching between power consumption and computing resources based on real-time workloads, which in turn affects the overall efficiency of the system.

3.3 Uneven load distribution leads to high temperature

Unequal load distribution is a common problem in multi-core processors and distributed computing environments. Some computing nodes are overloaded, causing a sudden increase in temperature in specific areas, while the rest of the nodes have lighter loads, failing to balance the computational burden reasonably. This uneven load phenomenon often leads to abnormal temperature rise in the system, affecting stability.

Table 3. Reasons and Effects of Unequal Load Distribution Issues

Problem	Reason	Effect
Uneven distribution of load	The load scheduling algorithm failed to accurately evaluate task load and computing resource requirements, resulting in uneven load distribution.	Partial nodes are overloaded, calculation pressure is concentrated, other nodes are idle, local area temperature is too high, and the heat dissipation system is under high pressure.
Insufficient utilization of computing resources	The resource scheduling algorithm cannot dynamically adjust resource allocation based on the current load, resulting in centralized use of computing resources.	The temperature in the high load area rises rapidly, while other low load areas are not fully utilized, and the heat dissipation system cannot evenly distribute heat.
Uneven heat dissipation	The design of the heat dissipation system is unreasonable, the heat dissipation mechanism has not been optimized, and some nodes have poor heat dissipation.	The heat of high load nodes cannot be dissipated in a timely manner, causing the temperature to rise, resulting in hardware frequency reduction, performance degradation, or failure.

As shown in Table 3, the fundamental reason for uneven load distribution is that the load scheduling algorithm cannot reasonably allocate computing tasks to various computing nodes, resulting in some nodes being overloaded and generating excessive heat. Due to the limitations of the cooling system, this heat was not transferred to other areas in a timely manner, resulting in localized high temperatures. Continuous high temperatures can cause damage to equipment, potentially leading to system failures, reduced performance, and accelerated aging of electronic components.

4. Computer System Load Optimization Method Based on Physics Principles

4.1 Real time load monitoring and prediction

Real time load monitoring and prediction are the foundation for achieving dynamic load optimization. By accurately tracking the workload in real-time, it is possible to quickly grasp the fluctuation of the load and implement reasonable allocation of system resources. Real time load monitoring usually relies on sensors and monitoring software, which collect various indicators such as CPU, memory, network bandwidth, and disk IO to dynamically reflect the load status of the system. The prediction of workload relies on in-depth analysis of historical data, research on task characteristics, and the current system load situation to predict the short-term trend of load, in order to optimize resources in advance. Common load forecasting methods include time series based forecasting, machine learning models, and physics based thermodynamic models. In load forecasting, commonly used mathematical models include linear regression models and exponential smoothing models.Assuming that $\mathrm{L(t)}$ represents the load of the system at time t, based on the load $\mathrm{L(t-1)}$, $\mathrm{L(t-2)},\ldots,$ of the previous period, a simple linear regression model can be constructed for prediction:

$$\mathrm{L(t)} = \alpha\mathrm{L(t-1)} + \beta\mathrm{L(t-2)} + \cdots + \epsilon \tag{1}$$

Among them, α, β is the regression coefficient, and ϵ is the error term. When there are complex load fluctuations, more complex nonlinear regression models or deep learning based prediction algorithms can be used to enhance the model's ability to predict sudden loads.

In addition, based on principles of physics, thermodynamic models can also be used to predict fluctuations in loads. There is a certain relationship between the heat generated by the computing nodes Q in the system and the negative load L. Assuming that the heat release of each computing node is proportional to the load, the following formula can be used:

$$Q = k \cdot L \tag{2}$$

Among them, k is the thermal effect coefficient of the node. According to this model, when the change of working load is detected, the corresponding change of heat energy can be predicted, and then the pressure of the cooling system can be estimated and adjusted in advance. Through real-time monitoring and prediction, the system can make appropriate adjustments before load fluctuations occur, preventing excessive load concentration or overload conditions, ensuring optimized use of computing resources and stable temperature management.

4.2. Collaborative utilization of heterogeneous computing resources

Heterogeneous computing resources refer to different types of hardware resources in computing systems, such as CPU, GPU, FPGA, TPU, etc. By leveraging the collaborative efficiency of these heterogeneous computing resources during task allocation, the computational efficiency and overall performance of the system can be significantly improved. The load optimization method based on physics principles promotes the intelligence level of task allocation by combining the performance characteristics and power consumption characteristics of computing resources. In order to efficiently and concurrently utilize heterogeneous resources, it is necessary to conduct in-depth analysis of the processing performance and power consumption characteristics of various types of hardware. For example, for CPU and GPU, assuming that the computing power of the CPU is linearly related to power consumption:

$$P_{cpu} = \alpha \cdot C_{cpu} \tag{3}$$

Among them, α is a proportional coefficient, and C_{cpu} represents the computing power of the CPU. Similarly, for GPU, a similar formula can be used to describe:

$$P_{gpu} = \beta \cdot C_{gpu} \tag{4}$$

In heterogeneous resource scheduling, the key is to select appropriate resources based on the computational characteristics of the task. For high parallelism tasks (deep learning computing), tasks can be allocated to resources with strong parallel computing capabilities such as GPUs, while for computationally intensive tasks, CPU should be prioritized.

A common heterogeneous computing scheduling algorithm is a task allocation algorithm based on load balancing, which typically uses multi-objective optimization methods to simultaneously consider the utilization and power consumption of computing resources. For example, in the optimization model based on Lagrange multiplier method, assuming that task T needs to be allocated to multiple resources, its goal is to strike a balance between minimizing total power consumption and maximizing computing power:

$$\min \sum_{i=1}^{n} (\lambda_i \cdot P_i + \mu_i \cdot C_i) \tag{5}$$

Among them, λ_i and μ_i are weight coefficients, and P_i and C_i are the power consumption and computing power of resource i, respectively. By precisely adjusting the objective function, intelligent task allocation in heterogeneous computing resource environments has been achieved, thereby improving the computational efficiency of the system and achieving effective control of energy consumption. Through the coordinated use of heterogeneous computing resources, computing tasks can be executed synchronously on different types of resources, thereby significantly enhancing the overall performance of the system, preventing overloading of specific computing nodes, and reducing the risk of thermal energy concentration.

4.3 Temperature sensitive load scheduling

By real-time tracking of temperature and flexible allocation of computing tasks, temperature sensitive workload management aims to prevent system overheating and optimize computing performance. With the widespread application of multi-core processors and high-performance computing systems, the problems of uneven load distribution and local temperature rise have gradually become prominent, posing challenges to the reliability of the system and the service life of equipment. Therefore, a management strategy that combines temperature and load is particularly important. In this strategy, real-time monitoring of the temperature and load status of each computing node is required.Assuming that the load of the i th node in the system is $\mathrm{L_i}$ and the temperature is $\mathrm{T_i}$, the temperature change of the node can be predicted using the following temperature calculation formula:

$$T_i(t) = T_i(t-1) + \alpha \cdot L_i \cdot \Delta t \tag{6}$$

Among them, α is the thermal effect coefficient and $\Delta \mathrm{t}$ is the time interval. Based on the load and temperature changes of each node, the load scheduling strategy can be dynamically adjusted to ensure that the temperature of each node does not exceed the safe upper limit.

The basic strategy of load scheduling is a temperature based feedback control mechanism. On nodes with heavy loads, if the temperature approaches the preset threshold $\mathrm{T_{max}}$, the system will automatically migrate tasks to nodes with lighter loads to avoid local overheating.Specifically, assuming that the total system load $\mathrm{L_{total}}$ needs to be allocated to n nodes, the relationship between the load change rate and temperature change of each node can be scheduled through the following optimization model:

$$\min \sum_{\mathrm{i}=1}^{\mathrm{n}} (\gamma_{\mathrm{i}} \cdot \mathrm{L_i} + \delta_{\mathrm{i}} \cdot \mathrm{T_i}) \tag{7}$$

Among them, γ_{i} and δ_{i} are scheduling weight coefficients, and $\mathrm{L_i}$ and $\mathrm{T_i}$ respectively represent the load and temperature of node i. By solving this optimization problem, it is possible to ensure that the load of the entire system fluctuates within a reasonable range and that the temperature index remains within safe limits. Temperature sensitive load scheduling methods can significantly prevent overheating of computing nodes, thereby enhancing the reliability and operational efficiency of computer systems. In addition, this method also helps to evenly distribute thermal energy, optimize the efficiency of the cooling system, and avoid equipment damage and performance degradation.

5. Conclusion

By introducing principles of physics, real-time monitoring, prediction, and temperature sensitive scheduling strategies based on load, the load optimization methods of computer systems have been significantly improved. By utilizing the fundamental principles of thermodynamics and energy conversion, more accurate predictions of system load fluctuations have been made, and computing resources have been allocated more reasonably. At the same time, efficient management of system energy consumption and temperature control has been achieved, especially in multi-core processors and distributed systems. Physics based load optimization techniques have successfully solved the problems of uneven load distribution, resource surplus, and heat dissipation, significantly improving computational efficiency and system stability. With the increasing demand for computing and continuous technological innovation, these physics based optimization methods will play a more important role in high-performance computing, injecting momentum into the development of intelligent and high-performance computing platforms, technological innovation, and sustainable development.

Reference

[1] *Wu Z, Lu Y, Xu Q, et al. Load optimization control of SJTU-WEC based on machine learning. Ocean engineering, 2022(Apr. 1):249.*

[2] *Wasa K, Talaka K, Dominik Wilczyński. Designing of the Electromechanical Drive for Automated Hot Plate Welder Using Load Optimization with Genetic Algorithm. Materials, 2022, 15(5):1787.*

[3] *Yan X, Zuo H, Hu C, et al. Load Optimization Scheduling of Chip Mounter Based on Hybrid Adaptive Optimization. Modeling and Simulation of Complex Systems, 2023, 3(1):11.*

[4] *Yan X, Zuo H, Hu C, et al. Load Optimization Scheduling of Chip Mounter Based on Hybrid Adaptive Optimization Algorithm. Complex System Modeling and Simulation, 2023, 3(1):1-11.*

[5] *Sansanwal S, Jain N. Inquisitive Genetic-Based Wolf Optimization for Load Balancing in Cloud Computing. applied computer systems, 2023, 28(1):170-179.*

[6] *Q. Hu, "Research on Dynamic Identification and Prediction Model of Tax Fraud Based on Deep Learning," 2025 2nd International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS), Hassan, India, 2025, pp. 1-6.*

[7] *F. Liu, "Architecture and Algorithm Optimization of Realtime User Behavior Analysis System for Ecommerce Based on Distributed Stream Computing," 2025 International Conference on Intelligent Communication Networks and Computational Techniques (ICICNCT), Bidar, India, 2025, pp. 1-8.*

[8] *F. Liu, "Transformer XL Long Range Dependency Modeling and Dynamic Growth Prediction Algorithm for E-Commerce User Behavior Sequence, " 2025 2nd International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS), Hassan, India, 2025, pp. 1-6.*

[9] *Su H, Luo W, Mehdad Y, et al. Llm-friendly knowledge representation for customer support[C]//Proceedings of the 31st International Conference on Computational Linguistics: Industry Track. 2025: 496-504.*

[10] *Lu, C. (2025). Application of Multi-Source Remote Sensing Data and Lidar Data Fusion Technology in Agricultural Monitoring. Journal of Computer, Signal, and System Research, 2(7), 1-6.*

[11] *Ye, J. (2025). Optimization of Neural Motor Control Model Based on EMG Signals. International Journal of Engineering Advances, 2(4), 1-8.*

[12] *Liu, Y. (2025). Use SQL and Python to Advance the Effect Analysis of Financial Data Automation. Financial Economics Insights, 2(1), 110-117.*

[13] *Sun, Q. (2025). Research on Cross-language Intelligent Interaction Integrating NLP and Generative Models. Engineering Advances, 5(4).*

[14] *Zhu, P. (2025). The Role and Mechanism of Deep Statistical Machine Learning In Biological Target Screening and Immune Microenvironment Regulation of Asthma. arXiv preprint arXiv:2511. 05904.*

[15] *Liu, B. (2025). Design and Implementation of Data Acquisition and Analysis System for Programming Debugging Process Based On VS Code Plug-In. arXiv preprint arXiv: 2511. 05825.*

[16] *Ding, J. (2025). Research On CODP Localization Decision Model Of Automotive Supply Chain Based On Delayed Manufacturing Strategy. arXiv preprint arXiv:2511. 05899.*

[17] Wu Y. Software Engineering Practice of Microservice Architecture in Full Stack Development: From Architecture Design to Performance Optimization. 2025.

[18] Wu Y. Optimization of Generative AI Intelligent Interaction System Based on Adversarial Attack Defense and Content Controllable Generation. 2025.

[19] Sun J. Quantile Regression Study on the Impact of Investor Sentiment on Financial Credit from the Perspective of Behavioral Finance. 2025.

[20] Wang Y. Application of Data Completion and Full Lifecycle Cost Optimization Integrating Artificial Intelligence in Supply Chain. 2025.