Modelling and Optimizing for Distributed Systems with limited Resource Migration Capacity

Haipeng Zhang*, Ningning Lu, Ruiliang Song and Jie Li
Beijing R&D Center, The 54th Research Institute of CETC, Beijing, China

*Corresponding author email: zhpeng@mail.ustc.edu.cn

Abstract. This paper discusses the resource scheduling of some distributed systems, such as logistics system and the mobile system. By analysing the characteristic of the resource allocation, we found the resource or load of the distributed systems has limited migration capacity, but the load can be transmitted between neighbour nodes. By diverting resource or load between neighbour nodes, the system can reach a more balanced state. Then a model, shown as a constrained optimization problem, is presented to describe the load balancing problem of the distributed systems. The constrained terms are processed by the penalty function, and then an algorithm based on penalty function is presented to obtain an optimal solution of the constrained optimization problem. Finally, we design some experiments to verify the load model and the optimization algorithm proposed in this paper.

1. Introduction
Distributed systems[1-3], which may have different structures, resources, operational modes, service strategies, etc., have been studied and applied for decades. By moving services from central node to distributed nodes, which are more close to the clients, the systems can quickly respond to the requests and provide much better services. The nodes in the systems may have different software resources, hardware resources, or physical resource, but how to use the resources more fair and efficient always attracts many attention from the academia and the industry. Generally, the optimal resource utilization in the distributed system is NP-hard, the traditional algorithms to solve the problem is based on the basis of equity, such as Round Robin and modified Round Robin [4, 5]. The classical Round Robin algorithm, which means that the node in the system has the same chance to be selected, is more suitable for the system with homogeneous nodes[4]. The modified Round Robin algorithm is designed for the system with heterogeneous nodes, which means the node with different capacity may have different possibility to be selected to provide services [5]. Be differ from traditional algorithms, the algorithm based on heuristic methods is another way to solve the resource utilization problem. Inspired by social spider, a meta heuristic algorithm was proposed in the work [6] to minimize overall makes pan with effective load balancing. The meta heuristic algorithm can also be used to find the task-to-virtual machine mapping for the cloud system [7].

The resource models for the system in the works are based on an assumption that the resource has unlimited migration capacity. However, the assumption does not apply to all conditions, such as the frequency resource of the base station in the mobile system[8].

The rest of this paper is organized as follows. In Section 2, the distributed systems are analysed, then a load model is proposed, which is shown as a constrained optimization problem. In Section 3, a load optimization algorithm based on penalty function is presented to solve the constrained optimization...
problem. In Section 4, some simulations are designed to verify the model and the load optimization algorithm proposed in this paper. Finally, the conclusions are presented in Section 5.

2. System Model

In this section, we will analyse the character of the distributed systems, and then a load model with limited resource migration capacity is introduced. There are many reasons for limited resource migration capacity. First, the resource cannot be transited, for example, the radio frequency resource in the mobile system. The base station accesses users by air interface, which use the radio frequency resource according to different physical layer protocol, such as FDMA, TDMA, CDMA. The base station cannot transmit the radio frequency resource to others, but the station has an area where the users can handover to the neighbour station, as shown in figure 1.

![Figure 1. The mobile system.](image1)

Second, the resource is limited to transmit for QoS, for example, the streaming service system, which has different servers and provide services to clients spread all over the world. The proxy servers are located close enough to provide services for the clients to meet the QoS requirement. When the current proxy server has too many clients to serve, it can transmit the client to other proxy server which can meet the QoS requirement, as shown in figure 2.

![Figure 2. The streaming service system.](image2)

Third, the cost of unlimited resource migration is too high. The logistics system has many warehouses distributed across the country, as shown in figure 3. The warehouse provides goods to clients who are close to the warehouse. The transportation of goods is dependent on freight automobiles, but the long-distance transport should be avoided for increasing cost. Therefore, the transportation between warehouses should be limited.
Based on the analysis listed above, the resources or the load shifted between neighbour nodes can be acceptable. Then a model is presented to describe the load model, in which the resource or the load can be shifted between neighbour nodes. A detailed definition of parameters used and their description are shown as follows.

- \( N \) Number of the nodes in the system.
- \( M_{ij} \) The load which can be transferred between node \( i \) and node \( j \), \( i \neq j \), \( M_{ij} = M_{ji} \).
- \( x_{ij} \) The load which can be migrated from node \( i \) to node \( j \), \( i \leq j \). If \( i = j \), \( x_{ij} \) denotes that node \( i \) can’t transfer to others. \( M_{jj} - x_{ij} \) denotes the load which can be migrated from node \( j \) to node \( i \).
- \( x_i \) The load of node \( i \), \( x_i = \sum_{m=1}^{i-1} (M_{mi} - x_{mi}) + \sum_{m=i}^{N} x_{im} \)

Generally, the variance of load is used to evaluate the load of the system. The lower the variance of numbers is, the better the load balancing state is. Furthermore, the system will be in a load balancing state when \( \sum_{i=1}^{N} x_i^2 \) gets the minimum, which means the load of different nodes has a minimum deviate degree. Then a theoretical load model to describe the distributed system with limited resource migration capacity is presented as follows.

\[
\min \sum_{i=1}^{N} x_i^2
\]

s.t. \( 0 \leq x_{ij} \leq M_{ij}, 1 \leq i < j \leq N \) (1)

As shown in equation (1), the load balancing problem of distributed system has turn into a constrained optimization problem. But the problem is quite difficult to obtain the explicit solution.

### 3. A Novel Algorithm Based on Penalty Function

Based on the above analysis, we have got that the load balancing problem is shown as a constrained optimization problem. The penalty function method is often used to solve the constrained optimal problem, which will transform the problem into an unconstrained problem[9]. A novel load balancing algorithm based on penalty function is proposed to solve the problem shown in equation (1) in this section.

Define the auxiliary function to transform the constrained optimal problem shown in equation (1) into an unconstrained problem.

\[
F(x, \sigma) = \sum_{i=1}^{N} x_i^2 + \sigma \left( \sum_{1 \leq i < j \leq N} \left[ \max(0, -x_{ij}) \right]^2 + \sum_{1 \leq i < j \leq N} \left[ \max(0, x_{ij} - M_{ij}) \right]^2 \right) 
\]

\[
= \sum_{i=1}^{N} \left( \sum_{m=1}^{i-1} (M_{mi} - x_{mi}) + \sum_{m=i}^{N} x_{im} \right)^2 + \sigma \left( \sum_{1 \leq i < j \leq N} \left[ \max(0, -x_{ij}) \right]^2 + \sum_{1 \leq i < j \leq N} \left[ \max(0, x_{ij} - M_{ij}) \right]^2 \right) \quad (2)
\]
\( \sigma \) denotes a huge number, which is used to avoid the value of \( x_j \) choosing a value inaccurately. Based on the model and the auxiliary function shown in equation (1, 2), a novel algorithm based on penalty function to solve the load balancing problem is introduced in Table 1.

**Table 1.** An algorithm based on penalty function.

| Step 1 | Given the initial value \( x^0, x^0 = |x_j^0| \); |
|--------|--------------------------------------------------|
| Step 2 | Initialize the value of \( \sigma, c, \epsilon, k = 1; \) |
| Step 3 | Solve \( \min F(x_j, \sigma_j) \), obtain the minimum value \( x^d \); |
| Step 3.1 | Set \( \frac{\partial F(x_j, \sigma_j)}{\partial x_j} = 0 \), obtain the value of \( x_j \) |
| Step 3.2 | If \( x_j < 0 \), set \( x_j = 0 \) |
| Step 3.3 | If \( x_j > M_j \), set \( x_j = M_j \) |
| Step 4 | If \( \max_{i,j} |x^d_i - x^d_j| < \epsilon \), stop the algorithm, \( \sigma_{k+1} = c \cdot \sigma_k, k = k + 1 \), go to Step 2 |

Where \( c \) denotes amplification factor, \( \epsilon \) denotes tolerated error.

**4. Validation**

In this section, we design an experiment to validate the load balancing algorithm proposed in Section 3. We use the distributed system scenario as shown in figure 4, and which parameters are set as shown in Equation (3) and Figure 5.

\[
\begin{align*}
&x_1 = 100, x_2 = 100, x_3 = 100 \\
&x_4 = 100, x_5 = 480, x_6 = 80 \\
&x_7 = 80, x_8 = 60, x_9 = 30 \\
&x_{10} = 60, x_{11} = 30, x_{12} = 30, x_{13} = 15
\end{align*}
\]

(3)

![Figure 4. The validation condition with 13 nodes.](image-url)
The clients, which are considered are the load of the node, are distributed evenly around the node. We use the algorithm based on penalty function proposed in Section 3 to optimize the system’s load performance. After optimization, the load of the system is shown as Equation (4) and Figure 6.

\[
\begin{aligned}
    x_1 &= 102, x_2 = 103, x_3 = 116 \\
    x_4 &= 108, x_5 = 355, x_6 = 106 \\
    x_7 &= 93, x_8 = 88, x_9 = 36 \\
    x_{10} &= 68, x_{11} = 37, x_{12} = 35, x_{13} = 18
\end{aligned}
\] (4)

As shown in Figure 5-7, the system’s load performance has been improved by using the algorithm proposed in Section 3. The larger the variance is, the more the numbers move from the mean, which means the system has a worse load balancing performance. The load variance is 14186 in the initial state, while the value is 7127.6 after optimization, which is 49.76% lower. The results show that the algorithm we proposed in Section 3 can improve the load performance of the distributed system with limited resource migration capacity.
5. Conclusions
In this paper, we analysed the resource scheduling problem of distributed systems, and obtained that the system has limited resource migration capacity for many reasons, such as high cost, QoS requirement, and unprocurable operation. Then a model is presented to describe the resource or the load scheduling problem of the distributed systems, which is shown as a constrained optimization problem. By using the penalty function method, the model can be transformed into an unconstrained optimization problem. Then a load balancing algorithm based on penalty function is presented to solve the load balancing problem in the distributed system. Finally, the simulation results show that the load balancing algorithm based on penalty function produces positive influence on system’s load performance.

Acknowledgments
This work was supported by the manned space advanced research project under Grant No. 060401, advanced research project of civil aerospace technology under Grant No. B0105 and the Development Fund of CETC 54 under Grant No. SXX18628X026.

References
[1] Yu Zhang, Xiaofei Liao, Hai Jin, Guang Tan, Geyong Min 2015 Inc-Part: Incremental Partitioning for Load Balancing in Large-Scale Behavioral Simulations. IEEE Trans. On Parallel and Distributed Systems, vol 26 pp 1900-1909
[2] Fatemeh Ebadifard, Seyed Morteza Babamir 2017 A PSO-based task scheduling algorithm improved using a load-balancing technique for the cloud computing environment. Concurrency and Computation: Practice and Experience. Vol 30 pp 1-16
[3] Xiaoying Tan, Yuchun Guo, Yishuai Chen, Wei Zhu 2018 Accurate inference of user popularity preference in a large-scale online video streaming system. Science China (Information Sciences) vol 61
[4] F. Alam, V. Thayananthan, and I. Katib 2016 Analysis of Round-robin Load-balancing Algorithm with Adaptive and Predictive Approaches, Proceedings of UKACC 2016 (Belfast UK: IEEE)
[5] Pandaba Pradhan, Prafulla Ku. Behera B N B Ray 2016 Modified Round Robin Algorithm for Resource Allocation in Cloud Computing Procedia Computer Science, vol85 pp.878-890
[6] V. M. Arul Xavier, S. Annadurai 2019 Chaotic social spider algorithm for load balance aware task scheduling in cloud computing Cluster Computing vol 22 pp 287-297
[7] Thanaa A. Alnusairi, Ashraf A. Shahin Yassine Daadaa 2018 Binary PSOGSA for Load Balancing Task Scheduling in Cloud Environment. International Journal of Advanced Computer Science and Applications vol 9 pp 255-264
[8] F.D. Calabrese L Wang E. Ghadimi G. Peters 2018 Learning Radio Resource Management in RANs: Framework, Opportunities, and Challenges IEEE Communications Magazine vol 56 pp 138-145
[9] Jiang Min Zhiqing Meng Gengui Zhou Rui Shem 2018 A Smoothing Penalty Function Algorithm for Two-Cardinality Sparse Constrained Optimization Problems, Proceedings of CIS 2018