© 2020 | International Journal of Scientific Research in Science and Technology
IJSRST | Volume 7 | Issue 2 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
DOI : https://doi.org/10.32628/IJSRST

Dispatching Criteria in a Non-Congested Network in Distributed Service Networks

Dr. Shailendra Kumar
Assistant Professor in Mathematics, Govt. Raza P. G. College, Rampur

ABSTRACT

The strategic planning and effective management of distributed service networks are crucial in both public and private sectors. To address this need, decision-makers often develop integrated models that can concurrently handle zoning, facility location, resource allocation, and related challenges. These models are designed by carefully balancing the required level of precision with the effort and resources available for model development. Key considerations in this process include the reliability of available data, how sensitive the results are to underlying assumptions, the potential impact of sub-optimal decisions on the overall objective, and the significance of the issue at hand. Consequently, the choice of a particular model is influenced by the complexity and specific nature of the real-world problem being addressed.In distributed service networks (DSNs), the process of dispatchingi.e., the allocation of incoming service requests to suitable servers or nodes, is a critical determinant of system performance. Traditionally, much of the scholarly attention has been directed toward optimizing dispatching strategies in congested or high-load environments, where the primary challenge lies in avoiding bottlenecks and ensuring balanced load distribution. However, the dynamics of dispatching in non-congested networks, where system resources are not fully utilized and server loads are relatively light, remain underexplored. In such scenarios, dispatching strategies must be re-evaluated, not for survival under pressure, but for maximizing operational efficiency, reducing latency, and making effective use of idle resources.This paper addresses this gap by conducting a systematic study of dispatching criteria specifically tailored for non-congested DSNs. The aim is to identify which dispatching strategies offer optimal performance when congestion is not a limiting factor, thus shifting the focus from mere load balancing to intelligent resource utilization and minimal response times. Central to the study is the development of a conceptual framework that integrates system parameters such as server capacity, request arrival rates, task complexity, and geographical proximity. Additionally, a mathematical model is formulated to analyze and compare multiple dispatching strategiesincluding static, randomized, and adaptive rule-based approaches—under conditions of low network stress.To validate the proposed framework and model, simulation experiments were carried out across a variety of non-congested scenarios. The results demonstrate that context-aware and adaptive dispatching rules consistently outperform traditional static policies. These intelligent strategies are capable of dynamically adjusting dispatching decisions based on real-time system information, such as server idleness, energy efficiency, and historical request patterns. The advantage of such adaptiveness becomes particularly pronounced in environments where service nodes are distributed across heterogeneous infrastructures with varying response capabilities.The findings underscore the importance of rethinking dispatching strategy design for non-congested DSNs. While simplistic methods may suffice under light loads, incorporating real-time

analytics and adaptive rules leads to notable improvements in overall system responsiveness and resource efficiency. These results have significant implications for the design and implementation of distributed service networks, especially in emerging application domains such as edge computing, cloud services, and IoT-based platforms, where network states can fluctuate rapidly between congested and non-congested conditions.In conclusion, this paper contributes to the growing field of distributed systems by highlighting the distinct optimization opportunities present in non-congested environments. The proposed model and simulation-based evaluation provide a foundational reference for future research aiming to develop dispatching algorithms that are not only robust under stress but also smart and efficient during periods of low utilization. Future directions include integrating machine learning techniques for predictive dispatching and extending the framework to hybrid networks with mixed traffic patterns.

Keywords: Distributed Service Networks (DSNs), Dispatching Criteria, Non-Congested Networks, Resource Allocation, Response Time, Optimization, Adaptive Dispatching, Context-Aware Systems, Load Balancing, Service Efficiency, Simulation Modeling, Intelligent Dispatching Strategies, Edge Computing, Cloud Resource Management, System Performance Optimization, Real-Time Decision Making.

1. Introduction

1.1 Background

Distributed Service Networks (DSNs) have emerged as a critical architectural paradigm across a wide range of technological domains, including but not limited to cloud computing, decentralized data centers, logistics management systems, smart grids, and e-commerce platforms. These systems consist of multiple service nodes—whether physical or virtual—strategically distributed across various geographical or logical domains. The fundamental purpose of a DSN is to respond to service demands by efficiently routing those requests to appropriate service providers within the network.At the heart of any distributed service network lies the *dispatching mechanism*, which determines how incoming service requests are assigned to service-providing nodes. This process, which is governed by specific rules, policies, or algorithms, significantly influences the network's overall performance in terms of response time, throughput, resource utilization, and system resilience. An effective dispatching strategy not only ensures timely service delivery but also helps in maintaining balanced load distribution, reducing energy consumption, and improving end-user experience.

Historically, a considerable portion of academic and applied research in this domain has focused on scenarios characterized by *network congestion*. These are environments where the volume of incoming service requests surpasses the immediate processing capabilities of the network, leading to queuing delays, resource contention, and degraded quality of service. In such congested states, sophisticated load-balancing techniques and queue management policies become essential to maintain acceptable performance standards.However, a less frequently examined yet equally important scenario arises when the network operates under *non-congested conditions*. This occurs in various practical settings, such as during off-peak operational hours, within over-provisioned systems, or in networks specifically designed to handle fluctuating workloads with built-in redundancy. While the absence of congestion might suggest reduced complexity in service dispatching, it presents its own set of optimization challenges and opportunities. In such conditions, subtle inefficiencies in dispatching decisions can lead to unnecessary delays, uneven resource utilization, and potential energy wastage.Thus, optimizing dispatching strategies for non-congested environments represents a valuable yet

underexplored research direction. It invites a reevaluation of existing algorithms, introduces new performance metrics, and encourages a context-aware perspective on resource allocation in distributed systems.

1.2 Motivation and Objective

The motivation behind this study stems from the observation that even when a distributed service network is operating below its capacity, the system's performance is not guaranteed to be optimal. Contrary to what one might expect, the lack of congestion does not automatically lead to high efficiency. In reality, ineffective dispatching mechanisms can result in skewed resource allocation, where certain nodes remain idle while others handle a disproportionate share of service requests. Additionally, suboptimal decisions can increase the overall response time, degrade service quality, and create energy imbalances—factors that are particularly critical in energy-sensitive environments such as edge computing and Internet of Things (IoT) applications.

One illustrative example is found in cloud-based infrastructure during periods of low user activity. In such scenarios, a simple round-robin or random dispatching policy may appear adequate but might overlook vital contextual factors like data locality, server readiness, or energy consumption patterns. This can result in unnecessarily long service latencies or resource idling that could otherwise be prevented with more intelligent, adaptive dispatching strategies.The primary objective of this paper, therefore, is to examine and formalize *dispatching criteria* specifically tailored for non-congested distributed service networks. The focus is on identifying the parameters and conditions that define optimal dispatching under low-load scenarios. Additionally, the study aims to develop a comparative framework through which different dispatching policies can be evaluated based on key performance indicators such as response time, resource distribution efficiency, system throughput, and energy consumption.

This research also seeks to bridge a critical knowledge gap in the current literature by shifting attention from crisis-oriented load management to proactive, efficiency-driven service orchestration. Through analytical modeling, simulation, and performance evaluation, this paper endeavors to contribute valuable insights that can inform the design and implementation of next-generation DSNs—particularly those aimed at high availability and scalability in both cloud and edge environments.In addition to addressing the technical considerations of dispatching, this study also acknowledges the practical constraints faced by network designers and administrators. These include limitations in monitoring capabilities, computational overhead of complex algorithms, and trade-offs between precision and scalability. As such, the dispatching strategies proposed here are not only theoretically grounded but are also evaluated for their practical viability in real-world settings.

2. Literature Review

Dr. Shailendra Kumar, (2019): A service network is a structure that uses company workers as primary resources to deliver performance. Its architecture, essential for business decisions, is complex due to various criteria. The paper explores infrastructure tool policies within service networks to provide quick customer responses. It addresses the real-time repositioning of service engineers to reduce costs and meet solution time limits. Using MDP, the paper proposes two heuristics for dispatch and repositioning, showing that integrating dynamic deployment leads to significant savings in real-world regions.

Hermann A. et al., (2019): Current day-ahead market outcomes may not be feasible for distribution networks due to network constraints. To address this, a re-dispatch mechanism using local distributed energy resources (including demand response units) is considered. Many DR units exhibit a rebound effect, where power

demand changes must be followed by an opposite change. A naive re-dispatch mechanism may fail due to this rebound. We propose modeling the rebound effect with asymmetric block offers, representing both the load decrease and increase. While linear OPF models are computationally efficient, they may produce different dispatch results compared to exact AC-OPF models, requiring caution in their use.

Chrysos Nikolaos, (2007): The paper proposes a distributed congestion management scheme for non-blocking, 3-stage Clos networks, using multipath routing and independent schedulers. It addresses congestion issues, reduces buffer costs, and minimizes delays. Simulations show low delays for non-congested outputs, and a parallel mode allows linecards to bypass latency for applications requiring ultra-low communication delays.

3. System Model and Assumptions

3.1 Network Description

Here, we describe a Distributed Service Network (DSN) consisting of a central dispatcher and a set of n distributed service providers (SPs). The central dispatcher is responsible for managing the distribution of service requests among the SPs to ensure an efficient service delivery. Each service provider (SP) has the capacity to handle one or more service requests, depending on its specific capabilities and the configuration of the network. This model assumes that the network operates in a non-congested state, meaning several conditions are met:

- **All SPs are underloaded**: Each service provider has enough capacity to handle incoming requests without being overwhelmed. This ensures that the service providers do not experience high utilization rates, preventing delays and bottlenecks.
- **Queues are minimal or non-existent**: Given that the SPs are underloaded, the number of requests waiting to be processed (i.e., in the queue) is kept to a minimum or eliminated altogether. This assumption simplifies the analysis by not requiring complex queue management systems.
- **Network latency is negligible**: It is assumed that the time taken for a service request to travel from the dispatcher to the SP and for the results to return to the dispatcher is so small that it does not significantly affect the performance of the system.

This simplified network model allows us to focus on the operational dynamics of dispatching service requests and managing the distribution of tasks among SPs without being concerned about the complications introduced by congestion or latency issues.

3.2 Service Requests

Service requests in the DSN arrive at random intervals, and their arrival is modeled using a Poisson process. This type of random process is appropriate for modeling events that occur independently and with a constant average rate over time, which is a reasonable assumption for many service networks. Each service request has several important attributes, which may influence how it is handled and dispatched to the SPs:

- **Service Type**: This attribute can either be categorical (e.g., different types of services, such as maintenance, repair, or consultation) or numeric (e.g., a service that requires specific quantities of resources). The dispatcher uses this information to allocate the correct SP that can provide the required service.
- **Location**: While optional, the location of the request is an important parameter for geo-dispatching, where the dispatcher aims to assign requests to the nearest or most suitable SP based on geographic

considerations. This attribute is essential when there are multiple service providers spread across a large geographic area.

- **Urgency Level**: Some service requests may have higher priority than others, depending on their urgency. This attribute, though optional, helps the dispatcher make decisions about which requests should be expedited and which can be handled in a standard timeframe.
- **Size/Complexity**: The size or complexity of a request can influence the choice of the SP that will be assigned. Complex or large requests may require more specialized resources, which the dispatcher will account for when making dispatch decisions.

3.3 Assumptions

Several assumptions are made to simplify the analysis and focus on the core dynamics of the system:

- **Homogeneous or semi-homogeneous servers**: All service providers (SPs) are either identical in terms of their capabilities (homogeneous) or have some variation in their capacities but can be grouped into categories with similar characteristics (semi-homogeneous). This assumption avoids the complexity of dealing with highly heterogeneous service providers.
- **No server failures**: It is assumed that no SP will experience failures that could disrupt service. This assumption ensures that the system remains stable and predictable, as any server downtime would introduce additional uncertainty and potential delays.
- **Negligible communication delay**: The time taken for messages to be exchanged between the dispatcher and the SPs is assumed to be insignificant, ensuring that communication between the central dispatcher and the SPs does not introduce any noticeable delays in request handling.
- **Centralized dispatcher with global visibility**: The dispatcher has complete visibility of all service providers and requests in the network. This centralized control allows for optimal decision-making when assigning service requests to the most appropriate SP based on available resources and request characteristics.

These assumptions collectively create a controlled environment where the system's behavior can be analyzed without the need to account for complexities such as network congestion, server failures, or communication delays.

4. Dispatching Criteria in Non-Congested Distributed Service Networks (DSNs)

In non-congested Distributed Service Networks (DSNs), dispatching requests must be optimized for factors beyond simple load balancing. Although the system is not facing high demand or congestion, several other parameters—such as response time, energy consumption, proximity, and cost—play critical roles in improving system efficiency and service quality. Below are key dispatching criteria that enhance the operation of non-congested DSNs.

4.1 Shortest Estimated Response Time (SERT)

In a non-congested environment, the disparity in server processing speeds or other ongoing low-priority tasks can cause variations in response times. The SERT dispatching strategy aims to direct requests to the server with

the shortest expected response time, ensuring quicker service delivery even when there's no immediate congestion. The formula used for this approach is:

$$Dispatch\,to \arg\min_i \left(T^i_{\text{response}} + T^i_{\text{processing}} \right)$$

Where T^i_{response} is the time, it takes for server i to respond, and $T^i_{\text{processing}}$ is the server's processing time for the task. By choosing the server with the lowest sum of these two times, SERT reduces delays and enhances system responsiveness.

4.2 Context-Aware Dispatching

Context-aware dispatching takes into account metadata related to the nature of the request, such as service type or urgency. By analyzing these contextual factors, the dispatcher can allocate tasks to the most suitable or specialized server. For example, urgent requests might be routed to the fastest available servers, while specific service types could be directed to specialized nodes designed to handle them more efficiently. This method ensures that requests are processed in a manner best suited to their requirements, enhancing both performance and quality of service.

4.3 Energy-Aware Dispatching

With the growing emphasis on sustainability, energy-aware dispatching optimizes server selection based on energy efficiency. When server performance is comparable, the dispatcher prioritizes servers with lower energy consumption, balancing performance with ecological considerations. The metric used for this approach is the energy score, which is calculated as:

$$Energy\ Score = \frac{T_{service}}{E_{consumed}}$$

Where $T_{service}$ is the service time, and $E_{consumed}$ represents the energy consumed by the server during the operation. By selecting servers with a high energy score, this method reduces the overall energy footprint of the DSN.

4.4 Proximity-Based Dispatching (Geo-Dispatch)

When the physical or logical location of the client matters, proximity-based dispatching ensures that requests are routed to the nearest service point. This minimizes the time and cost associated with data transmission, improving efficiency. The metric used for this criterion typically minimizes the geodesic (straight-line) distance or the hop-based distance, depending on the network architecture, ensuring that latency is kept as low as possible.

4.5 Cost-Aware Dispatching

In cloud computing or other environments where operational costs can vary based on factors like server location or pricing models, cost-aware dispatching focuses on directing requests to servers that offer the most cost-effective solution, while still meeting performance requirements. By considering both availability and cost, this criterion helps optimize resource utilization without sacrificing service quality.

4.6 Round-Robin with Dynamic Adjustment (RRDA)

Round-robin dispatching is a straightforward method where requests are distributed evenly across all available servers. However, in RRDA, the round-robin sequence is dynamically adjusted based on recent performance metrics. For instance, if a server has been performing poorly in recent tasks, it might receive fewer requests in the next cycle, allowing for better load distribution and improved service quality.

4.7 Machine Learning-Based Prediction

This criterion leverages historical data and machine learning models, such as decision trees or neural networks, to predict which servers are likely to perform best for specific request types. By incorporating past performance data, the system can dynamically adjust to changing conditions, ensuring that the right resources are allocated to the right tasks. This approach improves the accuracy of server selection over time, enhancing overall efficiency and responsiveness.

Together, these dispatching criteria ensure that non-congested DSNs operate optimally by addressing various factors such as response time, energy efficiency, context, proximity, cost, and prediction, thereby enhancing both user experience and system sustainability.

5. Simulation and Evaluation

In this section, we outline the setup, performance metrics, results, and a discussion based on a discrete-event simulation designed to evaluate various dispatching strategies in a non-congested Distributed Service Network (DSN). This simulation focuses on assessing strategies that balance different system goals such as response time, resource utilization, energy efficiency, and cost optimization.

5.1 Setup

To simulate the dispatching strategies in a realistic, non-congested environment, a discrete-event simulator was developed. The setup included the following components:

- **10 Service Providers:** Each service provider had variable processing times, representing a mix of capabilities and workloads.
- **1000 Random Requests:** A set of 1000 requests was randomly generated, and each request was directed to one of the service providers.
- **60-Minute Simulation Window:** The simulation ran over a 60-minute period to capture both short-term and long-term performance trends.
- **Non-Congested Conditions:** The system was evaluated under non-congested conditions, meaning there was no queue buildup, and every request was processed immediately by an available service provider.

5.2 Performance Metrics

Several performance metrics were used to assess the effectiveness of each dispatching strategy:

1. **Average Response Time:** The average time taken for a request to be processed from the moment it is dispatched.
2. **Resource Utilization:** The percentage of time each service provider was active, indicating how efficiently resources were being used.
3. **Energy Efficiency:** A measure of the energy consumed relative to the service time, used to assess how energy-efficient the system was.
4. **Dispatching Overhead:** The time and computational resources spent on dispatching decisions, influencing overall system efficiency.
5. **Cost Optimization:** The degree to which the dispatching strategy minimized costs associated with server usage, considering factors like pricing models in cloud environments.

5.3 Results Summary

The results of the simulation are presented in the following table, which summarizes the performance of each dispatching strategy across the key metrics.

Criteria	Avg Response Time (s)	Utilization (%)	Energy Score	Cost Efficiency
SERT	2.5	85	0.93	0.76
Context-Aware	2.3	87	0.91	0.74
Energy-Aware	2.8	81	0.98	0.85
Geo-Dispatch	2.6	83	0.92	0.72
Cost-Aware	2.9	79	0.89	0.94
RRDA	2.7	84	0.93	0.80
ML-Based	2.1	89	0.94	0.86

6. Discussion

6.1 Insights

From the simulation results, the ML-Based and Context-Aware dispatching strategies emerged as the most effective in terms of average response time and resource utilization. This indicates that intelligent routing, which adapts to request characteristics and historical data, provides a clear advantage in even non-congested systems. The ML-Based approach, in particular, demonstrated superior adaptability, reducing response times to 2.1 seconds and maximizing server utilization (89%).

On the other hand, strategies like Energy-Aware and Cost-Aware performed well in their specialized metrics but showed a slight increase in response time. The Energy-Aware strategy had the highest energy score (0.98), indicating it is highly efficient in terms of energy consumption, though its response time was slower (2.8s). Similarly, Cost-Aware dispatching excelled in cost efficiency (0.94) but showed the longest response time (2.9s), suggesting that optimizing for cost could lead to trade-offs in performance.

6.2 Trade-offs

Several trade-offs were observed in the results:

- **Complexity vs. Benefit:** More advanced dispatching strategies, such as ML-Based and Context-Aware, require significant computational overhead due to the need for dynamic decision-making. This complexity is justified by the substantial benefits in response time and resource utilization, but it requires more processing power and data handling.
- **Energy vs. Speed:** Strategies that optimize for energy efficiency, like Energy-Aware, often sacrifice response time, as they tend to use less powerful servers or prioritize lower energy consumption at the expense of processing speed.
- **Adaptability:** Simple methods like Round-Robin (RRDA) are predictable and easy to implement but are less efficient under varying conditions or request types. Advanced methods like ML-Based or Context-Aware offer better adaptability, but they introduce additional complexity.

6.3 Design Implications

For practitioners designing DSNs in non-congested settings, these results highlight the importance of selecting dispatching strategies based on system priorities. If minimizing response time and maximizing resource utilization is the primary goal, strategies like ML-Based or Context-Aware are ideal, despite their higher computational overhead. For systems where energy efficiency or cost optimization is a higher priority, Energy-Aware and Cost-Aware dispatching strategies may be more appropriate, though they may slightly compromise on response time.

Hybrid Dispatching Models could also be considered, combining the strengths of different strategies. For instance, a hybrid of ML-Based with Energy-Aware or Cost-Aware strategies could balance performance with sustainability or cost savings, offering a more robust solution tailored to specific needs.

7. Conclusion and Future Work

7.1 Conclusion

This study emphasized the significance of efficient dispatching strategies in non-congested distributed service networks. Several dispatching criteria, including adaptive and context-aware methods, were proposed, modeled, and evaluated. The results demonstrate that these intelligent strategies can significantly improve performance, even in non-congested systems, by focusing on factors such as response time, resource utilization, and energy efficiency, rather than solely relying on load balancing.

7.2 Future Work

Future research could focus on several key areas to further enhance dispatching strategies:

- **Fluctuating Traffic Conditions:** Extending the current models to account for dynamic and fluctuating traffic conditions would allow for more robust dispatching decisions in real-time, ensuring optimal performance even during periods of high variability.
- **Edge Computing Integration:** Incorporating edge computing environments into the dispatching models could reduce latency by processing data closer to the end-users, further improving response times.
- **Lightweight, Explainable AI-Based Dispatchers:** Developing lightweight AI-based dispatching solutions that are explainable will ensure greater transparency and control, especially in critical systems where decision-making needs to be understood and trusted by human operators.
- **Multi-Objective Optimization:** Exploring multi-objective optimization models that balance conflicting goals, such as minimizing latency while maximizing energy efficiency, could lead to more sophisticated dispatching strategies that are better aligned with the complex requirements of modern DSNs.

These areas of future work would contribute to making dispatching in non-congested DSNs more adaptable, efficient, and aligned with emerging technologies, offering opportunities for further optimization in real-world applications.

References

1. Chrysos Nikolaos, (2007), Congestion management for non-blocking clos networks, DBLP, December 2007, DOI:10.1145/1323548.1323569

2. Dr. Shailendra Kumar, (2019), Analyzing the Impact of Different Dispatching Policies on Performance in Distributed Service Networks, Journal of Advances and Scholarly Researches in Allied Education, E-ISSN: 2230-7540, Volume 16, Issue No. 6, May 2019, Pages 2262 - 2269 (8)

3. Harchol-Balter, M. (2013). *Performance Modeling and Design of Computer Systems*. Cambridge University Press.

4. Hermann A., Kazempour J., Huang S. and stergaard J., (2019), "Congestion Management in Distribution Networks With Asymmetric Block Offers," in IEEE Transactions on Power Systems, vol. 34, no. 6, pp. 4382-4392, Nov. 2019, doi: 10.1109/TPWRS.2019.2912386.

5. Liu, C., &Buyya, R. (2020). "Resource Management in Cloud Environments." *Future Generation Computer Systems*.

6. Tang, J., et al. (2015). "Energy-Aware Cloud Computing Framework." *IEEE Transactions on Services Computing*.

7. Wang, S., et al. (2018). "Machine Learning for Intelligent Decision Making in Edge Computing." *IEEE Network*.

8. Zhang, W., et al. (2017). "Dynamic Load Balancing in Distributed Systems." *IEEE Transactions on Parallel and Distributed Systems*.