Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 7, No. 3, 2026

Load Balancing Techniques in Multi-Sink Wireless Sensor Networks

Zexin Li, Guanhong Chen, Yadong Gong *, Chengjin Zhou

School of Computer Science and Software, Zhaoqing University, Zhaoqing 526061, China

* **Corresponding author**: (Email: gongyadong@zqu.edu.cn)

Abstract: Wireless sensor networks (WSNs) rely on battery-powered sensor nodes, making load balancing essential for prolonging network lifetime and ensuring reliable data delivery. Multi-sink WSNs can mitigate the hot-spot problem of single-sink architectures by distributing data collection points, but they also introduce complex imbalance among nodes, paths, cluster heads, and sink service regions. This paper reviews load balancing in multi-sink WSNs by first defining major load types and balancing objects, including communication, energy, path, cluster-head, and sink-region loads. Then, this paper analyzes imbalance formation from network topology, sink placement, node-sink association, routing strategy, traffic dynamics, and residual energy variation. Existing methods are classified into sink deployment, node association, load-aware routing, clustering, mobile sink scheduling, and intelligent optimization approaches, with their advantages and limitations compared. Finally, key challenges are discussed, including dynamic load awareness, low-overhead state maintenance, multi-objective trade-offs, static-mobile sink coordination, and lightweight adaptive mechanisms for practical deployment.

Keywords: Wireless sensor network, load balancing, multiple sinks, sink deployment.

1. Introduction

Wireless sensor networks (WSNs) have been extensively employed in environmental monitoring, industrial automation, precision agriculture, structural health inspection, disaster warning, and other sensing-oriented applications. A typical WSN consists of a large number of sensor nodes that are deployed to collect physical information and report it to one or more sink nodes. Since sensor nodes are usually powered by limited batteries and are difficult or even impossible to recharge after deployment, energy efficiency becomes one of the most critical design concerns. Among different sources of energy consumption, wireless communication is often the dominant one, especially in multi-hop data collection scenarios where nodes not only transmit their own sensed data but also forward packets generated by others. Consequently, the workload of different nodes may vary significantly according to their locations, forwarding roles, and traffic conditions. If such workload is not properly balanced, some nodes may exhaust their energy much earlier than others, leading to reduced network lifetime, degraded connectivity, and unstable data delivery. Therefore, load balancing is not a secondary optimization issue in WSNs, but a fundamental mechanism for maintaining long-term and reliable network operation.

In conventional single-sink WSNs, all sensing data are eventually delivered to a unique sink node. This simple architecture facilitates network management, but it also creates a typical many-to-one traffic convergence pattern. Nodes located close to the sink are usually responsible for forwarding a large amount of traffic from remote nodes, while nodes farther away may only generate and transmit their own data. As a result, the forwarding burden and energy consumption around the sink become much heavier than those in other parts of the network. Once these heavily loaded nodes deplete their batteries, the sink may be isolated from the rest of the network even though many outer nodes still retain sufficient residual energy. This phenomenon is commonly described as the hot-spot or energy-hole problem. It reveals an essential limitation of single-sink data collection: network performance is not determined only by the total remaining energy, but also by whether energy consumption and forwarding tasks are spatially balanced.

Multi-sink WSNs have been introduced as an effective architecture to alleviate the above limitations. By deploying multiple sink nodes, the network can shorten the average transmission distance, reduce the number of forwarding hops, distribute data collection points, and relieve the traffic pressure that would otherwise concentrate around a single sink. In addition, multiple sinks provide more flexible choices for node association and routing, which is beneficial for improving scalability in large-scale monitoring areas. Nevertheless, the presence of multiple sinks does not automatically guarantee load balancing. If sink nodes are placed improperly, some sinks may cover excessively large or dense service regions. If ordinary nodes simply connect to the nearest sink or select routes only according to the minimum hop count, traffic may still be concentrated on a few sinks, paths, or relay nodes. Moreover, dynamic events, uneven node distribution, residual energy variation, and cluster-head overload may further intensify local imbalance. Thus, the key issue in multi-sink WSNs is not merely how to increase the number of sinks, but how to coordinate sink deployment, node-sink association, routing, clustering, and mobility control to achieve balanced load distribution.

This paper focuses on load balancing techniques in multi-sink WSNs and provides a systematic review from both problem-definition and method-classification perspectives. First, the load balancing problem is clarified by identifying different load types and balancing objects, including node load, path load, cluster-head load, and sink-region load. Then, the major causes of load imbalance are analyzed in terms of network topology, sink placement, node association, routing strategy, traffic dynamics, and residual energy evolution. On

this basis, existing techniques are reviewed from six aspects: sink deployment optimization, node-sink association, load-aware routing, clustering-based load balancing, mobile sink scheduling, and intelligent load balancing decision methods. Furthermore, different categories of approaches are compared with respect to their advantages, limitations, and suitable application scenarios. Finally, several open challenges and future research directions are discussed, including dynamic load perception, low-overhead state maintenance, multi-objective trade-offs among energy balance, delay and reliability, coordination between static and mobile sinks, and lightweight adaptive mechanisms for practical deployment.

2. Problem Definition

In a multi-sink wireless sensor network, load balancing cannot be reduced to the simple idea of making all nodes consume energy at the same rate. Although energy consumption is a central concern, the load in such networks is distributed across several interrelated dimensions: individual sensor nodes, forwarding paths, cluster heads, and sink service regions. A method that balances one dimension may still leave another dimension highly uneven. For example, assigning an equal number of nodes to different sinks may balance the apparent sink load, but if most traffic is forwarded through a small set of relay nodes, the network may still suffer from premature energy depletion and local connectivity failures. Therefore, before discussing load balancing techniques, it is necessary to clarify what types of load exist, how imbalance is formed, and which metrics should be used to evaluate balancing performance.

2.1. Load Types and Balancing Objects

In wireless sensor networks, load first appears as communication load, which refers to the amount of data transmission, reception, and forwarding undertaken by sensor nodes. A node not only sends its own sensed data but may also act as a relay for other nodes, especially in multi-hop data collection scenarios. Since radio communication is usually the dominant source of energy consumption in WSNs, communication load is closely transformed into energy load. The more packets a node forwards, the faster its battery is depleted, and the more likely it becomes a bottleneck that limits the operational lifetime of the entire network. Thus, node-level load balancing mainly aims to avoid persistent overuse of a few sensors while many others remain underutilized.

Beyond the load of individual nodes, multi-sink WSNs involve several higher-level load objects. Path load describes the pressure placed on routing paths that are repeatedly selected for data delivery; when shortest paths or minimum-hop routes are used too frequently, the relay nodes on these paths may become overloaded even if the source nodes are evenly distributed. Cluster-head load is another important form of load in hierarchical networks, because cluster heads are responsible for member management, data aggregation, intra-cluster coordination, and sometimes inter-cluster forwarding. In addition, multi-sink networks introduce sink-region load, which refers to the difference among sinks in terms of the number of served nodes, received traffic volume, covered area, and forwarding burden imposed on nearby nodes. These load types are not isolated: an oversized sink region may increase path load, and an overloaded cluster head may further intensify local energy imbalance.

Accordingly, load balancing in multi-sink WSNs should be understood as a multi-layer problem. At the node level, the objective is to prevent individual sensors or relays from consuming energy much faster than others. At the path level, the goal is to distribute forwarding tasks among alternative routes rather than repeatedly using a small number of “best” paths. At the cluster level, balancing focuses on reasonable cluster formation, cluster-head rotation, and fair allocation of aggregation tasks. At the sink-region level, the key issue is whether different sinks serve comparable amounts of traffic and whether their surrounding areas avoid excessive forwarding pressure. Since different algorithms usually act on different layers, any meaningful comparison of load balancing schemes must first specify the primary balancing object being optimized.

2.2. Formation Mechanisms of Load Imbalance

Load imbalance in multi-sink WSNs is partly rooted in network topology and sink placement. Sensor nodes are rarely distributed in a perfectly uniform manner in real deployments; terrain, deployment method, obstacles, and monitoring requirements may all create dense and sparse regions. If sinks are placed without considering such spatial heterogeneity, some sinks may naturally cover larger or denser areas than others. Moreover, because sensor nodes usually have limited communication ranges, nodes located in certain positions may have to carry more transit traffic simply due to their topological advantage. For instance, nodes near a sink, near a narrow communication corridor, or at the boundary between two service regions may become unavoidable relays. In this sense, the initial placement of sinks and nodes establishes the basic load distribution pattern of the network.

Node-sink association and routing decisions further amplify or mitigate this initial imbalance. A common strategy is to associate each node with the nearest sink or the sink reachable through the minimum number of hops. This strategy is simple and often reduces average transmission cost, but it may also cause certain sinks to receive a disproportionate amount of traffic. Similarly, shortest-path routing can minimize the cost of individual data transmissions, yet when many nodes independently select the same optimal path, several intermediate nodes may be repeatedly used and rapidly drained. This problem is especially serious in multi-hop networks, where nodes close to sinks often forward data for a large number of upstream nodes. Therefore, a locally efficient decision, such as choosing the nearest sink or the shortest route, does not necessarily lead to global load balance.

The imbalance is also dynamic rather than fixed. In event-driven monitoring applications, traffic may suddenly increase in a specific region when an event occurs, causing temporary congestion and accelerated energy consumption around that area. Meanwhile, the residual energy of nodes changes continuously as the network operates. A routing path that is initially reasonable may become unsuitable after several relay nodes lose a large portion of their energy. Link quality may also fluctuate due to interference, environmental conditions, or node failures, forcing traffic to shift to alternative paths and creating new hotspots. Consequently, load balancing in multi-sink WSNs should not be treated as a one-time optimization problem. Instead, it requires adaptive mechanisms that can respond to traffic variation, energy depletion, and topology evolution with acceptable control overhead.

2.3. Load Balancing Objectives and Evaluation Metrics

The objectives of load balancing in multi-sink WSNs are multidimensional. At the most basic level, a balancing mechanism should reduce excessive energy consumption at heavily used nodes and delay the death of critical relays. At the same time, it should distribute data traffic among multiple sinks so that no single sink region becomes persistently overloaded. Another important objective is to alleviate hotspot and energy-hole problems, especially around sinks or along frequently used forwarding paths. However, energy-related goals cannot be considered alone. In many applications, data must still be delivered within an acceptable delay, with sufficient reliability and without excessive signaling. Thus, practical load balancing is essentially a trade-off among energy fairness, network lifetime, traffic distribution, delay, reliability, and protocol overhead.

Several metrics can be used to evaluate these objectives. Network lifetime is one of the most widely used indicators, and it may be defined according to the time until the first node dies, a certain percentage of nodes fail, or the network becomes disconnected. Residual energy variance is useful for measuring whether energy consumption is evenly distributed across nodes. The number of forwarded packets per node can directly reflect relay burden and help identify overloaded intermediate nodes. For sink-level evaluation, the amount of data received by each sink, the number of associated nodes, and the size of each sink service region can reveal whether traffic is fairly distributed among sinks. In addition, average end-to-end delay, packet delivery ratio, and control overhead are necessary to assess whether a balancing method achieves energy improvement at the cost of unacceptable latency, poor reliability, or excessive maintenance messages.

A multi-metric evaluation perspective is therefore essential. A scheme may extend network lifetime by routing traffic through longer but less congested paths, yet this may increase delay and energy consumption for certain flows. Another scheme may equalize the number of nodes associated with each sink, but it may fail to reduce the forwarding burden of nodes located near the sinks. Likewise, a highly adaptive method may respond well to dynamic traffic changes, while the state exchange required for adaptation may consume considerable energy. For this reason, load balancing performance should not be judged by a single metric. A more rigorous evaluation should examine whether the proposed mechanism balances the intended load object, whether it improves overall network sustainability, and whether the additional cost introduced by the mechanism remains acceptable for resource-constrained WSN environments.

3. Key Technologies for Load Balancing

3.1. Sink Deployment Optimization and Regional Load Balancing

The deployment of sink nodes determines the initial load distribution pattern of a multi-sink wireless sensor network. In a single-sink network, traffic naturally converges toward one collection point, which makes nodes near the sink responsible for disproportionate relay tasks. By contrast, a multi-sink architecture can divide the monitoring field into several service regions, reduce the average transmission distance, and shorten multi-hop forwarding paths. Nevertheless, the benefit is not obtained merely by increasing the number of sinks. If sinks are placed too close to one another, their service regions overlap excessively and remote areas remain underserved; if they are placed too far from high-density sensing regions, nodes in those regions may still experience heavy forwarding pressure. Therefore, sink deployment should be regarded as a regional load-balancing problem rather than a simple coverage problem. Its key purpose is to make the number of served nodes, generated traffic volume, communication distance, and relay burden among different sink regions as even as possible.

Typical sink deployment methods can be grouped into geometric partitioning, grid- or Voronoi-based partitioning, facility-location modeling, and heuristic optimization. Geometric and grid-based methods divide the monitoring field into regular subregions and place sinks near the geometric centers or traffic centers of these subregions. Voronoi partitioning is often used to assign each node to the closest sink and to analyze the spatial boundary of sink service regions; this idea is related to centroidal Voronoi tessellation and Lloyd's algorithm, which iteratively adjusts representative points according to regional centroids [1]. Facility-location models provide a more formal way to describe sink placement. For example, the classical p-median model proposed by Hakimi aims to select a fixed number of facilities so that the total distance from demand points to facilities is minimized [2], and this modeling idea can be adapted to sink placement by treating sensor nodes or traffic hotspots as demand points. When the search space becomes large or the objective function includes multiple constraints, heuristic algorithms are often adopted. Genetic algorithms [3], particle swarm optimization [4], and ant colony optimization [5] have been widely used as general-purpose optimization tools and can be employed to search for sink locations that reduce communication cost, balance regional traffic, or improve network lifetime.

The main advantage of deployment optimization is that it improves load distribution at the network planning stage. A well-designed sink layout can reduce the dependence on long routes and prevent some regions from becoming overloaded from the very beginning. However, deployment optimization also has clear limitations. Many deployment schemes assume that node locations, traffic generation rates, and link conditions are known in advance, which may be unrealistic in randomly deployed or harsh environments. Moreover, static sink placement cannot fully respond to node failures, energy depletion, event bursts, or temporal changes in data generation. Consequently, sink deployment should not be treated as a one-time solution. In practical multi-sink WSNs, it is more effective when combined with online mechanisms such as adaptive node-sink association, load-aware routing, cluster-head rotation, or mobile sink scheduling.

3.2. Node-Sink Association and Aggregation Load Balancing

Node-sink association answers a basic but crucial question in multi-sink WSNs: to which sink should a sensor node send its data? This issue is different from routing. Association determines the final data collection destination, whereas routing determines the forwarding path toward that destination. The simplest association rule is to connect each node to the nearest sink or the sink with the minimum hop

count. Such a rule is easy to implement and introduces little control overhead, but it may produce seriously unbalanced aggregation loads. For instance, a sink located near a dense sensing region may receive far more packets than other sinks, and the nodes around that sink may rapidly become bottleneck relays. In addition, minimum-distance association ignores residual energy and traffic dynamics; as a result, a seemingly efficient association at the beginning of network operation may become unsuitable after some nodes consume much of their energy.

Load-aware association strategies extend the association decision from distance-based selection to multi-factor selection. A node may choose its sink according to hop count, residual energy along candidate paths, link quality, estimated congestion level, current sink load, or the number of nodes already assigned to each sink. In this sense, association can be formulated as a dynamic load allocation problem. When one sink region becomes overloaded, some boundary nodes may be reassigned to neighboring sinks if the additional routing cost is acceptable. Similar ideas appear in geographic and data-centric routing studies. For example, GPSR uses location information to support scalable forwarding decisions [6], while directed diffusion constructs data delivery paths according to interests and gradients rather than fixed addresses [7]. Although these protocols were not designed solely for multi-sink load balancing, their design logic shows that data delivery decisions can be made adaptively according to local state and communication demand. In a multi-sink scenario, this adaptivity can be further used to distribute data aggregation pressure among multiple sinks.

The strength of node-sink association is that it directly controls the service scope of each sink. Compared with sink deployment, it is more flexible; compared with route-level adjustment, it works at a higher decision layer and can prevent aggregation imbalance before packets enter heavily loaded paths. It is particularly useful in networks where sink coverage regions overlap or where nodes can reach more than one sink through alternative multi-hop paths. However, association also brings several challenges. First, nodes need updated information about sink load, path quality, and local residual energy, and maintaining such information consumes energy. Second, frequent sink switching may cause route oscillation, packet reordering, and additional control messages. Third, a locally beneficial reassociation may shift congestion to another area if global load conditions are not considered. Therefore, an effective association mechanism should include hysteresis, switching thresholds, or cost penalties so that load balancing is achieved without sacrificing association stability.

3.3. Load-Aware Routing and Path Load Balancing

Even if sinks are well deployed and nodes are reasonably associated with sinks, routing may still create load imbalance. In many WSNs, shortest-path or minimum-hop routing repeatedly selects the same relay nodes because these nodes appear to provide the lowest immediate transmission cost. Over time, these relays consume energy much faster than their neighbors and may die early, cutting off the data delivery paths of other nodes. This phenomenon is especially serious near sinks, around region boundaries, and along narrow communication corridors. Hence, load-aware routing focuses on how packets reach the selected sink while avoiding the long-term overuse of specific nodes or paths. Its essential goal is to distribute forwarding tasks across multiple feasible routes so that energy consumption and relay pressure are spatially balanced.

A common approach is to incorporate residual energy into route selection. Instead of always selecting the path with the smallest hop count, a node may prefer a next hop with higher residual energy, lower queue length, or better expected transmission quality. Energy-aware routing proposed by Shah and Rabaey introduced the idea of selecting paths probabilistically according to energy-related costs, so that traffic can be spread over multiple routes rather than concentrated on one optimal path [8]. Directed diffusion also supports path reinforcement and can potentially maintain multiple gradients toward data sinks [7]. In multi-sink WSNs, such mechanisms can be extended by considering both the route cost to a sink and the load condition of intermediate nodes. If two paths have similar lengths, choosing the path with higher residual energy or lower forwarding burden can significantly postpone the emergence of energy holes.

Multipath routing is another important way to achieve path load balancing. By maintaining several candidate paths, a node can split traffic, rotate forwarding paths, or switch to backup routes when the primary path becomes congested or energy-poor. Braided multipath routing, developed in the context of directed diffusion, constructs partially disjoint alternative paths to improve resilience without requiring fully node-disjoint routes [9]. This idea is valuable for multi-sink WSNs because strict disjointness is often difficult in dense but energy-constrained networks. In addition, topology-based protocols such as TTDD build a grid-like forwarding structure to support data dissemination toward multiple mobile sinks [10]. Although TTDD mainly addresses sink mobility, its structured forwarding idea also illustrates how path organization can reduce repeated network-wide flooding and distribute forwarding responsibilities more systematically.

However, load-aware routing must carefully balance energy saving, latency, reliability, and control overhead. A route that avoids low-energy nodes may be longer and thus introduce more transmissions. Multipath routing can improve robustness but requires additional route discovery and maintenance. Frequent route changes may also increase packet loss or destabilize forwarding decisions. Therefore, practical load-aware routing should avoid excessive global state maintenance and rely as much as possible on local indicators, such as residual energy, neighbor queue length, link quality, and recent forwarding count. More importantly, routing should cooperate with sink association. If a node is associated with a lightly loaded sink but the path toward that sink passes through highly depleted relays, the overall load-balancing effect may still be poor. Thus, in multi-sink WSNs, routing and association should be jointly considered rather than optimized in isolation.

3.4. Clustering Structure and Cluster-Head Load Balancing

Clustering introduces a hierarchical organization into WSNs. Ordinary nodes send data to cluster heads, cluster heads aggregate or compress the received data, and then the processed data are forwarded to one or more sinks. In multi-sink WSNs, clustering can reduce redundant transmissions and simplify network management because local data aggregation decreases the amount of traffic entering inter-

cluster routes. This is particularly useful in periodic monitoring applications, where neighboring sensors often generate correlated readings. Nevertheless, clustering also creates a new load-balancing object: the cluster head. A cluster head usually performs member coordination, data reception, aggregation, and long-distance forwarding, so it consumes more energy than ordinary nodes. If cluster heads are not selected or rotated properly, they may become local bottlenecks even when the overall sink distribution is reasonable.

Classical clustering protocols provide important design foundations. LEACH randomly rotates the role of cluster head among nodes to distribute energy consumption over time [11]. Its key insight is that cluster-head responsibility should not be fixed, because static cluster heads deplete energy rapidly. HEED improves cluster-head selection by considering residual energy and intra-cluster communication cost, which makes the clustering process more energy-aware and more suitable for heterogeneous node conditions [12]. In multi-sink WSNs, these ideas can be extended by adding sink-related factors. For example, a node with high residual energy may still be a poor cluster-head candidate if it is far from all sinks or located on a congested inter-cluster path. Therefore, cluster-head selection should jointly consider residual energy, node degree, distance to member nodes, distance or hop count to candidate sinks, and expected forwarding burden.

The main advantage of clustering is that it transforms flat data collection into hierarchical load management. It can reduce channel contention, localize control messages, and support scalable data aggregation. Yet cluster-based load balancing is sensitive to cluster size, cluster-head rotation frequency, and inter-cluster routing strategy. Oversized clusters overload their heads; undersized clusters reduce aggregation benefits and increase the number of inter-cluster transmissions. Frequent cluster reconstruction consumes energy, while infrequent reconstruction may leave depleted cluster heads in service for too long. In addition, in multi-sink networks, cluster heads near a sink or near the boundary between sink regions may forward traffic for multiple clusters, becoming hidden bottlenecks. Thus, clustering should be combined with adaptive cluster sizing, energy-aware cluster-head rotation, and balanced cluster-to-sink routing.

3.5. Mobile Sink and Spatio-Temporal Load Balancing

Mobile sinks extend load balancing from the spatial dimension to the temporal dimension. In fixed-sink networks, nodes near a sink repeatedly forward packets and are therefore more likely to form hotspots. A mobile sink changes its data collection position over time, allowing different parts of the network to communicate with the sink at different moments. When the sink approaches a region, nodes in that region can transmit data over shorter paths or even directly, reducing the forwarding burden on distant relays. This mechanism is especially useful for large-scale WSNs, sparse networks, or applications where data can tolerate some collection delay. The fundamental idea is not to remove load, but to distribute it across both space and time so that no small group of nodes permanently carries the majority of traffic.

Mobile sink trajectories can be fixed or adaptive. Fixed trajectories, such as circular paths, straight lines, grid tours, or predefined patrol routes, are simple and predictable. They reduce scheduling complexity because nodes can estimate when the sink will arrive nearby. However, fixed trajectories may fail to respond to uneven event distribution or sudden traffic bursts. Adaptive trajectories, by contrast, adjust sink movement according to residual energy, buffer occupancy, event location, or regional traffic intensity. Data MULEs proposed by Shah, Roy, Jain, and Brunette use mobile entities to collect data from sensors and physically carry them to access points, showing that controlled mobility can reduce energy consumption in sparse sensor networks [13]. Luo and Hubaux further demonstrated that sink mobility, when jointly considered with routing, can help prolong network lifetime by changing the traffic convergence point [14]. These studies support the idea that mobility can fundamentally reshape energy consumption patterns in WSNs.

The dwell-point strategy is a practical compromise between continuous movement and full network traversal. Instead of visiting every node, a mobile sink stops at selected points where it can collect data from nearby nodes or cluster heads. Dwell points may be chosen according to node density, energy distribution, traffic demand, or coverage of communication ranges. In multi-mobile-sink scenarios, coordination becomes even more important. Multiple mobile sinks can divide the monitoring field into subregions, assign collection tasks, and coordinate routes to avoid redundant visits. Such coordination resembles a task allocation and path planning problem. If designed properly, multiple mobile sinks can reduce data waiting time, balance collection responsibilities, and improve scalability in large monitoring areas.

Despite these advantages, mobile sink methods are not universally suitable. Mobility introduces path planning complexity, requires synchronization between nodes and sinks, and may increase data collection latency. If nodes must buffer data until a sink arrives, buffer overflow and delayed reporting may occur in event-intensive applications. Moreover, disseminating the sink's current position or future trajectory can generate extra control traffic. For real-time monitoring, emergency detection, or industrial control, relying solely on mobile sinks may be risky. A more balanced design is to combine static sinks with mobile sinks: static sinks handle delay-sensitive traffic and provide stable access points, while mobile sinks visit high-load or energy-critical regions to relieve long-term forwarding pressure.

3.6. Intelligent Load-Balancing Decision Methods

Intelligent decision methods should not be viewed as an isolated category parallel to deployment, association, routing, clustering, and mobility. Rather, they are optimization tools that can be embedded into different stages of multi-sink WSN load balancing. The load-balancing problem is naturally multi-objective and dynamic: energy, delay, reliability, packet delivery ratio, control overhead, and regional fairness often interact with one another. Rule-based methods are simple and lightweight, but they may not perform well when topology, traffic, and link quality change simultaneously. Intelligent methods provide a way to search, infer, or learn better decisions under complex conditions, especially when the relationship between decision variables and network performance is nonlinear.

Several intelligent algorithms are relevant to multi-sink WSNs. Genetic algorithms can encode sink locations, cluster-head sets, or routing paths as chromosomes and evolve them

through selection, crossover, and mutation [3]. Particle swarm optimization treats candidate solutions as particles moving in the search space and is often suitable for continuous optimization problems such as sink placement or mobile sink trajectory design [4]. Ant colony optimization uses artificial pheromone updating to guide path search and can be adapted to energy-aware routing or balanced next-hop selection [5]. Fuzzy logic is useful when routing or clustering decisions must combine imprecise indicators such as "high residual energy," "short distance," and "good link quality"; the fuzzy set theory introduced by Zadeh provides the theoretical basis for such reasoning [15]. Reinforcement learning, particularly Q-learning, enables nodes or sinks to improve decisions through interaction with the environment, and it can be used for adaptive sink selection, next-hop selection, or mobile sink scheduling [16].

The practical value of intelligent methods depends heavily on their implementation cost. WSN nodes usually have limited computation, storage, and energy, so algorithms requiring heavy centralized training or frequent global information exchange may be unsuitable for direct deployment. In addition, learning-based methods may converge slowly when network conditions change rapidly, and inaccurate state information can mislead the decision process. Therefore, future intelligent load balancing should move toward lightweight, distributed, and low-state-dependence designs. A promising direction is to place relatively complex computation at sinks, gateways, or edge servers, while ordinary sensor nodes execute simple local rules derived from the intelligent model. In this way, intelligent optimization can improve adaptability without overwhelming the resource-constrained sensing layer.

4. Challenges and Future Research Directions

Several fundamental challenges remain that must be addressed to translate load-balancing gains from theoretical models to practical, long-lived multi-sink WSN deployments as follows:

4.1. Dynamic Load Awareness and Low-Overhead State Maintenance

Load in multi-sink WSNs is inherently dynamic, fluctuating with events, energy depletion, link quality, and routing adaptations. Mechanisms relying solely on initial topology or infrequent updates fail to capture real-time network states, while continuous reporting of comprehensive metrics incurs prohibitive control overhead.

Future research should therefore develop lightweight, selective load-awareness strategies based on local observations—such as packet forwarding rates, queue variations, and neighbor energy trends—supplemented by event-triggered rather than periodic updates. Hierarchical state maintenance, where cluster heads or sink-side controllers aggregate regional indicators, offers a promising balance between awareness and cost, provided that compact yet expressive load descriptors can be designed for resource-constrained nodes.

4.2. Multi-Objective Trade-Off among Energy Balance, Delay, and Reliability

Load balancing is frequently treated as an energy-centric problem, yet practical applications demand simultaneous consideration of delay, reliability, throughput, and control overhead, as these objectives are tightly coupled rather than independent. Optimizing energy alone may increase hop counts and latency, while aggressive sink selection for traffic uniformity may force nodes through unstable links.

Future work should frame load balancing as a context-aware multi-objective optimization problem with adaptive weighting mechanisms that shift priorities according to network state and application requirements—for instance, favoring latency during energy-abundant phases and shifting toward energy redistribution as imbalance grows. Evaluation must likewise move beyond single metrics such as first-node-death time toward comprehensive assessment of residual energy distribution, load variance, delivery ratio, delay, and overhead.

4.3. Cooperative Load Scheduling between Static and Mobile Sinks

Static sinks provide stability and predictable routing but risk persistent hotspots, whereas mobile sinks can physically redistribute load yet introduce trajectory complexity, buffering delays, and route obsolescence. When designed independently, their advantages may cancel out: mobile sinks may redundantly serve well-covered regions while overloaded areas remain neglected.

A hybrid cooperative model is needed, wherein static sinks handle baseline periodic traffic while mobile sinks are dynamically dispatched to regions exhibiting energy depletion, congestion, or event bursts. Key challenges include coordinating service regions and collection schedules, integrating load prediction with trajectory planning, and managing handover costs and routing stability during mode transitions.

4.4. Lightweight Adaptive Mechanisms for Practical Deployment

Many existing schemes assume idealized conditions—accurate localization, homogeneous energy, reliable links, and global topology knowledge—that rarely hold in real deployments, where irregular placement, asymmetric links, battery aging, and node failures are commonplace. Complex centralized algorithms often prove infeasible for distributed implementation on low-power nodes.

Future research should prioritize robust, self-adaptive designs based on limited local knowledge, employing simple rules such as relay rotation, energy-penalized path selection, and sink-switching limits. Edge-assisted architectures, where sinks or gateways execute heavier optimization while nodes perform lightweight forwarding, represent a practical path forward. Ultimately, validation must occur under realistic conditions including bursty traffic, imperfect state information, and irregular topologies, ensuring that deployability is weighted as heavily as theoretical optimality.

5. Conclusion

This paper reviews load balancing in multi-sink WSNs across four dimensions: balancing objects, formation mechanisms, key techniques, and design principles. Multi-sink architectures offer traffic distribution and path-shortening advantages over single-sink networks, yet balanced operation requires joint optimization of sink

deployment, node-sink association, routing, clustering, and sink mobility. Load balancing is inherently a multi-layer problem spanning node energy, path pressure, cluster-head workload, and sink-region load; optimizing one layer alone may create bottlenecks elsewhere. Each technical category has distinct scope and limitations: deployment optimization shapes initial distribution but cannot handle dynamic events; association adjustment risks routing instability; load-aware routing trades delay for control; clustering may localize bottlenecks at cluster heads; mobile sinks face trajectory and scheduling constraints; and intelligent methods require lightweight adaptation for resource-constrained nodes. Future research should therefore move beyond isolated mechanism optimization toward cooperative, adaptive, low-overhead designs that jointly balance energy efficiency, delay, reliability, and scalability.

Acknowledgment

This work was supported in part by the Research Program of Zhaoqing University under Grant fw202512.

References

[1] Lloyd, S. P. (1982). Least squares quantization in PCM. IEEE Transactions on Information Theory, 28(2), 129–137. https://doi.org/10.1109/TIT.1982.1056489

[2] Hakimi, S. L. (1964). Optimum locations of switching centers and the absolute centers and medians of a graph. Operations Research, 12(3), 450–459. https://doi.org/10.1287/opre.12.3.450

[3] Holland, J. H. (1975). Adaptation in natural and artificial systems. University of Michigan Press.

[4] Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. In Proceedings of ICNN'95—International Conference on Neural Networks (pp. 1942–1948). IEEE. https://doi.org/10.1109/ICNN.1995.488968

[5] Dorigo, M., Maniezzo, V., & Colorni, A. (1996). Ant system: Optimization by a colony of cooperating agents. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 26(1), 29–41. https://doi.org/10.1109/3477.484436

[6] Karp, B., & Kung, H. T. (2000). GPSR: Greedy perimeter stateless routing for wireless networks. In Proceedings of the 6th Annual International Conference on Mobile Computing and Networking (pp. 243–254). ACM. https://doi.org/10.1145/345910.346760

[7] Intanagonwiwat, C., Govindan, R., & Estrin, D. (2000). Directed diffusion: A scalable and robust communication paradigm for sensor networks. In Proceedings of the 6th Annual International Conference on Mobile Computing and Networking (pp. 56–67). ACM. https://doi.org/10.1145/345910.346730

[8] Shah, R. C., & Rabaey, J. M. (2002). Energy aware routing for low energy ad hoc sensor networks. In 2002 IEEE Wireless Communications and Networking Conference (pp. 350–355). IEEE. https://doi.org/10.1109/WCNC.2002.993474

[9] Ganesan, D., Govindan, R., Shenker, S., & Estrin, D. (2001). Highly-resilient, energy-efficient multipath routing in wireless sensor networks. ACM SIGMOBILE Mobile Computing and Communications Review, 5(4), 11–25. https://doi.org/10.1145/509408.509411

[10] Ye, F., Luo, H., Cheng, J., Lu, S., & Zhang, L. (2002). A two-tier data dissemination model for large-scale wireless sensor networks. In Proceedings of the 8th Annual International Conference on Mobile Computing and Networking (pp. 148–159). ACM. https://doi.org/10.1145/570645.570664

[11] Heinzelman, W. R., Chandrakasan, A., & Balakrishnan, H. (2000). Energy-efficient communication protocol for wireless microsensor networks. In Proceedings of the 33rd Annual Hawaii International Conference on System Sciences. IEEE. https://doi.org/10.1109/HICSS.2000.926982

[12] Younis, O., & Fahmy, S. (2004). HEED: A hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks. IEEE Transactions on Mobile Computing, 3(4), 366–379. https://doi.org/10.1109/TMC.2004.41

[13] Shah, R. C., Roy, S., Jain, S., & Brunette, W. (2003). Data MULEs: Modeling a three-tier architecture for sparse sensor networks. In Proceedings of the First IEEE International Workshop on Sensor Network Protocols and Applications (pp. 30–41). IEEE. https://doi.org/10.1109/SNPA.2003.1203317

[14] Luo, J., & Hubaux, J.-P. (2005). Joint mobility and routing for lifetime elongation in wireless sensor networks. In Proceedings IEEE INFOCOM 2005 (pp. 1735–1746). IEEE. https://doi.org/10.1109/INFCOM.2005.1498444

[15] Zadeh, L. A. (1965). Fuzzy sets. Information and Control, 8(3), 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X

[16] Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. Machine Learning, 8(3–4), 279–292. https://doi.org/10.1007/BF00992698