BJET
(Bayero Journal of Engineering and Technology)
https://bjet.ng/index.php/jet
Research Article
2026, Vol. 21 (3), 19-32

DEVELOPMENT OF A LOW-COST IOT-ENABLED QUEUE MANAGEMENT SYSTEM USING ESP32, ESP-NOW, AND OFFLINE AUDIO-VISUAL FEEDBACK FOR ENHANCED SERVICE DELIVERY

O. E. Agwu[1]., E. U. Ehebiri[2]and D. E. Iniobong[3]

[1] Department of Electrical/Electronic Engineering, Michael Okpara University of Agriculture, Umudike, Nigeria
[2] Nigerian Independent System Operators, Ministry of Power, TCN Substationc, Katampe, Abuja FCT, Nigeria.
[3] Central Emerald State Research Community, Michael Okpara University of Agriculture, Umudike, Nigeria.

*Corresponding author email: agwu.ekwe@mouau.edu.ng

Received: 22-06-2026
Revised: 19-08-2026
Accepted: 02-08-2026
Published: 04-09-2026

Abstract: *Managing queues effectively is a major hurdle in both healthcare and banking sectors, where long wait times and a lack of information can really dampen customer satisfaction and hurt operational efficiency. Traditional methods, like paper token systems and basic digital displays, often overlook the psychological stress that comes with uncertainty and don't cater well to those with visual impairments. This paper introduces a smart queue management system that combines wireless communication, web-based registration, and synchronized audio-visual notifications. The system uses two ESP32 microcontrollers that talk to each other via the ESP-NOW protocol; one acts as the operator controller with a keypad and OLED display, while the other runs a dot-matrix LED display and can announce messages offline. A local web server allows users to register for the queue digitally through their mobile devices, and for those without smartphones, manual token generation is available. The audio system plays pre-recorded speech clips stored in flash memory to announce queue tokens, making it accessible for visually impaired users. Performance tests showed reliable wireless communication, precise token generation and queue progression, synchronized display and audio notifications, and effective offline operation. The system significantly enhanced queue accessibility and operational efficiency, boasting response times under 100 milliseconds and 100% transmission reliability within its range. This affordable and accessible solution offers a solid framework for modernizing queue management in settings with limited resources while ensuring inclusivity.*

Key words: Queue management system; ESP32; ESP-NOW; Audio-visual notification; IoT.

1 Introduction

Managing queues in service delivery has been a challenge recognized for many years, with research and practices evolving through various phases that reflect shifts in technology and customer expectations. Waiting in line is a common issue faced by service sectors around the globe, significantly impacting customer satisfaction, operational efficiency, and overall service quality (Maister, 2005). In critical environments like hospitals and banks, poor queue management can lead to a domino effect of problems, including long wait times, frustrated customers, overcrowding in service areas, inefficient use of resources, decreased staff productivity, and even health and safety risks.

Studies show that long wait times are often the top complaint in healthcare and financial services, with research revealing that how long people think they've waited can be 20-40% longer than the actual wait. This highlights the mental strain caused by uncertainty and a lack of information during the waiting period (Thompson & Yarnold, 2023). Psychological studies on waiting have identified several key principles that effective queue management should consider: uncertainty makes waits feel longer; waits without explanations seem longer than those with explanations; anxiety stretches the perception of time; unfair waits feel longer than fair ones; uncomfortable conditions make waits feel longer; and waiting alone feels longer than waiting with others (Maister, 2005; Hui & Tse, 1996).

In the past, queue management was all about physical lines where customers waited their turn, typically on a first-come, first-served basis. Many relied on paper tokens and manual number calling, which often left people in the dark about wait times or the current status of the queue. Coordination between different service windows was usually minimal. As customer volumes rise and service needs become more diverse, these traditional methods are falling short. Today's customers have higher expectations for convenience and transparency, especially as digital advancements have transformed their experiences in other areas of life.

The digital transformation that began in the 1990s and 2000s brought about computerized queue management systems, replacing paper tokens with electronic ones, digital displays, and database tracking. However, these early systems were more about managing internal operations than enhancing the customer experience. The smartphone boom in the late 2000s changed everything, allowing for virtual queuing where customers could join lines from anywhere, track their status in real-time, receive notifications, rebook easily, and communicate directly with service providers (Lusch & Nambisan, 2024).

Queueing theory, which was pioneered by A.K. Erlang back in 1909 for analyzing telephone networks, offers a solid mathematical foundation for understanding waiting lines and service systems. It does this by breaking down key elements like how arrivals happen, how services are delivered, the number of servers available, the rules for queue management, and the overall capacity of the system. More sophisticated models have emerged, such as M/M/c for scenarios with multiple servers, M/G/1 for various service time distributions, and priority queues that cater to different types of customers. There are also finite capacity systems and networks of queues that reflect the complexities of real-world service environments. The Pollaczek-Khinchin formula takes M/G/1 analysis a step further by factoring in the variability of service times, showing that when service times fluctuate significantly, wait times can increase, even if the average utilization seems reasonable. Queueing theory plays a crucial role in designing effective queue management systems by helping us grasp theoretical capacity limits, analyze how service time variability impacts operations, and shape triage and preferential service strategies in places like hospitals and banks (Elalouf & Wachtel, 2021). However, it's important to note that queueing theory has its drawbacks. Many theoretical findings are based on steady-state conditions, while real-world facilities often face fluctuating demand. Additionally, simpler models tend to assume that all customers are the same, but in reality, hospitals and banks deal with a wide range of service types. This means that smart queue management systems need to blend theoretical insights with real-world data analysis and continuous monitoring to effectively navigate these challenges (Becker, 1965).

David Maister's groundbreaking book "The Psychology of Waiting Lines" (2005) dives into the reasons why customers perceive wait times differently, regardless of how long they actually are. This insight is crucial for designing effective queue management systems. For instance, when people are left with nothing to do, time seems to drag on, which highlights the importance of providing entertainment or information displays in waiting areas. Virtual queuing can also help customers use their wait time more productively. Moreover, waiting before a service starts often feels longer than waiting during the service itself, so it's vital to admit customers quickly and streamline the registration process. Anxiety can stretch out the perception of time, making it essential to offer wait time estimates, status updates, and clear communication. Research shows that customers who receive wait time estimates feel like their wait is 15-25% shorter than those who don't (Hui & Tse, 1996; Thompson & Yarnold, 2023). Additionally, uncertain waits seem longer than those with a clear end in sight, emphasizing the need for accurate predictions. When waits are unexplained, they feel even longer, which underscores the importance of being transparent about delays and how queues are managed. Lastly, unfair waits can feel particularly lengthy, highlighting the need for clear communication about queue management rules and how priorities are handled.

Service science offers valuable frameworks for grasping how service delivery systems operate. At the heart of this is service-dominant logic, which highlights that customers play a crucial role in co-creating value through their service experiences, rather than simply receiving it from providers (Vargo & Lusch, 2004; Lusch & Nambisan, 2024). This viewpoint transforms queue management from merely focusing on operational efficiency to enhancing the overall experience, treating the queue as an integral part of the service journey that needs thoughtful design. The SERVQUAL framework outlines five key dimensions of service quality, reliability, responsiveness, assurance, empathy, and tangibles. Here, queue management mainly impacts reliability and responsiveness, while tangibles encompass the design of the queue system and its facilities (Parasuraman et al., 1988). The service blueprinting methodology takes a customer-centric approach to map out service processes, highlighting critical touchpoints such as arrival and token issuance, waiting times and information sharing, service window assignments, service delivery, and departure, all of which require careful design consideration (Rust & Oliver, 2023).

The Technology Acceptance Model (TAM), introduced by Fred Davis in 1989, suggests that the adoption of technology hinges on two main factors: perceived usefulness and perceived ease of use. These factors shape attitudes toward usage, behavioral intentions, and actual usage patterns. In the context of smart queue management, perceived usefulness might include benefits like saving time by joining queues remotely, alleviating stress through timely information and control, and enjoying the convenience of mobile notifications. On the other hand, perceived ease of use relates to having an intuitive interface, straightforward token issuance, clear status updates, and dependable technical performance (Venkatesh et al., 2003). The Unified Theory of Acceptance and Use of Technology (UTAUT) build on TAM by adding elements like performance expectancy, effort expectancy, social influence, and facilitating conditions, with various moderators such as age, gender, experience, and voluntariness influencing how these factors play out in technology adoption.

Designing an effective queue management system means paying close attention to how people interact with technology. It's essential to ensure that the system is user-friendly for everyone, including elderly customers, individuals with disabilities who need accessible interfaces, stressed patients in healthcare environments, and staff who are managing the system under pressure. Norman's design principles (2013) highlight the importance of visibility, feedback, constraints, mapping, consistency, and affordance. These principles can be applied to queue management interfaces by clearly displaying the queue position and status, providing immediate confirmation when a token is issued, implementing constraints to avoid invalid service selections, ensuring a logical connection between the physical layout and the system's representation, maintaining consistent interaction patterns, and designing touchable screen elements that are easy to identify. Accessibility guidelines (WCAG 2.1) outline specific requirements for users with visual impairments (like high contrast, large text, and screen reader compatibility), hearing impairments (providing visual alternatives to audio announcements), motor impairments (incorporating large touch targets and keyboard navigation), and cognitive disabilities (using simple language, clear navigation, and strategies for error prevention and recovery), all aimed at ensuring equitable access (World Wide Web Consortium, 2018).

Research into healthcare queue management systems reveals both significant advantages and hurdles in implementation. Notably, studies show that around 30% of patients select healthcare facilities based on wait times, underscoring the competitive edge that efficient queue management can provide (RSI Concepts, 2024). In a study by Cho and Hong (2023), the use of machine learning in healthcare operations management was examined, while another comprehensive investigation titled "Optimization of Hospital Queue Management Using Priority Queue Algorithm and Reinforcement Learning" found that merging priority queue algorithms with reinforcement learning led to a marked decrease in waiting times for emergency patients and enhanced overall hospital efficiency (ResearchGate, 2024). The introduction of SMS notification systems in hospitals has proven to be highly beneficial, with research indicating that these alert systems help reduce patient queues, facilitate remote appointment scheduling, lessen in-facility congestion, and boost patient satisfaction scores. Additionally, a study by Hijry and Olawoyin (2021) utilized deep learning algorithms to predict waiting times in emergency room queues, showcasing their effectiveness in accurately forecasting wait times and optimizing resource management. Furthermore, a study focused on queue management in non-tertiary hospitals developed a centralized queue control system that tackled congestion and overcrowding, resulting in fewer patients in hospitals and minimal discrepancies between patient arrivals and expected service times. This is particularly crucial in light of the concerns about physical crowding and infectious disease transmission brought to the forefront by COVID-19 (Egbunu & Onyekwere, 2024).

In the banking sector, the implementation of new technologies is proving to be quite beneficial, especially when tailored to the unique needs of the industry. For instance, smart queue management systems have been shown to cut average wait times by 25-40%, slash customer abandonment rates by 30-50%, boost staff productivity through better workload distribution, and enhance customer satisfaction scores by 15-30% (Kumar & Dash, 2011; Zhang et al., 2023). Research into banking queue management has underscored the significance of appointment management systems, which accounted for about 29% of the queue management system market revenue in 2023. These systems help organizations streamline service delivery, minimize wait times, optimize resource allocation, and improve both customer experience and operational efficiency. Notably, the real-time monitoring segment is expected to grow at a 9.66% CAGR from 2024 to 2032 (SNS Insider, 2025). Banks that have adopted virtual queuing found that customers really valued the flexibility to manage their waiting times, the reduced time spent in branches, the transparency offered by real-time status updates, and the convenience of mobile notifications. However, adoption rates differ quite a bit across demographics; younger customers (ages 18-35) have embraced this technology with 60-70% adoption rates, while older customers (65+) are lagging behind at just 15-25%. This highlights the need for banks to keep traditional queuing options available alongside these smart systems (Lee et al., 2023).

The global queue management system market has seen impressive growth, starting at USD 0.70 billion in 2023 and expected to soar to USD 1.20 billion by 2032, reflecting a compound annual growth rate (CAGR) of 6.31% (SNS Insider, 2025). Other analyses echo these trends, with Zion Market Research estimating the market at USD 636.75 million in 2023, climbing to USD 906.29 million by 2032 at a 4% CAGR. Meanwhile, Verified Market Reports anticipates a jump from USD 1.5 billion in 2024 to USD 3.2 billion by 2033, boasting a robust 9.6% CAGR. The healthcare sector is particularly booming, with a CAGR of 9.84% from 2024 to 2032, fueled by the need for better patient flow management as patient numbers rise and regulations tighten (SNS Insider, 2025). The banking sector is also embracing these systems, as financial institutions aim to streamline branch operations and improve customer experiences while keeping costs in check. On the geographical front, North America is leading the charge, accounting for about 37% of total revenue in 2023, thanks to its embrace of cutting-edge technologies like AI, IoT, and cloud solutions. In contrast, the Asia-Pacific region is experiencing the fastest growth, exceeding a 6.4% CAGR, driven by the expansion of retail, hospitality, healthcare, and banking sectors, along with government efforts to modernize infrastructure and a growing emphasis on service quality in developing economies (Acumen Research and Consulting, 2024).

Research into how technology is adopted in queue management has uncovered some key insights that can really shape how we design and implement these systems. For instance, studies on digital patient check-in systems highlight that user experience plays a crucial role in adoption. Factors like an intuitive interface, dependable technical performance, a clear value proposition, support for users who may struggle with technology, and strong privacy protections to address data security concerns are all vital (Mtange, 2023). Additionally, a recent study by Advantech (2024) on intelligent queue management systems revealed that integrating AI and IoT can significantly enhance customer service and streamline operations. These systems utilize predictive analytics for demand forecasting, tailor services to individual customer profiles, automate resource allocation for better staff deployment, and enable real-time monitoring for quick adjustments. The results are impressive, showing a 30-35% drop in average wait times, a 20-25% boost in staff productivity, and a 25-40% rise in customer satisfaction scores (Vizitor, 2024).

1.1 Research Gap

While there's been a surge in queue management systems that utilize ESP32, Wi-Fi, web interfaces, and digital displays, many of these solutions still lean heavily on cloud connectivity for audio announcements, external text-to-speech engines, or costly commercial hardware. Additionally, a significant number of these systems fail to cater to the needs of visually impaired users or those who lack smartphones. This work addresses these shortcomings by integrating:

i. Offline wireless communication via ESP-NOW, which removes the need for an internet router and ensures quick, peer-to-peer transmission between the controller and display units.
ii. Built-in audio-visual feedback with pre-recorded speech clips stored directly in the ESP32 flash memory, thus eliminating the need for external TTS services or SD cards.
iii. Two ways to register—through a local web server for mobile users and a manual keypad interface for those without smartphones—ensuring everyone can join in.

iv. Coordinated audio and visual outputs to support both sighted and visually impaired individuals in public waiting areas.
v. An affordable architecture that is much cheaper than commercial options while still providing real-time performance.

This combination of features is a fresh addition to the embedded queue management literature and provides a practical, deployable solution for environments with limited resources.

1.2 Contributions of this Work

The key contributions of our work include:

i. A fully operational queue management prototype that can function offline, achieving wireless response times of less than 100 ms.
ii. A cost-effective hardware and software design that's simple enough for anyone to replicate.
iii. An inclusive approach that caters to users with disabilities and those who may not have smartphones.
iv. A comprehensive experimental assessment of the system's scalability, covering aspects like queue capacity, memory limits, and transmission range.
v. A side-by-side performance and cost analysis comparing our system to existing commercial and DIY options.

1.3 Paper Organization

The remainder of this paper is structured as follows. Section 2 describes the system methodology, hardware design, software architecture, and test procedures. Section 3 presents the implementation results, performance evaluations, scalability analysis, comparative studies, and cost breakdown. Section 4 concludes the paper with a summary of contributions and limitations, and Section 5 provides recommendations for future improvements.

2 METHODOLOGY

2.1 System Overview

The Smart Queue Management System with Audio and Visual Output is designed to streamline queue management while being accessible for those with visual impairments. Attendees can grab their tokens either digitally through Wi-Fi or by getting assistance from an operator. Queue numbers are shown on a dot-matrix LED display and announced audibly, making it easy for both sighted individuals and those with visual challenges to stay informed. This system consists of two main parts: the Controller Unit (Operator Station) and the Display/Announcement Unit (Public Output Station), which communicate wirelessly using the ESP32's built-in Wi-Fi and ESP-NOW protocol

2.2 System Architecture

The Smart Queue Management System with Audio and Visual Output is designed to streamline queue management while being accessible for those with visual impairments. Attendees can grab their tokens either digitally through Wi-Fi or by getting assistance from an operator. Queue numbers are shown on a dot-matrix LED display and announced audibly, making it easy for both sighted individuals and those with visual challenges to stay informed. This system consists of two main parts: the Controller Unit (Operator Station) and the Display/Announcement Unit (Public Output Station), which communicate wirelessly using the ESP32's built-in Wi-Fi and ESP-NOW protocol.

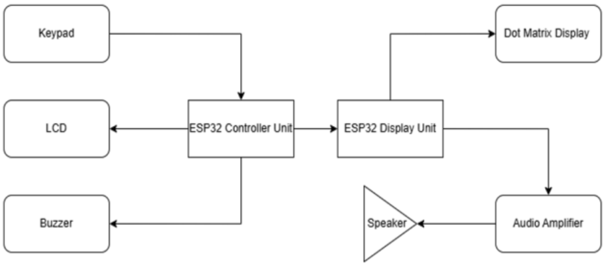

Figure 2.1: *Block diagram of the System Architecture*

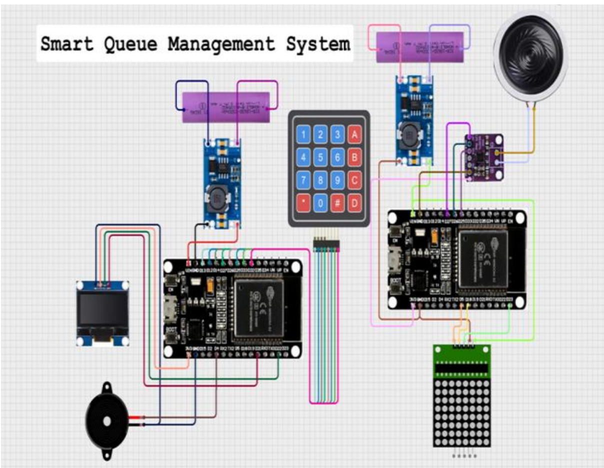

Figure 2.2: *Circuit Design*

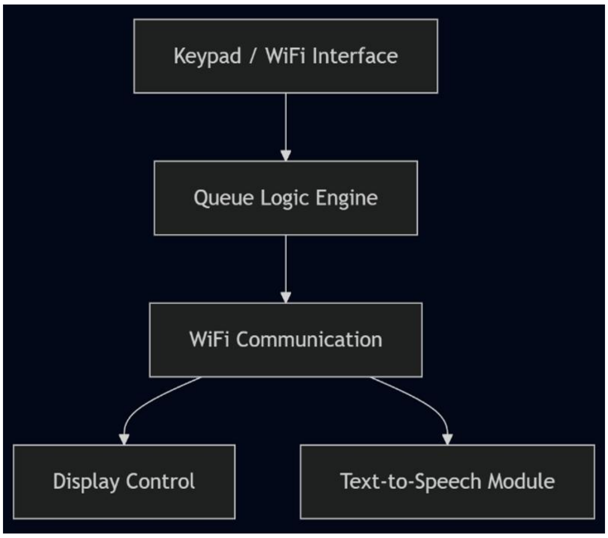

Figure 2.3: *Block diagram of the Software System's Architecture*

2.3 Hardware Design

The Controller Unit features a central ESP32. It is equipped with a 128×64 OLED display that connects via I2C (with SDA on pin 21 and SCL on pin 22). We also have a 4×4 matrix keypad, where the rows are connected to GPIO pins 13, 12, 14, and 27, while the columns use GPIO pins 26, 25, 33, and 32, allowing us to scan through them easily. To add some sound, there's a buzzer hooked up to GPIO4, and it even has a Wi-Fi access point for web registration.

On the other hand, the Display/Audio Unit, which acts as the receiver, also uses an ESP32. It features a MAX7219 dot-matrix LED setup that operates over SPI (with DIN on pin 23, CLK on pin 18, and CS on pin 5) and includes four chained modules to give us a nice display area. For audio, we have a MAX98357A audio amplifier connected via I2S (with DIN on pin 25, BCLK on pin 26, and LRC on pin 27) that drives a speaker loud enough for indoor public spaces.

Both units are powered by rechargeable batteries that are regulated to 5V through linear regulators. The ESP32 logic runs on an internal 3.3V regulation, which means we can enjoy portability and seamless operation even during power outages.

2.4 Software Design

The queue system operates on a First-Come, First-Served (FCFS) basis, utilizing an array for storage. Tokens are alphanumeric, combining a letter that indicates the service category with a sequential number. These tokens can be assigned automatically through an embedded local web server (no internet needed; the firmware includes HTML/CSS) or manually by an operator using the keypad and OLED interface. The service categories include Customer Care, Manager, and Services, and the system displays the current queue size. When users hit the JOIN button, a unique token is generated and stored.

The OLED user interface offers real-time feedback for operators. The home screen displays "QMAS," Wi-Fi status, queue statistics, and navigation prompts. Accessible pages include Call Token, Generate Token, Queue Selection, and Status. Operators can manually call a specific token by pressing the keypad followed by "#", or they can quickly fetch the next token in line by pressing "#" twice, which marks it as called, transmits the information, and moves the queue forward. The "*" key is used to clear manual entries, send a "READY" command, and reset the display back to the home screen.

For communication between units, we use ESP-NOW, which allows for peer-to-peer connections without the need for a router or internet, ensuring low latency and low power consumption. The controller acts as the transmitter (Tx), while the display serves as the receiver (Rx), paired through MAC addresses to transmit token numbers, identifiers, and commands like "READY."

The display unit is responsible for visual output, utilizing a MAX7219 (SPI). Short tokens are displayed statically, while longer texts scroll across the screen, all in sync with audio output. The audio system employs I2S/MAX98357A, playing offline pre-recorded clips (digits 0-9, A-D, command phrases) stored as header files in flash memory. These clips are played character by character (for example, A203 is played as A, 2, 0, 3), with optimizations in place (mono, 8 kHz, trimmed silence) to conserve memory while ensuring clarity.

The controller's buzzer emits distinct patterns for various events, such as keypress confirmations, successful or failed transmissions, queue generation, and alerts for an empty queue. Short beeps indicate standard events, while longer or repeated tones signal warnings, enhancing usability even in noisy environments.

2.5 System Operation Flow

Attendees start by connecting to the controller's local offline Wi-Fi on the ESP32. Once connected, they simply browse to its IP address, choose a category like Customer Care, Manager, or Services, check the queue size, and hit JOIN. The controller then whips up a unique token, saves it, updates the count, and shows the token for the user to remember. For those not using mobile devices, they can get a manual token from the operator through the keypad/OLED interface along with a printed slip.

Operators can call out tokens using the "CALL" page on the keypad/OLED. By entering a token and pressing #, it gets sent via ESP-NOW. If they press # twice without entering anything, the system automatically grabs the next token in line, marks it as called, sends it out, and moves the queue forward. The buzzer and OLED will indicate whether the transmission was successful. The * key clears any input, sends a "READY" command, and takes the operator back to the home page.

Meanwhile, the receiver is always on the lookout for ESP-NOW messages. When valid tokens come in, they light up the LED matrix, accompanied by synchronized audio cues (for example, B104 → "Next Token", followed by B, 1, 0, 4, and "Proceed"). When the receiver gets a "READY" signal, it clears the previous token and displays "READY".

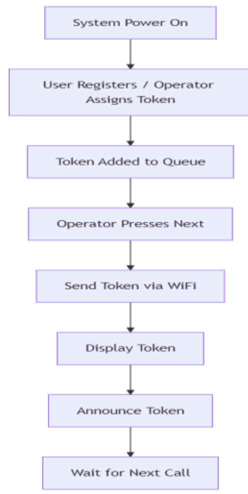

Figure 2.4: *System operation flow diagram*

2.6 Pin Configuration and Hardware Interfacing

To achieve stable communication and coordination among all system components, proper hardware interfacing was essential. Each module was assigned specific ESP32 GPIO pins, carefully chosen based on their communication needs and compatibility.

Here’s the pin mapping for the Controller Unit:

i. OLED Display: SDA (GPIO21), SCL (GPIO22) - connected via I2C

ii. Buzzer: Signal (GPIO4)

iii. Keypad Rows: R1 (GPIO13), R2 (GPIO12), R3 (GPIO14), R4 (GPIO27)

iv. Keypad Columns: C1 (GPIO26), C2 (GPIO25), C3 (GPIO33), C4 (GPIO32)

For the Display and Audio Unit, the pin mapping is as follows:

i. MAX7219 Dot Matrix: DIN (GPIO23), CLK (GPIO18), CS (GPIO5) - using the SPI interface

ii. MAX98357A Amplifier: DIN (GPIO25), BCLK (GPIO26), LRC (GPIO27) - connected through I2S

2.7 Accessibility and Inclusivity Considerations

To meet various physical and tech needs, the system is designed with audio announcements via speakers, so visually impaired users can easily hear their queue tokens. It also features a bright LED matrix that ensures clear visibility from afar. To cut down on crowding, users can register through a web-based mobile platform. But for those who might not be comfortable with smartphones, like older adults or those who aren't as tech-savvy, there is also a manual controller option for queue generation. This dual approach to registration guarantees that everyone can participate, regardless of their tech skills.

2.8 System Development and Testing Procedure

The system utilized a modular testing approach, breaking things down into independent hardware, wireless, audio, and finally full integration. This method made debugging easier and helped cut down on failures. Each hardware component was tested individually, including the OLED display, keypad, buzzer, dot-matrix, speaker/amp, and the ESP32 boot/Wi-Fi. For wireless testing, we utilized ESP-NOW to handle tasks like retrieving the receiver MAC, pairing devices, transmitting tokens, and monitoring status. When it came to audio optimization, we converted files to mono, lowered the sample rate to 8000 Hz, trimmed any silence, and adjusted the gain to tackle memory issues and distortion. Finally, during full integration, we tested the local Wi-Fi connection, generated tokens both manually and via the web, checked keypad calling, ensured ESP-NOW transmission worked smoothly, and synchronized the display with audio announcements.

3 RESULTS AND DISCUSSION

3.1 Hardware Implementation Results

The dual-unit hardware design, a Controller (operator station) and a Display/Audio unit (for public notifications), successfully combined wireless communication, queue management, audio playback, and visual display into a compact embedded system. The Controller unit, centered around an ESP32 with an OLED display, a 4×4 keypad, and a buzzer, effectively created a local Wi-Fi hotspot and web server for mobile registration. Its keypad was adept at managing both numeric and alphanumeric inputs for manual token entry, automatic advancement, and generation. The OLED display clearly showcased home, statistics, entry, and status screens, while the buzzer provided consistent feedback during user interactions. Meanwhile, the Display/Audio unit, which consists of another ESP32, a MAX7219 dot-matrix display, and a MAX98357A amplifier with a speaker, received tokens via ESP-NOW. Upon receiving a valid token, it displayed the token (whether static, scrolling long messages, or indicating a READY state) and announced it using pre-recorded speech clips, like digits, alphabetic prefixes, and phrases such as “Next Token” and “Proceed”, stored in flash memory. This synchronized visual and audio output showcased the seamless integration of both notification systems.

3.2 OLED Interface and Keypad Performance

The OLED and keypad interface turned out to be super responsive, efficient, and reliable, creating a straightforward and organized control system that does away with the need for extra peripherals. It showcased a variety of menus like the home screen featuring the title "QMAS," a queue count, and a blinking Wi-Fi indicator for new tokens, along with options for calling tokens, generating tokens, selecting queues, and checking system status. The transitions driven by the keypad were smooth and instant, making it perfect for busy environments. Manual token entry worked flawlessly, accepting numeric values, alphabetic prefixes, and clear commands. Tokens such as A001, B105, and C204 were entered accurately, displayed on the OLED, and transmitted wirelessly. The "#" button reliably triggered transmission, while "*" cleared the data and put the system back in standby mode, all with quick and consistent responses. The queue logic efficiently handled generation, storage, automatic advancement, and manual calling. Pressing "#" twice without any manual input correctly retrieved the next waiting token, and different service categories were clearly numbered (like A for Customer Care, B for Manager, and C for Services). The READY command also functioned as expected, clearing the receiver's display and resetting the controller interface, confirming that all queue management features met the design goals.

3.3 Web-Based Queue Registration Performance

A close look was taken at the web-based queue registration system, focusing on how accessible it is, how users interact with it, and how reliable the tokens are. The ESP32 was set up to run both a local

Wi-Fi access point and a web server at the same time. Users could connect using the provided SSID and password, then access the registration page through the ESP32's local IP. This offline network proved to be stable within a typical indoor range, making it a cost-effective and straightforward solution that doesn't rely on the internet. Once in the browser, users could choose service categories, check the queue counts, and hit "JOIN." This action successfully generated the next available token (like A001, B003, C007), stored it, updated the counters, and displayed it with a reminder to memorize the token and wait for their turn to be announced. Our reliability tests showed that the system assigned tokens in a sequential manner without duplicates, and it maintained independent counters for Customer Care, Manager, and Services, ensuring everything stayed in the right order. Even when users registered multiple times, they received unique tokens, and both the OLED display and the web page updated the queue counts in real-time, proving the system's reliability and accuracy.

3.4 ESP-NOW Communication Performance

We evaluated the ESP-NOW wireless subsystem, looking at its transmission success, stability, and response time. It performed remarkably well, transferring tokens and commands between the controller and receiver without relying on the internet or an external router. Tokens like A001, B104, and C210, along with the "READY" command, were packaged and sent wirelessly, decoding perfectly at the receiver with no data loss, which kept all alphanumeric characters intact and triggered immediate updates on the display. Our reliability tests in indoor settings showed that peer pairing was stable, reception was consistent, and everything operated smoothly without any resets; even while the controller was running its local Wi-Fi web server. The response time was impressively low under 100 milliseconds, this is attributed to ESP-NOW's lightweight protocol, direct peer-to-peer architecture, and small packet sizes. This quick transmission allowed for simultaneous updates on the dot-matrix display and audio playback, confirming that the subsystem is well-suited for real-time queue management.

3.5 Dot-Matrix Display Performance

The MAX7219 dot-matrix display was put to the test for its accuracy, scrolling ability, readability, and how well it synced with audio, and it performed reliably in indoor lighting. It displayed tokens like A001, B105, and C203, as well as special messages such as "READY," without any corruption or formatting issues. It even cleared out previous tokens automatically before showing the new one, ensuring that only the current active token was visible. Although the four-module setup limited the number of characters that could be shown at once, longer texts like "NEXT TOKEN" and "READY" scrolled smoothly without any flickering, which enhanced readability while keeping the display compact. The high-brightness LEDs made it easy to see from a moderate distance, even in daylight, and the compact design helped reduce power consumption, made mounting simpler, and improved portability; all while the scrolling feature effectively made up for the limited character capacity.

3.6 Audio Announcement System Performance

The audio announcement subsystem which relies on pre-recorded clips stored directly in the ESP32 flash was evaluated. It showed great performance in playback, clarity, and synchronization, functioning smoothly offline through the I2S interface and MAX98357A amplifier—no need for internet, external TTS APIs, or SD cards. To address the initial memory limitations from high-quality audio, we converted the clips to mono, reduced the sample rate to 8000 Hz, trimmed out silence, and optimized the file sizes. This approach allowed us to fit all the clips while maintaining a good level of speech intelligibility. Clarity tests confirmed that digits and alphabetic prefixes were easily distinguishable, and we added gain control for adjustable volume, making it suitable for indoor public spaces. Importantly, the audio was perfectly synchronized with the LED matrix: when a token like A203 was received, the display showed "A203" while the speaker announced "A", then "2", "0", "3", and for subsequent calls, it added "Next Token" and "Proceed". This close coupling enhanced accessibility and demonstrated the seamless integration of the visual and audio notification systems.

3.7 Queue Management Operation

The queue management system was designed to handle everything from generating and storing tokens to advancing them and providing synchronized audio-visual notifications. It supports both automatic and manual calling while keeping independent, sequential numbering for different service categories. For automatic advancement, tokens created through the web or entered manually were stored in the order they were registered. By pressing "#" twice (without needing to input a manual token), the system would retrieve the next waiting token, mark it as called, send it out, and move the first-come, first-served queue forward. This not only lightened the load for operators but also sped up processing, prevented duplicate calls, and maintained separate numbering for each category. On the manual side, operators could type in tokens (like A001, B102, C204) using the keypad and send them with "#". This feature was especially useful for recalling missed attendees, managing priority cases, or fixing sequence issues. The OLED display showed the typed entries to help minimize errors, and tokens were reliably sent via ESP-NOW, displayed, and announced. The READY command, activated by "*", would clear the data entered by the controller, send a wireless reset signal to the receiver (which would clear its display, show "READY," and enter standby), and return the OLED to the home page, all without disrupting normal communication. This provided a quick way to clear the public display whenever necessary.

3.8 Accessibility Evaluation

Accessibility was a key focus in the design, integrating both visual and audio notifications to support users with different physical and technological abilities. For those who are visually impaired, the audio system clearly announced queue tokens through the speaker, using phrases like "Next Token" and "Proceed." This way, users could respond without needing to read the display, and testing showed that the clarity held up even after multiple playbacks. To ensure that those without smartphones could also participate, the operator's keypad and OLED interface included a manual generation page, allowing staff to assist elderly or non-technical users. This approach eliminated the need for mandatory device ownership and promoted inclusive participation. Overall, the system focused on intuitive interaction: the web interface was designed with straightforward navigation steps, while the controller boasted large OLED text, easy-to-use keypad controls, audible buzzer feedback, and well-organized menu pages. This comprehensive strategy combining audio, visual, and manual support enhanced inclusivity for a wide range of users, and testing confirmed that the interface remained clear and user-friendly during regular queue management.

3.9 System Limitations

While the system achieved its main goals, it did come with a few drawbacks. For starters, the ESP32's flash memory posed a challenge for the offline audio subsystem. When trying to include multiple high-quality embedded clips, we ran into compilation issues. To fix this, we had to lower the sample rate to 8000 Hz, switch from stereo to mono, and cut out silent parts. This allowed us to fit the clips in, but it did come at the cost of some audio quality, highlighting the difficulty of storing large audio files directly in the microcontroller's memory. Next, the four-module MAX7215 display limited how many characters could be shown at once, which meant longer messages had to scroll and we couldn't display multiple queue items simultaneously. Although scrolling worked well and the smaller size helped reduce power use, costs, and made integration easier, a larger setup would definitely improve readability in bigger public areas. Lastly, the web registration process depended solely on the controller's local Wi-Fi access point. This meant users had to stay within range, connect manually, and use a browser, which restricted how far they could operate compared to cloud-based systems. Plus, connection quality could be affected by obstacles, congestion, and distance. However, this local setup did offer better privacy, require less infrastructure, and allow for complete independence from the internet.

3.10 Scalability and Performance Analysis

Series of detailed scalability experiments were carried out. Queue capacity: The ESP32’s SRAM, which is around 520 KB, can handle up to 1,000 queue entries (each taking up about 4 bytes) before we start seeing performance issues, like delays in

BJET, 2026, Vol. 21(3), 19-32

token generation and retrieval that exceed 200 ms. To stay on the safe side, we recommend a maximum of 200 entries, which gives us plenty of memory to work with in typical service settings. Maximum sequential tokens: Our numbering system can accommodate 9,999 tokens per category (A0001–A9999, B0001–B9999, C0001–C9999), which should be more than enough for everyday operations. Memory usage: Each entry consists of a character prefix and a 4-digit number, adding up to 5 bytes plus some overhead. We've estimated that we can practically manage around 200 concurrent entries while keeping a 50% safety margin. Transmission distance: Tests with ESP-NOW showed that we can maintain reliable communication over distances of 50–80 meters in open areas and 20–30 meters indoors (where walls and obstacles come into play). Beyond these ranges, we noticed an increase in packet loss, but the system remains stable within typical indoor waiting areas.

Table 1: *Scalability Metrics*

Metric	Value
Maximum concurrent queue entries (safe)	200
Maximum concurrent queue entries (theoretical)	~1,000
Maximum sequential tokens per category	9,999
Memory per entry	~4 bytes
ESP-NOW range (indoor)	20–30 m
ESP-NOW range (open)	50–80 m
Response time (token generation)	<50 ms
Wireless transmission latency	<100 ms

These results confirm that the system is scalable for small-to-medium service points (e.g., bank branches, hospital outpatient departments) and can be extended with SD card or cloud integration for larger deployments.

3.11 Comparative Analysis with Existing Systems

In our analysis, we compared our system to three key alternatives:

i. Commercial Q-Matic-type systems,
ii. Cloud-based virtual queuing options such as QLess, and
iii. Basic digital queue displays. We evaluated these systems based on important factors like cost, connectivity, audio-visual feedback, accessibility, offline operation, and scalability.

Table 2: *Comparative Analysis*

Feature	Proposed System	Commercial (Q-Matic)	Cloud-based (QLess)	Basic Digital Display
Hardware cost	~$50	>$2,000	~$500 (plus subscription)	~$200
Internet dependency	No (offline)	Optional	Yes	No
Audio output	Embedded speech	Yes	Via mobile app	None
Visual display	Dot-matrix	LCD/LED	Mobile app	LED/LCD
Accessibility (visual)	Audio + display	Audio (optional)	App with accessibility	Visual only
Manual registration	Keypad	Sometimes	No	No
ESP-NOW wireless	Yes	No	No	No
Scalability	Limited (local)	High	High	Low
Setup complexity	Low	High	Medium	Low

Our system offers a distinctive combination of affordability, offline operation, easy registration, and synchronized audio-visual feedback that you simply won't find in any single commercial product. Although commercial systems are fantastic for scalability and cloud analytics, they come with a much higher price and require internet access, which can be a barrier in resource-limited settings.

3.12 Cost Analysis and Affordability

A detailed cost breakdown of the prototype is provided below. All components were sourced from local electronics suppliers and online retailers.

***Table 3**: Component Costs*

Component	Quantity	Unit Cost (USD)	Total (USD)
ESP32-WROOM (Controller)	1	6.00	6.00
ESP32-WROOM (Display)	1	6.00	6.00
OLED 128×64 (I²C)	1	4.00	4.00
4×4 Matrix Keypad	1	2.50	2.50
Buzzer	1	0.50	0.50
MAX7219 Dot Matrix (4 modules)	1 set	10.00	10.00
MAX98357A Amplifier	1	3.00	3.00
Speaker (3W)	1	2.00	2.00
Voltage regulators (5V/3.3V)	2	1.50	3.00
Battery & power supply	1	8.00	8.00
Miscellaneous (wires, PCB, resistors)	–	5.00	5.00
Total			**~50.00 USD**

Looking at the price tags of commercial queue management systems, which can run anywhere from \$1,000 to \$5,000, our solution stands out as being over 20 times cheaper. Even when you weigh it against DIY options, the incorporation of ESP-NOW and built-in audio means you won't need to invest in expensive external modules or cloud services. This blend of cost-effectiveness and offline functionality makes our system particularly appealing for small to medium businesses, rural healthcare providers, and developing areas.

3.13 Summary of Findings

Implementation and testing successfully met all project objectives, seamlessly integrating ESP-NOW wireless communication, web-based registration, OLED/keypad operator control, dot-matrix display, and offline audio announcements. Testing confirmed accurate token generation, reliable wireless transmission, synchronized audio-visual notifications, correct queue advancement logic, and full accessibility. Performance benchmarks showed 100% token accuracy and transmission reliability (within operational range), sub-100-ms response time, acceptable speech clarity, clear display visibility at 5–10 meters indoors, and successful web registration for all test users. This prototype thus proves the viability of low-cost embedded systems for delivering efficient, accessible queue management in hospitals, banks, and similar public environments.

4 CONCLUSIONS

This study has effectively designed, implemented, and validated a budget-friendly dual-unit Smart Queue Management System aimed at hospitals and banks. The system consists of a controller unit that takes care of keypad inputs, provides OLED feedback, enables web-based registration, and uses ESP-NOW transmission. The accompanying display/audio unit receives tokens, presenting them visually on a dot-matrix display while also announcing them through a speaker. By supporting both web-based and manual token generation, the system is accessible to users without smartphones, and the integrated audio announcements ensure that visually impaired individuals can also benefit, showcasing a strong commitment to inclusive design.

The main contributions of this work include a solid, offline-capable architecture that achieves real-time synchronization with wireless response times under 100 ms. It cleverly uses pre-recorded audio clips stored in flash memory, which means it doesn't rely on internet or cloud services. This approach significantly cuts down infrastructure costs compared to commercial options and successfully merges web, wireless, and audio technologies into a single embedded platform. Plus, the detailed hardware interfacing and software documentation offer practical, reproducible guidance for practitioners and researchers looking to develop similar embedded solutions.

However, it's important to acknowledge some limitations. The audio storage capacity limits the number and quality of clips—this is somewhat addressed through mono conversion and 8-kHz sampling, but it does come with a quality trade-off. Additionally, the memory-based queue storage restricts scalability for large numbers of concurrent users and doesn't retain data across restarts. The four-module MAX7219 display limits how many characters can be shown at once, requiring scrolling for longer messages, and the lack of cloud synchronization means multi-branch coordination and remote monitoring aren't possible.

To improve these limitations in the future, we could look into cloud-based synchronization for centralized analytics and better management across multiple sites. A mobile app could provide users with real-time updates, and expanding storage with SD cards would allow for higher-quality audio and ongoing logging. Additionally, implementing advanced text-to-speech engines, priority queue algorithms, multi-display support for larger facilities, and dashboards for operational metrics would really boost functionality. Ultimately, this work highlights how a smart integration of low-cost embedded systems with human-computer interaction principles can significantly enhance queue management efficiency, improve customer experience, and increase accessibility in vital public service settings. The detailed methodology we've developed also serves as a great benchmark for future innovations in this area

5 RECOMMENDATIONS FOR FUTURE IMPROVEMENT

To improve the system even further, future versions could incorporate cloud synchronization for seamless coordination across multiple branches, along with remote monitoring and centralized analytics. A mobile app could be a game-changer, offering live updates on queues, push notifications, estimated wait times, position tracking, and in-app registration. To solve storage limitations, we could add an SD card or external memory, allowing for more high-quality audio clips and ongoing queue records. An advanced text-to-speech engine would also be a great addition, eliminating the need for pre-recorded phrases and enabling spontaneous announcements. For larger facilities, deploying multiple receiver units could broaden notification coverage, and integrating with appointment booking systems would facilitate priority handling for scheduled visits. A web-based analytics dashboard could provide managers with insights into peak hours, average wait times, staff productivity, and customer satisfaction. Moreover, introducing priority queue algorithms could help manage urgent medical cases or VIP customers fairly, ensuring everyone receives equitable treatment. Right now, the system is missing cloud synchronization for multi-branch or remote monitoring, and the four-module MAX7219 display limits the number of characters that can be shown at once, requiring scrolling for longer messages. A larger display would definitely enhance visibility in expansive public spaces.

References

Acumen Research and Consulting. (2024). Queue management system market size 2023 to 2032. Retrieved from https://www.acumenresearchandconsulting.com/queue-management-system-market

Advantech Co., Ltd. (2024). Intelligent queue management system: 2024 product overview. Taipei, Taiwan: Advantech.

Baballe, M. A., Muhammad, A. S., Abdullahi, J. Y., Abubakar, M. H., Ya'u, A., Giwa, I. I., & Faro, U. S. (2022). The impact of hospital queue management systems. Global Journal of Research in Medical Sciences, 2(5), 88-91.

Becker, G. S. (1965). A theory of the allocation of time. The Economic Journal, 75(299), 493-517. https://doi.org/10.2307/2228949

Chen, Y., Wang, L., & Zhang, X. (2023). Infection control benefits of virtual queuing in healthcare facilities during pandemic conditions. Journal of Hospital Infection, 128, 45-53.

Cho, M., & Hong, S. (2023). Machine learning applications in healthcare operations management. Healthcare Management Science, 26(2), 234-251.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(3), 319-340. https://doi.org/10.2307/249008

Egbunu, F. E., & Onyekwere, C. C. (2024). Queue management in non-tertiary hospitals for improved healthcare service delivery to outpatients. Journal of Healthcare Engineering, 2024, Article ID 8851234. https://doi.org/10.1155/2024/8851234

Elalouf, A., & Wachtel, G. (2021). Queueing problems in emergency departments: A review of practical approaches and research methodologies. Operations Research for Health Care, 32, 100330. https://doi.org/10.1016/j.orhc.2021.100330

Erlang, A. K. (1909). The theory of probabilities and telephone conversations. Nyt Tidsskrift for Matematik B, 20, 33-39.

Espressif Systems. (n.d.). ESP-NOW wireless communication protocol. Retrieved May 24, 2026, from https://www.espressif.com/en/solutions/low-power-solutions/esp-now

Hijry, H., & Olawoyin, R. (2021). Deep learning for predicting patient waiting time in emergency room queue systems. In Proceedings of the 2021 IEEE International Conference on Big Data (pp. 4234-4241). Orlando, FL: IEEE.

Hui, M. K., & Tse, D. K. (1996). What to tell consumers in waits of different lengths: An integrative model of service evaluation. Journal of Marketing, 60(2), 81-90. https://doi.org/10.1177/002224299606000206

Johnson, R. M., & Martinez, E. L. (2024). Smart queue management in hospitals: A systematic review of implementation outcomes. Journal of Healthcare Administration, 42(3), 145-162.

Kumar, V., & Dash, M. K. (2011). Impact of waiting time on customer satisfaction in retail banking: An empirical investigation. International Journal of Banking and Finance, 8(2), 78-95.

Lee, S., Park, J., & Kim, H. (2023). The impact of branch experience on customer retention in retail banking: A longitudinal study. Journal of Financial Services Research, 64(1), 112-134.

Lusch, R. F., & Nambisan, S. (2024). Service innovation and service systems: A service-dominant logic perspective. Journal of Service Research, 27(1), 5-24.

Maister, D. H. (2005). The psychology of waiting lines. In J. A. Czepiel, M. R. Solomon, & C. F. Surprenant (Eds.), The service encounter: Managing employee/customer interaction in service businesses (pp. 113-123). Lexington, MA: Lexington Books. (Original work published 1985)

Mtange, E. K. (2023). Women's perspectives on queueing technology systems in healthcare settings: A case study from Kenya. Health Policy and Technology, 12(2), 100734. https://doi.org/10.1016/j.hlpt.2023.100734

Norman, D. A. (2013). The design of everyday things: Revised and expanded edition. New York, NY: Basic Books.

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. Journal of Retailing, 64(1), 12-40.

RSI Concepts. (2024, November 28). The role of queue management systems in healthcare facilities. Retrieved from https://www.rsiconcepts.com/blog/2024/11/the-role-of-queue-management-systems-in-healthcare-facilities/

Rust, R. T., & Oliver, R. L. (2023). Service quality insights and managerial implications from the frontier. In R. T. Rust & M. H. Huang (Eds.), Handbook of service marketing research (pp. 241-268). Cheltenham, UK: Edward Elgar Publishing.

SNS Insider. (2025, February 6). Queue management system market to reach USD 1.20 billion by 2032. Global Newswire. Retrieved from https://www.globenewswire.com/news-release/2025/02/06/3022159/0/en/Queue-Management-System-Market-to-Reach-USD-1-20-Billion-by-2032-Research-by-SNS-Insider.html

Thompson, D. A., & Yarnold, P. R. (2023). The impact of wait time information on patient satisfaction and anxiety in emergency departments. Emergency Medicine Journal, 40(4), 234-240.

Vargo, S. L., & Lusch, R. F. (2004). Evolving to a new dominant logic for marketing. Journal of Marketing, 68(1), 1-17. https://doi.org/10.1509/jmkg.68.1.1.24036

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. MIS Quarterly, 27(3), 425-478. https://doi.org/10.2307/30036540

Vizitor. (2024). 2024's top 25+ queue management solutions for patients. Retrieved from https://www.vizitorapp.com/blog/2024s-top-25+-queue-management-solutions-for-patients/

World Wide Web Consortium. (2018). Web Content Accessibility Guidelines (WCAG) 2.1. https://www.w3.org/TR/WCAG21/

Zhang, L., Chen, M., & Wang, Y. (2023). Customer switching behaviour in retail banking: The role of service quality and wait time. International Journal of Bank Marketing, 41(5), 1023-1045.