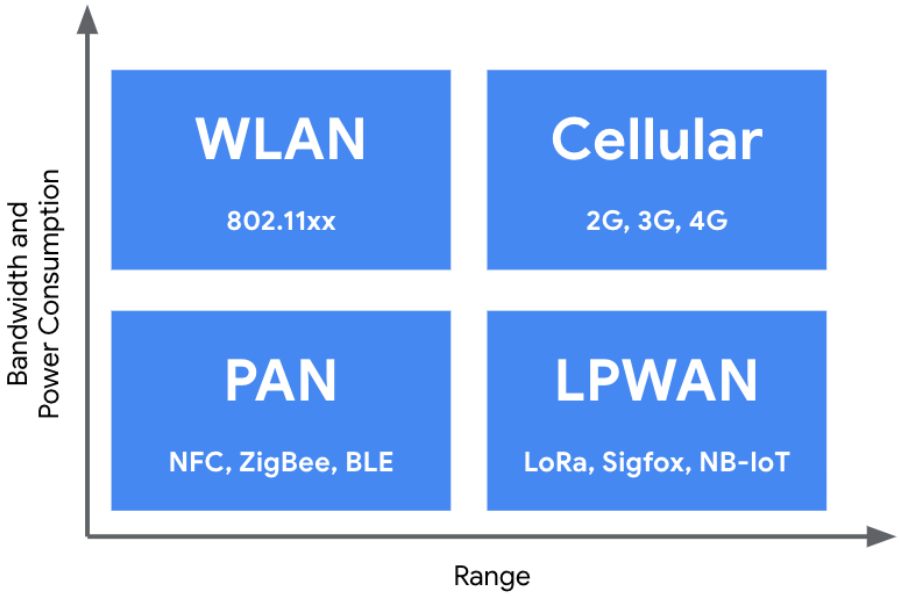
TinyML Communication Challenges & Technologies for Continuous Monitoring



About This Reading

Clearly, it is obvious that we have to monitor for model drift to ensure that the data distributions do not shift from what we expect when the ML model is serving embedded inference in the field. But to ensure that the data distribution has not changed, we need to monitor the TinyML devices that are in the field. This is where it becomes crucial for us to understand the pros and cons of different communication methodologies we have access to for TinyML devices. Continuously monitoring for data is very expensive because it drains the battery life extremely fast. In practice, nearly 50% of TinyML device budget can sometimes be allocated just for communication purposes over the device's projected lifetime. Given that we allocate so much energy resource for communication, it is crucial to understand the different technologies that exist, so that we can make informed decisions when we want to continuously monitor.

Overview of Communication Technologies

Over the past decade, we have seen massive improvements in the capabilities of wireless technologies. Twenty years ago, the prospect of millions of people transmitting photos, streaming videos online, and sending text messages over-the-air simultaneously would have been unthinkable. Nowadays, this situation is the norm, all enabled by increasingly sophisticated communication technologies. Given that intelligent IoT systems, like those leveraging TinyML, are network-enabled, it is good for us to get a general understanding of what communication technologies exist, how they work, and how they differ from one another.

Communication technologies are generally compared using certain metrics. Most commonly these are (1) functionality, (2) data speed, and (3) operating range. Depending on our system requirements, these metrics can help us to determine which communication technology is most appropriate for our given application. For example, utilizing a long range and high throughput communications protocol to send small amounts of data over a relatively short range (say, 10 meters) would be unnecessary, and likely consume other resources, like power, unnecessarily.

Peer-to-peer Communication

The first important type of wireless technology to discuss is peer-to-peer communication. This is a form of direct communication - for example, connecting wireless headphones to your smartphone, or using AirDrop to send files on Apple devices. The most prevalent protocol in this space is Bluetooth Classic. Bluetooth has a relatively short operating range, typically around 10 m, which also means that it is a low-power protocol in comparison to other technologies. More recently, we have seen Bluetooth Low-Energy (BLE) and Zigbee become popular alternatives, which are even more power efficient technologies than Bluetooth Classic. One of the disadvantages of these technologies is their relatively low data speeds. Thus, Bluetooth is advantageous for short-range communications with low data transmission rates.

The next most prevalent peer-to-peer technology is WiFi Direct. This is not quite the same as WiFi, since it does not require an access point, but uses the same frequency (2.4 or 5 GHz) for communication between devices and has comparable speed and throughput. The major advantage of WiFi Direct over Bluetooth is that data speeds are around 100x faster, but this comes at the cost of increased power consumption.

Another peer-to-peer communication technology that is becoming increasingly commonplace is Near Field Communication (NFC). This is the technology that is being used for contactless payment systems - every time you buy a coffee by tapping your credit card against the screen instead of inserting the card, or paying for something with ApplePay, a smartwatch, or similar item. NFC communicates by using an electromagnetic field generated by two coupled coils, instead of emitting radio waves like in other forms of wireless technologies. The main benefit of NFC is that it can be entirely passive, requiring zero power. This is why it can be utilized by credit cards even though we never need to charge our credit cards. Wireless charging mats also utilize NFC to charge your devices. One of the caveats of NFC is that it only works over extremely short ranges, on the order of a few centimeters, and so is not suitable for even most short-range applications.

Mesh Technologies

Moving on from peer-to-peer communications, which are *one-to-one* networks, we now look at mesh networks, which are *many-to-many* networks. There are two subsets of mesh technologies that we will look at. Firstly, technologies that offer low-power, short-range, and low data rates, and secondly, those offering low-power, long range communications.

Low Power + Short Range Mesh Technologies

There are four that fall into this range: BLE, Zigbee, Z-Wave, and 6LoWPAN. These are ideal for battery-powered devices that need to transmit information over a relatively short range.

One of the major benefits of mesh technologies is that they allow communication through other nodes in the network, which essentially act like middlemen. Instead of two nodes, A and B, needing to be directly connected in a peer-to-peer network to communicate with each other, in a mesh network, these can communicate between a third node, C. Thus, with a large enough network, data could still be sent over relatively long distances using a short-range and low-power communication technology.

BLE is the most common of these communication technologies, and is what was used on the Arduino Nano BLE Sense used in our prior courses. This allows data speeds up to 1 Mbps (Bluetooth Classic provides 2-3 Mbps) and mesh sizes for up to 32,767 devices. The operating distance for BLE is somewhere between 50-100 m, and so this could allow for meshes that span many miles if configured properly. Unless you have a good reason to use a different communication technology, BLE is often the best choice. It is the most widely supported of the technologies, easy to set up and configure, and consumes very little power.

Zigbee offers similar capabilities to BLE and works at the same frequency (2.4 GHz). However, one thing that makes it stand out is that it can support mesh sizes up to a staggering 65,000 devices. Zigbee is typically used for home automation technologies, and is thus widely supported in this application area for things like smart lighting, thermostats, energy monitoring, security systems, doorbells, etc.

Z-Wave is a competing technology for Zigbee and BLE. The main thing that stands out about Z-Wave is its lower frequency of 908 MHz for communication. A lower frequency means longer operating range, and also since it does not use the saturated 2.4 GHz band that many devices, such as microwaves, tend to use, there is significantly less interference. The disadvantage of this technology is the reduced data transmission rate as a result of the lower communication frequency. Also, the maximum supported mesh size of Z-Wave is only 232 devices, which is significantly less than BLE or Zigbee, but probably still sufficient for most applications.

6LoWPAN is a competitor for Zigbee in the space of home automation, and is differentiated by its use of an IP-based network similar to WiFi.

Low power + long distance technologies

Many important IoT applications require low-power and long-distance communications, and the same is likely to be true for TinyML. Often, the type of network used for this application is called a Low Power Wide Area Network (LPWAN). This type of network is desirable in situations where data is being collected over a wide area and needs to be transmitted to a location for uploading to the cloud. The most promising wireless technologies for this application are LoRa, NB-IoT, and LTE-M.

LoRa/LoRaWAN stands for long range wireless area network, and allows for communication for distances up to 6 miles while still consuming relatively little power. This wireless technology was developed by Semtech in 2012 and has become popular in recent years since the rise of the IoT device. LoRa uses sub-GHz range frequencies for transmission, meaning it has relatively low data rates but a long operating range and reduced interference. Since LoRa is not a cellular technology, it cannot communicate with mobile networks, making it less complex, cheaper, and simpler to set up. A LoRa gateway device will be required at the end of the network if the data is to be submitted to the cloud for data storage and processing.

NB-IoT is another popular technology in this space, and is a cellular technology. This makes it more complex to set up, more expensive, and also more power-consuming. However, it provides direct access to the internet and also higher quality cellular connections. Currently, NB-IoT is only supported in Europe and not the U.S., and is only capable of transmitting very small amounts of data.

LTE-M covers the applications that are not suitable for LoRa or NB-IoT, being specifically designed for long-distance cellular access that also require high data rates. It is significantly less power-consuming than the standard LTE used in smartphones and is also cheaper to implement due to the reduced bandwidth.

Which Should I Use?

The decision of which communication protocol is suitable for a given project is dependent on the context of the problem. Will you be using large numbers of devices that will be far apart and require cellular access? Or perhaps only a few devices that will each be placed in a single room in your house? Once you understand your specific use case, you can estimate what parameters you might need and then select the most suitable wireless technology based on those estimates.