Implementing M2M applications via GPRS, EDGE and UMTS

Klaus-Dieter Walter, M2M Alliance e.V., Aachen, Germany.
E-mail: kdw@ist1.de

1 What is M2M?

"M2M" is the abbreviated form of "Machine-to-Machine". The term is used to refer to machine-to-machine communication, i.e., automated data exchange between machines. ("Machine" may also refer to virtual machines such as software applications.) Viewed from the perspective of its functions and potential uses, M2M is causing an entire "internet of things", or internet of intelligent objects, to emerge.

Finding a neutral answer to the question "What is M2M?" that could apply to all manufacturers or suppliers is, however, not quite so simple. For GSM mobile network suppliers – in other words, traditional mobile phone network providers who put the infrastructure for mobile phone networks in place – the answer could involve, for example, automated data transmission via SMS/GPRS, or remote maintenance and teleservice applications. The activities of suppliers of GSM wireless modems (used in many M2M applications for data transmission via cellular networks) would lead them to interpret M2M in the same way. Meanwhile, manufacturers of Bluetooth or ZigBee radio chips generally take M2M to mean AMR (Automatic Meter Reading: in other words, wireless transmission of consumption data) or similar applications. As far as semiconductor manufacturers are concerned, there is immense market potential for these kinds of applications (hence the philosophy, "every radiator needs a radio chip") - and the list of examples of perspectives on M2M just goes on.

With such a wide range of potential applications, it is hardly surprising that, in 2003, analysts in the market research company Forrester deemed the M2M market to be "the biggest growth market of the next five to twenty years". To understand how they came to this conclusion, consider the following: The SMS and MMS messages sent by and to approximately six billion people around the world equate to extremely healthy sales figures for cell phone network providers; in 2003, approximately 20 billion SMS and 30 million MMS messages were sent. As well as this, worldwide there are around 50 billion machines which – at least in theory – are enabled for M2M applications.

M2M concepts are ideal for forming the basis of solutions for a large number of tasks, particularly in vertical markets. However, experience has shown that these kinds of markets develop less rapidly than horizontal markets for mass products. As a result, the prognoses of the market researchers in 2003 were doubtlessly too optimistic. To date, an M2M growth market has not established itself worldwide. However, there are a number of highly encouraging sectors within the market – performance data transmission, telematics, monitoring and RFID, to name but a few.

On closer inspection, however, M2M has merely become a new buzzword for demanding applications involving telemetry (automatic remote transmission of any measured data) and

SCADA (Supervisory, Control and Data Acquisition). In contrast to telemetry and SCADA-based projects, the majority of M2M applications are broadly based on established standards, particularly where communication protocols and transmission methods currently in use are concerned. Telemetry applications involve completely proprietary solutions that, in some cases, have even been developed with a specific customer or application in mind. M2M concepts, meanwhile, use open protocols such as TCP/IP, which are also found on Internet and local company networks. The data formats in each case are similar in appearance.

1.1 The basic M2M concept

First and foremost, an M2M application is characterized by three fundamental features: 1. the data end point (DEP); 2. the communication network; and 3. the data integration point (DIP).

The DEP and DIP can be used to integrate any subsystem, for example, or a complete process (X) into an IT application (Y). Figure 1.1 shows the basic interrelationship between the three elements. This kind of solution is also known as "End-to-End M2M". The process (X) and application (Y) form the actual function end points.

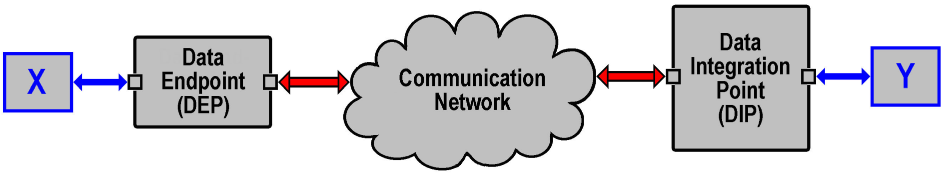

Figure 1.1: Basic concept of an M2M application

Generally speaking, a data end point (DEP) refers to a compact microcomputer system, one end of which is connected to a process, or to a higher-level subsystem via special interfaces. The other end is connected to a communication network. In most MSM applications, there are several DEPs. On the other hand, a typical M2M application only has one data integration point - saying this, however, it is possible to envisage M2M applications with several DIPs. There are no hard and fast rules governing implementation of a DIP: A function unit of this kind can, for example, be formed by an Internet server or a special software application on the host computer of a traffic control system.

The flow of information in an M2M application is also not necessarily server-oriented. Instead, direct communication paths between the DIP and DIP are supported, as well as direct and indirect connections between the individual DEPs, also known as P2P (Peer-to-Peer) connections.

2 The M2M communication network

As mentioned above, the communication network in an M2M application is the central connection component between a DEP (data end point) and DIP (data integration point). In terms of physical components, this kind of network can be established using a LAN, Wi-Fi, telephone network/ISDN, GSM mobile network, or similar.

2.1 GSM, GPRS and EDGE

Wireless M2M connections can be established over larger distances (in some cases, the entire world) by means of GSM mobile connections (GSM = Global System for Mobile Communications). In this case, the physical network takes the form of a cell phone network from a provider such as T-Mobile or Vodafone. The provider offers access to a cell phone network as part of a contractual agreement, in which use of the network incurs fees – that can sometimes be considerable – to be paid to the provider. In return, it is the job of the provider to ensure that cell phone connections operate without any problems.

GSM mobile networks, based on digital transmission technology, were introduced to Europe in the 1990s, when they were known as D and E networks. GSM technology was developed with the aim of establishing a mobile telephone system that would both grant participants the freedom to access their network throughout Europe and enable compatible voice and data services. The GSM network infrastructure consists of four components:

1. Cell phone (technical term: Mobile Station = MS)**.** The MS is comprised of an antenna, a sending and receiving unit, a power supply, a speaker and a microphone. In addition, the MS usually includes a display on which the telephone number of the caller and short messages can be viewed. A SIM card is located inside the MS. This is supplied by the network provider as part of the contract, and is intended to prevent unauthorized usage of the MS.

2. Cellular network transmission system (technical term: Base Station Subsystem = BSS). This consists of a central controller (BSC: Base Station Controller), which monitors the network connections and, if necessary, initiates handovers. Different base stations are connected to a BSC (most control tens or even hundreds of stations). Each base station operates one or more cells (technical term: Base Transceiver Station = BTS), to which the antennae are connected.

3. Network Subsystem (NSS) or **Core Network Subsystem** (CSS). The NSS consists of the MSC (Mobile Switching Center), constituting the telephone exchange itself and the interface between the wireless network and telephone network. Additionally, the NSS includes a Visitor Location Register (VLR) that stores information about all of the mobile participants under the jurisdiction of the wireless network. The HLR (Home Location Register), meanwhile, stores information on all participants who are customers of the network provider. The AUC (Authentication Center) is responsible for authentication, while the optional EIR (Equipment Identity Register) stores information on the serial numbers of the Mobile Stations in use.

4. Operation and Maintenance Center (OMC) The OMC is used for controlling and monitoring the cellular network as a whole.

In a GSM network, fundamentally this involves a circuit-switched network for digital voice communication. Voice signals are digitized, and the resulting data is embedded in time slots and then transmitted. On the receiving end, the data is converted into analog signals that can then be output via the speaker of an MS. During the early stages of development, CSD (Circuit Switched Data) was the only form of data transmission available; using this method, data transmission can take place at 9,600 bps within the time slot. On a GSM network, bills are based on the individual time slots. Accordingly, the volume of CSD-based data transmitted or the density of the conversation is of no relevance, since bills are based solely on the duration of the connection.

A few years ago, the GSM infrastructure was expanded to include GPRS (General Packet Radio Service), which enables packet-based data transmission on GSM networks. GPRS was developed with the intention of enabling mobile access to the Internet. 2.5G ("second and a half generation") networks are also frequently alluded to within the context of GPRS: This name was created in imitation of the fact that UMTS was already known as a "3G" network technology. A new billing method was developed for GPRS, which uses the actual volume of data transmitted as a basis. The connection duration has no effect on this. In other words, if only 1,000 characters are transmitted via GPRS within an 8-hour period, the participant need only pay for these 1,000 characters; the connection duration does not enter the equation at all. (Similarly, surfing the web via GPRS is charged on the basis of the data volume rather than the amount of time spent on the web browser.) However, some providers tend to round the data volume up to the nearest billing unit: this varies from provider to provider and tariff to tariff. Vigilance is vital where GPRS tariffs for an M2M application are concerned; otherwise, the rounding up of units by the provider could result in unexpected - and considerable - costs being incurred.

Multislot class	**Downlink slots**	**Uplink slots**	**Active slots**
1	1	1	2
2	2	1	3
3	2	2	3
4	3	1	4
5	2	2	4
6	3	2	4
7	3	3	4
8	4	1	5
9	3	2	5
10	4	2	5
11	4	3	5
12	4	4	5

Table 2.1: GPRS downlink and uplink slots

Features of GPRS data transmission are categorized into classes (known as multislot classes or sometimes GPRS classes). The individual classes are subdivided to indicate the number of time slots that are bundled for data transmission purposes. (In this context, the time slots are the same as those used for GSM voice or CSD data transmission.) A distinction is also drawn here between "downlink" and "uplink" slots: Table 2.1 provides an overview of these. The most common multislot classes are 1, 8, 10 and 12. Today, most GSM/GPRS modem modules use class 10. The GPRS coding scheme used in each case must be used to calculate the actual data transmission speed. There are four coding schemes: CS1, CS2, CS3 and CS4. The network determines the coding scheme used to be used for a connection dynamically on a case-by-case basis. The network utilization has an important role to play here. Table 2.2 provides an overview of the possible transmission speeds. (The values in the column for eight time slots are only theoretical, as currently eight time slots cannot be used for a GPRS connection in any GSM network.) In other words, the potential transmission speed for a GPRS connection is calculated on the basis of the multislot class and the coding scheme used in each case. One of the factors governing which scheme is used in an active connection is the current network utilization. With four time slots, the actual maximum speed is 85.6 kbps, but frequently only 53.6 kbps is reached. In practice, using a class 10 modem you can only really rely upon achieving a transmission speed of between 36.2 kbps and 85.6 kbps in the downlink.

Time slots →	**1**	**4**	**8 (theoretical)**
CS1	9.05 kbps	36.2 kbps	72.40 kbps
CS2	13.4 kbps	53.6 kbps	107.2 kbps
CS3	15.6 kbps	62.4 kbps	124.8 kbps
CS4	21.4 kbps	85.6 kbps	171.2 kbps

Table 2.2: GPRS coding schemes

Presently, EDGE technology is being used to further advance the GPRS service. EDGE stands for Enhanced Data Rates for GSM Evolution: An enhancement of GPRS, it is used to achieve higher rates of data transmission. Currently, standard rates range between 150 and 200 kbps. EDGE manages to increase this by using a more efficient modulation scheme (8PSK instead of GMSK as with GSM/GPRS). Under ideal circumstances (GPRS coding scheme CS4), GPRS can transmit 21.4 kbps within a single time slot. With the altered modulation scheme present in EDGE technology, it is possible to achieve a speed of 48 kbps per time slot. Therefore, with eight time slots EDGE can (in theory) reach 384 kbps. Up until now, EDGE has been introduced in 75 countries – and as one of the very first European cellular network providers, TIM (Telecom Italia Mobile) has done so under the name TIM Turbo. The service has also been implemented in Austria, Switzerland, the Netherlands, France, the Czech Republic, Slovakia, Poland, Latvia, Lithuania, Croatia, Turkey and, since March 2006, Germany (albeit exclusively with T-Mobile).

It is the costing model of GPRS and EDGE that makes them most attractive for M2M applications. A large number of M2M applications transmit relatively low data volumes, so the potential transmission speed has little significance for them. However, it is important that a permanent connection between the data end points (DEP) and the data integration point (DIP) can be provided without incurring additional costs.

2.2 UMTS

UMTS (Universal Mobile Telecommunications System) is a relatively recent third-generation (3G) cellular network standard, which enables data transmission at a considerably higher rate than is possible with the GSM standard. It was developed by the ETSI in 1998 specifically for universal cellular network telecommunication. The intention is for this standard to replace previous mobile communication via GSM, as it is applied in D and E networks, with a wider range of services. Particularly in the multimedia sector, thanks to its high transmission rates UMTS should be able to support high-capacity multimedia services. In addition to voice and audio services, this includes fast transmission of not only data, graphics and text, but also moving images and videos. As yet, UMTS is still not suitable for M2M applications as it does not meet the relevant hardware requirements: UMTS end devices are currently only available in the form of cell phones and extremely cost-intensive PC expansion cards. A small number of modern M2M applications do use UMTS for image-based data transmission; in such cases the DEP generally takes the form of an industrial PC.

3 Application examples

From the wide range of M2M applications – both those that have already been implemented and those which have the potential to be – two examples where the GSM mobile network is particularly useful as a method of communication will be described here: tracking and infrastructure monitoring.

3.1 Tracking

An M2M application with rapidly increasing market volumes is known as a tracking application (following and monitoring mobile objects). In most cases, mobile objects are taken to mean vehicles of any kind – whether on land, in the air or on water. However, it is possible to monitor people or mobile objects which do not move under their own power (such as containers) in the same manner. The Global Positioning System (GPS) forms the technological basis for these monitoring activities.

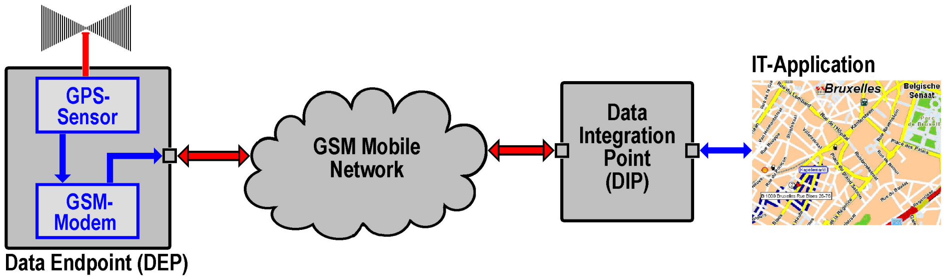

Figure 3.1: Basic concept of an M2M-based tracking application

GPS is a satellite-supported navigation system developed by the United States Department of Defense for determining the position of objects or people anywhere in the world. Put into operation in July 1995, GPS was originally designed for position detection and navigation in a military context (weapons systems, warships, airplanes, etc.). However, today it is also used in civilian applications such as sea and air travel, car navigation systems, navigation in the outdoors, surveying, and logistics.

GPS is based on satellites that continually send out signals: GPS receivers can determine their position from the time delay between transmission and reception of these signals. In theory, three satellites emitting signals should be sufficient to determine the precise position and height. In practice, however, GPS receivers do not have a clock that is precise enough to be able to determine delay times correctly, and for this reason require signals from a fourth satellite. It is not only the receiver's position that can be determined on the basis of GPS signals - their speed can be calculated too. Generally speaking, this is carried out by measuring the Doppler effect, or the numerical differentiation of a location according to time. It is also possible to detect the direction in which the receiver is moving; this can then serve as a kind of compass or as a method of navigation in electronic maps. Since a GPS receiver is always in contact with at least four satellites, the system uses a total of 24 satellites, each of which makes two complete orbits of the earth per sidereal day at an altitude of 20,183 km. There are 6 orbital planes across which the satellites are evenly distributed, so that four are assigned to each. The planes have 55° inclination with respect to the equatorial plane and are each rotated away from one another by 60°. Therefore, a satellite is positioned above the same point on the earth every 23 hours, 55 minutes and 56.6 seconds.

The basic concept of a tracking application is displayed in Figure 3.1. In a positioning application of this kind, the data end point (DEP) is essentially comprised of a GPS sensor and a GSM modem module. The GPS sensor transmits the GPS data to the GSM modem in NMEA 0183 format. The modem then sends the position data to the data integration point (DIP) via a cellular network. From this point, the positioning data can be relayed to higher-level IT applications. In many cases, data relating to the current position of a mobile object is incorporated into electronic map information. This then enables the position of an object to be displayed on a map. In Figure 3.1, the GSM mobile network is used for transmitting GPS data. If the mobile entity in question is a vehicle or object that is located outside of a GSM network (such as an airplane, seagoing vessel or container), a satellite connection with a global range can still be used to transmit position data. There are various providers of these kinds of global transmission services.

3.2 Infrastructure monitoring

Today, countless applications use complex networks of devices, in which most systems perform their services 24 hours a day, 7 days a week, without any human monitoring. Should an individual subsystem fail, the entire application will be damaged, at least in part. However, detecting the precise failure as promptly as possible is a problem. A typical example in the field of IT is an individual switch within a large Ethernet LAN. If this switch fails to perform its duty, certain computers within the LAN will no longer be accessible, or will be unable to communicate via the LAN. If, with one of these computers, the company's e-mail server is involved, considerable problems may arise. Therefore, ideally each switch and the accessibility of the e-mail server (as well as each of the company's other servers) should be permanently monitored. Every fault must be immediately reported to the appropriate administrator, who should generally be able to rectify it within as short a time as possible, as quick response times are possible within a company's premises. Presumably, the longest duration of time that would elapse in such cases would be the time leading up to a user noticing a malfunction and informing the administrator.

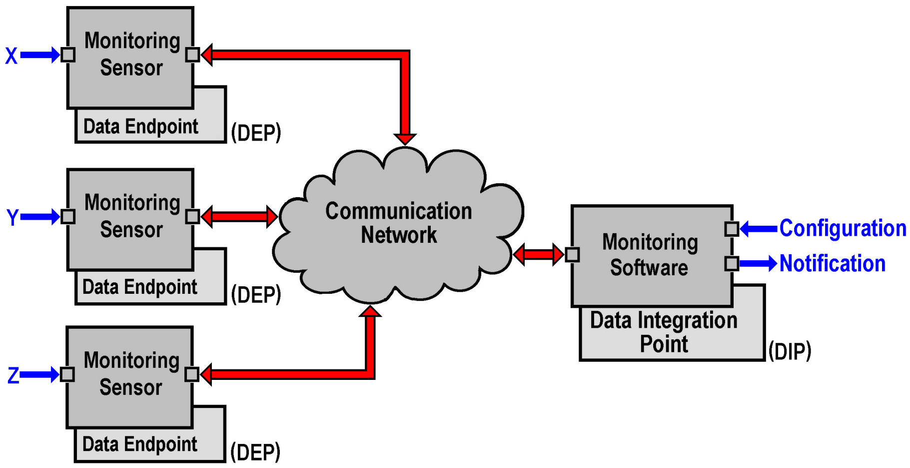

Figure 3.2: Concept for M2M-based infrastructure monitoring

This kind of situation is considerably more difficult where widely distributed applications are concerned. Let us take the example of the infrastructure for an energy supplier, which consists of a number of distributed pumping, transformer and substations used to operate the water and electricity supply for a region unattended. Another example is the toll billing system for trucks on German freeways. In this case, over 300 fixed control bridges installed on freeways monitor passing traffic to ensure that tolls are paid. If this kind of distributed infrastructure component were to fail, quick action would be required. It would be of critical importance to detect the failure as quickly as possible and inform the appropriate personnel immediately. In this case, longer downtimes would be expected due to the time taken for service personnel to travel to the fault site. The operator of a distributed application must often wait until the failure comes to the attention of its subscribers and the control station receives a message. In the event of a power failure, this usually takes very little time and users do not notice that a toll bridge is down.

Type of notification	**Cause**
Error	Complete failure of an infrastructure component
Warning	Short-term interruption of normal operation
OK	Infrastructure component is operating without errors
Escalation	Extremely long downtime for a particular infrastructure component
Change	Change detected in the behavior of an infrastructure component

Table 3.1: Examples of different types of notifications in infrastructure monitoring applications

Whether within a company building or distributed across an entire country, using an M2M-based infrastructure monitoring solution (i.e., "infrastructure monitoring") enables failures in individual function units to be detected considerably more quickly. As is well known, the quicker a failure and its cause are detected, the shorter the total downtime. In many cases, fault signals on individual modules are even able to detect the imminent failure of individual system components (by means of an appropriate signal lamp, for example). These kinds of visual warning signals do, however, frequently go unnoticed.

Figure 3.2 illustrates the use of an M2M-based solution for infrastructure monitoring. The data end point (DEP) in each case permanently checks the availability of any given infrastructure component by means of special monitoring sensors. Any potential failures can then be detected immediately by the DEP concerned. The individual DEPs are connected to a monitoring software application by means of the communication network: This application is used on the data integration point (DIP). It receives failure, error and fault messages from the individual DEPs with respect to the infrastructure to be monitored (X, Y and Z in Figure 3.2). In addition to the connection to the DEPs, the DIP’s monitoring software has two other interfaces, for configuration and notification. Among the facets of the configuration interface is the ability to determine who is responsible for the system and when. This generally makes it possible to envisage a number of different notification scenarios (table 3.1). This interface also enables the monitoring behavior to be defined precisely. The notification interface sends the message in each case (SMS, e-mail, automatic fax sending, or calling a telephone number and playing an audio file).

4 Summary and outlook

M2M concepts based on GSM, GPRS, EDGE and UMTS afford considerable potential for both growth and saving on costs. Ultimately, most objects are still not capable of communicating with one another. At the present time, however, the lack of standardization in interfaces and data formats between data end points and the data integration point continues to pose a problem.