Chapter 1

Embedded network systems

1.1 Introduction

Continuing advances in integrated circuit technology have enabled the integration of computation and communication capabilities into devices which monitor or control physical processes. Digital controllers and sensors are found in automobiles, home appliances, factories, aircraft, cellular telephones, video games, and environmental monitoring systems. Indeed, the vast majority of processors now being manufactured are used in embedded applications (i.e., having connection to physical processes) rather than in what would ordinarily be thought of as a computer. Many are networked within the confines of a local control system, typically in master/slave configurations. However, advances in wireless technology and in the understanding of distributed systems are now making possible far more elaborate compositions of embedded systems that may function as the connection of the Internet to the physical world. Embedded network systems (ENS) are poised to become pervasive in the environment with the potential for far-reaching societal changes that have hitherto been the subject of science fiction. It is the purpose of this book to lay out the foundations of this technology, the emerging design principles and applications, and some of the interesting societal questions raised by ENS. This chapter provides some examples of ENS, discusses relevant technological trends, explores some implications of scaling to very large numbers of ENS nodes, and places the technology in a historical context.

Figure 1.1 is the block diagram of an ENS node. Essential hardware components are some type of energy supply (e.g., battery or external connector), a sensor and/or actuator (e.g., microphone or speaker, camera or display), a processing unit (signal processing and storage capability), a communications device (e.g., radio or Ethernet port), and packaging. Some of these may be integrated together, while others may be discrete components. The components are connected by data buses (dashed) and power lines, with various devices to provide interfaces. Also essential is the software that enables the management of the platform resources, signal processing, and external communications. The basic characteristic of such devices is that it is their aggregation into a network that provides the required functionality.

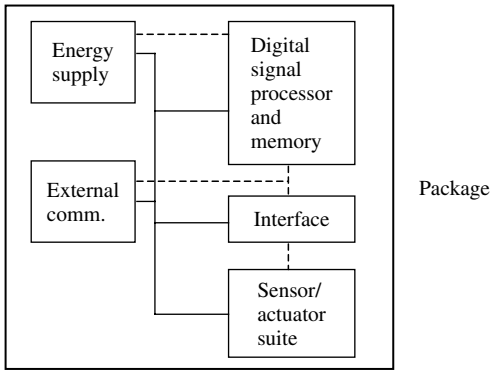

Figure 1.1 ENS node block diagram.

ENS networks provide distributed network and Internet access to sensors, controls, and processors embedded in equipment, facilities, and the environment. Integrated circuit technology now enables the construction of sensors, radios, and processors at low cost and with low power consumption, enabling mass production of sophisticated but compact systems that link the physical world to networks. Scales range from local to global, with applications including medicine, security, factory automation, environmental monitoring, and condition-based maintenance. Compact size and low cost allow ENS to be embedded and distributed at a small fraction of the cost of conventional wired sensor and actuator systems. This can enable there to be hundreds or thousands of sensors per user, resulting in many new system design challenges.

Centralized methods for sensor networking present heavy demands on cable installation and network bandwidth. However, by processing at source and conveying decisions rather than raw data, the burden on communication system components, networks, and human resources is drastically reduced. This same observation holds true whenever there are relatively thin communications pipes between a source and the end network, or when dealing with large numbers of devices. The physical world can generate an unlimited quantity of data to be observed, monitored, and controlled, but there are finite resources that can be put into wireless telecommunications infrastructure. Moreover, the end user needs to be presented with some redacted form of this stream or suffers from information overload as the number of sensors increases. Thus, from the perspectives of both network resources and the finite capacity of the end user, scaling to large numbers of devices implies that sensor nodes will become increasingly autonomous, doing a large fraction of the data processing and decision-making *in situ*.

In this overview, two scenarios are considered illustrating different aspects of the design tradeoffs. In the first, an autonomous network of sensors is used to monitor events in the physical world for the benefit of a remote user connected via the web. In the second, a space is instrumented to enable interactions with people. A general architecture is depicted in Figure 1.2, for now neglecting the details of how services are actually supported by Internet-connected devices. To supply some service, two different clusters of sensor nodes are connected through their respective gateways to the Internet. The nodes are assumed to be addressable either through an Internet

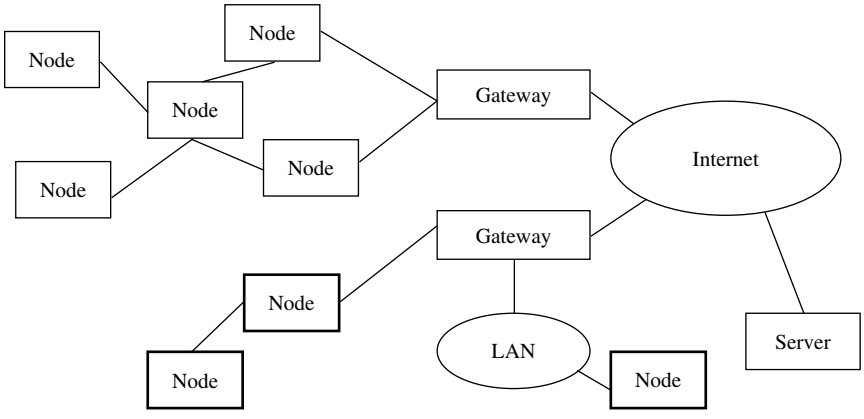

Figure 1.2 Generic ENS network architecture.

protocol (IP) address or some attribute (location, type, etc.), and are distinguished from pure networking elements in that they contain some combination of sensors and/or actuators. That is, they interact with the physical world. The gateway may itself be a sensor node similar to other nodes in its cluster, or it may be entirely different, performing, e.g., extra signal processing and communications tasks and having no sensors. In the cluster in the top left of Figure 1.2, nodes are connected by a multihop network, with redundant pathways to the gateway. In the bottom cluster, nodes may be connected to the gateway through multihop wireless networks or through other means such as a wired local area network (LAN). The nodes in the different clusters may all be of one type, or they may be different within or across clusters. In a remote monitoring situation, there may be part of the target region with no infrastructure, and thus the multihop network must self-organize, while in other parts of the region there may already be assets in place that are accessible through a pre-existing LAN. There is no requirement that these assets be either small or wired; the point is to make use of all available devices for providing the desired service.

The remainder of this chapter describes some design heuristics, discusses a number of different research efforts to deploy large networks in areas without infrastructure support, radio frequency identification (RFID) tag systems, some possibilities for enacted public spaces, and presents technological trends within a historical context.

1.2 ENS design heuristics

A description of some of the fundamental physical constraints for sensing, detection, communication, and signal processing cost is provided in later chapters. Implicit in any discussion of constraints is that an optimization problem emerges in which the quality of some basket of services is traded against the resource costs. The basic design constraints that emerge for ENS are:

(1) There are many situations in which reliable detection demands sensors in close proximity to a physical event, causing numbers to scale (e.g., physical obstructions

to cameras). With large numbers of sensors, the type of information obtained is qualitatively different than that obtained with remote arrays.
(2) Sensors, radios, and signal processing will all ride the integrated circuit technology curve down in cost, but batteries and other energy sources improve in cost only slowly with time.
(3) Communications energy cost per bit is in many instances many orders of magnitude larger than the energy required for making decisions at source, and communication is limited in its efficiency by fundamental limits, whereas the processing cost is to first order limited only by current technology.
(4) Human labor does not scale; networks must be self-organizing to be economical.
(5) Scaling with physical responsiveness demands hierarchy, with distributed operation at lower layers and increased centralized control at higher layers.

Note that hierarchy does not necessarily imply heterogeneous devices. If one considers human organizations, the native processing abilities are roughly equal at all levels. What differs most in progressing up the chain is that different information is processed, i.e., information at different levels of abstraction/aggregation. Moreover, commands progressing down the chain also differ in their level of abstraction, from policies down to work directives, with varying scope for interpretation. This flexibility enables lower levels to deal with local changes in the situation much faster than if a central controller needed to be consulted for each action, while enabling global goals to be pursued. With machines, of course, there is the possibility of providing the devices with highly differentiated abilities at different levels of the hierarchy. These can bring important advantages, e.g., a backbone long-range high-speed communication pipe can greatly reduce latency compared to relying only on multi-hop links. Thus, while the logical rather than physical hierarchy is arguably much more important in enabling scalability, it behooves the designer of large-scale systems to consider both. Homogeneity is in any case impractical in long-lived systems composed of integrated circuit components; as for the Internet, the architecture must accommodate the addition of successive generations of more powerful components.

1.3 Remote monitoring

To make the discussion more concrete, consider an application requiring identification of particular classes of signal sources passing through a remote region. These sources may be vehicles, species of animals, pollutants, seismic events on Mars, or, on a smaller scale, enzyme levels in the bloodstream or algal blooms in the ocean. In any case, it is assumed that there is no local power grid or wired communications infrastructure, but long-range communication means exist for getting information to and from a remote user. Then, in laying out a network such as the one depicted in Figure 1.2, both energy and communications bandwidth may be critical constraints. As noted above, when the network must scale in the number of elements, this effectively means that much of the signal processing must be performed locally. For example, in studying the behavior of animals in the wild, a dense network of acoustic sensors may be employed. The nodes contain templates for the identification of the species emitting calls. Nodes that make a tentative identification can then alert their immediate neighbors so that the location of

the animal can be roughly determined by triangulation. Infrared and seismic sensors may also be used in these initial identification and location processes. Then finally other nodes may be activated to take a picture of the source location so that a positive identification can be made. This hierarchy of signal processing and communications can be orders of magnitude more efficient in terms of energy and bandwidth than sending images of the entire region to the gateway. Further, the interaction of diverse types of nodes can more simply lead to automation of most of the monitoring work, with humans only brought into the loop for the difficult final visual pattern recognition on preselected images. On positive identification, the audio and infrared files corresponding to the image can be added to a database, which may subsequently be mined to produce better identification templates. Note that the long-range communication link (via the gateway) potentially enables the full use of web-accessible utilities, so that the end-user need not be present in the remote location, and databases, computing resources, and the like may all be brought to bear on interpreting the (processed) data.

There is tension between experimental apparatus for initially exploring an application domain and what will actually be needed for large-scale deployment. Because networked sensors have hitherto been very expensive, relatively few array data sets are available for most identification purposes, and sensors have typically been placed much further from potential targets than is possible with ENS. This means paradoxically that initially fairly powerful nodes need to be constructed to conduct large-scale experiments to collect raw data, so that suitable identification algorithms can be developed using the resulting data set. Likewise, in experimenting with different networking algorithms, it is desirable from the point of view of software development initially to provide a platform with considerable flexibility.

Example 1.1 Evolution of a habitat monitoring system

An overview of one set of sensor deployments in the James Reserve in the San Jacinto mountains near Palm Springs CA is provided in Figure 1.3. There are very large elevation changes and consequently a broad set of species, some endangered, represented in a small area. Early sensor deployments included cameras in bird nest boxes, and a moss camera that records its growth and change in appearance with moisture. A first phase of wireless sensor network deployment focused on measuring microclimates, with the results correlated with plant growth.

However, it became apparent very quickly that logistical tasks such as battery replacement that are easy with ten nodes become extremely tedious as numbers scale, given that sustained deployments are required. Further, studies of plant growth and animal behavior could benefit immensely from having a broad set of sensors that span from roots to tree tops. Many of the interesting locations are heavily shaded, rendering solar power an unattractive power source. Having some automated infrastructure would be very valuable in such an application, both for moving sensors and supplying them with energy and communications. A system of robotic nodes that move along wires at treetop level was devised (see also Chapters 12 and 13). The nodes can lower a secondary platform to the forest floor to allow sensing in 3-space, and the repositioning or replacement of nodes, as depicted in the schematic in Figure 1.4. Since these nodes have access to reliable energy supplies, higher end processors can be used. The energy and communications resources also allow high-quality imagers to be used, with elevation above the underbrush providing a large viewing volume.

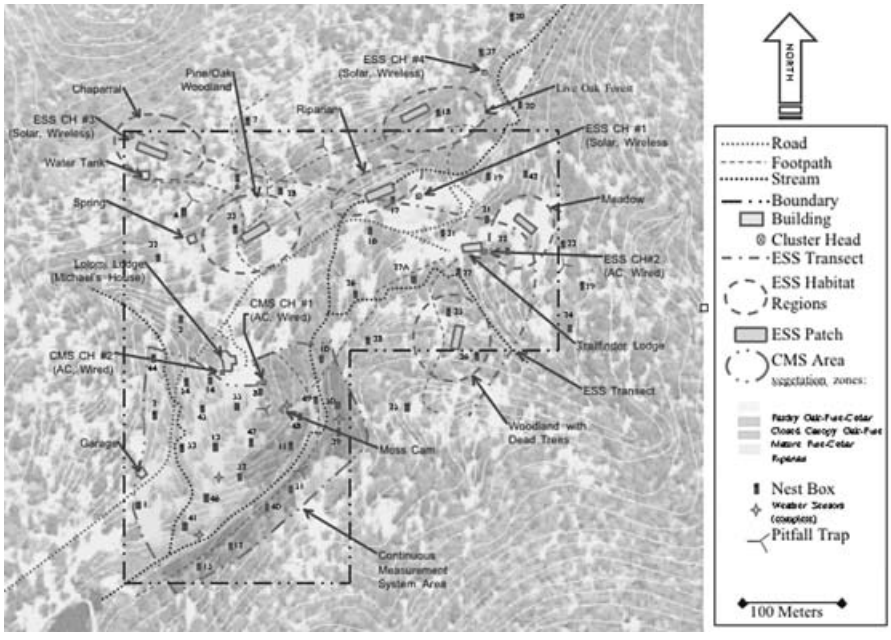

Figure 1.3 Early sensor deployments (Courtesy of Dr. Michael Hamilton).

Figure 1.4 Interaction of robotic elements with fixed sensor nodes.

Several such transects are planned to enable sustained study of factors thought to be important to the health of the ecosystem.

1.4 RFID

RFID provides a unique identifier to the item that is tagged. The devices may be in one of several categories:

- Passive tag, ID only: the tag has no power source of its own; it provides a unique resonant response to the interrogation signal.
- Passive RF response, energy for on-board memory: the radio responds only when interrogated, but new information can be written to the tag.
- Active tag: the tag has the ability to send RF signals.
- Active tag with sensors: the tag is a sensor node with a unique identity.

Active tags have been used for a long time for such purposes as tracking wild animals to infer behaviors. Passive tags are now used extensively in animal husbandry, and will increasingly be applied to the industrial supply chain for tracking of items at various stages of production and distribution. Very small tags will be applied to individual consumer items for such purposes as inventory control and automated check-out. Their small size demands that readers be in close proximity to retrieve data. Active tags may be attached to shipping pallets or containers to enable longer-range reading, and may include sensors to determine the conditions under which the items were shipped (temperature, vibration, tampering). In the limit, these may form, for example, a multihop sensor network among shipping containers.

The complete system consists of a set of tags, one or more interrogators, and a backend data management system. Even with purely passive tags this system is a sensor network, with the interrogators exciting responses and receiving as information the identity of the tags in range. Unlike a system of bar codes and readers, RFID systems provide the capability of identifying not only the category of an item but also its unique identity. Verification of individual identity may enable financial transactions based upon that identity. Further, a decreased sensitivity to orientation and the ability to read tags at a distance and through obstructions enables automation of many functions.

Since the tag readers have a limited range, the act of reading also provides position information on the tagged item. A sequence of readings can therefore yield a history of the motion of a tagged item (or person). Thus while the tag might be designed principally to provide identification, there are secondary inferences that can be made from the data retrieved by the interrogator network. The locations that a tagged individual visits can, for example, be used to infer shopping behavior (e.g., what displays within a store most captured attention). Moreover, the combination of determining identity and position enables binding of information to a tagged object if that object is observed by multiple sensors. Thus, while it is a difficult signal processing task to determine whether some particular individual or object is within the field of view of a camera, once tagged the uncertainty can be eliminated (in effect, identity is broadcast) and the image added to some database concerning that individual or object.

Clearly there are large privacy concerns associated with RFID tags and consumer products. If these become embedded in shoes and clothing, and interrogators are widely placed throughout urban environments, there could be a huge intrusion into the privacy of the individual. To deal with such concerns, a kill switch has been proposed for such tags, so that at the point of sale the tags are deactivated. Laws have also been proposed to govern the uses that can be made of information collected by RFID readers. However, the debate on privacy versus economic efficiency is not yet settled.

1.5 Enacted spaces

We have all seen visions of the future (or mythological past) in which the physical environment responds to gestures or verbal commands: doors open, music plays, lights go low, computers spew obscure facts, weapons systems fire, or powerful spells are cast. An enacted space is one in which embedded systems facilitate interactions with people and machines. This may take the form of a system of electronic tags and readers which interact with some database so that the desired information can be extracted (e.g., to seek out compatible people in a club based on profiles, or to automate retail shopping) or a system to assist navigation through building complexes to desired items (e.g., museums or warehouses), or a system that supports interactive entertainment (e.g., performance art or group techno-games).

Consider a robotic gaming example, in which the space physically reconfigures in response to the activities of robotic teams controlled by human players. Instead of seeing a purely virtual environment, players would "see" and "hear" through cameras and microphones on the robots within the game, with some robots controlled directly through joy sticks and others run by remote agents. Players might bring up different windows on their control panels to get views from different robots as they attempt to coordinate diverse elements. Some game elements might be virtual (e.g., explosions) but would nonetheless produce physical response (e.g., loss of particular functions, crumbling of buildings each of whose structural elements have some active components). Robots could manipulate physical elements to construct barriers, structures, or miniature cities in which the game would be carried out. Flying elements could be approximated through a network of wires allowing nearly complete mobility in 3-space.

A larger-scale system could have human players in an instrumented space that reacts to player actions. The costumes and props carried by the players may include sensors, readers, and signal sources that interact with the displays and activated surfaces in the game zone, as well as with the devices on other players or robotic elements. Many of the necessary elements have already been developed in other contexts (e.g., laser tags, virtual reality games, museum displays that change based on sound and images). Others, such as weaving arrays of electronic elements into fabrics, are still at the stage of interesting research ideas. Issues of cost and how to author a system with so many sophisticated components in an intuitive fashion also remain challenging. However, there is interest in creating artificial realms of this type for training for hazardous jobs as well as for entertainment.

1.6 Historical context

The close intertwining of processing and networking is a central feature of systems that connect the physical and virtual worlds. Research is now proceeding both in the design of small, specialized nodes that could potentially be deployed in very large numbers, and in the creation of dense networks of larger nodes that can be used to learn more about the types of networking, sensing, and signal processing that will be needed in future systems. It is apparent even from the two application examples that a broad range of skills is required to design systems that network the physical world. Later chapters of this book will therefore provide introductory treatments of topics ranging from propagation of signals all the way up to applications to illuminate the large and exciting design space available with ENS. To set the stage, this chapter concludes with a brief historical overview of the technology.

The western technological revolution has its roots in Classical Greek rationalism divorced from its social pessimism, fused instead with a belief in progress in human affairs. That is, Nature is governed by laws, these laws are best discovered through experimentation and the use of Reason, and understanding of these laws leads to progress towards a great destiny. When combined with the competitive short-term ambitions of city-states and later nation-states this potent mix has helped to launch a scientific age that began in the Italian Renaissance and has since become the most powerful force for social and economic change in the world. The direction of science and technology is shaped by the philosophical, economic, and political imperatives of the day, and in turn profoundly affects society. Thus in embarking on the creation of information technologies with a large potential for societal impact it is well to consider desired and undesired outcomes sooner rather than later.

The electronic digital age had its beginnings with telegraphy in the 1840s, with human beings acting as actuators, sensors, and processors. The story of communications has been one of gradual replacement of human functions with computational elements and switches, accompanied, as usual with machine technology, by higher speeds and lower costs. A prime driving force is described by Moore's Law: the processing capability of computing devices doubles roughly every 18 months. This is a consequence of economic forces rather than silicon technology, as it has held in some form from mechanical relay computers, through vacuum tubes, and finally to integrated circuits in silicon. There is an apparently unquenchable market for computation and an expectation that it will become less expensive with time, but investment must be recovered in one generation of devices before the new generation can be launched. Resources are marshaled to achieve the expectation, and thus Moore's Law continues. This evolutionary series of generations produces revolutionary results: with 66 generations in a century, processing power increases by a factor of 2^{66} (nearly 10^{20})! A startling consequence is that more artificial computational elements will be produced in the current year than in all previous human history. Projecting forward, there appear to be no insurmountable technological barriers to continue this growth for the next human generation, with many alternative technologies being explored should the limits of silicon be reached. Thus, it will be possible to support computations and systems of vastly increased complexity over time.

Integrated circuit fabrication technology may be regarded as a means to very reliably make small things in large numbers at low cost. This enables the construction of arrays of sensors and actuators in a technology known as MEMS: microelectromechanical systems. MEMS is completely compatible with computation and communication in the same small module. ENS on a chip is thus a present rather than a distant possibility, with the question being not whether it can be done but what kinds of ENS chips make sense. The feature sizes in integrated circuits are approaching those of cells so that it is not far-fetched to suppose that the next great technological revolution will be the merging of biological systems and information technology; many universities have established bioengineering departments with exactly this expectation. If this nanotechnology can then ride the Moore's Law curve, the consequences will be far-reaching in all aspects of human endeavor.

Yet there exist fundamental limits in activities that impinge upon the physical world. Energy technology is mature, with storage densities improving only slowly over time. The energy required to affect physical objects does not change at all over time, and there are other limits in the realms of communications and the ability to observe the environment accurately. Of course, even technology that is evolving at a rapid rate will also be limited at any given point in time. Thus engineers and computer scientists must make design tradeoffs. The remainder of this book addresses the underlying physical factors that affect the design of ENS, performance criteria and some algorithms which attempt to meet them, applications, and social implications of the technology.

1.7 Further reading

The following article highlights a few sensor network applications and describes various design tradeoffs:

G. J. Pottie and W. J. Kaiser, "Wireless integrated network sensors," *Comm. ACM*, **43**(5): 51–58, 2000.

It is one of several articles on this topic in a special issue.

A comprehensive study on ENS commissioned by the National Research Council is

D. Estrin, *Embedded Everywhere: A Research Agenda for Networked Systems of Embedded Computers*. National Academy Press, 2001.

It contains a survey of technology, potential applications, and future directions. One of the recommendations is a study of the ethical, legal, and social implications of ENS.

Some of the sensor deployments at the James San Jacinto Mountains Reserve can be monitored on-line at:

www.jamesreserve.edu.

The interplay of ideas in Western philosophy, religion, the cultural and practical arts, and the clash of personalities and peoples are treated at length in the series of books:

W. Durant and A. Durant, *The Story of Civilization*, Vols. I–XI. Simon and Shuster, 1954.