Introduction to Industrial Control Networks

Brendan Galloway and Gerhard P. Hancke, *Senior Member, IEEE*

***Abstract*—An industrial control network is a system of interconnected equipment used to monitor and control physical equipment in industrial environments. These networks differ quite significantly from traditional enterprise networks due to the specific requirements of their operation. Despite the functional differences between industrial and enterprise networks, a growing integration between the two has been observed. The technology in use in industrial networks is also beginning to display a greater reliance on Ethernet and web standards, especially at higher levels of the network architecture. This has resulted in a situation where engineers involved in the design and maintenance of control networks must be familiar with both traditional enterprise concerns, such as network security, as well as traditional industrial concerns such as determinism and response time. This paper highlights some of the differences between enterprise and industrial networks, presents a brief history of industrial networking, gives a high level explanation of some operations specific to industrial networks, provides an overview of the popular protocols in use and describes current research topics. The purpose of this paper is to serve as an introduction to industrial control networks, aimed specifically at those who have had minimal exposure to the field, but have some familiarity with conventional computer networks.**

***Index Terms*—industrial, control, networks, fieldbus.**

I. Introduction

IN the past decades the increasing power and cost-effectiveness of electronic systems has influenced all areas of human endeavour. This is also true of industrial control systems. Initially, control of manufacturing and process plants was done mechanically - either manually or through the use of hydraulic controllers. As discrete electronics became popular, the mechanical control systems were replaced by electronic control loops employing transducers, relays and hard-wired control circuits. These systems were large and space consuming, often requiring many kilometres of wiring, both to the field and to interconnect the control circuitry. With the invention of integrated circuitry and microprocessors, the functionality of multiple analogue control loops could be replicated by a single digital controller. Digital controllers began to steadily replace analogue control, although communication to the field was still performed using analogue signals. The movement toward digital systems resulted in the need for new communications protocols to the field as well as between controllers. These communications protocols are commonly referred to as fieldbus protocols. More recently, digital control systems started to incorporate networking at all levels of the industrial control, as well as the inter-networking of business and industrial equipment using Ethernet standards. This has resulted in a networking environment that appears similar to conventional networks at the physical level, but which has significantly different requirements.

This paper serves as an introduction to industrial control networks. Industrial networking concerns itself with the implementation of communications protocols between field equipment, digital controllers, various software suites and also to external systems. The specific requirements and methods of operation of industrial networks will be discussed and contrasted with those of conventional networks. Many aspects of the operation and philosophy of industrial networks has evolved over a significant period of time and as such a history of the field is provided. The operation of modern control networks is examined and some popular protocols are described. Although viewed as a mature technology, industrial networks are constantly under development and some current research areas are discussed.

It will be shown that industrial networks cover a large domain and are of increasing importance to fields such as manufacturing and electricity generation. They are highly specialised and make use of a variety of protocols that have been tailored to fulfil the rigorous requirements that result from implementing real-time control of physical equipment. Due to the fact that reliance on automation in the industrial environment is constantly growing, the prevalence of industrial networks is increasing and industrial networks are becoming further integrated with conventional technologies such as the Internet, greater numbers of professionals are required to interact with industrial networks in some way. While specialised knowledge is required for the development, installation, operation and maintenance of such networks, an understanding of the basic principles by which industrial networks function and the requirements that they fulfil is of use to those new to the field or who may interact with industrial networks in a less direct manner.

II. Industrial Network Basics

A. Commercial versus Industrial Networks

Although recent advances in industrial networking such as the incorporation of Ethernet technology have started to blur the line between industrial and commercial networks, at their cores they each have fundamentally different requirements. The most essential difference is that industrial networks are connected to physical equipment in some form and are used to control and monitor real-world actions and conditions [1]. This has resulted in emphasis on a different set of Quality of Service (**QoS**) considerations to those of commercial networks, such as the need for strong determinism and real-time data

B. Galloway is with the Department of Electrical, Electronic and Computer Engineering, University of Pretoria

G.P. Hancke is with the Information Security Group, Royal Holloway, University of London and the Department of Electrical, Electronic and Computer Engineering, University of Pretoria. email: ghancke@ieee.org

Manuscript received August 17, 2010; revised February 12, 2012; revised June 11, 2012.

TABLE I
TYPICAL DIFFERENCES BETWEEN INDUSTRIAL AND CONVENTIONAL NETWORKS

	Industrial	**Conventional**
Primary Function	Control of physical equipment	Data processing and transfer
Applicable Domain	Manufacturing, processing and utility distribution	Corporate and home environments
Hierarchy	Deep, functionally separated hierarchies with many protocols and physical standards	Shallow, integrated hierarchies with uniform protocol and physical standard utilisation
Failure Severity	High	Low
Reliability Required	High	Moderate
Round Trip Times	250 μs - 10 ms	50+ ms
Determinism	High	Low
Data Composition	Small packets of periodic and aperiodic traffic	Large, aperiodic packets
Temporal Consistency	Required	Not required
Operating Environment	Hostile conditions, often featuring high levels of dust, heat and vibration	Clean environments, often specifically intended for sensitive equipment

transfer. Reference [2] discusses several of the requirements of industrial networks in comparison to commercial Ethernet networks. The differences between typical conventional and industrial networks mentioned above are summarised in Table I and expanded upon in detail below.

1) Implementation: Industrial networks are employed in many industrial domains including manufacturing, electricity generation, food and beverage processing, transportation, water distribution, waste water disposal and chemical refinement including oil and gas. In almost every situation that requires machinery to be monitored and controlled an industrial control network will be installed in some form. Each industry presents its own set of slightly different but generally similar requirements, which can be broadly grouped into the following domains [3]: discrete manufacturing, process control, building automation, utility distribution, transportation and embedded systems.

Discrete manufacturing assumes that the product being created exists in a stable form between each step of the manufacturing process. An example would be the assembly of automobiles. As such the process can easily be divided into cells, which are generally autonomous and cover a reasonably small physical area. Interconnection of each cell is generally only at a high level, such as at the factory floor controller. Process control on the other hand involves systems that are dynamic and interconnected, such as steel smelting and electricity generation. Such systems require interconnection at a lower level and the availability of all plant equipment to function. Building automation covers many aspects such as security, access control, condition monitoring, surveillance and heating or cooling. The criticality of the information being gathered is generally lower and the networks are geared more towards supervision and monitoring than control. The large variation in building topology and automation requirements usually results in large variation in network architecture from installation to installation.

Utility distribution tends to resemble discrete manufacturing networks in their requirements, despite the fact that the controlled equipment tends to be interconnected. This is mainly because of the large physical distance covered by the distribution system, which makes interconnectivity of the control network more difficult but also increases the time it takes for conditions at one cell to influence another. Transportation networks also cover large distances as they deal with the management of trains, monitoring of highways and the automation of traffic controllers. Due to the significant presence of humans within the systems to be controlled, their safety requirements can be quite high. Finally, embedded systems generally involve the control of a single discrete piece of machinery, such as the control networks found in cars. Such networks cover a very small physical area, but tend to have demanding environments and a very high safety requirement.

2) Architecture: Industrial networks generally have a much deeper architecture than commercial networks. Whereas the commercial network of a company may consist of branch or office Local Area Networks (**LANs**) connected by a backbone network or Wide Area Network (**WAN**), even small industrial networks tend to have a hierarchy three or four levels deep. For example, the connection of instruments to controllers may happen at one level, the interconnection of controllers at the next, the Human Machine Interface (**HMI**) may be situated above that, with a final network for data collection and external communication sitting at the top. Different protocols and/or physical media often are used in each level, requiring gateway devices to facilitate communication. Improvements to industrial networking protocols and technology have resulted in some flattening of typical industrial hierarchies, especially in the combination of the higher layers. Often however, the network architecture is not flattened as much as is possible, in order to retain correlation to the functional hierarchy of the controlled equipment. For example, power islands within a power generating utility will retain independent control networks in order to retain a logical separation between units both at mechanical and control level. Examples of typical network architectures are given in Figure 1.

3) Failure Severity: Due to the fact that industrial control networks are connected to physical equipment, failure of a system has a much more severe impact than that of commercial systems. The various effects of failure of an industrial network are stressed in [1] and can include damage to equipment, production loss, environmental damage, loss of reputation and even loss of life. Although not always caused by control system failure, numerous industrial disasters such as the Fukashima Daiichi nuclear disaster in 2011 give examples of the impact of a severe industrial failure.

4) Real Time Requirements: The speed at which processes and equipment operate requires data to be transmitted, processed and responded to as close to instantly as is possible.

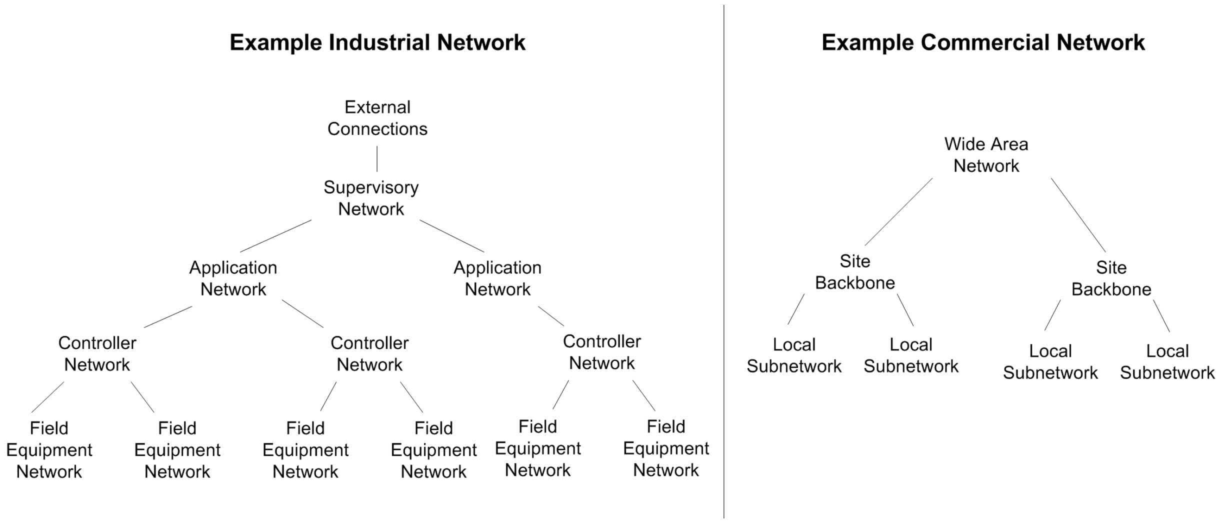


Fig. 1. Illustration of the difference in industrial and commercial network architectures

A general rule is that response time should be less than the sample time of data being gathered. For example, motion control applications have response time requirements in the region of 250 μs to 1 ms [4], although less stringent processes may only require response times of 1 ms - 10 ms. It is also shown in [5] and [6] that delays in information delivery can severely impact the performance of control loops, especially in the case of closed loop systems. Commercial networks tend not to have any response time requirements - if they do they are usually in the range of tens of hundreds of millisecond seconds, or rather seconds. Higher levels of the hierarchy of an automation network tend to have progressively lower time requirements and at the highest levels begin to resemble commercial networks.

5) Determinism: Not only must data used in the lowest levels of an industrial network be transmitted in real time, it must also be done in a predictable or deterministic fashion. For a network to be deterministic it must be possible to predict when a reply to a transmission will be received. This means that the latency of a signal must be bounded and have a low variance. The variance of the response time of a signal is often referred to as jitter. Low jitter is required due to the fact that variance in time has a negative effect on control loops. The derivative and integral portions of a control loop are affected by time variation and digital signal processing methods such as Fast Fourier Transforms require fixed intervals between sampled data. Commercial networks are as a whole not affected by jitter as severely as industrial networks are. Some exceptions to this do exist, such as in voice over Internet protocols, which require low jitter to transport speech. Voice over Internet can still be implemented on standard networks as it simply discards data with a high jitter as speech can withstand a relatively high data loss and still remain legible. Such a solution is not appropriate for industrial use and determinism must be built into industrial network protocols.

6) Data Size: Data packets transmitted in industrial levels are generally quite small, especially at low levels in the architecture where only a single measurement or digital value may need to be transmitted, along with some overhead information. Such transmissions are often only a few bytes in size, such as the transmission of a single binary state or a sixteen bit value. Commercial networks on the other hand regularly transmit kilobytes or more of data, with packet sizes starting at a minimum of 64 bytes. This difference requires significantly different protocols within the network stack, focussed on the transmission of smaller data packets.

7) Periodic and Aperiodic Traffic: Industrial networks require the transmission of both periodically sampled data and aperiodic events such as change of state or alarm conditions. As discussed above, these signals must be transmitted within a set time period. The sampling period used to collect and transmit data may vary from device to device according to control requirements and aperiodic data may occur at any time. To facilitate such transmissions, clocks and bus contention protocols are implemented in industrial network protocols at a low level to ensure that all data transfer occurs in a timely manner. No such considerations exist in commercial networks where data transmission is implemented as 'best effort' and may involve a random delay before data is transmitted.

8) Temporal Consistency and Event Order: There is a need in industrial networks to determine the time at which transmissions occurred and the order of events within a network, especially in the case of aperiodic transmissions. This is achieved using timestamps and synchronised clocks. The ability to guarantee the order and temporal consistency of data delivery is usually not a part of commonly implemented networking protocols such as the Transmission Control Protocol/Internet Protocol (**TCP/IP**).

9) Ruggedness: Industrial networks are implemented in a wide variety of physical locations, often experiencing adverse conditions such as moisture, dust, heat and vibration. In order to withstand such harsh conditions, equipment must be ruggedised with high intrusion protection ratings to prevent damage to equipment from liquids and dust. This contrasts strongly to commercial networks which are, as a whole, located in clean, temperature controlled environments.

B. Information Types

The information which is transmitted in industrial networks is defined as control-, diagnostic and safety information in [7]. Control information is sent between instruments and controllers and is either the input or output of a control loop implemented in a controller. As such, it has strong real-time and deterministic requirements. Examples of control information would include actuator position, tank levels, fluid flow or drive speed.

Diagnostic information is other sensory information collected, but not acted on, by the control system. This information is generally used to monitor the health of plant equipment, examples being the current pulled by a motor or the temperature of a bearing. The term diagnostic information can evoke some confusion, as information regarding the status of the communications medium, instrumentation or control equipment is referred to as network diagnostics. Since diagnostic information is generally not acted on in real-time by the control system, it can also be referred to as monitoring information. Monitoring information has much lower real-time requirements than control information, as it only needs to recorded or displayed and not responded to. Monitoring information does however still require temporal consistency and minimal data loss.

Safety information is used to implement critical functions, such as the safe shutting down of equipment and the operation of protection circuits. It therefore has not only strong real-time requirements, but also requires a high reliability - for example having safety integration levels of two or higher. In the past all, of these functions were implemented in separate networks, but more recently control and monitoring functions have been implemented using a single network. Due to the higher cost involved with implementing the required reliability of safety networks as well as their limited application mean that safety networks are still implemented separately.

Information which has been captured, stored and made available for off-line retrieval is referred to as historic information. This may include control, monitoring or safety information, which physically exists in the plant, as well as abstract values that may be useful for analysis such as setpoints or calculated values. A dedicated historian device is generally used for this purpose.

C. Industrial network components: PLC, SCADA and DCS

Industrial networks are composed of specialised components and applications, such as Programmable Logic Controllers (**PLCs**), Supervisory Control and Data Acquisition (**SCADA**) systems and Distributed Control Systems (**DCSc**). It is the communication within and between these components and systems that industrial networks are primarily concerned with.

1) PLC: PLCs are specialised, computer-based, solid-state electronic devices that form the core of industrial control networks. Sometimes referred to as programmable controllers (**PCs**), PLC is the preferred nomenclature to avoid confusion with the abbreviation for personal computer. Initially developed to meet requirements specified by the Hydramatic Division of General Motors in 1968, PLCs were first used to replace hard-wired relay logic circuits [8]. Some of the major initial requirements set forth were that the devices should be easily programmed and reprogrammed; easily maintained and repaired; smaller in size and cheaper than the relay circuits they would replace; capable of operating within a plant and capable of communicating with central data collection systems.

PLCs have developed significantly in the intervening time and are now available with a wide range of cost and capabilities. Modern PLCs have the ability to perform both binary and analogue input and output, as well as implement proportional, integral and derivative control loops. PLCs generally consist of a power supply, processor, input/output module and communication module. These modules are usually separate and interchangeable, especially in larger, more powerful PLCs. This modularity allows for easier maintenance, as well as greater flexibility of installation - more than one module of each type and modules with different functionality can be combined according to the requirements of the system to be controlled. The development and implementation of PLCs was the first step towards the highly interconnected industrial control networks in use today.

The unique requirements that PLCs address has resulted in a distinct field of research, particularly into design methods and programming languages. This research has resulted in several standards, the most influential of which are International Electrotechnical Commission (**IEC**) standards 61131 and IEC 61499 [9]. IEC 61131 defines five programming languages for use in PLCs - Ladder Diagram, Sequential Function Chart, Function Block Diagram, Structured Text and Instruction List. These languages range from simple graphical representation of relay circuits in Ladder Diagrams, to the assembler-like Instruction List and the high level programming language of Structured Text. IEC 61499 defines different function blocks, their interconnections and their application in PLC program design.

PLC programs are usually written on a computer and many manufacturers have released development environments to aid in program development. There is also a movement towards graphic-based control loop creation to allow for easier programming, with the graphics then being automatically converted into a high level programming language. The actual programming of a PLC is done using specialised programming software, either by utilising a physical connection to a dedicated programming port on the device, or through a network to which the PLC is attached. The programming software often forms part of the development environment, which may also include other features such as the ability to communicate instructions to the PLC, or to view internal variables on a running PLC for debugging and troubleshooting purposes.

2) SCADA: A SCADA system is a purely software layer, normally applied a level above control hardware within the hierarchy of an industrial network. As such, SCADA systems do not perform any control, but rather function in a supervisory fashion [10]. The focus of a SCADA is data acquisition and

TABLE II
SUMMARY OF THE DIFFERENCES BETWEEN A DISTRIBUTED CONTROL SYSTEM (DCS) AND SUPERVISORY CONTROL AND DATA ACQUISITION (SCADA) SYSTEM

DCS	SCADA
Process driven	Event driven
Small geographic areas	Large geographic areas
Suited to large, integrated systems such as chemical processing and electricity generation	Suited to multiple independent systems such as discrete manufacturing and utility distribution
Good data quality and media reliability	Poor data quality and media reliability
Powerful, closed-loop control hardware	Power efficient hardware, often focussed on binary signal detection

the presentation of a centralised Human Machine Interface (**HMI**), although they do also allow high level commands to be sent through to control hardware - for example the instruction to start a motor or change a setpoint. SCADA systems are tailored towards the monitoring of geographically diverse control hardware, making them especially suited for industries such as utilities distribution where plant areas may be located over many thousand square kilometres.

The control hardware that communicates with a SCADA is referred to as an Remote Terminal Unit (**RTU**) and is usually a type of specialised PLC. The device to which the RTUs communicate is known as a Master Terminal Unit (**MTU**). The remote location of RTUs imposes many restraints on the system and is a core aspect of the manner in which SCADA systems are designed. Data communication over such long distances often involves using third-party media such as telephone lines or cellular telephony. These media are often unreliable or have bandwidth limitations. As such, SCADA systems tend to be event-driven rather than process-driven with a focus on reporting only changes in the state of the monitored system rather than sending a steady stream of process variables. For example, an event-driven system would send a binary value indicating that flow through a pipe has dropped below a predefined threshold, whereas a process-driven system would regularly transmit an analogue value containing the flow through the pipe. This allows a reduction in the number of communications sent and lowers bandwidth requirements. SCADA software also needs to take unreliable communications media into account and needs to be able to implement features such as recording the last known value of all variables in the system and determining data quality.

Power supply to RTUs in remote locations is also a concern and RTUs are generally very power efficient. This is often achieved by limiting the processing capability of the device, or through more sophisticated methods such as sending the processor to sleep unless some change is detected. In the past many RTUs only performed rudimentary control, although advances in processor efficiency now mean most RTUs are capable of at least open-loop control.

Environmental conditions also play a large part in RTU specification and RTUs generally have to be extremely durable and reliable in order to withstand harsh field conditions. This is not to say that SCADA systems are only used to communicate with remote equipment - they may be used in situations where both local and remote equipment is present, or where only a supervisory level of control over equipment is required such as factory-level control or building automation. When local equipment is connected, normal PLCs are generally used and communication is usually through some form of fieldbus connected to multiple PLCs rather than through a direct connection using external communications.

A SCADA system usually consists of two application layers - client applications which present the HMI, and server applications which co-ordinate and record data being displayed by the clients as well as manage communication with control devices. The server may function as an MTU, or receive data from one or more dedicated MTUs to which it communicates. The server functions may also be implemented on redundant computers to improve reliability. Client and server applications communicate using Ethernet and communications models such as client-server, server-server or producer-consumer may be implemented.

In addition to the actual server and client software, SCADA systems also consist of other supporting software tools, such as the engineering tools required to configure and troubleshoot the SCADA. Most SCADA systems also contain some method of forwarding data to other applications such as plant historians; Object Linking and Embedding (**OLE**) for Process Control (**OPC**) being the predominant technology for this purpose.

Being purely software based, SCADA systems are heavily affected by standard Information Technology (**IT**) trends, such as advances in the operating systems and computer hardware on which the software runs. This creates situations in which SCADA software can quickly become obsolete as IT evolves [11]. This is especially problematic due to the fact that the control hardware to which the SCADA interfaces usually have life cycles several times that of the computer equipment. This can lead to situations where the communication is implemented using hardware and drivers which are viewed as obsolete and are not compatible with newer computers and operating systems. As such the life cycle of the entire SCADA system is an important consideration. Due to the increased use of conventional IT equipment, information and network security is also a growing concern.

3) DCS: A DCS resembles a SCADA in function, as it is a software package that performs communication with control hardware and presents a centralised HMI for controlled equipment. The difference between the two is often subtle, especially with advances in technology allowing the functionality of each to overlap. The key difference between the two is that DCSs are process-driven rather than event-driven and they generally focus on presenting a steady stream of process information. This means that although the two systems appear similar, their internal workings may be quite different. For example, a DCS may simply poll a controller to obtain whatever data is required to be displayed, rather than maintain records of all last known plant values. To this effect, a much higher level of interconnection both between the software layer and the control hardware, as well as between controllers, is evident. DCSs are also not as concerned with determining the quality of data, as communication with control hardware is much more reliable. As a whole, control hardware consists of traditional PLCs, often with very powerful processors implementing multiple closed-loop controls. This makes a DCS less suitable for geographically distributed systems, but more suitable for highly-interconnected local plants such as chemical refineries, power stations and other process domains.

The high level of interconnection between DCS software and control hardware usually also allows a single engineering tool to be used to both program the controllers and configure the software layer. Many DCSs are marketed as a complete hardware and software package by a single vendor due to the ability to implement such functionality. The use of a single package greatly reduces commissioning time, as a monitored value only needs to be configured once for it to be defined in both the hardware and software, although it also tends to restrict the DCS to use of control hardware from a single vendor only.

On the whole, DCSs and SCADAs use very similar technologies and have a similar architecture at higher levels. DCSs are also usually implemented using computers that communicate with the plant equipment either directly or through a bus, server applications that co-ordinate data and client applications that display data. DCSs are similarly very heavily affected by changes in the IT landscape and have similar security requirements to SCADA.

4) Summary: Specialised programmable electronic controllers form a core part of industrial networks, as they are usually responsible for the actual implementation of the control and protection logic used to operate the plant to which they are connected. Much of industrial networking concerns itself with methods by which information can be transferred between field devices and controllers, between controllers themselves and between controllers and software packages responsible for such functions as providing an HMI or an engineering interface. Such software packages are usually classified as being either a SCADA or a DCS. Although the functionality of both types of software may often overlap, the major differences between the two are summarised in Table II. Both software types are highly affected by advances in conventional information technology and vulnerable to malicious interference at the network level.

III. Origins and Development

The core of industrial networking consists of fieldbus protocols, which are defined in the IEC standard 61158 as “a digital, serial, multidrop, data bus for communication with industrial control and instrumentation devices such as - but not limited to - transducers, actuators and local controllers”. Although fieldbus was originally conceived to be a replacement for the traditional two-wire signalling techniques such as 4-20 mA and 0-10 V used at the lowest level of an industrial control system, the technology has expanded and now presents functionality that can be used at many different levels of a control installation.

According to [12], industrial control networks can be broken up into three distinct generations with varying levels of compatibility. The first consists of traditional serial-based fieldbus protocols, the second of Ethernet-based protocols and the latest generation, which has begun to incorporate wireless communications technologies. The incorporation of Ethernet technology has resulted in a growing similarity between the once distinct fieldbus and Internet technologies. This has given rise to new terms such as industrial control networking, which encompasses not only the functions and requirements of conventional fieldbus, but also the additional functions and requirements that Ethernet-based systems present.

Many articles have been written about the long and somewhat controversial development of fieldbus systems, often by people intimately involved in the development or standardization processes. These include [3], [13], [7] and [12]. This section will cover the main points in the development of industrial control networks, but the reader is encouraged to refer to the cited texts for a more detailed history.

A. FieldBus

Several precursors to what are now known as fieldbus systems were originally in development as early as the 1970s. The development of industrial communications protocols began due to both end-user requirements as well as the appearance of new technologies, which were adapted to industrial settings. Technologies such as programmable microcontrollers and digital signal processors allowed for the replacement of purely analogue control loops with digital controllers such as PLCs. The creation of the Open System Interconnection (**OSI**) seven layer model by the International Standards Organisation (**ISO**) aided significantly in defining and creating communications protocols and services. Advances in local area networking and Medium Access Control (**MAC**) resulted in much more flexible and powerful communications protocols. The concept of Computer Integrated Manufacturing (**CIM**) was developed by the United States National Bureau of Standards, which sought to define a hierarchical structure for the use of computers at all levels of industrial automation. The Manufacturing Automation Protocol (**MAP**) project was created by General Motors and the Technical and Office Protocol (**TOP**) project was created by Boeing, in attempts to create standard communications profiles within the CIM hierarchy. TOP profiles were defined to facilitate communications between business and technical offices, while MAP focussed on communications between factory controllers and control cells. The concept of Mini-MAP or MAP/Enhanced Performance Architecture (**MAP/EPA**) incorporated the factory automation interconnect system specification developed in Japan to define communications profiles within control cells. The Manufacturing Message Specification (**MMS**) was also developed as part of the MAP project. At the lowest level, between controllers and instruments, there was a need to reduce the wiring requirements of traditional signalling. This requirement led to the development of protocols that would be termed fieldbus in 1985 [3].

Many fieldbusses were developed in parallel, both in universities and by various control system vendors, to meet requirements defined by various users in various industry sectors. Due to the large initial investment and relatively long lifetime of control systems, end users preferred open protocols. Open protocols ensured greater availability of compatible instruments and controllers, as well as increased support over the life of the equipment. The developers of the protocols also realised that creation of open protocols allowed the cost of developing a protocol to be shared by companies searching for similar solutions. Proprietary protocols fell away on the whole in favour of open protocols. The standardization of a protocol has many benefits, such as an image of reliability and stability, and strengthening market position [13]. The developers of the various protocols motivated to have their protocols standardised and the pre-eminent fieldbus protocols of today were soon recognised as the national standards of their countries of origin. Examples include PROcess FIeld BUS (**PROFIBUS**) in Germany, Factory Instrumentation Protocol (**FIP**) in France and P-Net in Denmark.

At around this time, the IEC appointed a committee to define an international fieldbus standard. The Instrumentation Society of America (**ISA**) in the United States also appointed a committee to define an American standard. The ISA committee decided to cooperate with the IEC committee, rather than develop a new American standard. The IEC defined a need for fieldbus technologies at two levels: the H1 fieldbus with a low data rate for the connection of sensors in process control, and the H2 fieldbus with a high data rate for manufacturing or for interconnection of several H1 networks. Several protocols were submitted to the IEC committee for consideration, PROFIBUS and FIP being the two strongest contenders [13]. The ISA decided to define requirements in order to aid in their decision. At this time, fieldbus was not envisioned as a real-time system and much of the additional functionality available in modern fieldbus systems was not considered [3]. The emphasis was placed more on what a fieldbus should be able to achieve, rather than how it should achieve it, which became a major stumbling block in coming years.

Although PROFIBUS and FIP were both strong contenders, they both used very different approaches and neither perfectly fulfilled the requirements for an international fieldbus standard. While both used similar hardware, utilising serial RS-482 over Shielded Twisted-Pair to communicate - in the same manner that many other fieldbus protocols of the time did - their communications philosophies and contention management strategies were very different. PROFIBUS was based on a distributed control idea and in its original form supported an object-oriented vertical communication according to the client-server model in the spirit of the MAP/MMS specification. FIP, on the other hand, was designed with a central, but strictly real-time capable, control scheme and with the newly developed producer-consumer or publisher-subscriber model for horizontal communication. The differences between the client-server model and the producer consumer model are described in detail in Section IV-A3. Attempts were made by both parties to strengthen their fieldbus systems in order to meet the IEC requirements. FIP was expanded to become WorldFIP (**WFIP**) which added client-server functionality and the Interoperable Systems Project (**ISP**) attempted to demonstrate how PROFIBUS could be enhanced with the producer-consumer communication model. In the meantime, the IEC began to define their own standard.

After several years no significant progress had been made. The work-in-progress IEC standard had become increasingly complex and unwieldy [13], while the ISP project had been disbanded before reaching a mature state. This lack of progress prompted the American branches of the WorldFIP and ISP projects to combine into the Fieldbus Foundation. The goal of the Fieldbus Foundation was to develop an American fieldbus protocol called Foundation Fieldbus (**FF**), which would combine the bus access scheme of FIP with the application level of PROFIBUS. At this point, the question of fieldbus standardisation had moved beyond the technical requirements, as many of the existing fieldbus systems had become firmly entrenched in industry. Recognising this, the European Committee for Electrotechnical Standardisation (**CENELEC**) published several standards which elevated the existing national standards to European Standards. After lobbying by the British national committee, several American Protocols such as FF and Control Area Network (**CAN**) were also added to the European standards.

Work had continued by the IEC committee during this time, and after the dissolution of the ISP project and establishment of the Fieldbus Foundation, the draft standard began to resemble more a combination of FF and WFIP than PROFIBUS and WFIP. Fearing that PROFIBUS would begin to lose market share to FF, PROFIBUS proponents managed to block the presentation of the new standard with a minimum vote [13]. Although it should be noted that the new standard still contained several flaws, the move sparked outrage and no small amount of controversy. In an effort to reach a compromise, the IEC eventually moved to publish all the existing standards as-is in IEC 61158. This resulted in a large and rather unwieldy standard (well over 4000 pages long), and IEC standard 61784 has since been published in an attempt to clarify the situation. The only IEC-developed portion of the standard was 61158-2, which defined the physical layer and has been adopted by most fieldbusses that provide intrinsic safety. Since then, the standards have been updated to reflect changes to the various protocols, as well as to incorporate some new protocols that fulfil the requirements of fieldbus systems. A timeline of fieldbus development is given in Figure 2. The majority of the newer standards are Ethernet based - the impact of the incorporation of Ethernet into industrial networking will be discussed in Section III-B and some Ethernet protocols will be discussed in Section IV. This situation has both advantages and disadvantages. The large number of protocols leads to a lot of confusion, especially to those unfamiliar with the field. This is only exacerbated by the existence of proprietary protocols used by control systems vendors that are based on open protocols. Vendors and the fieldbus institutions would all have users believe that *their* fieldbus is the best solution for any and all industrial communications needs, making it even harder to distinguish the true differences between the various protocols. The differences do exist though, as can be

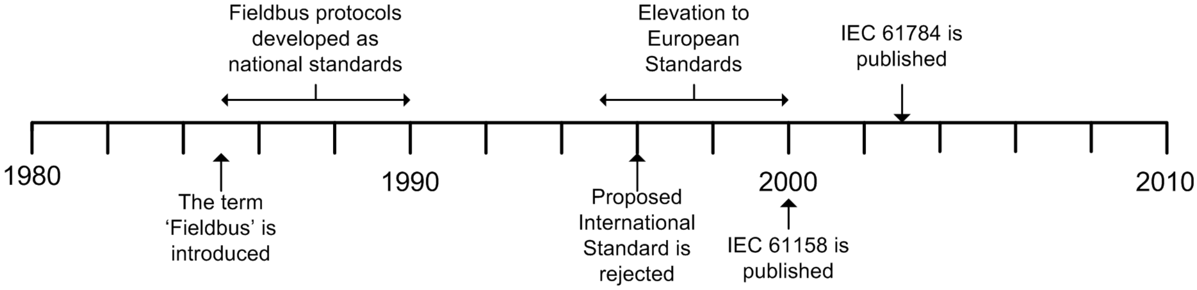

Fieldbus protocols developed as national standards
Elevation to European Standards
IEC 61784 is published
1980
The term 'Fieldbus' is introduced
1990
Proposed International Standard is rejected
2000
IEC 61158 is published
2010

Fig. 2. Timeline of Fieldbus development

seen by the difficulties experienced in attempting to create a protocol robust enough to be seen as the international standard. If such a protocol had been successfully developed, it would likely have been large and complex, increasing the cost of equipment and the configuration requirements of implementing it in differing applications. By having a diverse selection of protocols available, a fieldbus can be chosen for a specific task at a cheaper price and lower complexity. While this prevents the interoperability of all equipment, most companies attempt to make use of a minimum number of vendors in any case to minimise on training requirements and spares holding. In addition, technology such as OPC has helped significantly in communication between different systems, albeit at higher levels. Overall, the decision to create a compromise standard was the best available, as users are provided with standardisation that aids with longevity and support for their equipment, while still being able to implement a system suitable for their requirements that has the best price and lowest complexity. This is especially true when the requirements of the system are low and a simpler protocol is sufficient.

B. Ethernet Fieldbus

Although Ethernet, as part of the TCP/IP and User Datagram Protocol (**UDP**) stack, quickly became the prevalent standard for home and office use, it initially did not gain much acceptance in industrial areas. This was mainly due to the fact that it was designed with very different network QoS considerations, as discussed in Section II-A. However, advances in Ethernet technology have made the medium more suited to industrial use. The result has been a trend towards Ethernet-based fieldbus protocols, especially at H2 level. The increased data rates of newer Ethernet standards (for example 802.3u Fast Ethernet) make it easier to create real-time Ethernet protocols, as the transmission and retransmission times are significantly shorter. The implementation of full-duplex Ethernet lines allows for data transmission and reception to occur simultaneously, easing bus arbitration difficulties. Another advance that has allowed Ethernet to be considered for industrial use is the introduction of switched networks as opposed to the older hub based networks. Network hubs simply relay signals received on one port out onto all other ports, resulting in a physical medium that is very congested. Switched networks relay data received one port only onto the ports on which the recipients of the data are located. This allows some of the bus arbitration to occur within the switch, as they can buffer incoming data until it can be transmitted further. It should be noted, however, that the buffering can result in serious delays, especially in congested networks. In fact, it is shown in [14] that hub-based networks outperform switched networks at low loads, due the lack of switching delays. One method to alleviate some of the switching delays is to use pass-through switches instead of store-and-forward switches. Store-and-forward switches buffer an entire incoming packet before attempting to retransmit it, also allowing the switch to examine the packet and check it for errors. Pass-through switches examine the header of the received packet and begin retransmitting the data before the packet has been completely received. This results in significantly smaller switching delays (especially in the case of multiple switches) at the cost of allowing corrupted packets to be retransmitted rather than being discarded as they would by a store-and-forward switch. Significant research is also being undertaken into methods by which network delays can be modelled and compensated for [15].

Just because technology had arisen that made Ethernet more suitable for industrial use did not in itself mean that Ethernet should be used in industrial environments, especially because existing serial-based protocols had already been developed to address industrial communications requirements. However, it can be shown that the use of Ethernet presents several advantages, which justify the development of real-time Ethernet protocols for use as Ethernet fieldbusses. By using the existing Ethernet standards as a foundation, the advantages of Ethernet can be incorporated into the newer protocols. This includes the large amount of research that has gone into developing Ethernet as a standard, as well as the cheap and readily-available Ethernet hardware. The use of Ethernet also allows a flattening of the vertical hierarchy within a control network, simplifying the configuration requirements. It also allows for easier interconnection between business and industrial networks in order to relay process and control information to interested parties. In fact, it is possible to run business and industrial applications on a single network, although this is not advised for both network loading and security reasons. Through the use of standardised Internet applications such as eXtensible Markup Language (**XML**), Hypertext Transfer Protocol (**HTTP**) and File Transfer Protocol (**FTP**) is also possible for non real-time communications, such as configuration and maintenance activities, to be implemented. An example of this is the Electronic Device Description Language

(**EDDL**), which allows for the configuration and calibration of smart instrumentation through a standard XML interface. Another advantage that Ethernet offers is the ability to use technologies such as link spanning to implement redundant communication paths. It should be noted that even the rapid spanning tree protocol is not able to revert to a redundant path quickly enough to satisfy the real-time requirements of industrial Ethernet. As a result the majority of the redundancy protocols available for industrial use are proprietary [16].

The introduction of Ethernet into the field of industrial networking also presented some new challenges. The existing Ethernet standards had to be extended or modified to meet the stringent requirements of industrial networks. This was achieved at various levels of the IP stack, and using various approaches. Some of these will be discussed in Section IV. Of major concern with the incorporation of Ethernet technology is network security, which will be discussed in Section V-B. Backwards compatibility with existing fieldbus protocols is also an issue. Many of the newer Ethernet-based fieldbus protocols are extensions of existing protocols and various compatibility philosophies have been implemented. These are classified into four categories by [12]. The first is full compatibility at higher layer protocols, such as exists with Foundation Fieldbus HSE, MODBUS/TCP, Ethernet/IP and P-Net on IP to name but a few. This approach is especially prevalent in building automation fieldbusses. Another approach is compatibility of data objects and models, such as is the case with PROFINET. This approach requires the use of proxy hardware to allow communication between the fieldbus media. A lesser amount of compatibility is offered through the use of application layer profiles from existing protocols, as is implemented in Ethernet Powerlink and EtherCAT. With these protocols, the CANopen application layer is implemented to retain compatibility with existing device profiles, but compatibility with CANopen itself is not possible. Lastly, completely new protocols have been developed for Ethernet that have no relationship with any existing protocols and have forgone any compatibility. Examples of such protocols are Ethernet for Plant Automation and Time-Critical Control Network (**TCNet**).

After the compromise standard of IEC 61158 had been finalised late in the year 2000, the standardisation committee began on work defining the requirements and operational profiles of real-time Ethernet. While some might have hoped that the move towards Ethernet within the automation industry might result in the development of a single fieldbus standard where the original standardisation effort had failed, it became apparent from the structure of the working groups formed and their goals that another compromise standard was the most likely outcome [17]. This was the most likely outcome for a number of reasons. The standardisation situation for industrial Ethernet greatly resembled that of the initial fieldbus standardisation effort in the 1980s and would likely encounter the same difficulties and delays if the same initial approach was taken. Due to the fact that the majority of the Ethernet fieldbus protocols are extensions of existing serial protocols, most vendors provided and continue to provide upgrade paths from serial to Ethernet for their existing installations. This resulted in the new fieldbusses becoming as entrenched as their predecessors as each was the logical move forward from their predecessors. The work of the standardisation committee has therefore focused more on the refinement of the existing standards and identifying methodologies to address the new challenges Ethernet presented as a medium. Four new working groups were established in addition to the existing maintenance group and function block group. These groups are the following: a group to handle industrial cabling requirements; a group to handle the implementation of real-time communication without straying from the specifications of the original IEC 802-3 Ethernet specification; a group concerned with the implementation of safety functions using Ethernet and the final group concerned with cyber security.

Ethernet has, however, not become the *de facto* medium for fieldbus, at least not at all levels. In fact, it is possible that serial fieldbusses will always have a place in industrial networks. This is because of the increased cost of Ethernet fieldbusses compared to serial fieldbusses, as well as the distance limitations imposed on copper Ethernet cables such as CAT 5e. The increased cost is mainly due to the need for Ethernet switches to connect fieldbus components, whereas serial fieldbus components can usually be connected to a simple terminal block or star coupler. Although the price of Ethernet switches has dropped significantly since Ethernet was first implemented, the extra ruggedisation and redundancy required for field implementation, as well as the fact that each instrument requires a port on a switch, can rapidly escalate the cost of installation. This is especially true for installations with signal counts in the thousands. Serial fieldbusses can be implemented at distances of over a kilometre over copper, whereas Ethernet must be implemented using more expensive fibre cables to transport data more than a few hundred metres. These limitations have made Ethernet particularly unsuited to application at H1 level. It has, however, become very popular at H2 level, which generally covers shorter distances and requires the interconnection of fewer components. In these situations, the increased cost of Ethernet can be weighed against the increased data speeds, interconnectivity with commercial protocols and the easily implementable redundancy that Ethernet offers.

IV. Industrial Network Protocols

A. *Fieldbus Operation*

1) Network Stack: In 1984 the ISO defined the seven layer OSI reference model which consists of physical, data-link, network, transport, presentation, session and application layers. Each of the layers describes the services required to send information from one application to another, as well as interfaces between the layers in order to aid with the interconnection of standards. The physical layer concerns itself with the physical transmission of data over a medium; the data-link layer with the organisation of data and detection of transmission errors; the network layer with how data is routed from one application to another; the transport layer with transparent transfer of data; the session layer with organising and synchronising data exchange; the presentation layer

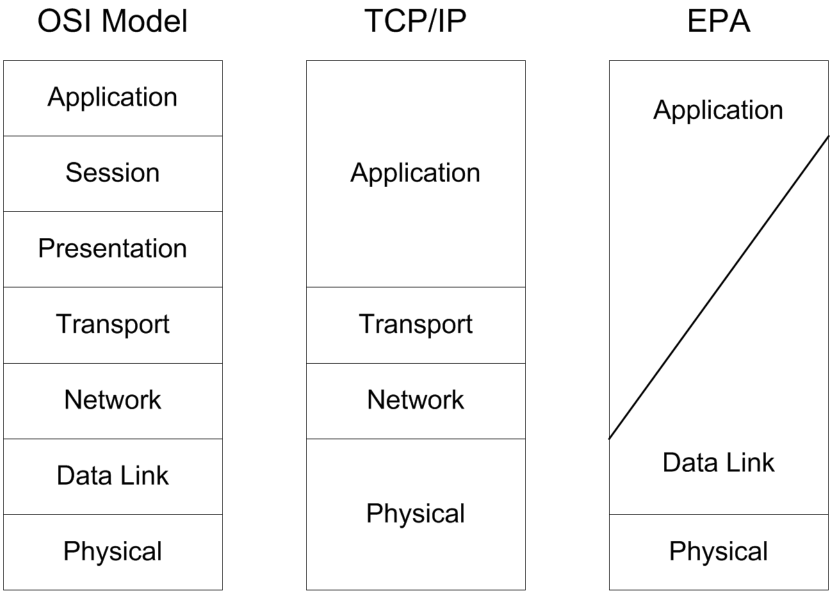


Fig. 3. Comparison of network stack configurations

with the transformation of syntax and the application layer with the management of the communication. While not often implemented as-is in realised protocols, the reference model is used as a benchmark for information exchange and is useful for analysing the manner in which various protocols operate. In reality, most communication protocols do not encounter all the problems described in the reference model, or choose to combine the requirements of one or more layer for the sake of simplicity. For example, the TCP/IP protocol consists only of physical, network, transport and application layers. This protocol is still fully functional and is used throughout the Internet and as the basis for real-time Ethernet. The majority of serial fieldbusses, including all of those defined in IEC 61158 work according to the reduced model defined in MAP/EPA. The MAP/EPA model consists of only three layers - physical, data-link and application, as shown in Figure 3. The main reason behind the introduction and utilisation of this reduced model is to reduce the delays introduced by passing information between layers and processing it at each layer [18]. Network layer functionality is generally not required, or is implemented at application level if information must be passed from one network to another. The small size of data being transmitted means that the transport layer can also be omitted, although the size of a packet of data in the application layer is then limited to that of the packet size in the data-link layer. The organisation of data exchange is implemented in the data-link layer to ensure determinism. While the strict requirements of fieldbus mean that the services presented by each layer are very similar, a variety of methods have been implemented to achieve them. The method of implementation is generally the biggest difference between each of the fieldbus protocols.

Real-time Ethernet implementations are all based on the four layer TCP/IP model, with some modifications to achieve determinism. Real-time requirements can be achieved through one of three approaches [19]. Common across all approaches is the use of Ethernet cabling and TCP and UDP for non real-time communications. Modification of the TCP/IP stack may be done only at the application level to use standard data packets, the transport level may be modified to use custom ethertypes for real-time communications, or the Ethernet data-link layer may be modified to apply mechanisms and infrastructure that allow for real-time communication. These approaches are called ‘on top of IP’, ‘on top of Ethernet’ and ‘modified Ethernet’ respectively, as shown in Figure 4.

When real-time Ethernet is implemented on top of IP, the application layer is responsible for scheduling communication such that the communication requirements are met. This makes it possible for communication to occur outside of network boundaries, and for external networks to be used for communication with remote devices. Such communication can, however, introduce non-deterministic delays and the scheduling device must be equipped with adequate resources.

Should the implementation be on top of Ethernet, the physical Ethernet layer remains unchanged, but custom ethertypes are defined alongside standard types such as IP. Both standard and custom ethertypes can be used within the network, but the network equipment and connected devices must have knowledge of the custom protocol. Often the custom ethertypes will be given dedicated bandwidth or priority within the network.

Direct modification to Ethernet mechanisms are usually made to enable non-standard topologies such as rings or busses to be implemented. Switching and routing functionality may often be implemented at device level, or the hardware of network equipment may be modified to manage the topology correctly. Such an implementation requires that all of the connected equipment be compatible at hardware level.

2) Bus Access: When fieldbus was first developed, it was viewed as a wiring simplification solution and its implementation was treated as a media access problem [20]. While this disregarded much of the application level requirements that have since originated, media access is still an important part

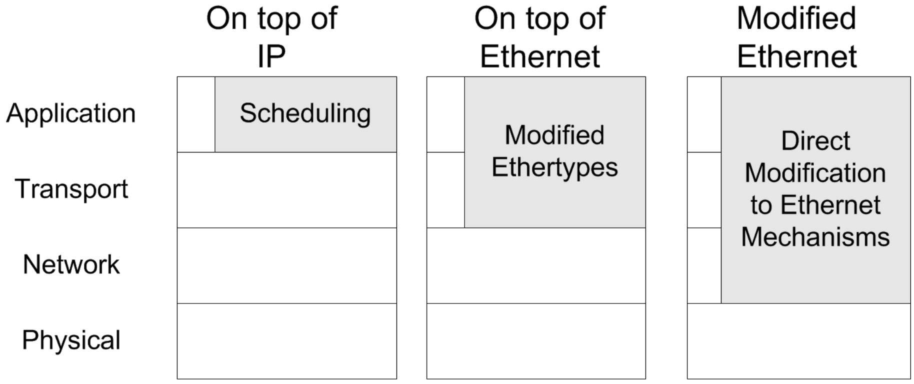


Fig. 4. Industrial Ethernet stack implementations

of fieldbus operation, especially with respect to maintaining determinism. The majority of fieldbusses control access to the transmission medium through the use of a bus master, although some operate without controlled access. The majority of those that operate without access control make use of Carrier Sensing Multiple Access with Collision Avoidance (**CSMA-CA**). CSMA-CA works through a contention algorithm that allows the message with the highest priority to be transmitted in the event of two devices attempting to transmit simultaneously. The bus is synchronised by a clock and before each message is transmitted the device first transmits the priority of the message as a binary integer. If the transmitting device detects that another device has transmitted a one at the same time it has transmitted a zero during the contention segment, it stops transmitting and waits for the next available transmission slot. This does place some limitations on the system, in that data priority must be discrete across the devices to ensure that no two devices are able to transmit at the same time. CSMA can also not be implemented using RS-485, since RS-485 is a balanced medium and does not allow the transmission of a zero to be 'over-ridden' by the transmission of a one.

When access to the transmission medium is controlled, two approaches can be taken - either the control is centralised, or it is decentralised. Decentralised control is usually implemented through a token passing system wherein the station holding the token is allowed to determine who transmits data. Token passing can add a significant amount of overhead to a network, since the token needs to be passed along even if the device that receives it has no need to arbitrate the bus at the time. However, token passing has been shown to be highly efficient in heavily-loaded networks, where the overhead is insignificant in comparison to the amount of transmitted data and the possibility of delays due to bus contention is high [21]. Centralised data control has a single device that is responsible for determining when transmissions occur, although many fieldbusses allow for redundant implementation of the master in order to improve reliability.

Controlled systems are, by their nature, suited to periodic traffic, with the bus master knowing ahead of time when data is expected and permission to transmit must be given. Periodic traffic often makes up the greater part of the traffic on a fieldbus, especially in DCSs and other process-oriented systems. However, aperiodic traffic must still be catered for in both event and process driven systems as it often contains data of a critical nature, such as notification of an alarm or other fault. Aperiodic transmission is handled in a number of different ways, although the basic premise of each method is to leave a set amount of bandwidth open in which aperiodic traffic is allowed to transmit. This can be done through leaving open slots in a transmission cycle that a device may request use of, or by having a field within every data packet left open for transmission of aperiodic data alongside periodic information.

3) Information Distribution: The distribution of information within an industrial network is interlinked with bus access, since the bus controller needs to know from where data is supplied and where it is required, as well as when it should be transmitted to ensure that scheduling of data is done properly. In order to maintain determinism, the method used to deliver data to a controller or instrument must be predetermined and managed. Various formats for packaging data have been developed, some specifically for use in fieldbusses, some co-opted from other applications. MMS, Simple Network Message Protocol (**SNMP**) and IEC 870-5 are each used by various bus protocols to determine the composition of a data packet and the format of data within a packet.

Two main methods are used for information distribution. The first is the client-server distribution model, which operates in the same manner as client-server interaction in traditional network applications. The client sends a message to the server, requesting that it fulfils some service - in this case, to provide a packet of information. The server then replies with a message that fulfils the request. Client-server is often used in conjunction with a token-passing bus access strategy. Each bus master can poll other devices and request necessary information as well as allow the devices to reply before passing the token on to the next master. This can result in a data mismatch between bus masters if they each request the same data from a device and the state of the data changes between the requests [20]. Variations of the client-server model are also implemented, such as client-multiserver in which the server acts as a proxy and obtains the required information through its own client-server requests to other devices; third-party client-server, in which the principal server has another server reply to the request from the client on its behalf and multiconfirmation client-server in which the server may reply to the client several times in order to fulfil the requested service.

The other method of data distribution is the publisher-subscriber model, which operates either with an information 'push' or an information 'pull'. In a pull publisher-subscriber model, the publishing manager will send a request that a publishing device transmit some information. Subscriber devices which require the information are individually responsible for listening for the response. The publisher will then broadcast the required information to the entire network, allowing the listening subscriber devices to receive it. In a push model on the other hand, no publishing manager is used. Subscribers will use client-server interaction to link themselves directly to a publisher. The publisher itself determines when it shall transmit and does so using an unconfirmed transmission.

There are several differences between the client-server and publisher-subscriber methods of information distribution. Publisher-subscriber requires that a broadcast capability be present in the network as it does not specify the addresses of the subscribers when transmitting. It also requires that subscribers be able to receive information they did not specifically request. The publisher-subscriber model is better suited for event-driven traffic, especially in the case of the push configuration where the publisher that detected the event is responsible for transmission of the related information. Client-server is better suited for process data, as a client will only receive data related to an event if it specifically requests it from a server. This requires that server applications require some method of indicating to clients that unexpected data, such as an aperiodic transmission, is available.

B. Protocol Overview

1) Controller Area Network: Controller Area Network (**CAN**) was originally developed by Bosch in the early 1980s for use in automobiles. It uses CSMA-CA for bus contention, which requires it to use an unbalanced, non-return-to-zero coding scheme, in this case RS232, for physical transmission. The publisher-subscriber model is used for information distribution. CAN is defined in ISO 11898 and only specifies the physical and data-link layers. Due to its lack of high level functionality, such as the provision of an application layer, CAN itself is unsuited for industrial automation. It is however used as the basis for other fieldbus protocols that define their own higher level services above the CAN specification. Examples of such protocols include CANopen, DeviceNet, ControlNet and Smart Distributed System. The CAN protocol specifies eight byte data exchanges to ensure short maximum bus access time and has a maximum speed of 500 kbits/s. As such, it and its derivatives are more suited for use at H1 level.

2) CANopen: CANopen is a high level expansion of CAN for use in automation developed by Bosch before being handed over to the CAN in Automation Organisation, which now manages the protocol. CANopen benefits from a strong European presence and vendor independence [22] and was defined in the European EN 50325 standard along with other CAN based protocols. The CANopen standard defines a wide variety of application profiles for specific implementations, such as motion control, building door control and medical equipment.

3) ControlNet and DeviceNet: ControlNet is also an application layer expansion of the CAN protocol, and also defined in EN 50325. Originally developed by Allen-Bradley (now Rockwell Automation), it has since been handed over to the Open DeviceNet Vendor Association (**ODVA**) for management. ControlNet implements the Common Industrial Protocol (**CIP**) application layer and is optimised for cyclical data exchange, making it more suited to process systems. As its name suggests, it was developed specifically for transmission of control data and has a high emphasis on determinism and strict scheduling. One notable feature of ControlNet is its built-in support for fully redundant cabling between devices.

DeviceNet is a variant of ControlNet, with a focus on device-to-device communication. In most respects it is very similar to ControlNet as it was also originally developed by Allan-Bradley and is now maintained by the ODVA. Both protocols have a large American user base [22] and are noted for their cost-effectiveness [21].

4) EtherNet/IP: Ethernet Industrial Protocol (not to be confused with the Ethernet Internet protocol) is an Ethernet-based implementation of the CIP application layer on top of TCP/IP. Originally developed by Rockwell Automation, it is maintained by the ODVA along with the other CIP fieldbusses. The use of the CIP application layer allows for a tight integration between the three fieldbusses and communication between them can be implemented through the use of gateway devices. Although not strictly deterministic, EtherNet/IP delivers real-time performance through the use of prioritised messages and clock synchronisation using the IEEE 1588 protocol. These considerations are combined with a full-duplex switched architecture, which prevents delays due to collisions. Actions in the network are based on planned timing as opposed to actual timing in order to counter delays encountered within the network stack [19]. The EtherNet/IP standard is defined in IEC 61784-1.

5) PROFIBUS: PROFIBUS is arguably one of the most well-known and widely-implemented fieldbusses, due to its endorsement by Siemens. PROFIBUS was developed by a consortium of various German companies and institutions and was one of the first fieldbusses to be created. Originally managed by various regional organisations, these were joined together to form PROFIBUS International which is now tasked with the maintenance of the standard, as defined in EN 50170, IEC 61158 and IEC 61784. Different profiles are defined within PROFIBUS, each for different applications. Non-deterministic high level communications between cells are catered for by PROFIBUS-FMS, while low level communication is realised using PROFIBUS Distributed Periphery (**DP**). Other variants are PROFIBUS Process Automation (**PA**), which is designed specifically for use in hazardous areas and is intrinsically safe, PROFIdrive for motion control and PROFIsafe for safety systems [23]. All the variants implement a token-passing bus access strategy with multiple masters able to poll other devices for information, the main difference being the application profiles defined in each. This allows for a high degree of interoperability between the different busses. PROFIBUS is mainly implemented using RS485 at the physical layer, except for PROFIBUS-PA, which makes use of the IEC 61158-2 physical layer to achieve intrinsic safety by limiting current on the bus.

6) PROFINET: PROFINET, defined in IEC 61158 and IEC 61784 is the adaptation of PROFIBUS data models and objects onto Ethernet and is also maintained by PROFIBUS International. PROFINET is available in two variants - Component Based Architecture (**CBA**), envisioned for use as an H2 fieldbus and Input/Output (**IO**) for use as an H1 fieldbus. PROFINET makes use of remote procedure calls (**RPC**) and the distributed component object model (**DCOM**) for communications in the range of 50 ms - 100 ms, as well as modified ethertypes for real-time communication. The use of modified ethertypes means that PROFINET is realised on top of Ethernet. Both RPC and DCOM were originally developed as part of the Microsoft Windows network stack. PROFINET-CBA is implemented through the use of component descriptor files, which abstract the services provided by a device with the intention that the realisation of the communication be implemented separately to promote vendor independence [24]. PROFINET-IO works similarly with application and communication relationships defined in a general station description file. PROFINET also allows for high level applications such as asset management to be implemented. Compatibility to PROFIBUS, as well as INTERBUS and DeviceNet, is achieved through the use of proxy devices.

7) INTERBUS: INTERBUS is a RS485 based fieldbus standard defined in EN 50254 and IEC 61158, developed and maintained by Phoenix Contact in Germany. It operates using a ring topology with a single bus master. Each device in the ring is connected in a point-to-point fashion; receiving, amplifying and passing on messages to the next device in the bus. This architecture means that there are no arbitration delays and it uses its 500 kbits/s transmission rate very efficiently. It also means that the bus is highly deterministic. A nested implementation is also allowed up to sixteen levels deep, with local branches connected to each terminal on the bus. At the lowest level, transmitters and actuators are connected through an INTERBUS Loop [25]. As such, INTERBUS is able to fulfil both low and medium level communication requirements and is particularly suited for connecting remote input/output modules. Its implementation as an H2 network is, however, limited by the speed of the bus.

8) WorldFIP: WorldFIP was developed as an expansion of the original FIP protocol in an attempt to fulfil the requirements for an international fieldbus. Originally developed by a conglomeration of French institutions, it is now managed and maintained by the WorldFIP Organisation. Much like PROFIBUS, it was one of the first fully-fledged fieldbusses to be developed and is recorded in EN 50170, IEC 61158 and IEC 61784. WorldFIP is notable in that it was the first fieldbus to implement a producer-consumer model and contains built-in support for redundant cabling. It is also fairly unique in that it consists of only a single variant intended for use at both H1 and H2 levels and can operate at either 31.25 kbit/s, 1 Mbit/s or 2.5 Mbit/s depending on requirements.

9) Foundation Fieldbus: Foundation Fieldbus can be seen as a combination of PROFIBUS and WorldFIP and was developed by the American Fieldbus Foundation in response to the delays encountered with establishing an international fieldbus standard. Despite its American origins, it was included in the European EN 50170 standard and consequently in the IEC 61158 and 61784 standards. Developed to address low level requirements in process industries, the original Foundation Fieldbus is now referred to as Foundation Fieldbus H1 due to the advent of Foundation Fieldbus Safety Instrumented Functions for use in safety applications and Foundation Fieldbus High Speed Ethernet for higher level applications. FF H1 makes use of the producer-consumer model of WorldFip and the device interfaces developed by the ISP [26]. Producer-consumer communication is used for cyclical data, unscheduled data transfer is managed through client-server communications and unscheduled multicast is possible for event notification. The protocol specifies the intrinsically safe IEC 61158-2 physical layer that operates at 31.25 kbit/s and is able to supply power to field devices. This ability does, however, require dedicated power supply and power conditioning modules to be connected to each bus [27].

10) Foundation Fieldbus HSE: Developed by the Fieldbus Foundation, Foundation Fieldbus High Speed Ethernet was designed to address the need for H2 level communications within the Fieldbus Foundation's protocol suite. One of the first Ethernet-based fieldbusses developed [28], HSE is fully compatible with H1 at the application level and for all intents and purposes is simply an implementation of the H1 protocol over the faster physical medium. The implementation is on top of the TCP/IP stack, with additional use of standard IP interfaces such as dynamic host configuration protocol and simple network management protocol [29]. As with other Ethernet-based fieldbusses, the use of switched networks is a prerequisite and redundancy can be implemented. Connectivity of H1 busses directly onto an HSE backbone can be achieved through the use of linking devices.

11) P-Net: P-Net is a low level fieldbus of Danish origin that is defined in EN 50170 and IEC 61158. Like many other low level fieldbusses, P-Net makes use of RS485 as a transmission medium, but has several distinguishing features. P-Net is particularly focussed on small installations with an emphasis on cost-effectiveness and efficiency. The code required for a device to communicate over P-Net has a very small footprint and can be implemented without the need for specialised communication chips - this reduces the cost of devices as well as the delays encountered by passing information from a communications module to a processor. Due to this, up to 300 transactions can occur a second despite the transmission rate of the bus being set at 76.8 kbit/s. Other distinguishing features are a focus on bus segmentation to allow for concurrent transmission on a single bus, while still retaining direct addressing between bus segments. Multiple masters are allowed per bus segment, with a client-server information distribution module. Contention is managed through the use of virtual token passing, which eliminates some of the overhead associated with token passing networks. Process data is also transmitted in standard international units rather than as digital values to minimise data conversion at higher levels [30].

12) HART: T he Highway Addressable Remote Transducer (**HART**) protocol was developed by Rosemount and handed over to the HART Communications Foundation for management. HART was not developed as a fieldbus in the strictest sense, although it can be implemented as such. HART operates by modulating an analogue 4-20 mA signal using frequency shift keying with an amplitude of ±0.5 mA to transmit data at 1200 bits/s. HART is able to operate in the manner of a standard fieldbus with up to 15 devices connected in a parallel multidrop configuration, in which the 4-20 mA signal simply provides power and all communication is digital. However, a purely digital HART configuration is too slow for most control tasks due to the extremely low data speed [31]. Instead, HART is normally implemented using either point-to-point communication or using dedicated time division multiplexer hardware which provides access to specific point-to-point connections when requested. In such configurations, plant data is transmitted as a continuous analogue signal, with digital communication reserved for application level communication. As such, HART provides a communications architecture that greatly resembles traditional analogue configurations but which also allows for the implementation of device descriptor files and other smart-instrumentation functionality.

13) OPC: Although not a fieldbus protocol, OLE for Process Control (OPC) forms part of many industrial networks at higher levels by providing a standardised interface for communication of industrial data. Maintained by the OPC Foundation, the original OPC standard (now referred to as OPC Data Access) uses RPC and DCOM to allow real-time communication of process values over Ethernet with a client-server model. Several other variants of OPC have also been developed, including OPC Historical Data Access which allows for retrieval of stored values, OPC Data Exchange for two-way communication using a server-server model and OPC XML Data Access which uses XML for communication. DCOM exchanges are difficult to secure due to their use of random ports for each transaction and require both parties to be located on the same network domain, which is not always possible to implement. As such, OPC is usually combined with tunnelling software that performs local transactions with the OPC interface and transmits them through a secured virtual private network.

14) Other Fieldbus Protocols: Recently, several new real-time Ethernet-based fieldbus protocols have been ratified by the IEC and added to the 61158 and 61784 standards. Due to their relative youth, they have yet to achieve significant market penetration or the level of academic attention given to the more established protocols. These protocols include Ethernet PowerLink defined by Berneker & Rainer and defined by the Ethernet Powerlink Standardisation Group; EtherCAT defined by Beckhoff and supported by the EtherCAT Technology Group; TCNet developed by Toshiba; Ethernet for Plant Automation developed in China and Vnet/IP developed by Yokogawa. Several proprietary fieldbus standards have also been released by various device vendors, usually consisting of extensions of existing standards to provide additional functionality and security specific to the operation of their equipment.

V. Current Research Areas

A. Wireless Technology

There is currently a trend within industrial networking to implement fieldbus protocols using wireless technologies [7]. There are many parallels between the current movement towards wireless and the previous movement towards Ethernet. As with Ethernet, it was decided that the reutilisation of existing standards was preferable to the development of new physical and data-link layers specifically for industrial use, as it allowed the existing research and manufacturing base to be exploited in order to decrease development time and costs. Technologies that make use of unlicensed bandwidth are the most popular, such as Wireless Local Area Network (**WLAN**) IEC 802.11, IEC 802.15.1 known as Bluetooth and IEC 802.15.4 which is used as the basis of the ZigBee protocol. The benefits of wireless technology are clear - a further reduction in the amount of wiring required for communication, which in turn reduces installation costs.

Wireless is also particularly suitable for hazardous environments or installation on moving equipment where cabling may be easily damaged or restrict the operation of the machinery to be monitored. Faster commissioning and reconfiguration can also be realised [32]. However, it can be said that standard wireless technology is even less suited to industrial use than Ethernet was and adaptation of the existing technology for real-time communication is the subject of much research. For example, [33] describes attempts to implement PROFIBUS over wireless. Both [32] and [34] discuss the use of wireless technology in industrial automation at length and the reader is encouraged to consult them for a deeper understanding of the field.

Much like Ethernet, the existing wireless technology was developed for use outside of industry and no considerations for real-time response or determinism are inherent in the media. Wireless faces additional challenges that need to be addressed for industrial application [32]. Wireless is highly susceptible to interference from a variety of sources, which causes transmission errors. Within the transmission channel itself, effects such as multi-path fading and intersymbol interference are present. Interference from other transmission channels is also possible, such as might occur at the boundaries between two wireless fieldbusses. Environmental electromagnetic emissions may also affect wireless transmission, such as those produced by large motors and electrical discharges. Thermal noise can negatively affect transmission, as can the Doppler-shift induced by rapidly moving equipment. Such interference is often transient in nature, resulting in bursts of data and affecting the reliability and determinability of the transmission. Wireless transmission radii are limited by transmission strength and negatively affected by path-fading, the degree of which is determined by environmental factors. This makes it difficult to design a wireless network for industrial use without first determining the path-fading coefficient throughout the intended usage area.

The limited distance over which wireless transceivers can operate, combined with the use of carrier sensing to determine when it is safe to transmit, may also result in what is referred to as a 'hidden terminal' problem, where two devices located out of the range of each other try and communicate with a third device that is located between them without knowledge of the other's actions. Wired carrier sensing technologies such as Ethernet are able to avoid such problems by ensuring that each device has knowledge of all others to which it is connected, for example by limiting the total length of cable allowed between any two stations. Even with careful planning and device location, such knowledge cannot be guaranteed in a wireless medium. Wireless transceivers are also only able to operate at half-duplex, as their own transmissions would overpower any signal they might be intended to receive.

Physical overhead on a wireless system is also significant in comparison to wired systems, as most wireless protocols require the transmission of predetermined data sequences before or during data transmission in order to evaluate and correct the effects of noise on the received information. Security of wireless transmission is also of concern, as physical access to the transmission medium cannot be restricted. Many wired fieldbusses are also able to make use of passively-powered field devices by supplying the energy required for the device's operation over the transmission medium. The existing wireless technologies have no such capability and provision for energy to remote devices is a concern, as is the energy efficiency of the remote devices.

In addition to difficulties in realising general reliability and timeliness requirements, the characteristics of wireless transmission can negatively affect specific fieldbus methodologies. Fieldbusses often utilise unacknowledged transmission, since the probability of data not being received at all is relatively low. Such a strategy is unsuitable for wireless where the possibility of nonreception of a broadcast is significantly higher. This is especially troublesome in the case of token-passing networks, where the loss of the token may result in the bus needing to reinitialise to re-establish which device is the current master. Since interference is generally not uniform, some equipment may receive a broadcast while others do not. This can result in data inconsistency across a network in which the producer-consumer model is utilised. The half-duplex operation of wireless also means that carrier sensing with collision avoidance is not possible and a protocol such as CAN cannot be implemented.

Several techniques can be implemented to improve the performance of wireless in industrial application. Hidden node problems can be solved by adding a handshake system to the network, in which permission to transmit must be requested and granted before transmission may occur. This allows the receiver to inform all other devices in its range, some of which may be out of the transmitter's range, that it is expecting a transmission and requires the channel to be kept open. This does however add significant overhead to the channel, especially in the case of small data packets, where the initialisation of transmission may require more time and data than the actual information to be communicated. Interference can also be combated in a number of manners. Error correcting codes can be added to data that will not be acknowledged, at the price of increased overhead, and retransmission requests can be sent for data that is acknowledged.

Retransmission requests only add overhead to the channel when a transmission fails, but the time required to retransmit may delay other transmissions. Retransmission may also be unsuccessful for a significant period due to the bursty nature of interference. A combination of error correction and retransmission requests can also be implemented. Since interference is often localised, exploitation of spatial diversity can be achieved by using multiple, physically separate antennas. In instances where multiple antennas cannot be implemented, devices may also attempt to route data through third parties in the hope that clear channels exist between the third device and each of the two devices attempting to communicate. More advanced error mitigation strategies may also be implemented, such as deadline awareness and increased error correcting overhead for retransmitted signals.

Each of the various technologies being investigated for wireless use has its own advantages and disadvantages. Bluetooth is typically used over short ranges of less than 10 m and uses very little power. A master-slave structure is implemented to provide some contention management and ad-hoc networks are the expected usage. It also implements a frequency-hopping algorithm to minimise interference and to allow multiple Bluetooth networks to operate within the same physical area. A variety of different packet types are specified, with differing lengths, coding strategies and retransmission allowances. Like Bluetooth, ZigBee also focusses on low power transmissions over relatively short distances, but is tailored towards static networks with infrequent transmissions and small packet sizes. ZigBee devices can be either fully functional or feature reduced functionality. Fully functional devices are able to communicate in a peer-to-peer manner and act as contention masters for reduced devices. Reduced devices can only communicate with master devices, through managed and unmanaged contention systems. WLAN is technically a collection of standards, each defining various physical layers and media access control strategies. Examples of this are 802.11b, 802.11g and 802.11n, each of which feature differing modulation schemes and data throughputs. 802.11e is also under development with the goal of providing better support for time-critical functions. WLAN networks can be implemented ad-hoc, or, more popularly, through a central access point. WLAN features much higher data rates that Bluetooth or ZigBee, but is very inefficient when transmitting small data packets [32].

Research into the adaptation of wireless technologies has been ongoing for more than a decade into a variety of topics such as quality of service provisions, media access protocols, security, energy efficiency, scalability, network planning methodologies, error control, mobility, scalability, routing algorithms and the integration of wireless into existing wired systems [34]. Commercial industrial wireless systems are only just beginning to appear and the field can still be considered to be in its infancy. An example of a commercial system is the wireless interface for sensors and actuators developed by ABB.

Open protocols are also beginning to emerge and are nearing readiness for commercial adoption. Three protocols for wireless communication have recently been approved as IEC standards, namely ISA100.11a, WirelessHART and Wireless Networks for Industrial Automation - Process Automation (**WIA-PA**), in standards 62734, 62591 and 62601 respectively. The three standards share several common features, such as the use of the IEC 802.15.4 physical layer [35]–[37] also used in the ZigBee Protocol. These protocols overcome one of the major weaknesses of ZigBee by modifying the 802.15.4 media access control functionality to implement frequency hopping [38]. WIA-PA retains full compatibility with the 802.15.4 physical standard, whereas ISA100.11.a and WirelessHART do not [37].

The protocols are intended for use in communicating with field instruments and fulfil a similar purpose to that of H1 fieldbusses. Although the terminology used to describe specific components differs from standard to standard, all of the standards are defined to cater for a similar set of devices. These are security and network management devices, gateway devices, routing devices, non-routing devices and handheld devices. The various instruments connect in a self-organising hybrid star/mesh network, which is controlled by the network and security management devices. The management devices are powerful, wired devices, which interface to the wireless portion of the network through the gateway device. The gateway device can also be implemented as a protocol converter, making use of a wired fieldbus protocol to facilitate deterministic communication between the gateway and any controllers [39]. The mesh portion of the network is realised by the routing devices, which in turn connect nearby non-routing devices through the star portion of the network.

Despite the similar operational philosophy of the protocols, they feature different network stacks and are incompatible. Some of the key differences are that WIA-PA and ISA100.11a allow for some of the network management functionality to be implemented in the routing devices, while WirelessHART only allows for centralised management by the management device. WIA-PA implements a two-level data aggregation system, ISA100.11a a single level of aggregation and WirelessHART does not specify any aggregation functionality. All three standards specify a time synchronization function to allow for time division multiple access to the communications medium, with ISA100.11a having an adjustable timeslot aligned to international atomic time. WirelessHART and WIA-PA use fixed timeslots of 10ms aligned to coordinated universal time [37].

The implementation of wireless industrial networks will likely remain an active research area for a significant time, especially due to the fact that wireless communication is still developing and new technologies will need to be adapted for industrial use. At this time, the main use envisioned for wireless in industrial networks is as part of hybrid systems where last-mile communications at H1 level are implemented wirelessly [40], which is the manner in which the current set of standards are intended to be used.

In summary, the major advantages being pursued in the development of wireless industrial networks are

- Lower cabling costs
- Installation of wireless instruments in locations where cables may be restrictive, impractical or vulnerable
- Faster and simpler commissioning and reconfiguration

For these advantages to be realised, existing wireless protocols are being adapted to provide the following features.

- Resistance to heavy interference on the transmission medium
- Provision of deterministic, real-time communication along unreliable, non-static routes
- Energy efficient wireless devices

The three most promising open standards which aim to fulfil the requirements for wireless industrial networks are WPA-IA, WirelessHART and ISA100.11a.

B. *Security*

Security in industrial networks bears a strong resemblance to that of commercial networks due to the growing overlap of the technologies used in both. While many of the same threats exist to both networks, the additional requirements and considerations of industrial networks mean that security may often be more difficult to implement. The goal of network security is to provide confidentiality, integrity of information, availability, authentication, authorisation, auditability, nonrepudiability and protection from third parties [41]. The lack or loss of these features can result in a situation where a failure of the network may occur.

The failure of an industrial network can have severe repercussions, as detailed in Section II-A3. Such failure could be accidental, or caused by malicious intent. Prevention of these failures is provided by reliability and security respectively, although the two aspects of the systems are tightly interlinked - security flaws can be viewed as reliability flaws that are exploited deliberately [42]. However, where the network itself cannot, or has not, addressed these flaws through its own reliability considerations, additional measures must be put in place to prevent access to the flaws and increase the security of the system. Securing industrial networks has become a prerequisite for securing critical infrastructure at a national level. This is true for all industrialised nations and a greater dependence on the development and implementation of industrial network security is realised as greater levels of automation and computer-dependence is implemented within chemical processing, utility distribution and discrete manufacturing [43], [44].

During the initial implementation and development of digital automation systems, a policy of ‘security through obscurity’ [41] was seen as adequate protection. Control networks were often physically separate from any other systems and employed technology rarely encountered outside of the industrial environment. At this time the main threats to the integrity of a system were from accidental interference or from the malicious actions of a disgruntled worker [45].

As the nature of control systems has changed, this situation has changed dramatically, with new vulnerabilities that are inherent to control systems and the equipment on which they are based. Controllers have become computer based, equipment is networked and may be accessible over the Internet, commodity IT solutions are becoming increasingly popular, open protocols have found widespread use, the size and functionality of control systems is increasing, a larger and more highly skilled IT workforce has become available and cybercrime has become a serious threat [46].

As Ethernet became the dominant technology within the higher levels of automation systems and the expected number of external connections to industrial networks grew, the need for security was recognised. At first, the main threats were seen as being incidental to the technology in use, with most security considerations aimed at preventing accidental exposure of the industrial network to conventional threats. Possible intruders to the network were viewed mainly as a nuisance rather than as serious opponents, with talk of 'teenage hackers' [47] and 'mischievous adversaries' [48]. The majority of incidents caused by security failures were not directly targeted at the affected systems - for example the loss of servers and HMI computers due to the spread of malicious software from corporate networks, or the failure of communications paths to RTUs due to third-party channels becoming compromised by a conventional virus.

This has recently changed, with skilled, knowledgeable cyber-terrorist organisations now posing the greatest threat to industrial networks. This Advanced Persistent Threat (APT), i.e. skilled adversaries who target and repeatedly try to attack systems, is most evident in the recent Stuxnet virus. Termed a 'cyber-weapon of mass destruction' [49], the virus shows an alarming degree of sophistication and specialist knowledge [50], [51]. The virus was composed of three components, each with a specific function. The first, termed the 'dropper', propagated itself through computer systems, mainly through the use of flash drives. The dropper was capable of determining whether software used to program PLCs was installed on any computer it infected. If this was the case, the dropper replaced certain libraries within the PLC programming software with compromised versions of the library. This allowed the virus to examine code being sent to, or read from a PLC in order to identify specific target PLCs. Once the specific PLCs had been identified and connected to, the purpose of the dropper was to deliver the other two components onto the PLC itself. This was achieved by appending segments of machine code to valid communication from the programming software and then hiding the additional segments when machine code was retrieved from the PLC, effectively creating the first known PLC 'rootkit'. The malicious code was designed to slowly degrade the physical integrity of specific centrifuges, most likely installed at a nuclear enrichment plant in Iran, by minutely affecting the acceleration and deceleration of the centrifuge arms. In addition, the code contained pre-recorded snippets of the correct operation of the centrifuges, which were reported back to operators and engineers at the plant in order to prevent them from detecting that any equipment had been compromised.

The level of sophistication shown in the engineering of the virus required specific knowledge of the physical equipment in the plant, the control loops in place and the architecture of the control network. The effects of the virus could have been considerably worse - malicious code of a similar nature could easily cripple a country's infrastructure by forcing equipment in utilities to shut down or damage itself. It can therefore be seen that the security of industrial systems is of critical concern and is an ongoing research area, especially by government agencies and other oversight committees. The governing bodies of the various fieldbus standards and the academic institutions associated with each are also heavily invested in order to gain competitive advantage.

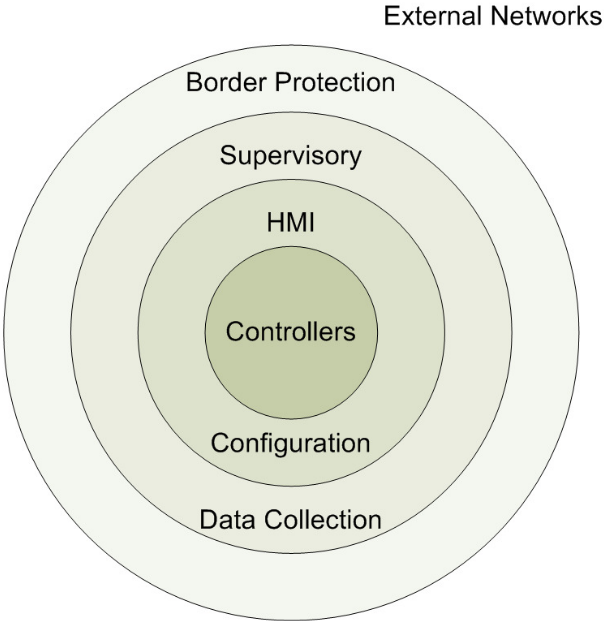


Fig. 5. Example Defense in Depth network structure

Security should be implemented at all layers of the control network, with each layer further isolating subsequent layers from external threats. Such an approach is referred to as 'defense in depth', with the most critical equipment being the most protected [1]. Such a layered network implementation is shown in Figure 5.

The outermost layer of security should prevent unauthorised access to the network itself from external sources. In the past this was trivial, as industrial networks were generally stand-alone systems. The growing amount of integration with business networks has made this a much more complex requirement. Plant data might be required by engineers or other employees working on the business network, information concerning the plant may be needed at other plants or at central locations and vendors may need dedicated remote access to assist with troubleshooting.

Firewalls are generally used to restrict electronic access to the network, and Virtual Private Networks (**VPNs**) may be used to establish remote connections. Firewalls are available with a variety of capabilities, ranging from simple devices, which block communication based on source or destination addresses to powerful devices, which are able to inspect the contents of communication and dynamically decide whether information should be passed on or blocked. At the minimum, a firewall should be placed between the industrial network and any external network to which it connects. However, a single firewall may often be inadequate, depending on the level of access that is required. For example, high level devices such as plant historians often pose a challenge to single firewall installations. If the historian is located on the industrial network many client devices on the business network must be given access to the industrial network to communicate with the historian. Alternatively, the historian could be placed on the business network and be granted access to all the devices on the industrial network from which it gathers data. In either scenario, the firewall must be configured to be very open, with a high level of interaction allowed between the business and industrial networks.

The solution is to utilise a DeMilitarised Zone (**DMZ**) firewall configuration, which makes use of two firewalls placed in series between the two networks. Any equipment that requires communication with both the business and industrial networks is placed between the two firewalls, within the DMZ. Each firewall can then be configured to allow the required level of interaction into the DMZ, but blocking any communication attempts from the business network directly to the industrial network and vice versa. An example of this implementation is shown in Figure 6. This configuration is not foolproof, as the servers located in the DMZ may still allow an intruder access to the industrial network if they are compromised. However, it is easier to make sure that the DMZ servers are sufficiently impervious to attack so as not to be compromised than it is to ensure the same level security across the whole of the process and business networks. Physical access to the industrial network should also not be overlooked - network equipment, computers and controllers should be housed in areas with limited physical access for approved personnel only.

No network can be rendered impenetrable through access control alone. Networks should ideally demonstrate an absence of reaction to malicious access [52]. The system itself should therefore be configured to minimise the effects of malicious access to the system. Unused ports on switches and routers should be disabled, as should data access capabilities of USB ports on computers within the network. User accounts and passwords should also be in place on all the equipment, to prevent unauthorised operation of the device should either physical or electronic access to it be gained. Software installed on devices should be kept up-to-date and operating systems should be patched to mitigate vulnerabilities. Such actions are often referred to as ‘hardening’ the equipment. Access control and boundary security mechanisms such as firewalls are also not as effective at countering insider threats, i.e. authorised persons acting in malicious ways. This threat is best dealt with by organisational means, like clearly delimiting employee responsibility, auditing and logs of actions and other organisational security measures.

In addition to the hardening of equipment, communications channels between devices also need to be secured. Cryptographic algorithms form a core part of securing communications in commercial networks, as they provide data confidentiality, integrity and authentication. The use of conventional network equipment means that many established technologies such as the IP Security and Secure Socket Layer protocols can be used at higher levels. Unfortunately, the nature of control equipment makes implementation of security features at lower levels problematic. Industrial equipment generally has a much longer life cycle than that found in corporate networks, and has much higher reliability requirements. As such, the technologies used in industrial networking equipment are generally mature and proven at the time of installation - by the end of the equipment’s life-cycle it may be several generations older than the latest technology [41].

Security threats evolve at the rate of the latest technology and older equipment often lacks the capacity to implement current best-practice security algorithms within real-time constraints. Factors such as key length and algorithm complexity are limited by processing power when attempting to implement any form of cryptography. In addition, other aspects of low level industrial protocols make implementation of security difficult. The low data transfer rate of many protocols means that they would be adversely affected by the additional overhead required for secure communication. Conventional cryptographic mechanisms are also very sensitive to all levels of electronic noise [42].

Conventional security protocols such as IP Security, Secure Socket Layer and VPN are not practical for use in low level industrial automation networks due to their lack of support for multicast- and broadcast transmissions [53]. Key distribution is also problematic in the use of cryptographic algorithms in industrial networks, as cryptographic keys may be needed by thousands of devices. Various approaches to key distribution have been discussed, for example loading keys onto physical storage and installing them at each device [48], or distributing keys electronically at install time when other configuration settings are loaded onto an instrument [54]. Many of the key distribution methods envisioned involve a high level of manual intervention during the commissioning of the equipment and fail to consider the lifetime of the keys. The length of the key and the algorithm in use determine the length of time it would require to decrypt sensitive information, and the two are normally matched to the expected lifetime of the data to be protected.

In terms of data confidentiality in industrial networks, the required lifetime may be of a short duration, if it is required at all. Authentication, on the other hand, needs to be maintained for the life of the equipment, which is generally several years. Due to the limited processing power and bandwidth in industrial networks, algorithms cannot be implemented that are able to deliver such long lifetimes. Therefore, the key will need to be replaced before the minimum amount of time in which it would be possible to decrypt the algorithm and deduce the key. To manually facilitate key replacement in large systems would be impractical, especially if equipment is only able to implement cryptographic algorithms with lifetimes measured in days or weeks. The practical implementation of secure communications within the lowest levels of industrial networks is currently a topic into which much research is being done, as many aspects such as effective key management remain an open problem [55].

Another research area which is receiving a lot of attention is the identification of vulnerabilities of existing protocols and equipment [56], [57], as well as on methodologies by which to

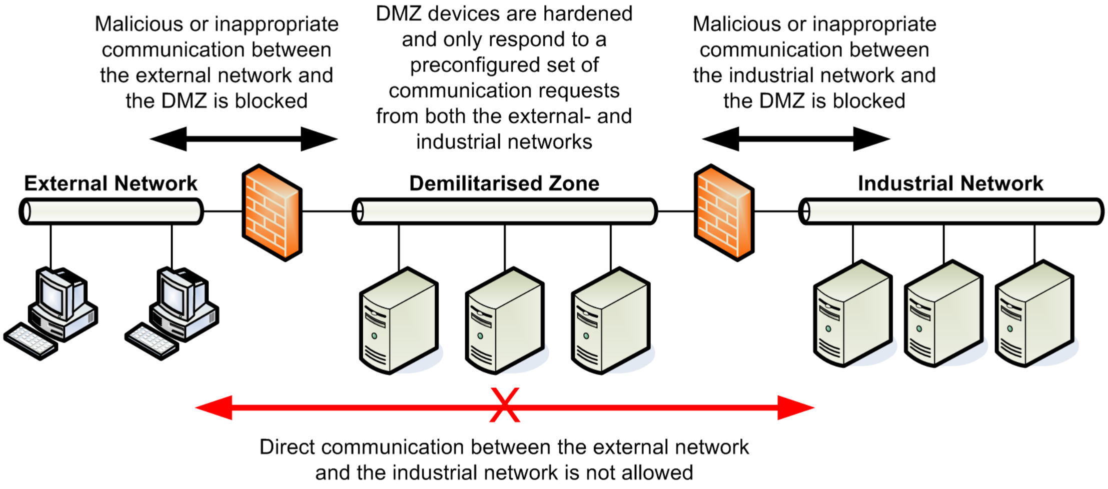

Malicious or inappropriate communication between the external network and the DMZ is blocked
DMZ devices are hardened and only respond to a preconfigured set of communication requests from both the external- and industrial networks
Malicious or inappropriate communication between the industrial network and the DMZ is blocked
External Network
Demilitarised Zone
Industrial Network
Direct communication between the external network and the industrial network is not allowed

Fig. 6. Example of DMZ Implementation

analyse existing networks in order to detect and mitigate vulnerabilities. These methodologies generally focus on detecting chains of vulnerabilities [58] or developing attack trees [59], as overcoming even low levels of security on a network often involves exploiting a series of several vulnerabilities before effecting a meaningful compromise. Such analysis is vital in the formulation of an effective security policy, which is often one of the most difficult aspects of successfully securing a network. Not only does the creation of a security policy require careful analysis of equipment and protocols, the means of addressing identified vulnerabilities must be balanced against cost and practicality of execution. It is important to remember that a security implementation should not interfere with the operation of personnel or equipment, else it will likely be circumvented by its users [60].

In summary, network security is becoming an increasingly important part of industrial networking in order to ensure

- Confidentiality of equipment operation and configuration
- Resistance to incorrect or malicious actions

There is no set method by which security can be implemented, and security cannot ever be said to be perfect, due to the possible presence of undiscovered vulnerabilities. Some of the aspects of industrial networks make implementing security difficult are

- Industrial equipment often has limited processing power and long lifecycles
- The application of patches and security updates may not be possible due to availability requirements
- The definition and implementation of border protection often involves multiple parties with different goals, priorities and skillsets.
- Security provisions cannot be allowed to negatively affect the correct operation of the control system
- Conventional security measures are often not applicable or practical within an industrial context

VI. Lessons to be Learnt

There are a number of lessons that can be learnt from an examination of the history of fieldbus protocols and the manner in which they have developed. The failed attempt at an international serial fieldbus standard highlights many of the possible pitfalls that can be encountered should a standardisation process become delayed or excessively influenced by market interests. The importance of open standards is also evident - despite the plethora of standards that are available, there is still a reasonable amount of interoperability provided by protocol converters and gateway devices. Such interoperability would not have been possible had the protocols been proprietary or restricted to use by specific manufacturers. Proprietary protocols would also have increased the cost and complexity of installing and operating industrial networks, due to additional licensing fees and intellectual property concerns.

When implementing an industrial network, designers should be aware of the core differences that exist with relation to commercial networks, especially when considering architecture, real-time requirements, determinism, temporal consistency and event order. The need for low latency communication in an industrial measurement and control environment is rather clear, but minimising the time taken for data to be transmitted between entities does in itself not satisfy measurement and control conditions. It is as important to determine when data was transmitted and the order of transmissions, even from different points of origin, as this is crucial in identifying and isolating events.

A programmable logic controller (PLC) is responsible for the lower layer logic and functionality in an industrial network. The life cycle of these devices are generally long as they are specially designed to be robust and reliable. Careful consideration must therefore be given to the capabilities of these devices when a system is first implemented, as it is unlikely for there to be a regular opportunity to upgrade or replace a PLC, as opposed to a regular client computer in a commercial network. PLCs should be specified to contain enough resources to allow for future network upgrades. At the same time, designers should also consider the use of proprietary systems, which remain despite attempts to standardise and define open protocols, and the impact this will have on future system development. The use of proprietary SCADA or

DCS with system-specific PLCs results in a situation where the distributor or provider is essentially responsible for improving the system, a situation in which a client might be unable to respond to a quickly developing threat like a system error or security vulnerability.

Some Ethernet-based industrial network protocols are extensions of previous bus-based protocols. Although it is to be expected that the lower network layers would differ, the level of compatibility between the these protocols at higher protocol layers differs from technology to technology. Some technologies are fully compatible, while others offer limited compatibility by means of compatible data object and models, or application layer profiles. System designers should keep in mind that further proxy or translator hardware might be required to interface between Ethernet and bus networks at the application layer. Unfortunately it is very difficult to predict what level of compatibility future protocols with existing, which is also a concern during network design.

Examination of the security aspect of industrial networks, as well as the attitude often associated with it in the past also shows the dangers of complacency and assumption. Both serial and Ethernet based fieldbus protocols were developed without any significant security features, despite the criticality of the equipment to which control networks are connected, and the growing awareness of security vulnerabilities in related fields. The manner in which the Stuxnet worm targeted software and communications protocols specifically intended for industrial use shows that security features should be a top priority. Wireless fieldbus protocols do not suffer from this lack of security, partly because wireless transmission is inherently insecure and the technologies on which wireless fieldbus protocols are based were developed to overcome this shortfall. The developers of wireless fieldbus protocols do appear to have learnt from the security shortfalls of previous generations of fieldbus and have extended the security functionality of the base technology.

In industrial networks, where performance is crucial, introducing additional functionality comes at a cost and trade-offs must be considered. Careful consideration must be given to which security services are implemented, and new threats must be identified and addressed. As discussed in the previous paragraph, security in industrial networks was at first an afterthought. Access control and integrity mechanisms that prevent unauthorised modification of network parameters is an obvious requirement and was once considered to be adequate security. However, in recent times confidentiality has also become important, as information about industrial processes become an attractive target for commercial competitors looking to improve their own industrial processes. In addition to technical security services, organisations should implement an accepted information security management system, such as detailed in ISO/IEC 27001. This means that the organisational processes are in place to deal with security issues as they arise, which is especially useful in industrial networks where new security threats can be identified at any time as research in this area increases.

The development of wireless fieldbus also show that a wide range of areas in which innovation is possible still exist, even in a field as established and mature as that of control hardware and industrial networking.

VII. Conclusion

The field of industrial networking is of vital importance to the continued operation of all forms of industry in which physical equipment must be controlled. Since the advent of the first fieldbus protocols, industrial networks have become widely implemented and are being used to a greater degree to fulfil a wide variety of control, safety and plant monitoring requirements.

Industrial networks offer a wide range of benefits that can be realised through their installation - reduction of cost and commissioning time through the use of low level fieldbusses, easier maintenance and configuration through the use of smart instruments that can perform application level communication, high levels of communication between controllers through the use of high level fieldbusses, and a greater overall integration both within a control system and with outside networks. However, it also has its disadvantages - greater levels of complexity increase the difficulty of troubleshooting; a greater level of understanding is required to configure and maintain control networks; the large variety of standards could make design choices more difficult and lower the level of interoperability between device vendors, and the greater level of integration exposes control networks to attack by malicious parties. On the whole, the benefits outweigh the disadvantages and control networks in some shape or form are constantly achieving a greater level of market penetration. By employing a proper degree of understanding of the technologies involved to create a thorough user requirements specification, it is possible to obtain a control network that is robust and well-suited to the equipment to which it is attached.

The technologies used to control and monitor plants have continually evolved and continue to do so, both affecting and affected by user requirements as additional capabilities and performance become available. Protocols ranging from fully mature and developed to those still in their infancy are available and supported. The long life-time of industrial networking equipment combined with the capability of the original low level fieldbusses means that combinations of these technologies can be found in a single installation.

Technological advancements from related fields such as computing, electronic communication and the Internet have been adapted for industrial use in order to save costs and make use of existing research. The adoption of the Ethernet physical standard and the ongoing adoption of wireless physical standards have resulted in a greater level of interconnection between industrial and commercial networks. The use of standards such as TCP/IP, HTTP and XML has resulted in a further blurring of the lines between traditional- and industrial networking. However, the two should not be confused - despite their growing resemblance they each fulfil fundamentally differing requirements. Due to this there is a growing need for engineers and technicians who understand not only the operation of the underlying commercial technology but also the strict and specific needs of the industrial environment and the operation of industry-specific protocols and standards. This is especially true in the case of network security where industrial networks are becoming increasingly vulnerable to threats native to their adapted technological base. Such concerns have traditionally been the realm of information technology professionals, but knowledge of both commercial best-practice and industrial requirements is needed to maximise security without compromising on the growing functionality requirements.

References

[1] K. Stoufer, J. Falco, and K. Scarfone, "Guide to industrial control systems (ICS) security," National Institute of Standards and Technology, Final Public Draft, Sep 2008.

[2] J.-D. Decotignie, "A perspective on Ethernet-TCP/IP as a fieldbus," in *IFAC international conference on fieldbus systems and their application*, Nov 2001, pp. 138–143.

[3] J.-P. Thomesse, "Fieldbus technology in industrial automation," *Proceedings of the IEEE*, vol. 93, no. 6, pp. 1073–1101, June 2005.

[4] P. Neumann, "Communication in industrial automation - what is going on?" in *Control Engineering Practice*. Elsevier Ltd, 2006, vol. 15, pp. 1332–1347.

[5] M. S. Branicky, S. M. Phillips, and W. Zhang, "Stability of networked control systems: Explicit analysis of delay," in *Proceedings of the American Control Conference*. AACC, Jun 2000, pp. 2352–2357.

[6] F. li Lian, J. Moyne, and D. Tilbury, "Network design considerations for distributed control systems," *IEEE Transactions on Control Systems Technology*, vol. 10, no. 2, pp. 297–307, Mar 2002.

[7] J. R. Moyne and D. M. Tilbury, "The emergence of industrial control networks for manufacturing control, diagnostics, and safety data," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 29–47, Jan 2007.

[8] K. T. Erickson, "Programmable logic controllers," *IEEE Potentials*, pp. 14–17, Feb/Mar 1996.

[9] G. Frey and L. Litz, "Formal methods in PLC programming," in *IEEE International Conference on Systems, Man, and Cybernetics*, vol. 4, 2000, pp. 2431–2436.

[10] A. Daneels and W. Salter, "What is SCADA?" in *International Conference on Accelerator and Large Experimental Physics Control Systems*, 1999, pp. 339–343.

[11] J. D. McDonald, "Developing and defining basic SCADA system concepts," in *Rural Electric Power Conference*, 1993, pp. B31–B35.

[12] T. Sauter, "The three generations of field-level networks - evolution and compatibility issues," *IEEE Transactions on Industrial Electronics*, vol. 57, no. 11, pp. 3585–3595, Nov 2010.

[13] M. Felser, "The fieldbus standard, history and structures," October 2002, presented at Technology Leadership Day 2002, organised by MICROSWISS Network.

[14] R. Viégas, R. A. M. Valentim, D. G. Texira, and L. A. Guedes, "Analysis of protocols to ethernet automation networks," in *SICE-ICASE International joint Conference*, 2006, pp. 4981 – 4985.

[15] R. A. Gupta and M.-Y. Chow, "Networked control system: Overview and research trends," *IEEE Transactions on Industrial Electronics*, vol. 57, no. 7, pp. 2527–2535, Jul 2010.

[16] K. Hansen, "Redundancy ethernet in industrial automation," in *10th IEEE Conference on Emerging Technologies and Factory Automation*, vol. 2, Sept 2005, pp. 941–947.

[17] M. Felser and T. Sauter, "Standardization of industrial ethernet - the next battlefield?" in *Proceedings of the 2004 IEEE International Workshop on Factory Communication Systems*, Sept 2004, pp. 413–420.

[18] R. Patzke, "Fielbus basics," *Computer Standards and Interfaces*, vol. 19, pp. 275–293, 1998.

[19] M. Felser, "Real-time ethernet – an industry perspective," *Proceedings of the IEEE*, vol. 93, no. 6, pp. 1118–1129, June 2005.

[20] J. P. Thomesse, "A review of the fieldbuses," *Annual Reviews in Control*, vol. 22, pp. 35–45, 1998.

[21] F.-L. Lian, J. R. Moyne, and D. M. Tilbury, "Performance evaluation of control networks," *IEEE Control Systems Magazine*, pp. 66–83, Feb 2001.

[22] A. McFarlane, "Fieldbus review," *Sensor Review*, vol. 17, no. 3, pp. 204–210, 1997.

[23] PROFIBUS International, "PROFIBUS system description," http://www.profibus.com/nc/downloads/downloads/profibus-technology-and-application-system-description/display/, 2010.

[24] PROFIBUS International, "PROFINET system description," http://www.profibus.com/nc/downloads/downloads/profinet-technology-and-application-system-description/display/, 2009.

[25] "INTERBUS basics," http://www.interbus.de/get.php?object=497, year =2001,.

[26] "FOUNDATION fieldbus," http://www.samson.de/pdf_en/l454en.pdf, Nov 1999.

[27] "Fieldbus wiring guide," http://www.relcominc.com/pdf/501-123\%20Fieldbus\%20Wiring\%20Guide.pdf.

[28] S. Vitturi, "On the use of ethernet at low level of factory communication systems," *Computer Standards and Interfaces*, vol. 23, pp. 267–277, 2001.

[29] S. J. Vincent, "FOUNDATION fieldbus high speed ethernet control system," http://www.fieldbusinc.com/downloads/hsepaper.pdf, 2001.

[30] The International P-NET User Organization, "The P-Net fieldbus for process automation," http://www.p-net.org/download/590004.pdf, 1996.

[31] SAMSON AG, "HART communications," http://www.samson.de/pdf_en/l452en.pdf, Dec 1999.

[32] T. Brooks, "Wireless technology for industrial sensor and control networks," in *Sensor for Industry, 2001, Proceedings of the First ISA/IEEE Conference*, 2001, pp. 73–77.

[33] J. Kjellsson, A. E. Vallestad, R. Steigmann, and D. Dzung, "Integration of a wireless I/O interface for PROFIBUS and PROFINET for factory automation," *IEEE Transactions on Industrial Electronics*, vol. 56, no. 10, pp. 4279–4287, Oct 2009.

[34] A. Willig, "Recent and emerging topics in wireless industrial communications: A selection," *IEEE Transactions on Industrial Informatics*, vol. 4, pp. 102–124, May 2008.

[35] "The ISA100 standards - Overview and Status," www.isa.org/isa100, International Society of Automation, Tech. Rep., 2008.

[36] A. Kim, F. Hekland, S. Petersen, and P. Doyle, "When HART goes wireless: Understanding and implementing the wirelesshart standard," in *Emerging Technologies and Factory Automation, 2008. ETFA 2008. IEEE International Conference on*, Sept 2008, pp. 899–907.

[37] W. Liang, X. Zhang, Y. Xiao, F. Wang, P. Zeng, and H. Yu, "Survey and experiments of WIA-PA specification of industrial wireless network," *Wireless Communications and Mobile Computing*, vol. 11, no. 8, pp. 1197–1212, Aug 2011.

[38] H. Hayashi, T. Hasegawa, and K. Demachi, "Wireless technology for process automation," in *ICCAS-SICE, 2009*, aug. 2009, pp. 4591 –4594.

[39] T. Zhong, M. Zhan, Z. Peng, and W. Hong, "Industrial wireless communication protocol WIA-PA and its interoperation with foundation fieldbus," in *Computer Design and Applications (ICCDA), 2010 International Conference on*, vol. 4, June 2010, pp. 370–374.

[40] S. Aslanis, C. Koulamas, S. Koubias, and G. Papadopoulos, "Architectures for an integrated hybrid (wired/wireless) fieldbus," Master's thesis, University of Patras.

[41] D. Dzung, M. Naedele, T. P. Von Hoff, and M. Creavtin, "Security for industrial communication systems," *Proceedings of the IEEE*, vol. 93, no. 6, pp. 1152–1177, Jun 2005.

[42] D. Serpanos and J. Henkel, "Dependability and security will change embedded computing," *Embedded Computing*, pp. 103–105, Jan 2008.

[43] D. J. Teumim, *Industrial Network Security, 2nd Edition*, 2nd ed. USA: International Society of Automation, 2010.

[44] E. D. Knapp, *Industrial Network Security: Securing Critical Infrastructure Networks for Smart Grid, SCADA, and Other Industrial Control Systems*, 1st ed. USA: Syngress/Elsevier, 2011.

[45] E. Byres and J. Lowe, "The myths and facts behind cyber security risks for industrial control systems," presented at the VDE Kongress, Berlin, Germany, 2004.

[46] A. A. Cárdenas, S. Amin, and S. Sastry, "Research challenges for the security of control systems," in *Proceedings of the 3rd conference on hot topics in security*. Berkeley, CA, USA: USENIX Association, 2008, pp. 6:1–6:6.

[47] J. Pollet, "Developing a solid SCADA security strategy," in *Sensors for Industry Conference, 2002. 2nd ISA/IEEE*, Nov 2002, pp. 148–156.

[48] C. Schwaiger and A. Treytl, "Smart card based security for fieldbus systems," in *Proceedings of the 2003 IEEE Conference on Emerging Technologies and Factory Automation*, vol. 1, Sept 2003, pp. 398–406.

[49] R. Lagner, "Cracking stuxnet - a 21st century cyberweapon," http://www.ted.com/talks/ralph_langner_cracking_stuxnet_a_21st_century_cyberweapon.html, Apr 2011.

[50] N. Falliere, L. O. Murchu, and E. Chien, "W32.stuxnet dossier," Symantec Security Response, Tech. Rep., Feb 2011, revision 1.4.

[51] A. Matrosov, E. Rodionov, D. Harley, and J. Malcho, "Stuxnet under the microscope," ESET, Tech. Rep., 2011, revision 1.31.

[52] T. Novak and A. Gerstinger, "Safety- and security-critical services in building automation and control systems," *IEEE Transactions on Industrial Electronics*, vol. 57, no. 11, pp. 3614–3621, Nov 2010.

[53] W. Granzer, F. Praus, and W. Kastner, "Security in building automation systems," *IEEE Transactions on Industrial Electronics*, vol. 57, no. 11, pp. 3622–3630, Nov 2010.

[54] J. Åkerberg and t. y. m. v. n. p. Mats Björkman, booktitle=Proceedings of the 2009 IEEE Conference on Emerging Technologies Factory Automation.

[55] V. M. Igure, S. A. Laughter, and R. D. Williams, "Security issues in SCADA networks," *Computers and Security*, vol. 25, pp. 498–506, 2005.

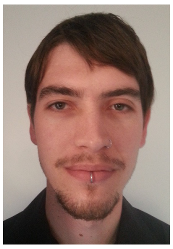
[56] R. C. Parks and E. Rogers, "Vulnerability assessment for critical infrastructure control systems," *IEEE Security and Privacy*, pp. 37–43, Nov/Dec 2008.

[57] M. Cheminod, A. Pironti, and R. Sisto, "Formal vulnerability analysis of a security system for remote fieldbus access," *IEEE Transactions on Industrial Informatics*, vol. 7, no. 1, pp. 30–40, Feb 2011.

[58] M. Cheminod, I. C. Bertolotti, L. Durante, P. Maggi, D. Pozze, R. Sisto, and A. Valenzano, "Detecting chains of vulnerabilities in industrial networks," *IEEE Transactions on Industrial Informatics*, vol. 5, no. 2, pp. 181–193, May 2009.

[59] E. J. Byres, M. Franz, and D. Miller, "The use of attack trees in assessing vulnerabilities in SCADA systems," 2004.

[60] D. Geer, "Security of critical control systems sparks concern," *Technology News*, pp. 20–23, Jan 2006.



Brendan Galloway Brendan Galloway (B.Eng) is currently employed as a control system engineer at a South African utility company. He received a Bachelors degree in Computer Engineering at the University of Pretoria (South Africa) in 2008 and is currently pursuing an Honours degree in the same field. His main interests are in industrial control networks, with specific focus on security and the integration of next-generation technology.

Gerhard Hancke Dr Gerhard Hancke (B.Eng , M.Eng , PhD, SMIEEE, MIET, CSCIP) is currently a Fellow with the Information Security Group (ISG) at Royal Holloway, University of London (RHUL). He received a Bachelor and Masters of Engineering degrees in Computer Engineering from the University of Pretoria (South Africa) in 2002 and 2003, and a PhD in Computer Science for the Security group at the University of Cambridge's Computer Laboratory in 2008. Subsequently, he worked four years for the ISG Smart Card Centre at RHUL as lead researcher/engineer where he managed the RF/Hardware Laboratory and was involved in the evaluation, development and integration of smart card systems. His main interests are the security of smart tokens and their applications, embedded/pervasive systems and mobile technology.