CONTROLLING INFRASTRUCTURE PERFORMANCE: THE THEORY AND PRACTICE OF CONTROL SYSTEMS INNOVATION

Paul Nightingale[*], Tim Brady, Andrew Davies and Jeremy Hall (CoPS Innovation Centre, SPRU and CENTRIM, Universities of Sussex and Brighton, Falmer, UK)[ℵ]

Abstract: This paper explores the role of control systems in the evolution of critical infrastructure. It explains how control systems co-ordinate the flow of goods, traffic, materials, funds, or information through complex supply, production or distribution systems. The paper examines how they increase productivity by improving the utilisation of installed capacity, creating economies of system that are distinct from the traditional economies of scale, speed and scope. A framework is developed that explains what sectors they are important in, and how innovation in component technologies relates to organisational changes. It is illustrated by case studies of three sectors: elevators, telecommunications and investment banking. (Keywords: Innovation, Technical Systems, Control).

Introduction

In a number of sectors – such as air traffic control, telecommunications, and railways– new, IT based control systems are improving efficiency and reliability. This paper develops a framework which identifies which sectors they are important in, and how they improve system performance. The framework suggests control systems are important sectors based around large-scale production and distribution systems or networks, where they improve system performance and achieve system goals (such as safety and reliability) by improving the routing and scheduling of traffic.[1]

This paper develops a new framework for explaining how control systems achieve this capacity utilisation by combining several strands of literature: Firstly, Chandler's research on how firms use control to achieve economies of scale and scope at higher levels of throughput. Secondly, Beniger's research on how innovations in control enabled firms to manage higher levels of output as system's complexity increases. Thirdly, Hughes' framework in which control systems co-ordinate the performance of large technical systems.

These issues are re-examined now because the incorporation of software and microprocessors into control systems in the 1980s and 1990s has fundamentally changed control, improving capacity-utilisation and generating new cost reductions based around economies of system (Davies 1994, 1996).[2] As a result, sectors where control was previously bureaucratic; such as banking, or electrical; such as telecommunications, have been transformed (Davies 1996, Nightingale and Poll 2000). The paper is divided into three parts. In Section 2 we present an integrated framework which explains how control systems facilitating improvements in systems performance. The explanation is then illustrated, in

[*] Corresponding Author, Email: p.nightingale@sussex.ac.uk

[ℵ] The authors would like to thank Mike Hobday, Ed Steinmeuller & Ken Edwards for their advice. The work was undertaken within the ESRC funded Complex Product System Innovation Centre.

[1] Where the system is the unit of analysis and can refer to a manufacturing plant, a logistics network, a utility infrastructure or a firm.

[2] These are distinct from traditional scale and scope economies, for example, if one thinks of a complex, fibre-optic telecommunications network of a constant size delivering a constant range of products, improvements in routing of calls can increase the capacity of the network and therefore reduce unit costs. These productivity increases are not due to economies of speed as traffic is already travelling at the speed of light. Similarly, as the size of the network and the range of products have not changed, they are not due to economies of scale or scope.

Section 3, by case studies of a technological artefact (elevators), a large utility (telecommunications network) and a large service firm (investment banking). The framework and case studies are reviewed in Section 3.

Capacity Utilisation and Control: Chandler, Beniger and the Economics of Throughput

Chandler has shown how the growth of large firms was linked to methods for controlling resources. In industries with the technological and organisational features that allowed fast, high-volume flows to turn low-cost inputs into high-value outputs large firms grew by investing in high fixed-cost technology. However, potential cost advantages could only be realised if high enough levels of throughput are maintained to spread the costs widely. Consequently, administrative co-ordination and techniques for monitoring and co-ordinating of flows in complex organisations were required (1990, p.24).

These techniques had been developed by the railroad and telegraph industries in the 1870s because "*... unless the movement of trains and the flow of goods were carefully monitored and co-ordinated, accidents occurred, lives were lost and goods moved slowly and with uncertainty*" (1992, p.264). Consequently, the railroads developed ways to measure, quantify and control how changes in inputs affect outputs. For example, the *ton-mile* was used to cost the transportation of goods and the *operating ratio* was used to measure operating success. Similarly, mass-market distribution industries developed *stock-turn* to relate turnover to investment. In production industries *earnings-to-sales* was used to measure the effectiveness of operations, while du Pont developed *turn-over* to measure the flow of materials through the production process (Chandler and Daems 1979).

The diffusion of these improved control techniques meant that by 1914:
'...American mass producers had developed techniques to account for the profits resulting from administrative co-ordination and had devised ways to allocate resources systematically. One weakness … was that the flows … were not yet calibrated to quick unexpected changes in short-term demand. This weakness was painfully exposed during the sharp economic recession of 1920-21. Nearly all the mass marketers and producers suffered inventory crisis. *Those companies with the most complex systems of supply, production and distribution suffered the most.*' (Chandler and Daems 1979:39, emphasis added).

Why did the companies with 'the most complex systems of supply, production and distribution' suffer the most? The increases in organisational complexity were consequences of the systemic co-ordination of specialised activity cells. Beniger has argued that as these systems increase in a) size, b) energy consumption, c) organisational and technical complexity, d) utilisation, and e) processing and transportation speed, coherence can break down and unintended interactions make control more difficult - often producing catastrophic failures (1986). These problems of coherence and control meant that rather than exploiting increased economies of scale, per-mile operating costs in railroads of the 1850s *increased* with size (Beniger 1986, p.227). The resulting crises of control, where 'innovations in information processing and communication technologies lagged behind those of energy and its application to manufacturing and transportation', then acted as focusing devices for innovation (1986:427).

Thus, a crisis in safety on the railroads was solved by the development of bureaucratic organisations. Crises in increasingly complex distribution systems in the 1850s were solved by improvements in telephony, the telegraph and postal reforms (ibid.). Crises of control in production in the 1860s produced innovation in the organisation of materials processing (ibid.). Finally, crises of control in consumption and marketing in the 1880s produced innovations in mass media and advertising (ibid.). Beniger suggests that control innovations are fundamental to the growth of large technical systems, and that these systems reach growth-bottlenecks where lack of control makes further cost reductions impossible without creating unacceptable *risks*.

Thomas Hughes (1983) defines technical systems in terms of being: '*...constituted of related parts or components... connected by a network, or structure... The interconnected components ... are often centrally controlled, and usually the limits of the system are established by the extent of this control. Controls are exercised in order to optimise the system's performance and to direct the system towards the achievement of goals... Because the components are related by the network of interconnections, the state, or activity, of one component influences the state, or activity, of other components in the system...*' (1983:5)

This definition highlights the four features of systems that differentiate them from their environment: components, the architecture of their interconnections, the control subsystem, and the overall function. Like Chandler, Hughes gives particular importance to the *increasing returns to adoption* whereby increasing the size of the system, increases the utilisation of capacity (and therefore economic returns) by mixing different customer needs to flatten-out troughs in demand – measured by the 'load-factor'. For example, the electricity utilities in Chicago expanded: '*... to encompass the diversity of loads that brought a fuller round-the-clock utilisation of generating equipment. A utility manager with a peak load caused by rush-hour use of electric streetcars learned not to expand the traction load. Instead, the utility reached out like a tree in a dark forest stretching its limbs into the sustaining sunlight. When sustenance for the load-hungry utility ... was the night-shift operation of a chemical plant, the system's distribution lines reached in that direction. System builders knew that the diversity of load allowed load management, ... and a lowering of the unit costs was likely to be found in a large geographical area where the population engaged in a wide variety of energy-consuming activities.... Expansion was not a drive for undifferentiated size: it was a purposeful move to lower the cost of energy.*' (1983:463)

This expansion required sophisticated control technologies to be developed to allocate electricity to different customers and ensure the load factor was well managed (1983:73). The load factor in turn allowed managers to direct the growth and use of large technical systems to increase capacity utilisation and therefore profits.

The Sectoral Importance of Control

The economic returns on strategic investments in systems or networks depend on capacity utilisation. We would therefore expect control to be important in sectors where large-technical systems produce reductions in unit costs and where control can make a substantial difference to how well this capacity is utilised which depends on:

- Firstly, the size and complexity of the system. In general, large numbers of components, linkages and states generate a greater number of alternative routings through the systems with varying economic characteristics.
- Secondly, how *systemically inter-dependent* the traffic through the systems or network is. i.e., the extent to which the behaviour of one item of traffic has economic implications for other items of traffic. When traffic can be stored and at very low levels

of traffic-flow this inter-dependence is reduced and the benefits of improved control decline.

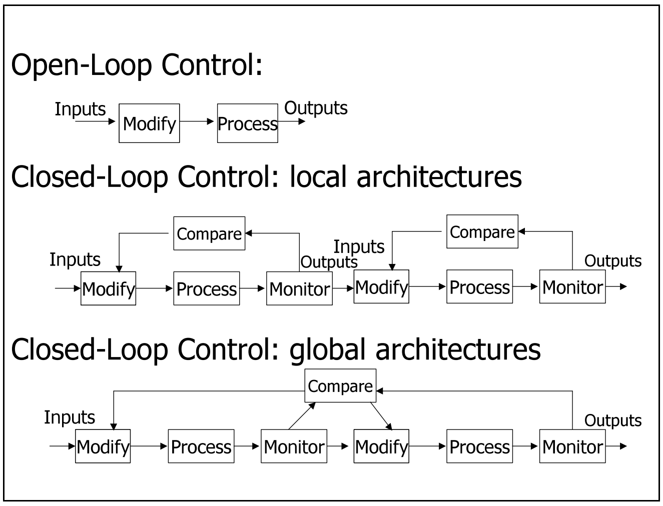
- Thirdly, the relative speed of traffic flow compared to the speed of control. If the time taken to calculate the efficient traffic routing is longer than the speed of the traffic control becomes very inefficient, and can cause the system to oscillate wildly.
- Fourthly, the *balancing* of productivity and reliability.

As a consequence, the greater the emphasis on routing fast moving, rapidly changing, economically inter-dependent traffic through complex, high-reliability, large scale, high load-level, capital intensive networks and systems, the greater the economic emphasis on control innovation will be.

Controlling Systems to Improve Capacity Utilisation

The word control comes from the Latin *contrarotulare,* meaning to compare '*against the rolls*', - i.e., comparing behaviour with the officially prescribed Roman policy written down on rolls of papyrus. When applied to technology it implies modifications that bring actual and intended behaviour together. Control is required when a match between actual and intended performance cannot be reliably maintained. For example, performance might change over time due to changing outputs or inputs;[3] processes might disproportionately amplify small differences in inputs; it might require optimisation or fine-tuning in use; or performance may be too complicated to predict ex ante. In these instances control systems monitor, compare and modify sub-components in a co-ordinated way to ensure that the over-all system behaves as intended. There are two forms of control. Firstly, in open-loop control an actuating device directly controls the inputs to the component and ensures that they are within the correct range required to produce the desired outputs. Secondly, in closed-loop-feedback control the output is measured and a feedback-loop is used to compare this measurement to its intended output and the inputs are modified to optimise the outputs. In these closed-loop control systems a model of the process being controlled is used to mediate the relationship between actual and intended performance (See Diagram 1).

Diagram 1



Improvements in the performance and capacity utilisation of a system are made possible by innovations that improve the accuracy, speed, scope and reliability of the control system.

[3] Changing inputs can corresponds to changes in the environment; changes in the inputs (such as different quality feedstocks), and changes in the components themselves (such as corrosion).

- *Accuracy:* When the processes being modelled are complex simplified models (i.e., a less true models) are often more practical. If the new model can improve the specificity of the changes to inputs it can allow more precise control over outputs.
- *Speed:* The accuracy of the model can also be improved by increasing the speed of calculation. This is particularly important if the time taken for the technologies' inputs to be modified is long relative to the period in which significant changes occur in the system being controlled.
- *Scope:* Control systems can also be improved by expanding the scope of control i.e., the extent to which control is local or global. Control is local when it only covers a single or small number of components within a technical system, while global corresponds to control over a larger number of interdependent components. With inter-dependent processes global optimisation are associated with increased efficiency.
- *Reliability:* There is a trade-off involved between scope, accuracy, speed and the cost of calculation. Larger more global optimisations will require a more complex model, which are often less accurate – creating economic costs, and centralised control architectures create reliability issues if control problems cause extensive rather than local downtime. Moreover, larger optimisation calculations are more time consuming which can create inaccuracies if the system changes state during the time taken for the calculation.

Control systems are typically multi-component and multi-technology with different technologies appropriate in different situations. People are superior to machines at understanding complex relationships between inputs and outputs and at working-out where the implicit model diverges from the system being controlled. This is why high-reliability-systems, such as air traffic control, nuclear power stations, trading rooms and ballistic missile systems have manned control rooms. Electronic control systems on the other hand are generally better at simulating discrete processes and rapidly performing large calculations – which is why telecommunications switches are electronic. The nature of control is not only dependent on the underlying technologies, but also on how they architecturally relate to the system being controlled.

Changes in the control architecture can have important implications for system performance. In centralised control, system performance is monitored and controlled through a few high-capacity control centres. In distributed control, greater control is exercised at the periphery of the system, typically by human operators and decentralised control technologies. The importance of the control architecture creates an important distinction between control systems that monitor systems carrying physical traffic and ones monitoring information traffic (see table 1). In systems where the traffic is physical, such as trains on a network, an *additional* information gathering system must be placed on top control the process. The two systems architectures must be optimised separately. However, when the traffic through the system is information, such as telephone data packets or banking transactions, the *same* infrastructure can be used to carry the traffic and the control function. This allows the performance of two systems to be *optimised together* and new more efficient architectures to be introduced, producing radical increases in capacity and utilisation. These productivity increases are distinct from traditional economies of scale, scope, and speed and have been termed economies of system (Davies 1994, 1996).

Table 1. Examples of the distinction between Physical and Informational Traffic

	Air Traffic Control	Telecommunications
Traffic	Physical (planes)	Information (data packets)
Control	Information	Information

This analysis of control technologies suggests that control innovations will increase capacity utilisation if they optimise performance over a larger number of interdependent components, controlling a larger more complex system, or increase the accuracy of the model, the speed of control or its reliability. The previous analysis of Chandler, Beniger and Hughes suggests that these improvements increase capacity utilisation by firstly, overcoming control constraints on the size of the system. This increases the scale of system being controlled and reduces unit costs. And secondly, by increasing the number of products and services being provided. This reduces unit costs by having a better economic mix and load diversity.

The overlap between the two levels of analysis suggests that further improvements are possible based around economies of system:

- Firstly, because improved control allows a more timely and specific matching of inputs to outputs, it can improve the speed and routing of traffic, increasing capacity.
- Secondly, increased precision in the routing and processing of traffic produces a greater range of possible outputs. For example, improved control over a chemical plant can be used to adjust the inputs to its processes more selectively and producing a larger range of more specific products. The control systems can then be used to optimise the economic-values of those new outputs and thereby increase productivity.
- Thirdly, innovation in control systems can allow performance-improving changes in systems' architecture.

In short, the framework would suggest that control innovation produces improvements in the routing of traffic, which allows new products and services to be introduced, and improved control enables the (architectural) optimisation and management of these outputs.

Section 4: Illustrative Case Studies

Control System and the Provision of Vertical Transportation

This first case study explores how improvements in the routing and scheduling of elevators in tall buildings has increased the capacity of vertical transportation systems, allowing *fewer* elevators to maintain the *same level of capacity*.[4] Innovation in elevators is important because providing sufficient elevator capacity without large numbers of elevator shafts filling rentable space on the lower floors has become the main constraints on the height of tall buildings and today depends on control technologies. A tall building requires 2 elevators for every 3 floors, so a 90-story building needs about 60 elevators (Lacob 1997:106).

Large vertical transportation networks can be thought of as systems that take passengers as their inputs and transfers them to different floors. The capacity of elevator systems can be increased in two main ways - physically increasing capacity or improving control. Increasing the number of elevator wells or the number of people each elevator holds reduces the amount of rentable space. Double-decker elevators serving alternate odd and even floors (i.e., the Time-Life building in Chicago) can increase capacity but have inherent limitations. As does increasing the speed of elevator cars, which requires large amounts of energy and can cause nausea.[5] Consequently, research now involves finding ways to de-couple elevator cars from their ropes allowing cars to share the same network of shafts.

[4] Rather than the same number of elevators serving a larger population.

[5] The world's fastest elevators, such as the Mitsubishi elevator in Yokohama Landmark tower, have top speeds of around 40ft per second. The elevators in the Sears Tower in Chicago have been

The main way that capacity is increased is by improving the routing of cars and specialising their functions (Barney et al 1985). This reduces waiting times, minimises travel distances, decreases costs and increases the number of passengers processed. Since the time and distance travelled depend on the number of stops and direction changes, the most obvious method of increasing passenger throughput is to control passengers entering the elevator so that they are all travelling in the same direction - going up or going down. Passengers then share their rides and avoid travelling to undesired floors. Similarly, express lifts can be used to move passengers to sky lobbies during the early morning demand up peak where they disembark and take slower lifts to their local floors - allowing shorter lift wells serving a smaller number of floors.

When elevators were first introduced in the 1860s concern about public safety meant that elevator attendants manually controlled each elevator. In 1925, the first automatic control system was introduced in St. Luke’s Hospital in Chicago and by the 1930s a group dispatcher system was developed where the position of elevators within the building was monitored manually and the closest elevator assigned to the call. This reduced waiting times, energy consumption and increased the speed of transfer. The first automatic elevator was introduced in 1948. In the 1950s automatically controlled elevators were introduced using relay technologies. These allowed a *collect and select* strategy for individual cars whereby passengers would choose their own floors. The technology was initially introduced to maintain a reliable service in response to an elevator attendants' strike in Chicago which had paralysed the city’s tall buildings.

The introduction of solid state electronics in the 1960s and microprocessors in the 1980s allowed group control to be automated. This initially involved controlling two cars as a team but gradually the number of cars increased as improvements in the speed and accuracy of the control systems allowed a more global optimisation of performance. Group control systems evolved so that passengers would not be automatically assigned the nearest car if this would have adverse effects on throughput. Throughput is most important in the early morning up-peak, just before the working day starts, when most passengers want to move from the central lobby into the higher floors. In this period local zoning can be used to allow different elevator cars to be assigned to different floor zones, reducing the number of stops, which increases the speed of the journey and increases capacity.[6]

Group control also allows the allocation of cars to be optimised according to variables such as waiting time, journey time, energy use etc., and new services to be introduced, such as floor access restrictions in dual use buildings. For example, in the Hancock centre (Chicago) commercial and residential passengers can be separated ensuring a more secure environment. The overlaying of microprocessor technologies over older solid state relays in the mid-1980s typically decreased average waiting time by 25% (Beebe 1995). This reduced the time taken to calculate optimal car allocations and allowed the intelligent gathering and use of data on passenger throughput which could be use to reduce the number of elevators required to process a given population of users. Over the 1980s control evolved from 8-bit systems that ensured an equal interval between call and arrival, to 16-bit predictive control systems that minimised waiting time. By 1985 energy saving systems were introduced and the 1990s saw

slowed down due to passenger complaints. Japanese firms are attempting to overcome the problem of inner ear distortions at high speeds by pressurising the elevator system.

[6] Elevators returning from higher floors will then bypass passengers waiting to go down in order to bring passengers up faster.

the arrival of expert system and genetic algorithm technologies that allowed individualised floor customisation and produced a further 15-20% increase in capacity. Between 1977 to 1997 typical control systems increased in memory 1000 fold from 10kb to 10Mbytes.

In the 1990s the introduction of 32-bit control systems allowed real time distance reduction calculations to be performed fast enough not to effect service performance. This increasing capacity utilisation by optimising the behaviour of the system in real time, rather than the system as it was when destinations were selected. Cars can be moved to pre-empt traffic and zoned during idle periods to minimise travelling time. Current technologies enable the cars to be fully grouped using fuzzy logic and artificial intelligence software. This predicts and optimises passenger throughput based on passenger data that is updated in real time and produced 5-10% improvements in maximum capacity.

New control techniques, such as Dynamic Zoning, have continued to produce capacity improvements (12% by simulation) in the 1990s (Chan et al 1998). Improved AI technologies in the late 1990s have allowed group control systems to move *beyond real-time control* and predict through-put. Modern control systems are able to learn patterns of use, and assign cars as required as well as allowing increasingly tailored service (Sond Liu 1996). For example, if a meeting is held at the same time and place each week, the control system will learn to have a car ready.

These increases in capacity have been used to reduce the number of elevators required to service a given building population. As the capacity of the whole elevator system became dependent on the control system, it became bundled in, and critical for service provision, creating an increased emphasis on reliability. Now that improving the allocation of cars is reaching its performance limits, the ability to maintain throughput is dependent on maintaining reliability. Consequently, control systems now monitor system performance and alert maintenance of any impending problems – an architectural shift from reactive to proactive maintenance. In 1987 Hitachi introduced a MAS (Maintenance Auto Station) system for monitoring elevators on a 24-hour, 365-day basis. In 1988 Otis introduced the remote elevator monitoring system (REM) that monitored the performance of elevators and linked it to a fast-response, 24-hour call out line. Data on elevator use could then be used to statistically analyse mean-time-to-failure so that parts could be replaced during regular night-time maintenance before they fail - decreasing systems down-time.

Since effective maintenance depends on the statistical analysis of performance, it is improved by having a larger number of elevators in the sample. As a consequence, the locus of control has moved from the car, to the group, to the building and on to groups of buildings. A system developed by Hitachi in 1994 minimises maintenance downtime by remotely monitoring the behaviour of 17,000 elevators. This HERIOS (Hitachi Elevator Remote and Intelligent Observation System) measures the performance of some 42 items within the lift system and analyses their failure rates to provide pre-emptive maintenance. This data is used to control and classify maintenance work on site. Since some 40% of all elevator maintenance in microcomputer controlled elevators involves monthly checks on control equipment the ability of HERIOS to carry out this remotely significantly reduces the downtime, provides real time monitoring and prompt failure recovery.

Control Systems and Innovation in Telecommunication

This second case study analyses how telecommunications firms have exploited control systems to increase the profitable capacity utilisation of telecommunications networks. Telecommunications networks are systems that take voice or data traffic as their inputs and outputs and transfer them between customers. Since telecommunications traffic cannot be easily stored, effective and reliable control and management of the load factor is essential to improve capacity utilisation and ensures that the high fixed cost network is being used as fully as possible (Davies, 1994; Davies, 1996).

Control systems in telecommunications have evolved through three main phases: a centralised control hierarchy based on analogue transmission between the 1880s and 1980s, a centralised control hierarchy based on digital transmission in the 1980s and 1990s, and more recently through then a shift towards internet protocol (IP) technologies that allow more decentralised routing and control of traffic. This case study will show that control systems have allowed significant increases in capacity utilisation by increasing the accuracy, speed, scope and reliability.

The telecommunications systems that existed between 1880 and 1984 communicated telegraph, telex and voice telephony by transmitting electrical signals in analogue waveform. The high-cost of switching relative to transmission produced a centrally-managed hierarchical system in which millions of telephone terminals were linked vertically to increasingly complex and higher-capacity local and long-distance switches, and then on to trunk exchanges at the apex of the hierarchy. Telephone networks were designed to have sufficient installed capacity to accommodate peak traffic volumes - or the load factor, with the load on the network defined as the call arrivals per second multiplied by the average call duration in seconds.[7] The load factor could be used to graphically show the utilisation of capacity and related unit costs over a given period of time.[8]

Human operators and electromechanical circuit switches were used to route traffic through local exchange operations, while long-distance traffic was routed through higher-capacity switches between local exchanges. Circuit switches kept the circuit open for the duration of the call and used the same transmission path as the message to control the movement and charging of traffic. As a consequence circuit switched networks are heavy users of network capacity and it was impossible to connect more than five links in tandem without the quality of the analogue signals deteriorating. In the mid 1980s increases in the volume of traffic and the variety of different services carried produced changes in network management and control. New digital control technologies were introduced between 1984 and 1994 which reinforced the centralised, hierarchical architecture of the older circuit-switched telephone system. They were used to control traffic between local and long-distance operations with digital switches at lower levels in the hierarchy becoming distribution points for communication to the centre.

In the late 1980s traffic volume increased and service reliability became increasingly important. As a consequence capacity utilisation was improved by reducing the cost of centralised control using two related technological innovations. The first involved improving the routing of digital traffic by removing control information from the message and transmitting it along a separate high-speed packet switching network - called Common Channel Signalling System No. 7. With this new technology several hundred circuits could

[7] The unit of load known as the erlang, after A.K. Erlang, the inventor of traffic theory.

[8] The actual installed capacity was greater than the actual peak traffic requirements to prevent requests for services exceeding the capacity of the system.

be controlled by a pair of signalling channels along a particular route. This increased capacity by speeding up call set-up times from 20 seconds (for an average 3-minute call using the old analogue signalling system) to only 3-4 seconds on the digital signalling system. This allowed valuable local and long-distance switching capacity to be used by other calls even though the telephone conversation had opened up a transmission circuit.

The second control innovation was the introduction of the intelligent network (IN) and its high speed signalling system, following research by Bell Communications Research (Bellcore). Intelligent networks changed the control architecture and enabled traditional telecom carriers to improve network performance and offer new customised services rapidly and efficiently. The Intelligent Networks use databases to provide centralised control of the routing, storing and manipulation of traffic and services for traditional operators, with control software in databases called service control points (SCP). The SCP determine how calls should be charged and transmitted and increase capacity by using a high-speed signalling system to interrogate the network databases for instructions about routing, while the ordinary voice and data messages are processed by the switch. The IN incorporates operation support systems capabilities, such as planning, maintenance and provisioning functions, which assist in managers in monitoring network performance and clearing service problems. These provide centralised management of traffic and allow new services to be delivered to residential (e.g. call forwarding) and corporate customers (e.g. virtual private networks). Moreover, the high-speed signalling system architecture of the control system in the IN provides increases in capacity utilisation.

The introduction of control systems on top of installed networks reinforced the network architecture of the previous analogue switching technology. By the early 1994, problems with mixing telephony, voice, data and images in centralised telecommunications networks, coupled with higher usage, overloaded existing systems. A number of disruptions to major networks were initiated by the installation of new software to increase network capacity.[9]

The development of Internet protocol (IP) packet-switching technologies offered enormous increases in capacity utilisation when accompanied by a shift towards a distributed control architecture. So while packet-switching technologies were introduced on top of old style centralised telephone systems, their real cost advantages were realised in distributed networks that offer many different routes and no centralised communications switch. In these packet-switched distributed networks the message is divided into packets and individually transmitted through multiple routes before being re-assembled at their destination. The packets themselves are automatically routed through the network via routers and bandwidth is allocated so that different messages share the same line.

Optical switching technologies provide the most recent and dramatic improvement in the capacity of telecommunications networks. In the most advanced telecommunications networks, user traffic is transferred in IP packets through optical fibres for high-volume long-distance transmission. But at each switching node, digital signals converted into optical form on a wavelength of light are converted back into electronic form for switching and forwarded to the next node. Although a single wavelength of optical fibre can carry traffic at 155Mbits/s, 2.5Gbits/s or 9.6Gbits/s, improvements in the capacity of optical network are

[9] In January 1990 AT&Ts network crashed when the New York switch became overloaded. MCI WorldComs network crashed in August 1999 after new software was installed. BTs network crashed in February 2000 after three control access gateways failed independently.

held up by the conversion into electronic format for switching.[10] Therefore, before end-to-end optical signalling is possible this switching bottleneck will have to be overcome and all devices in the network will have to work under heavy load conditions, placing greater importance on effective control architectures.

As modern economies become increasingly dependent on telecommunications services, optical networks offering massive bandwidth capacity and high-speed connections to services must be reliable and secure. Internet services have changed the usage of networks producing huge congestion, making traffic unpredictable, because of large peaks and troughs in demand. To overcome such bottlenecks, the focus of research has shifted from supporting centrally-managed traditional circuit-switched systems to developing distributed IP networks capable of dynamically changing to meet traffic needs (Awde, 2000). Genetic algorithms are combined to create control systems that are self-adjusting, so that traffic can be automatically re-routed from congested to free parts of the network. Intelligent, software-controlled IP networks can make decisions about traffic routing. Optical switching allows bandwidth to be remotely switched to a given node for a specified period to accommodate fluctuations in demand. Switches located at key nodes in the network feedback local intelligence to the network management centre where traffic is automatically controlled indicating a new control architecture is being born that can increase overall available capacity by 30-50 per cent, giving large cost advantages to new entrants.

Risk Management Systems in Investment Banking

This case study explores how investment banks use risk management control systems when trading securities such as bonds, equities, and derivatives. Investment banks can be thought of as systems that take financial products (funds, bonds etc.) as their inputs, un-bundle[11] them into their component parts, and financially engineer the parts to create contracts (the outputs) that are traded within a heavily regulated marketplace (Nightingale and Poll 2000). The profitability of the contract is determined by its size, margin and risk, where the risk reduces the value of the transaction to take into account the likelihood and distribution of any losses.[12] Risk management systems are used to control the allocation of resources, price products accurately and monitor and control how their values change over time.

These systems comprise IT hardware, sophisticated mathematical models and organisational auditing and control processes that ensure the control measures are in place and are being properly used. The nature and size of contracts being produced depends on balancing likely profits and losses. Each transaction has a safety-margin where the risk is acceptable and risk management systems are used to ensure that the *calculated* risk position corresponds as closely as possible to the *real* risk positions so that *actual* profits come closer to the *potential* profits. Thus capacity is dependent on calculating prices and risk, while utilisation is dependent on how well that risk can be approximated and managed.

Up until the late 1980s risk-analysis was largely bureaucratic. During the 1990s risk-analysis was transformed by advances in financial theory and new information and communications technologies. In particular, the development of the theory behind financial arbitrage by Modigliani and Miller and the development of models for pricing contingent

[10] "The effect is that the capacity of optical systems is expected to grow more rapidly than Moore's Law describes for electronics – that is a doubling every 18 months or less" (Gannon, 2000).

[11] Unbundling refers to the separation of contracts into their component risk and profit distributions.

[12] See Nightingale and Poll (2000) for details.

claims, such as the Black-Scholes-Merton option pricing model. Calculated risk can be brought closer to actual risk in four main ways:

- Firstly as trades are highly interdependent (or covariant), the bank's risk exposure is dependent on all the trades the bank is party to and is improved by the *widening the scope* of analysis and managerial control. If risk exposure is analysed and controlled locally in London, New York and the Far East the bank may still make losses if all three find themselves exposed on the same trading position.
- Secondly, because market movements make the calculated and actual risk positions diverge, *increasing the speed* of analysis reduces the uncertainty surrounding the risk exposure probability-distribution and allow traders to get '*closer to the edge*' increasing capacity utilisation.
- Thirdly, because the value of a contract is dependent on a range of interacting and changing variables, improvements in the *accuracy* of the models used to calculate pricing strategies can reduce uncertainty, increase profits and allow risk to be analysed at finer resolutions. Part of this processes has involved the development of more sophisticated mathematical models of pricing strategies and the typical natural traders of the 1980s are now being replaced by PhD physicists and mathematicians acting as quantitative analysts who can create more specific, tailored products.
- Fourthly, since the ability to value and price contracts is dependent on the ability to analyse risk, there is a great emphasis on ensuring systems reliability.[13]

In the late 1980s investment banks introduced small-scale control systems to monitor and automate pre-existing risk-management processes. Since the accuracy of risk exposure calculations is dependent on the size of the statistical sample, the control systems were driven towards increases in size as this improved accuracy and allowed traders to manage larger and more complex trades. During the early 1990s the systems began to increase in computing power and grew in size from about 100 to several thousand Unix machines, undergoing various qualitative architectural changes in the process (see Nightingale and Poll 2000).

The most important of these changes in architecture was the transfer of risk analysis from back office mainframes that had computed risk positions overnight, to the front-desk Unix systems which computed risk positions at the end of the day. This produced two main benefits. Firstly, it increased the speed of control and created a more accurate match between the calculated and actual risk position increasing profits. Secondly because risk information was available at the end of the day rather than the following morning, it allowed exposure to be actively managed and passed around the world from London to New York to the Far East and back to London 24 hours a day. This gives traders a more global understanding of the bank's exposure, which they can actively manage while the local markets shuts down overnight, generating substantial economies of system unobtainable to banks operating as local entities.

Architectural innovations also occurred in the control technologies themselves. Their importance made risk-management systems business critical as their failure can produce substantial losses. As a consequence ensuring their reliability is vital for ensuring capacity utilisation and additional economies of system can be found in maintenance. Typically a centralised 24-hour systems monitoring base will analyse the behaviour of the infrastructure

[13] This issue became important in the SWAPs markets during the 1980s where a number of high profile (and an even larger number of low profile) traders collected large bonuses on the basis of large profits, before leaving firms with extensive hidden exposures.

and allow the bank to move from reactive to proactive maintenance. For example, if a problem is found when the London markets open it can be fixed locally and then by logging-on to the global system changes can be made so that the same problem does not confront traders in New York when they start work. This reduces the amount of down time where traders are uncertain about how to price their trades and therefore increases capacity utilisation.

The sophisticated database technologies and mathematical models embedded in modern risk management systems allow complex contracts to be priced and then continuously valued as the economic environment changes. These technologies allow bankers to aggregate large numbers of financial contracts and then unbundled into their separate risk classes, before regrouping and financially engineering the resulting components. These fine grain distributions can be used to produce novel products with more precise characteristics that better match customer requirements. They also allow improved internal capital allocations, more precise loss reserves and more accurate risk monitoring across various levels of aggregation. Since *the ability to do this accurately* depends on *the size of the sample* the process is subject to increasing returns. The larger the sample, the more accurate the model used to control how the unbundled risk profiles are matched to customer requirements will be. As a consequence large banks with sophisticated systems and large amounts of data can produce more sophisticated products than smaller banks.[14] The top 10 largest investment banks now corner over 90% of the world's cross-boarder securities market, and the size of the OTC and exchange traded derivatives market grew from about $10 trillion in 1990, to about $68 trillion in 1998 (BIS 1999).

Risk information can also be aggregated at various levels and used to analyse the banks' performance and risk and banks have developed metrics to quantify and measure system performance. This can be seen in the changing terminology used to describe risk exposures, such as Value-at-Risk (V-a-R) and Risk-Adjusted-Return-on-Capital (RAROC). These allow risk to be analysed in the back-office at various levels of aggregation, and provide ways for senior officers to monitor and compensate their staff. Thus, investment banks can now be managed as socio-technical systems.

Conclusion

This paper has examined the economic role of control systems in increasing the capacity utilisation of elevators, telecommunications networks, and investment banks. Control innovation is important in sectors that are dependent on the routing of fast moving, rapidly changing, economically interdependent traffic-load through complex, high-reliability, large-scale, high-load systems and networks, because it enables large reductions in unit costs without expanding the scale or scope of systems.

The framework suggested that control systems increase the capacity utilisation of high fixed-cost networks and systems by bringing *actual* performance closer to *potential* performance by monitoring component outputs and adjusting component inputs. There are four main paths to improved control: Firstly, improving the accuracy of the model, as when elevator control systems used new genetic algorithm models. Secondly, improving the scope of control and moving to a more global (rather than local) control architecture, as when investment banks shifted from local to global control and management of risk. Thirdly,

[14] The process is similar to the improvements in products that come from introducing more precise machine tools in manufacturing.

improving the speed of control, as when investment banks use real-time risk analysis rather than relying on day old data. Fourthly, improving the reliability of control, as when elevator firms shifted from reactive to proactive maintenance by monitoring car performance in real time.

These innovations in control technologies in turn produced changes in the systems they were controlling. The analysis of Chandler, Beniger and Hughes suggested that improvements in control would allow systems to increase in size and complexity and be able to deal with a wider range of product lines or services. This suggestion was confirmed in all the case studies. However, Davies (1994) suggested that a whole range of new economies of system that are distinct from the traditional economies of scale, speed and scope may also be found. These included:

- Improvements in the utilisation of systems capacity that follow improvements in the routing of traffic through complex systems. This was demonstrated in the telecommunications and elevator sectors.
- The ability to more specifically control changes in the inputs to components allows systems to generate a wider range of routings. These can then be used to generate new products and services, as the improved ability to price and measure the risk of financial products allowed banks to produce innovative and more specialised products that better matched customer requirements.
- Improvements in control systems increased the reliability of systems performance, as when elevator companies used statistical data on performance to feed into their R&D.
- Changes in control allowed radical changes in architectures, as when telecommunications networks moved from centralised to decentralised control.

This has only been a initial positioning paper that attempts to explore the role of control system innovation in the evolution of infrastructure technologies. Further work will explore further innovation in what is largely an invisible part of the infrastructure of the new economy.

References

Awde, P. "'Intelligent' systems are set to become self-managing', *Financial Times,* 15 March 2000.

Barney, G. C., Ed *Elevator Technology* 6, IAEE, 1995

Barney G. C., and S. M. Dos Santos Ed (1985) *Elevator Traffic Analysis and Control* IEE, London

Baxter, A., (1996), A New Mission for Control, *Financial Times*, 13th September

Beebe, J. R., (1980) *Lift Management*, unpublished PhD Thesis, UMIST, Manchester, UK

Beniger, J.,(1986), The Control Revolution' Harvard University Press: Cambridge MA

BIS (2000), Annual Report, Bank of International Settlements, Geneva

Chandler, A. D.,, (1990), *Scale and Scope: the Dynamics of Industrial Capitalism* Belknap Press

Chandler, A. D., (1992), Corporate Strategy, Structure and Control Methods in the United States During the 20th Century *Industrial and Corporate Change*, 1,2,263-284

Davies, A., (1994) *Telecommunications and Politics,* London, Pinter

Davies, A., (1996), Innovation in Large Technical Systems *Industrial and Corporate Change* 5, 4, 1143-1180

Hughes T, (1983), *Networks of Power: Electrification in Western Society 1880-1930*, Baltimore, Johns Hopkins

Hughes T, (1987), The Evolution of Large Technical Systems, in W.,E Bijker, Thomas, P Hughes and Trevor J, Pinch, *The Social Construction of Techn0ological Systems.,* MIT Press, Cambridge Mass

Lacob, M., (1997) Elevators on the Move, *Scientific American*, October, 106-108

Nightingale, P., and Poll, R., (2000), Innovation in Investment Banking: The Dynamics of Control Systems in the Chandlerian Firm, *Industrial and Corporate Change,* 9, 113-141

Nightingale, P., and Poll, R., (2000b), Innovation in Services: The Dynamics of Control Systems in Investment Banking, in '*Innovation Systems and the Service Economy*, Eds: Metcalfe and Miles, Kluwer Academic Press,

Sasaki, K., Markon, S., ahnd Makagawa, M., (1996), 'Elevator Group Supervisory Control System Using Neural Networks', *Elevator World*, 2nd, 1st

Shepard, S., (1994), ‘Traffic Control in Over-Saturated Conditions’, *Transport Reviews,* Vol. 14, 1, 13-43
So, A. T. P., and Lui, S. K., (1996), ‘Advanced Elevator Technologies’, Elevator World, July, 13-17
Yates, J., (1989), ‘*Control Through Communication*’, John Hopkins University Press, Baltimore, MD.